Published in final edited form as:

J Chem Inf Model. 2020 March 23; 60(3): 1253–1275. doi:10.1021/acs.jcim.9b01080.

Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2

Devendra K. Dhaked,

Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, Maryland 21702, United States;

Wolf-Dietrich Ihlenfeldt

Xemistry GmbH, D-61479 Glashütten, Germany;

Hitesh Patel, Victorien Delannée, Marc C. Nicklaus

Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, Maryland 21702, United States;

Abstract

We have collected 86 different transforms of tautomeric interconversions. Out of those, 54 are for prototropic (non-ring–chain) tautomerism, 21 for ring–chain tautomerism, and 11 for valence tautomerism. The majority of these rules have been extracted from experimental literature. Twenty rules, covering the most well-known types of tautomerism such as keto–enol tautomerism, were taken from the default handling of tautomerism by the chemoinformatics toolkit CACTVS. The rules were analyzed against nine different databases totaling over 400 million (non-unique) structures as to their occurrence rates, mutual overlap in coverage, and recapitulation of the rules' enumerated tautomer sets by InChI V.1.05, both in InChI's Standard and a Nonstandard version with the increased tautomer-handling options 15T and KET turned on. These results and the background of this study are discussed in the context of the IUPAC InChI Project tasked with the redesign of handling of tautomerism for an InChI version 2. Applying the rules presented in this paper would approximately triple the number of compounds in typical small-molecule databases that would be affected by tautomeric interconversion by InChI V2. A web tool has been created to test these rules at <https://cactus.nci.nih.gov/tautomerizer>.

Corresponding Author: Marc C. Nicklaus – Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, Maryland 21702, United States; mn1@mail.nih.gov.

The authors declare no competing financial interest.

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jcim.9b01080>

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.9b01080>.

Spreadsheet S1: Occurrences, InChI Pass, Fail and other data for Standard InChI (XLSX)

Spreadsheet S2: Occurrences, InChI Pass, Fail and other data for Nonstandard InChI (XLSX)

Spreadsheet S3: Array-based tautomeric rule recapitulation for both Standard and Nonstandard InChI (XLSX)

Spreadsheet S4: Tautomeric examples received from Noel O'Boyle and Roger Sayle with our analysis added (XLSX)

Spreadsheet S5: Rule overlap data (XLS)

SMIRKS of tautomeric transforms and all scripts used to generate results (ZIP)

Tables S1 and S2 contain lists of CACTVS "ens transform" command flags used with each transform and of literature references for new transforms, respectively (PDF)

Brief analysis of, and difficulties encountered in, adapting the CACTVS-based rules to the chemoinformatics toolkits CDK and RDKit, as well as four modified JAVA classes of the CDK source code (ZIP)

Graphical Abstract

INTRODUCTION

Tautomerism—the existence of multiple possible forms of the same molecule that are capable of interconverting via an intramolecular movement of atoms—is a ubiquitous chemical phenomenon, especially in organic chemistry. Used without a further qualifier, the term is typically meant to designate prototropic tautomerism; i.e., the moving atom is a hydrogen. Other though rarer forms of tautomerism (valence tautomerism, movement of larger groups) are known. In another case of arguably imprecise usage of terminology in the field of tautomerism, interconversion and equilibrium of structures that involves closing and opening of a ring are called ring–chain tautomerism even though many of these cases are really a variant of prototropic tautomerism in that they involve movement of a proton. On the other hand, cyclizations without proton movements do occur such as in tetrazole–azide tautomerism. Tautomerism can occur in neutral or in charged molecules, and it can lead to an equilibrium that involves a zwitterion. We refer to the recent book chapter by Kleinpeter¹ about NMR-based studies of tautomerism for an excellent overview and many detailed examples of various types of tautomerism.

However, these definitions do not really answer the question of when tautomerism becomes relevant, or an issue, in different areas of chemistry, structural biology, etc. Scouring the literature and websites such as Wikipedia, one typically finds definitions such as that tautomers must “readily interconvert” or that it involves “facile migration of a proton.” But what is the unit of “readily”? Where lies the border between facile and difficult? Closer inspection of the concept reveals that “tautomerism” is surprisingly nonquantitative and that its meaning and scope in practical terms is essentially in the eye of the beholder. The main point is that tautomerism is not an immutable property of a molecule (such as molecular weight) but strongly condition-dependent. Temperature, solvent, pH, presence of impurities that can act as catalysts, packing forces in crystals, and other factors all influence interconversion rates and where the equilibrium lies. Tautomerism of the very same molecule can therefore mean very different things to the synthetic organic chemist, the computational quantum chemist, the maintainer of a compound registration system connected with a database of millions of molecules, or the developer of chemoinformatics software.

Tautomeric equilibria can, in principle, be computed based on relative energies via quantum mechanical (QM) computations at a high enough level of theory. Such computations are relatively straightforward, and have been frequently reported, for a vacuum environment as well as for various solvents utilizing solvation models such as the Polarized Continuum Model. Apart from the question of whether these continuum models can fully recapture possible involvement of solvent molecules in the proton migration, such as proton shuttling by water molecules, they are still prohibitively expensive for analyses of more than a handful of molecules.

The situation is in some sense even worse in chemoinformatics. While it is not deemed out of the ordinary or unacceptably onerous to spend days of CPU time on QM computations, a standard scenario in chemoinformatics is that an entire database of many thousands if not millions of molecules needs to be processed. This means that no more than, say, 1 CPU *second* can be spent on calculating everything that needs to be known about the tautomerism of an individual compound. Only rule-based approaches can currently achieve this; any physics-based algorithm is out of the question. The necessity of using rule-based approaches does however offer opportunities, too: (1) It is generally accepted that these rules can be expected to reproduce experimental results and/or higher-level computations only in a statistical sense, i.e., in the (hopefully vast) majority of cases. (2) They can be easily modified. (3) They can be developed by, and be based on, very different chemoinformatics approaches. (4) There can, at least in principle, be different sets of such rules for different conditions.

Tautomerism is not just an academic topic. It is a real-life issue with potential economic and even health-related consequences for chemical companies and their customers, database providers, drug developers, and crystallographers. As we and others have shown, tautomerism analysis of large sample databases may, depending on the detailed tautomerism transform rules used, turn up thousands of cases in which two different products (possibly sold at different unit prices) are declared as just different tautomers of the same molecule (“stuff in the bottle”) by the chemoinformatics rules.² In a similar vein, the fact that Warr³ reviewed the different approaches to handling of tautomerism used by 27 software vendors and database providers shows that there is a great diversity of views, approaches, and most likely outcomes in this field. After all, no one would write a review on how to calculate molecular weight. Different tautomers of the same molecule typically yield different predicted values for logP, hydrophobicity, pK_a , solubility, electrostatic potential, similarity index, etc., which may be a severely confounding factor in drug design.⁴ Finally, X-ray crystallography is also affected by incomplete, or even incorrect, handling of tautomerism, especially for the small-molecule ligand in protein–ligand complexes. The fact that hydrogen positions are not usually resolved in structures solved above the ultrahigh resolution limit (~ 0.8 Å) leads to placement of hydrogens in PDB structures based on chemical assumptions if not default settings of software. Martin⁴ discussed examples where a minor or less stable tautomer was found in the macromolecular binding site. Neutron diffraction data provide visibility of protons but are still a rarity in the PDB (<200 structures out of >157,000 PDB structures as of the time of this writing).

Tautomerism is also not a rare occurrence in organic chemistry. We showed previously that among approximately 103 million compounds aggregated from 150 or so small-molecule databases more than 66% of the molecules are susceptible to some kind of tautomerism, based on a subset of the transform rules presented in this work.⁵

Another area of chemoinformatics (though not unrelated to the foregoing) in which tautomerism has become increasingly important in the past 15 years or so is that of compound identifiers and structure representations. The difference between these two concepts as well as their frequent overlap in practice has been detailed elsewhere.⁶ The widely used SMILES strings⁷ are neither designed to be, nor are in practice, tautomer invariant, notwithstanding the fact that SMILES are often used for compound set deduplication and database overlap analyses—with usually incorrect results compared to the outcome if tautomerism had been taken into account. However, if no practical and comprehensive tautomer-invariant approach is available, there is no easy way to determine if something may be wrong with the results of the analysis⁴; i.e., we have a bit of a Catch-22 situation, where the resolution of one issue requires the other and vice versa.

Realizing this, the International Chemical Identifier (InChI) and its hashed version, the InChIKey,^{8,9} initially developed at NIST and subsequently sanctioned by IUPAC, were from the beginning (early 2000s) intended to be tautomer invariant. The way the InChI algorithm was coded, however, implemented tautomer invariance only partially. The issues are two-fold: (1) Well-known types of tautomerism such as keto–enol tautomerism are not active by default in the so-called Standard InChI but need to be turned on by the user, yielding a Nonstandard InChI[Key].¹⁰ (2) Many rarer types of tautomerism (such as 1,4-oxime/nitroso tautomerism) are not covered at all by the current InChI algorithm.⁸ In recognition of these shortcomings, an IUPAC InChI Working Group was initiated in 2012, tasked with developing recommendations for the Redesign of the Handling of Tautomerism in InChI V2.¹¹ (One of the current authors [M.C.N.] is chairperson of this Working Group.) The present work can therefore be seen as an important scientific backdrop for the Working Group's final decision and output though it is not constitutively dependent on the IUPAC project.

It is important to emphasize that in all these chemoinformatics efforts and resources involving tautomeric transforms, invariance, and enumeration the goal is usually not (or certainly not only) to predict the one, or the few, low-energy “canonical” tautomer(s) of a molecule even though reliably transforming exotic high-energy tautomeric forms into a low-energy standard form is certainly desirable. A task at least equally important in practice is to make sure that input structures encountered in, for example, substance registration systems are recognized as already in the database, even if drawn as a very “strange,” i.e., from a physics point of view high-energy, tautomer. A similar task arises in any large-scale merge of small-molecule databases, for example, in the context of corporate mergers.

All in all, we surmise that tautomerism, although in principle a well-known phenomenon, is “unfinished business” in chemistry in several respects: (1) At the QM level: Recapitulating the entirety of the condition-dependency of tautomerism is an unsolved challenge, and large-scale exploration of tautomerism at the QM level is in its infancy.¹² (2) At the

chemoinformatics level: The rule-based approaches cover only a subset of the physically possible types of tautomerism; attempts at predicting low-energy tautomer(s) based on rapid chemoinformatics approaches have so far proven unsatisfactory and/or not sufficiently general.⁵ (3) At the experimental level: While numerous experimental studies of tautomerism exist, they do not represent a systematic corpus of analyses and therefore present challenges for constructing training sets for computational approaches in that their methodologies, reported experimental details, and degree of quantitateness of results vary greatly.

To help address the latter issue of (lack of) systematic experimental data, we have created, and made publicly available a tautomer database comprising more than 2800 tautomeric tuples extracted from publications reporting experimental studies of tautomerism of small molecules. It is available for free download from <https://cactus.nci.nih.gov/download/tautomer/>. Details of the creation, curation, and structure of this database as well as numerous analyses of its contents are reported in the accompanying publication.¹³ We will call it for short “Tauto DB” in the following. Data from this Tauto DB have been used in the generation of tautomeric rules reported in this paper, and conversely, it has been augmented by information about novel types of tautomerism that were identified after the initial Tauto DB creation activities.

The goal of the present study is to present a comprehensive set of tautomeric transforms. Each should either be well-known and frequently encountered (such as keto–enol tautomerism), which will be termed “common” (rules) in the following, or supported by experimental evidence if it is a more “rare” type of tautomerism observed in specific cases.

One may ask about the relevance of such rare transforms for chemoinformatics tools such as InChI that are designed for general application to a wide variety of data sets in a wide variety of situations. One should keep in mind that if such a comprehensive set of tautomeric transforms spanning common to rare rules is applied to the very large databases of small organic molecules (100 million or more compounds) nowadays available, one finds examples of molecules that are amenable to that type of transformation even for the “rarest” of rules. This means that by virtue of simple combinatorial expansion of analogs of such example structures (provided the modifications would not affect the matching of the transform’s substructure patterns), millions more of molecules amenable to that transform could be easily constructed. In this sense, “rarity” of a rule is to some extent a function of the contents of existing databases.

Finally, it needs to be emphasized that the aforementioned IUPAC Working Group’s task is not to provide an implementation of any tautomeric rule in computer code (for an eventual InChI V2). The Working Group is solely tasked with providing recommendations of what types of tautomerism should be included in InChI V2 based on *chemical* grounds and not how these recommendations should be implemented. The fact that at an eventual coding stage it may become unavoidable to modify details of the transformation behavior of rules, say, for algorithm efficiency reasons or due to potential conflicts with other existing or new InChI features and extensions of coverage, is acknowledged but likewise not topic of this study.

