RDKit: State of the Toolkit
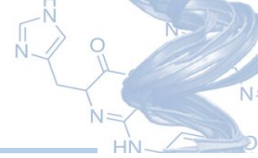
2022 UGM edition

Greg Landrum
@dr_greg_landrum
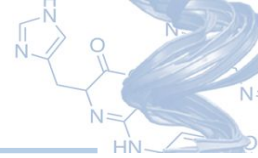
Community

The heart of any *successful* open-source project
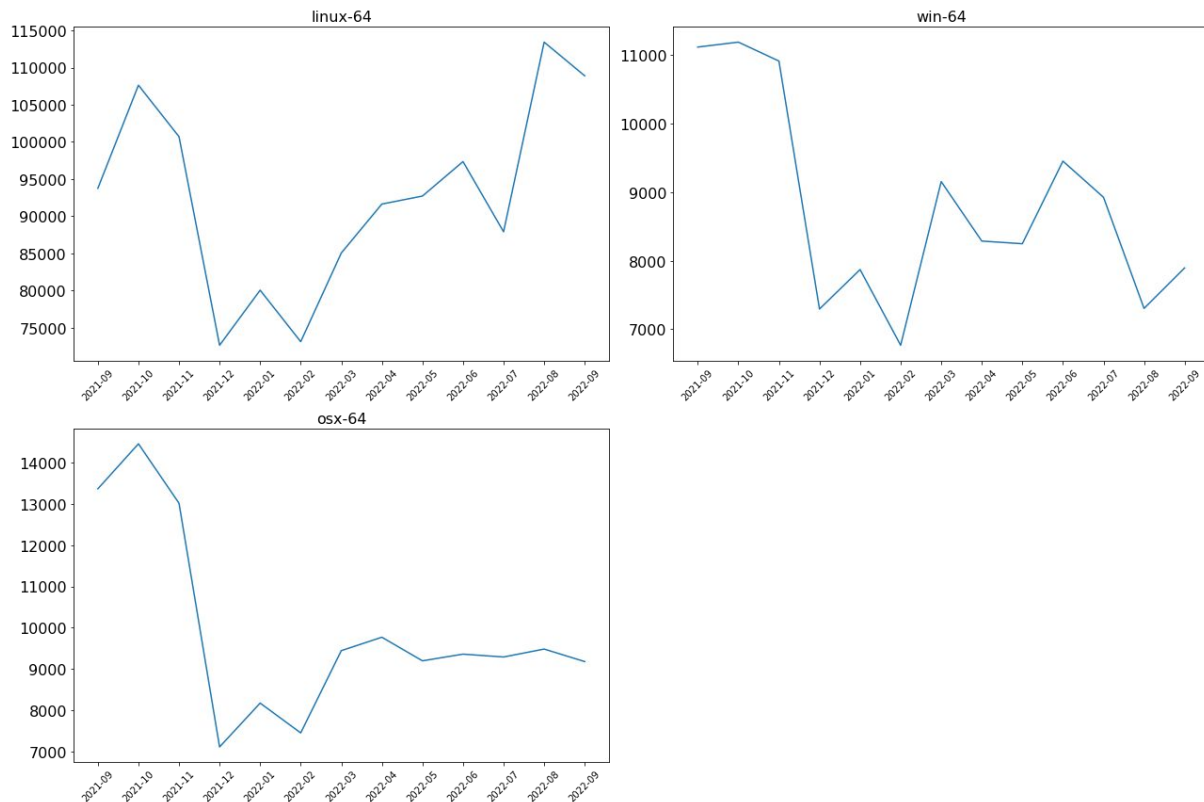
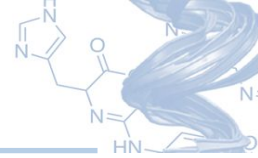# Adoption / usage

Always tricky to figure out with open source
tools, but let's try.

