

Revisiting ChemAxon's Name-to-Structure Tool in Patent Data

Dr Andrew Hinton Ph.D

A.Hinton@digital-science.com

Agenda

- ▶ The Data Environment – SureChem in Brief
- ▶ Recap on 2009 Analysis
- ▶ 2011 Analysis
 - Coverage
 - Rates of Agreement
 - ChemAxon Version Contributions
- ▶ Conclusions

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!**

http://www.surechem.org

SureChem

the new choice for **Chemical Patent Search**

Search

Help

About

a.hinton@digital-science.com | [Preferences](#) | [Saved Searches](#) | [Logout](#) | [Admin](#)

Sign up for a SureChem subscription or free trial



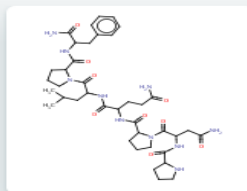
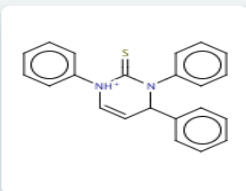
SureChem is making patent chemistry easier and accessible. SureChem indexes the *full text* of patent documents, delivers new structures within 24 hours of patents being issued, and lets users export and keep their data.

Our [SureChem Portal](#) enables users to search US, EP and WO patents along with MEDLINE and Japan patent abstracts quickly and cost-effectively and export both structure and patent results to their desktop.

SureChem's unique [Web Service](#) and [Database](#) products enable researchers to perform batch screens and analyses of proprietary compounds against the patent chemistry landscape, all in-house.

[Click here](#) or [email us](#) for more information.

New Chemical Structures This Week



SureChem Database Statistics

Total Patents	20,274,492
Total Unique Structures	11,826,195
Last Update	Tuesday 10 May
	more »



