Molecular standardization

June 1, 2016

The molecular standardization processes of ChemAxon's Standardizer and the Open Source library MolVS have been evaluated on a test suit of 37 compounds containing functional groups that are often represented differently. The 37 compounds are represented in multiple forms that should ideally be transformed into the same representation by the standardization process.

For each method a table is produced showing the input and output structures as drawn by RDKit for the different forms. The table also displays two fingerprint similarities between all pairs of output structures. If the standardization is successful the fingerprints should be identical. Furthermore, the set of 177 RDKit descriptors is calculated for each output structure and the number of differing descriptors is reported for each pair of output structures.

1 Results Summary

ChemAxon's Standardizer fails to identify the different forms of 14 out of the 37 structures as similar, while MolVS fails for 3 structures. The failures of MolVS are all for certain functional groups, nitro, phosphine and diazo, while Standardizer fails in the canonicalization of 9 tautomers, as well as for 5 functional groups.

As shown in the tables below, most of the standardization failures result in significant loss of fingerprint similarity and a large number of RDKit descriptors with different values. Hence, the failures will have an impact on the results of chemoinformatic analysis.

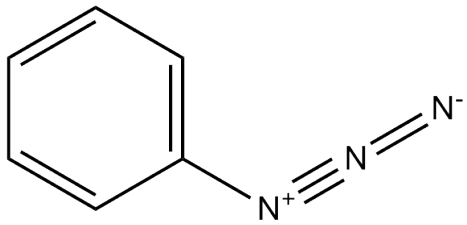
2 Standardization with ChemAxon Standardizer

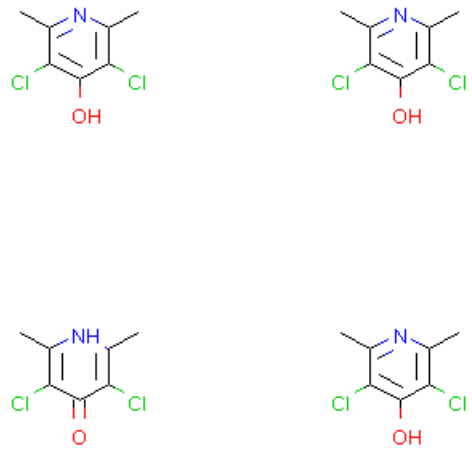
The ChemAxon standardizer was used with 7 actions, applied in the following order, together with the set of transformations defined in the GUI:

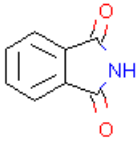
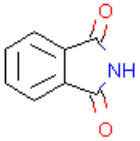
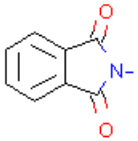
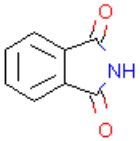
- removesolvents
- stripsalts

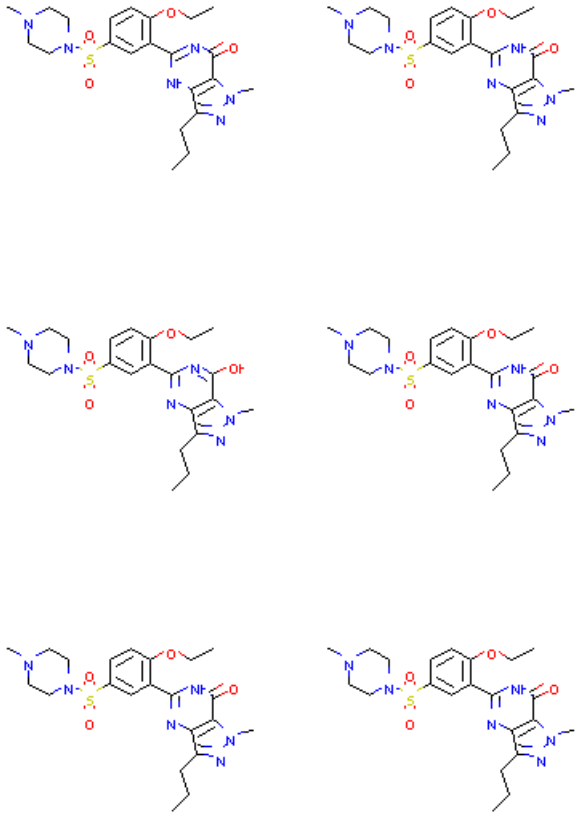
- removefragment:method=keeplargest
- neutralize
- aromatize
- tautomerize
- mesomerize

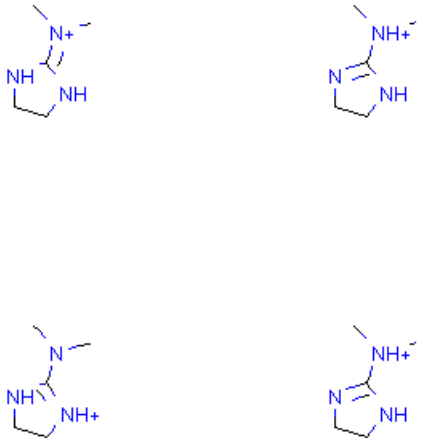
Two out of the failures to transform functional groups (Idx 7 and 35) originates from RDKit being unable to accept the valence of the structures returned by Standardizer. The output structure of molecule 7 is displayed below.

