

Dirty (Data) Dancing

Benchmarking ChemAxon's name-structure tool on SureChem patent data

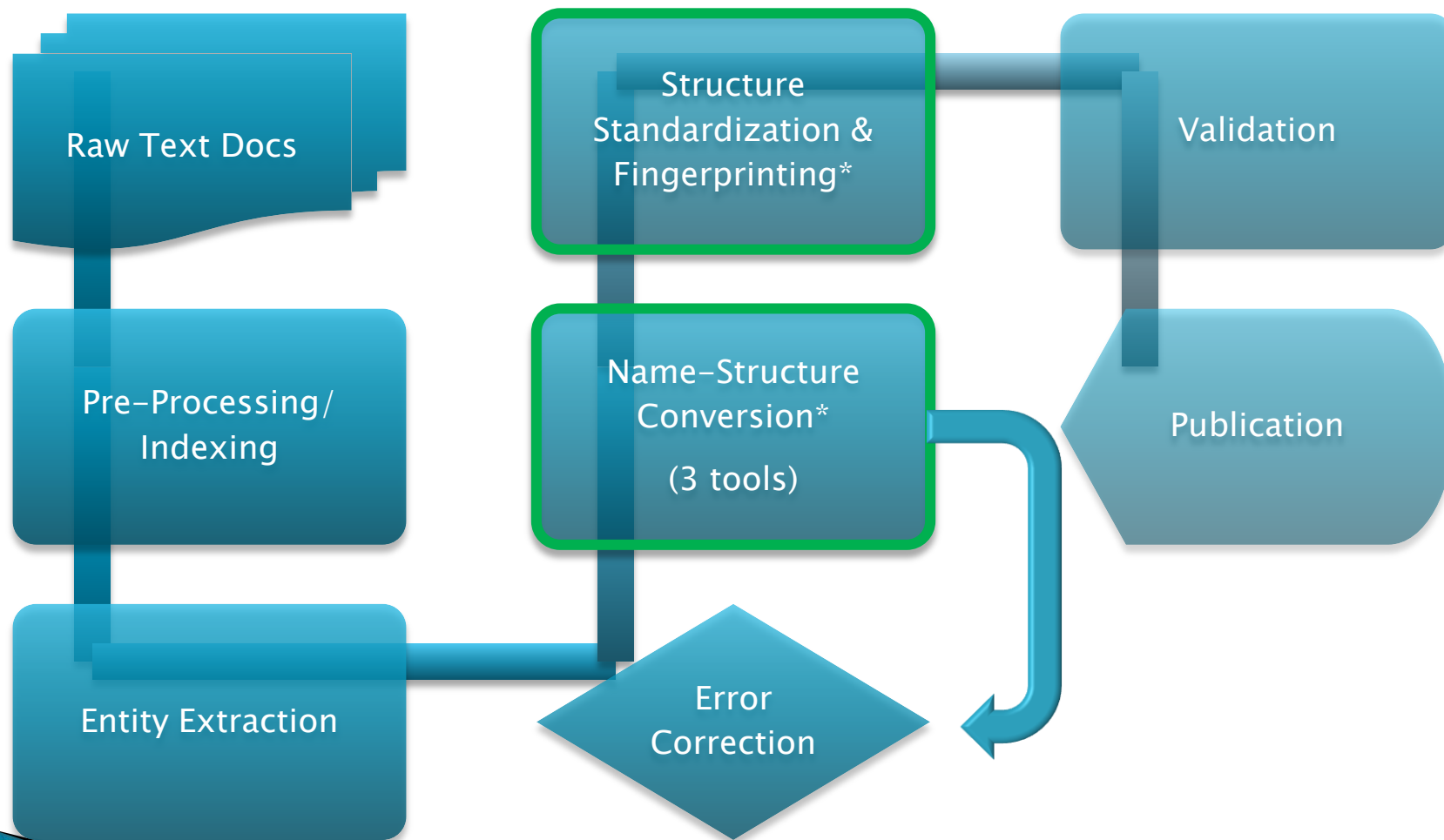
Agenda

- ▶ The Data Environment – SureChem in Brief
- ▶ Overall Numbers
- ▶ Rates of Agreement
- ▶ Unique ChemAxon contributions
- ▶ Precision Estimates
- ▶ Conclusions

SureChem At a Glance

- ▶ Compounds extracted from USPTO, EPO, WO full text, MEDLINE and Japan patent abstracts
- ▶ 11M unique structures
- ▶ 17.6M patent docs/18.5M MEDLINE abstracts
- ▶ 1M manually curated structures from images (USPTO)
- ▶ Available via Portal, Web Service or In-House Database

SureChem Workflow



*Includes use of 3rd party software
SureChem

Name-to-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

Benchmark Data Set

Chemical entities extracted from 900
pharmaceutically relevant patents

Data Element	Output
Chemical Entities	101,061
Converted to Structures	59,582
Conversion Rate	58.9%
Standard SMILES	55,374

Using four name-to-structure conversion tools yields a combined conversion rate of nearly 60% vs. around 40% for each tool on its own

