JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

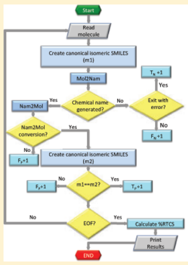
New Benchmark for Chemical Nomenclature Software

Edward O. Cannon*,[†]

[†]OpenEye Scientific Software, 9 Bisbee Court Suite D, Santa Fe, New Mexico 87508, United States

S *Supporting Information*

ABSTRACT: We propose a new, robust benchmark, called Percentage Round Tripping of Canonical Isomeric SMILES (%RTCS), for assessing the ability of chemical nomenclature software to convert chemical structures to names and chemical names to structures. The benchmark is based on a string comparison between canonical isomeric SMILES generated from the original structure and the resultant structure from round tripping. Using the latest version of the OpenEye chemical nomenclature toolkit, Lexichem v2.1.0, we report %RTCS values of over 92% on average for a variety of challenging compound collections.



$$\%RTCS = 100 \text{ x}\left(\frac{T_P}{T_P + F_P + F_N + T_N}\right)$$

Metric	Database				
	Maybridge	MDDR	NCI	PubChem	Wombat
%RTCS	98.69%	88.54%	92.32%	93.66%	89.54%
T_P	63036	98434	231038	28559368	47648
F_P	325	5551	4951	1089780	3907
F_N	511	7187	14262	843509	1659
Total	63872	111172	250251	30492657	53214

INTRODUCTION

The current explosion in chemical nomenclature software available on the market is both a blessing and a curse. While on one hand the users have a vast variety of programs to choose from, on the other hand it is very difficult to rationally decide which is the most appropriate software tool for a given task. Therefore, the existence of relevant benchmarks of performance is important.

In general, the purpose of a benchmark is to assess the effectiveness of a method in addressing a problem, and a variety of benchmarks can be applied to chemical nomenclature software. Quantitatively, the software could be assessed on its speed; how long does it take for a database of structures to be converted into names and back? Qualitatively the scope of the software can be assessed. For example does the software have a graphical user interface (GUI), is it command line only, or does it have an application programming interface (API)? If it does have an API, what programming languages are supported? Does the software work across multiple platforms such as Windows, Mac OS X, and Linux? Is the software multithread compatible? From a chemical perspective, what are the different classes of chemical compounds that are supported from both a name and structure perspective? Does the software support different nomenclature styles: Chemical Abstracts Service (CAS),[1] Traditional (names such as caffeine), MDL/Beilstein,[2] AutoNom,[3] OpenEye,[4] and the International Union of Pure and Applied Chemistry (IUPAC).[5] From a cheminformatics perspective it is important to know which file formats the software can accept as input and output. Possible formats include the following: Chemical Markup Language (CML),[6] simplified molecular input line entry specification (SMILES),[7] canonical SMILES,[4,8−11] canonical isomeric SMILES,[4,8,11] the IUPAC International Chemical Identifier (InChI),[12] Mol2,[13] and the Structure-Data file (SDF) (see Table 1).[14] While this list is not exhaustive, it does illustrate a large number of ways in which chemical nomenclature software can be assessed.

Table 1. Common Molecular Representations

line notation	connection tables
Wiswesser Line Notation (WLN)[29]	MDL Molfile[30]
Sybyl Line Notation (SLN)[31]	CML[6]
SMILES[7]	Tripos mol2 file[13]
canonical SMILES[4,8−11]	SDfile[14]
canonical isomeric SMILES[4,8,11]	Reaction-Data file (RDfile)[14]
InChI[12]	
InChI key[12]	
representation of organic structures description arranged linearly (ROSDAL)[32]	

All the assessment criteria listed above are universally applicable to every type of chemical nomenclature software. However, there is no universally accepted standard to evaluate their accuracy. This article describes the development of a new universal benchmark standard to measure the accuracy of chemical nomenclature software through the conversion of chemical structures into English IUPAC chemical names and back into their chemical structures, which is referred to as "round tripping". Such a benchmark is of use to scientists interested in comparing how well existing chemical nomenclature software performs on applications like converting chemical names from patents or journals into structures or measuring the percentage of correctly converted compounds into names (% RTCS) from screening libraries of compounds. By assessing the ability of different chemical nomenclature software[15−21] using a universal benchmark, researchers can make better informed decisions on which software is most appropriate for their needs.

A number of previous publications have evaluated the accuracy and reliability of existing chemical nomenclature software to convert

