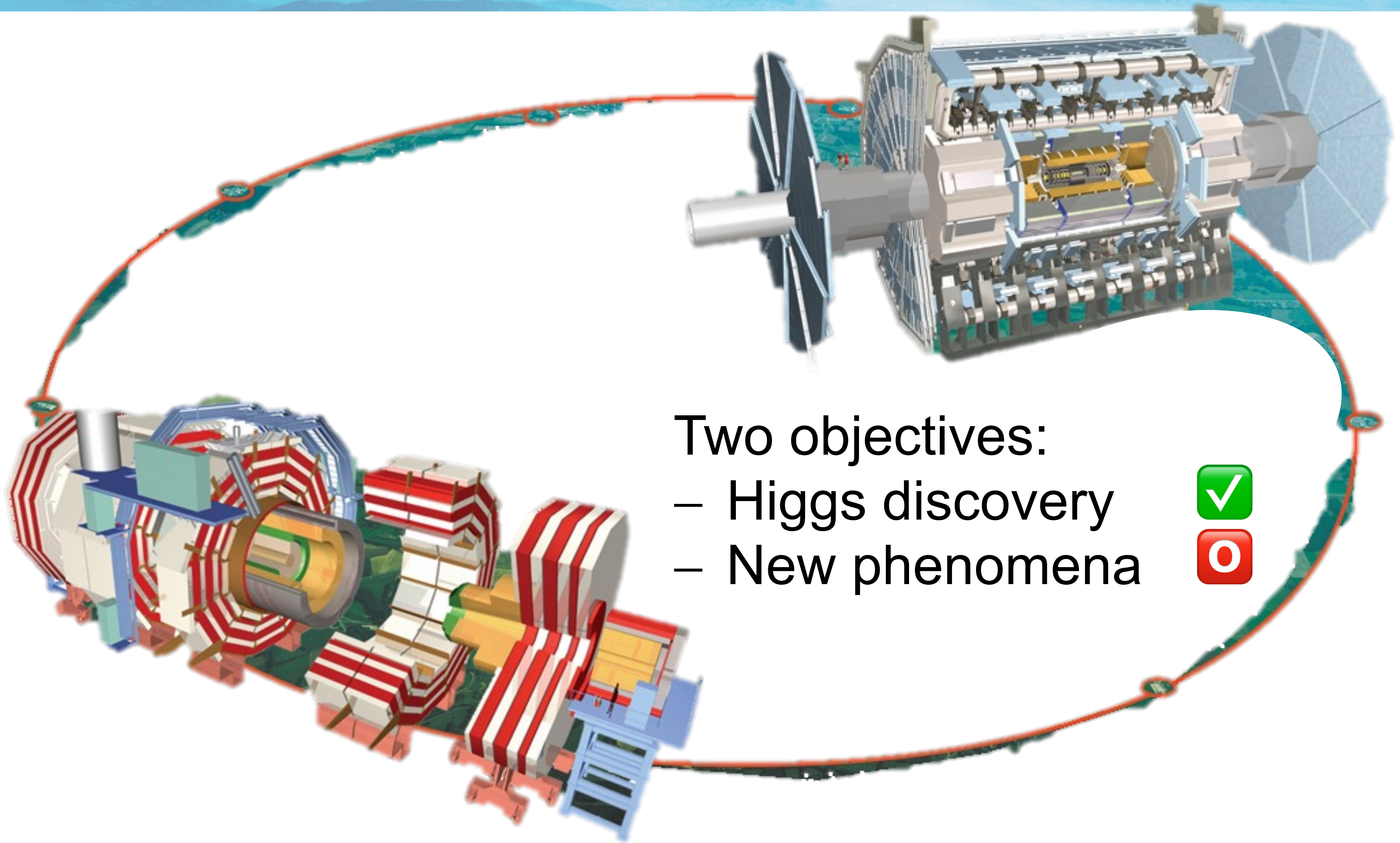
ML@HEP

Tobias Golling

UNIVERSITÉ DE GENÈVE

FACULTY OF SCIENCE

The Large Hadron Collider (LHC)



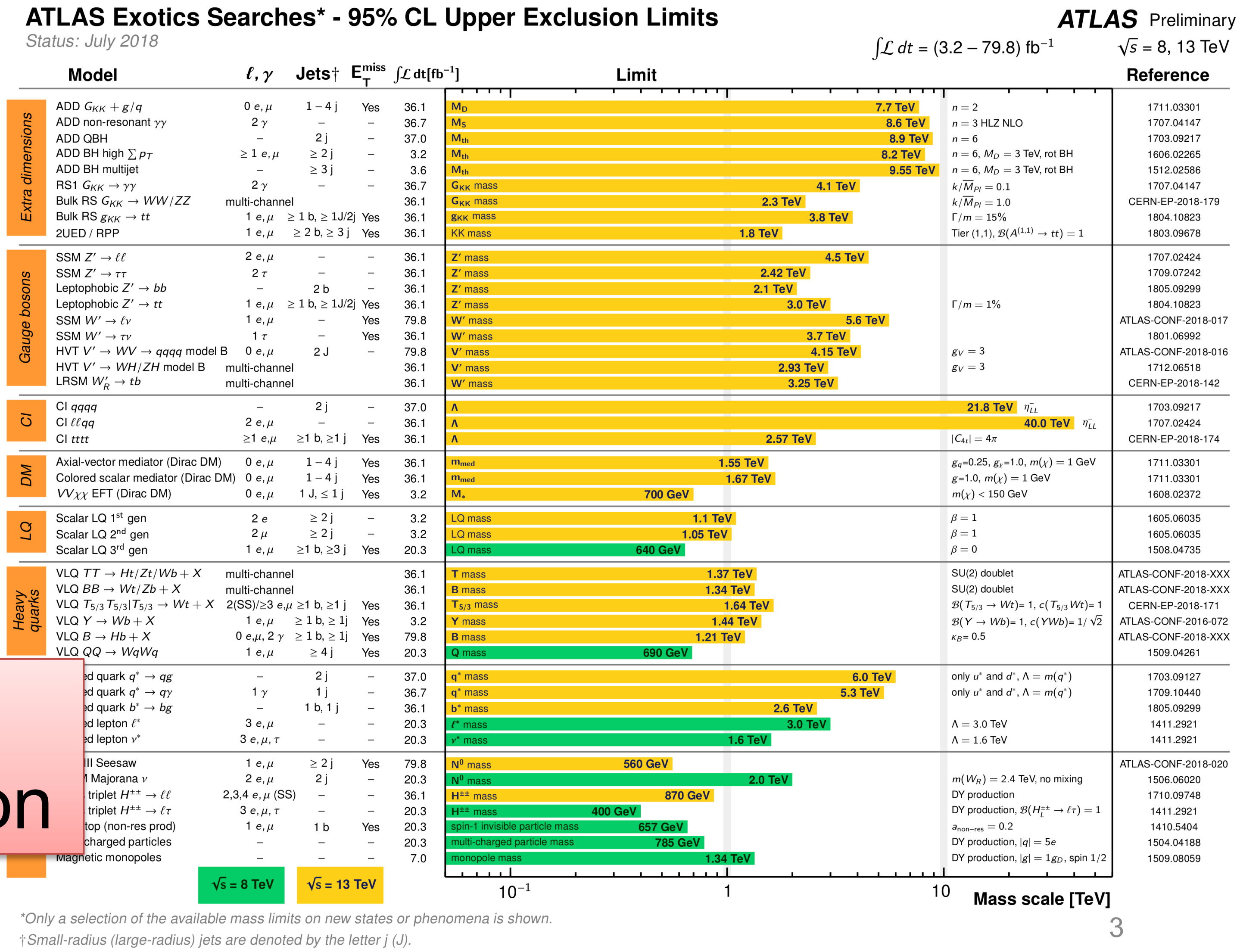
Two objectives:
- Higgs discovery ✅
- New phenomena ⭕

