Extraction, Analysis, Atom Mapping, Classification and Naming of Reactions from Pharmaceutical ELNs

Roger Sayle, Daniel Lowe, Noel O'Boyle
NextMove Software, Cambridge, UK
Michael Kappler, *Hoffmann-La Roche, Nutley, NJ, USA*
Nick Tomkinson, *AstraZeneca, Alderley Park, UK*

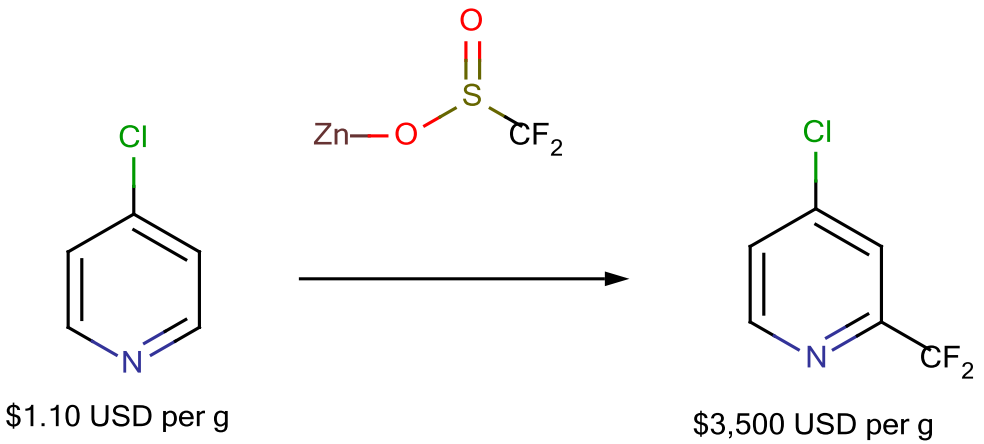
6th Joint Sheffield Conference on Chemoinformatics, Tuesday 23rd July 2013

INTRODUCTION

- Pharmaceutical ELNs contain a wealth of synthetic chemistry knowledge, particularly on failed reactions, often not described in the literature or available from public sources.
- This presentation describes several of the technical informatics challenges encountered during the process of exploiting reaction information from ELNs.

MOTIVATION #1: ECONOMICS

- The primary motivation for reaction informatics is reducing cost of goods, through higher yields, fewer failed reactions and more direct synthetic routes.



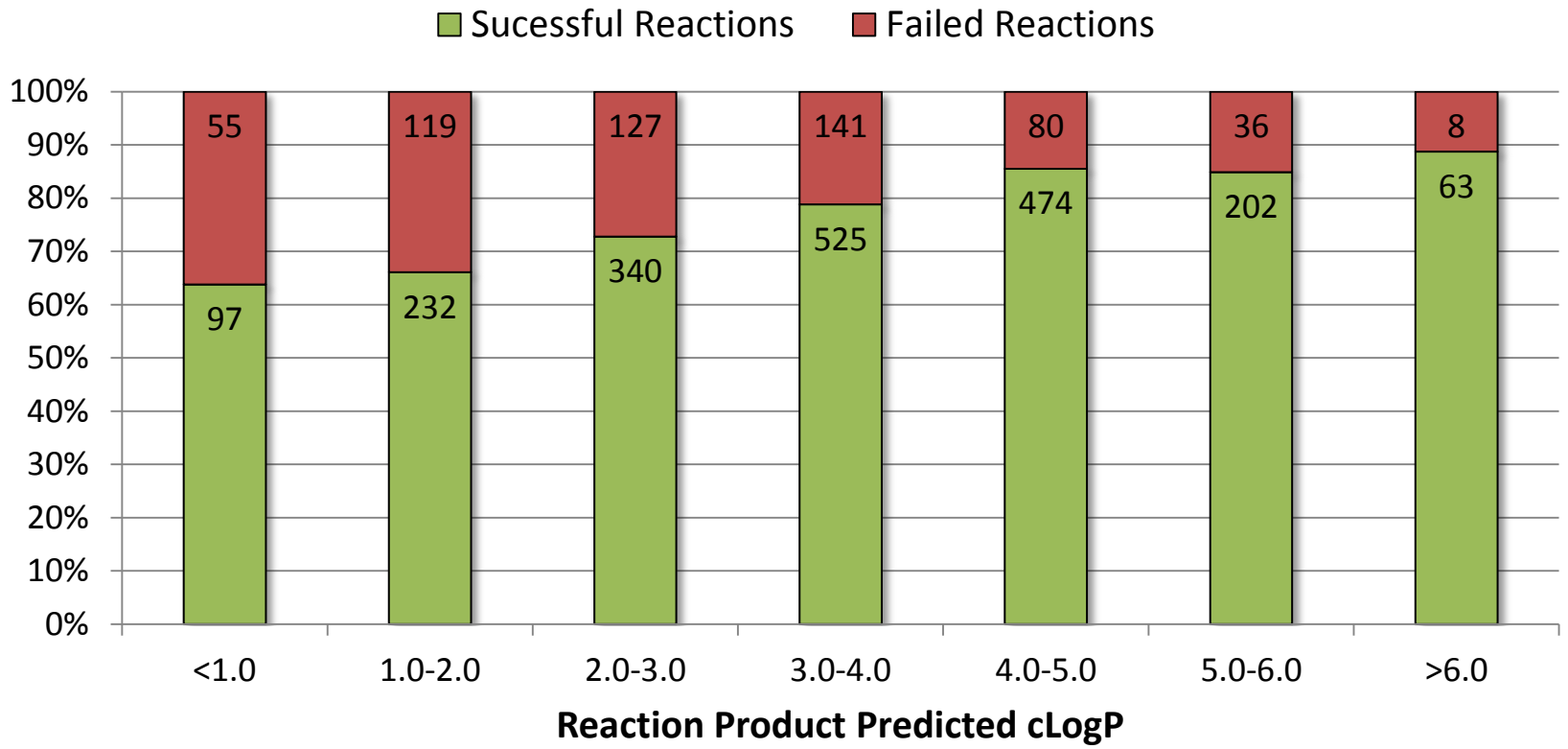
Yuta Fujiwara et al., "Practical and innate carbon-hydrogen functionalization of heterocycles", *Nature* Vol. **492**, pp. 95-99, 6th December 2012.