Received: January 23, 2012
Published: April 7, 2012

dx.doi.org/10.1021/ci3000419 | *J. Chem. Inf. Model.* 2012, 52, 1124−1131

names to structures and structures to names. In 1990 Wisniewski et al.[3] tested the capabilities of AutoNom name to structure conversion on a limited set of random samples taken from the Beilstein database. The software achieved a 63% correct conversion rate. More recently, Brecher[22] published a paper on "Name=Struct",[15] CambridgeSoft's automated system for name to structure conversion. Brecher rated the software's performance based on its comprehensiveness over a number of catalogs produced by commercial chemical vendors and 150000 synonyms within the ChemFinder WebServer.[23] A percentage was assigned to the total names that were correctly interpreted after individual examination. In 2006, Eller[24] reviewed the quality of 303 published chemical names randomly selected from 4 journals. The assessment took place by comparing chemists' abilities to construct names manually for these structures against names generated by 3 commercial nomenclature programs. The published and generated names were manually assessed using three criteria: (1) "No name", the software was unable to generate a name for the given structure; (2) "Unacceptable", the name generated for the structure was considered ambiguous; (3) "Unambiguous", was assigned to all other names. One of the main drawbacks with these approaches are that the data sets are randomly generated subsets, which are not reproducible. It is also very difficult to evaluate whether a name does represent the correct structure because the results are not validated against the input, and it is difficult to quantify the quality of the names as there could be more than one correct name dependent on the rules used in name construction. In Eller's work, the dependence on manual assessment of name creation and structure generation is time-consuming, and therefore this approach is not applicable when evaluating very large data sets of compounds. Another approach was introduced by Lowe et al.[25] They compared the: main, charge, stereochemical, and isotopic layers of their InChI output from their name to structure toolkit OPSIN with standard InChIs previously generated in PubChem,[26] the ChEBI database,[27] and the ChemBridge Catalogue.[28] While this approach is better, it does not take into account tautomeric information. The authors also noted issues with validating the stereochemical layer on the ChemBridge Catalogue.

Thus, the correct way of evaluating accuracy of chemical nomenclature software requires a rigorous examination of the final output through round tripping. In order to develop a robust benchmark standard to measure the accuracy, two questions need to be addressed:

- How can chemical structures be represented, and which is the best format?
- How can round tripping accuracy be defined?

In order to answer the first question, we have to consider the various chemical formats available. Other than chemical names or registration identifiers, two structural representations are available to chemists: line notations or connection tables (see Table 1). The distinction between these two representations is that line notations use alphanumeric characters to encode the molecular structure (the atoms, atom types, bonds, and stereochemistry if applicable, into a molecular graph, in which the atoms represent the vertices and the bonds the edges), in a single line. However, connection tables encode the molecular structure using two sections: an atom block, which in its simplest form, lists the atom numbers of the atoms in the structure along with its element type, and a bond block which lists all the bonds between the atoms in the structure and their bond type.

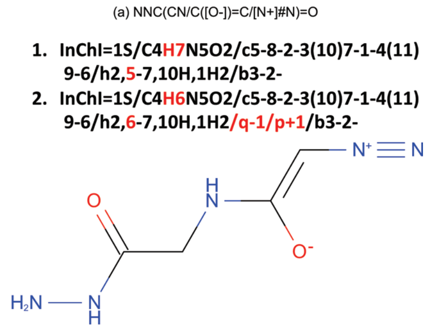
Connection Table Representations. While connection table representations have the advantage that they can completely encode all of the atoms and bonds of a structure into a graph and can easily be extended to allow additional information such as free electrons and stereochemistry, they have a number of disadvantages that do not make them a suitable choice for round tripping. First, connection table representations have much larger storage requirements. Second, the mixing of both coordinate and connectivity information can lead to redefinitions of the structure due to changes in the coordinates, for example in the case of conformers.

Linear Notation Representations. All linear notations shown in Table 1 have concise simple linear code representations of a structure, which are unambiguous. However, WLN, SLN, SMILES, ROSDAL, and InChI key do not provide a unique representation of a chemical structure, and canonical SMILES does not encode stereochemistry. Therefore the only two appropriate representations are canonical isomeric SMILES and InChI. For the present benchmark, we have chosen canonical isomeric SMILES as the preferred conversion format. While the main focus of this paper is not concerned with a direct comparison of these two formats, it is worthy of note the reasons for choosing canonical isomeric SMILES over InChI (for a summary see Table 2).

Table 2. Canonical Isomeric SMILES and Standard InChI Comparsion

feature	canonical isomeric SMILES	standard InChI
human readable	yes	no
reproducible	yes	no
tautomer support	yes	no
verbose	no	yes

Some of the key advantages of canonical isomeric SMILES over InChI are that they are human readable, less verbose than InChI, support additional stereochemistry such as octahedral and square planar geometries, can encode information on Markush structures, and retain bond order information, unlike InChI where bond order is ignored except when analyzing stereochemistry and H-migration. SMILES strings explicitly represent the position of electrons, in the case of radicals, InChI does not. InChI has problems with the reproducibility of the same InChI using two different molfiles encoding the same structure but using different atom numbering (see Figure 1 and Figure S1 for the molfiles). While this problem was fixed in InChI version 1.0.3 onward, there may be other similar examples not noted. Such a problem does not exist when using



(a) NNC(CN/C([O-])=C/[N+]#N)=O

1. InChI=1S/C4H7N5O2/c5-8-2-3(10)7-1-4(11)9-6/h2,5-7,10H,1H2/b3-2-
2. InChI=1S/C4H6N5O2/c5-8-2-3(10)7-1-4(11)9-6/h2,6-7,10H,1H2/q-1/p+1/b3-2-

Figure 1. Two different InChIs produced for the same structure.

OpenEye canonical isomeric SMILES.[33] SMILES can specify chirality completely or partially for any structure. Unlike standard InChI, tautomers are explicitly represented in SMILES (see Figure 2).

