Twisted SMILES? Proposed Conformation Codes

John W. McLeod and Daniel J. Peters

Defence R&D Canada

Presentation to MUG'03, February 25-28, 2003

Conference sponsored by
Daylight Chemical Information Systems Incorporated

Submitted Monday February 10, 2003
Minor fixes submitted February 13&18, 2003

Presentation scheduled for
Tuesday February 25, 2003 at 2:45PM

Chemical nomenclature is easy.

Three Steps to Coding a Molecule

- To specify constitution:
  - Use an atom numbering
- To specify configuration:
  - Use cosets
- To specify conformation:
  - Use dihedral angles

It's that simple.

Three Steps to Coding a Molecule (continued)

- Atom order
  - fixes molecule's bond pattern
- Coset
  - fixes atom's neighborhood
- Dihedral angle
  - fixes bond's double-neighborhood

One implication is strict order:
Constitution → Configuration → Conformation

Basic Limitation (“easy” clarified)

- The "easy" claim refers to chemical nomenclature applied to a valence, graph, model. That is:
  - Structures requiring a molecular orbital description are still problematical; but,
  - If you can build a molecule from a modeling kit, you can name it

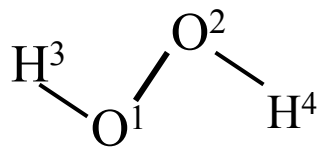
We didn’t say *chemistry* was easy

Atom order fixes molecule's bond pattern

- Any practical code for constitution must involve the following two things
  - An explicit or implicit numbering 1,2,…,n of all the atoms in the molecule; and,
  - A representation of the bonds joining the atoms

Atom order fixes molecule's bond pattern (continued)

- The *adjacency matrix* orders the atoms down its main diagonal
- Unintuitive codes result



$$\begin{pmatrix} O & 1 & 1 & 0 \\ 1 & O & 0 & 1 \\ 1 & 0 & H & 0 \\ 0 & 1 & 0 & H \end{pmatrix}$$

Possible codes:

O1101O0110H0010H ← read the matrix in row major order

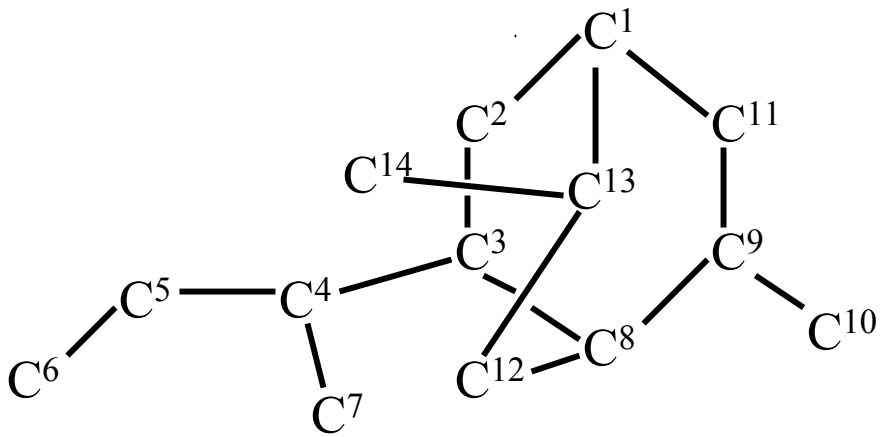
O2H2:2,3/4 ← list atoms-connected-to-first/atoms-connected-to-second …

Atom order fixes molecule's bond pattern (continued)

- SMILES orders the atoms from left to right in the linear code
- This is equivalent to diagonally ordering atoms in a matrix

C12C C(C(C C)C) C(C(C)C1) CC2C

1 2 3 4 5 6 7 8 9 10 11 12 13 14

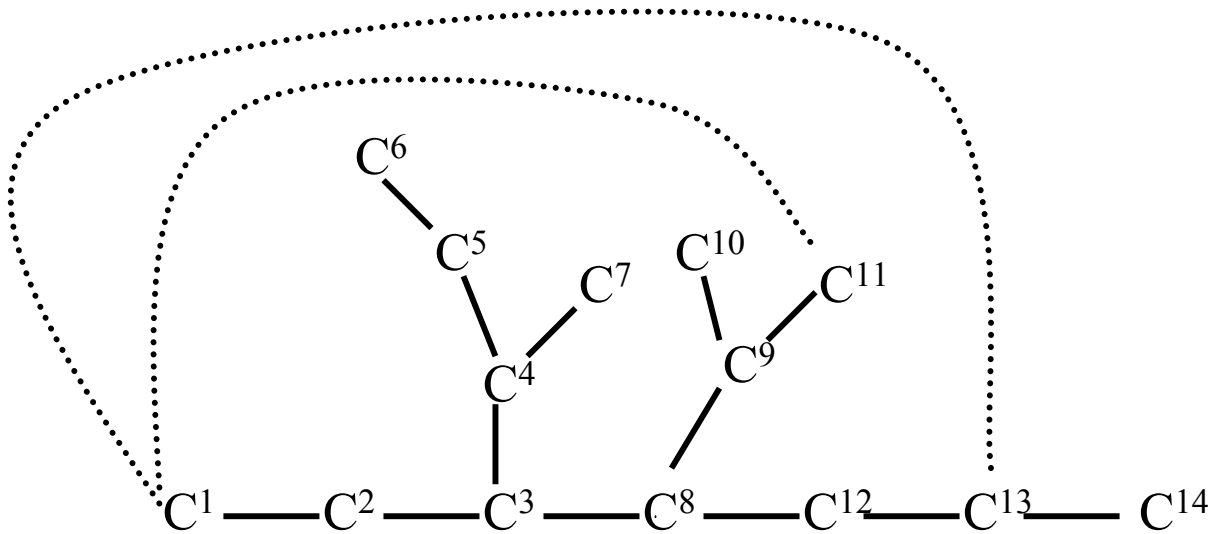


In fact, there is a natural correspondence:
SMILES ←→ Adjacency Matrix

Atom order fixes molecule's bond pattern (continued)

- Best SMILES codes use a *spanning tree* from the molecule, laying it out in a Swiss Army Knife pattern

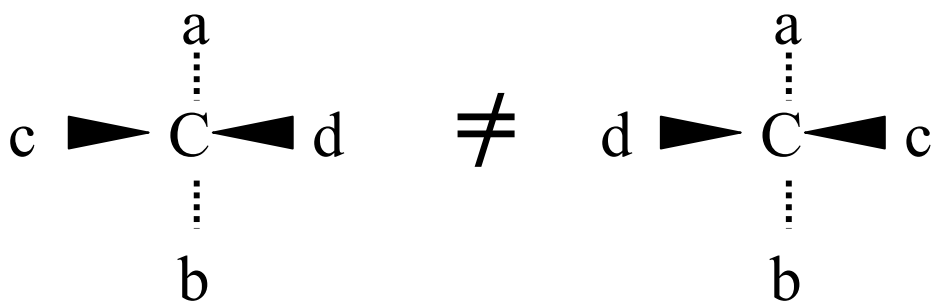
C12C C(C(C C)C) C(C(C)C1) CC2C
1 2 3 4 5 6 7 8 9 10 11 12 13 14



Main path is the "handle"; "tools" fold left to right,
each branch announced by a parenthesis.
This gives a branch-first, depth-first, left-first walk.

