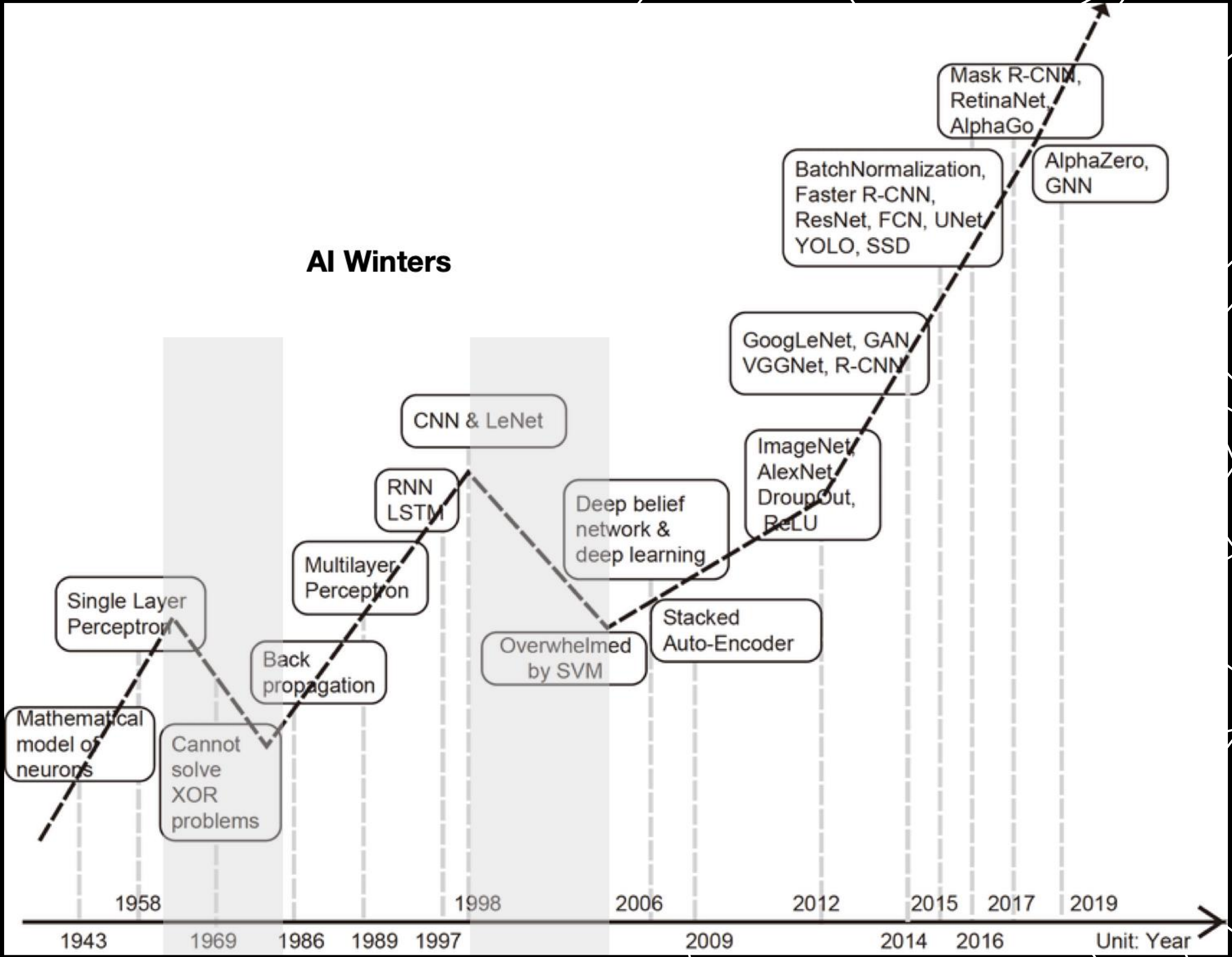
AI SUSTAINABILITY

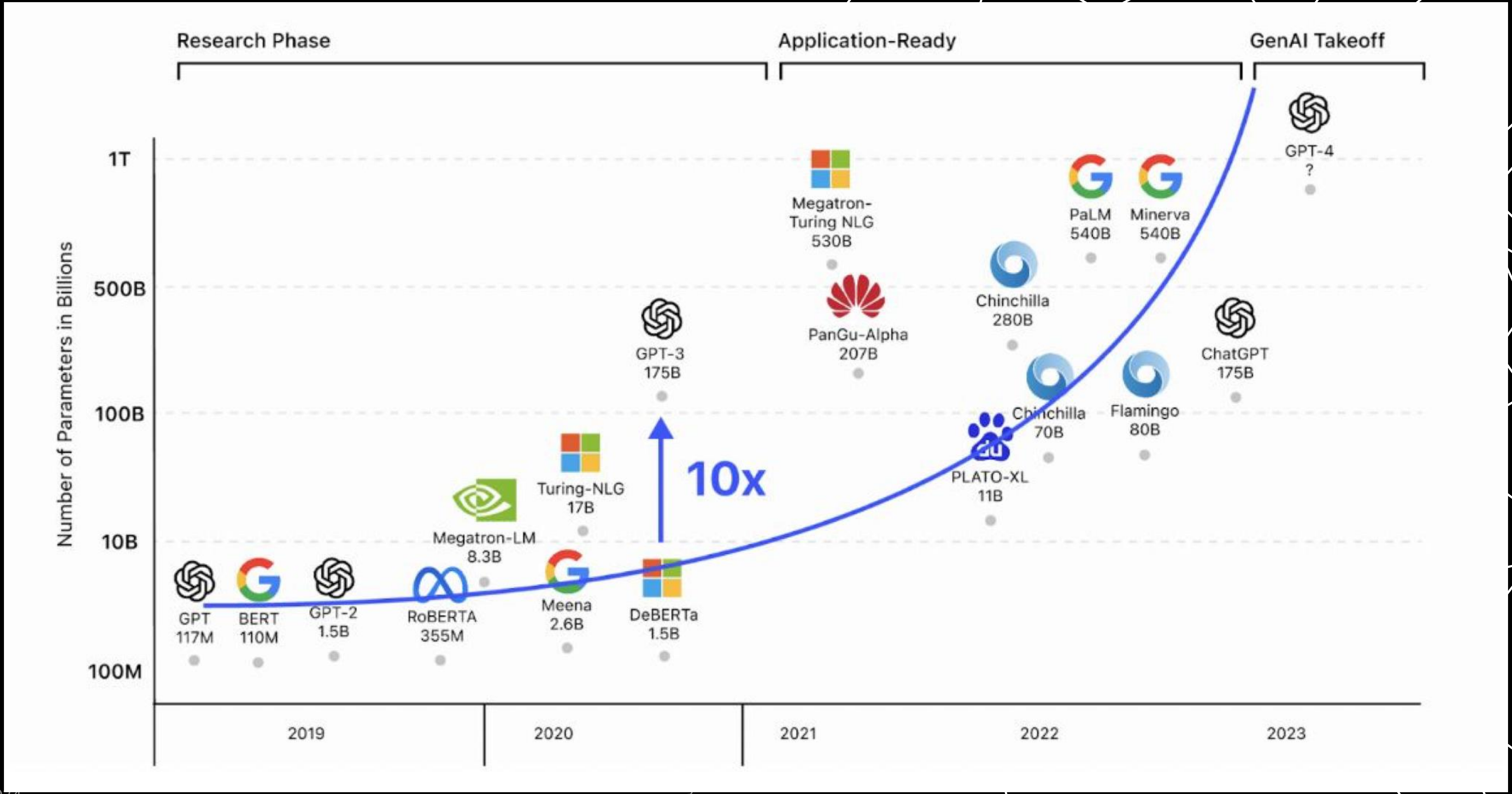
Sofia Vallecorsa – CERN

December 11th, 2024

