Six Not-So-Easy Pieces: Tautomers, Nucleic Acids, Inorganics, Reactions, Polymers and SMIRKS (in RDKit).

Roger Sayle, John Mayfield, Noel O'Boyle

NextMove Software, Cambridge, UK

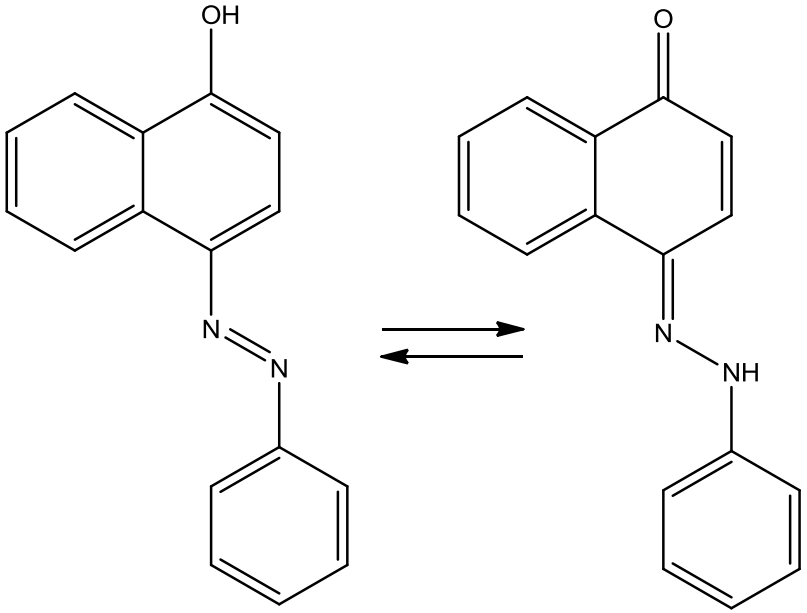
5th RDKit User Group Meeting, Novartis, Basel, Switzerland, Thursday 27th October 2016

QUICK OVERVIEW

1. **Tautomers**
    - Tautomer matching without enumeration
2. **Nucleic Acids**
    - RDKit support for RNA and DNA in PDB, HELM and FASTA.
3. **Inorganics**
    - Biovia's valence table revisions
4. **Reactions**
    - Some comments on expert rule sets
5. **Polymers**
    - InChI's recent support for polymers
6. **SMIRKS**
    - Standardizing interpretation/semantics of SMIRKS

1. TAUTOMERS

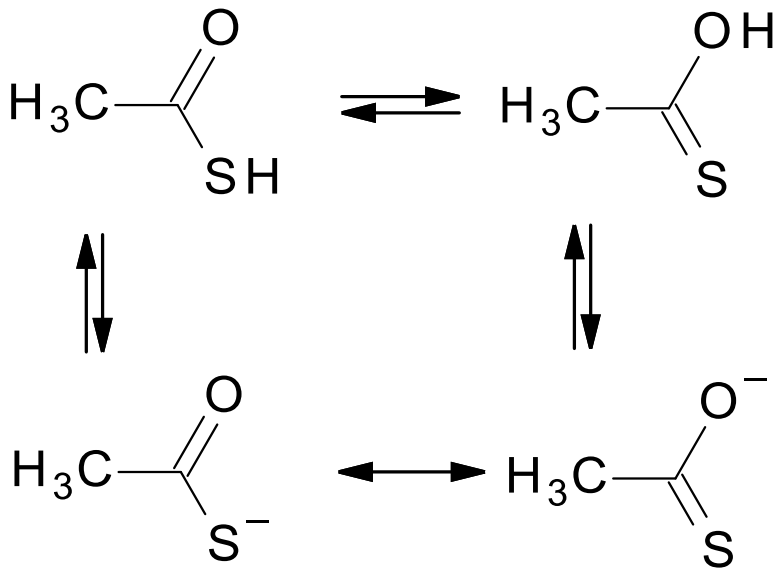
- Tautomers are molecular isomers that easily interconvert by migration of hydrogen atoms[1].



