Evaluating standardisation rules for reaction classification

Ingvar Lagerstedt
John Mayfield
Roger Sayle

NextMove Software, Cambridge, UK

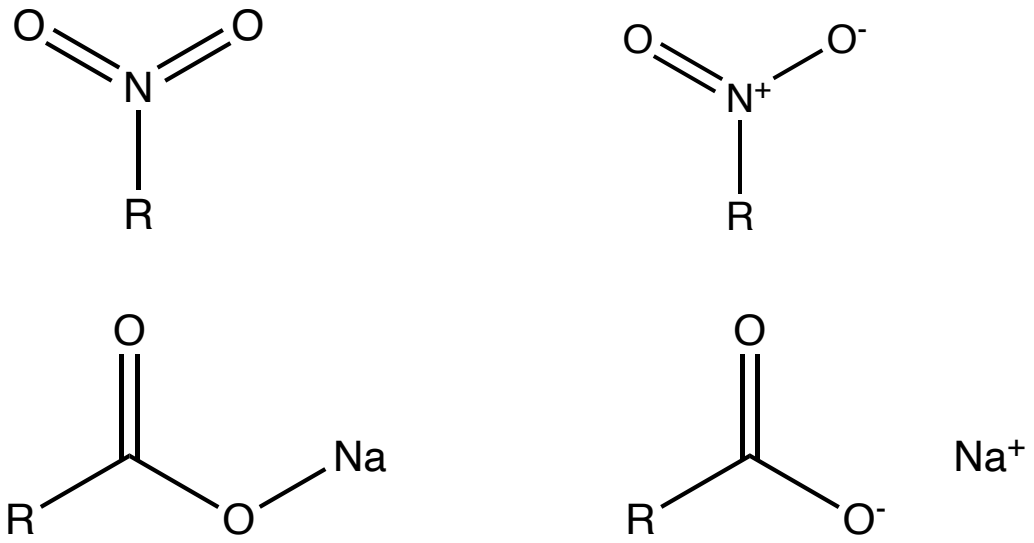
ACS Fall 2023, San Francisco, CA, Wednesday 16th August 2023

OUTLINE

- Standardisation
- Tautomers
- Data
- Tools
- NameRxn and use of standardised reactions

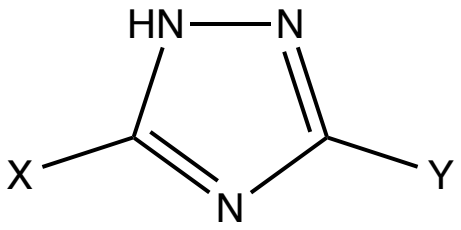
STANDARDISATION

Standardise the representation of chemical compounds to decide do you want nitro groups to be pentavalent or charge separated, if salts are bound or not.

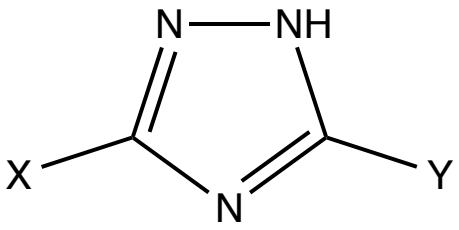


STANDARDISATION – WHAT FOR

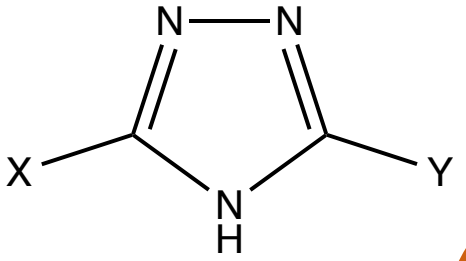
If standardising for lookup, aim for high symmetry. In 1,2,4-triazole pick *4H*. When looking at reactions, picking the dominant tautomeric form in the reaction medium may be preferable, in which case there is a choice of 1*H* and 2*H*. If interested in how the molecule interacts in the body, dissolved in water (or weak saline solution) could be of interest.



1*H*



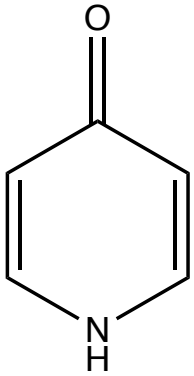
2*H*

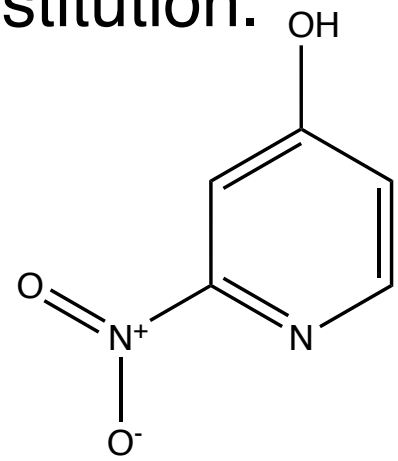


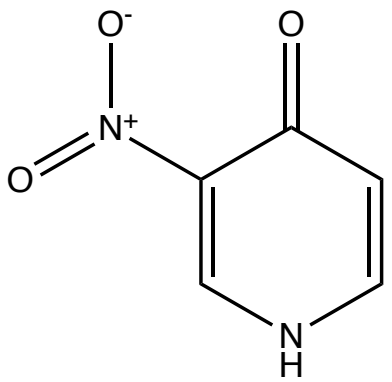
4*H*

SUBSTITUENT EFFECTS

Substituents can significantly change the tautomer distribution, though to what degree varies. For 4-pyridone in water, 2-substitution has a much larger effect than 3-substitution.