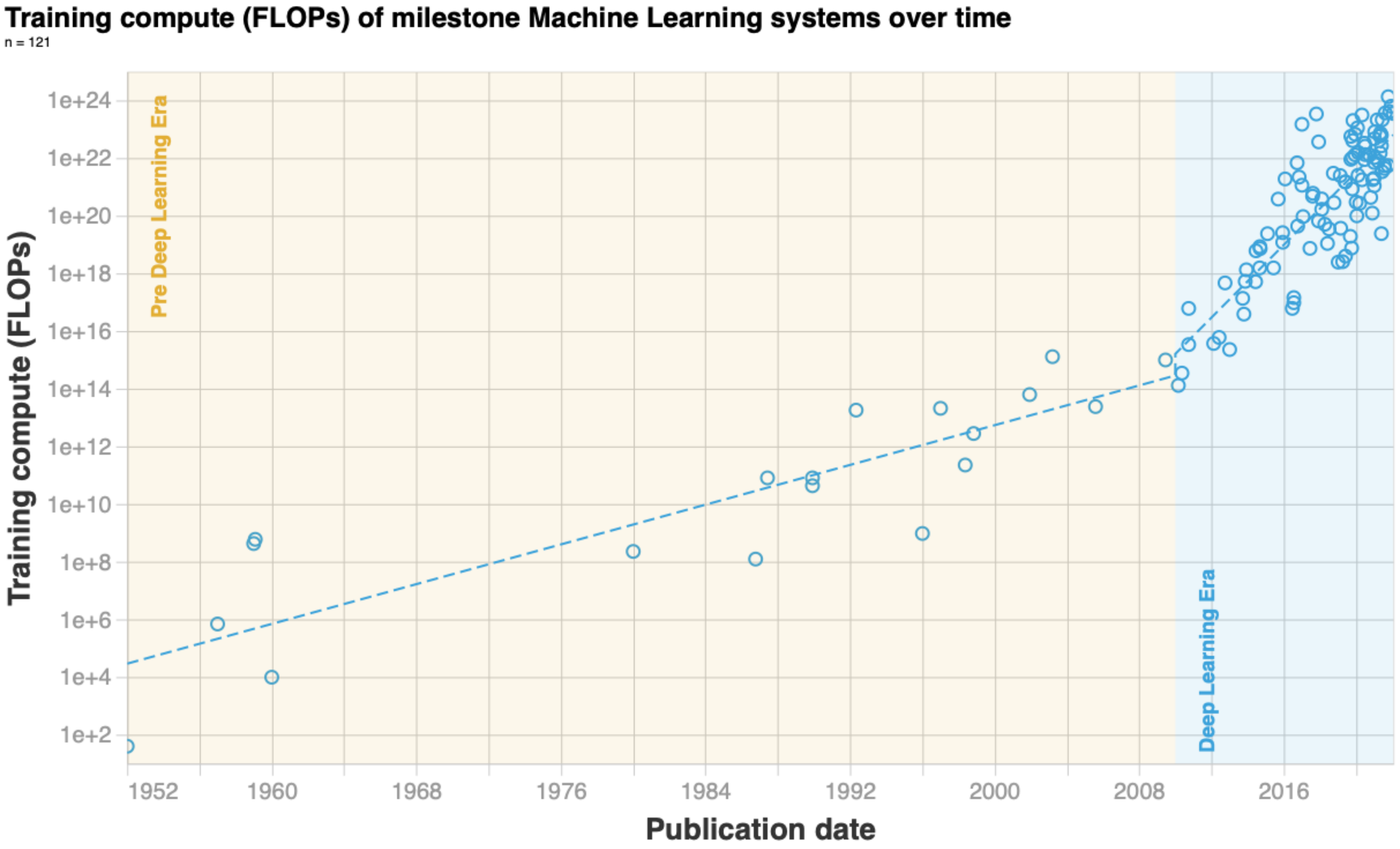
A BIT OF HISTORY



THEN … TAKEOFF

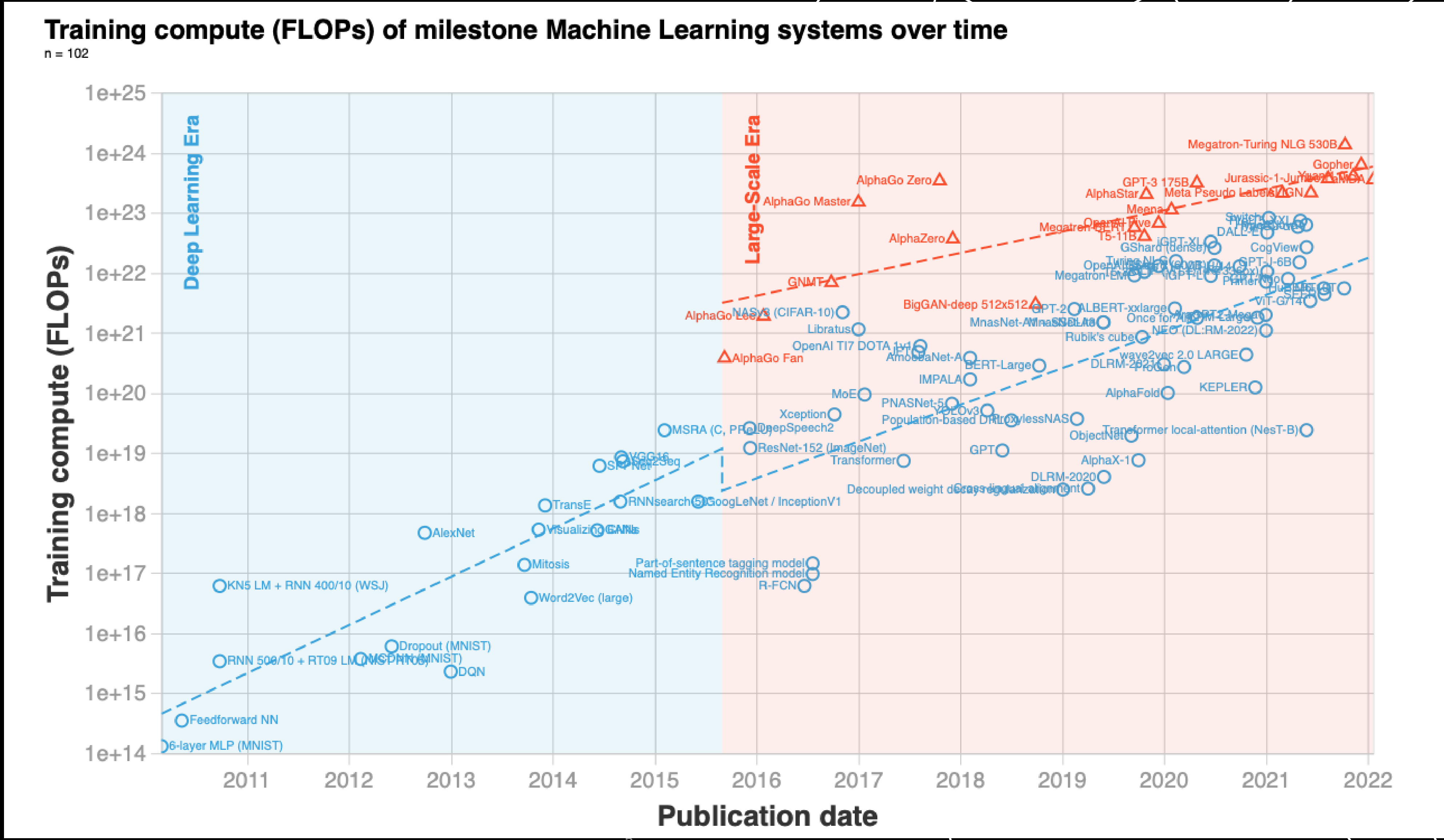


COMPUTING COST



(Sevilla et al., 2022)

