Efficient Searching and Similarity of Unmapped Reactions: Application to ELN Analysis

Roger Sayle, Ed Griffen, Thierry Kogej and David Drake
NextMove Software, Cambridge, UK
AstraZeneca, Alderley Park, UK
AstraZeneca, Molndal, Sweden

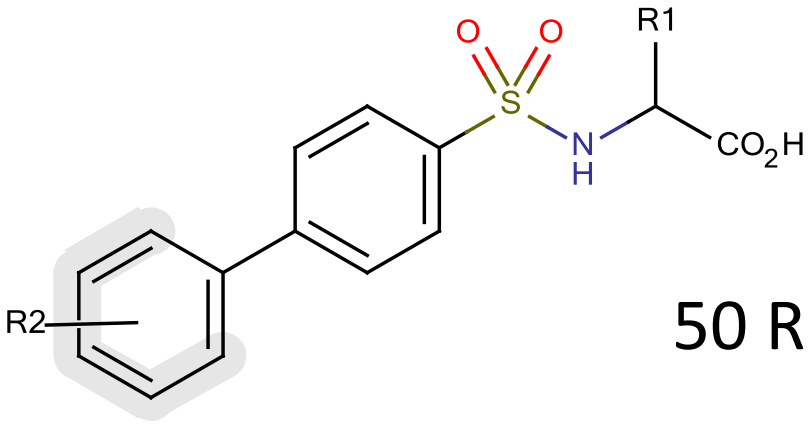
ACS National Meeting, San Diego, USA 27th March 2012

MOTIVATION

- Better understanding of chemical reactions can be used to improve the design-test-make cycles in the pharmaceutical industry, both for in-house synthesis and outsourcing to CROs.
- Although small molecule informatics is well supported, the computer representation, storage, manipulation, analysis, QSAR and property prediction of chemical reactions is relatively unexplored.

MOTIVATING EXAMPLE FROM GSK

- Stephen D. Pickett, Darren V.S. Green, David L. Hunt, David A. Pardoe and Ian Hughes, "Automated Lead Optimization of MMP-12 Inhibitors using a Genetic Algorithm", ACS Medicinal Chemistry Letters, Vol. 2, pp. 28-33, 2011.

