Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

Experimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples

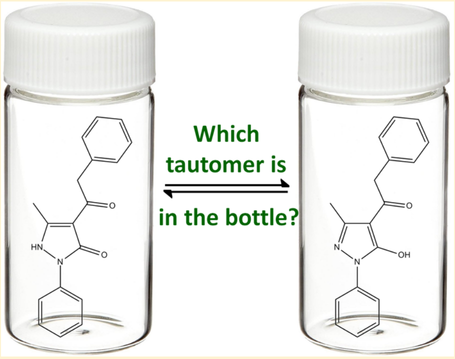
Laura Guasch,[†] Waruna Yapamudiyansel,[†] Megan L. Peach,[§] James A. Kelley,[†] Joseph J. Barchi, Jr.,[†] and Marc C. Nicklaus*,[†]

[†]Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland 21702, United States

[§]Basic Science Program, Chemical Biology Laboratory, Leidos Biomedical Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland 21702, United States

Supporting Information

ABSTRACT: We investigated how many cases of the same chemical sold as different products (at possibly different prices) occurred in a prototypical large aggregated database and simultaneously tested the tautomerism definitions in the chemoinformatics toolkit CACTVS. We applied the standard CACTVS tautomeric transforms plus a set of recently developed ring−chain transforms to the Aldrich Market Select (AMS) database of 6 million screening samples and building blocks. In 30 000 cases, two or more AMS products were found to be just different tautomeric forms of the same compound. We purchased and analyzed 166 such tautomer pairs and triplets by ^{1}H and ^{13}C NMR to determine whether the CACTVS transforms accurately predicted what is the same "stuff in the bottle". Essentially all prototropic transforms with examples in the AMS were confirmed. Some of the ring−chain transforms were found to be too "aggressive", i.e. to equate structures with one another that were different compounds.



INTRODUCTION

Chemical and pharmaceutical companies as well as government agencies and larger projects in academia (such as dedicated screening centers) maintain compound registry systems as a central part of their compound management efforts. Such systems typically consist of a database of all compounds of interest to the organization linked to company reports, biological screening data, stock numbers in warehouse storage, external as well as intracompany shipping data, the compounds' status in the drug development pipeline, etc. Usually, newly synthesized compounds are added to the database one-by-one directly by bench chemists. At the other end of the scale, company mergers may require combining structure databases that each number in the millions of molecules.

An important issue is how to ensure that chemical structures added to the registry system are represented correctly and that possible duplication of structures is recognized immediately. For a bench chemist, the most acceptable visual representation of a chemical compound is a two-dimensional plan of the three-dimensional structure following the valence bond model.[1] However, this representation has many limitations for chemoinformatics tasks for which well-defined data structures are needed and ambiguities in the interpretation of the structure representation have to be avoided. For storing and retrieving information about chemical compounds with a computer, chemical structure diagrams are typically transformed into linear strings of characters or into two-dimensional matrices listing all the atoms and their bonds. The chemoinformatics representations and identifiers most widely used today are MOL/SD,[2] SMILES strings,[3] InChI and InChIKey,[4,5] and CAS Registry Numbers (CAS RN).[6] (For the distinction between connection table-type and identifier-type chemical structure representations, see, e.g., refs 7 and 8.) Of course, these representations themselves have, to a varying extent, limitations for expressing the full chemical and physical understanding of a molecule when compared to a more complete molecular orbital-based description. However, the calculation of identifiers is a very fast and efficient process that can therefore be applied to very large numbers of compounds.

Identifier calculation involves some degree of structure "normalization" in the conversion of a two-dimensional chemical sketch into a linear identifier, and structure registration systems vary considerably in how rigorously they approach this task. This step can be quite complex because there are different ways of drawing and handling tautomers, salts, charged species, stereoisomers, etc., in the computer representation of molecules. General structure checks and normalization steps include comparing the molecular formula with the structure, standardizing functional groups as well as bonds to metal atoms, and adding hydrogen atoms. This is in preparation for the very important next step, which is to check whether the compound truly is new or is already present in the database. After all, resynthesizing a compound that is already

Received: June 9, 2016
Published: September 26, 2016

DOI: 10.1021/acs.jcim.6b00338

available in the organization's repository or can be commercially acquired is typically a waste of resources. Likewise, misassignment of a structure to a sample, whether based on tautomerism or other factors, can lead to serious consequences in the commercial context.[9] Depending on whether the registration system is structure- or sample-centric, a registry number is assigned to (only) a new compound, and supplementary data such as its melting point is added. Other publicly accessible (free or commercial) databases such as ChemSpider,[10] the Beilstein/Reaxys database,[11] PubChem,[12] the Chemical Abstracts Service (CAS) REGISTRY,[6] and ChEMBL[13] use the same structural registration principles in systems that collect compounds from published literature, patents, supplier catalogs, or other sources.

To reiterate, one of the most important components in the registration process is the correct handling of uniqueness of the chemicals represented in the database. Uniqueness is in fact a nontrivial concept in chemoinformatics. One of the major issues in this context is tautomerism: the existence of multiple possible forms of the same molecule that are capable of interconverting via an intramolecular movement of atoms, typically a hydrogen atom (thus termed prototropic tautomerism). There are other, rarer, types of tautomerism such as valence tautomerism that are not discussed here. The structure normalization and registration process can (but does not always) include a calculation of the "canonical" tautomer for a compound.[14] Proton migration can be accompanied by the formation of new, and/or breaking of existing, rings, in which case it is usually called ring−chain (RC) tautomerism.[15] The equilibrium of these reactions is strongly dependent on environmental factors such as pH, temperature and solvent. Additionally, small amounts of acid, base, water, or other catalytically active impurities in the sample can greatly affect the equilibration rate. Tautomeric equilibration times can therefore range from subsecond to months, which makes time-on-the-shelf an important additional parameter in the discussion and handling of tautomerism for real samples. In fact, a sample can be a mixture of tautomers, and thus the registered compound may be better described by a ratio of tautomers than by a single tautomer.

It needs to be emphasized at this point that tautomerism, by virtue of its nature as a chemical reaction involving bond breaking and formation, is really a quantum-mechanical (QM) effect. As such, it can in principle only be accurately handled computationally with molecular orbital calculations. With current software, however, it is entirely nontrivial to incorporate the above-mentioned environmental conditions in QM runs. Additionally, such QM runs can easily take days to weeks for a single molecule, even on modern hardware. This is obviously not a feasible approach for large databases, where one has maybe one second on average to process each entry—including its tautomeric analysis! Instead, rapid chemoinformatics approaches are typically used in practice. These approaches are rule-based and employ mathematical methods (often based on graph theory for operating on connection tables) rather than being derived from physical first principles. It has to be clear that the best that can currently be achieved by these rule-based approaches is that they will be "correct" (if they could be compared with accurate experimentation and/or QM computations) only in a statistical sense, i.e. for most but not for all cases; and that examples can most likely be found, or constructed, for which these rules give a thoroughly wrong answer.

Several chemoinformatics tools exist that can enumerate all possible tautomers, generate a canonical tautomeric form of a compound, and recognize tautomerism (and handle it appropriately) in structure and substructure searches.[16]

In other words, it should not matter which tautomeric form is used as a search query because the software should recognize, and account for, the possibility of tautomerism in the compound. However, it is possible that such rules for the enumeration of tautomers may be too aggressive and not realistic from an organic chemists' viewpoint,[17] i.e. they may declare structures to be tautomers which in reality have a high energy barrier for interconversion and can be isolated as different, stable compounds. Also, the rule set may not cover all types of tautomerism. To the best of our knowledge, such rules are not usually based on, or verified by, specific experimental analyses. Handling tautomerism well has been shown to significantly impact the success of drug design,[18] but only a few experimental observations of tautomerism explicitly conducted in the context of chemoinformatics have been reported.[19,20] This paper aims to provide experimental verification of the chemoinformatics-based handling of the tautomerism of a set of more than three hundred compounds.

Tautomerism is not a rare phenomenon in databases. Based on our chemoinformatics approaches, we found, in a previous study, that prototropic tautomerism is possible for more than two-thirds of the unique structures in our Chemical Structure Database (CSDB), an aggregated database of over 103 million chemical structure records.[17] In a more recent study, we found ring−chain tautomerism to be possible for more than 8% of structures in the AMS database and for an average of 16% of compounds in a set of natural product and approved drug databases.[21] Even earlier studies had pointed out that commercial databases contain pairs of tautomers registered under different catalog numbers which may even be sold at different prices.[22]

Here we present a comprehensive study to evaluate the tautomerism overlap in a commercial database. It consists of a chemoinformatics analysis to detect pairs (or larger multiples) of tautomers of the same molecule, followed by a ^{1}H and ^{13}C NMR spectroscopy analysis for the purpose of experimental validation. This is the first time to our knowledge that such a study has been conducted. The goal of this analysis is twofold: (1) to investigate how many cases of the same chemical being sold as different products (at possibly different prices) occur in a large aggregated screening sample database that is presumably representative of other such databases offered elsewhere and 2) to test, and possibly experimentally validate or reject, the tautomerism definitions in the chemoinformatics toolkit CACTVS.[23,24] Apart from the general interest we hope this analysis will have for the field, analyzing the CACTVS tautomerism rules is of particular interest to us as it underlies much of our chemoinformatics work, including most of the services offered to the public on our web server at https://cactus.nci.nih.gov.