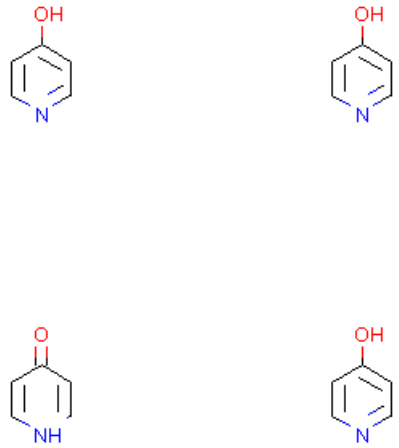


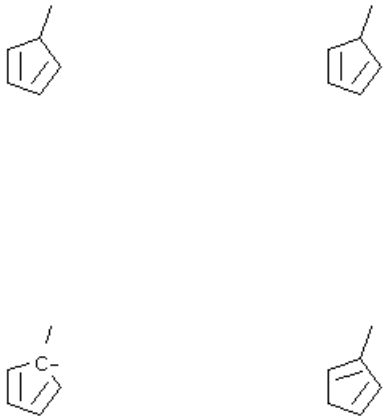
Idx	Name	Structure In and Out	Comment	TanTop	DiceMorgan	No. of DescDiff
1	clopidol			[1.0]	[1.0]	[0]

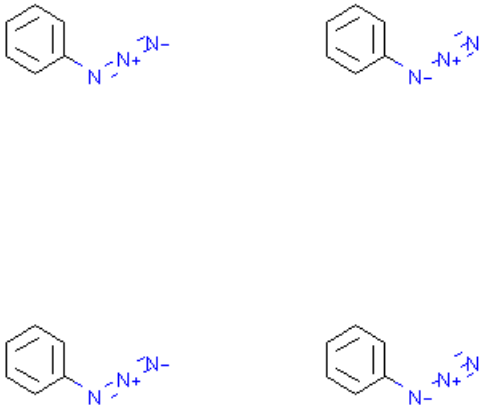
2	phthalimide	   		[1.0]	[1.0]	[0]

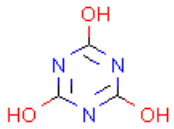
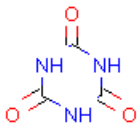
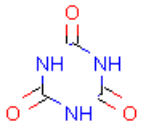
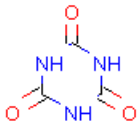
3	Viagra			[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

4	quatAmMesomer			[1.0]	[1.0]	[0]

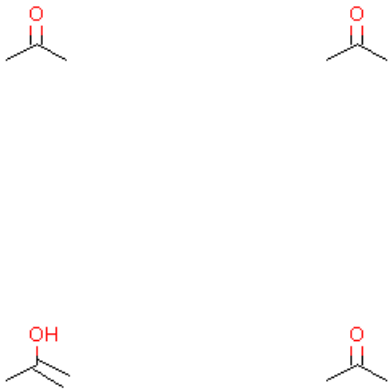
5	pyridinol			[1.0]	[1.0]	[0]

6	aromaticAnion			[0.68]	[0.24]	[36]

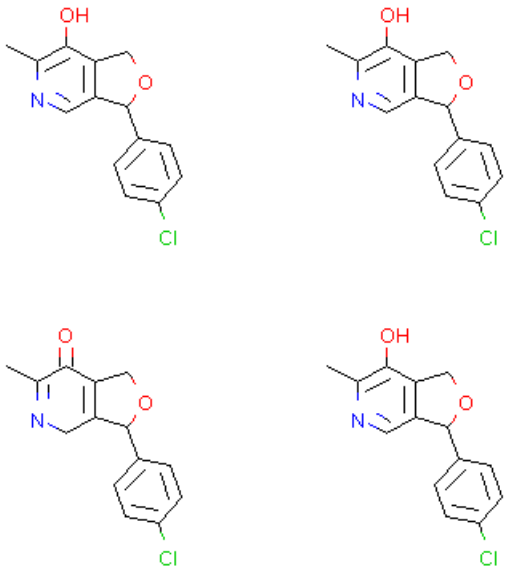
7	azide			[1.0]	[1.0]	[0]

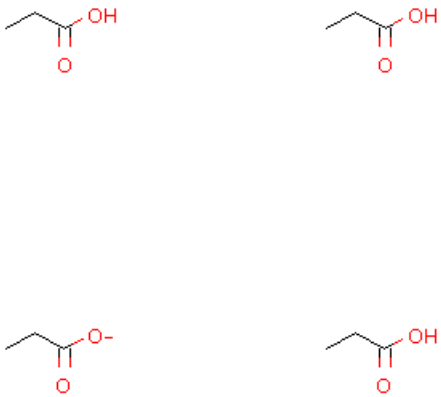
8	triazinol	   		[1.0]	[1.0]	[0]

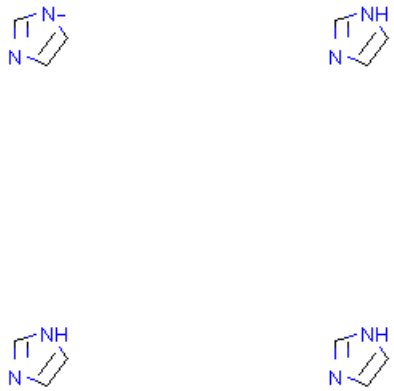
9	phosphate	   		[1.0]	[1.0]	[0]

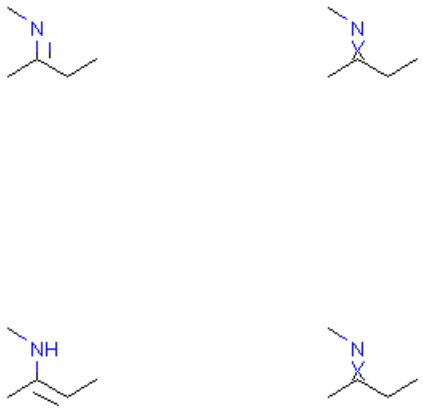
10	Propenol			[1.0]	[1.0]	[0]