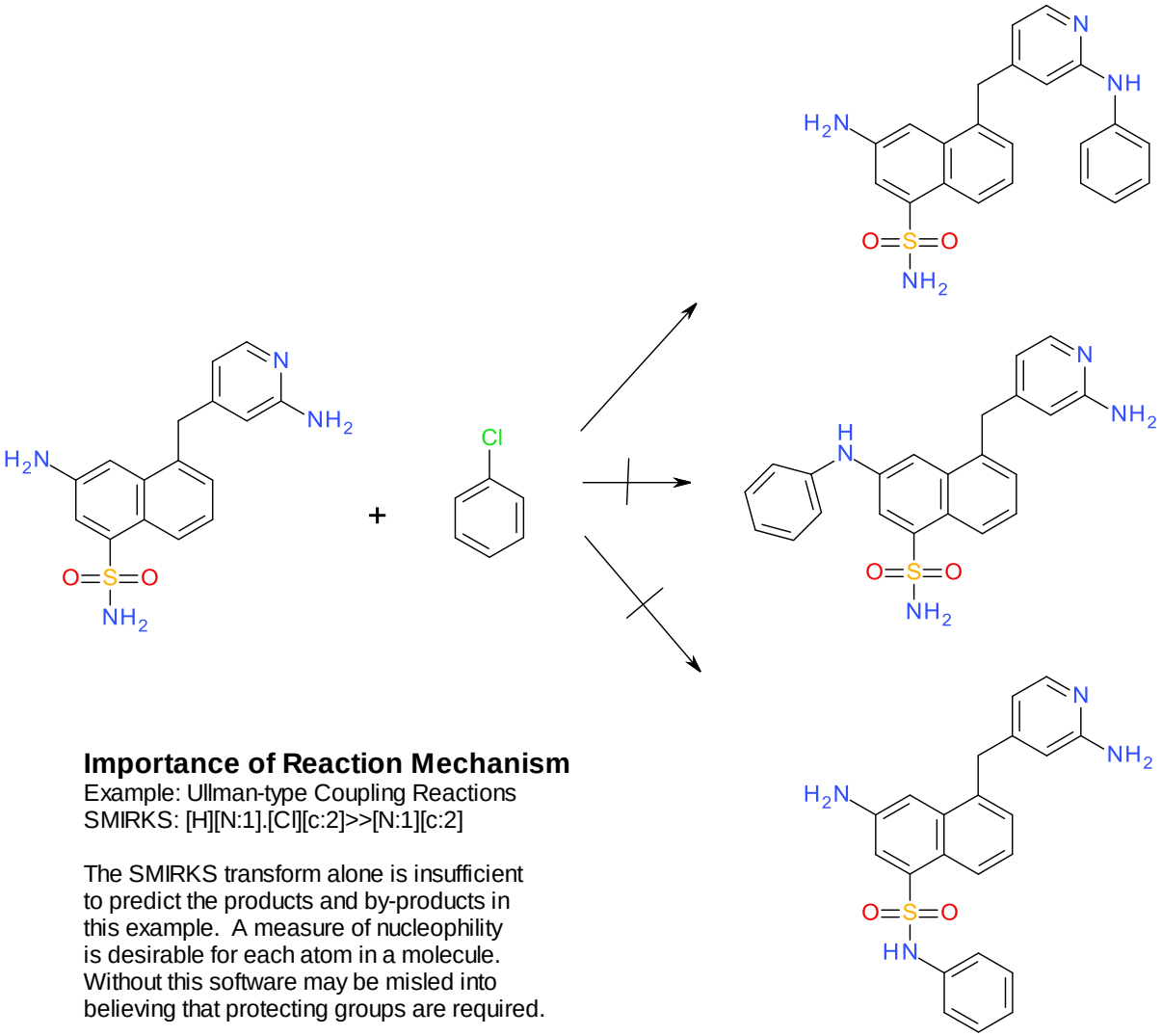


50 R1 x 50 R2 = 2500 cmpds

LIBRARIES IN PRACTICE

- Objective was to be diverse complete, so variations of standard route and conditions were required to avoid compromising diversity.
- Despite best efforts, 566 not made due to poor reactivity, unanticipated side reactions or product stability compared to 26 not assayed, 28 assay failed and 176 inactive, giving 1704 assay results.
- Interestingly, R1=21&R2=7 and R1=31&R2=25 were the two most active compounds (pIC50=8), R1=21&R2=25 had a pIC50=7.8 and R1=31&R2=7 was never made!

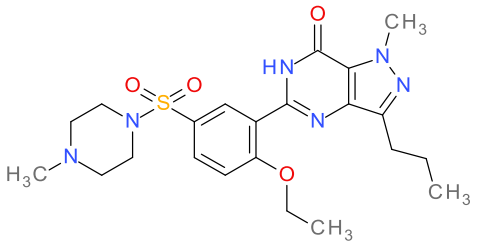
TRANSFORMS VS. REACTIONS



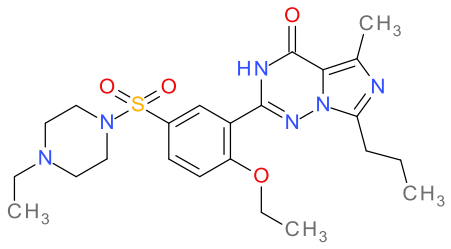
Importance of Reaction Mechanism
Example: Ullman-type Coupling Reactions
SMIRKS: [H][N:1].[Cl][c:2]>>[N:1][c:2]

The SMIRKS transform alone is insufficient to predict the products and by-products in this example. A measure of nucleophility is desirable for each atom in a molecule. Without this software may be misled into believing that protecting groups are required.