MOTIVATION #2: YIELD PREDICTION

- Nadine *et al.*[1] hypothesize that low LogP is a major cause of synthesis failure in parallel synthesis of combinatorial libraries.
- GSK data set from Pickett *et al.*[2] of 2500 Suzuki couplings in a 50x50 library of MMP-12 inhibitors.
  - 1704 compounds measured, mean logP = 3.56 (1.44)
  - 566 compounds not made, mean logP = 2.83 (1.52)
  - Student's t-test for different distributions, $p<2x10^{-22}$.

1. Nadine, Hattotuwagama and Churcher ,"Lead-Oriented Synthesis: A New Opportunity for Synthetic Chemistry", *Angew. Chem. Int. Ed*, 51:1114 2012.
2. Pickett et al., *ACS Med. Chem. Lett.* 2(1):28, 2011

MOTIVATION #2: YIELD PREDICTION



The clear trend between Suzuki coupling success rate and predicted octanol-water partition co-efficient.

MOTIVATION #3: VIRTUAL LIBRARIES

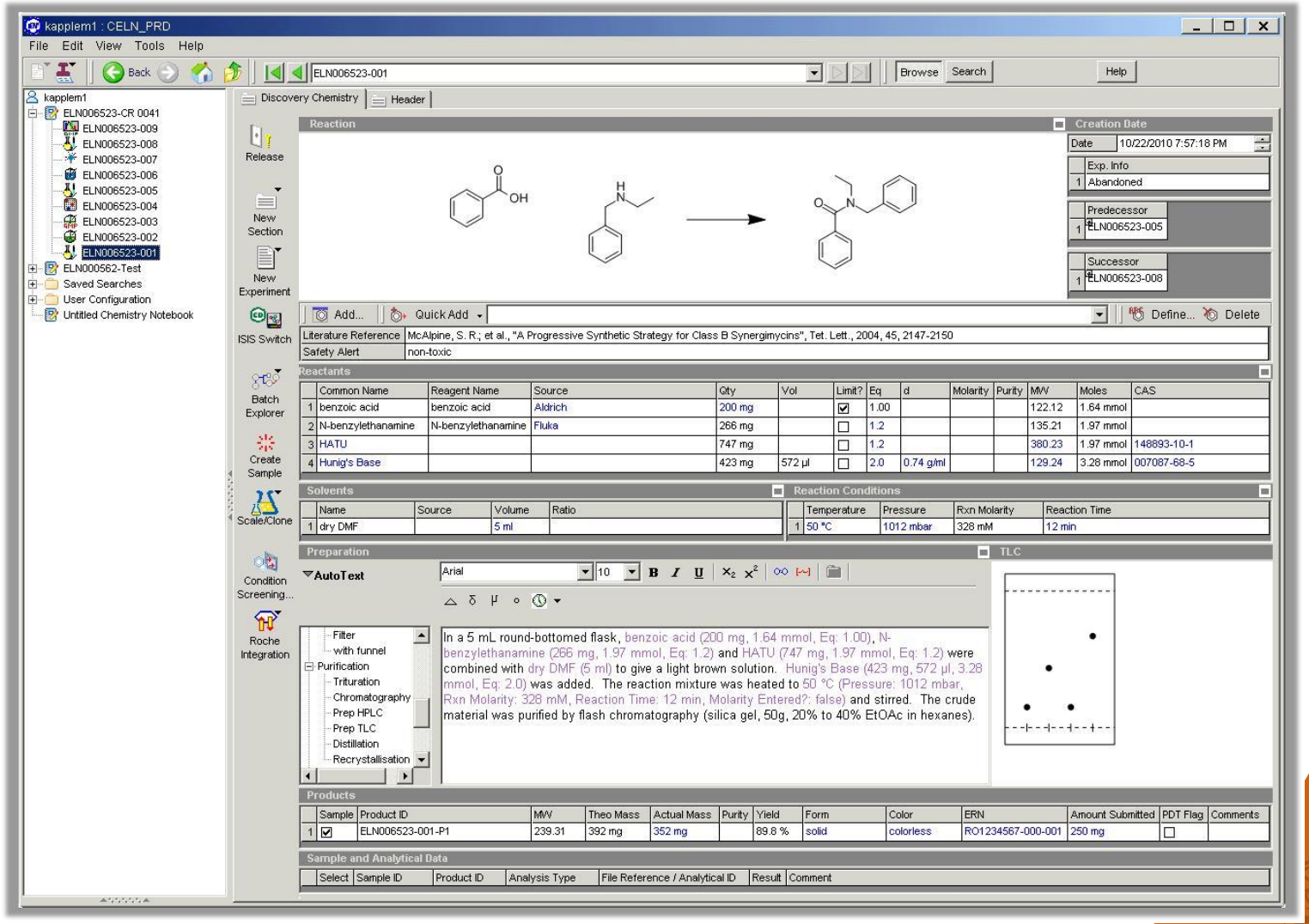
- Statistics of ELN reactions may be used to redefine RECAP-style rules for retro synthesis [1,2]
- For example, reduction of nitro groups to amines is one of the top 5 most common ELN reactions, but nitration reactions are amongst the rarest.
- This implies that nitro containing compounds are purchased as reagents rather made in-house.

1. Xiao Qing Lewell, Duncan B. Judd, Stephen P. Watson and Michael M. Hann, "RECAP – Retrosynthetic Combinatorial Analysis Procedure", *JCICS* 38(3):511-522, 1998.
2. Vainio, Kogej and Raubacher, "Automated Recycling of Chemistry for Virtual Screening and Library Design", *J. Chem. Inf. Model*. 52(7):1777-1786, 2012.

THE CHALLENGES

1. Export of the data from the ELN.
2. High fidelity conversion to other file formats.
3. Reaction normalization/standardization.
4. Reaction identity (canonicalization).
5. Reaction naming and classification.

TYPICAL ELN EXPERIMENT



DATABASE EXPORT

- Typically, ELNs are implemented as complex schemas within relational databases (Oracle), supporting transactions, auditing and security privileges.
- Not uncommonly the vendor provided functionality or APIs for data export are slow and/or buggy.
- In addition to reactions and structures, there is often a requirement to export all associated data, including textual and numeric data, tables, even LCMS and NMR spectra.

