Trustworthy Neural Sampler

Tae-Geun Kim

Guo-Liang Ma
Fudan University

Lingxiao Wang
RIKEN iTHEMS & U Tokyo

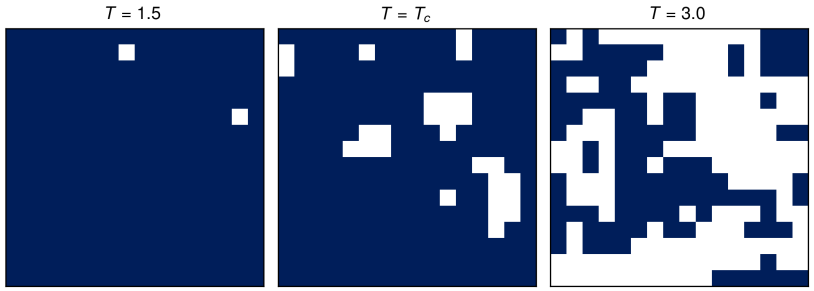
Bridging Machine Learning and Statistical Field Theory

ECT*, Villa Tambosi, Trento

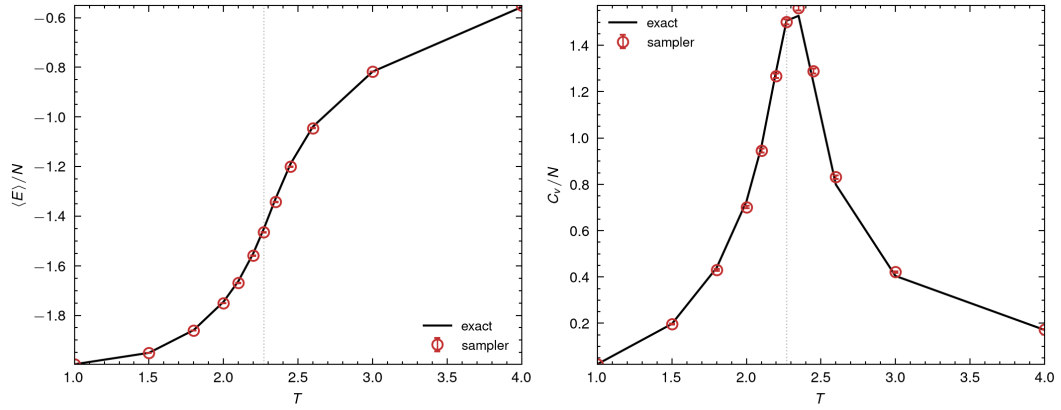
2026.09.09

axect.github.io

A trained neural sampler

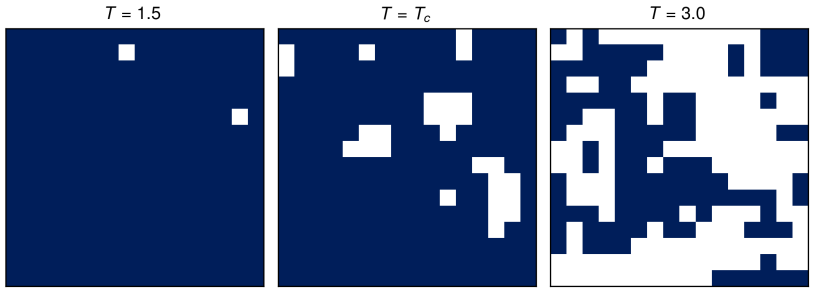


2D Ising, $L = 16$, periodic. One sample per T.



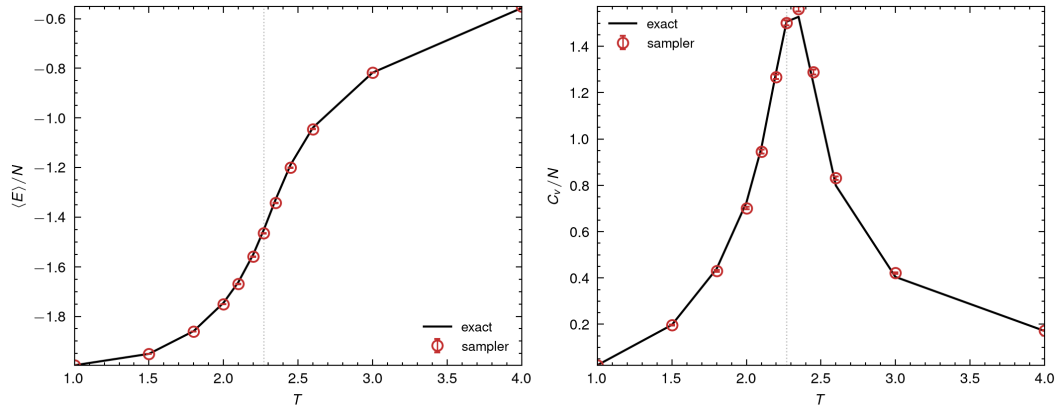
Black: exact. $n = 32768$ per T.

A trained neural sampler



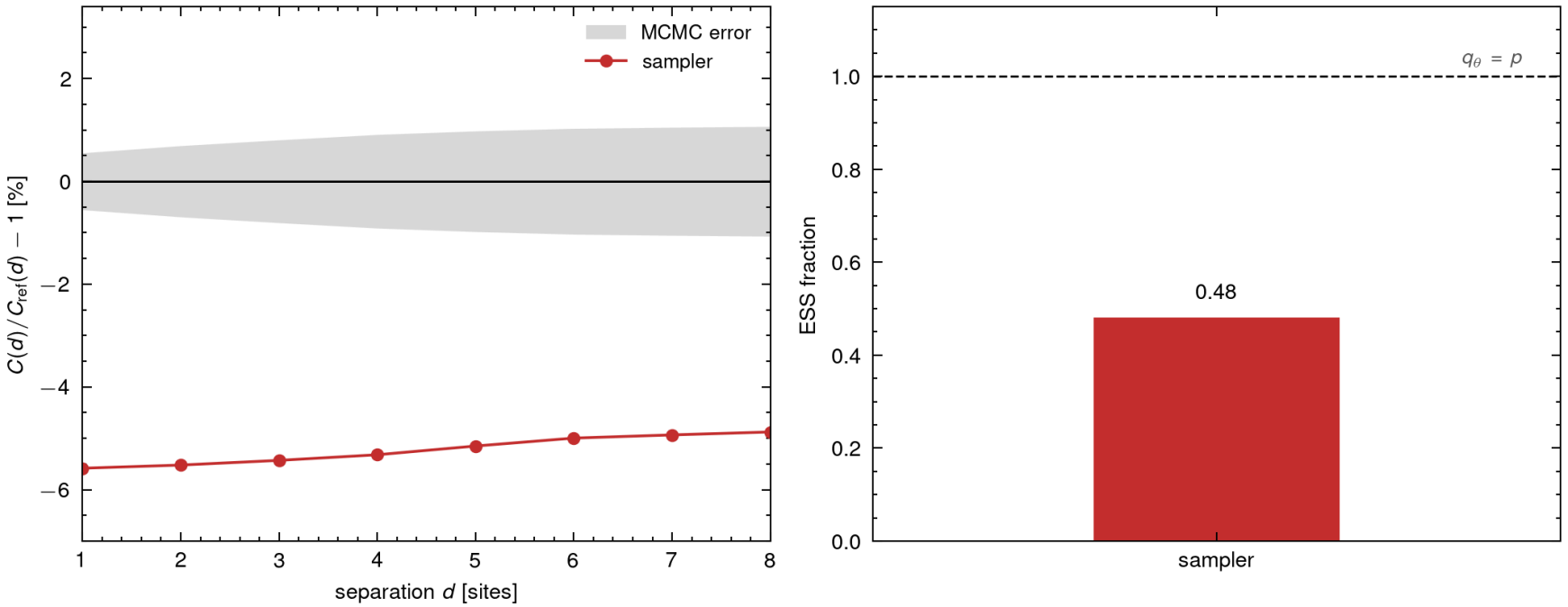
2D Ising, $L = 16$, periodic. One sample per T.

