EXPERIMENTAL ENTRÉE

A TASTING BOARD OF LHC SCIENCE WITH MACHINE LEARNING

Daniel Murnane
Niels Bohr Institute, University of Copenhagen

UNIVERSITY OF COPENHAGEN

BERKELEY LAB

Danish Data Science Academy

DISCLAIMER

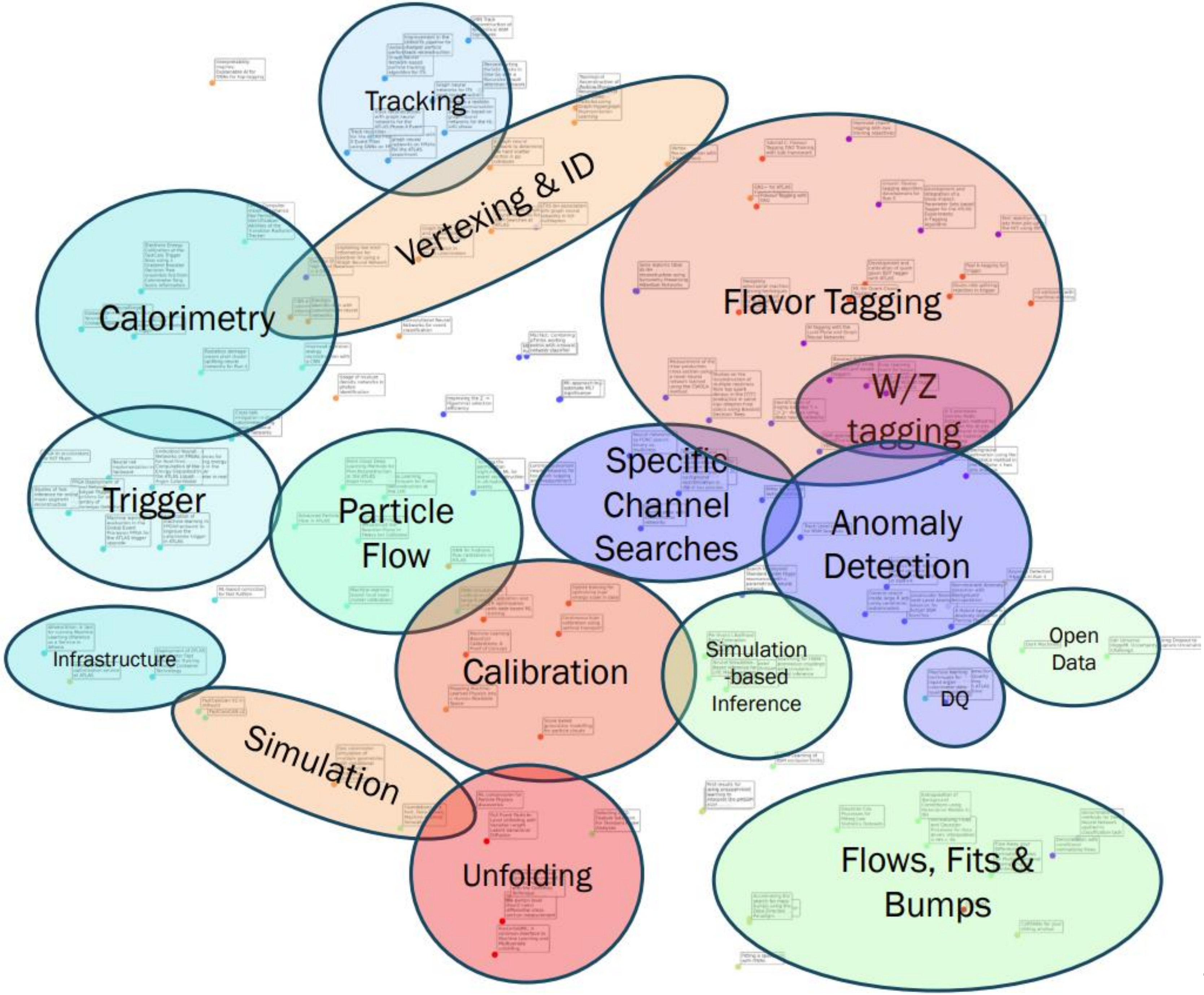
- I am an ATLAS ML Forum co-convener – but this is not the “official position” of the ML Forum
- I don’t speak for Niels Bohr Institute or Berkeley Lab or Danish Data Science Academy or ATLAS or, really, anybody! I’m going to give this talk from my personal vantage point
- I have tried to keep my personal interests at the door in selecting the examples in the talk. In fact, I find *all* of them interesting, but I think they are also an accurate representation of the breadth of in-production ML

WHAT DOES ML AT THE LHC "LOOK LIKE" TODAY?

- There are many incredibly sophisticated ideas emerging (e.g. at this workshop!), but I'm not going to deeply explain *any* machine learning ideas
- My focus here is: *What is being used right now to produce rigorous results released by a collaboration?*
- I embed every abstract from CDS that includes ML terminology in the title or abstract, released by a major* LHC collaboration within the last 12 months

* "Major" is biased, apologies if I have missed any work

WHAT DOES ML AT THE LHC "LOOK LIKE" TODAY?



See this image in high def!
https://github.com/murnanedaniel/ML4Jets-Overview/blob/main/notebooks/tsne_kmeans_clustering_high_res.pdf

DATA SCIENCE IN THE DISCOVERY PIPELINE

Simulation				Reconstruction					Analysis	
Matrix-element Calculation	Parton-shower / Hadronization	Detector Simulation	Digitization	Topoclusters & Spacepoints	Track Finding & Fitting	Jet Tagging & Vertexing	Particle ID & Particle Flow	Calibration	Likelihood Fitting	Unfolding

Numerical Integration

Markov Chain Monte Carlo

Topological clustering

Kalman Filtering & Fitting

Conformal Fits & Hough Transform

Statistical Techniques, Bayesian Inference

ML TODAY & TOMORROW IN THE DISCOVERY PIPELINE

Simulation

Matrix-element Calculation

Parton-shower / Hadronization

Detector Simulation

Digitization

Reconstruction

Topoclusters & Spacepoints

Track Finding & Fitting

Jet Tagging & Vertexing

Particle ID & Particle Flow

Calibration

Analysis

Likelihood Fitting

Unfolding

Generative Models: GANs, VAEs, Normalizing Flows and Diffusion

Metric Learning, Object Condensation

Deep Full Event Reconstruction

CNNs, Graph Neural Networks & Transformers

Symmetric ML & Equivariance

Autoencoders & Anomaly Detection

Omnifold and Likelihood-free Inference

ML TODAY & TOMORROW IN THE DISCOVERY PIPELINE

Simulation				Reconstruction					Analysis	
Matrix-element Calculation	Parton-shower / Hadronization	Detector Simulation	Digitization	Topoclusters & Spacepoints	Track Finding & Fitting	Jet Tagging & Vertexing	Particle ID & Particle Flow	Calibration	Likelihood Fitting	Unfolding

Generative Models: GANs, VAEs, Normalizing Flows and Diffusion

Metric Learning, Object Condensation

Deep Full Event Reconstruction

CNNs, Graph Neural Networks & Transformers

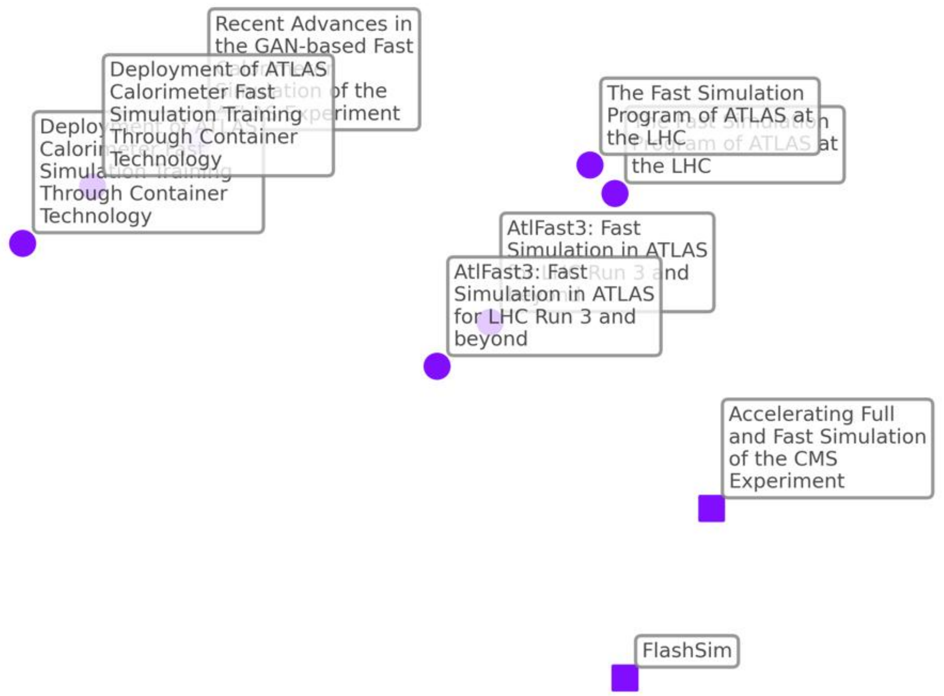
Symmetric ML & Equivariance

Autoencoders & Anomaly Detection

Omnifold and Likelihood-free Inference

For information about many of these applications…
https://iml-wg.github.io/HEPML-LivingReview/

SIMULATION

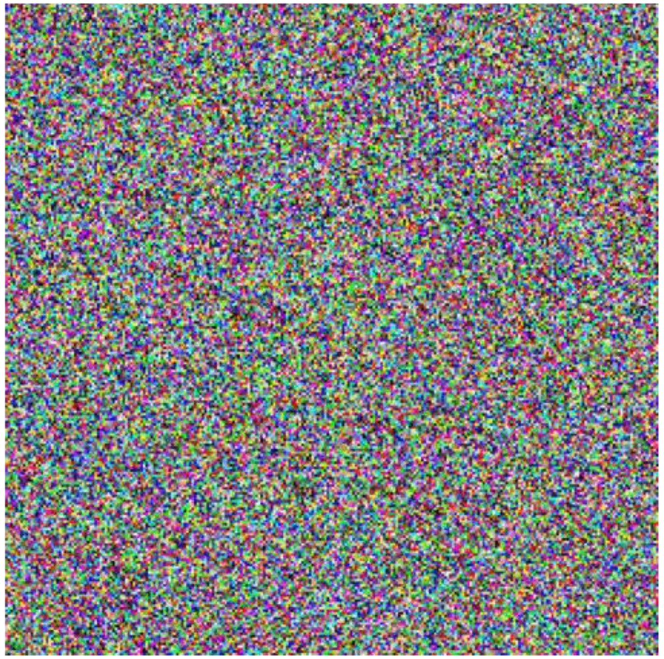


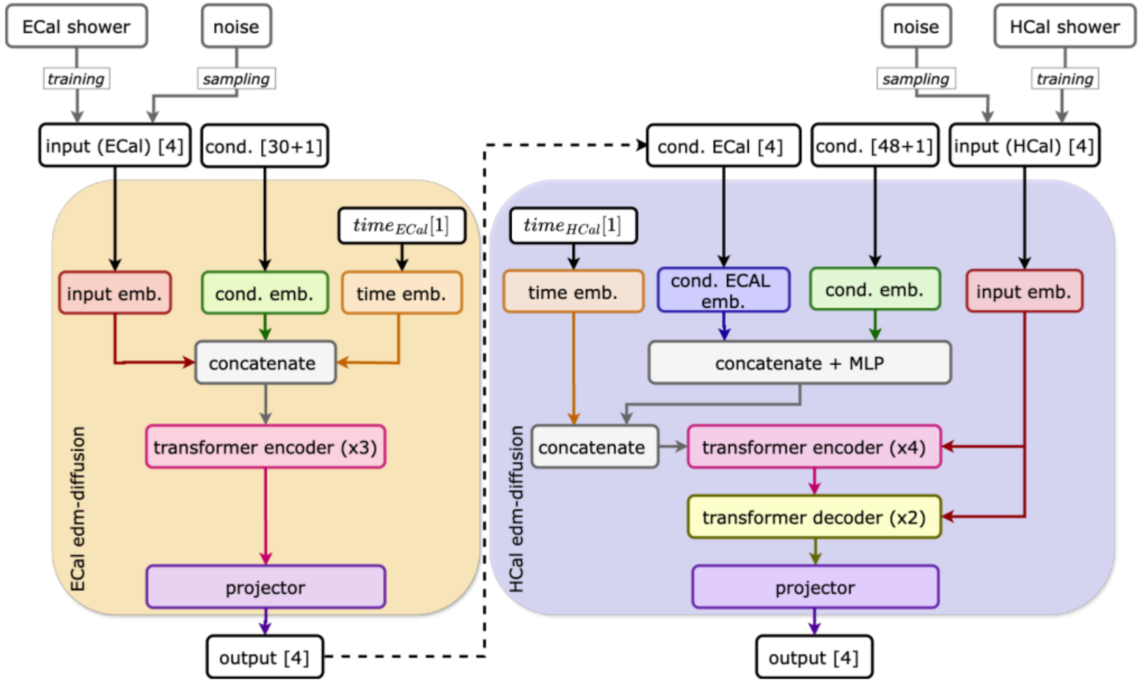
Recent Advances in the GAN-based Fast
Deployment of ATLAS Calorimeter Fast Simulation Training Through Container Technology
Through Container Technology
The Fast Simulation Program of ATLAS at the LHC
AtlFast3: Fast Simulation in ATLAS for LHC Run 3 and beyond
Accelerating Full and Fast Simulation of the CMS Experiment
FlashSim

