chem-bla-ics

PubChem-CDK

Egon Willighagen

Published May 11, 2009

Citation
Willighagen, E. (2009). PubChem-CDK. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/8bpsk-rb857

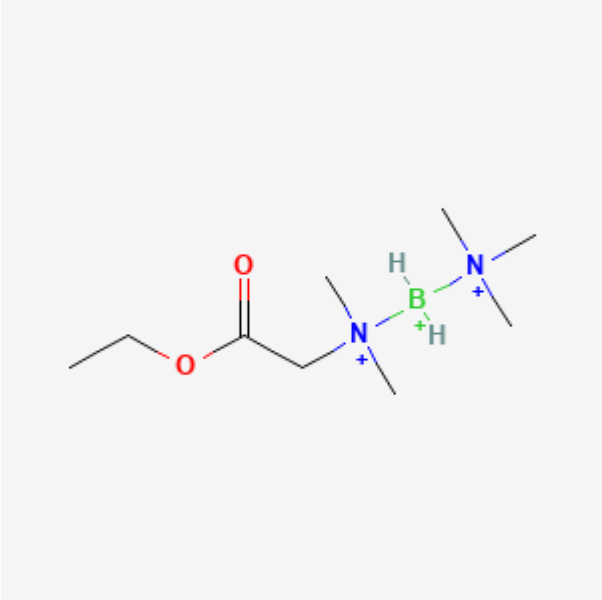
Keywords
Cdk, Pubchem

chem-bla-ics

PubChem-CDK is a project that runs CDK code on the PubChem data. As we speak, a groovy script reads about 100 PubChem Compounds XML entries per second into the database. Mind you, not the SDF they distribute which uses a custom extension to overcome the limits of the real MDL SDF format.

Right now, it has run the atom type perception algorithm on about 1M compounds, and has a pretty good coverage of the *organic chemistry* domain. I will analyze the results statistically soon, but will likely use this data first to add some missing atom types to CDK 1.2.x. BTW, did you know only ***three*** carbon atoms failed? A C^{4-} (CID:156031), a C^{3+} (CID:161072), and a C^{2+} (CID:161073). Would your cheminformatics library know what their properties are?

It is really nice way of browsing PubChem, BTW. For example, did you know there are several boron compounds which have a substructure [N+]-[B+]-[N+]? Yes, three positive charges, *next* to each other? For example (CID:3612285):



Well, neither did I. How was it synthesised? What are the spectral properties? How do they stabilise it? What magic counter ion? PubChem, unfortunately, does not have links to primary literature, and there is no free source for that available. A failure in chemistry. The source points to ChemDB, but the entry in that database does not shed light on this either.

Anyway, more on this later. Much more, as I plan to run many CDK algorithms on this code.