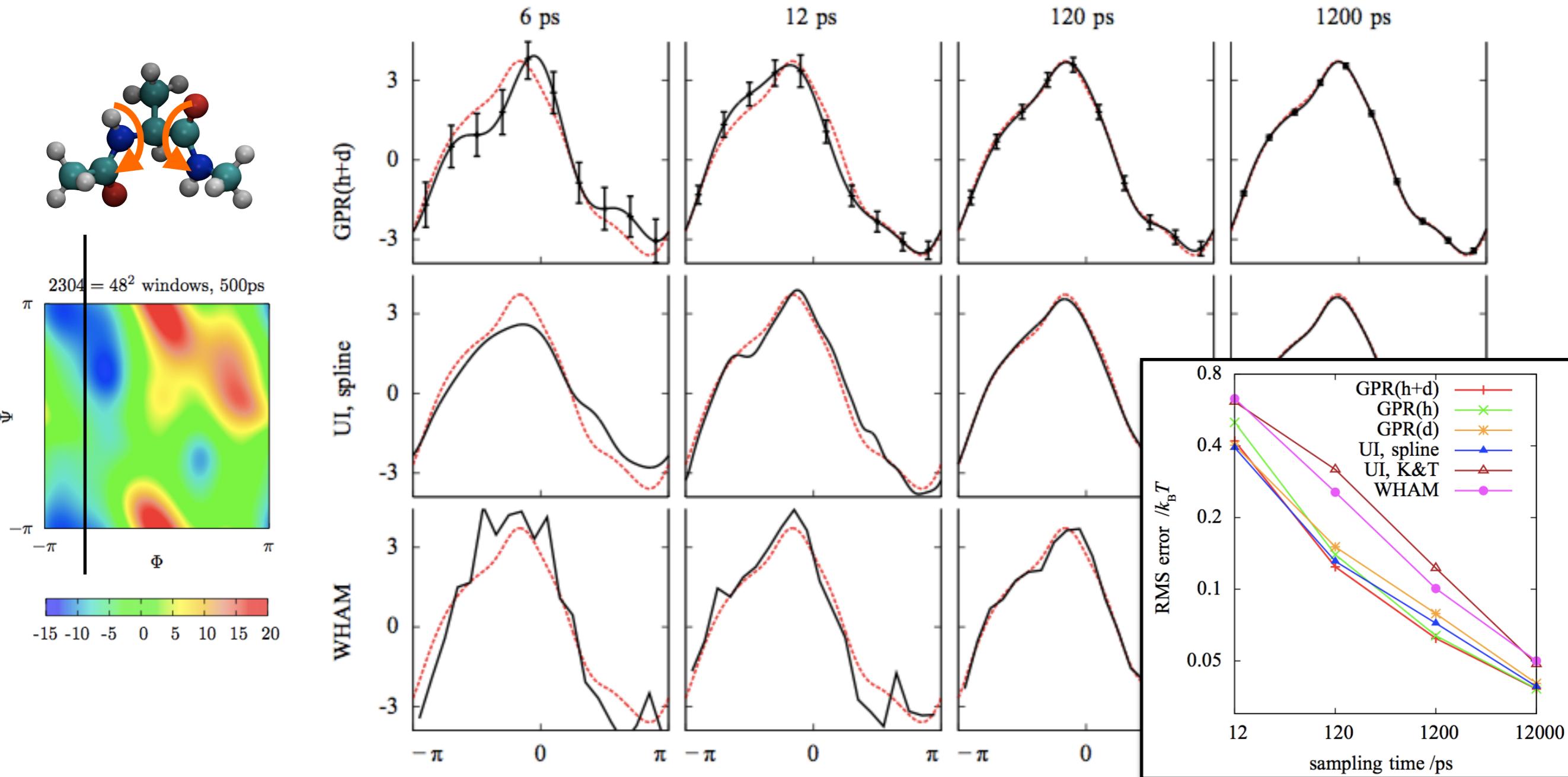


# 1D reconstruction (dialanine slice)

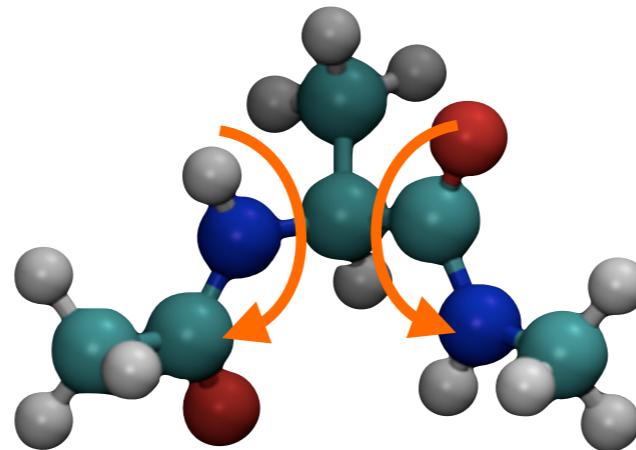
$$\bar{y}(x) = \mathbf{k}_{f,f'}^T (\mathbf{K}_{f',f'} + \sigma_\nu^2 \mathbf{I})^{-1} \mathbf{y}'$$

$$[\mathbf{k}_{f,f'}]_i = \partial k(x, x_i) / \partial x_i$$

Gaussian Process regression



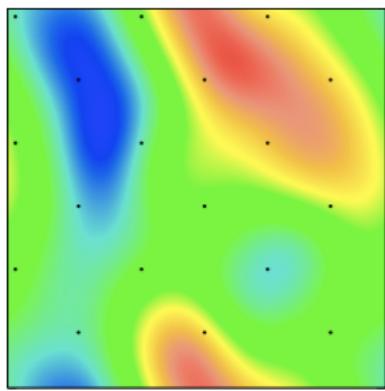
# 2D example (dialanine)



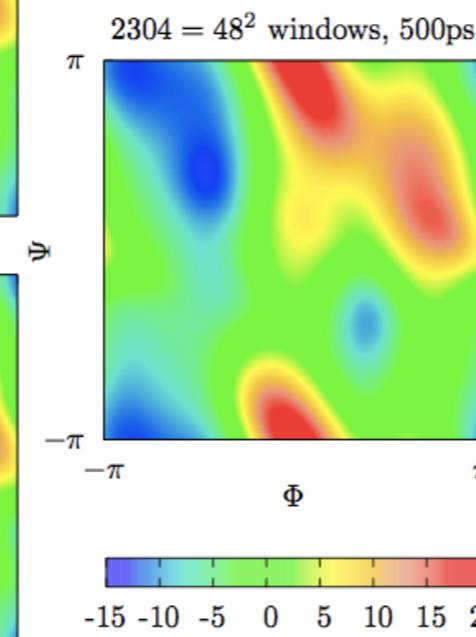
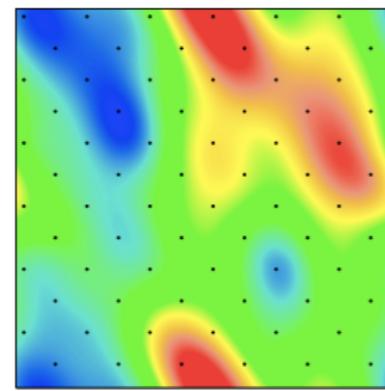
Least squares  
kernel fit  
(LSRBF)

Gaussian process  
(GPR)

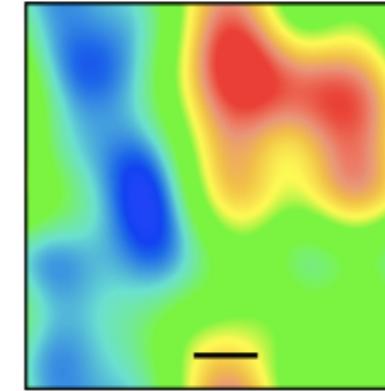
$18 = 2 \times 3^2$  windows, 500ps



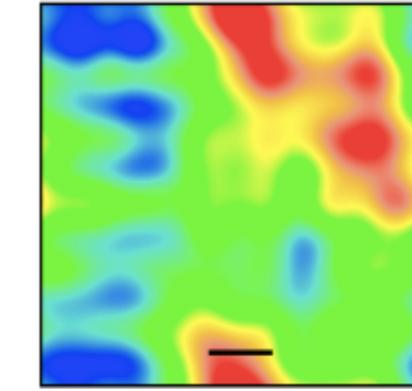
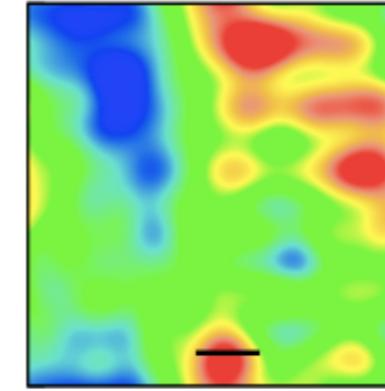
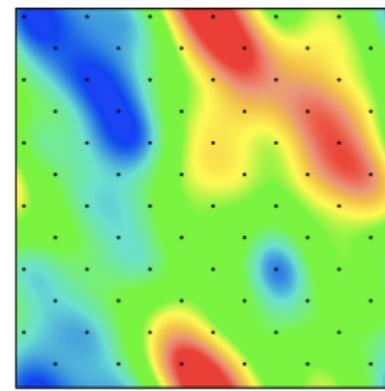
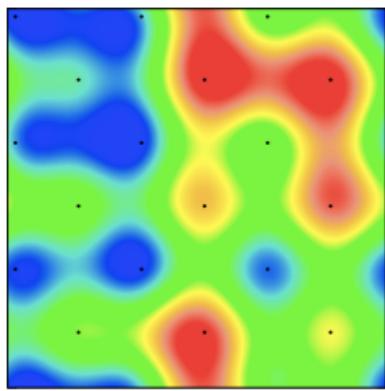
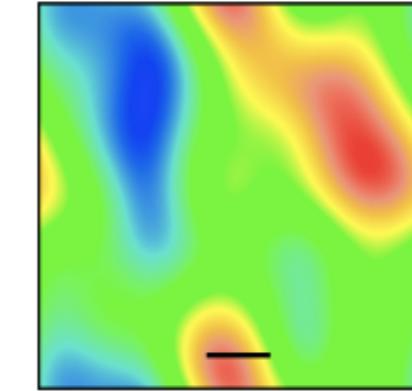
$72 = 2 \times 6^2$  windows, 500ps



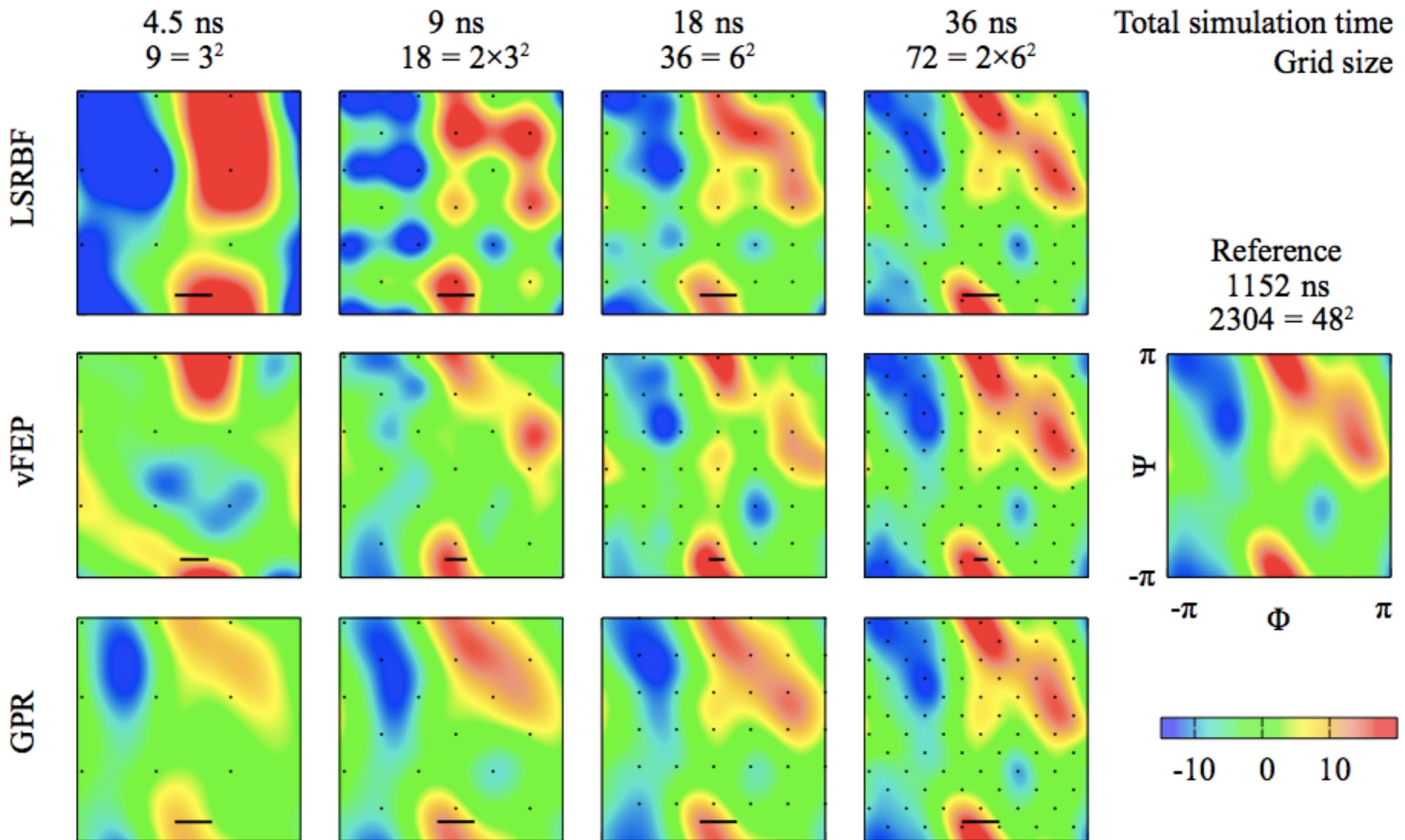
0.15 ps



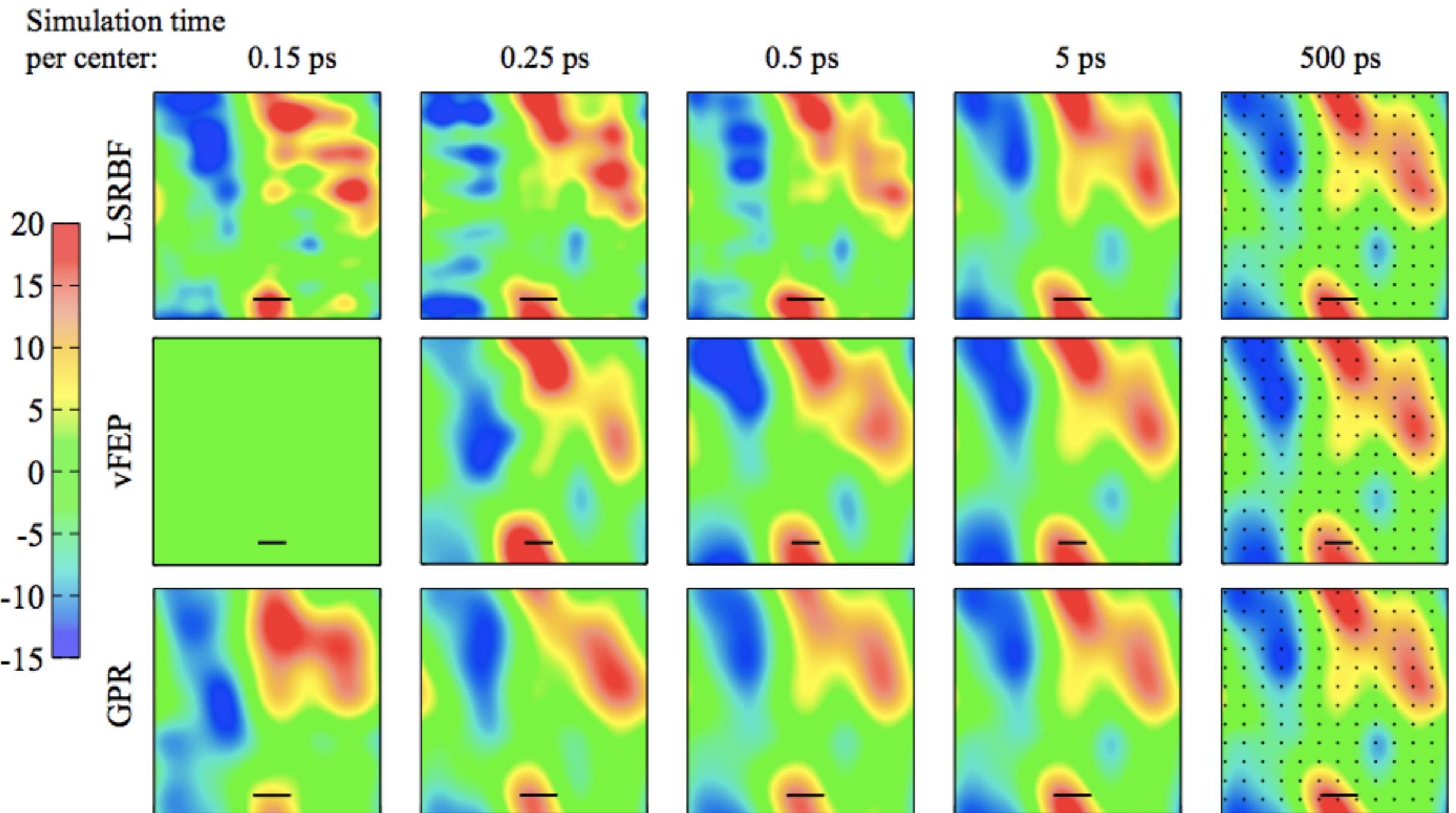
0.25 ps



# 2D: effect of grid size

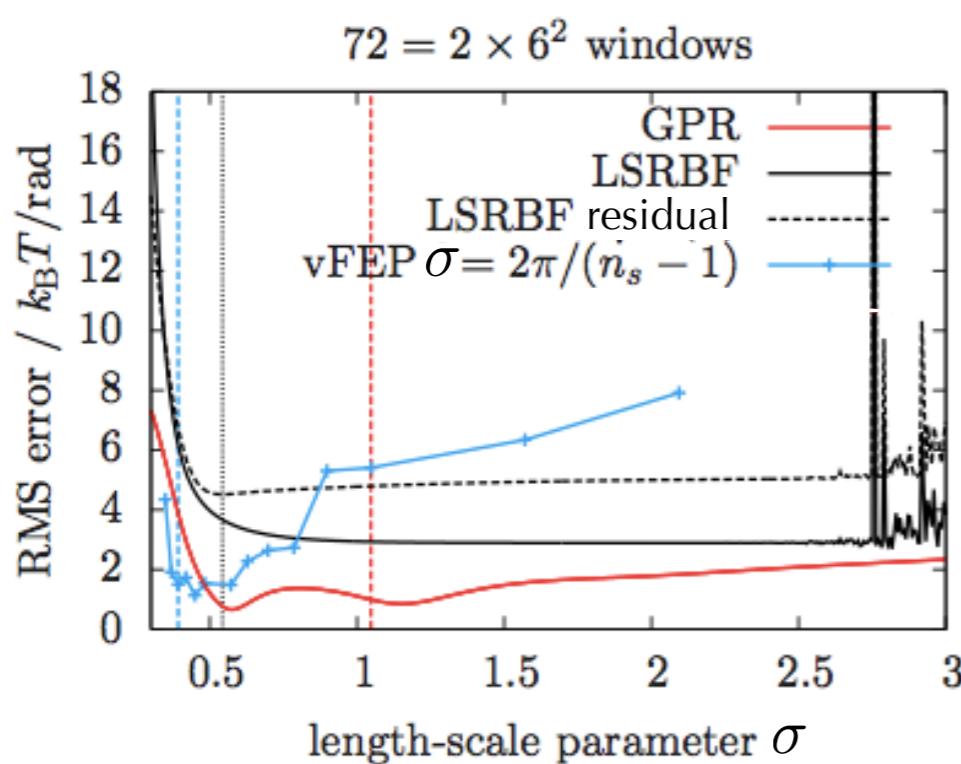
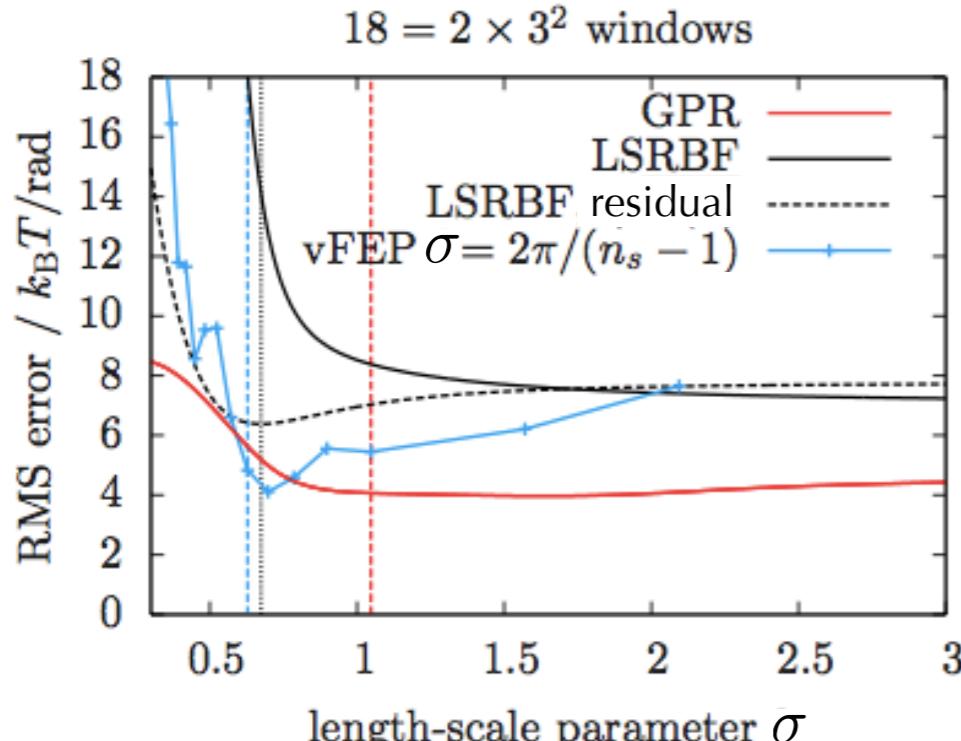


# 2D: effect of trajectory length (noise)

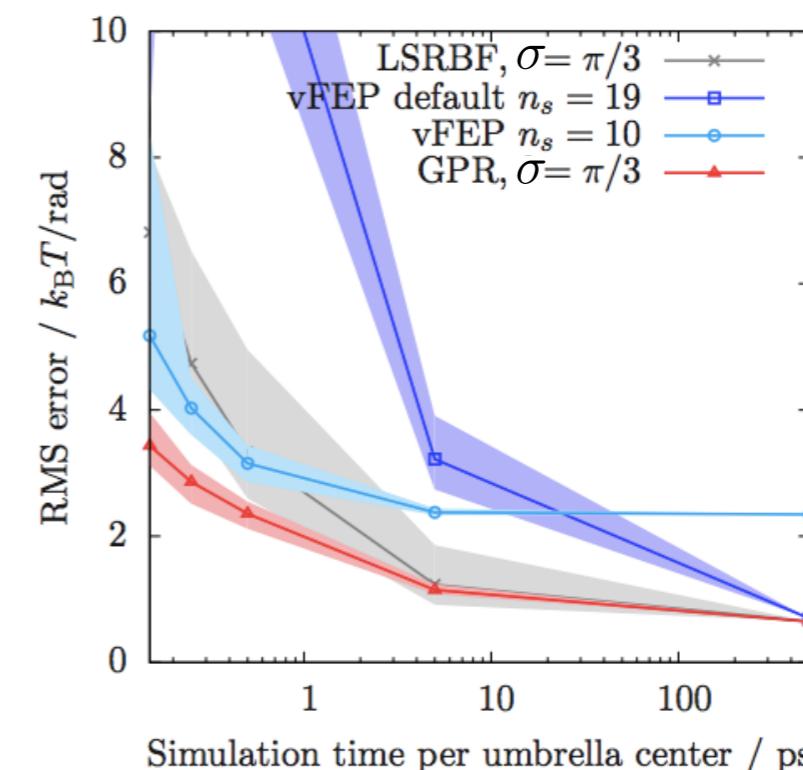
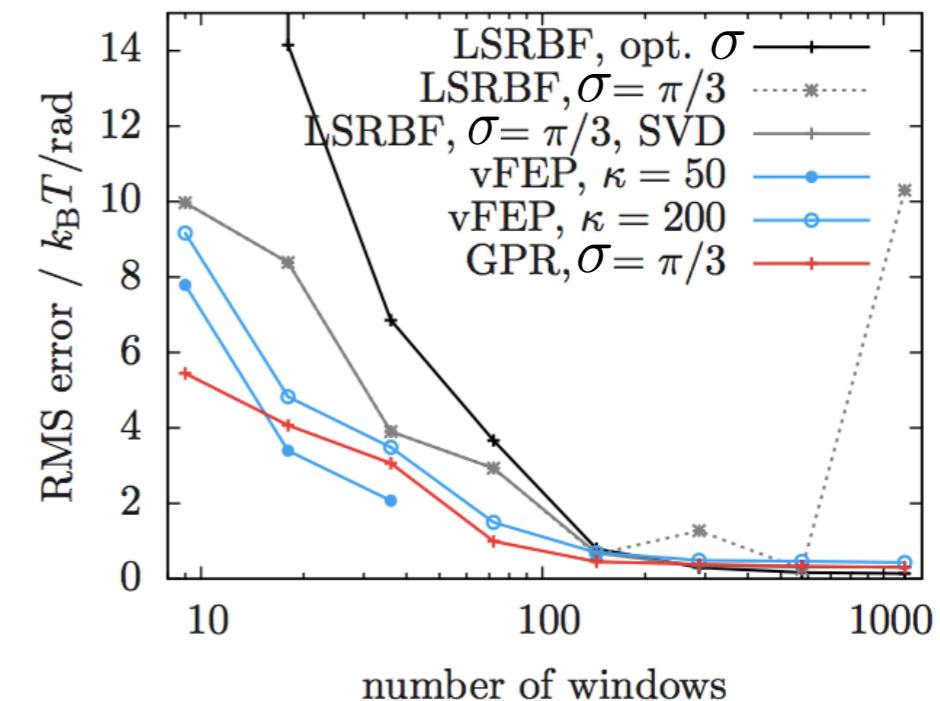


# GPR vs LSRBF

Should we optimise  
the length scale?

