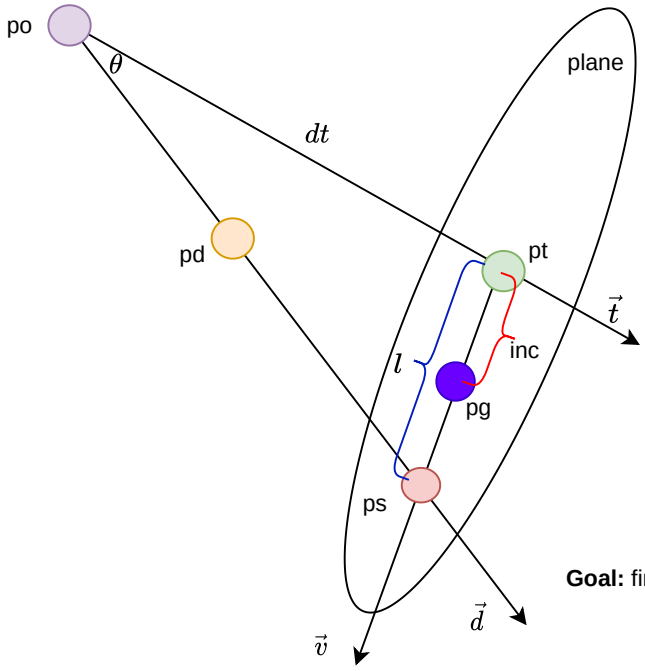
MOSS Bonds Variations With Surrounding Sampling

Mathematics, (can be ignored or look back which might be easy to understand..)

Compute the coordinates of the intersection point inside the plane



3D space, points, po, pd, pt, ps

vector $\vec{t} = po \rightarrow pt$
vector $\vec{d} = po \rightarrow pd$

θ is the angle between $\vec{t}$ &$\vec{d}$
Note: this angle should be the acute angle, otherwise, point ps is on the other side of pt

vector $\vec{t}$ is perpendicular to the plane

ps is the intersection point between $\vec{d}$ and the plane

vector $\vec{v} = pt \rightarrow ps$

dt is the modulus/scalar distance of vector $\vec{t}$
l is the modulus of vector $\vec{v}$

Goal: find the point pg inside the plane along the vector $\vec{v}$ at the increment inc

Then,

A

$$\varepsilon = \frac{\sec(\theta) \cdot dt}{\left|\left|\vec{d}\right|\right|} = \frac{\text{length}(po \rightarrow ps)}{\text{length}(po \rightarrow pd)} = \frac{ps_x - po_x}{d_x} = \frac{ps_y - po_y}{d_y} = \frac{ps_z - po_z}{d_z}$$

where if $any(d_{xyz}) = 0$, its corresponding axis $ps_{\text{axis}} = po_{\text{axis}}$

Thus, $ps = [\varepsilon \cdot d_x + po_x, \varepsilon \cdot d_y + po_y, \varepsilon \cdot d_z + po_z]$

B

$$l = \tan(\theta) \cdot dt$$

define, λ is the ratio between distance inc and l, $\lambda = \frac{\text{inc}}{l} \Rightarrow \lambda = \frac{pg_x - pt_x}{v_x} = \frac{pg_y - pt_y}{v_y} = \frac{pg_z - pt_z}{v_z}$

where $\vec{v} = [v_x, v_y, v_z] = [ps_x - pt_x, ps_y - pt_y, ps_z - pt_z]$

Thus, $ps = \left[\frac{pg_x - pt_x}{\lambda} + pt_x, \frac{pg_y - pt_y}{\lambda} + pt_y, \frac{pg_z - pt_z}{\lambda} + pt_z\right]$

Technically,

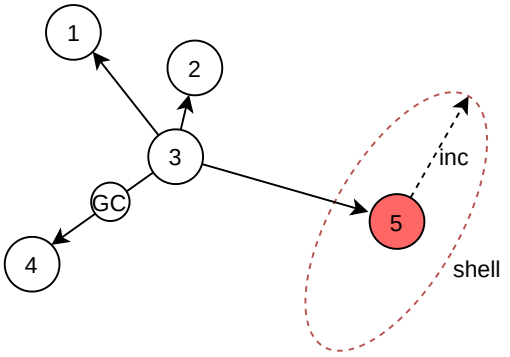
$$\text{acute } \theta = \begin{cases} 89^\circ & \text{if } \theta = 90^\circ \\ 1^\circ & \text{if } \theta = 0^\circ \end{cases}.$$

Simply compare ps in A & B,

$$pg_x = \varepsilon \cdot \lambda \cdot d_x + \lambda \cdot po_x + (1 - \lambda) \cdot pt_x$$
$$pg_y = \varepsilon \cdot \lambda \cdot d_y + \lambda \cdot po_y + (1 - \lambda) \cdot pt_y$$
$$pg_z = \varepsilon \cdot \lambda \cdot d_z + \lambda \cdot po_z + (1 - \lambda) \cdot pt_z$$

Bonds Variations on surroundings

Assume a molecule contains 5 atoms, GC means its geometry center.



