Can you trust your neural samplers?

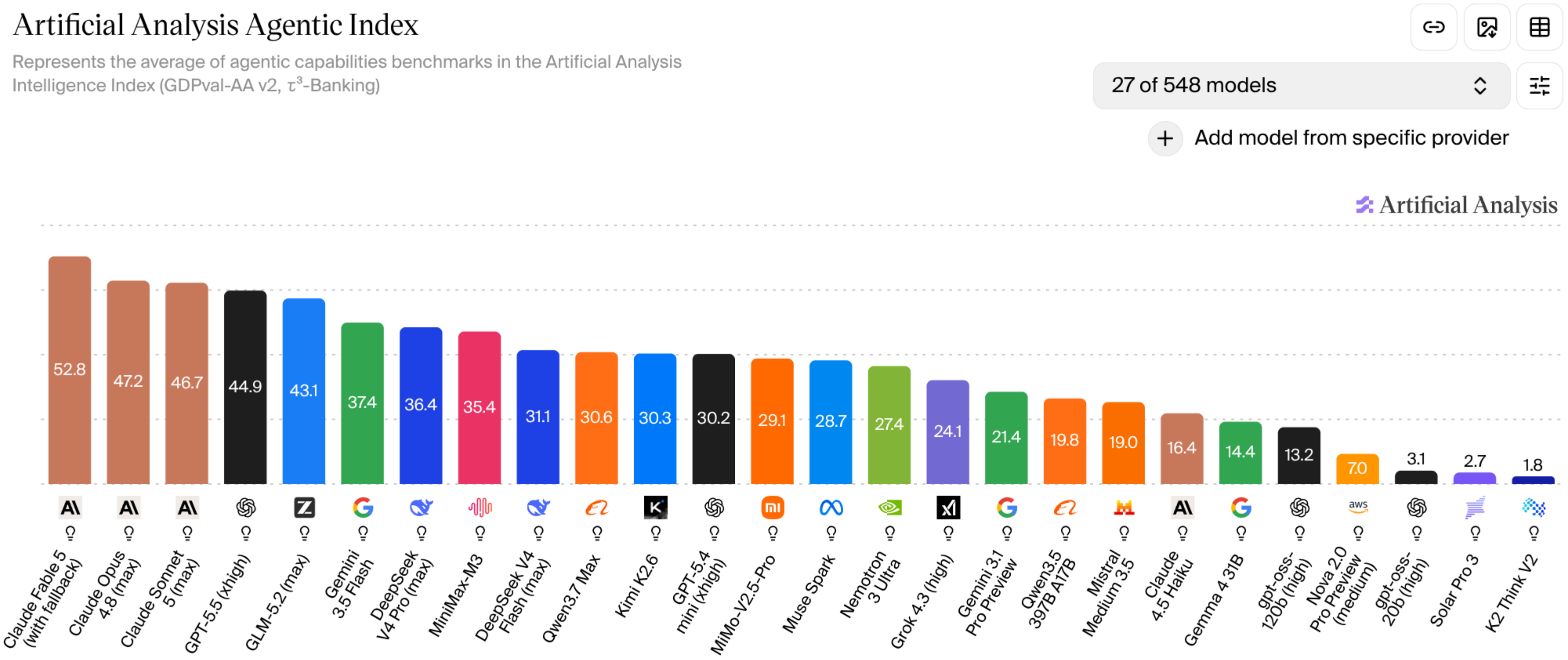
Tae-Geun Kim (金泰瑾)

With **Guo-Liang Ma** (Fudan University) · **Lingxiao Wang** (RIKEN iTHEMS & UTokyo)

AI Agents for Nuclear Physics Workshop

2026.07.05

The age of AI agents

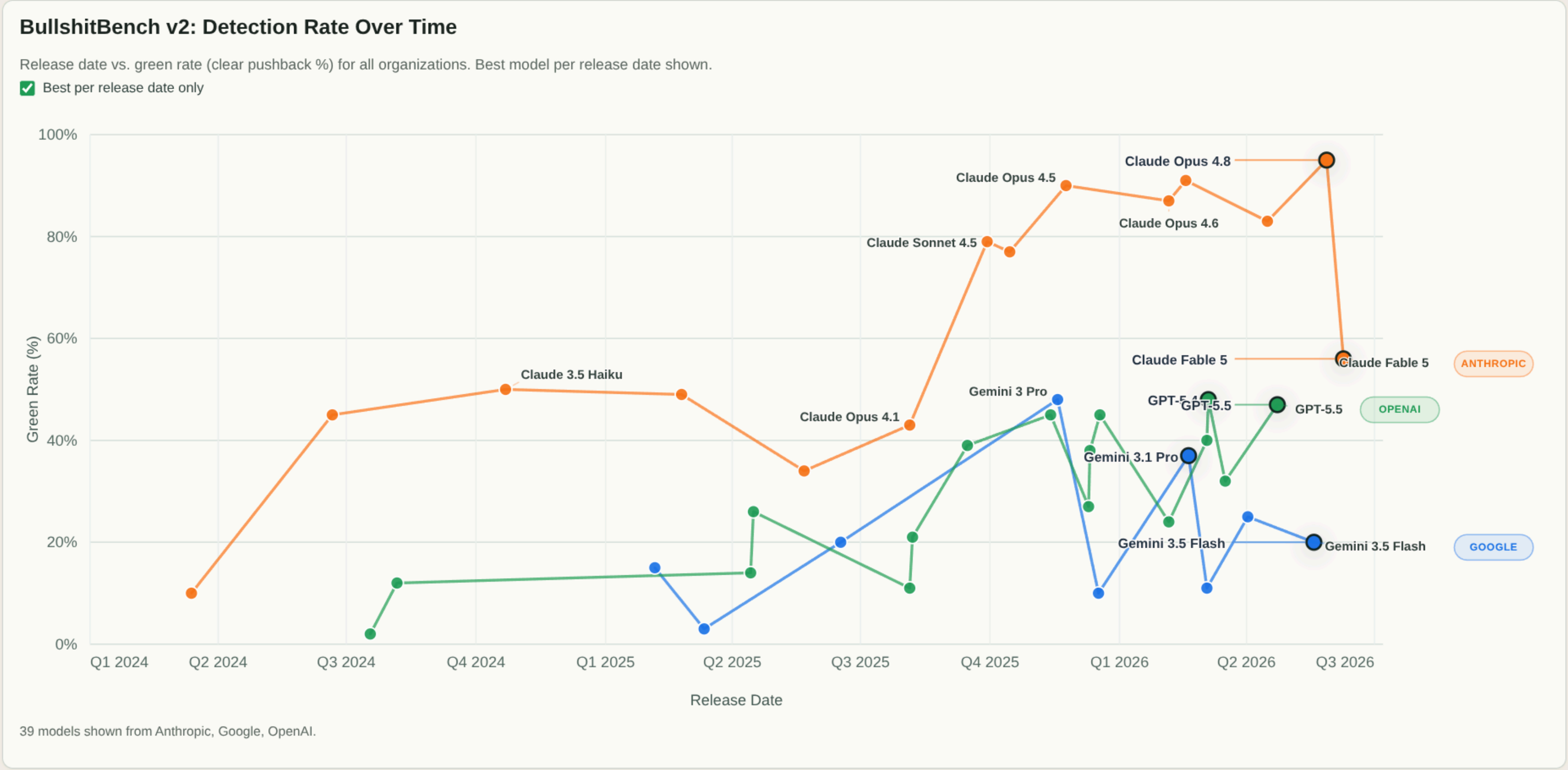


Artificial Analysis Agentic Index (average of agentic-capability benchmarks). Source: artificialanalysis.ai.

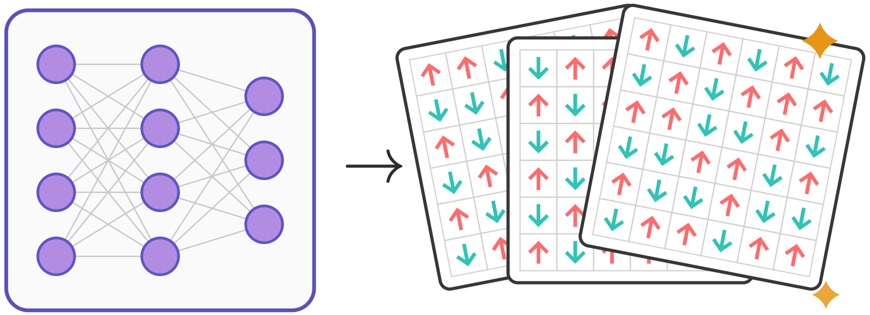
We hand them more and more of the work. **But can we trust what they give back?**

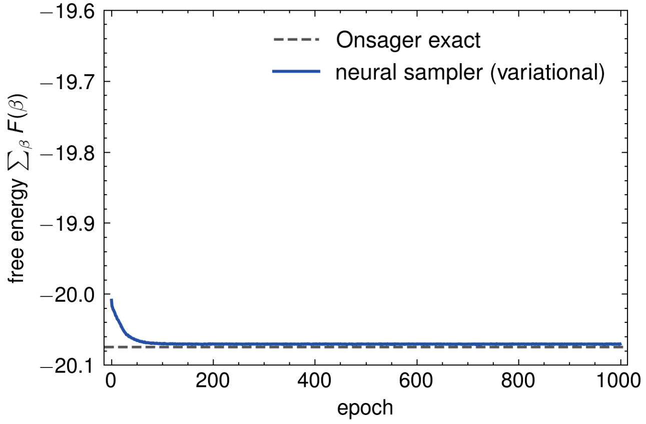
Can you trust them?

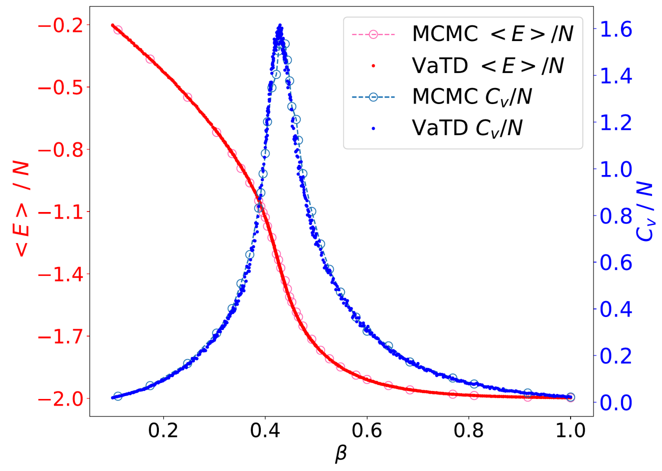
[github.com/petergpt/bullshit-benchmark]



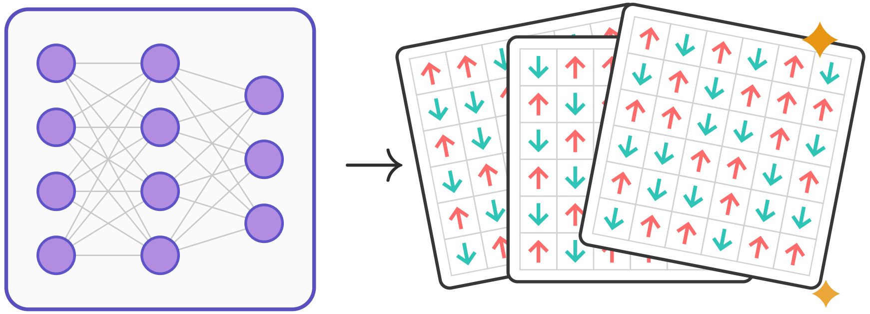
Let's see a neural sampler

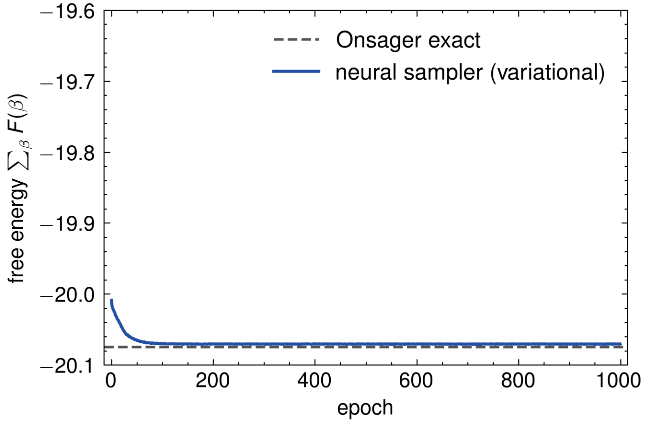






[Li, Zhang, Pan, PRL (2025)]

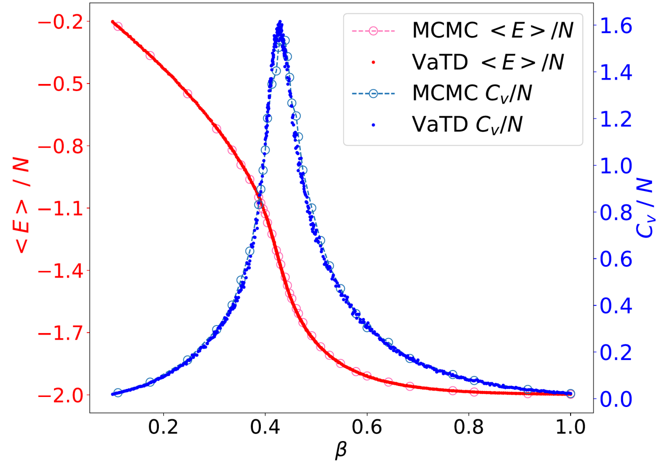




THE TWO QUESTIONS OF THIS TALK

1 Which test actually tells you a sampler is **trustworthy?**

2 And how, when there is **no reference** to check against?



[Li, Zhang, Pan, PRL (2025)]