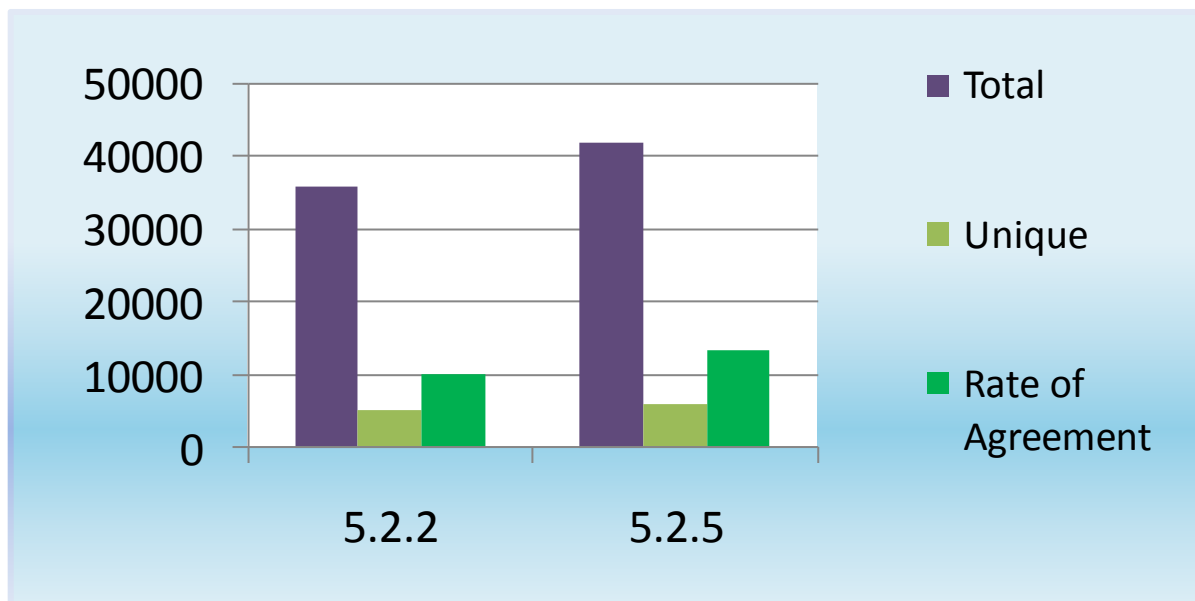
Overall and Pairwise Conversion Rates

ChemAxon v5.2.5

	Tool 1	Tool 2	Tool 3	ChemAx	Unique
Tool 1	41931	29943 (3737)	30781 (10428)	29674 (8609)	3541
Tool 2	29943 (3737)	39330	26844 (6780)	27450 (7513)	3617
Tool 3	30781 (10428)	26844 (6780)	35387	28173 (10619)	872
ChemAx	29674 (8609)	27450 (7513)	28173 (10619)	42156	6128

- **Bold** shows total conversion for each tool
- Pairwise comparison shows overlap & differing conversions ()
- “Unique” – number of names converted by just that tool.