GENERATIVE MODELS

- In HEP, simulation is often **very expensive:** Feynman integrals over next-to-next-to…-leading order, interactions with complicated materials, high numbers of particles, noise
- Could we skip the full numerical simulation of underlying physics?
- This is a relatively immature space – models that generate realistic new data are hard to train, unstable, and not yet well-suited to all data structures

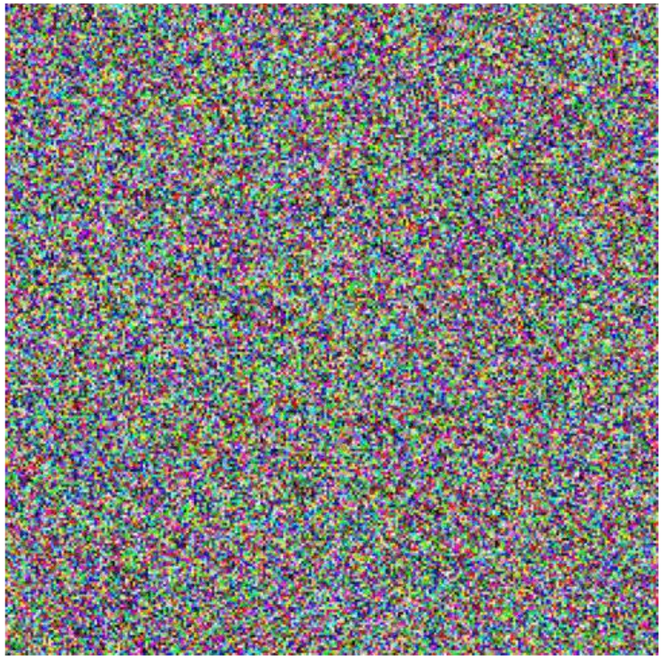
COLLISION SIMULATION WITH DIFFUSION

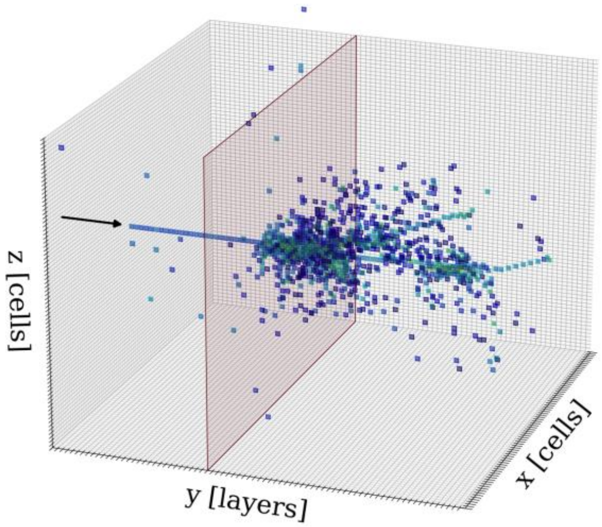




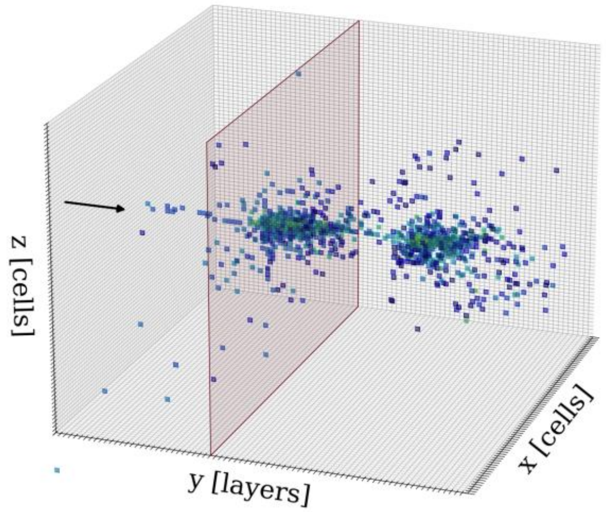
Link

COLLISION SIMULATION WITH DIFFUSION





Full Simulation



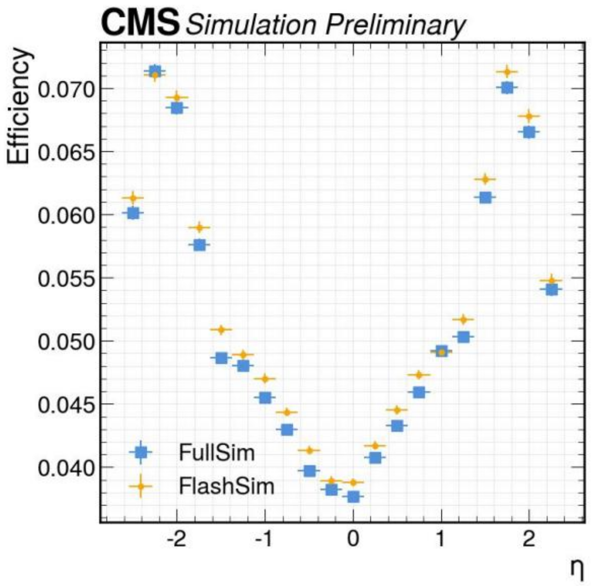
Generative Model

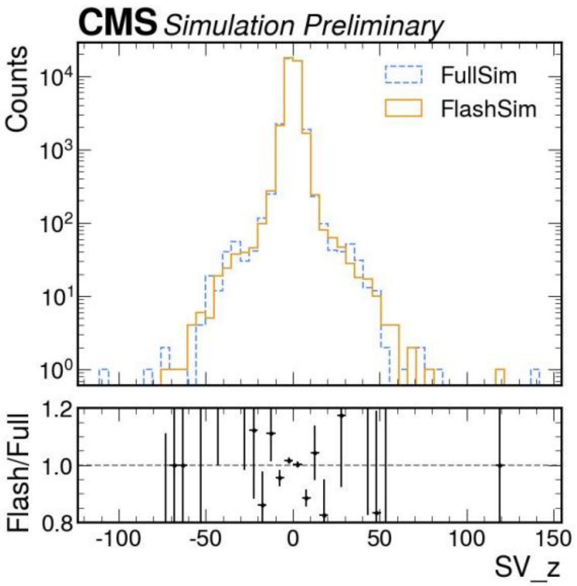
Link

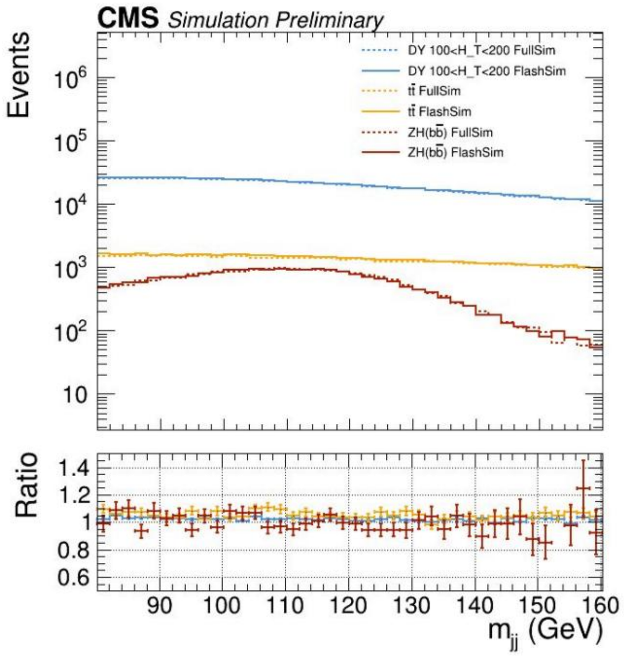
FLASHSIM

- Goal: Produce NANOAOD events at O(100)Hz with equivalent accuracy as existing FastSim
- NANOAOD has some hundreds of objects, upon which FlashSim trains Conditional Normalizing Flows with Flow Matching
- Each object, for each source, is generated with a dedicated model
- Event generation broken up into several steps – primary vertex, pile-up, secondary vertices (upper plots), jets, fake jets, and so on
- Some 300-30000x faster than FullSim

https://cds.cern.ch/record/2913372



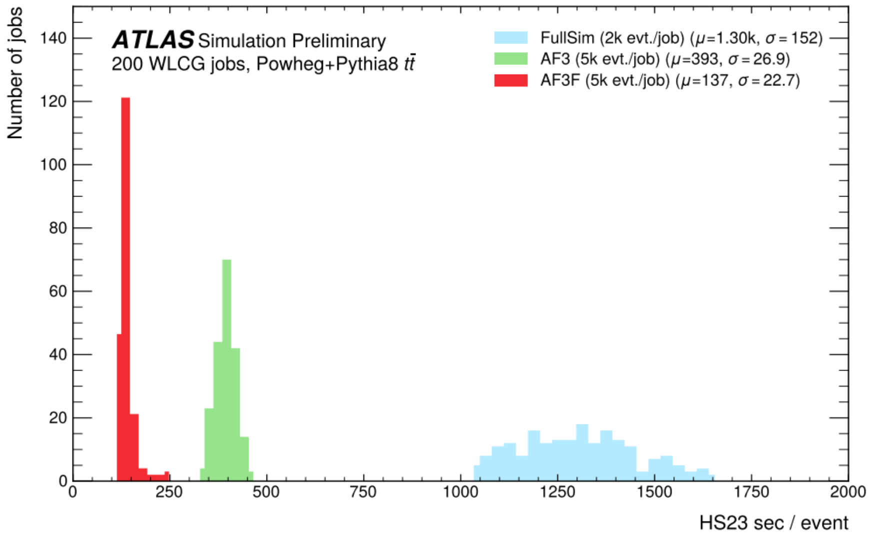


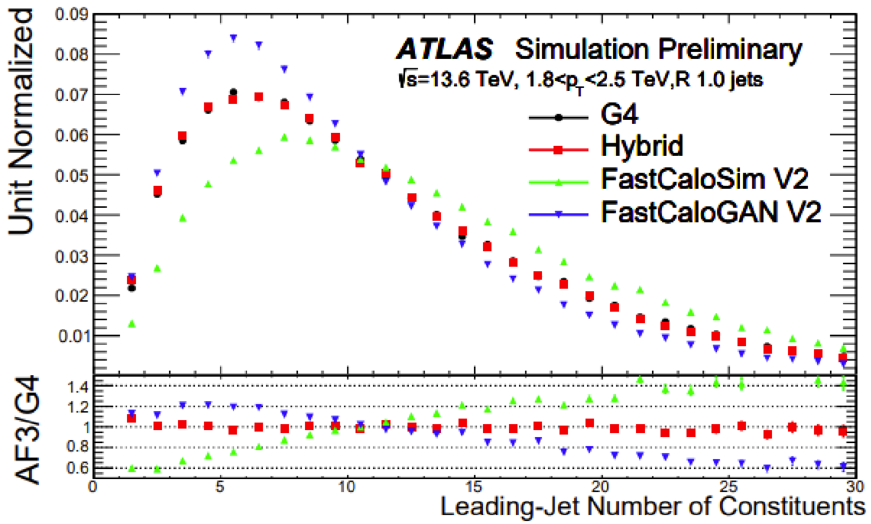


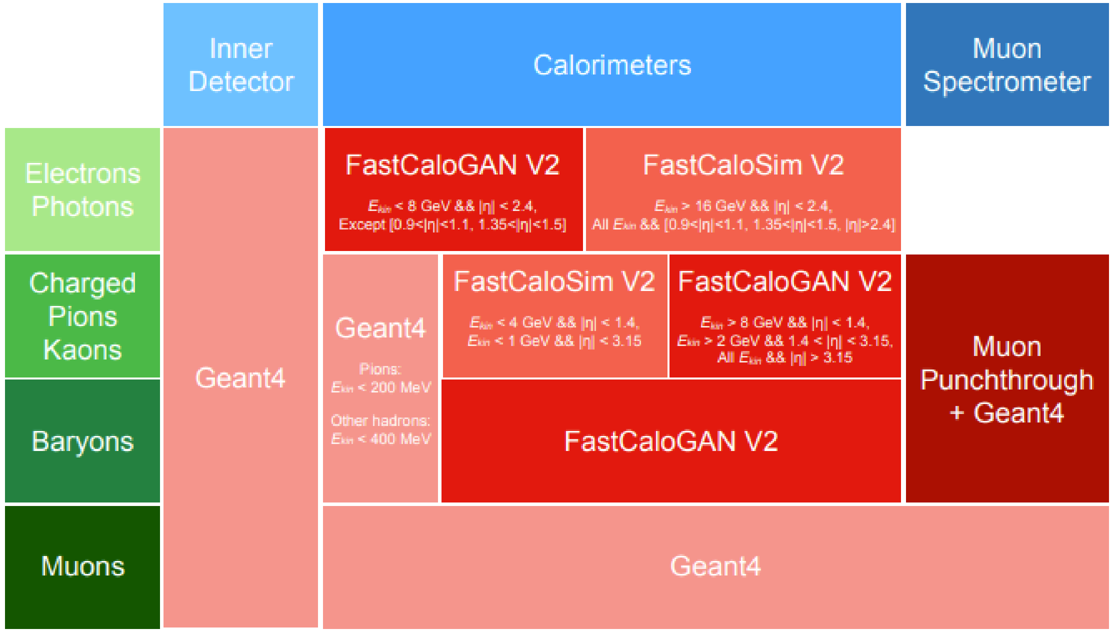
FAST SIMULATION AT ATLAS

https://cds.cern.ch/record/2911769

- Fast simulation at ATLAS is in production as the AtlFast3 system
- A collection of full sim, GANs, and simplified simulations
- E.g. tracking done with FATRAS leads to AtlFast3 being ~10x faster than full sim
- However, resolution accuracies in the FATRAS configuration have errors of O(10%), more work needed to improve