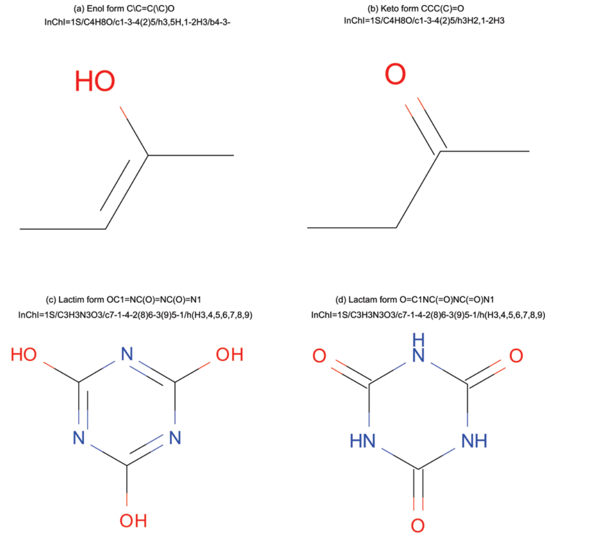

(a) Enol form C\C=C(\C)O
InChI=1S/C4H8O/c1-3-4(2)5/h3,5H,1-2H3/b4-3-
HO
(b) Keto form CCC(C)=O
InChI=1S/C4H8O/c1-3-4(2)5/h3H2,1-2H3
O
(c) Lactim form OC1=NC(O)=NC(O)=N1
InChI=1S/C3H3N3O3/c7-1-4-2(8)6-3(9)5-1/h(H3,4,5,6,7,8,9)
HO
N
OH
N
N
OH
(d) Lactam form O=C1NC(=O)NC(=O)N1
InChI=1S/C3H3N3O3/c7-1-4-2(8)6-3(9)5-1/h(H3,4,5,6,7,8,9)
O
H
N
O
HN
NH
O

Figure 2. Tautomer representations.

In answer to the second question, the benchmark introduced in this paper, is called Percentage Round Tripping of Canonical Isomeric SMILES, abbreviated to %RTCS. This benchmark represents the percentage of a database that can be converted from its original canonical isomeric SMILES back into the same canonical isomeric SMILES after the structure to name and name to structure conversion has taken place. The advantage of this approach over round tripping starting from a chemical name is that the start and end point are unique, unlike chemical name representation, where there are multiple styles of name that could be chosen.

■ MATERIALS AND METHODS

This section addresses the round tripping benchmark and how accuracy is defined in this context. Lexichem[16] the OpenEye chemical nomenclature toolkit is introduced. A procedure based on unique parent ring systems is discussed for the purpose of identifying how difficult it is to round trip a database of compounds. A method to measure the effectiveness with which chemical nomenclature software converts names to structures and structures to names is given. All results outlined in this paper have been obtained using the Lexichem toolkit version 2.1.0. Speed performance metrics have been obtained using the C++ version of the toolkit.

Lexichem. Lexichem is capable of converting chemical structures to chemical names and chemical names back to chemical structures. It is written in C++[34] and has been wrapped using SWIG[35] to Java,[36] C#,[37] and Python.[38] Lexichem supports a number of nomenclature styles: IUPAC 79/93/2005,[5] Chemical Abstracts (CAS),[1] Traditional, MDL/Beilstein,[2] and AutoNom.[3] Lexichem also supports foreign language translation in over 10 different languages.

Round Tripping Benchmark. Figure 3 shows the work flow of the round tripping procedure. In the first stage, structures are made unique, by converting them to canonical isomeric smiles. In the second stage, Lexichem attempts to convert the structures into names. If a chemical name is not generated, the work flow stops and the next structure is read. If a chemical name is generated, Lexichem then attempts to convert the name back into a structure. If a structure is generated, it is then canonicalized, and a string comparison is made between the original canonical isomeric SMILES and that generated after round tripping.

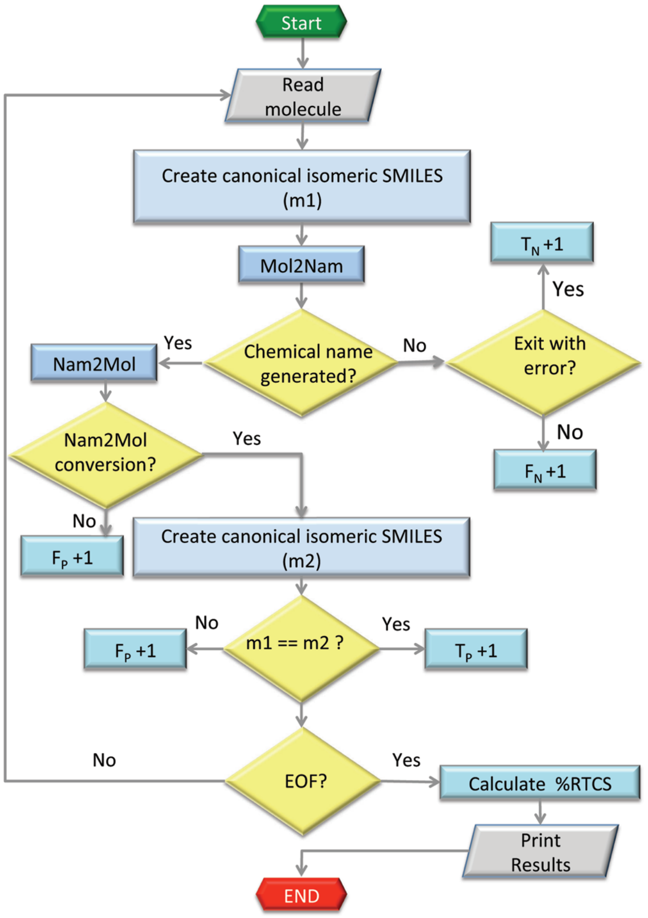


Figure 3. Flowchart showing the necessary steps to perform round tripping.