LHC interim evaluation

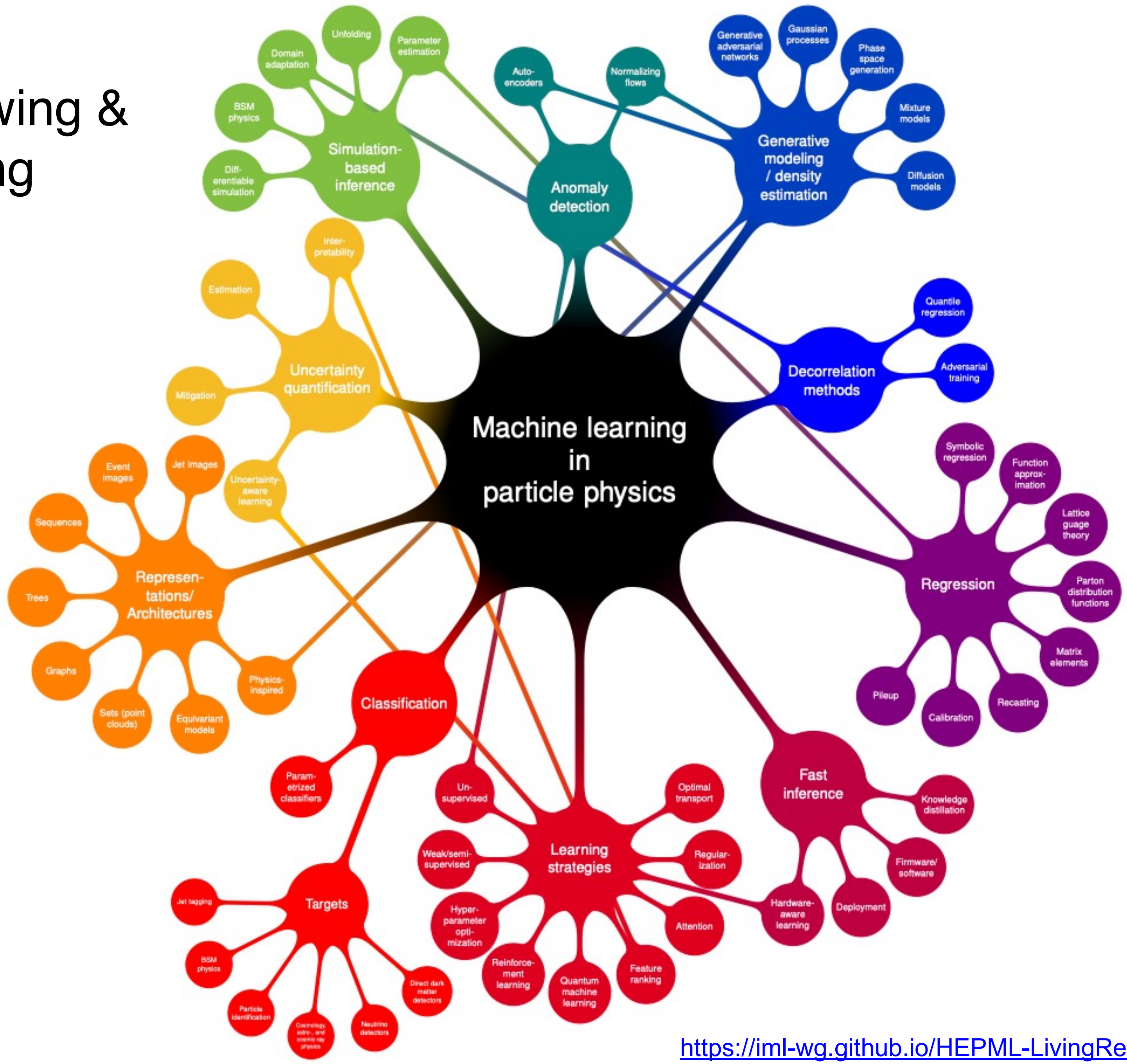
Physics beyond the SM is **not around the corner**

Slow-growth era of LHC: energy & luminosity

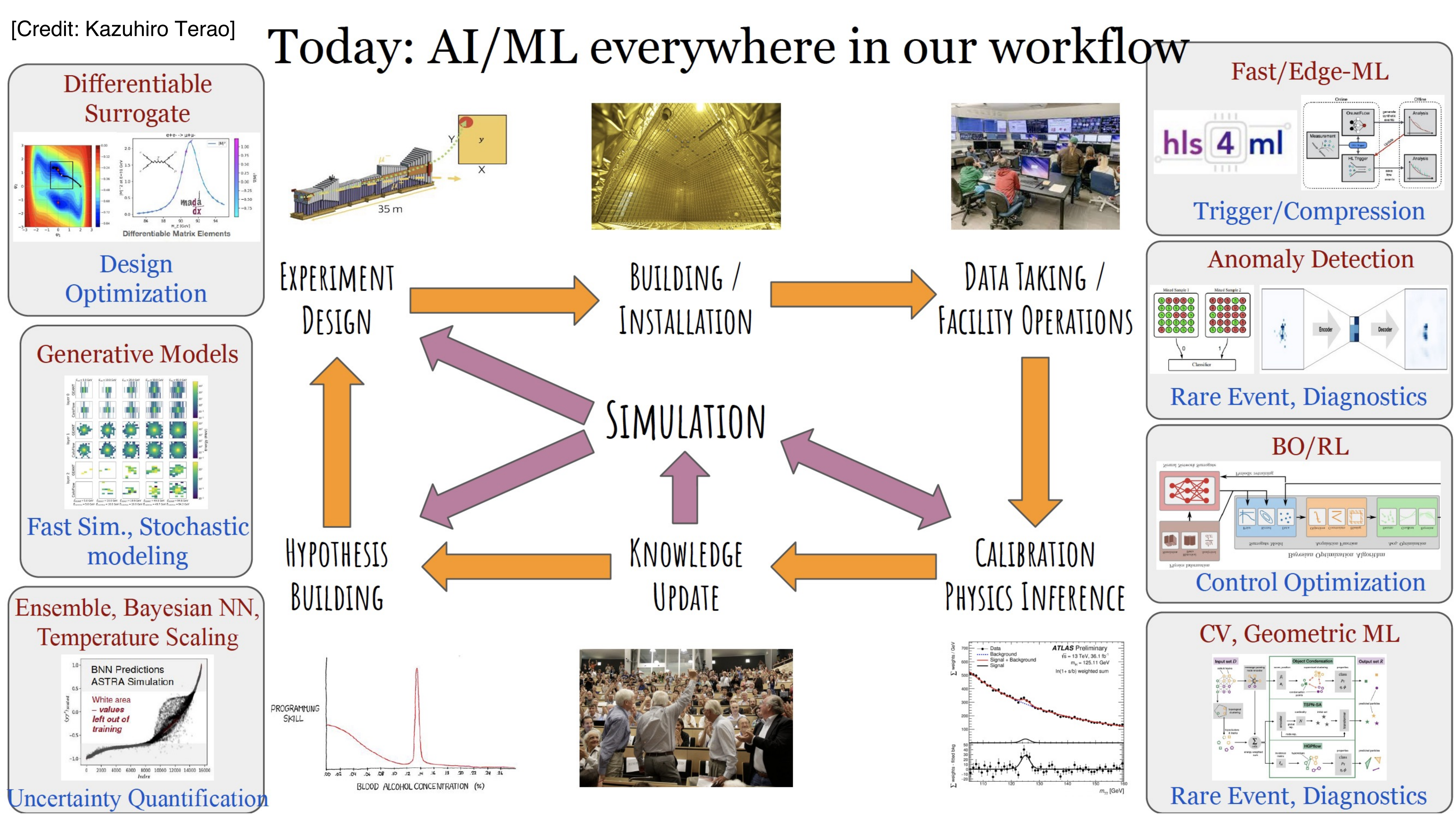
Opportunity !
Turning crank → innovation



