RDKit: State of the Toolkit

2023 UGM edition

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

What’s new in the last year?

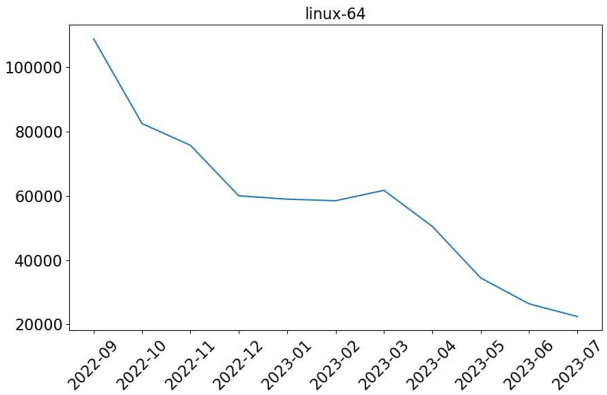
That comes later :-)

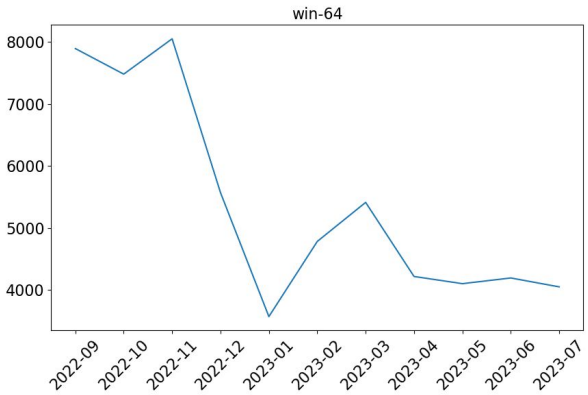
First let’s talk about the state of the toolkit.

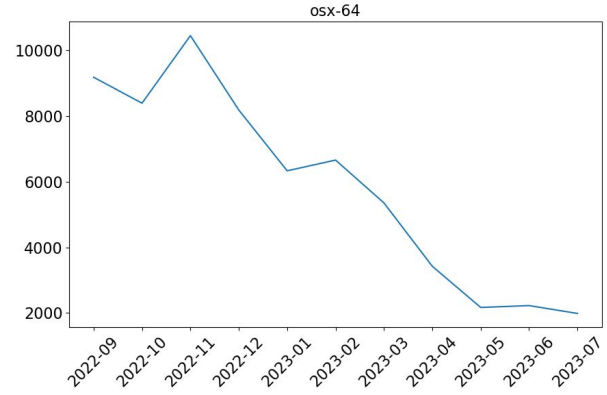
Adoption / usage

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 (by operating system)