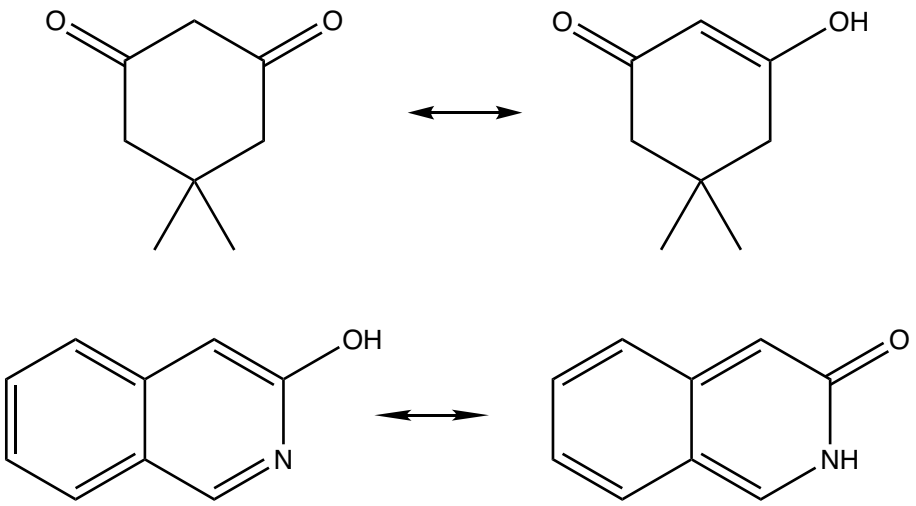




A 2-nitro group changes the preferred form to pyridol in water, but a 3-nitro group has little impact.

SOLVENT EFFECTS

Two examples where the solvent has a significant impact, in each the form on the left is dominant in chloroform, the form on the right is dominant in water