While our experience has shown that CACTVS provides one of the most comprehensive sets of tautomeric transformations among chemoinformatics tools,[17] this does not by itself guarantee that all possible types of tautomerism, or even just of prototropic tautomerism, that have been experimentally observed[25] are covered by the current CACTVS rules. While the approach taken in this study had by necessity to be limited to the currently available rule set, investigating whether broadening of the rule set may yet better represent, e.g., compound identity in large databases, will be the topic of future studies.

Finding the optimal chemoinformatics approach to tautomerism is also of central importance to the InChI and InChIKey identifiers.[26] The efforts reported in this study will find application to, and were to some extent motivated by, the IUPAC project "Redesign of Handling of Tautomerism for InChI V2" (Project No.: 2012-023-2-800).[27]

■ RESULTS AND DISCUSSION

Chemoinformatics Analysis. The data set used for this study was the Aldrich Market Select (AMS) database from ChemNavigator/Sigma-Aldrich.[28] The 2012-09 version of the AMS used for this study comprised over six million building blocks and screening compounds available from more than 60 individual suppliers worldwide. The AMS website interface consolidates the catalogs from all suppliers into a single database, applies a check for structure uniqueness, and facilitates the ordering and shipping process by allowing customers to place a single order for compounds and building blocks from multiple suppliers.

NCI/CADD Chemical Structure Identifiers[29] were generated for all structures in the AMS database. These identifiers are based on hashcodes calculated by CACTVS. This family of identifiers allows one to represent a chemical structure with sensitivity turned off or on to the following five chemical features: fragments (F), isotopes (I), charges (C), tautomers (T), and stereochemistry (S). The naming scheme behind these identifier designations has been explained elsewhere.[29] For the present work, we selected the "FICTS" and "FICuS" identifiers (out of the possible $2^5 = 32$ possible variants). The FICTS identifier is a very close representation of the original input structure. It is sensitive to fragments (such as counterions), isotopes, charges, and stereochemistry in the input structure as well as to the specific tautomer drawn. The FICuS identifier, in which the FICTS identifier's uppercase letter "T" has been replaced by a lowercase "u" (standing for "unsensitive"), is insensitive to tautomerism (but sensitive to all four other features), meaning that different tautomers are given the same FICuS hashcode. The FICuS hashcode thus comes closest to how a chemist perceives a compound, and it is conceptually similar to the InChIKey identifier (though neither algorithmically nor in format; and the handling of tautomerism is done differently with InChIs[4,5]).

We used the FICuS and FICTS structure identifiers for searching for tautomeric pairs in the AMS database. Basically, a conflict in this context is defined as a set of compounds (most often a pair, but 3-, 4-, or 5-tuples were also observed) in which all members have the same FICuS identifier but different FICTS identifiers. Thus, according to the chemoinformatics analysis, they are the same molecule simply represented in different tautomeric forms. Next, we enumerated all possible prototropic and ring–chain tautomers for the compounds in each conflict, using rules encoded in CACTVS as SMIRKS transforms.[17,21,30] The transforms were applied iteratively to the initial compounds and to all resulting new tautomeric structures until no additional tautomers were found. This process produces a full tautomer network for each conflict, with tautomer structures as nodes/vertices and tautomeric transformations as edges/connections.

A set of 62 869 molecules, which represents 1.09% of the AMS database, was identified as being involved in tautomeric interconversions with other molecules in the AMS database. This percentage is similar to the tautomer overlap rate of up to 0.5% found by Trepalin et al. in commercially available compound collections,[22] and to the overlap rates of between 0 and 2% found for the set of databases comprising CSDB.[17] This suggests that the AMS database is representative of other large databases in terms of its tautomeric duplication rate. The total number of conflicts was 31 155. The vast majority of the tautomeric cases identified consisted of two molecules (i.e., tautomeric pairs). There were smaller numbers of triplets (514 conflicts), quadruplets (21 conflicts), and even one quintuplet. Figure 1 shows an example of a tautomeric triplet involving amidine-imidine tautomerism. We found a subset of 16 cases where different tautomers of a compound were available at different prices for the same quantity from the same chemical supplier, with price differences of up to \$469/g. These cases mainly involve imidazole and pyrazole rings, in spite of the fact that the prototropic tautomerism of imidazoles and pyrazoles is well-known. These tautomeric duplications occur with a limited number of original chemical suppliers, so one wonders if they used compound registration software deficient in this regard and/or lacked appropriate QC for the generated computer databases.

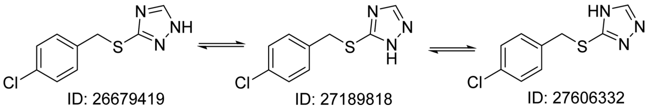


Figure 1. Triplet example of amidine–imidine tautomerism (covered by Rule 5 (Table 1)). The AMS structure ID is shown for each compound.

Tautomerism can change the stereochemistry of a compound through inversion of stereobonds and/or stereocenters.[17] There is, however, no specific tautomeric chemoinformatics rule for interconversion between stereoisomers. The application of one tautomeric transformation can add or eliminate the presence of one stereobond or one stereoatom. However, the application of two consecutive tautomeric transformations can re-establish the stereochemistry of the compound but with the opposite chirality. We observed that 40% of the tautomeric conflicts found in the AMS database involve changes in stereochemistry (which we have termed stereoconflicts[17]). These occurred via two different scenarios: (a) only one stereoisomer of the tautomeric pair has its chiral centers defined or (b) both stereoisomers have their stereochemistry defined but they have opposite stereobonds (E/Z) and/or opposite stereoatoms (R/S). Most of the stereoconflicts were due to an undefined stereobond representation. This paper will not discuss stereochemistry in further detail but these preliminary observations indicate that stereochemistry definitions in commercial databases may still be a serious issue.

Once the tautomeric conflicts were identified, we determined which chemoinformatics rule(s) described each tautomeric transformation. This was done by first enumerating all possible tautomers of each compound by applying two sets of transformations: (a) the default set of transforms available in CACTVS which covers a wide range of common as well as rarer prototropic tautomer transforms[17] and (b) our new set of ring–chain rules.[21] Both sets of rules are listed in Table 1. In addition, a tautomer network for each compound was generated to represent the interconversion pathways between tautomers.

Table 1. Frequency of Observation of Tautomeric Conflicts in the AMS Database Involving the Prototropic Rules and Ring−Chain Rules and the Number of Conflicts Selected from Each Rule for Experimental Evaluation

	tautomerism rules[a]	total conflicts (AMS)[b]	%	selected molecules (AMS)	conflicts
		Prototropic Rules			
Rule 2	1.5 (thio)keto/(thio)enol	731	2.9	24	12
Rule 3	simple (aliphatic) imine	561	2.2	52	26
Rule 4	special imine	120	0.5	23	11
Rule 5	1.3 aromatic heteroatom H shift	2,392	9.5	35	17
Rule 6	1.3 heteroatom H shift	9,143	36.3	124	61
Rule 7	1.5 (aromatic) heteroatom H shift (1)	6,826	27.1	92	45
Rule 8	1.5 (aromatic) heteroatom H shift (2)	2,204	8.7	32	16
Rule 9	1.7 (aromatic) heteroatom H shift	1,970	7.8	44	22
Rule 10	1.9 (aromatic) heteroatom H shift	788	3.1	26	13
Rule 11	1.11 (aromatic) heteroatom H shift	138	0.5	15	7
Rule 12	furanones	322	1.3	32	16
Rule 13	ketene/ynol exchange	*not found*			
Rule 14	ionic nitro/aci-nitro	*not found*			
Rule 15	pentavalent nitro/aci-nitro	*not found*			
Rule 16	oxime/nitroso	2	0.0	2	1
Rule 17	oxime/nitroso via phenyl	*not found*			
Rule 18	cyanic/isocyanic acids	*not found*			
Rule 19	formamidinesulfinic acids	*not found*			
Rule 20	isocyanides	*not found*			
Rule 21	phosphonic acids	*not found*			
		Ring−Chain Rules			
Rule RC1	3-exo-trig	*not found*			
Rule RC2	4-exo-trig	*not found*			
Rule RC3	5-exo-trig	136	37.8	8	4
Rule RC4	6-exo-trig	79	21.9	8	4
Rule RC5	7-exo-trig	1	0.3	2	1
Rule RC6	5-exo-dig	12	3.3	8	4
Rule RC7	6-exo-dig	19	5.3	8	4
Rule RC8	7-exo-dig	1	0.3	0	0
Rule RC9	5-endo-trig	26	7.2	16	8
Rule RC10	6-endo-trig	86	23.9	18	9
Rule RC11	7-endo-trig	*not found*			

[a] In the naming of the ring−chain rules, the initial number refers to the number of atoms in the ring, exo and endo refer to exocyclic and endocyclic ring closure processes, and dig (digonal/sp) and trig (trigonal/sp2) refer to the hybridization state of the electrophilic carbon.[21,31] [b] *Not found*: no example of a conflict involving this rule was found in the AMS.