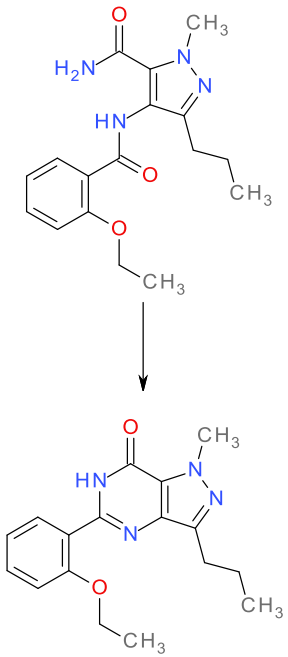
BEYOND DRUG GURU

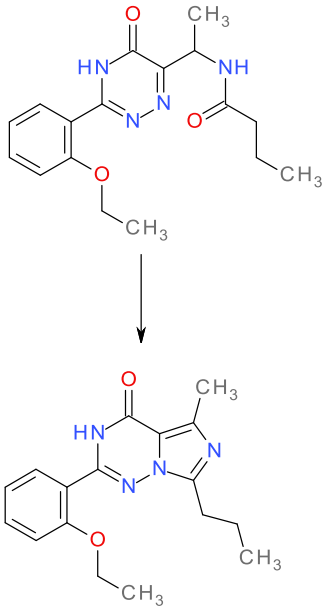


sildenafil (viagra)



vardenafil (levitra)





PROPOSED SOLUTION

- Analyze the hundreds of thousands of reaction examples described in in-house Electronic Laboratory Notebooks (ELNs).
- A typical use case would be to plot reaction yield for related reactions against catalysts and/or solvents used to determine the likely best or poor conditions.
- ELNs, unlike the literature, are strongly biased to the types of reactions/chemistries used in pharma.
- Most importantly, one of the only sources of negative data (reactions that have failed).

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. Improved reaction depiction.
6. Intuitive reaction searching.
7. Reaction similarity.
8. Reaction classification/clustering.

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.

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.

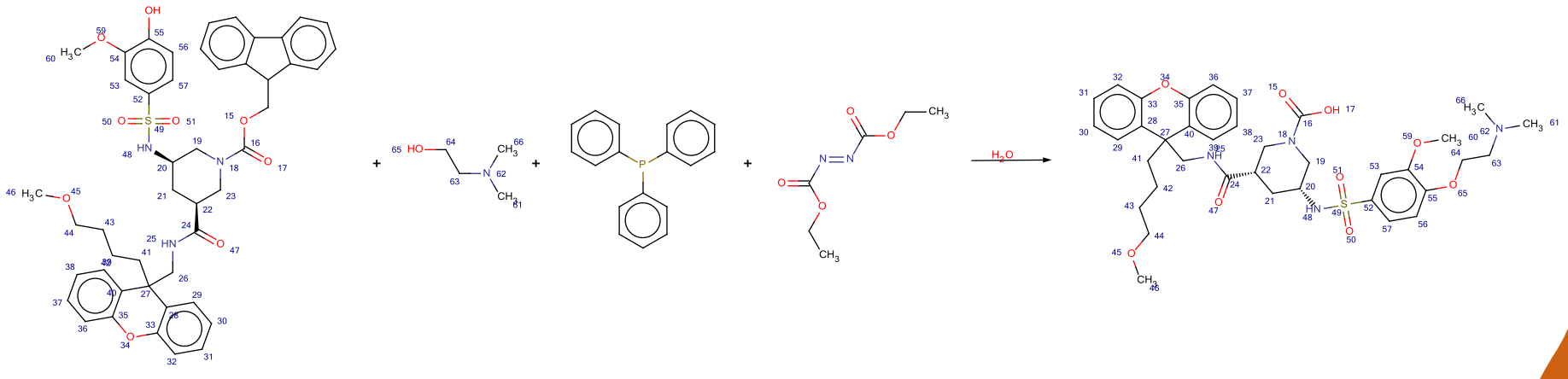
REACTION IDENTITY

- Continuing the notion of standardization, one might ask when are two reactions the same (duplicates).
- Two reactions that have matching agents (catalysts and solvents), reactants and products are considered identical, yet two that have the same reactants and products are considered the “same”.
- Whether a component is a reactant, catalyst, solvent or reagent may be consistently defined by atom-mapping; reactants contribute atoms to the product.

