Michele Ceriotti, michele.ceriotti@epfl.ch

Hands-on exercises on a Lennard-Jones 38 cluster

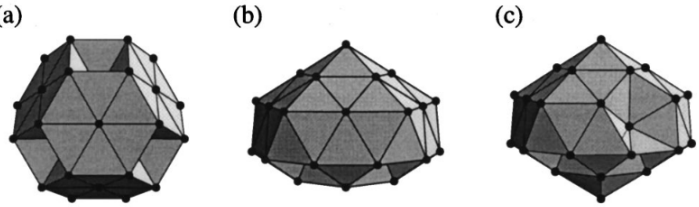
The system that will be examined is a clusters of 38 atoms interacting with a simple Lennard-Jones (LJ) potential:

$$U(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right]$$

where ϵ is the well depth and $2^{\frac{1}{6}}\sigma$ is the equilibrium separation for a diatomic molecule. This potential describes dipole fluctuation attractive interactions that decay as r^{-6}, and a somewhat arbitrary r^{-12} repulsive wall at short inter-atomic separations, that models the Pauli repulsion between electron clouds. The Lennard-Jones potential is a good model for the interaction between noble gases atoms, but here we will use it just as an inexpensive model of an isotropic pair-wise interaction between atoms. In all the exercises reduced units will be used, that correspond to measuring energies and temperatures in units of ϵ, distances in units of σ, time in units of $t^* = \left(\frac{\epsilon}{m\sigma^2}\right)$, and temperature in units of $T^* = \left(\frac{k_B}{\epsilon}\right)$. In practice this amounts at setting to one most constants: in principle all results can be scaled to the physical values for a particular system by setting the appropriate mass, well depth and equilibrium distance.

A LJ cluster provides a particularly useful model system since for small LJ clusters a complete enumeration of the minima and transition states allows a detailed view of the potential energy landscape [1]. In particular the LJ_{38} , the cluster which we study here, has a double-funnel landscape: the global minimum is a face-centered-cubic (fcc) truncated octahedron (Fig. 1(a)) and the second lowest energy minimum is an incomplete Mackay icosahedron (Fig. 1(b)).

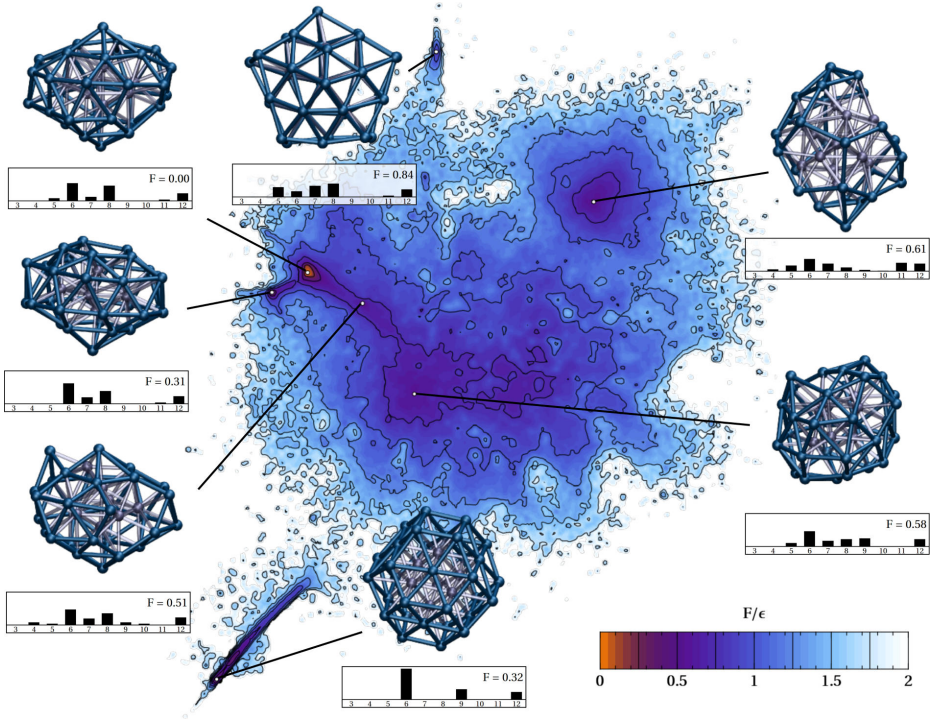
Figure 1: (a) The LJ_{38} global minimum, an fcc truncated octahedron. (b) and (c) Second lowest energy minimum of LJ_{38}. Figure adapted from [2].



