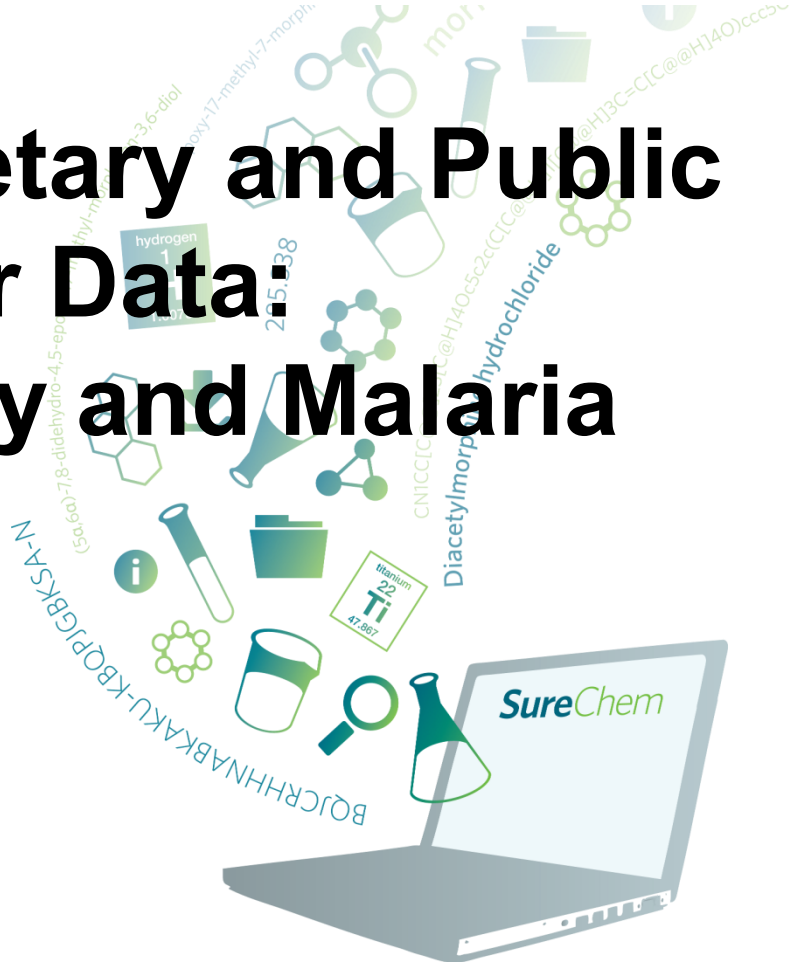
Integrating Proprietary and Public Molecular Data: Patent Chemistry and Malaria

Dr Andrew Hinton
Ph.D.

ICIC Oct 2011



DIGITAL science

www.surechem.org

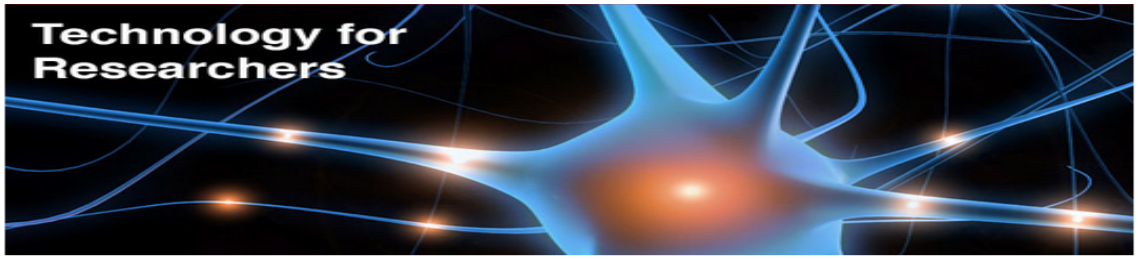
SureChem

Who is SureChem ?

DIGITAL science

nature publishing group

macmillan



A new kind of scientific information company

The digital age has created new opportunities for the way we do scientific research - from how we manage our labs and analyse data, to how we link information and make funding decisions. We're here to help you make the most of those possibilities.

LATEST NEWS

BioData Acquires LabLife, Empowering Life Scientists with Advanced web-based Research Management service

TEL AVIV, September 13, 2011 — BioData Ltd. announced today that it has acquired all the assets of LabLife Software Inc., furthering its strategy to become the leading research and lab management software and service provider. The acquisition follows an investment in BioData by Digital Science in December 2010. Digital Science is a division of Macmillan Publishers Ltd, owners of Nature Publishing Group.

→ More

EVENTS

ICIC — International Conference for the Information Community

Date: *23-26 Oct 2011*
Type: *Conference*
Country: *Barcelona, Spain*
Digital Science will be presenting and sponsoring at this years ICIC conference. Andrew Hinton, Cheminformatics Developer at SureChem, will be holding a talk on Integrating Proprietary and Public Collections of Pharmaceutically Relevant Small Molecules in the Patent Literature with Reference To Neglected Diseases.

→ More

BLOG

Welcoming Labtiva, and unveiling their product ReadCube

Today we're happy to announce both our investment in Boston based start-up Labtiva as well as the global release of their first product 'ReadCube' — a software application designed to provide researchers a more efficient and intuitive means of managing and discovering relevant literature in their field. The company joins Digital Science's portfolio of research tools as the first content management offering.

→ More

PRODUCTS

readcube — Readcube is a free, cross-platform desktop application that enables researchers to create and manage their personal content library, and intuitively discover new literature in their field through daily recommendations.

FigShare — FigShare allows researchers to publish all of their data in a citable, searchable and sharable manner.

1DegreeBio — 1DegreeBio is an open-access platform that enables researchers to source and review the highest-quality products for their research needs.

SYMPLECTIC ELEMENTS — Symplectic Elements automatically records research outputs for academics and integrates with other research management systems.

SureChem — SureChem Portal enables advanced chemical search of the world's patent literature.

BioData — BioData offers a modular, web-based laboratory research management system.