The second step was to search for the shortest transformation pathway possible between each pair of tautomers in each tautomeric conflict, i.e., the minimum number of transformation steps to get from one tautomer to the other. In this way, the evaluation of the tautomeric rules makes more sense from a statistical as well as an energetic point of view: a tautomeric interconversion with one or two steps (i.e., small energetic barriers to overcome) usually has a higher likelihood of occurring under standard conditions than one with more transformation steps. On the basis of this shortest-path analysis, we observed that the majority (81.2%) of the tautomeric conflicts required only one transformation step, though some transformations between tautomeric forms did require more: 17.4% of the cases required two steps, 0.9% required three steps and 0.1% required more than three steps.

Figure 2a shows an example of a keto−enol tautomerism conflict found between compound **1** and compound **7**. Figure 2b shows the tautomer network of this conflict; each vertex represents a tautomer and each line is a transformation rule. There are a total of 12 tautomers (shown in the Supporting Information) that can be enumerated for the two structures by applying the CACTVS rules. In this example, the shortest path between compound **1** and compound **7** is the line colored in green which corresponds to Rule 7 (1.5 (aromatic) heteroatom H shift (1)). Alternatively, one tautomer can be transformed into another through the application of several transformations, i.e. different tautomeric pathways through the tautomeric network.

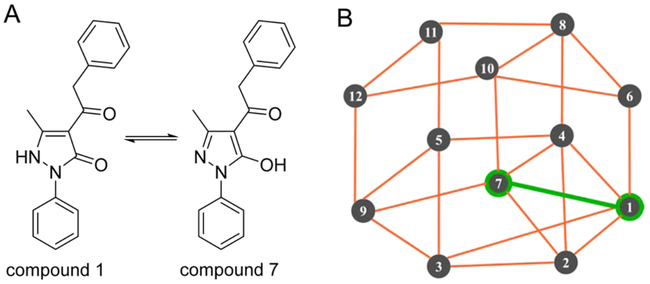


Figure 2. (A) Example of a keto−enol tautomerism conflict. (B) Tautomer network of this tautomeric pair. Each vertex represents a tautomer, and each line is a transformation rule. The shortest path between **1** and **7** is marked in green and corresponds to Rule 7 (1.5 (aromatic) heteroatom H shift (1)).

The majority of transformations between tautomers are of the prototropic type. The conflicts we found in the AMS database involved a subset of 12 out of the 20 prototropic rules as shown in Table 1. (Note that there is no longer a Rule 1 as it has been merged into Rule 6.) These 12 rules fall under the category of basic prototropic rules; we found no examples of conflicts with rarer prototropic tautomers involving groups such as cyanuric acids or phosphonic acids. Rules 6 and 7 are the most common transformations observed in the conflicts, while there were only two cases involving Rule 16. The distance that the hydrogen atom migrates (compare Rule 6 to Rule 11) correlated with the frequency of observation. As we expected from the application of tautomeric rules in other databases, ring−chain tautomerism is in the minority compared to prototropic tautomerism; nevertheless we found examples of tautomeric conflicts for 8 of the 11 ring−chain rules.

The ring−chain rules are very specific and selective.[21] A ring−chain transformation can only be encoded by one type of rule, as opposed to the prototropic rules where we have seen that the same transformation can be achieved via different pathways. Table 2 shows the set of alternative prototropic transformations for each type of tautomeric conflict we identified in the AMS database. The diagonal of the matrix represents tautomeric transformations that can only occur via one rule. Rules 6 and 7 are very general; for example we found cases that can be transformed by Rules 3, 4, 5, 7, 9, 12, or 16 as alternate pathways to Rule 6. This is because the SMIRKS transformation in Rule 6 is very general and tolerates any heteroatom (N, S, O) at the positions of the movement of the proton. At the other extreme, Rule 16 is very specific and represents only oxime/nitroso tautomerism. The individual assessment of the prototropic rules is complicated by these varying levels of specificity. Thus, instead of looking at the chemoinformatics rule applied for each tautomeric conflict, we were more interested in analyzing the type of tautomer being formed. The minimum moiety required for tautomerism in a molecule consists of three atoms able to produce the minimum 1,3 proton shift (besides the hydrogen, which is treated as implicit). Taking into account the topology of these three atoms, one obtains different types of tautomerism. For example, if atom 1 is an oxygen bound to a carbon (atom 2), and the carbon is bound to another carbon (atom 3), we have keto−enol tautomerism (O═C−C ↔ O−C═C). To cover all the basic prototropic tautomeric transforms, atom 1 can be either oxygen, nitrogen, or sulfur; and atoms 2 and 3 can be either carbon or nitrogen. In combination, this produces 12 different types of basic prototropic tautomerism. We classified the conflicts according to these 12 types of tautomerism (with the applicable transform rule(s) from Table 1 given in parentheses): keto−enol or thioketo−thioenol (Rule 2), imine−amine (Rules 3 and 4), amide−imide or thioamide−iminothiol (Rules 5 and 6), amidine−imidine (Rules 5 and 6), nitroso−oxime or thionitroso−thiooxime (Rules 16 and 17), azo-hydrazone (Rules 7 and 8), nitrosamine−diazohydroxide or thionitrosamine−diazothiol (Rule 6), and diazoamino−diazoamino (Rule 6).

Table 2. Matrix of Alternative Transformations of Prototropic Rules for the Set of Tautomeric Conflicts Identified in the AMS Database and Selected for NMR Evaluation[a]

	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8	Rule 9	Rule 10	Rule 11	Rule 12	Rule 16
Rule 2						13			3			
Rule 3			6		24						6	
Rule 4		6			11							
Rule 5					17			4			1	
Rule 6		24	11	17	6	1		4			17	1
Rule 7	13				1	11	16	1	9	1		1
Rule 8						16						
Rule 9				4	4	1		13	1	4		
Rule 10	3					9		1	5	1		
Rule 11						1		4	1	3		
Rule 12		6		1	17							
Rule 16					1	1						

[a] The color code indicates the relative frequency of each conflict, from light = infrequent to dark = most frequent.

NMR Analysis.

A set of 337 compounds (see SI for the 2D structure diagrams), consisting of 127 prototropic tautomeric pairs, 5 prototropic tautomeric triplets, and 34 ring−chain tautomeric pairs, was selected for NMR experiments. The aim of this analysis was to determine the identity or difference of the samples in the pair or triplet by comparing both ^{1}H NMR and ^{13}C NMR spectra between the individual compounds in each tautomeric conflict.

In comparing the NMR spectra of a conflict, ideally one would have one of two possible scenarios: either the two compounds will have the same spectra or they will have different spectra. However, this comparison can become more complicated because spectra do not always have fully resolved peaks indicating only a single tautomer. Therefore, there is another scenario applicable in both situations: the sample shows additional peaks due to the presence of impurities, a mixture of tautomers or an entirely different molecule. We thus saw that we can obtain what we called "simple spectra" or "complex spectra". We classified the comparison of spectra for each conflict into four categories: (a) same tautomers simple spectra (ST_SS), (b) same tautomers complex spectra (ST_CS), (c) different tautomers simple spectra (DT_SS), and (d) different tautomers complex spectra (DT_CS). Each conflict is compared twice, with the proton NMR spectra and the carbon NMR spectra. In the following, we show and discuss the conclusions drawn from the combined proton and carbon NMR comparisons but the Supporting Information provides the full individual results for each comparison type for each conflict.