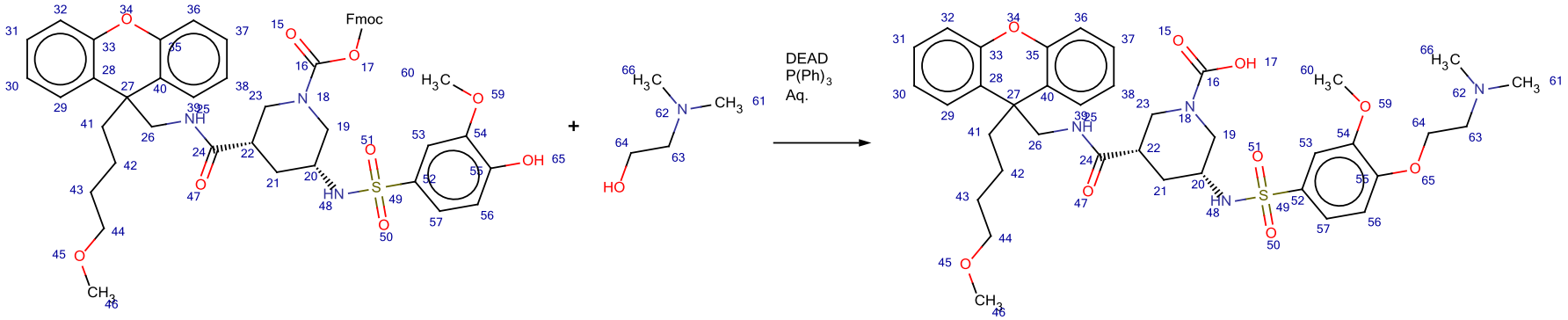
IMPROVED REACTION DEPICTION

- The 2D display of reactions can be improved by suitably aligning the reactants and products.
  - Molecules should be rotated or flipped (adjusting chirality) so that the substructures are consistent across an arrow.
  - The ordering reactants in coupling reactions should match their ordering in the product.
  - Ideally, the configuration of angles, torsions and chiral bonds should be consistent across a reaction arrow.
- Agents, solvents, salts and even protecting groups may be easier to interpret textually (superatoms).

