Algorithmic analysis of Cahn-Ingold-Prelog rules of stereochemistry: Proposals for revised rules and a guide for machine implementation

Robert M. Hanson, John W. Mayfield, Mikko J. Vainio, Andrey Yerin, Dmitry Redkin, and Sophia Musacchio

Abstract

The most recent version of the Cahn-Ingold-Prelog rules for the determination of stereodescriptors as described in *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013* (the "Blue Book") were analyzed by an international team of cheminformatics software developers. Algorithms for machine implementation were designed, tested, and cross-validated. Deficiencies in Sequence Rules 1b and 2 were found, and proposed language for their modification is presented. A concise definition of an additional rule ("Rule 6," below) is proposed, which succinctly covers several cases only tangentially mentioned in the 2013 recommendations. Each rule is discussed from the perspective of machine implementation. The four resultant implementations are supported by validation suites in 2D and 3D SDF format as well as SMILES. The validation suites include all significant examples in Chapter 9 of the Blue Book, as well as several additional structures that highlight more complex aspects of the rules not addressed or not clearly analyzed in that work. These additional structures support a case for the need for modifications of the Sequence Rules.

Introduction

In the 60+ years since the introduction of Cahn-Ingold-Prelog Sequence Rules in 1956,[1] the "CIP Rules" have become an integral part of chemical nomenclature, providing a way to identify the spatial arrangement of atoms of a molecule using simple mostly atom- or bond-based stereodescriptors. Over the course of this time, various authors have pointed out deficiencies in the rules and proposed solutions in the form of modifications and subrules,[2,3] to the point where today we have eight distinct Sequence Rules: 1a, 1b, 2, 3, 4a, 4b, 4c, and 5*.*

In order to provide a single resource summarizing the state of the evolving rules, the International Union of Pure and Applied Chemists (IUPAC) published the first comprehensive description of the CIP Rules in *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013* (referred to below as "BB 2013").[4] This description, an impressive 197-page chapter with more than two hundred examples, presents a complete set of CIP Rules, along with detailed procedures for their application.

The "CIP descriptors" generated by these rules are primarily intended for use in chemical nomenclature. It would be possible to use them in other applications, such as the removal of redundant stereo specifications, the determination of structure equivalence, or canonical labeling. However, CIP-based algorithms may be more resource consuming and complex than simpler, more efficient cheminformatics algorithms developed specifically for those purposes.

We note that discoveries of deficiencies in the original rules have been made before in the context of developing computer-based implementations. This was certainly the case in 1993, when subrules 4a, 4b, and 4c were proposed by a group developing a stereochemical module for the LHASA computer-aided synthesis analysis program.[5,6]

Not surprisingly, predating the open-source collaborative environment and not having a concise reference in hand, machine implementations of the CIP Rules to date have been only marginally successful. Certainly, many software developers have implemented the CIP Rules to one extent or another, but recent analysis of available software packages clearly demonstrates that there is much disagreement among these implementations, even for relatively simple compounds, among several highly respected software packages.[7]

In the spring of 2017, concurrent discussions started in two forums, the IUPAC Blue Book Project,[8] focused on preparation of errata for BB 2013, and the Blue Obelisk Group,[9] focused on implementation issues of CIP rules in algorithmic form. The result of these lively discussions has been a reason for coordinated and thorough analysis of the CIP Rules, with the aim of concurrent development and improvement of four software packages: Jmol[10], Centres[11], ChemSketch[12], and Balloon[13,14]. Our joint efforts ensured that multiple, validated, independent machine implementations of the CIP Rules are made available to the cheminformatics community, as well as to anyone interested in stereochemistry and chemical nomenclature. In fact, the results of our work have already been incorporated into an interactive web site utilizing JSmol.[15]

As we sought consensus when there were issues and questions as to interpretation of the rules or correctness of BB 2013 examples, we turned to the use of finite acyclic digraphs,[3] the assumption being that analysis of digraphs should always be the final arbiter in any dispute relating to stereochemistry. The use of digraphs in our discussions allowed for one of three possible conclusions: (a) that the disagreement was due to different interpretations of CIP rules among software developers, (b) that there was a problem with an algorithm or its implementation in code, or (c) that the CIP rules themselves were flawed. In fact, all three of these possibilities were encountered in the process of coming to consensus, including the discovery of a small number of errors in the Blue Book, two minor flaws in the CIP rules, and a proposal for a new rule.

Preliminary Considerations

Before discussing the Sequence Rules, we note that BB 2013 is inconsistent in its use of terminology in relation to duplication of nodes. In the discussion that follows, and in our proposed revision of these rules, we use the following terminology exclusively:

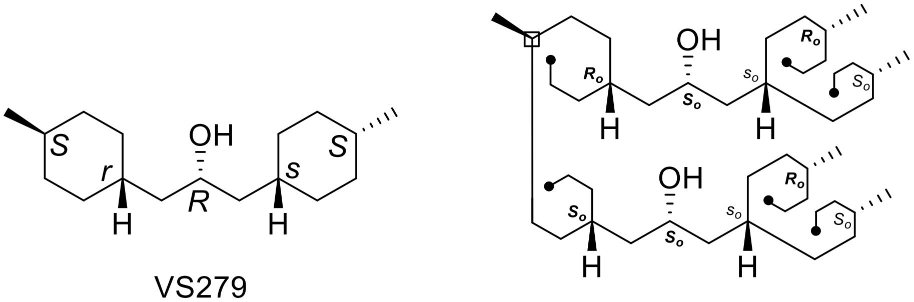
duplicate node	A digraph node that has been added as a copy of a “real” atom.
duplicated atom	A digraph node that represents the real atom that has been duplicated.

While the considerations below are mostly discussed in terms of stereogenic centers, the same procedures are valid for other stereogenic units, including double bonds, odd- and even-cumulenes, and atropisomers. Structure numbers starting with "VS" refer to compounds in the validation suite, available as supporting material and at https://cipvalidationsuite.github.io/ValidationSuite.

Digraphs and Auxiliary Descriptors. The importance of using finite acyclic directed graphs ("digraphs") and auxiliary descriptors (temporary assignments made in the process of determining a specific center's descriptor) was introduced by Prelog and Helmchen in 1982 in relation to cyclic structures[3] and emphasized later by Mata et al.[5] in relation to the process of developing Rule 4:

> *If, for a clear analysis of the ligands, they must be converted into hierarchical digraphs, then the comparison must always be done considering the hierarchical digraph, not the real ligand.*

Our consensus interpretation of this statement is that, at least for cyclic compounds and compounds with more than one stereocenter, a unique digraph must be generated for each center in question. This includes double-bond as well as axial stereochemistry and is true whether the center in question is in the ring or not. Temporary auxiliary descriptors are assigned solely on the basis of this digraph and may or may not be the "final" descriptors ultimately used to describe those centers using their own digraphs. Figure 1 shows an example (VS279) where only a minority of the auxiliary descriptors are the same as the final descriptors for the corresponding atoms.

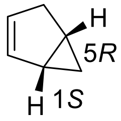


VS279

Figure 1. A compound with its simplified digraph for the determination of the *S* stereodescriptor on the far-left ring carbon. Only four of the ten auxiliary descriptors on the digraph are their final values.

