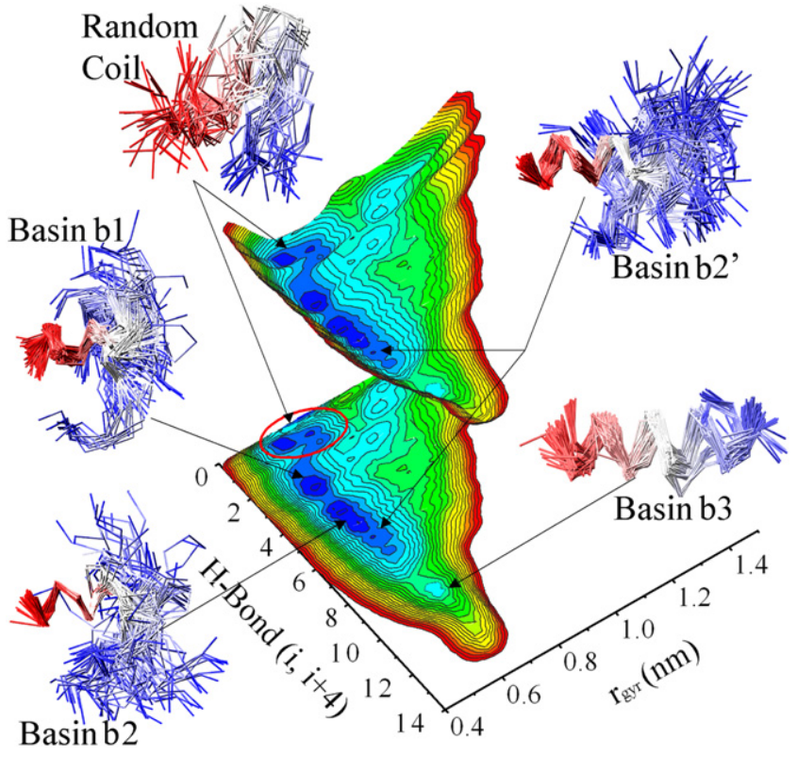
Importance Sampling and Umbrella Sampling

Carlo Camilloni

Sampling

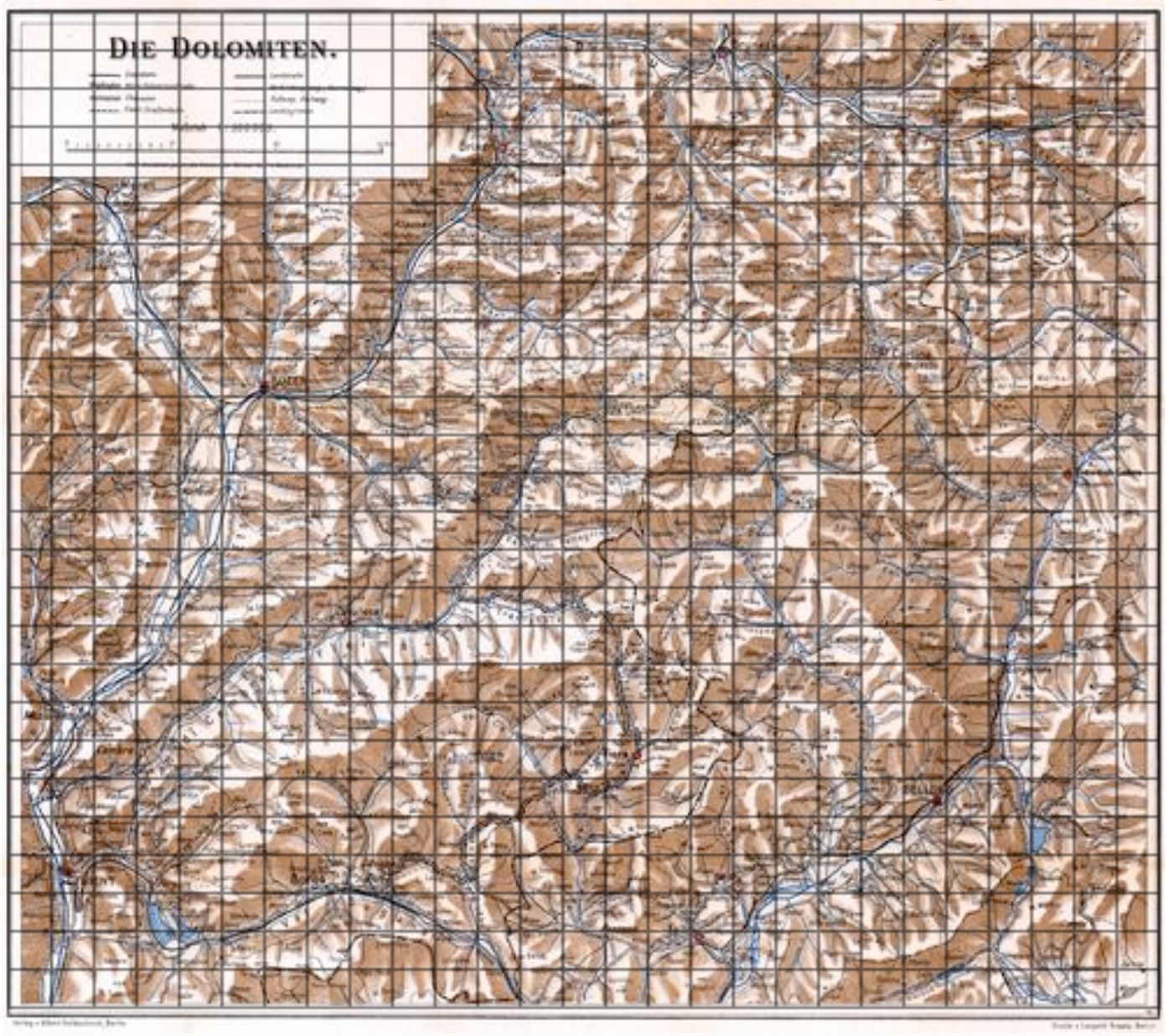
Sampling allows to learn some information on a large population by merely looking at a comparably small number of individuals.

In the case of a disordered proteins for example we would like to now the distribution of the secondary structure populations and that of the radius of gyration, while maybe on a folded protein the probability of observing “minorities” like the so called excited states.



Importance vs Random Sampling

The number of people living on the Dolomites?



Sampling

In the any thermodynamic ensemble the goal of sampling is that of acquiring an accurate estimate of functions of the partition function.

in the canonical ensemble

Maxwell-Boltzmann Distribution of the velocities

$$P(q,p) = \frac{e^{-H(q,p)/kT}}{\int e^{-H(q,p)/kT} dqdp} = \frac{e^{-U(q)/kT}}{\int e^{-U(q)/kT} dq} * \frac{e^{-K(p)/kT}}{\int e^{-K(p)/kT} dp} = P(q) * P(p)$$

$$P(q) = \frac{e^{-U(q)/kT}}{\int e^{-U(q)/kT} dq}$$

Configurational probability

Any configurational equilibrium observable is an integral calculated over the configurational probability

Importance Sampling

$$< O(q) >_p = \int P(q)O(q)dq = \int \frac{P(q)O(q)\rho(q)}{\rho(q)}dq = < \frac{O(q)P(q)}{\rho(q)} >_\rho$$

In order to sample the former integrals we can in principle:

1. Enumerate all the possible conformations and then calculate the needed function
2. Generate random conformations (with a uniform probability distribution)
3. Generate conformations from any other known probability distribution

We have ‘importance sampling’ every time we are not using the uniform probability distribution

MC and MD are special case of importance sampling.

Importance Sampling

In MC with metropolis and MD with a thermostat we in the condition for which in the limit t->Infinity the system is sampling

$$< O(q) >_p = \int \frac{P(q)O(q)P^{MD}(q)}{P^{MD}(q)} dq = < \frac{O(q)P(q)}{P^{MD}(q)} >_{P^{MD}} = < O(q) >_{P^{MD}}$$

This assumption is correct only if the MD algorithm is correct

Free Energy and Biases

$$P(q) = \frac{e^{-U(q)/kT}}{\int e^{-U(q)/kT} dq}$$

Configurational probability

The problem of this quantity is that it is a bit highly dimensional (3N atoms) so not exactly intuitive.

We are usually interested in more intuitive informations like the probability to observe a radius of gyration, or an angle, or a RMSD, etc. These properties, that are functions of the positions are generically called **collective variables s(q)**

$$P(s) \propto \int dq e^{-\frac{U(q)}{k_B T}} \delta(s - s(q))$$

From this mono or few dimensional probability distribution we can now define a free energy landscape

$$F(s) = -k_B T \log P(s)$$

If the above relation is used then the free energy is estimate but for an additive constant

$$P'(s) \propto \int dq e^{-\frac{U(q)+V(s(q))}{k_B T}} \delta(s - s(q)) \propto e^{-\frac{V(s(q))}{k_B T}} P(s)$$

$$F'(s) = -k_B T \log P'(s) = F(s) + V(s) + C$$

$$P(q) \propto P'(q) e^{\frac{V(s(q))}{k_B T}} \qquad w \propto e^{\frac{V(s(q))}{k_B T}}$$

The addition of a biasing potential reweighs the probability of observing all conformations and has an additive effect on the free energy

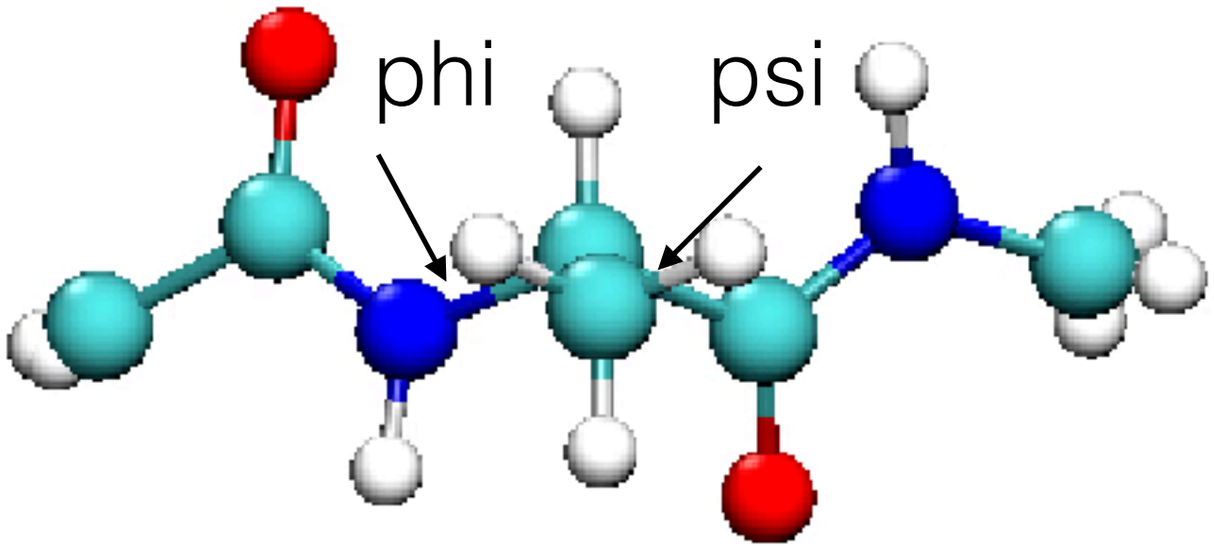
If the bias is constant in time and the sampling is complete it is possible to remove its effect by assigning a new weight to each sampled frame.

