Parallel Tempering and Population Annealing for Rough Free Energy Landscapes

Jon Machta

University of Massachusetts Amherst,
Santa Fe Institute

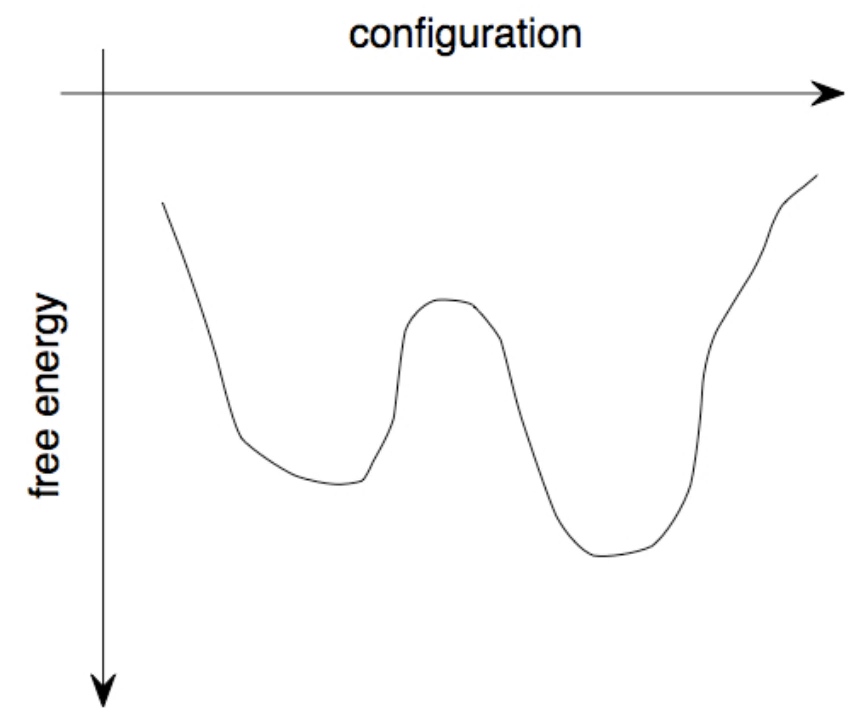
Melbourne, July 26, 2010

Collaborators

- Stephan Burkhardt
- Burcu Yucesoy
- Helmut Katzgraber

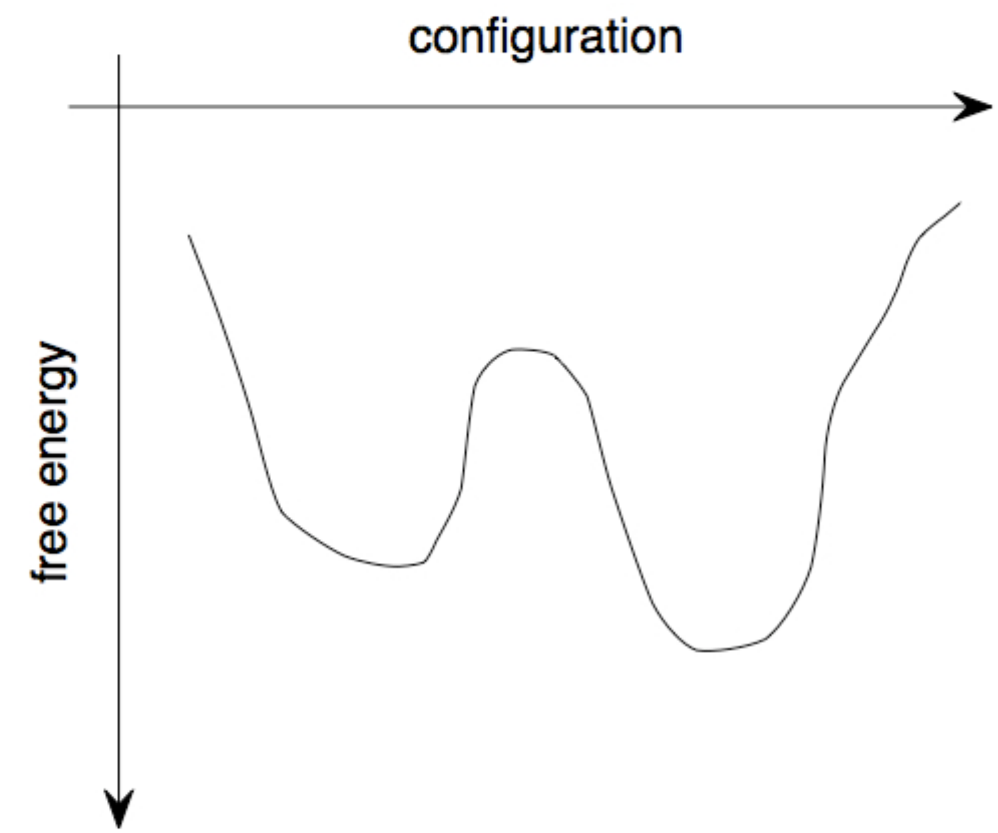
Motivation

- How to sample equilibrium states of systems with rugged free energy landscapes, e.g. spin glasses, configurational glasses, proteins, NP-hard combinatorial optimization problems.



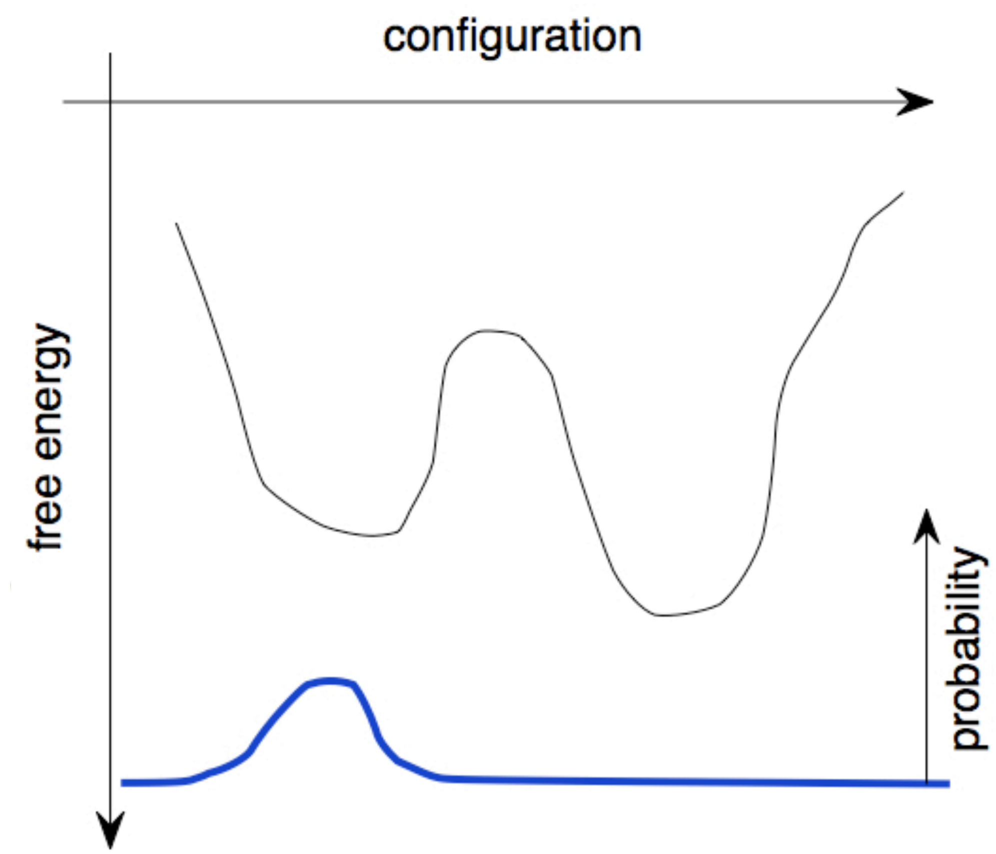
Problem

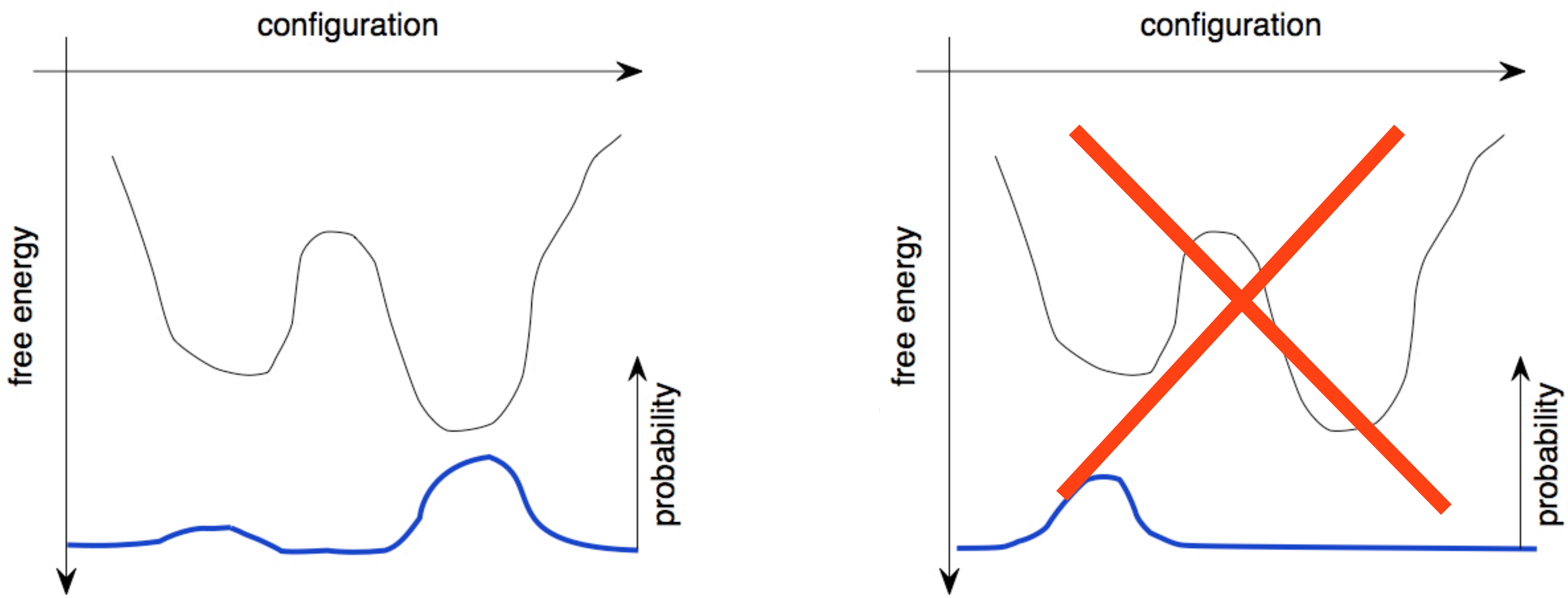
- Markov chain Monte Carlo (MCMC) at a single temperature such as the Metropolis algorithm gets stuck in local minima.



Problem

- Markov chain Monte Carlo (MCMC) at a single temperature such as the Metropolis algorithm gets stuck in local minima.

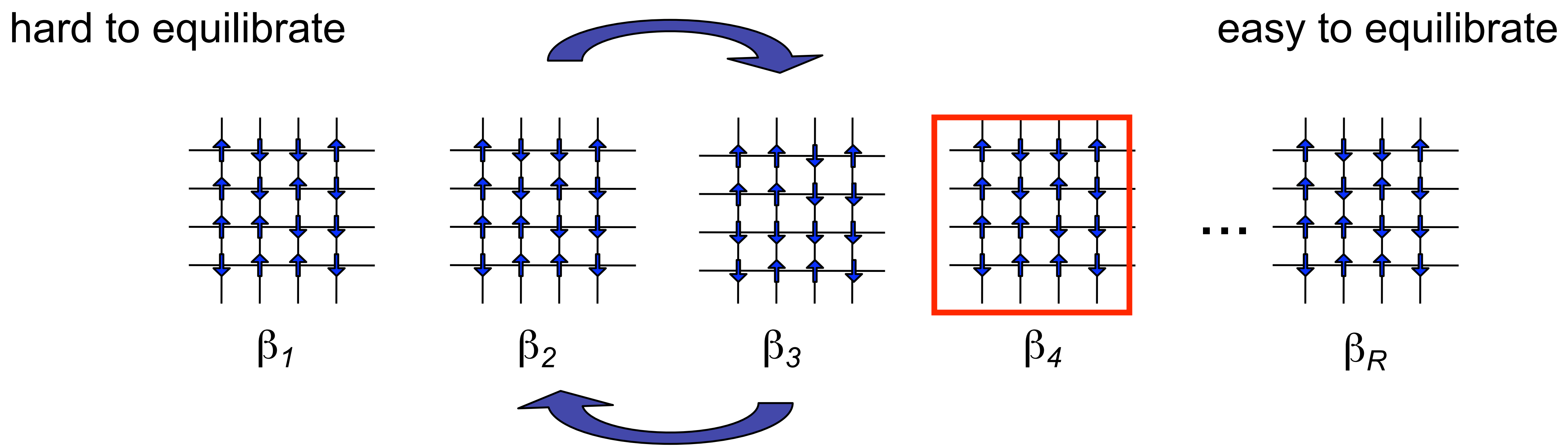



configuration
free energy
probability
configuration
free energy
probability

Outline

- Introduction to Parallel Tempering (aka Replica Exchange Monte Carlo)
- Strengths and Weaknesses of PT (in several simple landscapes)
- Introduction to Population Annealing
- PA vs PT
- Conclusions

Parallel Tempering



•*R* replicas at inverse temperatures $\beta_1 > \beta_2 > \ldots > \beta_R$ (each with the same couplings).