COMPUTING COST

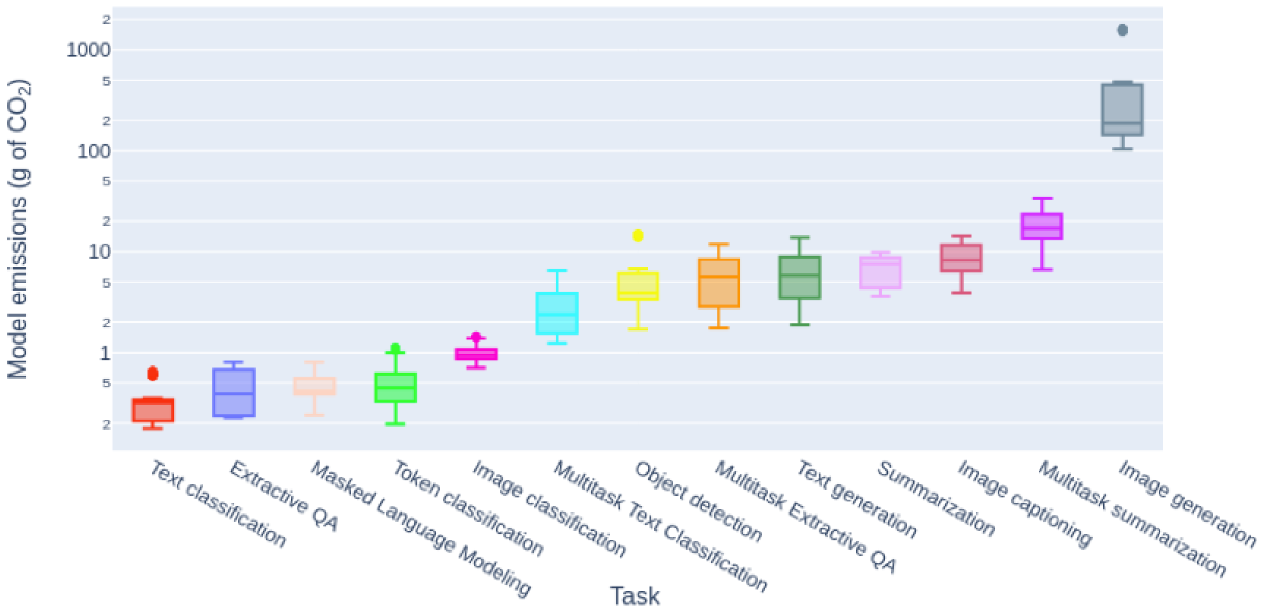


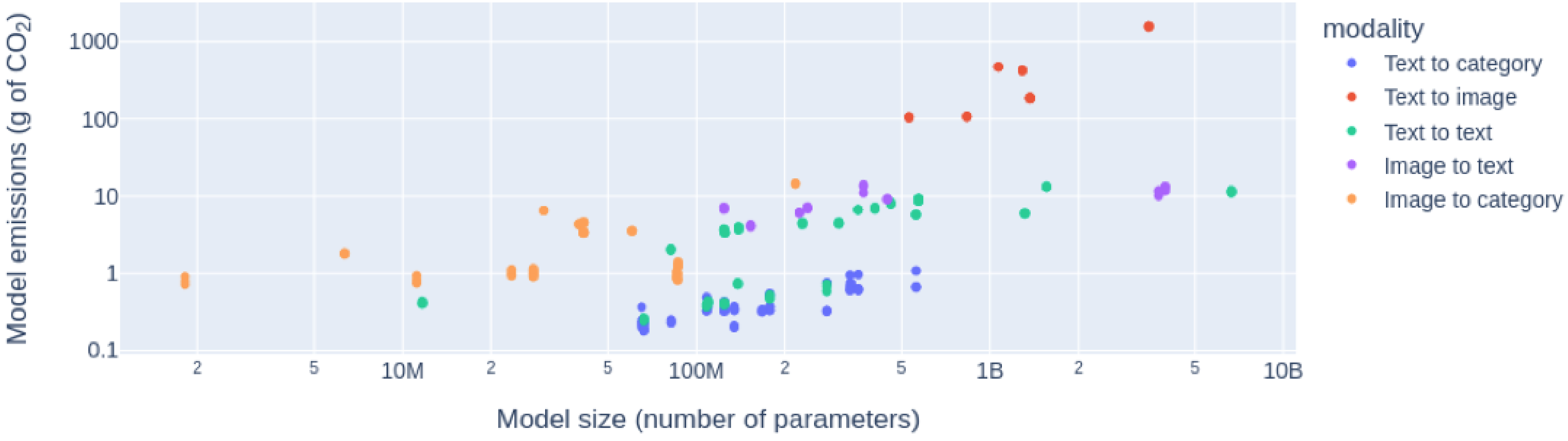
METRICS ARE AVAILABLE

AI is getting more expensive in terms of resources and carbon footprint.

But what does it mean exactly in terms of sustainability?

What about AI in HEP?





Power Hungry Processing: Watts Driving the Cost of AI Deployment? arXiv:2311.16863v

DEEP LEARNING IS IN PRODUCTION IN RUN3

Machine Learning since LEP years

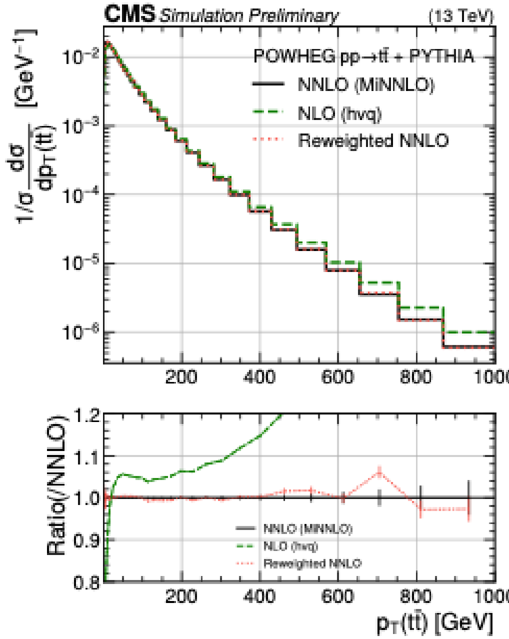
Mostly for classification & regression

Since early 2000s a multiplicity of applications through Deep Learning...

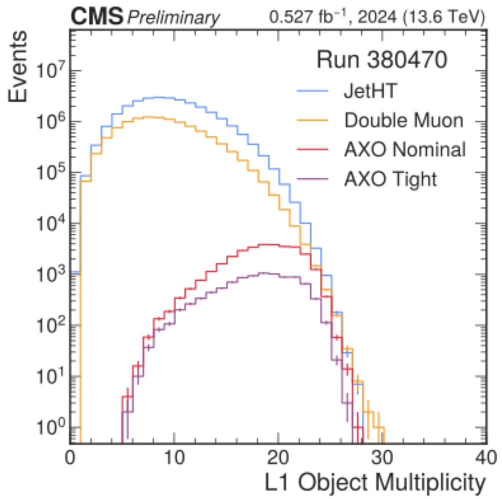
Many are in production for Run 3!