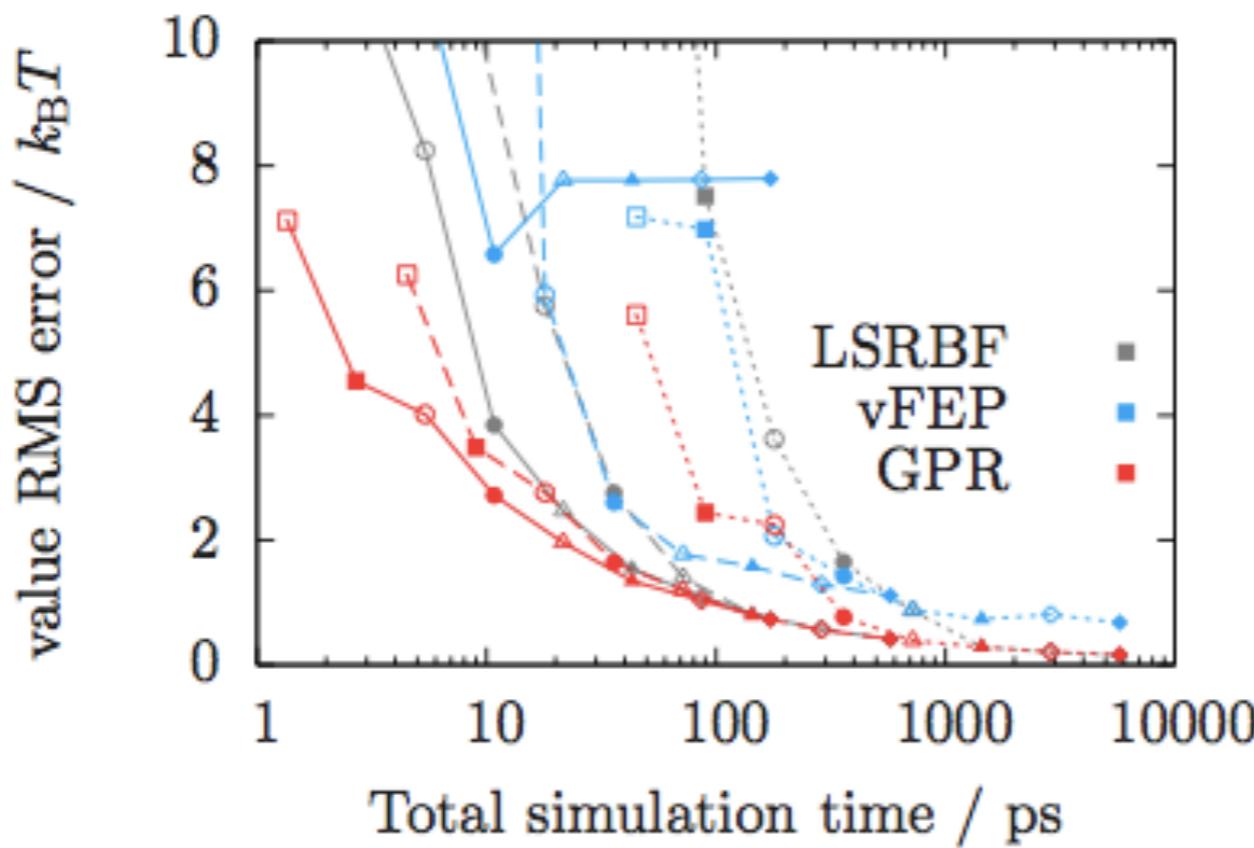
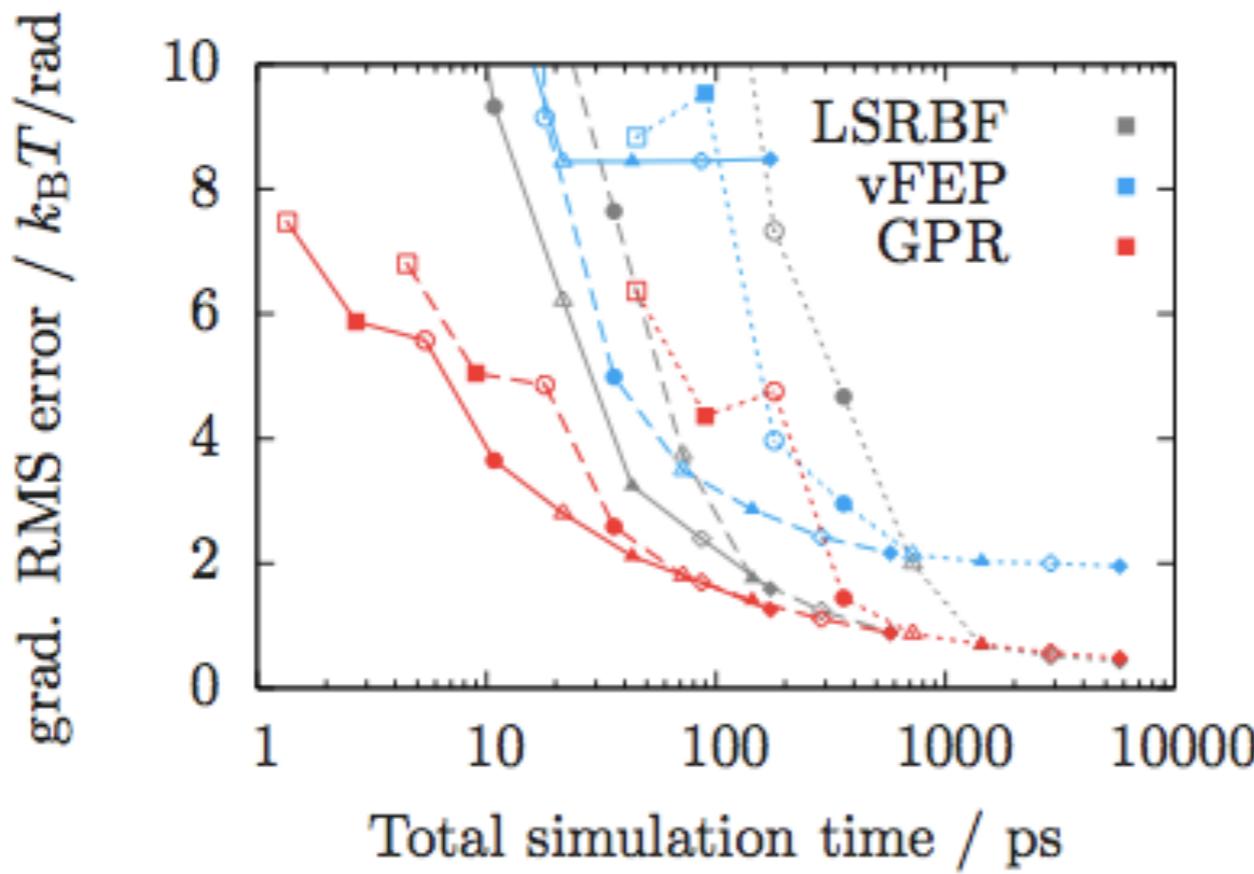


SVD for LSRBF ?



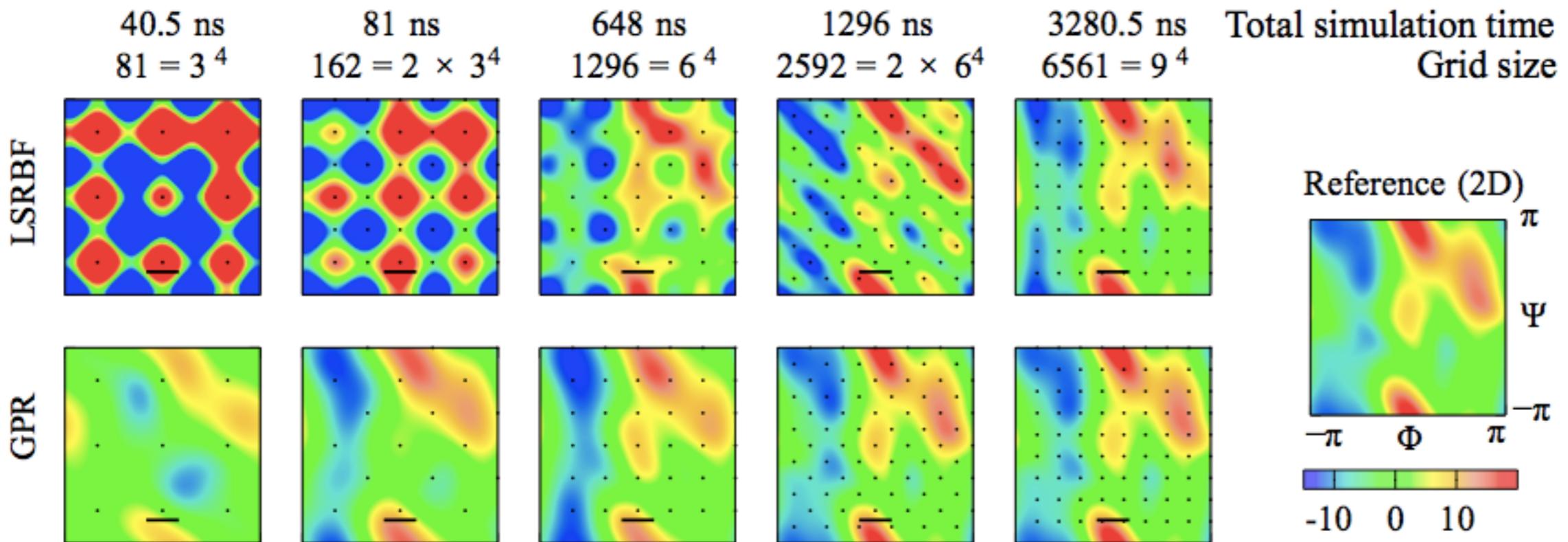
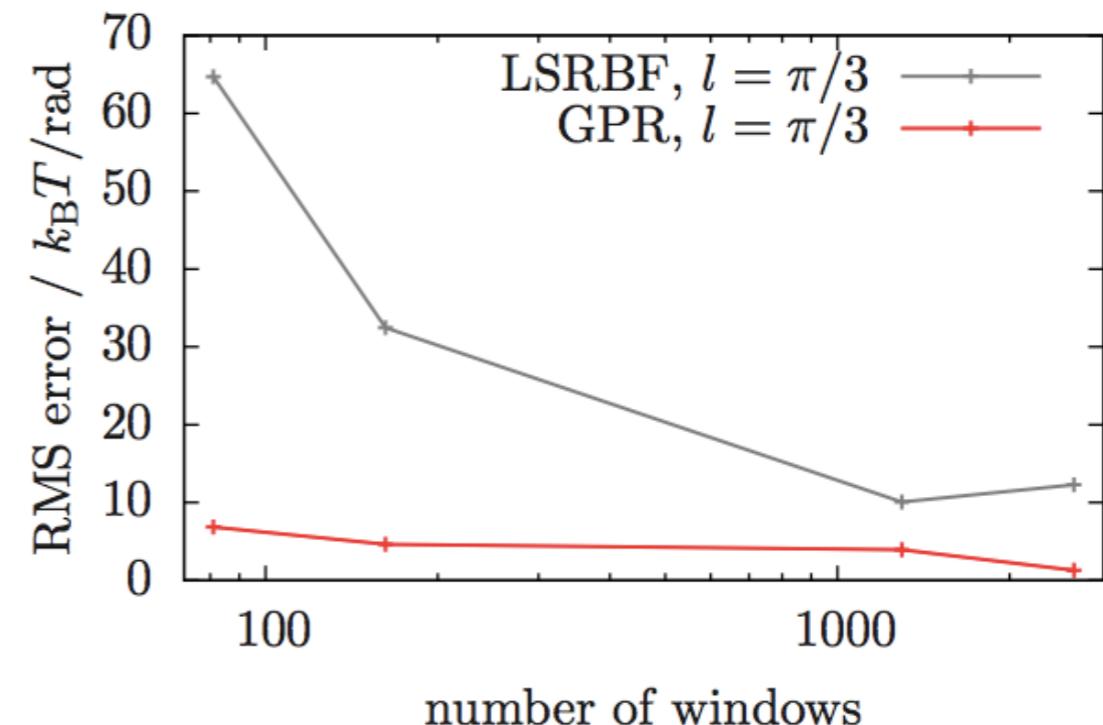
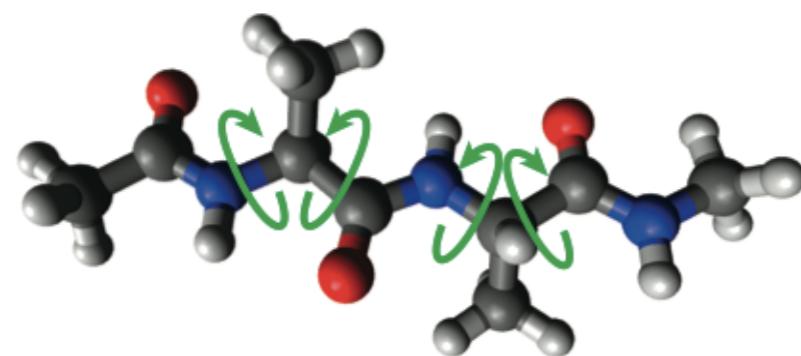
Noise  
tolerance

# Error vs total effort



GPR: denser grid  
with noisier data  
always better

# 4D: tripeptide



# Lecture 6: Nested Sampling

## Gábor Csányi

Engineering Laboratory  
 UNIVERSITY OF  
CAMBRIDGE

# What can we do with a fast atomistic model ?

Energy model



Explore *all* relevant configurations



Pressure-Temperature-Composition  
(equilibrium) phase diagram

# Probabilistic curve fitting

Data  $\{x_i\}$ ,  $\{y_i\}$ , fit function  $f(x; \alpha, \beta, \gamma)$

Likelihood:

$$P(\alpha, \beta, \gamma) \propto e^{-\sum_i |y_i - f(x_i; \alpha, \beta, \gamma)|^2}$$

Maximising likelihood often leads to overfitting

To choose between alternative models, instead of maximising likelihood, compute integral

$$\int d\alpha d\beta d\gamma P(\alpha, \beta, \gamma)$$

# Probability and statistical mechanics

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}$$

$D$ : data  
 $M$ : model

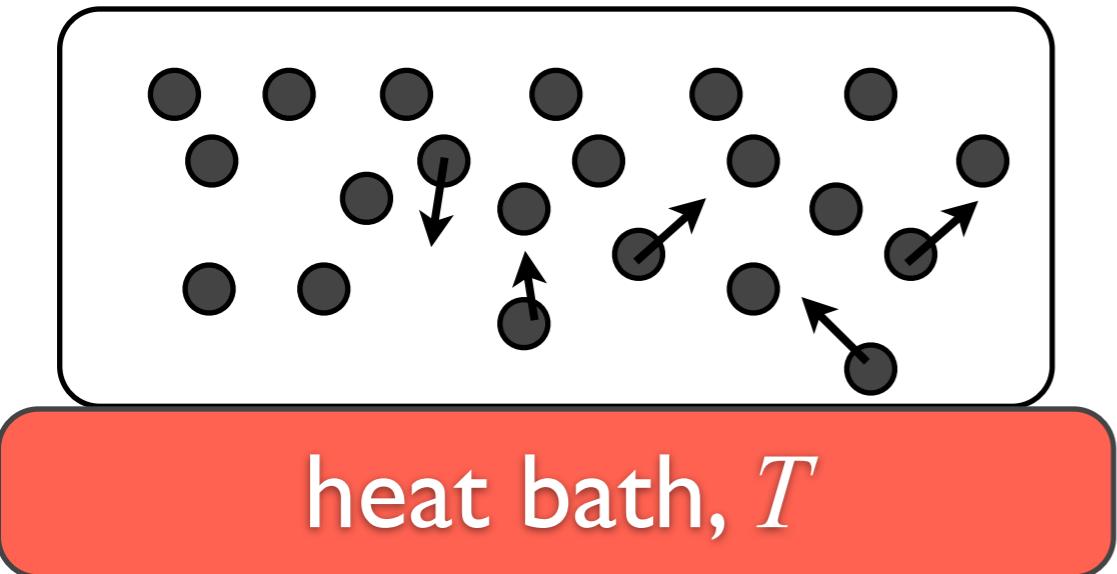
For choosing between models, compute “evidence”

$$P(D|M) = \int d\theta P(D|\theta, M)P(\theta|M) = \int d\theta L(\theta)\pi(\theta)$$

$$\begin{array}{ccc} \theta & \longleftrightarrow & q, p \\ L(\theta) & \longleftrightarrow & \exp(-\beta E(q, p)) \\ \ln L & \longleftrightarrow & -\beta E \\ P(D|M) & \longleftrightarrow & Z \end{array}$$

John Skilling in 2006 invented **Nested Sampling** to compute  $P(D|M)$

# Boltzmann distribution



Microscopic system of  $N$  particles:  
 $3N$  positions  $q$  and momenta  $p$   
Energy  $E(q, p)$   
each state equally likely

Statistical mechanics:  $P(q, p) \sim \exp(-E(q,p)/kT)$

Observable  $A$ :

$$\langle A \rangle = \frac{1}{Z} \int dq dp A(q, p) \exp(-E/kT) = \int dt A(q(t), p(t))$$

Ergodicity:  $q(t), p(t)$  from a process whose stationary distribution is Boltzmann : MCMC, MD, etc.

# The partition function

$$Z(V, \beta) = \int dqdp \exp(-\beta E)$$

$V$ : volume  
 $\beta$ :  $1/kT$

Access to  $Z$  would be useful:

Internal energy:  $U(V, \beta) = \langle E \rangle = -\frac{\partial \ln Z(V, \beta)}{\partial \beta}$

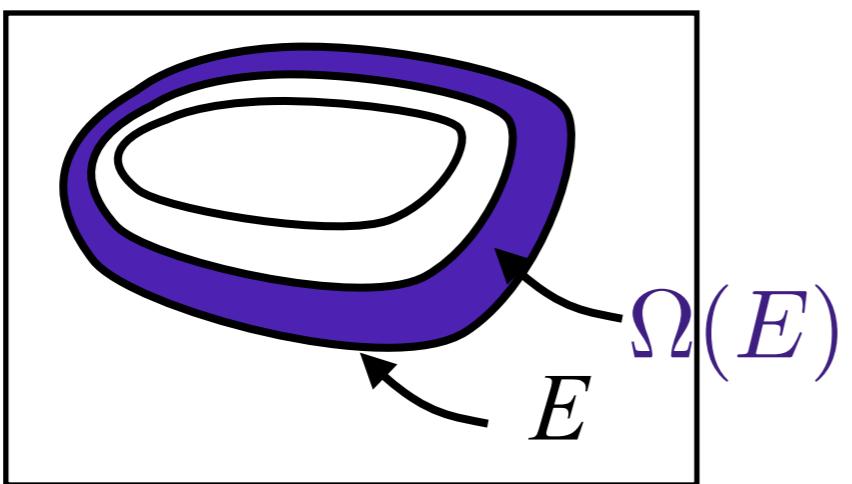
Free energy:  $F(V, \beta) = -\frac{1}{\beta} \ln Z(V, \beta)$

Heat capacity:  $C_V = \left( \frac{\partial U}{\partial T} \right)_V = -\frac{\partial}{\partial T} \frac{\partial \ln Z}{\partial \beta}$

$Z$  widely held to be “intractable” for nontrivial systems

# Obtaining the partition function

$$\begin{aligned} Z(N, V, \beta) &= Z_m(N, \beta) \int d\mathbf{q} e^{-\beta E(\mathbf{q})} \\ &= Z_m(N, \beta) \int dE \Omega(E) e^{-\beta E} \end{aligned}$$

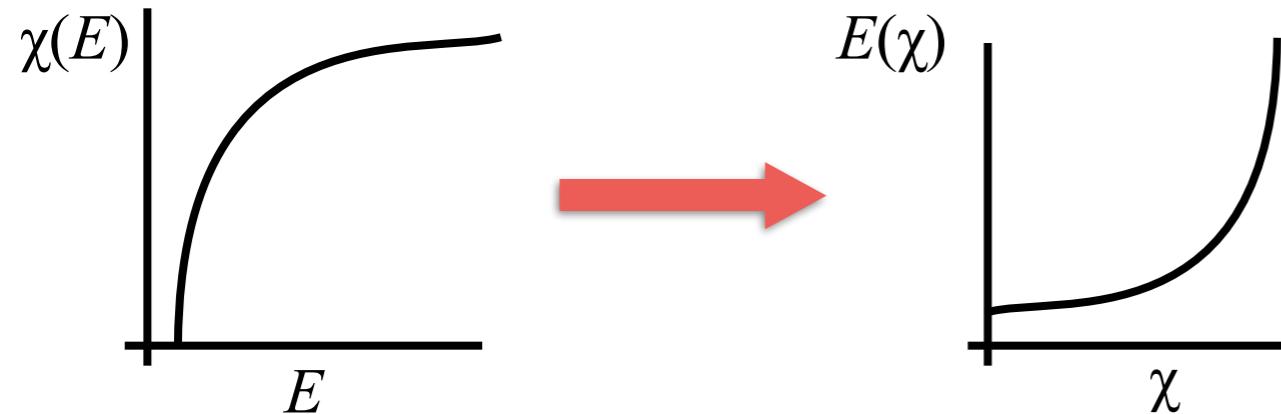


“Density  
of states”

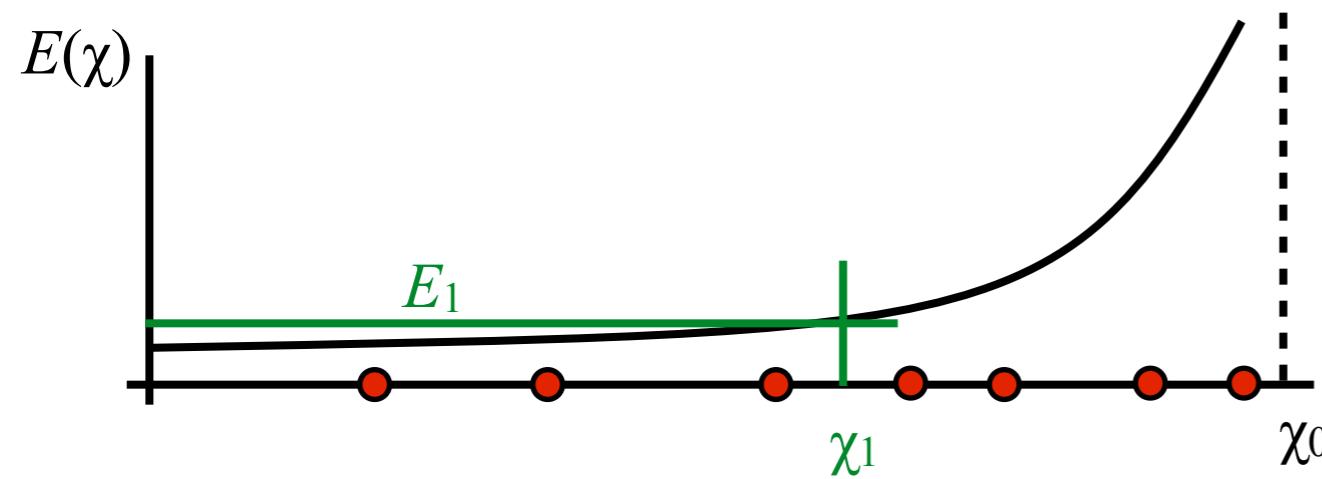
Consider cumulative density:

$$\chi(E) = \int_{-\infty}^E dE' \Omega(E')$$

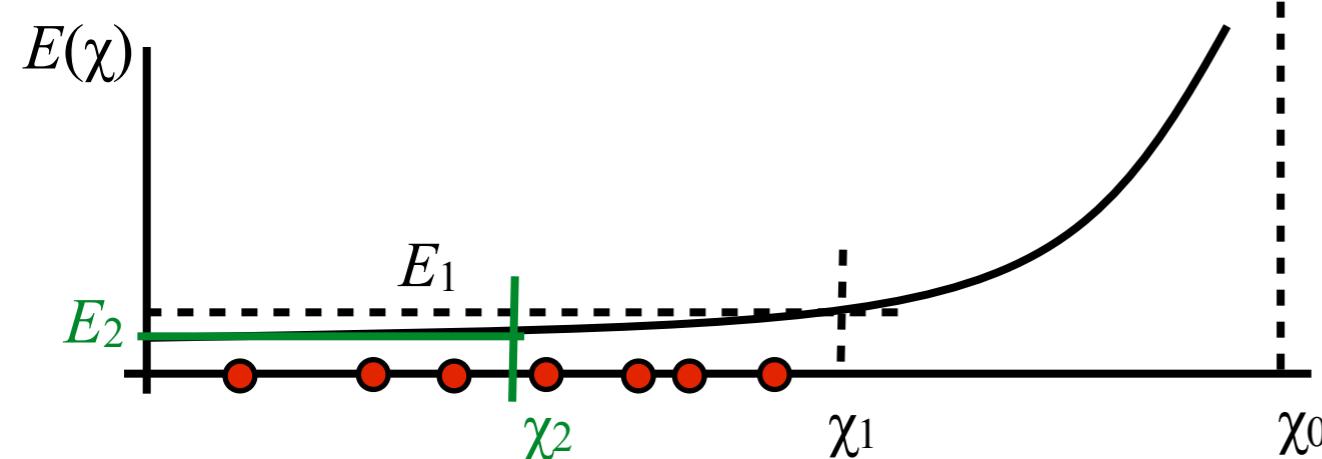
# The main trick



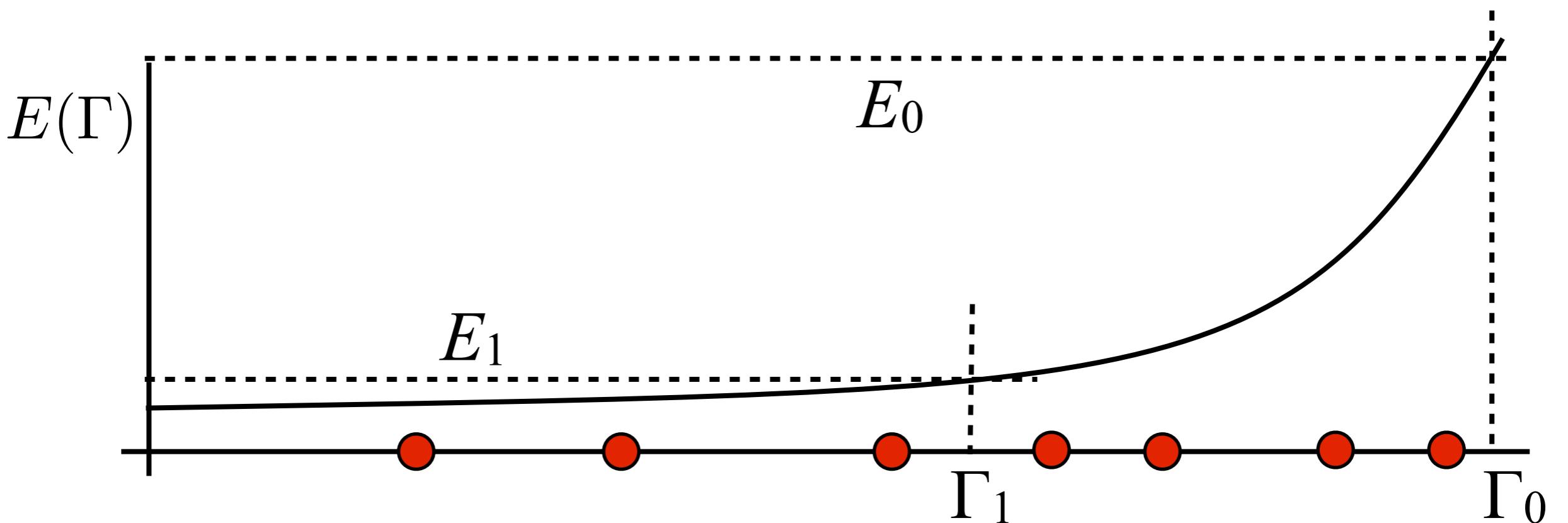
- At  $E = \infty$ , we have an ideal gas,  $\chi_0 = V^N$
- Instead of  $\chi(E)$ , compute  $E(\chi)$