REACTION DEPICTION



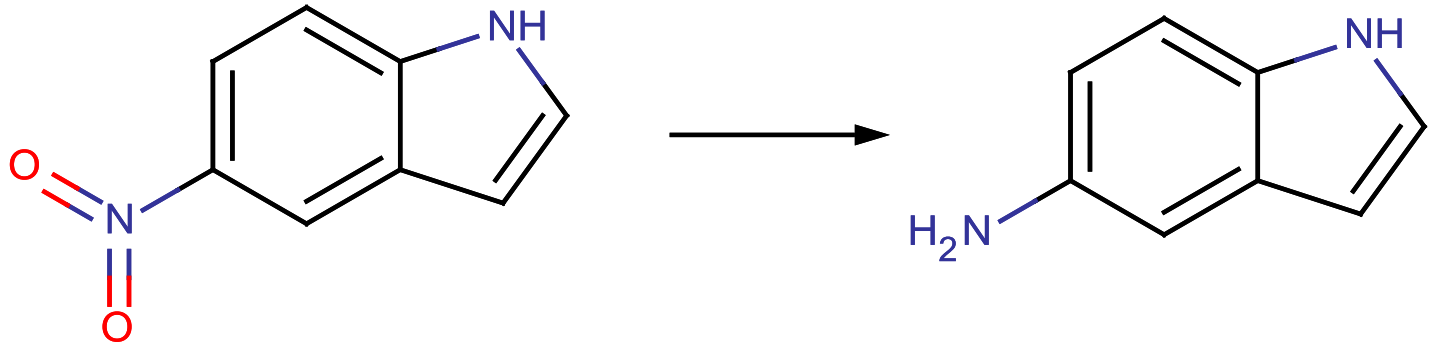
REACTION DEPICTION



RXN SEARCHING: MAKE OR BREAK

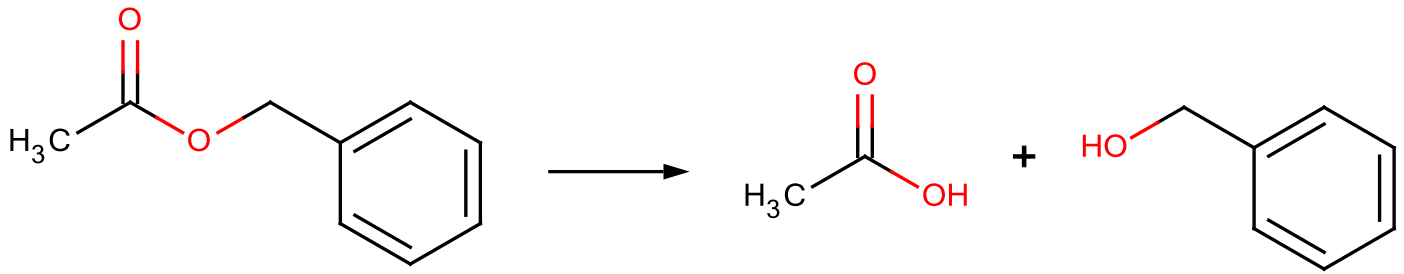
- In order to simplify the learning curve for exploiting reaction databases, an easy to use query mechanism has been developed for frequent use-cases.
- Query: Retrieve reactions that make indoles.
- Traditional interfaces provide either substructure search or require reaction centers/atom mapping.
- A more convenient (robust) mechanism is to check whether (or count) that a drawn substructure appears in the products but not in either the reactants or agents.

AVOIDING UNINTERESTING RESULTS



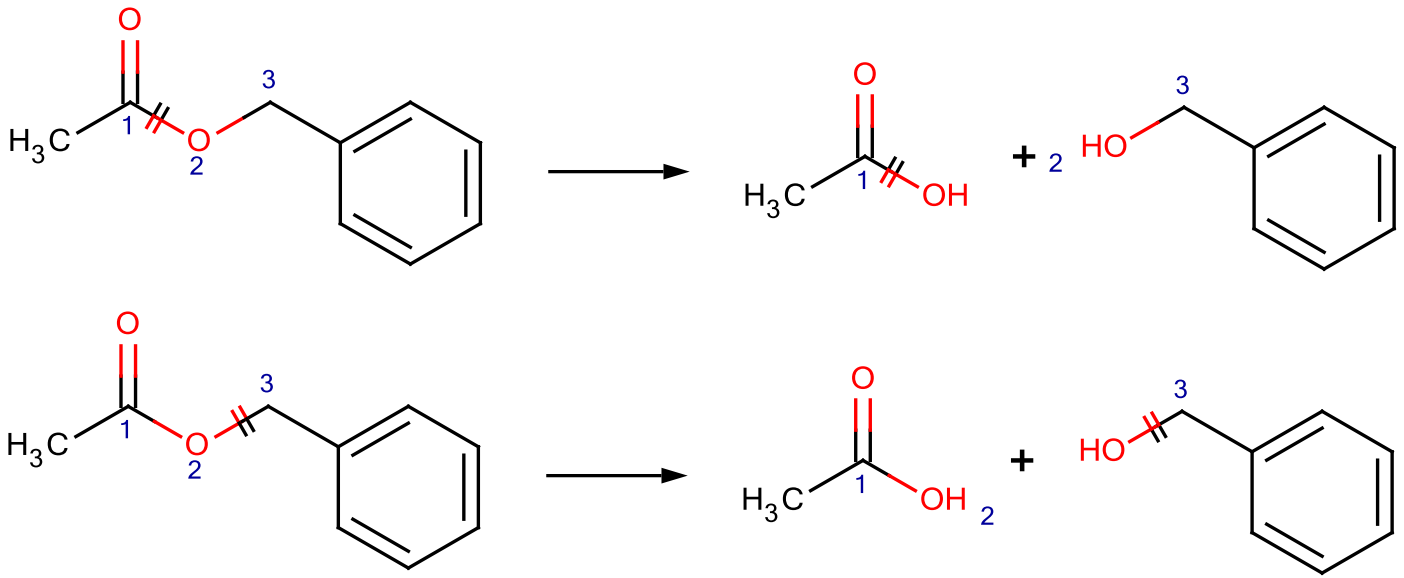
ATOM MAPPING AMBIGUITY

- Mechanistic atom mappings depend on mechanism.



