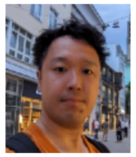
Lattice 2023, Fermilab, Illinois, USA, August 1st, 2023
https://indico.fnal.gov/event/57249/contributions/270657/
31 July -> 2 Aug

2023 LATTICE

arXiv: 2306.11527

Equivariant transformer is all you need

Akio Tomiya (IPUT Osaka, Assistant Prof.)



+ Yuki Nagai (JAEA, Senior Scientist)

MLPhYS Foundation of "Machine Learning Physics"
Grant-in-Aid for Transformative Research Areas (A)

Program for Promoting Researches on the Supercomputer Fugaku
Large-scale lattice QCD simulation and development of AI technology

akio_at_yukawa.kyoto-u.ac.jp

Outline

1. What is Machine learning?
2. Transformer and Attention
3. Target system
4. Equivariant Attention
5. SLMC (self-learning MC)
6. Results

Short summary

Akio Tomiya

Towards to simulate LQCD: equivariant attention

- We propose Attention blocks for physical systems!
  - Machine learning for physics (Monte-Carlo + neural net approx)
  - It keeps field rotation/translation symmetry (equivariant)
  - It can capture *non-local correlation* while CNN-type is hard to do
- We perform self-learning Monte-Carlo with the attention for "O(3) Yukawa system" system in condensed matter
  - *Not for gauge system*. Only for global symmetry
- We find that the attention layers improve acceptance rates systematically for increasing the number layers
  - It shows scaling behavior as in large language models

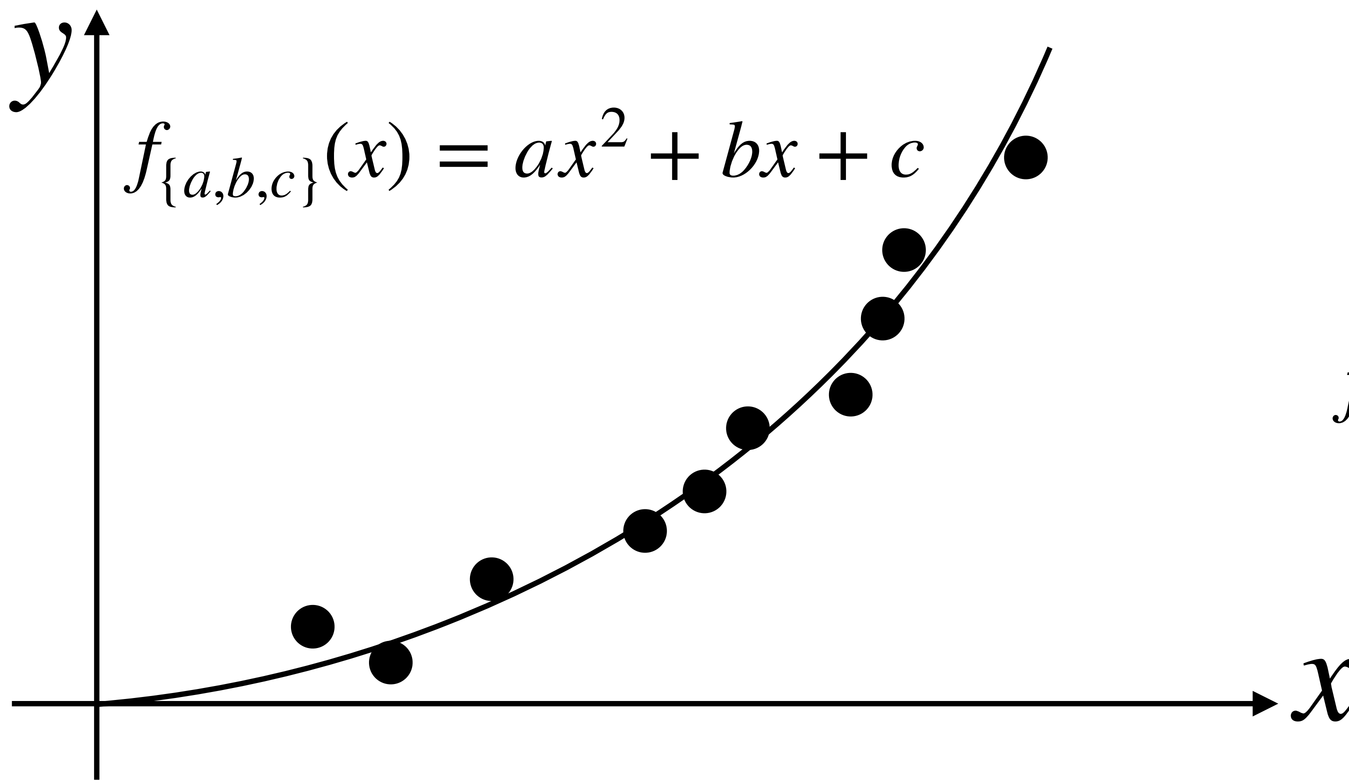
(c.f. We have proposed **gauge symmetric convolution** applied on 4d Full QCD, see arXiv: 2103.11965)

What is machine learning?
Symmetry?

What is machine learning?

E.g. Linear regression (supervised learning)

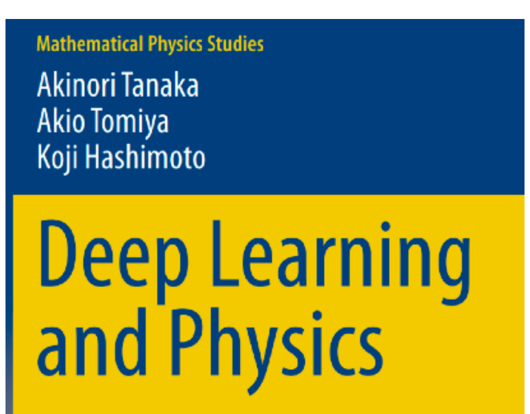
Data: $D = \left\{ (x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \cdots \right\}$



$f_{\{a,b,c\}}(x) : \ \mathbb{R} \to \mathbb{R}$

$a, b, c,$ are determined by minimizing E
(training = fitting by data)

$$E = \frac{1}{2} \sum_d \left| f_{\{a,b,c\}}(x^{(d)}) - y^{(d)} \right|^2$$



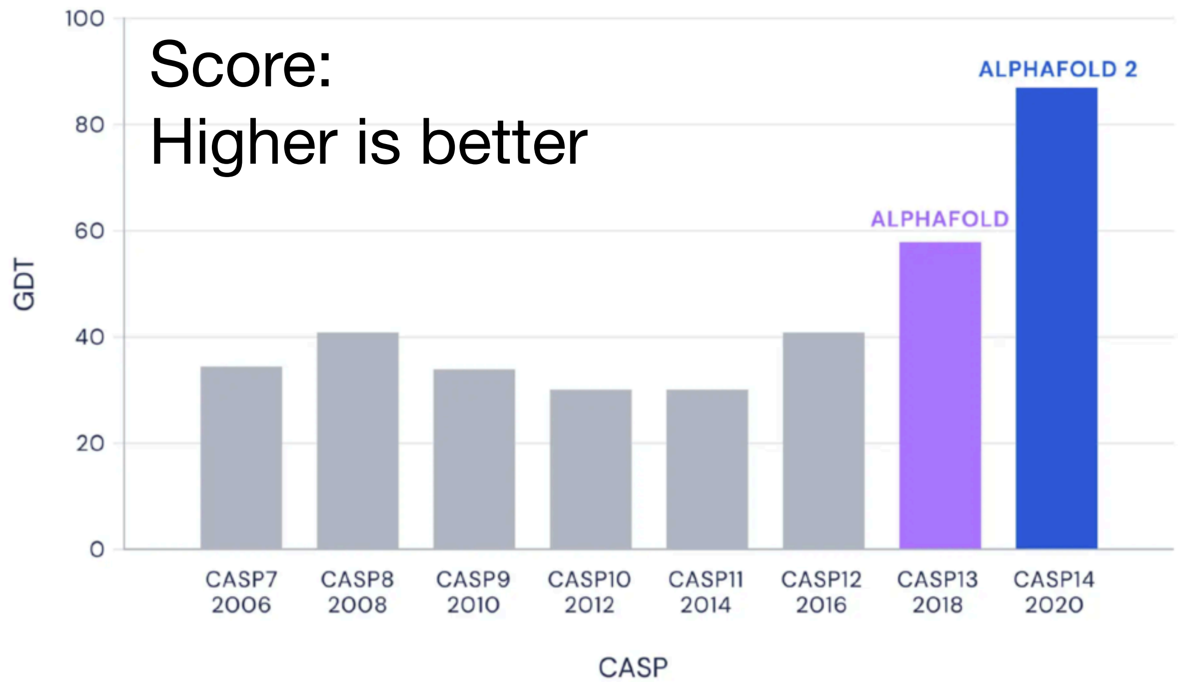
In physics language, variational method (with fitting)

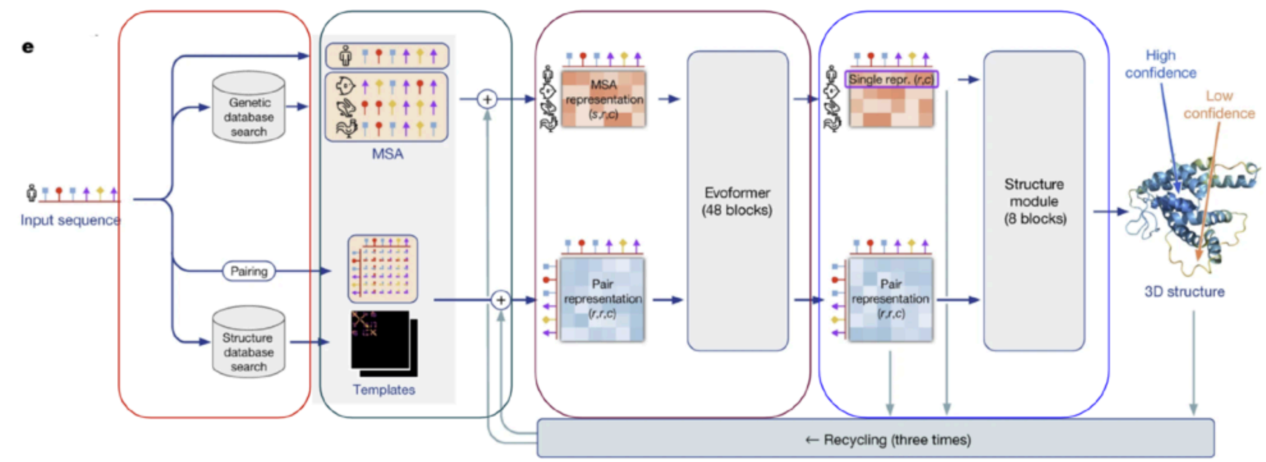
Data determines a function form

Equivariance and convolution

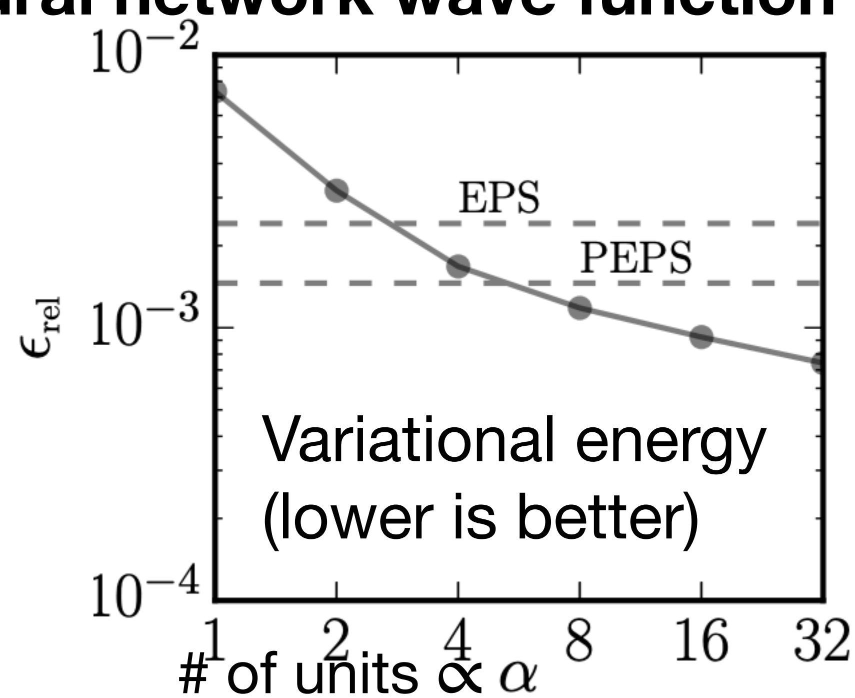
Neural network works quite well in natural science

