Accelerating Scientific Discovery with AI

Prof. Dr. Gregor Kasieczka
Email: gregor.kasieczka@uni-hamburg.de
@kasieczka.bsky.social / Gregor Kasieczka

be.hep Solstice Meeting — 19.12.2024

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

KISS

FSP CMS

PUNCH 4NFDI

DASHH

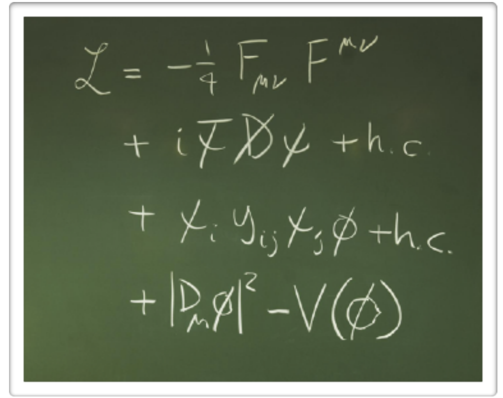
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

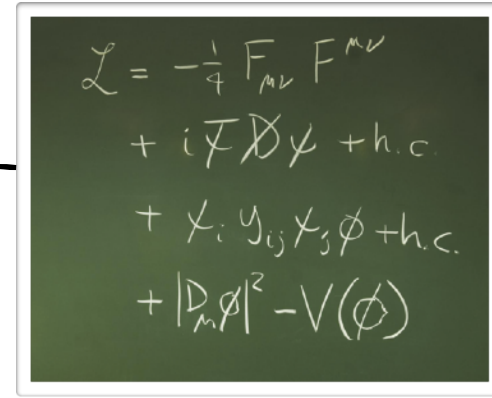
CDCS
CENTER FOR DATA AND COMPUTING IN NATURAL SCIENCES

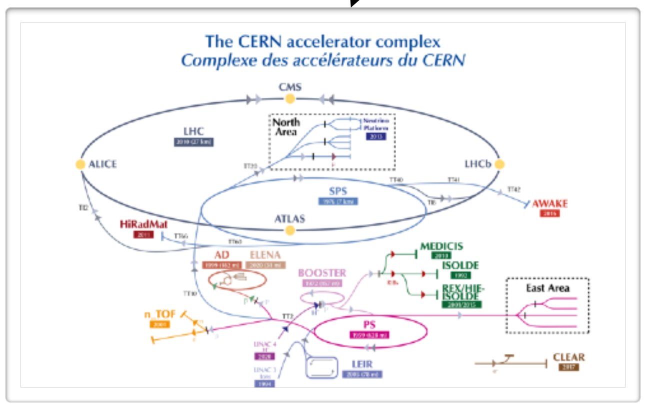
PIER
Partnership of Universität Hamburg and DESY

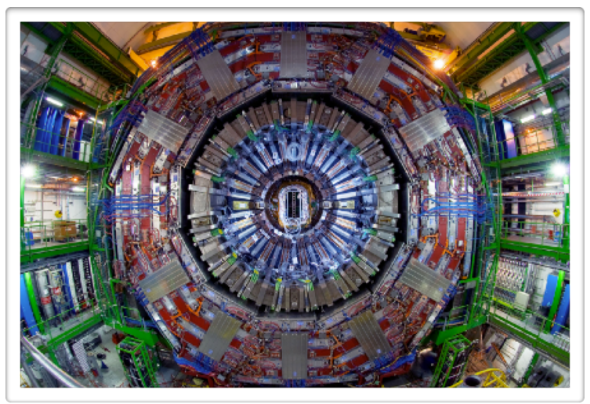
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG



First principle, quantum theoretical model

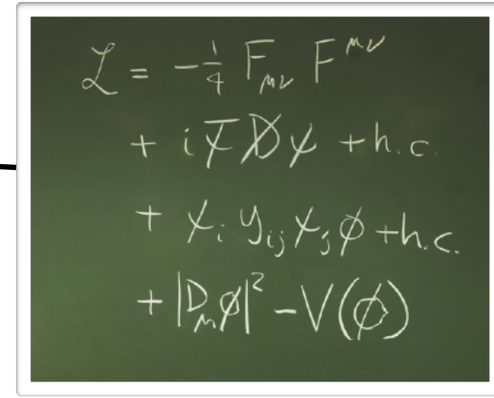


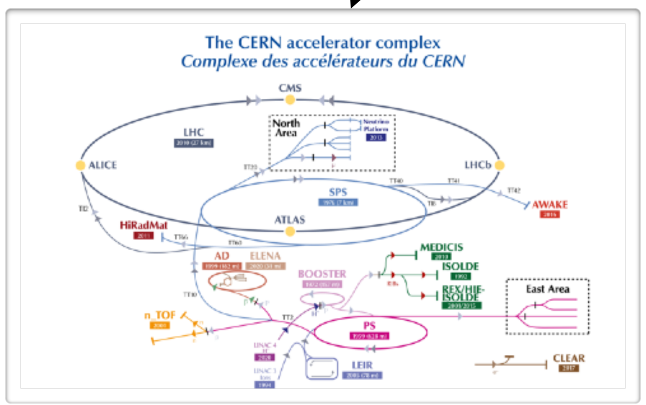


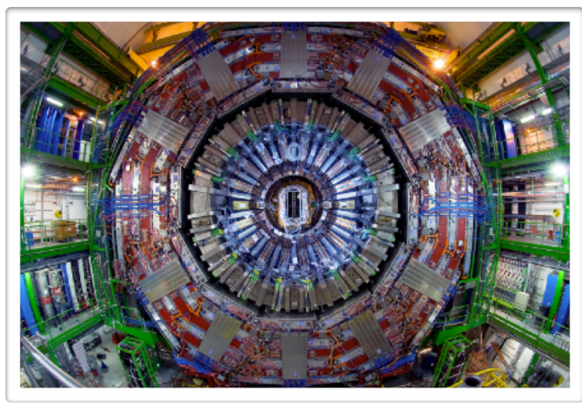


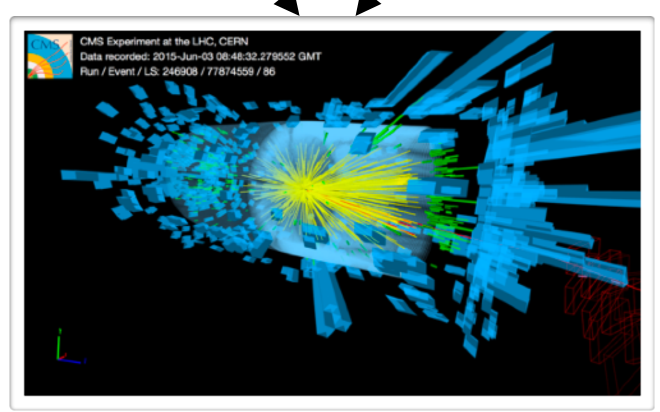
Colliders with
40 million events/second,
detectors with
100 million read-outs,

and massive theory-driven
simulation codes

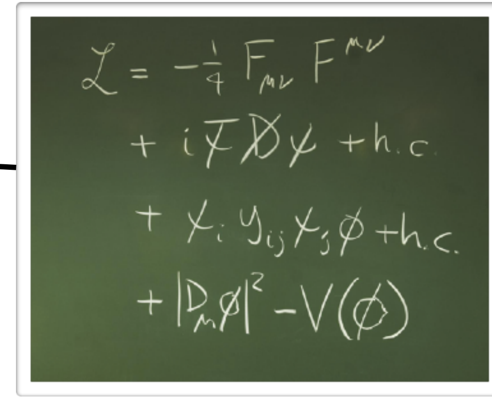


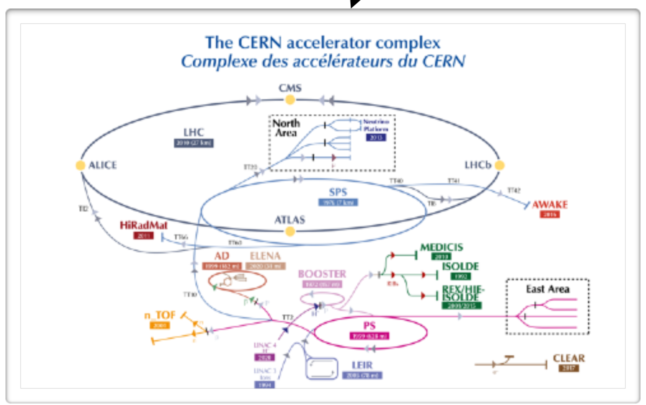


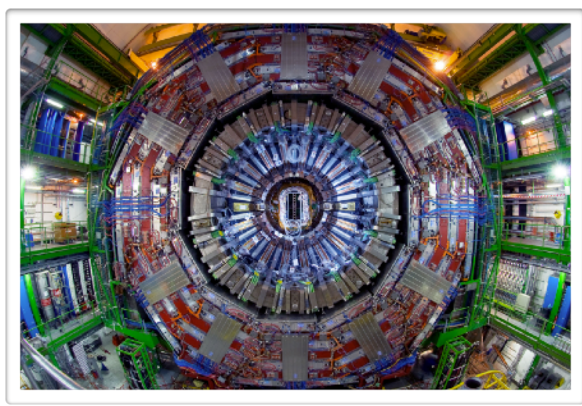


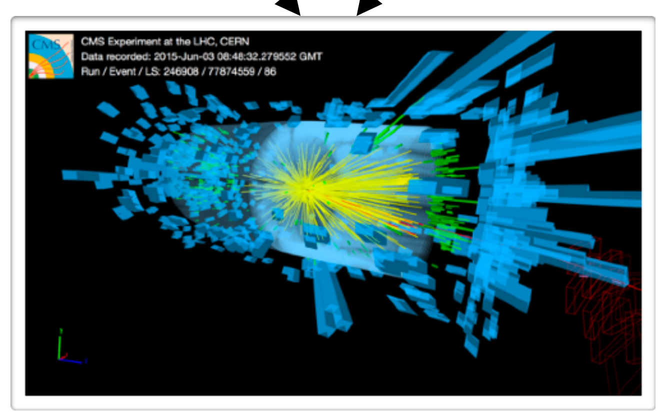


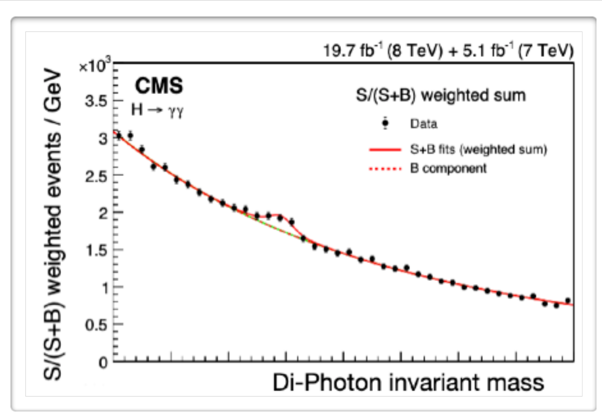
Complex reconstruction chain to turn low-level read-outs into high-level physics objects



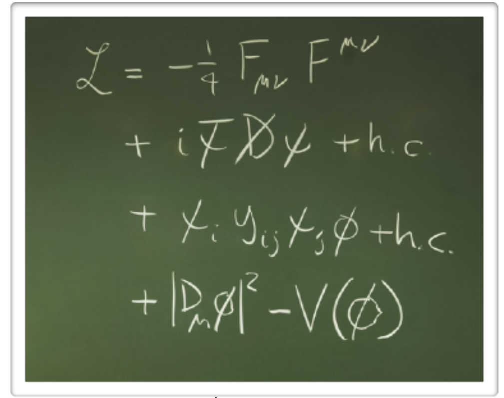




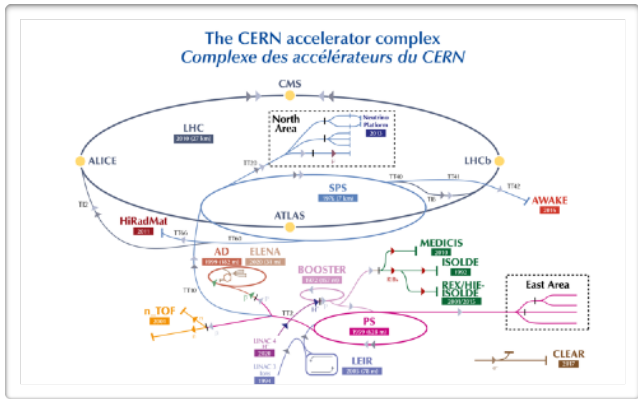




Sophisticated final statistical analysis



Inference / SBI

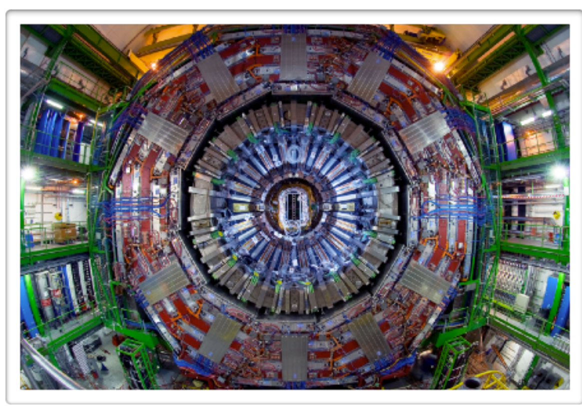


Experiment Design

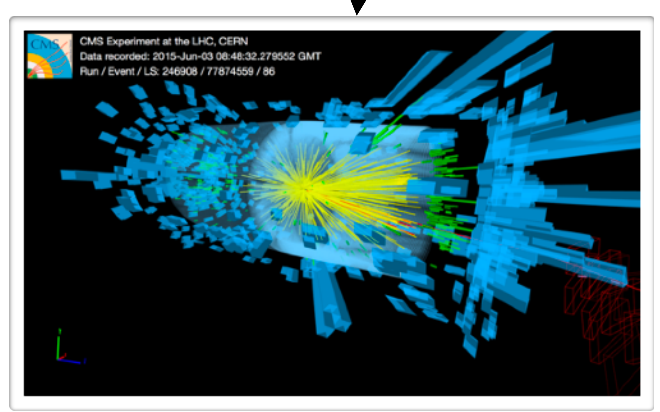
AI

Simulation

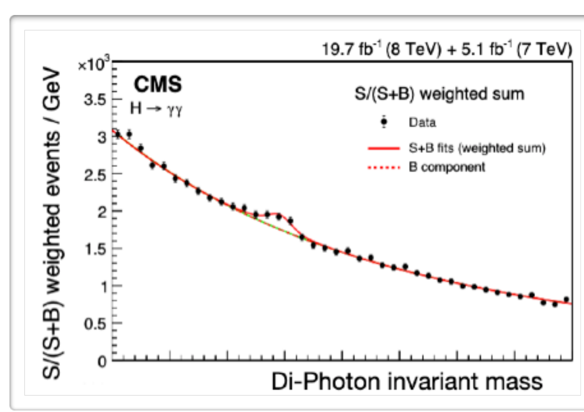
Triggers & DAQ



Tagging
Reconstruction

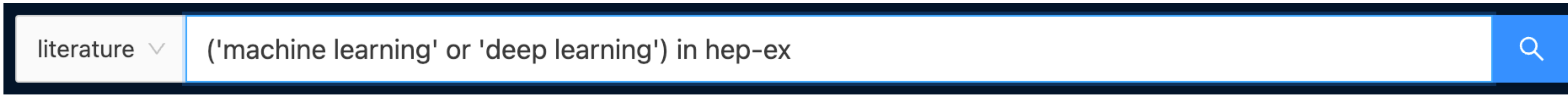


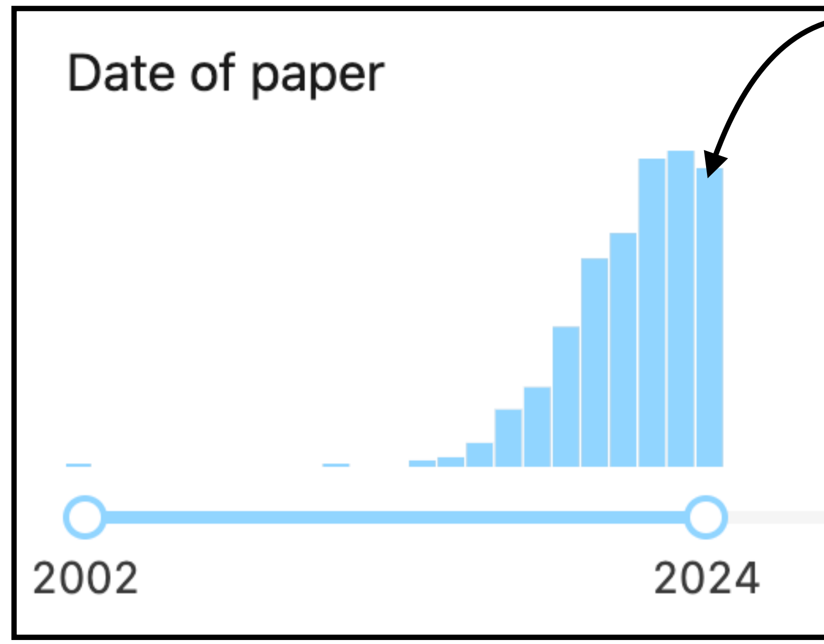
Unfolding
Anomaly Detection



AI models are studied for all aspects of modern fundamental physics

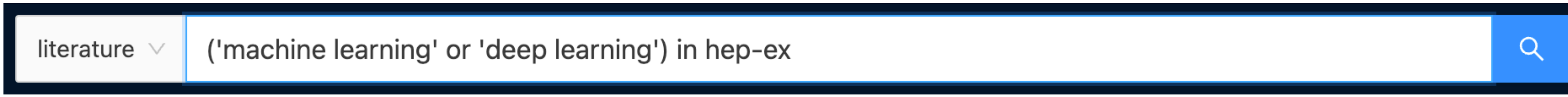
Setting the stage

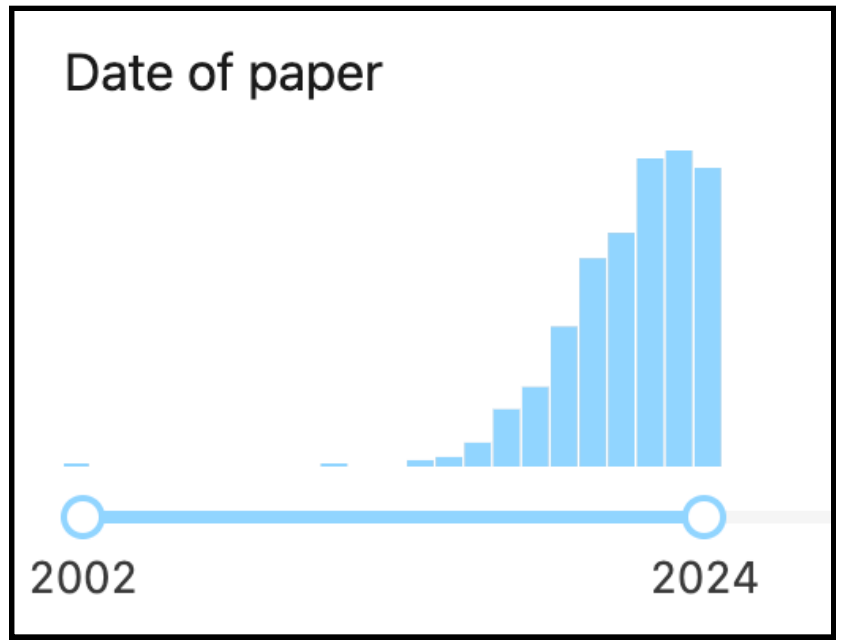


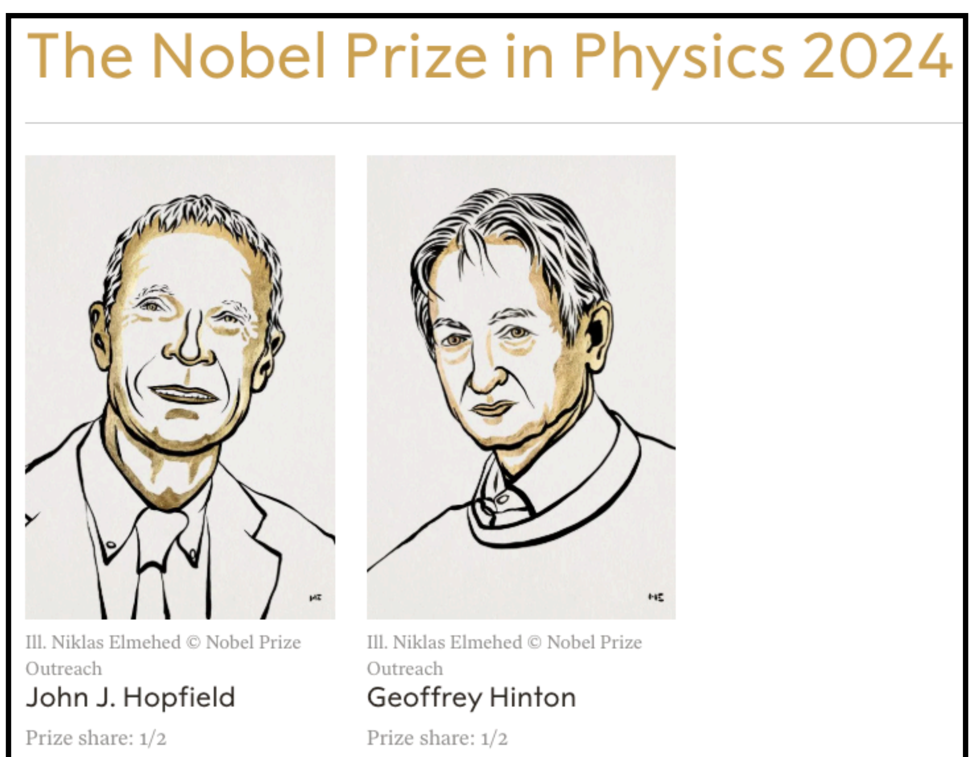


On the way to 200+ papers in 2024

Setting the stage



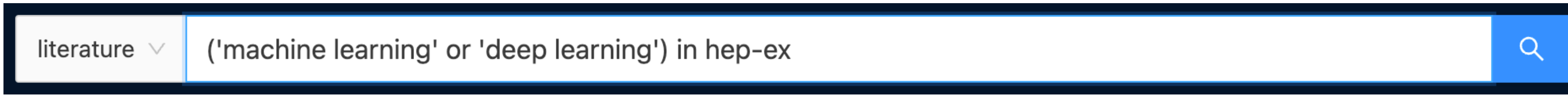


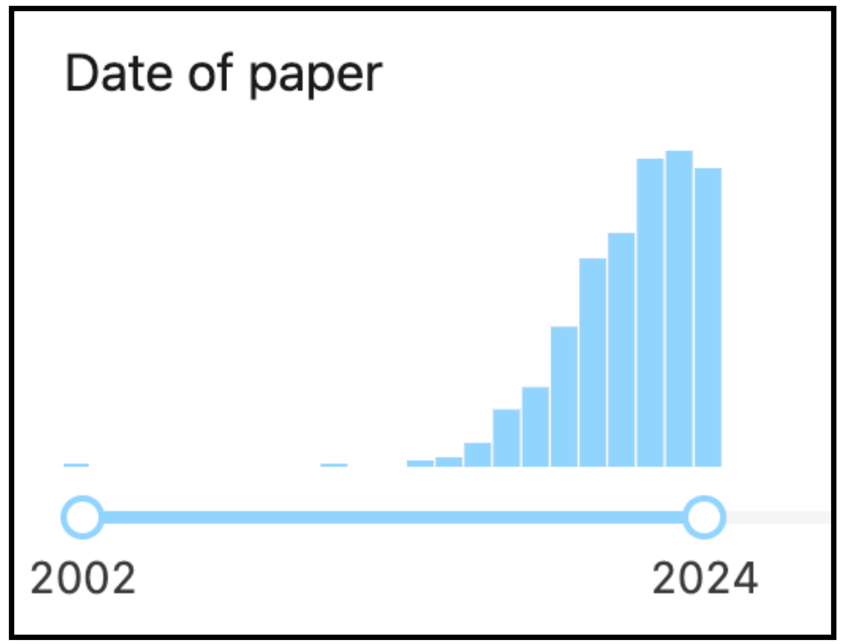


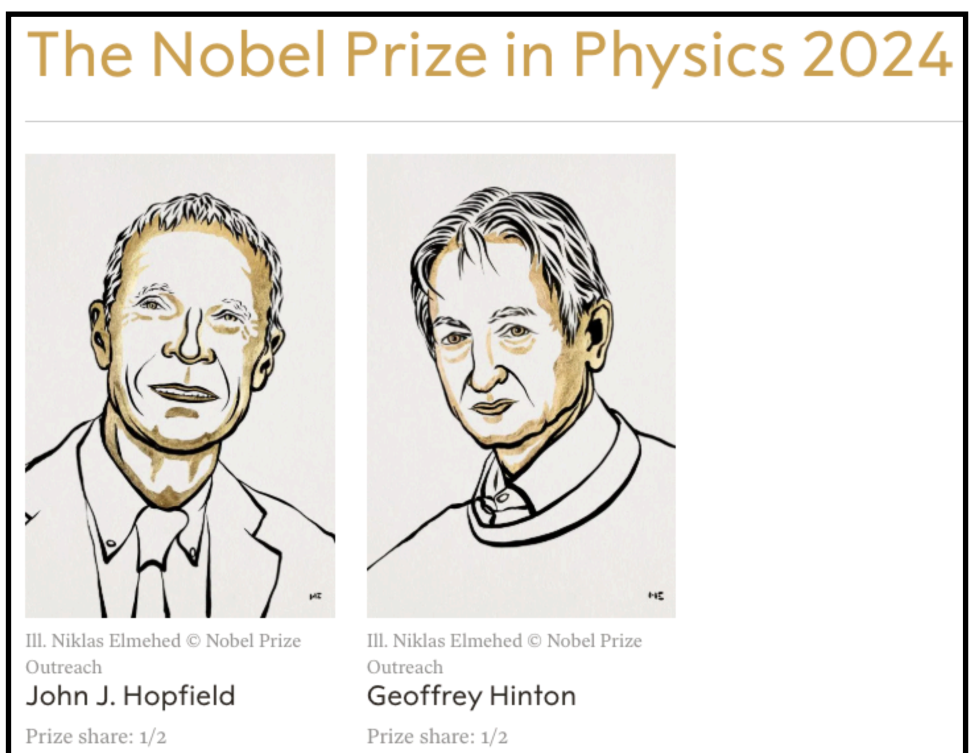
Use of ML is rapidly rising in fundamental physics

