DE LA RECHERCHE À L'INDUSTRIE

Accelerating the computation of finite temperature properties with Machine-Learning Assisted Canonical Sampling

Aloïs CASTELLANO **François BOTTIN** **Johann BOUCHET**

CEA, DAM, DIF, F-91297 Arpajon, France

1.1 Introduction : Canonical Average

In the classical limit, at a temperature T, the average of an observable O can be written as

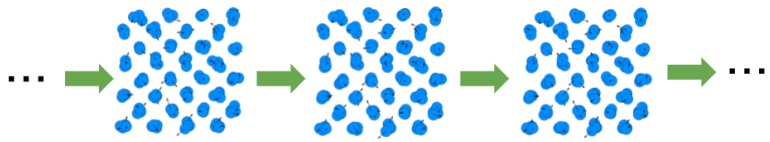
$$\langle O \rangle = \frac{1}{Z} \int \mathrm{d}\mathbf{R} O(\mathbf{R}) e^{-\beta V(\mathbf{R})}$$

In practice, $\mathbf{R}$ is high dimensional, so we use a finite number of sample to approximate $\langle O \rangle$

$$\langle O \rangle \approx \sum_n w_n O(\mathbf{R}_n) \quad \text{with} \quad \sum_n w_n = 1$$

1.1 Introduction : *Ab Initio* Molecular Dynamics

Integrate Newton's equations of motion



With a thermostat, it is possible to generate configurations according to the canonical ensemble at a temperature T

$$\langle O \rangle = \sum_n \frac{1}{N_{\text{tot}}} O(\mathbf{R}_n) \qquad w_n = \frac{1}{N_{\text{tot}}}$$

Very powerful but high computational cost

Needs from 10^4 to 10^5 forces evaluations to compute some properties

$\Rightarrow$ can take up to months if forces are computed using DFT
We need methods to accelerate the generation of configurations

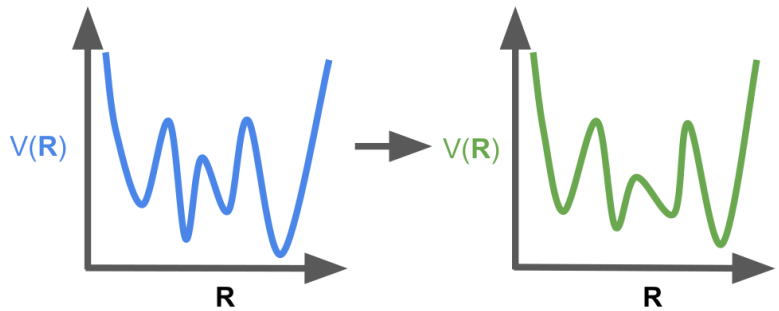
2.1 Machine Learning Interatomic Potential

Recently, progress in Machine-Learning Interatomic Potential (MLIP) allows a drastic acceleration with near-DFT accuracy

- **Less expensive than DFT**
- **Malleable : can describe almost any potential surface**
- **Data-driven : Needs a careful selection of the training dataset**

Generating the training dataset can need AIMD $\Rightarrow$ expensive

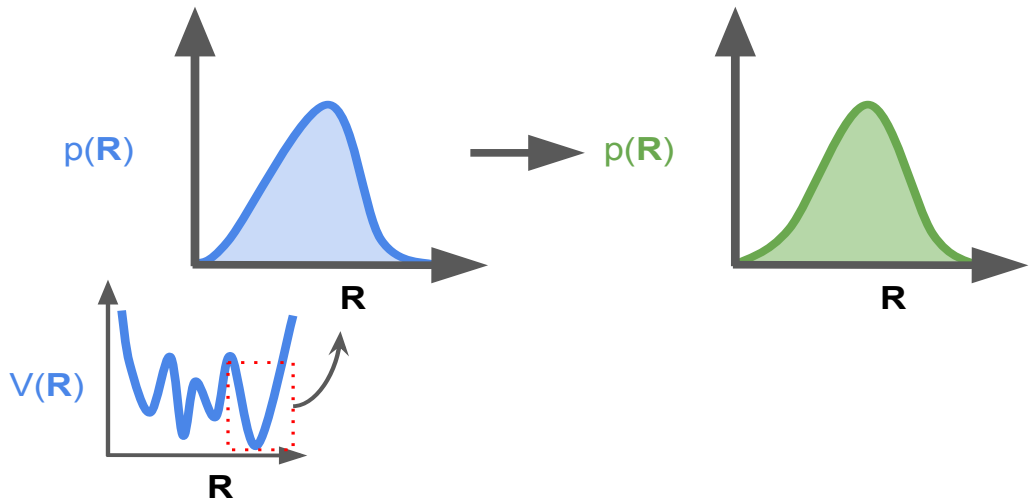
$$V^{\mathrm{DFT}}(\mathbf{R}) \Rightarrow V^{\mathrm{MLIP}}(\mathbf{R})$$



$$\langle O \rangle \approx \sum_n w_n^{\mathrm{DFT}} O^{\mathrm{DFT}}(\mathbf{R}) \Rightarrow \langle O \rangle \approx \sum_n w_n^{\mathrm{MLIP}} O^{\mathrm{MLIP}}(\mathbf{R})$$

We can use the MLIP potential as a distribution function

$$\frac{1}{Z}e^{-\beta V^{\mathrm{DFT}}(\mathbf{R})} \Rightarrow \frac{1}{Z}e^{-\beta V^{\mathrm{MLIP}}(\mathbf{R})}$$



$$\langle O \rangle \approx \sum_n w_n^{\mathrm{DFT}} O^{\mathrm{DFT}}(\mathbf{R}) \Rightarrow \langle O \rangle \approx \sum_n w_n^{\mathrm{MLIP}} O^{\mathrm{DFT}}(\mathbf{R}_n)$$

How to obtain the best weights w_n ?