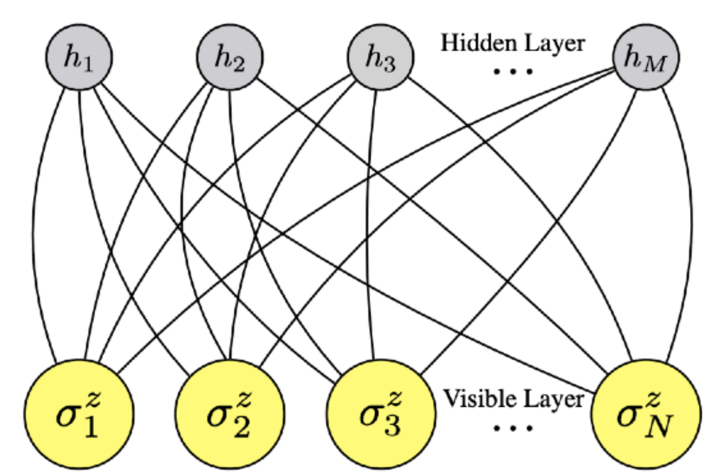
Protein Folding problem (AlphaFold2, John Jumper+, Nature, 2020+), Transformer





Neural network wave function for many body (Carleo Troyer, Science 355, 602 (2017))





Neural net + "Expert knowledge" → Best performance

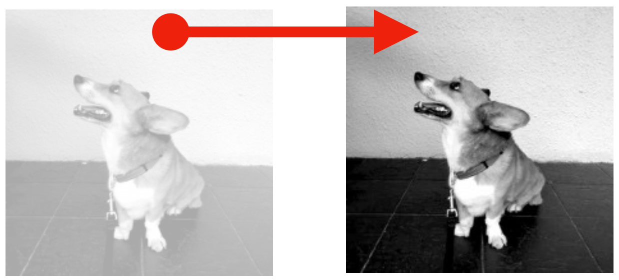
https://horomary.hatenablog.com/entry/2021/10/01/194825
https://www.aist.go.jp/aist_j/press_release/pr2020/pr20201111/pr20201111.html

Equivariance and convolution

Akio Tomiya

Knowledge ∋ Convolution layer = trainable filter, Equivariant

Filter on image



shift to right

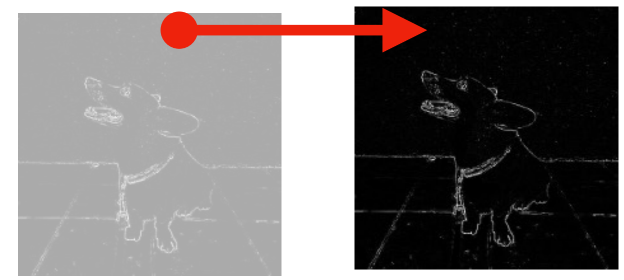
$*$

Laplacian filter

0	1	0
1	-2	1
0	1	0

(Discretization of ∂^2)

$=$

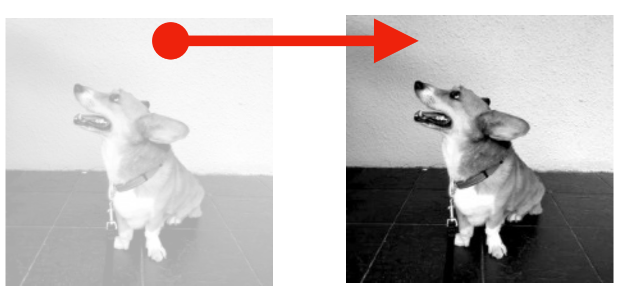


shift to right

Edge detection

Translational operation is *commutable* with filtering (equivariant)

Convolution layer



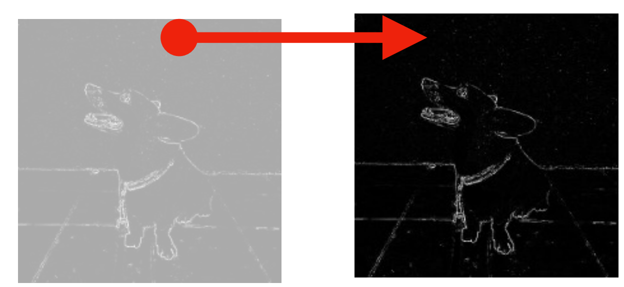
shift to right

$*$

Trainable filter

w_{11}	w_{12}	w_{13}
w_{21}	w_{22}	w_{23}
w_{31}	w_{32}	w_{33}

$=$



shift to right

Fukushima, Kunihiko (1980)
Zhang, Wei (1988) + a lot!

Translational operation is *commutable* with convolutional neurons (equivariant)

This can be any filter which helps feature extraction (minimizing loss)

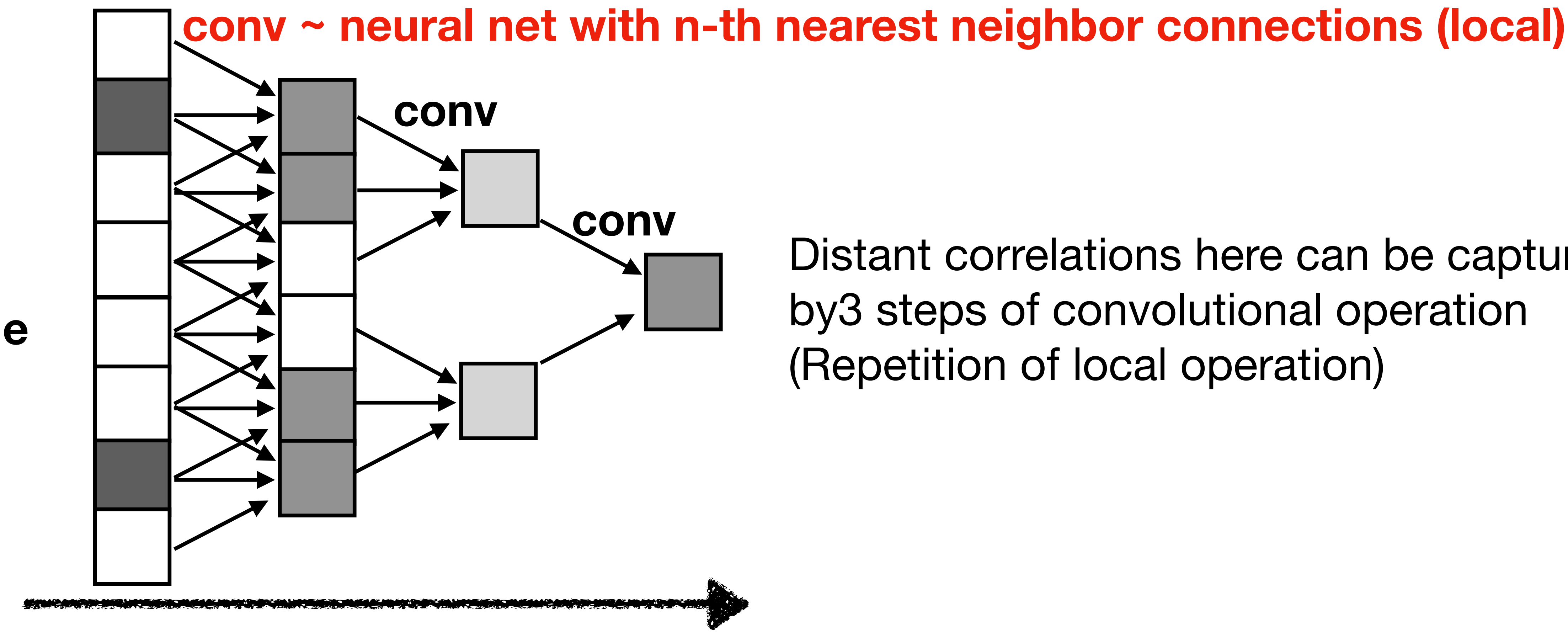
Equivariance reduces data demands. Ensuring symmetry (plausible Inference)

Equivariance and convolution

Convolutional Neural network have been good job but local

Convolutional neural layers in neural networks keep translational symmetry,
it can be generalized to any continuous/discrete symmetry in the theory. It helps generalization.

e.g.
1d Input image



Distant correlations here can be captured by3 steps of convolutional operation (Repetition of local operation)

However, 1 step of convolutional layer can pick up only local correlation and representability of neural networks is limited. Global correlations are sometimes important.
How can we overcome these difficulties?

Transformer and Attention

Transformer and Attention

Attention layer used in Transformers (GPT, Bard)

arXiv:1706.03762


arXiv:1706.03762v6 [cs.CL] 24 Jul 2023

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

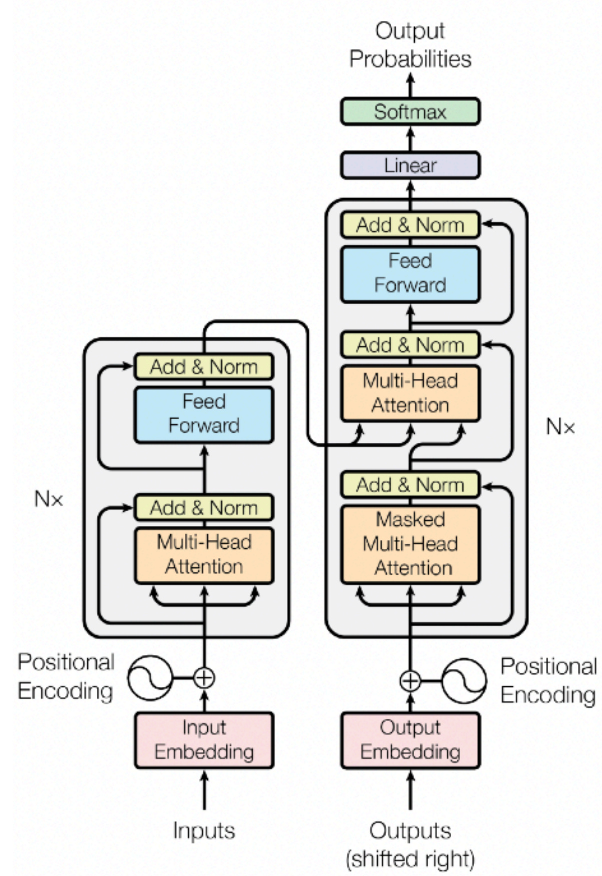


Figure 1: The Transformer - model architecture.

OpenAI ChatGPT

Bard

DeepL

Attention layer (in transformer model) has been introduced in a paper titled
"Attention is all you need" (1706.03762)
State of the art architecture of language processing.
Attention layer is essential.