Generation of a complete digraph, including all auxiliary descriptors, including *seqcis* and *seqtrans*, is required prior to Rule 3. The entire sequence of all rules must be carried out for each auxiliary center on the *same* digraph. Generation of auxiliary descriptors must start from the highest sphere, proceeding toward the root. In this way, all auxiliary descriptors in higher spheres than the one being determined are already assigned. This is sufficient, as the descriptor for an auxiliary center does not depend upon any descriptor between it and the root. The auxiliary center ligand leading to the digraph root can always be ranked by Rule 1a exclusively, as auxiliary centers are offset from the root of a digraph, and so the path back to the root is always unique in connectivity and atomic numbers.

It is common to use simplified digraphs that do not show hydrogen or “phantom” atoms. However, from an implementation perspective, simplified digraphs must be used with caution. In complex examples, for instance, a critical aspect of the algorithm must be finding the *first* difference in two ligands, and this may not be obvious from a simplified digraph. For example, there is no rule that “real atoms have higher priority than duplicate nodes.” This is intentional. While generally true, this statement hides the fact that duplicate nodes lose to their real counterparts only *in the next higher sphere*, where their associated phantom atom, with atomic number zero, always loses to any real atom. That being the case, *some other* consideration may be missed. An example of a compound for which proper consideration in this regard is required is the bicyclic alkene shown in Figure 2 (VS172). Naïve application of the pseudo-rule “real atoms have higher priority than duplicate nodes” leads to a 1*R* descriptor rather than correct 1*S* assigned in accord to Rule 1b*.*



VS172

Figure 2. (1*S*,5*R*)-bicyclo[3.1.0]hex-2-ene (VS172) may be improperly assigned 1*R* rather than 1*S* based solely on a simplified digraph.

Sequence Rules and Subrules. While BB 2013 specifies that each of Rules 1-5 must be applied “exhaustively in the order given,” it may not be clear what exactly “exhaustively” means or how this relates to subrules such as 1a and 1b. In fact, each subrule must be applied sequentially. Effectively, there are eight (nine, if one adds our proposed Rule 6) fully independent explicit Sequence Rules: 1a, 1b, 2, 3, 4a, 4b, 4c, and 5. The fact that these are not “Sequence Rules 1-8” is simply a result of the historical evolution of the Sequence Rules. “Exhaustively” simply means “until a decision is reached, or it is determined that no such decision is possible.”

Ranking and Comparing Ligands. As described clearly in BB 2013, application of all eight rules involves the same two processes: Ligands are ranked sphere by sphere, branch by branch in a breadth-first fashion. Then, in pairs, two ligands are compared atom by atom, in order of that ranking. In practice, it is not always necessary to completely rank a ligand, including hydrogen and phantom atoms. Ranking – at least through Rule 2 – can be carried out on a “need-to-know” basis, skipping whole sub-branches of the digraph where a decision has already been made.

Results and Discussion – Algorithmic analysis of CIP procedures

What follows is a discussion of each of the eight independent Sequence Rules from the perspective of machine implementation, along with a proposal for a ninth rule that we are calling “Rule 6.”

Rule 1a: *Higher atomic number precedes lower.*

On the face of it, Rule 1a sounds simple enough to implement. And it is, except for the special cases discussed in BB 2013 P-92.1.4.4 in relation to compounds and ions with multiple chemically equivalent Kekulé structures, such as benzene, pyridine or cyclopentadienyl anion. That section briefly introduces the idea of “atomic number averaging” for mancude-ring systems, which are “rings having (formally) the maximum number of noncumulative double bonds, e.g. benzene, indene, indole, 4H-1,3-dioxine”.[16] The idea is to average the atomic number of the duplicate node when it is involved in multiple resonance structures. Our reading of Section P-92.1.4.4 is that it is not a well-crafted guideline, with many unanswered questions. What exactly defines the pertinent cases for which this rule applies? What about acyclic cases such as allyl anion or acetoacetate? How exactly is the averaging to be done – Do we need to know the exact weighting of all the possible resonance structures, or is it sufficient to average over just the adjacent atoms involved in the electron delocalization?

Without going into detail here, suffice it to say that atomic number averaging is a difficult procedure to describe or implement, other than the case of all-carbon neutral species, for which it is unnecessary, and some simple heterocyclic systems, such as pyridine derivatives. Thus, consider substituted pyridine VS032/ VS033, Figure 3. Notice that without this consideration, Rule 1a gives two different results in this case, depending upon the choice of Kekulé structure. The correct result, *S*, derives from assigning the duplicate node for the aromatic nitrogen an averaged atomic number of 6.5, which loses to the unaveraged atomic-number 7 duplicate node of the imine nitrogen.

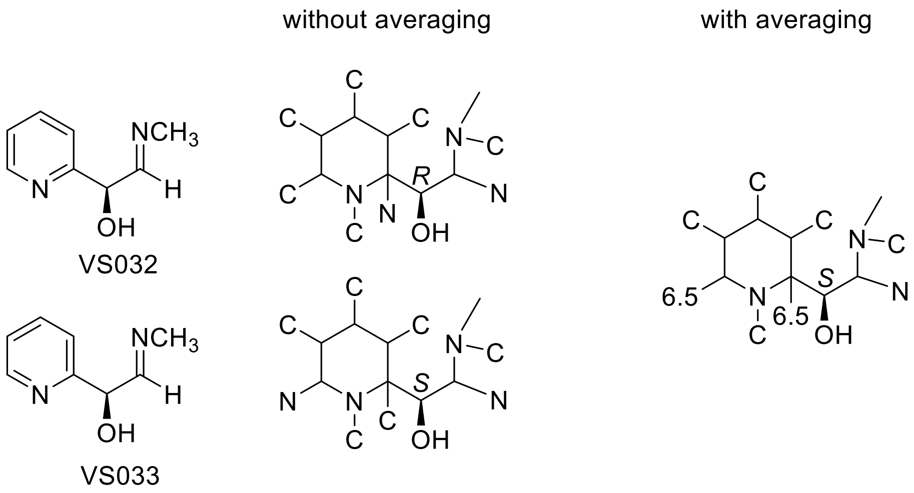


Figure 3. Without atomic number averaging in Rule 1a, the two chemically equivalent Kekulé structures give different assignments at the stereocenter. Atomic number averaging removes the issue.

Rule 1b: *A duplicate atom node whose corresponding nonduplicated atom node is the root or is closer to the root ranks higher than a duplicate atom node whose corresponding nonduplicated atom node is farther from the root.*

Shortly into our study it became clear that the current IUPAC recommendation for Rule 1b is not sufficient. The problem is that although Rule 1b was designed to solve a problem with ring-closure duplicate nodes,[18] the rule as stated also applies to multiple-bond duplicate nodes. As such, we again have in Rule 1b the same issue involving multiple Kekulé structures as for Rule 1a. The problem involves cases such as shown in Figure 4, where the application of Rule 1b as currently recommended can lead to the inappropriate introduction of stereodescriptors. The issue is just one specific case of more general problem of absent procedures to assign root distances for duplicates resulted from averaging of atomic numbers.