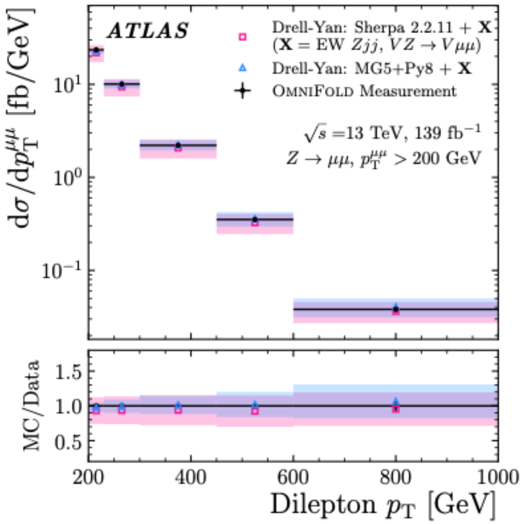
Today we also are interested in LLMs in all their shapes and forms

MC reweighting [1]



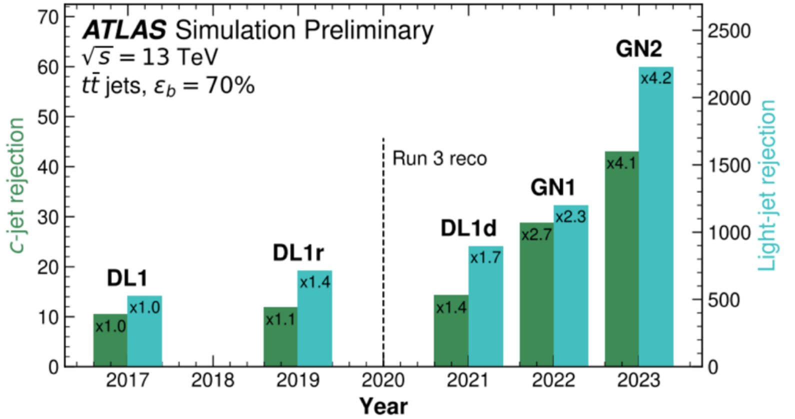
Trigger [3]

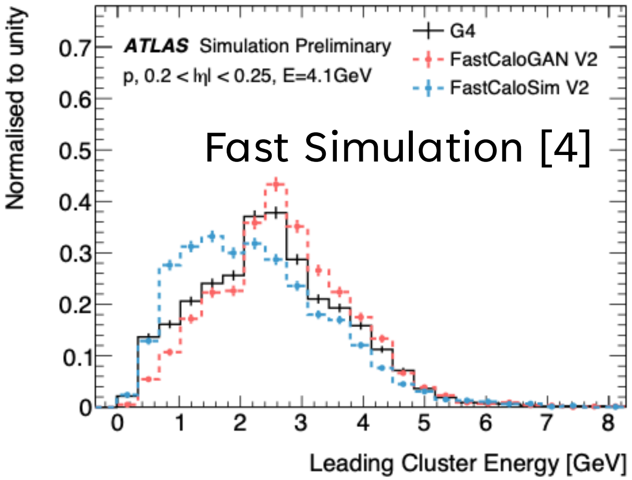




Unfolding [2]

Tagging [7]





Fast Simulation [4]

PID in ALICE O2 [5]

CMS Track selection [6]

...

WHAT IS SPECIAL ABOUT DL IN HEP?

In general highly optimized models

Out of the box models rarely worked

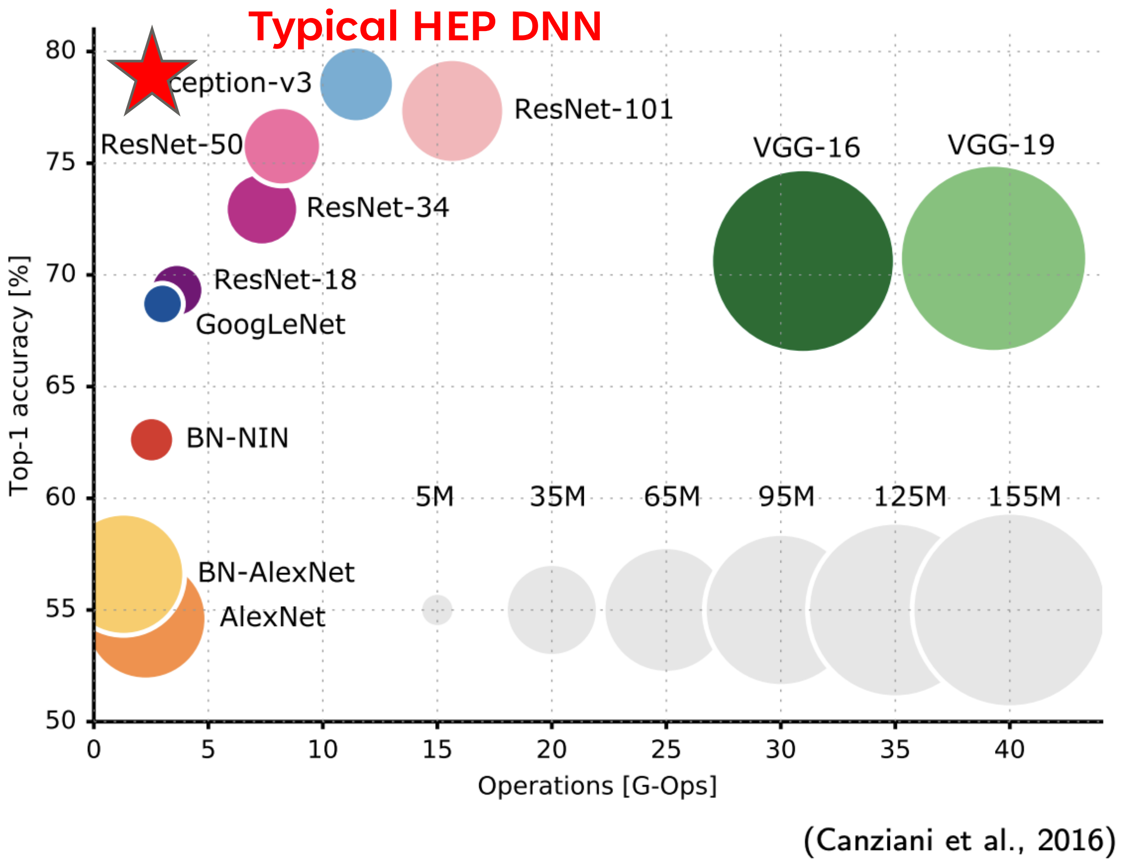
The broad range of applications leads to different computing requirements

Ex. ML in Real Time environment

Constraints on Latency

Constraints on Model Complexity

Constraints on the quality of data available



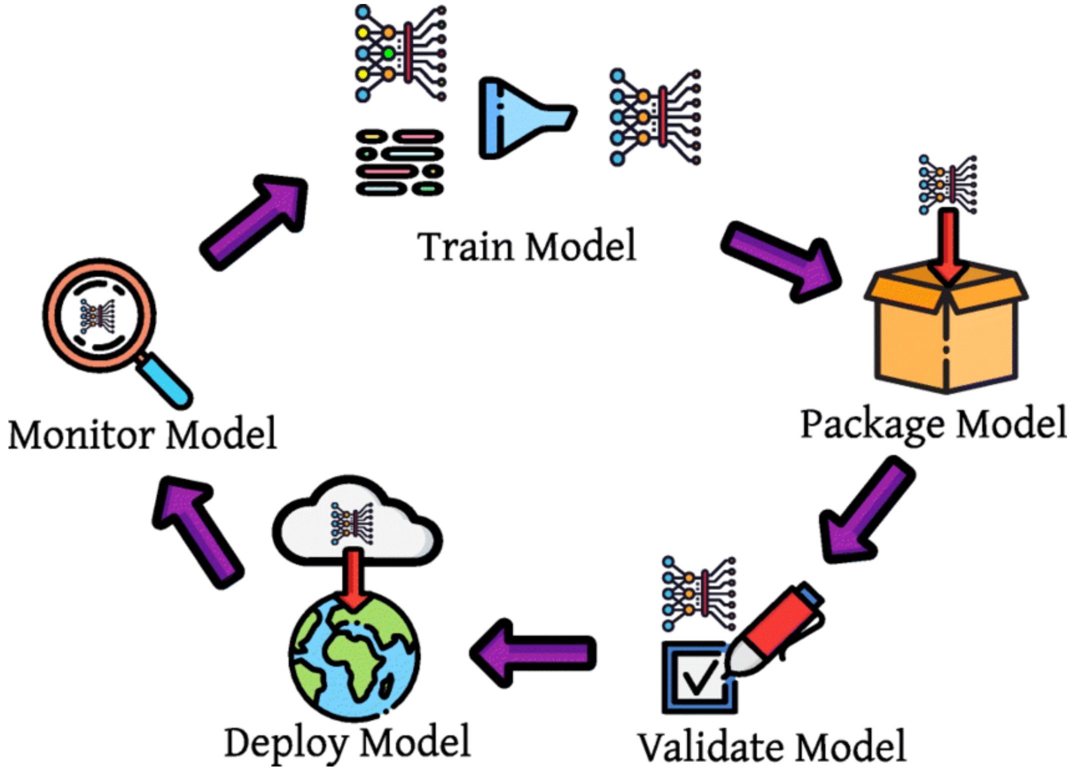
This plot (and/or similar) have been shown for years by many people!