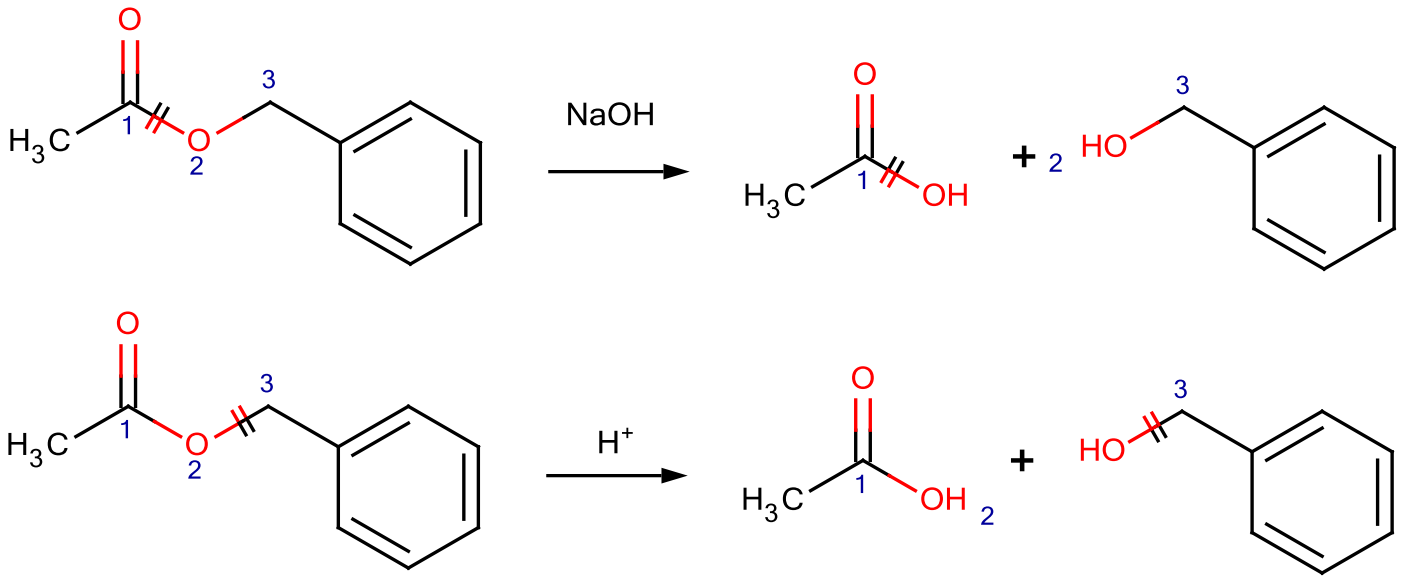
ATOM MAPPING AMBIGUITY

- Mechanistic atom mappings depend on mechanism.

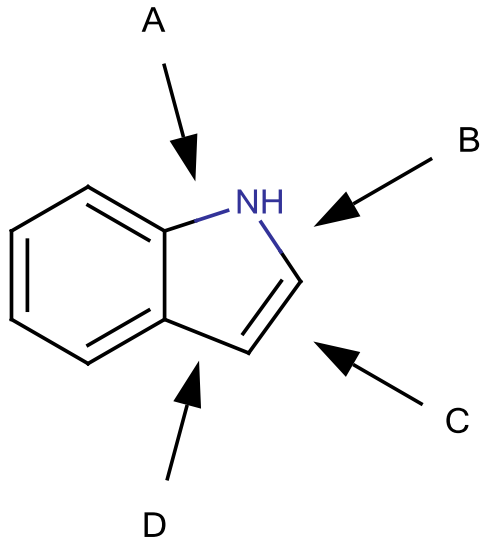


ATOM MAPPING AMBIGUITY

- Mechanistic atom mappings depend on mechanism.



BOND CHANGES IN INDOLE SYNTHESIS



Synthesis	A	B	C	D
Baeyer-Emmerling		M		
Bartoli		M		M
Bischler-Möhlau		M	C	M
Fischer		M	C	M
Fukuyama			M	
Hemetsberger	M			
Larock		M	C	M
Mandelung			M	
Nenitzescu	M			M
Reissert		M	M	
Larock		M	C	M

