Subject: Visualising SMILES string
Posted by amorrison on Thu, 18 Mar 2021 09:55:31 GMT
View Forum Message <> Reply to Message

Hi Thomas,

I have a smiles string with an attachment point that has been created in another application, e.g. C1CCC([*:1])CC1. Datawarrior fails to identify it as a valid smiles either on paste or convert structure to name. I'm assuming it's the '*:1' that is the issue. Can I modify the smiles so it can be interpreted by Datawarrior?

many thanks in advance,

Angus

Subject: Re: Visualising SMILES string
Posted by thomas on Sat, 20 Mar 2021 14:07:46 GMT
View Forum Message <> Reply to Message

Hi Angus,

if you replace [*:1] by *, then the atom is interpreted as wildcard query feature, which is drawn as a '?'.
In this case the new molecule is considered as sub-structure rather than as normal structure. The difference is
that a sub-structure may have query features and all unspecified atom valences are not implicitly considered
to be blocked by hydrogen atoms. Therefore, a single bonded oxygen is shown as -O in contrast to -OH in a normal molecule.

DataWarrior makes this distinction, which is not made by most other software, possibly, because the SMILES definition itself does not distinguish between these two structure flavors.

Thomas

Subject: Re: Visualising SMILES string
Posted by amorrison on Sun, 21 Mar 2021 04:00:31 GMT
View Forum Message <> Reply to Message

Thanks Thomas,

I've tried this. Doing a find and replace my smiles string becomes - C1CCC([*])CC1. If try a structure to name this is still not recognised as a valid smiles. If I remove the square bracket - C1CCC(*)CC1. I get the structure but lose the attachment point.

Sorry if this trivial I'm a med chemist and not so familiar with smiles syntax.

Thanks,

Angus

Subject: Re: Visualising SMILES string
Posted by thomas on Tue, 23 Mar 2021 15:38:22 GMT
View Forum Message <> Reply to Message

Dear Angus,

I have extended the SmilesParser within OpenChemLib and used by DataWarrior to support '*' and '?' as pseudo atom symbols. Both symbols can be used inside or outside square brackets. While '*' is inline with the opensmiles.org standard, '?' is not. Nevertheless, because of the missing distinction in the smiles syntax between query fragments and full molecules, I believe that '?' makes sense.

'*' creates a wild card atom, which is an atom query feature and only allowed in query fragments. Therefore, a smiles containing a '*' is atomatically perceived as a query structure rather than a full molecule. Its free valences are not considered as being filled with hydrogen.

'?' is converted into an atomicNo=0, which DataWarrior uses as attachment point, e.g. in R-groups after a SAR deconvolution. These atoms are meant to not exist. Their sole purpose is to carry the bond sticking out of the R-group.

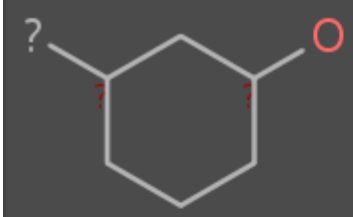
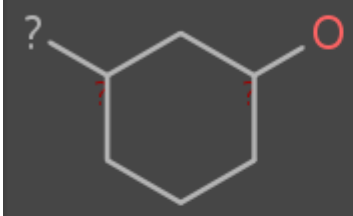
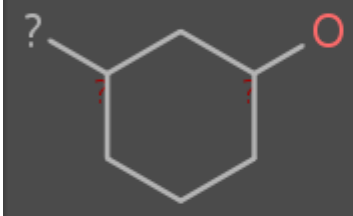
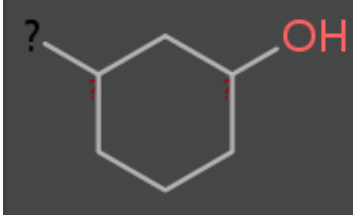
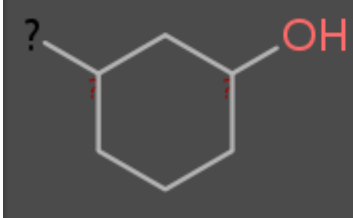
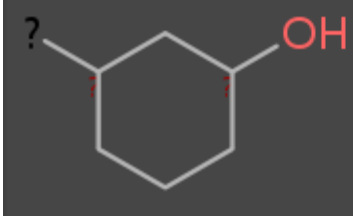
If you paste the these 6 rows of smiles:

C1CCC([*:1])CC1O
C1CCC([*])CC1O
C1CCC(*)CC1O
C1CCC([?:1])CC1O
C1CCC([?])CC1O
C1CCC(?)CC1O

into a new DataWarrior Window (newest dev release), then you get the following table:
Note that I have added an oxygen atom. If you see an 'H' at the oxygen, the structure is considered a molecule. If not, the structure is a query fragment with open valences.

File Attachments

1) t.png, downloaded 1339 times

Structure of Column 1	Column 1
	C1CCC([*:1])CC1O
	C1CCC([*])CC1O
	C1CCC(*)CC1O
	C1CCC([?:1])CC1O
	C1CCC([?])CC1O
	C1CCC(?)CC1O

Subject: Re: Visualising SMILES string
Posted by amorrison on Fri, 26 Mar 2021 08:31:47 GMT
View Forum Message <> Reply to Message

Thanks Thomas, works great!
Angus