chem-bla-ics

Processing the ChEBI MDL SD file with the CDK

Egon Willighagen

Published October 1, 2009

Citation
Willighagen, E. (2009). Processing the ChEBI MDL SD file with the CDK. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/2zprm-hs481

Keywords
Cdk, Groovy, Chebi

Bioclipse has a bug report about browsing the ChEBI SD file in its moltable editor. Some entries make Bioclipse crash (as reported), or just very sluggish as with my Dell superlapcomputer :)

So, I processed the file with a pure CDK 1.2.3 with this small piece of Groovy script:

```
import org.openscience.cdk.interfaces.*;
import org.openscience.cdk.io.*;
import org.openscience.cdk.io.iterator.*;
import org.openscience.cdk.*;
import org.openscience.cdk.tools.manipulator.*;

iterator = new IteratingMDLReader(
  new File("ChEBI_complete.sdf").newReader(),
  DefaultChemObjectBuilder.getInstance()
)
int i = 0;
boolean hasNext = true;
while (hasNext) {
  i++;
  long startTime = System.currentTimeMillis();
  hasNext = iterator.hasNext();
  IMolecule mol = iterator.next()
  long endTime = System.currentTimeMillis();
  formula = MolecularFormulaManipulator.getMolecularFormula(mol)
  long time = endTime - startTime;
  if (time > 99)
    println i + ": " + MolecularFormulaManipulator.getString(formula) +
            " (" + endTime + "-" + startTime + "=" + time + " ms)"

}
```

This script times reading of all entries and reports all that entries take more than 100 ms to read (in the scripting environment). There are surprising results: H2O takes 50 seconds, phosphate 100 seconds. So, I am quite certain it must be the reading of the metadata, and not the connection table. But, this I will explore in more detail now, hoping to come up with a patch for the CDK to speed up reading of such entries.

The full list of timings:

```
1: C10H2 (1254375053450-1254375052356=1094 ms)
152: C20HN7O6 (1254375054779-1254375054125=654 ms)
592: C3HO3 (1254375056604-1254375055499=1105 ms)
832: C9NO5 (1254375057016-1254375056823=193 ms)
879: C20N4 (1254375057381-1254375057039=342 ms)
1125: R (1254375058293-1254375057528=765 ms)
```

chem-bla-ics

```
1197: C20N7O6 (1254375058612-1254375058372=240 ms)
1198: C5NO3 (1254375058714-1254375058613=101 ms)
1243: C5NO4 (1254375063698-1254375058800=4898 ms)
1272: C21N7O16P3S (1254375067185-1254375063856=3329 ms)
1277: C23N7O17P3S (1254375067625-1254375067239=386 ms)
1282: C3NO2S (1254375070673-1254375067650=3023 ms)
1285: C3O3 (1254375071600-1254375070675=925 ms)
1290: C2O2 (1254375071802-1254375071608=194 ms)
1299: H2O (1254375122202-1254375071808=50394 ms)
1300: H (1254375136668-1254375122202=14466 ms)
1301: O2 (1254375145270-1254375136670=8600 ms)
1335: C15N6O5S (1254375150683-1254375145319=5364 ms)
1343: C10N5O13P3 (1254375298927-1254375150686=148241 ms)
1349: C2NO2 (1254375301391-1254375298953=2438 ms)
1351: C34N4O4 (1254375301659-1254375301396=263 ms)
1509: C6NO2 (1254375302753-1254375302011=742 ms)
1541: C19N7O6 (1254375303296-1254375302778=518 ms)
1543: C20HN7O7 (1254375303441-1254375303312=129 ms)
1609: C9N2O15P3 (1254375303740-1254375303558=182 ms)
1631: CHO2 (1254375303975-1254375303837=138 ms)
1632: C4O4 (1254375304127-1254375303976=151 ms)
1711: C21N7O14P2 (1254375310174-1254375304245=5929 ms)
1788: C6H3O9P (1254375310555-1254375310387=168 ms)
1798: C10H2N2O3S (1254375310705-1254375310588=117 ms)
1808: C6N3O2 (1254375312665-1254375310727=1938 ms)
1823: C10N5O14P3 (1254375318534-1254375312781=5753 ms)
1839: C5NO4 (1254375325988-1254375318583=7405 ms)
1840: C3O3 (1254375326249-1254375325989=260 ms)
1848: C10N5O7P (1254375336661-1254375326273=10388 ms)
1862: C5NOSR2 (1254375337336-1254375336699=637 ms)
1882: C3N2 (1254375337489-1254375337351=138 ms)
1893: C6O9P (1254375337626-1254375337501=125 ms)
1910: C3O6P (1254375337846-1254375337639=207 ms)
1934: H3N (1254375349713-1254375337921=11792 ms)
1977: O4S (1254375350045-1254375349902=143 ms)
1984: CN2O (1254375350174-1254375350050=124 ms)
2007: C5N (1254375350324-1254375350183=141 ms)
2015: C5N5O (1254375350493-1254375350329=164 ms)
2016: C2O (1254375350683-1254375350494=189 ms)
2018: C27N9O15P2 (1254375351927-1254375350684=1243 ms)
2020: H2O2 (1254375352124-1254375351928=196 ms)
2036: C17N3O17P2 (1254375352309-1254375352196=113 ms)
2095: C10N5O4 (1254375352578-1254375352394=184 ms)
2137: C14HO4R (1254375353331-1254375352646=685 ms)
```