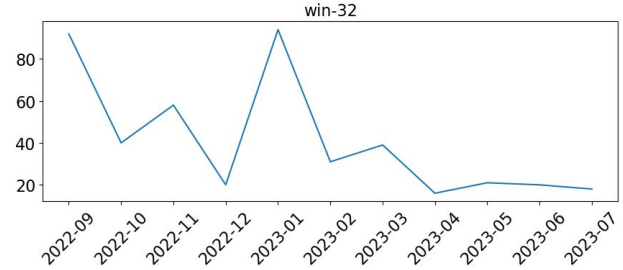


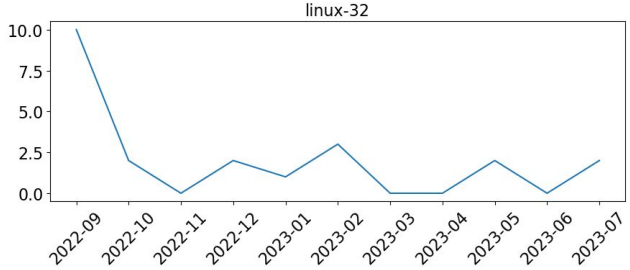


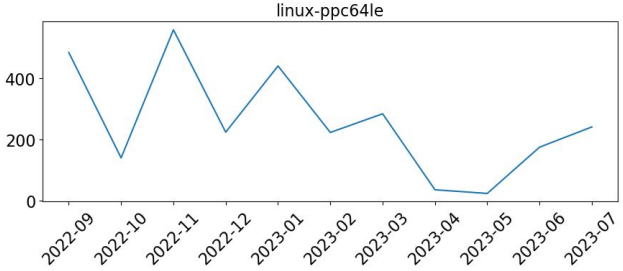


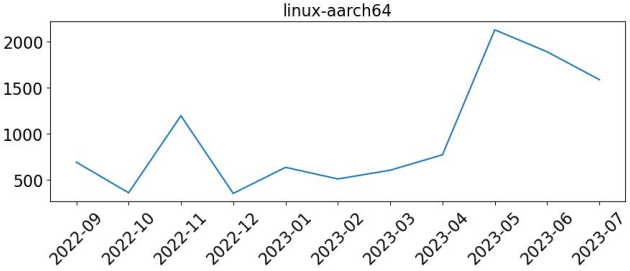
Last 12 months
Data collected using the `condastats` package

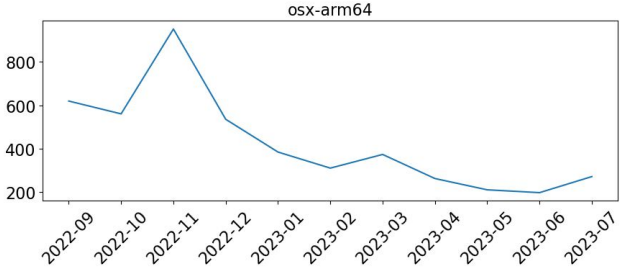
Usage: Conda install counts (by operating system)





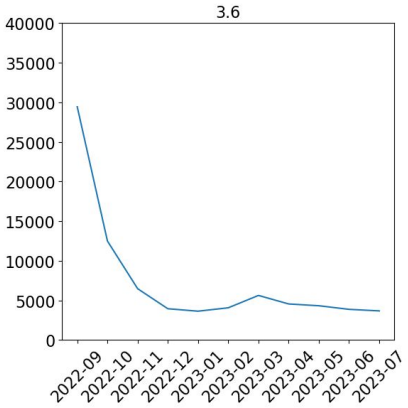


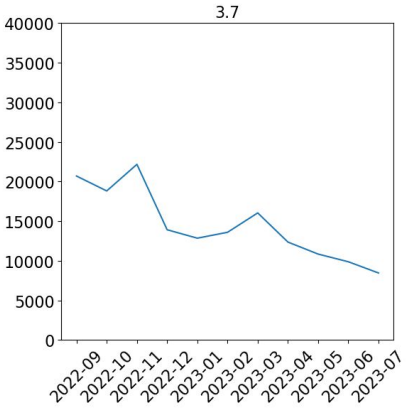


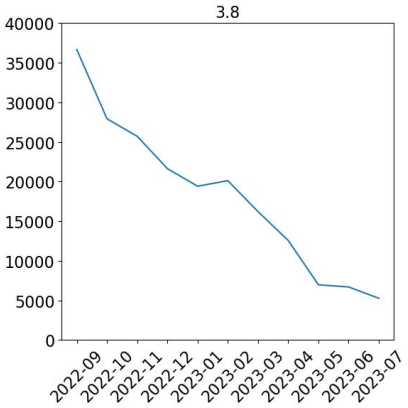


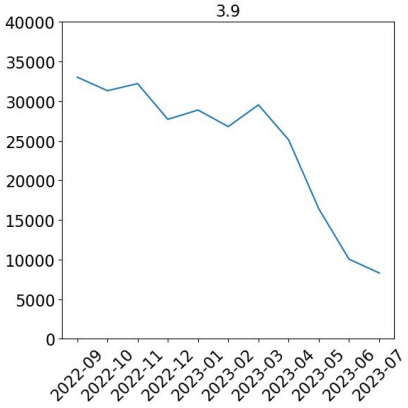
Less common operating systems / hardware combosd

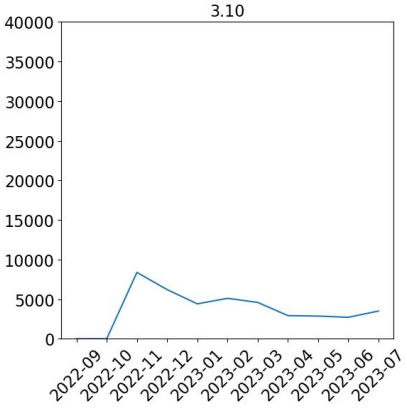
Usage: Conda install counts (by python version)





