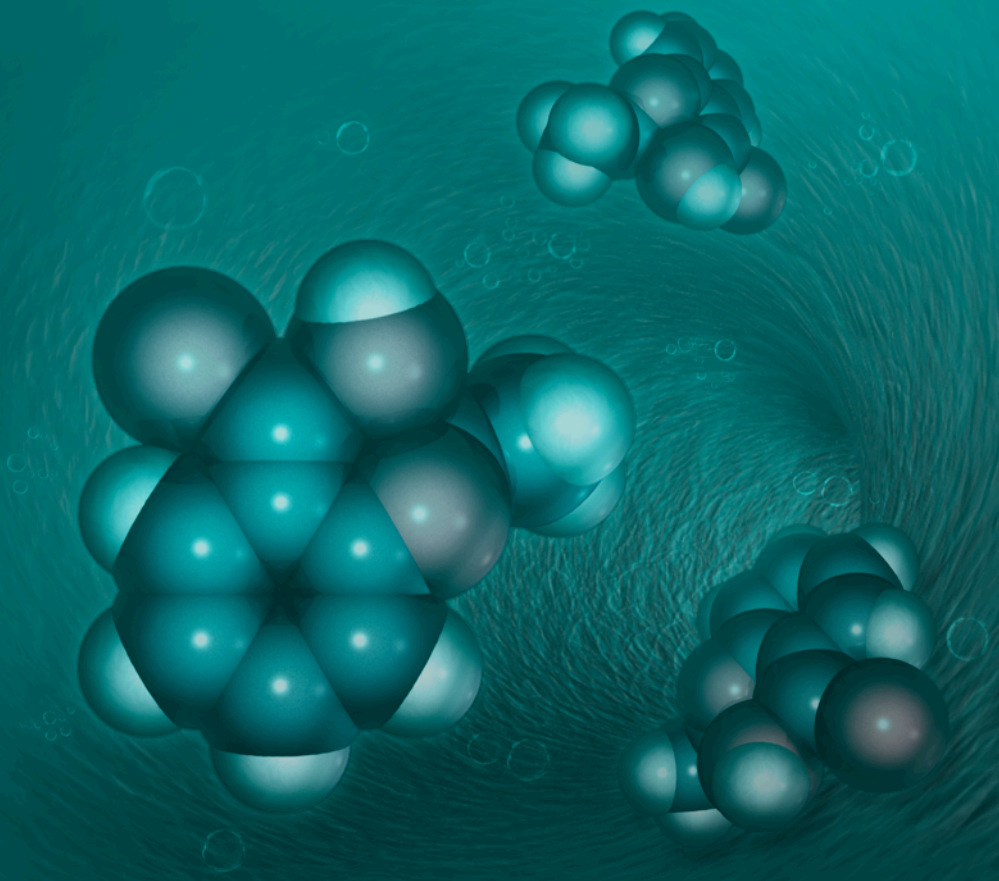
Patent Chemoinformatics with SureChEMBL

A talk-torial



George Papadatos
Senior Technical Officer
ChEMBL group, EBI
georgep@ebi.ac.uk

EMBL-EBI

Outline

- myChEMBL update
- SureChEMBL
- Talk-torial
  - Patent chemoinformatics with SureChEMBL and RDKit

myChEMBL: In a nutshell

- Ubuntu VM with:
  - Full ChEMBL db in Postgresql
  - RDKit database cartridge (latest stable version)
  - RDKit toolkit (latest stable version)
  - Custom ChEMBL web interface and web services
  - Web-based database/SQL browser
  - IPython (qtconsole & notebook)
    - Numpy, scipy, pandas, scikit-learn, etc.
    - Several IPython notebook examples
  - Accessible via http, ssh, postgres client, KNIME, etc.
  - Available in several formats

myChEMBL: Applications

- Centralised Resource
  - VM shareable across the local network
  - Access to standardised tools, services and data
- Application Development
  - Sandboxed VM, all source code available
- Learning
  - Lowers ‘activation barrier’ with pre-installed tools and examples
- Teaching, Training & Dissemination
  - IPython notebooks and KNIME workflows
  - 2nd Prize at ACS Teach-Discover-Treat competition

What's new in myChEMBL 19

- Ubuntu 14.04
- ChEMBL 19 database
- Latest python libs
- **Beaker server**
- 6 more IPython notebooks
  - Django ORM, Beaker, target druggability, drug ADME, machine learning, SureChEMBL
- Out later this week