Constantly growing & cross-connecting



https://iml-wg.github.io/HEPML-LivingReview/
https://github.com/jmduarte/Nomological_Net_ML_Particle_Physics



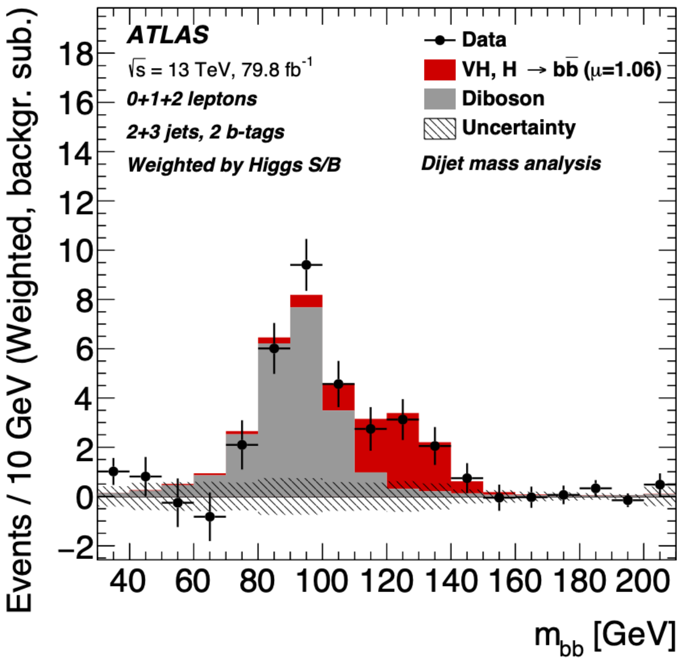
[Credit: Kazuhiro Terao]
Today: AI/ML everywhere in our workflow
Differentiable Surrogate
Differentiable Matrix Elements
Design Optimization
Generative Models
Fast Sim., Stochastic modeling
Ensemble, Bayesian NN, Temperature Scaling
BNN Predictions
ASTRA Simulation
White area – values left out of training
Uncertainty Quantification
35 m
EXPERIMENT DESIGN
BUILDING / INSTALLATION
DATA TAKING / FACILITY OPERATIONS
SIMULATION
HYPOTHESIS BUILDING
KNOWLEDGE UPDATE
CALIBRATION PHYSICS INFERENCE
PROGRAMMING SKILL
BLOOD ALCOHOL CONCENTRATION (%)
ATLAS Preliminary
Fast/Edge-ML
hls 4 ml
Trigger/Compression
Anomaly Detection
Encoder
Decoder
Classifier
Rare Event, Diagnostics
BO/RL
Bayesian Optimization Algorithm
Control Optimization
CV, Geometric ML
Object Condensation
Rare Event, Diagnostics

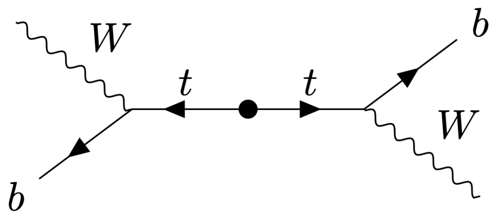
Success stories

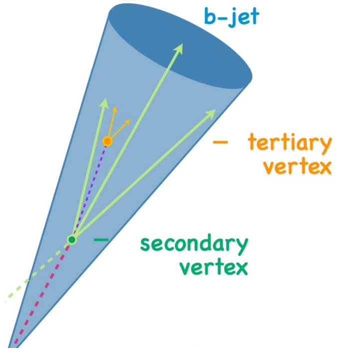
ML@HEP pioneer:

Flavor tagging

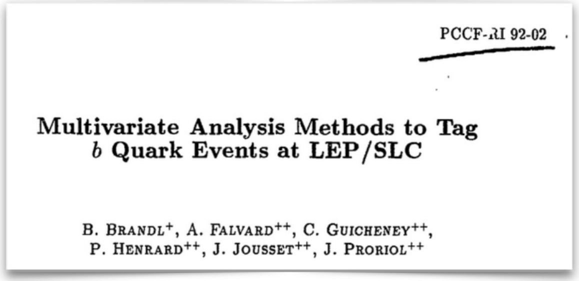
Enabler:
Higgs, top, new phenomena,…







Long history of ML in flavor tagging

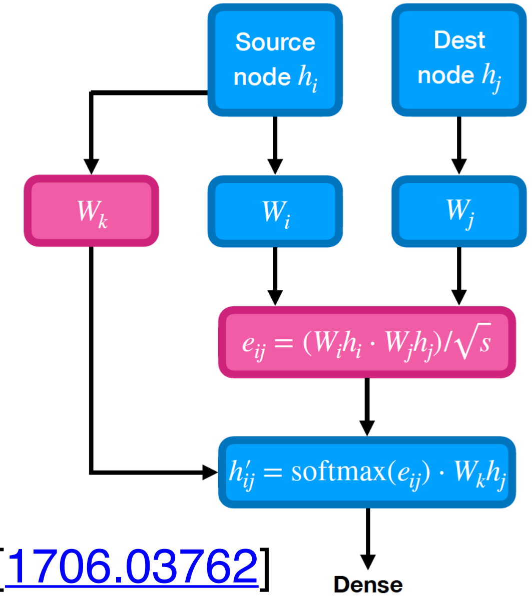


PCCF-RI 92-02

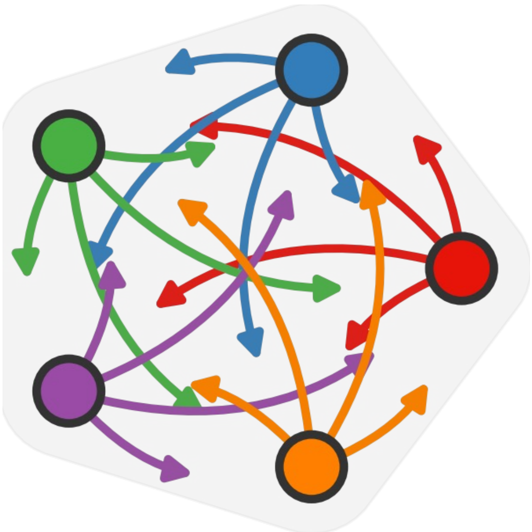
Multivariate Analysis Methods to Tag
b Quark Events at LEP/SLC

