ML Overview in CMS, HEP, and beyond

CMS JetMET Workshop
12 April, 2022

Jean-Roch Vlimant (California Institute of Technology)
jvlimant@caltech.edu @vlimant

Outline

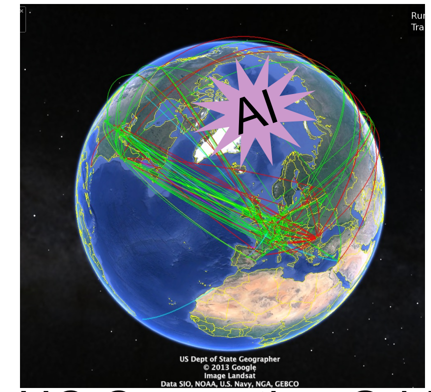
I. Machine Learning beyond classification
II. Geometric Deep Learning
III. Machine Learning in CMS
IV. Prospects of Deep Learning

Machine Learning beyond classification

control, reconstruction, simulation, …

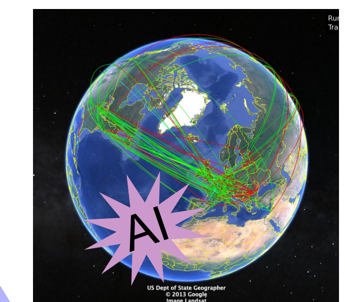
AI in HEP

Role of AI: accelerator control, data acquisition, event triggering, anomaly detection, new physics scouting, event reconstruction, event generation, detector simulation, LHC grid control, analytics, signal extraction, likelihood free inference, background rejection, new physics searches, ...

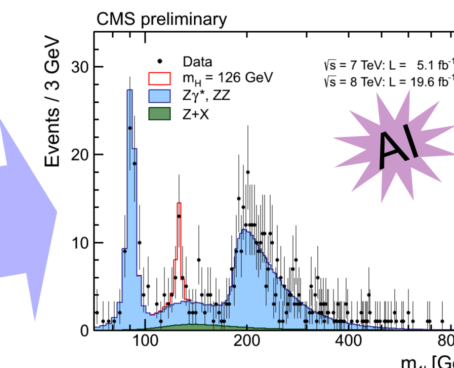


LHC Computing Grid
200k cores pledge to
CMS over ~100 sites

CERN Tier-0/Tier-1
Tape Storage
200PB total

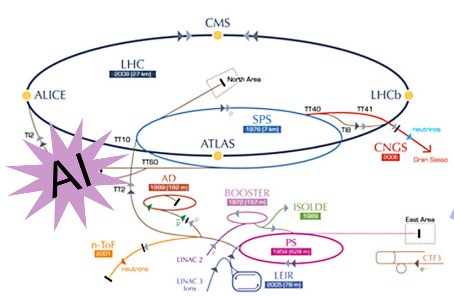


LHC Grid
Remote Access
to 100PB of data

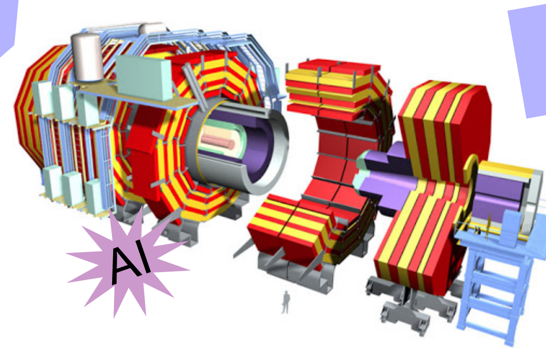


Rare Signal
Measurement
~1 out of 10^6

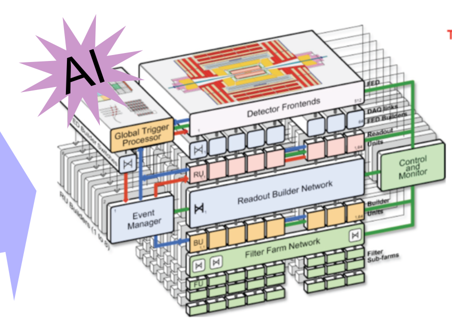
CERN Tier-0
Computing Center
20k cores



Large Hadron Collider
40 MHz of collision



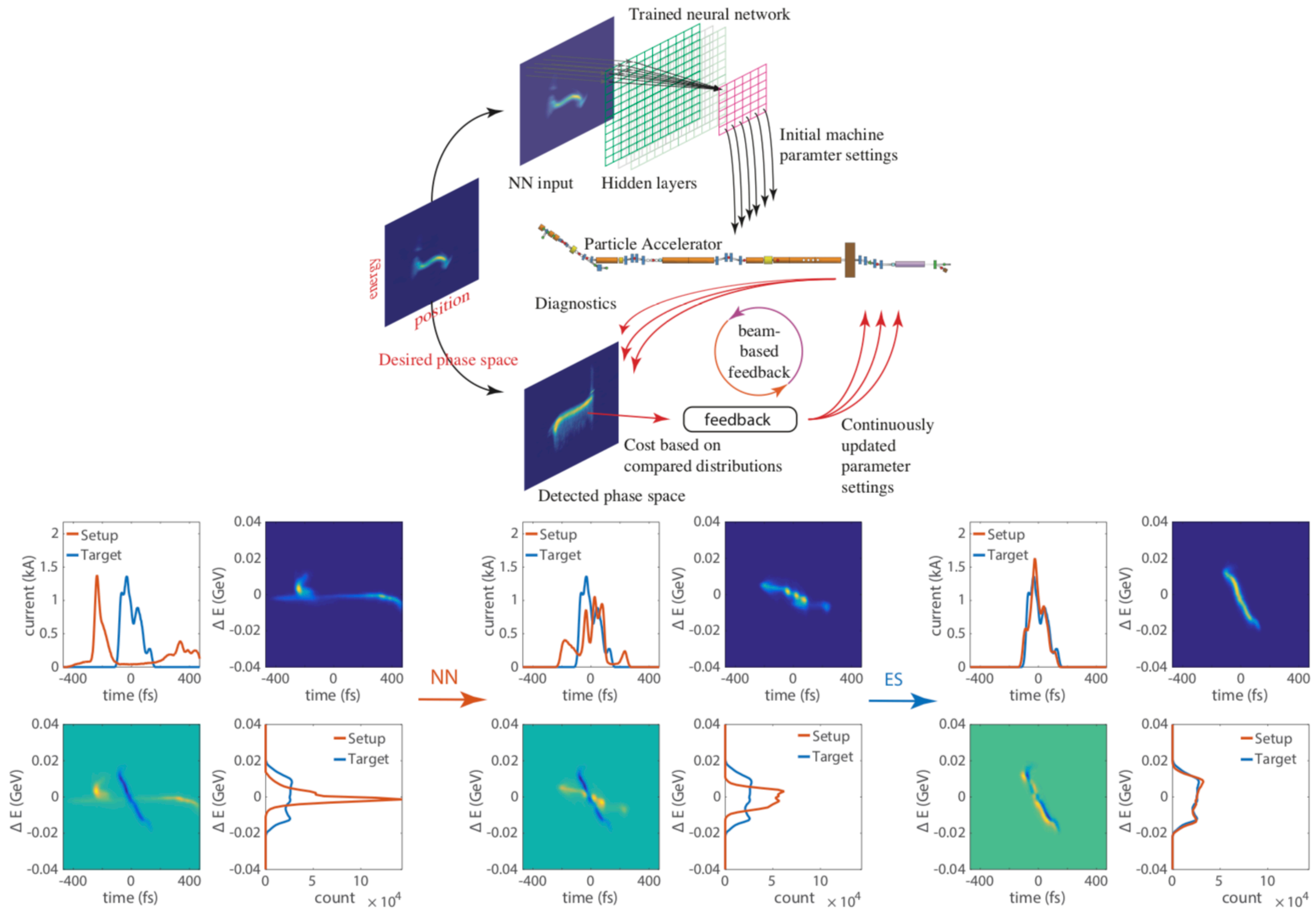
CMS Detector
1PB/s



CMS L1 & High-
Level Triggers
50k cores, 1kHz

Up to date listing of references:
https://github.com/iml-wg/HEPML-LivingReview

Producing the Data



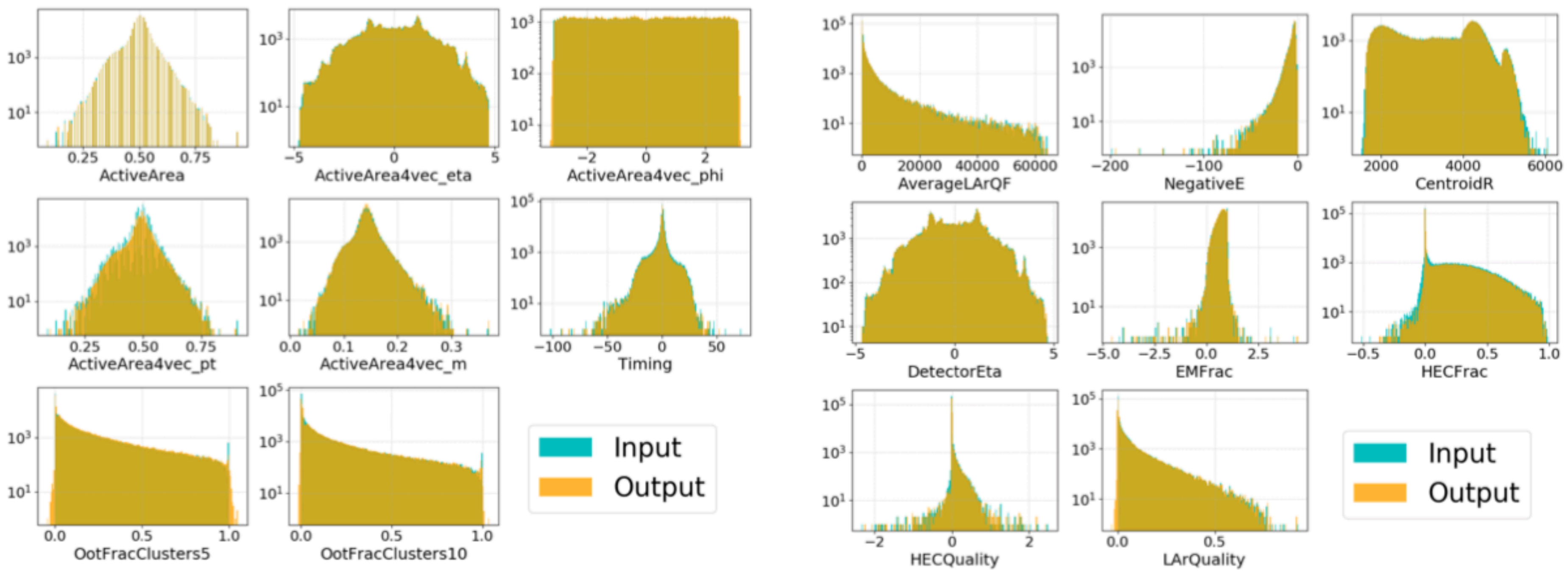
A. Scheinker, C. Emma, A.L. Edelen, S. Gessner [2001.05461]

- Machine learning can be used to tune devices, control beams, perform analysis on accelerator parameters, etc.
- Already successfully deployed on accelerator facilities.
- More promising R&D to increase beam time.
- Potential for detector control ?

Opportunities in Machine Learning for Particle Accelerators [1811.03172]
Machine learning for design optimization of storage ring nonlinear dynamics [1910.14220]
Advanced Control Methods for Particle Accelerators (ACM4PA) 2019 Workshop Report [2001.05461]
Machine learning for beam dynamics studies at the CERN Large Hadron Collider [2009.08109]
…

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Compressing Data

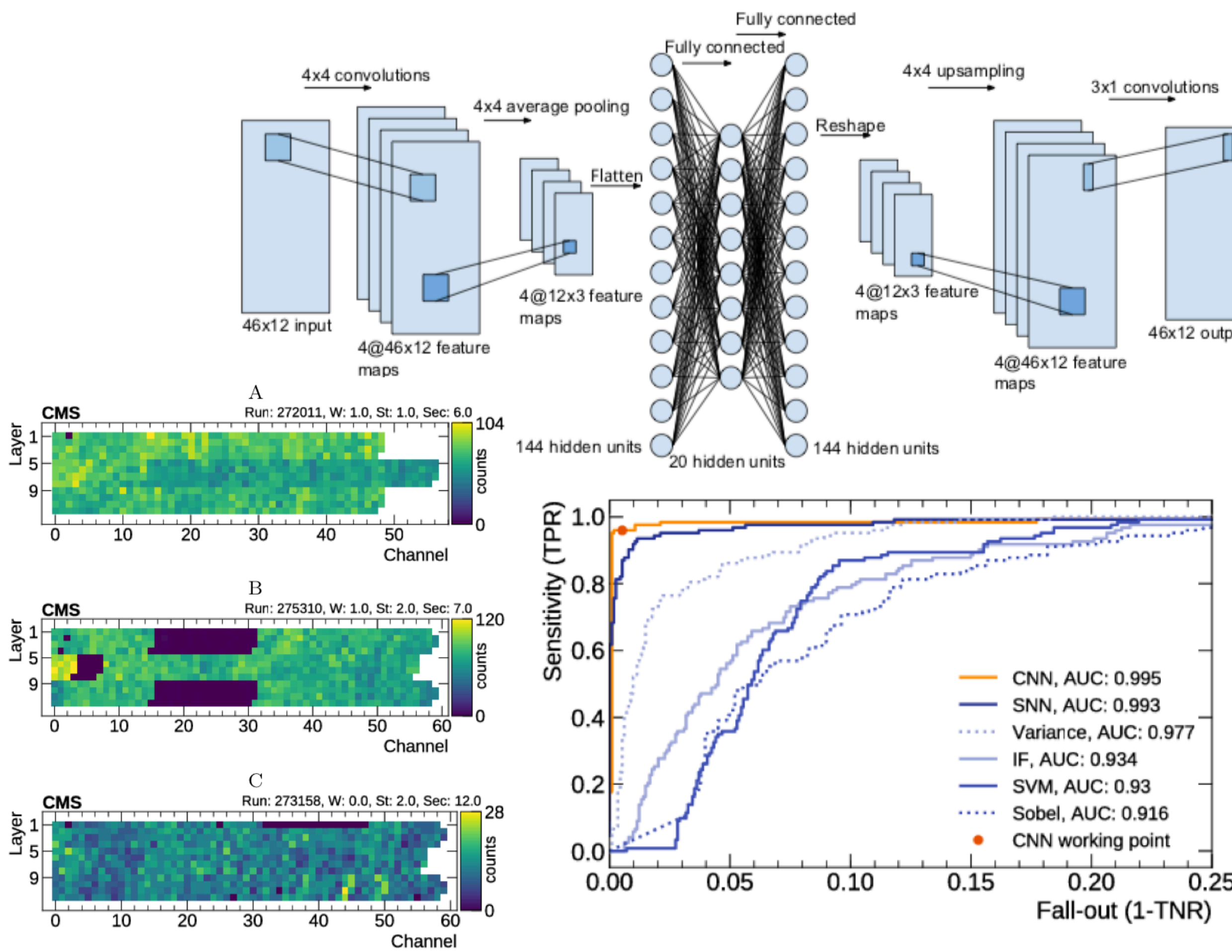


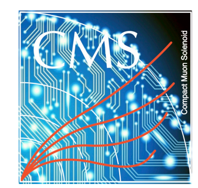
Deep Auto-Encoders for compression in HEP
http://lup.lub.lu.se/student-papers/record/9004751

- Rich literature on data compression of image with neural network.
- Make use of abstract semantic space for image compression.
- Image compression can suffer some loss of resolution.
- Saving on disk/tape cost. Potential in scouting strategies.
- R&D needed to reach the necessary level of fidelity.

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Cleaning Data





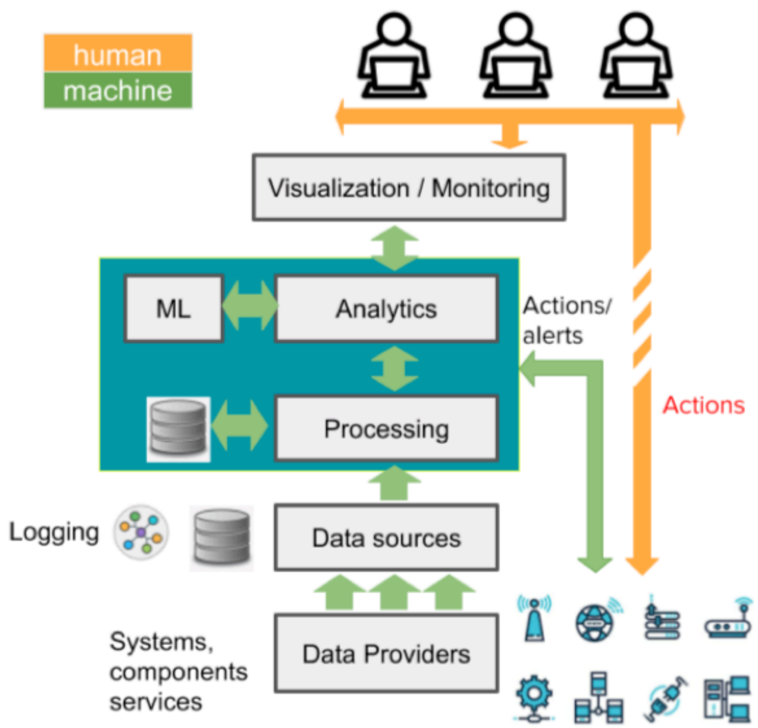
- Data quality is a person power intensive task, and crucial for swift delivery of Physics
- Machine learning can help with automation.
- Learning from operators, reducing workload.
- Continued R&D and experiment adoption.

A.A. Pol, G. Cerminara, C. Germain, M. Pierini, A. Seth
[doi:10.1007/s41781-018-0020-1]

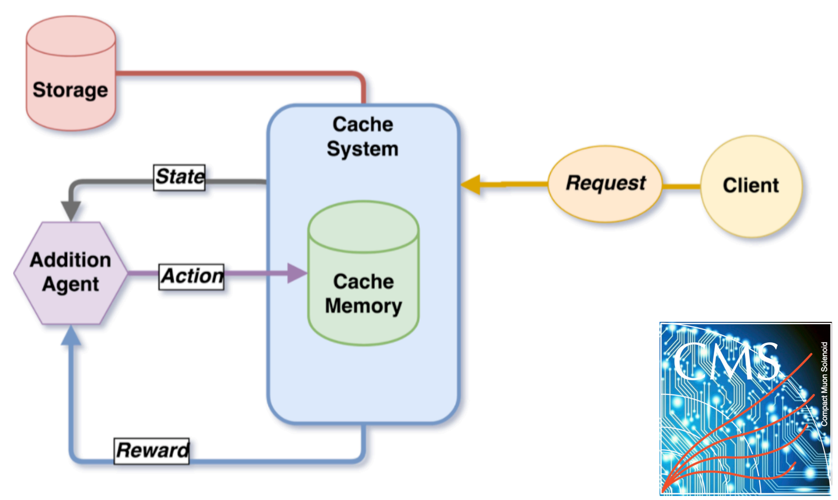
Towards automation of data quality system for CERN CMS experiment [doi:10.1088/1742-6596/898/9/092041]
LHCb data quality monitoring [doi:10.1088/1742-6596/898/9/092027]
Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider [1808.00911]
Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment [doi:10.1051/epjconf/201921406008]
…

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Managing Data



Operational Intelligence
[cds:2709338]



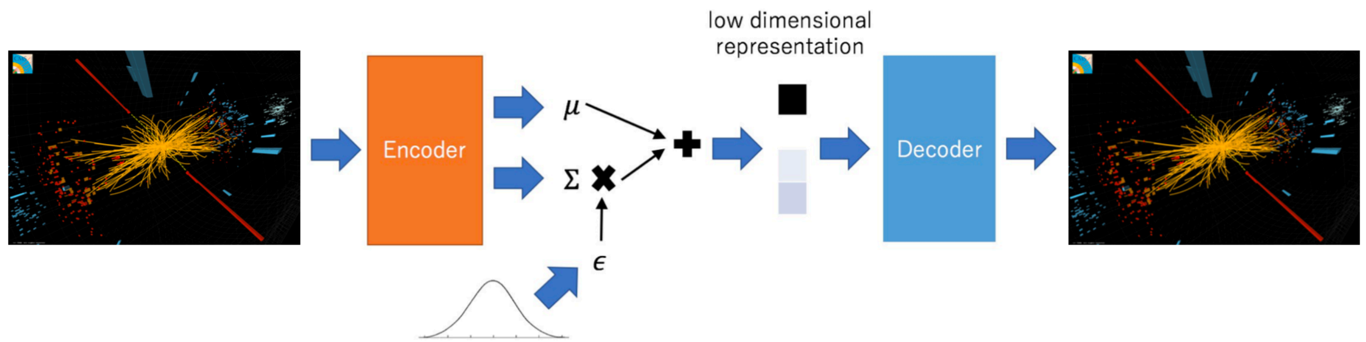
Cache Type	Throughput	Cost	Read on hit ratio	Band sat.	CPU Eff.
SCDL	**79.43%**	**50.68%**	21.22%	58.94%	58.75%
LFU	65.01%	104.73%	**33.29%**	**51.00%**	**60.92%**
Size Big	49.02%	111.73%	28.55%	54.40%	60.41%
LRU	47.15%	112.84%	27.64%	54.93%	59.90%
Size Small	46.71%	113.01%	27.39%	55.01%	59.73%

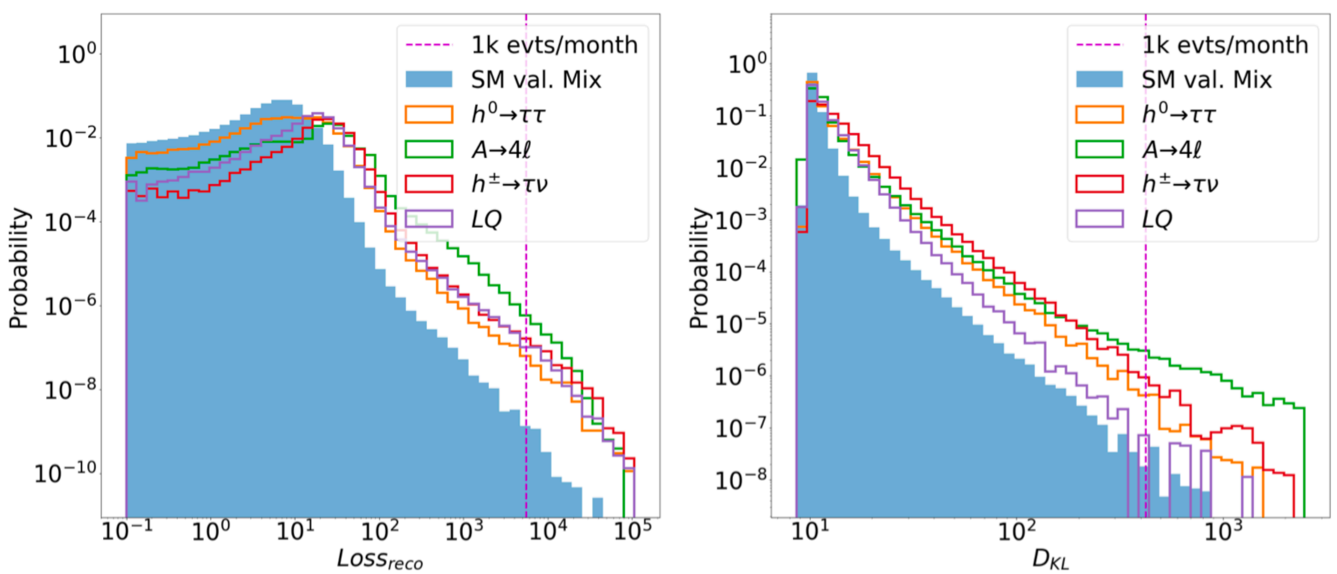
Caching suggestions using Reinforcement Learning
LOD 2020, in proceedings

- The LHC-grid is key to success of the LHC experiments.
- Complex ecosystem with dedicated operation teams.
- Person power demanding, and inefficient in some corner of the phase space.
- Potential for AI-aided operation.
- Lots of modeling and control challenges.
- R&D to increase operation efficiency.

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Detecting New Data





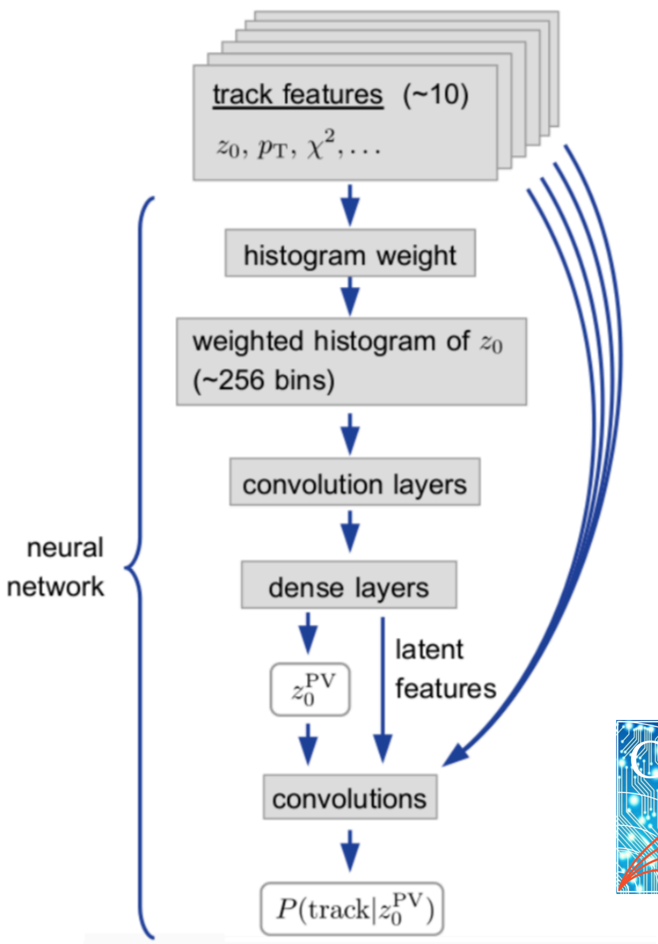
Use of variational auto-encoders directly on data to marginalize outlier events, for anomalous event hotline operation.
[doi:0.1007/JHEP05(2019)036]

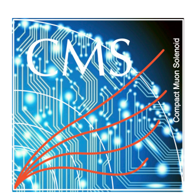
More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

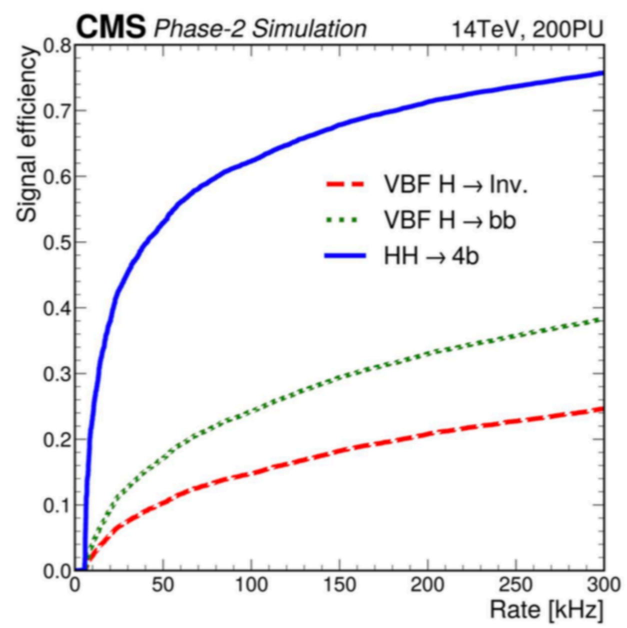
- Machine learning since long deployed in the trigger for selected signatures.
- Further potential for background trigger rate reduction.
- Emerging opportunity for triggering on unknown signatures : “a la Hotline”.
- More promising R&D and experiment adoption.