The formula to calculate %RTCS is

$$\%\text{RTCS} = 100 \times \left(\frac{T_P}{T_P + F_P + F_N + T_N} \right)$$

T_P, F_P, F_N, and T_N are abbreviations for true positives, false positives, false negatives, and true negatives respectively. T_P is the number of canonical isomeric SMILES that, when converted from a structure to a name and back, are identical to the starting canonical isomeric SMILES. The F_P value represents the number of structures that are converted to a chemical name, but when converted back produce a different canonical isomeric SMILES, or none at all. F_N gives the number of structures that fail to generate a chemical name. T_N represents the number of structures not converted to a name, causing the application to exit with an error.

One of the main advantages of %RTCS, is that it compares both the unique original and final structure after round tripping. This is not the case in the previous benchmark used for Lexichem.[4] The obvious error in the previous approach was that the generated structure could be different from the original structure but still be considered a T_P when it is actually a F_P. For example, the correct structure in Figure 4(a) with the green tick

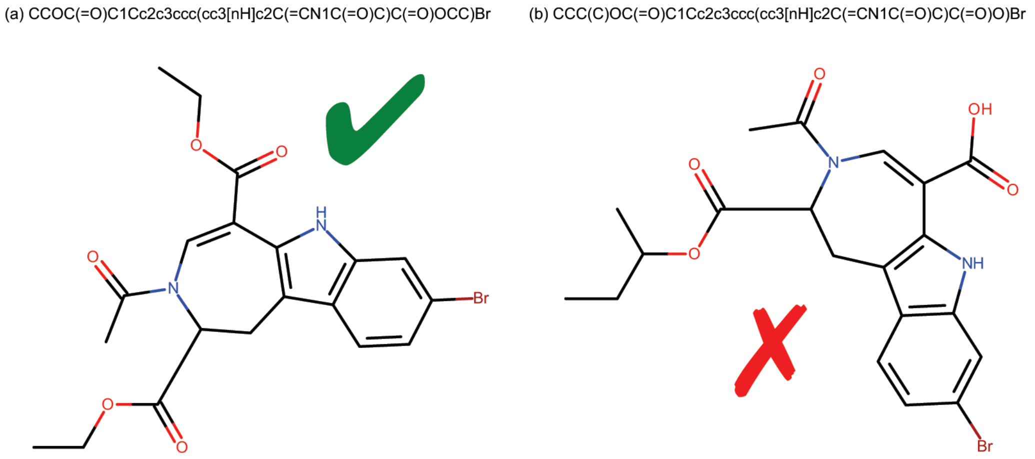

(a) CCOC(=O)C1Cc2c3ccc(cc3[nH]c2C(=CN1C(=O)C)C(=O)OCC)Br (b) CCC(C)OC(=O)C1Cc2c3ccc(cc3[nH]c2C(=CN1C(=O)C)C(=O)O)Br

Figure 4. 3-Acetyl-8-bromo-1,2,3,6-tetrahydro-azepino[4,5-b]indole-2,5-dicarboxylic acid diethyl ester was considered a pass in the old versions of Lexichem's benchmark.

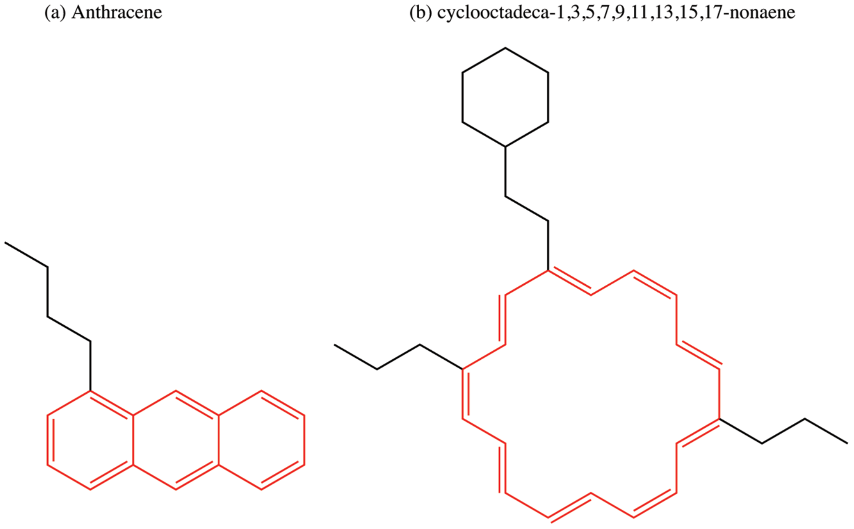

(a) Anthracene (b) cyclooctadeca-1,3,5,7,9,11,13,15,17-nonaene

Figure 5. Unique parent ring systems highlighted in red.

shows two ester groups. The incorrect structure Figure 4(b) shown with the red cross shows one ester group with branched alkyl chain attachments and one carboxylic acid group. In Lexichem's old benchmark the incorrect structure on the right was assessed as correct and thus considered a T_P, when it is in fact a F_P.