Fundamental ML developments inspired by physical systems

Setting the stage



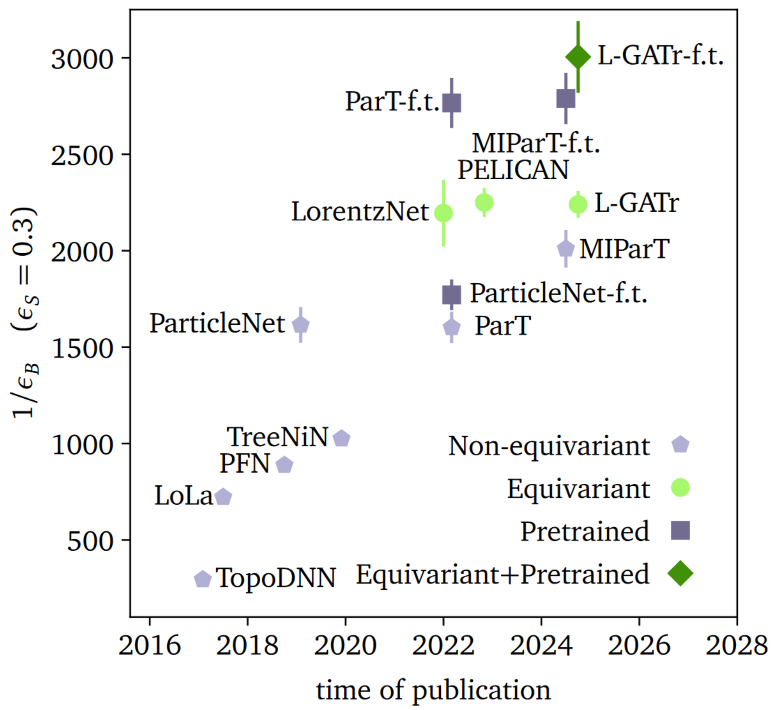




Use of ML is rapidly rising in fundamental physics

Fundamental ML developments inspired by physical systems

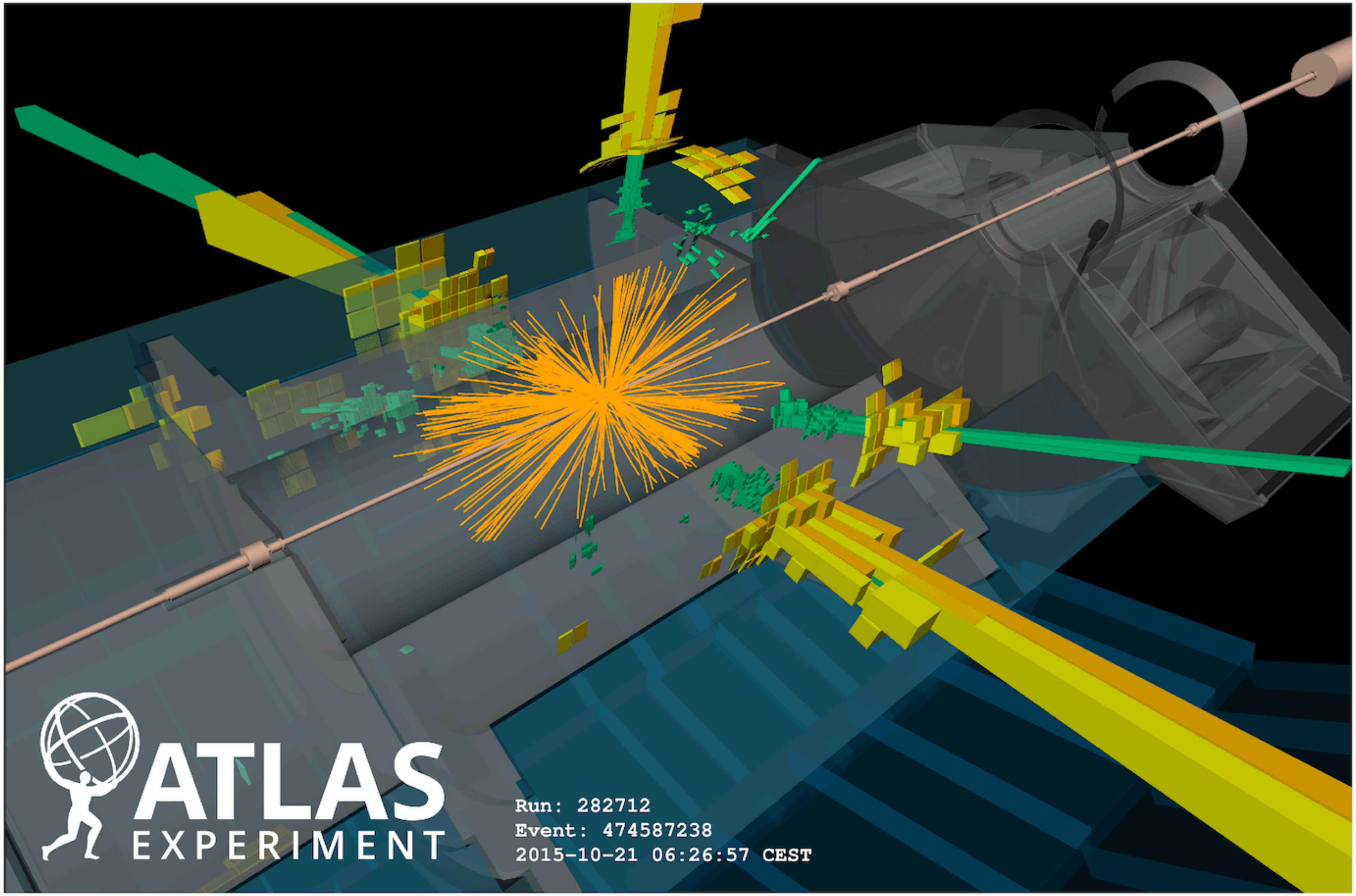
What next?



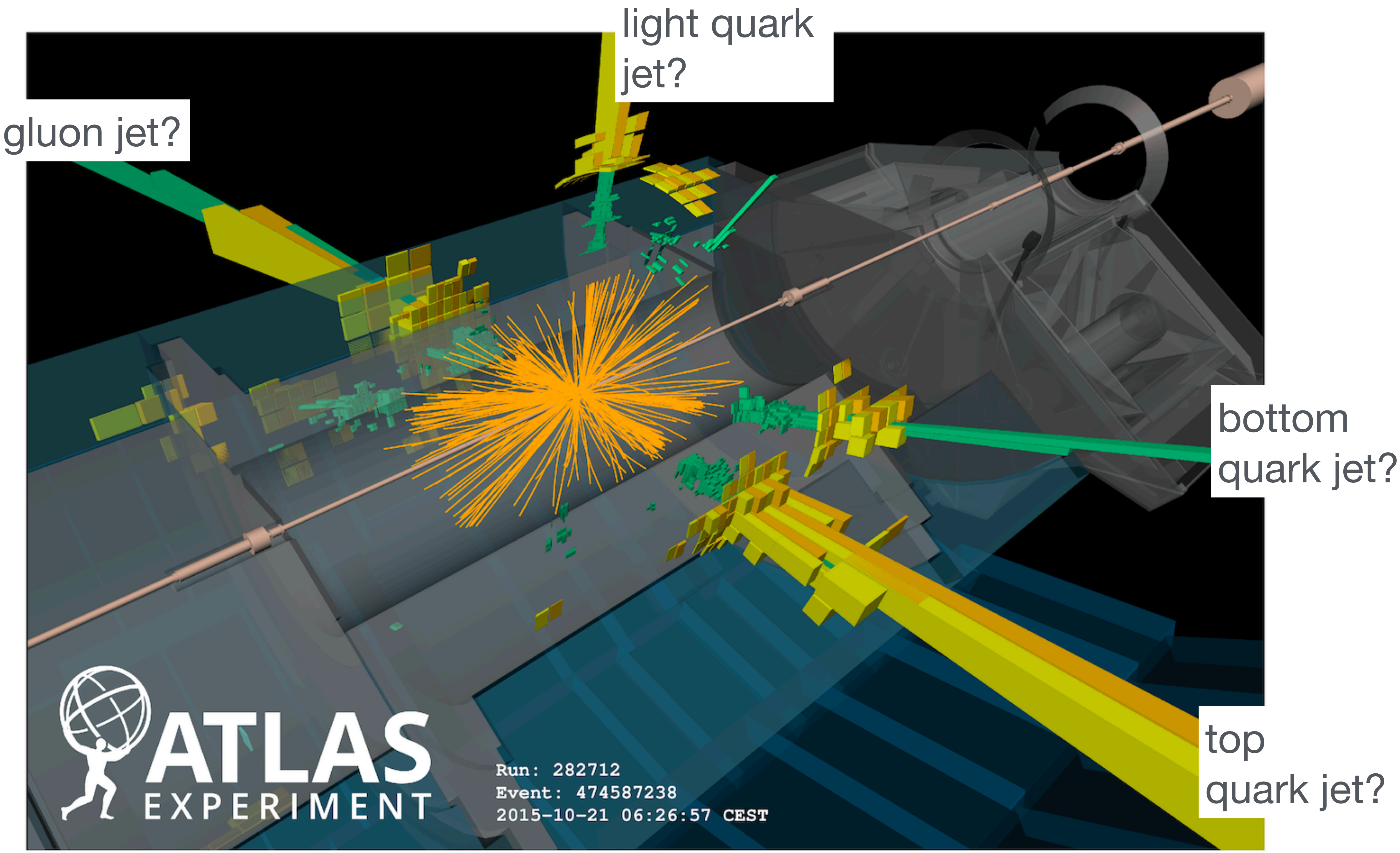
Physics or compute

The surrogate revolution

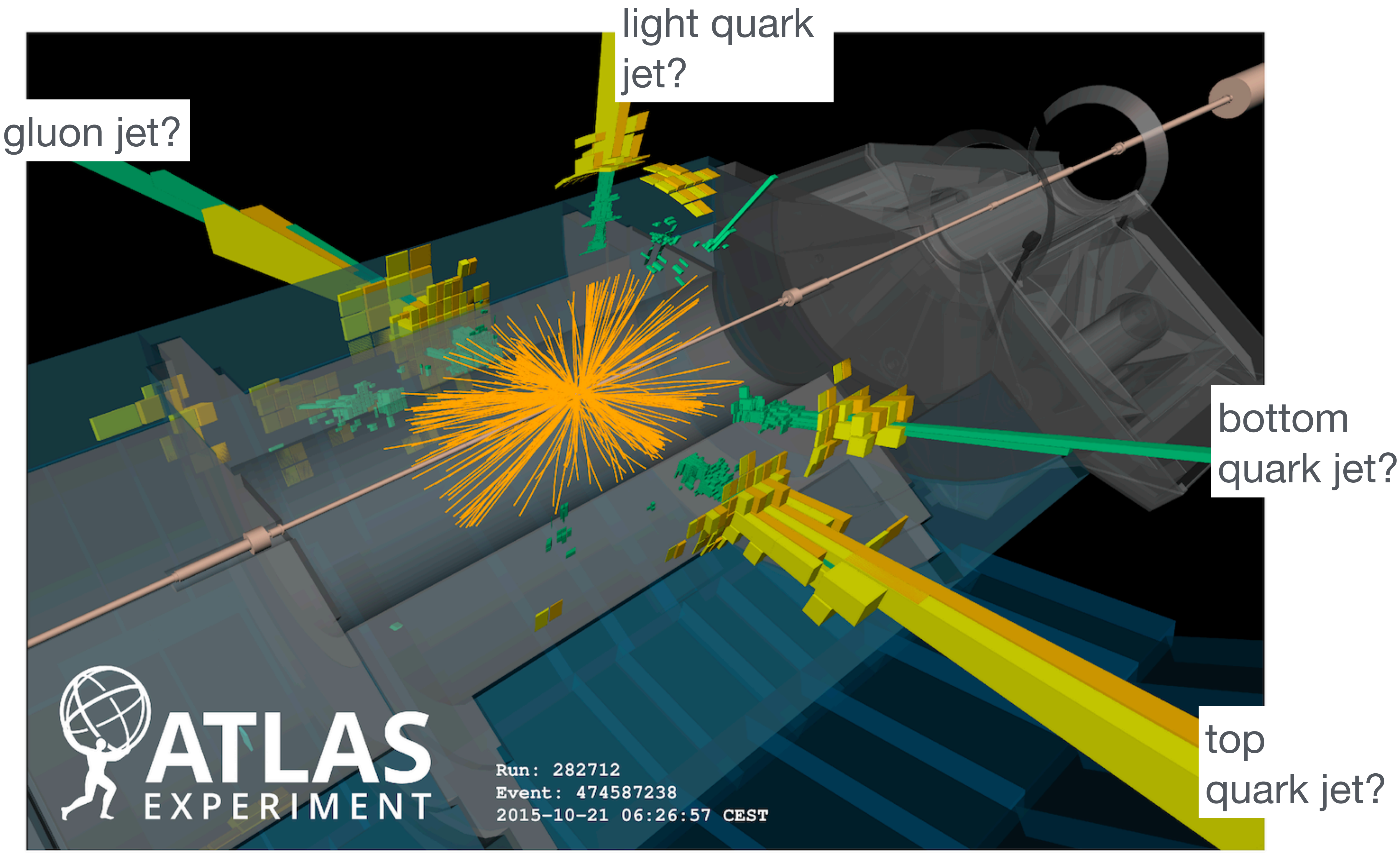
The rise of the AI physicst



A jet is a
collimated shower of particles in the detector

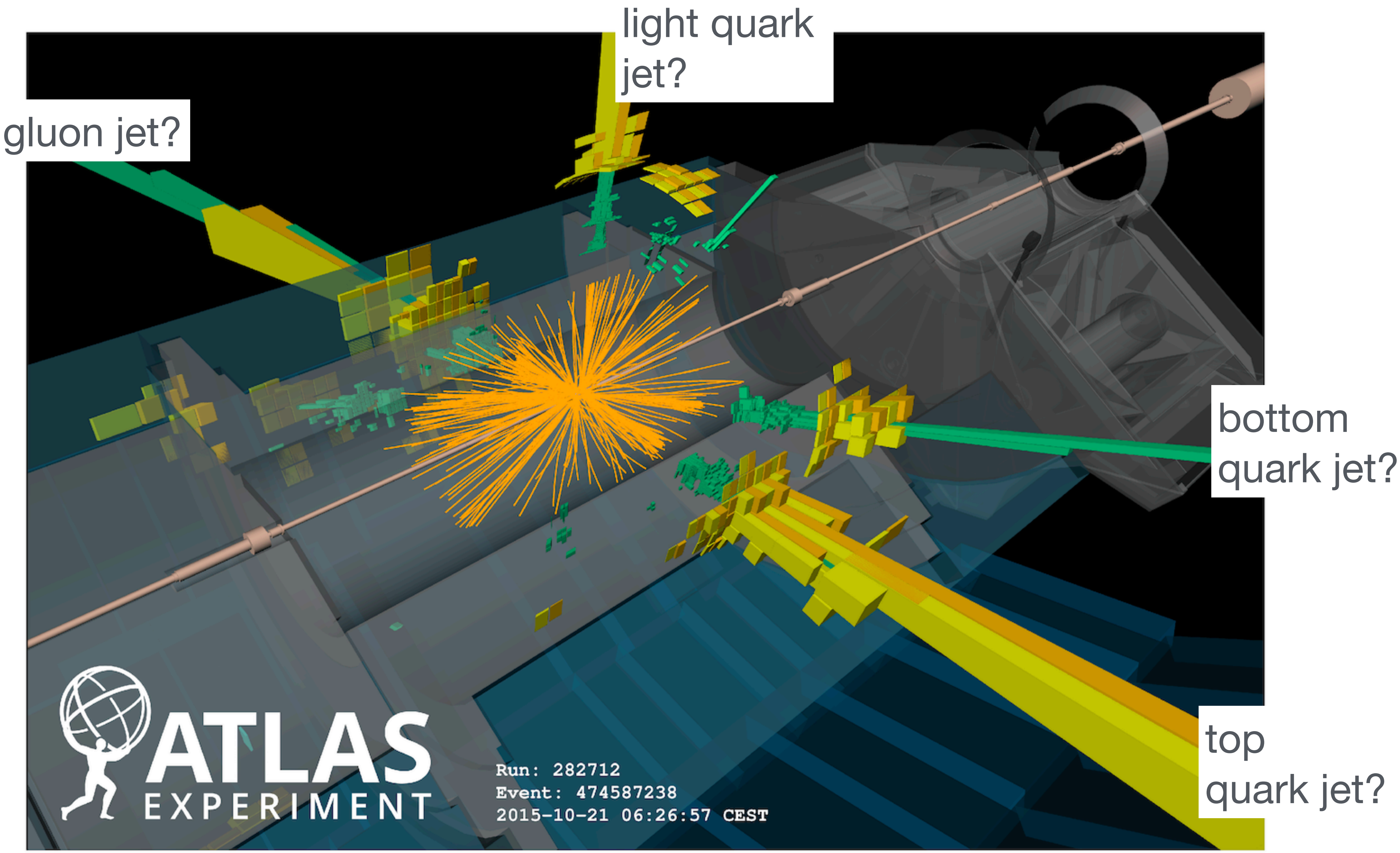


We want to know
which particle produced a jet

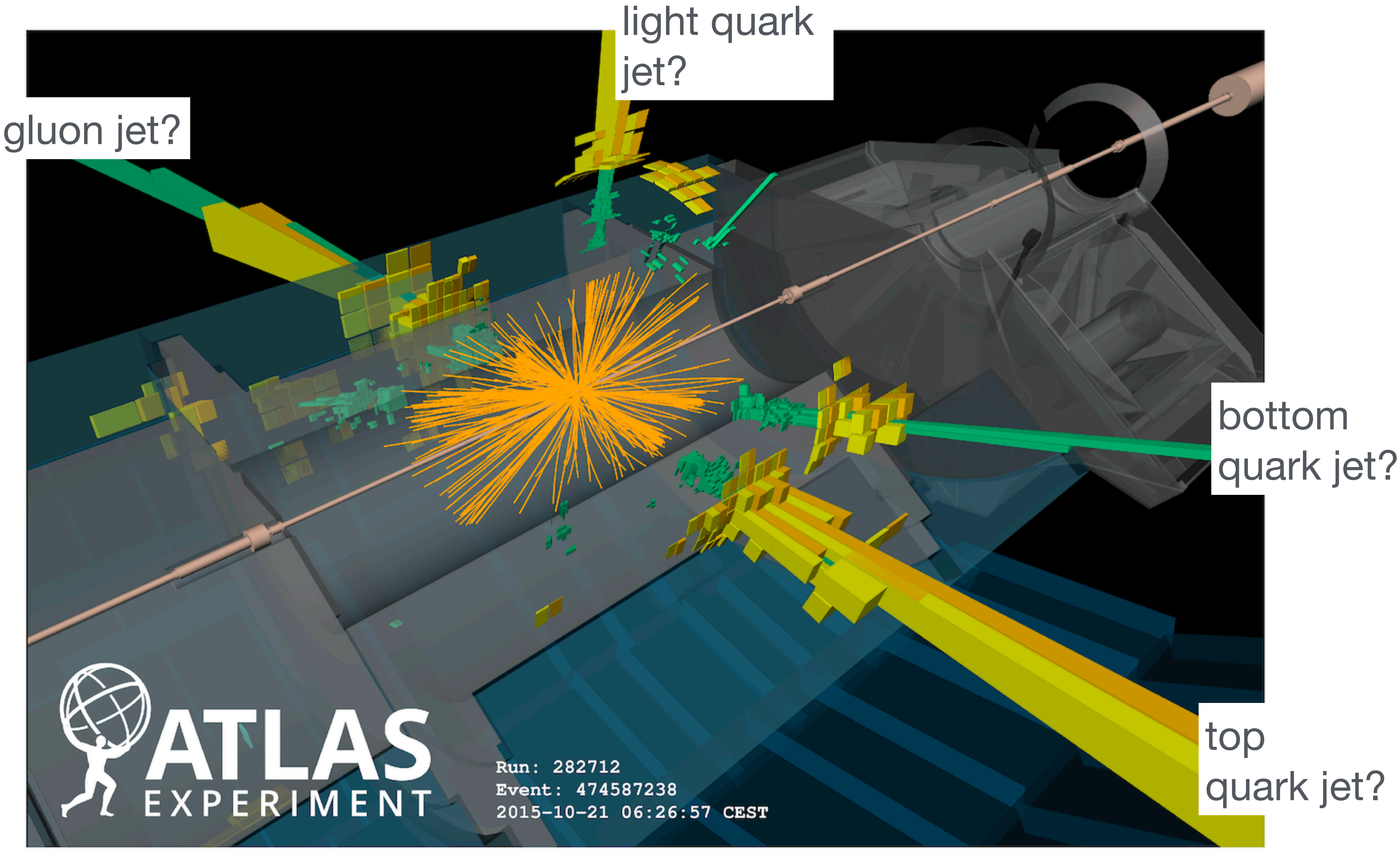


Why?

- Discover new particles
- Measure the Standard Model



Let’s focus on top quarks
(Modern taggers are multi-class)

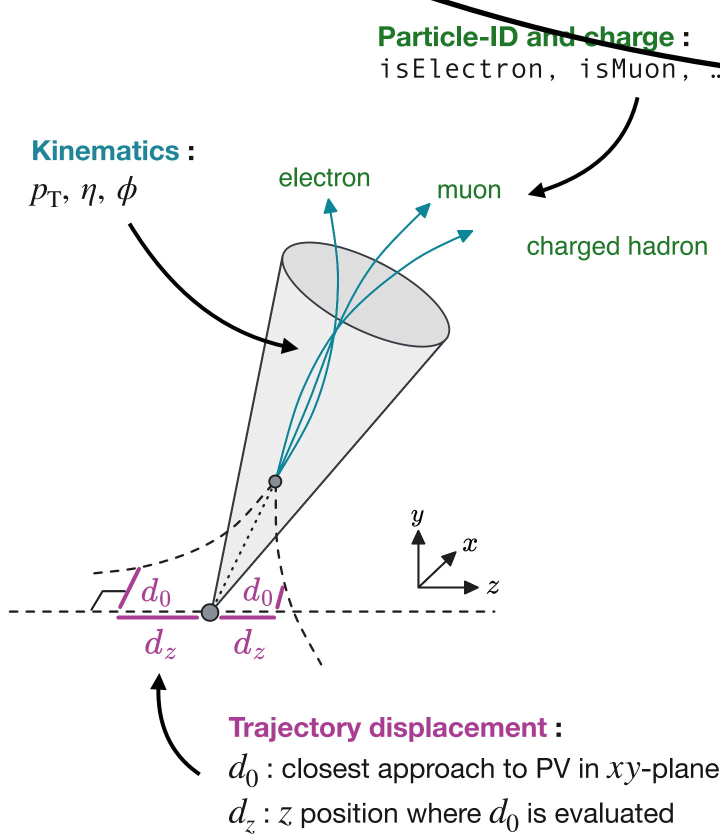


How to build ML algorithms for complex, heterogenous data?

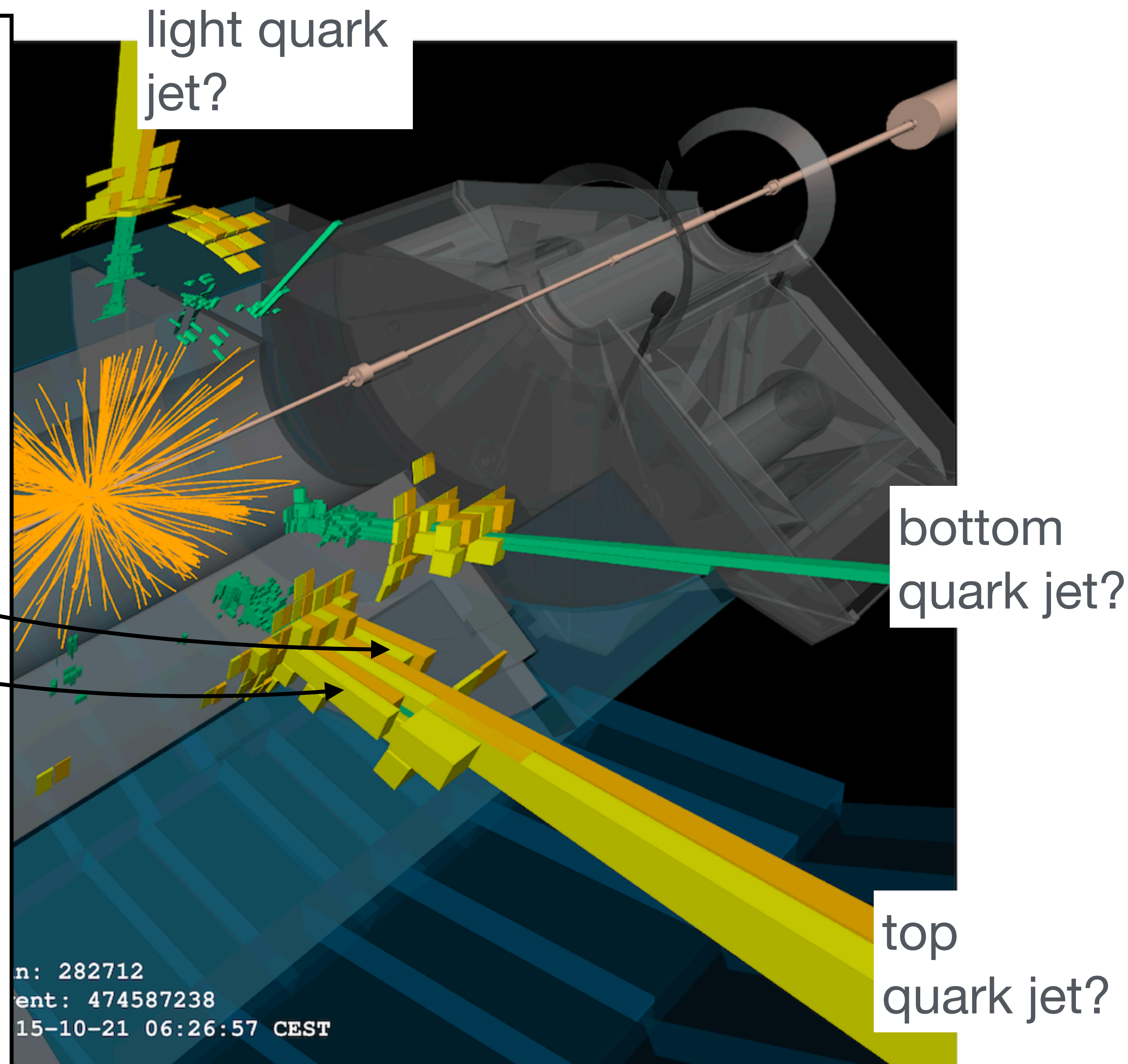
Data most naturally viewed as point cloud:

Each input (e.g. jet, event, ..) is a set of k-dimensional vectors (individual particles, hits, ..)

