Themis User Guide

version 1.0.0

Felippe Mariano Colombari

Themis: a software to assess association free energies via direct estimative of partition functions

Felippe M. Colombari, Asdrubal Lozada-Blanco, Kalil Bernardino,
Weverson R. Gomes and André F. de Moura

This manual is for Themis, version 1.0.0.

Table of Contents

1 About Themis **1**

2 Command line options **2**

3 Input files **3**

3.1 Coordinate files: `conf1.xyz` and `conf2.xyz` 3
3.2 Run control file: `INPUT` 3
3.3 Potential parameter files: `parameters1` and `parameters2` 5

4 Output files **7**

4.1 `energy.bin` file 7
4.2 `energy-sort.log` 7
4.3 `output.log` 7
4.4 `output-sort.log` 8
4.5 VMD script files: `surf_*.vmd` 8
4.6 `lowest_0001.xyz` 8
4.7 `grid_log.log` 9
4.8 `full_ensemble.xtc` 9
4.9 `point_0001_0001_0001.xyz` 10

5 Comparison with umbrella sampling **11**

6 Using QM interaction energies **12**

7 Performance benchmark **14**

8 References **15**

1 About Themis

Themis is a statistichal mechanics software designed to obtain the association thermodynamics of two structures (ions, molecules, crystals, nanoparticles, etc). It generates a configurational partition function by systematically sampling the phase space using discrete grids to perform translations and rotations of one structure around another. Interaction energy for each microstate can be obtained by one of the potentials implemented or by using external softwares.

Themis is a free software written in *Fortran 2003* language, being available at `https://github.com/colombarifm/themis` under the *GPLv3+* License. It runs under Linux environment with gfortran/gcc 5.4+ compilers. Since it was written in modules, new potential functions and analysis routines can be easily implemented.

Detailed information about the methodology and some applications of current implementation is on ChemRxiv:

Themis: a software to assess association free energies via direct estimative of partition functions.

Please, see also other papers that used earlier or adapted *Themis* versions:

Emergence of complexity in hierarchically organized chiral particles.
Science, v. 368, p. 642, 2020, doi: 10.1126/science.aaz7949.

Solvent effect on the regulation of urea hydrolysis reactions by copper complexes.
Chemistry, v. 2, p. 525, 2020, doi: 10.3390/chemistry2020032.

Graphitic Carbon Nitrides as Platforms for Single-Atom Photocatalysis.
Faraday Discussions, v. xxx, p. xxx-xxx, 2020, doi: 10.1039/C9FD00112C.

Ion pair free energy surface as a probe of ionic liquid structure.
The Journal of Chemical Physics, v. 152, p. 014103, 2020, doi: 10.1063/1.5128693.

Low-Temperature Phase Transitions of the Ionic Liquid 1-Ethyl-3-methylimidazolium Dicyanamide.
The Journal of Physical Chemistry B, v. 123, p. 9418, 2019, doi: 10.1021/acs.jpcb.9b07654.

Site-selective photoinduced cleavage and profiling of DNA by chiral semiconductor nanoparticles.
Nature Chemistry, v. 10, p. 821, 2018, doi: 10.1038/s41557-018-0083-y.

2 Command line options

Themis usage is done via Linux command line as follows:

```
themis [RUNTYPE] [GRID]
```

[RUNTYPE] options are:

`--run`

to start a new calculation;

`--rerun`

to calculate properties from interaction energies obtained previously. In this case, an `energy.bin` file will be read if these energy values were obtained with *Themis* or an `energy.log` file will be read if these energy values were obtained externally. While the former is useful to obtain thermodynamic properties using a different temperature from a previous calculation, the latter is useful to obtain thermodynamic properties using quantum chemistry interaction energies.

[GRID] options are:

`--shell <radius>`

indicates that translation moves will be performed on a spherical shell around the reference molecule (generated on the run). The real argument `<radius>` is the scaling factor for the radius (in Angstrom).