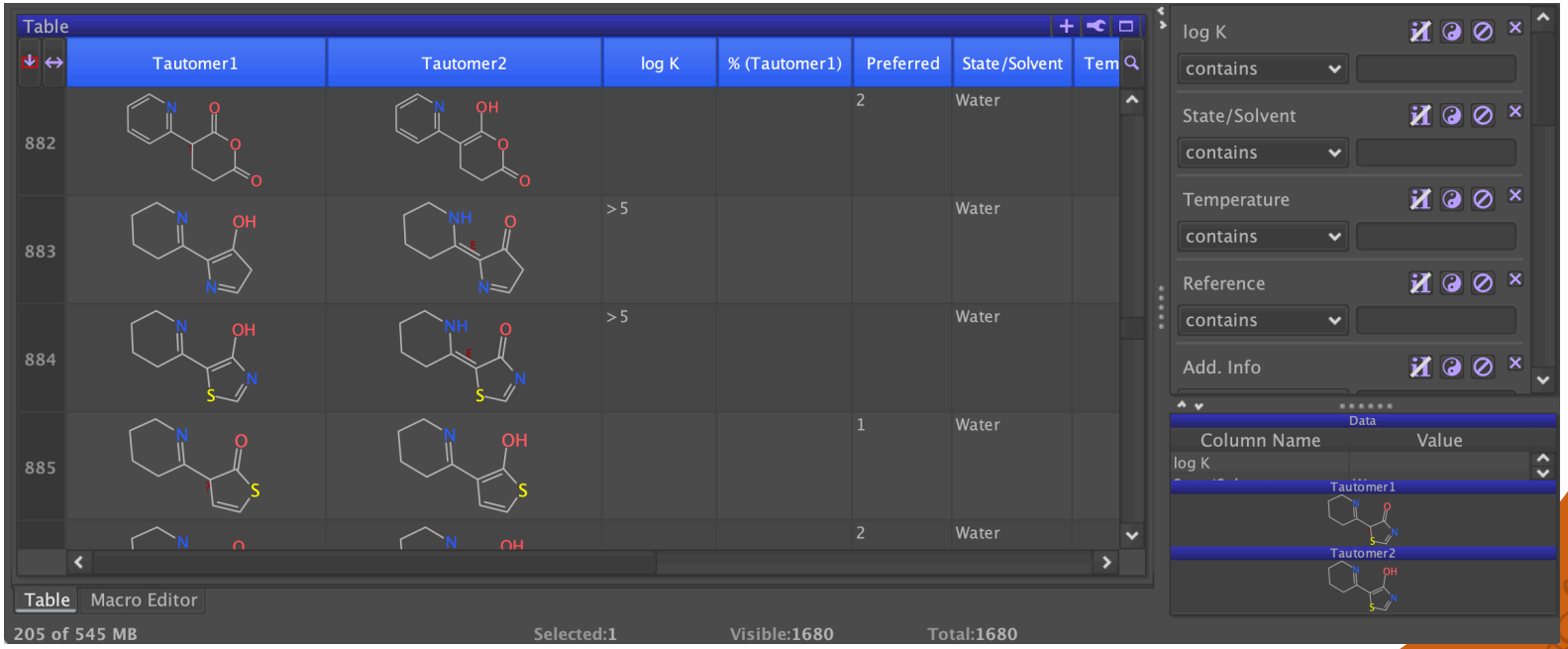
DATA SOURCES

There are limited public sources on tautomers, no doubt more is available in corporate collections. Here we have used Tautobase[1] which is based on a paper by Taylor and Kenny[2]. NIH has provided a public data set with 5977 tautomers in 2819 tuples[3], relatively few of those in water. A useful older source is the book by Elguero et al[4].

1. Oya Wahl and Thomas Sander, Tautobase: An Open Tautomer Database, Journal of Chemical Information and Modeling 2020 60 (3), 1085-1089, https://github.com/WahlOya/Tautobase
2. Peter J. Taylor† and Peter W. Kenny, The Prediction of Tautomer Preference in Aqueous Solution (Version 1.0), 2019, https://figshare.com/articles/preprint/The_Prediction_of_Tautomer_Preference_in_Aqueous_Solution_Version_1_0_/8966276
3. Devendra K. Dhaked, Laura Guasch, and Marc C. Nicklaus, Tautomer Database: A Comprehensive Resource for Tautomerism Analyses. Journal of Chemical Information and Modeling 2020 60 (3), 1090-1100, https://cactus.nci.nih.gov/download/tautomer/
4. J Elguero, C Marzin, AR Katritzky, P Linda, The Tautomerism of Heterocycles, Academic Press, New York, 1976

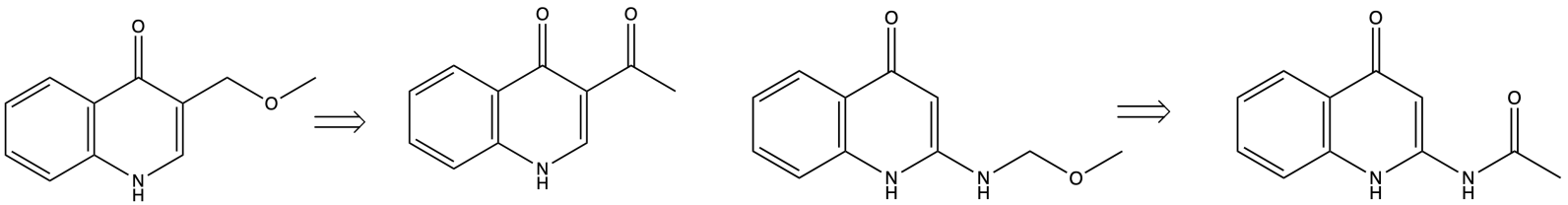
TAUTOBASE 1

Tautobase contains a collection of 1680 tautomer pairs in a variety of solvents taken from Taylor and Kenny. Of those 925 pairs are in water and another 403 has unknown/undefined solvent/state. The assignments are done from a mixture of measurements, LFER calculations, or calculation of the relative energy of the relevant tautomers.

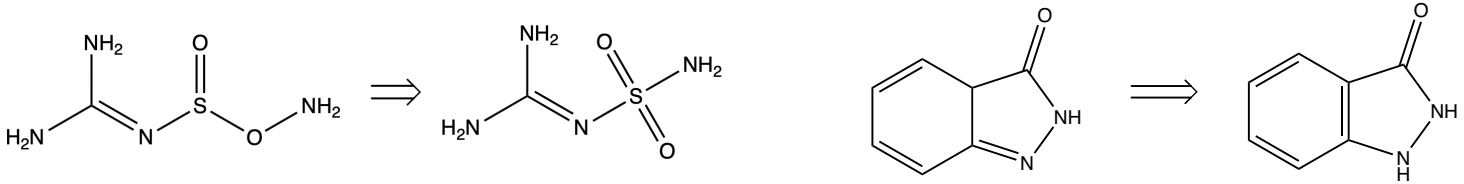


TAUTOBASE 2

We noticed that some of the Tautobase entries had functional groups that was somewhat unexpected, so decided to check the entries in the source. We have identified 16 cases where the entry or assignment in Tautobase do not match Taylor and Kenny, 3-4 % of entries looked at so far.



