

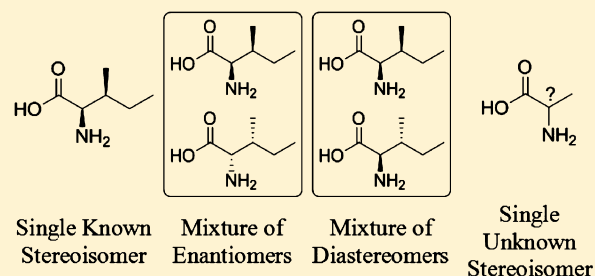
# Handling of Tautomerism and Stereochemistry in Compound Registration

Alberto Gobbi\* and Man-Ling Lee\*

Discovery Chemistry, Genentech Incorporated, 1 DNA Way, South San Francisco, California 94080, United States

## S Supporting Information

**ABSTRACT:** Automated registration of compounds from external sources is necessitated by the numerous compound acquisitions from vendors and by the increasing number of collaborations with external partners. A prerequisite for automating compound registration is a robust module for determining the structural novelty of the input structures. Any such tool needs to be able to take uncertainty about stereochemistry into account and to identify tautomeric forms of the same compound. It also needs to validate structures for potential mistakes in connectivity and stereochemistry. Genentech has implemented a Structure Normalization Module based on toolkits offered by OpenEye Scientific Software. The module is incorporated in a graphical application for single compound registration and in scripts for bulk registration. It is also used for checking compounds submitted by our collaborators via partner-specific Internet sites. The Genentech Structure Normalization Module employs the widely used V2000 molfile format to accommodate structures received from a wide variety of sources. To determine how much information is known about the stereochemistry of each compound, the module requires a separate stereochemical assignment. A structural uniqueness check is performed by comparing the canonical SMILES of a standard tautomer. This paper offers a discussion of the steps taken to validate the chemical structure and generate the canonical SMILES of the standard tautomer. It also describes the integration of the validation module in compound registration pathways.



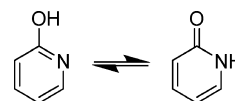
## INTRODUCTION

Pharmaceutical companies regularly purchase compound libraries from vendors, outsource chemical synthesis, and collaborate with Contract Research Organizations (CROs). The increased reliance on chemistry outsourcing<sup>1</sup> requires a robust and automated compound registration system to ensure the accurate and timely registration of compounds from external sources. A structure normalization process that handles tautomerism and stereochemistry is essential for validating structures and automating compound registration.

The principal task of a compound registration system is to correctly detect identical chemical structures as the same. This is a prerequisite for systematically organizing compounds. The uniqueness check can be done by matching new compounds with each compound in the database using slow graph matching algorithms.<sup>2–4</sup> Alternatively, algorithms have been developed to generate unique<sup>5–8</sup> or hashed<sup>9</sup> identifiers from chemical structures. The identifier of new compounds can be quickly compared to the precomputed identifiers in the database. For hashed identifiers, compounds with matching identifiers are compared to the input via a graph matching algorithm in a secondary step. The identifier based comparison has superseded the slower graph-based matching algorithms. Before generating the identifier, generally additional transformations are performed to normalize charges, remove counterions, and standardize the valence bond description of chemical groups, such as nitro groups. The comparison of chemical structures is

complicated by uncertainties concerning stereochemistry and by tautomeric forms. To handle these complications human intervention is often needed.

Tautomerism<sup>10,11</sup> arises when a compound can rapidly interconvert between chemical structures that differ in the location of one or more atoms (Figure 1). This means that a



**Figure 1.** Interconverting tautomers of pyridone due to the migration of a polar hydrogen atom.

registered compound could, in reality, be a mixture of tautomers. The ratio of the tautomers and rate of interconversion depend on the energy differences of the tautomers, their transition states, and on external factors, such as temperature, solvent, and pH. In principal such a mixture requires multiple structures to represent its composition.

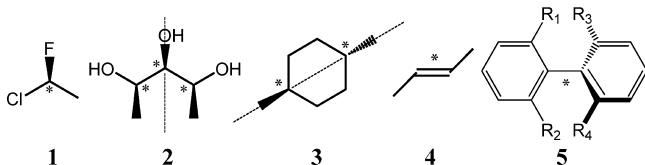
In contrast, stereoisomerism<sup>12,13</sup> arises when chemical structures have the same atomic connectivity with a different

**Special Issue:** 2011 Noordwijkerhout Cheminformatics

**Received:** July 16, 2011

**Published:** December 8, 2011

spatial arrangement. This is most frequently due to a different configuration of substituents around stereogenic centers. Figure 2 gives examples of molecules containing stereogenic centers



**Figure 2.** Molecules with stereogenic centers common to pharmaceutical small molecules. Symmetry planes relevant to the discussion are indicated by the dotted lines.

common to pharmaceutical compounds. Stereogenic centers are marked with stars. Molecule 1 is chiral because it cannot be superimposed with its mirror image. This is caused by the marked carbon atom with four different substituents.