`--user <file.xyz>`

indicates that translation moves will be performed on an user-defined grid read from `<file.xyz>`. It must be aligned with molecule 1 and can be generated using the `sas_grid` utility (found in the `utils/` folder.

Other valid options are `--help`, `--citation` and `--license`.

3 Input files

3.1 Coordinate files: `conf1.xyz` and `conf2.xyz`

Standard `.xyz` files containing the cartesian coordinates of both structures. For the water dimer, a dummy site (X) corresponding to water center of mass was used to define the rotation axis of MOL1.

```
     4

O        0.00000      0.06682      0.00000
H       -0.76677     -0.53032      0.00000
H        0.76677     -0.53032      0.00000
X        0.00000      0.00000      0.00000
```

3.2 Run control file: `INPUT`

Plain text file containing detailed intructions prior to calculation. It must contain the following keywords, in any order (the *:* symbol is the separator):

`rot1_factor :`

Parameter (p) used to generate the spherical grid used for reorientation moves. The number of points (n) obtained along the sphere surface by dodecahedron tessellation (as shown in Figure 1) is given by $n = 12 + 10 \times 3 \times (p-1) + 10 \times (p-2) \times (p-1)$. If one uses $p = 0$, the reorientation move will then correspond to align molecule 2 along Z-axis (1 reorientational move).

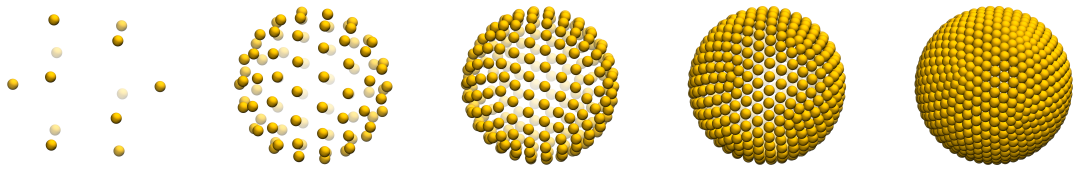

Figure 1. Translation grids obtained by sphere tesselation. From left to right: p = 1 (n = 12), p = 3 (n = 92), p = 5 (n = 252), p = 7 (n = 492) and p = 9 (n = 812).

`translation_factor :`

Same as `rot1_factor` if a spherical translation shell is used (`--shell <radius>`).

`rot2_factor` :

Corresponds to the number of rotation moves around the rotation axis of MOL2.

`rot2_range` :

Corresponds to the maximum rotation angle (in degrees) around the rotation axis of MOL2.

`temperature` :

Absolute temperature (in K) used to calculate all thermodynamic properties and probabilities.

`potential` :

Potential energy function selection. Current options are "`none`", "`lj-coul`" and "`bh-coul`".

`write_frames` :

Selects the format in which all valid frames will be written: "`XYZ`", "`MOP`" and "`none`". If "`MOP`" is selected, the optional character variable containing the first line of *MOPAC* input (`mopac_job`) is read.

`ref_mol1` :

Site of molecule 1 used for centering, according to `conf1.xyz` file.

`rot_ref_mol1` :

Site of molecule 1 that will build its rotation vector, according to `conf1.xyz` file.

`ref_mol2` :

Site of molecule 2 used for centering, according to `conf2.xyz` file.

`rot_ref_mol2` :

Site of molecule 2 that will build its rotation vector, according to `conf2.xyz` file.

`shortest_distance` :

Corresponds to the lowest intermolecular distance to consider the configuration as a valid one. Below such value (in Angstrom), molecular contacts are considered strongly repulsive and an interaction energy value of 10^{10} kJ/mol is attributed to such configuration. This is useful to avoid spending time calculating energies for unphisical configurations since the energy loop is skipped.

`write_xtc` :

Flag to enable the writting of all configurations to a `.XTC` file. WARNING: very large files can be generated ;)

`lowest_structures :`

Selects the number of lowest energy/highest probability structures to write after the run.

`mopac_job :`

String containing the header for mopac calculations. Enabled when "`write_frames : MOP`" is selected.