Usage: Conda install counts (by operating system)

Last 12 months
Data collected using the
`condastats` package

Usage: Conda install counts (by operating system)
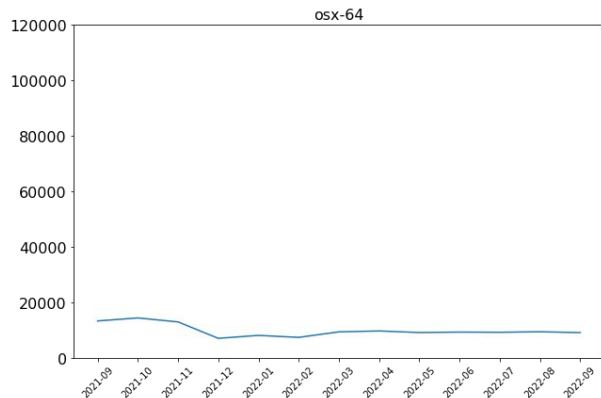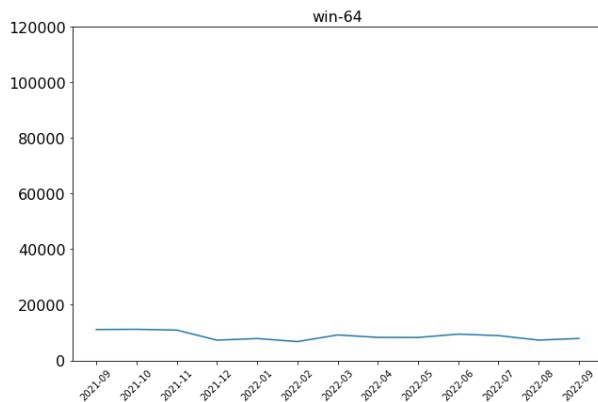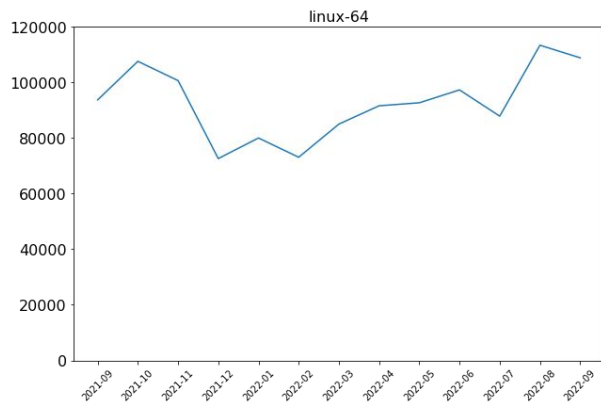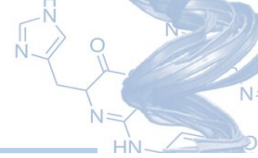

Last 12 months
Data collected using the
`condastats` package

Usage: Conda install counts (by operating system)