ENERGY CONSUMPTION IN AI LIFECYCLE

AI energy footprint needs to be assessed in the different steps of ML lifecycle

- Including cost for data gathering, storage and pre-processing

What about comparison to traditional techniques AI replaces?



Ex. Weather forecasting

1. Numerical models require O(hours) for one 10 days forecast

2. ECMWF model takes 2.5 min on a single GPU
Training takes 1 week using 64 A100 GPU
.. with 50 ensemble models (https://arxiv.org/pdf/2406.01465)

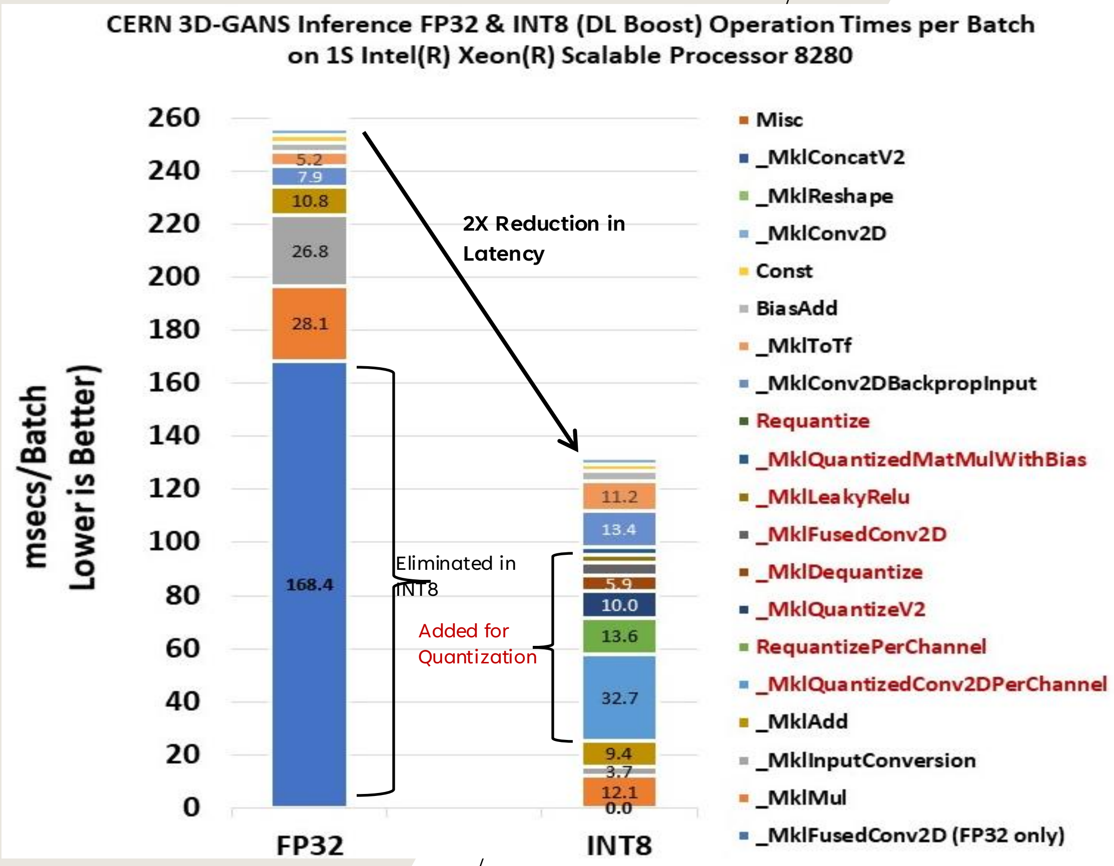
3. **Pangu-Weather (SoA) reports 11% better forecasting accuracy while being 10000x faster**
(https://arxiv.org/abs/2211.02556)

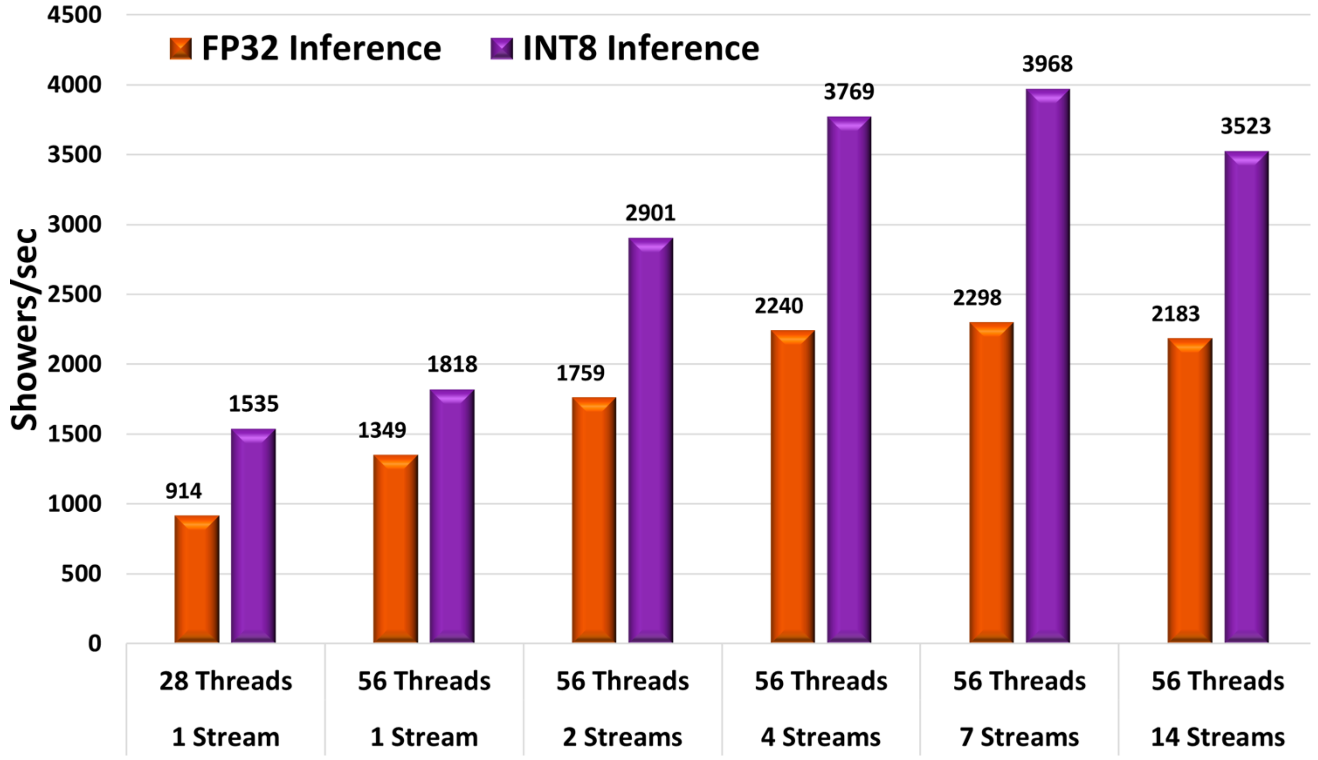
BACK IN 2021: FASTER THEN MONTE CARLO (...ON CPU!)

