Flow-based sampling for lattice field theories

Gurtej Kanwar
University of Bern

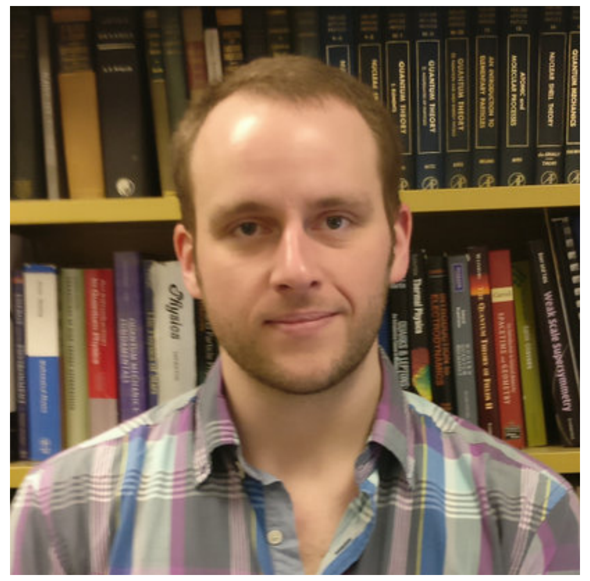
Lattice 2023
July 31, 2023 | Fermilab (Batavia, IL)

Massachusetts Institute of Technology

The NSF Institute for Artificial Intelligence and Fundamental Interactions


Phiala Shanahan


Denis Boyda


Dan Hackett


Fernando Romero-López


Julian Urban


Ryan Abbott


Betsy Tian

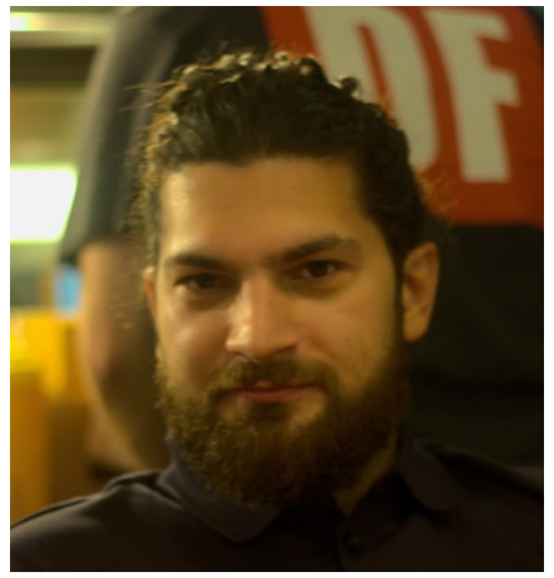
NEW YORK UNIVERSITY


Michael Albergo

DeepMind


Sébastien Racanière


Danilo Rezende


Aleksander Botev


Ali Razavi

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON


Kyle Cranmer

Motivations for flow-based sampling

Address **challenges in Monte Carlo** for Lattice Field Theory.

- Critical slowing down
- Topological freezing
- Parallelization
- Storage

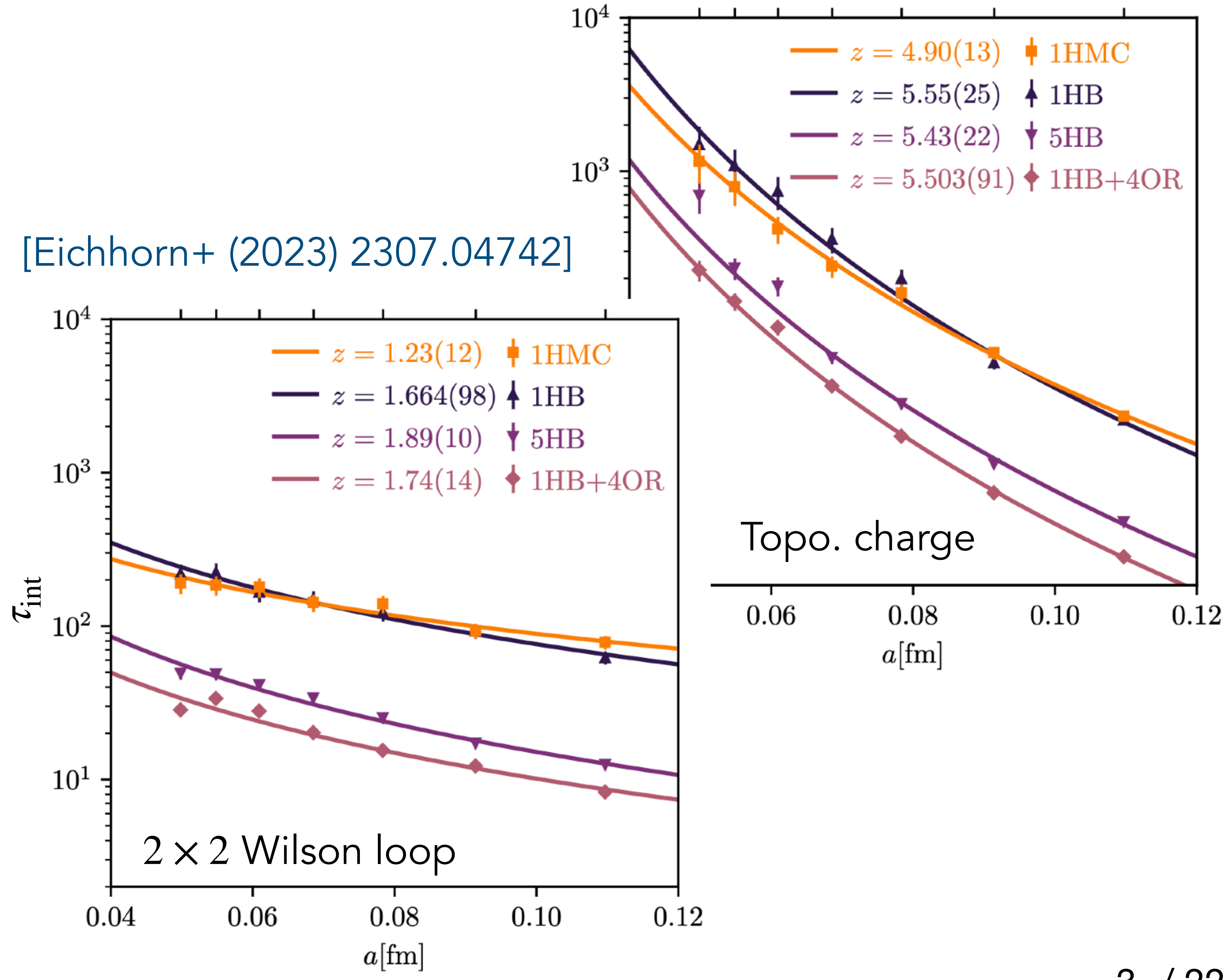
Goal: Boltzmann distribution for discretized field theory action

$$p(U) = e^{-S[U]}/Z$$

But also open the door to new paradigms.

- Estimate Z
- Draw correlated samples
- Parallel tempering
- Parameter scans
- ...

[Eichhorn+ (2023) 2307.04742]



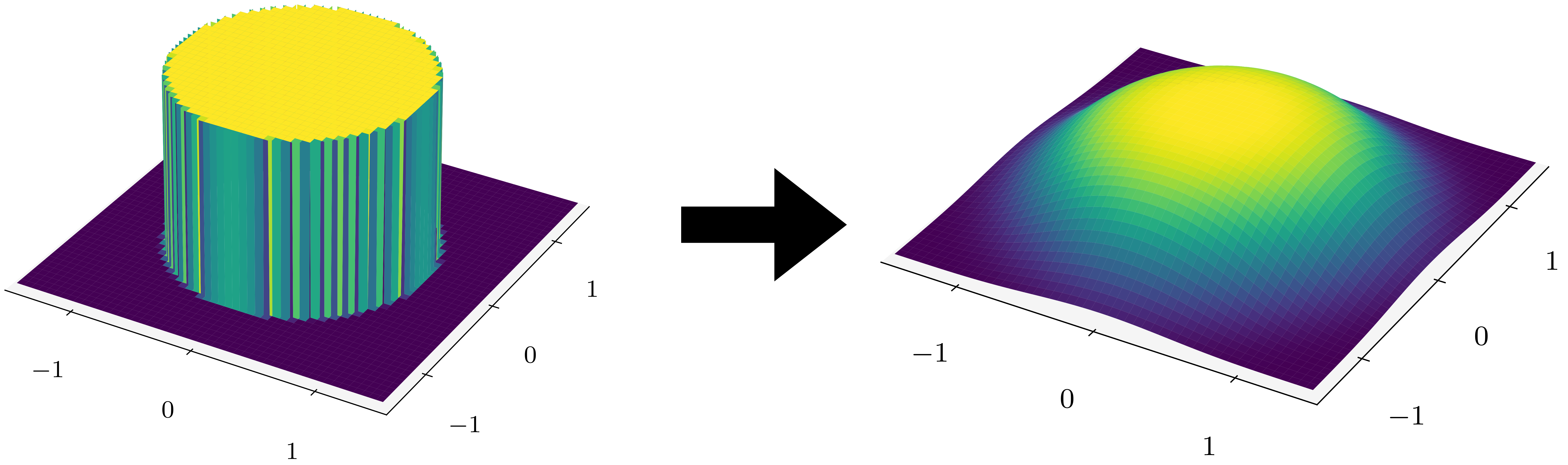
A taste of flow-based sampling

AKA a "normalizing flow"

[Tabak & Vanden-Eijnden CMS8 (2010) 217]
[Tabak & Turner CPA66 (2013) 145]

Box-Muller transform (Marsaglia polar form)

$$x' = \frac{x}{r}\sqrt{-2\ln r^2} \qquad y' = \frac{y}{r}\sqrt{-2\ln r^2}$$



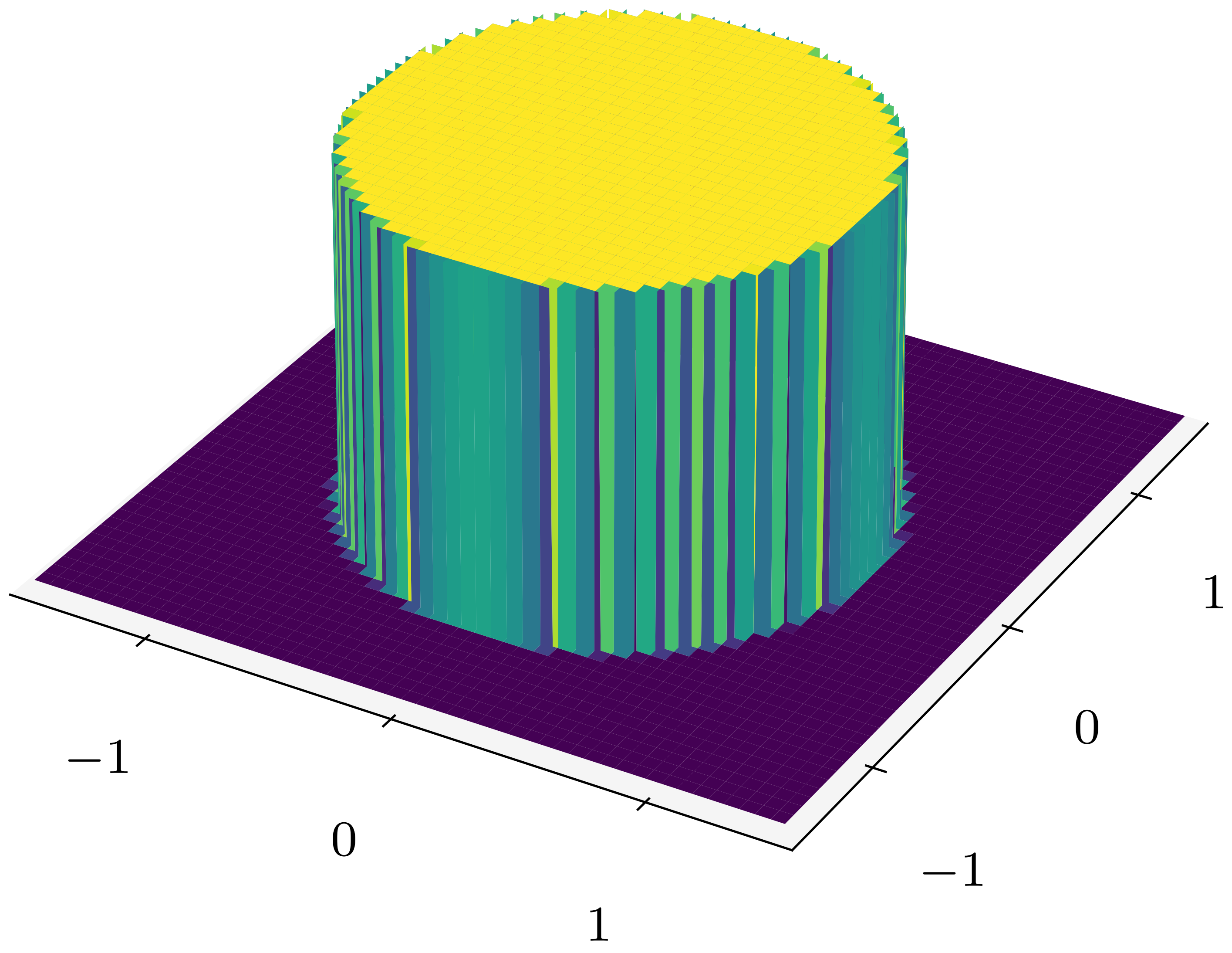
A taste of flow-based sampling

AKA a "normalizing flow"

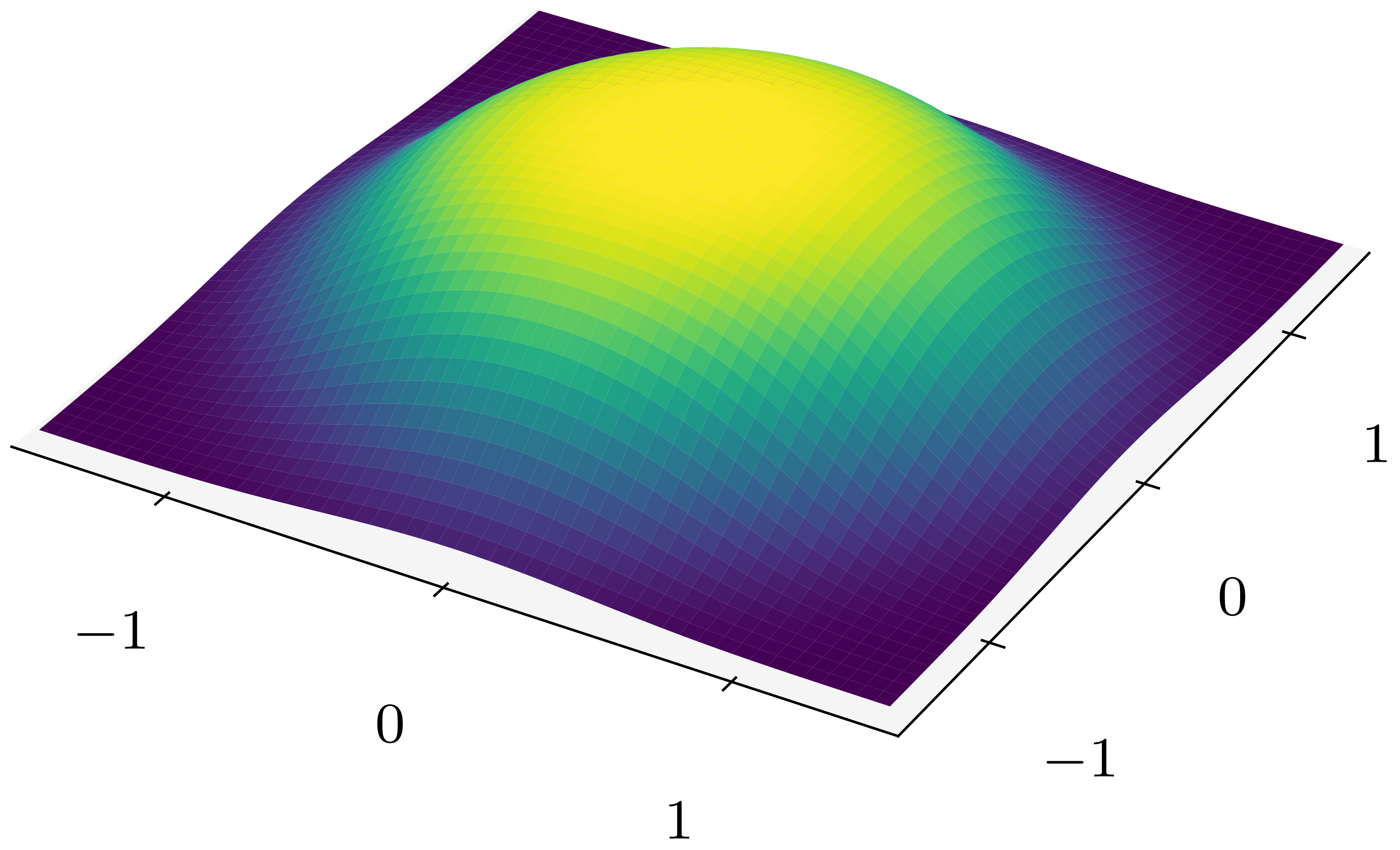
[Tabak & Vanden-Eijnden CMS8 (2010) 217]
[Tabak & Turner CPA66 (2013) 145]