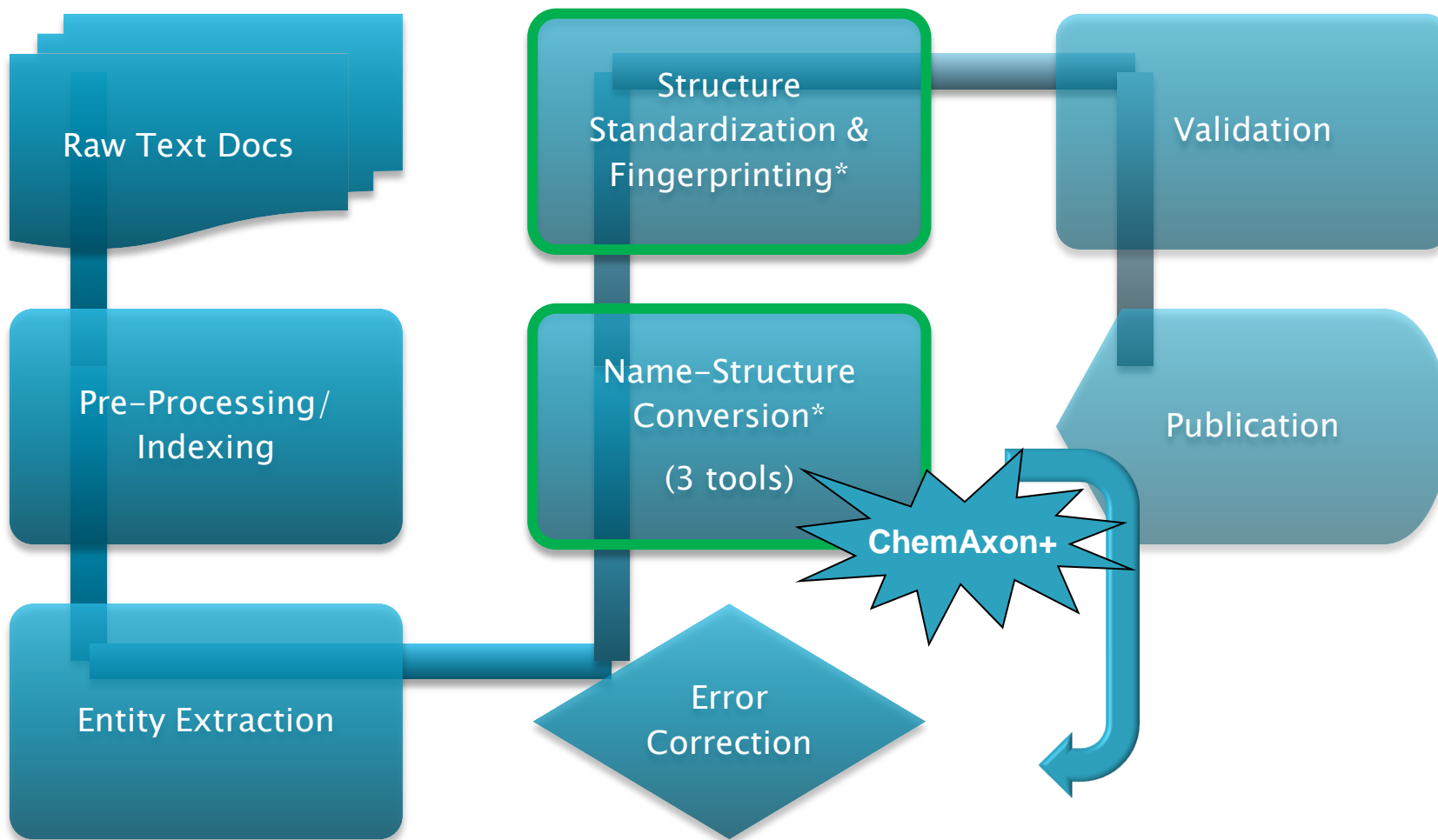
DIGITAL
science

[SureChem](#) – [Help](#) – [About](#) – [Terms & Conditions](#)
© 2011 Macmillan Publishers Limited. All Rights Reserved.



SureChem

SureChem Workflow



*Includes use of 3rd party software

Name-2-Structure Challenges

- ▶ **OCR errors**
 - Transposed letters and numbers, insertion of special characters instead of plain text, etc.
- ▶ **Name can't be resolved to a structure**
 - Inorganic compounds, chemical groups
- ▶ **Incorrect Nomenclature**
 - Ambiguous or incorrect nomenclature
 - 26% of chemical names in the literature are unacceptable and can't be converted into structures (GA Eller, *Molecules* 2006, 11, 915–928)
- ▶ **Chemical entity extraction false positives**
 - Chemical fragments, common words

