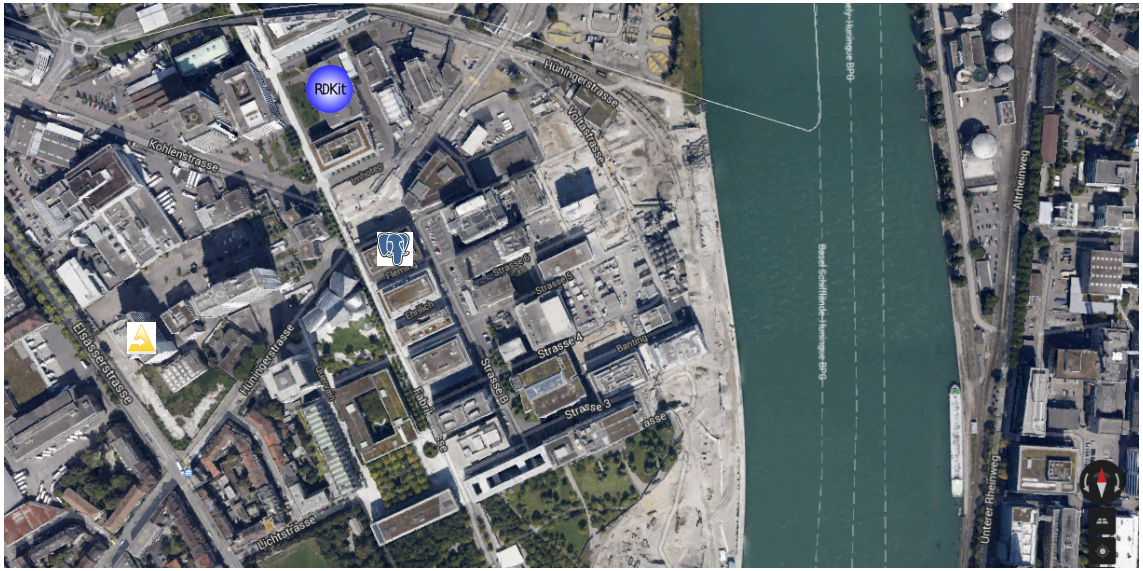


Open-source from/in the enterprise: the RDKit

Gregory Landrum

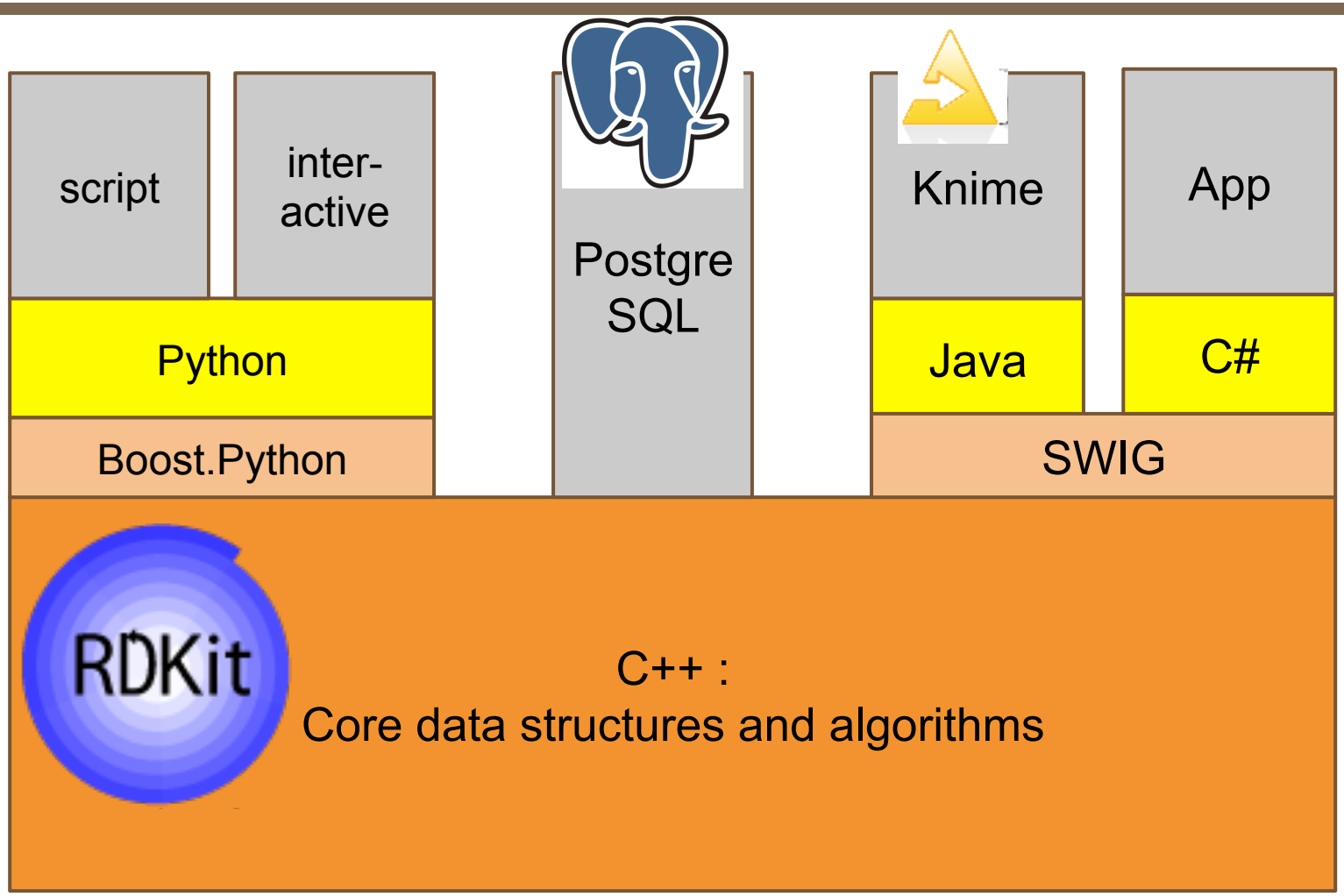
NIBR Informatics

Novartis Institutes for BioMedical Research, Basel, Switzerland

Outline

- RDKit integration with other open-source projects
  - Knime
  - PostgreSQL
  - IPython
  - Pandas
  - Lucene
- RDKit in NIBR, some case studies

What is this all about?