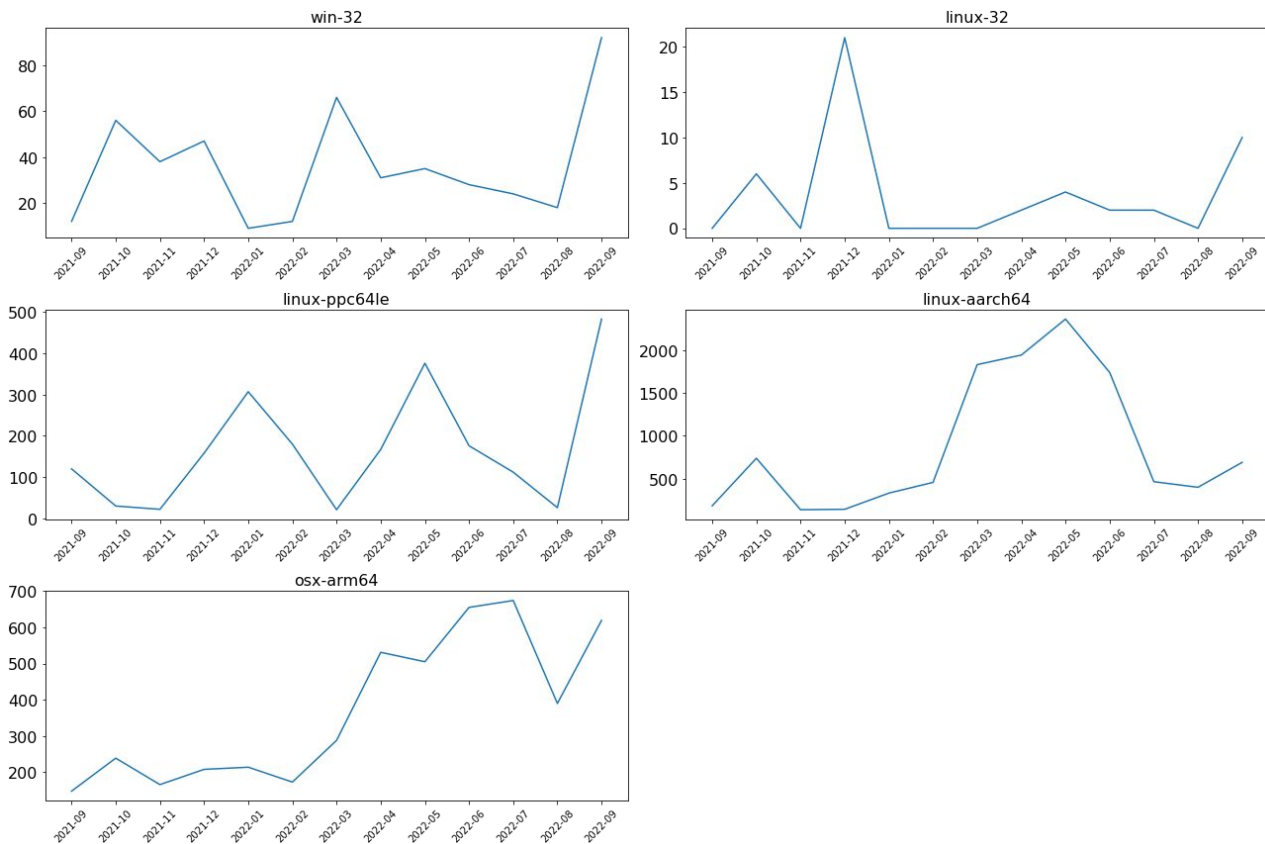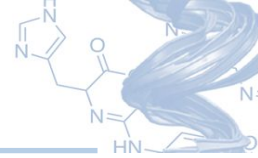Less common operating systems / hardware combos

