What do Transformers learn when trained via Masked Language Modelling?

An analysis in the framework of the Generalized Potts model

Federica Gerace

Joint work with:

Riccardo Rende

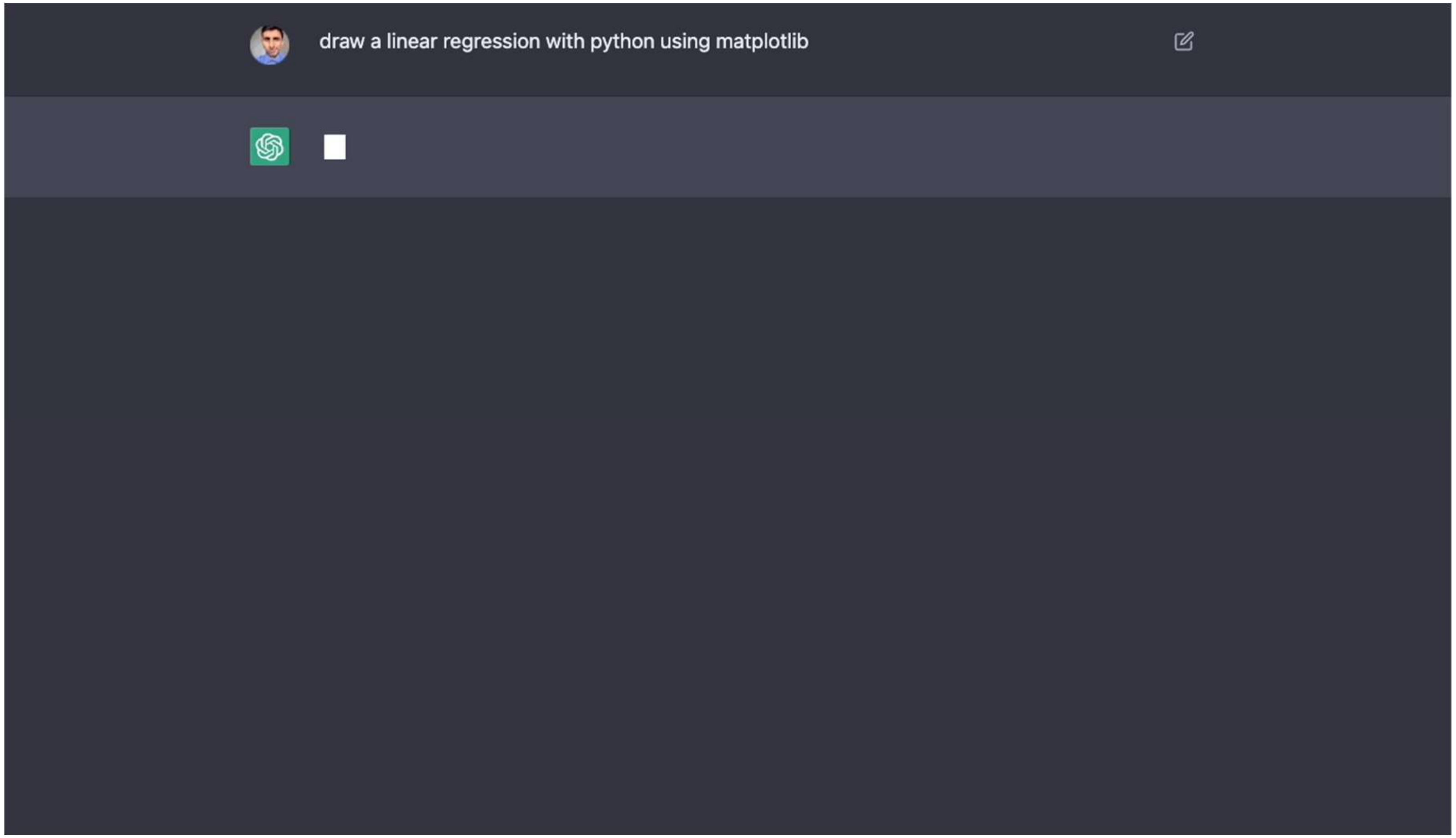
Alessandro Laio

Sebastian Goldt

SISSA

Bridging scales: At the crossroads among renormalisation group, multi-scale modelling, and deep learning - 16/04/2024

A famous Transformer example: Chat-GPT!



Artificial Neural Networks in short

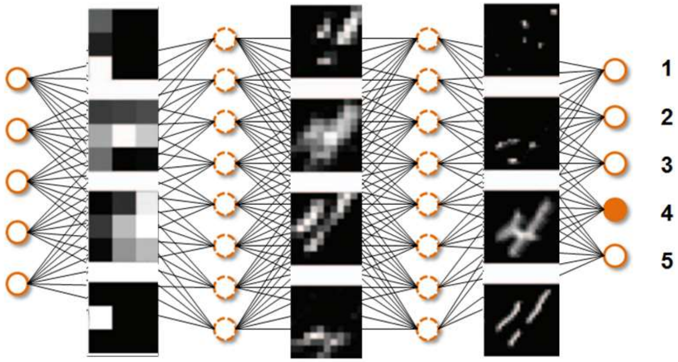
- **Training Phase:**

Training Set

$$\mathcal{D} = \{\boldsymbol{x}^{\mu}, y^{\mu}\}_{\mu=1}^{n}$$

$\boldsymbol{x}^{\mu} =$ 

$y^{\mu} = 4$



$= \hat{y}^{\mu}$

$$\boldsymbol{w}^{\mathrm{t+1,l}} \leftarrow \boldsymbol{w}^{\mathrm{t,l}} + \eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{w}^{\mathrm{t,l}}}$$

Loss Function

$$\mathcal{L} = \sum_{\mu=1}^{n} \ell\left(y^{\mu}, \hat{y}^{\mu}(\boldsymbol{w}; \boldsymbol{x}^{\mu})\right)$$

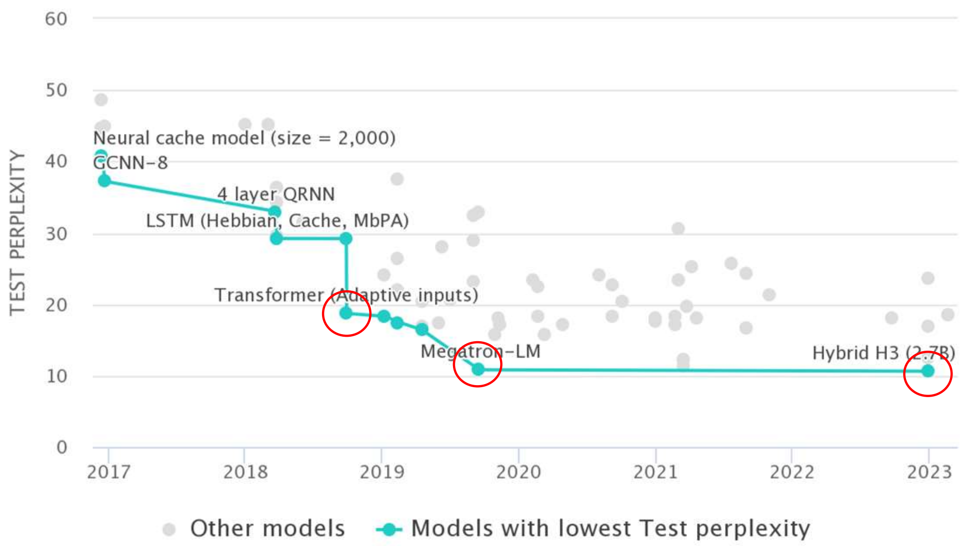
$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{w}|\mathcal{D})$$

- **Testing Phase:**

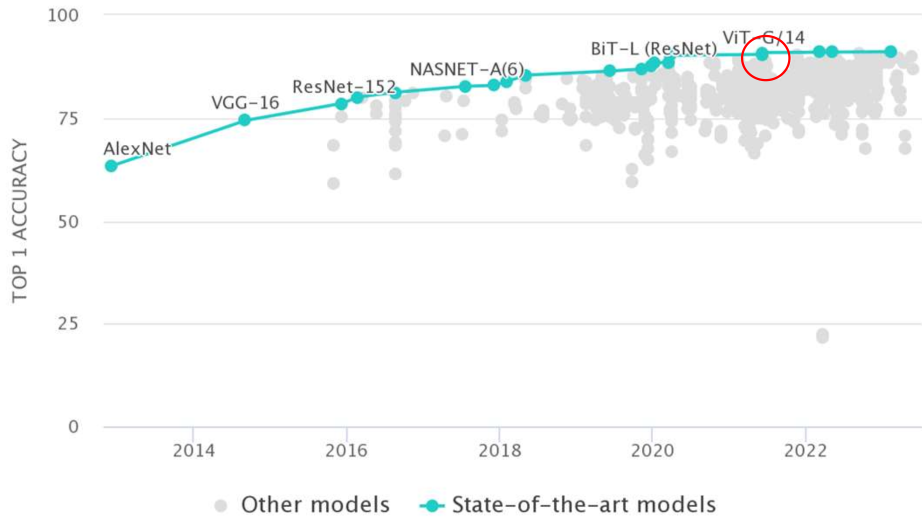
$$\epsilon_g = \mathbb{E}_{\{x_{new}, y_{new}\}}\left[\left(y^{new} - \hat{y}^{new}(\widehat{\boldsymbol{w}}; \boldsymbol{x}^{new})\right)^2\right]$$

