Research Ideas and Outcomes 10: e124884
doi: 10.3897/rio.10.e124884

Grant Proposal

The Chemistry Development Kit in 2024: improving cheminformatics research

Egon Willighagen‡, Marc A.T. Teunis§, Alyanne De Haan§

‡ Maastricht University, Maastricht, Netherlands
§ Hogeschool Utrecht, Utrecht, Netherlands

Corresponding author: Egon Willighagen (egon.willighagen@maastrichtuniversity.nl)

Reviewable v 1

Received: 07 Apr 2024 | Published: 17 Apr 2024

Citation: Willighagen E, Teunis MAT, De Haan A (2024) The Chemistry Development Kit in 2024: improving cheminformatics research. Research Ideas and Outcomes 10: e124884. https://doi.org/10.3897/rio.10.e124884

Abstract

Cheminformatics is the research field that deals with information about chemical systems. This includes the chemical structure which is used in computational chemistry where quantum chemistry is too complex. The Chemistry Development Kit (CDK) was one of the first Open Science libraries in chemistry, co-founded in The Netherlands. The source code goes as far back as 1997 and has been maintained for more than 25 years. The CDK is used by many tools in drug discovery, computational toxicology, and bioinformatics. This project will develop improvements to the core library and update tools using the CDK to use the latest release.

Keywords

cheminformatics, Chemistry Development Kit, Java, open science, OSF23.2.097

Project proposal

The vision for your project

The Chemistry Development Kit (CDK, research-software-directory.org/software/cdk) (Steinbeck et al. 2003, Steinbeck et al. 2006, Willighagen et al. 2017), rdkit, and OpenBabel are the three leading open source cheminformatics tools, powering much of the cheminformatics research. For example, the Chemistry Development Kit (CDK), such as the generic workflow platform KNIME (Beisken et al. 2013) and PaDEL-descriptor (Yap 2010) used in chemical property prediction research, each cited more than 1000 times. The CDK publications are cited more than two thousand times and the software is mentioned almost 50 thousand times on GitHub.

The most recent CDK release is CDK 2.9 of August 2023 (Mayfield et al. 2023). However, many tools using the CDK use an older version of the CDK. Each CDK release improves the use of open standards, such as SMILES, InChI, and open data, such as the latest IUPAC isotope masses and element names. Use of the latest CDK version improves the interoperability of the tools using it. Our intention is that important CDK-based tools use the latest CDK version, for which we will collaborate with the respective open science projects.

The vision of this project is two-fold. For the first part (see **WP1**), this project will improve the CDK library by introducing and making it compatible with newer Java programming language features. While developing these, the coverage of JavaDoc and unit testing with JUnit will be extended. Moreover, OSGi support will be improved, where a current limitation is that multiple OSGi bundles contain the same Java package, causing problems with OSGi-based software, like OpenChrom (Wenig and Odermatt 2010) and PathVisio (Kutmon et al. 2015).

The second part of this project will focus on updating tools using the CDK to the latest CDK version (currently 2.9, but later versions when they are released; see **WP2**). Tools of particular interest for updating are JChemPaint (Krause et al. 2000), AMBIT (Jeliazkova and Jeliazkov 2011), rcdk, OpenChrom, DECIMER (Rajan et al. 2023), KNIME, and PaDEL-descriptor. These downstream tools will need to be updated for changes in the application programming interfaces (APIs) of the CDK. Required API changes will be publicly shared and disseminated with the *Groovy Cheminformatics with the Chemistry Development Kit* book (egonw.github.io/cdkbook/). The applicants have a long-standing collaboration with the developers of most of the tools of interest.

Finally, we will present the results of this project at an open, international user group meeting (**WP3**). This meeting will be open to presentations from any project around CDK-based tools, both about the tools and about scientific research using those tools.

Project plan

The project plan is organized in three work packages (WP1, WP2, WP3, see also Table 1). **WP1** will be focusing on the CDK library itself, the others on updating the tools using the CDK. The CDK library is built with Apache Maven, but the Maven modules show a complex dependency tree, where the more core modules have fewer dependencies on third-party

libraries. This modularisation, however, needs upgrading. OSGi bundles are already created by Maven, but the bundle content does not yet match the standard and one Java package can be split over multiple bundles. Second, we wish to improve the maintainability of the library and improve the code coverage of the (unit) testing to at least 50% for all modules (statistics available at app.codecov.io/gh/cdk/cdk). Third, the CDK will be updated to compile with Java 21 (which it currently does not).

