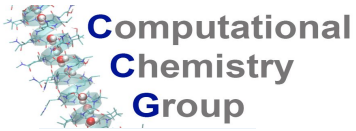


ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Intro to the RDKit

Greg Landrum, Ph.D.

AIDD, 19 Oct 2021

greglandrum

@dr_greg_landrum

Sneak preview

After these slides, I will be working through an Jupyter notebook with a bit of tutorial.

The notebook is here:
https://github.com/greglandrum/AIDD_RDKit_Tutorial_2021

If you just want to follow along without installing anything, you can use this mybinder link:
https://tinyurl.com/RDKit-AIDD-2021
or
https://mybinder.org/v2/gh/greglandrum/AIDD_RDKit_Tutorial_2021/HEAD

Me

- RDKit principle developer
- Now:
  - Senior scientist, Riniker Lab, ETH Zurich
  - T5 Informatics GmbH
  - Senior advisor, KNIME AG
- Before:
  - KNIME (Zurich/Konstanz)
  - Novartis (Basel)
  - startups (San Francisco Bay area)

The RDKit: an open-source toolkit for cheminformatics

What *is* a toolkit anyway?

A collection of tools for building things

some simple

some not

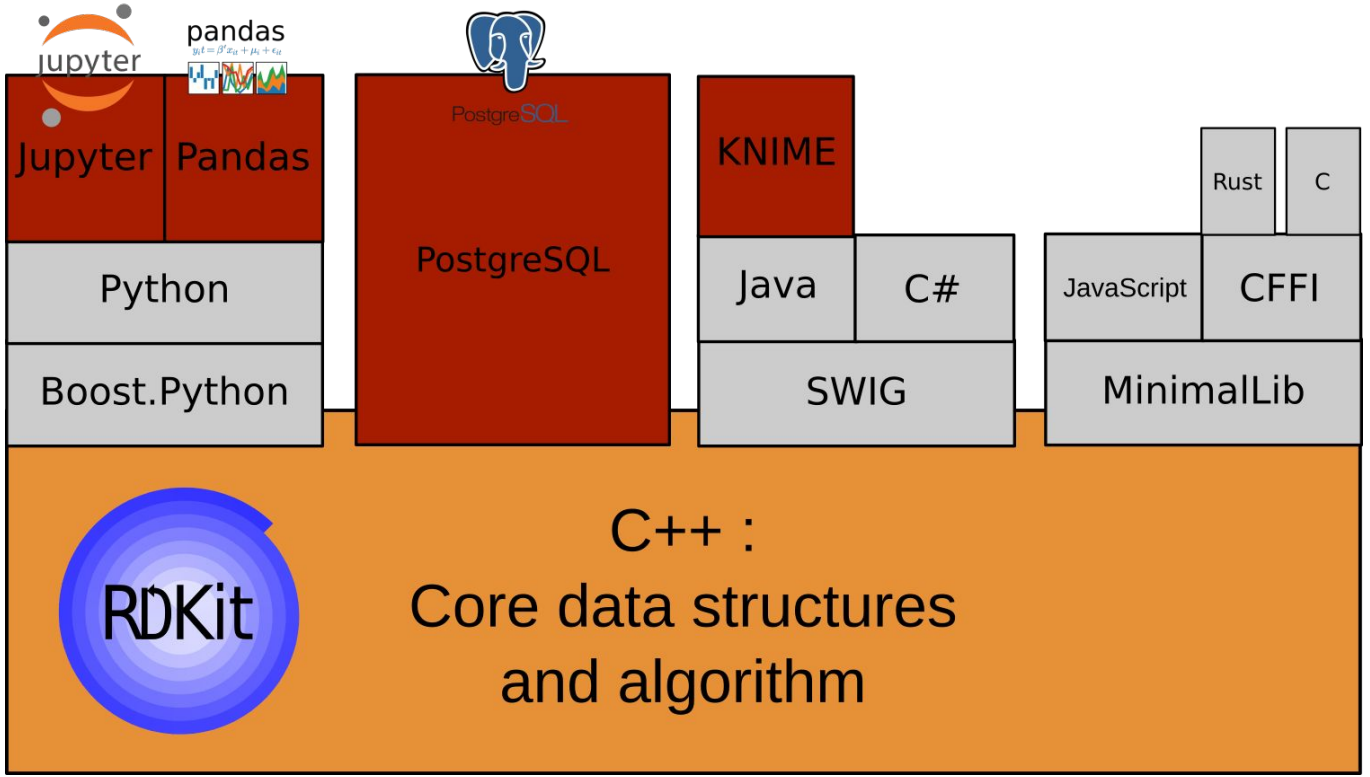
you'll use some all the time

some you'll never use and may not even know about

An open source toolkit for cheminformatics

- Business-friendly BSD license