3.2 Machine-Learning Assisted Canonical Sampling

$$p(\mathbf{R}) = \frac{1}{Z} e^{-\beta V(\mathbf{R})} \qquad q(\mathbf{R}) = \frac{1}{Z} e^{-\beta V(\mathbf{R})}$$

We can use the Kullback-Leibler divergence to define a measure of the similarity between two distributions

$$KL(p\|q) = \int \mathrm{d}\mathbf{R} q(\mathbf{R}) \ln \left[\frac{p(\mathbf{R})}{q(\mathbf{R})} \right] \geqslant 0$$

3.2 Machine-Learning Assisted Canonical Sampling

$$p(\mathbf{R}) = \frac{1}{Z} e^{-\beta V(\mathbf{R})} \qquad q(\mathbf{R}) = \frac{1}{Z} e^{-\beta V(\mathbf{R})}$$

We can use the Kullback-Leibler divergence to define a measure of the similarity between two distributions

$$KL(p\|q) = \int \mathrm{d}\mathbf{R} q(\mathbf{R}) \ln \left[\frac{p(\mathbf{R})}{q(\mathbf{R})} \right] \geqslant 0$$

With a few manipulations, this can be transformed to an equivalent free energy problem

Gibbs-Bogoliubov free energy $\widetilde{\mathcal{F}}$:

$$\widetilde{\mathcal{F}} = \mathcal{F}_0 + \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle \geqslant \mathcal{F}$$

3.3 Machine-Learning Assisted Canonical Sampling

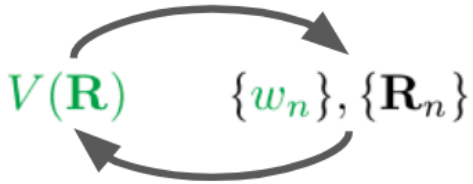
Gibbs-Bogoliubov free energy $\widetilde{\mathcal{F}}$:

$$\widetilde{\mathcal{F}} = \mathcal{F}_0 + \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle \geqslant \mathcal{F}$$

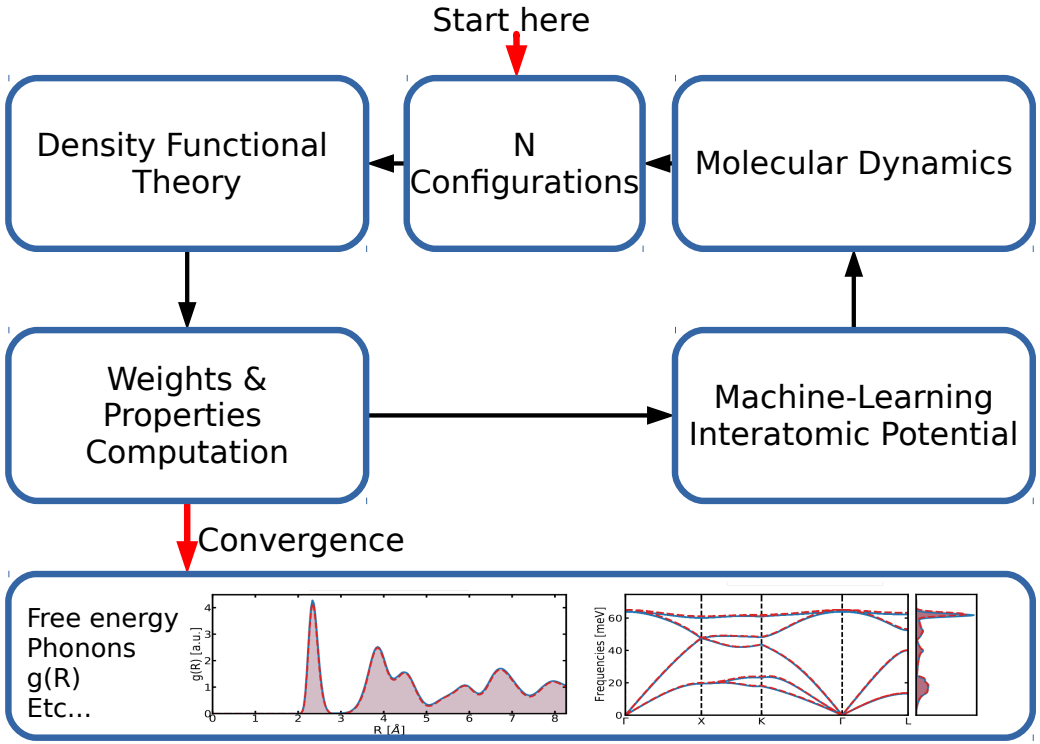
By minimizing this quantity with respect to the MLIP parameters γ, we assure the best approximation of the DFT distribution and free energy

$$\frac{\partial \widetilde{\mathcal{F}}}{\partial \boldsymbol{\gamma}} = 0$$

This can be solved in a self-consistent approach



3.4 Machine-Learning Assisted Canonical Sampling



Two sources of acceleration

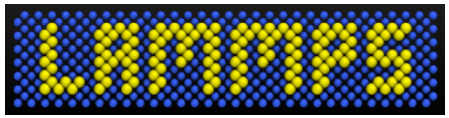
- **Reduced number of DFT calculations (100 to 200 in total)**
- **DFT calculations in parallel**

3.4 Machine-Learning Assisted Canonical Sampling

- The results are the configurations and the weights
- The properties are computed with DFT observables (energy, forces, electronic properties, . . .)
- The studied system is the DFT one, the approximation is made on the weights given to the configurations
- The method can be seen as a replacement of AIMD