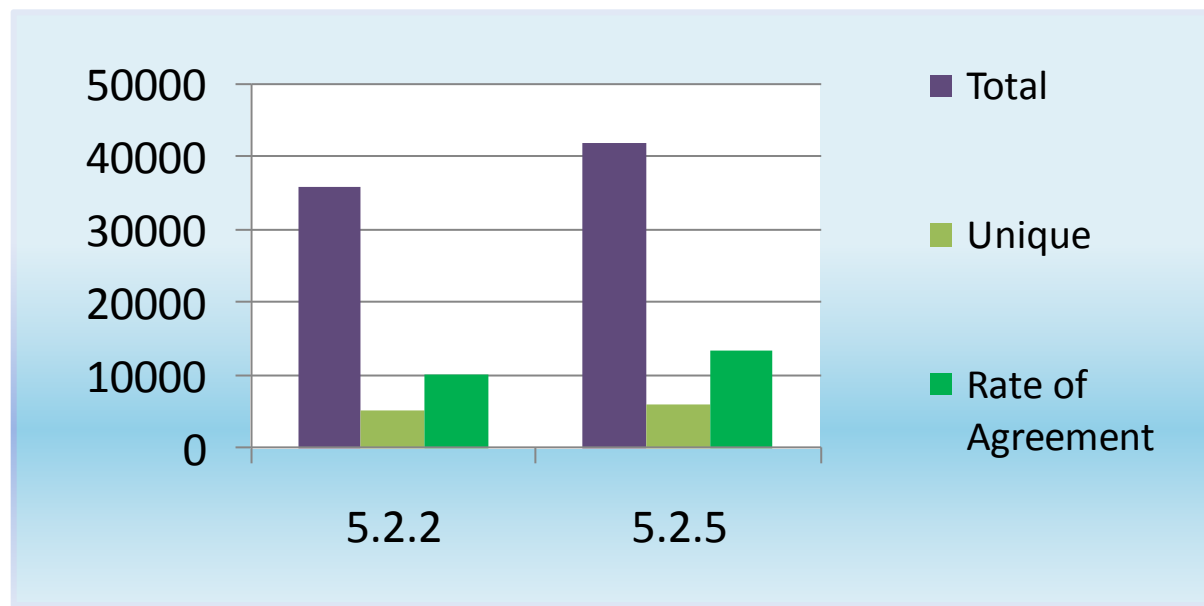
Name-2-Structure Corrections

- ▶ Basic OCR error correction heuristics (proprietary – pre process)
- ▶ Name fragmentation software (proprietary)
 - Improved heuristics dealing with erroneous white space in chemical names
- ▶ Advanced OCR error correction heuristics (proprietary – post process)
 - Automated spell correction (proprietary + 3rd party)
 - Allows for input of external name lists
 - Improved heuristics dealing with erroneous OCR characters in chemical names. 3rd party OCR chemical name correction software.

Recap on 2009 Benchmarking

Data Element	Output
No. Patents	900
Chemical Entities	101,061
Converted to Structures	59,582
Conversion Rate	58.9%
Standard SMILES	55,374

- Overall conversion: **+16.7%**
- Unique conversions: **+15.2%**
- 4-Tool Rate of Agreement: **+19.3%**



