News on TNs for ML and QC Simulation

Román Orús

DIPC, Ikerbasque, Multiverse

roman.orus@dipc.org

roman.orus@multiversecomputing.com

Today’s Menu

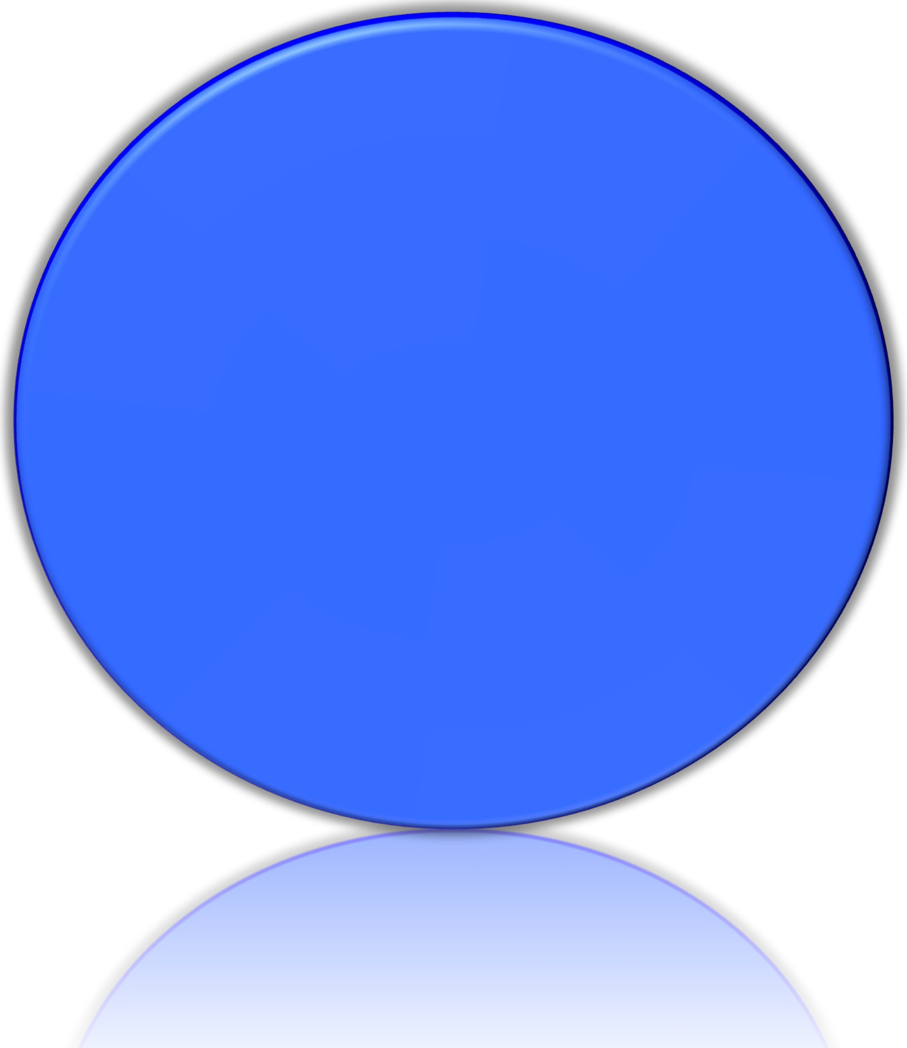
1) Simulating IBM’s kicked quantum Ising experiment with TNs
   arXiv:2309.15642

2) Compressing Large Language Models with TNs
   arXiv:2401.14109

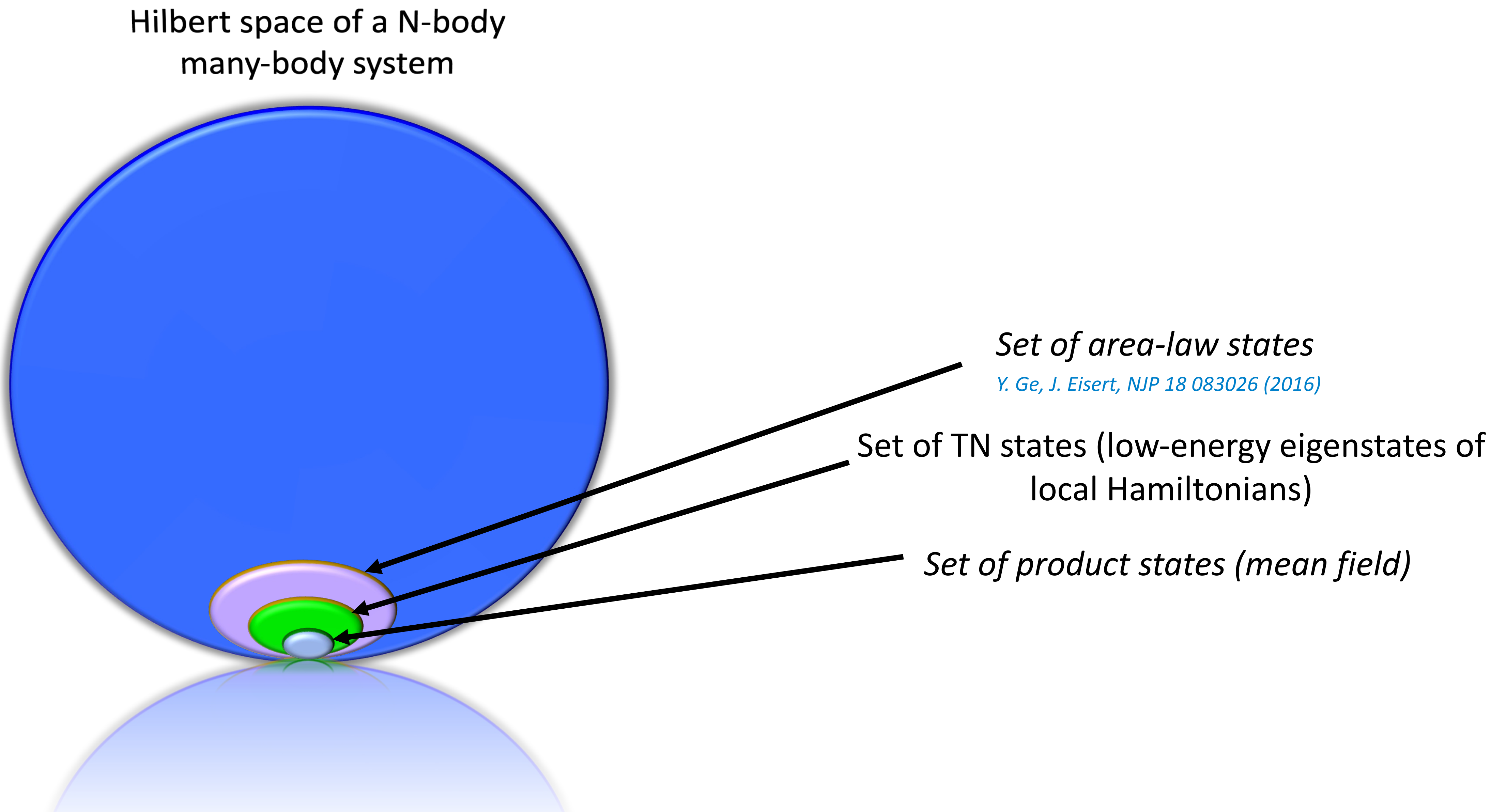
Tensor Networks

Hilbert space is a convenient illusion

Hilbert space of a N-body
many-body system



Hilbert space is a convenient illusion



Hilbert space is a convenient illusion

Hilbert space of a N-body many-body system

"Exploration" time ~ $O(10^{10^{23}})$ sec.

Compare to…

Age of the universe ~ $O(10^{17})$ sec.

Set of area-law states

Y. Ge, J. Eisert, NJP 18 083026 (2016)

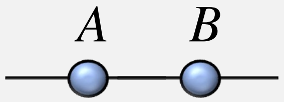
Set of TN states (low-energy eigenstates of local Hamiltonians)

Set of product states (mean field)

We need a language to target the relevant corner of quantum states directly

Tensor Networks

*e.g. RO, Annals of Physics **349** (2014) 117–158*

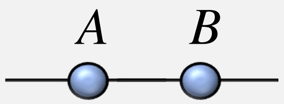
$A \cdot B \rightarrow$ 

$$|\Psi\rangle = \sum_{i's} \Psi_{i_1 i_2 \ldots i_N} |i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_N\rangle$$

p-level systems

Tensor Networks

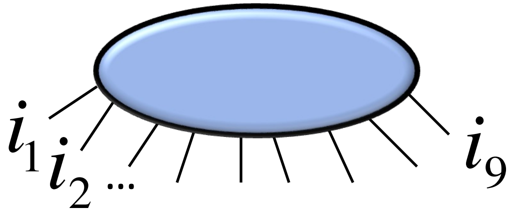
e.g. RO, Annals of Physics ***349*** *(2014) 117–158*

$A \cdot B \rightarrow$ 

$$|\Psi\rangle = \sum_{i's} \boxed{\Psi_{i_1 i_2 \ldots i_N}} |i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_N\rangle$$

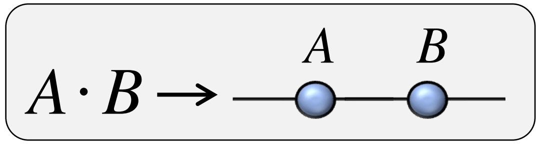
p-level systems

$$\Psi_{i_1 i_2 i_3 i_4 i_5 i_6 i_7 i_8 i_9}$$



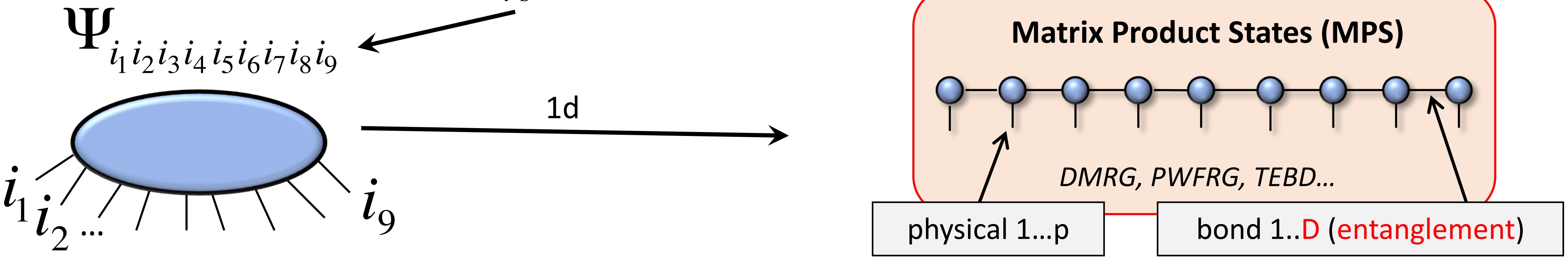
Tensor Networks

*e.g. RO, Annals of Physics **349** (2014) 117–158*



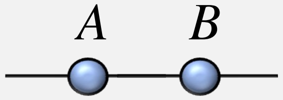
$$|\Psi\rangle = \sum_{i's} \boxed{\Psi_{i_1 i_2 \ldots i_N}} |i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_N\rangle$$

p-level systems



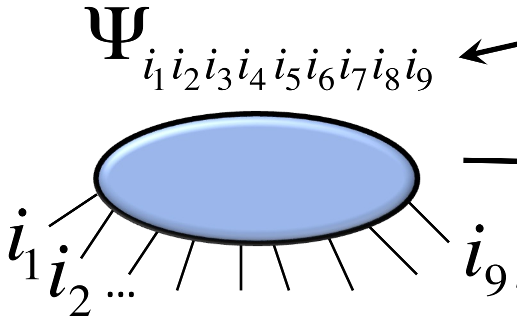
Tensor Networks

*e.g. RO, Annals of Physics **349** (2014) 117–158*

$A \cdot B \rightarrow$ 

$$|\Psi\rangle = \sum_{i's} \boxed{\Psi_{i_1 i_2 \ldots i_N}} |i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_N\rangle$$

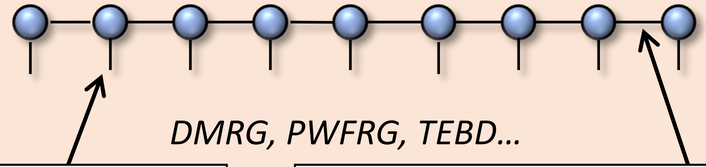
p-level systems



1d

2d, 3d...

Matrix Product States (MPS)

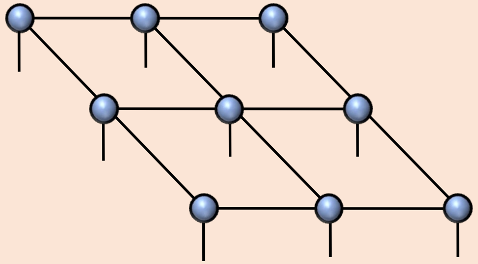


DMRG, PWFRG, TEBD…

physical 1…p

bond 1..D (entanglement)

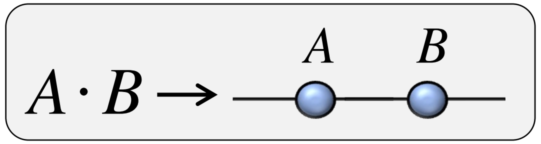
Projected Entangled Pair States (PEPS),
Tensor Product States (TPS)



Tensor Product Variational Approach, PEPS & iPEPS algorithms, Tensor-Entanglement Renormalization, TRG/SRG/HOTRG/HOSRG, gPEPS…

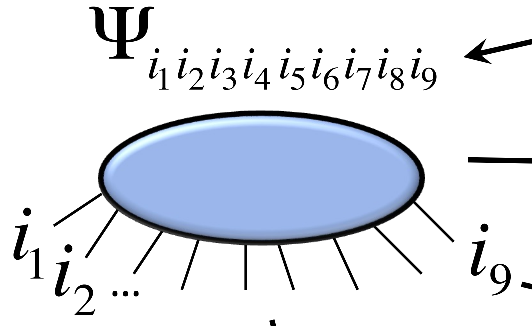
Tensor Networks

*e.g. RO, Annals of Physics **349** (2014) 117–158*



$$|\Psi\rangle = \sum_{i's} \boxed{\Psi_{i_1 i_2 \ldots i_N}} \, |i_1\rangle \otimes |i_2\rangle \otimes \cdots \otimes |i_N\rangle$$

p-level systems

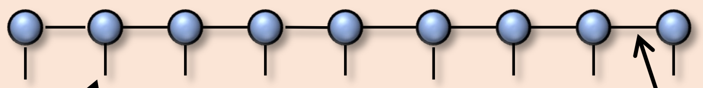


1d

2d, 3d...

