Machine Learning @ ATLAS

Tobias Golling, University of Geneva,

On behalf of the ATLAS Collaboration

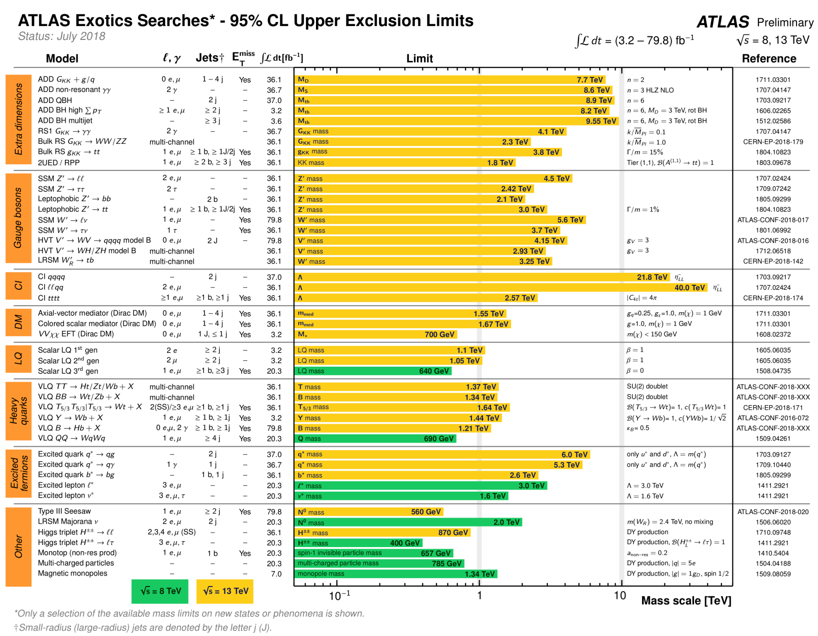
Game of Flavours, Dubrovnik, May 3 2019

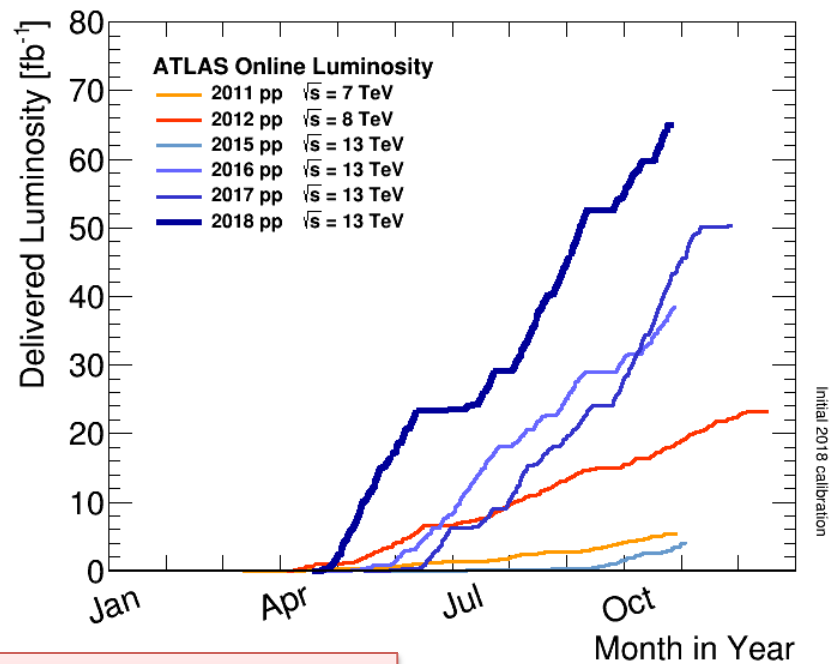
Overview

- *Representative* selection of ML@ATLAS
- Classification & related topics
  - Will not have time to discuss regression tasks
- Clustering & data structures
- Generative models

LHC interim evaluation

- Physics beyond the SM is **not around the corner**
- **Slow-growth era** of LHC has started: energy & luminosity
- **How to make rapid progress now?**

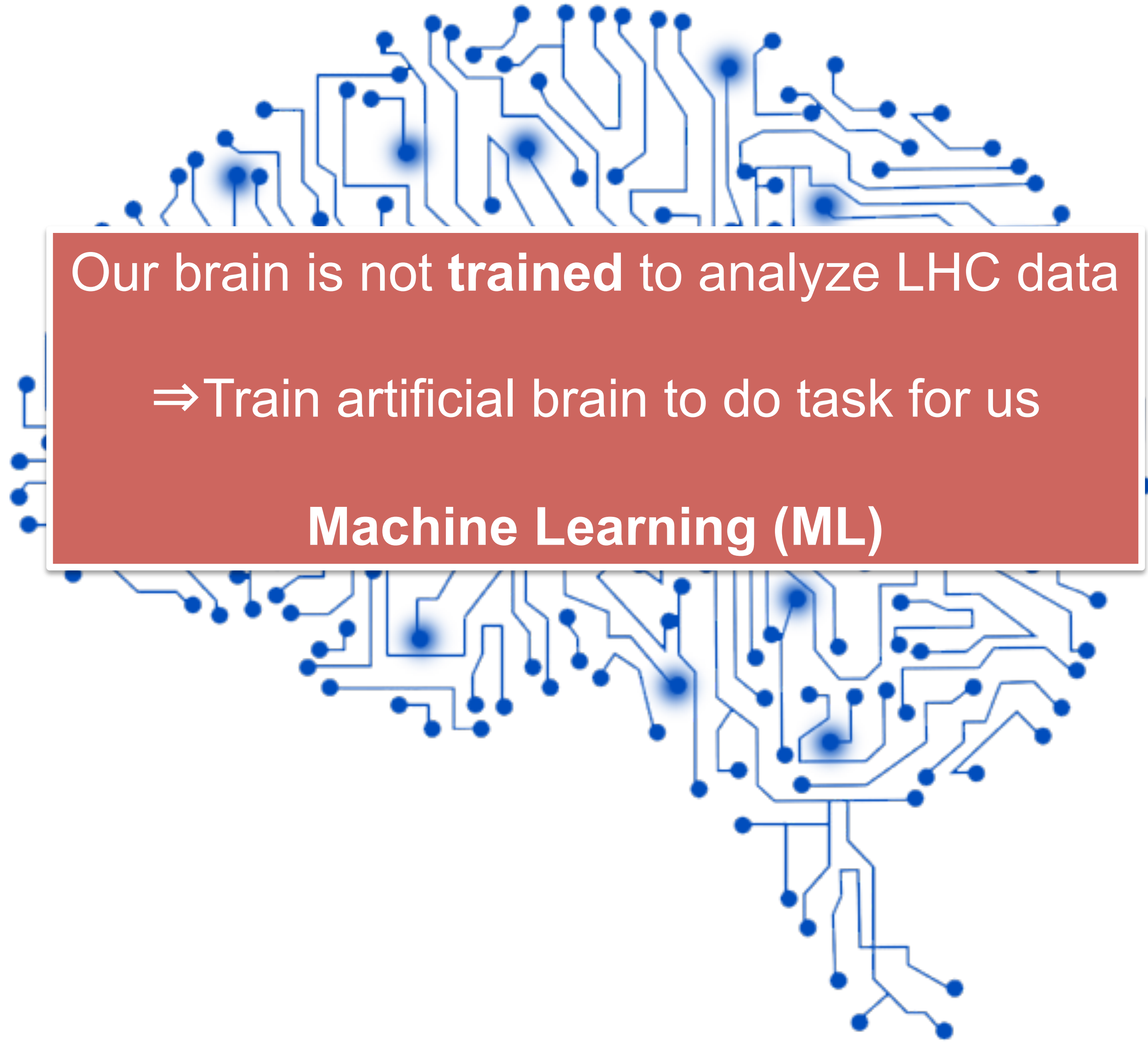




Opportunity !

Turning crank → innovation

In a nutshell



Our brain is not **trained** to analyze LHC data

⇒Train artificial brain to do task for us

Machine Learning (ML)

What is ML?

- Inspired by how the brain works
- Learning from examples
- Condensing information to “knowledge”

[There are other ways to define ML]

How can ML help?

- Low hanging fruit
  - Better
  - Faster
  - Easier / automated

- More profound changes to how we approach physics ?!

Uniqueness of HEP data for ML

- Simulation can produce **highly valuable labeled training data** for supervised learning
- We have a **theory model (SM)**
  - How to inject our domain knowledge into ML
- **Systematic uncertainties**
- HEP not only costumer but also *driver* of ML!

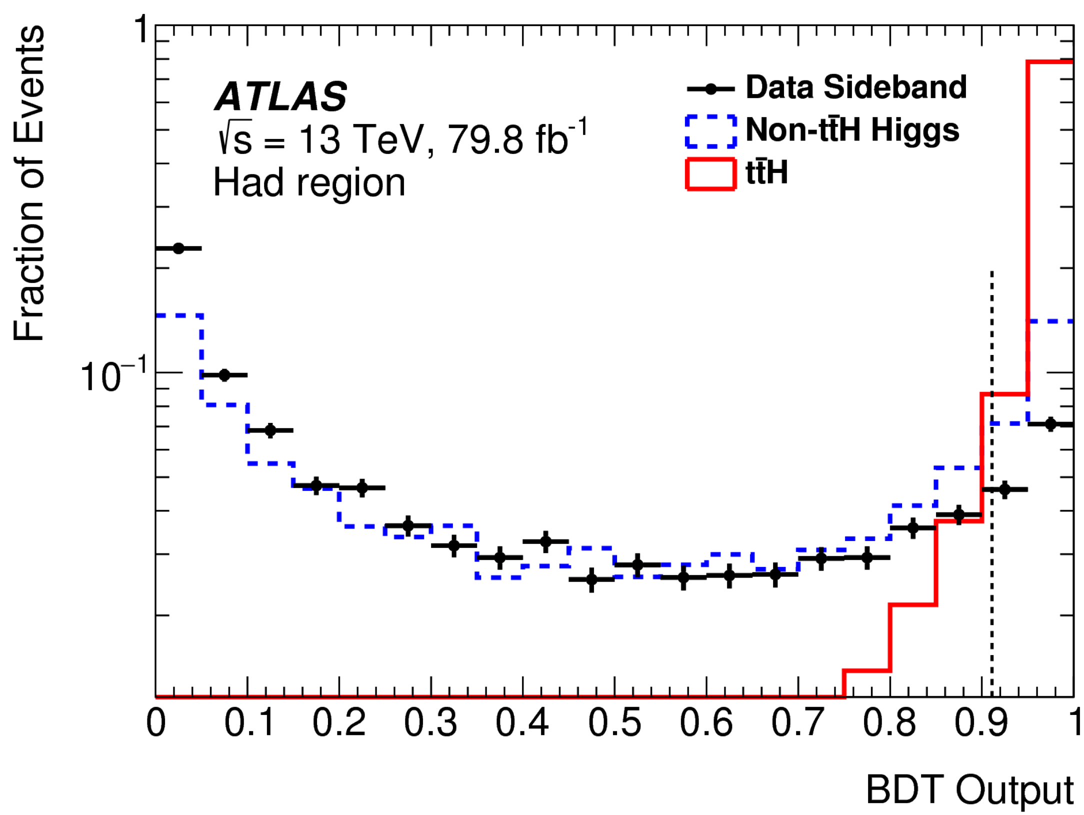
Classification with Machine Learning (better)

Classification task: eatable or not?

