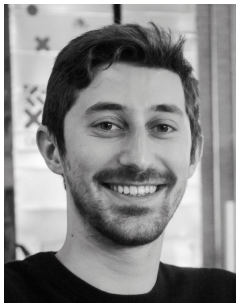
Massachusetts Institute of Technology

# Flow-based generative models for MCMC in lattice field theory[1]

Michael S. Albergo, **Gurtej Kanwar**, Phiala E. Shanahan

Center for Theoretical Physics, MIT



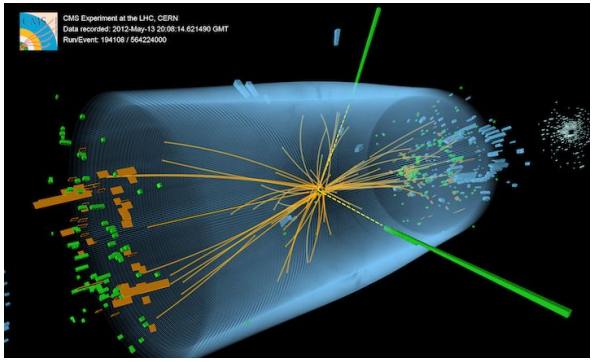
[1] [Albergo, GK, Shanahan 1904.12072]

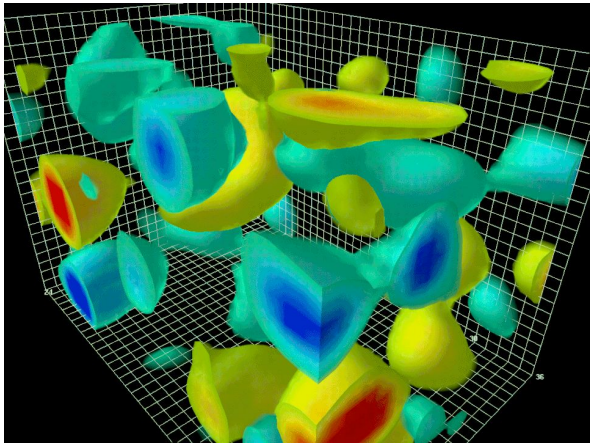
37th International Symposium on Lattice Field Theory
Wuhan, China (June 20, 2019)

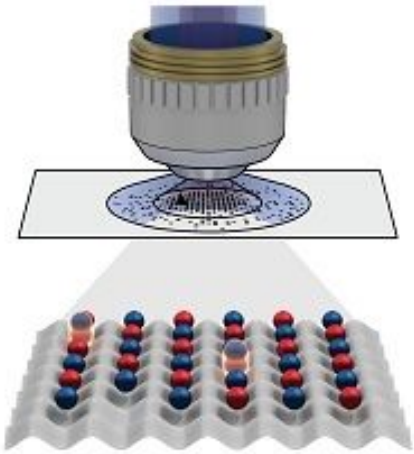
# Machine learning for lattice theories

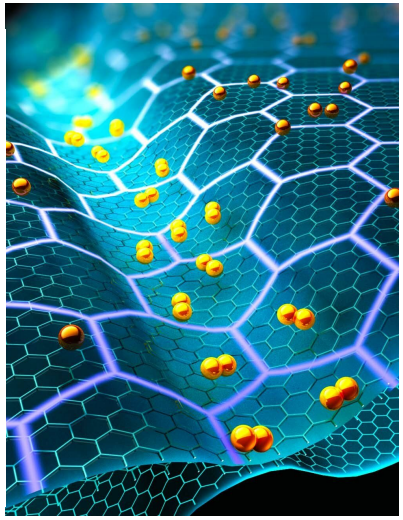
Quantum field theories

Real-world lattices

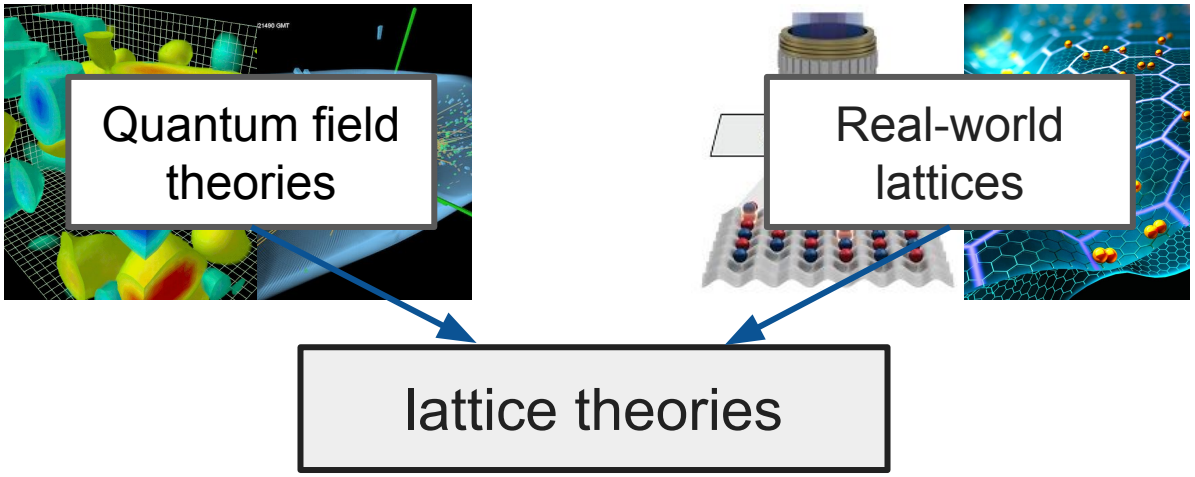


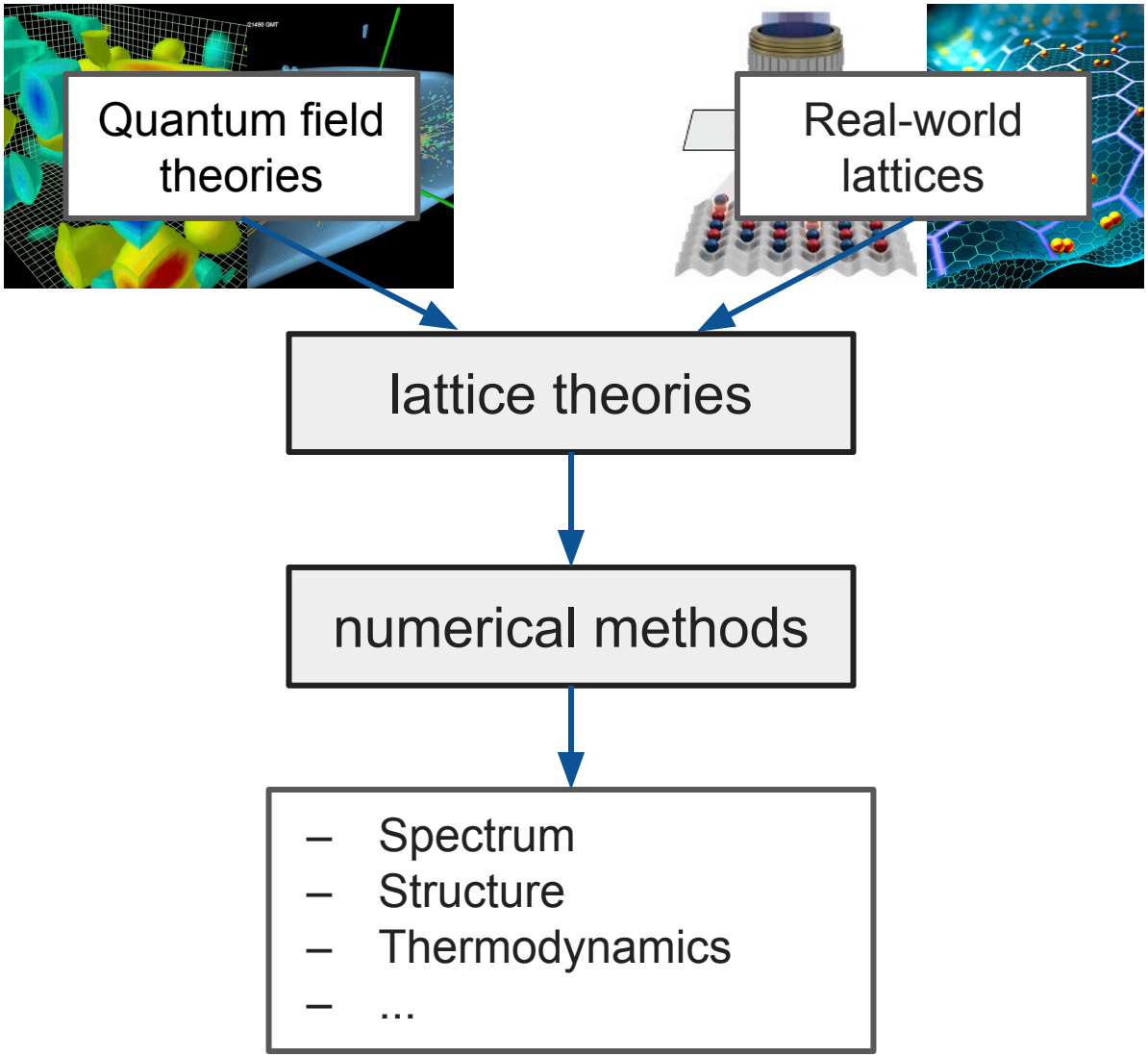




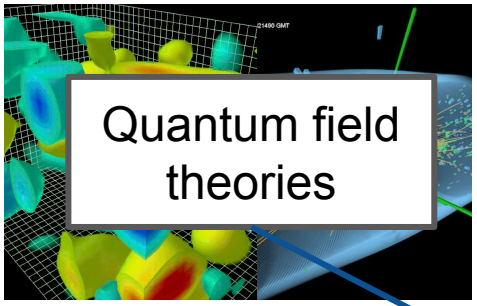
# Machine learning for lattice theories

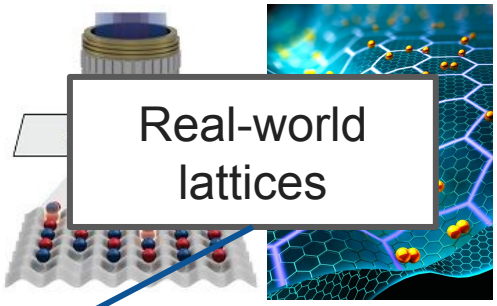


# Machine learning for lattice theories



# Machine learning for lattice theories



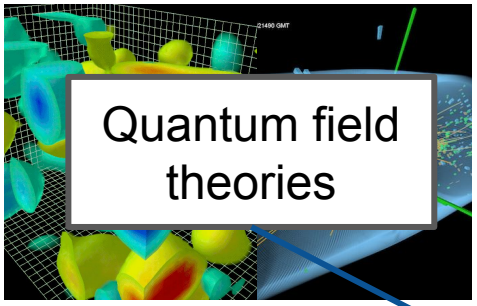

lattice theories

numerical methods

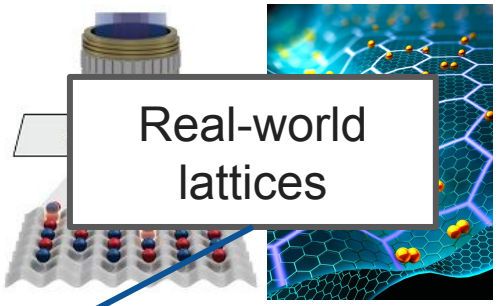
**hard to reach continuum limit / critical point in some theories**

- Spectrum
- Structure
- Thermodynamics
- ...