4.1 Implementation

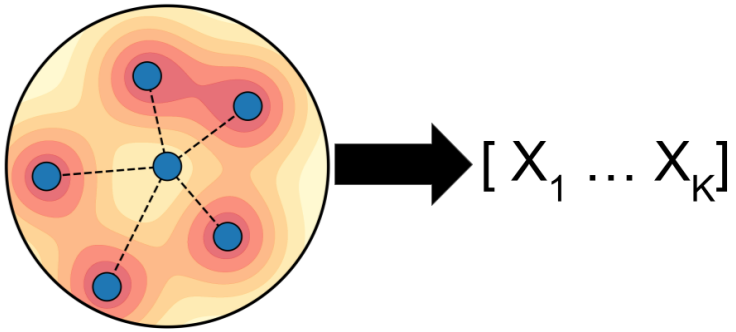
- **Atomic environment manipulated with** 

  manipulation of supercells, MD, etc . . .
- **MLIP computed using** 

- **Weights computed with pymbar**
- **DFT calculations with** 

4.2 SNAP potential from LAMMPS

The atomic environment is mapped to a descriptor space



A linear dependence is assumed between the descriptor space and the energy

$$V(\mathbf{R}) = \sum_I \underbrace{\sum_k \gamma_k X_k^I}_{V_{\mathrm{SNAP}}^I(\mathbf{R})} + V_{\mathrm{pair}}(\mathbf{R})$$

Fitting by weighted least squares

$$\boldsymbol{\gamma} = (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{W}\boldsymbol{Y}$$

- $\boldsymbol{X}$: **Descriptor (or derivatives)**
- $\boldsymbol{W}$: **weights**
- $\boldsymbol{Y}$: **DFT Energies, Forces and Stress**

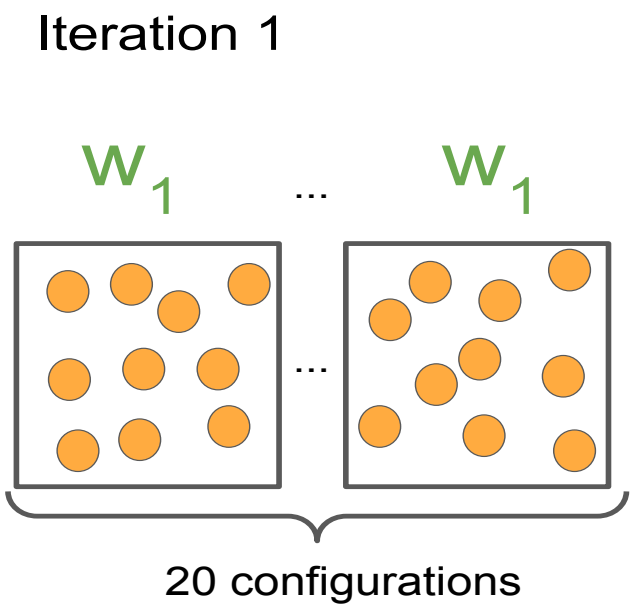
A. P. Thompson *et al* J. Comput. Phys. **285** 316-330 (2015)

4.3 Weights computation : MBAR

Multistate Bennett Acceptance Ratio (MBAR)

A free energy method that allows to compute the weights of all configurations generated during the simulation

We use the pymbar package



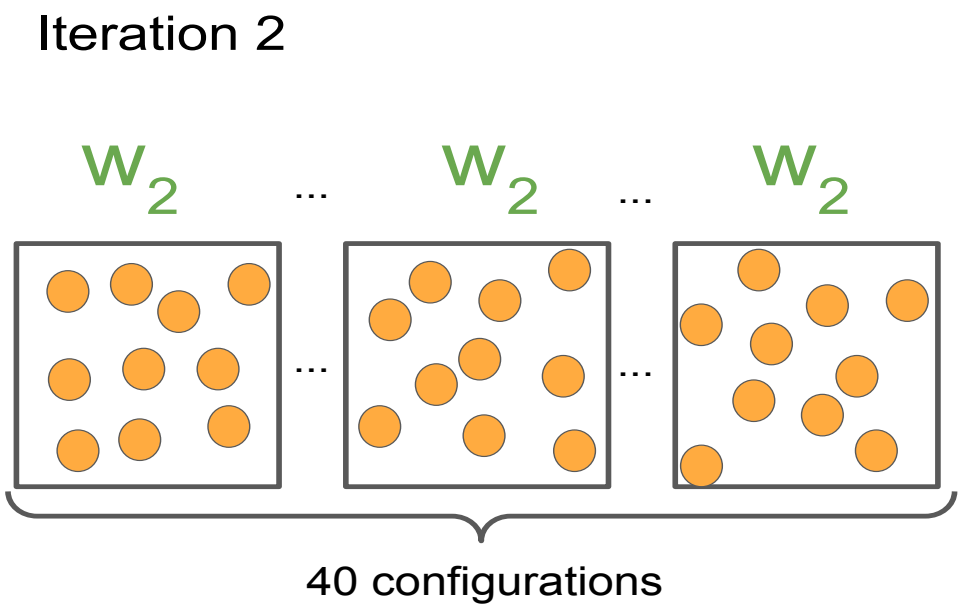
M. Shirts *et al*, J. Chem. Phys. **129** 124105 (2008)

4.3 Weights computation : MBAR

Multistate Bennett Acceptance Ratio (MBAR)

A free energy method that allows to compute the weights of all configurations generated during the simulation