`cutoff_distance :`

Maximum intermolecular distance to calculate interaction energies. Beyond this value, interaction energies are cutoff to 0. This could generate serious artifacts. If "`cutoff_values : 0.0`", all interactions are computed.

Thus, a simple and quick calculation for the water dimer mentioned earlier can be executed with the following `INPUT` file:

```
translation_factor :  2
rot1_factor :  2
rot2_factor :    120
rot2_range :  360.0
temperature :  300
potential :  lj-coul
ref_mol1 :  1
rot_ref_mol1 :  4
ref_mol2 :  1
rot_ref_mol2 :  2
shortest_distance :  1.2
write_xtc :  no
lowest_structures :  10
write_frames :  none
cutoff_distance :  0.0
```

3.3 Potential parameter files: `parameters1` and `parameters2`

Plain text files containing potential parameters used for energy calculations. Those files are read differently according to the potential used. For Lennard-Jones + Coulomb interaction potential (invoked by `potential : lj-coul`), one should provide q_i (e), σ_i $(\mathring{A})$ and ϵ_i (kJ/mol) parameters, according to Equation 1.

$$U_{ljc} = \sum_i \sum_{j<i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{1}{4\pi\varepsilon_0} \sum_i \sum_{j<i} \frac{q_i q_j}{r_{ij}} \tag{1}$$

where $\epsilon_{ij} = (\epsilon_i \cdot \epsilon_j)^{\frac{1}{2}}$ and $\sigma_{ij} = (\sigma_i \cdot \sigma_j)^{\frac{1}{2}}$. TIP3P parameter files for water are read as follows:

```
##   q       sigma     epsilon
OW  -0.834   3.15061   0.636386
HW   0.417   0.00000   0.000000
HW   0.417   0.00000   0.000000
XX   0.000   0.00000   0.000000
```

For Buckingham + Coulomb interaction potential, according to Matsui[1], one should invoke `potential : bh-coul` and provide q (e), A_i ($\mathring{A}$), B_i ($\mathring{A}$) and C_i ($\mathring{A}^3 kJ/mol$) parameters according to Equation 2.

$$U_{bhc} = \sum_i \sum_{j<i} \left\{ \left(\frac{-C_i C_j}{r_{ij}^6} \right) + f(B_i + B_j) \exp\left[\left(\frac{A_i + A_j - r_{ij}}{B_i + B_j} \right) \right] \right\} + \frac{1}{4\pi\varepsilon_0} \sum_i \sum_{j<i} \frac{q_i q_j}{r_{ij}} \tag{2}$$

where the quantity f corresponds to a standard force of 4.184 $kJ/mol/\mathring{A}$. Parameters for a TiO_2 unit must be provided as follows:

```
#    q       A        B        C
Ti   2.196   1.1823   0.0770   22.500
O   -1.098   1.6339   0.1170   54.000
O   -1.098   1.6339   0.1170   54.000
X    0.000   0.0000   0.0000    0.000
```

NOTE: It is important to highlight that atoms described in both `parameters1` and `parameters2` files must be in the same order as they appear in both `conf1.xyz` and `conf2.xyz` files. Parameters file must contain a header followed by one line for each atom descrived in structure files.

4 Output files

4.1 `energy.bin` file

Binary file containing interaction energy values for all microstates. Since all entries are written in the right loop sequence, they can be read using the `rerun` feature to calculate thermodynamic properties in a different temperature, or to write a different number of lowest-energy structures.

4.2 `energy-sort.log`

Contains interaction energy values and probabilities for the N most probable structures. By running *Themis* with the input files for the water dimer presented previously, and considering a spherical grid with radius = 2.8 $\AA$, one obtains

#int_energy(r2,r1,t)	r2	r1	t	prob.	sum prob.
-2.83500E+001	1	10	3	6.704E-004	6.704E-004
-2.83500E+001	1	4	9	6.704E-004	1.341E-003
-2.83453E+001	2	4	9	6.692E-004	2.010E-003
-2.83453E+001	2	10	3	6.692E-004	2.679E-003
-2.83453E+001	120	4	9	6.692E-004	3.348E-003
-2.83453E+001	120	10	3	6.692E-004	4.017E-003
-2.83312E+001	3	10	3	6.654E-004	4.683E-003
-2.83312E+001	119	10	3	6.654E-004	5.348E-003
-2.83312E+001	3	4	9	6.654E-004	6.014E-003
-2.83312E+001	119	4	9	6.654E-004	6.679E-003

which indicates that the 10 lowest-energy structures have interaction energies close to each other and presented a cumulative probability of 0.67 %.