$$J_i = \{\vec{p_1}, \ldots, \vec{p_n}\}$$



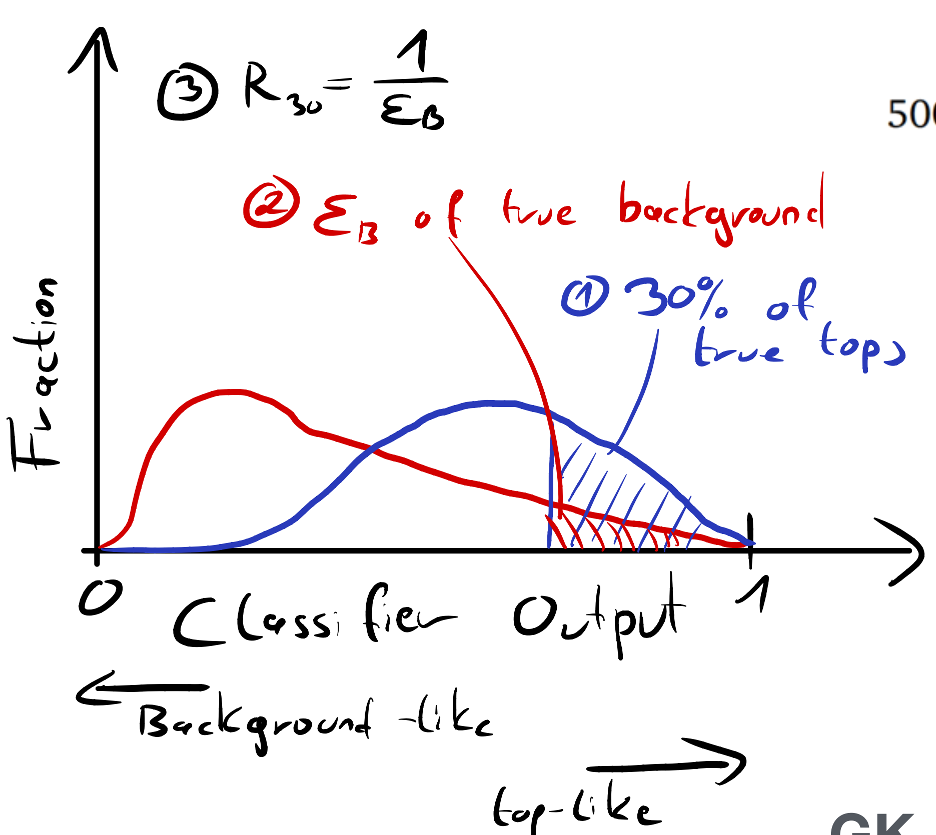
Example per-particle features

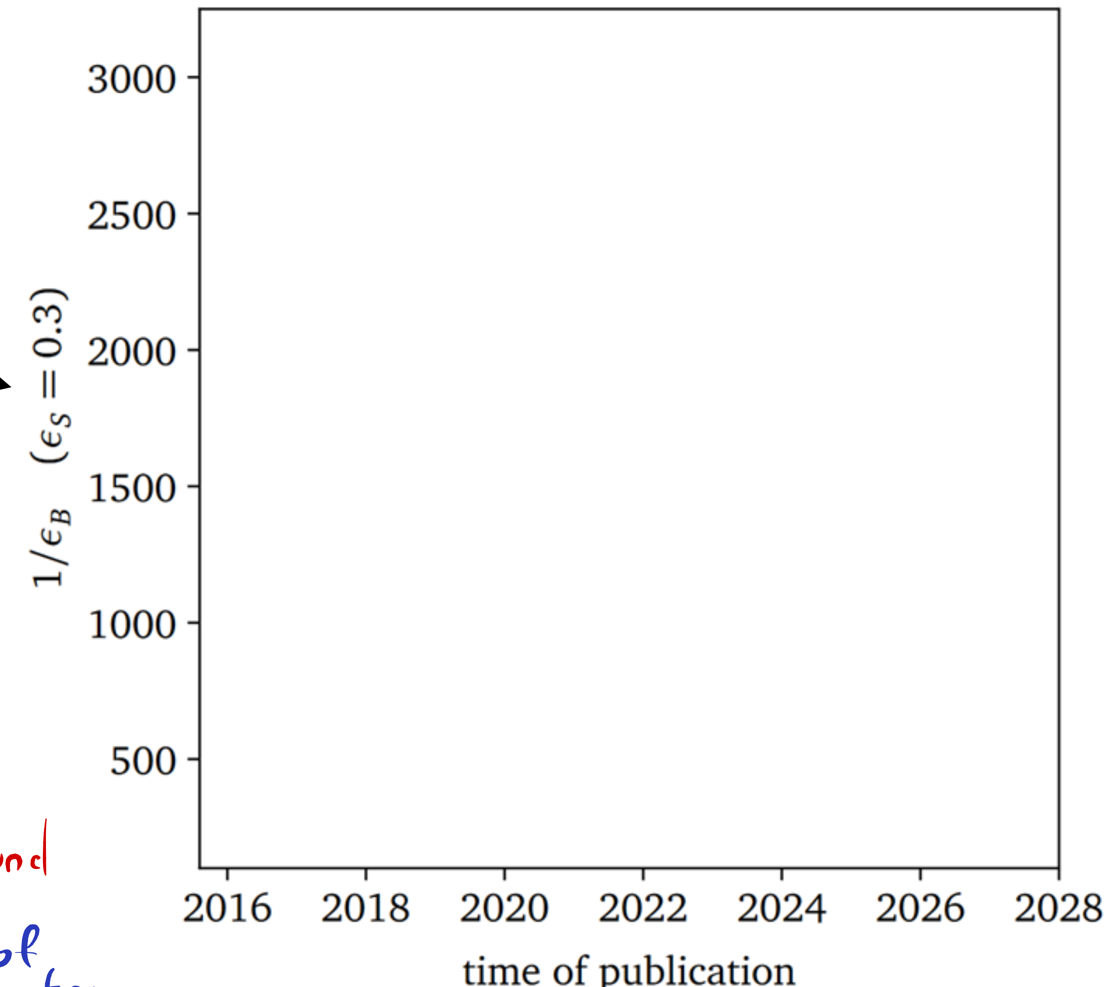


Komiske, Metodiev, Thaler 1810.05165; Birk, **GK**, et al 2312.00123;

Top Quark Tagging

Public simulated* benchmark dataset





The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)[1], T. Plehn (ed)[2], A. Butter[2], K. Cranmer[3], D. Debnath[4], B. M. Dillon[5], M. Fairbairn[6], D. A. Faroughy[5], W. Fedorko[7], C. Gay[7], L. Gouskos[8], J. F. Kamenik[5,9], P. T. Komiske[10], S. Leiss[1], A. Lister[7], S. Macaluso[3,4], E. M. Metodiev[10], L. Moore[11], B. Nachman,[12,13], K. Nordström[14,15], J. Pearkes[7], H. Qu[8], Y. Rath[16], M. Rieger[16], D. Shih[4], J. M. Thompson[2], and S. Varma[6]

1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Institut für Theoretische Physik, Universität Heidelberg, Germany
3 Center for Cosmology and Particle Physics and Center for Data Science, NYU, USA
4 NHECT, Dept. of Physics and Astronomy, Rutgers, The State University of NJ, USA
5 Jozef Stefan Institute, Ljubljana, Slovenia
6 Theoretical Particle Physics and Cosmology, King's College London, United Kingdom
7 Department of Physics and Astronomy, The University of British Columbia, Canada
8 Department of Physics, University of California, Santa Barbara, USA
9 Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
10 Center for Theoretical Physics, MIT, Cambridge, USA
11 CP3, Universitéxx Catholique de Louvain, Louvain-la-Neuve, Belgium
12 Physics Division, Lawrence Berkeley National Laboratory, Berkeley, USA
13 Simons Inst. for the Theory of Computing, University of California, Berkeley, USA
14 National Institute for Subatomic Physics (NIKHEF), Amsterdam, Netherlands
15 LPTHE, CNRS & Sorbonne Université, Paris, France
16 III. Physics Institute A, RWTH Aachen University, Germany

gregor.kasieczka@uni-hamburg.de
plehn@uni-heidelberg.de

July 24, 2019

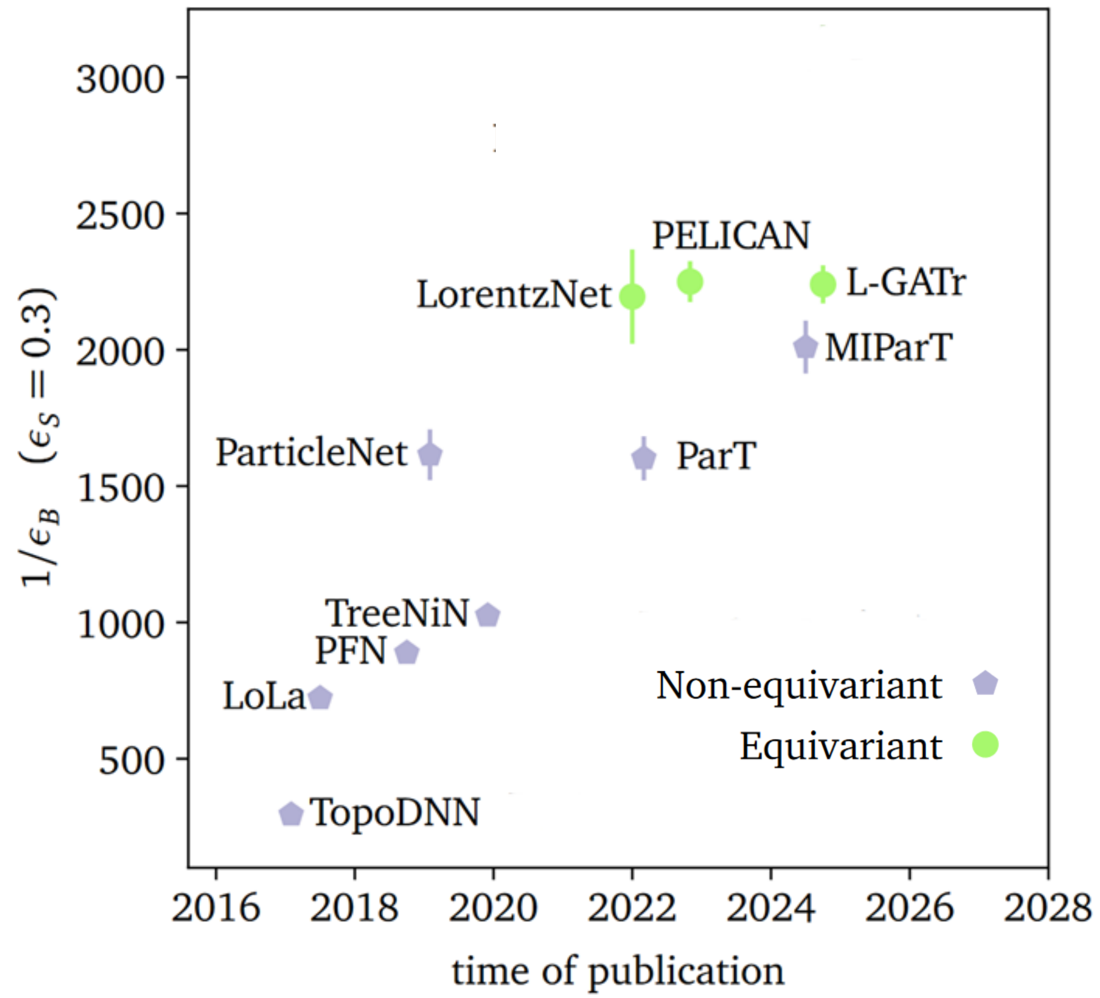
Abstract

Based on the established task of identifying boosted, hadronically decaying top quarks, we compare a wide range of modern machine learning approaches. Unlike most established methods they rely on low-level input, for instance calorimeter output. While their network architectures are vastly different, their performance is comparatively similar. In general, we find that these new approaches are extremely powerful and great fun.

GK, Plehn, et al 1902.09914

* techniques are also applied to collider data

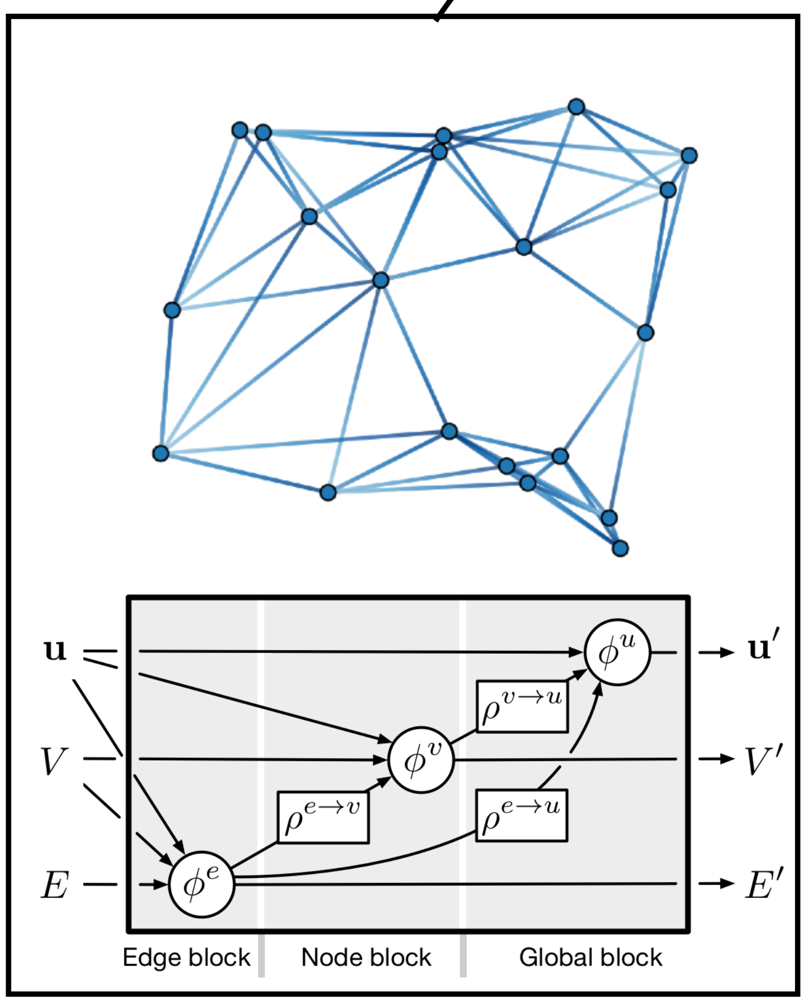
Top Quark Tagging

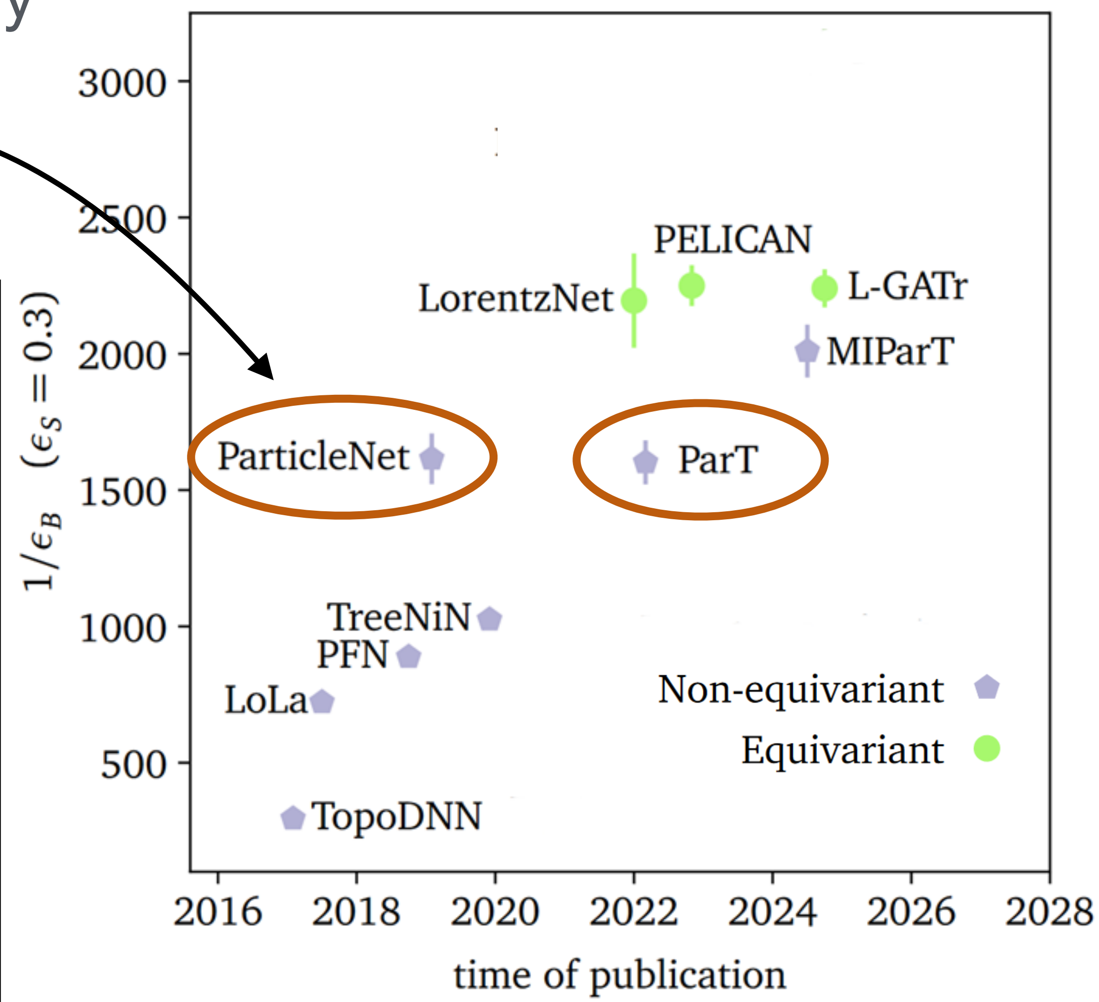


(*Subset of methods, graph from
V Breso at ML4Jets)

Top Quark Tagging

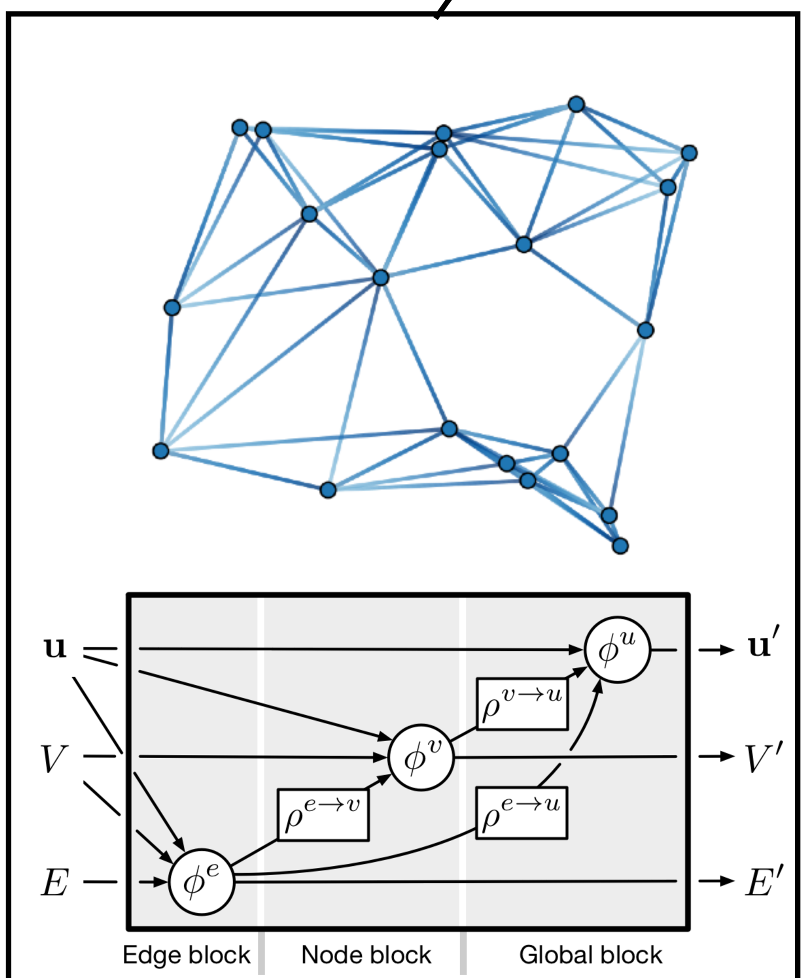
View data as graph based on geometry & use message passing

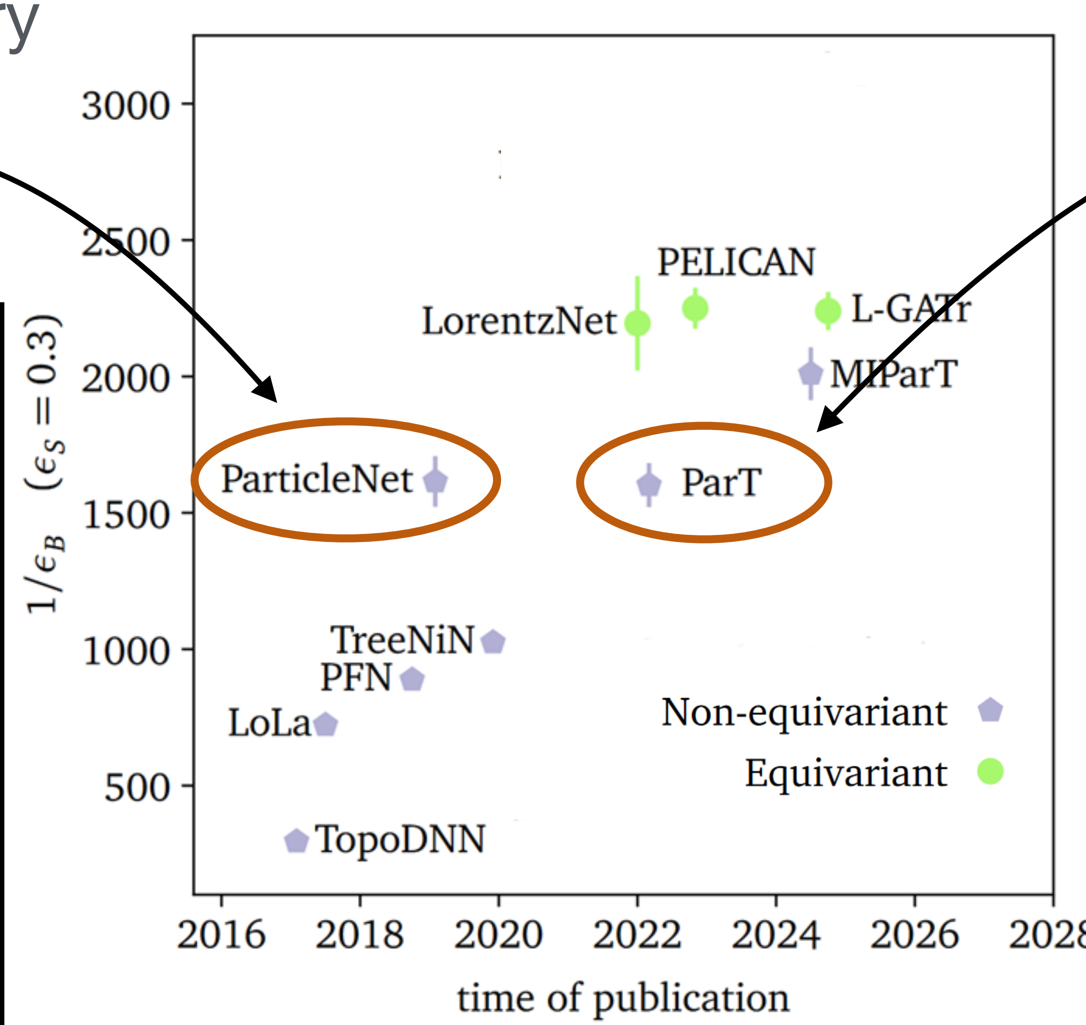




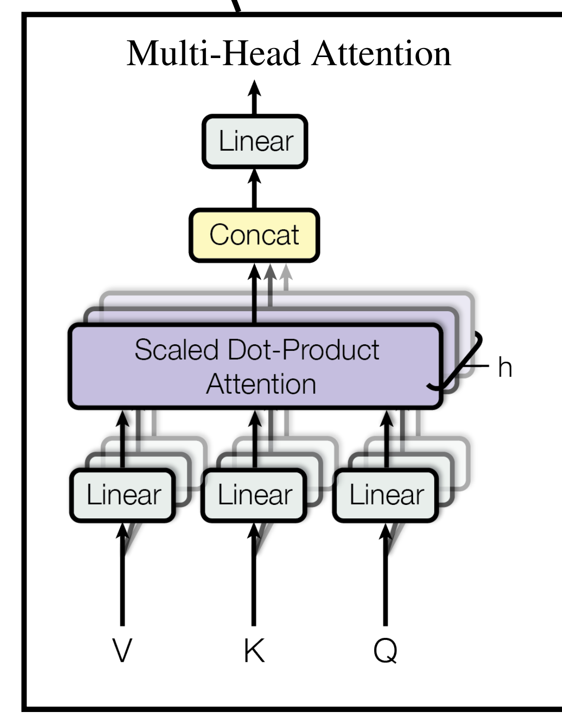
Top Quark Tagging

View data as graph based on geometry & use message passing





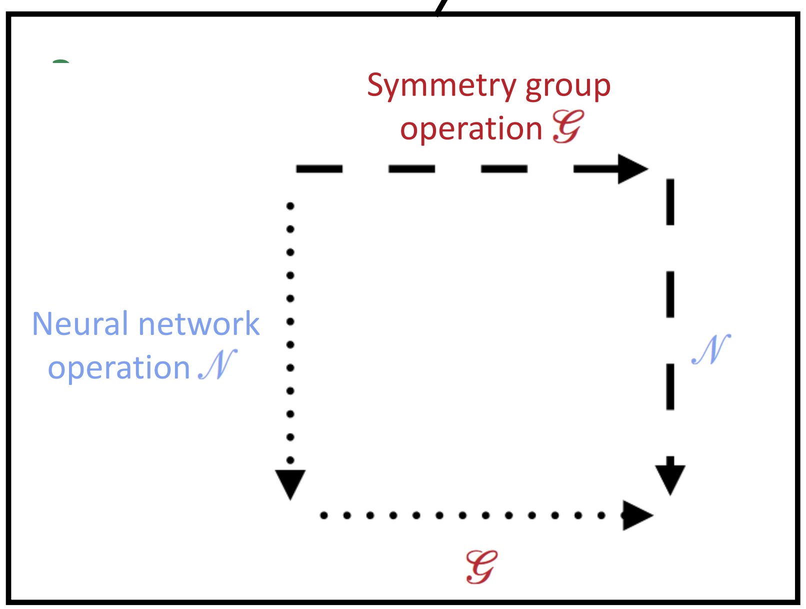
Learn relations from data using attention

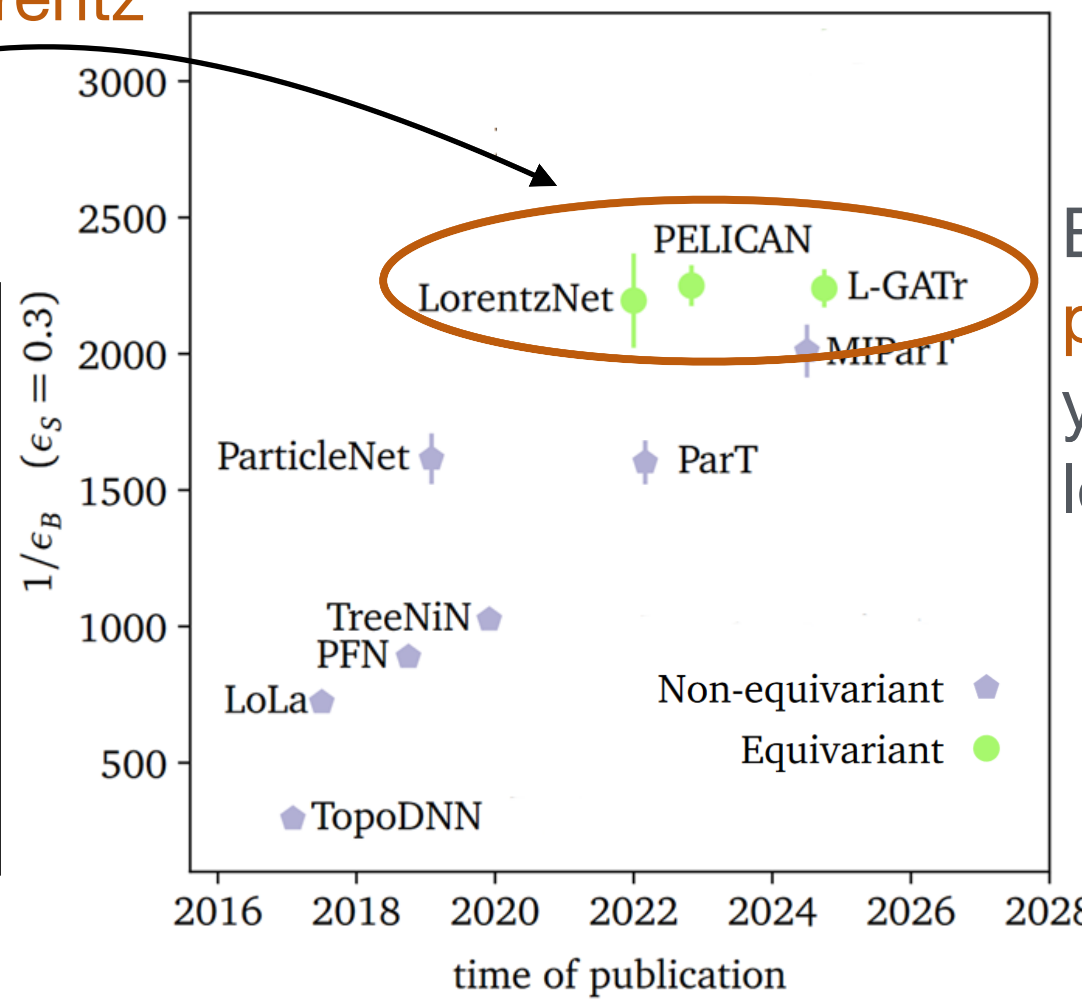


Qu, Gouskos 1902.08570; Qu et al 2202.03772; Gong et al 2201.08187; Shlomi et al 2007.13681; Bogatskiy et al 2211.00454;