Table 1.

Gantt diagram of project work timeline. In the months M3 and M5, two two-day hackathons (H) will be organized.

	M1	M3	M5	M7	M8	M11
WP1	H					
WP2				H		
WP3						UGM

Work package 2 (**WP2**) focuses on the tools using the CDK, particularly JChemPaint, AMBIT, ToxTree, rcdk, OpenChrom, DECIMER, KNIME, the ChemViz plugin of Cytoscape, and PaDEL-descriptor. JChemPaint is an open source chemical structure editor, but not currently based on the latest CDK version. KNIME is one of the tools using JChemPaint, for which the CDK extension cannot be fully updated until JChemPaint is. AMBIT is the open source chemistry database frequently used in the field of toxicology and will benefit from newer CDK functionality too. Toxtree, OpenChrom, rcdk, DECIMER, are other tools that will benefit from improvements of recent CDK versions. Updating the PaDEL-descriptor software will be the most challenging, but with its many users and citations will have a significant impact on molecular property prediction research.

WP3 will organize a final user group meeting (UGM) where the project results will be presented and where users of the CDK will be invited. The open UGM will consist of a session with presentations about projects using the CDK and a workshop where we will collect needed future CDK functionality and improvements.

Project roles and expertise

The funding will be used to fund the work of three researchers, **Alyanne De Haan** (orcid:0000-0003-0896-0906), **Marc Teunis** (orcid:0000-0002-3496-6669), and **Egon Willighagen** (orcid:0000-0001-7542-0286). Willighagen is co-founder of the Chemistry Development Kit and researcher at Maastricht University and has been involved in many other Open Science projects, involved in the Dutch Open Science Community, former editor of the Journal of Cheminformatics, and leader of various open scientific software projects. Teunis and De Haan have extensive knowledge about open educational resources, version control, and generally open science approaches to support research.

Open Science track record of the applicant

Dr Egon Willighagen has been active in Open Science for over 20 years, for example, contributing to projects like JChemPaint (since 1998; doi:10.3390/50100093), WikiPathways (since 2011; doi:10.1093/NAR/GKV1024), and (temporarily) leading projects like Jmol and coordinating the science in the EU FP7 project eNanoMapper (doi:10.3762/BJNANO.6.165), and co-founded the Chemistry Development Kit (in 2000; doi:10.1021/ci025584y). He is recognized for his work with the international Blue Obelisk Award (2007), a national runner-up Open Initiative Trophy (2021), and received a NWO Open Science Fund in 2022. For five years he was one of two Editor-in-Chief of the full CC-BY, highly ranked Journal of Cheminformatics (issn:1758-2946), which promotes Open Science in chemistry. At various National Plan Open Science events and meetings, Willighagen has provided input from a researcher's perspective and is co-founder of the Open Science Community Maastricht. A more complete list of his Open Science work can be found in his publication list: orcid.org/0000-0001-7542-0286.

Data management

Will this project involve re-using existing research data?

This project will focus on software, but if existing data is used, it will be open, to ensure compatibility with the open license of the CDK.

Will data be collected or generated that are suitable for reuse?

This is not expected.

After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? Are there possible restrictions to data sharing or embargo reasons?

This is not expected, but if this is the case, it will be shared via Zenodo.

Will any costs (financial and time) related to data management and sharing/preservation be incurred?

No. All the necessary resources (financial and time) to store and prepare data for sharing/preservation are or will be available at no extra cost.

Software sustainability

How large do you expect the community that will potentially use the software to be, and do you expect outside contributors to the software?

The Chemistry Development Kit has a wide user base that uses the CDK directly or indirectly via one of the CDK-based tools. The Java package “org.openscience.cdk” is mentioned almost 50 thousand times on GitHub, the software using the CDK cited many times, and searching “chemistry development kit” on Google Scholar finds more than two thousand hits. The potential is very significant.

How will the software be licensed and be made available for re-use?

The CDK is released under the GNU Lesser GPL license, which is similar to the GPL license but allows use in proprietary software, and only changes to the CDK itself need to be made available under the same open license. Our work on the CDK library will use the same LGPL license as the CDK.

Similarly, our updates in WP1 will be made available under the same license as the tool it improves. An overview of licenses of these tools is given in Table 2.