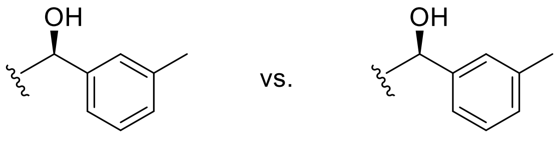


Figure 4. Two chemically equivalent Kekulé structures for the same molecular fragment. Assigning distance to root using the current Rule 1b gives the same problem as found for Rule 1a if an averaging criterion is not applied.

Our discussion revolved around how to address this issue algorithmically: (a) Should we implement an averaging scheme as in Rule 1a for all six duplicate nodes? (b) Do we omit multiple-bond duplicate nodes from consideration-- just assigning them "n/a" and skipping them entirely? (c) Should we assign the distance to the root of their sphere? (d) Do we assign the distance to the root of their attached atom? Our group decided that the first of these options was too complex, the second would lead to ambiguities in the algorithm, the third would not work, and the simplest and surest solution would be the last of these -- to assign to a multiple-bond duplicate node the distance to the root of its corresponding attached atom, not its corresponding duplicated atom.

Thus, the digraph for bis-(2-hydroxyphenyl)methanol is shown in Figure 5, where numbers are distances to the root assigned for each node needed for Rule 1b (not the usual atomic numbers seen on digraphs in BB 2013). The specific representation of the bonding is no longer significant, and the center is found to have no descriptor, as expected.

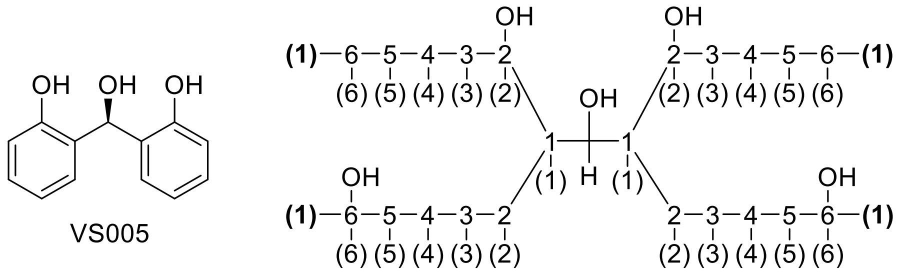


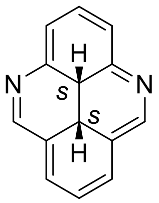
Figure 5. The proposed solution to the Rule 1b Kekulé problem is to assign the root distances for duplicate nodes to be the root distance of their attached atom, not that of the atom they duplicate. Digraph numbers are distances to the root assigned for each node. Duplicate nodes are in parentheses; bold numbers refer to ring-closure duplicates.

There is an additional aspect of Rule 1b we suggest revising. The original statement of Rule 1b includes an additional criterion ranking any duplicate node higher than any node that is not a duplicate node *for these purposes.*[17] This statement is omitted from BB 2013, ostensibly because it appears to be chemically irrelevant. However, it is important in the general statement of the rule topologically, and we suggest retaining it in CIP rules.

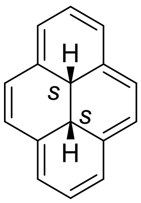
Thus, we propose the following new Rule 1b:

> ***Rule 1b (proposed): Lower root distance precedes higher root distance, where "root distance" is defined: (a) in the case of ring-closure duplicate nodes as the sphere of the duplicated atom; (b) in the case of multiple-bond duplicate nodes as the sphere of the atom to which the duplicate node is attached; and (c) in all other cases as the sphere of the atom itself.***

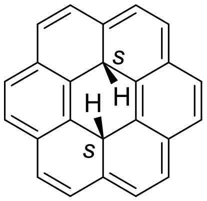
We note that an unresolved problem with Rules 1a and 1b involves the case of fully conjugated systems such as those shown in Figure 6 (VS215, VS217, and VS218). Each of these structures has at least two Kekulé representations, and yet they are not mancude-ring systems, so atomic number averaging is not formally called for by BB 2013 guidelines. Still, application of Rule 1a should not resolve the chirality. Rather, these descriptors must be decided by Rule 5, since the compound has two pseudoasymmetric centers. The solution to this issue, we propose, is to not focus on mancude systems at all. The essential feature is that there exists a fully conjugated cycle of double bonds of essentially any size.



VS215



VS217



VS218

Figure 6. Though not mancude-ring systems, these compounds have two or more equivalent Kekulé structures. They require atomic number averaging in Rule 1a and revised definition of root distance in Rule 1b in order to pass the decision to Rule 5, giving an (*s*,*s*) descriptor in each case.

Rule 2: *Higher atomic mass number precedes lower.*

The implementation problem in this case relates to comparisons where one atom has an isotope indicated and one does not, and also (again) when several alternative Kekulé structures are involved. The problem is that "mass number" is always an integer -- the sum of the number of protons and neutrons in the nucleus. This leaves open the question as to what to do in the case where an atom with isotope number indicated is compared with an atom of the same element that has no isotope number indicated, thus referring to a natural composition for the element. Conflicting examples are given in BB Section P-92.3 in relation to how to deal with this issue, where the term "mass number" is replaced with "atomic mass."

The consensus of our group was to recommend changing the language of Rule 2 to be specific, using exact isotopic mass when an isotope is indicated, and atomic weight when it is not. In addition, we recommend explicitly using a mass of 0 for all duplicate nodes, since if there is a mass issue related to a node, it will always be found first in relation to its real atom before its duplicated atom is ever checked in Rule 2. We propose:

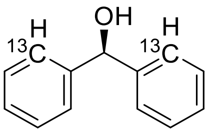
> ***Rule 2 (proposed): Higher mass precedes lower mass, where mass is defined in the case of a duplicate node as 0, an atom with isotope indicated as its exact isotopic mass, and in all other cases as the element's atomic weight.***

Atomic weights of the elements and isotopic abundances are IUPAC recommended values[18,19]. Using this modification, the example in BB 2013 P-92.3 still holds: ^{81}Br > Br > ^{79}Br. A second example in BB 2013 (p 1189 Example 2), however, places ^{125}I > I, even though natural iodine is 100% I-127; we consider this an erratum in BB 2013; the corrected version should read ^{125}I < I = ^{127}I. In addition, C (12.011) > ^{12}C (12.000),^{16}O (15.994) < O (15.999), and H(1.0079) > ^{1}H (1.0078).

The switch to atomic weight and exact isotope mass rather than integer mass number allows atoms of elements that are 100% one isotope naturally (^{9}Be, ^{19}F, ^{23}Na, ^{27}Al, ^{31}P, ^{45}Sc, ^{55}Mn, ^{59}Co, ^{75}As, ^{89}Y, ^{93}Nb, ^{103}Rh, ^{127}I, ^{133}Cs, ^{141}Pr, ^{159}Tb, ^{165}Ho, ^{169}Tm, ^{197}Au, ^{209}Bi, ^{231}Pa, and ^{232}Th) to be equivalent whether their isotope number is given explicitly or not, as is the case chemically. In addition, the elements Tc, Pm, Po, At, Rn, Fr, Ra, Ac, and all elements with atomic number > 92, have no natural abundance; their “atomic weight” found on the periodic table is just one of their integer isotope mass numbers, as though that isotope were 100% naturally abundant.