InChI=1S/C16H12N20/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H
InChI=1S/C16H12N20/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H

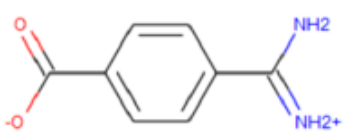
1. R. Sayle, "So you think you understand tautomerism?", JCAMD, 24(6-7):485-496, June 2010.

FRIENDS OF TAUTOMERS: MESOMERS & PROTOMERS

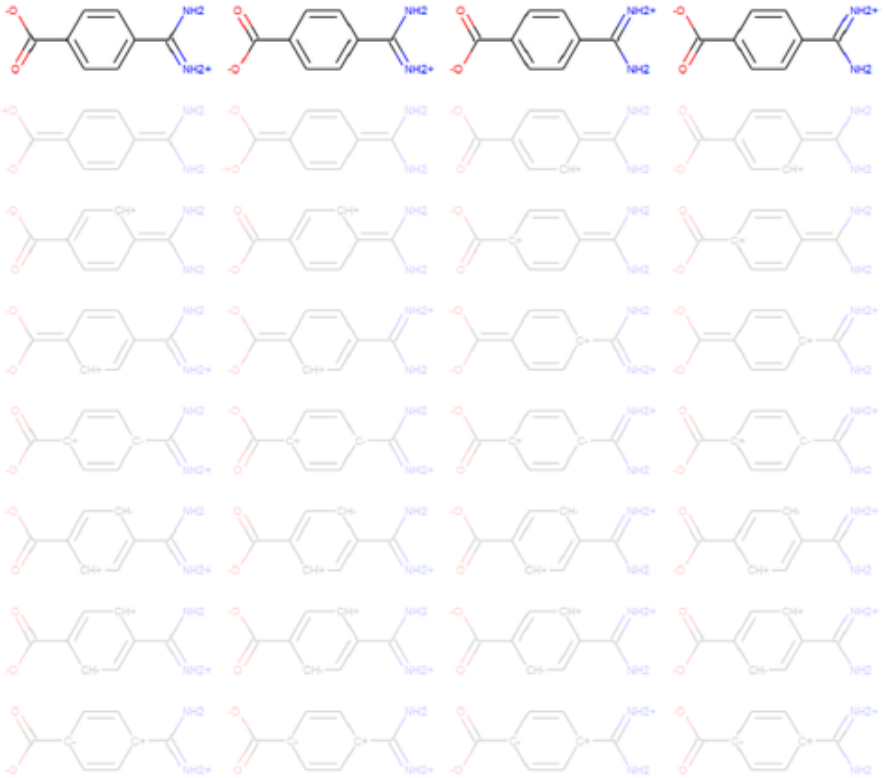


InChI=1S/C2H4OS/c1-2(3)4/h1H3,(H,3,4)
InChI=1S/C2H4OS/c1-2(3)4/h1H3,(H,3,4)/p-1

PAOLO'S SOLUTION: ENUMERATION



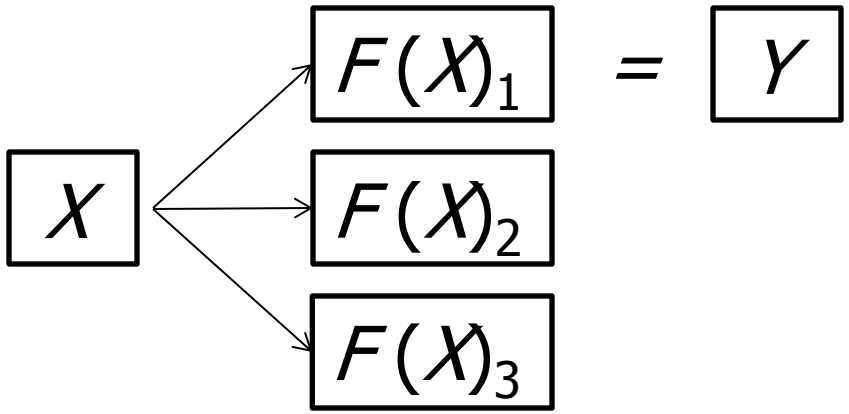
- The above molecule gives rise to 32 resonance structures
- Only the first 4 feature complete octets
- By default, the others are not generated, unless the ALLOW_INCOMPLETE_OCTETS, UNCONSTRAINED_CATIONS and UNCONSTRAINED_ANIONS flags are set