Box-Muller transform (Marsaglia polar form)

$$x' = \frac{x}{r}\sqrt{-2\ln r^2} \qquad y' = \frac{y}{r}\sqrt{-2\ln r^2}$$



Flow f



(Simple) Prior density:
$r(x, y)$

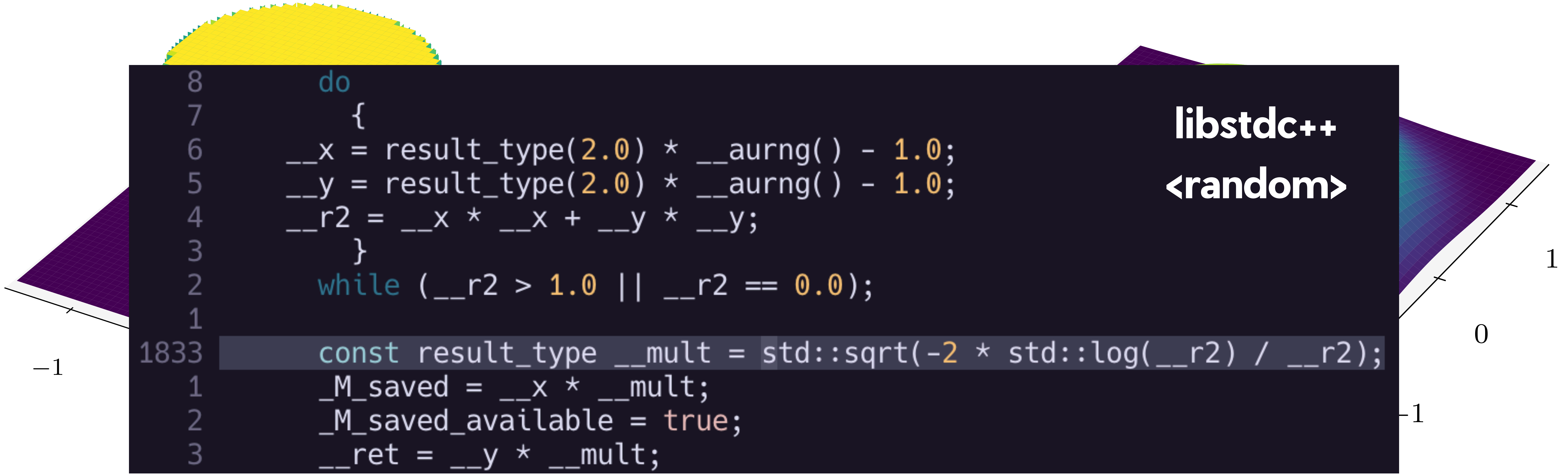
(More complex) Output density:
$q(x', y') = r(x, y)\,|\det J\,|^{-1}$

A taste of flow-based sampling

AKA a "normalizing flow"
[Tabak & Vanden-Eijnden CMS8 (2010) 217]
[Tabak & Turner CPA66 (2013) 145]

Box-Muller transform (Marsaglia polar form)

$$x' = \frac{x}{r}\sqrt{-2\ln r^2} \qquad y' = \frac{y}{r}\sqrt{-2\ln r^2}$$




```
    do
      {
    __x = result_type(2.0) * __aurng() - 1.0;
    __y = result_type(2.0) * __aurng() - 1.0;
    __r2 = __x * __x + __y * __y;
      }
    while (__r2 > 1.0 || __r2 == 0.0);

    const result_type __mult = std::sqrt(-2 * std::log(__r2) / __r2);
    _M_saved = __x * __mult;
    _M_saved_available = true;
    __ret = __y * __mult;
```

(Simple) Prior density:
$r(x, y)$

(More complex) Output density:
$q(x', y') = r(x, y)\,|\det J\,|^{-1}$

Machine-learned flows for LQFT

Machine learning + flows

[Rezende & Mohamed (2015) PMLR 37, 1530]

By making f learnable, we can approximate more complicated distributions.

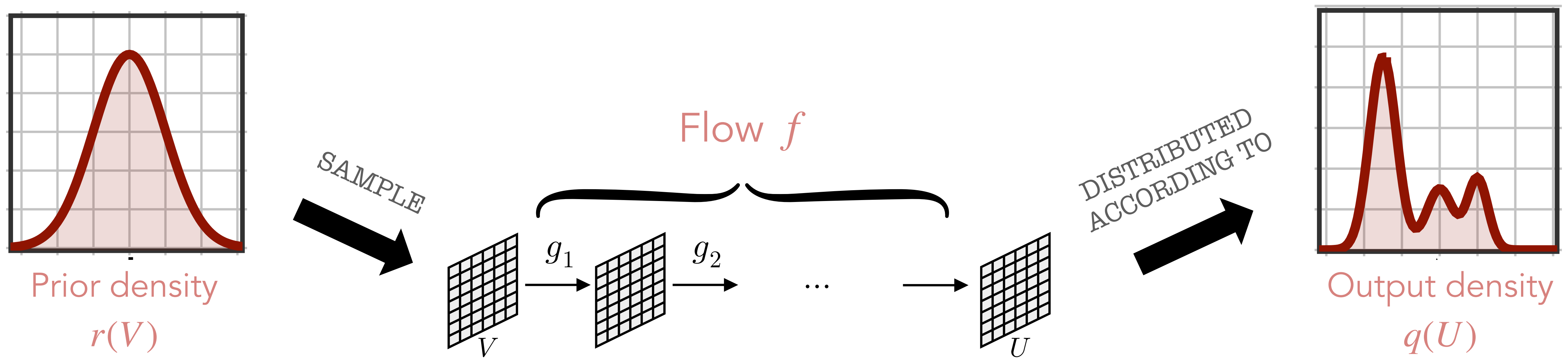
- Must be a **diffeomorphism** with **tractable Jacobian**

- Discrete learnable flows:

  [Dinh+ (2014) 1410.8516] [Dinh+ (2016) 1605.08803]

$$f = g_1 \circ \ldots \circ g_n$$
$$\det J = \det J_1 \cdot \ldots \cdot \det J_n$$

[**GK** (2021) PhD Thesis, MIT]



Note: U, V indicate generic fields: gauge fields, scalar fields, etc.

- Continuous learnable flows:

  [Chen+ (2018) 1806.07366] [Zhang+ (2018) 1809.10188]

$$f(V) = \int_0^T dt\, \nabla \varphi(U(t);t)\Big|_{U(0)=V} + V \qquad \ln \det J = -\int_0^T dt\, \nabla^2 \varphi(U(t);t)$$

The "trivializing map" is a special continuous flow
[Lüscher CMP293 (2010) 899]

Note: For compact spaces, derivatives and integrals should be appropriately modified to act in the space.

The early days

Trivializing map

[Lüscher CMP293 (2010) 899]

"Monge-Ampère flows" for scalar field theory

[Zhang+ (2018) 1809.10188]

"RealNVP coupling layers"

[Dinh+ (2016) 1605.08803]

Self-learning Monte Carlo

[Huang & Wang PRB95 (2017) 035105, Liu+ PRB95 (2017) 041101, … and many more …]

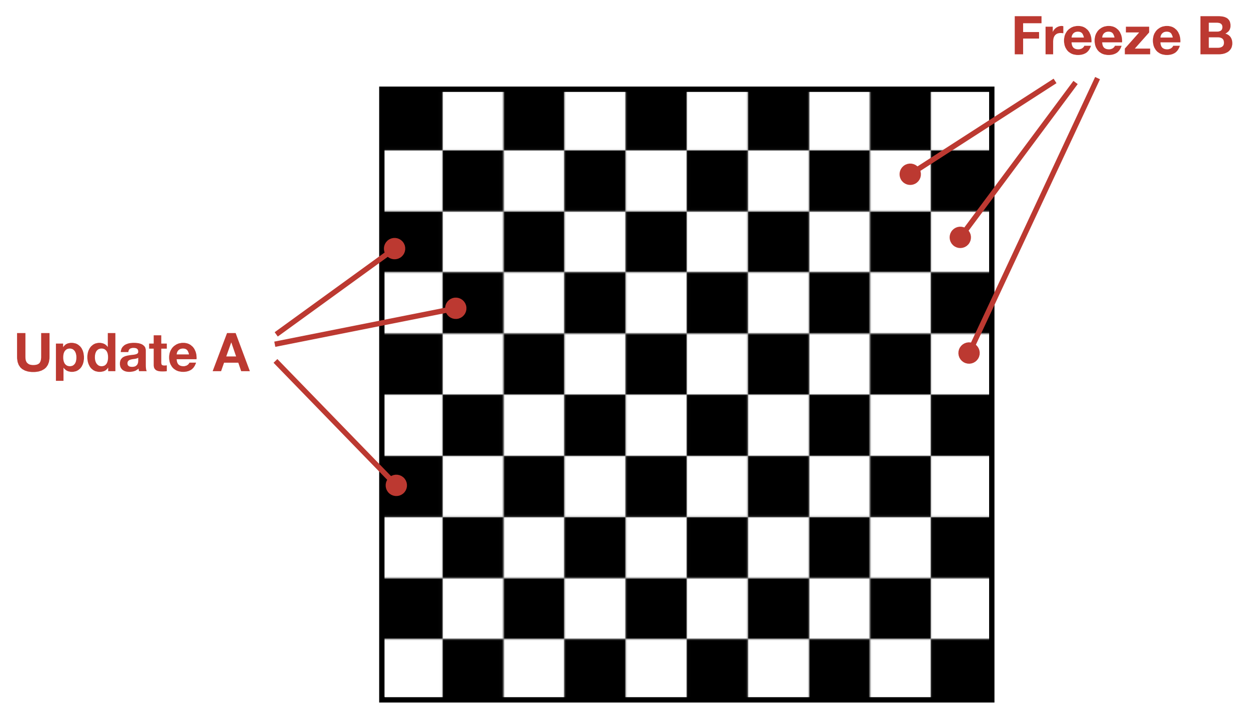
Coupling layers + Self-learning + Exactness

[Albergo, **GK**, Shanahan PRD100 (2019) 034515]

The early days

Scalar field $\phi(x) \in \mathbb{R}$, 1+1D spacetime

$$S[\phi] = \sum_x \partial_\mu \phi(x) \partial^\mu \phi(x) + \frac{M^2}{2}\phi(x)^2 + \lambda \phi(x)^4$$

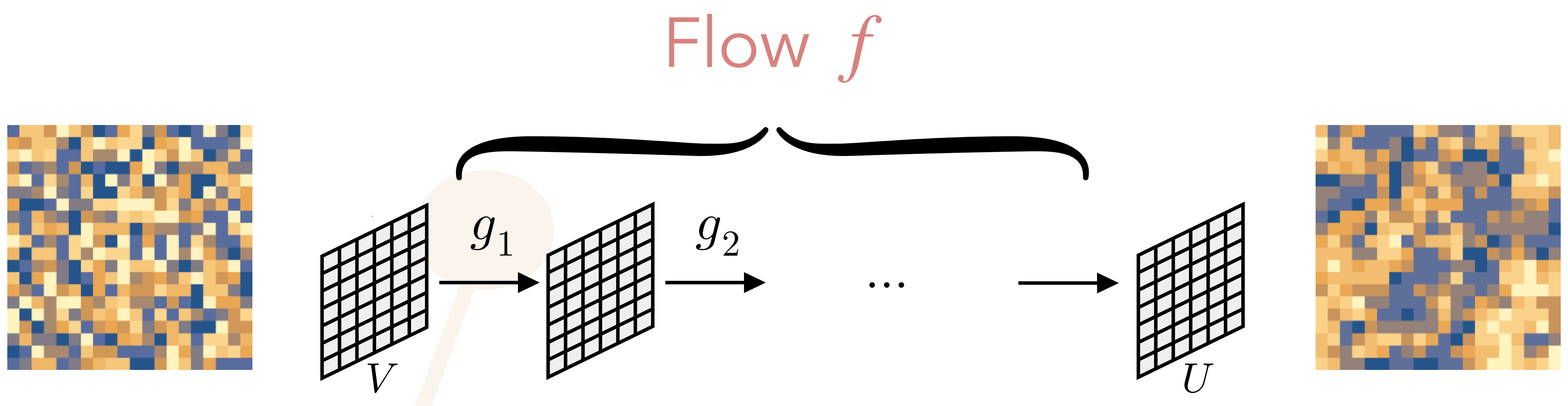


Checkerboard masking pattern m

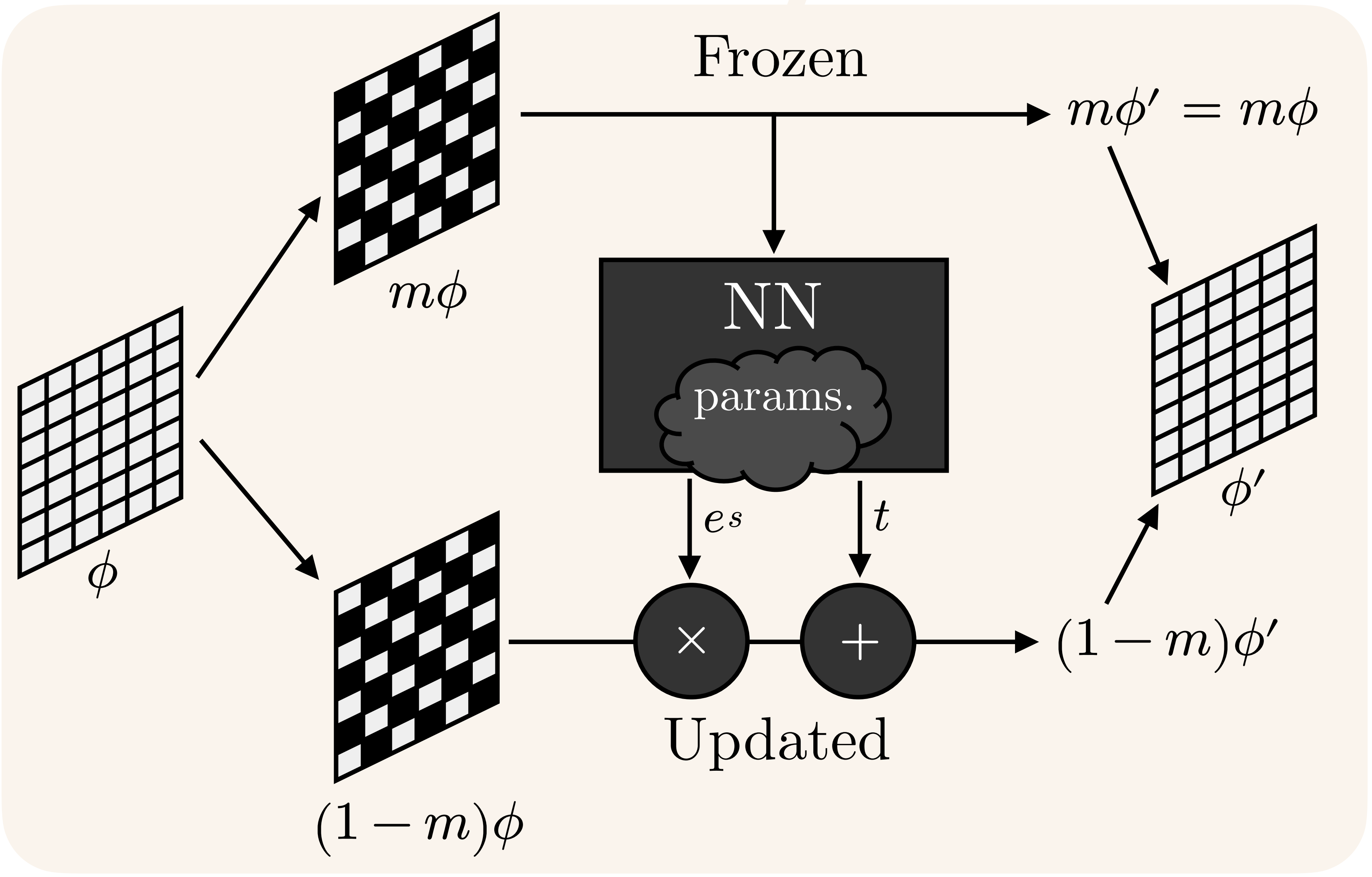
[Albergo, **GK**, Shanahan PRD100 (2019) 034515]

Machine learning jargon

Neural network (NN) = highly parameterized function approximator, usually a composition of linear + elementwise non-linear transformations



"Coupling layer" g_i



Tractable Jacobian

$$J_{ij} \equiv \partial \phi'_i / \partial \phi_j = \begin{bmatrix} I & \\ \blacksquare & \delta_{ij} e^{s_i} \end{bmatrix}$$

$$\implies \ln \det J = \sum_i s_i$$

The early days

Machine learning jargon

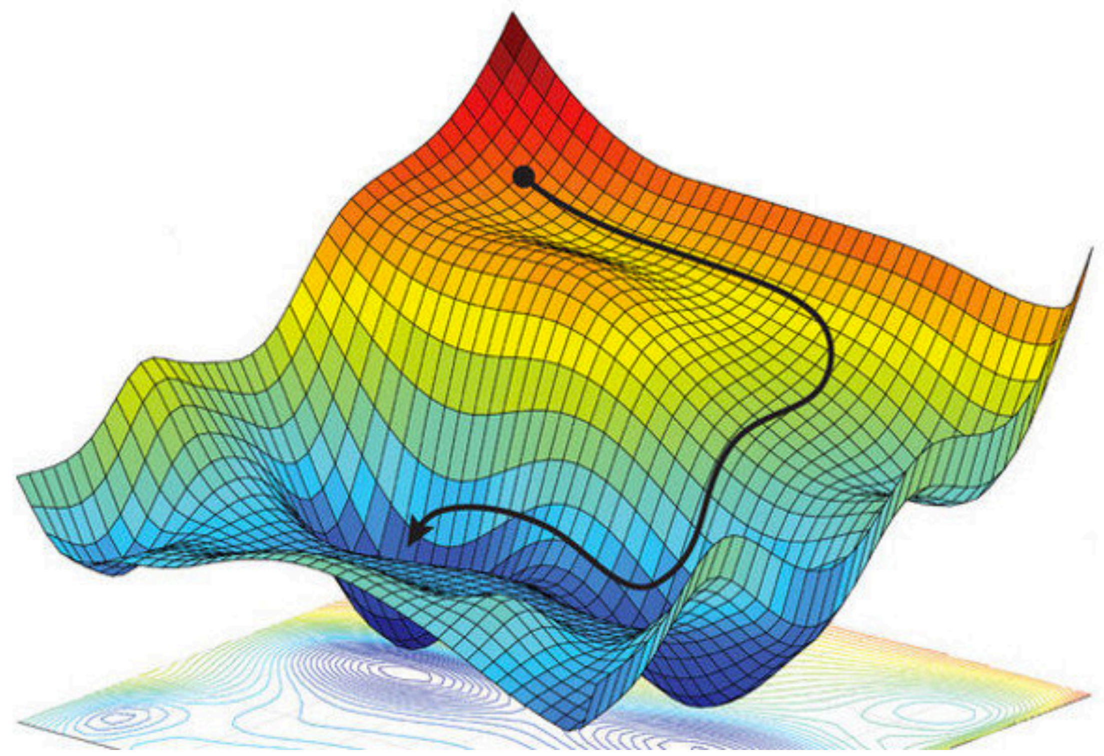
Training = optimization, typically by stochastic gradient descent
Loss function $\mathscr{L}$ = target function to be minimized