Data Triggering and Scouting

Vertex reconstruction at L1





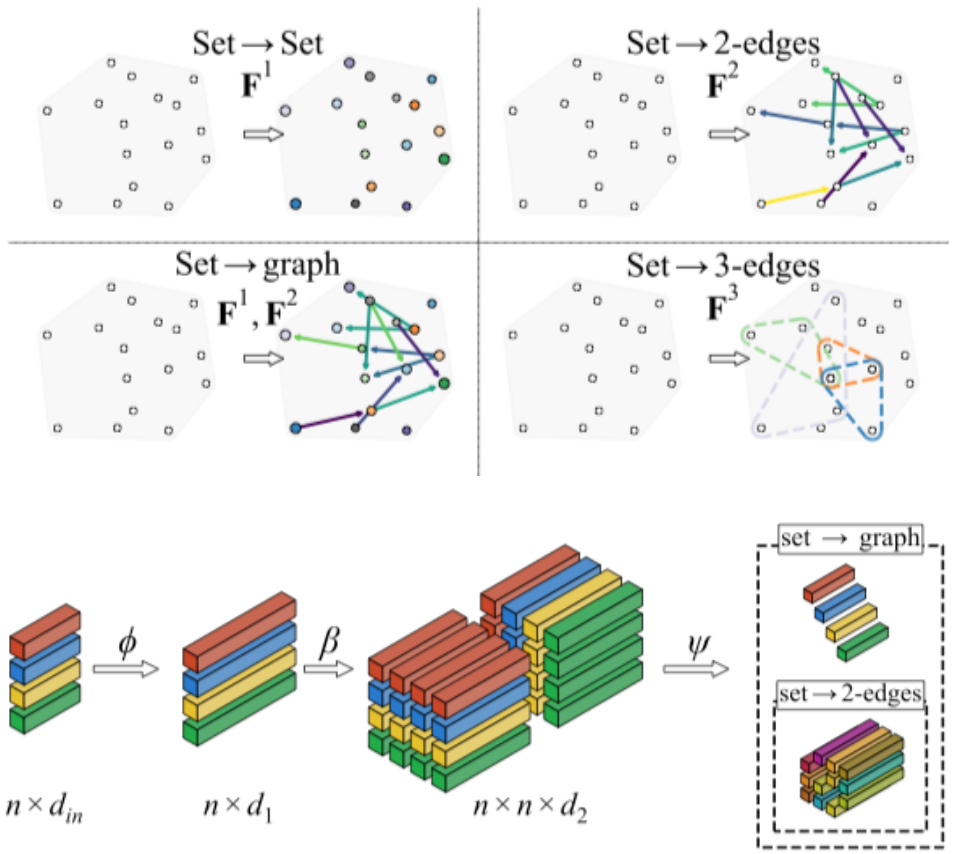


Anomaly detection at L1

Phase-2 upgrade of the CMS L1-Trigger
[cds:2714892]

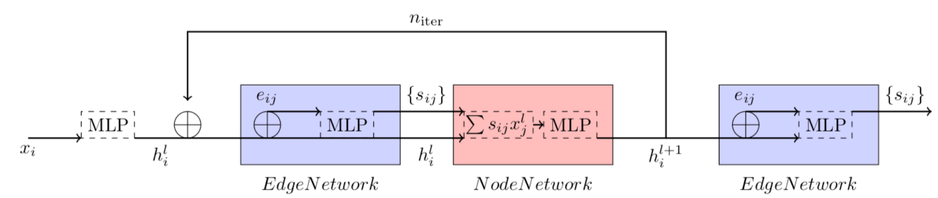
- Trigger benefit from fast reconstruction algorithms
- L1 needs FPGA implementation. hls4ml-enabled algorithms.
- Quality of selection increases with refinement of object reconstruction
- Having the best reconstruction is particularly important in scouting
- Balance between speed and accuracy

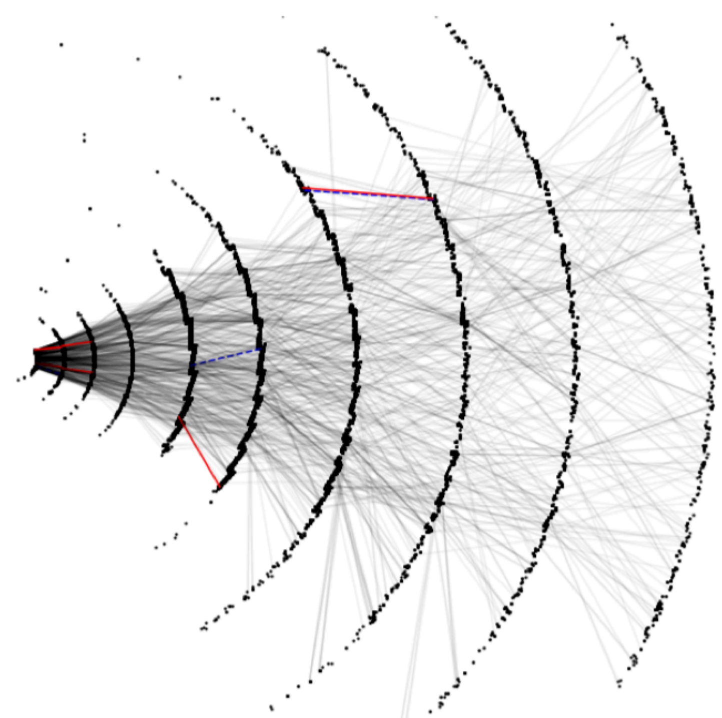
Reconstructing Data



Learning graphs from sets, applied to vertexing
[2002.08772]

GNN applied to charged particle tracking
[2007.00149]

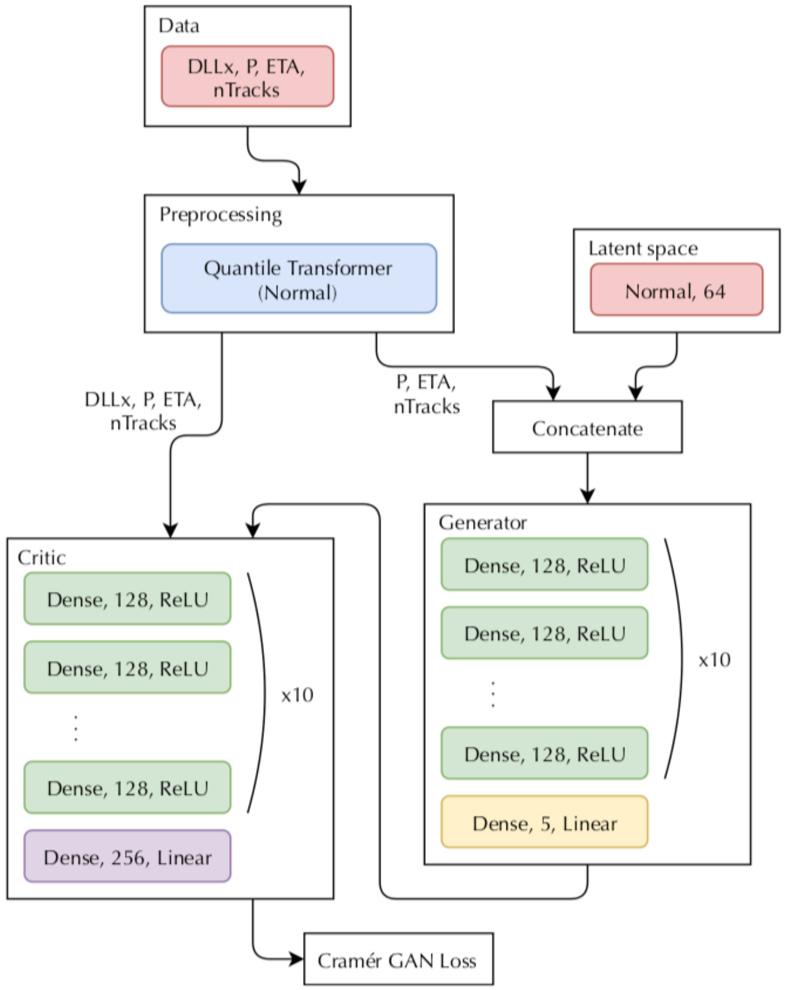


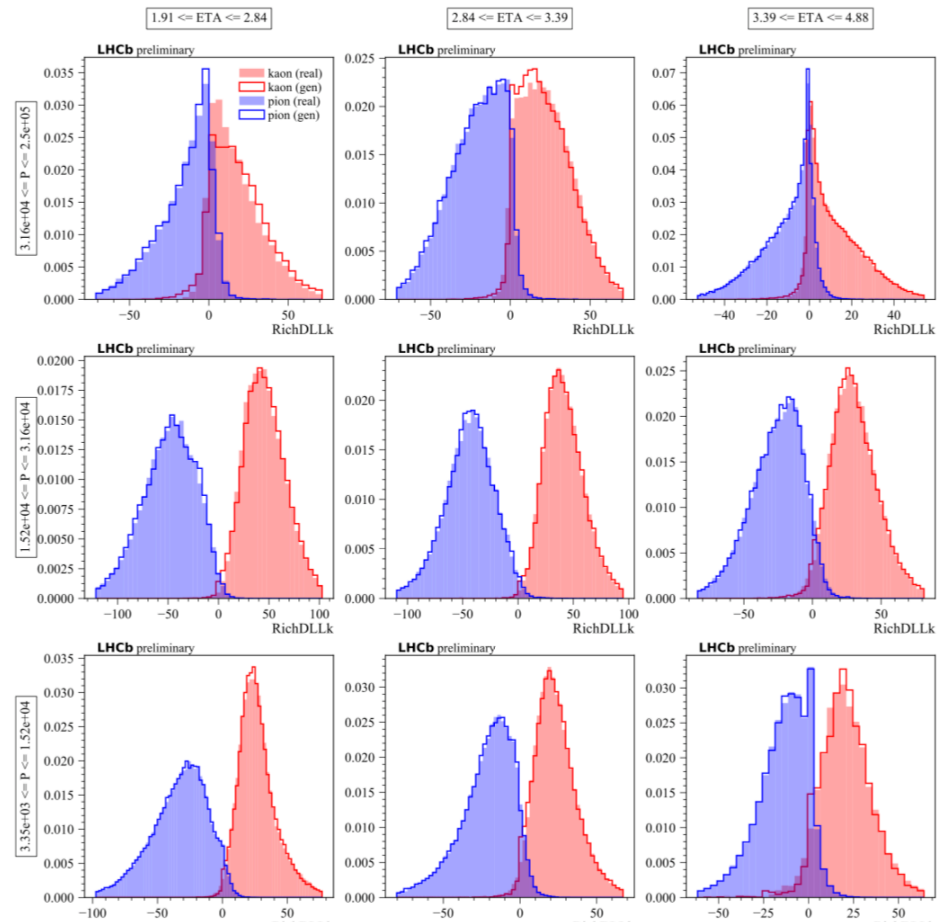


- Event reconstruction is pattern recognition to a large extend. Advanced machine learning techniques can help.
- Learn from the simulation, and/ or data.
- Learn from existing "slow reconstruction" or simulation ground truth.
- Automatically adapt algorithm to new detector design.
- Image base methods evolving towards graph-based methods.
- Accelerating R&D to exploit full potential.

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Simulating Data



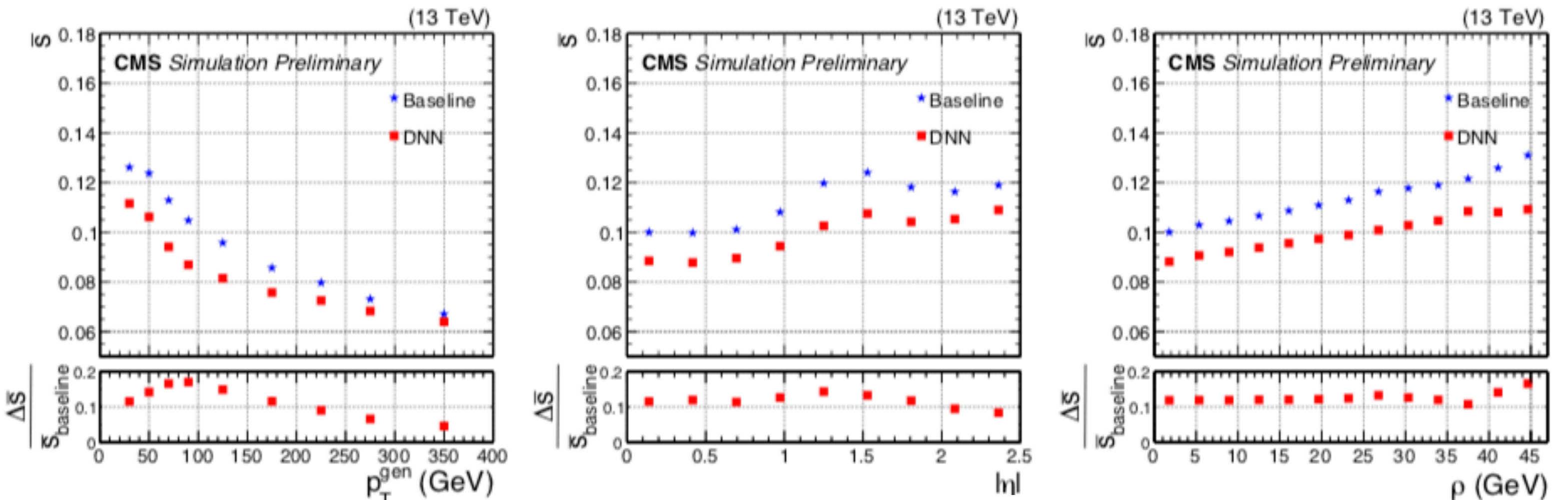


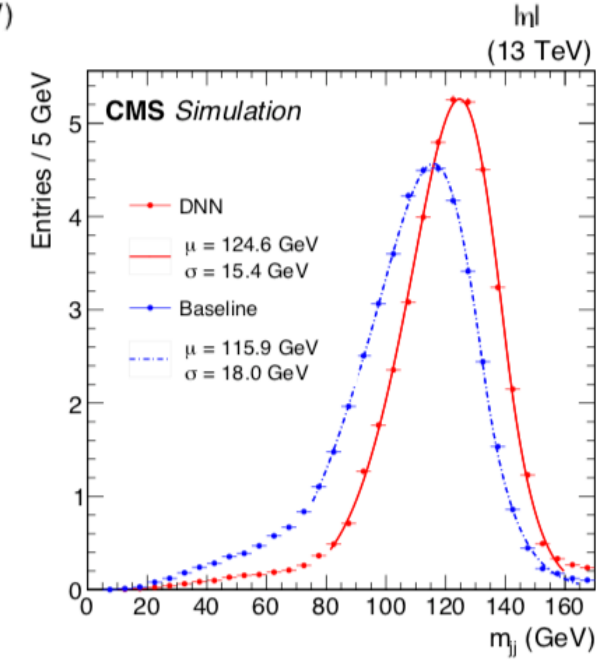
Generative Adversarial Networks for LHCb Fast Simulation [2003.09762]

- Fully detailed simulation is computing intensive.
- Fast and approximate simulators already in operation.
- Applicable at many levels : sampling, generator, detector model, analysis variable, etc
- Generative models can provide multiple 1000x speed-up.
- Careful study of statistical power of learned models over training samples.
- Many R&D, experiment adoption starting.

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Calibrating Data



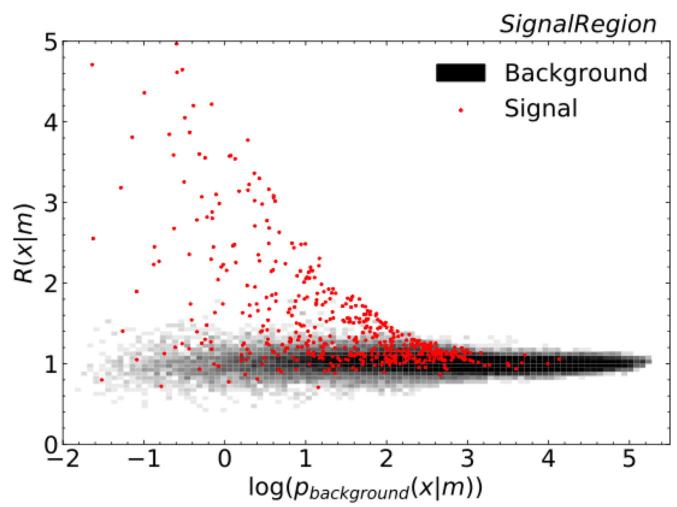


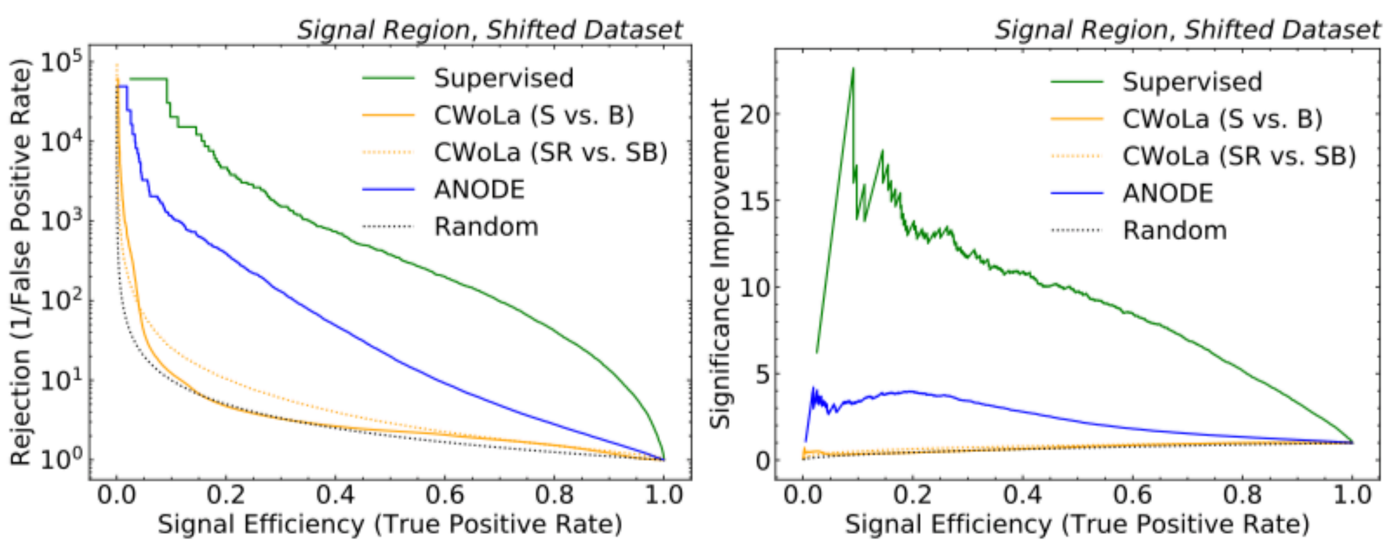
A deep neural network for simultaneous estimation of b jet energy and resolution [1912.06046]

- Energy regression is the most obvious use case.
- Learning calibrating models from simulation and data.
- Parametrization of scale factors using neural networks.
- Reducing data/simulation dependency using domain adaptation.
- Continued R&D

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

Analyzing Data



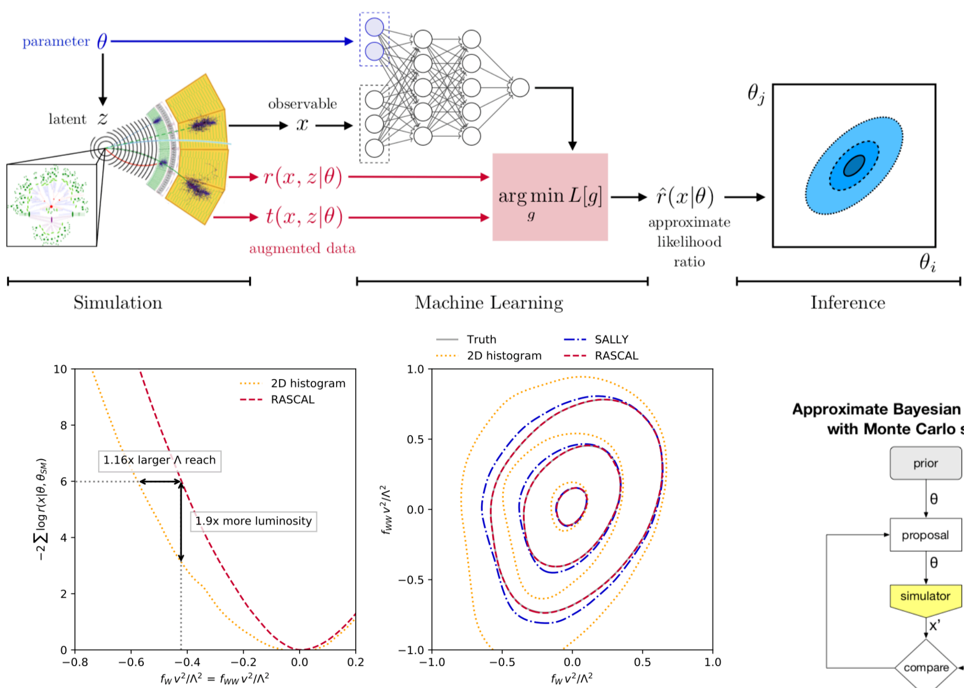


Use of masked autoregressive density estimator with normalizing flow as model-agnostic signal enhancement mechanism.
[doi:10.1103/PhysRevD.101.075042]

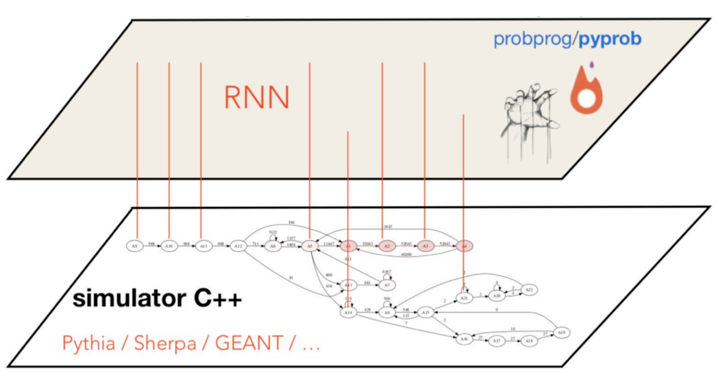
- Machine learning has long infiltrated analysis for signal/bkg classification.
- Increasing number of analysis with more complex DNN.
- Application to signal categorization, bkg modeling, kinematics reconstruction, decay product assignment, object identification, …
- Breadth of new model agnostic methods for NP searches.
- Continued R&D and experiment adoption initiated.

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

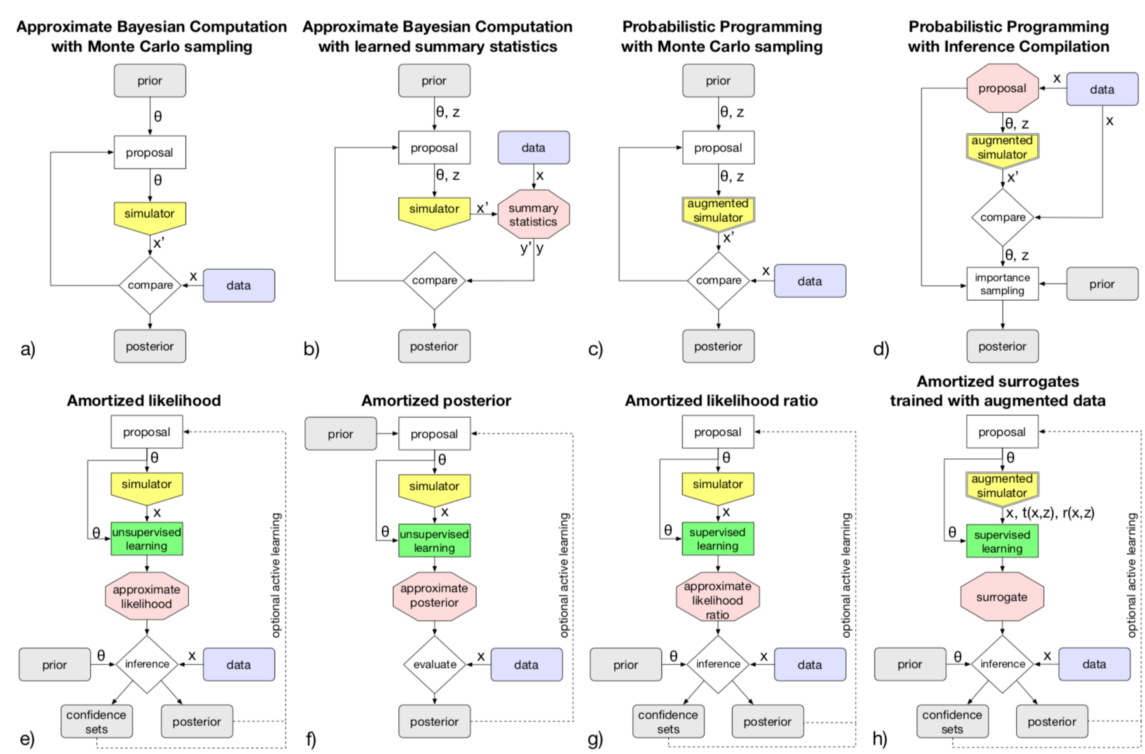
Theory Behind the Data



Constraining EFT with ML
[1805.00013]



https://github.com/probprog/pyprob



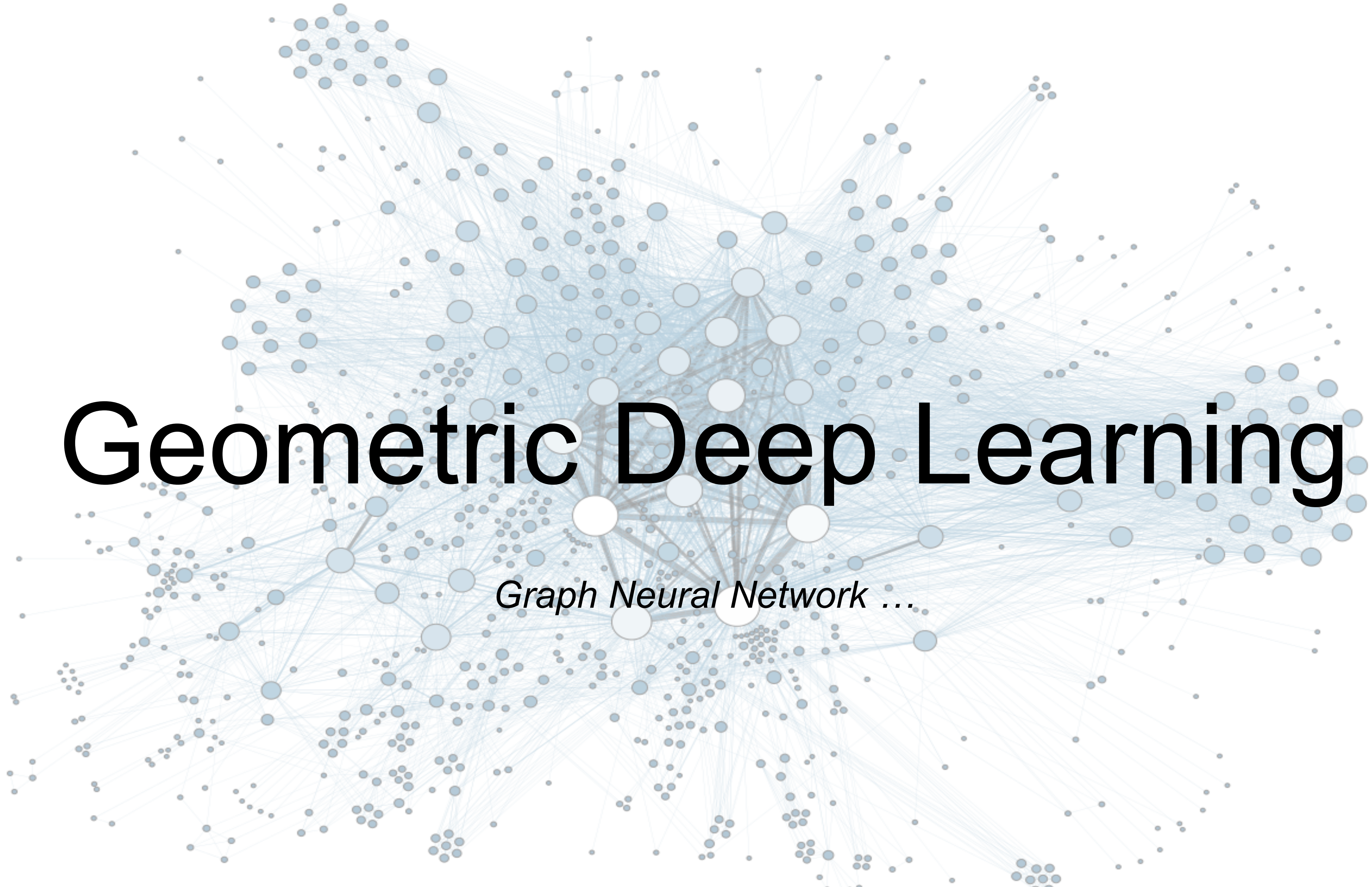
The frontiers of simulation-based inference
[1911.01429]

More of the relevant works at:
https://iml-wg.github.io/HEPML-LivingReview/

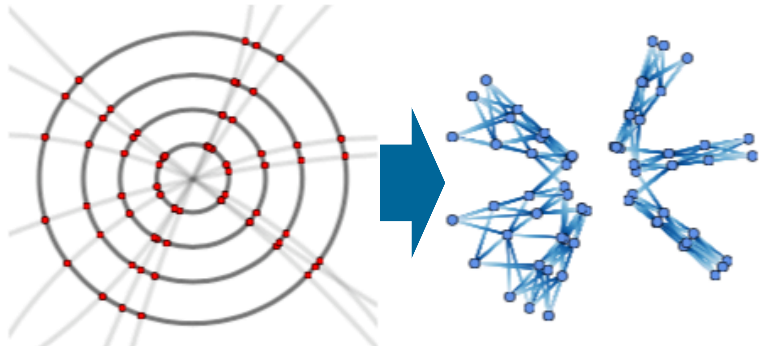
- Hypothesis testing is the core of HEP analysis.
- Intractable likelihood hinders solving the inverse problem.
- Going beyond the standard approach using machine learning and additional information from the simulator.
- More precise evaluation of the priors on theory's parameters.
- May involve probabilistic programming instrumentation of HEP simulator.
- R&D to bring this in the experiment.

Geometric Deep Learning

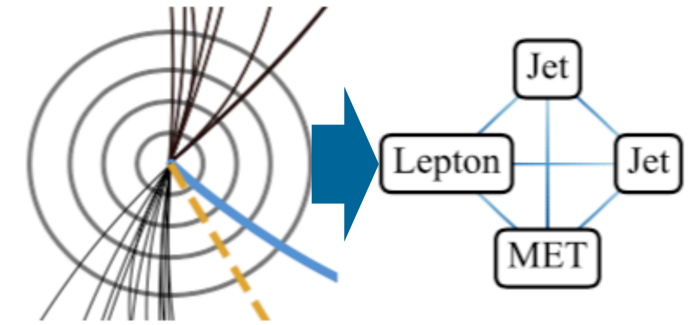
Graph Neural Network …



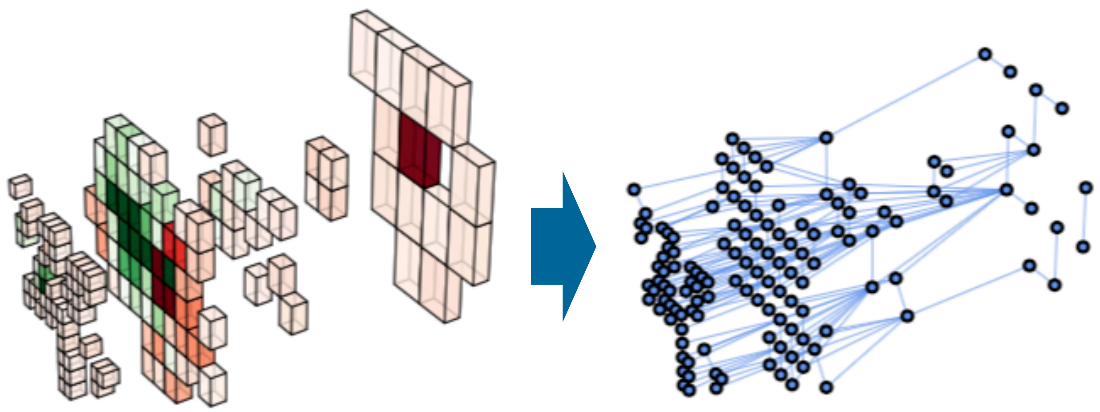
Graph Representation



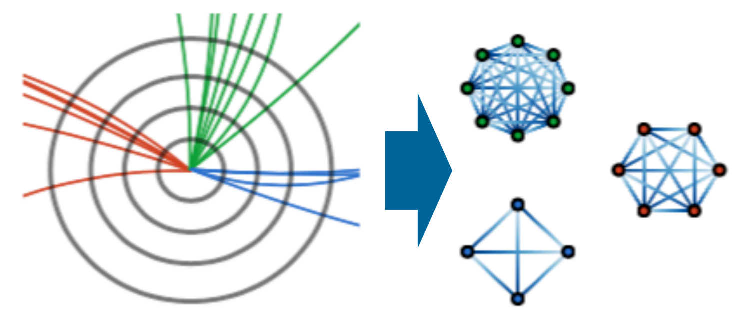
Hits in tracking detector



Objects in an event



Hits in calorimeter detector



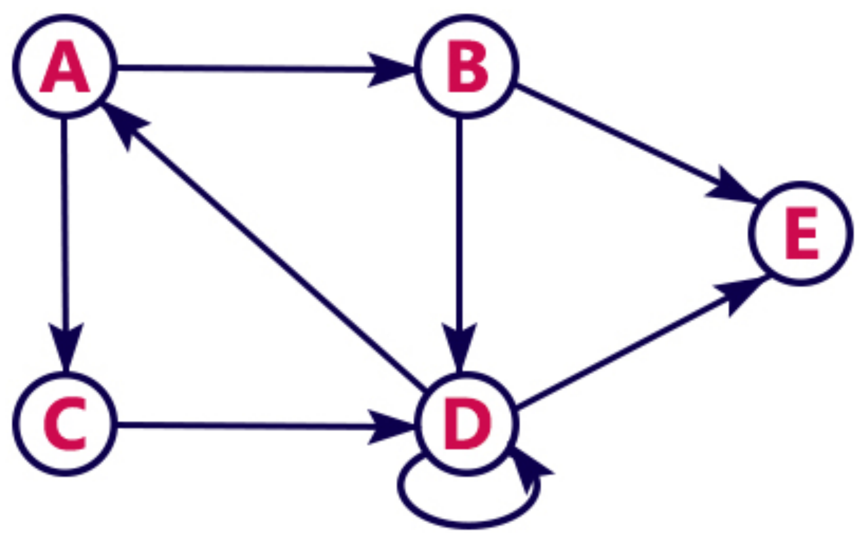
Object sub-structure in an event

Graph Neural Networks for Particle Physics reconstruction
[2007.13681], [2012.01249]

Heterogenous data fits well in graph/set representation.

Multiple CMS ML Forum presentations on GNN applications [Sept 30, 2020] , [Oct 20, 2021], [Nov 3, 2021] and reconstruction with ML [Feb 21, 2021], [Mar 10, 2021].