4.3 `output.log`

Contains thermodynamic data for each point of the translation grid, and also for the overall ensemble. Written in an extended `.XYZ` format containing extra field values for each grid point

(probability, free energy, energy and entropic penalty). For the same example, part of the file reads as

```
           42
      X (A)    Y (A)    Z (A)   t        PROB   A (kJ/mol)  -TS (kJ/mol)  E (kJ/mol)
X   2.38182  1.47205  0.00000   1  2.9885E-03  -1.0813E+01   6.7592E+00  -1.7573E+01
X   2.38182 -1.47205  0.00000   2  2.9885E-03  -1.0813E+01   6.7592E+00  -1.7573E+01
X   1.47205  0.00000  2.38182   3  5.8517E-02  -1.8233E+01   6.8417E+00  -2.5075E+01
X   1.47205  0.00000 -2.38182   4  1.3400E-03  -8.8128E+00   4.1958E+00  -1.3009E+01
X   0.00000  2.38182  1.47205   5  1.0097E-02  -1.3850E+01   6.7855E+00  -2.0636E+01
X   0.00000  2.38182 -1.47205   6  1.7473E-01  -2.0962E+01   4.3203E+00  -2.5282E+01
X   0.00000 -2.38182  1.47205   7  1.0097E-02  -1.3850E+01   6.7855E+00  -2.0636E+01
X   0.00000 -2.38182 -1.47205   8  1.7473E-01  -2.0962E+01   4.3203E+00  -2.5282E+01
   ...
X  -2.26525  0.86525 -1.40000  40  9.8346E-04  -8.0411E+00   5.5717E+00  -1.3613E+01
X  -2.26525 -0.86525 -1.40000  41  9.8346E-04  -8.0411E+00   5.5717E+00  -1.3613E+01
X  -2.80000  0.00000  0.00000  42  2.9720E-03  -1.0800E+01   6.9719E+00  -1.7772E+01
-------------------------------------------------------------------------------------
TOTAL OVER TRANSLATIONAL GRID      1.0000E+00  -1.5990E+01   7.6750E+00  -2.3665E+01
```

4.4 `output-sort.log`

Same as `output.log` but ordered from most probable point to the least probable point.

4.5 VMD script files: `surf_*.vmd`

Before a successful termination, *Themis* writes three `.vmd` script files containing instructions for reading the thermodynamic data along the translation grid from file `output.log`: `surf_free-energy.vmd`, `surf_energy.log` and `surf_entropic-penalty.log`. Besides rendering these thermodynamic properties, they also render the most stable structure (`lowest_0001.xyz` file) over such visualization.

4.6 `lowest_0001.xyz`

Cartesian coordinates of the most probable structure from the whole ensemble. The number of lowest structure files is defined by the user in the `INPUT` file.

```
  7
Energy =  -2.8350000E+01
O         0.0000     0.0000     0.0000
H        -0.0000    -0.7668    -0.5971
H        -0.0000     0.7668    -0.5971
X        -0.0000    -0.0000    -0.0668
O         1.4720     0.0000     2.3818
H         0.9611     0.0000     1.5551
H         0.7957     0.0000     3.0797
```

4.7 grid_log.log

File containing informations of each translation grid point: point number,number of rejected structures (due to atomic clashes) and spent time.

```
t point   rejected structures   time (s)
      1         0 of    5040      0.020
      2         0 of    5040      0.012
      3         0 of    5040      0.012
      4         0 of    5040      0.013
      ...
     38         0 of    5040      0.024
     39         0 of    5040      0.013
     40         0 of    5040      0.013
     41         0 of    5040      0.012
     42         0 of    5040      0.014
```

4.8 full_ensemble.xtc

Trajectory file in `.XTC` format containing the whole ensemble. Written if `INPUT` option `write_xtc` is enabled.

