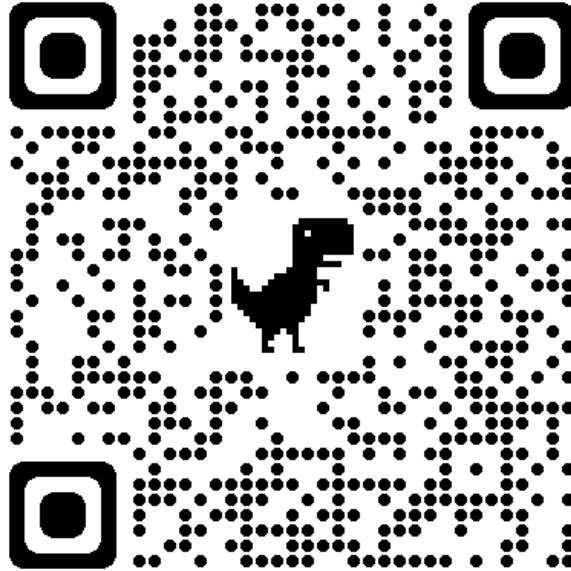


Doc-a-thon: Chemical representation best practices for humans and machines

ACS Spring 2023, Indianapolis

Vincent Scalfani, Leah McEwen, and Evan Bolton

Slides and Reference Documentation



Use QR Code, or Google “vfscalfani gists” or here is the direct link:

<https://gist.github.com/vfscalfani/1eb4c2b32e3b8730933bf77643bca0a1>

Goals for Today (1.5 hour discussion)

- Discuss use-cases and edge-cases where machine-interpretation of the graphical standards are limited.
- Evaluate selected parts of the current IUPAC recommendations [1, 2] and discuss if recommendations are still appropriate for both human and modern machine-interpretation of chemical structures.

[1] Graphical representation standards for chemical structure diagrams, J. Brecher, K. N. Degtyarenko, H. Gottlieb, R. M. Hartshorn, K.-H. Hellwich, J. Kahovec, G. P. Moss, A. McNaught, J. Nyitrai, W. Powell, A. Smith, K. Taylor, W. Town, A. Williams, A. Yerin, *Pure Appl. Chem.*, **2008**, 80(2), 277-410, <https://doi.org/10.1351/pac200880020277>

[2] Graphical representation of stereochemical configuration, J. Brecher, K. N. Degtyarenko, H. Gottlieb, R. M. Hartshorn, G. P. Moss, P. Murray-Rust, J. Nyitrai, W. Powell, A. Smith, S. Stein, K. Taylor, W. Town, A. Williams, A. Yerin, *Pure Appl. Chem.*, **2006**, 78(10), 1897-1970, <https://doi.org/10.1351/pac200678101897>

N.B. All images and excerpts in these slides are from reference [1] and [2]

Outcomes for Today

- Start to identify areas of IUPAC representation standards documentation that need to be updated (e.g., we disagree or to align with modern chemical drawing/cheminformatics software interpretation).

Follow-up work after discussion:

- A two page "short guide" to supplement existing graphical representation standards (focused on general chemist best practices for machine interpretation)
- An IUPAC project proposal to completely update IUPAC's graphical representation standards.

Discussion of Representation Standards “Software Cautions”

[1] Graphical representation standards for chemical structure diagrams, J. Brecher, K. N. Degtyarenko, H. Gottlieb, R. M. Hartshorn, K.-H. Hellwich, J. Kahovec, G. P. Moss, A. McNaught, J. Nyitrai, W. Powell, A. Smith, K. Taylor, W. Town, A. Williams, A. Yerin, *Pure Appl. Chem.*, **2008**, 80(2), 277-410, <https://doi.org/10.1351/pac200880020277>



disagree or ambiguous/incorrect, **needs revision**



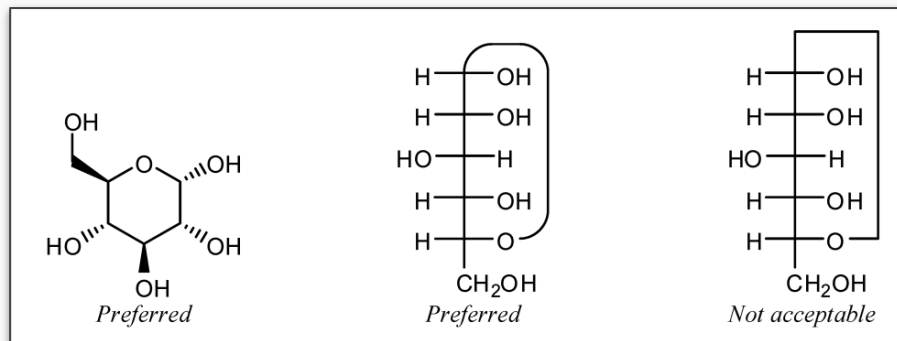
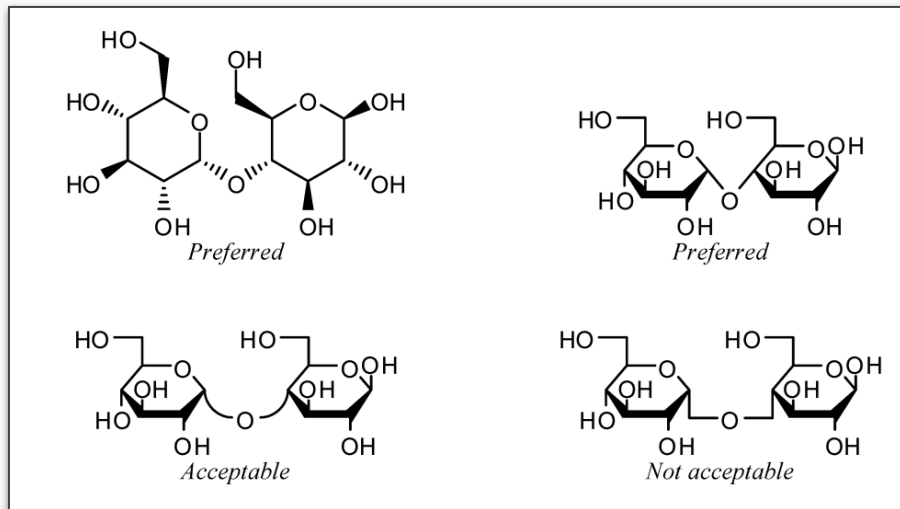
somewhat disagree or sometimes ambiguous/incorrect, **may need revision**