Top Quark Tagging

Constrain functions to be equivariant under the Lorentz group

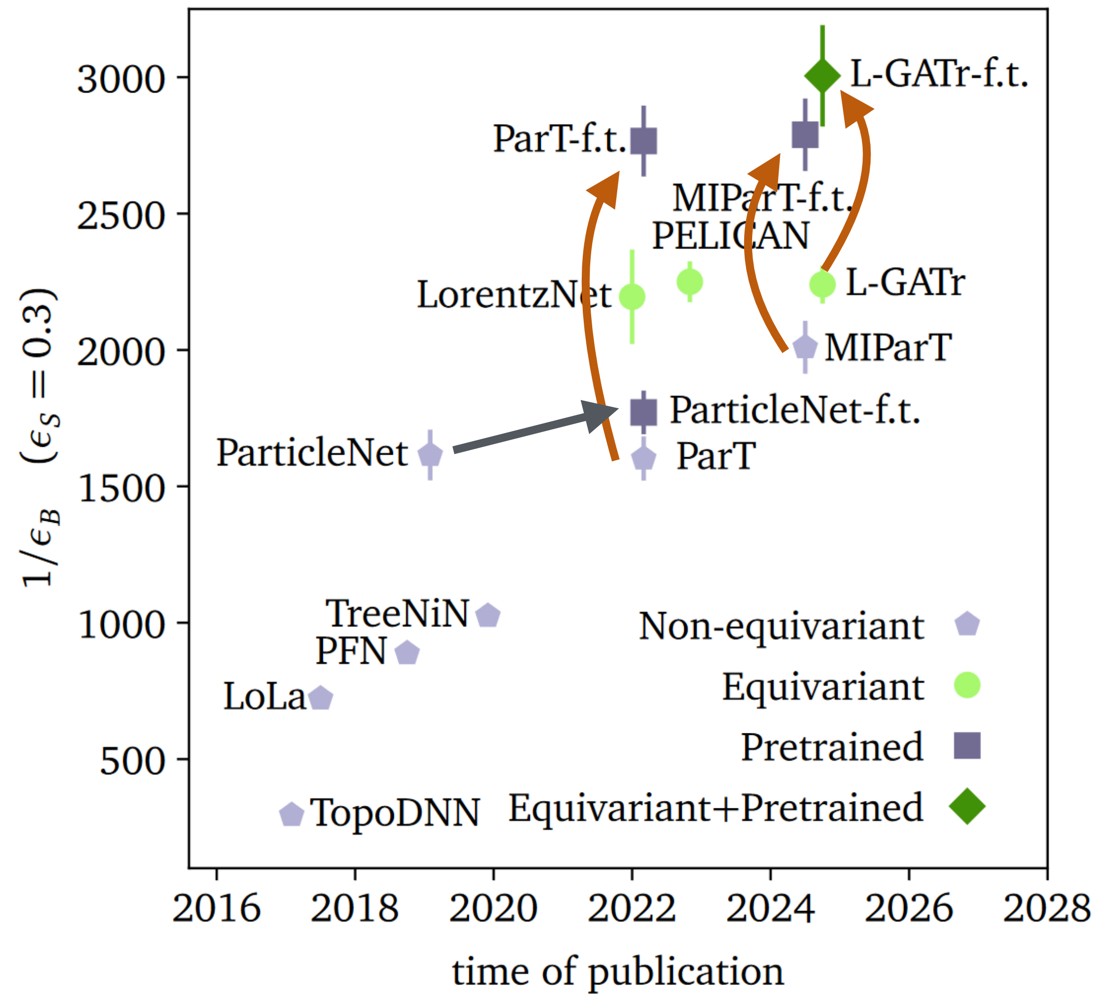




Explicitly injecting physics knowledge yields more efficient learning

Qu, Gouskos 1902.08570; Qu et al 2202.03772; Gong et al 2201.08187; Shlomi et al 2007.13681; Bogatskiy et al 2211.00454; Brehmer et al 2411.00446

Top Quark Tagging



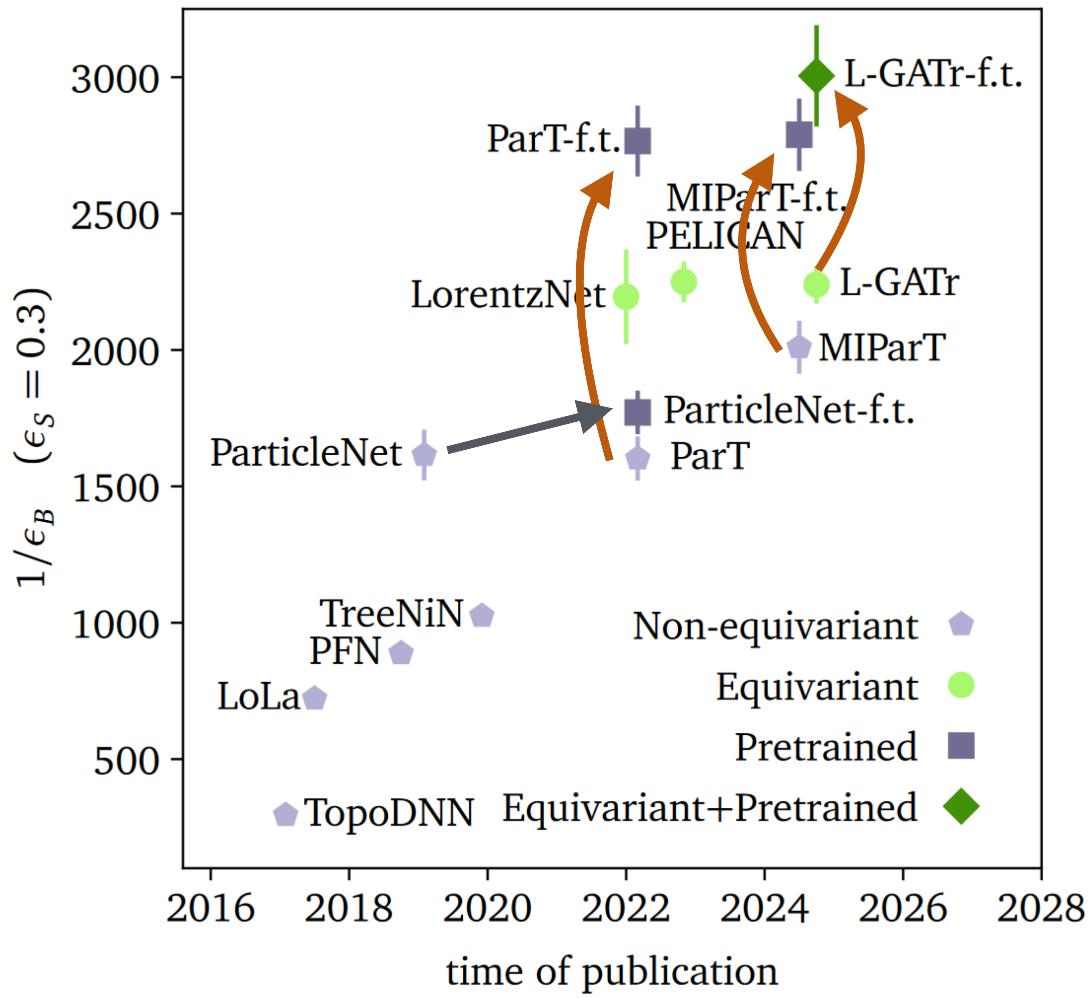
Fine tuning:
Pre-train on one dataset and recycle weights for training on new data

Yields substantial boost in performance (for transformer-based models)

Can combine with equivariance

Top Quark Tagging

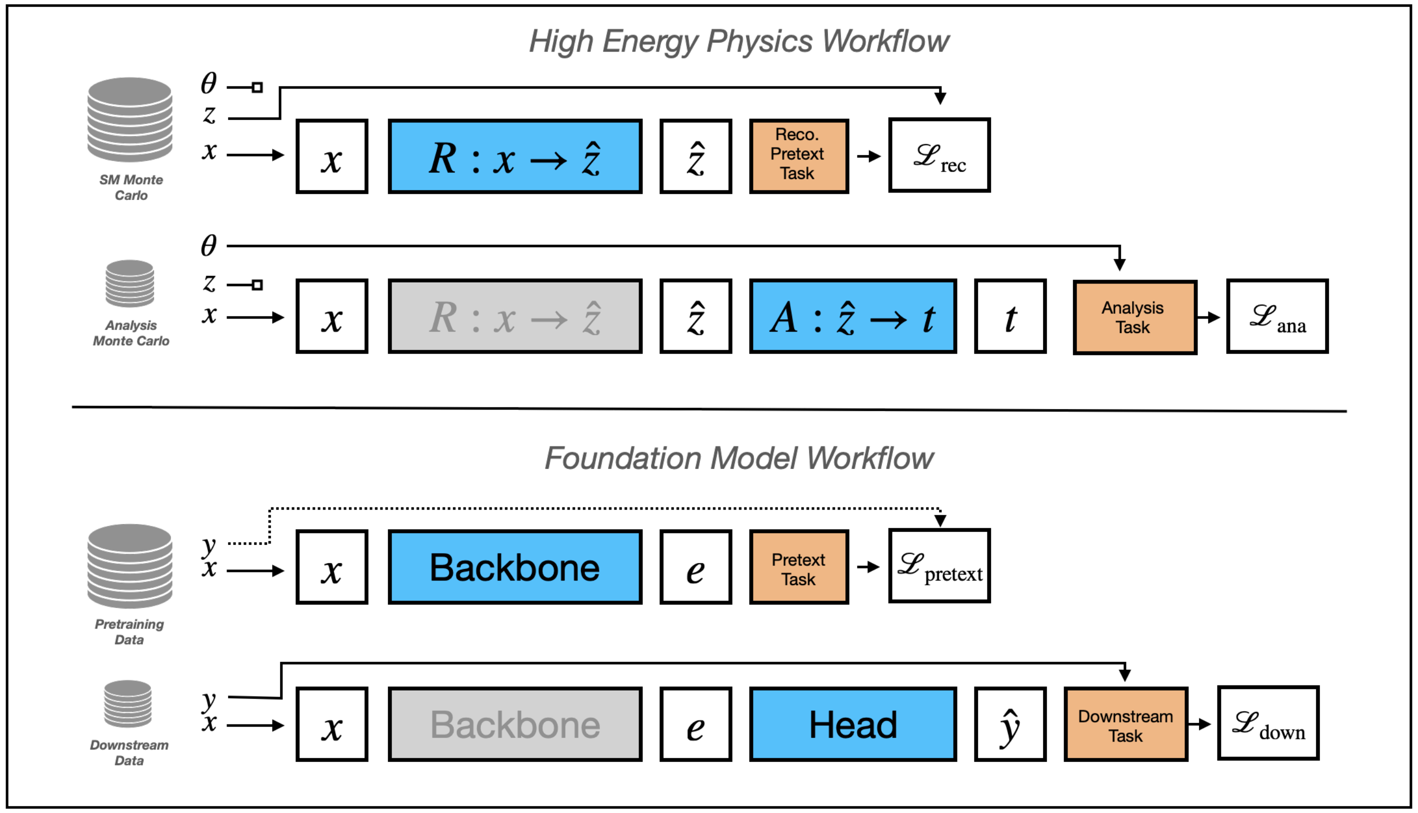
- Usefulness of open benchmark data
- Good representation of data pays off
- Transformers + physics rules
- Boost by fine-tuning across datsets



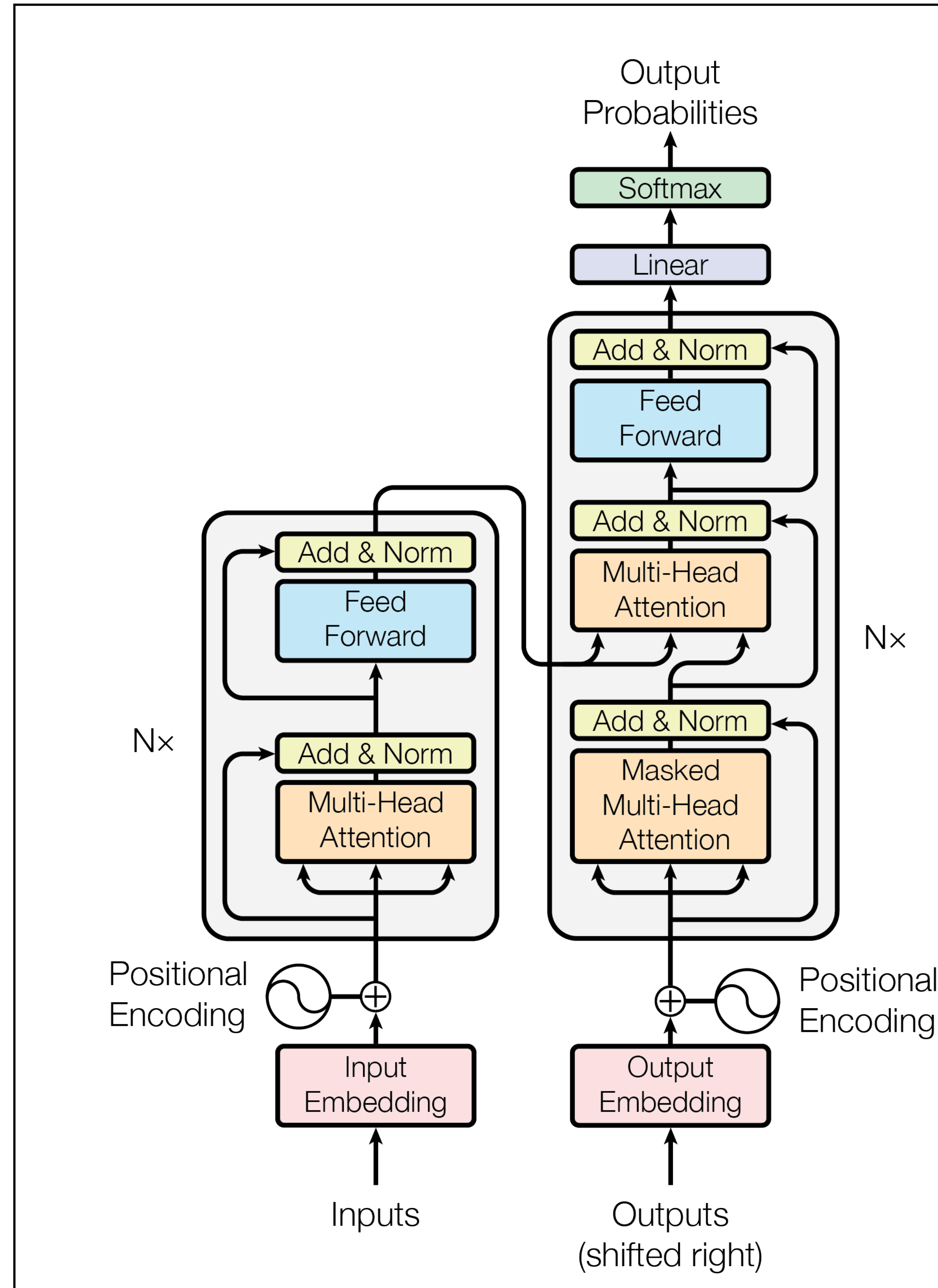
Important issues on application side

- Domain shift
- Calibrateable
- Compute cost

Foundation models



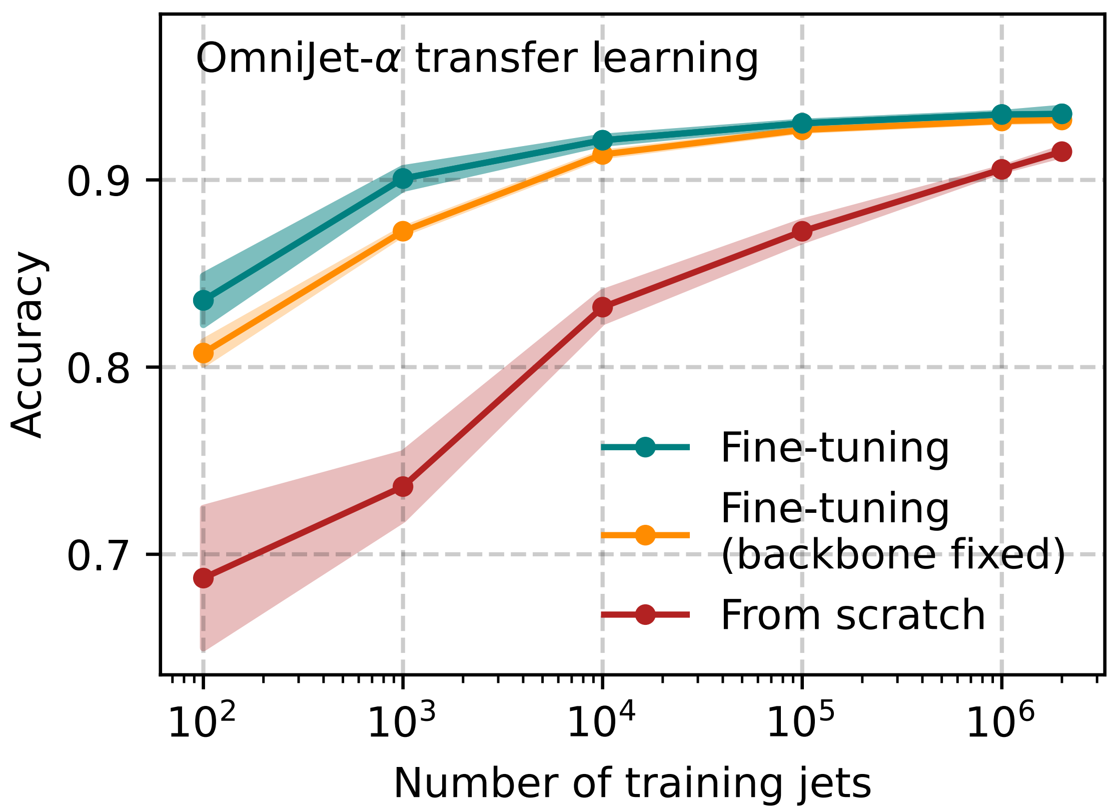
Foundation models extend finetuning broadly and centralise and re-use training



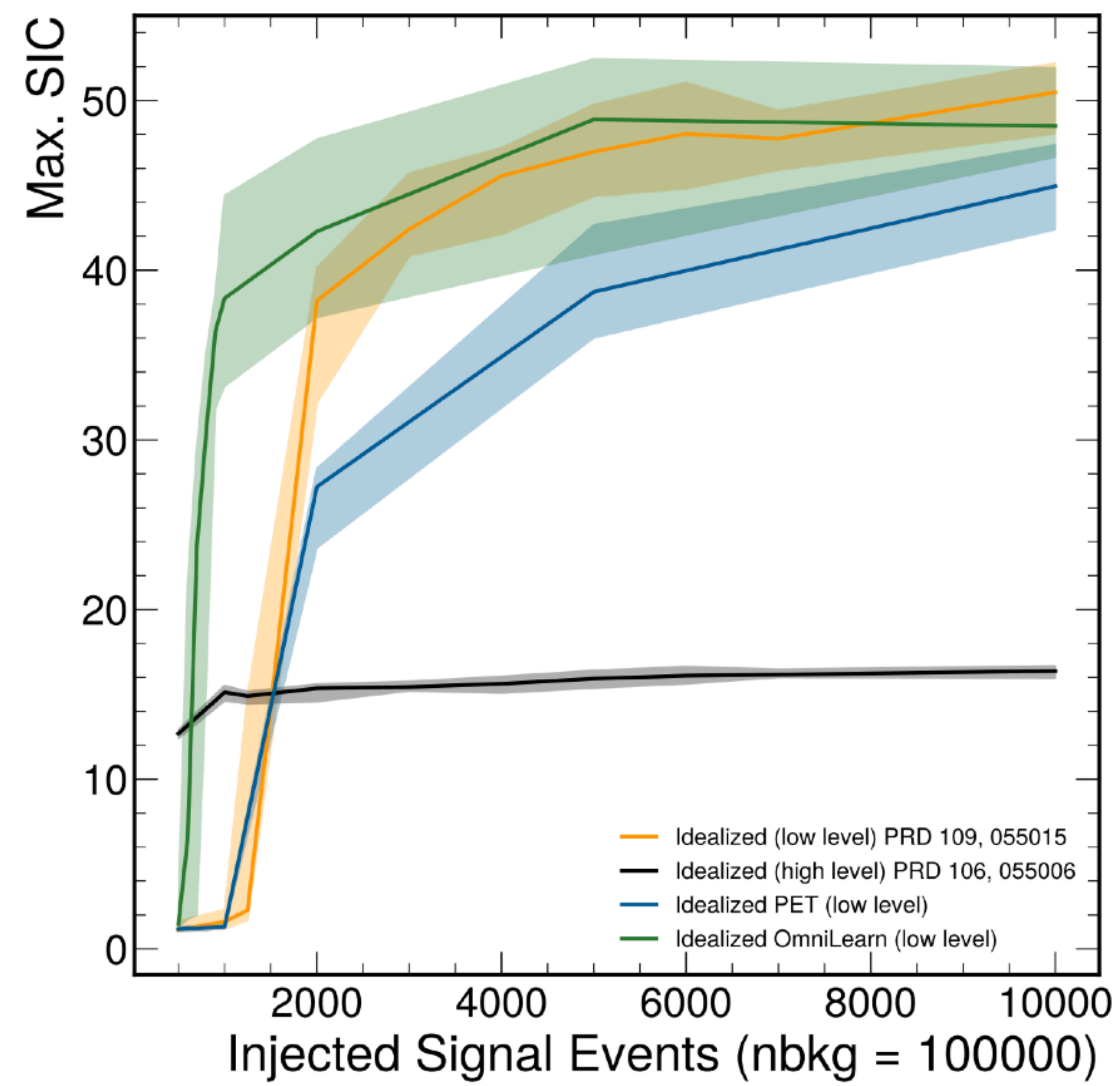
Largely based on autoregressive transformer architecture

Qu et al 2202.03772; Vigl et al 2401.13536

Foundation models

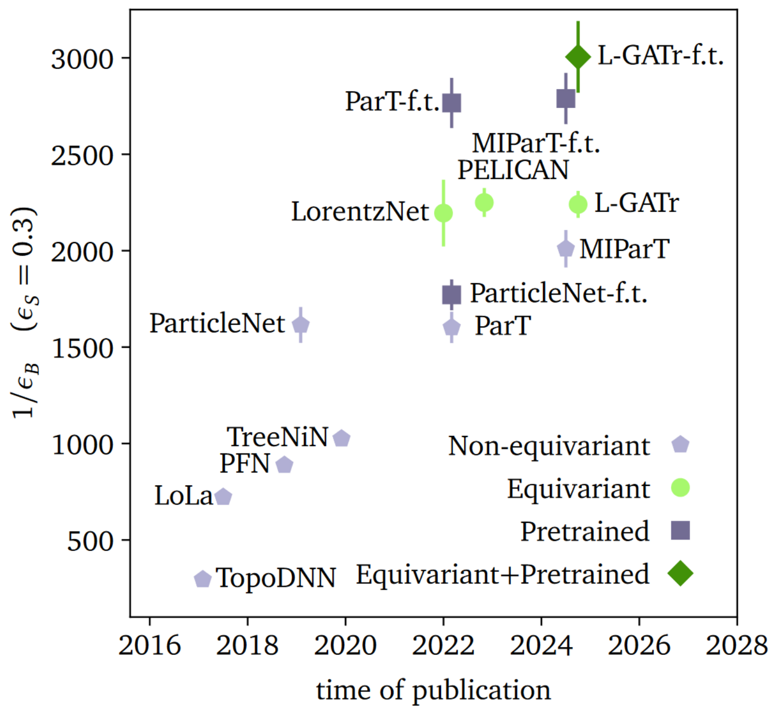


Pre-training on “cheap” unlabelled examples improves supervised classification data efficiency up to 100-1000x

