Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information: Fast exploration of Potential Energy Surfaces with a joint venture of Quantum Chemistry, Evolutionary Algorithms and Unsupervised Learning. Supplementary Information

Giordano Mancini,*,† Marco Fusè,‡ Federico Lazzari,† and Vincenzo Barone†

†*Scuola Normale Superiore, Piazza dei Cavalieri 7, 56125, Pisa, Italy*

‡*DMMT-sede Europa, Università di Brescia, viale Europa 11, 25121 Brescia, Italy*

E-mail: giordano.mancini@sns.it

Aspartic acid in the gas-phase

Effect of genetic operators

It can be easily observed that enforcement of diversity in the initial population *does* improve the speed of the exploration both for the best case scenario and on average, even in the presence of a large dispersion between single runs. Note that even for "missing" structures the nearest neighbour was not far from the arbitrary cut off of 0.2 Å and even in the worst case scenario these are high lying conformers. Moreover the missing structures do not change among different runs. This could have been expected taking into account the workflow of a general EA and, in particular, the way in which the Island Model tries to preserve diversity.

The results show that the mutation does speed up the exploration (as evidenced by the slightly higher number of missing structures and RMSD) and, at variance with crossover, plays a *critical* role. As a matter of fact, the global energy minimum is not properly located in the absence of mutation. This confirms that for PES explorations EAs behave more like "hill climbers" than "hyperspace samplers" (as is the case, for instance, in combinatorial optimization). This finding is reasonable if one thinks that the algorithm just works with a very limited sub set of degrees of freedom, so that very large populations would be needed to obtain an unbiased sampling of different regions of the search space. Note however that increasing the mutation rate above the default value (0.3 for parents and 0.5 for children) did not result in any significant benefit (data not shown).

Table S1: Results of the IM-EA runs on Asp without either crossover or mutation.

run	# calc.	# miss	RMSD_{max} (Å)	ΔE(kJ/mol)
no CO	3700	1	0.2068	21.956
no CO	2000	0	NA	NA
no CO	3300	0	NA	NA
no CO	3500	1	0.2483	54.229
no Mut	700	9	0.2279	0.
no Mut	700	11	0.2112	0.
no Mut	500	9	0.2482	0.
no Mut	1400	11	0.2141	0.

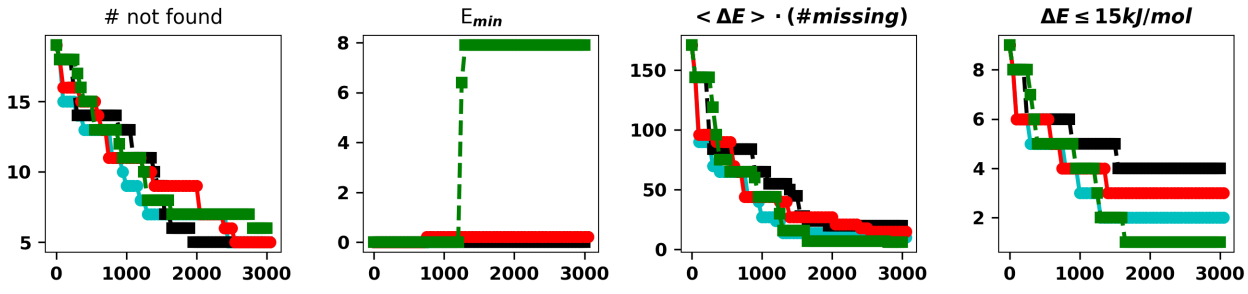

Figure S1: Summary of results for the IM-EA searches with DFTBA. The panels are analougous to those of Figure 2 in the main text

