

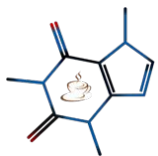
Cheminformatics for Natural Products with the CDK

Maria Sorokina, Aziz M. Yirik, Jonas Schaub, Christoph Steinbeck

Friedrich-Schiller University Jena, Germany



About us



Cheminformatics and Computational Metabolomics

Friedrich-Schiller-University, Jena, Germany

<https://cheminf.uni-jena.de>



**FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA**



Dr. Maria Sorokina

Research interests:
NP cheminformatics, databases,
enzyme promiscuity and metabolic
networks



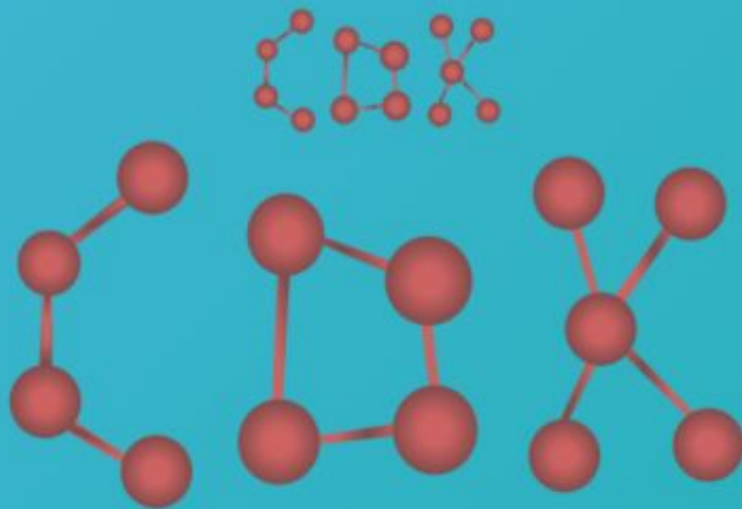
Mehmet Aziz Yirik

Research Interests:
Chemical Graphs, Graph
Isomorphism, Graph Generation



Jonas Schaub

Research interests:
NP and carbohydrate
cheminformatics, molecular
fragmentation algorithms



Chemistry Development Kit

Open Source modular Java libraries for Cheminformatics

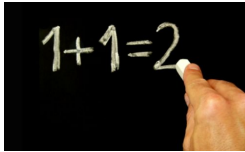
<http://cdk.github.io>

Java - a modern object-oriented language



- Java was created in 1995 for Sun Microsystems
- It is a **platform independent** language following the logic "Write Once, Run Everywhere"

Simple



Portable



Object-Oriented

{OOP}

Secure



Distributed



Robust



High Performance



Modern



CDK: Chemistry Development Kit

The Chemistry Development Kit (CDK) is a collection of modular Java libraries for processing chemical information (Cheminformatics). The modules are free and open-source and are easy to integrate with other open-source or in-house projects.

Willighagen et al. *J Cheminform* (2017) 9:33
DOI 10.1186/s13321-017-0220-4

 Journal of Cheminformatics

SOFTWARE

Open Access

The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching



Egon L. Willighagen^{1*}, John W. Mayfield², Jonathan Alvarsson³, Arvid Berg³, Lars Carlsson⁴,
Nina Jeliazkova⁵, Stefan Kuhn⁶, Tomáš Pluskal⁷, Miquel Rojas-Chertó⁸, Ola Spjuth³,
Gilleain Torrance⁹, Chris T. Evelo¹, Rajarshi Guha¹⁰ and Christoph Steinbeck¹¹



> 100 Contributors

Key people: Christoph Steinbeck,
Egon Willighagen, Dan Gezelter,
Rajarshi Guha, John Mayfield

Get Started with CDK!


- check the general documentation (Java API): it's very detailed <http://cdk.github.io/cdk/latest/docs/api/index.html>
- join the CDK mailing list <https://sourceforge.net/projects/cdk/lists/cdk-user>
- download the latest release JAR from GitHub
- OR use Maven or Gradle dependencies, which will automatically fetch the CDK modules


```
<dependency>
  <groupId>org.openscience.cdk</groupId>
  <artifactId>cdk-bundle</artifactId>
  <version>2.3</version>
</dependency>
```




Download

<> [Java API - JavaDoc](#)

 [Groovy Cheminformatics with the CDK - Book](#)

 [Chemistry Toolkit Rosetta Wiki - Code examples](#)

 [Mailing list archives](#)