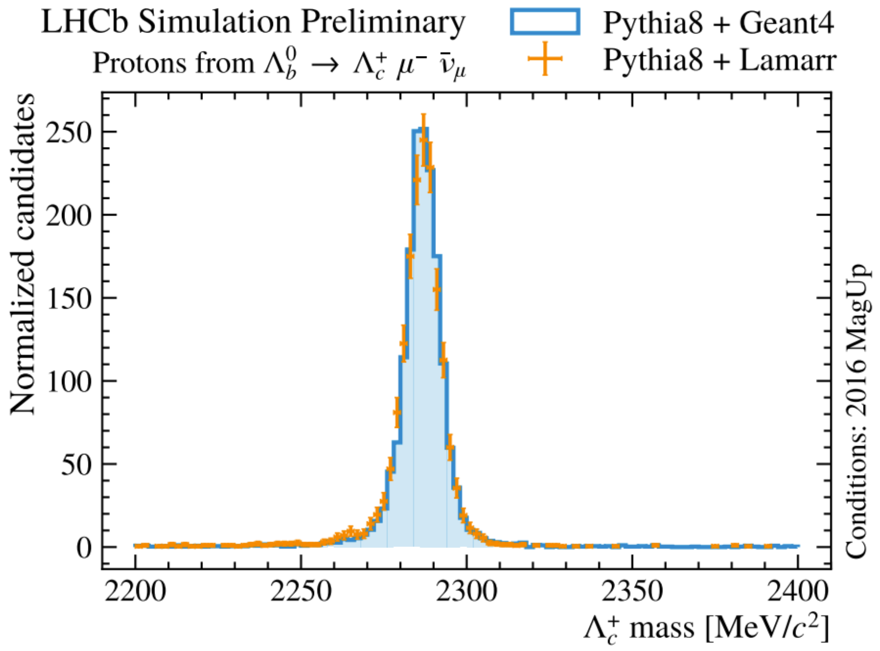






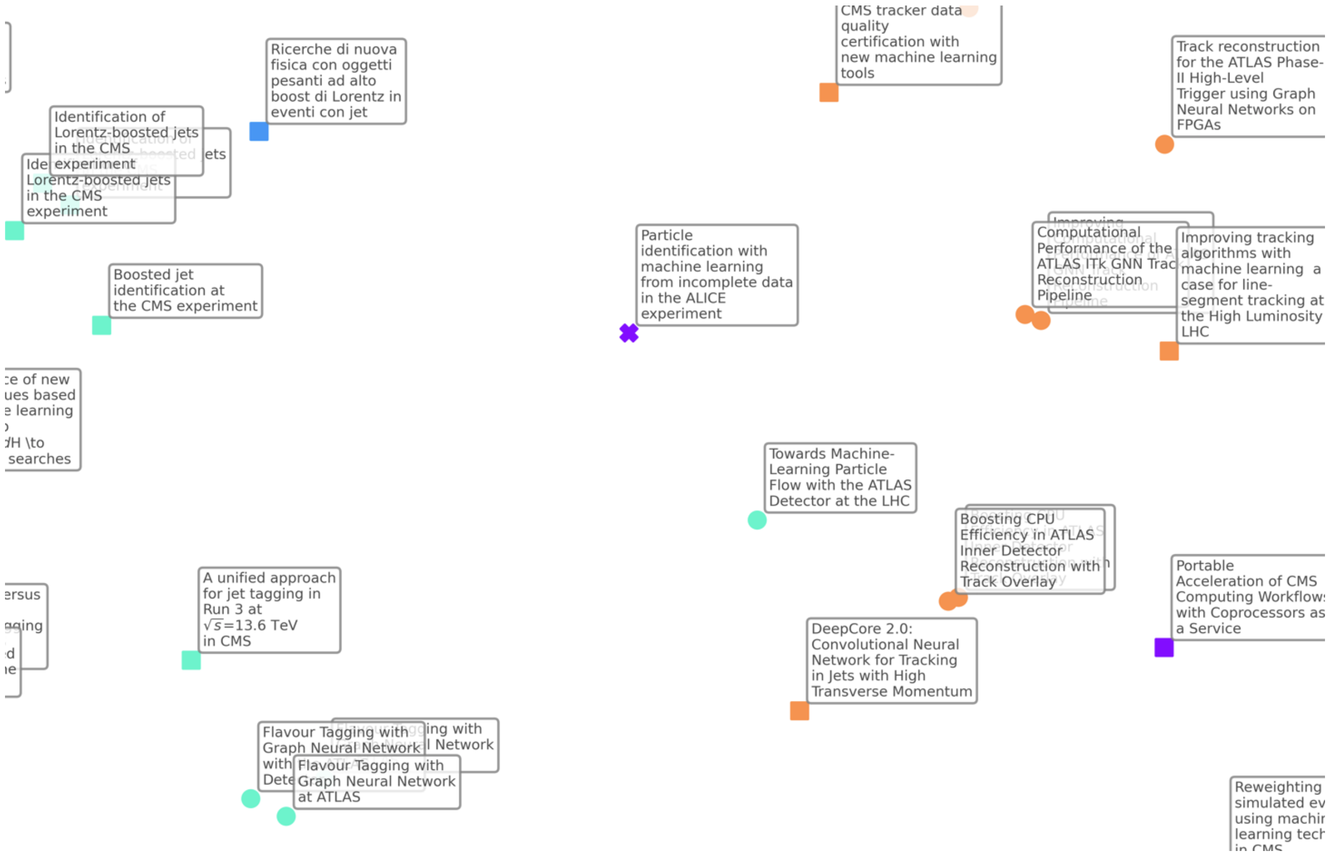
FLASH SIMULATION AT LHCB

- FlashSim at LHCb is called *LAMARR*
- In tracker, a sequence of BDTs (for ratios), NNs (for regression e.g. of efficiencies) and GANs for feature smearing
- Transformer-based GAN for ECAL generation
- 100x speed-up is *estimated* (though not yet shown) over full sim



https://cds.cern.ch/record/2875421
https://cds.cern.ch/record/2906203

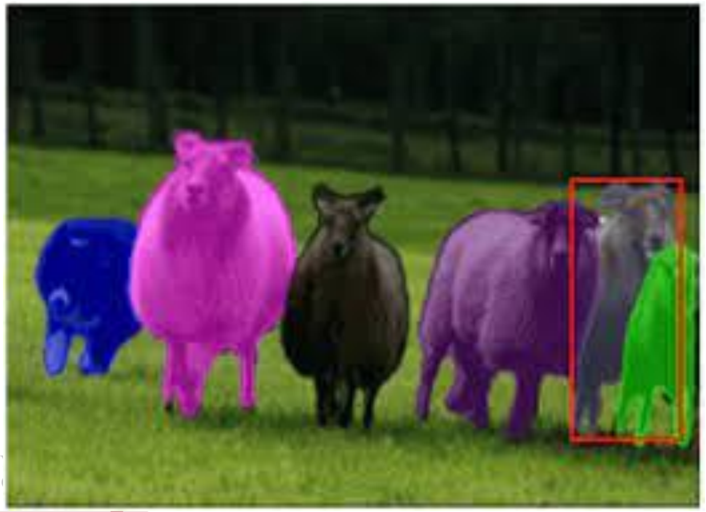
RECONSTRUCTION



SEGMENTATION

Typically two types of task:

- **Semantic segmentation** – This pixel in the image is a car
- **Instance segmentation** – This pixel in the image is sheep number 3

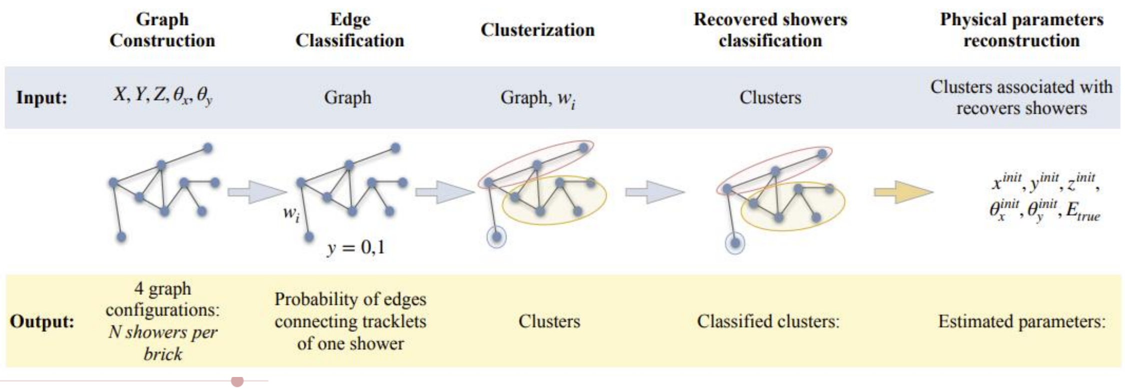


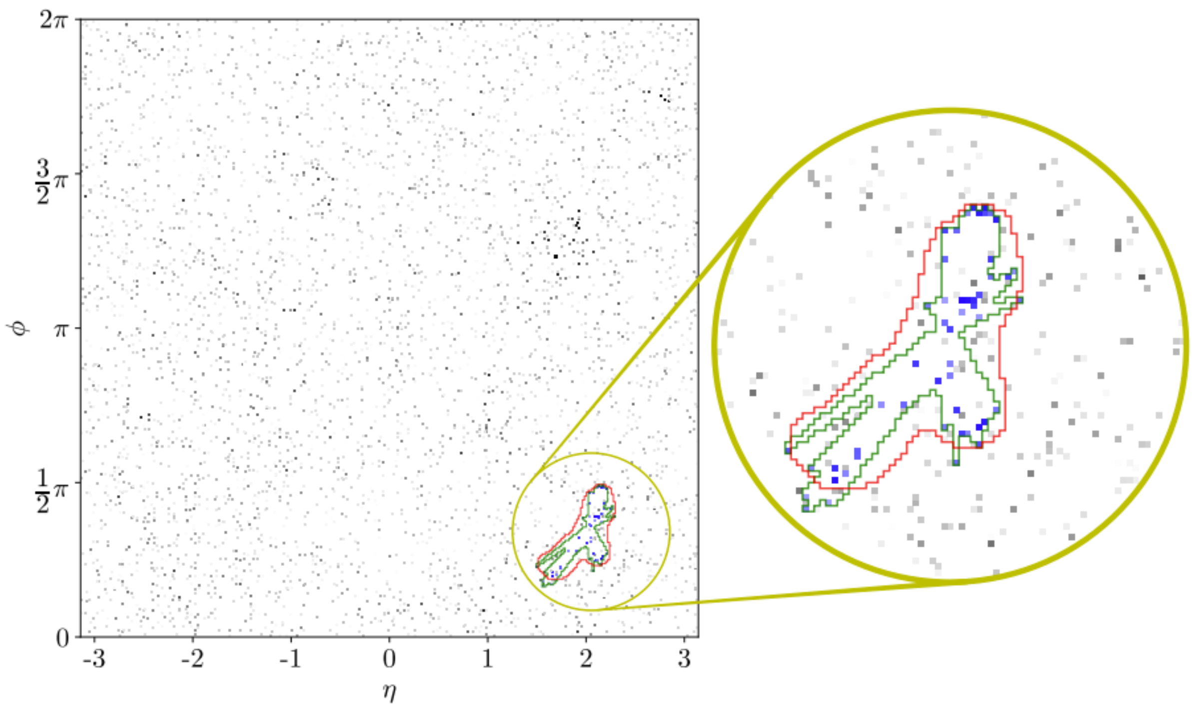
Some other exotic types of segmentation could be interesting to HEP

- E.g. **panoptic segmentation** (arXiv:1801.00868) combines semantic and instance segmentation:
- This pixel is particle 1 (instance), this pixel is particle 2 (instance), this pixel is just "noise" (semantic)

SEGMENTATION

- Has traditionally worked with image data (i.e. CNN models)
- Has been shown in reconstructing boosted Higgs jets as images
- Can also do this with GNNs – see *Belavin V et al*





Hits from Higgs + 3 QCD jets (and pileup). Higgs hits in blue.
Li J et al Arxiv: 2008.13529