The three marked carbon atoms in molecule 2 all have four different substituents. The carbon atom in the middle, however, has two substituents that are mirror images of each other. Atoms, such as the marked carbon atom in the middle of 2, have been termed pseudo-chiral.<sup>14,15</sup> The plane of symmetry going through the pseudo-chiral atom causes the molecule to be achiral. In the subsequent discussion, we refer to all stereogenic centers with four different substituents as asymmetric stereogenic centers.

Nonchiral substituted ring systems, such as 1,4 disubstituted cyclohexane 3, can have stereogenic centers with two identical substituents. This type of stereogenic center has been termed pseudo-chiral in some publications. However, newer publications suggest that the term pseudo-chiral should be reserved to acyclic systems and atoms with four different substituents.<sup>16</sup> In this paper, we refer to the stereogenic centers in molecule 3 as symmetric stereogenic centers. Additionally, bonds can be stereogenic centers too. Double bonds can give rise to cis and trans isomers, such as in molecule 4. The rotational barrier around a single bond in atropisomers, such as 5 in Figure 2, can yield chiral molecules in which the single bond is a stereogenic center.

Depending on the synthesis, a sample can consist of a single stereoisomer whose absolute stereochemistry is either known or unknown. In other cases, a sample can be a mixture of enantiomers or stereoisomers. Any automated registration system must discern and capture the stereochemistry from the user input and distinguish stereoisomers.

**Chemical Structure Representations.** A number of computer readable chemical structure representations have been published and are in use today. Blackwood et al. published a structure representation that is used in the Chemical Abstracts Service (CAS) registry file.<sup>17,18</sup> To describe interconverting tautomers, hydrogen atoms are assigned to the group of heavy atoms that form bonds to the given hydrogen atoms in any tautomeric form. To describe stereochemistry the CAS registry file combines stereogenic centers into groups with known absolute, known relative, or unknown stereochemistry. Groups with relative stereochemistry are designated as a single form or as mixtures of the two mirror images.

The International Union of Pure and Applied Chemistry (IUPAC) has recently endorsed the International Chemical Identifier (InChI) as a new format for structure representation.<sup>8,19,20</sup> The InChI identifier follows an approach similar to the one described by Blackwood et al.

In the mid-1990's, MDL Information Systems published the molfile V3000 structure representation format.<sup>21</sup> The V3000 format allows the combination of multiple stereogenic centers into groups of different types, similar to the CAS registry file format. Any degree of known absolute or relative stereochemistry can be stored in the molfile V3000 format. The identification and representation of tautomeric forms, however, are not explicitly handled in the V3000 format.

Despite the capability of capturing stereoisomerism and tautomerism, none of the above structure representation formats (CAS registry file, InChI, or molfile V3000) has become a widely used standard for data exchange in the pharmaceutical industry. The common exchange format for compound registration remains the molfile V2000 format.<sup>21</sup> To a lesser extent, delimited files containing structures in the SMILES line notation<sup>7,22</sup> are used. Because the V2000 molfile format is provided by all our compound sources, we have elected to use the molfile V2000 format as the standard input format for the compound registration system at Genentech.

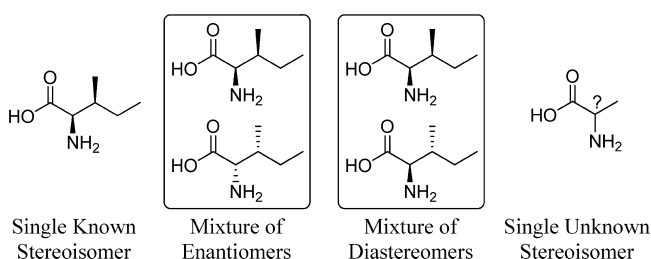
Neither the molfile V2000 nor SMILES format can represent interconverting tautomeric forms or ambiguities about stereoisomers in a "single" structure representation. We use a separate stereochemistry flag to augment the structural information contained within the molfile V2000. This flag is assigned by the chemist and allows the differentiation of compounds with and without uncertainties in stereochemistry. The flag is stored in a separate database field. For the novelty check of input structures, a standardized tautomer is generated, and its canonical isomeric SMILES is stored in the database. We refer to this SMILES as the CTISMILES. The CTISMILES is used along with the stereochemistry flag for the uniqueness check.

**Recognizing Tautomers.** The handling of tautomers in structure-based design and chemical information has to satisfy different requirements. In structure-based drug design, the accuracy of a model frequently depends on the tautomer used.<sup>11,23–25</sup> If the relevant tautomer is not known, it is possible to use multiple forms and assume that the tautomer making the best ligand–protein interaction is the most relevant one. In chemical information the requirement is to be able to uniquely identify an input structure independently from the input tautomer. Warr has published a comprehensive overview of the state of the art in tautomer handling in chemical information systems.<sup>26</sup> The uniqueness check can be accomplished by storing all tautomers, storing a tautomer independent description, such as the InChI code, or by extending the concept of a canonical structure identifier to the generation of a standardized tautomer. This paper follows the third approach by using the CTISMILES for identifying compounds.