Silver ion in aqueous solution

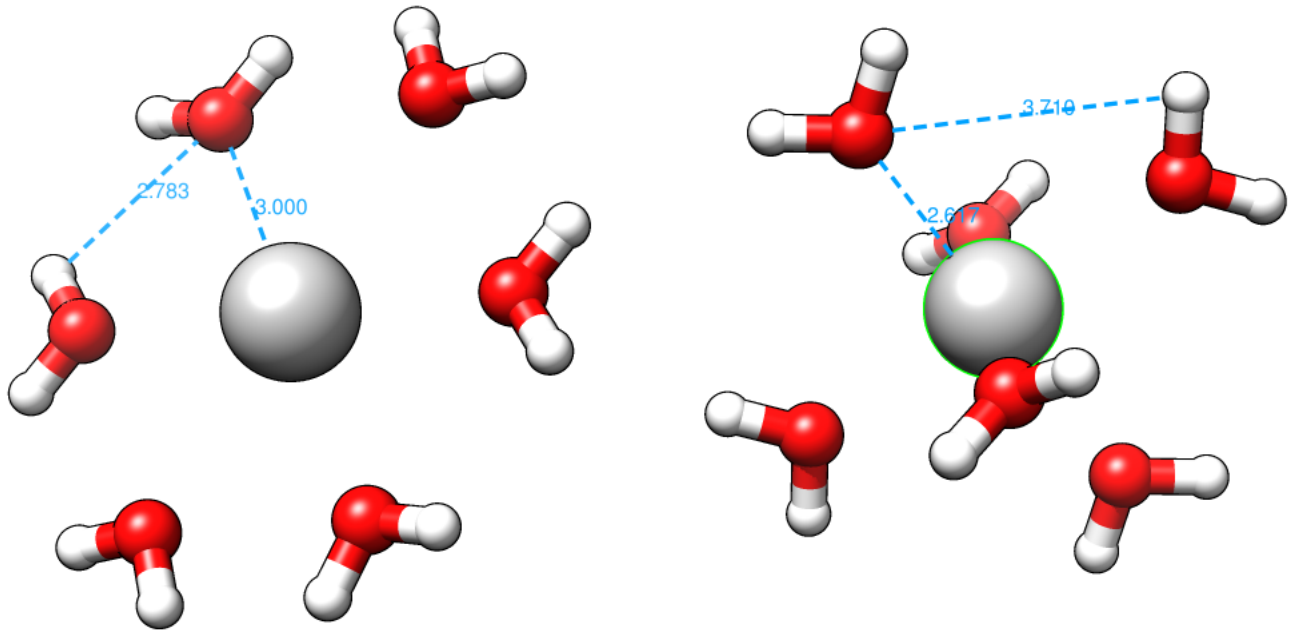

Figure S2: Starting templates for the $Ag^{+}(aq)$ searches. Left: planar geometry. Right: octahedral geometry. Ag−O and nearest neighbour O−H distances are shown in Å.

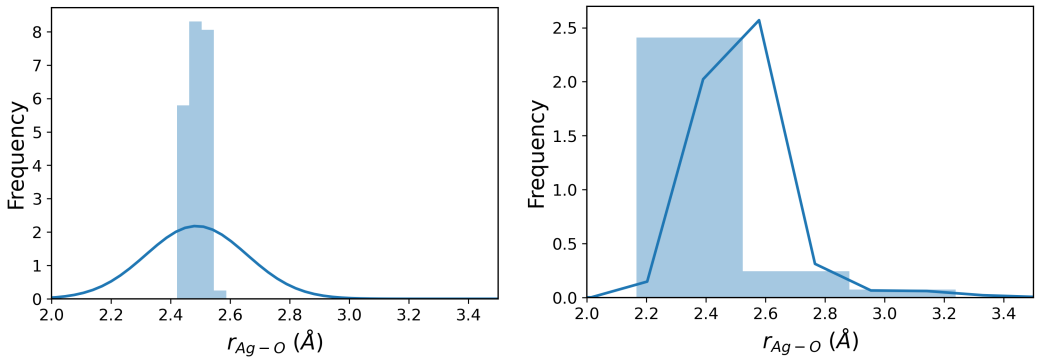

Figure S3: Distribution of ion-oxygen distances for the first CREST (left) and IM-EA/XTB search for the $Ag^{+}(aq)$ ion PES.