Figure 3a shows the distribution of spectra comparisons for the prototropic conflicts. In 93 cases the spectra were simple and identical (category ST_SS), indicating the samples represented the same tautomer. For instance, conflict 26 (Figure 4) is a keto−enol tautomerism conflict whose ^{1}H NMR spectra showed only the enol form (26_1) for both samples based on the chemical shift at 12.5 ppm assigned to the hydroxyl proton of the enol form. In 29 cases, though the same tautomer is clearly present in both samples, the spectra indicated something else is also present in the sample such as impurities or other tautomeric forms (category ST_CS). For example, in conflict 31 (Figure 5), despite a lot of impurities shown in the proton spectra, we can still identify the same tautomer in both samples. It is interesting to note that the pattern of impurities is the same for 31_1 and 31_2, perhaps implying that both samples may have ultimately come from the same source. We further discuss this issue below. The spectra of

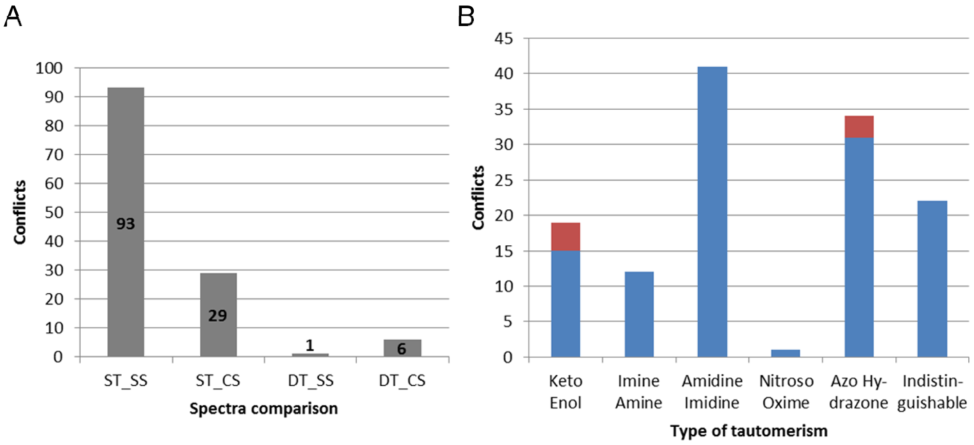


Figure 3. (A) Distribution of the NMR spectra comparisons for selected prototropic tautomeric conflicts: (ST_SS) same tautomers simple spectra; (ST_CS) same tautomers complex spectra; (DT_SS) different tautomers simple spectra; (DT_CS) different tautomers complex spectra. (B) Distribution of the type of prototropic tautomerism between the selected prototropic conflict types. Conflicts whose NMR spectra showed they are the same tautomer (ST_SS and ST_CS) are colored in blue, whereas conflicts whose NMR spectra showed that they are different tautomers (DT_SS and DT_CS) are colored in red.

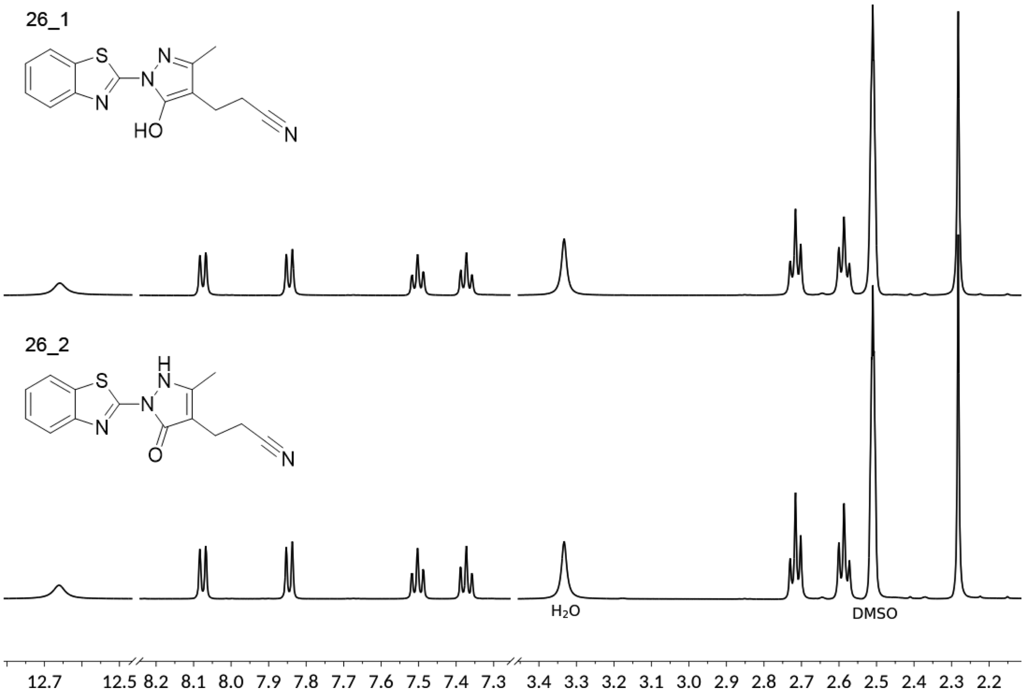


Figure 4. ^{1}H NMR spectra of conflict 26 involving keto−enol tautomerism. The comparison indicates that the samples in conflict 26 are in fact the same tautomer (ST_SS). Structures shown are the representations provided by the vendor.

conflict 131 had additional peaks that suggested that another isomer may be present in the sample; the carbon spectra showed duplication of some peaks at very similar chemical shift values.

We found only one single case in the category DT_SS, (conflict 30), with clearly different ^{1}H and ^{13}C NMR spectra that corresponded to different tautomers. The enol form was found in sample 30_1 and the keto form in sample 30_2. The physical appearance of these samples was slightly different, which also suggested the potential for different tautomers. Sample 30_1 had fine brown crystals whereas sample 30_2 was a dark yellow powder of relatively large particles. Six cases were assigned to the category DT_CS where we did not observe the same tautomer because the chemical shifts of at least one sample were unexpected for that particular chemical structure or it was a mixture of tautomers, as in conflicts 46 and 53.

Since the vast majority of the conflicts had the same spectra showing the same tautomers in both (or all three) samples, we cannot draw any conclusions as to which of the types, and specifically rules, of prototropic tautomerism may be too "aggressive" in the sense discussed above. It seems that at least the part of the current CACTVS rule set that could be tested with this analysis does indeed reproduce experimentally found tautomerism. Figure 3b shows the distribution of the selected prototropic conflicts between the different types of prototropic tautomerism. The few cases which had different spectra and not the same tautomer involve keto−enol and azo-hydrazone tautomerism. We labeled 22 conflicts in Figure 3b as

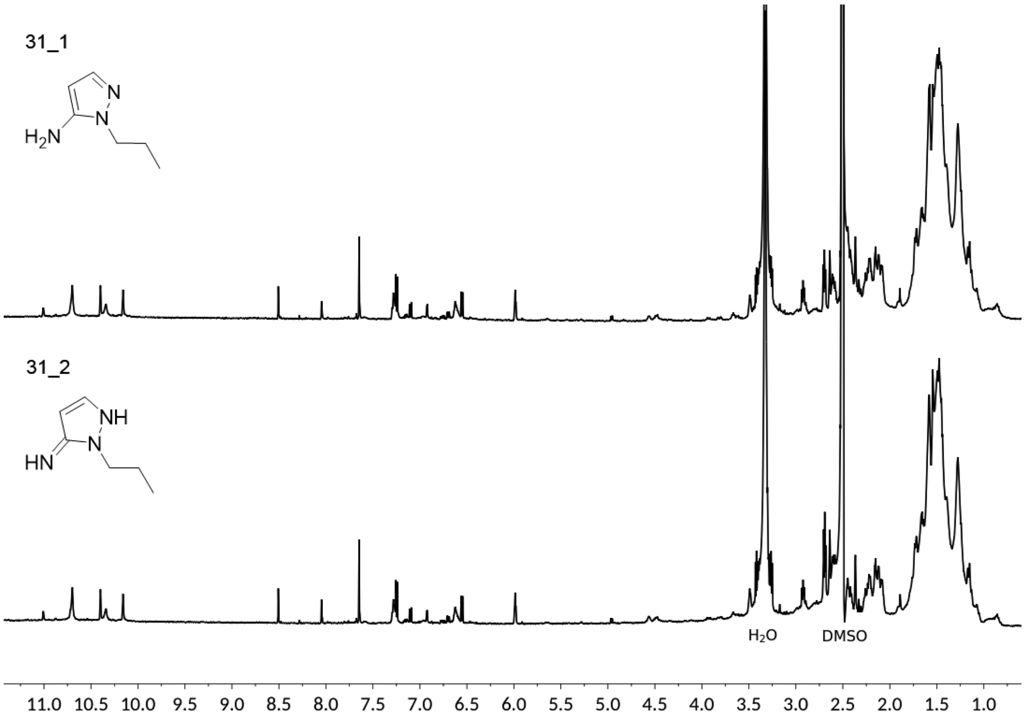


Figure 5. ^1H NMR spectra of conflict 31 involving imine−amine tautomerism. The comparison indicates that the samples in conflict 31 are the same tautomer, though many impurities are present in both samples (ST_CS).