Transformer and Attention

Akio Tomiya

Attention layer can capture non-local correlations

arXiv:1706.03762

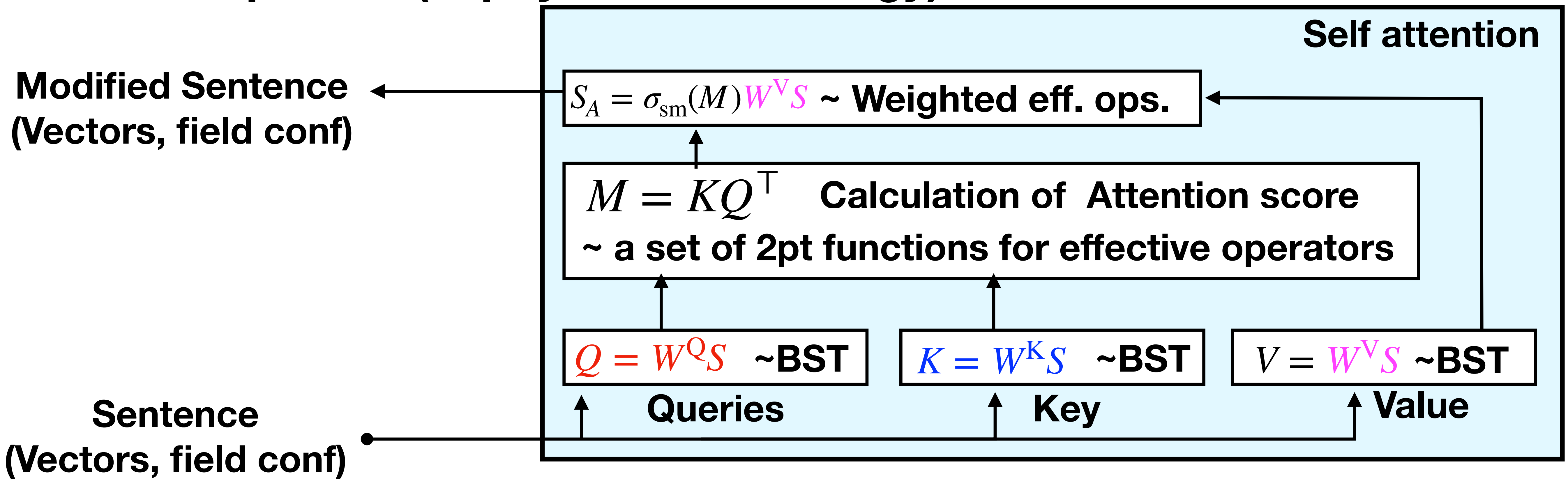
Modifier in language can be non-local

Eg. I am Akio Tomiya living in Japan, who studies machine learning and physics

In physics terminology, this is **non local correlation.**

The attention layer enables us to treat non-local correlation with a neural net!

<u>Schematic picture (in physics terminology)</u>



Akio Tomiya

Transformer and Attention

Transformer shows scaling lows (power law)

arXiv: 2001.08361

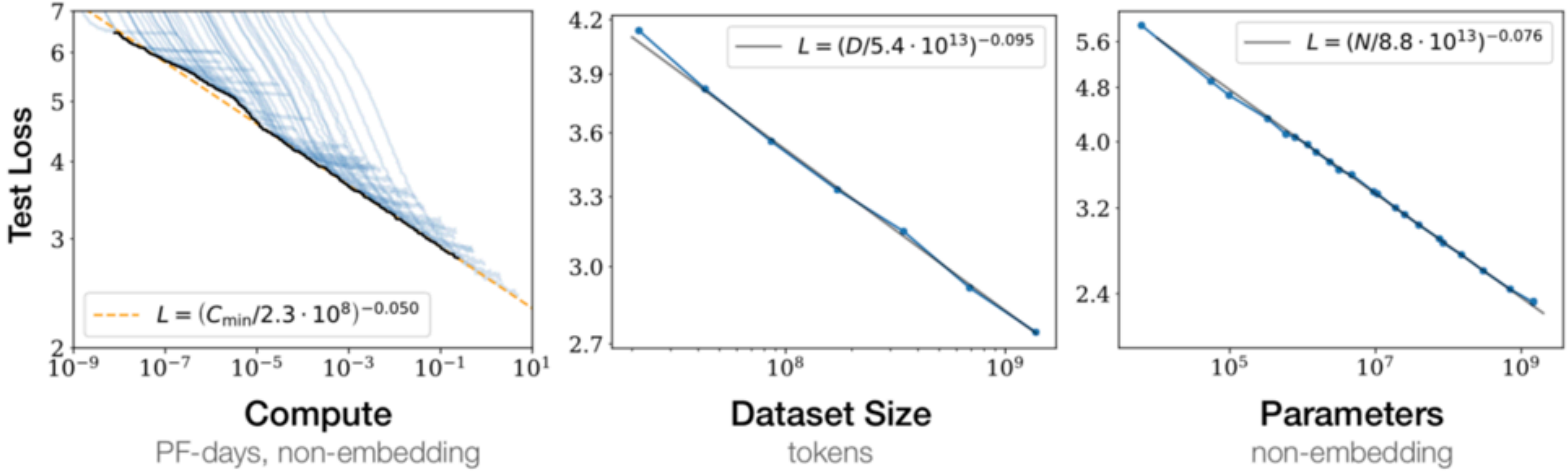


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

- **Transformers requires huge data**
  (e.g. GPT uses all electric books in the world)
  Because it has few inductive bias (no equivariance)
- **It can be improved systematically (obey scaling law)**

Transformer and Attention

Akio Tomiya

Physically symmetric Attention layer

Attention layer can capture global correlation
Equivariance reduces data demands for training

	Equivariance	**Capturable correlation**	**Data demmands**	**Applications**
Convolution ($\in$ equivariant layers)	Yes 👍	Local 😲	Low 👍	Image recognition VAE, GAN Normalizing flow
Standard Attention layer	No 😲	Global 👍	Huge 😲	ChatGPT Bard Vision Transformer arXiv:1706.03762
(This work) Physically *Equivariant* attention	Yes 👍	Global 👍	?	This work arXiv: 2306.11527

Target system and its symmetry

Target system and its symmetry

Akio Tomiya

Monte-Carlo + self-learning

Target system: Classical Heisenberg spin $\mathbf{S}_i$+ Fermion on 2d lattice

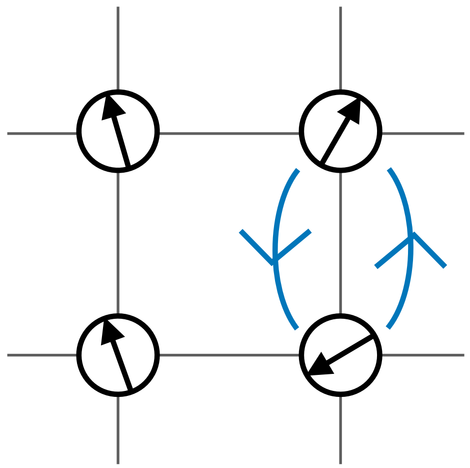
$$H = -t \sum_{\alpha,\langle i,j\rangle} (\hat{c}^{\dagger}_{i\alpha}\hat{c}_{j\alpha} + \mathrm{h.c.}) + \frac{J}{2}\sum_{i} \mathbf{S}_i \cdot \hat{\sigma}_i$$

Symmetries:

- Global O(3)

- Translational

- 90-deg rotation



$$[\hat{\sigma}_i]_\gamma \equiv \hat{c}^{\dagger}_{i\alpha}\sigma^{\gamma}_{\alpha\beta}\hat{c}_{i\beta}$$

$$\mathbf{S}_i = \begin{pmatrix} S_i^1 & S_i^2 & S_i^3 \end{pmatrix}^{\top}$$

$$S_i^{\mu} \in \mathbb{R}$$

In lattice QCD language, **Yukawa-theory with O(3) scalar field**

$\mathbf{S}_i$: 3 component scalar field on site i

$\hat{c}_{i\alpha}$: Fermion (annihilation op.) at site i with spin α

Toy model for QCD.

Target system and its symmetry

Akio Tomiya

Previous work

Target system: Classical Heisenberg spin $\mathbf{S}_i$+ Fermion on 2d lattice

$$H = -t \sum_{\alpha,\langle i,j\rangle} (\hat{c}^\dagger_{i\alpha}\hat{c}_{j\alpha} + \mathrm{h.c.}) + \frac{J}{2}\sum_i \mathbf{S}_i \cdot \hat{\sigma}_i$$

Integrate out fermions $\hat{c}^\dagger_{i\alpha}, \hat{c}_{i\alpha}$:

$$Z = \sum_{\{\mathbf{S}\}} \prod_n (1 + e^{-\beta(\mu - E_n(\{\mathbf{S}\}))})$$

Non-local. Difficult.
Time consuming

Using Hopping parameter expansion, we get a local effective model
This is used in a previous work:

$$H^{\mathrm{Linear}}_{\mathrm{eff}} = -\sum_{\langle i,j\rangle} J^{\mathrm{eff}} \mathbf{S}_i \cdot \mathbf{S}_j + E_0$$

Fit (J^{eff}, E_0)

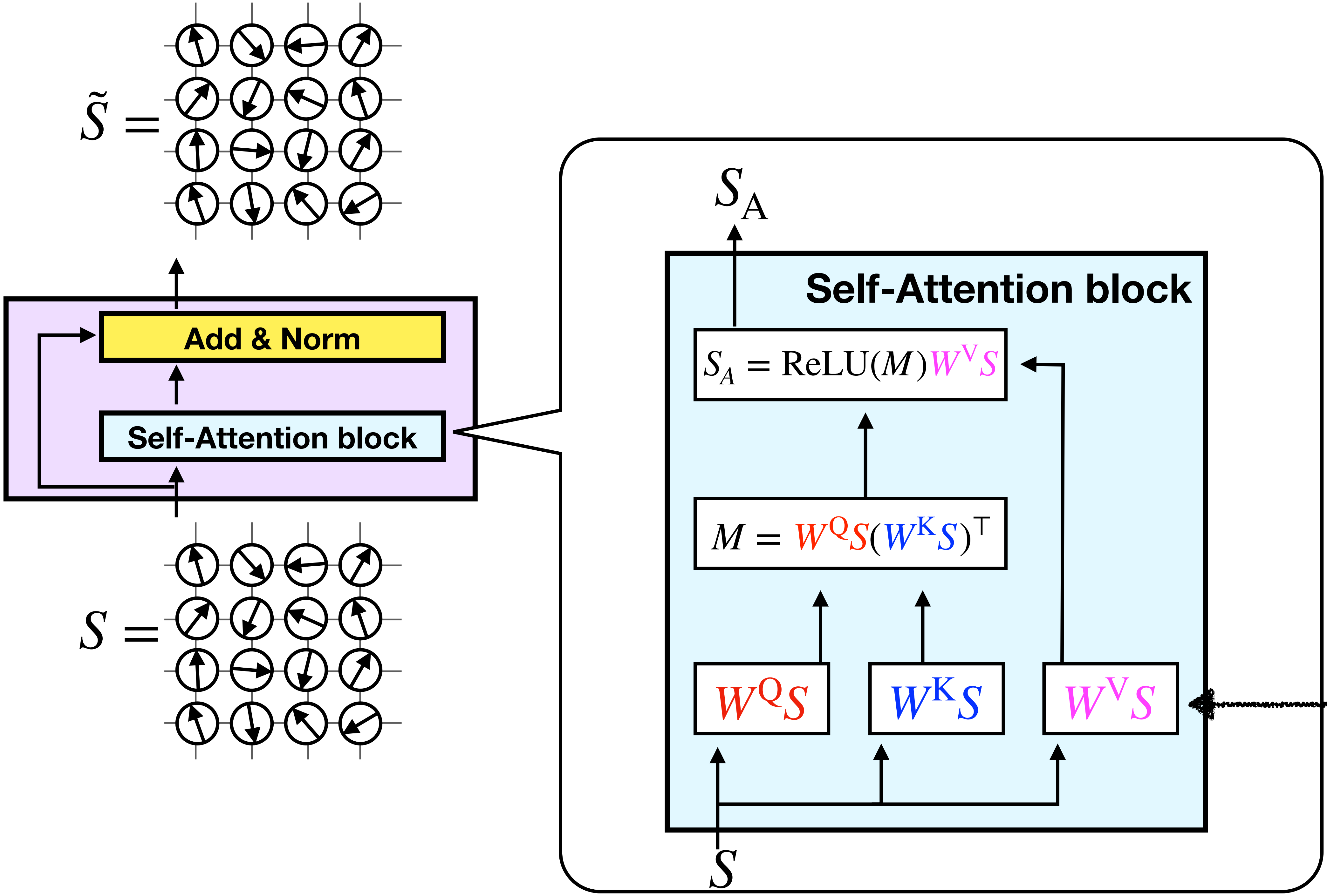
Equivariant attention

Self-learning Monte-Carlo

Akio Tomiya

arXiv: 2306.11527.