Boltzmann distribution

A system in equilibrium at temperature T is governed by the **Boltzmann distribution**:

$$P(s) = \frac{e^{-\beta H(s)}}{Z}, \quad Z = \sum_s e^{-\beta H(s)}, \quad \beta = 1/T$$

Every thermodynamic quantity descends from the **partition function** Z:

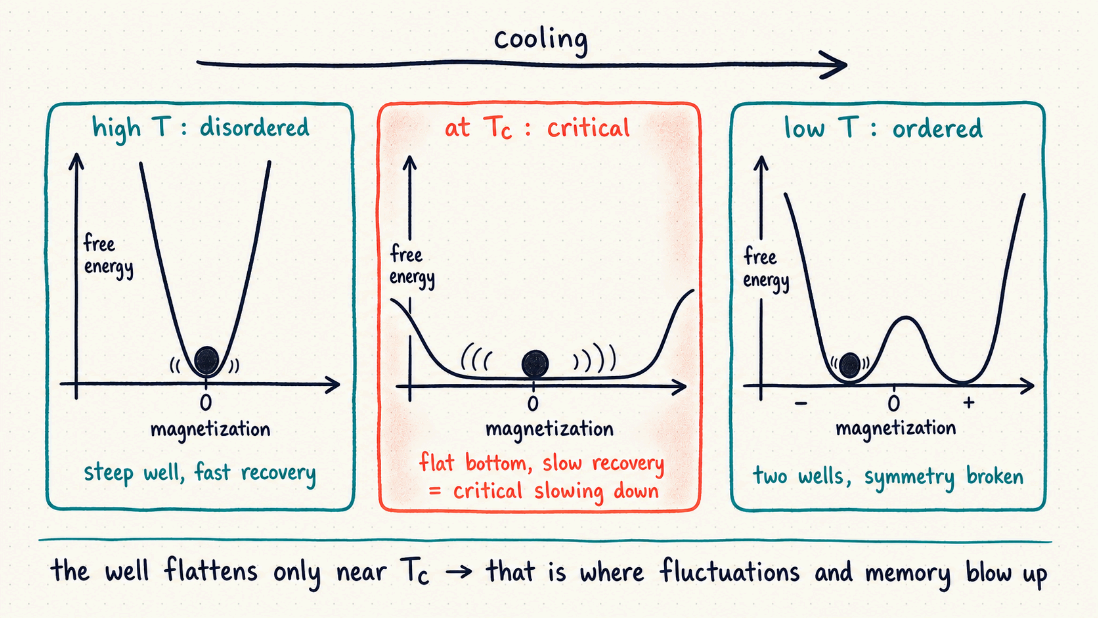
$$\langle E \rangle = -\frac{\partial \ln Z}{\partial \beta} \qquad C_v = \beta^2 \frac{\partial^2 \ln Z}{\partial \beta^2} \qquad F = -\beta^{-1} \ln Z$$

But Z sums over all 2^N states: 16 for a 2×2 lattice, but $2^{256} \approx 10^{77}$ for $L = 16$

Z can never be enumerated. **We must sample from $P(s)$.**

[Metropolis et al., J. Chem. Phys. (1953); Hastings, Biometrika (1970)]

- **Markov chain Monte Carlo (MCMC)** samples $P(s)$ from ratios alone, $P(s')/P(s) = e^{-\beta \Delta H}$, so it **never needs** Z. The trusted workhorse.



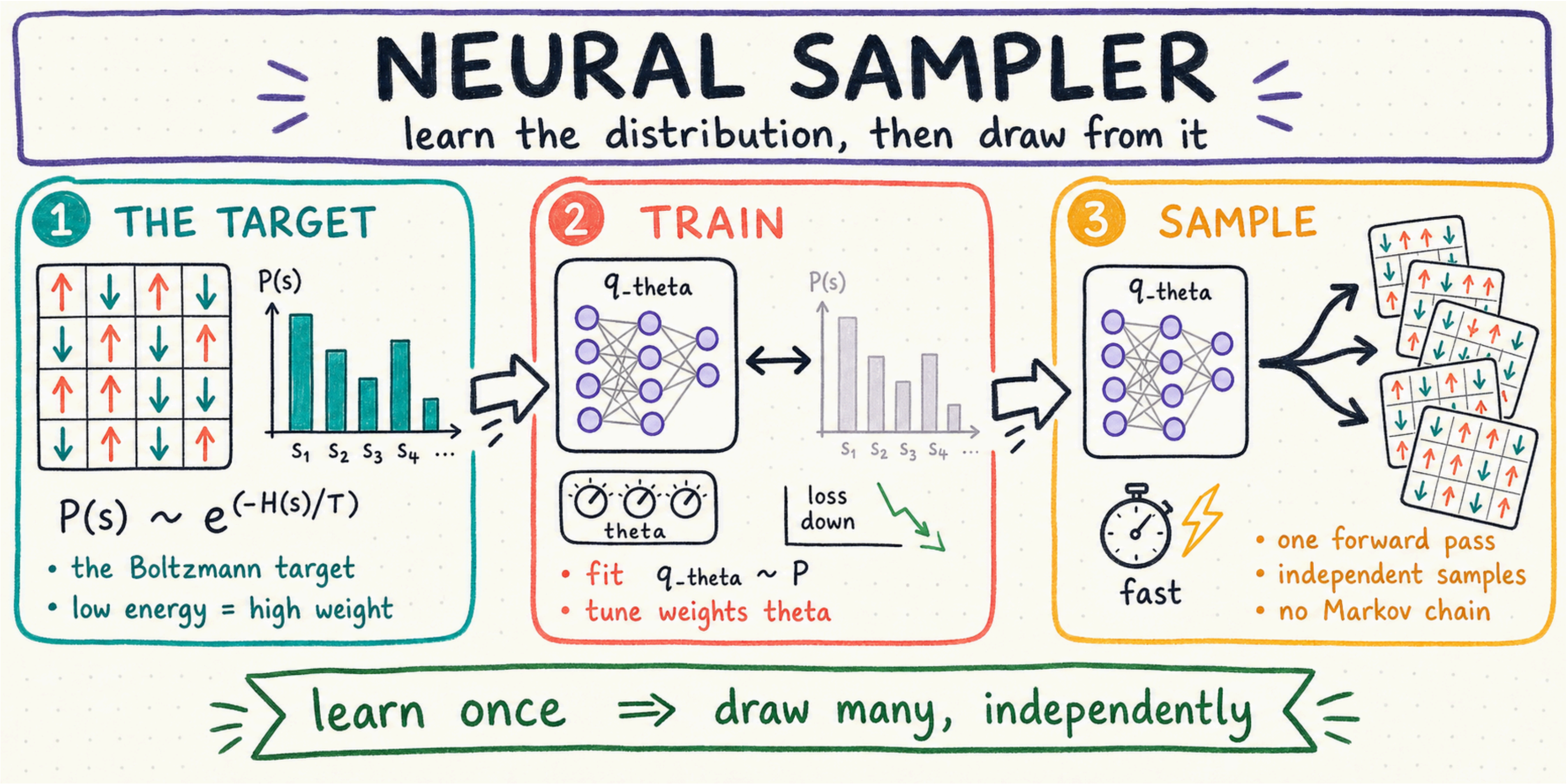
- **But: critical slowing down**

  Near T_c the basin flattens, the state random-walks, and the autocorrelation time diverges:

$$\tau \sim \xi^z \to \infty.$$

- Independent samples become astronomically costly **exactly at the transition we study**.

Cluster algorithms beat CSD but are hand-crafted per model, not universal. [Swendsen-Wang, PRL (1987); Wolff, PRL (1989)]

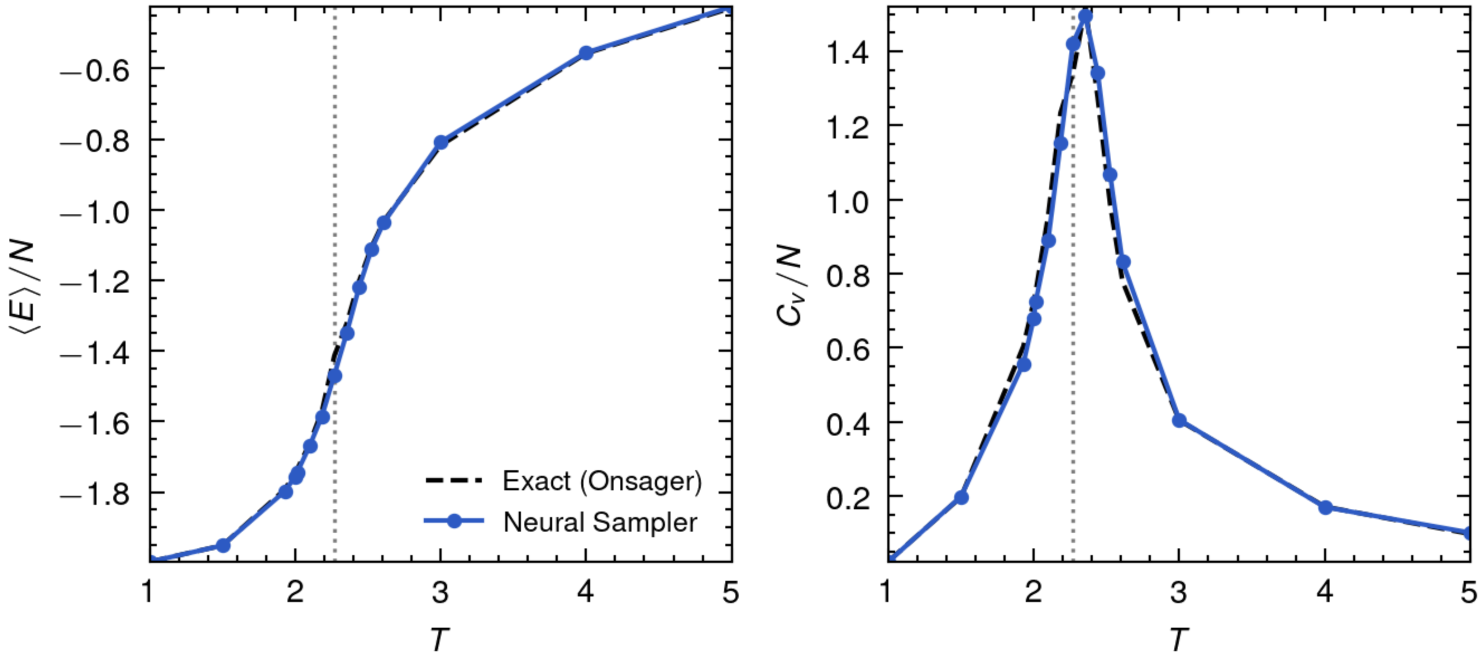


A network learns $q_\theta \approx P$ and draws **i.i.d. samples in one forward pass**: **no critical slowing down**.

Landmarks: VAN [Wu-Wang-Zhang, PRL (2019)], HAN [Białas et al., CPC (2022)], VaTD [Li, Zhang, Pan, PRL (2025)].

[VaTD: Li, Zhang, Pan, PRL (2025)]

- We train a neural sampler (PixelCNN) on the $L = 16$ **2D Ising model** and validate it ourselves.

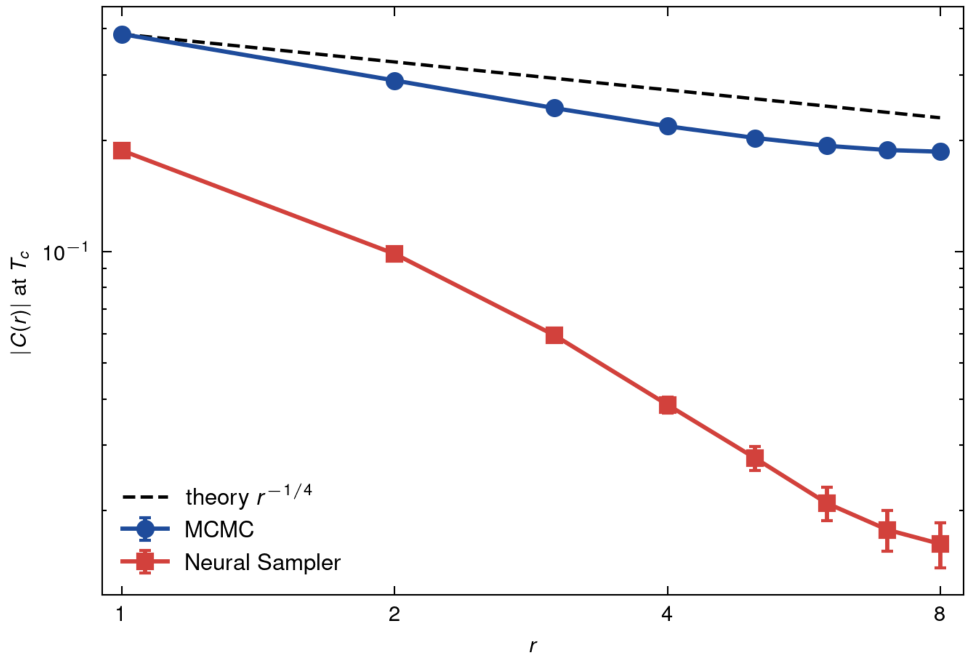


Thermodynamics matches: $\langle E \rangle$ and C_v track the Onsager exact solution.

A neural sampler on 2D Ising

[VaTD: Li, Zhang, Pan, PRL (2025)]

- We train a neural sampler (PixelCNN) on the $L = 16$ **2D Ising model** and validate it ourselves.