A necessary part of such an algorithm is the enumeration of the different tautomeric forms. As reviewed in Warr's publication,<sup>26</sup> several programs are available for enumerating tautomers. Some vendors provide standalone programs, *sdwash*<sup>27</sup> from the Chemical Computing Group and *MN.TAUTOMER*<sup>28</sup> from Molecular Networks, use a collection of tautomer transformation rules derived from Oellien et al.<sup>24</sup> for the enumeration of tautomer structures. Other vendors, such as ChemAxon<sup>29</sup> and OpenEye,<sup>30,31</sup> provide toolkits, which allow the enumeration of tautomers in a canonical order. At Genentech we have decided to use OpenEye's QuacPac toolkit.<sup>32,33</sup> The first tautomer generated by QuacPac is used as the standard tautomer in our registration system.

**Stereochemistry Flag.** The V2000 molfile format contains a binary chiral flag used to specify if drawn stereogenic

centers are known absolutely or only relative to each other. The flag has only two values and can thus not be used to distinguish stereogenic centers with unknown stereochemistry from such which are mixtures of the two conformations. To distinguish between various levels of known and unknown information on stereogenic centers using the V2000 molfile format, some pharmaceutical companies have augmented the molfile with stereochemistry flags containing values from a fixed vocabulary. The number of expressions can be large depending on the level of distinction.<sup>34</sup> In many cases the flags are not mutually exclusive and require complex rules that define the priority of each flag. The Normalization Module described in this paper uses a set of six mutually exclusive stereochemistry flags: “no stereo”, “single known stereoisomer”, “single unknown stereoisomer”, “mixture of enantiomers”, “mixture of diastereomers”, and “uncertain structure” (Figure 3).



**Figure 3.** Depending on the knowledge about the stereochemistry, a compound is assigned one of the flags shown above. A “no stereo” flag is used for compounds without stereogenic centers, and an “uncertain structure” flag is used for special cases, such as mixtures of regioisomers.

Assignment of the correct stereo flag is obvious for chemists, given their knowledge of the starting materials, route of synthesis, and purification method. Since they are well-defined and mutually exclusive, these flags permit significant automated checking. For structure records from external sources that do not provide the stereochemistry flag, an automatic assignment is made based on the customary practice of expressing stereochemistry in the molfile V2000 format. For partners who employ their own stereochemistry vocabulary, rules have been implemented translating their expressions to the appropriate Genentech stereochemistry flag.

The generation of the CTISMILES and the ability to validate stereochemistry information allow Genentech to perform batch registration of compounds from collaborating partners, CROs, and commercially acquired compounds. Since the collaborating partners and the CROs represent a significant portion of the chemistry resources of Genentech’s Small Molecule Drug Discovery, a natural next step was to integrate the batch compound registration into an automation framework. This automation has enabled our organization to absorb an increasing number of incoming compound records without delaying compound registration and, thus, the progress of projects.

## METHOD

The core Normalization Module at Genentech is implemented as an application programming interface (API). The principal task of this module is to generate a single component CTISMILES and the stereochemistry flag from the input structure information. The input is validated, and inconsistent structures are rejected. As normalization and validation rules are likely to improve over

time, it was important to separate the normalization steps into clearly defined submodules. The API is implemented in Java using the OEChem and QuacPac toolkits.

The rules are encoded in a xml file<sup>35</sup> (Figure 4). Each rule type is implemented as a Java class. Some rules are as simple as

```
<StructureRules>

<atomLabelCheck> <description>
  Generate error on invalid atoms.
</description>
</atomLabelCheck>

<wigglyBondCheck> <description>
  Reject wiggly bonds</description>
</wigglyBondCheck>

<checkChiral> <description>
  Inconsistent wedges</description>
</checkChiral>
...
<transform id='nitro'>
  [*:4]-[N:1](=[O:2])=[O:3]>[*:4]-[N+:1](-[O:-2])=[O:3]
  <description>
    Normalize Nitro Group.</description>
</transform>
...
<componentNormalizer> <description>
  Remove counter ions and solvents.
  Reject multi components.</description>
  <solvent smiles='O' name='Hydrate'/>
</componentNormalizer>
...
<valenceCheck> <description>
  Validate valences on listed atoms.
  </description>
  <acceptableFragments> <!-- exceptions -->
    <fragment name='Nitro'>
      [N+v4](=[Ov2])-[O-v1]</fragment>
    ...
  </acceptableFragments>
  <atom symbol='N'>
    <valence charge='0' values='3'>
    <valence charge='1' values='4'>
  </atom>
  ...
</valenceCheck>

<checkStructFlag/>
...
<tautomerize> <description>
  Generate CTISMILES.</description>
</tautomerize>

</StructureRules>
```

**Figure 4.** Configuration file for the Normalization Module. Each check is described by a separate xml element with check-specific attributes and subelements. Some elements have been omitted in this figure for clarity. The full file can be found in the Supporting Information.

checking for unknown atom labels or unsupported bond types (e.g. wiggly bonds), others are more complex. The more complex rule types are discussed below in order of their execution. The final rule is the generation of the CTISMILES.