Transformers are taking the show

- Language Modelling on WikiText-103



- Image Classification on ImageNet



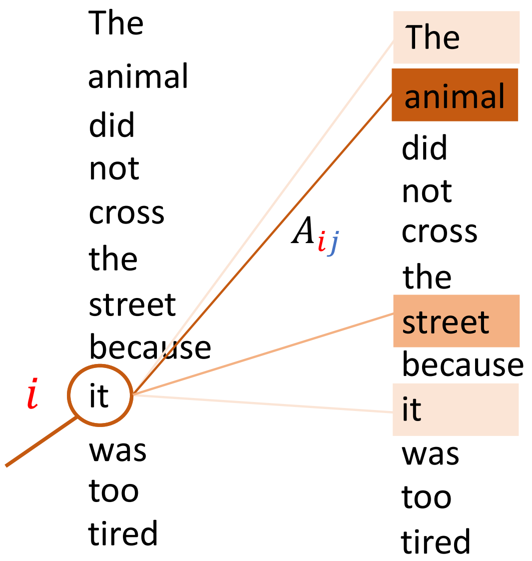
From https://paperswithcode.com/

Transformers learn the context

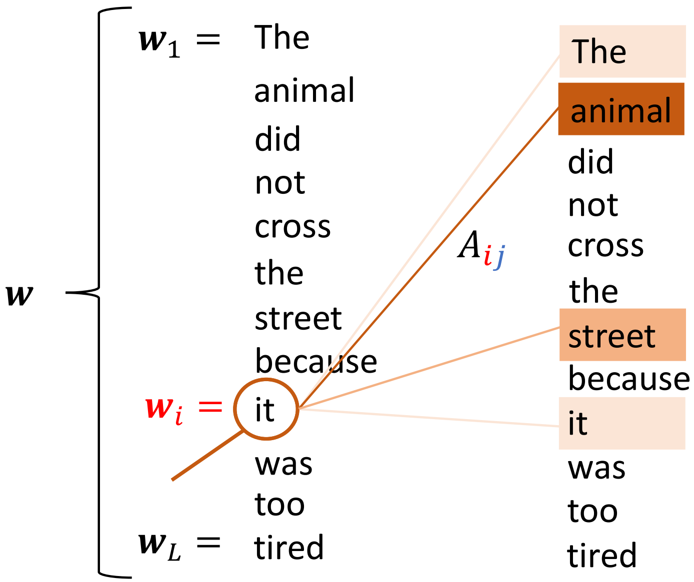
The animal did not cross the street because it was too tired.

The animal did not cross the street because it was too wide.

Self-Attention Mechanism



Self-Attention Mechanism



word | token | embedding | positional encoding

$$w_j \to t_j \to x_j = Et_j + p_j \in \mathbb{R}^D$$

$$A_{ij} = K(x_i, x_j) = softmax(x_i^T Q^T K x_j)$$

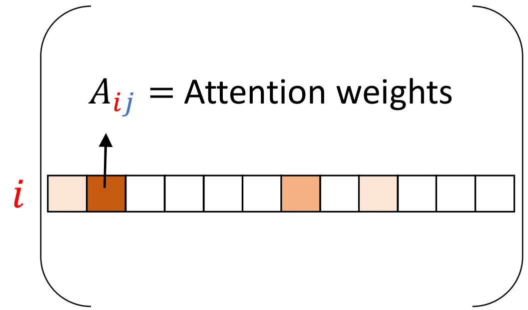
$$\mathbf{k}_j = K x_j \in \mathbb{R}^D \quad \text{key}$$

$$\mathbf{q}_i = Q x_i \in \mathbb{R}^D \quad \text{query}$$

$$x_i \longrightarrow \mathbf{h}_i = \sum_{j=1}^{L} A_{ij} V x_j$$

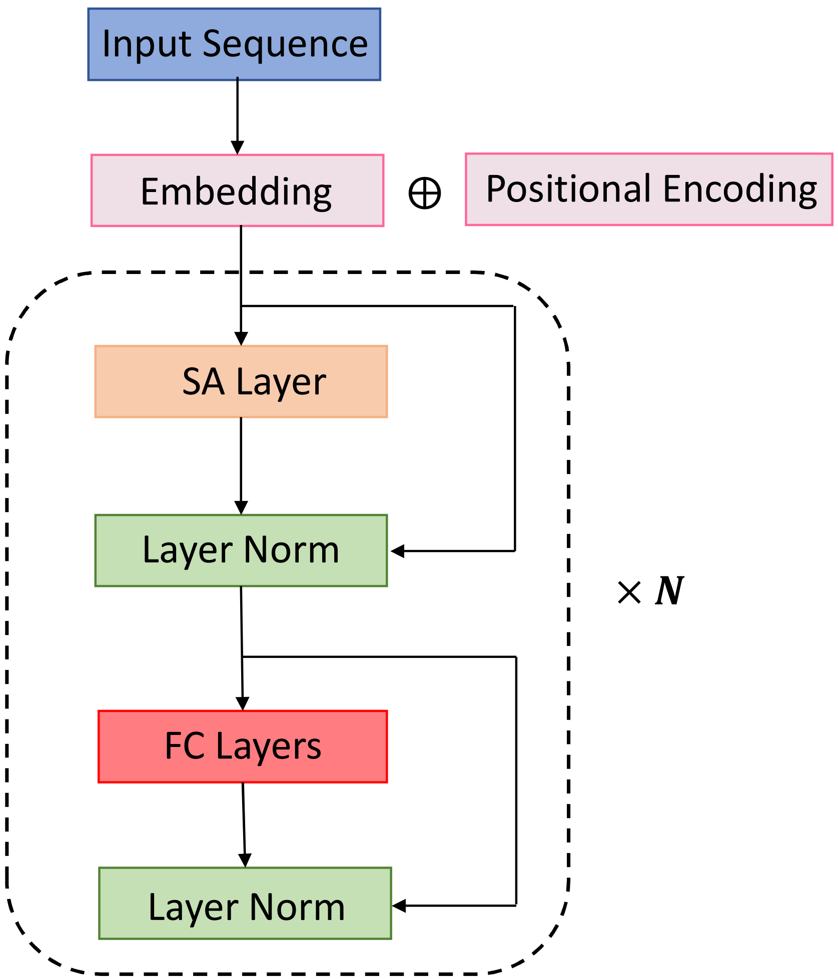
$$\mathbf{v}_j = V x_j \in \mathbb{R}^D \quad \text{value}$$

$A =$



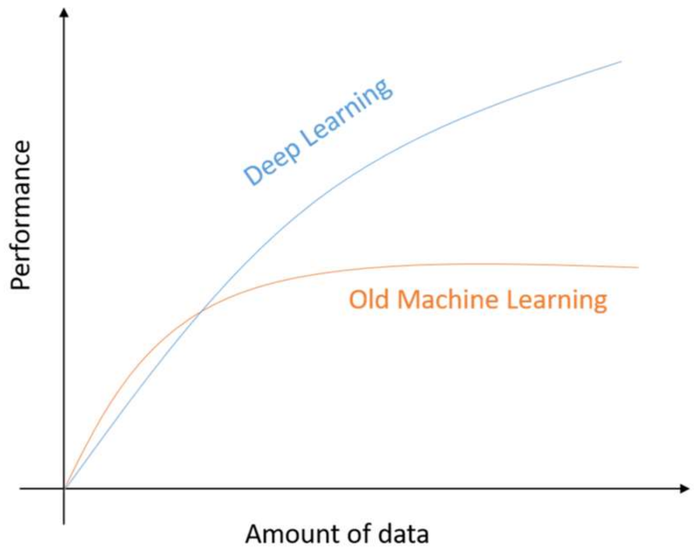
$\in \mathbb{R}^{L \times L}$ = **Attention Map**