It may seem that the switch to exact isotope mass from integer isotope number might be difficult to implement, requiring access to a complete table of isotopes, but that is not the case. It turns out that we can use integer isotope mass numbers provided we take account of just four anomalies: ^{16}O, ^{52}Cr, ^{96}Mo, and ^{175}Lu. These four isotopes are the only ones that have exact masses slightly below their element’s atomic weight even though their mass number is above it. For example, the exact mass of ^{16}O is 15.994, which is below the element’s atomic weight of 15.999, even though its mass number (16) is higher. In practice, this is no problem. We simply use integer isotope numbers, but reduce these four by 0.1 for the purpose of setting priorities. For example, since the atomic weight of oxygen is 15.999, when ^{17}O is compared to O, there is no problem -- 17 > 15.999. But when ^{16}O is specified, we use 15.9 instead of its actual value of 15.994. This allows ^{16}O to have the required lower priority than “O” itself, with atomic weight 15.999. No other isotopes have this problem, and we can just use their unadjusted integer mass number as a surrogate for isotopic mass. In this way, there is never a need to check a table of exact isotope masses.

Once Rule 1b is passed, duplicate nodes should play no role in deciding stereochemistry. The exclusion of duplicate nodes by assigning their mass to be zero guarantees this. It also removes an issue similar to the one discussed for Rules 1a and 1b, that different arrangements of double bonds in conjugates systems must not affect the chirality. For example, isotopically labeled alcohol VS007, shown in Figure 7, is achiral. To ensure this result, we simply assign all duplicate node masses to be zero, allowing the mass difference to be carried only by their corresponding duplicated atom.



VS007

Figure 7. The priority of an isotopically labeled atom in an aromatic ring in relation to Rule 1b must not depend upon the Kekulé structure used to depict the compound. To assure this, we propose using zero for the mass of all duplicate nodes.

Rule 3: *When considering double bonds and planar tetraligand atoms, 'seqcis' = 'Z' precedes 'seqtrans' = 'E', and this precedes nonstereogenic double bonds.*

As mentioned above, generation of a complete digraph, including auxiliary descriptors, is required prior to Rule 3 (not just Rule 4a, as mentioned in BB 2013). This is because *seqcis* and *seqtrans* also describe double bonds that involve two *diastereomorphic* ligands on the same end (with stereodescriptors RR and SR, for example). Inverting about a plane changes the comparison (to *SS* and *RS*, in this case), but this change does not reverse priorities.

Placement of Rule 3 before Rule 4a ensures that only enantiomorphic (*seqCis* and *seqTrans*) comparisons involving double-bonds and cumulenes with an odd number of double bonds are left to consider in Rules 4 and 5. From an implementation point of view, application of Rule 3 is simply the comparison of two ligand pairs, one pair on each end of an alkene or cumulene.

Rule 4a: *Chiral stereogenic units precede pseudoasymmetric stereogenic units, and these precede nonstereogenic units.*

That is, (*R* or *S)* > (*r* or *s*), (*M* or *P*) > (*m* or *p*), and (*seqCis* or *seqTrans*) > (*seqcis* or *seqtrans*), and that all of these have higher priority than digraph nodes with no auxiliary descriptor. The purpose of Rule 4a is to ensure that all comparisons in Rule 4b and later are of the same general type: *R* vs. *S*, *M* vs. *P*, or *seqCis* vs. *seqTrans* in Rules 4b; *r* vs *s* or *m* vs. *p* in Rule 4c; *R* vs. *S* or *M* vs. *P* in Rule 5. In addition, application of Rule 4a guarantees that the lists of ranked descriptors that are being compared in Rule 4b are of equal length. Implementation of Rule 4a is straightforward and needs no further discussion.

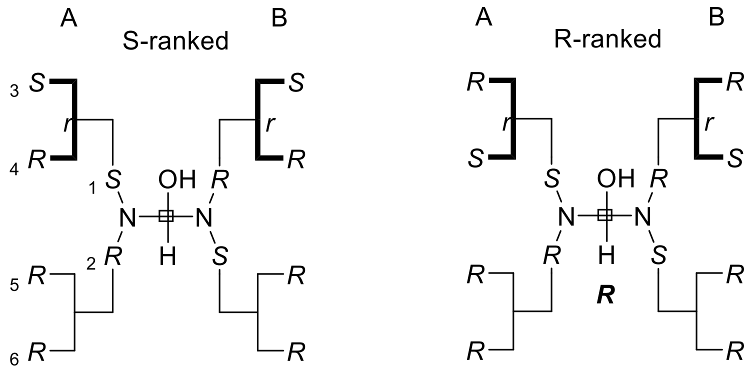
Rule 4b: *When two ligands have different descriptor pairs, then the one with the first chosen like descriptor pair has priority over the one with a corresponding unlike descriptor pair.*

Rule 4b is by far the most difficult rule to comprehend and implement. One simplification is that although Rule 4b, as stated in BB 2013, refers to all possible mixes of *R*/*S*, *M*/*P*, and *seqCis*/*seqTrans* descriptors, for implementation purposes, all auxiliary descriptors can be normalized by labeling them either *R* or *S*. For example, any of *R*, *M*, or *secCis* can be assigned *R* for the purpose of processing Rules 4b. In this way, all discussion can be expressed in terms of “equal” or “not equal” to a reference *R* or *S*, rather than “like” vs. “unlike”. It is critical that an implementation assign auxiliary descriptors involving double bonds -- *seqCis*/*seqTrans* and *M*/*P* -- to the sp^2 node closer to the root (or, alternatively, to both nodes equally). Otherwise the second phase of Rule 4b may fail.

The process for ranking ligands in Rule 4b is a more complex process than for previous rules, involving a two-stage process. First, the nodes are re-ranked in a way that may cross digraph branches. Second, the nodes are scanned in rank order for auxiliary descriptor similarity to both *R* and *S* reference descriptors. The higher priority ligand is the one with the highest score after *both of these comparisons* are made. We have found the easiest way to conceptualize this algorithmically is to “read” each of the four rankings (two for each ligand) as a series of 0s and 1s, which can be implemented as an integer, an array, or a bit set. In our examples, we will use integers, though our different implementations actually use different representations. The basic idea is to create four lists for each pair of ligands that can be compared together. If there is no winner, we go on to the next rule.

Ranking of ligand nodes starts with the creation of a new priority criterion for each node. This criterion must incorporate the full path from the root to this node, including all Rule 4a-priorities as well as the similarity or dissimilarity of the node’s descriptor to the reference. Pseudoasymmetric descriptors *r*, *s*, *m*, *p*, *seqcis*, and *seqtrans* are ignored in this process. As always, only previously identically-ranked nodes are resolved.

The example in Figure 8 illustrates the process used in Rule 4b for VS262. Note that branches change order, depending upon the reference. In this case, some of the branch orderings have already been set, due to a Rule 4a comparison, *r* vs. *ns.* The reading of the *S*-ranked ligand A, SRSRRR (read from centers 1-6, in order), is encoded as 101000 in base-2, giving a value of 40. Similar encoding gives 24 for *S*-ranked ligand B, 27 for *R*-ranked ligand A, and 43 for R-ranked ligand B. So ligand B, with a high score of 43, is given higher priority than ligand A, and the designator is *R*.