**Transformations of Specific Functional Groups.** Genentech scientists have agreed to several conventions for the consistent representation of certain classes of compounds, for example:

- Charged species are neutralized by adding or removing protons whenever possible.
- Bonds to first and second main group metals are separated, and the metals are charged.
- Nitro groups are represented with charged nitrogen and oxygen atoms.
- Sulfoxides are represented with no charge separation.

Table 1. Allowed Counts of Stereogenic Centers for Given Stereochemistry Flags

stereochemistry flag	asymmetric stereogenic center		symmetric stereogenic centers		stereogenic double bonds	
	$A_{\text{Spec}}$	$A_{\text{USpec}}$	$S_{\text{Spec}}$	$S_{\text{USpec}}$	$D_{\text{Spec}}$	$D_{\text{USpec}}$
no stereo	0	0	0	0	0	0
single known stereoisomer <sup>a</sup>	$k$	0	$m$	0	$n$	0
mixture of enantiomers	$A_{\text{Spec}} + A_{\text{USpec}} > 0$		$m$	0	$n$	0
mixture of diastereomers <sup>b</sup>	$k = A_{\text{Spec}} + A_{\text{USpec}}$		$m = S_{\text{Spec}} + S_{\text{USpec}}$		$n = D_{\text{Spec}} + D_{\text{USpec}}$	
single unknown stereoisomer <sup>a</sup>	$k = A_{\text{Spec}} + A_{\text{USpec}}$		$m = S_{\text{Spec}} + S_{\text{USpec}}$		$n = D_{\text{Spec}} + D_{\text{USpec}}$	

$A_{\text{Spec}}$ ,  $A_{\text{USpec}}$ : Number of specified and unspecified asymmetric stereogenic center.  $S_{\text{Spec}}$ ,  $S_{\text{USpec}}$ : Number of specified and unspecified symmetric stereogenic center.  $D_{\text{Spec}}$ ,  $D_{\text{USpec}}$ : Number of specified and unspecified stereogenic double bonds.  $k$ ,  $m$ , and  $n$  are integers  $\geq 0$ . Additional conditions apply for some stereochemistry flags. <sup>a</sup> $k + m + n > 0$ . <sup>b</sup> $k > 1$  or  $m + n > 0$ .

Each transformation is encoded as a SMIRKS string.<sup>36</sup> The transformations are applied to every input molecule. A complete list of transformations is included in the Supporting Information.

**Solvent and Counterion Removal.** Counterions and solvents are not considered in the uniqueness check and are removed as follows: A depth first search starting with the first atom in the input identifies the first component as the collection of atoms and bonds directly or indirectly connected to this first atom. After removing this component, the algorithm is repeated to perceive additional components. The canonical SMILES of each component is compared to a list of predefined counterions and solvents. Currently, 325 counterions and 36 solvents are recognized. After removing counterions and solvents, compounds that still consist of multiple components are rejected except in special situations, such as the registration of explicit mixtures for fragment screening. New salts or solvents are easily added to the database or configuration file should a compound be rejected.

**Valence Check.** To recognize drawing errors, the atomic valence of elements common to organic compounds is verified against a table of allowed charges and valences. For example, carbon atoms must be tetravalent and may not be charged. This type of validation is performed for the following atoms: H, Li, Na, K, Mg, Ca, B, C, Si, N, P, O, S, F, Cl, Br, and I. Fewer than 1 in 10 000 compounds in the Genentech database contain atoms which are not in this list.

**Stereochemistry Check or Assignment.** As described in the Introduction, the complete compound structure description at Genentech includes a stereochemistry flag. Chemists from Genentech and from CROs are required to provide a stereochemistry flag for compound registration. The stereochemistry of atoms in the molfile and the stereochemistry flag are both used for validation. For example, a compound flagged as “no stereo” may not have any stereogenic centers, while a compound with the “single known stereoisomer” flag must have explicit stereo information for all stereogenic centers.

The structure validation module recognizes symmetric and asymmetric stereogenic centers as well as stereogenic double bonds (cf. Figure 2). The OpenEye toolkit recognizes asymmetric stereogenic centers and stereogenic double bonds, but it does not recognize symmetric stereogenic center. Therefore we have implemented the following algorithm to recognize symmetric stereogenic centers:







- 1 A candidate atom is in a ring.
- 2 It is not an asymmetric stereogenic center.
- 3 It has three neighbors if it is a sulfur atom or it has four neighbors if it is a carbon or phosphorus atom.