Forewords on Graph



$$
\begin{array}{c|ccccc}
 & A & B & C & D & E \\
\hline
A & 0 & 1 & 1 & 0 & 0 \\
B & 0 & 0 & 0 & 1 & 1 \\
C & 0 & 0 & 0 & 1 & 0 \\
D & 1 & 0 & 0 & 1 & 1 \\
E & 0 & 0 & 0 & 0 & 0
\end{array}
$$

http://btechsmartclass.com/data_structures/graph-representations.html

A graph is composed of

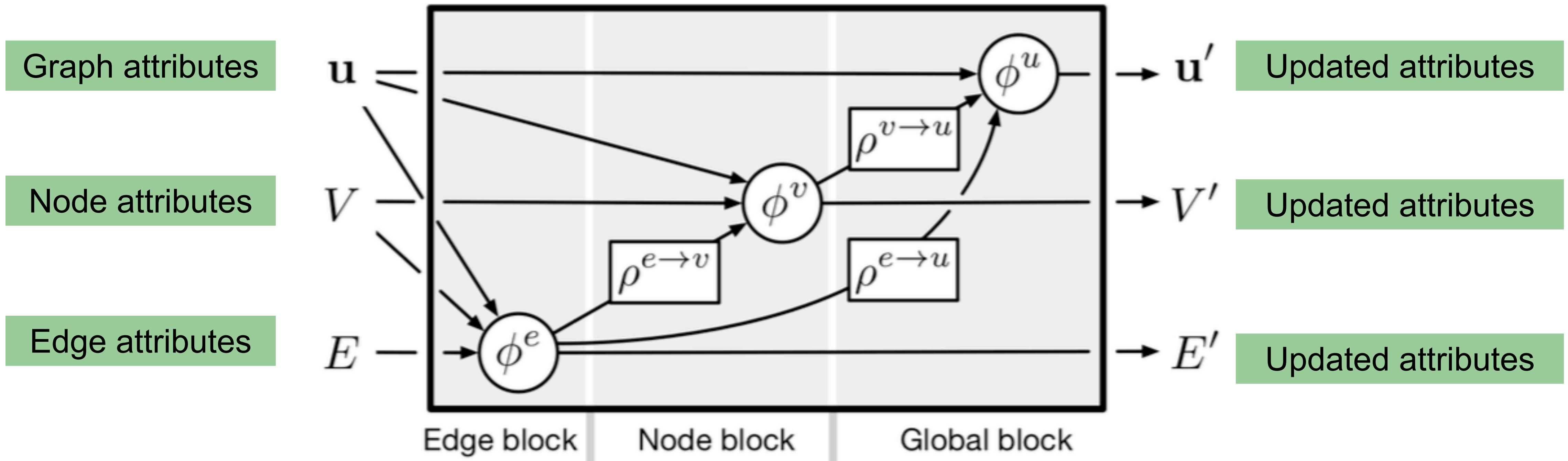
- **Nodes** that can be represented as a vector.
- **Edges** that can be represented with the adjacency matrix.

→ Flowing of information using matrix operations.

→ With machine learning on graphs, edges and nodes might acquire internal representations.

Graph Neural Networks Formalism

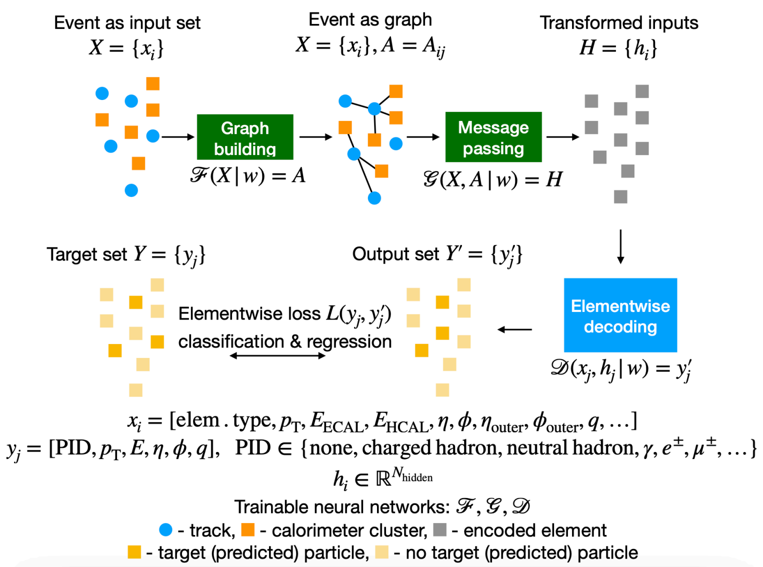
https://arxiv.org/abs/1806.01261

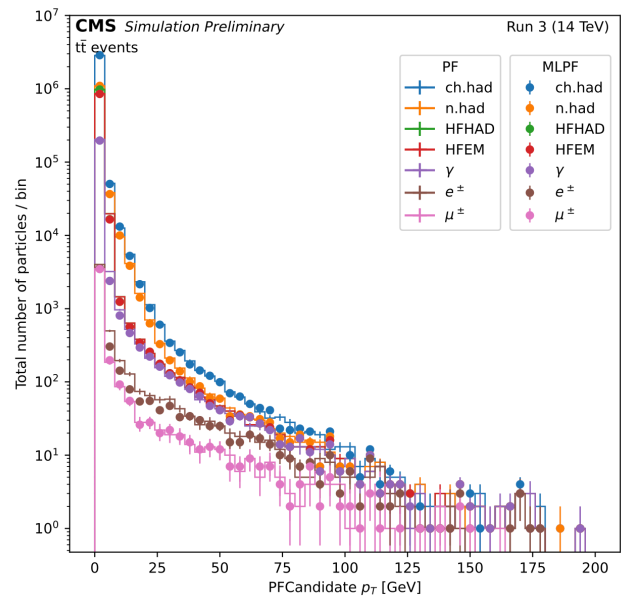


Lots of possibilities to operate on a graph.
Most available architectures can be expressed with Φ and ρ.

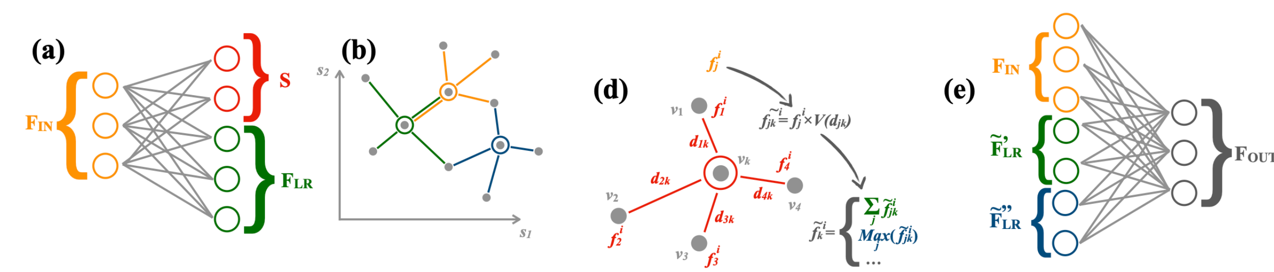
Readily software:
https://github.com/deepmind/graph_nets
https://github.com/rusty1s/pytorch_geometric
…

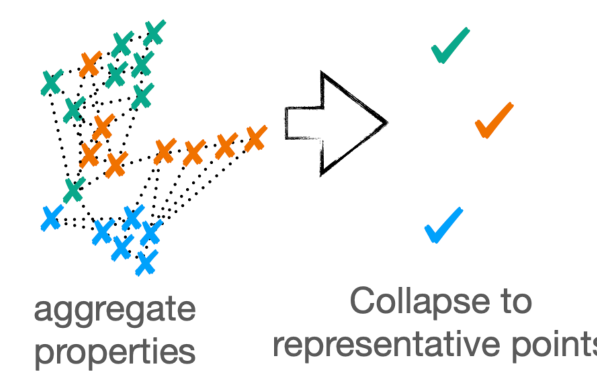
Geometric Deep Learning (I)

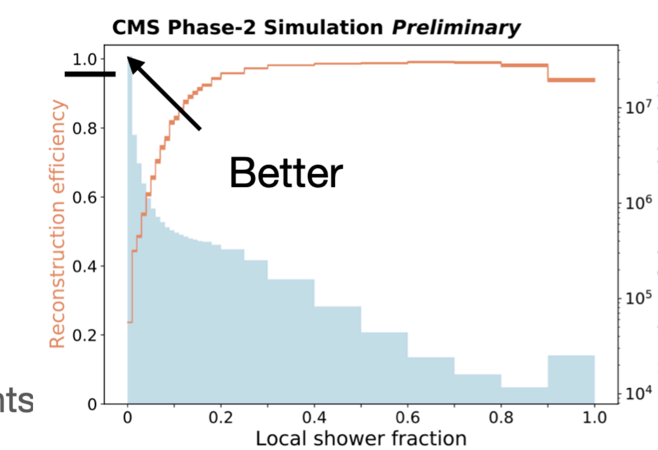




GNN for **particle-flow reconstruction**, https://indico.cern.ch/event/1094349/#8-dp-note-ml4pf. [2101.08578]

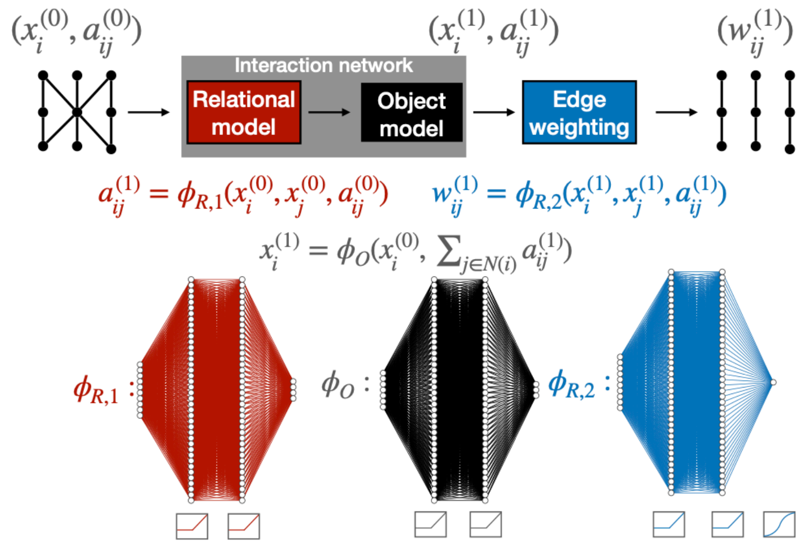


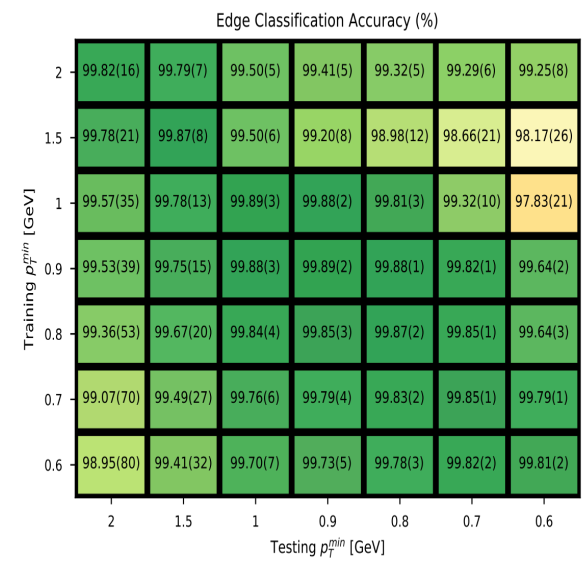




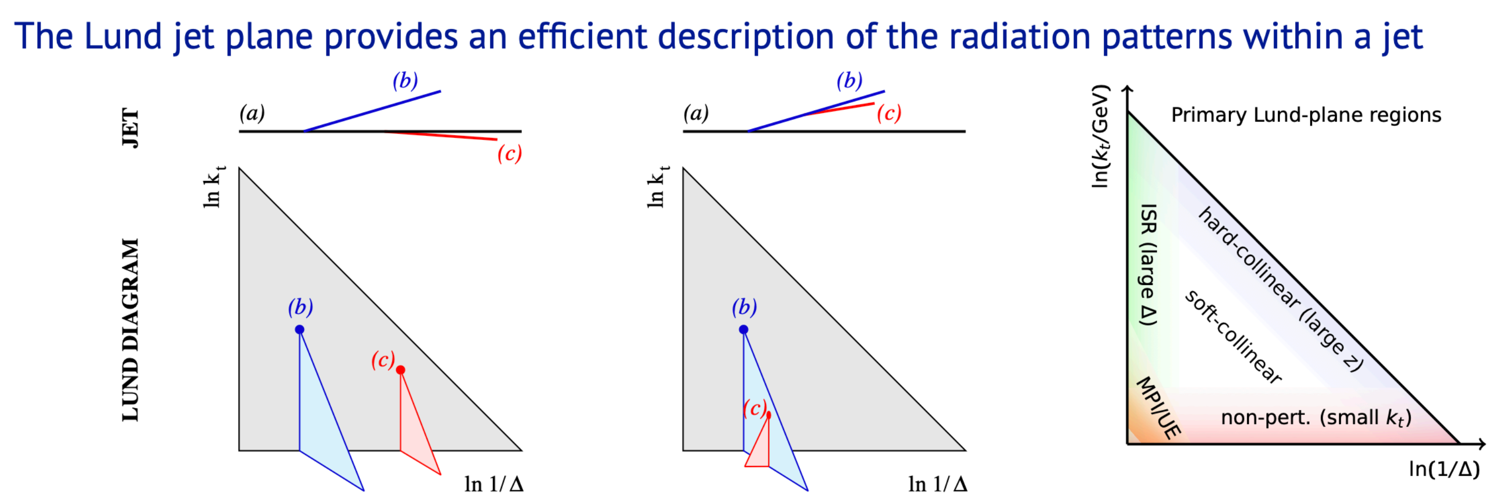
HGCAL Phase-2 **calorimeter reconstruction** using GarNet and Object condensation, [2106.01832]. [1902.07987] [2002.03605]

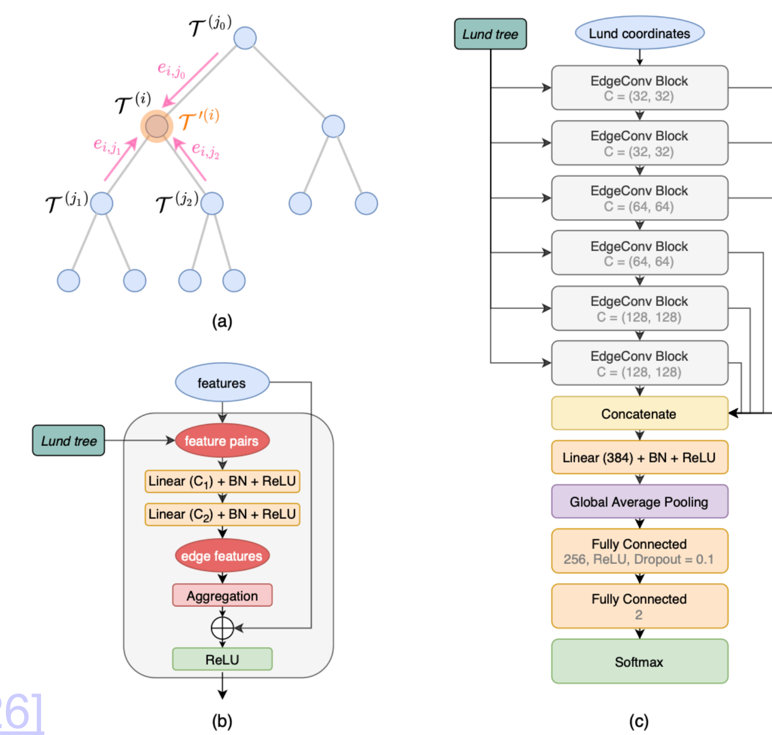
Geometric Deep Learning (II)





Charged particle **track reconstruction** with GNN, [2103.06995], [2103.16701]

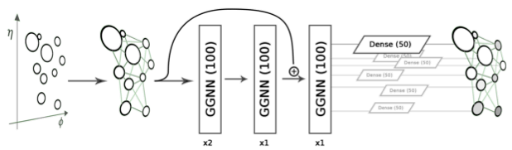




Jet tagging in the Lund plane, [2012.08526]

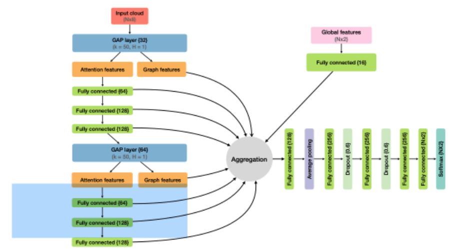
Geometric Deep Learning (III)

Gated graph neural network
(radius graph, ΔR = 0.3)



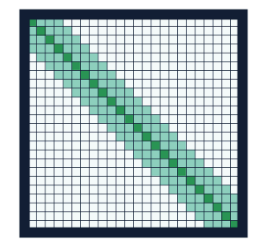
J. Arjona Martínez, O. Cerri, M. Pierini, M. Spiropulu and J. R. Vlimant
[[Eur.Phys.J.Plus 134 (2019) 7, 333]]

ABCNet
(kNN graph, k = 50)



V. Mikuni and F. Canelli
[[Eur. Phys. J. Plus 135, 463 (2020)]]

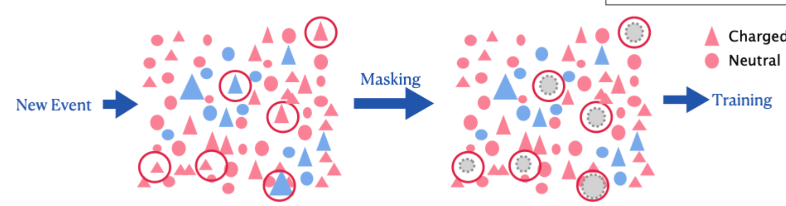
PUMA
(Sparse transformers with sliding window attention)



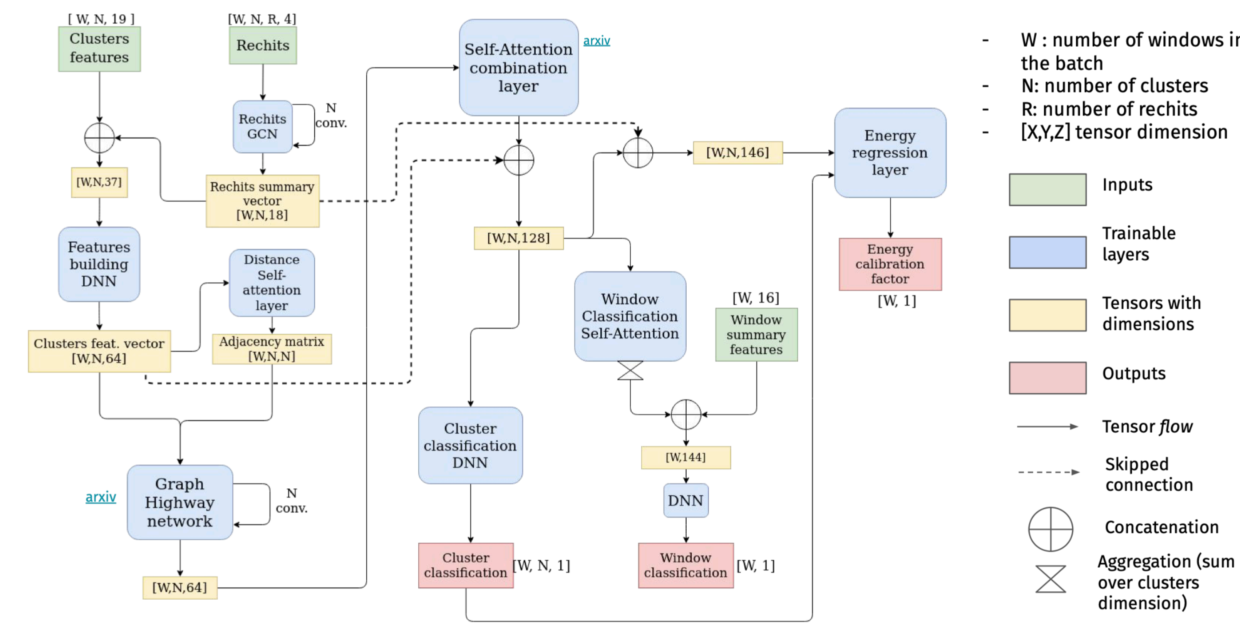
B. Maier, S. M. Narayanan, G. de Castro, M. Goncharov, C. Paus and M. Schott
[[arXiv:2107.02779]]

Semi-supervised GNN

Y. Feng
[talk@BOOST2021]



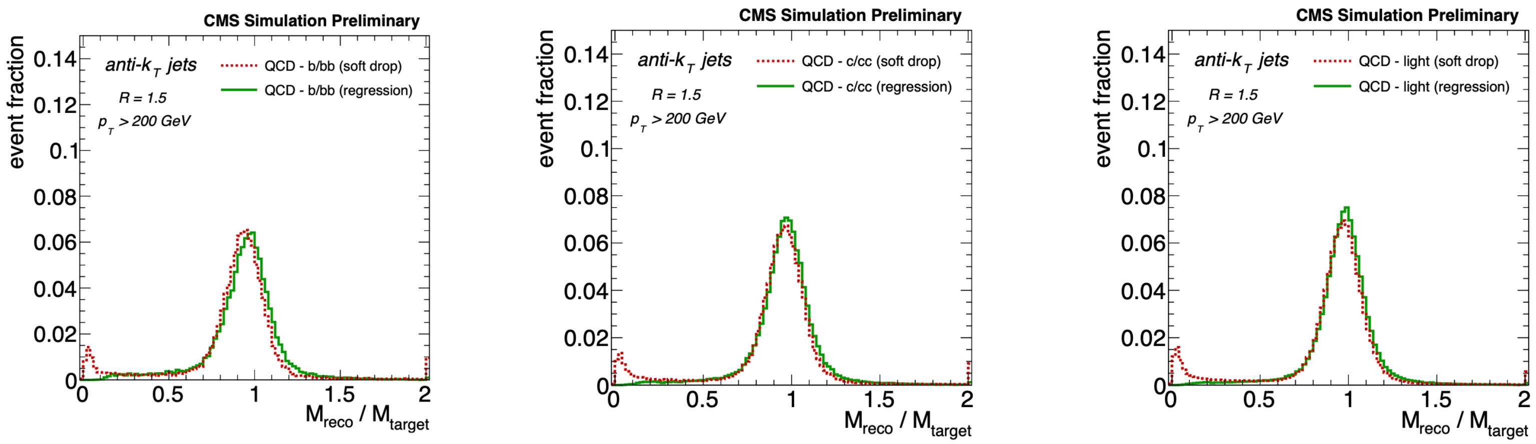
Pileup mitigation using graph neural network and transformers

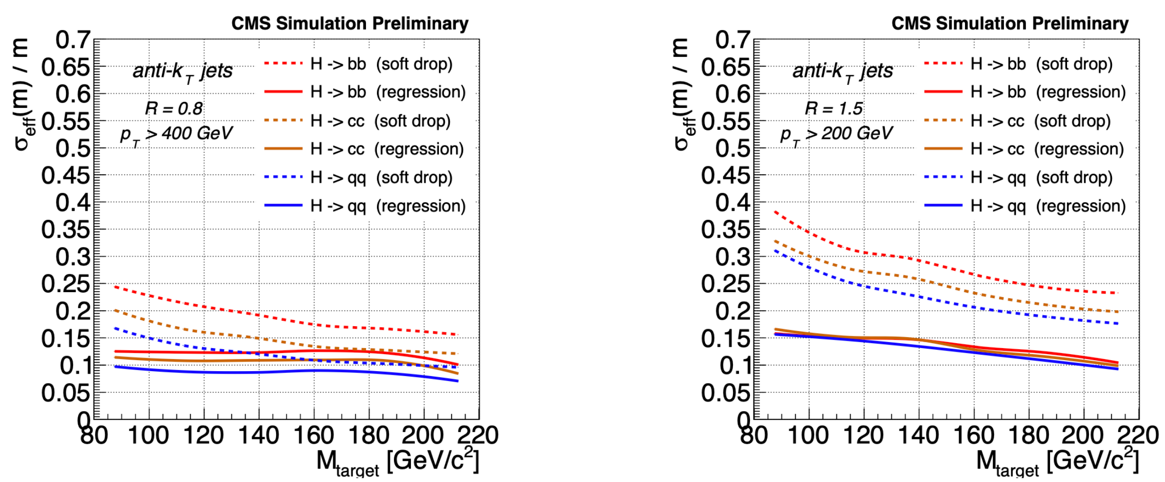


Single cluster classification: binary cross-entropy	⊕	Clusters classification by window: softF1 loss	⊕	Window classification: multiclass softmax	⊕	Energy regression: quantile regression loss
Standard loss for classification of the single cluster		SoftF1 loss balances the efficiency and purity of cluster selection in each window		Softmax between the different window classes mixed in the training (ele+gamma at the moment)		Optimize $(E-E_{Gen})^2$ for the mean, 25% and 75% quantiles

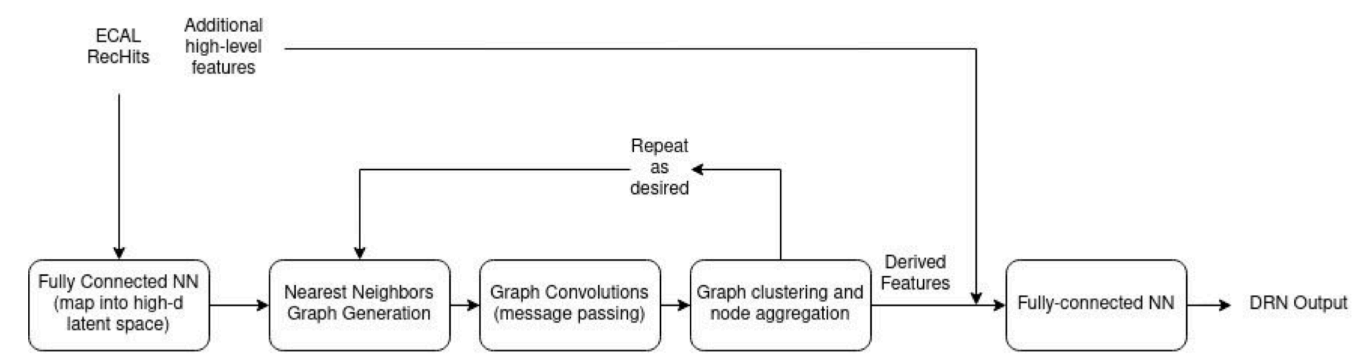
ECAL **superclustering** with machine learning

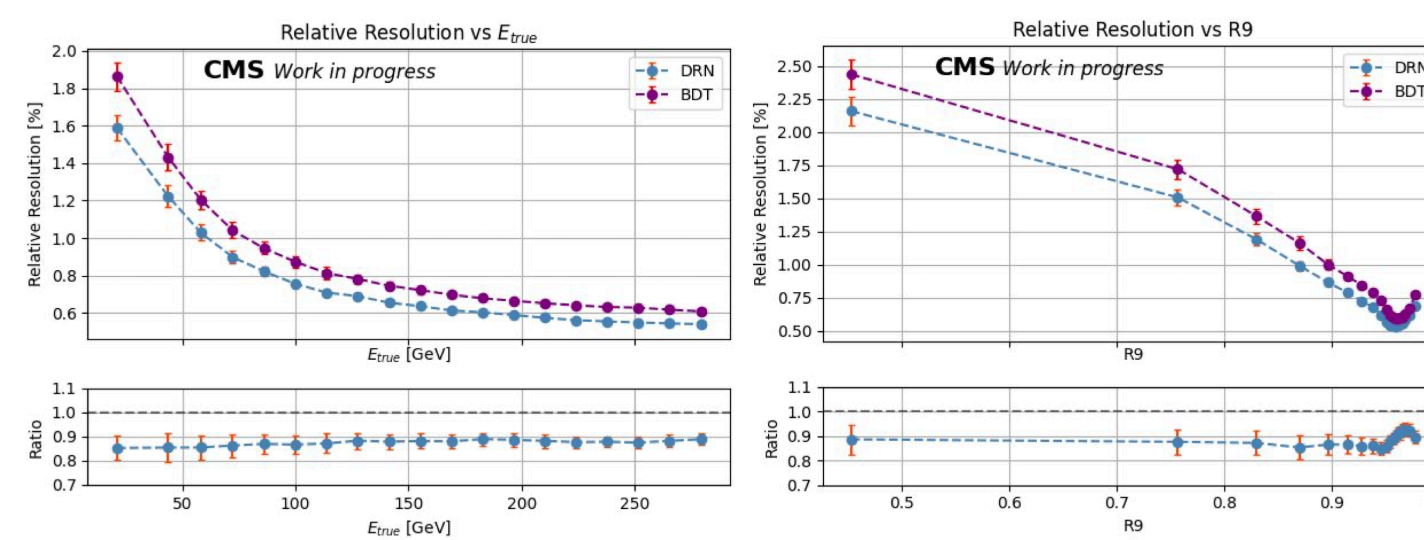
Geometric Deep Learning (IV)





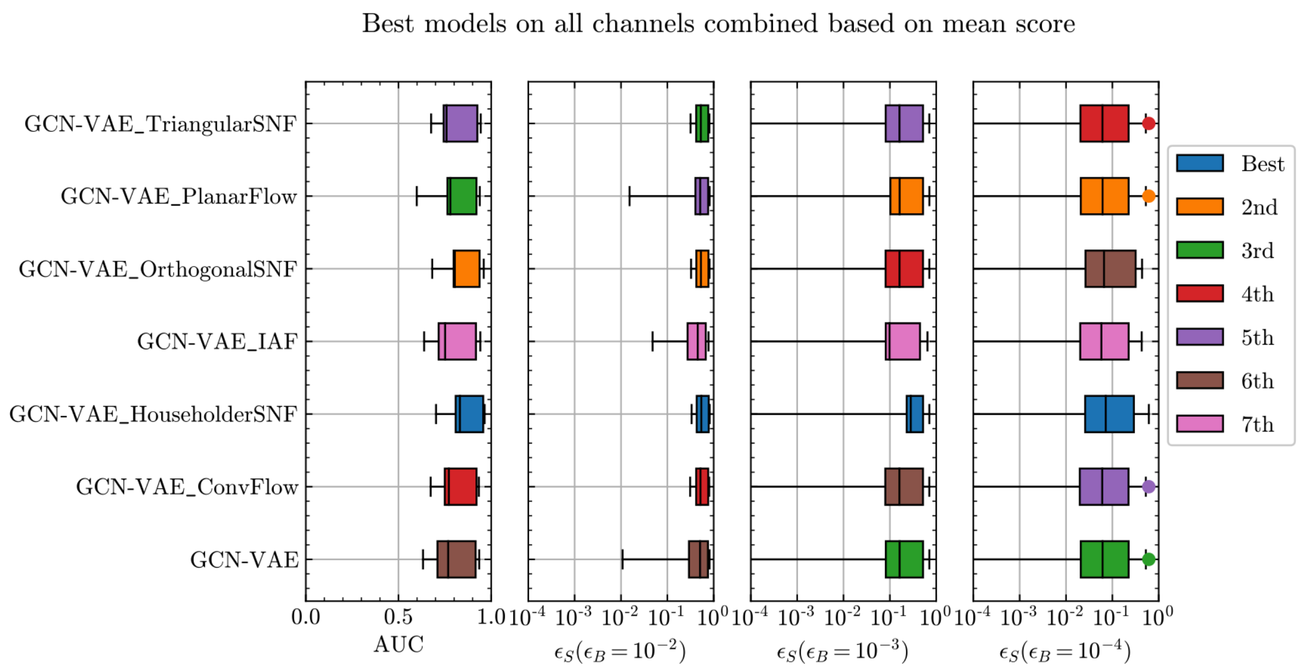
Jet mass regression using ParticleNet model, [2777006]