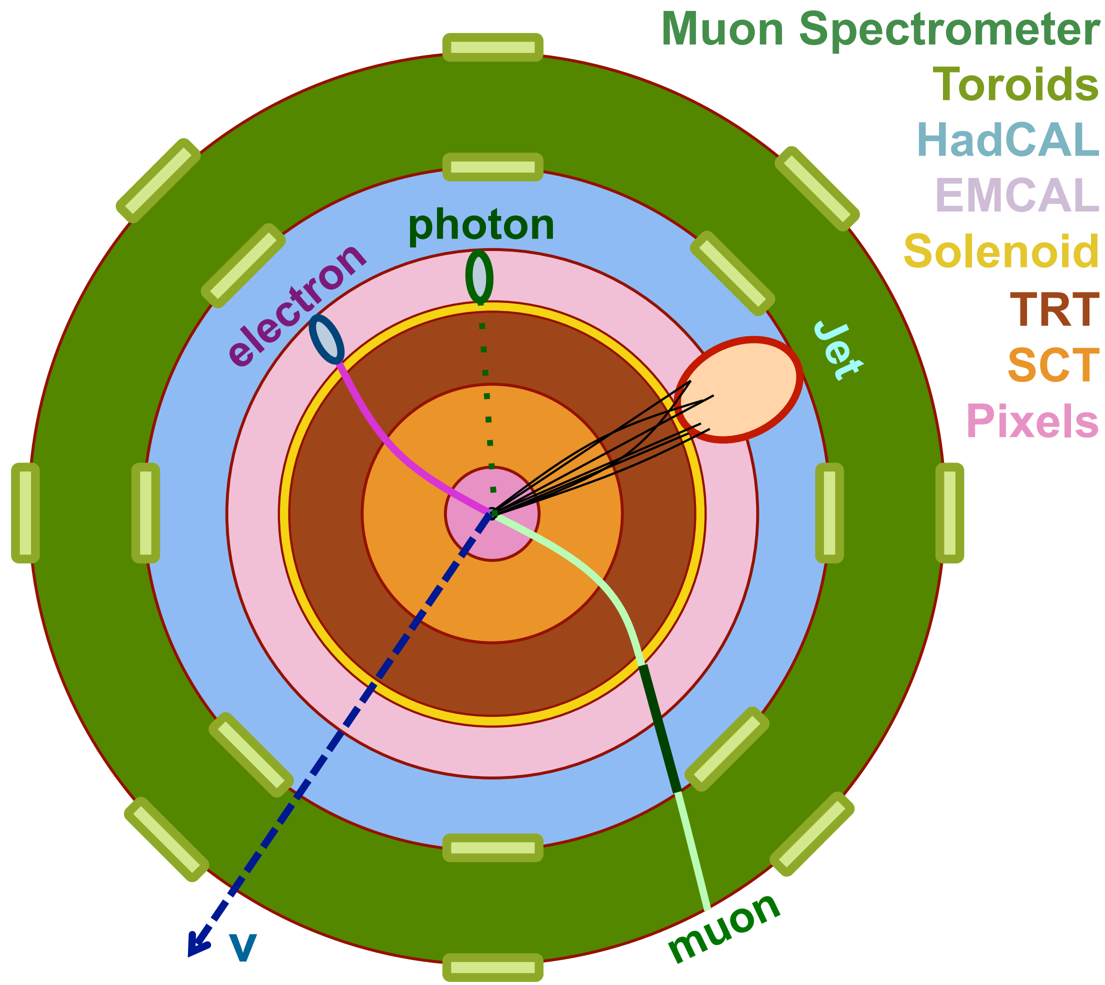
Event-level discrimination

- We've used ML for decades
- Recent example: ttH discovery
- Human-engineered features (here 38 input variables)
- Many more examples exist

[1806.00425]



Ideal test ground: physics object classification



- Large statistics
- Excellent modeling
- Good return/effort
- Validate in Control Region

Flavor tagging algorithms

1. Algorithm based on **track properties**

RNNIP

2. Algorithm based on **leptons**

3. Algorithm based on **secondary vertices** (SV)

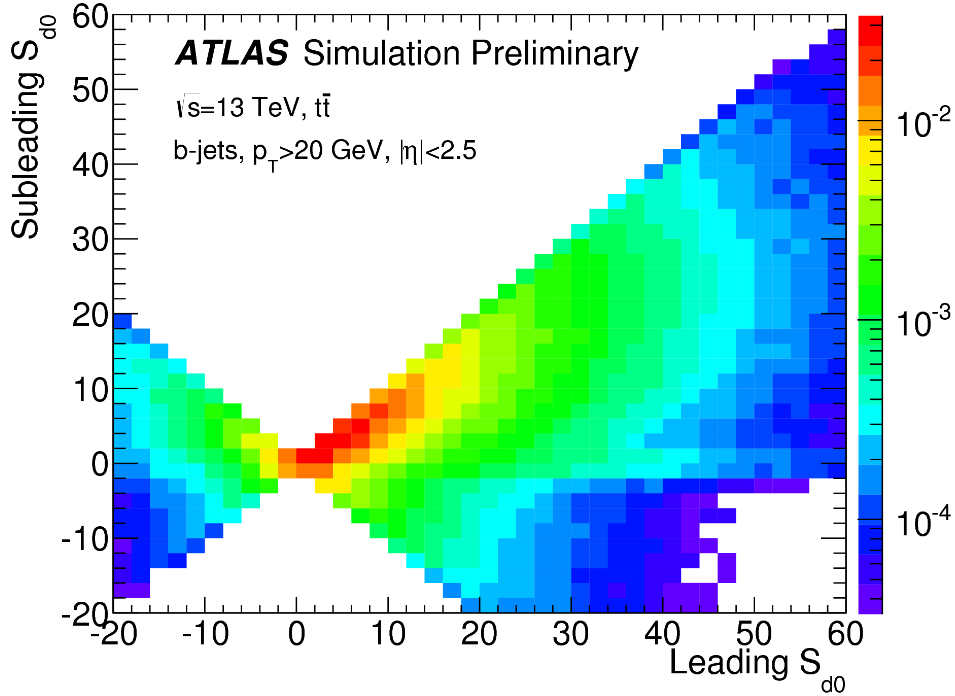
DL1

Algorithm combining all information

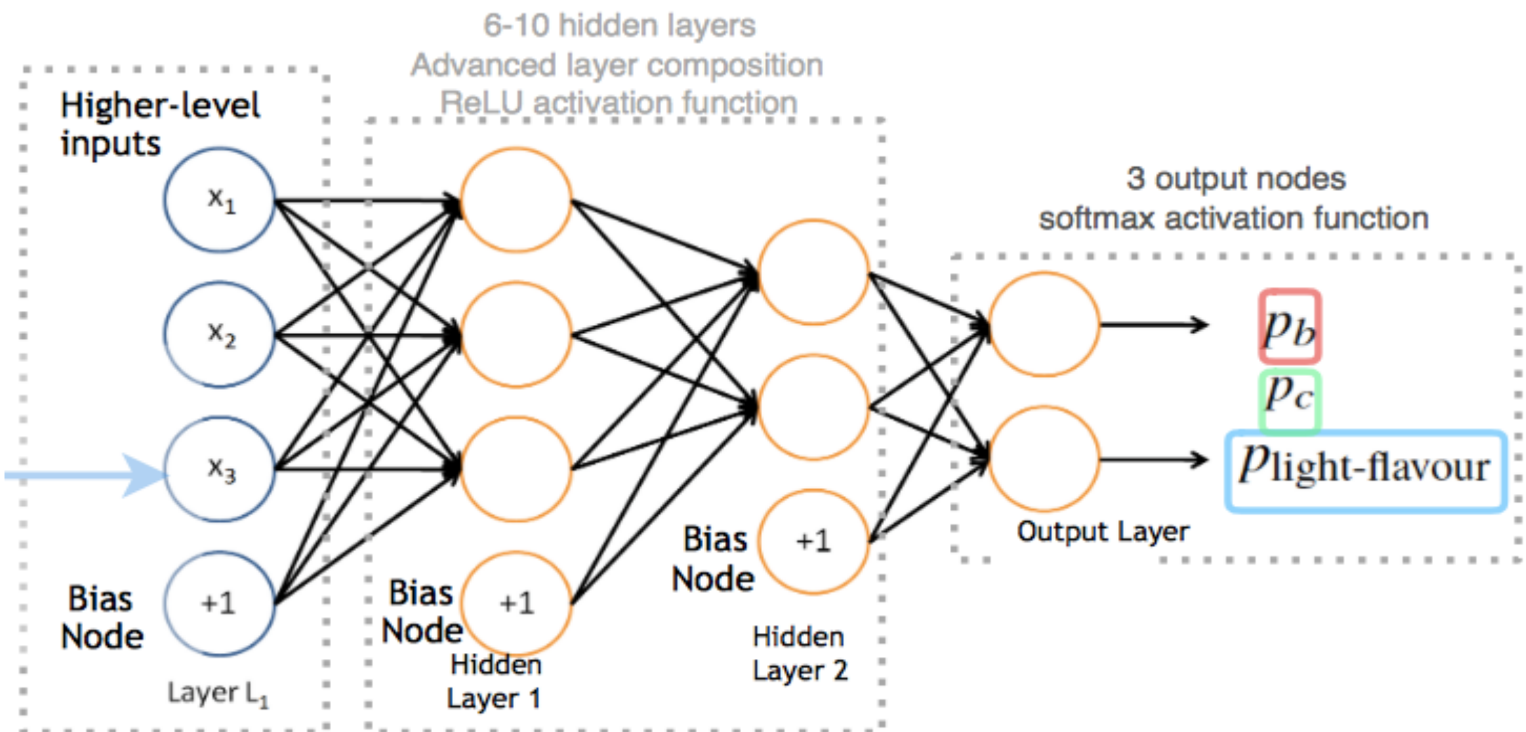
Recurrent NN: RNNIP

- Inputs: track properties of **arbitrary length**
- **Tracks as series**, ordered e.g. by d0 significance
- Exploits **track correlations**
  - Long Short-Term Memory (LSTM) used to preserve memory and combat vanishing gradient problems

[ATL-PHYS-PUB-2017-003]



Deep Learning: DL1



[ATL-PHYS-PUB-2017-013]

- Trained using MC truth labels
- **Multi-class output** (easily extendable to more classes)
- **Flexibility**: one training for all OP for b- & c-tagging

[Parenthesis: training challenge]

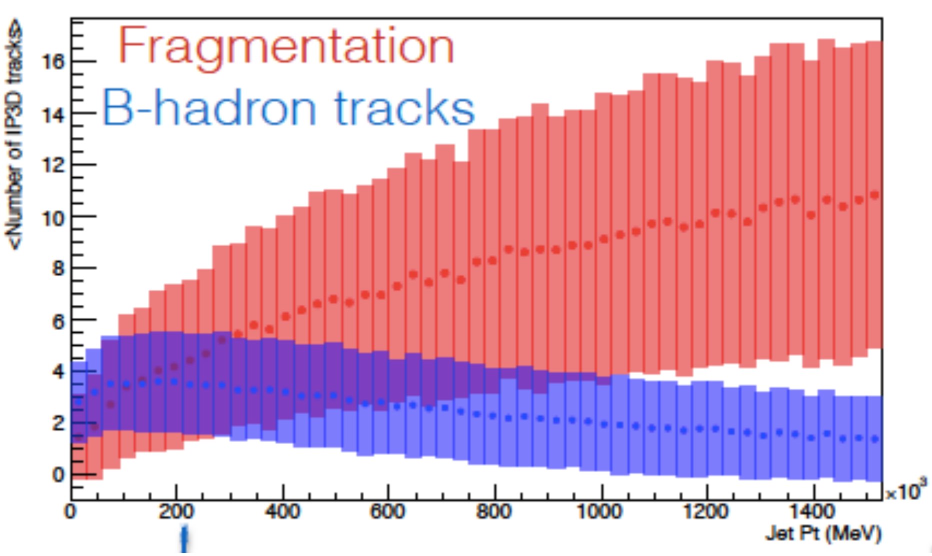
- How to find optimal hyperparameters
  - Brute force: grid search