Configuration fixes atom's neighborhood

- Configuration
  - Considers each atom's neighborhood
  - Uses constitutional atom order to locally order neighbors (i.e. ligands)
  - Detects distinct geometries for some neighborhoods; central atom then an *asymmetric center*



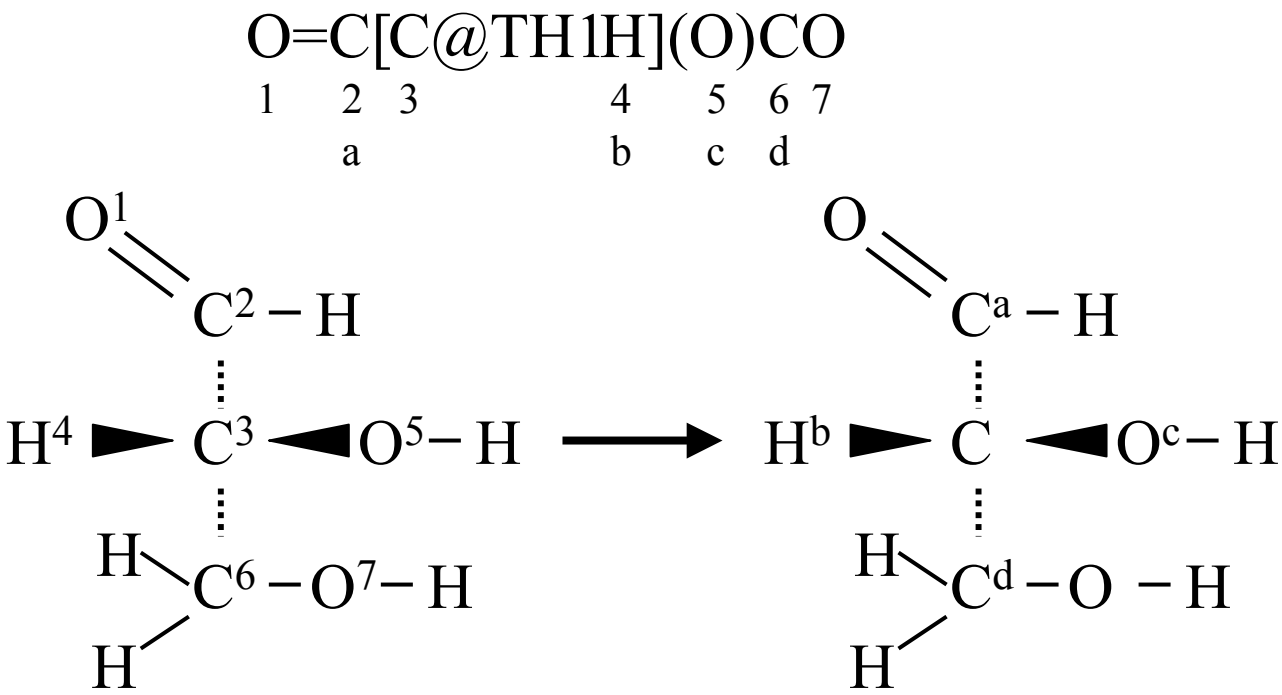
Configuration fixes atom's neighborhood (continued)

- Configuration is *not* about
  - Twist around bonds, or
  - Topological priority ordering of ligands, or
  - Clockwise/counter-clockwise sequence of ligands

Today, failure to correctly separate aspects of configuration from those of constitution and conformation is the most important obstacle to progress in chemical nomenclature.

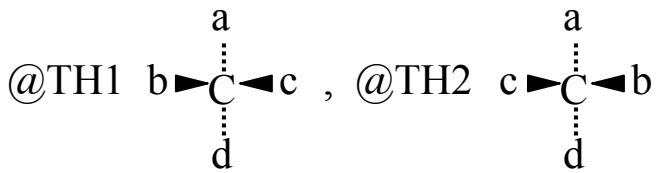
Configuration fixes atom's neighborhood (continued)

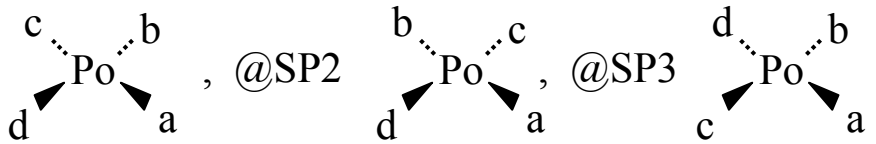
- General chiral specification
  - Orders the ligands a,b,c,… around an atom in accordance with the ordering 1,2,3,… of all the atoms in the molecule

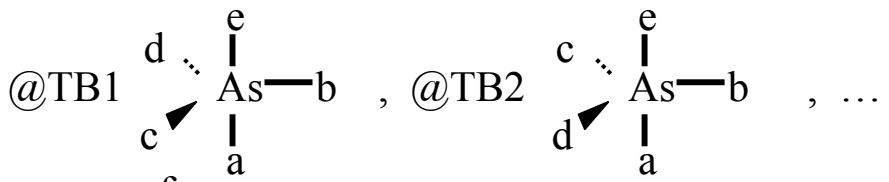


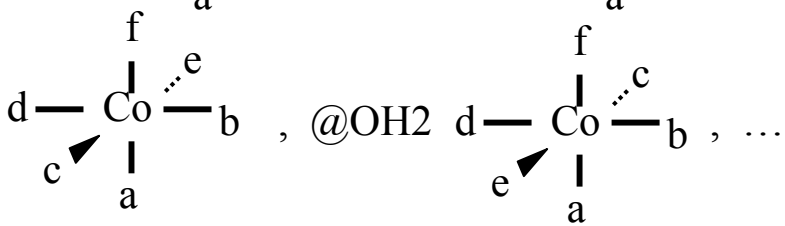
This, and not absolute priority, is the correct way to order an atom's ligands.