DIGITAL science

www.surechem.org

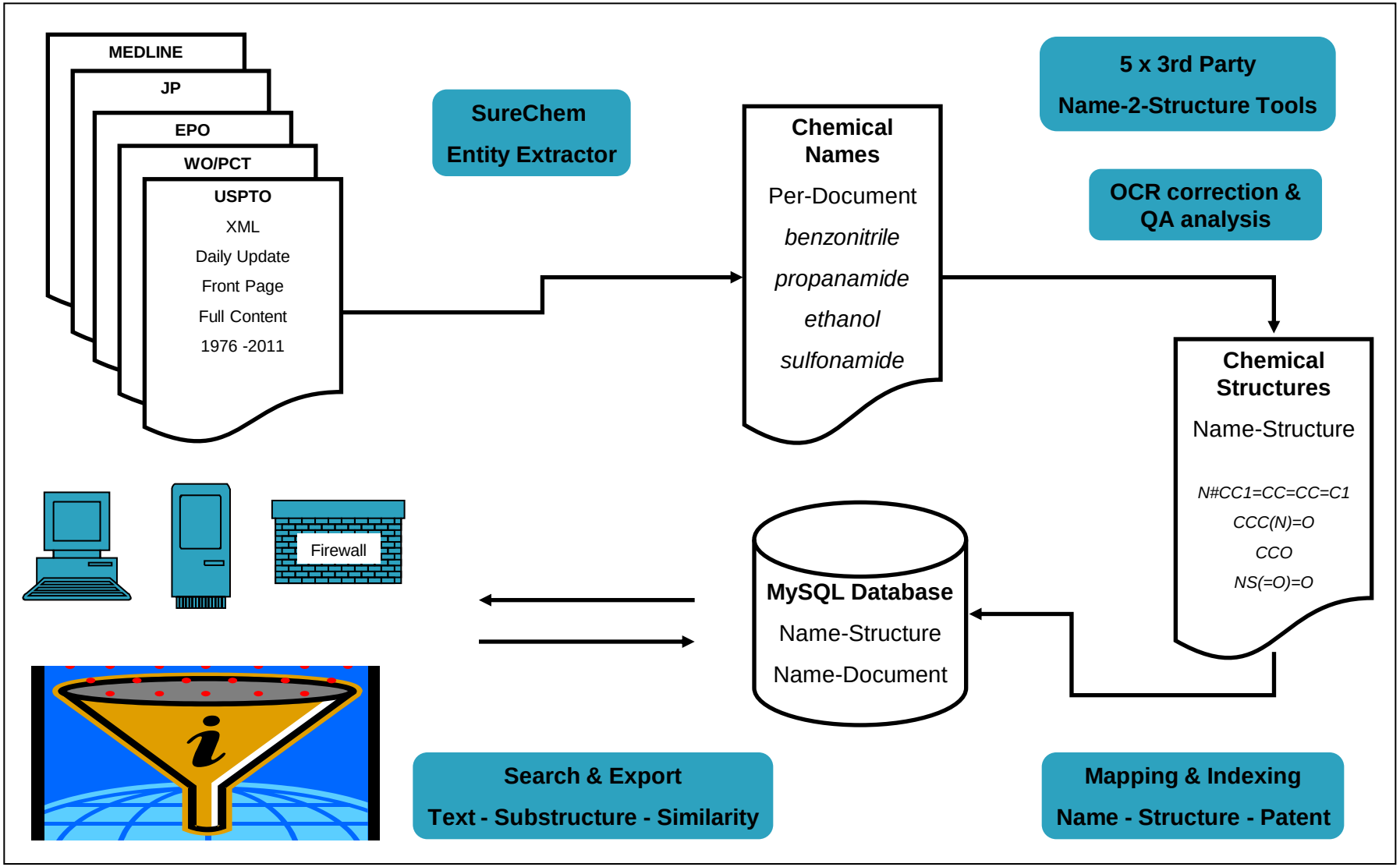
SureChem

SureChem at a Glance

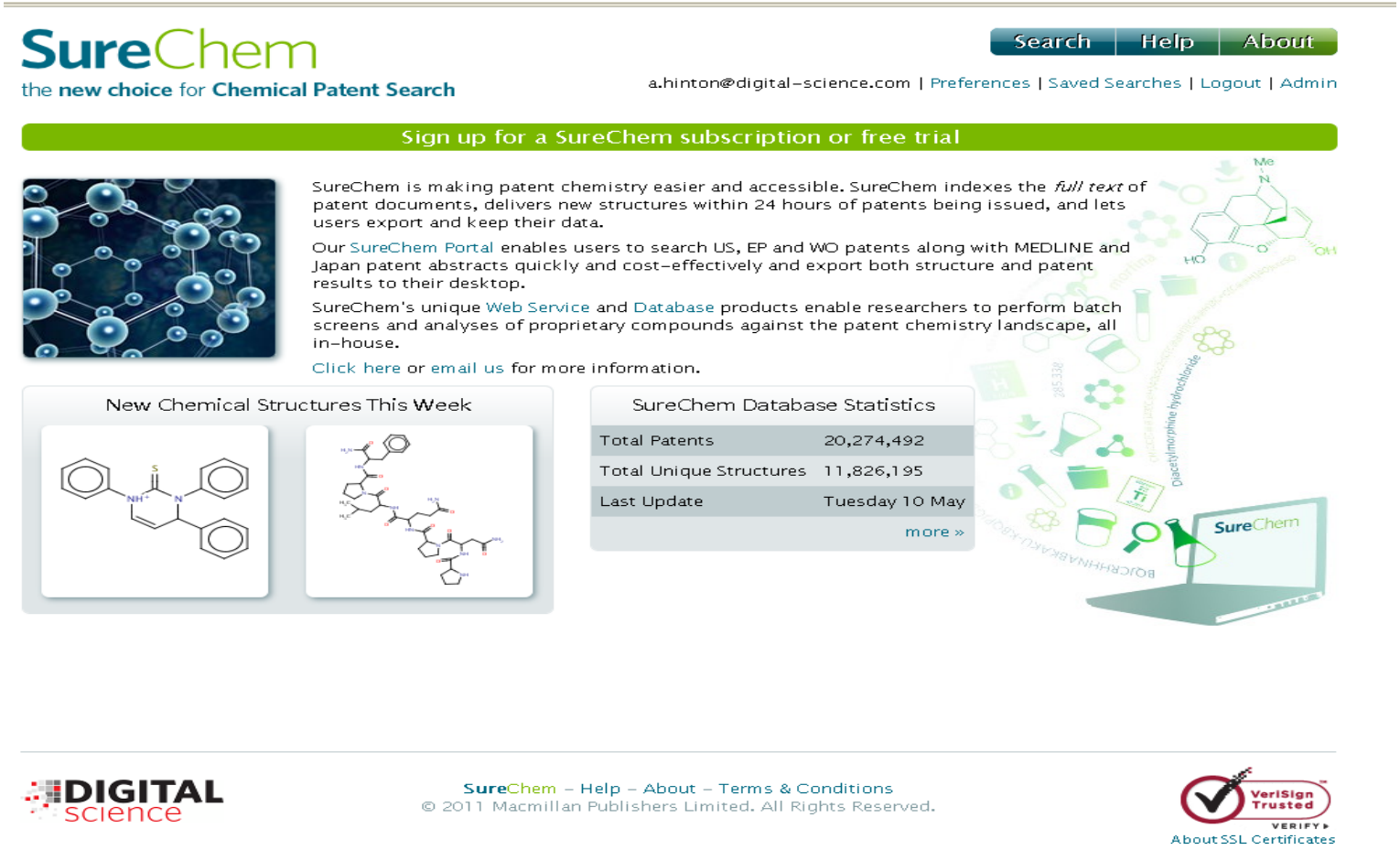
- Structure and text searchable database of
  - **USPTO** applications/grants (from 1976)
  - **EP** applications/grants (from 1986 soon to be from 1978)
  - **WO** applications (from 1978)
  - **JP** patent abstracts (from 1976)
  - **MEDLINE** abstracts