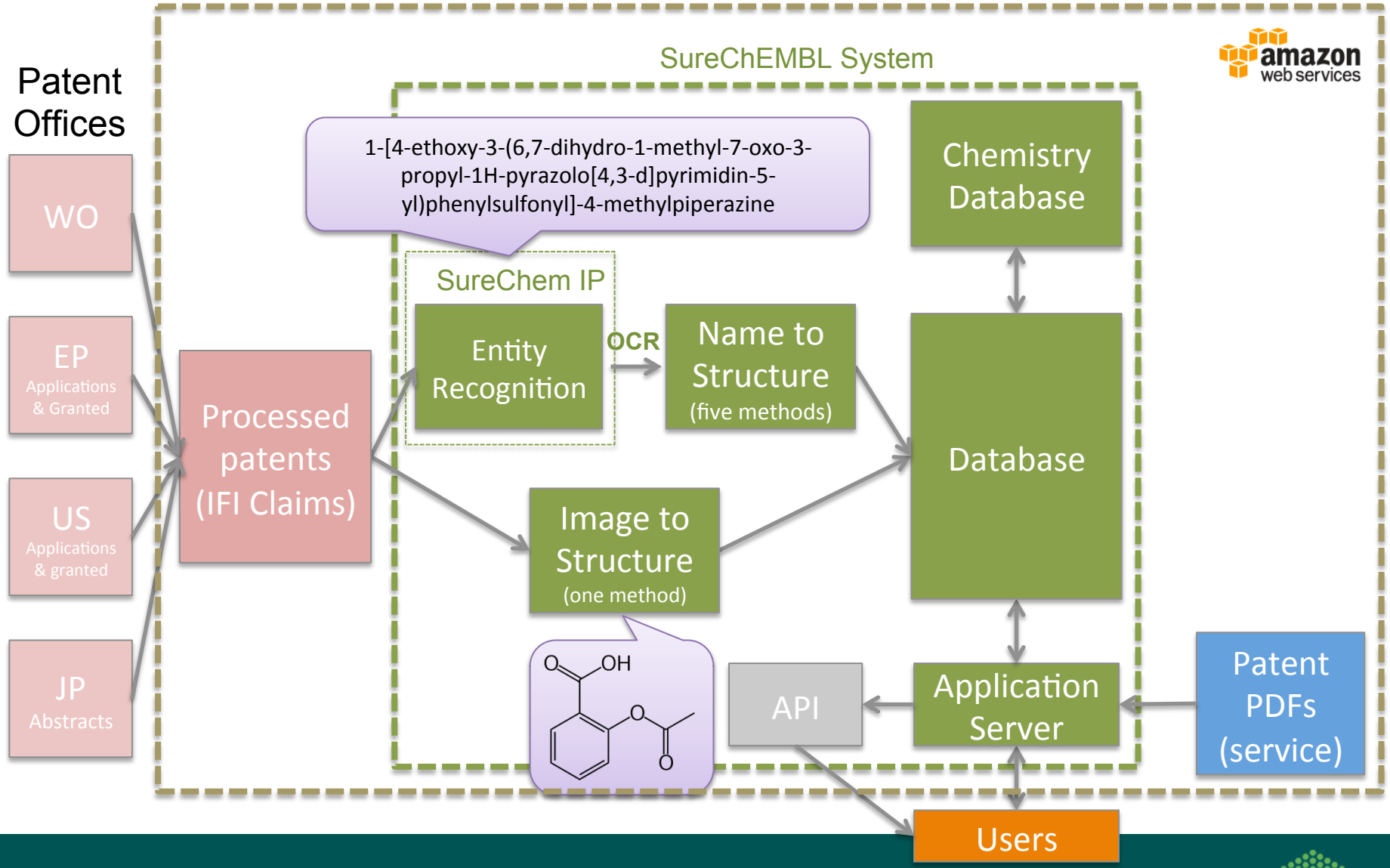
SURECHEMBL

SureChem => SureChEMBL

- December 2013 EMBL-EBI acquired SureChem – a leading ‘chemistry patent mining’ product from Digital Science, Macmillan Group
  - SureChem not aligned with core future academic business
- EMBL-EBI supported existing licensees during transition
- EMBL-EBI provides an ongoing, *open* resource to the entire community
  - Free and secure
  - No registration or login required
- Rebranded as **SureChEMBL**