TYPICAL ELN FIELDS IN RD FORMAT

```
$DTYPE REACTION:REACTION.CONDITIONS:TEMPERATURE:VALUE
$DATUM 120 &#xB0;C
$DTYPE REACTION:REACTION.CONDITIONS:TEMPERATURE:MINVAL
$DATUM  393.15
$DTYPE REACTION:REACTION.CONDITIONS:TEMPERATURE:MAXVAL
$DATUM  393.15
$DTYPE REACTION:REACTION.CONDITIONS:PRESSURE:VALUE
$DATUM 5 bar
$DTYPE REACTION:REACTANTS(1):NAME
$DATUM nicotinoyl chloride
$DTYPE REACTION:REACTANTS(1):CHEMICAL.STRUCTURE
$DATUM c1cc(cnc1)C(=O)Cl
$DTYPE REACTION:REACTANTS(1):FORMULA.MASS:VALUE
$DATUM 141.56
$DTYPE REACTION:REACTANTS(2):NAME
$DATUM (2R,3S)-3-methylpentan-2-ol
$DTYPE REACTION:REACTANTS(2):CHEMICAL.STRUCTURE
$DATUM CC[C@@H](C)[C@H](C)O
$DTYPE REACTION:REACTANTS(2):FORMULA.MASS:VALUE
$DATUM 102.17
```

```
$DTYPE REACTION:PRODUCTS(1):CHEMICAL.STRUCTURE
$DATUM CC[C@@H](C)[C@H](C)OC(=O)c1cccnc1
$DTYPE REACTION:PRODUCTS(1):NAME
$DATUM (2R,3S)-3-methylpentan-2-yl nicotinate
$DTYPE REACTION:PRODUCTS(1):MOLECULAR.WEIGHT:VALUE
$DATUM 207.27
$DTYPE REACTION:PRODUCTS(1):MOLECULAR.FORMULA
$DATUM C12H17NO2
$DTYPE REACTION:PRODUCTS(1):ACTUAL.MOLES:VALUE
$DATUM 2.16 mmol
$DTYPE REACTION:SOLVENTS(1):NAME
$DATUM 1-pentanol
$DTYPE REACTION:SOLVENTS(1):VOLUME:VALUE
$DATUM 5 mL
$DTYPE REACTION:SOLVENTS(1):R.VALUES
$DATUM R10,R20
```

FILE FORMAT CONVERSION

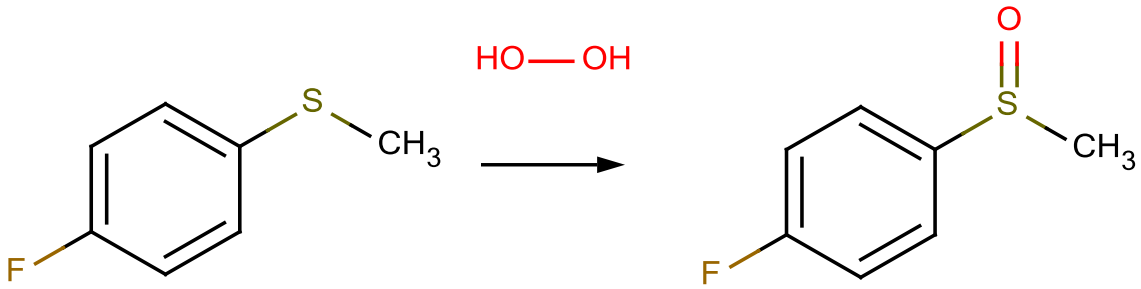
- The source format for many reactions is typically a sketch, in either CDX, CDXML, ISIS Sketch or Marvin file format.
- For data processing reactions are much easier to handle as reaction SMILES, MDL RXN or RD files and possibly even variants of MOL and SD file formats.
- Alas handling of reaction file formats is generally poorly handled by many cheminformatics tools.
- Additionally, reaction file formats can rarely encode all of the same information.

DECRYPTING CDX & CDXML FILES

- CambridgeSoft are to be congratulated for publicly documenting their CDX and CDXML file formats.
- Unfortunately this online ChemDraw developer resource is no longer being kept up to date.
- New tags: object 0x802b encodes “annotation”.
- Mistakes: “arrow” is encoded by object 0x8021.
- Proprietary property tags: USPTO’s “PageDefinition”.
- Support for reading and writing isotopic information in CDXML files has been contributed to Open Babel.

AGENTS AND CATALYSTS

How to express reaction role, i.e. molecules drawn above and below the reaction arrow.




```
CSc1ccccc1(F)>OO>CS(=O)c1ccc(cc1)F
```

Although standard MDL RXN files only capture reactants and products, a useful ChemAxon extension is to add a third count for agents. Downstream tools can optionally remove or reclassify these agents to produce strictly compliant MDL files.

VALENCE ISSUES

A tricky problem is working around the different MDL valence interpretations used by cheminformatics tools.

For example, it is not uncommon for the alchemical reaction **`[Pb]>>[Au]`** to become reinterpreted as **`[PbH2]>>[Au]`** after writing and re-reading from MDL formats. Such errors lose the distinction between sodium metal and sodium hydride, resulting in incorrect molecular formulae, or to distinguish metals from radicals, causing problems for substructure searching.

RICH TEXT FORMAT

- Exporting formatted text from electronic lab note books, such as experimental/preparation write-ups, requires converting Microsoft Rich Text Format (RTF).
- This can be translated into HTML or ASCII, and then long lines wrapped for inclusion in SD/RD data fields.
- Special code is required for handling non-U.S. character sets, and whilst Western European, Russian, Chinese and Japanese were expected, finding Arabic, Thai and Vietnamese text in major pharmaceutical ELNs came as a surprise.

SALT/COMPONENT GROUPING

- Keeping track of the intended number and formulae of reactants, products and agents in a reaction, requires preserving salt form associations.
- This is implemented by honoring the “group” information from the sketch as single disconnected components in MDL RXN and RD file output.
- These associations are traditionally lost in SMILES...