Tautobase 598 Tautobase 601

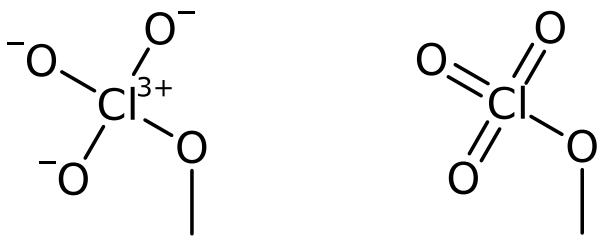


Tautobase 1167 Tautobase 1587

TOOLS

Tools for standardisation are available from various vendors and toolkits, and what form to adopt is to some degree a matter of style, as with nitro groups.

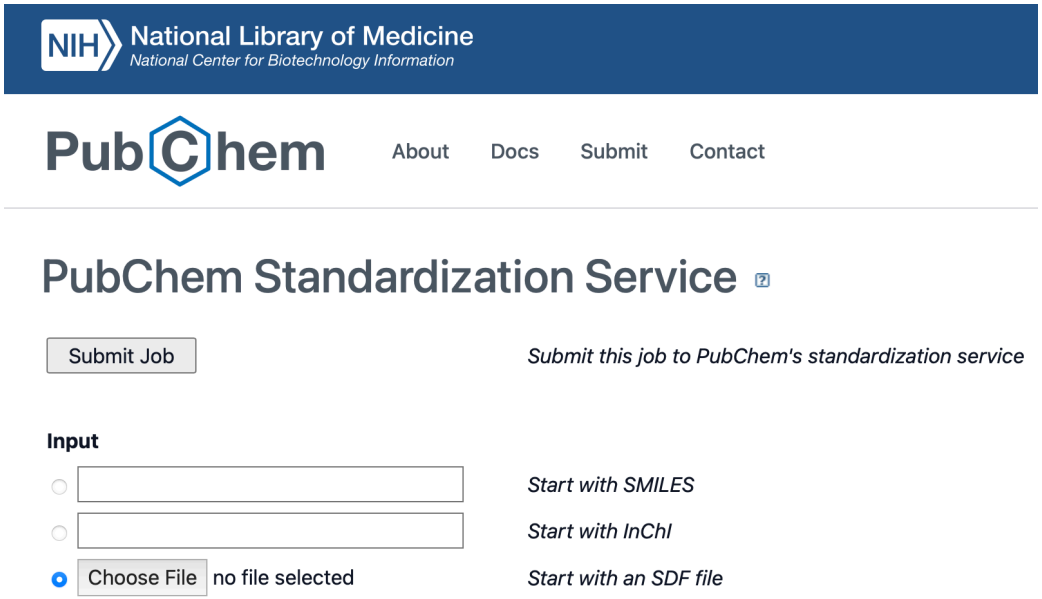
The choice of toolkit may restrict your choices, e.g., with RDKit[1] you may want to go with charge separated hypervalent halogens.



1. RDKit, https://www.rdkit.org/

TOOLS - PUBCHEM

PubChem has a standardisation pipeline[1] based on OpenEye's toolkit[2]. Their public web server only allows user to upload one structure at a time.



1. Hähnke, V.D., Kim, S. & Bolton, E.E. PubChem chemical structure standardization. J Cheminform 10, 36 (2018). https://doi.org/10.1186/s13321-018-0293-8, https://pubchem.ncbi.nlm.nih.gov/standardize/standardize.cgi
2. OpenEye toolkits, https://docs.eyesopen.com/toolkits/python/index.html

TOOLS – LILLYMOL AND SYNGENTA RULES

Corporate tools. Eli Lilly and Co., made some of their tools publicly available, as LillyMol[1]. The tools tend to be quite flexible, we use the following command line:

```
preferred_smiles -g rvnv5 -g guan -g Rguan -g azid -g msdur -g Rn+n- -g imidazole -g pyrazole -g triazole -g tetrazole -g isoxazole -g aminothazole -g pirazolone -g arguan -g ltlt -g ltltr.
```

There is a set of rules known as the "Syngenta rules"[2]. It is currently more limited than LillyMol.

Both LillyMol and Syngenta rules appear to be aiming for look up.