cresset

ENUMERATION VS. UNUMERATION

- Given a generator function F, which is a one-to-many mapping, many types of problem ask if $F(X) == Y$.



- Frequently, an efficient solution is to find the inverse many-to-one mapping, and ask if $F^{-1}(Y) == X$.

ALTERNATIVE APPROACH

- Automatically transform
  - NC(=N)c1ccc(cc1)C(=O)O
- Into the bond order, formal charge and polar hydrogen count suppressed SMARTS pattern
  - [#7D1]~[#6H0D3](~[#7D1])~[#6H0D3]1~[#6H1D2]~[#6H1D2]~[#6H0D3](~[#6H1D2]~[#6H1D2]~1)~[#6H0D3](~[#8D1])~[#8D1]
- Plus additional constraints on the total formal charge and the total polar hydrogen count.
  - polar_hcount – total_charge = 4

2. NUCLEIC ACIDS

- Support for DNA and RNA sequences in PDB, HELM and FASTA will shortly be contributed to RDKit.
- Only a minor set of changes to RDKit's APIs:
  - RDKit::SequenceToMol(string,sanitize,**flavor**)
  - RDKit::FASTAToMol(string,sanitize,**flavor**)
  - RDKit::PDBBlockToMol(string)
  - RDKit::HELMToMol(string)

  - RDKit::MolToSequence(mol)
  - RDKit::MolToFASTA(mol)
  - RDKit::MolToPDBBlock(mol)
  - RDKit::MolToHELM(mol)

BIOVIA VS. PISTOIA HELM ISSUES

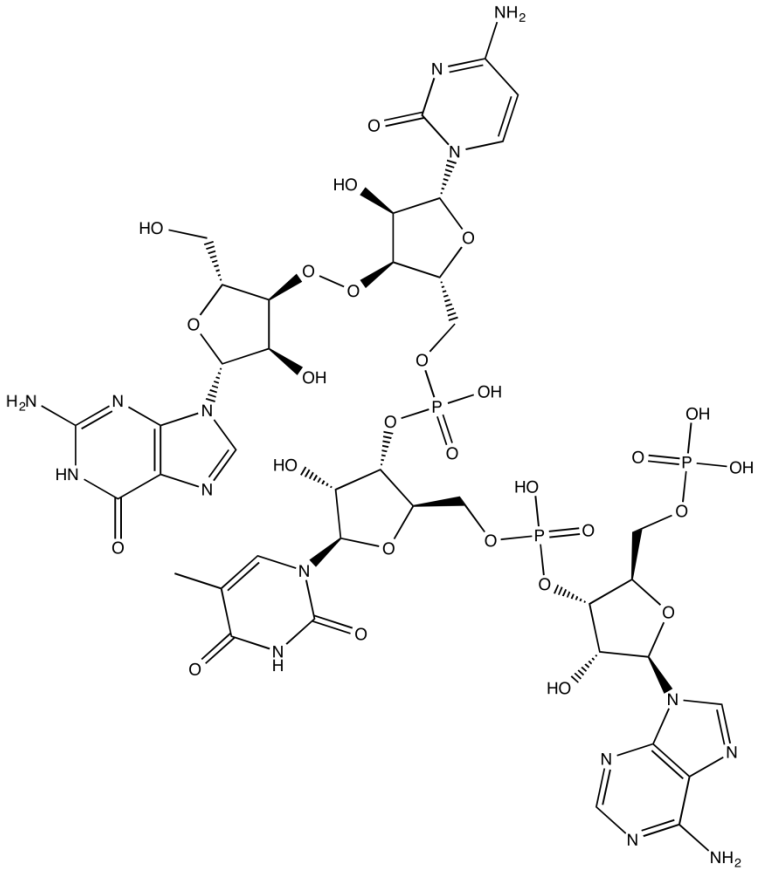
- RDKit's implementation doesn't have Biovia's bugs.
- The monomer phase problem in RNA registration:
  - Sequence: GATTACA
  - Interpretation: 5'-G-A-T-T-A-C-A-3'
  - Implied phosphates: 5'-G-P-A-P-T-P-T-P-A-P-C-P-A-3'
  - IUMB/PDB/Biovia: G-PA-PT-PT-PA-PC-PA
  - Pistoia HELM: GP-AP-TP-TP-AP-CP-A
- This discrepancy complicates RNA registration and leads to serious bugs in Biovia's HELM support.