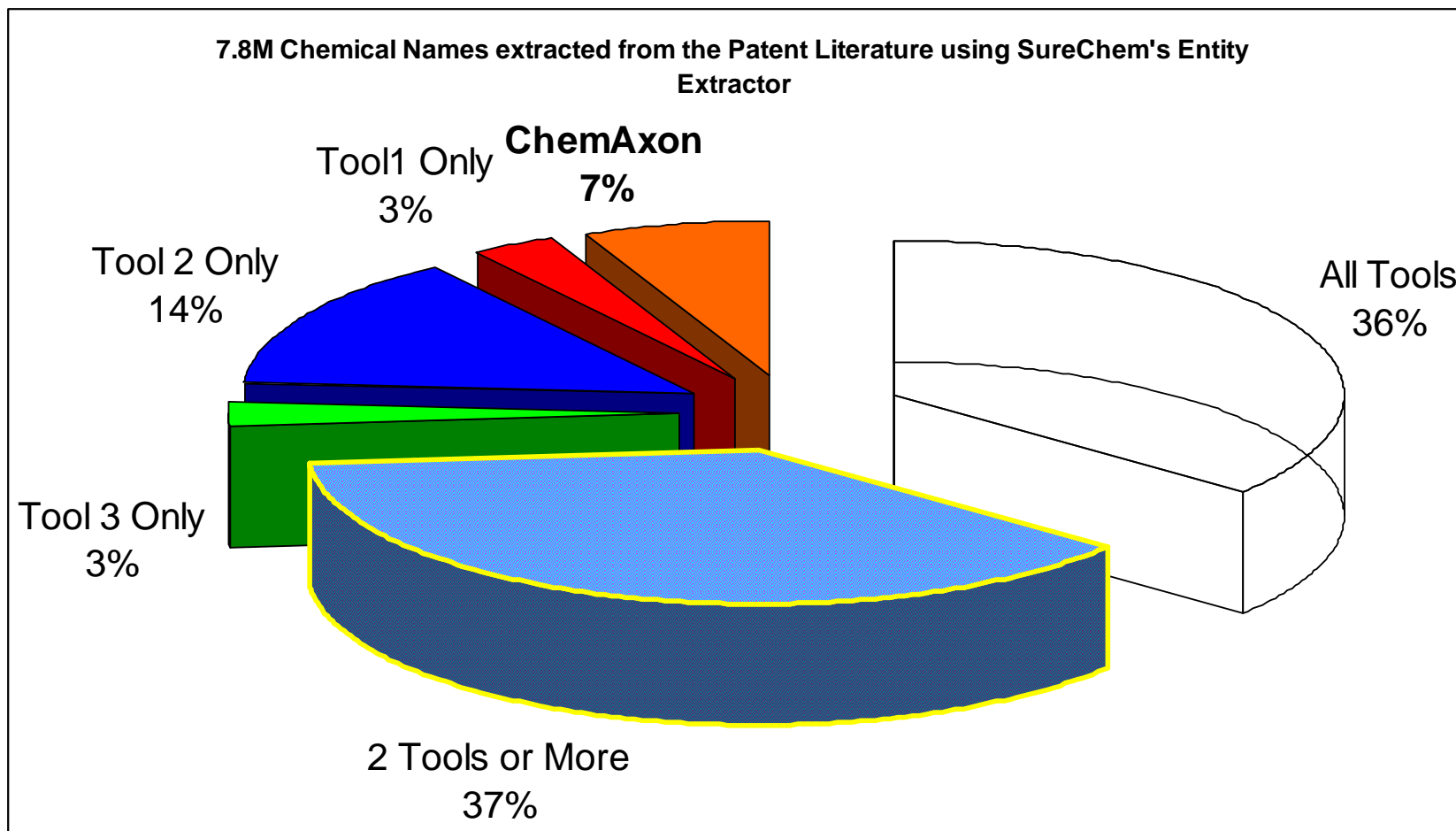
Benchmark Set 2011

Pharmaceutically relevant Chemical entities extracted from
20 Million Patents

Data Element	Output
No. of Patents	~20 M
Total Chemical Names	~34 M
Converted to Structures	~17 M
Conversion Rate	2/3
“Pharmaceutical” Structures	~10M
Pharma Structures -> Unique Names	~7.8 M
Unique Names -> Unique Structures	~6.8 M

Using four name-to-structure conversion tools yields an average additional 40% of conversions compared with a single tool alone

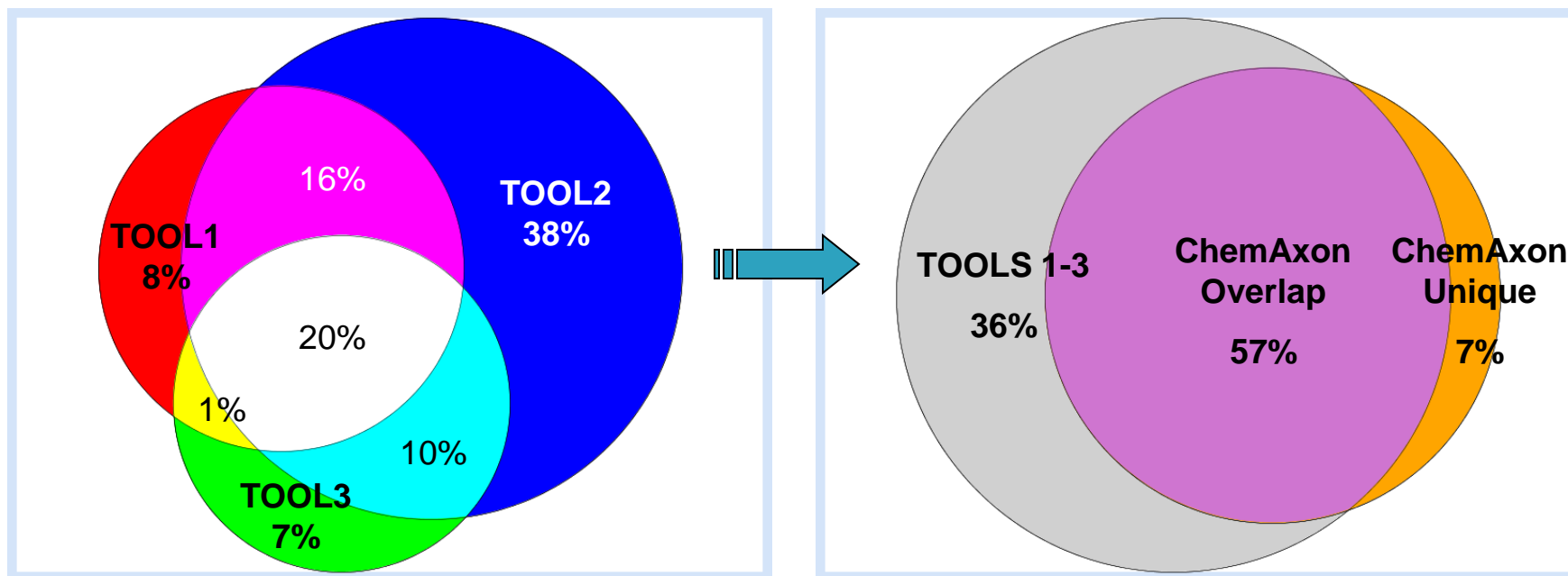
Coverage of 4 Leading N-2-S Conversion Tools



