Application of Machine Learning to Event Reconstruction & Analysis

Benjamin Nachman

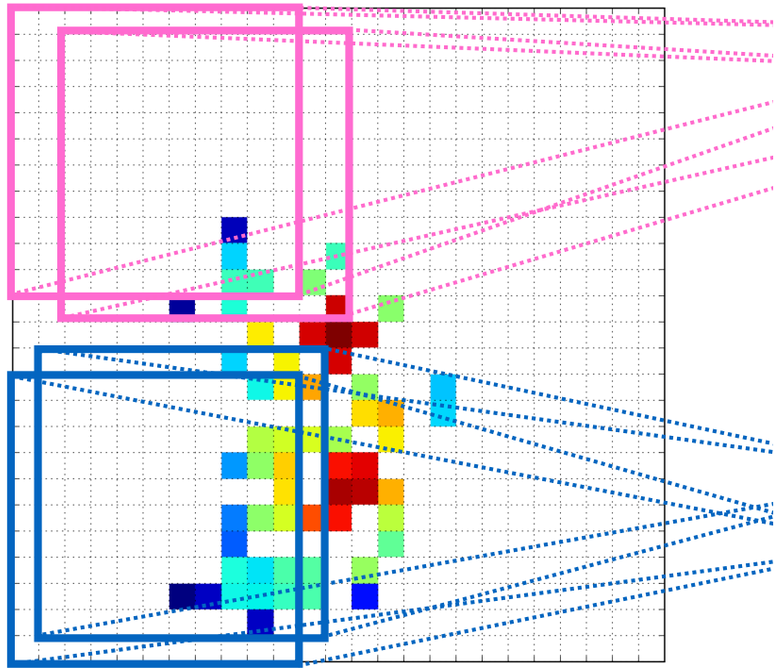
Lawrence Berkeley National Laboratory

bpnachman.com

bpnachman@lbl.gov

@bpnachman

bnachman



BERKELEY LAB

BIDS

BERKELEY INSTITUTE FOR DATA SCIENCE

HKUST IAS Workshop

January 14, 2022

Disclaimer

I have been asked to present a
“Application of Machine Learning to
Event Reconstruction & Analysis”

Disclaimer

I have been asked to present a
“Application of Machine Learning to
Event Reconstruction & Analysis”

This is a very broad topic and there is
no way I can do it justice in ~25 min.

Disclaimer

I have been asked to present a
"Application of Machine Learning to
Event Reconstruction & Analysis"

This is a very broad topic and there is
no way I can do it justice in ~25 min.

I will use the recent ML4Jets workshop
as a roadmap for this talk.

Disclaimer

I have been asked to present a
"Application of Machine Learning to
Event Reconstruction & Analysis"

This is a very broad topic and there is
no way I can do it justice in ~25 min.

I will use the recent ML4Jets workshop
as a roadmap for this talk.

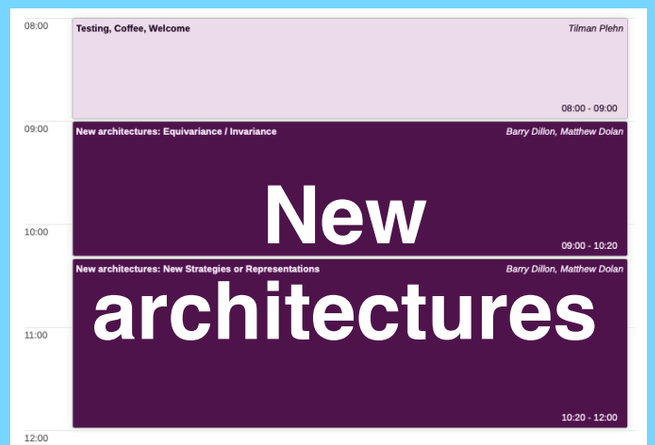
For a comprehensive review (anyone can contribute!) see:
https://iml-wg.github.io/HEPML-LivingReview/

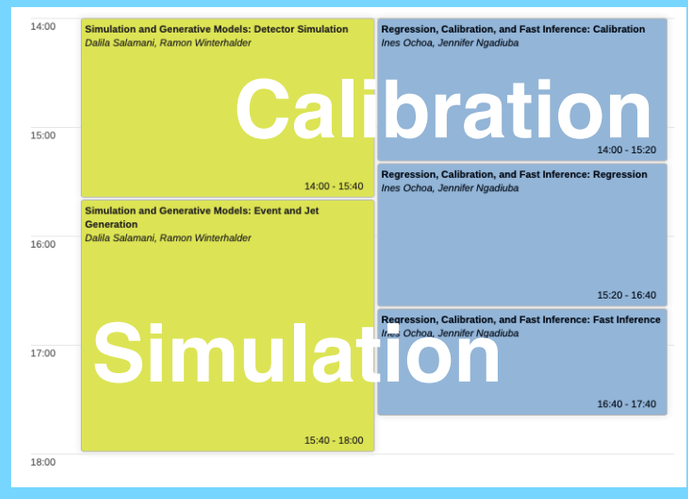
ML4Jets2021

The annual ML4Jets conference some months ago had 100 talks in three days (!)

N.B. most plots are links!

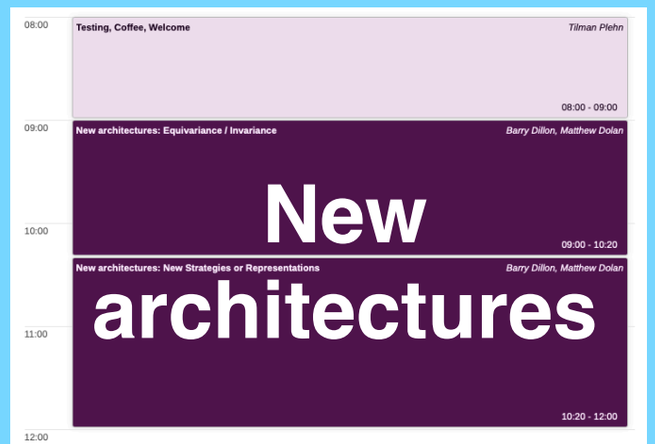
ML4Jets2021

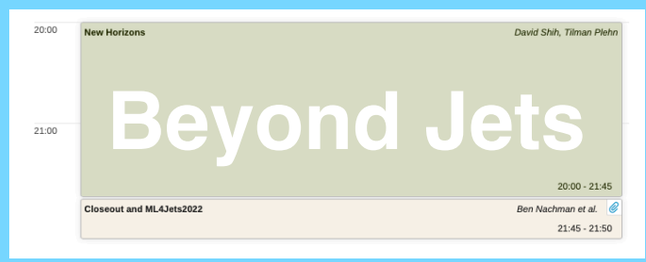




ML4Jets2021

I won't cover everything - just giving you a taste!





…my apologies in advance for not covering your / favorite topic.

New architectures

A **hot topic** in this area is **equivariance** / **invariance**

A NN is **equivariant** if it **commutes** with the symmetry group and a NN is **invariant** if the output is **unchanged** under symmetries of the inputs

New architectures

A **hot topic** in this area is **equivariance** / **invariance**

A NN is **equivariant** if it **commutes** with the symmetry group and a NN is **invariant** if the output is **unchanged** under symmetries of the inputs

Equivariant example

Learn features that transform under rotations in the same way as the inputs - then feed these into further layers

e.g. train a NN that takes as input all constituents inside a jet and outputs the true jet 3-vector.

see e.g. E. Catalina's ML4Jets talk.

New architectures

A **hot topic** in this area is **equivariance** / **invariance**

A NN is **equivariant** if it **commutes** with the symmetry group and a NN is **invariant** if the output is **unchanged** under symmetries of the inputs

Equivariant example

Learn features that transform under rotations in the same way as the inputs - then feed these into further layers

Invariant example

Event/jet constituents are permutation invariant - use Deep Sets, Graph Networks, Transformers, Attention, …

New architectures

A **hot topic** in this area is **equivariance** / **invariance**

A NN is **equivariant** if it **commutes** with the symmetry group and a NN is **invariant** if the output is **unchanged** under symmetries of the inputs

e.g. Deep Sets:

$$f(x_1, \ldots, x_n) = F\left(\sum_{i=1}^{N} \Phi(x_i)\right)$$

for permutation invariance

see 1810.05165

Invariant example

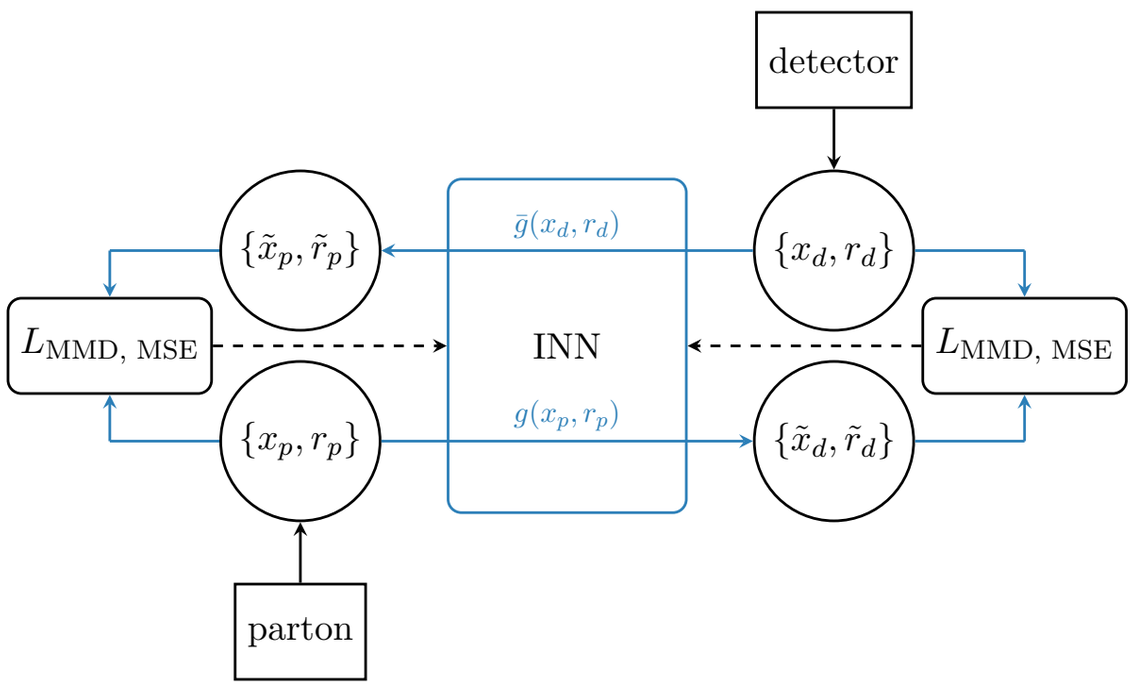
*Event/jet constituents are permutation invariant - use **Deep Sets**, Graph Networks, Transformers, Attention, …*

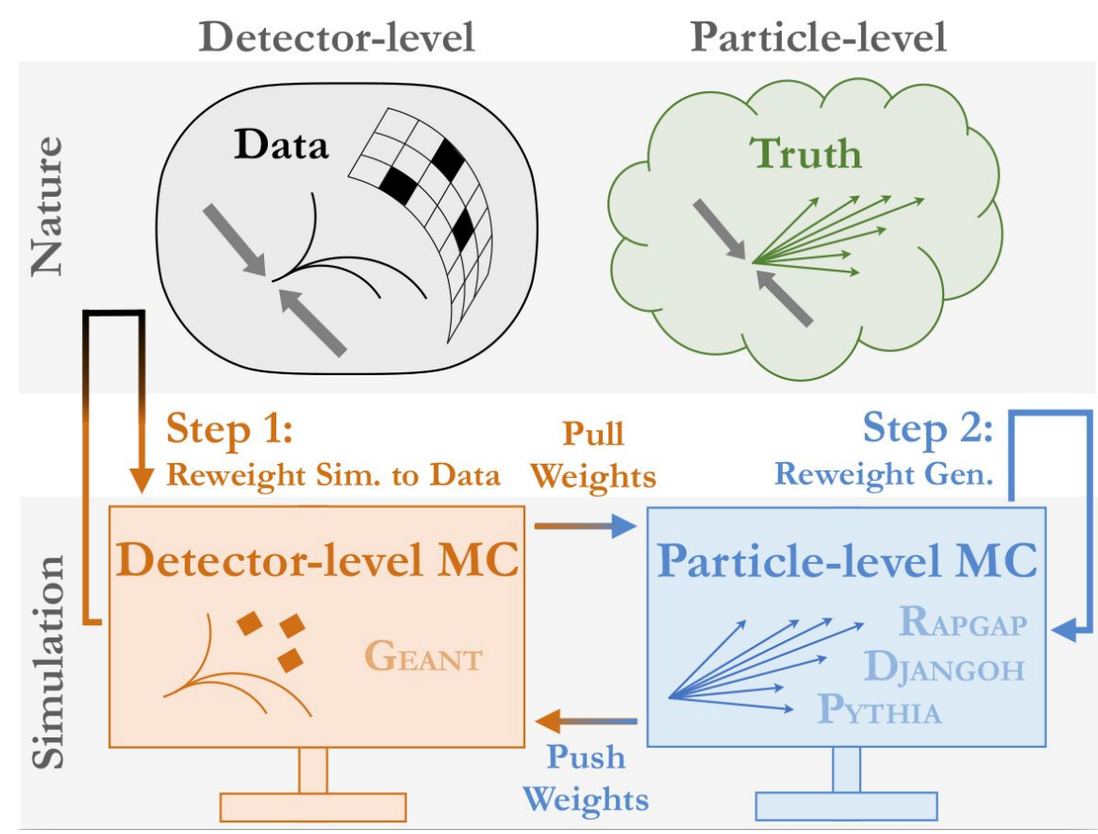
Measurements

Significant interest in ML-based unfolding for high-(or even variable-)dimensional, unbinned measurements

Measurements

Significant interest in ML-based unfolding for high-(or even variable-)dimensional, unbinned measurements





Generative model-based
2006.06685

Classifier-based
1911.09107

Measurements

Significant interest in ML-based unfolding for high-(or even variable-)dimensional, unbinned measurements

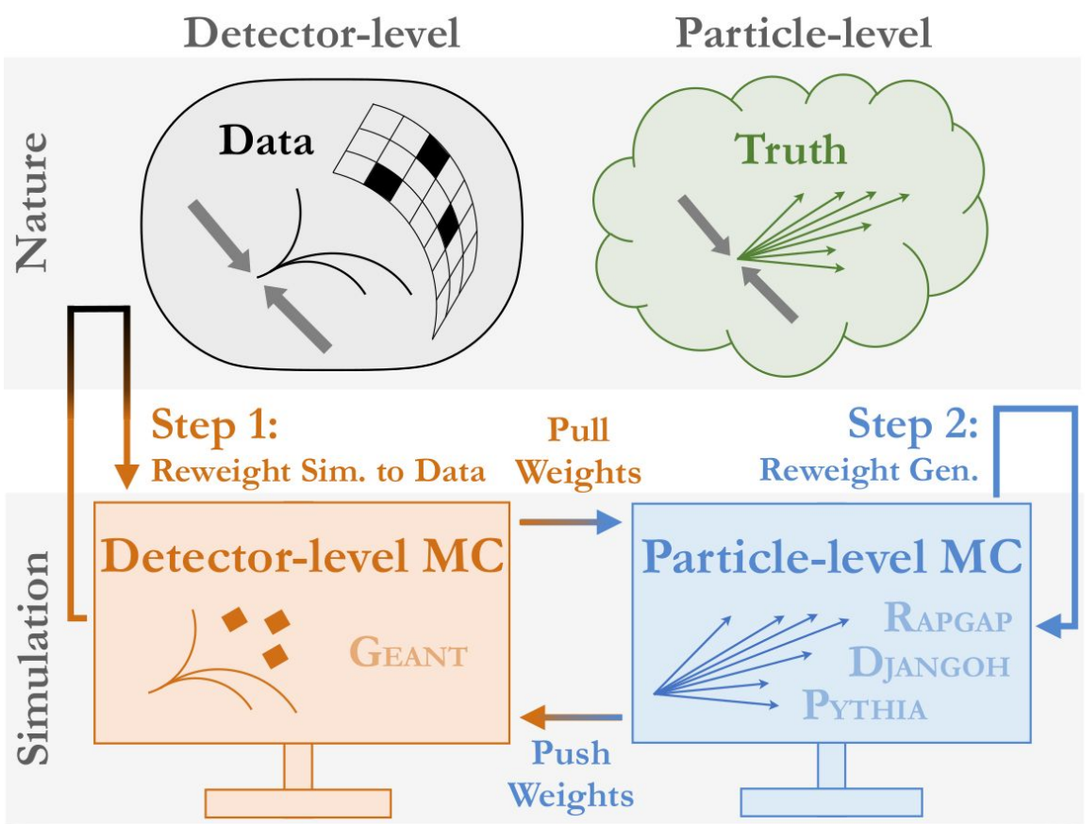
Fact: Neutral networks learn to approximate the likelihood ratio

or something monotonically related to it in a known way

Solution: train a neural network to distinguish the two datasets!

