



It's not just for Python: Interacting with chemical data using KNIME and RDKit

Daria Goldmann
KNIME AG

2018 RDKit UGM

KNIME Analytics Platform

The screenshot displays the KNIME Analytics Platform interface. The top bar shows the title "KNIME Analytics Platform - /Users/daria_knime/knime-demo". The left sidebar contains the "KNIME Explorer" and "Node Repository" panels. The main workspace shows a workflow titled "Machine Learning with Chemical Fingerprints". The workflow is divided into three sections: "Read", "Preprocess and Explore", and "Build and Score a Model". The "Read" section includes a "Table Reader" node. The "Preprocess and Explore" section includes "RDKit Salt Stripper", "RDKit Descriptor Calculation", "Color Manager", "Parallel Coordinates Plot (JavaScript)", and "RDKit Fingerprint" nodes. The "Build and Score a Model" section includes "Partitioning", "Random Forest Learner", "Random Forest Predictor", and "Scorer (JavaScript)" nodes. The "Scorer (JavaScript)" node is highlighted, and its description is shown in the right sidebar. The bottom panel shows the "Console" and "Node Monitor" tabs. The "Node Monitor" tab displays the state of the "Scorer (JavaScript)" node as "EXECUTED" and shows flow variables.

Machine Learning with Chemical Fingerprints.
This workflow demonstrates Model Building for a bioactivity data set with Random Forest learner and binary fingerprints. The data set represents a subset of 844 compounds evaluated for activity against CDPK1; 181 compounds inhibited CDPK1 with IC50 below 1uM and have "active" as their class. More information is available <https://www.ebi.ac.uk/cheminformatics/dataset>. See Set 19.

Scorer (JavaScript)
Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of a given attribute and its classification match. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The view of the node displays three tables, the first one is the confusion matrix with the number of matches in each cell. Row and column rates can be shown via a configuration setting; they are the number of correct predictions divided by the total number of records in the confusion matrix. Additionally, it

Node: Scorer (JavaScript) (0:261)
State: EXECUTED

Flow Variables: Port 0 Load data

Variable	Value
knime.workspace	/Users/daria_knime/knime-demo

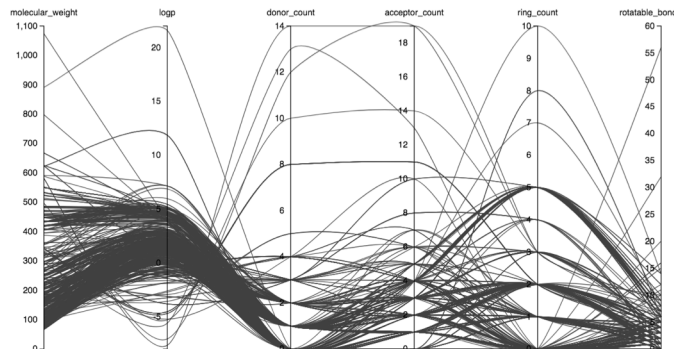
513M of 1077M

Guided Analytics in KNIME

KNIME Analytics Platform - /Users/daria_knime/knime-demo

Explore the input

Parallel Coordinates Plot



Node Description

Explore the input

<no description set>
In order to set a description browse the input node contained in the Wrapped Metanode and change its configuration.

Ports

Input Ports

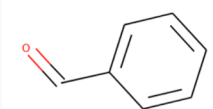
0 <no description set>
Change this label by browsing the input node contained in the Wrapped Metanode and changing its configuration.

Output Ports

0 <no description set>
Change this label by browsing the output node contained in the Wrapped Metanode and changing its configuration.

Node Repository

Show 10 entries

Keep	mol_original	patent_id	annotation_reference	schembl_id	smiles	molecular_weight
<input checked="" type="checkbox"/>		WO-2006112464-A1	benzoyl	SCHEMBL573	<chem>O=CC1=CC=CC=C1</chem>	106.12200164794922
<input checked="" type="checkbox"/>		WO-2006112464-A1	7-[4-(4-benz[b]thiophen-4-yl)-piperazin-1-yl] butoxy]-IH-	SCHEMBL1037592	<chem>O=C1NC2=C(C(=C1)C=CC(=O)CCCCN1CCN(CC1)C1=CC=CC3=C1C=CS3)=C2</chem>	433.5660095214844

1077M of 1441M

The problem

- You found an interesting patent
 - get an overview of the chemistry in the document
 - find those hidden key compounds
 - explore the scaffolds and chemistry further

Example

- Brexiprazole: “piperazine-substituted benzothiophenes for treatment of mental disorders”

Approach

- Interactively explore compounds and pick those interesting ones for further analysis
- Select possible key compounds
- Fuzzy MCS search? R-group decomposition?
- Interactive and reproducible

09_Explore_Patent_Data_and_Find_MCS

09ExplorePatentDataandFindMCS

This workflow allows to interactively explore the data available in a patent downloaded from SureChEMBL. The input data is available for download via https://workflows.knime.com/knime/rest/v4/repository/99Community/03RDKit_data/patents:data. Use your forum login to access the data

Tags: cheminformatics, RDKit, patent data, JavaScript Views

Workflow Image

09_Explore_Patent_Data_and_Find_MCS

This workflow allows to interactively explore the data available in a patent downloaded from SureChEMBL. The input data is available for download via https://workflows.knime.com/knime/rest/v4/repository/99Community/03RDKit_data/patents:data. Use your forum login to access the data and the data associated with it.