Idea

After bonds variations, based on the original reference (pmfzmat), new references can be made, then repeat the bonds variations steps, the molecule can be again sampled.

It is like variation on surroundings.

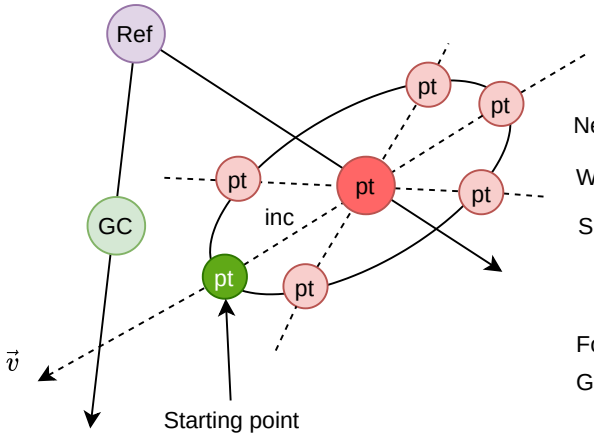
For example, atom 5 is chosen, at the given increment, a shell is created, new atom 5's position is confined on the 3D shell/sphere.

Reduce the 3D shell to the circle:

The circle is defined as the intersection between shell and plane, where the plane is perpendicular to the vector $Ref \rightarrow \mathrm{TargetPoint}$

However, there are infinite points on the circle.

The starting point is chosen as the intersection point with the circle and the vector $\vec{v}$.
Where, $\vec{v}$ can intersect with vector $Ref \rightarrow GC$



New positions of target points are the rotation results along the axis $ref \rightarrow pt$

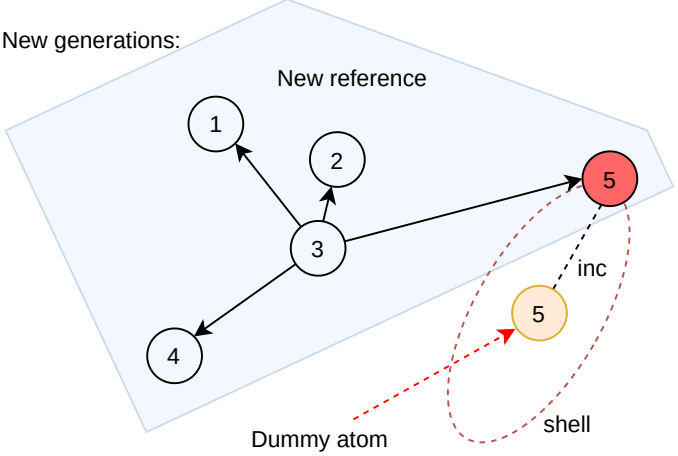
We define this rotation angle as the $vbrotate$ angle

Specially, for this plot, $vbrotate = 60^{\circ}$

For reference atom, any atom except target atom can be chosen

Geometry center is fixed

New generations:



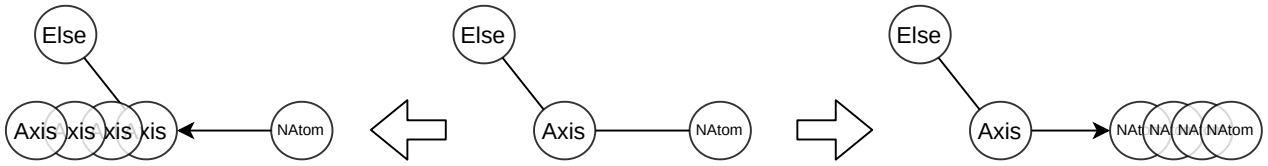
Assume new reference molecules are created.

By choosing one of those new references, do bonds variations.

Now, discuss conditions: $\text{new atom} = \begin{cases} \text{as the direction} \\ \text{as the origin} \end{cases}$

Shown in here, just let you know, which atom is "surroundingly" sampled.

Simplify the plot like,



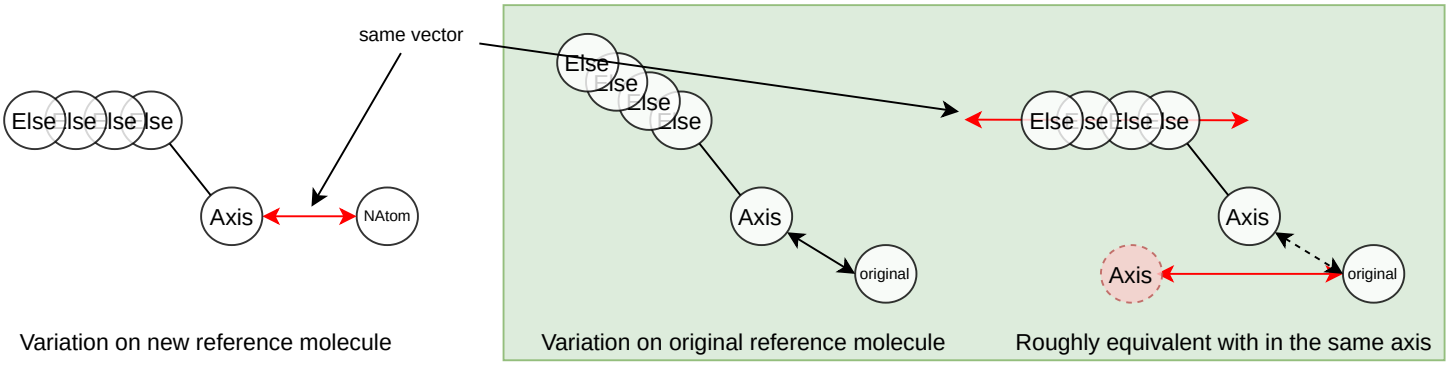
New atom as the origin point
varying **axis** at every increment