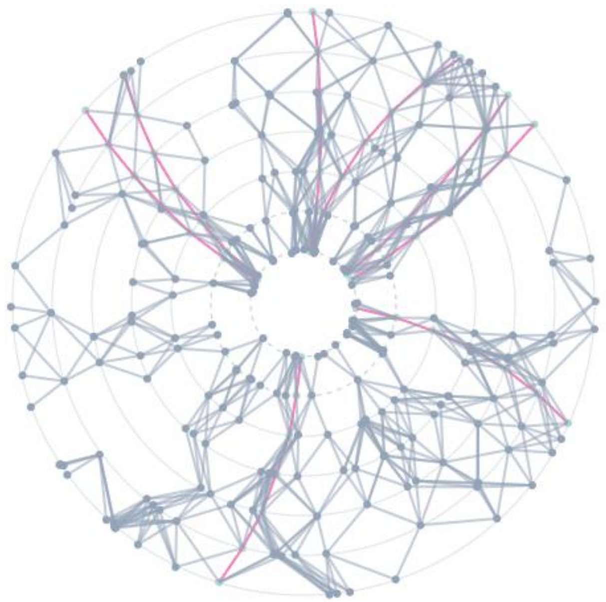
Belavin V et al, arXiv:2104.02040v6

TRACK RECONSTRUCTION WITH GRAPH NEURAL NETWORKS

Combinatorial Kalman Filter

TRACK RECONSTRUCTION WITH GRAPH NEURAL NETWORKS

Combinatorial Kalman Filter

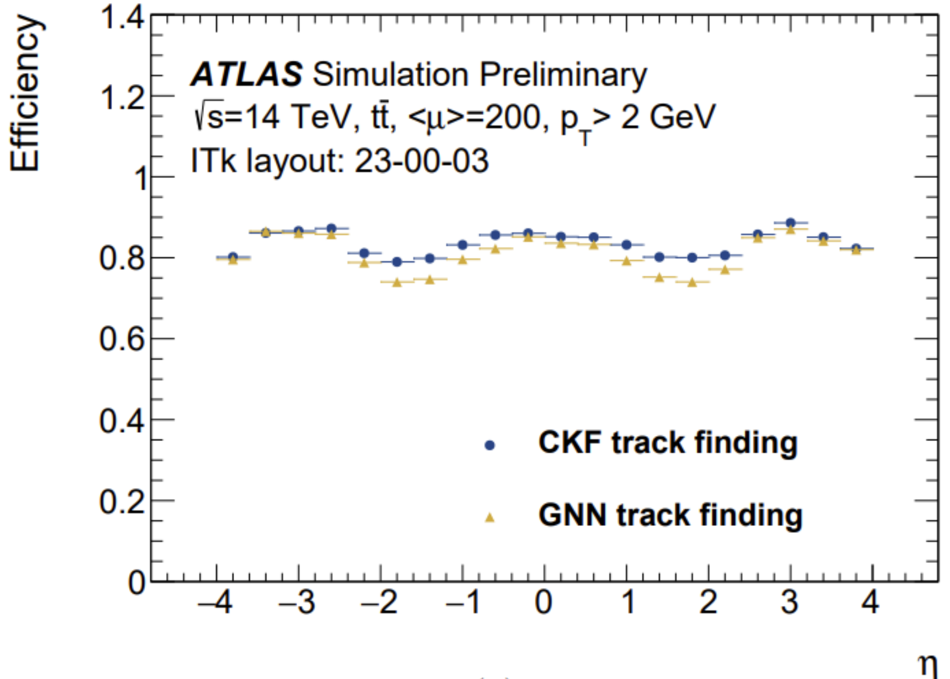


Graph Neural Network

Link

TRACKING: GNN TRACK FINDING

- Phase 2 ATLAS tracker will require fast processing of O(350,000) hits to find O(1k-2k) particles above 1GeV
- While there is work to port traditional (CKF) tracking to GPU, an ML approach would natively suit hardware acceleration
- Most mature implementation uses GNN, with physics performance competitive with CKF
- Computational performance measured at 500-700ms per event



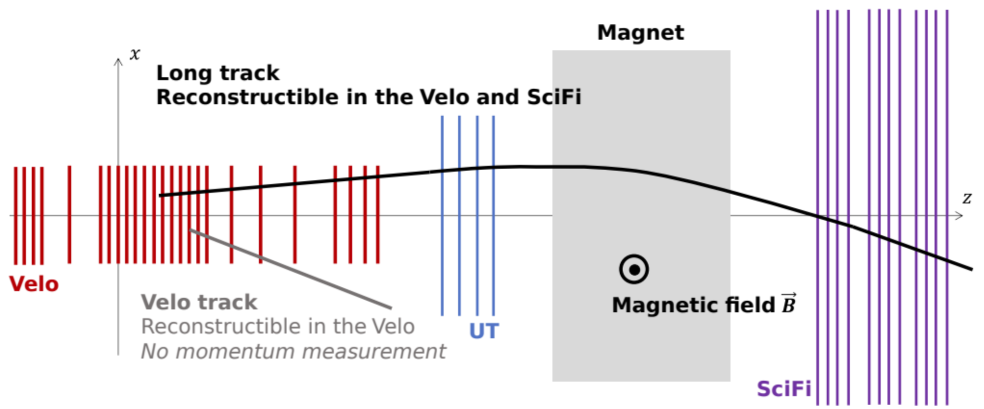
(a)

https://cds.cern.ch/record/2882507

Stage	Pipeline	
	Metric Learning (ms)	Module Map (ms)
1. Graph Construction	505	69
2. Edge Classification	108	323
3. Graph Segmentation	118	118
Sum	**731**	**510**

https://cds.cern.ch/record/2914282

TRACKING: GNN TRACK FINDING



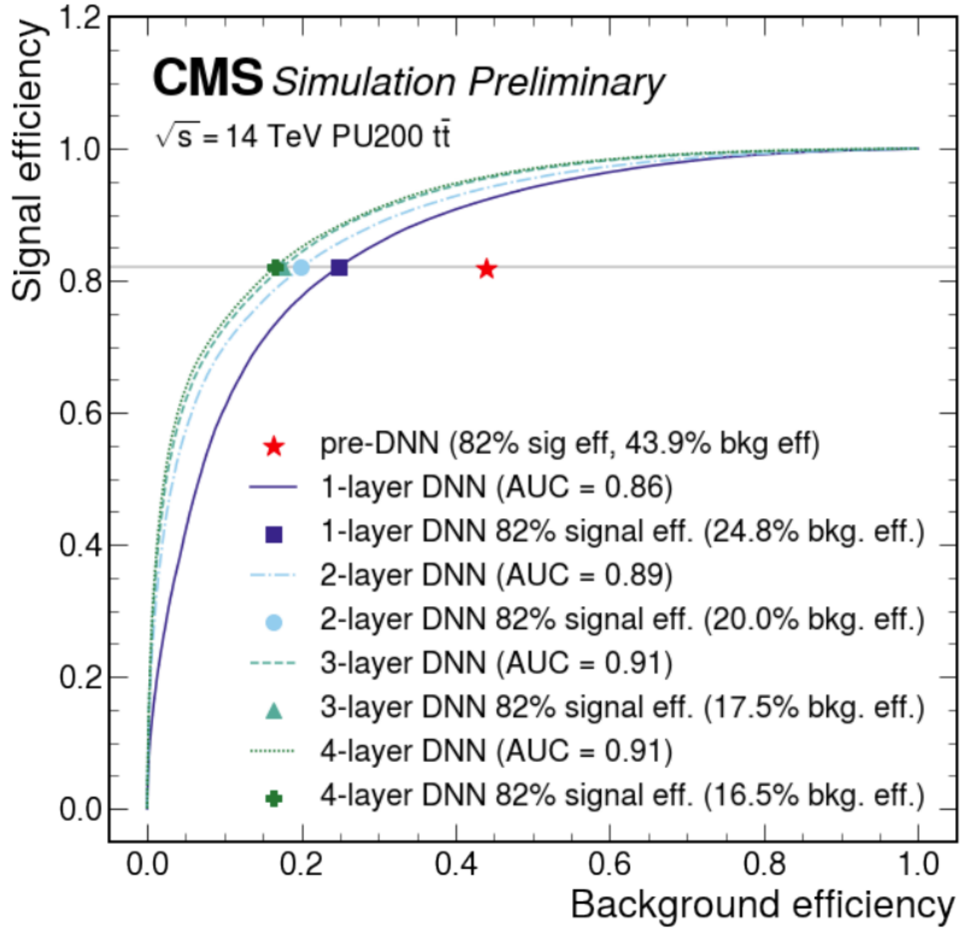
- HL-LHC LHCb detector to have new "Velo" tracker
- One project uses same approach as ATLAS GNN track finding: ETX4VELO
- Positron-electron pairs produce shared hits, and particles may leave multiple hits per layer
- Due to these complexities, new innovations are added, including GNN applied to "triplet graph"
- Almost universally improved physics performance over traditional tracking in Allen

Long category	Efficiency		Clone rate		Hit efficiency		Hit purity	
	Allen	ETX4VELO	Allen	ETX4VELO	Allen	ETX4VELO	Allen	ETX4VELO
No electrons	99.26	99.28 (99.51)	2.54	0.96 (0.89)	96.46	98.73 (98.90)	99.78	99.94 (99.94)
Electrons	97.11	98.80 (99.22)	4.25	7.42 (7.31)	95.24	96.54 (96.79)	97.11	98.46 (98.46)
From strange	97.69	97.50 (98.06)	2.50	0.92 (0.81)	97.69	98.22 (98.77)	99.34	99.68 (99.68)

https://arxiv.org/pdf/2406.12869

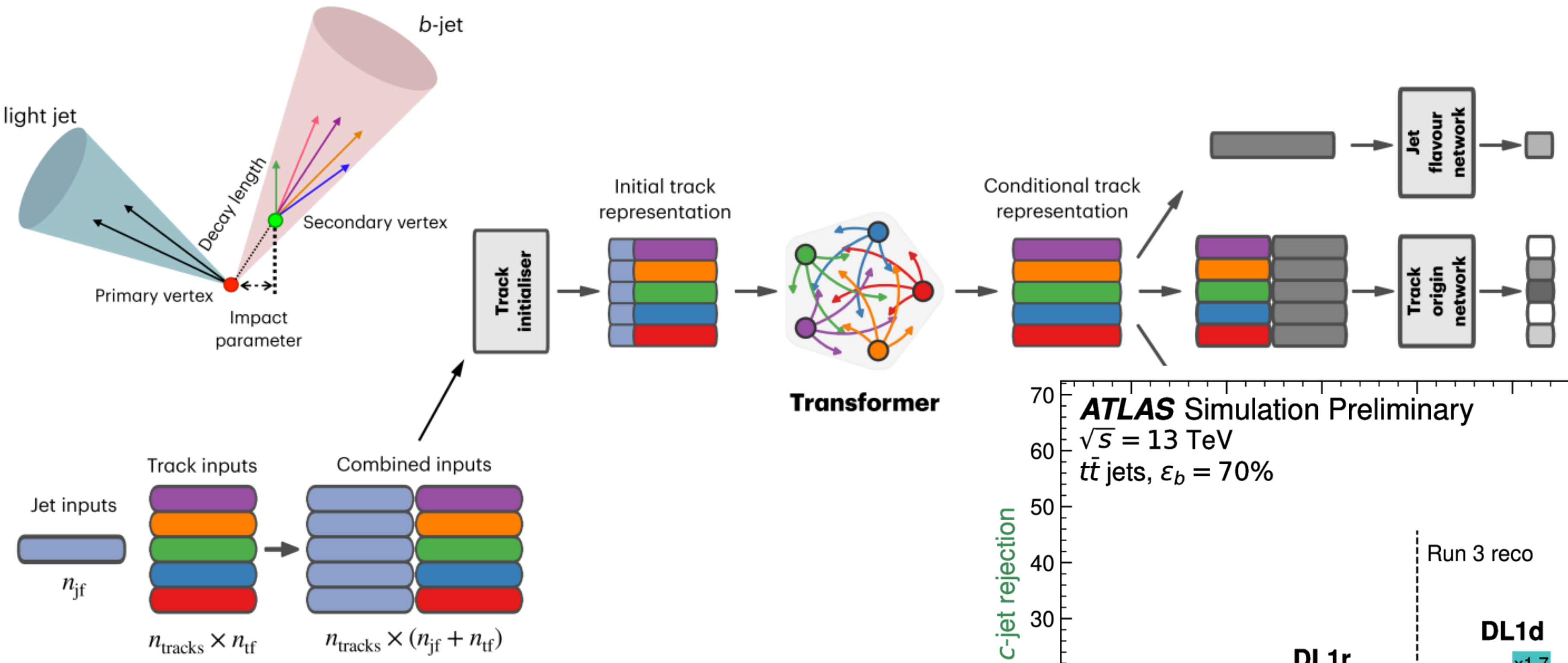
TRACKING: LINE SEGMENTS

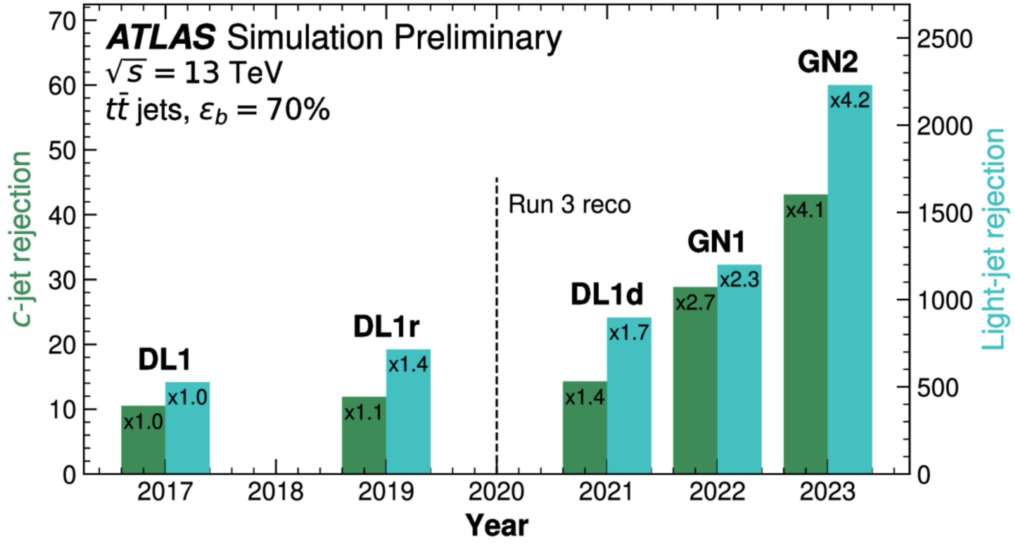
- In HL-LHC CMS tracking, seeds are built from progressively longer line segments (LS) in the outer tracker
- These LSs are based on physical heuristics, given minimum requirements of e.g. pT
- A first analysis released of ML applied to this track finding approach, shows significant reduction in fake tracks (>2x)
- Tracking efficiency of large radius tracks also boosted by 5-10%