RNA BUGS IN BIOVIA HELM SUPPORT

- From Sequence
  - "ACGT" drawn in Biovia has 4 phosphates, but the HELM editor interprets this sequence as having three.
  - Biovia's Mol file for this sequence places the phosphate at the 5' end, but the HELM string it generates has it at the 3'.
- HELM interpretation
  - RNA1{R(A).P.R(C)P.R(T)P.R(G)}$$$$ is exactly the same molecule as RNA1{R(A).PR(C).PR(T).PR(G)}$$$$ in HELM.
  - These strings have very different behaviour in Biovia, neither of which has the correct connection table.

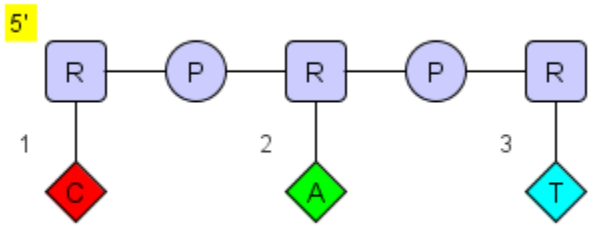
EXPANDED SCSR FILE

select molfile(mol('RNA1{R(A)P.R(T)P.R(C)P.R(G)}$$$$')) from dual;

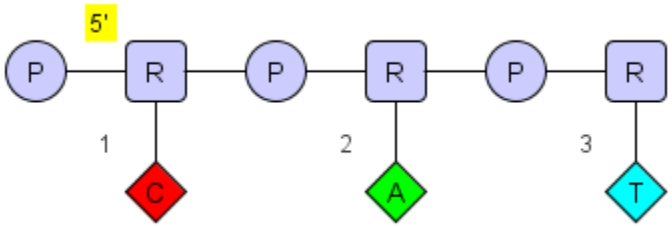


RDKIT HELM IMPLEMENTATION

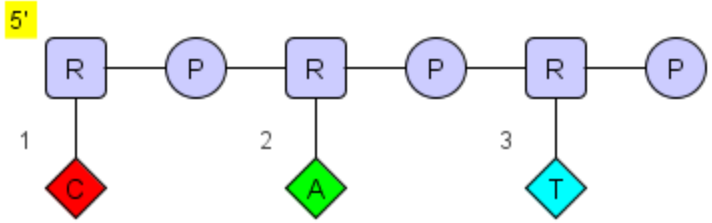
No Caps
Flavor =2



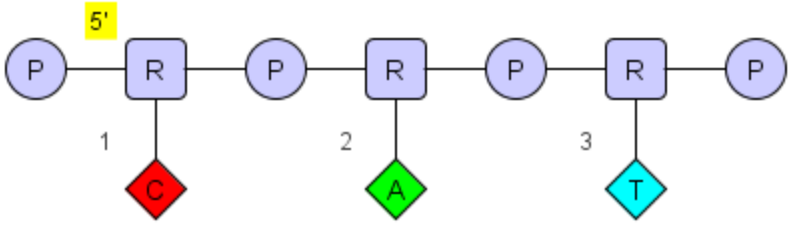
5’-Cap
Flavor = 3



3’-Cap
Flavor = 4



Both Caps
Flavor = 5

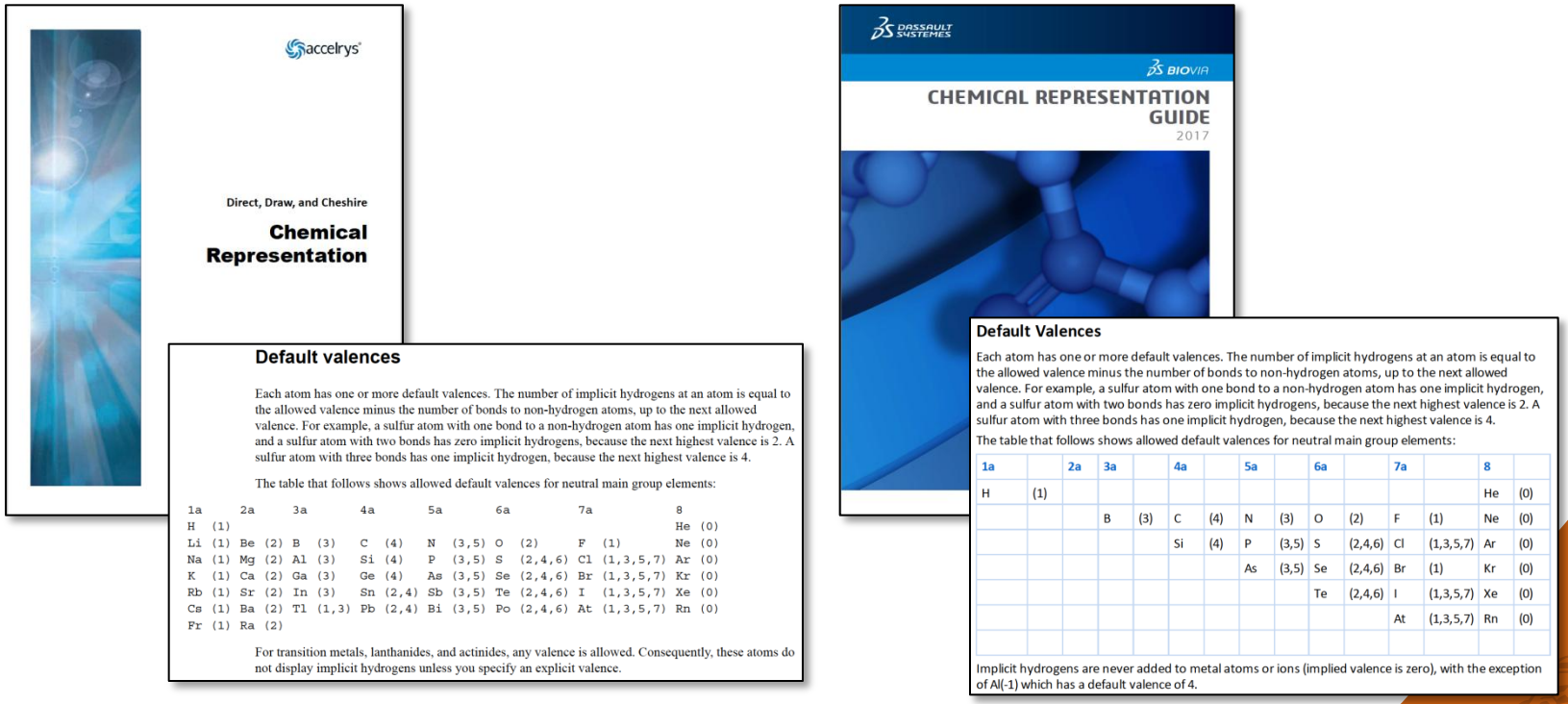


3. INORGANICS

- Sanitization failures: Insane or insanitary?
- Great recent advances in RDKit.
  - Hexafluorophosphate
  - Dess-Martin Periodinane
- Perhaps exceptions for remaining “platypi”.
  - Chlorine Dioxide O=[Cl]=O

MDL MOLFILE-AGEDDON

- Biovia 2017 changes the interpretation of MDL files.
- This affects over 213097 CIDs in PubChem!