New reference molecule

New atom as the direction point
varying **it** at every increment

Care: Variation may or may not include Else Atom(s).

Situations like varying only on Else atom(s) should **not** be considered.



Variation on new reference molecule

Variation on original reference molecule

Roughly equivalent with in the same axis

As you can see, varying only on Else Atoms is roughly equivalent to varying of the same directional vector on original reference molecule.
Sampling on them do not make big changes, although one of their atoms is different: NewAtom v.s. original.

Equation of Bonds Variations on rotation:

dual operation

$$VBr(n) = \frac{360}{\text{vbrotate}} \cdot A_n^1 \cdot O_2 \cdot \left[\sum_{\Omega_{\text{axis}}=1}^{n-2} C_{n-2}^{\Omega_{\text{axis}}} + \sum_{\Omega_{\text{na}}=1}^{n-2} C_{n-2}^{\Omega_{\text{na}}} \right]$$

rotation angle to generate new references

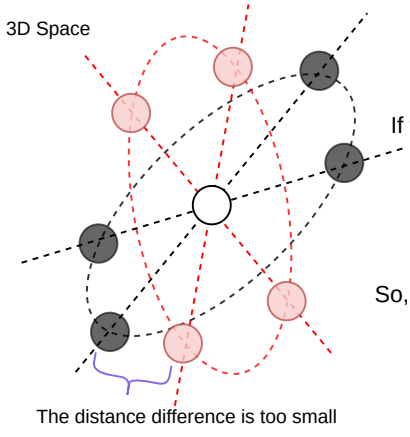
select one of atoms

origin on new-atom, varying includes axis atom

origin on axis atom, varying includes new-atom

Term "n-2" is different with "n-3" in function $VB(n)$ for dual operation.

Some discussions on the bonds variations rotation angle
For the selected atom, sampling on its surroundings. Apparently, different axis atoms will generate different shells.



If the rotation angle is too small, then many new reference molecules will be generated on the shell-circle.

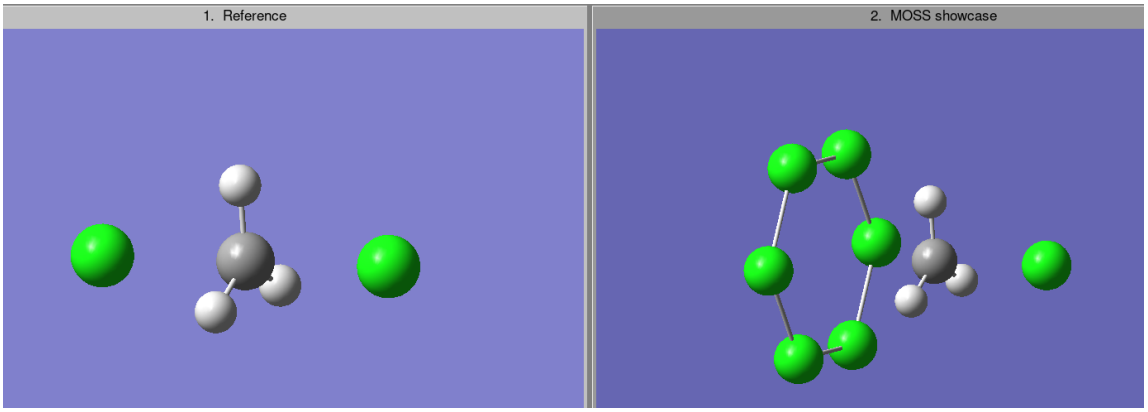
As a result, similarity may happen among those new reference molecules.