The building block of Transformers



Deep learning is data-hungry

> *"The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data we can feed to these algorithms."*
>
> Andrew Ng



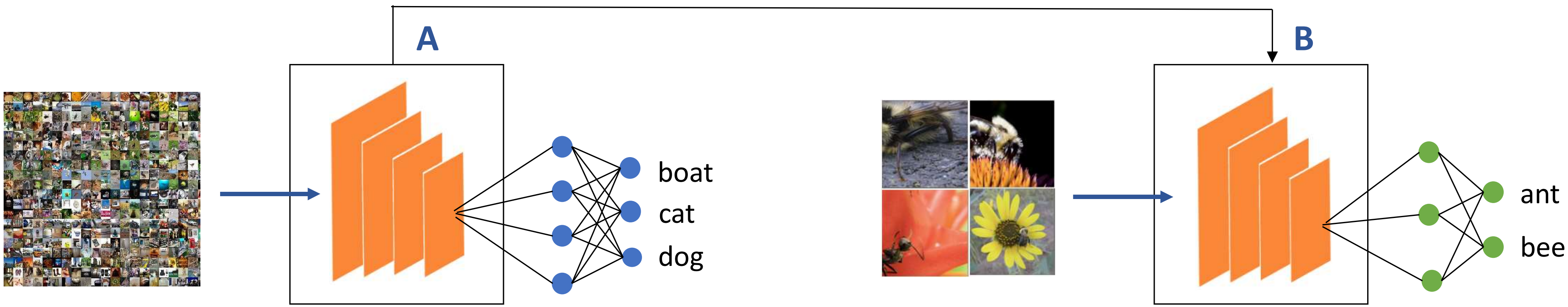
Alom et al, *A State-of-the-Art Survey on Deep Learning Theory and Architectures*, Electronics, 2019

A possible solution: Transfer Learning

DATA-ABUNDANT SOURCE TASK

DATA-SCARCE TARGET TASK

Feature map transfer



Gerace, Sarao Mannelli, Saglietti, Saxe, Zdeborová, *Machine Learning: Science and Technology, 2022*

Gerace, Doimo, Sarao Mannelli, Saglietti, Laio, *arXiv:2303.01429, 2023*

A possible solution: Transfer Learning

DATA-ABUNDANT SOURCE TASK

DATA-SCARCE TARGET TASK

- Masked Language Modelling (MLM)

Input: $\boldsymbol{x}^{\mu} =$ *We* $\boxed{\text{MASKED}}$ *milk chocolate* $\quad \mu = 1, \dots, M$

Label: $\boldsymbol{y}^{\mu} =$ *eat*

Goal: $\boxed{\text{MASKED}} \; = ?$

- Text Translation;
- Text Generation (Chat-GPT);
- Sentiment Analysis…

Some open questions

🤔 What self-attention learns with Masked Language Modelling?

🤔 How many samples are required to achieve good generalization performances?

A key ingredient: The Generalized Potts Model!

Rende, **Gerace**, Laio, Goldt, *Physical Review Research, 2024*

The Generalized Potts Model

$\boldsymbol{w} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_i, \dots, \boldsymbol{w}_L\}$ ⟶ $\boldsymbol{s} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_i, \dots, \boldsymbol{s}_L\}$

word:

$\boldsymbol{w}_i = street$

one-hot encoded Potts spin:

$\boldsymbol{s}_i = [s_{i1}, \dots, s_{ic}, \dots, s_{iC}] \in \{0,1\}^C$

C = size of the vocabulary

Each sequence is sampled from the Gibbs Measure of the Generalized Potts Model:

$$\mathcal{P}(\boldsymbol{s}) \propto \exp(-\beta \mathcal{H}(\boldsymbol{s}))$$

$$\mathcal{H}(\boldsymbol{s}) = \frac{1}{2} \sum_{i,j} J_{ij} \, \boldsymbol{s}_i^T \, U \, \boldsymbol{s}_j$$

$U \in \mathbb{R}^{C \times C}$ = Interaction among colors.

$J \in \{0,1\}^{L \times L}$ = Interaction among sites;

MLM with the Generalized Potts Model

Given a training set made of M *masked sequences* $\boldsymbol{s}^{\mu}_{\setminus i}$ with the corresponding *masked spin value* $\boldsymbol{s}^{\mu}_{i}$, e.g.:

$$\mathcal{D} = \left\{\boldsymbol{s}^{\mu}_{\setminus i}, \boldsymbol{s}^{\mu}_{i}\right\}_{\mu=1}^{M}$$

with each sequence $\boldsymbol{s}^{\mu}$ sampled from the *Generalized Potts Model*, e.g.: $\boldsymbol{s}^{\mu} \backsim \mathcal{P}(\boldsymbol{s}^{\mu}; J, U)$.

The *goal* is to achieve the lowest possible generalization loss, e.g.:

$$\epsilon_g = -\mathbb{E}_{\boldsymbol{s} \sim \underbrace{\mathcal{P}(\boldsymbol{s}^{new}; J, U)}_{\text{True Gibbs Measure}}} \left[\frac{1}{L} \sum_{i=1}^{L} \sum_{c=1}^{C} \overbrace{s_{ic}^{new}}^{\text{Exact masked spin value}} \log \Big(\underbrace{\hat{\mathcal{P}}_{ic}\,(\boldsymbol{s}^{new}; A, V)}_{\text{Transformer Gibbs Measure prediction}} \Big) \right]$$

Question 1

🤔 What self-attention learns with Masked Language Modelling?

Vanilla Transformer on Generalized Potts

Task:

$$\mathcal{D} = \left\{ \boldsymbol{s}^{\mu}_{\setminus i}, \boldsymbol{s}^{\mu}_{i} \right\}_{\mu=1}^{M}$$

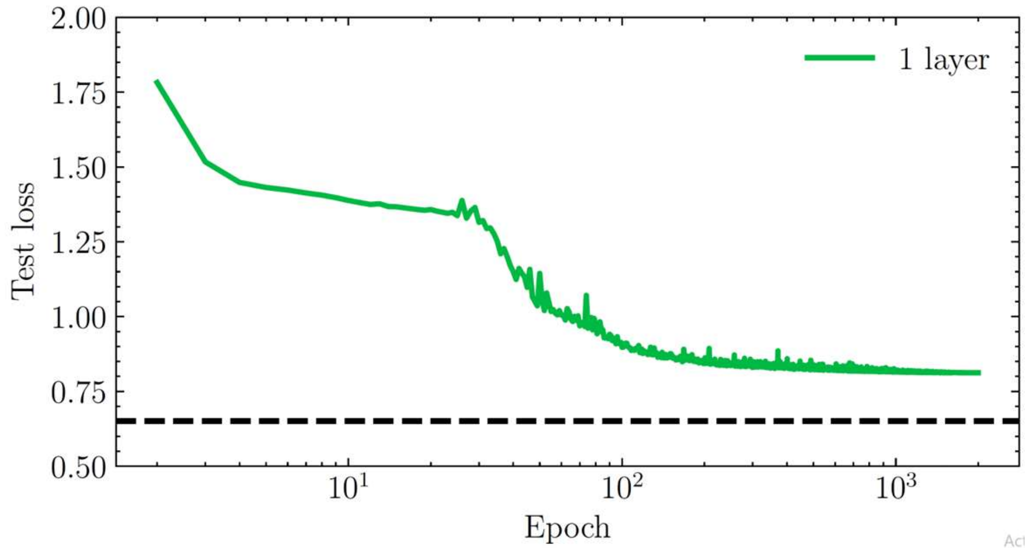
Test Loss:

Exact masked spin value

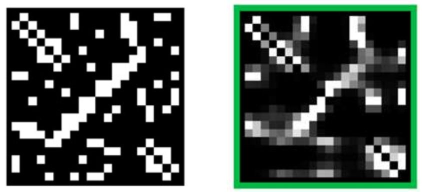
$$\epsilon_g = -\mathbb{E}_{\boldsymbol{s} \sim \mathcal{P}(\boldsymbol{s};J,U)} \left[\frac{1}{L} \sum_{i=1}^{L} \sum_{c=1}^{C} s_{ic} \log \left(\hat{\mathcal{P}}_{ic} \left(\boldsymbol{s}; A, V \right) \right) \right]$$

Transformer Gibbs Measure estimate

True Gibbs Measure



Attention Map



J $\quad$ $\hat{J}_{1-\text{layer}}$

Rende, **Gerace**, Laio, Goldt, *Physical Review Research, 2024*

Vanilla Transformer on Generalized Potts

Task:

$$\mathcal{D} = \left\{ \boldsymbol{s}^{\mu}_{\setminus i}, \boldsymbol{s}^{\mu}_{i} \right\}_{\mu=1}^{M}$$

