RDKit: State of the Toolkit

2024 UGM edition

Greg Landrum
@dr_greg_landrum@sciencemastodon.com
@greg_landrum.bsky.social

What’s new in the last year?

That comes later :-)

First let’s talk about the state of the toolkit.

Adoption / usage

Unlike with web apps or commercial software, this is tricky to figure out with open source tools, but let’s try.

Usage: Conda install counts (last release)

Type	Size	Name	Downloads
conda	6.3 MB	linux-64/rdkit-2024.03.5-py312h7b4b7d0_3.conda	11530
conda	6.1 MB	linux-64/rdkit-2024.03.5-py310h57e35d3_3.conda	6759
conda	6.4 MB	linux-64/rdkit-2024.03.5-py311h845bd92_3.conda	5210
conda	6.1 MB	linux-64/rdkit-2024.03.5-py38h890d753_2.conda	2374
conda	6.1 MB	linux-64/rdkit-2024.03.5-py38h890d753_3.conda	1983
conda	6.1 MB	linux-64/rdkit-2024.03.5-py39hc1ff0a3_3.conda	1978
conda	6.1 MB	linux-64/rdkit-2024.03.5-py310h57e35d3_2.conda	1810
conda	36.4 MB	linux-64/rdkit-2024.03.5-py310h5dbf55c_0.conda	1415
conda	6.4 MB	linux-64/rdkit-2024.03.5-py311h845bd92_2.conda	1345
conda	6.3 MB	linux-64/rdkit-2024.03.5-py312h7b4b7d0_2.conda	1341
conda	5.1 MB	osx-arm64/rdkit-2024.03.5-py312h619ea94_3.conda	1211
conda	5.3 MB	osx-64/rdkit-2024.03.5-py310h926f623_3.conda	1177
conda	4.1 MB	win-64/rdkit-2024.03.5-py312h9d9823f_3.conda	1162

Partial data. Unfortunately the condastats package no longer works

Usage: nuget downloads

nuget Packages Upload Statistics Documentation Downloads Blog

Sign in

Search for packages...

RDKitDotNetCore 2024.6.5

.NET Standard 2.0

ⓘ This package has a SemVer 2.0.0 package version: 2024.6.5+b8dbf52625aca4cf41ce776885cbbb4d893daffb.