Can you trust it?



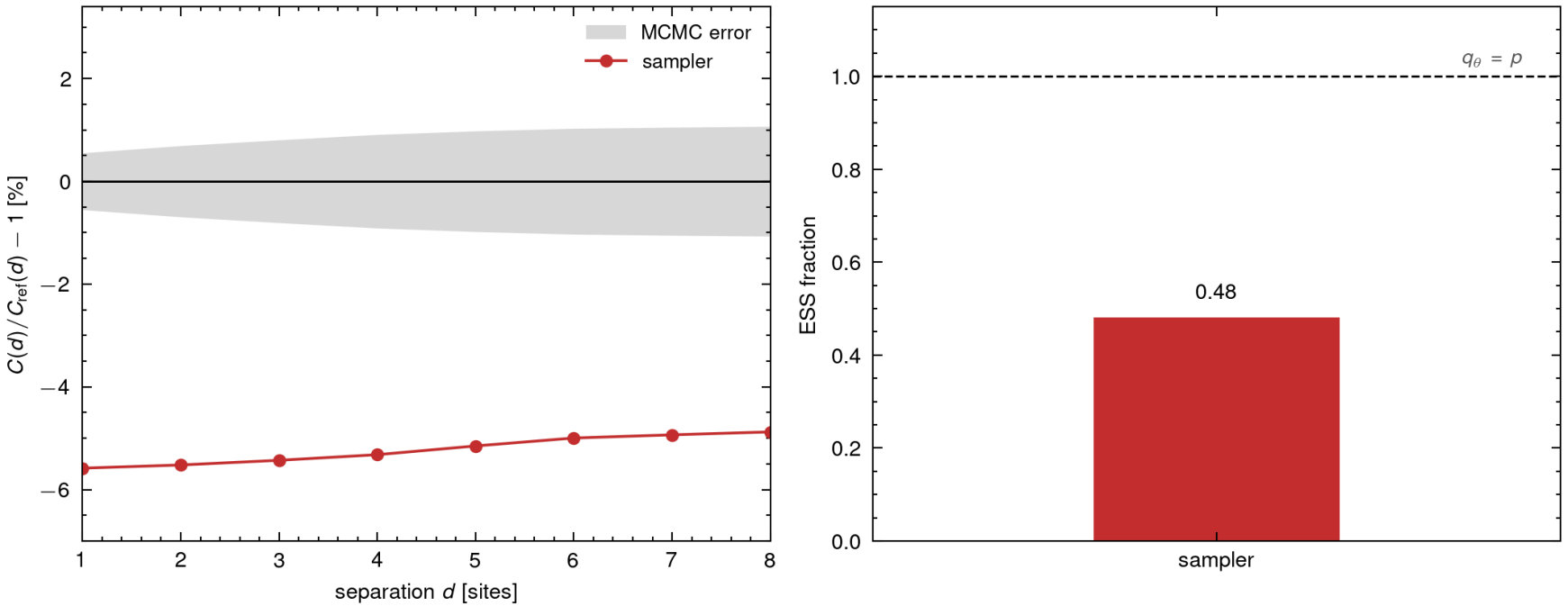
Black: exact. $n = 32768$ per T.

No



$T = T_c$. Left: connected correlator vs Swendsen-Wang, band = MCMC error. Right: ESS of $e^{-\beta H}/q_\theta$.

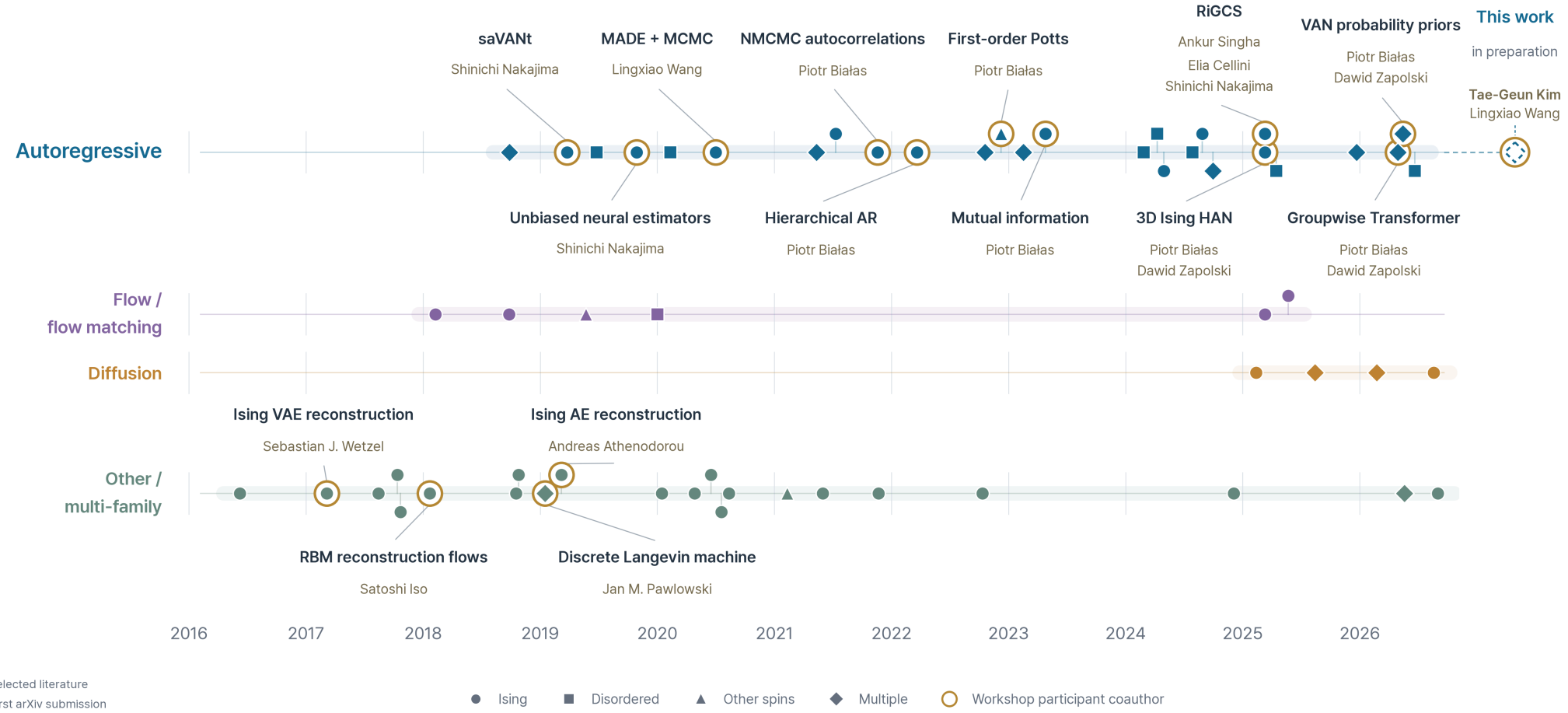
No



$T = T_c$. Left: connected correlator vs Swendsen-Wang, band = MCMC error. Right: ESS of $e^{-\beta H}/q_\theta$.

