Cambridge Cheminformatics Network Meeting

Setting up Cheminformatics Support for the Open Source Malaria Project

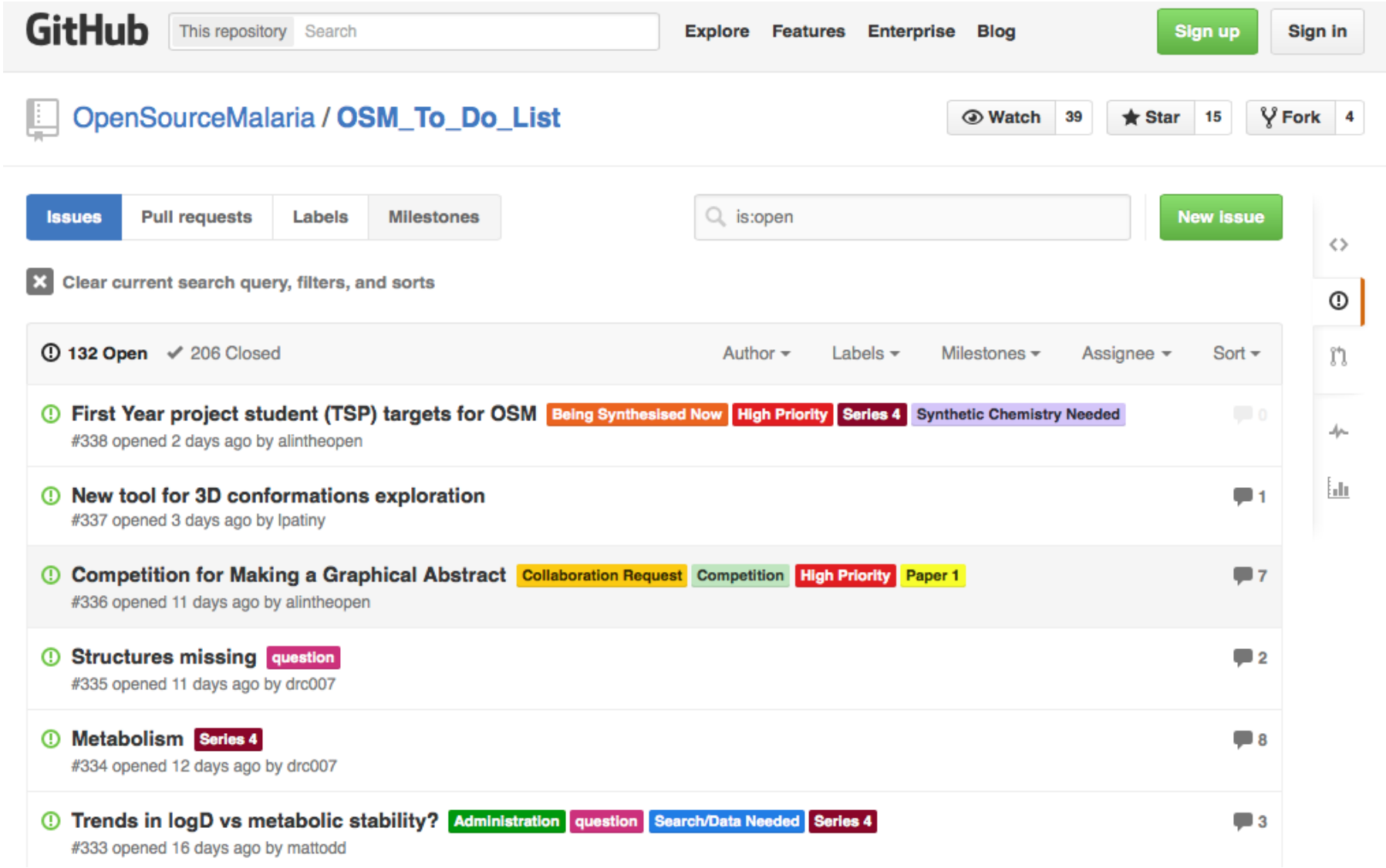
Chris Swain
Cambridge MedChem Consulting
http://www.cambridgemedchemconsulting.com

26 August 2015

Open Source Malaria

- The Open Source Malaria project is trying a different approach to curing malaria. Guided by open source principles, everything is open and anyone can contribute.
- http://opensourcemalaria.github.io/NewSite/
- This Landing Page aggregates the most recent activity in Open Source Malaria. In open source research all data and ideas are freely shared, anyone may participate as an equal partner and there will be no patents - think “Linux for Malaria Research”.

