## 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.

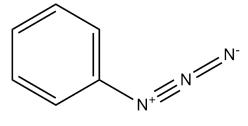
## 2 Standardization with ChemAxon Standardizer

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

- $\bullet \ \ remove fragment: method = keep largest$
- $\bullet$  neutralize
- $\bullet$  aromatize
- tautomerize
- $\bullet$  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
		CI OH CI OH				
1	clopidol	CI CI OH		[1.0]	[1.0]	[0]

		NH	NH			
2	phthalimide	N- O	NH	[1.0]	[1.0]	[0]

		N N S N N N N N	N N S N N N N N N N N N N N N N N N N N			
		N O O O O O O O O O O O O O O O O O O O	N N N N N N N N N N N N N N N N N N N			
		N N N N N N N N N N N N N N N N N N N	N N S N N N N N N N N N N N N N N N N N			
3	Viagra			[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

**~**1

$\infty$	

6	aromaticAnion	/ □∕>		[0.68]	[0.24]	[36]

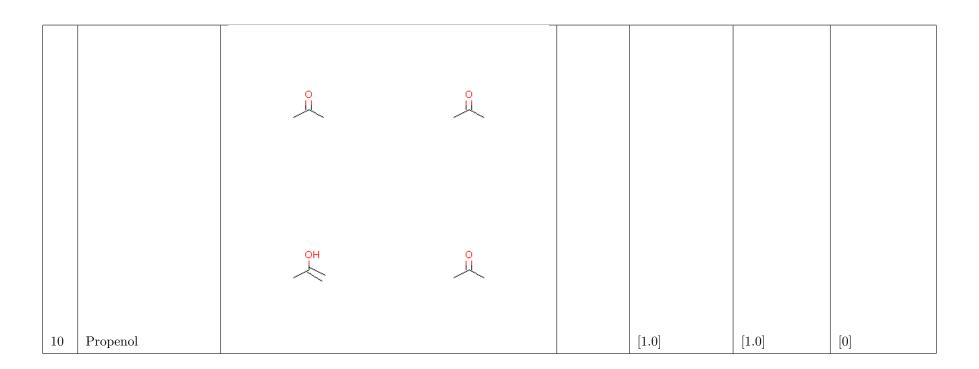
		N-N+N-	N-N+,N			
7	azide	N-N+N-	N-N+*N	[1.0]	[1.0]	[0]

		HO N OH	NH NH			
8	triazinol	NH NH NH	NH NH	[1.0]	[1.0]	[0]

ŀ	4
	,
ł	_

		-02P20 -0-0-	но <sub>Р</sub> О но он			
9 ph	nosphate	но <sub>р</sub> , о но он	но <sub>ЭР</sub> , О но он	[1.0]	[1.0]	[0]

$\vdash$	
2	



		<b>∕</b> PH3+	PH2			
13	phosphine	PH2	PH2	[1.0]	[1.0]	[0]

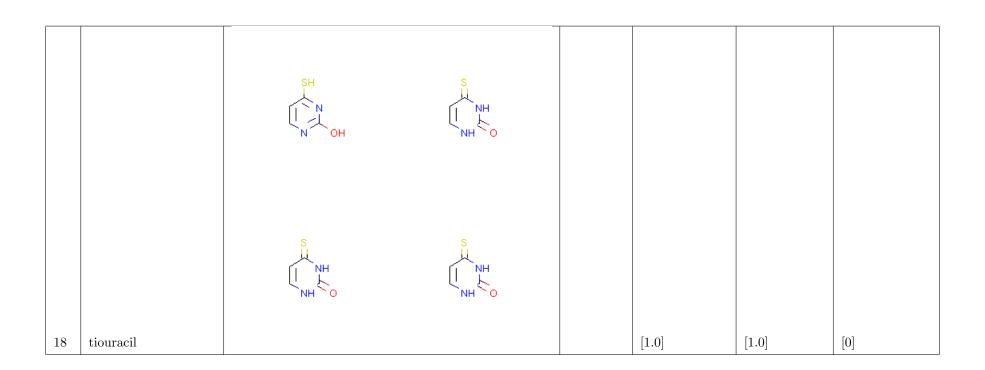
_		4	
`~	7		
_		•	

		OH O	ОН			
13	propaneAcid	0- 0	ОН	[1.0]	[1.0]	[0]