Last 12 months
Data collected using the
`condastats` package
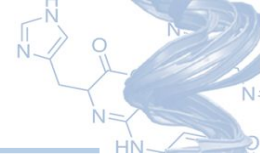
Usage: Conda install counts (by python version)
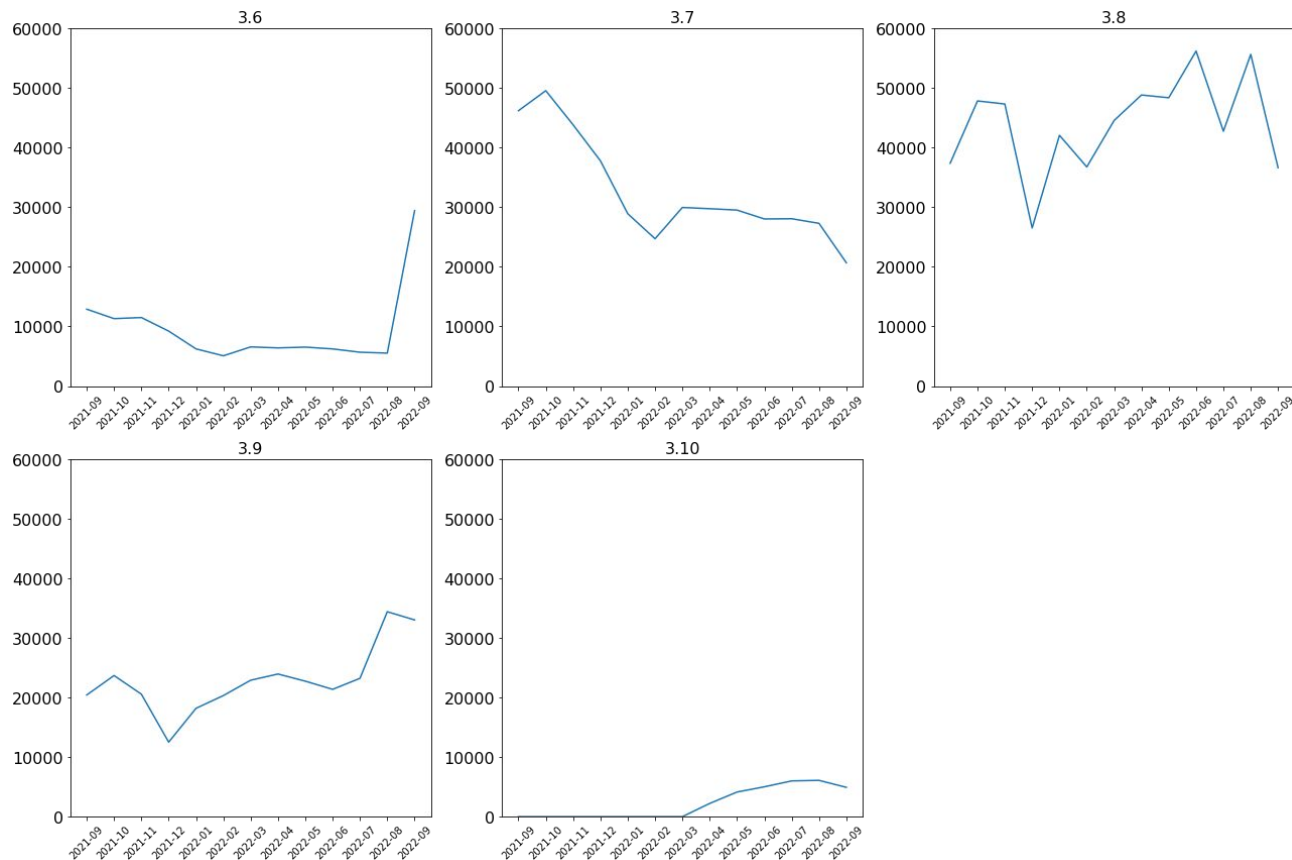
Last 12 months
Data collected using the
`condastats` package

Usage: PyPi
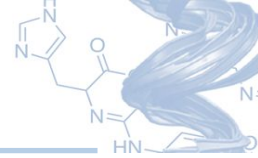
Daily Download Quantity of rdkit-pypi package - Overall

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

Something new: npm packaging

Thanks to Michel Moreau for getting this set up!

Something new: pypi packaging

Thanks to Chris Kuenneth for getting this set up!