- **12 million unique structures**
- **20 million patents, 18.5 million MEDLINE abstracts**
- **Structures indexed from full text of patent document**
- **Updated within 24 hours of patent publication!**

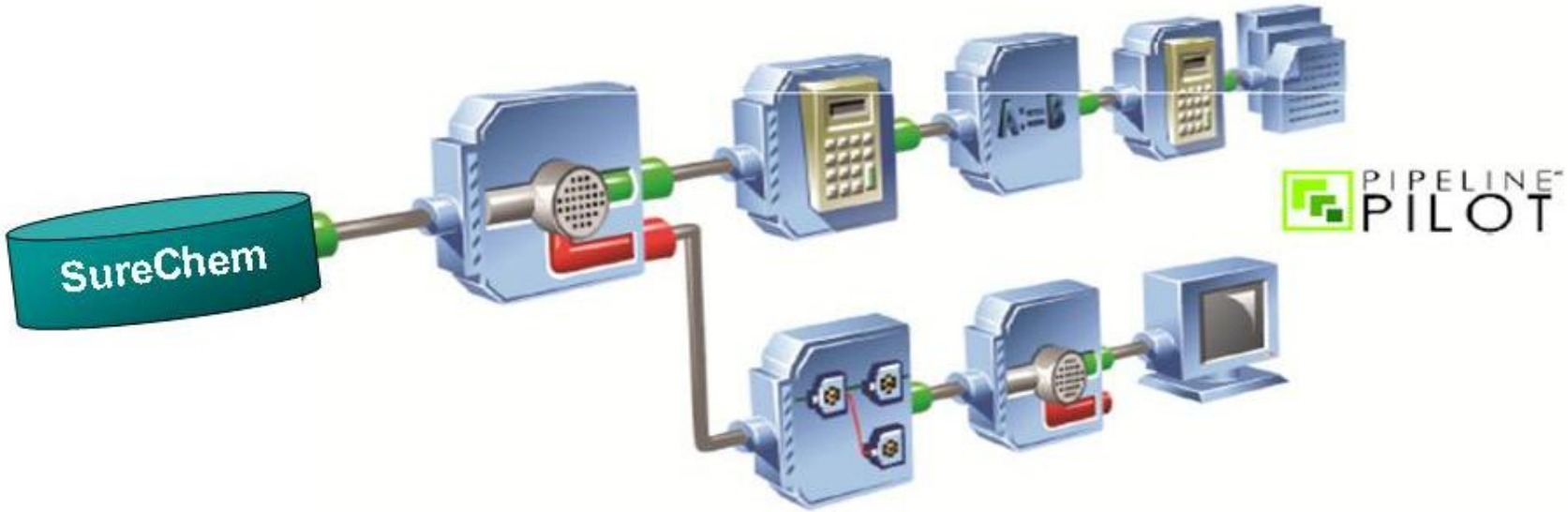
How do we do it



http://www.surechem.org

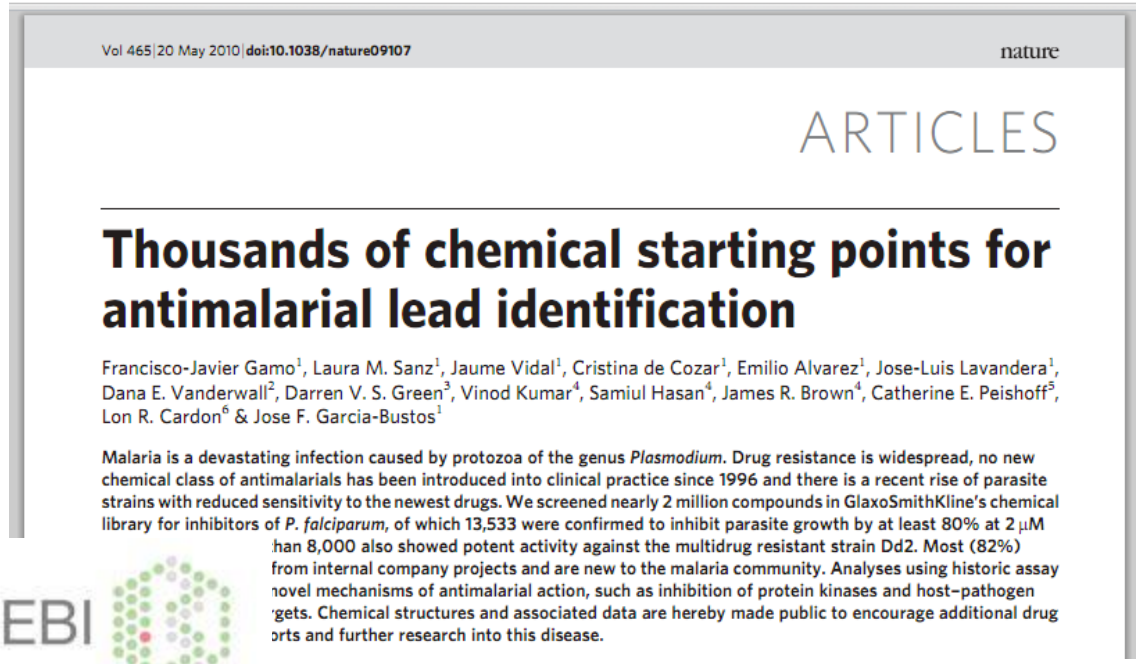


SureChem: An Accelrys Partner



www.accelrys.com

Background



Vol 465 | 20 May 2010 | doi:10.1038/nature09107 nature

ARTICLES

Thousands of chemical starting points for antimalarial lead identification

Francisco-Javier Gamo[1], Laura M. Sanz[1], Jaume Vidal[1], Cristina de Cozar[1], Emilio Alvarez[1], Jose-Luis Lavandera[1], Dana E. Vanderwall[2], Darren V. S. Green[3], Vinod Kumar[4], Samiul Hasan[4], James R. Brown[4], Catherine E. Peishoff[5], Lon R. Cardon[6] & Jose F. Garcia-Bustos[1]

Malaria is a devastating infection caused by protozoa of the genus *Plasmodium*. Drug resistance is widespread, no new chemical class of antimalarials has been introduced into clinical practice since 1996 and there is a recent rise of parasite strains with reduced sensitivity to the newest drugs. We screened nearly 2 million compounds in GlaxoSmithKline's chemical library