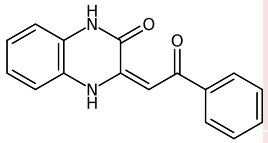
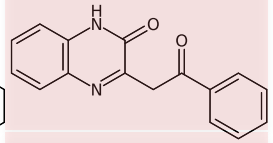
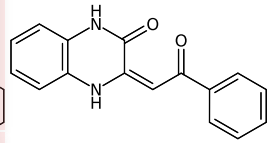
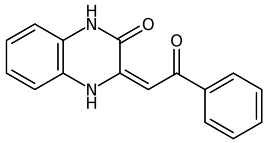
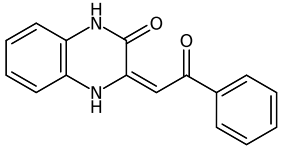
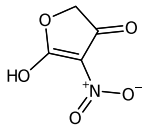
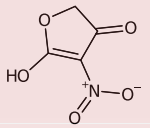
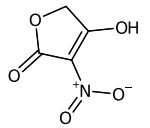
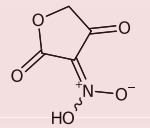
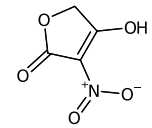
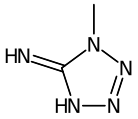
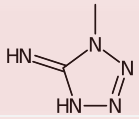
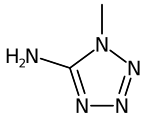
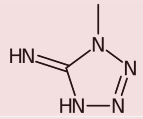
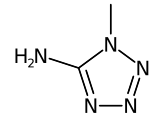
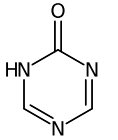
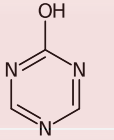
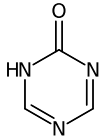
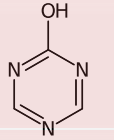
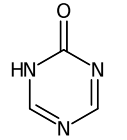
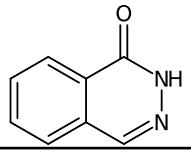
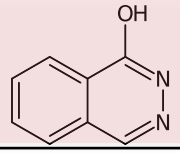
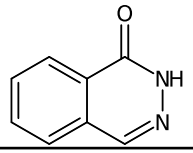
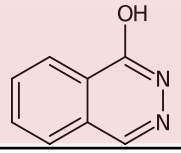
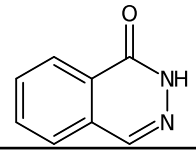
1. LillyMol, https://github.com/EliLillyCo/LillyMol
2. Private communication, Richard Lewis, Novartis

TOOLS - IBM TRANSFORMER 1

A team from IBM used their Transformer deep learning protocol to assign the major tautomer[1]. They have also used Tautobase as their source. They provided 5 examples, unfortunately they appear to assigned which tautomer that is dominant incorrectly in some cases, making comparisons difficult. There is also the issue if a third tautomer is dominant looking at a particular pair may not be relevant.

1. Cretu MT, Toniato A, Thakkar A, Debabeche A, Laino T, Vaucher AC. Standardizing chemical compounds with language models. ChemRxiv. Cambridge: Cambridge Open Engage; 2023; https://doi.org/10.26434/chemrxiv-2022-14ztf-v2

TOOLS – IBM TRANSFORMER 2

Test case	IBM reported truth	Tautobase major tautomer	IBM prediction	RxnFix prediction	Tautobase rule
					728
					353, 354
					1242
					1503
					1005

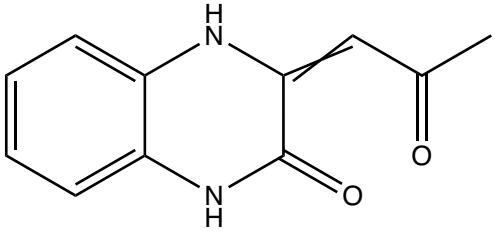
Shaded entries indicate wrongly extracted major tautomer, or wrongly predicted major tautomer. LillyMol converted the last two cases correctly. PubChem matched Tautobase for the last four cases, 728 was turned into a conjugated enol. XTB agrees with Tautobase for 728, 1005 and 1242, but not for 1503. The solvent in 353 and 354 is actually DMSO

TAUTOMATIC

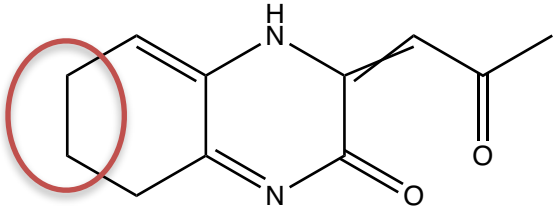
- Tautomatic[1] combines Ambit[2] for tautomer generation and XTB[3] for evaluation
- Recently released
- Can be used as a black box

1. Tautomatic https://github.com/xyzmjf/tautomatic.git
2. Ambit https://pubmed.ncbi.nlm.nih.gov/27481667/
3. XTB https://pubs.acs.org/doi/full/10.1021/acs.jctc.8b01176

AMBIT OBSERVATIONS



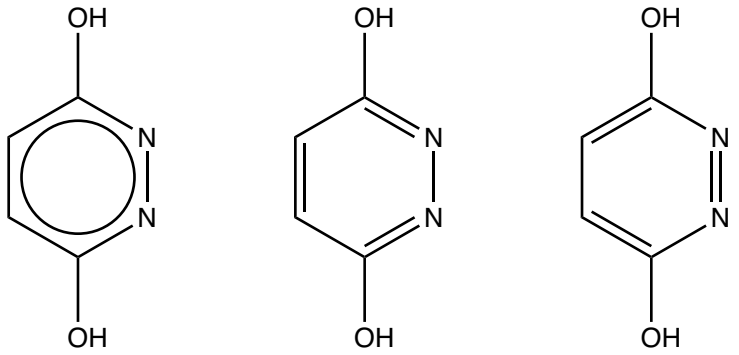
Tautobase 727



1 of 99 “tautomers” Ambit generated with 2 additional hydrogens

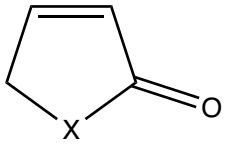
- Ambit may include C-tautomers, sometimes not.
- Some tautomers are less likely, XTB total energies can be up to 100 kcal/mol off.
- We have used multiple seed structures to Ambit and then removed duplicates - this seems to work well.