This turns the problem of **density estimation** (**hard**) into a problem of **classification** (**easy**)

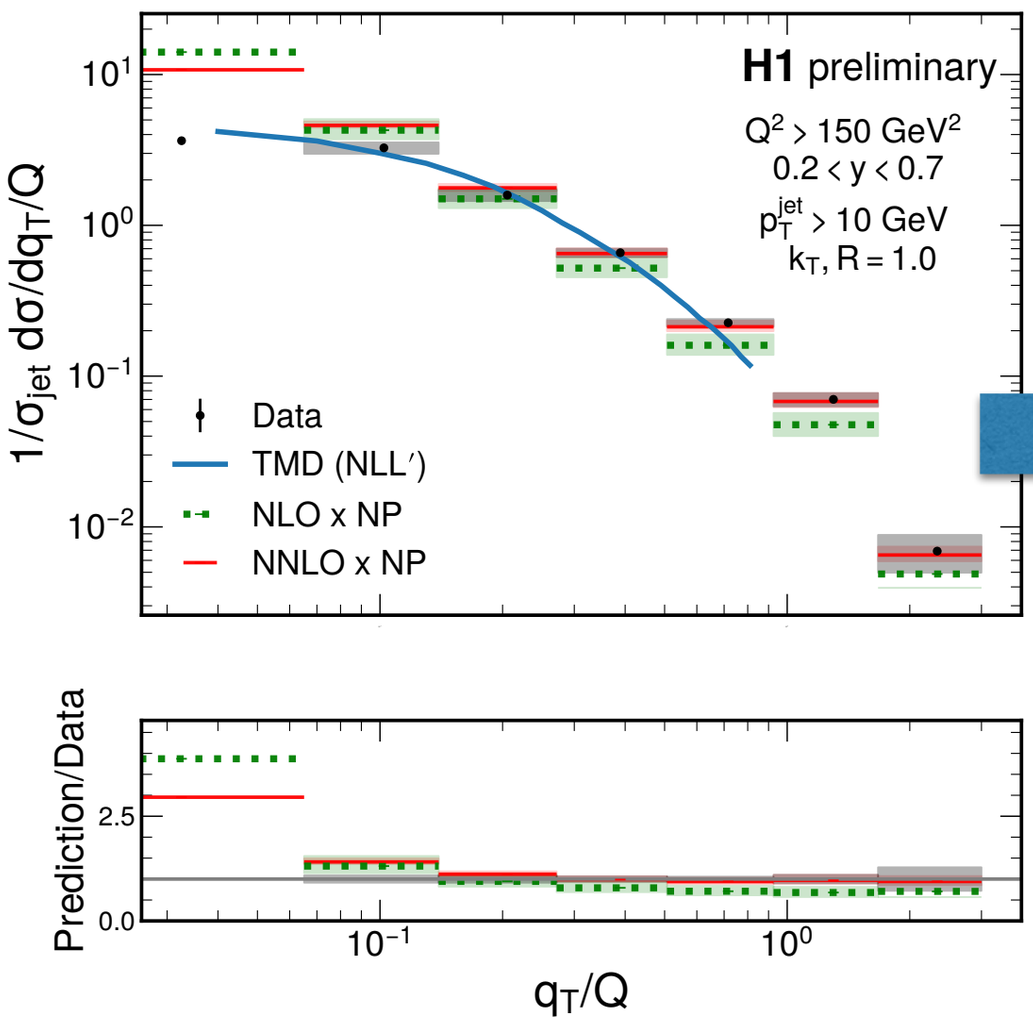
(this is a form of **likelihood-free inference**)



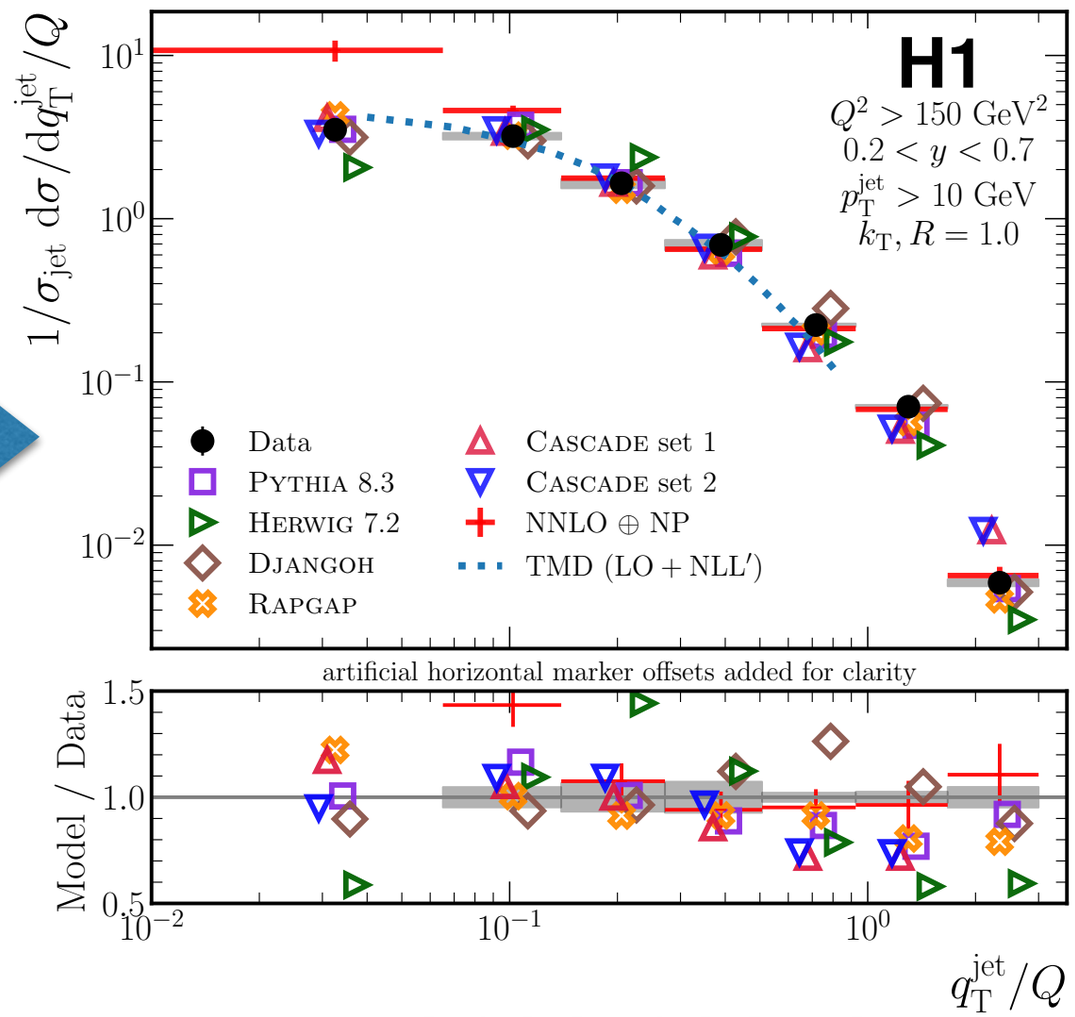
Classifier-based
1911.09107

Measurements

First application to collider data!



H1prelim-21-031



2108.12376
(very recent!)

1911.09107

Unbinned measurements - future

Publishing unbinned measurements is tricky - we have started a conversation about this in a recent paper. Feedback is most welcome!

2109.13243

Presenting Unbinned Differential Cross Section Results

Miguel Arratia,[a,b] Anja Butter,[c] Mario Campanelli,[d] Vincent Croft,[e] Dag Gillberg,[f] Kristin Lohwasser,[g] Bogdan Malaescu,[h] Vinicius Mikuni,[i] Benjamin Nachman,[j,k] Juan Rojo,[l,m] Jesse Thaler,[n,o] Ramon Winterhalder[p]

[a] *Department of Physics and Astronomy, University of California, Riverside, CA 92521, USA*
[b] *Thomas Jefferson National Accelerator Facility, Newport News, VA 23606, USA*
[c] *Institut für Theoretische Physik, Universität Heidelberg, Heidelberg, Germany*
[d] *University College London, London WC1E 6BT, UK*
[e] *Department of Physics and Astronomy, Tufts University, Boston, MA 02155, USA*
[f] *Department of Physics, Carleton University, Ottawa ON K1S 5B6, Canada*
[g] *University of Sheffield, Sheffield, S10 2TN, UK*
[h] *LPNHE, Sorbonne Université, Université de Paris, CNRS/IN2P3, Paris, France*
[i] *National Energy Research Scientific Computing Center, Berkeley, CA 94720, USA*
[j] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[k] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[l] *Nikhef Theory Group, Science Park 105, 1098 XG Amsterdam, The Netherlands*
[m] *Department of Physics and Astronomy, Vrije Universiteit Amsterdam, NL-1081 HV Amsterdam, The Netherlands*
[n] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[o] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[p] *Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium*

E-mail: bpnachman@lbl.gov

Abstract: Machine learning tools have empowered a qualitatively new way to perform differential cross section measurements whereby the data are unbinned, possibly in many dimensions. Unbinned measurements can enable, improve, or at least simplify comparisons between experiments and with theoretical predictions. Furthermore, many-dimensional measurements can be used to define observables after the measurement instead of before. There is currently no community standard for publishing unbinned data. While there are also essentially no measurements of this type public, unbinned measurements are expected in the near future given recent methodological advances. The purpose of this paper is to propose a scheme for presenting and using unbinned results, which can hopefully form the basis for a community standard to allow for integration into analysis workflows. This is foreseen to be the start of an evolving community dialogue, in order to accommodate future developments in this field that is rapidly evolving.

BSM

2112.03769

18

This is a topic that can fill an entire review!

Machine Learning in the Search for New Fundamental Physics

Georgia Karagiorgi,[1,*] Gregor Kasieczka,[2,†] Scott Kravitz,[3,‡] Benjamin Nachman,[3,4,§] and David Shih[5,¶]

[1] *Department of Physics, Columbia University, New York, NY 10027, USA*
[2] *Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*
[3] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[4] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*
[5] *NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, USA*

(Dated: December 8, 2021)

Machine learning plays a crucial role in enhancing and accelerating the search for new fundamental physics. We review the state of machine learning methods and applications for new physics searches in the context of terrestrial high energy physics experiments, including the Large Hadron Collider, rare event searches, and neutrino experiments. While machine learning has a long history in these fields, the deep learning revolution (early 2010s) has yielded a qualitative shift in terms of the scope and ambition of research. These modern machine learning developments are the focus of the present review.