Coverage of 4 Leading N-2-S Conversion Tools

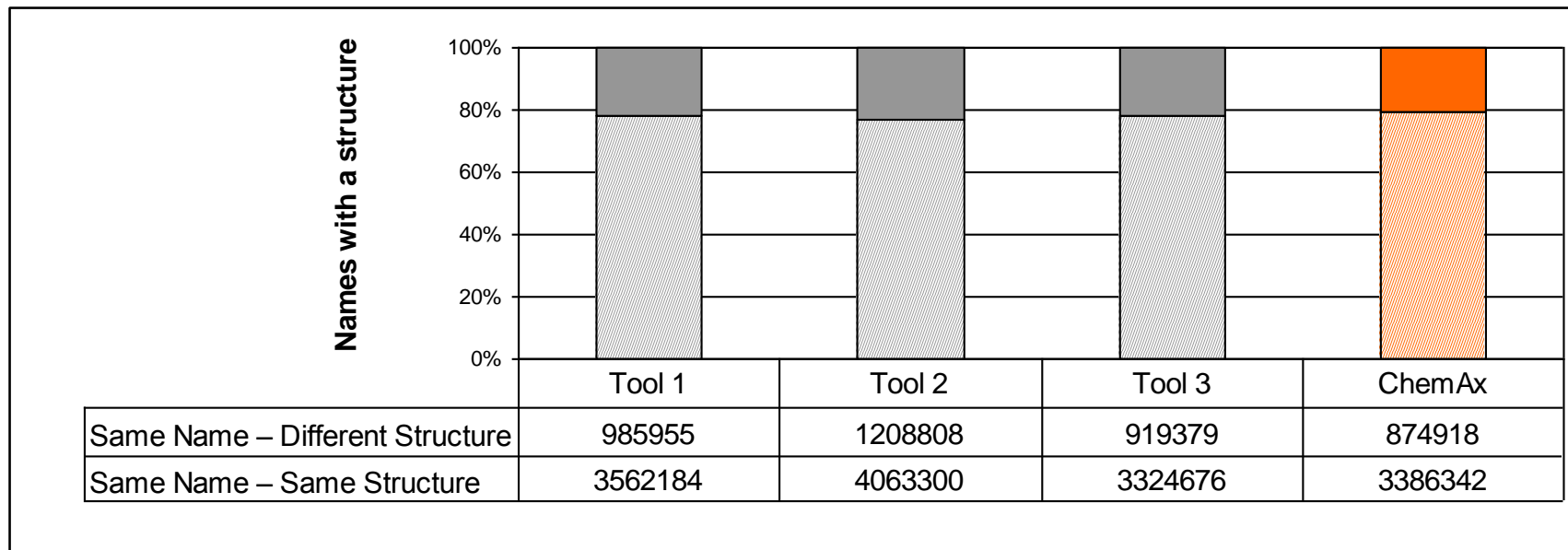
3 Tool Coverage ~4.8M names

ChemAxon converts 4.4M names



Using ChemAxon N-2-S conversion tools yields an additional **7%** of conversions or **0.5 M** structures

Rates of Structural Agreement



ChemAxon shows similar levels of structural variation as other N-2-S.

Rate of agreement suggest 1 in 5 chemical names found in a patent will have at least 2 or more different structures available to choose from!