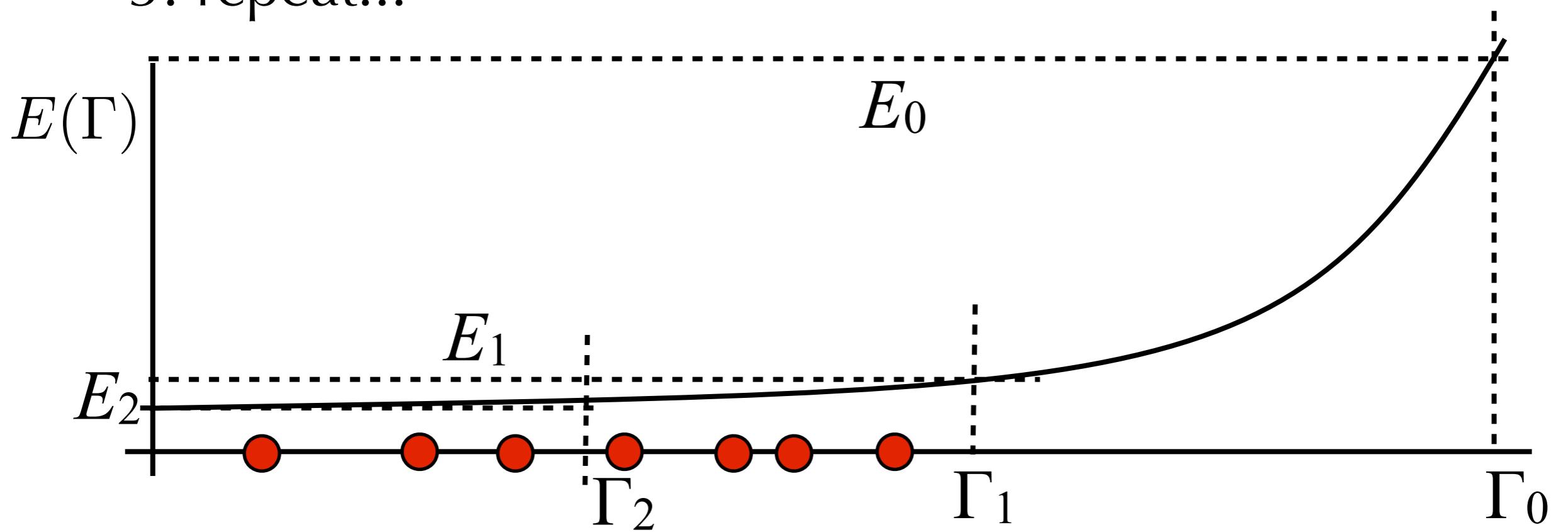
1. obtain K uniform samples such that  $E(q) < E_{\text{limit}}$
2. compute median:  $E(\chi_1) = E_1$ ,  $\chi_1 \approx \chi_0/2$ ,  $E_{\text{limit}} \leftarrow E_1$
3. repeat...



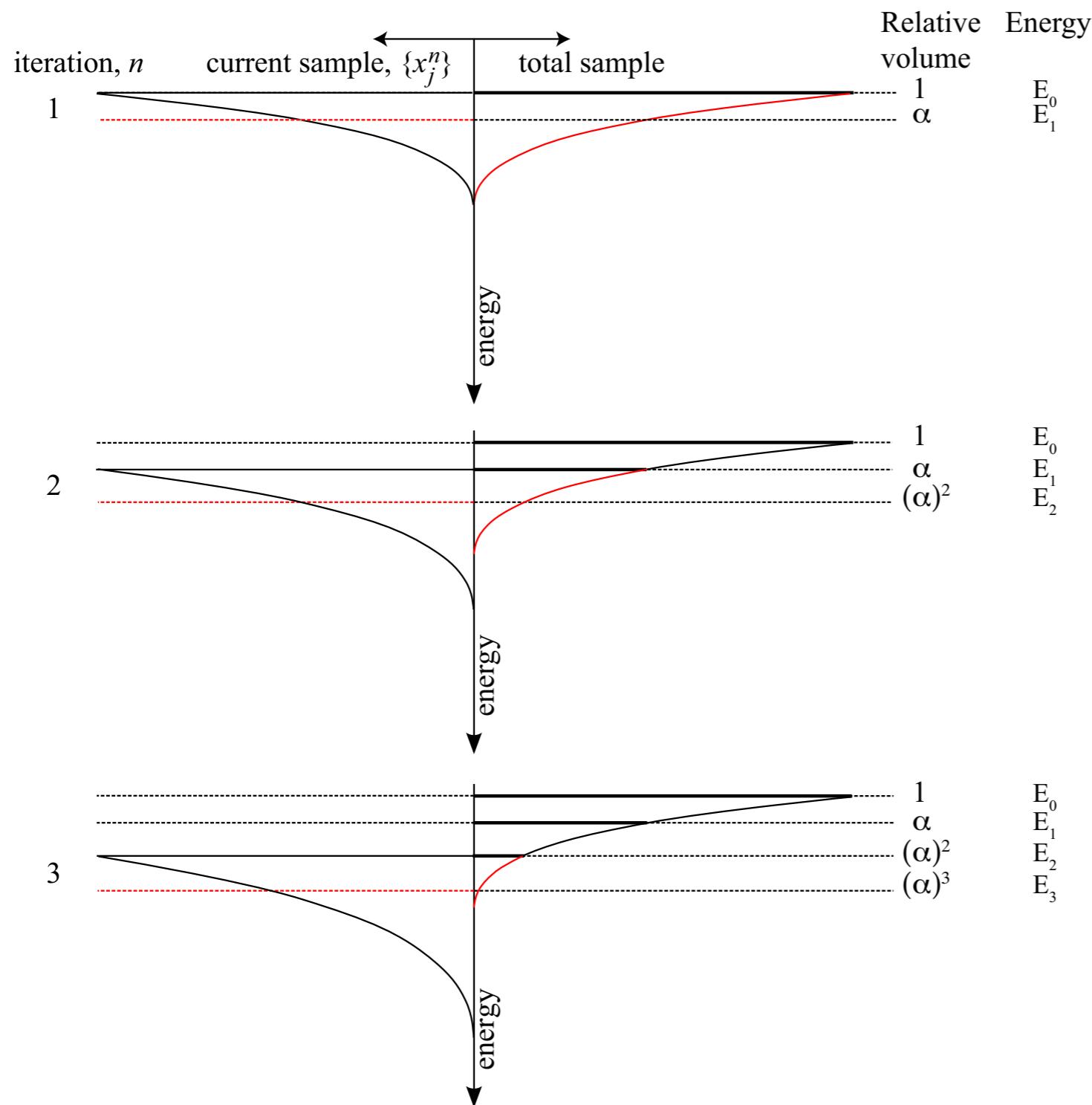
$$\begin{aligned} \chi_n &\approx \chi_0 \alpha^n \\ \Omega(E_n) &= \chi_0 (\alpha^{n-1} - \alpha^n) \\ Z &= Z_m \int dE \Omega(E) e^{-\beta E} \end{aligned}$$



1. obtain K uniform samples such that  $E(q) < E_0$
2. compute median:  $E(\Gamma_1) = E_1$ ,  $\Gamma_1 \approx \Gamma_0/2$
3. repeat...



# “Squeeze” down the energy distribution



$\alpha = 1/2$  (median) :  
keep half of  
distribution

$\alpha = K/(K+1)$  :  
throw away one  
sample / iteration

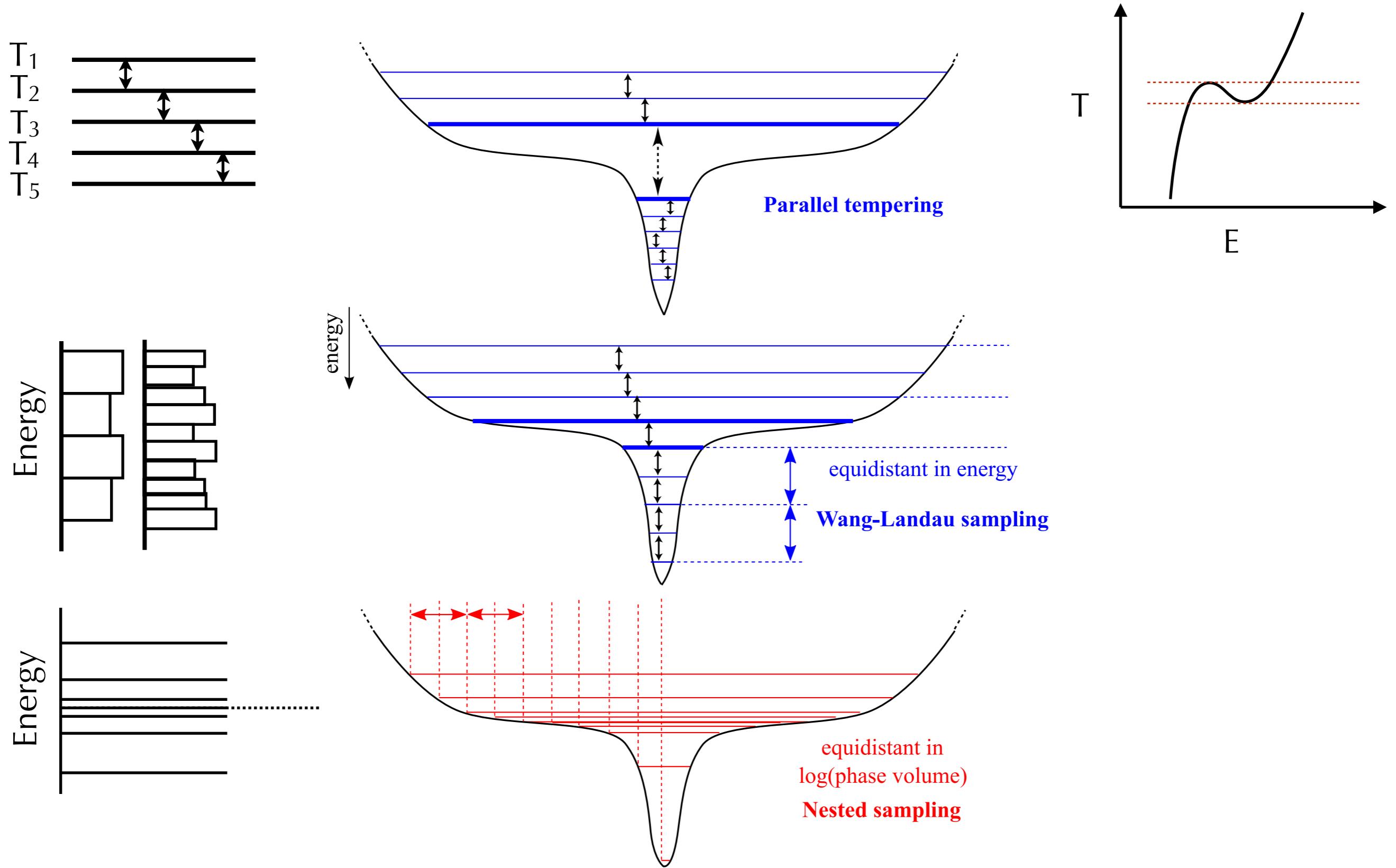
$$\Gamma_n \approx \Gamma_0 \alpha^n$$

$$\Omega(E_n) = \Gamma_0(\alpha^n - \alpha^{n+1})$$

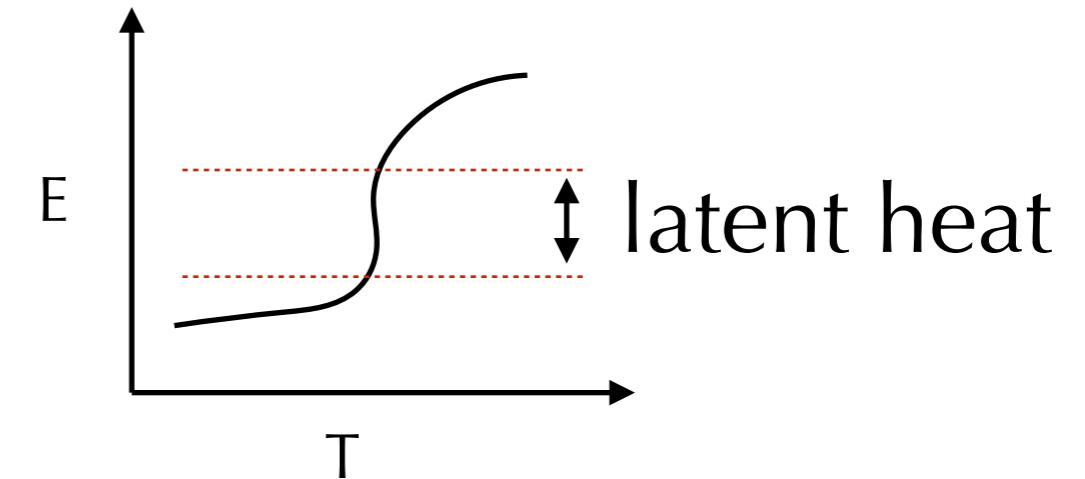
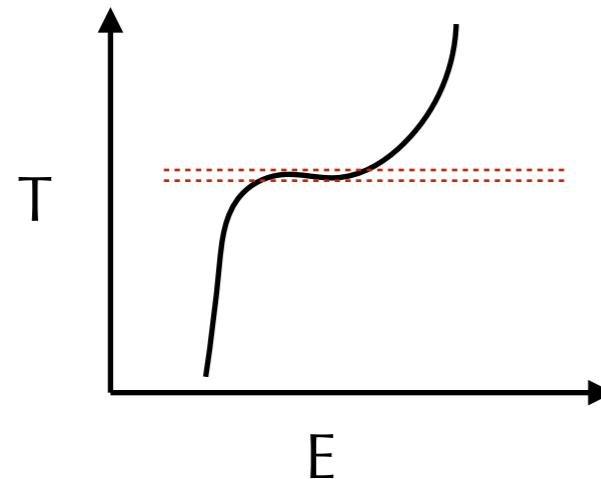
# Remarks

- How to obtain  $K$  uniform samples with  $E(\mathbf{q}) < E_n$  ?
  - Need a “black box” uniform sampler
  - Clone existing sample, walk  $L$  steps to decorrelate
  - MCMC: isotropic Gaussian rejection sampling (very dumb!)
- Two convergence parameters:  $K, L$  - *not independent*
- Error
  - “Resolution” in configuration space:  $1/K$
  - Cumulative error in  $\log Z \sim K^{-1/2}$
- Linear scaling cost with # of iterations and  $K$
- $\Omega(E)$  independent of  $T$
- Analogies with RESTART, Cross-Entropy method etc.

# Equilibration near a first order phase transition



# How to detect a phase transition



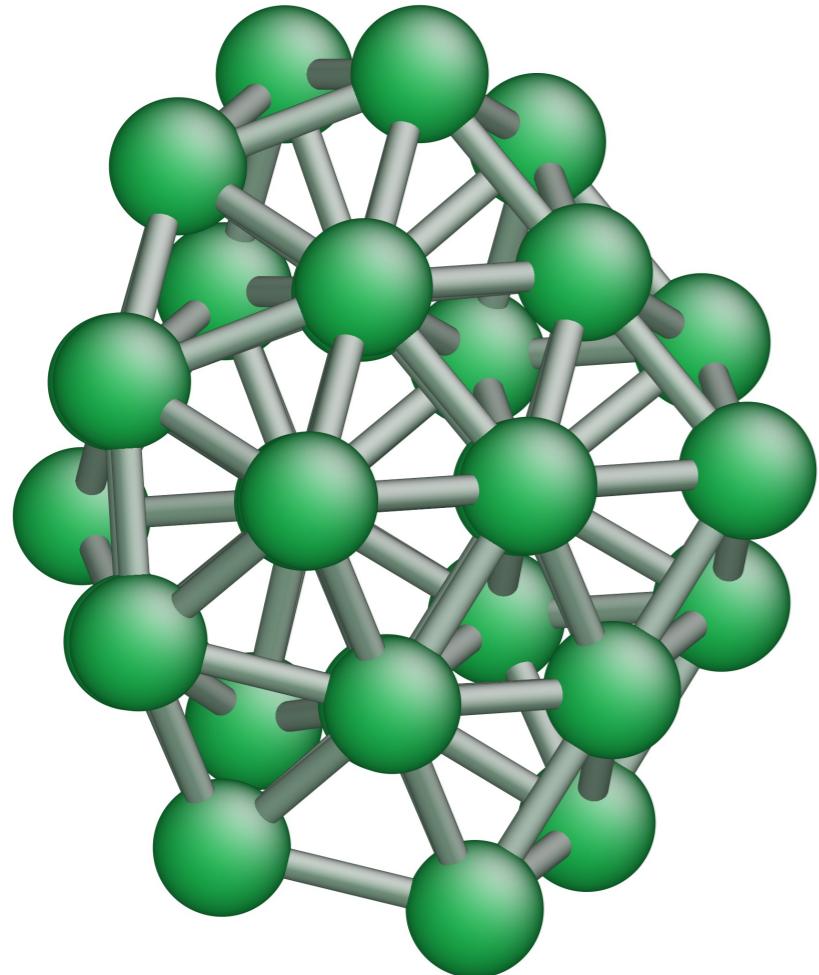
Heat capacity:  $C_v = \frac{\partial E}{\partial T}$

Evaluate using our samples:

$$C_v = \frac{\partial}{\partial T} \left( \frac{1}{Z(T)} \int dE E \Omega(E) e^{-E/kT} \right)$$

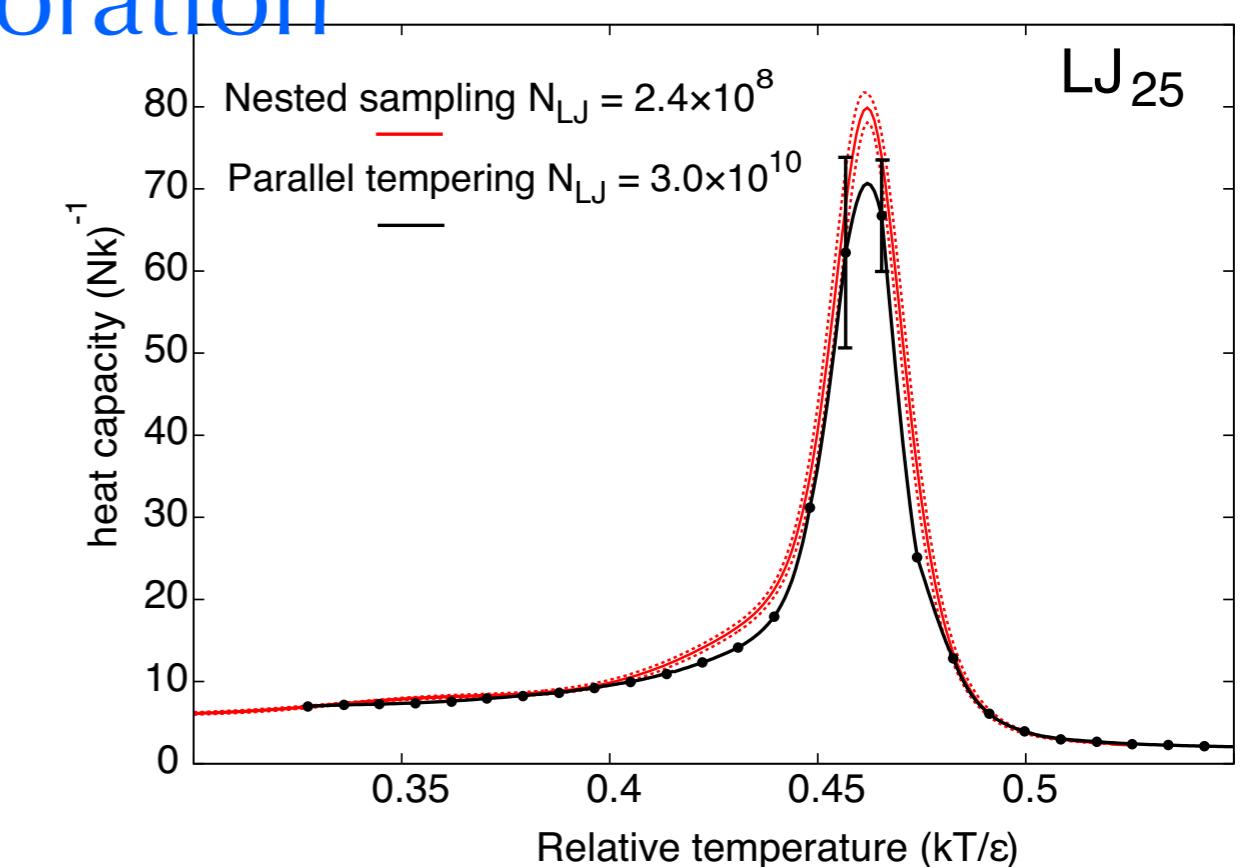
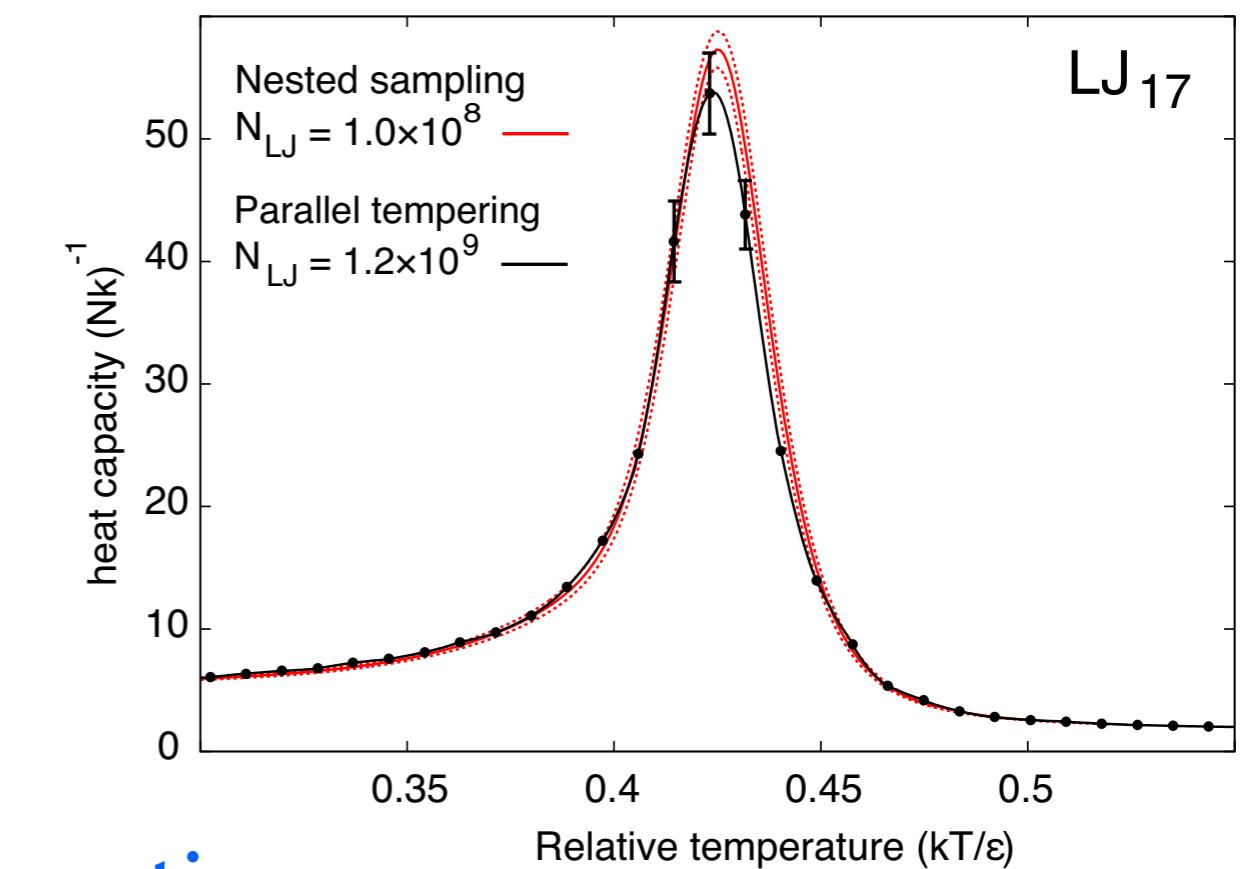
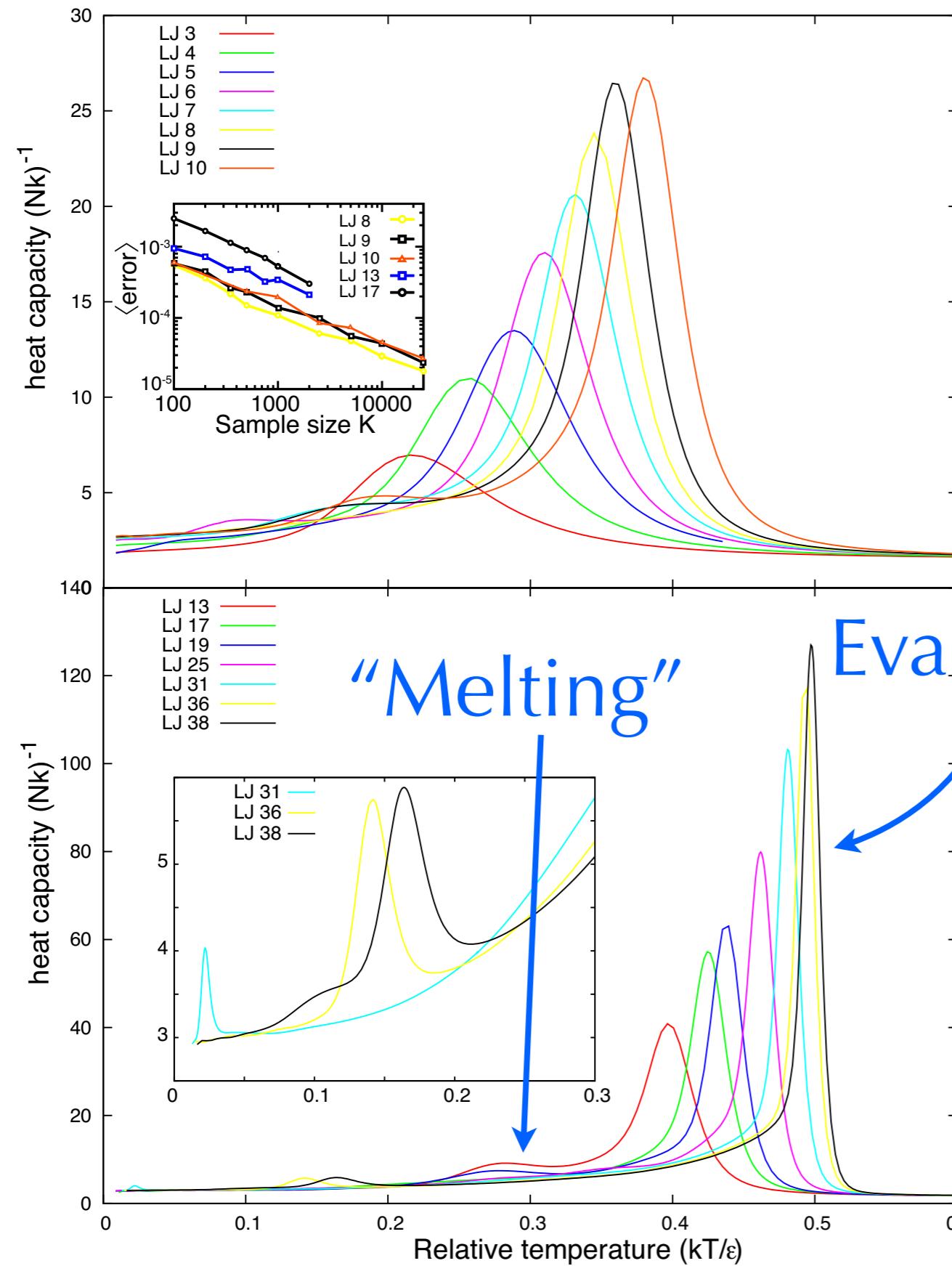
# Test case: Lennard-Jones clusters