METHODS AND DATA

Nomenclature.

The set of rules discussed in the following are subdivided into three classes, with concomitant naming conventions: (1) Prototropic transforms, called PT_*nn*_*mm*, with *nn* being the number of the rule and *mm* being a possible subversion. In the majority of cases, there is only one subversion, with subversion indicator 00. For example, the (one) rule encoding nitroso/oxime tautomerism is PT_16_00. (2) Transforms encoding ring-chain tautomerism are named RC_*nn*_*mm*, with *nn* and *mm* having the same meaning as above. (3) Rules based on valence tautomerism are named VT_*nn*_*mm*, with again *nn* and *mm* as above.

Identifiers, Hashcodes, and Algorithmic Approaches.

The analyses of tautomerism were performed with the chemoinformatics toolkit CACTVS.¹⁴ Version 3.4.6.33 and 3.4.8.6 of CACTVS were used. CACTVS allows the user to calculate a number of identifiers that are hash codes computed from a given chemical structure (in the parlance of CACTVS: Ensemble). These identifiers differ in that they are sensitive to different chemical features such as stereochemistry, presence of isotopically labeled atoms, formal charges in the input structure, etc. One of these features is tautomerism; i.e., if tautomerism invariance is turned on, the identifier returned by CACTVS is the same for all possible tautomers that can be enumerated based on the tautomeric rule set active at the time of execution. One such tautomer-invariant identifier is called E_TAUTO_HASH (the “E_” standing for: Ensemble property). Conversely, E_ISOTOPE_STEREO_HASH128 is an isotope-sensitive and stereosensitive but not tautomer-invariant ensemble hashcode with 128 bit length (default hashcode length is 64 bit), which was also used in some of the analyses reported below.

It is possible for the (experienced) user of CACTVS to change the set of rules that is active for a given identifier at any time, from limiting oneself to just one of, for example, the standard rules to addition of an arbitrary number of new rules.

It is possible but not mandatory to use identifiers or hashcodes in tautomerism-related algorithms in CACTVS. Enumeration, counting, structural comparisons, and other processing of generated tautomers can also be performed entirely at the ensemble level. This approach was also used in the analyses reported below.

Rules Expressed as SMIRKS.

All tautomeric transforms presented in the following are expressed as SMIRKS strings.¹⁵ It should be noted that CACTVS allows, and some of the rules use, CACTVS-specific extensions to the standard Daylight SMARTS syntax (most notably the atom attributes “e”: ring pi electron count of all ESSSR rings the atom is part of; “z”: required number of heteroatom neighbors; “a”: number of aromatic rings the atom is a member of; and “{}”: range for every attribute that can take a count). Similarly, the application of any SMIRKS for transforming a start structure into one or several result structures (performed by the CACTVS “ens transform” command) is governed by several flags and command parameters

that can have a significant influence on the outcome of the command execution. Note that in CACTVS, transform schemes can be applied in a bidirectional manner, i.e., both sides of the SMIRKS are independently matched and, if the match is successful, transformed to the other side. This mode is used for all tautomeric transforms. We list the flags used in the context of the tautomerism transforms in Table S1 in the Supporting Information. For more in-depth information, we refer to the CACTVS Full Reference manual.¹⁶ The Supporting Information also reports on (partially successful) attempts to adapt our rules to parse in the cheminformatics toolkits CDK and RDKit (whose default SMIRKS processing differs from CACTVS), by applying both limited source code modifications to these toolkits and minor changes to the SMIRKS.

The handling of stereocenters in CACTVS in the context of tautomerism is currently partially handled outside the rules themselves: (a) If one starts from an achiral compound, and generates a potential stereocenter, this stereocenter is made undefined. (b) If one starts from a stereocenter, and it changes, it is also made undefined. (c) Furthermore, in the tautomer set, all stereocenters which are flattened in any result structure are flattened in all compounds of the set, even if no transform touched it in the specific rule set applied to arrive at this compound.

Existing Rules.

CACTVS comes with a predefined set of currently 20 tautomeric transforms. All of them are prototropic rules. They are listed in Table 1 (rule example and SMIRKS) and Table 2 as rules PT_02_00 through PT_21_00. (There is no transform “PT_01_00” since Rule 1 was merged with another rule in the past.) This is the rule set the toolkit’s user will invoke when enumerating the full set of possible tautomers for a given start structure, in which case these 20 rules are applied in a *multi-step* reaction mode; i.e., if more than one interconvertible group exists, all intermediate structures are generated. All these intermediate structures are retained and are again subjected to all 20 transforms, etc., in an iterative and exhaustive manner, i.e., until no new (tautomer) structure is generated. This process, which can suffer from combinatorial explosion for complex molecules, can be limited to a user-defined maximum number of generated distinct tautomers, a maximum number of analyzed tautomers, and/or a maximum amount of CPU time per “ens transform” command execution. Note that this is the default procedure of CACTVS’s full tautomer enumeration, not the way we use these rules in the following analyses.

In those, our calculations are based on a single-step approach of producing tautomers from all possible matches of a given transform to a target structure. These generated tautomers are not subjected to rematching. We did not apply all transforms together at any step. Even though this vastly reduces the chances of combinatorial explosion, we used the following limits in light of the very large numbers of analysis and thus required CPU time: Maximum for generated tautomers was set to 10; CPU time per transform was set to 30 s.

The first 11 rules (PT_02_00 to PT_12_00) are generally the most common rules, each matching at least approximately 1% of any typical small-molecule database tested (Table 3). They comprise rules for keto–enol tautomerism as well as for hydrogen migration between heteroatoms, including those in aromatic systems, via odd-numbered H-shift paths

ranging in length from 3 to 11. Note that in spite of the name of PT_02_00, “1,5 (thio)keto–(thio)enol,” the vast majority of cases of keto–enol tautomerism are actually covered by rule PT_06_00, “1,3 heteroatom H-shift.”

New Rules.

All rules beyond PT_21_00 as well as all rules with a subversion greater than 00 (such as PT_11_01) are “new” in the sense that they do not exist in the standard CACTVS rule set. Three main sources—to some extent overlapping—were used to extract rules from (1) individual publications including book chapters, (2) the “Tauto DB¹³” mentioned above, which itself is a collection of cases of tautomerism extracted from experimental literature (though not with the primary objective of finding “new” types of tautomerism), (3) our previous work on ring–chain rules: A few rules covering well-known cases of ring–chain inter-conversions such as those of sugars (pentoses and hexoses) as well as a few other types involving 5- or 6-membered heterocyclic endocyclization were taken from Guasch et al.¹⁷ They are part of a larger set of transforms, which had been developed not primarily based on experimental work but the well-known work by Baldwin¹⁸ on rules to predict the relative facility of ring forming reactions. The entirety of these rules, RC_01_00 to RC_11_00 in both the current and Guasch’s nomenclature, which cover the majority of ring–chain tautomerism cases, have more than one variant (such as RC_05_00 to RC_05_04), yielding a total of 38 SMIRKS transforms (Table 4). From among these rules, we have included here RC_03_00, RC_04_01, RC_04_02, RC_09_00, and RC_10_00. Note that a few subrules were developed for this study as “relatives” of the original Guasch rules and thus were named as subrules of the RC_00_00 to RC_11_00 set: RC_03_03, RC_03_04, RC_04_04.

Where possible, we have evaluated at least two literature references providing experimental evidence for each of the new rules (number of references per rule: 1–5). For space reasons, this list of references is available as Table S2 in the Supporting Information. It is also available at https://cactus.nci.nih.gov/tautomerizer/rules_ref.html.

The guiding principles in the mostly manual process of creating the new rules from the literature sources are as follows:

1. Get a diverse set of molecules involved in the particular type of tautomeric equilibrium.
2. Identify the part of the molecule involved in the hydrogen migration (1,3 H-shift, 1,5 H-shift, etc.), ring closing, or ring opening (for ring–chain and valence tautomers).
3. Identify whether hydrogen migration involves any aromatic atom and/or any other polar group near the migrating hydrogen.
4. Identify whether during transformation any formal charge is created, removed, or preserved.
5. Write SMIRKS using DAYLIGHT and CACTVS attributes based on above-mentioned points. Test written SMIRKS on the diverse set of molecules we collected in the first step. Also check reproducibility of generation of reagent

and product side tautomers from each other, i.e., check if the matching and transformation using the left side as well as using the right side of the SMIRKS both work correctly.

6. Finally, we pulled out some examples from chemical databases in order to check what kind of hits we obtained. Whenever we saw some unusual hits, then the SMIRKS was modified to exclude such undesired hits.

Occurrence Rates and Databases Analyzed.

We define as the “occurrence rate” of each rule in a given database the number of records in that database that matched either the left side or the right side pattern of the rule’s SMIRKS (or both). No counting of possibly multiple matches of each pattern in an input molecule was performed.

Occurrence rates were determined for the databases listed in Table 5.

We chose these databases (Table 5) in order to cover a wide variety of types, sizes, and purposes of small-molecule collections, encompassing experimentally determined structures, drugs, commercially available screening samples, assayed compounds, and others. All databases are publicly available except (the organic part of) the CSD and CSDB. The latter is to a large part a combination of PubChem structures plus screening samples from the ChemNavigator *IRL* database.^{26,28}

The total size of the databases analyzed for the occurrence rate analyses was nearly 401 million. This is simply the sum of the counts of the individual databases. No attempt was made to reduce either the aggregated collection nor any individual database to a unique subset, not in the least because such a uniqueness analysis is dependent, among other things, on whether tautomeric deduplication is applied and if yes, by what rule set—which is after all the very thing we want to study in this project. It also simply represents the reality of many large databases, i.e., that the user encounters duplicate structures present in the database for a variety of reasons.

Tautomeric Conflicts.

We define a tautomeric conflict as the occurrence, in a given database, of two or more records labeled by the database provider as structurally different entries, whereas the set of tautomeric rules applied indicates that these structures are just tautomers of each other. For example, for a chemical products vendor, this would mean that our rules classify (structurally) different catalog items as compounds that are just drawn as different tautomers but in reality are “the same stuff in the bottle.” A straightforward way to detect such conflicts is to search for compounds in a database that have the same tautomer-invariant but different tautomer-sensitive hashcodes.

Orthogonality of Rules (Overlap Analysis).

We call two tautomeric rules orthogonal to each other if no molecule exists for which these two rules generate the same tautomer. While orthogonality of rules is desirable both in principle and in practice simply for efficiency and computer resource reasons, this is not

mandatory to make a rule set useful and fully applicable. (Even the standard CACTVS rules are by no means fully mutually orthogonal.) For example, more-complex molecules can have several paths of differing lengths by which the proton migration can occur, thus triggering more than a rule for that specific transformation. To determine orthogonality between two rules, we essentially proceed as follows: We analyze the cases in which a tautomer generated from the start structure by rule 1 was also generated by any other rules. We make sure we count only unique occurrences of this event. This ensures that the overlap count cannot exceed the size of the database analyzed or, expressed as percentage, cannot exceed 100%. The precise value of the overlap count for each rule pair is thus dependent on the database analyzed. For the most part, we do not see large variations between databases in the overlap percentages for sufficiently common rules.

Comparison of Rules with Handling of Tautomerism by Current InChI.

As mentioned above, an important aspect of, and significant part of the motivation for, this study was the assessment of the rules *vis-à-vis* current InChI (and by extension, InChIKey), v.1.05, and its handling of tautomerism. We therefore analyzed how comprehensively InChI recapitulates each of our rules. The first of these analyses was defined as the statistics of how many of the tautomers enumerated by a rule for each structure taken from a given database ("start structure") had the same InChI as the start structure. This is in principle a binned statistics: If a start structure has, say, five different rule-based enumerated tautomers, the degree of recapitulation can be 0, 1, 2, 3, 4, or 5. Since more-complicated molecules can have tens if not hundreds of rule-enumerated tautomers, explicit categorization of all possible different degrees of overlap would become unwieldy to the point of uselessness. We therefore simplified the categorization of InChI recapitulation for each rule into just three cases: No InChI match: none of the rule-generated tautomers had the same InChI as the start structure; Partial InChI match: At least two but fewer than all of the structures from the set of tautomers (including the start structure) had the same InChI as the start structure; Complete InChI match: All of the tautomers (including the start structure) had the same InChI (Table 6). We further condensed the cases of Partial InChI match and Complete InChI match into the class "Pass" while No InChI match was classified as "Fail."