F. Rehm, ICPRAM2021
in collaboration with Intel

FP32: 3DGAN is **38000x faster** than Monte Carlo
INT8: quantized 3DGAN is **68000x faster** than Monte Carlo

Post training quantization (INT 8):



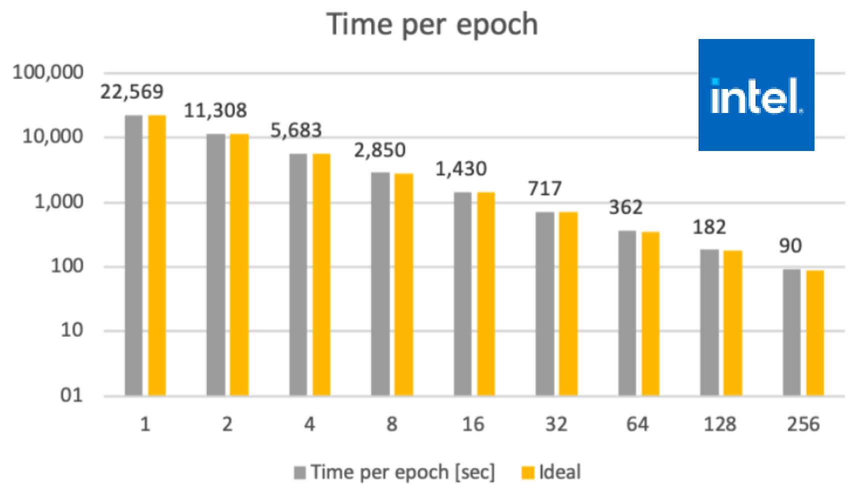


BACK IN 2021: OPTIMIZED TRAINING

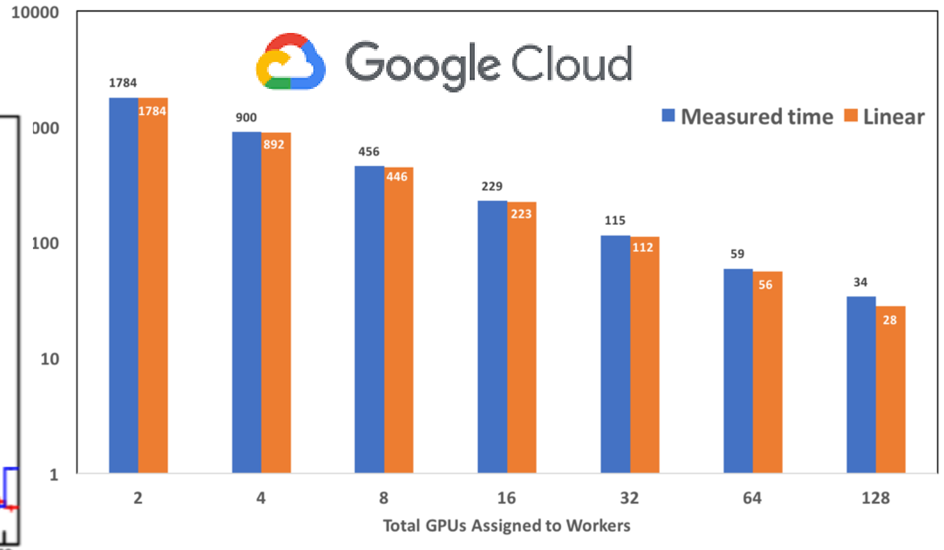
Training 3DGAN (3M parameters) takes ~7 days on a GPU

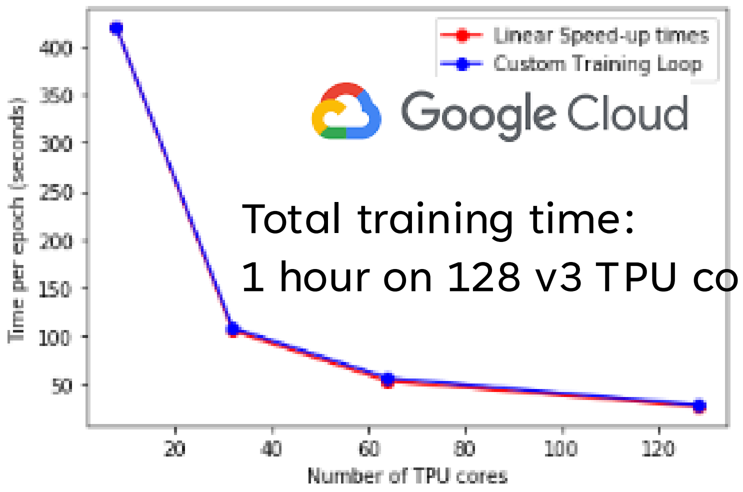
- Distributed training is essential
- Keep physics under control
- Optimise costs

Total training time: 3 hours on 256 Intel Xeons

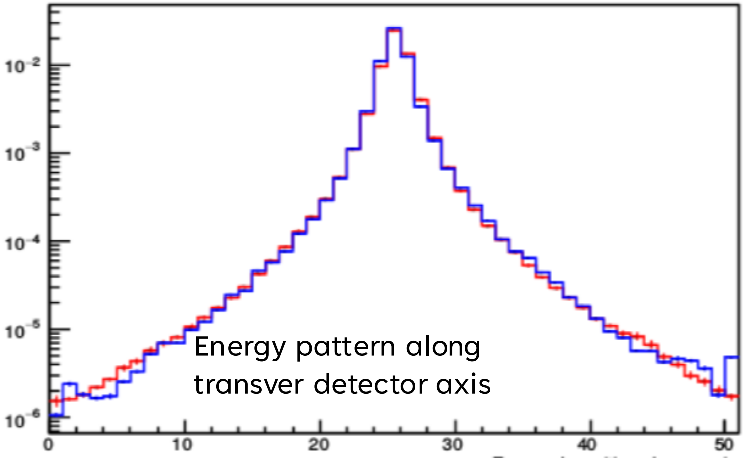


Total training time:
1 hour on 128 V100 GPUs





Total training time:
1 hour on 128 v3 TPU cores



Energy pattern along
transver detector axis

PATH TOWARD ENERGY EFFICIENT AI

STRATEGIC

Optimise **use case definition**

Optimise **integration** with existing software

Estimate classical tools **replacement savings**

Actively contribute to **existing** green **intiatives beyond HEP**

HARDWARE

Improve usage efficiency of available h/w

AI models are based on a few frameworks: optimising them impacts all use cases

Introduce **new h/w technologies** (dedicated accelerators, Quantum Computing…)

DEPLOYMENT

Optimise across the **AI lifecycle**

Optimise workloads **definition, scheduling**, …

Data centers choice: centralisation allows better resource managament

AI ARCHITECTURES

Improved/ compactified **data representation and computational graphs**

Foundation models

New **approaches to training**

Neural Architecture Search (NAS)

OPTIMISING DEPLOYMENT

Adoption of **cloud-based solutions** that offer better energy efficiency through optimized resource management.