B. Brandl[+], A. Falvard[++], C. Guicheney[++],
P. Henrard[++], J. Jousset[++], J. Proriol[++]

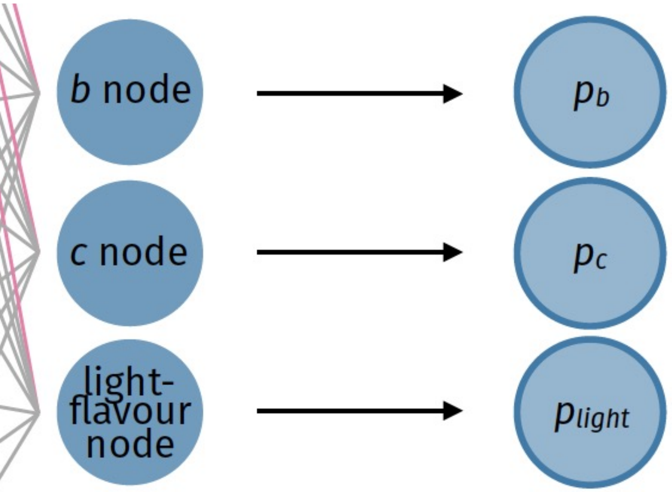
1992: Started with an **MLP** @LEP
2005: First ML b-tagging @hadron collider @D0
2007: CDF@Tevatron used **NN**
2012: ML @ATLAS: MV1
2015: **BDT** journey: MV2
2017: Back to **NN**: DL1
2017: CMS DL with DeepCSV
2019: CMS **ParticleNet**
2020: **Deep Sets**
2022: GN1 (**GNN**)
2023: GN2 (**Transformers**)
new training framework



[1706.03762]



Flexible multi-classification



$$D_b = \log\left[\frac{p_b}{f_c \cdot p_c + (1-f_c)\cdot p_u}\right]$$

$$D_c = \log\left[\frac{p_c}{f_b \cdot p_b + (1-f_b)\cdot p_u}\right]$$

[Introduced with DL1]

Category	Label
Top	bcq bqq bq cq
Higgs	bb cc VV*→qqqq
Z	bb cc qq
W	cq qq
QCD	g→bb g→cc b c others

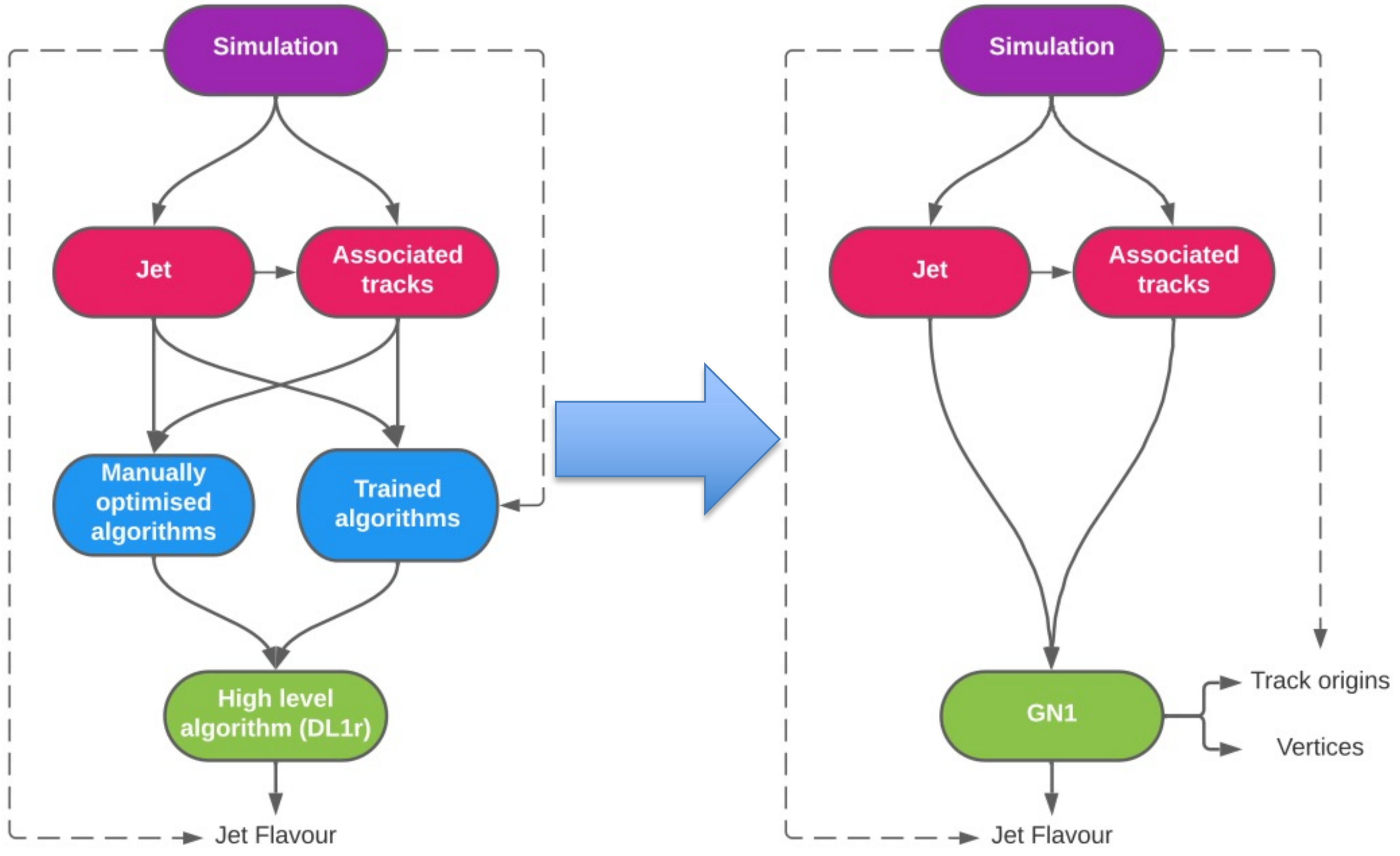
Training independent from algorithm tuning (f_c & f_b)

Extended to more classes

Streamlining architecture

Improved performance

Easier training & optimization



Hand-designed → End-to-end

[GN1]