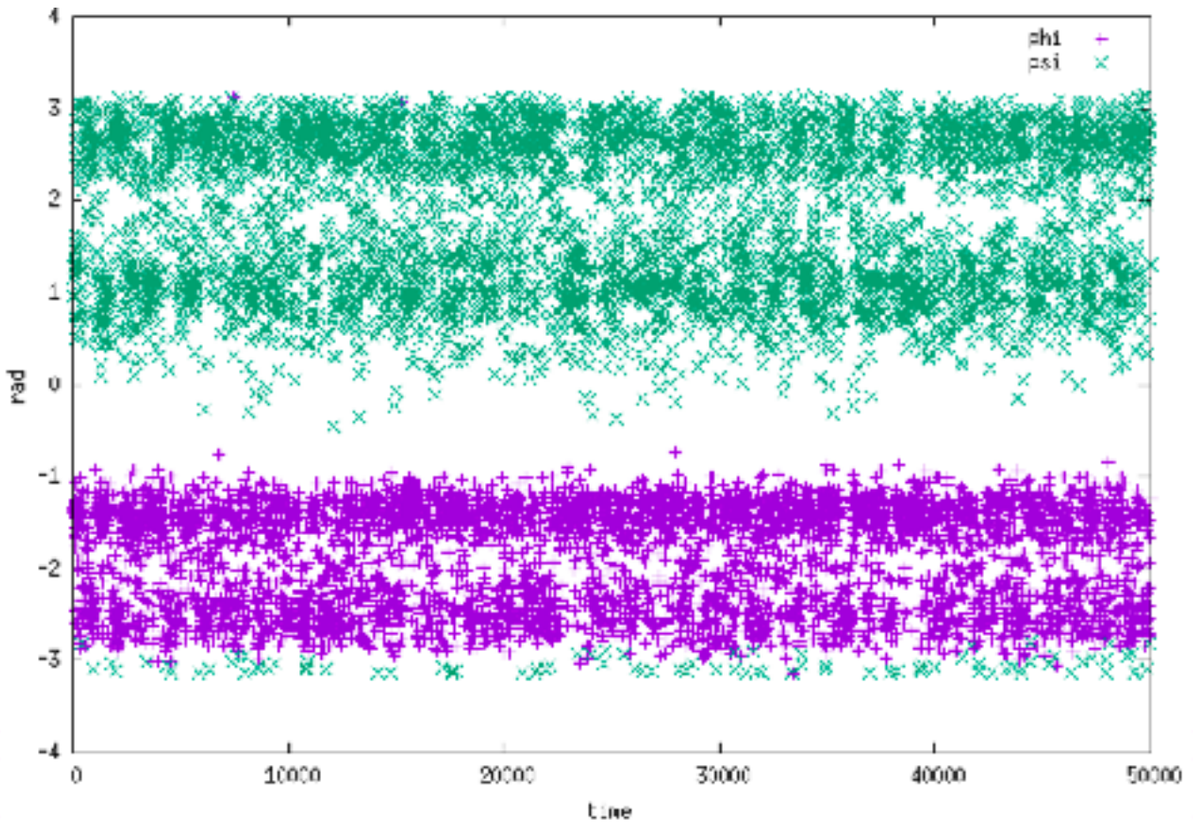
Umbrella Sampling

The original idea of umbrella sampling is that of building a probability distribution that will help sampling the wanted integral. Such a probability should bridge the unknown free energy with an optimal distribution keeping a good overlap with both. The optimal solution is to use the free energy itself.

In this way one could do random sampling in optimally selected regions of the phase space. For example within a range of free energies high enough to just overcome a free energy barrier.

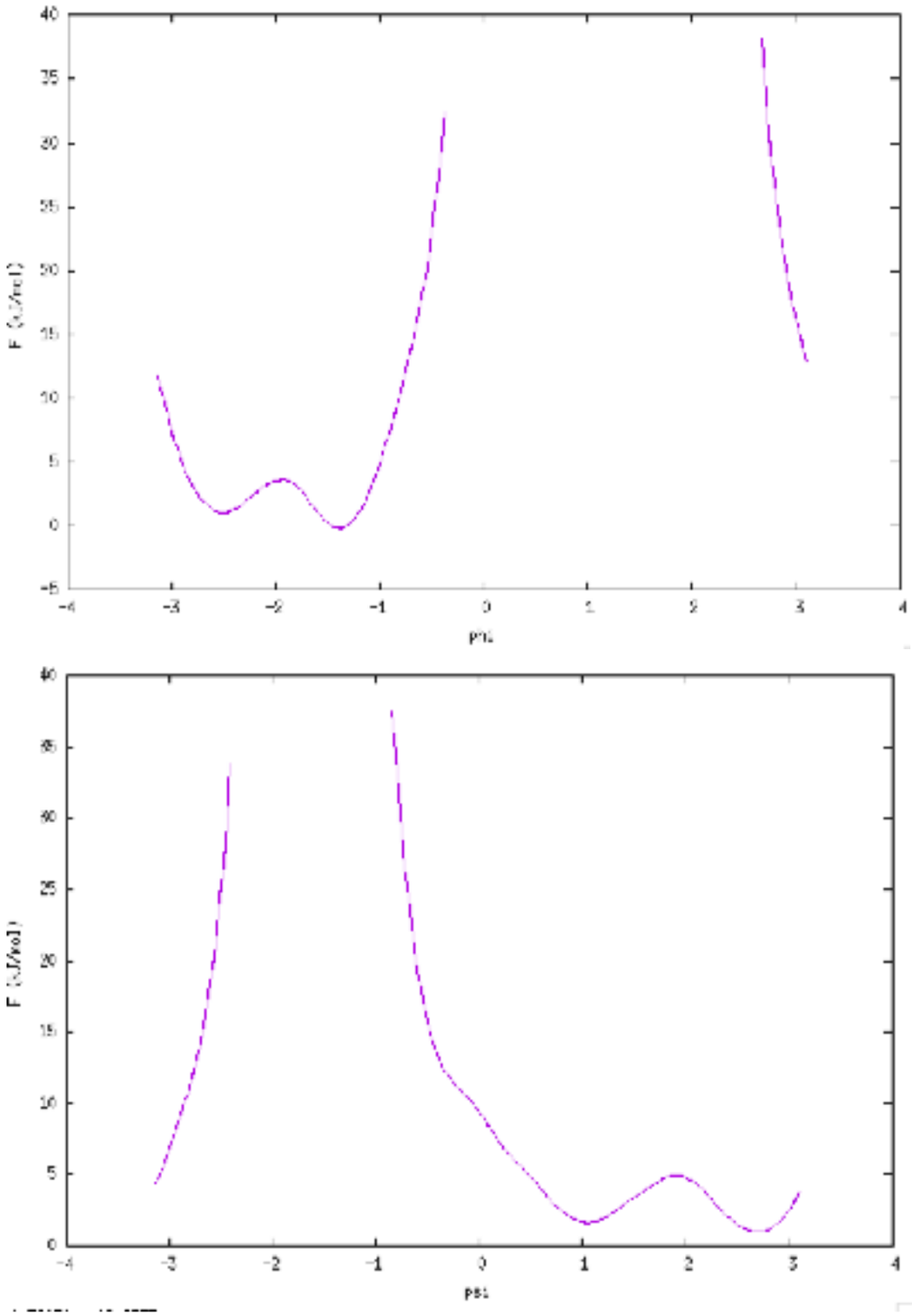
Alanine Dipeptide

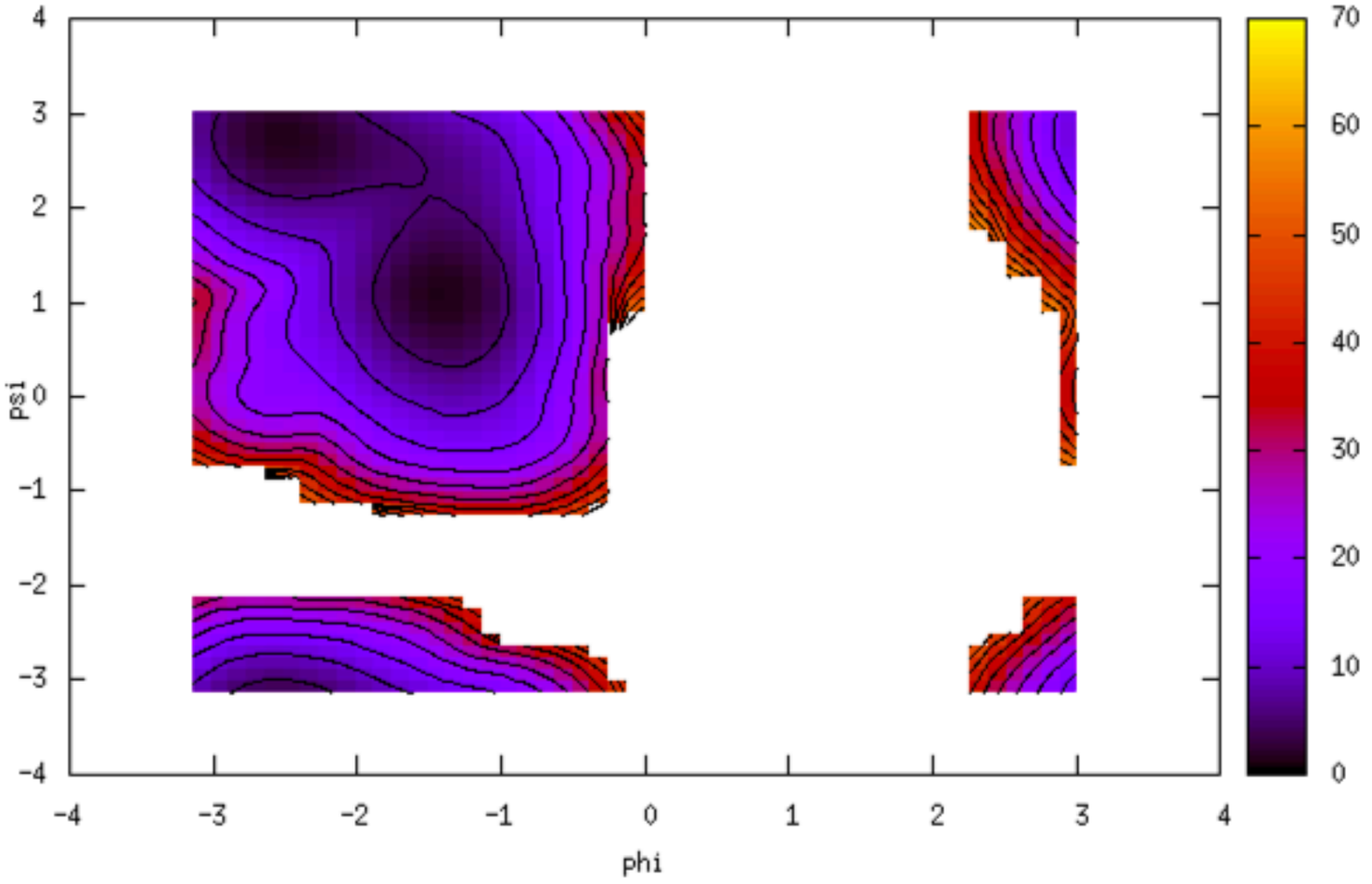





```
# vim:ft=plumed
MOLINFO STRUCTURE=../aladip.pdb

phi: TORSION ATOMS=@phi-2
psi: TORSION ATOMS=@psi-2
hhphi: HISTOGRAM ARG=phi STRIDE=10 GRID_MIN=-pi GRID_MAX=pi GRID_BIN=200 BANDWIDTH=0.1
hhpsi: HISTOGRAM ARG=psi STRIDE=10 GRID_MIN=-pi GRID_MAX=pi GRID_BIN=200 BANDWIDTH=0.1
hh: HISTOGRAM ARG=phi,psi STRIDE=10 GRID_MIN=-pi,-pi GRID_MAX=pi,pi GRID_BIN=50,50 BANDWIDTH=0.2,0.2
ffphi: CONVERT_TO_FES GRID=hhphi TEMP=298
ffpsi: CONVERT_TO_FES GRID=hhpsi TEMP=298
ff: CONVERT_TO_FES GRID=hh TEMP=298
DUMPGRID GRID=ffphi FILE=ffphi.dat
DUMPGRID GRID=ffpsi FILE=ffpsi.dat
DUMPGRID GRID=ff FILE=ff2d.dat
PRINT ARG=phi,psi FILE=colvar.dat STRIDE=10
```

