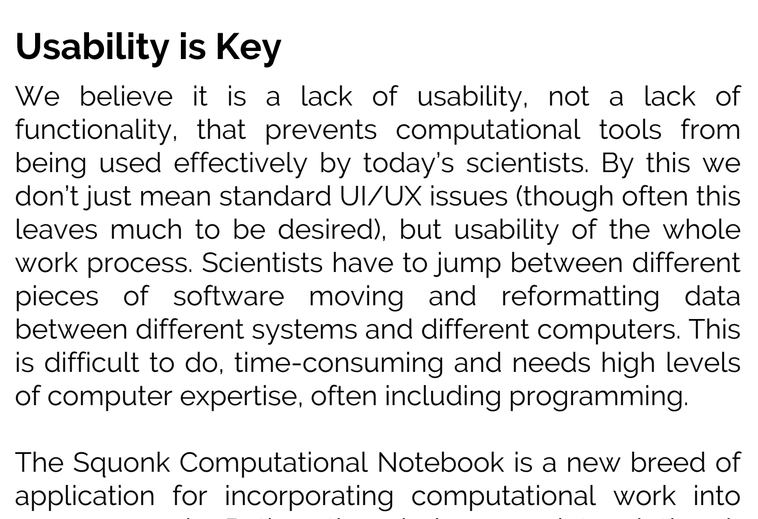
Squonk Computational Notebook

Informatics Matters Ltd.

Usability is Key

We believe it is a lack of usability, not a lack of functionality, that prevents computational tools from being used effectively by today's scientists. By this we don't just mean standard UI/UX issues (though often this leaves much to be desired), but usability of the whole work process. Scientists have to jump between different pieces of software moving and reformatting data between different systems and different computers. This is difficult to do, time-consuming and needs high levels of computer expertise, often including programming.

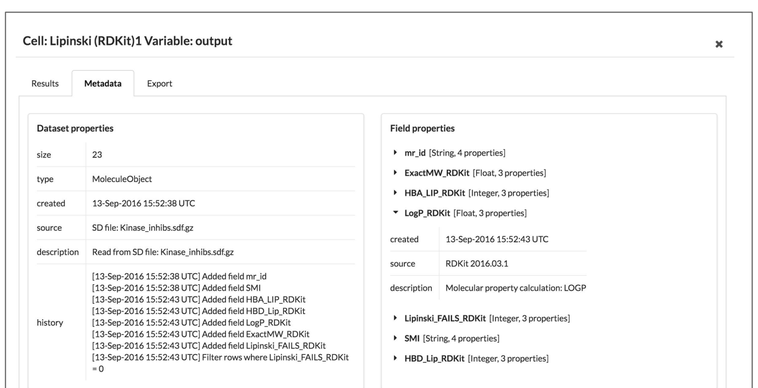
The Squonk Computational Notebook is a new breed of application for incorporating computational work into your research. Rather than being a point solution it addresses the whole workflow process by providing a single place to work, a simple user interface targeted at the normal scientist not the computational geek, and access to powerful, best of breed tools, both open source and commercial.



Example workflow illustrating analysing molecular properties for a set of structures provided in SD file format. Properties are calculated using RDKit and ChemAxon toolkits.

Reproducibility and Traceability

It has become recognised that there is a crisis of reproducibility in science (1). The Squonk Computational Notebook addresses this head on by recording sources of data and transformations performed on that data and incorporating this information into the workflow definition and the metadata associated with the results. This allows a record of computational work to be generated as you work, providing the equivalent of the ELN for chemical synthesis.



Metadata view that is associated with each dataset. This documents where data came from and what operations were performed on it, by what version of software and when the operation was performed..