[Kullback & Leibler Ann. Math. Statist. 22 (1951) 79]

Self-training using Kullback-Leibler divergence between $p(U) = e^{-S[U]}/Z$ and $q(U)$

$$\mathscr{L} \equiv D'_{\mathrm{KL}}(q\,||\,p) = \int \mathscr{D}U\, q(U)\left[\log q(U) - \log e^{-S[U]}\right]$$
$$= \int \mathscr{D}U\, q(U)\left[\log q(U) + S(U)\right] \geq -\log Z$$

$$\vec{\omega}' = \vec{\omega} - \epsilon \nabla_{\vec{\omega}} \mathscr{L}$$



[Image credit: 1805.04829]

Exactness by reweighting or Metropolis

[Albergo, **GK**, Shanahan PRD100 (2019) 034515] [Nicoli+ PRE101 (2020) 023304]

$$p_{\mathrm{acc}}(U \to U') = \min\left(1, \frac{p(U')}{q(U')}\frac{q(U)}{p(U)}\right)$$

The early days

[Kullback & Leibler Ann. Math. Statist. 22 (1951) 79]

Self-training using Kullback-Leibler divergence between $p(U) = e^{-S[U]}/Z$ and $q(U)$

$$\mathscr{L} \equiv D'_{\mathrm{KL}}(q\,||\,p) = \int \mathscr{D}U\, q(U)\big[\log q(U) - \log e^{-S[U]}\big]$$
$$= \int \mathscr{D}U\, q(U)\big[\log q(U) + S(U)\big] \geq -\log Z$$

Exactness by reweighting or Metropolis

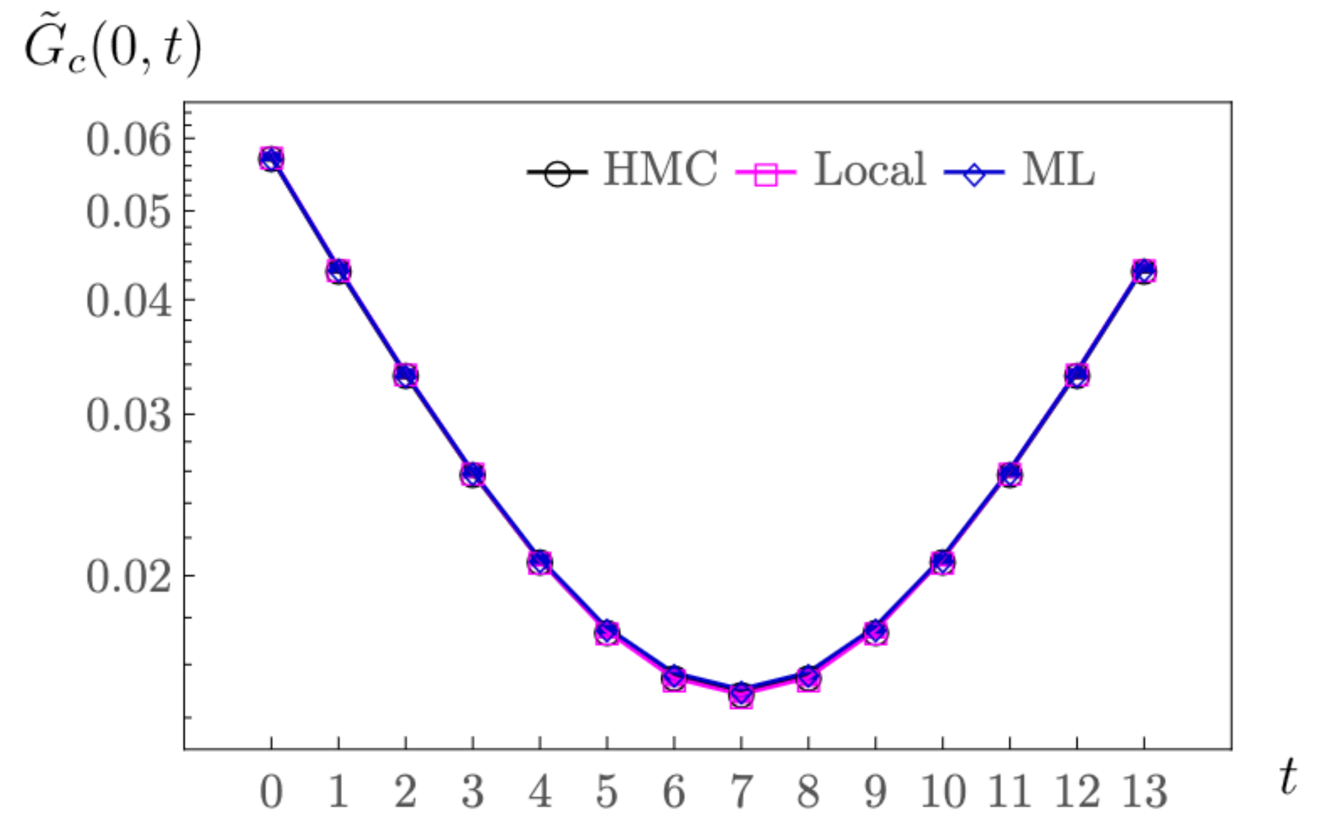
[Albergo, **GK**, Shanahan PRD100 (2019) 034515] [Nicoli+ PRE101 (2020) 023304]

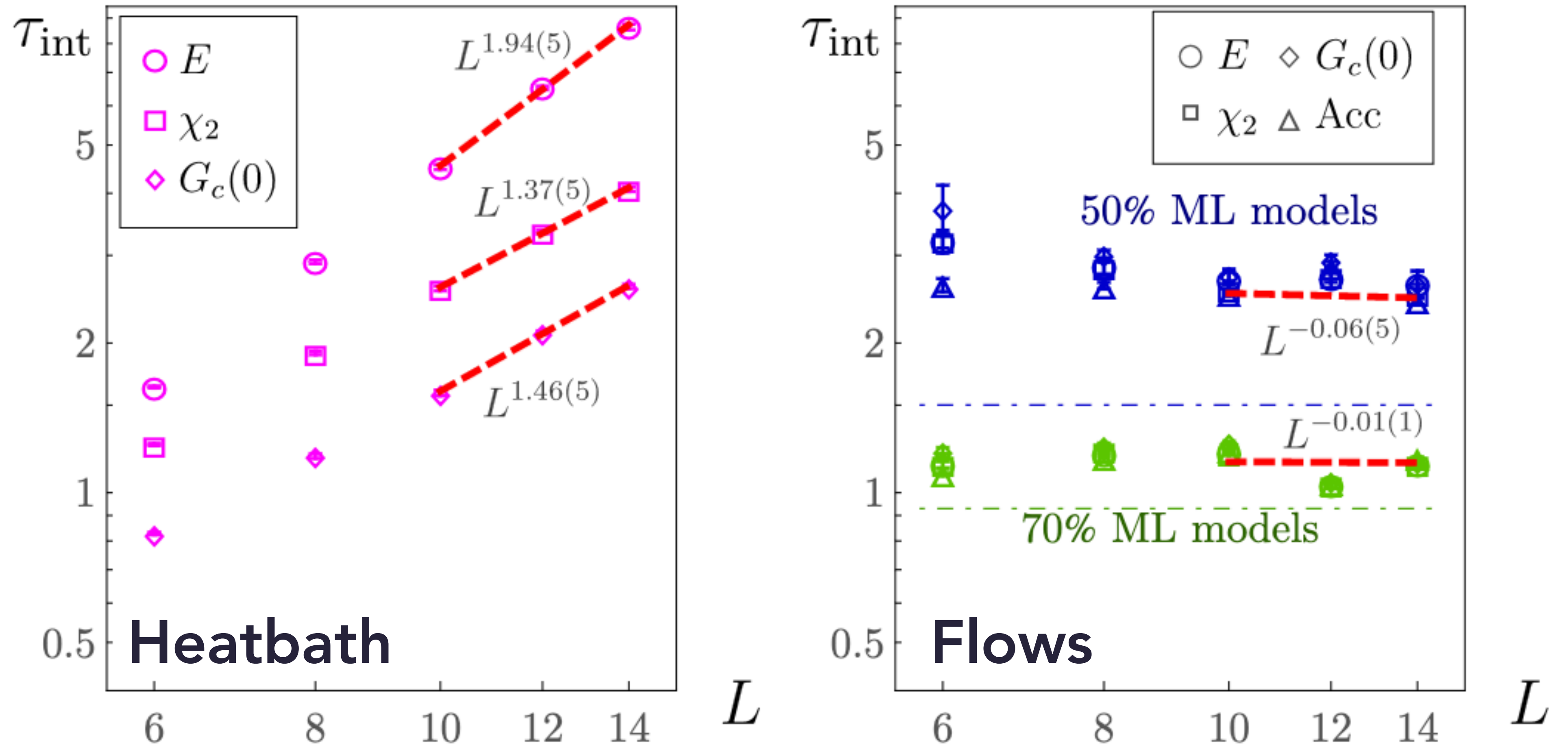
$$p_{\mathrm{acc}}(U \to U') = \min\left(1, \frac{p(U')}{q(U')}\frac{q(U)}{p(U)}\right)$$

Machine learning jargon

Training = optimization, typically by stochastic gradient descent

Loss function $\mathscr{L}$ = target function to be minimized





The early days

Machine learning jargon

...astic gradient descent
...o be minimized

Self-tra...
betwee...

$\mathscr{L} \equiv$ ...

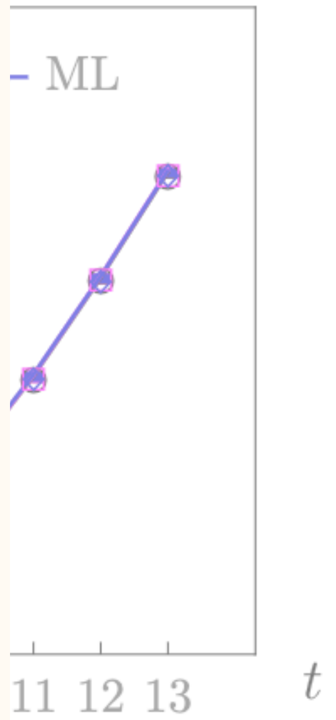
Exactn...

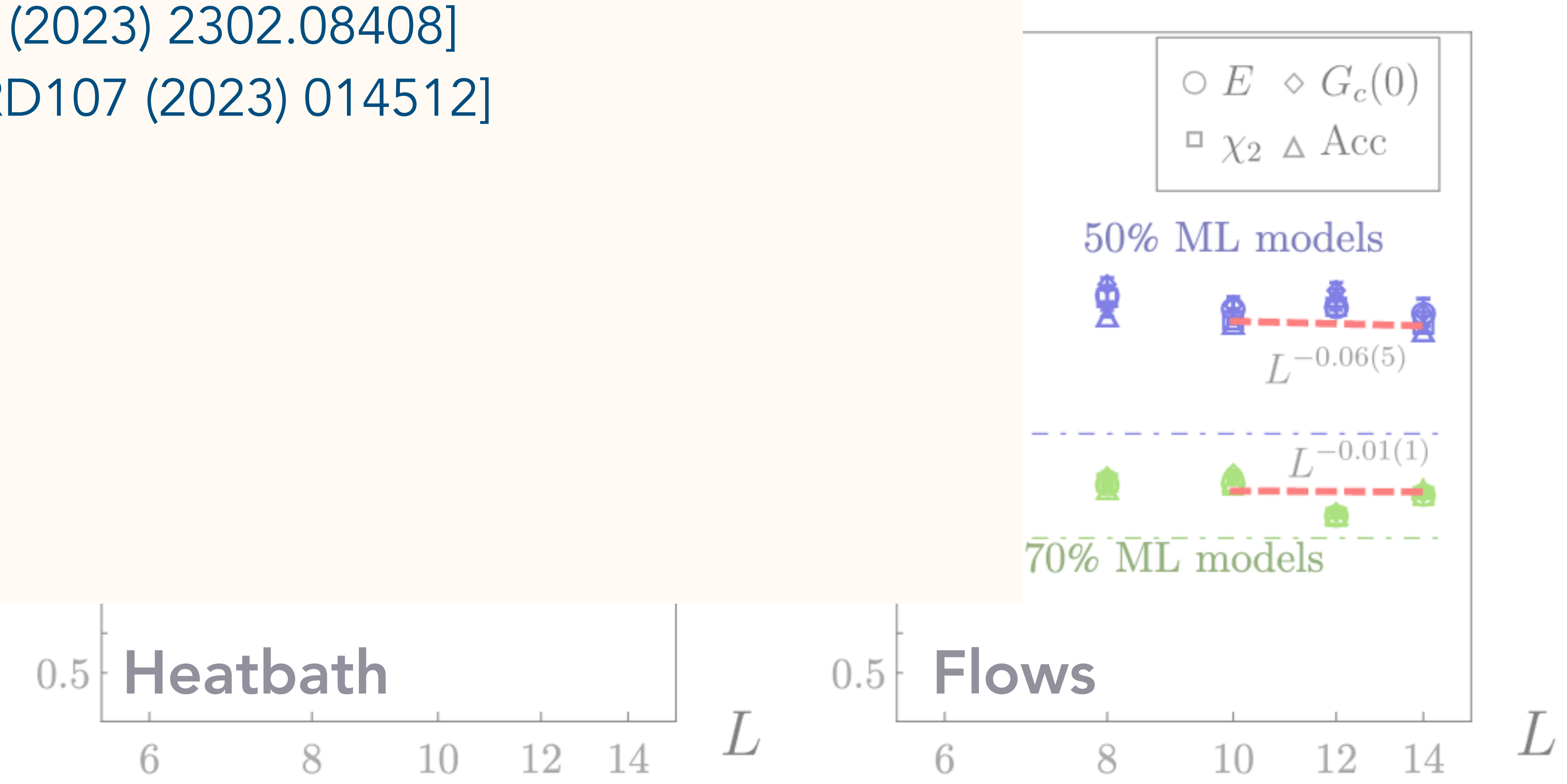
[Albergo, ...

Superseded by many recent scalar field theory results:

- Lattice size up to $L = 64$
- Smaller lattice spacings
- Broken phase

[Del Debbio+ JHEP07 (2021) 2105.12481]
[de Haan+ NeurIPS (2021) 2110.02673]
[Nicoli+ PRL126 (2021) 032001]
[Caselle+ JHEP07 (2022) 015]
[Komijani & Marinkovic PoSLATTICE (2022) 019]
[Gerdes+ (2022) 2207.00283]
[Albandea+ (2023) 2302.08408]
[Singha+ PRD107 (2023) 014512]

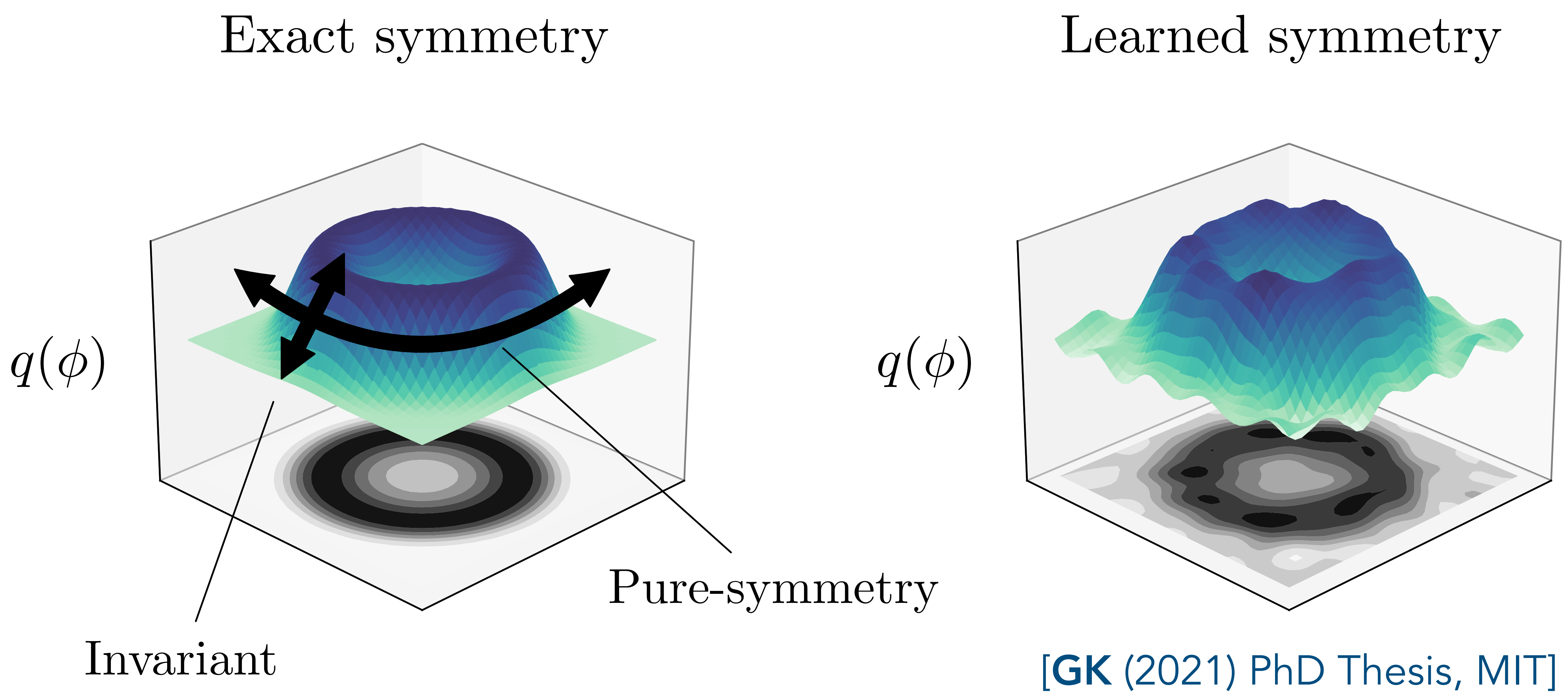




Incorporating symmetries

Symmetries…

- ✓ Reduce data complexity of training
- ✓ Reduce model parameter count
- ✓ May make "loss landscape" easier



[**GK** (2021) PhD Thesis, MIT]

Invariant prior + **equivariant** flow = symmetric model

$$r(t \cdot U) = r(U) \qquad f(t \cdot U) = t \cdot f(U)$$

[Cohen, Welling PMLR48 (2016) 2990]

Gauge symmetry

Distribution should be symmetric under $(\Omega \cdot U)_\mu(x) = \Omega(x) U_\mu(x) \Omega^\dagger(x + \hat{\mu})$.

