CAN WE AGREE ON THE STRUCTURE REPRESENTED BY A SMILES STRING? A BENCHMARK DATASET

Noel M. O'Boyle, John W. Mayfield, Roger A. Sayle

NextMove Software Ltd, Cambridge, UK

https://github.com/nextmovesoftware/smilesreading

Motivation

Our starting point is the axiom that a SMILES string represents a particular molecule. The job of a SMILES reader is to faithfully recreate that molecule.

We quantify to what extent different SMILES readers agree on the molecule represented by a SMILES string. Our goal is to improve the interoperability of SMILES strings by identifying ambiguities in the specification and by working with toolkit developers to resolve bugs.

Benchmark set 1: SMILES valence model

How many hydrogens are on the nitrogen in `N(C)(C)(C)C`? This atom type (*N4*) was tested, along with 60 other atom types. Disagreements with the specification [1] (and Dave Weininger's own code [2]) are listed below.

Avalon	Cl2 Cl4 Br2 Br4 I2 I4
BIOVIA Draw	Cl2 Cl4 Br2 Br4 I2 I4
Cactvs	N4.P4.S3.S5 (or N4*)
CDK	
CEX (Weininger)	
ChemDoodle	
ChemDraw	
Indigo†	
iwtoolkit	N4 Cl2 Cl3 Cl4 Cl5 Br2 Br3 Br4 I2 I4 (or P4 S3 S5*)
JChem	N4
KnowItAll	
OEChem	
Open Babel	
OpenChemLib	N4 Cl2 Cl4 Br2 Br4 I2 I4
RDKit†	P6 I3 I4

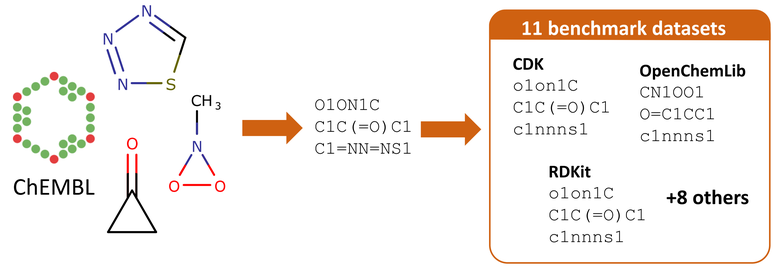
** If the default options are modified*
† Results exclude 17 atom types rejected by Indigo, and 19 rejected by RDKit

As a sanity check, the test was repeated but with hydrogen count specified, e.g. `[NH](C)(C)(C)C`. This is respected by all of the toolkits. Interestingly, Indigo no longer rejects any of the atom types.

Bibliography

1. **Daylight Theory Manual** *http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html*
2. Weininger, D. **Chemical EXchange 1.3.2**. *https://github.com/nextmovesoftware/CEX*
3. O'Boyle, N.M.; Mayfield, J.W. *https://www.slideshare.net/baoilleach/we-need-to-talk-about-kekulization-aromaticity-and-smiles*

Benchmark set 2: Aromatic SMILES for ChEMBL ring systems



The dataset contains **47463** unique ring systems derived from ChEMBL 23. Non-ring atoms were included if attached via double bonds, or via single bonds but only if from a non-carbon ring atom.

For each of the 11 benchmark datasets, every toolkit tested was required to:

1. read the SMILES
2. report any kekulization or parse errors
3. report the hydrogen count on each atom (if no error)

Comparison 1: Compare readers on the same dataset, from CDK

	Different H Count	Kekulization Failure
Avalon	0	1
BIOVIA Draw	0	0
CDK	0	0
ChemDoodle	13*	
ChemDraw	7	25
Indigo†	456	23
iwtoolkit	91	69
JChem	5	8
OEChem	0	0
Open Babel	0	0
OpenChemLib	9	136
RDKit†	7	1

** It is not possible to distinguish between kekulization failures and differences in hydrogen count*
† Results exclude 8 structures rejected by Indigo, and 15 by RDKit

Myth bust: Do differences in aromaticity models create problems for SMILES readers? **No** – the problems are caused by kekulization algorithms that are not sufficiently robust. [3]

Comparison 2: Compare to Open Babel across all 11 datasets

By comparing to a particular reader across all datasets, corner cases and bugs can be identified. Here are results compared to Open Babel, counting how many SMILES resulted in different hydrogen counts or where one program gave an error but the other did not.

	Differences	Ignoring errors
Avalon	166	33
BIOVIA Draw	2837	21
CDK	205	24
ChemDoodle	4333	179
ChemDraw	1027	71
Indigo	6110 (6062*)	1769
iwtoolkit	6839	1179
JChem	318	43
OEChem	436	18
Open Babel	-	-
OpenChemLib	1367	89
RDKit	342 (235*)	50

** Differences ignoring errors about bad valence*

If we inspect the CDK results, we find that SMILES with contradictory stereobond symbols (e.g. `C1CCCCN2/C(=N\1)\CN=C2`) are accepted by Open Babel (with warning) but rejected by CDK. Another case is SMILES with stereobond symbols in aromatic rings; these are treated by Open Babel as explicit single bonds but by CDK as implicit bonds.

Conclusions

We believe this benchmark dataset to be a useful resource for the improvement of SMILES interoperability. For all of the toolkits tested, the results yield a treasure trove of corner cases and bugs. These results have already led to changes to Cactvs, CDK, ChemDoodle, iwtoolkit, KnowItAll, Open Babel and OpenChemLib.

We encourage any toolkit developers interested in improving SMILES interoperability to get in touch, or just download the benchmark at the URL above and try it out.

Acknowledgements and software versions

Thanks to the developers of many of the toolkits tested for interesting discussions, and Matt Swain for providing results. The toolkits tested were: Avalon 1.2, BIOVIA Draw 2018, Cactvs 3.4.6.25, CDK 2.1, CEX 1.3.2, ChemDoodle API 2.3.0, ChemDraw 16.0, Indigo 1.3.0b.r16, iwtoolkit Oct2017, JChem 17.23, KnowItAll 2018, OEChem Feb 2018, Open Babel (dev) May2018, OpenChemLib 2018.5.0, RDKit 2018.03.1.