Organisation of project uses GitHub

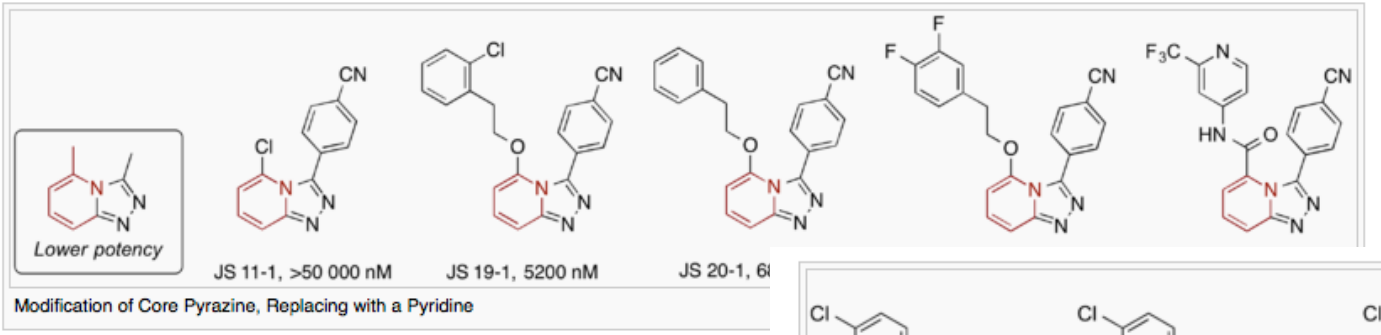


Organization of data

- Historically project data published on web pages
- All in public domain
- But……
- Challenging to collate all data associated with project
- Multiple compound identifiers
- Multiple data sources
- Not always clear which assay data refers to

Examples of data

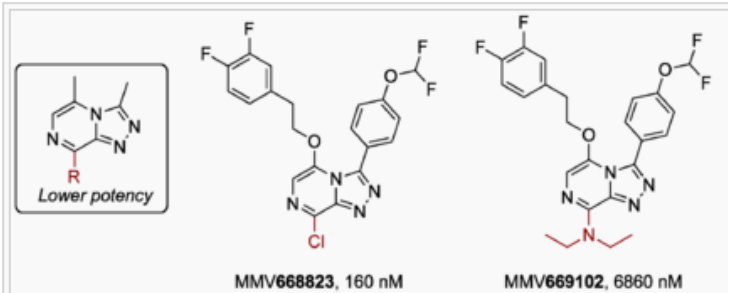
sponding trizaolopyrazine compounds.