Scale-invariant

Matrix Product States (MPS)

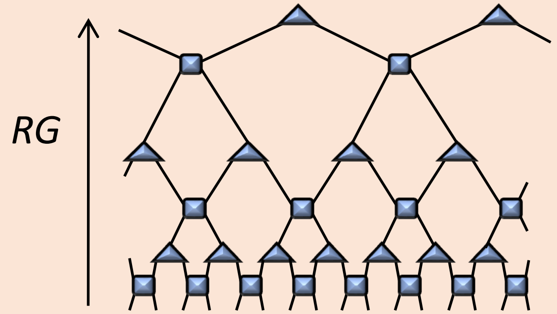


DMRG, PWFRG, TEBD...

physical 1...p

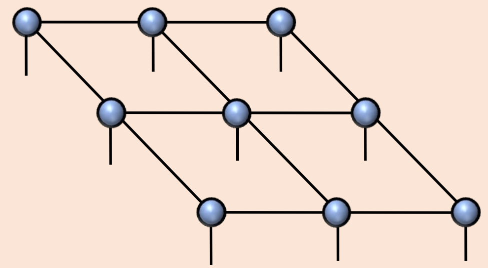
bond 1..D (entanglement)

Multiscale Entanglement Renormalization Ansatz (MERA)



AdS/CFT, Entanglement Renormalization

Projected Entangled Pair States (PEPS), Tensor Product States (TPS)



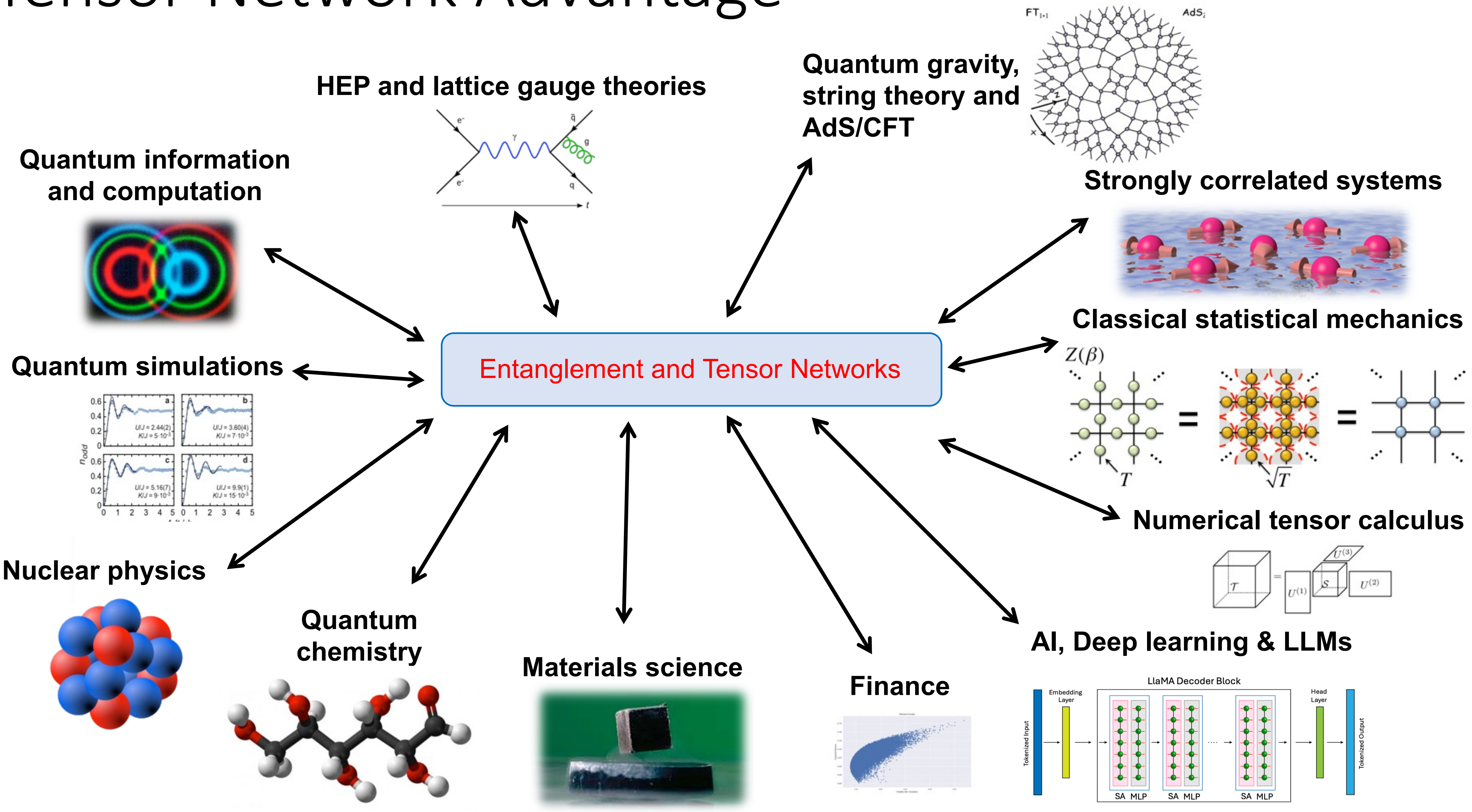
Tensor Product Variational Approach, PEPS & iPEPS algorithms, Tensor-Entanglement Renormalization, TRG/SRG/HOTRG/HOSRG, gPEPS...

Efficient O(poly(N)), satisfy area-law, low-energy eigenstates of local Hamiltonians

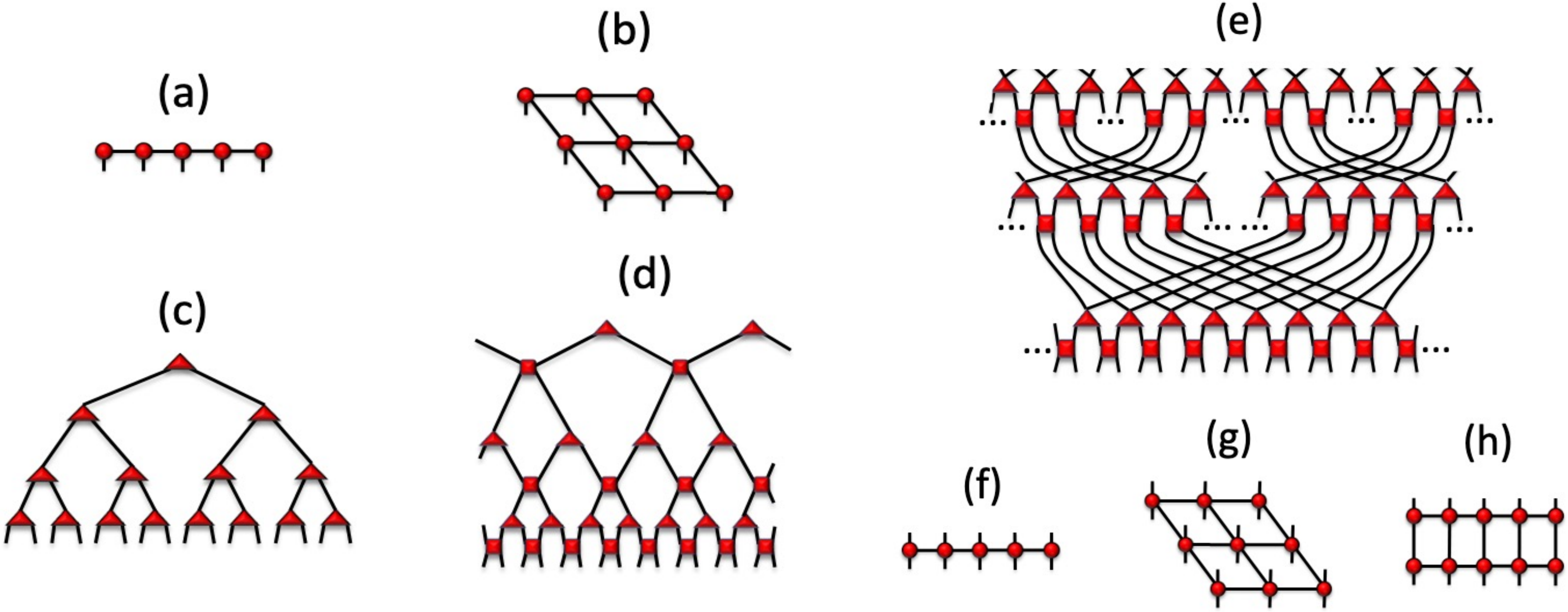
Tensor Network Advantage

Entanglement and Tensor Networks

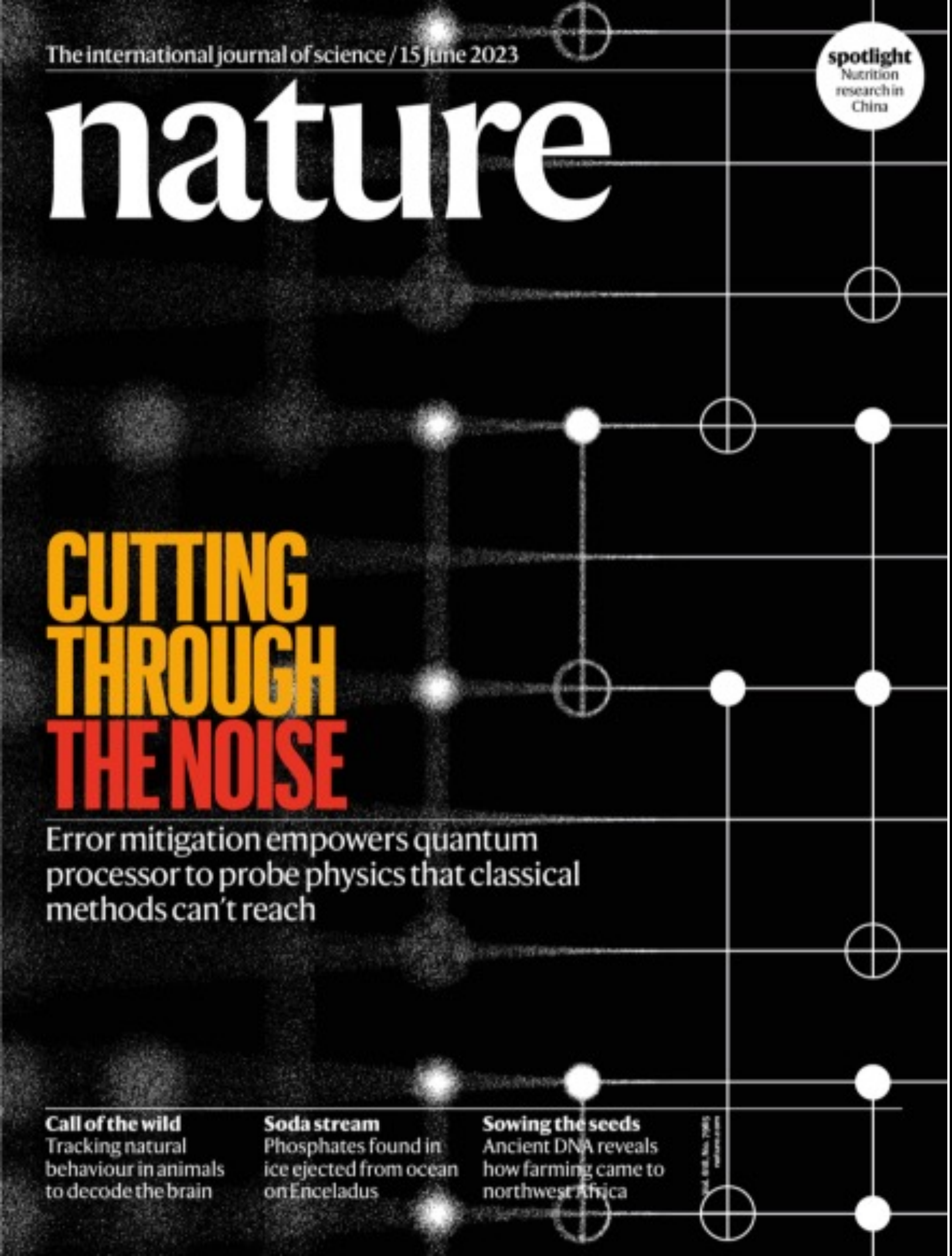
Tensor Network Advantage



Choose your fighter!



Simulating IBM’s QC



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open access | Published: 14 June 2023

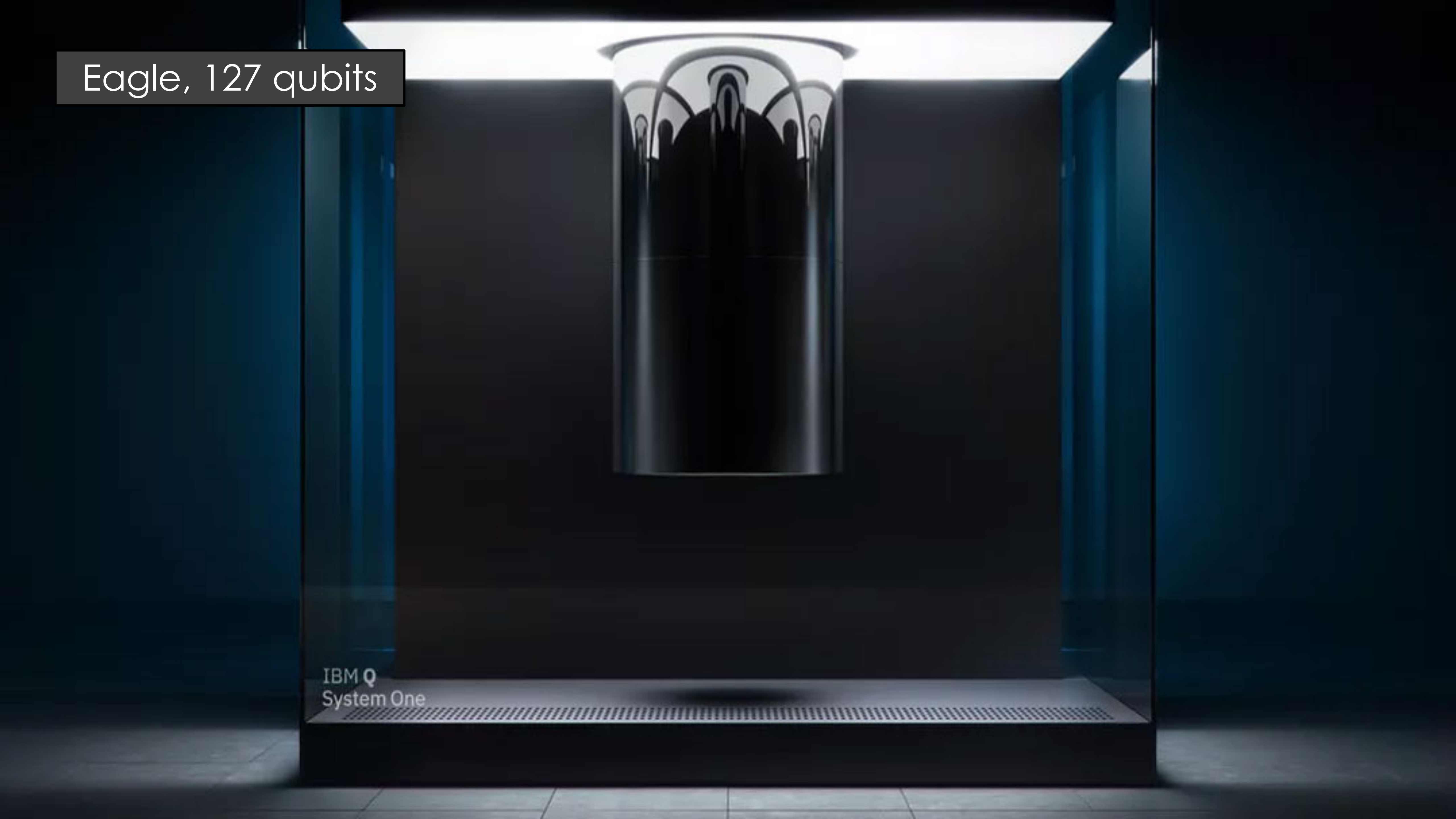
Evidence for the utility of quantum computing before fault tolerance

Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout van den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme & Abhinav Kandala