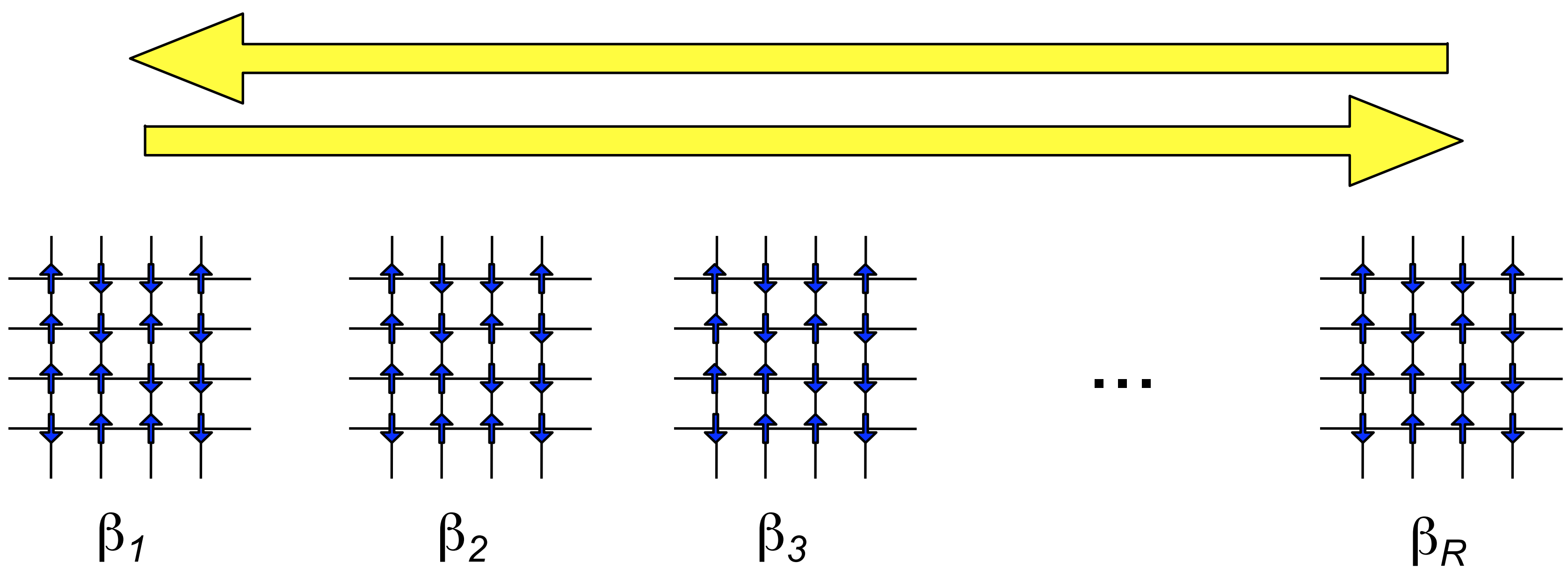
•MCMC (e.g. Metropolis) on each replica

•Exchange replicas with energies E and E' and temperatures β and β',with probability:

$$p_{\mathrm{swap}} = \min\left[1, e^{(\beta-\beta')(E-E')}\right]$$

Intuition

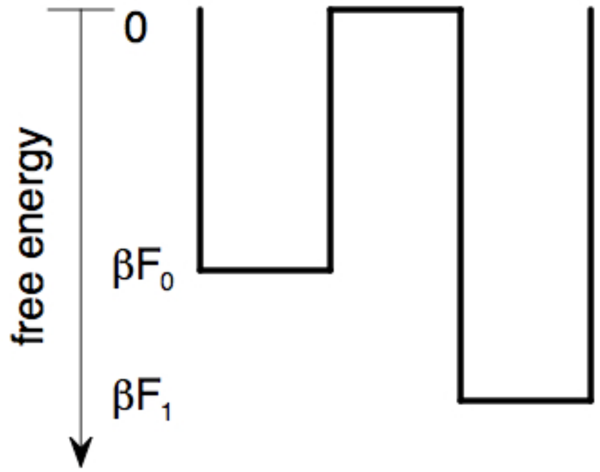
- Mixing is accelerated by “round trips” from low to high temperature and back.



  – Parallel tempering can be optimized by minimizing the equilibrium round trip time.

A simple landscape:

- Consider a model free energy landscape with two free energy minima separated by a high barrier.
  - JM, PRE **80**, 056706 (2009)



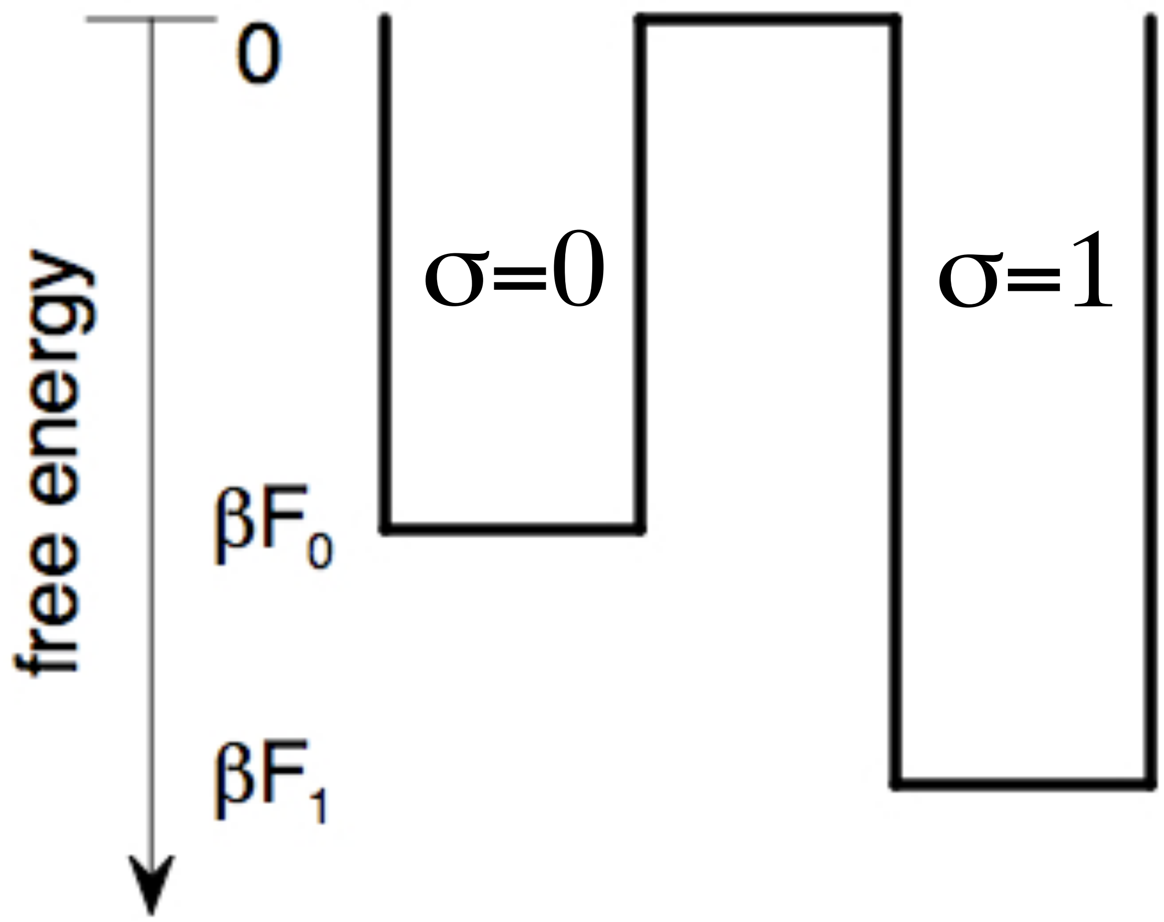
Two Well Model

$$\beta F_\sigma(\beta) = -\frac{1}{2}(\beta - \beta_c)^2(K + H\sigma)$$

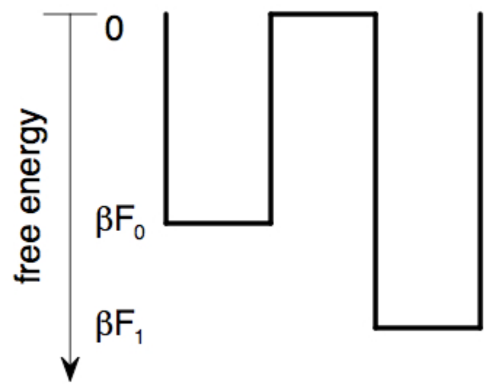
$$\sigma = 0, 1$$

$$\beta \delta F = -\frac{1}{2}(\beta - \beta_c)^2 H$$

$$\text{Prob}\left[\sigma = +1\right] = \frac{1}{1 + e^{-\beta\delta F}}$$



Parallel tempering for the two-well model



- Assumptions:
  - Fast equilibration within each well by standard MCMC.
  - No transitions between wells except at β_c where each well is equally probable.
  - Energy is normally distributed in each state; from thermodynamics:

$$\langle E \rangle = -(\beta - \beta_c)(K + H\sigma) \qquad \mathbf{Var}(E) = (K + H\sigma)$$

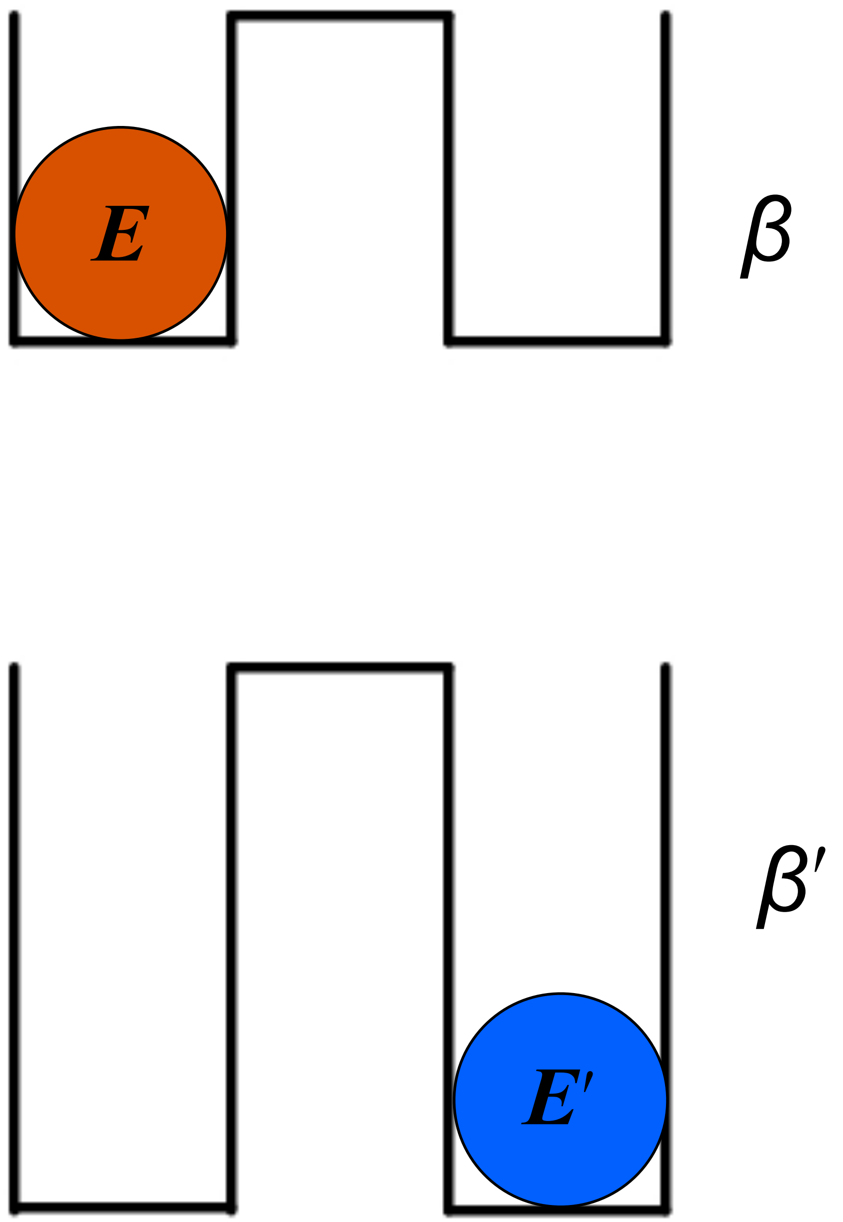
Replica exchange probabilities

- For the two-well model, replica exchange transition probabilities can be computed exactly. For symmetric wells ($H=0$):

$$p_{\text{swap}} = \min\left[1, e^{(\beta-\beta')(E-E')}\right]$$

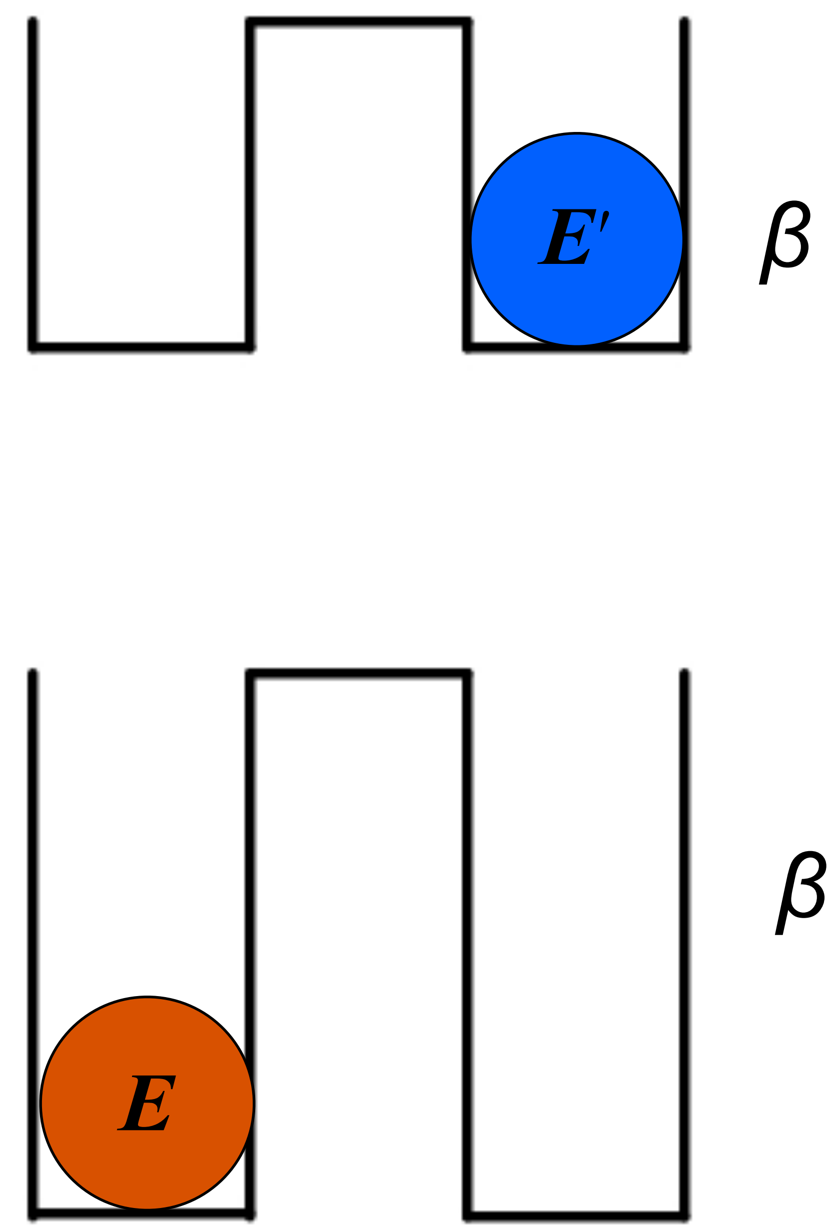
$$\langle E \rangle = -(\beta - \beta_c)K \ , \ \mathbf{Var}(E) = K$$

$$p_{\text{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$



Replica exchange probabilities

- For the two-well model, replica exchange transition probabilities can be computed exactly. For symmetric wells ($H=0$):