$$E_{LJ} = \sum_{ij} \frac{1}{(r_{ij})^{12}} - \frac{1}{(r_{ij})^6}$$

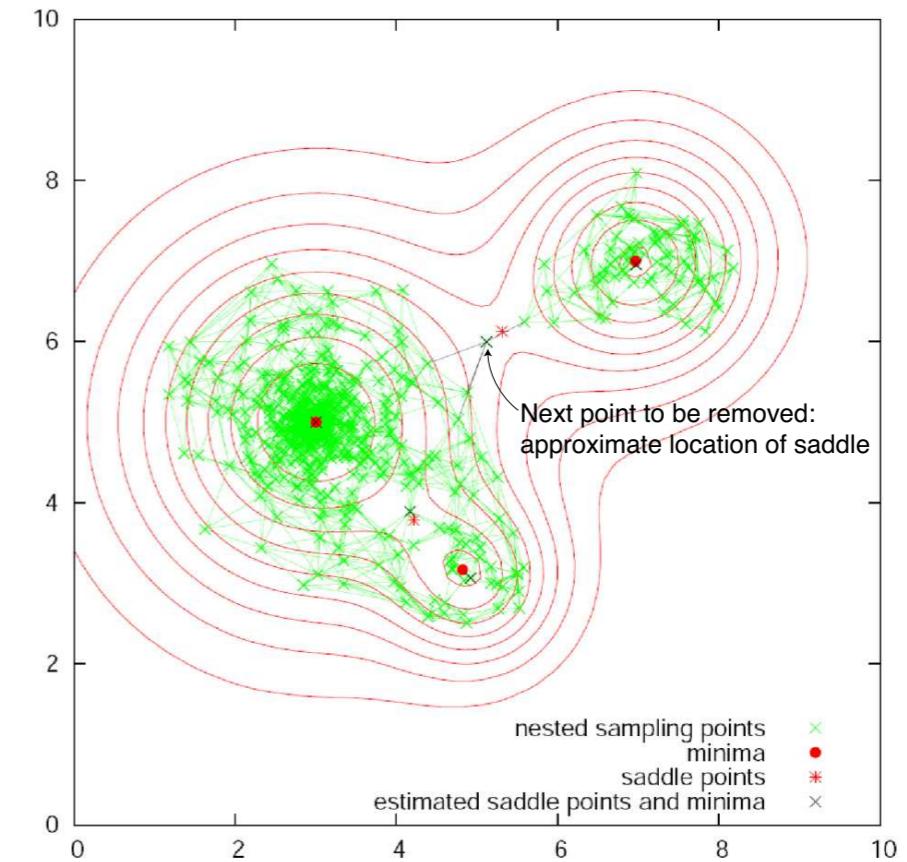
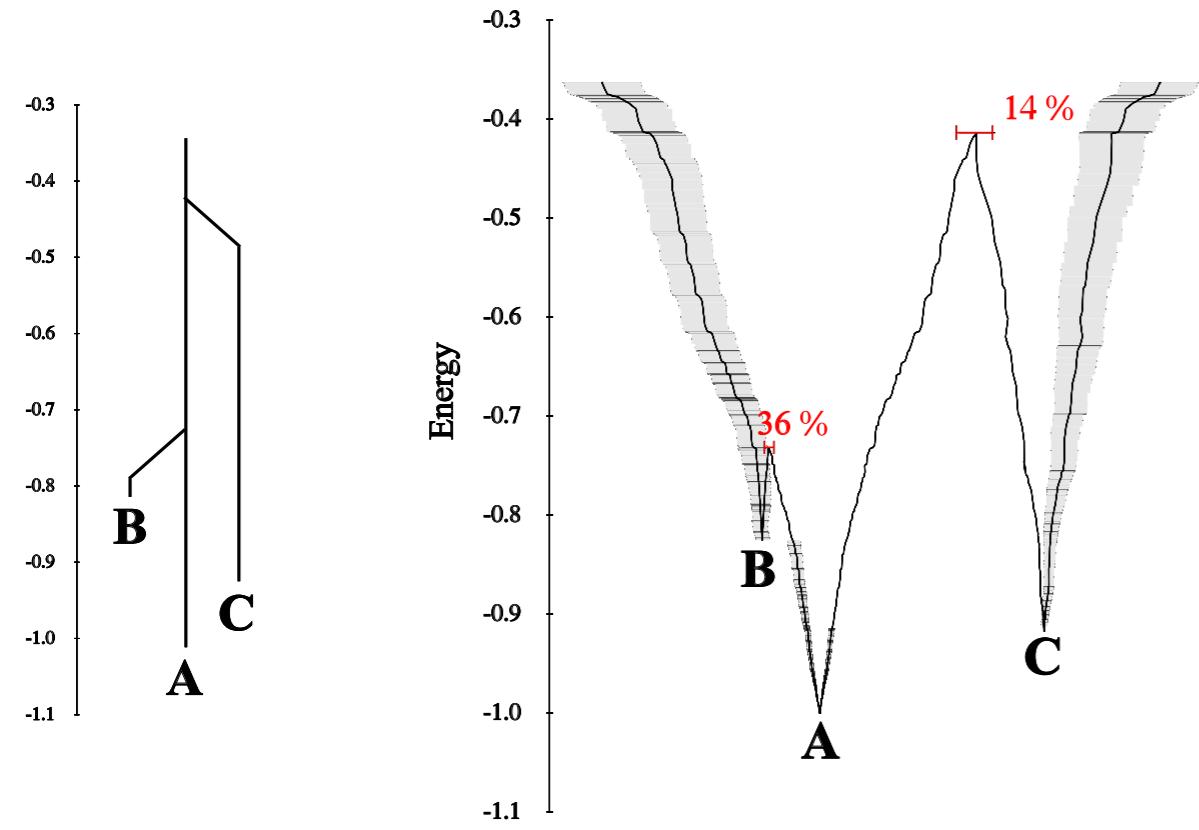
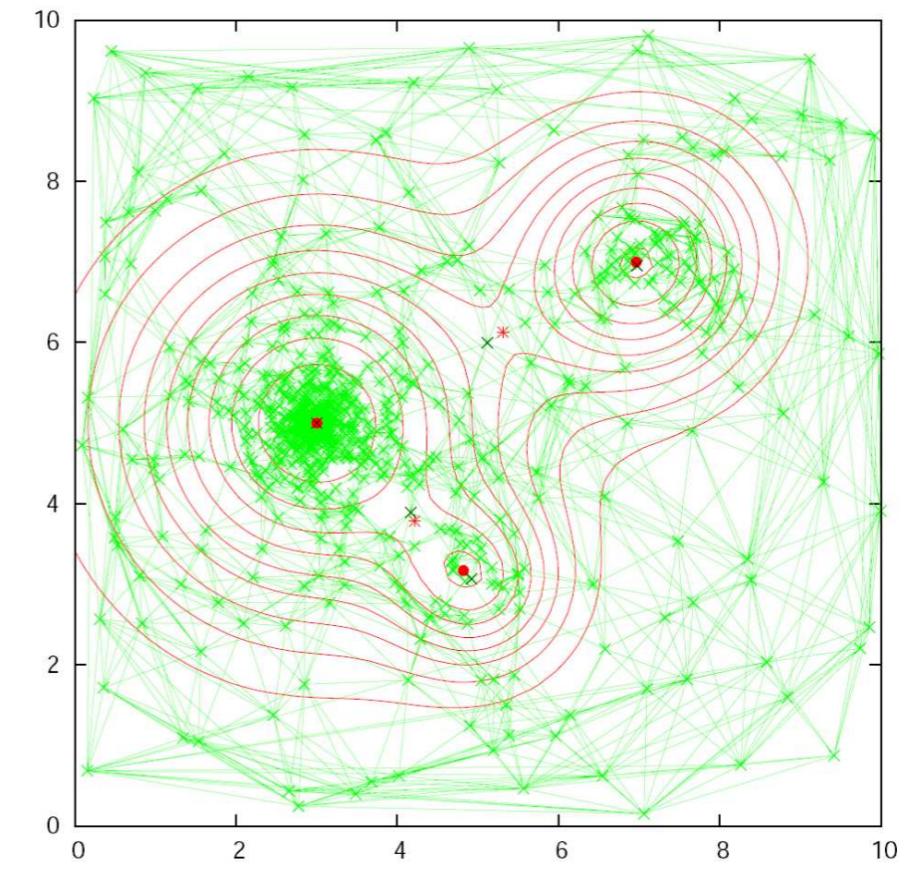
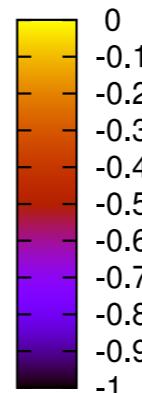
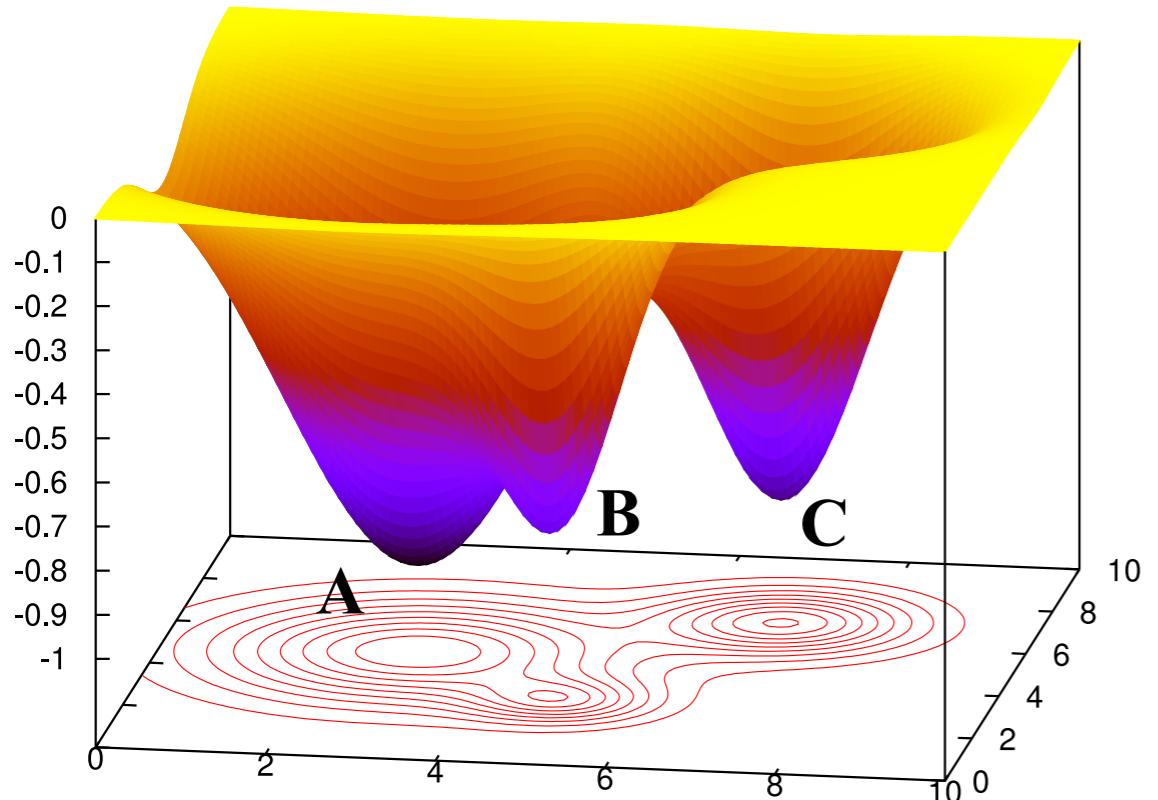


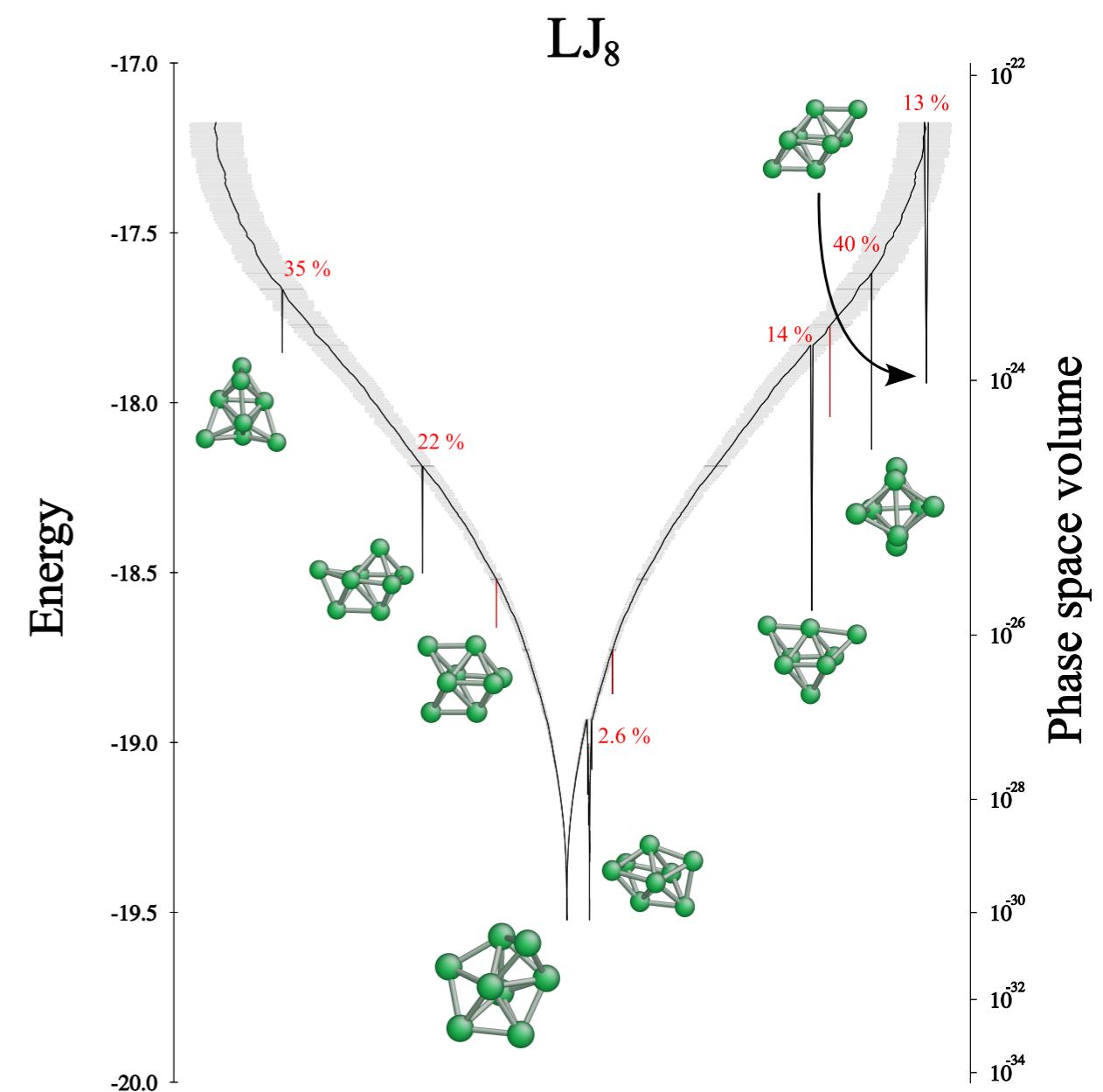
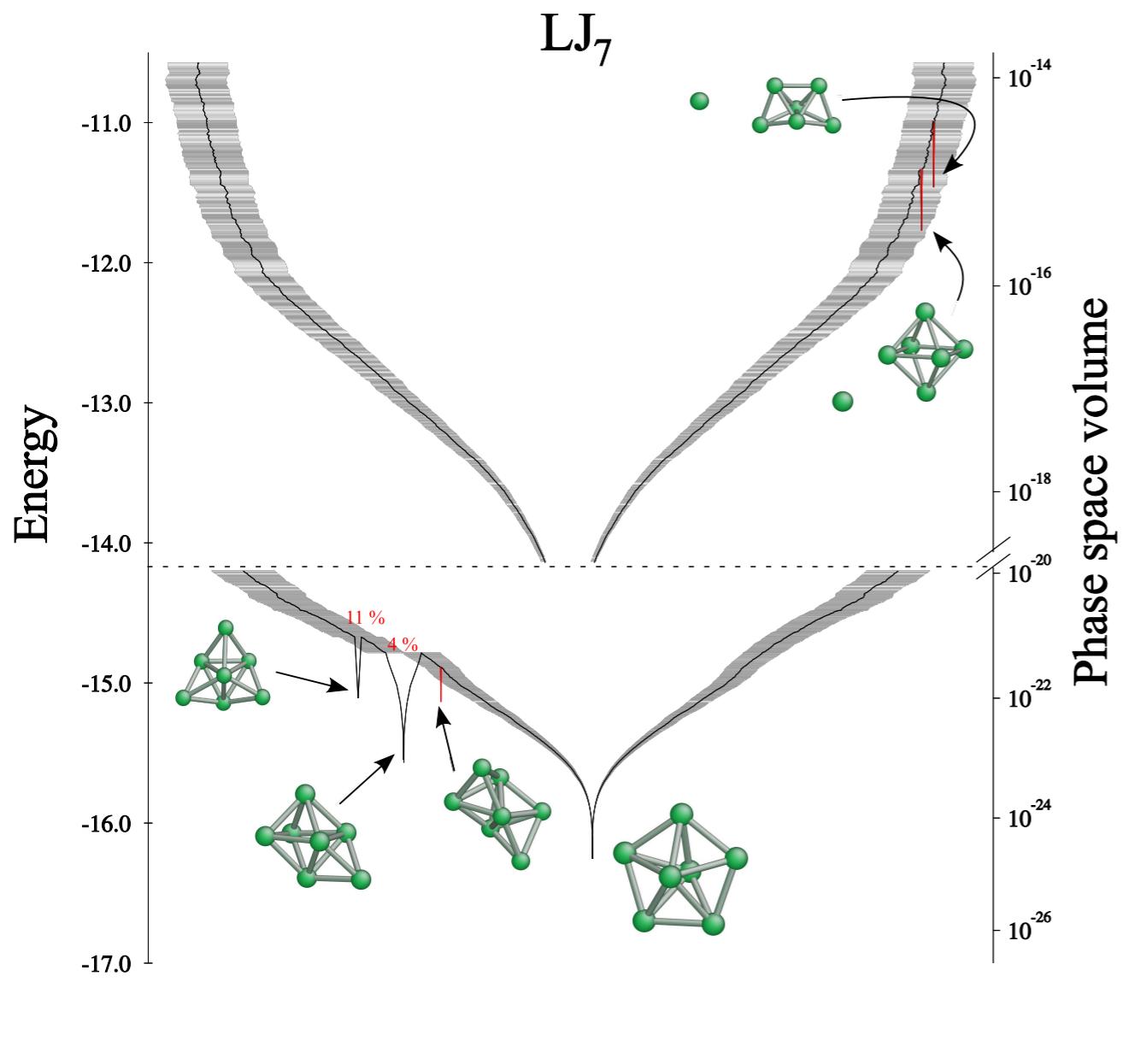
- Very fast to evaluate
- Complex energy landscape
- Rapidly growing number of local minima
- Standard test for structural optimisation and phase space search

# Example: Lennard-Jones cluster heat capacity



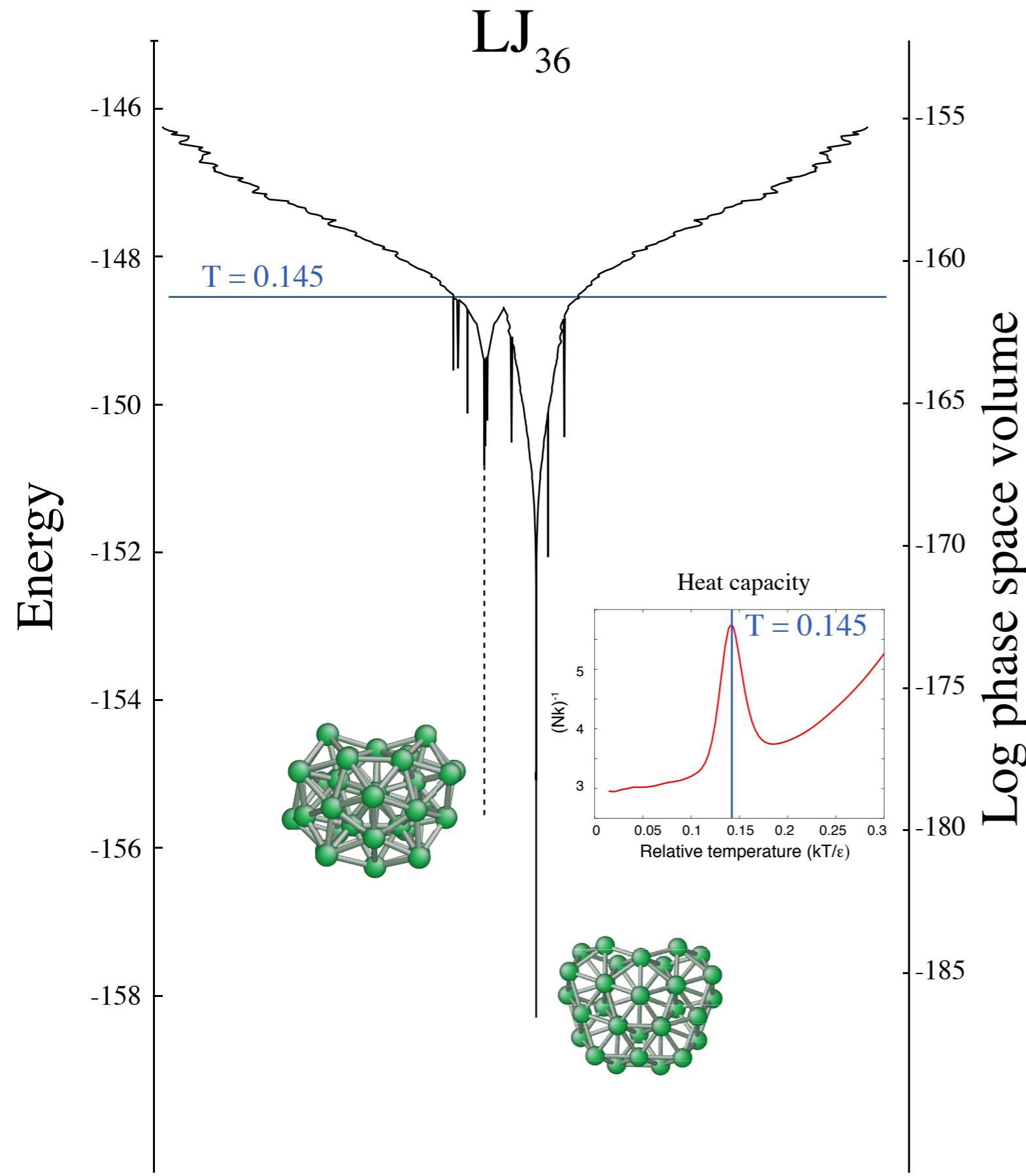
# Energy landscape charts



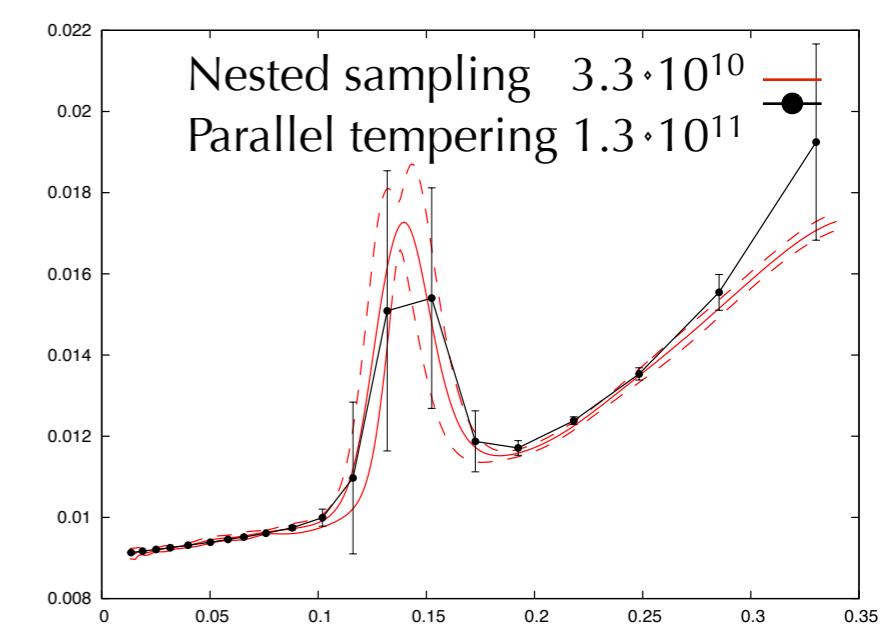


(exponential scaling of volume)