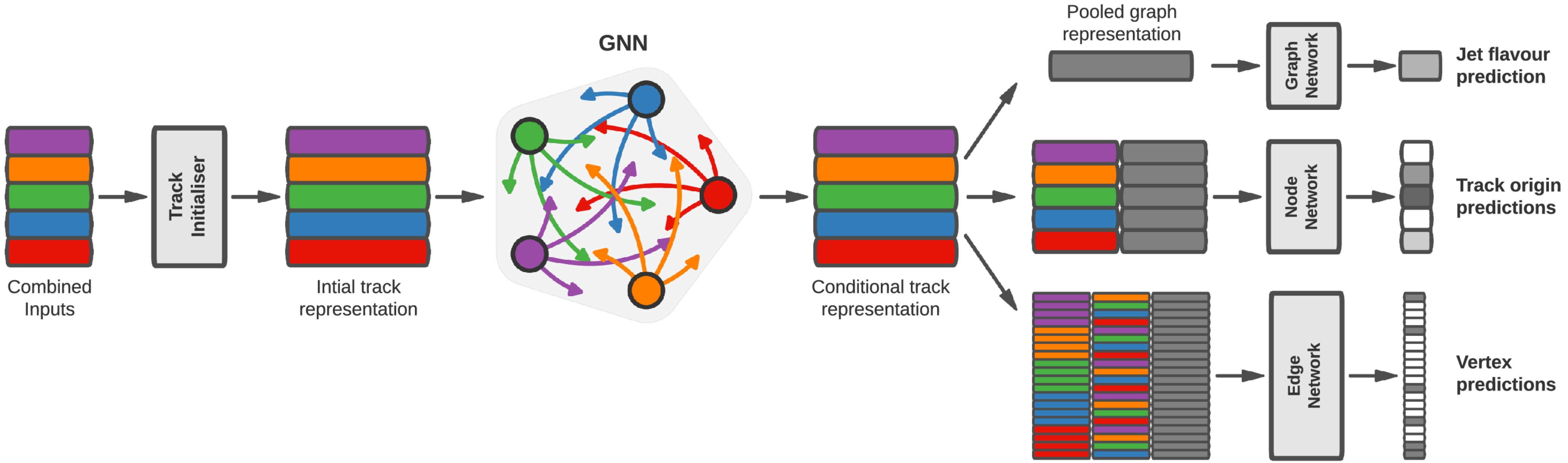
The benefit of auxiliary tasks

Performance

Interpretability

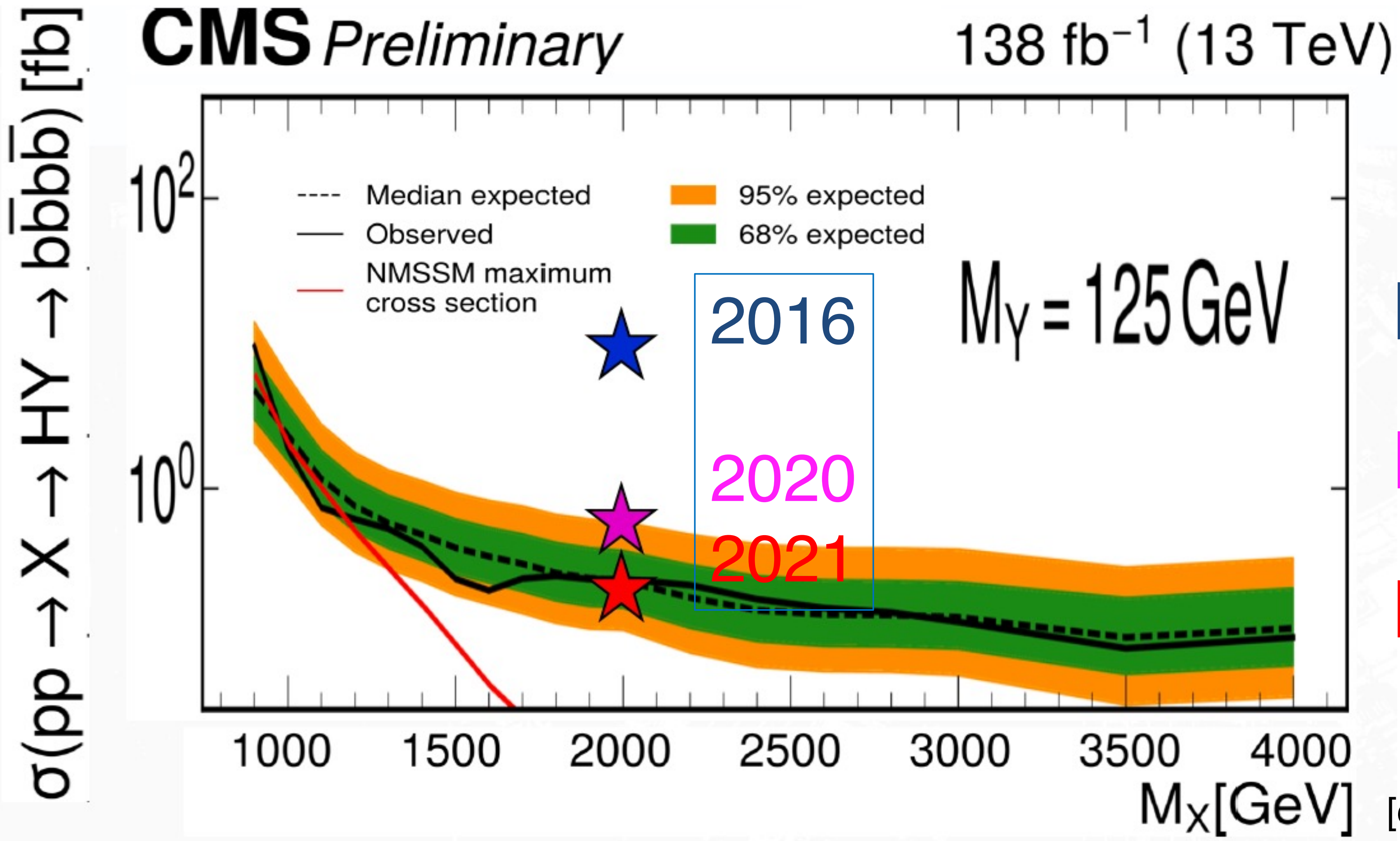
Track origin:

- Pileup
- Fake
- Primary
- FromB
- FromBC
- FromC
- FromTau
- OtherSecondary



[GN1]