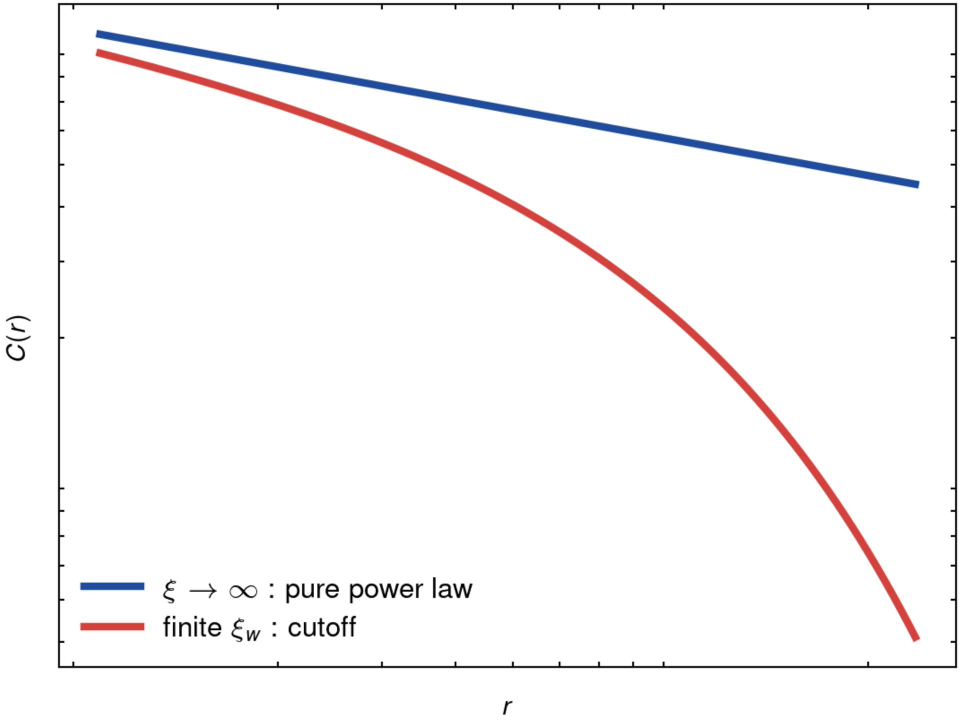
But $C(r)$ at T_c collapses: $\eta \approx 1.24$ (exact 0.25), $\xi \approx \frac{1}{4}$ of MCMC.

Why the tail collapses

[Fisher & Burford, Phys. Rev. (1967)]

- Near T_c, the correlation function has the **scaling form**:

$$C(r) \simeq A\, r^{-(d-2+\eta)}\, e^{-r/\xi}$$



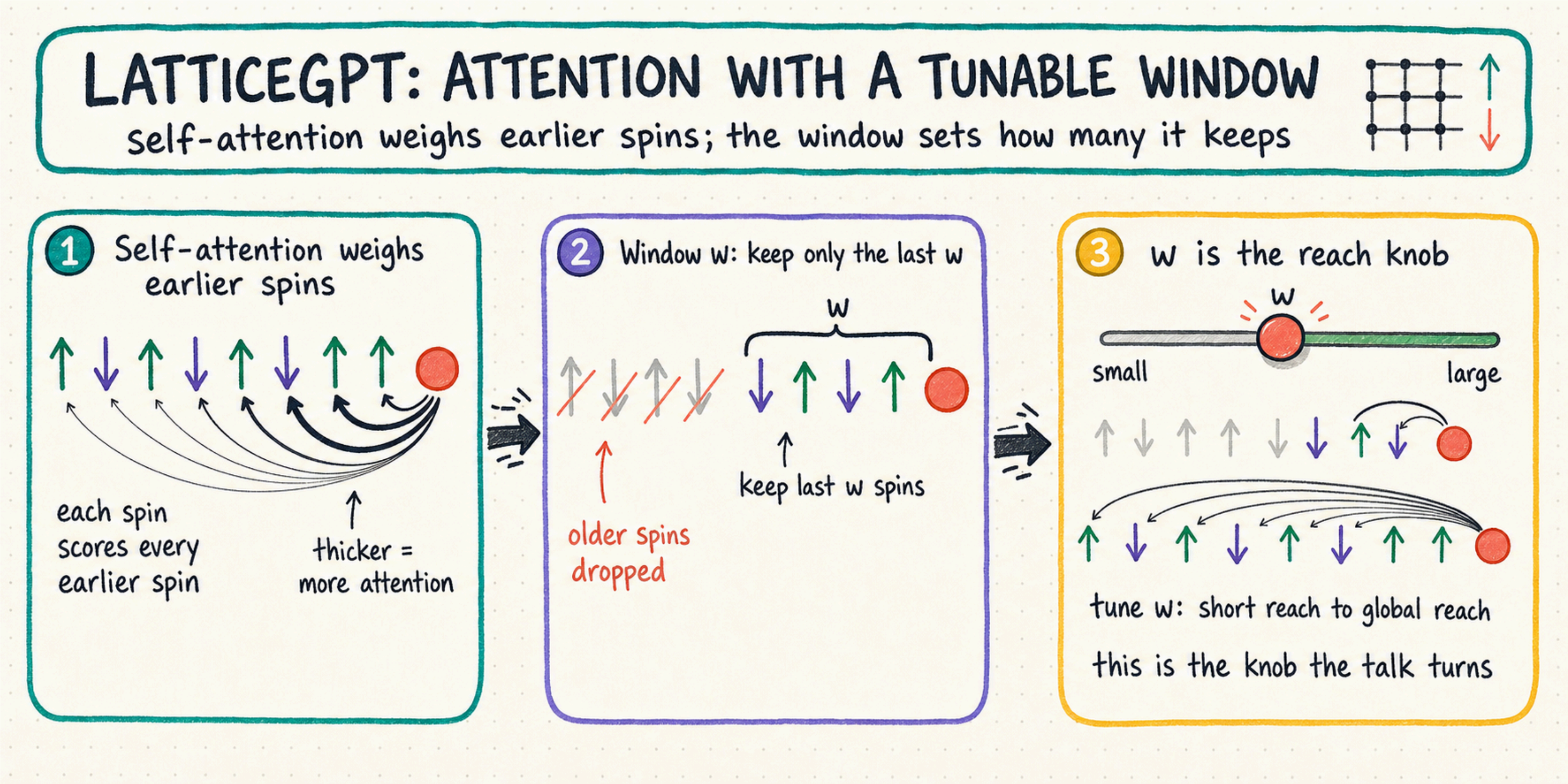
- $\xi \to \infty$ at the 2nd-order T_c: the exponential $\to 1$, leaving a **pure power law**. A well-trained model must follow it.

- finite ξ_w from limited reach: the model learns ξ only up to its window, so the **exponential cutoff** survives and $C(r)$ plunges once $r > \xi_w$.

The collapse is a **reach-set correlation length**: smaller reach $\to$ smaller ξ_w $\to$ earlier cutoff.

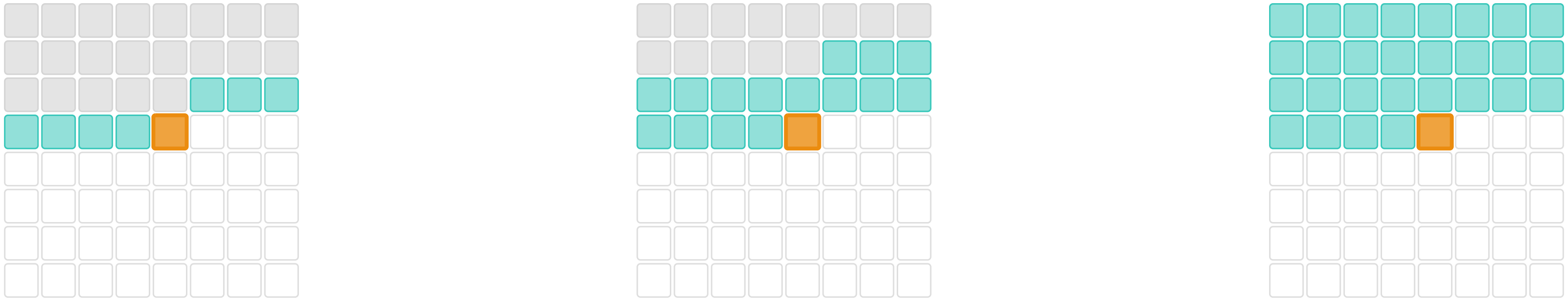
Give it more reach

- The collapse is a **reach problem**: a finite ξ_w set by how far the model can see. The fix is to **give it more reach**.
- But a CNN's reach is **not a knob**: its effective receptive field is far below the nominal RF, and cannot be set directly. (see supplements)
- So we designed a new model where the **communication range is an explicit dial**.


LATTICEGPT: ATTENTION WITH A TUNABLE WINDOW
self-attention weighs earlier spins; the window sets how many it keeps
1 Self-attention weighs earlier spins
each spin scores every earlier spin
thicker = more attention
2 Window w: keep only the last w
w
older spins dropped
keep last w spins
3 w is the reach knob
w
small
large
tune w: short reach to global reach
this is the knob the talk turns

LatticeGPT

The window w sets how far back a model looks when it draws the **current spin** (raster order, $L = 8 \times 8$).



$w = 8$

sees 7 of 28 earlier spins

$w = 16$

sees 15 of 28 earlier spins

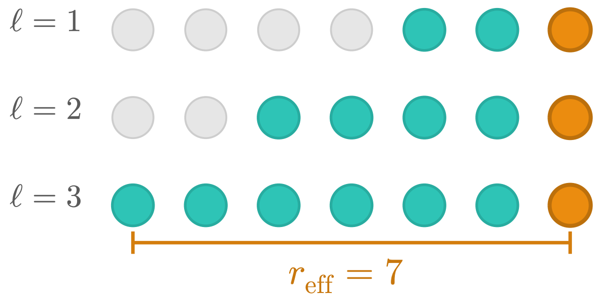
$w = 32$

sees all 28 of 28 earlier spins

seen (window) earlier, unseen not yet drawn

By tuning w, one architecture spans from a **local** model to a **global** one.

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



orange = spin being drawn **teal** = spins within reach

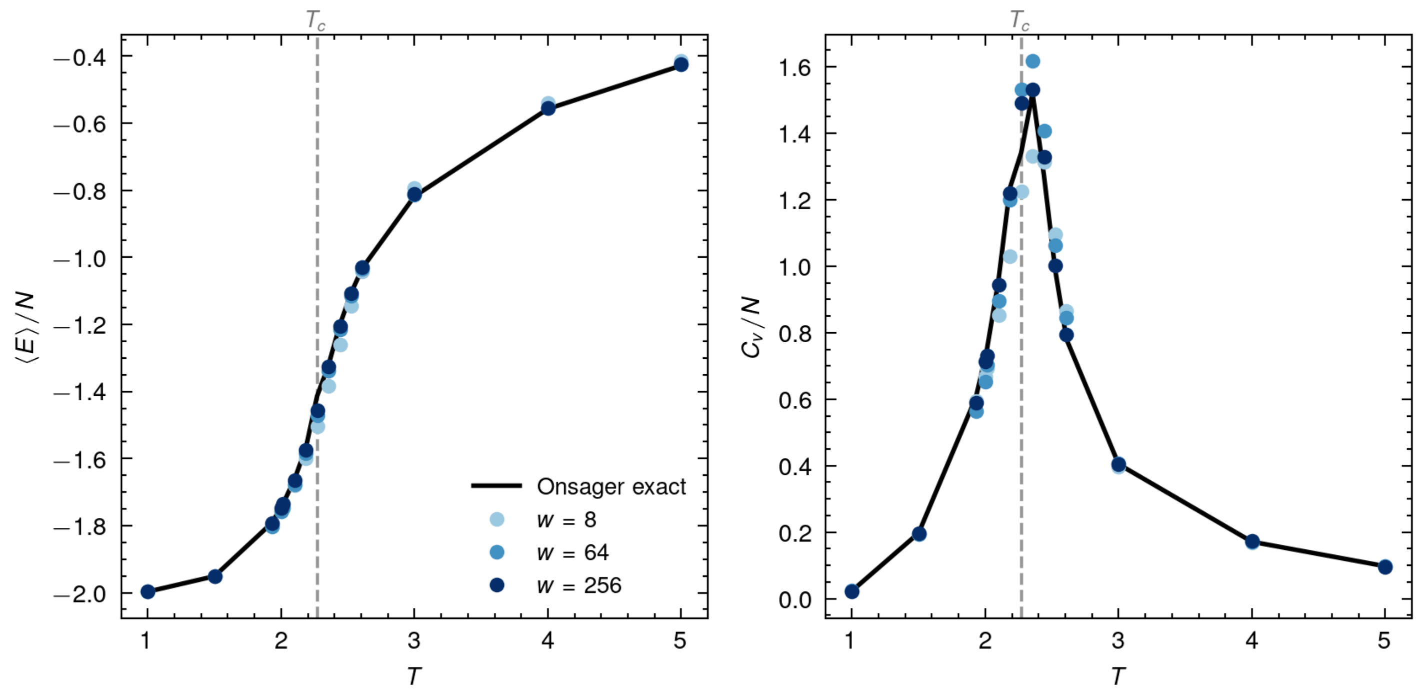
A self-attention window $w = 3$ lets each spin see the $w - 1 = 2$ spins before it. Every added layer reaches $w - 1$ further:

$$r_{\text{eff}} = \ell(w-1)+1$$

We **fix** $\ell = 6$ and sweep the window w:

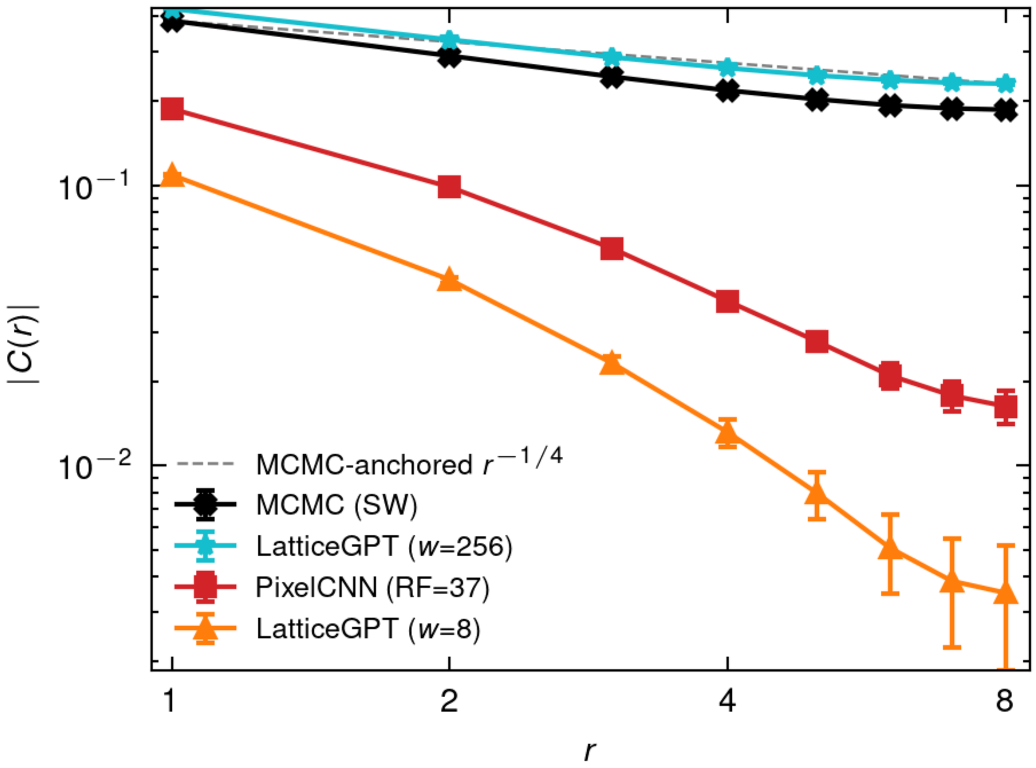
w	8	32	64	128	256
r_{eff}	43	187	379	763	1531

Thermodynamics can't tell them apart



The **local** ($w = 8$) and the **global** ($w = 256$) model both match the Onsager-exact energy and specific heat, and the variational loss converges to the same value. By the standard validation, **every window passes**.

But the correlations disagree



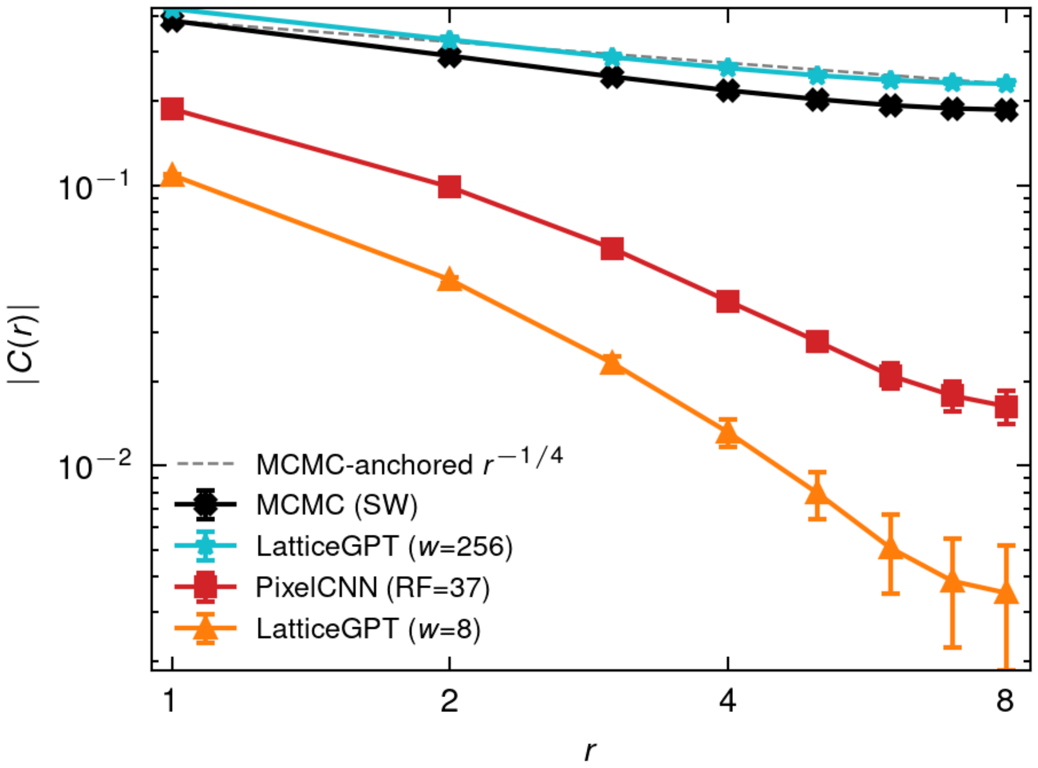
At T_c, the true correlation decays slowly,

$$C(r) \sim r^{-1/4}$$

- $w = 8$ (most local): collapses fastest.
- **PixelCNN** (RF $= 37$): wide nominal reach, yet it collapses too.
- $w = 256$ (global): holds $r^{-1/4}$ out to large r.

The gap **widens with distance r**: a more local model loses more of the long-range physics.

But the correlations disagree



At T_c, the true correlation decays slowly,

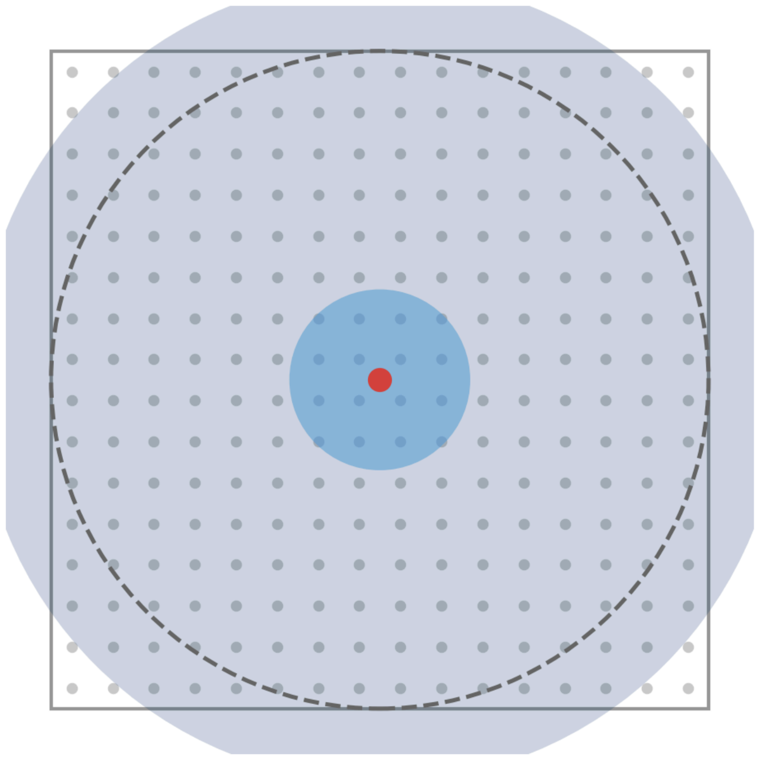
$$C(r) \sim r^{-1/4}$$

- $w = 8$ (most local): collapses fastest.
- **PixelCNN** (RF $= 37$): wide nominal reach, yet it collapses too.
- $w = 256$ (global): holds $r^{-1/4}$ out to large r.

The gap **widens with distance r**: a more local model loses more of the long-range physics.

> Thermodynamics passes every check, yet the critical physics breaks **silently**.

[Pelissetto & Vicari, Phys. Rep. (2002)]



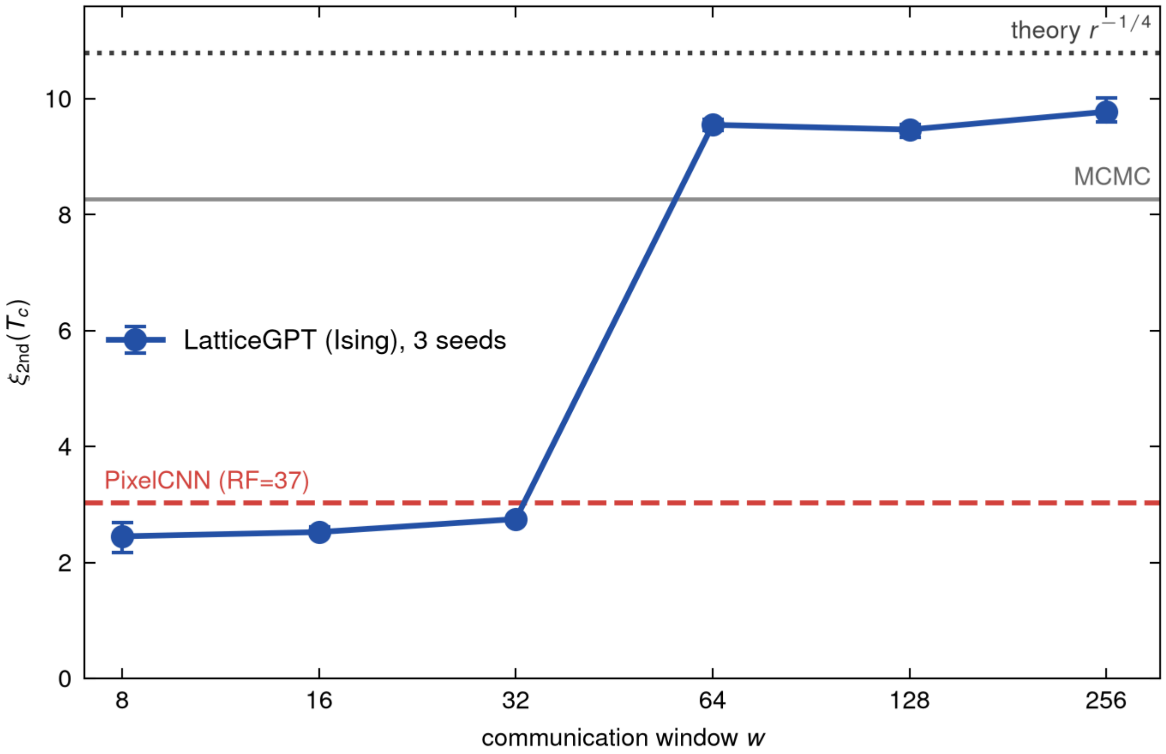
Instead of fitting a shape to $C(r)$, take a moment of it:

$$\xi_{2\mathrm{nd}}^2 = \frac{1}{2d} \frac{\sum_r r^2\, C(r)}{\sum_r C(r)}$$

- The **root-mean-square reach** of correlations: how far, from one spin, correlation actually extends (the disks at left).
- **No shape assumed**: a weighted sum over the $C(r)$ you already measured. No η, no exponential, no fit.

A local $w = 8$ reaches only ≈ 2 sites; a global $w = 256$ nearly fills the box. The natural yardstick is the half-width $L/2$.

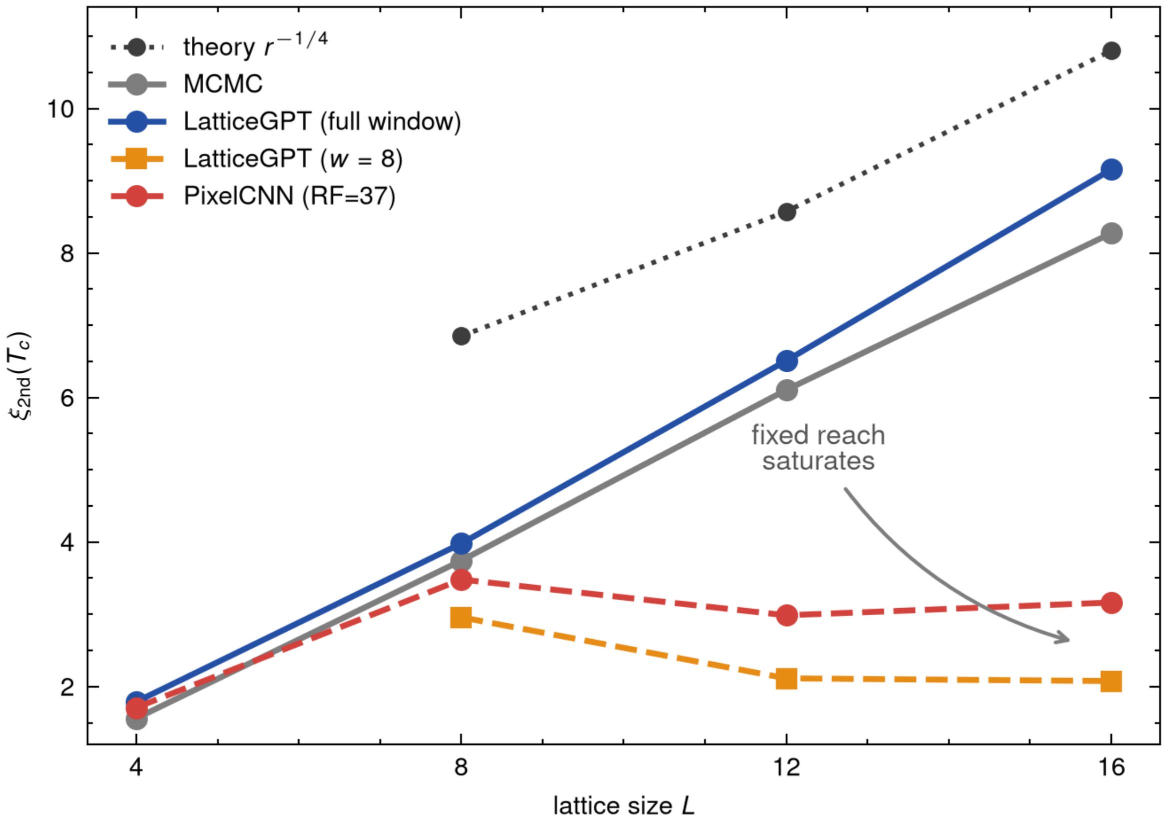
The window controls $\xi_{2\text{nd}}$



Measure $\xi_{2\text{nd}}(T_c)$ from each model's own samples, one window at a time (3 seeds):

- $w = 32 \to r_{\text{eff}} = 187$: just short of spanning the box, stuck at $\xi_{2\text{nd}} \approx 2.5$.
- $w = 64 \to r_{\text{eff}} = 379$: now covers the box, jumps to $\xi_{2\text{nd}} \approx 9.5$ (MCMC, $r^{-1/4}$).