HEAVY METALS (IT GOES BEYOND 111)

Atomic Number	Symbol	IUPAC	RDKit Ptable	RDKit SMILES	Toolkit Heaviest	Biovia 2017	SMARTS
103	Lr		Y	Y	Indigo	Lr	
104	Rf		Y	Y		'Rf'	
105	Db		Y			'Db'	
106	Sg		Y			'Sg'	
107	Bh		Y			'Bh'	
108	Hs		Y			'Hs'	
109	Mt		Y		ChemAxon	'Mt'	
110	Ds		Y			'Ds'	
111	Rg		Y		OEChem	'Rg'	
112	Cn		Y		OpenBabel	'Cn'	
113	Nh	2016?					N&h
114	Fl					'Fl'	
115	Mc	2016?					
116	Lv				CDK	'Lv'	
117	Ts	2016?					
118	Og	2016?					

4. REACTIONS

- Virtual library enumerations are overly optimistic.
  - SAVI, SCUBIDOO, Enamine, PLC, Evotec, Novartis?

- Transformations vs. Reactions is a major challenge.
  - Which reactions to use?
  - Reactions that do happen!
  - Reactions that don't happen!

WHICH REACTIONS?

- NCI/LHASA SAVI (14+6 reactions)
  - Suzuki Coupling 36262 out of 1.2M USPTO examples
  - Sulfonamide Schotten-Baumann 14348 out of 1.2M USPTO examples
  - Buchwald-Harwig 6040 out of 1.2M USPTO examples
  - Hiyama coupling 458 out of 1.2M USPTO examples
  - Fukuyama coupling 2 out of 1.2M USPTO examples
  - Liebeskind-Srogl coupling 0 out of 1.2M USPTO examples
- Hartenfeller (58 reactions)
  - #1 Pictet-Spengler reaction 7 out of 1.2M USPTO examples
  - #10+ #11 Azide-nitrile Huisgen-cycloaddition 5 out of 1.2M USPTO examples
  - #17 Pyridone synthesis 2 out of 1.2M USPTO examples
  - #20 Phthalazinone synthesis 16 out of 1.2M USPTO examples
  - #24 Friedlander quinoline synthesis 30 out of 1.2M USPTO examples
- Enamine REAL (43 reactions)
  - Thiourea to guanidine (14 out of 160M examples).

REACTIONS THAT DO HAPPEN!