"indistinguishable" because the *predicted* differences in the ^1H and ^{13}C chemical shifts between tautomers were very small or almost nonexistent. Similarity in chemical shifts between tautomers is, in some cases, due to symmetry and free rotation around single bonds. If the chemical context of a particular atom involved in the tautomeric transformation is the same within a distance of at least three surrounding atoms, not much difference can be expected in its ^1H or ^{13}C spectra. Some examples are conflict 12, 17, and 69 (Figure 6). These cases may have been resolvable with more elaborate approaches, which were, however, beyond the scope of this study. There is another group of conflicts such as 38 and 40, mostly azo-hydrazones, whose spectra were predicted to be almost the same for different tautomers but slight differences should still be apparent. These differences in the predicted ^1H spectra appear for one single atom involved in tautomerism whose chemical shift varies by, at most, one ppm.

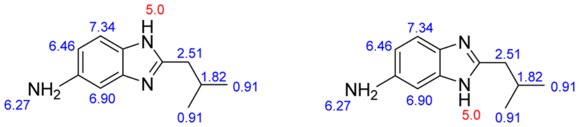


Figure 6. Conflict 69 is indistinguishable in standard NMR experiments. Predictions from ChemDraw show the same ^1H chemical shifts for both tautomers. The estimation quality of the chemical shifts is indicated by color: good in blue; exchangeable protons (less reliable predictions) in red.

The primary goal of this experimental analysis was not to assign all the peaks of every spectrum and determine the chemical structure of the tautomer present in each sample, or to predict the most favorable tautomer in each case. For some types of tautomerism, however, we were able to easily identify the tautomer present in the sample by looking for a particular chemical shift. For example, in conflicts 5, 6, 7, and 10, involving imine-amine tautomerism, the ^1H NMR spectra showed a peak around 9 ppm that corresponds to the proton bound to the nitrogen (i.e., imine form). Conflicts 29, 133, and 134, however, also involving an imine−amine transformation, do not show this peak at 9 ppm, which implies that the amine form is prevalent. For keto−enol tautomers, such as conflicts 11, 23, and 136, a peak around 200 ppm in the ^{13}C NMR spectra is indicative of the keto form, whereas a peak around 11.50 ppm in the ^1H NMR spectra suggests the enol form is present.

The distribution of spectra comparisons for the ring−chain conflicts (Figure 7a) showed fewer cases of identical tautomers than for the prototropic conflicts. While the majority of RC conflicts yielded the same tautomer (14 and 9 cases in categories ST_SS and ST_CS, respectively), there was a higher percentage of samples that did not show the same tautomer. It is interesting to note that the number of RC cases with simple spectra was close to the number of RC cases with complex spectra. Ring−chain tautomerism might thus be associated with a higher likelihood of a compound occurring as a mixture of tautomers than as a single tautomer, when compared to the situation with prototropic tautomerism. For instance, in conflict 679 from group ST_CS (Figure 8), the proton and carbon spectra had some differences because 679_2 had additional peaks that we assigned to a different tautomer. The 679_1 spectra may indicate the closed form and 679_2 could be a mixture of the closed and open forms.

Five ring−chain tautomerism cases were assigned to the category DT_SS because each sample contained a separate tautomer according to the NMR spectra. Figure 9 shows the proton NMR spectra of conflict 695. The peak assignation of 695_1 represents the open form whereas sample 695_2 is in the closed form. Those compounds, which are connected through Rule RC6 5_exo_dig, do not tautomerize. For the six conflicts in category DT_CS, we did not observe the same

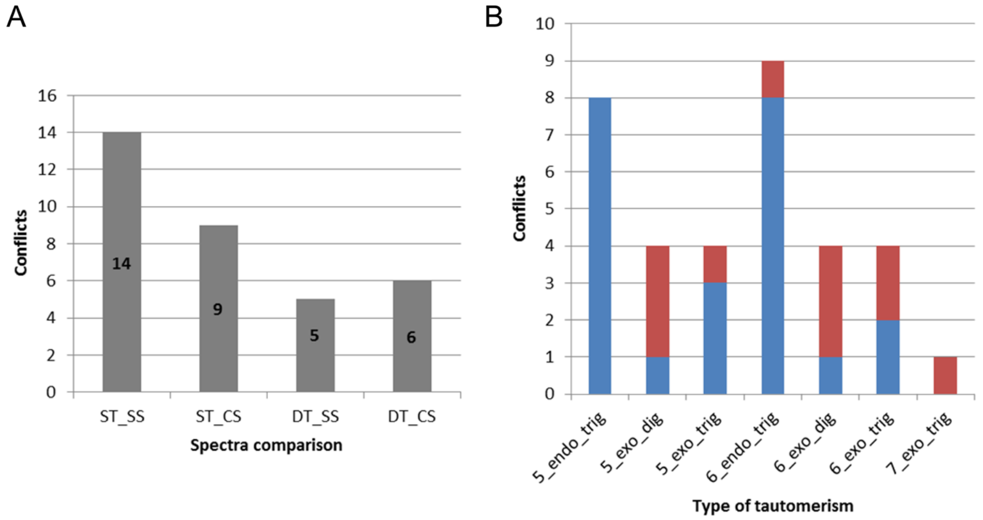


Figure 7. (A) Distribution of the NMR spectra comparisons for ring−chain tautomeric conflicts. (ST_SS) same tautomers simple spectra; (ST_CS) same tautomers complex spectra; (DT_DS) different tautomers simple spectra; (DT_CS) different tautomers complex spectra. (B) Distribution of the type of ring−chain tautomerism between the selected ring−chain conflict types. Conflicts whose NMR spectra showed the same tautomer (ST_SS and ST_CS) are colored in blue, whereas conflicts whose NMR spectra showed different tautomers (DT_SS and DT_CS) are colored in red.

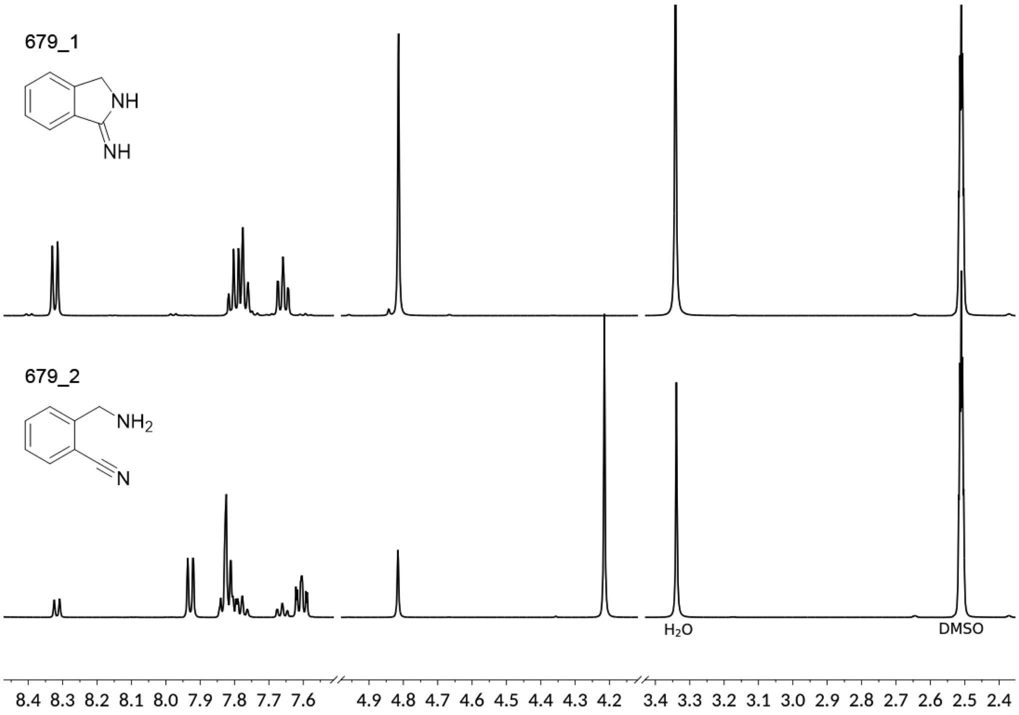


Figure 8. ^{1}H NMR spectra of conflict 679 involving ring−chain tautomerism. Sample 679_1 contained the closed-form tautomer whereas sample 679_2 was a mixture of the open and closed forms.

tautomer but the chemical shifts were unexpected for that particular chemical structure. These samples most likely contain entirely different compounds, as occurred in conflict 52 and conflict 134.