for inhibitors of *P. falciparum*, of which 13,533 were confirmed to inhibit parasite growth by at least 80% at 2 μM han 8,000 also showed potent activity against the multidrug resistant strain Dd2. Most (82%) from internal company projects and are new to the malaria community. Analyses using historic assay novel mechanisms of antimalarial action, such as inhibition of protein kinases and host–pathogen gets. Chemical structures and associated data are hereby made public to encourage additional drug orts and further research into this disease.

EMBL-EBI

gsk GlaxoSmithKline

NATIONAL LIBRARY OF MEDICINE

CDD COLLABORATIVE DRUG DISCOVERY

Francisco-Javier Gamo et al
***Nature* 465, 305–310 (20 May 2010)**

DIGITAL science

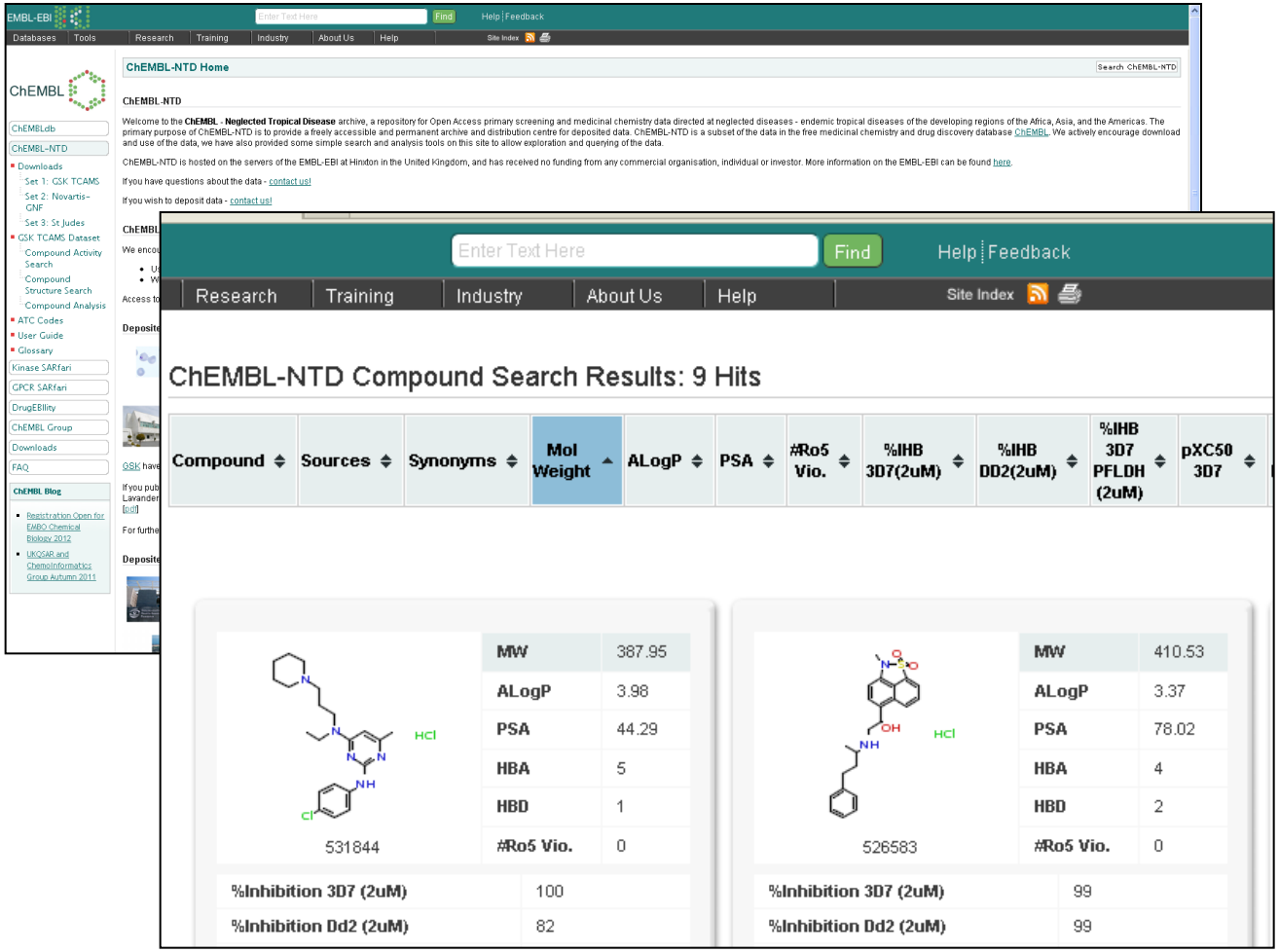
www.surechem.org

SureChem

ChEMBL-NTD

ChEMBL

https://www.ebi.ac.uk/chemblntd



GlaxoSmithKline

Genomics Institute of the Novartis Research Foundation

St. Jude Children's Research Hospital
ALSAC • Danny Thomas, Founder

DIGITAL science

www.surechem.org

Research Aims

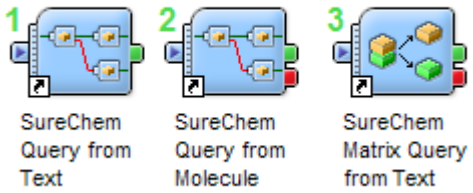
1. Generate a list of patents using the chemical compounds deposited in ChEMBL NTD for release to Research community