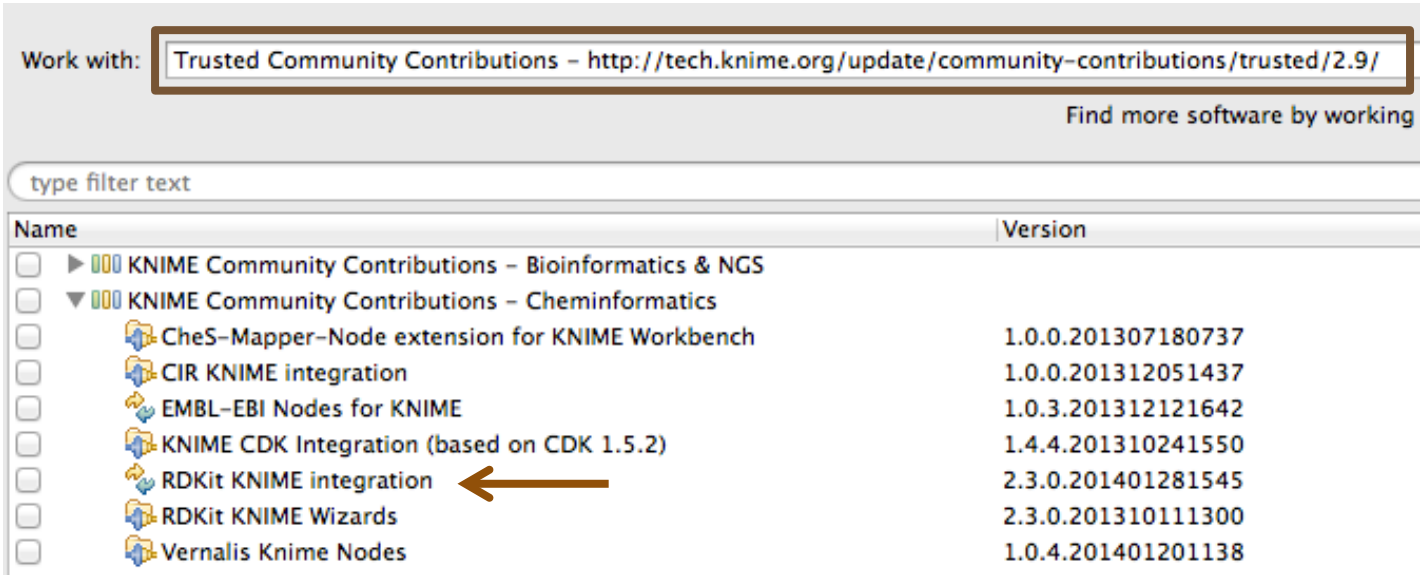


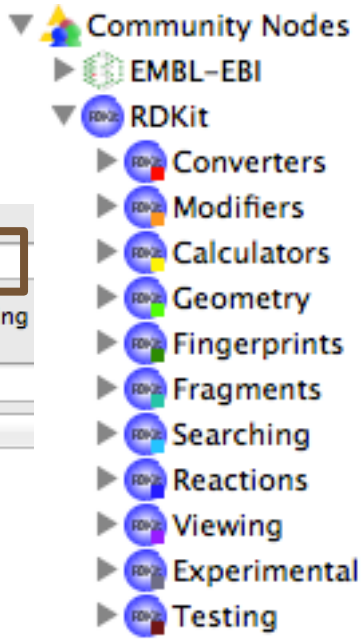
Exact same algorithms/implementations accessible from many different endpoints

NOVARTIS

Knime integration

- *Open-source* RDKit-based nodes for Knime providing cheminformatics functionality
- Trusted nodes distributed from knime community site





- Work in progress: more nodes being added (new wizard makes it *easy*)

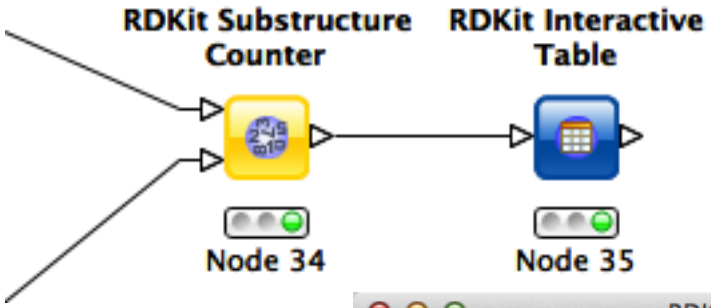
What's there?

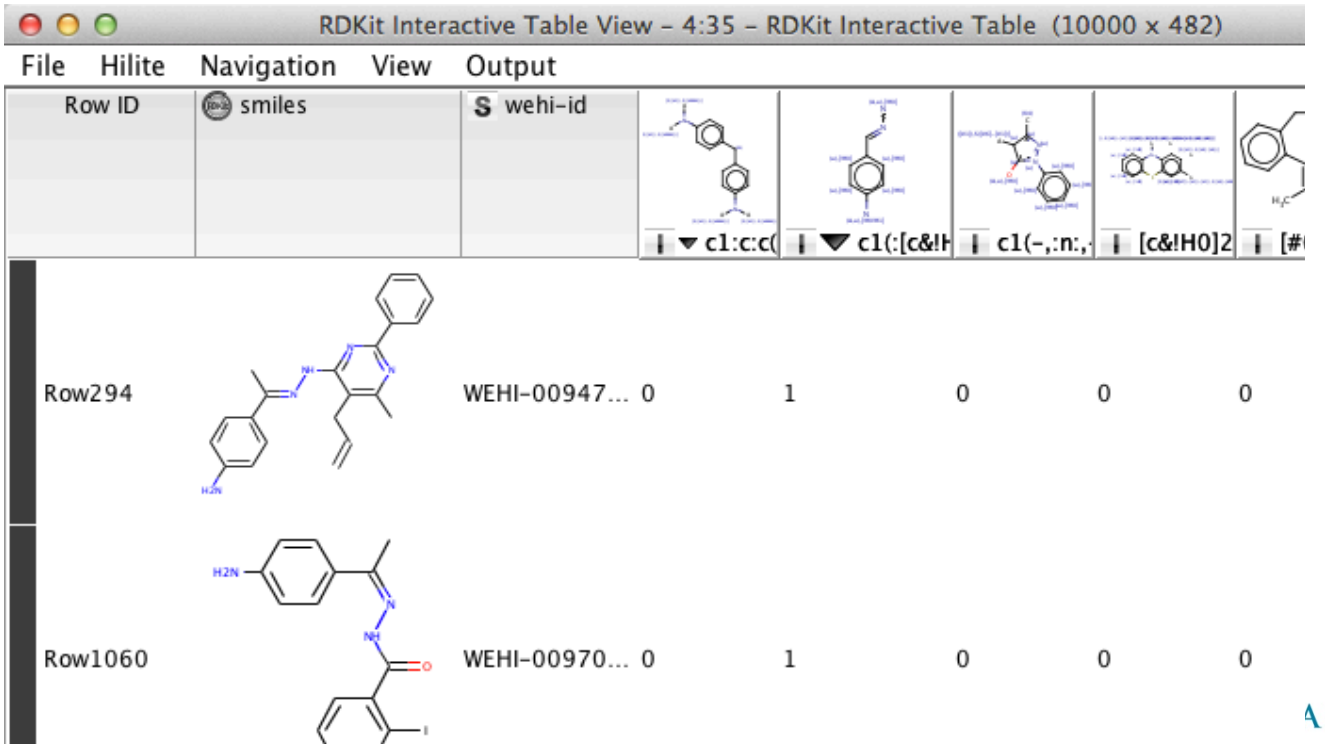
- Community Nodes
  - EMBL-EBI
  - RDKit
    - Converters
    - Modifiers
    - Calculators
    - Geometry
    - Fingerprints
    - Fragments
    - Searching
    - Reactions
    - Viewing
    - Experimental
    - Testing