4 It has two identical neighbors in a ring and two additional nonidentical neighbors (molecule 3, Figure 2).<sup>37</sup> This is determined by comparing the ring membership and the symmetry class of each neighbor of a candidate.

5 It is classified as a symmetric stereogenic center if there are one or more additional atoms fulfilling rules 1–4 in the same ring system.

The count of specified and unspecified stereogenic centers for each of these three types (symmetric, asymmetric, and double bonds) determines which of the stereochemistry flags are valid for the molecule. Table 1 shows the allowed counts for each stereochemistry flag. Inconsistencies between the entered stereochemistry flag and the counts as given in Table 1 result in a rejection of the registration. Table 2 shows some example

Table 2. Examples of Compounds with Counts of Stereogenic Centers<sup>a</sup>

Input	$A_{\text{Spec}}$	$A_{\text{USpec}}$	$S_{\text{Spec}}$	$S_{\text{USpec}}$	$D_{\text{Spec}}$	$D_{\text{USpec}}$	Consistent Flags	Assigned Flag
	1	0	0	0	0	0	SKS ME SUS	SKS
	0	1	0	0	0	0	ME SUS	ME
	1	1	0	0	0	0	ME MD SUS	MD
	0	0	0	0	1	0	SKS MD SUS	SKS
	1	0	2	0	0	0	SKS ME MD SUS	SKS
	0	1	0	2	0	0	MD SUS	MD

<sup>a</sup>Stereochemical flags consistent with the input and assigned for catalog compounds are given in the last two columns. Stereochemistry flags: SKS, single known stereoisomer; ME, mixture of enantiomers; MD, mixture of diastereomers; and SUS, single unknown stereoisomer. Other abbreviations are identical to Table 1.

input structures and the stereochemistry flags consistent with these inputs.

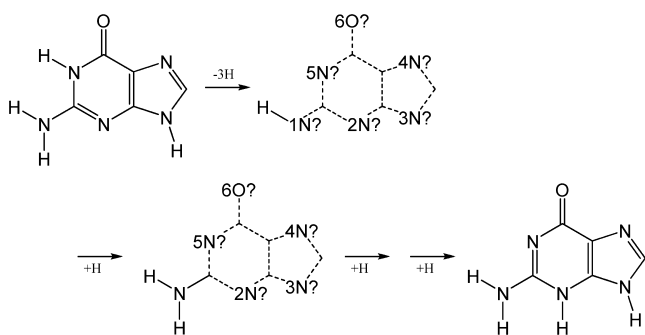
Compounds ordered from vendor catalogs have to be handled differently because vendors only provide molfiles without additional stereochemistry information. In this case, the stereochemistry flag is deduced from the molfile based on the customary practice of encoding the stereochemistry information in molfiles. The Normalization Module assumes that the stereochemistry is known absolutely if the stereo specification on a center in the molfile is present. Stereogenic centers with



unspecified stereochemistry are assumed to be mixtures of the two forms. The assignment of a compound as mixture of enantiomers or mixture of diastereomers is based on the count of the different types of stereogenic centers. Compounds with a single unspecified asymmetric stereogenic center having no unspecified double bonds or symmetric stereogenic centers are recognized as mixtures of enantiomers. All other inputs with unspecified stereogenic centers are recognized as mixtures of diastereomers. The assigned stereochemistry flag for example catalog inputs is given in the last column of Table 2. The underlying assumption may yield some incorrect assignments due to lack of information. For example, a compound which is a single unknown stereoisomer is incorrectly classified as a mixture of diastereomers. In general, this has not caused problems because positive results from biological experiments with catalog compounds are always repeated with resynthesized compounds. Thus, project-relevant decisions are based on compounds with full stereochemical information.

In compounds that are not flagged as single known stereoisomers, the stereochemistry information is removed from the SMILES after validating the flag. As the stereochemistry is not fully known for these compounds, the uniqueness check is performed using the SMILES string without stereochemistry information. If there are multiple registered compounds with SMILES and stereochemistry flag identical to those of the to-be registered compound, the registration module will issue a request for human inspection.

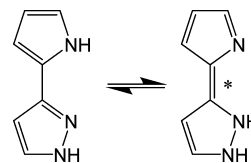
**Standard Tautomer Generation.** The OpenEye QuacPac toolkit generates all possible tautomers in a canonical order. QuacPac was configured to enumerate only those tautomers that involve the transfer of a hydrogen atom to and from heteroatoms. The rationale for applying this restriction is that tautomers involving hydrogen atoms on carbon are frequently stable at room temperature and do not interconvert. For cases with well documented equilibrium, such as the keto–enol tautomerism, specific transformations are performed using SMIRKS. The QuacPac algorithm<sup>32</sup> first removes hydrogen atoms from heteroatoms and unassigns the bond type to conjugated atoms (Figure 5, step 1). The first hydrogen atom is then added back to



**Figure 5.** QuacPac algorithm. Dotted lines represent bonds with unassigned type. Question marks are placed on atoms with unassigned hydrogen count. Numbers on atoms represent the rank order generated by the Morgan algorithm.