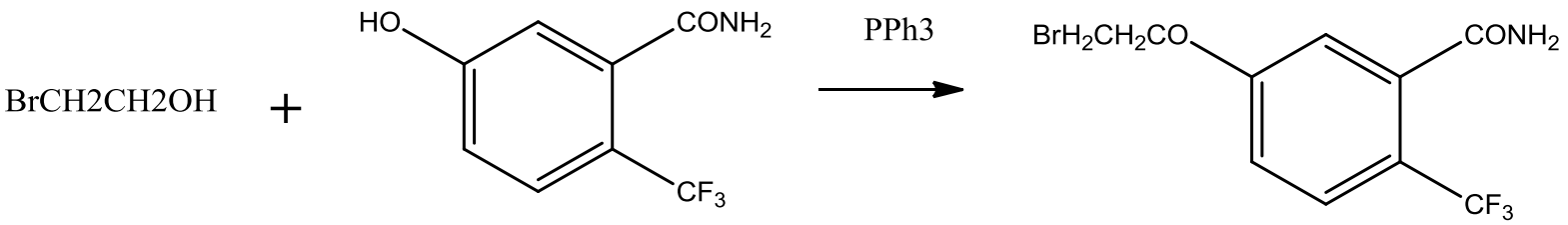
  …>CC(=O)[O-].[Cl-].[Cl-].[Fe].[K+].[Pd+2]…>…

  but can be retained via ChemAxon/GGA extensions.

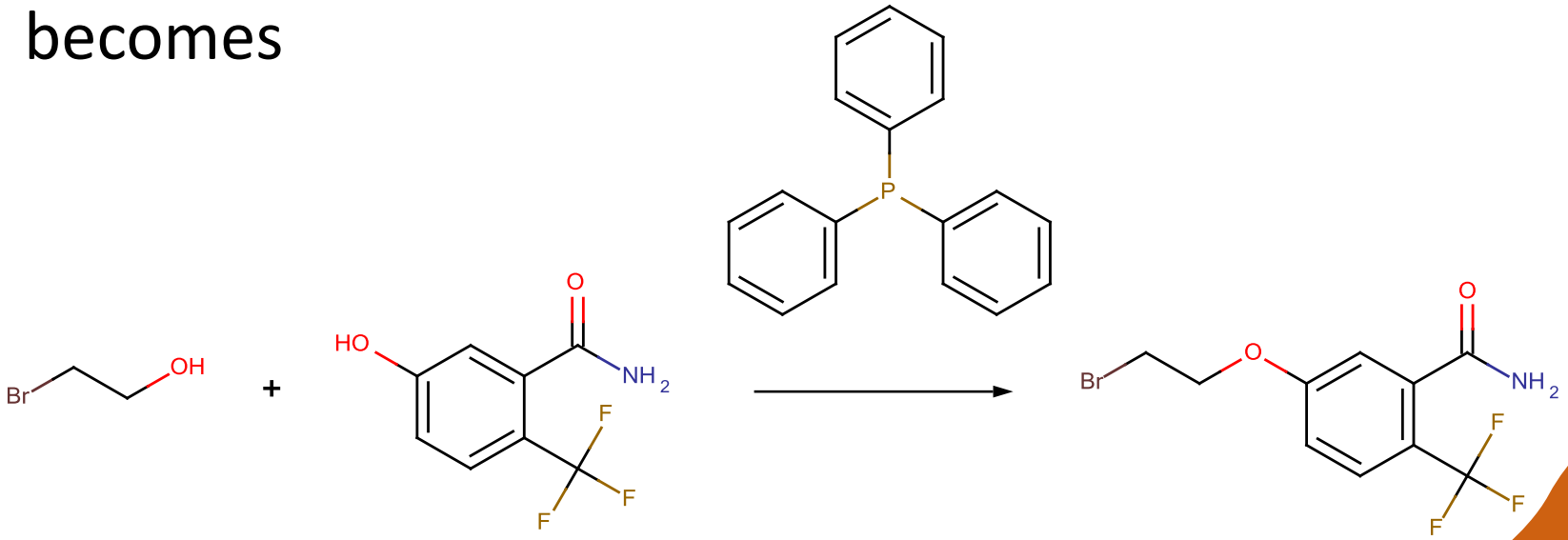
THREE TYPES OF SUPERATOM/LABEL

1. Recognized superatoms are expanded to their explicit all atom representations.
2. Unrecognized superatoms/labels that are bonded to a molecule are encoded as dummy asterisk atoms where the text is preserved as an MDL atom alias.
   - Support for writing MDL aliases contributed to RDKit.
3. Disconnected unrecognized labels are preserved as supplementary data fields in SD and RD file formats.

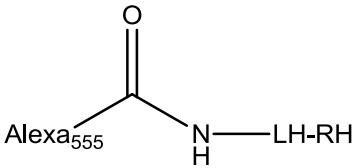
SUPERATOM EXPANSION



becomes



TEXT LABEL PRESERVATION

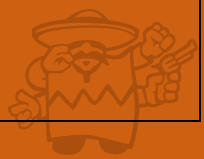


NMS-123456

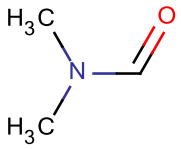
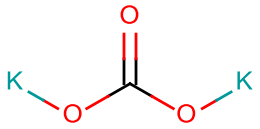
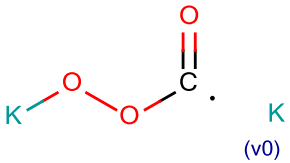
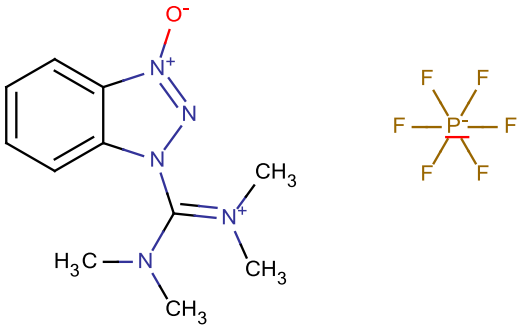
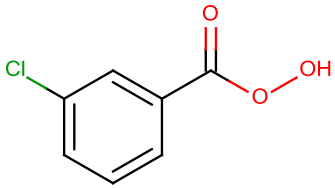
is converted to an MDL SD file as

```
NextMove07171309292D

  5  4  0     0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
    1.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000   -0.8660    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000    0.8660    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.5000    0.8660    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
  2  4  1  0  0  0  0
  4  5  1  0  0  0  0
A    1
Alexa555
A    5
LH-RH
M  END
> <LABELS>
NMS-123456
$$$$
```