		Z Z Z	N NH			
14	imidazole	N NH	N NH	[1.0]	[1.0]	[0]

		O: \$	O:				
16	sulfoxide	<b>ọ-</b> ∕S+,	Ö. .s.	[1.	0]	[1.0]	[0]

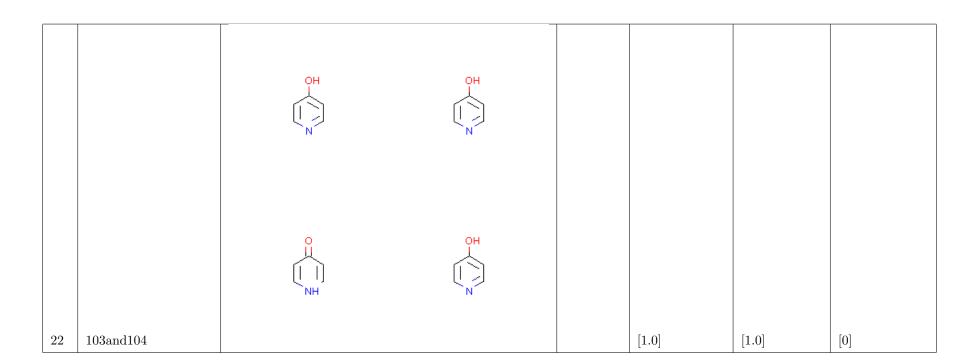
		0 <u>s</u> -NH2 0	O. S. NH2 O			
17	sulfonamide	0. snh- 0	0. 5. NH2 0	[1.0]	[1.0]	[0]



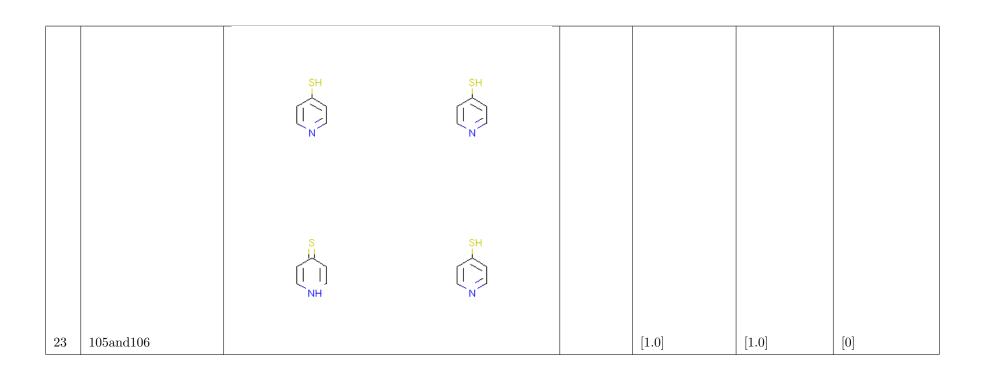
		-0 N+	-0 N+			
		HONNOH	HONNOH			
19	nitro	HO N+	HO N+	[0.67, 1.0, 0.67]	[0.56, 0.8, 0.6]	[58, 43, 56]

		NH NH	OH N			
		OH N	OH N			
20	107and108and109	O NH	OH N	[1.0, 1.0, 1.	0] [1.0, 1.0, 1.0]	[0, 0, 0]