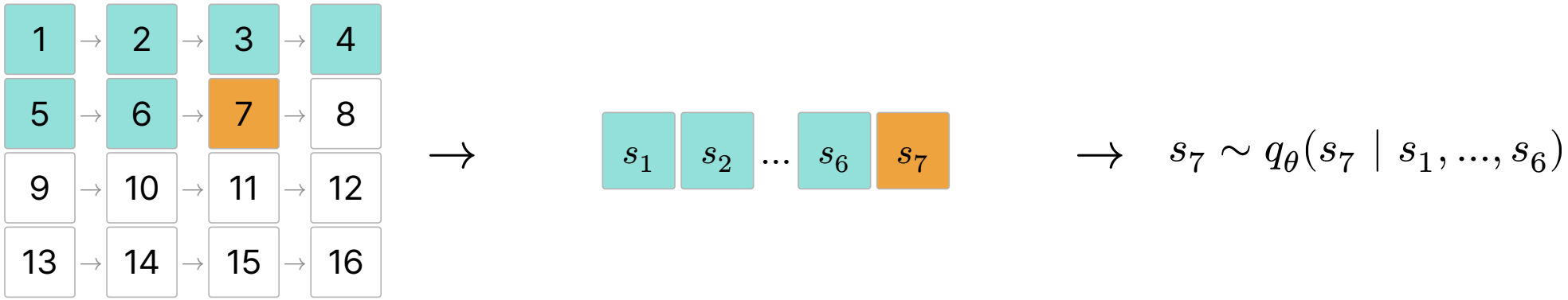
Failed to learn, or unable to learn?

Neural sampling on spin lattices



Autoregressive sampling

Generation order → **Spin sequence** → **Next spin**



One possible generation order. Teal: generated spins. Orange: next spin.

$$q_\theta(\boldsymbol{s}) = \prod_{t=1}^{N} q_\theta(s_t \mid s_1, \ldots, s_{t-1}), \quad \boldsymbol{s} = (s_1, \ldots, s_N)$$

s_t: the spin at position t in the chosen generation order.

Variational training

For data-free training of diffusion models, see the Nakajima group's talk.

$$p_\beta(\boldsymbol{s}) = \frac{e^{-\beta H(\boldsymbol{s})}}{Z(\beta)}, \quad \beta = \frac{1}{k_B T}$$

$$\mathcal{L}_\beta(\theta) = \mathop{\mathbb{E}}_{\boldsymbol{s}\sim q_\theta}\left[\log q_\theta(\boldsymbol{s}) + \beta H(\boldsymbol{s})\right] = D_{\mathrm{KL}}\left(q_\theta \,\|\, p_\beta\right) - \log Z(\beta)$$

VAN Wu, Wang & Zhang, PRL (2019)

Fixed temperature

$q_\theta(\boldsymbol{s}), \quad \min_\theta \mathcal{L}_\beta(\theta)$

VaTD Li, Zhang & Pan, PRL (2025)

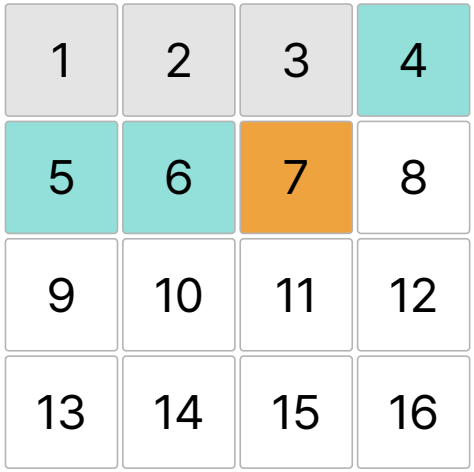
Temperature-conditioned model

$q_\theta(\boldsymbol{s} \mid \beta), \quad \min_\theta \mathop{\mathbb{E}}_{\beta\sim\rho}\left[\mathcal{L}_\beta(\theta)\right]$

ρ: training distribution over inverse temperature.

Training uses H and samples from q_θ; neither reference configurations nor Z is required.

Does the omitted past matter?



Target conditional

$$p_\beta(s_7 \mid s_1, \ldots, s_6)$$

Model conditional

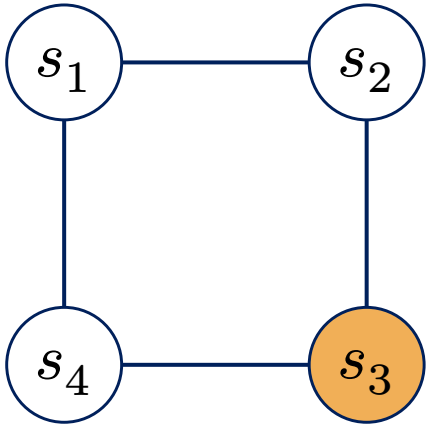
$$q_\theta(s_7 \mid s_4, s_5, s_6)$$

Spins 1 to 3 are generated, but not available as inputs.

Can the model match the target without s_1, s_2, s_3?

Spatial vs sequential conditioning

$$H = -J(s_1 s_2 + s_2 s_3 + s_3 s_4 + s_4 s_1), \quad K = \beta J$$



Generate: $1 \to 2 \to 3 \to 4$

$s_i \in \{-1, +1\}$

Spatial: condition on s_4

$$p(s_3 \mid s_1, s_2, s_4) = p(s_3 \mid s_2, s_4)$$

Sequential: sum over s_4

$$p(s_3 \mid s_1, s_2) \propto e^{K s_2 s_3} \sum_{s_4 = \pm 1} e^{K s_4 (s_1 + s_3)}$$