2. Prioritise Patents in terms of chemical and biological relevance associated with anti-Malaria treatment

3. Demonstrate that the additional data form SureChem add sbenefit to that currently accessible from ChEMBL NTD

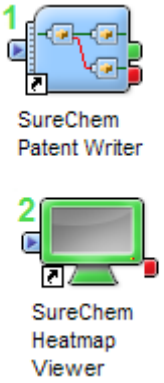
SureChem Pipeline Pilot API

1. ***Construct Query***
   - Text
   - Chemical
   - Text & Chemical



2. ***Search***
   - Select Patent Authority
   - Perform
   - e.g. Exact, Substructure, Similarity or Patent No.

3. ***Display Results***
   - Structures
   - Patent information
   - Heatmap



Chemical Search Sets

Set Name	Source	No. Unique of Structures	Search Type
Full	TCAMS+GNF+ StJudes +FDA	19699	Sim 0. 95
TCAMS_filtered	Med Chem Friendly TCAMS Structures	1484	Similarity 0.9
GNF_filtered	Med Chem Friendly GNF Structures	1102	Similarity 0.9
St Judes_filtered	Med Chem Friendly StJudes Structures	207	Similarity 0.9
FDA Drugs	Literature	17	Exact/ Sim/ SubS
TCAMS_Murcko	Scaffolds	415	SubS
TCAMS_Div	Cluster Representatives	2834	Similarity 0.9

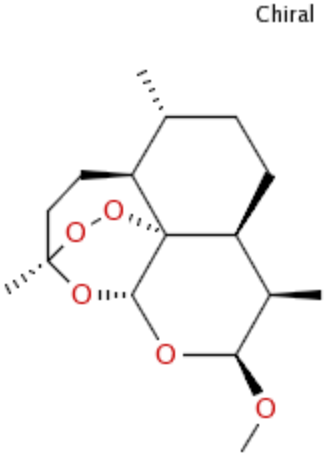
Filtering

- Remove unwanted compounds with structural features or PhysChem 'Lead-like' properties

- Filters
  1. (soft) Compounds removed containing only the very 'worst' Structural features
  2. (Medium) Compounds removed due to unwanted Structural features only
  3. (Strict) Compounds removed due to unwanted Structural & PhysChem features

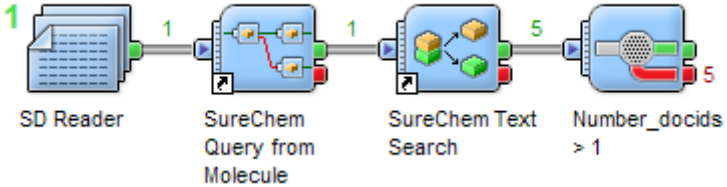
Pass Rate	#	Filter 1	Filter 2	Filter 3
TCAMS	13358	97%	56%	11%
GNF	5373	89%	51%	20%
St Judes	1523	91%	46%	13%