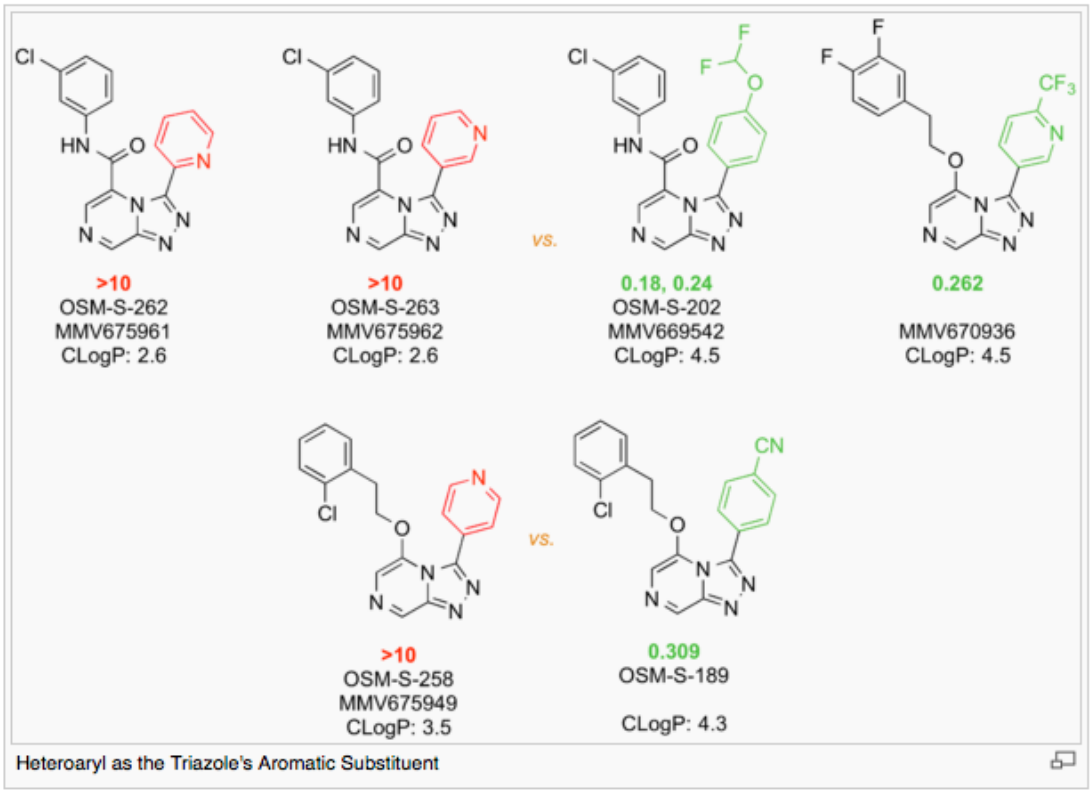
Modification of Core Pyrazine, Replacing with a Pyridine

cation of Pyrazine Substitution Pattern

ought possible that the pyrazine moiety of the triazolopyrazine could undergo aldehyde
s were made to the substitution of the "southernmost" (8-) ring C-H, all lowering potenc