the heteroatom with the lowest rank computed by an algorithm based on the publication by Morgan.<sup>5</sup> The first addition in Figure 5 determines the bond order from 1N to the carbon atom. These steps are repeated for the other hydrogen atoms. If a structure with invalid atomic valence is created by adding a hydrogen atom, the hydrogen is removed and added to the next higher ranking heteroatom. In most cases, the first tautomer

generated by QuacPac is used as the standard tautomer. If the tautomerization algorithm creates, destroys, or shifts stereogenic centers (Figure 6), the tautomer is ignored, and the next tautomer

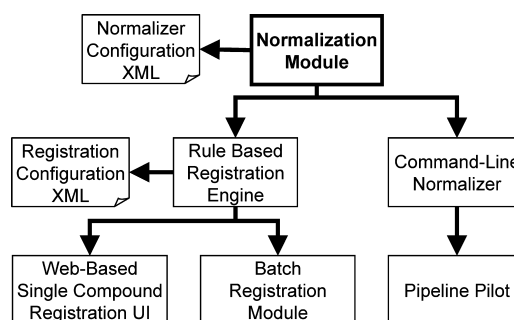


**Figure 6.** Example of tautomerism generating a stereogenic double bond between the two rings.

in the sequence is evaluated. This ensures that no information with respect to stereochemistry is lost in the selection procedure for the standard tautomer. Finally the canonical isomeric SMILES of this standard tautomer (CTISIMILES) is generated. This is used in combination with the stereochemistry flag for all novelty checks at Genentech.

## APPLICATIONS

The Normalization Module is used in the single compound registration system and in batch registration modules. It is also wrapped as a Pipeline Pilot component that is used by computational chemists for comparing novel ideas to compounds in our structure database (Figure 7). The rigorous structure

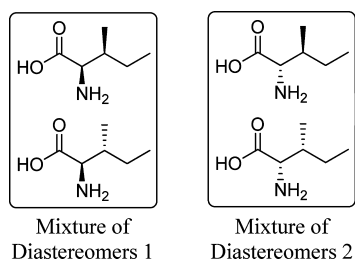


**Figure 7.** The Normalization Module is used whenever a novelty check needs to be performed on incoming structures. This can be in the Web-based single compound registration system, the batch registration module, or Pipeline Pilot.

checking and the normalization procedures enable automated structure registration pathways for combinatorial libraries, catalog compounds, and compounds from our external partners.

**Compound Registration.** Whenever a new compound is synthesized at Genentech or acquired from an external source, it is registered into the small molecule database as a new compound lot. This compound lot is assigned an existing parent substance based on its CTISIMILES and stereochemistry flag. A new parent substance record is created if no matching parent substance is found. The original input molfile is stored with the lot level information. The CTISIMILES is stored with the parent substance information.

For compounds that have no stereochemistry or are single known stereoisomers, the CTISIMILES uniquely identifies the parent substance. For the other stereochemical categories, the CTISIMILES does not contain any stereo information. For example, the two mixtures of diastereomers given in Figure 8 have the same CTISIMILES. In the single compound registration user interface, the registrar is asked to resolve any ambiguity by picking a specific parent substance after reviewing the



**Figure 8.** Multiple mixtures of diastereomers can exist with the same connectivity.

stereochemistry of existing compounds with the same structure and stereochemistry flag. Registration comments from previous lots are also shown. For compound lots added to the database as part of a batch registration process, a temporary assignment is made, and a request for a review by a project member is issued.

The single compound and batch registration applications use a common, rule-based, Registration Engine (Figure 7). The Registration Engine checks associated data, such as project assignment and lab journal reference, and uses the Normalization Module to validate the input structure and stereochemistry flag. If no errors are found, it inserts the information into the database.

This system has been in operation for two years at Genentech. During this period, thousands of compounds have been registered using the single compound user interface. Large compound libraries were registered using the batch registration. In one batch registration operation, over 1 000 000 compounds were loaded into the database in 20 h. After upgrading the hardware of the database server, registering a screening collection with similar size was accomplished within 4 h.

**Automation.** The infrastructure for automation consists of drop and archive directories and a script for file processing. The registrar can deposit any number of files into the drop directory. The processing script automatically scans the drop directory in regular intervals. It is executed using the Unix cron mechanism. If input files are found, the script validates the structures, registers the compounds, and moves the processed files along with the corresponding output and log files to the archive directory. Upon completion of the registration procedure a report is sent by e-mail to the Registrar and the Compound Management Group.

To further streamline the registration process, we have worked closely with the Genentech Information Technology department to set up secure partner-specific Web sites. These Web sites allow partner companies to upload files for registration. Uploaded files are immediately validated using the in-house Registration Engine. If the files do not comply with the agreed format or contain invalid structures, errors are reported back to the partner immediately.