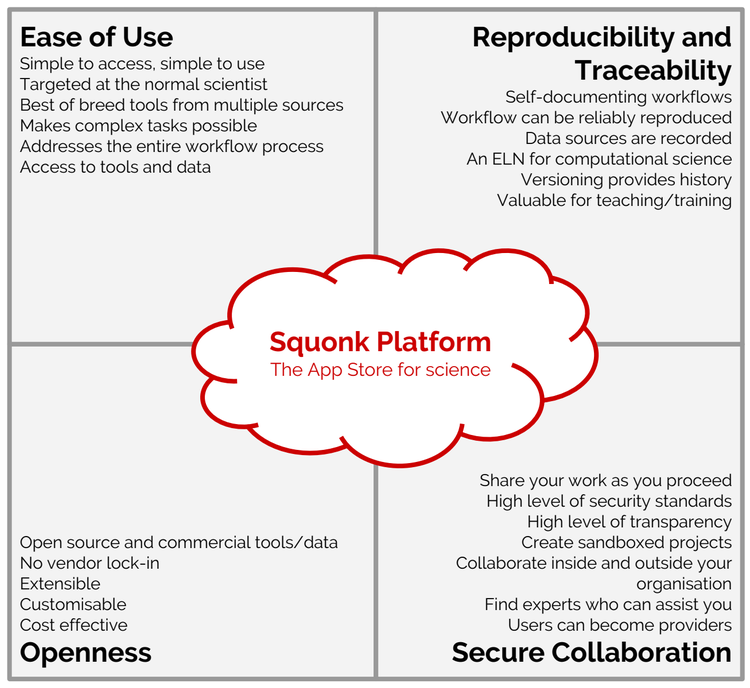
Cloud Hosted Architecture

A cloud hosted web based portal provides access to simple to use high performance tools. No need for internal IT departments or large support organisations. The subscription portal will include a free tier that gives access to a wide range of open source tools. The commercial tiers provide access to commercial tools, heavy computational power, privacy and enhanced collaboration.

A modern service based architecture provides high performance and extensibility allowing big jobs to be done easily and quickly.

Deployment options will also include VPC and in-house hardware.