Modification of the Triazolopyrazine's 8-position



Heteroaryl as the Triazole's Aromatic Substituent

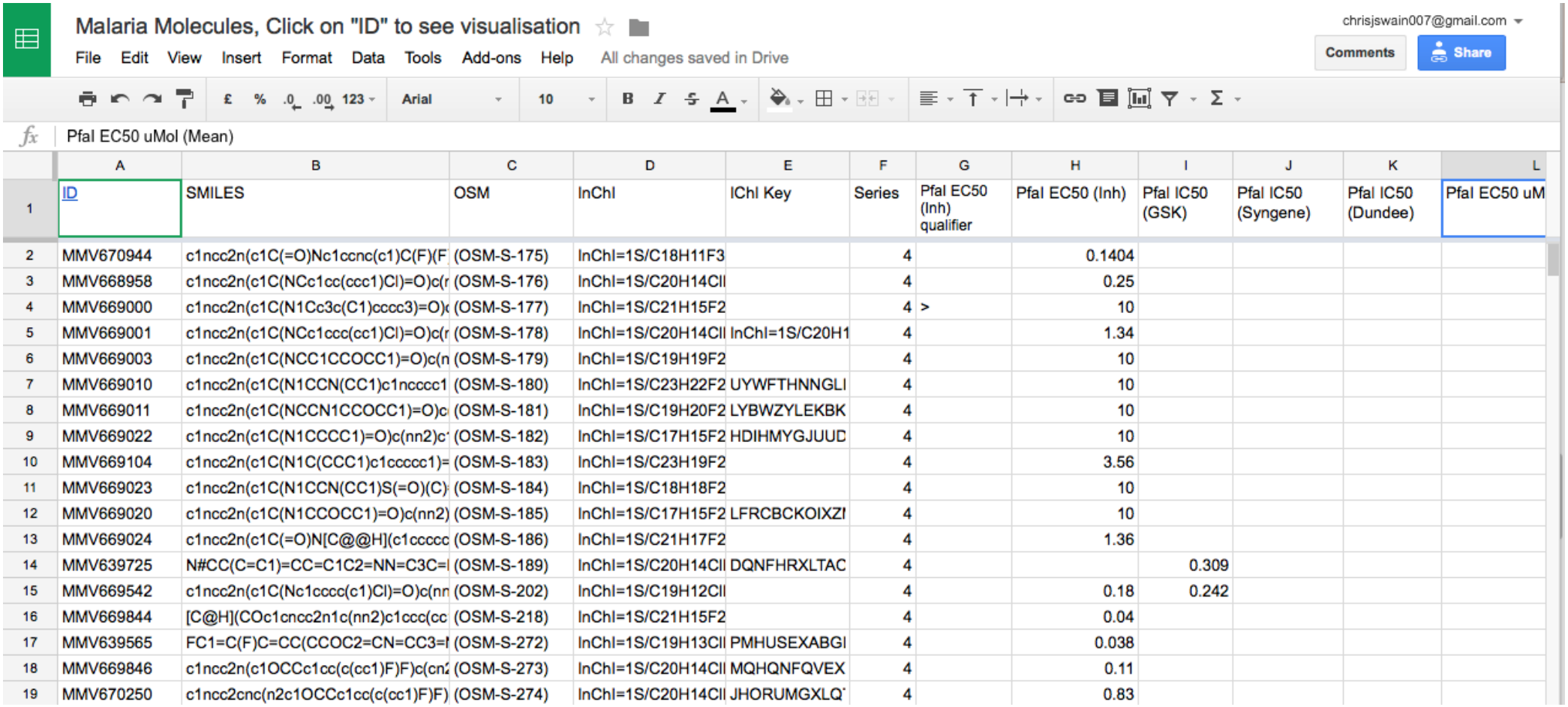
Stage 1:- Implement single data source

- Use Google spreadsheet
- Users comfortable with data entry
- Negligible set up or support required

Google doc spreadsheet

- Simple spreadsheet has allowed multiple users to enter data (BIG THANKS TO ALL!).
- https://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/edit#gid=510297618
- Currently contains 265 rows (compounds), and 46 columns (descriptors and expt data).
- All data can be downloaded as a tab delimited file here
  - http://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/export?format=tsv
- How can we leverage this data source.