CDK in action

Webservices:

- **CDK Depict:** <https://www.simolecule.com/cdkdepict/depict.html>
- **NP-likeness score, Sugar Removal Utility, NP databases:** <https://naturalproducts.net/>

Workflows:

- **MetFrag:** <https://msbi.ipb-halle.de/MetFrag/>
- CDK KNIME nodes <https://doi.org/10.1186/1471-2105-14-257>

Stand-alone tools:

- **CDK Descriptor GUI:** <http://www.rguha.net/code/java/cdkdesc.html>
- Scaffold Hunter: Visual analysis of large and complex data sets <http://scaffoldhunter.sourceforge.net>

And many many more....

naturalproducts.net Resources



NaPLeS - Natural Product Likeness Score calculator



About

Natural product-likeness of a molecule, i.e. similarity of this molecule to the structure space covered by natural products, is a useful criterion in screening compound libraries and in designing new lead compounds. This NP-likeness scorer has been trained on 315 916 natural products (NP) from various public databases, on a manually curated NP dataset used for the publication of the previous standalone NP-likeness scorer [1] and on 488 642 randomly selected synthetic molecules (SM) from the [ZINC](#) database.

Sugar moieties are removed from the molecules for the training and for the computation of the NP-likeness score as in [2].

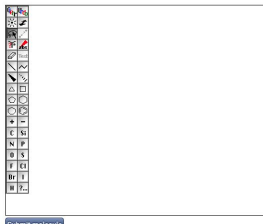
The stereochemistry is removed from the training set and from the user-submitted molecules.

Four ways to use the NP-likeness scorer:

- Upload a molecular file in one of the accepted formats (MOL, SDF or smiles). Maximum 1000 molecules per file.
- Paste a SMILES string of a molecule
- Draw a molecule
- Visualise the distribution of the NP-likeness score across natural products of public databases and taxonomy (bacteria, fungi and plants)

Upload a file

Draw a molecule



[Cheminformatics](#)



Find natural products

Molecule name, InChI, InChIKey, formula, COCONUT id, SMILES, chemical class

Search

Structure Search | Advanced Search

Home Browser Search Download Documentation

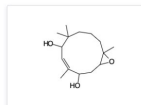
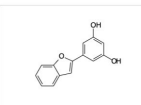
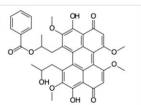
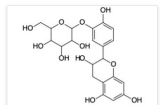
Natural Products Online is an open source project for Natural Products (NPs) storage, search and analysis. The present version hosts COCONUT, the COllection of Open Natural Products, one of the biggest and best annotated resources for NPs available free of charge and without any restriction.

Component Browser

Cards Table

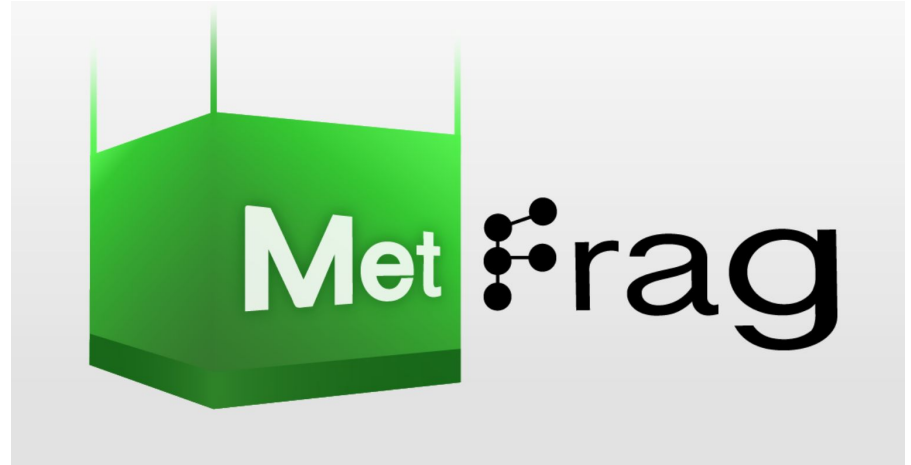
There are 401 624 unique natural products in the database. They are sorted by their annotation level, starting with the best annotated.

< 1 2 3 4 5 6 7 8 9 10 ... 16734 >



MetFrag

MetFrag is a freely available software for the annotation of high precision tandem mass spectra of metabolites which is a first and critical step for the identification of a molecule's structure. Candidate molecules of different databases are fragmented in silico and matched against mass to charge values. A score calculated using the fragment peak matches gives hints to the quality of the candidate spectrum assignment.