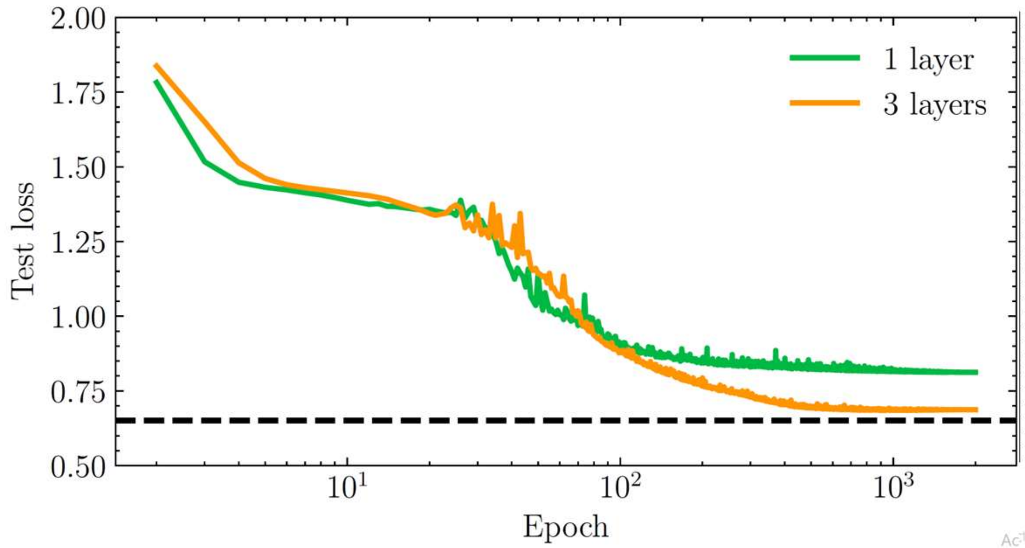
Test Loss:

Exact masked spin value

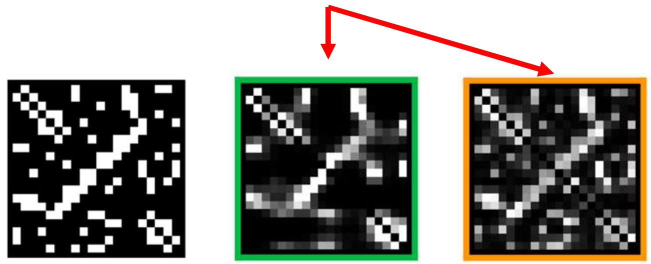
$$\epsilon_g = -\mathbb{E}_{\boldsymbol{s} \sim \mathcal{P}(\boldsymbol{s};J,U)} \left[\frac{1}{L} \sum_{i=1}^{L} \sum_{c=1}^{C} s_{ic} \log \left(\hat{\mathcal{P}}_{ic} \left(\boldsymbol{s}; A, V \right) \right) \right]$$

Transformer Gibbs Measure estimate

True Gibbs Measure



Attention Map



J $\quad$ $\hat{J}_{1-\text{layer}}$ $\quad$ $\hat{J}_{3-\text{layers}}$

Rende, **Gerace**, Laio, Goldt, *Physical Review Research, 2024*

Factored Self-Attention learns Generalized Potts

Input Sequence

$$\boldsymbol{s}_{\setminus i} = \{\boldsymbol{s}_1, \dots, ?, \dots, \boldsymbol{s}_L\}$$

Embedding $\oplus$ P. E.

$$\boldsymbol{x}_j = \boxed{\boldsymbol{s}_j} + \boldsymbol{p}_j \in \mathbb{R}^C$$

Factored Attention Layer

$$\mathbf{h}_i = \sum_{j=1}^{L} A_{ij} V \boldsymbol{x}_j = \sum_{j=1}^{L} softmax(\boldsymbol{x}_i^T Q^T K \boldsymbol{x}_j) V \boldsymbol{x}_j = \sum_{j=1}^{L} softmax\left((\,?\, + \boldsymbol{p}_i)^T Q^T K(\cancel{\boldsymbol{s}_j} + \boldsymbol{p}_j)\right) V(\boldsymbol{s}_j + \cancel{\boldsymbol{p}_j})$$

Bhattacharya et al., bioRxiv:2020.12.21.423882,2020
Viteritti, Rende, Becca, PRL, 2023

Transformer prediction

$$\hat{s}_{ic} \sim softmax(\boxed{h_{ic}}) \propto \exp\left(\sum_{j=1}^{L} A_{ij} (V \boldsymbol{s}_j)_c\right)$$

$$\boldsymbol{A_{ij}} \dashrightarrow \boldsymbol{\beta J_{ij}}$$

$$\boldsymbol{V} \dashrightarrow \boldsymbol{U}$$

Exact Potts conditionals

$$s_{ic} \sim \exp\left(\beta \sum_{j=1}^{L} J_{ij} (U \boldsymbol{s}_j)_c\right)$$

Rende, **Gerace**, Laio, Goldt, *Physical Review Research, 2024*

Vanilla Transformer on Generalized Potts

Task:

$$\mathcal{D} = \left\{ \boldsymbol{s}^{\mu}_{\setminus i}, \boldsymbol{s}^{\mu}_{i} \right\}_{\mu=1}^{M}$$

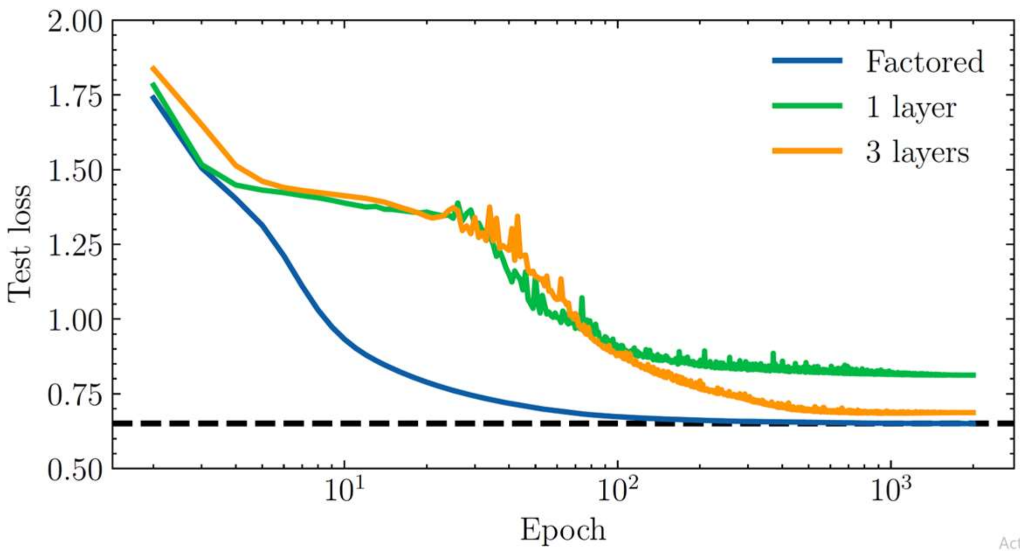
Test Loss:

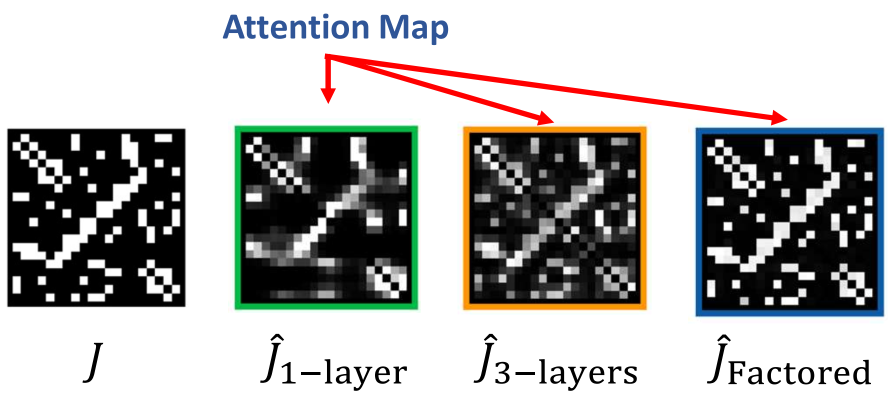
Exact masked spin value

$$\epsilon_g = -\mathbb{E}_{\boldsymbol{s} \sim \mathcal{P}(\boldsymbol{s};J,U)} \left[\frac{1}{L} \sum_{i=1}^{L} \sum_{c=1}^{C} s_{ic} \log\left(\hat{\mathcal{P}}_{ic}\left(\boldsymbol{s}; A, V\right) \right) \right]$$