We use the pymbar package



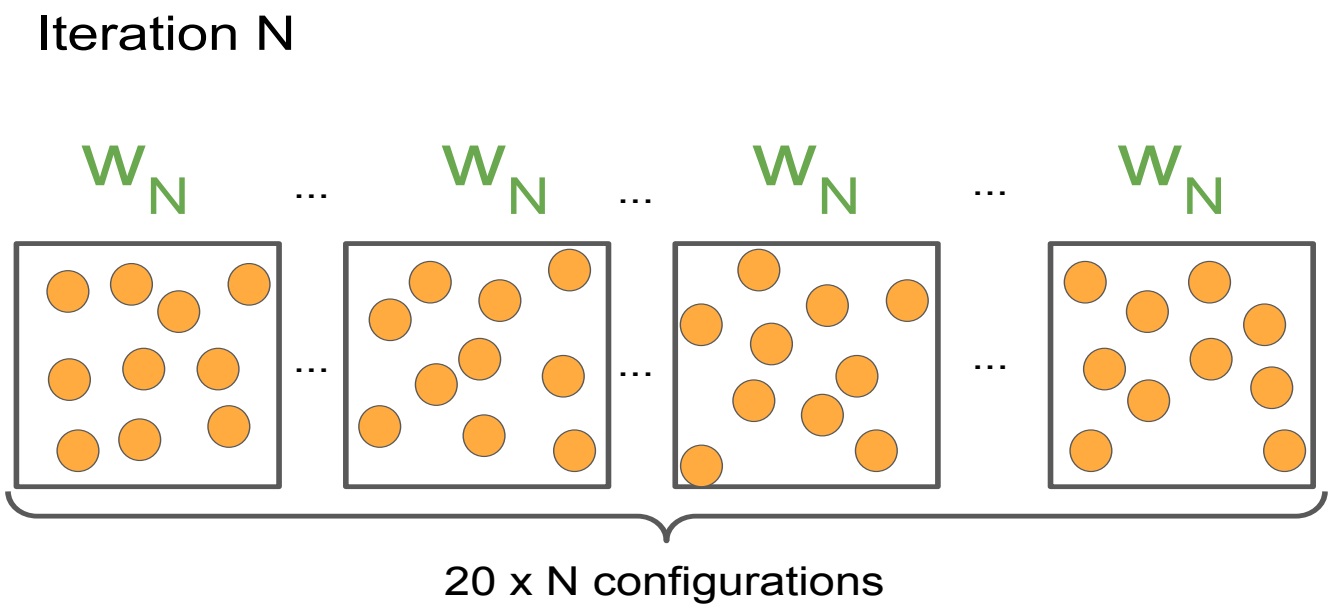
M. Shirts *et al*, J. Chem. Phys. **129** 124105 (2008)

4.3 Weights computation : MBAR

Multistate Bennett Acceptance Ratio (MBAR)

A free energy method that allows to compute the weights of all configurations generated during the simulation

We use the pymbar package



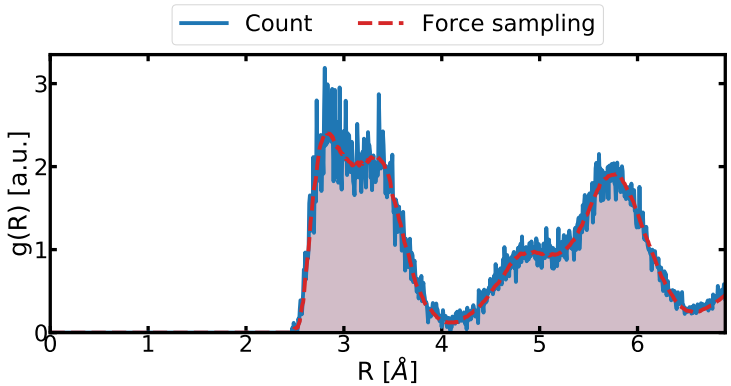
M. Shirts *et al*, J. Chem. Phys. **129** 124105 (2008)

4.4 Property computed

We can compute properties by writing them as canonical averages

PDF $g(R)$: force sampling

$$g(R) = \frac{\beta}{4\pi}\frac{V}{N^2}\left\langle \sum_i \sum_{j\neq i} \frac{1}{2}(\mathbf{F}_i - \mathbf{F}_j)\frac{\mathbf{R}_j - \mathbf{R}_i}{R_{ij}^3}\theta(R - R_{ij})\right\rangle$$



B. Rotenberg J. Chem. Phys **153** 150902 (2020)

Finite temperature phonons : TDEP

$$V_{\mathrm{TDEP}}(\mathbf{R}, T) = \frac{1}{2}\sum_{ij} \mathbf{u}_i \mathbf{\Theta}_{ij}(T)\mathbf{u}_j$$

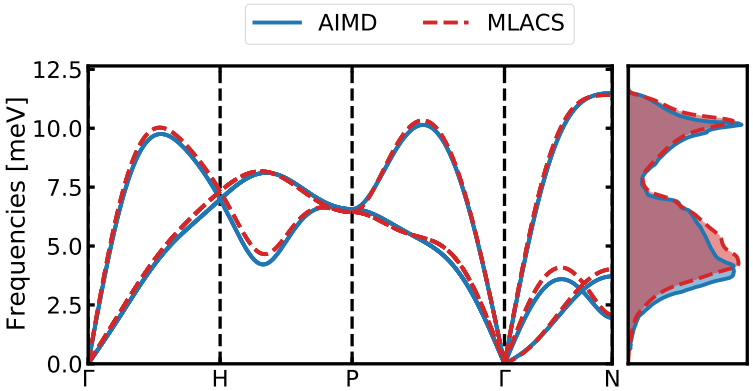
$$\mathbf{\Theta}_{ij}(T) = \frac{\sum_k \langle \mathbf{F}_i \mathbf{u}_k \rangle}{\sum_k \langle \mathbf{u}_j \mathbf{u}_k \rangle}$$