chem-bla-ics

```
2180: C3NO2 (1254375354199-1254375353469=730 ms)
2184: C9NO5 (1254375354480-1254375354270=210 ms)
2194: C6N4O2 (1254375356738-1254375354485=2253 ms)
2201: C21N7O17P3 (1254375359838-1254375356748=3090 ms)
2228: C6O2 (1254375360480-1254375359912=568 ms)
2240: CO2 (1254375363324-1254375360485=2839 ms)
2327: C5NO2S (1254375370536-1254375363612=6924 ms)
2348: C14N6O5S (1254375371522-1254375370558=964 ms)
2359: C9N2O9P (1254375372236-1254375371544=692 ms)
2367: C9N2O6 (1254375373614-1254375372265=1349 ms)
2370: C5N5 (1254375373975-1254375373615=360 ms)
2404: C10N5O5 (1254375374360-1254375374108=252 ms)
2413: C10N5O10P2 (1254375401639-1254375374373=27266 ms)
2454: C5O5 (1254375401831-1254375401688=143 ms)
2455: C5NO2S (1254375407807-1254375401832=5975 ms)
2470: C11N2O2 (1254375408251-1254375407815=436 ms)
2494: C4NO3 (1254375409200-1254375408373=827 ms)
2499: C5O10P2R (1254375412153-1254375409297=2856 ms)
2525: C4HO7P (1254375412777-1254375412293=484 ms)
2526: C4N2 (1254375414071-1254375412777=1294 ms)
2534: C21N7O14P2 (1254375417657-1254375414091=3566 ms)
2581: C3NO2 (1254375422072-1254375417745=4327 ms)
2638: C4NO4 (1254375424772-1254375422244=2528 ms)
2680: C5O14P3 (1254375426347-1254375424831=1516 ms)
2683: C3NO3 (1254375433063-1254375426353=6710 ms)
2702: C3HO6P (1254375433192-1254375433079=113 ms)
2749: C4N2O3 (1254375434106-1254375433445=661 ms)
2755: C10N4O8P (1254375434417-1254375434113=304 ms)
2756: C5NO2 (1254375436750-1254375434418=2332 ms)
2779: C4NO2S (1254375437847-1254375436759=1088 ms)
2803: C5N4 (1254375438991-1254375437968=1023 ms)
2832: C9NO2 (1254375439226-1254375439026=200 ms)
2844: C8HNO3 (1254375440463-1254375439238=1225 ms)
2856: C5O13P3R (1254375441336-1254375440497=839 ms)
2863: C10O6 (1254375442424-1254375441348=1076 ms)
2873: C10N5O8P (1254375442560-1254375442433=127 ms)
2898: C3HO3 (1254375443712-1254375442655=1057 ms)
2925: C8H4NO6 (1254375443886-1254375443729=157 ms)
3025: CO3 (1254375444508-1254375444131=377 ms)
3031: C10N5O11P2 (1254375444810-1254375444601=209 ms)
3038: C3NO2S (1254375449012-1254375444836=4176 ms)
3042: C4N2O2 (1254375449224-1254375449066=158 ms)
3060: C6O2 (1254375449433-1254375449274=159 ms)
3083: C17N4O9P (1254375450751-1254375449465=1286 ms)
```