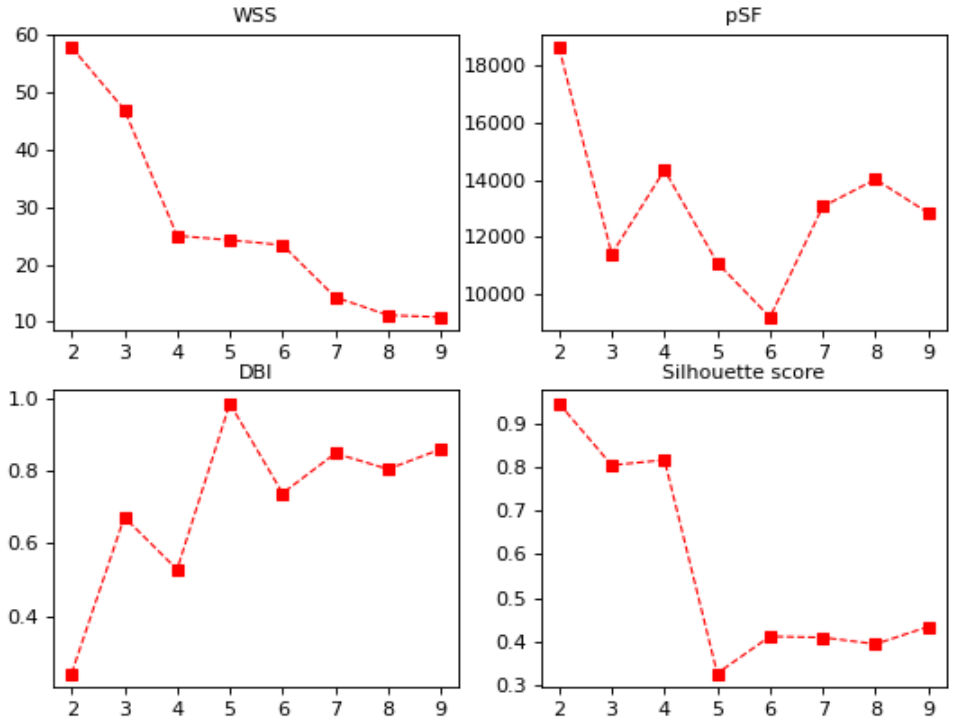

Figure S4: Clustering validation scores for the first IM-EA/XTB run as a function of the number of clusters k.