Pairwise Comparison of Rates Structural Agreement

No. of Name with the Same Structure	Tool 1	Tool 2	Tool 3	ChemAxon
Tool 1				
Tool 2	76.3%			
Tool 3	68.7%	67.8%		
ChemAxon	57.2%	66.9%	48.5%	

Relating other N-2-S Tools to ChemAxon by measuring frequency of matching structures

ChemAxon shows closest structural agreement to Tool2, showing greatest divergence with Tool3

Progress from 2009

Prior to 2009

- ▶ Compounds w/ 4 tool agreement: 33%
 - (up 4%)
- ▶ Compounds w/ 3 tool agreement: 24%
 - (down 2%)
- ▶ Compounds w/ 2 tool agreement: 19%
 - (even)
- ▶ Compounds w/ 1 tool agreement: 24%
 - (down 2%)

Decreases likely due to higher 4-way agreement

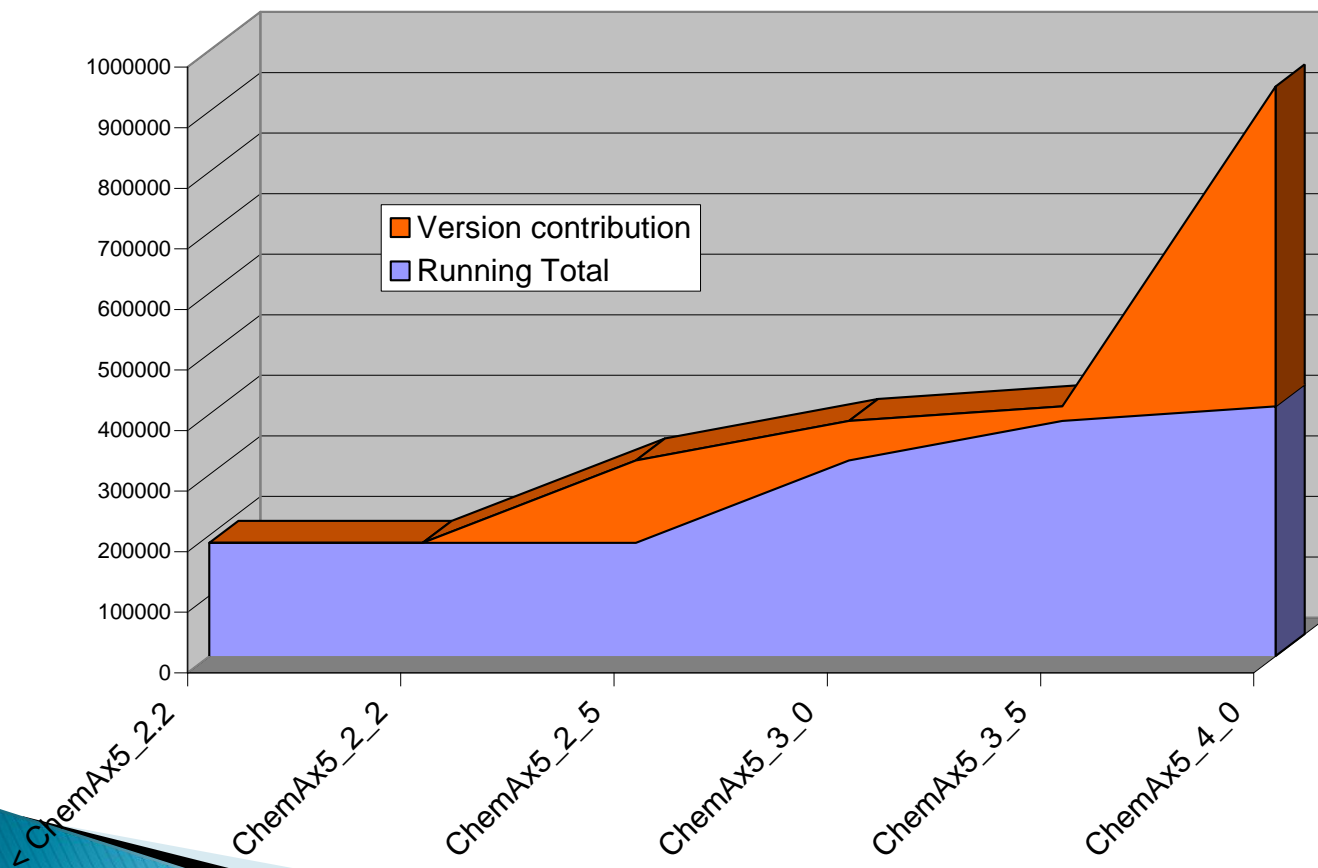
Today

- ▶ Compounds w/ 4 tool agreement: 36 %
 - (up 3%)