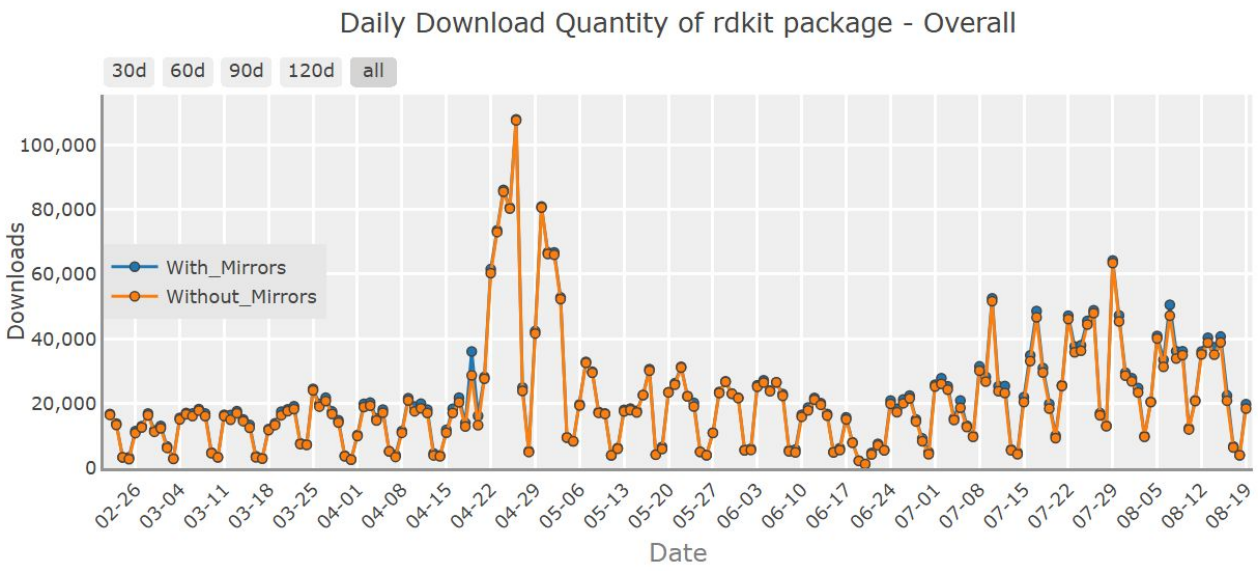
Downloads

Full stats →

Total	**4.0K**
Current version	**76**
Per day average	**2**

Thanks to Gareth Jones for setting this up

Usage: PyPi



Thanks to Chris Kuenneth for getting the pypi installs set up!

Last 120 days of data from https://pypistats.org/packages/rdkit-pypi

rdkit-js usage:

rdkit.js

A powerful cheminformatics and molecule rendering toolbelt for JavaScript

npm v2024.3.5-1.0.0

downloads 4.5k/week | downloads 20k/month | downloads 184k/year | total downloads 276k

Azure Pipelines succeeded | license BSD-3-Clause | DOI 10.5281/zenodo.13469390

Explore the docs »

Report Bug · Request Feature · Star Repository

Install

```
> npm i @rdkit/rdkit
```

Repository

github.com/rdkit/rdkit-js

Homepage

www.rdkitjs.com

Weekly Downloads

4’481

Version

2024.3.5-1.0.0

License

BSD-3-Clause

Thanks to Michel Moreau for getting this set up!

Beyond download counts: what about other approaches for looking at adoption?

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

Usage in commercial tools

- Amazon Web Services
- Collaborative Drug Discovery
- Cresset Software
- Dalke Scientific Software
- Datagrok
- Glysade
- MedChemica
- 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

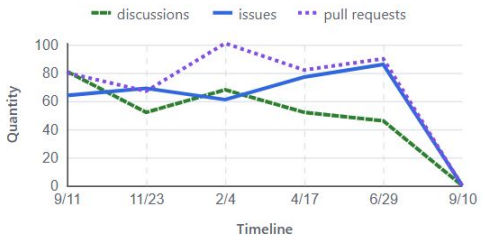
Github community stats

Community insights

Period: Last year

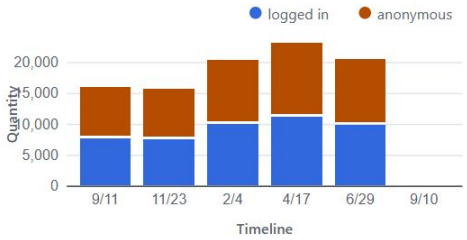
Contribution activity

Count of total contribution activity to Discussions, Issues, and PRs



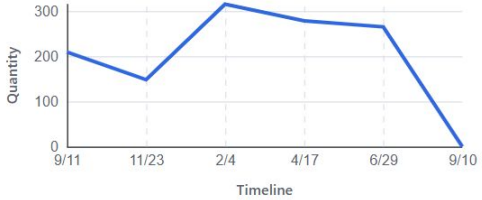
Discussions page views

Total page views to Discussions segmented by logged in vs anonymous users.



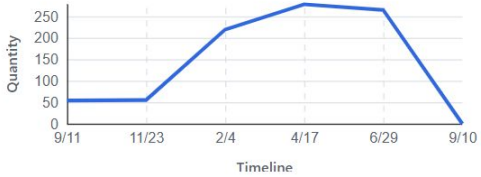
Discussions daily contributors

Count of unique users who have reacted, upvoted, marked an answer, commented, or posted in the selected period.



Discussions new contributors

Count of unique new users to Discussions who have reacted, upvoted, marked an answer, commented, or posted in the selected period.