N	೨	
1	$\sim$	



ľ			
ž		-	
u		(	
3	_		



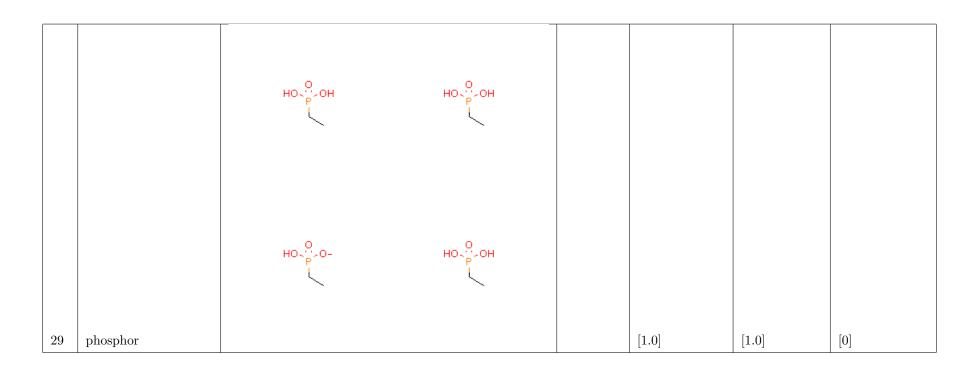
		О 3 5 ° О НО 3 ° О	О.S.5.О НО 3				
25	sulfon	0350 -035	OSS O HOSS	[1.0	)]	[1.0]	[0]

		N NH2 NH N	   NH NH2   N N			
26	triazine	N NH2	N NH NH2	[1.0]	[1.0]	[0]

		N N N N N N N N N N N N N N N N N N N	N N N NH				
27	tetrazole	N N N N N N N N N N N N N N N N N N N	N NH NH	[1.0]	[1.0]	[0]	

•	
	v
	_
_	_

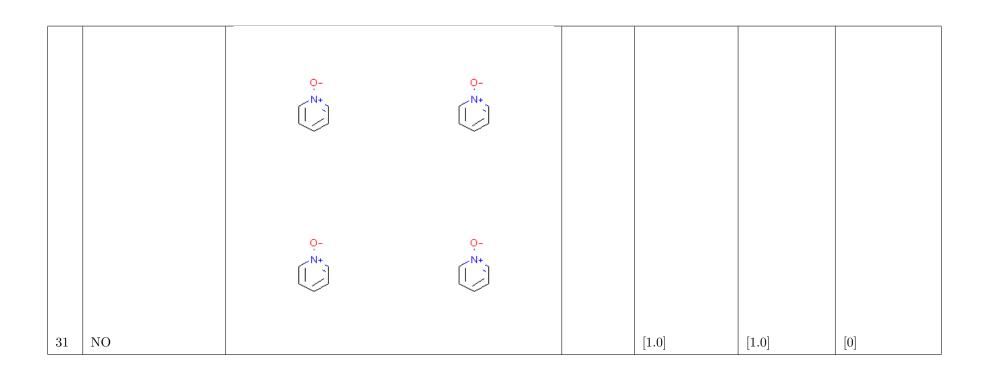
		OH L	OH			
28	phenol		OH	[1.0]	[1.0]	[0]



32	

		SH SH	S S S			
30	112and113	S ►s	S ►s	[1.0]	[1.0]	[0]

•	J	
	٠	
	·	

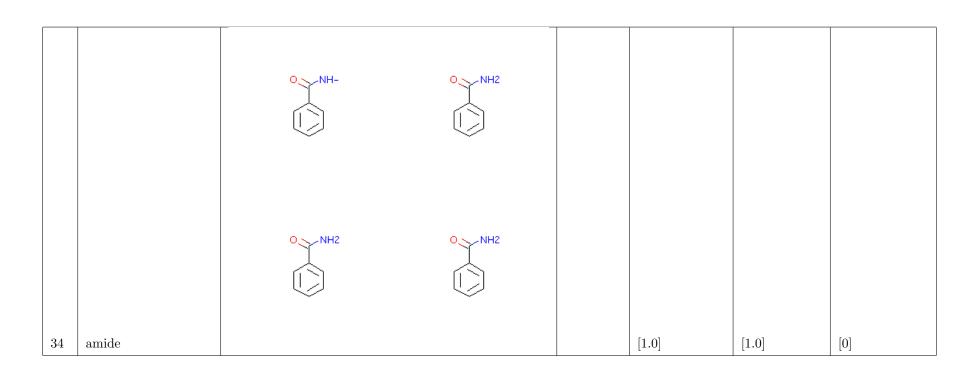


			0- N+: N			
32	diazo	, c. »+ . »	O- N+ N+ N	[1.0]	[1.0]	[0]

C	
C	л

		N OH	N <sup>-</sup> -O			
33	benzaldoxime	N.O	N <sup>-</sup> O	[1.0]	[1.0]	[0]

•		•	١	
3	_	•	•	
	0	ī	١	



## 3 Standardization with MolVS

The following 4 classes of MolVS were used:

- $\bullet$  Uncharger
- $\bullet \ Largest Fragment Chooser$
- $\bullet$  Standardizer
- Normalizer
- TautomerCanonicalizer

In addition to standardization, MolVS can also enumerate tautomers.