The jump is a **coverage threshold** near $r_{\text{eff}} \approx 190$: once the reach spans the 16×16 lattice, the critical correlations appear. Sharp and seed-robust.

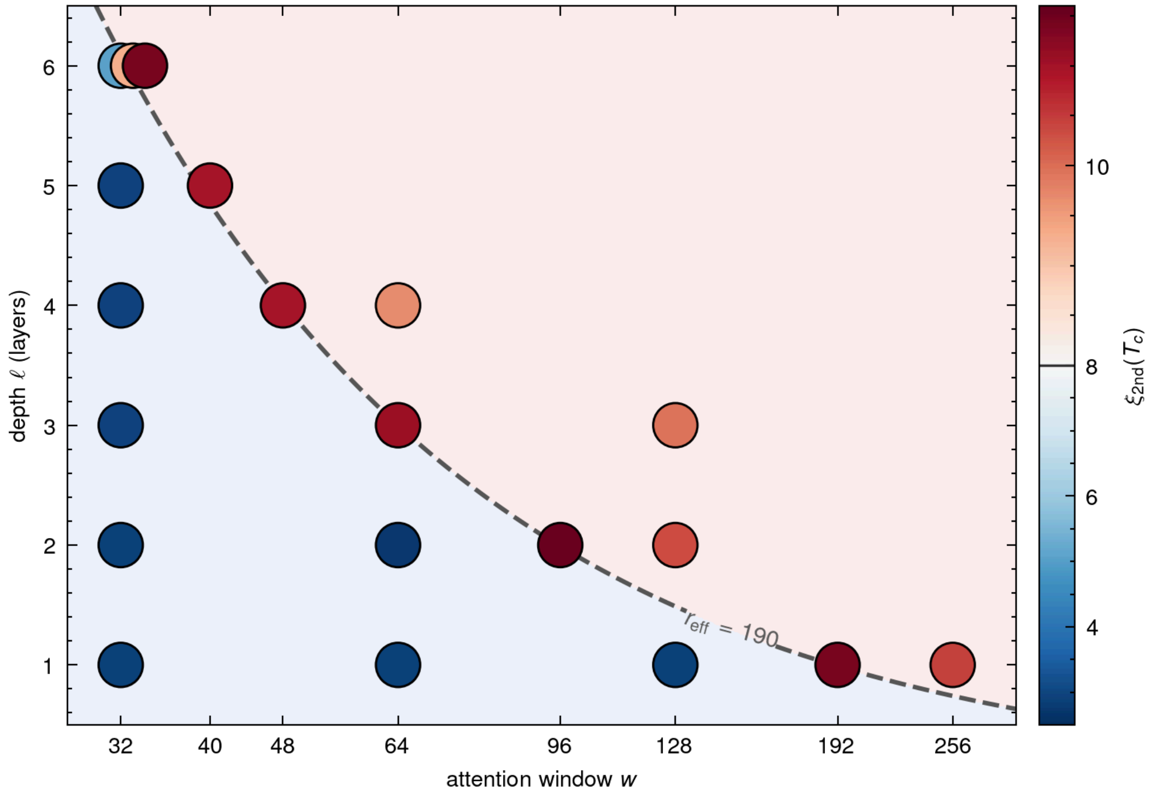


Now fix the reach and grow the lattice L:

- **Full window** tracks MCMC: $\xi_{2nd} \sim L/2$, growing with the box.
- $w = 8 \to r_{eff} = 43$: perfect on a 4×4 box (43 covers all 16 sites), but from 8×8 on it can no longer span the lattice, so ξ_{2nd} saturates near 2.5.

Trustworthiness is reach **relative to the lattice**: a fixed r_{eff} suffices at small L, hopeless at large L.

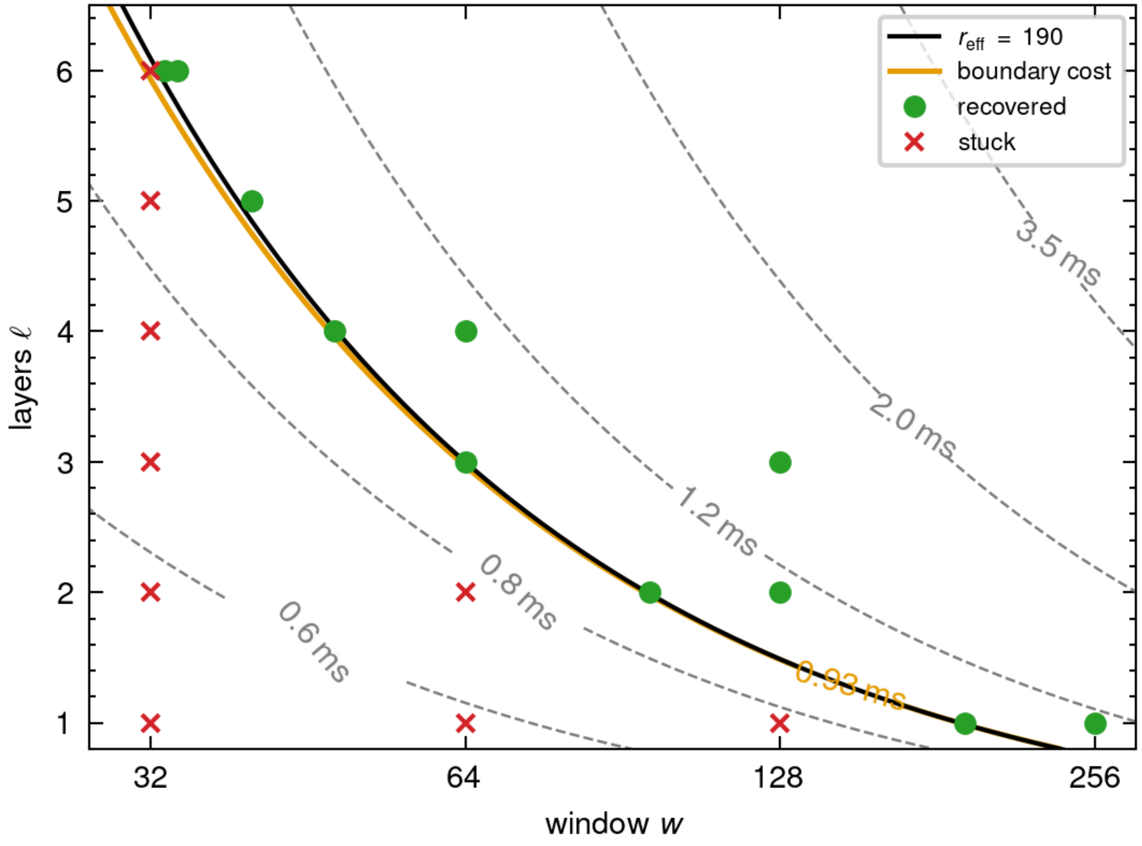
The boundary is a reach contour



Every (ℓ, w) run at **1000 epochs**, colored by its final ξ_{2nd}:

- a single contour $r_{\mathrm{eff}} = \ell(w-1)+1 = 190$ splits the plane;
- **recovered** above it, **stuck** below.

Width and depth trade off freely; only their product r_{eff} decides.

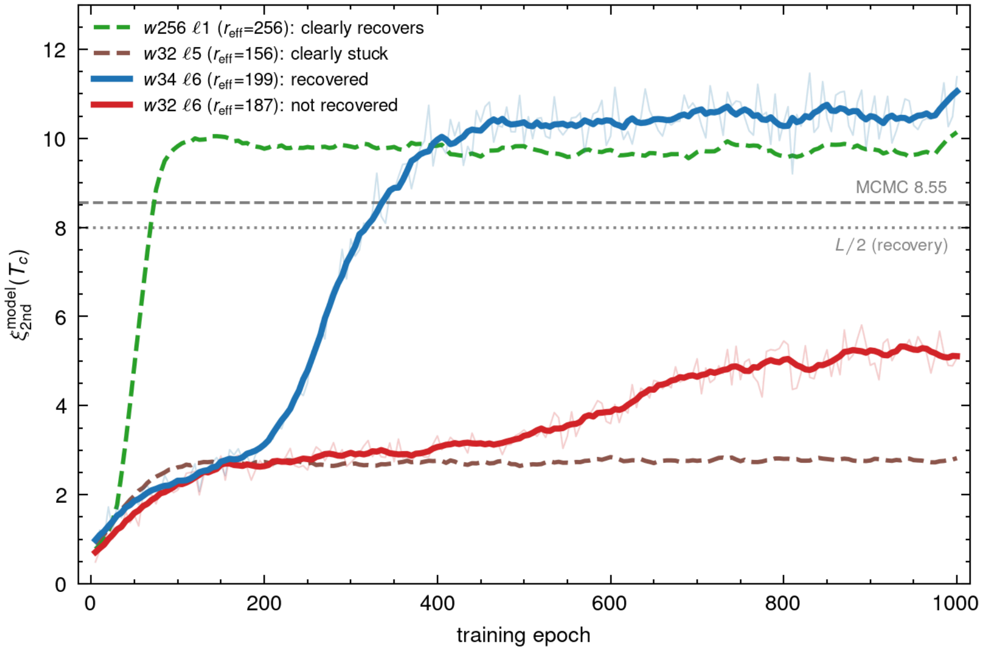


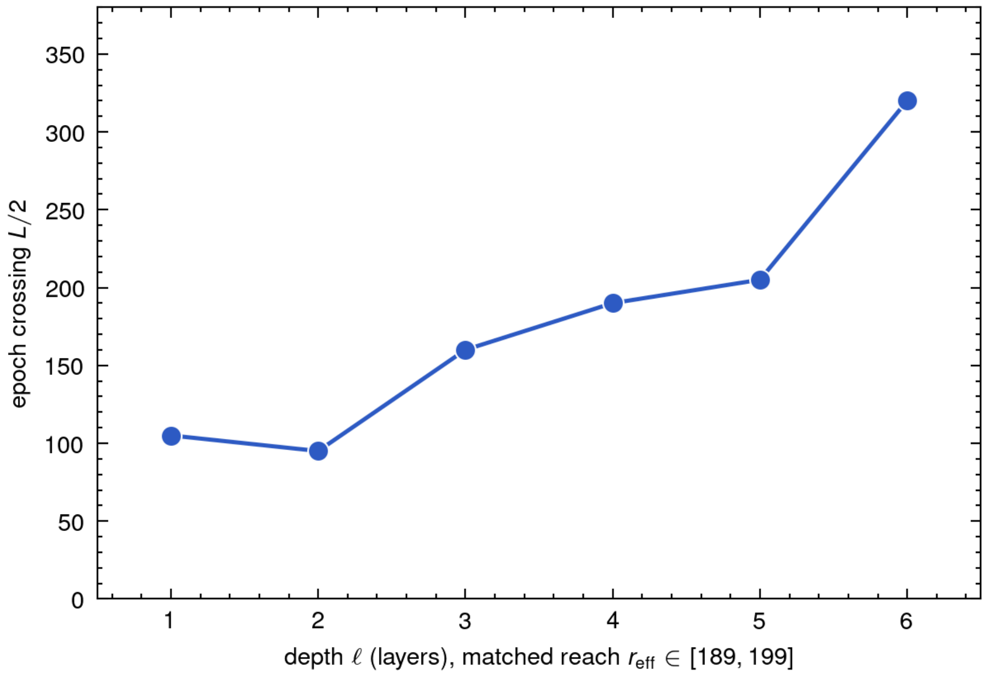
The same (ℓ, w) plane, overlaid with **iso-cost** lines (one-shot sampling time, dashed):

- recovery needs $r_{\text{eff}} \geq 190$ (black): a trustworthy sampler always **costs compute**, never free.
- the boundary runs along a **low, near-constant iso-cost** (orange): on it, sampling cost is set by **reach alone**, the same for any depth $\times$ width split.

Reach fixes the sampling price. What still separates models on the boundary is the training price.