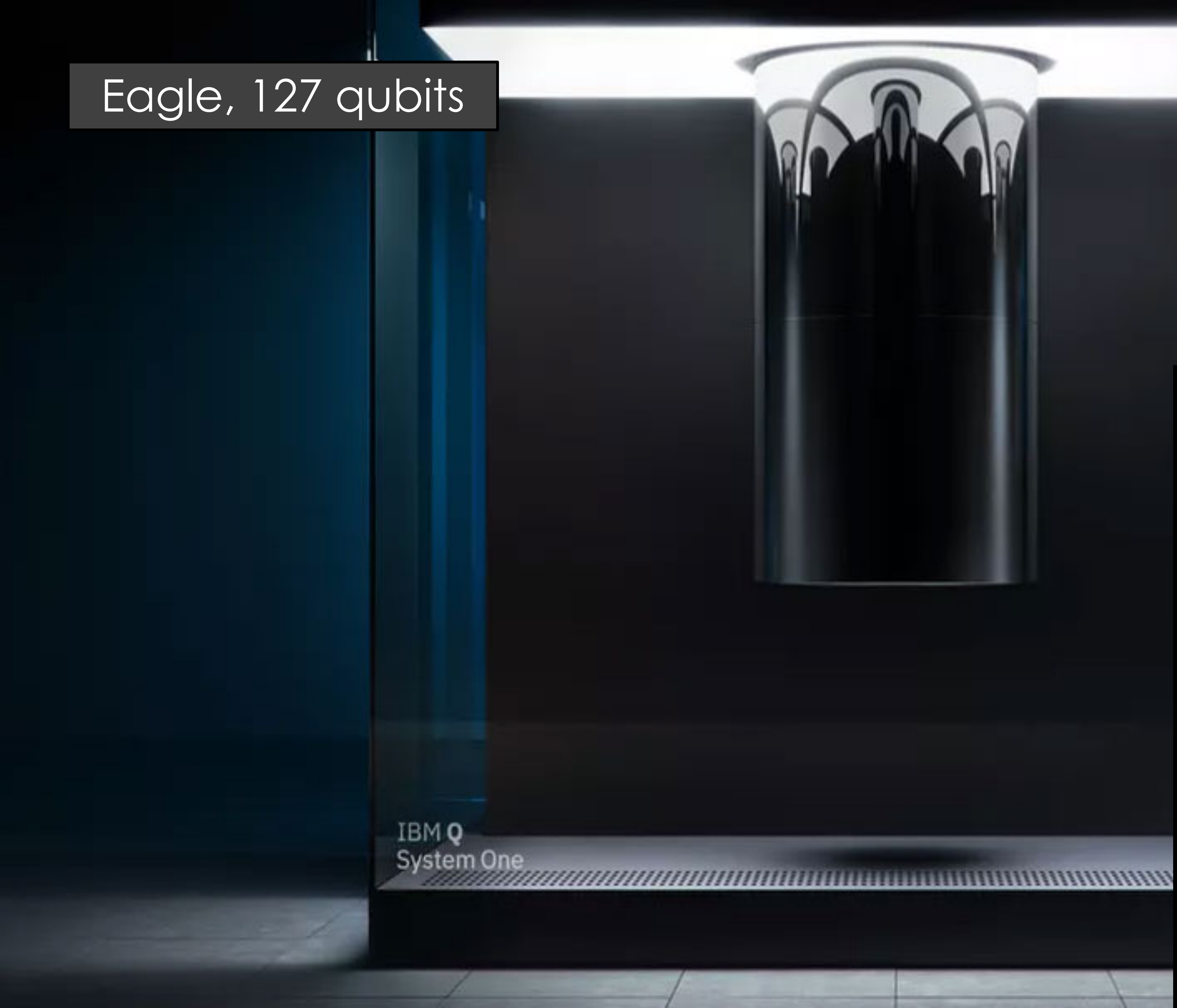
Nature **618**, 500–505 (2023) | Cite this article

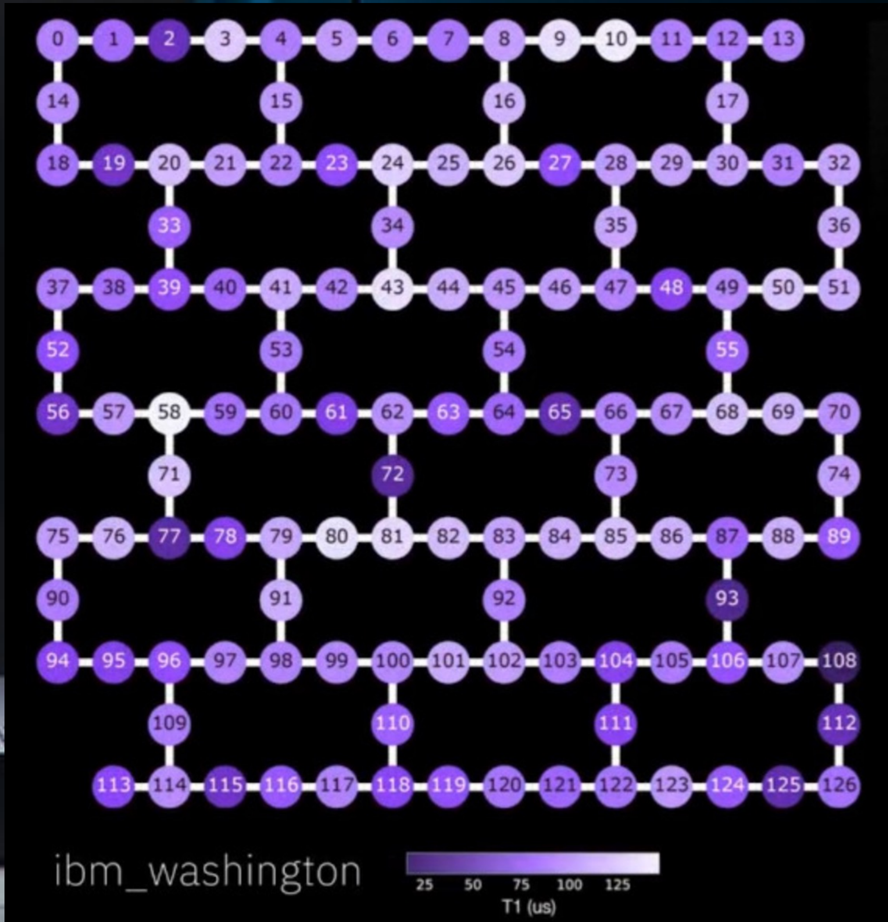
- IBM Quantum team published in June 2023 a quantum simulation paper using unprecedented error-mitigation techniques for a 127-qubit (''Eagle'') quantum processor.

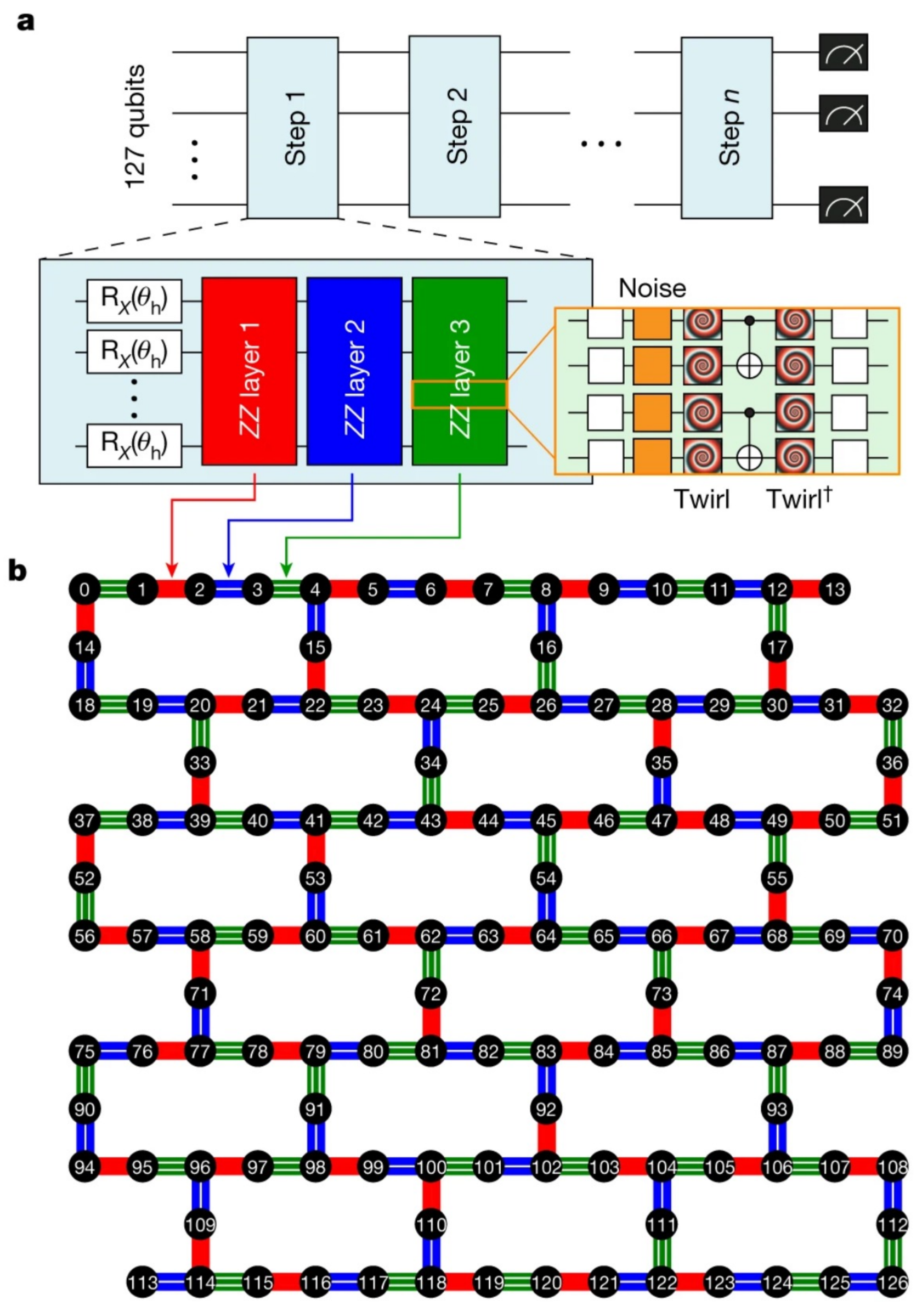
Eagle, 127 qubits



Eagle, 127 qubits





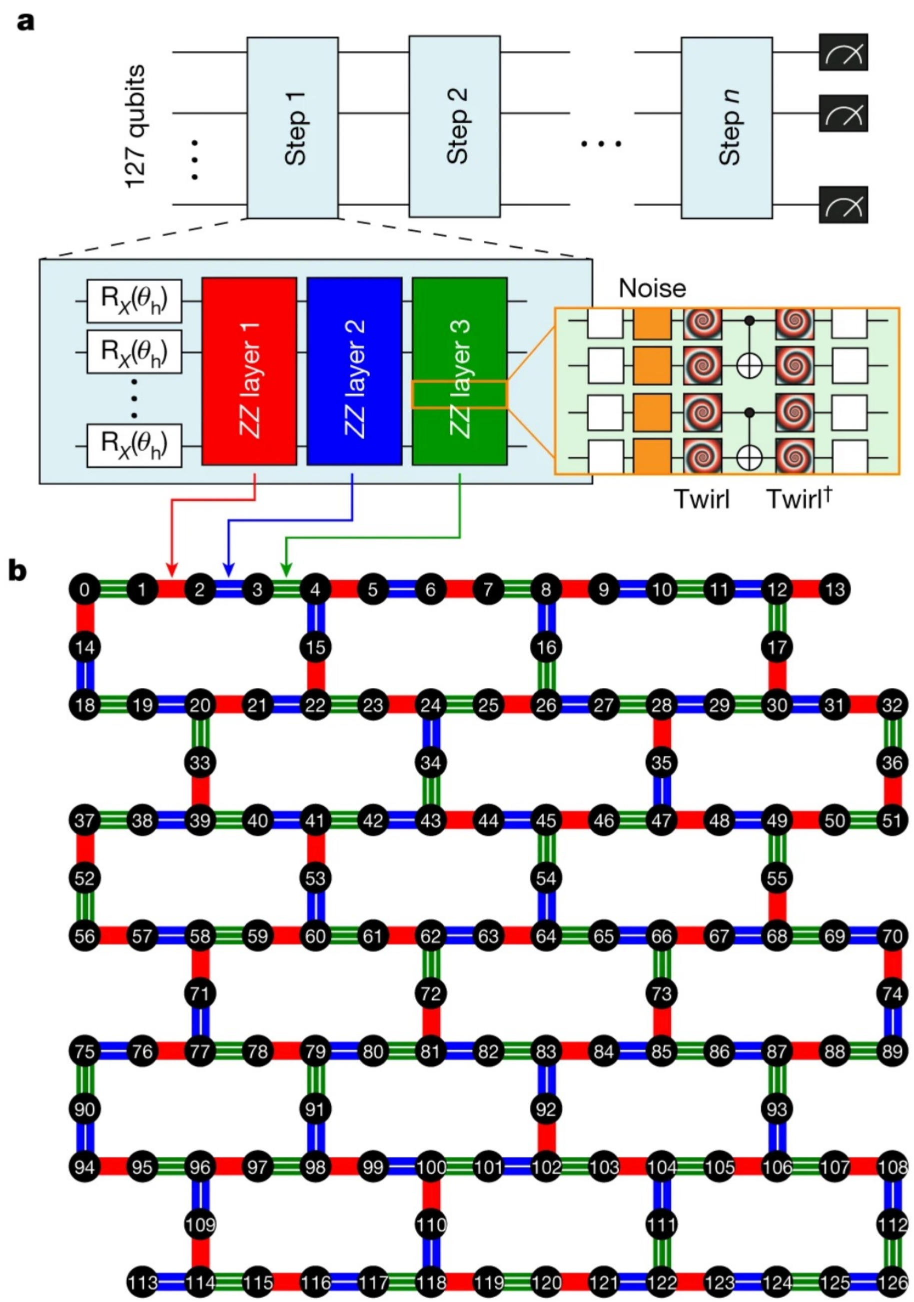


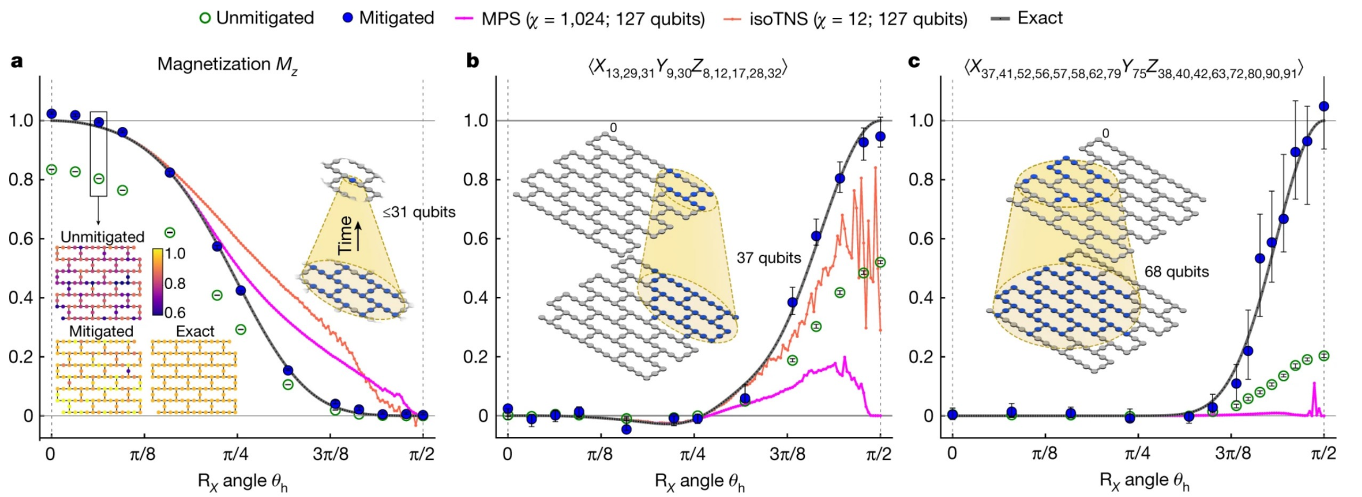
Spin-1/2 transverse field Ising model
on a heavy-hexagon lattice