Configuration fixes atom's neighborhood (continued)

- Different cosets correspond to distinct ways of distributing the ordered ligands around asymmetric centers' neighborhoods, interpreted as geometric figures
  - Tetrahedron: @TH1 

  - Planar square: @SP1 

  - Trigonal bipyramid: @TB1 

  - Octahedron: @OH1 

  - …

Configuration (summary)

- General Chiral Specification is a complete, mathematically sound, solution to the configuration coding problem

Configuration and stereo numbers

- General Chiral Specification fulfills the promise of stereo numbers
  - Alfred P. Feldman (1959): interprets @TH? cosets as bits, concatenates these into binary *stereo numbers*
  - James G. Nourse (1979): *configuration symmetry group* resolves reduced chirality for @TH? stereo numbers
  - McLeod (1976, 1986): complete mathematical theory for coset-based configuration codes (references next screen)
  - SMILES (1988): Extension to all possible @XX? cosets, complete integration into constitution using atom decoration

Configuration and stereo numbers (continued)

- McLeod, John W. “A Graph Theoretic Model for Configurational Isomerism”. Thesis B.Sc.(Honours, Mathematics), Acadia University, Wolfville, Nova Scotia, Canada, April 1, 1976.
  - The complete theory from an uncompromisingly mathematical viewpoint
- McLeod, John W. “Stereo Numbers, Cosets, and the Configuration Symmetry Group”. J. Chem. Inf. Comput. Sci. 1986, **26**; 77-83.
  - This time with pictures!

Conformation fixes bond's double-neighborhood

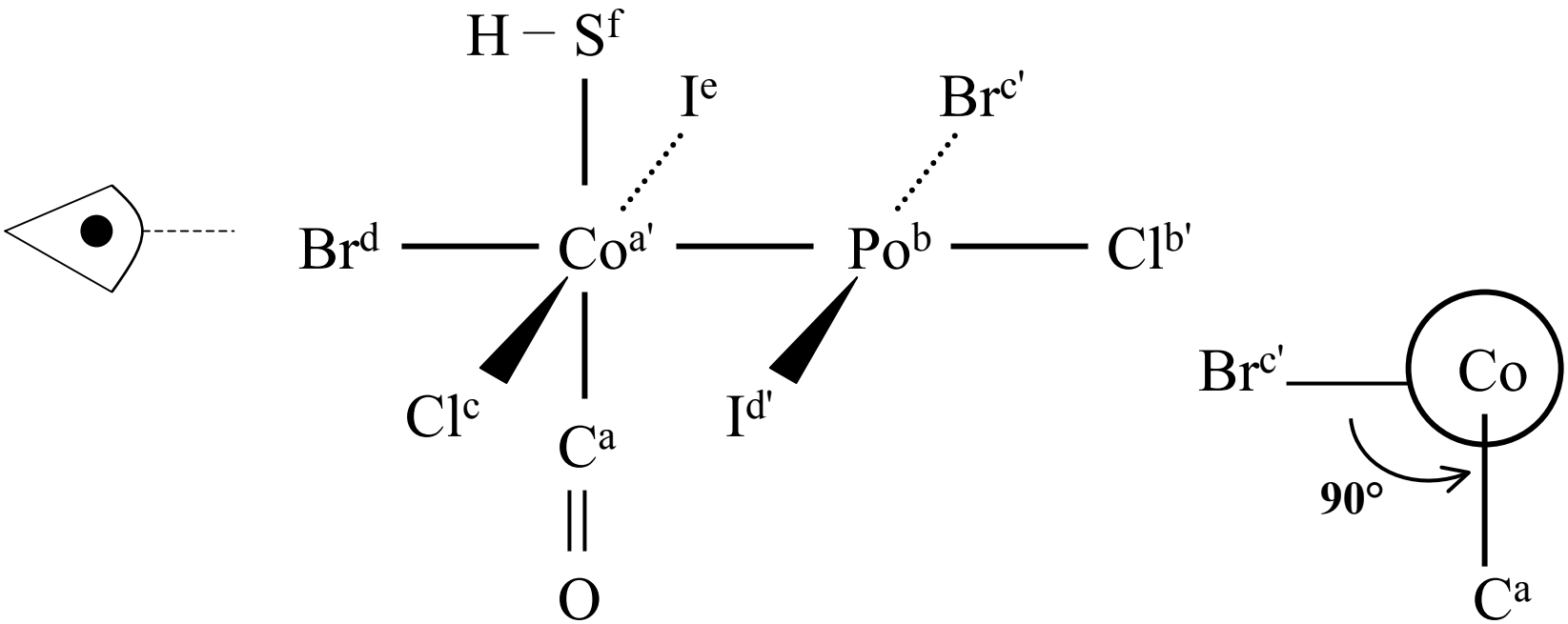
- Conformation
  - Considers a bond's double-neighborhood
  - Uses constitutional atom order to locally order ligands of both neighborhoods
  - Uses the neighborhoods' two coset (general chiral specification) codes to identify which ligands are coaxial with the bond
  - Measures the dihedral angle between the first non-coaxial ligands of the two neighborhoods
  - Assigns this dihedral angle to the bond as a conformation code

All of this is illustrated on the next slide.

Conformation fixes bond's double-neighborhood (continued)

O=C[Co@OH1] ([-@90.0] [Po@SP2] (Cl)(Br)I)(Cl)(Br)(I)S

1	2	3			4			5	6	7	8	9	10	11
	a				b						c	d	e	f
		a'						b'	c'	d'				



Conformation fixes bond's double-neighborhood (continued)

- Interpretation of previous slide
  - Double neighborhood of Co-Po bond is Co's neighborhood plus Po's neighborhood
  - SMILES-order generates a,b,c,d,e,f order for Co's ligands and a',b',c',d' order for Po's
  - Co's ligand b, the Po, is coaxial with the bond since it is attached to it; so ligand d, the Br, is coaxial too since ligands b and d are opposed in @OH1; so ligand a, the C, is the first non-axial ligand of Co
  - Po's ligand a', the Co, is attached to the bond and ligand b', the Cl, is opposed to it in @SP2; so ligand c', the Br is the first non-axial ligand of Po
  - Replacement syntax [-@90.0] for the Co-Po bond means that the dihedral angle from Br (far) counter-clockwise to C (near) is 90 degrees

Conformation fixes bond's double-neighborhood (continued)

Phew!