XTB OBSERVATIONS 1

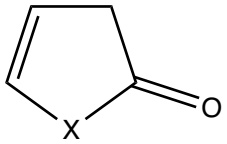


Total energy kcal/mol
−15426.75 −15416.26 −15416.26

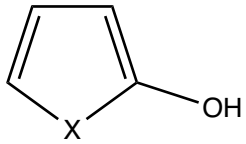
XTB OBSERVATIONS 2



a



b

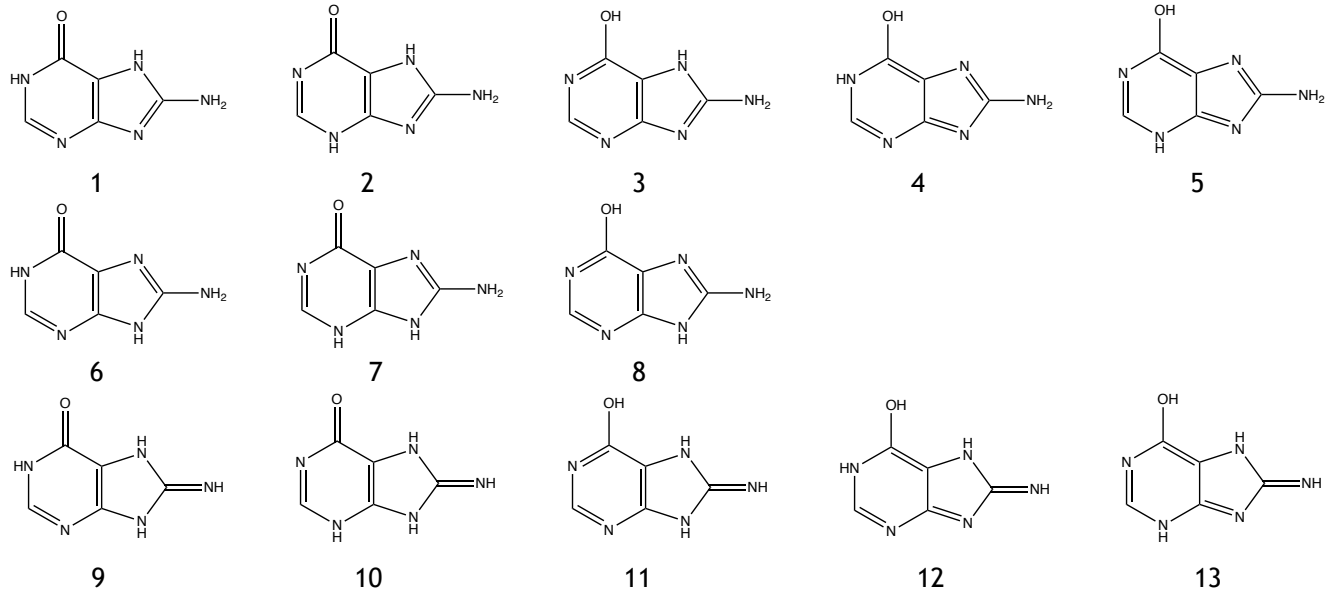


c

X	Taylor & Kenny a->b	XTB a->b	XTB b->c	Tautobase entries
NH	-1.0	2.4	-3.7	1,168,170
O	-2.1	1.0	-1.0	4,8,168,170
S	-2.9	-1.3	-1.4	9,168,170

Taylor and Kenny: a always dominant, c has not been detected in any solvent

8-AMINOHYPOXANTHINE



- RxnFix and PubChem standardises all 13 to 1
- LillyMol standardises 1,2,6,7,9 and 10 to 1; 3,8 and 11 to 3; and leaves 4,5,11 and 12 as is
- Syngenta rules do not standardise this set
- ChemDraw uses 3 for 8-aminohypoxanthine
- XTB assign 1 with the lowest energy by 1.05 kcal/mol

RXNFIX

Rules for RxnFix covers 336 molecules from 390 Tautobase rules. Internal test set contains 1244 tautomers, including cases where only the the kekulization of aromatic rings are different.

Using Tautomatic we generated ~10500 structures, removing duplicates and cases that had different hydrogen counts brought this down to ~2600, removing those C-tautomers that we deemed of less interest (breaking up benzene rings, exocyclic vinyl groups) reduced this to 1067 tautomers. We ran XTB on these.

In 224 out of 336 (67 %) does XTB agree with Taylor and Kenny about which is the major tautomer in water.