Basic

Pick a file
Select the patent "tag"

Preprocess and Explore

Generate Images
Explore the input - RDKit Descriptors
Explore the input - RDKit Descriptors - improved structure

Explore possible key compounds

RDKit Fragment
Generate Network from Distance Matrix
Network View and Filter

Explore MCS

MCS Search
MCS View Results

Download Results

Download File
Download File (from Repository)

EGF Value
Save the result

Optional: Remove the nodes when closing the workflow

By downloading the workflow you agree to our [terms & conditions](#).

Author
daria.goldmann

Path
/99_Community/03_RDKit/

Created on
19.09.2018 14:38:30

Last Uploaded on
25.09.2018 12:09:31

Rating
This workflow has not been rated yet

Rate this workflow:
☆☆☆☆☆

Tags
RDKit JavaScript Views cheminformatics

Requirements
KNIME Core 3.6.1

Use your KNIME forum login for access

Find the most interesting compounds

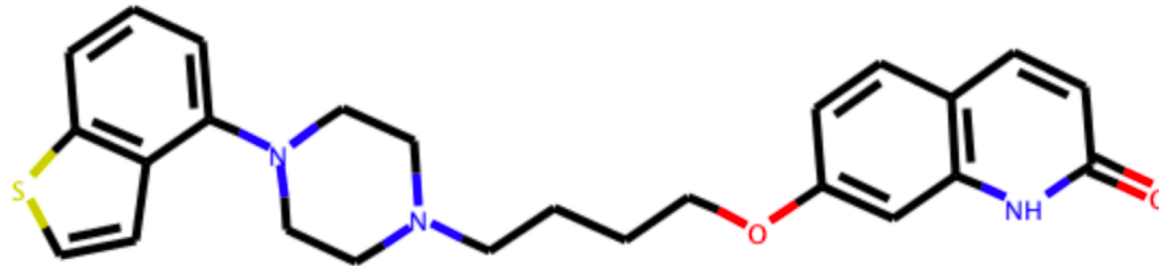
- By identifying those that have a large number of neighbors (=very similar compounds)
- By constructing a similarity based network and calculating "hub scores" to rank compounds

Hattori, K., Wakabayashi, H. & Tamaki, K. Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone. *J. Chem. Inf. Model.* **48**, 135–142 (2008).

https://en.wikipedia.org/wiki/HITS_algorithm

Example

- Brexiprazole: “piperazine-substituted benzothiophenes for treatment of mental disorders”



Brexiprazole

Use possible key compounds further?

- Approximate the Markush structure: take the whole cluster and do a fuzzy MCS
- Retrieve **all** compounds matching that substructure

Workflow Hub: workflows.knime.com

Welcome to the KNIME Community Workflow Hub, the place for the KNIME community to share, rate and comment on workflows. Here you can browse all of our example workflows. Invited community members may post their own workflows to share. If you didn't receive an invite, but want to share your workflows with the community, please [get in touch](#).

News

Tweets by @knime

KNIME @knime
#KNIME #Meetup in #Warsaw on September 24. Join us & our partner @EPAMSYSTEMS to learn more about #KNIME #Analytics Platform & hear about some #MachineLearning & #GuidedAnalytics use cases. Register at bit.ly/2JOGNfK #DataScience #OpenSource
16h

[Embed](#) [View on Twitter](#)

Most Recent Workflows

01_Guided_Analytics_for_ML_Automation uploaded by christian.dietz on 19.09.2018 11:24:45	★★★★★
01_Bioactivity_Prediction_Load_Local uploaded by daria.goldmann on 20.08.2018 17:30:06	☆
Model_Factory uploaded by daria.goldmann on 20.08.2018 17:30:06	★★★★★
03_Bioactivity_Prediction_Learn_All_Methods uploaded by daria.goldmann on 20.08.2018 17:30:06	☆
02_Bioactivity_Prediction_Transform uploaded by daria.goldmann on 20.08.2018 17:30:06	☆

[Show more...](#)

Top Rated Workflows

03_Simple_MMP_Example uploaded by knime_admin on 29.11.2017 16:36:44	★★★★★
---	-------

Random Workflows

03_Looping_over_all_columns_and_manipulation_of_each uploaded by knime_admin on 20.01.2018 11:23:25	☆
--	---

- Use your KNIME forum login for access

Summary

- It's easy to build a workflow to interactively explore the data in KNIME Analytics Platform
- With RDKit integration it's easy to build cheminformatics applications to guide the analysis of the data