chem-bla-ics

```
3088: C34FeN4O4 (1254375452873-1254375450848=2025 ms)
3111: C9N2O12P2 (1254375454560-1254375452939=1621 ms)
3119: CNO5P (1254375454774-1254375454563=211 ms)
3122: C9N3O14P3 (1254375454972-1254375454778=194 ms)
3184: C3O3 (1254375455362-1254375455053=309 ms)
3213: CO (1254375455489-1254375455375=114 ms)
3216: C3HO7P (1254375455662-1254375455490=172 ms)
3223: C9N2O6 (1254375455850-1254375455737=113 ms)
3239: C3NO3 (1254375458116-1254375455868=2248 ms)
3296: C9NO3 (1254375459575-1254375458250=1325 ms)
3306: S3R (1254375464014-1254375459596=4418 ms)
3313: C6O6 (1254375464701-1254375464016=685 ms)
3348: C14HO4 (1254375465102-1254375464766=336 ms)
3360: C12O11 (1254375465830-1254375465193=637 ms)
3364: N2 (1254375475917-1254375465872=10045 ms)
3371: C21N7O17P3 (1254375479243-1254375475920=3323 ms)
3377: C6N2O2 (1254375481175-1254375479306=1869 ms)
3379: C3O6P (1254375482278-1254375481176=1102 ms)
3390: O10P3 (1254375484356-1254375482286=2070 ms)
3403: C5N2O3 (1254375486975-1254375484451=2524 ms)
3499: C9NO3 (1254375487745-1254375487074=671 ms)
3502: C5O8P (1254375489044-1254375487747=1297 ms)
3532: C55MgN4O5 (1254375489206-1254375489097=109 ms)
3537: C5NO4 (1254375494872-1254375489209=5663 ms)
3546: Fe (1254375507646-1254375494892=12754 ms)
3554: C5N2O2 (1254375507934-1254375507650=284 ms)
3566: H2 (1254375508526-1254375508033=493 ms)
3576: C12ClN4O7P2S (1254375508737-1254375508548=189 ms)
3577: Mn (1254375511113-1254375508738=2375 ms)
3582: C11NO6P (1254375511249-1254375511120=129 ms)
3628: O7P2 (1254375554180-1254375511388=42792 ms)
3633: O4P (1254375659461-1254375554183=105278 ms)
3647: C12N4OS (1254375659706-1254375659481=225 ms)
3664: C8HNO6P (1254375661230-1254375659713=1517 ms)
3665: C9N4O8P (1254375661450-1254375661231=219 ms)
3679: Mg (1254375679426-1254375661513=17913 ms)
3859: C20N7O6 (1254375679768-1254375679522=246 ms)
3860: C19N7O6 (1254375680069-1254375679769=300 ms)
4026: Ca (1254375681849-1254375680582=1267 ms)
4029: CNOR (1254375682983-1254375681850=1133 ms)
4031: COR2 (1254375686384-1254375682984=3400 ms)
4038: Cl (1254375686610-1254375686387=223 ms)
4099: F (1254375687012-1254375686767=245 ms)
4138: H (1254375722496-1254375687100=35396 ms)
```