- Core data structures and algorithms in C++
- Python 3.x wrapper generated using Boost.Python
- Java and C# wrappers generated with SWIG
- JavaScript wrappers
- CFFI interface for usage from other languages
- 2D and 3D molecular operations
- Descriptor generation for machine learning
- Molecular database cartridge for PostgreSQL
- Cheminformatics nodes for KNIME (distributed from the KNIME community site: http://www.knime.org/rdkit)

Ecosystem



Exact same implementation regardless of where you are using it from

Details

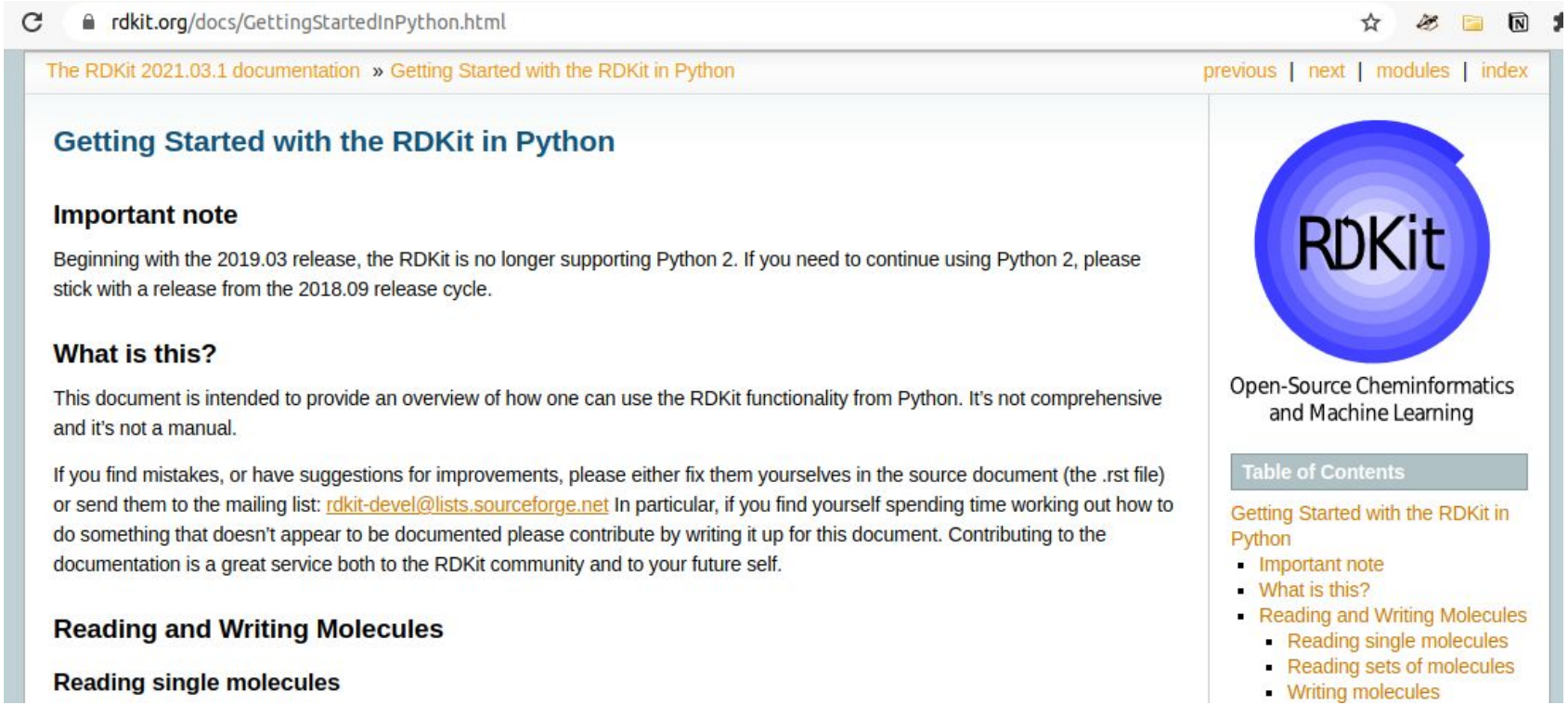
- http://www.rdkit.org
- Supports Mac/Windows/Linux
- Feature releases every 6 months
- Github (https://github.com/rdkit): Downloads, bug tracker, git repository, discussions
- Mailing lists at https://sourceforge.net/p/rdkit/mailman/, searchable archives available for rdkit-discuss and rdkit-devel
- Blog (https://greglandrum.github.io/rdkit-blog/): Tips, tricks, random stuff
- KNIME integration (https://github.com/rdkit/knime-rdkit): RDKit nodes for KNIME (also just from the community download site inside of KNIME)
- Twitter: @RDKit_org
- LinkedIn: https://www.linkedin.com/groups/8192558