Performance Metrics. While calculating %RTCS might seem straightforward, computing this measure over large databases is not always a trivial task. The diversity of the database as well as the time required to convert it are important additional performance metrics.

Unique Parent Ring Systems. A measure of the number of unique parent ring systems gives an indication of how difficult it is to convert a database of compounds into names and back into structures. This is because, for each ring system a new template or rule must be present to allow Lexichem to convert structures with that ring system to names and a new token (a morpheme, representing the smallest semantically meaningful unit in chemical nomenclature) and set of rules must be present in Lexichem's parser to convert that name back into a structure. The absence of either of these rules will result in Lexichem failing to round trip a structure with that ring system. Selection of the parent ring system in a structure is performed in accordance with the criteria for choosing a senior ring or ring system in the IUPAC nomenclature system. One of the limitations of this approach is that simple ring systems will be ignored, such as the cyclohexane ring in Figure 5. The number of unique ring systems in a data set is quick to calculate and gives an estimate of the number of rules required by Lexichem to encode all the parent ring system

Table 3. Unique Parent Ring Systems

	database				
	Maybridge	MDDR	NCI	PubChem	Wombat
unique parent ring systems	2400	10421	21077	183832	2686
database size/unique parent ring systems	26.61	10.67	11.87	165.87	19.81

Table 4. Round Tripping Results

	database				
metric	Maybridge	MDDR	NCI	PubChem	Wombat
%RTCS	98.69%	88.54%	92.32%	93.66%	89.54%
T_P	63036	98434	231038	28559368	47648
F_P	325	5551	4951	1089780	3907
F_N	511	7187	14262	843509	1659
Total	63872	111172	250251	30492657	53214

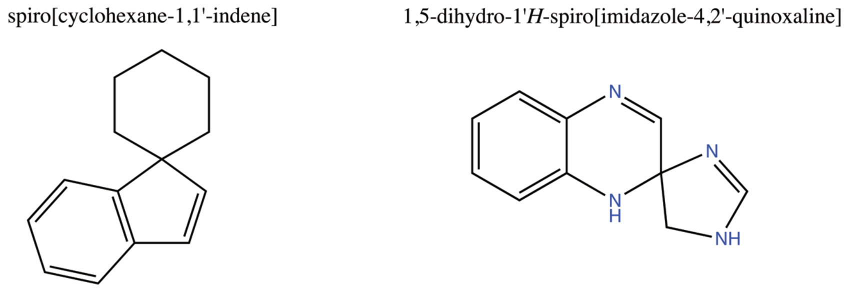

spiro[cyclohexane-1,1'-indene]
1,5-dihydro-1'H-spiro[imidazole-4,2'-quinoxaline]
N
N
H
N
NH

Figure 6. Monospiro compounds for which a structure cannot be generated from a name.

information in that database, which is a reflection of the difficulty of round tripping that database.

Speed Test. Wall-clock time has been used to measure the time taken to perform structure to name and name to structure conversions. The wall-clock time is made up of three components: input/output time, central processing unit (CPU) time, and the communication channel delay (if data are present on more than one machine).

Databases. Five databases have been chosen to generate % RTCS performance benchmarks. All databases were prefiltered using OpenEyes' OEChem toolkit high level structure reader function. A structure passes the validity test if the number of atoms is greater than or equal to one. Except for the 2784 null entries in the MDL Drug Data Report database, all the other databases passed the OEReadMolecule filter.

2003 Maybridge Catalog. The Maybridge Catalog[39] contains organic drug-like structures that tend to follow Lipinski's rule of 5.[40] In this work the 2003 Maybridge Catalog is used which contains 63872 structures.

2003 MDL Drug Data Report. The MDDR database is provided by Accelrys[41] and Thomson Reuters[41] and contains 113956 structures of biologically relevant structures. The database has information on the structure: registration number, name, generic name, activity index, class, and the phase in the research and development process. 2784 structures of the total 113956 were considered invalid by the OEReadMolecule() function, resulting in a data set of 111172 structures. Structure registration numbers for the structures processed in the MDDR can be found in Table S1. It should be noted that only 113842 of the structures in this data set had structure registration numbers (114 had no registration number). The remaining 2670 unprocessable structure registration numbers are listed in Table S2. These registration numbers had no atoms present in their entries. One randomly selected example taken from the 2670 entries is shown in Figure S2.

August 2000 National Cancer Institute. The August 2000 National Cancer Institute (NCI) database[42] is freely available online and contains 250251 unique SMILES strings canonicalized using Daylight's[8] rules. The SDfile version contains CAS Registry numbers and SMILES strings.

2003.2 WOMBAT. The 2003.2 version of the WOMBAT database[43] used in this work contains 53214 structures of bioactive compounds.

October 2011 PubChem. PubChem[26] is part of the NIH's Molecular Libraries Roadmap Initiative. It is comprised of three linked databases: Bioassay, Compound, and Substance. In this work a snapshot of the October 2011 version of the PubChem Compound database has been downloaded. It contains 30492657 chemical structures. Each entry contains information on the structure, numerous name descriptions (IUPAC, CAS, Systematic and traditional), and various additional chemical formats such as canonical, and canonical isomeric SMILES, InChI, and InChI key.