# Machine learning for lattice theories



Quantum field theories



Real-world lattices

lattice theories

numerical methods

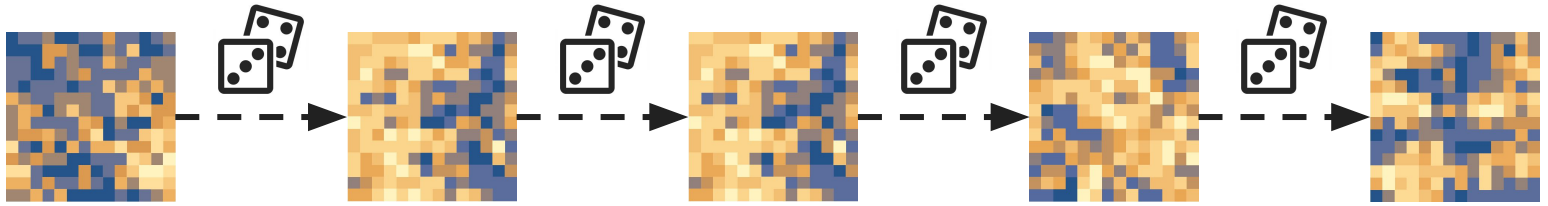
+ ML

hard to reach continuum limit / critical point in some theories

- Spectrum
- Structure
- Thermodynamics
- ...

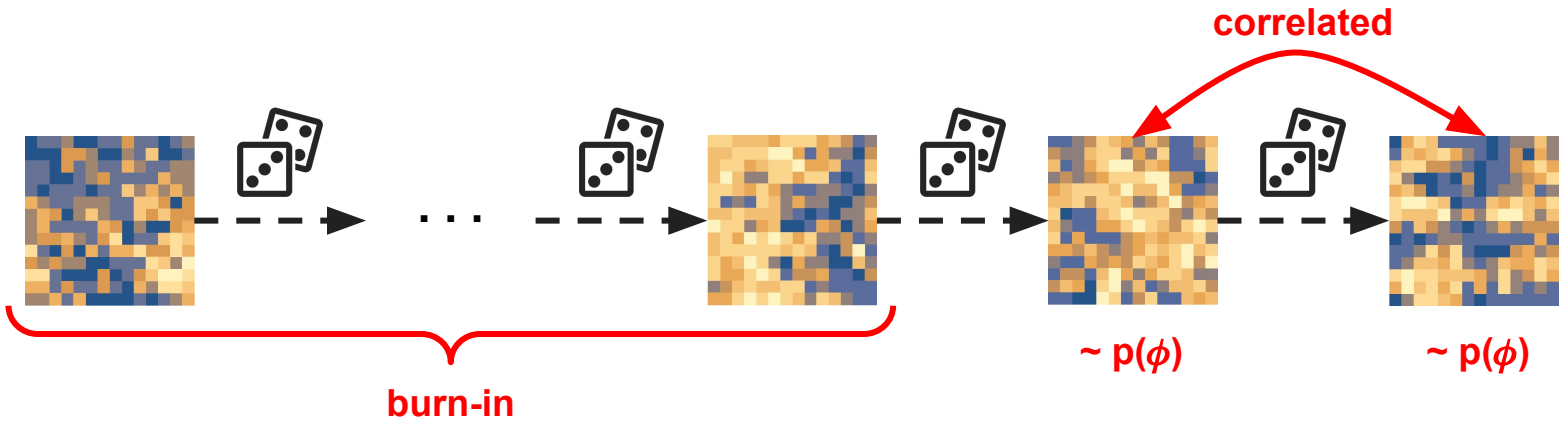
# Computational approach to lattice theories

- Markov Chain Monte Carlo allows estimating path integrals / partition functions



# Computational approach to lattice theories

- Markov Chain Monte Carlo allows estimating path integrals / partition functions
  - Need to wait for "burn-in period"
  - Need to take many steps before drawing independent samples

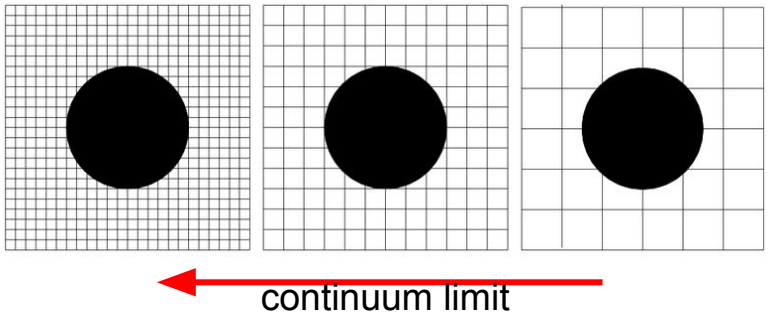



- Burn-in and correlations both related to Markov chain "autocorrelation time" → smaller autocorrelation time means **less computational cost!**

$$\tau_{\mathcal{O}}^{\mathrm{int}} = \frac{1}{2} + \lim_{\tau_{\mathrm{max}} \to \infty} \sum_{\tau=1}^{\tau_{\mathrm{max}}} \frac{\rho_{\mathcal{O}}(\tau)}{\rho_{\mathcal{O}}(0)}$$

# Critical slowing down

- As parameters in the action approach criticality, for Markov chains using local updates, **autocorrelation time diverges**



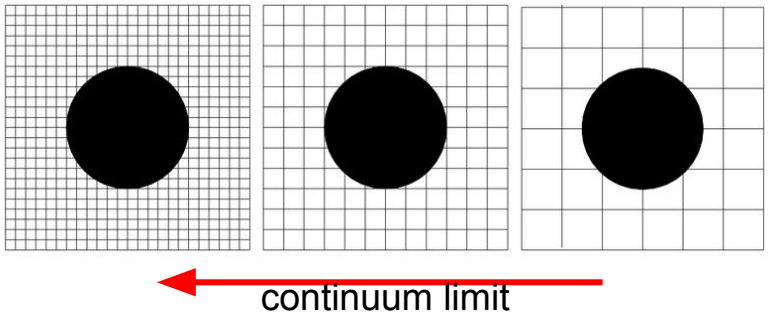

- Fitting $\tau^{\text{int}}$ to power law behavior gives dynamical critical exponents

$$\tau_{\mathcal{O}}^{\text{int}} = \alpha_{\mathcal{O}} L^{z_{\mathcal{O}}}$$

- Smaller dynamical critical exponent = cheaper, closer approach to criticality

# Critical slowing down

- As parameters in the action approach criticality, for Markov chains using local updates, **autocorrelation time diverges**



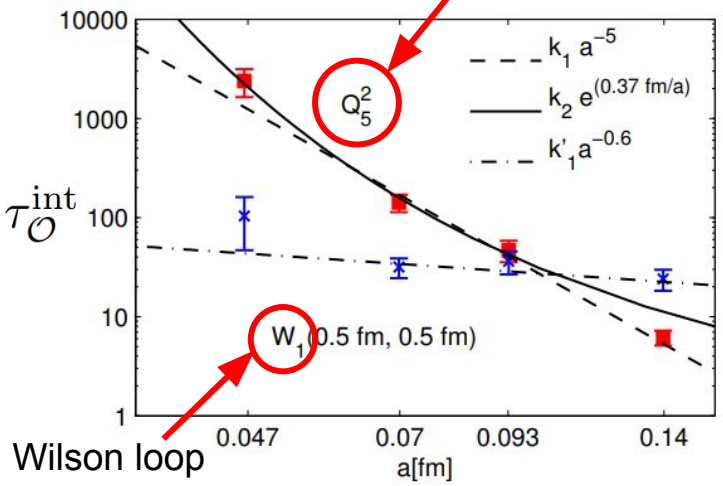

- Fitting $\tau^{\text{int}}$ to power law behavior gives dynamical critical exponents

$$\tau_{\mathcal{O}}^{\text{int}} = \alpha_{\mathcal{O}} L^{z_{\mathcal{O}}}$$

- Smaller dynamical critical exponent = cheaper, closer approach to criticality

topological charge

[Schaefer et al. / ALPHA collaboration 1009.5228]

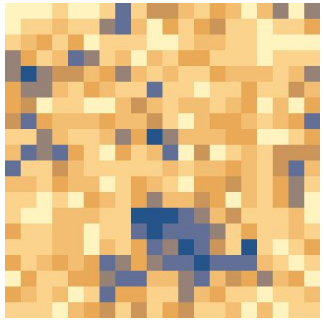



CSD also affects (naive simulation of) simpler models:

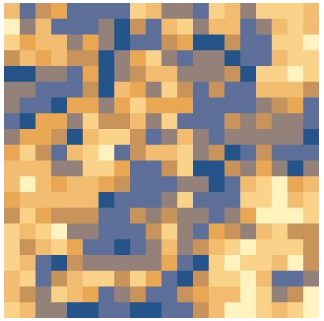
- $CP^{N-1}$ [Flynn, et al. 1504.06292]
- O(N) [Frick, et al. PRL 63, 2613]
- $\phi^4$ [Vierhaus doi:10.18452/14138]
- ...

# Sampling lattice configs ≅ generating images

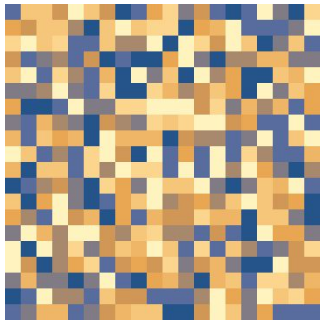
**likely**
(log prob = 22)



**likely**
(log prob = 5)



**unlikely**
(log prob = -6107)

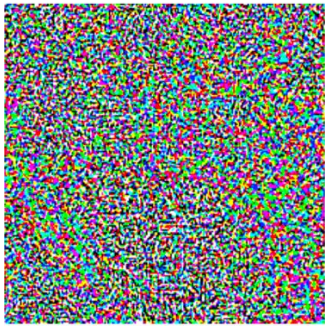


**likely**

[Karras, Lane, Aila / NVIDIA 1812.04948]

**likely**

**unlikely**



# Sampling lattice configs vs. generating images

✓ Probability density can be computed for a given sample (up to normalization)

$$p(..) = e^{-S(...)}/Z$$

✓ Physics distributions have many symmetries

✘ For lattice field theories, $10^9$ to $10^{12}$ variables per config

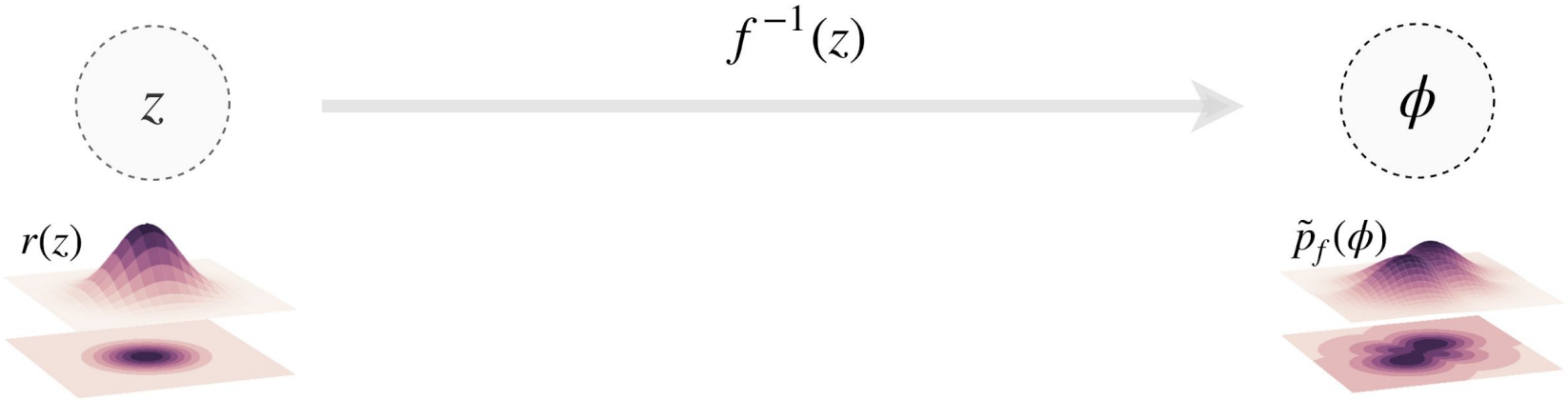
✘ Often few, e.g. O(1000), samples available (fewer than # vars!)