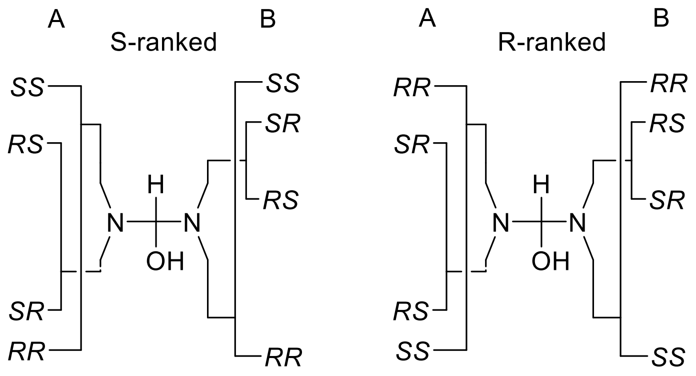


A: *SRSRRR*	B: *RSSRRR*	A: *SRRSRR*	B: *RSRSRR*
luluuu	*ulluuu*	*ullull*	*lulull*
$101000_2 = 40$	$011000_2 = 24$	$011011_2 = 27$	**$101011_2 = 43$**

Figure 8. Numerical rankings of ligands A and B in Rule 4b for VS262 requires using both *S* and *R* as the reference descriptor. Rule 4a has already sorted the main branches of the ligands, leaving only the bolded *r* branch for sorting. Ligand B, with the high score of 43, has the higher priority, and the descriptor is *R*.

The rationale behind Rule 4b becomes clear upon inspection of the algorithm. First, we are giving preference to ligands that have the most similarities closest to the root – both *RR* and *SS* (each encoded as the number 3 by one of the reference options) will be selected in preference to either *RS* or *SR* (each of which will be encoded as 2 by one of the options and 1 by the other). And yet, identical readings or opposite readings overall – such as *RRSRRS* vs. *SSRSSR*, which need to pass on to Rule 5 – are not distinguished, as they will be encoded as the same two numbers with one or the other reference options.

An important facet of the ranking in Rule 4b is that identically-ranked nodes can come from different branches of a ligand. So, for example, in Figure 9 we have a case (VS242), which has four “highest-ranked nodes” that must be sorted by *R* and *S* as a group. The result that the two ligands cannot be distinguished in Rule 4b.



A(*S*),B(*S*): *SSRRSRSR* A(R),B(R): *RRSSRSRS*

lluululu *lluululu*

$11001010_2 = 202$ $11001010_2 = 202$

Figure 9. In this case (VS242), ranking by references R and S involve crossing digraph lines, because in each case all four of the branches have “highest priority”. The ligands are not distinguished in Rule 4b.

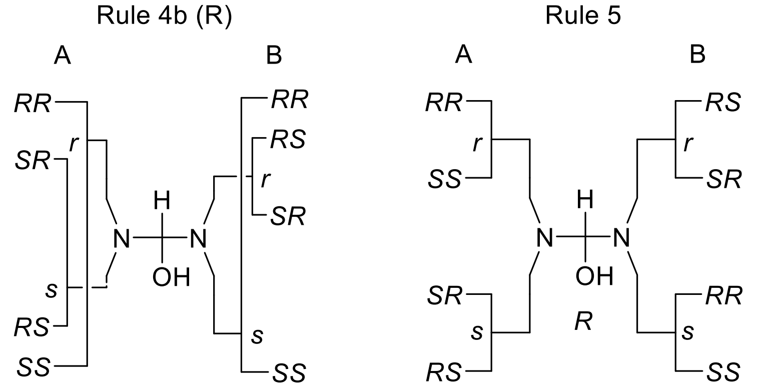
Note that all of the “rules for sorting” discussed in other works, such as, “The first step of this procedure is the critical choice of the first descriptor for each ligand,”[20] are unnecessary in terms of code implementation. These “critical choices,” such as determining the descriptor associated with the highest-ranking node or the descriptor that occurs the most in the set of highest-ranking nodes or sequentially evaluating both *R* and *S* as references, simply fall out of the mathematics of the numerical rankings described above. Thus, no such critical choices need to be implemented, though if they are, they might speed the processing.

Rule 4c: *‘r’ precedes ‘s’ and ‘m’ precedes ‘p’.*

If a center passes Rule 4b undecided, it means that there are only three possibilities: (1) There is no ligand chirality; (2) two or more ligands have identical chirality descriptors; or (3) the two ligands each have sub-branches with opposite chirality. Rule 4c takes care of case (3), where we assign *r* over *s*, and *m* over *p*. The implementation of Rule 4c is a straightforward extension of the implementation of Rule 4a.

Rule 5: *An atom or group with descriptor ‘R’, ‘M’, or ‘seqCis’ has priority over its enantiomorph ‘S’, ‘P’, or ‘seqTrans’.*

Rule 5 does a final check for enantiomorphic ligands. Note that implementation of Rule 5 is not just a check of the lists generated using the procedure of Rule 4b, as priorities may have changed after application of Rule 4c. Consider the digraph in Figure 10, which shows the digraph of Figure 9, ranked by *R*- and *S*- reference for both Rule 4b and Rule 5. Here we see that after application of Rule 4c, the sorting is changed, and Ligand A has preference over Ligand B by Rule 5.

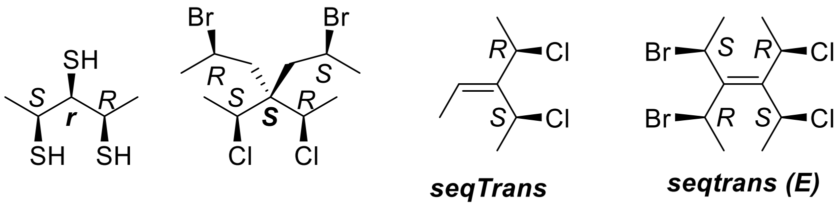


A(R),B(R): *RRSSRSRS*
A(*S*),B(*S*): *SSRRSRSR*

A(*R*): *RSRSRSSR* B(*R*): *RSRSSRRS*
A(*S*): *SRSRSRRS* B(*S*): *SRSRRSSR*

Figure 10. The same digraph as in Figure 9, for VS242, here also indicating the four pseudoasymmetric centers. In this case, sorting and listing of ligands gives different results after application of Rule 4c, which sorts both main ligand branches by *r* > *s*. The *R*-reference, Ligand A, with *RSRS**R***… has higher priority than B, with *RSRS**S**….* Due to the fact that both asymmeric ligands are their own enantiomorph, the final designation is *R* rather than *r.*

If all ligands are finally distinguished after application of Rule 5, an additional test should be done to count the number of pairs of enantiomorphic ligands. The final descriptor will be *r*/*s*, *m*/*p*, or, in the case of akenes, *seqCis*/*seqTrans,* if and only if this number is one (Figure 11), otherwise it will be *R*/*S*, *M*/*P*, or *seqcis/seqtrans (Z/E).*



VS236 VS244 VS229 VS245

Figure 11. (VS236) A typical center decided by Rule 5, with one pair of enantiomorphic ligands; (VS244) a doubly enantiomorphic center decided by Rule 5 to be *S*; (VS229) an alkene with one pair of enantiomorphic ligands is given the descriptor *seqTrans* (indicated as '*e*' in the validation suite) not *seqtrans* (*E*) because, upon reflection through a plane, it changes descriptor; (VS245) an alkene with two pairs of enantiomorphic ligands is given the usual *seqtrans (or E)* descriptor. Similar tests need to be made for both odd- and even-cumulenes.