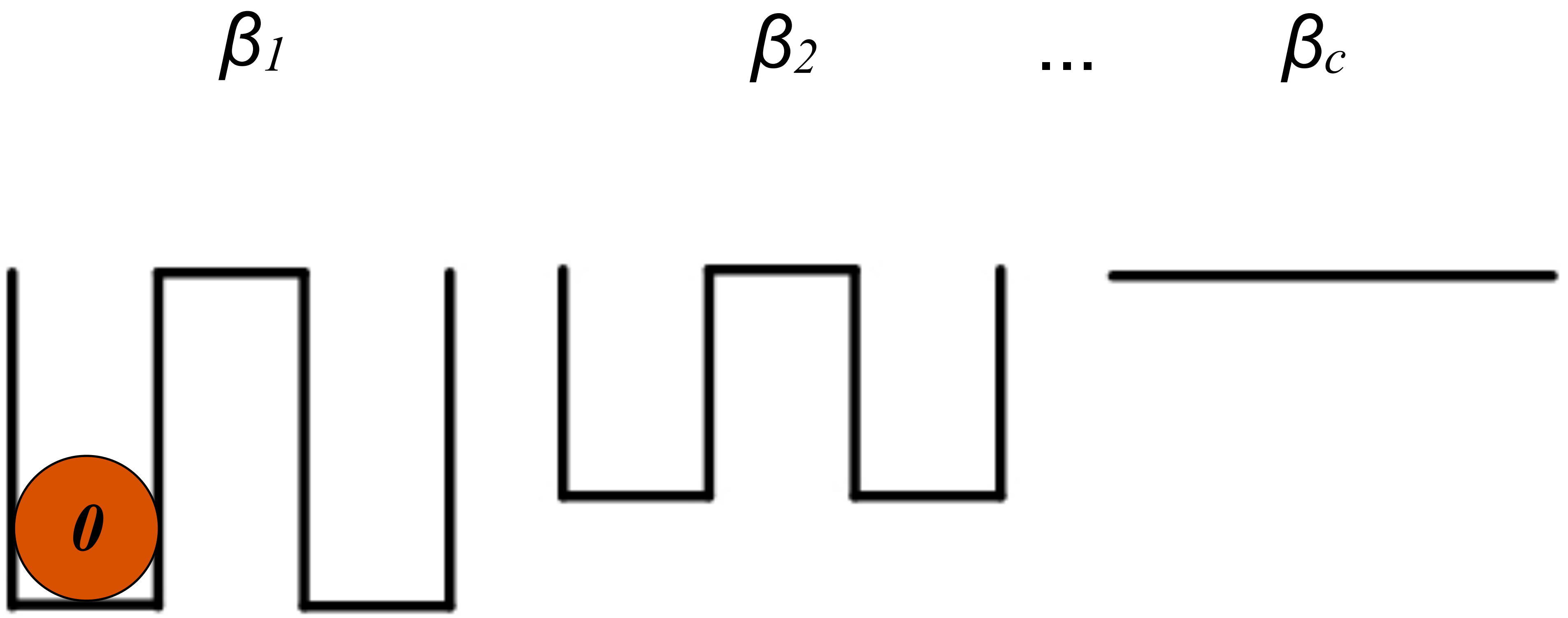


$$p_{\mathrm{swap}} = \min\left[1, e^{(\beta-\beta')(E-E')}\right]$$

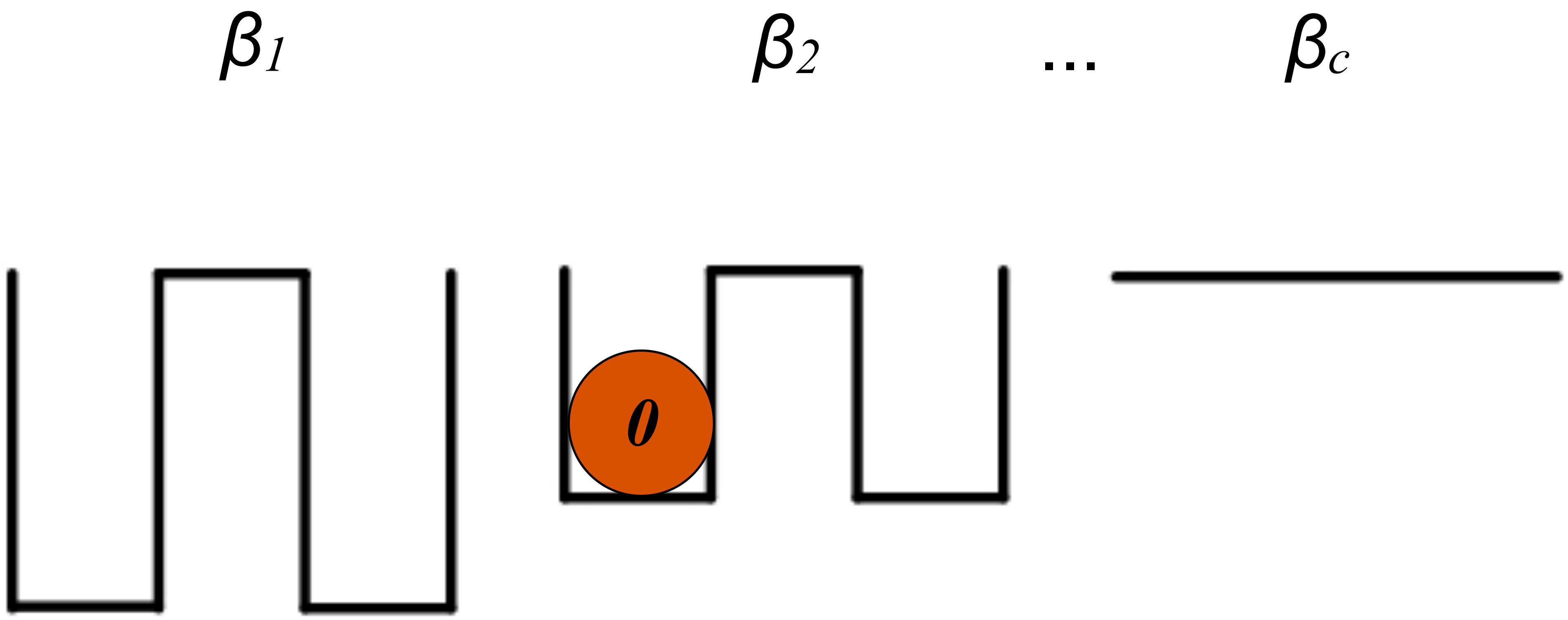
$$\langle E \rangle = -(\beta - \beta_c)K \ , \ \mathbf{Var}(E) = K$$