- Hard to use ML training paradigms that rely on existing samples from distribution

# Flow-based generative models

Using a change-of-variables, produce a distribution approximating what you want.

[Rezende & Mohamed 1505.05770]

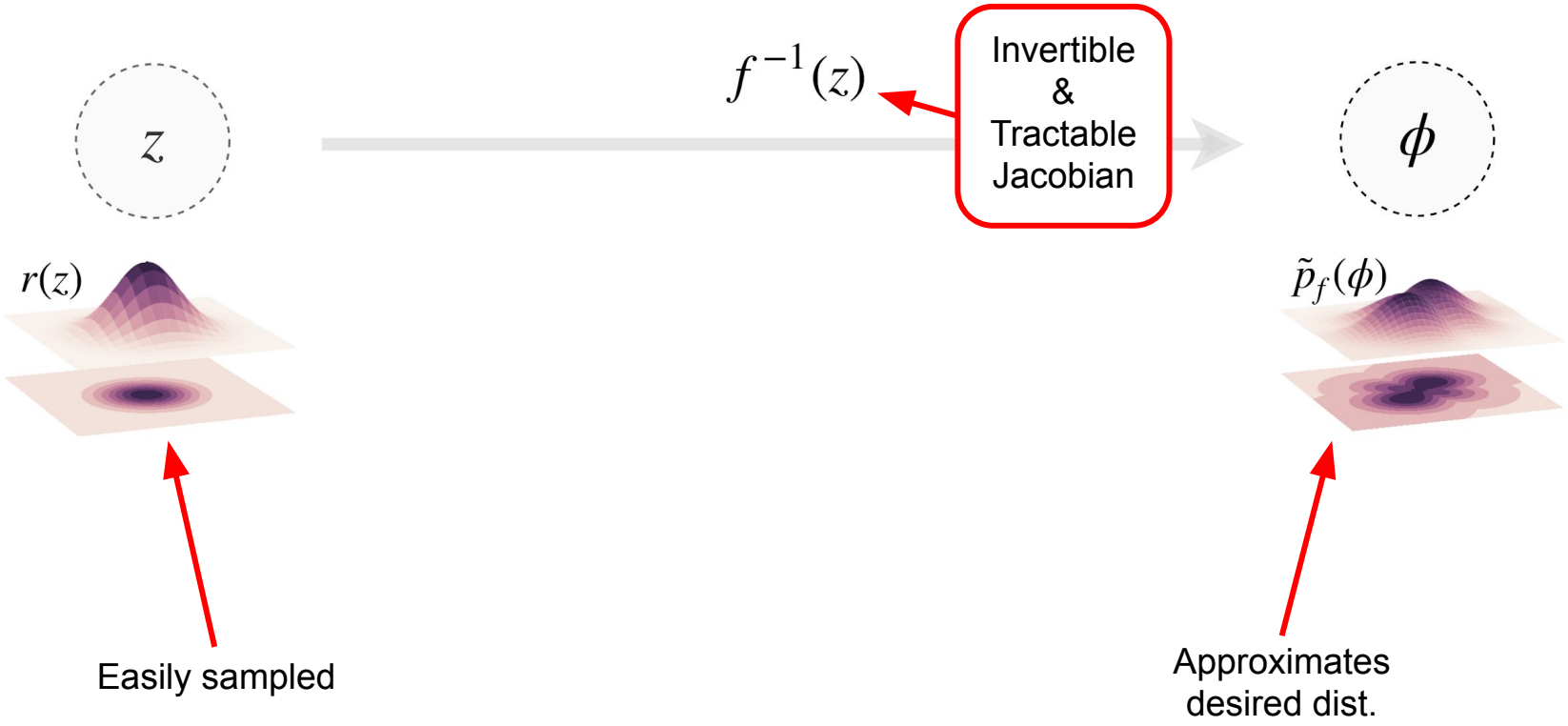


# Flow-based generative models

Using a change-of-variables, produce a distribution approximating what you want.

[Rezende & Mohamed 1505.05770]

$$\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)$$

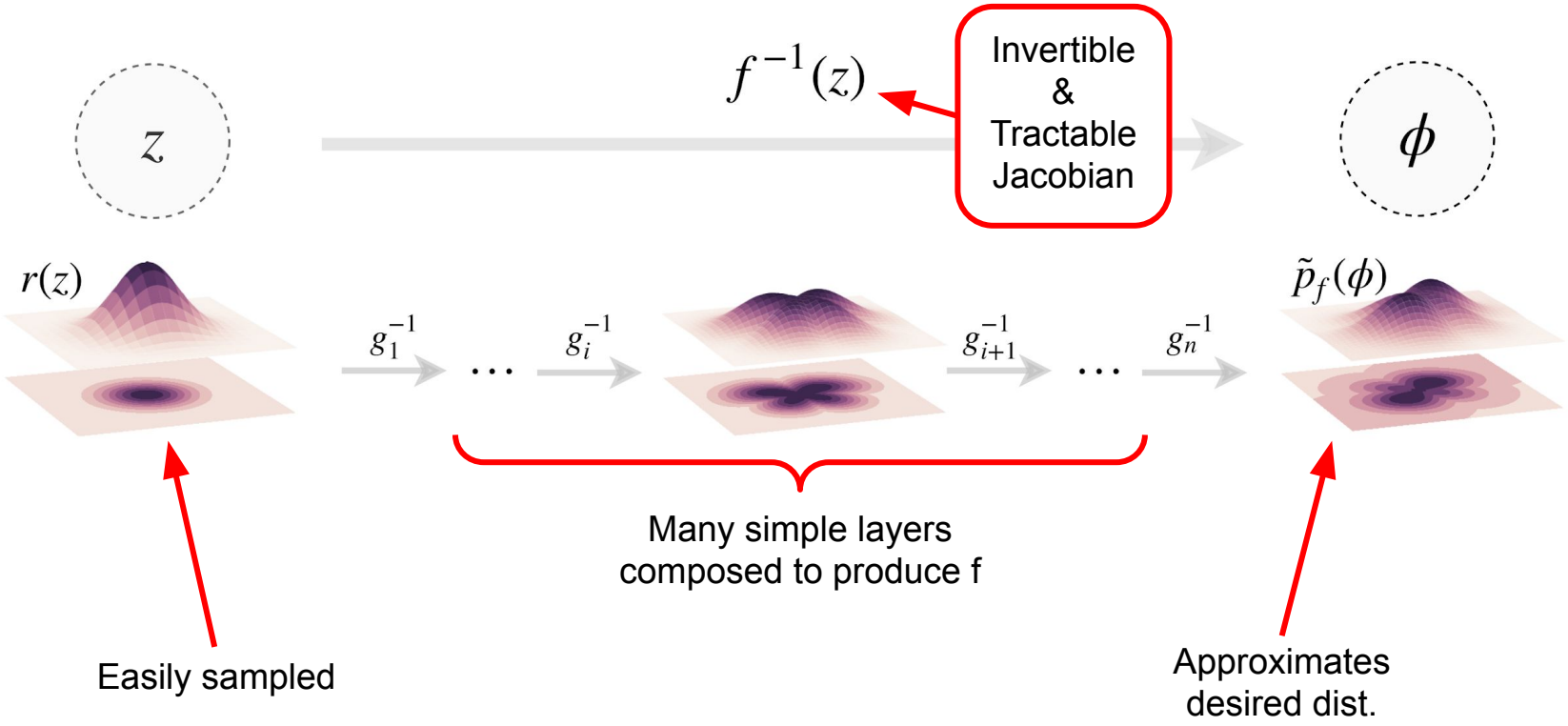


# Flow-based generative models

We chose real non-volume preserving (real NVP) flows for our work.

[Dinh et al. 1605.08803]

$$\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)$$

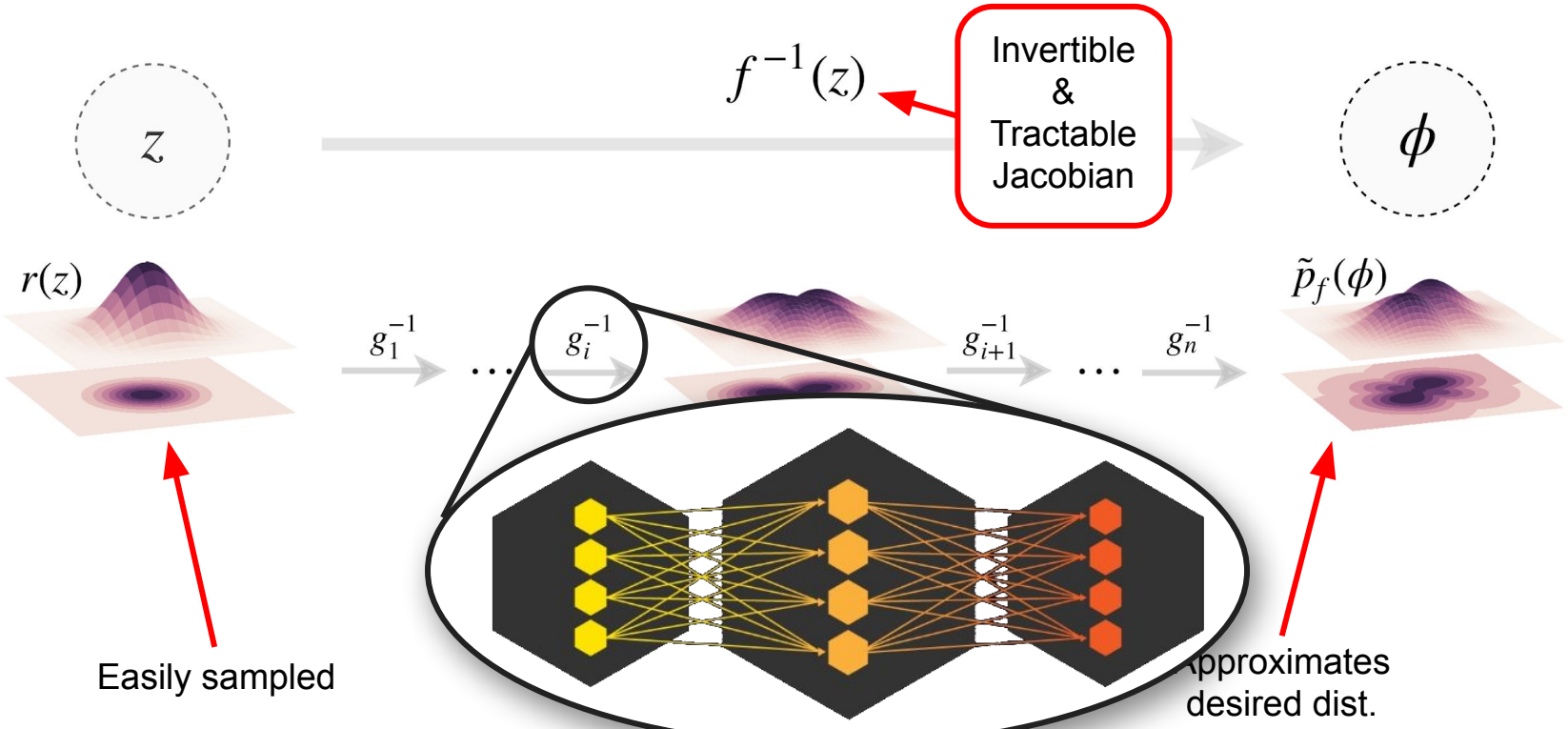


# Flow-based generative models

We chose real non-volume preserving (real NVP) flows for our work.