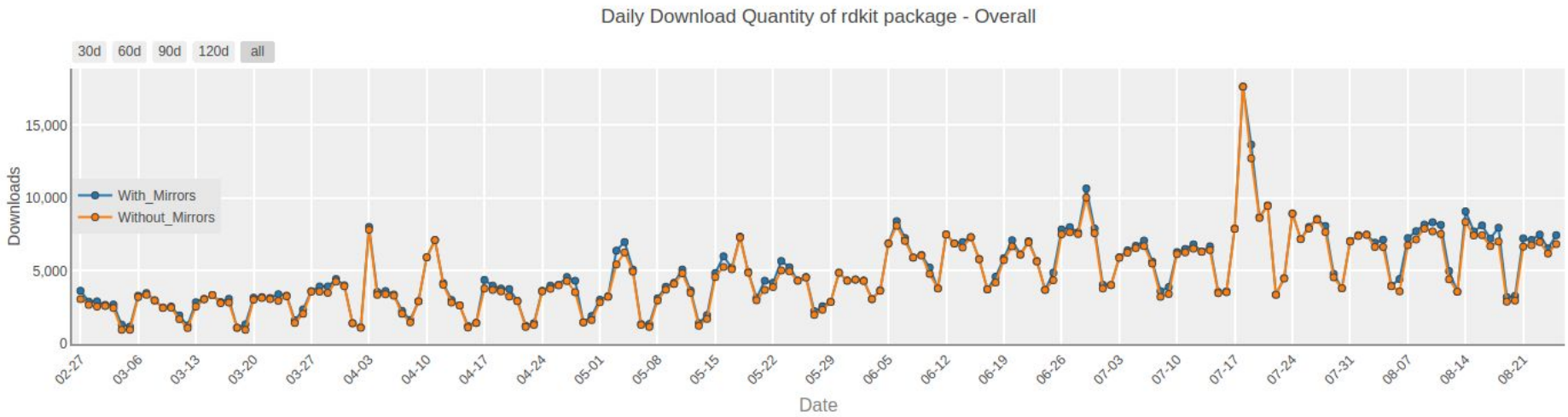






v3.11 was not available from condastats when I ran these queries

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/rdkit TS

2023.3.3-1.0.0 • Public • Published 9 days ago

Readme | Code Beta | 0 Dependencies | 2 Dependents | 57 Versions

rdkit.js

A powerful cheminformatics and molecule rendering toolbelt for JavaScript

npm v2023.3.3-1.0.0

downloads 3.2k/week | downloads 14k/month | downloads 155k/year | total downloads 202k

Azure Pipelines succeeded | license BSD-3-Clause | DOI 10.5281/zenodo.8254217

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

3,210

Version

2023.3.3-1.0.0

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 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
- 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

Other adoption measures

- Mailing lists: ~250 messages to rdkit-discuss from 2022.09 - 2023.08
- Google scholar: >2300 hits for "rdkit" in 2022, >2000 so far in 2023
- Searching github for `”from rdkit import Chem”` returns >27000 code results
- Each of the last nine in-person UGMs at capacity with 40-150 attendees