## SCOPE AND LIMITATIONS

The Normalization module described in this paper was developed to handle the automated registration of small organic molecules of pharmaceutical relevance. This results in the following restrictions:

- Mixtures of multiple compounds are rejected in the default configuration of the Normalization Module. In pharmaceutical research, experiments are usually done using single compounds. In some circumstances, experiments are done with mixtures of compounds. One example is crystallographic fragment based screening.<sup>38</sup> To be able to handle compound mixtures, the batch registration application was extended to allow mixtures of

compounds. All normalization rules except the enforcement of a single component are performed, and the dot disconnected CTISMILES is used as structure identifier. This relaxed validation is only applied when registering inputs that are known to be mixtures.

- The Normalization Module does not attempt to distinguish different formulations of an active ingredient. On the contrary, because counterions and solvents are removed, different formulations of the same compound are recognized as the same structure. Currently formulations are tracked in a database maintained by the Formulation Group. A future extension of the registration system might incorporate formulation information by storing formulations as mixtures of components while tracking the component ratio.
- Inorganic and organometallic compounds are not handled in the current Normalization Module. The separation of counterions and solvents is not a good choice for inorganic and organometallic compounds. Stereochemistry on pentavalent and hexavalent stereogenic centers is not supported. Additionally, there are many valence bond representations for metalorganic complexes, which would result in different CTISMILES. This has not been a problem for the small-molecule database at Genentech, as inorganic and organometallic compounds are currently not of interest to the drug discovery projects.
- Axial chirality is not supported in SMILES strings, and therefore, stereochemistry in allenes and atropisomers is not handled by the current Normalization Module. While allenes are reactive and not of high importance for drug discovery, atropisomers are being studied as drugs.<sup>39,40</sup> The two enantiomers of atropisomers might or might not be separable at room temperature depending on the size, type, and orientation of the substituents. While the energy barrier is well studied in biphenyls with simple substituents, a prediction for more complex compounds requires force field or quantum mechanical calculations. This issue needs to be addressed in a future version of the Normalization Module.

## SUMMARY

At Genentech, we have chosen to use the molfile V2000 structure file format combined with a stereochemistry flag to represent small molecule chemical structures because the V2000 molfile format remains the most widely used format for exchanging structural information. Six mutually exclusive stereochemistry flags allow chemists to precisely convey the level of certainty about the stereochemistry of the structures. The paper also describes the generation and use of a standard tautomer for uniquely identifying compounds regardless of the input tautomer. Input structures are validated and normalized using a configurable Normalization Module. Strict application of the validation and normalization rules has enabled the automated registration of compounds from external sources, allowing Genentech to streamline its compound flow and reduce project cycle times.

## ASSOCIATED CONTENT

### Supporting Information

The full configuration file for the Normalization Module. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## ■ AUTHOR INFORMATION

## Corresponding Author

\*E-mail: gobbilberto@gene.com; lee.man-ling@gene.com.

## Author Contributions

Both authors contributed equally.

## ■ ACKNOWLEDGMENTS

We would like to thank J. Seerveld, D. F. Ortwine, K. P. Clark, and J. J. Crawford for reviewing the manuscript and the Cheminformatics group at Genentech for many discussions. Specifically, we would like to thank J. Blaney, C. Goliva, C. Lu, B. Sellers, N. Skelton, and H. Zheng. The validation rules were reviewed with in-house chemists F. Cohen, J. Dotson, L. Gazzard, R. Mendonca, and J. Rudolph. Implementing the partner-specific Web sites would not have been possible without the full support of our partners and the Genentech Information Technology department. We also thank K. Boda, R. Sayle, and B. Tolbert from OpenEye Scientific Software for supporting us in integrating the OpenEye toolkits.