4.9 point_0001_0001_0001.xyz

File containing the structure of the microstate $t = 1$, $r1 = 1$, $r2 = 1$ in cartesian coordinates. These files are numbered according to the loop position. Written if `INPUT` option `write_frames = XYZ` is set. Microstates with intermolecular distances below the one defined by `shortest_distance` are skipped. **WARNING**: this option will create a very large number of files in the directory. ;)

5 Comparison with umbrella sampling

The sampling along the OW-OW separation coordinate (ξ) was carried by running multiple independent simulations with a biased umbrella potential that restrains the water-water intermolecular distance. For each run, a short energy minimization was carried in order to remove repulsive contacts (especially for short intermolecular distances). The sampling was carried out in vacuum using a stochastic dynamics integrator (T = 300 K). Lennard-Jones and electrostatic interactions were computed in direct space without a cutoff. Bonds were constrained using LINCS allowing a time step of 1 fs. ξ_i was sampled from 2 $\mathring{A}$ to 15 $\mathring{A}$ in 0.5 $\mathring{A}$ intervals. An umbrella force constant of 10^4 kJ/mol/nm was set for 3 $\mathring{A} \leq \xi_i \leq 15$ $\mathring{A}$, and 2×10^4 kJ/mol/nm for 2 $\mathring{A} \leq \xi_i < 3$ $\mathring{A}$. Calculations were performed using the GROMACS 5.1.2 package.[2,3] Entropic and energetic contributions to the potential of mean force were obtained by finite differences, considering simulations at different temperatures (Equation 3).

$$\Delta S(\xi) = -\left(\frac{\partial \Delta A(\xi, T)}{\partial T}\right) \tag{3}$$

Although free energy profiles presented differences smaller than 0.2 kJ/mol when 100 ns and 500 ns simulations were compared, energetic and entropic contributions presented poor convergence along the profiles for smaller simulation times (Figure 2).

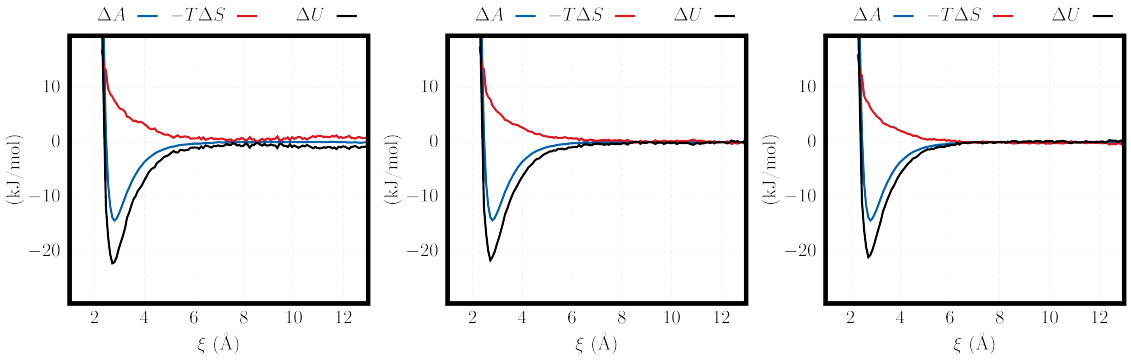

Figure 2. Convergence of water dimer PMF along the MD simulations. Profiles obtained for 100 ns (left), 300 ns (center) and 500 ns (right).

Note that MD simulations deliver highly correlated microstates an so it is not necessary to write and/or analyze all energy points. Thus, every 1000^{th} energy point was saved for analysis, resulting in a ensemble of 5×10^5 energy points *per* ξ distance and 6.55×10^7 configurations altogether.