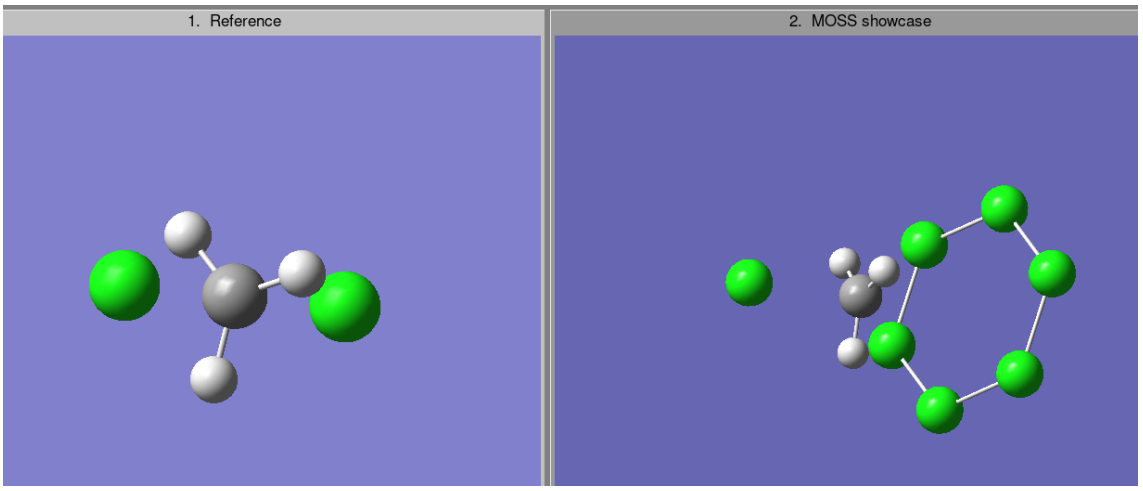
So, the rotation angle should become larger as the number of atoms in the reference molecule increase.

From the function, for n atoms molecule, if its $vbrotate$ angle is set to 60°, then

$$VBr(2) = 24$$
$$VBr(3) = 72$$
$$VBr(4) = 432$$
$$VBr(5) = 1680$$
$$VBr(6) = 5400$$

Showcase on Bonds Variations Rotation Sampling





Folder showcase-vbr is attached, which includes all surrounding sampling examples
For the meaning of file names, please take a look on this readme file.

file name format

showcase-vbr-2A-60d-2-sample3.txt

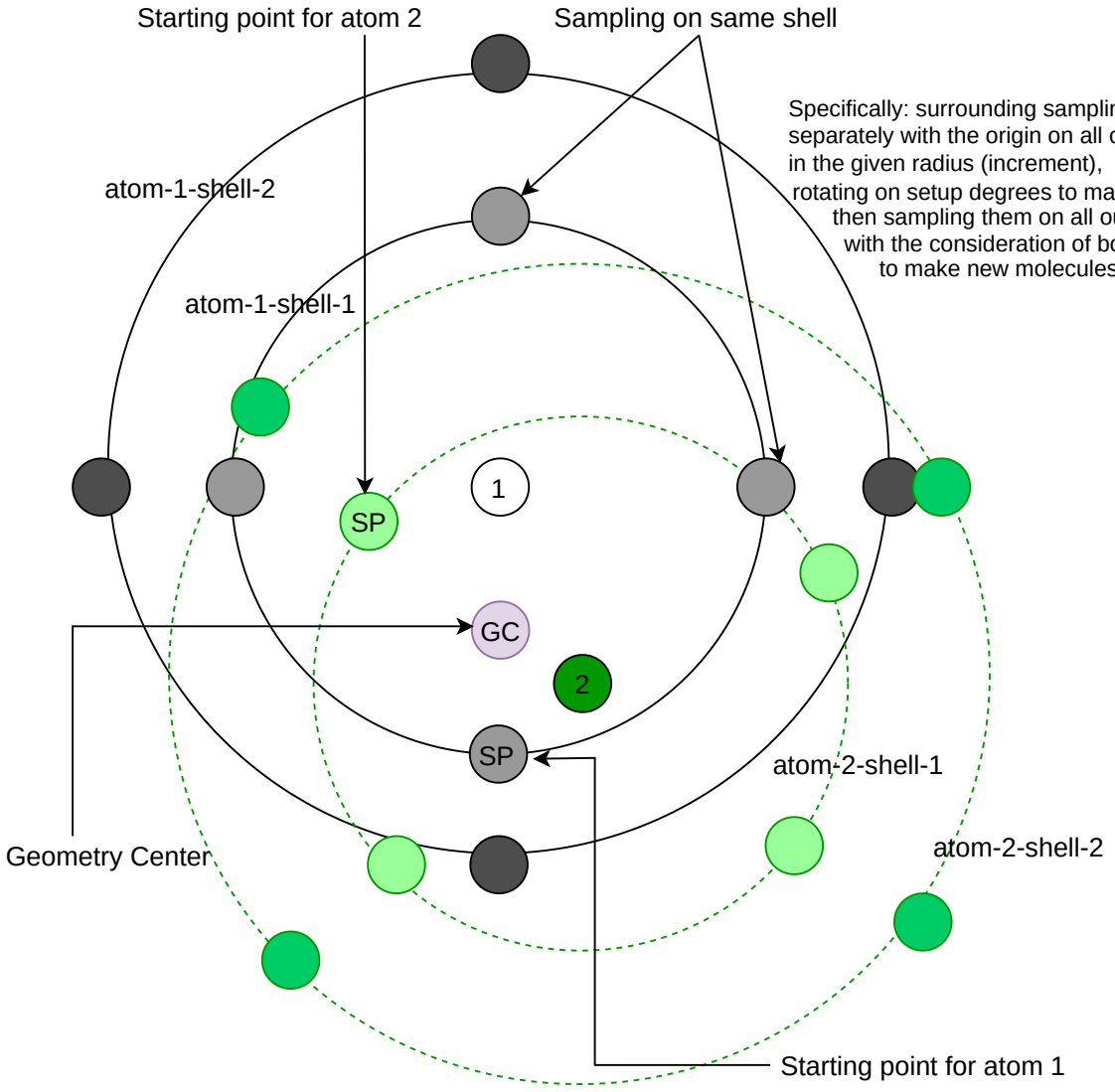
2A : surrounding sampling distance
60d : vbrotate angle in degree
2 : index of reference atom, second atom
sample3 : index of atom to be sampled, third atom