agree, still generally correct, **no revision needed**

GR-1.5 Bonds with bends [1]

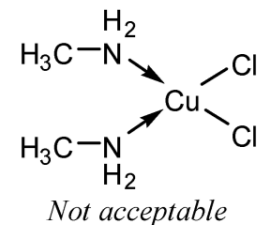
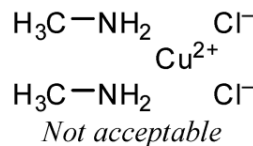
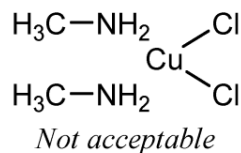
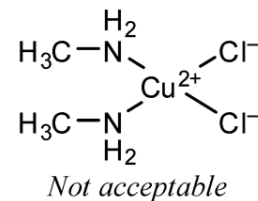
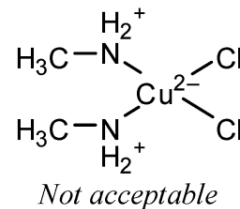
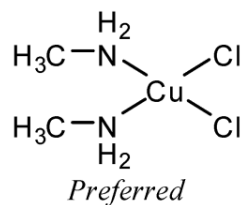
"SOFTWARE CAUTION: At the time of writing of this document, the authors know of no computer software that is able to represent bonds with smooth curves."



[1] <https://doi.org/10.1351/pac200880020277>

GR-1.7 Coordination Bonds [1]

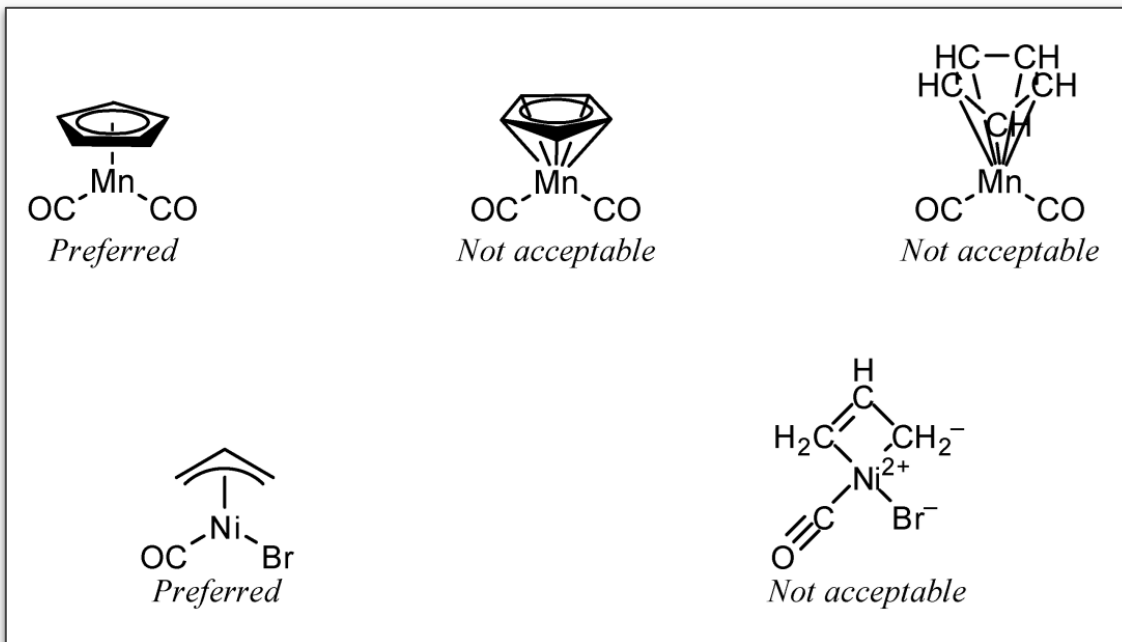
“SOFTWARE CAUTION: Some existing software may be unable to interpret properly coordination bonds drawn with single bonds and without charges, as recommended above. When creating chemical structure diagrams for use with such software, one of the otherwise “not acceptable” forms may in fact be the only way to produce a diagram that the software can understand.”