Impact on physics



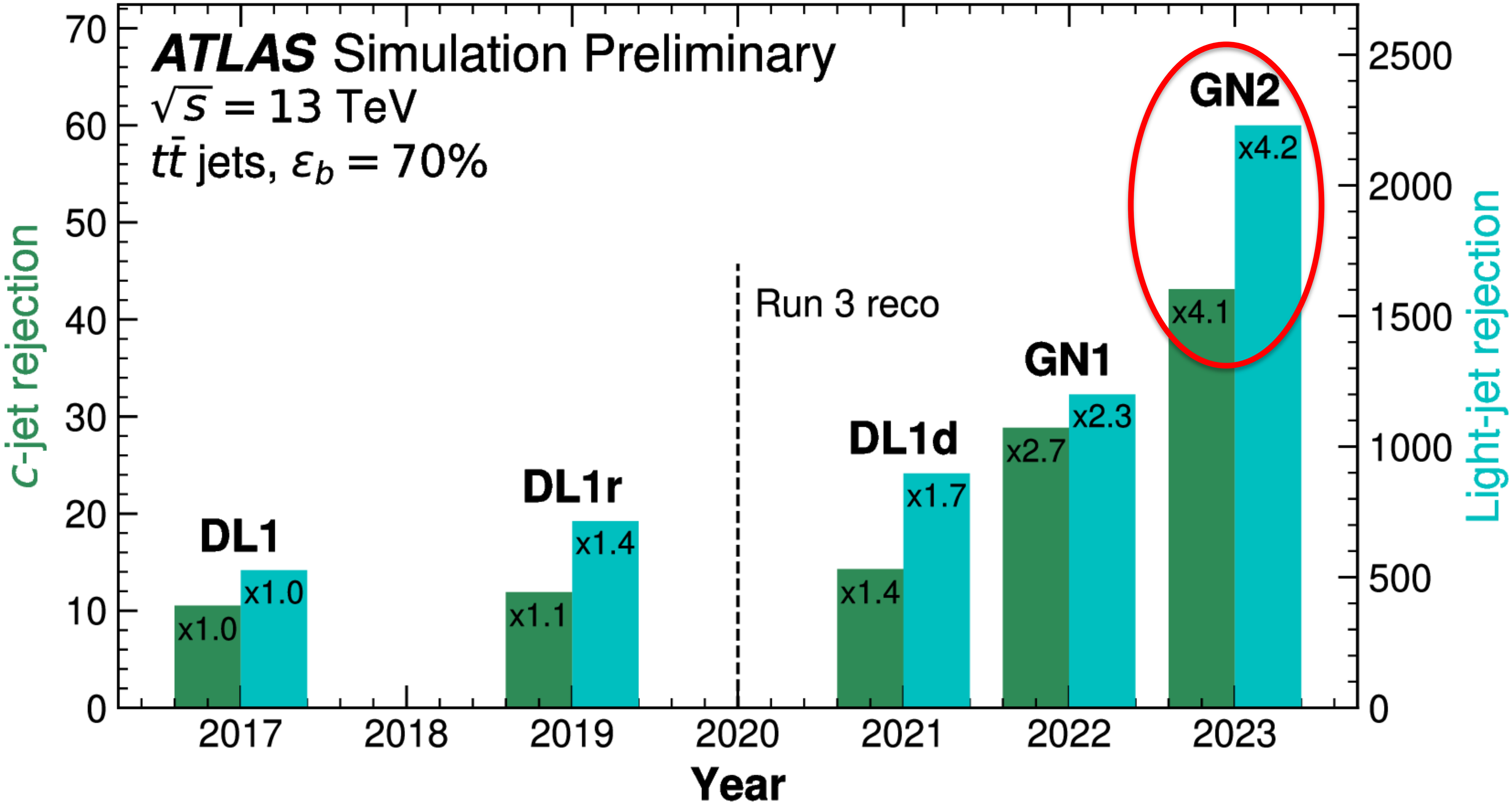
[1710.04960]

[CMS-B2G-20-004]

[2204.12413]

[credit: Loukas Gouskos]

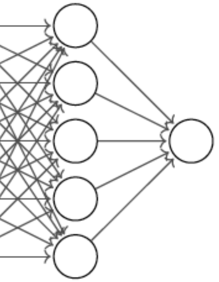
When are we reaching a plateau?



[Transformer-based GN2, see FTAG-2023-01]

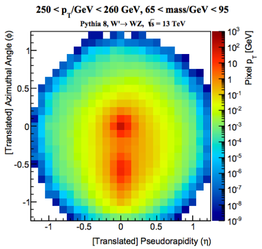
Evolving data representations in HEP

Arbitrary inputs
FF NN



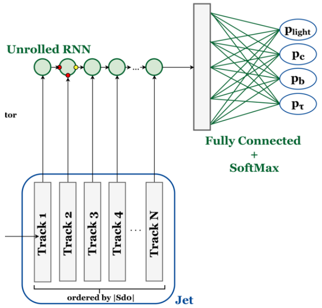
⇒

Images
CNN
[1511.05190]



⇒

Sequences
RNN
[ATL-PHYS-PUB-2017-003]

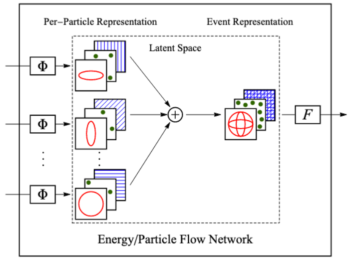


⇒

⇒

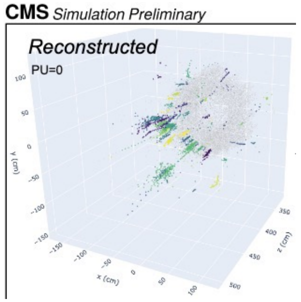
Point clouds

Deep Sets
[1810.05165]



⇒

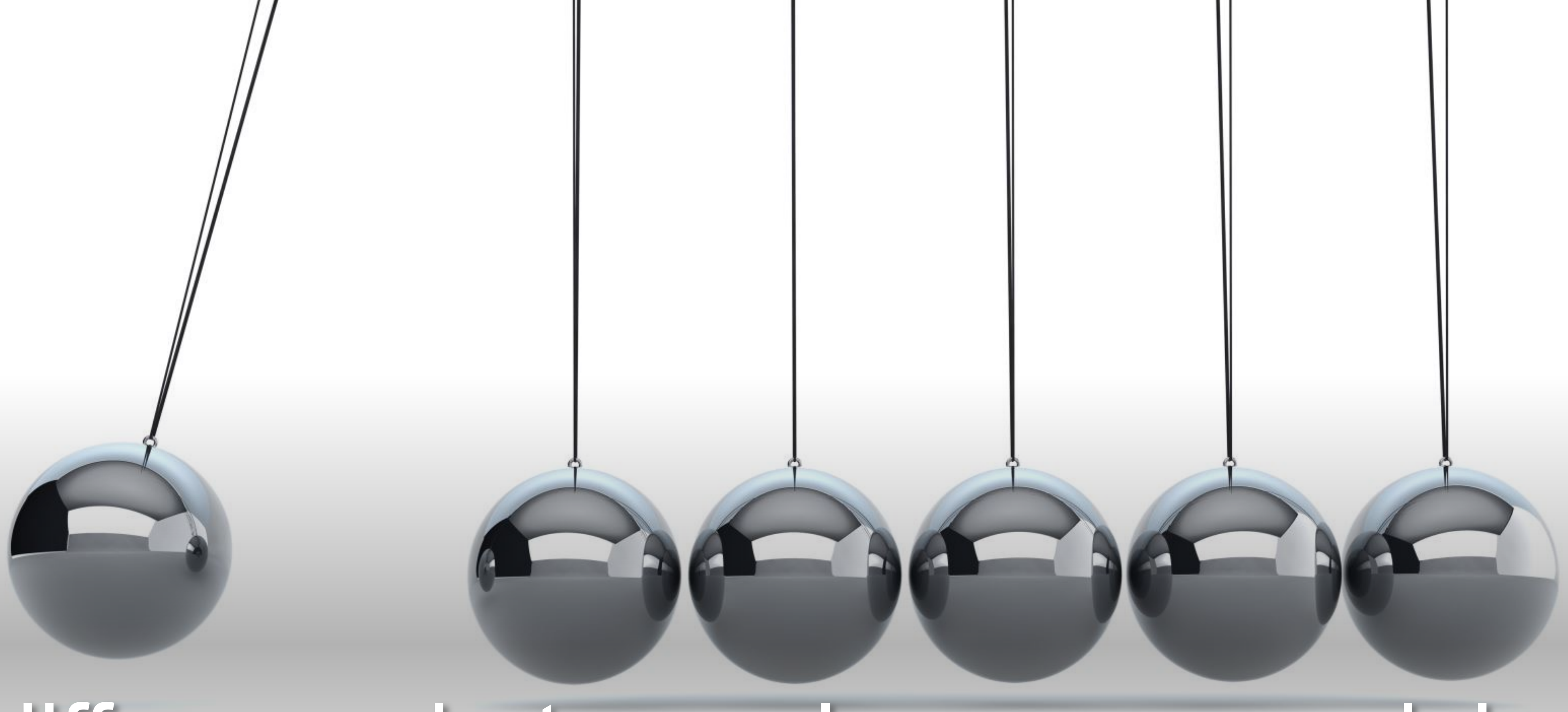
GNN / Transformer
[2203.01189]