a hidden dummy atom, geometry center, is not shown.

parsed as:
for reference molecule, take second atom as the origin,
with the geometry center to calculate direction for the starting point,
sequentially rotating 60 degree to sample the third atom.

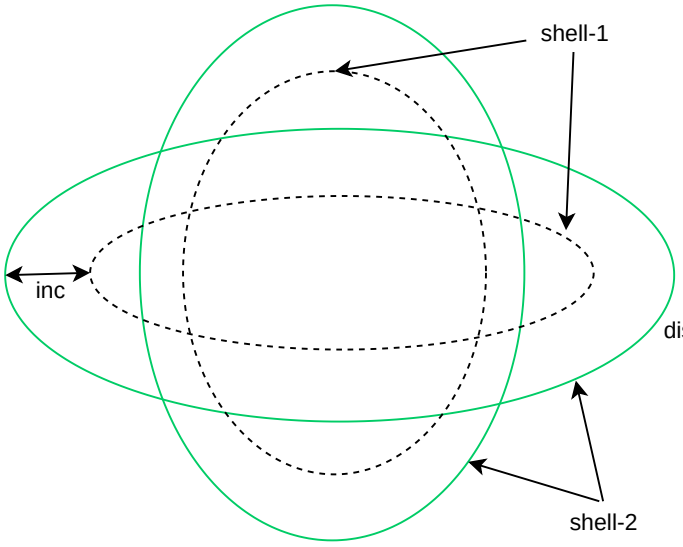
More examples for understanding..

If the molecule is in 2D planar, ignoring the other molecules, the plot would be like



Specifically: surrounding sampling on atom 1,
separately with the origin on all other atoms,
in the given radius (increment),
rotating on setup degrees to make new references,
then sampling them on all outer shells,
with the consideration of bonded information,
to make new molecules.

If it is 3D plot,

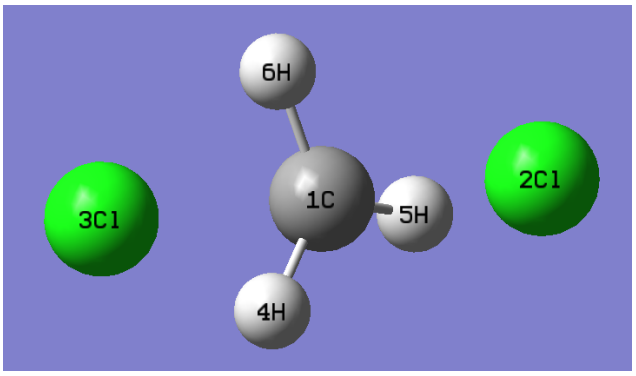


Samplings are performed on shells

distances between consecutive shells are determined by increment.

Selection of molecules for ML training

Reference molecule:



Two folders are included, by knowing the meaning of their file names,
using the attached script, "bash_process.sh", then,

(careful with the folder name, moss-vb-full is the sampling on original reference, moss-vbr-full: again samplings on surrounding samplings)

Examples

Training on VB sampling on atoms 1 or 3 in dual mode (sampling containing them).

check them:

```
xiang@:sn2$ for i in moss-vb-full/vb-dual-*1-3*; do echo $i; done
moss-vb-full/vb-dual-less-1-3+3+2+4+5-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2+4+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2+4-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2+5+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2+5-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+2-m52.txt
moss-vb-full/vb-dual-less-1-3+3+4+5+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+4+5-m24.txt
moss-vb-full/vb-dual-less-1-3+3+4+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+4-m24.txt
moss-vb-full/vb-dual-less-1-3+3+5+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3+5-m24.txt
moss-vb-full/vb-dual-less-1-3+3+6-m24.txt
moss-vb-full/vb-dual-less-1-3+3-m52.txt
moss-vb-full/vb-dual-more-1-3+3+2+4+5-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2+4+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2+4-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2+5+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2+5-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+2-m105.txt
moss-vb-full/vb-dual-more-1-3+3+4+5+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+4+5-m24.txt
moss-vb-full/vb-dual-more-1-3+3+4+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+4-m24.txt
moss-vb-full/vb-dual-more-1-3+3+5+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3+5-m24.txt
moss-vb-full/vb-dual-more-1-3+3+6-m24.txt
moss-vb-full/vb-dual-more-1-3+3-m154.txt
```