$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$

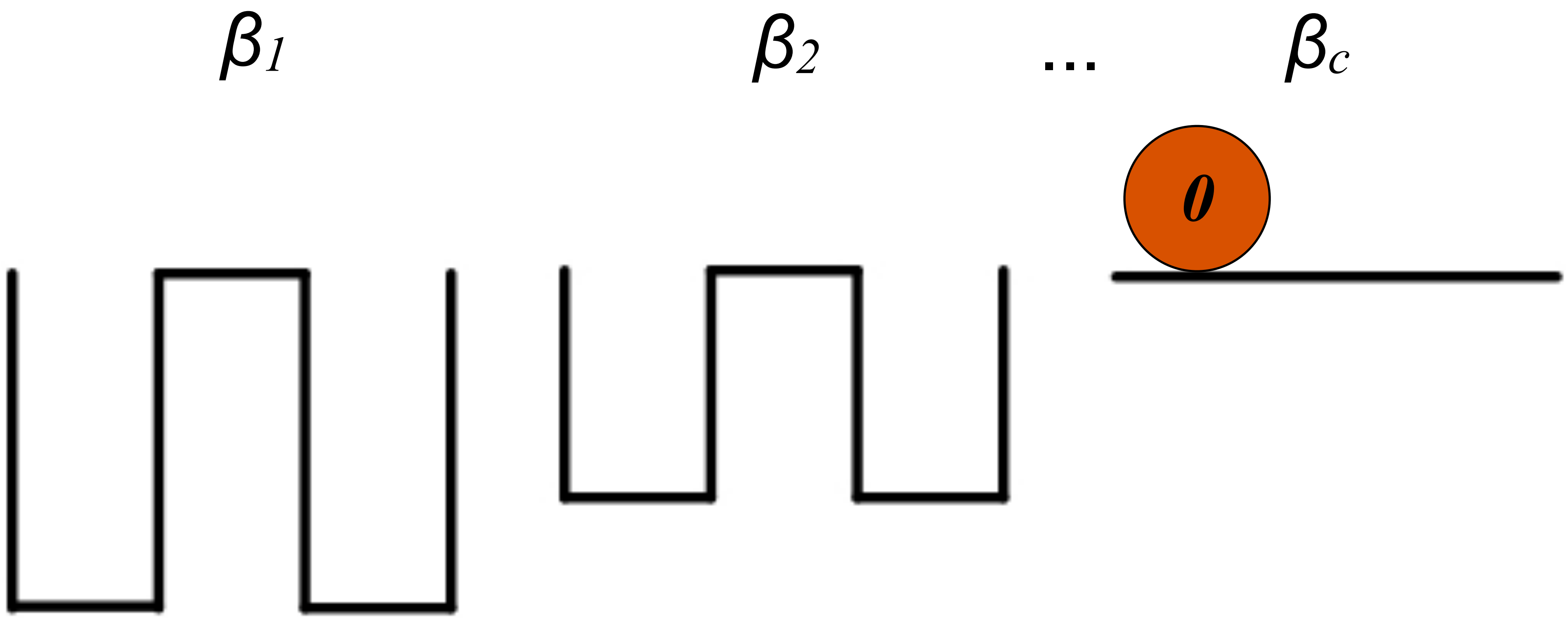
Moving between wells



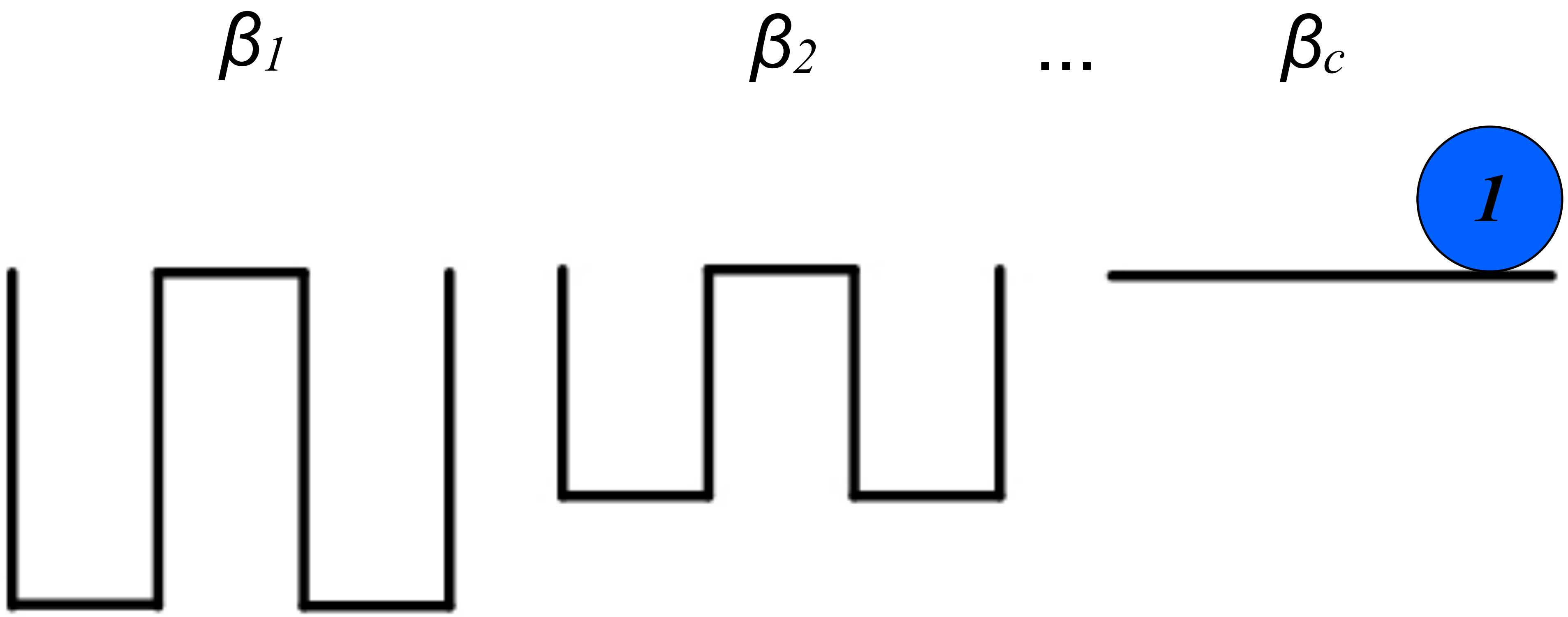
Moving between wells



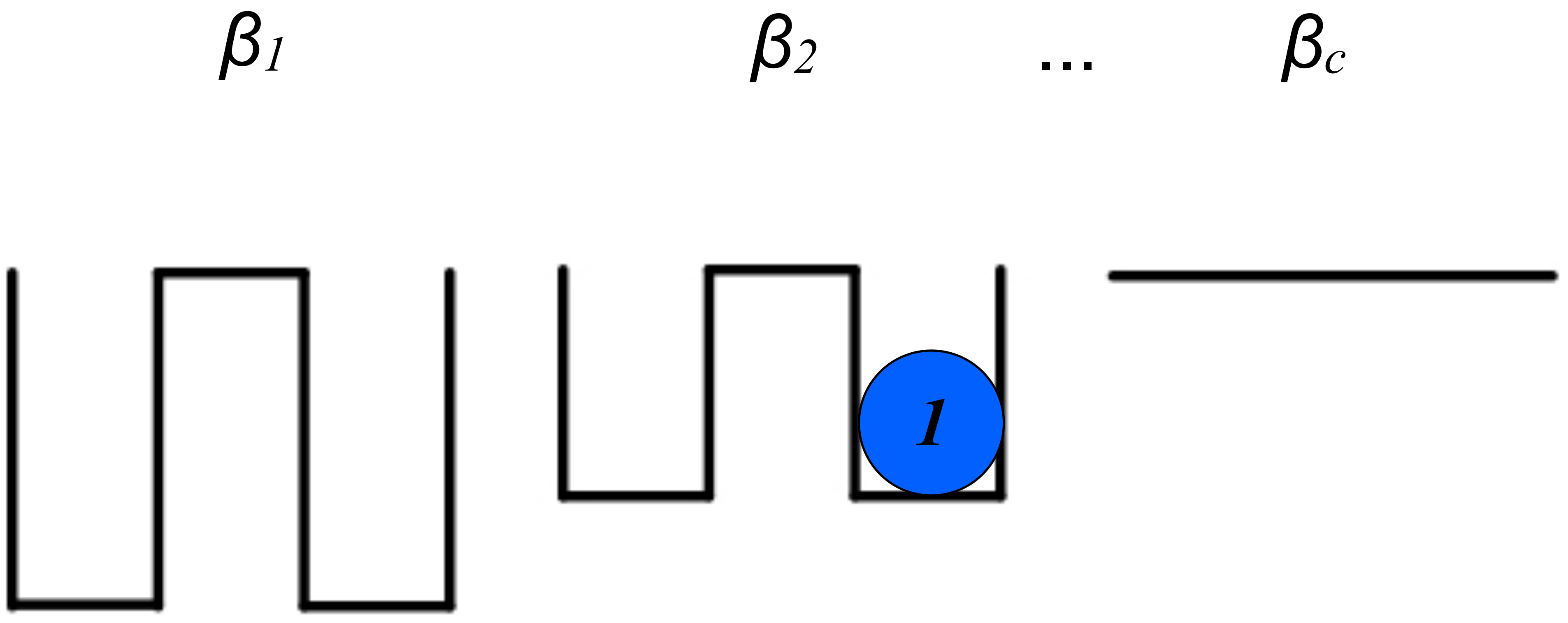
Moving between wells



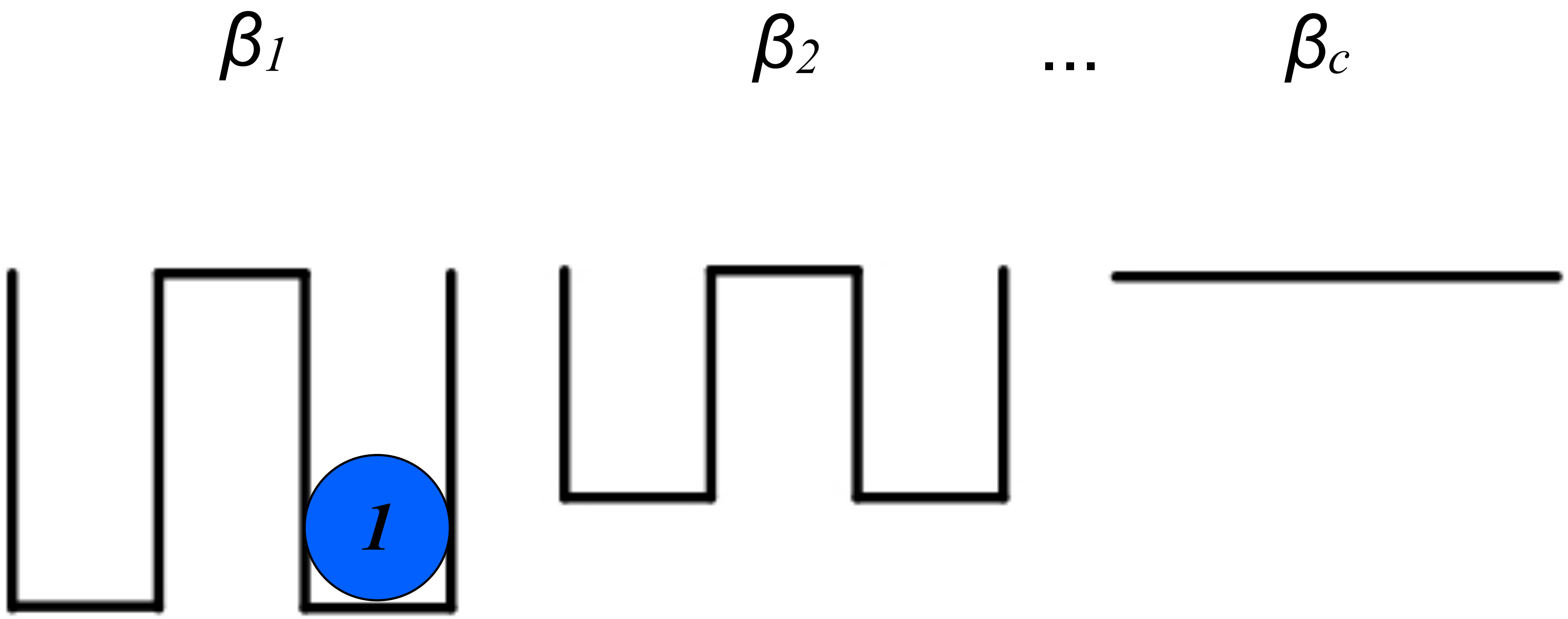
Moving between wells



Moving between wells

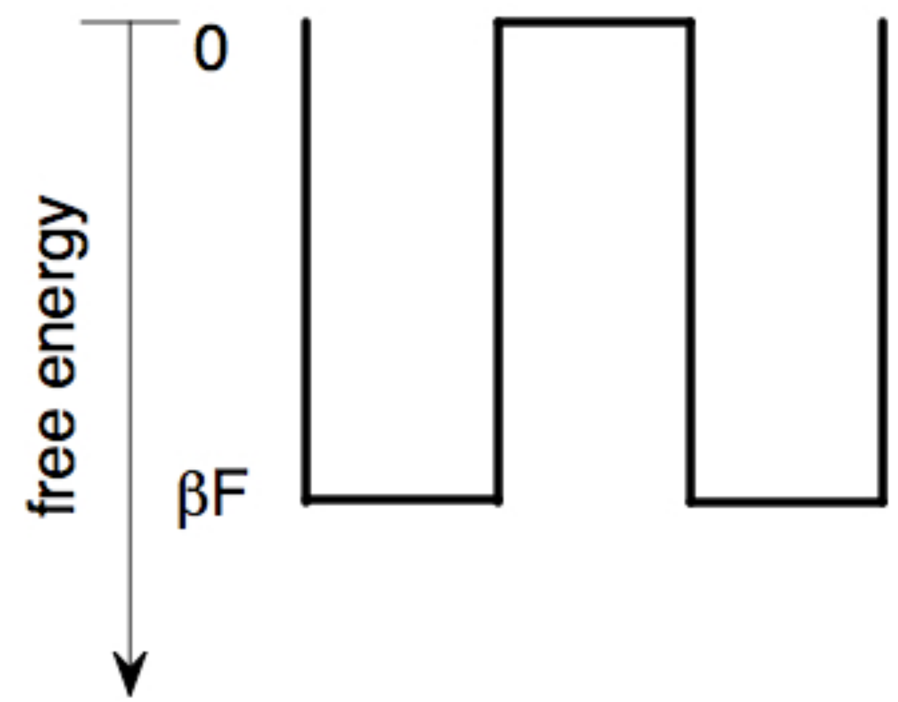


Moving between wells



Symmetric wells (H=0)

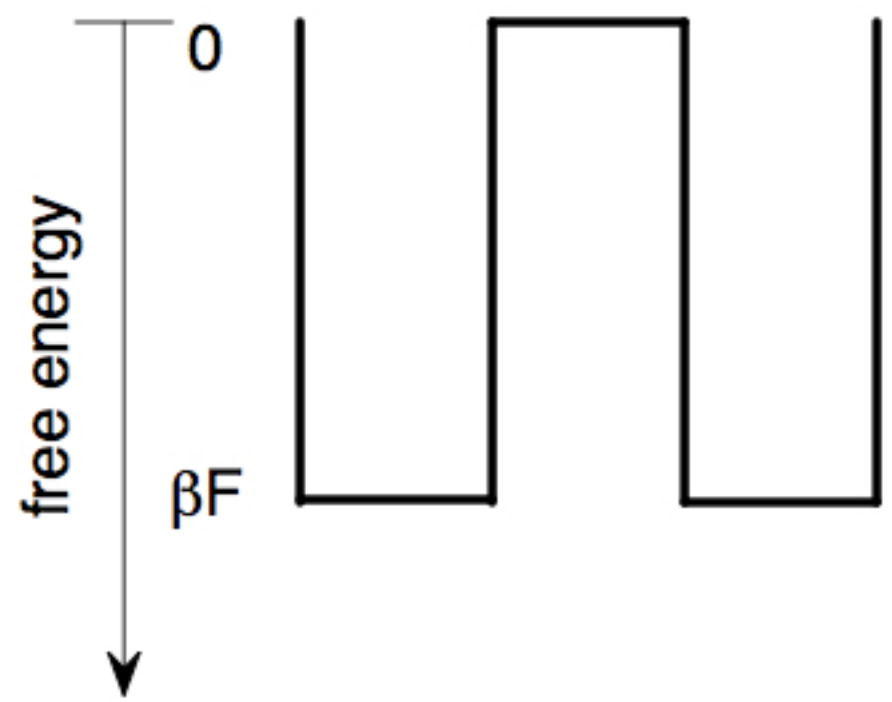
$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$



Symmetric wells (H=0)

- *Diffusion* of replicas

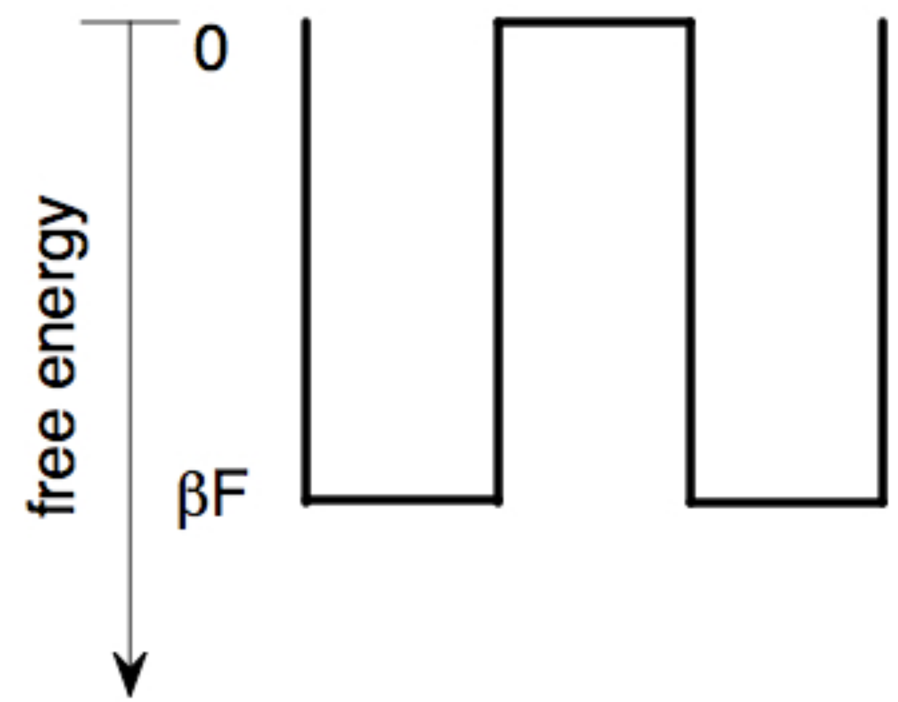
$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$



Symmetric wells (H=0)

- *Diffusion* of replicas
- Equilibration time τ is proportional to the <u>mean first passage time</u> for diffusion from β_0 to β_c with R equally spaced replicas

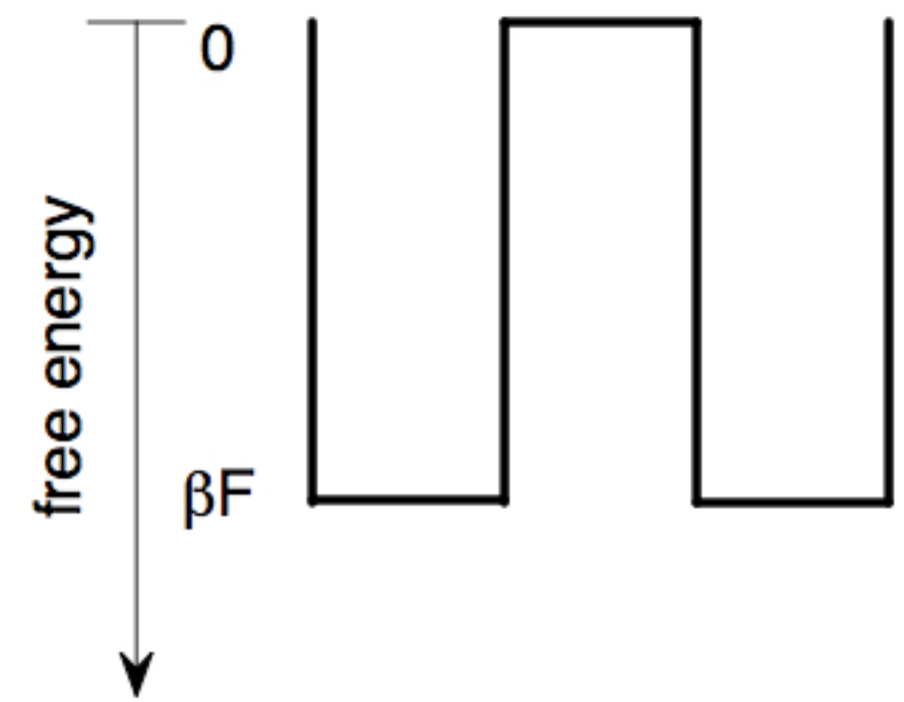
$$p_{\text{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$



Symmetric wells (H=0)

- *Diffusion* of replicas
- Equilibration time τ is proportional to the mean first passage time for diffusion from β_0 to β_c with R equally spaced replicas
- Optimum number of replicas balances diffusion time and replica exchange acceptance fraction.

$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$

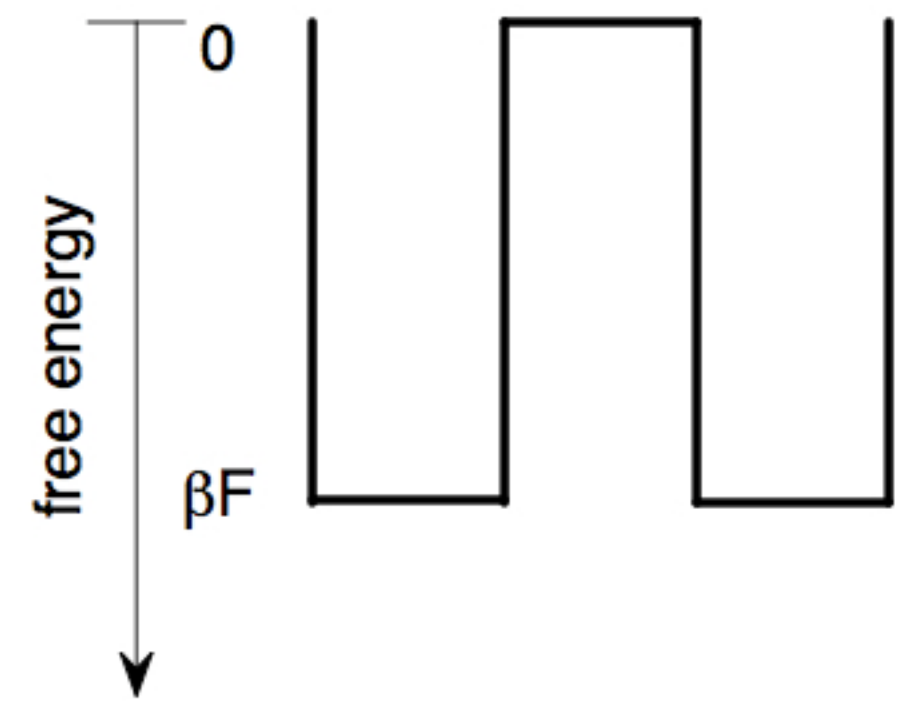


Symmetric wells (H=0)

- *Diffusion* of replicas

$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$

- Equilibration time τ is proportional to the <u>mean first passage time</u> for diffusion from β_0 to β_c with R equally spaced replicas



- Optimum number of replicas balances diffusion time and replica exchange acceptance fraction.