- ▶ Compounds w/ 3 tool agreement: 18 %
 - (down 6%)
- ▶ Compounds w/ 2 tool agreement: 20%
 - (up at 1%)
- ▶ Compounds w/ 1 tool agreement: 26%
 - (up 2%)

Some but not all Tools are converging in-terms of coverage

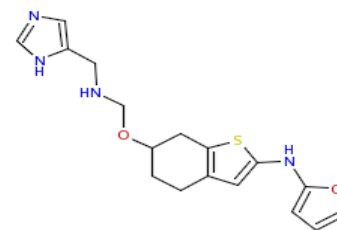
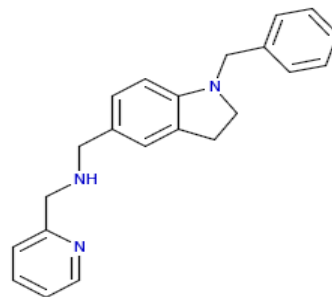
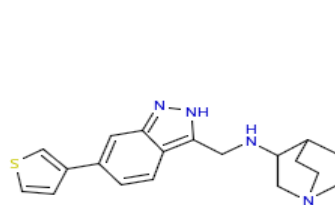
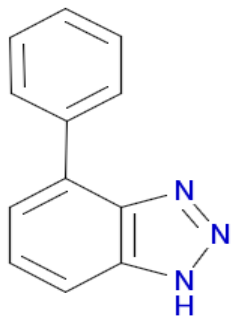
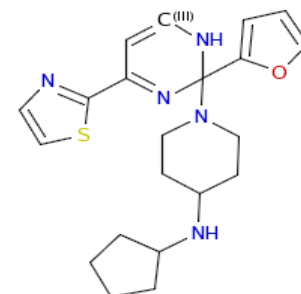
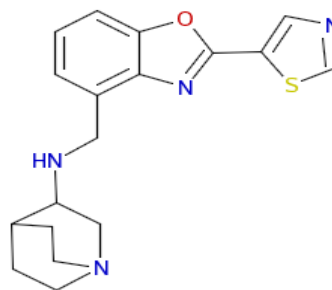
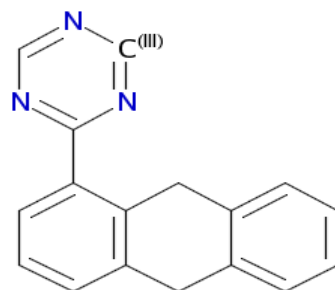
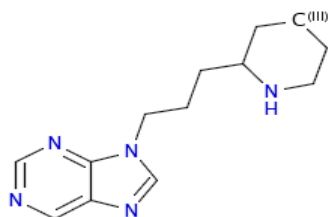
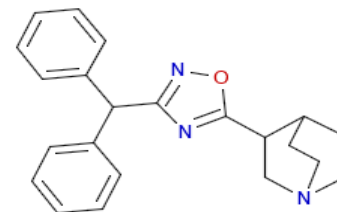
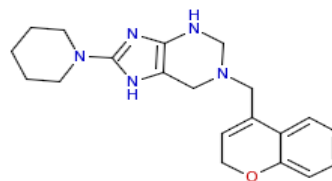
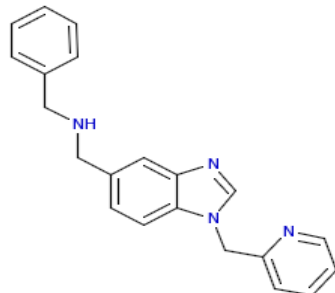
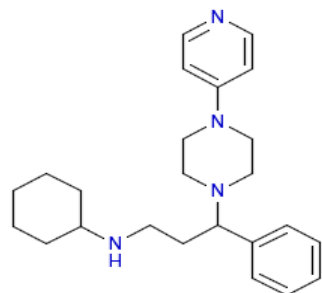
Contributions from Version Improvements in ChemAx N-2-S