Rebranding complete

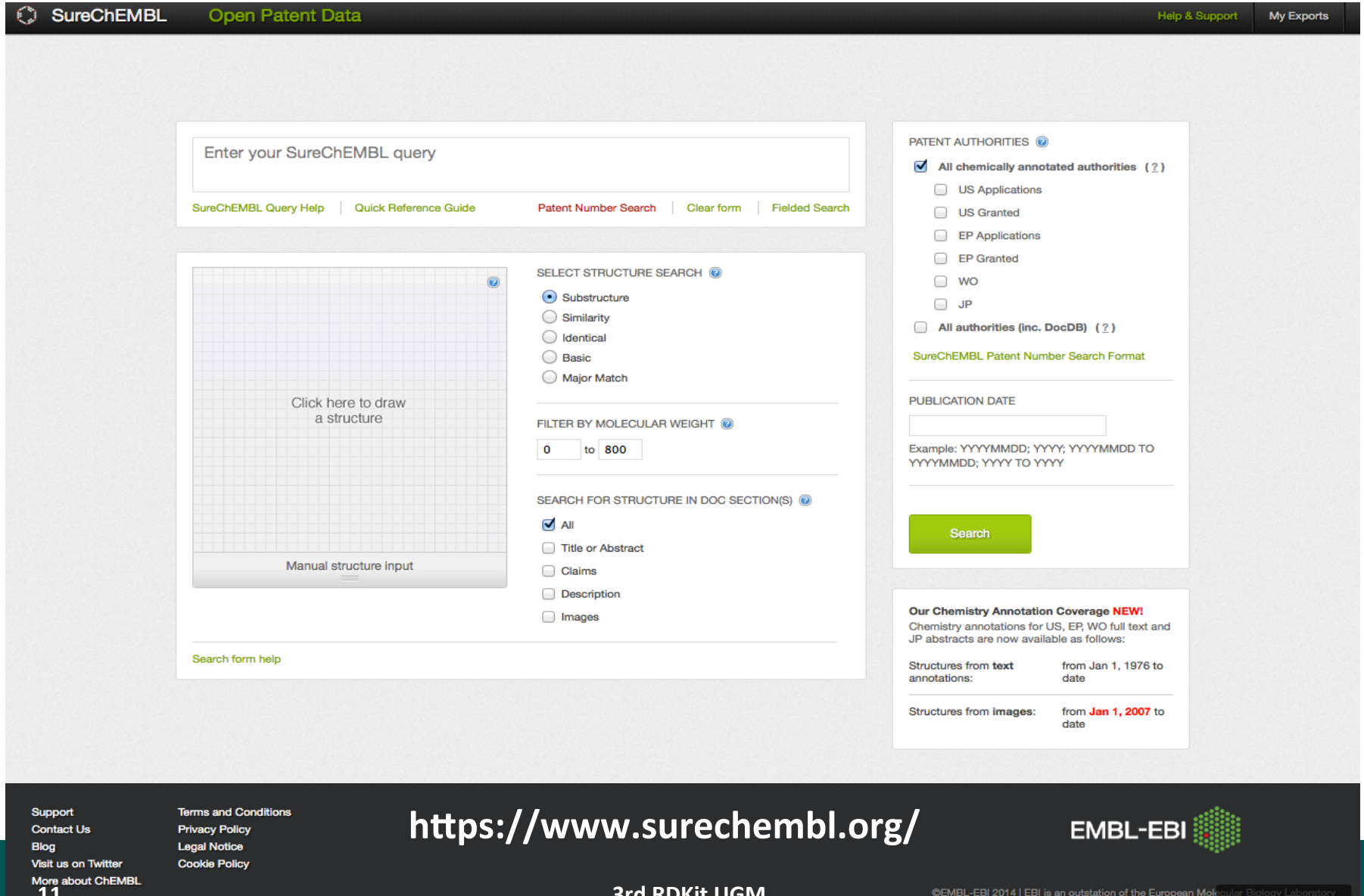
SureChEMBL data pipeline



SureChEMBL data content (last month)

- 15,668,225 unique compounds
- 12,888,125 *annotated* patents
- ~80,000 *novel* compounds extracted from ~50,000 new patents *monthly*
- 2–7 days for a published patent to be chemically annotated and searchable in SureChEMBL
- SureChEMBL provides search access to all patents (not just chemically annotated ones)
  - ~120M patents

SureChEMBL user interface



TALK-TORIAL!

Identify key compounds in patent docs

- What are key compounds in med. chem. patents?
  - ‘Best’, most promising/important compound
    - Drug candidate
    - Optimal property profile
    - Most active
- Automated identification of key compounds in patents
  - Using compound information only
  - Number of NNs, MCS extraction and R-group frequency

Hattori, K., Wakabayashi, H., & Tamaki, K. (2008). *JCIM*, *48*(1), 135–142. doi:10.1021/ci7002686
Tyrchan, C., Boström, J., Giordanetto, F., Winter, J., & Muresan, S. (2012). *JCIM*, *52*(6), 1480–1489. doi:10.1021/ci3001293