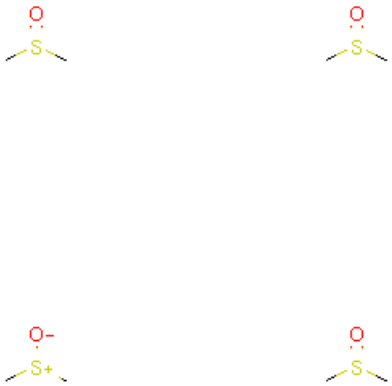
11	phosphine			[1.0]	[1.0]	[0]

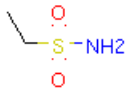
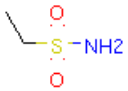
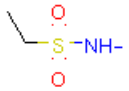
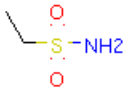
12	cicletanine			[1.0]	[1.0]	[0]

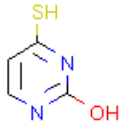
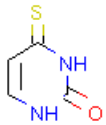
13	propaneAcid			[1.0]	[1.0]	[0]

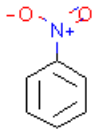
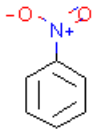
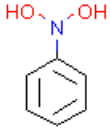
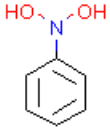
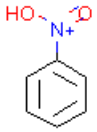
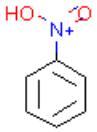
14	imidazole			[1.0]	[1.0]	[0]

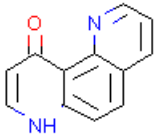
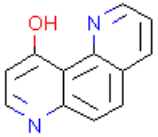
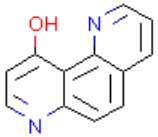
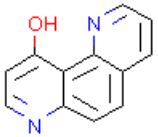
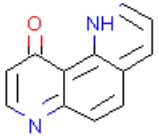
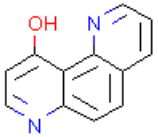
15	enamine			[1.0]	[1.0]	[0]

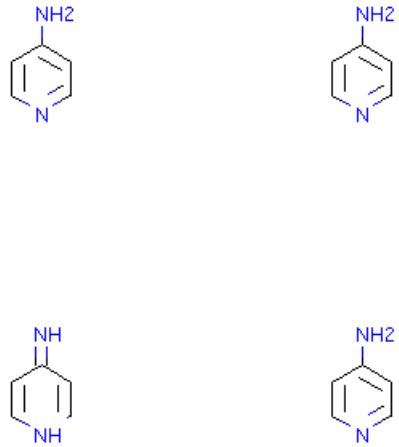
16	sulfoxide			[1.0]	[1.0]	[0]

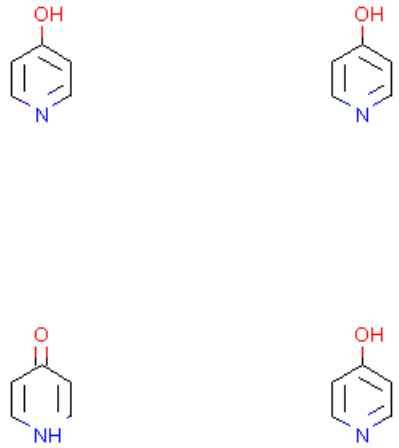
17	sulfonamide	   		[1.0]	[1.0]	[0]

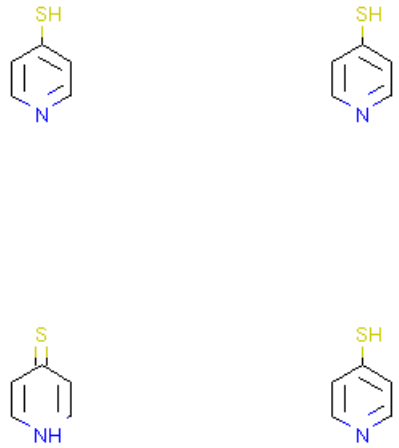
18	tiouracil	   		[1.0]	[1.0]	[0]

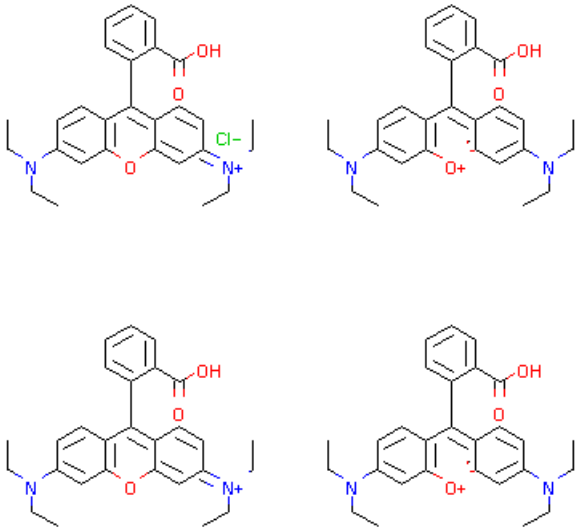
19	nitro	     		[0.67, 1.0, 0.67]	[0.56, 0.8, 0.6]	[58, 43, 56]