Google doc spreadsheet



Malaria Molecules, Click on "ID" to see visualisation

chrisjswain007@gmail.com

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

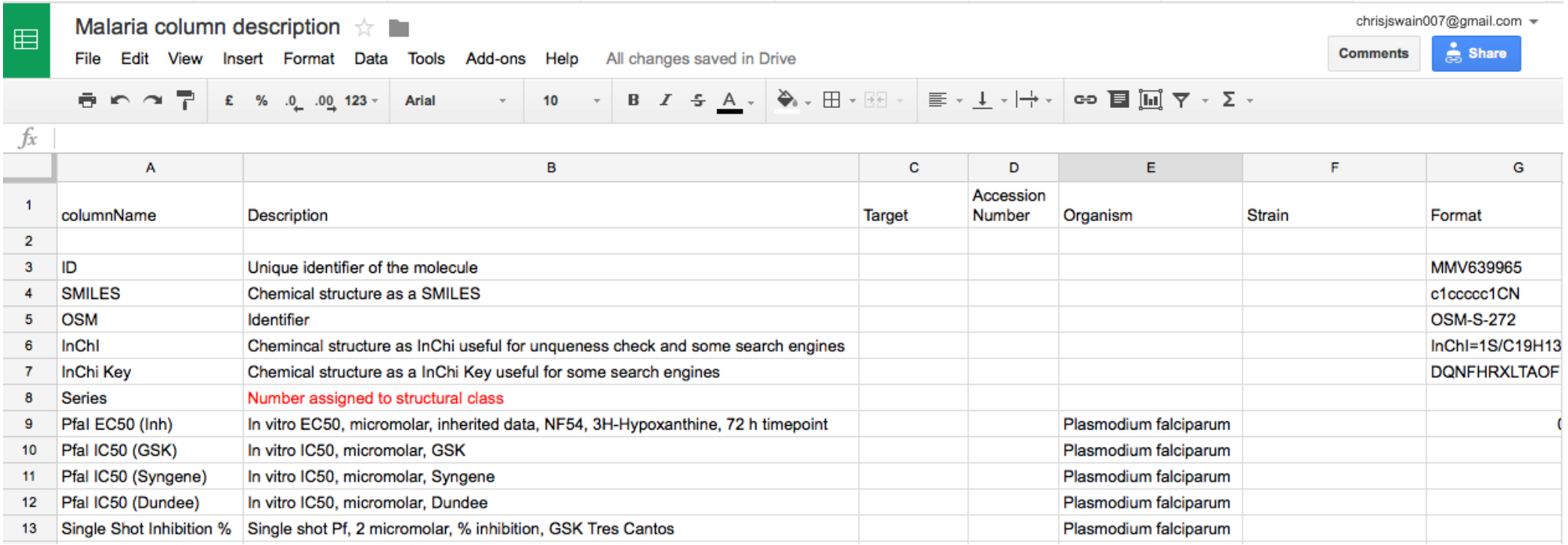
Comments Share

Arial 10

Pfal EC50 uMol (Mean)

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	SMILES	OSM	InChI	IChI Key	Series	Pfal EC50 (Inh) qualifier	Pfal EC50 (Inh)	Pfal IC50 (GSK)	Pfal IC50 (Syngene)	Pfal IC50 (Dundee)	Pfal EC50 uM
2	MMV670944	c1ncc2n(c1C(=O)Nc1ccnc(c1)C(F)(F	(OSM-S-175)	InChI=1S/C18H11F3		4		0.1404				
3	MMV668958	c1ncc2n(c1C(NCc1cc(ccc1)Cl)=O)c(r	(OSM-S-176)	InChI=1S/C20H14ClI		4		0.25				
4	MMV669000	c1ncc2n(c1C(N1Cc3c(C1)cccc3)=O)c	(OSM-S-177)	InChI=1S/C21H15F2		4	>	10				
5	MMV669001	c1ncc2n(c1C(NCc1ccc(cc1)Cl)=O)c(r	(OSM-S-178)	InChI=1S/C20H14ClI	InChI=1S/C20H1	4		1.34				
6	MMV669003	c1ncc2n(c1C(NCC1CCOCC1)=O)c(n	(OSM-S-179)	InChI=1S/C19H19F2		4		10				
7	MMV669010	c1ncc2n(c1C(N1CCN(CC1)c1ncccc1	(OSM-S-180)	InChI=1S/C23H22F2	UYWFTHNNGLI	4		10				
8	MMV669011	c1ncc2n(c1C(NCCN1CCOCC1)=O)c	(OSM-S-181)	InChI=1S/C19H20F2	LYBWZYLEKBK	4		10				
9	MMV669022	c1ncc2n(c1C(N1CCCC1)=O)c(nn2)c	(OSM-S-182)	InChI=1S/C17H15F2	HDIHMYGJUUD	4		10				
10	MMV669104	c1ncc2n(c1C(N1C(CCC1)c1ccccc1)=	(OSM-S-183)	InChI=1S/C23H19F2		4		3.56				
11	MMV669023	c1ncc2n(c1C(N1CCN(CC1)S(=O)(C)	(OSM-S-184)	InChI=1S/C18H18F2		4		10				
12	MMV669020	c1ncc2n(c1C(N1CCOCC1)=O)c(nn2)	(OSM-S-185)	InChI=1S/C17H15F2	LFRCBCKOIXZI	4		10				
13	MMV669024	c1ncc2n(c1C(=O)N[C@@H](c1ccccc	(OSM-S-186)	InChI=1S/C21H17F2		4		1.36				
14	MMV639725	N#CC(C=C1)=CC=C1C2=NN=C3C=I	(OSM-S-189)	InChI=1S/C20H14ClI	DQNFHRXLTAC	4			0.309			
15	MMV669542	c1ncc2n(c1C(Nc1cccc(c1)Cl)=O)c(nn	(OSM-S-202)	InChI=1S/C19H12ClI		4		0.18	0.242			
16	MMV669844	[C@H](COc1cncc2n1c(nn2)c1ccc(cc	(OSM-S-218)	InChI=1S/C21H15F2		4		0.04				
17	MMV639565	FC1=C(F)C=CC(CCOC2=CN=CC3=I	(OSM-S-272)	InChI=1S/C19H13ClI	PMHUSEXABGI	4		0.038				
18	MMV669846	c1ncc2n(c1OCCc1cc(c(cc1)F)F)c(cn2	(OSM-S-273)	InChI=1S/C20H14ClI	MQHQNFQVEX	4		0.11				
19	MMV670250	c1ncc2cnc(n2c1OCCc1cc(c(cc1)F)F)	(OSM-S-274)	InChI=1S/C20H14ClI	JHORUMGXLQ	4		0.83				