In addition to the above rule-specific InChI recapitulation analysis, we also looked at the overall InChI performance *vis-à-vis* all rules for each database, i.e., provide an overall picture how all molecules of databases behave relative to StdInChI and NonStdInChI (Table 7). Each molecule was evaluated by applying all 86 rules. We categorized a molecule's behavior to InChI into three main cases: (1) Complete pass: if start structure InChI matched with all enumerated tautomers generated by at least one rule but without any failure by another rule (i.e., only pass for one or more rules), (2) Partial pass: if start structure InChI matched with some but not all enumerated tautomers generated by at least one rule but without any failure by another rule (i.e., only partial pass for one or more rules), (3) Complete pass for one rule and partial pass for other: if start structure InChI matched with all enumerated tautomers generated by at least one rule and matched with fewer than all enumerated tautomers generated by any other rule but without any failure by any rule (i.e., molecule passes for one or more rules along with partial passes to other rule(s) too). In

addition to these three cases, one has three more cases if these scenarios combine with failure to any rule.

For reasons of efficiency, we set the maximum number of generated tautomers to 10. The number of cases observed for each rule for tautomer counts from 1 to 10 are given in Spreadsheet S1 and Spreadsheet S2 of the Supporting Information (columns T to AC). In practically all cases, the one-tautomer count was higher than any of the corresponding 2- to 10-tautomer counts and in many cases higher than the sum of the 2- to 10-tautomer counts. Out of the 400+ million structures analyzed from nine databases, there were a total of 0.63 million cases that generated 10 tautomers and thus indicate that there may be 11 tautomer(s). If any molecule generates more than 10 tautomers, these 11th and higher tautomer(s) will not affect the InChI success rate much because their InChI match or partial match will add to Total InChI pass (in 2/3 of the cases of 0.63 million). If the InChI of 11th and higher tautomer(s) fail along with all previous tautomers, then this will add to Total InChI fail (1/3 of 0.63 million)

We note here that the details of this analysis are more complicated than described here. For example, there were cases where the InChI calculation for the start structure itself or any of the enumerated tautomers failed. We refer the reader to Spreadsheet S1 and Spreadsheet S2 of the Supporting Information for the complete data plus more-detailed explanations of all columns of this analysis.

This analysis was performed separately both for Standard InChI as well as for Nonstandard InChI, where the tautomerism-related options KET and 15T were turned on. As for the previous analyses, the precise quantitative statistics are dependent on the database evaluated, i.e., are not an invariant of each rule per se.

Comparison with Tautomeric Systems Identified by Other Approaches.

We analyzed a set of 4158 tautomeric systems extracted from ChEMBL 24.1 via a SMILES-based tautomer hash.²⁹ We gratefully acknowledge receiving this set from Noel O'Boyle and Roger Sayle (NextMove Software, Cambridge, UK). It was generated with the following procedure: For each molecule, tautomeric systems were found using a flood-fill procedure to identify substructures that consisted solely of donor, acceptor, or sp² atom types as described by Sayle and Delany.³⁰ For each substructure, a SMILES-based tautomer hash was generated along with the canonical SMILES for the substructure. This allowed different tautomeric forms of the same substructure to be collated based on the tautomer hash.³¹

The set extracted from ChEMBL 24.1 contained tautomeric tuples ranging in size from 2 to 6. The majority of tuples (3824 cases) had 2 tautomers, plus tuples with 3 (311), 4 (19), 5 (3), or 6 (1) tautomers, respectively. We analyzed these systems as to which rule(s) and/or rule combination(s) could effect the transformation between the members of each tuple, or if the system was too complicated for this type of detailed analysis and could have led to a combinatorial explosion, we simply tested if any path was possible with our rules between the first and any other tautomer of a tuple. The table with these systems (and how often it was found in ChEMBL 24.1), as well as the results of our transform analysis, is available as Spreadsheet S4 in the SI.

Tautomerizer Web Service.

To offer a convenient way to test these rules with various input structures, and to simply offer to the public the capability of applying them to any user molecule, we have created a web tool called Tautomerizer on our web server at <https://cactus.nci.nih.gov/tautomerizer/>. In addition to the web page with the input form and Help and Introduction pages, individual rule's pages are provided that present an interconversion diagram for an example molecule, a brief summary of some of the experimental evidence we found, and references to such papers, as well as one Rules Sources page where we have assembled these references for all new rules (PT_00_22 and higher). The only molecular input format currently allowed is SMILES. The user can choose between single-step and multi-step execution of each rule. We note that in contrast to the standard enumeration of tautomers in CACTVS, which applies all transforms exhaustively and recursively (i.e., creates a complete tautomer network), this tool applies each transform by itself (though repeatedly if applicable and requested by the option "multi-step"). The user also can flexibly select which rule(s) should be activated for their molecule (Figure 1):

- "Activate all rules": Select all transforms (standard and new rules) to be applied to the input molecule.
- "Activate 20 standard rules": Select only the 20 standard transforms (rules 2 to 21).
- "Activate only new rules": Select only the 60+ new transforms (rules 22 and higher).
- "Enter your own rule as SMIRKS": This option allows one to enter one's own transform/rule for the Tautomerizer to apply to the input molecule. One can also use this option to test modifications of our transforms.
- "Activate custom rule set via following checkboxes": Manually select any number of transforms from the 80+ transforms to apply them to the input molecule.

For additional explanations and instructions, we refer to the Help page of the service.

Scripts and Other Code Used in This Project.

In addition to the SMIRKS of the tautomeric transforms, all scripts used to generate the results of the analyses outlined above are also provided. They are made available in the Supporting Information as CACTVS Tcl scripts. For the most part, these are pieces of code written in Tcl, the language used for one of the scripting interfaces of CACTVS. In addition, a number of Linux pipes were used.

RESULTS AND DISCUSSION

We have compiled a comprehensive set of tautomeric transform rules, based on a multitude of experimental references comprising research papers, reviews, book chapters, and other sources. We have tried to provide as comprehensive a coverage of possible types of tautomerism as possible; though of course, due to the nonsystematic nature of studies related

to tautomerism, there is no guarantee that yet other types could not be identified. It is also clear, as for example evidenced by the nonzero overlap between our rules, that the rules, being strictly pattern-based SMIRKS, could be structured differently to cover essentially the same chemistry of tautomerism.

Occurrence Rates.

Table 3 makes it clear that rules PT_ *nn*_00, with *nn* = 2 ...12, which we already labeled above as “common,” are indeed found to be applicable to large numbers of structures: greater than 1 million for each rule in the combined 401 million compound set (except PT_11_01 to PT_11_04). PT_06_00 (“1,3 heteroatom H-shift”) occupies the top spot, with more than 70% of the molecules analyzed being amenable to it. As already noted, this rule covers the vast majority of cases of keto–enol tautomerism, arguably the best-known type of tautomerism. Among the new rules, the first two, PT_22_00 (“imine/imine”) and PT_23_00 (“1,5 furanones”), stand out as also having a significant number of matches, more than 6 million out of 401 million. The new rules PT_11_01 to PT_11_04 involve long-range hydrogen migration via 1,13, 1,15, 1,17, and 1,19 H-shifts, respectively. Out of these, PT_11_01 had a significant count of about 0.5 million, and the others had counts in the range of 40,000–250,000, with a very approximate halving of the count for each increase in the migration length by two atoms.

Two ring–chain rules RC_03_00 and RC_04_01 are amenable to 62 and 23 millions molecules, and these rules deal with ring–chain tautomerism of pentose and hexose sugar-type molecules, respectively. Rule RC_04_02, which includes ring–chain tautomerism of warfarin-like molecules, had 21 million hits. In addition to these, rules RC_09_00 (5-membered endocyclization) and RC_10_00 (6-membered endocyclization) had matches to more than 1 million molecules. Out of 11 valence rules, only one rule VT_02_00 (tetrazole/azide interconversion) had a significant match rate, being amenable to 2.8 million molecules.

All other rules, whether prototropic, ring–chain, or valence tautomerism, show occurrence rates below 1%. Still, in absolute numbers, many of these rules have thousands of representatives in the 401 million combined database. Only 15 rules had fewer than 900 matches, and only one single rule, PT_40_00 (“tetrad phosphorus-carbon”), had consistently zero hits across all tested databases. This rule is one of a handful in our collection whose pattern requires a “nonstandard” element in the sense of not being part of the core elements found in drugs: H, C, N, O, S, F, Cl, Br. PT_40_00 requires P and so does PT_21_00, PT_34_00, RC_12_00, RC_18_00, RC_24_00, VT_09_00, and VT_10_00. Boron is required by rules RC_03_03, RC_16_00, and RC_18_00. Rule PT_38_00 requires Si. No rule requires or even contains any halogen. Migration of halogens, methyl, and other larger groups has been reported but was outside of the scope of this study.

Perhaps along these lines, we note the interesting fact that the CSD was devoid of examples for 12 rules—but so was the 6 times larger ChEMBL database (no examples for 15 rules) and also the yet approximately 5 times larger AMS (no examples for 14 rules). (There is significant overlap but not identity between these sets of example-free rules.) One can speculate that this may be due to the nature of the two latter databases, both being focused

on drug-like molecules, whereas the crystallographically solved structures in the CSD cover a larger spectrum of chemotypes.

Overlap between Rules.

To simplify the discussion, we focus on the numbers obtained for PubChem (Table 8), assuming that this largest of the analyzed public databases is representative of current chemical space in general. The entirety of our overlap analysis is available in Spreadsheet S5 in the Supporting Information. Table 8 shows that the vast majority of overlap is concentrated within the “common” subset of the standard rules (PT_02_00 to PT_12_00), not only in terms of absolute counts but also by percentage of each rule’s coverage counts for all databases subject to this analysis. In general, there was only limited qualitative difference in overlap statistics for the other eight databases vs those for PubChem. In terms of the largest differences, the overlap between PT_06 and PT_12 was in the range from 18.69% to 64.03%, between PT_11_02 and PT_11_04, it ranged from 31.33% to 100%.

As already noted, PT_06_00 is the most common of all rules. It is thus not surprising that it also had the highest number of cases of overlap with other rules: nearly 23 million. Next-prolific in this sense were PT_03_00, ~9M; PT_05_00, ~8M; and PT_09_00, ~7M. PT_06_00 cases were a near-complete superset of the cases for PT_05_00. Still, there were 41,653 molecules uniquely amenable to PT_05_00 vs PT_06_00, which represents a higher absolute number than for many of the truly rare rules, thus providing some *raison d’être* for it as a separate transform. In any event, it is one of the standard CACTVS rules, thus not up for modification, merging, or omission in the context of this study. PT_06_00 also covers about half of the cases amenable to each of PT_03_00, PT_04_00, and PT_12_00.

Conversely to these significant overlap numbers, there were numerous rules that showed no overlap at all with any other rule, among both the (rare) CACTVS standard rules (PT_13_00, PT_18_00, PT_20_00, and PT_21_00) and the 18 new prototropic rules (e.g., PT_24_00 and PT_34_00). All new RC rules did not show overlap except rules RC_03_03, RC_14_00, RC_15_00, RC_20_00, and RC_23_00, which showed overlap for molecules in the range of 0.10%–33%. All VT rules showed practically no overlap with any rule.

We note that there is a significant overlap between rules RC_04_01 and RC_04_02. This is intentionally accepted since both rules cover important classes of molecules that are capable of ring–chain tautomerism: RC_04_01, hexose sugars; RC_04_02, coumarin type structures such as warfarin; neither of which we wanted to lose.

Tautomeric Conflicts.

We have previously analyzed a medium-size database (~6 M records) as to its tautomeric conflicts, identifying more than 31,000 cases of such conflicts, and experimentally verified more than 100 of them.² We are quite certain that any large (i.e., multi-million record) database will similarly show thousands of tautomeric conflicts. The impact of such tautomeric conflicts depends on the nature of the database. It would appear more significant if a chemical vendor offers tautomers of the same compound under different unit prices in their catalog than if one finds such conflicts in collections such as PubChem, which itself is aggregated from many different compound sets and database sources. One can ask

additionally: Are there such conflicts even in significantly smaller databases, which may have been manually curated and one would assume to be easier to clean up tautomericallly? Such a comprehensive tautomeric conflict analysis including more detailed studies including dedicated experimental analysis by X-ray crystallography of previously studied tautomeric conflict pairs² in small-molecule crystals exceeds the scope of this paper and will be the topic of a separate publication. We note here qualitatively that we have not found any database so far without any tautomeric conflict.

Recapitulation of Rules' Enumerated Tautomer Sets by InChI.

The analyses show how well InChI recapitulates the behavior of our rules and paints an interesting and varied picture. We focus first on the numbers for PubChem. The numbers for the other analyzed databases are not fundamentally different, though we note that especially the smaller databases are more likely to have no examples at all for some of the rarer rules, which of course precludes the InChI-related analysis for these rules.