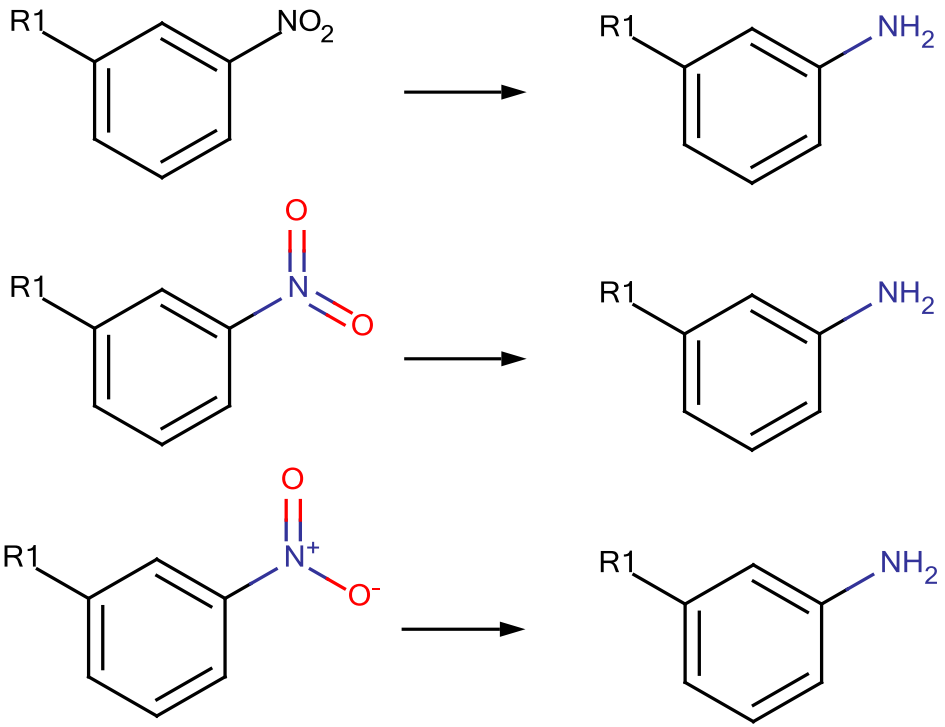


CHEMDRAW SUPERATOM CORRECTION

Chemist drew	Chemist Intended	Chemist got
DMF		
K2CO3		
HBTU		
mCPBA		

REACTION STANDARDIZATION

Chemists may draw the reaction components in multiple ways (e.g. tautomers and protonation states)



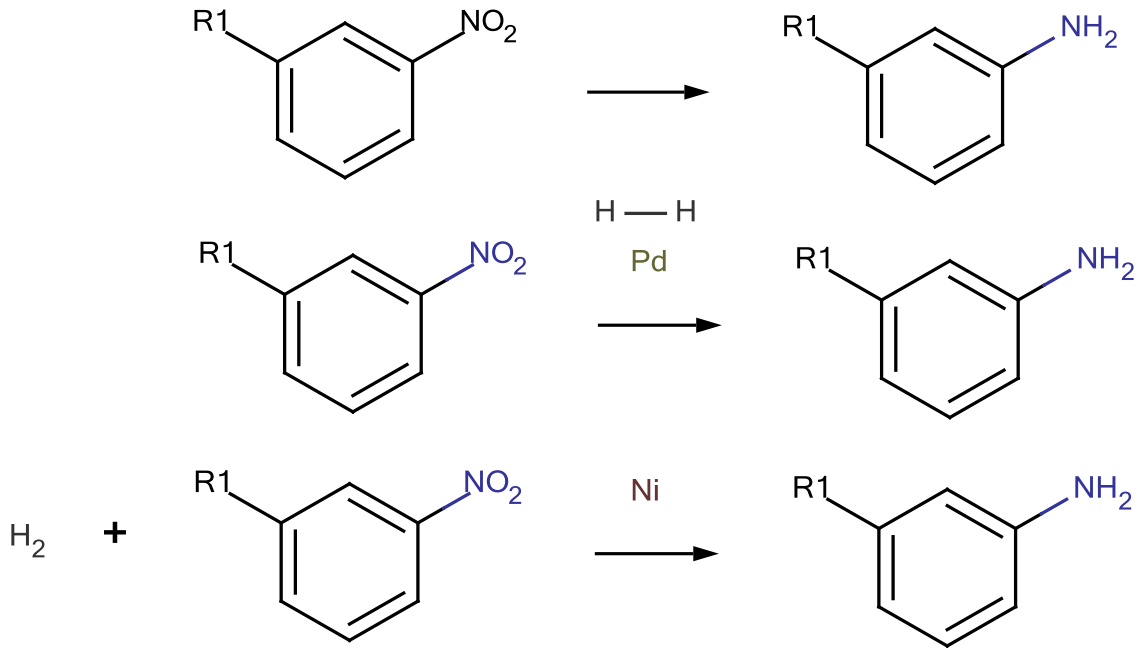
REACTION STANDARDIZATION

- An often overlooked aspect of ELNs is the need to enforce “business rules” to consistently represent a reaction, in the same way that normalized molecules are stored in registration systems.
- Pharmaceutical ELNs contain structures where nitros are represented arbitrarily, and even cases where azide representations differ on each side of an arrow.
- Unfortunately, the rules used for molecules (such as InChI) may be inappropriate for reactions, where metal co-ordination and radicals play a major role.

REACTION IDENTITY

ELNs frequently contain repeated reactions, duplicates.

We need define when two reactions are the same.



REACTION IDENTITY

- When translating from the experiment-centric view of an ELN to the reaction-centric view of a reaction database one asks when are two reactions the same.
- A pragmatic/operational definition might be that two experiments with identical sets of reactants and products, but differing quantities, conditions, catalysts and solvents are *variation*s of each other.
- Whether a component is a reactant, catalyst, solvent or reagent may be consistently defined by atom-mapping; reactants contribute atoms to the product.