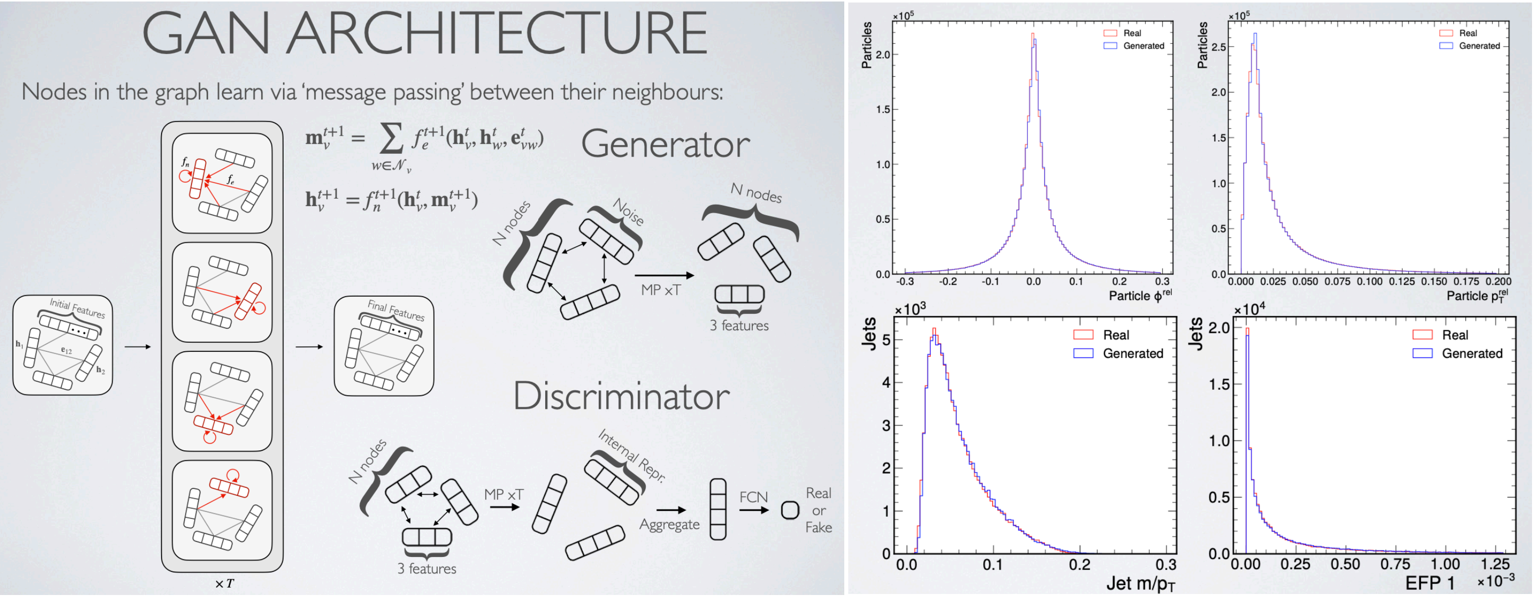


E/G **energy regression** using dynamic reduction network, [2003.08013]

Geometric Deep Learning (V)



Anomalous jet detection using graph convolution network variational auto-encoder with normalizing flow in the latent space, [2110.08508]

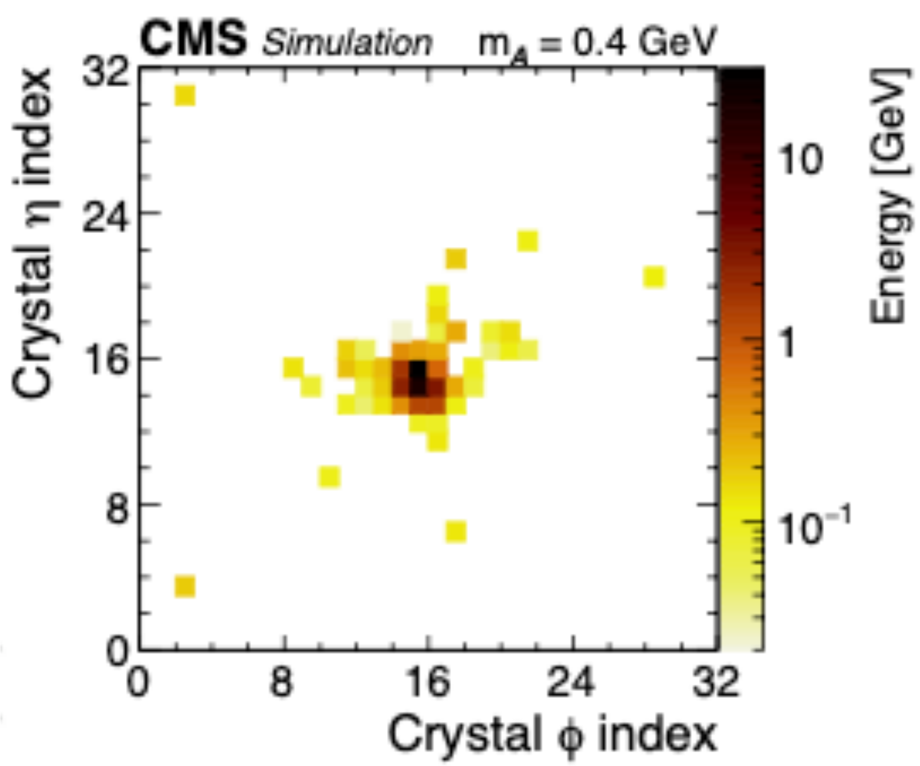


Jet particle-based **simulation** with message passing GNN generative adversarial network, [2012.00173]

ML in CMS

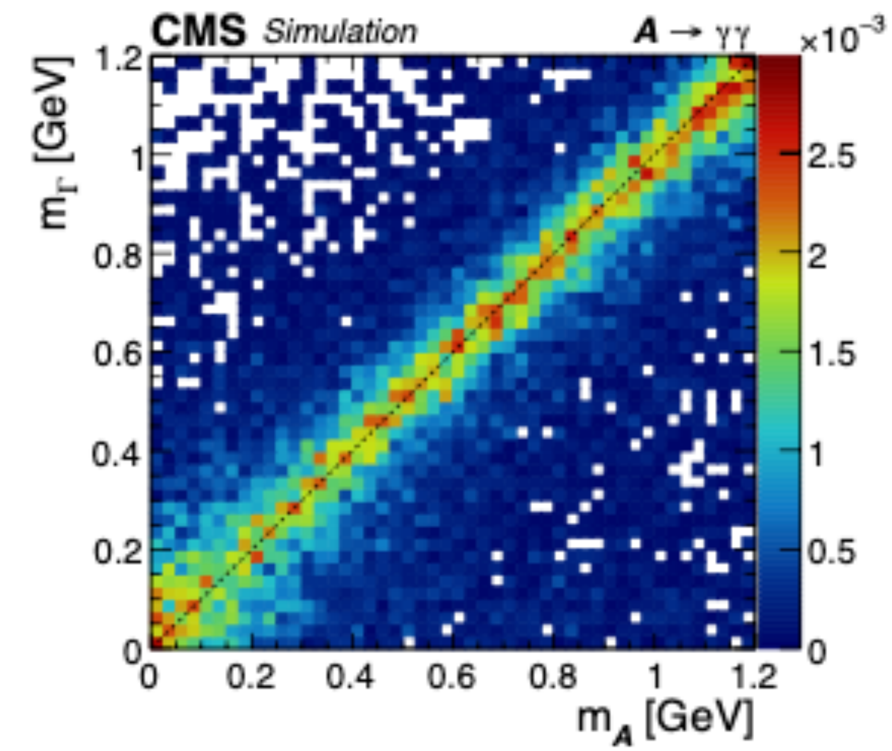
a selected pick of recent results …

di-photon Mass Regression



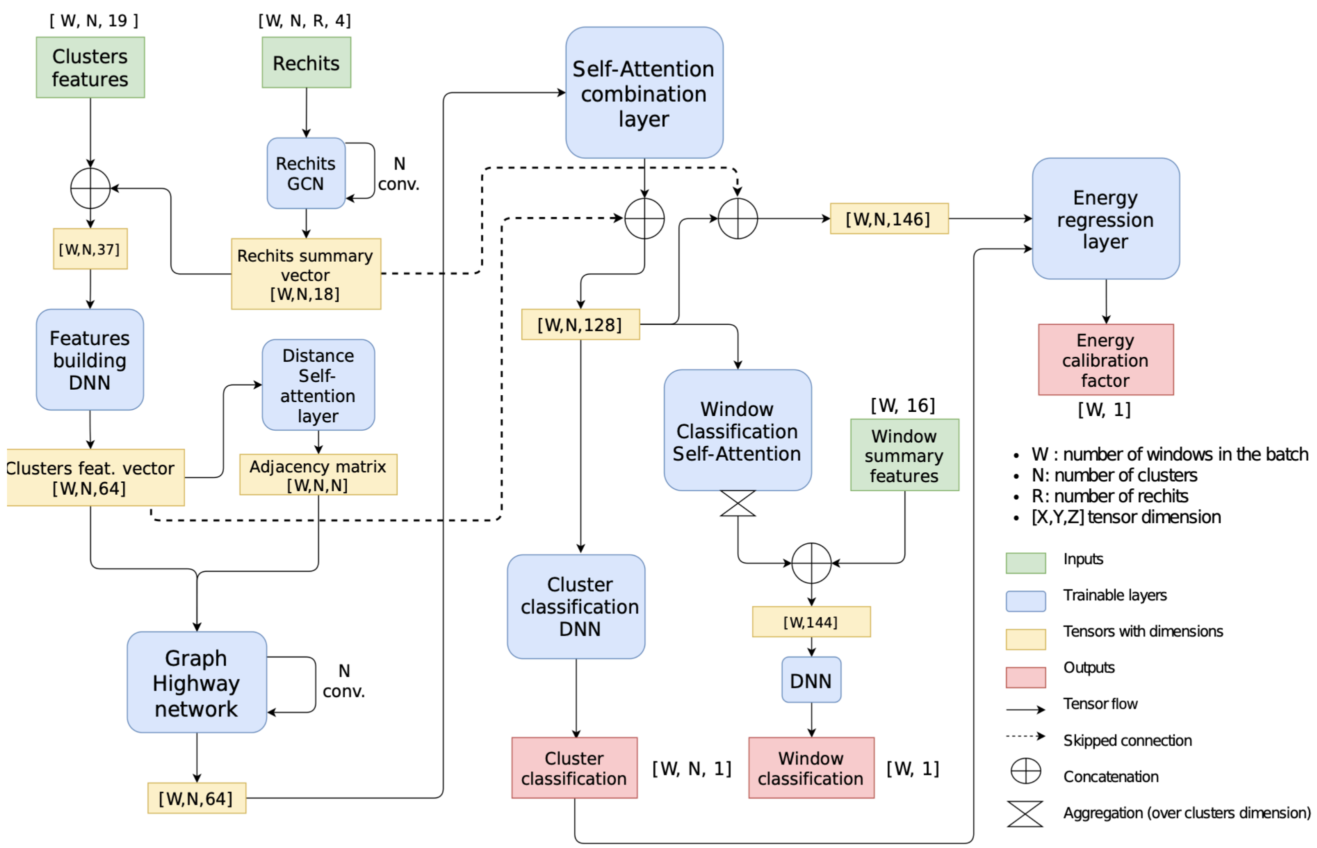
RESNET for mass regression
+
domain continuation a low mass

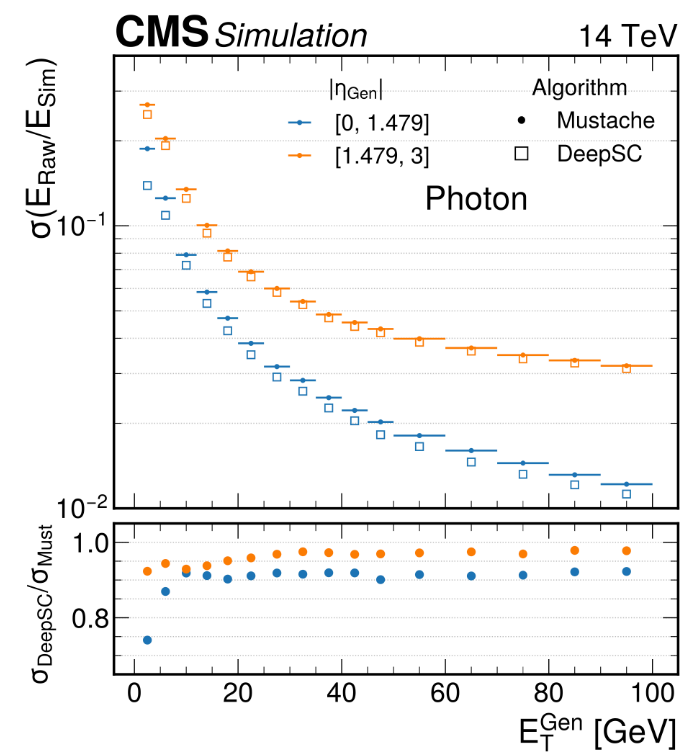
CMS Paper EGM-20-001 to appear soon

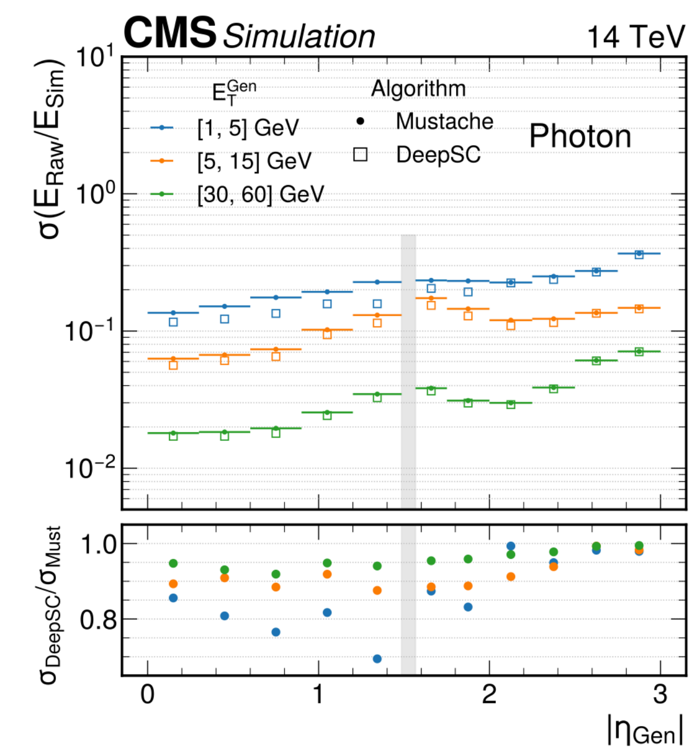


Learn the a/di-photon mass from the energy deposition at the Ecal surface.
Unprecedented reach at low mass.

Ecal Regression



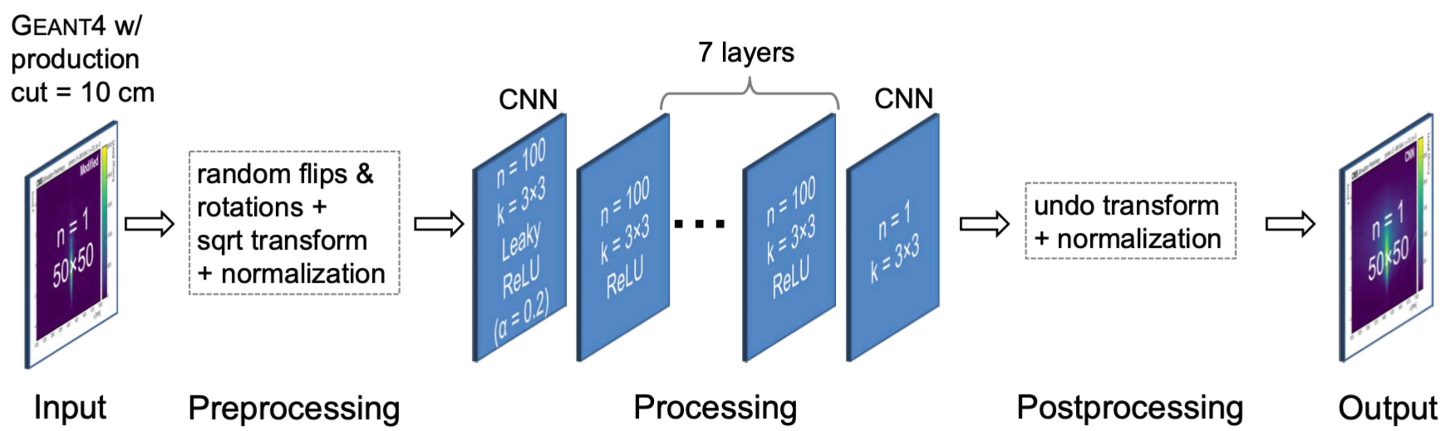




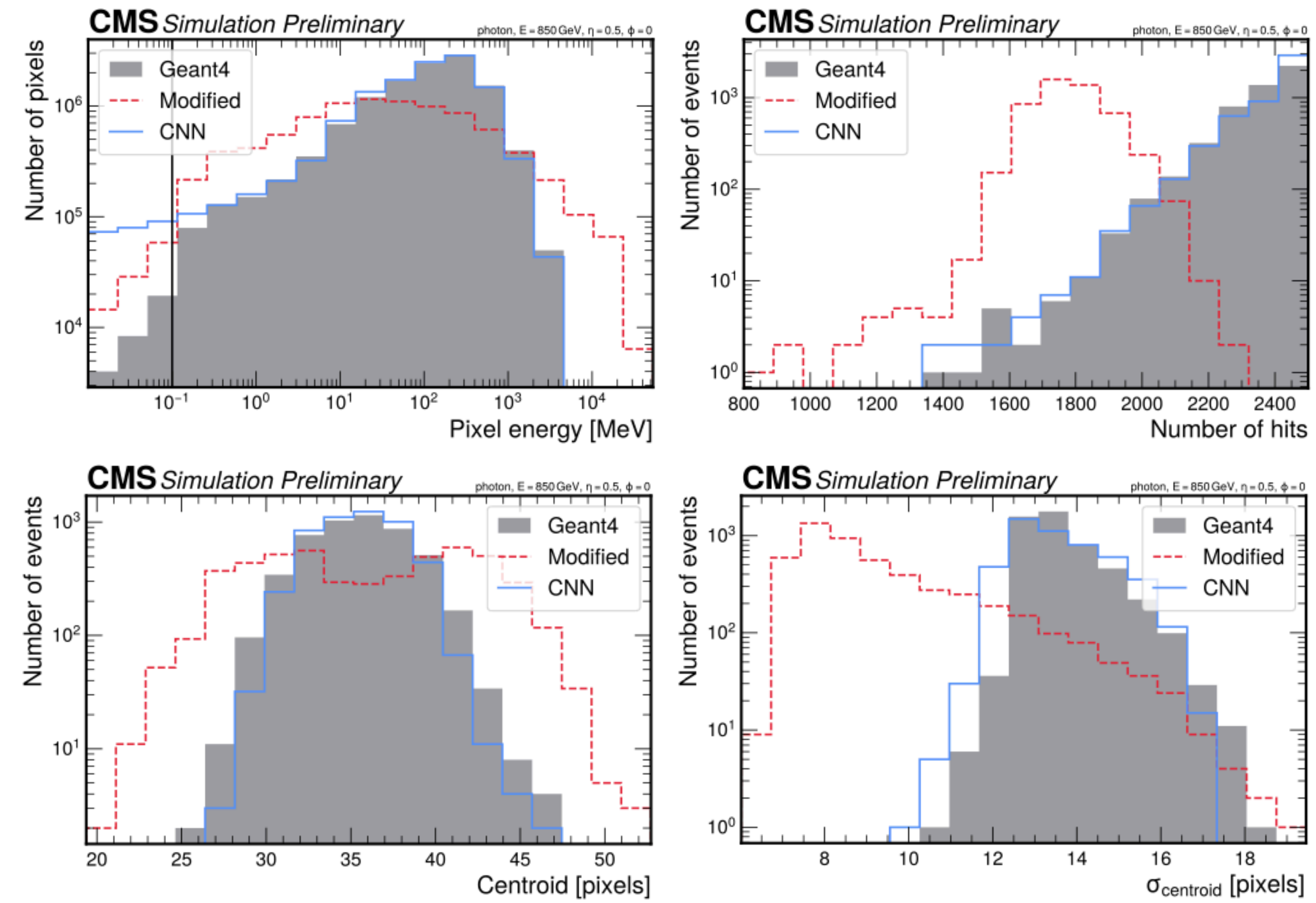
[cds:2803235]

- Graph-based model with self-attention trained to :
  - ✓seed-cluster classification
  - ✓super-cluster classification
  - ✓super-cluster energy regression
- Promising work in progress for calorimeter reconstruction

Super-resolution Simulation

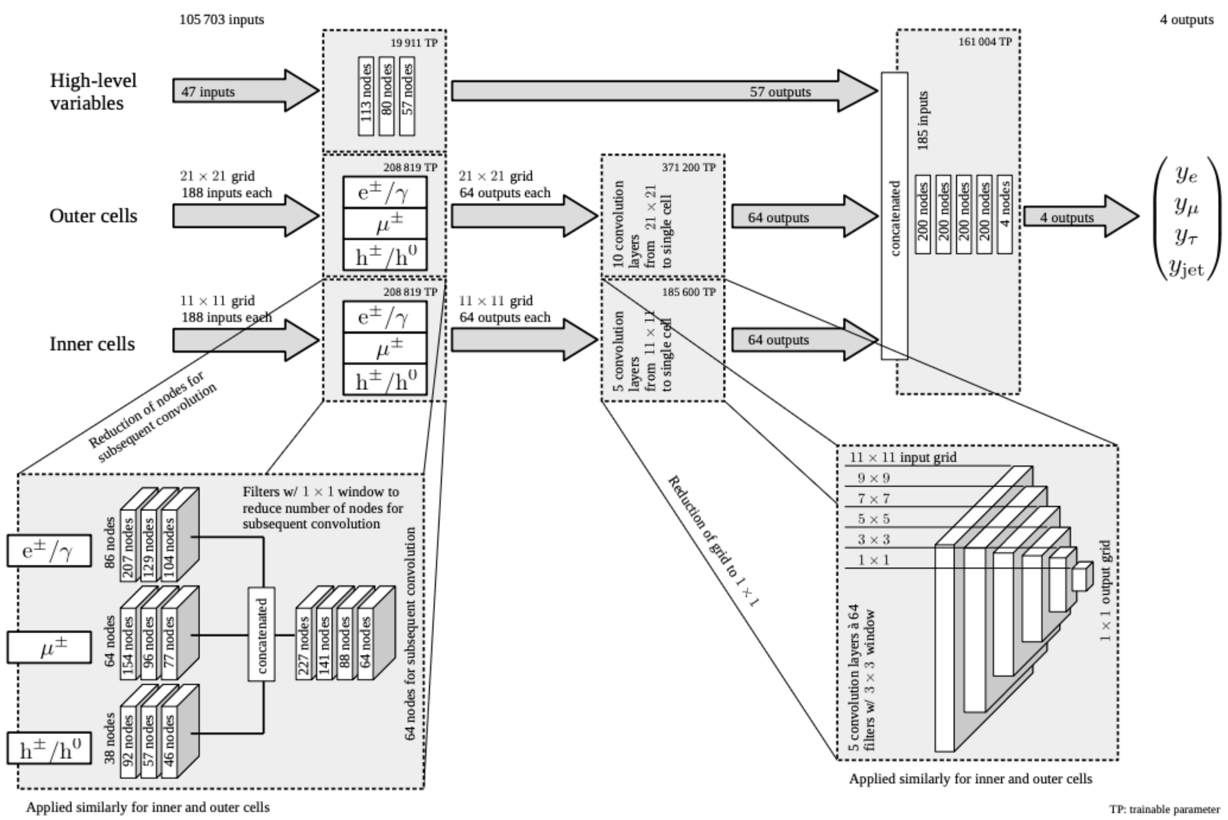


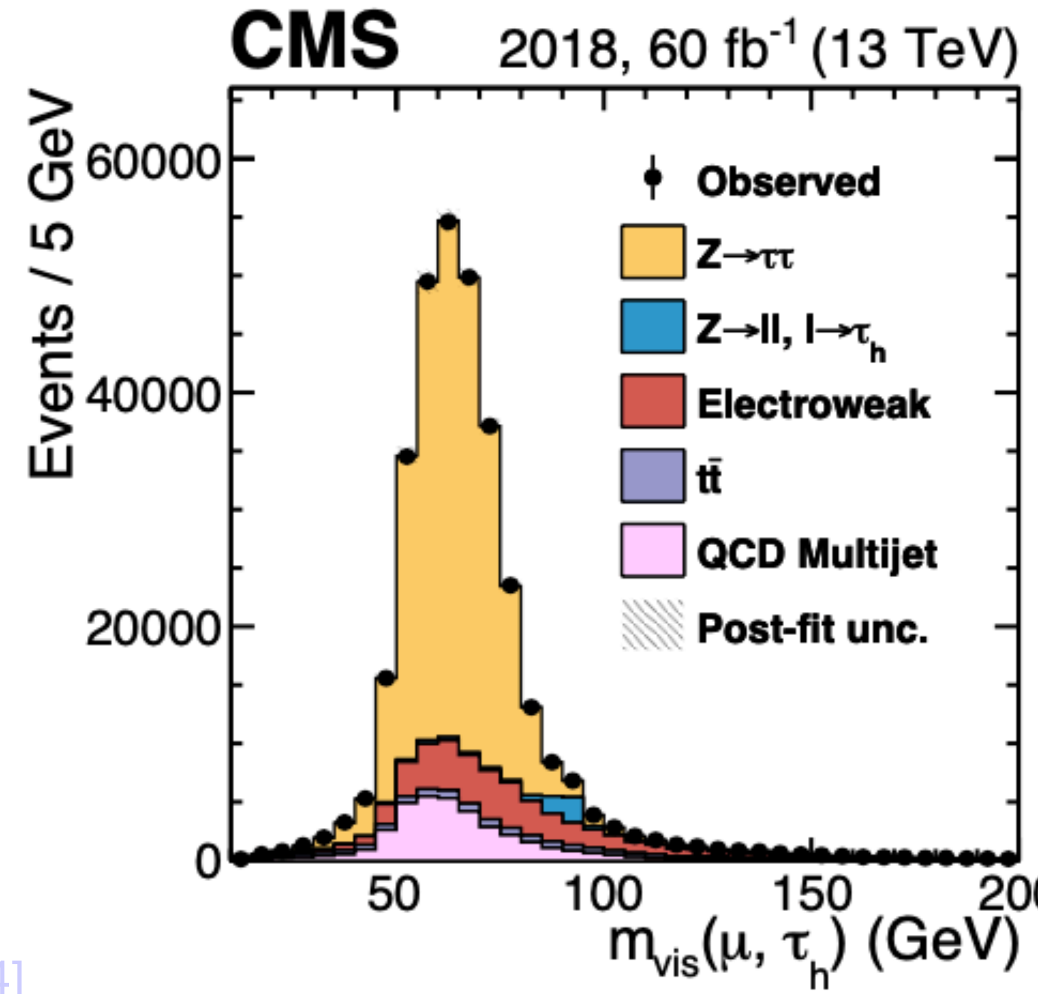
[cds:2802586]



- Run GEANT4 with loose parameters as low-quality input
- Learn the full precision high-quality output with CNN
- Model able to “denoise” and approach full precision

Hadronic Tau Identification

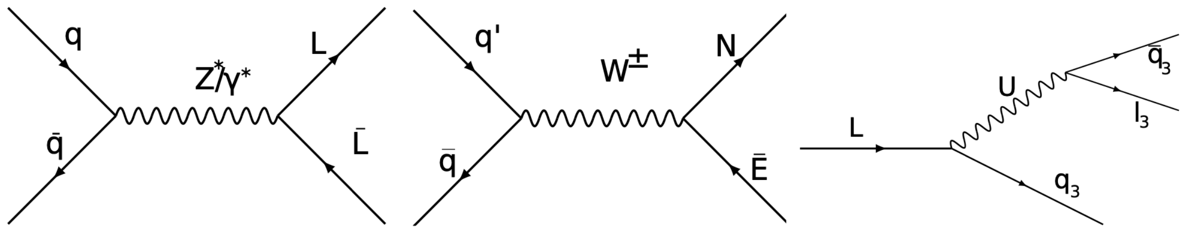


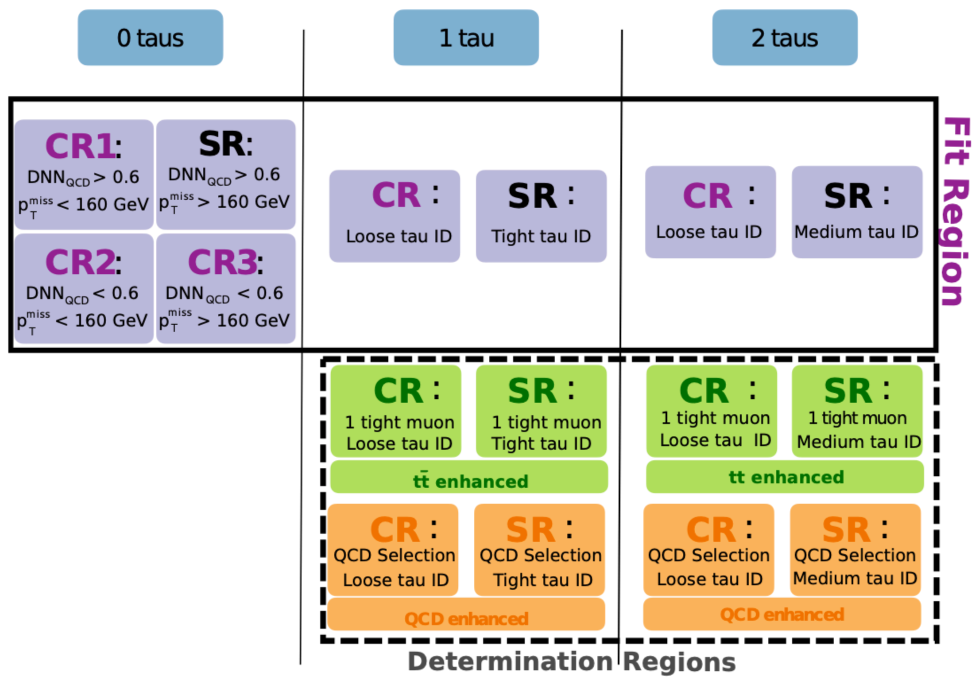


[cds:2800114]