6 Using QM interaction energies

Calculations using *Themis* + *MOPAC* were performed for the biphenyl dimers (please see *Themis* main reference for more details) in multiple steps. For each intermolecular distance:

i. Write *MOPAC* input files for all valid configurations using the following `INPUT` options along with the following coordinate files `conf1.xyz` and `conf2.xyz` for P-Biphenyl

```
translation_factor :  3
rot1_factor :  4
rot2_factor :  36
rot2_range :  360.0
temperature :  300.0
potential :  none
ref_mol1 :  23
rot_ref_mol1 :  1
ref_mol2 :  23
rot_ref_mol2 :  1
shortest_distance :  1.2
write_xtc :  no
lowest_structures :  10
write_frames :  MOP
mopac_job :  MOP PM7 1scf output threads=1 shift=1.0 itry=150
cutoff_distance :  0.0
```

```
        23

C   0.366897    0.635707    0.002701
C   1.764297    0.635707    0.002701
C   2.459097    1.842607    0.002701
C   1.762897    3.050407    0.007001
C   0.368897    3.051107    0.008901
C  -0.330403    1.846807    0.004401
H   2.305797   -0.308493   -0.003899
H   3.547197    1.841707   -0.000899
H   2.307297    3.991907    0.008801
H  -0.174803    3.993707    0.014001
H  -1.418903    1.845007    0.009001
C  -0.367803   -0.635193   -0.000699
C  -0.148703   -1.572593    1.012201
C  -0.847803   -2.776993    1.007401
C  -1.762103   -3.050793   -0.008599
C  -1.979403   -2.117193   -1.020699
C  -1.286003   -0.909493   -1.017699
H   0.563497   -1.355393    1.806201
H  -0.679203   -3.504793    1.798601
H  -2.305903   -3.992693   -0.011599
```

```
H  -2.691603   -2.331493   -1.814999
H  -1.451703   -0.180393   -1.808899
X   0.000000    0.000000    0.000000
```

By running *Themis*, a large number of numbered *MOPAC* files will be generated, starting from `point_0001_0001_0001.mop` until `point_0092_0162_0036.mop`. Configurations that presented intermolecular distances below 1.2 $\mathring{A}$ are not written.

ii. Run the single-point calculation for every `.mop` file. This can be done more efficiently using the GNU Parallel tool;[4]

iii. Once finished, a python script was used to extract the final heat of formation of every output file (numbered as described above) and generate an `energy.log` file containing all interaction energies. A large repulsive energy is attributed to configurations whose input files were not written or single-point calculations did not converge;

iv. Themis `--rerun` option must be used to read all required files used previously altogether with the energy values written at `energy.log` file. It will also calculate all thermodynamic properties and search for the most stable structures.

7 Performance benchmark

In order to analyze the effect of grid coarseness on both computation time and thermodynamic results, the association thermodynamics for (L)-CYS dimer was obtained using different grids for translation and $\text{rot}_{\text{point}}$. Considering $\text{nr}_2 = 120$ and a total of 167 separation distances (from 20.0 $\mathring{A}$ to 3.4 $\mathring{A}$, in 0.1 $\mathring{A}$ intervals), the number of microstates of the whole ensemble ranges from $\approx 3.5 \times 10^7$ (grid factors = 2) to $\approx 2.0 \times 10^{10}$ (grid factors = 10). This large difference results in wall-times ranging from ≈ 6 min to ≈ 2 days (as shown in Figure 3, top-left), which requires a compromise between computational cost and accuracy.