Attention block makes effective spin field with non-local BST



Smearing
Rot. equivariant
Trsl. equivariant
trainable!

Self-learning Monte-Carlo

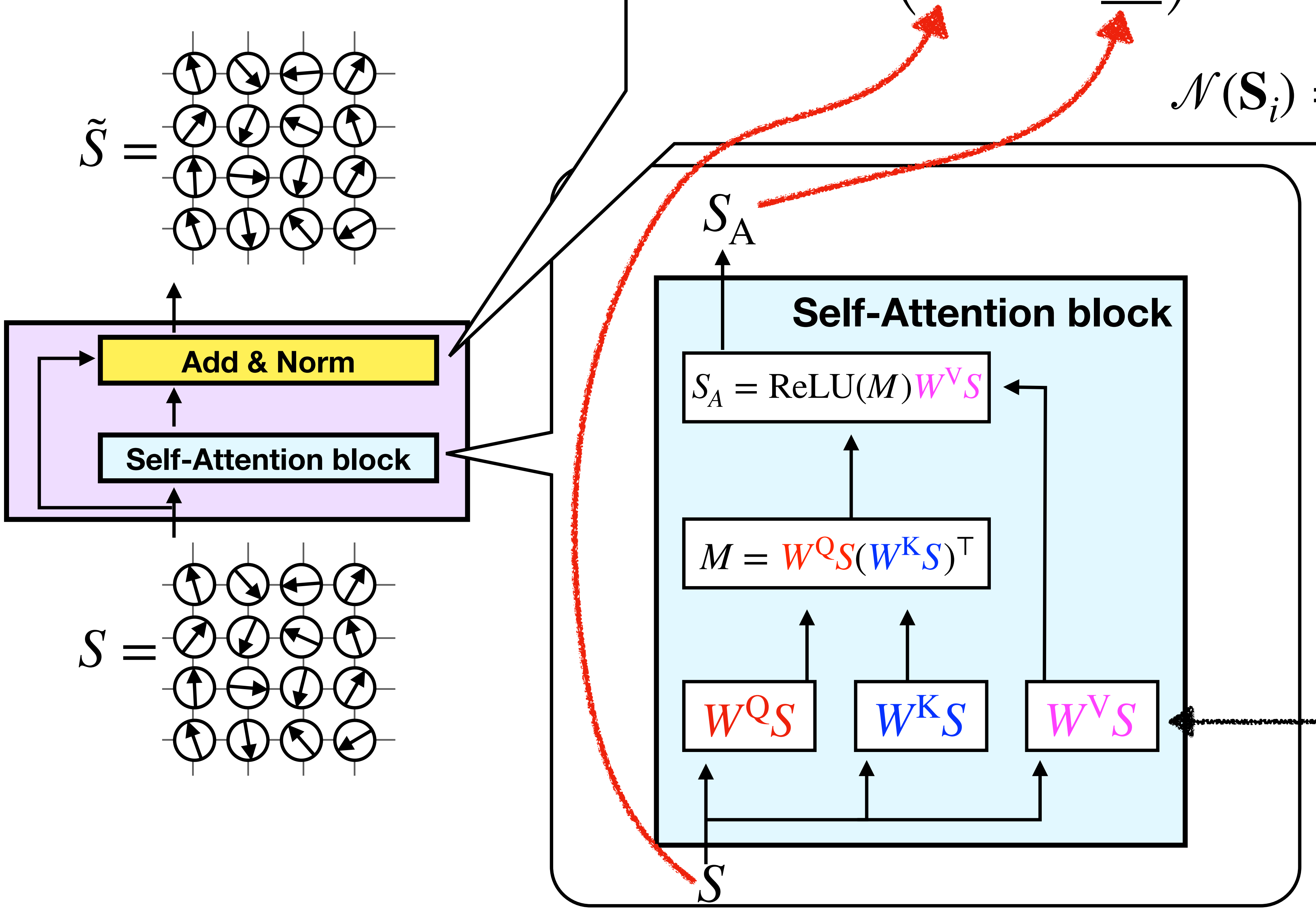
Akio Tomiya

arXiv: 2306.11527.

Attention block makes effective spin field with non-local BST

$$\mathbf{S}^{(l)} \equiv \mathcal{N}\left(\mathbf{S}^{(l-1)} + \underline{\mathbf{S}_{\mathrm{A}}}\right)$$ position-wise

$$\mathcal{N}(\mathbf{S}_i) = \mathbf{S}_i / \|\mathbf{S}_i\|$$



Smeared fields
Rot. equivariant
Trsl. equivariant
Skip connection
Normalized!

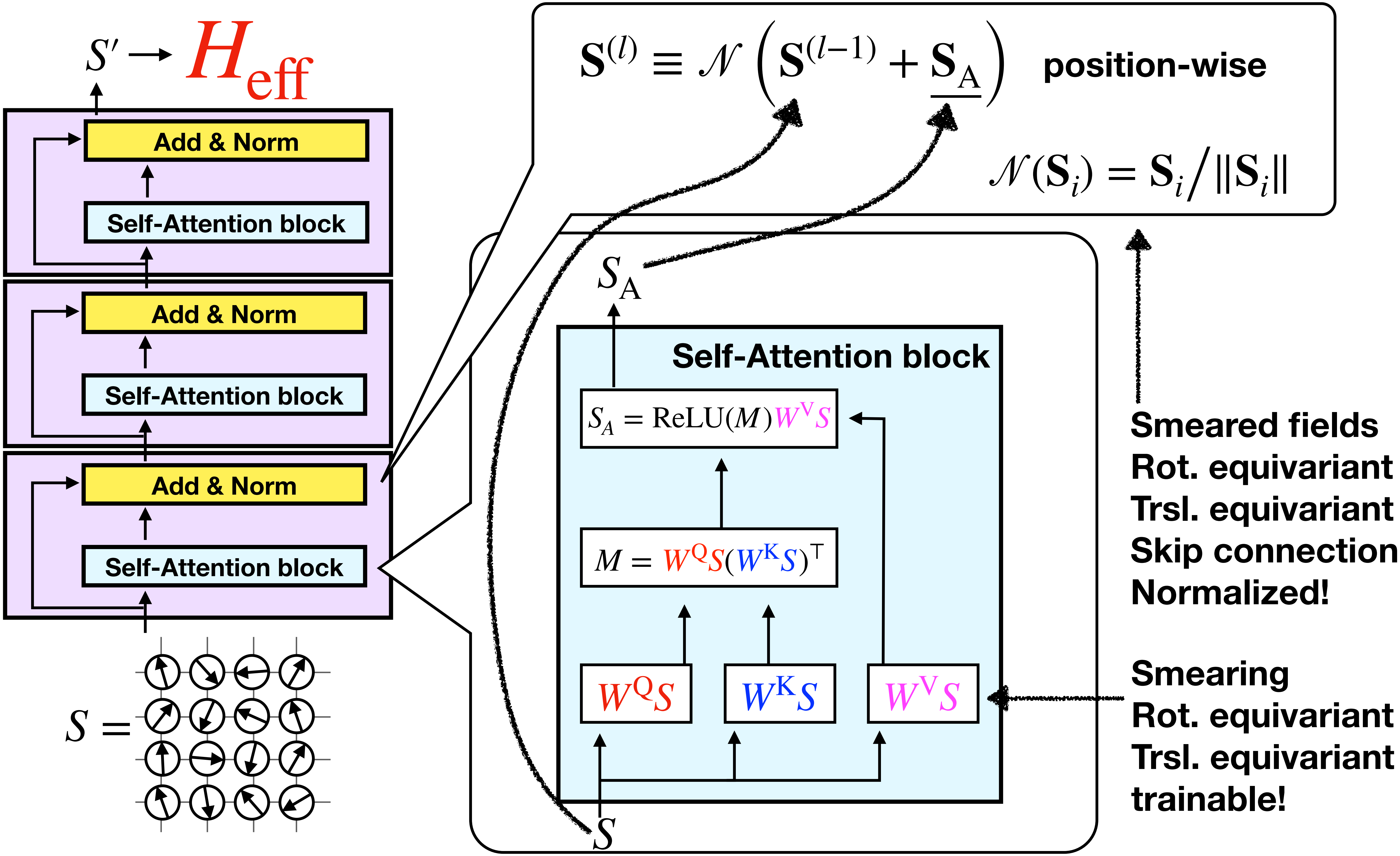
Smearing
Rot. equivariant
Trsl. equivariant
trainable!

Self-learning Monte-Carlo

Akio Tomiya

arXiv: 2306.11527.

Variational Hamiltonian with Equivariant Attention layers



SLMC
= Self-learning Monte Carlo
= MCMC with variational hamiltonian

Self-learning Monte-Carlo

SLMC = MCMC with an effective model

arXiv:1610.03137+

For statistical spin system, we want to calculate expectation value with

$$W(\{\mathbf{S}\}) \propto \exp[-\beta H(\{\mathbf{S}\})]$$

On the other hand, an effective model $H_{\mathrm{eff}}(\{\mathbf{S}\})$ can compose MCMC,

$\{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\}$ this distributes $W_{\mathrm{eff}}(\{\mathbf{S}\}) \propto \exp[-\beta H_{\mathrm{eff}}(\{\mathbf{S}\})]$

if the update「→」 satisfies the detailed balance. We can employ Metropolis test like

$$A_{\mathrm{eff}}(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min\left(1, W_{\mathrm{eff}}(\{\mathbf{S}'\})/W_{\mathrm{eff}}(\{\mathbf{S}\})\right).$$

Self-learning Monte-Carlo

Akio Tomiya

SLMC = MCMC with an effective model

arXiv:1610.03137+

For statistical spin system, we want to calculate expectation value with

$$W(\{\mathbf{S}\}) \propto \exp[-\beta H(\{\mathbf{S}\})]$$

On the other hand, an effective model $H_{\rm eff}(\{\mathbf{S}\})$ can compose MCMC,

$\{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\}$ this distributes $W_{\rm eff}(\{\mathbf{S}\}) \propto \exp[-\beta H_{\rm eff}(\{\mathbf{S}\})]$

if the update「→」 satisfies the detailed balance. We can employ Metropolis test like

$$A_{\rm eff}(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min\left(1, W_{\rm eff}(\{\mathbf{S}'\})/W_{\rm eff}(\{\mathbf{S}\})\right).$$

SLMC: Self-learning Monte-Carlo

We can construct *double* MCMC with $H(\{\mathbf{S}\})$ and $H_{\rm eff}(\{\mathbf{S}\})$

$$\{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \Rightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \longrightarrow \{\mathbf{S}\} \Rightarrow$$

with Metropolis-*Hastings* test: $A(\{\mathbf{S}'\}, \{\mathbf{S}\}) = \min\left(1, \frac{W(\{\mathbf{S}'\})}{W(\{\mathbf{S}\})} \frac{W_{\rm eff}(\{\mathbf{S}\})}{W_{\rm eff}(\{\mathbf{S}'\})}\right).$

- **Effective model can have fit parameters**
- **Exact! It satisfies detailed balance with $W(\{\mathbf{S}\})$**
- **It has been used for full QCD too (arXiv: 2010.11900, 2103.11965)**

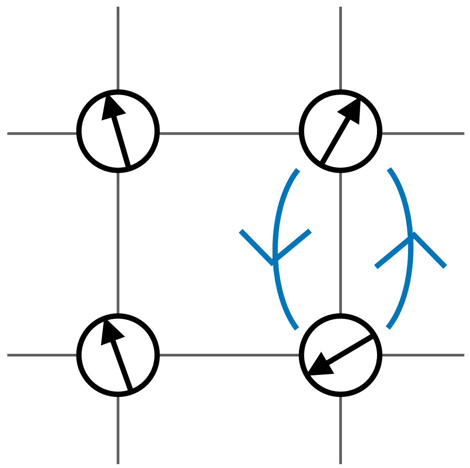
Self-learning Monte-Carlo

Monte-Carlo + self-learning

Target system: Classical Heisenberg spin $\mathbf{S}_i$+ Fermion on 2d lattice

$$H = -t \sum_{\alpha,\langle i,j\rangle} (\hat{c}^\dagger_{i\alpha}\hat{c}_{j\alpha} + \mathrm{h.c.}) + \frac{J}{2}\sum_i \mathbf{S}_i \cdot \hat{\sigma}_i$$

Symmetries:
- Global O(3)
- Translational
- 90-deg rotation



$$[\hat{\sigma}_i]_\gamma \equiv \hat{c}^\dagger_{i\alpha}\sigma^\gamma_{\alpha\beta}\hat{c}_{i\beta}$$

$$\mathbf{S}_i = \begin{pmatrix} S_i^1 & S_i^2 & S_i^3 \end{pmatrix}^\top$$

$$S_i^\mu \in \mathbb{R}$$

In lattice QCD language, **Yukawa-theory with O(3) scalar field**

$\mathbf{S}_i$: 3 component scalar field on site i

$\hat{c}_{i\alpha}$: Fermion (annihilation op.) at site i with spin α

Toy model for QCD.