- RDKit
  - Converters
    - RDKit From Molecule
    - RDKit To Molecule
    - RDKit From InChI
    - RDKit To InChI
    - RDKit From IUPAC
    - RDKit Canon SMILES
  - Modifiers
    - RDKit Add Hs
    - RDKit Remove Hs
    - RDKit Aromatizer
    - RDKit Kekulizer
    - RDKit Salt Stripper
  - Calculators
    - RDKit Descriptor Calculation
    - RDKit Calculate Charges
  - Geometry
    - RDKit Generate Coords
    - RDKit Optimize Geometry
    - RDKit Add Conformers
    - RDKit Open 3D Alignment
    - RDKit RMSD Filter
  - Geometry
  - Fingerprints
    - RDKit Fingerprint
    - RDKit Fingerprint Reader
    - RDKit Fingerprint Writer
    - RDKit Diversity Picker
  - Fragments
    - RDKit Molecule Fragmenter
    - RDKit Find Murcko Scaffolds
  - Searching
    - RDKit Substructure Filter
    - RDKit Molecule Substructure Filter
    - RDKit Dictionary Substructure Filter
    - RDKit Functional Group Filter
    - RDKit Substructure Counter
  - Reactions
    - RDKit One Component Reaction
    - RDKit Two Component Reaction
  - Viewing
    - RDKit Interactive Table
    - RDKit SMILES Headers
    - RDKit Highlighting Atoms
  - Experimental
    - RDKit R Group Decomposition

RDKit Interactive Table

- KNIME interactive table with molecules as column headers

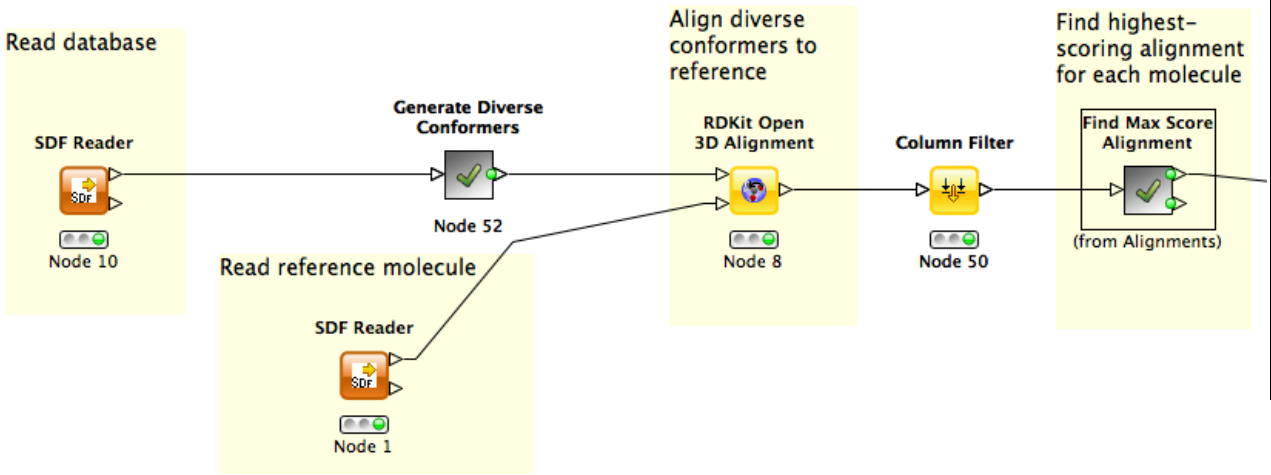


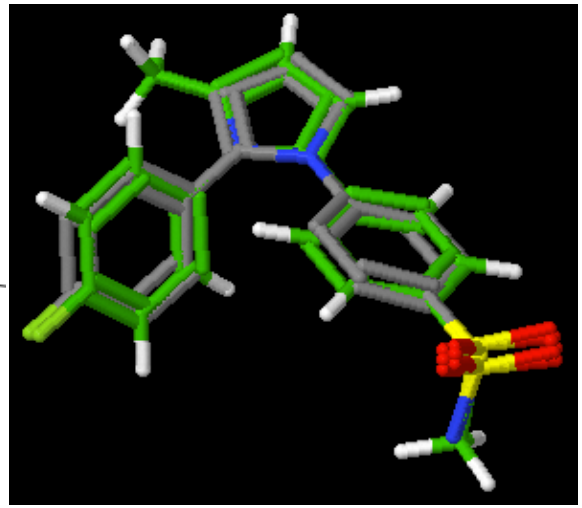


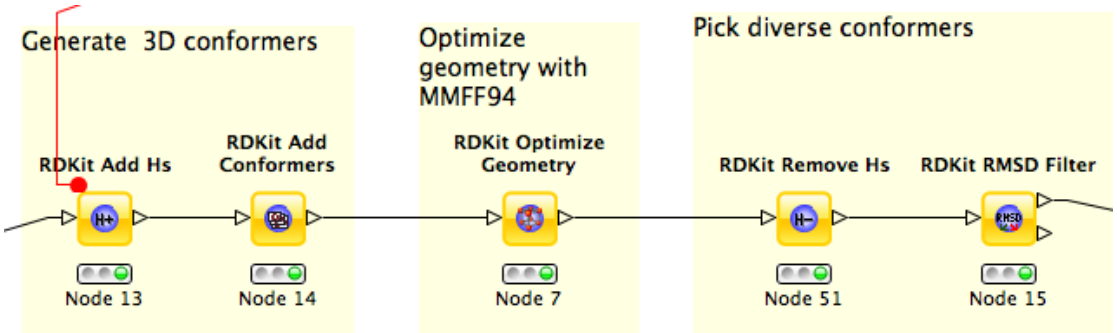
ARTIS

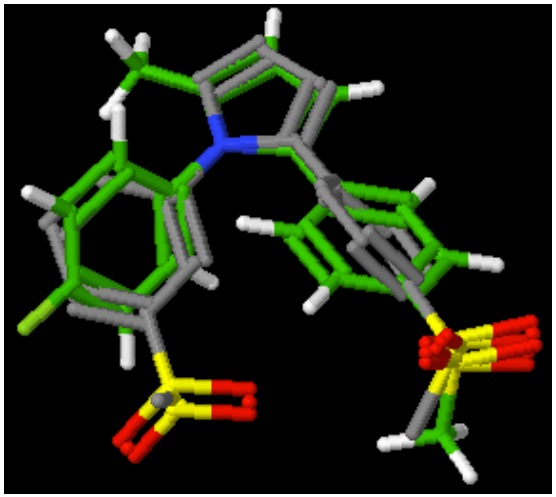
Functionality for working with 3D molecules

- Example: flexible molecule-molecule alignment









PostgreSQL integration

- PostgreSQL (http://www.postgresql.org): a robust, flexible, and extensible relational open-source database. Rich collection of extensions available
- RDKit “cartridge”:
  - Fast substructure and similarity search
  - Fingerprints (count-based and bit-vector):

    Morgan (ECFP-like), FeatMorgan (FCFP-like), RDKit (Daylight like), atom pair, topological torsion, MACCS
  - Standard molecule properties and descriptors
- Basis for myChEMBL (http://chembl.blogspot.co.uk/2013/10/chembl-virtual-machine-aka-mychembl.html) Ochoa, R., Davies, M., Papadatos, G., Atkinson, F., & Overington, J. P. (2014). myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics*, *30*(2), 298–300.