	$\sim$
۱	=
1	$\neg$
•	$\overline{}$

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

		NH	NH			
2	phthalimide	N- O	NH	[1.0]	[1.0]	[0]

		N N N N N N N N N N N N N N N N N N N	N N N N N N N N N N N N N N N N N N N			
		N OF N OF N	N N N N N N N N N N N N N N N N N N N			
		N N N N N N N N N N N N N N N N N N N	N N N N N N N N N N N N N N N N N N N			
3	Viagra			[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

. N	
<del>- 11-</del>	
ಌ	

		NH NH	N NH2+				
4	${ m quatAmMesomer}$	NHX NH+	N NH2+	[1.	.0]	[1.0]	[10]

42	
n-	

		OH N	O NH			
5	pyridinol	O NH	NH NH	[1.0]	[1.0]	[0]

	١,	
н		`
0		7
•		

6	${ m aromatic Anion}$	/ [_/>		[1.0]	[1.0]	[0]

	r	ς.	
1	ł	_	
,	•		

		N-N+N-	N-N+ <sup>N</sup> -			
7	azide	N N N TN-	N N+ N-	[1.0]	[1.0]	[0]

		HO N OH	NH NH O			
8	triazinol	NH NH	NH NH	[1.0]	[1.0]	[0]

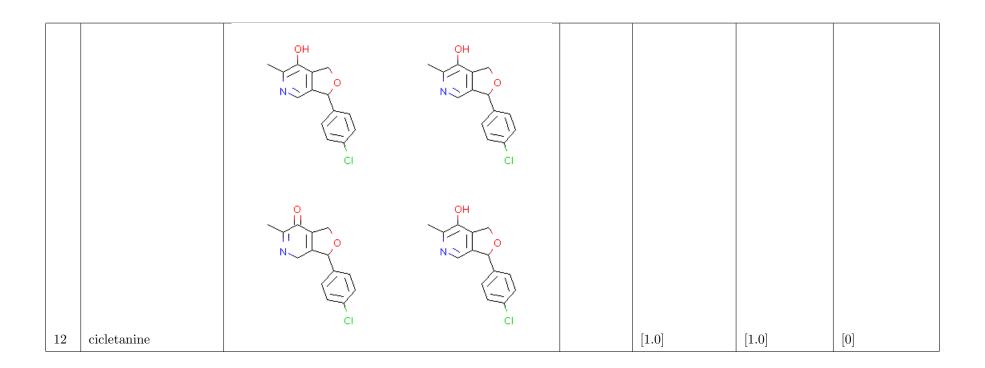
L	$\wedge$
5	
v	~

		-0>P>0 -0'-	о он но он о у ън			
9	phosphate	HO P OH	O S PH HO OH	[1.0]	[1.0]	[0]

		°	°			
10	Propenol	OH \_	<u> </u>	[1.0]	[1.0]	[0]

C	5	τ
		)

		<b>∕</b> PH3+	<b>∕</b> PH3+			
11	phosphine	PH2	PH2	[1.0]	[0.5]	[26]



ರಾ	
Ö	

		OH O	ОН			
13	m propane Acid	<b>~</b> <del>  0</del> <del>-</del> 0 <del></del>	<b>ОН</b> ОН	[1.0]	[1.0]	[0]