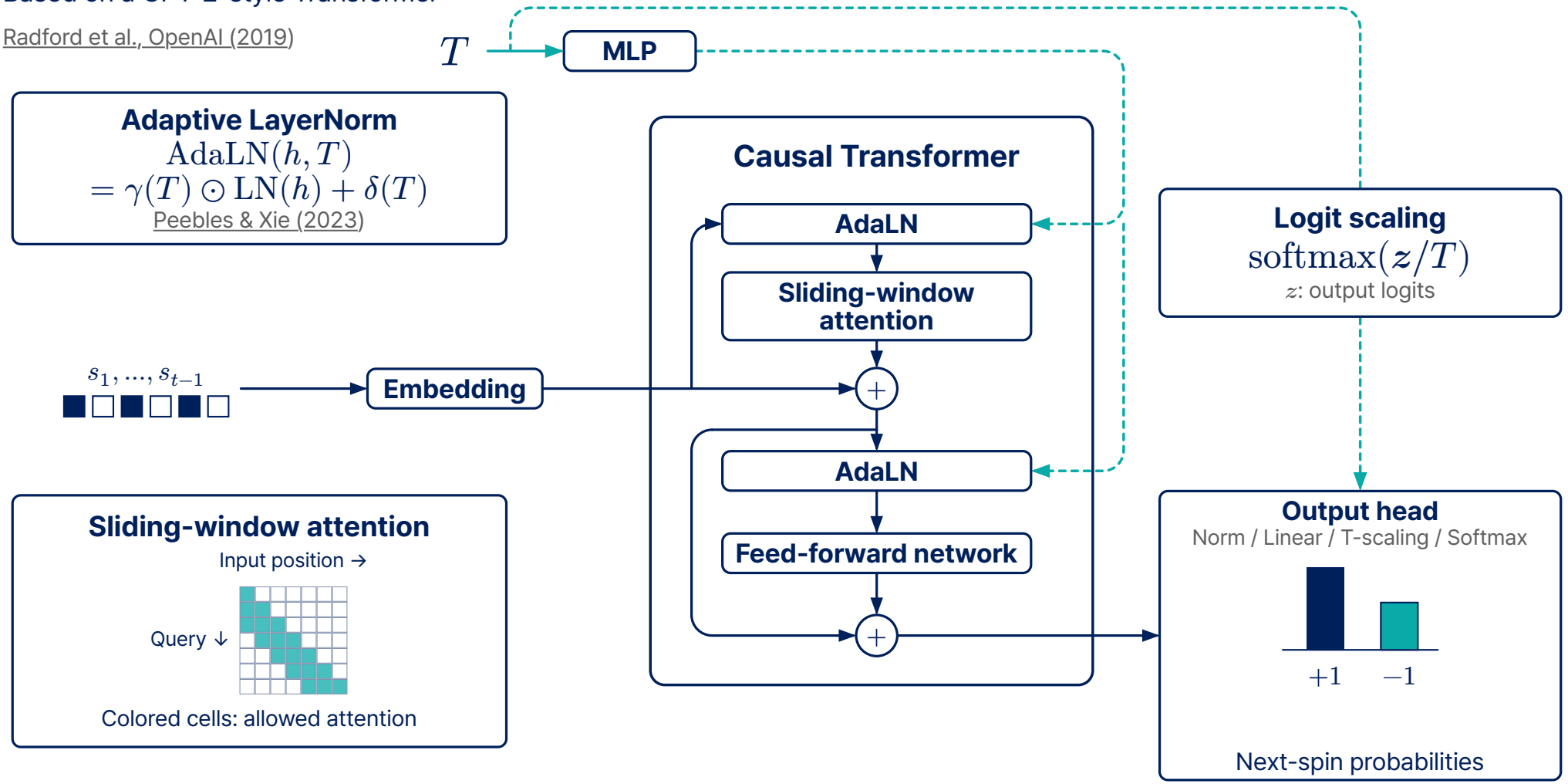
$$= e^{K s_2 s_3} 2 \cosh(K(s_1 + s_3))$$

The spatial Markov property does not imply a first-order Markov chain in generation order.

LatticeGPT

Based on a GPT-2-style Transformer

Radford et al., OpenAI (2019)