Nonstandard InChI (NonStdInChI)—the more relevant identifier for an eventual expansion of InChI to a version 2—delivered “Success” rates (as defined above) between 6% and nearly 100% (average of rates: 58%) for all of the common CACTVS rules (PT_02_00 – PT_12_00) (Table 6). Still, for only three rules was the rate greater than 90% and greater than 50% for only seven rules. Standard InChI (StdInChI) success drops by varied ratios, from a few percent to a factor of nearly 10 and falls to zero for PT_03_00, PT_04_00, and PT_12_00. Values above 1% success for NonStdInChI were found among rare CACTVS and the new rules for PT_16_00, PT_17_00, PT_18_00, PT_23_00, PT_42_00, PT_44_00, PT_48_00, and RC_14_00, with an additional smattering of a few nonzero success values below 1%. Again, StdInChI shows varied degrees of drop of success rates for these rules, including to zero. The significance of the 1.85% success rate for rule RC_14_00 is doubtful due to the small absolute number of examples found in PubChem (108, out of which two were recapitulated by either variant of InChI). We note that all rules with nonzero InChI success had more cases with Complete match (all rule-enumerated tautomers had the same NonStdInChI) than Partial match, sometimes by orders of magnitude. All other rules, be they prototropic, ring-chain, or valence tautomerism, are as noncovered by current InChI as they are rare in the databases analyzed (but see the caveat above pertaining to “rarity” of rules). The overall success rate across all rules was 50% for NonStdInChI and 37% for StdInChI, explained by the fact of much higher coverage of the common CACTVS rules in PubChem (and in all other databases). One should keep in mind, however, that both Complete match and Partial match (as defined above) contribute equally; thus, the values for full recapitulation (all enumerated tautomers had the same InChI) are somewhat lower.

We note that the two new rules with absolute occurrence counts well above 1 million (in both PubChem and CSDB), PT_22_00 (“imine/imine”) and PT_23_00 (“1,5-furanones”), showed InChI recapitulation rates for NonStdInChI below 2%: 0.047% and 1.89%, respectively. If nothing else, these two types of tautomerism are therefore calling for addition to any future version of InChI.

Assessing both Tables 6 and 7 together, one sees that even NonStandard InChI recapitulates only between a quarter and one-half of the cases covered by our rules, depending on how exactly one defines overlap between these two approaches.

Comparison with SMILES-Based Tautomer Hash Applied to ChEMBL 24.1.

The analysis of the set of 4158 tautomeric systems extracted from ChEMBL 24.1 via a SMILES-based tautomer hash²⁹ showed that our rules cover essentially all the tautomeric systems in that set. Apart from a handful of doubtful structures, six cases appeared to involve migration of an unspecified group or were categorized as simply the same molecule according to the (tautomer sensitive) CACTVS hashcode E_ISOTOPE_STEREO_HASH (presumably due to molecular symmetry/rotatable substructures). Practically all the ChEMBL tautomeric systems were covered by the standard CACTVS rules PT_02_00 through PT_21_00, with most everything actually being covered by PT_09_00 or below.

We also checked how InChI[Key] performed for these tautomeric systems. StdInChIKey failed (i.e., returned different InChIKeys) in about 28% of the cases. NonStdInChIKey with 15T and KET turned on was about four times better, i.e., failed in approximately 7% of the cases (bottoms of columns T and Z, respectively, in Spreadsheet S4 in the SI).

Assessment of Rarity of Rules.

While one might draw the conclusion that rare rules are in fact synonymous to “irrelevant rules” (particularly in the context of identifiers including InChI), one thing should be kept in mind: The occurrence rates are a function of the structure contents of the databases analyzed. For example, if a database focuses on drug-like small molecules, then it is less likely that very-long-range H-shifts are even possible based on maximum path lengths in molecules. A case in point is the nonoccurrence of any cases of 1,21 H-shifts and longer H-shifts in the ChEMBL subset discussed above³¹ and the rarity of, for example, 1,13, 1,17, and 1,19 H-shifts (one case each out of 4158), whereas PubChem, which is known to contain a broader spectrum of structures than just drug-like molecules,³² showed an occurrence rate of, for example, one out of 567 for 1,13 H-shift. By the same token, occurrence rates are a function of time: if in the future, chemotypes susceptible to a nowadays “rare” type of tautomerism become, for whatever reason, more “popular” (be it actually synthesized or generated *in silico*), then this rule would become less rare. It should not be forgotten that by a simple change of substitution patterns (if not negatively impacting the possibility for the specific type of tautomerism), a near infinite number of analogs of just one single example of a molecule susceptible to even the rarest type of tautomerism can be generated as a virtual library.

Assessment *vis-à-vis* Experimental and Physics-Based Computations.

We reiterate and re-emphasize here that none of these rules takes energetics of tautomers into account in any way, neither relative energies nor energy barriers to interconversion. There is no mechanism to make SMIRKS directly aware that “energy” even exists. One could, in principle, consider using a paradigm for expressing transform rules that allows one to incorporate more chemical knowledge such as CHMTRN/PATRN³³ in order to imbue the rules with at least some pragmatic basis for decision-making as to lower-energy vs

higher-energy tautomers. However, no attempts in this direction were made in the context of this study.

We do however mention here standard CACTVS functions such as a tautomer rating property `E_TAUTOMER_SCORE` as well as a canonic tautomer selection (`E_CANONIC_TAUTOMER`), which are based entirely on chemoinformatics approaches.

The true realm that allows for quantitative calculations of energies, and thus of an attempt of at least ruling out very high-energy tautomers if not prediction of likely experimentally observable tautomer(s), is that of quantum mechanical (QM) computations that permit one to break and reform bonds involving mobile hydrogens (or other migrating groups). Large-scale computations of millions of tautomers at the semiempirical level have recently been undertaken.¹² Attractive recent approaches combine a significant number of QM computations subsequently used as a training set for machine learning models, yielding neural network potentials with QM accuracy at force field computational cost.³⁴ We are exploring these kinds of approaches for our tautomerism-related work.

Still, for all these higher-level approaches, the limitation still holds that if these computations are done for a vacuum environment, they are likely to miss the important contribution of solvent to proton-shuttling in many cases. This is but one aspect of the difficulty of how to treat the influence of conditions on tautomeric equilibria, which persists no matter what approach and level of theory is used.

Impact on, and Distinction from, InChI V2.

In the context of this work being inspired by, and informing the decision of, the IUPAC Working Group on Handling of Redesign of Tautomerism for InChI V2, several points are worth reiterating. It needs to be remembered that whereas CACTVS is a full-fledged chemoinformatics toolkit, InChI's purpose is solely to calculate an identifier from an input structure, not to output an enumeration of many possible tautomers. Also, the part of the current InChI algorithm that provides tautomer invariance is based on a very different chemistry and algorithmic approach from CACTVS's handling of tautomerism.³⁵ Even though the recommendations by the Working Group will most likely be in the form of a set of SMIRKS describing the various types of applicable tautomerism transformations (i.e., all, or a subset, of the rules described in this publication), they will then need to be translated in the appropriate code of an eventual InChI V2 program/library by the developers (which will not be the Working Group). For a variety of reasons, not in the least computational efficiency, it is highly unlikely that the code of InChI V2 will contain a SMIRKS parser.

We reiterate that the current chemistry model of InChI bases its handling of tautomerism on migration of mobile hydrogens in an otherwise fixed connectivity of heavy atoms.³⁶ This is most appropriate for prototropic rules. Adding ring-chain tautomerism rules may therefore pose significant additional challenges, even though it would be desirable for InChI to handle, for example, the well-known ring-chain tautomerism of carbohydrates. Valence tautomerism may be entirely impossible to implement without a significant change in InChI's chemistry model. We note that our rules handle numerous cases of "poster children" of tautomerism or cases specifically mentioned as not covered by InChI V1: 2-hydroxypyridine 1-oxide,⁸ Rule

PT_41_00; pentose sugars, Rule RC_03_00; hexose sugars, Rule RC_04_01; warfarin, Rule RC_04_02 (for its ring-chain interconversions).

A concern about the 80+ rules presented here could be that they constitute a too-aggressive handling of tautomerism. More accurately, such a concern should be associated with the degree of applicability to compound databases, i.e., how often equating two (or more) tautomers with each other as the “same stuff” would be confirmed by other, non-SMIRKS-based, methods. Apart from the impossibility to do this even just computationally via QM approaches let alone experimentally for today’s databases approaching the billion-compound count, we need to remind the reader that tautomerism is not an immutable compound property but a phenomenon depending on conditions and even the very purpose of the tautomeric analysis. As we already mentioned above, the synthetic chemist will have something different in mind when talking about tautomerism than the chemical repository/catalog manager—and for perfectly valid reasons. There appears at this time no simple, affordable, and scientifically rigorous approach to fully reconcile the competing if not conflicting demands on any handling of tautomerism in the different areas of chemistry. Any decision taken in this context, such as by the IUPAC Working Group on the Redesign of Handling of Tautomerism in InChI V2, will therefore be a compromise based on practical considerations.

Given that we have shown that a broadening of the scope of tautomerism along the lines of the rule set presented here will increase the number of molecules susceptible to tautomerism in any typical database by up to 3-fold relative to Standard InChI (Table 7), it is clear that InChI V2 will not just be a fine-tuning of InChI V1 but a major change. One possibility to reconcile to some degree the conflicting demands on the InChI identifier would be to bracket, in a new (V2) InChI[Key], full tautomer invariance and full tautomer sensitivity within the same identifier. The layered structure of InChI would lend itself for this naturally, whereas a tripartite (new) format of InChIKey V2 could, for example, encode the tautomeric “parent” structure in the first two blocks, with the third block specifying the specific tautomer represented in the input structure. Searches by InChIKey could then be either fully tautomer invariant (using only the first two blocks) or tautomer specific (using all three blocks). Since the version of InChI is indicated in both InChI and InChIKey itself, it should be no problem to use V1 and V2 in parallel for many years; i.e., any new format of the identifier could be phased in gradually in much the same way that the chemical table³⁷ (CT) formats V2000 and V3000 have been coexisting for several years.

We finally note here that *current* InChI appears to be already tautomerically (too?) aggressive in the above sense for some structures: e.g., pralidoxime (O/N=C/C1=[N+](C)C=CC=C1) and its Z diastereomer (O/N=C\C1=[N+](C)C=CC=C1) have the same InChIKey (JBKPUQ-TUERUYQE-UHFFFAOYSA-O) in both the Standard version and with the 15T and KET options turned on. While we did not attempt tracing of the InChI code execution to see exactly where things become the same, based on an analysis with CACTVS rules, where we see the identity of NCI/CADD identifiers³⁸ (D894A9BE897FE4C8-FICuS-01-93, same tautomer invariant identifier FICuS for both) as a side effect of tautomeric transformation, we assume this effect is tautomerism-related for InChI[Key], too.

CONCLUSIONS

We have presented evidence that tautomerism is a widespread and important phenomenon. We deem it fair to say that one finds it everywhere one looks, and that it is indeed “unfinished business” in chemistry and chemoinformatics. We note that every single database we have analyzed so far, whether multimillion-structure in size or smaller (in the hundred thousand range), contained at least a handful of tautomeric conflicts based on our rules if not thousands of them.² Virtually all of the transformations one can derive from experimental literature have at least a handful of examples amenable to this rule in large small-molecule databases. No matter whether all or only a (significant) subset of the tautomerism types presented here is ultimately chosen to be incorporated in InChI V2, this will lead to a major change in the way InChI[Key] addresses tautomerism as well as in the values, and possibly format, of the identifier itself.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Noel O’Boyle and Roger Sayle for providing the extract of tautomeric systems from ChEMBL 24 to us and for useful discussions about these cases and tautomerism in general. We thank Thomas Sander and Oya Wahl for providing us with their Tautomer Codex³⁹ and its full reference list, which allowed us to generate a few additional rules. We thank Jeff Saxe for his help in setting up the Tautomerizer web tool on the CADD Group’s web server. M.C.N. thanks the members of the IUPAC Working Group on Redesign of Handling of Tautomerism in InChI V2 for contributions to, and valuable discussions of, the Group’s mission and the steps on the way to fulfill it. This work utilized the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). This work was in part supported by the Intramural Research Program of the National Institutes of Health, Center for Cancer Research, National Cancer Institute. D.K.D., H.P., V.D., and M.C.N. received funding from the NCI, NIH, Intramural Research Program. W.-D.I. received funding from Xemistry GmbH internal research budget. 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). Kleinpeter ENMR Spectroscopic Study of Tautomerism in Solution and in the Solid State. In Tautomerism; John Wiley & Sons, Ltd., 2013; pp 103–143. DOI: 10.1002/9783527658824.ch5.
- (2). Guasch L; Yapamudiyansel W; Peach ML; Kelley JA; Barchi JJ; Nicklaus MCExperimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples. *J. Chem. Inf. Model*2016, 56 (11), 2149–2161. [PubMed: 27669079]
- (3). Warr WATAutomerism in Chemical Information Management Systems. *J. Comput.-Aided Mol. Des*2010, 24 (6–7), 497–520. [PubMed: 20372974]
- (4). Martin YCLet’s Not Forget Tautomers. *J. Comput.-Aided Mol. Des*2009, 23 (10), 693–704. [PubMed: 19842045]
- (5). Sitzmann M; Ihlenfeldt W-D; Nicklaus MCTautomerism in Large Databases. *J. Comput.-Aided Mol. Des*2010, 24 (6–7), 521–551. [PubMed: 20512400]
- (6). Peach ML; Zakharov AV; Guasch L; Nicklaus MCCChemoinformatics. In *Comprehensive Biomedical Physics*2014, 6, 123–156.
- (7). Weininger DSILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model*1988, 28 (1), 31–36.
- (8). Heller S; McNaught A; Stein S; Tchekhovskoi D; Pletnev IInChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminf*2013, 5 (1), 7.