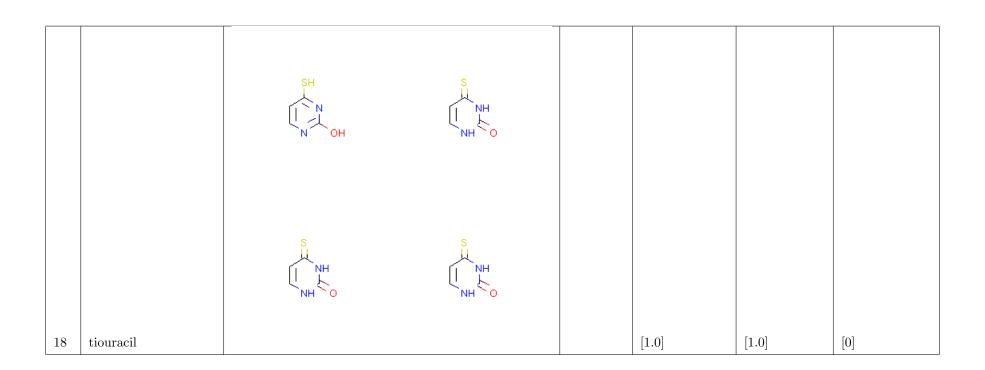
		Z Z Z	N NH			
14	imidazole	N NH	N NH	[1.0]	[1.0]	[0]

r		•		
			l	
	ř			

		N -	N I			
15	enamine	NH	N	[1.0]	[1.0]	[0]

		0: s	Q- _S+_,			
16	sulfoxide	о- _s+_	О- _S+ 、	[1.0]	[1.0]	[6]

		O. SNH2 O	0. S-NH2 0			
17	sulfonamide	0. 5. NH- 0	0. 5. NH2 0	[1.0]	[1.0]	[0]



		-0 N+	-0 N+D			
		HO NOH	HONNOH			
19	nitro	HO N+	HO N+	[0.67, 1.0, 0.67]	[0.56, 0.8, 0.6]	[58, 43, 56]

		NH	O N			
		OH N	NH NH			
20 1078	and108and109	NH NH	NH NH	[1.0, 1.0, 1.0]	[1.0, 1.0, 1.0]	[0, 0, 0]

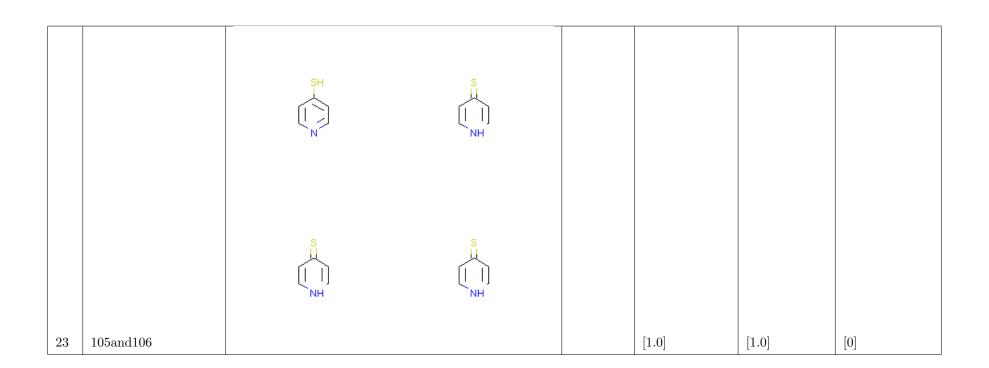
		•
	_	٠.

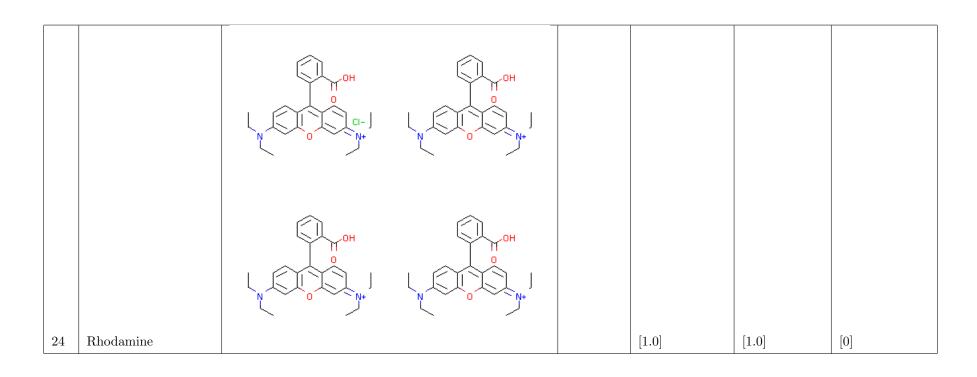
		NH2	NH NH			
21	aminopyrimidine	NH L NH	NH NH	[1.0]	[1.0]	[0]