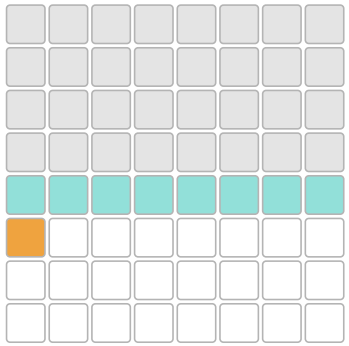
Controlling the reach

$$r = \ell(w-1) + 1$$

w: attention window ℓ: depth

$r = 8$

$(w = 8, \ell = 1)$

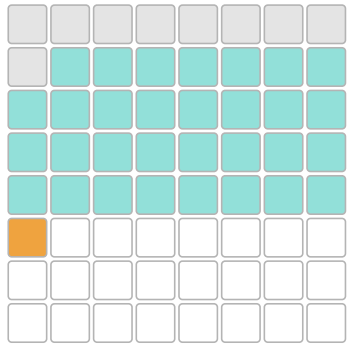


Increase reach

$\rightarrow$

$r = 31$

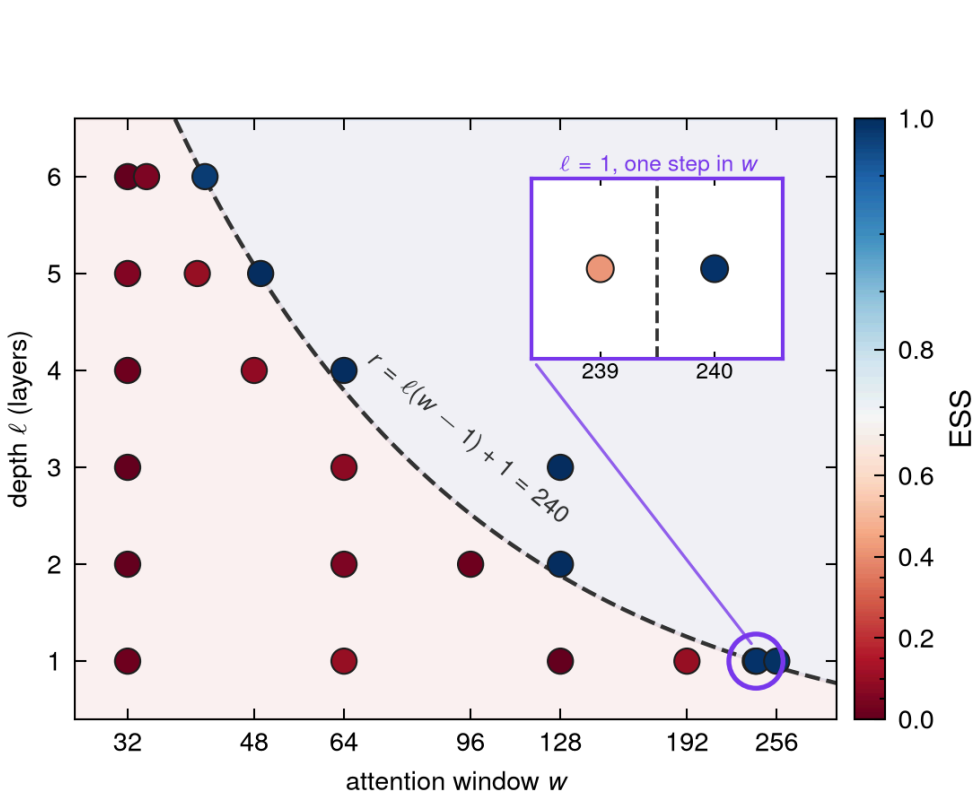
$(w = 16, \ell = 2)$

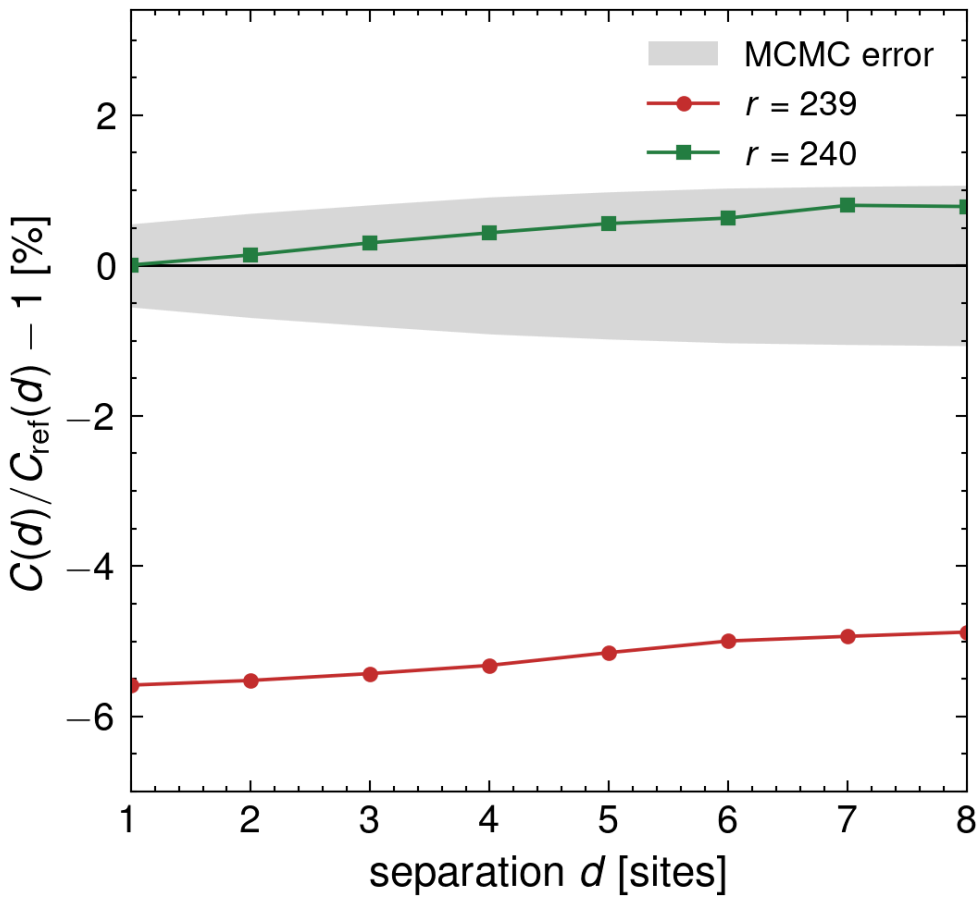


Current spin Within reach Outside reach Not yet drawn

Reach: the number of preceding spins that can affect the next-spin probability.

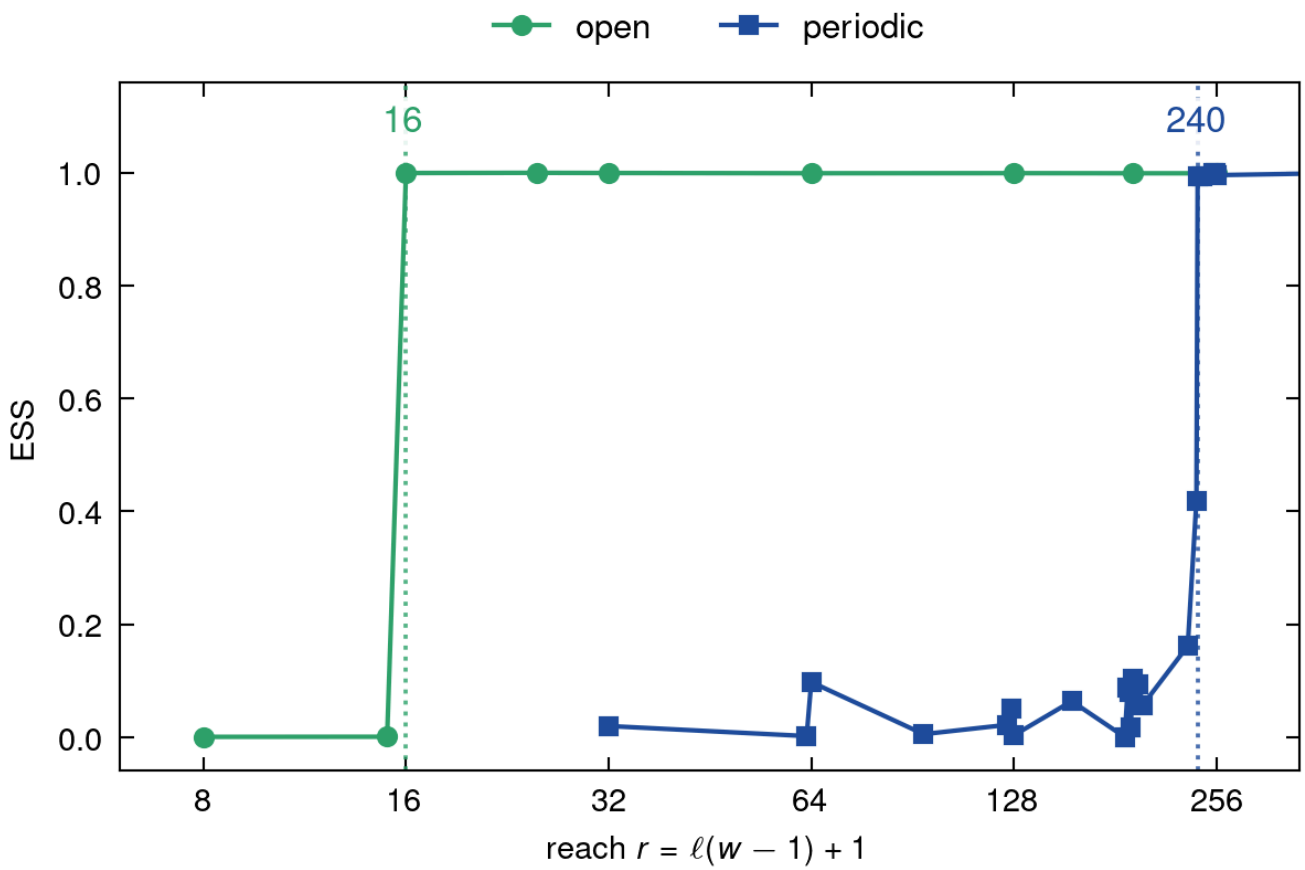
Varying window and depth





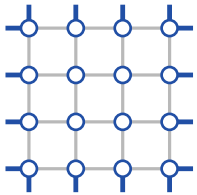
Ising, $L = 16$, periodic, raster order, $T = T_c$. Right: $r = 239$ and 240, one layer.

Changing the boundary conditions



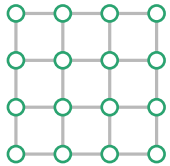
Ising, $L = 16$, raster order, $T = T_c$.

Periodic



↓

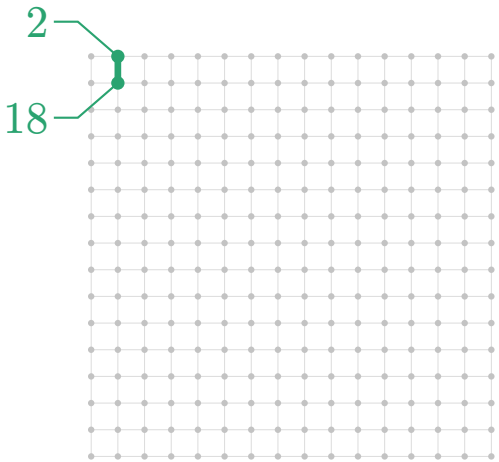
Open



Where do 16 and 240 come from?