Layers of the solutions

CLOUD NATIVE SUSTAINABILITY

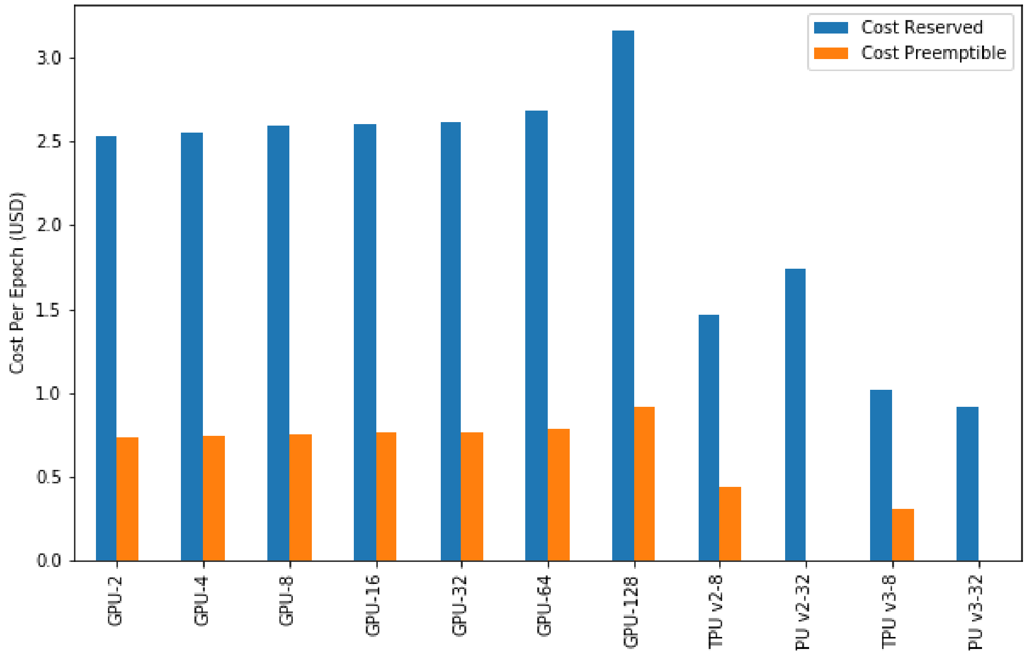
When considering solutions complimentary to the three foundations of sustainable cloud systems, we can divide solution considerations into three general areas:

1. Which data center to use, if there are multiple options available.
2. Where to place the workload once a data center is chosen.
3. How to manage the resources on the node allocated for a workload to run on.

All of these elements can be investigated further individually.

AREA	GOAL	EFFORTS
Multi Data Centers	Intelligently choosing which data center to schedule on according to environmental factors such as whether the region is powered by renewables, the region's Marginal Emissions Rate, Power Usage Effectiveness (PUE), time of day, etc.	Cluster Management
Within Data Center	Scheduling effectively according to workload, availability, and urgency of workload	Power Management, K8S Scheduler Plugin
Within a node	Optimizing resources to handle workload specifications (which may include performance parameters) while minimizing resource consumption	Node Tuning, Pod Scaling

NB: There is ongoing work along these lines at CERN by R. Rocha & his team



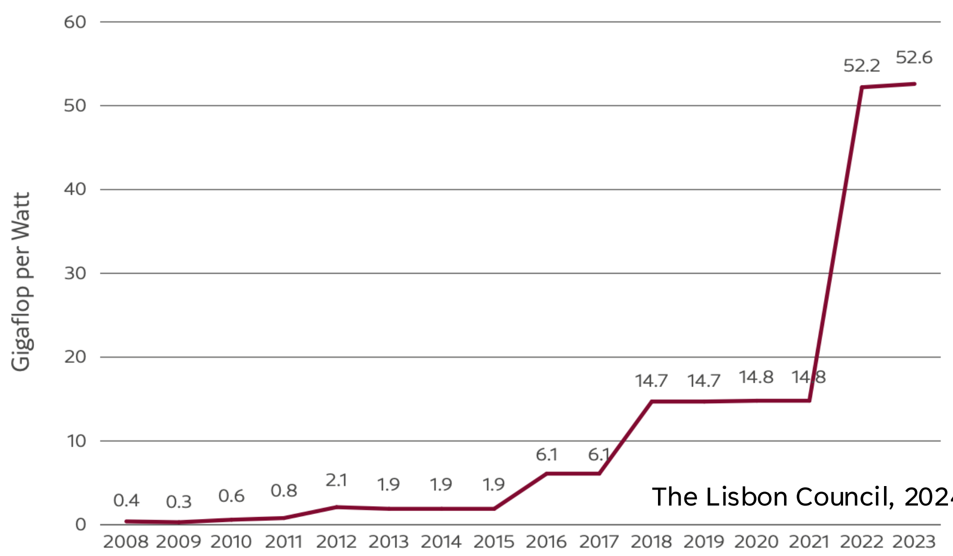
Our old study focused on cost .. Not a proxy for energy consumption!

Cardoso, Renato, et al. "Accelerating GAN training using highly parallel hardware on public cloud." EPJ Web of Conferences. Vol. 251. EDP Sciences, 2021.

INTRODUCING ADVANCED HARDWARE TECHNOLOGIES

New hardware is more efficient
(but we need to make sure AI platforms make the best out of it!)

Figure 2: Energy Efficiency of the Fastest Supercomputer in Gigaflop per Watt



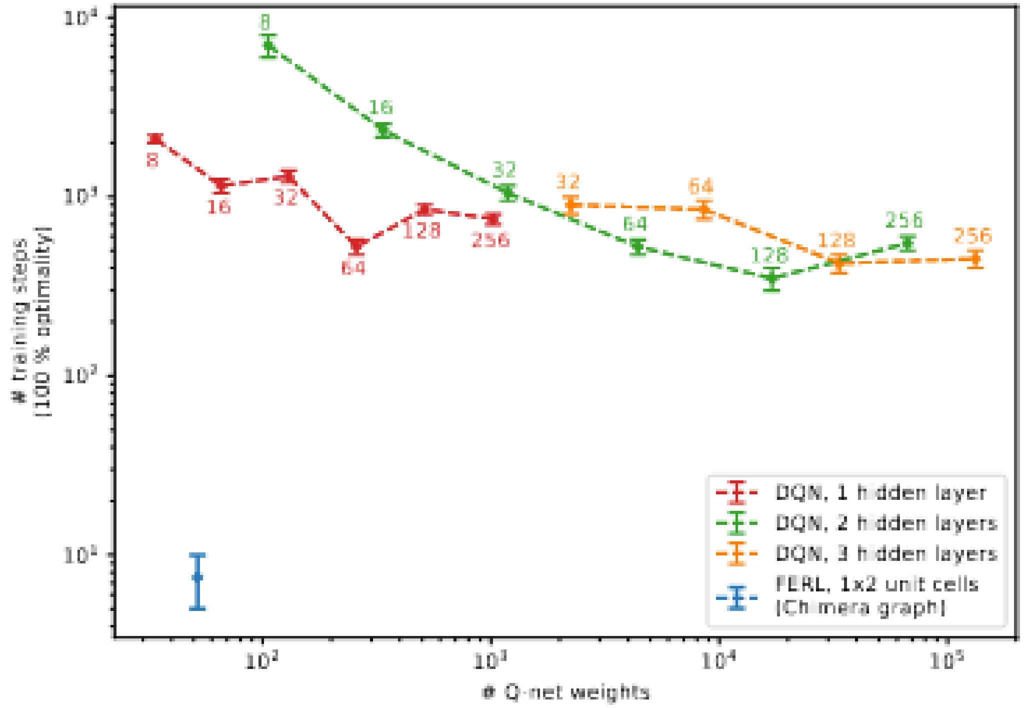
The Lisbon Council, 2024

New technologies can bring orders of magnitude improvements!

QC, Neuromorphic, Edge ...