In terms of implementation, the criterion for pseudoasymmetry at a tetrahedral center is that, when comparing otherwise identical ligands, there is an odd number of pairs that reverses priority when comparing like/unlike sequences using an *S* reference vs. using an *R* reference. So, for example, in Figure 11a, the right-hand *R* ligand has higher priority with the *R*-reference, but the left-hand *S* ligand has higher priority with the *S*-reference. The priority switches, and we have the normal outcome for Rule 5 – pseudoasymmetric. But in Figure 11b, priority switches twice, so the result is asymmetric.

In Figure 10, we have a different story. Sorting by *R* gives A > B. But sorting by *S* also gives A > B – no switch! A naïve application of Rule 5, only checking the two *R*-reference like/unlike lists, would have assigned *r* to that center. However, this center is asymmetric, not pseudoasymmetric, because each ligand *is its own enantiomorph*. The ultimate descriptor will be *R,* not *r*. The analysis is the opposite for alkenes and even-atom cumulenes (Figure 11 c and d). Thus, if the pair on only one end of the alkene reverses priority when using the *S* reference vs. using the *R* reference (Figure 11 c), then the result is the asymmetric *seqCis* or *seqTrans*; if neither or both pairs reverse priority (Figure 11d), then the result is the pseudoasymmetric *seqcis* or *seqtrans*.

Alternatively, one can simply repeat the entire process of assigning descriptors using the structure that is reflected through a mirror plane. If the resulting descriptor is reversed, then the final descriptor is *R/S*, *M/P*, or *seqCis*/*secTrans*; if not, then *r/s*, *m/p*, or *seqcis/seqtrans*.

A Proposal for Rule 6: Spiro and other axially-symmetric compounds. Early on in the development of the CIP system,[3] it was recognized that certain cases involving C_2, D_2, and C_3 point groups require additional consideration for assignment of stereodescriptors; an S_4 case was described later[3] (Figure 12). Specifically, an algorithm must distinguish between both enantiomers of compounds VS285, VS281, and VS283, and yet still deliver no stereodescriptors for cubane (VS009) and the S_4 compound VS012. Only simple spiro structure VS285 is mentioned in BB 2013, in section P-93.5.3.

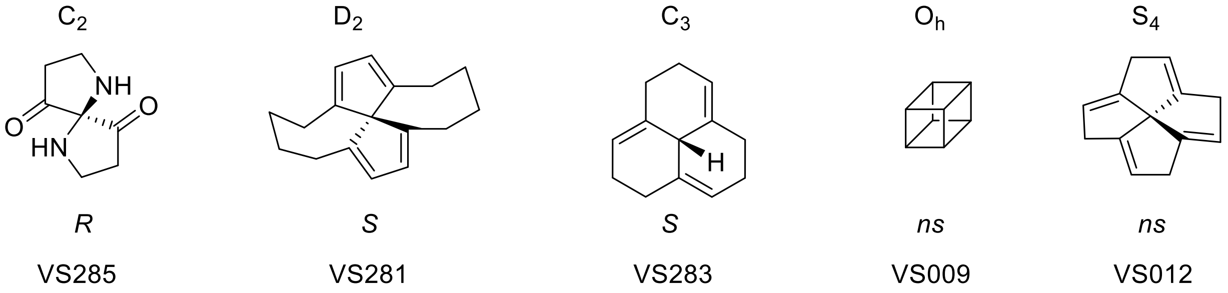


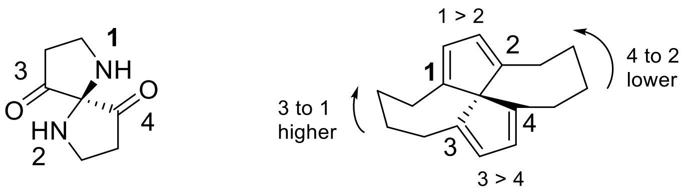
Figure 12. Five compounds for which the stereodescriptor is decided by Rule 6, along with their point group and final designation.

A relatively simple additional ninth Sequence Rule -- “Rule 6” -- takes care of all such cases:

Rule 6 (proposed): An undifferentiated reference node has priority over any other undifferentiated node

Cases to be considered by Rule 6 are identified by having, after application of Rule 5, two pairs of identical ligands or three or four identical ligands. Thus, in the first case in Figure 12, we have two identical amino ligands and two identical keto ligands; in the second and fifth cases, we have four identical ligands. In the third and fourth cases, we have three identical ligands.

The solution for all such cases (first proposed in 1966[2] for specific examples) is simply to select one node of any one of the undistinguished ligands for promotion to higher rank (Figure 13). Basically, by arbitrarily breaking the symmetry in this way, the problem is immediately resolved upon inspection of the digraph. We note for double spiran VS281, the result of application of the proposed Rule 6 generates the opposite descriptor to the one assigned previously.[2,3] We believe this is a due to either a misassignment or a typographical error.



VS285 *R* VS281 *R*

Figure 13. Two examples of the analysis of compounds by Rule 6, which sets Node 1 to be higher priority than Node 2. This single change decides also the priority 3 > 4, due to the presence of a ring connection from Node 3 back to Node 1 and from Node 4 back to Node 2.

Two outcomes of Rule 6 are possible:

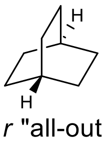
(a) After application of Rule 6, there are still two undistinguished ligands. Such will be the case, for example, with simple acyclic compounds, such as CH_2Cl_2 or $CHCl_3$.The center remains without descriptor.

(b) After application of Rule 6, all ligands are distinguished. The center receives a descriptor. Such will be the case only for compounds that have rings that involve the root atom and three or more ligands. A full application of Rule 6 tests all possible promotions, though this is necessary only for certain symmetries. Any matching *R* and *S* pairs are ignored; if a descriptor remains, it is valid.

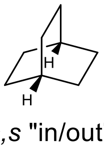
In terms of implementation of Rule 6, it may be noted that a descriptor will only be assigned by Rule 6 when there are three or more paths through the structure leading back to the root atom. Thus, a simple test that indicates fewer than three duplicate nodes with root distance 0 is sufficient for skipping Rule 6.

Note that Rule 6 must be applied in the determination of all auxiliary descriptors as well as in the final root-node determination, as for all other Sequence Rules, because it is possible for an auxiliary descriptor to result from C_3 symmetry (e.g. VS300). Even if it is determined that Rule 4a - 5 can be skipped due to the lack of auxiliary descriptors, Rule 6 should not be skipped without further evidence that Outcome (a) is the only possibility.