Usage in other open-source projects (updated 2021)
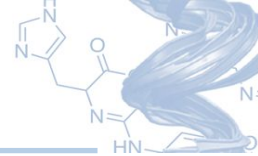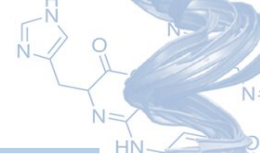
- Shape-IT - shape-based alignment
- DockOnSurf - high-throughput code to find stable geometries for molecules on surfaces
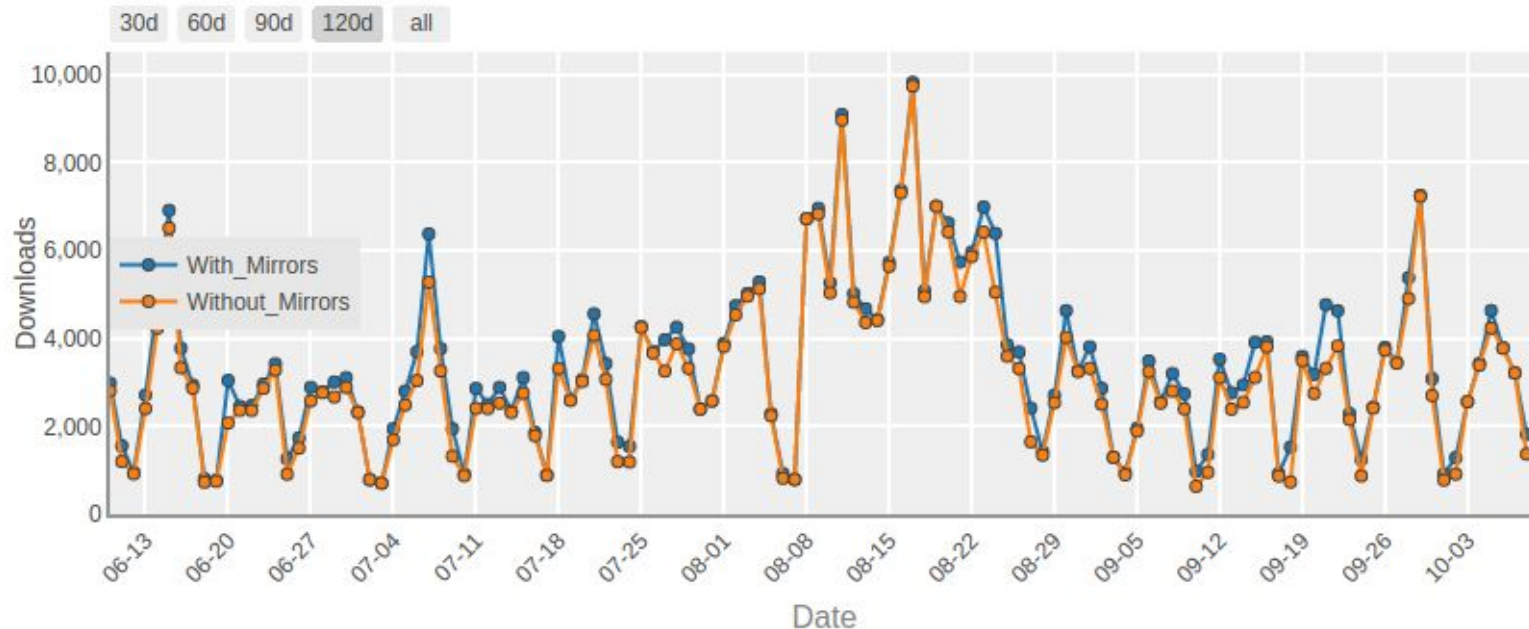- https://datamol.io/ - A Python library to intuitively manipulate molecules.
- Scopy - Python library for desirable HTS/VS database design
- ChEMBL Structure Pipeline - ChEMBL protocols used to standardise and salt strip molecules.
- FPSim2 - Simple package for fast molecular similarity searches.
- 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 ...
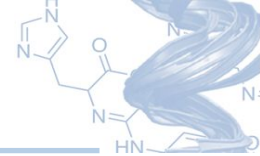
Usage in online tools/resources

- 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
- Cresset Software
- Dalke Scientific Software
- Glysade
- 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