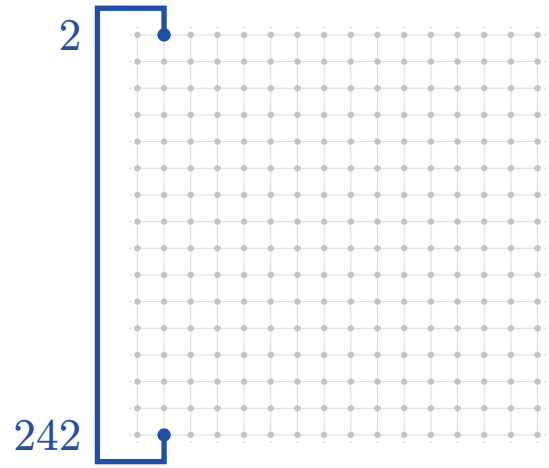
$$H(\boldsymbol{s}) = -J \sum_{\langle i,j \rangle} s_i s_j, \quad s_i \in \{-1, +1\}$$

Open



$$18 - 2 = 16 = L$$

Periodic

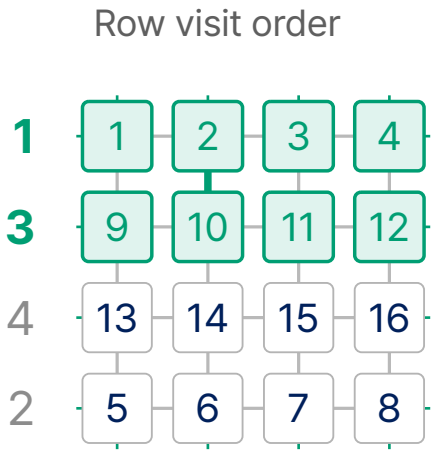


$$242 - 2 = 240 = L(L-1)$$

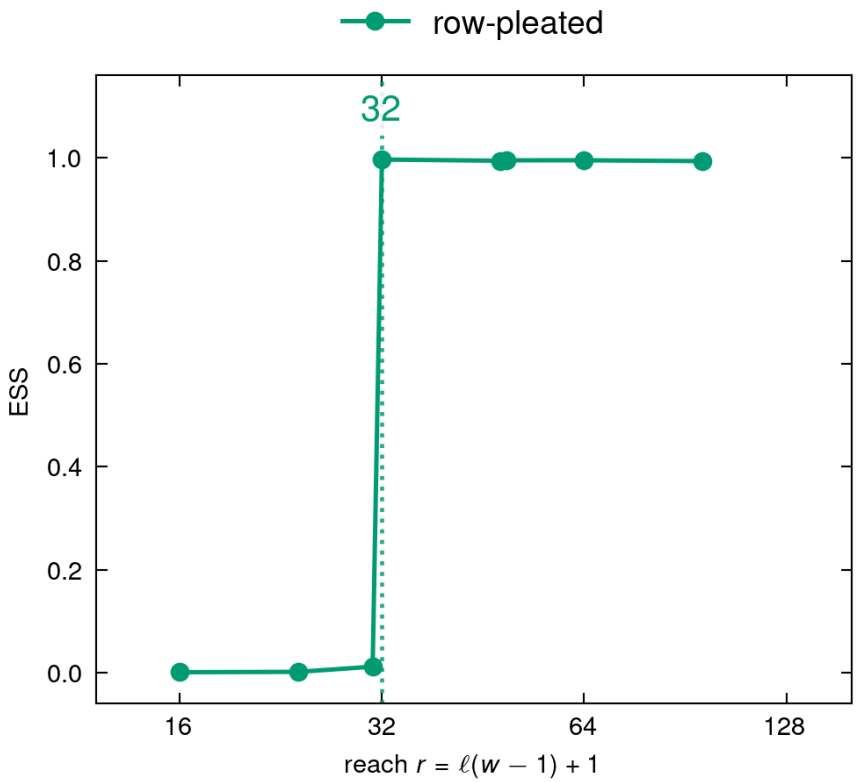
Raster numbering: left to right, top to bottom. Highlighted bonds join nearest neighbours.

Changing the orderings

Visit rows alternately from the top and bottom; traverse each row left to right.



$$(3-1)L = 2L$$



Ising, $L = 16$, periodic, $T = T_c$.

Bandwidth of an ordering

Boundary	Ordering	Largest bond separation
Open	Raster	$L = 16$
Periodic	Raster	$L(L-1) = 240$
Periodic	Row-pleated	$2L = 32$

$$b(\pi) = \max_{\{u,v\} \in E} |\pi(u) - \pi(v)|$$

[Chinn et al., J. Graph Theory (1982)]

$$\pi : V \to \{1, \ldots, N\}, \quad \pi(v) = \text{generation position of site } v$$

$G = (V, E)$: interaction graph of the Hamiltonian.

Reach and bandwidth

Theorem 1 (Bandwidth threshold)

For a given generation order, let the target be a strictly positive, finite-state pairwise Boltzmann distribution. No edge interaction can be absorbed into on-site terms.
Allowing arbitrary conditional distributions, an autoregressive model using only the preceding r spins can represent the target exactly if and only if

$$r \geq b(\pi)$$

$r = 7 < b(\pi)$



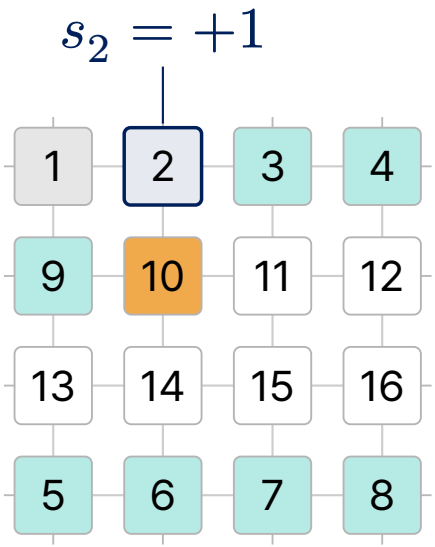
$r = 8 = b(\pi)$



Representability, not a guarantee of successful training.

Why the omitted spin matters

Choose a longest bond: sites 2 and 10.
Generate spin 10 with $r = 7$; hold all other past spins fixed.

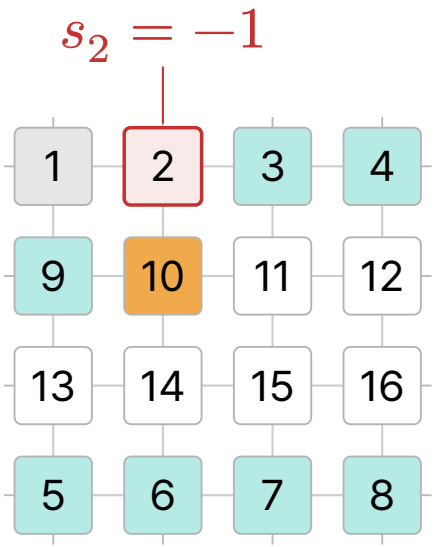


Model

$$q_\theta(s_{10} \mid s_3, \ldots, s_9)$$

Why the omitted spin matters

Choose a longest bond: sites 2 and 10.
Generate spin 10 with $r = 7$; hold all other past spins fixed.



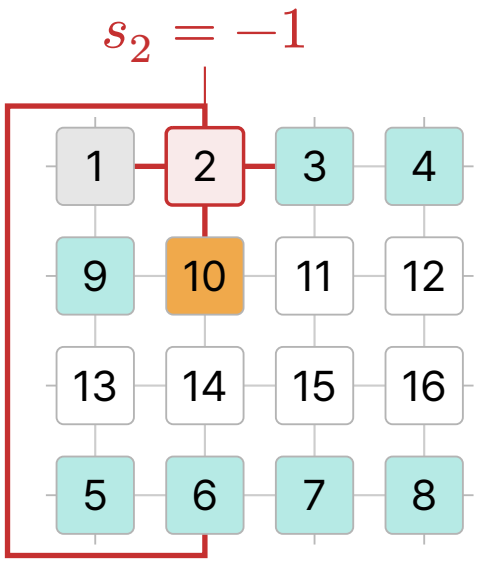
Model

$$q_\theta(s_{10} \mid s_3, \ldots, s_9)$$

Same input. Same prediction.

Why the omitted spin matters

Choose a longest bond: sites 2 and 10.
Generate spin 10 with $r = 7$; hold all other past spins fixed.



Model

$$q_\theta(s_{10} \mid s_3, \ldots, s_9)$$

Same input. Same prediction.