Self-learning Monte-Carlo

Previous work

Target system: Classical Heisenberg spin $\mathbf{S}_i$+ Fermion on 2d lattice

$$H = -t \sum_{\alpha,\langle i,j \rangle} (\hat{c}^\dagger_{i\alpha} \hat{c}_{j\alpha} + \mathrm{h.c.}) + \frac{J}{2} \sum_i \mathbf{S}_i \cdot \hat{\sigma}_i$$

Integrate out fermions $\hat{c}^\dagger_{i\alpha}, \hat{c}_{i\alpha}$:

$$Z = \sum_{\{\mathbf{S}\}} \prod_n (1 + e^{-\beta(\mu - E_n(\{\mathbf{S}\}))})$$

Non-local. Difficult.
Time consuming

(In previous work), using Hopping parameter expansion, we can get a local effective model:

$$H^{\mathrm{Linear}}_{\mathrm{eff}} = -\sum_{\langle i,j \rangle} J^{\mathrm{eff}} \mathbf{S}_i \cdot \mathbf{S}_j + E_0$$

Fit (J^{eff})

In SLMC,
Poor scaling, poor representability = poor acceptance!

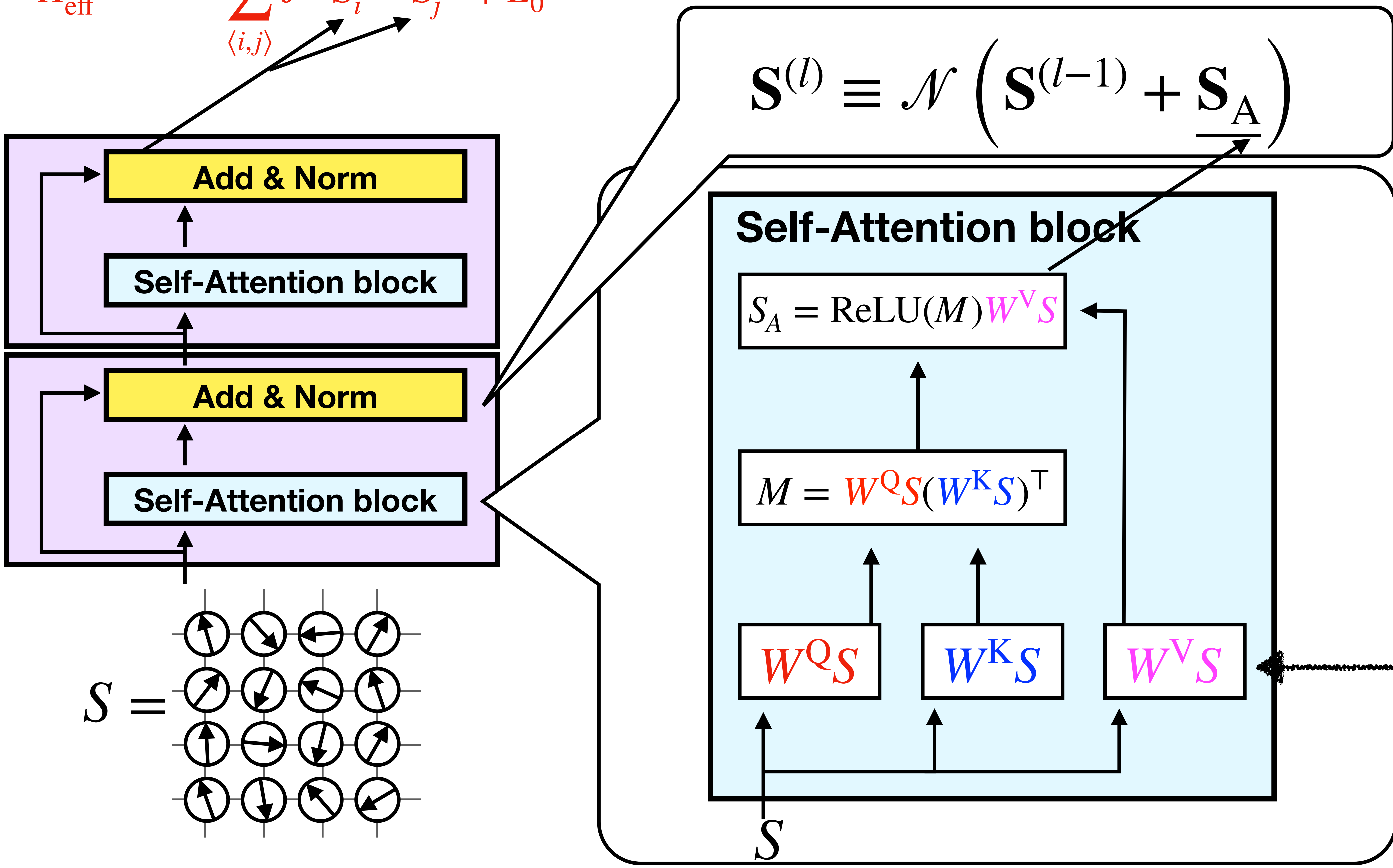
Self-learning Monte-Carlo

Equivariant Attention layer

arXiv: 2306.11527.

We can construct effective hamiltonian with output of Attention layer because “output of Attention = smeared fields with non-local correlation”

$$H_{\text{eff}}^{\text{Linear}} = -\sum_{\langle i,j \rangle} J^{\text{eff}} \mathbf{S}_i^{\text{eff}} \cdot \mathbf{S}_j^{\text{eff}} + E_0$$



Smeared fields
Rot. equivariant
Trsl. equivariant
trainable!

Smearing
Rot. equivariant
Trsl. equivariant
trainable!