Other adoption measures

- Mailing lists: ~100 messages to rdkit-discuss from 2023.09 - 2024.08
- Google scholar: >3200 hits for "rdkit" in 2023, >3000 so far in 2024
- Searching github for `”from rdkit import Chem”` returns >35000 code results
- UGM attendance!

Sustainability: the bus problem

https://commons.wikimedia.org/wiki/File:Postauto_susten.jpg

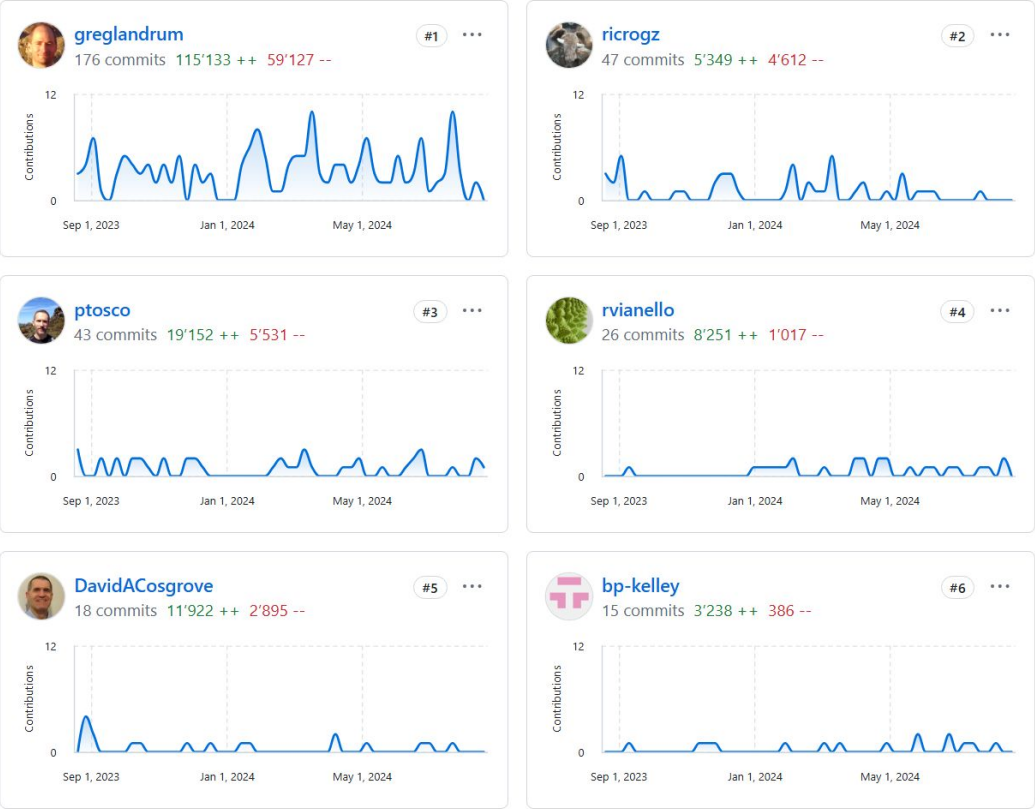
Sustainability: the bus problem

RDKit maintainers:

- Greg
- Brian Kelley (Relay Therapeutics)
- Ricardo Rodriguez Schmidt (Schrödinger)
- Paolo Tosco (Novartis)

Most frequent code contributors in the last year

From 20 Aug 2023 to 18 Aug 2024



Merged pull request contributors in the last year

['AAriam', 'Adam-maz', 'AnnaBruenisholz', 'BenoitClaveau', 'DavidACosgrove', 'JanCBrammer', 'JuniorSen', 'KollinRT', 'MarioAndWario', 'Old-Shatterhand', 'PatrickPenner', 'RPirie96', 'StLeonidas', 'UnixJunkie', 'ankane', 'apahl', 'bertiewooster', 'bjonnh-work', 'bp-kelley', 'buerbaumer', 'cdvonbargen', 'cho-m', 'christophhillisch', 'cthoyt', 'd-b-w', 'ddgunizar', 'dehaenw', 'df7cb', 'e-kwsm', 'ergo70', 'esiaero', 'frakyk', 'fwaibl', 'ghost', 'github-actions[bot]', 'greglandrum', 'iwyoo', 'johnmay', 'jones-gareth', 'kevingreenman', 'lounsbrough', 'mcs07', 'nbehrnd', 'nmaeder', 'padix-key', 'pechersky', 'ptosco', 'rachelnwalker', 'richardjgowers', 'ricrogz', 'rvianello', 'syedzayyan', 'tadhurst-cdd', 'tgaudin', 'vfscalfani', 'vslashg', 'whosayn']