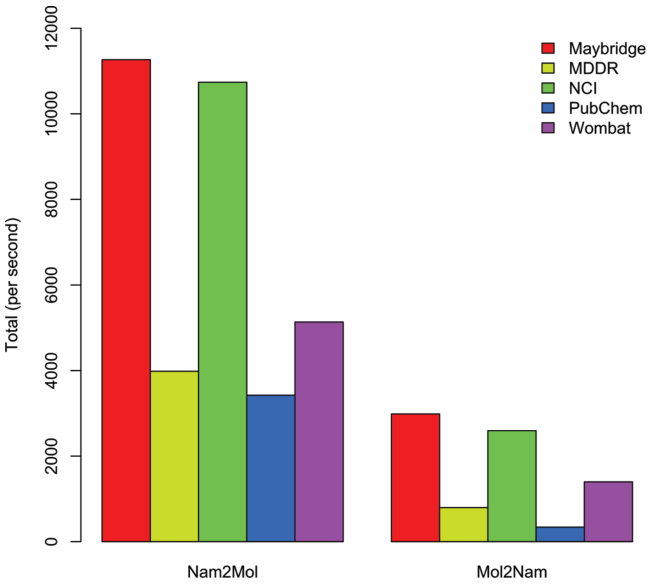


Figure 7. Rate of conversion of names to structures and structures to names per second on the respective databases using an Intel(R) Core(TM) i7 CPU 930 @ 2.80 GHz.

Table 5. Wall-Clock Time for Structure to Name and Name to Structure Conversion

	database				
	Maybridge	MDDR	NCI	PubChem	Wombat
Nam2Mol (secs)	5.67	27.90	23.30	8906.99	10.36
names per second	11265	3985	10740	3423	5136
Mol2Nam (secs)	21.39	139.73[a]	96.41	89388.76	38.03
Mol2Nam no CIP (secs)	20.75	85.14	97.22	30397.18	31.73
Structures per second	2986	796	2596	341	1399
structures per second no CIP	3078	1306	2574	1003	1677
structures with CIP stereo	0	51321	0	10415568	23555

[a] Removal of 4 large ring structures.

■ RESULTS AND DISCUSSION

Round Tripping. Table 3 shows the number of unique parent ring systems present in the 5 databases. The Maybridge

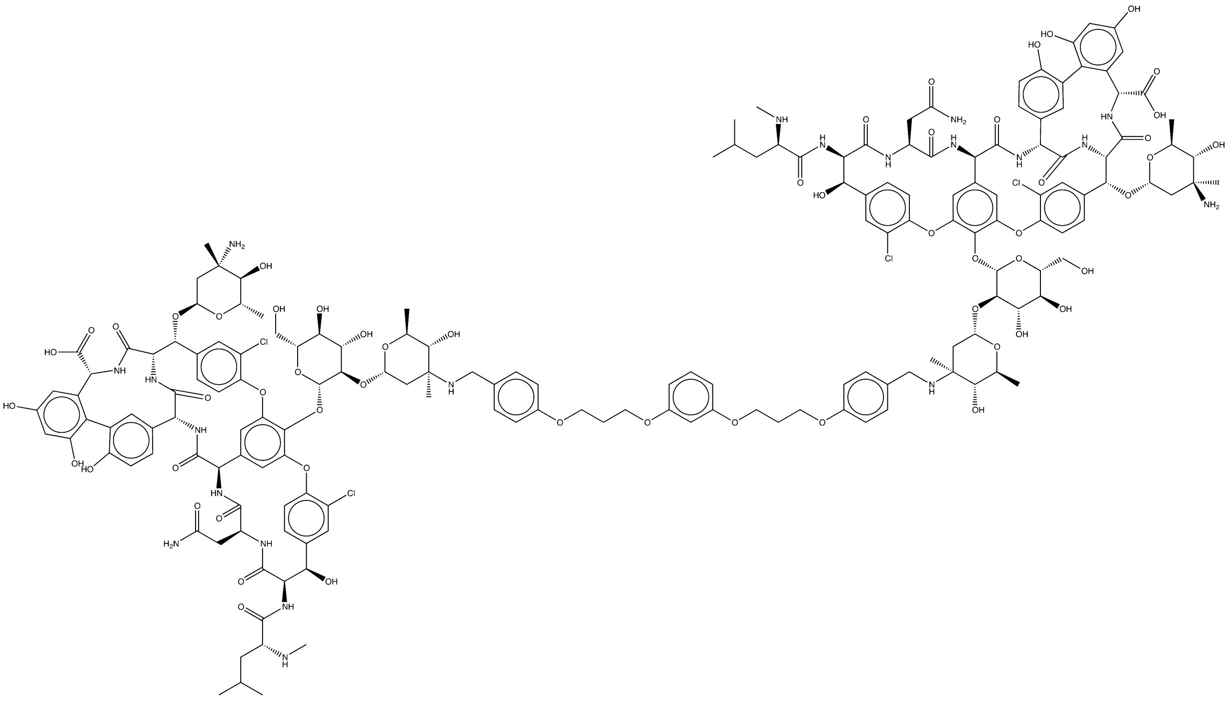


Figure 8. Example structure with a large number of stereogenic atoms. Mol2Nam processing time for this structure was 175.64 s.