$$H = -J \sum_{<i,j>} Z_i Z_j + h \sum_i X_i$$

Spin-1/2 transverse field Ising model
on a heavy-hexagon lattice

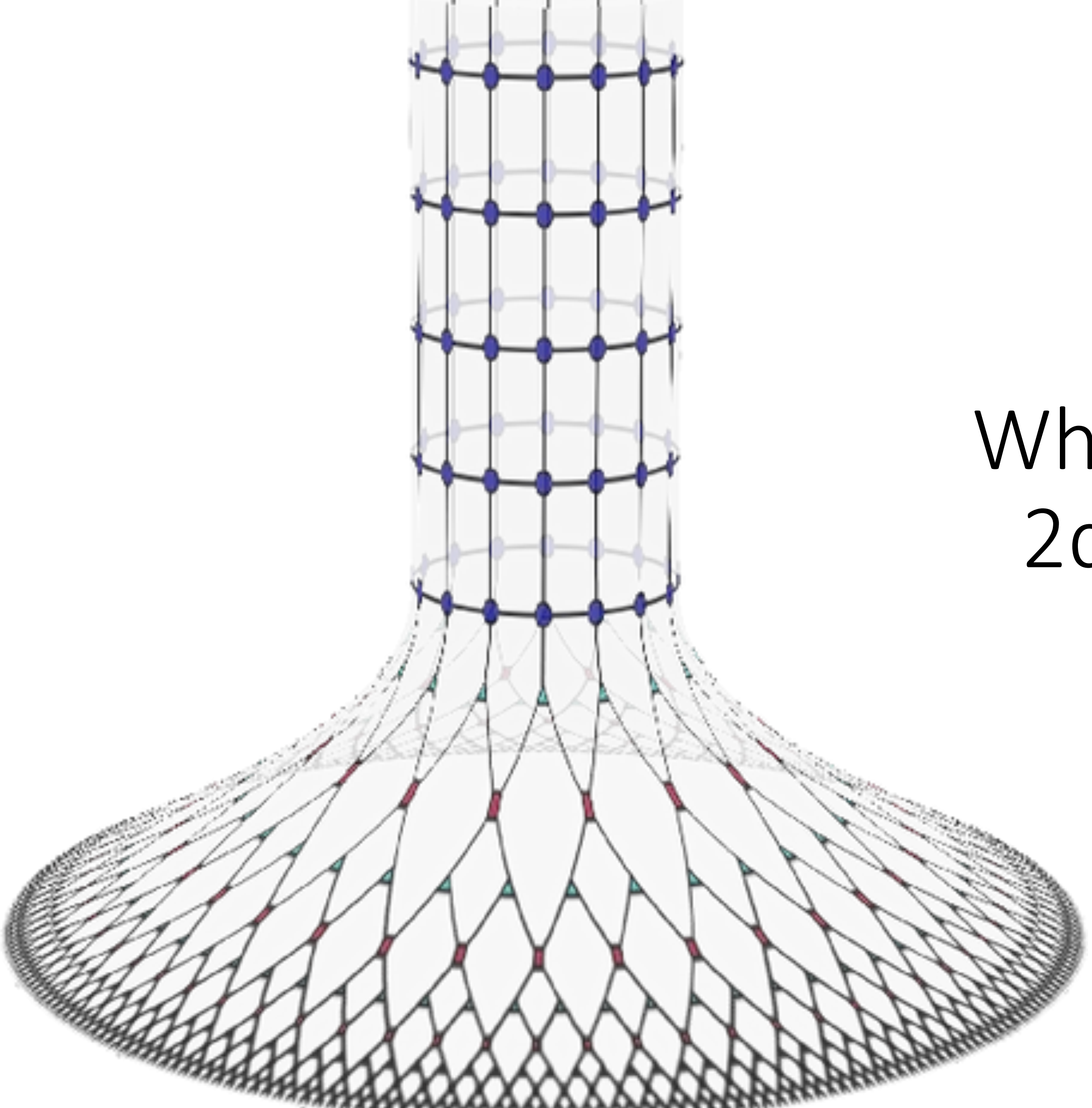
$$H = -J \sum_{<i,j>} Z_i Z_j + h \sum_i X_i$$





"the quantum computer provides correct results for which leading classical approximations such as pure- state based (...) tensor network methods break down".

"We have now reached reliability at a scale (...) which can provide utility beyond classical approximation methods."



What if we use other
2d TN algorithms?

PHYSICAL REVIEW B
covering condensed matter and materials physics

Highlights Recent Accepted Collections Authors Referees Search Press About

Universal tensor-network algorithm for any infinite lattice

Saeed S. Jahromi and Román Orús
Phys. Rev. B **99**, 195105 – Published 3 May 2019

Article References Citing Articles (13) **PDF** **HTML** **Export Citation**

ABSTRACT

We present a general graph-based projected entangled-pair state (gPEPS) algorithm to approximate ground states of nearest-neighbor local Hamiltonians on any lattice or graph of infinite size. By introducing the structural matrix, which codifies the details of tensor networks on any graphs in any dimension d, we are able to produce a code that can be essentially launched to simulate any lattice. We further introduce an optimized algorithm to compute simple tensor updates as well as expectation values and correlators with a mean-field-like effective environments. Though not being variational, this strategy allows to cope with PEPS of very large bond dimension (e.g., $D = 100$) and produces remarkably accurate results in the thermodynamic limit in many situations, and specially when the correlation length is small and the connectivity of the lattice is large. We prove the validity of our approach by benchmarking the algorithm against known results for several models, i.e., the antiferromagnetic Heisenberg model on a chain, star and cubic lattices, the hardcore Bose-Hubbard model on square lattice, the ferromagnetic Heisenberg model in a field on the pyrochlore lattice, as well as the three-state quantum Potts model in field on the kagome lattice and the spin-1 bilinear-biquadratic Heisenberg model on the triangular lattice. We further demonstrate the performance of gPEPS by studying the quantum phase transition of the $2d$ quantum Ising model in transverse magnetic field on the square lattice, and the phase diagram of the Kitaev-Heisenberg model on the hyperhoneycomb lattice. Our results are in excellent agreement with previous studies.

Our fighter ☺

Graph-PEPS algorithm (gPEPS)

- Simple tensor update
- Mean field environments
- Flexible to adapt to any lattice and any dimension
- Very accurate away from criticality

Simulating the IBM experiment

Works shortly after IBM's paper (all focused on 127 qubits)

[3] J. Tindall, M. Fishman, M. Stoudenmire, and D. Sels, Efficient tensor network simulation of ibm's eagle kicked ising experiment (2023), arXiv:2306.14887 [quant-ph].

[4] H.-J. Liao, K. Wang, Z.-S. Zhou, P. Zhang, and T. Xiang, Simulation of ibm's kicked ising experiment with projected entangled pair operator (2023), arXiv:2308.03082 [quant-ph].

[5] S. Anand, K. Temme, A. Kandala, and M. Zaletel, Classical benchmarking of zero noise extrapolation beyond the exactly-verifiable regime (2023), arXiv:2306.17839 [quant-ph].

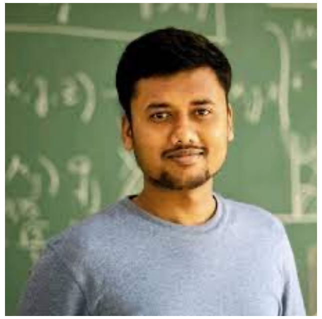
[6] T. Begusic, J. Gray, and G. K.-L. Chan, Fast and converged classical simulations of evidence for the utility of quantum computing before fault tolerance (2023), arXiv:2308.05077 [quant-ph].

[7] T. Begusic and G. K.-L. Chan, Fast classical simulation of evidence for the utility of quantum computing before fault tolerance (2023), arXiv:2306.16372 [quant-ph].

[8] M. S. Rudolph, E. Fontana, Z. Holmes, and L. Cincio, Classical surrogate simulation of quantum systems with lowesa (2023), arXiv:2308.09109 [quant-ph].

Our work (much beyond previous references)

Siddhartha Patra @DIPC+Multiverse



Saeed Jahromi @Multiverse

Sukhbinder Singh @Multiverse

Román Orús @DIPC+Multiverse

arXiv > quant-ph > arXiv:2309.15642

Quantum Physics

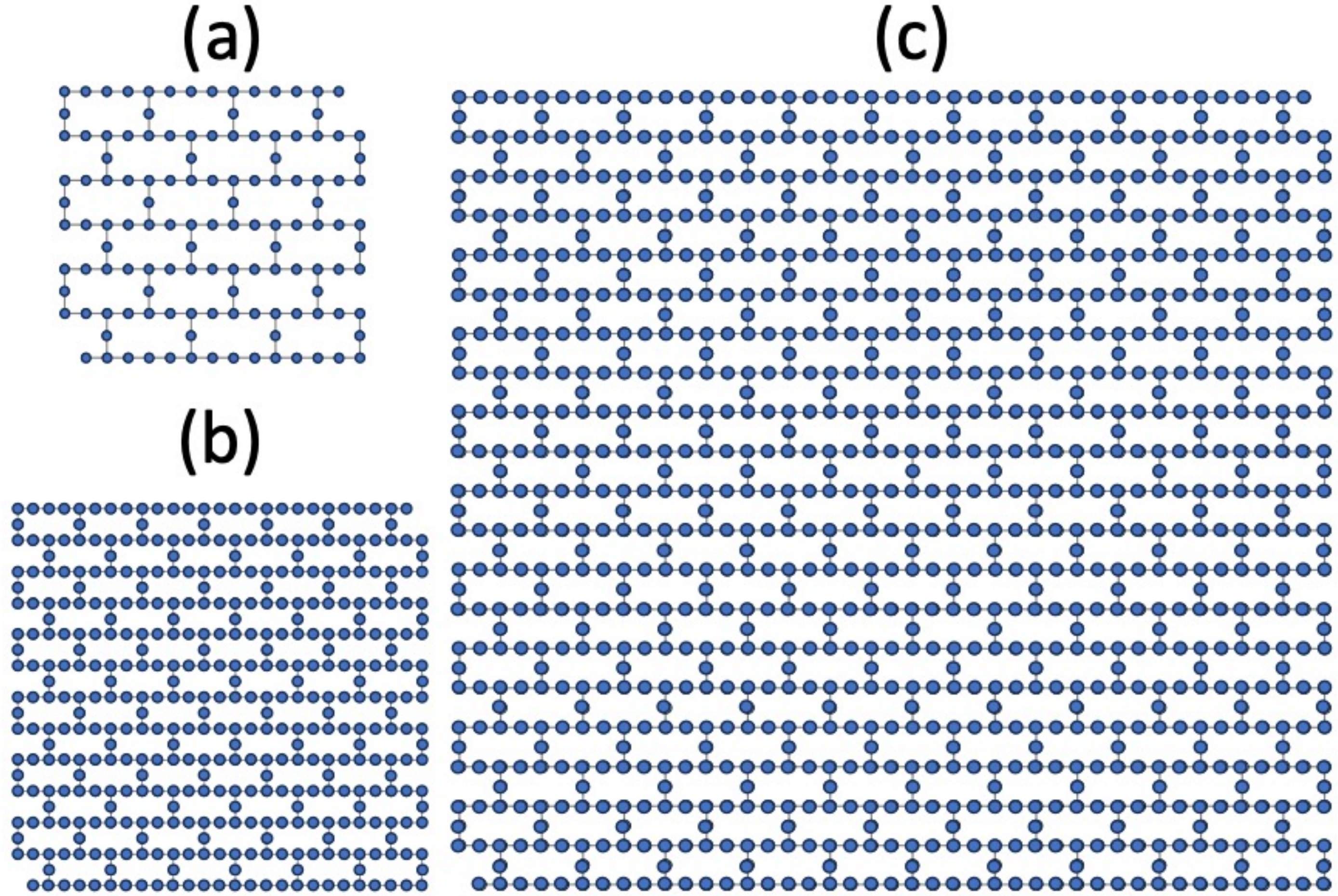
[Submitted on 27 Sep 2023 (v1), last revised 16 Oct 2023 (this version, v2)]

Efficient tensor network simulation of IBM's largest quantum processors

Siddhartha Patra, Saeed S. Jahromi, Sukhbinder Singh, Roman Orus

We show how quantum-inspired 2d tensor networks can be used to efficiently and accurately simulate the largest quantum processors from IBM, namely Eagle (127 qubits), Osprey (433 qubits) and Condor (1121 qubits). We simulate the dynamics of a complex quantum many-body system -- specifically, the kicked Ising experiment considered recently by IBM in Nature 618, p. 500-505 (2023) -- using graph-based Projected Entangled Pair States (gPEPS), which was proposed by some of us in PRB 99, 195105 (2019). Our results show that simple tensor updates are already sufficient to achieve very large unprecedented accuracy with remarkably low computational resources for this model. Apart from simulating the original experiment for 127 qubits, we also extend our results to 433 and 1121 qubits, and for evolution times around 8 times longer, thus setting a benchmark for the newest IBM quantum machines. We also report accurate simulations for infinitely-many qubits. Our results show that gPEPS are a natural tool to efficiently simulate quantum computers with an underlying lattice-based qubit connectivity, such as all quantum processors based on superconducting qubits.

