JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

Cite This: J. Chem. Inf. Model. 2019, 59, 4083−4086

pubs.acs.org/jcim

TeachOpenCADD-KNIME: A Teaching Platform for Computer-Aided Drug Design Using KNIME Workflows

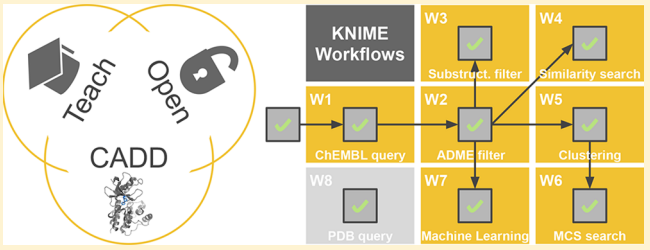
Dominique Sydow,[†,∥] Michele Wichmann,[†,∥] Jaime Rodríguez-Guerra,[†] Daria Goldmann,[‡] Gregory Landrum,[§] and Andrea Volkamer*[,†]

[†]In Silico Toxicology, Institute of Physiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

[‡]KNIME GmbH, Körtestr. 10, 10967 Berlin, Germany

[§]KNIME AG, Technoparkstr. 1, 8005 Zurich, Switzerland

ABSTRACT: Open-source workflows have become more and more an integral part of computer-aided drug design (CADD) projects since they allow reproducible and shareable research that can be easily transferred to other projects. Setting up, understanding, and applying such workflows involves either coding or using workflow managers that offer a graphical user interface. We previously reported the TeachOpenCADD teaching platform that provides interactive Jupyter Notebooks (talktorials) on central CADD topics using open-source data and Python packages. Here we present the conversion of these talktorials to KNIME workflows that allow users to explore our teaching material without any line of code. TeachOpenCADD KNIME workflows are freely available on the KNIME Hub: https://hub.knime.com/volkamerlab/space/TeachOpenCADD.



INTRODUCTION

In computer-aided drug design (CADD), computational tools are used to process and rationalize large and heterogeneous data sets involving small molecules and macromolecules. For this endeavor, open-access resources have gained momentum, especially for setting up complex workflows, since they enable modular, reproducible, and reusable research.

We recently reported the TeachOpenCADD[1] teaching platform (https://github.com/volkamerlab/teachopencadd) that provides learning material for CADD using open-source data and Python libraries. Central topics in CADD are covered in the form of interactive Jupyter Notebooks that contain both theory and code for each topic.

An alternative to code-based pipelines are workflow managers that allow the design of protocols via an intuitive drag-and-drop style graphical interface without the need for coding. KNIME[2,3] is a popular workflow manager for data science with several open-source modules for CADD,[4] while its usage ranges from small in-house applications such as compound library preparation to more complex workflow applications integrating chemical, pharmacological, and structural information. An example of the latter is 3D-e-Chem,[5,6] which allows, e.g., structure-based bioactivity data mapping of kinase inhibitors or structure-based GPCR−kinase cross-reactivity prediction.

Here we address users who aim to learn how to use KNIME for CADD applications as well as users who desire to study central CADD topics without necessarily learning how to code. We report the conversion of the TeachOpenCADD Python pipeline (talktorials T1−T8) to a KNIME workflow pipeline (workflows W1−W8). The KNIME pipeline is publicly available on the KNIME Hub: https://hub.knime.com/volkamerlab/space/TeachOpenCADD (current release: https://doi.org/10.5281/zenodo.3475086).

METHODS

KNIME (the Konstanz Information Miner) provides an open-source data analysis, reporting, and integration platform. KNIME enables users to create data workflows, execute selected analysis steps, and check intermediate and final results, models, and interactive views via a graphical user interface. Coding is not required, since the workflows are built up by stringing together small preimplemented code units (nodes) with defined, tested, and thus standardized functionalities, which can be configured with individual settings. In addition, KNIME offers functionalities to design complex workflows in a well-structured way via metanodes that encapsulate parts of a workflow.

This work was developed using KNIME version 4.0.0 and uses nodes from the KNIME Analytics Platform, KNIME Extensions, and Community Extensions by RDKit[3,7] and Vernalis[8] (RSCB PDB Tools).

RESULTS

The TeachOpenCADD KNIME pipeline consists of eight interconnected workflows (W1−W8) in the form of metanodes, each containing one CADD topic. The pipeline is illustrated using the epidermal growth factor receptor

Received: August 8, 2019
Published: October 15, 2019

DOI: 10.1021/acs.jcim.9b00662
J. Chem. Inf. Model. 2019, 59, 4083−4086

Figure 1. The TeachOpenCADD KNIME pipeline offers eight KNIME workflows covering central topics in CADD while using open-source data and KNIME nodes. This figure shows the graphical interface of KNIME, demonstrating the software's visual potential.