PostgreSQL integration

Substructure search

```
chembl_17=# select molregno,m from rdk.mols where
m@>'c1ccc2c(c1)C(=NN(C2=O)Cc3nc4cc(ccc4s3)C)CC(=O)O';
 molregno |                                m
----------+-------------------------------------------------------------
     7502 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12
    23364 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)cc(C(F)(F)F)c3s2)c(=O)c2ccccc12
    23439 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)cc(Cl)c3s2)c(=O)c2ccccc12
    23462 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)cc(F)c3s2)c(=O)c2ccccc12
    24192 | Cc1cc2nc(Cn3nc(CC(=O)O)c4ccccc4c3=O)sc2c(C)c1
    24190 | COc1cc2sc(Cn3nc(CC(=O)O)c4ccccc4c3=O)nc2cc1C(F)(F)F
    24194 | Cc1ccc2sc(Cn3nc(CC(=O)O)c4ccccc4c3=O)nc2c1
    24237 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)c(O)cc3s2)c(=O)c2ccccc12
    24331 | CC(c1nc2cc(C(F)(F)F)ccc2s1)n1nc(CC(=O)O)c2ccccc2c1=O
(9 rows)

Time: 112.325 ms
```

NOVARTIS

PostgreSQL integration

Similarity search

```
chembl_17=# select * from get_mfp2_neighbors('O=C(O)Cc1nn(Cc2nc3cc(C(F)
(F)F)ccc3s2)c(=O)c2ccccc12') limit 5;
 molregno |                          m                          |    similarity
----------+-----------------------------------------------------+--------------------
     7502 | O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12  |                  1
    24184 | O=C(O)Cc1nn(Cc2nc3ccc(C(F)(F)F)cc3s2)c(=O)c2ccccc12  | 0.859649122807018
    24153 | O=C(O)Cc1nn(CCc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12 | 0.830508474576271
    24152 | O=C(O)Cc1nn(Cc2nc3ccccc3s2)c(=O)c2cc(C(F)(F)F)ccc12  | 0.813559322033898
    24150 | O=C(O)Cc1nn(Cc2nc3ccccc3s2)c(=O)c2ccc(C(F)(F)F)cc12  | 0.813559322033898
(5 rows)

Time: 1222.426 ms
```

Notice that results come back in sorted order

NOVARTIS

PostgreSQL integration

Other functionality

```
chembl_17=# select mol_formula('O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12');
  mol_formula
---------------
 C19H12F3N3O3S
(1 row)
chembl_17=# select mol_logp('O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12');
 mol_logp
----------
   3.7004
(1 row)
chembl_17=# select mol_inchi('O=C(O)Cc1nn(Cc2nc3cc(C(F)(F)F)ccc3s2)c(=O)c2ccccc12');
mol_inchi
------------------------------------------------------------------------------------------
-----------------------------------------------
 InChI=1S/C19H12F3N3O3S/
c20-19(21,22)10-5-6-15-14(7-10)23-16(29-15)9-25-18(28)12-4-2-1-3-11(12)13(24-25)8-17(26)27
/h1-7H,8-9H2,(H,26,27)
(1 row)
```

PostgreSQL integration

Other functionality

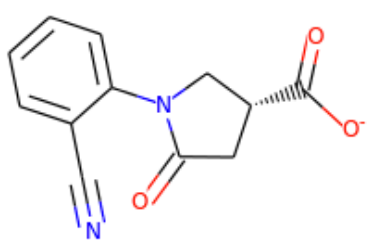
```
chembl_17=# select mol_to_ctab('CC'::mol);
                              mol_to_ctab
-----------------------------------------------------------------------
                                                                      +
      RDKit          2D                                               +
                                                                      +
   2  1  0  0  0  0  0  0  0  0999 V2000                              +
     0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0+
     1.2990    0.7500    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0+
   1  2  1  0                                                         +
 M  END                                                               +

(1 row)
```

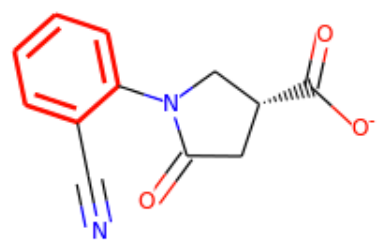
IPython notebok integration

- IPython: a very powerful interactive shell for python
  http://www.ipython.org
- IPython notebook: IPython in the browser, with graphics
  - combines code and output in one place
  - great tool for reproducible research
  - Example notebook with graphics.
- RDKit integration:
  - Display molecules, substructure matches, reactions, graphics from PyMOL

```
In [3]: Chem.SetHybridization(m)
        m
Out[3]:
```



```
In [4]: tmp=m.GetSubstructMatch(Chem.MolFromSmarts('c1cccccl'))
        m
Out[4]:
```



