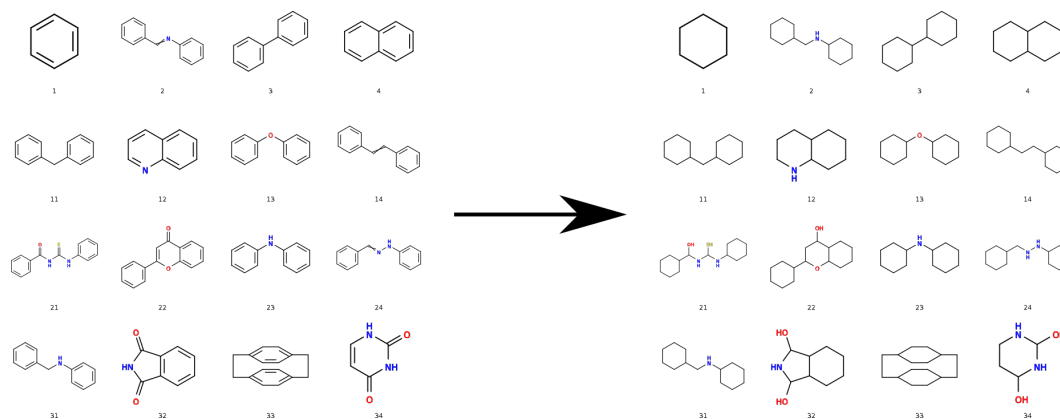


## 1 Background

The Bemis-Murcko scaffold<sup>1</sup> provided by DataWarrior<sup>2</sup> retains information about bond order and chirality. Sometimes, however, it suffices to retain only atom connectivity, like an assumption «there are only single bonds». Note, DataWarrior equally offers the export of Bemis-Murcko skeleton, however this simplifies e.g. the scaffold about an imidazole into one of cyclopentane.



## 2 Typical use

Running from the command line, the script processes one, or multiple SMILES strings listed in an input file (symbolized by text below) to report the processed SMILES strings back to the command line:

```
python saturate_murcko_scaffolds.py [-h] [-o] text
```

Instead of an input file, individual SMILES strings equally may be processed; to prevent their misinterpretation by the shell, enclose them by either single, or double quotes. Either approach elected, the result may be redirected into a permanent record with a user defined file name (flag `--outfile`, or shorter `-o`). Only modules the standard library of Python 3 provides (e.g., version 3.11.2) are used.

## 3 Example

For a collection of organic materials, the Bemis-Murcko scaffolds were extracted with DataWarrior (then release 5.0.0 for Linux, January 2019) as listing `test_input.smi` including higher bond orders (see folder `test_data`). The effect of the «artificial saturation» is easy to recognize while comparing the scaffold lists (fig. 1) in a difference view of the two `.smi` files.

<sup>1</sup>Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893 (<https://doi.org/10.1021/jm9602928>).

<sup>2</sup>Sander, T.; Freyss, J.; Von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473 (<https://doi.org/10.1021/ci500588j>). The

013	<chem>c(cc1)ccc10c1cccc1</chem>	013	<chem>C(CC1)CCC10C1CCCC1</chem>
014	<chem>C(c1cccc1)=C/c1cccc1</chem>	014	<chem>C(C1CCCC1)CC1CCCC1</chem>
015	<chem>c1cc2cc3cccc3cc2cc1</chem>	015	<chem>C1CC2CC3CCCC3CC2CC1</chem>
016	<chem>O=C(c1cccc1)c1cccc1</chem>	016	<chem>OC(C1CCCC1)C1CCCC1</chem>
017	<chem>c1c[nH]c2c1cccc2</chem>	017	<chem>C1C[NH]C2C1CCCC2</chem>
018	<chem>c(cc1)ccc1/N=N/c1cccc1</chem>	018	<chem>C(CC1)CCC1NNC1CCCC1</chem>
019	<chem>C(c1cccc1)=N/N=C/c1cccc1</chem>	019	<chem>C(C1CCCC1)NNCC1CCCC1</chem>
020	<chem>C(Cc1cccc1)c1cccc1</chem>	020	<chem>C(CC1CCCC1)C1CCCC1</chem>
021	<chem>O=C(c1cccc1)NC(Nc1cccc1)=S</chem>	021	<chem>OC(C1CCCC1)NC(NC1CCCC1)S</chem>
022	<chem>O=C1c(cccc2)c2OC(c2cccc2)=C1</chem>	022	<chem>OC1C(CCCC2)C2OC(C2CCCC2)C1</chem>
023	<chem>c(cc1)ccc1Nc1cccc1</chem>	023	<chem>C(CC1)CCC1NC1CCCC1</chem>
024	<chem>C(c1cccc1)=N/Nc1cccc1</chem>	024	<chem>C(C1CCCC1)NNC1CCCC1</chem>
025	<chem>O=C(C=CN1[C@@H]2OCCC2)NC1=O</chem>	025	<chem>OC(CCN1[C@@H]2OCCC2)NC1O</chem>
026	<chem>c1ccc2c(-c3cccc4cccc34)cccc2c1</chem>	026	<chem>C1CCC2C(-C3CCCC4CCCC34)CCCC2C1</chem>
027	<chem>c1ccc(C(c2cccc2)c2cccc2)cc1</chem>	027	<chem>C1CCC(C(C2CCCC2)C2CCCC2)CC1</chem>
028	<chem>c(cc1)cc2c1[nH]c1c2cccc1</chem>	028	<chem>C(CC1)CC2C1[NH]C1C2CCCC1</chem>
029	<chem>c(cc1)ccc1P(c1cccc1)c1cccc1</chem>	029	<chem>C(CC1)CCC1P(C1CCCC1)C1CCCC1</chem>
030	<chem>c1c(-c2cccc2)oc2c1cccc2</chem>	030	<chem>C1C(-C2CCCC2)OC2C1CCCC2</chem>
031	<chem>C(c1cccc1)Nc1cccc1</chem>	031	<chem>C(C1CCCC1)NC1CCCC1</chem>
032	<chem>O=C(c1c2cccc1)NC2=O</chem>	032	<chem>OC(C1C2CCCC1)NC2O</chem>
033	<chem>C(Cc1ccc(CC2)cc1)c1ccc2cc1</chem>	033	<chem>C(CC1CCC(CC2)CC1)C1CCC2CC1</chem>