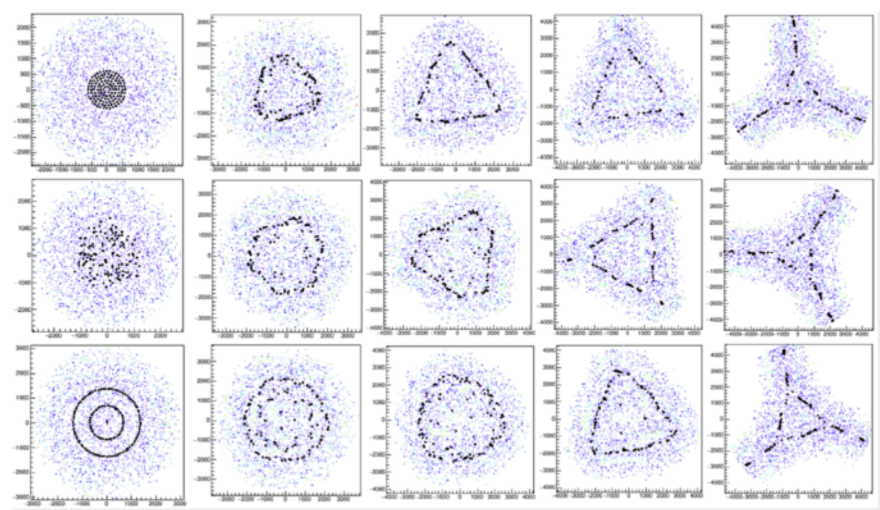


First generalisations across multiple tasks including anomaly detection shown

Birk, Hallin, **GK** 2403.05618; Mikuni, Nachman 2404.16091



Physics or compute

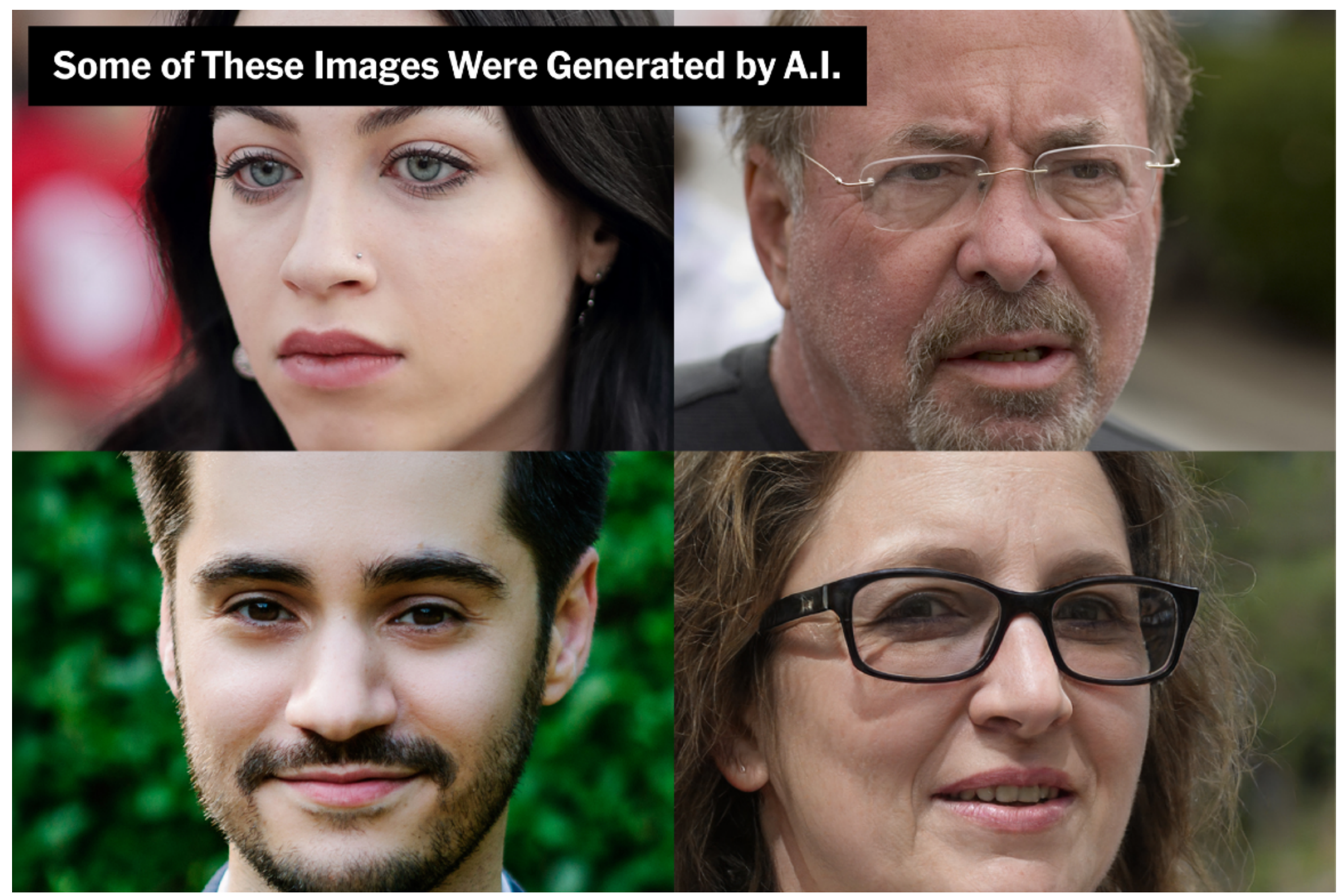


The surrogate revolution

The rise of the AI physicst

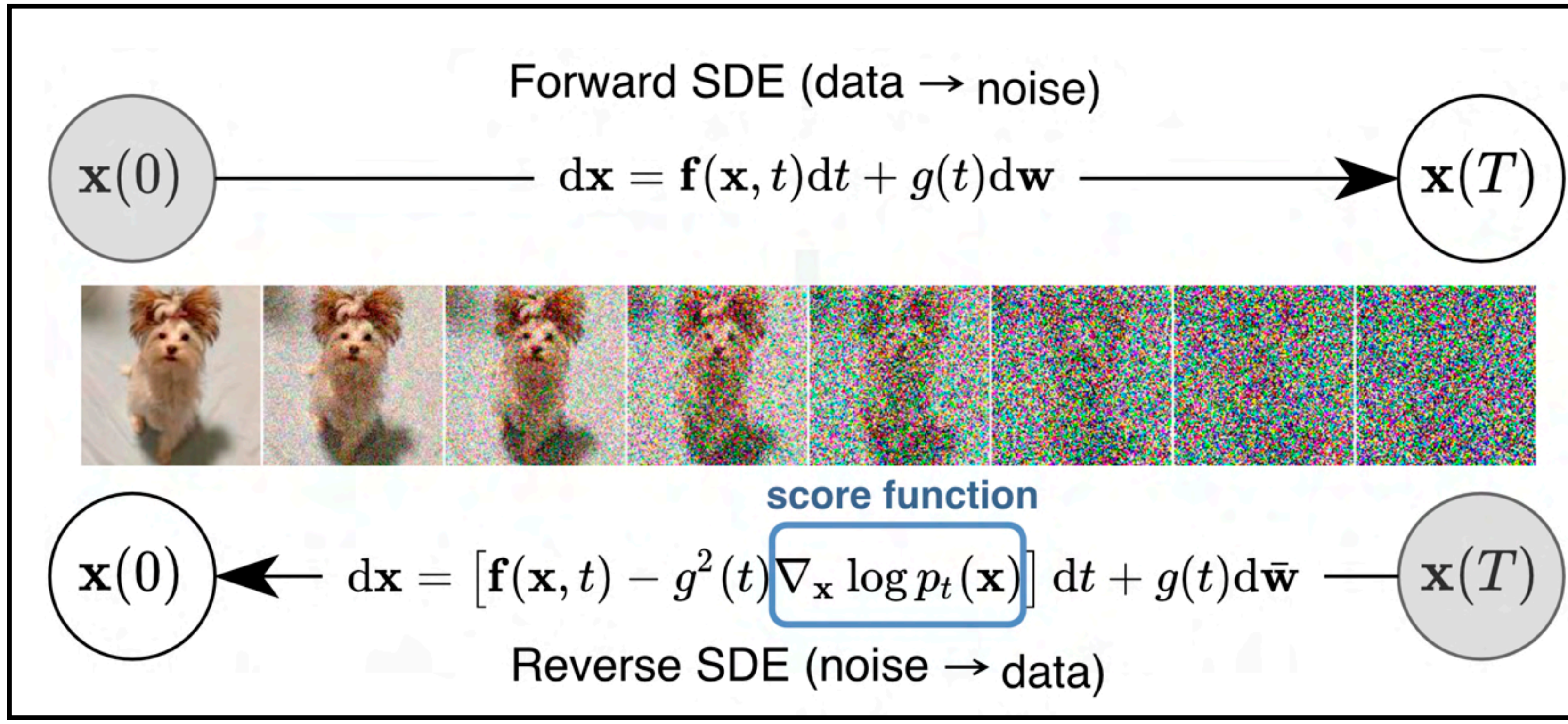
Generative Image Models

Massive progress in the generation of artificial images



2024

Generative Image Models

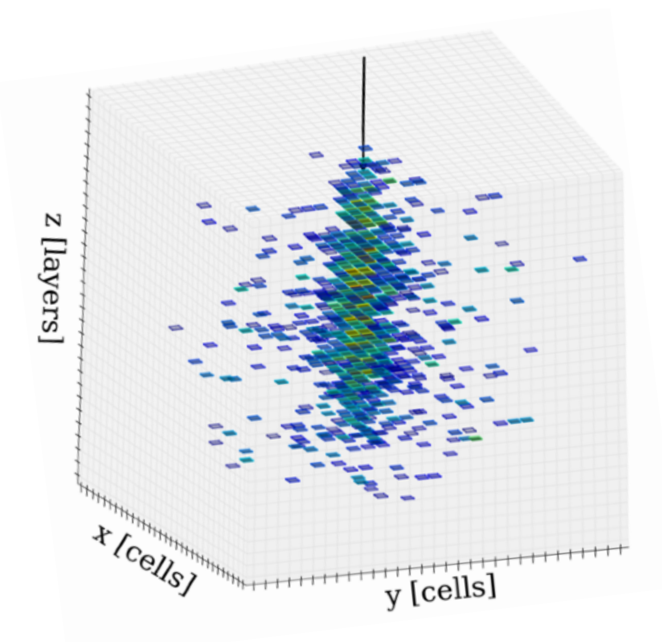


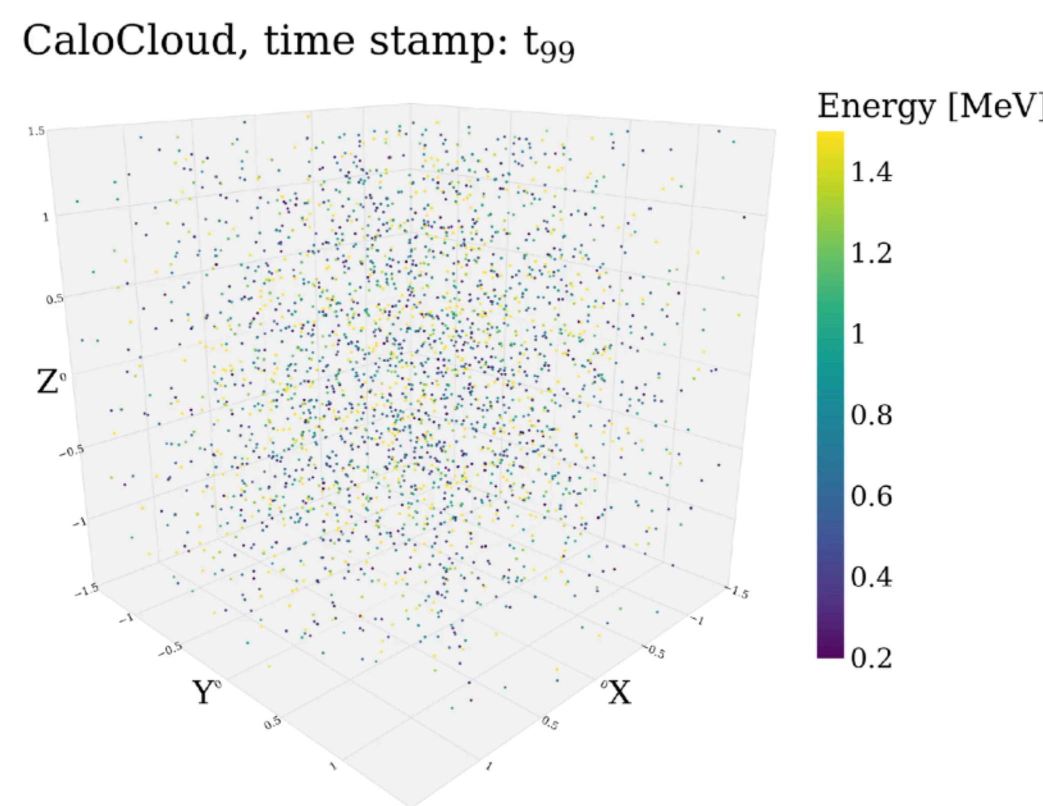
Massive progress in the generation of artificial images

Main driver: Normalising flows and diffusion

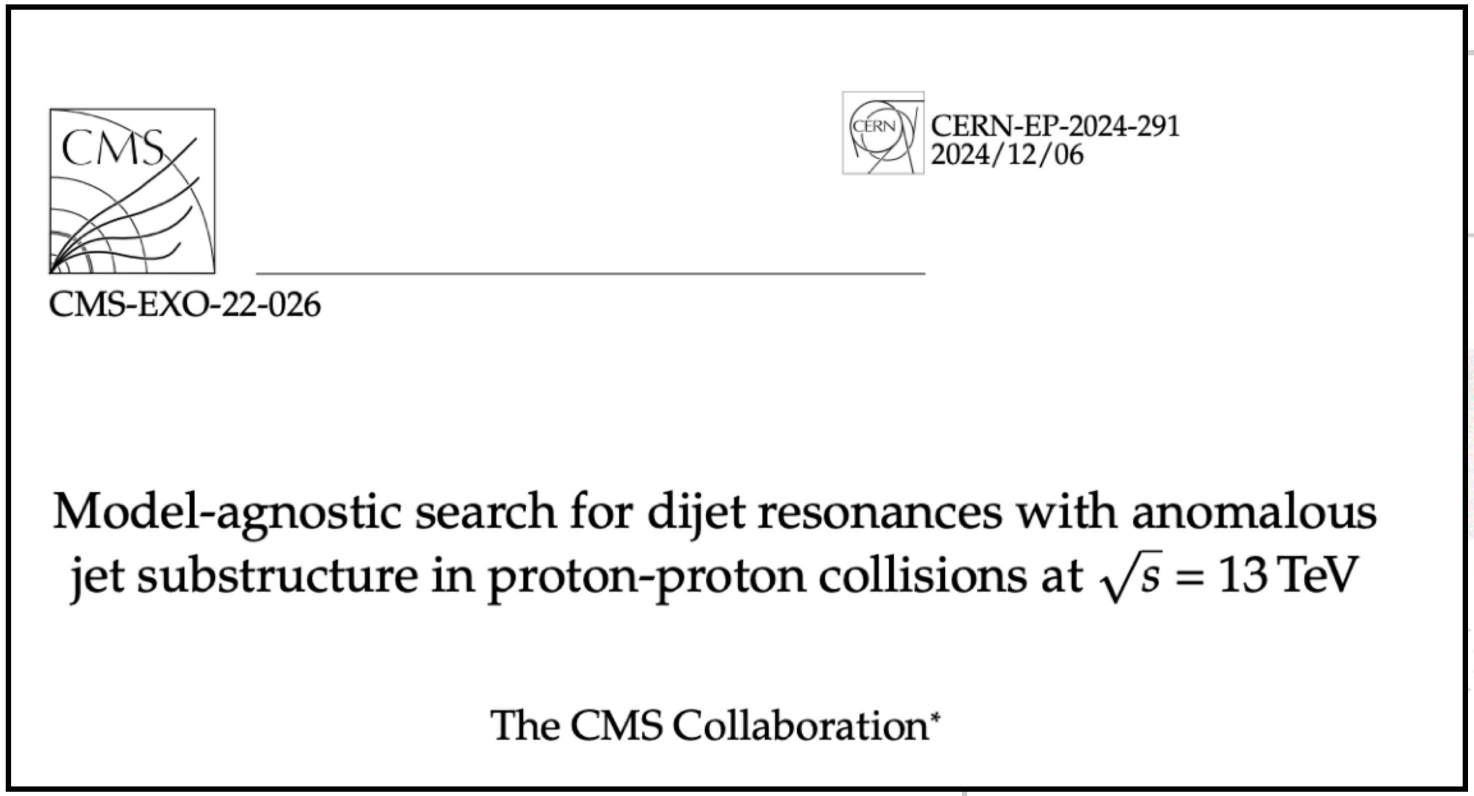
Example: surrogate model for particle interaction in high granularity detector.

Similar approaches for e.g. surrogates for cosmic airshowers. Relevent whenever simulation is expensive.





Generative Image Models

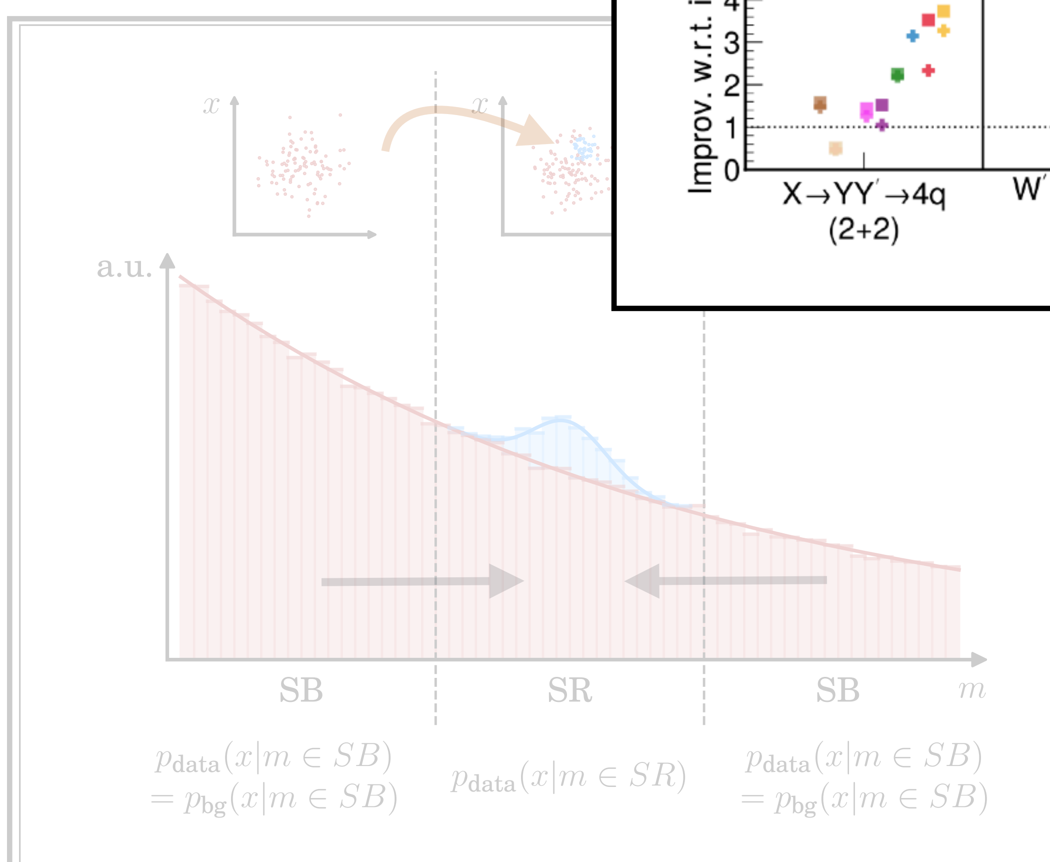

CMS-EXO-22-026

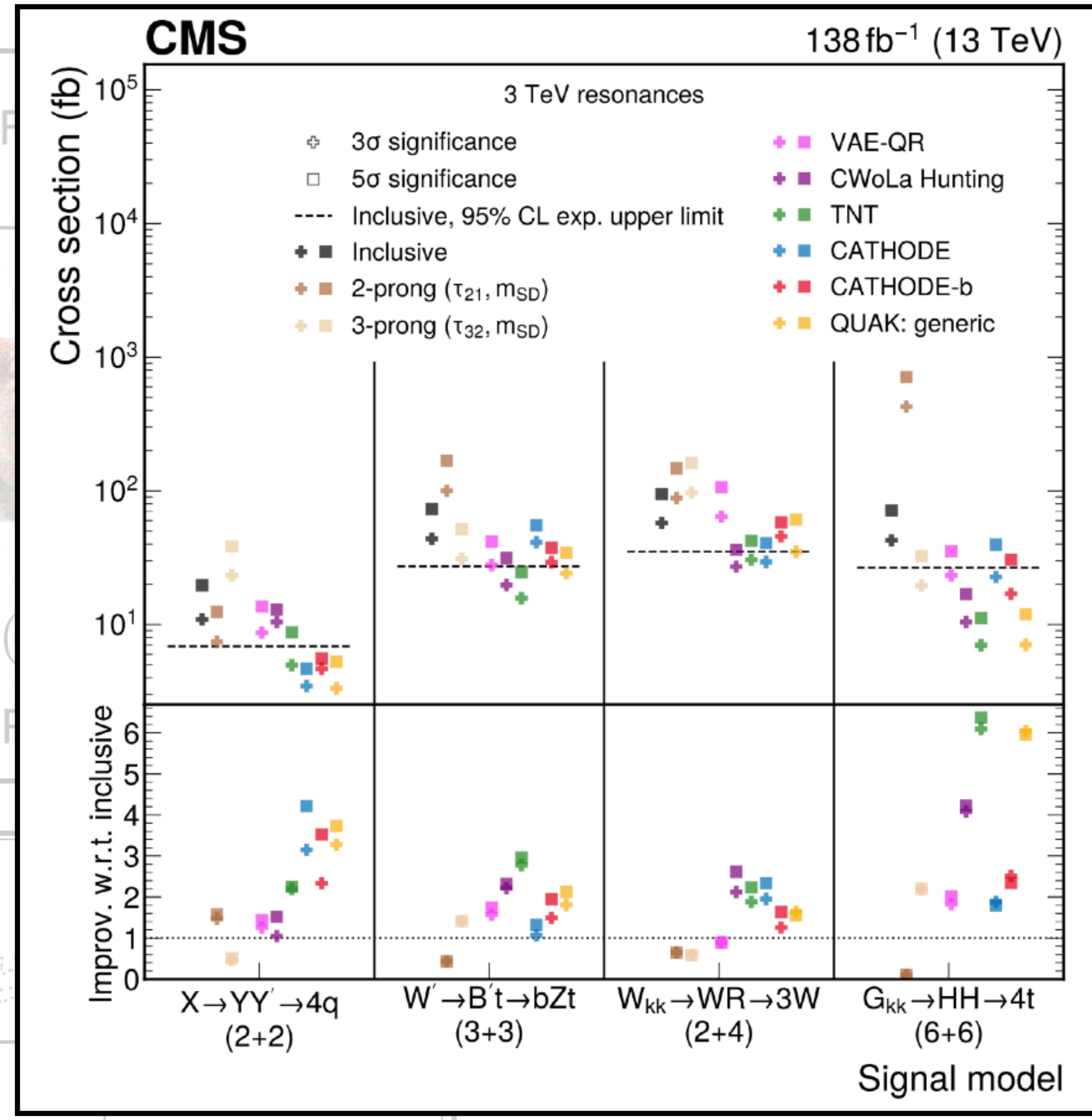
CERN-EP-2024-291
2024/12/06

Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV

The CMS Collaboration*

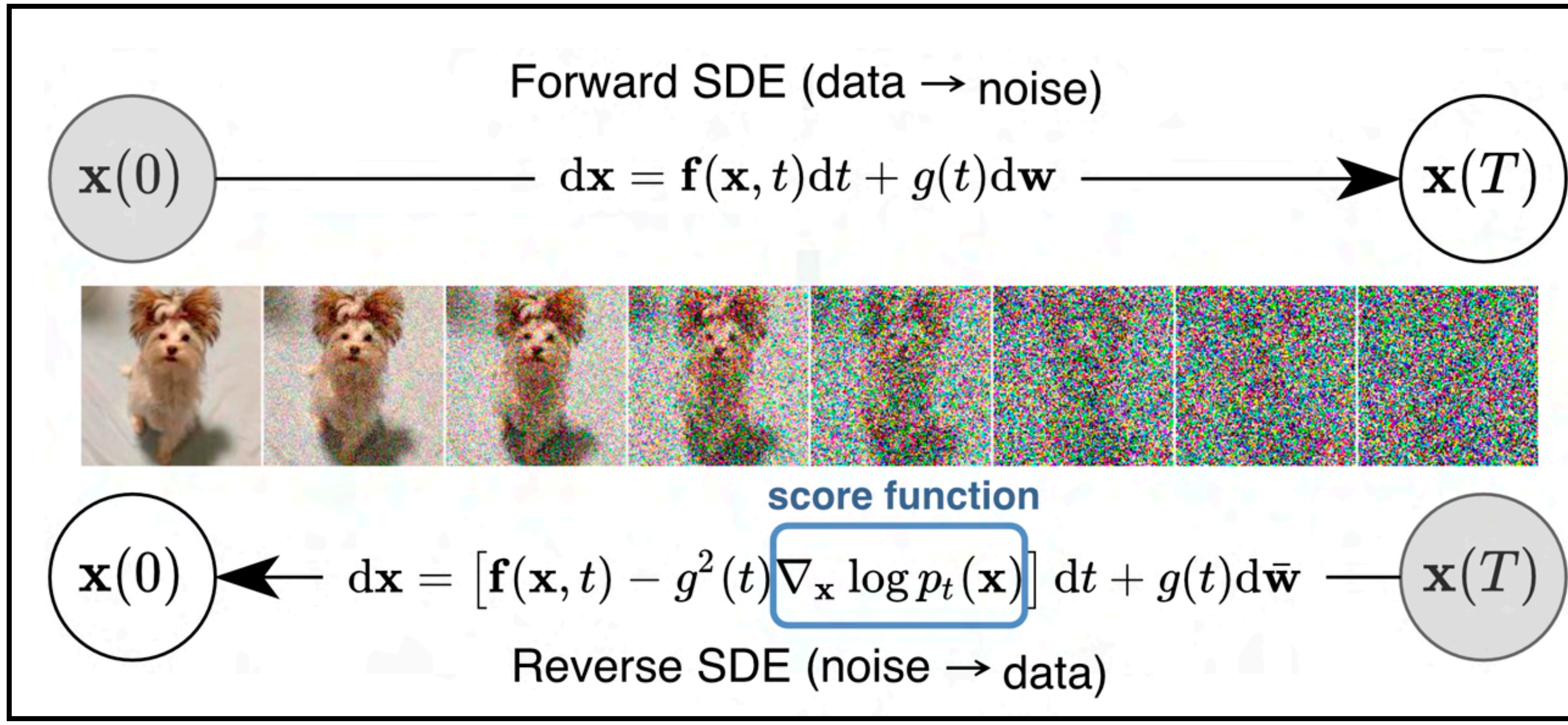
ant for in-situ learning:
se generative
odels to produce
high-dimensional
background
predictions directly
from data



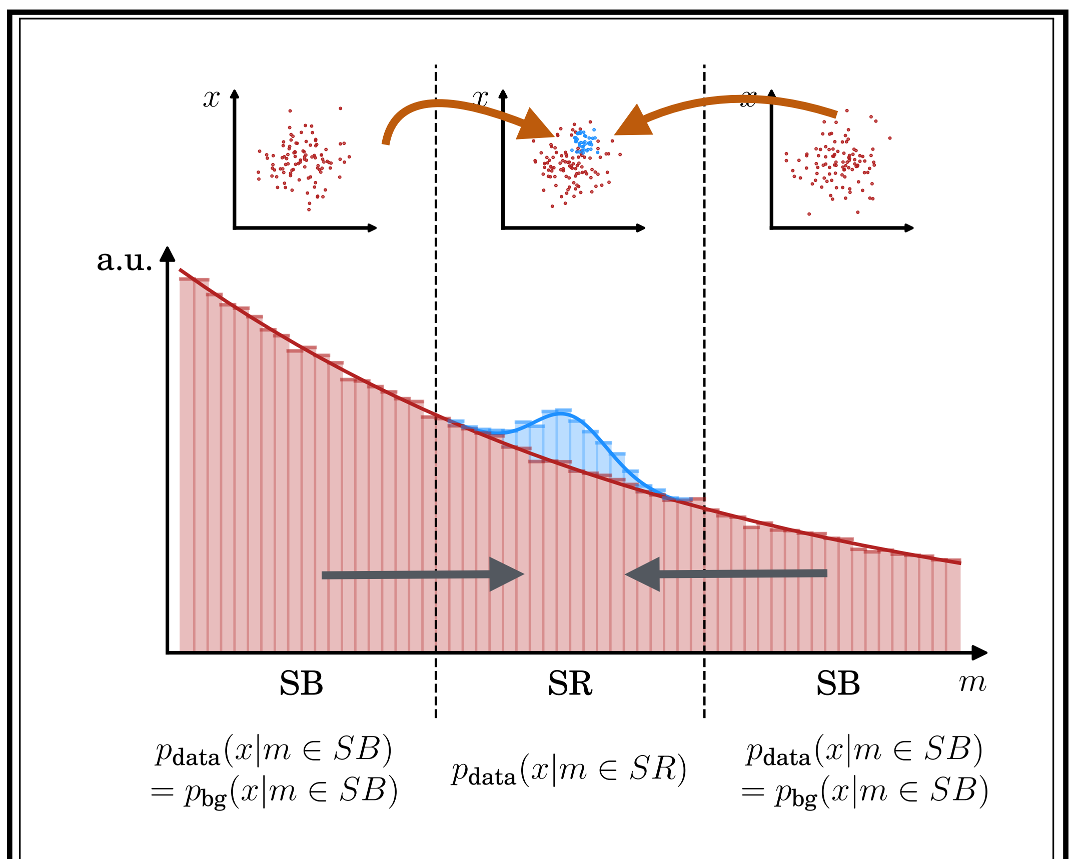


Currently explored in **anomaly detection**, but wider potential?

Generative Image Models



Generative process based on learning transport equation between noise and data.



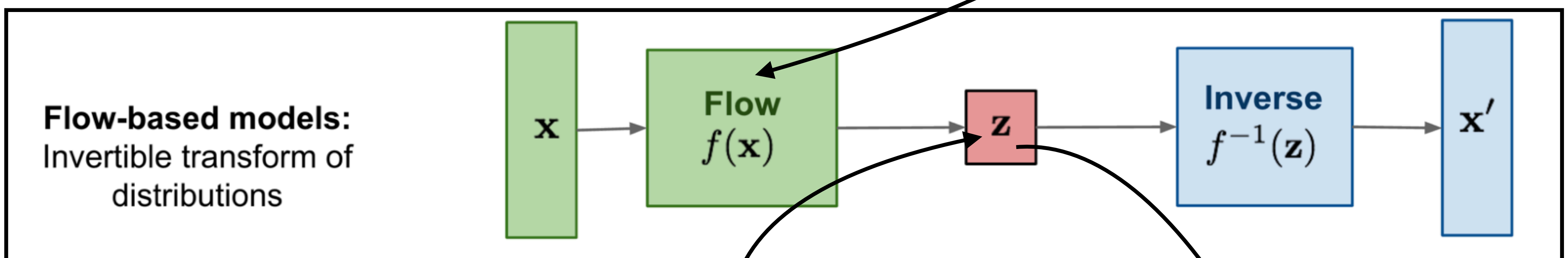
Also relevant for in-situ learning:
Can use generative models to produce high-dimensional background predictions directly from data

Currently explored in anomaly detection, but wider potential?