Gauge-invariant prior:

Uniform (Haar) distribution
$r(U) = 1$ works.

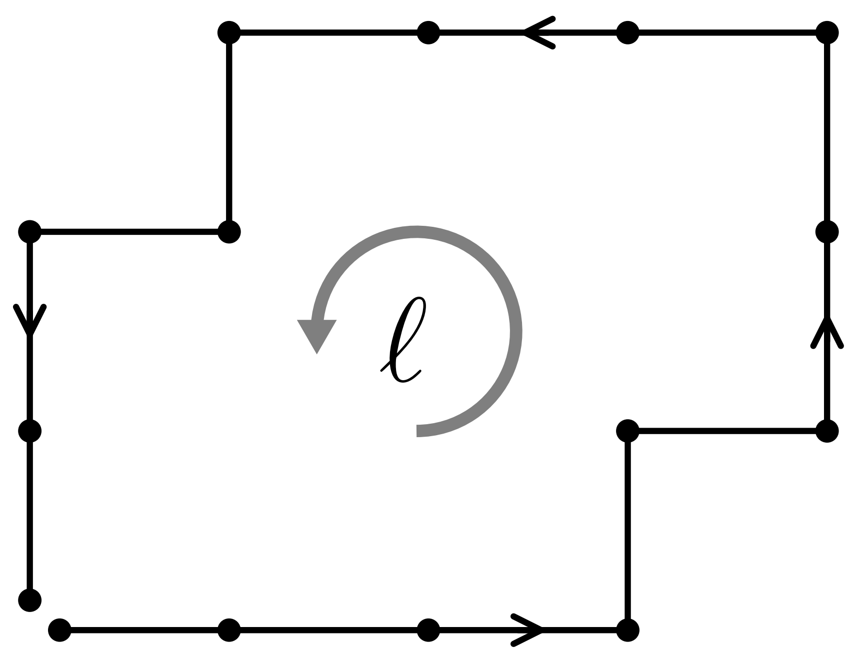
Gauge-equivariant flow:

Coupling layers act on
(untraced) Wilson loops.

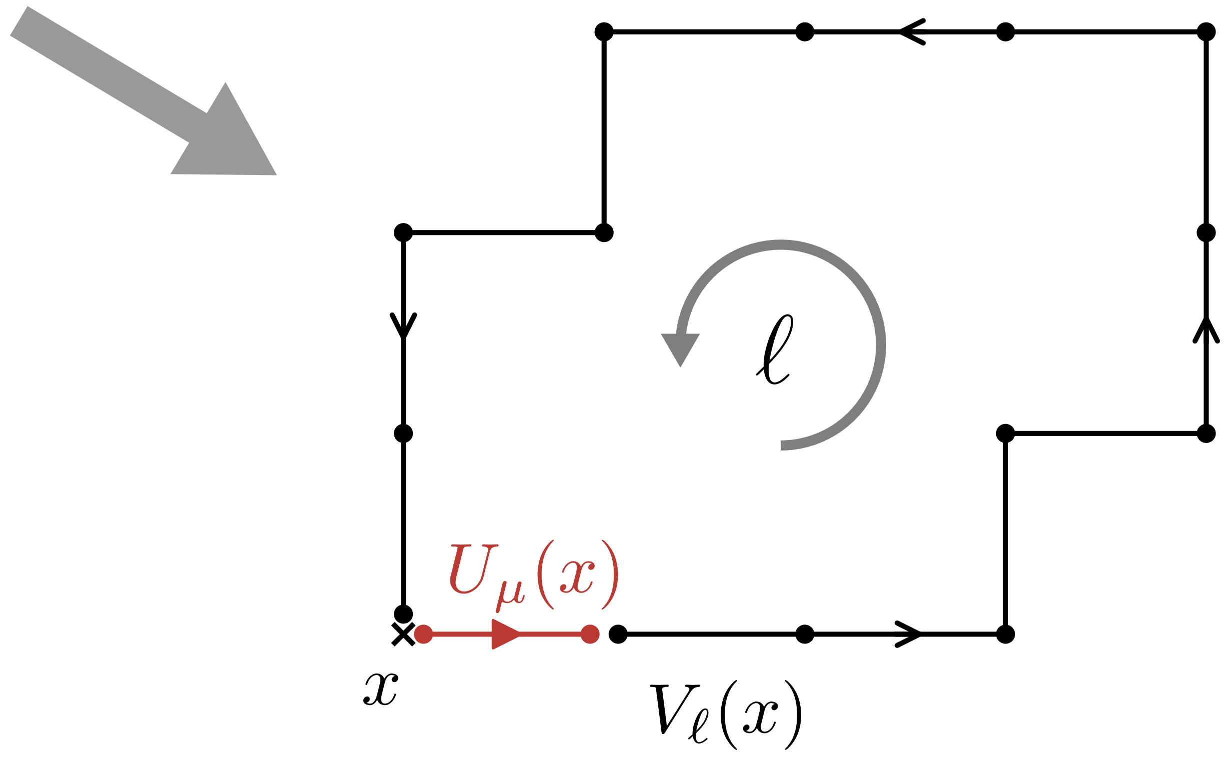
Loop transformation easier to satisfy.

Open loop

[**GK**, Albergo, … PRL125 (2020) 121601]



$$W_\ell(x) \xrightarrow{\text{Flow}} W'_\ell(x)$$



$$U'_\mu(x) = W'_\ell(x)\, V^\dagger_\ell(x)$$

Gauge symmetry

Distribution should be symmetric under $(\Omega \cdot U)_\mu(x) = \Omega(x) U_\mu(x) \Omega^\dagger(x+\hat{\mu})$.

Gauge-invariant prior:

Uniform (Haar) distribution
$r(U) = 1$ works.

Gauge-equivariant flow:

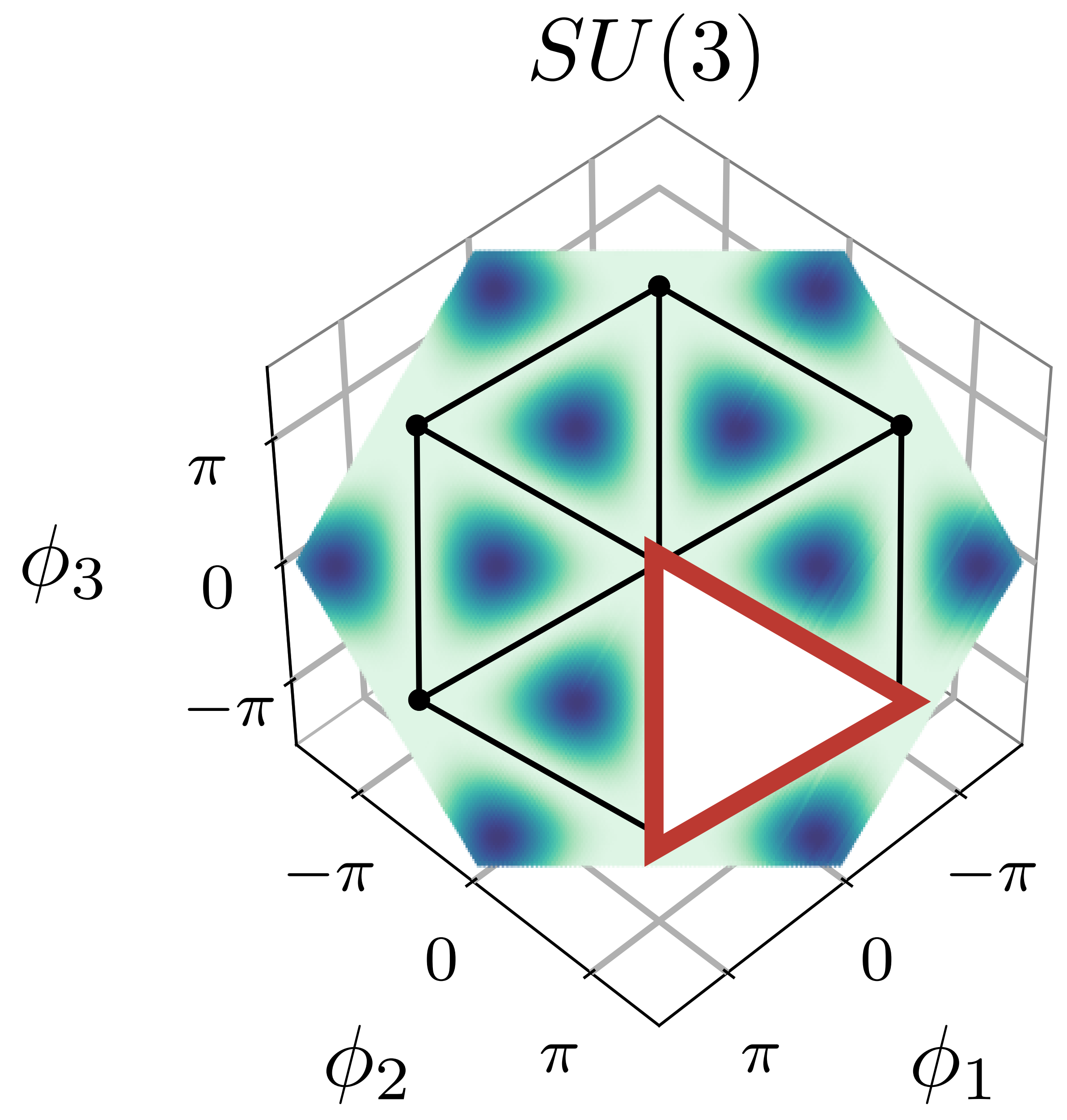
Coupling layers act on
(untraced) Wilson loops.

Loop transformation easier to satisfy.

Custom flows designed
for $U(1)$ and $SU(N)$ gauge
manifolds

[**GK**, Albergo, ... PRL125 (2020) 121601]

[Boyda, **GK**, ... PRD103 (2021) 074504]
[Rezende, ..., **GK**, ... PMLR119 (2020) 8083]



Gauge symmetry

Distribution should be symmetric under $(\Omega \cdot U)_\mu(x) = \Omega(x)U_\mu(x)\Omega^\dagger(x+\hat{\mu})$.

Several non-flow gauge-equivariant models and applications have emerged:

[Nagai & Tomiya (2021) 2103.11965]
[Favoni+ PRL128 (2022) 032003]
[Lehner & Wettig (2023) 2302.05419]
[Aronsson+ (2023) 2303.11448]
[Lehner & Wettig (2023) 2304.10438]

[A. Tomiya Mon 1:50]
[U. Wenger Mon 3:10]
[T. Wettig Mon 5:00]
[X.-Y. Jin Thur 2:30]

Gauge-i

Uniform (H

$r(U)$

K, Albergo, … PRL125 (2020) 121601]

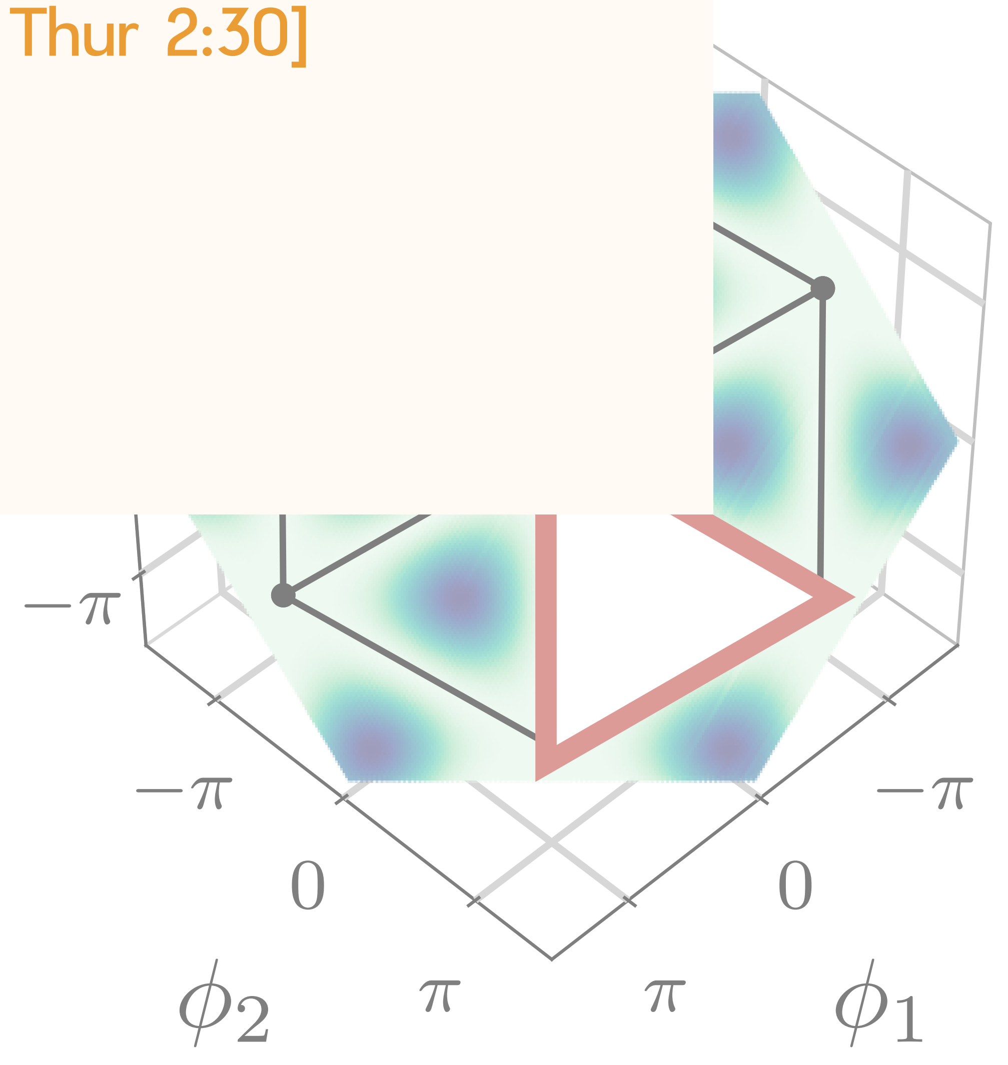
yda, **GK**, … PRD103 (2021) 074504]

zende, …, **GK**, … PMLR119 (2020) 8083]

Gauge-e

Coupling layers act on (untraced) Wilson loops.

Loop transformation easier to satisfy.