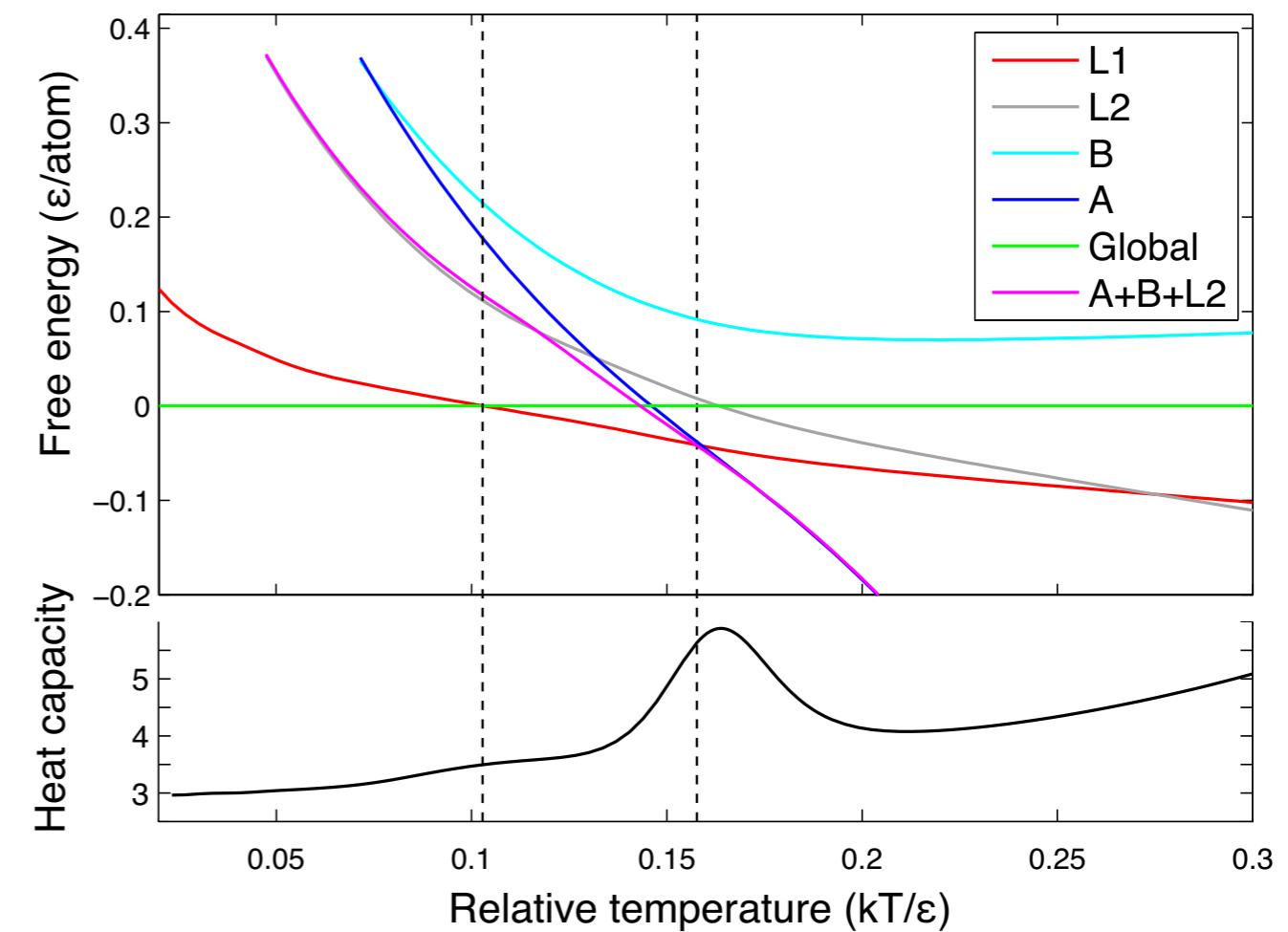
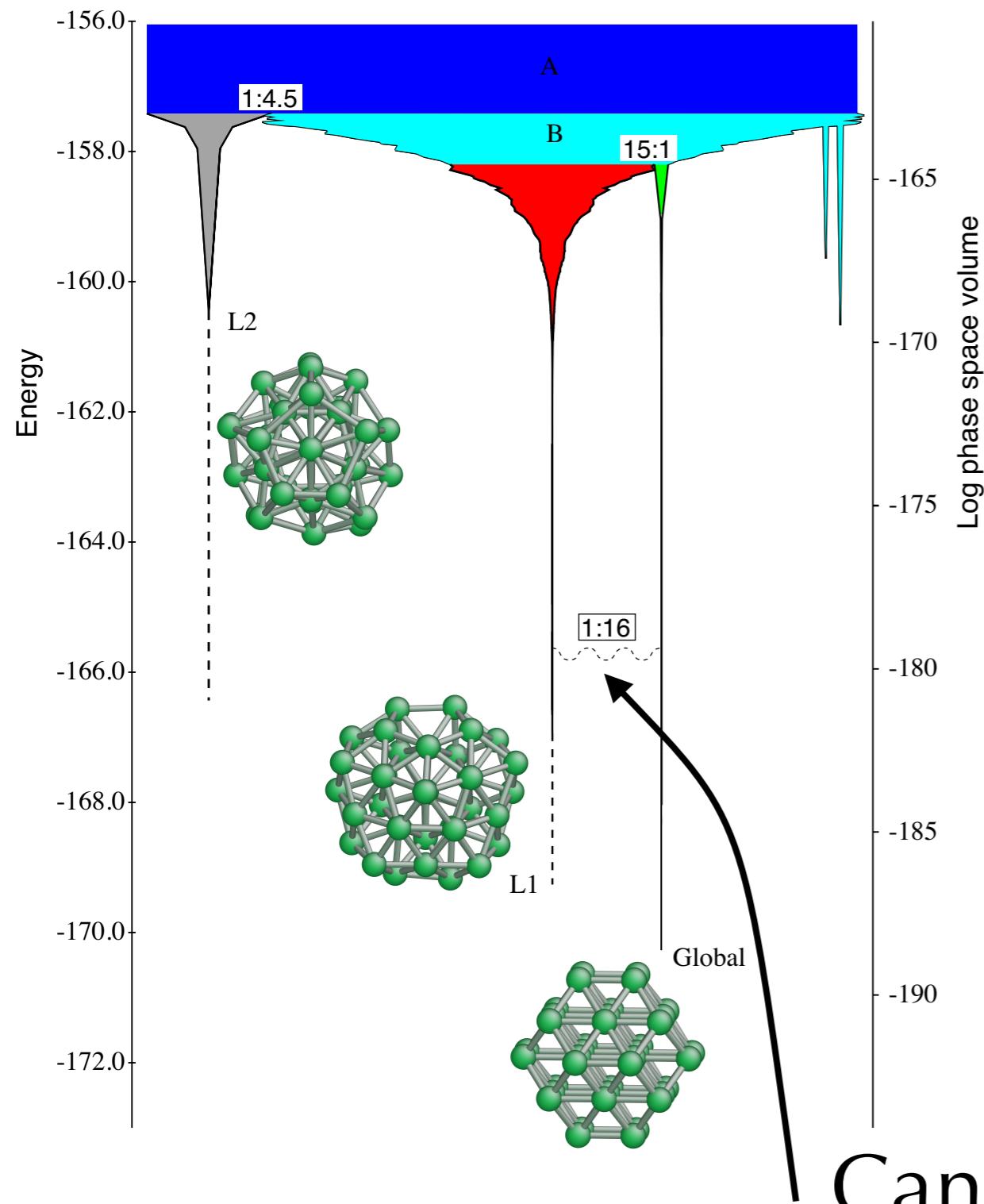
# LJ 36



LJ<sub>36</sub> Mackay/anti-Mackay



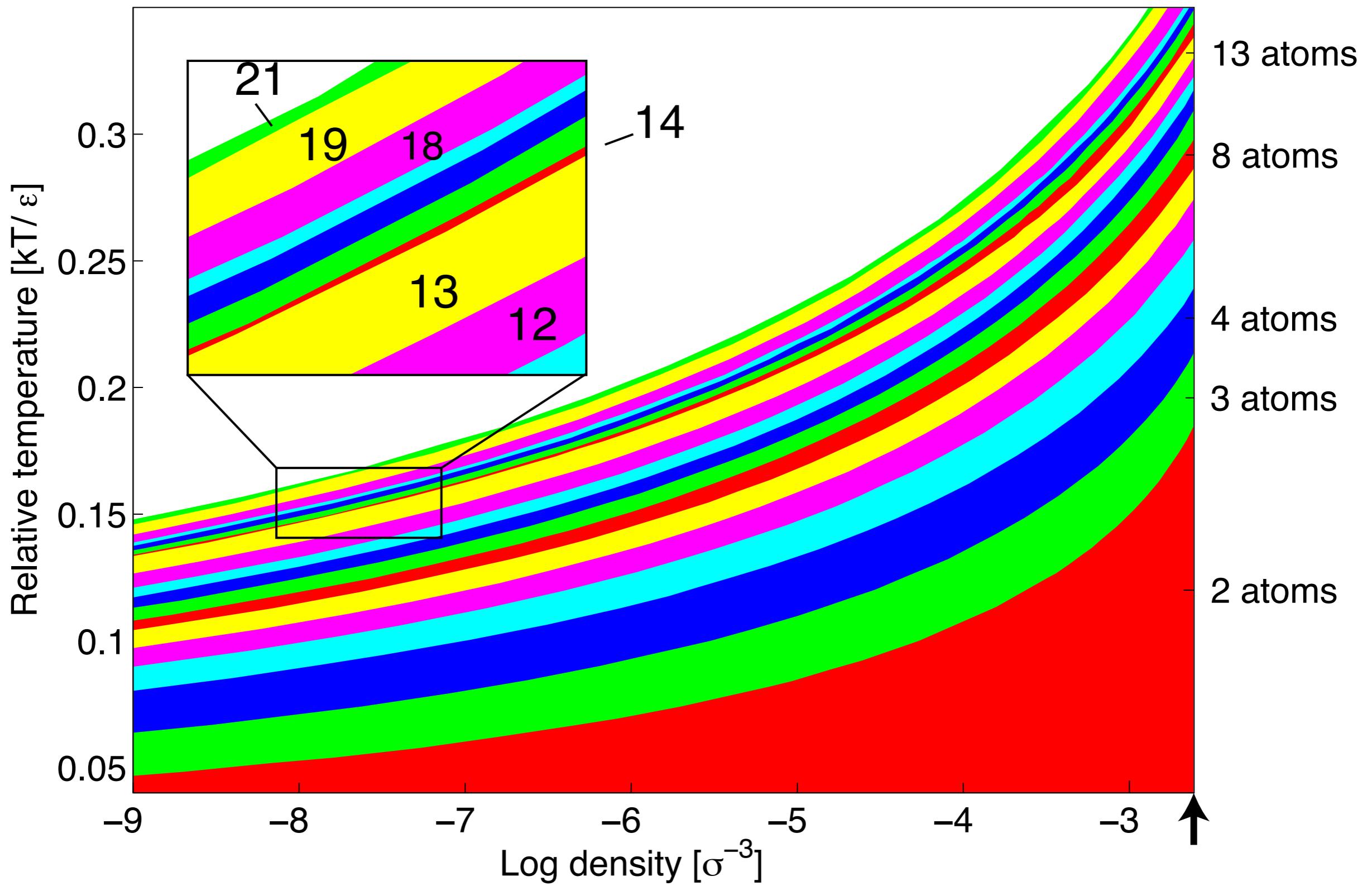
# LJ38



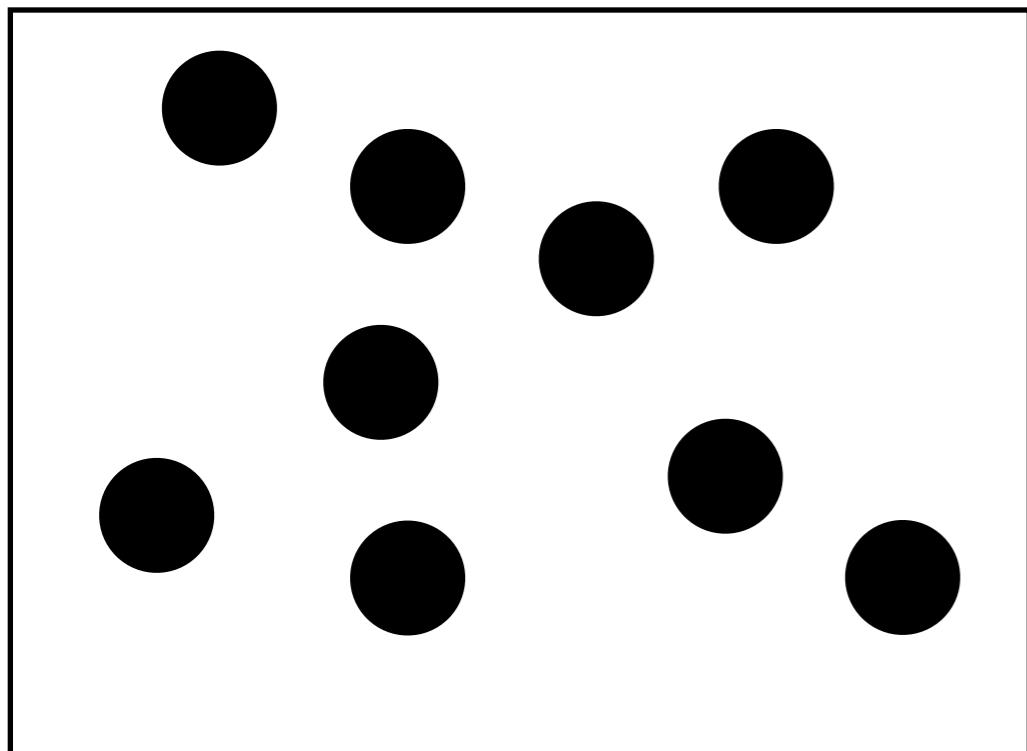
Candidate minimum energy path  
(Wales & Doye)

# Nanocluster phase diagram

Phase diagram of LJ clusters

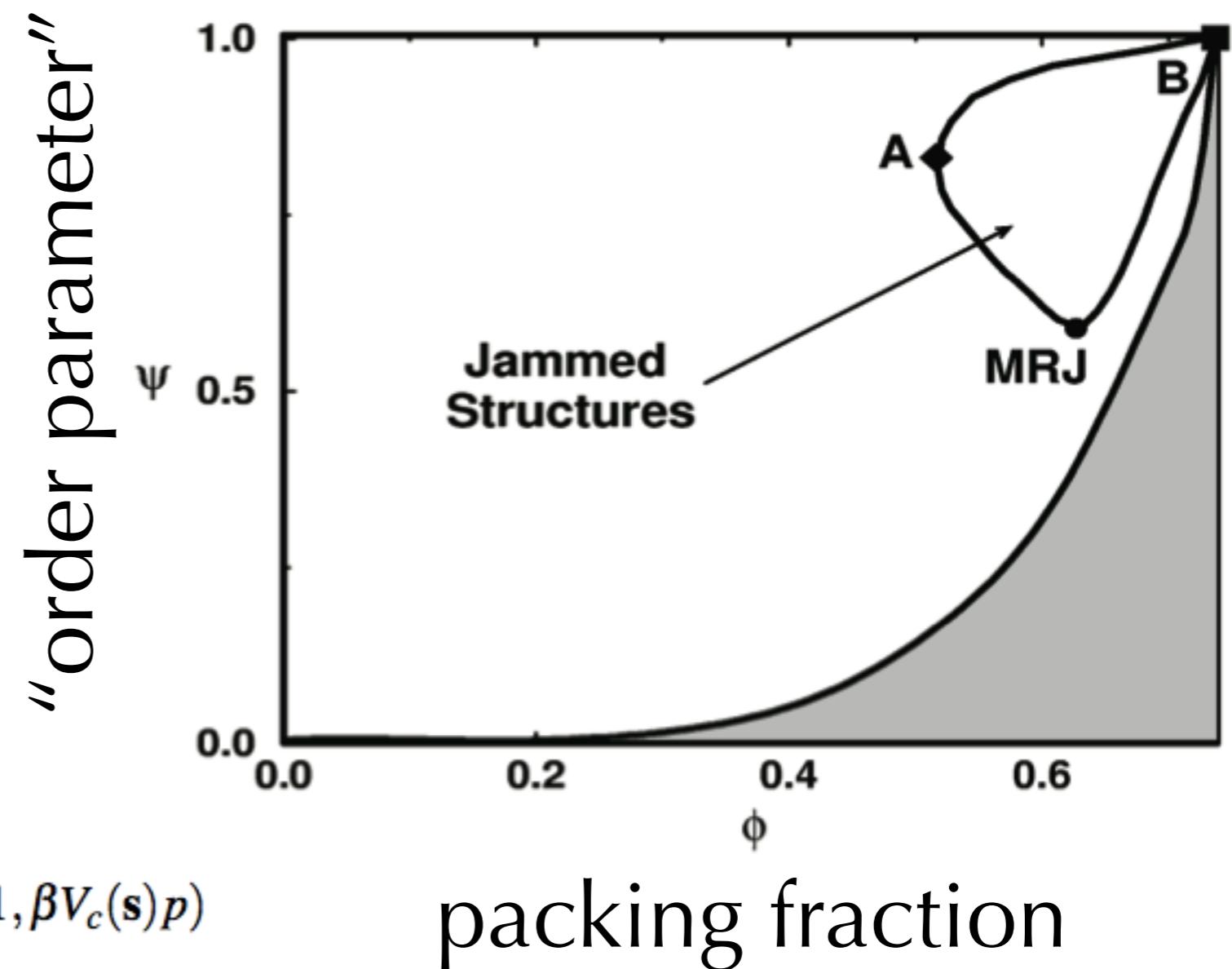


# Condensed phase: hard spheres

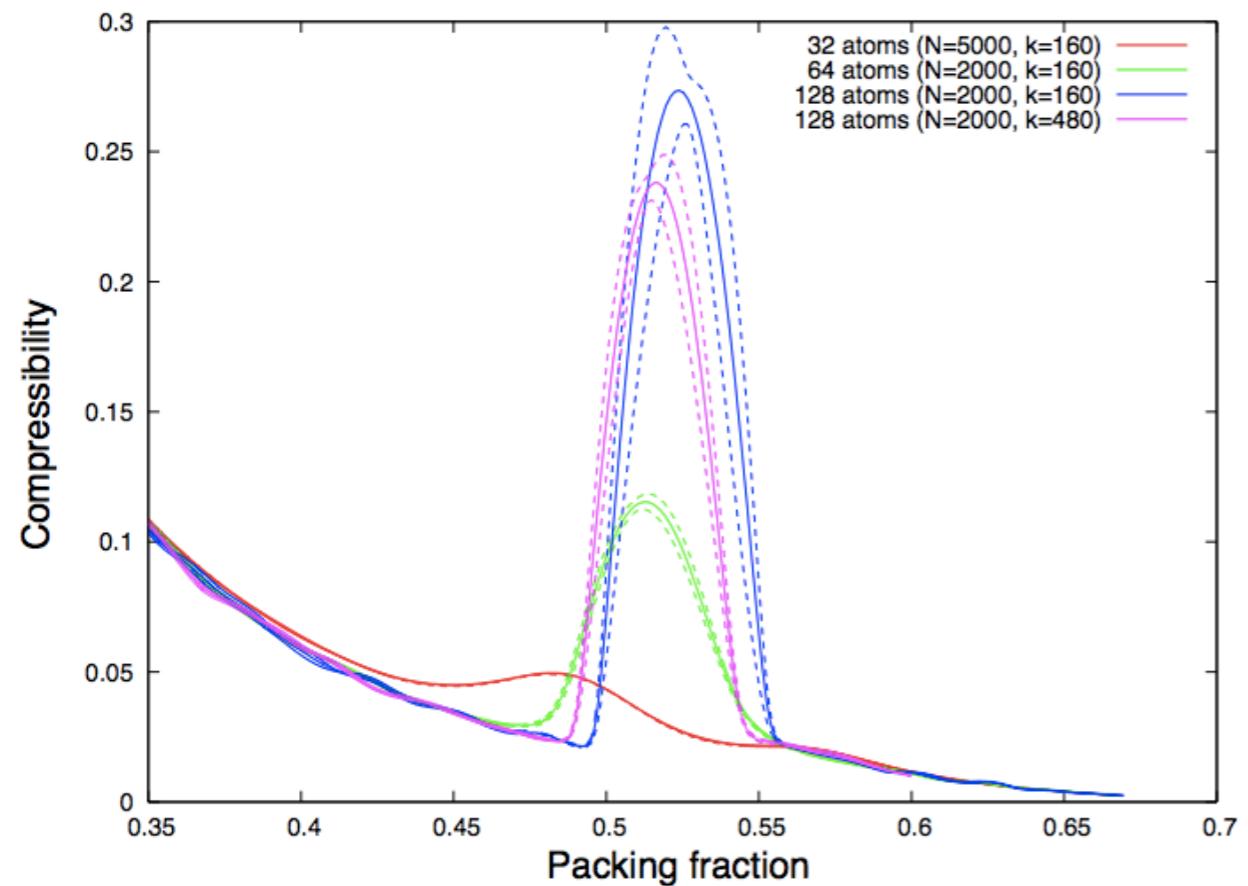


$$U(\mathbf{s}, V) = \begin{cases} 0 & \text{if } V \geq V_c(\mathbf{s}) \\ \infty & \text{if } V < V_c(\mathbf{s}) \end{cases}$$

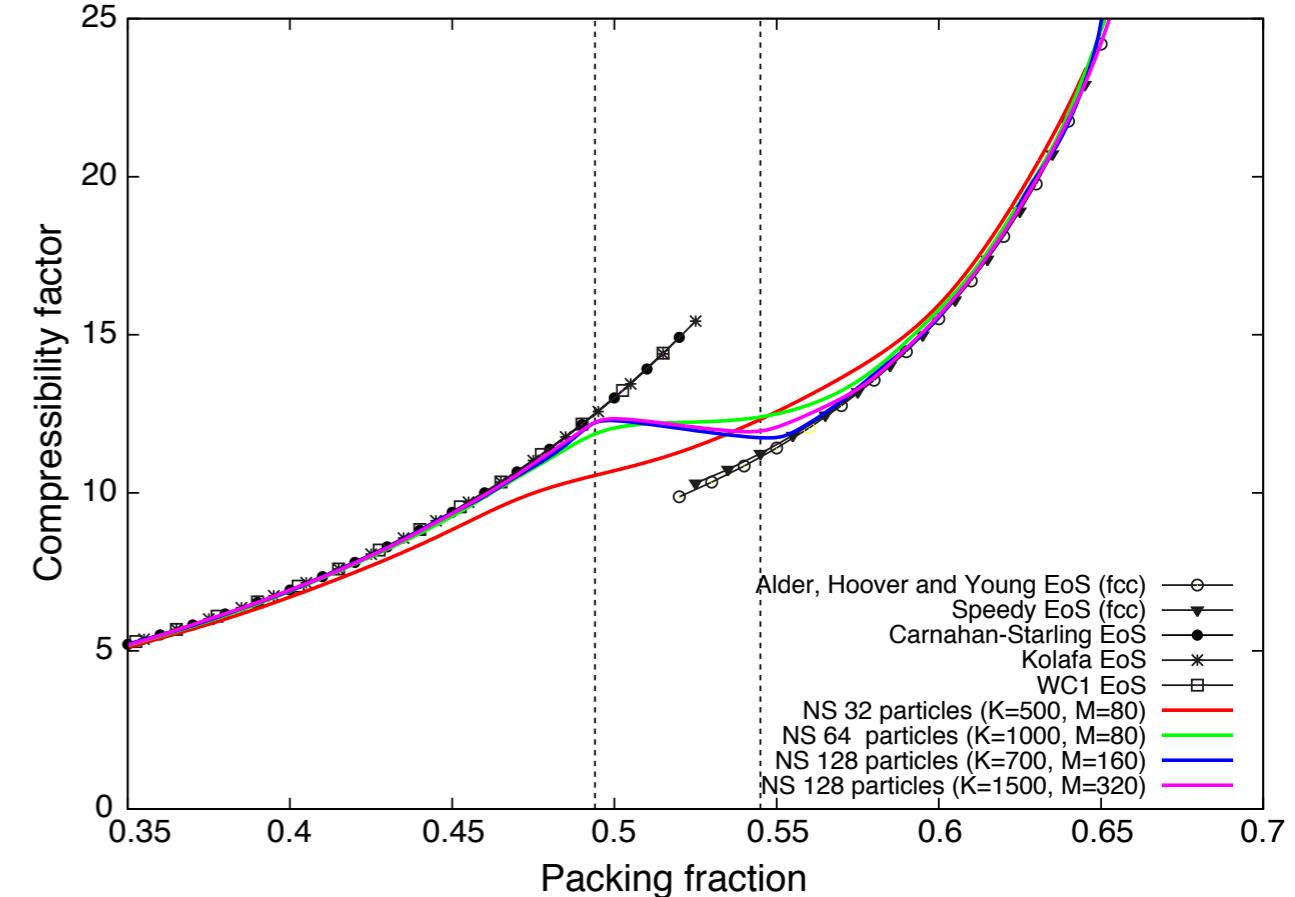
$$\Delta(\beta, p, N) = \frac{1}{\beta^N p^N} Z_p(\beta) \int_{(0,1)^{3N}} d\mathbf{s} \Gamma(N+1, \beta V_c(\mathbf{s}) p)$$