(a) Eagle (127), (b) Osprey (433), (c) Condor (1121)



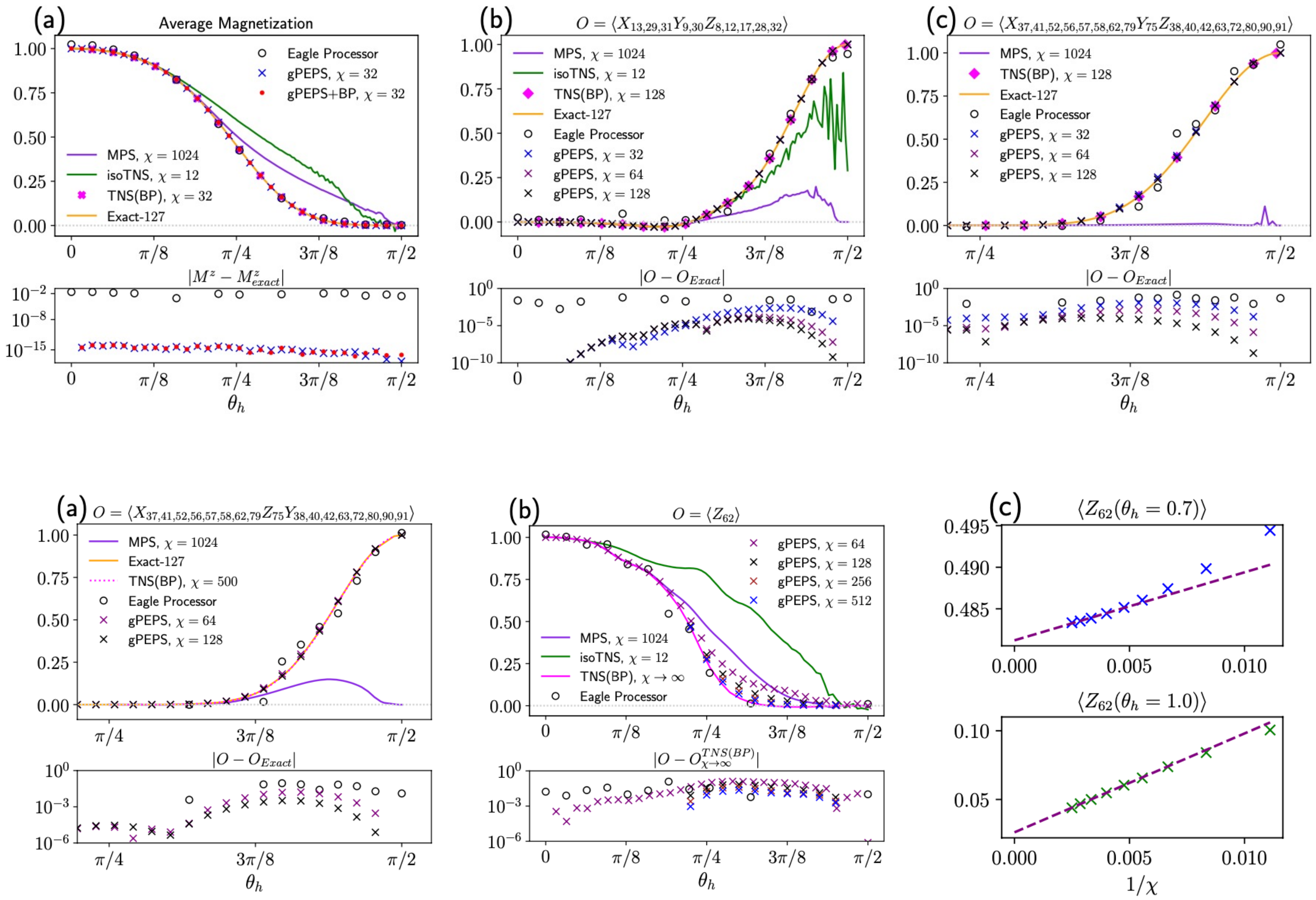
$$H = -J \sum_{<i,j>} Z_i Z_j + h \sum_i X_i$$

$$U(\theta_h) = \left(\prod_{\langle i,j \rangle} e^{i \frac{\pi}{4} Z_i Z_j} \right) \left(\prod_i e^{-i \frac{\theta_h}{2} X_i} \right)$$

$$|\psi(\theta_h, n)\rangle \equiv (U(\theta_h))^n \, |0\rangle^{\otimes m}$$

… and we also considered the thermodynamic limit, infinitely-many qubits

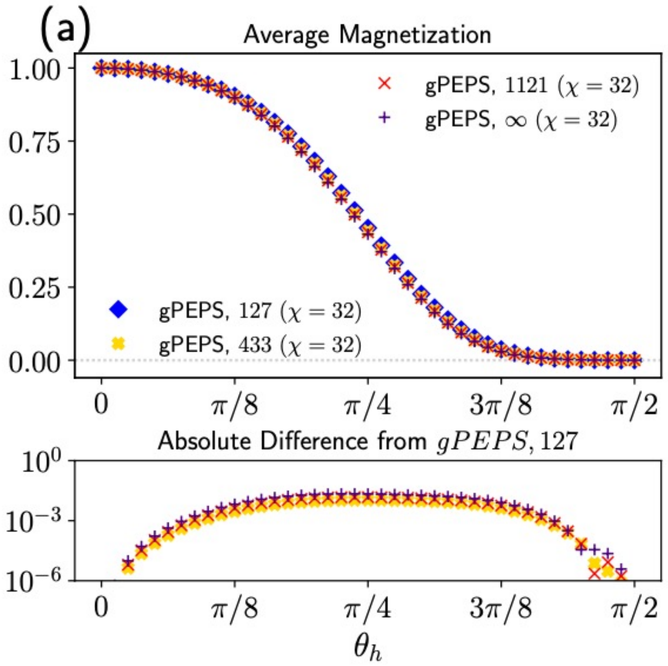
Eagle results, 127 qubits



- gPEPS simulated perfectly the system
- Unprecedented low error, even better than other TN techniques (eg TNS-BP)
- Average of 2s/point on a PC (QC was 5h/point). Superfast!!!
- Results hold for local and non-local observables

(approx 5 Trotter steps)

Larger systems

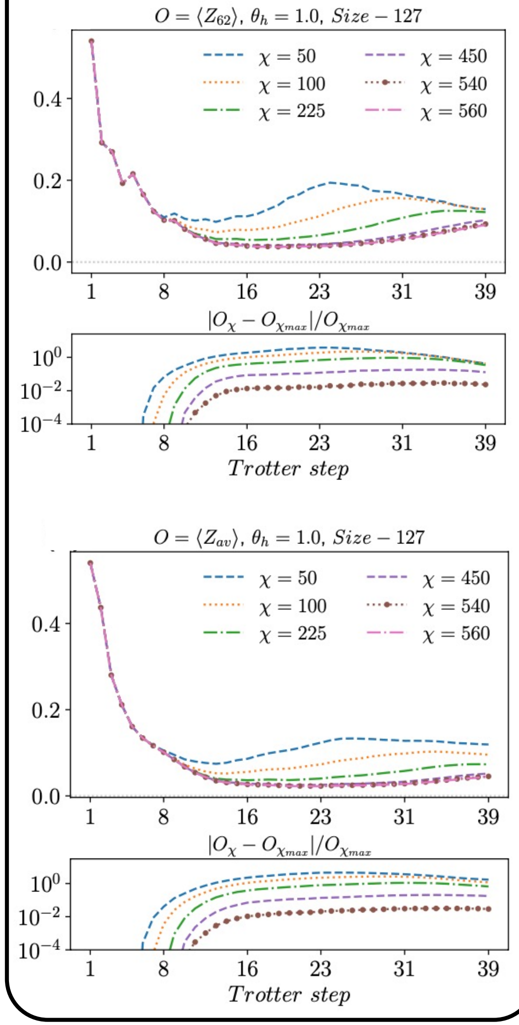


- Large accuracy also for larger systems: 127, 433, 1121 and infinitely-many qubits
- Simulations also extremely efficient