[Dinh et al. 1605.08803]

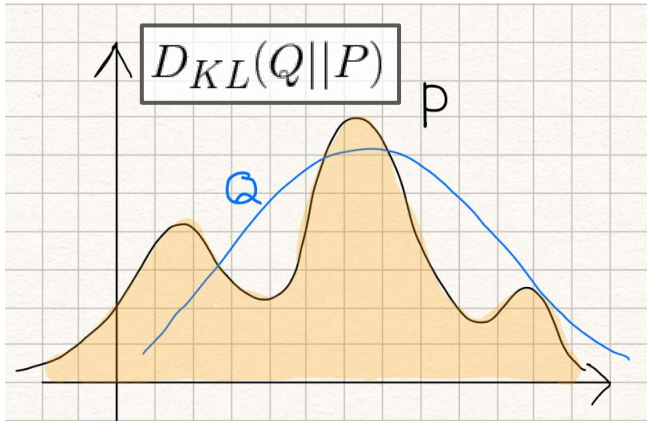
$$\tilde{p}_f(\phi) = \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|^{-1} r(z)$$



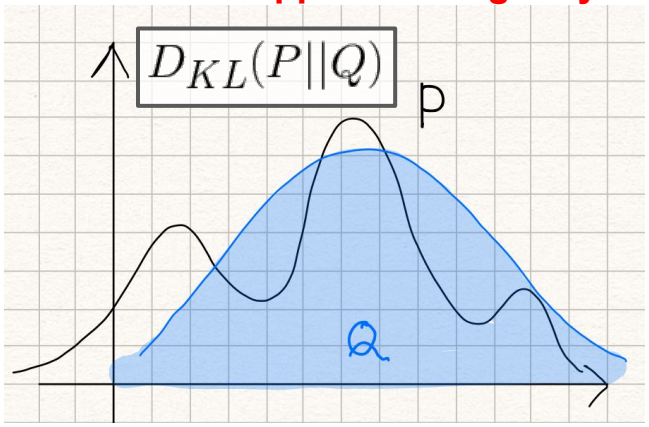
# Training by minimizing a loss function

- Desired distribution is known up to normalization: $p(\phi) = e^{-S(\phi)}/Z$
- KL divergence $D_{KL} \geq 0$ measures "distance" between distributions

**"badness" of approximating Q by P**




**"badness" of approximating P by Q**




[Shibuya, "Demystifying KL Divergence"]

- For our application, train to **minimize shifted KL divergence**

$$L(\tilde{p}_f) := D_{KL}(\tilde{p}_f||p) - \log Z$$

shift removes unknown normalization Z

[Zhang, E, Wang 1809.10188]

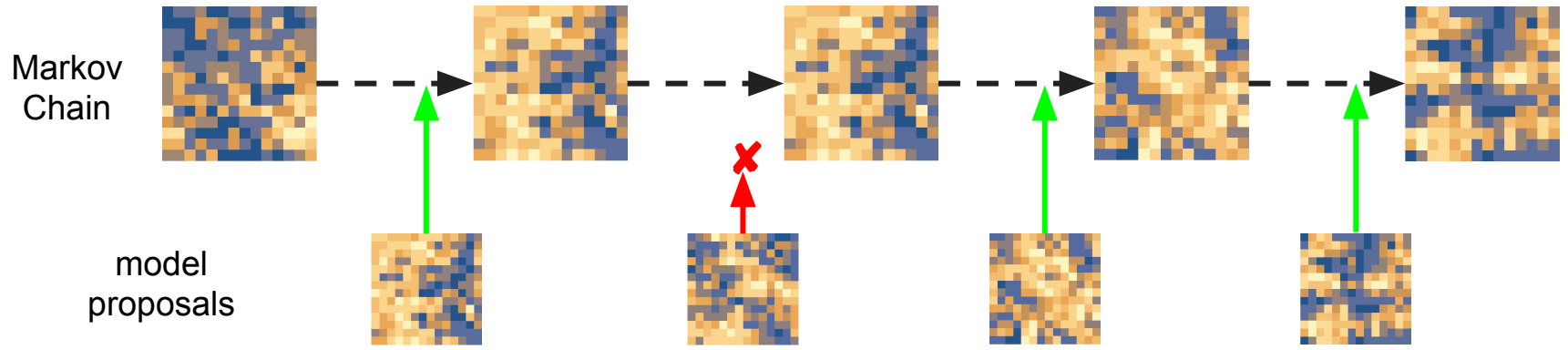
# Making things exact via MCMC

- Borrow idea from standard approach to lattice physics: Markov Chain Monte Carlo (MCMC)

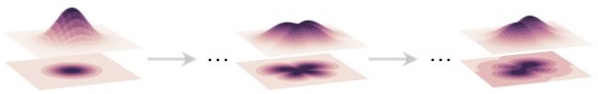
- Use generative model for proposals in a Metropolis-Hastings step

$$A(\phi^{(i-1)}, \phi') = \min\left(1, \frac{\tilde{p}(\phi^{(i-1)})}{p(\phi^{(i-1)})}\frac{p(\phi')}{\tilde{p}(\phi')}\right)$$

proposal independent of previous sample



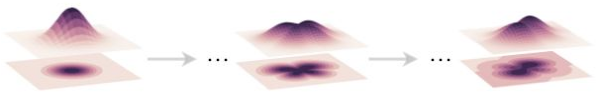
# Overview of algorithm

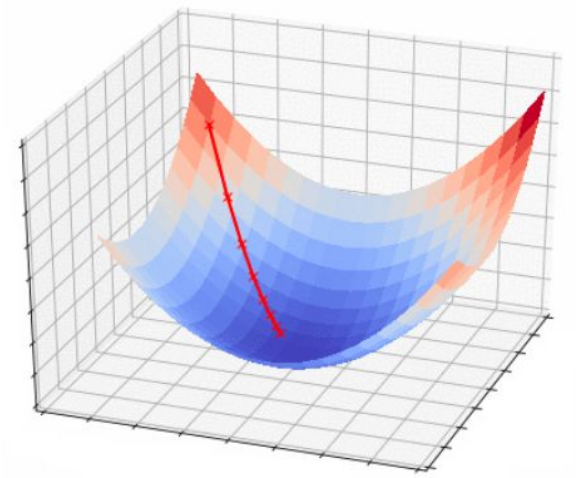


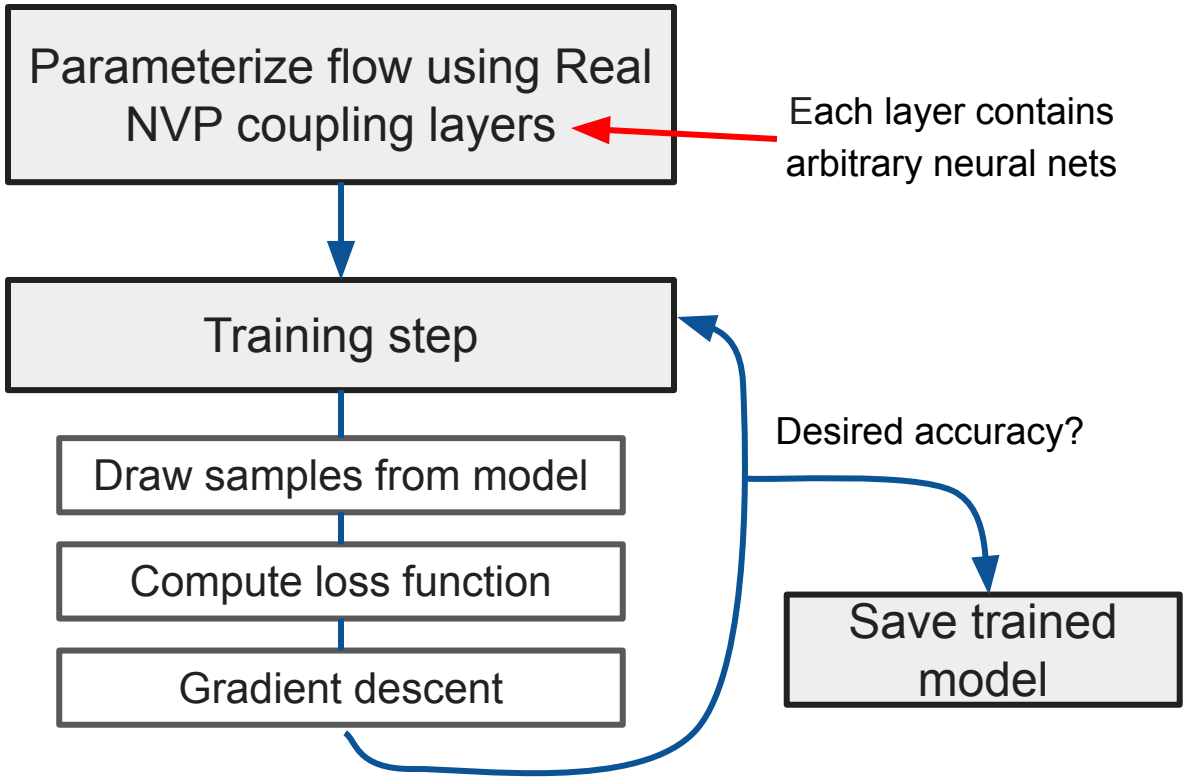
Parameterize flow using Real NVP coupling layers

Each layer contains arbitrary neural nets

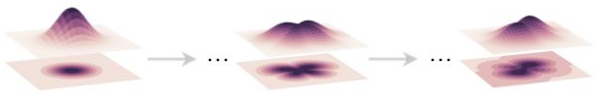
# Overview of algorithm

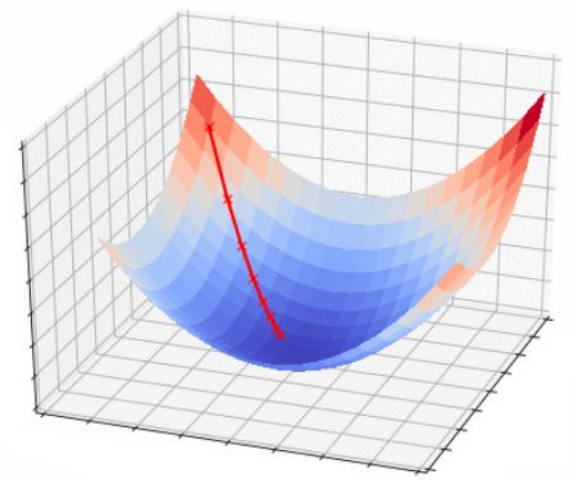


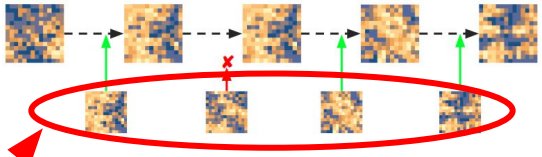




# Overview of algorithm







generating samples is "embarrassingly parallel"

Parameterize flow using Real NVP coupling layers

Each layer contains arbitrary neural nets

Training step

Draw samples from model

Compute loss function

Gradient descent

Desired accuracy?