https://cds.cern.ch/record/2882251

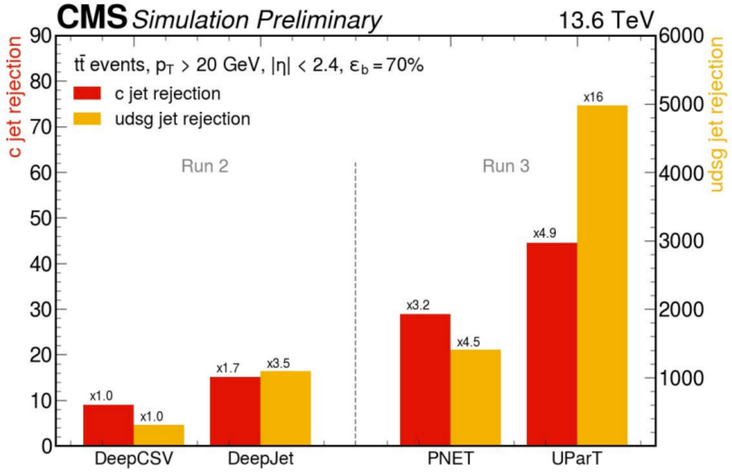
JET TAGGING WITH TRANSFORMERS



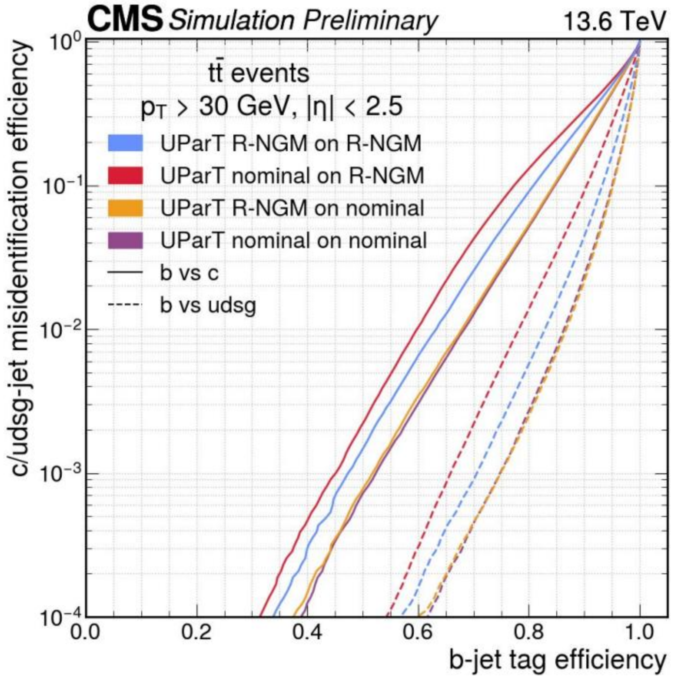


Link

TAGGING & JETS: UNIFIED TAGGING IN CMS



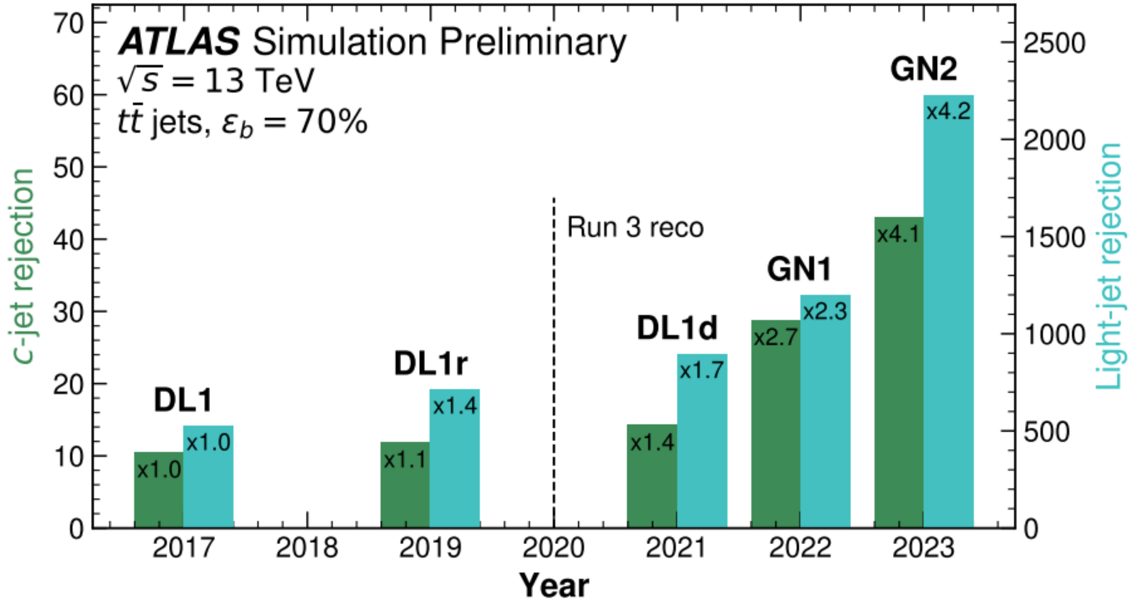
- Latest CMS jet tagging uses a form of ParticleTransformer called UParT
- Trained to predict b, c, tau, and s (for the first time) as well as energy regression and resolution quantiles – performance is significantly better than previous models (upper plot)
- UParT trained with adversarial attacks to increase robustness to mismodelling (lower plot)

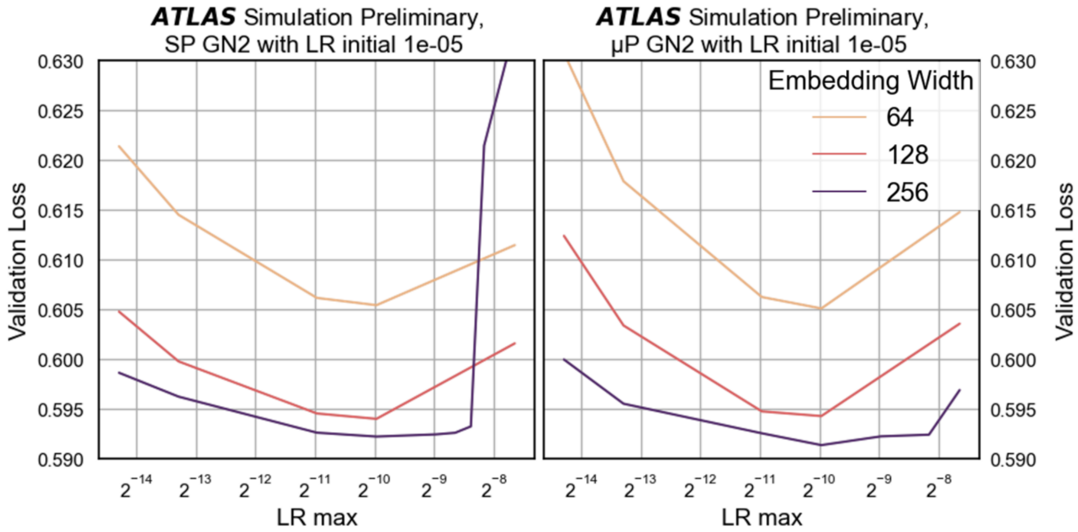


https://cds.cern.ch/record/2904702

TAGGING & JETS: UNIFIED TAGGING IN ATLAS

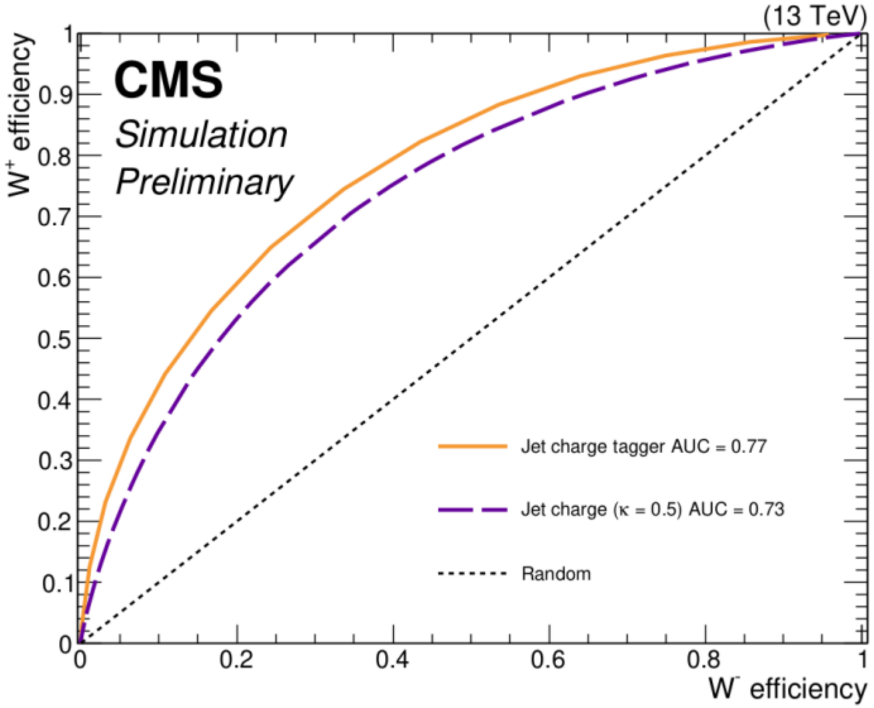
- Latest ATLAS tagger GN2 also grounded on transformer encoder
- Besides jet type classification, two further tasks performed: track origin and pairwise classification of shared track vertex
- As in UParT, dramatic improvement of GN2 over earlier taggers
- Sophisticated training with μP and μTransfer, for better weight updates at high learning rates, and better transfer of hyperparameters from small to large models, respectively





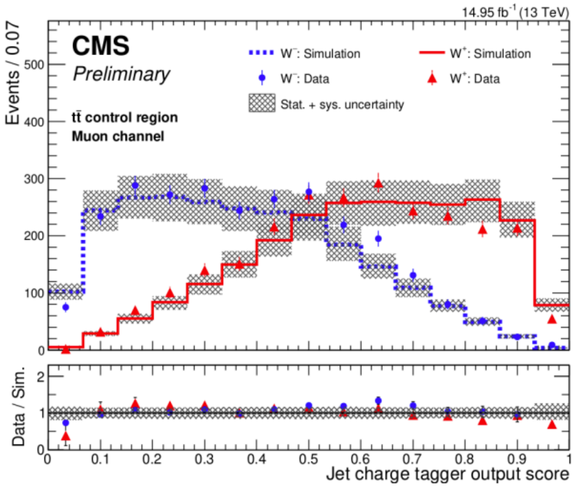
https://cds.cern.ch/record/2912358

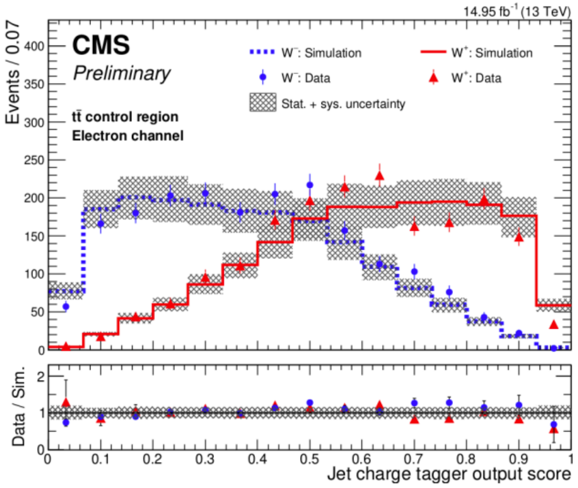
TAGGING & JETS: CHARGE CLASSIFICATION



https://cds.cern.ch/record/2904357

- W and Z jet classification appear to be a harder target, but one that CMS is studying
- Use a Dynamic GNN (i.e. the graph is constructed on-the-fly) – not UParT
- Very similar performance on data and simulation, without even adversarial training

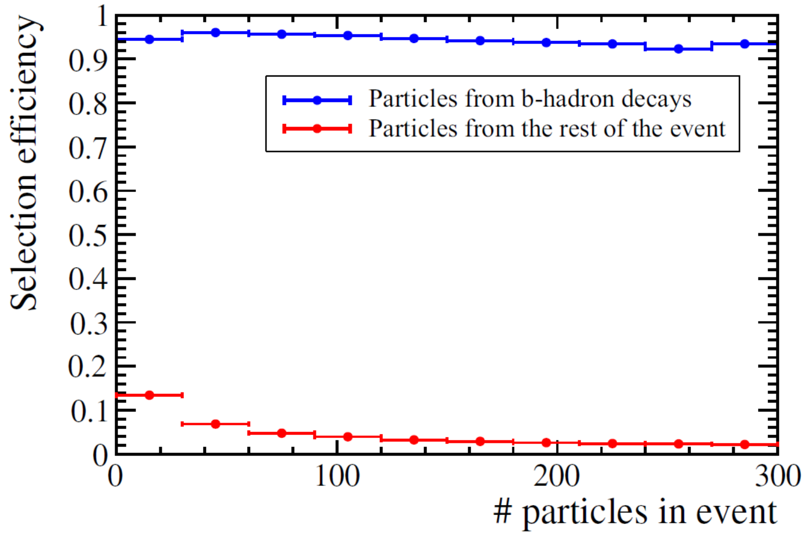




DEEP FULL EVENT INTERPRETATION

- In runs 4 and 5, LHCb aims to run triggerless, by storing only parts of each event's "full event interpretation" – a description of the decay chains of all particles
- Treats decay chain as a tree-style graph, with target of predicting the lowest common ancestor (LCA) for each leaf in the graph
- Similarities with Belle II FEI ML approaches
- By building the graph, and selecting only those branches from b-hadron decays, high efficiency preserved with significant reduction in event size

LHCb period	Num. vis. pp collisions	Num. tracks	Num. b hadrons	Num. c hadrons
Runs 1–2	~ 1	~ 50	$\ll 1$	$\ll 1$
Runs 3–4 (Upgrade I)	~ 5	~ 150	$\ll 1$	~ 1
Runs 5 (Upgrade II)	~ 50	~ 1000	~ 1	~ 5



https://link.springer.com/article/10.1007/s41781-023-00107-8

PARTICLE I.D. WITH MISSING DATA

- Particle ID in ALICE relies on many subdetectors, each of which may have inefficiency or mis-reconstruction
- Rather than throw away any data with suspected inefficiency, this work aims to train a robust particle ID network with "Feature Set Embedding" – embedding of features and their one-hot encoding
- Allows a geometric notion of "similar features" and enables missing feature inputs
- Considerable improvement in incomplete setting over standard particle ID algorithm

(a) 3 data samples with 4 attributes with different amounts of missing values.

id	momentum	TOF	TPC	TRD
1	0.1		3	
2	7	70	24	13
3		78		

(b) First particle.

key				value
1	0	0	0	0.1
0	0	1	0	3

(c) Second particle.

key				value
1	0	0	0	7
0	1	0	0	70
0	0	1	0	24
0	0	0	1	13

(d) Third particle.