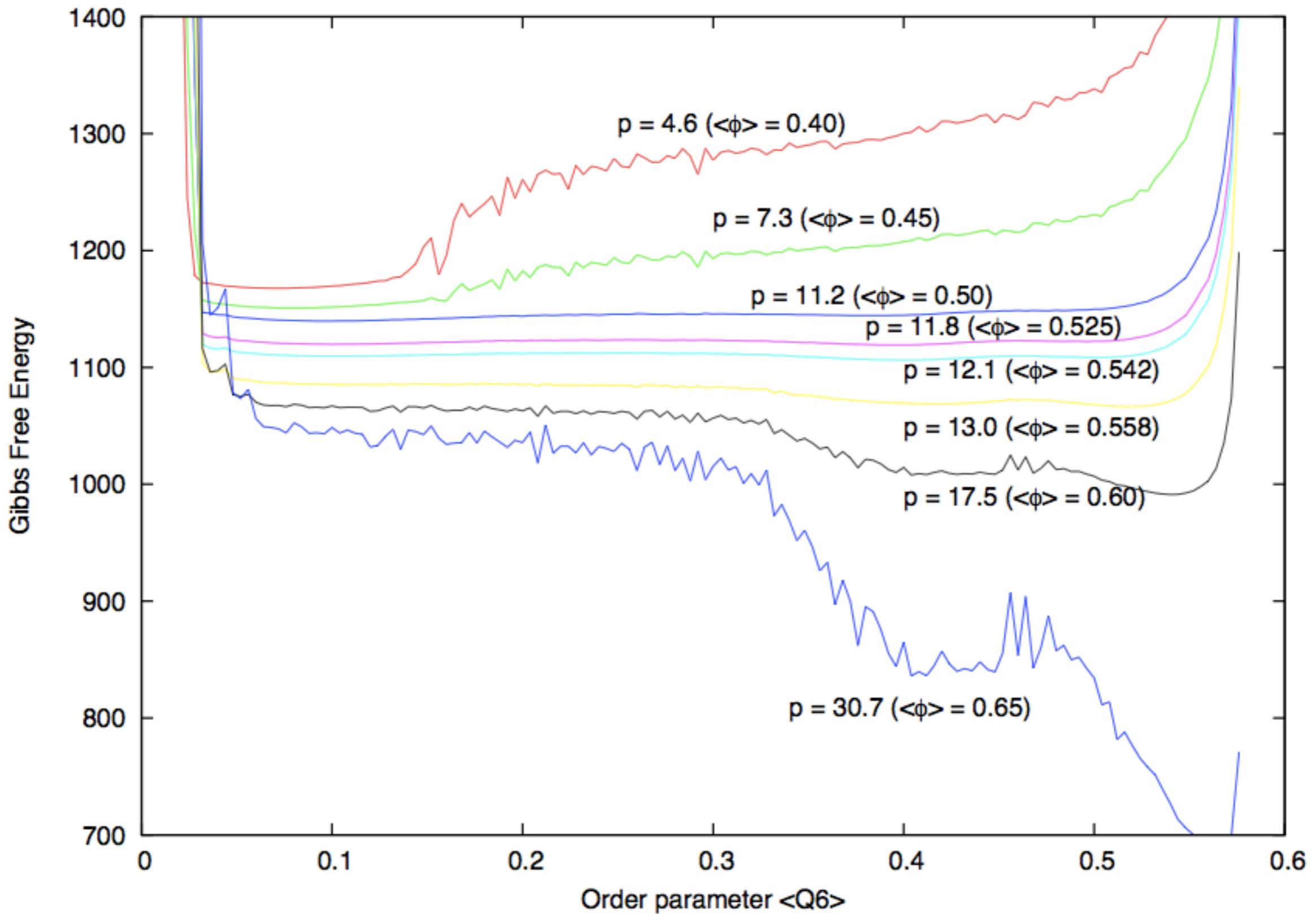
# Compressibility



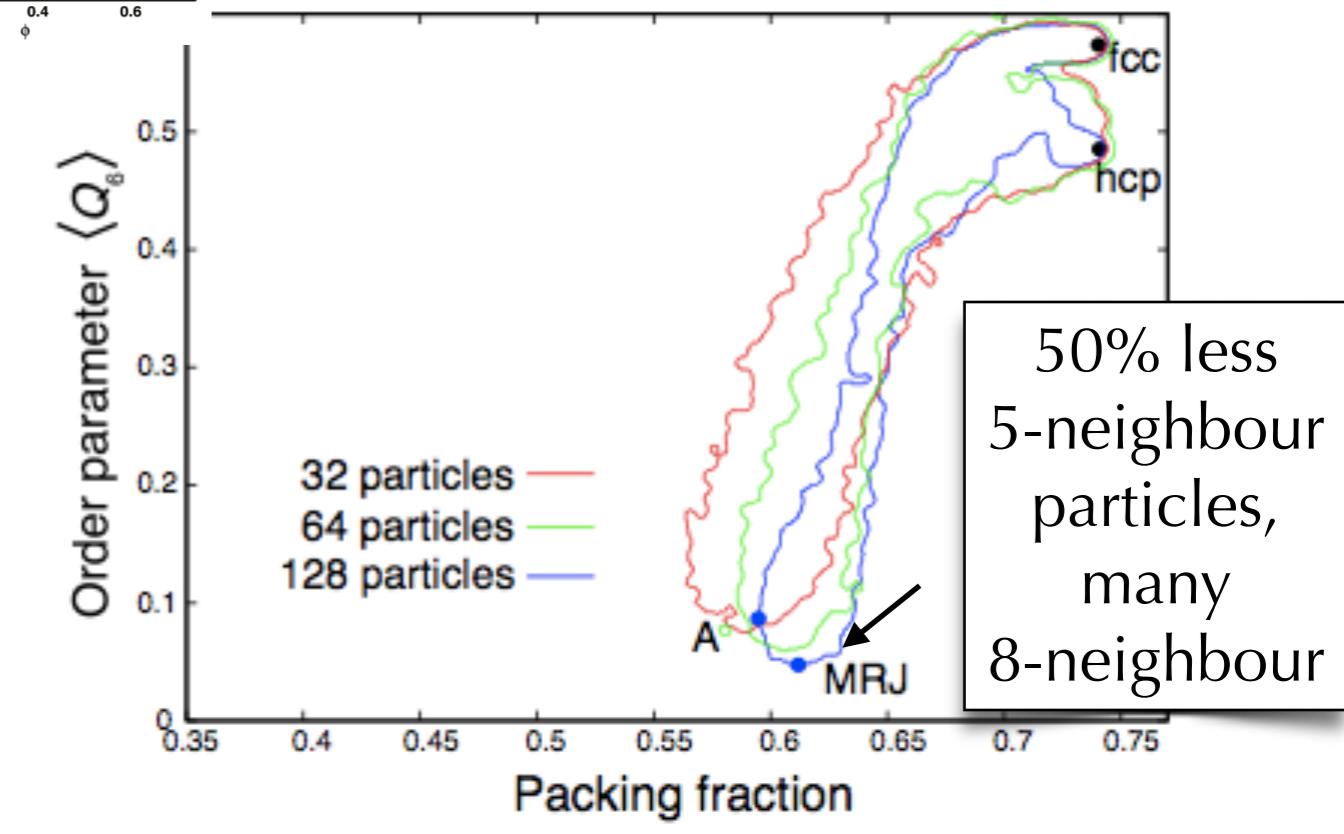
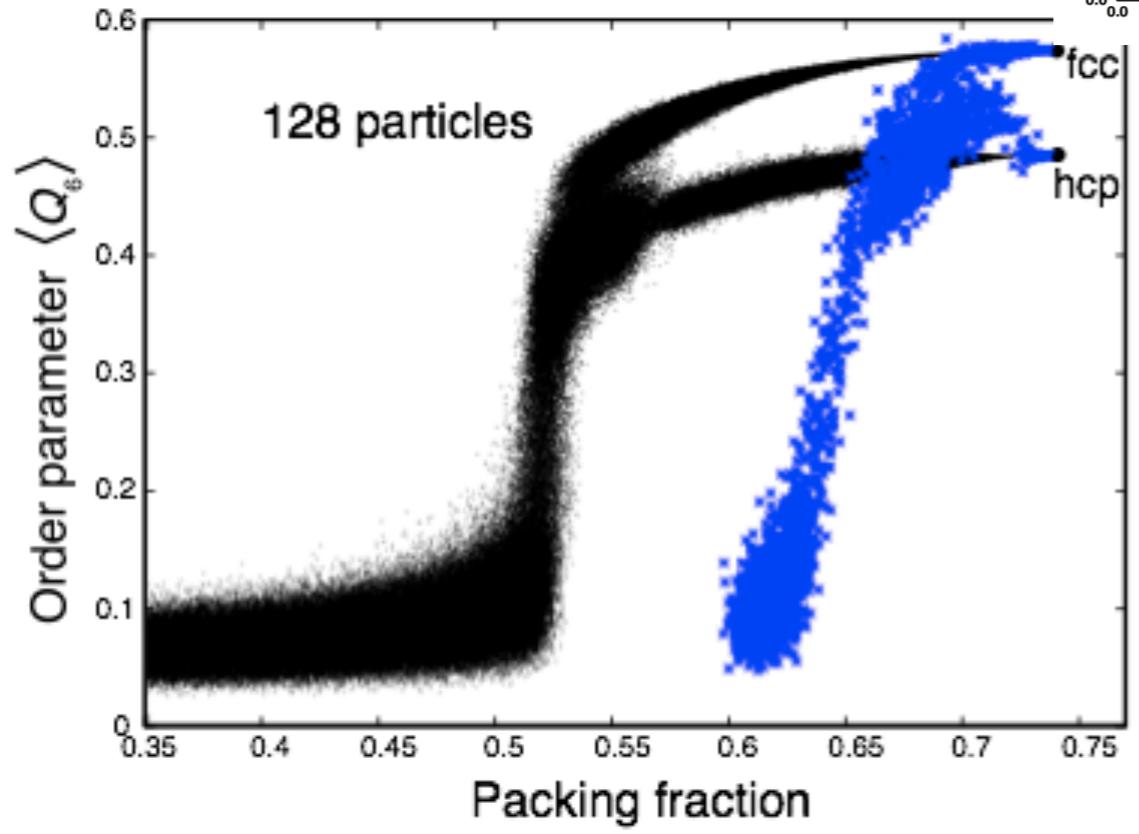
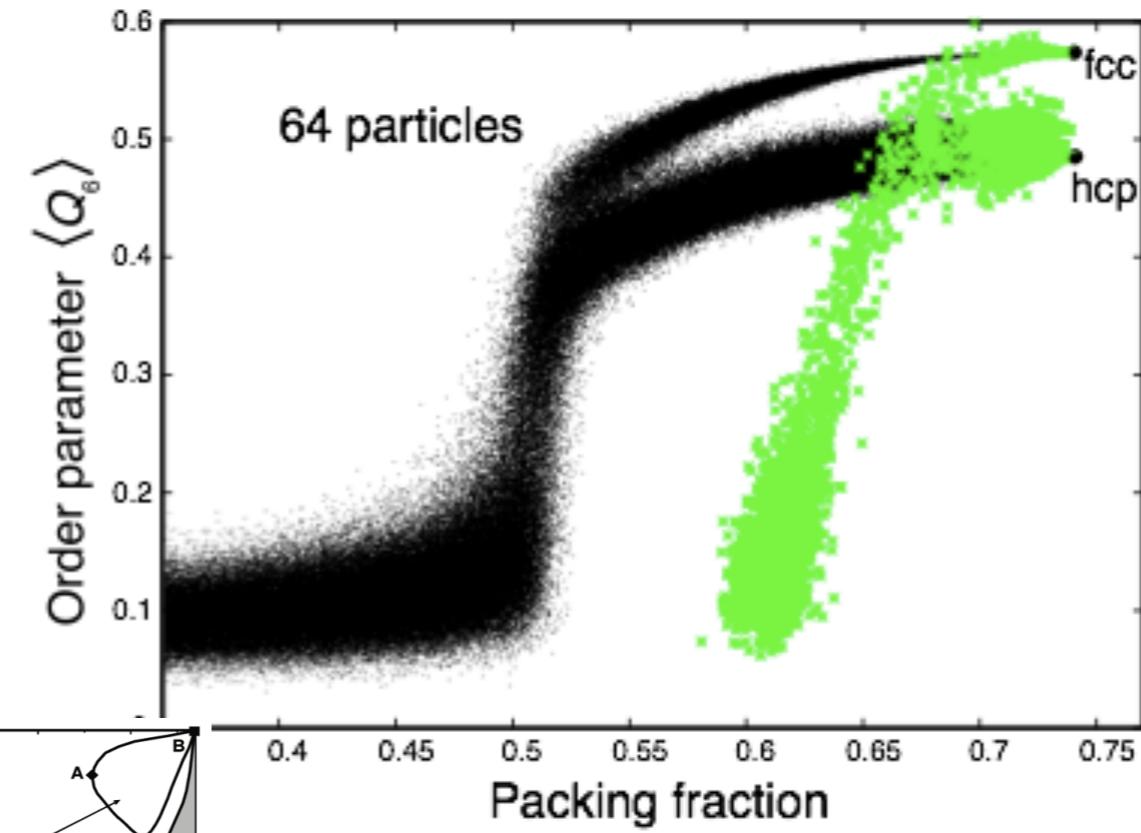
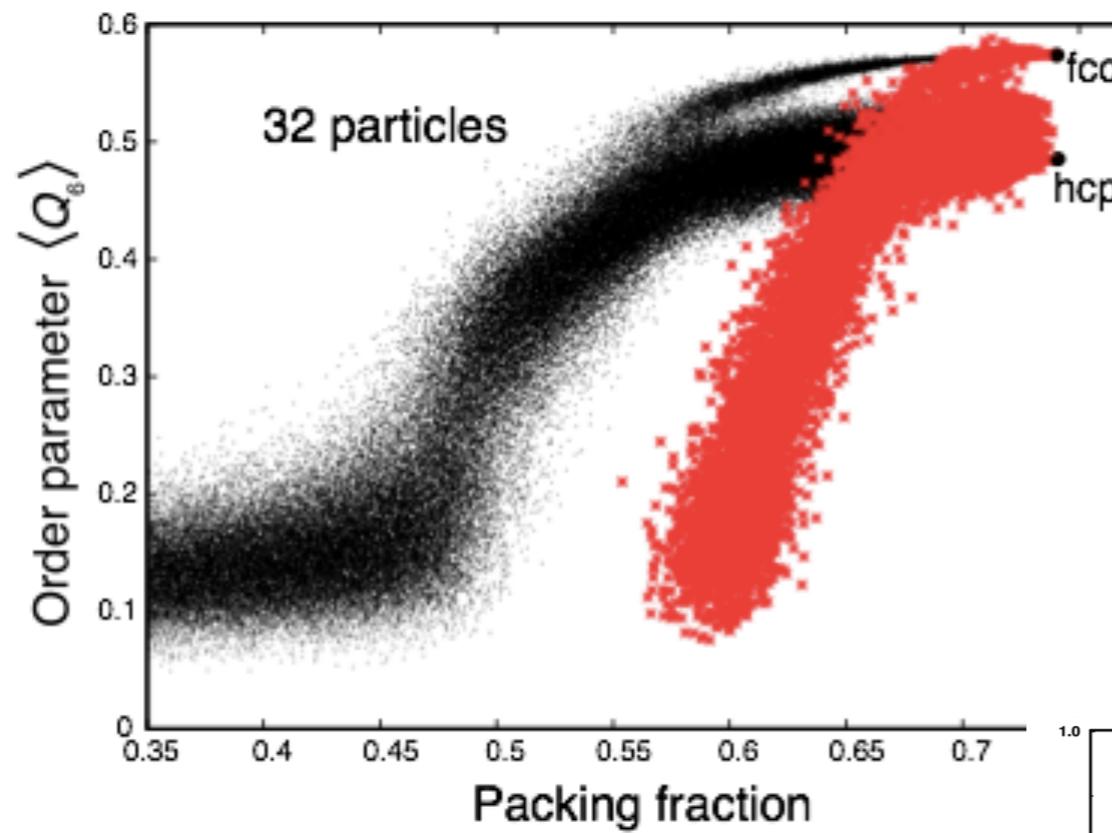
# Equation of state



# Free energy vs. Q6



# Jammed states



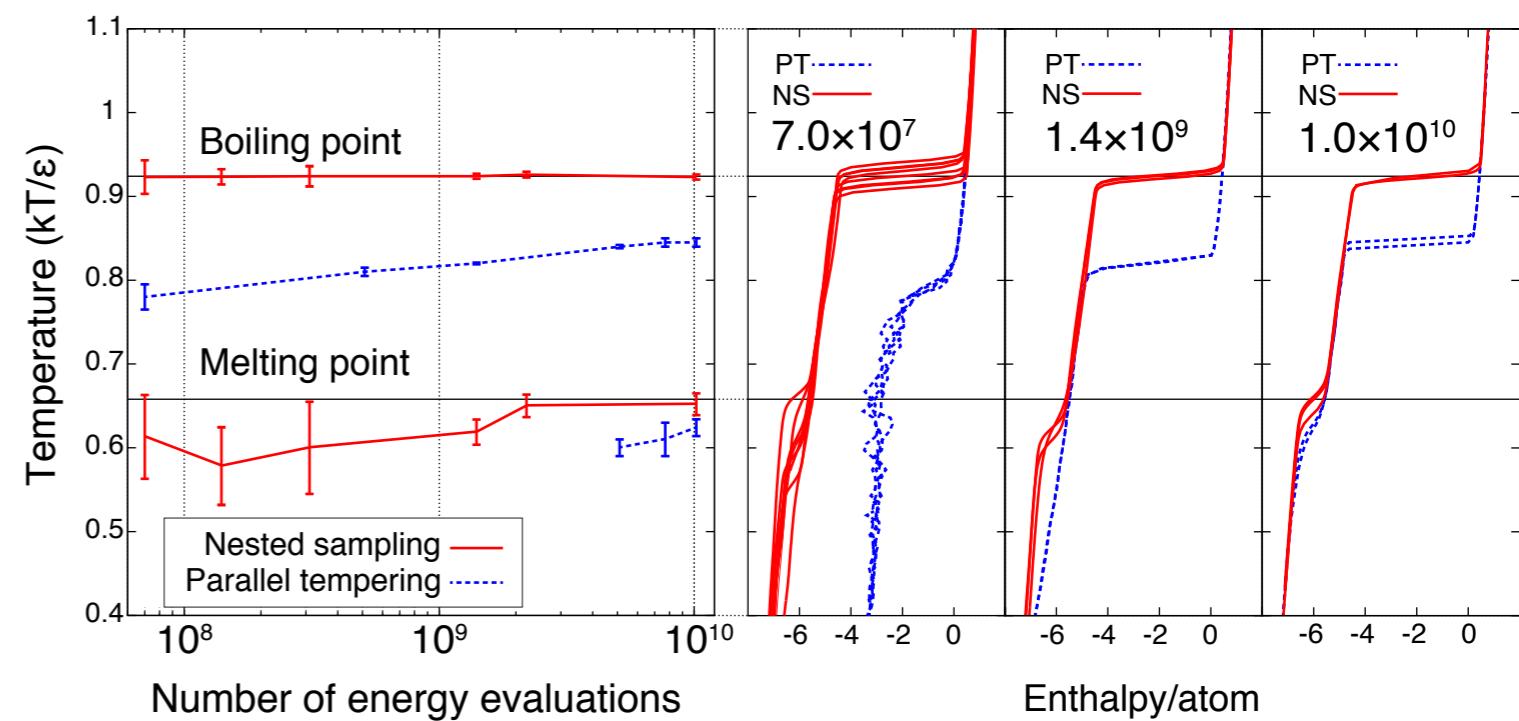
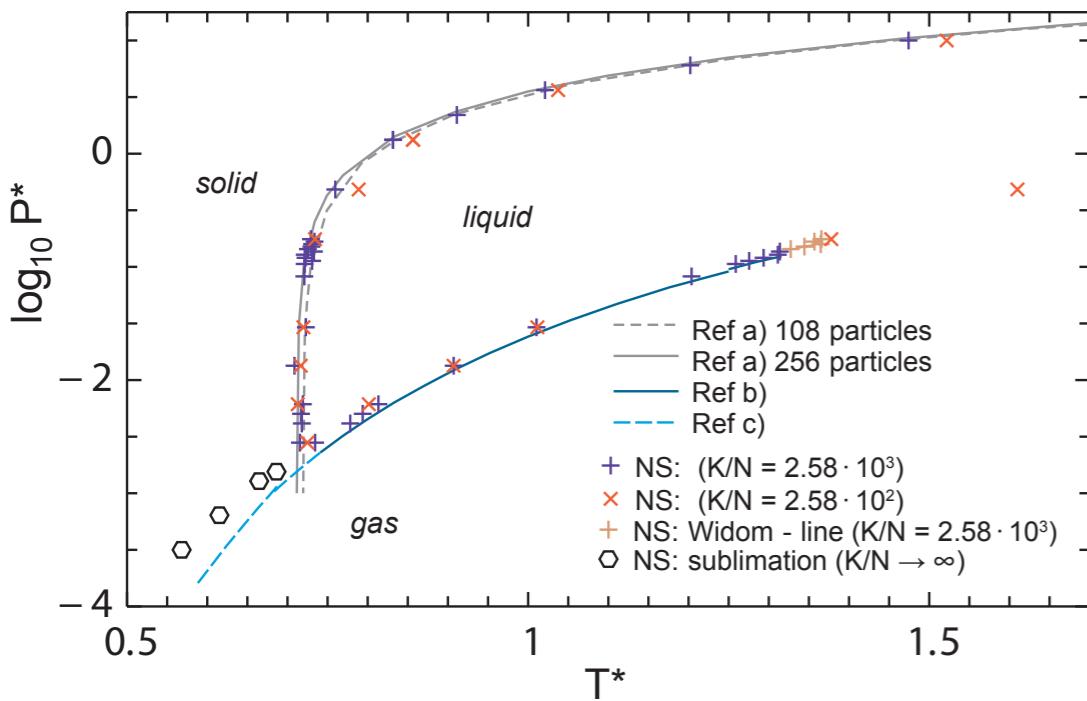
# Condensed phase, constant pressure

Fractional coordinates  $\mathbf{s}$  and lattice vector matrix  $\mathbf{h}$ :  $\mathbf{q} = \mathbf{hs}$      $V = |\mathbf{h}|$      $\mathbf{h}_0 = \mathbf{h}V^{-1/3}$

Enthalpy:  $H(p, \mathbf{s}, V, \mathbf{h}_0) = E(\mathbf{s}, V, \mathbf{h}_0) + pV$