That last sequence simply demonstrated
that a local conformation strategy (analogous
to that of SMILES for local configuration)
naturally yields pairs of knobs we can
use to twist the Tinker-toy around its sticks

We now address the problem of decorating
the bonds with their dihedral angle codes

Conformation code syntax

- Syntax for conformation code (no actual spaces)

<bond symbol> becomes

[< bond symbol> @ <dihedral angle indicator>]

<dihedral angle indicator> becomes one of …

<continuous dihedral angle>, or

<discrete dihedral angle>

<continuous dihedral angle>

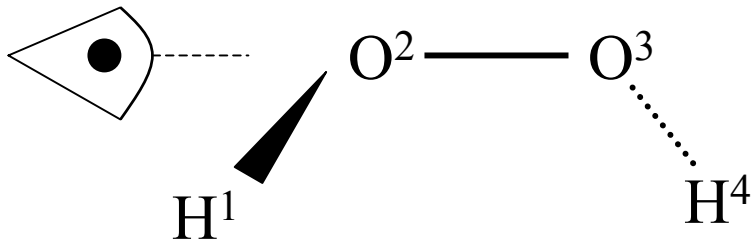
Real valued degrees: $0.0 \leq \mathbf{x} < 360.0$

<discrete angle indicator>

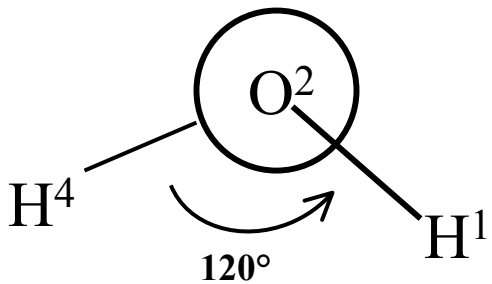
Duodecimal integer valued circle-divisions in [0,1,…,9,A,B] or variation

Conformation code semantics (continuous):

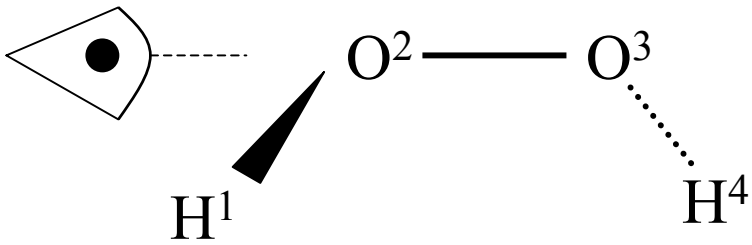
[H]O[-@120.0] O[H]
1 2 3 4



Dihedral angle for O^2- O^3 bond is 120.0 degrees counter-clockwise from far ligand to near ligand

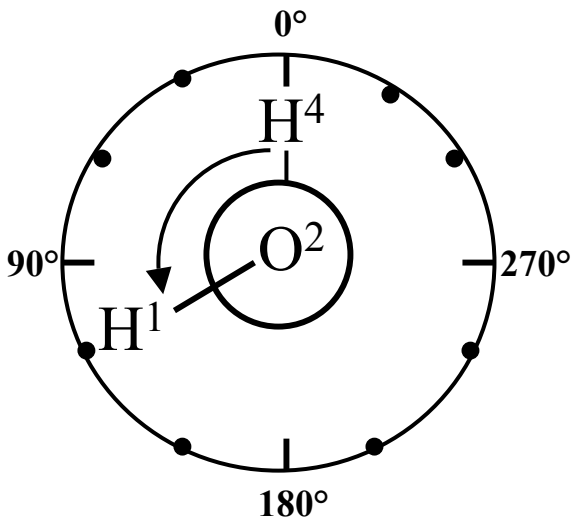


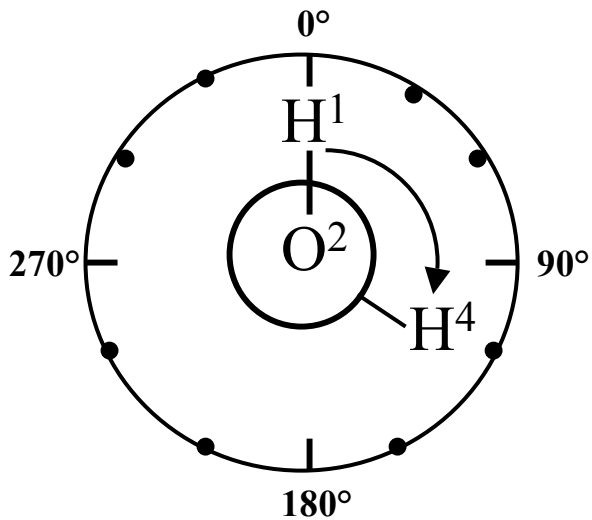
Conformation code semantics (continuous, continued)



[H]O[-@120.0]O[H]
1 2 3 4

Place far ligand “up”, then measure 120.0° counter-clockwise to near ligand. Equivalently, place near ligand “up”, then Measure 120.0° clockwise to far ligand. We call this thing a “bearing” at our day jobs.