RXNFIX 2

The RxnFix rules are written as SMIRKS-like patterns, part of the HazELNut[1] suite of tools. HazELNut can use our in-house toolkit as well as OpenEye's[2] and RDKit's[3] toolkits.

An example:

```
# 1,2,4-TRIAZIN-3-ONE: 2H preferred

[OD1h1+0:1][CD3h0+0:2]1=[ND2h0v3+0:3]
[ND2h0v3+0:4]=[CX3v4+0:5]
[CX3v4+0:6]=[ND2h0v3+0:7]1>>[Oh0:1]=[C:2]1-[Nh1:3]-
[N:4]=[C:5]-[C:6]=[N:7]-1
```

We found that our transforms work better if they are applied to localised structures. This makes it easy to control the atom valency. There is a comment in the PubChem paper they had seen something similar.

1. HazELNut, https://www.nextmovesoftware.com/hazelnut.html
2. OpenEye toolkits, https://docs.eyesopen.com/toolkits/python/index.html
3. RDKit, https://www.rdkit.org/

NAMERXN 101

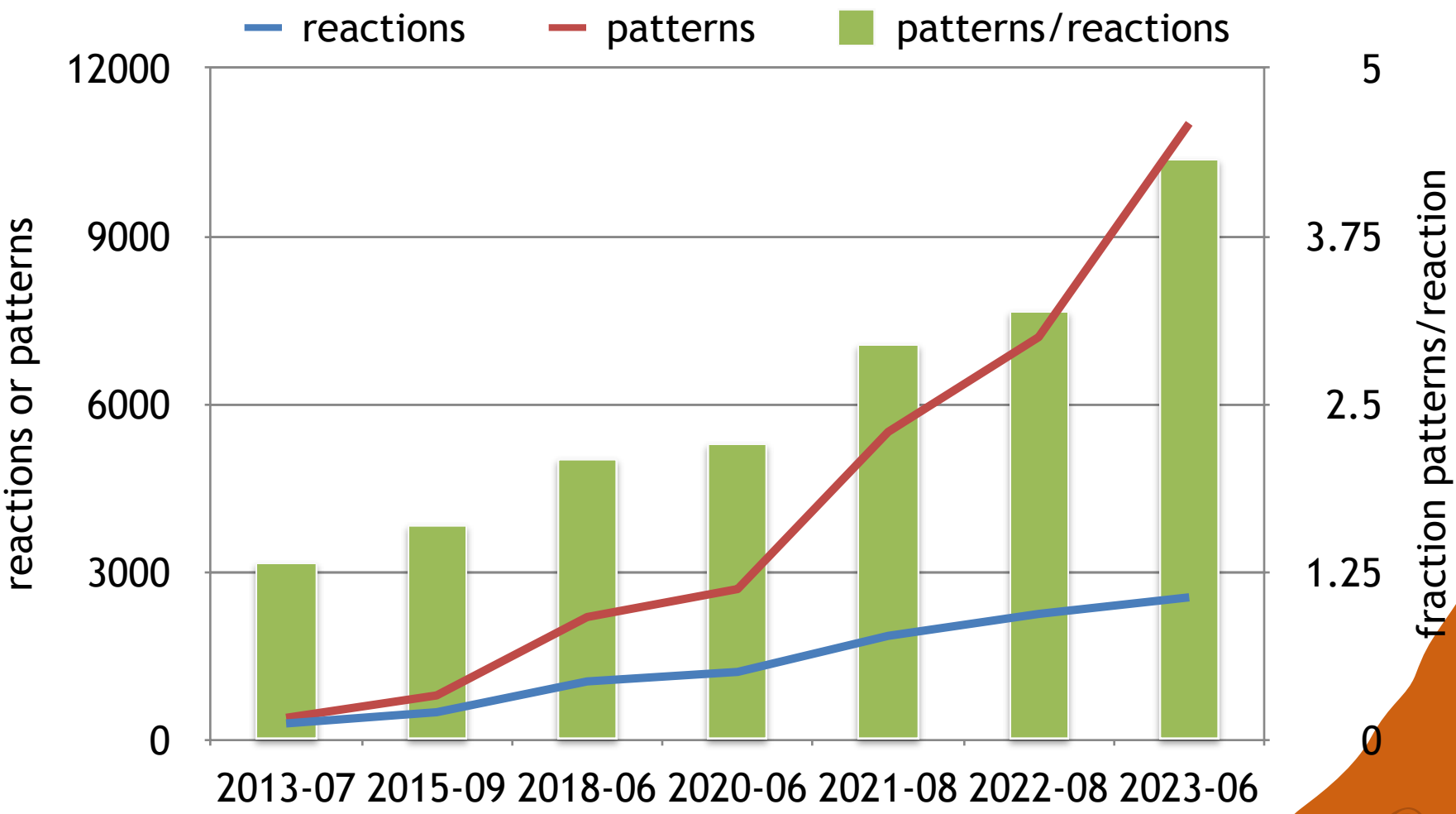
- NameRxn is a mechanism driven classifier
- Much of the heavy lifting is done by the canonicalisation process (i.e., not using MCS)
- Reactions are described in SMIRKS-like patterns
- All reactants and reagents are placed in a single pot and the sets of patterns are applied in turn.
- If a desired product is generated, the reaction (its mechanism and atom mapping) has been identified.
- `[#16D2v2h0+0:1].[OD1h1+0:2][OD2v2h0+0:3]>>[#16:1]=[Oh0:2].[Oh1:3]`
  `SULFANYL_TO_SULFINYL`