and Wombat databases have the least number of unique ring systems (1 unique ring system per 27 and 20 structures, respectively), suggesting that these databases require chemical nomenclature software to encode fewer rules and thus may be faster and more accurate to round trip. Both the MDDR and NCI have considerably more unique parent ring systems, suggesting these databases may take longer and be less accurate to round trip. While this is true in the case of the MDDR and NCI databases when comparing %RTCS values to the Maybridge database (see Table 4), it is not true with respect to the Wombat database. One explanation could be that the Wombat database has more diverse ring systems, or a larger number of structures with ring systems present in their structure other than the parent ring system for which Lexichem does not have rules to encode. The PubChem database has the greatest number of unique parent ring systems (183832); however, it also has the largest number of structures in its database (30492657). One may expect that based on the high ratio of database size to unique parent ring systems this database would give the highest %RTCS value of all the databases for chemical nomenclature software to round trip. Though looking at the round tripping results in Table 4 this is not the case (for the purpose of comparison, round tripping results for the previous version, version 2.0.2 of the Lexichem toolkit are shown in Table S3). It is clear that the large number of unique parent ring systems (183832) and hence rules make this database harder to predict than the likes of the Maybridge database, on which Lexichem, performs best with a %RTCS value of 98.69%. Lexichem performed least well on the MDDR database. This is not surprising when there is one new unique ring system for every 11 structures.

One of the advantages of Lexichem is that it has false positive rates of 0.51% (Maybridge), 4.99% (MDDR), 1.98% (NCI), 3.57% (PubChem), and 7.34% (Wombat). That is, Lexichem is not likely to produce an incorrect structure from a correct name, it is more likely to either return the correct structure outright or fail quickly. Of these false positives, spiro based compounds make up the vast majority of these name to structure failures; in particular, monospiro compounds with different components at least one of which is polycyclic (see Figure 6). For this class of compounds, no structure is generated from the name. This is an additional area of development that needs to be addressed in future releases of Lexichem.

From a false negative perspective, that is structures that do not get converted into chemical names, Lexichem performs well. F_N's make up 0.80%, 6.46%, 5.70%, 2.77%, and 3.12% of the Maybridge, MDDR, NCI, PubChem, and Wombat's total structures, respectively. Of these F_N's von Baeyer name (Section P23 of the IUPAC Nomenclature of Organic Chemistry Blue Book)[44] generation is the leading cause of these failures, giving rise to 467 (91.39%), 6608 (91.94%), 11512 (80.72%), 402998 (47.78%), and 1530 (92.24%) failed structure to name conversions for the Maybridge, MDDR, NCI, PubChem, and Wombat databases. For information on the number of false positives and negatives with rings systems in for each database, see Table S4.

Speed. To give an overall impression of the speed of Lexichem, we have compared the wall-clock time for the separate structure to name and name to structure conversions for each database (see Figure 7 and Table 5).

It is evident that converting chemical names into structures takes considerably less time than structure to name conversion. One of the main reasons why the MDDR and PubChem databases take a much longer time for structure to name conversion is because these databases have a large number of structures for which Cahn–Ingold–Prelog[45] (CIP) stereochemistry must be determined (see Figure 8).

Table 5 shows the wall-clock time in seconds for structure to name conversion for each database when no CIP stereochemistry is present. For databases where no CIP stereochemistry was present in the original database (Maybridge and NCI), the times are comparable to the original values. However, the processing time is now reduced for the MDDR, PubChem, and Wombat databases when CIP stereochemistry has been removed.

■ CONCLUSIONS

A new, more rigorous, benchmark to measure the performance of chemical nomenclature software called Percentage Round Tripping of Canonical Isomeric SMILES (%RTCS) has been introduced. The benchmark is based on a string comparison between the canonical isomeric SMILES of the starting structure and the final structure after name generation and conversion back to canonical isomeric SMILES. This benchmark is quick to calculate, scales well for large databases, and gives an easily interpretable numeric value in the form of a percentage, and unlike previous benchmarks, this benchmark has the advantage that it validates the original input against the generated output.

In this paper, the OpenEye chemical nomenclature toolkit, Lexichem, has averaged %RTCS values greater than 92% across the Maybridge, MDL Drug Data Report, National Cancer Institute, PubChem, and WOMBAT databases. By increasing the bar, we hope that others will use this benchmark and that it will become the universal performance benchmark when comparing chemical nomenclature software.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1. Processable MDDR registration numbers. Table S2. Unprocessable MDDR registration numbers. Table S3. Lexichem v2.0.2%RTCS values. Table S4. Count of false positives and negatives with ring systems. Figure S1. Example molfiles producing different InChI for the same structure using InChI version 1.0.2. Figure S2. Example of a MDDR structure with an unprocessable registration number. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ed.cannon@eyesopen.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The author would like to thank Dr. Bob Tolbert, Dr. Geoff Skillman, Dr. Paul Hawkins at OpenEye, and Dr. Cristina Sisu at Yale University, for proofreading this manuscript.

■ REFERENCES

(1) CAS, Chemical Abstracts Service, Columbus, OH, USA. http://www.cas.org/ [accessed October 2011].

(2) Laszlo, D. The Beilstein Structure Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 320−326.

(3) Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 324−332.

(4) *Lexichem TK* - C++, version 2.1.0, OpenEye Scientific Software, Santa Fe, NM, USA. http://www.eyesopen.com/docs/toolkits/current/html/Lexichem_TK-c++/index.html [accessed October 2011].