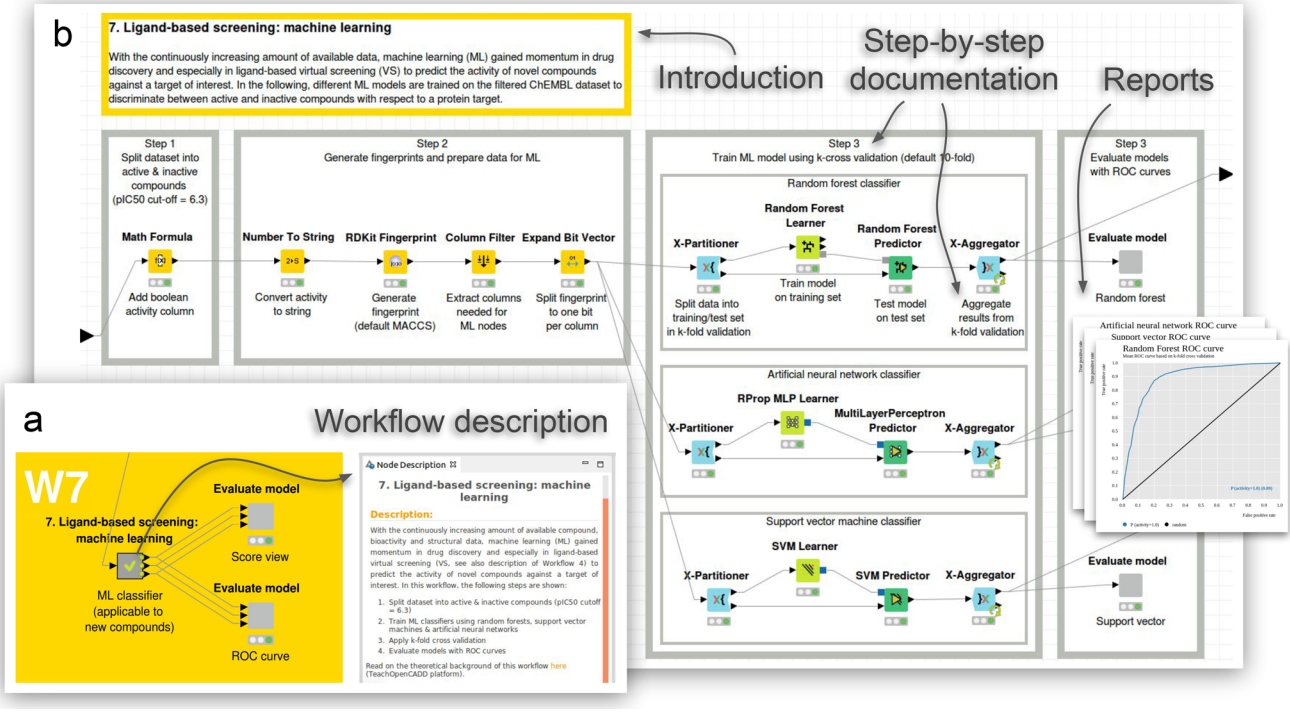


Figure 2. Workflow composition shown for workflow W7 (ligand-based screening: machine learning). (a) Each workflow metanode is labeled with a brief topic description and the main workflow steps. (b) The interior of each workflow metanode consists of an introduction, nodes organized in boxes per step, node documentation, and output reports.

(EGFR)[9,10] but can easily be applied to other targets of interest. Topics include how to fetch, filter, and analyze compound data associated with a query target and are briefly described in the following (Figure 1). For a detailed description, we refer the reader to the initial TeachOpenCADD publication.[1]

First, compound data for the query target EGFR are acquired from the ChEMBL web services[11] (W1)[12] and

DOI: 10.1021/acs.jcim.9b00662
J. Chem. Inf. Model. 2019, 59, 4083−4086

subsequently filtered for drug-likeness using Lipinski's rule of five (W2). This filtered data set forms the basis for the remaining workflows. Unwanted substructures that potentially cause toxicity or nonspecific assay interactions are detected (W3), and a similarity search for a ligand-based screen with the EGFR inhibitor gefitinib as the query[13] is conducted (W4). Compounds are grouped using a hierarchical clustering algorithm (W5),[14] whereupon the maximum common substructure is detected and visualized for the largest cluster (W6).[15] Additionally, machine learning approaches are employed to build models for active compound prediction (W7).[16] Lastly, ligand−EGFR complexes are fetched from the PDB web services[17] and filtered by criteria such as structure resolution (W8).[18] The last two previously reported talktorials, T9 and T10, were not translated to workflows because of their extensive use of PyMOL, which is currently not supported in KNIME.

The workflows can be examined and executed independently from each other or as a pipeline. As shown in Figure 2 for W7, each workflow is introduced with a brief topic motivation and grouped into multiple steps using gray boxes that contain a step description and all step-associated nodes labeled with task descriptions. Results from intermediate steps (e.g., filtered compound tables) or from final plotting nodes can be viewed interactively and configured easily using the nodes' graphical interface.

CONCLUSION

The TeachOpenCADD platform offers learning material on central topics of cheminformatics and structural bioinformatics. In the present work, teaching material was translated from code-based Jupyter Notebooks to KNIME workflows, which have several advantages. KNIME workflows (i) are knitted together from preimplemented nodes with standardized functionalities, (ii) are easy to understand because of the visual representation of their architecture, and (iii) permit a low-threshold entry for nonprogrammers to build customized pipelines.