Exact Structure Searching



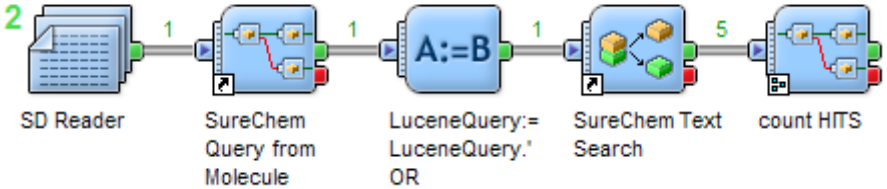
FDA Approved Drug: Artemether

0 patents



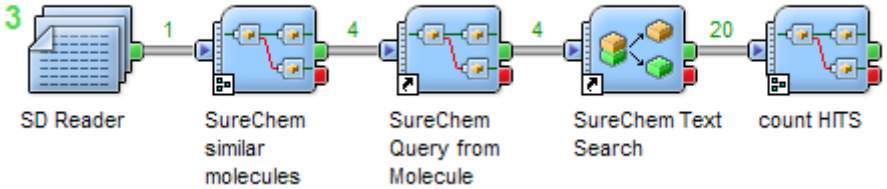
Exact Molecular Matching

841 patents



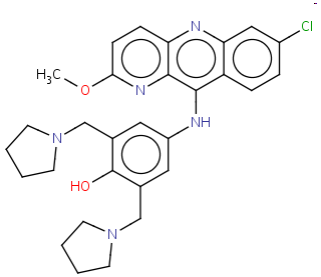
Exact Molecular or Name Matching

841 patents

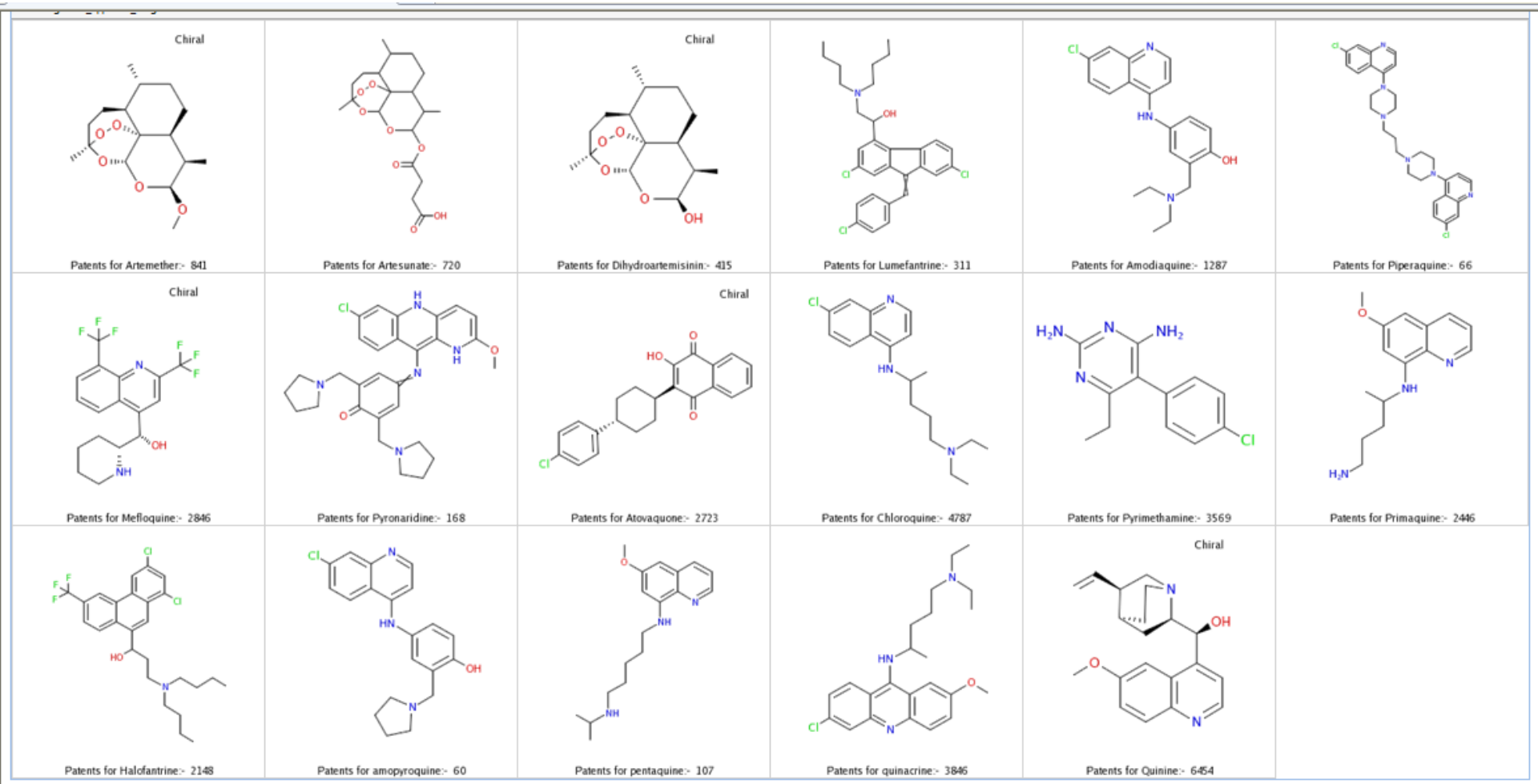


Non-Stereo Molecular Matching

Exact Structure Searching



pyronaridine

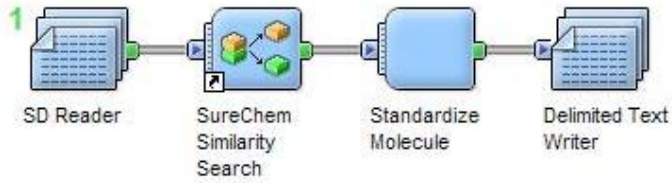


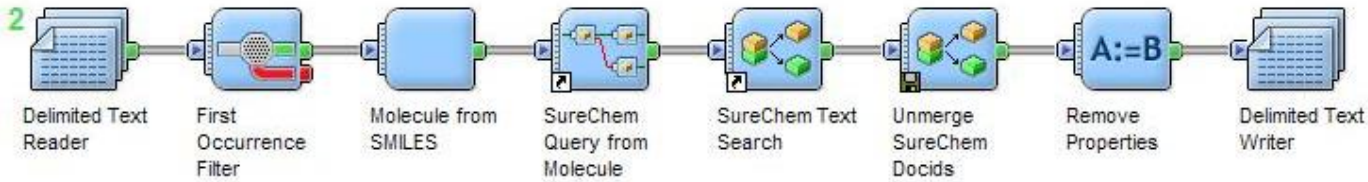
Prioritising Patents

- Any 50% of search structures can be found together in a single patent in 1% the patent hits

Total No of patent = 19029	Count of Individual Patents Containing
1 search structure	13714
Between 2 - 4 search structure	4100
Between 5 - 8 search structures	1026
Between 9 - 12 search structures	168
Between 13 - 16 search structures	21
All 17 search structures	0