Transformer Gibbs Measure estimate

True Gibbs Measure

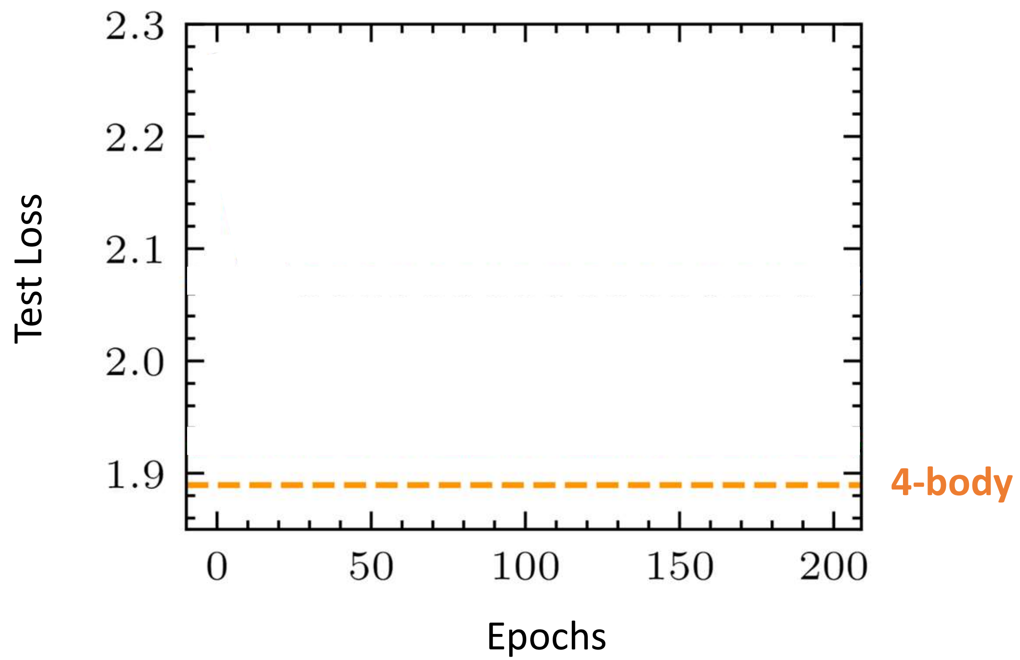




Rende, **Gerace**, Laio, Goldt, *Physical Review Research, 2024*

And if the interactions are of higher-order?

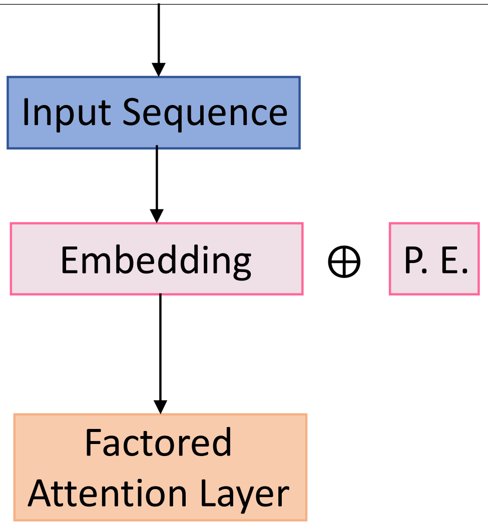
$$\boldsymbol{s} \sim \frac{1}{\mathcal{Z}} \exp\left(-\frac{\beta}{2}\sum_{ijkl}\sum_{\alpha\beta\gamma\delta} J_{ijkl} U_{\alpha\beta\gamma\delta} s_{i\alpha} s_{j\beta} s_{k\gamma} s_{l\delta}\right)$$

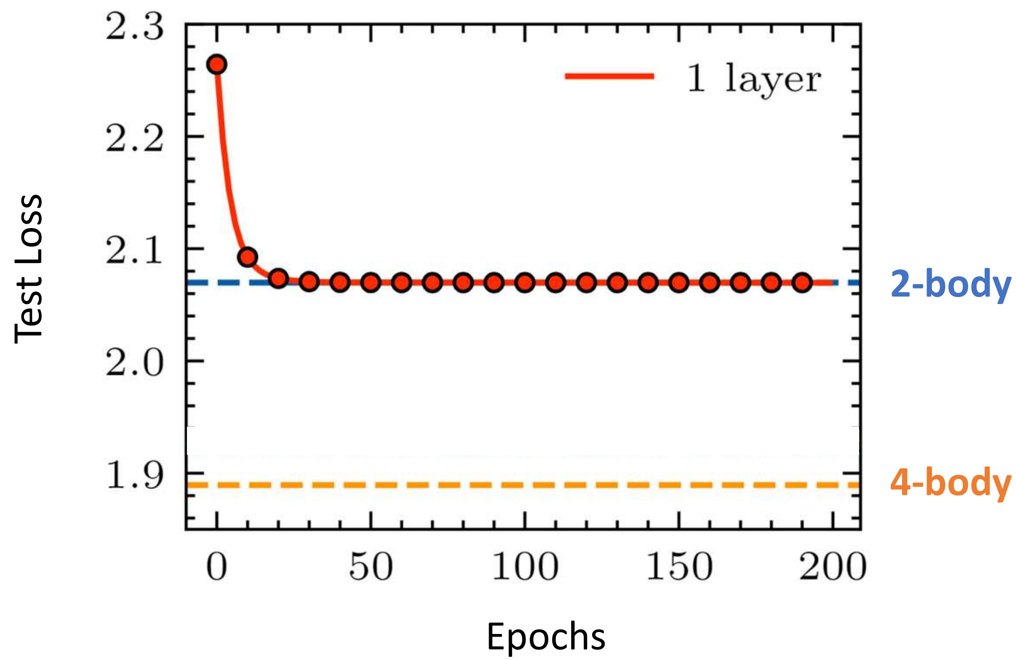


Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

And if the interactions are of higher-order?

$$\boldsymbol{s} \sim \frac{1}{\mathcal{Z}} \exp\left(-\frac{\beta}{2} \sum_{ijkl} \sum_{\alpha\beta\gamma\delta} J_{ijkl} U_{\alpha\beta\gamma\delta} s_{i\alpha} s_{j\beta} s_{k\gamma} s_{l\delta}\right)$$