I. INTRODUCTION

High Energy Physics (HEP) is entering a new data-driven era. For many decades, the Standard Model (SM) of particle physics has provided clear theoretical guidance to experiments, resulting in an extensive search program that culminated in the discovery of the Higgs boson [1, 2]. But while the SM is now complete, there are key experimental observations that compel the community to expand the search efforts for new particles and forces of nature beyond the SM (BSM). For example, the existence of dark matter and dark energy is well-established [3], as is the mass of neutrinos [4, 5], and the baryon-anti-

large amounts of data in many dimensions to find subtle patterns. Multivariate analysis has been commonplace in HEP for decades (see e.g. the thousands of citations to Ref. [8]), but the latest tools will qualitatively extend the sensitivity to *hypervariate* analysis whereby the entire phase space of available experimental information can be analyzed holistically.

In tandem with the growing data volume, a related challenge is the increasing need for efficient (in terms of computational time, power, and resource utilization) and accurate data processing for high-throughput applications. Efforts to that end include the development and acceleration of deep learning-based processing algorithms on

BSM

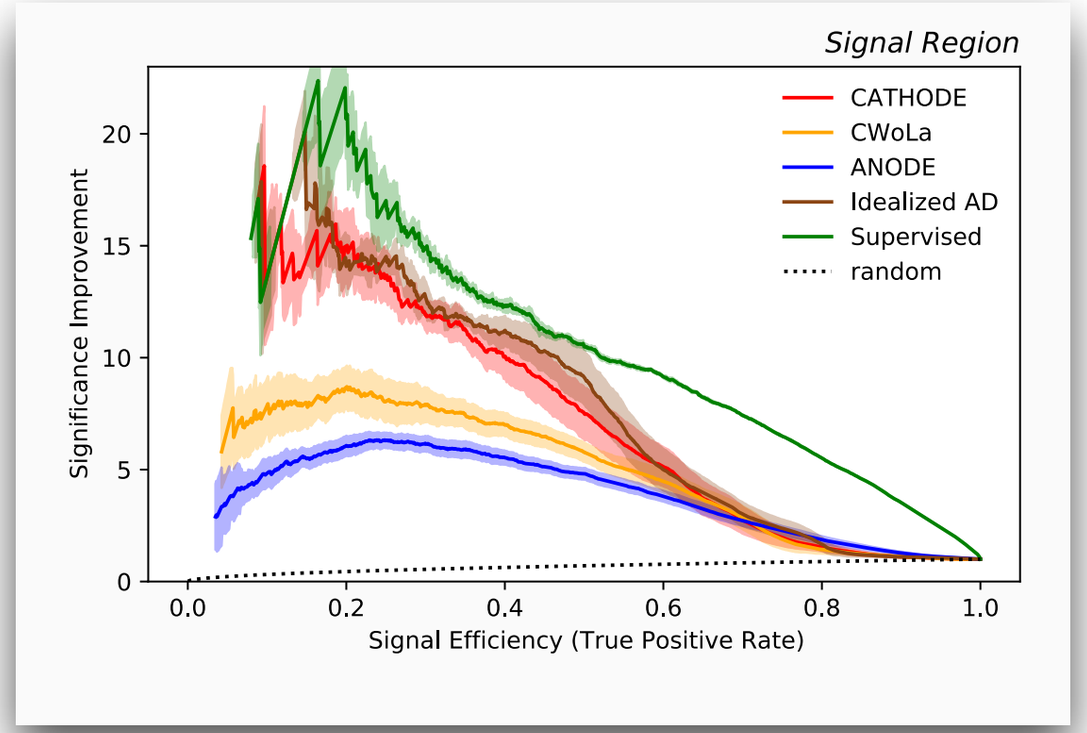
A **hot topic** in this area is **anomaly detection**

i.e. using unsupervised/weakly-supervised/semi-supervised learning to reduce signal/background model dependence

BSM

A **hot topic** in this area is **anomaly detection**

i.e. using unsupervised/weakly-supervised/semi-supervised learning to reduce signal/background model dependence



New methods are saturating **bounds** in some regimes

Context: LHC Olympics

Features: jet substructure

Methods:

CATHODE: density + classifier

CWoLa: classifier

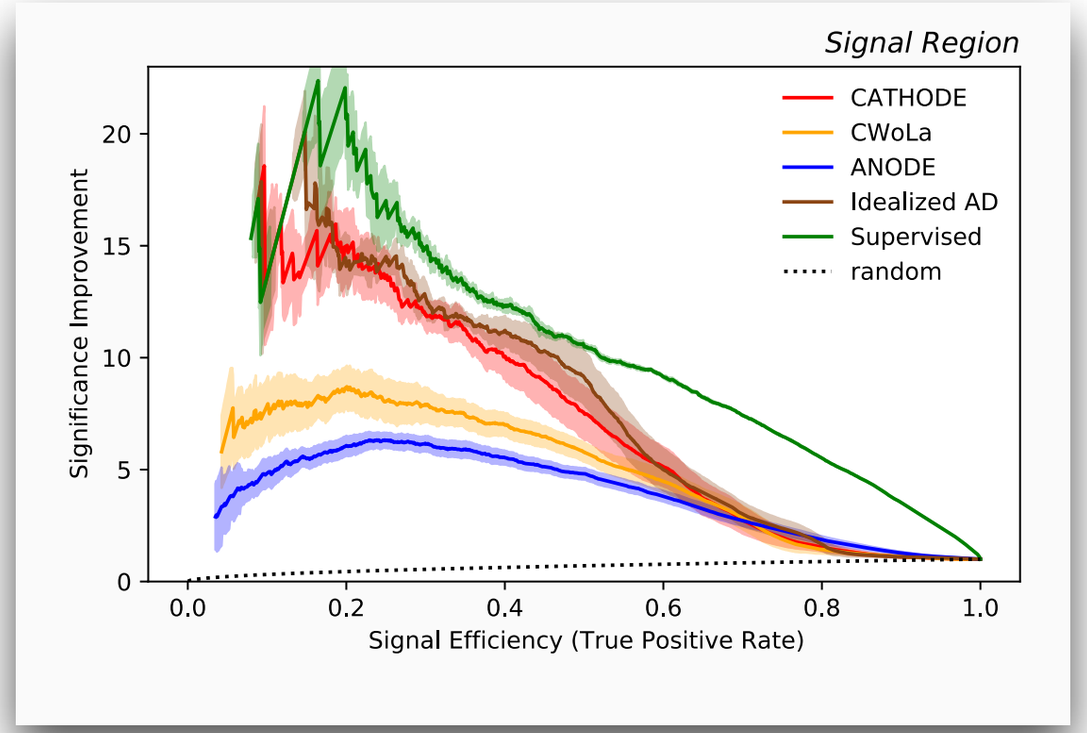
ANODE: density + density

(sideband model + signal region model)

BSM

A **hot topic** in this area is **anomaly detection**

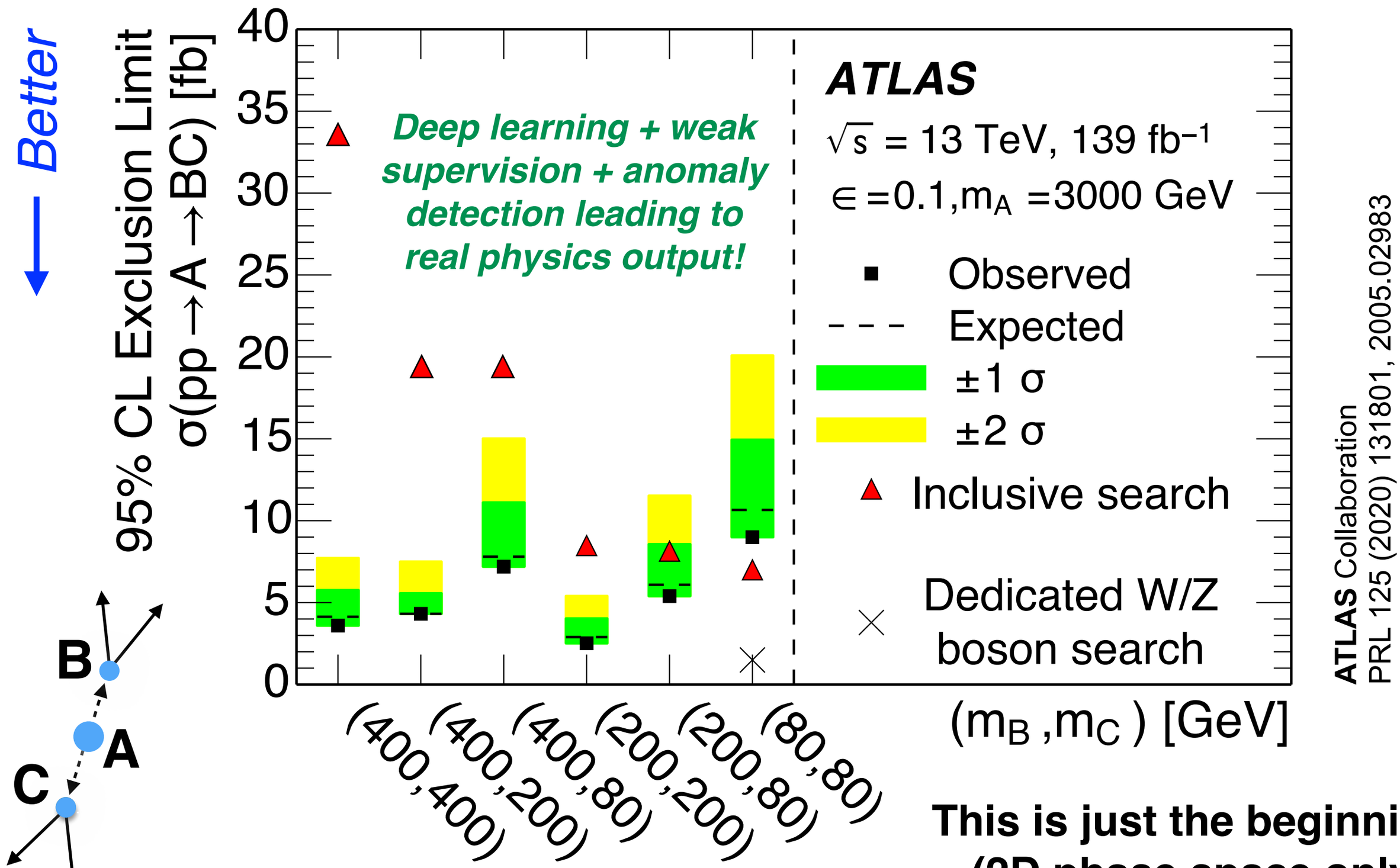
i.e. using unsupervised/weakly-supervised/semi-supervised learning to reduce signal/background model dependence