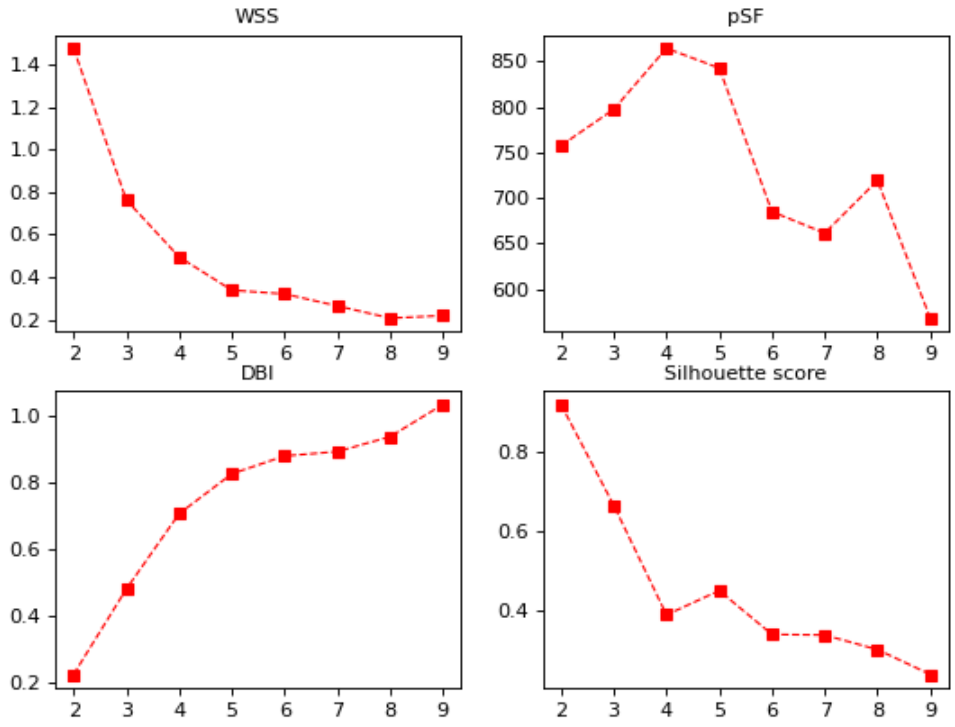

Figure S5: Clustering validation scores for the first CREST run as a function of the number of clusters k.

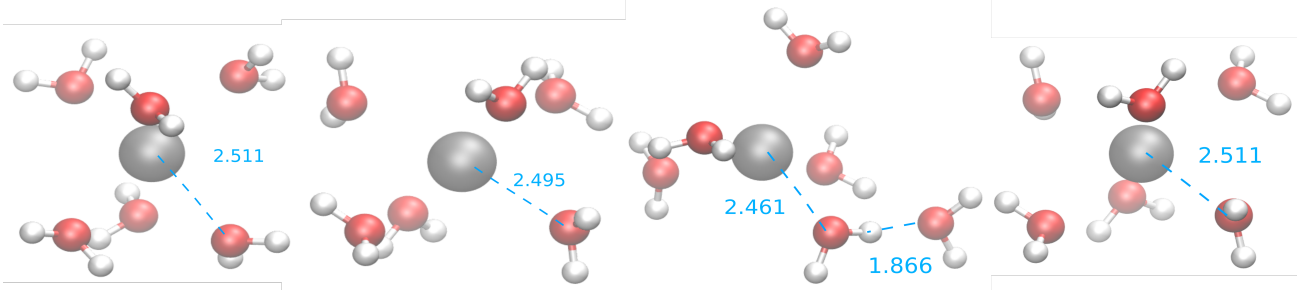

Figure S6: Medoids obtained from the first CREST run. Selected Ag−O and O−H distances are shown in Å.