There is thus a solid–solid transition at moderate temperatures and a subsequent solid–liquid transition at higher temperatures. The solid–solid transition occurs because the energy landscape in the high-temperature phase is flatter, resulting in a larger entropic contribution to the free energy of this structure, and to its stabilization at moderate temperatures relative to the *fcc* minimum [3]. Figure 2 shows a few selected configurations for this system, together with the free energy computed close to the solid-liquid transition temperature. The stability of different configurations depends dramatically on the simulation temperature, and the time scales for transitions is long, but accessible to direct

simulation thanks to the small size and inexpensive potential. This makes LJ_{38} an ideal example to show the strength and pitfalls of different sampling techniques in atomistic simulations.

Figure 2: A number of representative configurations of the LJ_{38} cluster are projected together with the free energy surface computed at $0.18T^*$. Figure adapted from [3]



Sketch-map analysis

In this exercise we will focus on post-processing the results of molecular dynamics simulations to determine machine-learning collective variables (CV) that can give you a better understanding of the configuration space of LJ_{38}, and of the shortcomings of conventional CVs even in such a simple system.

We will use the sketch-map method [3], and the suite of programs that can be downloaded from `http://epfl-cosmo.github.io/sketchmap`.

We will first extract a set $\{X_i\}$ of reference configurations, described in terms of the set of all coordination counts between 4 and 13, and find their sketch-map projection by minimizing the objective function

$$\chi^2 = \sum_{i,j=1}^{N} \left[s\left(|X_i - X_j|\right) - s\left(|x_i - x_j|\right) \right]^2 . \tag{1}$$

Having obtained this map, one can more easily find the out-of-sample embedding of the remaining points, and analyze the simulation based on the map obtained by machine-learning.

Extracting landmark points

Since the cost of minimizing iteratively Eq. (1) grows quickly with the number of reference points, so it is important to select a sub-set of the data from which one wants to build the map. The `mdcode` generates a set of configurations by running molecular dynamics, and also compute a smooth histogram of the coordination numbers of the atoms in the cluster

$$n_c = \sum_{i=1}^{N} e^{-\frac{(c_i - c)^2}{2\sigma^2}},$$

where c_i is the coordination number of atom i, defined in turn as

$$c_i = \sum_j \mathcal{C}\left(|\mathbf{q}_i - \mathbf{q}_j|\right), \qquad \mathcal{C}\left(d\right) = \begin{cases} 0 & d > r_0 \\ 1 & d < r_1 \\ (y-1)^2\,(2y+1) & r_1 < d < r_0, \quad y = \frac{d-r_1}{r_0 - r_1} \end{cases}.$$

The variable n_c corresponds approximately to the number of atoms in the structure with a coordination number around the value c. Run the code with

```
$ mdcode input.cv > log
```

Then, use `dimlandmark` to extract 500 reference points from the dataset:

```
$ awk '{if (NR%10==0) print $0}' out.all | \
   dimlandmark -D 10 -n 500 -w -unique -mode staged \
   -gamma 0.5 -wgamma 0.5 > lj38.lm
```

Read the help string of the program (`dimlandmark -h`) to get a description of the options. Note that we pre-select a subset of the data set to speed up the evaluation. The *staged* mode of selecting the landmarks is a two-step procedure in which a larger set of points is chosen, approximately uniformly spaced one from another. Then, the probability density in the D dimensional space is estimated, and it is used to select the n landmark points according to $P\left(X\right)^{\gamma}$. You can try to visualize the selected landmarks with `gnuplot`, to see how they change when the value of γ is modified

```
gnuplot> p 'lj38.lm' u 3:5 w p
```

Analyzing the distribution of data points

Sketch-map is based on restricting the similarity matching that underlies MDS methods to the range of distances that characterize adjacent meta-stable states. Compute the distribution of individual CVs, e.g. by

```
$ awk '{print $3}' out.all | \
      histogram -xi 0 -xf 25 -n 500 -t 0.1 > n6
```

and look at the amplitude of fluctuations around maxima of the histogram in each dimension. Note that there are multiple scales in the fluctuations in probability density – sharp peaks with a width of about 3 units, superimposed with broad features with much larger breadth of about 5 units. This is a common problem when analyzing atomistic data in a glassy free-energy landscape – multiple shallow minima are grouped together to form extended regions of similar, closely-related structures. Also, consider that fluctuations in D dimensions are approximately $\sqrt{D}$ times broader than their one-dimensional projections, so depending on the resolution one wants to achieve one might want to select a threshold parameter for sketch-map between 1.5 and 10 units.