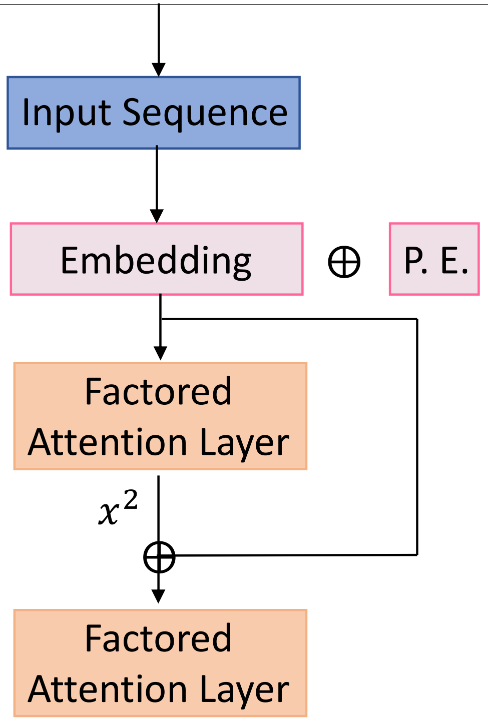


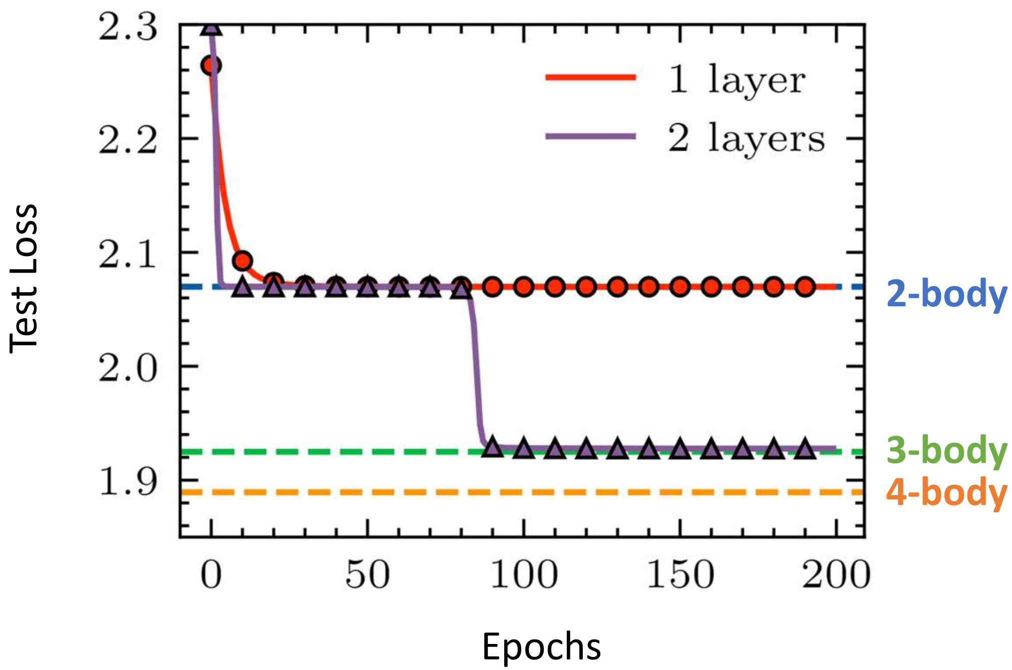


Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

And if the interactions are of higher-order?

$$\boldsymbol{s} \sim \frac{1}{\mathcal{Z}} \exp\left(-\frac{\beta}{2} \sum_{ijkl} \sum_{\alpha\beta\gamma\delta} J_{ijkl} U_{\alpha\beta\gamma\delta} s_{i\alpha} s_{j\beta} s_{k\gamma} s_{l\delta}\right)$$

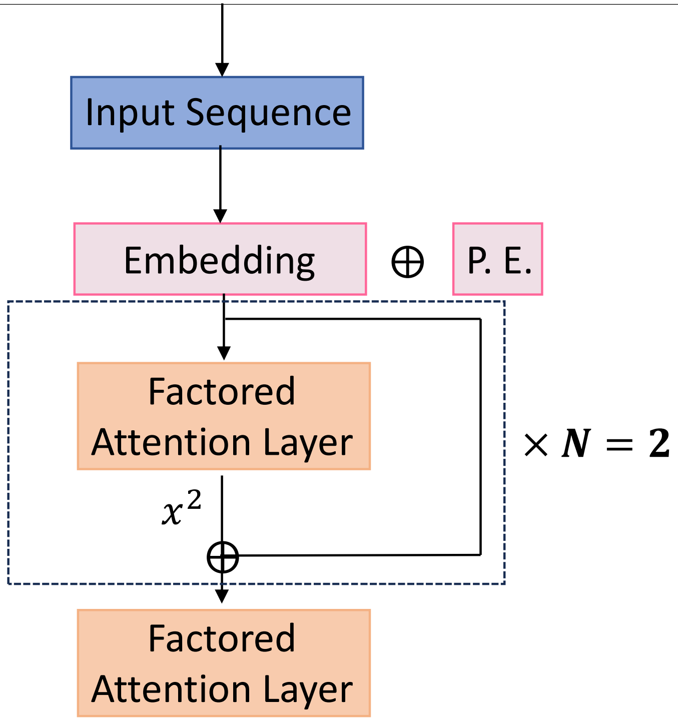


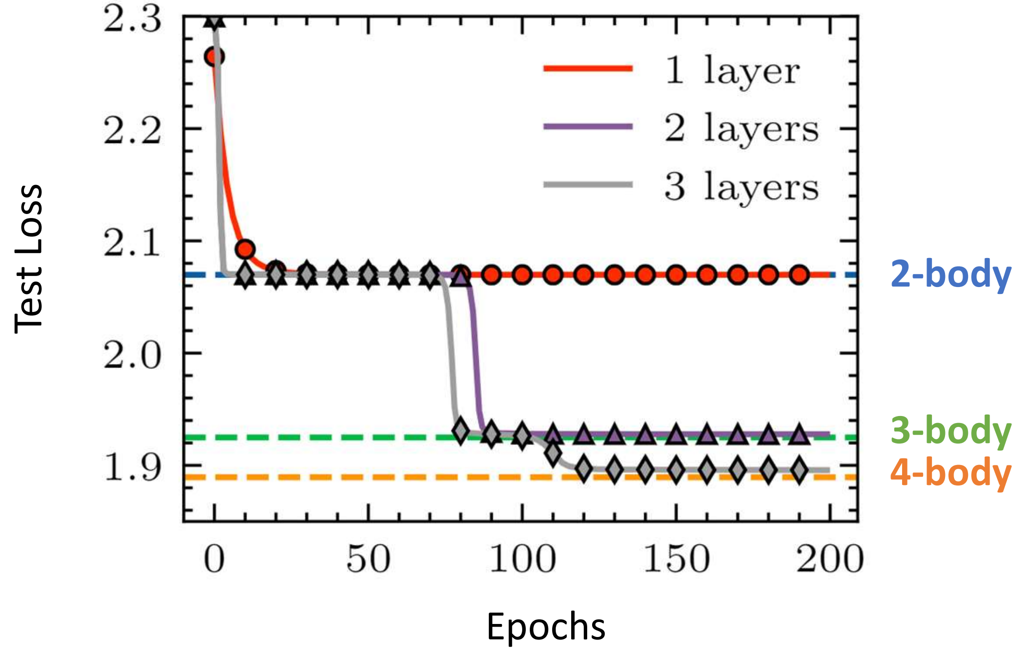


Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

And if the interactions are of higher-order?

$$\boldsymbol{s} \sim \frac{1}{Z}\exp\left(-\frac{\beta}{2}\sum_{ijkl}\sum_{\alpha\beta\gamma\delta} J_{ijkl}U_{\alpha\beta\gamma\delta}s_{i\alpha}s_{j\beta}s_{k\gamma}s_{l\delta}\right)$$