New methods are saturating **bounds** in some regimes

Key questions remain:
how to do model selection for unsupervised methods? How to best estimate the background? What about the non-resonant case?

BSM in data



This is just the beginning!
(2D phase space only)

Classification

1902.09914 + H. Qu

Top tagging landscape

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [30]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k
ParticleNet-Lite	0.984	0.937	1262±49			26k
ParticleNet	0.986	0.940	1615±93			366k
ParticleNeXt	0.987	0.942	1923±48			560k

Graph-based

State-of-the-art classification performance continues to improve! New tricks like self-attention, etc.

Calibration

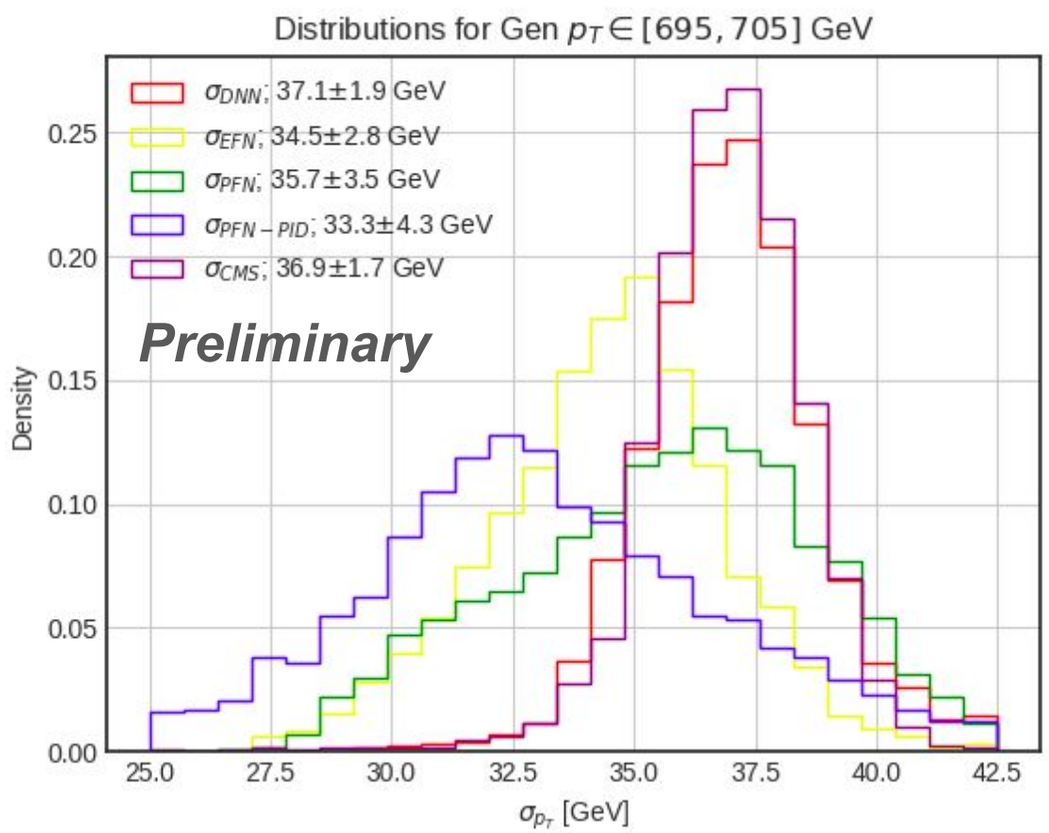
This is often set up as a regression task.

Innovation on many fronts, including the combination of various sources of information, mitigation of pileup, ...

Calibration

This is often set up as a regression task.

Innovation on many fronts, including the combination of various sources of information, mitigation of pileup, ...



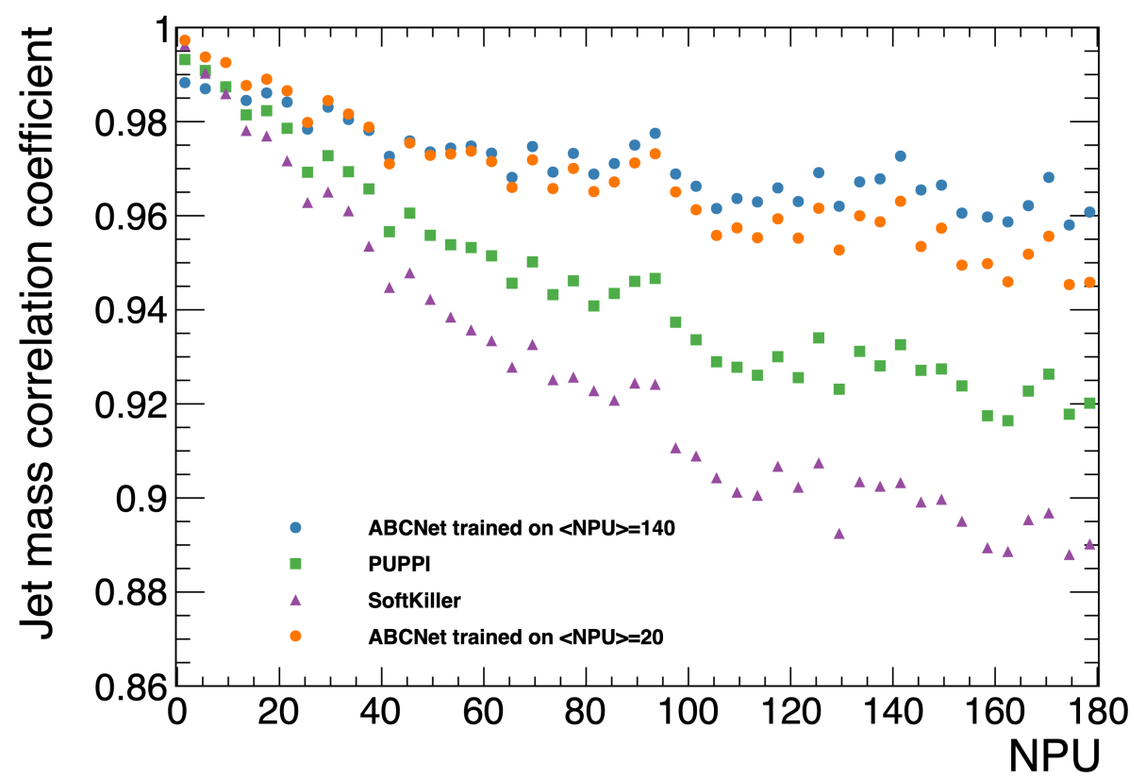
Ex: Prior-independent jet calibrations

Calibration

This is often set up as a regression task.

Innovation on many fronts, including the combination of various sources of information, mitigation of pileup, ...