REACTION CLASSIFICATION

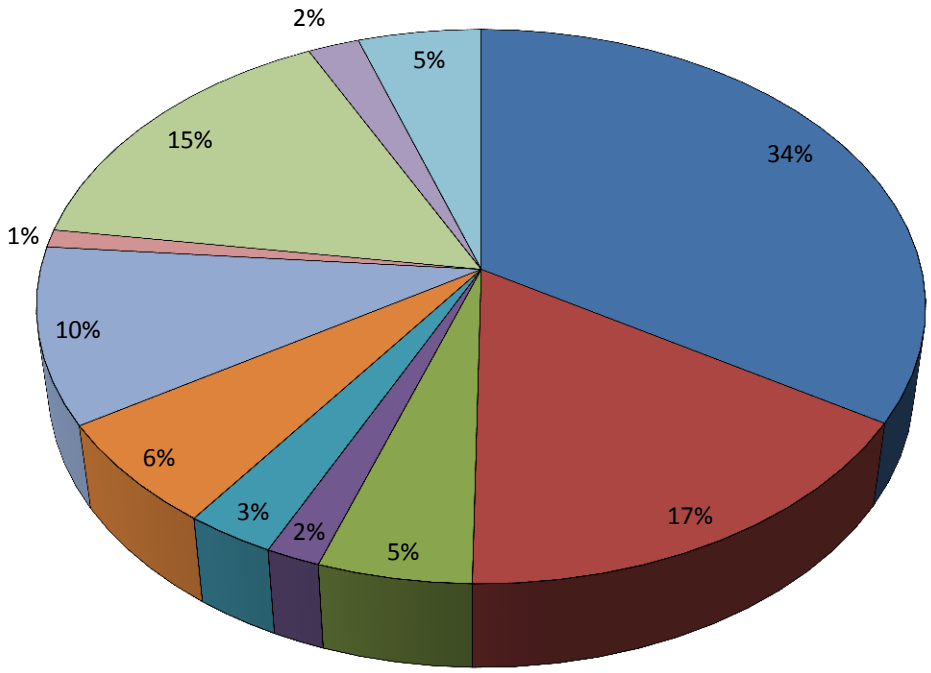
- For searching and analysis it is often convenient to algorithmically assign each reaction to a type, often a named reaction such as Negishi coupling, Diels-Alder cycloaddition, nitro reduction or chiral separation.
- This is implemented using a database of SMIRKS-like transforms that may be pre-compiled for efficient matching and portability across informatics toolkits.
- This approach provides both classification under the RSC's RXNO ontology and reaction atom mapping for component role assignment.

SIMPLE CLASSIFICATIONS

- Examples of simple reaction classifications

`A.B>>C`	Regular reaction
`A.B>>`	Failed reaction
`>>C`	Compound purchase
`A>>A`	Purification
`A.B>>A`	Separation (and chiral separation)
`A.B>>C.A`	Catalysts or unreacted reagents

CATEGORIZATION OF ELN REACTIONS



- Heteroatom alkylation and arylation
- Acylation and related processes
- C-C bond formations
- Heterocycle formation
- Protections
- Deprotections
- Reductions
- Oxidations
- Functional group conversion
- Functional group addition
- Resolution

1. J. Carey, D. Laffan, C. Thomson, M. Williams, *Org. Biomol. Chem.* 2337, 2006.
2. S. Roughley and A. Jordan, *J. Med. Chem.* 54:3451-3479, 2011.

REACTION ONTOLOGY

- Reactions are classified into a common subset of the Carey et al. classes and the RSC's RXNO ontology.
- There are 12 super-classes
  - e.g. 3 C-C bond formation (RXNO:0000002).
- These contain 84 class/categories.
  - e.g. 3.5 Pd-catalyzed C-C bond formation (RXNO:0000316)
- These contain ~300 named reactions/types.
  - e.g. 3.5.3 Negishi coupling (RXNO:0000088)
- These require >400 SMIRKS-like transformations.

EXAMPLE SMIRKS-LIKE TRANSFORM

- Reactions are specified as SMIRKS transformations:

```
[BrD1h0+0:1][#6:2].[#7X3v3+0:3][H]>>[#6:2][#7:3]
1.6.2 BROMO_N_ALKYLATION
```

- As demonstrated in the example above, these patterns may operate on explicit hydrogen atoms for brevity, but these are "compiled" via more efficient SMARTS-like patterns for matching during naming.
- The nitrogen match becomes "`[#7X3v3h>0+0:3]`".