Workflow

1. Read a file with chemistry extracted from the Levitra family of patents
   - US6566360[1]
2. Filter by different text-mining and chemoinformatics properties to remove noise and enrich the genuinely novel structures
3. Visualise the chemical space using MDS and dimensionality reduction
   - Identify and fix outliers in the chemistry space.
4. Find the Murcko and MCS scaffolds
5. Compare the derived MCS core with the actual Markush structure
6. Identify key compounds using structural information only
   - Count number of NNs per compound
   - R-Group decomposition and frequency of R-Groups

[1]Sayle, R. *et al*. (2012). *JCIM*, *52*(1), 51–62. doi:10.1021/ci200463r

Go to Notebook

Summary

- Export chemistry from SureChEMBL
- Filter to get to the claimed compounds
- Errors may occur due to OCR
- Markush structures can be approximated via an MCS core
- PoC to identify key compounds prospectively
  - Using open data and tools

surechembl-help@ebi.ac.uk

Acknowledgements

- ChEMBL team
  - John Overington
  - Jon Chambers
  - George Papadatos
  - Mark Davies
  - Nathan Dedman
  - Anna Gaulton
- Digital Science
  - Nicko Goncharoff
  - James Siddle
  - Richard Koks
- RDKit Community

Funding:

Innovative Medicines Initiative Joint Undertaking, grant agreement no. 115191 (Open PHACTS)

Wellcome Trust Strategic Award for Chemogenomics, WT086151/Z/08/Z

European Molecular Biology Laboratory

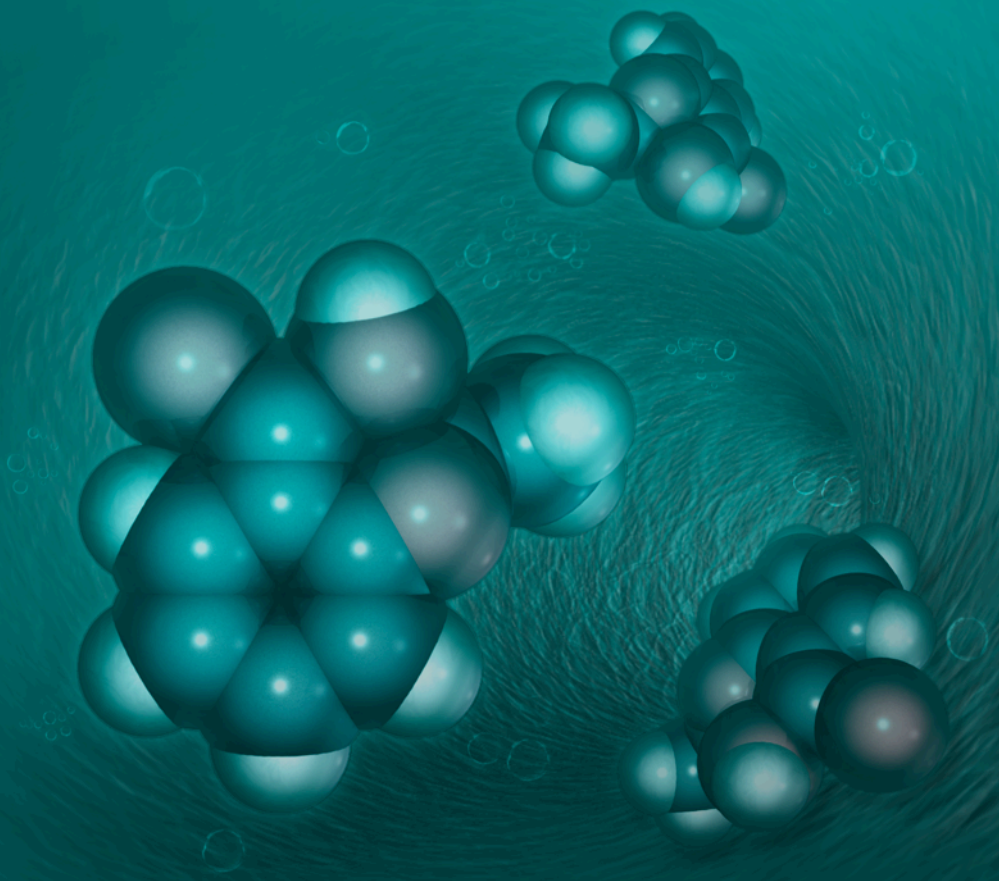
European Commission FP7 Capacities Specific Programme, grant agreement no. 284209 (BioMedBridges)

Software:

ChemAxon, DIGITAL science, ACD/Labs, keymodule, OpenEye Scientific Software, IFI CLAIMS® Patent Services, OPSIN, PerkinElmer, NextMove Software, Twin Dolphin Software

Patent Chemoinformatics with SureChEMBL

A talk-torial



George Papadatos
Senior Technical Officer
ChEMBL group, EBI
georgep@ebi.ac.uk

EMBL-EBI