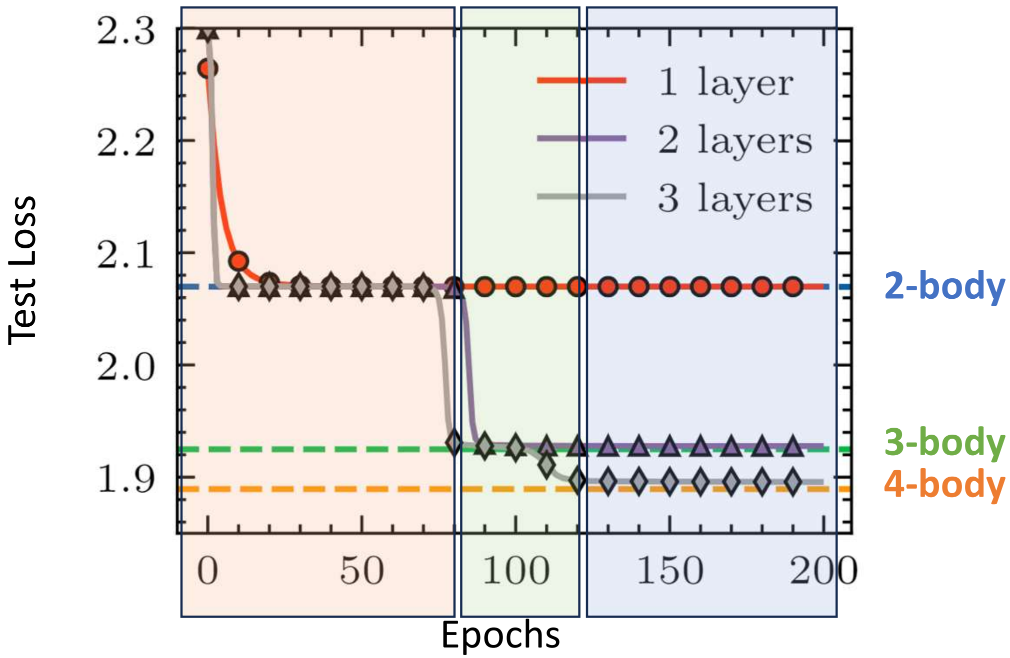


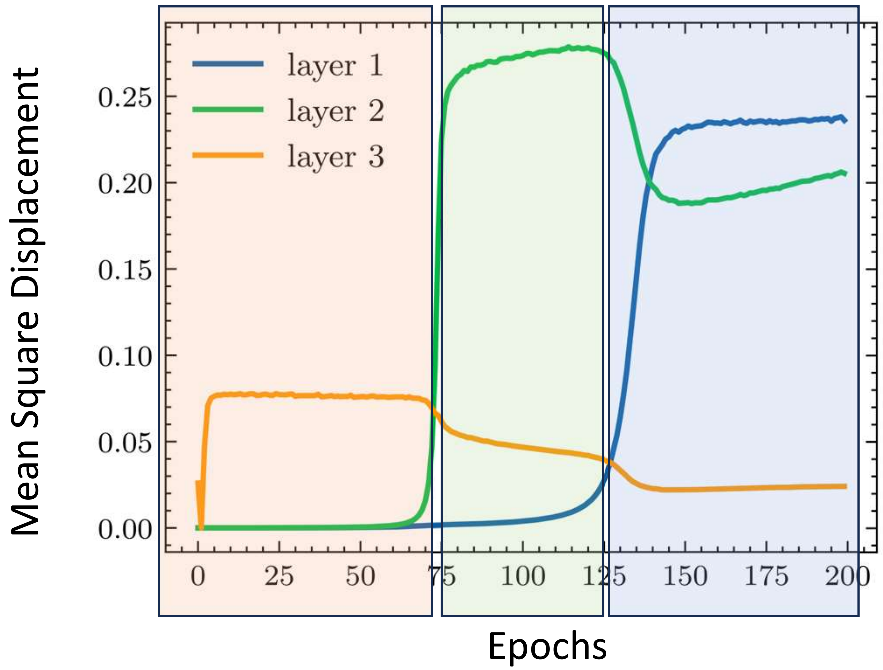


***N* layers learn interactions of order $N + 1$…**

Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

And if the interactions are of higher-order?



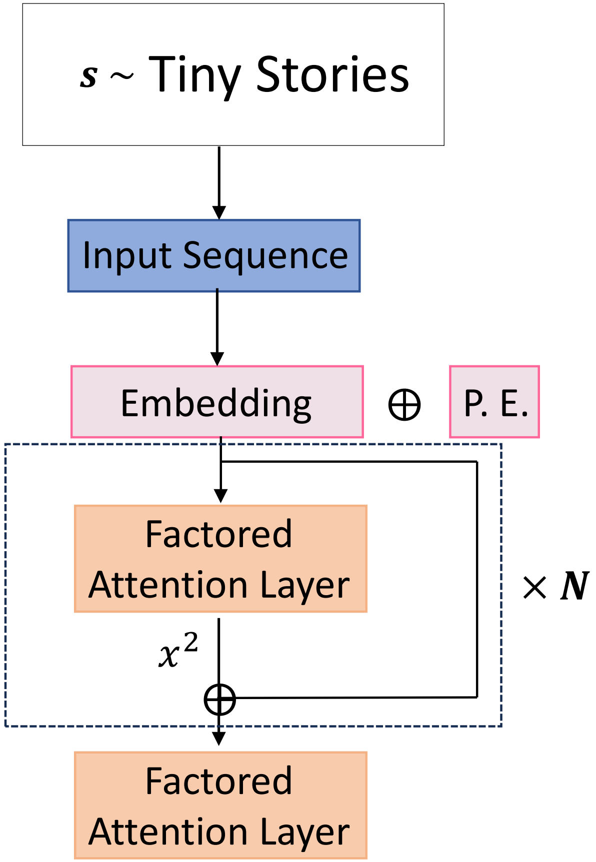


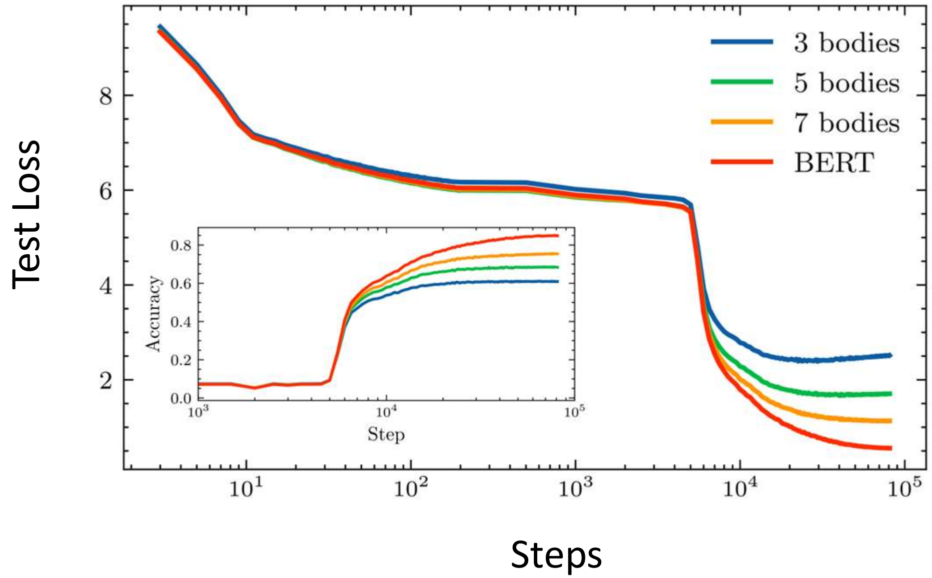
N layers learn interactions of order $N+1$...

$$MSD_l(t) = \frac{1}{L}\|A_l(t) - A_l(0)\|^2 + \frac{1}{C}\|V_l(t) - V_l(0)\|^2$$

Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

And if the interactions are of higher-order?





…and in real NLP datasets!

Rende, **Gerace**, Laio, Goldt, *soon on the ArXiv!*

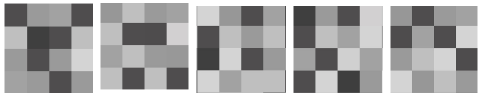
Question 2

🤔 How many samples are required to achieve good generalization performances?

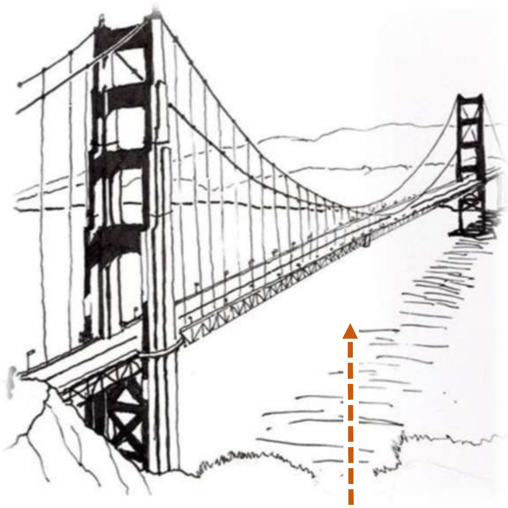
On the way to fill the gap

Machine Learning: $\boldsymbol{x} \sim\ ?$🤔

Statistical Physics: $\boldsymbol{x} \sim \mathcal{N}(0, \mathbb{I})$



Gardner, Derrida, *Journal of Physics A, 1988…*



Up to now: $\boldsymbol{x} \sim \sum_{k=1}^{K} p_k\, \mathcal{N}(\boldsymbol{\mu}_k, \Omega_k)$

Gerace, Loureiro, Mézard, Krzakala, Zdeborová, *ICML, 2020*
Loureiro, Gerbelot, Cui, Goldt, Krzakala, Mézard, Zdeborová, *NeurIPS, 2021*
Loureiro, Sicuro, Gerbelot, Pacco, Krzakala, Zdeborová, *NeurIPS, 2021*
Gerace, Loureiro, Stephan, Krzakala, Zdeborová, *PRE, 2023*
Sarao Mannelli, **Gerace**, Rostamzadeh, Saglietti, *arXiv:2205.15935, 2022*

A Simplified Gaussian Data Model

$$\boldsymbol{s} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_i, \dots, \boldsymbol{s}_L\}$$

$$\boldsymbol{m} = \{m_1, \dots, m_i, \dots, m_L\}$$

Generalized Potts Model:

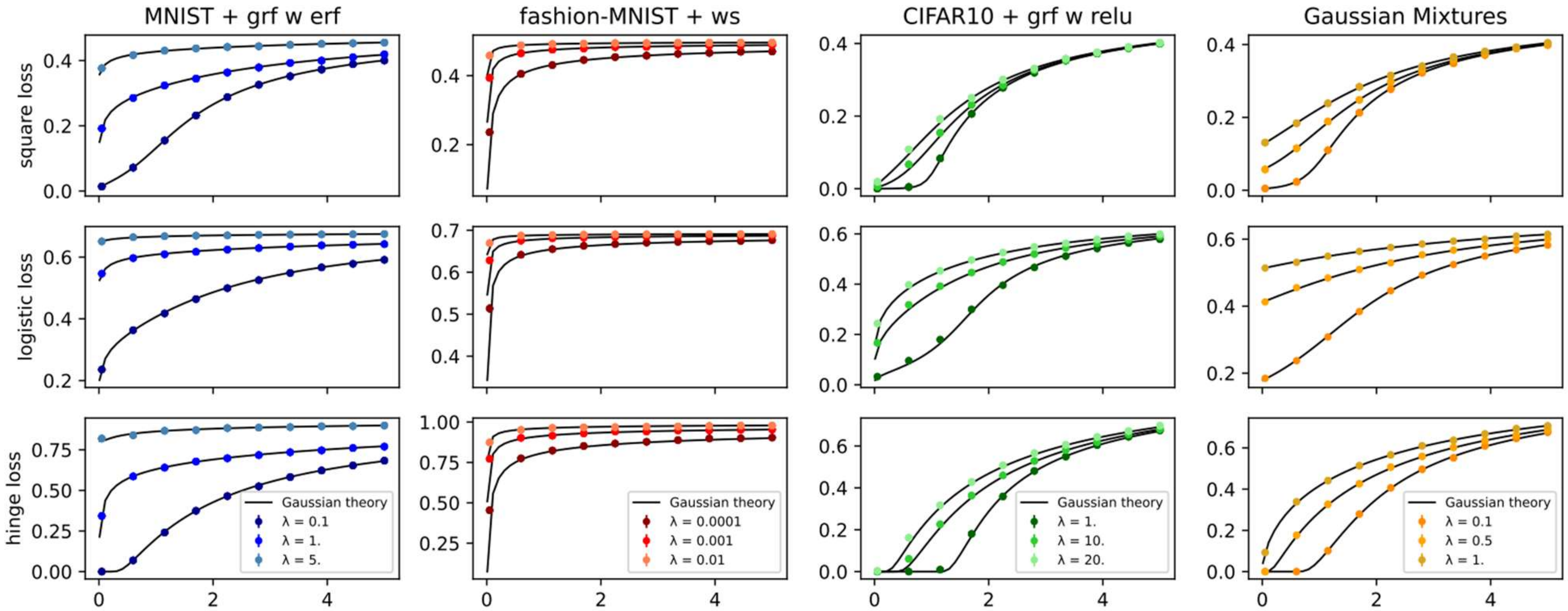
$$\mathcal{P}(\boldsymbol{s}) \propto \exp\left(\frac{\beta}{2}\sum_{i,j} J_{ij}\, \boldsymbol{s}_i^T U \boldsymbol{s}_j\right)$$

Gaussian Model:

$$\mathcal{N}(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\sum_{i,j} \Sigma_{ij}^{-1}\, m_i m_j\right)$$

Gaussian data are representative of real data

Single-Layer Neural Networks with Random Labels in classification tasks:



Number of Samples/Input Dimension

★ Shallow Neural Networks only care of data covariance with random labels!

Gerace, Loureiro, Stephan, Krzakala, Zdeborová, *PRE, 2023*

MLM from a Statistical Physics point of view

Masked Language Modelling:

Data:

$$\mathcal{D} = \left\{ \boldsymbol{m}^{\mu}_{\setminus i}, m^{\mu}_{i} \right\}_{\mu=1}^{M}$$

$$m_i^{\mu} \sim \mathcal{N}\left(m_i^{\mu} \middle| \boldsymbol{m}^{\mu}_{\setminus i};\ \Sigma_i^{-1} \right)$$

Empirical Risk Minimization:

$$\hat{A}_i = \underset{A}{\operatorname{argmin}}\ \mathbb{E}_{\mathcal{D}} \left[\frac{1}{2} \left(m_i^{\mu} - \sum_{j \neq i = 1}^{L} A_{ij} m_j^{\mu} \right)^2 \right]$$

Test on unseen data:

$$\epsilon_g = \mathbb{E}_{\{x^{new}, y^{new}\}} \left[\frac{1}{2} \sum_{i=1}^{L} \left(m_i^{new} - \sum_{j \neq i = 1}^{L} \hat{A}_{ij} m_j^{new} \right)^2 \right]$$

Teacher-Student:

Data:

$$\mathcal{D} = \{ \boldsymbol{x}^{\mu}, y^{\mu} \}_{\mu=1}^{M}$$

$$y^{\mu} \sim P(y^{\mu} | \boldsymbol{x}^{\mu}; \boldsymbol{T})$$

Empirical Risk Minimization:

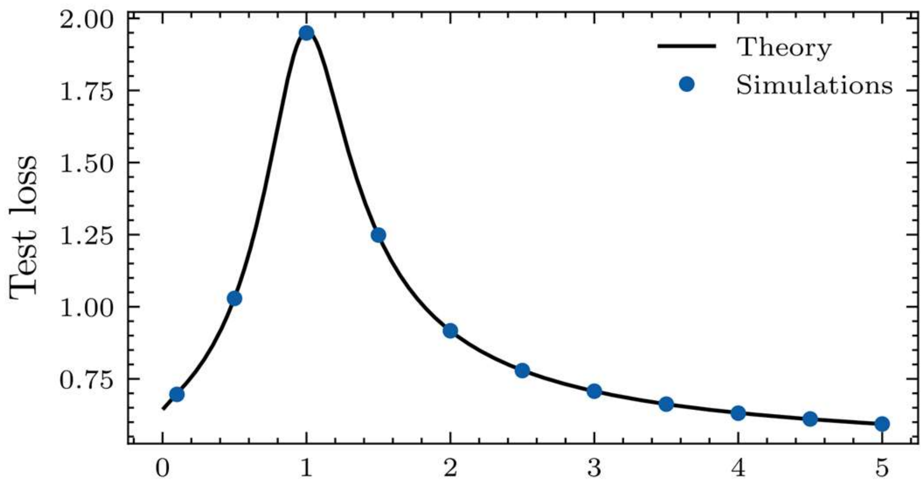
$$\widehat{\boldsymbol{W}} = \underset{W}{\operatorname{argmin}}\ \mathbb{E}_{\mathcal{D}} \left[\frac{1}{2} \left(y^{\mu} - \sum_{j=1}^{L} W_j x_j^{\mu} \right)^2 \right]$$

Test on unseen data:

$$\epsilon_g = \mathbb{E}_{\{x^{new}, y^{new}\}} \left[\frac{1}{2} \sum_{i=1}^{L} \left(y^{new} - \sum_{j=1}^{L} \widehat{W}_j x_j^{new} \right)^2 \right]$$

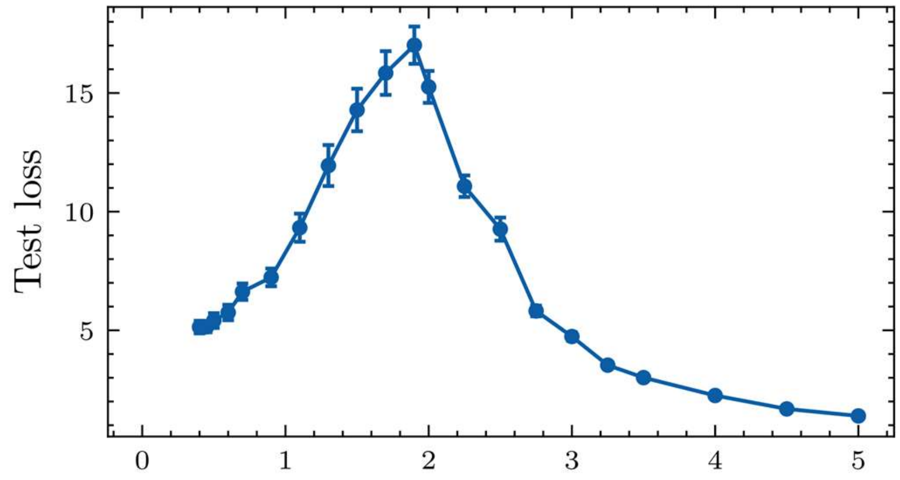
A New Generalization Behavior

Gaussian Model



Number of Samples/Sequence Length

Generalized Potts Model



Number of Samples/Sequence Length

★ Qualitative similar behavior!

To conclude few take home messages...

- Factored Attention learns the exact conditionals of the Generalized Potts Model;
- Deep Transformers sequentially learn high-order interactions in the input data;
- The interpolation peak appears in self-supervised learning too but it is now triggered by the noise inherent in the training data;
- There exist universality classes qualitatively describing the behavior of learning models on more complex data distribution.