- No off-the-shelf solution
- Toolsets exist, but no *instructions/theory*

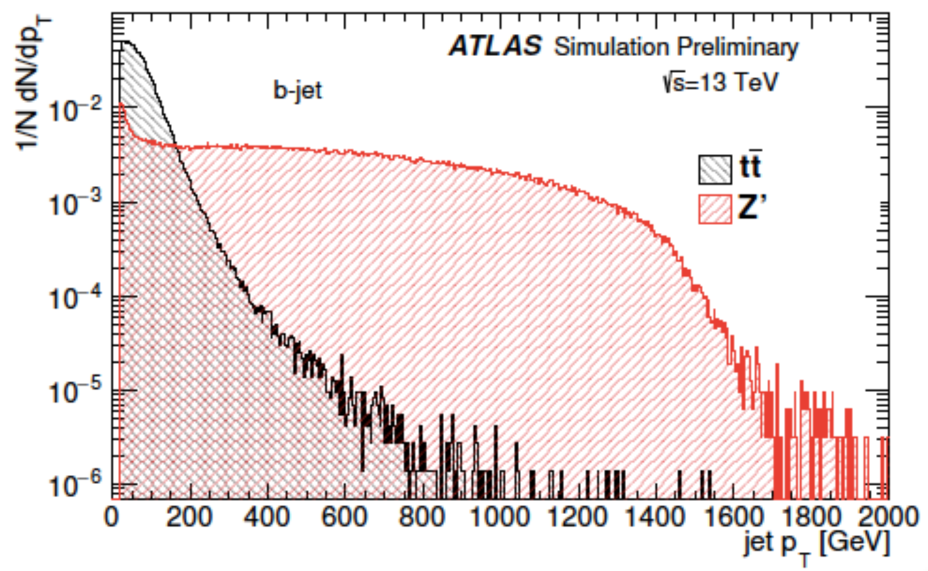
Improve Flavor tagging at highest p_T

Provide algorithm with adequate training statistics at high p_T:
Use $Z' \to b\bar{b}/c\bar{c}/q\bar{q}$ (made ~flat in p_T) instead of $t\bar{t}$

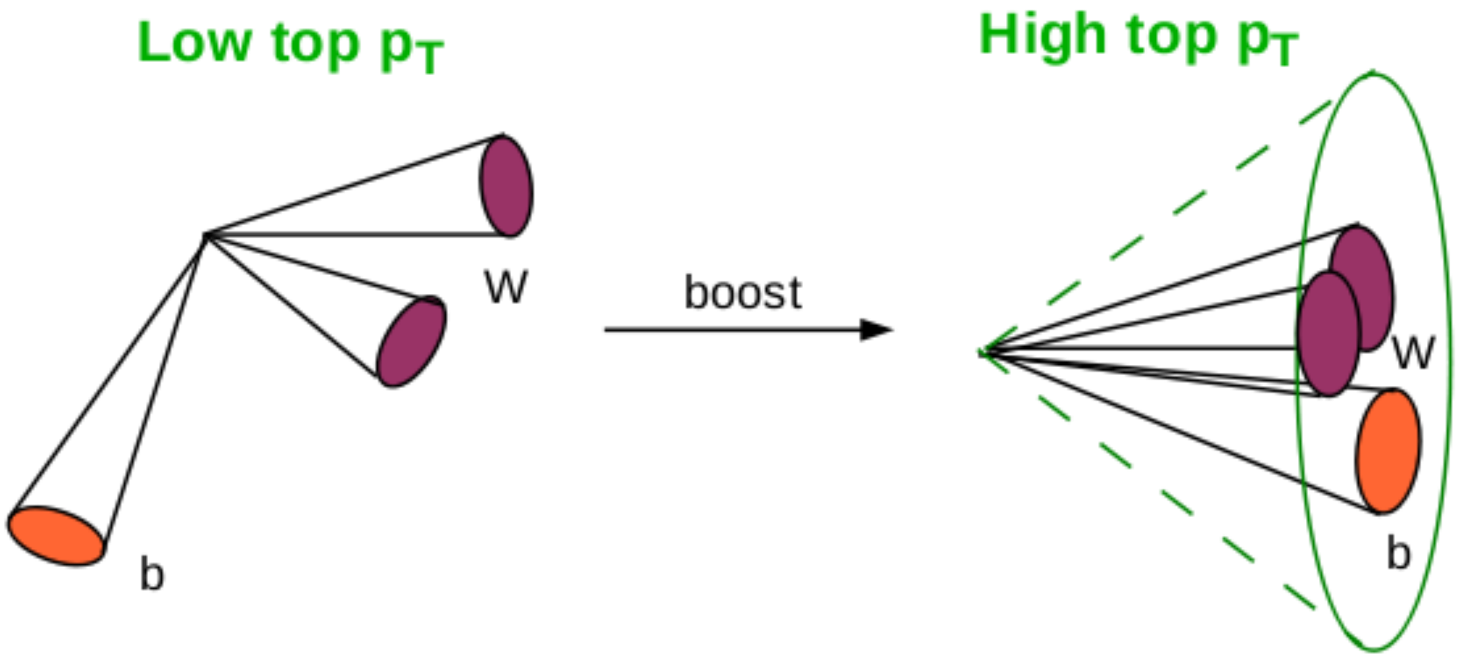
Flavor tagging very challenging at high p_T



[ATL-PHYS-PUB-2017-013]



Boosted object tagging

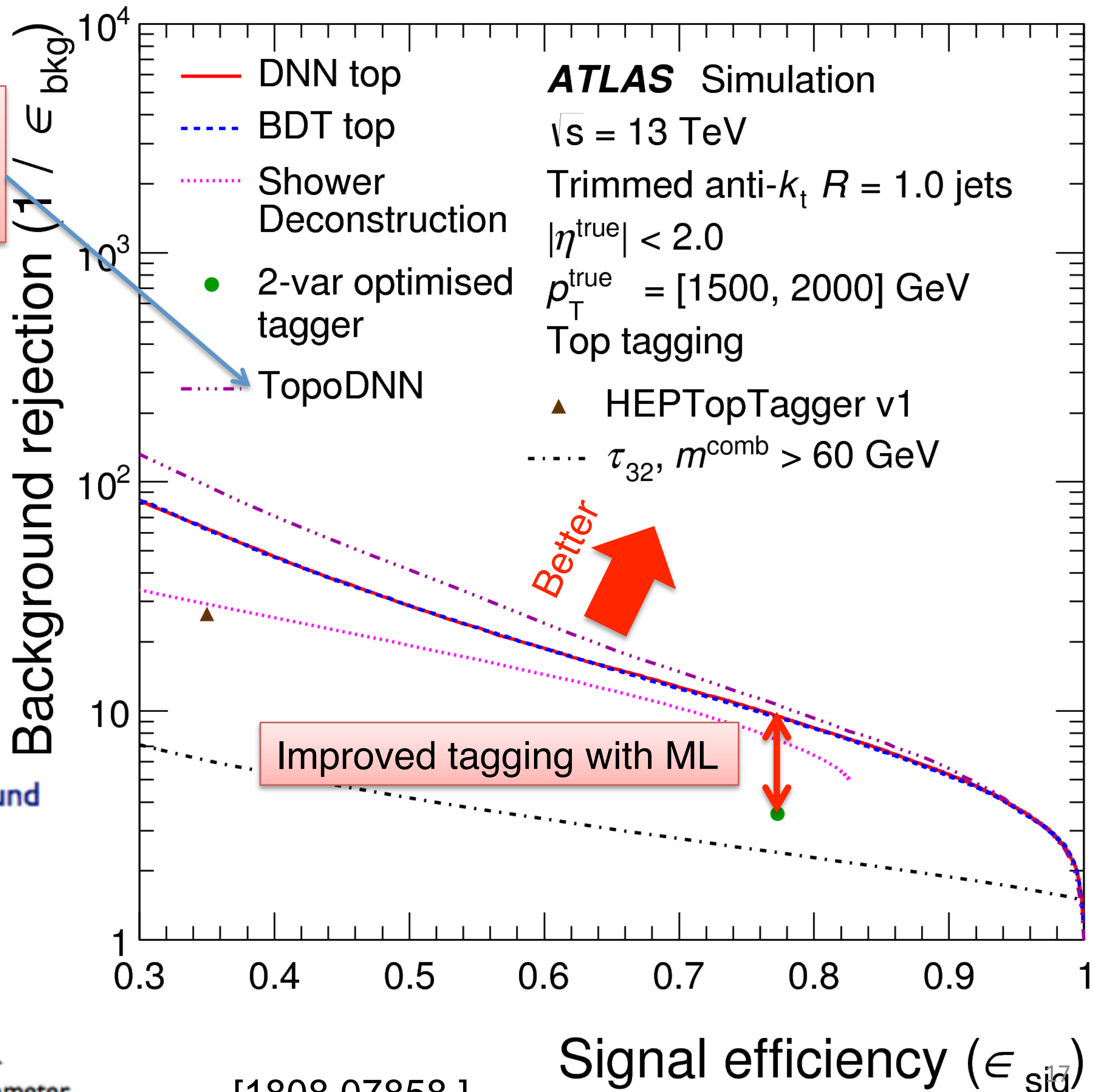


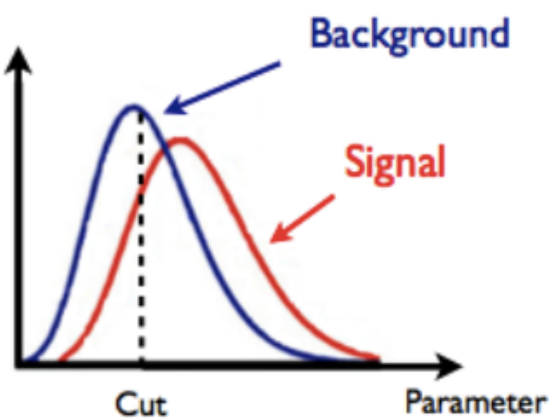
Best results with raw data as input [1704.02124]

Is there more information to extract?