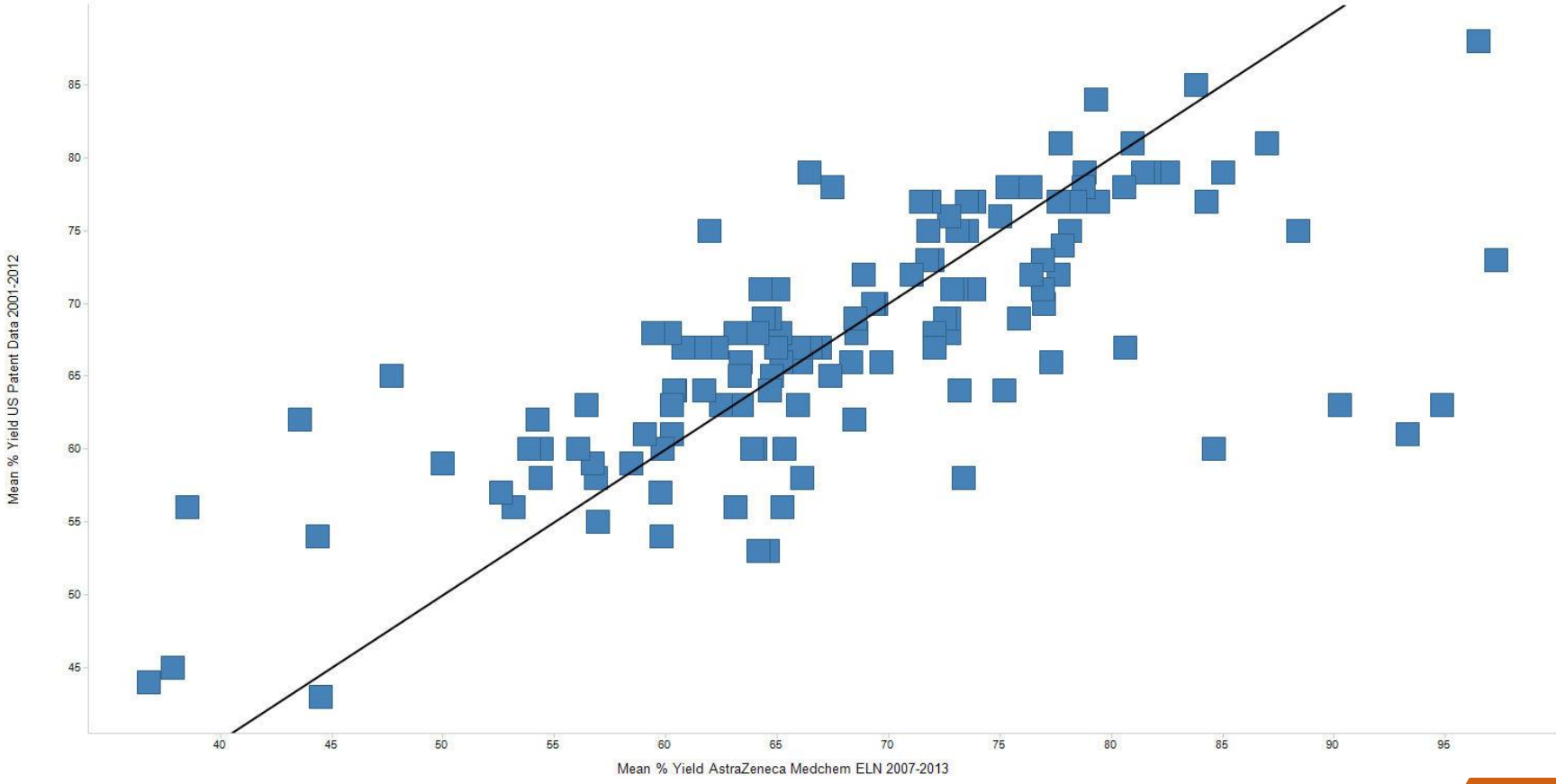
ETL* SUMMARY STATISTICS

	ELN Experiment Category	Fraction
A	NULL CLOBs	0.30%
B	Empty sketches	0.24%
C	No reaction (molecule?)	8.52%
D	No reactants	0.63%
E	No products	0.06%
F	Regular Reactions	88.93%
G	Markush Reactions	1.32%
	Total	100.0%

Export success for a typical pharmaceutical ELN is currently about 90.94% (see D+E+F+G above).

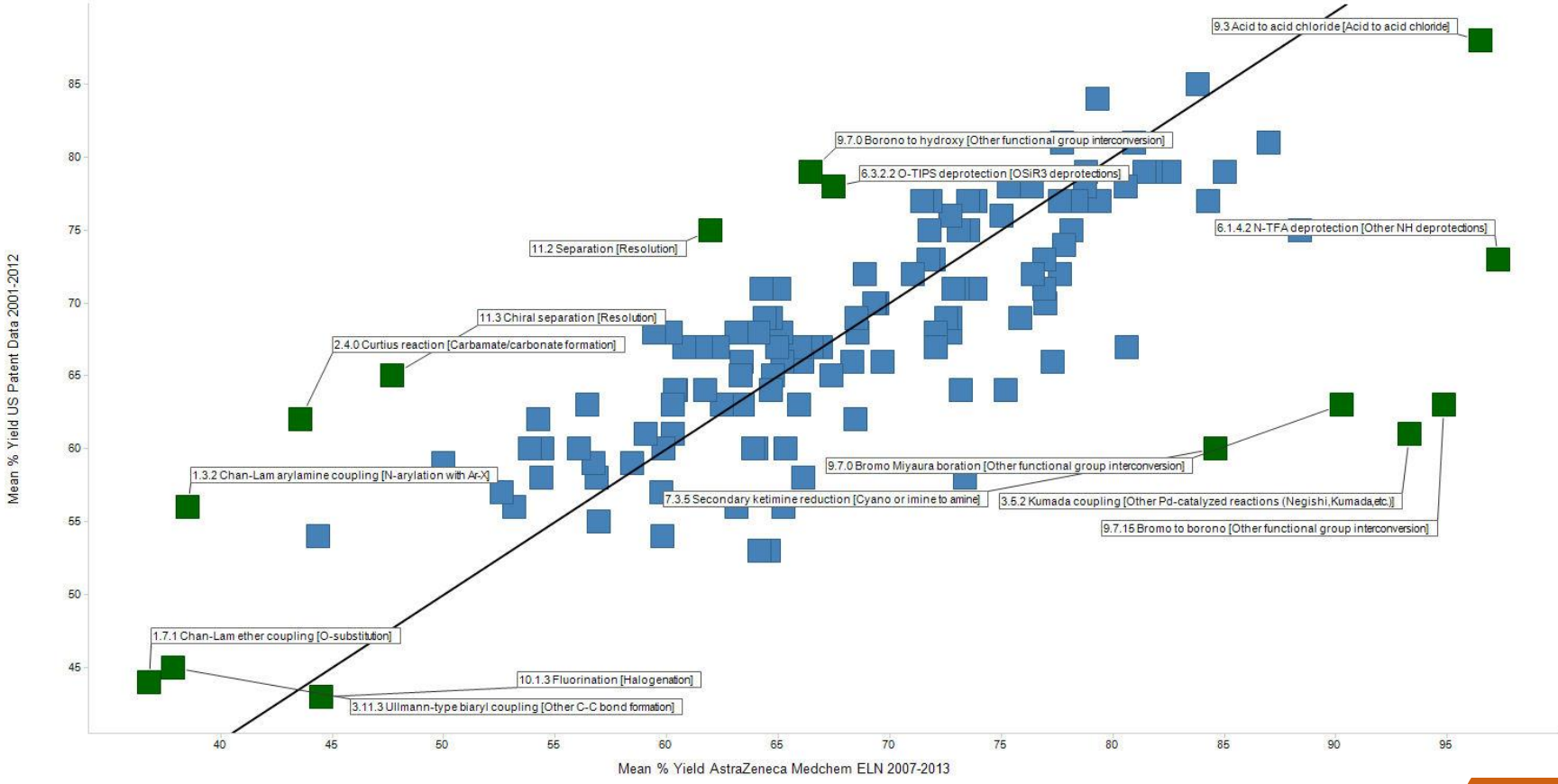
* Extract-Transform-Load (ETL), A data warehousing term.