CLASSIFICATION

- Reactions are classified into a common subset of the Carey et al.[1] classes
- There are 12 super-classes
  – e.g. 3 C-C bond formation (RXNO:0000002)
- These contain 72 classes/categories.
  – e.g. 3.5 Pd-catalyzed C-C bond formation (RXNO:0000316)
- The classes contains > 2500 named reactions
- Requiring > 11000 SMIRKS patterns
  - Tautomers, keto-enol, lactim-lactam, imidazoles, 1,2,4-triazoles, amidines and guanidines
  - Alternative representations for groups such as nitro and azides ([N+](=O)[O-] vs N(=O)=O)

1. John S. Carey, David Laffan, Colin Thomson, Analysis of the reactions used for the preparation of drug candidate molecules, Org. Biomol. Chem., 2006, 4, 2337–2347

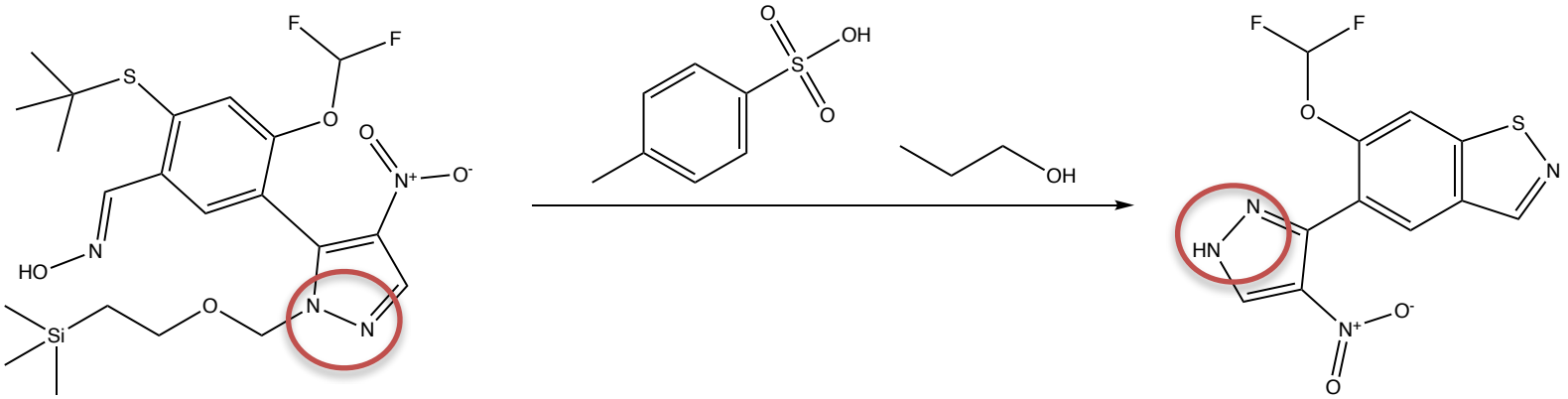
NAMERXN EVOLUTION



NAMERXN – STANDARDISATION MODE

- Added flag -phys that expects standardised reactions
- Skip ~300 patterns for alternative nitro and azide representations
- Work in progress to skip patterns for guanidine, triazoles, halopyridine to pyridone reactions
- Helps with selecting order of transforms for secondary deprotections, e.g., N-THP/N-SEM deprotection of pyrazoles

NAMERXN – STANDARDISATION MODE



Now classified as: 4.3.7;6.1.18 Isothiazole synthesis; N-SEM deprotection [S-containing heterocycle formation; NH deprotections]

CONCLUSIONS

- Is a single structure enough or do you need an ensemble
- What is the use case, lookup or representing some aspect of chemistry
- Carefully analyse all your data input
- By standardising tautomers we can ignore some reaction transforms and gain a slight improvement in reaction classification

ACKNOWLEDGEMENTS

- Thank you for listening
- NextMove Software team
  - John Mayfield
  - Rachel Pirie
  - Roger Sayle
- Alumni
  - Daniel Lowe
  - Noel O'Boyle
  - Richard Gowers
- Richard Lewis (Syngenta rules)

- Our customers
  - in particular those who provide feedback
- We are always looking for talented people to join NextMove Software
- https://www.nextmovesoftware.com/talks/ACS_2023_Fall_Lagerstedt_Standardisation.pdf