STANDARDIZED REPRESENTATIONS OF ELN REACTIONS FOR CATEGORIZATION AND DUPLICATE/VARIATION IDENTIFICATION

ROGER SAYLE AND DANIEL LOWE

NEXTMOVE SOFTWARE, CAMBRIDGE, UK
Overview

- Electronic Lab Notebooks (ELNs) are a rich source of experimental observations of the synthetic methods used by the pharmaceutical industry, especially failed and low yield reactions.
- Recent successes in data mining this information have provided insights to improving the productivity and reducing costs in medicinal chemistry programs.
- This presentation describes the challenge of a small but vital aspect of this process: reaction pivoting.
The pharmaceutical industry is increasingly making use of reaction data warehouses from ELN data to better share and learn from the experience of in-house and CRO chemists.
The clear trend between Suzuki coupling success rate and predicted octanol-water partition co-efficient.

BIG DATA MINING CONFIRMATION OF NADINE-CHURCHER HYPOTHESIS

On 16,335 Suzuki coupling reactions extracted from US patent applications between 2001 and 2012.

<table>
<thead>
<tr>
<th>LogP</th>
<th>Mean Yield</th>
<th>N Obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1.0</td>
<td>52.89%</td>
<td>196</td>
</tr>
<tr>
<td>1.0 – 2.0</td>
<td>56.02%</td>
<td>1155</td>
</tr>
<tr>
<td>2.0 – 3.0</td>
<td>56.72%</td>
<td>2881</td>
</tr>
<tr>
<td>3.0 – 4.0</td>
<td>58.14%</td>
<td>4071</td>
</tr>
<tr>
<td>4.0 – 5.0</td>
<td>57.26%</td>
<td>3186</td>
</tr>
<tr>
<td>5.0 – 6.0</td>
<td>59.25%</td>
<td>2126</td>
</tr>
<tr>
<td>&gt; 6.0</td>
<td>63.83%</td>
<td>2720</td>
</tr>
</tbody>
</table>

CATEGORIZATION OF ELN REACTIONS

PROBLEM SUMMARY

- Categorizing reactions and observing trends requires transforming data from the representations used in electronic lab notebooks, which are primarily designed to reliably and efficiently capture intellectual property, into “cleaned up” forms more suitable for analysis and analytics.
- Often overlooked is the “pivoting” from document or experiment-based view of data to the de-duplicated reaction (and variation) view of the same data.
**SIMPLEST EXAMPLE OF DUPLICATION**

- In a “declassified” subset of 18237 reactions from AstraZeneca’s medchem ELN, there only 11451 unique experiments (63%), without any reaction normalization.

- The most duplicated reaction (41 times) is the Buchwald-Hartwig amination below.
REACTION VARIATIONS IN REAXYS

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>With sodium tetrahydroborate; ethanol</td>
<td>T=20°C;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zbinden, Katrin Groebeke; Anselm, Lilli; Banner, David W.; Benz, Joerg; Blasco, Francesca; Decorel, Guillaume; Himber, Jacques; Kuhn, Bernd; Panday, Narendra; Ricklin, Fabienne; Risch, Philippe; Schlatter, Daniel; Stahl, Martin; Thomi, Stefan; Unger, Robert; Haap, Wolfgang</td>
</tr>
</tbody>
</table>

| Location in patent | Page, Page column 46 |

247th ACS National Meeting, Dallas, TX, Monday 17th March 2014
WHEN IS A MILLION REALLY A MILLION?

Unleashing over a million reactions into the wild
Posted on February 27, 2014 by daniel

Unlike with small molecules, there are currently no large sets of publically available reaction data.

To remedy this situation, we have extracted over a million reactions from United States patent applications (2001-2013) and the same again from patent grants (1976-2013). This contrasts to the original data release of “only” 420 thousand (from 2008-2011 applications) whilst I was in the PMR group.

The reactions are available as reaction SMILES or CML from here, as 7zip archives. The CML representation includes quantities and yields where these were found. A documentation zip provides further information on the format of the data. This data is made available under CC-Zero i.e. without copyright.

It is hoped that making this data resource available will facilitate analyses that require a large number of reactions.

Although this data set contains 1,196,165 reaction instances, there are only 744,043 unique reactions, even without any structure or role normalization.

http://nextmovesoftware.com/blog
COMPLICATION #1: MOLECULE NORMALIZATION

• Some duplicates are caused by alternate chemistry representations requiring normalization of SMILES.
• This problems can be solved by Reaction InChIs.

• EN01585-15

• EN01995-47
Phosphodiesterase-4 inhibitors

**Step 2 (Scheme 3): (4-methoxyphenoxy)acetamide oxime**

A mixture of the (4-methoxyphenoxy)acetonitrile product (5.0 g, 31 mmol) from step 1, hydroxylamine hydrochloride (4.3 g, 62 mmol) and sodium acetate (5.1 g, 62 mmol) in MeOH (100 ml) was stirred at r.t. for 2 h. The resulting mixture was filtered on Celite®, concentrated, stirred in CHCl₃ for 18 h and filtered. The resulting solution was concentrated to yield (4-methoxyphenoxy)acetamide oxime as a gum.