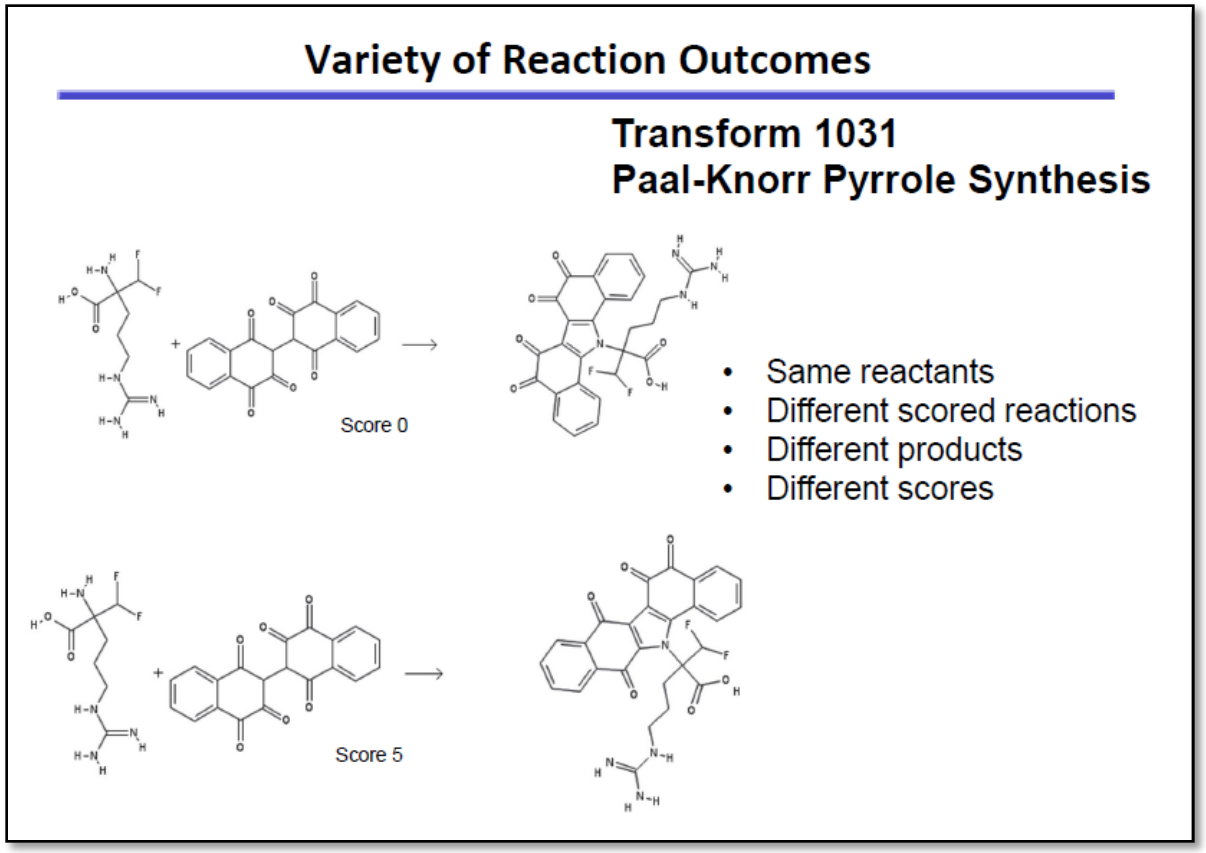
- Chloro Sonogashira Couplings (@ NCI)
  - The LHASA 'CHMTRN' rules for Transform 2267, Sonogashira couplings, (presented by Marc Nicklaus at Sheffield) states "Iodides are usually more reactive than bromide. Chlorides do not react".

Code	Name	Count	Yield
3.3.2	Bromo Sonogashira coupling	3717	49.3%
3.3.3	Chloro Sonogashira coupling	429	44.2%
3.3.4	Iodo Sonogashira coupling	2721	64.9%

- Isotopically-labelled Compounds (@Eli Lilly)
  - The LAAR reactions used to construct Lilly's PLC (Nicolaou et al. 2016) forbid the presence of isotopic labels in reactants.

REACTIONS THAT DON'T HAPPEN!

- Paal-Knorr Pyrrole Synthesis vs. Aldehydes/Ketones

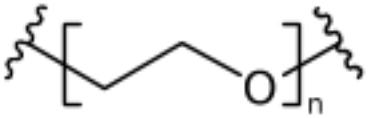


Of the 430 examples of Paal-Knorr pyrrole synthesis reported in US patent applications 2001-2012, exactly zero have more than the two reacting ketones/aldehydes.

5. POLYMERS

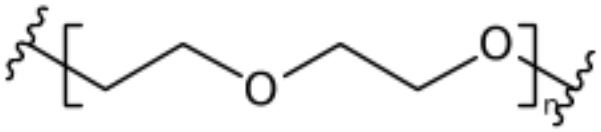
- The InChI Trust has added polymer support to InChI.
- Or has it?