chem-bla-ics

```
4163: C6NO2 (1254375722805-1254375722566=239 ms)
4166: C6N2O2 (1254375724837-1254375722807=2030 ms)
4167: Mg (1254375746423-1254375724838=21585 ms)
4229: O (1254375754305-1254375746586=7719 ms)
4254: H3O4P (1254375771602-1254375754367=17235 ms)
4263: K (1254375771850-1254375771608=242 ms)
4265: C5NO2 (1254375772195-1254375771852=343 ms)
4297: Na (1254375772801-1254375772310=491 ms)
4311: C4O3R (1254375773107-1254375772835=272 ms)
4313: S (1254375795116-1254375773109=22007 ms)
4356: Zn (1254375814849-1254375795263=19586 ms)
4424: C5O5 (1254375818351-1254375814892=3459 ms)
4453: C2 (1254375818489-1254375818369=120 ms)
4482: C6N3O2 (1254375819699-1254375818525=1174 ms)
4494: C (1254375821009-1254375819706=1303 ms)
4519: Co (1254375821358-1254375821068=290 ms)
4670: C11N2O2 (1254375821817-1254375821583=234 ms)
4677: C4H2O4 (1254375822301-1254375821824=477 ms)
4801: Ni (1254375822605-1254375822450=155 ms)
4912: C5N2O3 (1254375823778-1254375822655=1123 ms)
5060: C5O5 (1254375824119-1254375823908=211 ms)
5111: C6O6 (1254375824420-1254375824212=208 ms)
5143: Cu (1254375824613-1254375824502=111 ms)
5357: C6N4O2 (1254375826277-1254375824919=1358 ms)
5368: C9NO5 (1254375826504-1254375826289=215 ms)
5369: Fe (1254375826620-1254375826505=115 ms)
5380: C3HO6P (1254375827340-1254375826635=705 ms)
5398: Na (1254375827949-1254375827359=590 ms)
5400: K (1254375828174-1254375827951=223 ms)
5402: Zn (1254375844116-1254375828175=15941 ms)
5404: Ca (1254375845806-1254375844117=1689 ms)
5438: HO (1254375846125-1254375845836=289 ms)
5538: CH3 (1254375847891-1254375846233=1658 ms)
5548: Ca (1254375861972-1254375847928=14044 ms)
5560: H2N (1254375866398-1254375861980=4418 ms)
5693: C10O6 (1254375867526-1254375866499=1027 ms)
5814: O7P2 (1254375910579-1254375867608=42971 ms)
5869: C2NOR2 (1254375914124-1254375910633=3491 ms)
5871: C3NOSR2 (1254375914574-1254375914161=413 ms)
5873: C6N4OR2 (1254375916764-1254375914575=2189 ms)
5875: C11N2OR2 (1254375917143-1254375916766=377 ms)
5877: C4NO3R2 (1254375917710-1254375917145=565 ms)
5885: C6N2OR2 (1254375919573-1254375917716=1857 ms)
5889: C5NO3R2 (1254375920901-1254375919576=1325 ms)
```