Target

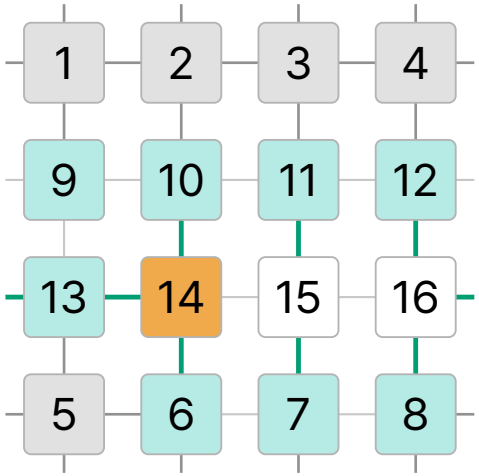
$$e^{+\beta J s_{10}} \longrightarrow e^{-\beta J s_{10}}$$

Different conditional probabilities.

Site 2 cannot couple to a site numbered above 10: the bond's span would exceed 8.
Summing over future spins therefore cannot remove its influence on spin 10.

Why the bandwidth is sufficient

Generating spin 14 with $r = b(\pi) = 8$. Spins 6 to 13 are within reach.

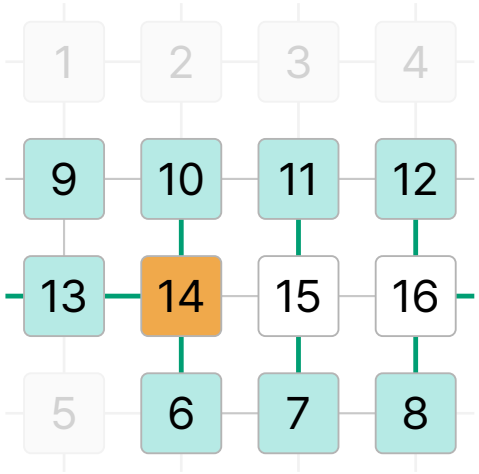


No bond to the undrawn spins

A bond from sites 1–5 to sites 14–16 would have span at least $14 - 5 = 9 > 8$.

Why the bandwidth is sufficient

Generating spin 14 with $r = b(\pi) = 8$. Spins 6 to 13 are within reach.

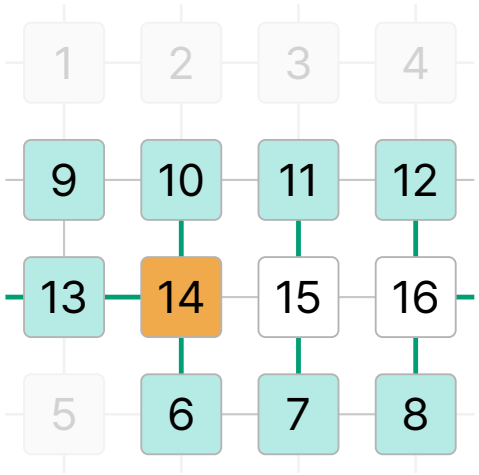


Same weight for every completion

All terms involving sites 1–5
depend only on already generated spins.

Why the bandwidth is sufficient

Generating spin 14 with $r = b(\pi) = 8$. Spins 6 to 13 are within reach.

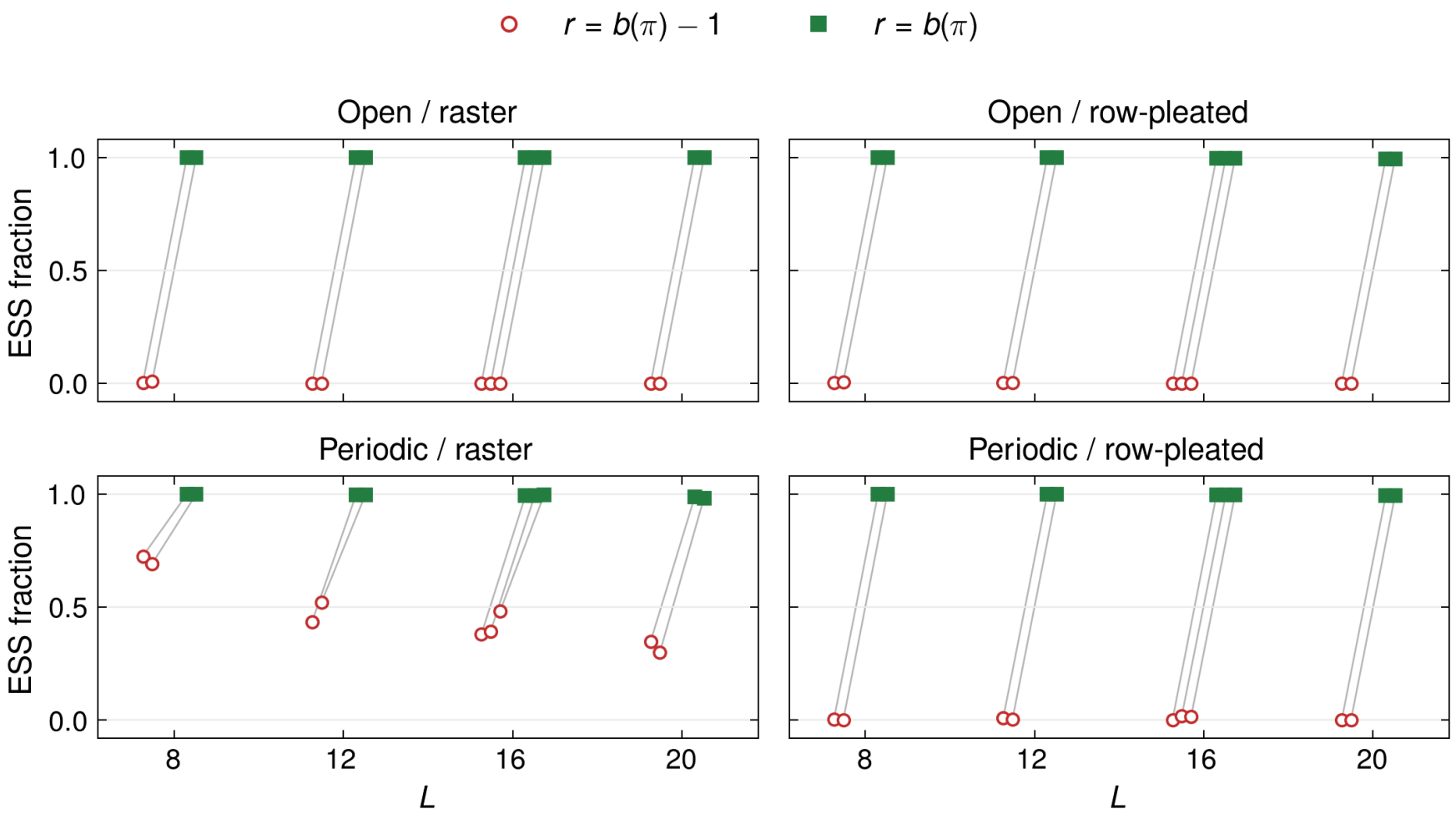


$$p(s_{14} \mid s_1, \ldots, s_{13}) = \frac{\cancel{e^{-\beta H_{\text{out}}}} \sum_{s_{15}, s_{16}} e^{-\beta H_{\text{rest}}}}{\cancel{e^{-\beta H_{\text{out}}}} \sum_{s_{14}, s_{15}, s_{16}} e^{-\beta H_{\text{rest}}}}$$

H_{out}: terms involving sites 1–5.
$H_{\text{rest}} = H - H_{\text{out}}$.