Based on our chemoinformatics tautomeric analysis, ring−chain tautomerism occurs less frequently than prototropic tautomerism. For the purpose of assessing the transform rules, we likewise have fewer tested examples. However, the specificity of the ring−chain chemoinformatics rules (RC transformations can only be encoded by one type of rule) allows us to analyze the results individually for each rule as shown in Figure 7b. A high number of conflicts involved Rules 5_endo_trig or 6_endo_trig, and the results of their NMR spectra comparison showed that most of the conflicts had the same tautomer. This suggests that the endo_trig rules can be reliably used for deduplicating molecules capable of ring−chain tautomerism. In contrast, Rules 5_exo_dig and 6_exo_dig appear somewhat "aggressive" at predicting naturally interconverting tautomers at least under standard conditions. The results for the exo_trig rules were too inconclusive to make a call whether those rules predict ring−chain tautomers well. Larger sets of data may be needed to answer this question; and

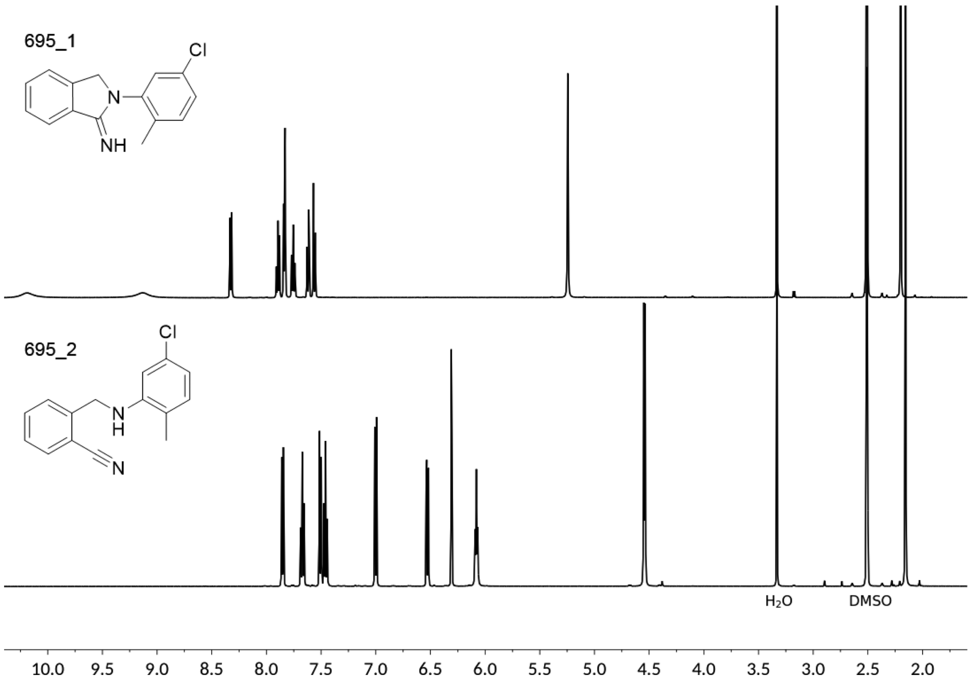


Figure 9. ^{1}H NMR spectra of conflict 695 involving ring−chain tautomerism. The comparison of these spectra indicates the samples are indeed different structures (DT_SS). Sample 695_1 contains the closed form, and sample 695_2 contains the open form.

Table 3. MS Results Comparing Trace Impurities between Selected Tautomeric Conflict Pairs

conflict ID	tautomerism type	rule[a]	physical aspect	^{1}H NMR[b]	^{13}C NMR[b]	MS[c]
6	imine−amine	R10	same	same	same	diff
11	keto−enol	R10	same	same	same	same
23	keto−enol	R6, R12	same	same	same	same
30	keto−enol	R6, R12	diff	diff	diff	diff
48	keto−enol	R2, R7	same	same	same	diff
61	imine−amine	R3, R6	same	same	same	diff
64	imine−amine	R3, R6	same	same	same	same
83	amidine−imidine	R5, R6	diff	same	same	same
92	keto−enol	R6, R12	diff	same	same	diff
94	amidine−imidine	R6	same	same	same	same
97	amidine−imidine	R5, R6	same	same	same	same
110	imine−amine	R7	same	same	same	same
126	imine−amine	R7, R8	same	same	same	same
133	imine−amine	R6	same	same	same	same
135	imine−amine	R9	same	same	same	same
136	keto−enol	R9	diff	same	same	diff
328	ring−chain	5_endo_trig	same	same	same	*d*
617	ring−chain	6_exo_trig	same	same	same	same
660	ring−chain	6_endo_trig	same	same	same	same
987	ring−chain	5_exo_trig	same	same	same	same

[a] Only one transformation is necessary for the interconversion of each of the conflicts. However, for some conflicts (e.g., 23, 30, ...), two alternative rules can be applied to produce the same transformation. [b] Here same vs diff refers to whether or not the tautomeric form appears to be the same in both samples. [c] Here same vs diff refers to whether or not the samples appear to come from the same upstream source. [d] See discussion of this conflict below in the Environmental Variables section.

it may well be that this type of ring−chain tautomerism is so structure-dependent that no general verdict can be reached.

It is interesting to compare the prevalence of RC tautomerism for our rules with Baldwin's rules.[31] Rules 5-exo-dig and 6-exo-dig, and the exo_trig rules, were predicted as favorable ring−chain closures by Baldwin. However, our experiments did not provide conclusive results especially when the geometry of the atom being attacked was linear (i.e., was of type "dig"). Baldwin suggested two types of behaviors for the endo_trig rules depending on the size of the ring being formed: Three- to five-membered rings formed were predicted as unfavorable whereas six to seven-membered rings formed were predicted as favorable. However, our results shows the same type of favorable interconversion whether for both five- and six-membered ring formed during the endocyclic ring closure.

Mass Spectrometric Analysis. We subjected a subset of the compounds that had been analyzed by NMR to further

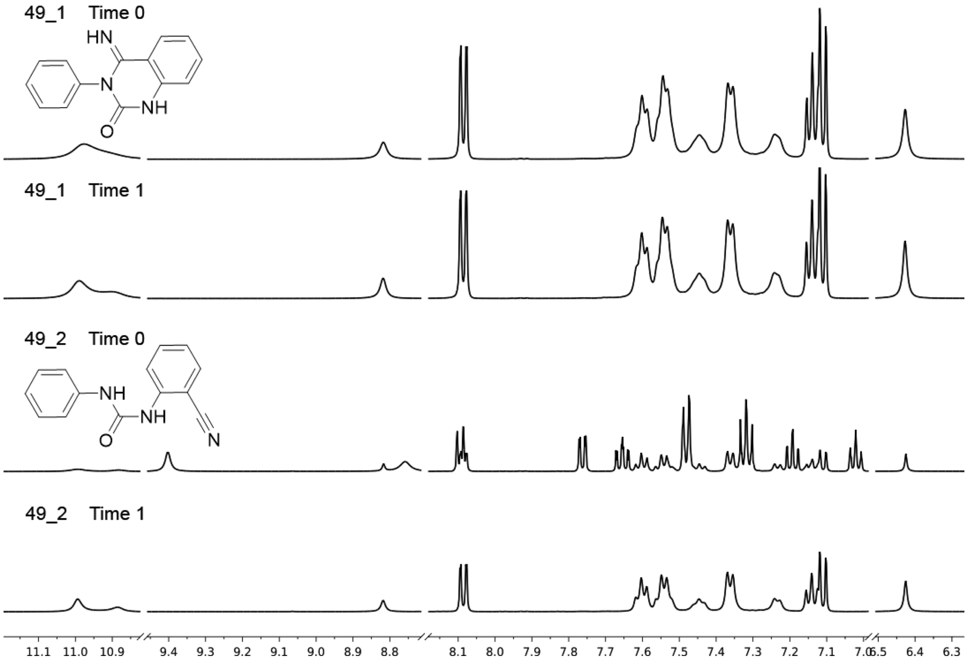


Figure 10. Change in time of ^{1}H NMR spectra of conflict 49 involving ring–chain tautomerism. Initially (Time 0), structure 49_1 was mainly the closed form whereas 49_2 was the open form with some closed form present. At Time 1 (2 weeks later), both samples showed mainly the closed form.

analysis with MS. All samples were independently analyzed at least twice using two ionization methods and direct-injection of the sample solution. LC/MS was employed if the compounds were amenable to it based on structure or if there were questions about the initial analyses, e.g. obvious impurities or unexpected results.