Long time evolutions

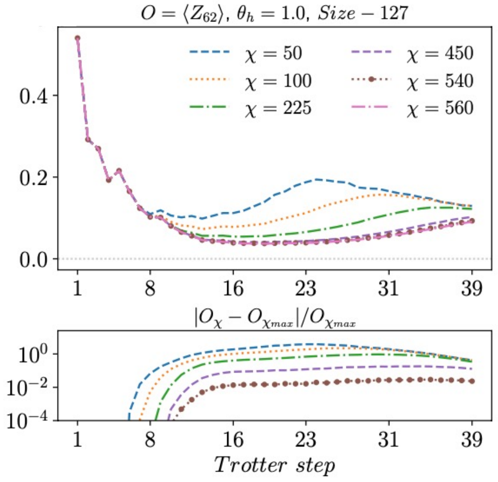
Long time evolutions

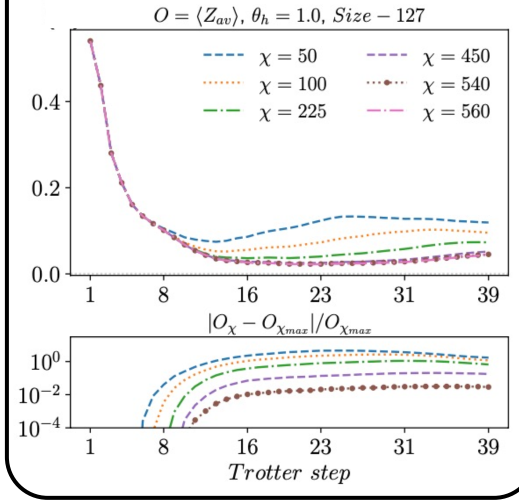
127 qubits



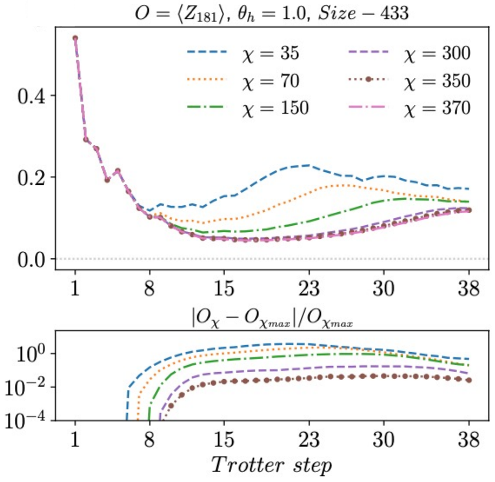
Long time evolutions

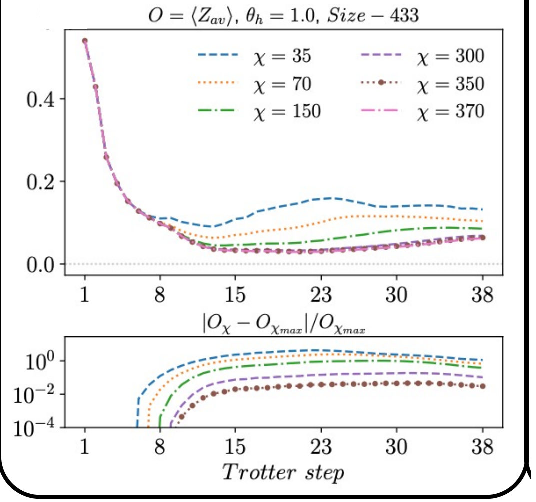
127 qubits





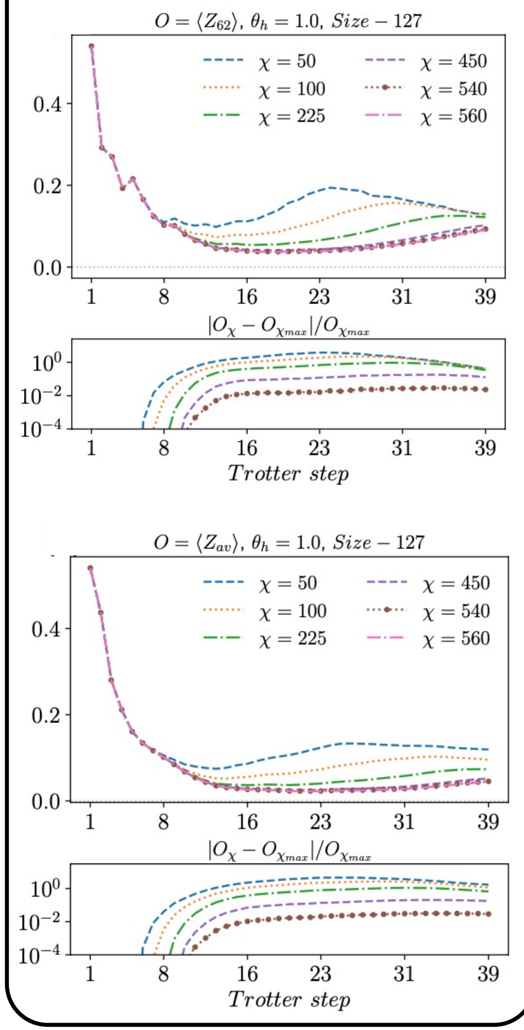
433 qubits



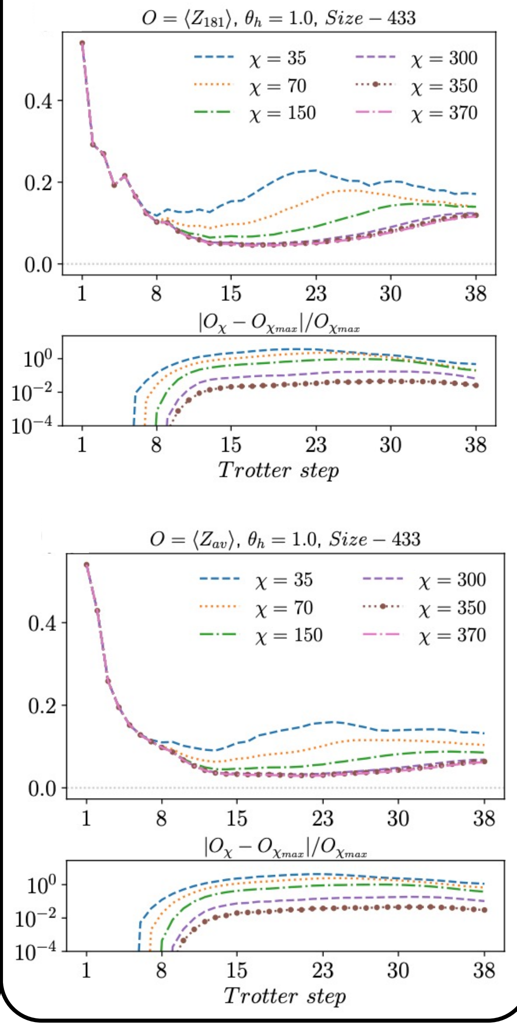


Long time evolutions

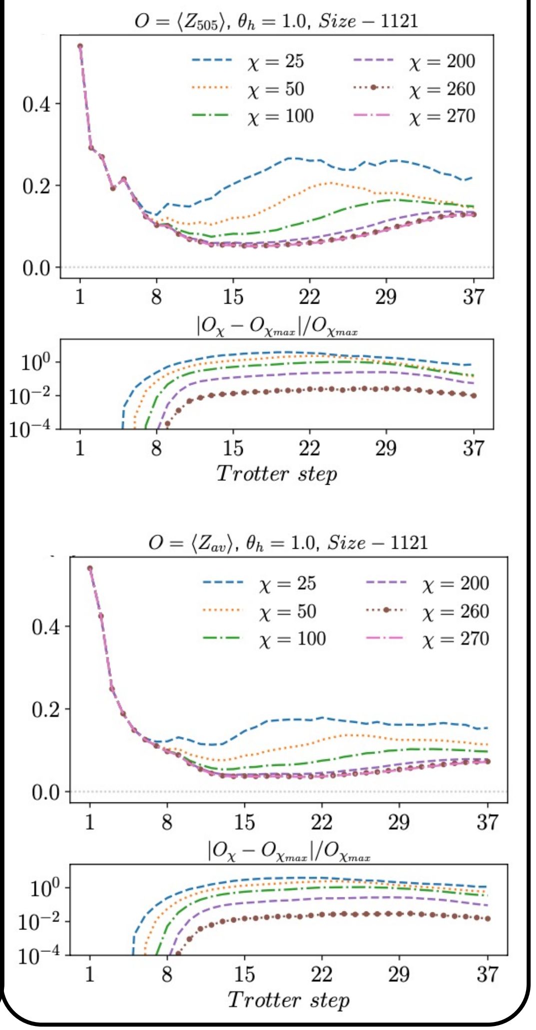
127 qubits



433 qubits

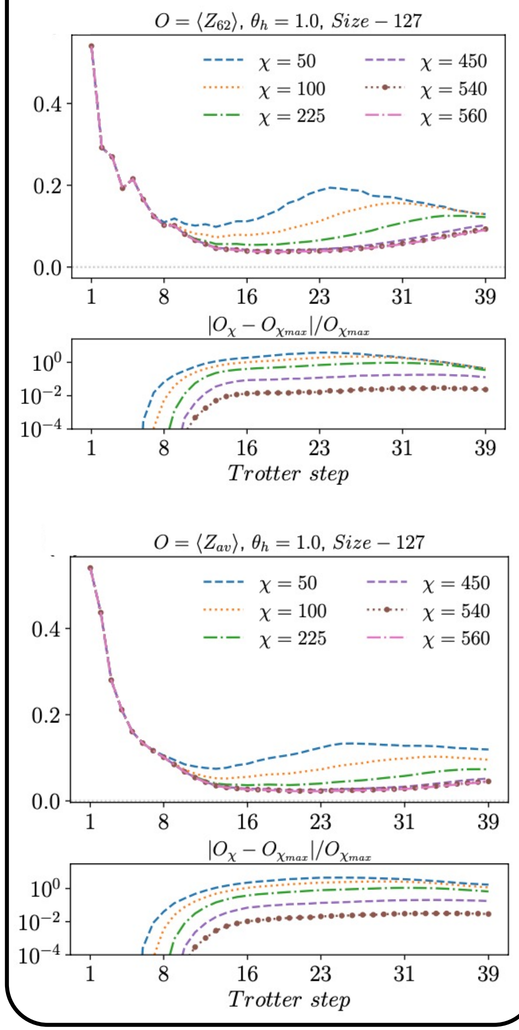


1121 qubits

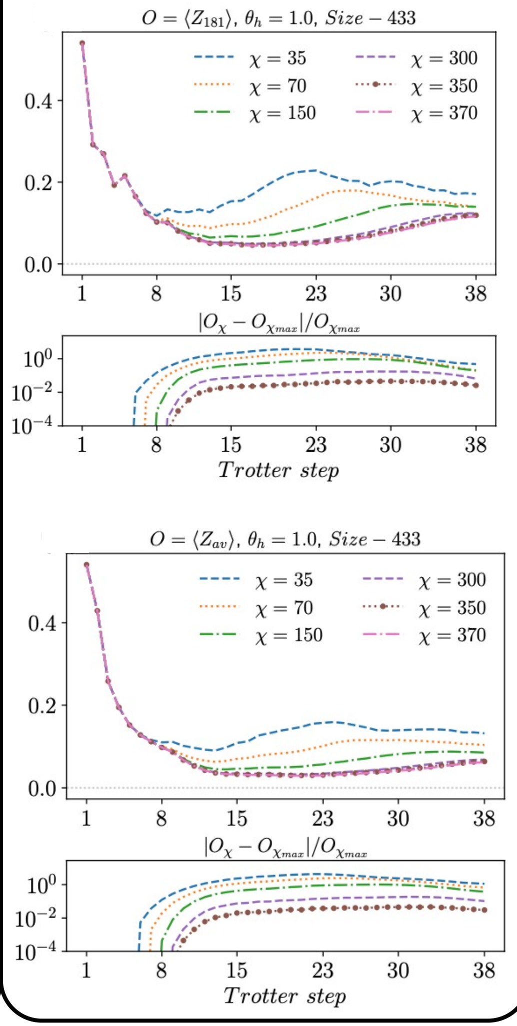


Long time evolutions

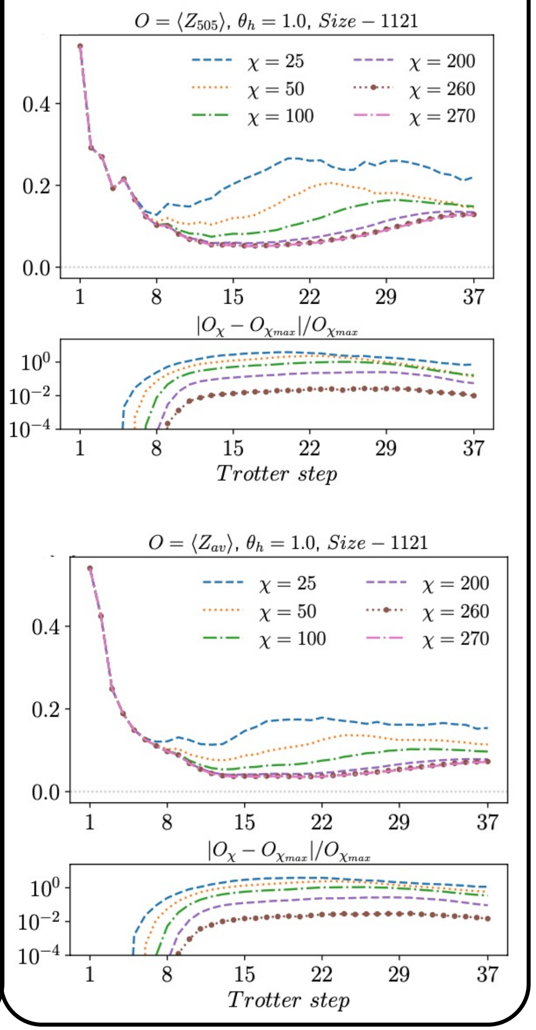
127 qubits



433 qubits



1121 qubits



- gPEPS still simulates perfectly the system, convergence with the PEPS bond dimension even for long times.
- Beyond all calculations done for this system with any technique
- Results hold for local and non-local observables
- Pushed to the memory limit of DIPC cluster, and still works!

What does all this imply?

- We simulated the original IBM experiment for (at least) 10x more qubits and 8x longer time evolutions.
- The original experiment is fine for testing error mitigation, but cannot claim advantage beyond classical simulation methods, in particular with respect to tensor networks.
- The Hamiltonian is too tailored to the topology of interactions in the quantum circuit.
- The heavy-hexagon lattice is one of the easiest 2d lattices to simulate (very close to a tree!)
- Can lattice-based quantum computers (superconducting, quantum dots…) reach the low level of noise required not to be simulated by an appropriate TN algorithm? (reminder: gPEPS was the simplest 2d algorithm we could think of).
- Lattice is a huge constraint. Why imposing a lattice at all? Why not going for lattice-free quantum processors? And lattice-free simulators?

Compressing
Large Language Models
(LLMs)

What is Generative AI?

Generative artificial intelligence is a type of artificial intelligence (AI) system capable of generating text, images, media, or other type of data.

Generative AI models learn the patterns and structure of their input training data, and then generate new data that has similar characteristics.

Impact goes from improving current solutions, to destroying humanity ☺.

ChatGPT

M

Write a monologue about future in heaven and hell from the perspective of a sentient cactus.

"I am a cactus, a being that has stood here for centuries, watching the world go by. I have seen the rise and fall of civilizations, the birth and death of countless living creatures. But there is one thing that has always fascinated me, and that is the concept of the afterlife.

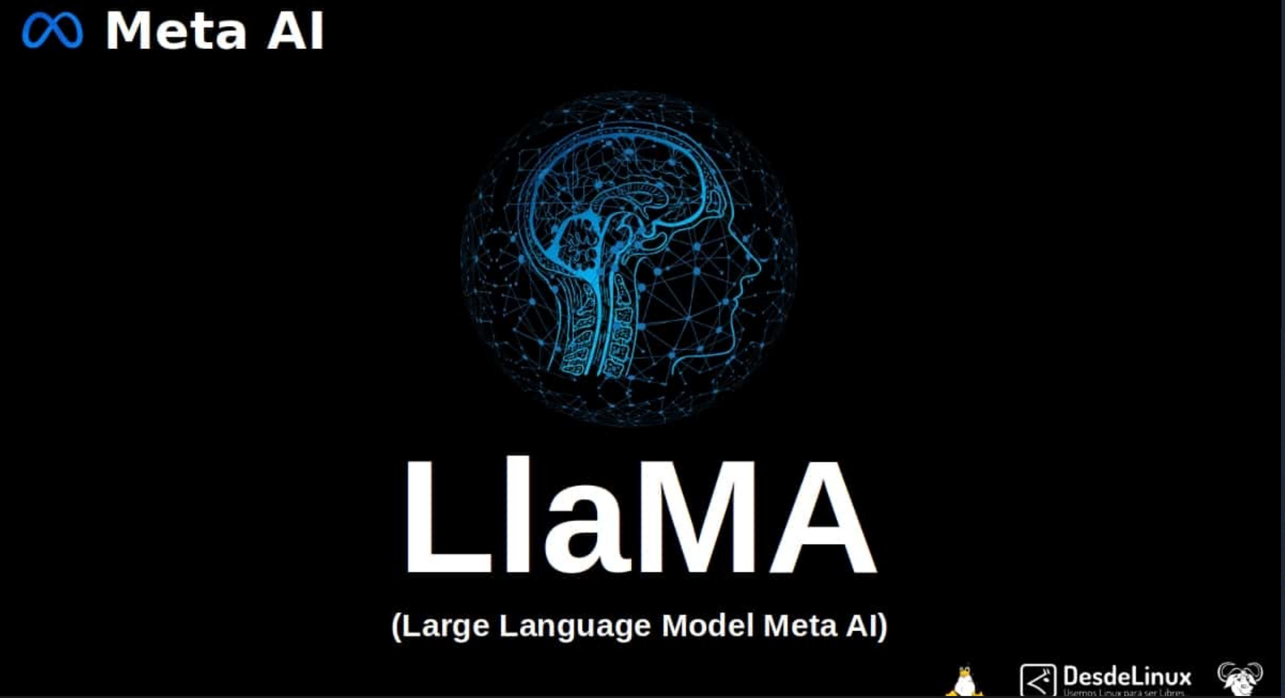
They say that when we die, we will either go to heaven or hell. And as a cactus, I cannot help but wonder what that would be like.

In heaven, they say there are gardens of paradise, where the flowers bloom all year round and the sun always shines. But what would that be like for a cactus? Would I be surrounded by lush greenery, or would I be in a desert oasis with nothing but sand and rocks?

And then there's hell, where they say there is fire and brimstone. But would that be so bad for a cactus? After all, I am a desert plant and I thrive in harsh conditions. Perhaps hell would be like a vacation for me.

But then again, I cannot help but think that heaven and hell are not so different. They are both places where we go after we die, and they are both places where we will be judged. And as a cactus, I cannot help but wonder what judgment would be passed upon me.

But perhaps it doesn't matter. After all, I am just a cactus, a simple plant that has stood here for centuries. I will continue to watch the world go by, and I will continue to wonder about the afterlife. But in the end, I will accept whatever fate is in store for me, be it heaven or hell."



Mark Zuckerberg
18h ·

Today we're releasing a new state-of-the-art AI large language model called LLaMA designed to help researchers advance their work. LLMs have shown a lot of promise in generating text, having conversations, summarizing written material, and more complicated tasks like solving math theorems or predicting protein structures. Meta is committed to this open model of research and we'll make our new model available to the AI research community.

61K 8.4K 2.7K

Like Comment Share

1000 WORD STORY

It was a typical day for Nate. He had just finished his classes at college and he was heading home. On his way back, he stopped by his favorite gaming store to pick up some snacks and drinks. As he walked in, he noticed that there were two people playing on one of the demo computers. They were both playing the popular multiplayer online battle arena (MOBA) game, League of Legends. The players were both using champions that Nate wasn't familiar with. One player was controlling a champion named "Okono" while the other controlled a champion called "Kayle".

The match started and it looked like Kayle was dominating. She was quickly gaining levels and items, while Okono was barely able to keep up. Nate watched the match closely and soon realized why Kayle was so much more

DALL·E 2

DALL·E

Edit the detailed description

Surprise me

Upload

the painting American Gothic, with two dogs holding pepperoni pizza instead of the farmers holding a pitchfork

Report issue

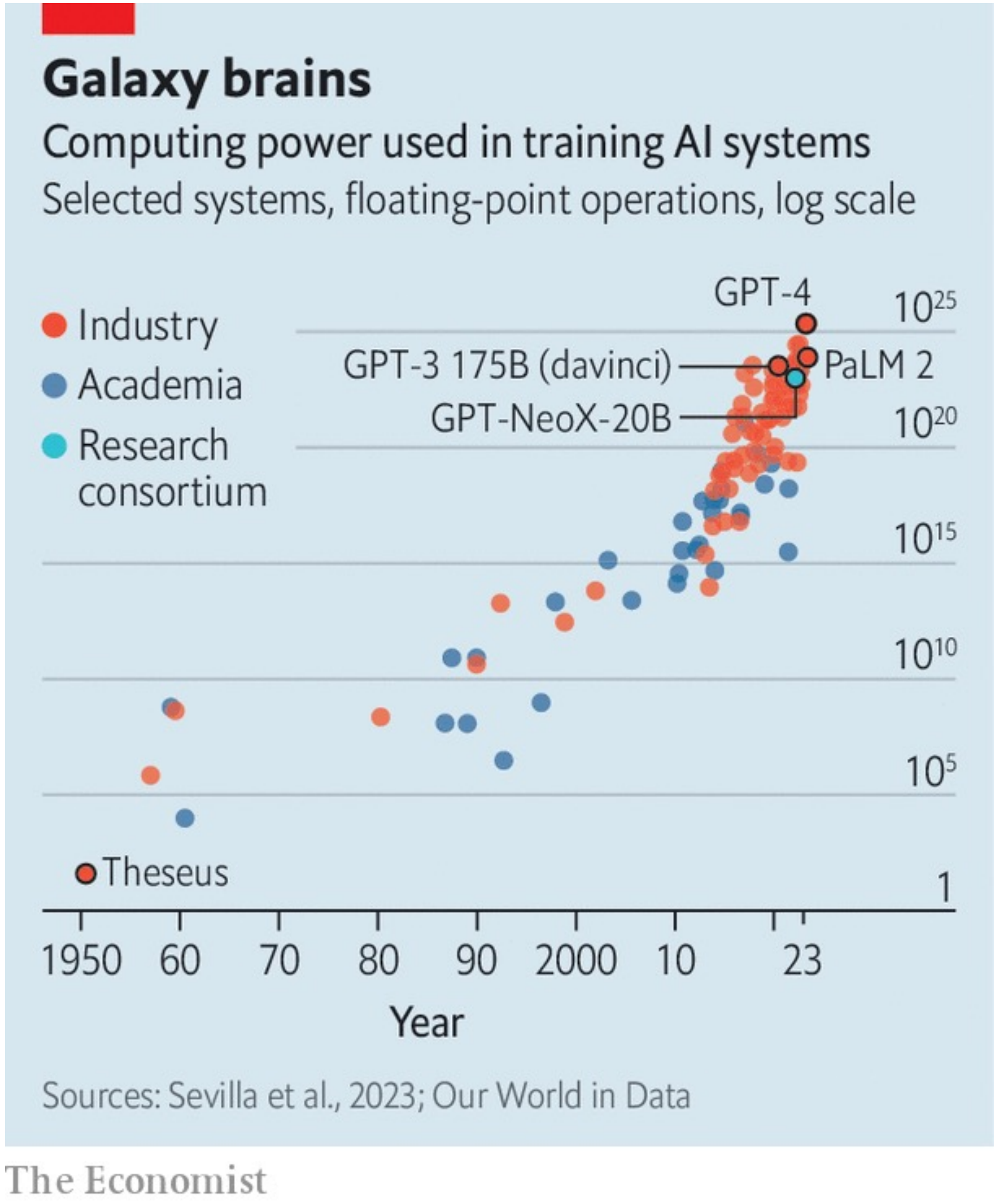
The problem with AI

The problem with AI

The cost is **huge**

Training ChatGPT-4 = 100M$ in electricity bill. And it still doesn't speak well!

Given the exponential demand, this is completely unsustainable.



Our Solution: Tensorize!

arXiv > cs > arXiv:1706.03762

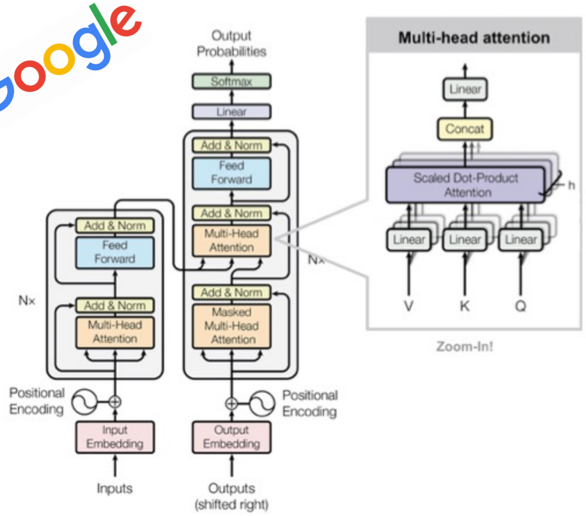
Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Our Solution: Tensorize!



