Subject: Improper valencies generated from smiles
Posted by Paul on Wed, 25 Nov 2026 21:25:22 GMT
View Forum Message <> Reply to Message

Hello Thomas,

I recently grabbed a million-structure subset of smiles from GDB-13 at https://gdb.unibe.ch/downloads/ that generates erroneous structures in DataWarrior.

I am using DW v06.07.01 (30-Okt-2026), on both MacOS (27.2.1) M1 Ultra Studio, and on Win 11 Pro 26H2.

I found over-valent atoms in 317 structures of 1 million, mostly oxygen, but also pentavalent carbon atoms.

Typically, a neutral oxygen is three-bonded in aromatic ring systems with other heteroatoms. In one case, a ring contains a carbon sequence of 2 doubly bonded C atoms, each triply bonded to other atoms.

I have used DW to generate its own smiles, then loaded those into DW from a text file, and the idcodes for the structures built from both sets of smiles are identical for the rendered structures (and the same errors). In some cases, the smiles/structures involve aromatic rings with protonated N atoms.

I took the problematic smiles to the cactus-cir site and converted them then opened those with DW, with no change in results.

The problematic smiles are in the attached gdb13_over_valent.txt file. The gdb13_1M.smi file is 23 MB, so I'll not upload that unless you'd like.

Hope this is sufficient information, and, as always, thanks for your work,

Paul

File Attachments

1) gdb13_over_valent.txt, downloaded 1 times