20	107and108and109	     		[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

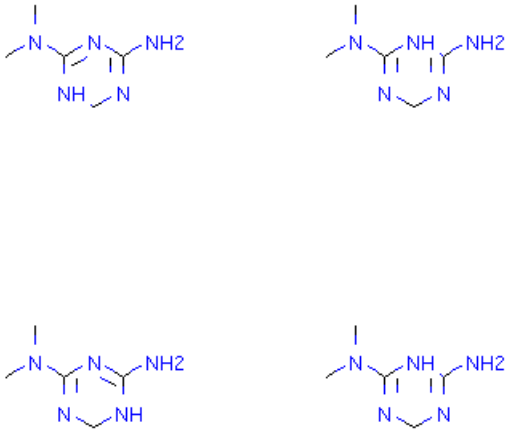
21	aminopyrimidine			[1.0]	[1.0]	[0]

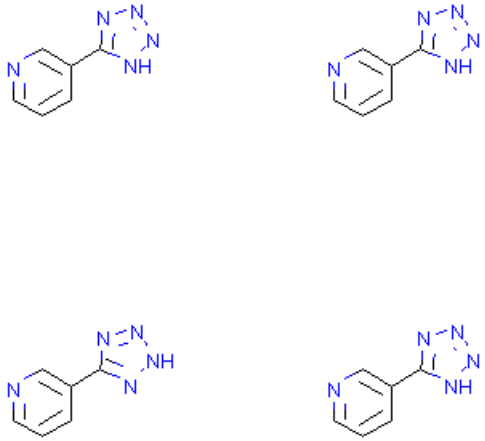
22	103and104			[1.0]	[1.0]	[0]

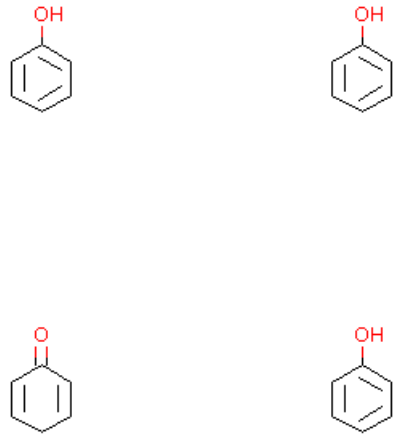
23	105and106			[1.0]	[1.0]	[0]

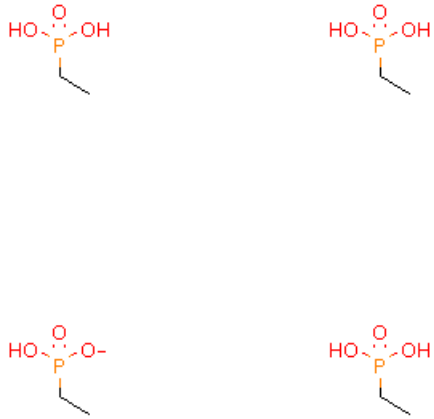
24	Rhodamine			[1.0]	[1.0]	[0]

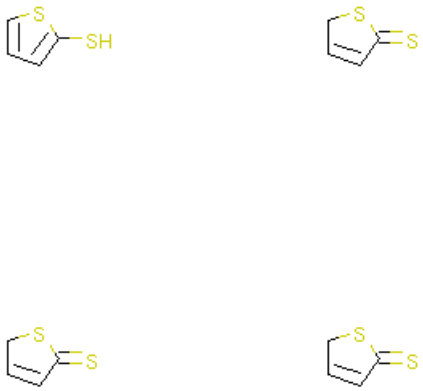
25	sulfon	   		[1.0]	[1.0]	[0]

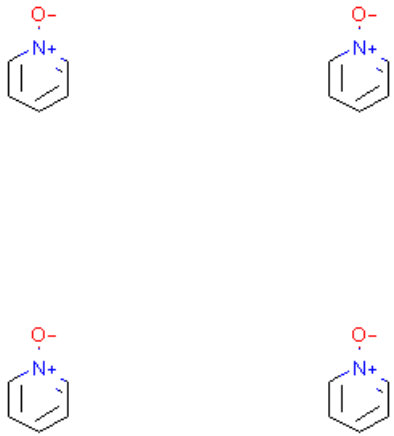
26	triazine			[1.0]	[1.0]	[0]

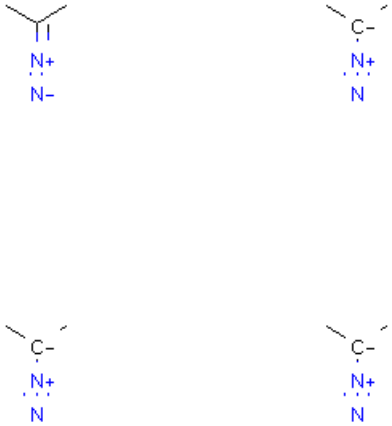
27	tetrazole			[1.0]	[1.0]	[0]

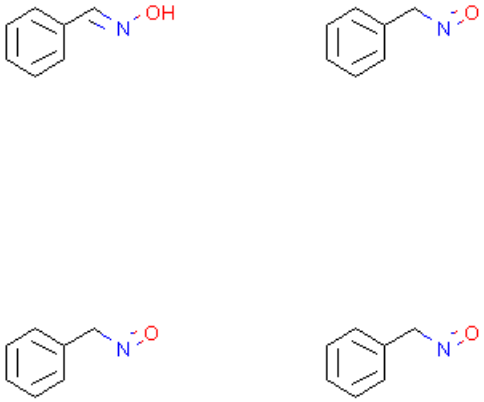
28	phenol			[1.0]	[1.0]	[0]