Our Solution: Tensorize!

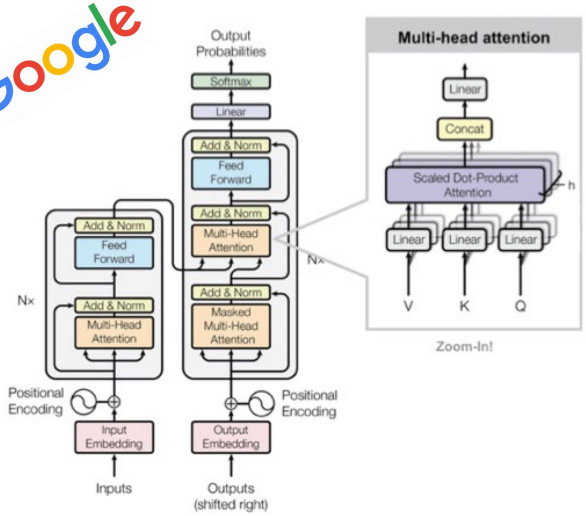

arXiv > cs > arXiv:1706.03762

Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

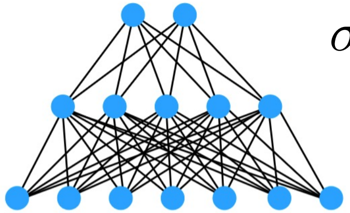
Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

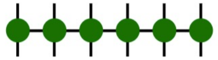


MPO Decomposition of Weigth Matrices

$$\sigma(Wx+b) \Rightarrow \text{TN}$$



$\approx$



See also Ma et al. arXiv:1906.09777 and Patel et al., arXiv:2208.02235


arXiv > cs > arXiv:1706.03762

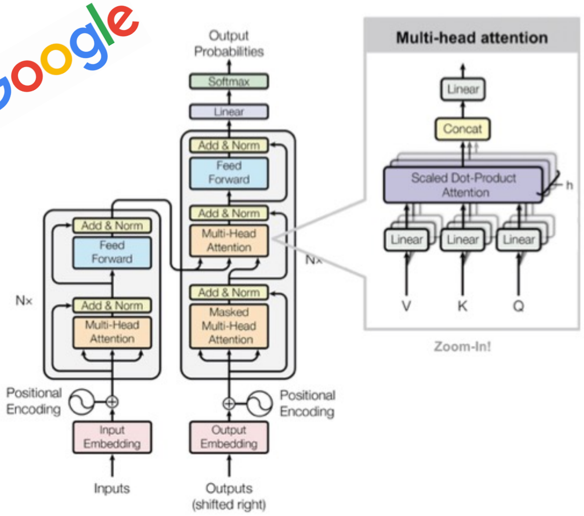
Computer Science > Computation and Language

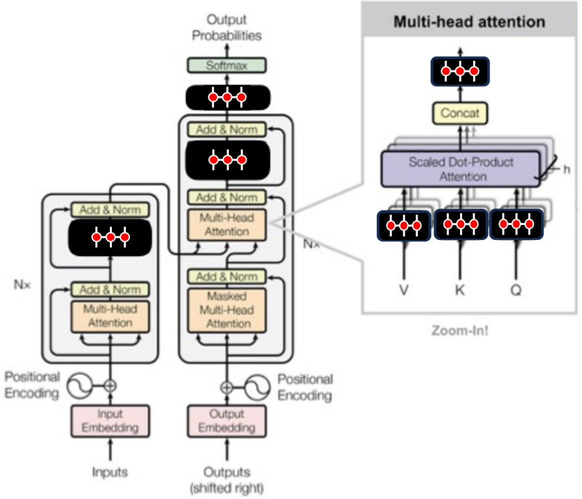
[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

Attention Is All You Need

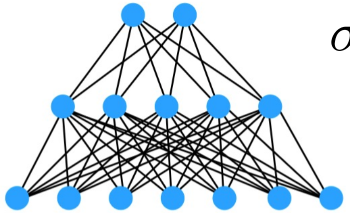
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Our Solution: Tensorize!



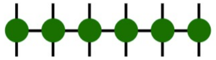


MPO Decomposition of Weigth Matrices



$$\sigma(Wx+b) \Rightarrow \text{TN}$$

$$\} \approx$$



See also Ma et al. arXiv:1906.09777 and Patel et al., arXiv:2208.02235

Our Solution: Tensorize!

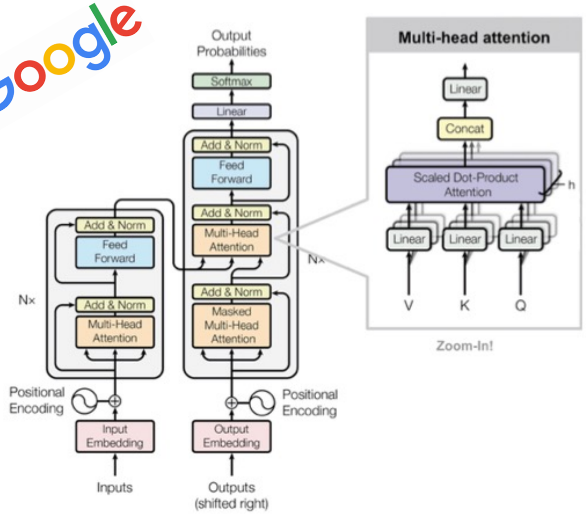
arXiv > cs > arXiv:1706.03762

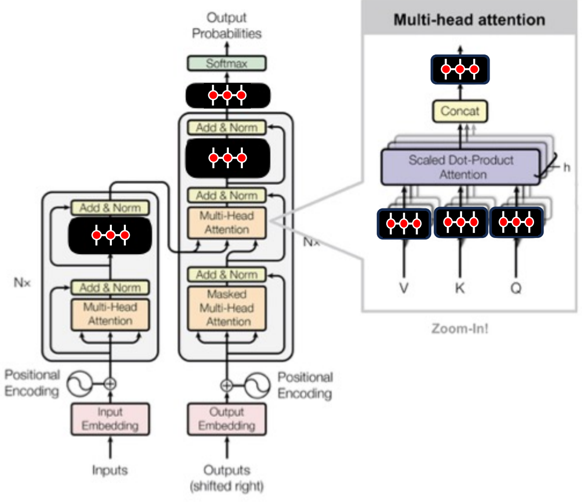
Computer Science > Computation and Language

[Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7)]

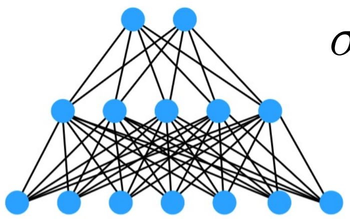
Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin



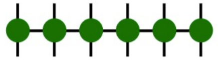


MPO Decomposition of Weigth Matrices



$$\sigma(Wx+b) \Rightarrow \text{TN}$$

$$\} \approx$$



See also Ma et al. arXiv:1906.09777 and Patel et al., arXiv:2208.02235

arXiv > cs > arXiv:2401.14109

Computer Science > Computation and Language

[Submitted on 25 Jan 2024]

CompactifAI: Extreme Compression of Large Language Models using Quantum-Inspired Tensor Networks

Andrei Tomut, Saeed S. Jahromi, Sukhbinder Singh, Faysal Ishtiaq, Cesar Muñoz, Prabdeep Singh Bajaj, Ali Elborady, Gianni del Bimbo, Mehrazin Alizadeh, David Montero, Pablo Martin-Ramiro, Muhammad Ibrahim, Oussama Tahiri Alaoui, John Malcolm, Samuel Mugel, Roman Orus

Large Language Models (LLMs) such as ChatGPT and LlaMA are advancing rapidly in generative Artificial Intelligence (AI), but their immense size poses significant challenges, such as huge training and inference costs, substantial energy demands, and limitations for on-site deployment. Traditional compression methods such as pruning, distillation, and low-rank approximation focus on reducing the effective number of neurons in the network, while quantization focuses on reducing the numerical precision of individual weights to reduce the model size while keeping the number of neurons fixed. While these compression methods have been relatively successful in practice, there's no compelling reason to believe that truncating the number of neurons is an optimal strategy. In this context, this paper introduces CompactifAI, an innovative LLM compression approach using quantum-inspired Tensor Networks that focuses on the model's correlation space instead, allowing for a more controlled, refined and interpretable model compression. Our method is versatile and can be implemented with – or on top of – other compression techniques. As a benchmark, we demonstrate that CompactifAI alone enables compression of the LlaMA-2 7B model to only 30% of its original size while recovering over 90% of the original accuracy after a brief distributed retraining.

Tensorizing LlaMA2 7B

Tensorizing LlaMA2 7B

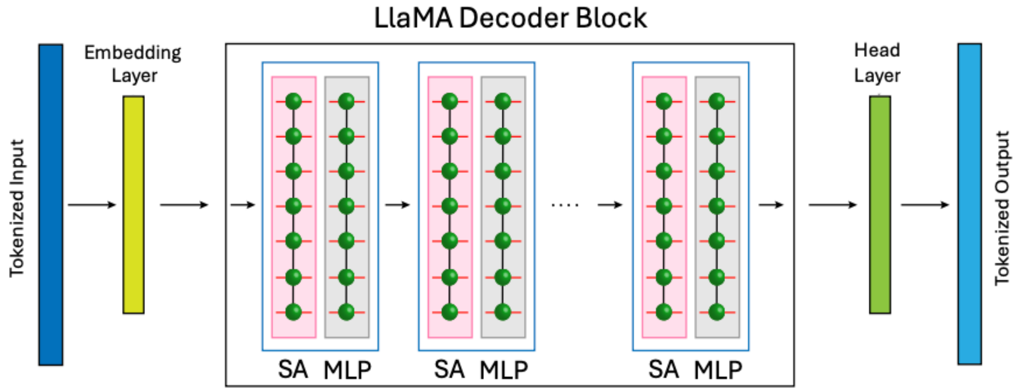
Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	**Model architecture:** **Pretraining Tokens:** 2 Trillion **Context Length:** 4096	**Data collection for helpfulness and safety:** **Supervised fine-tuning:** Over 100,000 **Human Preferences:** Over 1,000,000
13B		
70B		

Tensorizing LlaMA2 7B

Llama 2		
MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B 13B 70B	Model architecture: Pretraining Tokens: 2 Trillion Context Length: 4096	Data collection for helpfulness and safety: Supervised fine-tuning: Over 100,000 Human Preferences: Over 1,000,000

MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model

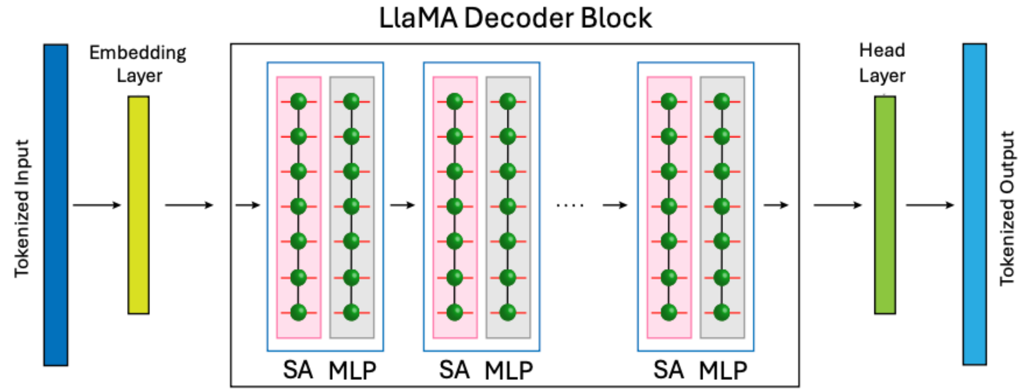


Tensorizing LlaMA2 7B

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture: Pretraining Tokens: 2 Trillion Context Length: 4096	Data collection for helpfulness and safety: Supervised fine-tuning: Over 100,000 Human Preferences: Over 1,000,000
13B		
70B		

MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



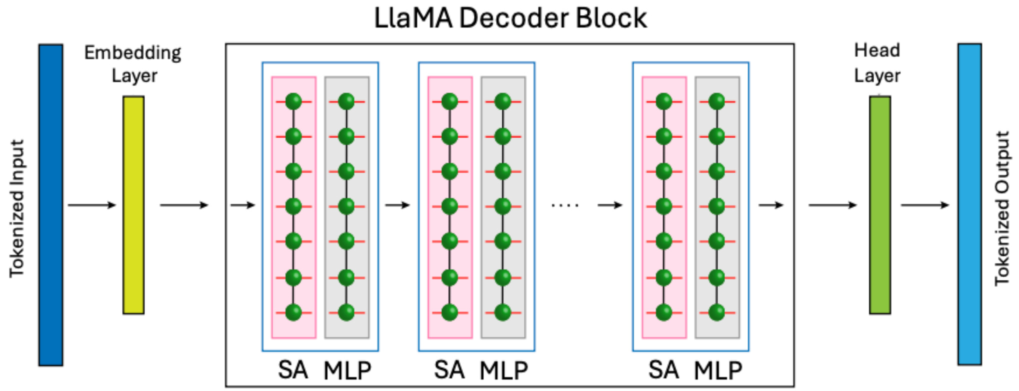
MPOs with bond dimension ≈ 200
Mild quantization from Float32 to Float16

Tensorizing LlaMA2 7B

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture: Pretraining Tokens: 2 Trillion Context Length: 4096	Data collection for helpfulness and safety: Supervised fine-tuning: Over 100,000 Human Preferences: Over 1,000,000
13B		
70B		

MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



MPOs with bond dimension $\approx$ 200
Mild quantization from Float32 to Float16

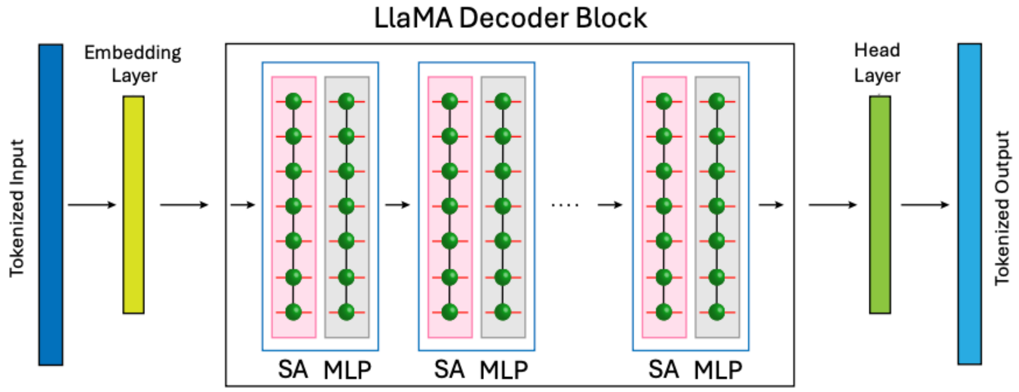
85% of memory reduction
(from 24Gb to 3.7Gb)

Tensorizing LlaMA2 7B

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

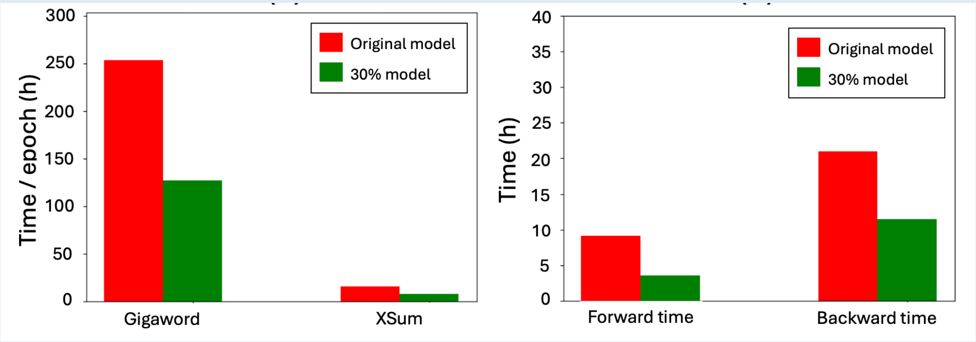
MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