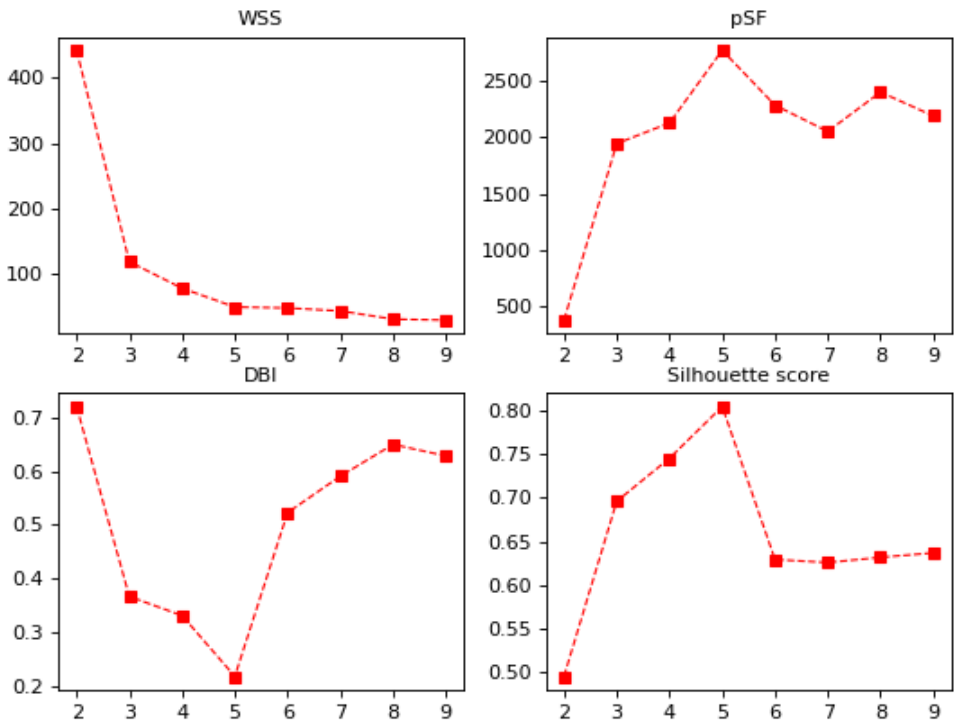

Figure S7: Clustering validation scores for the IM-EA/B3SC run as a function of the number of clusters k.

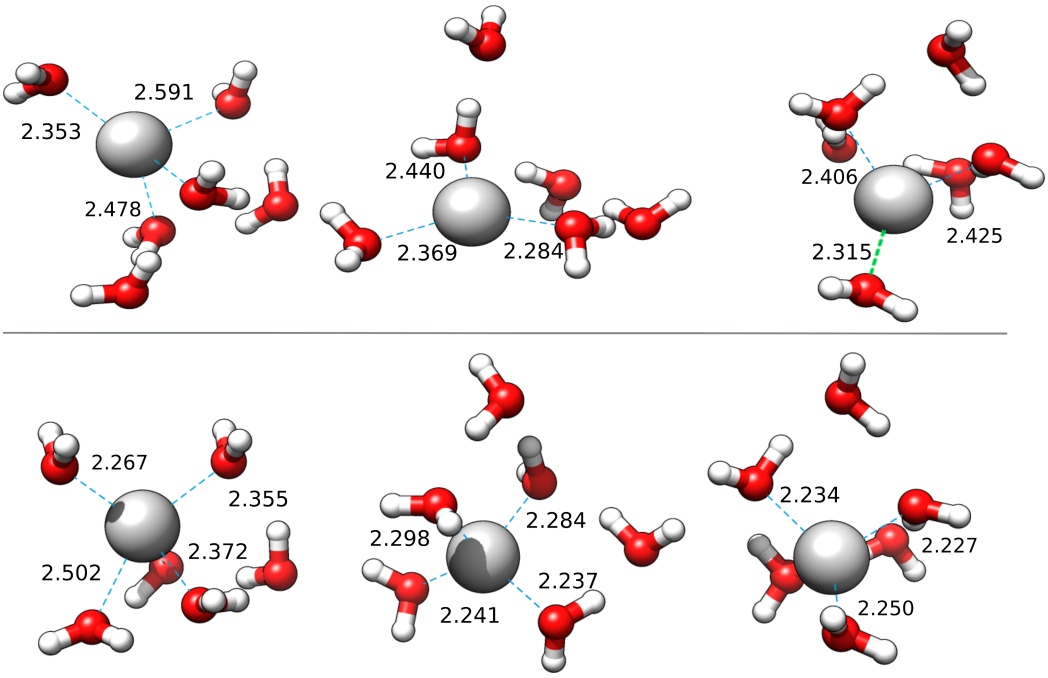

Figure S8: Geometries of medoids 1,2 and 4 from the fist IM-EA/XTB search re-optimized at the B3SC (top row) and B3LC (bottom row) level. Ag−O distances are shown in Å.