Next frontier:

- Robustness
- Versatility
- Transparency

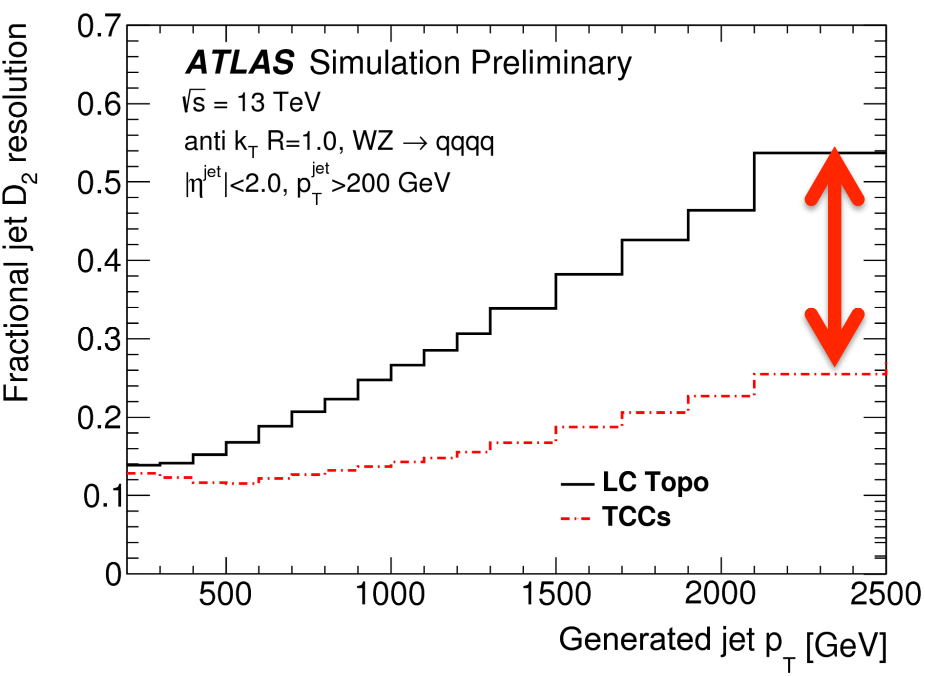


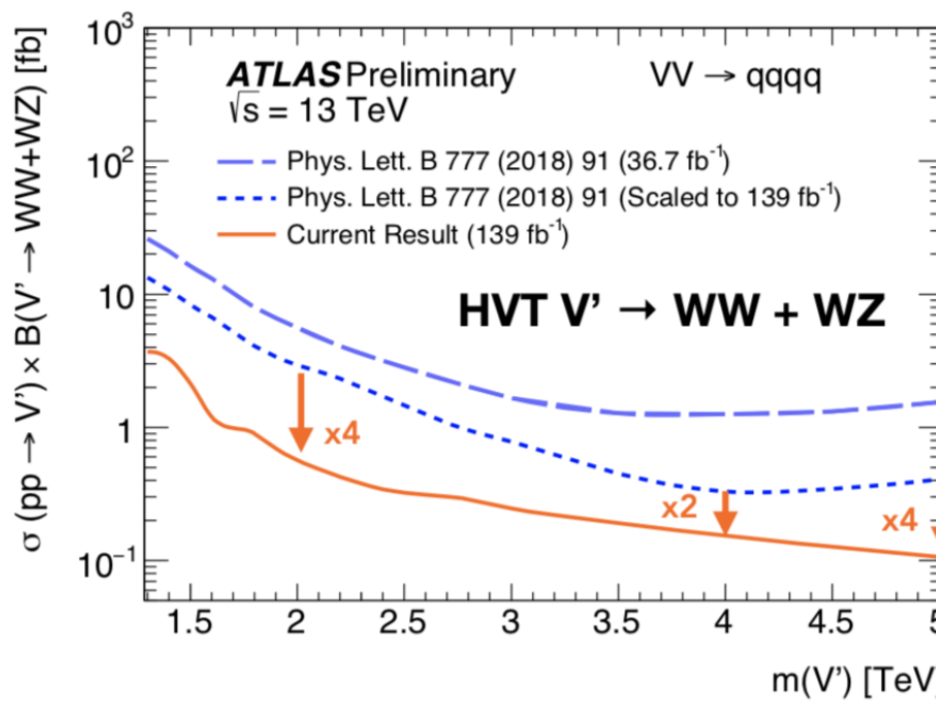


[1808.07858]

Boosted object tagging at highest p_T

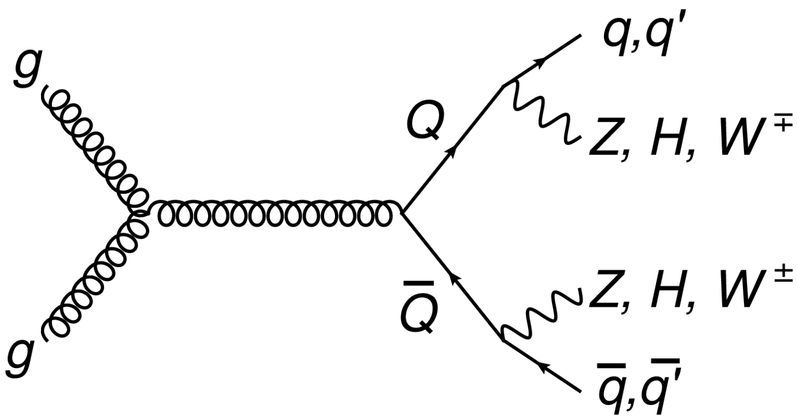
- Adequate training statistics at high p_T
- New features at high p_T: combine tracking & calorimeter info (Track-CaloCluster matching)

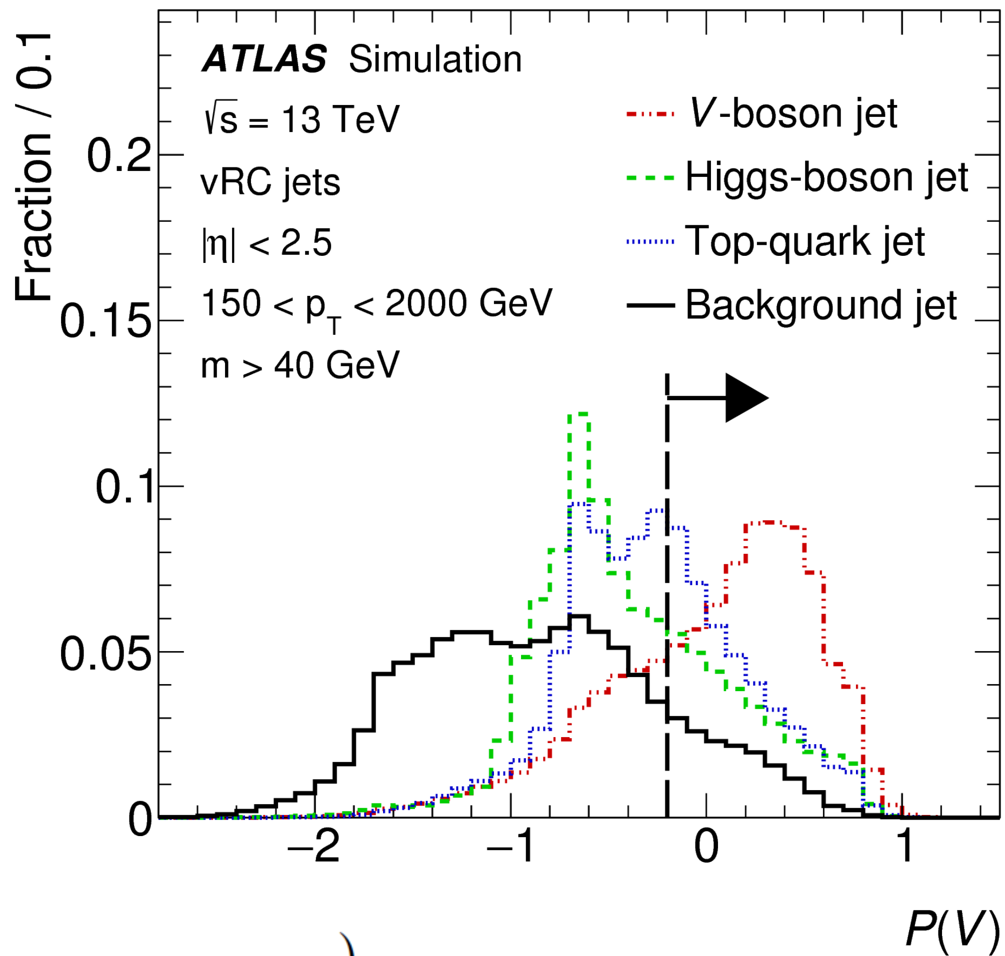




[ATLAS-CONF-2019-003]

Multi-class tagging with deep NN





- Discriminate (in context of VLQ search)
  - W/Z
  - Higgs
  - Top
  - QCD-jet

$$P(V) = \log_{10}\left(\frac{D^{V}_{\mathrm{DNN}}}{0.9 \cdot \boldsymbol{D}^{\mathbf{background}}_{\mathbf{DNN}} + 0.05 \cdot D^{t}_{\mathrm{DNN}} + 0.05 \cdot D^{H}_{\mathrm{DNN}}}\right)$$

[1808.01771]

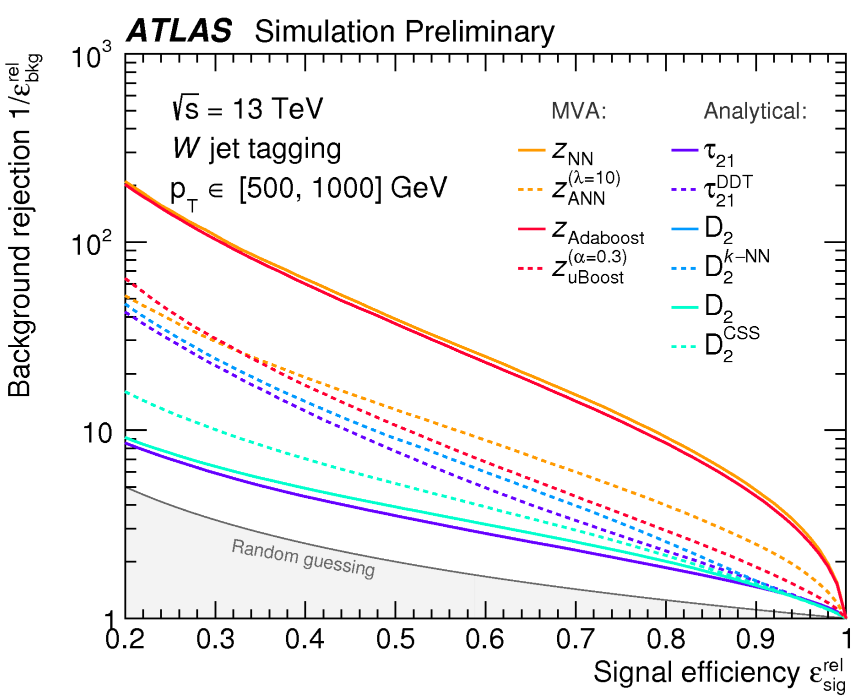
[ATL-PHYS-PUB-2018-014]

Mass decorrelation

- Minimize sculpting of background jet mass distributions
- Enable more robust background estimation
- **Adversarial NN** penalizes classifier if mass is learned:

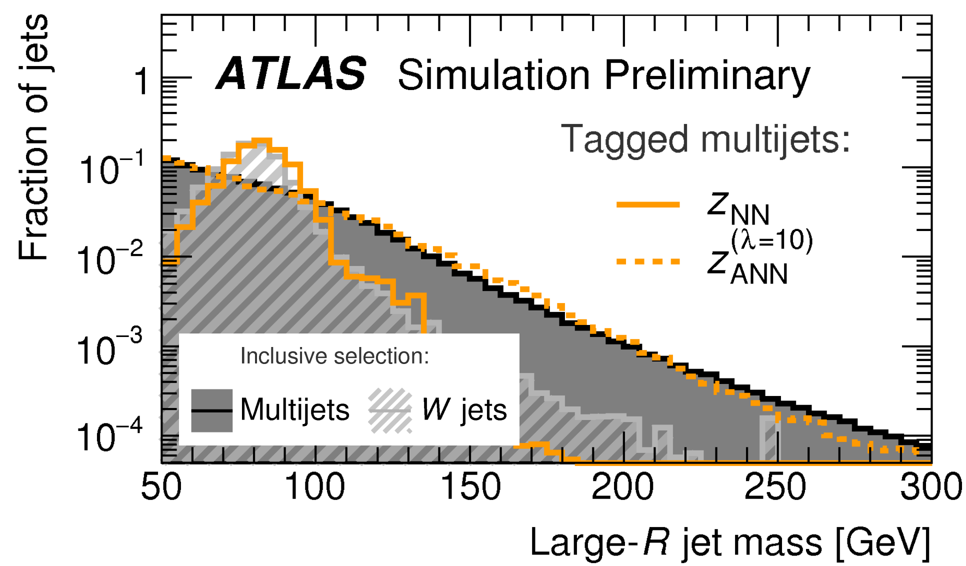
$$L = L_{\text{classifier}} + \lambda L_{\text{KL}}$$

\- Sacrifice performance



\+ Unsculpted BG

$\sqrt{s}$ = 13 TeV, *W* jet tagging
Cuts at ε_{sig}^{rel} = 50%

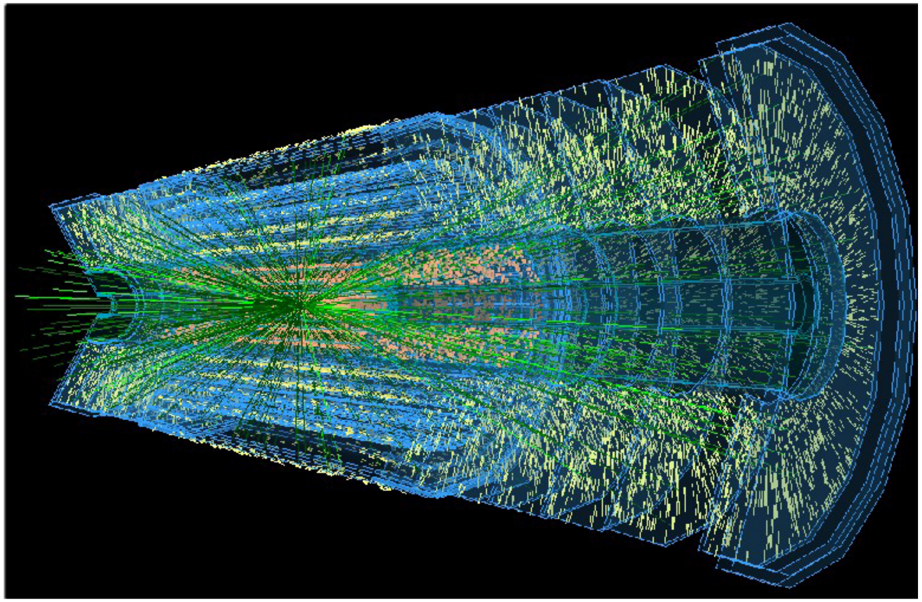


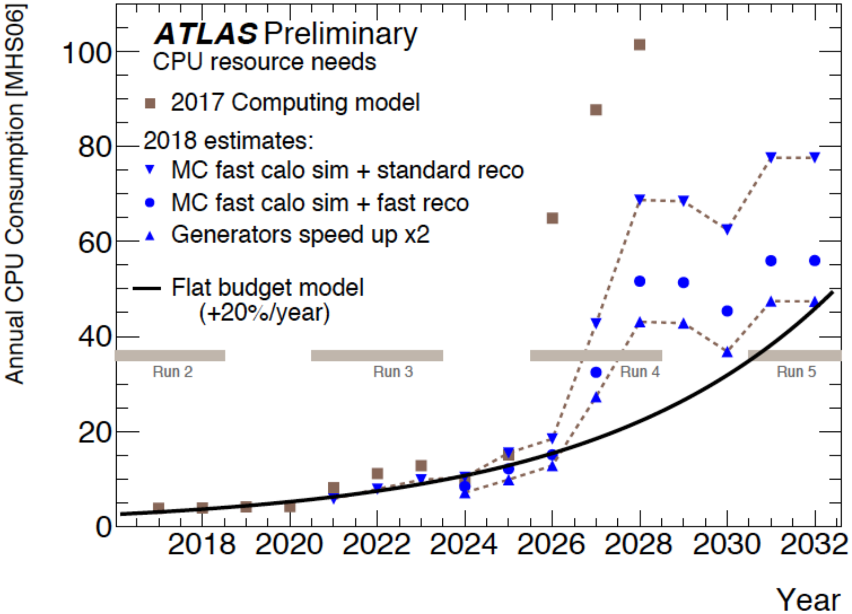
Mitigating impact of MC mismodeling (or pile-up dependence) is further application of ANN
Or train on data [1702.00414, 1708.02949]: correct labeling → **correct proportions**

Beyond classification: clustering & generative models (faster)

2026: High-Lumi-LHC tracking crisis

100’000 space-points
10’000 tracks





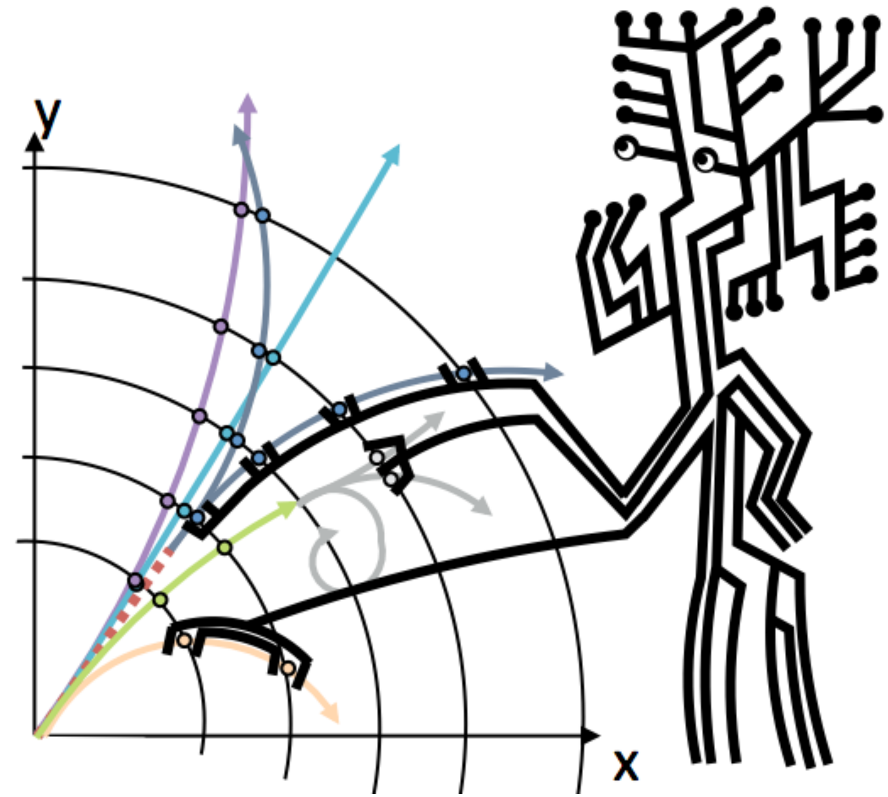
Current algorithm:
combinatorial approach = slow!

Reconstruction limited by tracking

TrackML challenge

- Can ML help?
- *HL-LHC* data set with ACTS
  - $t\bar{t}$, 200 pile-up

ACTS: public high-fidelity simulation [http://acts.web.cern.ch/ACTS/]



<u>2 phases:</u>

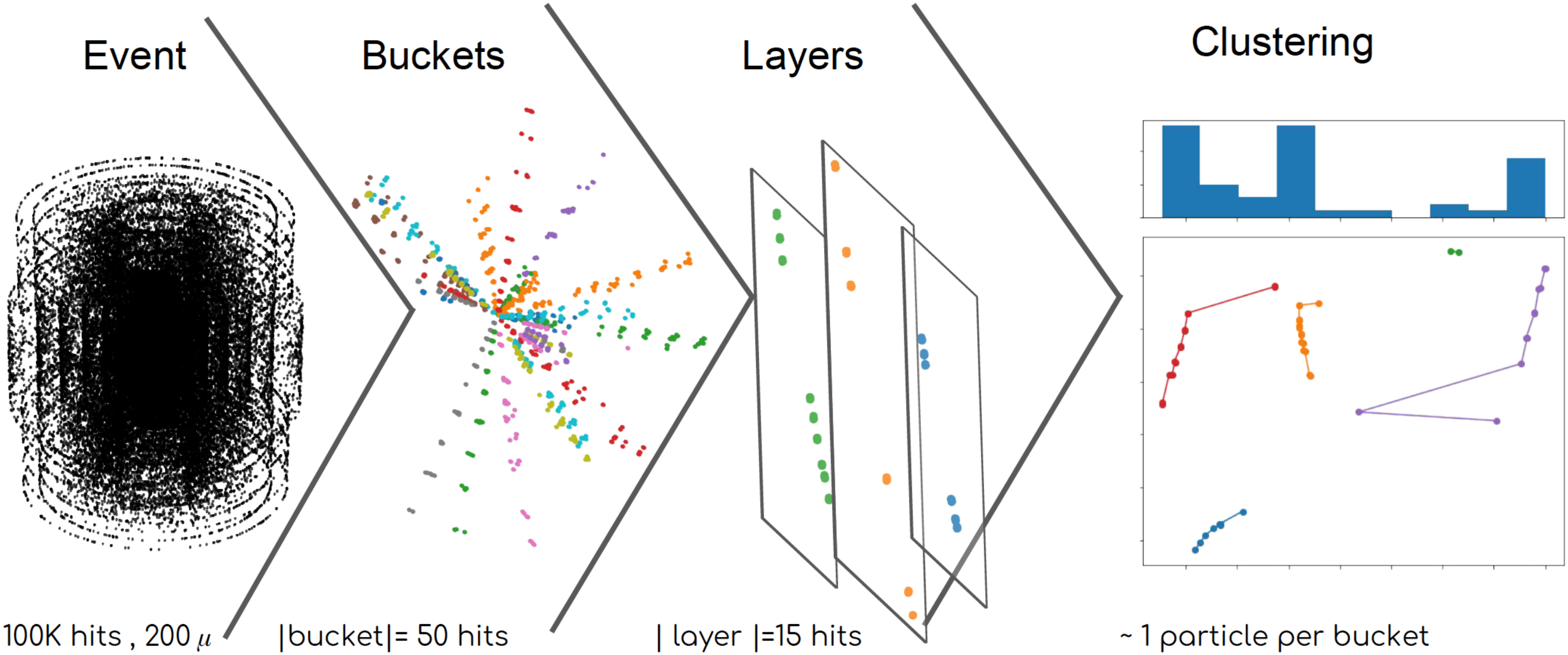
kaggle https://www.kaggle.com/c/trackml-particle-identification Accuracy

CodaLab https://competitions.codalab.org/competitions/20112 Throughput

Promising 2-Step Approach

1) Reduce combinatorial complexity

2) Track finding inside regions



[CTD/WIT2019 talk by Sabrina Amrouche]