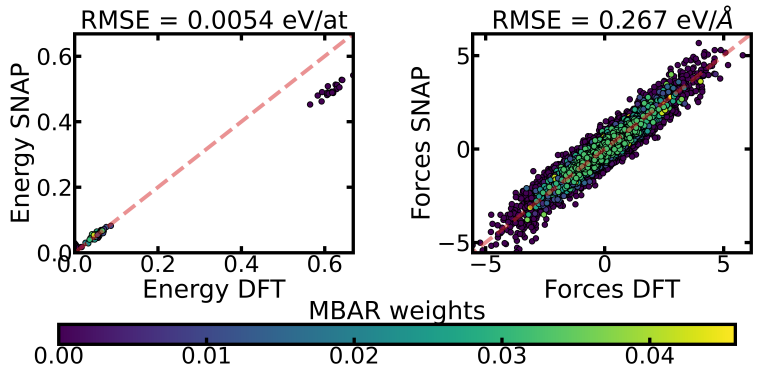
$\mathbf{\Theta}(T) \Rightarrow \omega(T)$ **renormalized phonons**

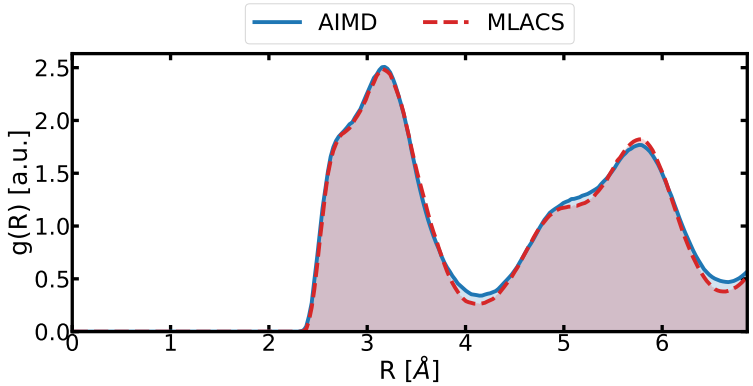
O. Hellman *et al*, PRB **84** 180301 (2011)

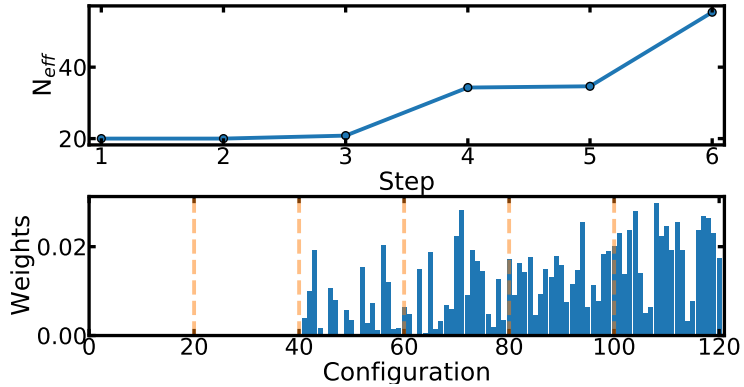
F. Bottin *et al*, CPC **254** 107301 (2020)