Support

- Web searches
- Mailing list
- Github discussions

- Commercial support
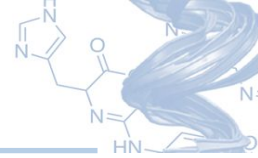
Community support

General
Welcome to RDKit Discussions!
 greglandrum

New   Top: All ▾   Label ▾   Filter ▾   **New discussion**

🔍 Search all discussions

Categories   ✎

🗨 View all discussions

🏛 Development

\# General

💡 Ideas

🗳 Polls

🙏 Q&A

👪 Show and tell

Discussions

↑ 1   🏛   **2022_09_1b1 (Q3 2022) Release**
greglandrum started 7 hours ago in Development
💬 0

↑ 1   🙏   **Wrong MMFF atom types with aromatic rings**
apiserchia asked 4 days ago in Q&A · **Unanswered**
✓ 2

↑ 1   🙏   **RDKIT MolFromSmiles**
VANIAQFB asked 7 days ago in Q&A · **Unanswered**
✓ 6
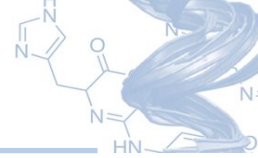
Other adoption measures

- Mailing lists: ~500 messages to rdkit-discuss from 2021.10 - 2021.10

- Google scholar: >1800 hits for "rdkit" in 2021, >1800 so far in 2022

- Searching github for `from rdkit import Chem` returns >32000 code results

- Each of the last eight in-person UGMs at capacity with 40-100+ attendees

- Last two virtual UGMs with 300+ attendees

Contributions to github issue tracker in the last year

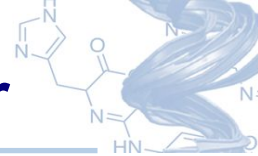AttilaVM C-nit DavidACosgrove ElricleNecro Emirali007 FabioUrbina Hong-Rui IchiruTake JLaff99 JackFang0815 Limsande LiuCMU MariaDolotova MaxDNG Mirrty OleinikovasV PigUnderRoof Polydynamical SGenheden SarahAvron SeongsangCHO Tong-Du TraceLD YuanyueLi abhik1368 adelenelai ale94mleon bddap bjonnh-work bp-kelley brilee bzoracler c-feldmann cdelv cottonkiet csu1505110121 d-b-w diogomart dskatov e-kwsm eguidotti fredludlow gedeck getuem greglandrum hogru hstern2 ignatovmg iwatobipen jaechang-hits jasondbiggs jingxingzhi jones-gareth jp-um kienerj kjelljorner kttien loluwot lucasmorin222 marcostenta matteoferla nigel-palmer nightcresendo nuzillard philopon pierred5 psireal9 ptosco rachelnwalker richardjgowers ricrogz rmrmg rvianello sagitter sauhaardac schallerdavid sebastianmdick shortydutchie sonial stephanielabouille steven-bioinfo tdudgeon thegodone vfscalfani wangyingxie wopozka xavierholt zacps
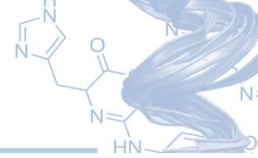
That's 88 different people

Sustainability: the bus problem
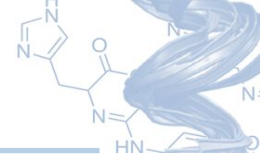
https://commons.wikimedia.org/wiki/File:Postauto_susten.jpg

Sustainability: the bus problem

RDKit maintainers:
- Greg
- Brian Kelley (Relay Therapeutics)
- Ricardo Rodriguez (Schrödinger)
- Paolo Tosco (Novartis)

Most frequent code contributors in the last year
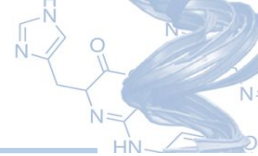
Oct 3, 2021 – Oct 3, 2022