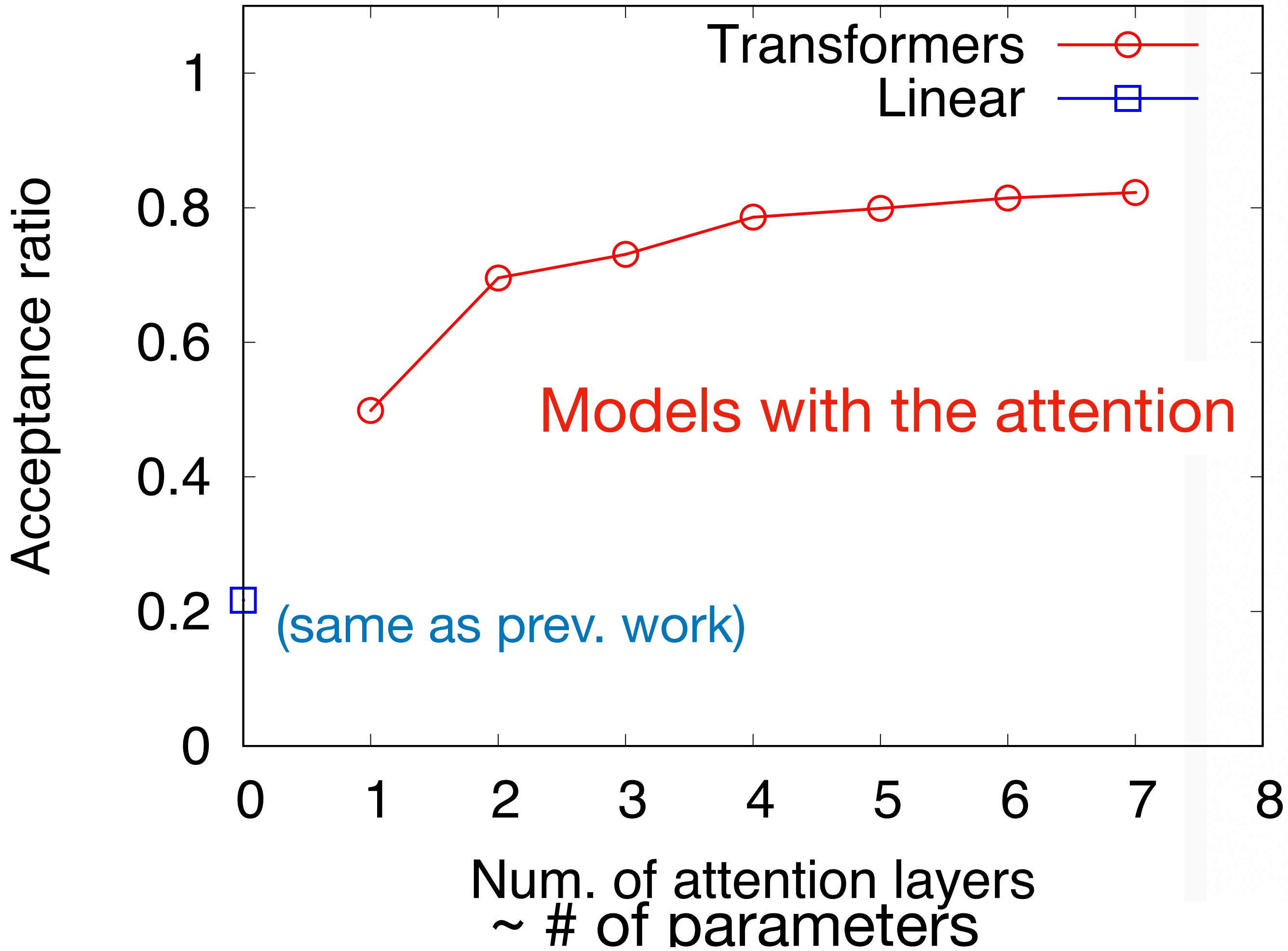
Results

Transformer and Attention

Akio Tomiya
arXiv: 2306.11527 + update

Acceptance rate is improved with # of layers

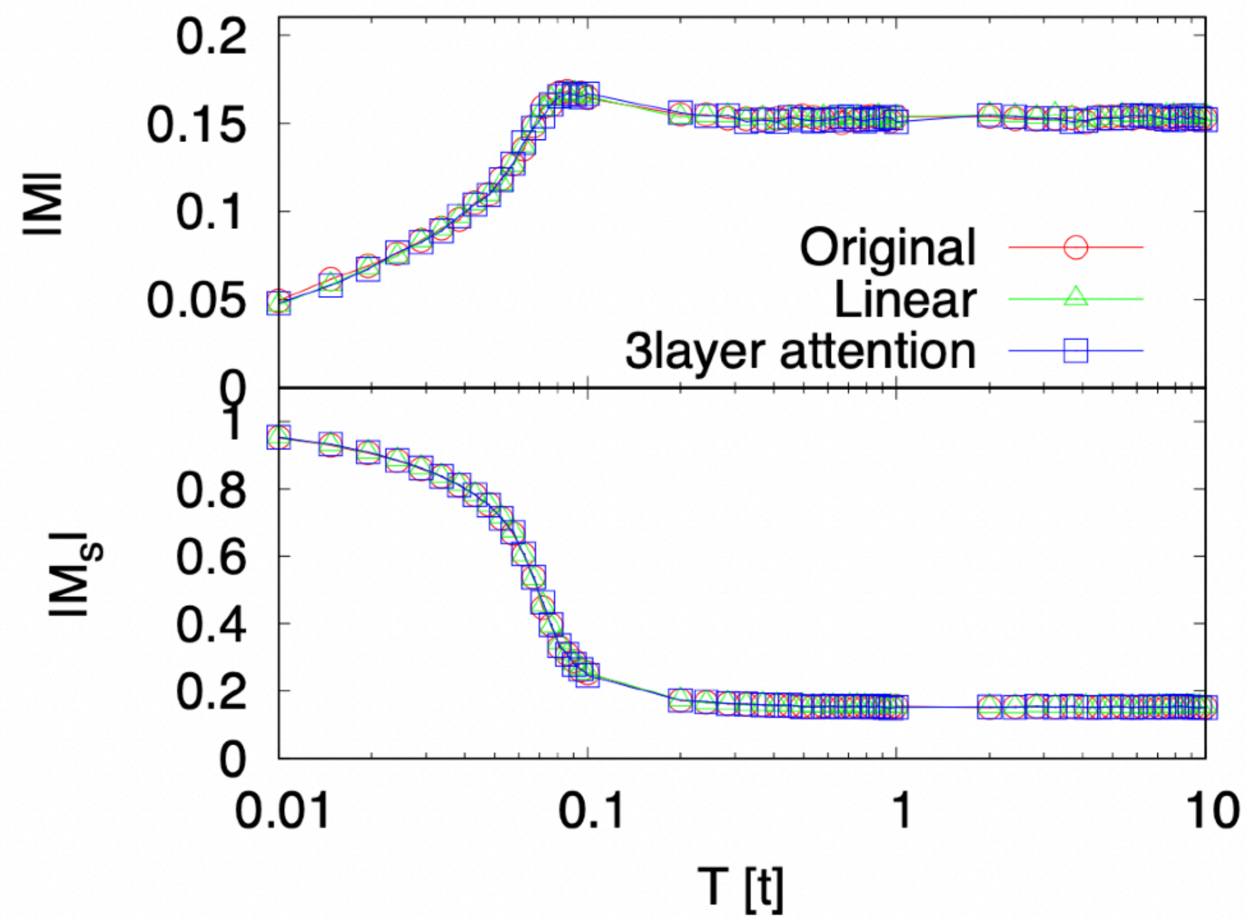
Acceptance rate



Note: As far as we tested,
CNN-type does not work in this case.

No improvements with increase of layers.
(Global correlations of fermions from
Fermi-Dirac statistics make acceptance bad?)

Observables



Physical values are consistent
(as we expected)

Nx=Ny=6
(Lattice sites)

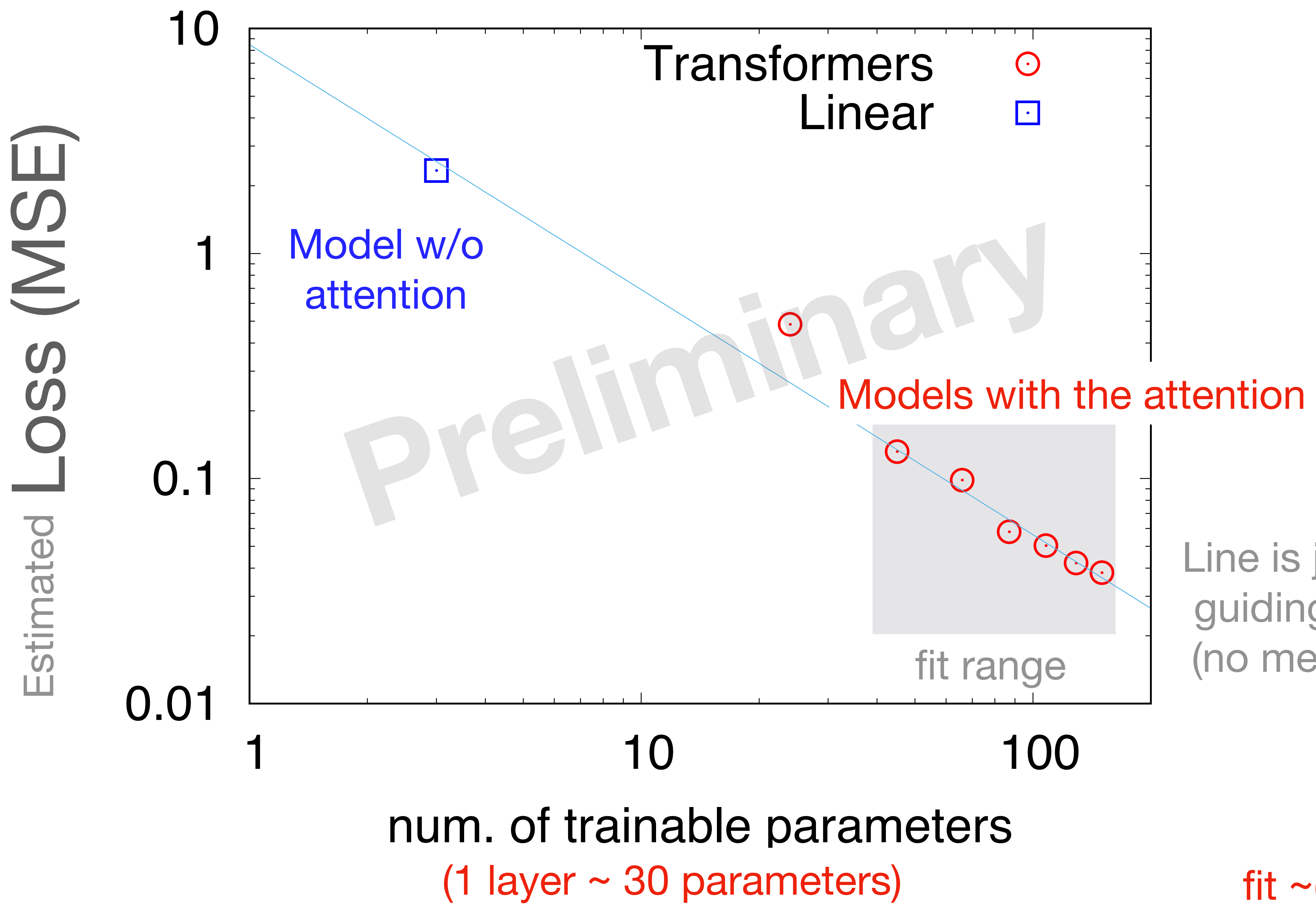
Akio Tomiya

Transformer and Attention

Acceptance rate -> MSE (~ loss), Scaling law (power law)

arXiv: 2306.11527 + update

$$\text{acceptance} = \exp\left(-\sqrt{\text{MSE}}\right)$$



Line is just for guiding eyes (no meaning)

fit ~(7.1/x)^(1.1)

Summary

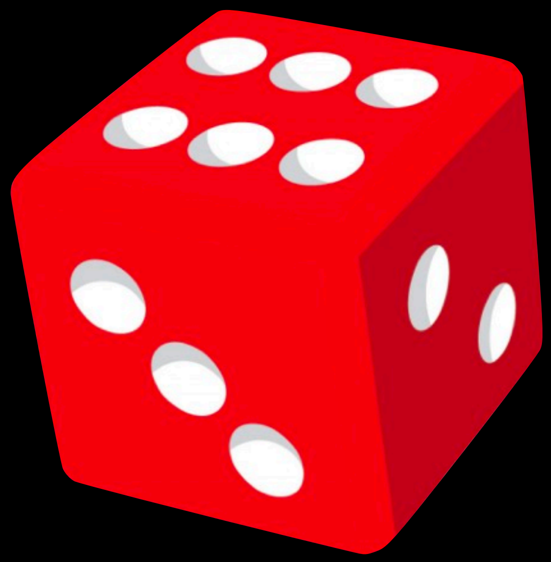
Physics + Machine learning

arXiv: 2306.11527

- Equivariance helps generalization of machine learning models
  Attention enables us to capture global correlations
- O(3) spin-fermion system can be efficiently simulated SLMC with Attention
  - In lattice QCD terminology, it is O(3) scalar + fermions
  - Increase of #of attention layers makes increase acceptance rate
  - Models with the CNN-type do not work (not showed)
  - SLMC with the equivariant Attention shows the scaling law
- Attention is all you need (?)
- Future work:
  - Apply ``equivariant attention” on full QCD
  - What is ``gauge equivariant attention”? Is it possible?
  - Can we marge it with gauge covariant convolution? (arXiv: 2103.11965)
  - Can we use this to the flow based sampling algorithm? (GomalizingFlow.jl)

MLPhys Foundation of "Machine Learning Physics" Grant-in-Aid for Transformative Research Areas (A)
Program for Promoting Researches on the Supercomputer Fugaku Large-scale lattice QCD simulation and development of AI technology
KAKENHI: 20K14479, 22H05112, 22H05111, 22K03539

Thanks!