[1] <https://doi.org/10.1351/pac200880020277>

GR-1.7.2 Coordination bonds to contiguous atoms [1]

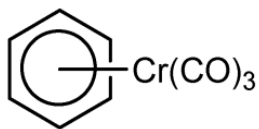
“SOFTWARE CAUTION: When working electronically, it is extremely important to specify a coordination bond appropriately, according to the capabilities of the software program you are using. If the coordination bond is specified inappropriately, the diagram may be interpreted as two disjoint fragments, with the coordination bond being interpreted as a normal bond to an unlabeled carbon atom.”



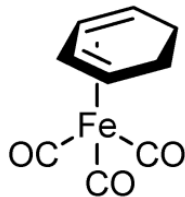
[1] <https://doi.org/10.1351/pac200880020277>

GR-1.7.2 Coordination bonds to contiguous atoms [1]

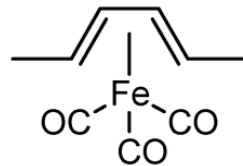
“SOFTWARE CAUTION: When working electronically, it is extremely important to specify a coordination bond appropriately, according to the capabilities of the software program you are using. If the coordination bond is specified inappropriately, the diagram may be interpreted as two disjoint fragments, with the coordination bond being interpreted as a normal bond to an unlabeled carbon atom.”



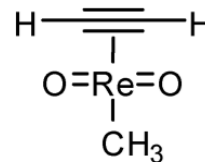
Preferred



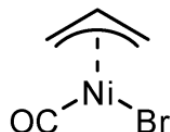
Preferred



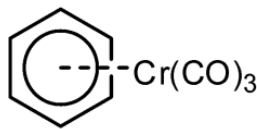
Preferred



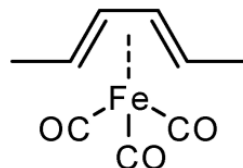
Preferred



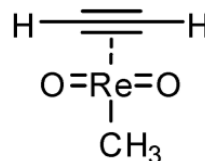
Not acceptable



Not acceptable



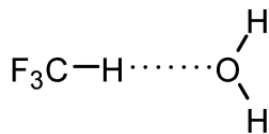
Not acceptable



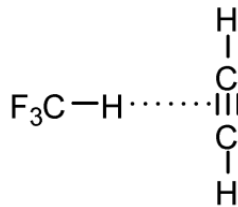
Not acceptable

GR-1.8 Partial Bonds [1]

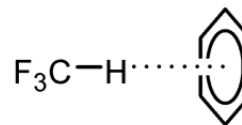
“SOFTWARE CAUTION: At the time of preparing this document, the authors are unaware of any computer software that can produce dotted bonds. In cases where it is not possible to produce true dotted bonds, it is acceptable to use dashed bonds instead.”



Preferred



Preferred

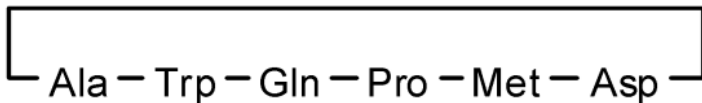


Preferred

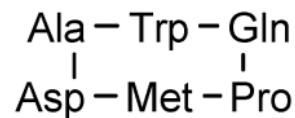
[1] <https://doi.org/10.1351/pac200880020277>

GR-2.2.1 Three-letter amino acid abbreviations [1]

“SOFTWARE CAUTION: At the time of writing of this document, few software programs are able to interpret bent bonds at all, and the authors know of no computer software that is able to fully interpret complex cyclic peptides as depicted above. If chemical structure diagrams of cyclic peptides are required for use within an electronic environment, the use of bent bonds may indeed be not acceptable in that case, and an alternative diagram may be required that avoids the use of abbreviations with two or more attachments.”



Preferred

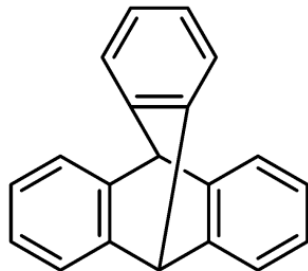


Not acceptable

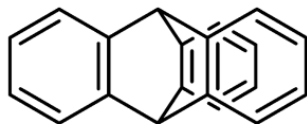
[1] <https://doi.org/10.1351/pac200880020277>

GR-3.3.4 Bridged Rings [1]

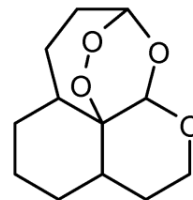
“SOFTWARE CAUTION: Although one hopes that no human would create diagrams with rings that exactly overlap in this manner, such overlap is regrettably common in diagrams produced by computer software. Authors are always encouraged to review the output of computer programs carefully, lest any atoms or bonds be completely obscured and the apparent meaning of a diagram be other than was intended.”