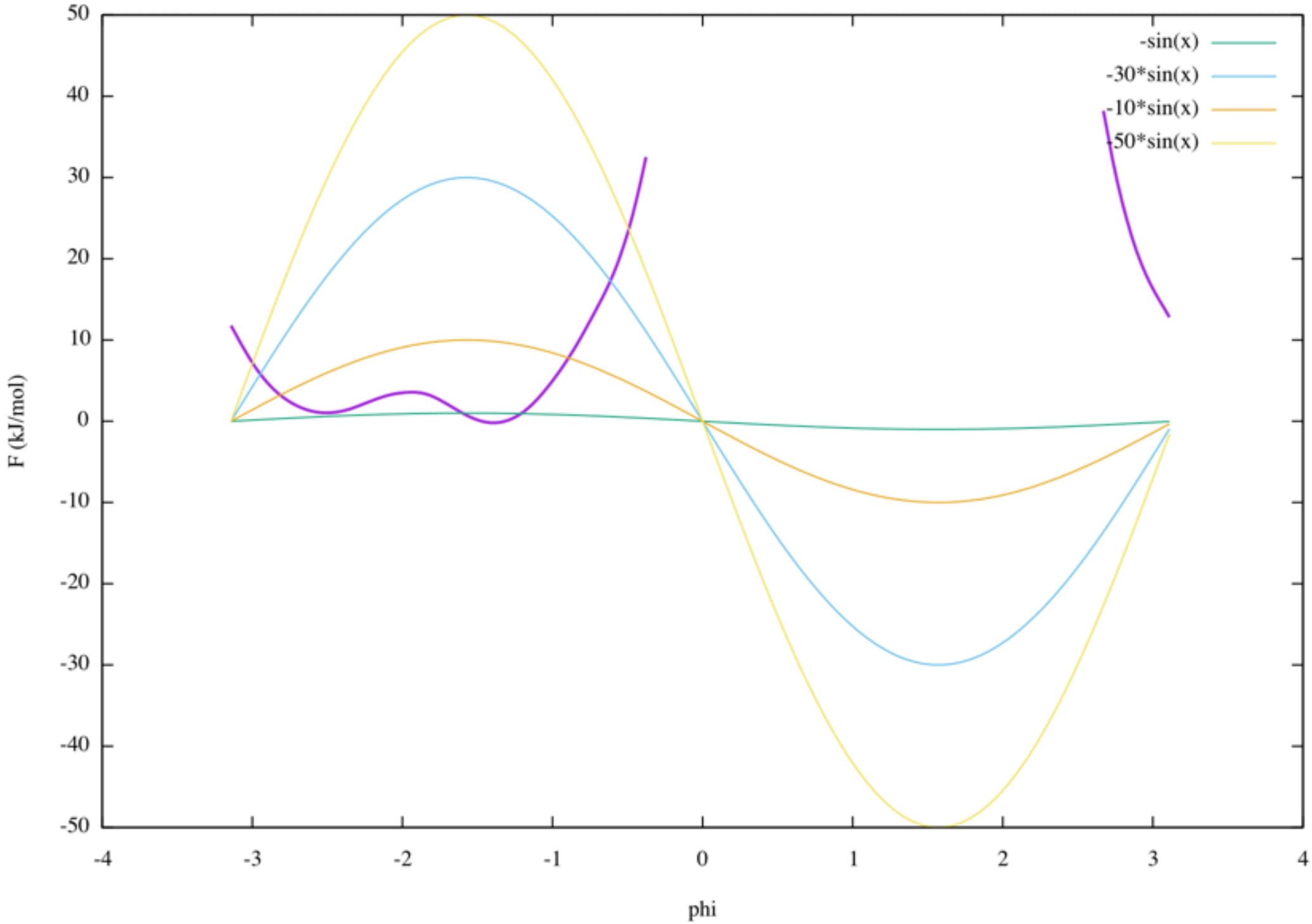
Alanine Dipeptide





Since these are periodic collective variables we could use a new ‘importance sampling’ defined as -sin(phi)

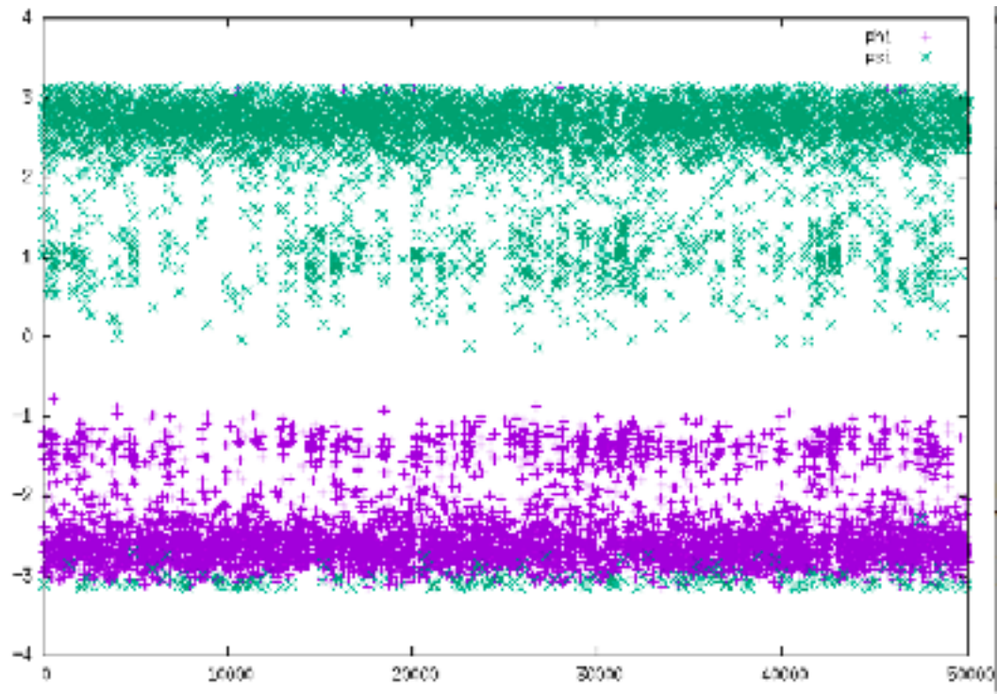
Alanine Dipeptide



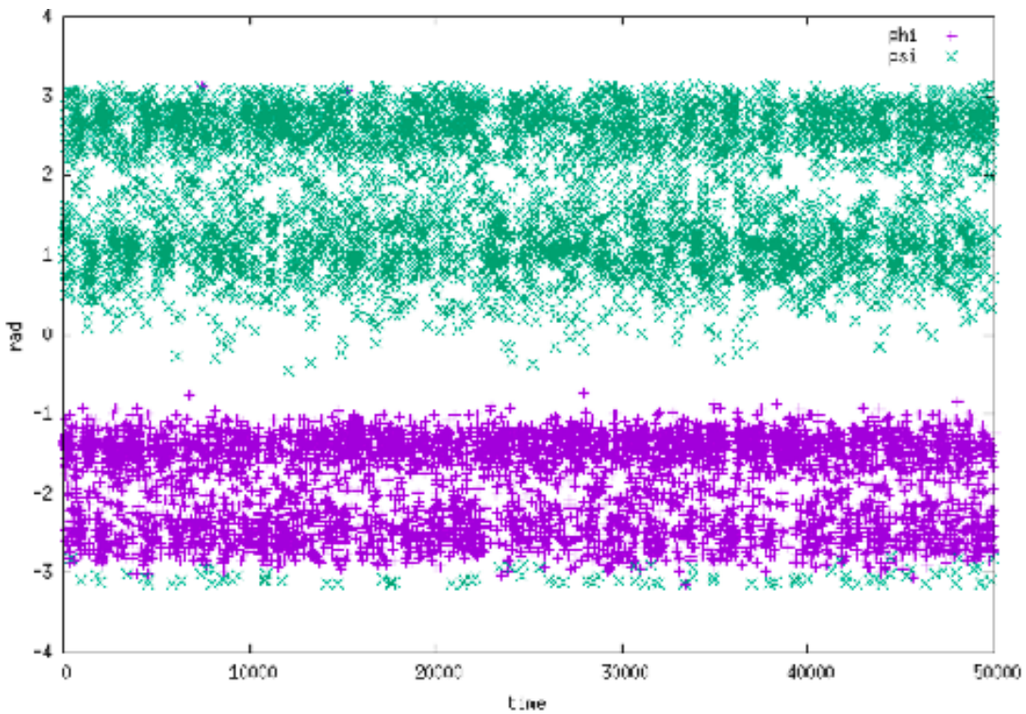
Adding a bias:

```
# vim:ft=plumed
MOLINFO STRUCTURE=aladip.pdb

phi: TORSION ATOMS=@phi-2

MATHEVAL ...
ARG=phi
LABEL=doubleg
FUNC=-10*sin(x)
PERIODIC=NO
... MATHEVAL

b: BIASVALUE ARG=doubleg

PRINT ARG=phi FILE=phi.dat STRIDE=10
```

