

# Clashes and Conflicts

A status report on CDK structure diagram generation (SDG).

by Christoph Steinbeck

Recently, a new 2D layout algorithm for molecular graphs was published in the Journal of Chemical Information and Modeling [1]. The new method by Clark et al., who work for the Chemical Computing Group (CCG) is implemented in CCG's Scientific Vector Language (SVL) and used in the Molecular Operating Environment (MOE) [2].

The MOE algorithm partitions the molecular graph in a number of steps:

- Hydrogen atoms are ignored until the final stages of the algorithm.
- Terminal atoms are ignored until the final stages, because their coordinates are implied by their more connected neighbor.
- Ring blocks are detected and their rings grouped together.
- Stereochemically active double bonds are identified and restricted appropriately.
- Sequences of certain atoms are treated as chains.
- The remaining atoms are paired or isolated.

After this partitioning step, each block is assigned a set of geometric constraints and afterwards a search is performed to obtain a globally optimal solution that satisfy these constraints. The authors claim that in contrast to sequential layout algorithms, their strategy avoids globally undesired outcomes.

The art of automatically generating aesthetically pleasing, publication quality 2D depictions of molecular structures is called Structure Diagram Generation (SDG). In one or the other form, it is part of almost any cheminformatics program or package. The state of technology in this field has been captured by Harald Helson from CambridgeSoft, Massachusetts, in 1999 [3]. I take the publication of the new MOE methods as a chance to summarize the status of the (SDG) algorithm implemented in the Chemistry Development Kit and to compare its performance to that of the MOE SDG. For the sake of simplicity, I will just distinguish between MOE-SDG and CDK-SDG in the following.

It is clear that the MOE-SDG performs well on a wide variety of molecules. Molecules which cause problems with the CDK-SDG can be as simple as hexaisopropoxybenzene (Figure 1) or be as large as buckminster fullerene (C60). For both cases, the MOE-SDG generates nice 2D diagrams.

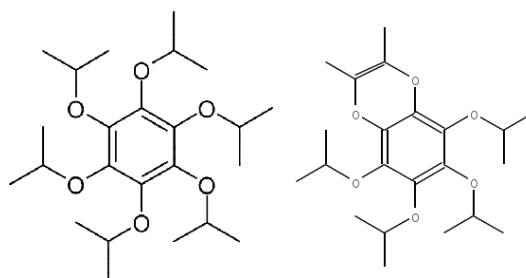


Figure 1: Layout of hexaisopropoxybenzene [4] by MOE-SDG (left) and CDK-SDG (right). It is a nice example of a simple molecule where only one of 64 possible combinations of local decisions on substituent placement is the optimal solution.

In most cases, the basic algorithm sketched above will do the job. In cases where certain ring systems impose problems, such as in the case of adamantane, a ring template library is used. In even worse cases, like C60, a 3D embedding algorithm is employed, where a first guess in 3D is obtained using distance geometry, followed by a truncated newton minimization of bond and angle deviations. The final projection into the plane is then done while minimizing atom overlap and placing potential substituents in the periphery.

The CDK-SDG uses a much simpler, sequential layout algorithm. It performs a partitioning of the initial structure into ring systems and aliphatic chains. In the case of ring systems, it distinguishes bridged, fused and spiro connections between rings, depending on whether more than two, two or just one atom is shared by two connected rings. The CDK-SDG also uses templates for ring systems, but the size of the template library is limited at the moment. Layout is first performed on the largest ring system and then the attached aliphatics are laid out. Then the next ring system is laid out, attached, and so on. In the end, an overlap resolver checks if there are any clashes and moves atoms a bit, if necessary. While the use of the template handler can be switched on or off, the use of the overlap resolver can not be disabled.

For testing the performance of the CDK-SDG, Alex Clark kindly provided me with a list of SMILES of notorious cases. A number of them are shown in Figure 2.

My first tests with this collection caused me to increase the amount by which atoms are moved in overlap cases, which lead to the output shown in Figure 2.