Promising early results

U(1) gauge theory in 1+1D

$$S(U) = -\beta \sum_x \sum_{\mu<\nu} \mathrm{Re}\, P_{\mu\nu}(x)$$

$$P_{\mu\nu}(x) = U_\mu(x) U_\nu(x+\hat{\mu}) U_\mu^\dagger(x+\hat{\nu}) U_\nu^\dagger(x)$$

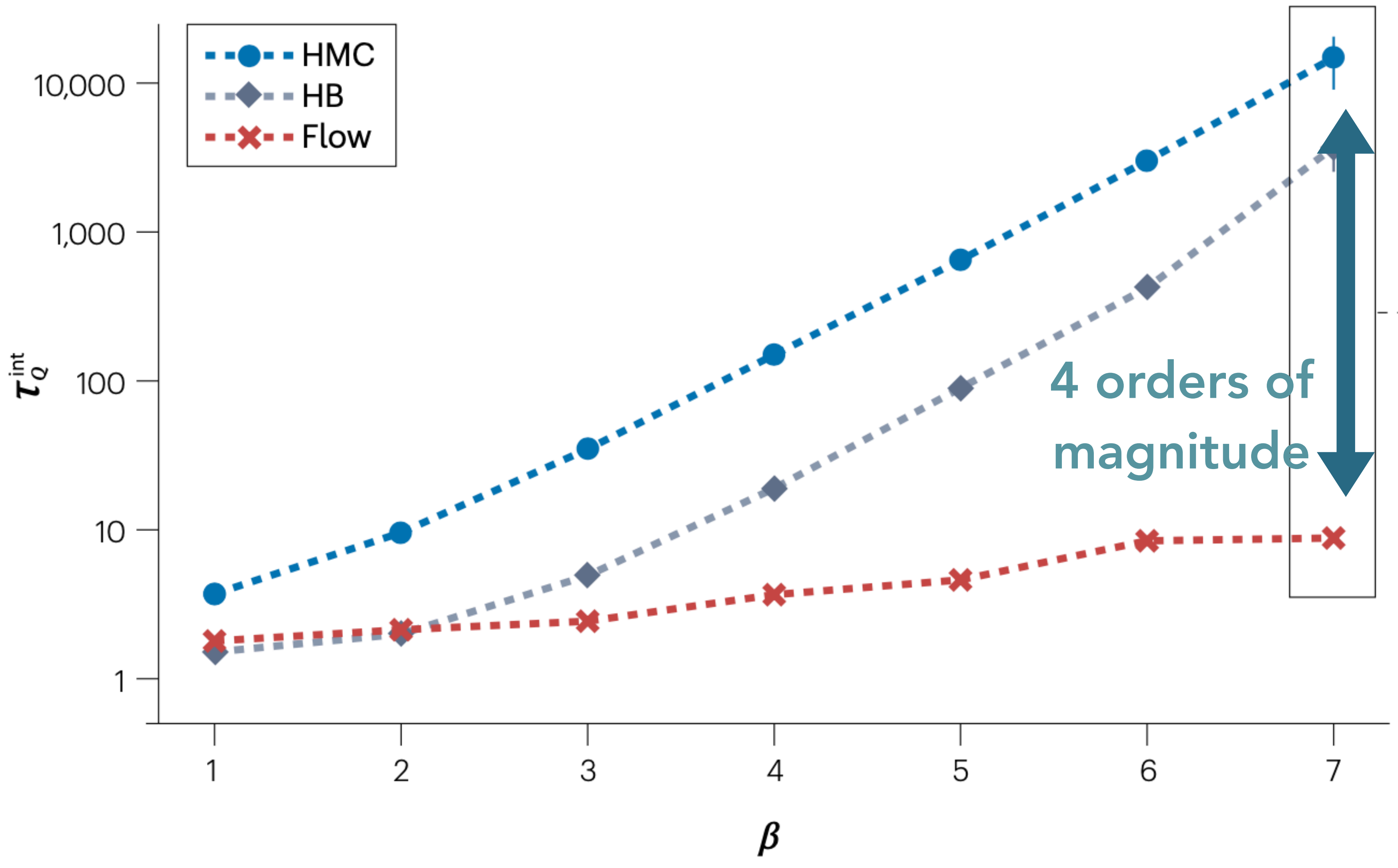
Topological observables

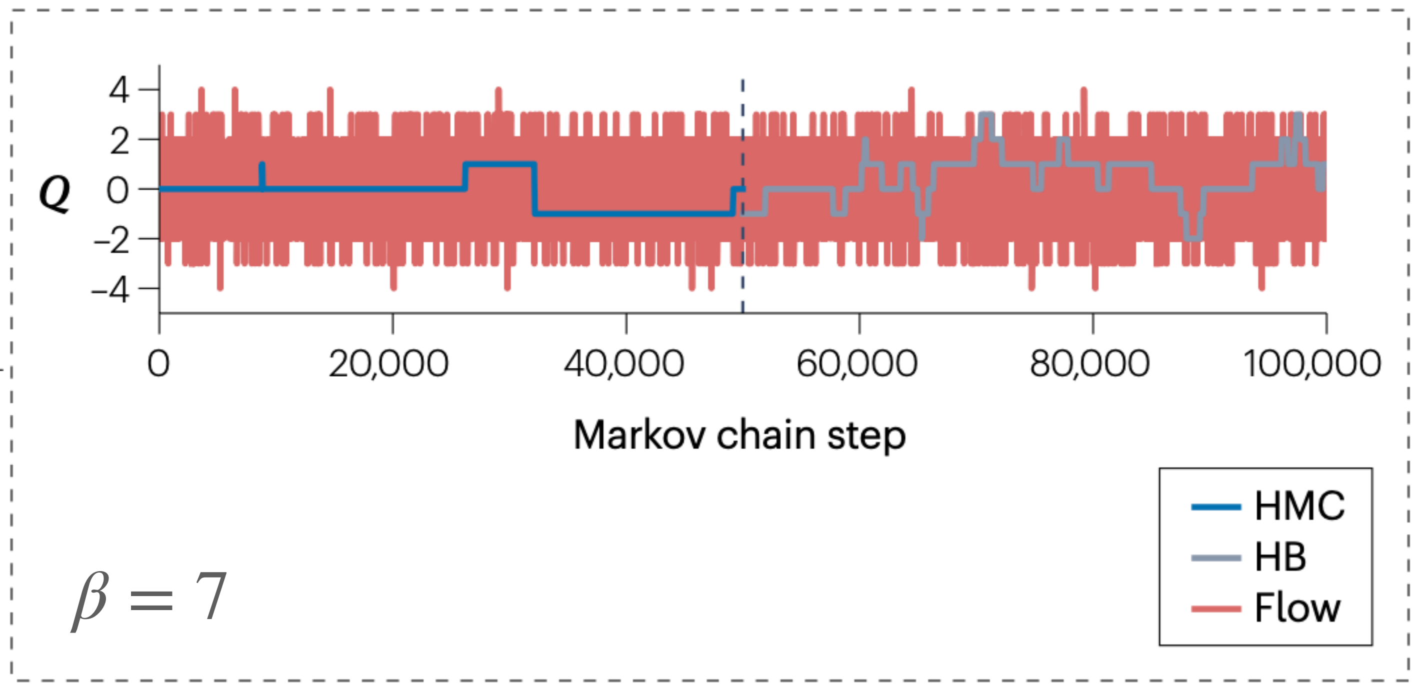
$$Q = \frac{1}{2\pi} \sum_x \arg(P_{01}(x))$$

$$\chi_Q = \langle Q^2 \rangle$$

Non-topological observables

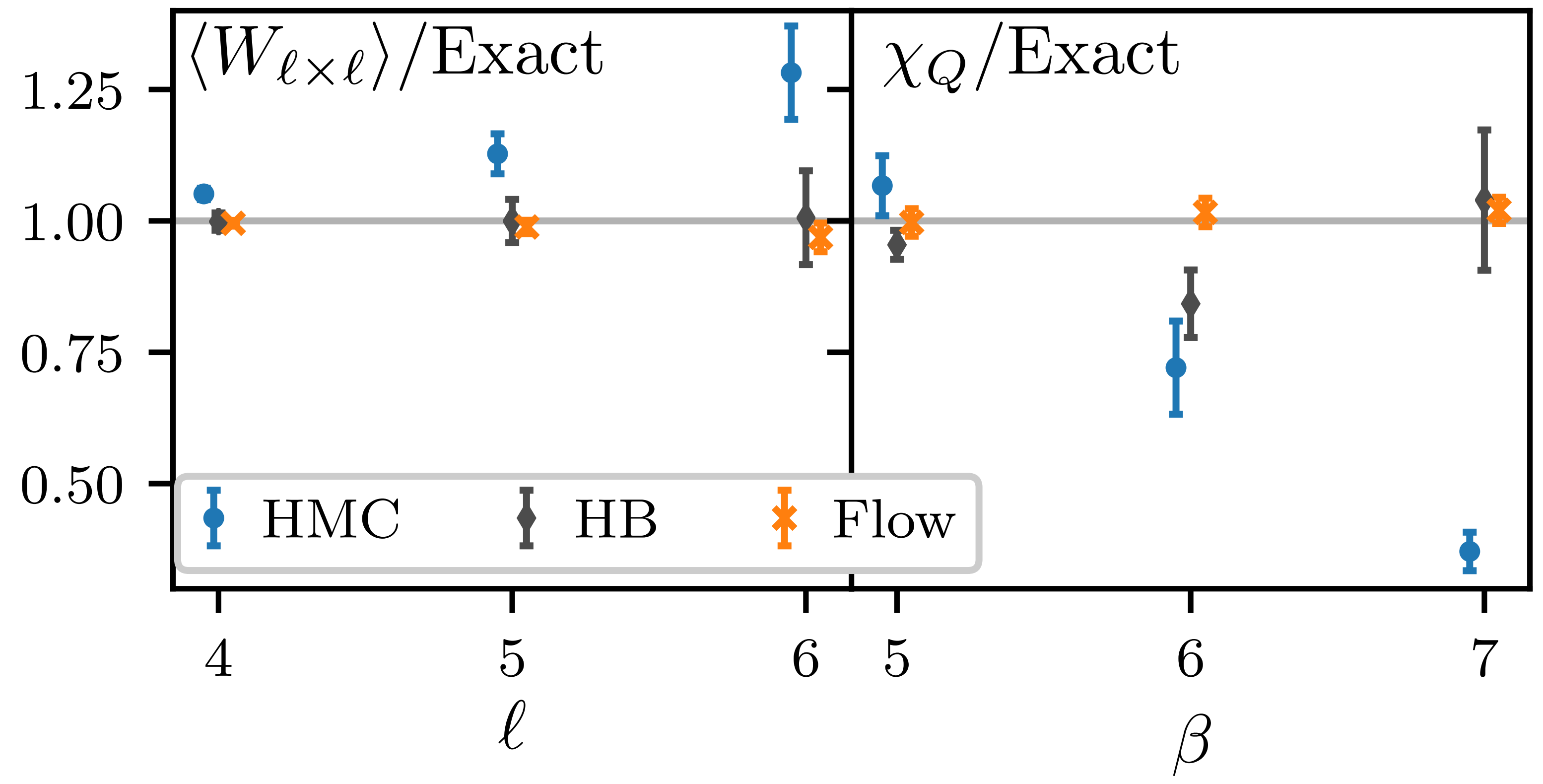
$W_{\ell\times\ell} = \ell \times \ell$ Wilson loop





[Cranmer, **GK**, Racanière, Rezende, Shanahan Nat. Rev. Phys. (2023) in production]

[**GK**, Albergo, … PRL125 (2020) 121601]



Where we stand

Beyond critical slowing down

New paradigms

- Partition functions (e.g. for thermodynamics) [C. Kirwan Poster]
- Parameter dependence [Gerdes+ (2022) 2207.00283] [Singha+ (2022) 2207.00980]
- Correlated samples (e.g. for Feynman-Hellmann)
- Faster parallel tempering

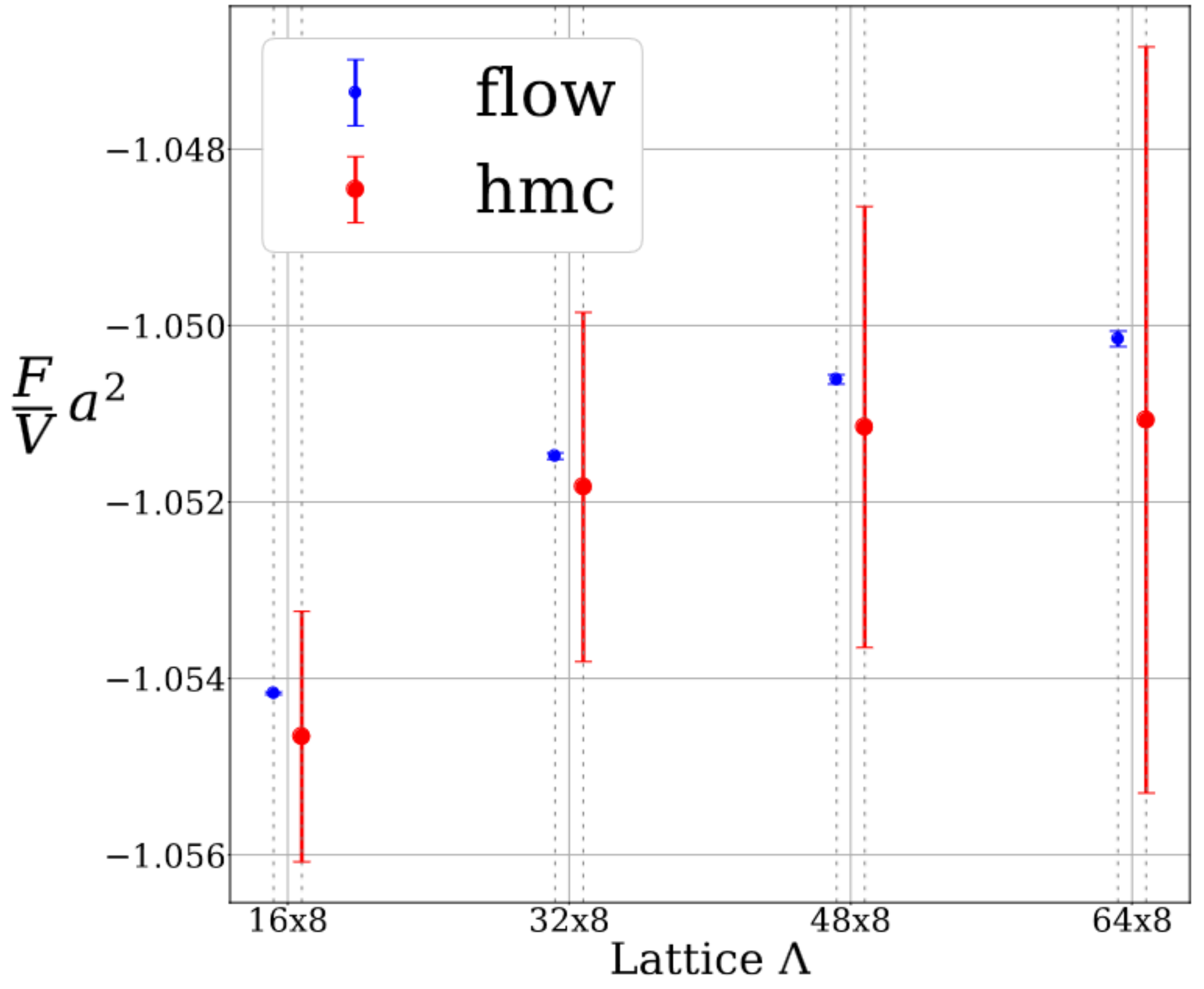
[D. Hackett Thur 3:10]

- Sign problems [Lawrence+ PRD103 (2021) 114509] [Rodekamp+ PRB106 (2022) 125139] [Pawlowski & Urban (2022) 2203.01243] [M. Rodekamp Thur 4:20] [Y. Lin Thur 4:40]

Practical gains

- Embarrassingly parallel sampling
- Storage-free ensembles

[Nicoli+ PRE101 (2020) 023304]
[Nicoli+ PRL126 (2021) 032001]



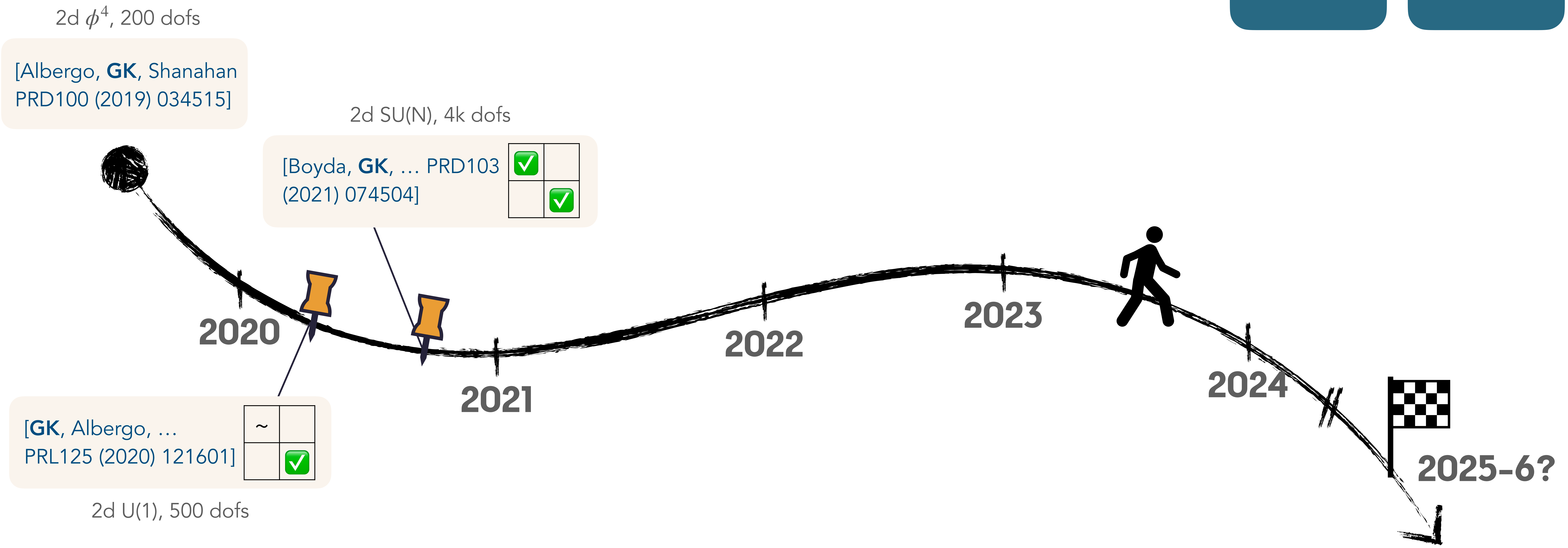
With $U_i \sim q(U)$,

$$\hat{Z} = \frac{1}{N}\sum_{i=1}^{N} e^{-S[U_i]}/q(U_i)$$

and $\hat{F} = -\log \hat{Z}$

Towards lattice QCD

SU(N) gauge symmetry	Dynamical fermions
3+1D	Large β / small a



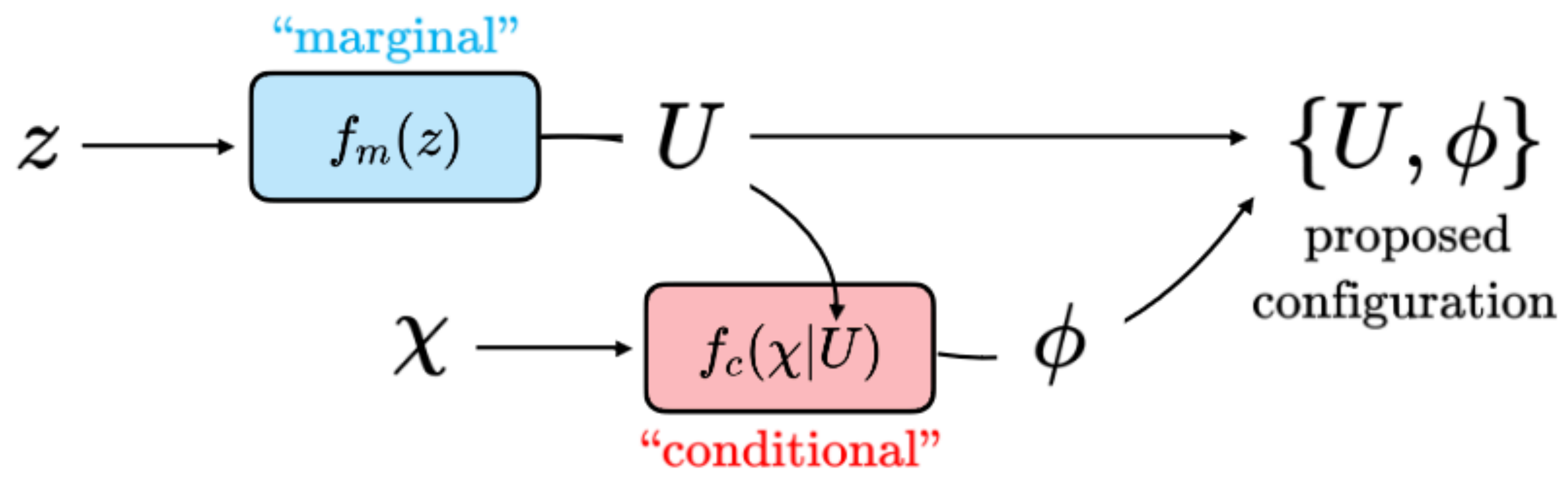
[Albergo, ..., **GK**, ... PRD104 (2021) 114507]
[Albergo, ..., **GK**, ... PRD106 (2022) 014514]
[Abbott, ..., **GK**, ... PRD106 (2022) 074506]

Towards lattice QCD: fermions

Pseudofermions highly effective in HMC, logical to use for flows also.