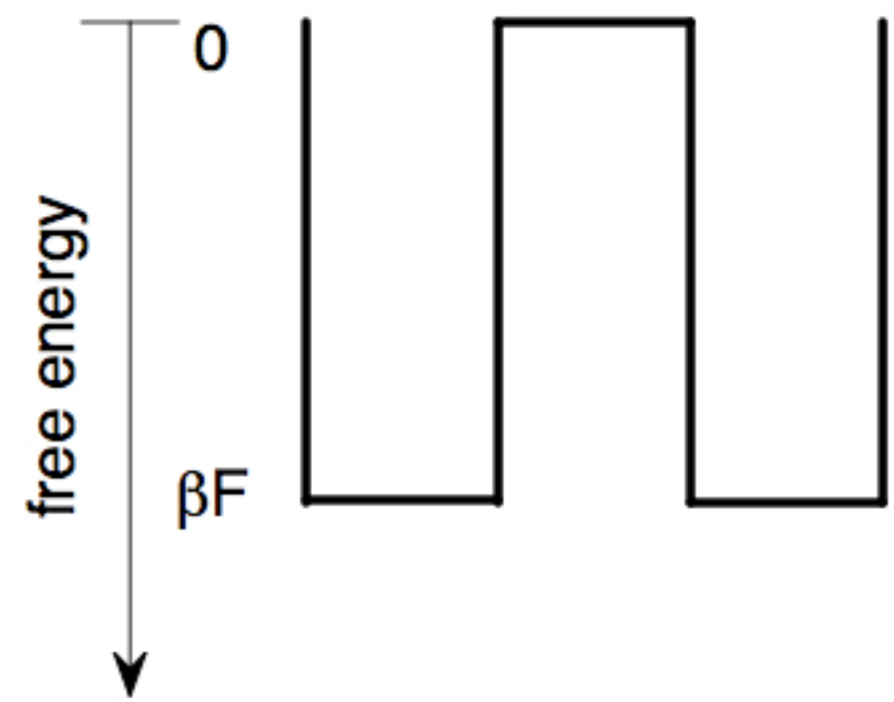
$$R_{\mathrm{opt}} = 1 + 0.594(\beta_0 - \beta_c)\sqrt{K} \sim \sqrt{-\beta F}$$

Symmetric wells (H=0)

- *Diffusion* of replicas
- Equilibration time τ is proportional to the <u>mean first passage time</u> for diffusion from β_0 to β_c with R equally spaced replicas
- Optimum number of replicas balances diffusion time and replica exchange acceptance fraction.

$$p_{\mathrm{swap}} = \frac{1}{2}\mathrm{Erfc}\left((\beta - \beta')\sqrt{K}\right)$$

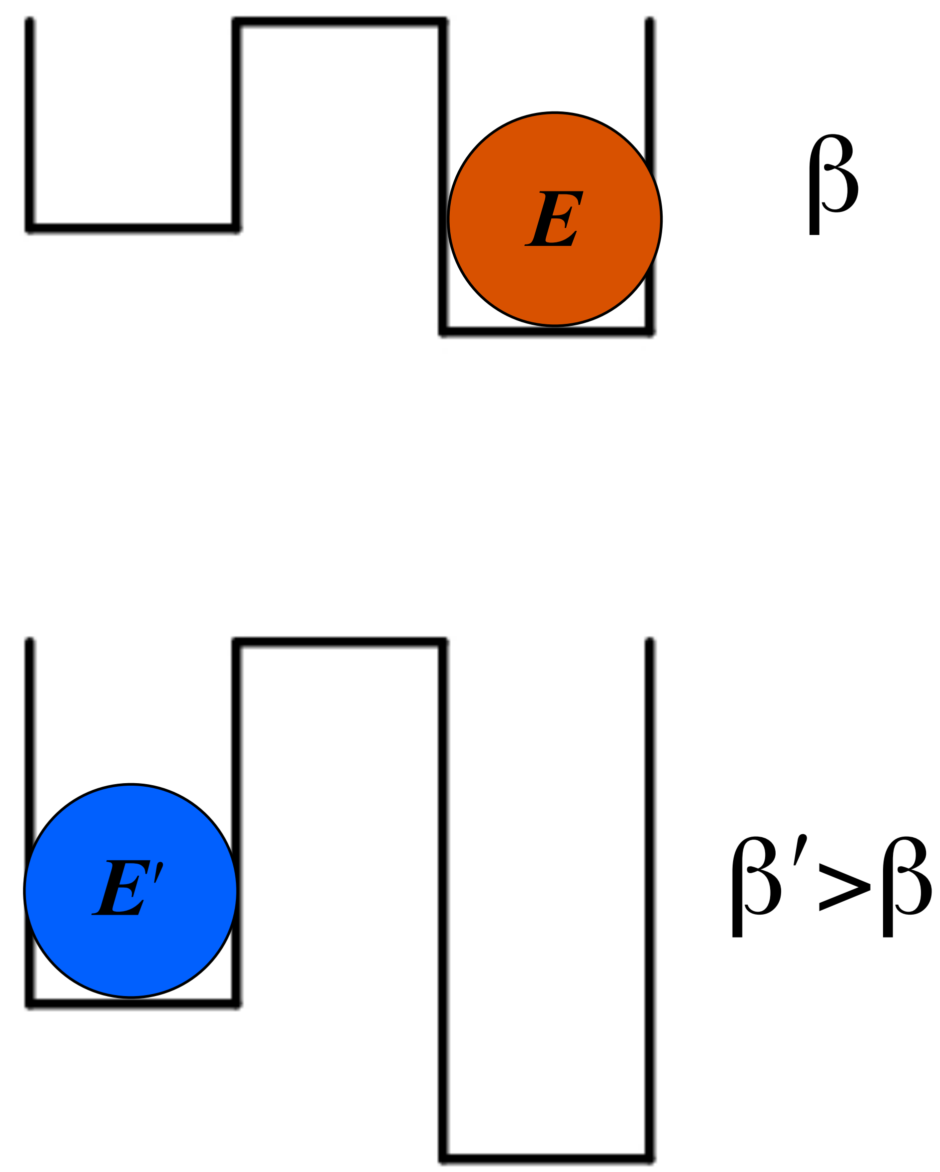
$$\boxed{\tau \sim R^2 \sim \beta F}$$



$$R_{\mathrm{opt}} = 1 + 0.594(\beta_0 - \beta_c)\sqrt{K} \sim \sqrt{-\beta F}$$

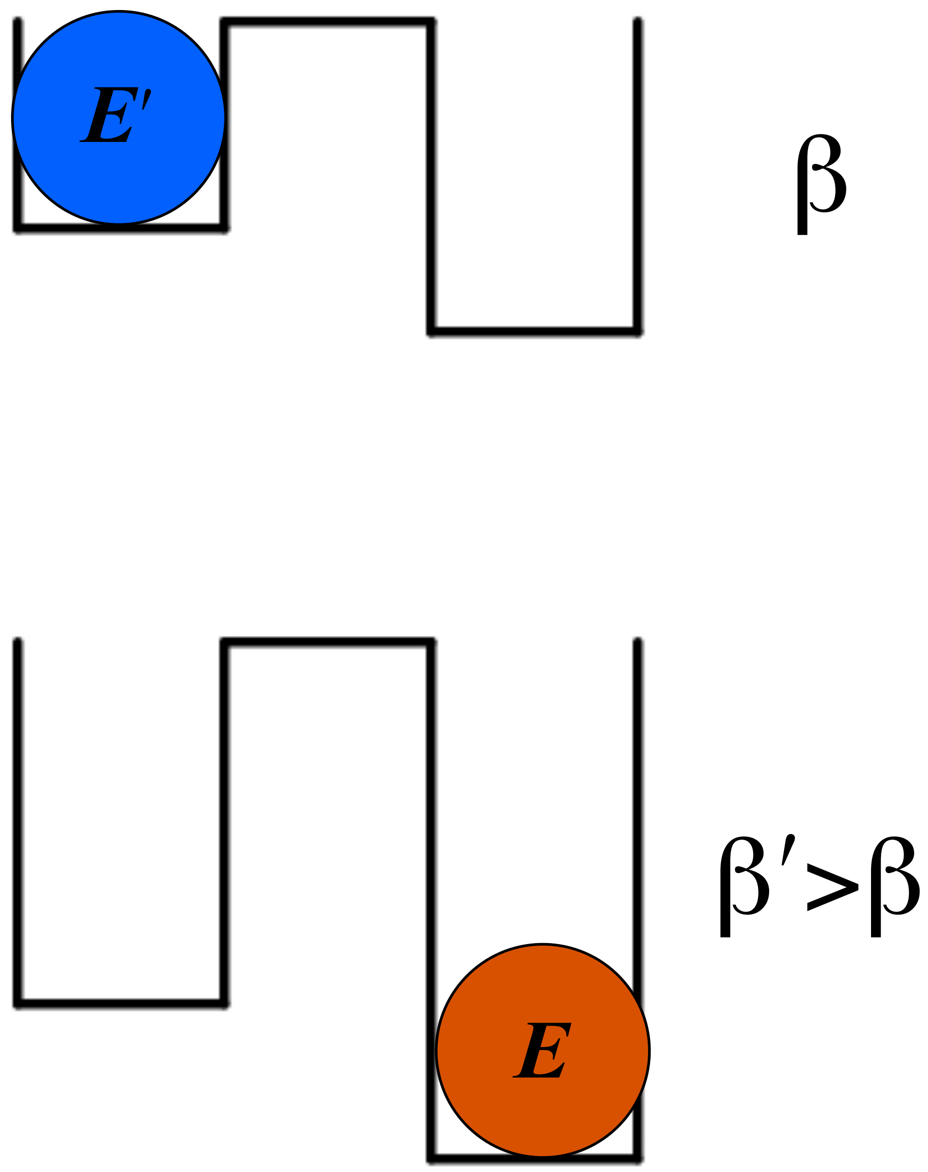
Asymmetric wells (H>0)

- Replica exchange is biased toward moving replicas in the deep well to lower temperature.
- *Drift and diffusion* instead of diffusion:
  - Replicas in the deeper well drift toward lower temperature and vice versa.
- Faster equilibration than for the symmetric case!



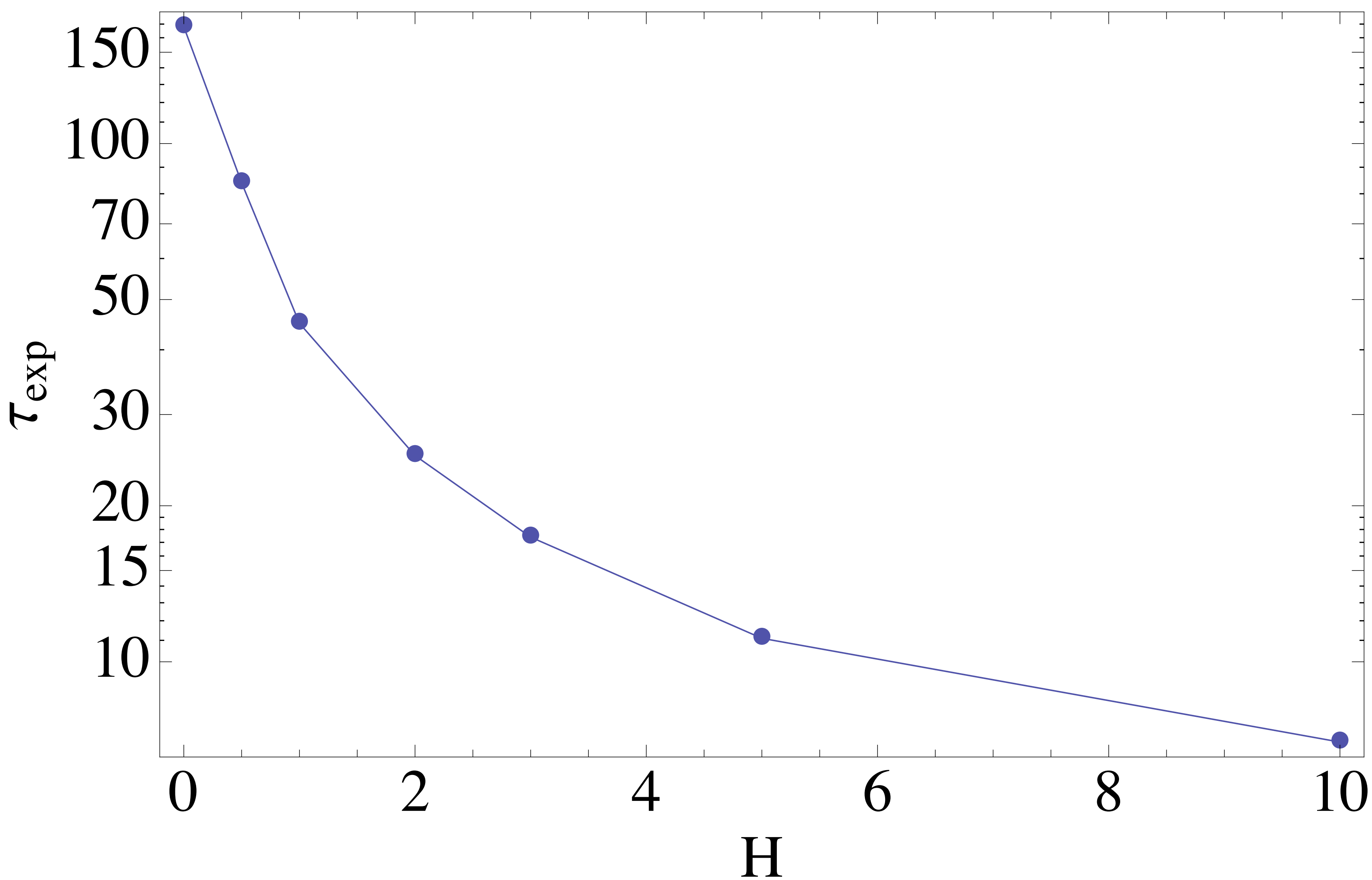
Asymmetric wells (H>0)

- Replica exchange is biased toward moving replicas in the deep well to lower temperature.
- *Drift and diffusion* instead of diffusion:
  - Replicas in the deeper well drift toward lower temperature and vice versa.
- Faster equilibration than for the symmetric case!