Save trained model

Markov chain using samples from model

# Toy model: scalar $\phi^4$ lattice field theory

- One real number $\phi(x) \in (-\infty,\infty)$ per lattice site x (2D lattice)

- Action: relativistic scalar with quartic coupling

$$S(\phi) = \sum_x \left( \sum_y \frac{1}{2}\phi(x)\Box(x,y)\phi(y) + \frac{1}{2}m^2\phi(x)^2 + \lambda\phi(x)^4 \right)$$

# Toy model: scalar $\phi^4$ lattice field theory

- One real number $\phi(x) \in (-\infty,\infty)$ per lattice site x (2D lattice)
- Action: relativistic scalar with quartic coupling

$$S(\phi) = \sum_x \left( \sum_y \frac{1}{2}\phi(x)\Box(x,y)\phi(y) + \frac{1}{2}m^2\phi(x)^2 + \lambda\phi(x)^4 \right)$$

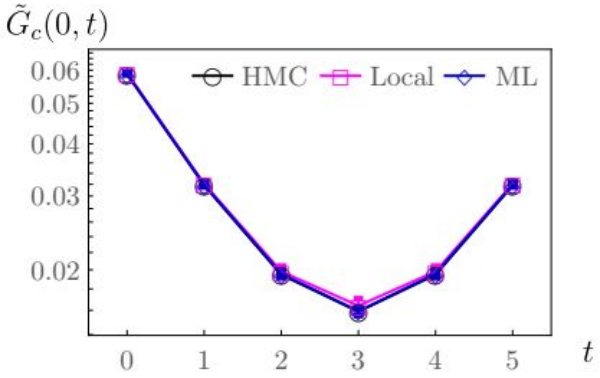
- 5 lattice sizes $L^2 = \{6^2, 8^2, 10^2, 12^2, 14^2\}$ with bare parameters tuned to follow a line of constant physics (symmetric phase)

| | E1 | E2 | E3 | E4 | E5 |
|---|---|---|---|---|---|
| $L$ | 6 | 8 | 10 | 12 | 14 |
| $m^2$ | $-4$ | $-4$ | $-4$ | $-4$ | $-4$ |
| $\lambda$ | 6.975 | 6.008 | 5.550 | 5.276 | 5.113 |
| $m_p L$ | 3.96(3) | 3.97(5) | 4.00(4) | 3.96(5) | 4.03(6) |

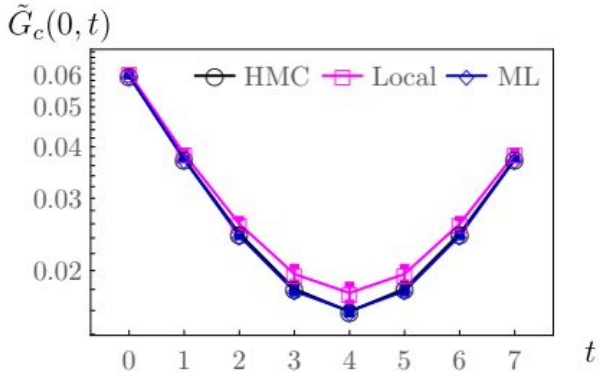
- **HMC** and **local Metropolis** compared against our **ML method**

# Comparing observables (1)

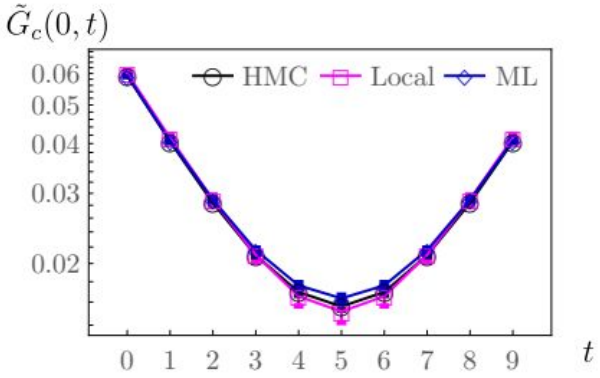
$$G_c(x) = \frac{1}{V} \sum_y \langle \phi(y)\phi(y+x) \rangle$$



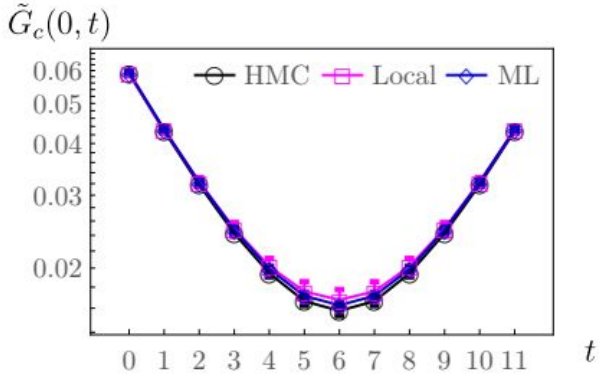

(a) $L = 6$ ensembles



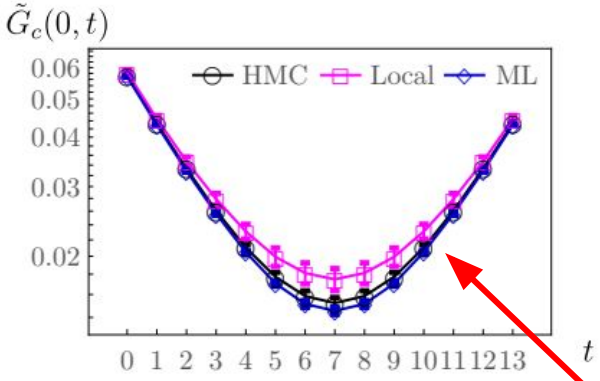

(b) $L = 8$ ensembles




(c) $L = 10$ ensembles




(d) $L = 12$ ensembles



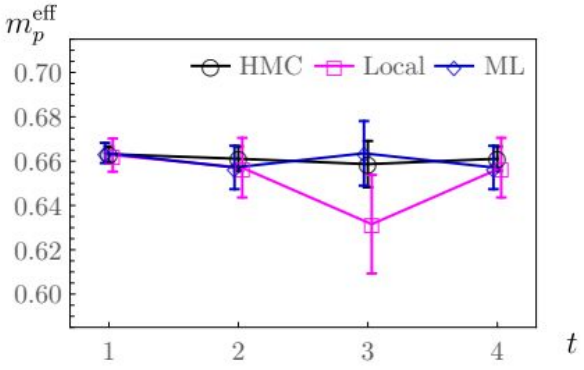

(e) $L = 14$ ensembles

correlation falls off with separation in both directions on periodic lattice

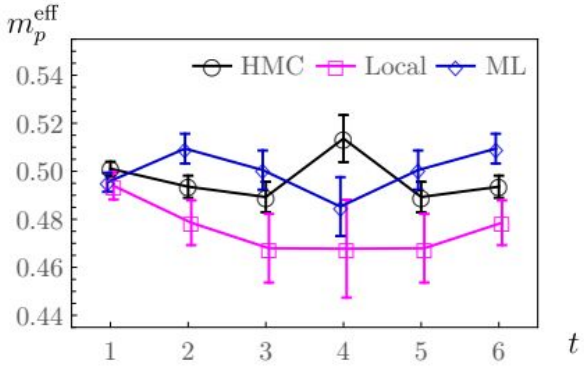
Green's functions...

# Comparing observables (2)

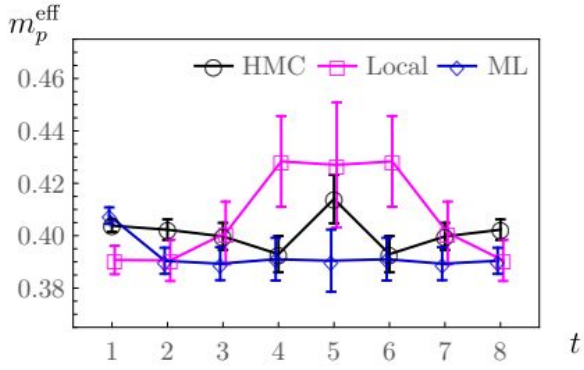
$$m_p = -\partial_t \log \left\langle \tilde{G}_c(0,t) \right\rangle$$



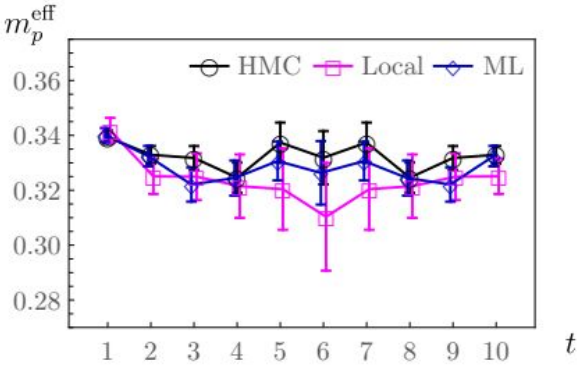

(a) $L = 6$ ensembles



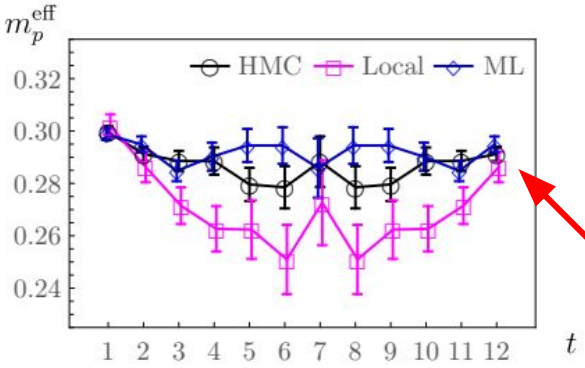

(b) $L = 8$ ensembles




(c) $L = 10$ ensembles




(d) $L = 12$ ensembles



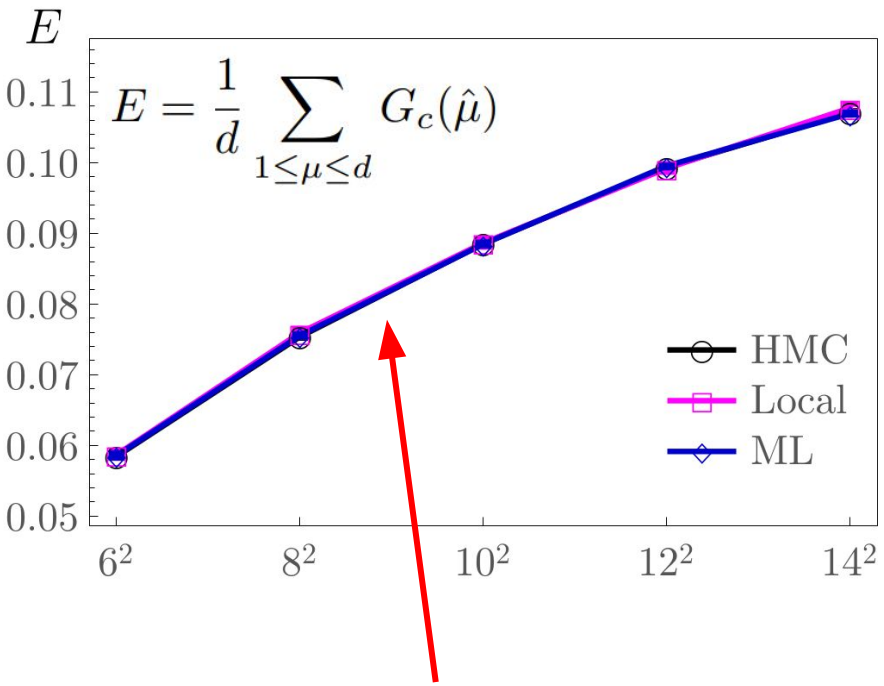

(e) $L = 14$ ensembles

effective pole mass plateaus to true pole mass

Green's functions...