## ■ REFERENCES

- (1) Ainsworth, S. Managing Outsourcing. *Chem. Eng. News* **2011**, 89, 48–51.
- (2) Figueras, J. Automorphism and equivalence classes. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 153–157.
- (3) Razinger, M.; Balasubramanian, K.; Munk, M. E. Graph automorphism perception algorithms in computer-enhanced structure elucidation. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 197–201.
- (4) Faulon, J.-L. Isomorphism, Automorphism Partitioning, and Canonical Labeling Can Be Solved in Polynomial-Time for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 432–444.
- (5) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, 5, 107–113.
- (6) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Chem. Soc.* **1974**, 96, 4834–4842.
- (7) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, 29, 97–101.
- (8) Stein, S. E.; Heller, S. R.; Tchekhovskoi, D. An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier. In Proceedings of the 2003 International Chemical Information Conference, Nimes, France, October 19–22, 2003; Collier, H., Ed.; Infonortics: Malmesbury, U.K., 2003; pp 131–143.
- (9) Wipke, W. T.; Krishnan, S.; Ouchi, G. I. Hash Functions for Rapid Storage and Retrieval of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1978**, 18, 32–37.
- (10) IUPAC Gold Book – Tautomerism; IUPAC: Research Triangle Park, NC; <http://goldbook.iupac.org/T06252.html> (accessed October 19, 2011).
- (11) Martin, Y. C. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, 23, 693–704.
- (12) IUPAC Gold Book – Stereoisomerism; IUPAC: Research Triangle Park, NC; <http://goldbook.iupac.org/S05983.html> (accessed October 19, 2011).
- (13) Eliel, E. L.; Wilen, S. H. Stereoisomers. *Stereochemistry of Organic Compounds*, John Wiley & Sons: New York, 1994; pp 49–70.
- (14) Nourse, J. G. Pseudochirality. *J. Am. Chem. Soc.* **1975**, 97, 4594–4601.
- (15) IUPAC Gold Book – pseudo-asymmetric carbon atom; IUPAC: Research Triangle Park, NC; <http://goldbook.iupac.org/P04921.html> (accessed October 19, 2011).
- (16) Chandrasekhar, S. Pseudoasymmetry: A final twist? *Chirality* **2008**, 20, 771–774.
- (17) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1977**, 17, 3–8.
- (18) Blackwood, J. E.; Blower, P. E.; Layten, S. W.; Lillie, D. H.; Lipkus, A. H.; Peer, J. P.; Qian, C.; Staggenborg, L. M.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 13. Enhanced handling of stereochemistry. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 204–212.
- (19) Stein, S.; Heller, S.; Tchekhovskoi, D. *The IUPAC Chemical Identifier – Technical Manual*, version 1.03 2010; IUPAC: Research Triangle Park, NC; <http://www.iupac.org/inchi/download/version1.03/INCHI-1-DOC.zip> (accessed October 19, 2011).
- (20) *The IUPAC International Chemical Identifier (InChI)*; IUPAC: Research Triangle Park, NC; <http://www.iupac.org/inchi/> (accessed October 19, 2011).
- (21) *CTFile formats* (July 2010); Acclerys: San Diego, CA; <https://community.accelrys.com/servlet/JiveServlet/download/3451-1-5658/ctfile.pdf> (accessed October 19, 2011).
- (22) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31–36.
- (23) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Rec. Sig. Trans.* **2003**, 23, 361–371.
- (24) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, 46, 2342–2354.
- (25) Milletti, F.; Storch, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, 49, 68–75.
- (26) Warr, W. A. Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.* **2010**, 24, 497–520.
- (27) *Command-line program and Pipeline Pilot component from Chemical Computing Group*; Chemical Computing Group: Montreal, Canada; <http://www.chemcomp.com/journal/sdtools.htm> (accessed October 19, 2011).
- (28) *Command-line program and Pipeline Pilot component from Molecular Networks*; Molecular Networks: Erlangen Germany; <http://www.molecular-networks.com/products/tautomer> (accessed October 19, 2011).
- (29) *API and Calculator-plugin from ; ChemAxon Kft.*: Budapest, Hungary; <http://www.chemaxon.com/products/calculator-plugins/tautomerization/> (accessed October 19, 2011).
- (30) *API from OpenEye, Inc.*; OpenEye, Inc.: Santa Fe, NM; <http://www.eyesopen.com/products/applications/quacpac.html> (accessed October 19, 2011).
- (31) Sayle, R. A. So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, 24, 485–496.
- (32) Sayle, R.; Delany, J. Canonicalization and Enumeration of Tautomers. In Proceedings of EuroMug99, Cambridge, U.K., October 28–29, 1999; [http://www.daylight.com/meetings/emug99/Delany/taut\\_html/index.htm](http://www.daylight.com/meetings/emug99/Delany/taut_html/index.htm); Daylight: Laguna Niguel, CA (accessed October 19, 2011).
- (33) *QuacPac Quality Atomic Charges, Proton Assignment and Canonicalization*, version 1.3.1 2008; OpenEye: Santa Fe, NM; <http://www.eyesopen.com/docs/html/quacpac/index.html> (accessed October 19, 2011).
- (34) In working with our collaborators we have seen vocabularies containing from 7 to over 15 stereochemistry flags.
- (35) The xml file is available as Supporting Information.
- (36) *Daylight Theory Manual*, V4.9 (2/1/2008); Daylight: Laguna Niguel, CA; <http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html> (accessed October 19, 2011).
- (37) On Sulfoxides the lone pair is treated as a neighbor atom.
- (38) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, 47, 3463–3482.
- (39) LaPlante, S. R.; Edwards, P. J.; Fader, L. D.; Jakalian, A.; Hucke, O. Revealing Atropisomer Axial Chirality in Drug Discovery. *ChemMedChem* **2011**, 6, 505–513.

(40) LaPlante, S. R.; D. Fader, L.; Fandrick, K. R.; Fandrick, D. R.; Hücke, O.; Kemper, R.; Miller, S. P. F.; Edwards, P. J. Assessing Atropisomer Axial Chirality in Drug Discovery and Development. *J. Med. Chem.* **2011**, *54*, 7005–7022.