Ex: Prior-independent jet calibrations



Ex: Graph-based Pileup Mitigation

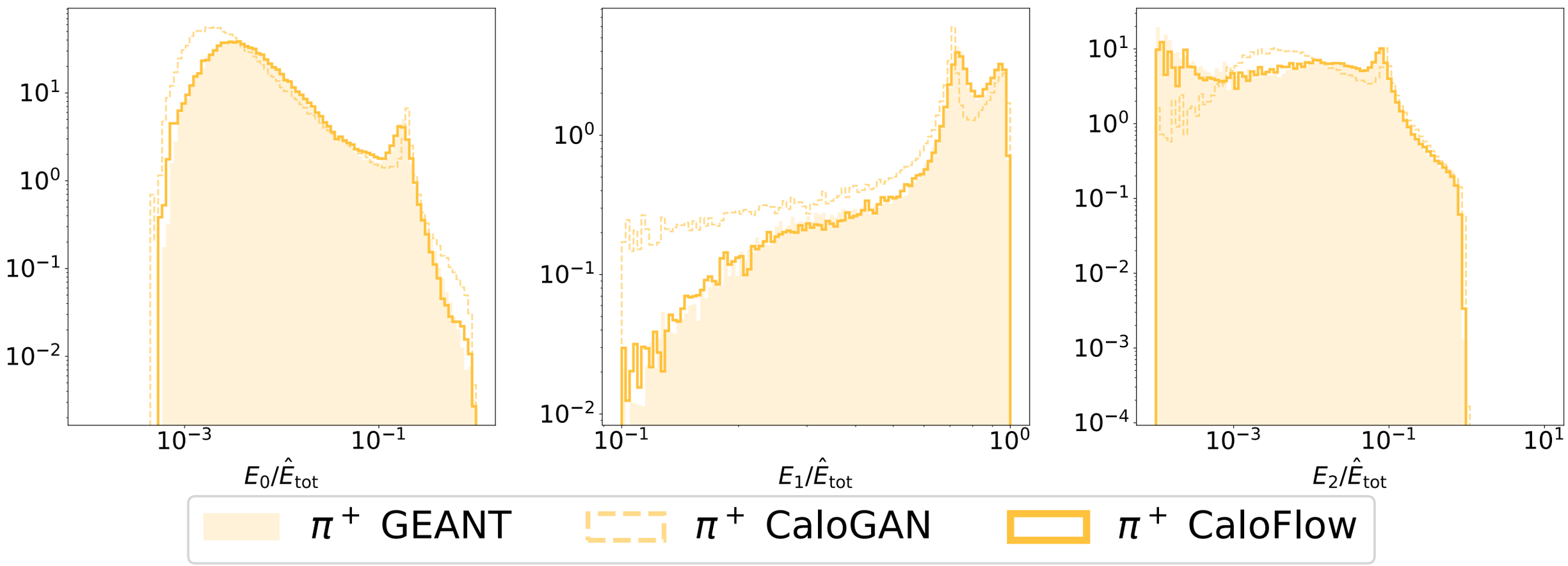
Simulation

A **hot topic** in this area is **fast calorimeter simulation**

Simulation

A **hot topic** in this area is **fast calorimeter simulation**

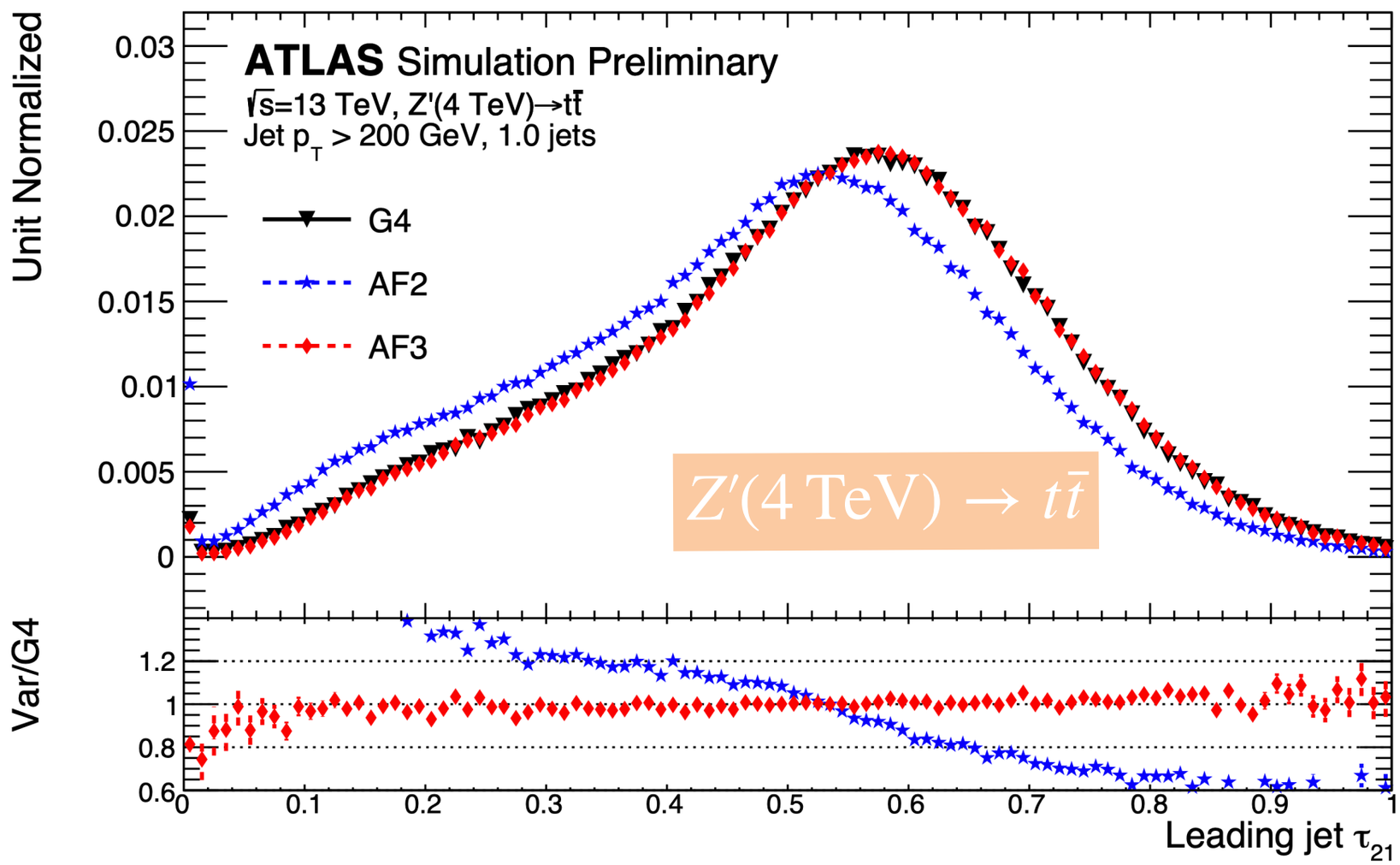
2106.05285



State-of-the-art with GANs and Normalizing Flows are reaching precision!

Simulation

Now with a full integration into a collider simulation!

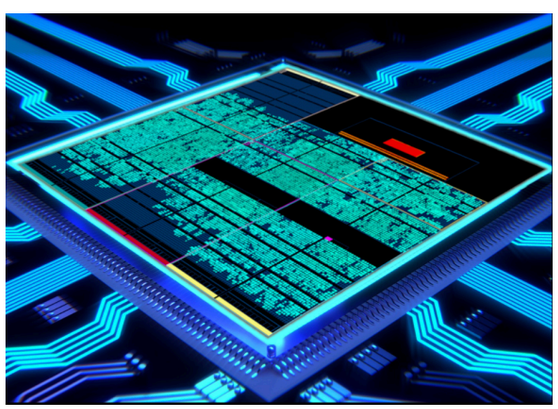


(AF3 uses a GAN for intermediate energies)

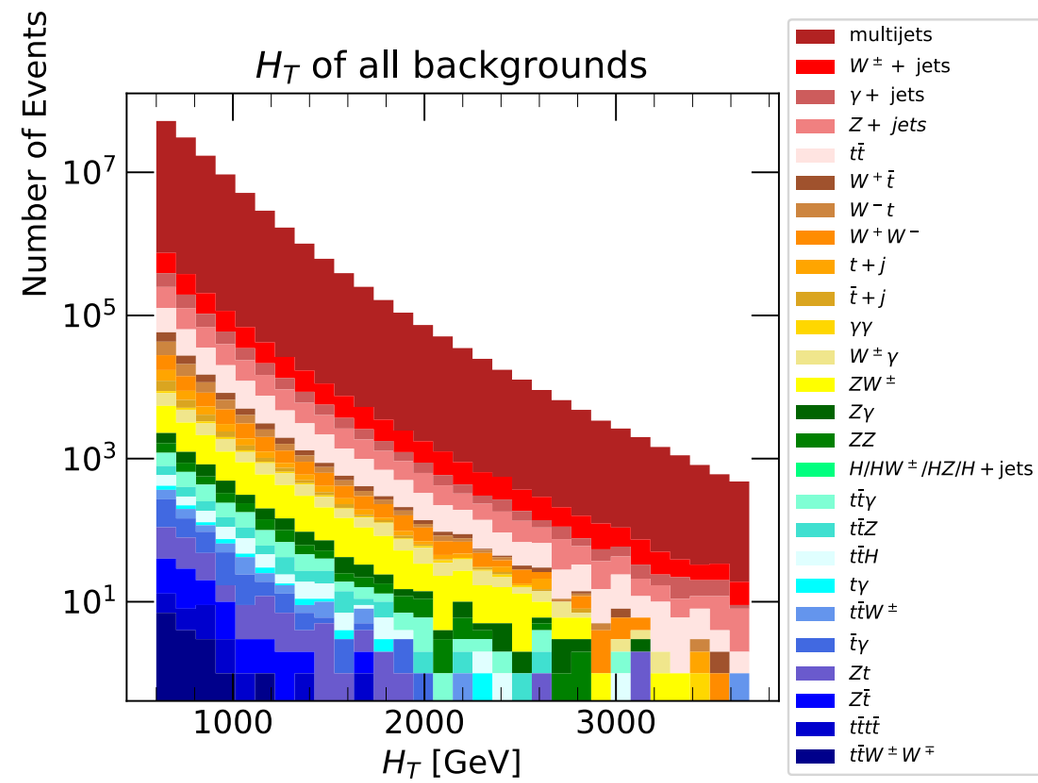
see yesterday's simulation session for more…

Datasets

LHC Olympics



Real-time Anomaly Detection



Dark Machines

+ more presented at ML4Jets and beyond!