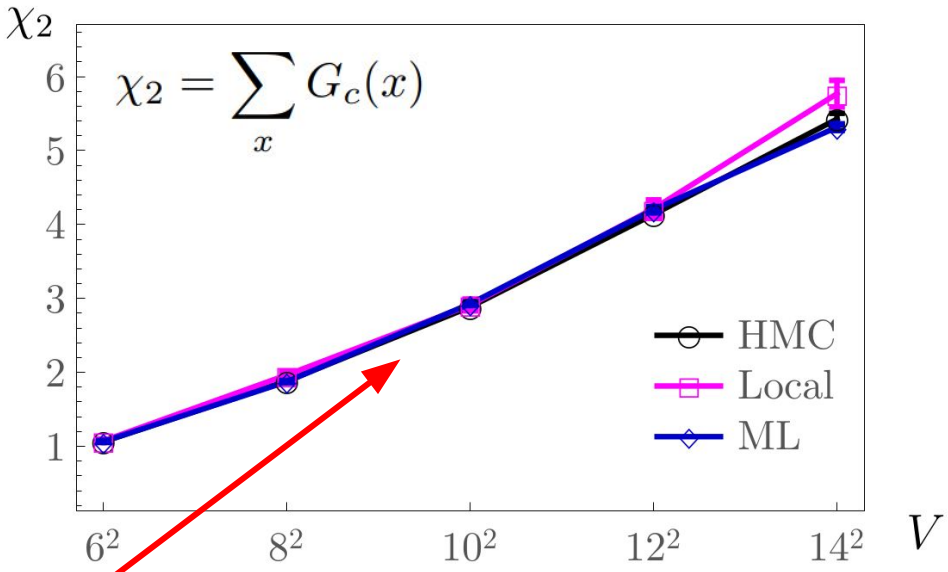
... and pole masses agree.

# Comparing observables (3)



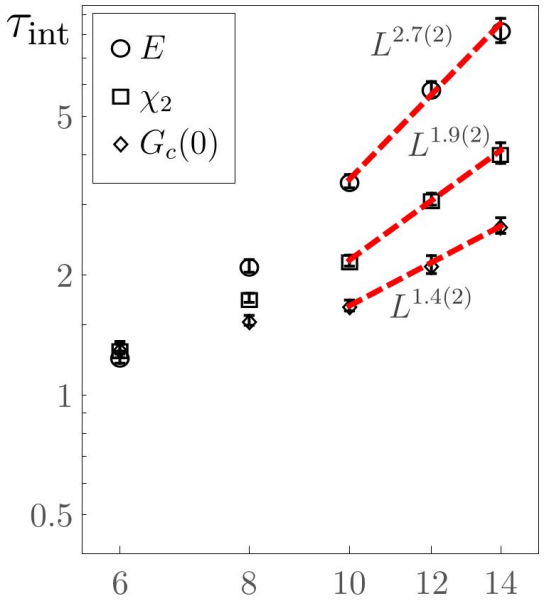

nearest neighbor response grows with shrinking lattice spacing

Ising energy and two-point susceptibility agree.



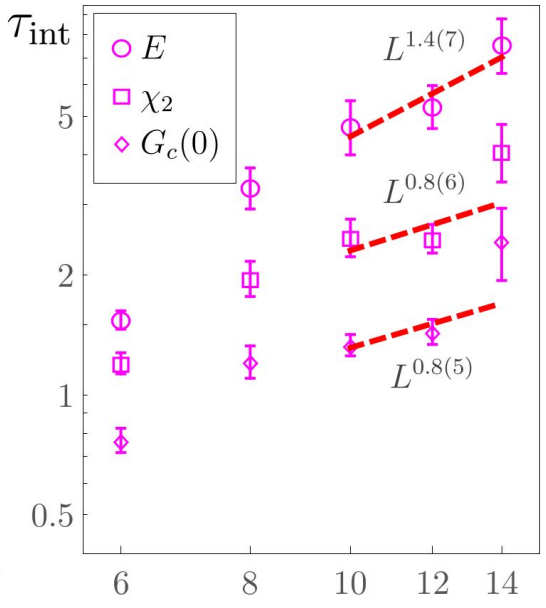

susceptibility (total lattice response to an impulse) diverges in the continuum limit

# Critical slowing down



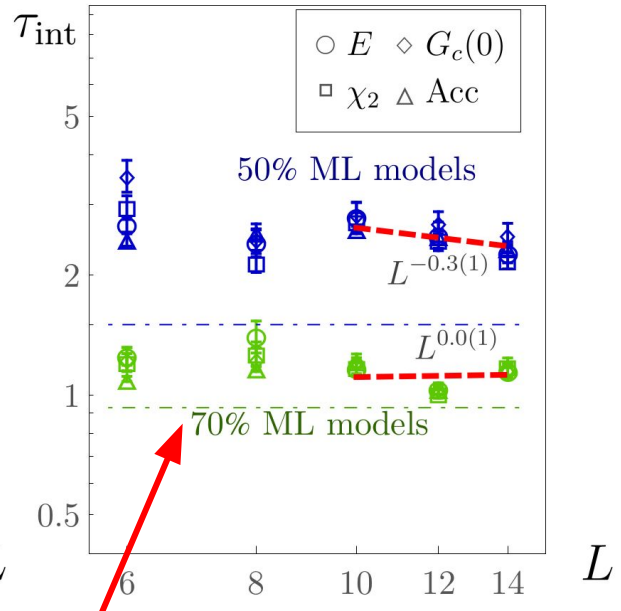

**HMC**

Dynamical critical exponents
$z_{\mathcal{O}} = 1.4$ to $2.7$




**Local**

Dynamical critical exponents
$z_{\mathcal{O}} = 0.8$ to $1.4$



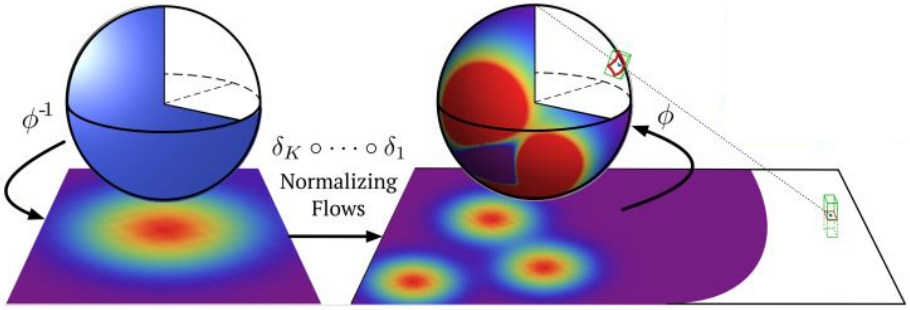

**ML models**

Dynamical critical exponents
compatible with zero

by spending time training up-front,
autocorrelations are fixed during sampling

# Towards gauge (and other) theories

- Real NVP only directly works on fields taking **real values** $\phi(x) \in (-\infty,\infty)$
- What about fields taking values in compact domains (gauge theories, O(N) models, etc.)?
  - Stereographic projection coupled with standard methods may work [Gemici, Rezende, Mohammed 1611.02304]

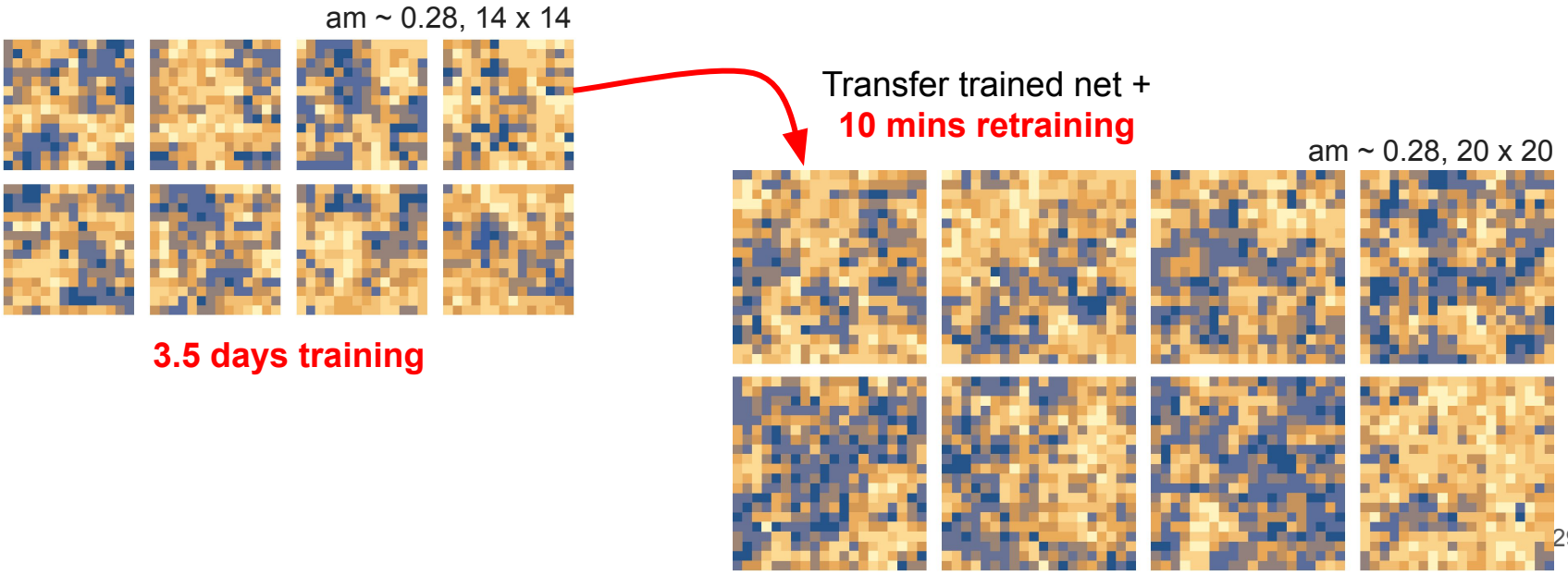



- What about discrete models (Ising, Potts, etc.)?
  - Some recent ideas emerging [Ziegler & Rush 1901.10548]

# Better choices for neural networks

- Our $\phi^4$ results use fully-connected neural networks, but Real NVP authors suggest convolutions, and hierarchical structure
  - Translational invariance, improved scaling
  - Preliminary results for $\phi^4$ indicates that this works!