Performance Improvements



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

Better Rates of Agreement

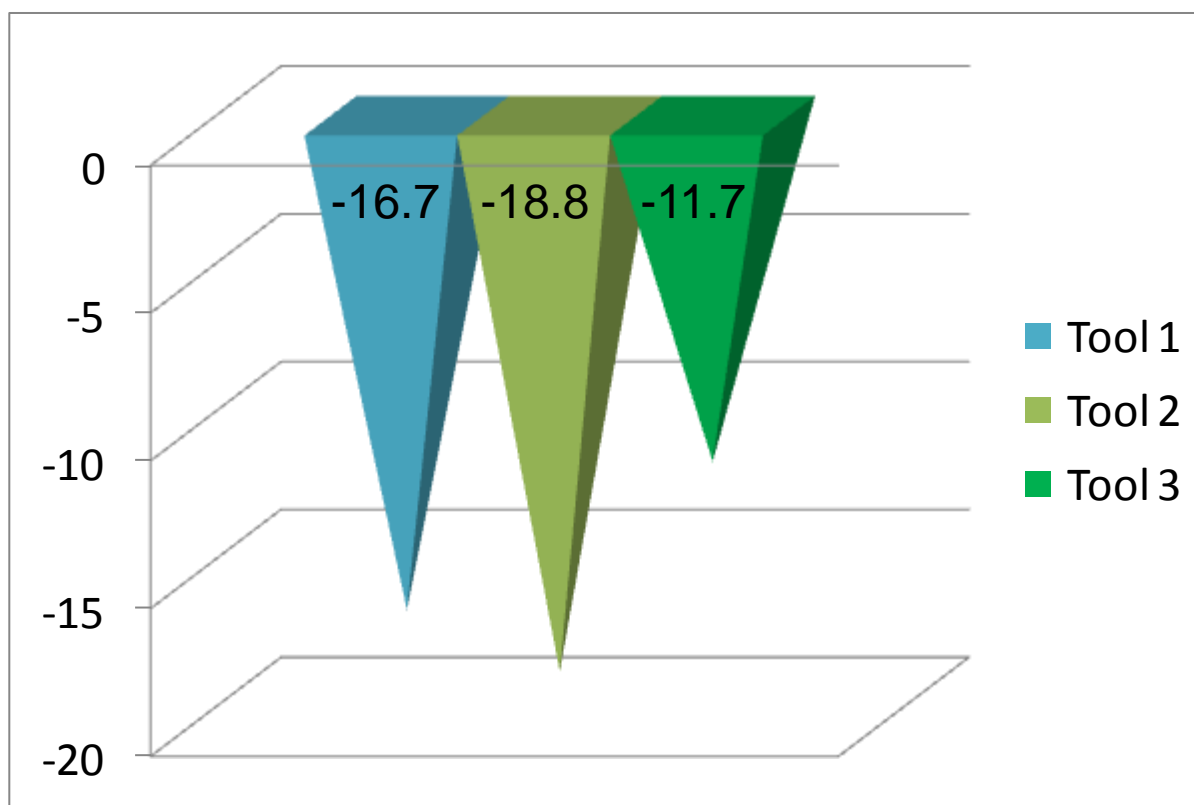
Percentage of Total Benchmark Set

- Compounds w/ 4 tool agreement: **33%**
(up from 29%)
- Compounds w/ 3 tool agreement: 24%
(down from 26%)
- Compounds w/ 2 tool agreement: 19%
(even at 19%)
- Compounds w/ 1 tool agreement: **24%**
(down from 26%)

Decreases likely due to higher 4-way agreement

Less differing structures per tool

Decrease in number of same names converted to different structures on a per-tool basis



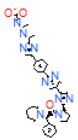
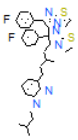
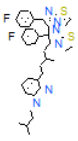
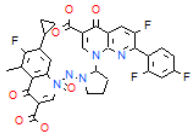
Differing Structures – Per Tool

	Tool 1	Tool 2	Tool 3	ChemAx	Unique
Tool 1	41931	29943 (3737)	30781 (10428)	29674 (8609)	3541
Tool 2	29943 (3737)	39330	26844 (6780)	27450 (7513)	3617
Tool 3	30781 (10428)	26844 (6780)	35387	28173 (10619)	872
ChemAx	29674 (8609)	27450 (7513)	28173 (10619)	42156	6128

- **Bold** shows total conversion for each tool
- Pairwise comparison shows overlap & differing conversions ()
- “Unique” – number of names converted by just that tool.

Unique ChemAxon Conversions

- ▶ 6128 unique structures, most of all tools
- ▶ That can mean more questionable structures
- ▶ **Many fragments**
- ▶ But also high value data
- ▶ **Estimated 1,800 structures appear to be whole exemplified compounds**

	chiral	1-(4-(4-(5-(2-((2S)-1-((2R)-2-phenyl-2-(1-piperidinyl)acetyl)-2-yl)H-imidazol-5-yl)-2-pyrimidinyl)phenyl)-H-imidazol-2-yl)ethyl)carbamate	687.8
	racemic	(4-[6-(4-Fluorophenyl)-6-[6-(4-Fluorophenyl)-imidazo[2,1-b]thiazol-5-yl]-imidazo[2,1-b]thiazol-5-yl]-1-N'2-(3-methylbutyl)-3-methyl)-1H-benzene-1,2-diamine	684.9
	racemic	4-[6-(4-Fluorophenyl)-6-[6-(4-Fluorophenyl)-imidazo[2,1-b]thiazol-5-yl]-imidazo[2,1-b]thiazol-5-yl]-1-N'2-(3-methylbutyl)-3-methyl)-1H-benzene-1,2-diamine	684.9
	racemic	N-1-(7-(1-cyclopropyl-6-fluoro-5-methyl-1,4-dihydro-4-oxo-1H-benzopyridin-3-carboxylic acid))-N-3-amino-(7-(1-(2,4-difluorophenyl)-6-fluoro-1,4-dihydro-4-oxo-1,8-naphthyridine-3-carboxylic acid))-pyrrolidine	679.6

Precision Estimates (5.2.2)

- ▶ Manual Review of 200 name/structures from ChemAxon uniquely generated structures
 - 70% unambiguously correct
 - 20% ambiguous or converted fragments
 - Might want to be a bit less lenient on these, but some users might want to retain fragments rather than have nothing
 - 10% are incorrect, due either to nature of names or how the tool handled them

Improved Rate of Agreement suggests improved precision for v5.2.5

Conclusions

- ▶ In a short time, ChemAxon has developed a tool that is comparable to longstanding competitors
- ▶ Improving rapidly
- ▶ We find ChemAxon's tool easiest to use, with a good range of settings options
- ▶ ChemAxon's n2s seems here to stay
- ▶ SureChem will be licensing it

Acknowledgements

- ▶ Richard Littin, SureChem/NetValue
- ▶ Daniel Bonniot de Ruisselet, ChemAxon
- ▶ Rita Vereb, ChemAxon
- ▶ Michael Faust, SureChem
- ▶ Gyorgy Pirok, ChemAxon
- ▶ Richard Koks, SureChem