**Preparation 168**

Synthesis of N-hydroxy-2-(4-methoxy-phenoxy)-acetamidine

**Add sodium acetate (5.1 g, 62 mmol) to 4-methoxyphenoxyacetonitrile (5.0 g, 31 mmol) and hydroxylamine hydrochloride (4.3 g, 62 mmol) in methanol (100 mL). Stir the resulting mixture at room temperature for 20 hours. Filter the resulting mixture through Celite, concentrate, stir in chloroform for 18 hours and filter. Concentrate the resulting solution to the title compound (5.1 g). LC-MS (m/e): 197 (M+1).**

Both InChI=1S/C9H12N2O3/c1-13-7-2-4-8(5-3-7)14-6-9(10)11-12/h2-5,12H,6H2,1H3,(H2,10,11)
COMPLICATION #2: SUPERATOMS

BrCH2CH2OH + \text{[Chemical structure]} \rightarrow \text{[Chemical structure]} \text{PPh3 } \rightarrow \text{[Chemical structure]}

becomes

\text{[Chemical structure]} \rightarrow \text{[Chemical structure]}
COMPLICATION #3: CORRECTING CHEMDRAW’S SUPERATOM EXPANSION

<table>
<thead>
<tr>
<th>Chemist drew</th>
<th>Chemist Intended</th>
<th>Chemist got</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMF</td>
<td><img src="image" alt="Chemist Intended DMF" /></td>
<td>^{2}H → M → F</td>
</tr>
<tr>
<td>K2CO3</td>
<td><img src="image" alt="Chemist Intended K2CO3" /></td>
<td>K (v0)</td>
</tr>
<tr>
<td>HBTU</td>
<td><img src="image" alt="Chemist Intended HBTU" /></td>
<td>^{3}H → B → U</td>
</tr>
<tr>
<td>mCPBA</td>
<td><img src="image" alt="Chemist Intended mCPBA" /></td>
<td></td>
</tr>
</tbody>
</table>

247th ACS National Meeting, Dallas, TX, Monday 17th March 2014
Complication #4: Reaction Role Normalization

- Some duplicates are caused by inconsistent reaction roles (reactants vs. agents) in the chemist’s sketch.

- EN00104-06

- EN00104-47

247th ACS National Meeting, Dallas, TX, Monday 17th March 2014
WHAT IS ATOM-MAPPING?

Mapping algorithm
ATOM-MAPPING FOR REACTION ROLE IDENTIFICATION/NORMALIZATION

- Atom mappings can distinguish reactants from solvents and catalysts, by whether they contribute atoms to the product(s).

but this requires reasonable atom mapping...

\[ \text{Reactants} \rightarrow \text{Products} \]
ATOM MAPPING + CLASSIFICATION

Percent of reactions with all product atoms mapped

- Atom mapping algorithms alone
- Combined with NameRXN

Result
Verified / Recognised by NameRXN (71%)

Marvin 6.0
ChemDraw 12
Consensus

247th ACS National Meeting, Dallas, TX, Monday 17th March 2014
COMPPLICATION #5: ELN SECTIONS

Would the full ELN reaction please stand up?

247th ACS National Meeting, Dallas, TX, Monday 17th March 2014
Complication #6: Agents Matter


Sandmeyer reaction
Benzenediazonium chloride heated with cuprous chloride dissolved in HCl to yield chlorobenzene.

\[ \text{C}_6\text{H}_5\text{N}_2^+ + \text{CuCl} \rightarrow \text{C}_6\text{H}_5\text{Cl} + \text{N}_2 + \text{Cu}^+ \]

Gatterman reaction
Benzenediazonium chloride is warmed with copper powder and HCl to yield chlorobenzene.

\[ \text{C}_6\text{H}_5\text{N}_2^+ + \text{CuCl} \rightarrow \text{C}_6\text{H}_5\text{Cl} + \text{N}_2 + \text{Cu}^+ \]
AZ EXAMPLES WITH SAME PARENT

- **EN01325-20**

- **EN01325-22**

- **EN01325-25**

- **EN01325-27**
EXAMPLES WITH SAME GRANDPARENT

• EN00930-16

• EN00930-25

• EN00930-60
RESULTS & LESSONS LEARNED

• Exact duplicates are extremely common.
  – 62.2% of USPTO data set is unique (744043/1196165).
  – Some large pharmaceutical ELNs < 50% unique.

• True reaction variations are relatively rare.
  – 684688 unique products without normalization.
  – 675026 unique tautomer/protomer products (56.4%).
  – Over 90% of deduplicated reactions have unique products.

• For many use cases the reaction InChI or the product InChI may be sufficient for use as a key to join on.
CONCLUSIONS

• Transforming the ELN content into analysis friendly data sets is perhaps more complicated than it might first appear.

• This presentation describes the challenge of “reaction pivoting” and several technical solutions devised by the authors to address this problem.

• The resulting de-duplicated and categorized reactions are useful in current work investigating the influence of reaction conditions on yield, and step choice and ordering in multi-step syntheses.
ACKNOWLEDGEMENTS

• Nick Tomkinson, AstraZeneca R&D, Alderley Park, UK.
• Thierry Kogej, Astra Zeneca R&D, Mölndal, Sweden.
• Mick Kappler, Hoffmann-La Roche, Nutley, NJ, USA.
• Plamen Petrov, AstraZeneca R&D, Mölndal, Sweden.
• Colin Batchelor, Royal Society of Chemistry, UK.

• Thank you for you time. Questions?