Column Description



Malaria column description

chrisjswain007@gmail.com

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Comments Share

	A	B	C	D	E	F	G
1	columnName	Description	Target	Accession Number	Organism	Strain	Format
2							
3	ID	Unique identifier of the molecule					MMV639965
4	SMILES	Chemical structure as a SMILES					c1ccccc1CN
5	OSM	Identifier					OSM-S-272
6	InChI	Chemincal structure as InChi useful for unqueness check and some search engines					InChI=1S/C19H13
7	InChi Key	Chemical structure as a InChi Key useful for some search engines					DQNFHRXLTAOF
8	Series	Number assigned to structural class					
9	Pfal EC50 (Inh)	In vitro EC50, micromolar, inherited data, NF54, 3H-Hypoxanthine, 72 h timepoint			Plasmodium falciparum		
10	Pfal IC50 (GSK)	In vitro IC50, micromolar, GSK			Plasmodium falciparum		
11	Pfal IC50 (Syngene)	In vitro IC50, micromolar, Syngene			Plasmodium falciparum		
12	Pfal IC50 (Dundee)	In vitro IC50, micromolar, Dundee			Plasmodium falciparum		
13	Single Shot Inhibition %	Single shot Pf, 2 micromolar, % inhibition, GSK Tres Cantos			Plasmodium falciparum		

Google doc spreadsheet

- Simple spreadsheet has allowed multiple users to enter data (BIG THANKS!).
- https://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/edit#gid=510297618
- Currently contains 265 rows (compounds), and 46 columns (descriptors and expt data).
- All data can be downloaded as a tab delimited file here
  - http://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/export?format=tsv
- How can we leverage this data source?

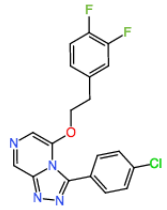
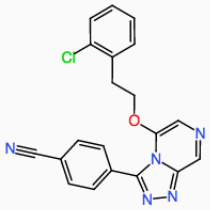
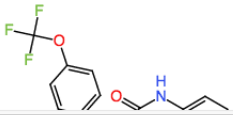
Web View

- One click access to data provided by chemifo.org (huge thanks to Luc Patiny)
- http://www.cheminfo.org/flavor/malaria/Display_data.html
- Imports data from live Google spreadsheet
- Calculates a variety of physiochemical properties on the fly.
- Can display data as table and a variiety of different plots.