see also https://iml.web.cern.ch/public-datasets

Exploring Data

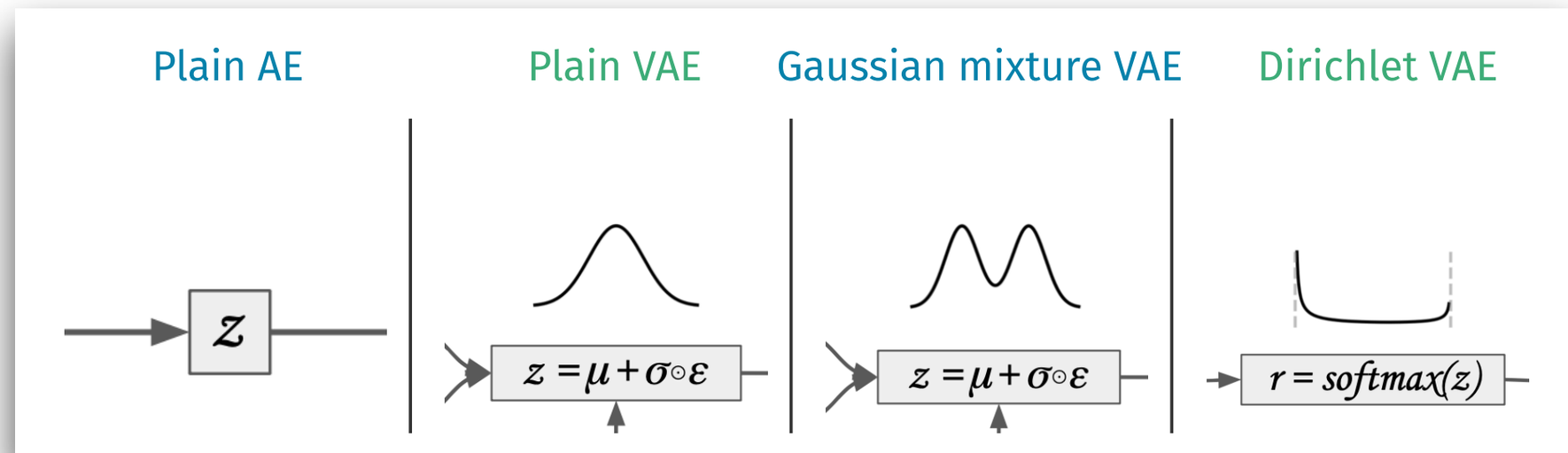
Discovering / categorizing **latent** structure in data

…this could be symmetries or multi-class components, etc.

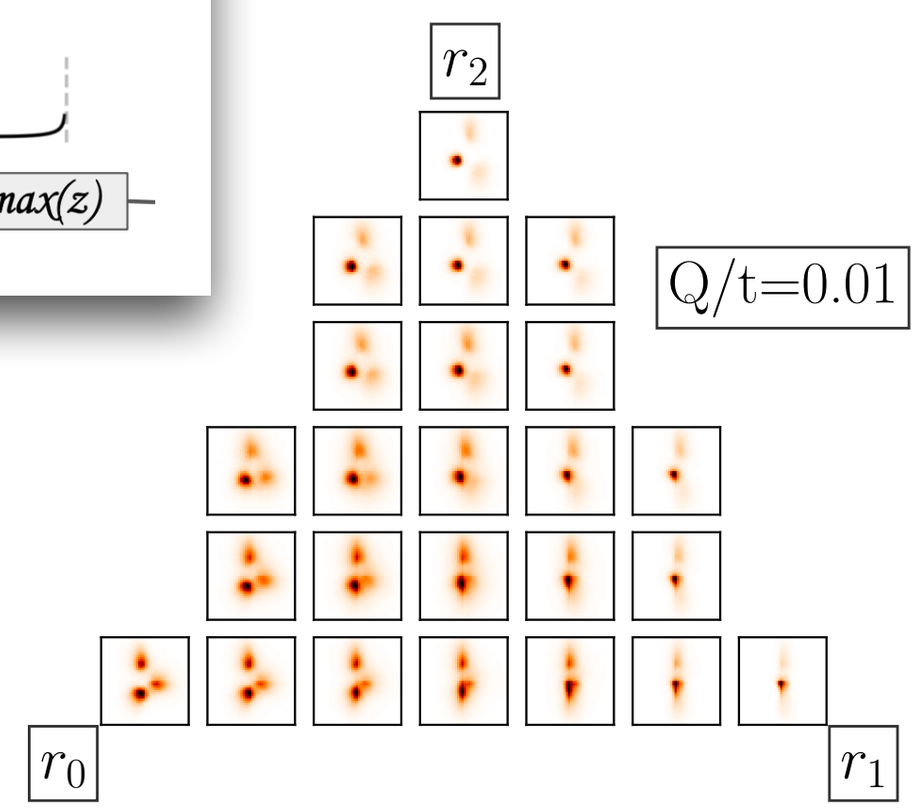
Exploring Data

Discovering / categorizing **latent** structure in data

…this could be symmetries or multi-class components, etc.



Imposing structure can lead to more interpretable latent spaces



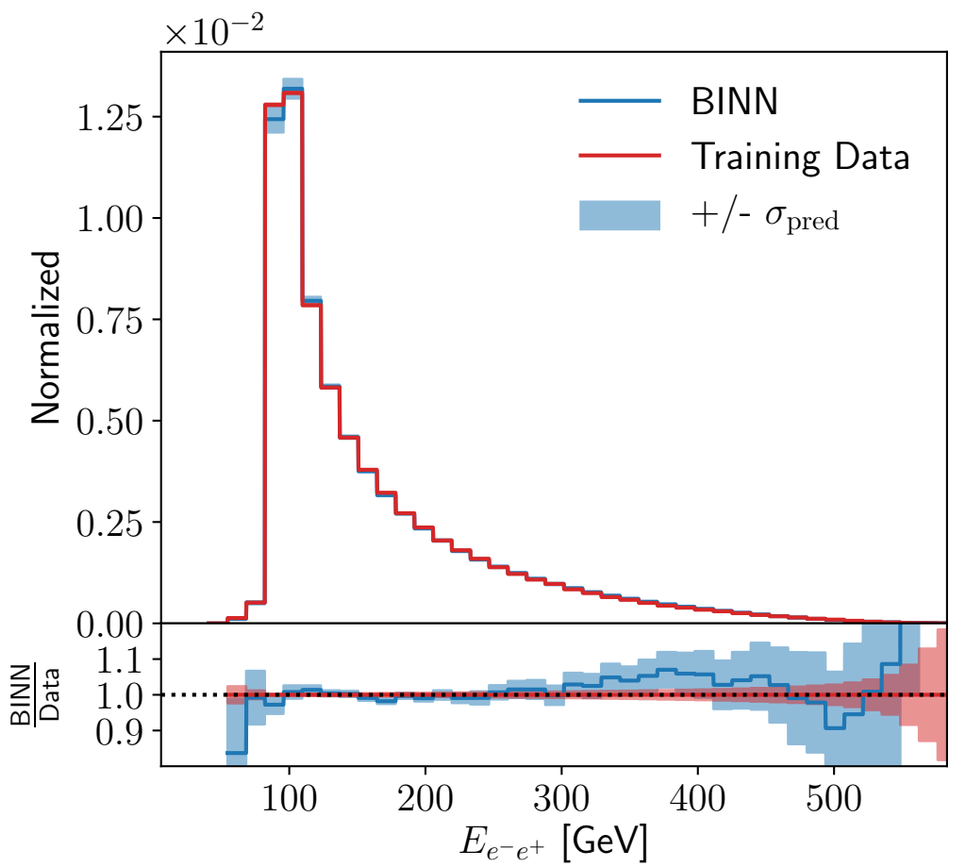
2104.0829

Interpretability and Uncertainties

Key questions: *what are uncertainties associated with neural networks? How to make networks use uncertainty information (uncertainty-aware)? How to make networks optimal with respect to downstream analysis (Inference-aware)?*