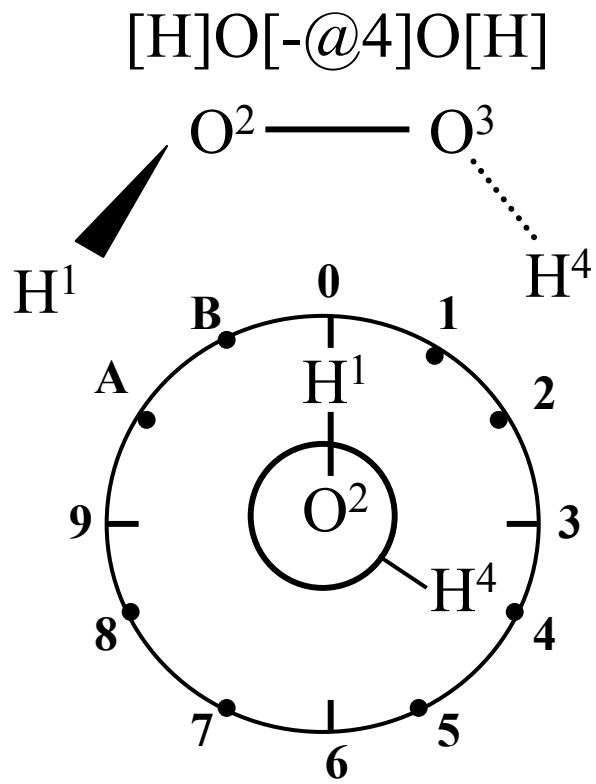




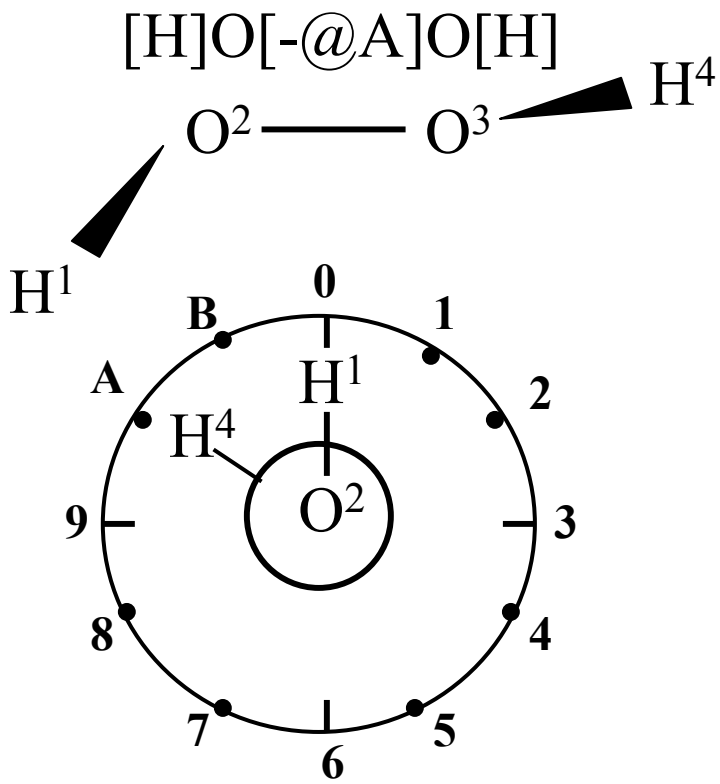
Conformation code semantics (continuous, continued)

- The “far ligand counter-clockwise to near ligand” interpretation respects the counter-clockwise sense of “@”
- The equivalent “near ligand clockwise to far ligand” interpretation corresponds to most standard dihedral angle definitions in chemistry
- We use the first “official” definition to avoid having to use “@@”; work with the second definition to preserve an intuitive clock analogy

Conformation code semantics (discrete)



120.0 degrees or 4 O'clock "bond time"

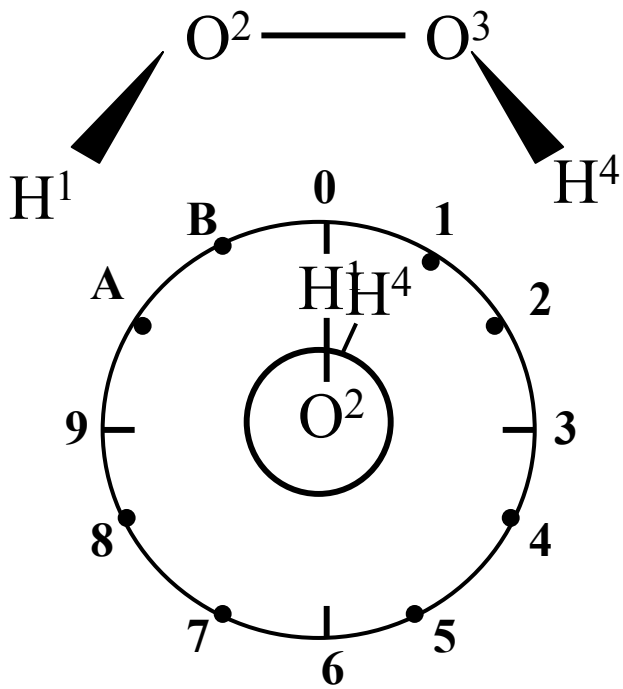


300.0 degrees or 10 O'clock "bond time"

Duodecimal (or "Dozenal") numbers divide the circle
twelve times (every 30.0 degrees)

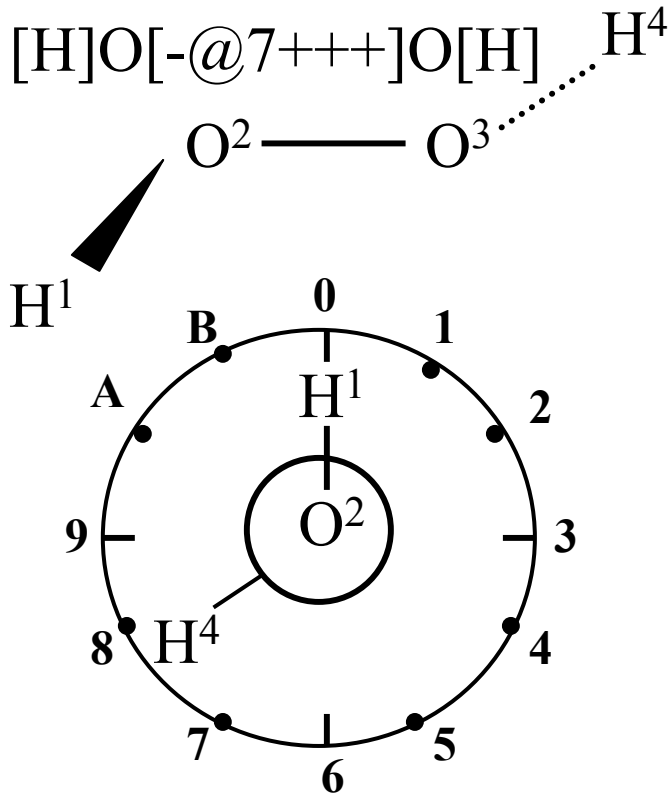
Conformation code semantics (discrete, continued)

[H]O[-@0++]O[H]



15.0 degrees or 12:30 O'clock "bond time"

[H]O[-@7+++]O[H]

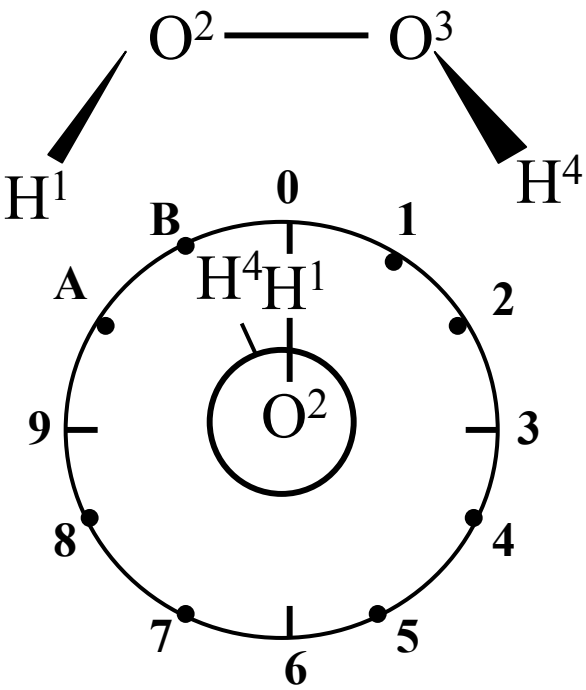


232.5 degrees or 7:45 O'clock "bond time"

The "+" suffix stands for 7.5 degrees or "15 minutes of additional bond time". This system divides the circle forty-eight times (every 7.5 degrees).