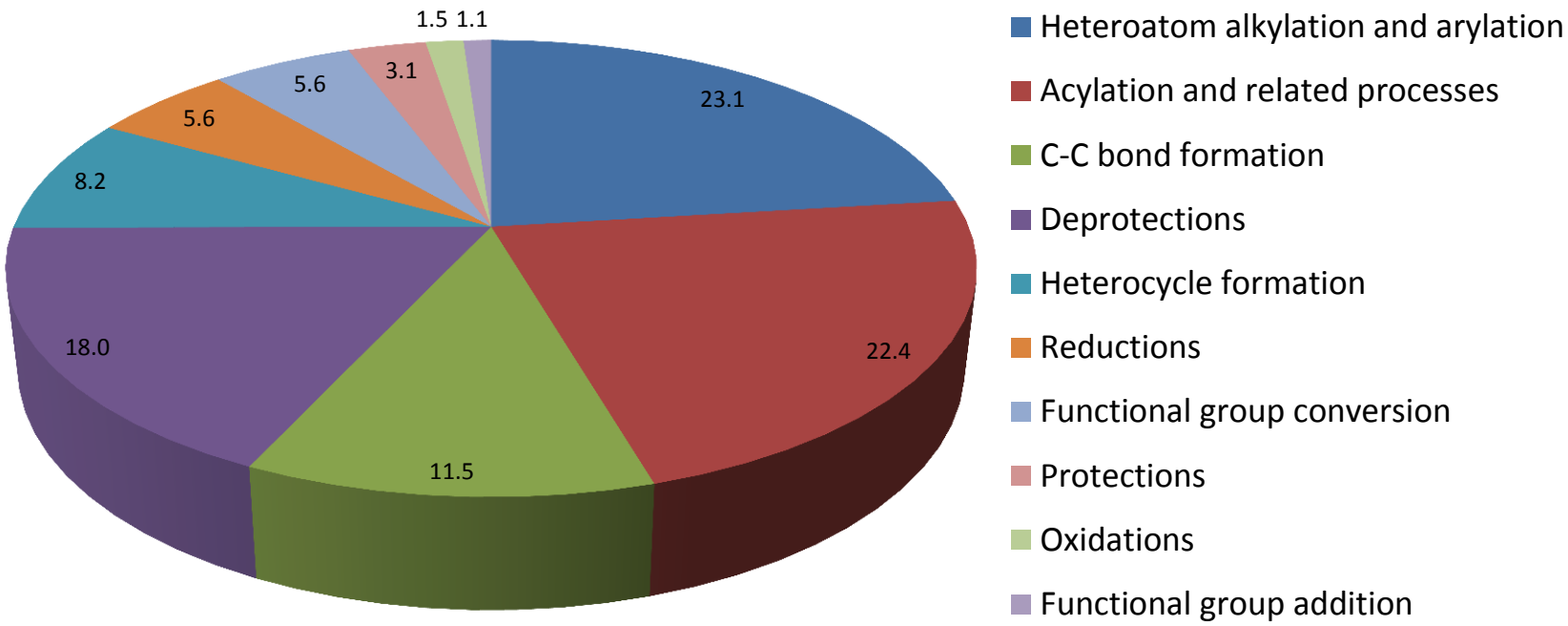
REACTION SIMILARITY

- To define the similarity between two reactions, we employ an ECFP variation of Daylight's "difference" fingerprints.
- Conceptually, a difference fingerprint encodes the differences between the fingerprint of the reactants and fingerprint of the products.
- We use differences in counted Scitegic/Accelrys ECFP fingerprints to capture the local environment of the reaction centre without requiring atom mapping.

REACTION CLASSIFICATION

- Although reaction similarity may be used to cluster and group mechanistically related reactions together, it is also convenient for them to be categorized by the type of reaction.
- Manually assignment into the RSC's RXNO ontology.
- InfoChem's $IC_{CLASSIFY}$ and $IC_{NAMERXN}$ tools.
- Algorithmic assignment of reaction class (compound purchase, purification, chiral separation) or recognized (named) reaction mechanism using SMIRKS transformations.

CATEGORIZATION OF REACTIONS



J. Carey, D. Laffan, C. Thomson, M. Williams, *Org. Biomol. Chem.* 2337, 2006.

S. Roughley and A. Jordan, *J. Med. Chem.* 54:3451-3479, 2011.

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

- Daniel Lowe, NextMove Software.
- Plamen Petrov and Mikko Vainio, AstraZeneca.
- Richard Bolton and Andrew Wooster, GSK.
- Peter Loew and Hans Kraut, InfoChem.
- Ethan Hoff, Abbott Laboratories.
- Stephen Roughley, Vernalis.

- Thank you for your time.