- (9). Heller SR; McNaught A; Pletnev I; Stein S; Tchekhovskoi DInChI, the IUPAC International Chemical Identifier. J. Cheminf2015, 7 (1), na DOI: 10.1186/s13321-015-0068-4.
- (10). InChI FAQ 6.4. InCHI Trust. <http://www.inchi-trust.org/technical-faq/#6.4> (accessed August 30, 2019).
- (11). IUPAC projects. https://iupac.org/projects/project-details/?project_nr=2012-023-2-800 (accessed February 2020).
- (12). ConstrQt White Paper. ChemAlive. <https://www.chemalive.com/2019/01/11/construqt-white-paper/> (accessed February 2020).
- (13). Dhaked DK; Guasch L; Nicklaus MCTautomer Database: A Comprehensive Resource for Tautomerism Including InChI V2. J. Chem. Inf. Model2020, DOI: 10.1021/acs.jcim.9b01156.
- (14). Ihlenfeldt WD; Takahashi Y; Abe H; Sasaki SComputation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. J. Chem. Inf. Model1994, 34 (1), 109–116.
- (15). Daylight SMIRKS Tutorial. https://daylight.com/dayhtml_tutorials/languages/smirks/ (accessed February 2020).
- (16). Extended Toolkit scripting command documentation. CACTVS Tcl Scripting Introduction.https://www.xemistry.com/docs/cactvs_full.pdf (accessed February 2020).
- (17). Guasch L; Sitzmann M; Nicklaus MCEnumeration of Ring–Chain Tautomers Based on SMIRKS Rules. J. Chem. Inf. Model2014, 54 (9), 2423–2432. [PubMed: 25158156]
- (18). Baldwin JERules for Ring Closure. J. Chem. Soc., Chem. Commun1976, No. 18, 734.
- (19). Wishart DS; Feunang YD; Guo AC; Lo EJ; Marcu A; Grant JR; Sajed T; Johnson D; Li C; Sayeeda Z; Assempour N; Iynkkaran I; Liu Y; Maciejewski A; Gale N; Wilson A; Chin L; Cummings R; Le D; Pon A; Knox C; Wilson MDrugBank 5.0: A Major Update to the DrugBank Database for 2018. Nucleic Acids Res. 2018, 46 (D1), D1074–D1082. [PubMed: 29126136]
- (20). Ligand Expo. Protein Data Bank. <http://ligand-expo.rcsb.org/index.html> (accessed July 2019).
- (21). The Cambridge Crystallographic Data Centre (CCDC). <https://www.ccdc.cam.ac.uk/> (accessed December 2018).
- (22). Gaulton A; Hersey A; Nowotka M; Bento AP; Chambers J; Mendez D; Mutowo P; Atkinson F; Bellis LJ; Cibrián-Uhalte E; Davies M; Dedman N; Karlsson A; Magariños MP; Overington JP; Papadatos G; Smit I; Leach ARThe ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45 (D1), D945–D954. [PubMed: 27899562]
- (23). Aldrich Market Select. <https://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html> (accessed July 2019).
- (24). SureChEMBLOpen Patent Data. <https://www.surechembl.org/search> (accessed July 2019).
- (25). PUBCHEM<https://pubchem.ncbi.nlm.nih.gov/> (accessed October 2018).
- (26). NCI/CADD iRL-Based Database of Commercially Offered Screening Compounds. https://cactus.nci.nih.gov/download/ncicadd_irl/ (accessed July 2019).
- (27). Chemical Structure DataBase (CSDB): It is the aggregated database of structures collected by the Computer-Aided Drug Design (CADD) Group of the National Cancer Institute (NCI), <https://cactus.nci.nih.gov/chemical/structure> (accessed July 2018).
- (28). Chem Navigator. <https://www.chemnavigator.com/> (accessed February 2020).
- (29). MolHash. GitHub. <https://github.com/nextmovesoftware/molhash> (accessed February 2020).
- (30). Sayle R; Delany JMolHash PowerPoint. https://www.daylight.com/meetings/emug99/Delany/taut_html/sld001.htm (accessed February 2020).
- (31). Sander TPPrivate Communication, 2019.
- (32). Kim S; Thiessen PA; Bolton EE; Chen J; Fu G; Gindulyte A; Han L; He J; He S; Shoemaker BA; Wang J; Yu B; Zhang J; Bryant SHPubChem Substance and Compound Databases. Nucleic Acids Res. 2016, 44 (D1), D1202–D1213. [PubMed: 26400175]
- (33). Adapting CHMTRN (ChEMistry TRAnslator) for a New Use10.26434/chemrxiv.11439984.v1 (accessed January 2020).
- (34). Smith JS; Isayev O; Roitberg AEANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. Chem. Sci2017, 8 (4), 3192–3203. [PubMed: 28507695]

- (35). InChI_Source_Code_Documentation. https://www.inchi-trust.org/download/103/InChI_Source_Code_Documentation_v1.0.pdf (accessed February 2020).
- (36). InChI-1-doc.zip <http://www.inchi-trust.org/download/105/INCHI-1-DOC.zip> (accessed February 2020).
- (37). CTFile Formats. Dassault Systèmes. http://help.accelrys.com/ulm/onelab/1.0/content/ulm_pdfs/direct/reference/ctfileformats2016.pdf (accessed February 2020).
- (38). Sitzmann M; Filippov IV; Nicklaus MCInternet Resources Integrating Many Small-Molecule Databases 1. SAR QSAR Environ. Res2008, 19 (1–2), 1–9. [PubMed: 18311630]
- (39). Wahl O; Sander TTautobase: An Open Tautomer Database. J. Chem. Inf. Model2020, DOI: 10.1021/acs.jcim.0c00035.

Tautomerizer - Predict tautomers based on 80+ rules

[Introduction](#) | [Form](#) | [Individual Rule Pages](#) | [Rules Sources](#) | [Help](#)

Enter the structure in SMILES format

1. Input Structure SMILES: [Structure Editor](#)

Submit

2. Single step or Multi step:

☒ Single step ☐ Multi step

3. Activate rules:

- ☒ Activate all rules
- ☐ Activate 20 standard rules
- ☐ Activate only new rules
- ☐ Enter your own rule as SMIRKS:
- ☐ Activate custom rule set via following checkboxes:

Submit

Figure 1.
Screenshot of the web service Tautomerizer.

Table 1.

Representative Tautomeric Transform Reactions and Their SMIRKS^a

Rule number	Rule example Name	SMIRKS
PT_02_00	1,5 (thio)keto/(thio)enol	[O,S,Se,Te,X1:1]=[Cz1H0:2][C:5]=[C:6][CX4z0,NX3:3][#1:4]>>[#1:4][O,S,Se,Te,X2:1][Cz1:2]=[C:5][C:6]=[Cz0,N:3]
PT_03_00	simple (aliphatic) imine	[#1,a,O:5][NX2:1]=[Cz{1-2}:2][CX4R{0-2}:3][#1:4]>>[#1,a,O:5][NX3:1][(#1:4)][Cz:2]=[C:3]
PT_04_00	special imine	[Cz0R0X3:1]([C:5])=[C:2][Nz0:3][#1:4]>>[#1:4][Cz0R0X4:1]([C:5])[c:2]=[nz0:3]
PT_05_00	1,3 aromatic heteroatom H-shift	[#1:4][N:1][C:e6:2]=[O,NX2:3]>>[NX2,nX2:1]=[C,c:e6:2][O,N:3][#1:4]
PT_06_00	1,3 heteroatom H-shift	[CX{2-3}z{0-1},N,n,S,s,O,o,Se,Te:1]=[NX2,nX2,CX3,c,Pp:2][N,n,S,O,Se,Te:3][#1:4]>>[#1:4][CX4z{0-1},N,n,S,O,Se,Te:1][NX2,nX2,CX3z{0-1},c,Pp:2]=[N,n,S,s,O,o,Se,Te:3]
PT_07_00	1,5 (aromatic) heteroatom H-shift (1)	[nX2,NX2,S,O,Se,Te:1]=[C,c,nX2,NX2:6][C:c:5]=[C,c,nX2:2][N,n,S,s,O,o,Se,Te:3][#1:4]>>[#1:4][N,n,S,O,Se,Te:1][C,c,nX2,NX2:6]=[C,c:5][C,c,nX2:2]=[NX2,S,O,Se,Te:3]
PT_08_00	1,5 (aromatic) heteroatom H-shift (2)	[n,s,o:1]=[c,n:6][c:5]=[c,n:2][n,s,o:3][#1:4]>>[#1:4][n,s,o:1][c,n:6]=[c:5][c,n:2]=[n,s,o:3]
PT_09_00	1,7 (aromatic) heteroatom H-shift	[nX2,NX2,S,O,Se,Te,Cz0X3:1]=[C,c,NX2,nX2:6][C,c,NX2,nX2:5]=[C,c,NX2,nX2:2][C,c,NX2,nX2:7]=[C,c,NX2,nX2:8][N,n,S,s,O,o,Se,Te,CX4z0:3][#1:4]>>[#1:4][N,n,S,O,Se,Te,Cz0X4:1][C,c,NX2,nX2:6]=[C,c:5][C,c,NX2,nX2:2]=[C,c,NX2,nX2:7][C,c,NX2,nX2:8]=[NX2,S,O,Se,Te,CX3z0:3]
PT_10_00	1,9 (aromatic) heteroatom H-shift	[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,nX2:6][c,nX2:7]=[c,nX2:8][c,nX2,C:9]=[n,N,O:10]>>[n,N,O:2]=[C,c,nX2:3][c,nX2:4]=[c,nX2:5][c,nX2:6]=[c,nX2:7][c,nX2:8]=[c,nX2:9][n,O:10]
PT_11_00	1,31 (aromatic) heteroatom H-shift	[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,nX2:6][c,nX2:7]=[c,nX2:8][c,nX2,C:9]=[c,nX2:10][c,nX2:11]=[nX2,NX2,O:12]>>[NX2,nX2,O:2]=[C,c,nX2:3][c,nX2:4]=[c,nX2:5][c,nX2:6]=[c,nX2:7][c,nX2:8]=[c,nX2:9][c,nX2:10]=[c,nX2:11][nX2,O:12][#1:1]