You may also want to compute the histogram of D-dimensional distances. Typically one wants to select a cutoff somewhere before the maximum of the distribution, which is dominated by long-range features

```
$ awk '{if (NR%10==0) print $0}' out.all > tmp
$ dimdist -D 10 -P tmp -maxd 20 -nbin 200 -lowmem > lj38.histo
```

Sketch map uses a sigmoid function defined as

$$s(r) = 1 - \left(1 + \left(2^{a/b} - 1\right)(r/\sigma)^{a}\right)^{-b/a}.$$

You may want to plot it superimposed with the histogram of pairwise distances to get a feeling of how different sets of distances are transformed by the function:

```
gnuplot> f(r,s,a,b)=1-(1.0+(2**(a*1.0/b)-1)*(r/s)**a)**-(b*1.0/a)
gnuplot> p 'j38.histo' u 1:($2*10) w l, f(x,5,8,1)
```

Run sketch-map

Being an iterative minimization scheme, sketch-map requires a decent starting configuration and an optimizer that can get out of local minima to approach a global optimum. Without getting into details, you can use a simple script that takes care of the optimization procedure. Remember to delete the two-lines header of `lj38.lm` before proceeding.

```
$ sketch-map.sh
Please enter the dimensionality of input data 10
Are points weighted [y/n]? y
Please enter the periodicity of input data [0 if non-periodic] 0
Please enter the input data file name lj38.lm
Please enter the output data prefix lj38.5_8_1-5_2_2
Please enter high dimension sigma, a, b [e.g. 6.0 2 6 ] 5 8 1
Please enter low  dimension sigma, a, b [e.g. 6.0 2 6 ] 5 2 2
```

You can then remove intermediate files and diagnostics, and assemble a file with just the low-dimensional coordinates of the landmarks

```
$ rm log global.* lj38.5_8_1-5_2_2.*[0-9]
$ awk '!/#/{print $1, $2}' lj38.5_8_1-5_2_2.gmds > lj38.ld
```

- Visualize the projected points, and verify by coloring how the sketch-map coordinates separate clearly points with different n_ks

  ```
  gnuplot> p '< paste lj38.ld lj38.lm' u 1:2:5 w p pt 7 lt pal
  ```

- [OPTIONAL] Try to run projections with different parameters – σ, a, b – and verify how much the projection changes

Out-of-sample embedding

Having obtained a map that assigns to each of the high-dimensional landmark points a corresponding embedding, one can proceed to project all the data in the original trajectory. You should use the utility `dimproj` to do so, specifying the high and low-dimensional references, the sketch-map parameters and giving in input the full trajectory:

```
$ dimproj -D 10 -d 2 -P lj38.lm  -p lj38.ld -w -grid 15,16,151 \
   -cgmin 3 -fun-hd 5,8,1 -fun-ld 5,2,2 < out.all > \
   out.proj 2> /dev/null
```

Make sure to discard the error log: this is development code and outputs a lot of junk that you don't need to worry about at this stage.

- Compute the free energy from the sketch-map projection. A (long) one-liner to do this is as follows

```
$ awk '{print $1, $2}' out.proj | ndhistogram -d 2 -g -xi -15,-15 \
   -xf 15,15 -n 150,150 -t 0.2,0.2 -adaptive 0.25 | \
   awk -v kt=0.168 'BEGIN{print "# s1  s2  F(s1,s2)" } \
   !/#/{ if (NF==0) print ""; else printf "%15.7e %15.7e %15.7e\n", \
   $1, $2, -kt*log($3) } ' >  smap.fes
```

Bibliography

[1] David Wales. *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.

[2] Jonathan P. K. Doye, Mark A. Miller, and David J. Wales. The double-funnel energy landscape of the 38-atom lennard-jones cluster. *The Journal of Chemical Physics*, 110(14):6896, 1999.

[3] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Demonstrating the transferability and the descriptive power of sketch-map. *Journal of Chemical Theory and Computation*, 9(3):1521–1532, March 2013.