Equilibration time vs well asymmetry

$$\beta \delta F = -\frac{1}{2}(\beta - \beta_c)^2 H$$



symmetric *asymmetric*

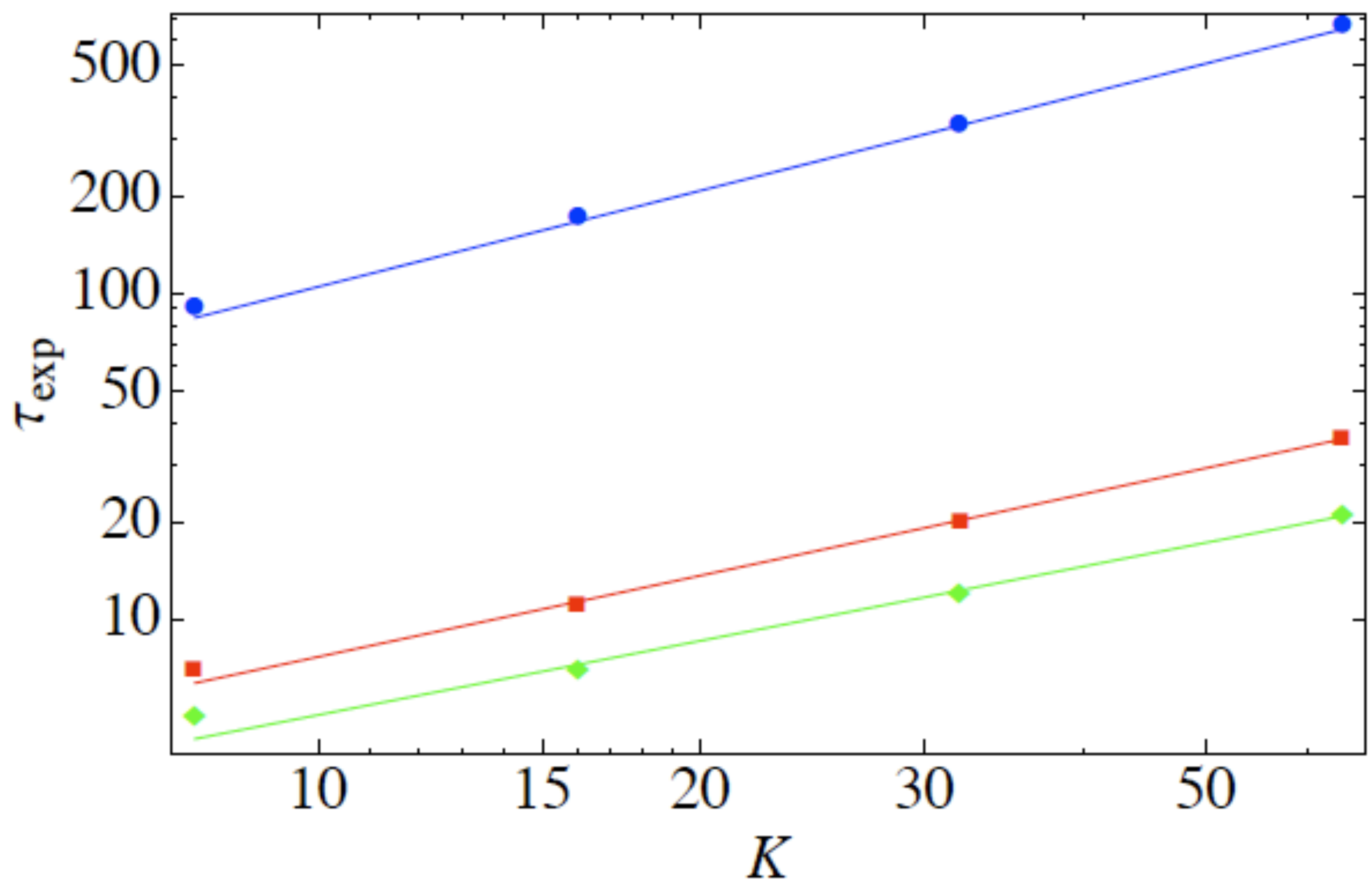
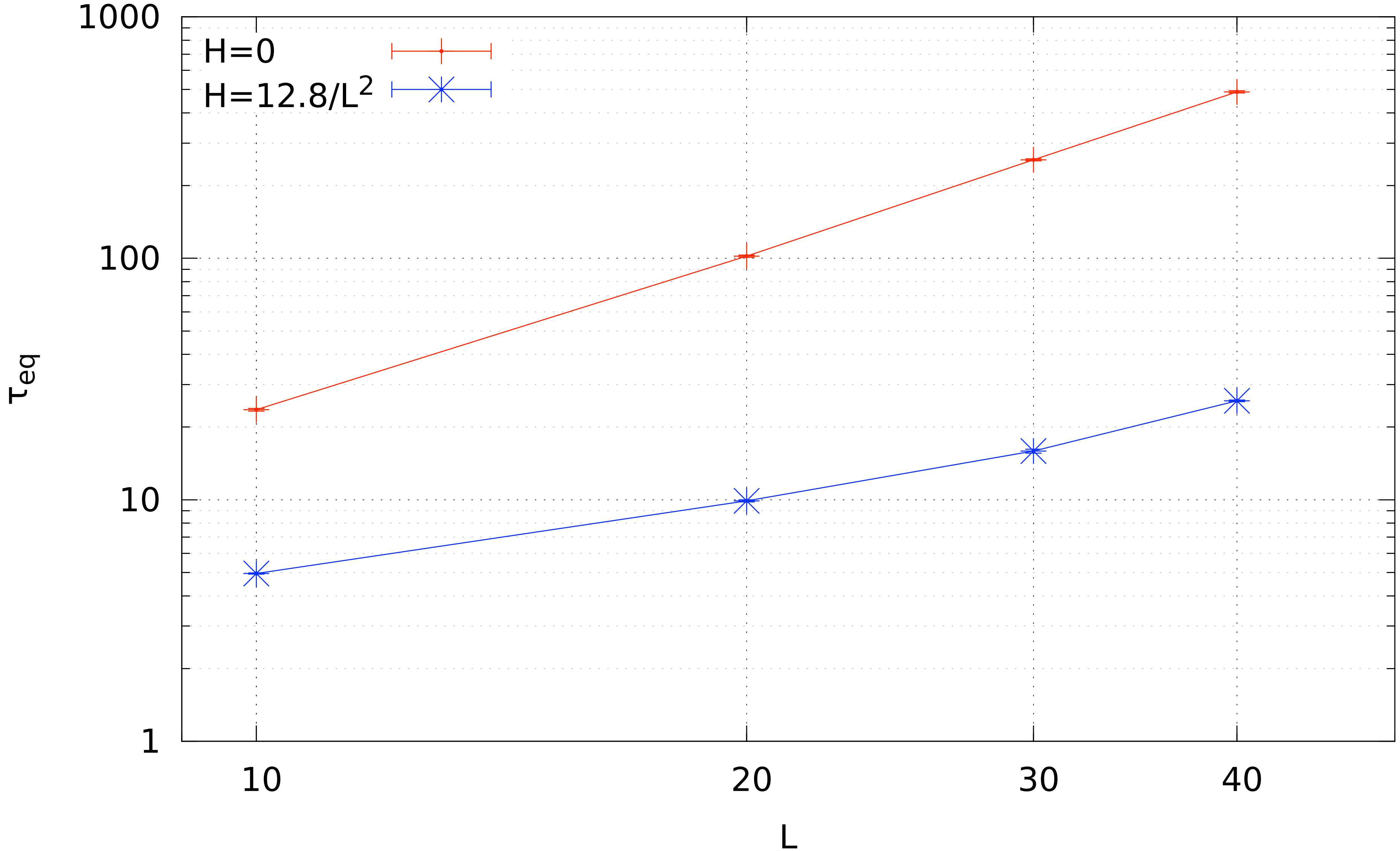


FIG. 2: The exponential autocorrelation time τ_{exp} for the fraction of sites in the deep well vs. the well depth parameter K for $H = 10$ (green diamonds), $H = 5$ (red squares) and $H = 0$ (blue circles). The lines are best power law fits, $\tau_{\mathrm{exp}} \sim K^x$ with $x = 0.76$, 0.83 and 0.99 for $H = 10$, 5 and 0, respectively.

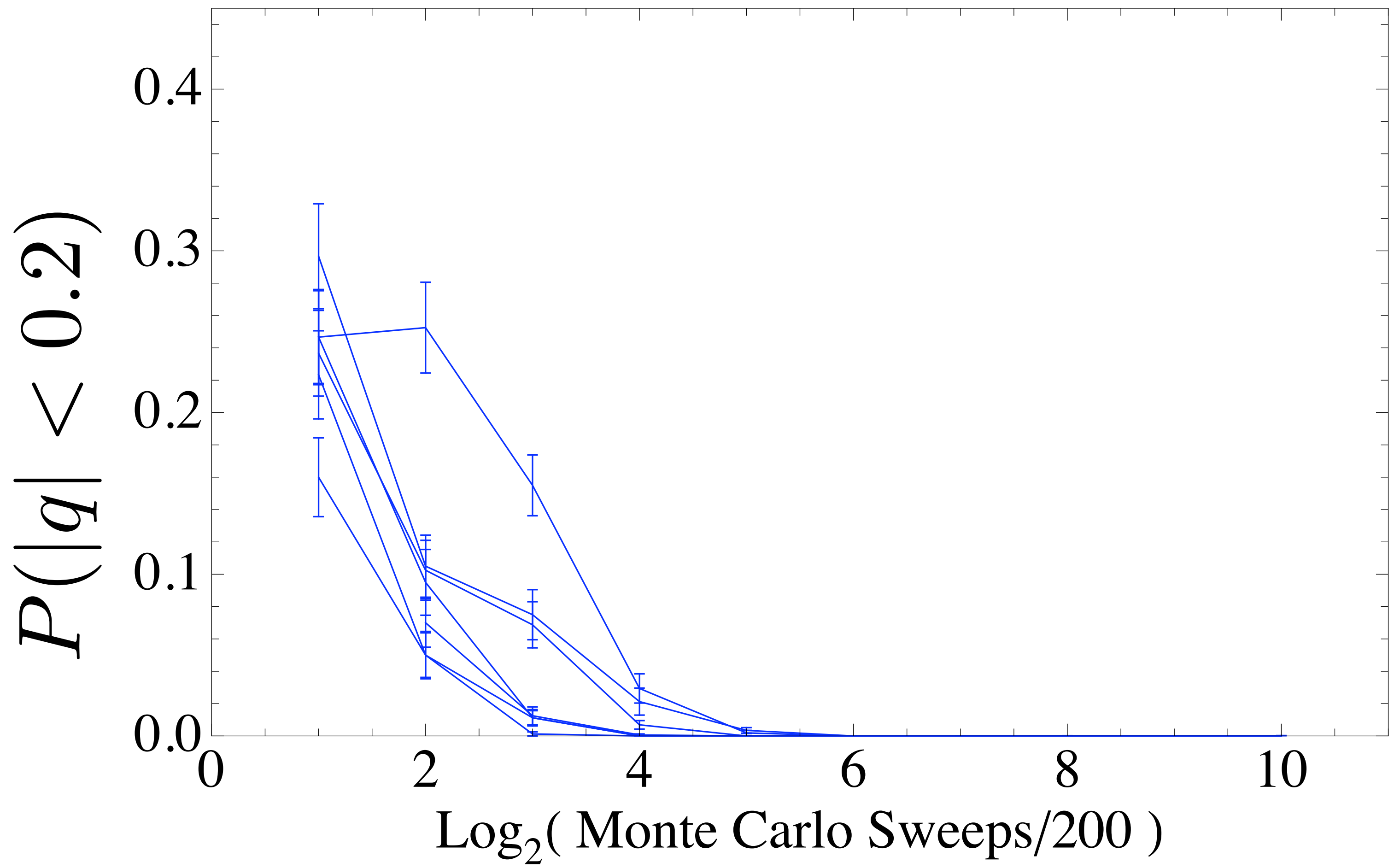
PT for the ferromagnetic Ising model



Exponential autocorrelation time for the magnetization for the low temperature Ising model with and without a temperature dependent external field.