As you can see, the command (wildcards) is correct, sampling is performed on bonds 1-3, either shrinking or expanding samples are containing them with all others

Use script combine them to make a new total full file:

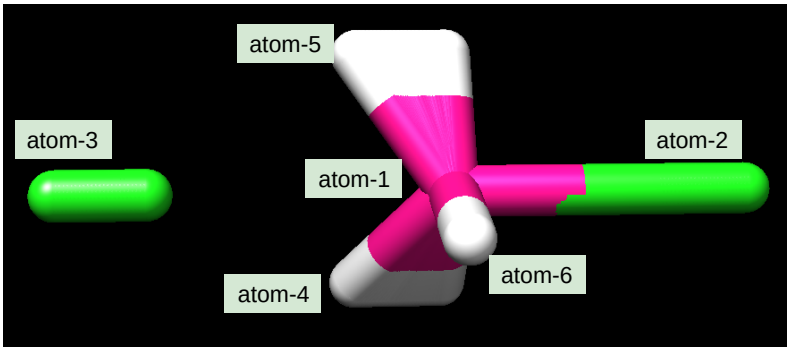
```
xiang@:sn2$ ./bash_process.sh combine moss-vb-full/vb-dual-*1-3*
Note: DONE: new file: combine-vb.txt
```

Then this file can be split by our early script, "backupFileProcess.py", which is attached.

If you want to have a look at the sampling result on a certain file,

```
xiang@:sn2$ ./bash_process.sh showcase moss-vb-full/vb-dual-more-1-3+3+2+4+5-m24.txt
Note: DONE: new file: showcase-vb.xyz
```

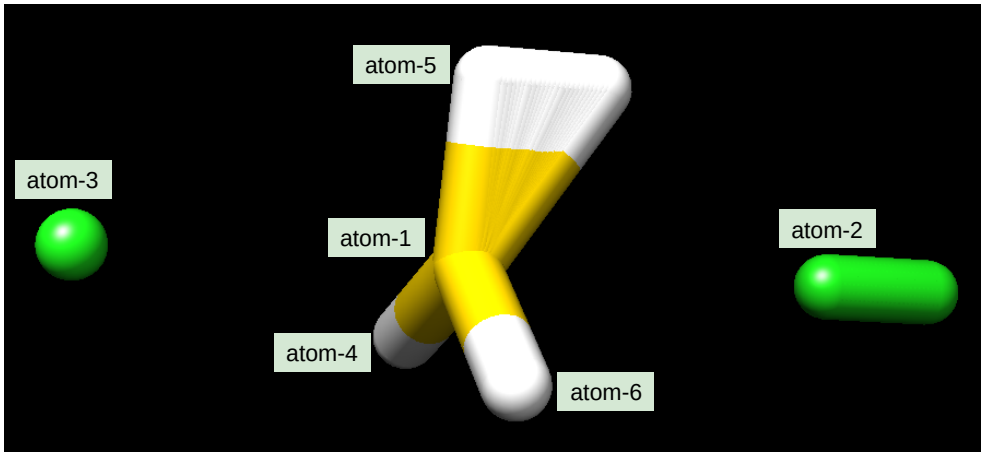
can you see the file type is in xyz? Which can be opened by VMD or Chimera program.



As you can see, the sampling is performed on atoms 3, 2, 4 and 5, in axis 1->3 expanding, 24 molecules are generated, however, atoms 1 and 6 are fixed at their position...

Another plot, this script will never overwrite any files, so it is safe to continuous combine files

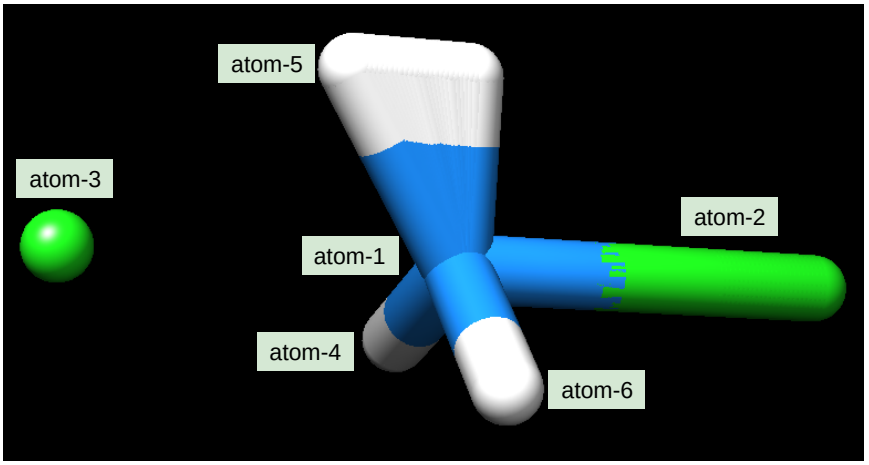
```
xiang@:sn2$ ./bash_process.sh showcase moss-vb-full/vb-dual-more-1-2+2+5-m24.txt
Note: DONE: new file: showcase-vb-1.xyz
```



