Session:
CINF: Successful Projects Fueled by Open Source Tools

Title:
How the rcdk Enbables Open Source Cheminformatics in R: From fingerprints to mass spectra

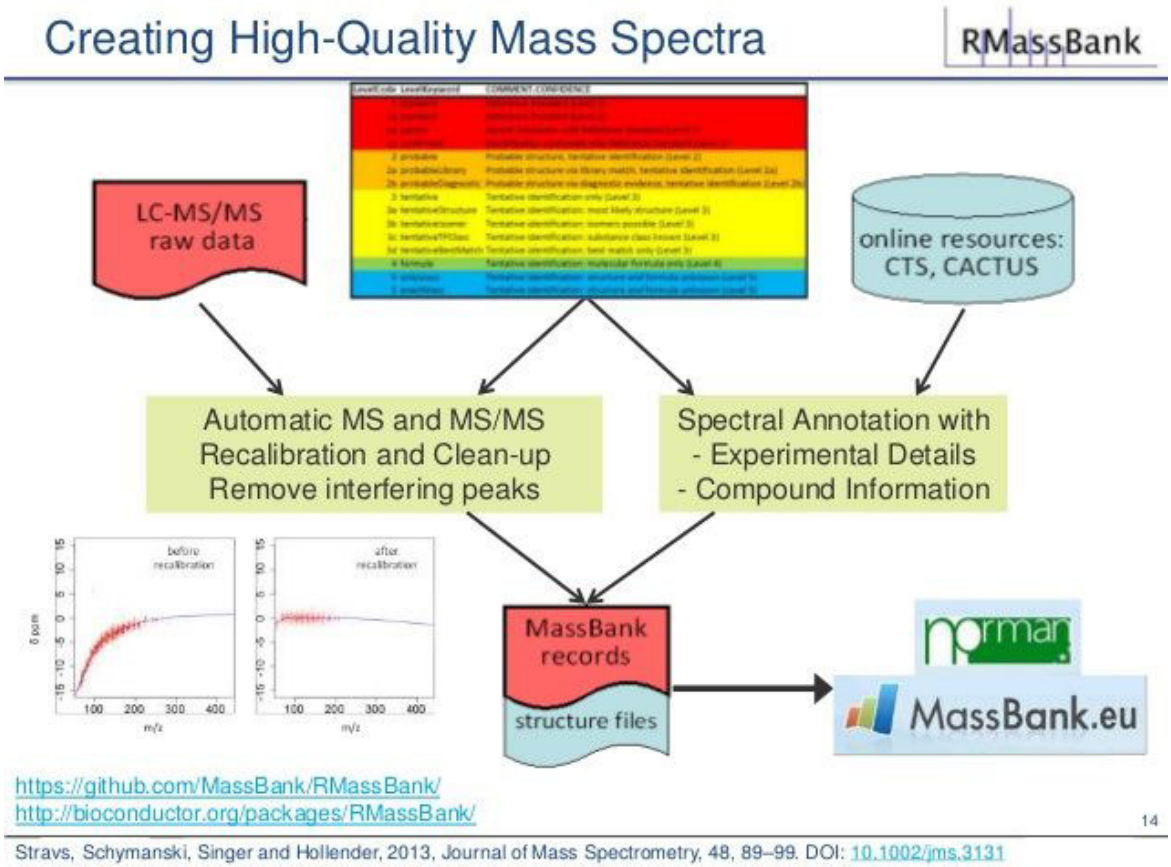
Rajarshi Guha, Emma L. Schymanski, Tobias Schulze, Michael A. Stravs

RG: Vertex Pharmaceuticals, 50 Northern Ave, Boston, MA 02210
ELS: Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6, avenue du Swing, L-4367 Belvaux, Luxembourg. emma.schymanski@uni.lu
TS: UFZ - Helmholtz Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany. tobias.schulze@ufz.de
MAS: Eawag: Swiss Federal Institute of Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland, michael.stravs@eawag.ch