Ising spin glass, Gaussian disorder, L=8, T/T_c=0.2

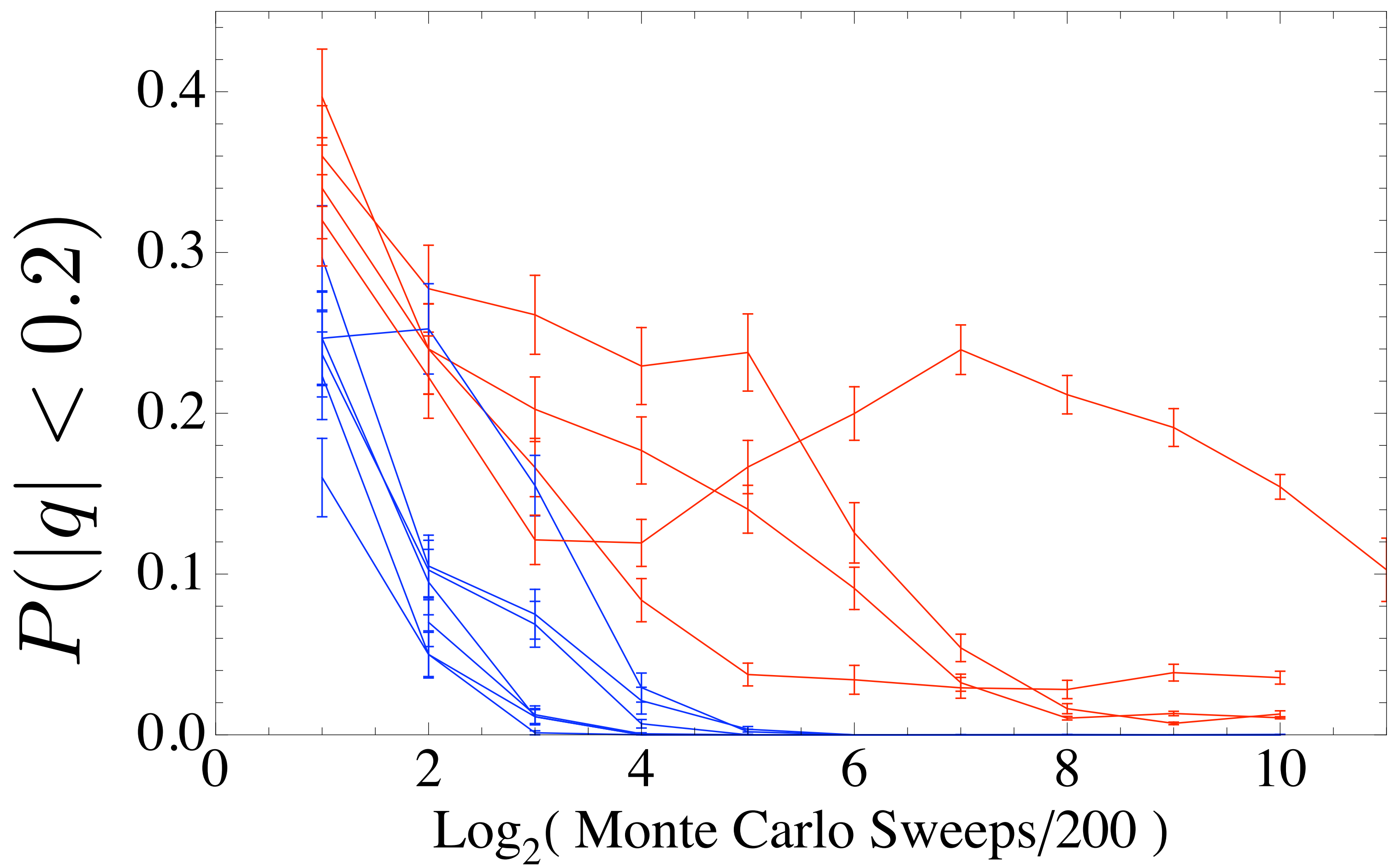
$P(|q| < 0.2) \approx 0$



Ising spin glass, Gaussian disorder, L=8, T/T_c=0.2

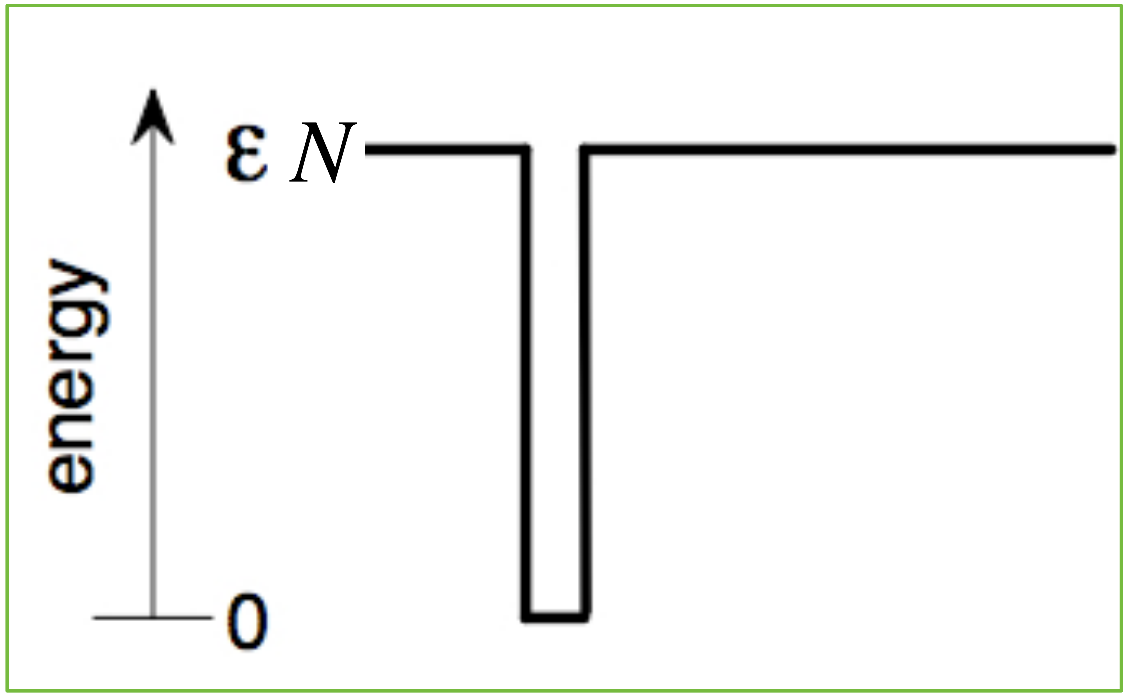
$P(|q| < 0.2) \approx 0$

$P(|q| < 0.2) > 0$



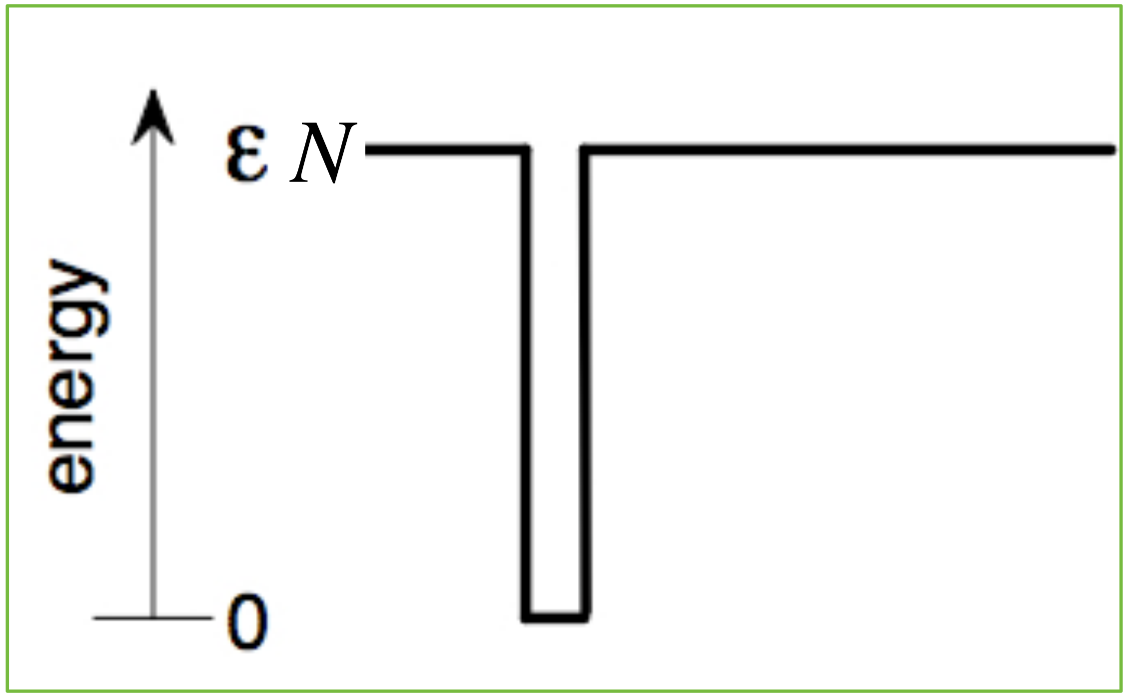
Weakness of Parallel Tempering

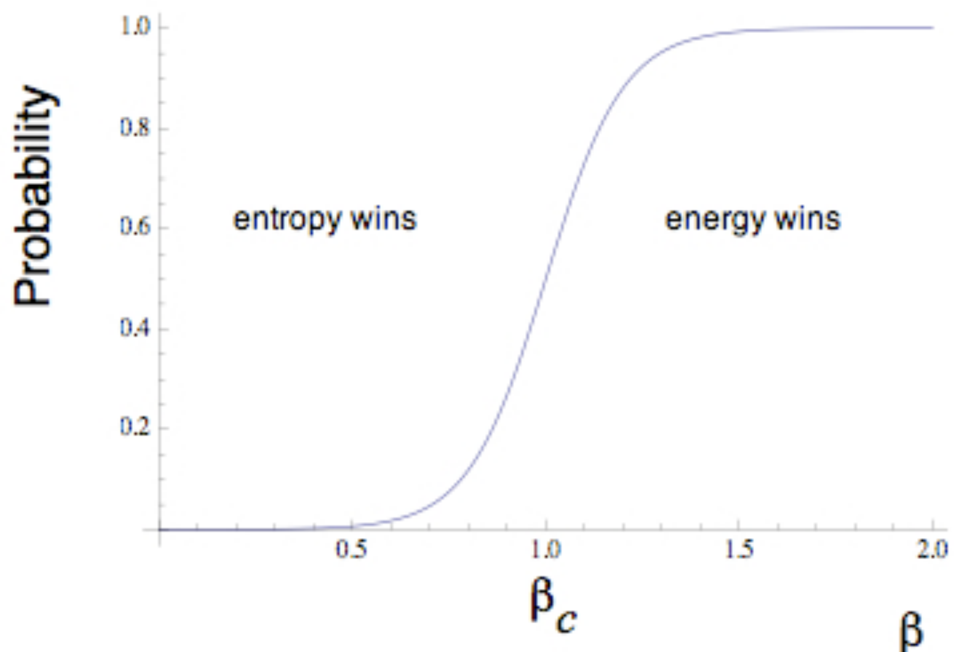
- "First-order" free energy landscapes. The simplest and worst case is the "golf-course" potential:
  - e^N microstates,
  - Nearly all have energy $\varepsilon N > 0$
  - An small fraction, $e^{-\varepsilon\beta_c N}$ are ground states with energy zero. There is a pseudo phase transition at β_c
  - Probability to be in a ground state:



Weakness of Parallel Tempering

- "First-order" free energy landscapes. The simplest and worst case is the "golf-course" potential:
  - e^N microstates,
  - Nearly all have energy $\varepsilon N > 0$
  - An small fraction, $e^{-\varepsilon\beta_c N}$ are ground states with energy zero. There is a pseudo phase transition at β_c
  - Probability to be in a ground state: $$\frac{1}{1+e^{-(\beta-\beta_c)\epsilon N}}$$





Equilibration on the Golf Course

Equilibration on the Golf Course

- The only advantage of parallel tempering is crude parallelism.

Equilibration on the Golf Course

- The only advantage of parallel tempering is crude parallelism.
- Each replica independently searches for grounds states and sends them to low temperature when they are found.

Equilibration on the Golf Course

- The only advantage of parallel tempering is crude parallelism.
- Each replica independently searches for grounds states and sends them to low temperature when they are found.
- Efficiency exceeds that of single temperature MCMC only to the extent that there are more replicas looking for grounds states than cold replicas ($\beta<\beta_c$) needing them.

Equilibration on the Golf Course

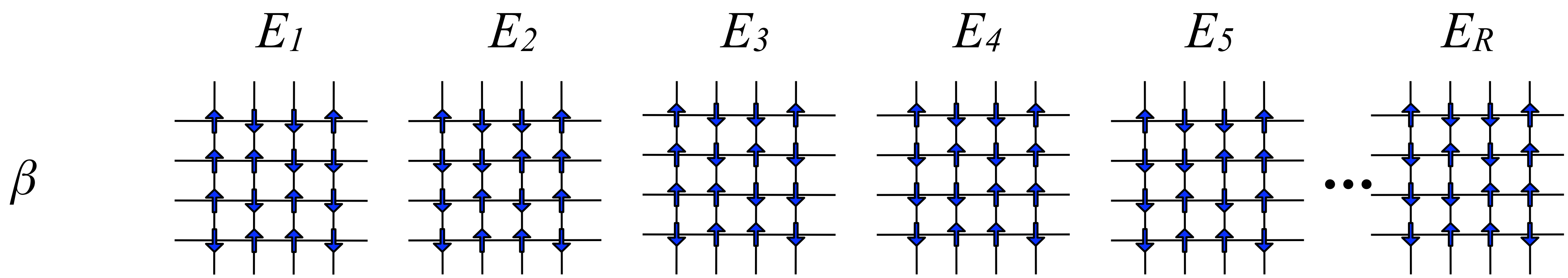
- The only advantage of parallel tempering is crude parallelism.
- Each replica independently searches for grounds states and sends them to low temperature when they are found.
- Efficiency exceeds that of single temperature MCMC only to the extent that there are more replicas looking for grounds states than cold replicas ($\beta<\beta_c$) needing them.
- *Before equilibration round trip-time is misleading.*

Population Annealing

K. Hukushima and Y. Iba, in *THE MONTE CARLO METHOD IN THE PHYSICAL SCIENCES: Celebrating the 50th Anniversary of the Metropolis Algorithm*, edited by J. E. Gubernatis (AIP, 2003), vol. 690, pp. 200–206.

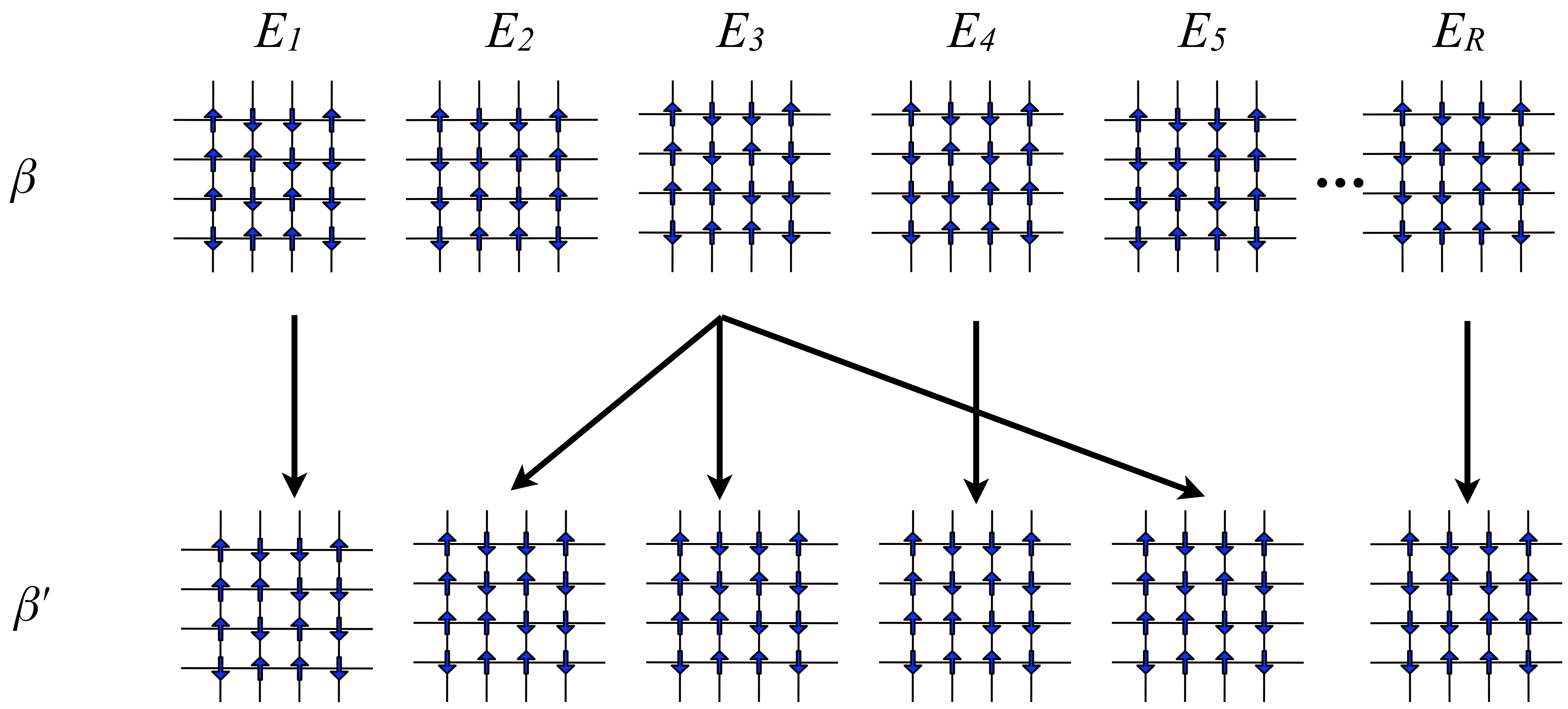
- Modification of simulated annealing for equilibrium sampling.
- A population of replicas is cooled according to an annealing schedule. Each replica is acted on by a MCMC (e.g. Metropolis) at the current temperature.
- During each temperature step, the population is differentially resampled according to Boltzmann weights to maintain equilibrium.

Population Annealing



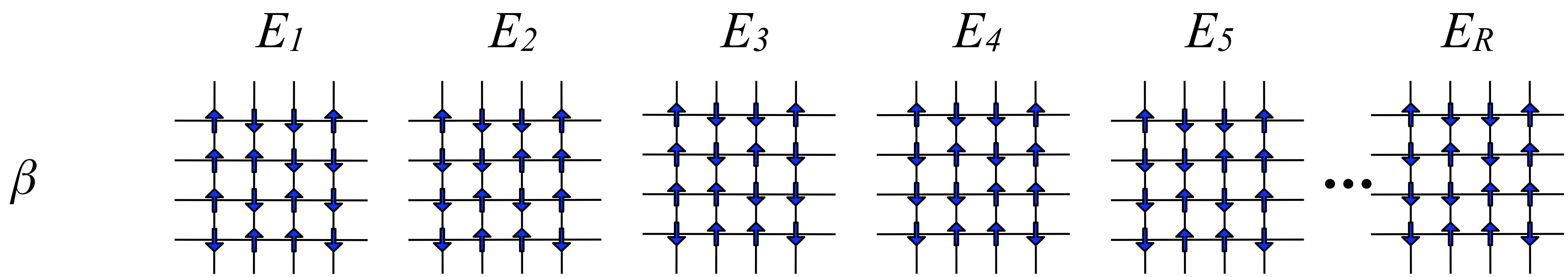
Population annealing = simulated annealing with differential reproduction of replicas

Population Annealing



Population annealing = simulated annealing with differential reproduction of replicas

Temperature Step

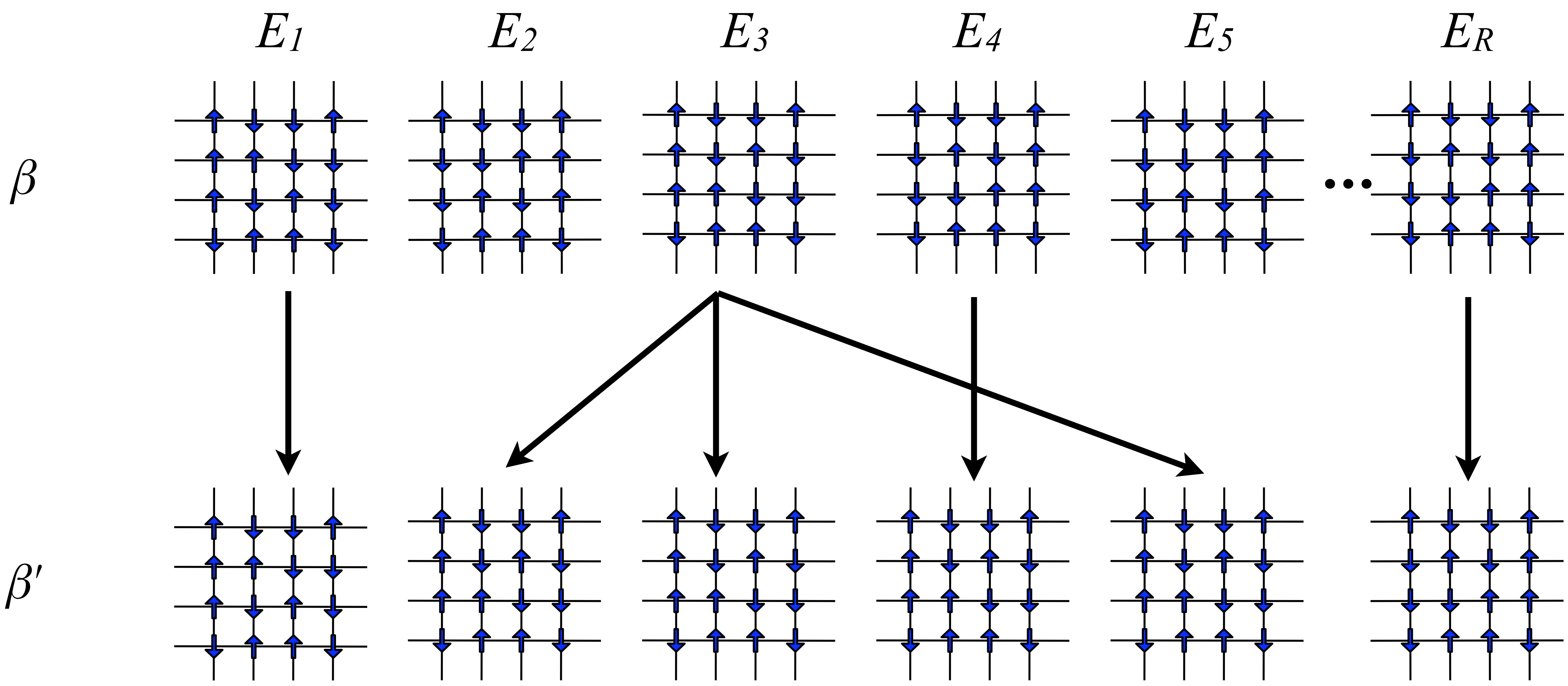


$$\tau_j(\beta, \beta') = \frac{\exp\left[-(\beta' - \beta)E_j\right]}{Q(\beta, \beta')}$$

$$Q(\beta, \beta') = \frac{\sum_{j=1}^{R} \exp\left[-(\beta' - \beta)E_j\right]}{R}$$

Replica j is reproduced n_j times where n_j is a Poisson random number with mean τ_j.