	•
_	
_	-

	OH N	O NH			
22 103and104	NH NH	O NH	[1.0]	[1.0]	[0]

L	J
II.	





0	

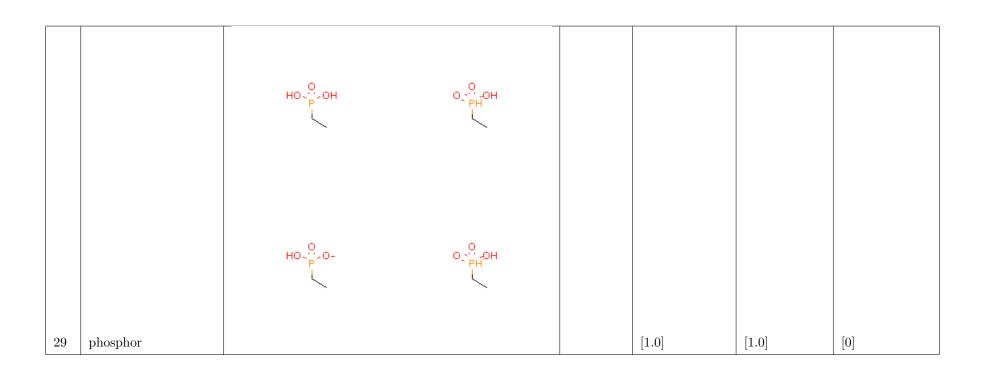
		О ) S 2 О НО ) S	O.S.O HO			
25	sulfon	0)s <sup>2</sup> 0	0.35-0 HO_S	[1.0]	[1.0]	[0]

$\sim$		)
7		٠
•	J	٥

		N NH NH	N N N N N N N N N N N N N N N N N N N			
27	tetrazole	N N N N N N N N N N N N N N N N N N N	N N N N N N N N N N N N N N N N N N N	[1.0]	[1.0]	[1]

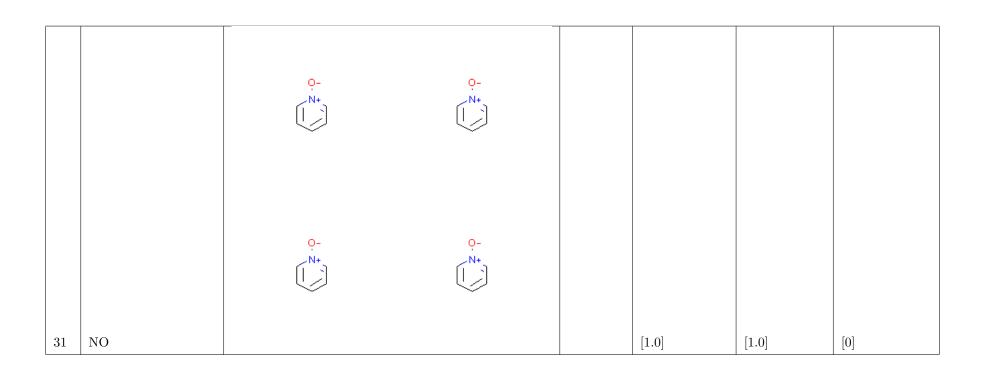
÷

	OH OH	OH				
28 phenol	<u> </u>	OH 	[1	1.0]	[1.0]	[0]



		SH SH	SH SH			
30	112and113	s =s	SH SH	[1.0]	[1.0]	[0]

	•
	_



$\neg$			
	-	•	J

		N+ N+ N-	N+ N+ N-			
32	diazo	0. + N+ N	0-+ N+ N	[0.24]	[0.27]	[34]