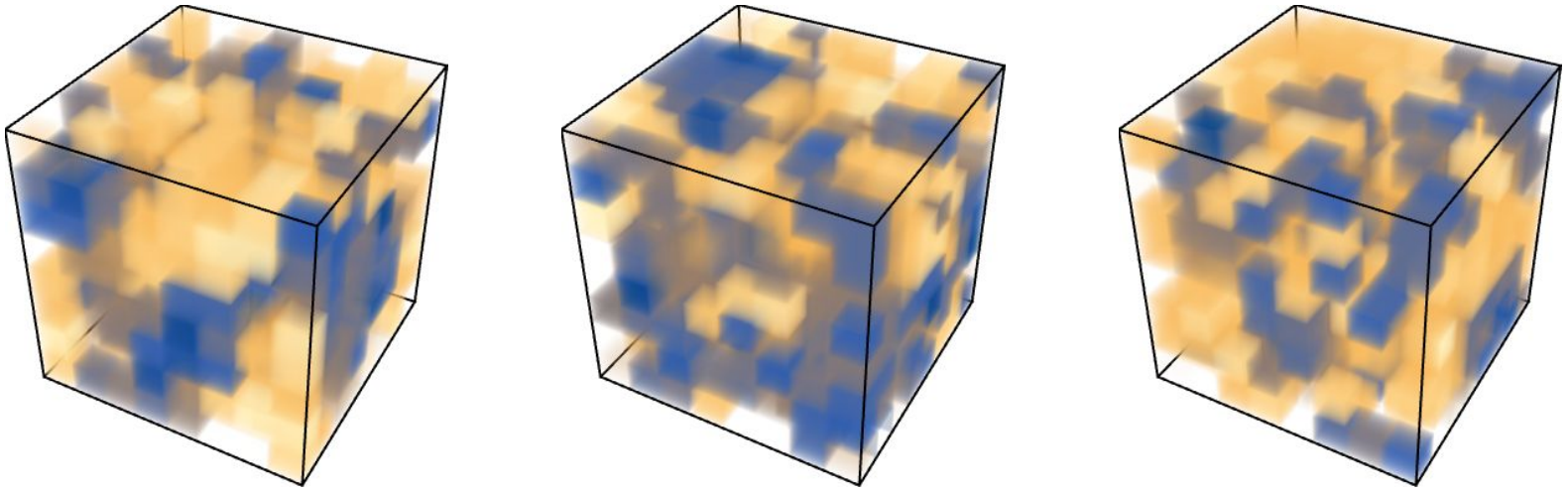
- Convolutions also make scaling physical volume easy



# Towards higher dimensions

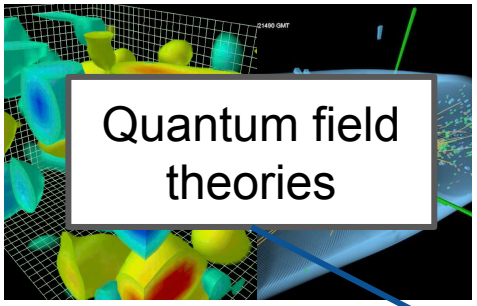
- Costs scale up, but no theoretical obstacle
- Preliminary: 3D $\phi^4$ easily accessible, (solvable) memory bottleneck for 4D

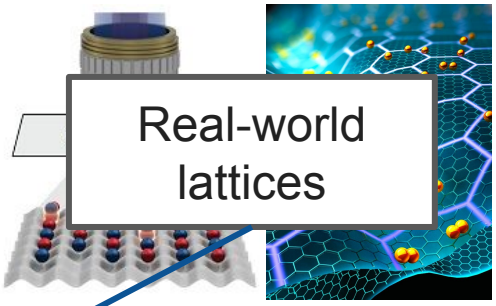
30% acc, no hyperparameter tuning required



Samples generated for $\phi^4$ theory with V=$8^3$, $m^2$=-6.0, λ=14.590
mL ~ 4, matching CSD investigation of [Vierhaus, Thesis, doi:10.18452/14138]

# Machine learning for lattice theories



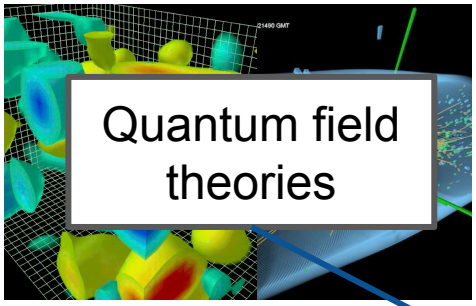




lattice theories

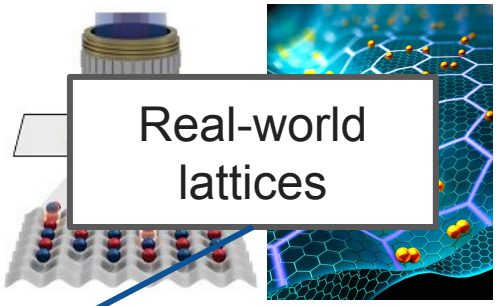
numerical methods

- Spectrum
- Structure
- Thermodynamics
- ...

# Machine learning for lattice theories



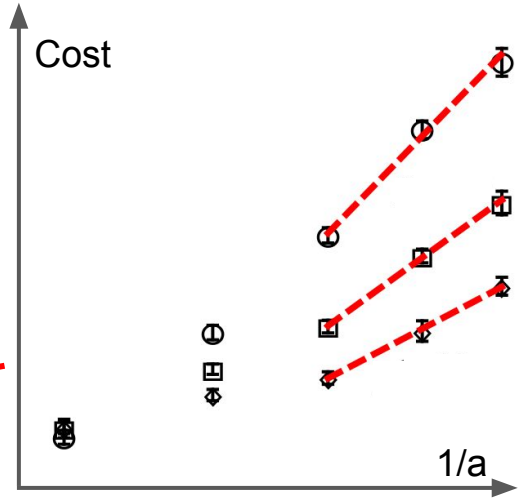
Quantum field theories



Real-world lattices

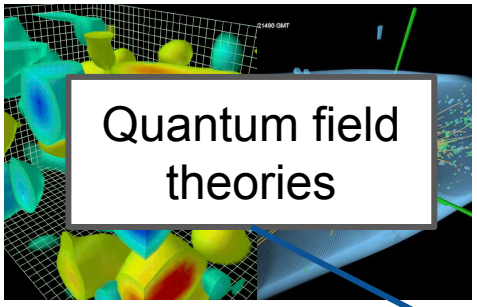
lattice theories

numerical methods

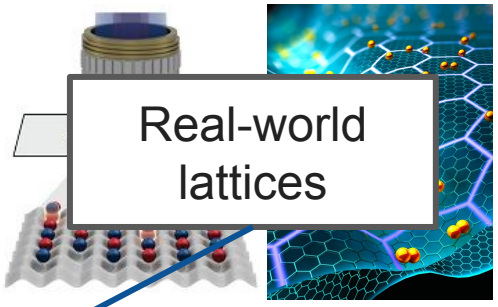



- Spectrum
- Structure
- Thermodynamics
- ...

# Machine learning for lattice theories



Quantum field theories

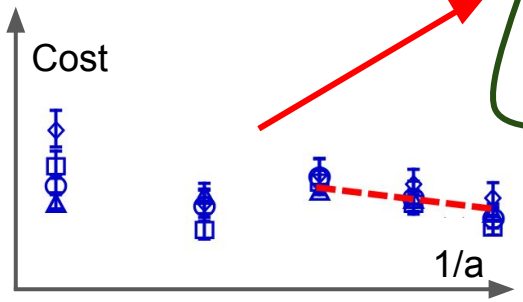


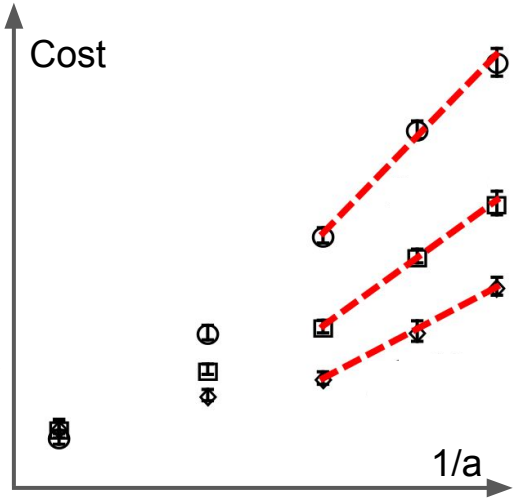
Real-world lattices

lattice theories

numerical methods

+ ML




- Spectrum
- Structure
- Thermodynamics
- ...

# Backup slides

# Image generation via ML

1. Likelihood free methods: [Goodfellow et al. 1406.2661]

   E.g. Generative Adversarial Networks (GANs)

   ✘ Needs many real samples

   ✘ No associated likelihood for each produced sample

2. Autoencoding: [Kingma & Welling 1312.6114]

   E.g. Variational Auto-Encoders (VAEs)

   ✔ Good for human interpretability

   ✘ Same issues as GANs

[Shen & Liu 1612.05363]

3. Flow-based: [Rezende & Mohamed 1505.05770]

   E.g. Normalizing flows

   ✔ Exactly known likelihood for each sample

   ✔ Can be trained with samples from itself

# Image generation via ML

1. Likelihood free methods: [Goodfellow et al. 1406.2661]

   E.g. Generative Adversarial Networks (GANs)

   ✘ Needs many real samples

   ✘ No associated likelihood for each produced sample

2. Autoencoding: [Kingma & Welling 1312.6114]

   E.g. Variational Auto-Encoders (VAEs)

   ✔ Good for human interpretability

   ✘ Same issues as GANs

[Shen & Liu 1612.05363]