While MS cannot necessarily distinguish between tautomers that have the same molecular weight, we were interested in exploring the possibility of analyzing the pattern of trace impurities in different samples of the same tautomer in order to provide a "fingerprint" that might suggest whether the samples may (or may not) have come from the same upstream source. This analysis was therefore not intended to provide a comprehensive "forensic investigation" of the complete sample set by MS but to deliver a first impression of what such an analysis might yield. We investigated a set of 20 conflicts (40 samples), distributed over the various types of tautomerism and transform rules, and all having simple spectra not obviously containing other compounds or large amounts of impurities. The purity stated by the suppliers for these samples ranged between 90 and 95%.

The results, comparing the MS with both the NMR results and the physical appearance of the samples, are shown in Table 3. The full set of MS data for each compound is provided in the Supporting Information. In all but one case the NMR results had shown that the compounds in each conflict pair were the same tautomer. The conclusions we reached from the MS results for these samples were distributed approximately 2:1 between possibly the same primary source (13 cases) vs most likely two different sources (6 cases). In the case of conflict 30, the only prototropic case of different tautomers with simple spectra (DT_SS), the samples had different physical appearances, and the MS results (Table 3) showed that these likely came from different sources as well.

We found that the physical appearance of the compound samples was not necessarily predictive of either their tautomeric form or their original source either. In conflict 83, one sample was in the form of fine white crystals and the other appeared as light yellow particles, but the NMR spectra were identical and the MS results showed evidence of the same byproduct or decomposition product in each sample to the same extent. Conversely, six of the conflicts were identical in appearance but showed traces of having originated from different sources.

Even though this was a limited analysis by MS, the results already showed that this is a complex problem. If two different suppliers followed the same published synthetic procedure to the letter (and perhaps even used starting materials from the same source), similar impurity patterns may not be so unexpected. However, as it is logistically impossible for an end buyer to establish any kind of comprehensive "chain of custody" of commercially acquired samples, it is conceivable that samples originating from "the same bottle" might become different tautomers through different handling, storing, and transportation conditions; or conversely that samples originating from different primary sources could become more similar to one another through environmental conditions promoting or accelerating tautomerization. The following section briefly discusses these parameters in the context of the history of a sample before it is actually used in the buyer's laboratory.

Environmental Variables. Time—i.e., time on the shelf during which the tautomerization reaction can proceed—is definitely an important variable in tautomerism. We measured the NMR spectra of some of the samples 2 weeks apart and found some cases in which the tautomeric ratio had changed. For conflict 49, shown in Figure 10, the first NMR spectrum obtained suggested that 49_1 was mainly the closed form, and 49_2 was the open form with some closed form present. However, 2 weeks later we observed the same proton and carbon spectra for both samples: 49_2 had converted to mainly the closed form as found initially for sample 49_1. For conflict 201, we observed the opposite effect—instead of converging with time, 201_1 and 201_2 diverged to different spectra. The carbon spectra were very clean, and in the beginning 201_1 and 201_2 showed the open form; however, 2 weeks later 201_2 had become a mixture of products, and the spectra no longer looked the same. It would not be chemically plausible to attribute this event to tautomerism; instead it is most probable that different amounts and/or types of impurities may have led to some reaction or decomposition in one sample but not the other. We did not however further investigate this case.

For conflict 328, both the ^{13}C and ^{1}H NMR spectra indicated that the two samples in this conflict pair were exactly the same (with the NMR spectra favoring the open form tautomer 328_2 in both samples). Conversely, the MS spectra of conflict 328 (Figure 11) did not show evidence of any tautomers of the registered (and sold) molecule, instead only the possible starting materials of the sample: (4-dimethylamino)-benzaldehyde and 2-amino-1-(4-nitrophenyl)ethanol). The most obvious interpretation is that the samples had decomposed in the intervening time between the two experiments, as the MS analyses were done several months after the NMR experiments had been conducted. It is also possible that the acidic conditions of the ionization reaction had hydrolyzed the sample during the MS analysis itself. Other studies have pointed out this effect of "disappearing compounds" in sample collections,[32] and these anecdotal examples again highlight the importance, well-known in the high-throughput screening literature,[33] of verifying the quality and status of any compound sample before it is used in an experiment.

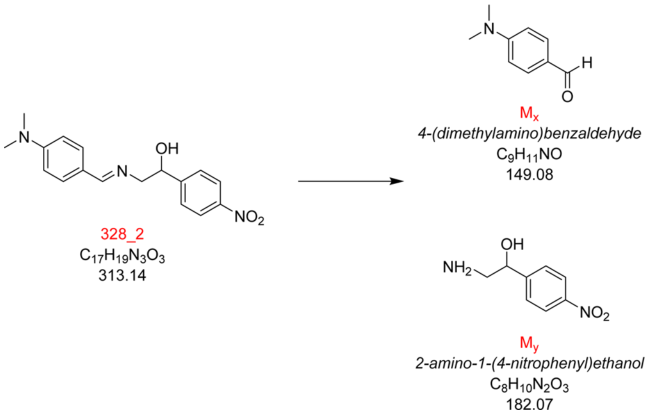


Figure 11. Results of the LC/MS analysis of conflict 328, showing apparent decomposition of the sample into two precursors or hydrolysis products. The masses listed for the molecular weights are the monoisotopic masses (relevant to MS) although the actual ions seen in the MS are MH^+.

As mentioned already, the relative ratio of the tautomers of any compound is highly dependent on the environmental conditions including temperature, solvent, solute, pH, concentration, etc. In this study, we determined the NMR spectra under only one condition, at room temperature in DMSO as the solvent. As the results for the prototropic tautomerism cases were decisive in the sense that the majority of the conflicts had shown to have the same tautomer in both (or all three) samples, we did not repeat the NMR experiments under different conditions. For example, heating up samples that had already been shown to be the same compound, would not under reasonable assumptions be expected to generate different compounds. For some ring−chain tautomerism conflicts whose NMR spectra were different, it might be interesting to measure the samples again under different environmental conditions; however this was outside the scope of this study.

■ CONCLUSION

The identification of tautomeric conflicts in real (i.e., nonvirtual) sample databases presents a useful scenario for experimental analysis of chemoinformatics rules encoding tautomeric transforms. We identified a set of 62 869 molecules in a prototypical screening sample database as being tautomeric pairs or multiples on the basis of prototropic and ring−chain chemoinformatics rules. This set included example conflicts for 20 out of the total of 31 chemoinformatics rules we employed in our combined approach to prototropic and ring−chain tautomerism. For most of the prototropic conflict cases, the spectra indicated that the different commercial products were in fact the same compound. The comparison of ring−chain tautomer spectra produced a somewhat different picture than for the prototropic results in that we found that Rules 5_exo_dig and 6_exo_dig appear to be too aggressive for applications such as compound deduplication in sample databases whereas the endo_trig rules do seem reliable in describing ring−chain transformations.

No examples constituting a conflict were found for the remaining 11 rules in this database, which prevents us from reaching any conclusions as to their appropriateness for deduplication of compound collections (other than their relative rarity). To expand the coverage of the rules to these missing 11 cases, it might be possible to synthesize a particular tautomer with one route, and a different one via another route, and then perform the NMR spectroscopic analysis as above. However, dedicated synthesis was entirely beyond of the scope of this study.

This analysis indicates that our chemoinformatics rules appear to be better at recognizing tautomeric transformations that lead to the same "stuff in the bottle" than many standard vendor representations in databases of commercially available compounds. Improvements in the structure normalization process that handles tautomerism and stereochemistry are essential for correct compound registration. Modern approaches and software allow the rapid calculation of unique, tautomer-invariant identifiers that greatly facilitate the detection of tautomeric forms. We would argue that applying these types of chemoinformatics approaches to all chemical databases would be beneficial to providers, users, and sample buyers alike in order to improve database quality in terms of avoiding or at least annotating tautomeric duplication.

■ EXPERIMENTAL SECTION

Selection of Tautomer Pairs for Experimental Evaluation. A subset of the tautomeric conflicts was selected considering the following criteria: (a) coverage of the rule set, with the goal of including as many different rules as possible; (b) shortest transformation path, prioritizing conflicts where a one-step transformation occurs in order to minimize ambiguities in the analysis; (c) chemical diversity, based on clustering by linear fingerprints; (d) solubility, based on calculation of logS and logP; (e) availability from the same supplier or vendor catalog (since this could be considered a more "serious" case of tautomeric conflict as these "different" products are likely sold at different unit prices); and (f) likelihood of being distinguishable by NMR. We applied these criteria to the Aldrich Market Select (AMS) database of 6 million screening samples and building blocks. We placed sample orders for a total of 371 samples with Sigma-Aldrich (Milwaukee, Wisconsin, USA) at an average price per sample of $78. Experiencing a typical delivery attrition rate of about 10%, we received a total of 337 compounds comprising 127 prototropic tautomeric pairs, 5 prototropic tautomeric triplets, and 34 ring−chain tautomeric pairs.