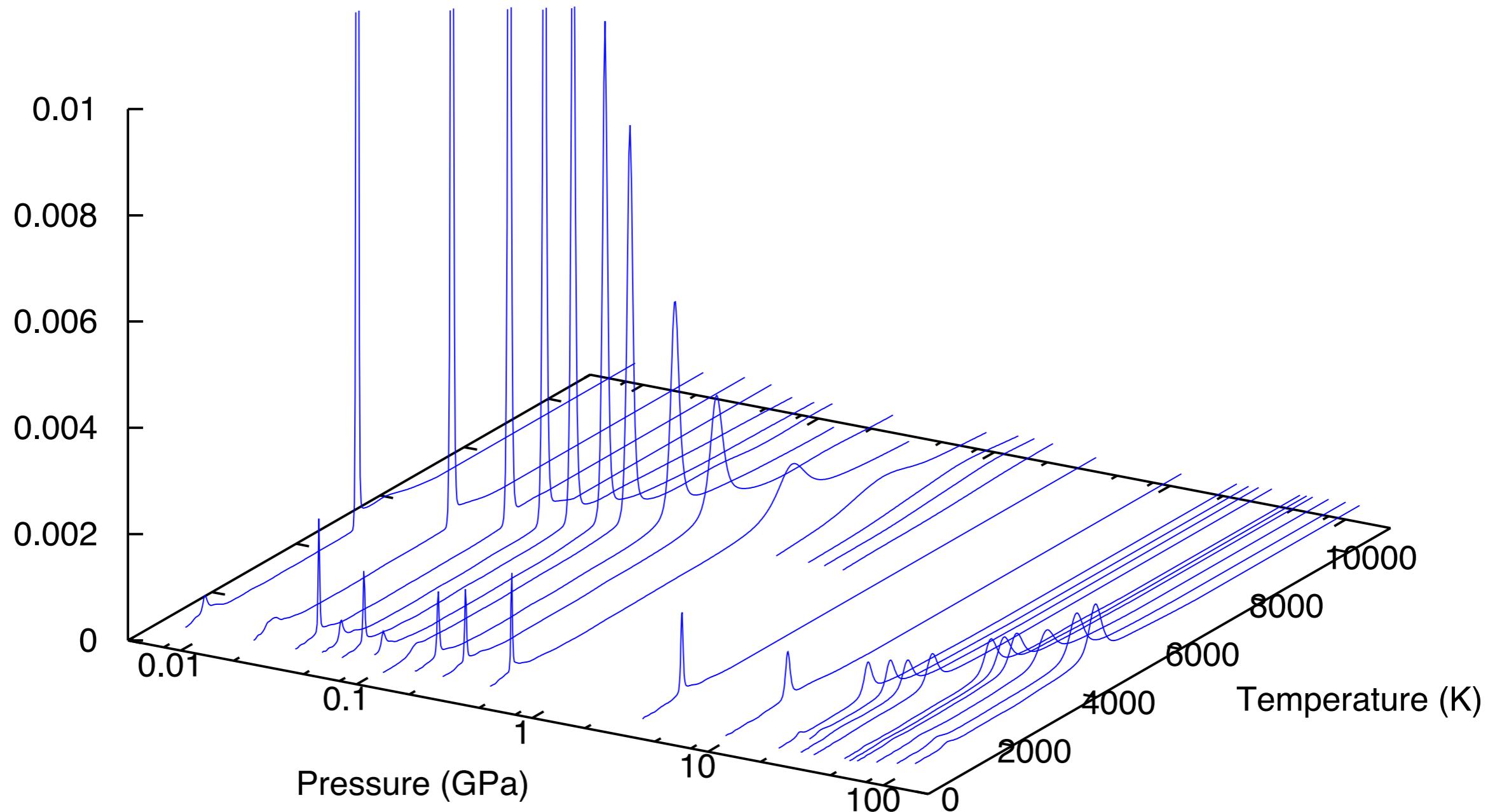
$$\begin{aligned}\Delta(N, p, \beta) &= Z_m \beta p \int_0^\infty dV V^N \int d\mathbf{h}_0 \delta(|\mathbf{h}_0| - 1) \times \int d\mathbf{s} e^{-\beta(E(\mathbf{s}, V, \mathbf{h}_0) + pV)} \\ &= \frac{V_0^{N+1}}{N+1} \sum_i (\chi_{i-1} - \chi_i) e^{-\beta H_i} + \frac{\Gamma(N+1, \beta p V_0)}{(\beta p)^N}\end{aligned}$$

## Periodic Lennard-Jones system, 64 particles



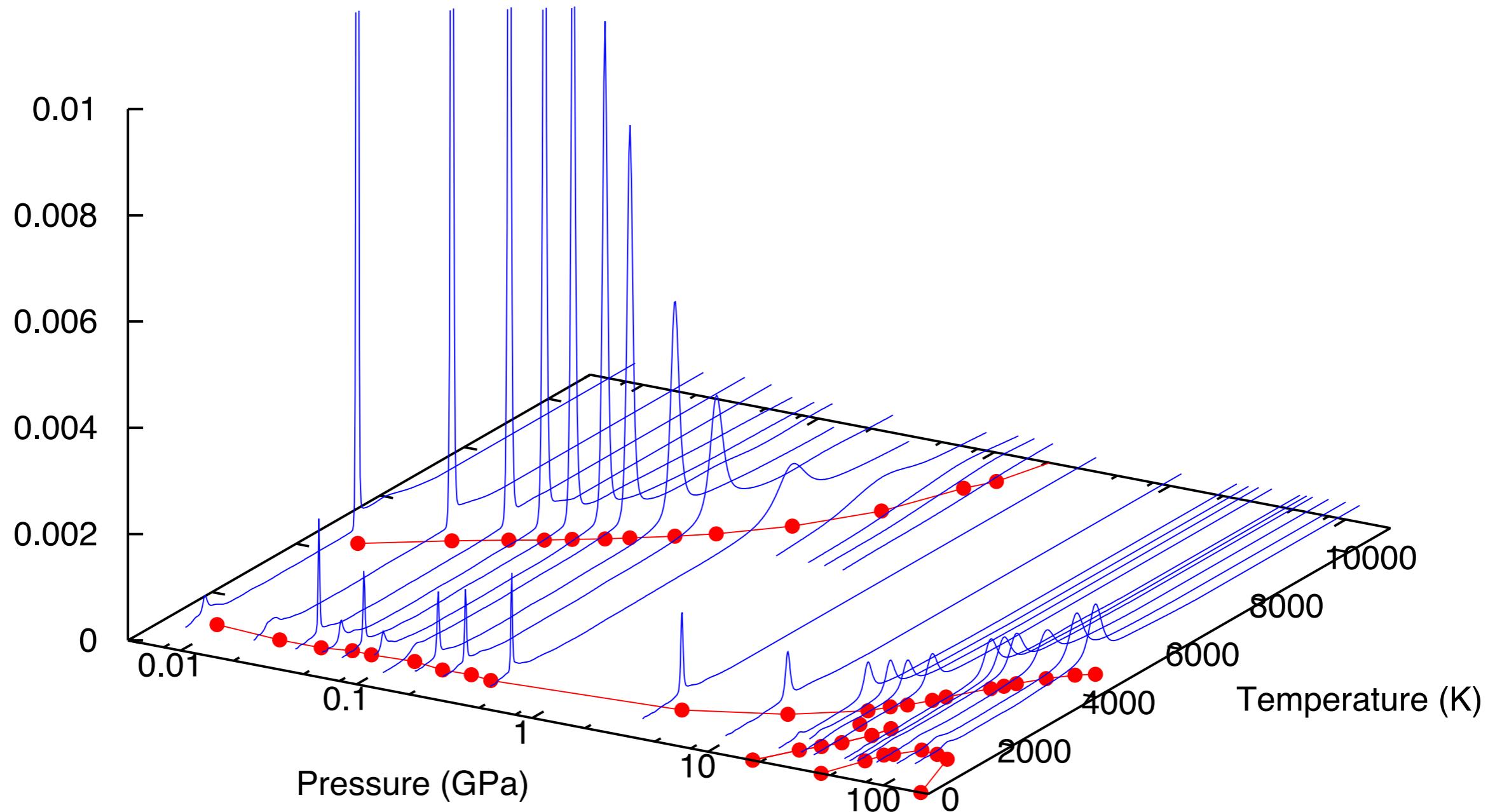
# Compute experimental observables:

## Heat capacity of EAM aluminium

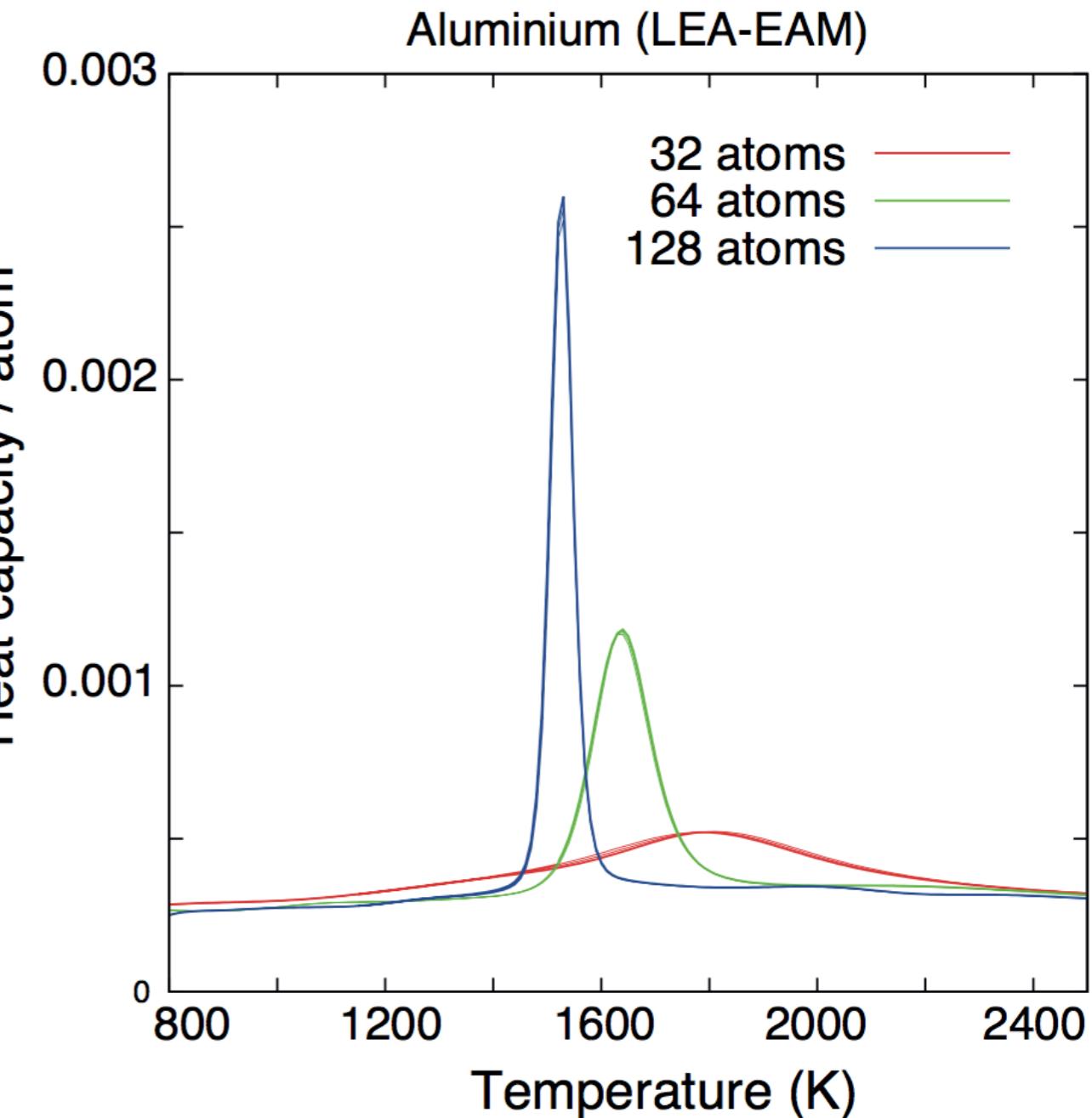
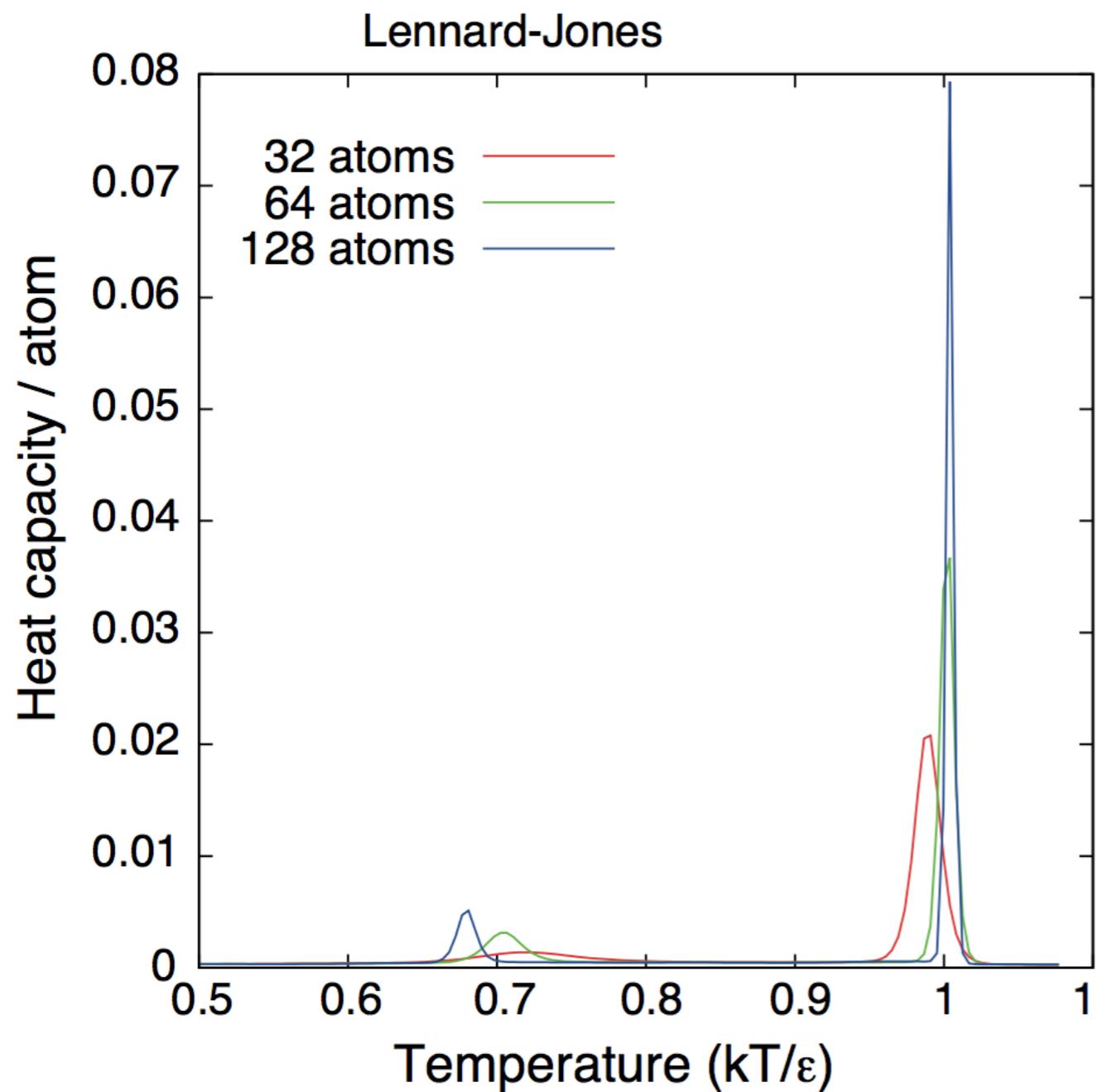


# Compute experimental observables:

## Heat capacity of EAM aluminium

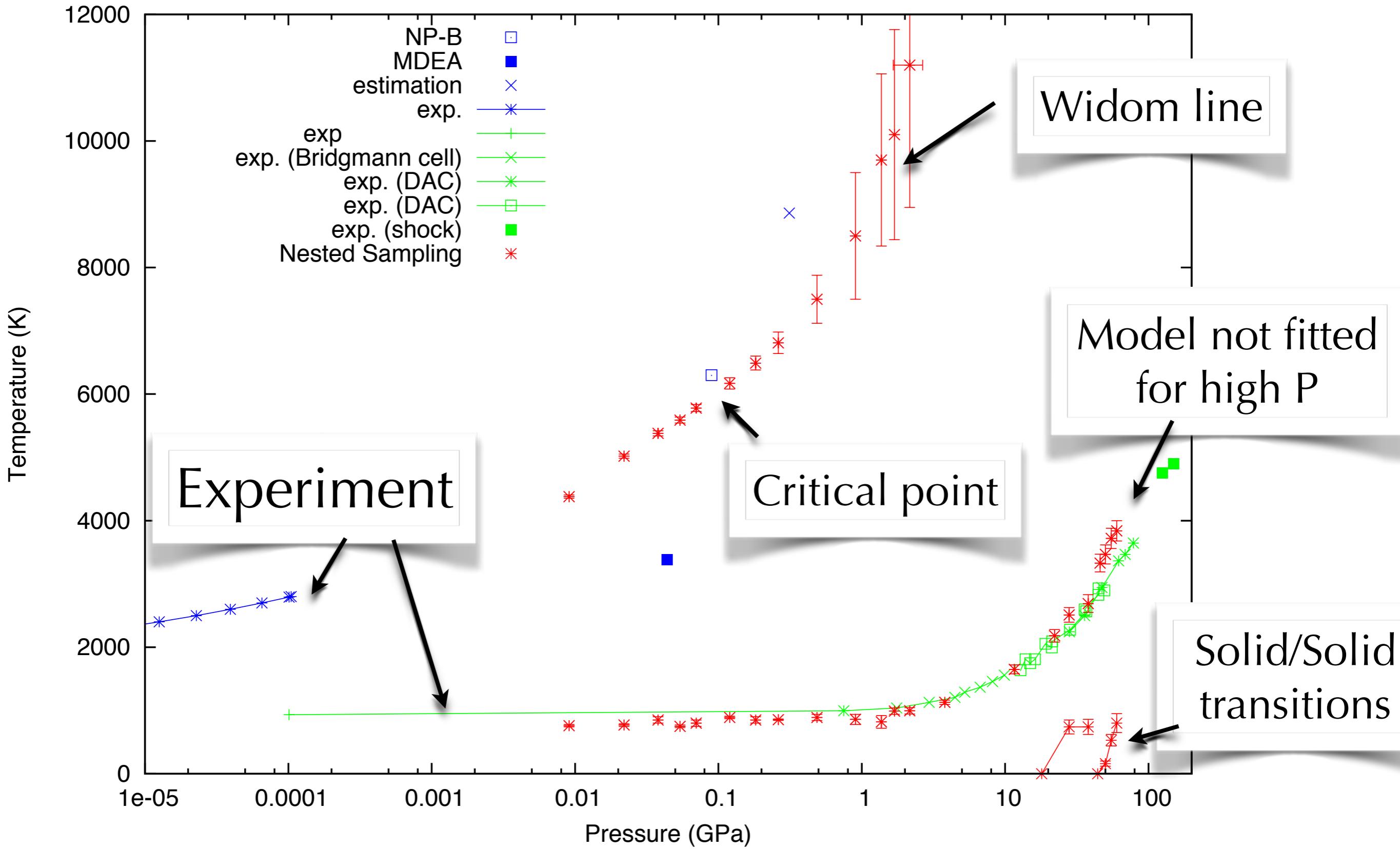


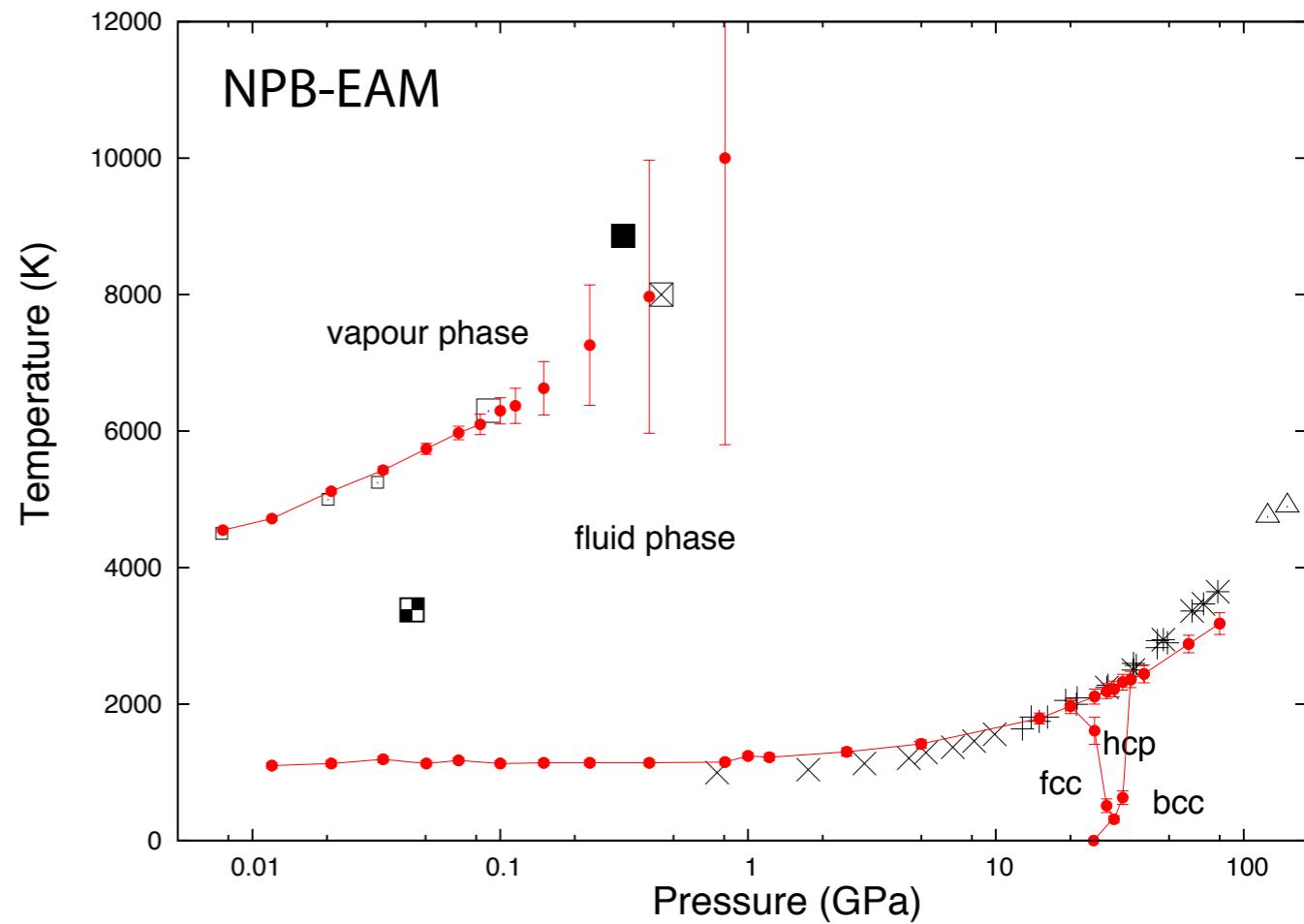
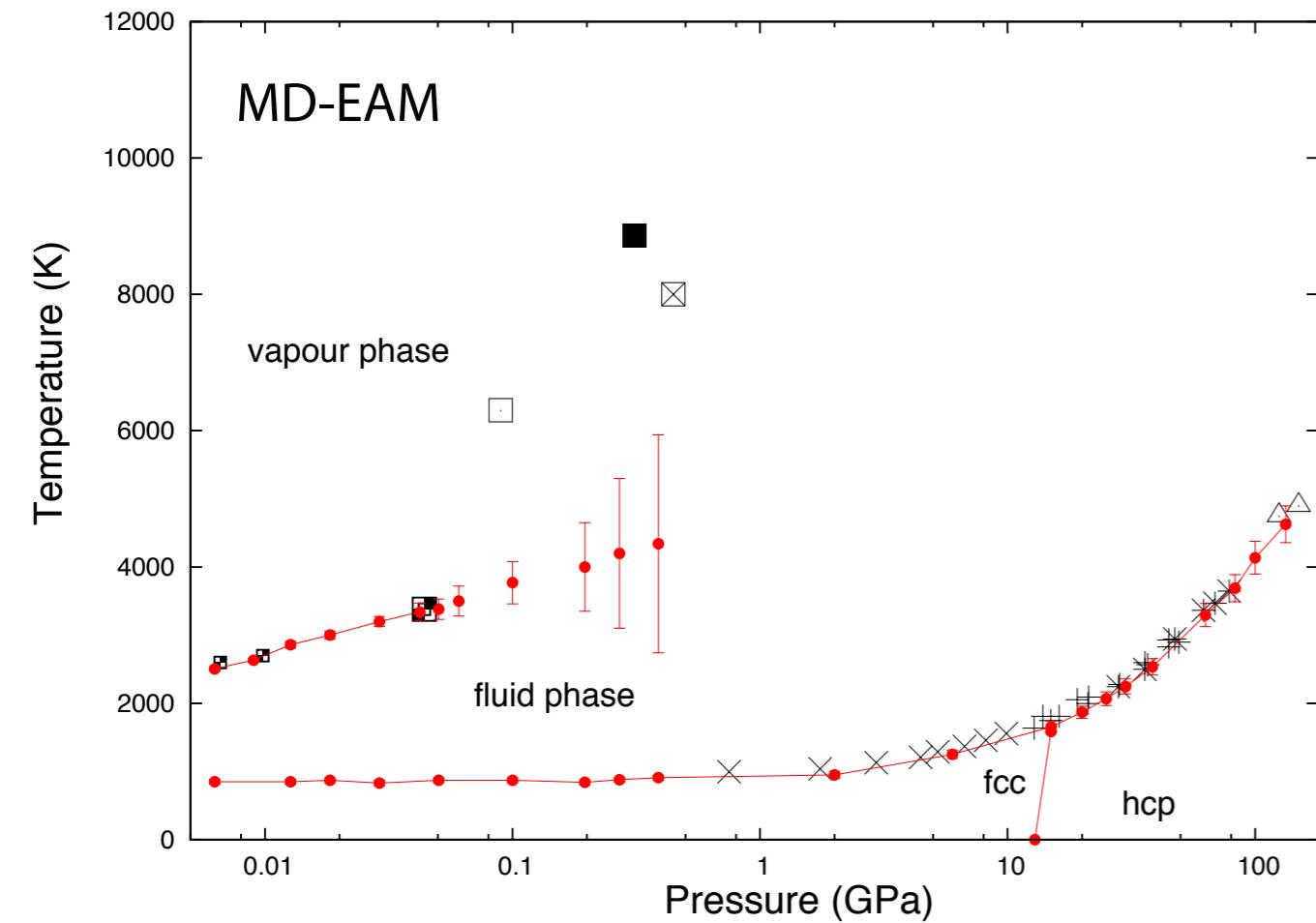
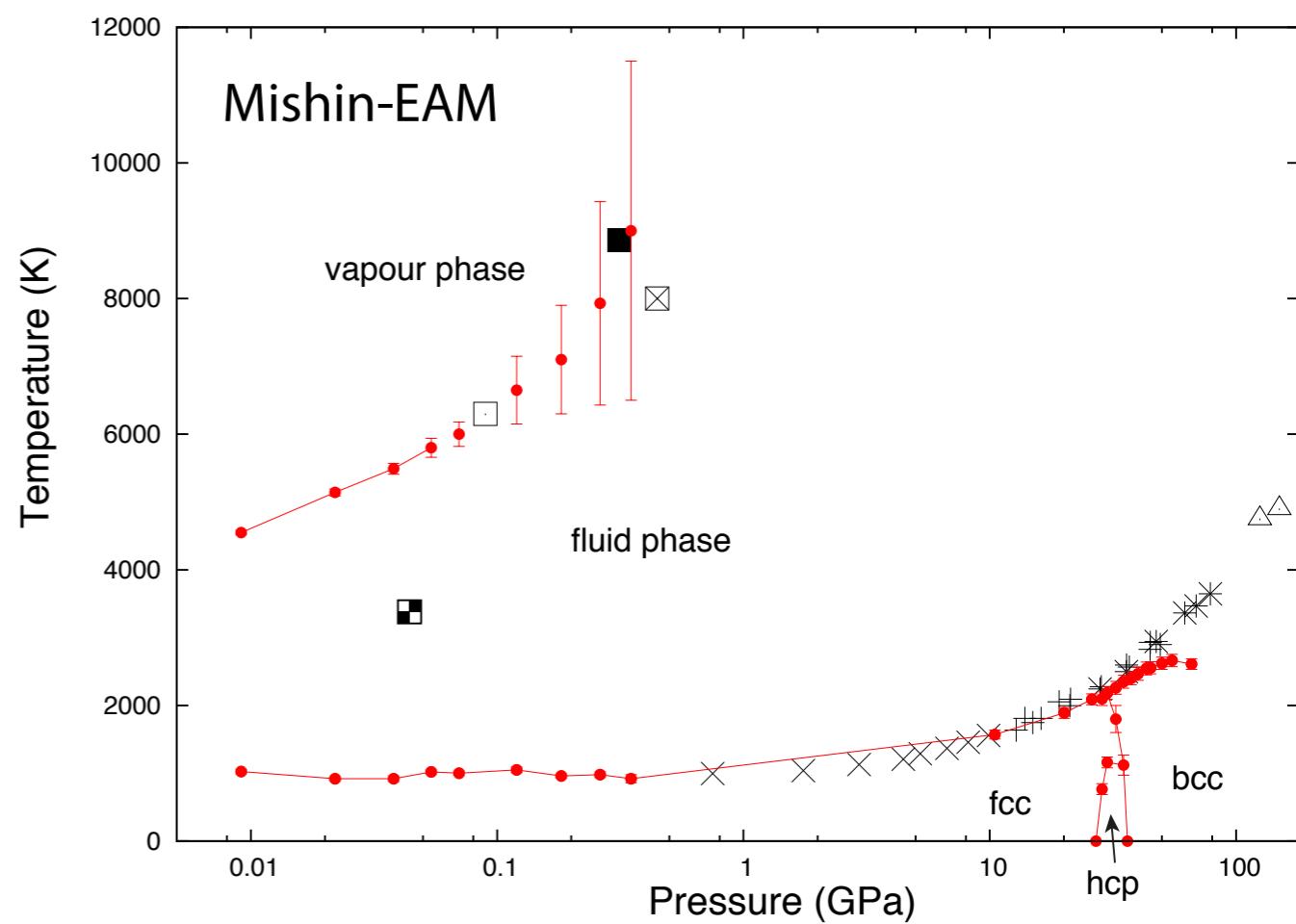
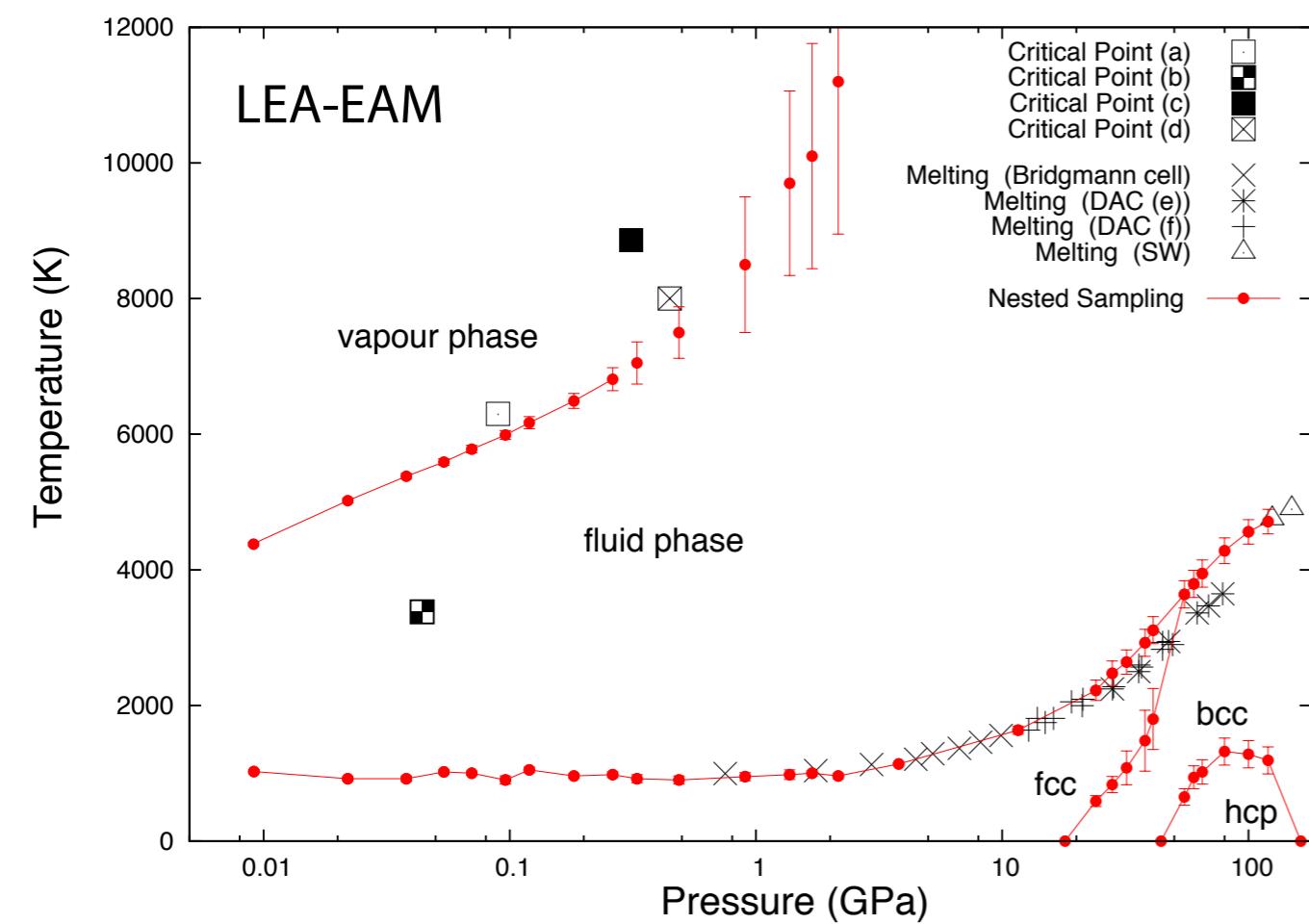
# Finite size effects