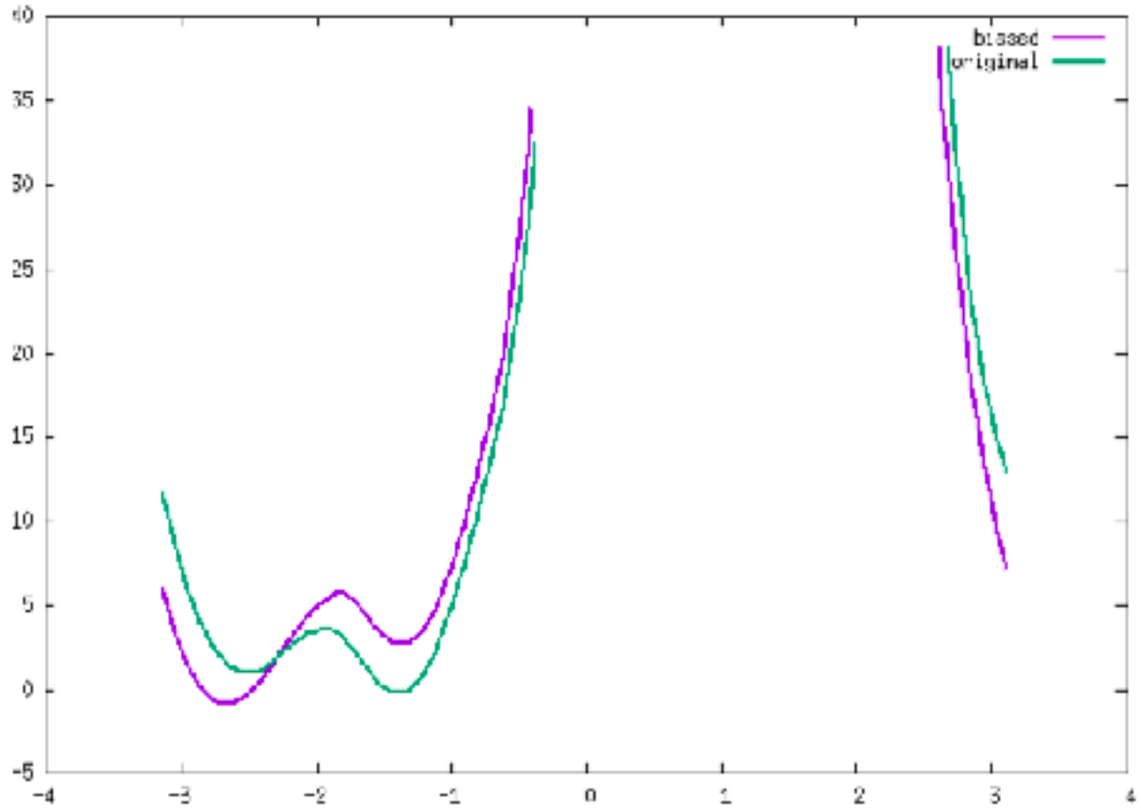
Biased

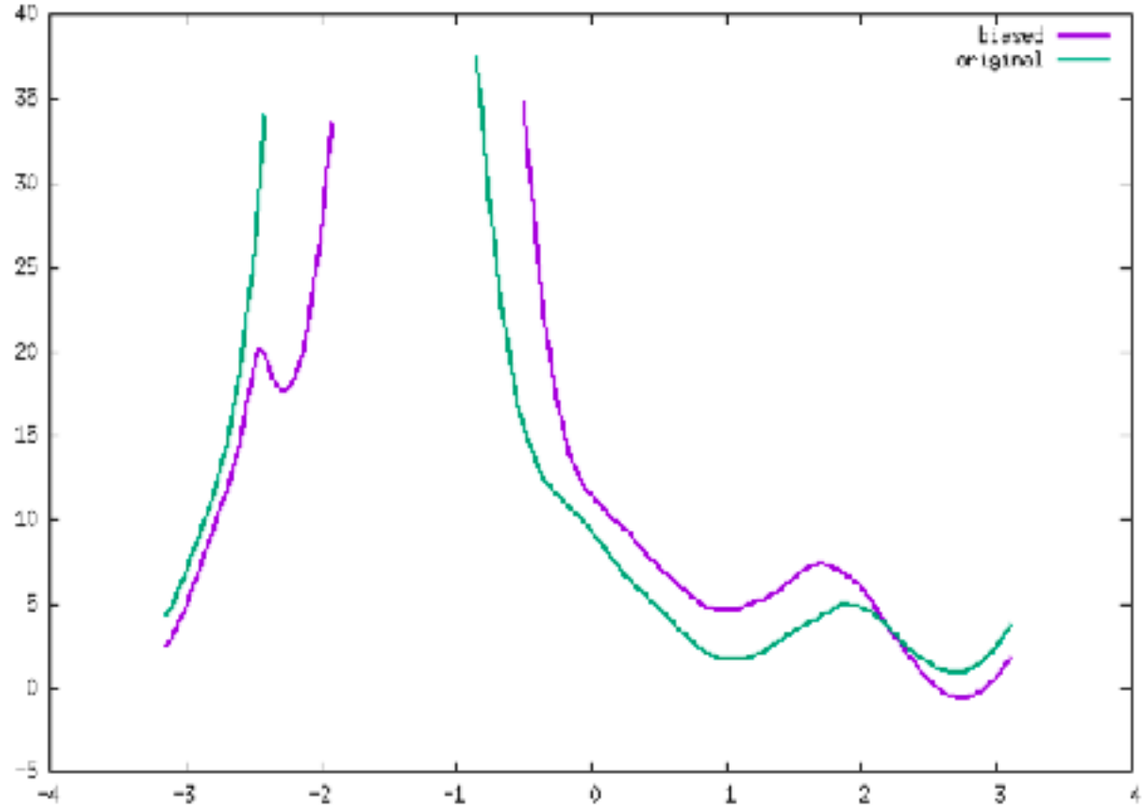


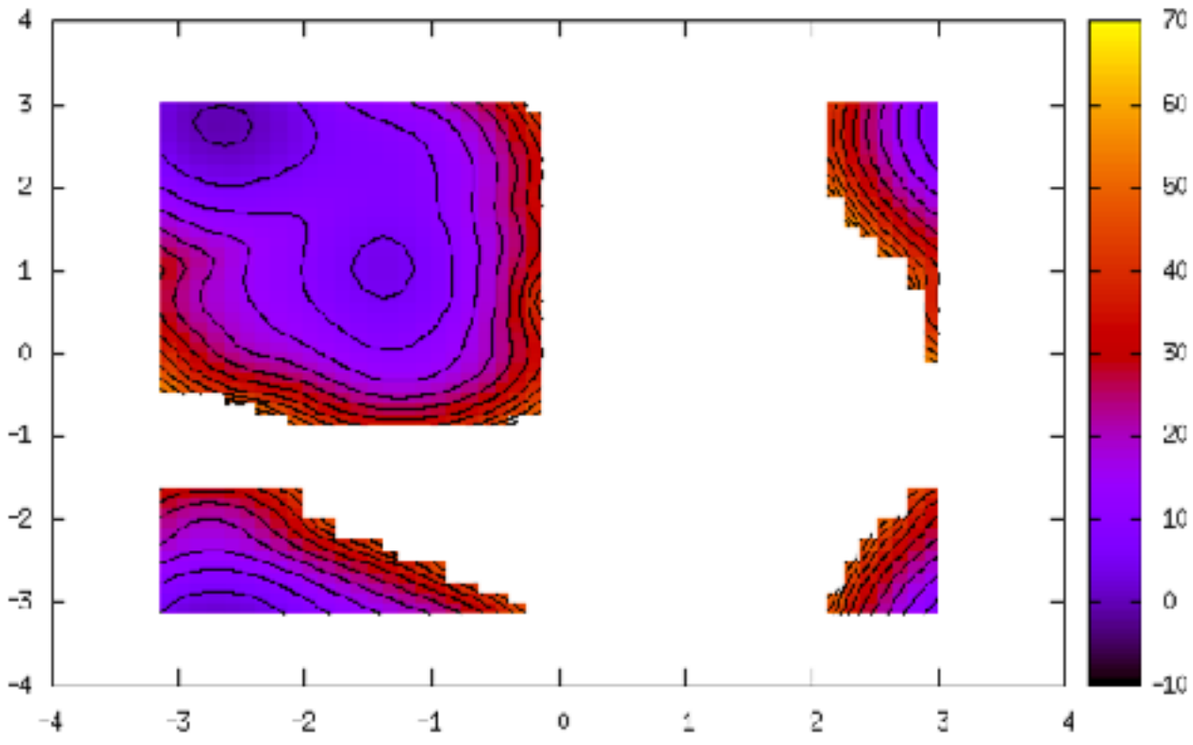
Original

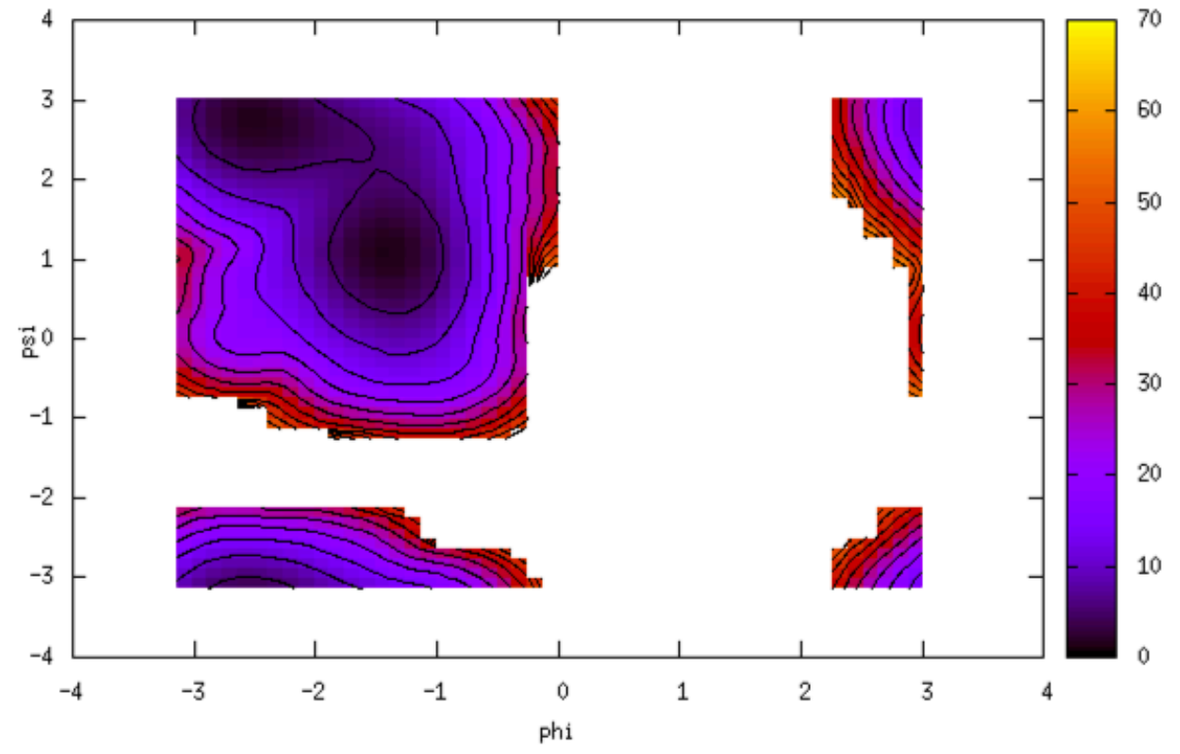


Adding a bias:





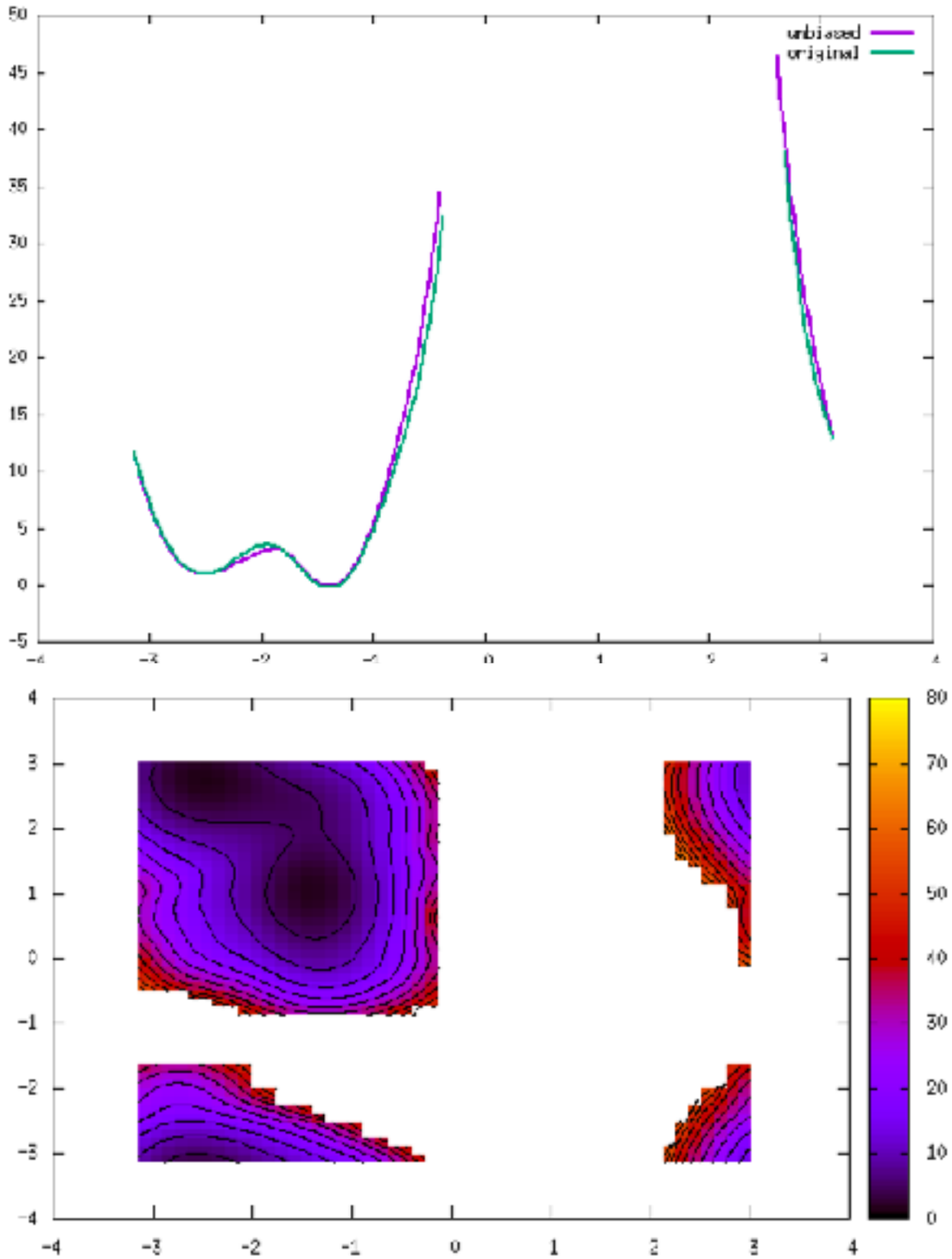


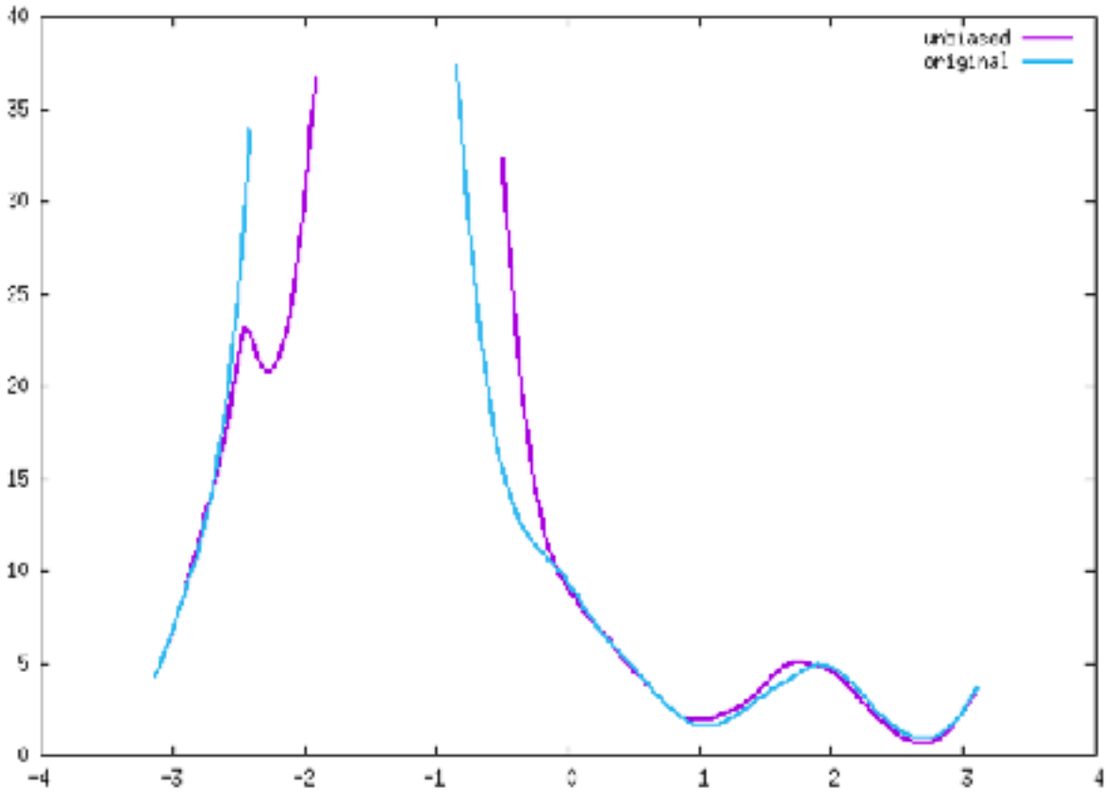


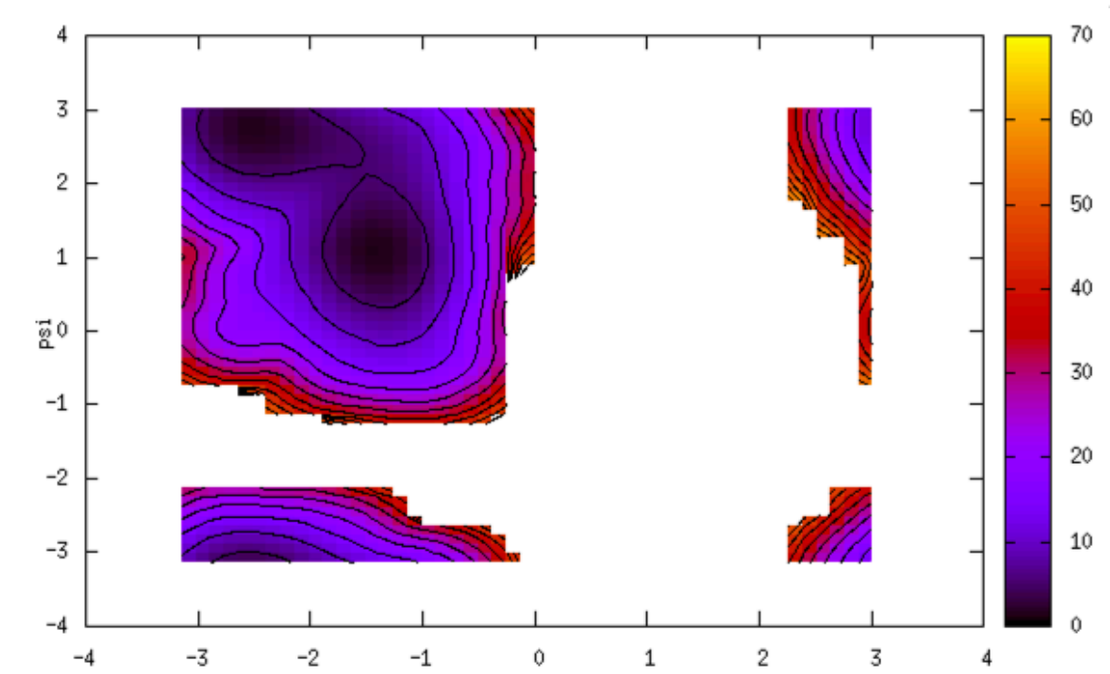
Removing a bias:

```
hh: HISTOGRAM ARG=phi,psi STRIDE=10 GRID_MIN=-pi,-pi GRID_MAX=pi,pi GRID_BIN=50,50 BANDWIDTH=0.2,0.2
ffphi: CONVERT_TO_FES GRID=hhphi TEMP=298
ffpsi: CONVERT_TO_FES GRID=hhpsi TEMP=298
ff: CONVERT_TO_FES GRID=hh TEMP=298
DUMPGRID GRID=ffphi FILE=ffphi.dat
DUMPGRID GRID=ffpsi FILE=ffpsi.dat
DUMPGRID GRID=ff FILE=ff2d.dat

MATHEVAL ...
ARG=phi
LABEL=doubleg
FUNC=-10*sin(x)
PERIODIC=NO
... MATHEVAL

b: BIASVALUE ARG=doubleg

as: REWEIGHT_BIAS TEMP=298

hhphiu: HISTOGRAM ARG=phi STRIDE=10 GRID_MIN=-pi GRID_MAX=pi GRID_BIN=200 BANDWIDTH=0.1 LOGWEIGHTS=as
hhpsiu: HISTOGRAM ARG=psi STRIDE=10 GRID_MIN=-pi GRID_MAX=pi GRID_BIN=200 BANDWIDTH=0.1 LOGWEIGHTS=as
hhu: HISTOGRAM ARG=phi,psi STRIDE=10 GRID_MIN=-pi,-pi GRID_MAX=pi,pi GRID_BIN=50,50 BANDWIDTH=0.2,0.2 LOGWEIGHTS=as
ffphiu: CONVERT_TO_FES GRID=hhphiu TEMP=298
ffpsiu: CONVERT_TO_FES GRID=hhpsiu TEMP=298
ffu: CONVERT_TO_FES GRID=hhu TEMP=298
DUMPGRID GRID=ffphiu FILE=ffphiu.dat
DUMPGRID GRID=ffpsiu FILE=ffpsiu.dat
DUMPGRID GRID=ffu FILE=ff2du.dat

PRINT ARG=phi,psi,b.bias FILE=colvar.dat STRIDE=10
```