Sample of ~1 M Chemical names only converted by ChemAxon



Significant improvement seen between versions 5.3.5 and 5.4.0!

Prominent Scaffolds New to ChemAx v5.4.0



Conclusion

- ▶ In a short time, ChemAxon has developed a N-2-S tool that is comparable to longstanding competitors
- ▶ Improving rapidly
- ▶ We find ChemAxon's tool easiest to use, with a good range of settings options
- ▶ Significant improvement seen between versions 5.3.5 and 5.4.0!
- ▶ ½M names (i.e. 7% of the sample corpus) are only recognised-converted by ChemAxon!
- ▶ With Four name-to-structure tools yields extra 40% conversion

Acknowledgements

- ▶ Daniel Bonniot de Ruisselet, ChemAxon
- ▶ Nicko Goncharoff, SureChem Digital Science
- ▶ Richard Koks, SureChem Digital Science
- ▶ James Siddle, SureChem Digital Science

- ▶ **SureChem EU Sales Manager**
Elisabeth Piveteau, e.piveteau@digital-science.com

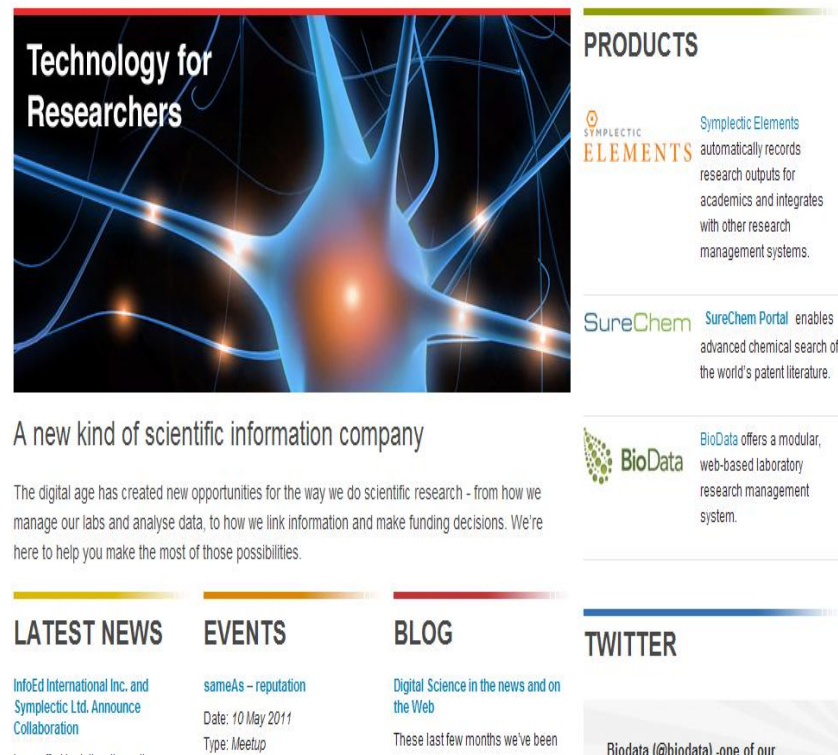
- ▶ **SureChem US Sales Manager**
Mark Moir, m.moir@digital-science.com

<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



Technology for Researchers

PRODUCTS

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.

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

InfoEd International Inc. and Symplectic Ltd. Announce Collaboration

EVENTS

sameAs - reputation
Date: 10 May 2011
Type: Meetup

BLOG

Digital Science in the news and on the Web
These last few months we've been

TWITTER

Biodata (@biodata) - one of our