Uranium BCC 1200K - 128 atoms



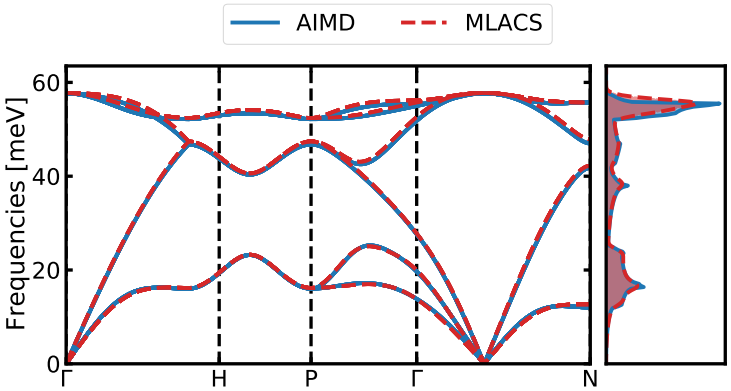


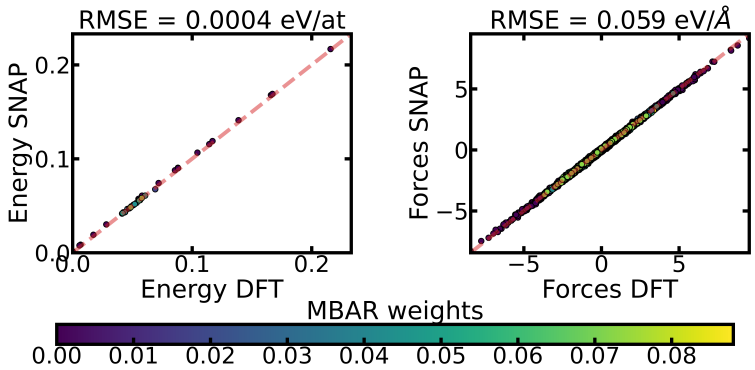


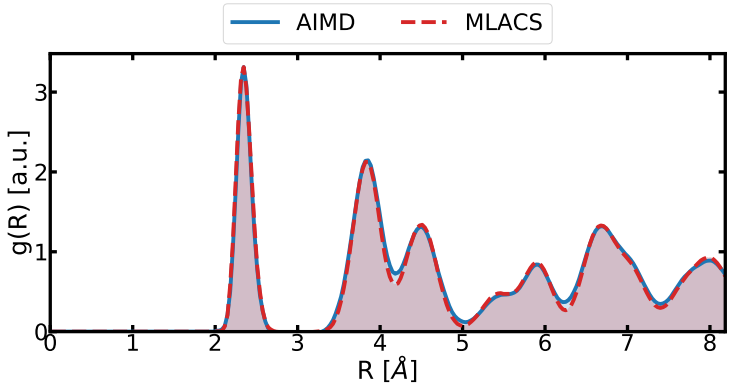


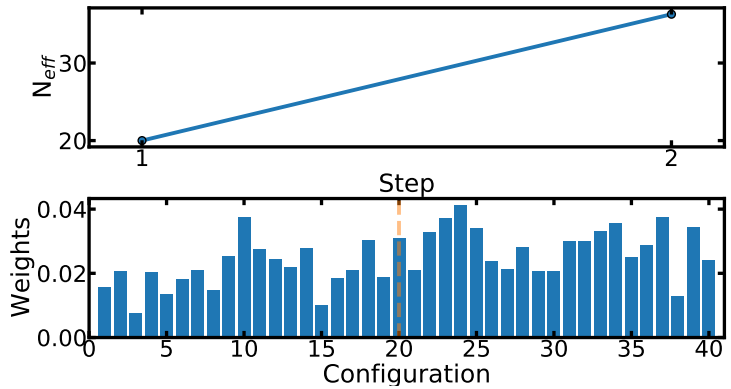
Number of configurations computed : AIMD 4000 - MLACS 140

Silicium 900K - 216 atoms



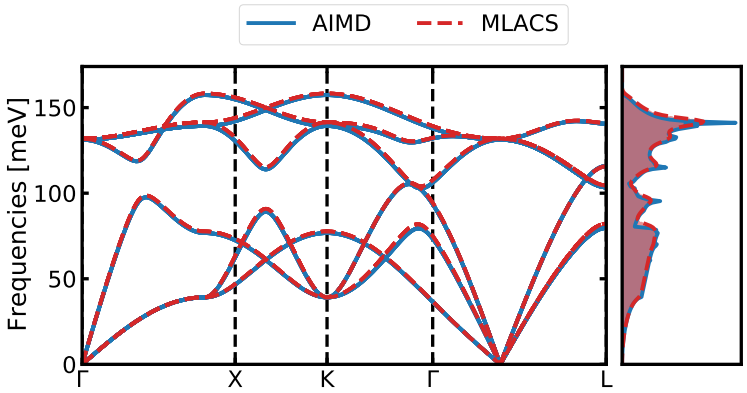


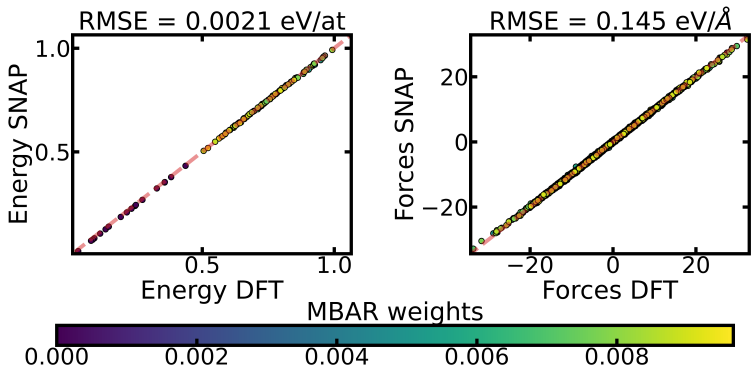


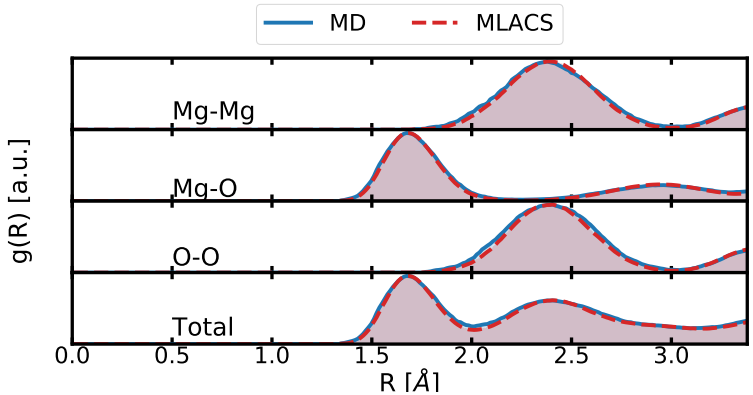


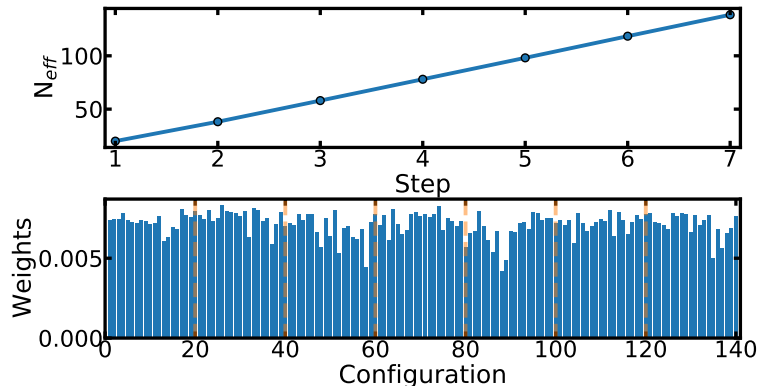
Number of configurations computed : AIMD 4000 - MLACS 60