As you can see, the sampling is performed on atoms 2 and 5, in axis 1->2 expanding, 24 molecules are generated,

Continue..

```
xiang@:sn2$ ./bash_process.sh showcase moss-vb-full/vb-dual-less-1-2+2+5-m24.txt
Note: DONE: new file: showcase-vb-2.xyz
```

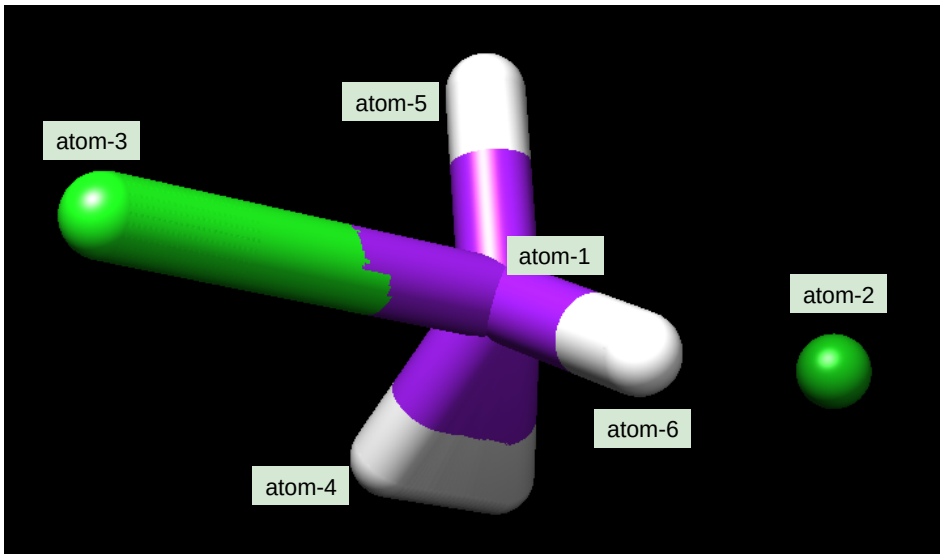


Can you see that we are combining "less" samplings..
its axis is in 1->2 shrinking
but in axis 2->1 expanding..

Sampling on dual mode always contains axes atoms.

continue of single mode...

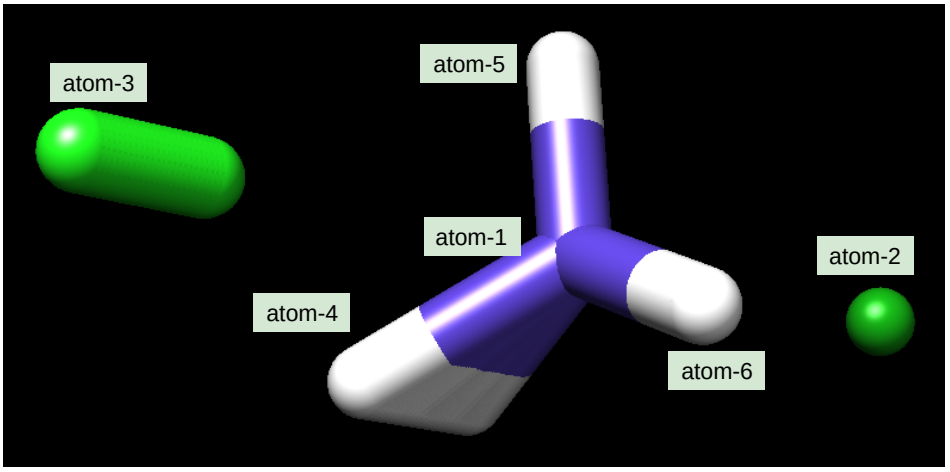
```
xiang@:sn2$ ./bash_process.sh showcase moss-vb-full/vb-single-more-1-2+3+4-m24.txt
Note: DONE: new file: showcase-vb-3.xyz
```