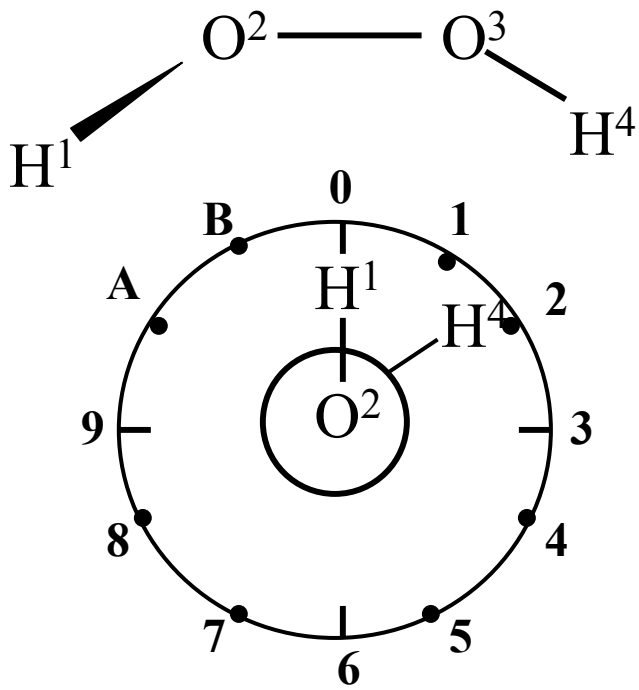
Conformation code semantics (discrete, continued)

[H]O[-@B:1]O[H]



332.5 degrees or 11:05 O'clock "bond time"

[H]O[-@1:A]O[H]

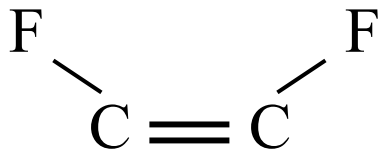


55.0 degrees or 1:50 O'clock "bond time"

Two duodecimal digits separated by a colon could divide the circle 144 times (every 2.5 degrees). Various schemes of discretization might be used.

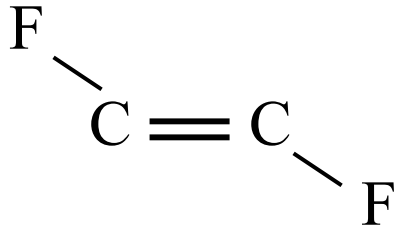
Frozen bond conformation

- Isomerism determined by twists around bonds, even frozen bonds, can and should be coded as conformation



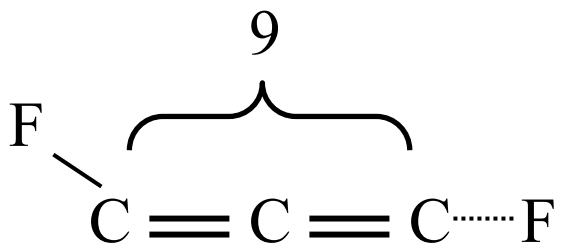
F/C=C\F Equivalent to

FC[=@0]CF



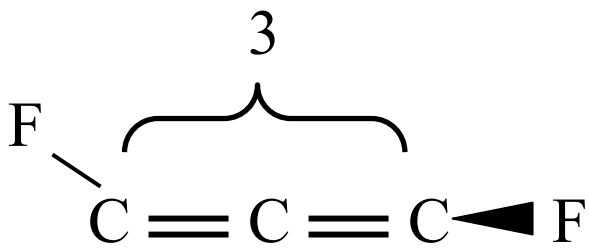
F/C=C/F Equivalent to

FC[=@6]CF



FC=[C@AL1]=CF Equivalent to

FC[=@0]C[=@9]CF



FC=[C@AL2]=CF Equivalent to

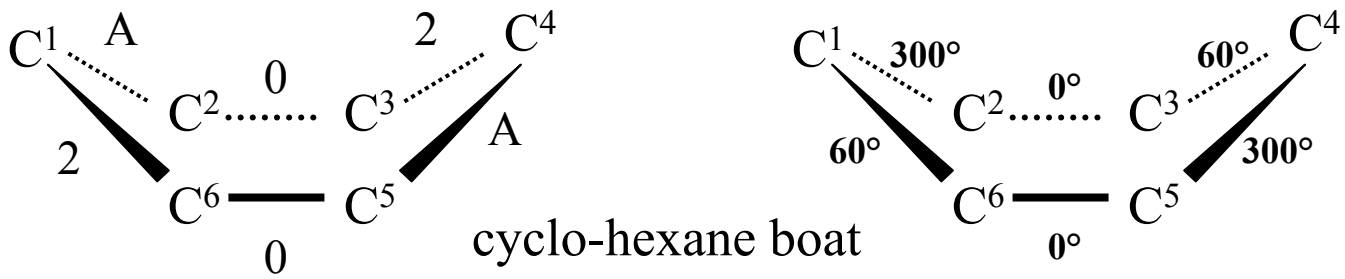
FC[=@3]C[=@0]CF

Note: only the sum of dihedral angles values matters in a straight chain

Unstrained ring conformation

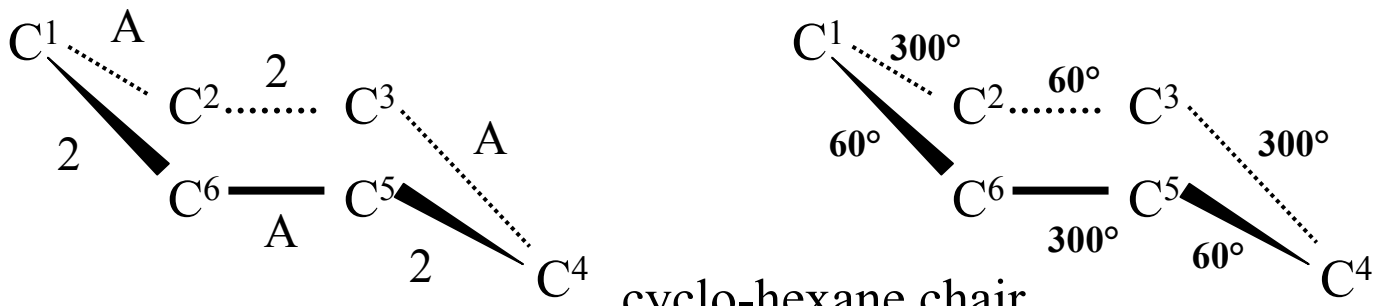
- Unstrained structures, including rings, usually feature dihedral angles of a half, quarter, or third of the circle
- So simple duodecimal codes dividing the circle twelve times are often sufficient