The TeachOpenCADD KNIME pipeline is suitable for self-study training and classroom teaching but can also serve as a starting point for workflows in research projects. TeachOpenCADD is open for contributions and ideas from the community with regard to both Jupyter Notebooks and KNIME workflows.

AUTHOR INFORMATION

Corresponding Author

*E-mail: andrea.volkamer@charite.de.

ORCID

Dominique Sydow: 0000-0003-4205-8705
Michele Wichmann: 0000-0002-7441-1561
Jaime Rodríguez-Guerra: 0000-0001-8974-1566
Daria Goldmann: 0000-0002-4793-8579
Gregory Landrum: 0000-0001-6279-4481
Andrea Volkamer: 0000-0002-3760-580X

Author Contributions

‖D.S. and M.W. share first authorship.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

A.V. and D.S. received funding from the Deutsche Forschungsgemeinschaft (Grant VO 2353/1-1). A.V. received funding from the Bundesministerium für Bildung und Forschung (Grant 031A262C). J.R.-G. received funding from the Stiftung Charité (Einstein BIH Visiting Fellow Project). M.W. received funding from the "SUPPORT für die Lehre" Program (Förderung innovativer Lehrvorhaben) of Freie Universität Berlin.

REFERENCES

(1) Sydow, D.; Morger, A.; Driller, M.; Volkamer, A. TeachOpenCADD: A Teaching Platform for Computer-Aided Drug Design Using Open Source Packages and Data. *J. Cheminf.* **2019**, *11*, 29.

(2) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, 2008; pp 319−326.

(3) Fillbrunn, A.; Dietz, C.; Pfeuffer, J.; Rahn, R.; Landrum, G. A.; Berthold, M. R. KNIME for Reproducible Cross-Domain Analysis of Life Science Data. *J. Biotechnol.* **2017**, *261*, 149−156.

(4) Mazanetz, M. P.; Goode, C. H. F.; Chudyk, E. I. Ligand- and Structure-Based Drug Design and Optimization Using KNIME. *Curr. Med. Chem.* **2019**, DOI: 10.2174/0929867326666190409141016.

(5) McGuire, R.; Verhoeven, S.; Vass, M.; Vriend, G.; de Esch, I. J. P.; Lusher, S. J.; Leurs, R.; Ridder, L.; Kooistra, A. J.; Ritschel, T.; de Graaf, C. 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *J. Chem. Inf. Model.* **2017**, *57*, 115−121.

(6) Kooistra, A. J.; Vass, M.; McGuire, R.; Leurs, R.; de Esch, I. J. P.; Vriend, G.; Verhoeven, S.; de Graaf, C. 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery. *ChemMedChem* **2018**, *13*, 614−626.

(7) RDKit Nodes for KNIME. https://www.knime.com/rdkit (accessed May 15, 2019).

(8) Roughley, S. Five Years of the KNIME Vernalis Cheminformatics Community Contribution. *Curr. Med. Chem.* **2018**, DOI: 10.2174/0929867325666180904113616.

(9) UniProt Entry for EGFR. https://www.uniprot.org/uniprot/P00533 (accessed May 16, 2019).

(10) Chen, J.; Zeng, F.; Forrester, S. J.; Eguchi, S.; Zhang, M.-Z.; Harris, R. C. Expression and Function of the Epidermal Growth Factor Receptor in Physiology and Disease. *Physiol. Rev.* **2016**, *96*, 1025−1069.

(11) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A LargeScale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−7.

(12) Adapting KNIME Workflow Example to Extract Bioactivities for a Target ID. KNIME EXAMPLES Server under 50_Applications/30_RESTful_ChEMBL/ 03_ChEMBL_Bioactivity_Search (accessed May 18, 2019).

(13) DrugBank Entry for Gefitinib. https://www.drugbank.ca/drugs/DB00317 (accessed May 16, 2019).

(14) Adapting KNIME Workflow Example to Cluster Molecules Using RDKit Nodes. KNIME EXAMPLES Server Under 99_Community/03_RDKit/01_Clustering (accessed May 24, 2019).

(15) Adapting KNIME Workflow Example Created by Daria Goldmann. KNIME Introduction and Training Session on 2019-01-21 at Volkamer Lab in Berlin: 0 × 1_Maximum_Common_Substructure (accessed Jan 21, 2019).

(16) Adapting KNIME Workflow Example Created by Daria Goldmann. KNIME Introduction and Training Session on 2019-01-21 at Volkamer Lab in Berlin: 0 × 2_Machine_Learning (accessed Jan 21, 2019).

DOI: 10.1021/acs.jcim.9b00662
J. Chem. Inf. Model. 2019, 59, 4083−4086

(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−42.
(18) Adapting KNIME Workflow Example to Download and Save PDB Queries using Vernalis Nodes. KNIME EXAMPLES Server Under 99_Community/04_Vernalis/01_PDB_Query_Downloder_-and_Save_Locally (accessed May 24, 2019).

DOI: 10.1021/acs.jcim.9b00662
J. Chem. Inf. Model. 2019, 59, 4083−4086