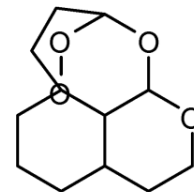
Preferred



Not acceptable



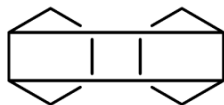
Preferred



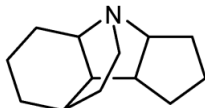
Not acceptable



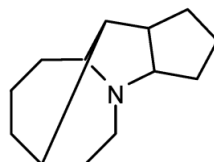
Preferred



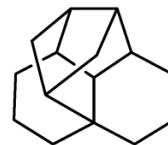
Not acceptable



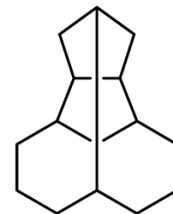
Preferred



Not acceptable



Preferred

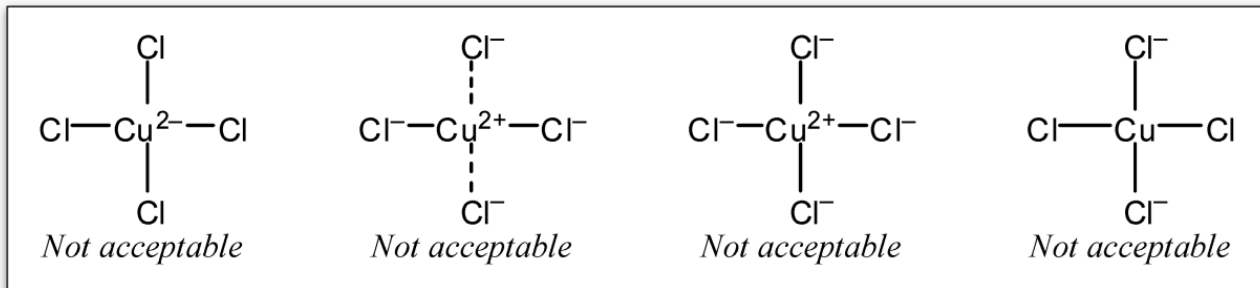
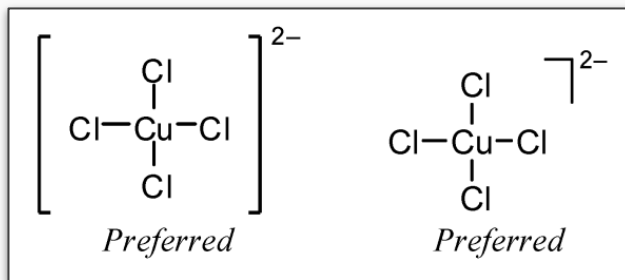


Not acceptable
[when representing
the molecule at left]

[1] <https://doi.org/10.1351/pac200880020277>

GR-5.7 Polyatomic ions [1]

“SOFTWARE CAUTION: When working electronically, it is extremely important to specify delocalized charges appropriately, according to the capabilities of the software program you are using. [...] Given those weaknesses, software programs appear to most commonly prefer the structural depiction that localizes the negative charge on the metal atom. If you need to store structures of this type in an electronic format, please consult the documentation for the software programs you are using for recommendations on the best methods to use with that package.”

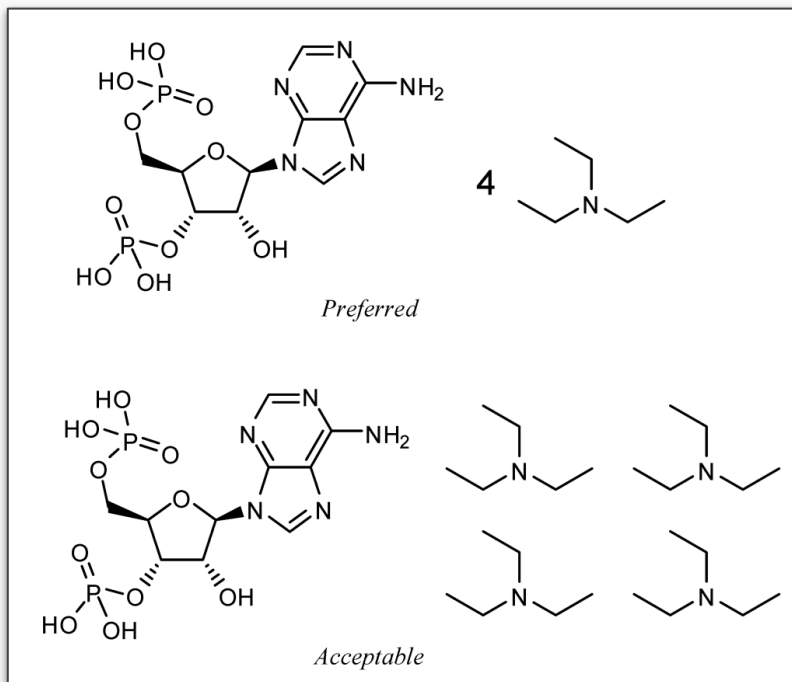


[1] <https://doi.org/10.1351/pac200880020277>



GR-7.2 Positioning of components [1]