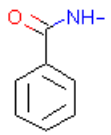
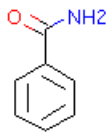
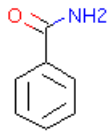
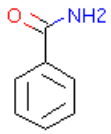
29	phosphor			[1.0]	[1.0]	[0]

30	112and113			[1.0]	[1.0]	[0]

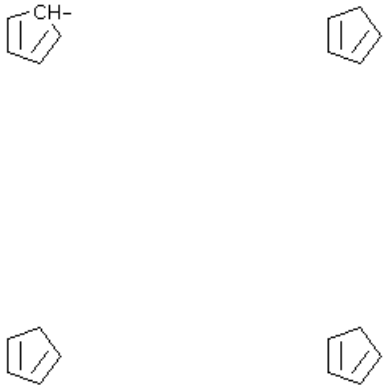
31	NO			[1.0]	[1.0]	[0]

32	diazo			[1.0]	[1.0]	[0]

33	benzaldoxime			[1.0]	[1.0]	[0]

34	amide	   		[1.0]	[1.0]	[0]

35	thiopurine	   		[1.0]	[1.0]	[0]

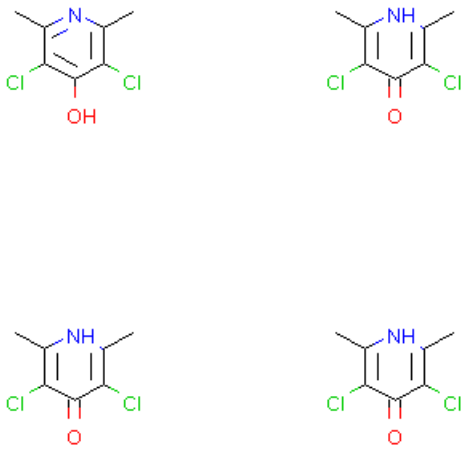
36	cyclopentadieneAnion			[1.0]	[1.0]	[0]

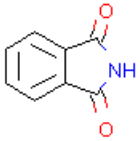
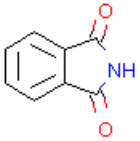
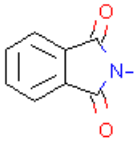
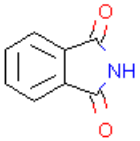
3 Standardization with MolVS

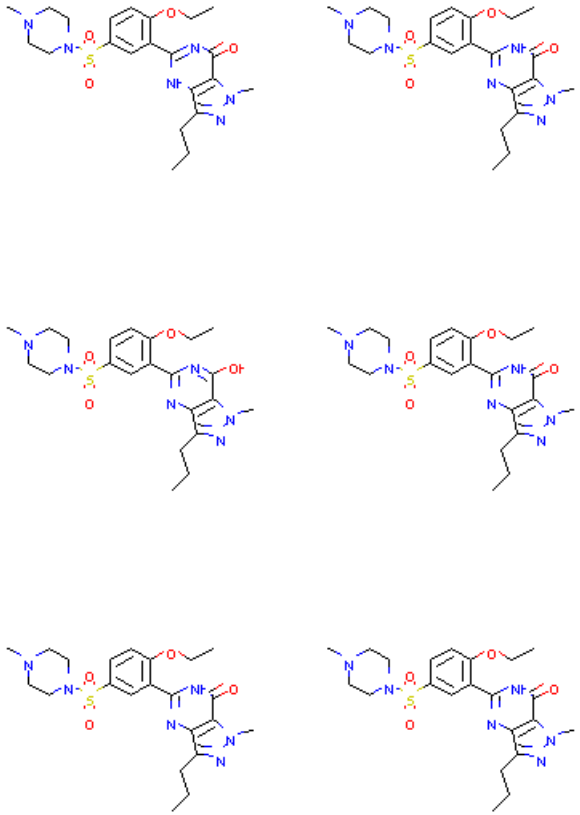
The following 4 classes of MolVS were used:

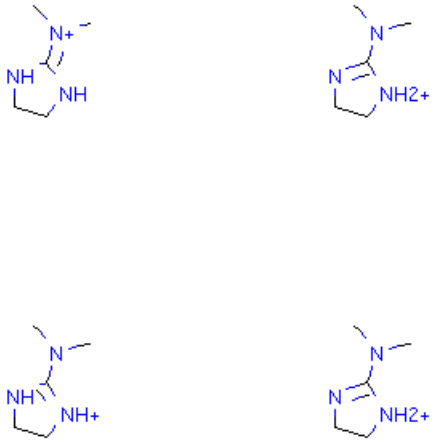
- Uncharger
- LargestFragmentChooser
- Standardizer
- Normalizer
- TautomerCanonicalizer

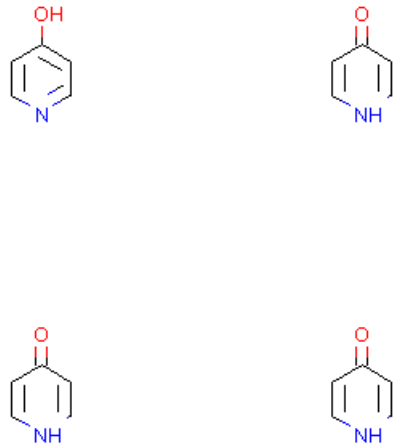
In addition to standardization, MolVS can also enumerate tautomers.