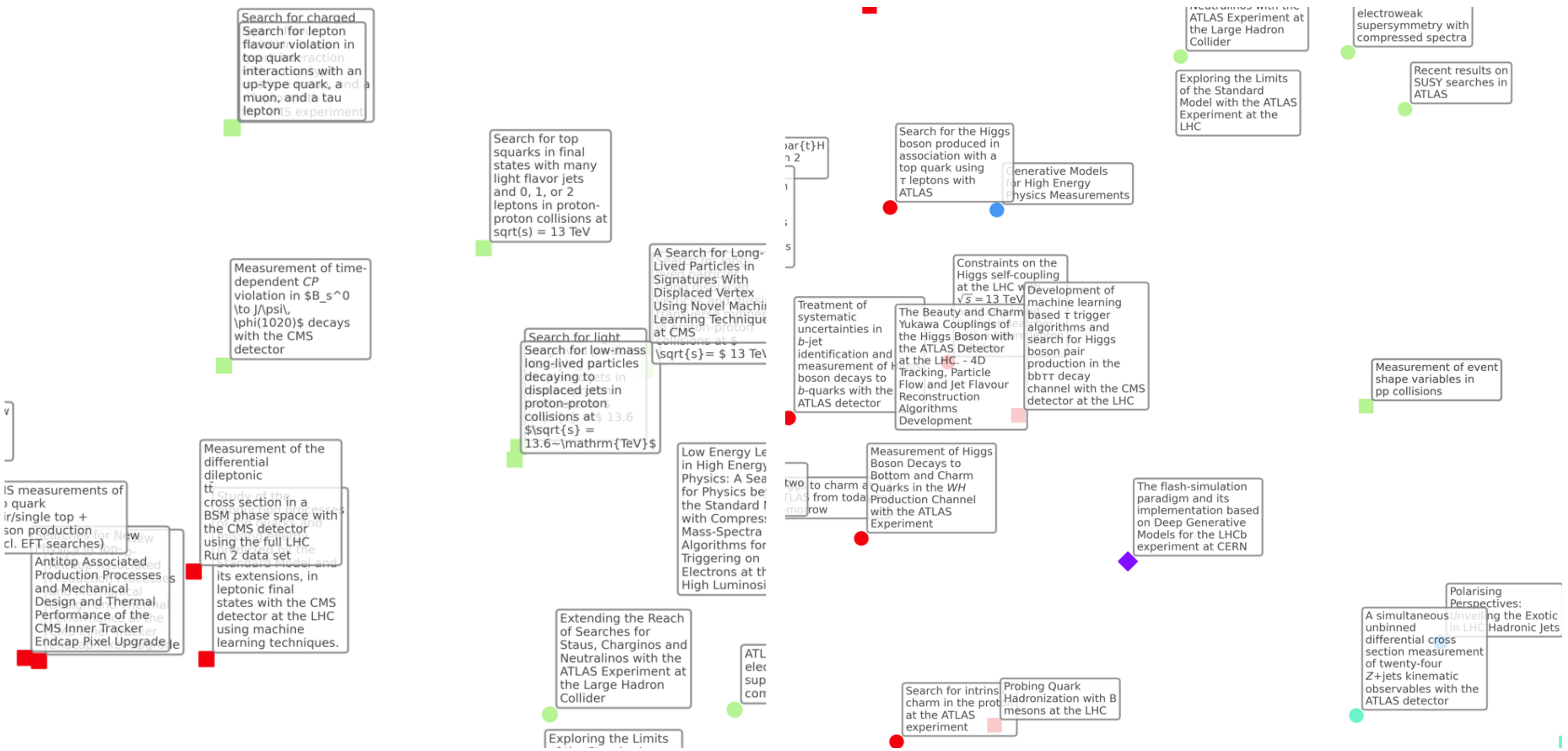
key				value
0	1	0	0	78

(d) Pion identification on data including incomplete examples.

Model	Precision	Recall	F_1
Standard	**99.99 ± 0.01**	78.37 ± 0.01	87.87 ± 0.87
Ensemble	97.47 ± 0.25	99.46 ± 0.21	98.45 ± 0.04
Mean	97.31 ± 0.07	99.52 ± 0.07	98.40 ± 0.01
Proposed	97.49 ± 0.06	**99.54 ± 0.05**	**98.50 ± 0.02**
Regression	97.33 ± 0.06	99.49 ± 0.07	98.40 ± 0.04

https://cds.cern.ch/record/2893904

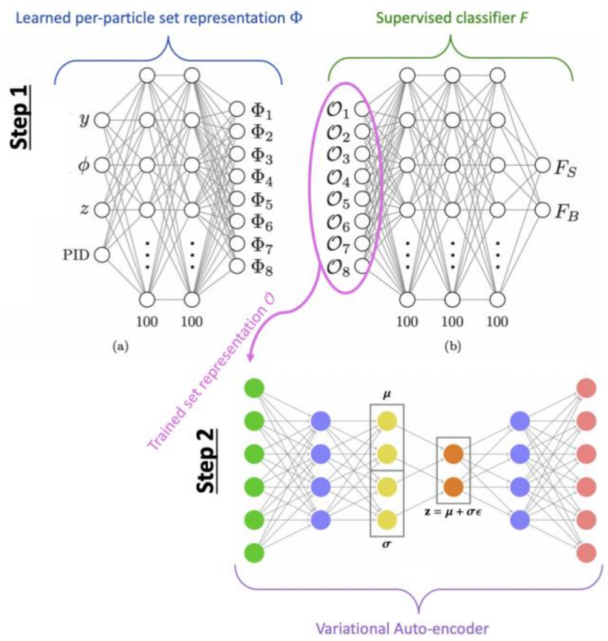
ANALYSIS

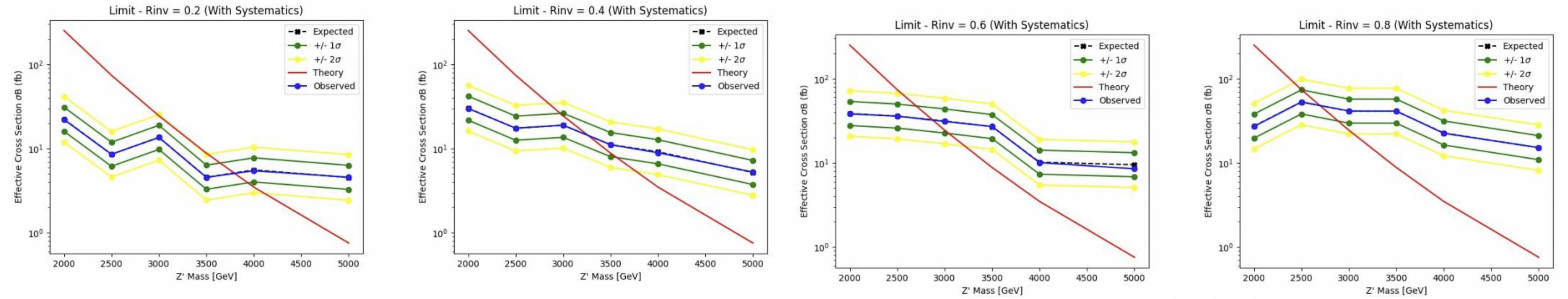


Search for lepton flavour violation in top quark interactions with an up-type quark, a muon, and a tau lepton
Search for top squarks in final states with many light flavor jets and 0, 1, or 2 leptons in proton-proton collisions at sqrt(s) = 13 TeV
Measurement of time-dependent CP violation in $B_s^0 \to J/\psi\, \phi(1020)$ decays with the CMS detector
Search for low-mass long-lived particles decaying to displaced jets in proton-proton collisions at $\sqrt{s} = 13.6~\mathrm{TeV}$
Measurement of the differential dileptonic tt cross section in a BSM phase space with the CMS detector using the full LHC Run 2 data set
Antitop Associated Production Processes and Mechanical Design and Thermal Performance of the CMS Inner Tracker Endcap Pixel Upgrade
Extending the Reach of Searches for Staus, Charginos and Neutralinos with the ATLAS Experiment at the Large Hadron Collider
Exploring the Limits
Search for the Higgs boson produced in association with a top quark using τ leptons with ATLAS
Treatment of systematic uncertainties in b-jet identification and measurement of boson decays to b-quarks with the ATLAS detector
The Beauty and Charm Yukawa Couplings of the Higgs Boson with the ATLAS Detector
Development of machine learning based τ trigger algorithms and search for Higgs boson pair production in the bbττ decay channel with the CMS detector at the LHC
Measurement of Higgs Boson Decays to Bottom and Charm Quarks in the WH Production Channel with the ATLAS Experiment
The flash-simulation paradigm and its implementation based on Deep Generative Models for the LHCb experiment at CERN
Probing Quark Hadronization with B mesons at the LHC
Exploring the Limits of the Standard Model with the ATLAS Experiment at the LHC
Recent results on SUSY searches in ATLAS
electroweak supersymmetry with compressed spectra
Measurement of event shape variables in pp collisions
A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector
Polarising Perspectives: Unveiling the Exotic Hadronic Jets

BSM SEARCHES: SEMI-SUPERVISED LEARNING

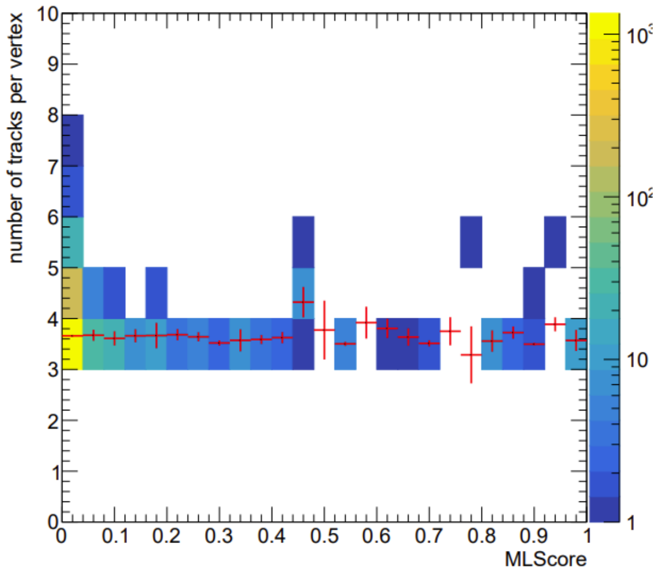
- Supervised PFN coupled with VAE applied to final encodings: *ANTELOPE* model
- Used to find possible anomalous signal regions in top mass and Z′ mass, in a dark sector "Hidden Valley" scenario
- ATLAS bump hunts in those SRs place bounds on Z′ mass of 2-5 TeV

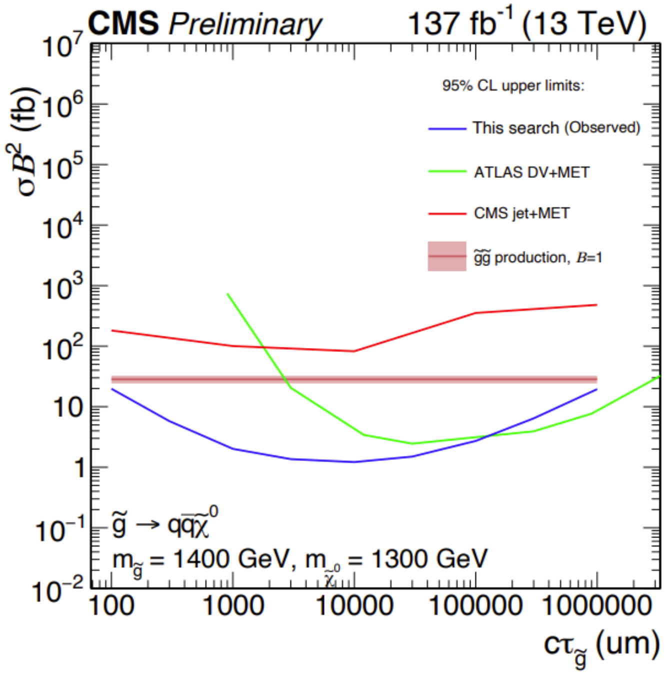




https://cds.cern.ch/record/2907718?ln=en

BSM SEARCHES: LLP LIMITS WITH GNN





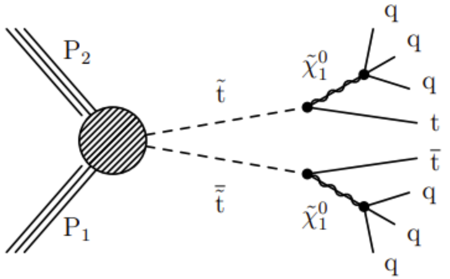
- Search for LLPs in SUSY scenario, with background rejection improved by Interaction Network
- Validation of background estimation performed with ABCDisCo, with *distance correlation* in the loss function:

$$L = L_{bce} + \lambda_{reg} \cdot L_{reg} + \lambda_{dcorr} \cdot L_{dcorr},$$

- Loss encourages *MLScore* out of the GNN to be model-independent (as seen by flat performance across number of tracks)
- Improves on previous ATLAS & CMS gluino mass exclusions