C[-@2]1[-@A]C[-@0]C[-@2]C[-@A]C[-@0]C1



cyclo-hexane boat

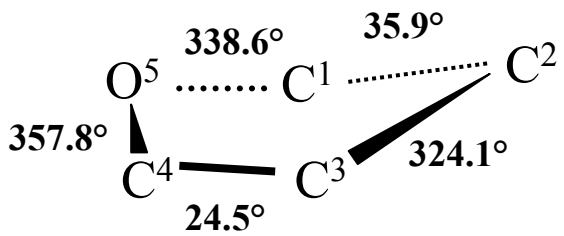
C[-@2]1[-@A]C[-@2]C[-@A]C[-@2]C[-@A]C1



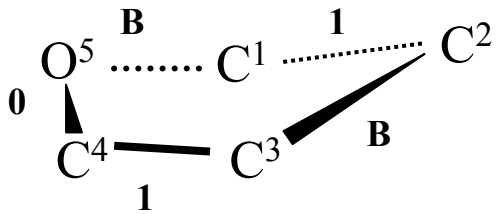
cyclo-hexane chair

Strained ring conformation

- Strained ring structures require real valued angles for complete description
- Discrete codes can divide the ring more finely, and capture more nuance, at the cost of a longer code



C[-@338.6]1[-@35.9]C[-@324.1]C[-@24.5]C[-@357.8]O1

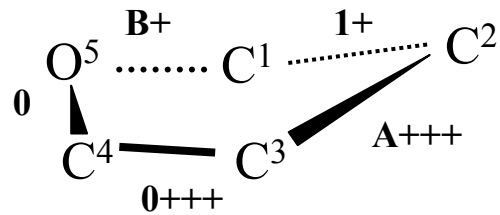


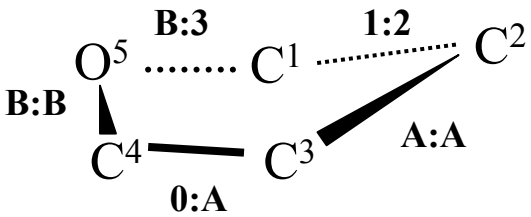
C[-@B]1[-@1]C[-@B]C[-@1]C[-@0]O1

Maximum discretization data loss: 15°

C[-@B+]1[-@1+]C[-@A+++]C[-@0+++]C[-@0]O1

Maximum discretization data loss: 3.75°





C[-@B:3]1[-@1:2]C[-@A:A]C[-@0:A]C[-@B:B]O1

Maximum discretization data loss: 1.25°

Strain, bond length, and molecular geometry

- Twisted SMILES is already very close to giving us a complete geometric description of a molecule – even when we have strain (e.g. five atom ring on the previous slide)
- An obvious omission is bond lengths.
  - They could be inserted into the expanded bond descriptor, e.g. [- length @ dihedral angle]
  - Alternatively, the bond order plus the atom species will often be enough to specify a length (from a look-up table)
- For purposes of determining how well the codes fix geometry, assume from now on that bond lengths are known.

Strain Hints

- Ideally (i.e. in the absence of strain), the bond angles at each atom are specified by the configurational information from standard SMILES (ideal geometry of the @XX?).
- Even in some strained cases (e.g. cyclo-pentane) the explicit dihedral angles of Twisted SMILES are enough to fully specify the geometry.
  - 9 relative degrees of freedom for the 5 C atoms of cyclo-pentane
  - 5 fixed by bond lengths, 5 by dihedral angles – overspecified!
- Sometimes this will not be the case.
- If we want Twisted SMILES to specify the full geometry of any molecule, we can add “Strain Hints”.

Strain Hints (continued)

- Strain Hints are
  - The missing piece of conformational information.
  - “Hints” because much can be deduced from the dihedral angles already presented.
- Two possible (nearly equivalent) schemes are immediately apparent:
  - Scheme 1: Add strain information to atom codes
  - Scheme 2: Add strain information to bond codes

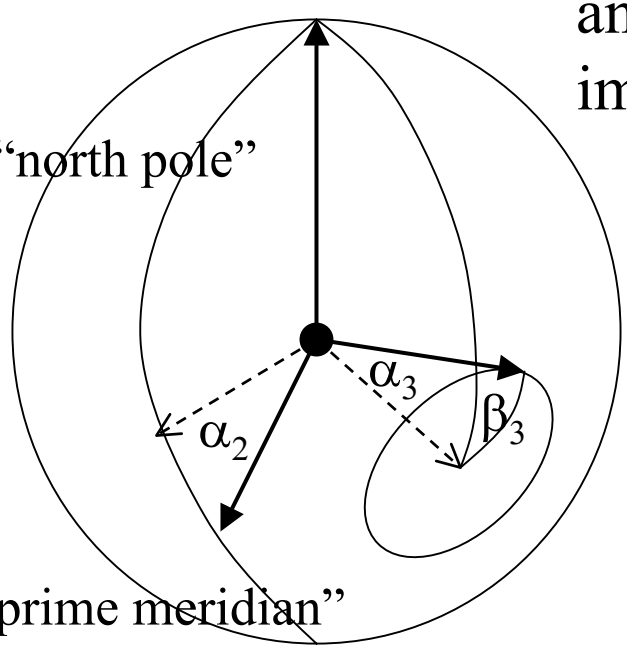
Atom-Based Strain Hints

2n-3 angles are required for n explicit ligands. All the explicit bond angles at the atom are immediately specified.