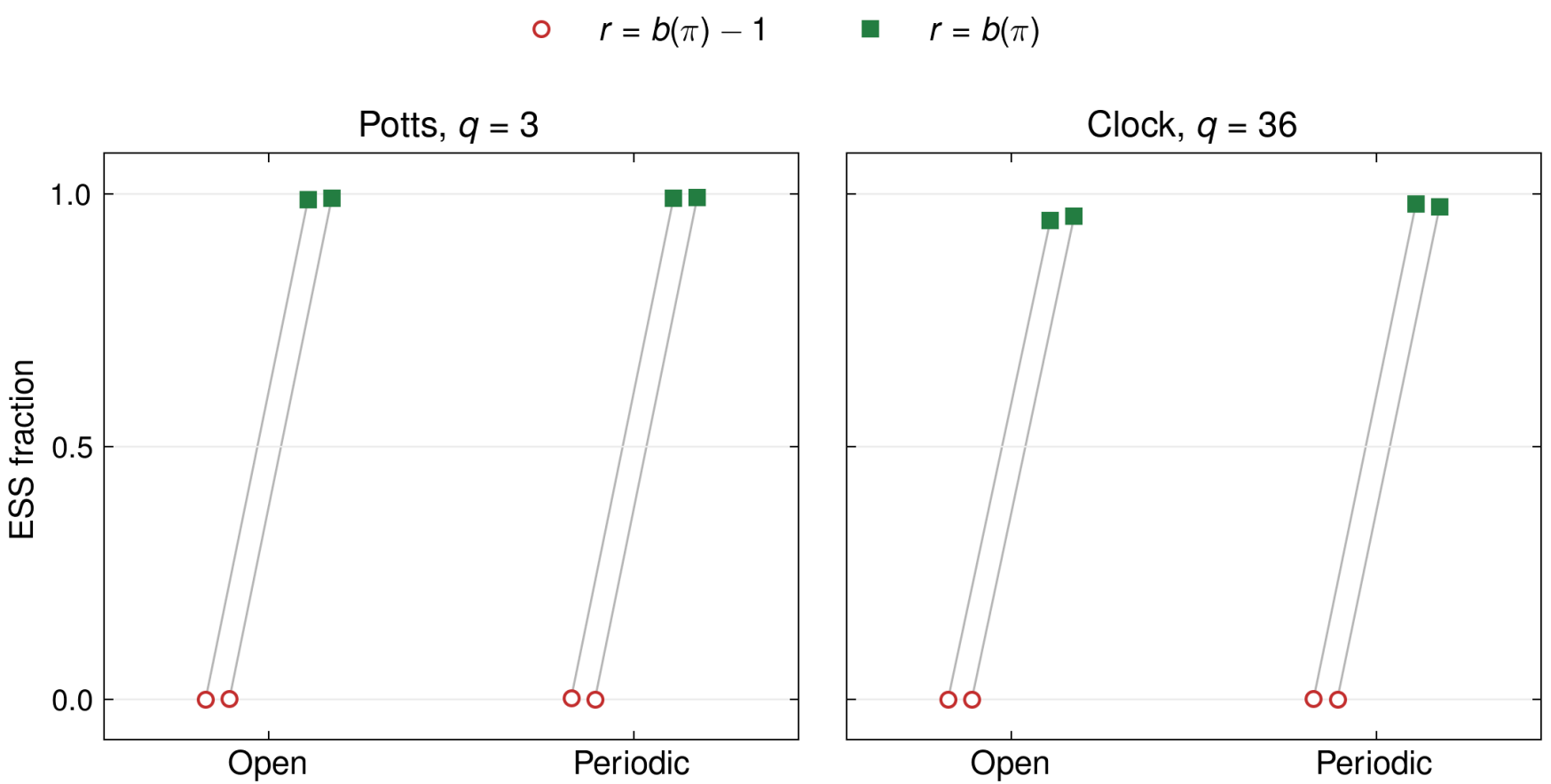
Spins 6–13 are sufficient.

Across boundaries and orderings



Ising, one layer. Each line joins a matched seed pair at $r = b(\pi) - 1$ and $r = b(\pi)$; 36 pairs.

Beyond Ising



$L = 16$, one layer, row-pleated ordering with $b(\pi) = 32$. Each line joins a matched seed pair; 32,768 samples per model.

What can we do with the theorem?

[Chinn et al., J. Graph Theory (1982)]

Before training

$$r < b(\pi)$$

Insufficient reach

Optimize the ordering

$$B(G) = \min_{\pi} b(\pi)$$

Minimize the required reach

Optimal ordering

[Chvátalová (1975); Li, Tao and Shen (1981)]

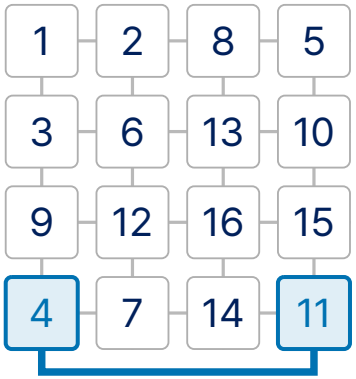
Boundary	G	$B(G)$
Open	$P_L \square P_L$	L
Periodic	$C_L \square C_L$	$2L-1$

P_L: path; C_L: cycle; $\square$: Cartesian product. $L \geq 3$.

Optimal orderings are not unique

[Karakhanyan (1982); Riordan (1998)]

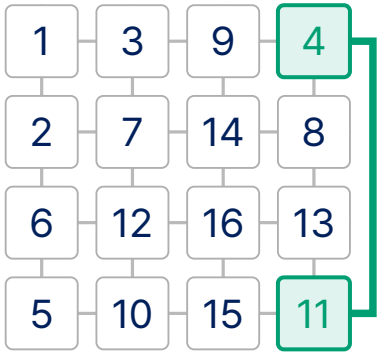
Karakhanyan-Riordan



$$b(\pi_1) = 7$$

Ours

Shell-pleated



$$b(\pi_2) = 7$$

4×4 periodic lattice; $B(C_4 \square C_4) = 7$.

Sampling cost with a shorter context

Measured · $L = 16$ · A40

Preliminary

Global: raster, $w = 256$ · Local: shell-pleated, $w = 31$
One layer, width 192, 549,864 parameters in each model.

Samples	Sampling (ms)		Sampling (MiB)	
	Global	Local	Global	Local
1 at T_c	33.47	30.88	24.37	19.78
512 at T_c	334.85	88.59	221.24	50.48
4096 at T_c	2499.73	545.31	1620.94	269.80
$8T \times 512$	2502.09	545.07	1620.94	269.80

Local training · A40
91.65 min · 334 epochs

Training peak memory
Global 12.072 GiB (RTX 4090) · Local 2.135 GiB (A40)

Estimated · larger lattices

4096 samples · one layer · width 192 · FP32

L	Attention window w		KV cache (GiB)	
	Global: L^2	Optimal: $2L - 1$	Global	Optimal
32	1024	63	6	0.369
64	4096	127	24	0.744
128	16384	255	96	1.494

Optimal: $r = w = B(G)$. KV tensors only, not total GPU memory.

Summary

Can you trust it?

1. **A few matching observables are not enough.**
   Energy and specific heat can hide a correlation error.

2. **Check the reach before training.**
   $r \geq b(\pi)$: exact representability under the stated assumptions, not a training guarantee.

3. **Optimize the ordering to reduce sampling cost.**
   $B(G) = \min_\pi b(\pi)$: shorter context and a smaller KV cache.

Code coming soon!

github.com/Axect

Public release with the paper.