MgO - 400GPa - 8000K - 64 atoms



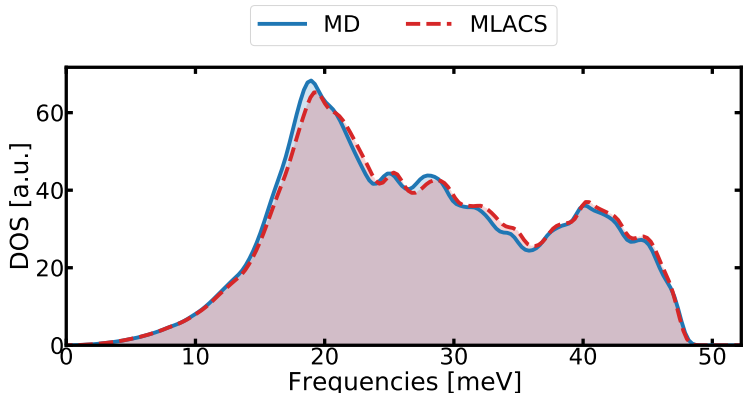


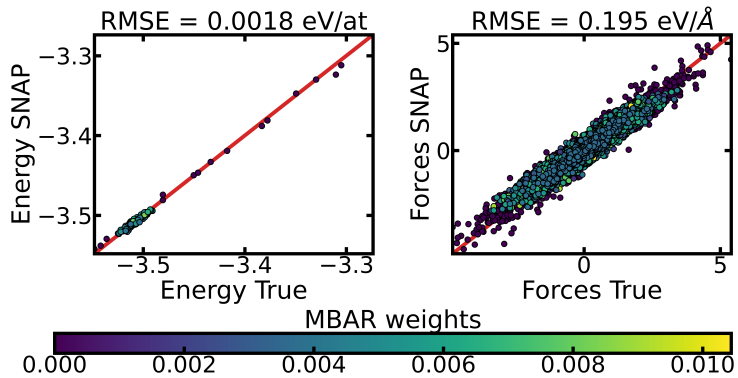


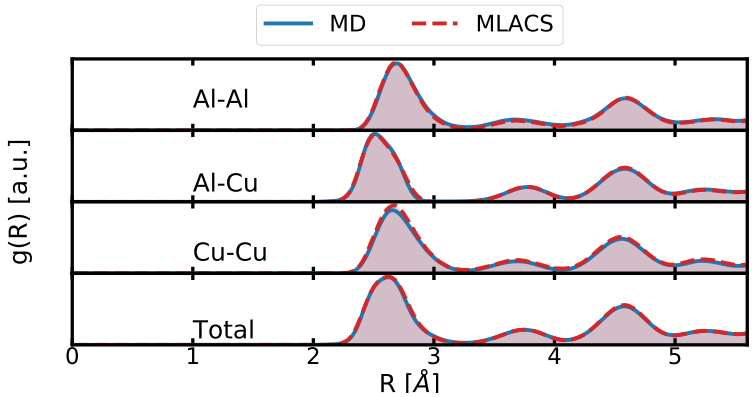


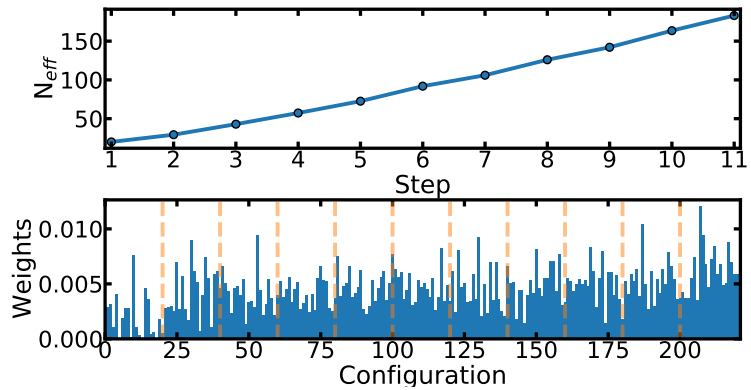
Number of configurations computed : AIMD 7000 - MLACS 160