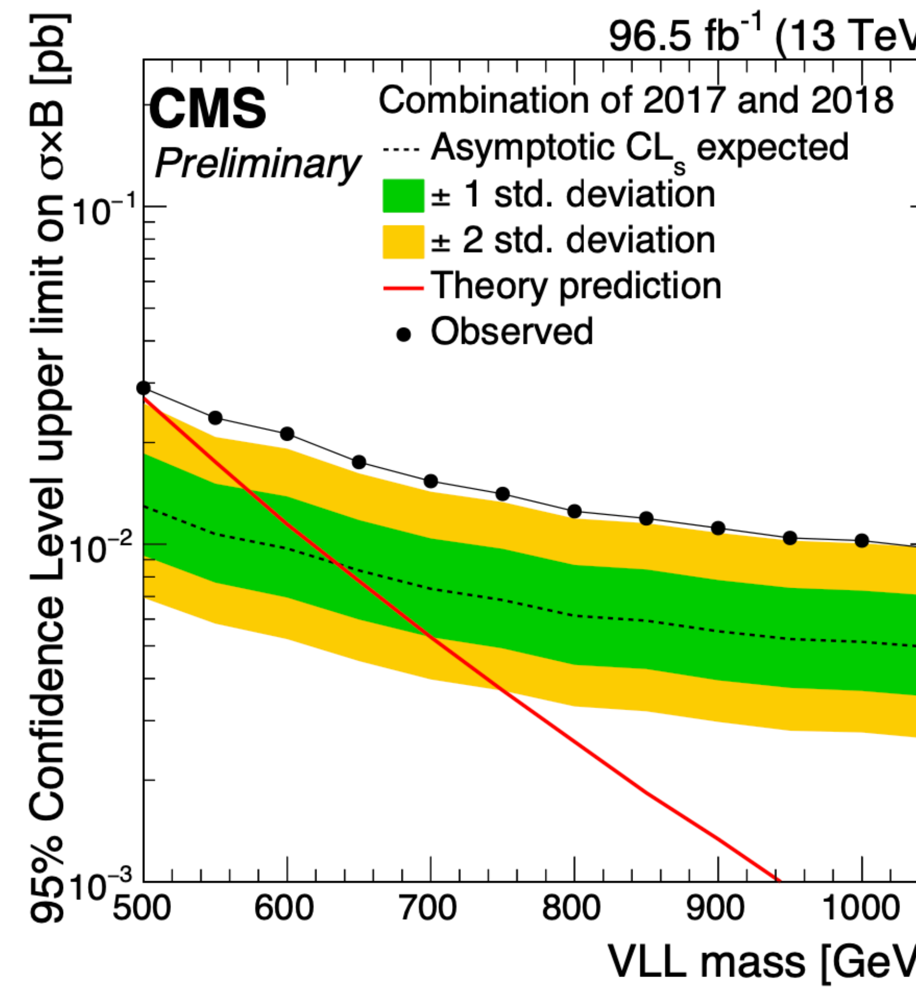
- Combines jet features and particle-image features
- CNN model to classify hadronic tau
- Much reduced fake rate
- More hadronic taus in analysis

Vector-Like Lepton Pair Search



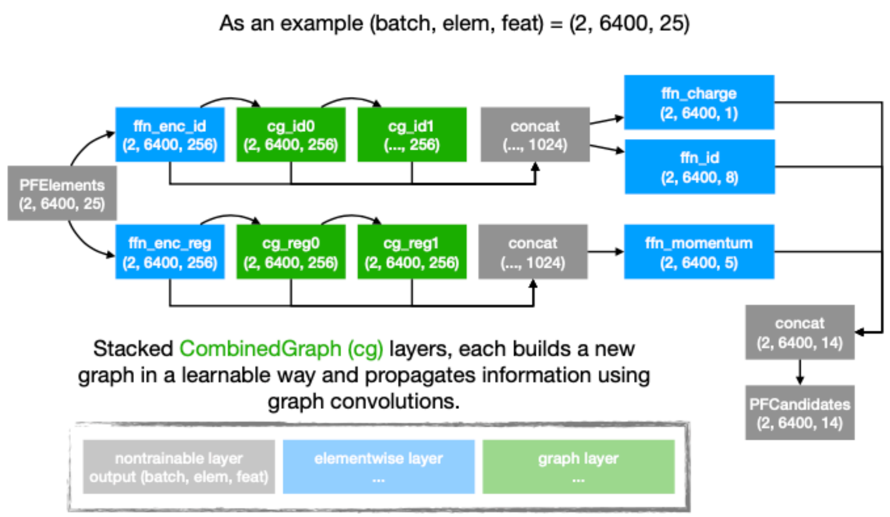


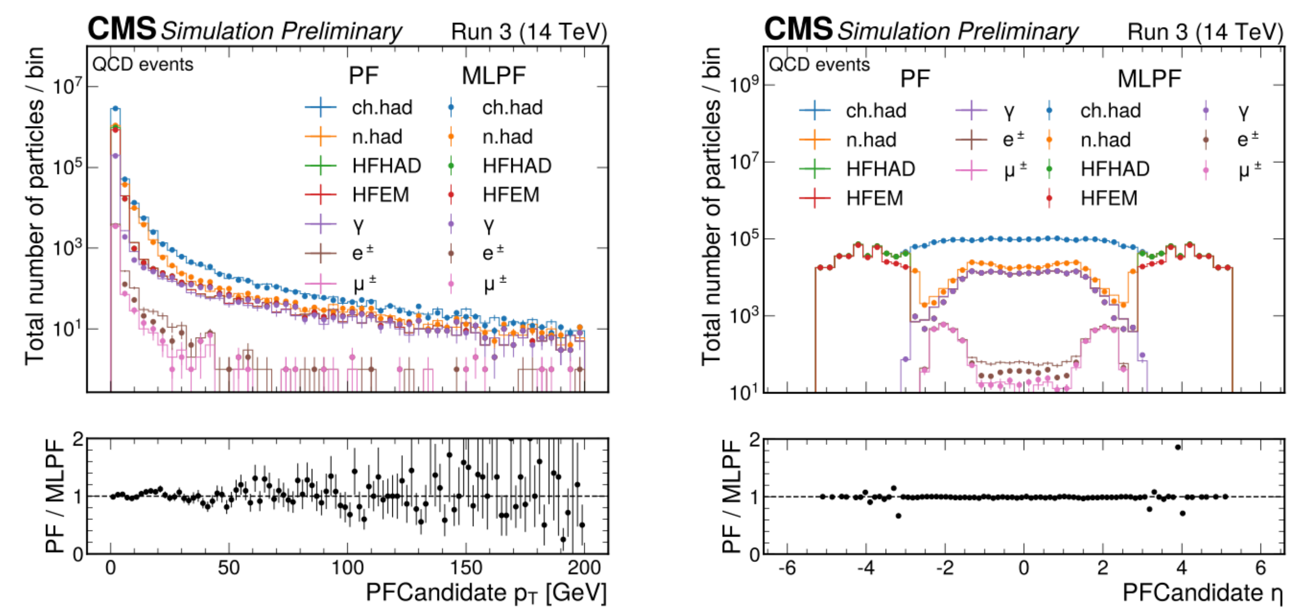
[cds:2803736]

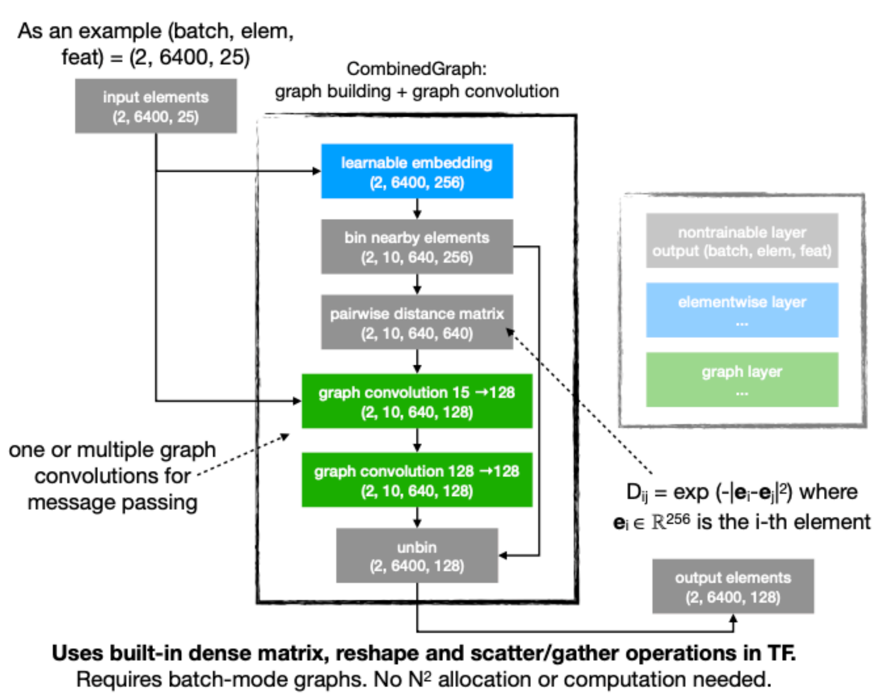


- At least 3b jets and two third generation leptons in final state
- DeepTau [cds:2800114] method used for tau identification.
- Attention-based graph model [2001.05311] working on final state objects acts as classifier used for signal categorization.
- State of the art deep-learning in state of the art NP search

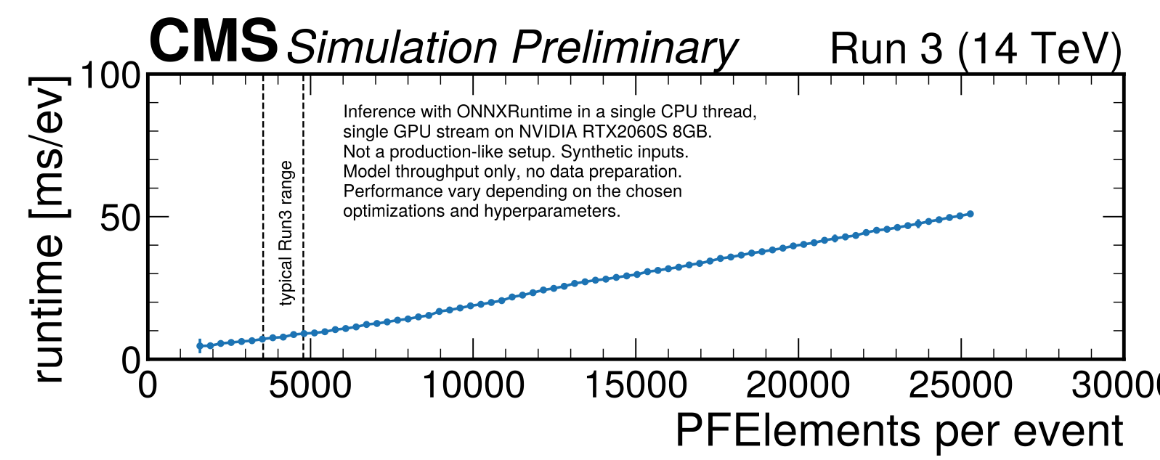
Particle-Flow Reconstruction





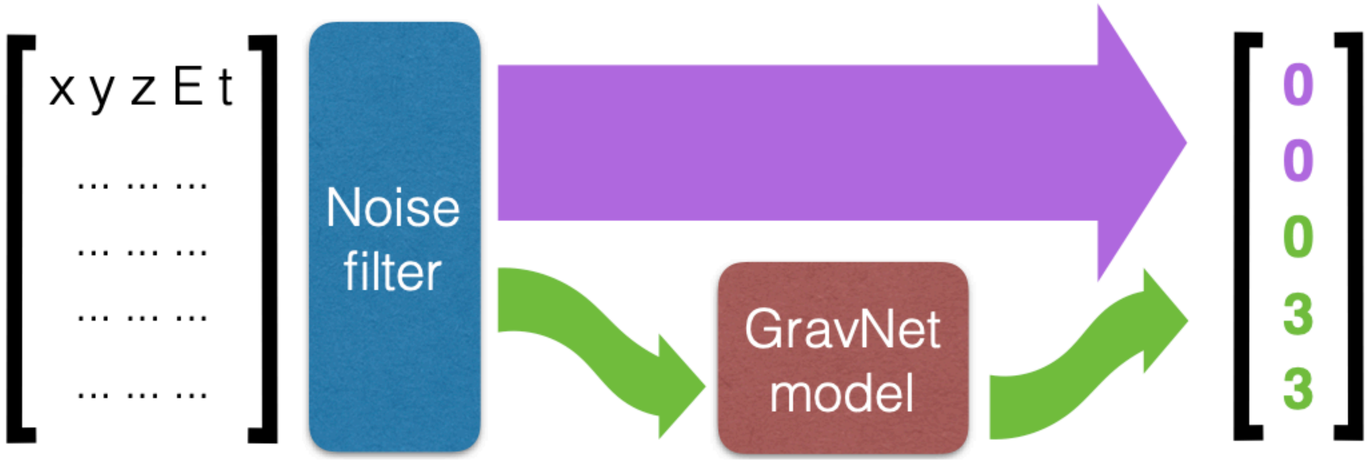


[cds:2802826]

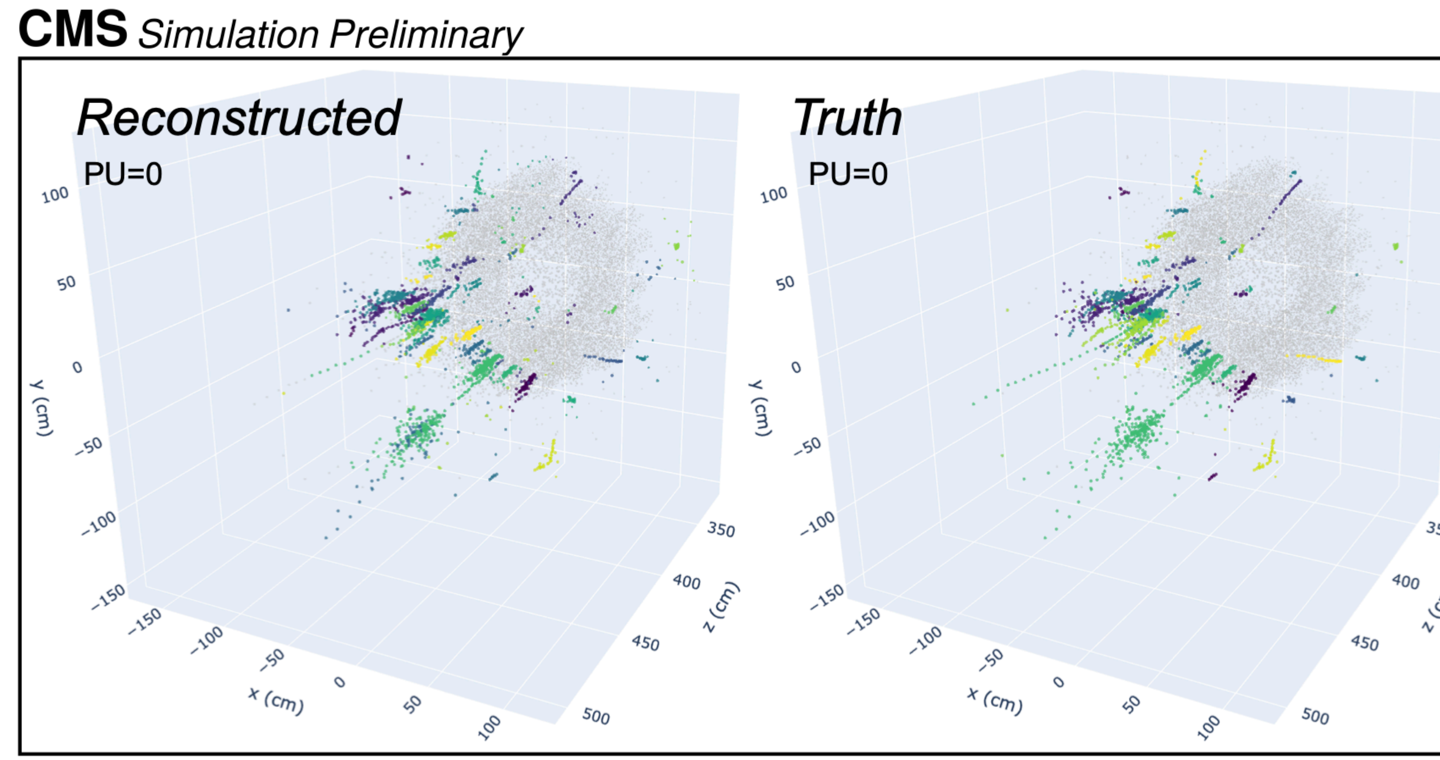


- Set of tracks & clusters in input to graph-based model
- Classify sub-set of graph nodes as particle candidates
- Regress parton kinematics from candidate
- Model almost matching classical algorithm
- Execution time quasi-linear with pile-up

Particle Reconstruction at HL-LHC

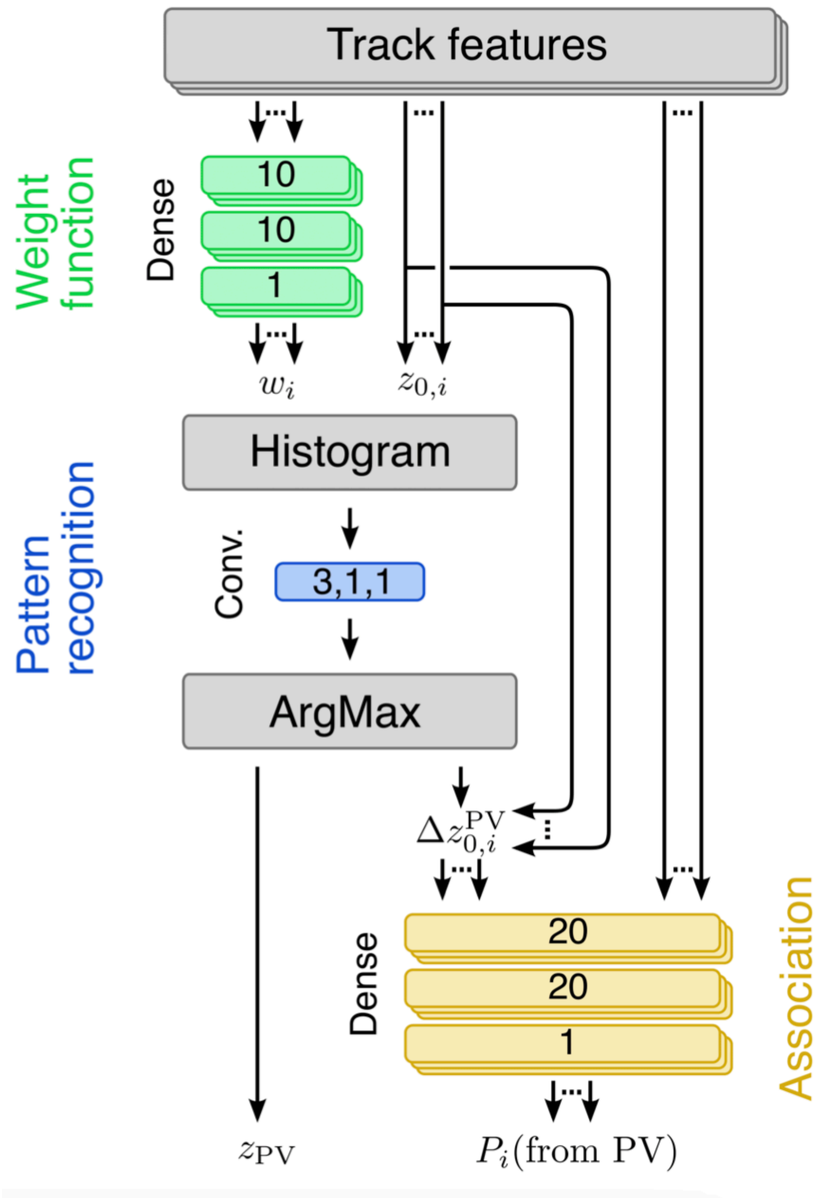


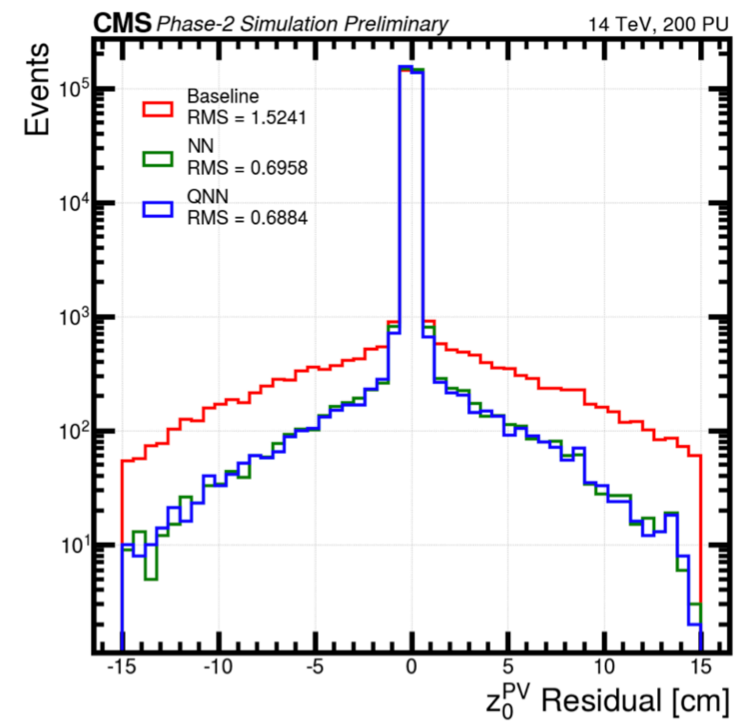
[cds:2803236]

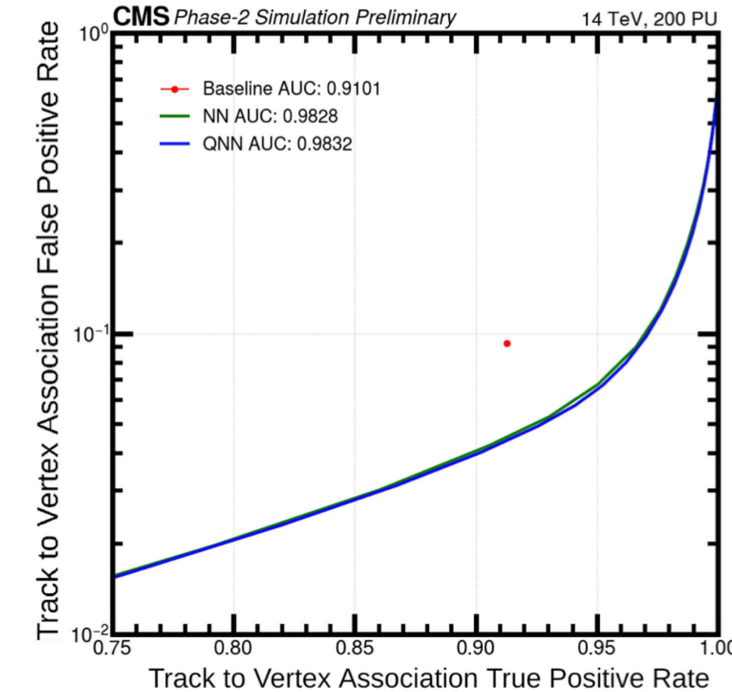


- High-Granularity Calorimeter (HGCAL) provides fine-grained description of energy deposition
- Graph-based models [2106.01832] using object condensation loss [2002.03605] trained to perform cell-to-particle association
- Stepping stone towards ML-based particle reconstruction in HGCAL

Vertexing at L1 at HL-LHC







[cds:2801638]

- Tracks reconstructed at L1 used in input
- Model regress position of primary vertex and track-PV assignment
- Quantized/pruned model can efficiently deploy on FPGA

Prospects for Deep Learning

a few words of wisdom …

Interpretability

Interplay between deep learning and science is key.
Use Physics knowledge to produce better models.
Use models to learn Physics knowledge.

Uncertainty Quantification

Propagation and estimation of uncertainties are keys.
Uncertainty-aware models.
Uncertainty-predicting models.
Uncertainty-improving models.

Computation Aspect

Computational cost of Science is key.

Adapt to heterogenous computing environment.

Hardware-aware model optimization.

Publication Plans

Publishing in peer-reviewed journal is key.
Importance of open-data samples.
Flexibility in experiments to publish work in progress.

Summary

➡Modern machine learning a.k.a Deep Learning goes much beyond classification.

➡GDL is most promising for many applications.

➡Novel Deep Learning are being adopted in CMS. Many more upcoming results.

➡The future of AI4HEP is interpretable, quantifiable, runnable and publishable …

DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

mPP

4pc

A Definition

"Giving computers the ability to learn without explicitly programming them" A. Samuel (1959).

Is fitting a straight line machine learning ?
Models that have enough capacity to define its own internal representation of the data to accomplish a task : **learning from data.**

In practice : a statistical method that can extract information from the data, not obviously apparent to an observer.

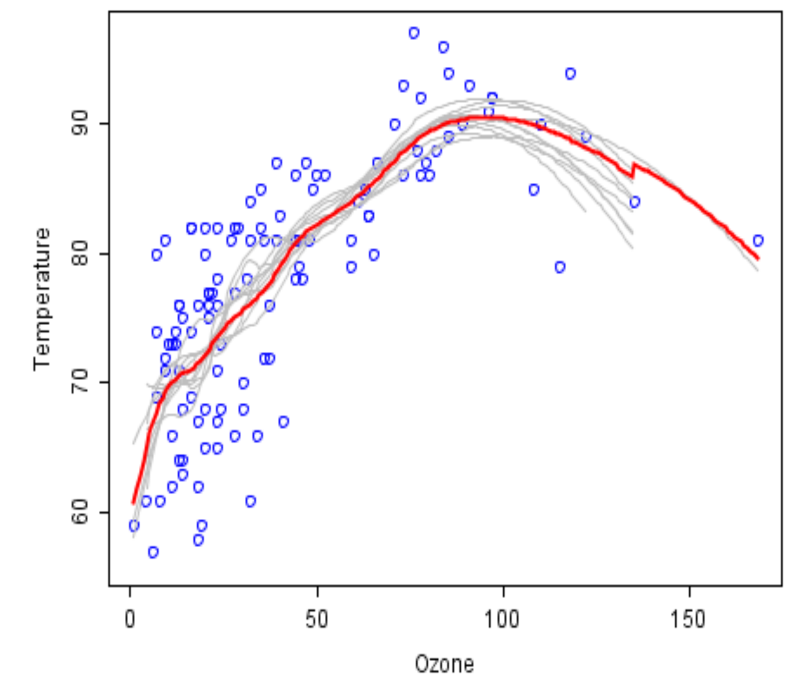
→ Most approach will involve a **mathematical model** and a cost/reward function that needs to be **optimized.**

→ The more **domain knowledge** is incorporated, the better.

Supervised Learning

- Given a dataset of samples, a subset of features is qualified as **target**, and the rest as **input**
- Find a **mapping from input to target**
- The mapping should **generalize to any extension** of the given dataset, provided it is generated from the same mechanism

$$dataset \equiv \{(x_i, y_i)\}_i$$
$$find \;\; function \;\; f \;\; s.t. \;\; f(x_i) = y_i$$



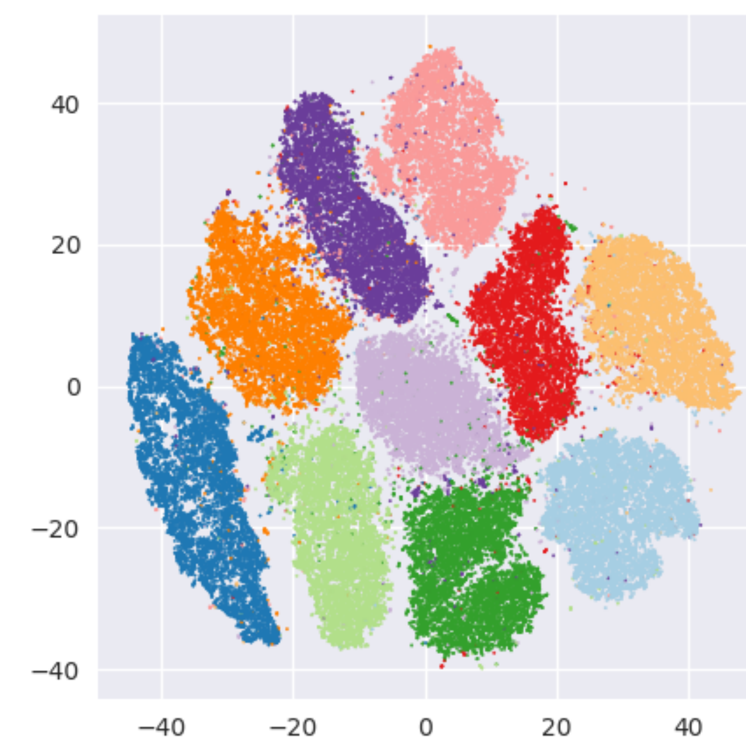
- Finite set of target values :
  → **Classification**
- Target is a continuous variable :
  → **Regression**

Unsupervised Learning

- Given a dataset of samples, but there is no subset of feature that one would like to predict
- Find mapping of the samples to a lower dimension manifold
- The mapping should generalize to any extension of the given dataset, provided it is generated from the same mechanism

$$dataset \equiv \{(x_i)\}_i$$
$$find\ \ f\ \ s.t.\ \ f(x_i) = p_i$$

- Manifold is a finite set
  → **Clusterization**
- Manifold is a lower dimension manifold :
  → **Dimensionality reduction, density estimator**



Reinforcement Learning

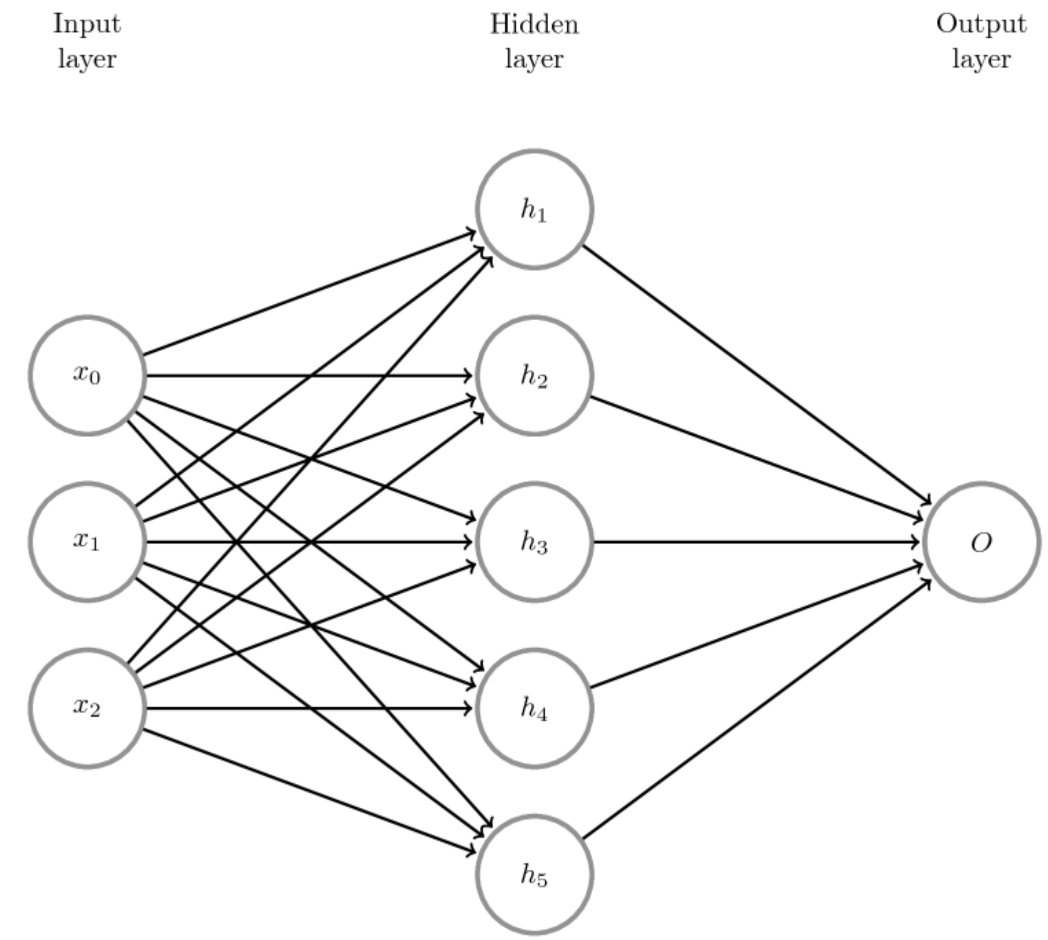
- Given an **environment** with multiple states, given a reward upon action being taken over a state
- Find an **action policy to drive** the environment toward maximum cumulative reward

$$s_{t+1} = Env(s_t, a_t)$$
$$r_t = Rew(s_t, a_t)$$
$$\pi(a|s) = P(A_t = a | S_t = s)$$
$$find\ \pi\ s.t. \sum_t r_t\ is\ maximum$$