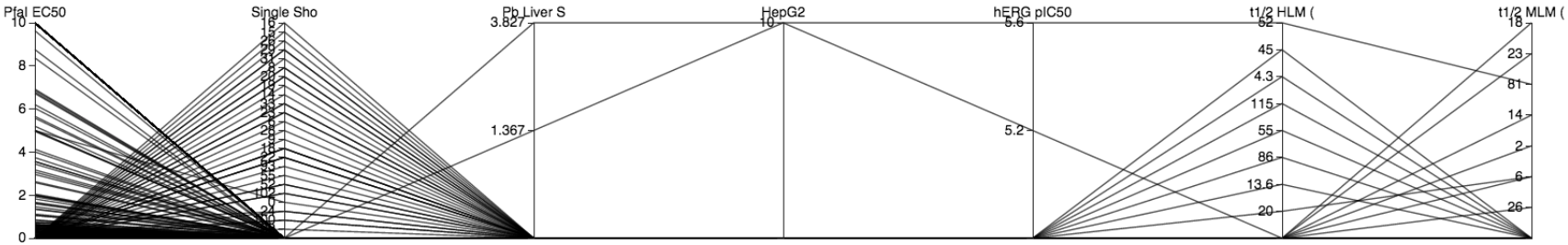
WebView

DISPLAY DATA

Reload data

Molfile	ID	smiles	OSM	InChI	IChI Key	Series	Pfal EC50 (I…	Pfal EC50 (Inh)	Pfal IC50 (G…	Pfal IC50 (S…	Pfal IC50 (D…	Pfal EC50 u…	Single Shot I…	Ion Regulati…	Pb Liver Stage	HepG2	hERG pIC50	t1/2 HLM (M…	t1/2 HLM (Inh)	t1/2 M
	MMV639565	FC1=C(F)C=…	(OSM-S-272)	InChI=1S/C1…	PMHUSEXA…	4		0.038				0.038								
	MMV639725	N#CC(C=C1)…	(OSM-S-189)	InChI=1S/C2…	DQNFHRXLT…	4			0.309				99							
	MMV657963	O=C(NC1=C…	OSM-S-271	InChI=1S/C1…	KPPYAXXED…	4		0.234				0.234								