Community

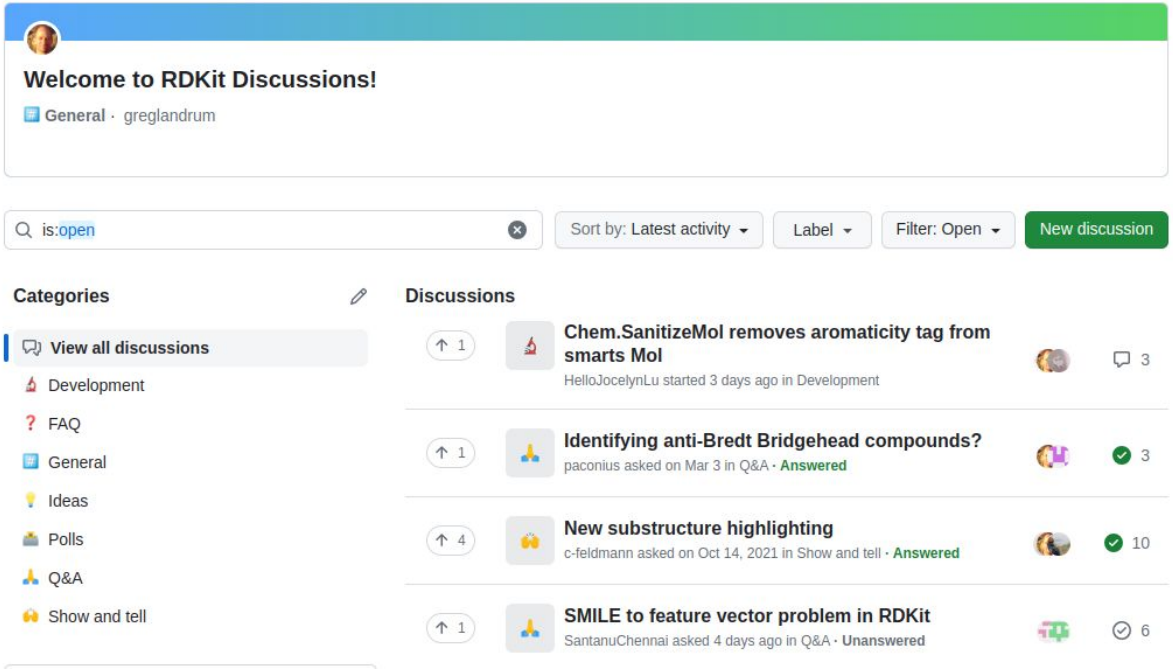
The heart of any
successful open-source
project

Support

- Web searches
- Mailing list
- Github discussions

- Commercial support

Community support



Github community stats

Community insights

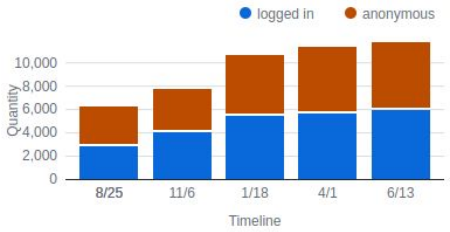
Period: Last year

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.

Contributions to github issue tracker in the last year

AlanKerstjens Arch4ngel21 AttilaVM Boilermaker14 ChemRMB CreamyLong
DavidACosgrove Efim-Shats Hikoyu Hong-Rui JLVarjo JackFang0815 KrisVolkova
Leocontreas LiuCMU MariaDolotova OleinikovasV SPKorhonen StLeonidas UnixJunkie
ValeryPolyakov andresilvapimentel autodataming bddap ben-ikt bjonnh-work bp-kelley
bradakta bwolfe-benchling bzoracler cdvonbargen chloechow chmnk dangthatsright
davidegraff davidoskky diogomart eguidotti eloyfelix gayverjr gedeck giordano greglandrum
jasondbiggs jepdavidson jmyounk jones-gareth juius kienerj koalaaaaaaaaaa kovalp
lavoisiermod lhyuen liushili0319 lounsbrough lpravda luwei0917 maclandrol mapengsen
mcneela mpagni12 oleksii-dukhno-bayer pablo-arantes peastman ptosco pwging13
rachelnwalker radchenkods rmrmg roccomoretti sagitter sakoht shortydutchie
sitanshubhunia spparel trallnag vfscalfani zpincus

That's 78 different people

How you can contribute/help: non-developers

- Use the code in your own projects and provide feedback:
  - Good bug reports
  - Ideas for improvements
  - Positive feedback via the mailing list/Github discussions
- Answering questions on the mailing list/Github discussions
- Improve the documentation
  - in-code documentation
  - the “Getting started in Python” book
  - the “RDKit Book” reference
  - the “Cookbook”
- Write blog posts (either your own or for the RDKit blog)
- Contribute interesting scripts/libraries for the Contrib folder
- Pay someone else to work on RDKit code[1]

[1] It’s generally a good idea to check with Greg or one of the maintainers before adding significant new functionality.