Rule number	Rule example Name	SMIRKS
PT_11_01		[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2,C:5]=[c,nX2,C:6][c,nX2:7]=[c,C,nX2:8][c,C,nX2:10][c,nX2:12][c,C,nX2:13]=[nX2,NX2,O:14]>>[NX2,nX2,O:2][c,nX2,C:3][c,nX2:6]=[c,C,nX2:5][c,C,nX2:NX2:7][c,C,nX2:NX2:9]
PT_11_02		[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2,C:5]=[c,nX2,C:NX2:6][c,nX2:8][c,C,nX2:NX2:9]=[c,C,nX2:NX2:10][c,nX2,C:11]=[c,C,nX2:12][c,C,nX2:13]=[c,C,nX2:14][c,C,nX2:15]=[nX2,NX2,O:16]>>[NX2,nX2,O:2][c,nX2,C:3][c,nX2:C:4]=[c,C,nX2:5][c,C,nX2:NX2:6]=[c,C,nX2:NX2:7][c,C,nX2:8]=[c,C,nX2:NX2:9][c,C,nX2:NX2:10]=[c,C,nX2:11][c,C,nX2:12]=[c,C,nX2:13]
PT_11_03		[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2,C:5]=[c,nX2,C:6][c,nX2:7]=[c,C,nX2:8][c,C,nX2:10][c,nX2:12][c,C,nX2:13]=[c,C,nX2:14][c,C,nX2:15]=[c,C,nX2:16][c,nX2:17]=[nX2,NX2,O:18]>>[NX2,nX2,O:2][c,nX2,C:3][c,nX2:C:4]=[c,C,nX2:5][c,C,nX2:6]=[c,C,nX2:NX2:7][c,C,nX2:8]=[c,C,nX2:NX2:9][c,C,nX2:10]=[c,C,nX2:11][c,C,nX2:12]=[c,C,nX2:13][c,C,nX2:14]=[c,C,nX2:15]
PT_11_04		[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2,C:5]=[c,nX2,C:NX2:6][c,nX2:8][c,C,nX2:NX2:9]=[c,C,nX2:NX2:10][c,nX2,C:11]=[c,C,nX2:12][c,C,nX2:13]=[c,C,nX2:14][c,C,nX2:15]=[c,C,nX2:16][c,nX2:17]=[nX2,NX2,O:18]>>[NX2,nX2,O:2][c,nX2,C:3][c,nX2:C:4]=[c,C,nX2:5][c,C,nX2:6]=[c,C,nX2:NX2:7][c,C,nX2:8]=[c,C,nX2:NX2:9][c,C,nX2:10]=[c,C,nX2:11][c,C,nX2:12]=[c,C,nX2:13][c,C,nX2:14]=[c,C,nX2:15][c,C,nX2:16]=[c,C,nX2:17]
PT_12_00		[#1:1][O,S,N:2][c,C:2:r5:3]=[C:r5:4][c,C:r5:5]>>[O,S,N:2]=[C:2:r5:3][C:r5:4]([#1:1])(C:r5:5)
PT_13_00		[O,S,Se,Te:X1:1]=[C:2]=[C:3][#1:4]>>[#1:4][O,S,Se,Te:X2:1][C:2][#1:3]
PT_14_00		[#1:1][C:2][N+3]([O:-5])=[O:4]>>[C:2]=[N+3]([O:-5])(O:4)[#1:1]
PT_15_00		[#1:1][C:2][N:3]([O:5])=[O:4]>>[C:2]=[N:3]([O:5])(O:4)[#1:1]
PT_16_00		[#1:1][O:R:2][N+0z1:3]=[CX3:4]>>[O:R:2]=[N+0z1:3][CX4:4][#1:1]
PT_17_00		[#1:1][O:2][Nz1:3]=[C:4][C:5]=[C:6][C:7]=[O:8]>>[O:2]=[Nz1:3][C:4]=[c:5][c:6]=[c:7][O:8][#1:1]
PT_18_00		[#1:1][O:2][C:3][N:4]>>[O:2]=[C:3]=[N:4][#1:1]
PT_19_00		[#1:1][O,N:2][C:3]=[S,Se,Te:4]=[O:5]>>[O,N:2]=[C:3][S,Se,Te:v(2-4):4][O:5][#1:1]

Author Manuscript

Rule number	Rule example Name	SMIRKS
PT_30_00	sulfonamide to sulfonic acid	[#1:1][N:2][N+:3](O-:5)=O:4>>[O:4]>>[N+:3](O-:5)[O:4][#1:1]
PT_31_00	sulfone to sulfonic acid	[#1:1][CX4z1:2][CX3:3]=[CX3:4][CX3:5]=[CX3:ia:6][SX4:7]([O:4])=[O:5][CX4:7]([O:4])
PT_32_00	sulfone to sulfonic acid	[#1:1][CX4:4\$(C)[CX{3-4}]]=-[OX{1-2}]:2[OX2:3][NX1:4]>>[CX3:5\$(C)[CX{3-4}]]=-[OX{1-2}]:2[OX2:3][NX2:4][#1:1]
PT_33_00	sulfone to sulfonic acid	[#1:1][CX4:2][CX3:3]=[C:\$(CX3)[CX{2-3}]]=-[N:O]:4[OX2:5][NX1:6]>>[CX3:2]=[CX3:3][C:\$(CX3)[CX{2-3}]]=-[N:O]:4[OX2:6][#1:1]
PT_34_00	sulfone to sulfonic acid	[#1:1][CX4:2][PX4:3]=[C:\$(CX{2-3}z2)-[PX{3-4}]]:4>>[CX3:2]=[PX4:3][C:\$(CX{3-4}z2)-[PX{3-4}]]:4[OX2:4][#1:1]
PT_35_00	sulfone to sulfonic acid	[Sv2X2:1][OX2:2][#1:3]>>[Sv4X3:1]([O:1:3])=[OX1:2]
PT_36_00	sulfone to sulfonic acid	[CX3:1]=[NX2:2][OX2:3][#1:4]>>[CX3:1]=[NX3+2]([OX1-3])[#1:4]
PT_37_00	sulfone to sulfonic acid	[NX2:1]=[CX3z{2-3}:2][SX2:3][OX2:4][#1:5]>>[#1:5][NX3:1][CX3z{2-3}:2]=[SX2+3][OX-4]
PT_38_00	sulfone to sulfonic acid	[#1:1][CX4:ia:2][CX3:ia:3]=[NX3+4][SX4-5]([NX3:7])(NX3:8)=O:6>>[CX3:ia:2]=[CX3:ia:3][NX3:4][SX4:5]([NX3:7])(NX3:8)[OX2:6][#1:1]
PT_39_00	sulfone to sulfonic acid	[CX3.NX2:1]=[NX3+2]([O-:3])(CX4:4)[#1:5]>>[#1:5][CX4.NX3:1][NX3+2]([O-:3])=[CX3:4]
PT_40_00	sulfone to sulfonic acid	[#1:1][PX4:2]=[C:\$(CX3)[PX4+]:3][CX3z1:4]=[O:5]>>[PX3:2][C:\$(CX3)[PX4+]:3]=[CX3z1:4][OX2:5][#1:1]
PT_41_00	sulfone to sulfonic acid	[#1:1][SX2.NX3.OX2:IR:2][CX3:cr{5-6}:3]=[NX3+T{5-6}:4][OX1-5]>>[SX1.NX2.OX1:IR:2]=[CX3:cr{5-6}:3][NX3r{5-6}:4][OX2:5][#1:1]

Rule number	Rule example Name	SMIRKS
PT_42_00		[#1:1][CX4:4]1[NX3.O.S.Se:5][CX3:6](=[O:7])[CX3:2]=[CX3:a0:3]1>>[#1:1][CX4:2]1[CX3:a0:3]=[CX3:4][NX3.O.S.Se:5][CX3:6]1=[O:7]
PT_43_00		[#1:1][CX4:2][c:5]1=[c:9]2[c:8]=[c:7][c:6]=[c:11][c:10]2=[c:4][#8:3]1>>[#1:1][CX4:4]1[#8:3][CX3:5](=[CX3:c2:2])[c:9]2=[c:10]1[c:11]=[c:6][c:7]=[c:8]2
PT_44_00		[#1:7][CX4:\$([C]C)[C]#N:).\$(C)[C](=[O])[O]:6[c:5]1=[c:R1:4][c:3]=[c:2][nX3:1]1>>[#1:7][CX4R1:4]1[CX3:3]=[CX3:2]([CH3:3))([CH3:4))=[CX3:\$([C]C)[C]#N:).\$(C)[C](=[O])[O]:6
PT_45_00		[#1:1][CX4:2]([CH3:3))([CH3:4))([CX3R1r(5-8):c:z0:5]=[CX3R1r(5-8):c:6][CR(1-):c:7]>>[CX3:2]([CH3:3))([CH3:4))=[CX3R1r(5-8):z0:5][CX4R1r(5-8):6]([CR(1-):7))/#1:1
PT_46_00		[#1:8][CX4:\$([C]S)(=[O])[O]:7)[C:1]1=[C:6][C:5]=[NX2:+0:4][C:3]=[C:2]1>>[#1:8][N:4]1[C:3]=[C:2][C:1](=[CX3:\$([C]S)(=[O])[O]:7))([C:6]=[C:5]1
PT_47_00		[#1:10][CX4:8]1[#7X2:9]=[CX3:7][c:6]2=[c:5]1[c:4]=[c:3][c:2]=[c:1]2>>[#1:10][#7X3:9]1[c:7]=[c:6]2[c:1]=[c:2][c:3]=[c:4][c:5]2=[c:8]1
PT_48_00		[#1:12][OX2:10][c:2]1=[c:3][c:4]=[c:5][c:6]2=[c:1]1[C:8](=[O:11])[O:7][C:9]2>>[OX1:10]=[C:2]1[C:3]=[C:4][CX4:5]([#1:12])[C:6]2=[C:1]1[C:8](=[O:11])[O:7][C:9]2
PT_49_00		[#1:9][OX2:8][NX3R1r5:1]([aR(1-):r(5-):2])[eR1r5:5]=[cR1r5:4]([aR(1-):r(5-):3])[CX3:6]=[O:7]>>[OX1-8][NX3+R1r5:1]([a.AR(1-):r(5-):2))=CR1r5:5][CX3R1r5:4]([a.AR(1-):r(5-):3))=[CX3:6][OX2:7]1/#1:9
RC_03_00		[#1:1][O.N.S.Se.Te:2][#6R1:c:3]1[*:4]~[*:7]~[R1:6][O.N.S.Se.Te:R:5]1>>[O.N.S.Se.Te:2]=[C:IR:3][R(0-1):4]~[R(0-1):7][IR:6][O.N.S.Se.Te:5]1/#1:1
RC_03_03		[#1:1][OX2:2][BX3:3]([OX2])[cr6:4][cr6:5][CX3:6]=[OX1:7]>>[OX2:2]1[BX3:3]([OX2])([cr{5-6}:4][cr{5-6}:5][CX4:6]1[OX2:7]1/#1:1
RC_03_04		[#1:1][OX2:2][CX4:IR:3][CX4:4][CX4:5][CX3:c:6]=[CX3:c:\$([C]C)[CX3](=[OX1])[OX2]:7]>>[OX2:2]1[CX4:3][CX4:4][CX4:5][CX4:6]1[CX4:\$([C]C)[CX3](=[OX1])[OX2]:7]1/#1:1

Author Manuscript

Rule number	Rule example Name	SMIRKS
RC_18_00		[#1:1][OX2:2][CX4:3][c:4]=[c:5][P:6]=[OX1:7]>>[OX2:2][CX4:3][c:4]=[c:5][P:6][OX2:7][#1:1]
RC_19_00		[#1:1][CX4:2]([NX3+:3]([O-:5)=[O:4])[CX4:6][CX4:7][CX3:8]=[CX3:9]>>[CX3:2]([NX3+:3]([O-:5)=[O:4])[CX4:6][CX4:7][CX3:8][CX4:8][CX4:9][#1:1]
RC_20_00		[#1:1][OX2,NX3:2][CX4:R,CD4:R2:3][CX4:4][NX3+:5]([OX1-:7])=[CX3:6]>>[OX2,NX3:2][CX4:3][CX4:4][NX3:5]([OX2:7][#1:1])[CX4:6]1
RC_21_00		[#1:1][CX4:2]([NX3+:3]([O-:5)=[O:4])[CX4:6][CX4:7][CX3:8]=[CX3:9]>>[CX4:2]1([NX3+:3]([O-:5)=[O:4])[CX4:6][CX4:7][CX3:8][CX4:8][CX4:9][#1:1]
RC_22_00		[#1:1][OX2:2][NX2:3]=[CX3:4][CX4:5][NX3+:6]([OX1-:7])=[CX3:8]>>[OX1-:2][NX3+:3]1=[CX3:4][CX4:5][NX3:6]([OX2:7][#1:1])[CX4:8]1
RC_23_00		[#1:1][OX2:2][CX4:3][CX4:4][CX4:5][NX3+:6]([OX1-:7])=[CX3:8]>>[OX2:2]1[CX4:3][CX4:4][CX4:5][NX3:6]([OX2:7][#1:1])[CX4:8]1
RC_24_00		[#1:12][O:10][C:9][cr6:6]=[cr6:1][PX3z0:7]>>[#1:12][PX3z0:7]>>[#1:10][C:9][cr{5-6}:6]=[cr{5-6}:1]1
VT_01_00		[OX1:1]=[CX3R1:2][CX3R1:3]=[SX{1-2}z{0-1}];!R:4>>[OX2:1]1[c:2]=[c:3][SX{2-3}z{1-2}:4]1
VT_01_01		[SX1:1]=[CX3:2][CX3:3]=[SX1]>>[SX2:1]1[C:2]=[C:3][SX2:4]1
VT_02_00		[NX2,nX2:1]=[CX3,cz{2-3}:2][NX2z1:3]=[NX2z2+:4]=[NX1-:5]>>[NX3:1]1[Cz{2-3}:2]=[NX2z1:3][NX2z2:4]=[NX2z2:5]1

Author Manuscript

J Chem Inf Model. Author manuscript; available in PMC 2021 September 23.