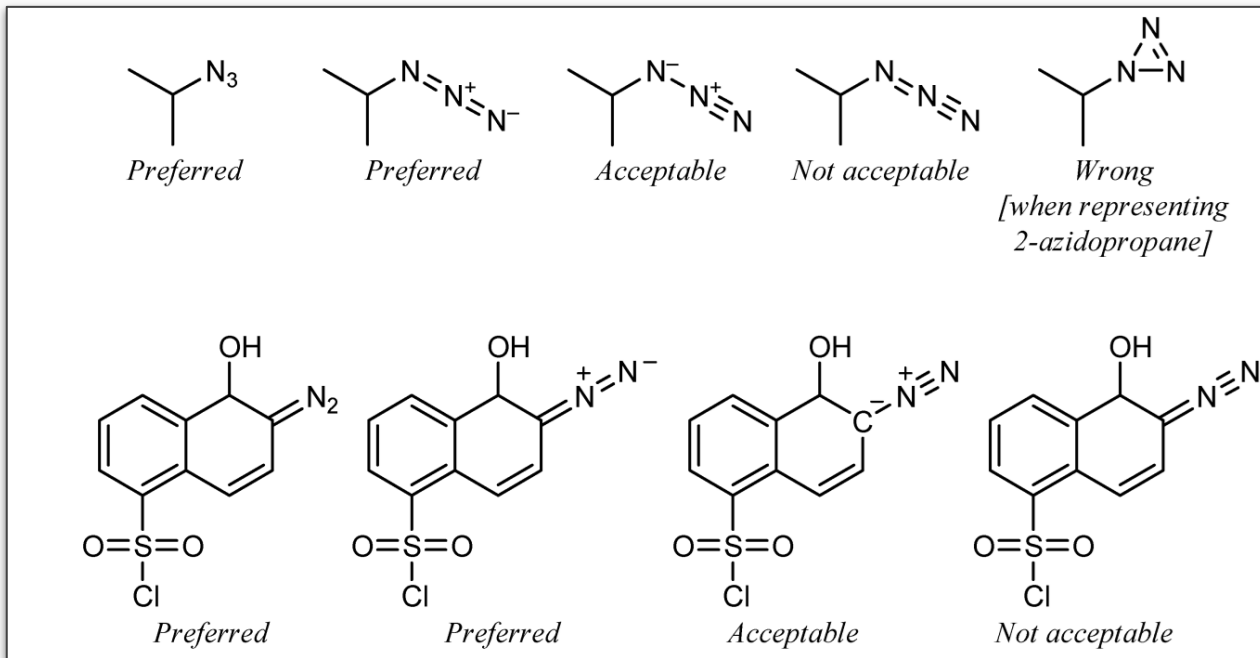
“SOFTWARE CAUTION: At the time of writing this document, there are few computer software programs that are able to recognize stoichiometric multipliers. If chemical structure diagrams of multi-component salts are required for use within an electronic environment, it may be necessary to depict all of the components explicitly.”



[1] <https://doi.org/10.1351/pac200880020277>

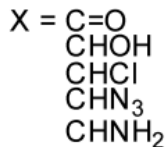
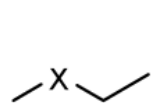
GR-8.1 Nitrogen Compounds [1]

“SOFTWARE CAUTION: Although both depiction styles may be used for depicting azides and diazo compounds, different software programs may interpret the two styles as referring to chemically different compounds. It may be necessary to use only the Preferred style when working in electronic environments.”

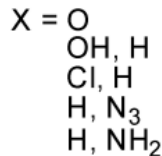
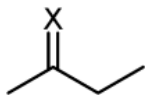


GR-9.1 Small substituents [1]

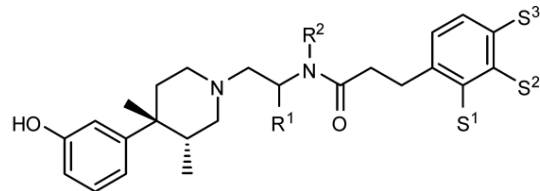
"SOFTWARE CAUTION: Variable lists of all sorts pose many problems for interpretation by computer software. At the time of writing this document, even the simplest forms can be recognized by only a few software programs, and extended variable lists are not supported by any software. Although these forms are extremely compact and can be a boon to printed publication, it will usually be necessary to draw all forms explicitly when using current programs."



Preferred



Not acceptable



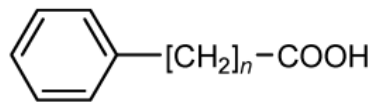
Compd.	R ¹	R ²	S ¹	S ²	S ³	% inhib.
...						
8	iPr	H	H	H	OH	71
9	iPr	H	H	H	OH	11
10	iPr	H	H	H	H	28
11	iPr	H	H	OH	H	20
12	iPr	H	OH	H	H	25
13	iPr	H	H	H	OH	6
14	iPr	H	H	H	OH	15
15	iPr	H	H	H	F	26
16	iPr	H	H	OH	OH	31
17	iPr	H	H	OCH ₃	OH	42
18	iPr	H	H	H	OCH ₃	16
19	H	H	H	H	OH	11
20	CH ₃	H	H	H	OH	20
21	H	CH ₃	H	H	OH	0
22	CH ₃	CH ₃	H	H	OH	1
23	Ph	CH ₃	H	H	OH	7

Preferred [29]

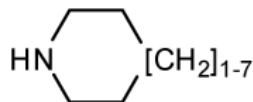
[1] <https://doi.org/10.1351/pac200880020277>

GR-9.3 Variable chain length and ring size [1]

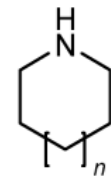
“SOFTWARE CAUTION: At the time of writing this document, there is no software that supports the use of curved bonds. Although this representation is acceptable for general depiction, it must be considered not acceptable in cases where the resulting diagram will be further modified in electronic format.”