Actual direction

Unstrained direction

First ligand defines “north pole”



Third ligand needs two angles to specify strain:

α_3 : how far from unstrained position

β_3 : which direction (bearing on surface of sphere)

Second ligand defines “prime meridian”

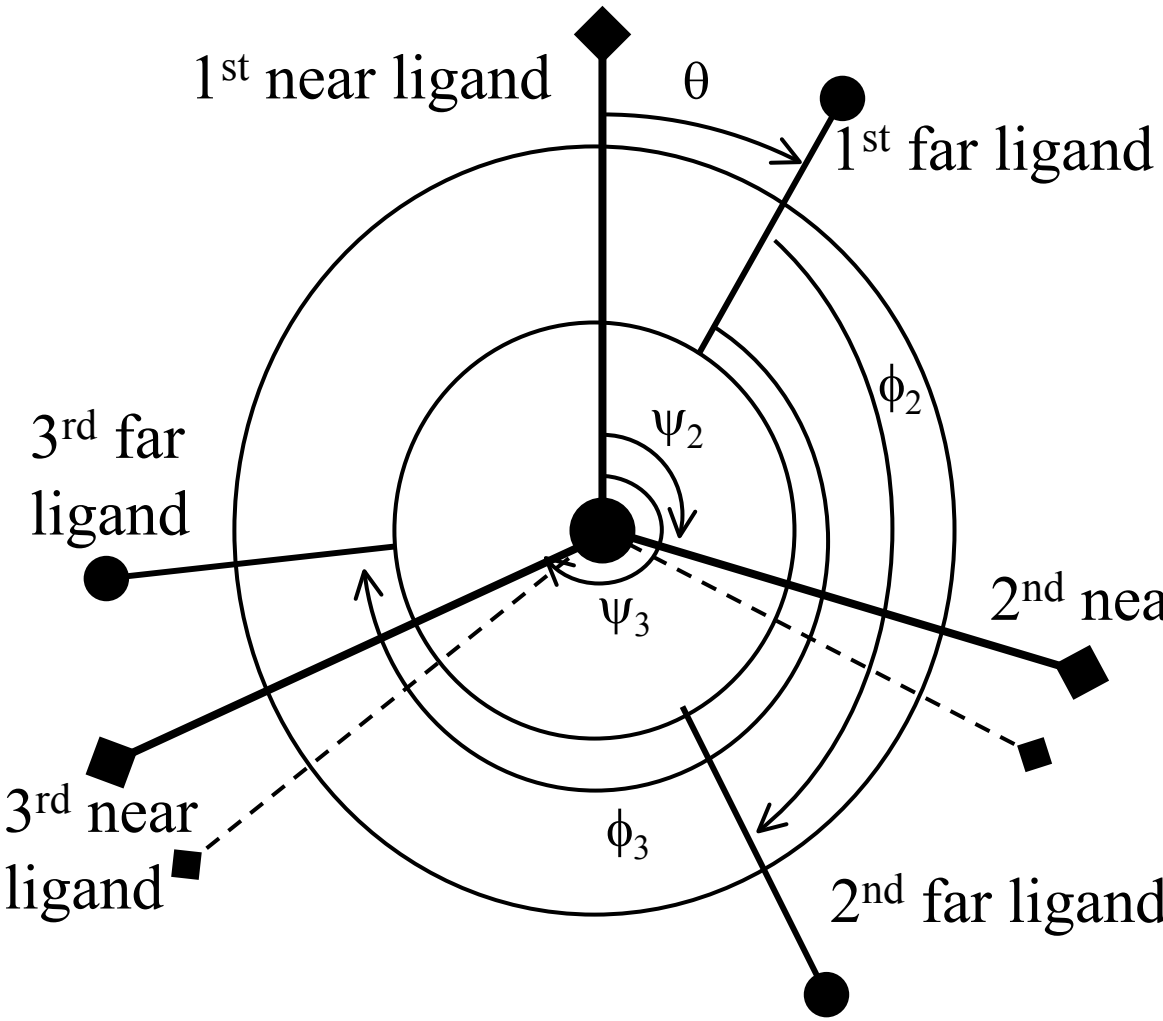
- needs one angle α_2 to specify strain

Atom-Based Strain Hints (continued)

- Atom based strain hints require specifying the deviation from the expected location of a number of points on a sphere

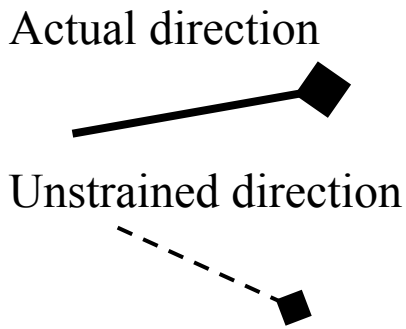
- This task is not trivial, but is a familiar sort of job; cartographers do it all the time

Bond-based Strain Hints



θ: The usual dihedral angle of Twisted SMILES.

ψ_i , ϕ_i: Angles to add to (or subtract from) θ to get all other dihedral angles.



Atom-based vs Bond-based

- Atom-based strain hints
  - are clearly sufficient for a complete geometric specification of the molecule (except for implicit hydrogen)
- Bond-based strain hints
  - are more “natural” in the sense that strain hints are motivated by *conformational* considerations, and thus belong with bonds
  - are equivalent to atom-based strain hints, except where we have atoms with only two explicit ligands – in which case we can add a (one-angle) atom-based strain hint where desired.

Nomenclature: are we there yet

- The problem of nomenclature can be considered "solved" if:
- (1) equivalent beasts have the same name
  - Canonicalization (i.e. resolving reduced chirality) should prove no harder for conformation than it was for configuration
- (2) different beasts have different names
  - Assertion: It is impossible to find two distinct molecules that code the same way in (continuous) Twisted SMILES with Strain Hints
  - Conjecture: It is difficult to find two distinct molecules that code the same way in Twisted SMILES even *without* Strain Hints
    - Challenge: We'd love to see an example!
- (3) names are descriptive
  - On top of the (considerable!) descriptive capabilities of standard SMILES, Twisted SMILES tells you *a lot* about a molecule's geometry

A parting shot

Dare we suggest … IChI?

Acknowledgments

- The first acknowledgement is of a different category
  - To the Google search engine for the mapping, on Friday October 11, 2002
  - "Wisswesser" → "Strings and Things" by John Bradshaw
- Some of the people without whose support and constructive criticism this project would not have happened are
  - John Bradshaw, Colin Cameron, Jim S. Kennedy, Marcel Lefrancois, Bruce McArthur, Garfield Mellema, Gordon Murray, Warren Nethercote, Bill Roger, Michelle Shaw, Dave Weininger
- And most importantly
  - (JM) Peggy, Alice, Eric, and Ian
  - (DP) Kim and Zachary

Contact

john.mcleod@drdc-rddc.gc.ca

dan.peters@drdc-rddc.gc.ca