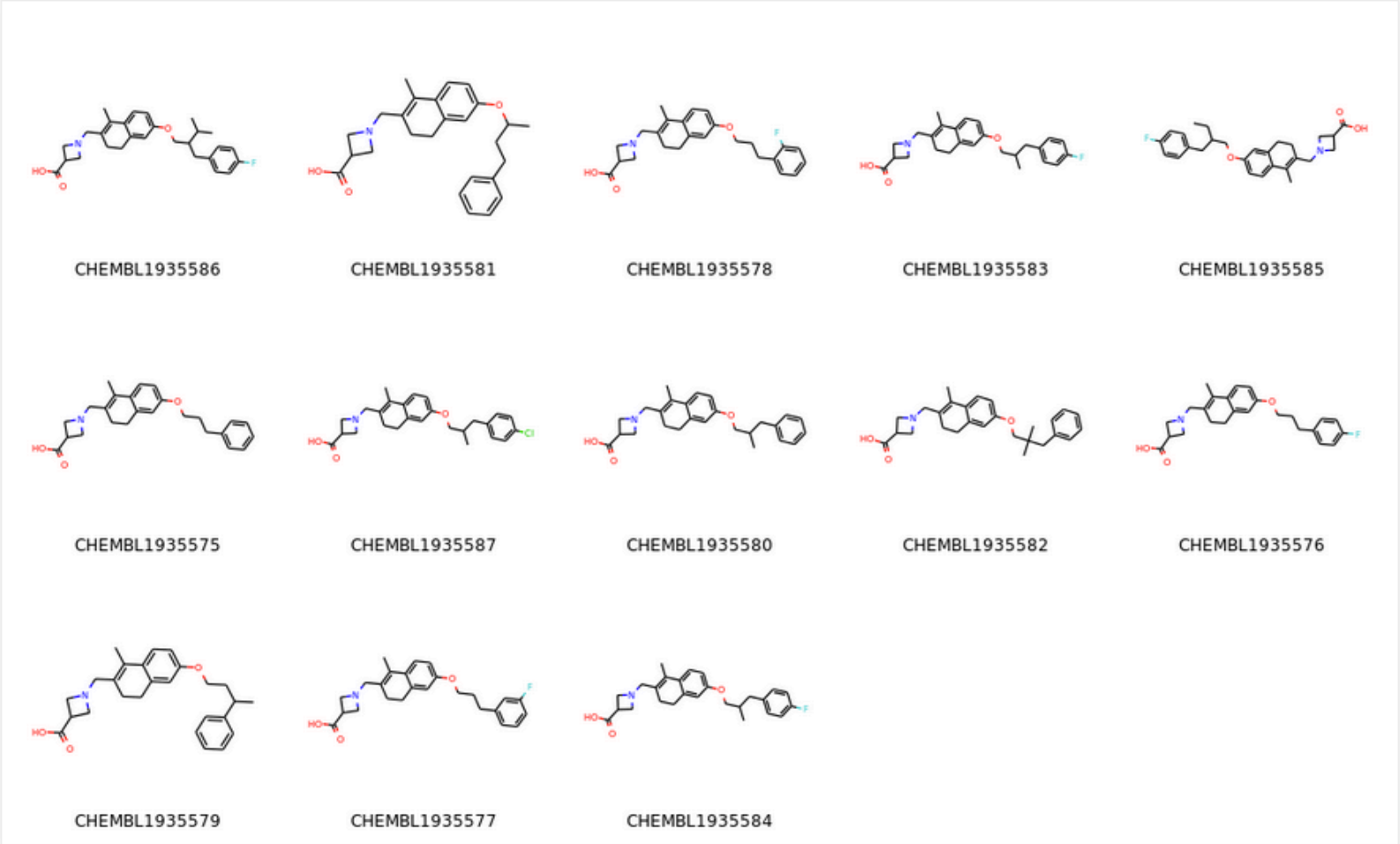
NOVARTIS

IPython notebook integration:
Molecule tables

IP[y]: + RDKit

In [8]:
```
ids,smis= zip(*cmpds)
mols = [Chem.MolFromSmiles(x) for x in smis]
Draw.MolsToGridImage(mols,molsPerRow=5,legends=ids)
```

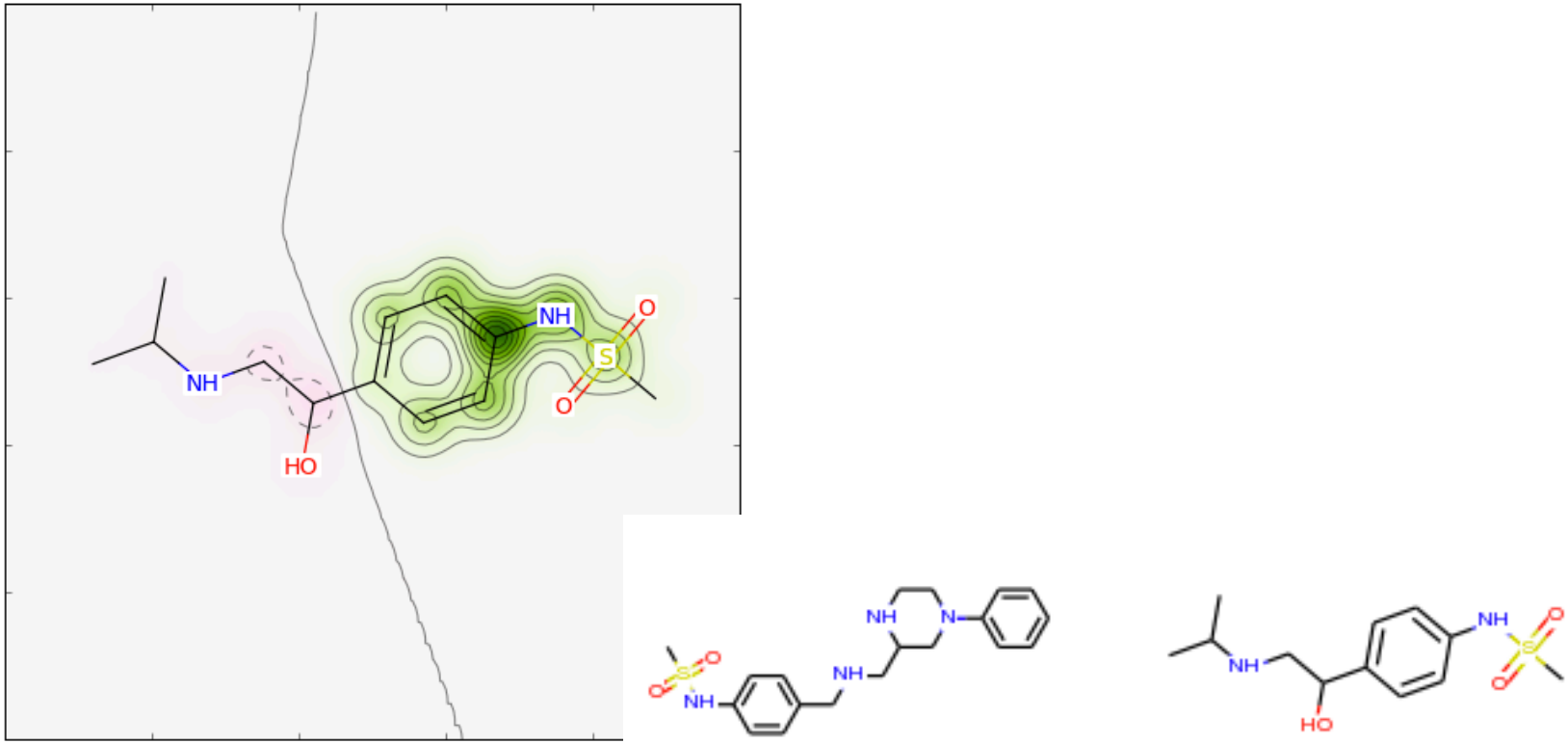
Out[8]:



http://rdkit.blogspot.ch/2014/02/more-on-datasets-ii.html

IPython notebook integration:
Similarity Maps

```
In [56]: SimilarityMaps.GetSimilarityMapForFingerprint(ms[0],ms[16],SimilarityMaps.GetTTFingerprint)
Out[56]: (<matplotlib.figure.Figure at 0x109786850>, 0.31663113006396593)
```



Riniker, S. & Landrum, G. A. J Cheminf (2013). http://www.jcheminf.com/content/5/1/43