In normal MD simulations of phase coexistence, there are **interfaces**, leading to much larger finite size effects

# Phase diagram of EAM aluminium, $10^9$ calls

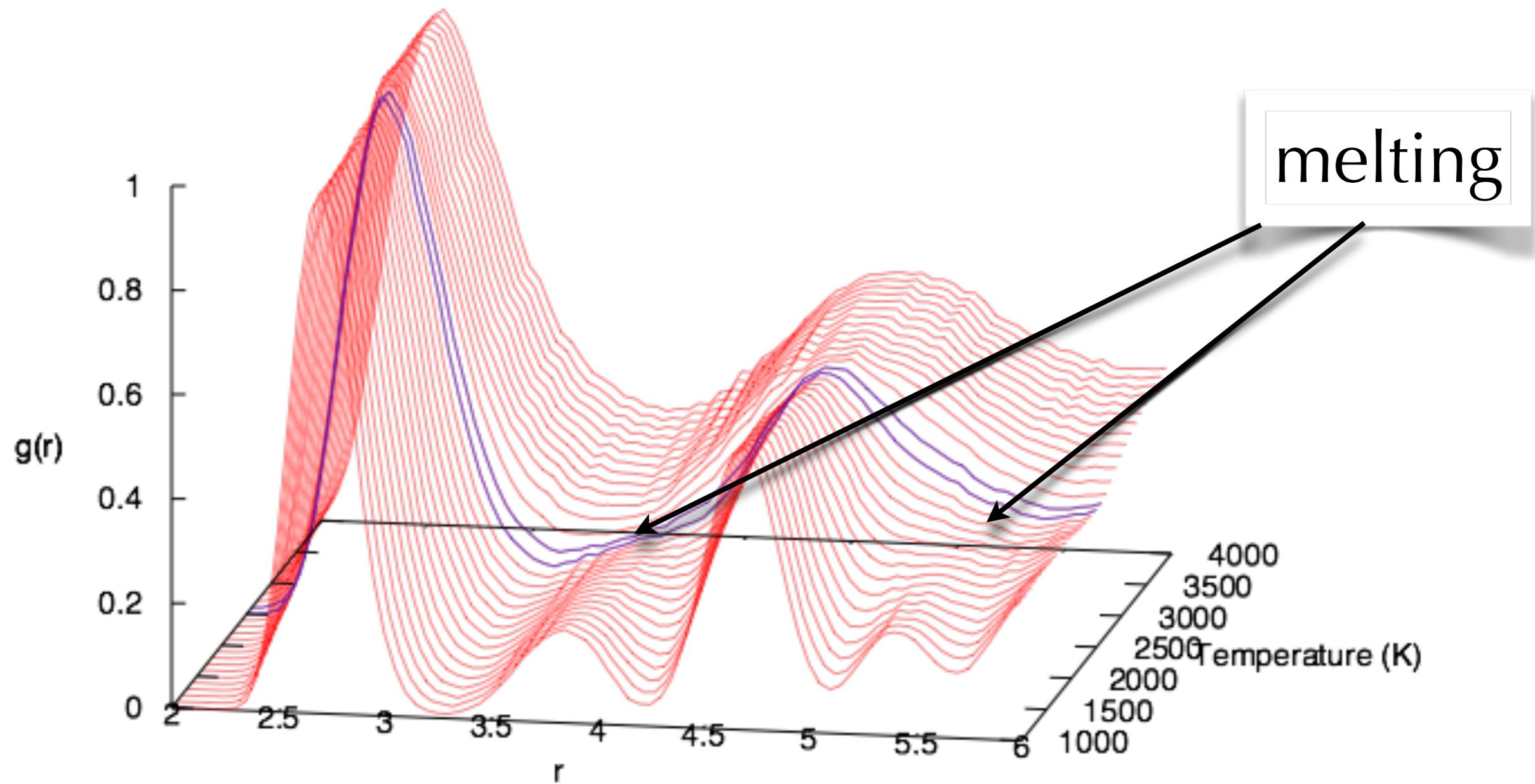




# Radial distribution function

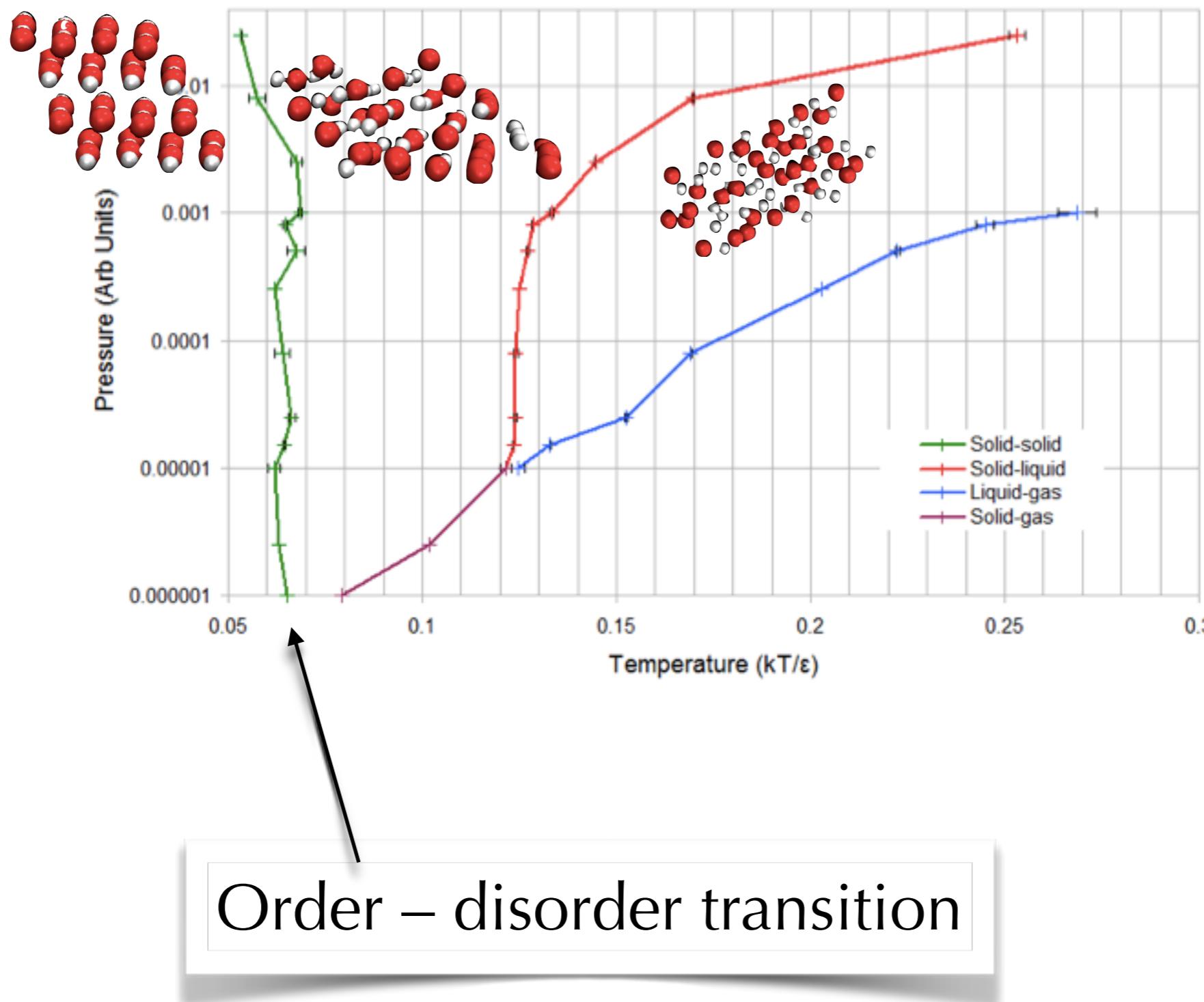
Samples suitable for Z ought to yield all interesting observables

$P = 38 \text{ GPa}$

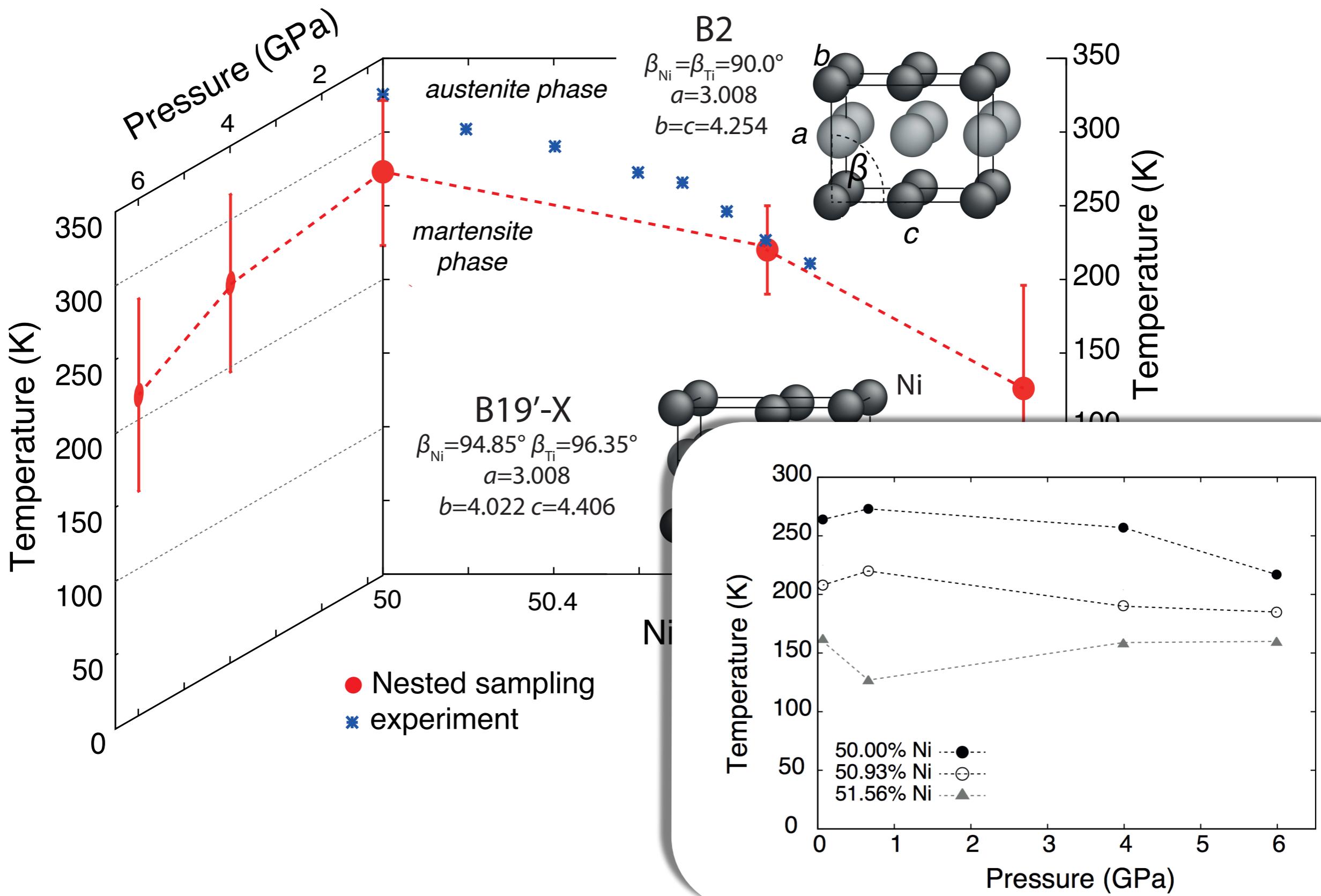


# Binary Lennard-Jones

$$\epsilon_{AB} = 2\epsilon_{AA} = 2\epsilon_{BB}$$

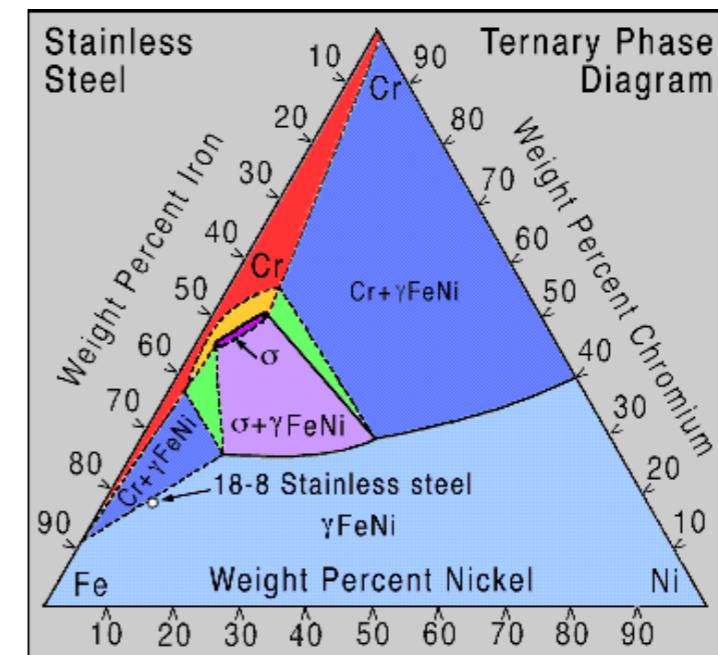
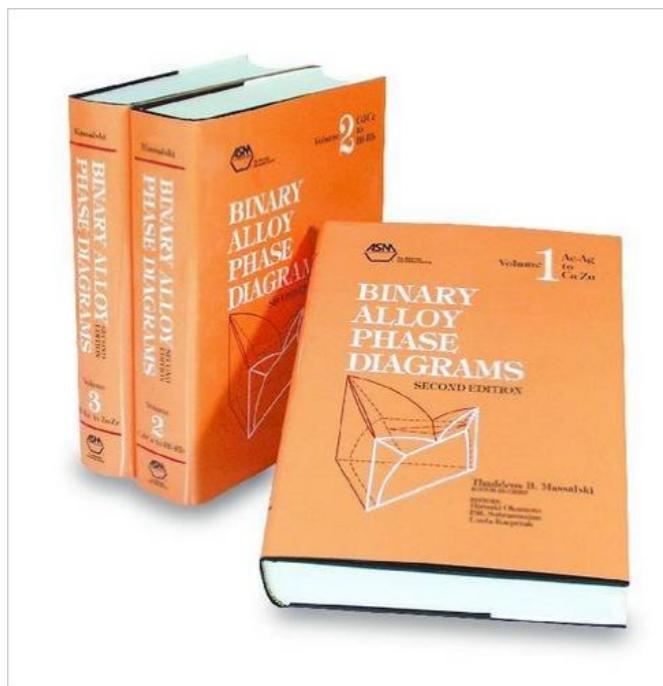


# NiTi shape memory alloy



# Phase diagram perspective

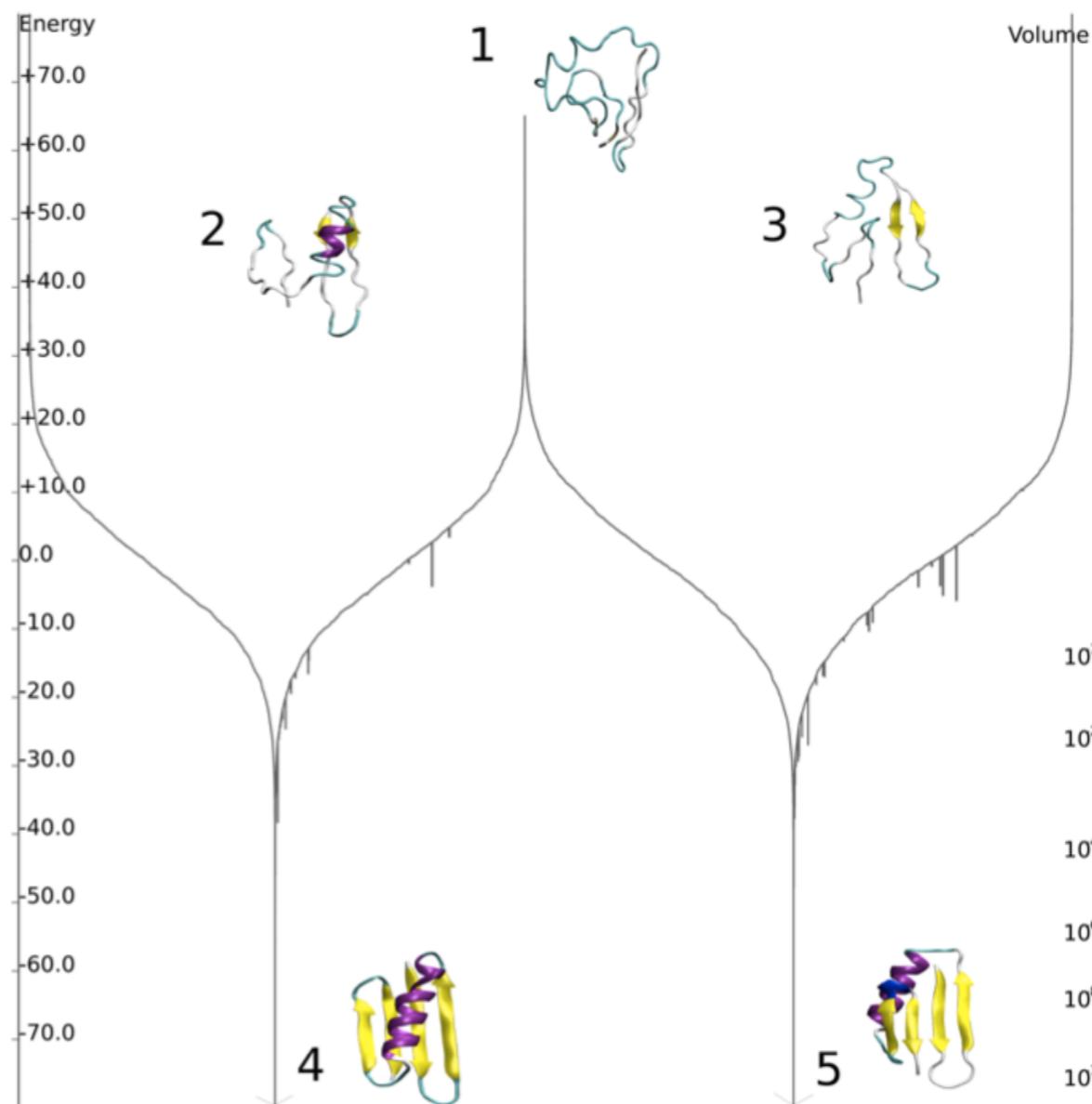
- 1 Billion energy evaluations (feasible)
- High-throughput computation (massively parallel)
- Optimise models to reproduce simple phase diagrams
- Ternary and multi-component alloys



# Protein landscapes

Landscape  
of protein G

using a coarse  
grained model



David Wild (Warwick, Systems Biology)

# Still early days for Nested Sampling in materials:

- Origin: J Skilling *J Bayesian Anal* **1**, 833 (2006)
- LJ clusters: L B Pártay et al. *J Phys Chem B* **114**, 10505 (2010)  
S Martiniani et al. *Phys Rev X* **4**, 031034 (2014)
- Hard Spheres: L B Pártay et al. *Phys Rev E* **89**, 022302 (2014)
- Water: H Do et al, *J Chem Phys* **135**, 174105 (2011)
- Proteins: N Burkoff et al., *Biophysical Journal*, **102**, 878-886, (2012)

## Vision

- Complete phase diagrams at the “touch of a button”
- High-throughput materials science
- Explore complete “model parameter space”

## Some open mathematical problems:

- Convergence rates for macroscopic number of local minima (glassy systems) ?
- Efficient uniform samplers ? Currently need  $\sim 10^9$  energy evaluations
- How can we make use of known good order parameters ?