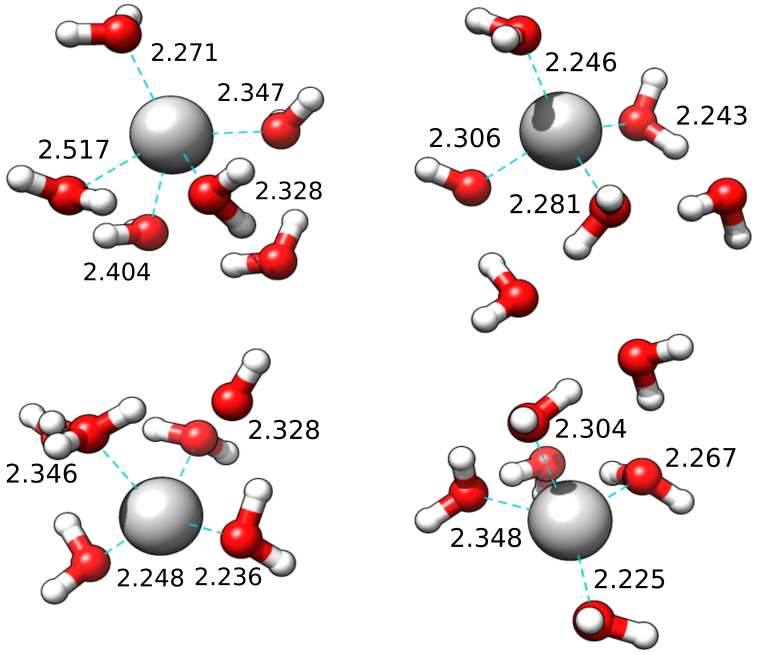

Figure S9: Geometries of medoids 1-4 from the fist IM-EA/XTB search reoptimized at the BLC level.Ag−O distances are shown in Å.

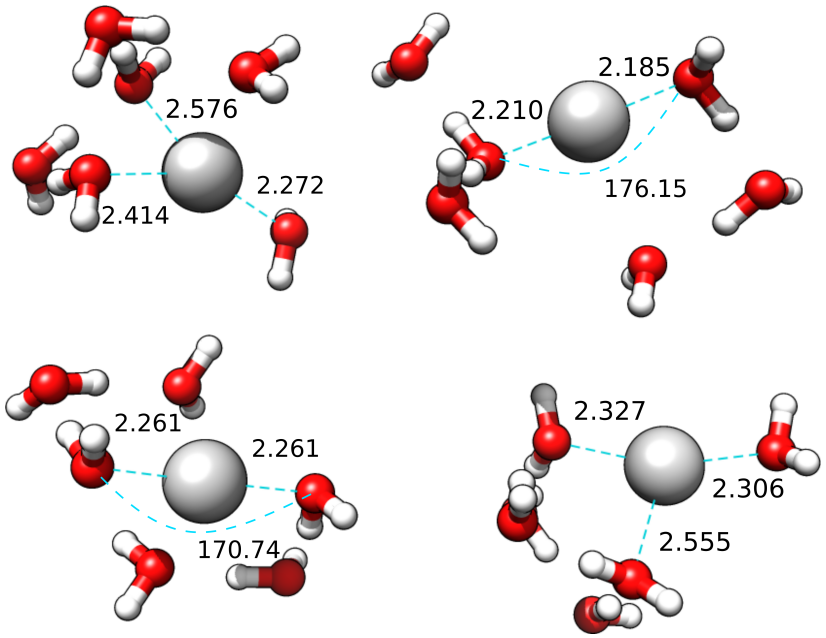

Figure S10: Geometries of medoids 1-4 from the fist IM-EA/XTB search re-optimized at the BSC level with the full basis set.Ag−O distances are shown in Å; for linear structures the O−Ag−O angle is shown as well

.

