

Reaction enumeration and machine learning enhancements for the open-source pipelining solution CDK-Taverna 2.0

Andreas Truszkowski^{1,3}, Stefan Neumann², Achim Zielesny³, Egon Willighagen⁴ and Christoph Steinbeck¹

¹ Chemoinformatics and Metabolism, European Bioinformatics Institute (EBI), Cambridge, UK

² GNWI – Gesellschaft für naturwissenschaftliche Informatik mbH, Oer-Erkenschwick, Germany

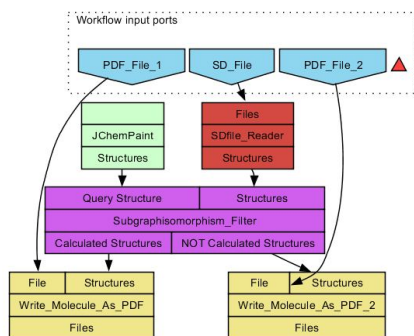
³ University of Applied Sciences of Gelsenkirchen, Institute for Bioinformatics and Chemoinformatics, Recklinghausen, Germany

⁴ Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

Abstract

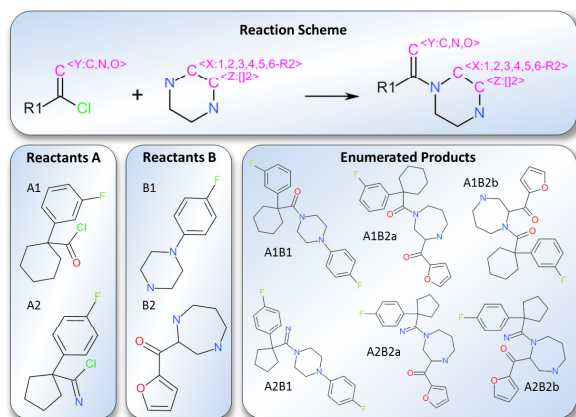
Pipelining or workflow tools allow for the Lego™-like, graphical assembly of I/O modules and algorithms into a complex workflow which can be easily deployed, modified and tested without the hassle of implementing it into a monolithic application.

The CDK-Taverna project aims at building an open-source pipelining solution through combination of different open-source projects such as Taverna [1], the Chemistry Development Kit (CDK) [2,3], the Waikato Environment for Knowledge Analysis (WEKA) [4] or Bioclipse[5]. A first integrated version of CDK-Taverna was recently released to the public [6].



The Lego™-like architecture allows a drag & drop arrangement of specific workers for tailored workflows.

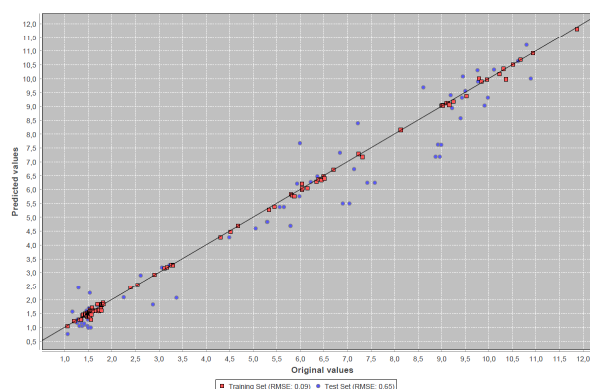
Current developments refactor all workers as well as the complete setup on the basis of Taverna 2.2.1 and CDK 1.3.7 which themselves introduce major improvements to the whole platform.



Combinatorial chemistry support: Enumeration of products for a template reaction and specific reactant libraries.

In addition the CDK is enhanced with specific functions and options for reaction enumeration based on a reaction template and corresponding reactant libraries. Reaction enumeration supports combinatorial chemistry approaches in the drug discovery process.

Three-layer perceptron-type neural network



The machine learning workers enable easy-to-use evaluation and visualisation techniques for data analysis.

Furthermore the machine learning section is established using the WEKA library. The CDK-Taverna plugin provides different workers for classification, regression and clustering as well as evaluation and result visualization workers.

Features

There are more than **190 workers** in preparation:

- **Input:** Molfiles, SD files, RXN files, SMILES, CML, JChemPaint structures, ARFF, XRFF
- **Output:** Molfiles, SD files, RXN files, SMILES, CML, PNG, JPEG, PDF, ARFF, XRFF
- **Iterative file reading:** SD files, RXN files (supports large file sizes)
- **Filter:** Substructures, salts, atom types
- **Reaction enumeration** with advanced options
- **Calculation** of more than 90 QSAR descriptors
- **Clustering, Regression and Classification** methods

CDK-Taverna 2.0 works on Microsoft Windows and Mac OS X 32/64bit operating systems and will be released soon.

References:

- [1] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004, 20(17):3045-3054.
- [2] Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen E: The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences* 2003, 43(2):493-500.
- [3] Steinbeck C, Hoppe C, Kuhn S, Guha R, Willighagen EL: Recent Developments of The Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design* 2006, 12(17):2111-2120.
- [4] Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P, Witten IH: WEKA - Experiences with a Java Open-Source Project, *Journal of Machine Learning Research* 2010, 11:2533-2541
- [5] Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Steinbeck C, Wikberg JE: Bioclipse: An open rich client workbench for chemo- and bioinformatics. *BMC Bioinformatics* 2007, 8(59).
- [6] Kuhn T, Willighagen EL, Zielesny A, Steinbeck S: CDK-Taverna: an open workflow environment for cheminformatics. *BMC Bioinformatics* 2010, 11:159.