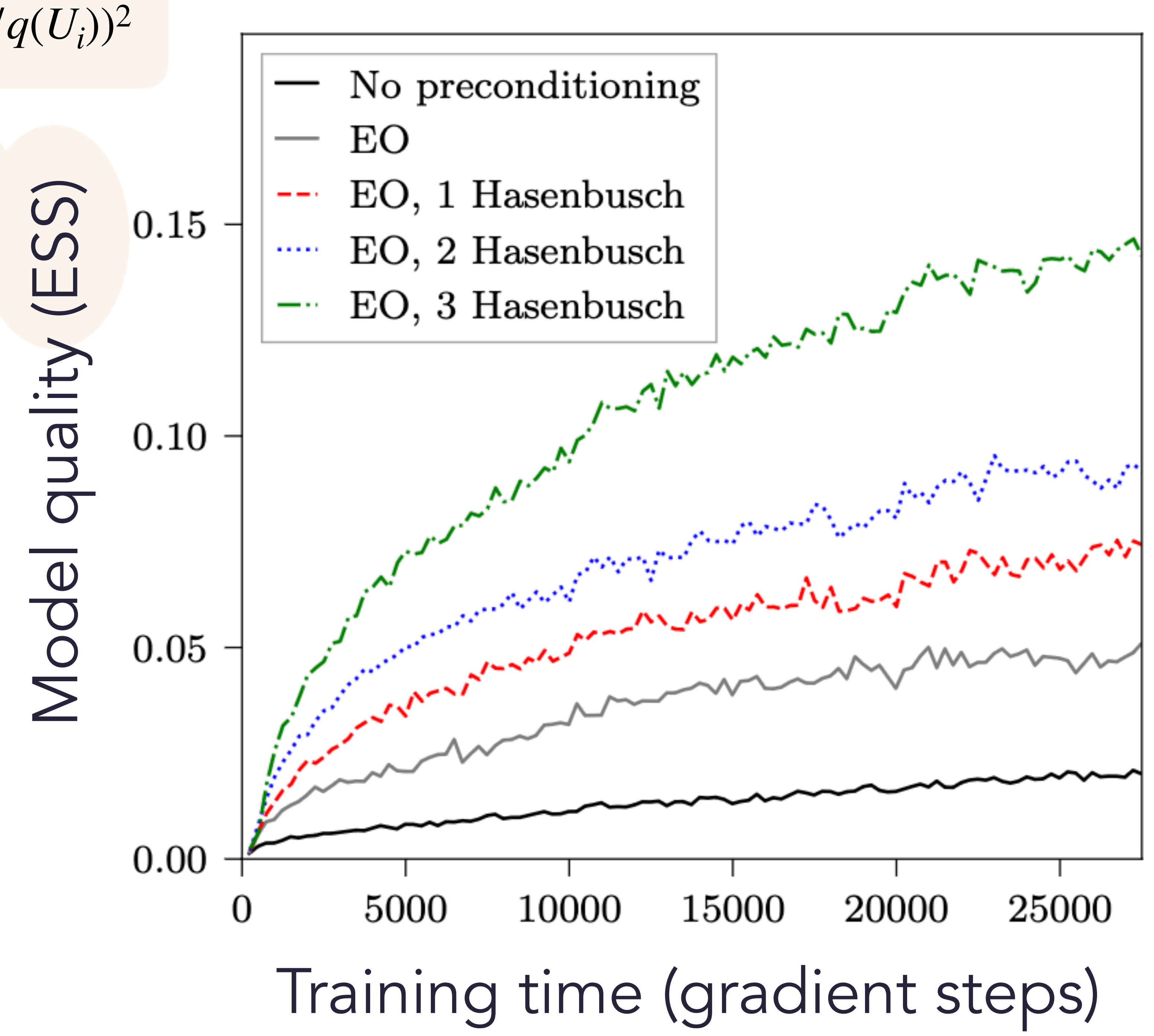
Separate coupling layers for gauge field and PFs can be composed arbitrarily

- **Simplest case:** marginal + conditional model



- **Preconditioning** works equally well for flows
- Modified Metropolis allows averaging away noise in the conditional flow

$$\mathrm{ESS} = \frac{(\frac{1}{N}\sum_i p(U_i)/q(U_i))^2}{\frac{1}{N}\sum_i (p(U_i)/q(U_i))^2}$$



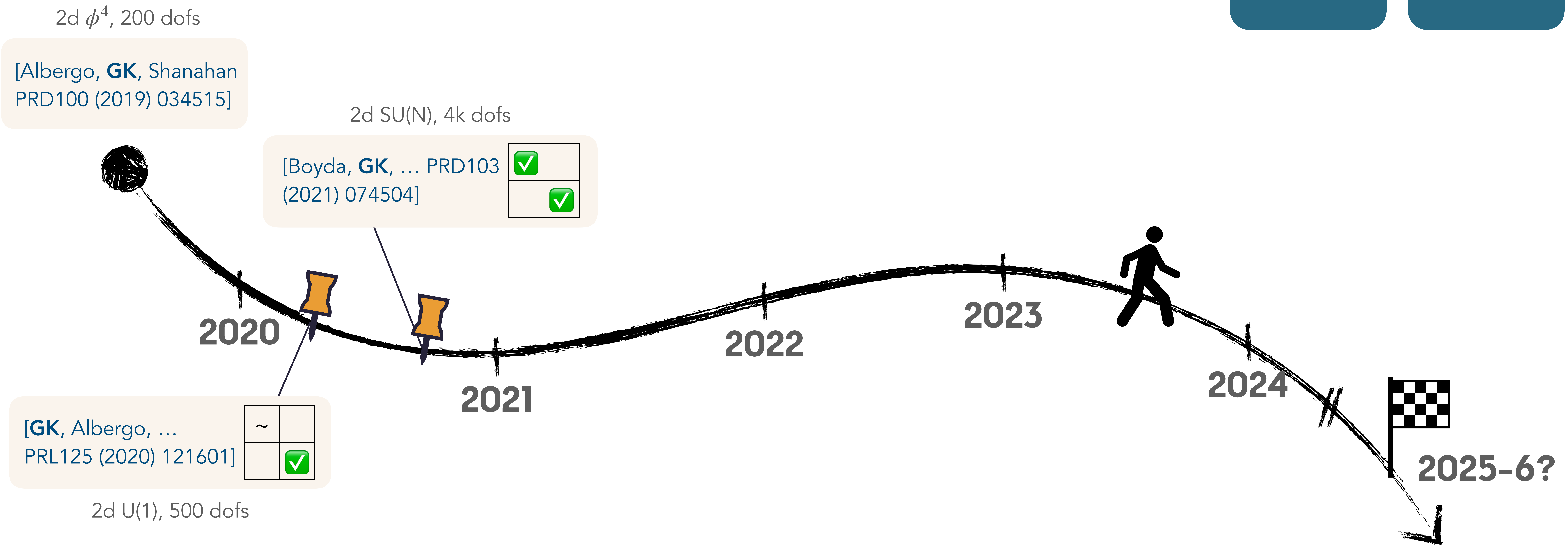
Towards lattice QCD

SU(N) gauge symmetry

Dynamical fermions

3+1D

Large β / small a



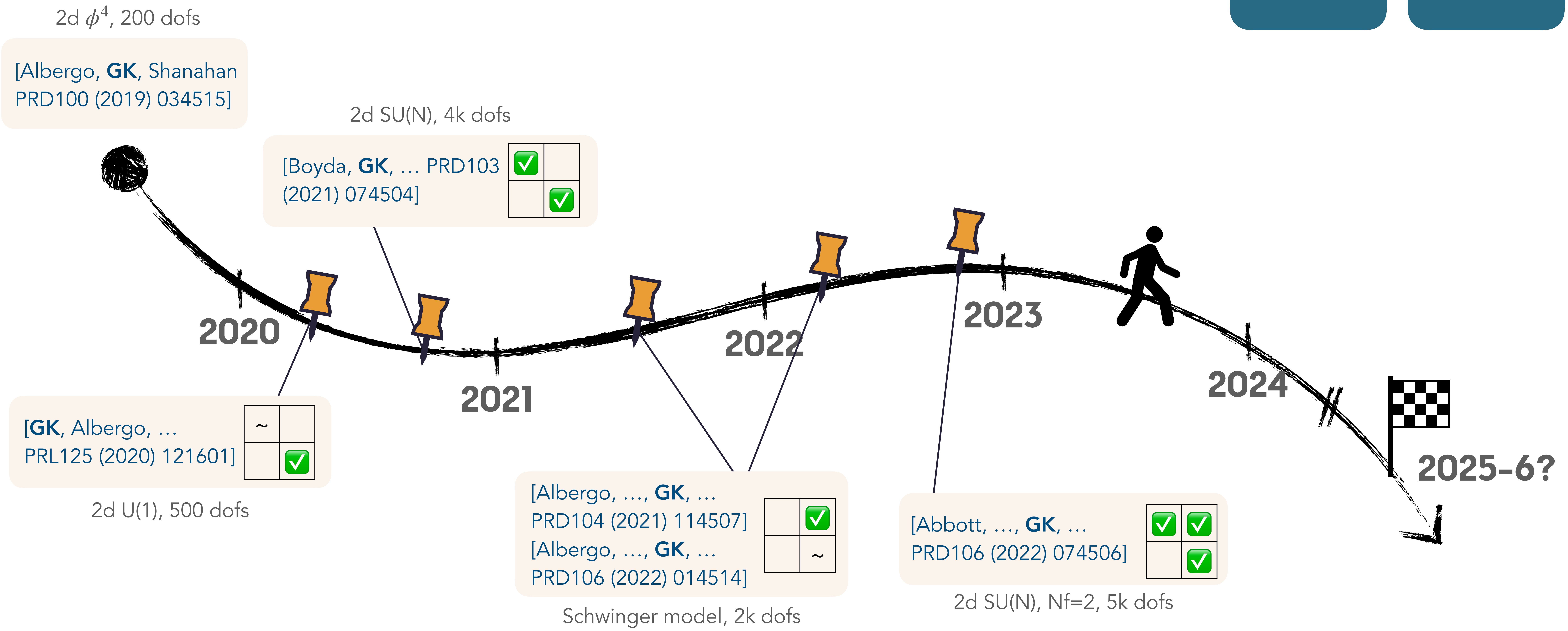
Towards lattice QCD

SU(N) gauge symmetry

Dynamical fermions

3+1D

Large β / small a



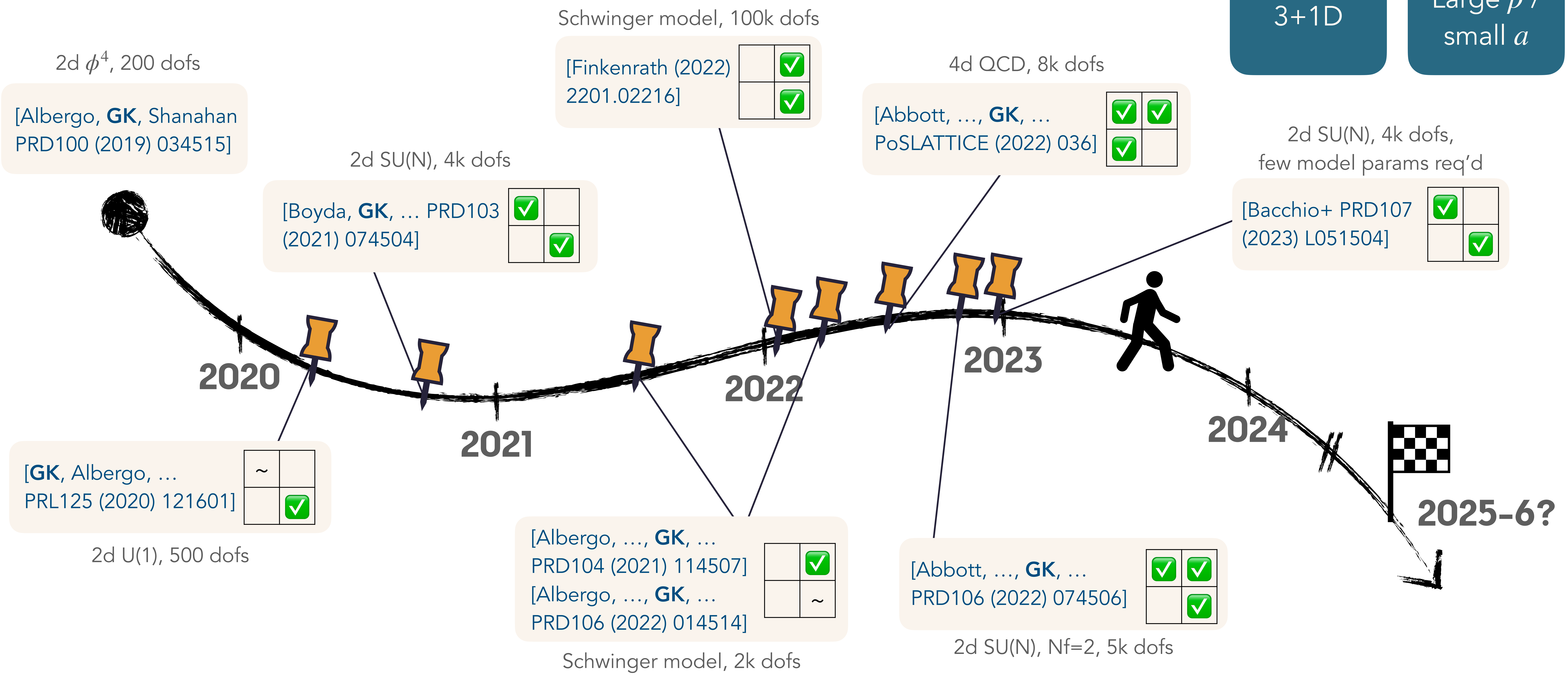
Towards lattice QCD

SU(N) gauge symmetry

Dynamical fermions

3+1D

Large β / small a



Frontiers of development

Frontiers of development

Opinions

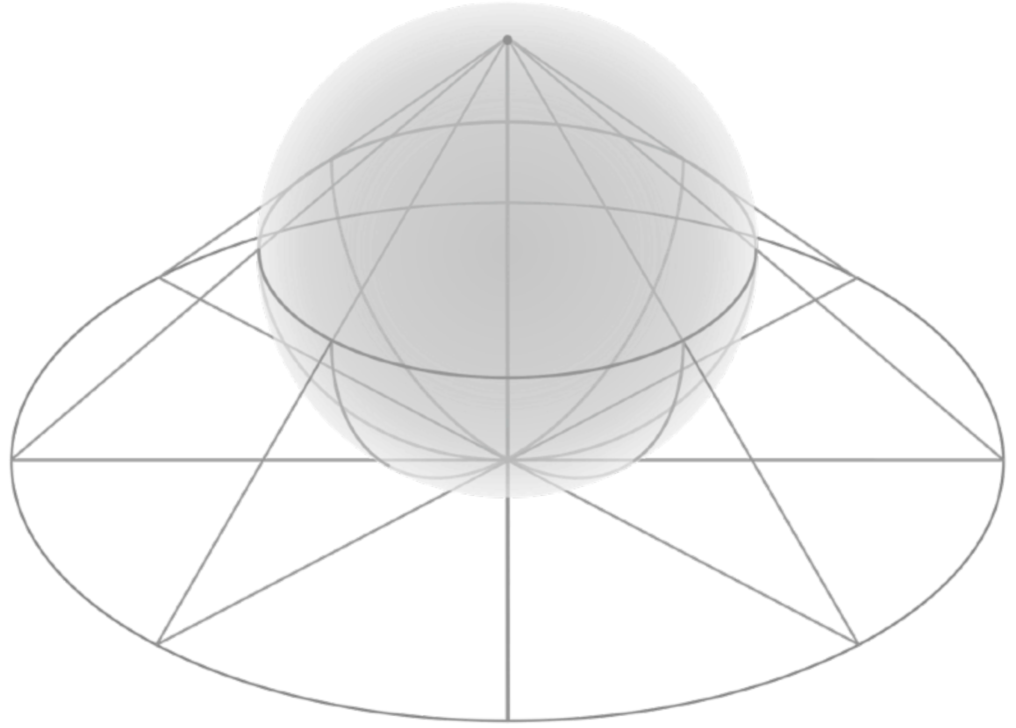
An interesting detour on the way to QCD

CP^{N-1} theory formulated on a 2D lattice:

Fields: $z(x) = (z_1(x), \ldots, z_N(x)),\ |z(x)|^2 = 1$

Lattice action: $S[z] = -\beta \sum_{x,\mu} |z^\dagger(x) \cdot z(x+\hat{\mu})|^2$

Riemann sphere $\cong CP^1$



Shares many properties with $SU(N)$ YM…

- Asymptotic freedom
- Confinement
- Topology, instantons

[d'Adda, Lüscher, Vecchia NPB146 (1978) 63]

[Witten NPB149 (1979) 285]

$$q = \frac{1}{2\pi} \epsilon_{\mu\nu} \partial_\mu A_\nu = \frac{i}{2\pi} \epsilon_{\mu\nu} \overline{D_\mu z} \cdot D_\nu z \,.$$

$$z_\alpha(x) = \frac{\lambda u_\alpha + [(x_1 - a_1) - i(x_2 - a_2)]\, v_\alpha}{(\lambda^2 + (x-a)^2)^{1/2}} .$$

No efficient sampling algorithm known for $N > 2$. We should solve this problem with flows.