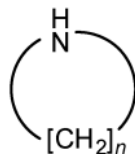
Preferred



Preferred



Preferred

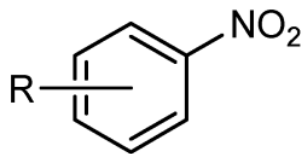


Acceptable

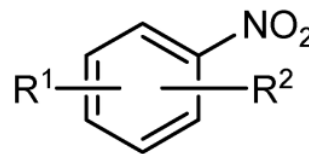
[1] <https://doi.org/10.1351/pac200880020277>

GR-9.4 Variable attachment location [1]

“SOFTWARE CAUTION: When working electronically, it is extremely important to specify the variable attachment correctly, according to the capabilities of the software program you are using. If the variable attachment is specified incorrectly, the structure may be interpreted as two disjoint fragments, with the variable bond being interpreted as a normal bond to a carbon atom. In addition to losing the intended variability, this misinterpretation will also add an additional CH₄ to the structures perceived formula: CH₃ for the “methyl group” at the center of the ring, and one more H for the location in the ring that was not substituted as intended.”



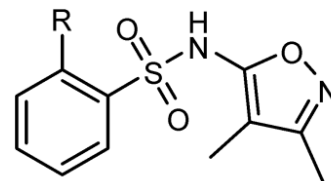
Preferred
[representing any of five molecules
(three unique molecules)]



Preferred
[representing any of 20 molecules
(10 unique molecules)]

GR-9.5 Large substituents [1]

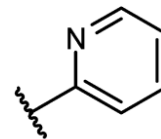
“SOFTWARE CAUTION: At the time of writing this document, there is no software that is able to interpret this sort of tabular data. Nonetheless, it remains as an acceptable way of depicting large substituents and large collections of substituents in printed form, even if it may be necessary to draw all individual molecules explicitly if they are to be further modified by computer software.”



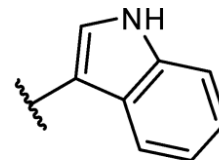
Compd.	R
--------	---

1	Ph
---	----

2	
---	--



3	
---	--



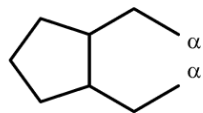
Preferred [30]

[1] <https://doi.org/10.1351/pac200880020277>

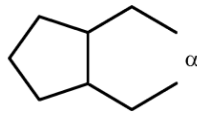


GR-11. Annotations [1]

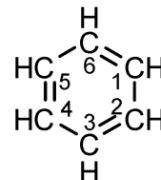
"SOFTWARE CAUTION: Some software programs may have special ways to create annotations so that they remain associated with their target objects, or so that they may be indicated as having a specific meaning. If you are using annotations in an electronic environment, you may want to take advantage of the features that your chemical drawing software offers in this area."



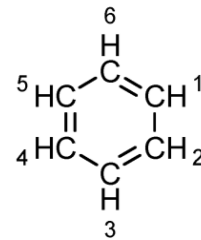
Preferred



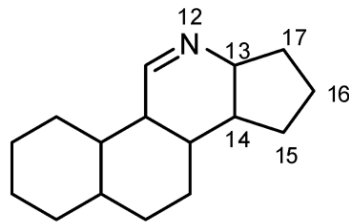
Not acceptable



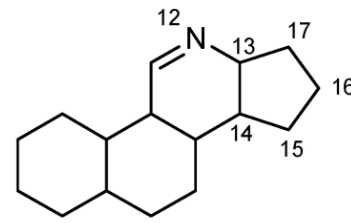
Preferred



Not acceptable



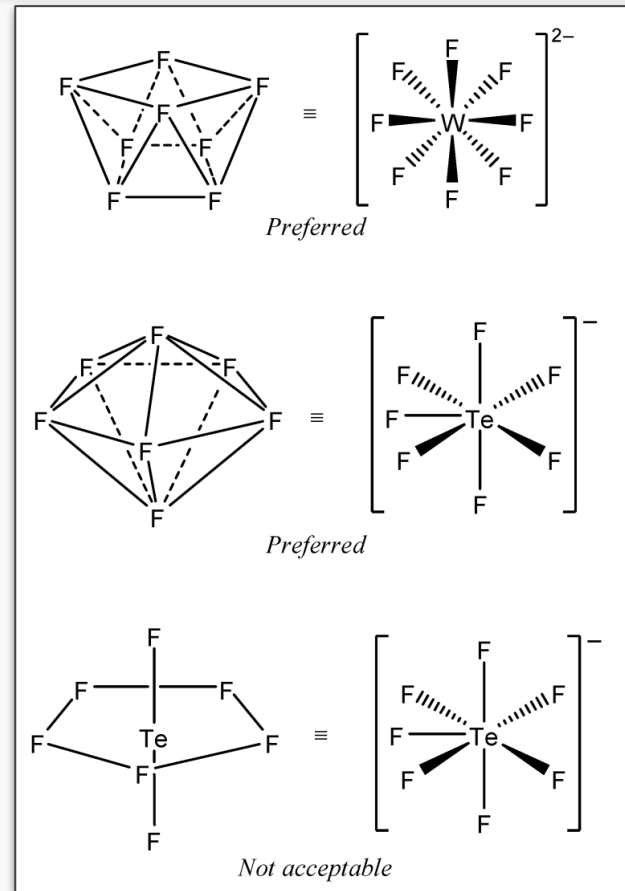
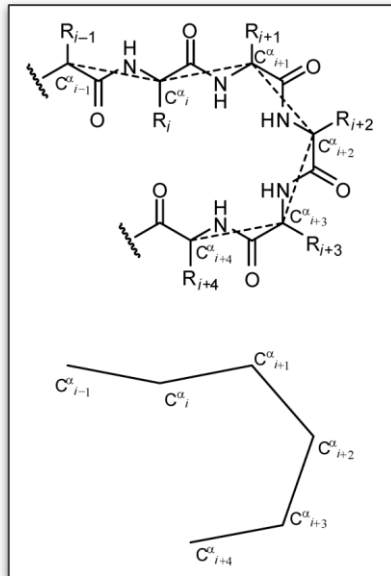
Preferred