Reach decides, depth costs





1 **Reach sets the outcome: whether it recovers, and how fast it samples.**

2 **Depth sets only the training cost: deeper needs more epochs.**

Outline

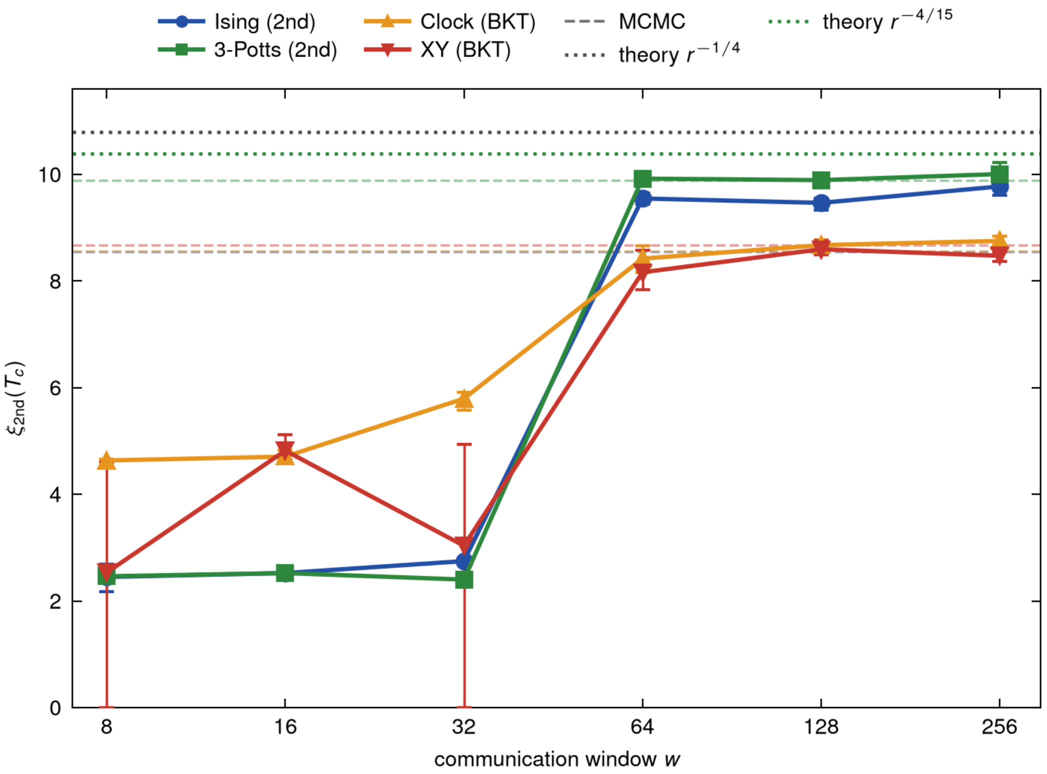
Four systems, bulk passes

System	Interaction	Transition class	$\langle E \rangle$ error at T_c
Ising ($q = 2$)	$-J \sum_{\langle ij \rangle} s_i s_j$	continuous (2nd order)	**≤ 4.5%**
3-Potts	$-J \sum_{\langle ij \rangle} \delta_{\sigma_i \sigma_j}$	continuous (2nd order)	**≤ 3.0%**
Clock ($q = 36$)	$-J \sum_{\langle ij \rangle} \cos(\theta_i - \theta_j)$	BKT (infinite order)	**≤ 0.6%**
XY	$-J \sum_{\langle ij \rangle} \cos(\theta_i - \theta_j)$	BKT (infinite order)	**≤ 3.9%**

Bulk energy stays within **a few percent** at every window (worst $\leq 4.5\%$): by the standard energy gate, all four **ship**.

Truth: Onsager exact (Ising), Swendsen-Wang MCMC (rest). All $L = 16$, 300 epochs.

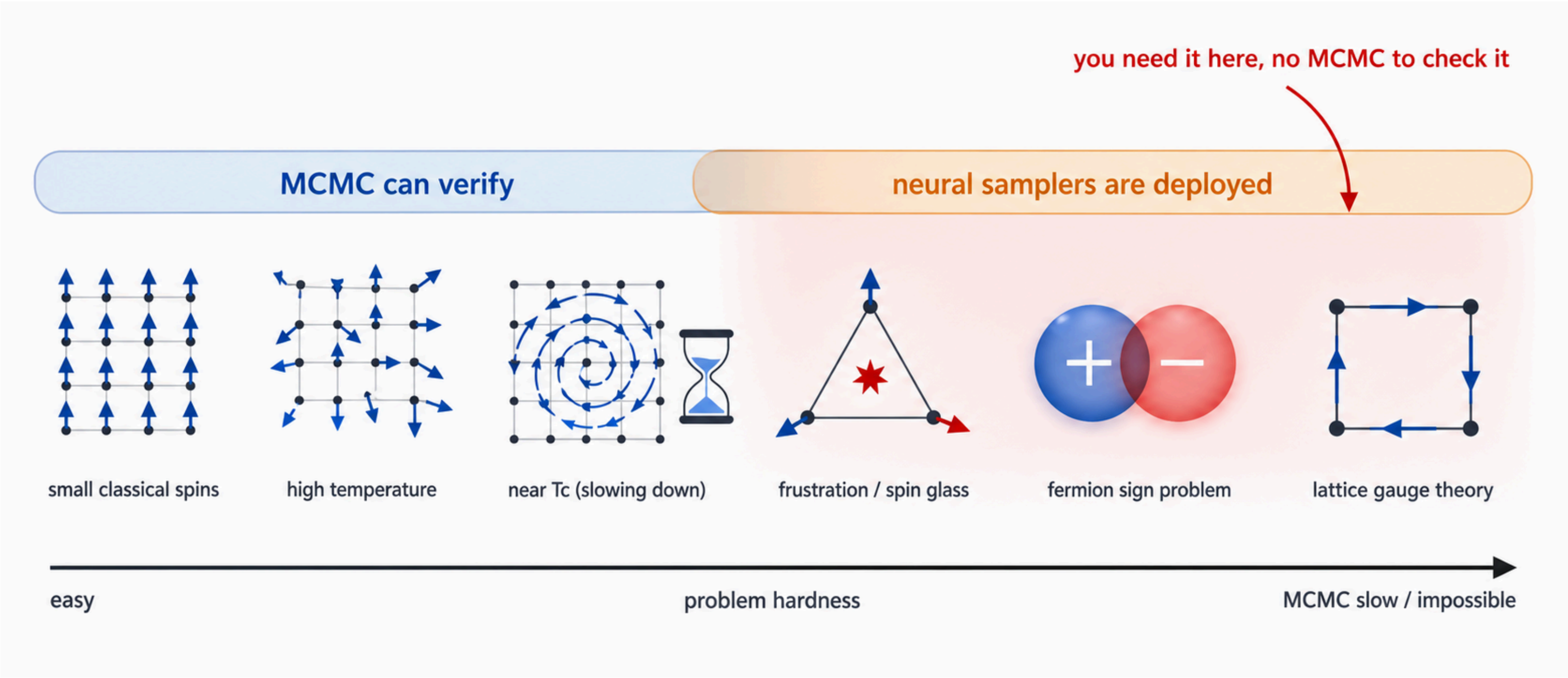
But $\xi_{2\mathrm{nd}}$ still jumps



Four systems, two classes, **one pattern**: broken at small w, then a jump to each system's own truth by $w \approx 64$. The break and the diagnostic are **not Ising-specific**.

Outline

Where you need it, MCMC can't check



Neural samplers are deployed where MCMC can't reach: frustration, the sign problem, gauge theory. There, **no reference exists to check against**. Trust must come from the model's **own samples**.

One number, read entirely from the model's own samples:

$$\rho = \xi_{2\mathrm{nd}}^{\mathrm{model}}(T_c)/(L/2) \geq 1$$

Reference-free (no MCMC)

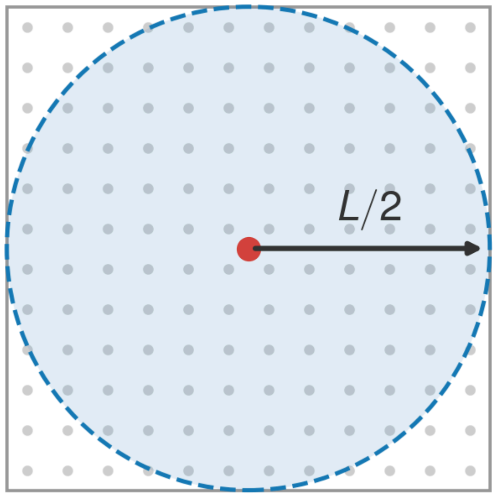
No reference simulation, no ground truth. $\xi_{2\mathrm{nd}}$ is read off the model's own samples at its own susceptibility (χ) peak, so no exact T_c is needed either.

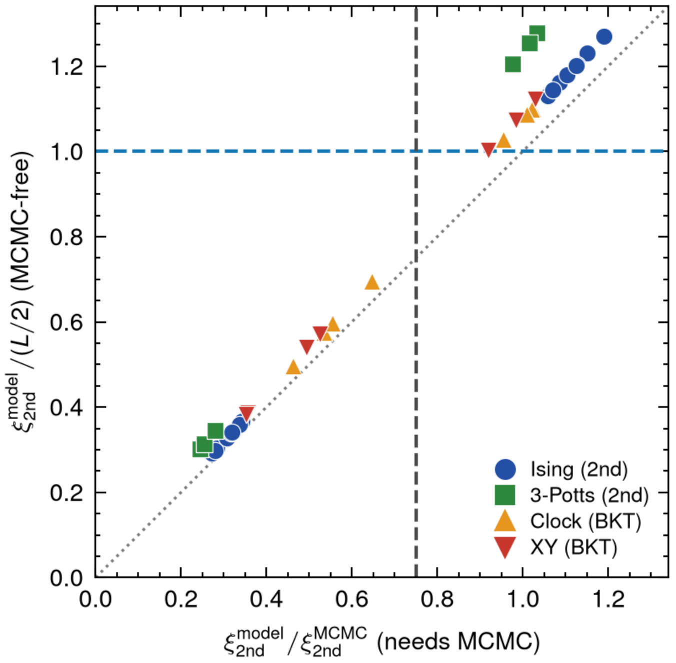
Fit-free

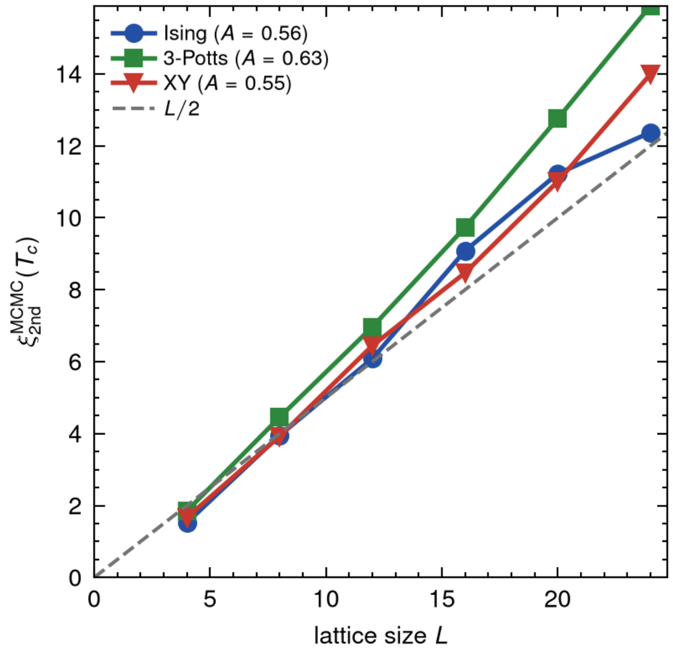
A moment of $C(r)$, not an exponential fit: no assumed functional form, no biased exponent.

Trust the model iff $\rho \geq 1$. But what sets the scale $L/2$?