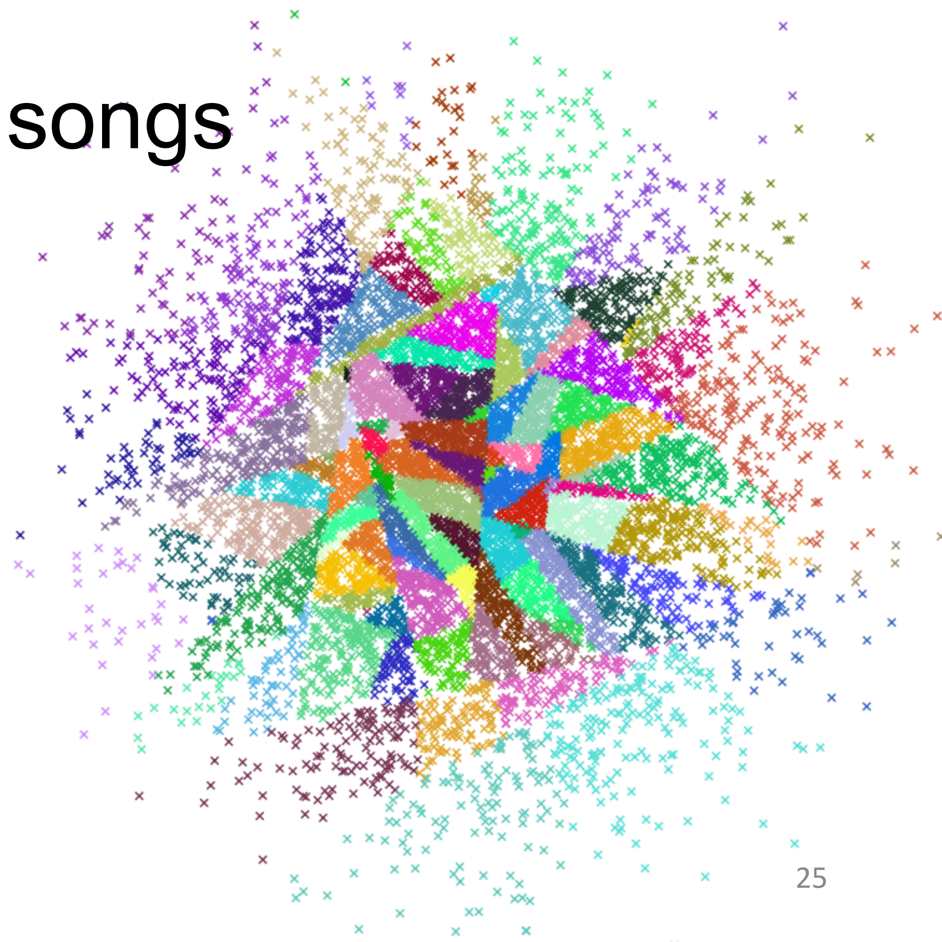
Inspiration from spotify

- Approximate Nearest Neighbors: https://github.com/spotify/annoy

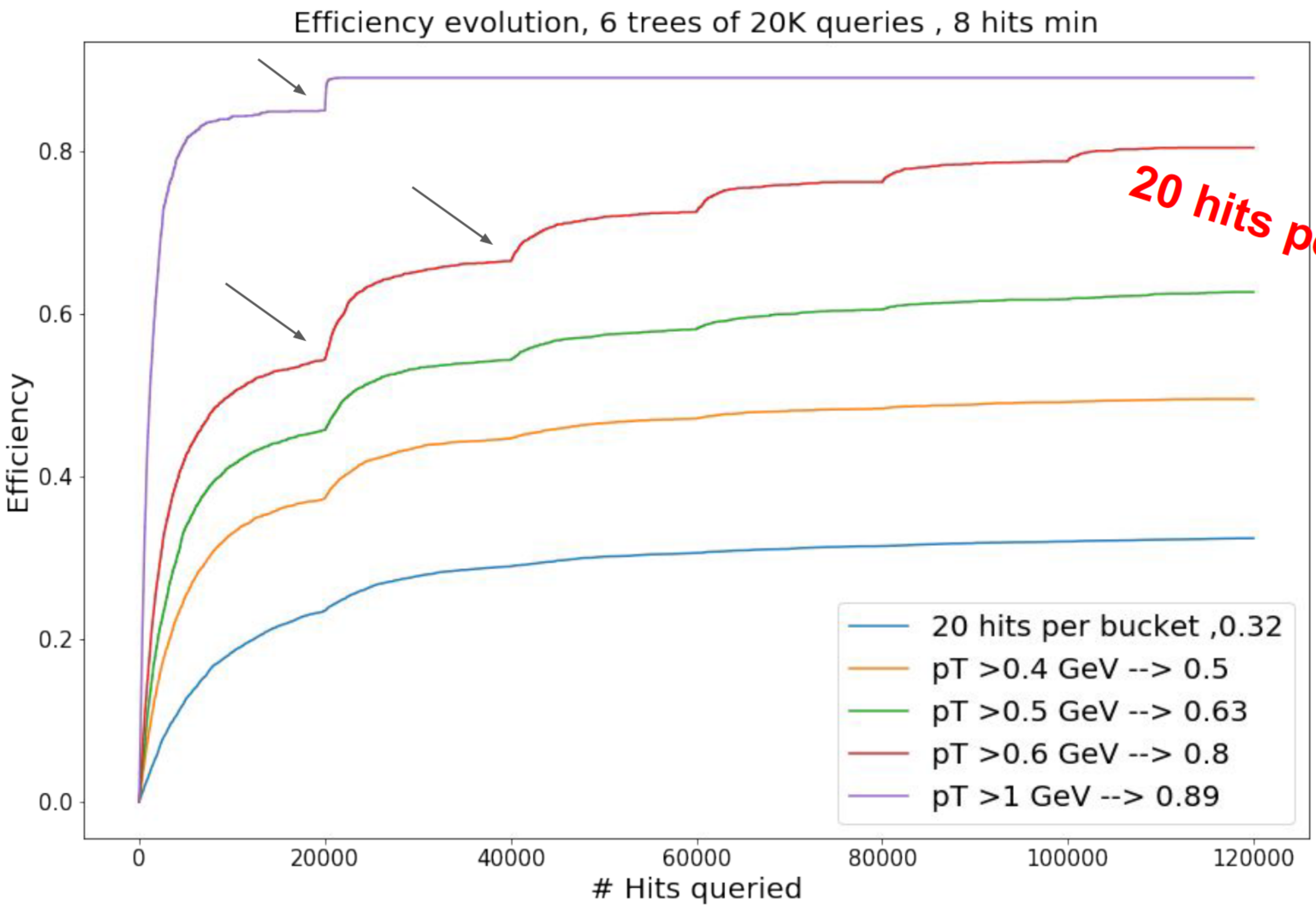
- < 0.1ms to get n similar songs
  - high-dimensional space

- Unsupervised

- Bucket definition based on angular distance



Bucketing provides fast & efficient seeding



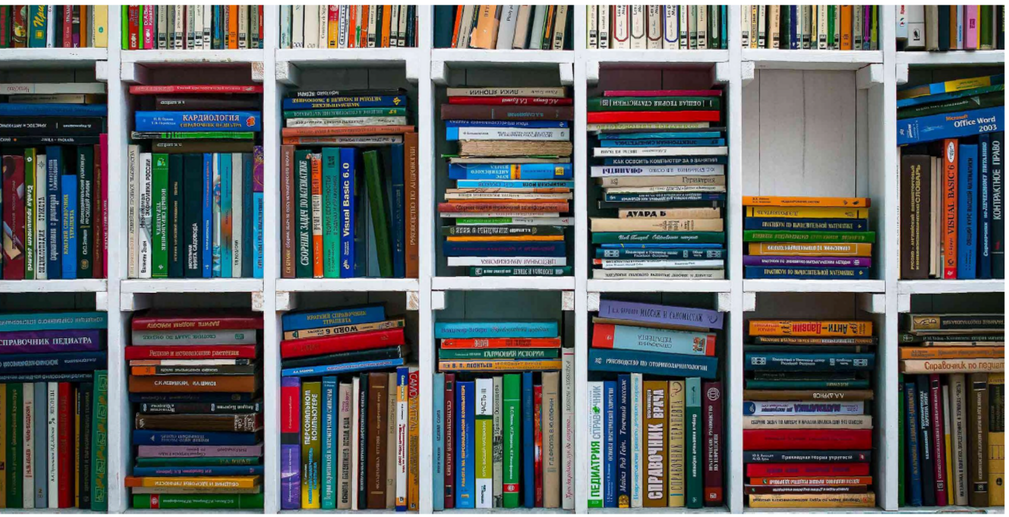
20 hits per bucket

0.07 ms per query

58

Data Preparation

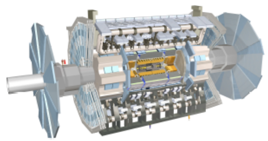
Data structure = Layout for memory

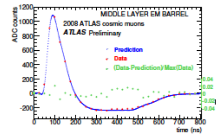


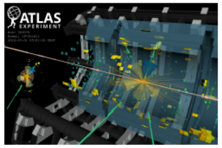
vs

- Most tracks are largely contained within a bucket
- Reduce problem to track finding inside buckets

Large-scale and high-fidelity simulation



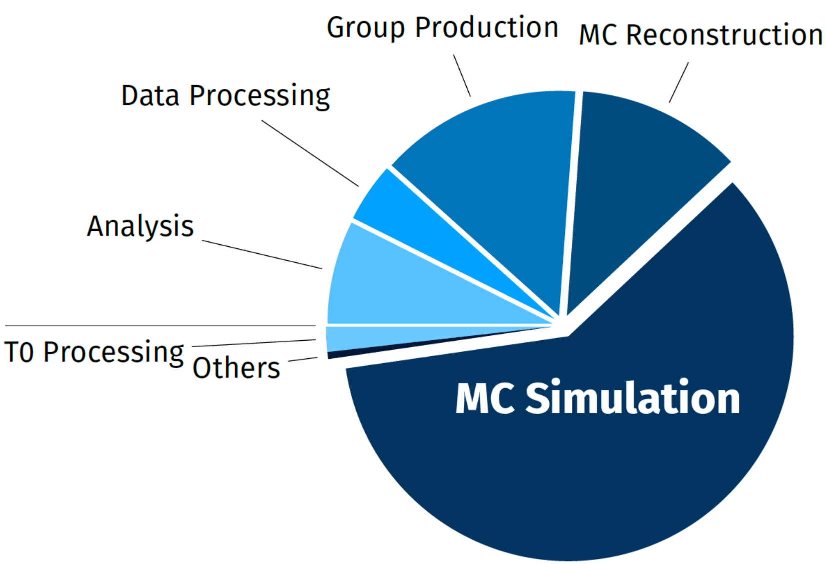




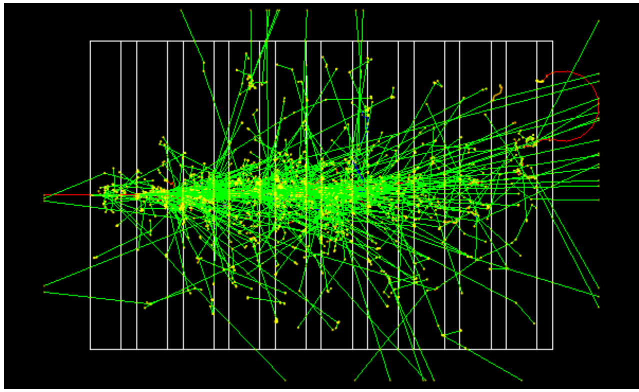
ROOT
Data Analysis Framework

Event generation → **Detector simulation** → Digitization → Reco. → Physics analysis

Bottleneck!