Functionality[1]

- Fingerprints
- Descriptors
- Reactions
- MCS
- Enhanced stereochemistry
- Molecular standardization
- Depiction
- Diversity picking
- Tight integration with Jupyter and pandas
- Conformation generation
- 3D descriptors
- UFF and MMFF94/MMFF94S
- Open3D Align
- Feature map vectors
- Pharmacophore embedding

[1]A not-quite-random selection

Documentation



Documentation

Getting started in Python

Reading sets of molecules

Groups of molecules are read using a Supplier (for example, an rdkit.Chem.rdmolfiles.SDMolSupplier or a rdkit.Chem.rdmolfiles.SmilesMolSupplier):

```
>>> suppl = Chem.SDMolSupplier('data/5ht3ligs.sdf')
>>> for mol in suppl:
...   print(mol.GetNumAtoms())
...
20
24
24
26
```

You can easily produce lists of molecules from a Supplier:

```
>>> mols = [x for x in suppl]
>>> len(mols)
4
```

or just treat the Supplier itself as a random-access object:

```
>>> suppl[0].GetNumAtoms()
20
```

Two good practices when working with Suppliers are to use a context manager and to test each molecule to see if it was correctly read before working with it:

```
>>> with Chem.SDMolSupplier('data/5ht3ligs.sdf') as suppl:
...   for mol in suppl:
...     if mol is None: continue
...     print(mol.GetNumAtoms())
...
20
24
24
26
```

Documentation

Cookbook

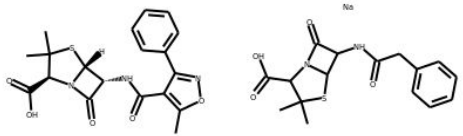
Black and White Molecules

Author: Greg Landrum
Source: https://gist.github.com/greglandrum/d85d5693e57c306e30057ec4d4d11342
Index ID#: RDKitCB_1
Summary: Draw a molecule in black and white.

```
from rdkit import Chem
from rdkit.Chem.Draw import IPythonConsole
from rdkit.Chem import Draw
```

```
ms = [Chem.MolFromSmiles(x) for x in ('Cc1onc(-c2ccccc2)c1C(=O)N[C@@H]1C(=O)N2[C@@H](C(=O)O)C(C)(
Draw.MolsToGridImage(ms)
```

```
IPythonConsole.drawOptions.useBWAtomPalette()
Draw.MolsToGridImage(ms)
```



Documentation

Reference

```
>>> Chem.MolFromSmiles('CC(=O)OC').GetSubstructMatches(Chem.MolFromSmarts('[z{1-}]'))
((1,), (4,))
>>> Chem.MolFromSmiles('CC(=O)OC').GetSubstructMatches(Chem.MolFromSmarts('[D{2-3}]'))
((1,), (3,))
>>> Chem.MolFromSmiles('CC(=O)OC.C').GetSubstructMatches(Chem.MolFromSmarts('[D{-2}]'))
((0,), (2,), (3,), (4,), (5,))
```

SMARTS Reference

Note that the text versions of the tables below include some backslash characters to escape special characters. This is a wart from the documentation system we are using. Please ignore those characters.