Quantum computing accelerates the training of a classical RL agent

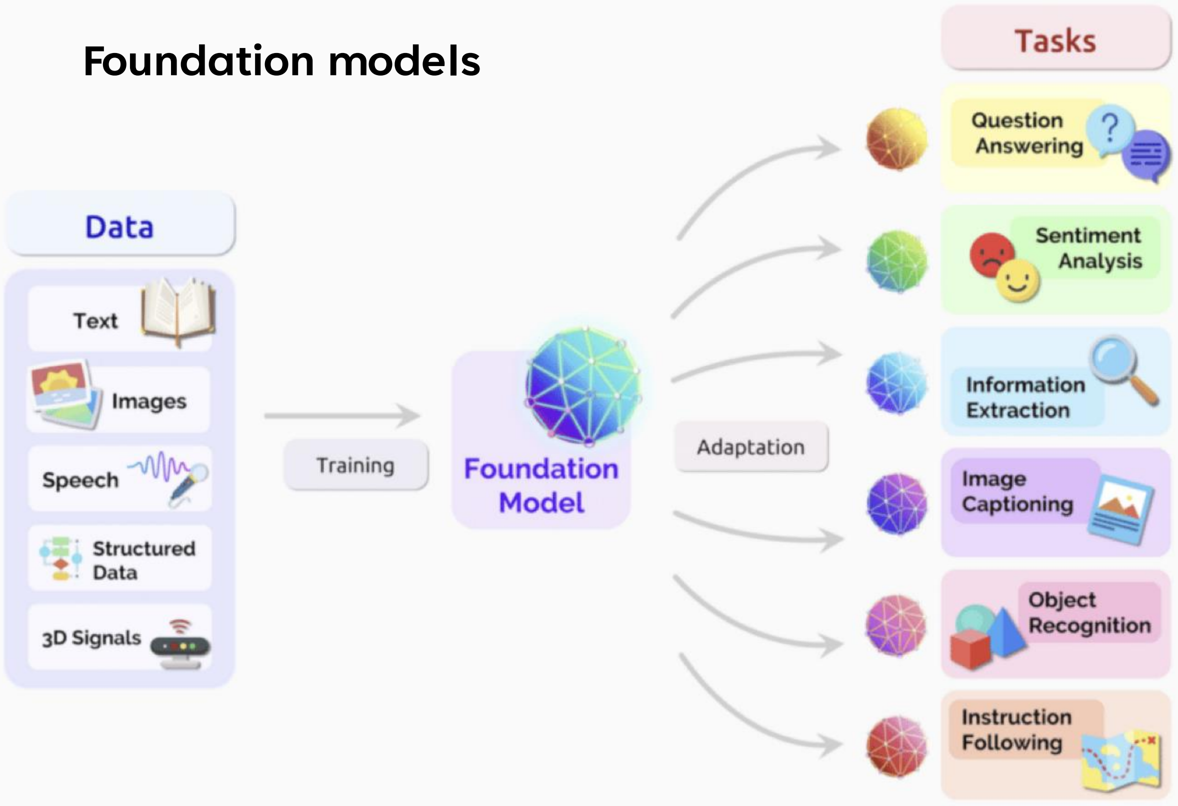
It could be used today!

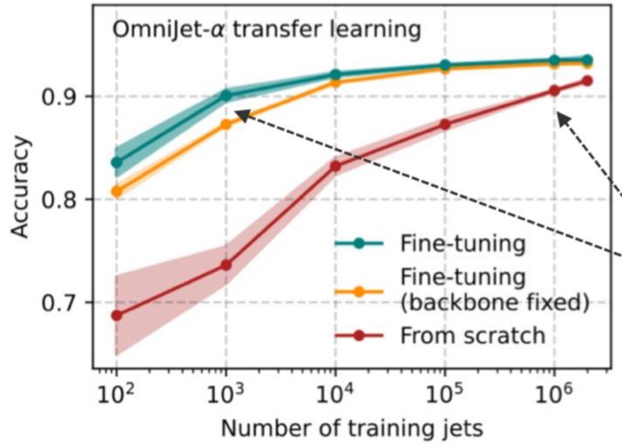


M. Schenk, et al., **Hybrid actor-critic algorithm for quantum reinforcement learning at CERN beam lines.** arXiv:2209.11044

A NEW APPROACH TO AI AND NEW TRAINING STRATEGIES

Foundation models





Anna Hallin et al.
arxiv: 2403.05618

Pre-trained model requires only 1000 training events to reach the same accuracy level that the "from scratch" model reaches with 1M events

Improved training techniques
https://arxiv.org/pdf/2307.00368

Table 1: Comparison of accuracy and energy consumption achieved with standard training (ST) and our energy-aware method (*EAT*).

	GTSRB				CIFAR-10				CelebA			
	ResNet18		VGG16		ResNet18		VGG16		ResNet18		VGG16	
	ST	*EAT*	ST	*EAT*	ST	*EAT*	ST	*EAT*	ST	*EAT*	ST	*EAT*
Accuracy	0.91	0.93	0.90	0.89	0.92	0.90	0.91	0.88	0.76	0.78	0.77	0.78
E. ratio	0.76	0.55	0.69	0.63	0.73	0.61	0.67	0.53	0.68	0.63	0.63	0.54
E. decrease%	-	**27.63**	-	**8.69**	-	**16.43**	-	**20.89**	-	**7.35**	-	**14.28**

SUSTAINABLE AI THROUGH A MULTI-TIERED APPROACH

AI is quickly becoming a major workload for HEP

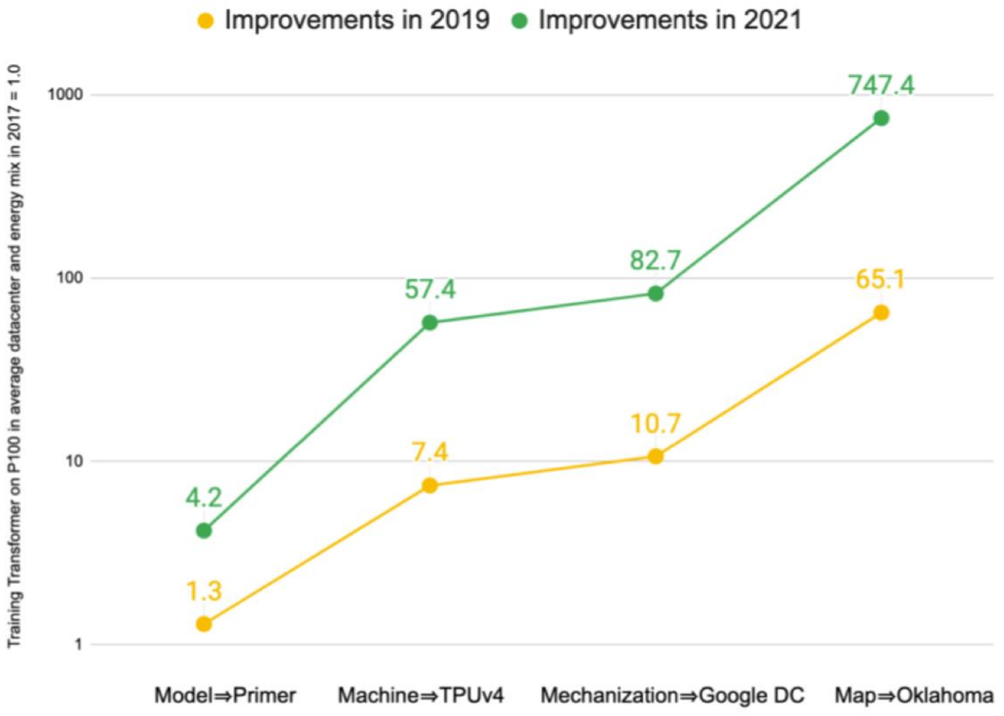
AI Energy sustainability is a multi-faceted problem that **deserves an initiative on its own**

HEP expertise should be leveraged to generate impact in the broader AI field

We should in any case strive towards building collaboration between **AI researchers, environmental scientists, and policymakers to address energy sustainability.**

Standardized metrics would be a place to start

HOW DIFFERENT ASPECTS CONTRIBUTE TO IMPROVEMENT



Patterson, David, et al. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." (2022).

THANK YOU

Sofia Vallecorsa

Sofia.Vallecorsa@cern.ch

HEP REFERENCES

[1] CMS Collaboration, "Reweighting of simulated events using machine learning techniques in CMS", CMS Physics Analysis Summary CMS-PAS-MLG-24-001, 2024.

[2] ATLAS Collaboration, "A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector", 2024. arXiv:2405.20041. Submitted to *Phys. Rev. Lett.*

[3] CMS Collaboration, "2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger", CMS Detector Performance Note CMS-DP-2024-059, 2024.

[4] ATLAS-SIM-2023-004

[5] EPJ Web of Conferences 295, 09029 (2024)

[6] EPJ Web of Conferences 295, 03001 (2024)

[7] ATL-PHYS-PROC-2024-081