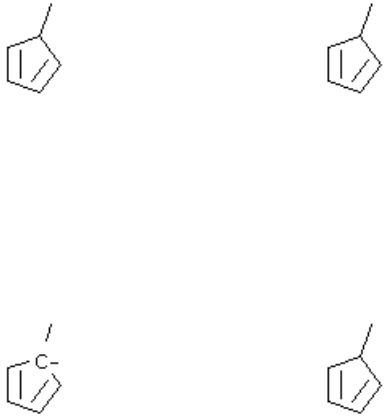
Idx	Name	Structure In and Out	Comment	TanTop	DiceMorgan	No. of DescDiff
1	clopidol			[1.0]	[1.0]	[0]

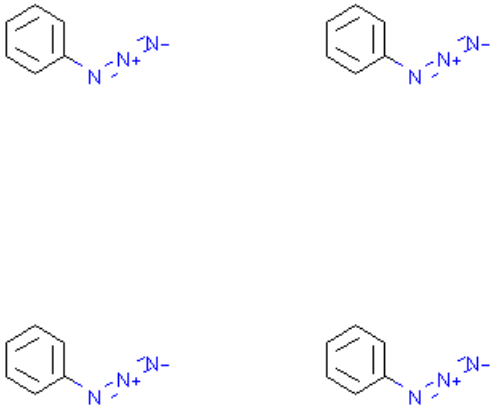
2	phthalimide	   		[1.0]	[1.0]	[0]

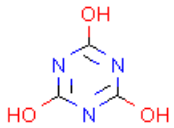
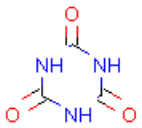
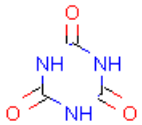
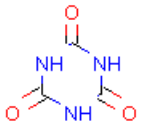
3	Viagra			[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

4	quatAmMesomer			[1.0]	[1.0]	[10]

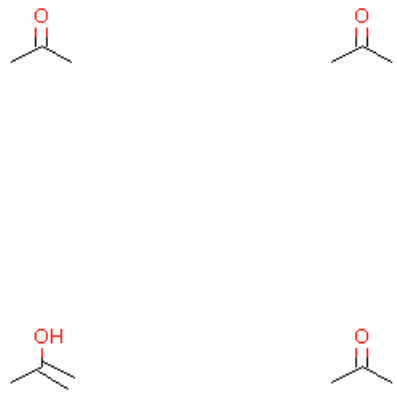
5	pyridinol			[1.0]	[1.0]	[0]

6	aromaticAnion			[1.0]	[1.0]	[0]

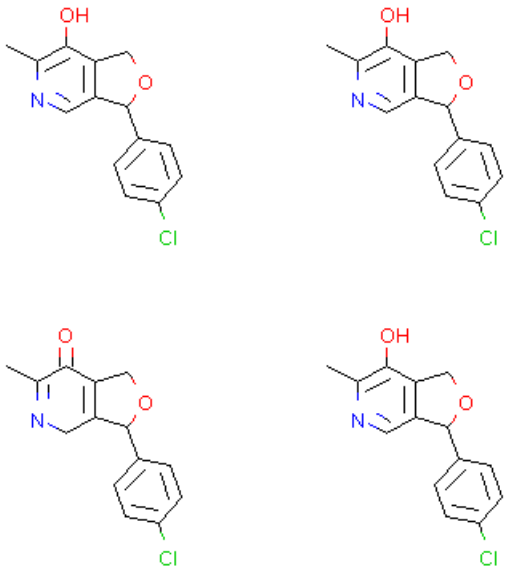
7	azide			[1.0]	[1.0]	[0]

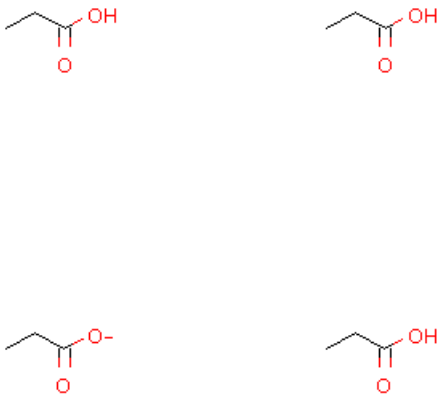
8	triazinol	   		[1.0]	[1.0]	[0]

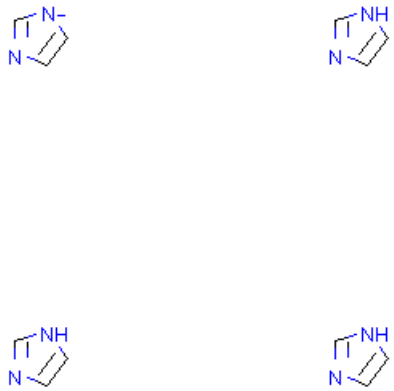
9	phosphate	   		[1.0]	[1.0]	[0]

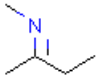
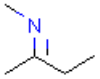
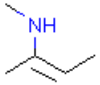
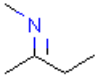
10	Propenol			[1.0]	[1.0]	[0]

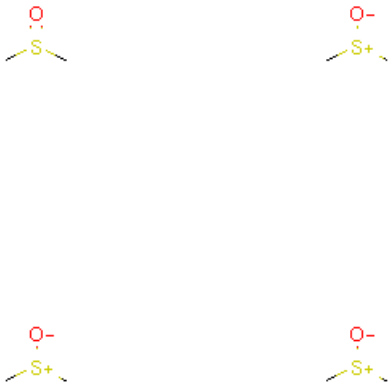
11	phosphine			[1.0]	[0.5]	[26]

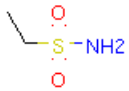
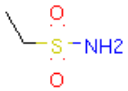
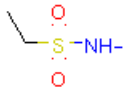
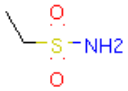
12	cicletanine			[1.0]	[1.0]	[0]