Removing a bias:

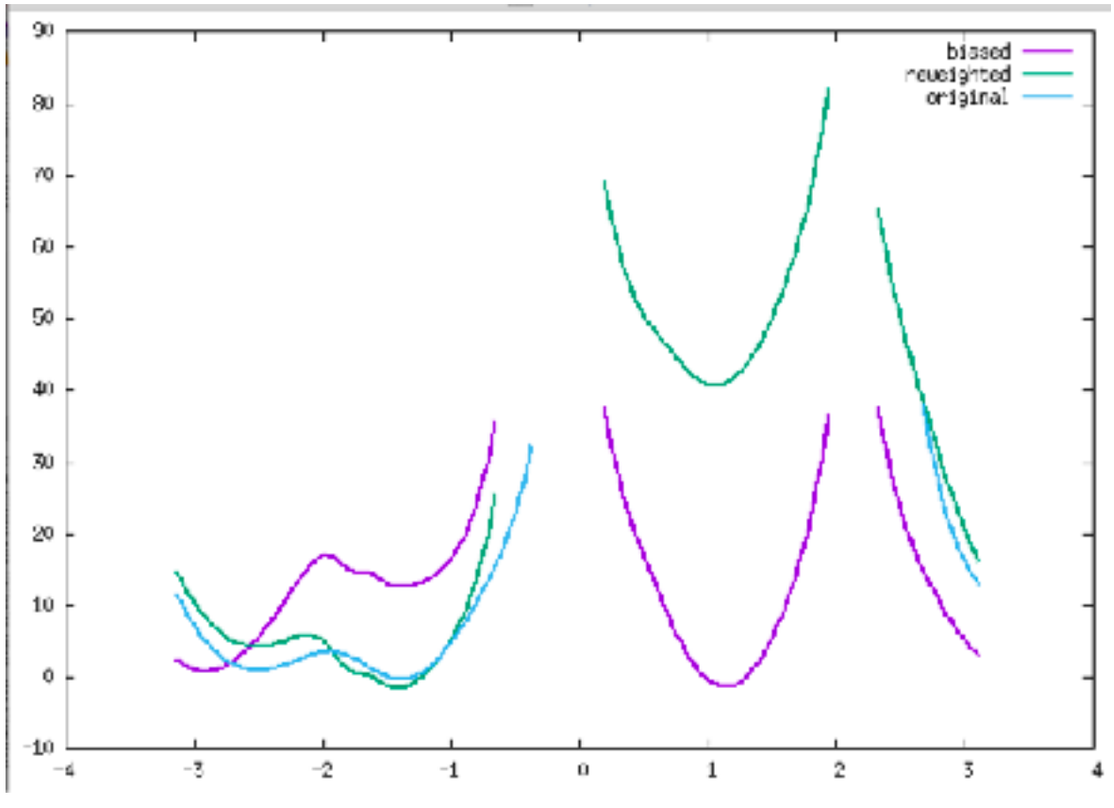




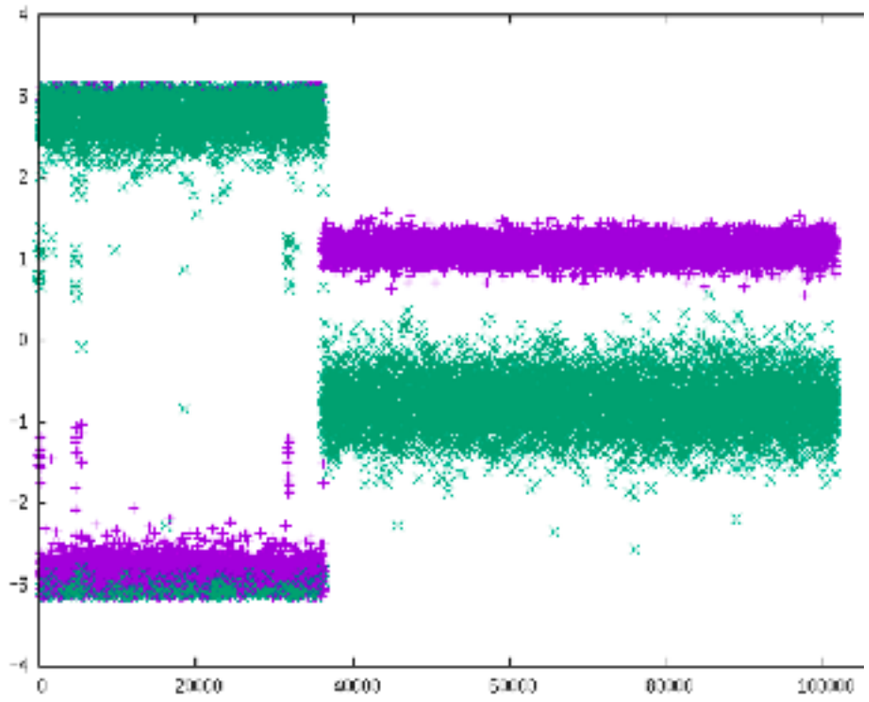


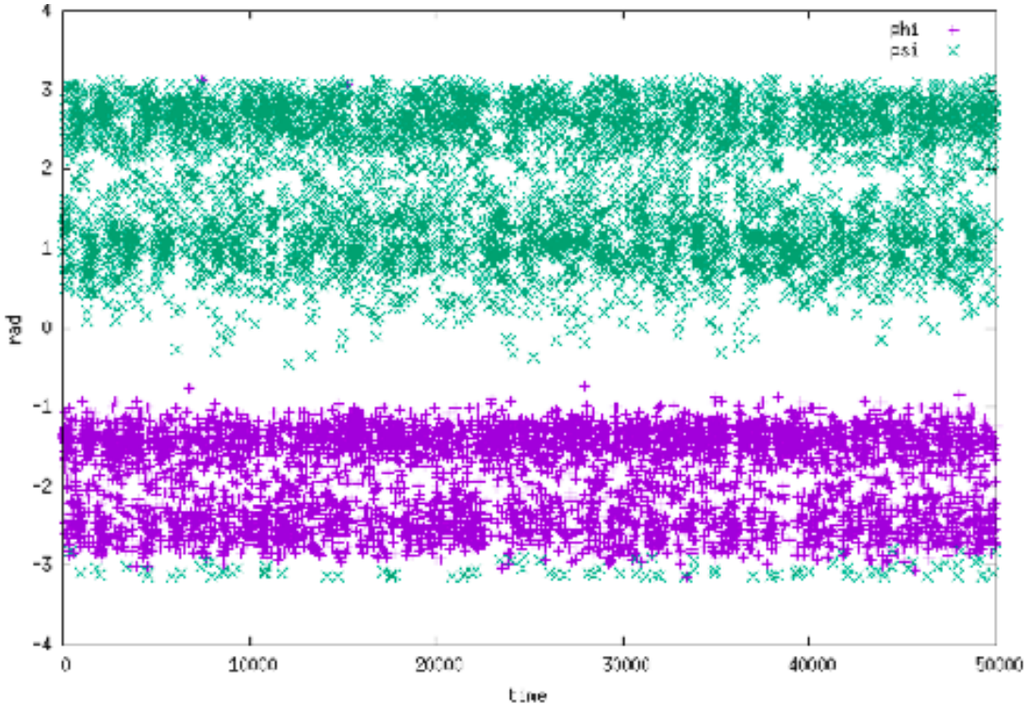
Adding a bias:

if we now use a stronger bias we can leave the local minimum but even doubling the sampling time we are just trapped in a new minimum. The two distributions don't overlap.



Original

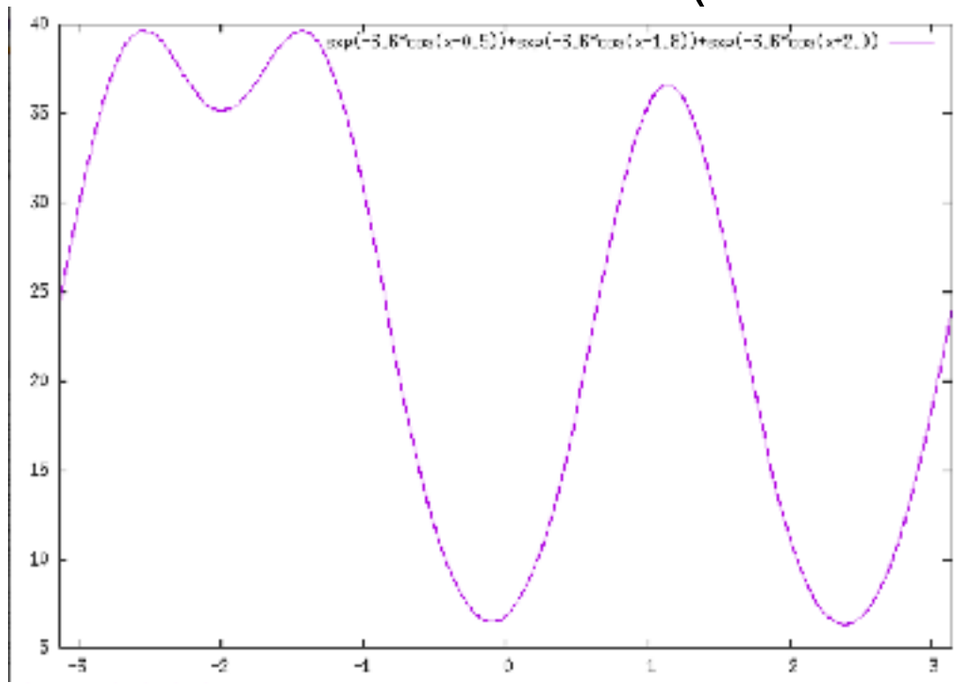




Adding a bias:

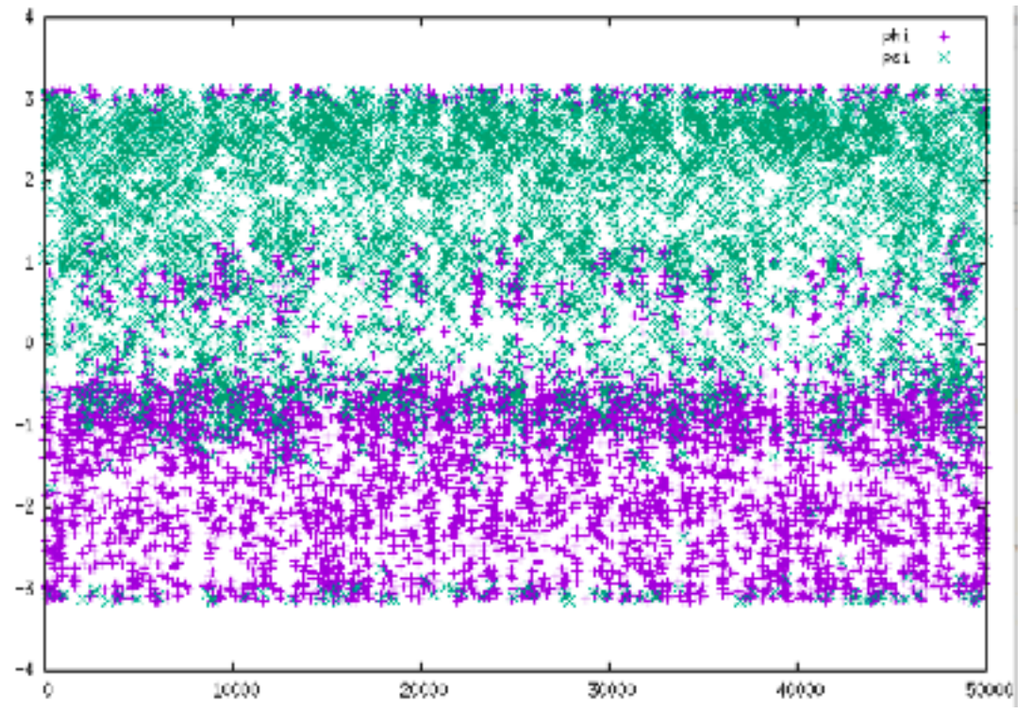
So the main issue of using ‘umbrella sampling’ is how to build a biasing potential that can bridge the real unknown probability distribution and a target distribution easy to sample.

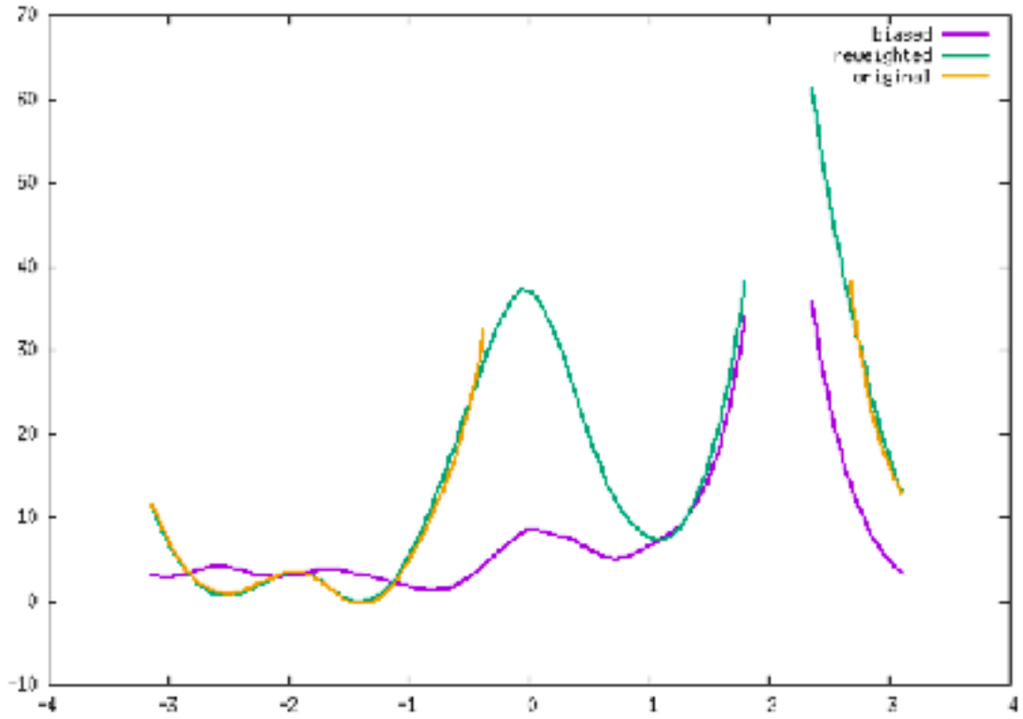
In the case of alanine dipeptide we can now add an additional bias in the new minimum and hope to be able to sample almost uniformly everywhere. Since we have three basins we can use three von Mises functions (circular Gaussians) $e^{k\cos(x-\mu)}$

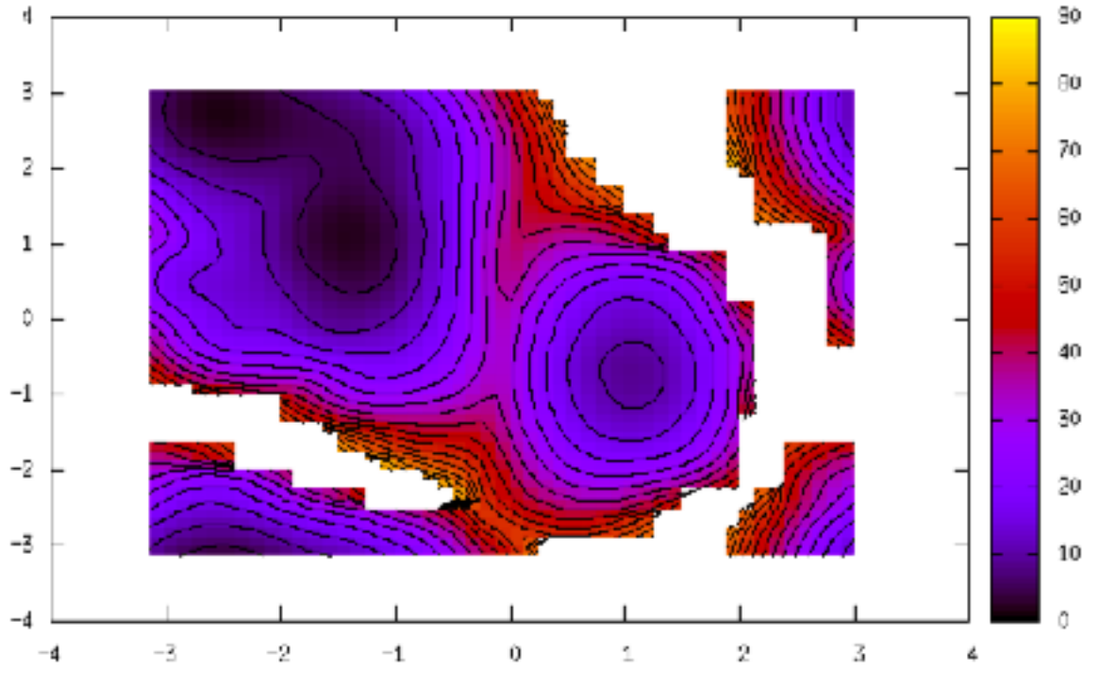


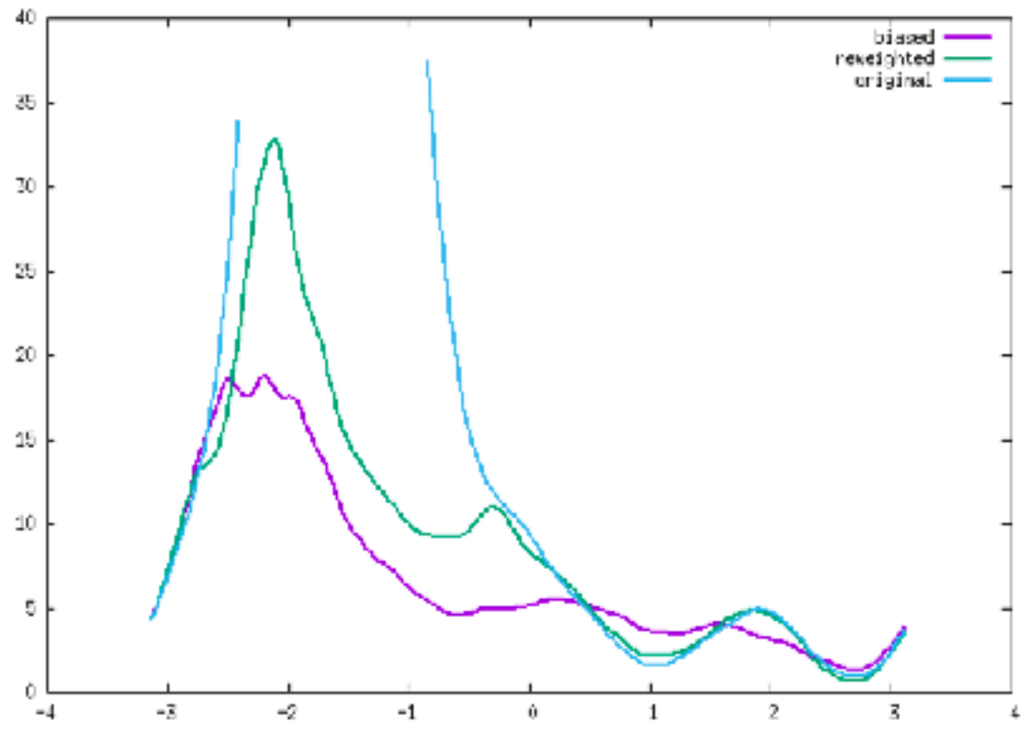

```
# vim:ft=plumed

MOLINFO STRUCTURE=aladip.pdb

phi: TORSION ATOMS=@phi-2

MATHEVAL ...
ARG=phi
LABEL=doubleg
FUNC=exp(-3.6*cos(x-0.5))+exp(-3.6*cos(x-1.8))+exp(-3.6*cos(x+2.))
PERIODIC=NO
... MATHEVAL

b: BIASVALUE ARG=doubleg

PRINT ARG=phi FILE=phi.dat STRIDE=10
```

Adding a bias:





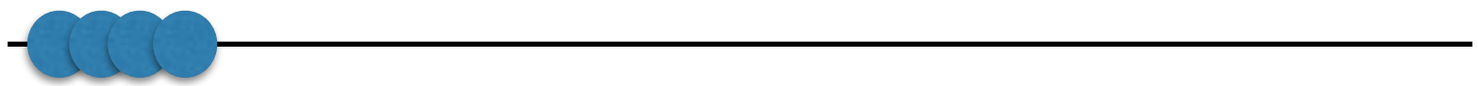




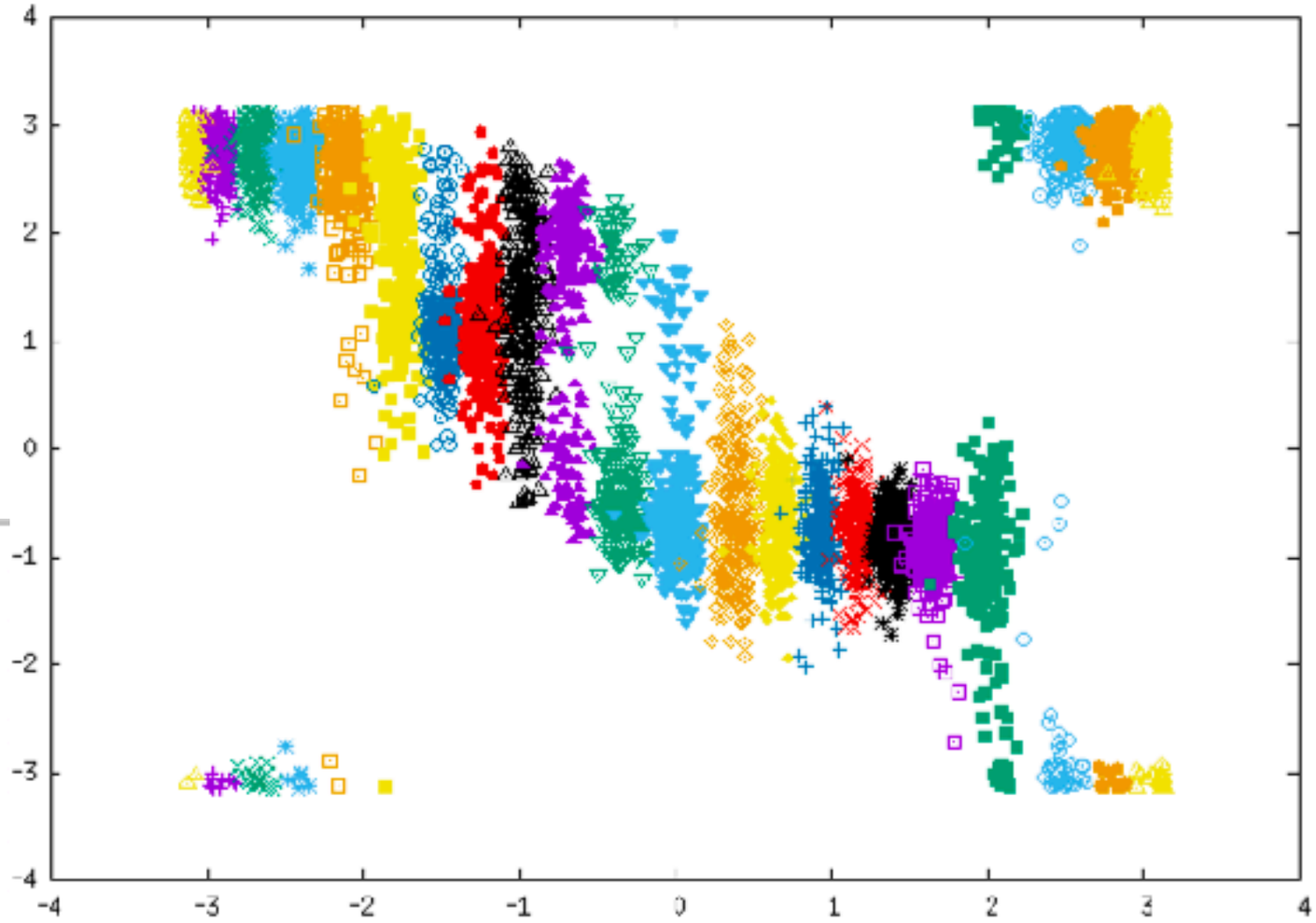
Umbrella Sampling: windows implementation

The WHAM solution to the problem of umbrella sampling consists to force the sampling to be uniform along one or more collective variables, with the hypothesis that the sampling in all other directions is converged.

By adding a sufficiently strong quadratic potential it is possible to restrain the sampling in the surrounding of a point in the CV space. Neighbour window should be close enough that there is overlap in the sampling. (if the value of CV is constrained like with shake this is the blue moon ensemble)

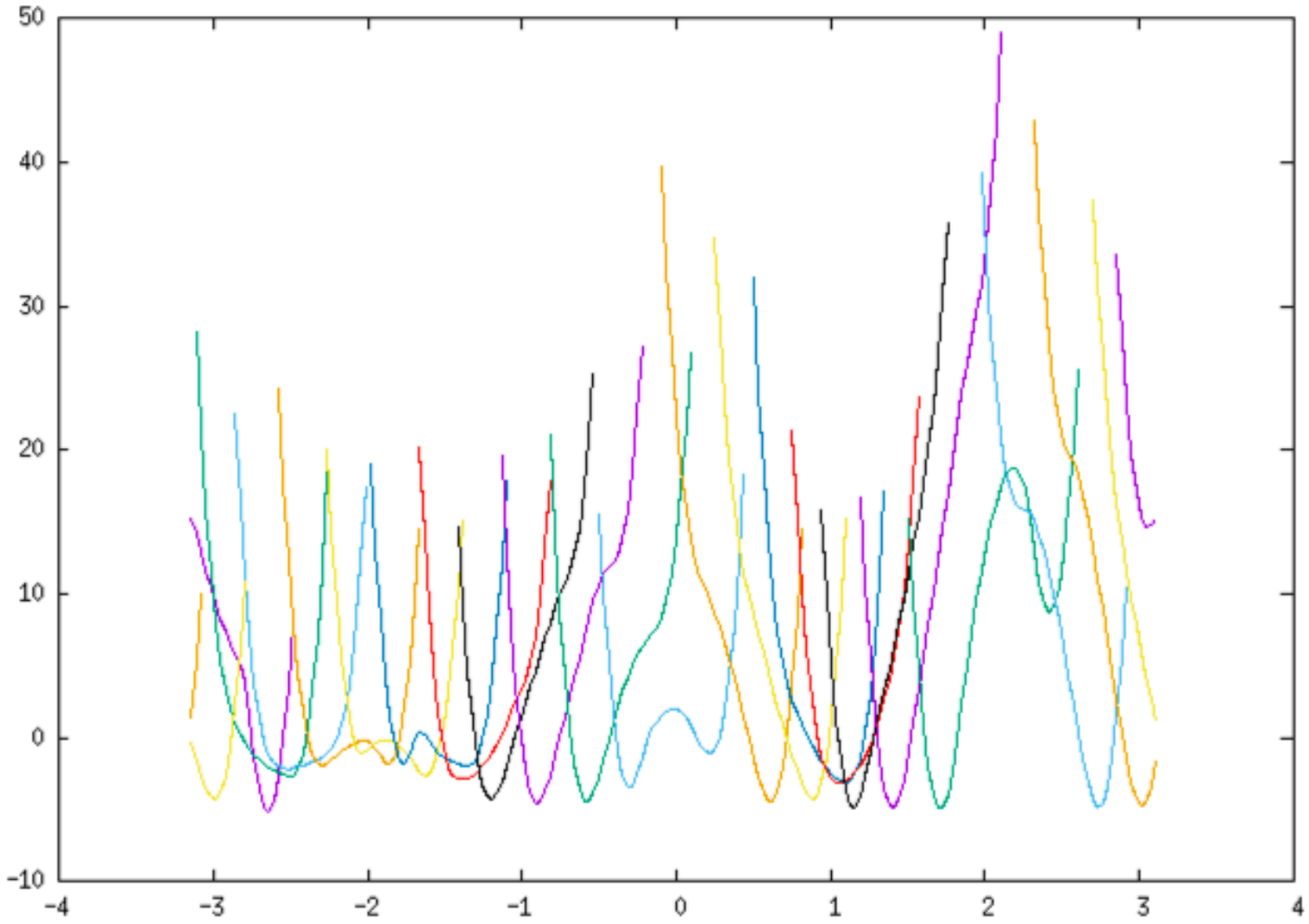


Umbrella Sampling: windows implementation




```
AT=-3
for i in `jot - 0 20`; do

cat >plumed.$i.dat << EOF

MOLINFO STRUCTURE=aladip.pdb

phi: TORSION ATOMS=@phi-2
psi: TORSION ATOMS=@psi-2

restraint-phi: RESTRAINT ARG=phi KAPPA=500.0 AT=$AT

PRINT ARG=phi,psi FILE=colvar.dat STRIDE=10

EOF

gmx_mpi mdrun -plumed plumed.$i.dat -nsteps 250000 -x traj$i.xtc -nb cpu -v -s short.tpr -cpi state -noappend

AT=`echo 'scale=1; '$AT'+0.3' | bc`;
done
```

Merging windows



Tomorrow we will see how to merge the statistics using WHAM

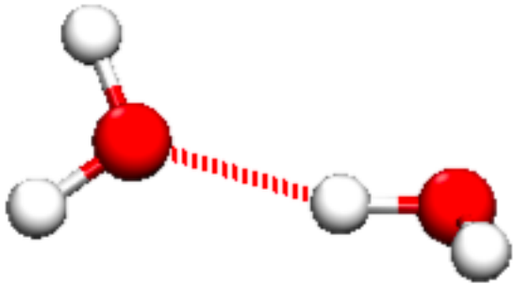
Umbrella sampling

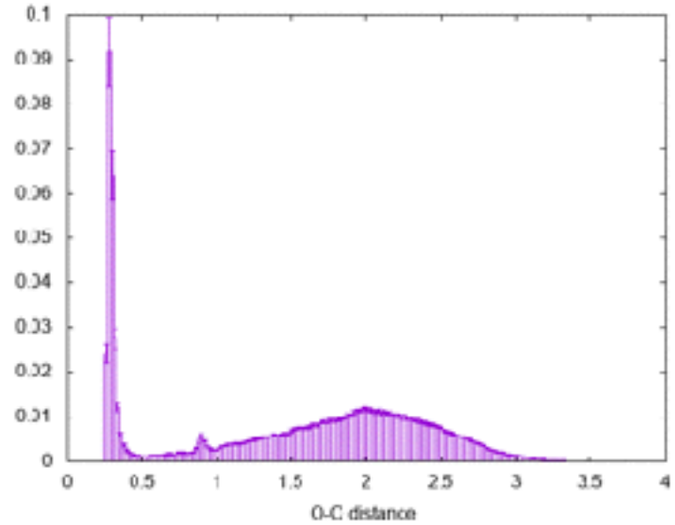
- The sampling of each window should be tested for
  - de-correlation
  - convergence

- Umbrella sampling can be coupled to other methods to enhance the sampling:
  - Replica-exchange (exchange between neighbour windows, this is essentially free)
  - Parallel Tempering (Run each window at many temperature)

This Afternoon:

- Adding and accounting for constant biases:





We will add first a weak linear or quadratic bias and observe that we obtain a very similar result. If we use a too strong bias we will not be able to sample anymore some region and so the reweigh will not work anymore (i.e. with WALLS). Eventually we will use a good estimate as an EXTERNAL potential on a GRID.

This Afternoon:

Then we will repeat the same exercises on Alanine Dipeptide using analytical potential defined using MATHEVAL.