Atoms

Primitive	Property	"Default value"	Range?	Notes
a	"aromatic atom"			
A	"aliphatic atom"			
d	"non-hydrogen degree"	1	Y	extension
D	"explicit degree"	1	Y	
h	"number of implicit hs"	>0	Y	
H	"total number of Hs"	1		
r	"size of smallest SSSR ring"	>0	Y	
R	"number of SSSR rings"	>0	Y	
v	"total valence"	1	Y	
x	"number of ring bonds"	>0	Y	
X	"total degree"	1	Y	
z	"number of heteroatom neighbors"	>0	Y	extension
Z	"number of aliphatic heteroatom neighbors"	>0	Y	extension

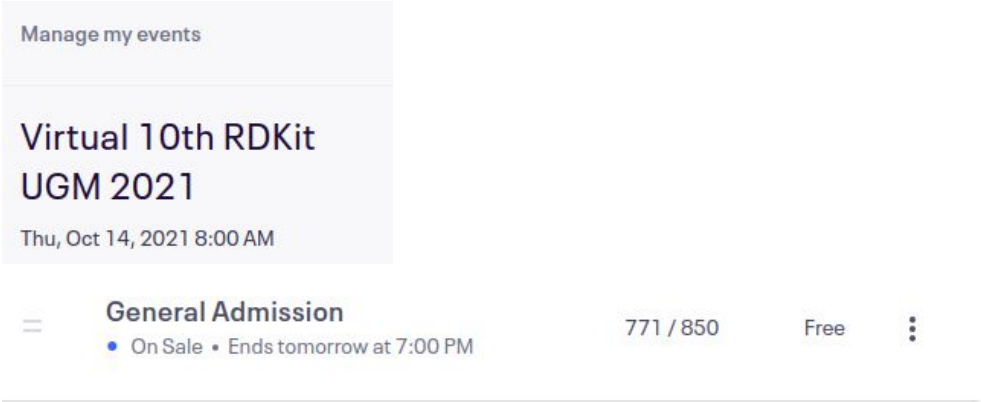
Support

- Web searches
- rdkit-discuss mailing list
- Github discussions

- Commercial support

Community

- Mailing lists: ~600 messages to rdkit-discuss from 2020.10.11- 2021.10.12
- Google scholar: >1200 hits for "rdkit" in 2020, >1400 so far in 2021
- Searching github for `from rdkit import Chem` returns >30000 code results
- Each of the last eight UGMs at capacity with 40-100+ attendees



Usage in other open-source projects (updated 2021)

- Shape-IT - shape-based alignment
- ChEMBL Structure Pipeline - ChEMBL protocols used to standardise and salt strip molecules.
- FPSim2 - Simple package for fast molecular similarity searches.
- https://datamol.io/ - A Python library to intuitively manipulate molecules.
- Scopy - Python library for desirable HTS/VS database design
- stk (docs, paper) - a Python library for building, manipulating, analyzing and automatic design of molecules.
- OpenFF - Open source approach for better force fields
- gpusimilarity - GPU implementation of fingerprint similarity searching
- Samson Connect - Software for adaptive modeling and simulation of nanosystems
- mol_frame - Chemical Structure Handling for Dask and Pandas DataFrames
- mmpdb 2.0 - matched molecular pair database generation and analysis
- CheTo - Chemical topic modeling
- OCEAN - web-tool for target-prediction of chemical structures which uses ChEMBL as datasource
- Coot - software for macromolecular model building, model completion and validation
- DeepChem - deep learning toolkit for drug discovery
- sdf2ppt - Reads an SDFile and displays molecules as image grid in powerpoint/openoffice presentation.
- chemfp
- PYPL - Simple cartridge that lets you call Python scripts from Oracle PL/SQL.
- WONKA - Tool for analysis and interrogation of protein-ligand crystal structures
- OOMMPPAA - Tool for directed synthesis and data analysis based on protein-ligand crystal structures
- chemicalite - SQLite integration for the RDKit
- django-rdkit - Django integration for the RDKit
- … more ...

Usage in commercial tools

- Cresset Software
- Dalke Scientific Software
- NextMove Software
- Schrödinger
- SCM
- Wolfram Research

Disclaimer: this info is from public statements made by people from those companies.
I almost certainly have forgotten someone

Usage in online tools/resources

- ChEMBL
- ZINC
- Google Patents
- PDBe
- Enamine
- TeachOpenCADD

Disclaimer: this info is from public statements made by people associated with those projects. I almost certainly have forgotten someone

Acknowledgements

- Everyone who has contributed code, questions, answers, bug reports, etc
- People who have funded RDKit development (directly or indirectly)
- The others in our community who've been pushing the idea and adoption of open source

Thanks!

Let’s actually use the RDKit

The notebook is here:

https://github.com/greglandrum/AIDD_RDKit_Tutorial_2021

If you just want to follow along without installing anything, you can use this mybinder link:

https://tinyurl.com/RDKit-AIDD-2021

or

https://mybinder.org/v2/gh/greglandrum/AIDD_RDKit_Tutorial_2021/HEAD