chem-bla-ics

```
5895: C6N3OR2 (1254375922306-1254375920904=1402 ms)
5900: C5NO4 (1254375925689-1254375922310=3379 ms)
5902: C5NO4 (1254375930626-1254375925693=4933 ms)
5903: C5NO4 (1254375933920-1254375930626=3294 ms)
5906: C4NO4 (1254375934593-1254375933924=669 ms)
5907: C4NO4 (1254375935274-1254375934594=680 ms)
5909: C4NO4 (1254375936451-1254375935274=1177 ms)
5911: C9NOR2 (1254375936575-1254375936453=122 ms)
5913: C3NO2R2 (1254375940197-1254375936577=3620 ms)
5914: C3NOSeR2 (1254375940307-1254375940197=110 ms)
5920: C6NOR2 (1254375940705-1254375940311=394 ms)
5925: C5N2O2R2 (1254375942662-1254375940737=1925 ms)
5926: C4NO2R2 (1254375943012-1254375942662=350 ms)
5939: C4O4 (1254375943199-1254375943061=138 ms)
5993: C2O2 (1254375943413-1254375943287=126 ms)
6082: Zn (1254375958993-1254375943449=15544 ms)
6453: C10N5O13P3 (1254376116554-1254375959193=157361 ms)
6574: CHO2 (1254376116903-1254376116743=160 ms)
6706: C5O5 (1254376117248-1254376117131=117 ms)
7032: C3O4 (1254376117689-1254376117435=254 ms)
7104: C5NO3R2 (1254376118481-1254376117843=638 ms)
7252: C5NOSR2 (1254376118731-1254376118630=101 ms)
7411: C3O3 (1254376120056-1254376119011=1045 ms)
7465: CR (1254376121752-1254376120212=1540 ms)
7627: CNO (1254376122089-1254376121887=202 ms)
7741: C12ClN4OS (1254376122320-1254376122156=164 ms)
7858: CO2R (1254376122547-1254376122436=111 ms)
7891: Fe4S4 (1254376122844-1254376122585=259 ms)
8178: Mn (1254376124399-1254376122960=1439 ms)
8338: C4NO2 (1254376124643-1254376124480=163 ms)
9219: C4NO6P (1254376127494-1254376124951=2543 ms)
9234: C9N3O14P3 (1254376127605-1254376127498=107 ms)
9235: C10N5O14P3 (1254376132596-1254376127605=4991 ms)
9305: C3NO6P (1254376143823-1254376132629=11194 ms)
9311: C6O6 (1254376144017-1254376143825=192 ms)
9402: C5NOSR (1254376144332-1254376144214=118 ms)
9427: C4O2R2 (1254376144645-1254376144419=226 ms)
10281: O10P3 (1254376146646-1254376144942=1704 ms)
10308: C2OR (1254376148204-1254376146659=1545 ms)
10453: CHOR (1254376148682-1254376148321=361 ms)
10506: HOR (1254376149933-1254376148793=1140 ms)
10589: C21N7O17P3 (1254376150485-1254376150182=303 ms)
10602: C5N2O2R (1254376150727-1254376150503=224 ms)
10604: C5N2OR2 (1254376150946-1254376150728=218 ms)
```

chem-bla-ics

```
10614: C5N2O2 (1254376151170-1254376150949=221 ms)
10617: C5N2OR (1254376151389-1254376151171=218 ms)
10641: C3O6P (1254376152229-1254376151404=825 ms)
10656: C16OR (1254376152505-1254376152235=270 ms)
10688: C5O5 (1254376155565-1254376152582=2983 ms)
10690: C3NO5PR2 (1254376166856-1254376155566=11290 ms)
10729: C12N4O7P2S (1254376167090-1254376166889=201 ms)
10748: C3NOR2 (1254376171309-1254376167097=4212 ms)
10756: C3O4 (1254376172041-1254376171320=721 ms)
10760: C9N2O15P3 (1254376172162-1254376172045=117 ms)
10786: OR (1254376172419-1254376172169=250 ms)
10828: C2NO2R (1254376173256-1254376172429=827 ms)
10830: C2NOR (1254376174784-1254376173258=1526 ms)
10883: C9NO2R2 (1254376175110-1254376174827=283 ms)
10899: NR (1254376202420-1254376175114=27306 ms)
10902: C2O (1254376203938-1254376202454=1484 ms)
10914: C9NO5 (1254376204203-1254376203942=261 ms)
11203: C6O6 (1254376204716-1254376204522=194 ms)
11226: C4HO7P (1254376205190-1254376204732=458 ms)
11680: C5NO2SR (1254376205649-1254376205392=257 ms)
11681: C5NOSR (1254376205900-1254376205650=250 ms)
11916: H (1254376206483-1254376206109=374 ms)
12216: C5NOR2 (1254376208374-1254376206750=1624 ms)
12217: C4N2O2R2 (1254376209040-1254376208375=665 ms)
12680: COSR2 (1254376209601-1254376209314=287 ms)
13478: C25HN2O19 (1254376210152-1254376209951=201 ms)
13662: C5O8P (1254376210482-1254376210379=103 ms)
```