Input file for Gaussian 16 geometry optimization at the B3LC level starting from the geometry of medoid 4 of the B3SC search.

```
#P B3LYP/gen pseudo=read test
 empiricaldispersion=gd3
 opt=tight scrf=(cpcm,read)

silver cluster DFT opt

  1 1
Ag     0.000   0.000   0.000
O     -0.101  -1.527   1.918
H      0.736  -1.321   2.366
H     -0.165  -2.485   1.887
O     -0.188  -1.352  -2.045
H      0.642  -1.668  -2.410
H     -0.869  -1.969  -2.328
O      0.204   1.544   2.028
H      0.136   2.491   1.870
H     -0.535   1.306   2.597
O      2.249  -0.263   2.811
H      2.699  -0.140   3.651
H      1.691   0.519   2.666
O     -1.423   1.791  -0.870
H     -2.385   1.772  -0.858
H     -1.168   2.199  -1.702
O      2.453  -0.162  -0.004
H      3.096   0.477  -0.323
H      2.677  -0.339   0.922

H 0
 cc-pvTZ
 ****
O 0
 cc-pvTZ
 ****
Ag 0
S    1   1.00
      1.4801580              1.0000000
S    1   1.00
      0.6538510              1.0000000
S    1   1.00
      0.1244880              1.0000000
S    1   1.00
      0.0492640              1.0000000
S    1   1.00
      0.0160000              1.0000000
P    1   1.00
```

```
      0.0756940              1.0000000
P     1   1.00
      0.0237230              1.0000000
D     1   1.00
      0.0660000              1.0000000
 ****

Ag 0
ECP46SDF 3 46
F-component
1
2  1.000000  0.000000
S-F
2.
2  0.800000  7.901670
2  0.400000 -2.271510
P-F
2
2  0.800000  8.033490
2  0.400000 -1.857460
D-F
2
2  0.400000  3.404570
2  0.200000 -0.179570

Nsph=1

0. 0. 0. 6.0
```

In the case of small-core computations, both the pseudopotential and the basis are available in the internal library of G16, so that the basis set section of the above input is replaced by the following one.

```
H 0
 jun-cc-pvTZ
 ****
O 0
 jun-cc-pvTZ
 ****
Ag 0
 sdd
 ****

Ag 0
 sdd
```

Bash script to submit a IM-EA run for Aspartic Acid

```
#!/bin/bash
#PBS -l select=1:ncpus=08 -q q02curie -N asp_mut_R1

module load python/3
module load xtb/6.4.1

export ROOT=/home/${USER}/bigd/ga_test/
export WORK=/home/${USER}/bigd/ga_test/aminoacids/AsparticAcid/new_mut
export BIN=/home/${USER}/from_bigdata/ga_test/src/
export PYTHONPATH="$PYTHONPATH:$BIN"
export SCRATCH="/local/scratch/g.mancini"
export LOCALDIR="$SCRATCH/$PBS_JOBID"
export TPL="$WORK/asp_ref_order.tpl"
export xtbTPL="$WORK/asp_xtb.tpl"

echo "ENVIRONMENT SET"

cd $WORK
rm -rf $WORK/Asp_xtb_d6_mut_R1 asp_mut_R1.e* asp_mut_R1.o*
echo $PBS_JOBID
rm    -rf $LOCALDIR
mkdir -p $LOCALDIR
cp xtb_shell.sh $xtbTPL $TPL $LOCALDIR
cd $LOCALDIR

echo "working area set"

python3 $BIN/conf_GA_parser.py -I $TPL -g 1 4 4 13 4 6 6 9 9 11 13 15 -N Asp_xtb_d6_mut_
 -n 50 -C 100 -s 0.5 -c 0.6 -m 0.5 -M 0.3 -V 4 --cutoff 0.75 \
 -F -l "opt=(verytight,MaxCycles=100,ModRedundant)" -X "rotation3"\
 -i 4 -p 4 "maxdist" "cosine" \
 --xTB $xtbTPL --hof 0.1 --var 10.\
 -X "rotation2"
 1> >(tee $WORK/Asp_xtb_d6_mut_R1.out) 2> >(tee $WORK/Asp_xtb_d6_mut_R1.out)

cp fitness_* $WORK
mkdir Coord Out
mv *coord Coord
mv *out Out
bzip2 Coord/*
bzip2 Out/*
cd ..
mkdir -p $WORK/Asp_xtb_d6_mut_R1
rsync -az $LOCALDIR/Coord $LOCALDIR/Out $WORK/Asp_xtb_d6_mut_R1/
cp init_pop* $WORK/Asp_xtb_d6_mut_R1/
if [ $ret -ne 0 ] ; then
        echo "error copying files, $LOCALDIR on `hostname`"
else
```

```
        rm -rf $LOCALDIR
fi
cd $WORK
echo ‘hostname‘
echo "All Done"

exit 0
```