The CDK is a Java library for cheminformatics and is employed in a variety of applications. Given that many problems in cheminformatics involve analysis of collections of molecules, accessing cheminformatics in a statistical modelling environment enables flexible modelling and analysis of chemical information. R is an open source statistical modelling and programming environment and provides extensive support for data munging and machine learning, but does not support cheminformatics natively. As a result, a number of packages have been developed to provide such capabilities within the R environment. In this talk we discuss the rcdk, an R package that makes the CDK accessible from within R, After a brief discussion of the design of the rcdk [1], we briefly survey the ecosystem of packages that has grown up around rcdk to enable chemical data science within R. We then describe how these efforts have fed into larger, originally third-party R-based projects that require cheminformatics functionality, using the example of RMassBank [2]. Building on the rcdk functionality, the automated cheminformatic and spectral curation workflow has enabled the release of over 16,000 open mass spectra to the open spectral library MassBank [3] - which itself surfaces the structural information to the public using CDK libraries. The seamless integration between rcdk and the CDK enables consistent cheminformatics functionality for end users and developers alike. These once independent projects have now grown into a (spontaneous) geographicaly distributed, collaborative development between the CDK, rcdk, RMassBank and MassBank teams, driven mainly by communications via Github, showing that the ability to combine open source projects leads to products that are more than the sum of their parts.

[1] https://github.com/CDK-R/cdkr
[2] https://github.com/MassBank/RMassBank/
[3] www.massbank.eu

Extra notes just for the record … not for inclusion in the abstract …



https://www.slideshare.net/EmmaSchymanski/environmental-cheminformatics-for-unknown-id-uc-davis-nov-2018
This was not quite the slide I was after, but I'm thinking a modified version of this is one of the "changeover" slides towards (R)MassBank

Emma's thoughts / contributions /ideas (left below for the record)

- RMassBank built on core functionality available in the rcdk to do cheminformatic manipulations (formula calculation, mass calculation, etc) and built a suite of extensions into RMassBank
- The fact that this was available in R meant we could make a package compliant with BioConductor (we were using a method that was fast, efficient, but licensed and would have been an external dependency)
- We also build on Babel functionality (but irrelevant, I think, for your abstract).
- What has happened since then: rcdk is effectively indisposable now for mass spectral workflows for small molecules (metabolomics, environmental, forensics, toxicology), we have a huge ecosystem of mass spectral packages (some coming from proteomics even) and you (rcdk) are the interface to the cheminformatics

- With the depiction, you've added the visualisation layer that is essential for us to view structures in "high throughput" workflows
- What it's also now enabled is that through rcdk we can reproduce in R what other methods can reproduce in Java with the "true" CDK - this has opened up incredible interoperability for us and extended beyond R to MassBank and MetFrag (https://msbi.ipb-halle.de/MetFragBeta/) where we create records in RMassBank and render the structures live from the SMILES we use in RMassBank in the real MassBank: https://massbank.eu/MassBank/

What this means in short for your abstract: you may not have much room, but what you could do would be to adjust the last sentence to specifically use perhaps the availability of cheminformatics to the mass spectral community as the major case study of "the ability to combine open source projects leads to products that are more than the sum of their parts"
I have some awesome slides showing this, with all the packages we use in them (the rcdk is not obvious, but I can make it so).
Here's a "showcase" presentation, maybe the content doesn't mean much to you without background, but this is, effectively, all done/available in R ..
https://www.slideshare.net/EmmaSchymanski/setac-rome-nontarget-screening-for-chemical-discovery
And here's one with more details
https://www.slideshare.net/EmmaSchymanski/environmental-cheminformatics-for-unknown-id-uc-davis-nov-2018
(I'm thinking especially e.g. a modified version of slide 51 in that talk (in green are mostly R packages, rcdk isn't obvious but it's a cornerstone of many of them …).