Artificial Neural Network

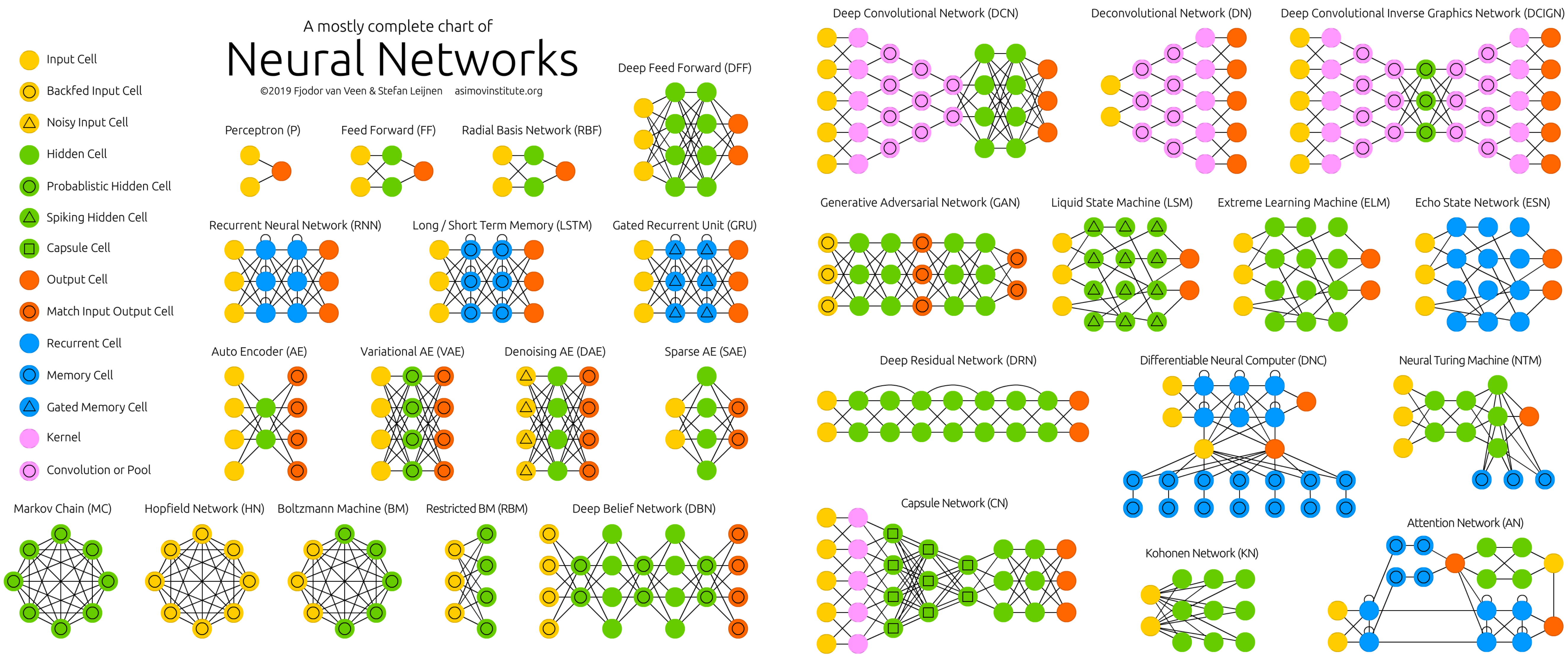
- **Biology inspired** analytical model, but **not bio-mimetic**
- Booming in recent decade thanks to large dataset, increased computational power and theoretical novelties
- Origin tied to logistic regression with change of data representation
- Part of any “deep learning” model nowadays
- Usually large number of parameters trained with stochastic gradient descent

$$h = \phi(Ux + v)$$
$$o(x) = \omega^T h + b$$
$$p_i \equiv p(y = 1 | x) \equiv \sigma(o(x)) = \frac{1}{1 + e^{-o(x)}}$$
$$loss_{XE} = -\sum_i y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)$$



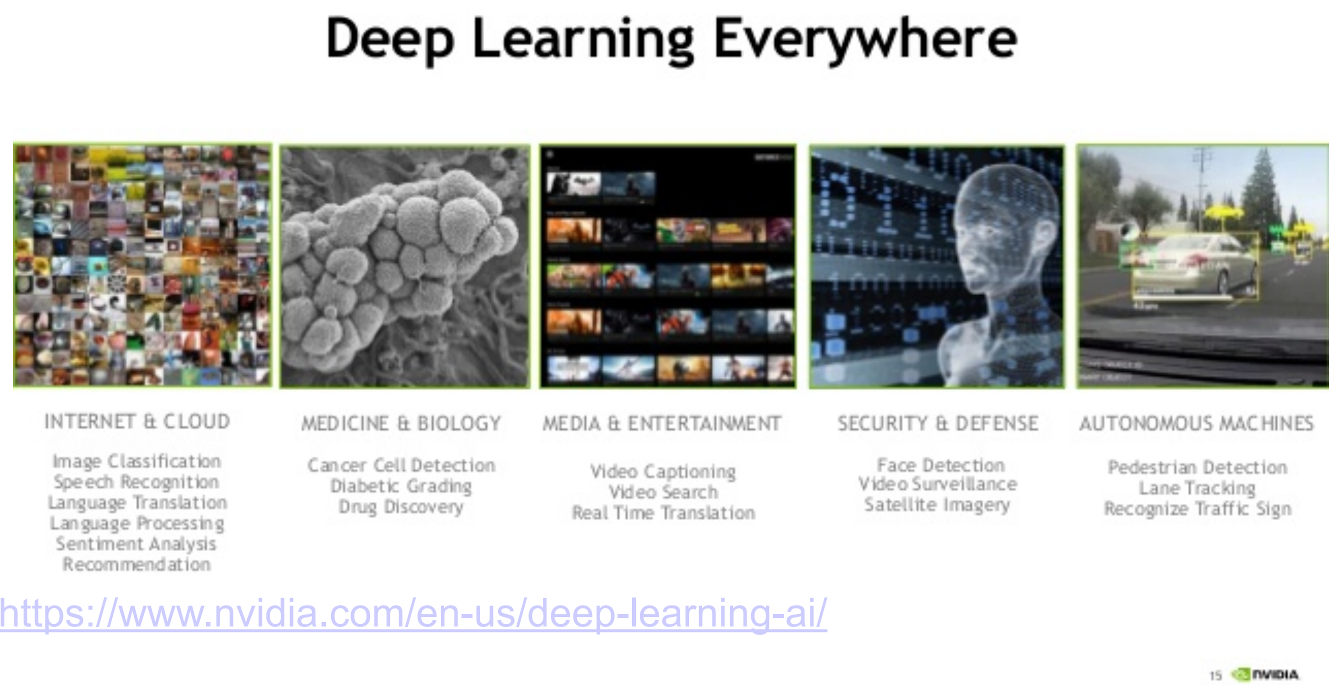
Neural Net Architectures

http://www.asimovinstitute.org/neural-network-zoo

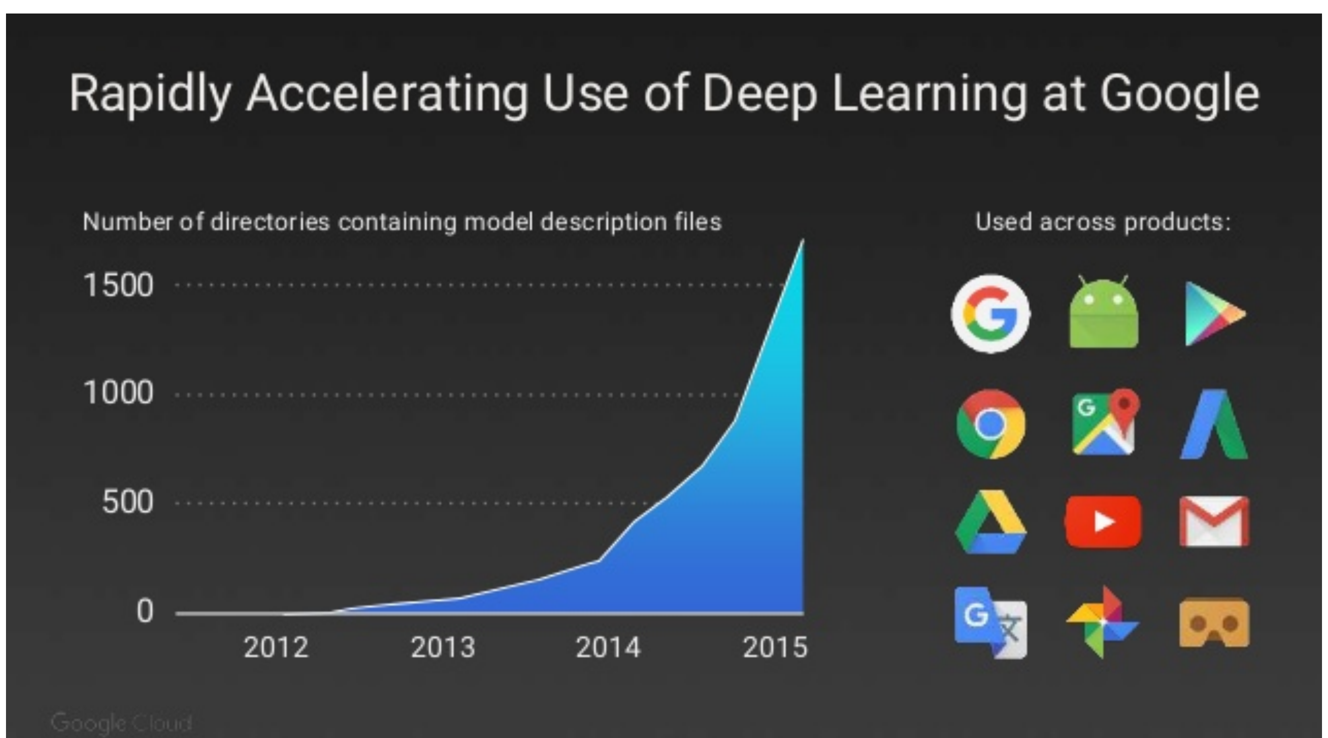


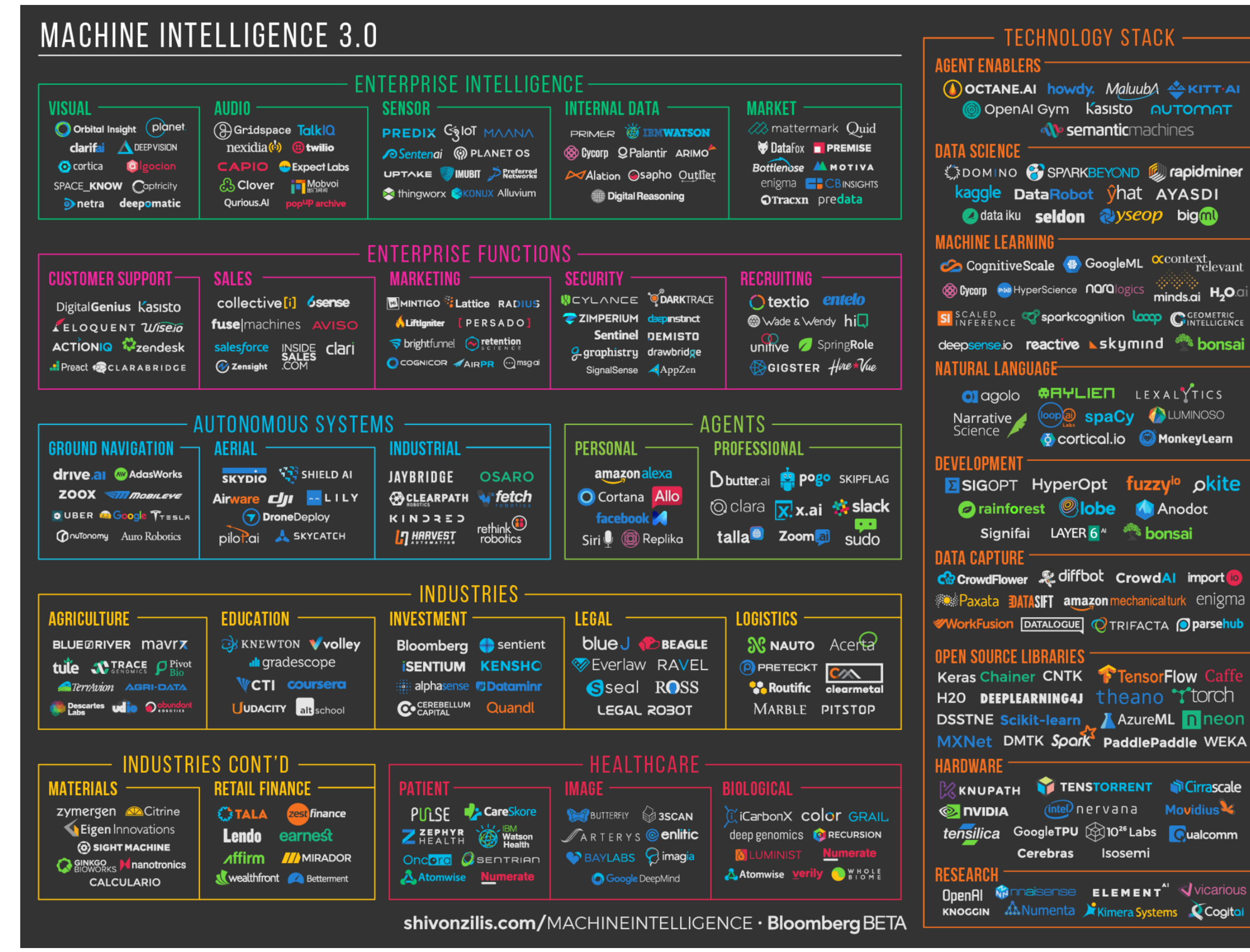
- Does not cover it all : densenet, graph network, ...

Machine Learning in Industry



https://www.nvidia.com/en-us/deep-learning-ai/

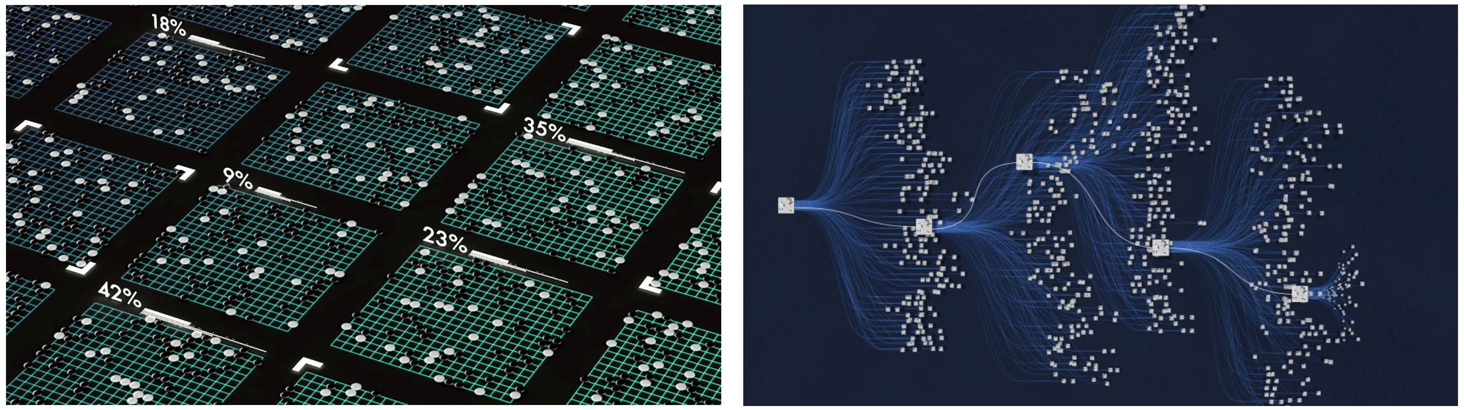


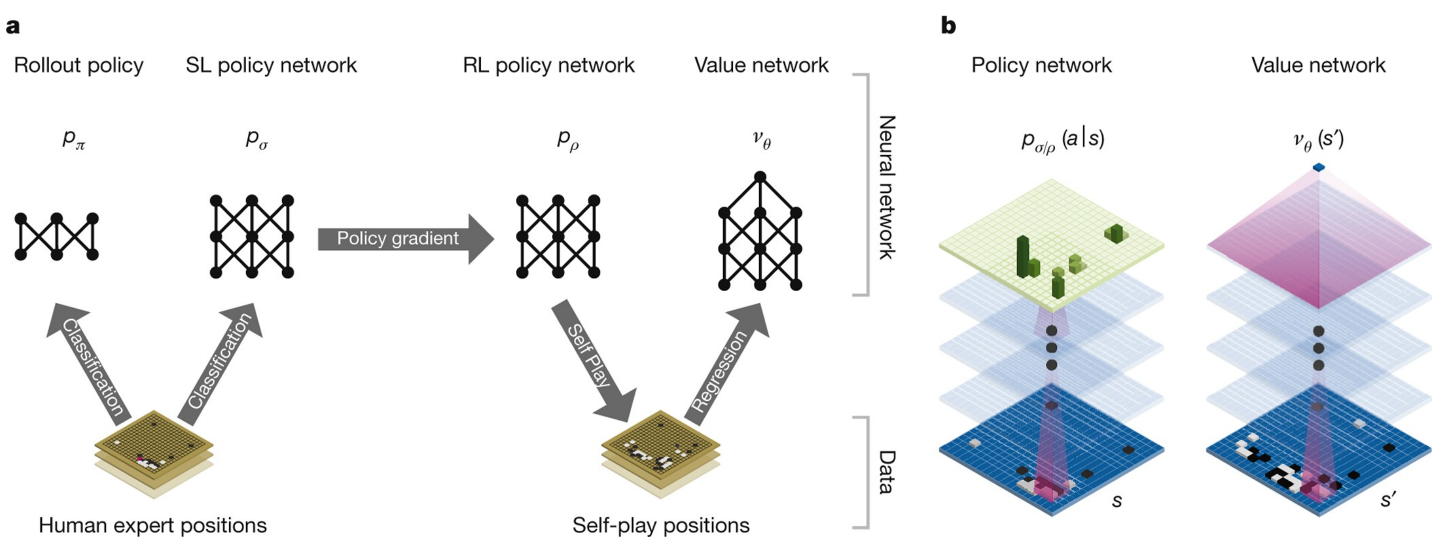


http://www.shivonzilis.com/machineintelligence

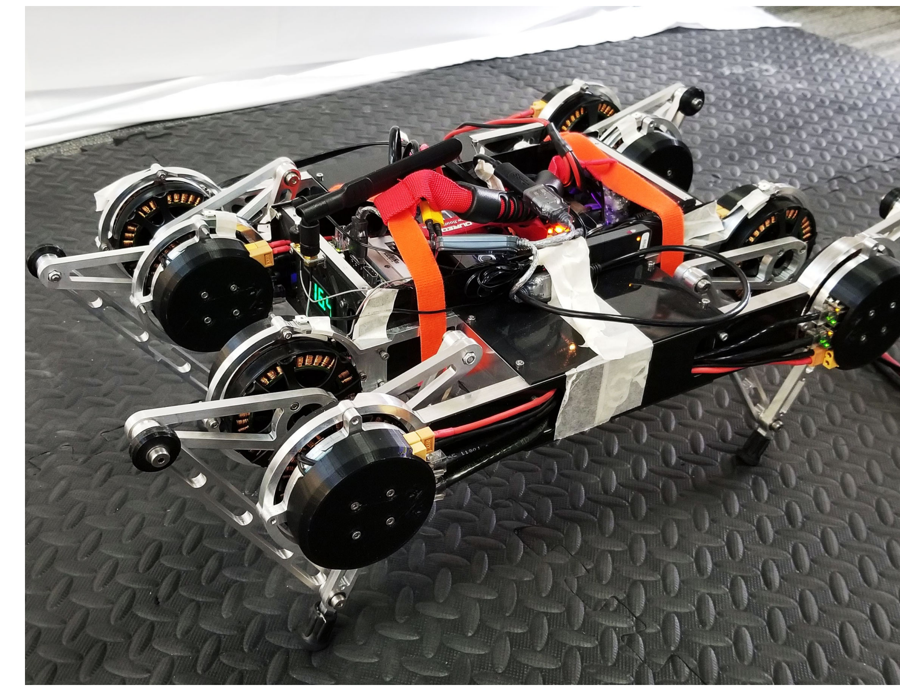
Prominent skill in industry nowadays. Lots of data, lots of applications, lots of potential use cases, lots of money. Knowing machine learning can open significantly **career horizons.**

Learning to Control





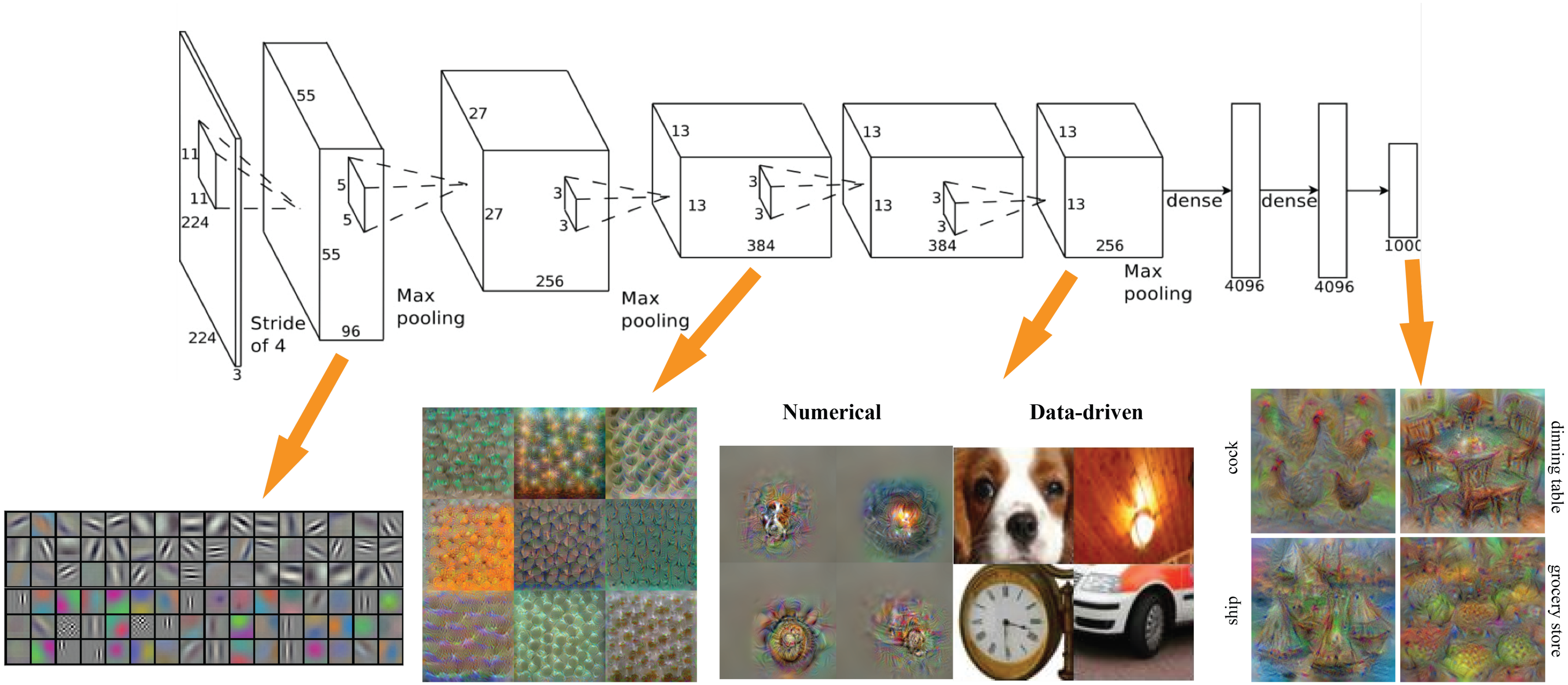
Mastering the game of Go with deep neural networks and tree search, https://doi.org/10.1038/nature16961



Learning to Walk via Deep Reinforcement Learning
https://arxiv.org/abs/1812.11103

Modern machine learning **boosts control technologies**.
AI, gaming, robotic, self-driving vehicle, etc.

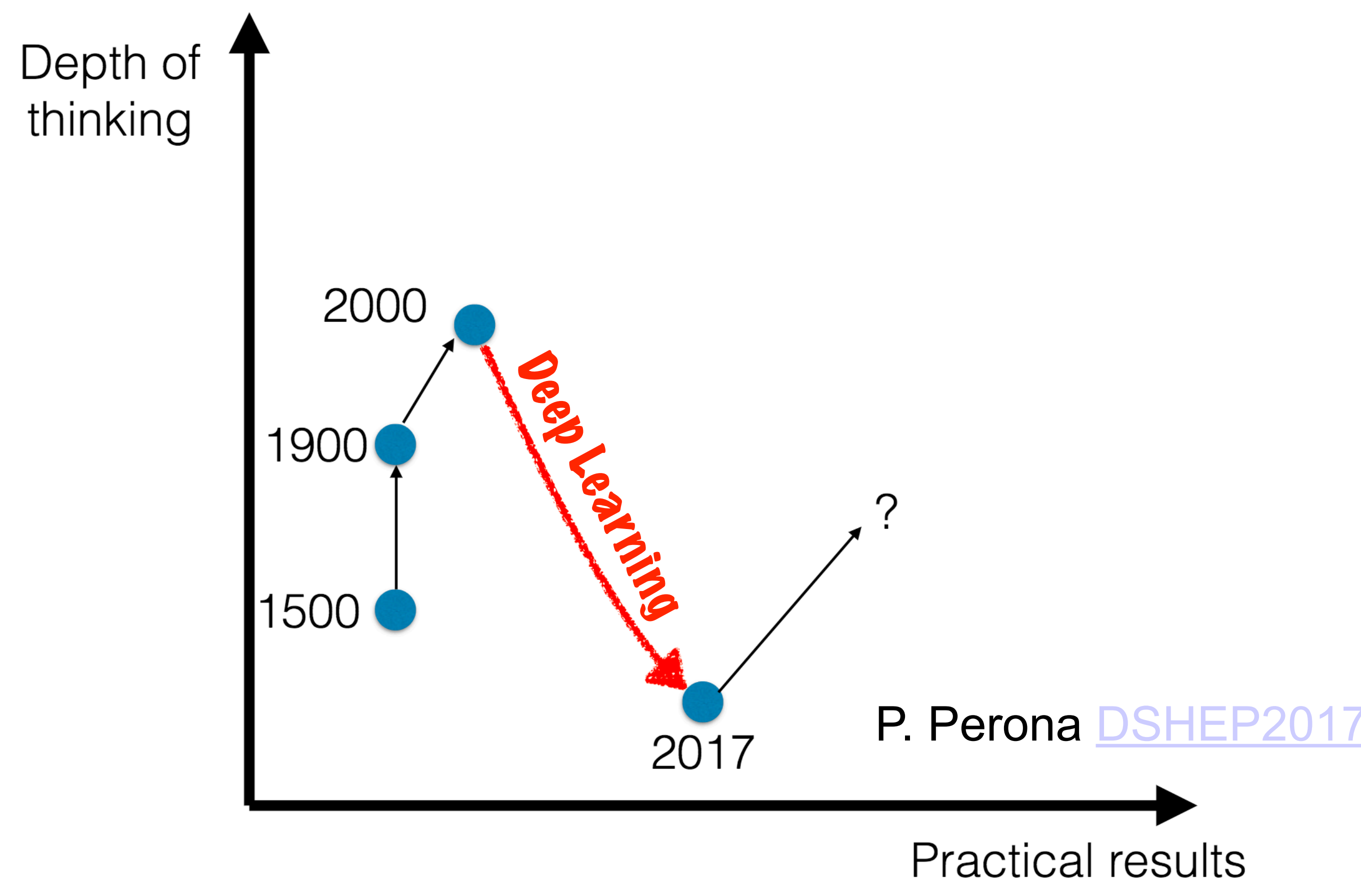
Learning from Complexity



Conv 1: Edge+Blob   Conv 3: Texture   Conv 5: Object Parts   Fc8: Object Classes

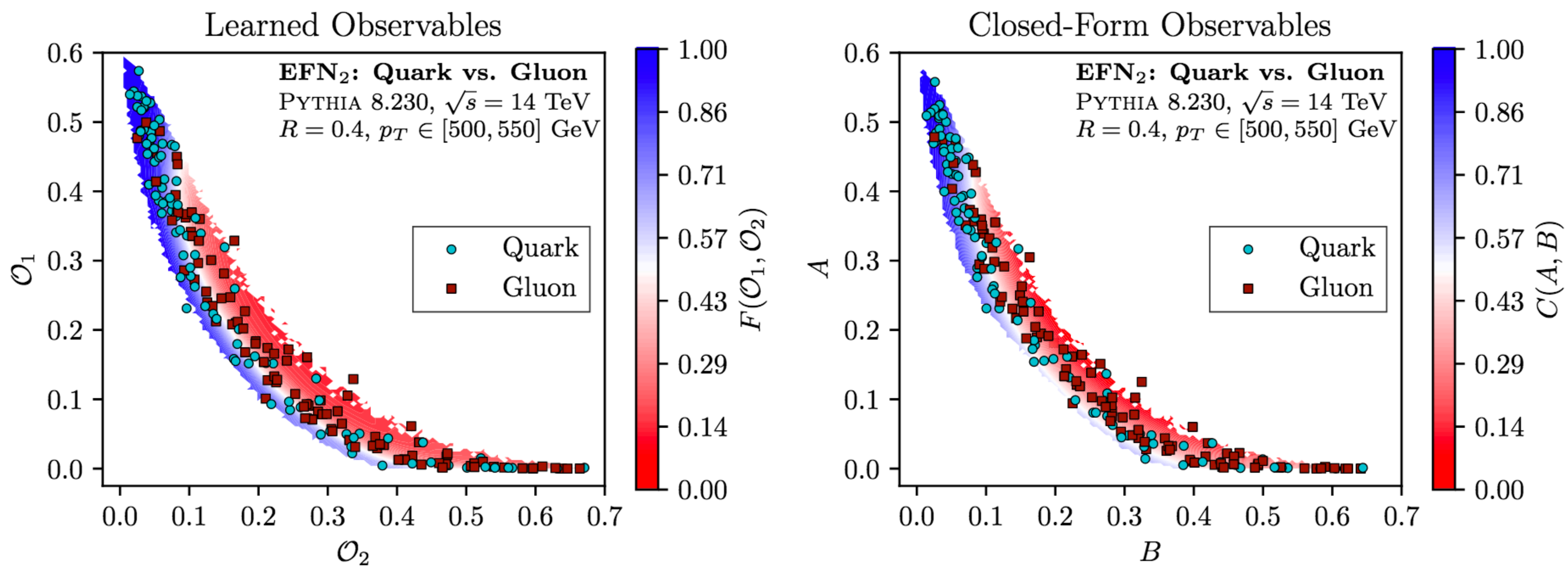
Machine learning model can **extract information from complex dataset.**
More classical algorithm counter part may
take **years of development.**

The Black-box Dilemma



Deep learning may yield great improvements.
Having the “best classification performance” is not always sufficient.
Forming an understand of the processes at play is often crucial.