Interpretability and Uncertainties

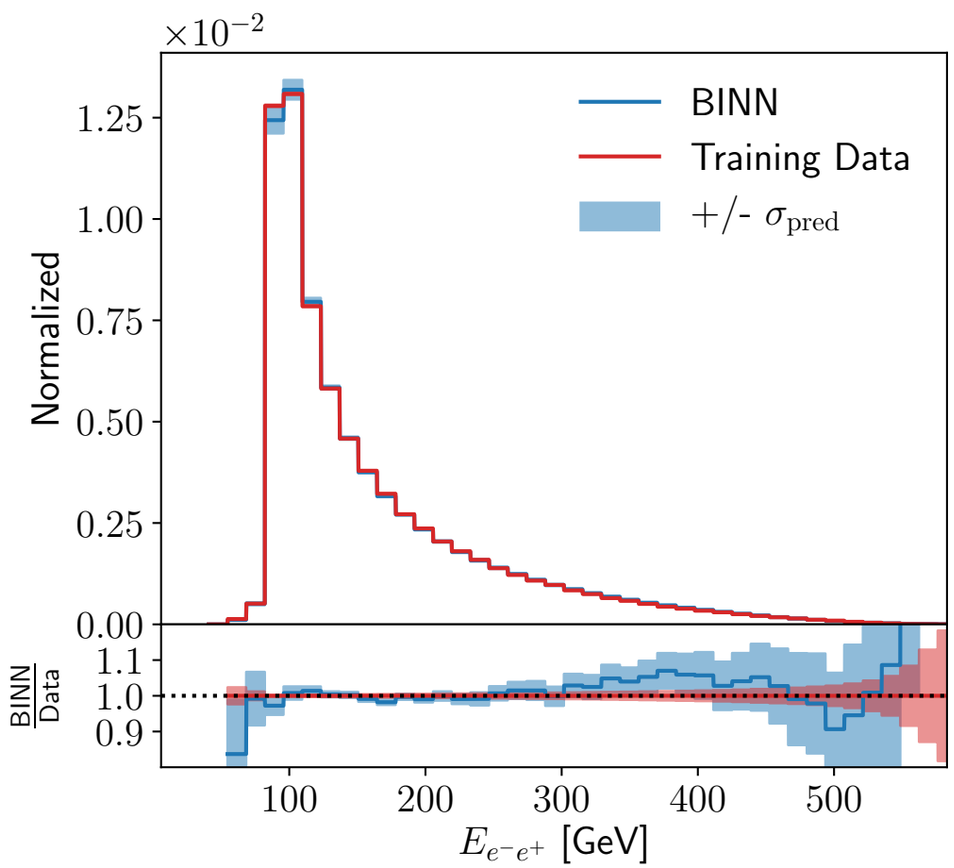
Key questions: *what are uncertainties associated with neural networks? How to make networks use uncertainty information (uncertainty-aware)? How to make networks optimal with respect to downstream analysis (Inference-aware)?*

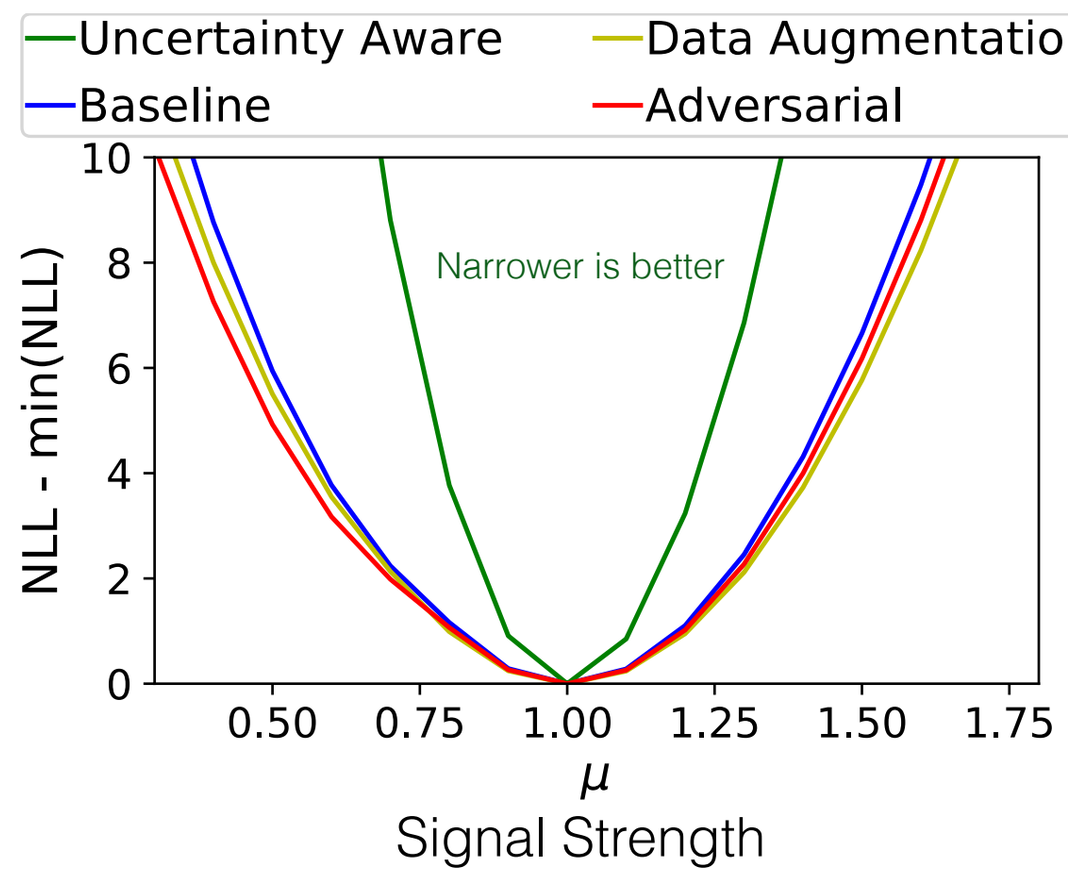


Bayesian generative models, parameterized uncertainty networks, …

Interpretability and Uncertainties

Key questions: *what are uncertainties associated with neural networks? How to make networks use uncertainty information (uncertainty-aware)? How to make networks optimal with respect to downstream analysis (Inference-aware)?*





Bayesian generative models, **parameterized uncertainty networks**, …

Data analysis in NP/HEP

Theory of everything

↓

Physics simulators

↓

Detector-level observables

↓

Pattern recognition ⟷ Pattern recognition

Nature

↓

Experiment

↓

Detector-level observables

↓

Data analysis in NP/HEP + ML

Theory of everything

Fast simulation / phase space

Parameter estimation / unfolding

Physics simulators

Detector-level observables

Pattern recognition

Nature

Experiment

Online processing & quality control

Detector-level observables

Pattern recogn

Data curation

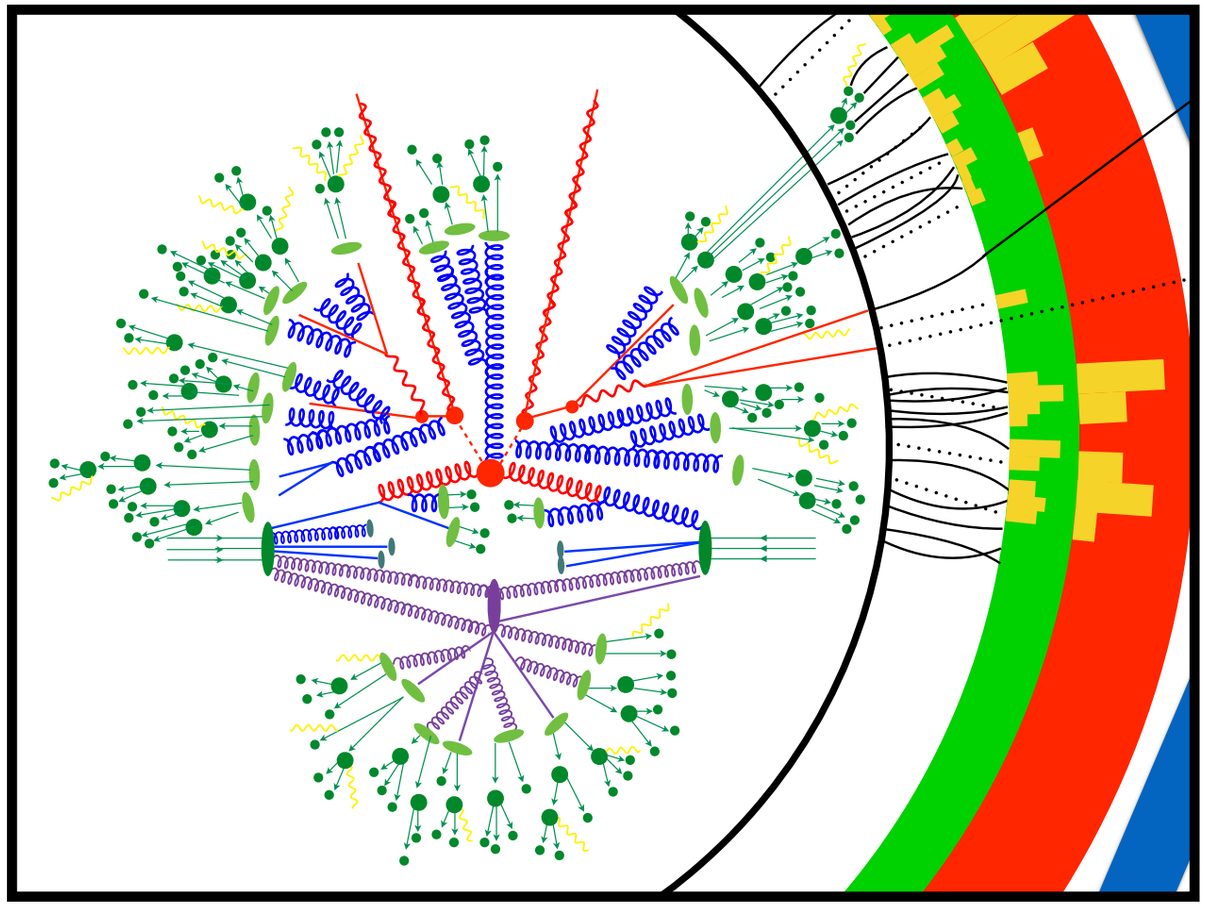
Classification to enhance sensitivity

"signal" versus "background"

calibration
clustering
tracking
noise mitigation
particle identification
...

Conclusions and outlook

Deep learning has a great potential to **enhance**, **accelerate**, and **empower** collider analyses



Due to the limited time, I was only able to cover a small selection of new ideas and results

Backup