Squonk's Multiple Dimensions

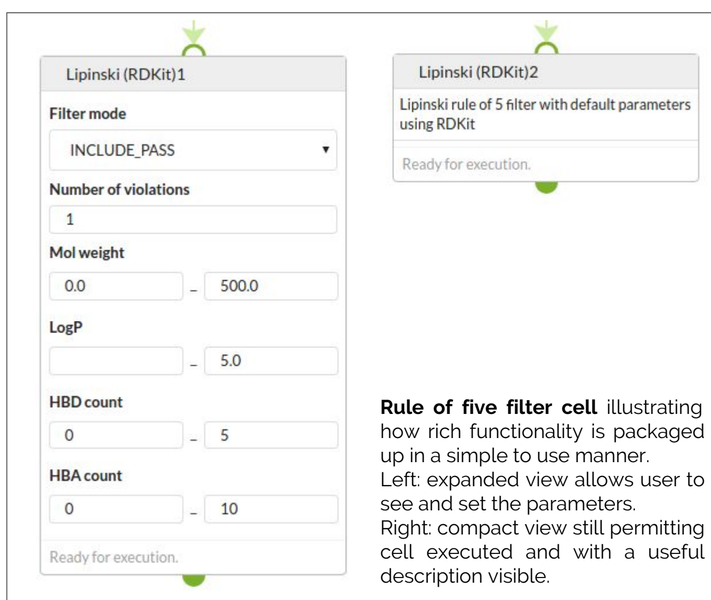


Current Workflows

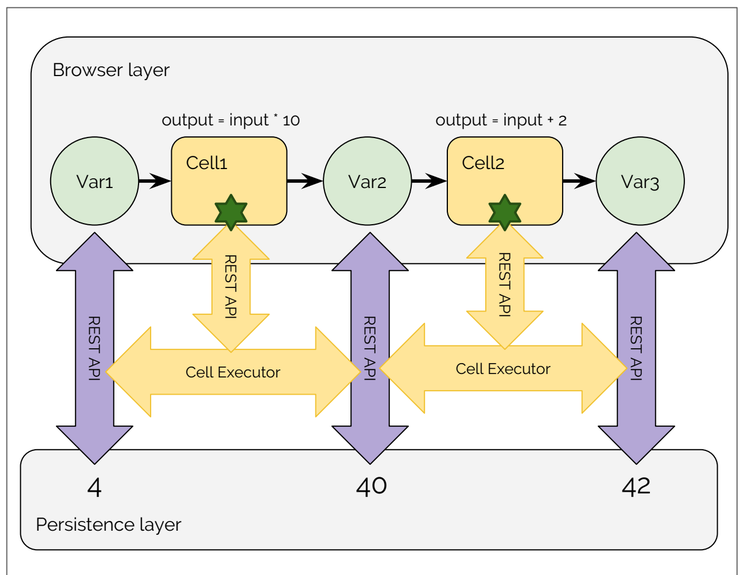
Based around the familiar concepts of wiring workflow components together on a canvas we provide an intuitive but powerful approach to executing workflows. The core components are referred to as "cells". These can be data processing cells, where data is processed using back-end services with user specified options, or they can be data visualisation cells allowing the user to view and interact with the data.

Currently workflows are primarily based around computational chemistry and cheminformatics, but the platform will also support biological and 'omics data. Current functionality includes:

- Data import/export
- Data transformation
- Molecular property prediction
- Property based filters
- Virtual screening
- Clustering
- Chemical databases
- Groovy and Python scripting
- Execution of Docker containers
- Conformer generation
- Library enumeration
- Tabular/SAR displays
- Charts & visualisation
- 3D molecular display



Rule of five filter cell illustrating how rich functionality is packaged up in a simple to use manner.
Left: expanded view allows user to see and set the parameters.
Right: compact view still permitting cell executed and with a useful description visible.



Execution model. Cells from the UI tier are executed by the user which results in a REST web service call to execute the cell. Execution involves retrieving the input variable values, operating on them and writing back new variable values to the persistence store. Whilst this simplistic example assumes simple mathematical operation on numbers, in reality variables are normally complex datasets of molecules or other artefacts.

Access Best of Breed Tools

At the heart of the Squonk Computational Notebook are best of breed toolkits that we integrate and make easy to use. We provide the interoperability between the toolkits. This includes commercial and open source tools.

Current tools that we have integrated or are in the process of being integrated include:

<u>Open source</u>

- RDKit - cheminformatics toolkit (2)
- Chemistry Development Kit- cheminformatics toolkit (3,4)
- OpenChemLib - cheminformatics toolkit from Actelion (5)
- 3DMol from University of Pittsburgh (6)
- D3 Javascript visualisation library (7)

<u>Commercial</u>

- ChemAxon's Marvin and JChem (8)
- CPSign predictive modelling tools from GenettaSoft (9)

Library enumeration being performed on reactants sourced from a database search using the RDKit cartridge, the reactants filtered using a Rule of Three filter and reacted using a amide formation reaction using ChemAxon's Reactor. The enumerated products are analysed using a Ghose filter and the properties assessed using scatter plot and heatmap visualisers.

Collaboration

Being web based with all data being centrally located, collaboration becomes easy. Notebooks can be shared with others, and multiple people can work on a single Notebook with the built in versioning system preventing one user destroying another user's work.

Become involved

Squonk is a community of users with a strong collaborative ethos. You can become involved with this as a user or as a provider.
Currently we are in private alpha testing stage, planning for a public beta later in the year. If you are interested in trying us out then get in touch. If you have some great tools or ideas that could be incorporated into Squonk then also let us know and we'll work with you to incorporate them.

For more information contact us at
info@informaticsmatters.com

References

1. Baker, M. Nature 533, (2016), 452–454.
2. http://www.rdkit.org
3. Steinbeck et al. Current Pharmaceutical Design (2006), 12, (17), 2111-2120
4. https://github.com/cdk/cdk
5. https://github.com/Actelion/openchemlib
6. Rego, N & Koes, D. Bioinformatics (2015) 31 (8): 1322-1324
7. https://d3js.org/
8. https://www.chemaxon.com/
9. http://www.genettasoft.com/