Table 2.

Tautomeric Transform Classification

Rule number	Rule name
Existing CACTVS rules (Prototropic tautomerism)	
PT_02_00	1,5 (thio)keto/(thio)enol
PT_03_00	simple (aliphatic) imine
PT_04_00	special imine
PT_05_00	1,3 aromatic heteroatom H-shift
PT_06_00	1,3 heteroatom H-shift
PT_07_00	1,5 (aromatic) heteroatom H-shift (1)
PT_08_00	1,5 (aromatic) heteroatom H-shift (2)
PT_09_00	1,7 (aromatic) heteroatom H-shift
PT_10_00	1,9 (aromatic) heteroatom H-shift
PT_11_00	1,11 (aromatic) heteroatom H-shift
PT_12_00	1,3 furanones
PT_13_00	keten-inol exchange
PT_14_00	ionic nitro/aci-nitro
PT_15_00	pentavalent nitro/aci-nitro
PT_16_00	nitroso/oxime
PT_17_00	oxime/nitroso via phenol
PT_18_00	cyanic/iso-cyanic acids
PT_19_00	formamidinesulfonic acid
PT_20_00	isocyanide
PT_21_00	phosphonic acid
New rules (Prototropic tautomerism)	
PT_11_01	1,13 (aromatic) heteroatom H-shift
PT_11_02	1,15 (aromatic) heteroatom H-shift
PT_11_03	1,17 (aromatic) heteroatom H-shift
PT_11_04	1,19 (aromatic) heteroatom H-shift
PT_22_00	imine/imine
PT_23_00	1,5 furanones
PT_24_00	1,4 N-oxide/N-hydroxide
PT_25_00	1,6 N-oxide/N-hydroxide (1)
PT_26_00	1,6 N-oxide/N-hydroxide (2)
PT_27_00	acene
PT_27_01	thiophene analogue of acene
PT_28_00	nitro/aci-nitro via aromatic ring (1)
PT_29_00	nitro/aci-nitro via aromatic ring (2)
PT_29_01	o-tolualdehyde
PT_30_00	nitramide/N-nitronic acid
PT_31_00	sulfone-based aliphatic compounds
PT_32_00	nitrile/ketenimine 1,3 H-shift

Rule number	Rule name
PT_33_00	nitrile/ketenimine 1,5 H-shift
PT_34_00	triad phosphorus-carbon
PT_35_00	sulfenyl/sulfinyl
PT_36_00	oxime/nitrone
PT_37_00	sulfenyl/S-oxide
PT_38_00	silane-hemiaminal/silanoic amide
PT_39_00	nitrone/azoxy or Behrend rearrangement
PT_40_00	tetrad phosphorus-carbon
PT_41_00	pyridine 1-oxide/1-hydroxypyridine
PT_42_00	³ - / ⁴ -pyrro(thio/seleno)lin-2-one
PT_43_00	phthalan/isobenzofuran
PT_44_00	2-substituted-pyrrole
PT_45_00	isoindole/isoindolenine
PT_46_00	4-picoline
PT_47_00	isopropylidenecycloalkane/isopropylcycloalkene
PT_48_00	benzofuranone
PT_49_00	N-hydroxyindole
Existing rules ¹⁷ (Ring-chain tautomerism)	
RC_03_00	5_exo_trig
RC_04_01	6_exo_trig
RC_04_02	6_exo_trig
RC_09_00	5_endo_trig
RC_10_00	6_endo_trig
New rules (Ring-chain tautomerism)	
RC_03_03	boronic acid/oxaborole
RC_03_04	5_exo_trig
RC_04_04	6_exo_trig
RC_12_00	5_endo_tet or iminophosphorane/benzoxazaphospholine
RC_13_00	6_endo_dig
RC_14_00	thiadiazoline rearrangement
RC_15_00	5_exo_trig
RC_16_00	boryl/borate
RC_17_00	boryl/borate
RC_18_00	5_exo_tet or hydroxyphosphorane
RC_19_00	nitroolefin/1,2-oxazine N-oxide
RC_20_00	5_endo_trig
RC_21_00	cyclobutane/enamine
RC_22_00	5_endo_trig
RC_23_00	6_endo_trig
RC_24_00	λ^5 -/ λ^3 -phosphane
New rules (Valence tautomerism)	
VT_01_00	monothio-o-benzoquinone/benzoxathiete

Rule number	Rule name
VT_01_01	α -dithione and 1,2-dithiete
VT_02_00	tetrazole/azide
VT_03_00	isothiocyanate/triazinethione
VT_04_00	tetrazine/azodiazole
VT_05_00	1,2,3-triazole/diazoamidene
VT_06_00	norcaradiene/cycloheptatriene or benzene-oxide/oxepin
VT_07_00	phospha-münchnones
VT_08_00	1,2,3,4-tetrazinium/azodiazonium
VT_09_00	phosphinoimine/diazaphosphazole
VT_10_00	phosphine/phosphonium salt

Table 3.

Occurrence of Transforms in 400+ Million Small Molecules

Rule number	Occurrence ^a	Occurrence rate ^a (%)
PT_02_00	3,349,074	0.84
PT_03_00	56,988,782	14.21
PT_04_00	7,426,264	1.85
PT_05_00	34,290,440	8.55
PT_06_00	295,316,597	73.64
PT_07_00	31,141,877	7.77
PT_08_00	4,964,189	1.24
PT_09_00	146,537,974	36.54
PT_10_00	10,279,720	2.56
PT_11_00	3,149,927	0.79
PT_11_01	514,098	0.13
PT_11_02	244,544	0.06
PT_11_03	144,249	0.04
PT_11_04	45,213	0.01
PT_12_00	20,131,770	5.02
PT_13_00	27,983	0.01
PT_14_00	222,839	0.06
PT_15_00	227,189	0.06
PT_16_00	1,120,680	0.28
PT_17_00	6613	<0.01
PT_18_00	3975	<0.01
PT_19_00	4040	<0.01
PT_20_00	1733	<0.01
PT_21_00	176,295	0.04
PT_22_00	6,305,306	1.57
PT_23_00	7,410,570	1.85
PT_24_00	49,477	0.01
PT_25_00	5471	<0.01
PT_26_00	10,752	<0.01
PT_27_00	32,266	0.01
PT_27_01	99	<0.01
PT_28_00	1,291,000	0.32
PT_29_00	539,360	0.13
PT_29_01	49,814	0.01
PT_30_00	24,296	0.01
PT_31_00	363	<0.01
PT_32_00	542,830	0.14
PT_33_00	359,601	0.09
PT_34_00	1523	<0.01

Rule number	Occurrence ^a	Occurrence rate ^a (%)
PT_35_00	7568	<0.01
PT_36_00	721,416	0.18
PT_37_00	298	<0.01
PT_38_00	5	<0.01
PT_39_00	19,699	<0.01
PT_40_00	0	0.00
PT_41_00	53,562	0.01
PT_42_00	791,945	0.20
PT_43_00	6055	<0.01
PT_44_00	36,207	0.01
PT_45_00	65,414	0.02
PT_46_00	294	<0.01
PT_47_00	31,593	0.01
PT_48_00	2140	<0.01
PT_49_00	1175	<0.01
RC_03_00	62,261,031	15.53
RC_03_03	1650	<0.01
RC_03_04	112,442	0.03
RC_04_01	23,185,626	5.78
RC_04_02	21,983,105	5.48
RC_04_04	2303	<0.01
RC_09_00	1,389,202	0.35
RC_10_00	1,069,887	0.27
RC_12_00	104	<0.01
RC_13_00	250,829	0.06
RC_14_00	239	<0.01
RC_15_00	979	<0.01
RC_16_00	5	<0.01
RC_17_00	34	<0.01
RC_18_00	251	<0.01
RC_19_00	10,304	<0.01
RC_20_00	2464	<0.01
RC_21_00	10,353	<0.01
RC_22_00	2541	<0.01
RC_23_00	994	<0.01
RC_24_00	1077	<0.01
VT_01_00	2938	<0.01
VT_01_01	3726	<0.01
VT_02_00	2,881,841	0.72
VT_03_00	1347	<0.01
VT_04_00	7	<0.01
VT_05_00	1502	<0.01

Rule number	Occurrence ^a	Occurrence rate ^a (%)
VT_06_00	82,695	0.02
VT_07_00	4303	<0.01
VT_08_00	57	<0.01
VT_09_00	1	<0.01
VT_10_00	5	<0.01

^aOccurrence was analyzed across the nine databases listed in Table 5.

Table 4.

Current Naming of Guasch's Ring-chain Rules^a

Guasch's numbering	Rule name	Guasch's rule variant(s)	Numbering in current nomenclature	Used in this paper
RC1	3_exo_Trig	1	RC_01_00	–
RC2	4_exo_Trig	1	RC_02_00	–
RC3	5_exo_Trig	3	RC_03_00 to RC_03_02	RC_03_00
RC4	6_exo_Trig	4	RC_04_00 to RC_04_03	RC_04_01, RC_04_02
RC5	7_exo_Trig	5	RC_05_00 to RC_05_04	–
RC6	5_exo_Dig	3	RC_06_00 to RC_06_02	–
RC7	6_exo_Dig	4	RC_07_00 to RC_07_03	–
RC8	7_exo_Dig	5	RC_08_00 to RC_08_04	–
RC9	5_endo_Trig	3	RC_09_00 to RC_09_02	RC_09_00
RC10	6_endo_Trig	4	RC_10_00 to RC_10_03	RC_10_00
RC11	7_endo_Trig	5	RC_11_00 to RC_11_04	–

^aRule variants had been differentiated in the Guasch nomenclature by adding apostrophe(s) to the numbering (e.g., Guasch's rules RC4, RC4' and RC4'' correspond to RC_04_00, RC_04_01, and RC_04_02, respectively in the current nomenclature).

Table 5.

List of Databases Used for Transform Analyses

Name	Size (Compounds)	Accessibility	Reference
Drugs (DrugBank)	10,632	Public	19
PDB ligands	29,877	Public	20
CSD organics	319,204	Private	21
ChEMBL	1,820,035	Public	22
AMS screening samples	8,409,644	Public	23
SureChEMBL (Patents)	19,334,472	Public	24
PubChem	96,502,282	Public	25
ChemNav	131,901,120	Public	26
CSDB	142,706,819	Private	27

Table 6.
Observation of Standard and Nonstandard InChI Pass and Fail for Each Rule (PubChem)^a

Rule number	NonStdInChI				StdInChI	
	Partial InChI match	Complete InChI match	InChI fail	InChI success rate ^b (%)	InChI success rate ^b (%)	
PT_02_00	184,177	488,776	523,910	56.22	7.96	
PT_03_00	51,385	767,997	11,986,691	6.40	0.00	
PT_04_00	1029	209,430	1,647,601	11.33	0.00	
PT_05_00	1078	7,636,482	12,601	99.82	99.65	
PT_06_00	19,888,198	29,908,744	12,979,805	79.32	68.50	
PT_07_00	69,248	7,214,629	473,749	93.88	37.56	
PT_08_00	14,684	952,904	88,997	91.55	90.14	
PT_09_00	3,658,356	4,095,444	24,507,221	24.03	11.00	
PT_10_00	16,184	1,270,264	435,923	74.68	22.55	
PT_11_00	3559	204,614	328,702	38.76	33.07	
PT_11_01	661	62,012	107,340	36.85	4.26	
PT_11_02	766	12,005	69,209	15.57	13.96	
PT_11_03	877	6664	43,900	14.65	11.90	
PT_11_04	768	6719	9699	43.54	43.12	
PT_12_00	44,432	2,217,528	1,325,881	63.04	0.00	
PT_13_00	0	0	5701	0.00	0.00	
PT_14_00	0	0	88,485	0.00	0.00	
PT_15_00	0	0	88,503	0.00	0.00	
PT_16_00	76	22,247	367,745	5.72	0.00	
PT_17_00	1	321	1837	14.91	0.14	
PT_18_00	0	23	1849	1.23	1.12	
PT_19_00	0	5	1615	0.31	0.37	
PT_20_00	0	0	586	0.00	0.00	
PT_21_00	0	0	26,502	0.00	0.00	
PT_22_00	1189	224	2,992,839	0.05	0.04	
PT_23_00	347	23,263	1,225,994	1.89	0.00	
PT_24_00	0	0	15,746	0.00	0.00	