IPython notebook integration:
PyMol

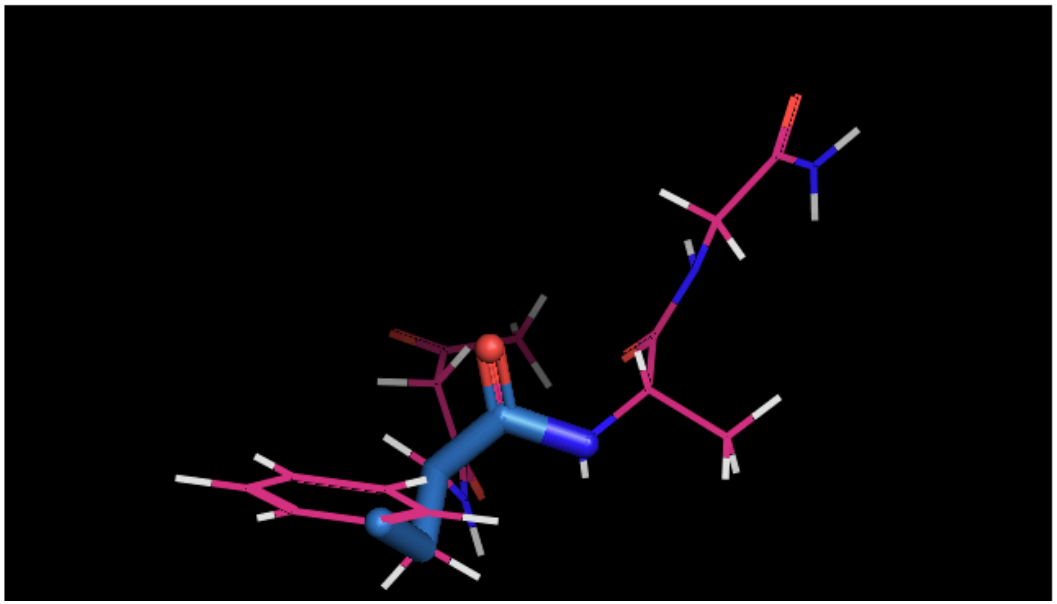
IP[y]: + RDKit

In [12]:
```
nm = Chem.Mol(matchingMols[3])
GetFF=lambda x,confId=-1:AllChem.MMFFGetMoleculeForceField(x,AllChem.MMFFGetMoleculeProperties(x),confId=conf
Id)
AllChem.ConstrainedEmbed(nm,core1,getForceField=GetFF)
rms = float(nm.GetProp('EmbedRMS'))
print 'RMS:',rms


v.ShowMol(core1,molB=Chem.MolToMolBlock(core1,kekulize=False),name='core')
v.SetDisplayStyle('core','sticks')
v.ShowMol(nm,'constrained embed',showOnly=False)
v.GetPNG(preDelay=2)
```

```
RMS: 0.0407469103071
```

Out[12]:



http://rdkit.blogspot.ch/2013/12/using-allchemconstrainedembed.html

Pandas integration

pandas + RDKit

- Pandas: library for working with data tables in Python. Integrates well with matplotlib and ipython

  http://pandas.pydata.org/
- RDKit integration:
  - Load smiles tables or SD files into Pandas data tables
  - Adds molecule columns to existing tables with smiles/SD columns
  - Enables substructure filters on tables
  - Integration with IPython notebook to render molecules

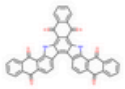
Pandas integration

pandas + RDKit

Molecules in tables

```
In [6]: PandasTools.AddMoleculeColumnToFrame(data,smilesCol='smiles',molCol='molecule',includeFingerprints=False)
        data.head(2)
```

Out[6]:

	smiles	mutagenic	molecule
cas			
2475-33-4	O=C1c2ccccc2C(=O)c3c1ccc4c3[nH]c5c6C(=O)c7ccccc7C(=O)c6c8[nH]c9c%10C(=O)c%11ccccc%11C(=O)c%10ccc9c8c45	0	
820-75-7	NNC(=O)CNC(=O)\C=N\#N	1	None

Substructure filters

```
In [11]: data.groupby(data['molecule'] >= nitroso).describe().unstack()
```

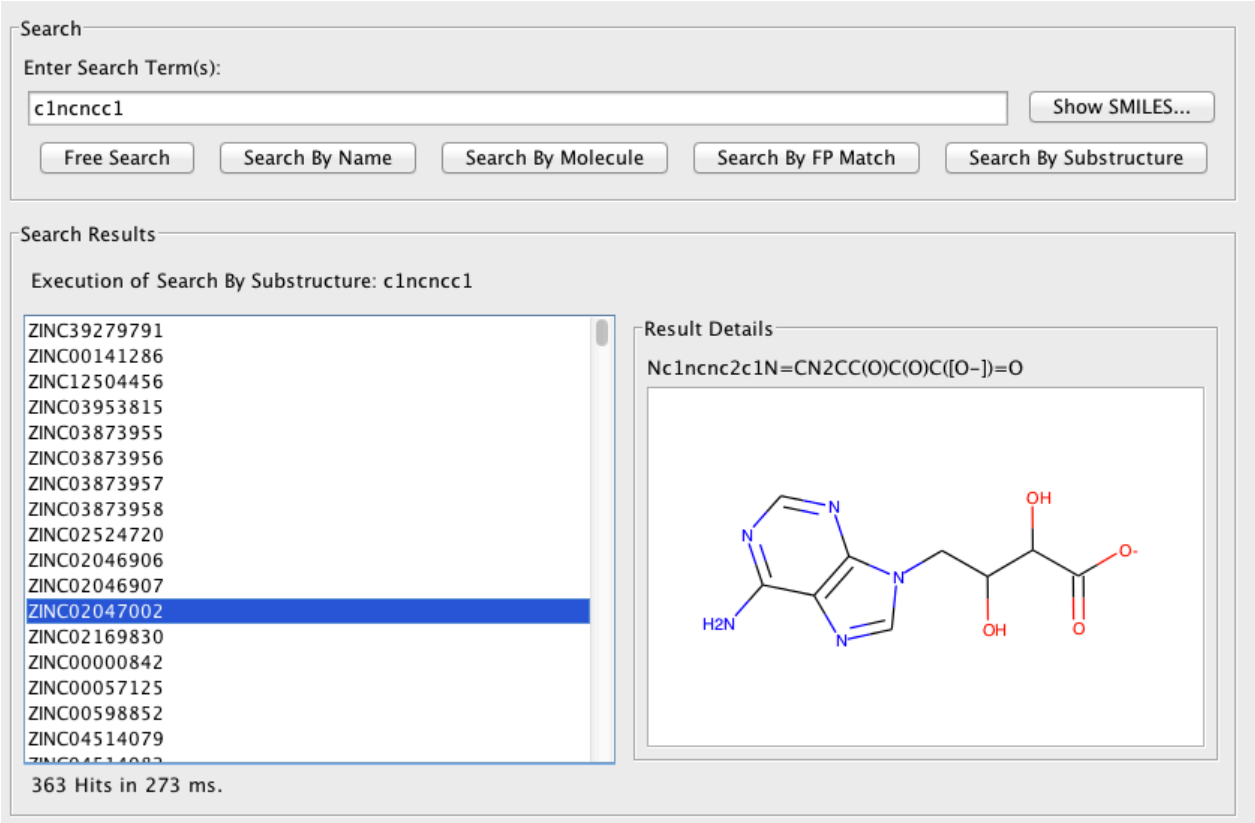
Out[11]:

	mutagenic							
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
molecule								
False	5217	0.461760	0.498583	0	0	0	1	1
True	1233	0.838605	0.368044	0	1	1	1	1

http://nbviewer.ipython.org/github/rdkit/UGM_2013/blob/master/Tutorials/pandastools/Pandas_RDKit_UGM.ipynb

Lucene integration

- Still in the experimental stage
- Adds substructure search functionality with fingerprint screenout to Lucene
- Includes demo app for testing