Generative Image Models

Bijective, easy-to-calculate Jacobian determinant

(Some) generative models directly learn likelihood of data p(x)



Gaussian latent space

Can evaluate f(x) as likelihood

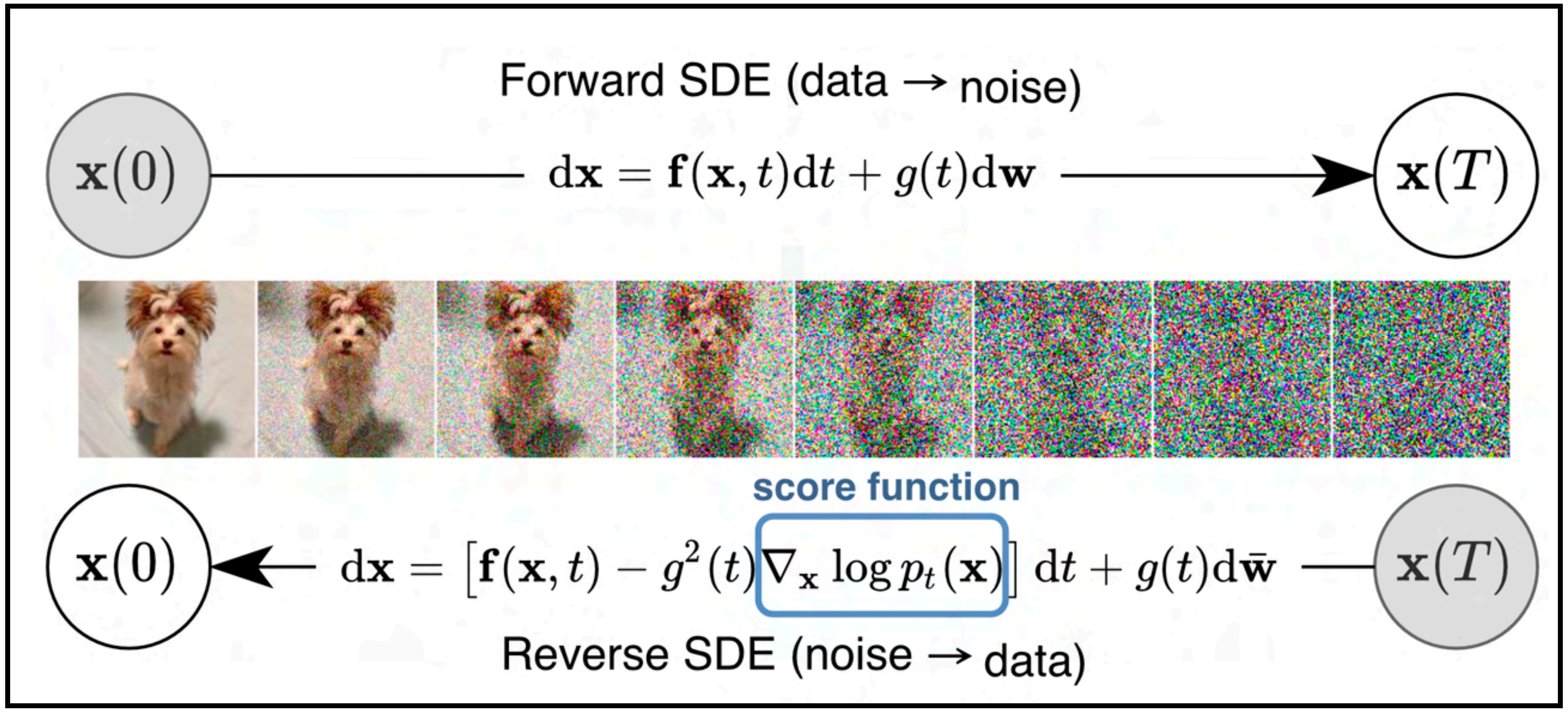
Can use to build likelihood ratios or build likelihood function for measurements

Potential gains from thinking of data not as individual samples but as density function p(x)?

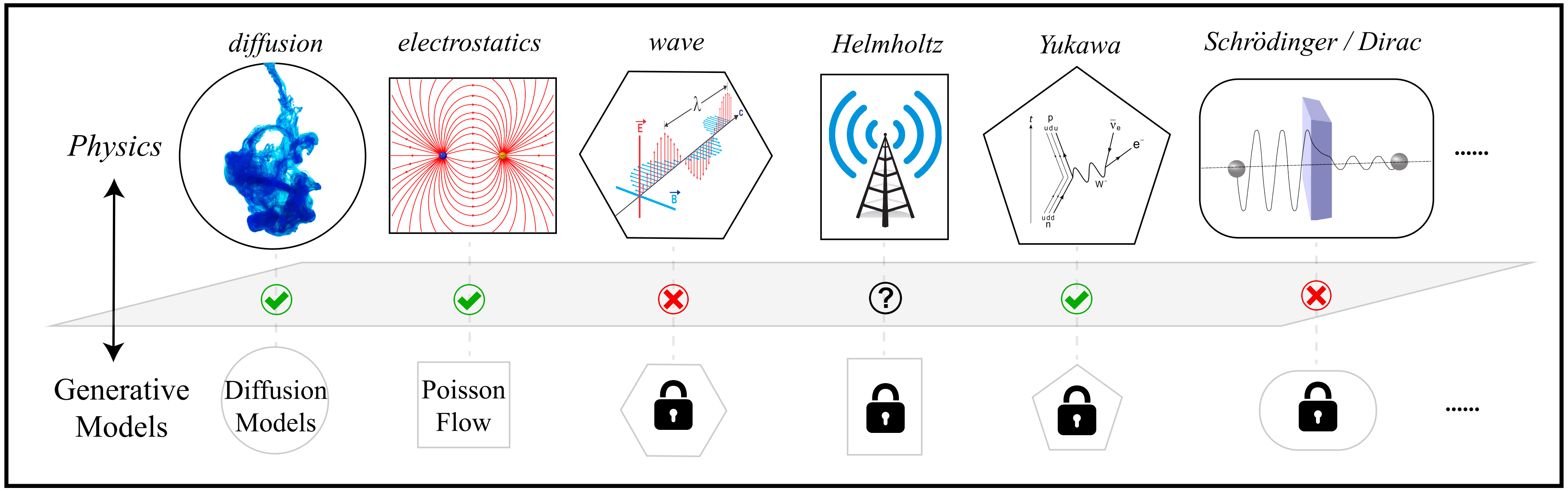
Currently explored in anomaly detection, but wider potential?

Generative Image Models

Generative process based on learning transport equation between noise and data.

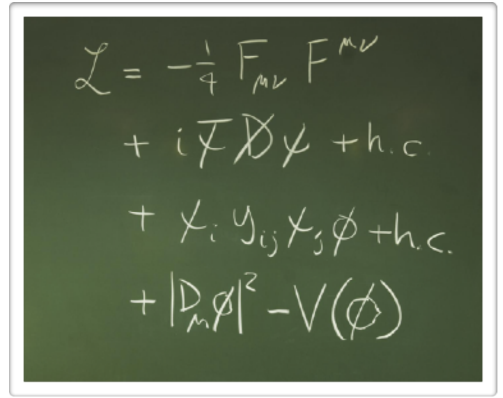


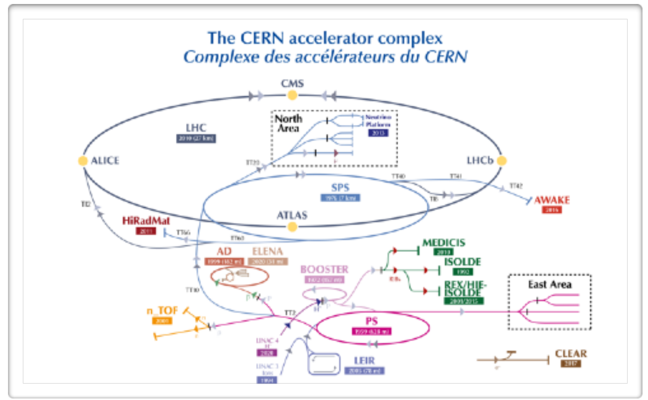
Room for importing more physics knowledge.

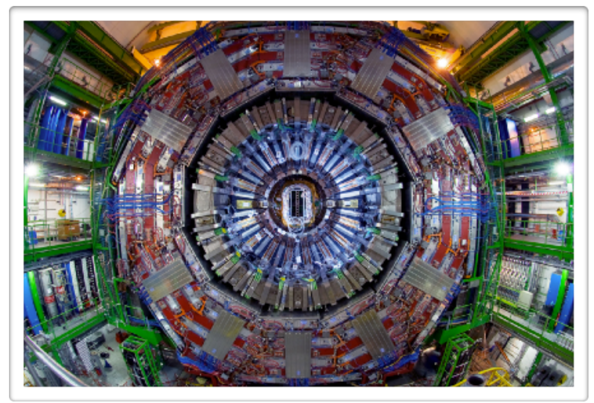


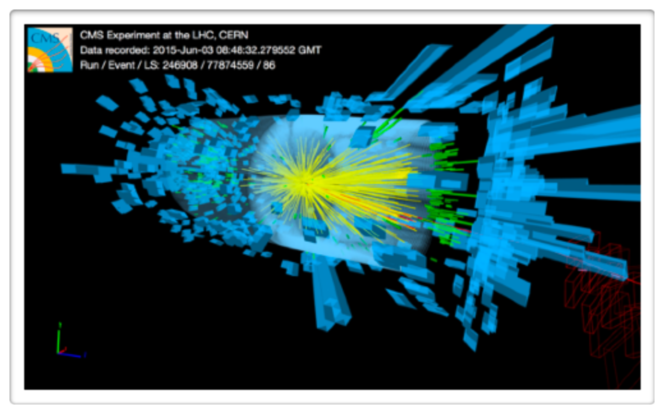
2304.02637

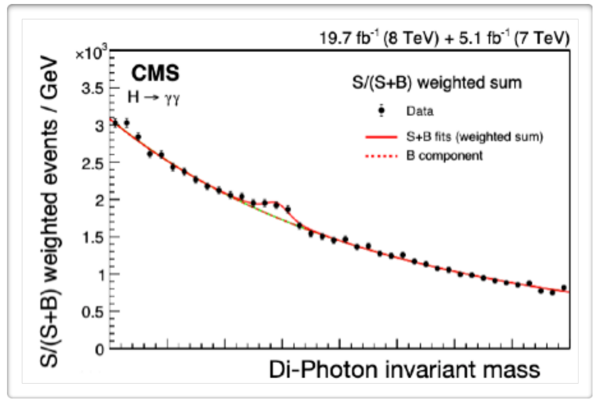
Differentiable versions of all steps in the processing chain

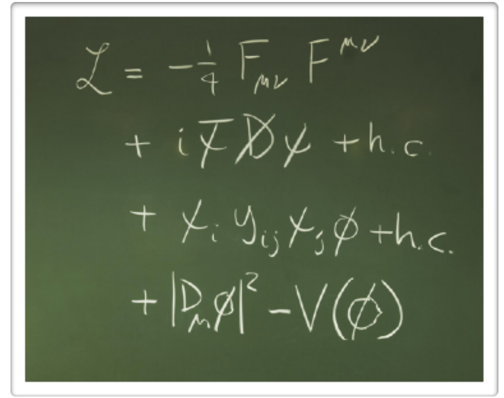


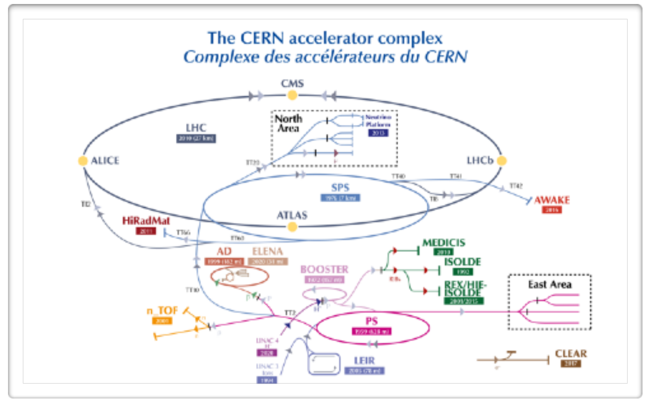


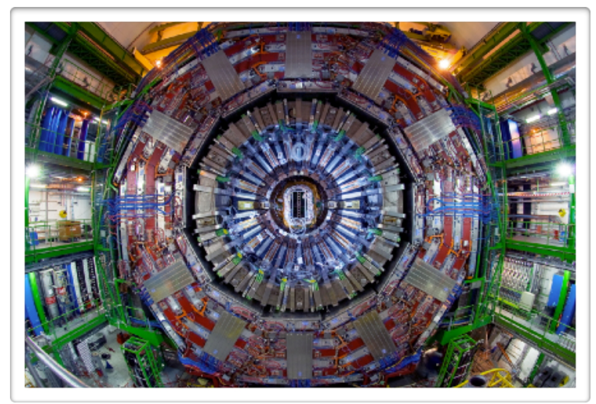


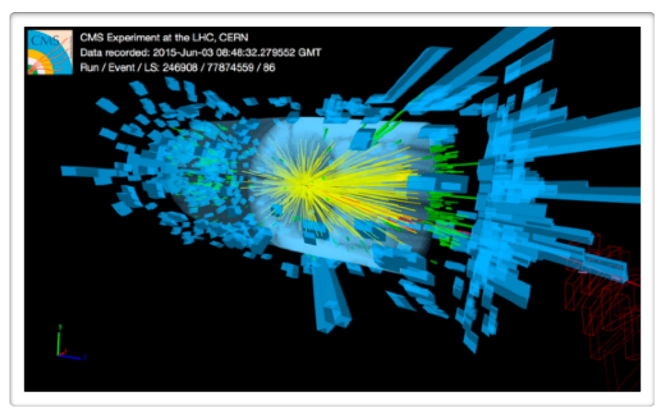


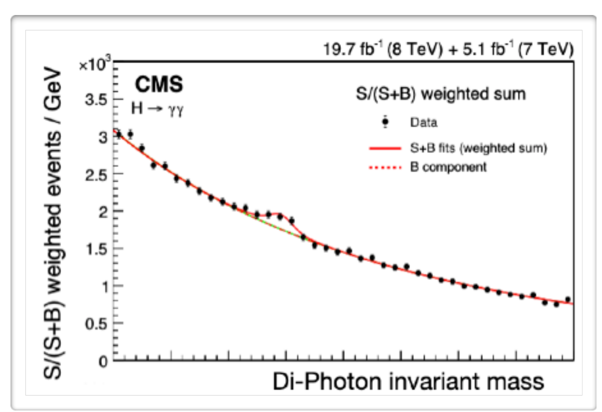










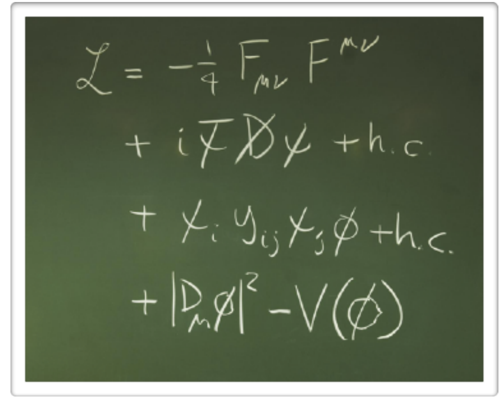


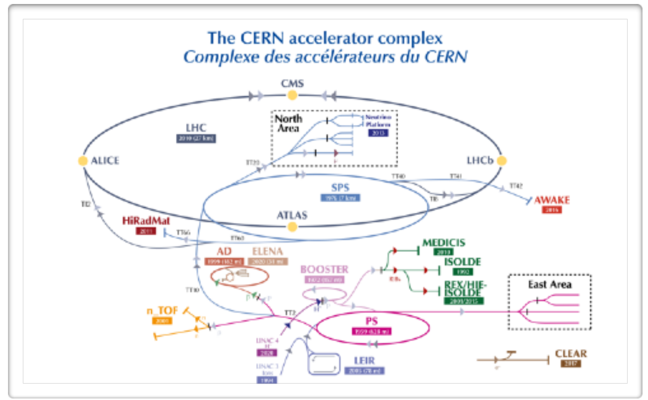
Differentiable versions of all steps in the processing chain

Either as ML-based (surrogate) models

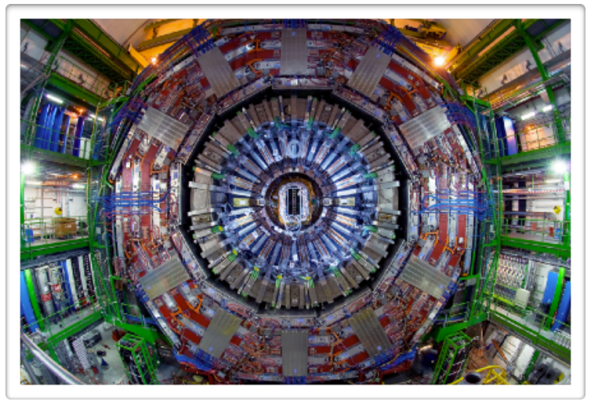
or via e.g. differentiable programming

Heinrich, Kagan 2308.16680

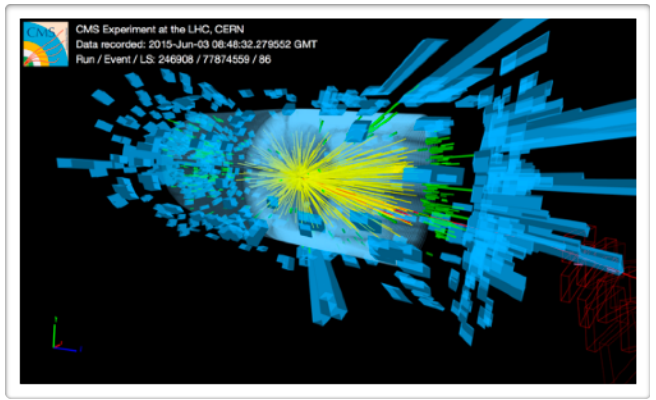




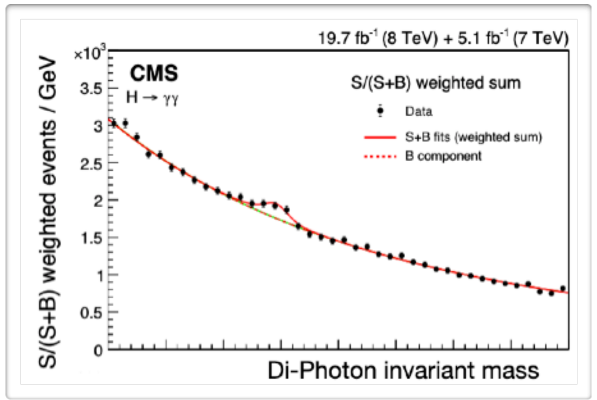
Differentiable versions of all steps in the processing chain



Either as ML-based (surrogate) models

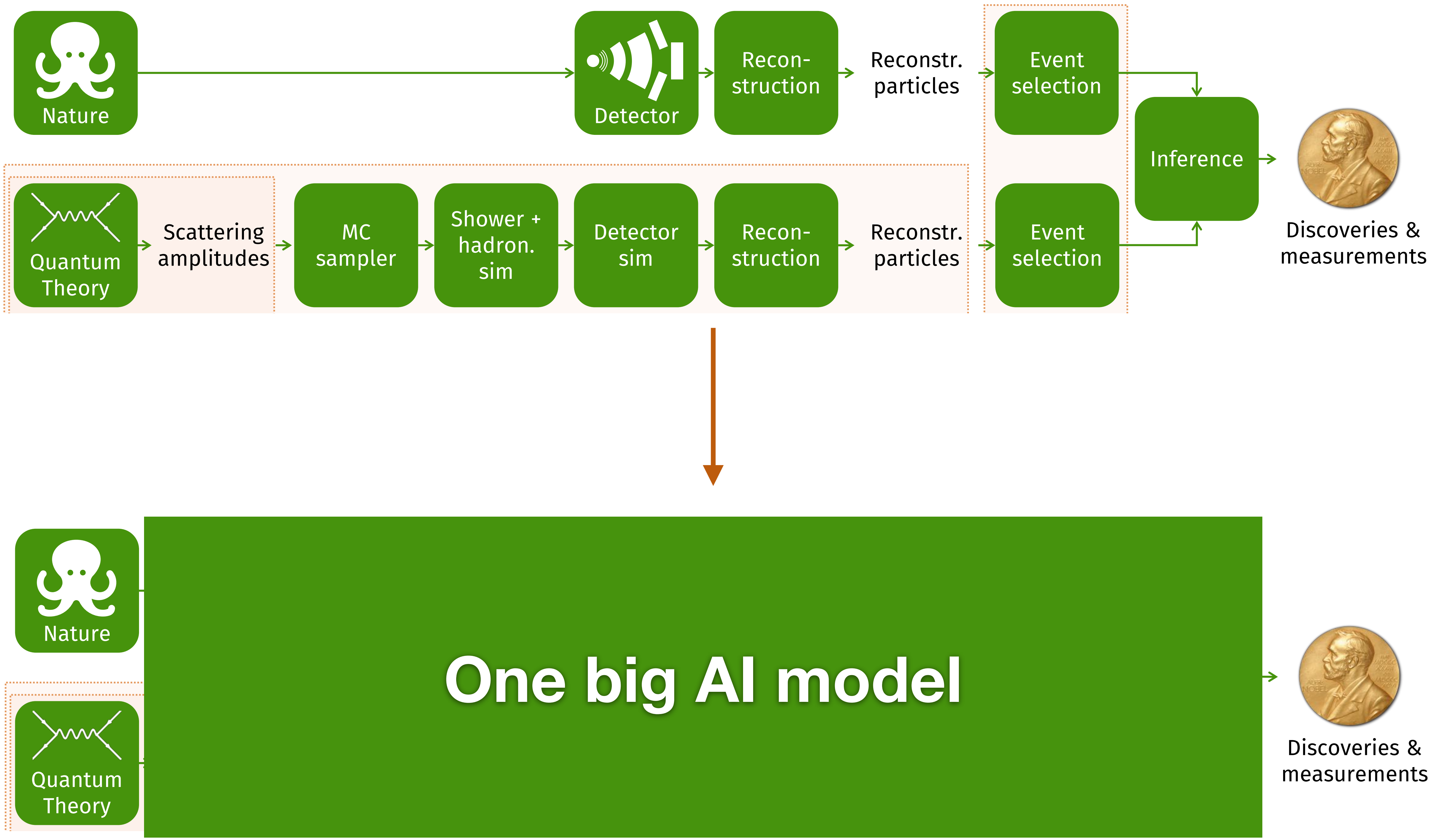


or via e.g. differentiable programming



What can we do with this?

End-to-end learning



End-to-end learning

Some thoughts:

- Sensitivity gain (optimise low-level reconstruction parameters for specific measurment)

End-to-end learning

<u>Some thoughts:</u>

- Sensitivity gain
- Compute cost (execution might be cheaper, training depends on re-usability)

End-to-end learning

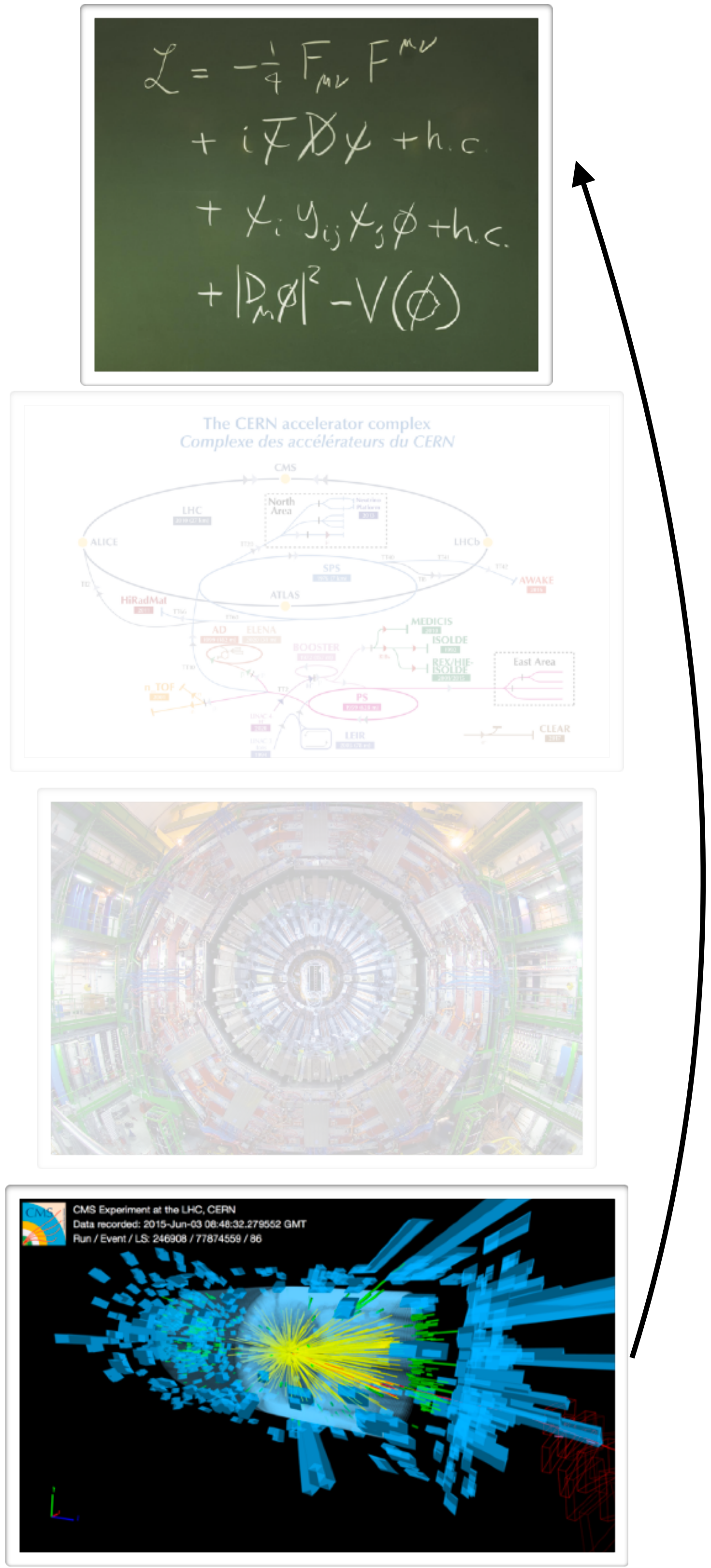
Some thoughts:

- Sensitivity gain
- Compute cost
- Interpretability (ultimate black-box)
- Weak learning signal (scaffolding?)

Direct correspondence e.g. between hard partons and jets.