MPOs with bond dimension ≈ 200
Mild quantization from Float32 to Float16

85% of memory reduction
(from 24Gb to 3.7Gb)

Retraining & F/W pass (inference) times
(Gigaword & Xsum databases)

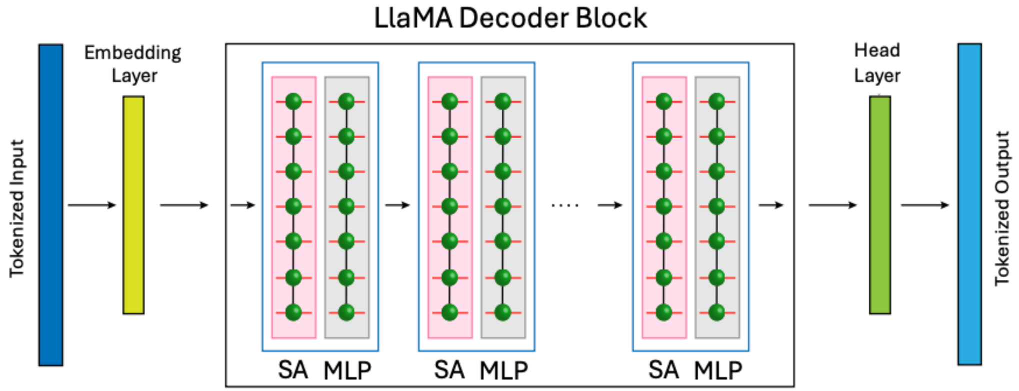


Tensorizing LlaMA2 7B

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

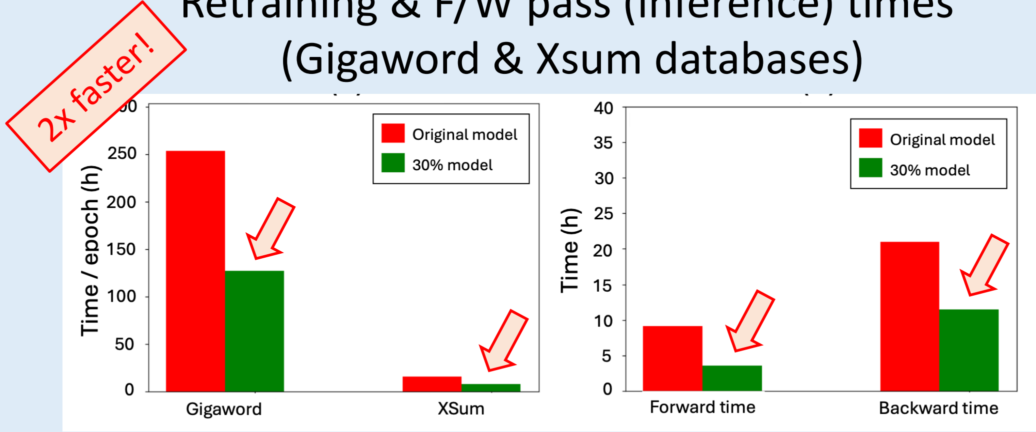
MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



MPOs with bond dimension $\approx$ 200
Mild quantization from Float32 to Float16

85% of memory reduction
(from 24Gb to 3.7Gb)

Retraining & F/W pass (inference) times
(Gigaword & Xsum databases)

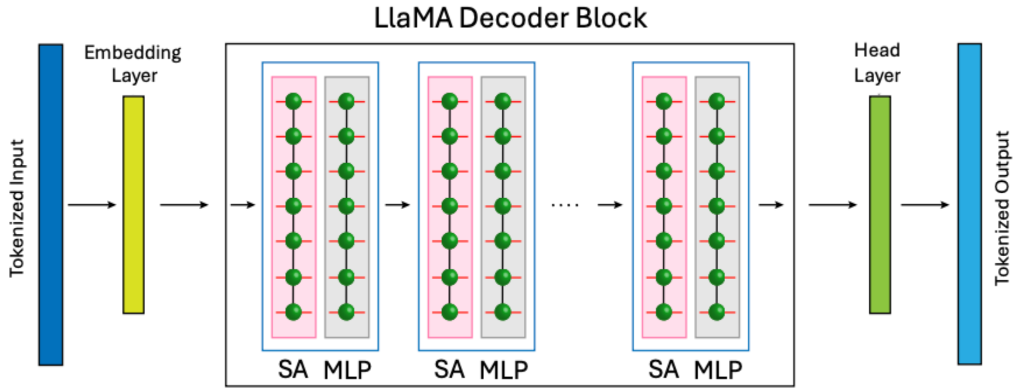


Tensorizing LlaMA2 7B

Llama 2

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture: Pretraining Tokens: 2 Trillion Context Length: 4096	Data collection for helpfulness and safety: Supervised fine-tuning: Over 100,000 Human Preferences: Over 1,000,000
13B		
70B		

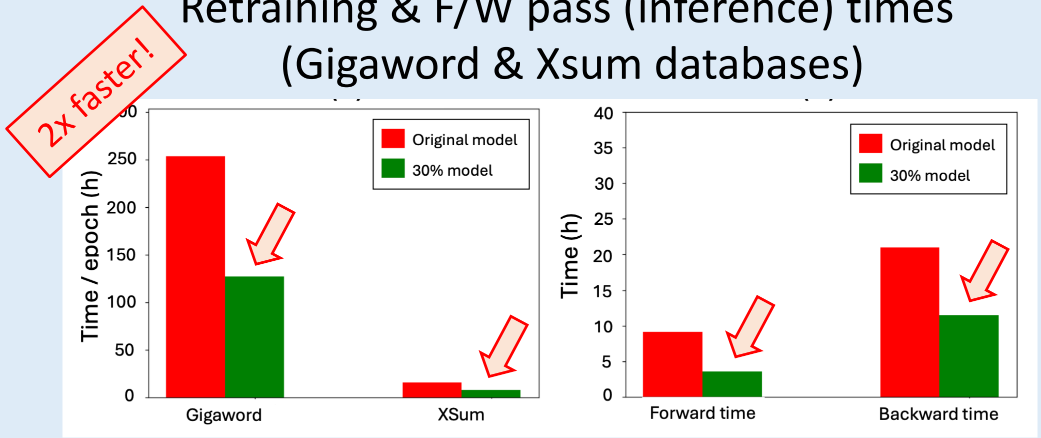
MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



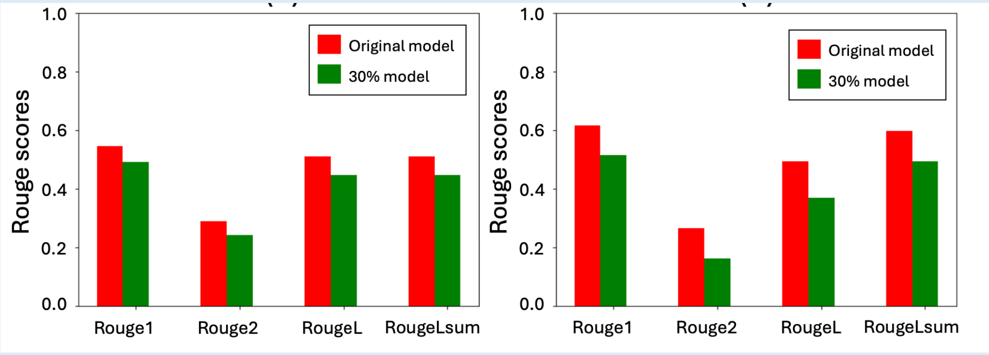
MPOs with bond dimension $\approx$ 200
Mild quantization from Float32 to Float16

85% of memory reduction
(from 24Gb to 3.7Gb)

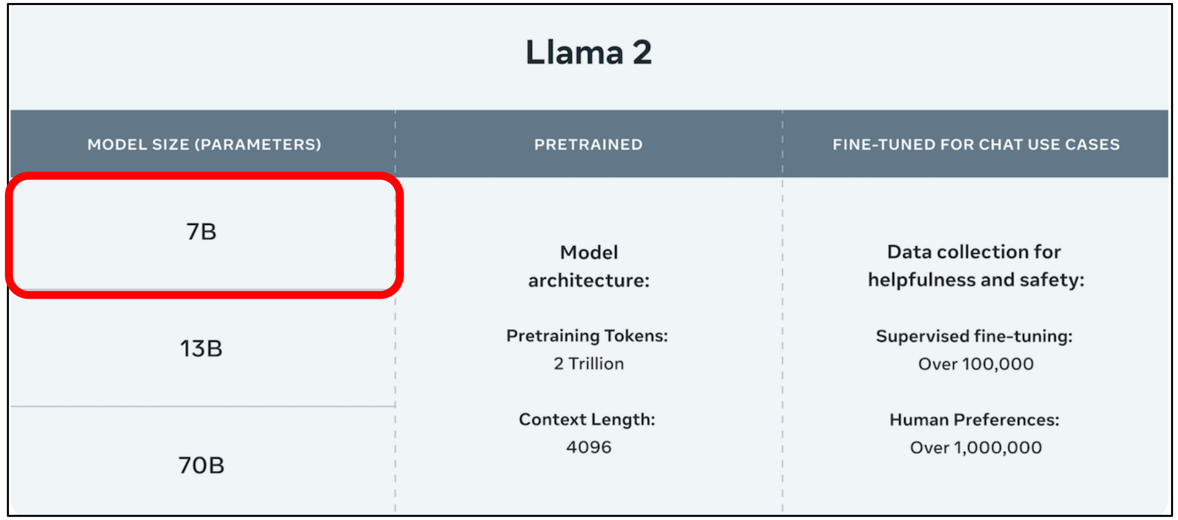
Retraining & F/W pass (inference) times
(Gigaword & Xsum databases)



Accuracy Scores (Gigaword & Xsum)

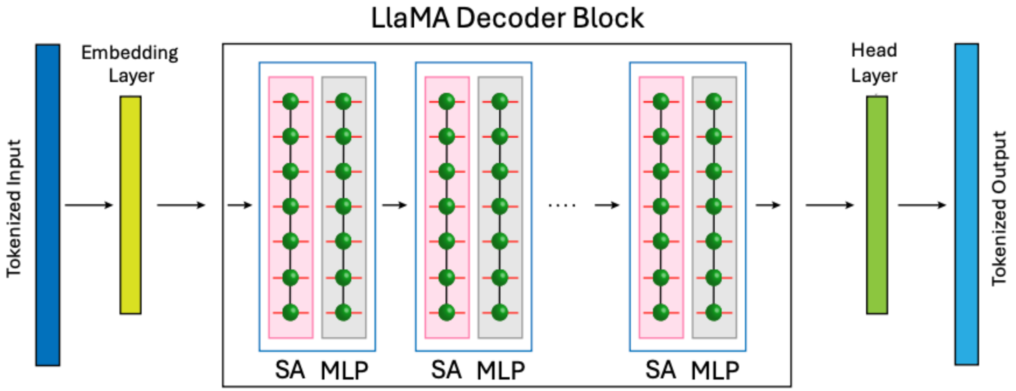


Tensorizing LlaMA2 7B



Llama 2		
MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

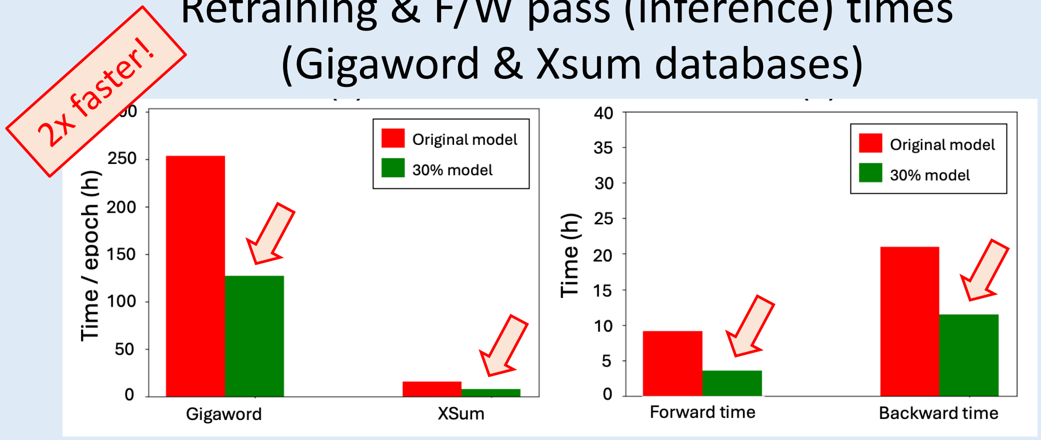
MPOs at Self Attention and Multi-layer Perceptron Layers for a pre-trained model



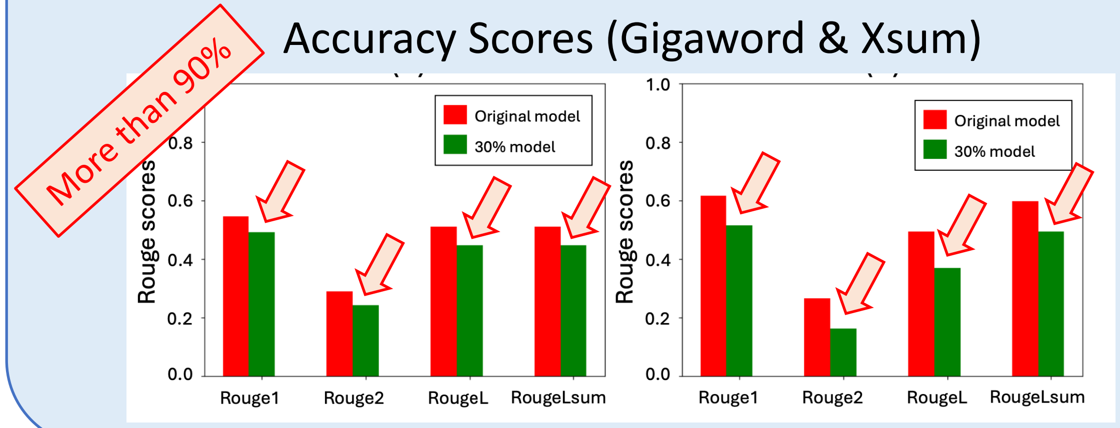
MPOs with bond dimension $\approx$ 200
Mild quantization from Float32 to Float16

85% of memory reduction
(from 24Gb to 3.7Gb)

Retraining & F/W pass (inference) times
(Gigaword & Xsum databases)



Accuracy Scores (Gigaword & Xsum)



What does all this imply?

- We compressed LlaMA2 7B LLM using TNs down to 15% size keeping more than 90% of accuracy.
- This saves a huge amount of energy costs in AI, and also in training and inference times (2x faster).
- It is perfectly compatible with other “standard” compression techniques in AI (quantization, distillation, pruning, LoRa…). Also more controllable and interpretable than these approaches.
- We had to come up with distributed training a TN over an large amount of data among several GPUs.
- The compression also allows to deploy LLMs on premises, without cloud access.
- What about more complex TN structures and compression & training techniques?

THANK YOU!

Donostia International Physics Center

roman.orus@dipc.org

MULTIVERSE
COMPUTING

www.multiversecomputing.com

roman.orus@multiversecomputing.com