[A. Nada Mon 2:50]
[J. Marsh-Rossney Poster]

Continuous and discrete flows

Continuous flows

- **Jacobian:** integrate an ODE ⚠️
- **Parameterization:** scalar function $\varphi(U)$
- **Symmetries:** make $\varphi(U)$ **invariant**

$$f(V) = \int_0^T dt\, \nabla\varphi(U(t);t)\Big|_{U(0)=V} + V$$

Discrete flows

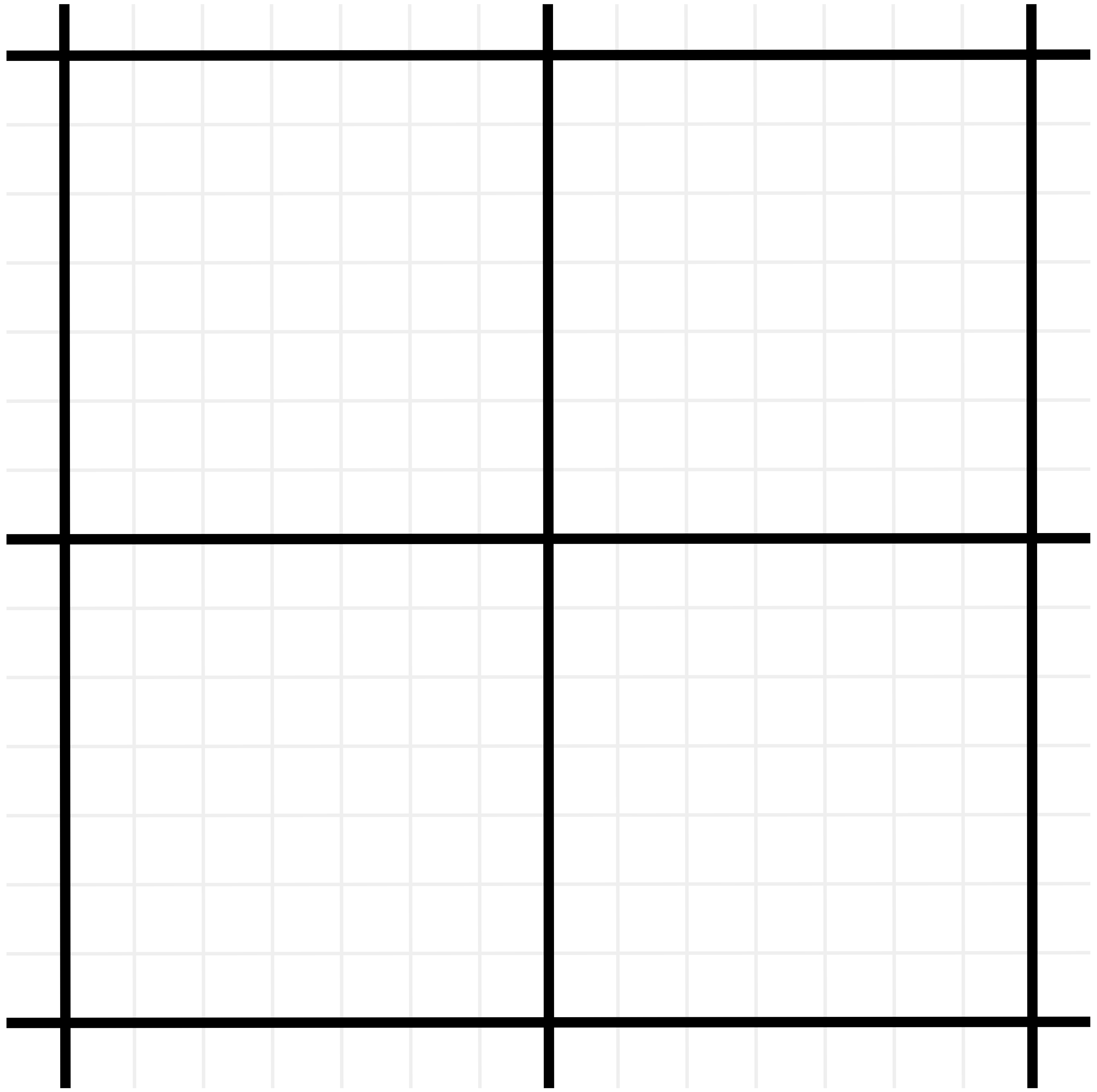
- **Jacobian:** exact upper triangular
- **Parameterization:** hand-crafted diffeomorphisms ⚠️
- **Symmetries:** make transformations **equivariant**

$$f = g_1 \circ \ldots \circ g_n$$

Advantages are not in conflict. We should investigate options taking the best of both.

Masked residual layers are promising. [Abbott, ..., **GK**, ... 2305.02402]