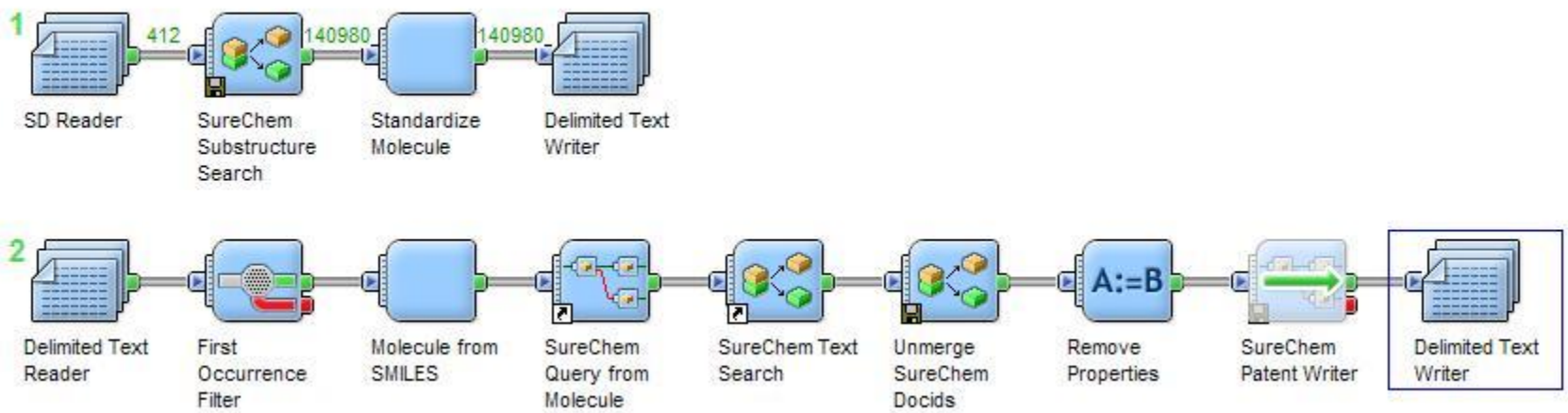
Similarity Search Results





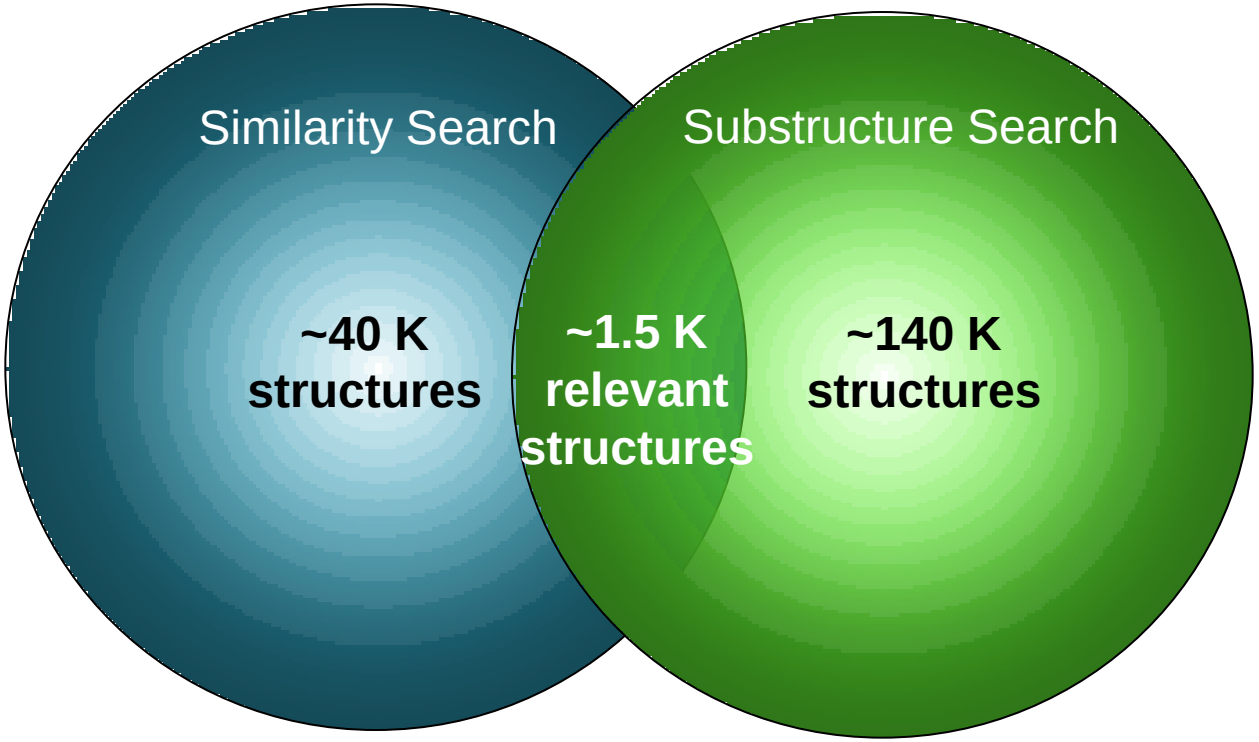
Set Name	No. of Search Structures	Tanimoto Cut Off	No. Structure Hits	No. Patents Hits
Full	19699	0. 95	39,733	496,162
TCAMS_filtered	1484	0.9	10,753	185,278
GNF_filtered	1102	0.9	11,475	119,955
St Judes_filtered	207	0.9	1,042	14,526
TCAMS_Div	2834	0.9	1,955	11,758

Substructure Searching



Set Name	No. of Search Structures	No. Structure Hits	No. Patents Hits
FDA	11	7574	50,893
TCAMS_Murcko	412	140980	534,520

Combined Chemical Result Sets



3068 Patent from 1542 structures with overlapping chemistry from scaffold substructure and close overall similarity

Structures with 10% of patents also continuing the term 'Malaria' or 'Falciparum'