Rule 6 also allows assignment of stereodescriptors for centers with axial symmetry, such as those involving allenes and biphenyls. No further special consideration of high-symmetry compounds or groups is necessary. Interestingly, an easy extension of Rule 6 solves the heretofore unresolved issue of “in/out” stereochemistry such as shown in Figure 14. These compounds can be treated successfully simply by taking into account the temporary re-assignment of auxiliary descriptors after each promotion of a node. However, we stop short of recommending this modification at this time, as it only applies to rather esoteric structures mostly of only theoretical interest. Its implementation would result in additional (unnecessary) descriptors for common all-*r* “all-out” structures, many of which are known compounds and have already been arbitrarily assumed to have no stereodescriptors.



r,r "all-out"



s,s "in/out"

Figure 14. Stereodescriptors for (*r,r)*- and *(s,s)*-bicyclo[2.2.2]octane arise from allowing a recalculation of auxiliary descriptors in the processing of Rule 6. Similar descriptors are found for all-out and in/out tetrahedrane, cubane, adamantane, and related structures.

Implementations

The CIP rules defined in BB 2013 and all relevant publications were analyzed in the development of algorithmic procedures for generation of stereodescriptors. Following this analysis, it was concluded that several refinements are necessary for the CIP rules to ensure generation of the descriptors described in BB 2013.

Specific software implementations are listed in Table 1, with detailed implementation characteristics given in Table 2. The four implementations were developed independently. Two are open-source; all are freely available as executable versions. As can be seen, all implementations have limitations. However, outside of a few specific limitations, all four implementations give the same results for all compounds in the validation set and agree in all cases with descriptors provided in BB 2013 or its errata.

Table 1. Software Availability, Features, and Limitations

Jmol (RMH, https://jmol.sourceforge.net, https://chemapps.stolaf.edu/jmol/jsmol/cip.htm)	
Availability	Open-source; both Java and JavaScript versions (v. 14.29.15)
Features	Visual; interactive 3D models; easy building and adaptation of models (click atom to invert chirality); can be made part of any web page in the form of a JSmol app, run on a server using JmolData.jar, or run as a stand-alone Java application as Jmol.jar. The Jmol implementation of CIP rules leverages Jmol's substantial features in the area of 3D molecular data handling and visualization, including Jmol conformational SMARTS[21] searching[22] to effectively identify potential atropisomer and helicene chirality. Scripting proves easy annotation of 3D interactive models with descriptors (*LABEL %[chirality]*) and the CIP Rules leading to those (*LABEL %[cipRule]*), atom selection (*select chirality == "R"*), and structured return of CIP information (*calculate chirality; print _M.CIPInfo*). 2D-to-3D and SMILES-to-3D conversion are provided by the NCI/CADD Chemical Resolver.[23]
Balloon (MJV, http://users.abo.fi/mivainio/balloon)	
Availability	Closed-source, freeware; C++ (v1.6.6)
Features	Balloon is a command-line program for 2D to 3D conversion using distance geometry and conformational sampling using a genetic algorithm. The chirality perception code, amounting to around 2K lines, is implemented in C++. The algorithm searches the input molecular graph for possible stereogenic centers and generates a digraph for each center. Conformational sampling is possible with input chirality retained.
ACD/Name and ACD/ChemSketch (AY, http://www.acdlabs.com/products/draw_nom)	
Availability	Closed-source; freeware and commercial versions; Delphi (v. XXX)
Features	ACD/Name generates chemical names for structures indicating stereodescriptors in names in accord with IUPAC recommendations. Specifies stereoconfiguration with E/Z, R/S and P/M descriptors supporting tetrahedral and trigonal pyramidal centers, cumulene and atropisomer axial chirality. Supports both 2D structures with stereobonds and 3D structures. ACD/ChemSketch is a desktop full featured chemical drawing program allowing stereodescriptor generation using ACD/Name's stereochemistry procedures.
Centres (JWM, http://www.github.com/simolecule/centres)	
Availability	Open-source; 2-Clause BSD, Java (v. XXX)
Features	Centres is a library specifically for CIP labelling. It provides an abstract API allowing integration on top of multiple back-end toolkits (CDK[24–27], OPSIN[28], JChem Base[29]). The underlying toolkit handles the file processing and describes the stereochemistry as a data structure. Centres then generates CIP descriptors.

Table 2. Features and limitations of the implementations

Features		Jmol	Balloon	ACD ChemSketch	Centres	Examples
Software Details	Language	Java/JS	C/C++	Delphi	Java	
	Free to Use	Y	Y	Y[1]	Y	
	Open-Source	Y	N	N	Y	
Priority Rules	Rule 1a	Y	Y	Y	Y	VS013-170
	Rule 1b[2]	Y	Y	Y	Y	VS171-174
	Rule 2[2]	Y	Y	Y	Y	VS175-187
	Rule 3	Y	Y	Y	Y	VS188-195,246-248
	Rule 4a	Y	Y	Y	Y	VS249-251,269-272,277,278
	Rule 4b	Y	Y	Y	Y	VS196-204,252-263,265,268-272,279
	Rule 4c	Y	Y	Y	Y	VS273-279,296-298
	Rule 5	Y	Y	Y	Y	VS205-300
	Rule 6[2]	Y	Y	Y	Y	VS280-300
Features	no stereo	Y	Y	Y	Y	VS001-009,012
	seqcis/seqCis distinction	Y[3]	Y	Y[3]	Y	VS229,246-248, 299
	large cyclic π systems	Y	Y[4]	Y	Y	VS215-218
Geometry Support	odd-cumulene (chirality axis)	Y	Y	Y	Y	VS078,079,120,141,144,166,231,232,243,287
	even-cumulene (planar)	Y	Y	Y	Y	VS063,118,135,154,164
	atropisomer (chirality axis)	Y	N	Y	Y	VS023,055,057,072,073,086,158
	helicial chirality	Y	N	N	N	VS010,011
	fullerene chirality	N	N	N	N	
	chirality plane	N	N	N	N	
	inorganic configugation index	N	Y	N	Y	

[1] Structure size restriction in freeware

[2] proposed Sequence Rules revision or addition

[3] reports z/e for seqCis/seqTrans, Z/E for seqcis/seqtrans

[4] uses aromaticity flags in lieu of full detection

Methods - Validation

Key to our development process was the production of a robust validation suite that could be used by any developer to ensure that a CIP implementation follows the rules to whatever degree that software claims to implements them. Though scattered attempts have been made to develop such model collections, primarily for in-house testing of various software packages, the supplemental material to this paper includes the first fully tested openly available collection of models that can test the full range of issues presented in the CIP Sequence Rules. The 300 models are in annotated SDF format, providing both 2D and 3D models, as well as SMILES.[22] Standard SDF data annotations provide reported (or corrected) descriptors for all relevant models in Chapter 9 of BB 2013, along with a significant number of additional “challenges” for developers. Accompanying this collection is a table that correlates the specific test models with specific sections of BB 2013 Chapter 9 as well as specific aspects of the CIP rules that are targeted by this particular test.

The four implementations generate the same stereodescriptors for each model in the validation set. The analysis largely agrees with the descriptors provided in BB 2013, differing only in relation to a few already-identified minor analysis or typographical errors in the published version of BB 2013.