⇒

…

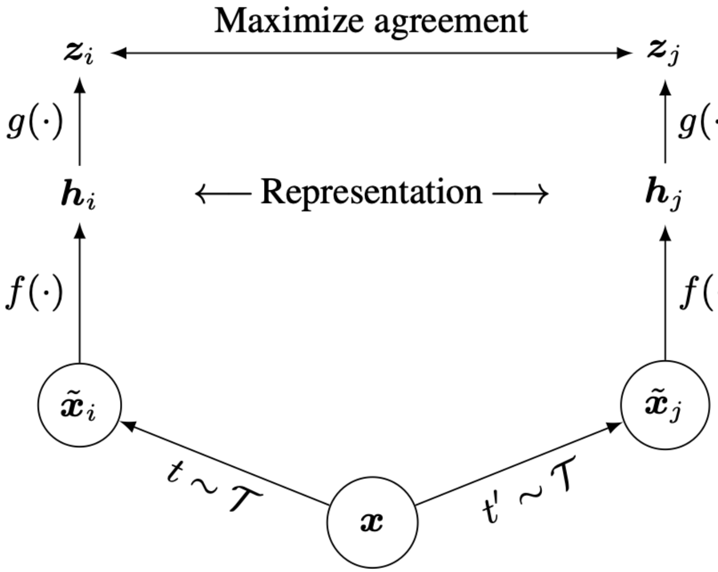
Physics-aware AI
[the edge of science]



The difference between language models & PP?

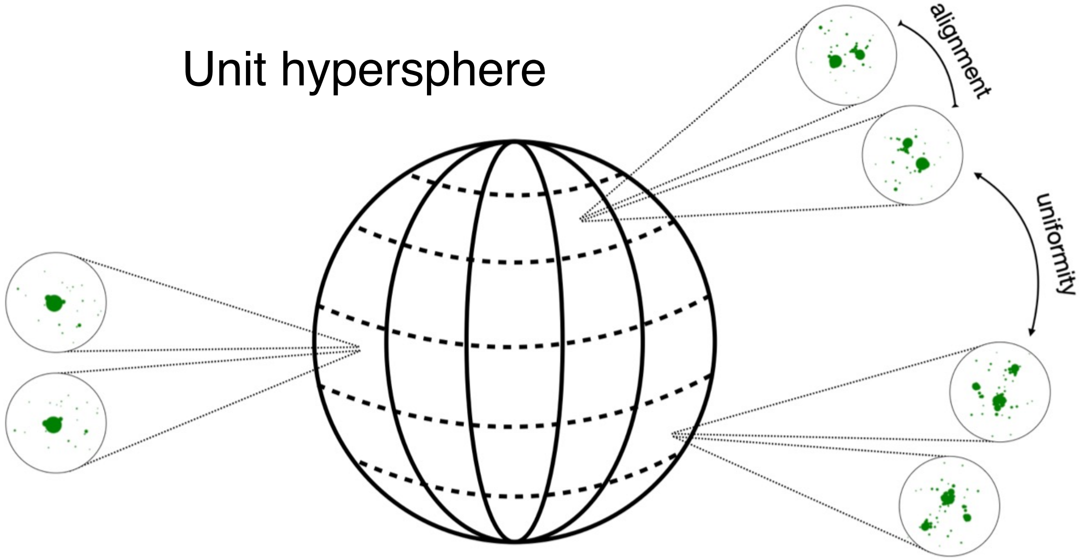
We have a model

Invariance to transformation: contrastive learning



Projection head

Encoder

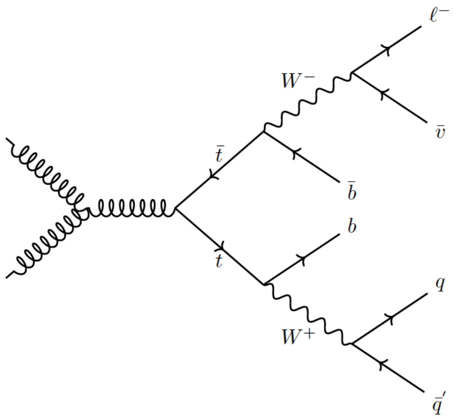


$$s(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|} = \cos\theta_{ij}$$

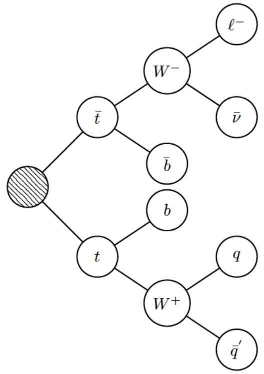
Augmentation	$\epsilon^{-1}(\epsilon_s=0.5)$	AUC
none	15	0.905
translations	19	0.916
rotations	21	0.930
soft+collinear	89	0.970
all combined (default)	181	0.979

[JetCLR [2108.04253] (based on SimCLR Hinton et al.)]