Link: <https://ipb-halle.github.io/MetFrag/>



MetFrag

In silico fragmentation for computer assisted identification of metabolite mass spectra

Database Settings

Database:

Parent Ion:

Neutral Mass: Search ppm:

Formula:

Identifiers:

Candidate Filter & Score Settings

Fragmentation Settings & Processing

Mzppm:

Mzabs:

Mode:

Tree depth:

Group candidates

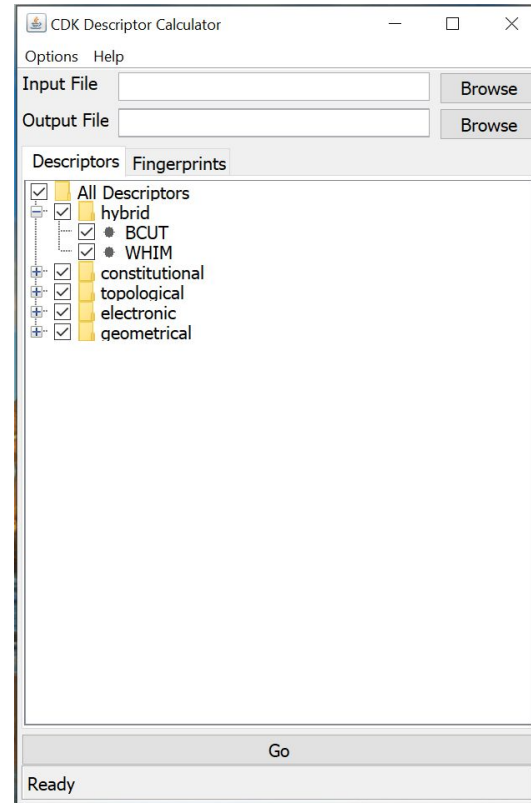
MS/MS Peak list

```
90.97445 681
106.94476 274
110.02750 110
115.98965 95
117.98540 384
124.93547 613
124.99015 146
125.99793 207
133.95592 777
143.98846 478
144.98846 478
```

CDK Descriptor Calculator

- Calculate multiple fingerprints and molecular descriptors easily
- Developed by one of the CDK developers, Rajarshi Guha

Link: <http://www.rguha.net/code/java/cdkdesc.html>



An abstract geometric network diagram consisting of numerous black dots (nodes) connected by thin black lines (edges). The nodes are arranged in a complex, interconnected pattern, forming various polygonal shapes. Some nodes are larger than others, and the overall structure is dense and intricate, resembling a complex graph or a network of relationships. The diagram is positioned in the upper right and lower right areas of the slide, with a large white space on the left containing the text.

Intermediate questions?

CDK hands-on!

CDK examples in GitHub: https://github.com/mSorok/CDK_secondary_metabolites_in_plants_2021

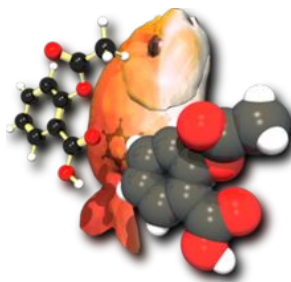
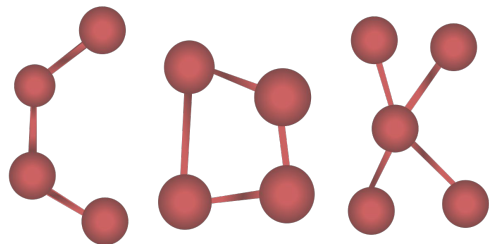
- a. Basics
- b. *AtomContainer* manipulation:
 - i. Molecular descriptors calculation
 - ii. Fingerprints

CDK in the world: other cheminformatic libraries

A very nice article on Open Source Cheminformatic Toolkits:

<https://www.macinchem.org/scientificsoftware/toolkits.php>

- **RDKit**: Python, great and active community, lots of functionalities. Try it out: <http://www.rdkit.org>
- **OpenBabel**: chemical file formats manipulation (http://openbabel.org/wiki/Main_Page)
- **OpenChemLib**: JavaScript, multiple tools, especially adapted for web. Try it out: <http://www.cheminfo.org/> or <https://www.npmjs.com/package/openchemlib>
- and many more....!



Open-Source Cheminformatics
and Machine Learning



Question time

Acknowledgements