**Figure 1:** Difference view of the SMILES strings of a Murcko scaffold *prior* (left hand column) and *after* an «artificial saturation» (right hand column). The processing affects explicit bond order indicators, e.g. double bond (equality sign, e.g., line #14), triple bond bond (number sign #, not shown); or about implicit aromatization (lower case to upper case) for atoms of carbon, nitrogen, oxygen (depicted); or phosphorus, sulfur (not depicted). Stereochemical indicators about double bonds will be removed (e.g., slashes in lines #18 and #19). Descriptors of stereogenic centers (@-signs, e.g., line #25) and charges (not shown) are copied verbatim.

The work can be illustrated by OpenBabel<sup>3</sup> with instructions to the command line in the pattern of

```
obabel -ismi test_input.smi -O test_input_color.svg -xc10 -xr12 -xl --addinindex
```

to generate a .svg file (vector representation), or

```
obabel -ismi test_input_sat.smi -O test_input_sat_color.png -xc10 -xr12 -xl
↪ --addinindex -xp 3000
```

to generate a bitmap .png with structure formulae depicted in a grid of 10 columns by 12 rows.

It is remarkable how well OpenBabel's displays the molecular structures with advanced motifs. In addition to those shown in the first illustration of this guide, see sub-folder test\_data for a more extensive survey (e.g., the scaffold of cyclophane [entry #33], sparteine [#38], or adamantane [#50]).

program, (c) 2002–2023 by Idorsia Pharmaceuticals Ltd., is freely available under <http://www.openmolecules.org>. For the source code (GPLv3), see <https://github.com/thsa/datawarrior>.

<sup>3</sup>[www.openbabel.org](http://www.openbabel.org) For the most recent documentation, see <https://open-babel.readthedocs.io/en/latest/ReleaseNotes/ob310.html>

## 4 Known peculiarities

The script provides «saturation» by dropping explicit information related to double and triple bonds which SMILES encode (=, # regarding bond order; / (forward slash), \ (backward slash) regarding (*cis*)-(*trans*) relationship around double bonds). While processing double bonds of e.g., ketones to yield secondary alcohols, the script refrains from the assignment of new CIP priorities and a corresponding label. It then depends on the program used for a visualization, if an explicit wedge is used (e.g., OpenBabel), or the absence of information is highlighted (e.g., as question mark in DataWarrior, or the project of CDK depict<sup>4</sup>) as ambiguous. Absolute configuration of stereogenic centers (indicated in SMILES with the @ sign) already assigned in the input however is retained.

For a selection of elements (C, N, O, P, S), the implicit description of aromatic systems (e.g., as c1ccncc1 in pyridine, c1c[nH]cc1 in pyrrol) is recognized. To offer a «saturation», these characters returned as upper case characters to yield e.g., piperidine (C1CCNCC1) and pyrrolidine (C1C[NH]CC1).

The script equally preserves up to one single negative, or single positive charge of these five elements (e.g., [O-]c1ccccc1 about the phenolate anion, and C[N+](c1ccccc1)(C)C about *N,N,N*-trimethylbenzenaminium cation). Here, it can be sensible to «sanitize» the results this script provides by other libraries as e.g. RDKit.<sup>5</sup>

The capitalization of the five characters is constrained to prevent non sensible transformations of e.g., an (implicitly) aromatic atom of tin [sn] into the invalid form [SN]. Though the script is going to write tin as [Sn], an adjustment of valence for elements written with two characters is beyond the current scope of the script.

A SMILES string may describe more than one molecule. Thus, the concatenation with "." (period character) as seen for example in descriptions of co-crystals like about 1,4-benzoquinone and hydroquinone, C1=CC(=O)C=CC1=O.c1cc(ccc1O)O, is retained. The example is resolved as C1CC(O)CCC1O.C1CC(CCC1O)O.

## 5 License

Norwid Behrnd, 2019–23, GPLv3.

---

<sup>4</sup><https://www.simolecule.com/cdkdepict/depict.html> For the mentioned annotation of CIP labels, change No Annotation (second pull down menu from the left) to CIP Stereo Label.

<sup>5</sup>For an overview about the freely available RDKit library, see [www.rdkit.org](http://www.rdkit.org). An introduction into the topic of «molecular sanitization» is provided in the section of this very title in the on-line [RDKit Book](#).