In axis 1->2 expanding, sampling atom 3 & 4

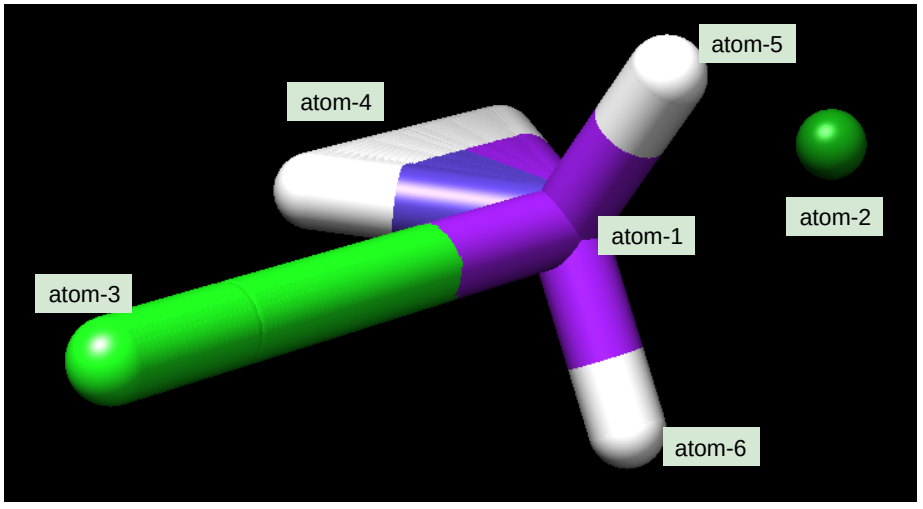
As the comparison

```
xiang@:sn2$ ./bash_process.sh showcase moss-vb-full/vb-single-more-2-1+3+4-m24.txt
Note: DONE: new file: showcase-vb-4.xyz
```



In axis 2->1 expanding, sampling atom 3 & 4

In full



Sampling on single mode will never contain axes atoms

I guess upto now, you should understand the meaning of their file names..

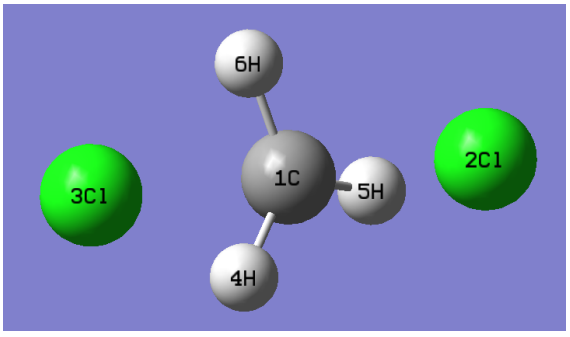
To get script usage:

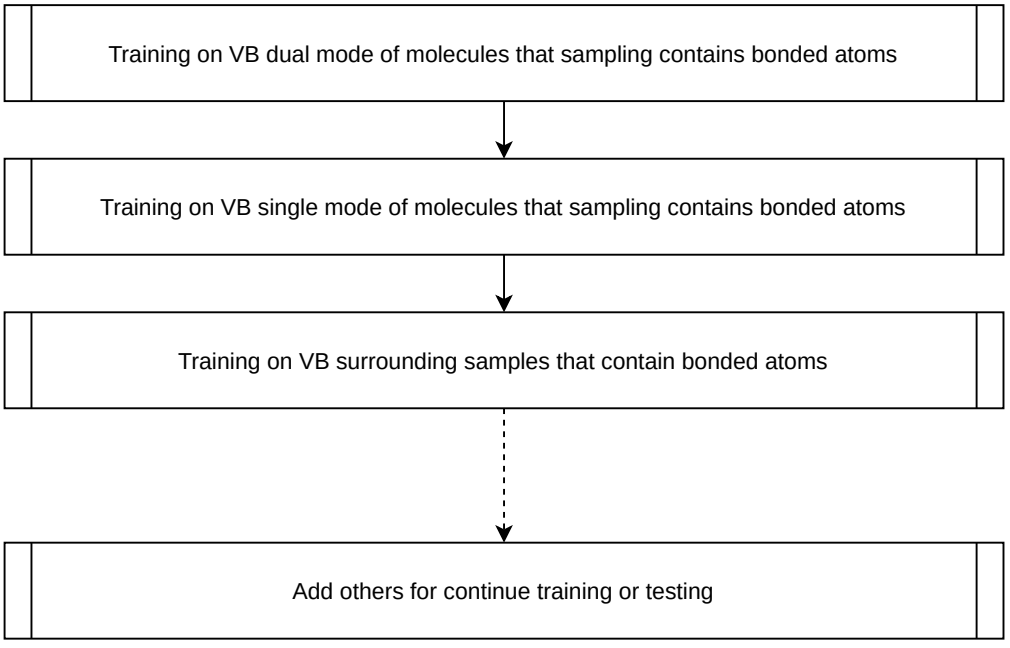
```
xiang@:sn2$ ./bash_process.sh
Process MOSS Generations

Usage:

    1) ./bash_process.sh  [showcase]  [vb-text-files]
    => convert vb text files to a single xyz file for program visualiztion

    2) ./bash_process.sh  combine  [vb-text-files]
    => combine vb text files to a single vb text file for split-training
```

Some protocol for molecule selections

For our reference molecule, bonded atoms are 1,4,5 and 6,
free atoms are 2 and 3

With the knowledge that MOSS cannot well handle non-bonds info.





Summary:

number of atoms : 6
increment : 0.03 Angstrom
vbrotate angle : 60 degree

VB(6) = (dual=900, single=450, total=1350)
number of new molecules with fully bonds variations: 33363

VBr(6) = 5400
after bonds filtration: valid: VBr(6) = 3600
number of new molecules in new references: 229891

The End