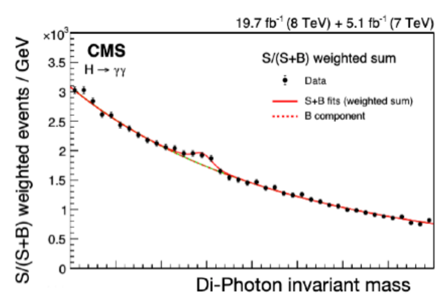
Similar correspondences/learning targers to train/understand very deep models

Inference



Goal: Learn parameters of theory (e.g. couplings) directly from high-dimensional data

Inference



Summary Statistics

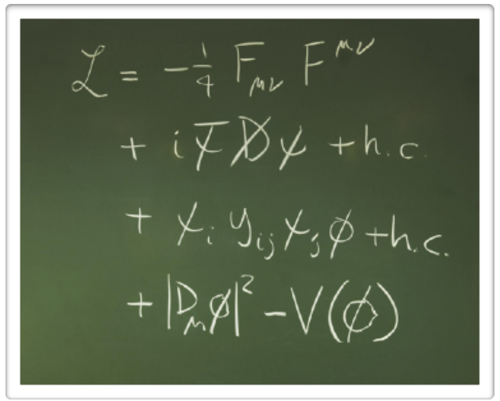
Likelihood Learning (e.g. flows or cINNs)

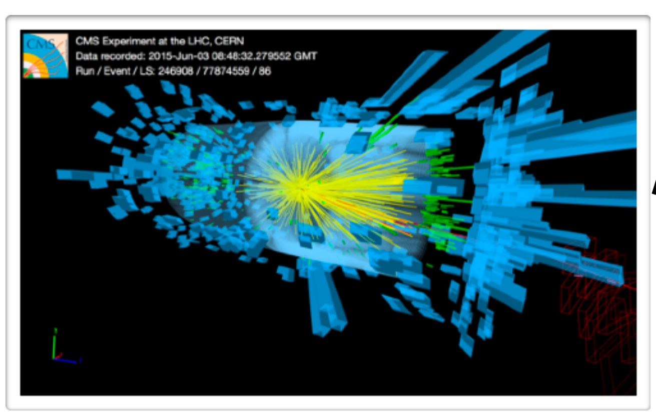
Likelihood ratio trick (e.g. CARL, swyft)

Integration (e.g. MadMiner)

Cranmer, Brehmer, Louppe 1911.01429; Cranmer, Brehmer 2010.06439

Inference

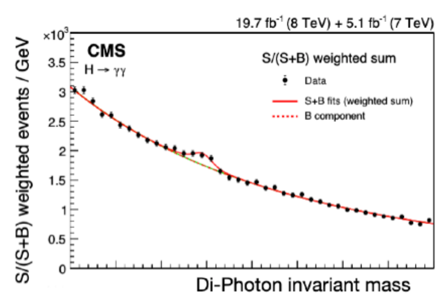




Goal: Learn parameters of theory (e.g. couplings) directly from high-dimensional data

No exact likelihood, but forward simulations available: likelihood-free / simulation based inference

Inference



Summary Statistics

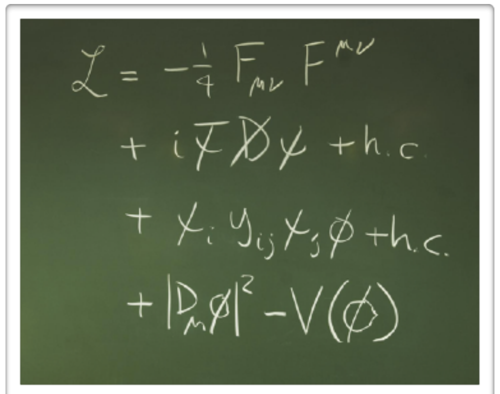
Likelihood Learning (e.g. flows or cINNs)

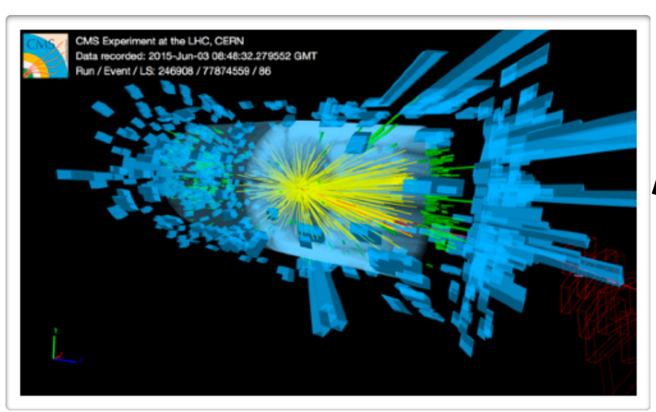
Likelihood ratio trick (e.g. CARL, swyft)

Integration (e.g. MadMiner)

Cranmer, Brehmer, Louppe 1911.01429; Cranmer, Brehmer 2010.06439

Inference

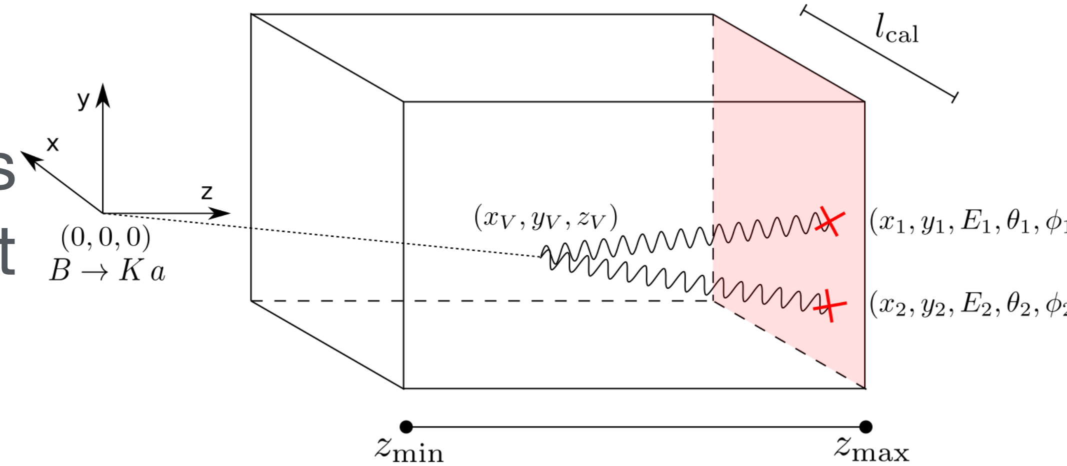


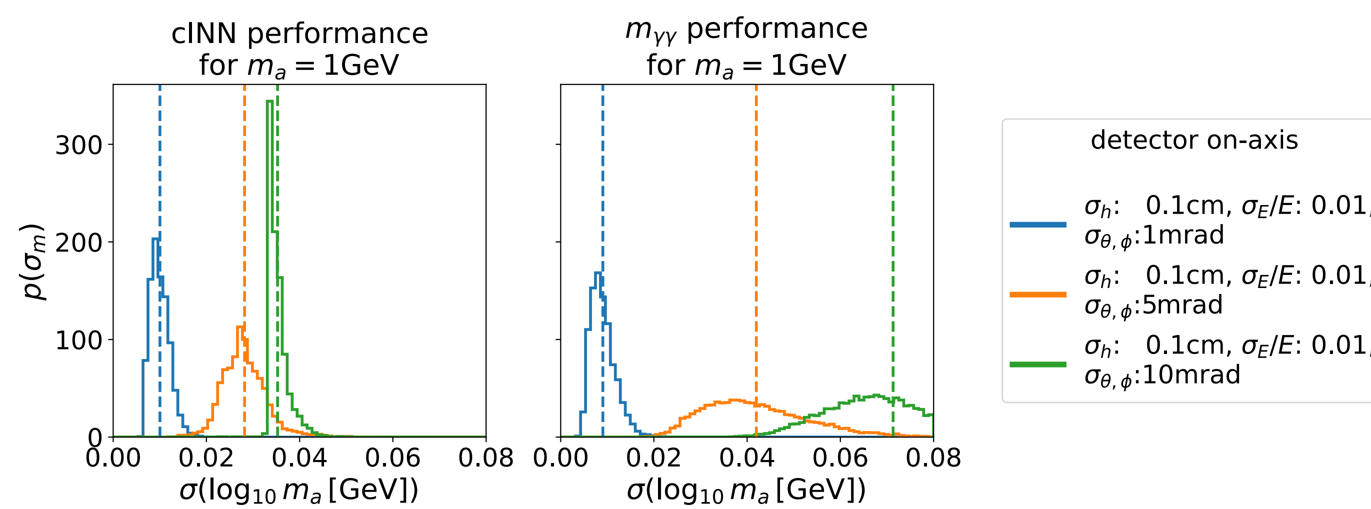


Inference

Reconstructing axion-like particles from beam dumps using cINN approach

Infer axion mass from measurement

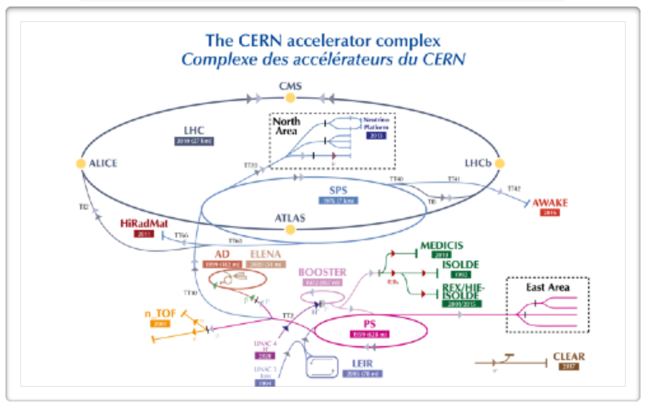


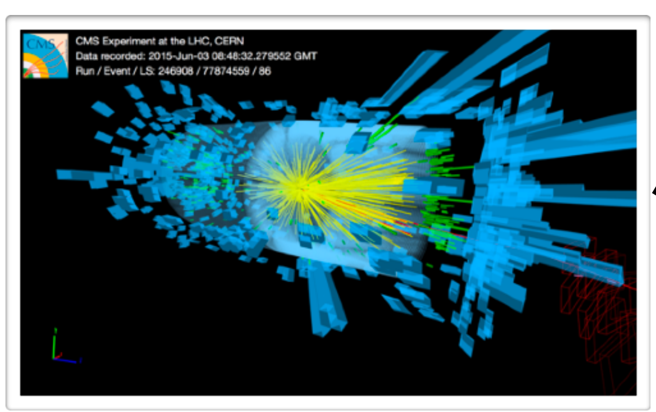


More stable vs resolution than traditional approach

Morandini, Ferber, Kahlhoefer 2308.01353; See also recent ATLAS result: 2412.01600

Experiment Design

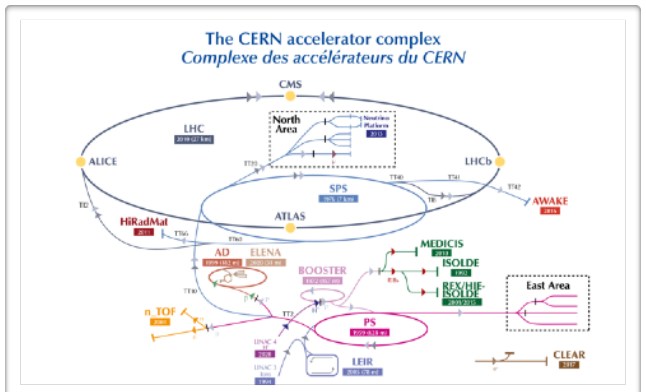


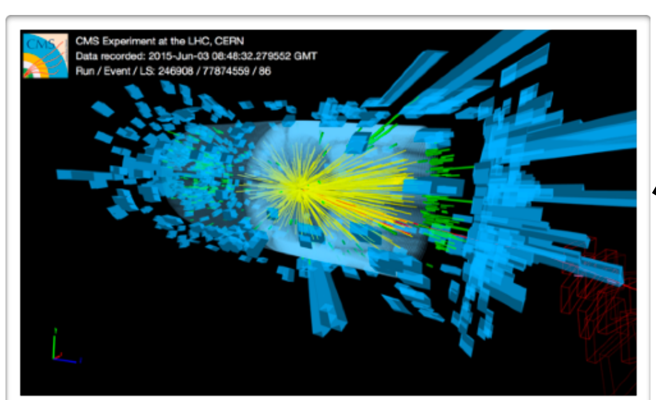


Experimental Design

Automatically learn to arrange sensors given a physics target

Experiment Design



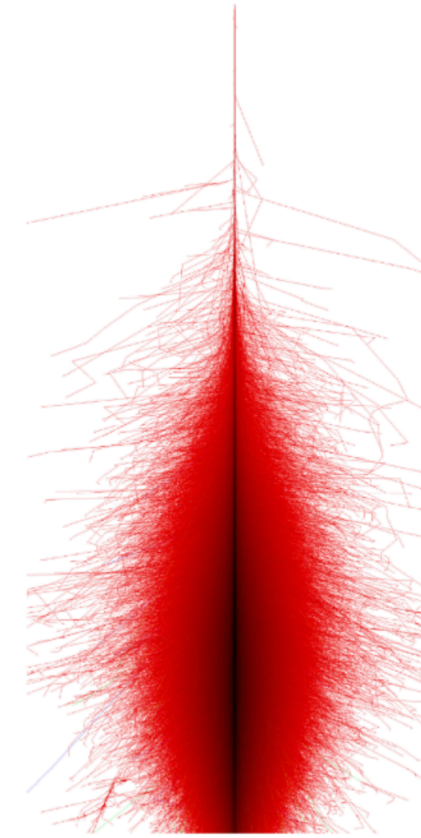


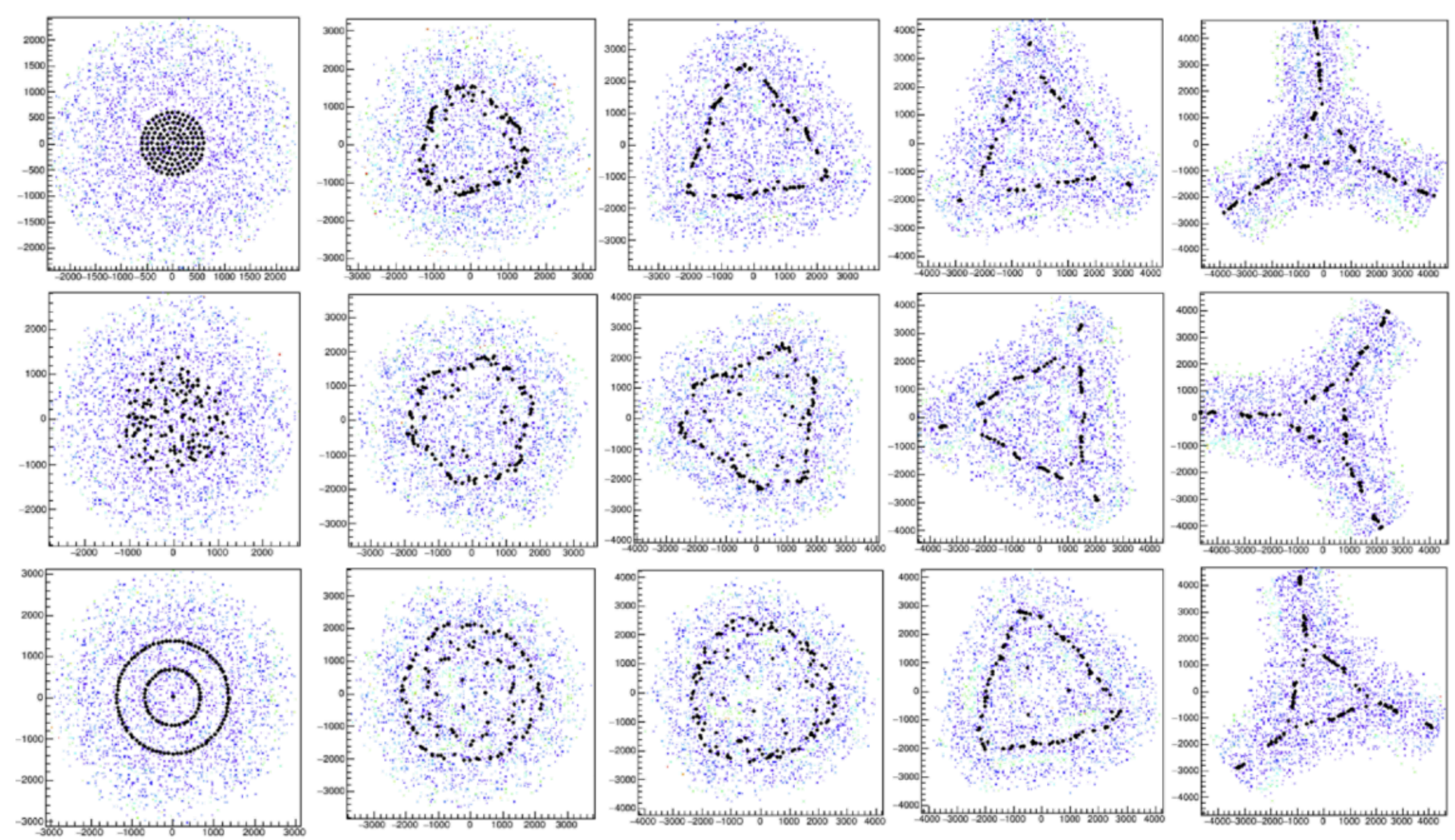
Experimental Design

Automatically learn to arrange sensors given a physics target

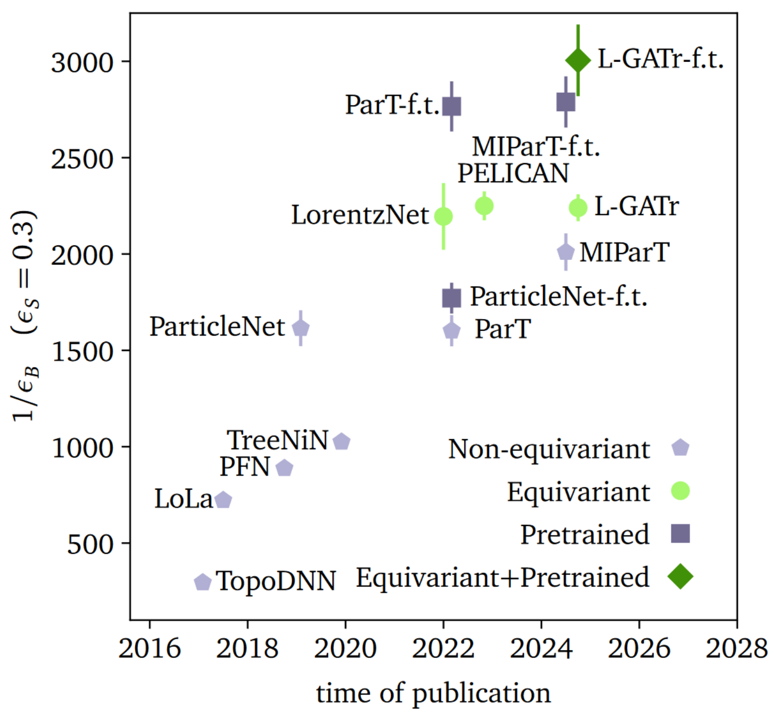
Example tuning positions of detectors for a gamma ray observatory

(What about future colliders?)

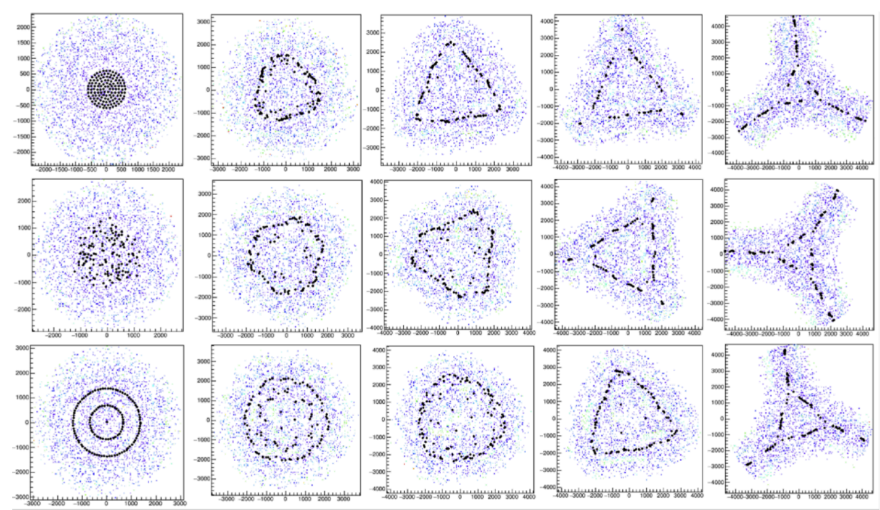




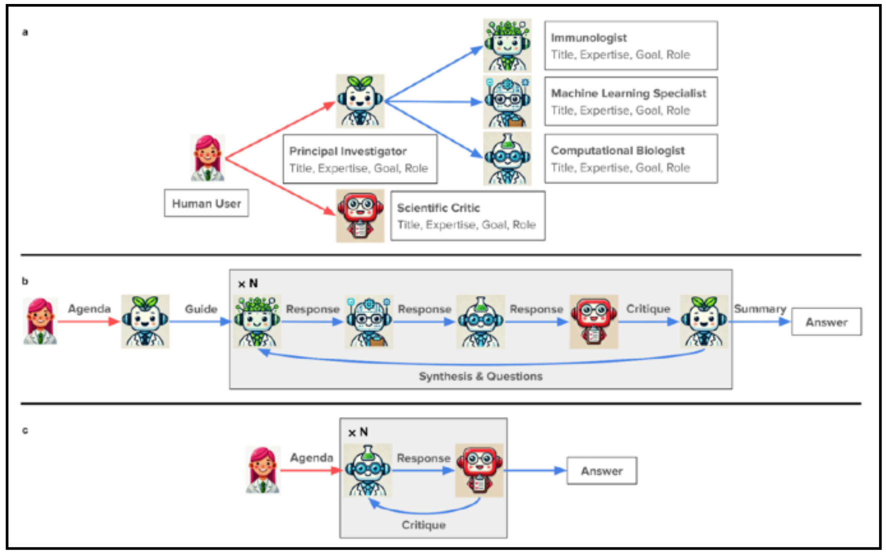
https://mode-collaboration.github.io/'; Dorigo et al 2310.01857



Physics or compute



The surrogate revolution

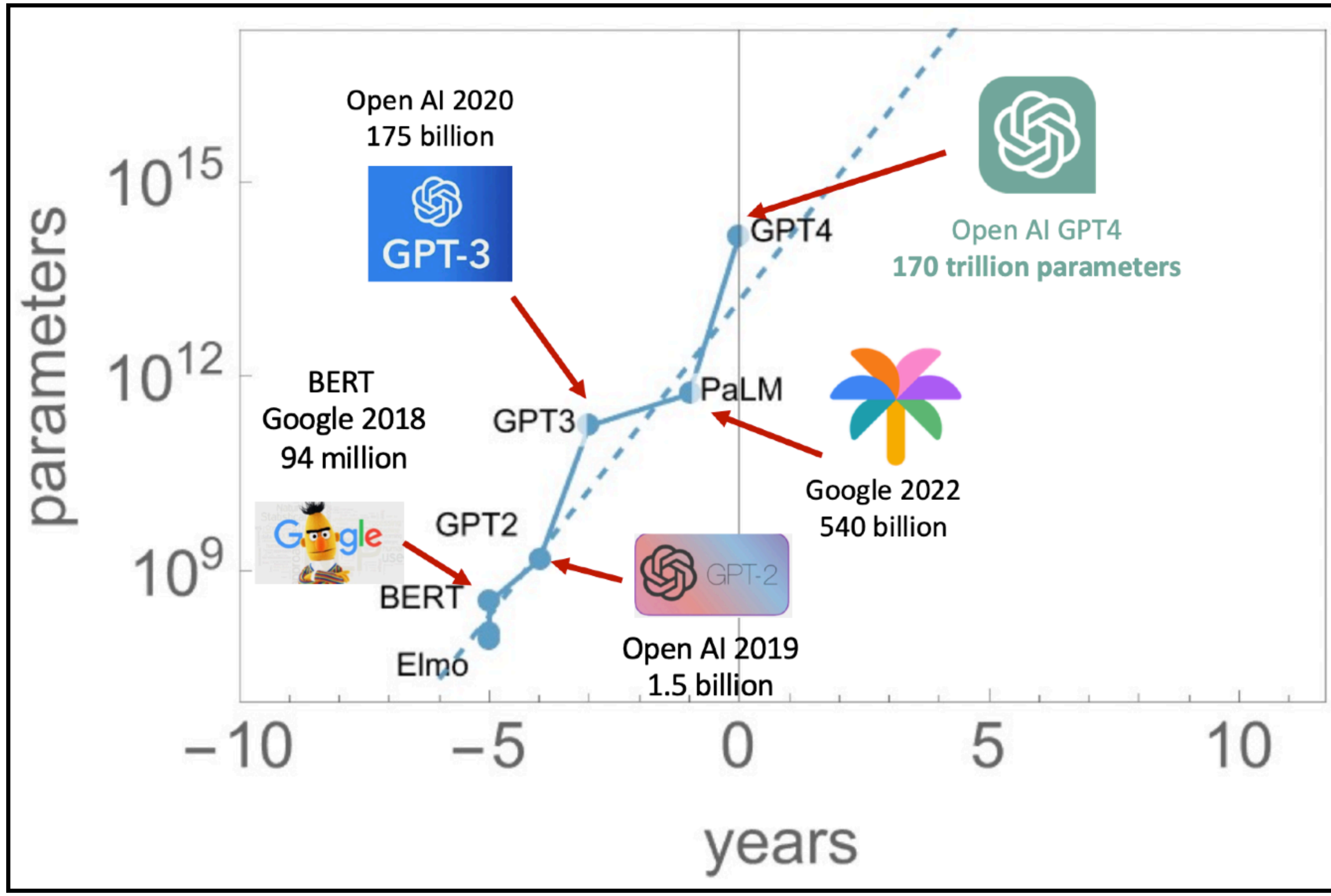


The rise of the AI physicst

Large Language Models

Most impressive growth:
Large language models

Impact for physics?



Schwartz EuCAIFCon ’24;

Large Language Models

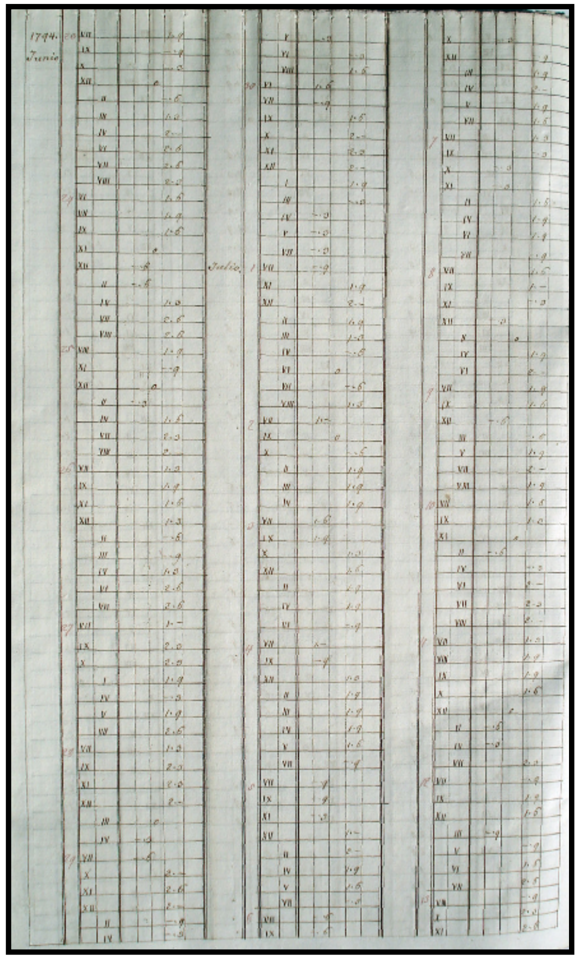
Most impressive growth:
Large language models

Impact for physics?