Not acceptable

GR-12 Pseudobonds [1]

“SOFTWARE CAUTION: At the time of writing this document, there are few computer software programs that are able to recognize pseudobonds of any sort. If chemical structure diagrams of such molecules are required for use within an electronic environment, it may be necessary to eschew pseudobonds and instead depict all of the atoms and bonds explicitly.”



[1] <https://doi.org/10.1351/pac200880020277>

Discussion of Representation Standards “Software Cautions”

[2] Graphical representation of stereochemical configuration, J. Brecher, K. N. Degtyarenko, H. Gottlieb, R. M. Hartshorn, G. P. Moss, P. Murray-Rust, J. Nyitrai, W. Powell, A. Smith, S. Stein, K. Taylor, W. Town, A. Williams, A. Yerin, *Pure Appl. Chem.*, **2006**, 78(10), 1897-1970,
<https://doi.org/10.1351/pac200678101897>



disagree or ambiguous/incorrect, **needs revision**

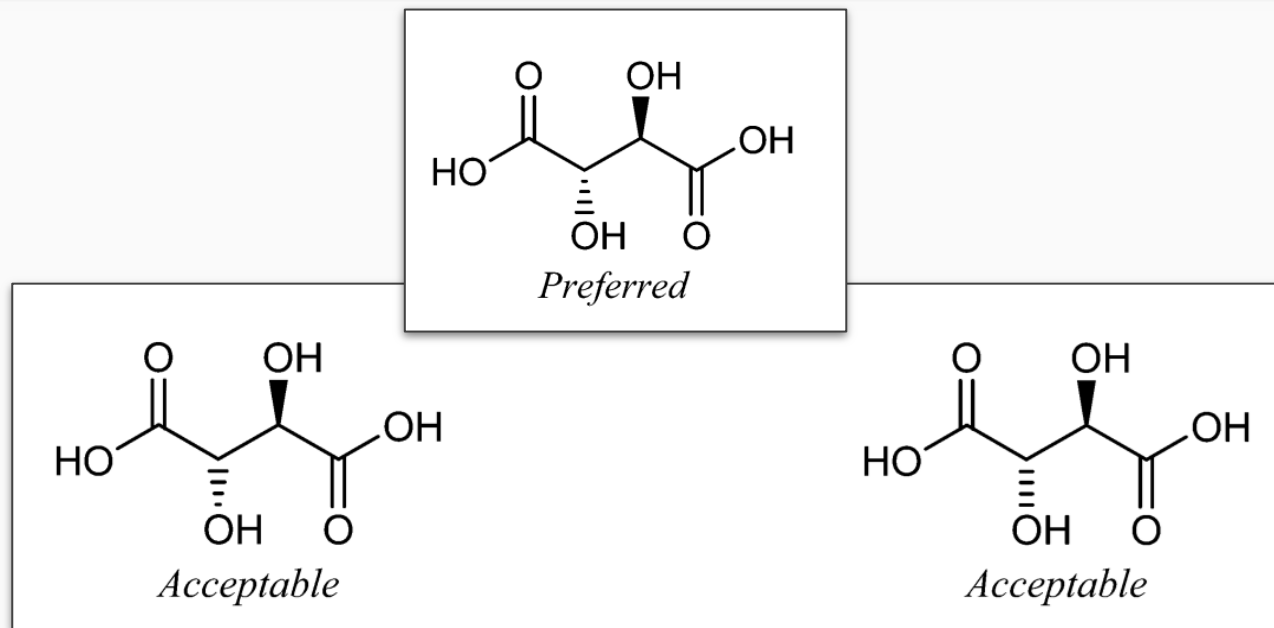


somewhat disagree or sometimes ambiguous/incorrect, **may need revision**



agree, still generally correct, **no revision needed**

ST-0.3 Hashes, dashes, and wedges [2]



“At best, an author using bonds of these types may cause a moment of confusion as a reader figures out which convention was intended. In extreme cases, a reader (or chemical software) might actually interpret the structure incorrectly—for example, by interpreting an unwedged hashed bond in the convention where it would represent relative (and not absolute) configuration, as discussed in ST-0.8 and ST-6.”

ST-0.8 Mixtures of diastereoisomers [2]



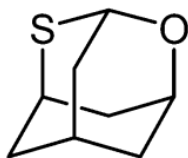
"There are a variety of obscure conventions for depicting arbitrary mixtures of diastereoisomers. The most common (but still obscure) of these conventions says that bold and hashed wedges should be used for absolute configuration, **unwedged bold and unwedged hashed lines represent relative configuration and racemic character, while hollow wedges and dashed lines represent relative configuration only** [7,8].