Dominated by Geant shower simulation!



Fast Calorimeter Simulation

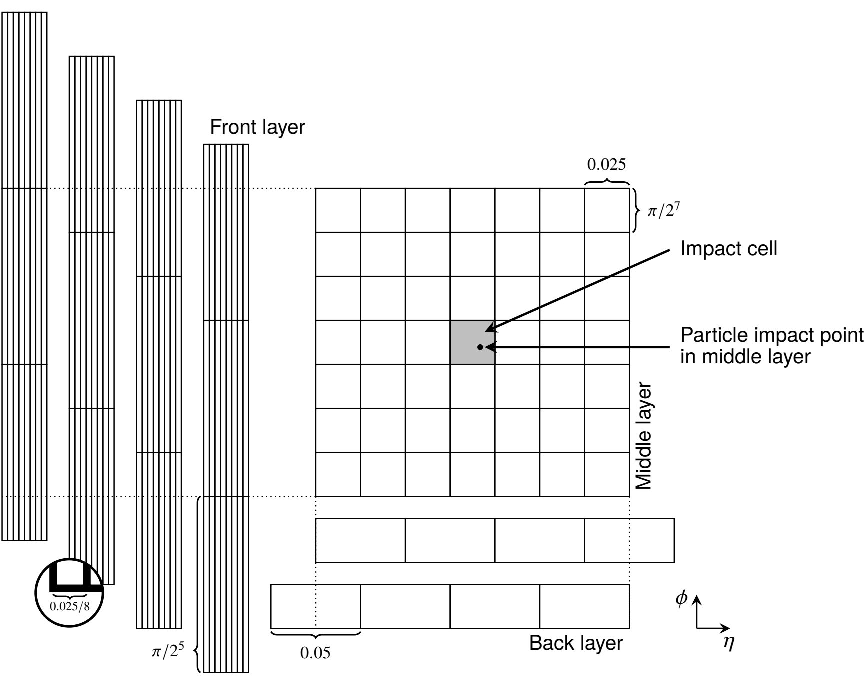
- Geant 4 too slow
- FastCaloSim V1 [ATL-PHYS-PUB-2010-013] used for years
- Improved FastCaloSim V2 [ATL-SOFT-PUB-2018-002] using PCA
- FastCaloSim fast enough but still not accurate enough for all simulation needs

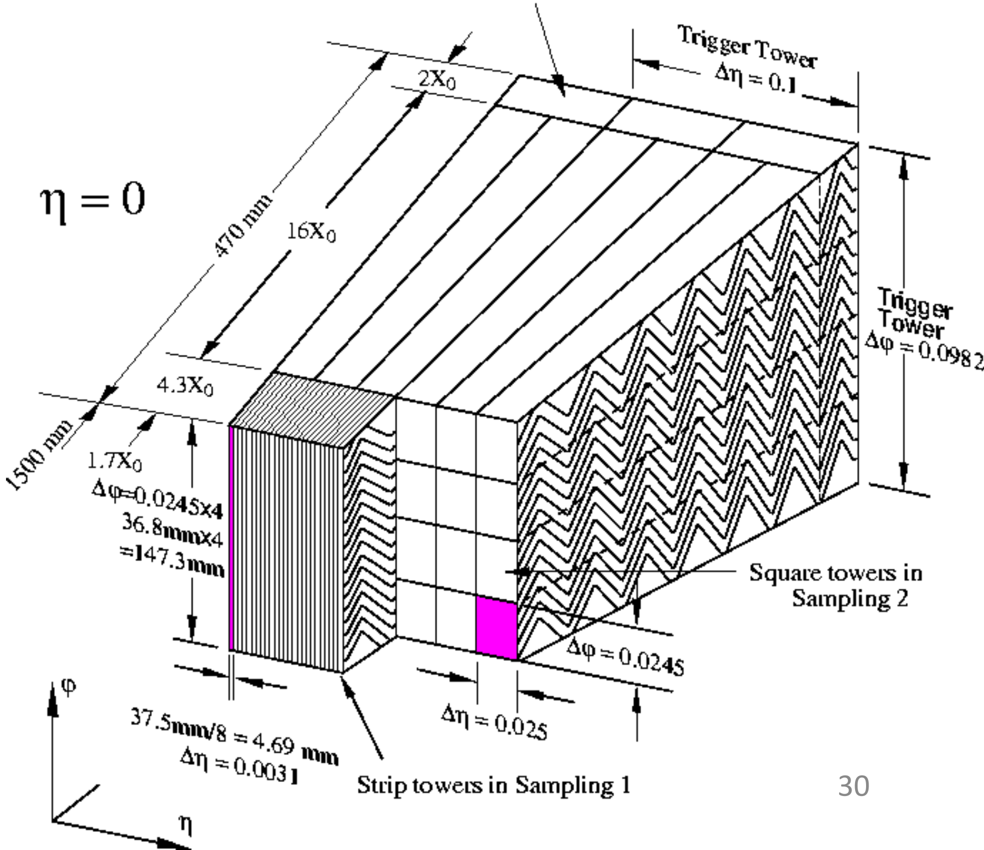
- Objective: generative models to simulate calorimeter showers [ATL-SOFT-PUB-2018-001]

- Challenges:
  - Non-uniform geometry
  - Sparse data
  - Large dynamic range: tails

The ATLAS EM calorimeter

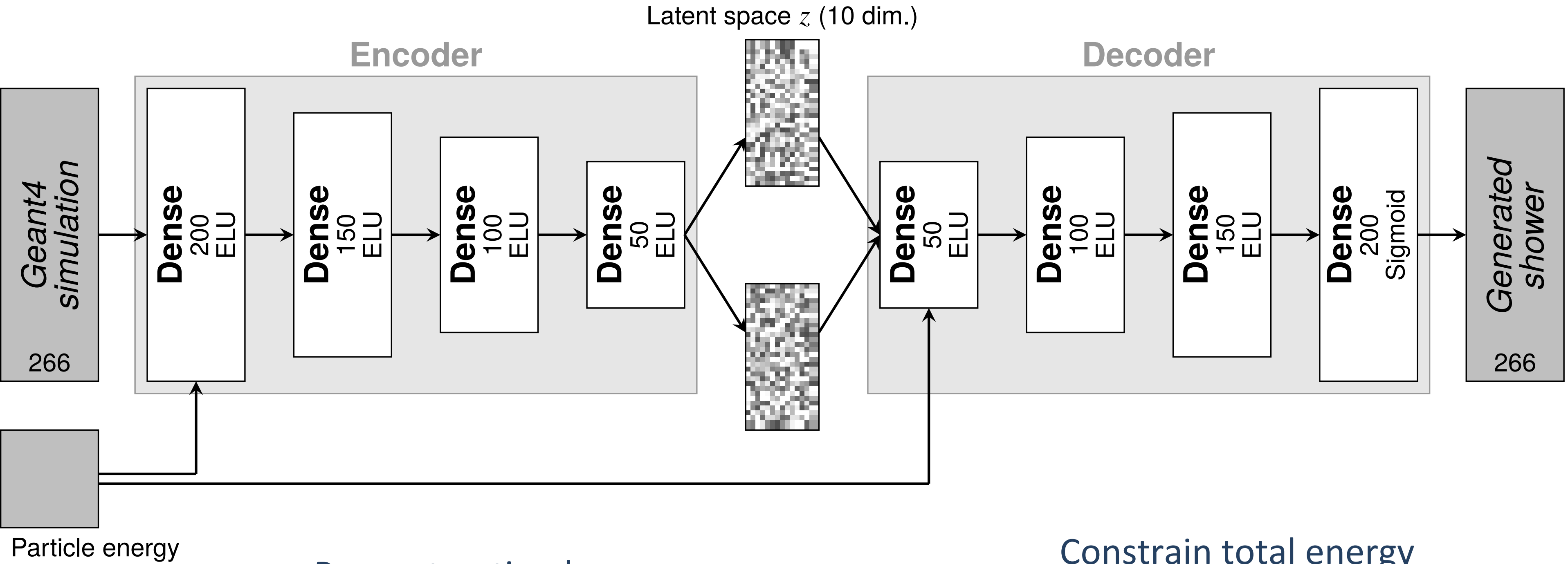
- Train and validate using G4 simulation of photons for the ATLAS geometry
  - Discrete particle energies logarithmically spaced between 1 and 260 GeV
  - Uniformly distributed in 0.20 < |η| < 0.25





Deep Generative Models (VAE, GAN)

Variational Auto-Encoder (VAE) architecture: [ATL-SOFT-PUB-2018-001]



Reconstruction loss

Constrain total energy

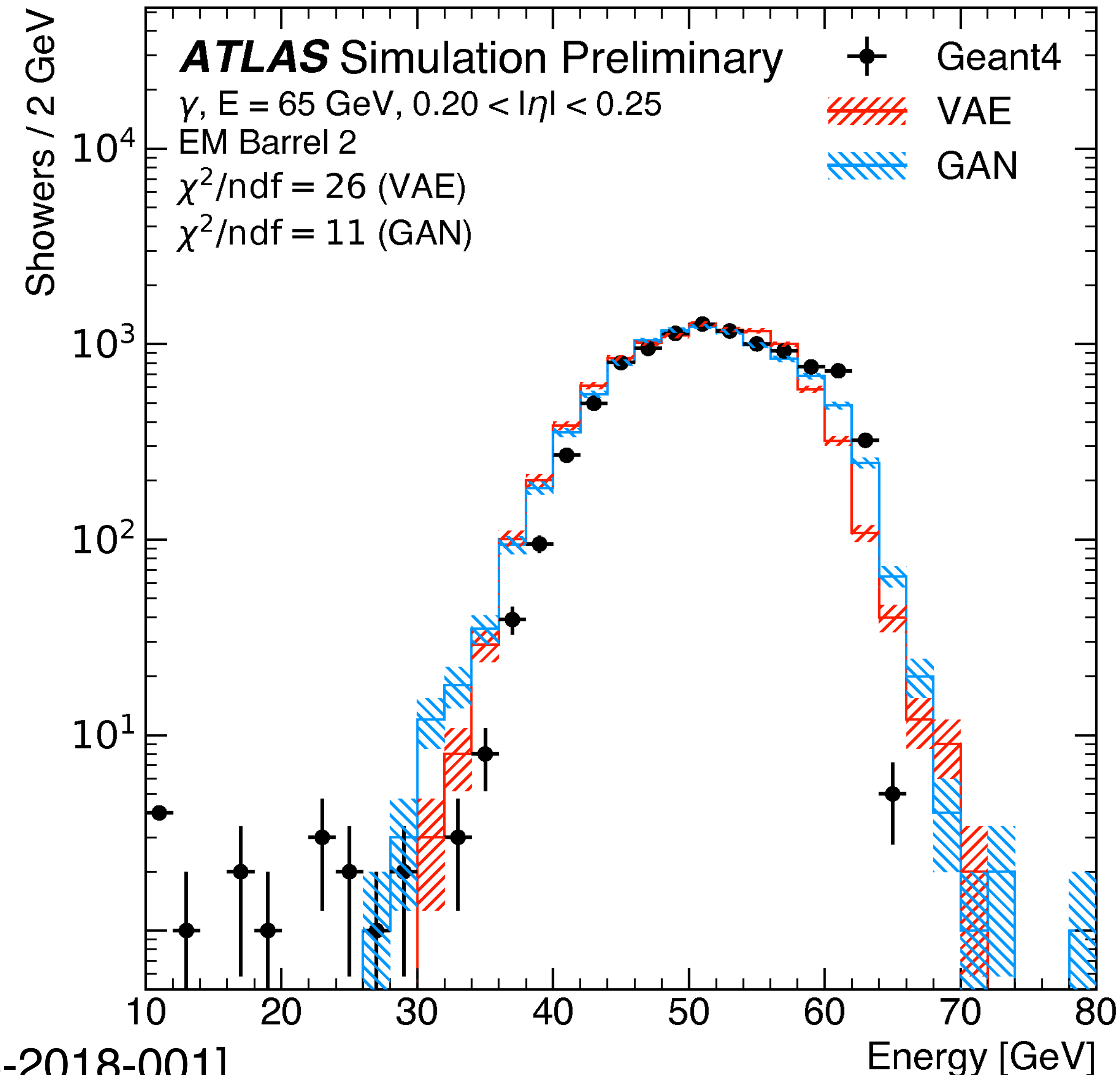
$$L_{\mathrm{VAE}}(x, \tilde{x}) = \boxed{w_{\mathrm{reco}}} E_{z\sim q_\theta(z|x)}[\log p_\phi(x|z)] - \boxed{w_{\mathrm{KL}}} \mathrm{KL}(q_\theta(z|x)||p(z)) + \boxed{w_{E_{\mathrm{tot}}}} L_{E_{\mathrm{tot}}}(x, \tilde{x}) + \sum_i^M \boxed{w_i} L_{E_i}(x, \tilde{x}).$$