A possible future 🏁



Storage-free inner configurations

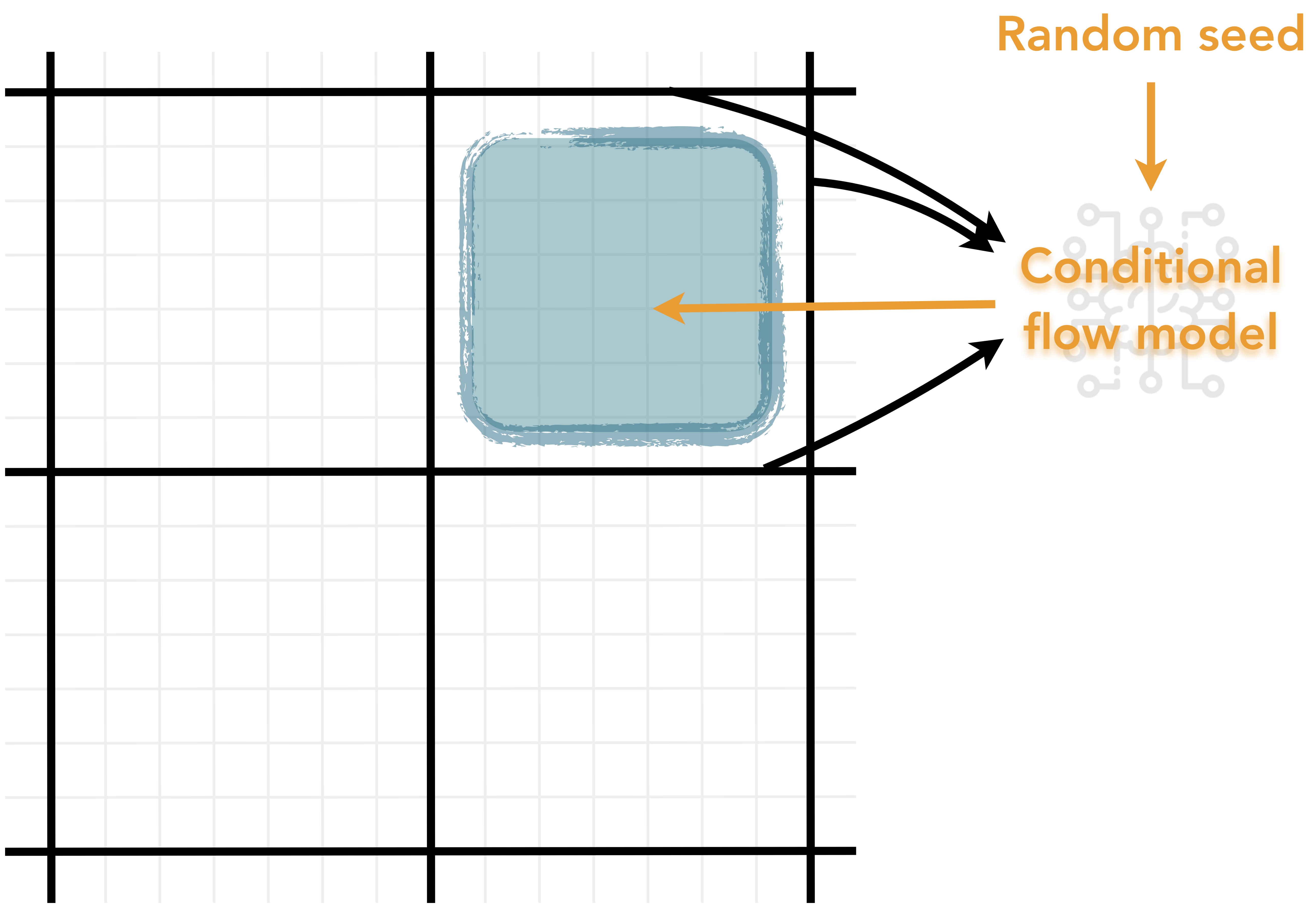
+

Partition function estimation
for efficient outer sampling

+

Exponential error reduction
by multi-level estimates

A possible future



Storage-free inner configurations

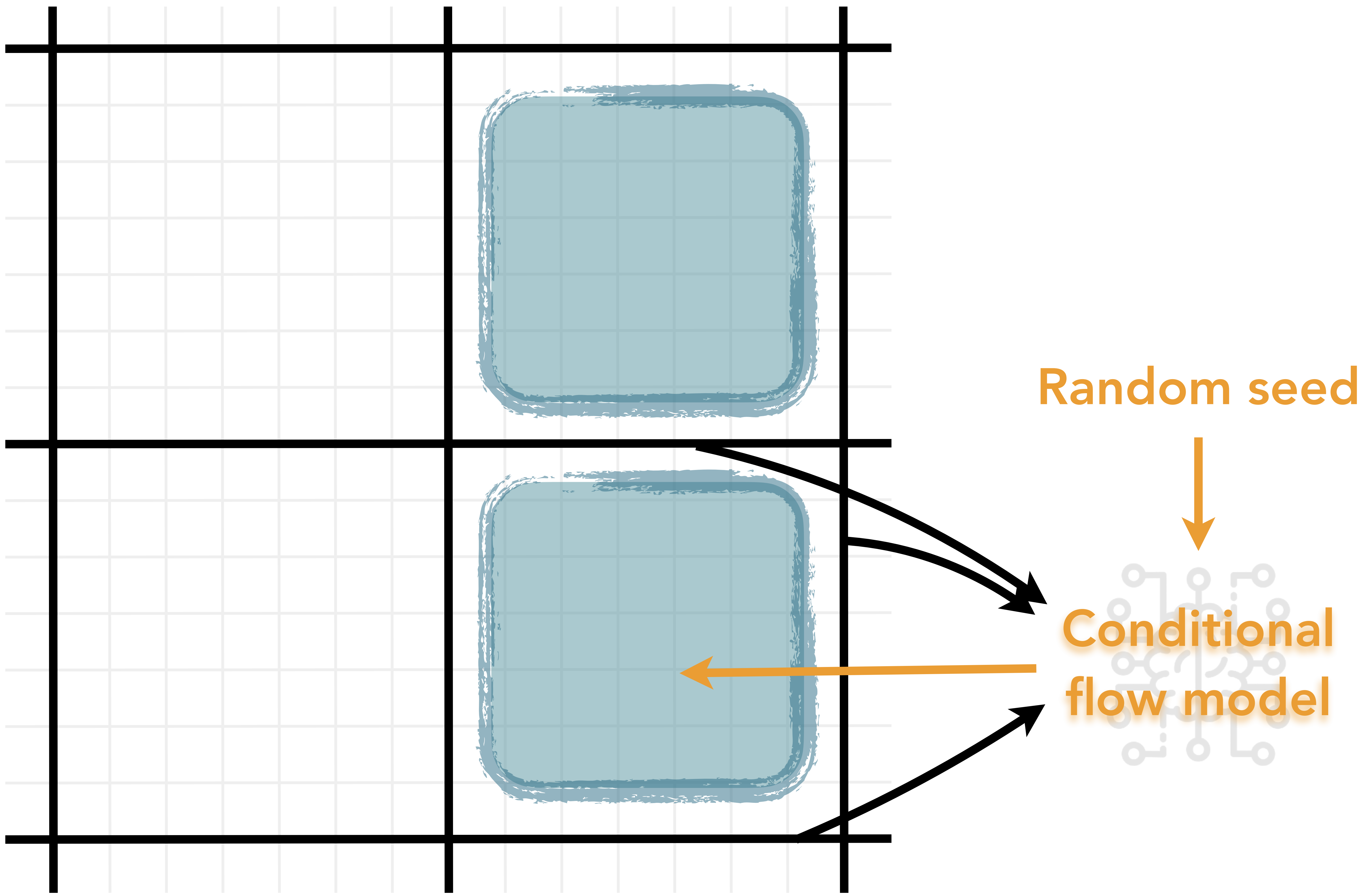
+

Partition function estimation for efficient outer sampling

+

Exponential error reduction by multi-level estimates

A possible future 🏁



Storage-free inner configurations

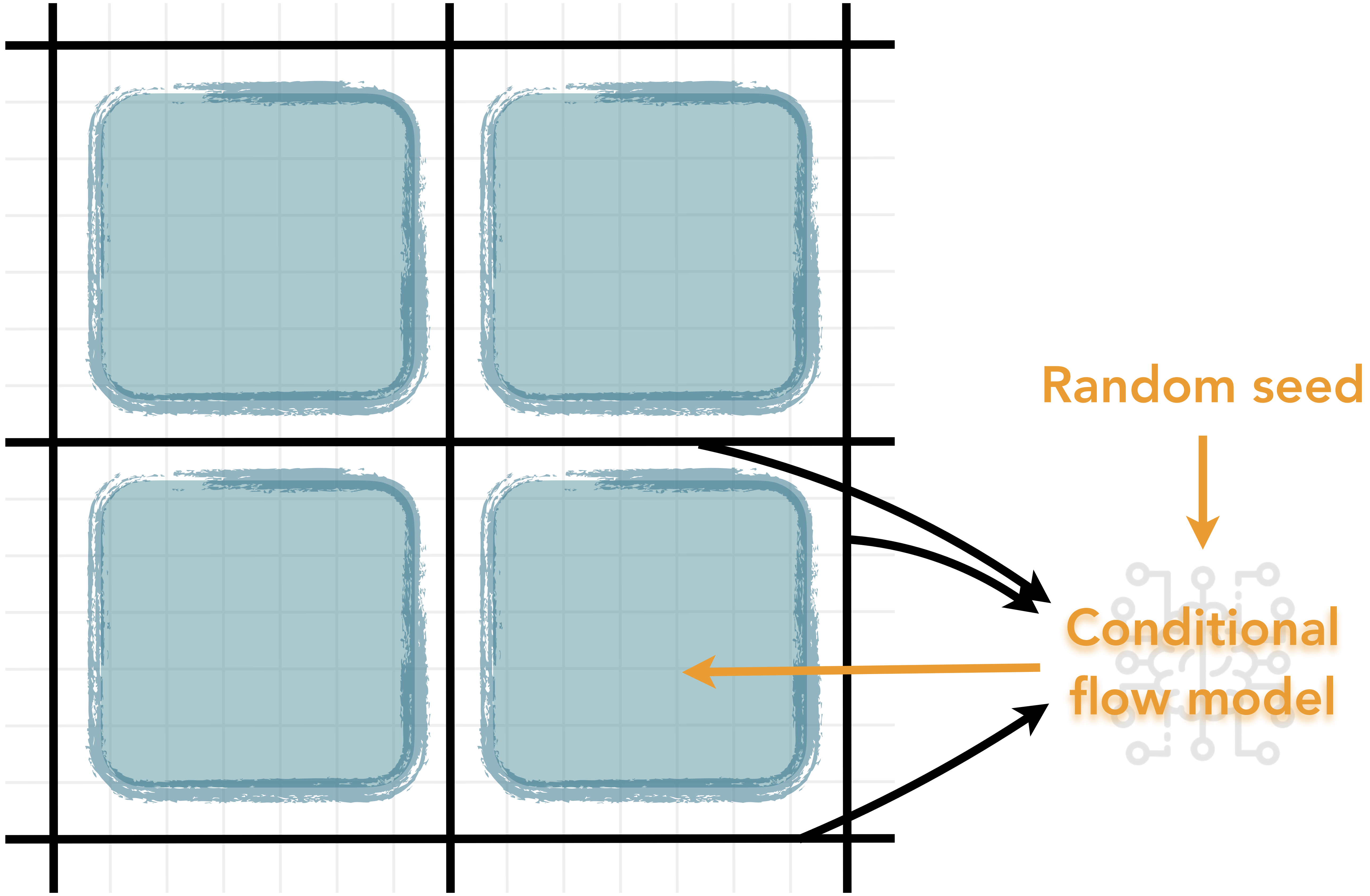
+

Partition function estimation for efficient outer sampling

+

Exponential error reduction by multi-level estimates

A possible future 🏁



Storage-free inner configurations

+

Partition function estimation
for efficient outer sampling

+

Exponential error reduction
by multi-level estimates

A possible future

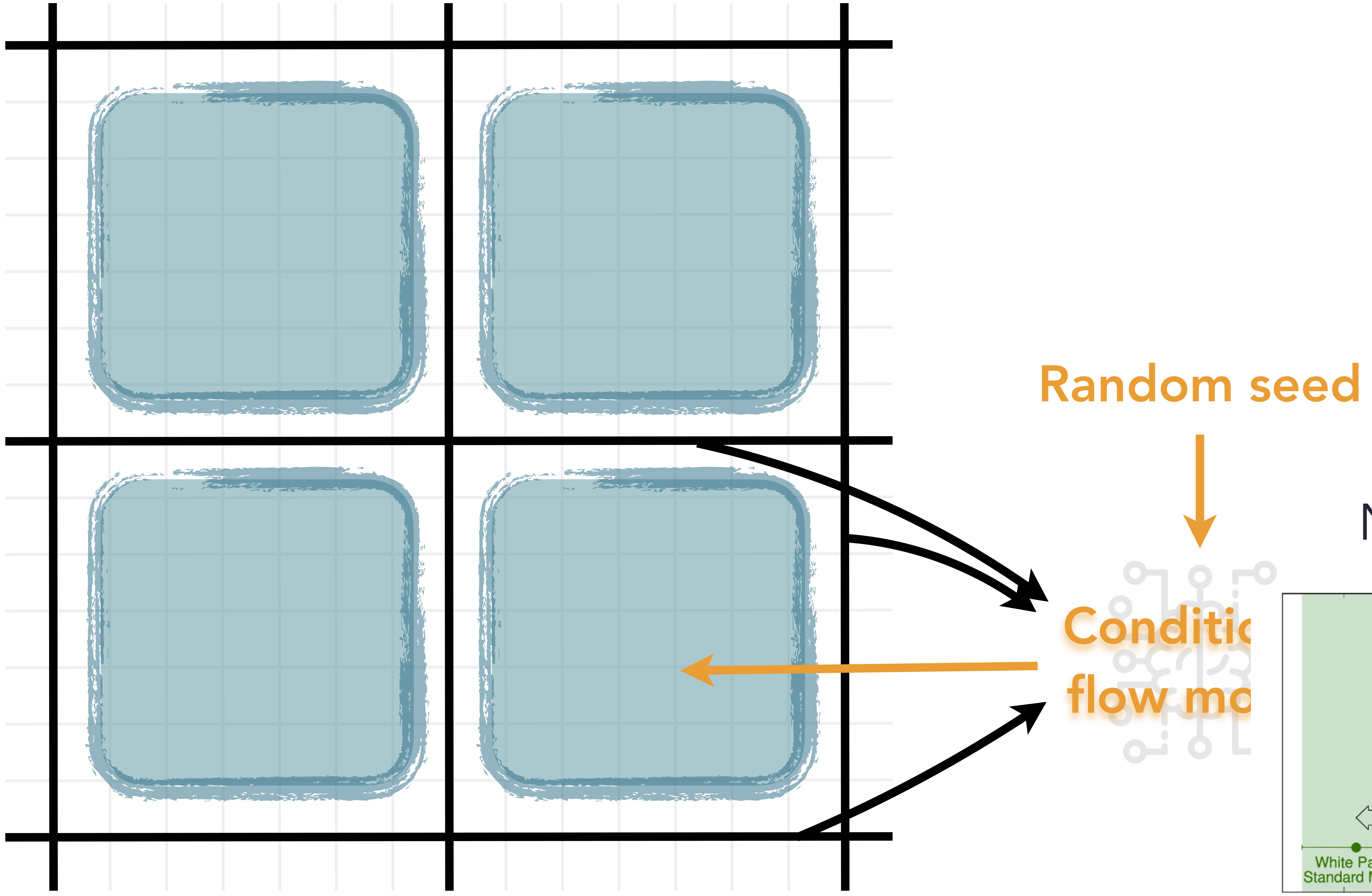
Storage-free inner configurations

+

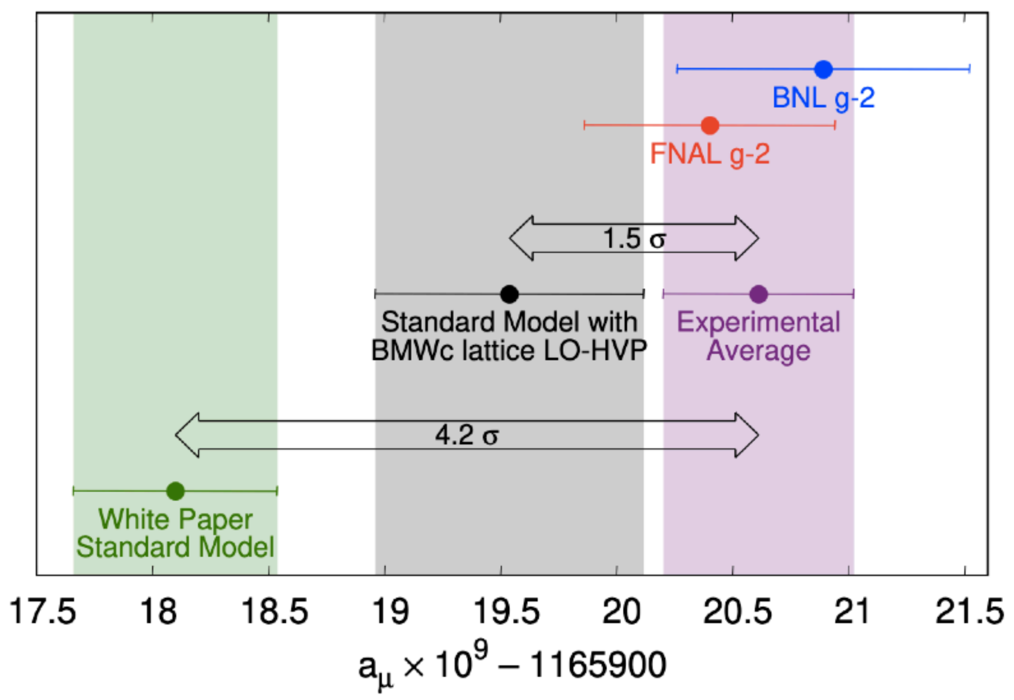
Partition function estimation for efficient outer sampling

+

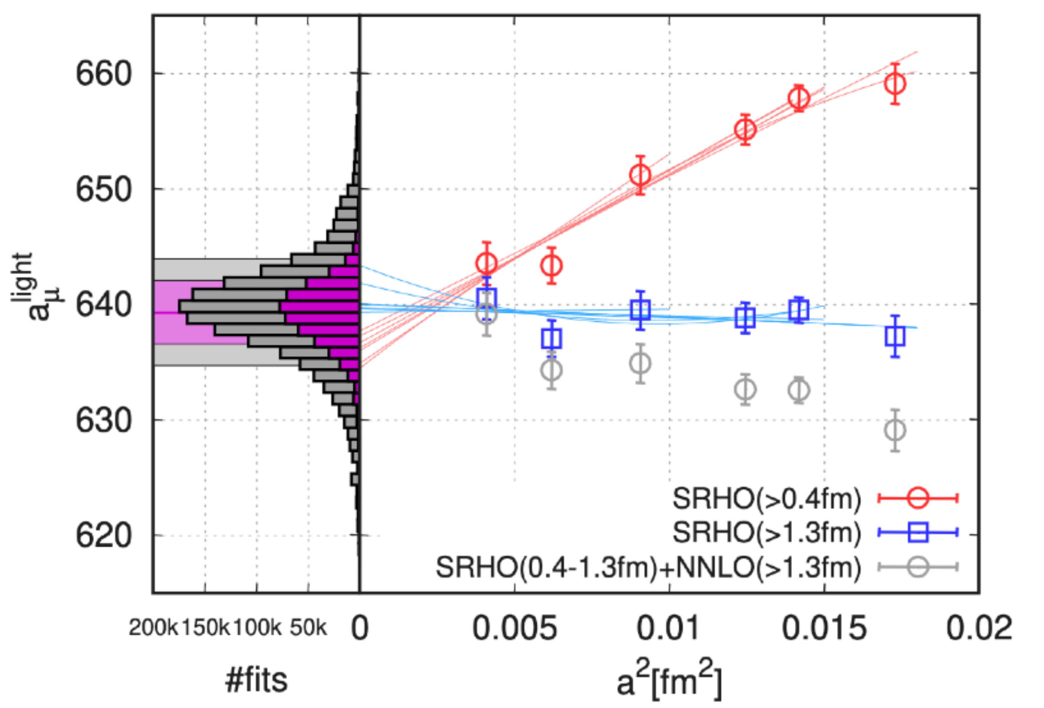
Exponential error reduction by multi-level estimates



Novel and more precise QCD studies, e.g. …



[Aoyama+ Muon g-2 WP (2020) 2006.04822]
[Lellouch, Moriond 2022]



[Borsanyi+ Nature593 (2021) 51]

Thanks! Questions?

What about volume scaling?

[Abbott+ "Aspects of scaling and scalability…" 2211.07541]

Fixed models will always* scale exponentially poorly with the **physical volume**.

- Expect variance of log reweighting factors to scale as $(L/\xi)^d$
  Scaling relation $\mathrm{ESS}(V) = \mathrm{ESS}(V_0)^{V/V_0}$, where $V_0 \sim \xi^d$

* in a direct sampling scheme

- This says nothing about scaling towards the continuum limit!

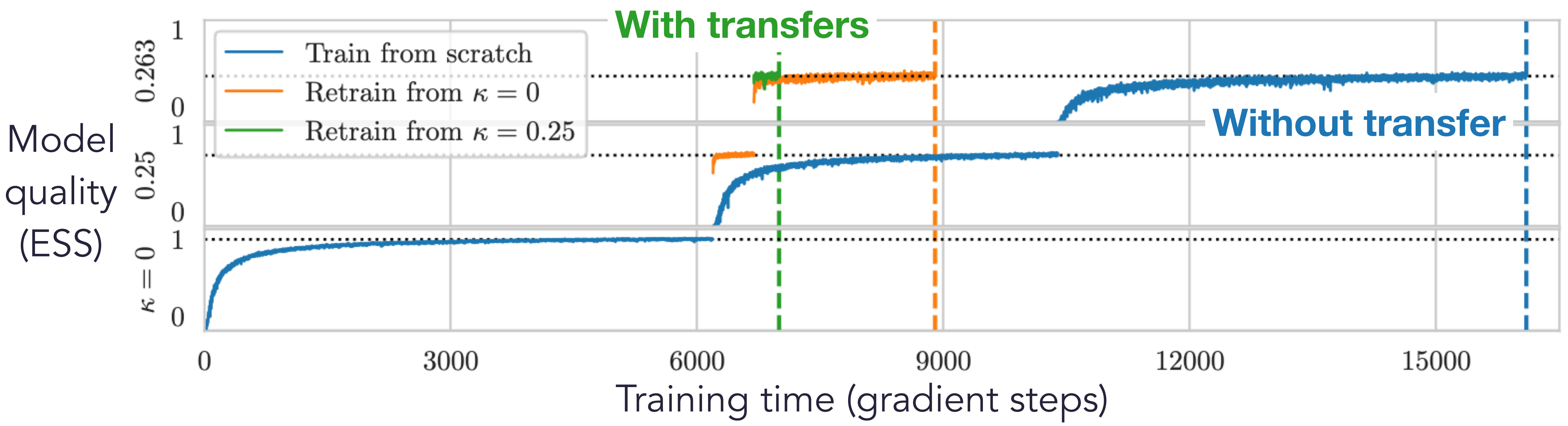
We should be thinking about targeting boxes of size $\approx \xi^d$.

- For larger volumes, hybrid/multilevel sampling schemes should be used

Transfer learning

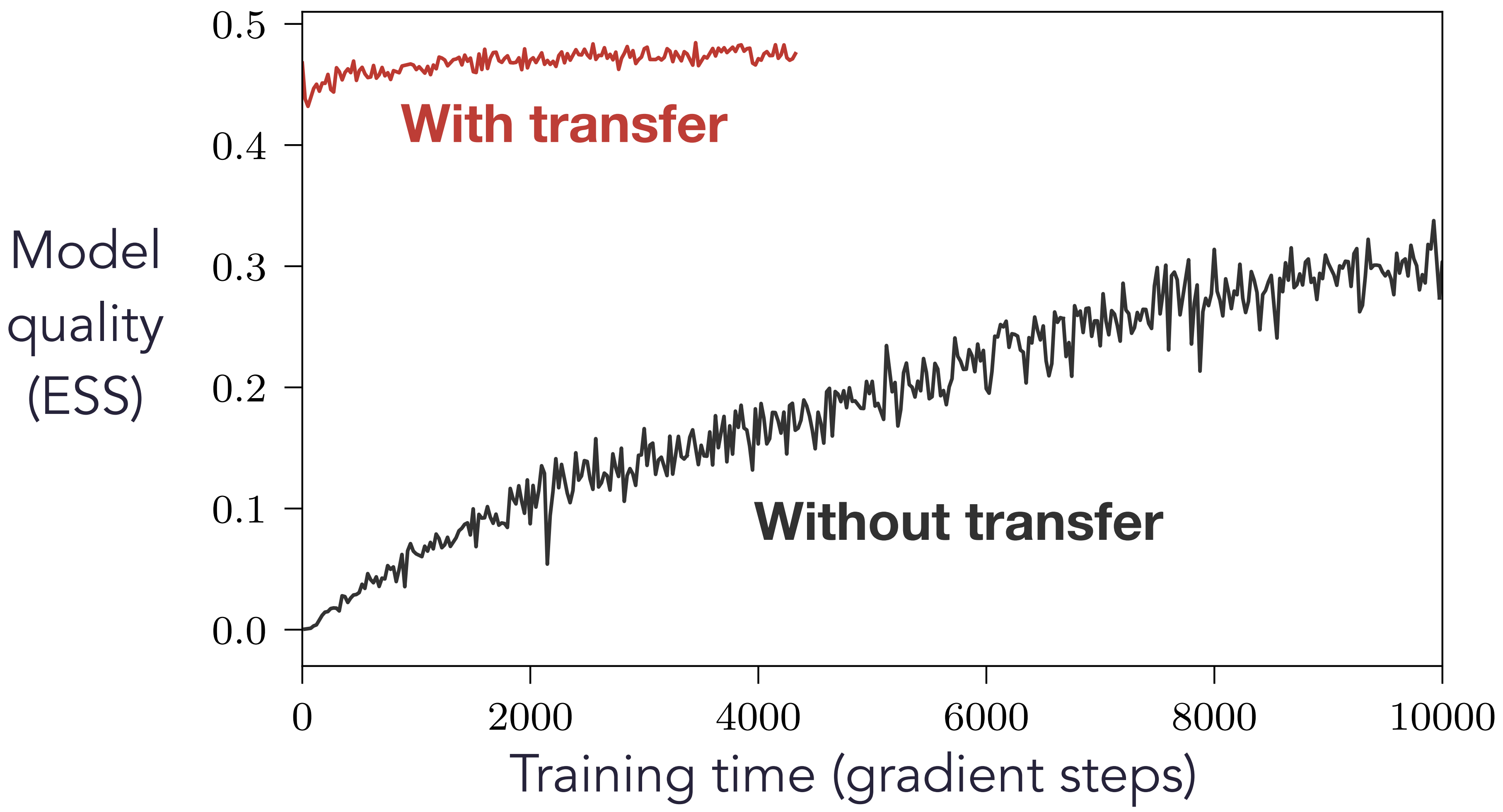
Both **parameter transfer** and **volume transfer** are highly effective for lattice field theory.

[Abbott+ "Aspects of scaling and scalability…" 2211.07541]



- Schwinger model [U(1) gauge theory + fermions]
- Parameter transfer $\kappa = 0 \to 0.25 \to 0.263 (\kappa_{\rm cr})$

[Boyda, **GK**, … PRD103 (2021) 074504]

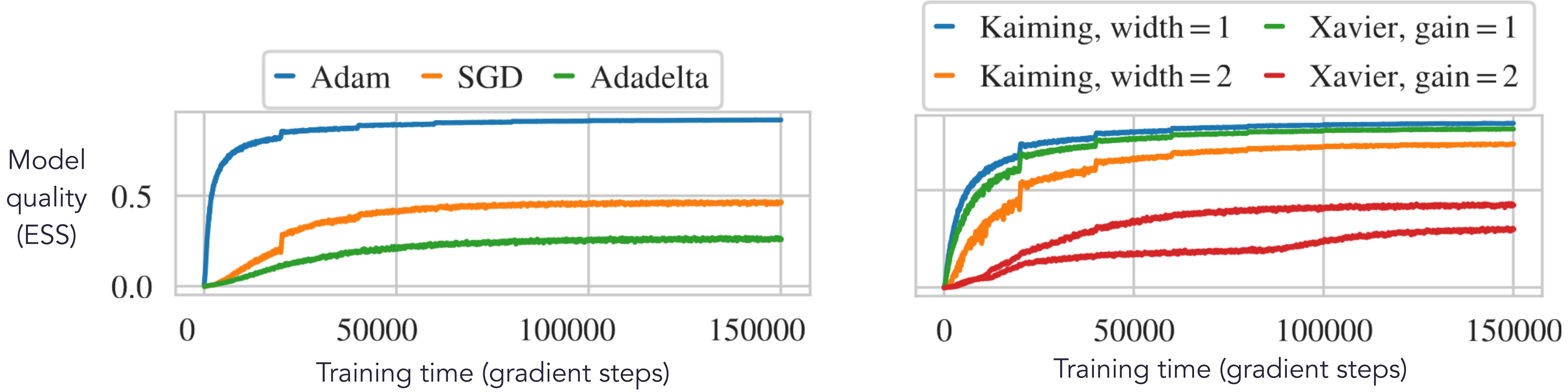


- SU(N) gauge theory
- Volume transfer $8 \times 8 \to 16 \times 16$ (red)
- Directly start at 16×16 (black)

Hyperparameters can make a big difference

Optimization algorithm, hyperparameters,
and initialization have strong effects on training rate.

[Abbott+ "Aspects of scaling and scalability..." 2211.07541]



Flows vs the trivializing map

Trivializing map: [Lüscher CMP293 (2010) 899]

- Interpolation between simple theory $r(V)$ and target theory $p(U) = e^{-S[U]}/Z$

- Phrased as a transport problem, formally solvable

$$f(V) = \int_0^T dt\, \nabla \varphi(U(t);t)\Big|_{U(0)=V} + V$$

$$\ln \det J = -\int_0^T dt\, \nabla^2 \varphi(U(t);t)$$

Note: For compact spaces, derivatives and integrals should be appropriately modified to act in the space.

- Expansion in t to estimate required potential $\varphi(U(t);t)$
  Not very computationally beneficial…

[Engel & Schaefer CPC182 (2011) 2107]