Sustainability: the bus problem

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

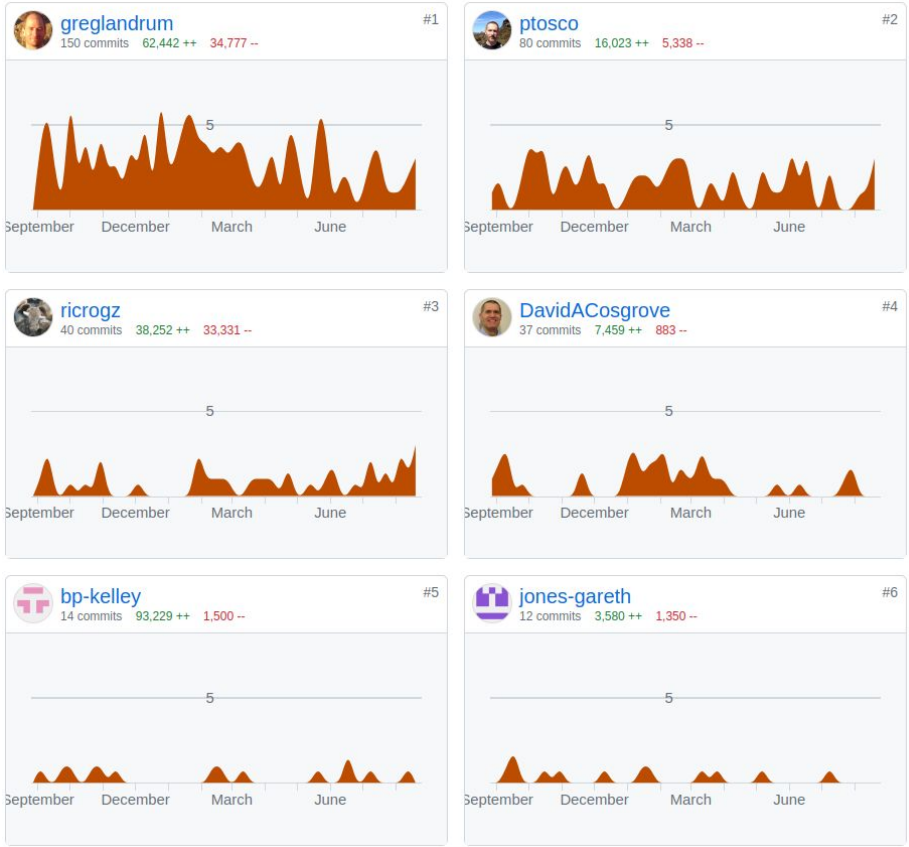
Sustainability: the bus problem

RDKit maintainers:

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

Most frequent code contributors in the last year

Aug 26, 2022 – Aug 26, 2023



Merged pull request contributors in the last year

DavidACosgrove EmmaHovhannisyan2 HalflingHelper JLVarjo OleinikovasV PatWalters RPirie96 SiPa13 alexwahab althonos autodataming bertiewooster bjonnh-work bp-kelley cdvonbargen clarezhu d-b-w dessygil e-kwsm e-mayo eloyfelix fwaibl gedeck giordano github-actions[bot] gosreya greglandrum hadim irenazra jasondbiggs jkhales jminuse jones-gareth juius kazuyaujihara kmnis kuelumbus maksbotan manangoel99 markf94 mbanck mwojcikowski philopon proteneer ptosco rachelnwalker ricrogz roccomoretti rvianello santeripuranen sroughley swamidass tadhurst-cdd thegodone thomp-j timothyngo vandan-revanur vedranmiletic vfscalfani yy692

That's 60 different people

Maintenance work in the last year

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

For the 2023.03 and 2023.09 releases, there have been >45 “cleanup” issues/PRs merged:

Greg Landrum 15
Paolo Tosco 13
Ric 5
David Cosgrove 3
Riccardo Vianello 2
github-actions[bot] 1
Vedran Miletić 1
Rocco Moretti 1
Juuso Lehtivarjo 1
Jonathan Bisson 1
Iren Azra Azra Coskun 1
Gareth Jones 1
Eisuke Kawashima 1
Dan N 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)

Thinking about changing the RDKit release model

Motivation: make new functionality available sooner

Current:

- 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

Possible alternative:

- Major releases twice a year, e.g. **2023.09**
  - Possibly including backwards-incompatible changes
- Minor releases every 4-6 weeks, e.g. 2023.09.**2**
  - Include bug fixes (can change results)
  - Include backwards-compatible new features

State of the RDKit?