An earlier test with the CDK-SDG-based PDF-Generator by Rajarshi Guha (<http://cheminformatics.indiana.edu/~rguha/code/java/>) had shown that the CDK-SDG tackles a large variety of layout cases quite nicely. The cases collected



- [7] InChI=1/C26H50N4O2/c1-13-27(17(5)6)21-22(28(14-2)18(7)8)26(32)24(30(16-4)20(11)12)23(25(21)31)29(15-3)19(9)10/h17-20,31-32H,13-16H2,1-12H3.
- [8] InChI=1/C18H10F5P/c19-11-5-1-2-10-16(11)24(17-12(20)6-3-7-13(17)21)18-14(22)8-4-9-15(18)23/h1-10H.
- [9] InChI=1/C22H18O8/c23-19(24)17(20(25)26)9-15-11-5-1-2-6-12(11)16(10-18(21(27)28)22(29)30)14-8-4-3-7-13(14)15/h1-8,17-18H,9-10H2,(H,23,24)(H,25,26)(H,27,28)(H,29,30).
- [10] InChI=1/C17H14Cl2N2O2/c18-11-5-6-15-13(9-11)17(12-3-1-2-4-14(12)19)21(7-8-23-17)10-16(22)20-15/h1-6,9H,7-8,10H2,(H,20,22).
- [11] InChI=1/C32H34Br2O3/c1-31-13-12-25-22(26(31)14-20(17-35)30(31)37)11-10-21-15-23(29-27(33)8-5-9-28(29)34)24(16-32(21,25)18-36)19-6-3-2-4-7-19/h2-10,17,20,22-26,36H,11-16,18H2,1H3.
- [12] InChI=1/C20H32O5.ClH/c1-7-17(4)11-12(21)20(24)18(5)10-8-9-16(2,3)14(18)13(22)15(23)19(20,6)25-17;/h7,13-15,22-24H,1,8-11H2,2-6H3;1H.
- [13] InChI=1/C16H19NO4/c1-20-16(19)14-12-8-7-11(17-12)9-13(14)21-15(18)10-5-3-2-4-6-10/h2-6,11-14,17H,7-9H2,1H3.
- [14] InChI=1/C13H10Cl4O4/c1-20-13(21-2)11(16)7-5(18)3-4-6(19)8(7)12(13,17)10(15)9(11)14/h3-4,7-8H,1-2H3.
- [15] InChI=1/C10H15NO2S/c1-9(2)7-3-4-10(9)6-14(12,13)11-8(10)5-7/h7H,3-6H2,1-2H3/q+2.
- [16] InChI=1/C7H12O5/c1-10-7-5(9)6-4(8)3(12-7)2-11-6/h3-9H,2H2,1H3.
- [17] InChI=1/C21H21NO6/c1-22-7-6-11-8-15(25-2)16(26-3)9-13(11)21(22)19(23)12-4-5-14-18(28-10-27-14)17(12)20(21)24/h4-5,8-9,20,24H,6-7,10H2,1-3H3.
- [18] InChI=1/C26H38N2O4/c1-3-7-25(8-4-1)23-27-11-15-29-19-21-31-17-13-28(24-26-9-5-2-6-10-26)14-18-32-22-20-30-16-12-27/h1-10H,11-24H2.
- [19] J.J. Irwin and B.K. Shoichet. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1):177-82, 2005.

**Web page:**

<http://cdknews.org/>

**Editors-in-Chief:**

Egon Willighagen [egonw@users.sf.net](mailto:egonw@users.sf.net) and  
Christoph Steinbeck [steinbeck@users.sf.net](mailto:steinbeck@users.sf.net)

**Editorial Board:**

Andreas Bender, Christoph Steinbeck, Egon Willighagen, Rajarshi Guha, Rich Apodaca and Uli Fechner.

CDK News is a publication of the Chemistry Development Kit (CDK) project. All articles are copyrighted with GNU's FDL by the respective authors. Submissions can be submitted via the web page.

**CDK Project web pages:**

<http://cdk.sourceforge.net/>

<http://www.chemistry-development-kit.org/>

<http://cdk.sf.net/wiki/>