Experimental Beta Option **-Polymers**



PEG_n

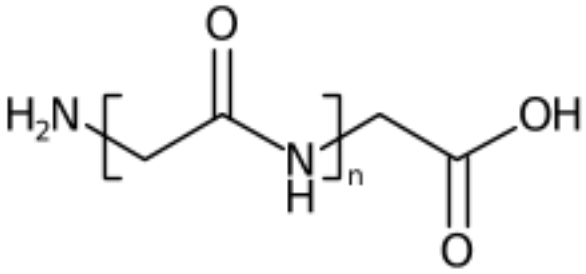
InChI=1B/C2H4O/c1-2-3-1/h1-2H2/z101-1-3(1,2,1,3,2,3)
InChIKey=IAYPIBMASNFSPL-GCGQHNKHBA-N



PEG_2n

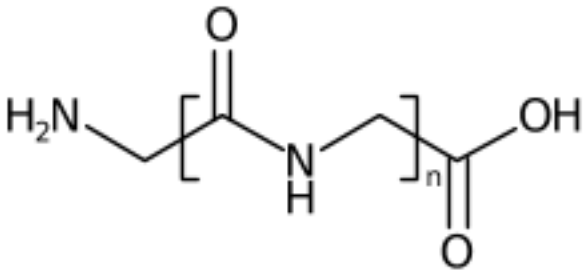
InChI=1B/C4H8O2/c1-2-6-4-3-5-1/h1-4H2/z101-1-6(1,2,1,5,2,6,3,4,3,5,4,6)
InChIKey=RYHBNJHYFVUHQT-UEXHMUNTBA-N

INCHI POLYMERS
CAPPING GROUP FRAME SHIFT



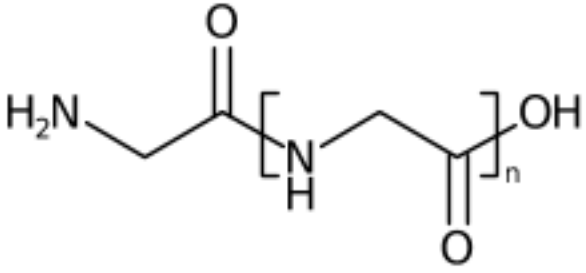
pGly_1

InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-1,3,6-7(2-6,5-1)
InChIKey=YMAWOPBAYDPSLA-NPFRMHEKBA-N



pGly_2

InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-2-3,6-7(1-3,4-2)
InChIKey=YMAWOPBAYDPSLA-LCMLVUGIBA-N



pGly_3

InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-2,4,6,8(3-6,9-4)
InChIKey=YMAWOPBAYDPSLA-YYSJEQPSBA-N

6. SMIRKS

- SMIRKS is Daylight's Reaction Transform Language.
- SMILES is nearly portable, SMARTS is improving, but...
- The RSC's CVSP platform contains rules such as:

[N-,P-:1][C:2]=[N+,P+:3]>>[N,P:1]=[C:2][N,P:3]

1. Daylight SMIRKS requires SMILES on the RHS, but allows SMARTS on the LHS.
2. In SMIRKS, a property is matched/modified if specified.

[N-,P-:1][C:2]=[N+,P+:3]>>[*+0:1]=[C:2][*+0:3]

NEED AN OPENSMIRKS STANDARD?

- Product Primitives
  - SMILES Primitives
    - #N Atomic Numbers
    - N Isotopic Mass
    - +N/-N Formal Charge
  - SMARTS Primitives
    - hN/HN (Implicit) Hydrogen Count
    - A/a Aliphatic/Aromatic

ACKNOWLEDGEMENTS

- The RDKit Hackers at Novartis (and T5)
  - Greg Landrum
  - Nadine Schneider
  - Joann Prescott-Roy
- The team at NextMove Software
  - John Mayfield
  - Noel O'Boyle
  - Daniel Lowe
- And many thanks for your time!