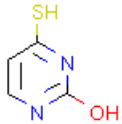
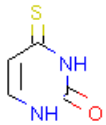
13	propaneAcid			[1.0]	[1.0]	[0]

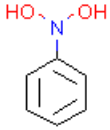
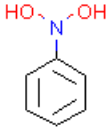
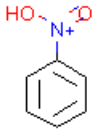
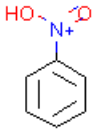
14	imidazole			[1.0]	[1.0]	[0]

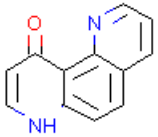
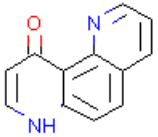
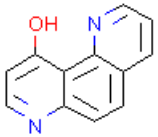
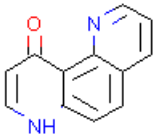
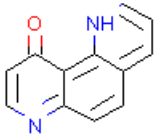
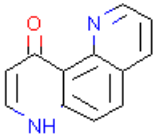
15	enamine	   		[1.0]	[1.0]	[0]

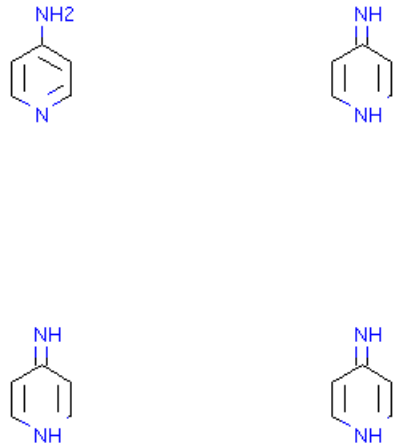
16	sulfoxide			[1.0]	[1.0]	[6]

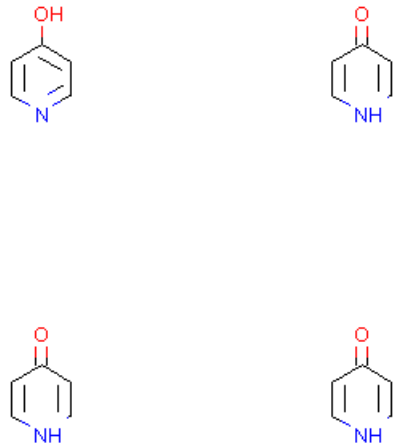
17	sulfonamide	   		[1.0]	[1.0]	[0]

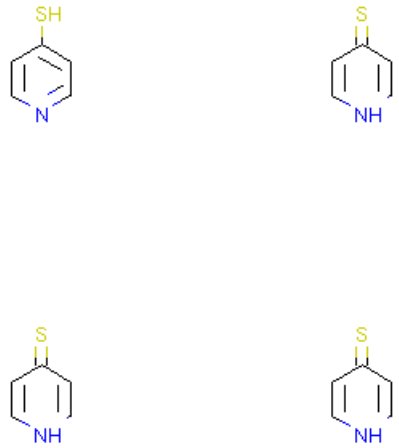
18	tiouracil	   		[1.0]	[1.0]	[0]

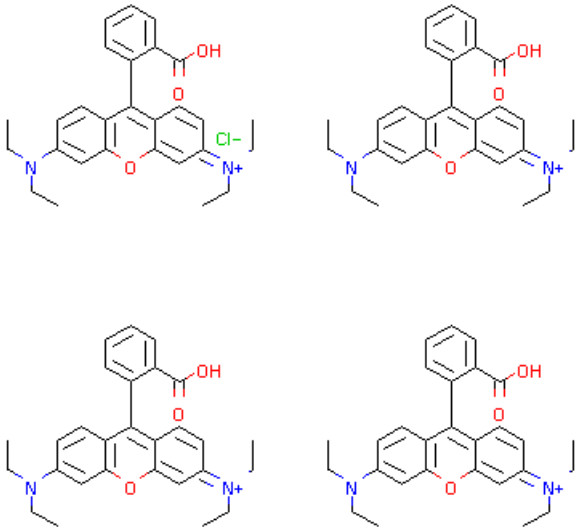
19	nitro	    		[0.67, 1.0, 0.67]	[0.56, 0.8, 0.6]	[58, 43, 56]

20	107and108and109	     		[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

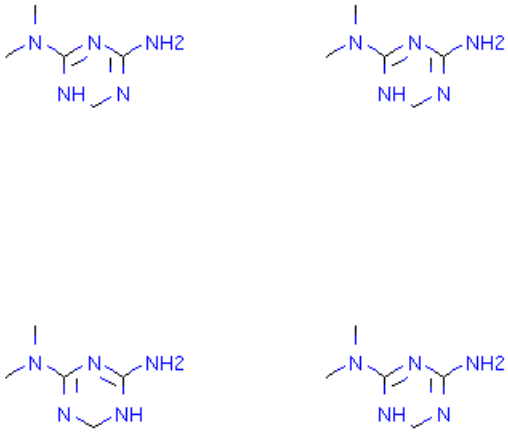
21	aminopyrimidine			[1.0]	[1.0]	[0]

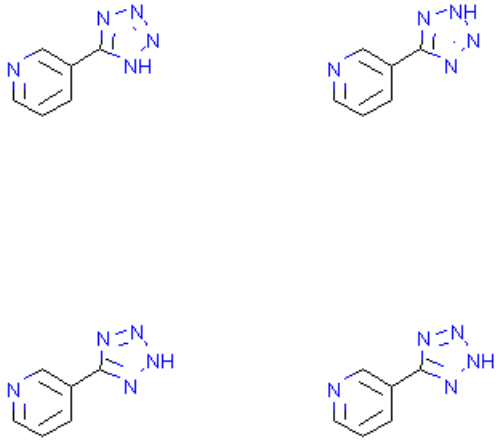
22	103and104			[1.0]	[1.0]	[0]