Kullback-Leibler divergence

Constrain energy fractions in layers

Deep Generative Models (VAE, GAN)

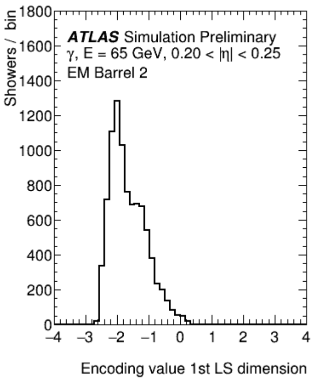
Validation: promising

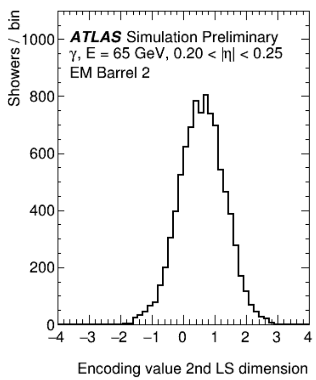


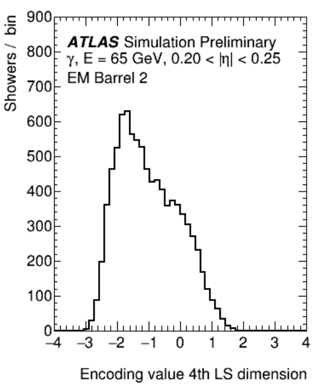
[ATL-SOFT-PUB-2018-001]

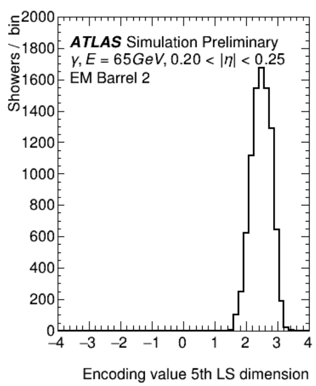
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-003/

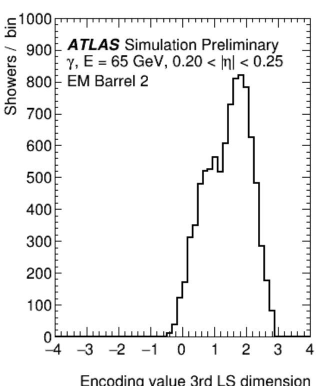
VAE Latent Space







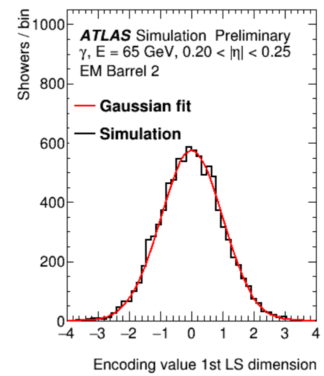


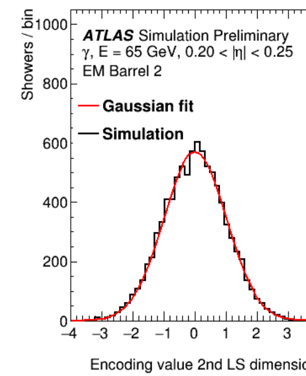


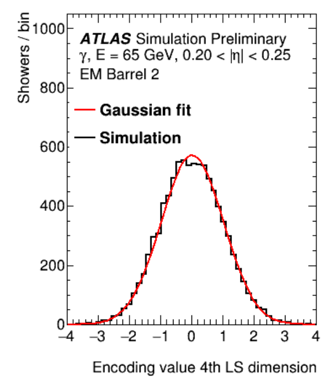
Non-Gaussian 5D latent space

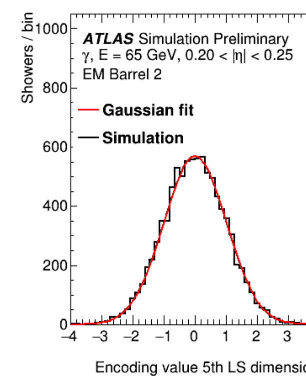
Inverse Autoregressive Flow [1606.04934]: latent space more Gaussian

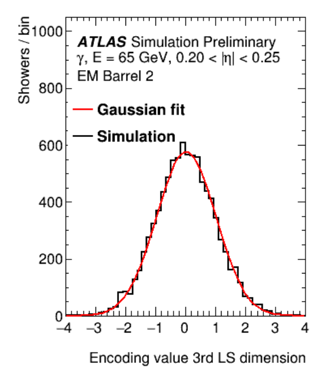
Allows one to sample from Gaussian distribution











Integration into ATLAS (C++) Software

- Implemented with Lightweight Trained Neural Network (LWTNN) [https://github.com/lwtnn/lwtnn]
- Flag to switch trained model
- DNNCaloGAN same speed as FastCaloSim V2
  - 65 GeV single photon: G4 → FCS : 10 seconds → 70 ms
- LWTNN takes <1 ms per shower

- Model trained on particles with fixed energy, but interpolates well to other energies

- Only central eta for now
- Plan: use higher uniform granularity for full eta range

The ML revolution in HEP has started

Relevance for flavour tagging

- Exciting time – room for **creativity**
  - Opportunity for young researchers to think outside the box
- Fruitful interdisciplinary work

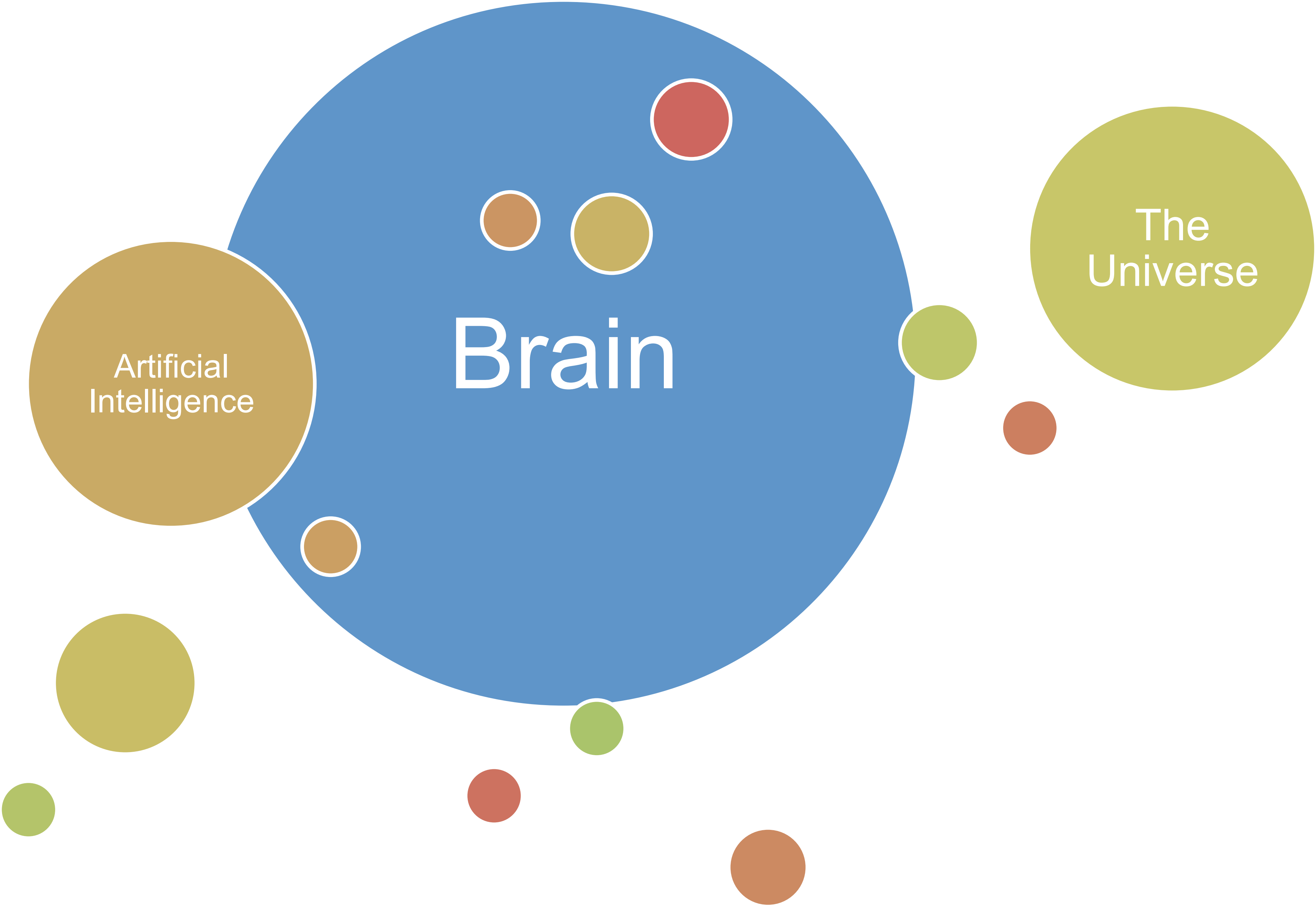
…but also hard work from proof of principle to realization in ATLAS

- Better, faster, more automated,…

…but also maintainability & memory footprint

- Trend from "ML in analysis" to "ML for simulation, reconstruction, trigger"
- "Raw data" vs. "human-engineered features"
- Domain knowledge vs. what machine learns

- New frontiers
  - Latency: trigger
  - Specialized hardware: FPGA, GPU, Custom DL chips,…
  - Anomaly detection
  - Interpretability
  - …


Brain
Artificial Intelligence
The Universe

Backup