NOVARTIS

RDKit in NIBR

- Extensive use by CADD, informaticians, and IT
- Lots of convenience code/wrappers for accessing internal data sources and tools
- Combined with the Avalon toolkit (another NIBR-supported open-source project), provides the underpinning for many of our global chemistry-based applications

The Avalon toolkit

- C/Java cheminformatics toolkit
- Primary author: Bernd Rohde (NIBR Informatics Basel)
- http://sourceforge.net/projects/avalontoolkit/
- Functionality:
  - Canonical SMILES
  - Avalon fingerprint (highly optimized substructure fingerprint)
  - Molecular standardization (STRUCHK)
  - 2D Coordinate generation
  - Tomcat webapp for 2D rendering

- The RDKit has (optional) Python bindings for much of the functionality

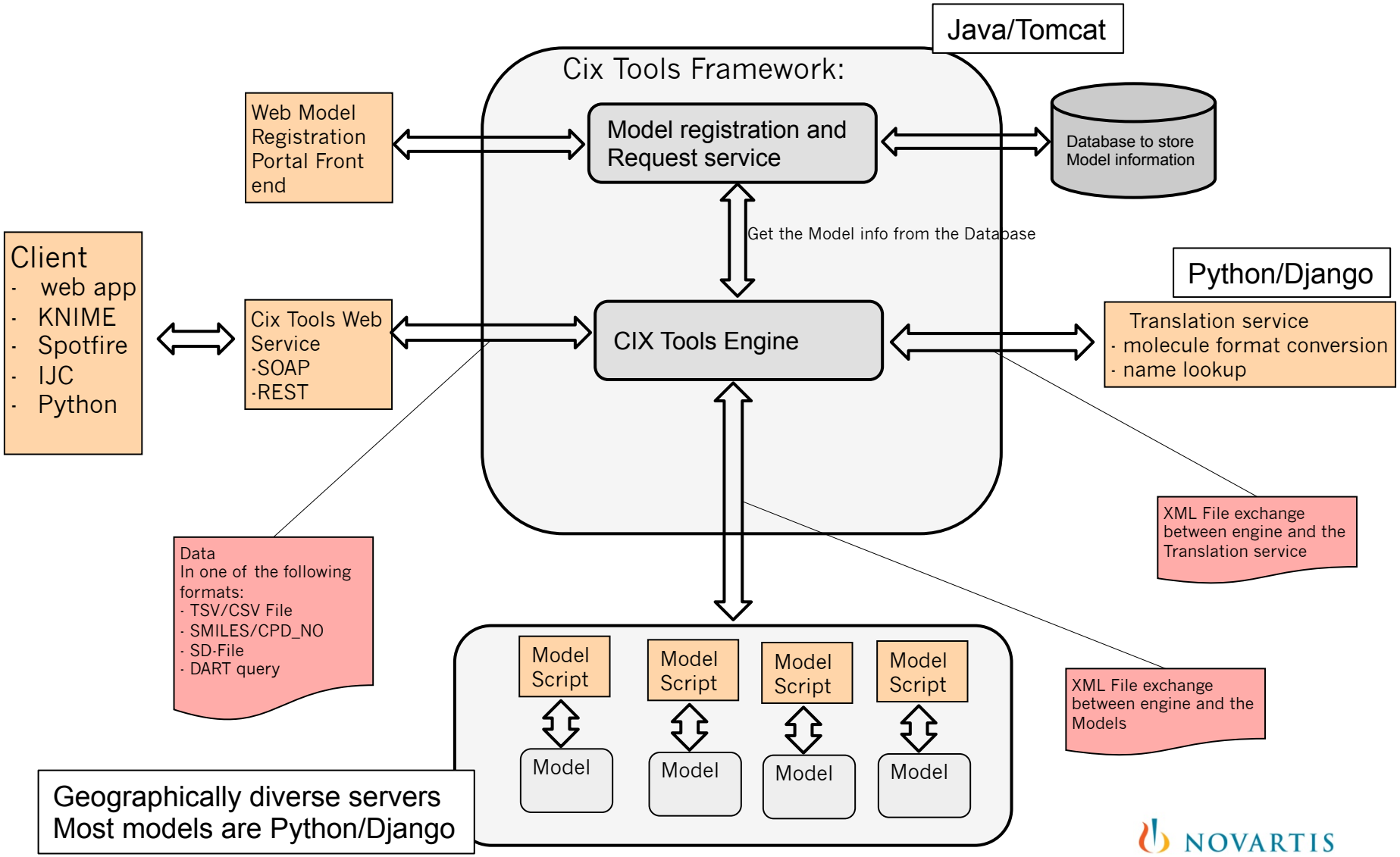
RDKit in NIBR

Case study 1: CIx Framework

- "Service bus" for cheminformatics/CADD services
- Handles format conversions for input/output automatically

  i.e. callers can provide SMILES input to a service/model wants CTABs with 3D coordinates
- Supports versioning of models/services
- Tight integration with scientific tools (e.g. Tibco Spotfire, Knime, Instant JChem, etc.)
- Enables trivial addition of "chemical intelligence" to web apps
- Makes it easy to globally deploy models: once a new model/service (or new version of a model/service) is registered with the Framework, it is instantly globally accessible

CIx Framework architecture



RDKit in NIBR

Case study 2: Small-Molecule Registration

NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH + RDKit

- Internally developed web application for compound registration
- C#-based web services writing to Oracle
- RDKit + Avalon toolkit for structure standardization
- RDKit + InChI used for structure-key calculation
- Calls out to CIx Framework for standard computed properties

- Independent (but validated) Python implementation of standardization and structure-key calculation for standalone use

RDKit in NIBR

Case study 3: QSAR Toolkit

- Descriptor calculator providing access to all available internal descriptors
- Tools for pulling assay data from our data warehouse
- Standardized model-building
- Standardized reporting for evaluation and peer review
- Packaging for deployment via CIx Framework
- Model Watchdog:

  Pulls most recent data, generates predictions, creates report showing evolution of model accuracy over time

RDKit in NIBR

Case study 4: Similarity Server

- Central PostgreSQL database with easily available compounds
  - in-house available
  - available from reliable vendors
- Kept up-to-date
- Substructure search
- Similarity search with various fingerprints:
  - Avalon
  - Morgan2, Morgan3, FeatMorgan2
  - Atom Pairs, Topological Torsions
- Web services interface
- Available to chemists via one of their standard desktop tools

- Currently deploying a new version based on chemfp

RDKit in NIBR

Case study 5: Compound Series and Favorites

- Central system to store definitions of the chemical series that project teams are working on
- Captures annotations and relationships
- Searchable, including Markush search (using ChemAxon cartridge)
- Accessible via web services
- RDKit used for scaffold validation