ANALYSIS OF ELN REACTION YIELDS



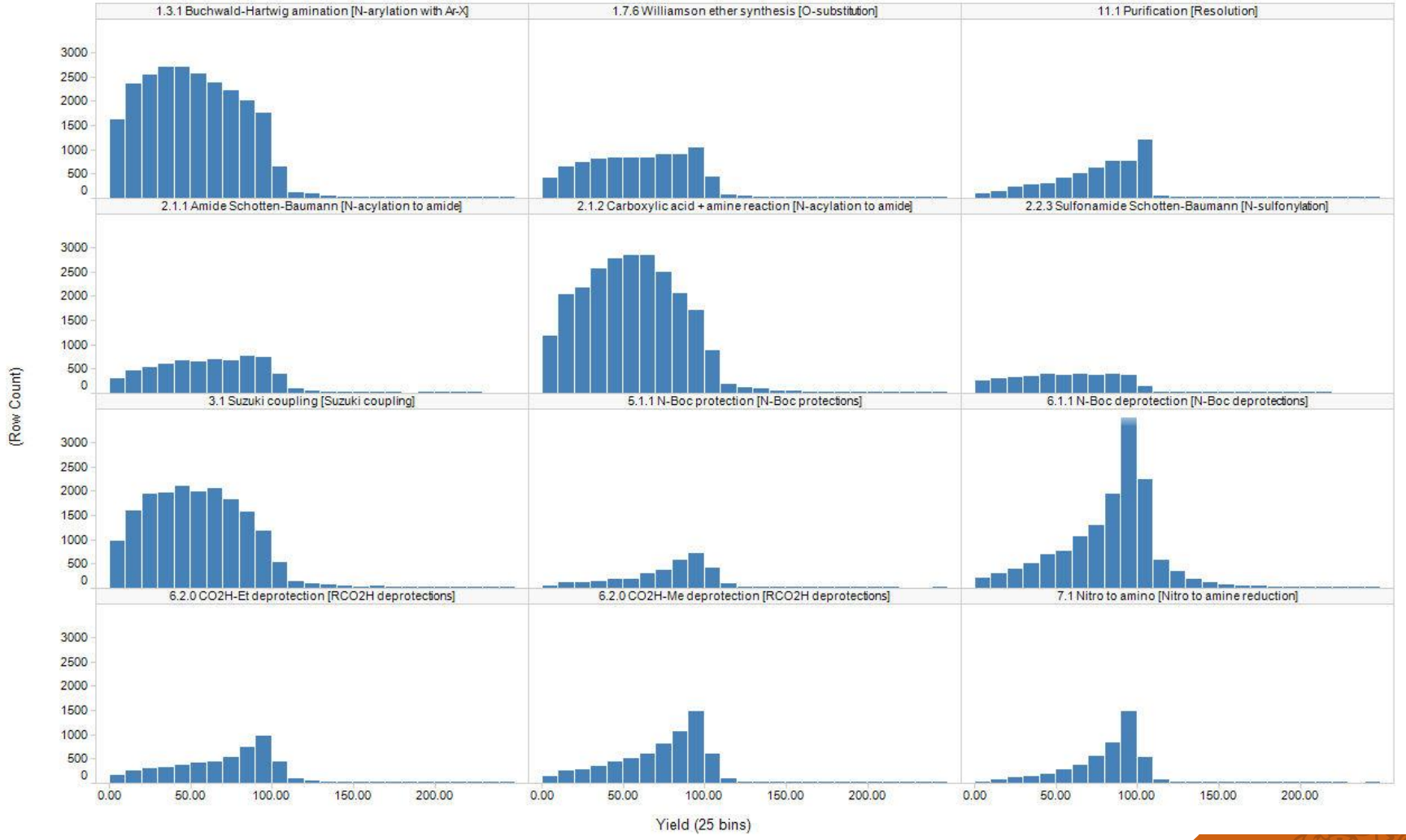
Data courtesy of Nick Tomkinson, AstraZeneca RDI, Alderley Park, UK.

ANALYSIS OF ELN REACTION YIELDS



Data courtesy of Nick Tomkinson, AstraZeneca RDI, Alderley Park, UK.

REACTION YIELD DISTRIBUTIONS



CONCLUSIONS

- In an attempt to better understand and hopefully improve the productivity of synthetic chemists, new computational methods have been developed to process “real world” organic reaction data.
- The fruits of this work now enable medicinal chemists and informaticians to make greater use of the wealth of information in their in-house ELNs.

ACKNOWLEDGEMENTS

- Anna Paola Pellicioli and Greg Landrum, Novartis,CH.
- Ethan Hoff and Manli Zheng, AbbVie, IL, USA.
- Colin Batchelor, RSC, Cambridge, UK.
- David Drake, AstraZeneca, Alderley Park, UK.
- Plamen Petrov, AstraZeneca, Molndal, SE.
- Daniel Stoffler, Hoffman-La Roche, Basel, CH.
- Pat Walters, Vertex Pharmaceuticals, MA, USA.
- Andrew Wooster, GSK, RTP, NC, USA.

- Thank you for your time.

CO-ORDINATE HANDLING

- Several approaches for handling 2D co-ordinates
  - Preserve original co-ordinates as drawn by chemist
  - Center and/or rescale for downstream depiction tools
  - Regenerate all 2D co-ordinates algorithmically
  - Clean-up long short/bonds introduced by superatom expansion, attempting to preserve original orientation.

NESTED REACTIONS

- Some ELN configurations allow more than one reaction or experiment per lab notebook page, i.e. multiple reaction sketches in different “tabs”.
- Two solutions to this breaking of the 1-to-1 mapping between COLLECTION_ID and reaction include:

1. Nested Reactions: Using the MDL RD file’s ability to embed each reaction step as data fields in a single record.
2. Splitting Reactions: Where each reaction has its own record, possibly duplicating shared data.

META DATA FIELDS

- In addition to the data explicitly recorded by the chemist and captured by the ELN, it is also often useful to export meta-data from an ELN schema.
- This includes data fields such as experiment creation data, creation modification date, experiment status (open/closed), chemist name, chemist user id, etc.

INCREMENTAL UPDATES

- An important feature for a production data warehouse for ELN reaction data is the ability to keep its contents valid/fresh with live data.
- The can be implemented by supporting incremental updates that export only those creations that have been modified or created since a given date or in a range of dates.
- One subtlety is the handling of “closed” experiments (i.e. those signed off by a supervisor), whose status change date need not match the last modified date.

REACTION ROLE ASSIGNMENT

- Normalized reaction roles may be assigned via reaction atom mapping algorithms.
- Reaction components that contribute atoms to the product are defined to be reactants, and the remaining components as catalysts and solvents.

- Hence, `c1cccc1[N+](=O)[O-].[Ni]>>c1ccccc1N`

  may be canonicalized as

  `c1ccccc1[N+](=O)[O-]>[Ni]>c1ccccc1N`

FUTURE WORK

- Preserving superatom definitions as S groups in V2000 and V3000 format Mol/RXN files.
- Enhanced stereochemistry (non-tetrahedral and axial chiralities for catalyst optimization).
- Improved support for Marvin and ISIS sketches.
- Support for IDBS e-Workbook ELN.