That's 57 different people

Maintenance work in the last year

We started tracking maintenance/cleanup work with the 2019.09 release.

For the 2024.03 and 2024.09 releases, there have been >50 “cleanup” issues/PRs merged:

Greg Landrum 24
Riccardo Vianello 9
Paolo Tosco 8
Anna Brünisholz 6
nmaeder 4
Ricardo Rodriguez 3
Eisuke Kawashima 2
tadhurst-cdd 1
Yakov Pechersky 1
Théophile Gaudin 1
Michael Cho 1
Matt Swain 1
Jan C. Brammer 1
Hussein Faara 1
Christoph Berg 1
Brian Kelley 1

Roadmap

Future work tends to be determined by what's needed for active projects or requests that come out of the community. So there's not much of a roadmap.

Still, some parts of the way forward are pretty obvious...

✔ ~~Making sure all the pieces required to build a good compound registration system are there~~

Making sure all the pieces required to build a good corporate chemical database are there

Better support for polymers and organometallics

Performance improvements

Ongoing improvements to the conformer generator

Ongoing refactoring and code cleanup

Taking big steps forward…

Some things are hard...

Technology changes (i.e. taking advantage of new C++ or Python versions) is tricky: which operating systems/compilers are people using?

Is it safe to remove old code that seems peripheral or redundant with functionality provided better by other packages?

There are some larger API changes to clean up old mistakes and improve performance and safety that it would be nice to make.

We really, really want to avoid the Python 2/Python 3 situation, so we can’t just make arbitrary changes.

… what we're doing about it

Try to minimize hard external dependencies

Be conservative about language versions/features

Announce deprecations at least one major release in advance

"Backwards incompatible changes" doc

Version-compatibility report (for commercial support customers)

Changing the RDKit release model

Motivation: make new functionality available sooner

Previous:

- Feature releases twice a year, e.g. **2023.03**
  - Possibly including backwards-incompatible changes
- Patch releases every 4-6 weeks, e.g. 2023.03**.2**
  - Only bug fixes, but these can still change results

Changing the RDKit release model

Motivation: make new functionality available sooner

Previous:

- Feature releases twice a year, e.g. **2023.03**
  - Possibly including backwards-incompatible changes
- Patch releases every 4-6 weeks, e.g. 2023.03**.2**
  - Only bug fixes, but these can still change results

Current:

- Major releases twice a year, e.g. **2024.03**
  - Possibly including backwards-incompatible changes
- Minor releases every 4-6 weeks, e.g. 2024.03.**2**
  - Include bug fixes (can change results)
  - Include backwards-compatible new features

Possible upcoming big changes

- Fixing the Hydrogen mess
- Changing the default stereo perception algorithm

More info on these in the What’s New notebook

Summary: great stuff

- Steadily growing numbers of people using and building things with the RDKit
- Steady progress on adding new features, fixing bugs, and cleaning up old stuff

Summary: less great

- The group of people making *regular* contributions to the RDKit itself is pretty static
- A steadily growing user community is a mixed blessing: there are a bunch of unanswered (or partially answered) questions in GitHub Discussions and a similarly large number of open issues that aren't really issues

We need:

- More people involved with the code (both C++ and Python) itself
- More people actively helping in the Discussions
- Someone to help triage open issues and handle the ones that aren't actually issues

I don't know how to solve this, maybe *we* can figure it out.

Ok, enough of that, let’s look at what’s new

The notebook I’m using will be in the UGM github repo:
https://github.com/rdkit/UGM_2024