XTB template for Aspartic Acid

```
! cmd line start
!  --chrg 0 --parallel 4 --cma
! cmd line end
$coord angs
  0.000   0.000   0.000     N
  1.013   0.000   0.000     H
 -0.340   0.955   0.000     H
 -0.507  -0.728  -1.155     C
  0.109  -0.582  -2.057     H
 -0.552  -2.234  -0.866     C
 -1.021  -2.784  -1.688     H
 -1.141  -2.430   0.034     H
  0.831  -2.804  -0.666     C
  1.872  -2.227  -0.871     O
  0.782  -4.085  -0.235     O
  1.697  -4.393  -0.140     H
 -1.874  -0.178  -1.536     C
 -2.334   0.862  -1.140     O
 -2.516  -0.968  -2.428     O
 -3.350  -0.527  -2.654     H
$end
! coordinates end
$constrain
   dihedral: 3,1,4,6,auto
   dihedral: 1,4,6,9,auto
   dihedral: 4,6,9,11,auto
   dihedral: 1,4,13,15,auto
   dihedral: 4,13,15,16,auto
   dihedral: 6,9,11,12,auto
$end
! template end
```

Gaussian 16 DFTBA template for Aspartic Acid

```
#p dftba test nosymm pop=none geom=print
! route section end

Aspartic acid in vacuum

0 1
! header section end
 n
 h   1 hn2
 h    1 hn3          2 hnh3
 c    1 cn4          3 cnh4          2 dih4
 h    4 hc5          1 hcn5          2 dih5
 c    4 cc6          1 ccn6          2 dih6
 h    6 hc7          4 hcc7          1 dih7
 h    6 hc8          4 hcc8          1 dih8
 c    6 cc9          4 ccc9          1 dih9
 o    9 oc10         6 occ10         4 dih10
 o    9 oc11         6 occ11         4 dih11
 h   11 ho12         9 hoc12         6 dih12
 c    4 cc13         6 ccc13         9 dih13
 o   13 oc14         4 occ14         1 dih14
 o   13 oc15         4 occ15         1 dih15
 h   15 ho16        13 hoc16         4 dih16

! variable list
hn2          1.012929
hn3          1.014037
hnh3         109.605
cn4          1.456276
cnh4         110.721
dih4        -122.018
hc5          1.102596
hcn5         112.850
dih5         -36.464
cc6          1.534513
ccn6         110.587
dih6          83.905
hc7          1.094262
hcc7         111.382
dih7         174.521
hc8          1.093655
hcc8         110.265
dih8          55.468
cc9          1.509717
ccc9         111.631
dih9         -65.183
oc10         1.207413
occ10        125.968
```

```
dih10        -9.234
oc11        1.352441
occ11        111.494
dih11        171.631
ho12        0.970213
hoc12        107.341
dih12        179.186
cc13        1.521790
ccc13        112.036
dih13        172.577
oc14        1.203871
occ14        124.915
dih14         14.538
oc15        1.353711
occ15        112.339
dih15       -168.021
ho16         0.969644
hoc16        107.222
dih16       -176.917
! coordinates end

*  1  4 * F
*  4 13 * F
* 13 15 * F
*  4  6 * F
*  6  9 * F
*  9 11 * F

@/home/${USER}/bigd/ga_test/dftba.prm

! template end
```