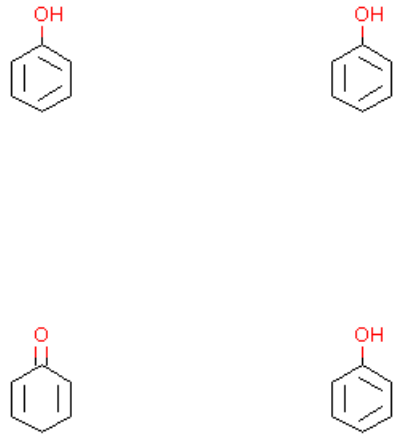
23	105and106			[1.0]	[1.0]	[0]

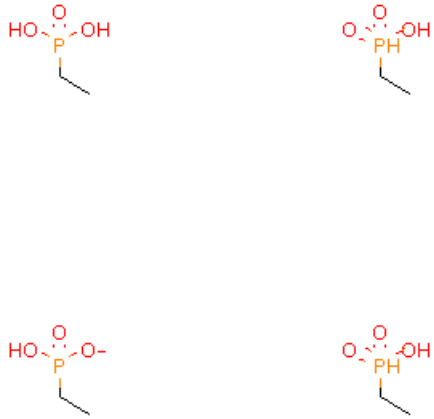
24	Rhodamine			[1.0]	[1.0]	[0]

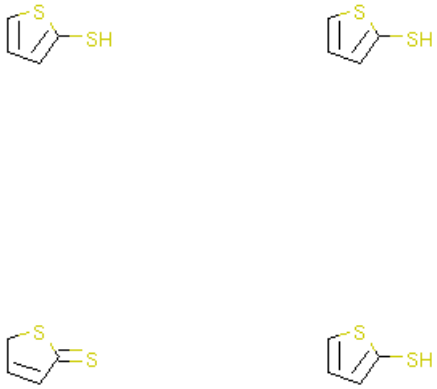
25	sulfon	   		[1.0]	[1.0]	[0]

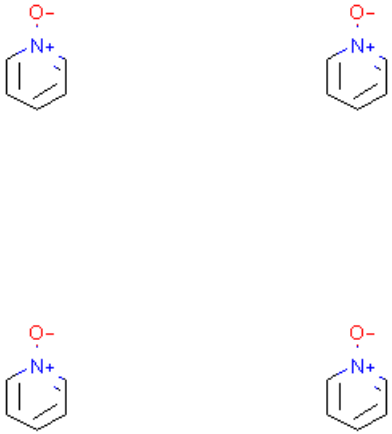
26	triazine			[1.0]	[1.0]	[0]

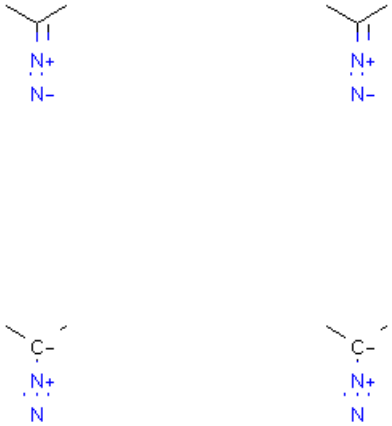
27	tetrazole			[1.0]	[1.0]	[1]

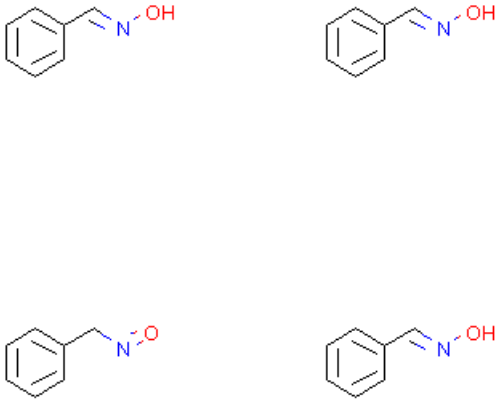
28	phenol			[1.0]	[1.0]	[0]

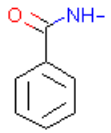
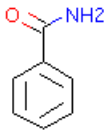
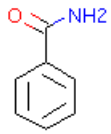
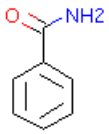
29	phosphor			[1.0]	[1.0]	[0]

30	112and113			[1.0]	[1.0]	[0]

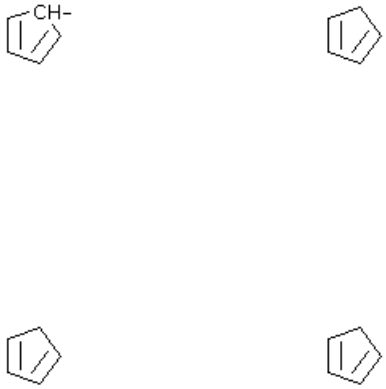
31	NO			[1.0]	[1.0]	[0]

32	diazo			[0.24]	[0.27]	[34]

33	benzaldoxime			[1.0]	[1.0]	[0]

34	amide	   		[1.0]	[1.0]	[0]

35	thiopurine	   		[1.0]	[1.0]	[0]

36	cyclopentadieneAnion			[1.0]	[1.0]	[0]