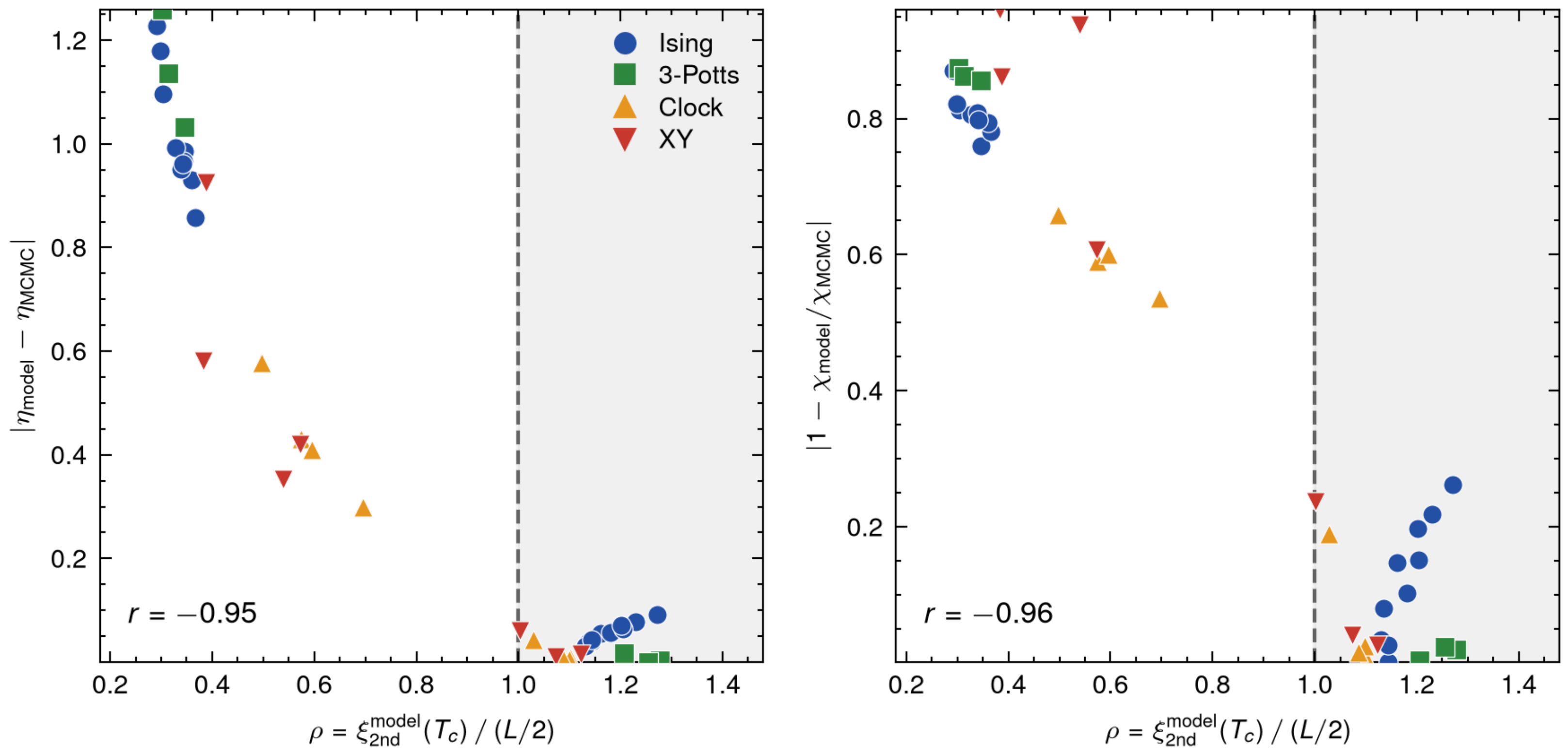
$L/2$ reaches the far edge





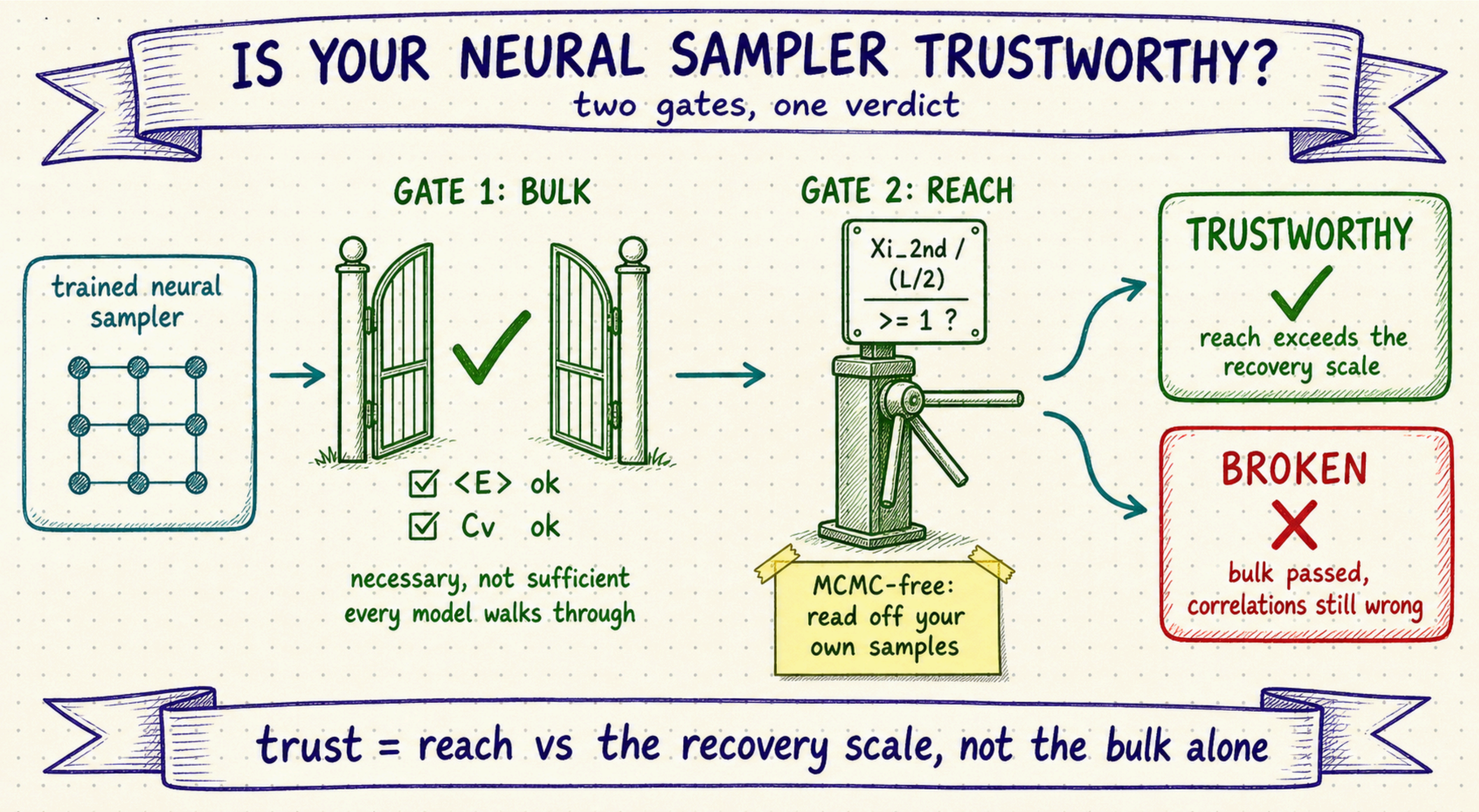


$L/2$ is the farthest a spin can reach on a periodic box (left). At criticality the true $\xi_{\mathrm{2nd}}(T_c) \approx AL$ with $A \approx 0.5$ to 0.6 (right), so $\rho \to 1$ marks a **fully-correlated box**. The MCMC-free rule matches the MCMC one (middle): **same verdict, no reference**.



Across four systems and 41 checkpoints, ρ orders two observables it was **not built from**: the shape-exponent error $|\eta - \eta_{\text{MCMC}}|$ (left) and the susceptibility error $|1 - \chi_{\text{model}}/\chi_{\text{MCMC}}|$ (right, form-free). Both collapse onto ρ and vanish as $\rho \to 1$. A **reference-free** number predicts the reference errors.

Outline



IS YOUR NEURAL SAMPLER TRUSTWORTHY?
two gates, one verdict
trained neural sampler
GATE 1: BULK
<E> ok
Cv ok
necessary, not sufficient
every model walks through
GATE 2: REACH
Xi_2nd / (L/2) >= 1 ?
MCMC-free: read off your own samples
TRUSTWORTHY
reach exceeds the recovery scale
BROKEN
bulk passed, correlations still wrong
trust = reach vs the recovery scale, not the bulk alone

1 A neural sampler can match **every bulk number** ($\langle E\rangle$, C_v) while its **critical correlations silently break**.

2 One **reference-free** gauge catches it: $\rho = \xi_{\mathrm{2nd}}^{\mathrm{model}}(T_c)/(L/2) \geq 1$, from the model's own samples. No MCMC, no fit, no exact T_c.

3 Recovery is set by one number, reach $r_{\mathrm{eff}} = \ell(w-1)+1$: below its threshold nothing recovers, and depth or budget shift only the **timing**. The same threshold governs Ising, Potts, Clock, and XY.

Trust a neural sampler by its reach, not by its energy.

Supplements

Importance reweighting

- **Goal** Estimate true-Boltzmann averages $\langle O \rangle_P$ from neural samples $s \sim q_\theta$ (with $q_\theta \approx P$ but $q_\theta \neq P$).
- **Importance-sampling identity**

$$\langle O \rangle_P = \sum_s P(s)\, O(s) = \sum_s q_\theta(s) \underbrace{\frac{P(s)}{q_\theta}(s)}_{w(s)} O(s) = \frac{\langle w\, O \rangle_{q_\theta}}{\langle w \rangle_{q_\theta}}$$

with $w(s) = e^{-\beta H(s)} / q_\theta(s)$ (the unknown Z cancels between numerator and denominator).

- **Estimators** draw M samples $s_1, \ldots, s_M \sim q_\theta$, weight each $w_i = e^{-\beta H(s_i)} / q_\theta(s_i)$:

$$\langle O \rangle_P \approx \frac{\sum_i w_i\, O(s_i)}{\sum_i w_i}, \quad Z \approx \frac{1}{M} \sum_i w_i = \langle w \rangle, \quad F = -T \ln \langle w \rangle$$

- **When it works** Effective sample size $\mathrm{ESS} = \left(\sum_i w_i\right)^2 / \sum_i w_i^2 \leq M$ counts how many of the M draws effectively contribute: near M when $q_\theta \approx P$, but $\mathrm{ESS} \ll M$ if q_θ misses high-P regions $\rightarrow$ huge variance, silent failure. (the trust question)

[Pelissetto & Vicari, Phys. Rep. 368, 549 (2002)]

r^2 in real space is $-\partial_k^2$ in Fourier space, so the second moment is the curvature of the structure factor at $k=0$. The lattice Ornstein-Zernike form turns that curvature into ξ.

Structure factor (one FFT), in the lattice Ornstein-Zernike form:

$$S(\boldsymbol{k}) = \sum_{\boldsymbol{r}} C(\boldsymbol{r})\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{S(0)}{1+\xi^2\hat{k}^2}, \quad \hat{k} = 2\sin(k/2)$$

Its curvature at $k=0$ is ξ, exactly the second moment:

$$\xi^2 = -\frac{1}{2d}\frac{\nabla_k^2 S(0)}{S(0)} = \frac{1}{2d}\frac{\sum_{\boldsymbol{r}} |\boldsymbol{r}|^2 C(\boldsymbol{r})}{\sum_{\boldsymbol{r}} C(\boldsymbol{r})}$$

Read that curvature off the two lowest modes, $k=0$ and $k_{\text{min}} = 2\pi/L$:

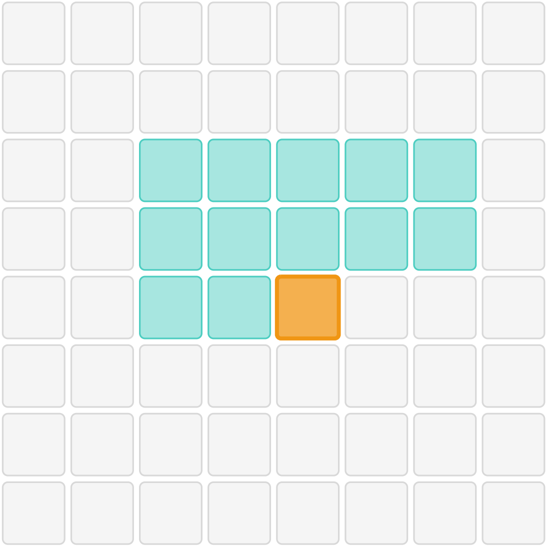
$$\frac{S(0)}{S(k_{\text{min}})} = 1 + \xi^2\hat{k}^2_{\text{min}}$$

$$\xi_{\text{2nd}} = \frac{1}{2\sin(k_{\text{min}}/2)}\sqrt{\frac{S(0)}{S(k_{\text{min}})} - 1}$$

One FFT, two numbers, **no fit**. This $\xi_{\text{2nd}}(T_c)$, from the model's own samples, is the number the rest of the talk uses.

Valid even though OZ is not exact? Yes. The identity curvature $=$ second moment holds exactly, with no OZ assumption. The Lorentzian only interpolates the two lowest modes; the anomalous (η) deviation enters at $O(k^4)$, so the estimator is exact as $L\to\infty$ ($O(1/L^2)$ bias at finite L), a controlled fit-free measurement rather than an OZ fit.

Communication range

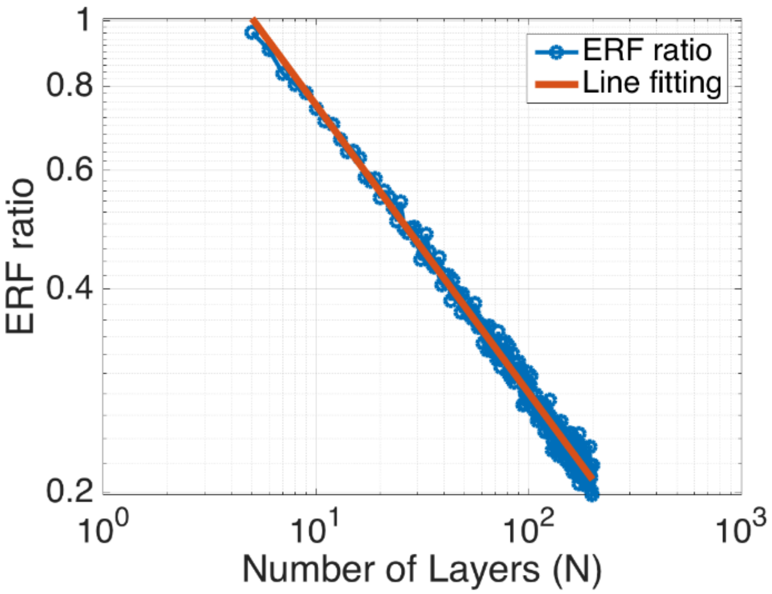


- **Thermodynamics is forgiving**
  $\langle E\rangle$ and C_v are short-range and averaged, so even a short-reach model reproduces them.

- **$C(r)$ is not**
  It probes correlations at each distance; the error grows with r, worst at T_c. What bulk hides, $C(r)$ exposes.

- **So: communication range?**
  A spin's correlations reach only as far as it can see when it is generated.