RDKit in NIBR

Case study 5: Compound Series and Favorites

Scaffold

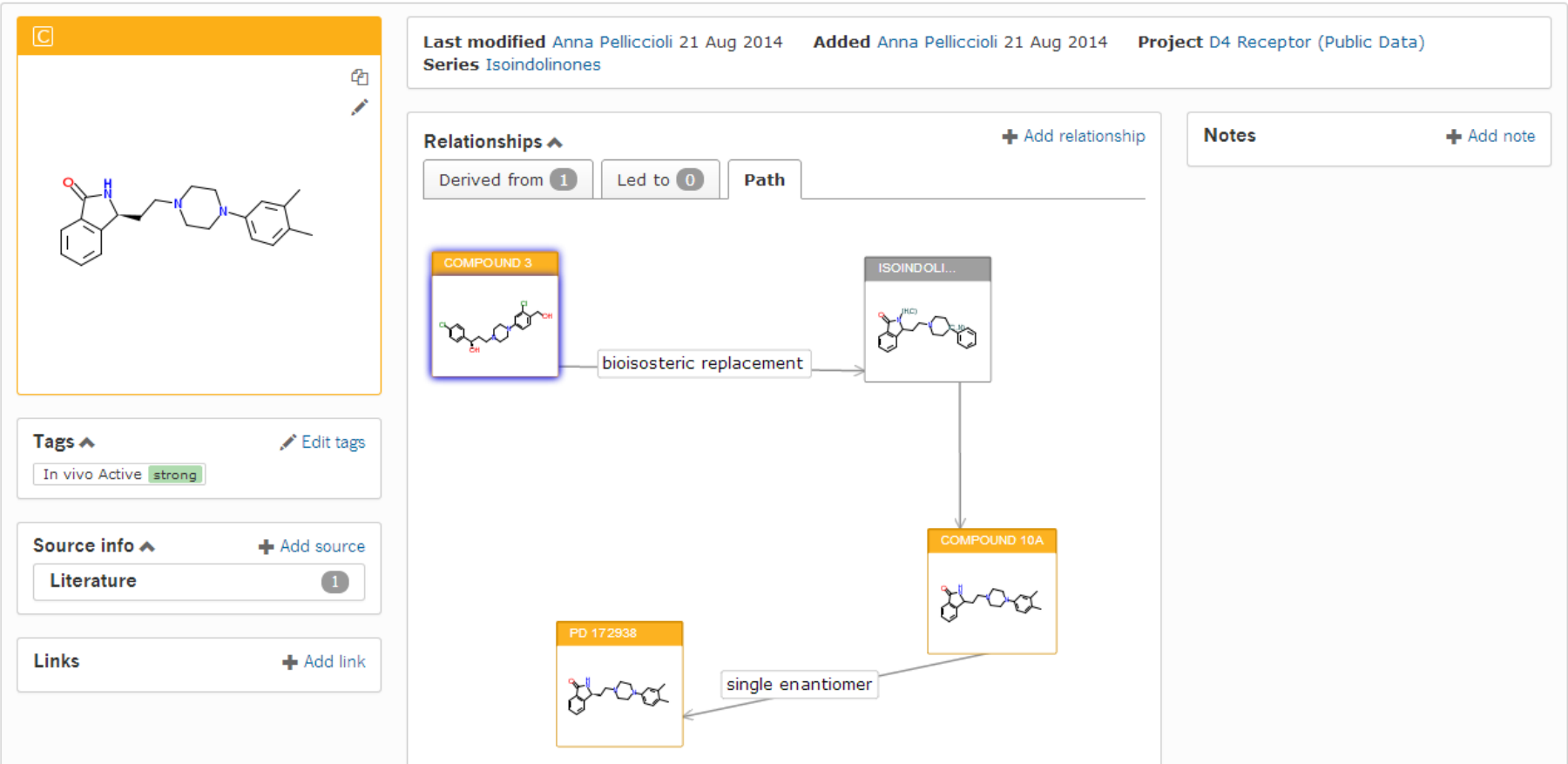
Compound

RDKit in NIBR

Case study 5: Compound Series and Favorites

NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH + RDKit

PD 172938



NIBR Open Source

Something new

opensource.nibr.com

Apps | Boost | DART-stage | DART | STL | Open in Papers | Other

NIBR Open Source | NIBR | Contact Us | NOVARTIS

Open Source at NIBR

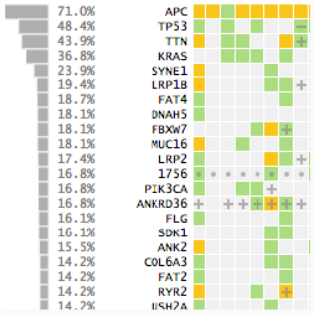
The Novartis Institutes for BioMedical Research (NIBR) is pioneering new informatics tools for drug discovery. We believe in the power of open-sourced, global collaboration for the greater good. Join us to help patients worldwide. Read about the work we do.

Interested in working for NIBR Engineering?

At NIBR, you'll be at the forefront of technology – helping to shape it, develop it, and make it impactful. Partnering with scientists, our engineers create cutting-edge, state-of-the-art solutions that accelerate drug discovery and ultimately improve patients' lives.

Learn More »

GridVar



GridVar is a jQuery plugin that visualizes multi-dimensional datasets

RDKit

RDKit

A collection of cheminformatics and machine-learning software written in

GitHub, Inc. [US] https://github.com/novartis

Boost | DART-stage | DART | STL | Open in Papers

Search or type a command | Explore | Gist | Blog | Help

Novartis

Find a repository...

gridvar JavaScript ★4 ⑂2

GridVar is a jQuery plugin that visualizes multi-dimensional datasets as layers organized in a row-column format.

Updated a month ago

Acknowledgements

http://www.rdkit.org

- General:
  - Remy Evard (NIBR/Informatics)
  - Richard Lewis (NIBR/GDC)
  - Tom Digby (NIBR/Legal)
  - Peter Gedeck (NIBR/GDC)
  - Nik Stiefl (NIBR/GDC)
- RDKit Community
  - Roger Sayle (NextMove): PDB Parser
  - Andrew Dalke (Dalke Scientific): FMCS
  - Paolo Tosco (University of Turin, now Cresset): MMFF94, Open3DAlign
  - Jameed Hussain (GSK, now CCG): Fraggle, mmpa
- Pandas, scikit-learn:
  - Sereina Riniker (NIBR/Informatics, now ETH)
  - Nikolas Fechner (NIBR/Informatics)
- Knime:
  - Manuel Schwarze (NIBR/Informatics)
  - Thorsten Meinl (knime.com)
  - Bernd Wiswedel (knime.com)
- SMR
  - Thomas Mueller (NIBR/Informatics)
  - Thomas Veith (NIBR/Informatics)
  - Dave Cotter (NIBR/Informatics)
- QSAR Toolkit:
  - Peter Gedeck (NIBR/GDC)
  - Nikolas Fechner (NIBR/Informatics)
- CIx Framework
  - Sandra Mueller (NIBR/Informatics)
  - Joerg Muehlbacher (NIBR/CPC)
  - Riccardo Vianello (NIBR/Informatics)
- Compound Series & Favorites
  - Anna Pelliccioli (NIBR/Informatics)
  - Manuel Schwarze (NIBR/Informatics)
  - Recca Chatterjee (NIBR/Informatics)
  - Mikhail Rybalkin (NIBR/Informatics)
  - Roman Bolshev (NIBR/Informatics)
- NIBR Open Source
  - Ken Robbins (NIBR/Informatics)
  - Dennis Jen (NIBR/Informatics)
  - Mark Schreiber (NIBR/Informatics)

NOVARTIS