Physics Knowledge



P. Komiske, E. Metodiev, J. Thaler, [1810.05165]

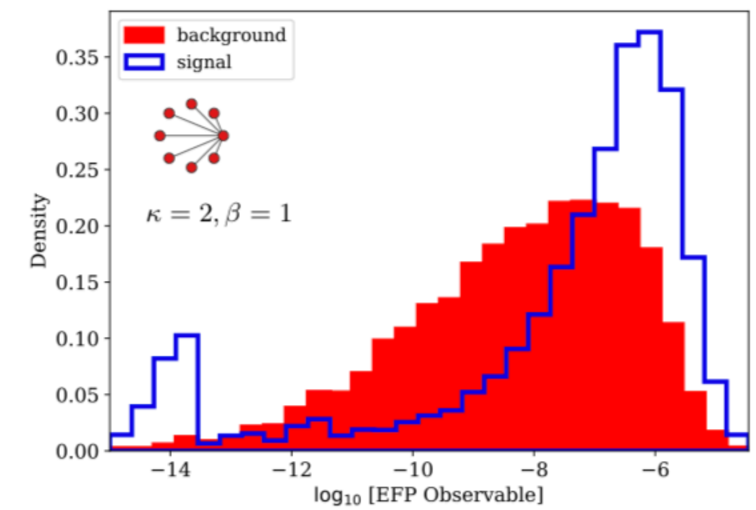
Machine Learning can **help understand Physics**.

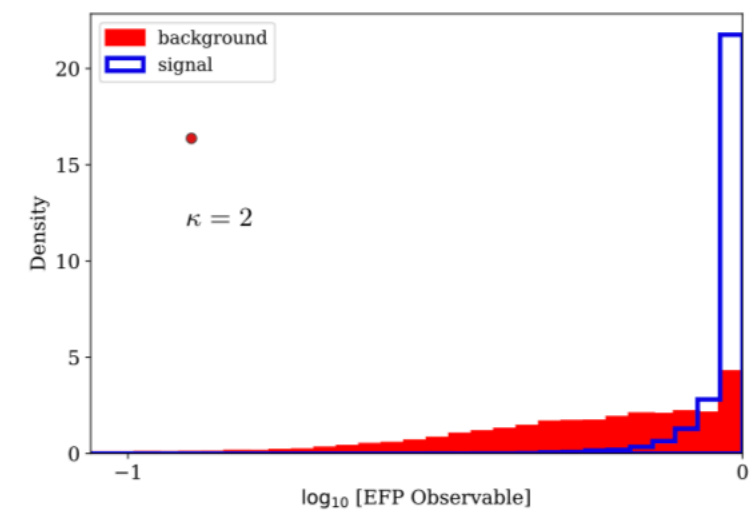
Learning Observables

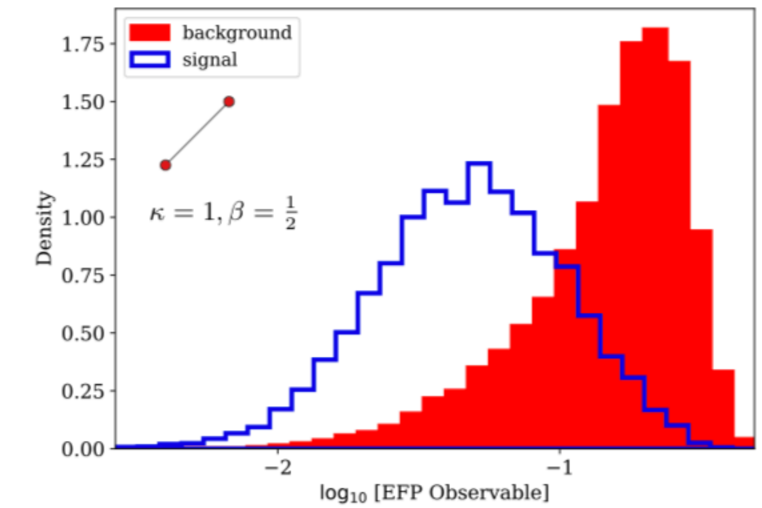
Electron classification performance

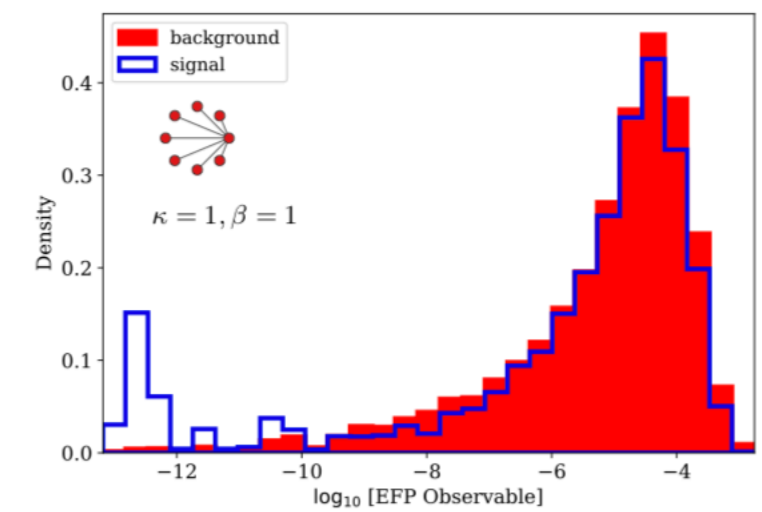
Base		Additions (κ, β)		(AUC)
7HL				0.945
7HL	$+M_{\text{jet}}$			0.956
7HL		$(1, \frac{1}{2})$		0.970
7HL	$+M_{\text{jet}}$	$(1, 1)$	$(1, \frac{1}{2})$	0.971
7HL		$(2, -)$		0.970
7HL	$+M_{\text{jet}}$	$(2, 1)$	$(2, -)$	0.971
CNN				0.972

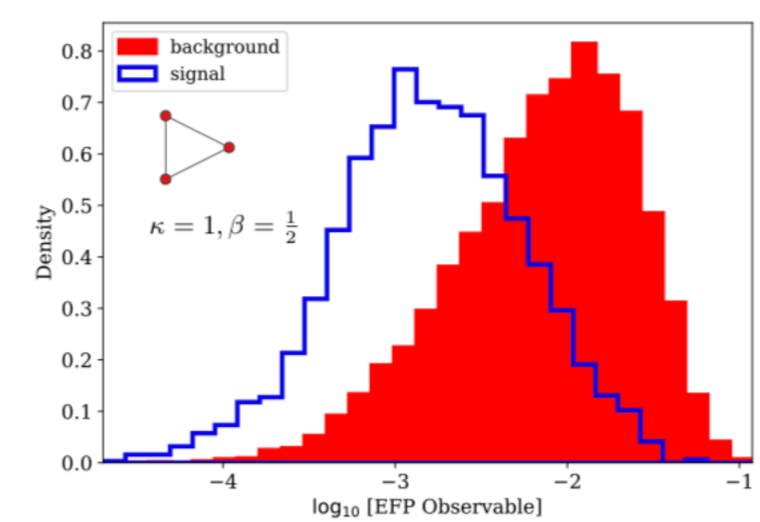
[2010.11998] [2011.01984]





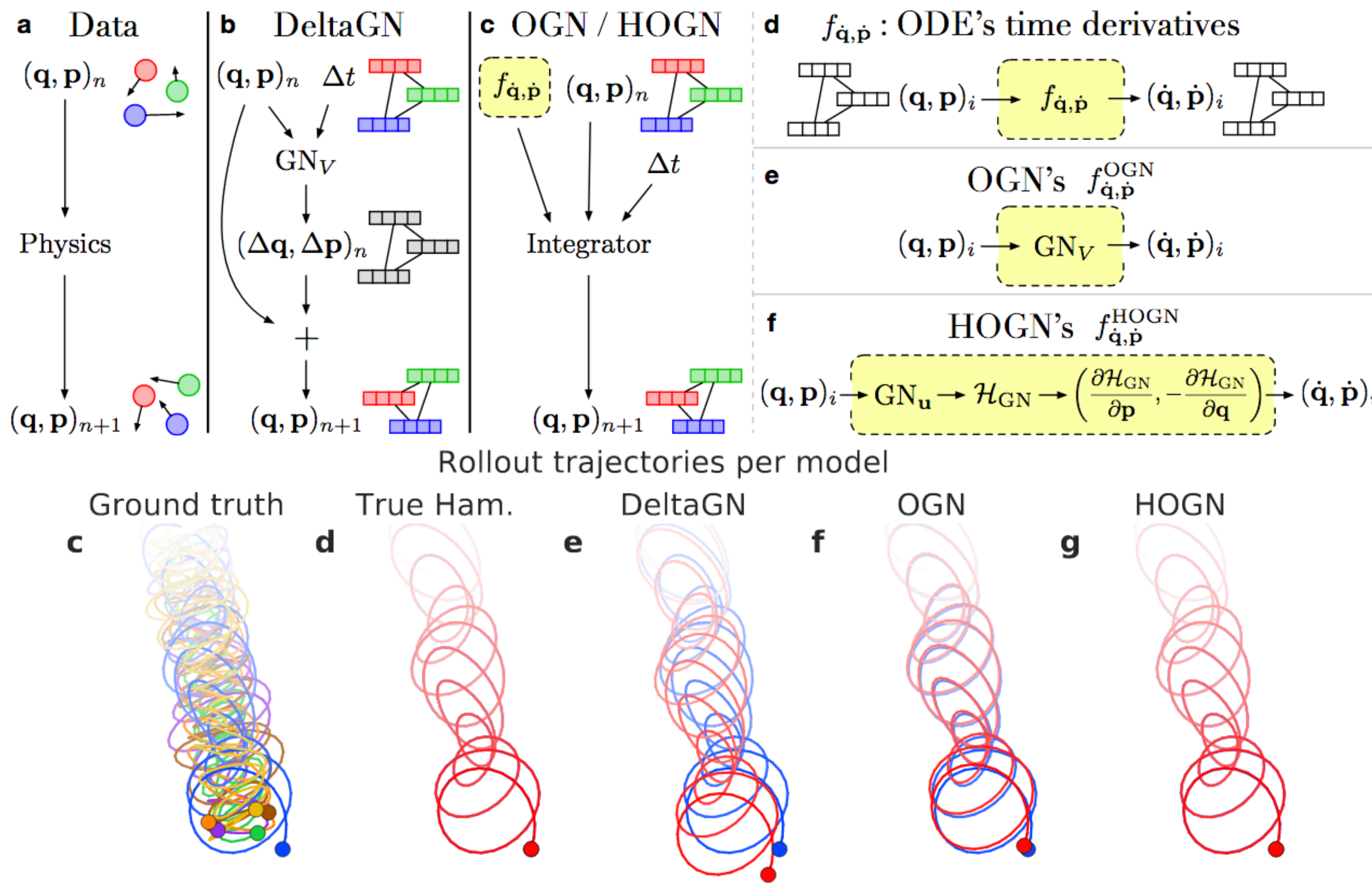






Search in the space of functions using decision ordering.
Simplified to the energy flow polynomial subspace.
Extract set of EFP that matches DNN performance.

Use Physics



A. Sanchez-Gonzalez, V. Bapst, K. Cranmer, P. Battaglia [1909.12790]

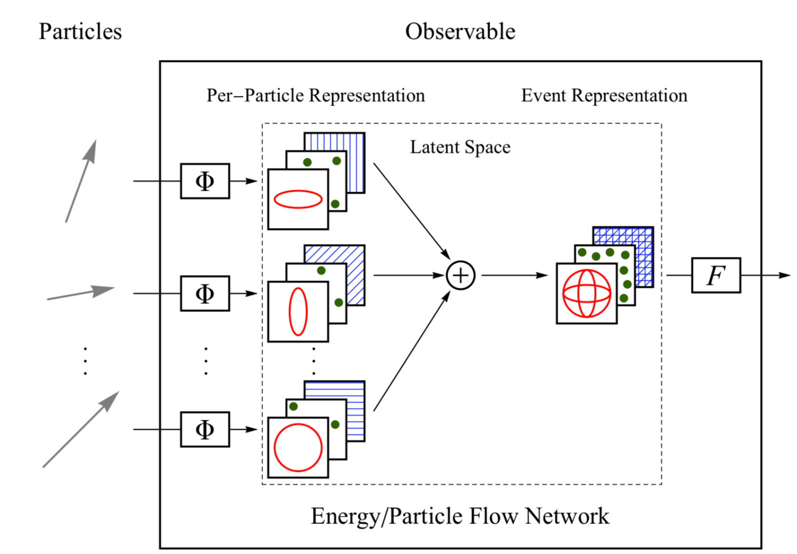
Let the model **include Physics principles** to master convergence

Inductive Bias

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \widetilde{k}_{\mu,j} = k_{\mu,i} \begin{pmatrix} 1 & 0 & \cdots & 0 & C_{1,N+2} & \cdots & C_{1,M} \\ 0 & 1 & & \vdots & C_{2,N+2} & \cdots & C_{2,M} \\ \vdots & \vdots & \ddots & 0 & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & C_{N,N+2} & \cdots & C_{N,M} \end{pmatrix}$$

$$\tilde{k}_j \xrightarrow{\text{LoLa}} \hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ w_{jm}^{(E)} E(\tilde{k}_m) \\ w_{jm}^{(d)} d_{jm}^2 \end{pmatrix}$$

Lorentz Learning Layer
[1707.08966]



Deep set
[1810.05165]

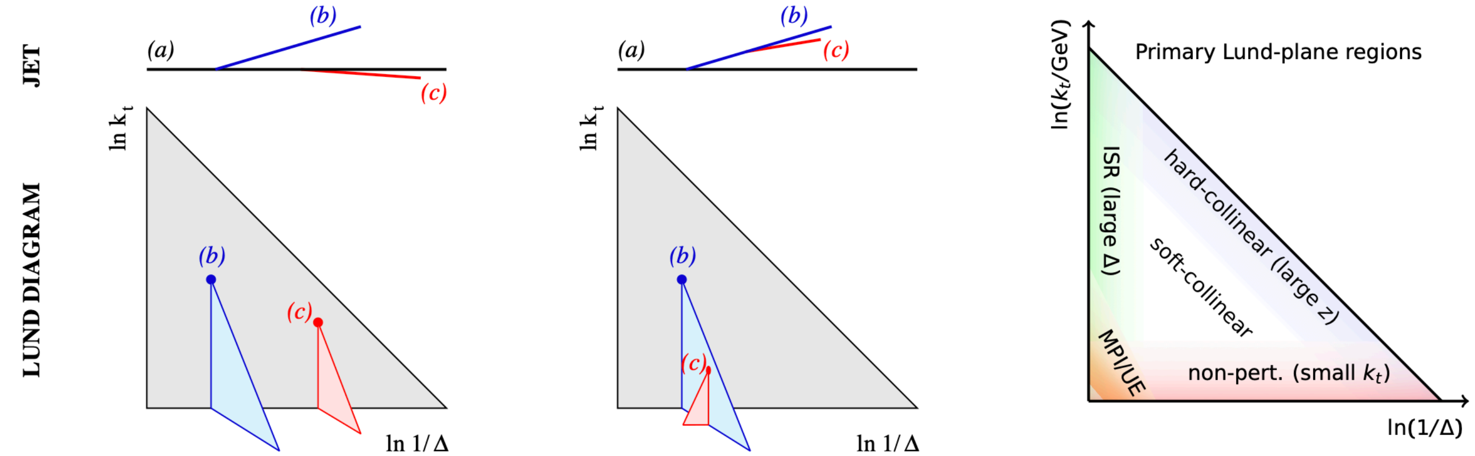
$$\mathscr{F}_i \mapsto W \cdot \left(\mathscr{F}_i \oplus \mathscr{F}_i^{\otimes 2} \oplus \sum_j f\left(p_{ij}^2\right) \cdot p_{ij} \otimes \mathscr{F}_j \right)$$

Lorentz group quivariant networks
[2006.04780]

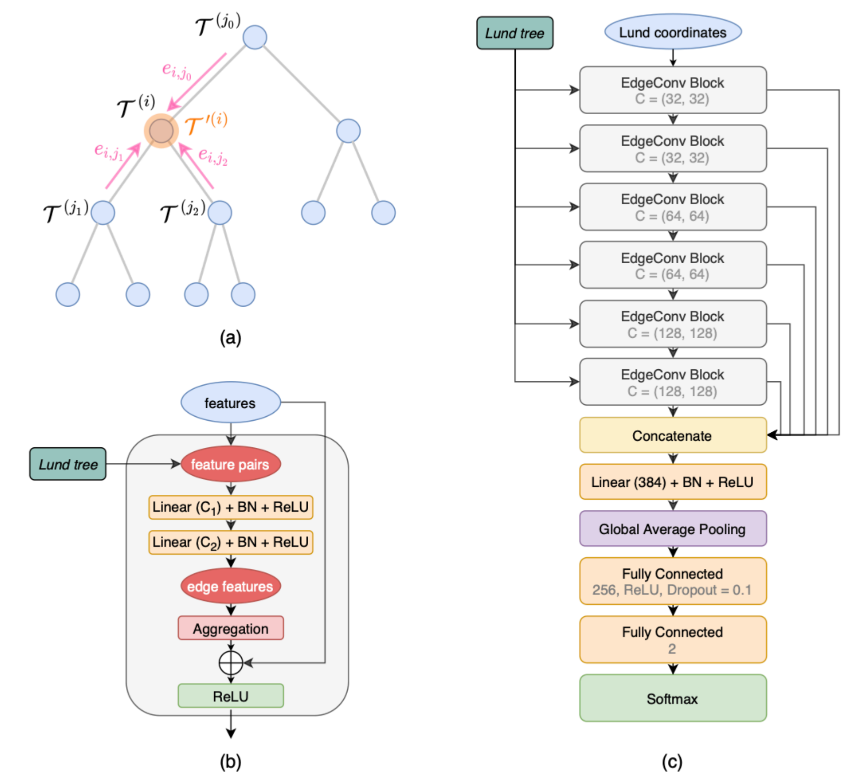
Embed the symmetry and invariance in the model.
Economy of model parameters.

Jet Tagging

The Lund jet plane provides an efficient description of the radiation patterns within a jet

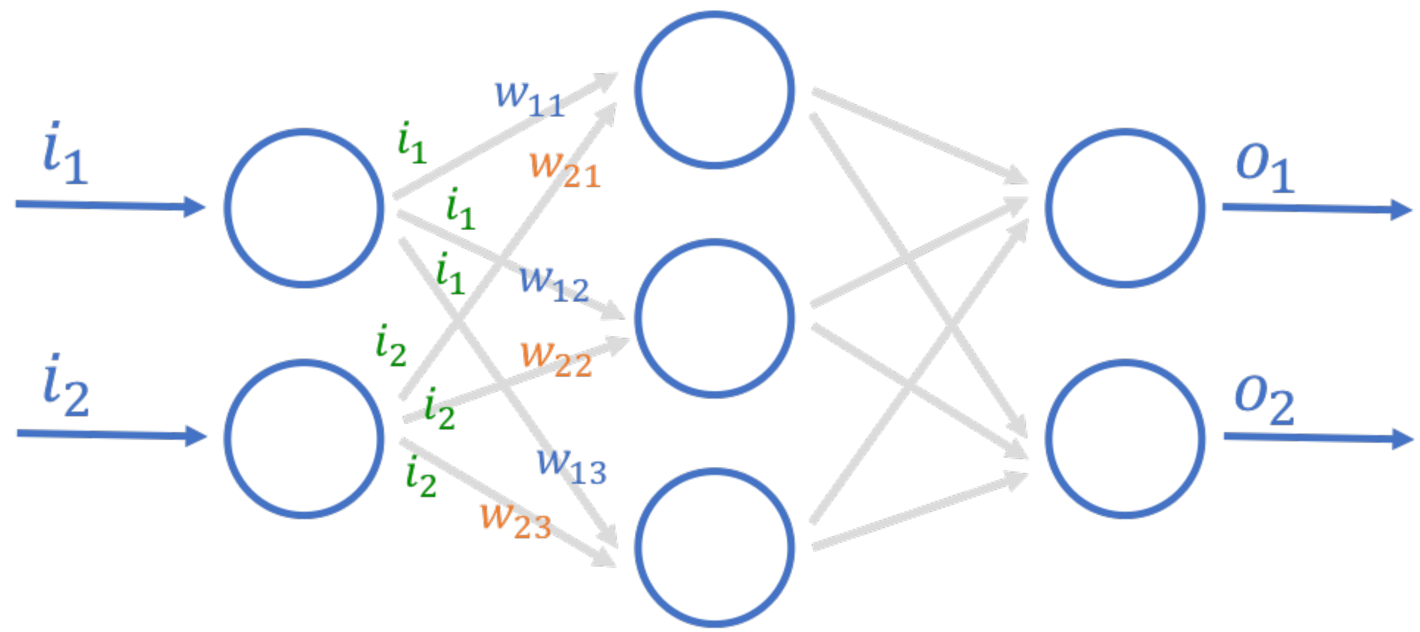


[2012.08526]



Graph-based models have recently achieved state-of-the-art jet tagging performance on benchmarks, and in analysis. Still a very rich field, in particular in developing inductive bias in the model (symmetry, invariance, …). Kinematic regression, substructure assignment, … also possible thanks to model flexibility.

Operation Vectorization

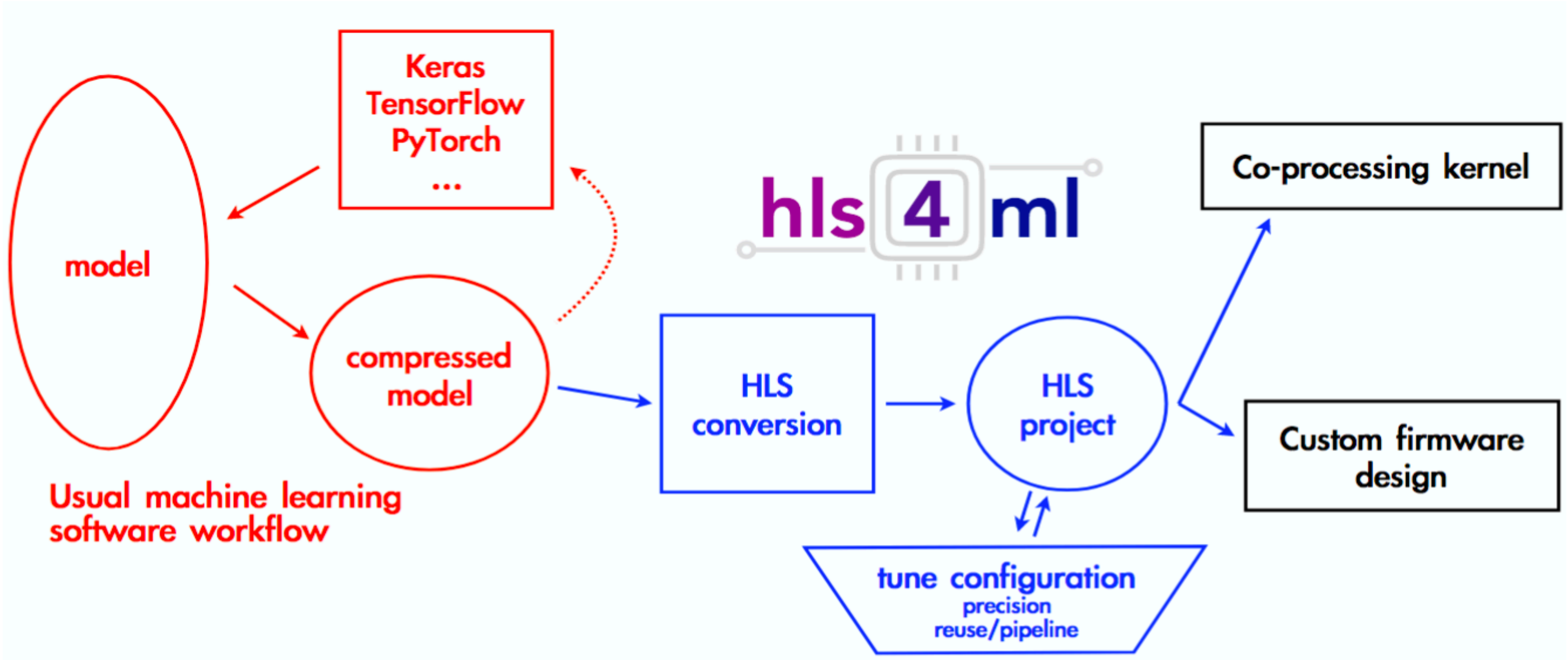


ANN ≡ matrix operations ≡ parallelizable

$$\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} (w_{11} \times i_1) + (w_{21} \times i_2) \\ (w_{12} \times i_1) + (w_{22} \times i_2) \\ (w_{13} \times i_1) + (w_{23} \times i_2) \end{bmatrix}$$

Computation of prediction from artificial neural network model can be **vectorized to a large extend.**

Hyper-Fast Prediction



Synthesizing FPGA firmware from trained ANN

https://fastmachinelearning.org/hls4ml/

J. Duarte et al.[1804.06913]

Artificial neural network model can be **executed efficiently on FPGA**, GPU, TPU, ...

Inference Engines

"On-Board accelerator"

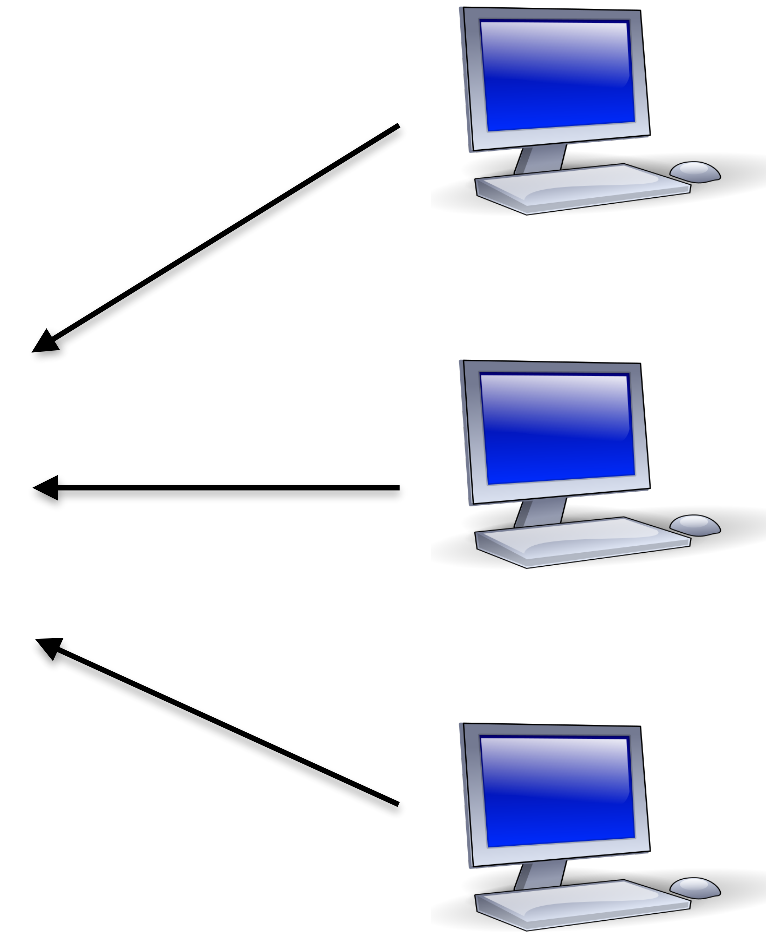
CPU
- Small models
- Small datasets
- Useful for design space exploration

GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

TPU
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

"Remote accelerator"

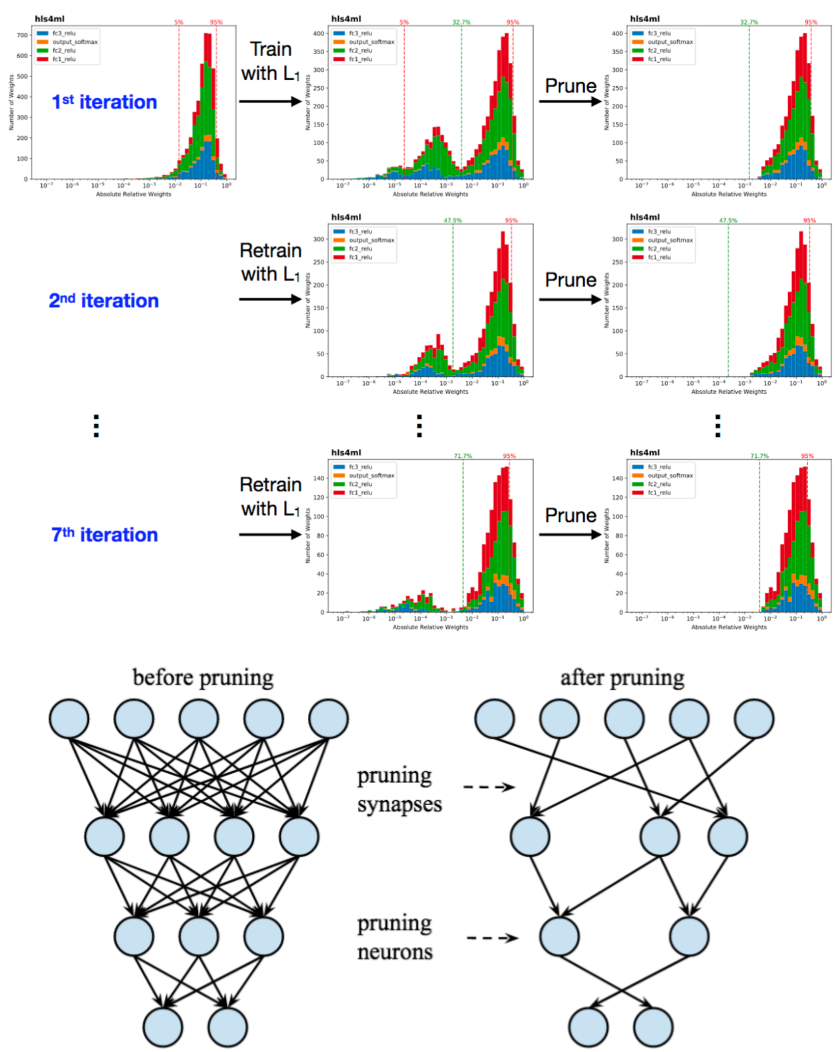


[1811.04492], [2007.10359], [2007.14781]

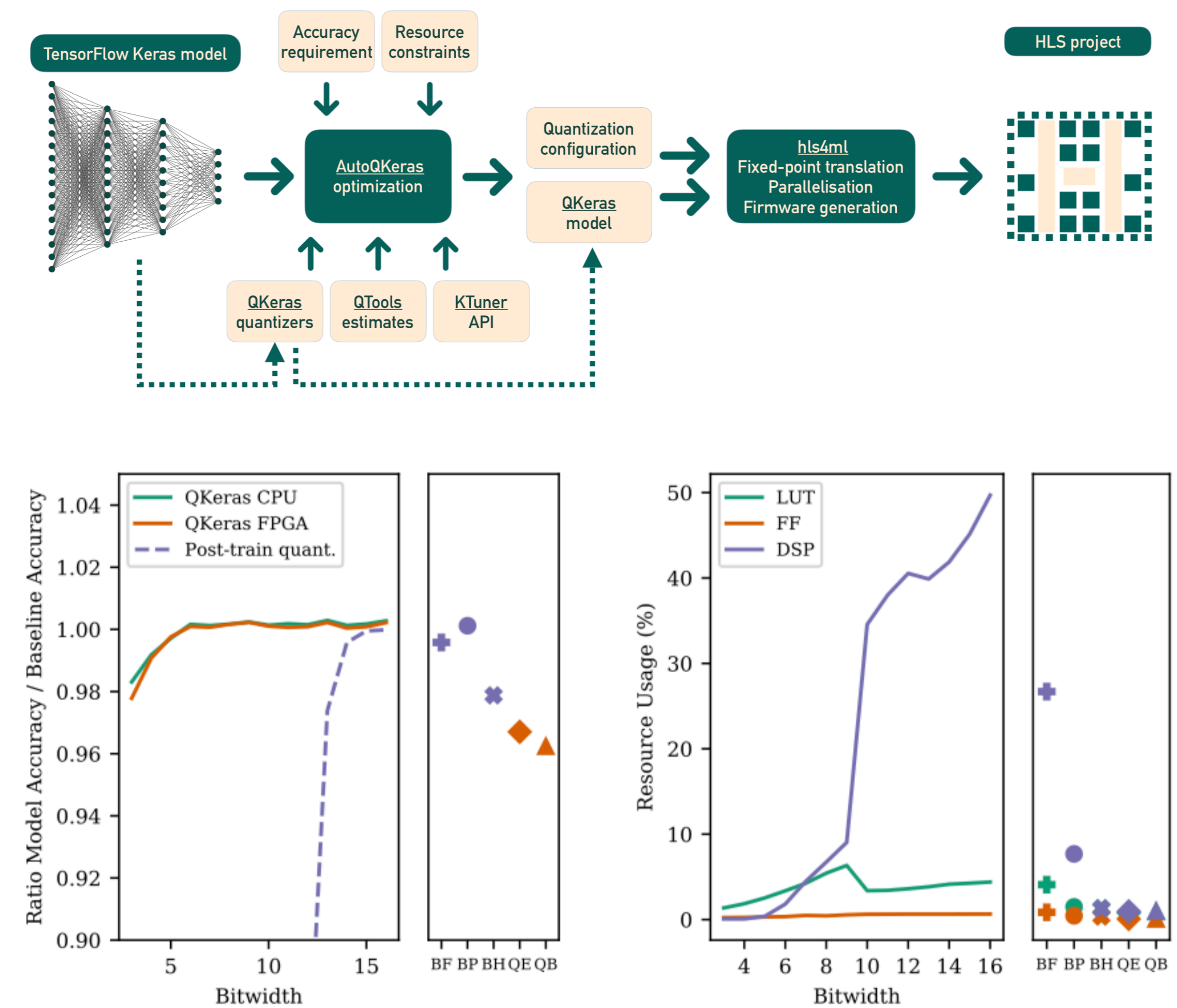
Growing list of deep learning accelerators.
Location of the device is driven by the environment (Trigger, Grid, HPC, …).