3. Flow-based: [Rezende & Mohamed 1505.05770]

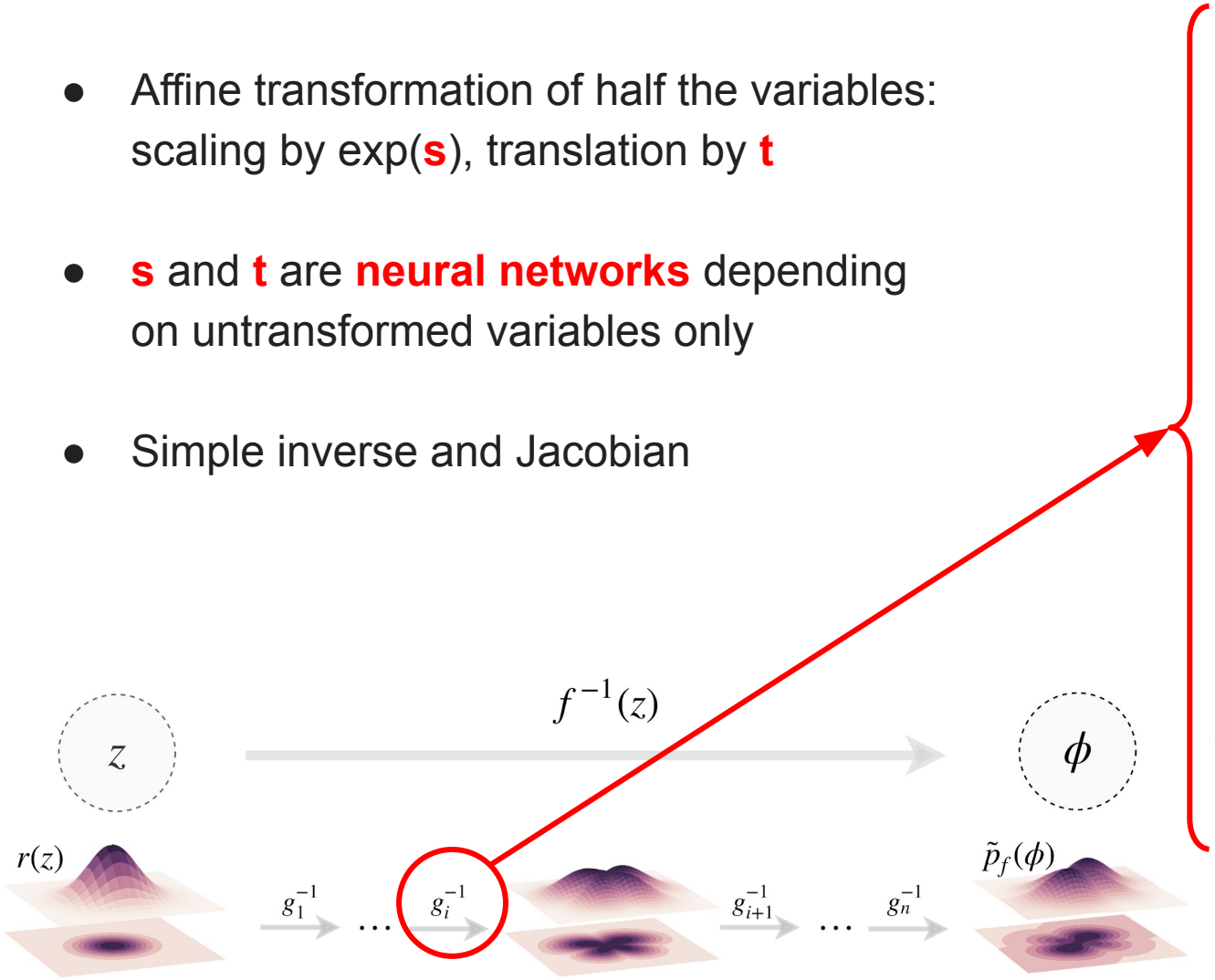
   E.g. Normalizing flows

   ✔ Exactly known likelihood for each sample

   ✔ Can be trained with samples from itself

# Real NVP coupling layer

- Affine transformation of half the variables: scaling by exp(**s**), translation by **t**
- **s** and **t** are **neural networks** depending on untransformed variables only
- Simple inverse and Jacobian



# Loss function: shifted KL divergence

- Desired distribution is known up to normalization: $p(\phi) = e^{-S(\phi)}/Z$

- For our application, train to **minimize shifted KL divergence**

$$L(\tilde{p}_f) := D_{KL}(\tilde{p}_f||p) - \log Z$$
$$= \int \prod_j d\phi_j \, \tilde{p}_f(\phi) \left(\log \tilde{p}_f(\phi) + S(\phi)\right)$$

shift removes unknown normalization Z

- This loss allows **self-training**: sampling with respect to model distribution $\tilde{p}_f(\phi)$ to estimate loss

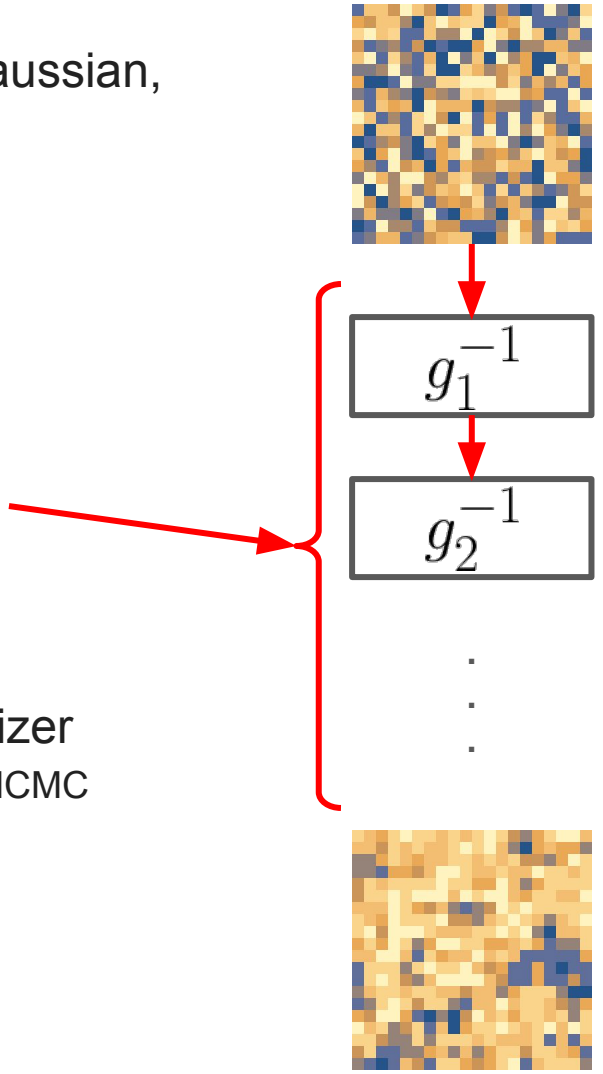
# ML model for scalar lattice field theory

- Prior distribution chosen to be uncorrelated Gaussian, i.e. for each site x,

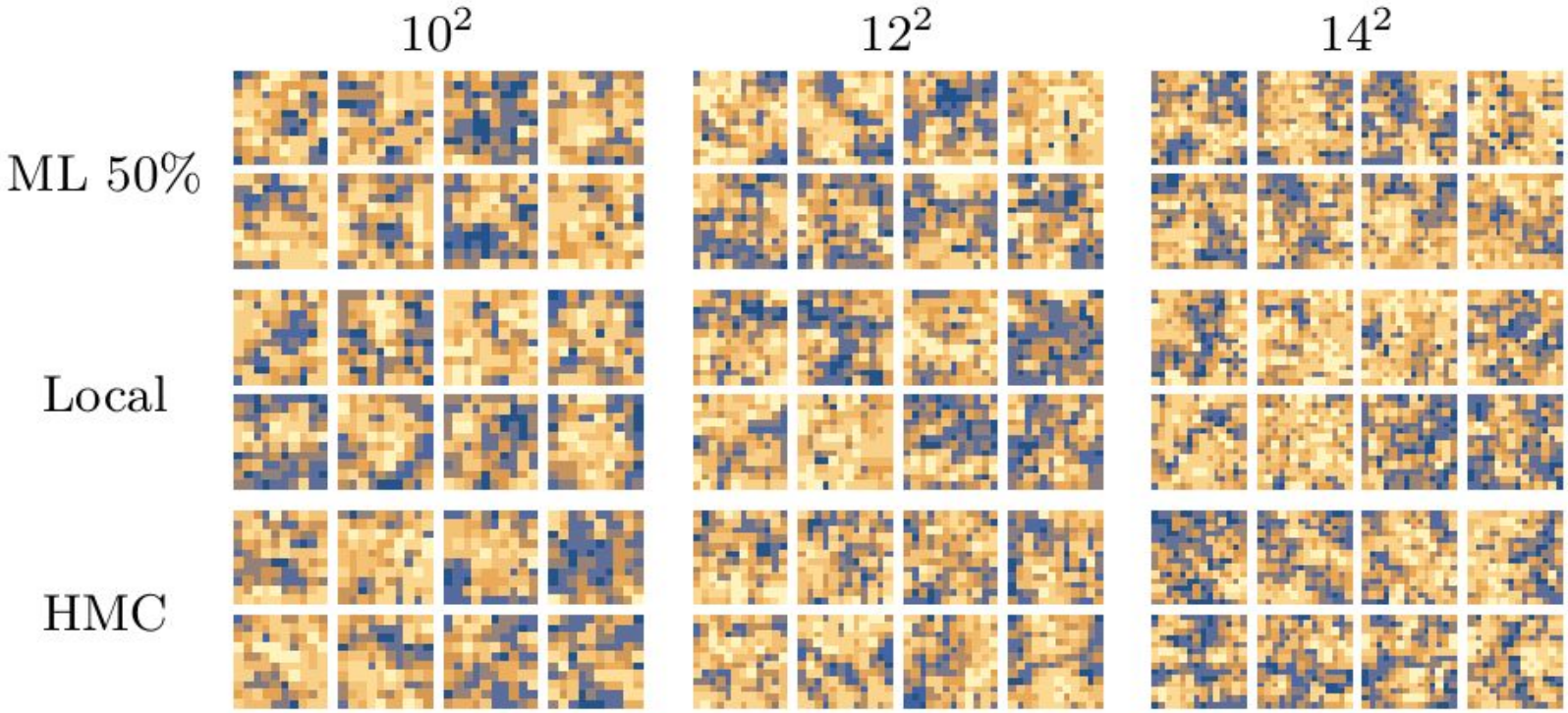
$$\phi(x) \sim \mathcal{N}(0,1)$$

- Real NVP model:
  - 8-12 Real NVP coupling layers
  - Alternating checkerboard pattern for variable split
  - 2-6 fully connected layers with 100-1024 hidden units

- Trained using shifted KL loss with Adam optimizer
  - Target fixed acceptance rate in Metropolis-Hastings MCMC



# Samples from ML model vs standard algorithms



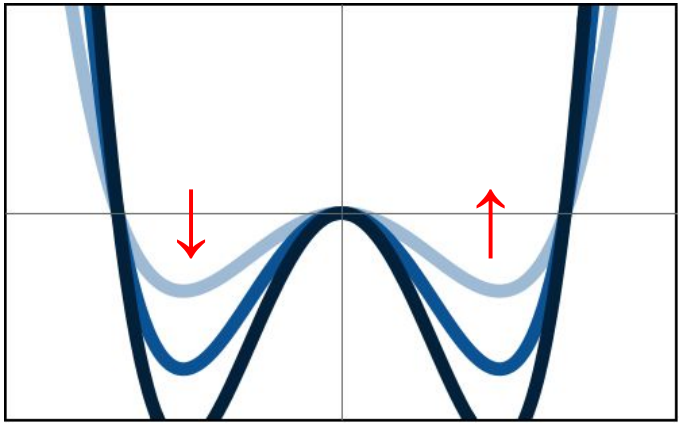

By eye, ML model produces **varied samples** and **correlations** at the right scale

# Physical limits of scalar $\phi^4$ lattice field theory

$$S(\phi) = \sum_x \left( \sum_y \frac{1}{2}\phi(x)\Box(x,y)\phi(y) + \frac{1}{2}m^2\phi(x)^2 + \lambda\phi(x)^4 \right)$$

$\lambda \to \infty$
$m^2/\lambda < 0$

Ising



$\lambda \to 0$

Gaussian