Focus of AI in Physics on data analysis:

Better algorithms to handle numerical data
→Fuels discovery and insight
→Much more to do
→But generally not ideal for language models

Historical astronomical observation data



Large Language Models

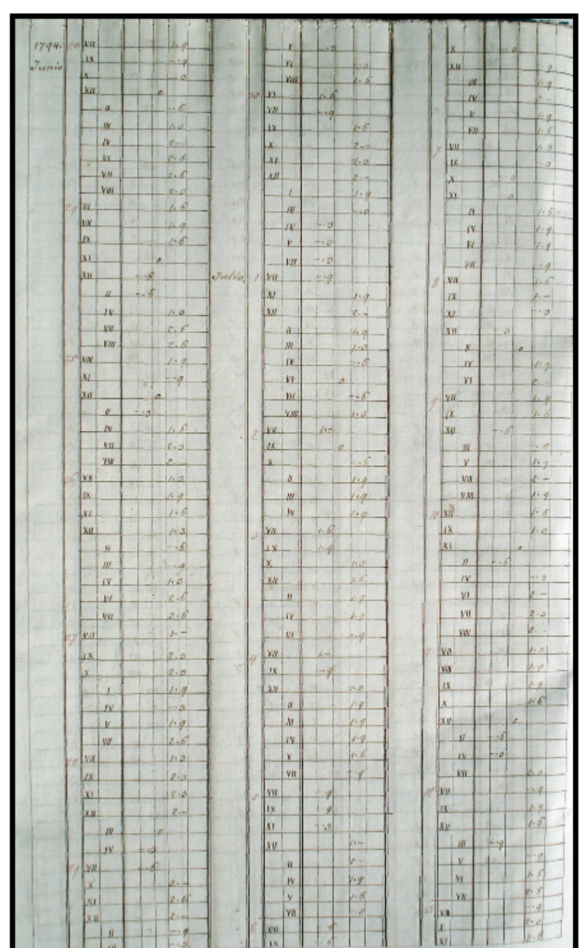
Most impressive growth:
Large language models

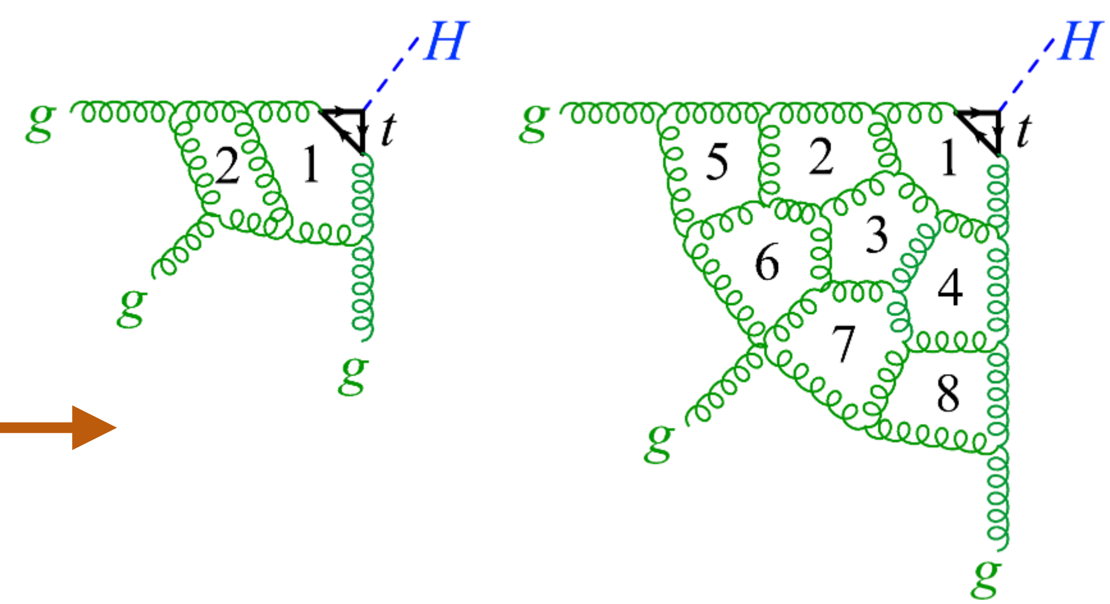
Impact for physics?

Focus of AI in Physics on data analysis:

Better algorithms to handle numerical data

What about symbolic problems instead?





2405.06107

Large Language Models

Most impressive growth:
Large language models

Impact for physics?

Focus of AI in Physics on data analysis:

Better algorithms to handle numerical data

What about symbolic problems instead?

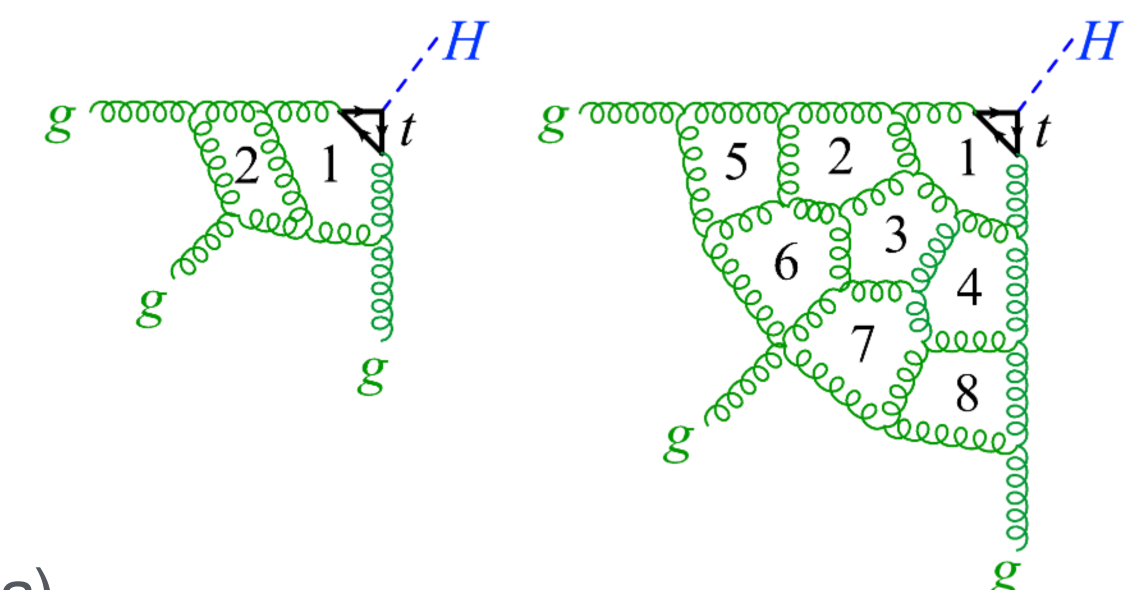
Example:
Generalised Polylogarithms

$$\mathcal{S}[\mathcal{F}^{(L)}] = \sum_{l_{i_1},\ldots,l_{i_{2L}} \in \mathcal{L}_m} C^{l_{i_1},\ldots,l_{i_{2L}}}\, l_{i_1} \otimes \cdots \otimes l_{i_{2L}}$$

Integer coefficients

Symbols
(functions of momenta)

$$\mathcal{L}_{\mathrm{3gFF}} = \{a, b, c, d, e, f\}$$



2405.06107

Large Language Models

Most impressive growth:
Large language models

Impact for physics?

Focus of AI in Physics on data analysis:

Better algorithms to handle numerical data

What about symbolic problems instead?

Example:
Generalised Polylogarithms

$$\mathcal{S}[\mathcal{F}^{(L)}] = \sum_{l_{i_1},\ldots,l_{i_{2L}} \in \mathcal{L}_m} C^{l_{i_1},\ldots,l_{i_{2L}}}\, l_{i_1} \otimes \cdots \otimes l_{i_{2L}}$$

Integer coefficients

Symbols
(functions of momenta)

$$\mathcal{L}_{\text{3gFF}} = \{a, b, c, d, e, f\}$$

$$\begin{aligned}
\mathcal{S}[\mathcal{F}^{(1)}_{\text{3gFF}}] &= (-2)\Big[b\otimes d + c\otimes e + a\otimes f + b\otimes f + c\otimes d + a\otimes e\Big]\,,\\
\mathcal{S}[\mathcal{F}^{(2)}_{\text{3gFF}}] &= 8\Big[b\otimes d\otimes d\otimes d + c\otimes e\otimes e\otimes e + a\otimes f\otimes f\otimes f\\
&\qquad + b\otimes f\otimes f\otimes f + c\otimes d\otimes d\otimes d + a\otimes e\otimes e\otimes e\Big]\\
&\quad + 16\Big[b\otimes b\otimes b\otimes d + c\otimes c\otimes c\otimes e + a\otimes a\otimes a\otimes f\\
&\qquad + b\otimes b\otimes b\otimes f + c\otimes c\otimes c\otimes d + a\otimes a\otimes a\otimes e\Big]\,.
\end{aligned}$$

3-gluon form factor at 1- and 2-loop

(Billions at 8-loop)

2405.06107

Large Language Models

Focus of AI in Physics on data analysis:

Most impressive growth:
Large language models

Better algorithms to handle numerical data

Impact for physics?

What about symbolic problems instead?

Example:
Generalised Polylogarithms

$$\mathcal{S}[\mathcal{F}^{(L)}] = \sum_{l_{i_1},\ldots,l_{i_{2L}} \in \mathcal{L}_m} C^{l_{i_1},\ldots,l_{i_{2L}}} \, l_{i_1} \otimes \cdots \otimes l_{i_{2L}}$$

Integer coefficients

Symbols
(functions of momenta)

$$\mathcal{L}_{\mathrm{3gFF}} = \{a, b, c, d, e, f\}$$

Arch. \ Train. size	7.3M	7M	6M	5M	4M	3M	2M	1M
8 layers, $d = 1024$	98.8%	98.7%	98.2%	97.5%	96.7%	94.8%	90.8%	78.2%
8 layers, $d = 512$	96.2%	97.4%	98.4%	96.6%	95.3%	93.8%	88.5%	36.7%
6 layers, $d = 1024$	98.6%	98.9%	98.0%	97.9%	96.7%	94.8%	90.3%	58.5%
6 layers, $d = 512$	95.2%	96.6%	96.9%	95.8%	94.4%	94.5%	87.9%	34.8%
4 layers, $d = 1024$	99.1%	98.9%	98.3%	97.9%	96.6%	94.9%	89.9%	39.1%
4 layers, $d = 512$	48.5%	96.0%	94.1%	48.3%	94.6%	81.7%	55.3%	33.9%

ML task: Predict coefficients (and understand learned rules)

2405.06107

Large Language Models

Most impressive growth:
Large language models

Impact for physics:
Numerics
Symbolic

What else?

Large Language Models

Most impressive growth:
Large language models

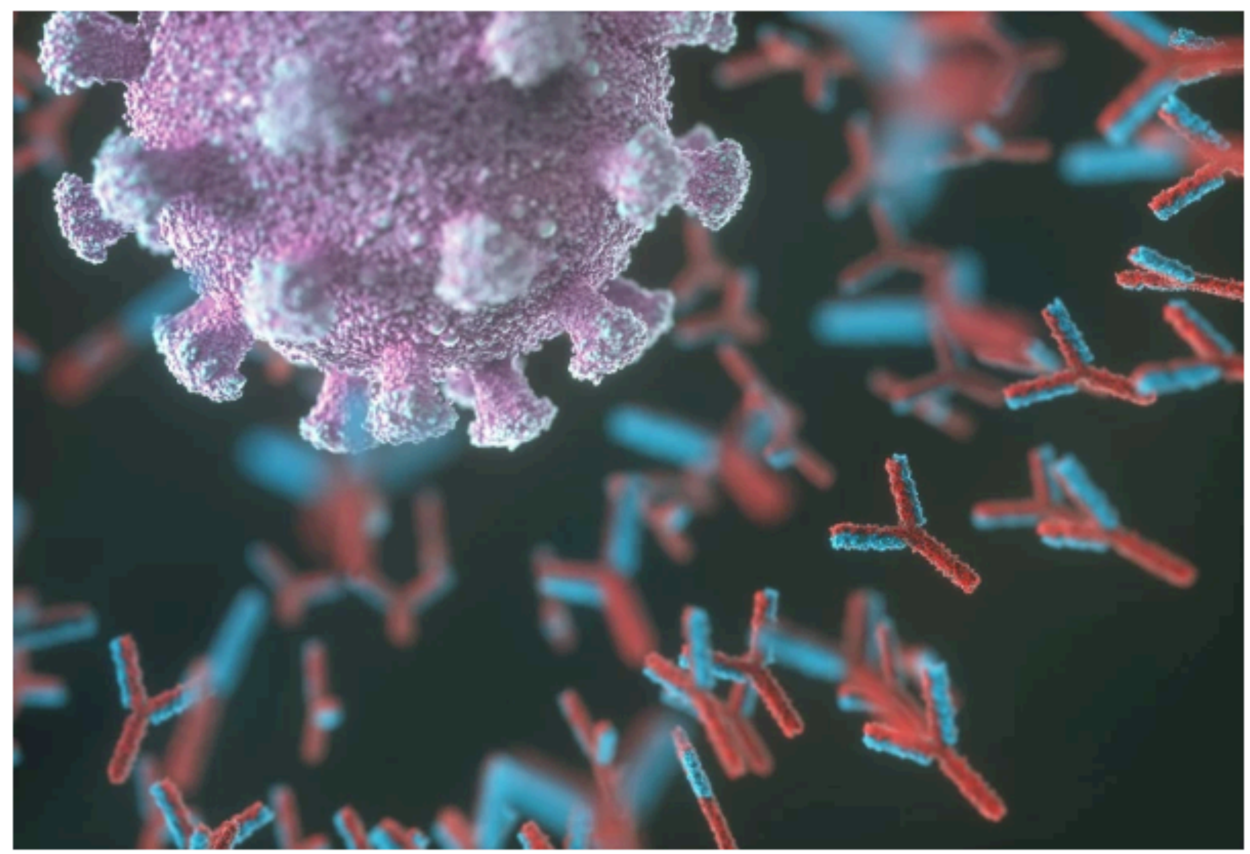
Impact for physics:
Numerics
Symbolic

What else?

Virtual lab powered by 'AI scientists' super-charges biomedical research

Could human–AI collaborations be the future of interdisciplinary studies?

By Helena Kudiabor

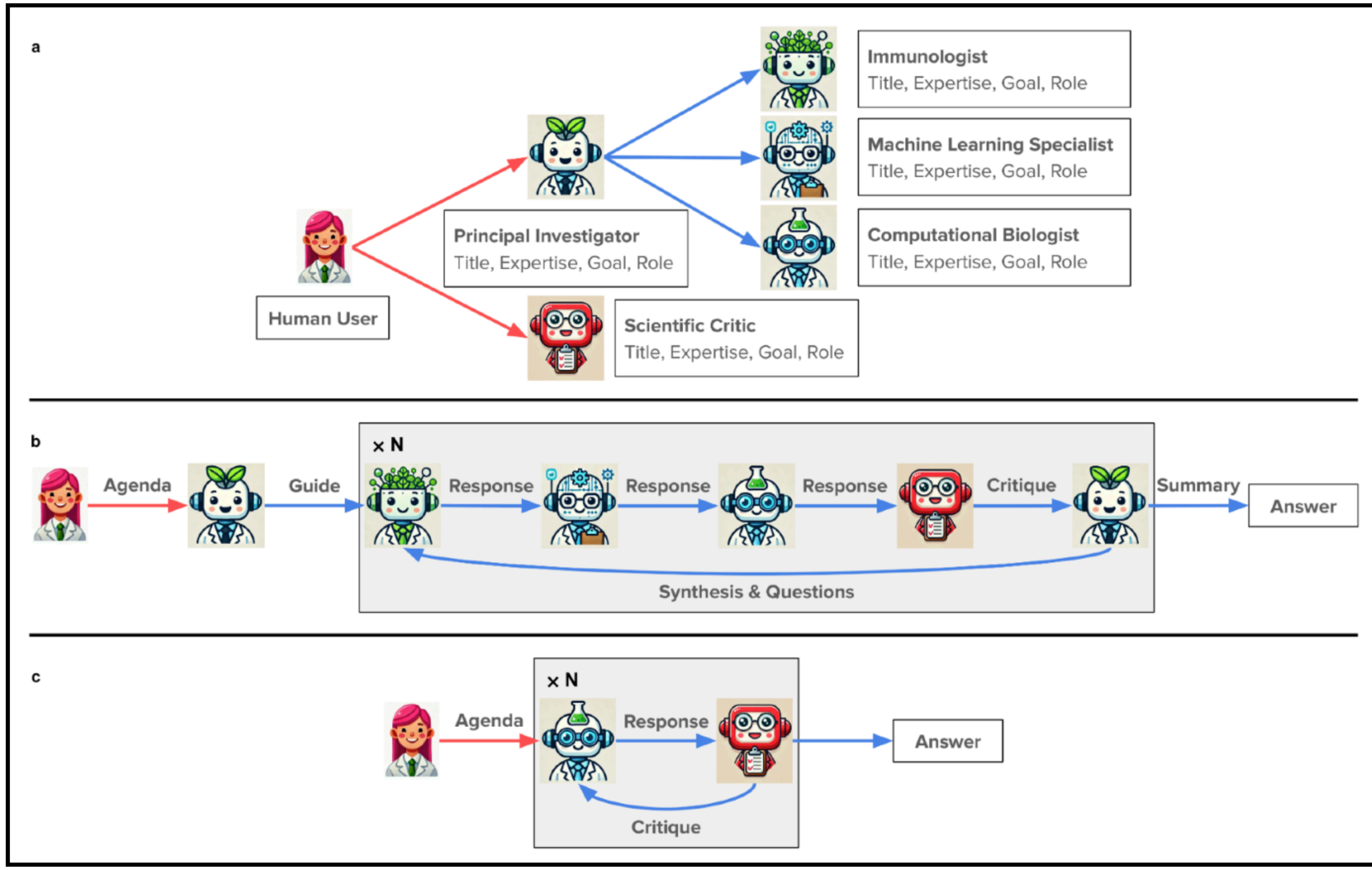


Large Language Models

Most impressive growth:
Large language models

Impact for physics:
Numerics
Symbolic

Include agent-based models as collaborators?



https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1

Closing

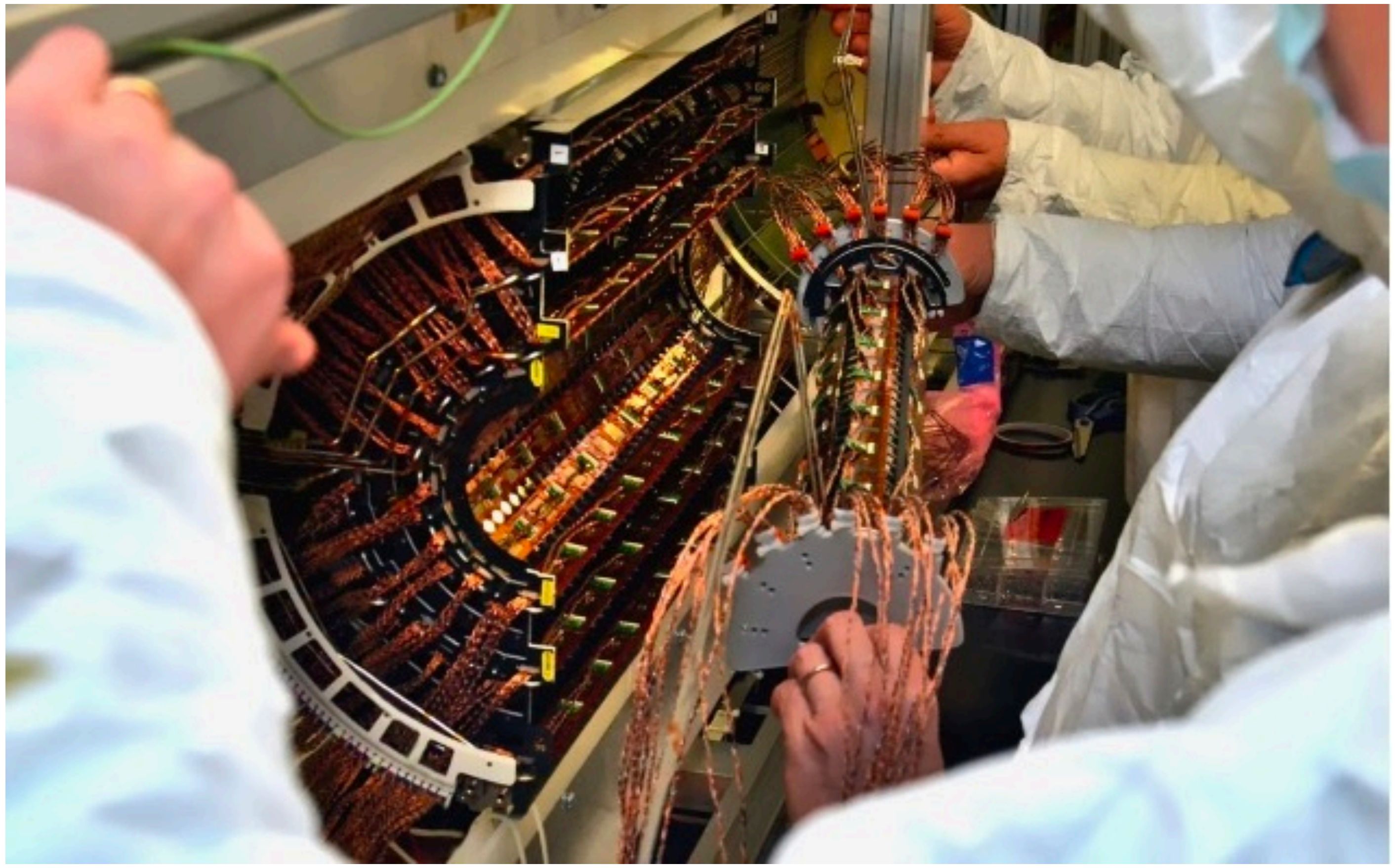
Tools

Frequent statement: Isn't machine learning just a tool?

Tools

Frequent statement: Isn't machine learning just a tool?

Yes, but so is e.g. the CMS pixel detector or a telescope

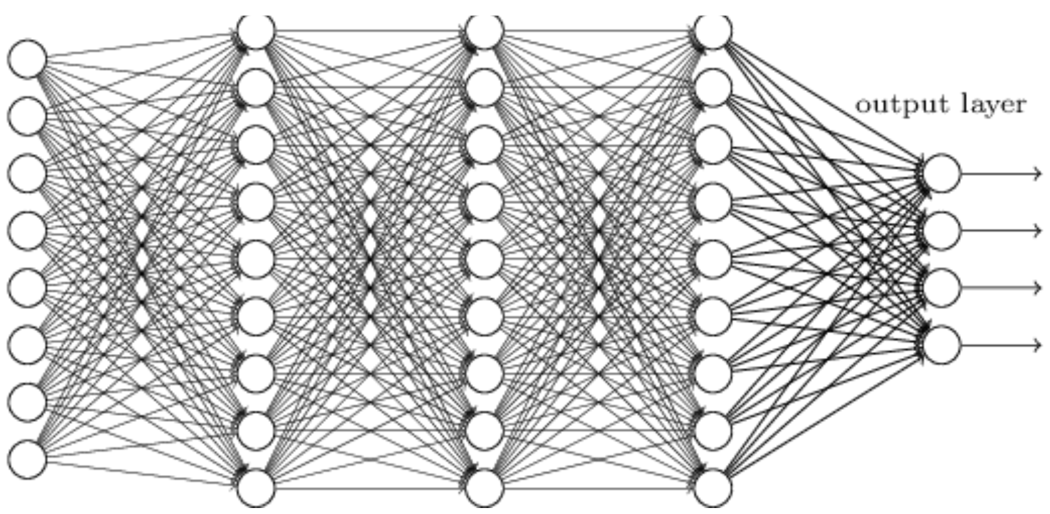


Tools

Frequent statement: Isn't machine learning just a tool?

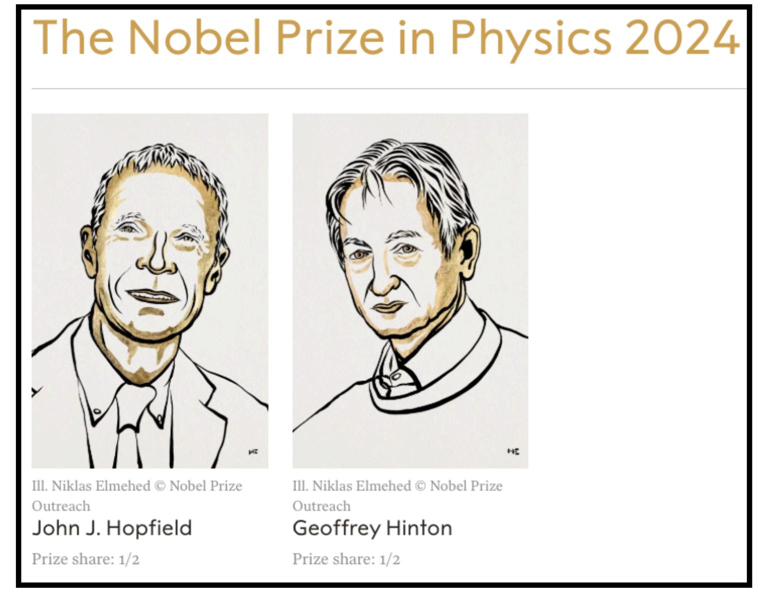
Yes, but so is e.g. the CMS pixel detector or a telescope

Simple architectures are like hammers: Easy to apply (and potentially automate)

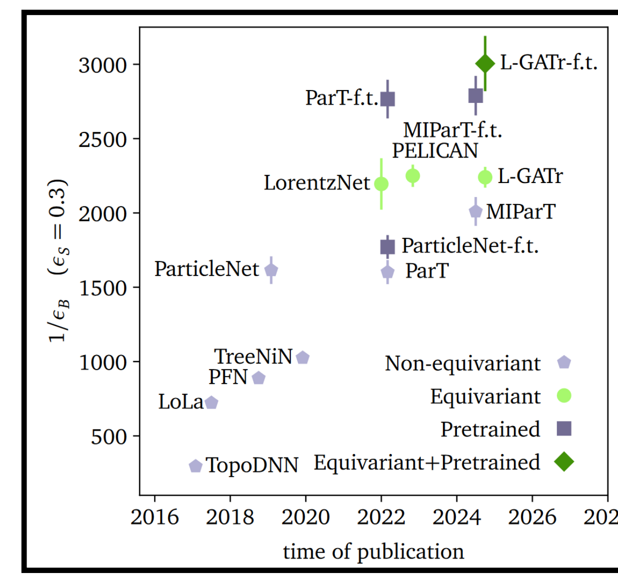


Tools

Frequent statement: Isn't machine learning just a tool?



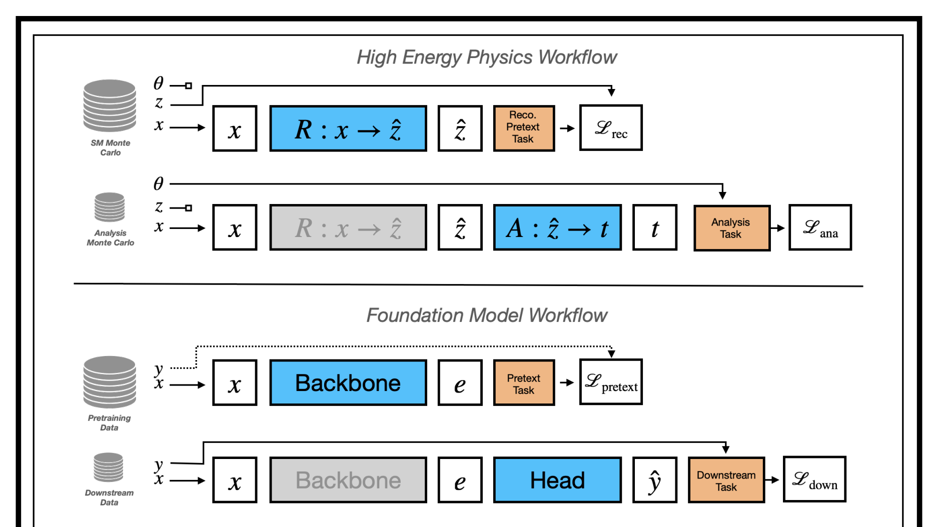
Yes, but so is e.g. the CMS pixel detector or a telescope



Simple architectures are like hammers: Easy to apply (and potentially automate)

Complex learning setups or architectures encoding physics are needed to maximise potential of big data

More like a small experiment



Conclusions

- Best performance reached by compute (e.g. attention mechanism) combined with large-data pre-training & physics insight
- Generative models/likelihood learning as flexible approach for many components of data analysis and far beyond
- Foundation models have the potential for synergies between problems/ experiments/communities
- Large language models offer new directions e.g. for symbolic approaches or collabiration with AI agents

Thank you

‘Small’ AI in Physics workshop: ML4Jets in Hamburg in 2023

Look forward to ACAT 2025
https://indico.cern.ch/e/acat2025

Rebuilding the Scientific Process: AI at the Heart of Theory, Experiment, and Computation in High-Energy and Nuclear Physics