Conclusions

We have developed and implemented specific algorithms that handle all nuances of all nine of the Sequence Rules for tetrahedral centers and double bond stereochemistry, two of which are open-source Java applications (Centres and Jmol, which also has a JavaScript equivalent). All are freely available in binary format. We have developed a robust testing suite for algorithms that implement the CIP Sequence Rules. This suite covers a wide variety of possible issues that might arise in the course of any algorithm development.

In addition, in the process of this work, we have identified and addressed a number of issues with the eight Sequence Rules as presented *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013* (1a, 1b, 2, 3, 4a, 4b, 4c, and 5) and propose simple solutions, including a new Rule 6. In particular, we provide a definitive method of implementing Kekulé averaging in Rules 1a, 1b, and 2. We propose a simple solution to working with isotope masses in Rule 2. We argue for the necessity for fully elaborated auxiliary descriptors prior to Rule 3 and describe the order in which they must be generated. We demonstrate a simple way to implement the *like/unlike* analysis in Rules 4b and 5. Most significantly, we propose three modifications to the Sequence Rules:

Proposed New Rule 1b: Lower root distance precedes higher root distance, where "root distance" is defined: (a) in the case of ring-closure duplicate nodes as the sphere of the duplicated atom; (b) in the case of multiple-bond duplicate nodes as the sphere of the atom to which the duplicate node is attached; and (c) in all other cases as the sphere of the atom itself.

Proposed New Rule 2: Higher mass precedes lower mass, where mass is defined in the case of duplicate nodes as 0, atoms with isotope indicated as their exact isotopic mass, and in all other cases, as their element's atomic weight.

Proposed New Rule 6: An undifferentiated reference node has priority over any other undifferentiated node.

Acknowledgments

The authors thank International Union of Pure and Applied Chemistry (IUPAC) for permission to use and provide chemical structures from BB 2013 as part of the validation suite. RMH acknowledges Tram Thi Bich Bui for her early work with the structure database. AY thanks the European Commission Taxation and Customs Union for funding ACD/Labs to digitize chemical structures and names from BB 2013 into SDF format.

References

1. Cahn RS, Ingold CK, Prelog V. The specification of asymmetric configuration in organic chemistry. *Experientia*. 1956;12(3):81-94. doi:10.1007/BF02157171
2. Cahn RS, Ingold C, Prelog V. Specification of Molecular Chirality. *Angew Chemie Int Ed English*. 1966;5(4):385-415. doi:10.1002/anie.196603851
3. Prelog V, Helmchen G. Basic Principles of the CIP-System and Proposals for a Revision. *Angew Chemie Int Ed English*. 1982;21(8):567-583. doi:10.1002/anie.198205671
4. CHAPTER P-9. Specification of Configuration and Conformation. In: *Nomenclature of Organic Chemistry*. Cambridge: Royal Society of Chemistry; 2013:1156-1292. doi:10.1039/9781849733069-01156
5. Mata P, Lobo AM, Marshall C, Johnson AP. The CIP Sequence Rules: Analysis and proposal for a revision. *Tetrahedron: Asymmetry*. 1993;4(4):657-668. doi:10.1016/S0957-4166(00)80173-1
6. Mata P, Lobo AM, Marshall C, Johnson AP. Implementation of the Cahn-Ingold-Prelog System for Stereochemical Perception in the LHASA Program. *J Chem Inf Comput Sci*. 1994;34(3):491-504. doi:10.1021/ci00019a004
7. Mayfield J, Lowe D, Sayle R. CINF 17: Comparing Cahn-Ingold-Prelog Rule Implementations: The need for an open CIP. https://www.slideshare.net/NextMoveSoftware/cinf-17-comparing-cahningoldprelog-rule-implementations-the-need-for-an-open-cip. Accessed April 2, 2018.

8. Corrections, Revisions and Extensions for the Nomenclature of Organic Chemistry - IUPAC Recommendations and Preferred Names 2013 (The IUPAC Blue Book). https://iupac.org/projects/project-details/?project_nr=2015-052-1-800.

9. BlueObelisk. https://blueobelisk.github.io.

10. Jmol. http://jmol.sourceforge.net. Accessed April 2, 2018.

11. May JW. Cheminformatics for genome-scale metabolic reconstructions (doctoral thesis), 2015. https://doi.org/10.17863/CAM.15987

12. ACD/ChemSketch Freeware. 2018. http://www.acdlabs.com/products/draw_nom/draw/chemsketch/.

13. Vainio MJ, Johnson MS. Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model*. 2007;47(6). doi:10.1021/Ci6005646

14. Puranen JS, Vainio MJ, Johnson MS. Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. *J Comput Chem*. 2010;31(8). doi:10.1002/jcc.21460

15. Study Aids Chemistry220js. http://ursula.chem.yale.edu/~chem220/chem220js/StudyAids.html#Stereochemistry. Accessed March 31, 2018.

16. Mancude-ring systems. https://goldbook.iupac.org/html/M/M03695.html. Accessed March 31, 2018.

17. Custer RH. Mathematical statements about the revised CIP-system. *Match*. 1986;21(3):3-31. http://match.pmf.kg.ac.rs/electronic_versions/Match21/match21_3-31.pdf.

18. Meija J, Coplen TB, Berglund M, et al. Atomic weights of the elements 2013 (IUPAC Technical Report). *Pure Appl Chem*. 2016;88(3). doi:10.1515/pac-2015-0305

19. Rosman KJR, Taylor PDP. Isotopic compositions of the elements 1997 (Technical Report). *Pure Appl Chem*. 1998;70(1):217-235. doi:10.1351/pac199870010217

20. Mata P, Lobo AM. The Cahn, Ingold and Prelog System: eliminating ambiguity in the comparison of diastereomorphic and enantiomorphic ligands. *Tetrahedron: Asymmetry*. 2005;16(13):2215-2223. doi:10.1016/j.tetasy.2005.05.037

21. Daylight Theory Manual. http://www.daylight.com/dayhtml/doc/theory/index.html. Accessed March 31, 2018.

22. Hanson RM. Jmol SMILES and Jmol SMARTS: Specifications and Applications. *J Cheminform*. 2016;8(1):50. doi:10.1186/s13321-016-0160-4

23. Ihlenfeldt WD, Voigt JH, Bienfait† B, Oellien F, and Nicklaus MC, Enhanced CACTVS Browser of the Open NCI Database. *J. Chem. Inf. Comput. Sci.* 2002;42 (1), pp 46-57. doi: 10.1021/ci010056s

24. Willighagen EL, Mayfield JW, Alvarsson J, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform*. 2017;9(1):33. doi:10.1186/s13321-017-0220-4

25. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr Pharm Des*. 2006;12(17):2111-2120. doi:10.2174/138161206777585274

26. May JW, Steinbeck C. Efficient ring perception for the Chemistry Development Kit. *J Cheminform*. 2014;6(1):3. doi:10.1186/1758-2946-6-3

27. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci*. 2003;43(2):493-500. doi:10.1021/ci025584y

28. Lowe DM, Corbett PT, Murray-Rust P, Glen RC. Chemical Name to Structure: OPSIN, an Open Source Solution. *J Chem Inf Model*. 2011;51(3):739-753. doi:10.1021/ci100384d

29. JChem Base, Commercial Java library provided by ChemAxon. https://chemaxon.com/products/jchem-engines. Accessed May 9, 2018.