Contributions: **Commits** ▾

greglandrum #1
137 commits 59,838 ++ 23,808 --

ptosco #2
34 commits 6,544 ++ 1,914 --

ricrogz #3
27 commits 1,831 ++ 810 --

DavidACosgrove #4
23 commits 30,727 ++ 11,080 --

bp-kelley #5
19 commits 91,786 ++ 512 --

IchiruTake #6
15 commits 1,442 ++ 1,347 --
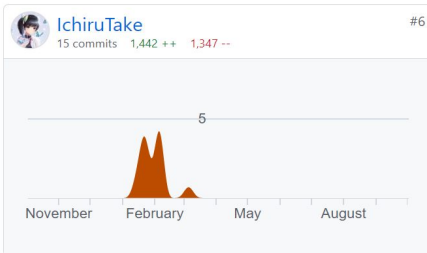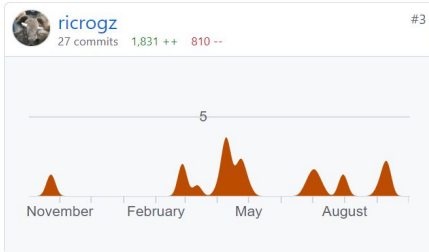
Merged pull request contributors in the last year

Richard Gowers, Niels Kristian Kjærgård Madsen, Santeri Puranen, Jeff van Santen, gedeck, David Cosgrove, Aleš Erjavec, Eisuke Kawashima, Toshiki Kataoka, Alan Kerstjens, Rachel Walker, Maciej Wójcikowski, Chris Kue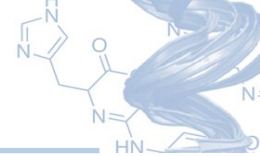nneth, Kaushalesh Shukla, Guy Rosin, Dan N, Hyeonki Hong, Paolo Tosco, Christoph Hillisch, Jon Sorenson, Michel Moreau, Gareth Jones, Ric, Alex Rebert, Kazuya Ujihara, Brian Kelley, Steve Roughley, Riccardo Vianello, Greg Landrum, Cédric Bouysset, Kevin Burk, Sreya Gogineni, Ichiru Take, Emanuele Guidotti, Ivan Tubert-Brohman, Mosè Giordano
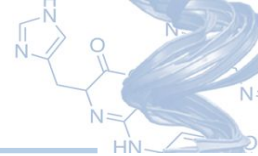
That's 36 different people

Maintenance work in the last year

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

For the 2022.03 and 2022.09 releases, there have been >50 "cleanup" issues/PRs merged:

Greg Landrum 14
Ichiru Take 13
Paolo Tosco 6
Eisuke Kawashima 5
Riccardo Vianello 3
Ric 3
David Cosgrove 2
Brian Kelley 2
Michel Moreau 1
Guy Rosin 1
Alex Rebert 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...
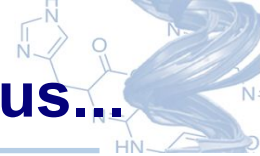
Making sure all the pieces required to build a good compound registration system are there
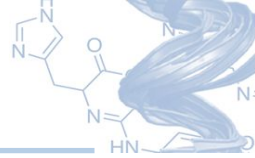
Making sure all the pieces required to build a good corporate chemical database are there

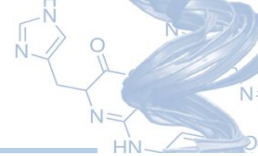Better support for polymers and organometallics

Performance improvements

Ongoing refactoring and code cleanup

We still should be able to take 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, but we can't just make arbitrary changes

Really, really want to avoid the Python 2/Python 3 situation
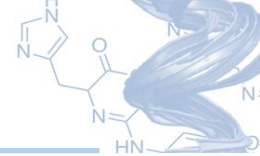
… what we're doing about it
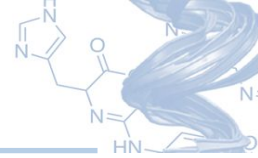
Try to minimize hard external dependencies

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)

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_2022