"At this point, there does not seem to be any consensus for a preferred depiction style. Given the lack of consensus among different systems and the poor general acceptance of any given one, the safest approach currently is to depict multiple structure diagrams explicitly, optionally accompanied by additional descriptive text as discussed in ST-6.3."

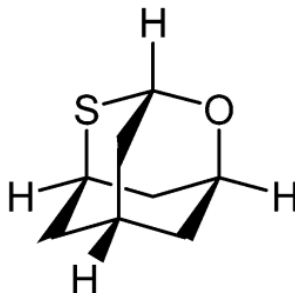
[2] <https://doi.org/10.1351/pac200678101897>

ST-3. Use of perspective to indicate configuration [2]

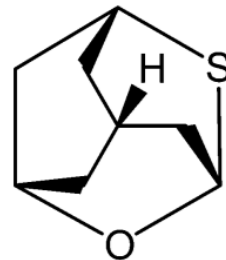
"At the current time, few computer-based software packages will recognize perspective at all, let alone recognize it accurately and completely. For this reason, although perspective-based diagrams can be quite useful for humans, they should still be avoided in situations where accurate computer interpretation is important."



Preferred



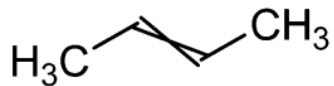
Acceptable



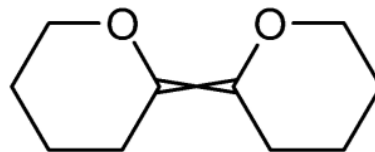
Acceptable

ST-4.4 Double bonds of unspecified configuration [2]

“A crossed double bond has sometimes been used to indicate unspecified double bond configuration or a mixture of configurations. This type of bond is not considered acceptable for general use, although it may still be required by some computer software.”



Not acceptable



Not acceptable

ST-6.2 Absence of indicators indicates configuration as drawn [2]

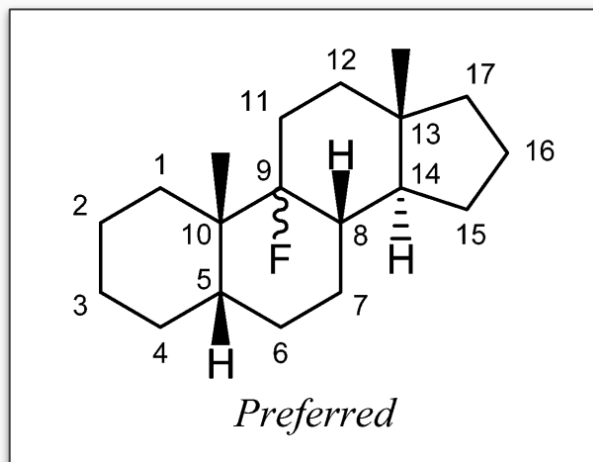
"In the absence of any other sort of indicators, a structural diagram is expected to represent the specific molecular entity drawn. If stereochemical hashed wedged and/or solid wedged bonds are present at all stereogenic centers in the diagram, it will represent a single stereoisomer. If no stereochemical information is present in the diagram, it represents a molecular entity with unknown configuration."



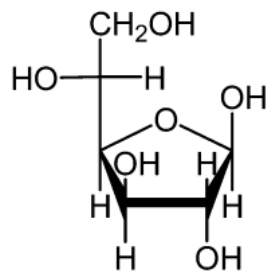
[2] <https://doi.org/10.1351/pac200678101897>

ST-0.4 Wavy bonds [2]

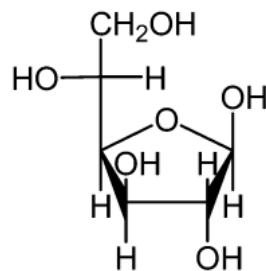
"Historically, wavy bonds have frequently been used to represent an unspecified mixture containing enantiomers or diastereoisomers. It is preferred to use a plain bond to depict an unspecified mixture, just as it is preferred when depicting single molecular entities with unspecified configuration. If the nature of the mixture needed further emphasis, a wavy bond would remain acceptable for mixtures as well."



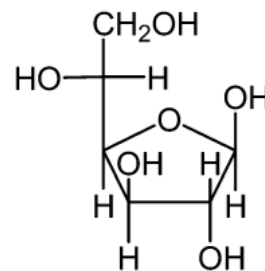
ST-1.9/10 Haworth and Mills depictions [2]



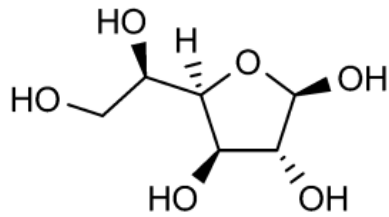
Acceptable



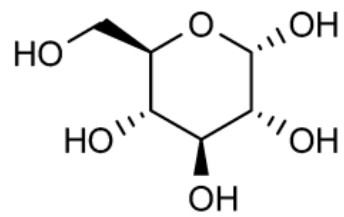
Not acceptable



Acceptable



Preferred



Preferred

Wrap up

Take home messages?

Where do we go from here?

Follow-up work after today:

- A two page "short guide" to supplement existing graphical representation standards (focused on general chemist best practices for machine interpretation)
- An IUPAC project proposal to completely update IUPAC's graphical representation standards.
- Other thoughts?