(5) IUPAC, International Union of Pure and Applied Chemistry, Research Triangle Park, NC, USA. http://www.iupac.org/ [accessed October 2011].

(6) Murray-Rust, P. Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928−942.

(7) 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.

(8) SMILES, Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, USA. http://www.daylight.com/about/index.html [accessed October 2011].

(9) Chemical Computing Group, Montreal, Quebec, Canada. http://www.chemcomp.com/aboutccg-contact.htm [accessed October 2011].

(10) *MolSoft*, Molsoft LLC, San Diego, CA, USA. http://www.molsoft.com/contacts.html [accessed October 2011].

(11) Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. L. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(12) McNaught, A. The IUPAC International Chemical Identifier:-InChI. *Chem. Int.* **2006**, *28*, 12−15.

(13) Mol2, Tripos, St. Louis, MO, USA. http://tripos.com/data/support/mol2.pdf [accessed October 2011].

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

(15) Brecher, J. *Name=Struct*: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 943−950.

(16) Sayle, R. Foreign Language Translation of Chemical Nomenclature by Computer. *J. Chem. Inf. Model.* **2009**, *49*, 519−530.

(17) *ACD/Name*, version 12, Advanced Chemistry Development, Inc., Toronto, ON, Canada. http://www.acdlabs.com/ [accessed October 2011].

(18) *IUPAC DrawIt*, Bio-Rad Laboratories, Hercules, CA, USA. http://www.bio-rad.com/ [accessed October 2011].

(19) *Name<>Structure*, ChemAxon, Budapest, Hungary. http://www.chemaxon.com/ [accessed October 2011].

(20) *Name to Structure*, InfoChem, Munich, Germany. http://infochem.de/ [accessed October 2011].

(21) *NameExpert*, ChemInnovation Software, San Diego, CA, USA. http://www.cheminnovation.com/ [accessed October 2011].

(22) PerkinElmer, Inc., 75 Nicholson Ln, San Jose, CA,USA. [accessed March 2012].

(23) *ChemFinder*, CambridgeSoft Corporation, Cambridge, UK. http://www.cambridgesoft.com/ [accessed October 2011].

(24) Eller, G. A. Improving the Quality of Published Chemical Names with Nomenclature Software. *Molecules* **2006**, *11*, 915−928.

(25) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, and Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51*, 739−753.

(26) Bolton, E.; Wang , Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry (ARCC)*; ACS COMP division: 2008; Vol. *4*, Chapter 12, pp 217−241

(27) *ChEBI*, European Bioinformatics Institute, Hinxton, Cambridge, UK. http://www.ebi.ac.uk/chebi/ [accessed October 2011].

(28) *ChemBridge Database*, ChemBridge Co., San Diego, CA, USA. http://www.chembridge.com/ [accessed October 2011].

(29) Wiswesser, W. J. *A Line-Formula Chemical Notation*; Crowell: New York, 1954.

(30) Accelrys, Inc., San Diego, CA, USA. [accessed October 2011].

(31) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.

(32) Barnard, J. M.; Jochum, C. J.; Welford, S. M. In *A Universal Structure/Substructure Representation for PC-Host Communication*; Warr, W. A., Ed.; ACS Symposium Series; Chemical Structure Information Systems: American Chemical Society: Washington, DC, 1989; Vol. *400*; pp 76−81.

(33) *OEChem TK* - C++, version 1.7.6, OpenEye Scientific Software, Santa Fe, NM, USA. http://www.eyesopen.com/docs/toolkits/

current/html/OEChem_TK-c++/index.html [accessed October 2011].

(34) Stroustrup, B. *The C++ Programming Language*; Addison-Wesley: One Jacob Way, Reading, MA, 1985.

(35) Beazley, D. M. SWIG: An Easy to Use Tool for Integrating Scripting Languages with C and C++. In *Proceedings of the 4th USENIX Tcl/Tk Workshop*, Monterey, CA, USA July 6−10, 1996.

(36) *Oracle Java*, Oracle Co., Redwood Shores, CA, USA. http://www.oracle.com/ [accessed October 2011].

(37) *C#*, Microsoft Co., Redmond, WA, USA.

(38) van Rossum, G.; de Boer, J. Interactively Testing Remote Servers Using the Python Programming Language. *CWI Quarterly* **1991**, *4*, 283−303.

(39) *Maybridge Online Catalog*, Thermo Fisher Scientific, Inc., Maybridge, Trevillett, Tintagel, Cornwall, UK. http://www.maybridge.com/ [accessed October 2011].

(40) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(41) MDDR licensed by Accelrys, Inc, San Diego, CA, USA. [accessed October 2011].

(42) *NCI* Open Database Compounds, NCI/CADD Group, National Cancer Institute, Bethesda, MD, USA. http://cactus.nci.nih.gov/ [accessed October 2011].

(43) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Mracec, Z. S. M.; Oprea, T. I. 9. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co: Weinhem, 2005.

(44) IUPAC, Nomenclature of Organic Chemistry. http://old.iupac.org/reports/provisional/abstract04/BB-prs310305/CompleteDraft.pdf [accessed October 2011].

(45) Cahn, R. S.; Ingold, C. K.; Prelog, V. Specification of Molecular Chirality. *Angew. Chem., Int. Ed.* **1966**, *5*, 385−415.