$Al_{0.50}Cu_{0.50}$ Bond Order Potential - Solid Solution 600K



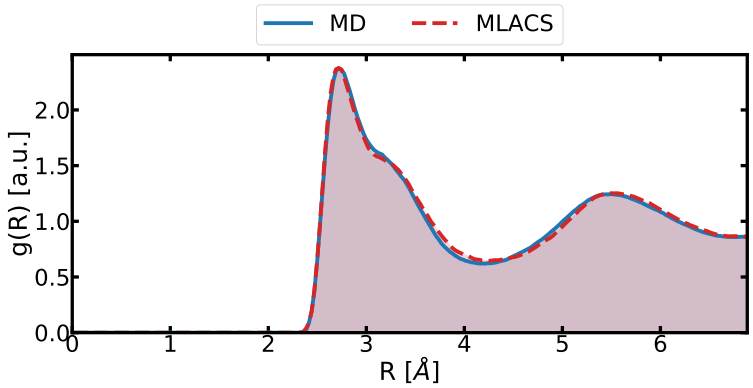


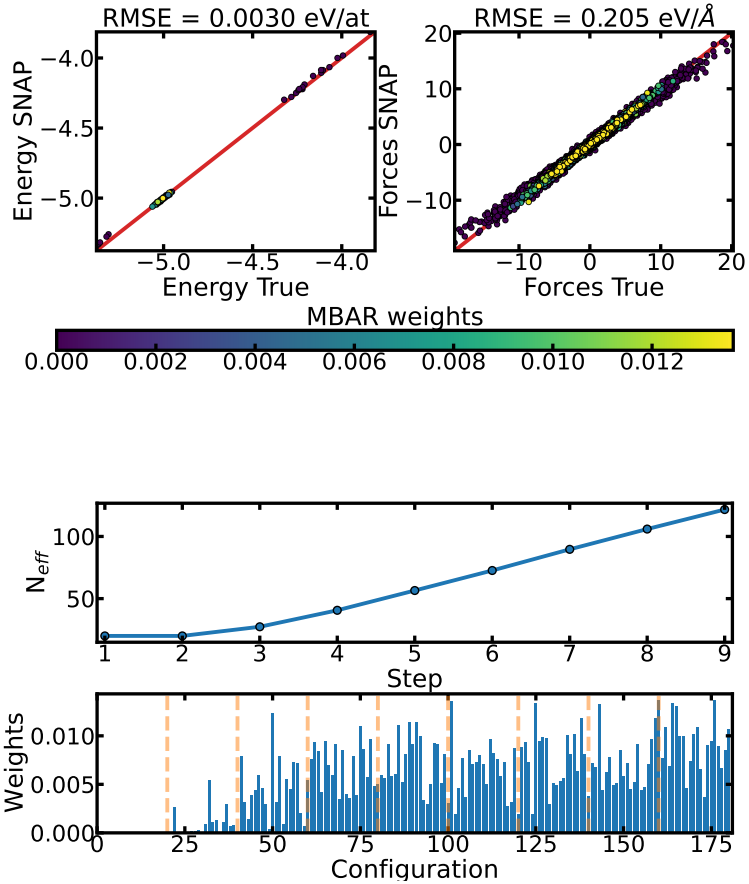




Number of configurations computed : MD 6000 - MLACS 240

Uranium MEAM - Liquid 2500K

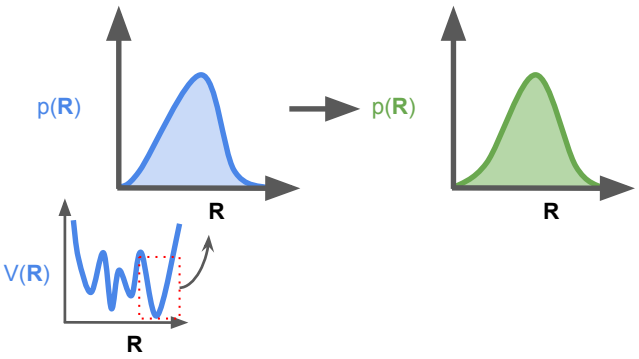


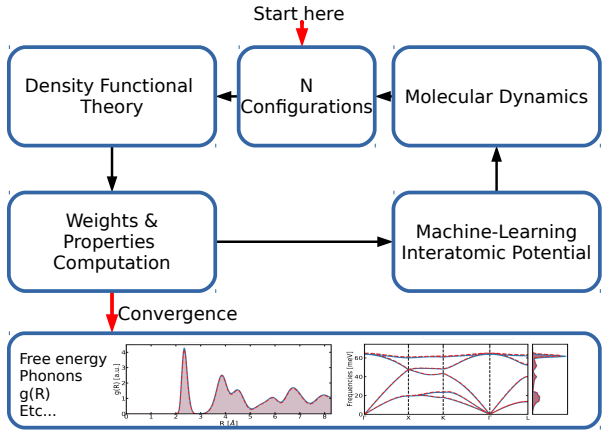


Number of configurations computed : MD 15000 - MLACS 200

5.4 Summary and some perspectives

Machine-Learning Assisted Canonical Sampling





- **MLIP** is seen as a probability distribution
- Self-consistent procedure to sample canonical distribution
- Fast and near-DFT accuracy
- Allows to compute electronic averages
- Allows to compute accurate free energy at the DFT level

Thank you for your attention

DE LA RECHERCHE À L'INDUSTRIE

Gibbs-Bogoliubov free energy

$$\widetilde{\mathcal{F}} = \mathcal{F}_0 + \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle \geqslant \mathcal{F}$$

gradient

$$\frac{\partial \widetilde{\mathcal{F}}}{\partial \gamma_k} = \langle V(\mathbf{R}) \rangle \langle X_k \rangle - \langle V(\mathbf{R}) X_k \rangle - \langle V(\mathbf{R}) \rangle \langle X_k \rangle + \langle V(\mathbf{R}) X_k \rangle$$

Weighted Least-Squares Property

$$\langle V(\mathbf{R}) \rangle = \langle V(\mathbf{R}) \rangle$$

5.5 Machine-Learning Assisted Canonical Sampling

Cumulant expansion of the free energy

$$\mathcal{F} = \mathcal{F}_0 + \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle - \frac{\beta}{2} \left[\langle \left(V(\mathbf{R}) - V(\mathbf{R}) \right)^2 \rangle - \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle^2 \right] + \dots$$

One can recognise the Gibbs-Bogoliubov free energy

$$\mathcal{F} = \widetilde{\mathcal{F}} - \frac{\beta}{2} \left[\langle \left(V(\mathbf{R}) - V(\mathbf{R}) \right)^2 \rangle - \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle^2 \right] + \dots$$

The second order cumulant of the free energy difference can be used as a measure of the accuracy of the simulation (as well as a correction to the free energy)

$$\Delta\mathcal{F}^{(2)} = -\frac{\beta}{2} \left[\langle \left(V(\mathbf{R}) - V(\mathbf{R}) \right)^2 \rangle - \langle V(\mathbf{R}) - V(\mathbf{R}) \rangle^2 \right]$$