Rule number	NonStdInChI				StdInChI	
	Partial InChI match	Complete InChI match	InChI fail	InChI success rate ^b (%)	InChI success rate ^b (%)	InChI success rate ^b (%)
PT_25_00	0	0	2214	0.00	0.00	0.00
PT_26_00	0	0	4101	0.00	0.00	0.00
PT_27_00	0	0	14,785	0.00	0.00	0.00
PT_27_01	0	0	31	0.00	0.00	0.00
PT_28_00	0	0	305,195	0.00	0.00	0.00
PT_29_00	0	0	195,131	0.00	0.00	0.00
PT_29_01	19	108	24,802	0.51	0.00	0.00
PT_30_00	0	0	9586	0.00	0.00	0.00
PT_31_00	0	0	165	0.00	0.00	0.00
PT_32_00	0	0	61,800	0.00	0.00	0.00
PT_33_00	0	0	105,513	0.00	0.00	0.00
PT_34_00	0	1	717	0.14	0.00	0.00
PT_35_00	0	0	2,882	0.00	0.00	0.00
PT_36_00	0	0	361,348	0.00	0.00	0.00
PT_37_00	0	0	117	0.00	0.00	0.00
PT_38_00	0	0	5	0.00	0.00	0.00
PT_39_00	0	15	7524	0.20	0.01	0.01
PT_40_00 ^c	0	0	0	0	0.00	0.00
PT_41_00	0	0	20,966	0.00	0.00	0.00
PT_42_00	105	6,120	431,113	1.42	0.00	0.00
PT_43_00	0	0	3078	0.00	0.00	0.00
PT_44_00	0	165	9434	1.72	0.00	0.00
PT_45_00	0	0	28,726	0.00	0.00	0.00
PT_46_00	0	0	150	0.00	0.00	0.00
PT_47_00	0	0	12,360	0.00	0.00	0.00
PT_48_00	1	29	443	6.34	0.00	0.00
PT_49_00	0	0	447	0.00	0.00	0.00
RC_03_00	0	0	8,300,320	0.00	0.00	0.00
RC_03_03	0	0	632	0.00	0.00	0.00
RC_03_04	0	0	40,862	0.00	0.00	0.00

Rule number	NonStdInChI				StdInChI	
	Partial InChI match	Complete InChI match	InChI fail	InChI success rate ^b (%)	InChI success rate ^b (%)	
RC_04_01	0	0	4,028,848	0.00	0.00	
RC_04_02	0	0	3,666,752	0.00	0.00	
RC_04_04	0	0	752	0.00	0.00	
RC_09_00	0	0	274,785	0.00	0.00	
RC_10_00	0	0	203,731	0.00	0.00	
RC_12_00	0	0	31	0.00	0.00	
RC_13_00	0	0	55,989	0.00	0.00	
RC_14_00	0	2	106	1.85	1.85	
RC_15_00	0	0	529	0.00	0.00	
RC_16_00	0	0	3	0.00	0.00	
RC_17_00	0	0	10	0.00	0.00	
RC_18_00	0	0	83	0.00	0.00	
RC_19_00	0	0	5982	0.00	0.00	
RC_20_00	0	0	995	0.00	0.00	
RC_21_00	0	0	5950	0.00	0.00	
RC_22_00	0	0	960	0.00	0.00	
RC_23_00	0	0	482	0.00	0.00	
RC_24_00	0	0	335	0.00	0.00	
VT_01_00	0	0	869	0.00	0.00	
VT_01_01	0	0	1474	0.00	0.00	
VT_02_00	0	0	463,075	0.00	0.00	
VT_03_00	0	0	631	0.00	0.00	
VT_04_00	0	0	3	0.00	0.00	
VT_05_00	0	0	742	0.00	0.00	
VT_06_00	0	0	31,722	0.00	0.00	
VT_07_00	0	0	1769	0.00	0.00	
VT_08_00	0	0	40	0.00	0.00	
VT_09_00	0	0	1	0.00	0.00	
VT_10_00	0	0	2	0.00	0.00	
Overall^d				50.31	37.39	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

^b Partial and Complete InChI match columns are shown only for NonStdInchi. InChI success rate = ("Complete match" + "Partial match")/(Occurrence of rule).

^b The rules with InChI success rate of 0.00 (= 0/Occurrence of rule) indicate that the cases of InChI pass for them is 0.

^c No cases were found for rule PT_40_00, i.e., the InChI success rate of 0.00 is thus assigned to what would be strictly speaking the value 0/0.

^d Overall percentage calculated by summing up the numbers for all rules, not as average of the rate percentages.

Table 7. Standard and Nonstandard InChI Recapitulation across All Rules (InChI used: V.1.05)^a

Database	Complete pass		Complete pass for at least one rule and partial pass for other	Tautomeric molecules count	Overall InChI recapitulation ^b (%)	Overall strict InChI recapitulation ^c (%)
	For any applicable rule	Partial pass				
Drugbank	1,042	100	375	StdInChI 7427	62.11	14.03
	965	1431	700			
PDB ligands	3494	360	1354	22,939	69.83	15.23
	3402	4794	2615			
CSD organics	16,807	3379	2351	153,091	35.28	10.98
	16,469	11,127	3872			
ChEMBL	207,453	36,033	48,316	1,398,045	70.64	14.84
	304,087	246,541	145,095			
AMS	1,126,213	289,808	116,649	6,358,861	73.38	17.71
	1,657,392	1,030,261	445,996			
SureChEMBL	1,802,766	268,598	517,010	12,621,006	62.21	14.28
	1,949,348	2,006,240	1,307,812			
PubChem	10,516,304	1,417,527	1,580,535	67,262,970	66.36	15.63
	14,270,022	12,801,744	4,050,060			
ChemNav	17,418,383	4,447,222	1,500,175	105,565,942	80.30	16.50
	33,623,754	22,336,554	5,438,461			
CSDB	17,154,105	4,534,817	1,694,508	115,696,900	79.08	14.83
	36,928,720	23,633,538	7,547,799			
NonStdInChI						
Drugbank	2016	157	582	7427	81.88	27.14
	658	1909	759			
PDB ligands	5484	502	2169	22,939	83.47	23.91
	2305	5841	2847			
CSD organics	45,556	5690	7982	153,091	65.10	29.76
	12,143	20,702	7592			

Database	Complete pass		Partial pass		Tautomeric molecules count	Overall InChI recapitulation ^b (%)	Overall strict InChI recapitulation ^c (%)
	For any applicable rule	Complete pass for at least one rule and partial pass for other					
ChEMBL	330,685	43,749	98,588		1,398,045	83.44	23.65
	<i>219,892</i>	<i>299,824</i>	<i>173,848</i>				
AMS	1,534,982	306,656	307,735		6,358,861	81.57	24.14
	<i>1,263,143</i>	<i>1,126,724</i>	<i>647,711</i>				
SureChEMBL	2,917,438	366,712	866,922		12,621,006	75.69	23.12
	<i>1,419,390</i>	<i>2,512,248</i>	<i>1,470,143</i>				
PubChem	15,900,675	1,826,999	2,973,696		67,250,941	77.94	23.64
	<i>11,266,876</i>	<i>15,119,739</i>	<i>5,328,079</i>				
ChemNav	22,942,776	4,617,529	3,328,166		105,565,942	86.64	21.73
	<i>25,734,204</i>	<i>25,121,432</i>	<i>9,719,674</i>				
CSDB	23,447,796	4,921,978	3,883,624		115,679,596	86.76	20.27
	<i>28,119,041</i>	<i>27,699,416</i>	<i>12,295,971</i>				

^aThe first row of the three columns "Complete pass", "Partial pass", and "Complete pass for one rule and partial pass for other" for each database shown here contains numbers without failure by any other rule, whereas the second row for each database (in italics) shows the results for the cases with failures included. For more detailed explanation of these columns and failure-containing data added, please refer to the third spreadsheet in the SI.

^b"Overall InChI recapitulation" is the percentage of the sum of the six columns named "Complete pass", "Partial pass", and "Complete pass for one rule and partial pass for other" and three columns that failed relative to the tautomeric molecules of that database.

^c"Overall strict InChI recapitulation" is the percentage of molecules where input InChI matches with all enumerated tautomers generated by at least one rule (Complete pass) relative to tautomeric molecules of that database.

Table 8.

Overlap Matrix of Rules between PT_00_02 and PT_12_00^a

Rule	PT_02_00	PT_03_00	PT_04_00	PT_05_00	PT_06_00	PT_07_00	PT_08_00	PT_09_00	PT_10_00	PT_11_00	PT_11_01	PT_11_02	PT_11_03	PT_11_04	PT_12_00
PT_02_00	0	0	0	0	1343	657,926	0	1476	18,921	9	13,303	16	661	5	0
PT_02_00%	0	0	0	0	0.11	54.97	0	0.12	1.58	0	1.11	0	0.06	0	0
PT_03_00	0	0	1,735,457	0	6,182,856	0	0	176,504	0	0	0	0	0	0	437,568
PT_03_00%	0	0	13.55	0	48.28	0	0	1.38	0	0	0	0	0	0	3.42
PT_04_00	0	1,735,457	0	0	1,008,035	0	0	30,101	0	0	0	0	0	0	0
PT_04_00%	0	93.4	0	0	54.25	0	0	1.62	0	0	0	0	0	0	0
PT_05_00	0	0	0	0	7,609,348	256	4	208,141	10	2270	5	1907	0	2331	0
PT_05_00%	0	0	0	0	99.46	0	0	2.72	0	0.03	0	0.02	0	0.03	0
PT_06_00	1343	6,182,856	1,008,035	7,609,348	0	12,855	107	6,535,999	35	91,673	32	3456	19	3241	1,590,402
PT_06_00%	0	9.85	1.61	12.12	0	0.02	0	10.41	0	0.15	0	0.01	0	0.01	2.53
PT_07_00	657,926	0	0	256	12,855	0	1,052,699	3147	194,262	18,257	19,561	680	31,885	443	0
PT_07_00%	8.48	0	0	0	0.17	0	13.57	0.04	2.5	0.24	0.25	0.01	0.41	0.01	0
PT_08_00	0	0	0	4	107	1,052,699	0	463	66,218	17,977	1287	562	30,875	431	0
PT_08_00%	0	0	0	0	0.01	99.61	0	0.04	6.27	1.7	0.12	0.05	2.92	0.04	0
PT_09_00	1476	176,504	30,101	208,141	6,535,999	3,147	463	0	940	48,064	175	9663	225	4007	222
PT_09_00%	0	0.55	0.09	0.65	20.26	0.01	0	0	0	0.15	0	0.03	0	0.01	0
PT_10_00	18,921	0	0	10	35	194,262	66,218	940	0	559	44,081	10,553	4921	1350	0
PT_10_00%	1.1	0	0	0	0	11.28	3.84	0.05	0	0.03	2.56	0.61	0.29	0.08	0
PT_11_00	9	0	0	2270	91,673	18,257	17,977	48,064	559	0	706	13,320	29,694	3677	0
PT_11_00%	0	0	0	0.42	17.07	3.4	3.35	8.95	0.1	0	0.13	2.48	5.53	0.68	0
PT_11_01	13,303	0	0	5	32	19,561	1287	175	44,081	706	0	18,876	6996	1818	0
PT_11_01%	7.82	0	0	0	0.02	11.5	0.76	0.1	25.92	0.42	0	11.1	4.11	1.07	0
PT_11_02	16	0	0	1907	3456	680	562	9663	10,553	13,320	18,876	0	3228	8930	0
PT_11_02%	0.02	0	0	2.33	4.21	0.83	0.69	11.78	12.87	16.24	23.02	0	3.94	10.89	0
PT_11_03	661	0	0	0	19	31,885	30,875	225	4921	29,694	6996	3228	0	3387	0
PT_11_03%	1.28	0	0	0	0.04	61.95	59.98	0.44	9.56	57.69	13.59	6.27	0	6.58	0
PT_11_04	5	0	0	2331	3241	443	431	4007	1350	3677	1818	8930	3387	0	0
PT_11_04%	0.03	0	0	13.56	18.85	2.58	2.51	23.3	7.85	21.39	10.57	51.94	19.7	0	0

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Rule	PT_02_00	PT_03_00	PT_04_00	PT_05_00	PT_06_00	PT_07_00	PT_08_00	PT_09_00	PT_10_00	PT_11_00	PT_11_01	PT_11_02	PT_11_03	PT_11_04	PT_12_00
PT_12_00	0	437,568	0	0	1,590,402	0	0	222	0	0	0	0	0	0	0
PT_12_00%	0	12.19	0	0	44.32	0	0	0.01	0	0	0	0	0	0	0

^aCommon CACTVS rules plus variants, for PubChem only.