Model Compression

Fast inference of deep neural networks in FPGAs for particle physics [1804.06913]

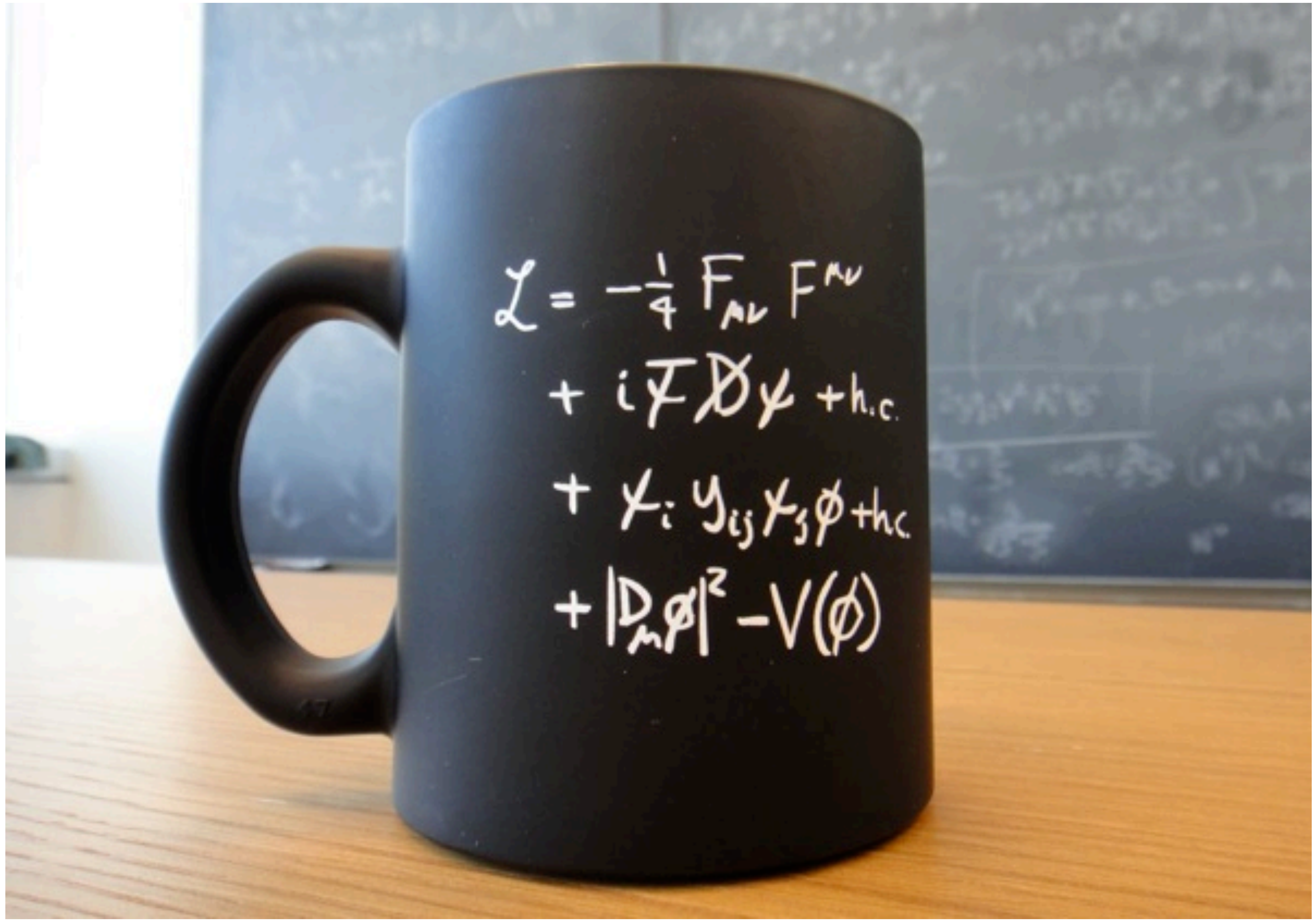


Automatic deep heterogeneous quantization of Deep Neural Networks for ultra low-area, low-latency inference on the edge at particle colliders [2006.10159]



Model inference can be accelerated by reducing the number and size of operations.

The Standard Model

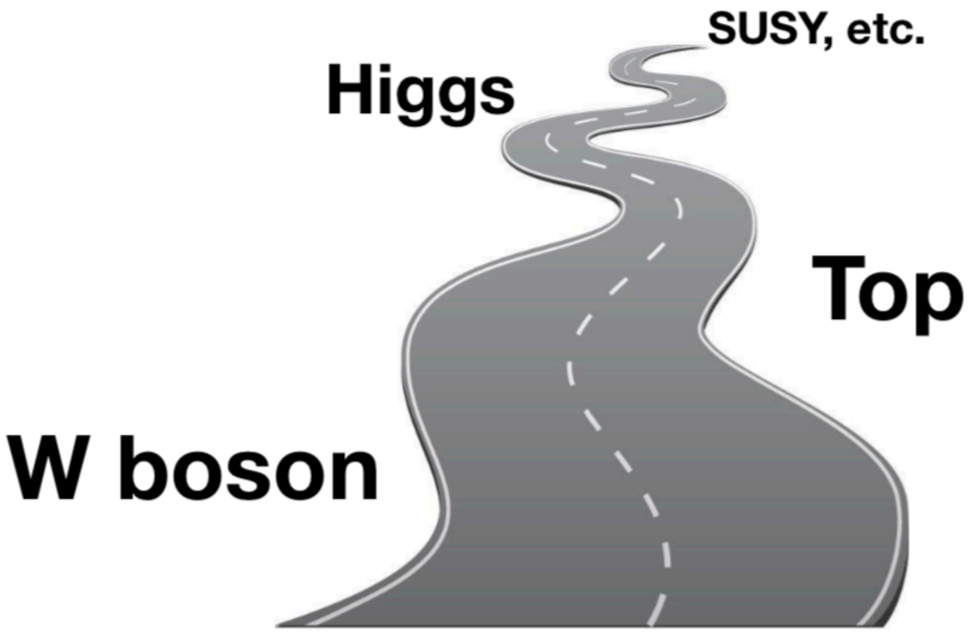


Well demonstrated effective model.
Good amount of detailed, **"labelled" simulation available**.

The Sea Beyond Standard Model

Slide: A. Wulzner [H&N]

HEP yesterday



"Almost" **Simple H_1**

Focus on **few sharply-defined** alternative models (e.g., the Higgs)

Case-by-case design of **optimal test**

HEP today

"Very" **Composite H_1**

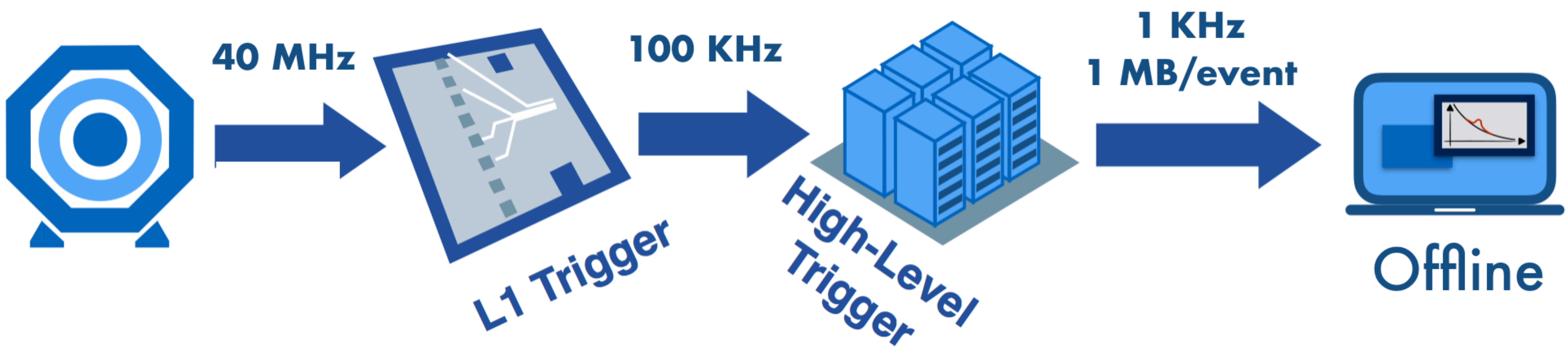
Huge set of alternatives

Case-by-case optimisation **unfeasible**

The **right H_1** likely **not yet formulated**

Event Triggering

Select what is important to keep for analysis.
Ultra fast decision in hardware and software.

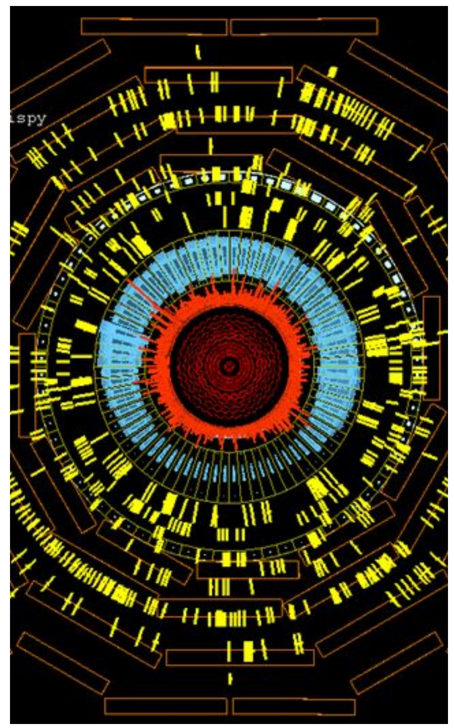


Reconstruction of the event under limited latency / bandwidth.
Better resolution help lowering background trigger rates,
Faster algorithms helps making more refined decisions.

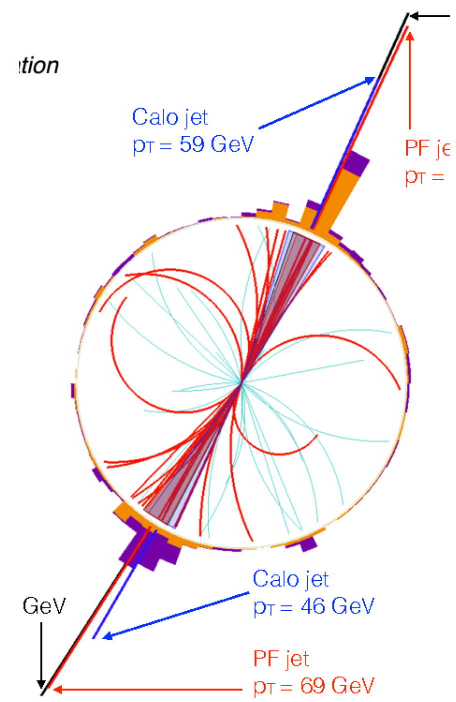
Reconstructing Collisions

Detector Data

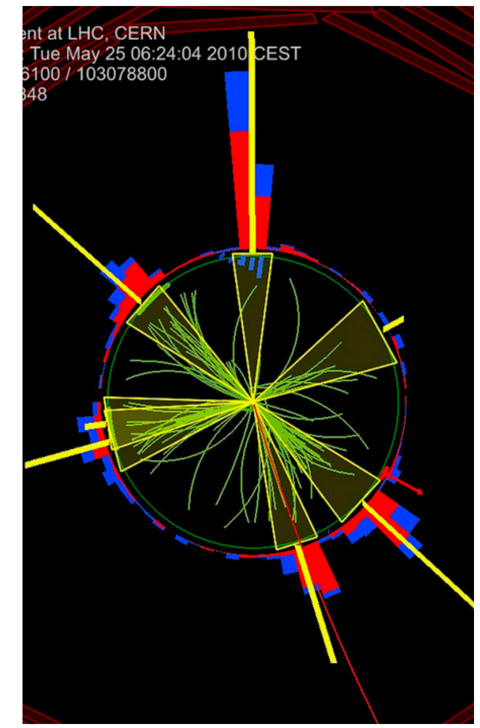
Local reconstruction



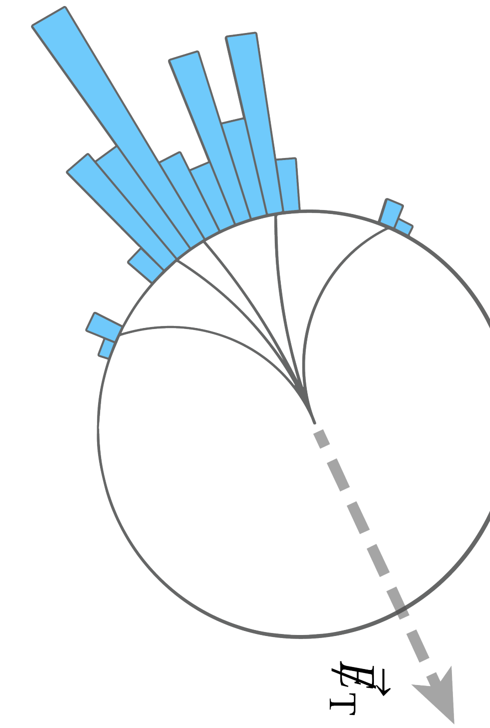
Particle representation



Jet Clustering



High level features



Event Processing

Dimensionality reduction

Globalization of information

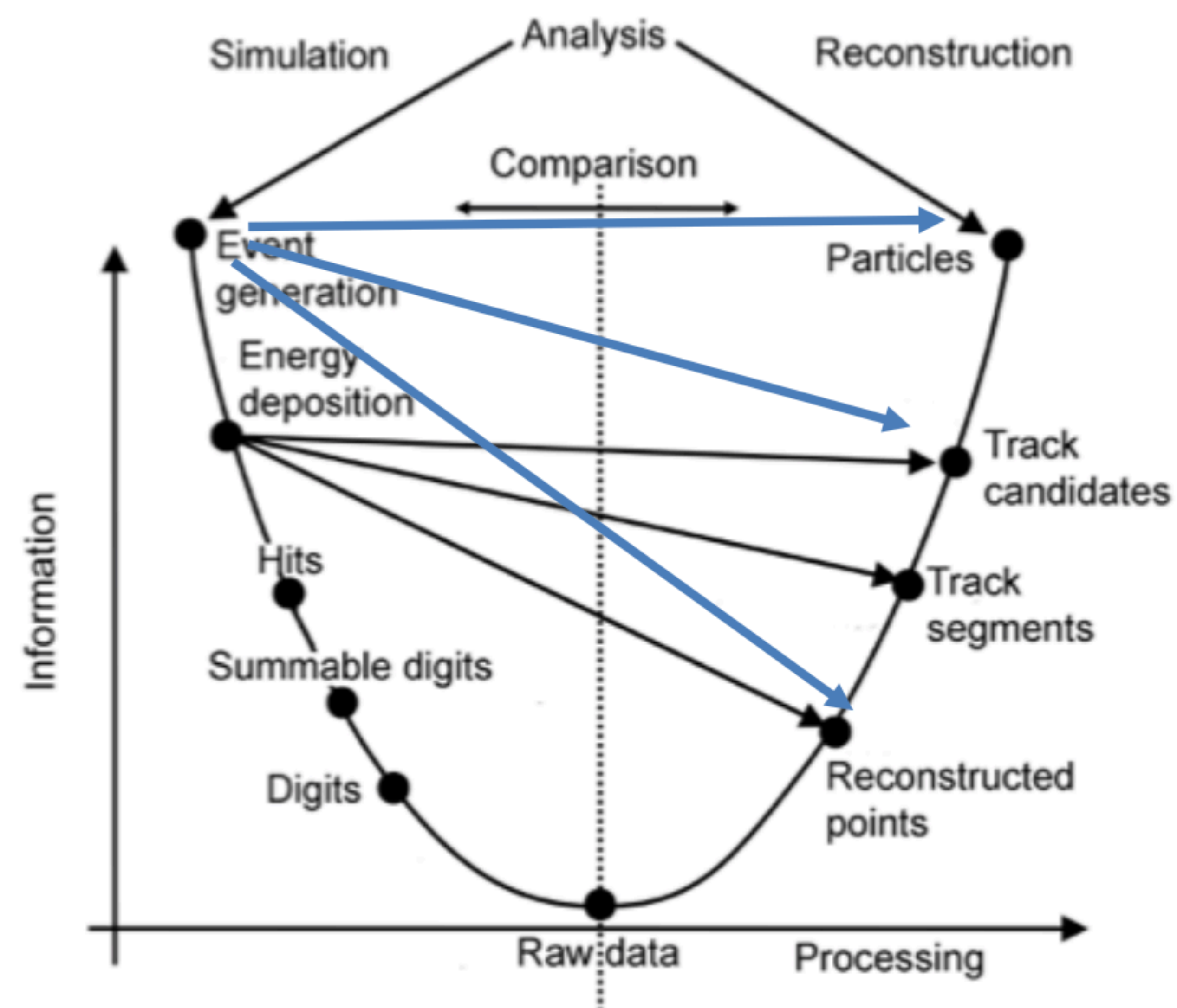
From detector signal to high-level features using **mostly pattern recognition**.
Complex and **computing intensive** series of tasks.

Simulating Collisions

Madgraph, Pythia, Sherpa, ...	**Event Generator**: compute predictions of the standard models to several orders of expansion in coupling constants (LO, NLO, NNLO, ...) using proton density functions.
Pythia, ...	**Hadronization**: phenomenological model of the evolution of hadrons under the effect of QCD.
GEANT 4, GEANT V	**Material simulator**: transports all particles throughout meters of detector, using high resolution geometrical description of the materials.
Homegrown software	**Electronic emulator:** converts simulated energy deposits in sensitive material, into the expected electronic signal, including noise from the detector.

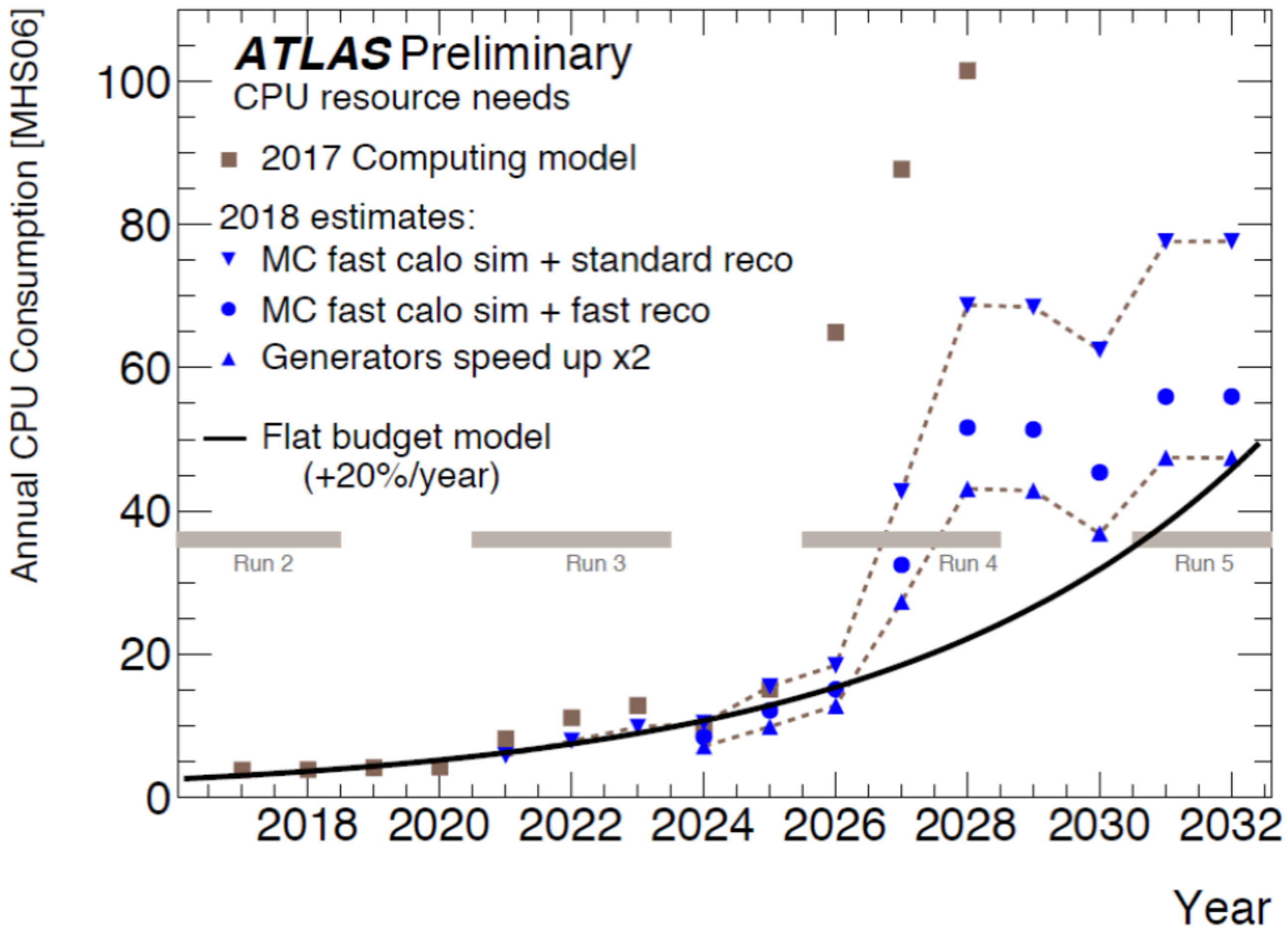
Non-differentiable, **computing intensive** sequence of **complex simulators** of the signal expected from the detectors.

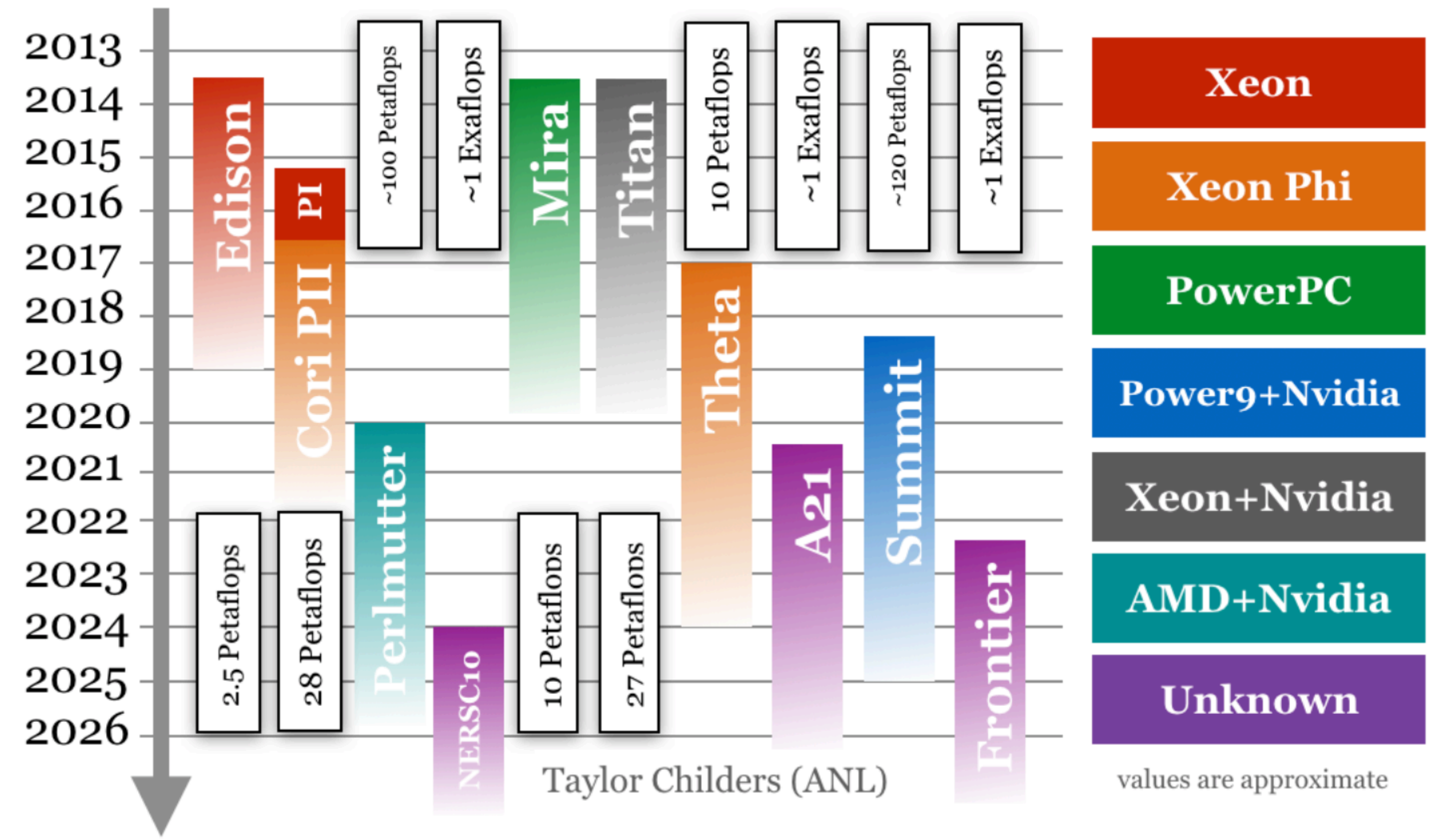
Reconstruction ∘ Simulation ~ Identity



Simulation aims at predicting the outcome of collisions.
Reconstruction aims at inverting it.
Multiple ways to connect intermediate steps with deep learning.

The Computing Cost of Science

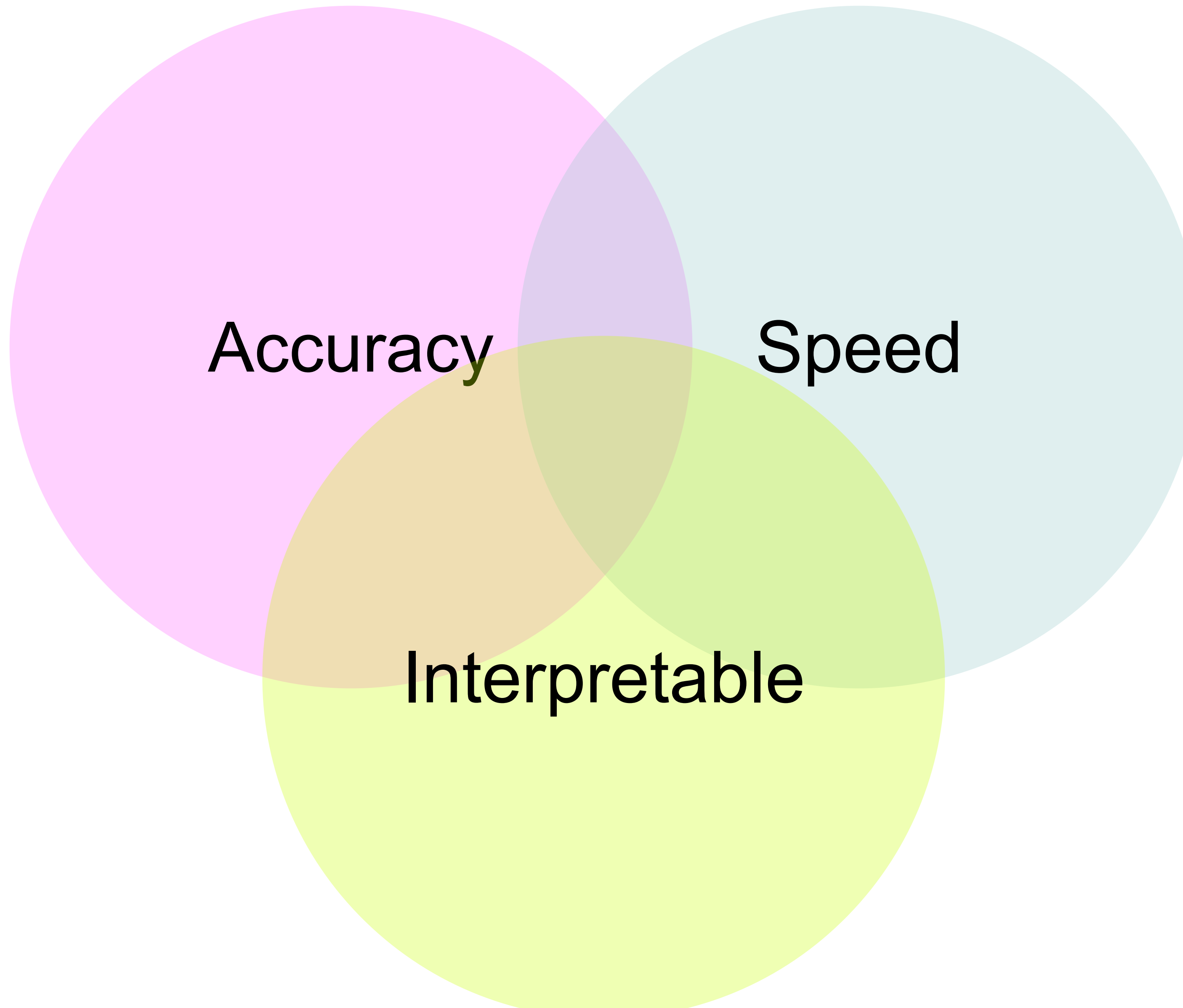




https://indico.cern.ch/event/822126/contributions/3500169/

Ever growing needs for computing resource.
Slowdown of classical architecture, over growth of GPU architecture.

Possible Utilizations



- **Fast surrogate** models (trigger, simulation, etc) ; even better if more accurate.
- **More accurate** than existing algorithms (tagging, regression, etc) ; even better if faster.
- Model performing **otherwise impossible tasks** (operations, etc)