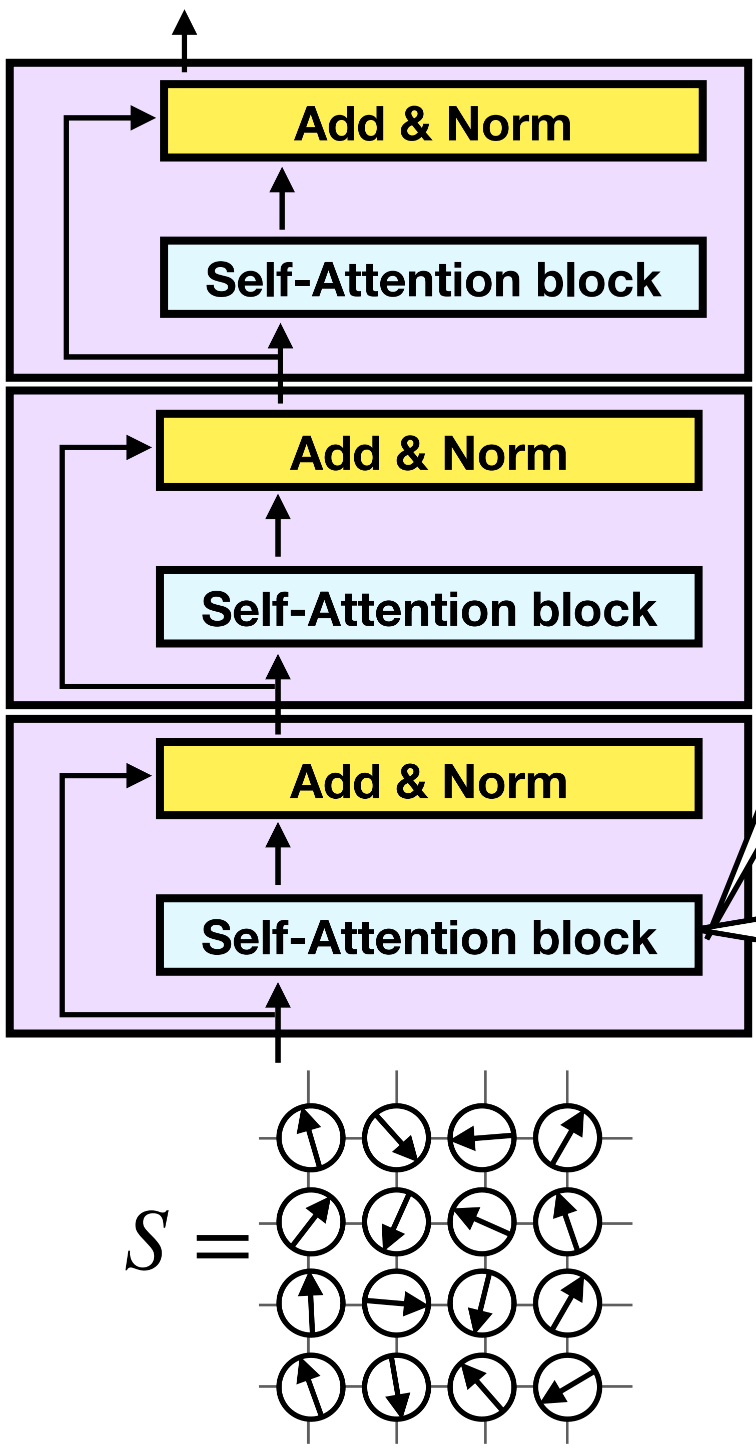
Details

Self-learning Monte-Carlo

Attention layer

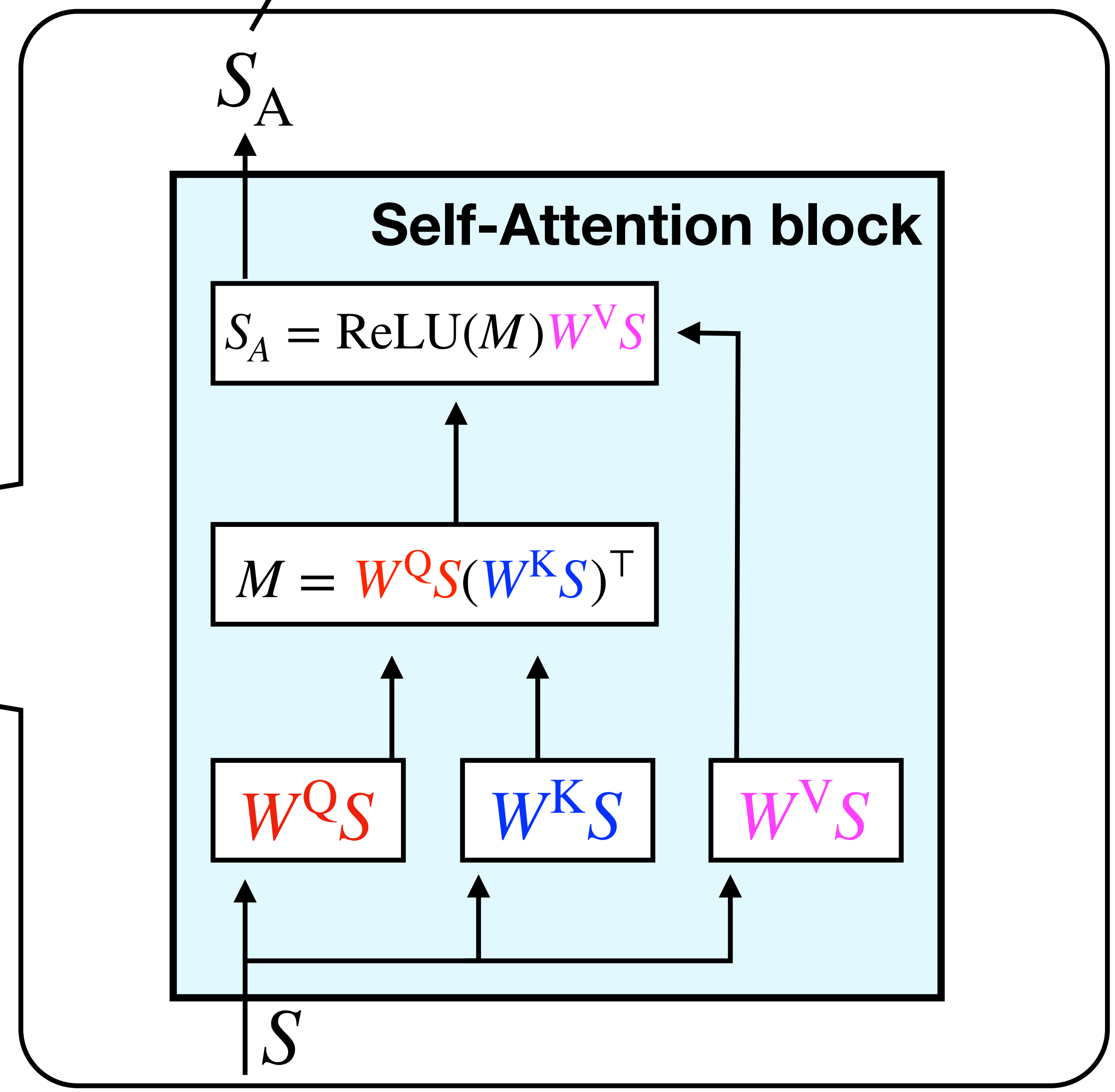
arXiv: 2306.11527.

$$S' \rightarrow H_{\text{eff}} = \text{tr}[S'(JS')^{\top}]$$



$$\mathbf{S}^{(l)} \equiv \mathcal{N}\left(\mathbf{S}^{(l-1)} + \underline{\text{SelfAttention}^{\text{spin}}_{\theta^{(l)}}}(\mathbf{S}^{(l-1)})\right)$$

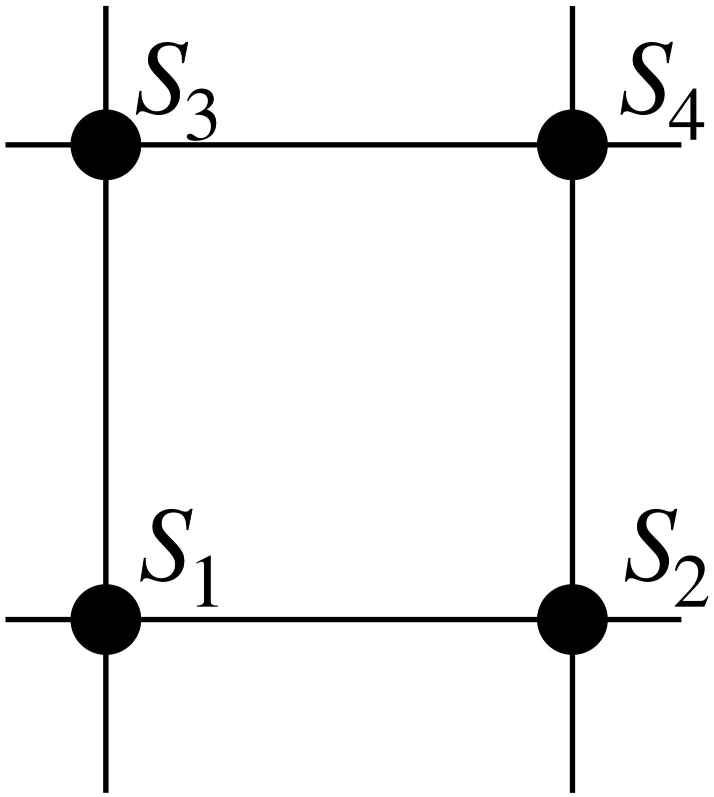
$$\mathcal{N}(\mathbf{S}_i) = \frac{\mathbf{S}_i}{\|\mathbf{S}_i\|}$$



Self-learning Monte-Carlo

Equivariant under spin-rotation & translation

arXiv: 2306.11527.



$$S_i^\top = \begin{pmatrix} s_i^1 & s_i^2 & s_i^3 \end{pmatrix}^\top$$

$$|S_i| = \sqrt{(s_i^1)^2 + (s_i^2)^2 + (s_i^3)^2}$$
$$= 1$$

3 component scalar, normalized

$$\mathbf{S} = \begin{pmatrix} S_1^\top & S_2^\top & S_3^\top & S_4^\top \end{pmatrix}^\top$$

S_i : Classical Heisenberg spin at site i

$\mathbf{S}$: A spin configuration

Gram matrix

$$G \equiv \mathbf{S}^\top \mathbf{S} = \begin{pmatrix} S_1^\top S_1 & S_1^\top S_2 & S_1^\top S_3 & S_1^\top S_4 \\ S_2^\top S_1 & S_2^\top S_2 & S_2^\top S_3 & S_2^\top S_4 \\ S_3^\top S_1 & S_3^\top S_2 & S_3^\top S_3 & S_3^\top S_4 \\ S_4^\top S_1 & S_4^\top S_2 & S_4^\top S_3 & S_4^\top S_4 \end{pmatrix}$$

- G is a matrix for coordinate but not for spin.
- Spin rotation for Si keeps G invariant.

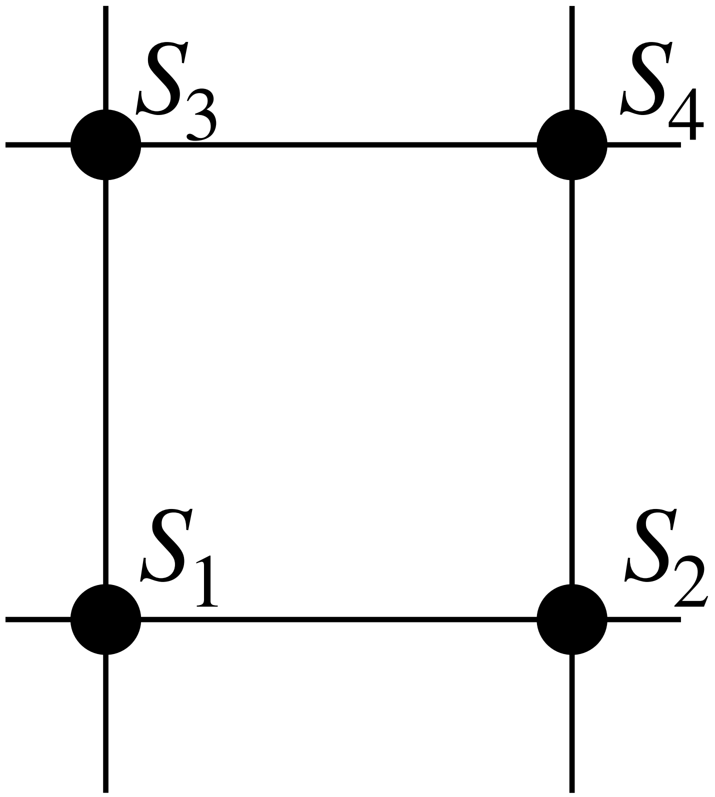
If an effective hamiltonian is a function Gram matrix, it has rotational symmetry

Self-learning Monte-Carlo

Equivariant under spin-rotation & translation

Akio Tomiya

arXiv: 2306.11527.



$$S_i^\top = \begin{pmatrix} s_i^1 & s_i^2 & s_i^3 \end{pmatrix}^\top$$

$$|S_i| = \sqrt{(s_i^1)^2 + (s_i^2)^2 + (s_i^3)^2}$$
$$= 1$$

3 component scalar, normalized

$$\mathbf{S} = \begin{pmatrix} S_1^\top & S_2^\top & S_3^\top & S_4^\top \end{pmatrix}^\top$$

- Local weighted sum over neighbors
= "Smeared spin" with parameters
~ "Block spin sum" with parameters

$$\tilde{S}_i^\alpha = \sum_{l=0} W_l^\alpha S_{i+l} \qquad \alpha = \mathrm{Q, K, V}$$