NMR Analysis. NMR spectra were obtained on a Bruker Avance III-500 spectrometer operating at 500 and 125 MHz for ^{1}H and ^{13}C, respectively, equipped with a cryogenic triple resonance probe. Approximately 3 mg of each sample (~10 mM concentration) were dissolved in 100% DMSO-d_6 and NMR data were collected with the probe temperature set to 298 K. One-dimensional spectra were recorded with standard pulse sequences with between 16 and 64 scans and a recycle delay of 1 s for ^{1}H spectra or 512−1024 scans and a recycle delay of 2 s for ^{13}C spectra. NMR data were processed using the MNova NMR software (Mestrelab, Escondido, CA).

MS Analysis. The mass spectra were obtained as follows. A selected subset of the tautomeric pairs was subjected to comparative analyses by direct sample introduction or flow-injection analysis (FIA) mass spectrometry and by LC/MS, where feasible. Solid samples were accurately weighed (±0.003 mg) on a Thermo-Cahn C-35 electrobalance and a stock solution of 1.00 mg/mL concentration was made by dissolution in the appropriate high-purity solvent (i.e., CH_3OH, CH_3CN, CH_2Cl_2, H_2O, DMSO) or combination of solvents. An aliquot of the stock solution was further diluted to a concentration of 25 μg/mL in 1:1 LC-MS grade CH_3OH/H_2O and a 1.0- to 5.0-μL aliquot of this diluted solution was used for mass spectrometric and chromatographic analysis. Low resolution, positive ion MS analyses were carried out on an Agilent LC/MSD single quadrupole system, equipped with an in-line diode-array UV detector, to assess compound identity and homogeneity. Initial analyses were carried out in FIA mode with the sample injected directly into the LC/MSD using 1:1 CH_3OH/H_2O containing 0.1% CH_3COOH at a flow rate of 300 μL/min. Where feasible, samples were additionally analyzed by LC/MS using a narrow-bore (100 × 2.1 mm), small-particle (3.5-μm), Zorbax Rapid-Resolution reversed-phase C_{18} column coupled with a C_{18} guard column (12.5 × 2.1 mm) eluted with a 5−90% gradient of CH_3OH/H_2O containing 0.1% CH_3COOH at a flow rate of 300 μL/min. All samples were analyzed using both electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI) modes, and the resulting mass spectra were averaged and background-subtracted using the standard ChemStation software (ver. B.02.01-SR2). Full scan mass spectra, as well as both the total-ion chromatogram (TIC) and the UV-chromatogram, were used to assess compound purity and similarity. The full scan (210−400 nm) diode-array UV spectra for both FIA and LC/MS analyses of each tautomer were also generated and compared to assess similarity.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00338.

Enumeration of 12 tautomers from the tautomer network example in Figure 2 (tautomers_Figure 2.cdx). Full list of chemoinformatic, supplier, and experimental information for prototropic and ring−chain tautomeric conflicts selected for experimental evaluation (conflict_lists.xlsx). 2D structures and comparison of ^{1}H and ^{13}C NMR spectra for selected prototropic tautomeric conflicts (prototropic_NMR.pdf). 2D structures and comparison of ^{1}H and ^{13}C NMR spectra for selected ring−chain tautomeric conflicts (ring−chain_NMR.pdf). Mass spectral, chromatographic, and UV data for selected tautomeric conflicts (mass_spec.pdf) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mn1@helix.nih.gov.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Wolf-Dietrich Ihlenfeldt for his help with the chemoinformatics toolkit CACTVS, and Dr. Markus Sitzmann for his help with the calculation of identifiers for the AMS database. We also gratefully acknowledge Dr. Christopher C. Lai for his expert assistance in analytical sample preparation and for LC/MS analysis of a portion of the samples. This work was funded in part by the Intramural Research Program of the National Institutes of Health, at the Center for Cancer Research, National Cancer Institute, and in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

■ REFERENCES

(1) Langmuir, I. THE ARRANGEMENT OF ELECTRONS IN ATOMS AND MOLECULES. *J. Am. Chem. Soc.* **1919**, *41*, 868−934.

(2) Dalby, A.; Nourse, J.; Hounshell, W.; Gushurst, A.; Grier, D.; Leland, B.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(3) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(4) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminf.* **2013**, *5*, 7.

(5) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, 23.

(6) Chemical Abstracts Service (CAS). https://www.cas.org/ (accessed Mar 29, 2016).

(7) Peach, M. L.; Zakharov, A. V.; Guasch, L.; Nicklaus, M. C. Chemoinformatics. In *Comprehensive Biomedical Physics*; Brahme, A., Ed.; Elsevier: Amsterdam, 2014; vol. *6*, pp 123−156.

(8) Representation of Chemical Compounds. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH Verlag GmbH, 2008; pp 21−26.

(9) Jacob, N. T.; Lockner, J. W.; Kravchenko, V. V.; Janda, K. D. Pharmacophore Reassignment for Induction of the Immunosurveillance Cytokine TRAIL. *Angew. Chem., Int. Ed.* **2014**, *53*, 6628−6631.

(10) ChemSpider. http://www.chemspider.com/ (accessed Mar 29, 2016).

(11) Beilstein Database. https://www.reaxys.com/ (accessed Mar 29, 2016).

(12) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed Mar 29, 2016).

(13) ChEMBL Database. https://www.ebi.ac.uk/chembl/ (accessed Mar 29, 2016).

(14) Sayle, R. A. So You Think You Understand Tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, *24*, 485−496.

(15) Valters, R. *Ring-Chain Tautomerism*, softcover reprint of the original first 1985 ed.; Springer: Boston, MA, 2013.

(16) Warr, W. A. Tautomerism in Chemical Information Management Systems. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 497−520.

(17) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. C. Tautomerism in Large Databases. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 521−551.

(18) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduction Res.* **2003**, *23*, 361−371.

(19) Martin, E.; Monge, A.; Duret, J.-A.; Gualandi, F.; Peitsch, M. C.; Pospisil, P. Building an R&D Chemical Registration System. *J. Cheminf.* **2012**, *4*, 11.

(20) Guasch, L.; Peach, M. L.; Nicklaus, M. C. Tautomerism of Warfarin: Combined Chemoinformatics, Quantum Chemical, and NMR Investigation. *J. Org. Chem.* **2015**, *80*, 9900−9909.

(21) Guasch, L.; Sitzmann, M.; Nicklaus, M. C. Enumeration of Ring−Chain Tautomers Based on SMIRKS Rules. *J. Chem. Inf. Model.* **2014**, *54*, 2423−2432.

(22) 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.

(23) Ihlenfeldt, W.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical-Properties in CACTVS - an Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(24) Xemistry Chemoinformatics. http://www.xemistry.com/ (accessed Mar 29, 2016).

(25) Kleinpeter, E. NMR Spectroscopic Study of Tautomerism in Solution and in the Solid State. In *Tautomerism: Methods and Theories*; Antonov, L., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2013; pp 103−143.

(26) Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer Identification and Tautomer Structure Generation Based on the InChI Code. *J. Chem. Inf. Model.* **2010**, *50*, 1223−1232.

(27) IUPAC|International Union of Pure and Applied Chemistry Project Details. http://iupac.org/projects/project-details/?project_nr=2012-023-2-800 (accessed Apr 11, 2016).

(28) Aldrich Market Select. https://www.aldrichmarketselect.com/ (accessed Mar 29, 2016).

(29) Sitzmann, M.; Filippov, I. V.; Nicklaus, M. C. Internet Resources Integrating Many Small-Molecule Databases. *SAR QSAR Environ. Res.* **2008**, *19*, 1−9.

(30) 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.

(31) Baldwin, J. E. Rules for Ring Closure. *J. Chem. Soc., Chem. Commun.* **1976**, *18*, 734−736.

(32) Kozikowski, B. A.; Burt, T. M.; Tirey, D. A.; Williams, L. E.; Kuzmak, B. R.; Stanton, D. T.; Morand, K. L.; Nelson, S. L. The Effect of Room-Temperature Storage on the Stability of Compounds in DMSO. *J. Biomol. Screening* **2003**, *8*, 205−209.

(33) MacArthur, R.; Leister, W.; Veith, H.; Shinn, P.; Southall, N.; Austin, C. P.; Inglese, J.; Auld, D. S. Monitoring Compound Integrity with Cytochrome P450 Assays and qHTS. *J. Biomol. Screening* **2009**, *14*, 538−546.