Parallel coordinates

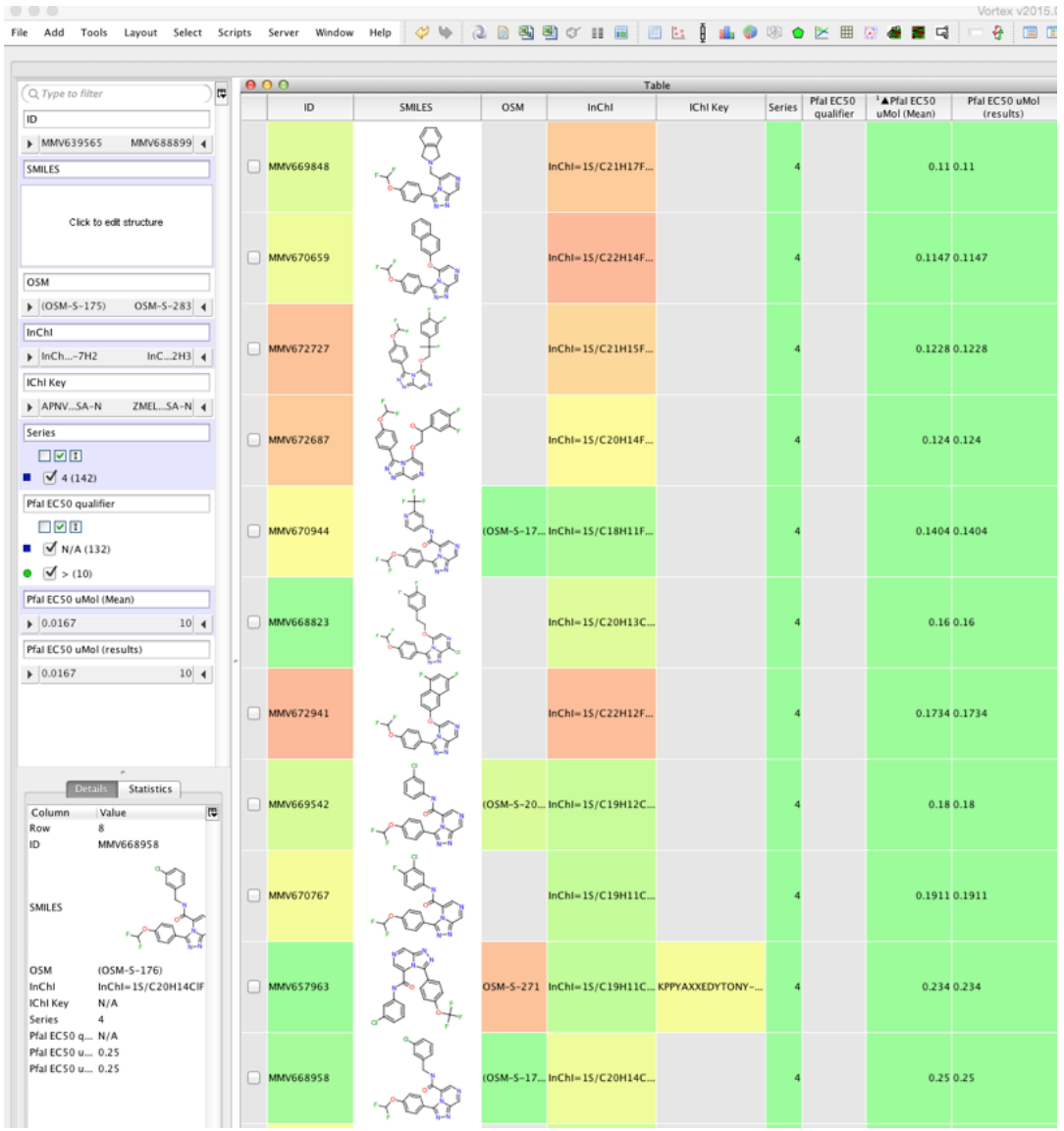


Vortex

- Vortex is a high performance data visualisation and analysis platform
- Written script to automatically import from Google spreadsheet
- http://macinchem.org/reviews/vortex/tut26/scripting_vortex26.php

Vortex

```
# Python imports
import urllib2
import urllib
import csv
import sys
from com.xhaus.jyson import JysonCodec as json

# Vortex imports
import com.dotmatics.vortex.util.Util as Util
import com.dotmatics.vortex.mol2img.jni.genImage as genImage
import com.dotmatics.vortex.mol2img.Mol2Img as mol2Img
import com.dotmatics.vortex.table.VortexTableModel as vtm
import jarray
import binascii
import string
import os

mystr = "http://docs.google.com/spreadsheets/d/
1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/export?format=tsv"

myreturn = urllib2.urlopen(mystr).read()
list1 = myreturn.split('\n')

TableName = "OSMData"

# Get column names
column_names = list1[0].split('\t')

rows = []
for i in list1[1:]:
    row = i.split('\t')
    rows.append(row)


arrayToWorkspace(rows, column_names, TableName)a
```

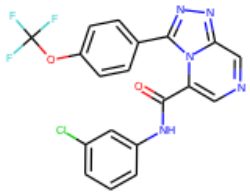


iPython Notebook

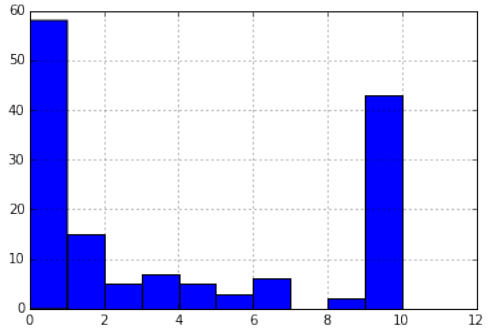
- iPython Notebooks are becoming increasingly popular formats for combining combine code execution, rich text, mathematics, plots, chemical structures etc
- http://www.macinchem.org/reviews/osm/osmipython.php

iPython Notebook

```
!wget http://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3lSuJ1jaR7AJa8hgGsc/export?
format=tsv
from rdkit.Chem import AllChem as Chem
from rdkit.Chem.Draw import IPythonConsole
from rdkit.Chem import PandasTools
from rdkit.Chem import Draw
import pybel


import pandas as pd
datafile = pd.read_table('./export?format=tsv')
smiles = datafile['SMILES'].loc[2]

mol = Chem.MolFromSmiles(smiles)
mol
```



```
datafile['Pfal EC50 uMol (Mean)'].hist()
```



Everybody has access

- Everybody has live up to the minute access to the data
- All the data

- Why not give it a try and then contribute your findings and suggestions to the Open Source Malaria project.