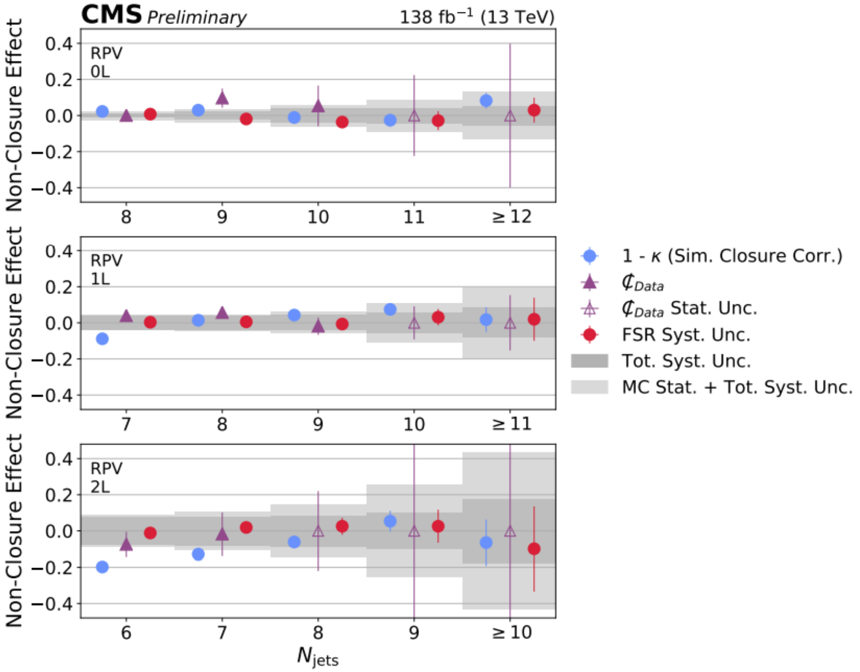
https://cds.cern.ch/record/2889341

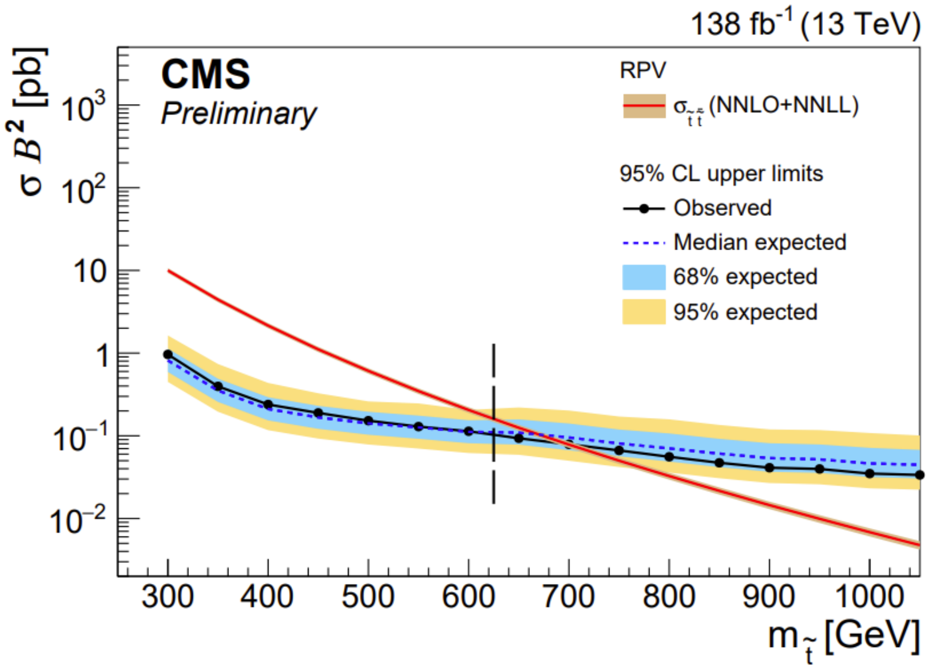
https://cds.cern.ch/record/2899862



BSM SEARCHES: RPV SUSY

- An analysis studying signals of RPV SUSY top squarks goes further in decorrelation
- Proposes ABCDisCoTEC: includes non-closure tests in the loss function
- This approach is proven to maintain the closure tests, and place new bounds on top squark masses

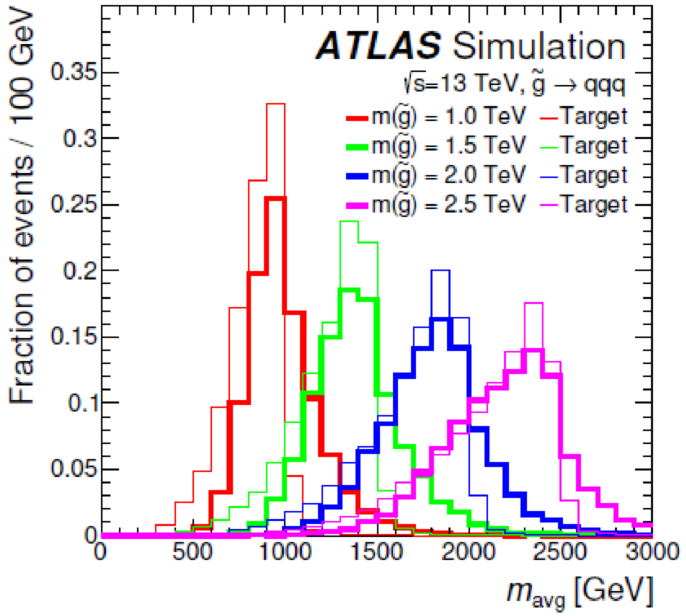


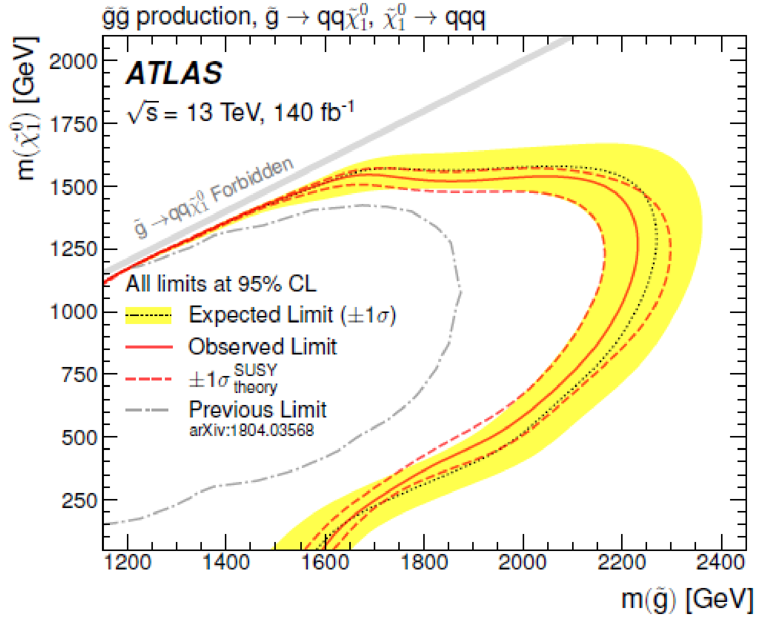


BSM SEARCHES: RPV SUSY

- Signal regions are required to have seven or more high-pT jets
- This leads to high combinatorics, so a transformer is trained to assign each jet to one of the gluino candidates (or a non-signal)
- Plot on right gives assignments passing score vs. target number of assignments across average gluino mass, with good agreement
- Jet counting approach allows strong limits to be placed on gluino mass in RPV SUSY

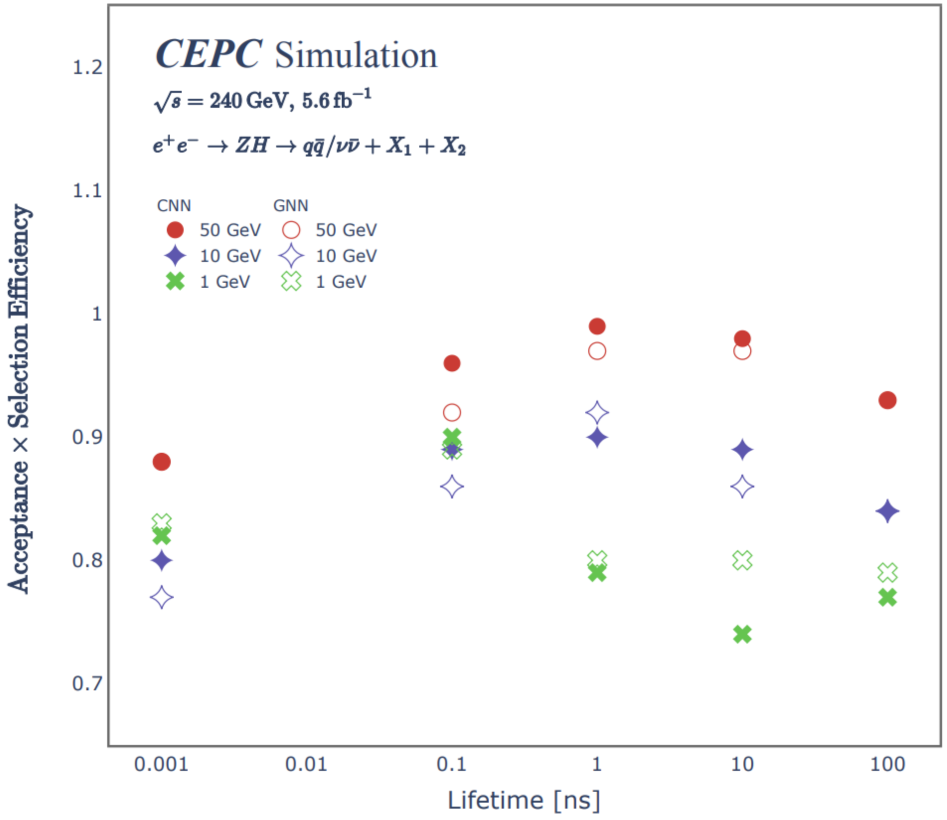
https://cds.cern.ch/record/2901186





BSM SEARCHES: HIGGS DECAYS TO LLPS

- CEPC sensitivity study, using low-level calorimeter and inner tracker features in a heterogeneous LorentzNet-like GNN
- Expected to improve branching ratio limits by three orders of magnitude compared with ATLAS+CMS (using an order of magnitude fewer Higgses)
- CEPC can run triggerless, and with ML is expected to have very high efficiency for LLP-type signatures