- **But PixelCNN sees the whole lattice**
  Masked convolutions stack to $\mathrm{RF} = 37$, spanning 16×16. On paper it should capture long-range correlations, so why doesn't it?



Effective receptive field: Luo, Li, Urtasun & Zemel, NeurIPS (2016).

- **Effective $\ll$ nominal**
  A CNN's effective receptive field is far smaller than the nominal one, concentrated near the center.

- **And it shrinks with depth**
  The effective fraction falls as $\sim 1/\sqrt{N}$ with depth N.

- **A reach we cannot set**
  So $\mathrm{RF} = 37$ overstates PixelCNN's real reach, and reach is not something we can directly set or measure in a CNN.

- **We need a new instrument**
  An architecture where communication range is an explicit knob. $\rightarrow$ **LatticeGPT** (attention window w).

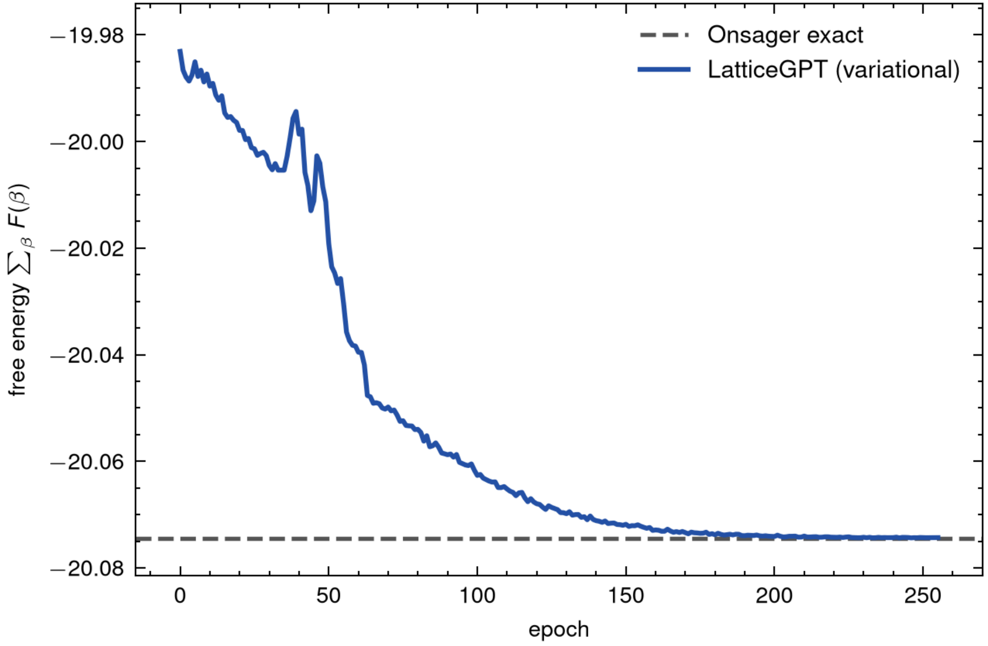
Variational, Temperature-Differentiable (VaTD)

No data. minimize the free energy by **REINFORCE**, with β a **continuous input** (whole band, not a grid).

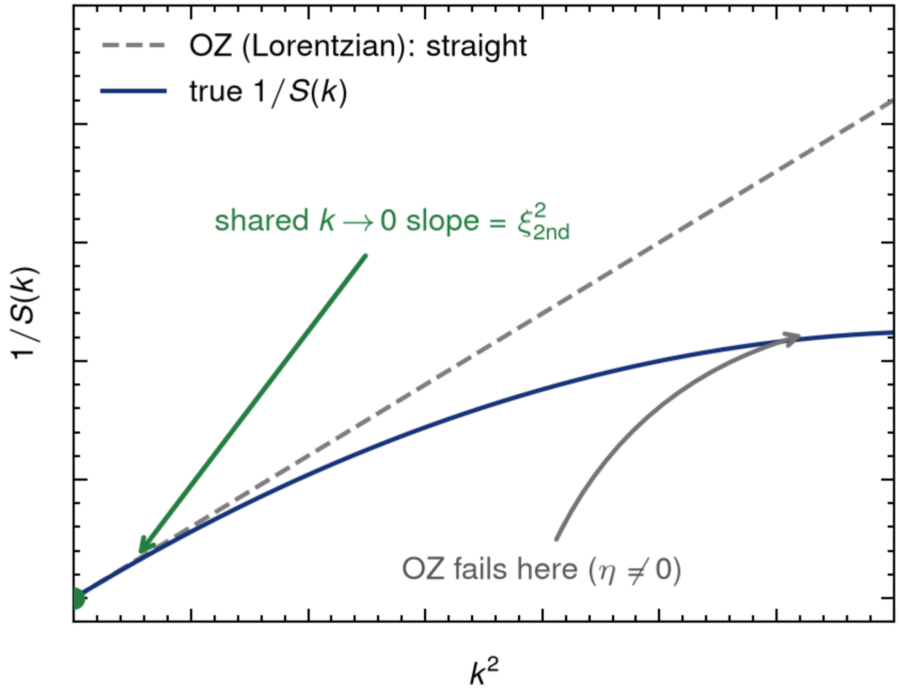
$$\mathcal{F}_\beta[q_\theta] = \langle E \rangle_{q_\theta} + T\langle \ln q_\theta \rangle_{q_\theta} \geq F(\beta)$$

Autoregressive: exact likelihood, **i.i.d.** samples (no Markov chain).

Setup. Windowed Transformer on a **periodic** lattice ($d_{\text{model}} = 192$, 6 heads, up to 6 layers); reach set by the window w. Optimizer **SPlus** [arXiv:2506.07254], scheduler **ExpHyperbolicLR** [**T.-G. Kim**, arXiv:2407.15200] 300 epochs.



The summed variational free energy converges to the **Onsager-exact** value.



The real-space moment (last slide) is also the $k \to 0$ curvature of the **structure factor** $S(k) = \sum_r e^{i\boldsymbol{k}\cdot\boldsymbol{r}} C(r)$ (what scattering measures; $S(0) = \chi$, the susceptibility):

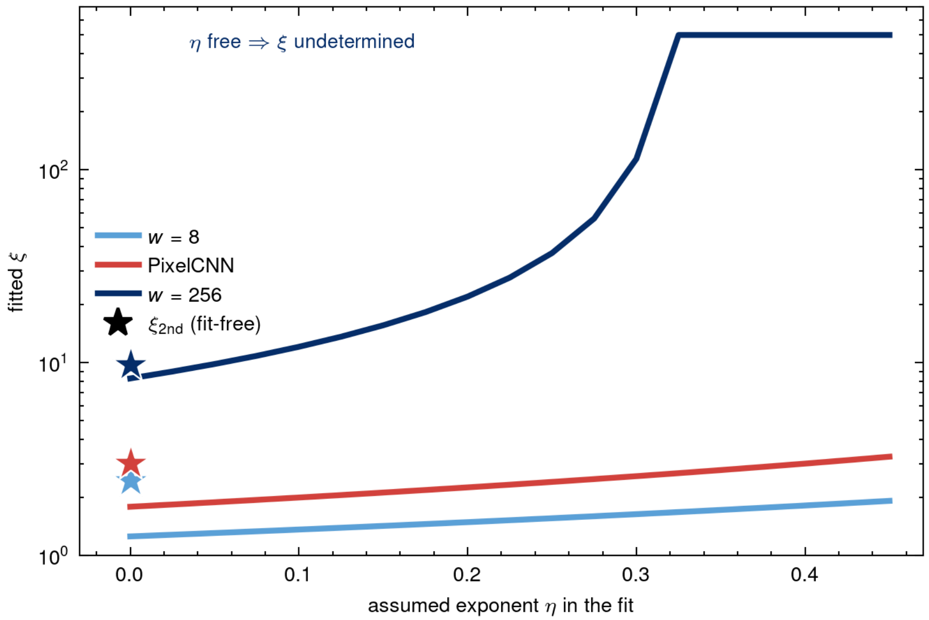
$$S(k)/S(0) = 1 - \xi^2_{\mathrm{2nd}}\, k^2 + O(k^4)$$

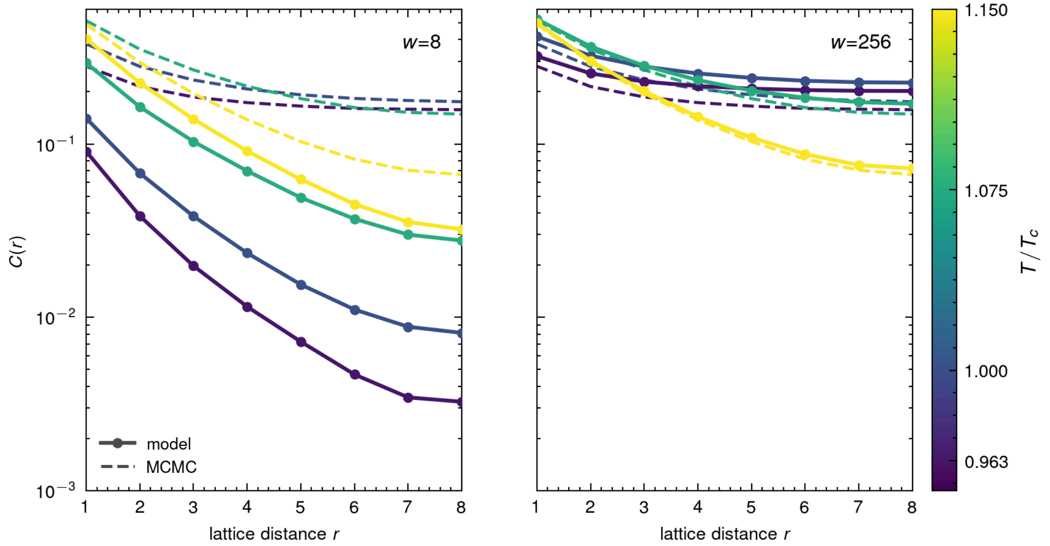
This is the textbook **second-moment correlation length**, with a long lineage:

- correlation length from scattering $S(k)$, [Ornstein & Zernike, 1914]
- the second-moment effective range vs the true exponential range, [Fisher 1964; Fisher & Burford 1967]
- the standard lattice definition, [Pelissetto & Vicari 2002]

OZ (1914) approximated $S(k)$ by a straight line (dashed); ξ_{2nd} needs only the $k \to 0$ slope, so it holds even where that approximation fails.

Why not just fit ξ?



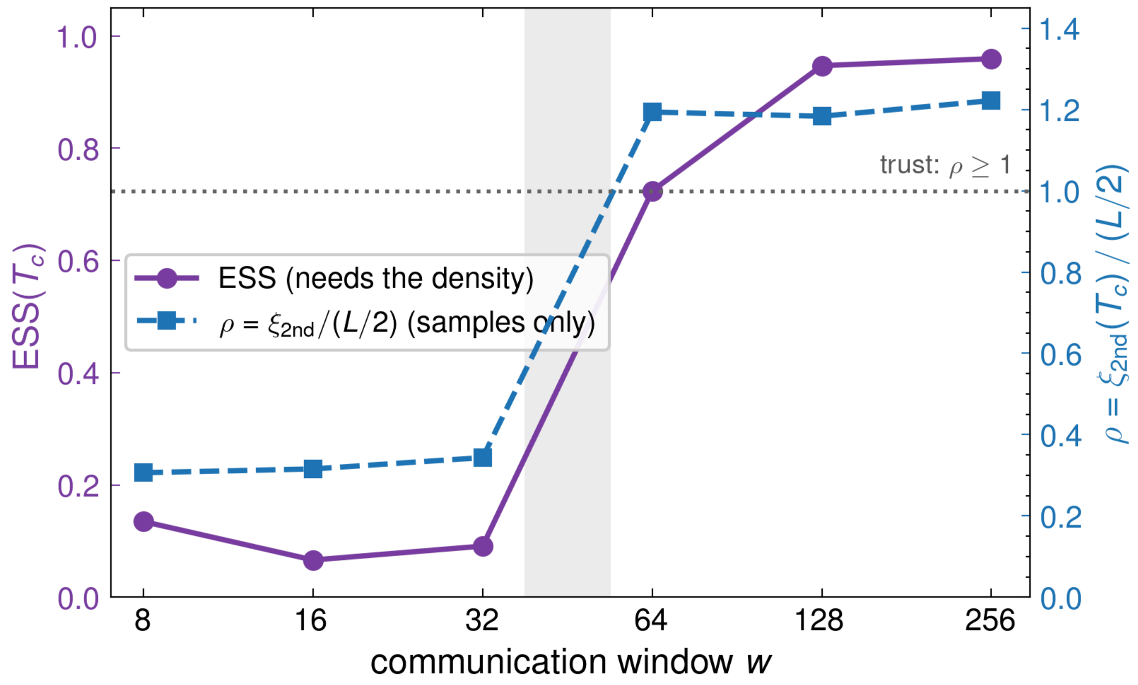


η and ξ are degenerate. Leave η free and the fitted ξ runs from 1 to ∞. To get ξ you must already know η.

$C(r)$ swings with T. Miss the exact T_c and the fit anchors on the wrong curve entirely.

Any exponential fit must **assume both η and T_c**. We want a length that assumes **neither**.

[ESS: Nicoli et al., PRE (2020)]



On the $\ell = 6$ Ising window sweep, the effective sample size (ESS) and $\rho = \xi_{2nd}/(L/2)$ **agree**: both jump together as the reach crosses the lattice.

ρ's edge: samples-only

ESS needs the model's exact density q_θ. ρ reads $C(r)$ off the **configurations alone**, so it works for **any sampler**, even the intractable-density frontier (diffusion, flows, black-box) where ESS cannot be formed.