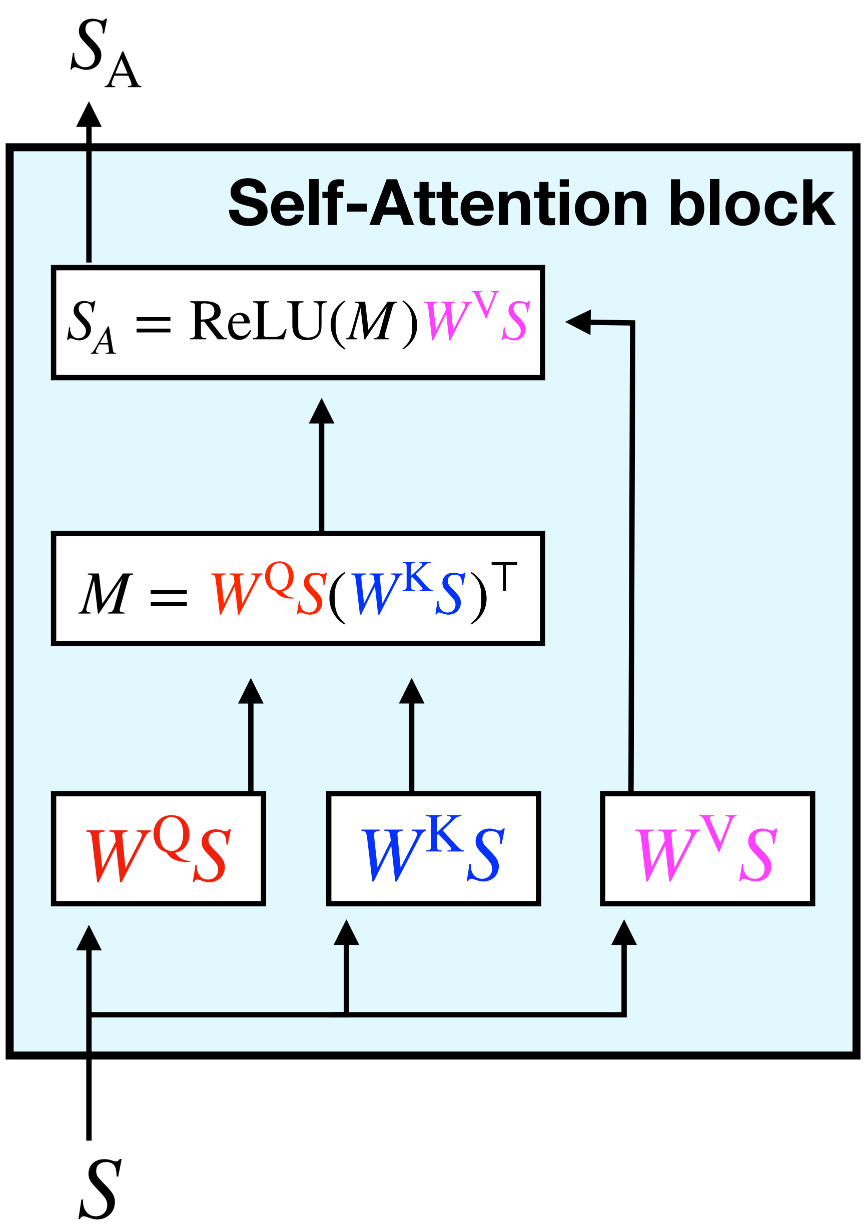
$W_l^\alpha \in \mathbb{R}$: trainable

Translationally equivariant
Rotatinally equivariant

Self-learning Monte-Carlo

Equivariant under spin-rotation & translation

arXiv: 2306.11527.



$$\mathbf{S} = \begin{pmatrix} S_1^{\top} & S_2^{\top} & S_3^{\top} & S_4^{\top} \end{pmatrix}^{\top}$$

$$S_i^{\top} = \begin{pmatrix} s_i^1 & s_i^2 & s_i^3 \end{pmatrix}^{\top}$$

$$\tilde{S}_i^{\alpha} = \sum W_l^{\alpha} S_{i+l}$$ **"Smeared spin"**

Gram matrix with smeared spin

$$M = \tilde{G}^{\alpha} \equiv (\tilde{\mathbf{S}}^{\alpha})^{\top} \tilde{\mathbf{S}}^{\alpha} \quad \alpha = \mathrm{Q}, \mathrm{K}, \mathrm{V}$$

Translationally covariant
Rotatinally invariant

$$S_A = \mathrm{ReLU}(M) W^{\mathrm{V}} S$$
$$= \mathrm{ReLU}(M) \tilde{S}^{\mathrm{V}}$$

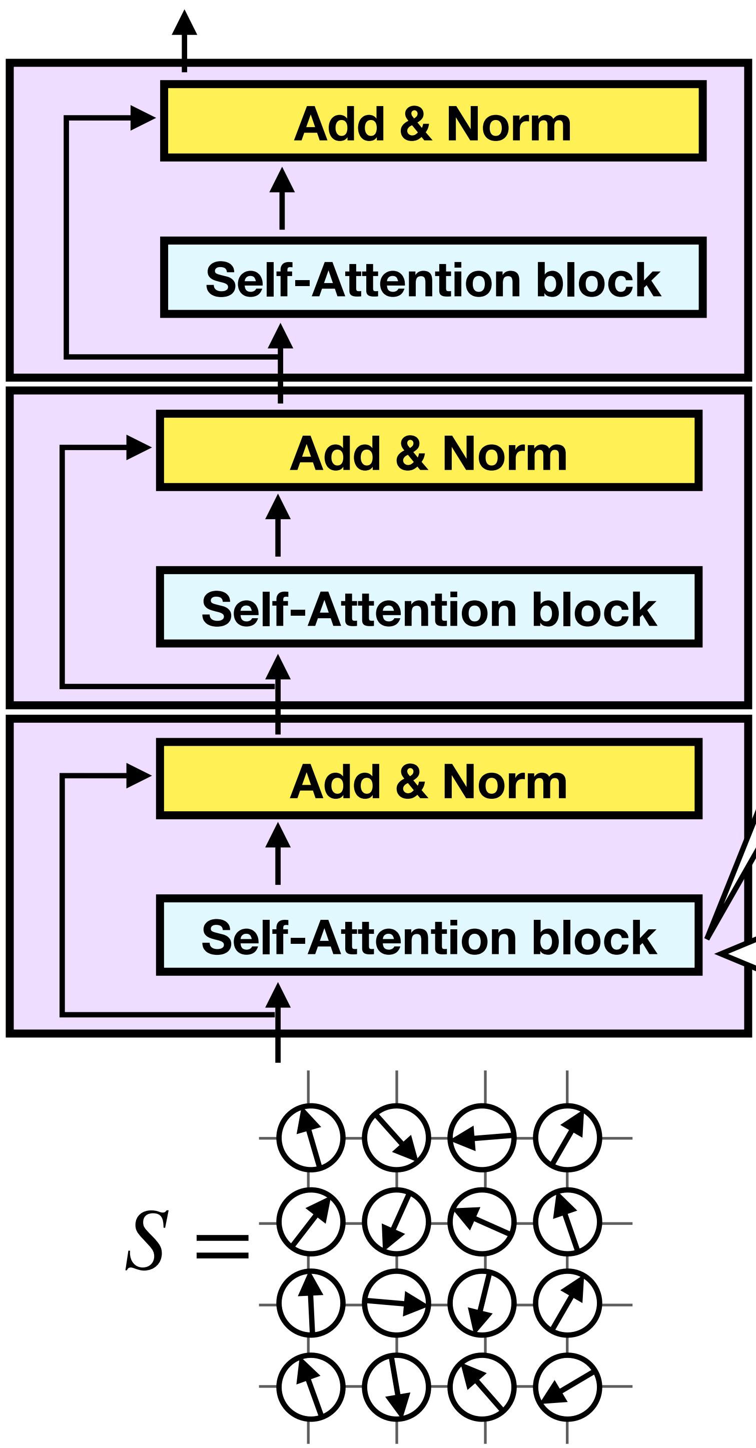
Self-learning Monte-Carlo

Akio Tomiya

Attention layer

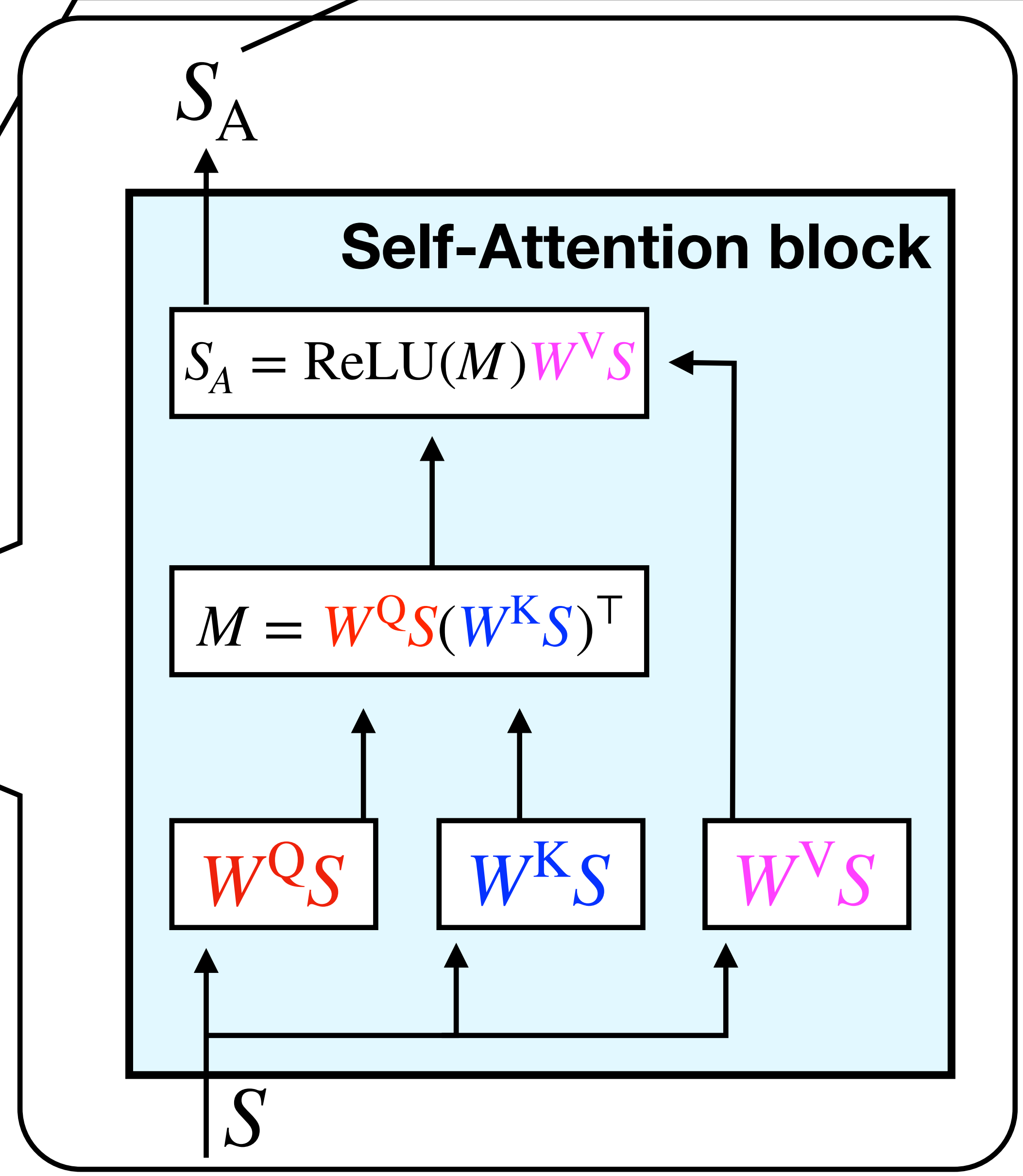
arXiv: 2306.11527.

$$S' \to H_{\text{eff}} = \mathrm{tr}[S'(JS')^\top]$$



$$\mathbf{S}^{(l)} \equiv \mathcal{N}\left(\mathbf{S}^{(l-1)} + \underline{\mathrm{SelfAttention}^{\mathrm{spin}}_{\theta^{(l)}}}(\mathbf{S}^{(l-1)})\right)$$

$$\mathcal{N}(\mathbf{S}_i) = \frac{\mathbf{S}_i}{\|\mathbf{S}_i\|}$$



Note that,
If W = 0,
it fall back to
identity operation

Effective Hamiltonian
fall back to
a linear model

Transformer and Attention

Training scheme

Akio Tomiya

arXiv: 2306.11527.