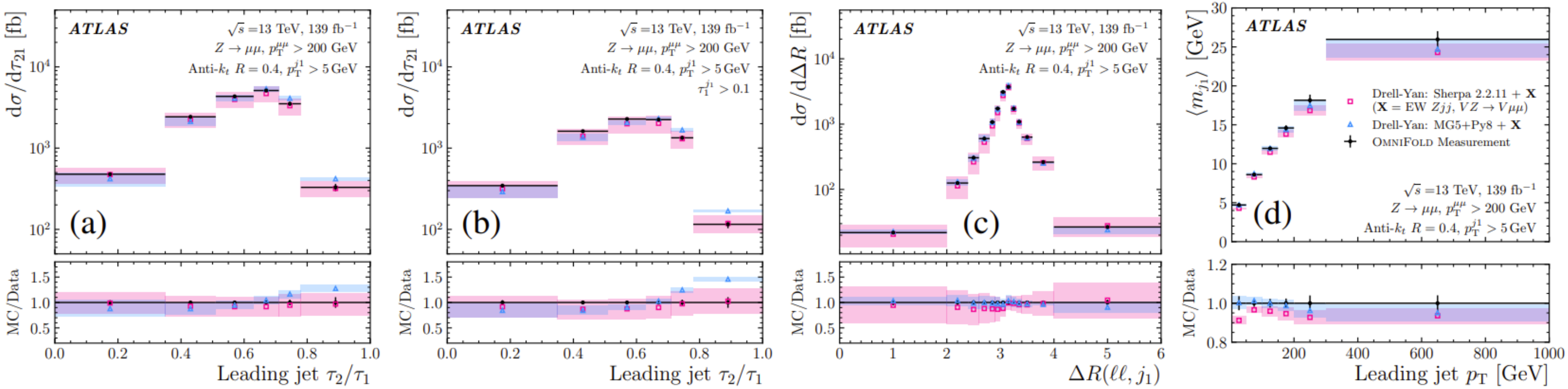


https://cds.cern.ch/record/2913609

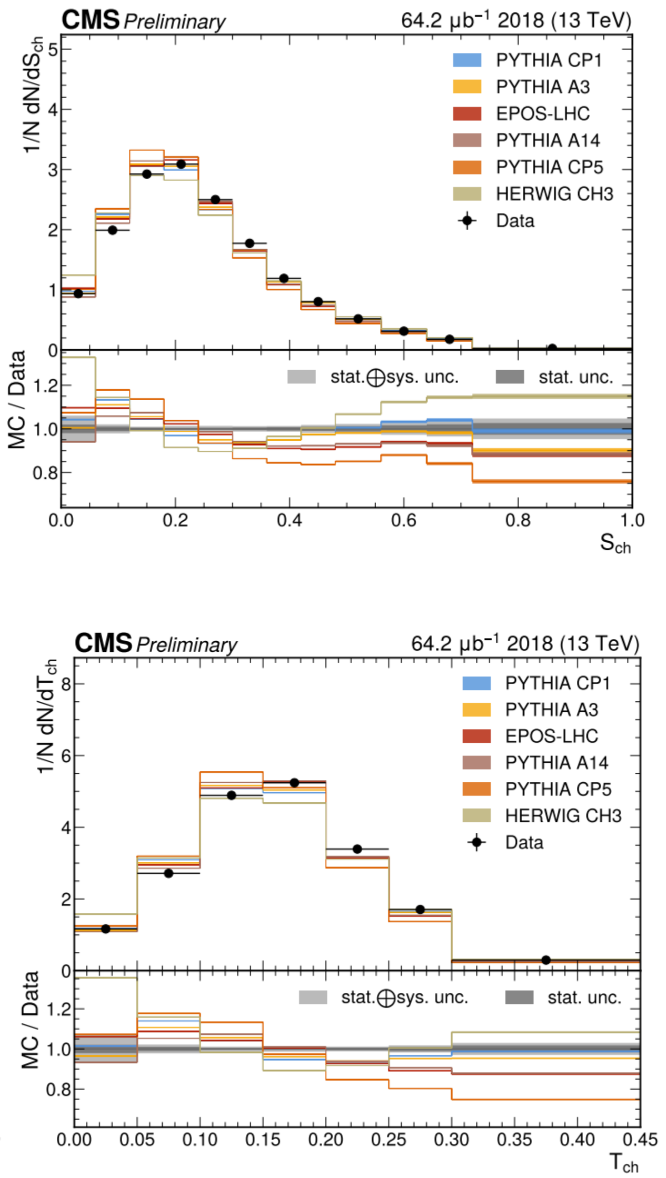
UNFOLDING: UNBINNED DIFFERENTIAL CROSS SECTIONS MEASUREMENT IN Z+JETS

https://cds.cern.ch/record/2899105

- Uses OmniFold to produce first high-dimensional (24 observables) unbinned measurement of fiducial cross sections
- ML initialization uncertainties are handled by training 100 copies of each network to take the median weight
- Statistical uncertainties are handled by perturbing input samples and re-performing full analysis, across 250 variation weights
- 25 independent components are used to calculate systematic uncertainties
- Since the "released weighting" is per-event at particle level, new observables can be created from the measured 24, e.g. the below plot shows good agreement even with derived observables



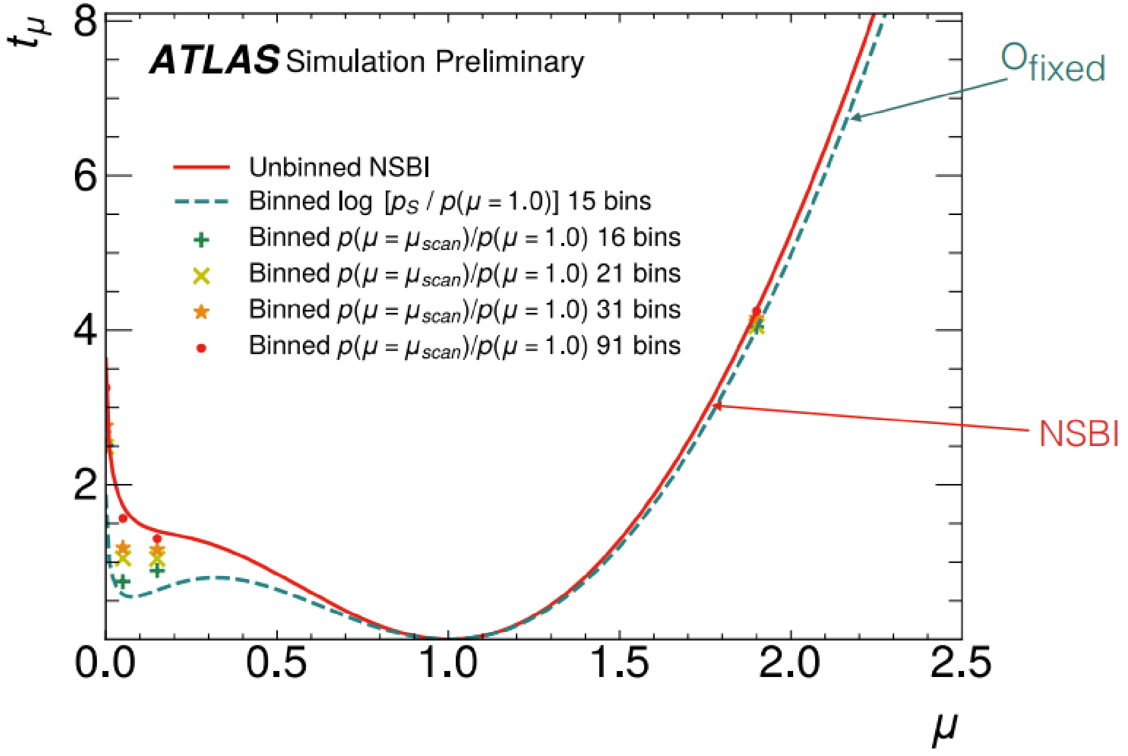
UNFOLDING: MULTIFOLDING EVENT SHAPE VARIABLES



- Event shape variables often require good modelling of non-perturbative effects – variables include sphericity, isotropy, thrust and broadening
- This study unfolded eight event shape variables in an unbinned approach with OmniFold
- Two examples here: sphericity (upper) and thrust (lower)
- Actually, no generators match all unfolded data points
- Again (as in track functions), points to specific places to work on improving modelling

https://cds.cern.ch/record/2911731

PARAMETER ESTIMATION: NEURAL SIMULATION-BASED INFERENCE

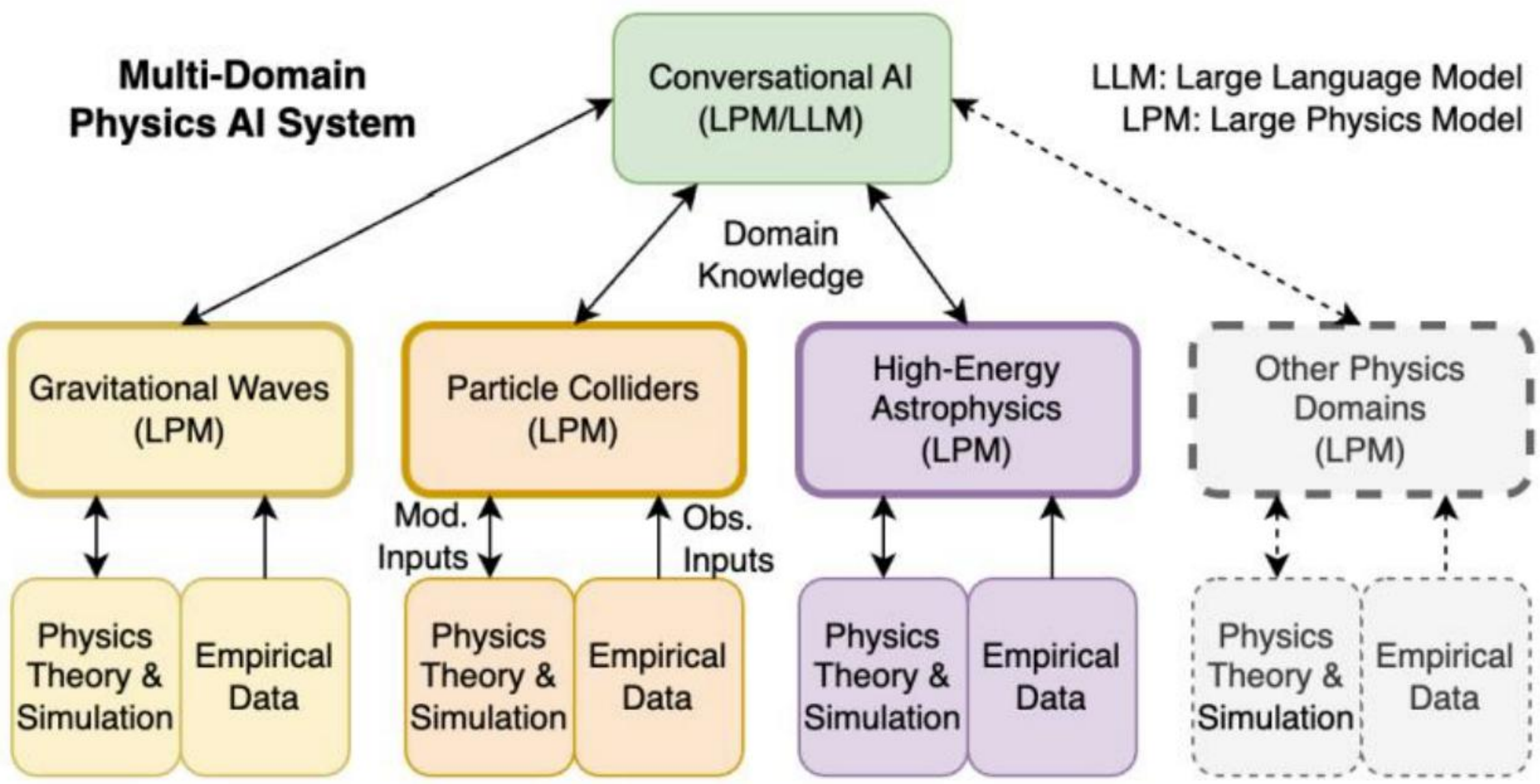


- Can apply similar classifier-based techniques as in OmniFold, but to parameter estimation – necessary in cases with quantum interference
- Neural Simulation Based Inference applied to ATLAS Higgs to 4-lepton channel
- Systematics handled with 3 trainings per nuisance parameter
- At small parameter differences, quantum interference important, so NSBI has higher test sensitivity than binned approaches

https://cds.cern.ch/record/2915357

OUTRO

FOUNDATION MODELS: PARTICLE LANGUAGE MODELS?



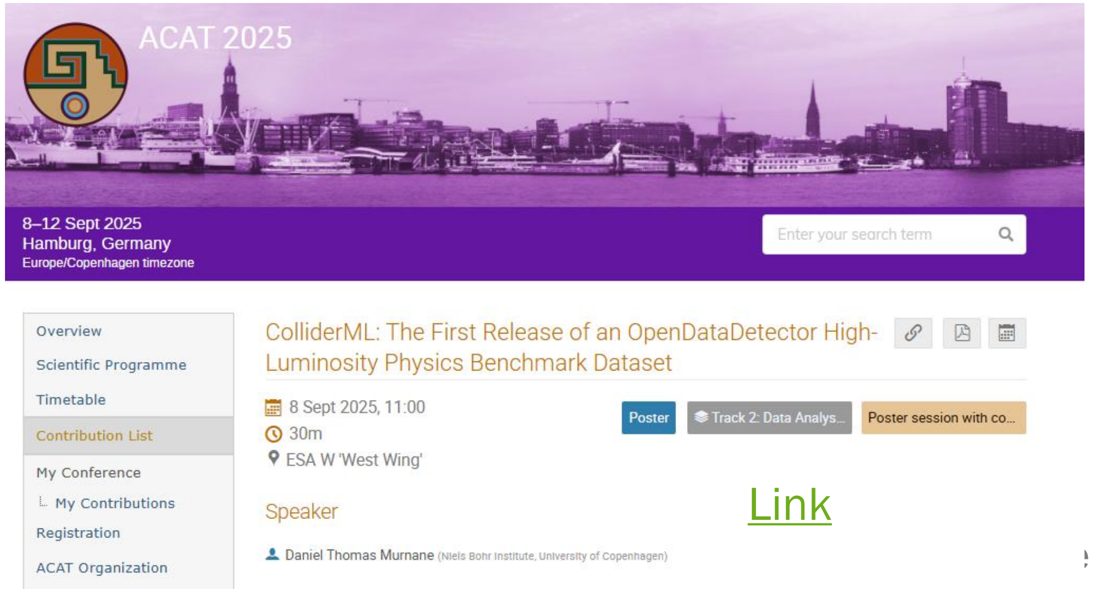
Link

BUT WE ARE MISSING OPEN DATA

TrackML
(2019)
100Gb

JetClass
(2022)
400Gb

ColliderML
(2026)
100Tb



[Link]

CONVERGENCES & DIVERGENCES

- Simulation is clearly converging towards a need for ML, with ATLAS and LHCb converging on GANs and CMS diverging towards normalizing flows
- Reconstruction generally converging towards GNNs for tracking and transformers for jet physics, in particular multi-tasking and multi-modal models that can handle uncertainties and mismodelling
- Self-supervised models are emerging in all experiments, either for scouting trigger channels, or for defining signal regions. Weakly-supervised methods already converged to ideas such as ANODE, but newer self-supervised models not yet converged on a best path (supervised task with VAE in latent space, masked prediction a la foundation model, something else?)
- All experiments converging to early use of advanced ML in analysis (e.g. transformers for jet matching), but diverging on *how* it is used
- Unfolding converging towards unbinned (in particular the OmniFold formalism), with more than one experiment discovering modelling discrepancies with this approach

NEW & OLD CHALLENGES

- Simulation: Solved as a patchwork. New challenge may be one-shot conditional generator-to-simplified-AOD
- Decorrelations and model-independence: An old challenge that actually now seems to have some serious tools to tackle (DisCo(TEC), adversarial attacks, feature pair embeddings)
- Oversampling, undersampling, SMOTE, weighting: Handling class imbalances and event weighting seems to be a mostly solved challenge. There are more sophisticated ideas appearing such as SMOTE, which open the question back up a little
- Quantifying uncertainty: I'm not convinced that any of the highlights in this talk have "solved" uncertainty quantification. If anything, this will get harder and harder as models get more sophisticated
- Sharing data, models and weights in an unbinned era: As we (hopefully!) move closer to sharing lower-level features, unbinned observables, even encoded pieces of events, we may need to re-think our data storage and sharing infrastructure