DIGITAL science

SureChem

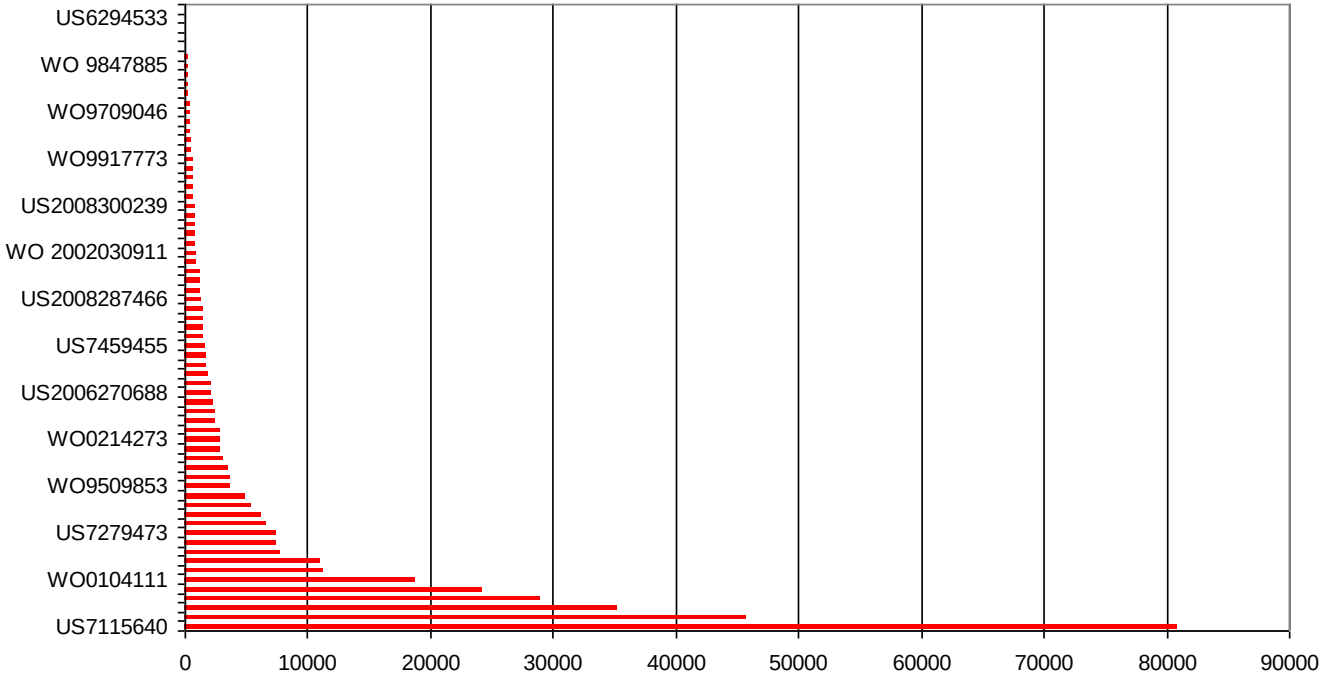
Prioritising Patents by Chemistry

- Metrics
  - The count of co-occurrence of different hits within a single patent.
  - Frequency of a given structure within patent compared to baseline.

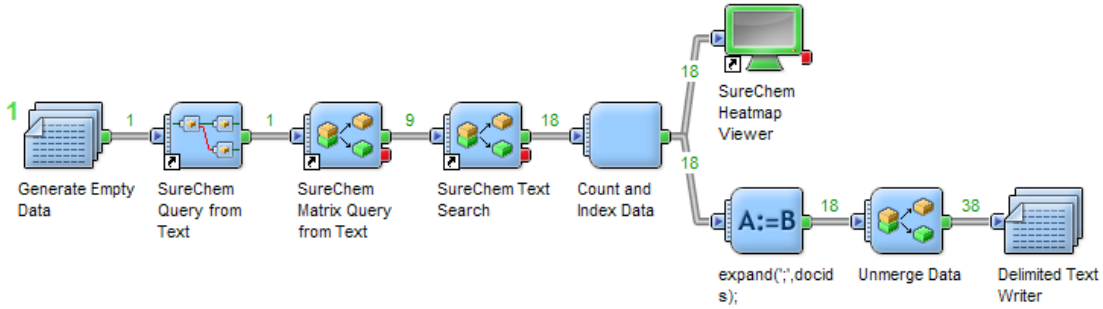
- Metrics + other approaches
  - Use of chemical filters and rules of thumb
  - Combined substructure, similarity and keyword
  - Natural Language Processing approach to identify exemplified compounds

Patent Result Sets

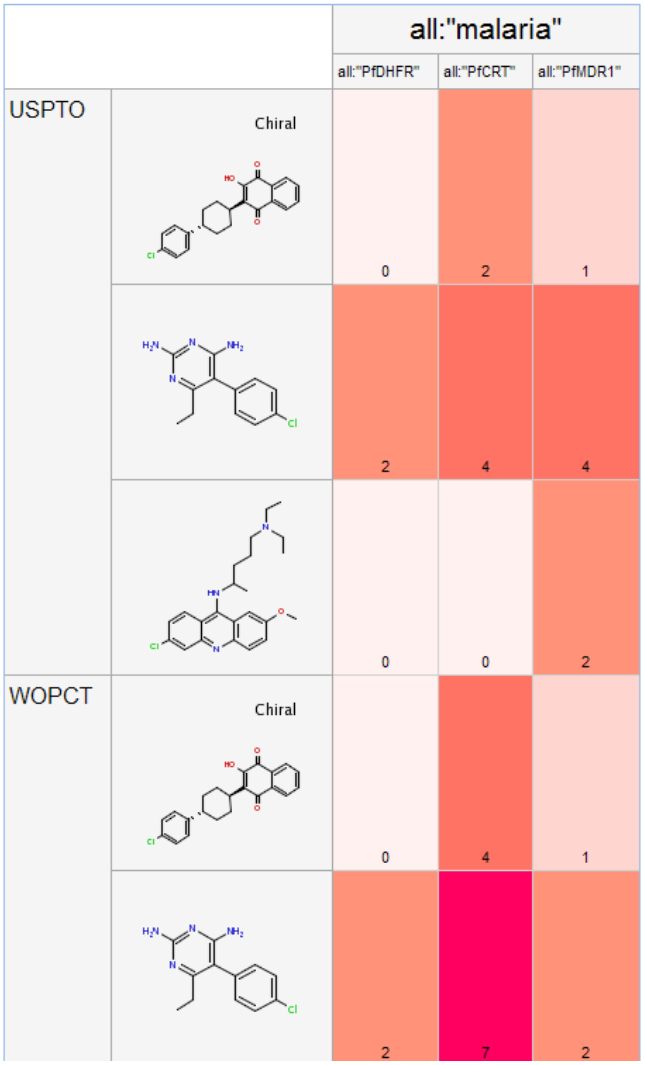
- TCAMS author's identified 404 patents with structures present in in the dataset



Combined Chemical and Biological Searches on Patents



- Genes for P. falciparum
  - PfCRT CQ resistance transporter
  - PfMDR1 P-glycoprotein homolog 1
  - PfDHFR dihydrofolate reductase
- Structures
  - Atovaquone
  - Pyrimethamine
  - quinacrine
- Keyword
  - Malaria



		all:"malaria"		
		all:"PfDHFR"	all:"PfCRT"	all:"PfMDR1"
USPTO	Chiral	0	2	1
		2	4	4
		0	0	2
WOPCT	Chiral	0	4	1
		2	7	2

J Yuan et al; Science 333, 724.

DIGITAL science

www.surechem.org

SureChem

Future Data Release to Malaria Community

- Make publicly available structures and patent IDs

- Likely sets
  - Structures from ‘Exact’ search of full ChEMBL-NTD
  - Hits from combined Substructure & Similarity searches
  - Patents filtered by relevant keywords
  - Patents from top bio-active molecules (to-do)

- Discuss with key stake holders as to mode and method of release

Acknowledgements

- Dr John Overington, ChEMBL
- Dr Ashely George GSK
- Dr Yingyao Zhou GNF

- Digital-Science
  - Nicko Goncharoff
  - Richard Koks
  - James Siddle

http://www.digital-science.com

E: info@digital-science.com

London office:

Digital Science
The Macmillan Building
4 Crinan Street
London
N1 9XW

A.Hinton@digital-science.com