Table 2.
Licenses of source code repositories of various tools using the Chemistry Development Kit library.

Name	Source Code License	Code repository
KNIME (CDK Nodes)	LGPL	github.com/cdk/nodes4knime
rcdk	LGPL	github.com/CDK-R/cdkr
AMBIT	LGPL	sourceforge.net/projects/ambit/
JChemPaint	LGPL	github.com/JChemPaint/jchempaint
ToxTree	GPL v2	sourceforge.net/p/toxtree/
DECIMER	MIT	github.com/Kohulan/DECIMER-Java
ChemViz (Cytoscape)	LGPL	github.com/RBVI/chemViz2
OpenChrom	EPL 1.0	github.com/Openchrom/openchrom
PaDEL-descriptor	public domain	yapcwsoft.com/dd/padeldescriptor/

What measures are needed to make the software appropriate for long-term (re-)use by third parties?

CDK releases are made on GitHub (github.com/cdk/cdk/releases), archived on Zenodo (doi:10.5281/zenodo.592588), and distributed via Maven Central. Tools using the CDK have various solutions (see the previous section). For example, rcdk is released via the CRAN network for R software, and KNIME has a custom release and distribution network.

Patches to update these tools will be made available via GitHub pull requests and archived on Zenodo.

What expertise do you expect to be needed to make the software appropriate for long-term re-use by third parties? Is this expertise available?

Expertise is needed about version control, continuous building, code refactoring, software testing, software build systems, packaging standards, and other standards for software development. The applicants have this expertise, as demonstrated in their research output.

While it is not expertise needed, for formal releases of software, we will depend on the release managers of the respective packages. Some tools, like KNIME, have a fixed release schedule, but many do not. For the CDK itself, the release manager currently is Dr. John Mayfield. Deployments of CDK 'snapshot' releases to Maven Central are done by the main applicant. Of the tools we seek to improve, we know most developers personally (see co-authored articles in the publication lists of the applicants).

Funding program

Open Science Fund

Grant title

The Chemistry Development Kit in 2024: improving cheminformatics research

Hosting institution

Maastricht University

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C (2013) KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics 14 (1). https://doi.org/10.1186/1471-2105-14-257
- Jeliazkova N, Jeliazkov V (2011) AMBIT RESTful web services: an implementation of the OpenTox application programming interface. Journal of cheminformatics 3: 18. https://doi.org/10.1186/1758-2946-3-18

- Krause S, Willighagen E, Steinbeck C (2000) JChemPaint - Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. Molecules 5 (12): 93-98. https://doi.org/10.3390/50100093
- Kutmon M, van Iersel M, Bohler A, Kelder T, Nunes N, Pico A, Evelo C (2015) PathVisio 3: An Extendable Pathway Analysis Toolbox. PLOS Computational Biology 11 (2). https://doi.org/10.1371/journal.pcbi.1004085
- Mayfield J, Willighagen E, Guha R, Torrance G, Ujihara K, Rahman SA, Alvarsson J, Williamson M, Gražulis S, Katzel D, Pluskal T, Uli, Linn X, Wei YC, Szisz D, Kochev N, Clark A, Berg A, Bach E, Jeliazkova N, Stephan R, Plante J, Jönsson K, Dole K, Stueker O, De Pineda Gutierrez NA, Wenk M, Kaibioinfo, AndyHowlettGitHub, Katsubo D (2023) cdk/cdk: CDK 2.9. Zenodo https://doi.org/10.5281/zenodo.8270947
- Rajan K, Brinkhaus HO, Agea MI, Zielesny A, Steinbeck C (2023) DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nature communications 14 (1): 5045. https://doi.org/10.1038/s41467-023-40782-0
- Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. Journal of Chemical Information and Computer Sciences 43 (2): 493-500. https://doi.org/10.1021/ci025584y
- Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E (2006) Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. Current Pharmaceutical Design 12 (17): 2111-2120. https://doi.org/10.2174/138161206777585274
- Wenig P, Odermatt J (2010) OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. BMC Bioinformatics 11 (1). https://doi.org/10.1186/1471-2105-11-405
- Willighagen E, Mayfield J, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo C, Guha R, Steinbeck C (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics 9 (1). https://doi.org/10.1186/s13321-017-0220-4
- Yap CW (2010) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry 32 (7): 1466-1474. https://doi.org/10.1002/jcc.21707