Temperature Step



$$\tau_j(\beta, \beta') = \frac{\exp\left[-(\beta' - \beta)E_j\right]}{Q(\beta, \beta')}$$

$$Q(\beta, \beta') = \frac{\sum_{j=1}^{R} \exp\left[-(\beta' - \beta)E_j\right]}{R}$$

Replica j is reproduced n_j times where n_j is a Poisson random number with mean τ_j.

Strengths of Population Annealing

- Massively parallel
- Direct measurement of the free energy

$$-\beta_k F(\beta_k) = \sum_{\ell=k}^{K} \ln Q(\beta_{\ell+1}, \beta_\ell) + \beta_K F(\beta_K) \qquad Q(\beta, \beta') = \frac{\sum_{j=1}^{R} \exp\left[-(\beta' - \beta) E_j\right]}{R}$$

- Comparable efficiency to parallel tempering for Ising spin glasses.

Weighted Averages

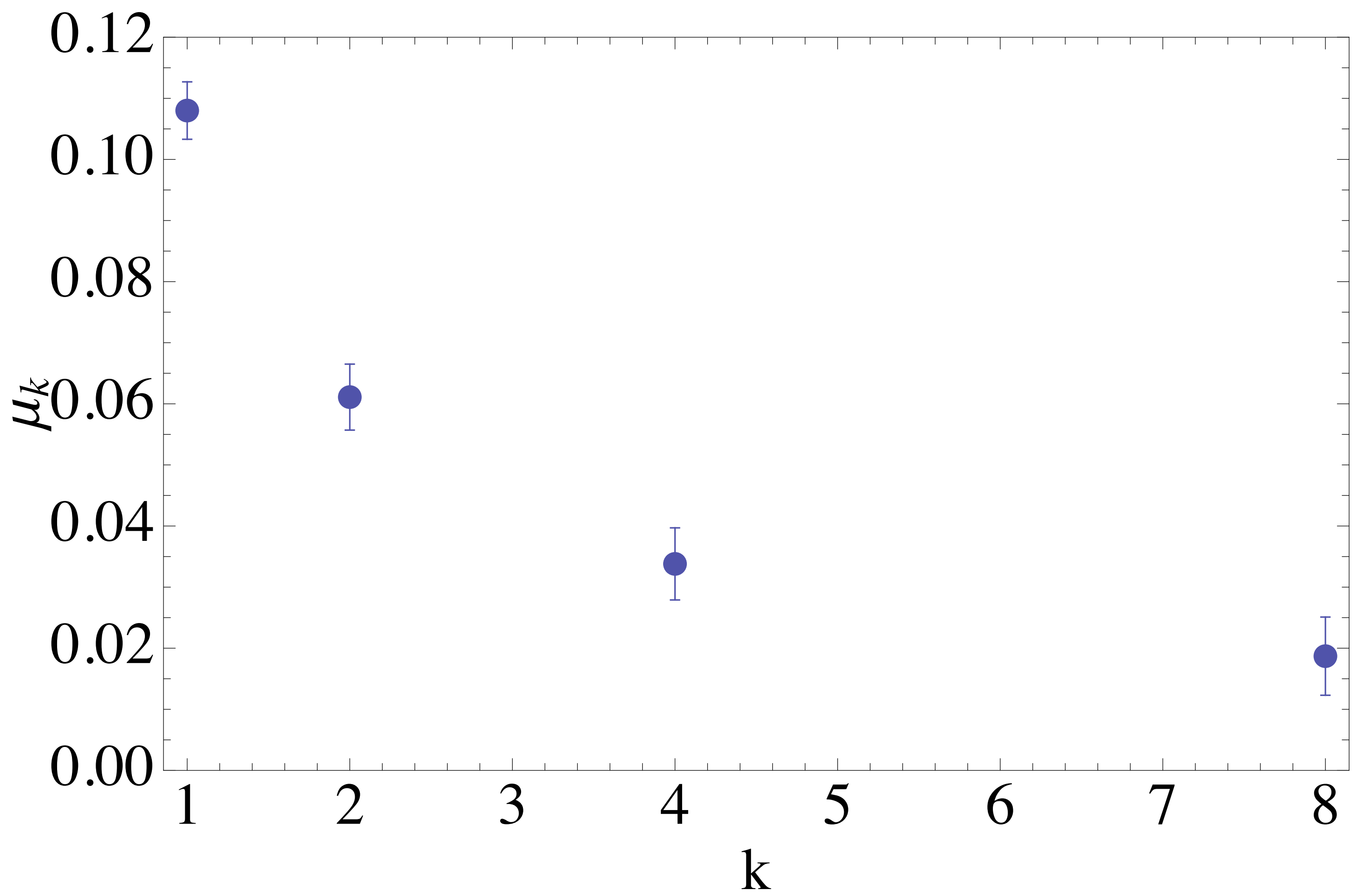
JM, arXive:1006.0252

- Results using small number of replicas are biased.
- Results from independent runs can be combined and biases reduced using weighted averages.
- Observables from each run weighted by the exponential of the free energy estimator for that run.

$$\overline{A}(\beta) = \sum_{m=1}^{M} \tilde{A}_m(\beta)\omega_m(\beta) \qquad \omega_m(\beta) = \frac{e^{-\beta \tilde{F}_m(\beta)}}{\sum_{i=1}^{M} e^{-\beta \tilde{F}_i(\beta)}}$$

- Number of runs needed for unbiased results determined by variance of the free energy estimator--an intrinsic measure of equilibration.

bias

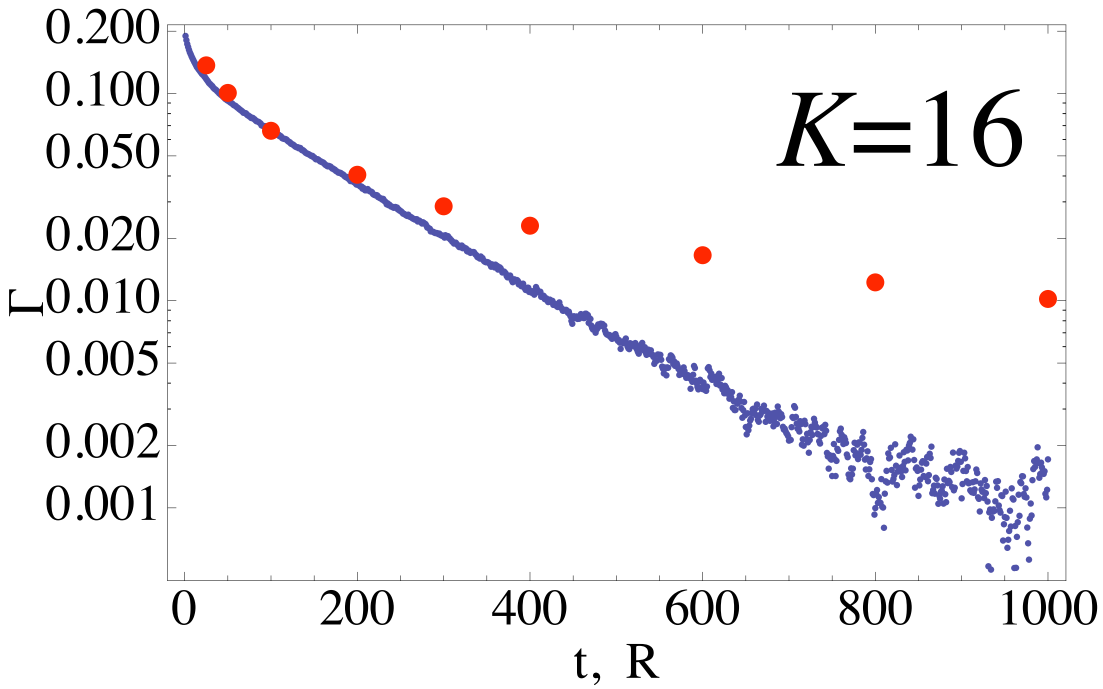


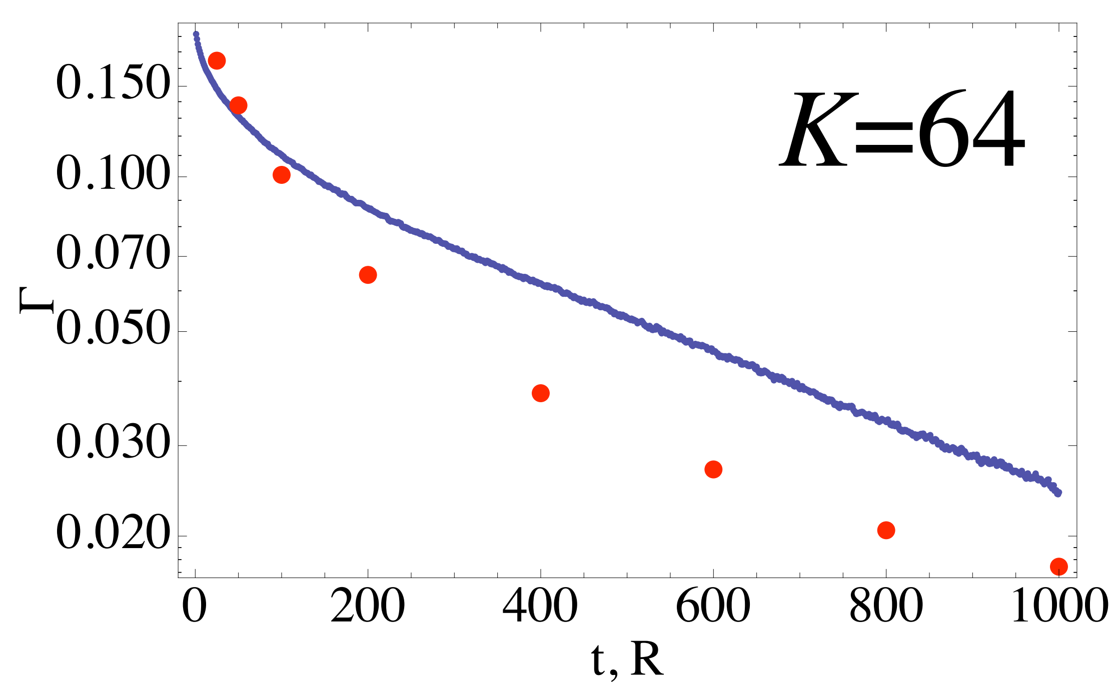
number of runs

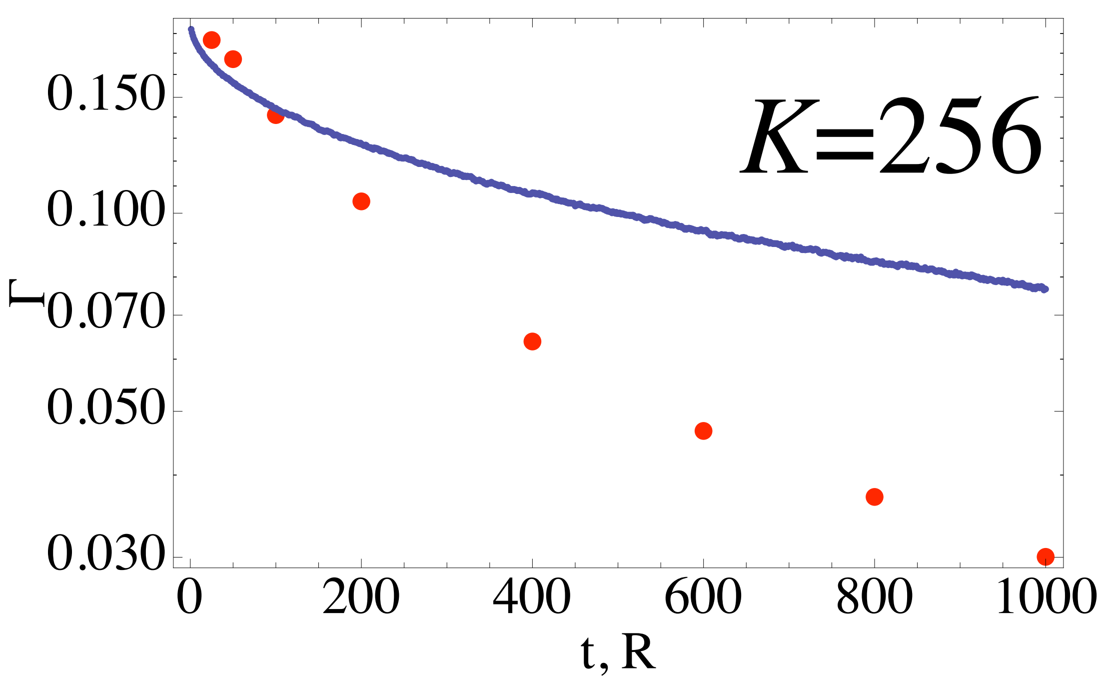
Bias in the energy for the 1D Ising spin glass as a function of ensemble size.

Population Annealing		**Parallel Tempering**
Sequential Monte Carlo		Markov Chain Monte Carlo
# Replicas (R) *space*	$\approx$	#Sweeps (t) *parallel time*
#Temperature steps (T) *parallel time*	$\approx$	#Replicas (R) *space*

Compare PA and PT for the Two-Well Model







● = PA

— = PT

H=0.1

$$\Gamma = \langle \mathrm{Prob}\left[\sigma = +1\right] \rangle - \langle \mathrm{Prob}\left[\sigma = +1\right] \rangle_{\mathrm{eq}}$$

Population Annealing		**Parallel Tempering**
Sequential Monte Carlo		Markov Chain Monte Carlo
# Replicas (R) *space*	$\approx$	#Sweeps (t) *parallel time*
#Temperature steps (T) *parallel time*	$\approx$	#Replicas (R) *space*
$(A - A_{eq}) \approx R_0/R$		$(A - A_{eq}) \approx e^{(-t/\tau)}$
$R_0 \sim K^{1/2}$ $T \sim K^{1/2}$		$\tau \sim K$ $R \sim K^{1/2}$

Conclusions

- Both parallel tempering and population annealing are efficient at finding equilibria between free energy minima with large basins of attraction separated by large barriers.
  - For PT efficiency is greatest when the minima are highly asymmetric
- Neither method is useful, except via brute force parallelism, for the case of free energy minima with small basins of attraction.
- Population annealing is comparably efficient to parallel tempering and has several features to recommend it:
  - Highly parallel
  - Direct measurement of free energy
  - Built-in equilibration criteria (variance of the free energy estimator)