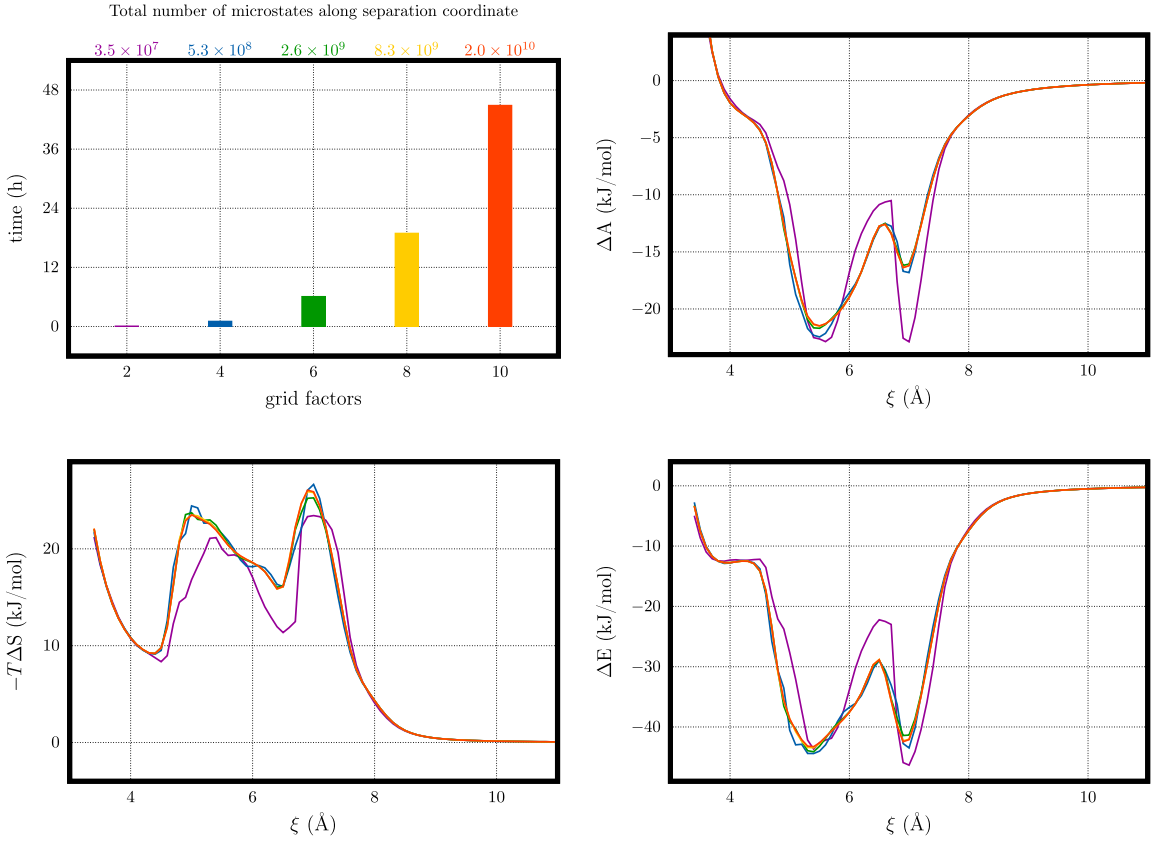

Figure 3. Comparison of calculation wall-time (in hours) and thermodynamic properties as a function of the grid coarseness for the association of (L)-CYS dimers.

As one can notice, the cheapest calculation (grid factor = 2, purple curves) resulted in thermodynamic profiles considerably different, due to poorly sampling phase space regions with higher entropic loss. For grid factors = 4 (blue curves), although results are improved, one can still observe noticeable differences in comparison to the more costly calculations. On the other hand, for grid factors >= 6, only small differences are observed, indicating a good convergence for all thermodynamic profiles.

8 References

1. Matsui, M.; Akaogi, M. *Molecular Dynamics Simulation of the Structural and Physical Properties of the Four Polymorphs of TiO_2.* **Molecular Simulation**, v. 6, p. 239, 1991.

2. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *GROMACS: Fast, flexible, and free.* **Journal of Computational Chemistry**, v. 26, p. 1701, 2005.

3. Abraham, M. J.; Murtola, T.; Schulz, R.; Pll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers.* **SoftwareX**, v. 1, p. 19, 2015.

4. Tange, O. *GNU Parallel - The Command-Line Power Tool.* **;login: The USENIX Magazine**, v. 36, p. 42, 2011.