Subject: feature suggestion: .sdf export with explicit hydrogens
Posted by nbehrnd on Tue, 07 Sep 2021 08:10:50 GMT
View Forum Message <> Reply to Message

Dear Thomas,

for the already present export of structures generated by DW as an .sdf file (File -> Save Special -> SD-File), I would like to suggest an extension; namely that these files contain explicit hydrogens.

In the sketcher, the definition of e.g., fluorochlorobromomethane CHBrClF yields either (R), or (S) isomer. The .sdf generated however accounts for four atoms only where five (then including hydrogen) may be anticipated. Irrespective if the structures are flat 2D objects (e.g., result of generating a set of random molecules), a complete set of atom connectivities were helpful for processing the .sdf outside DW in other programs. While Chemistry -> Generate Conformers triggers the generation of a separate .dwar file from which the .sdf exported accounts for the hydrogen previously missing, I would welcome hydrogens included in the .sdf independently of this query.

Norwid

Subject: Re: feature suggestion: .sdf export with explicit hydrogens
Posted by thomas on Thu, 23 Sep 2021 09:19:08 GMT
View Forum Message <> Reply to Message

Dear Norwid,

structures in DataWarrior files are stored as canonical encoded strings (idcodes). Therefore, it doesn't matter, whether a user draws a structure with or without hydrogen atoms. They are immediately converted into the canonical form, which is the one without hydrogen atoms. Tetrahedral stereo configurations are encoded as atom parity, such that when displaying the structure, one of the bonds will be drawn as up or down bond to account for the correct parity. This bond may be different from the one originally used to define the parity. As a consequence, when molecules are exported within SD-files, there are no simple hydrogens, but the stereo information is never lost. In order to display or export molecules with the originally supplied coordinates, DataWarrior stores coordinates of all idcode atoms as a separately encoded string. For 3D-coordinates there is a special handling: idcodes are still the same canonical ones without hydrogen atoms, but the coordinate encoding contains 3D-coordinates even for implicit hydrogen atoms, because for conformers hydrogen positions may be important, e.g. for docking. Therefore, SD-Files, if exported with 3D-coordinates, contain hydrogen atoms.

If I provided a setting to also write hydrogen atoms into exported 2D-SD-Files, then I would need to generate new potentially sub-optimal coordinates for the added formerly implicit hydrogen atoms. For 2D-SD-Files it seems common practice to not include implicit hydrogen atoms. Which application do you have in mind that requires hydrogen atoms?

Thomas

Subject: Re: feature suggestion: .sdf export with explicit hydrogens
Posted by nbehrnd on Sat, 25 Sep 2021 16:01:32 GMT
View Forum Message <> Reply to Message

Dear Thomas,

based on recent work with DataWarrior, I retract this feature suggest.

DataWarrior's default assignment of an idcode about a structure is a simplified, yet complete description of a molecule in terms of atom connectivity, as well as stereo chemistry. Thus -- already at the level where libraries of molecules are generated -- DW is able to assign e.g., a a full InChI string, despite at this stage, exported .sdf files lack lines with an explicit label "hydrogen". These .sdf aim for the exchange of information with other programs; including their own (independent from DW) assignment of full InChI strings.

The (re)generation of 3D structures by the conformer generator departing from such a "library .sdf" serves a different purpose. As you mentioned, these subsequent structures are the ones to be used in docking experiments. And these are the .sdf aiming for a visualization of the molecule's shape/geometry in a viewer like Jmol with explicit hydrogen atoms.

As noticed, the set of InChI strings assigned to molecules at level of library generation, and the one about structures past the conformer generation, may differ from each other. This however is because the conformer generator resolves ambiguity if an entry in a "library stage .sdf" did not describe a stereogenic center in either (R), or (S); an axis in (P), or (M); a double bond in (E), or (Z) configuration. The conformer generator's permutation of these choices may be caused by lack of this particular information, i.e. an unknown stereo configuration, or because the .sdf read intentionally describes a racemate (the "enhanced stereo recognition" the structure editor's documentation describes).

Norwid