Encode physics into a GNN

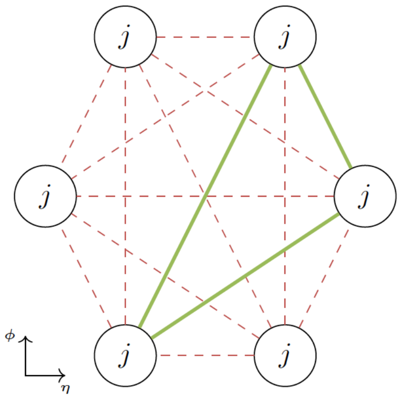


(a) Feynman diagram

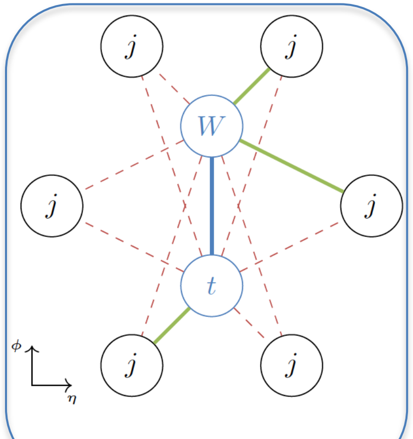


(b) Node and edge graph

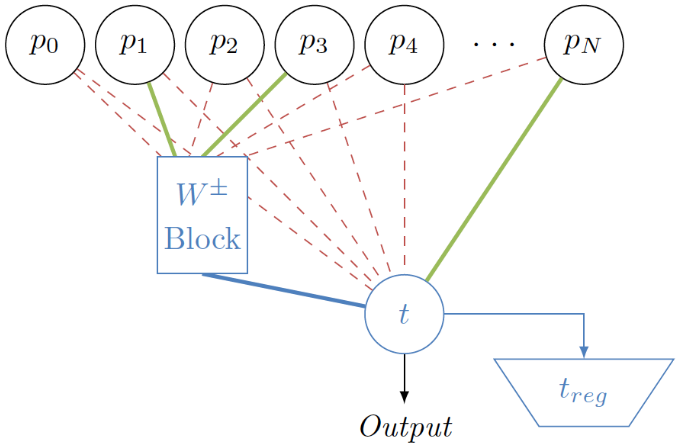
Encode information by leaving out edges

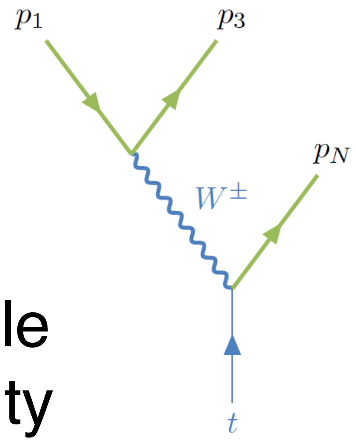


(a) Fully connected graph



(b) Topograph



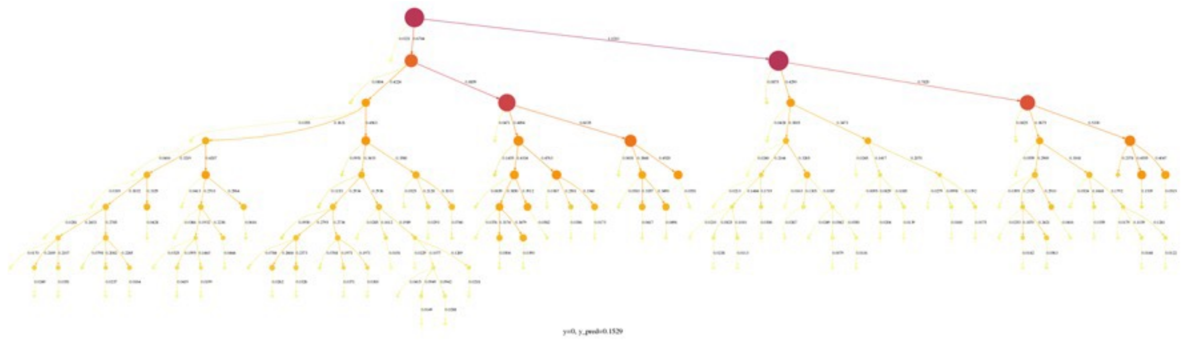


Modular
Generalizable
Interpretability
Combinatorics solving
Downstream tasks

[2303.13937]

Inject physics knowledge into AI

[1702.00748, 1711.02633]

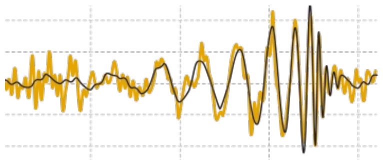


Tree structure of sequential recombination jet algorithms as Recursive NN

- Symmetries [rotation, translation, permutation,…]
  - Lorentz layers [2006.04780, 2201.08187]
  - GNNs: permutation symmetry [Energy flow network, ParticleNet]
  - PELICAN [2211.00454]
- Auxiliary tasks: energy conservation,…
- Observable construction with ML [1902.07180]

ML interpretability for science

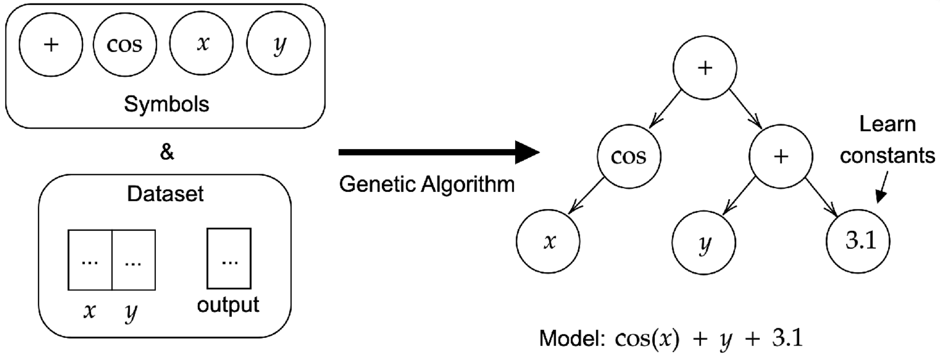
Science



$$h = \frac{2G}{c^4}\frac{1}{r}\frac{\partial^2 Q}{\partial t^2}$$

Computer vision

???

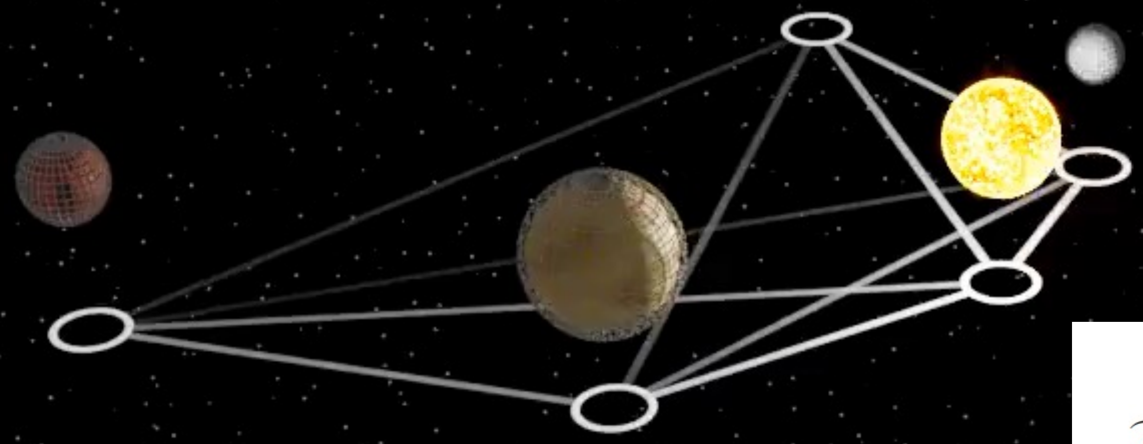


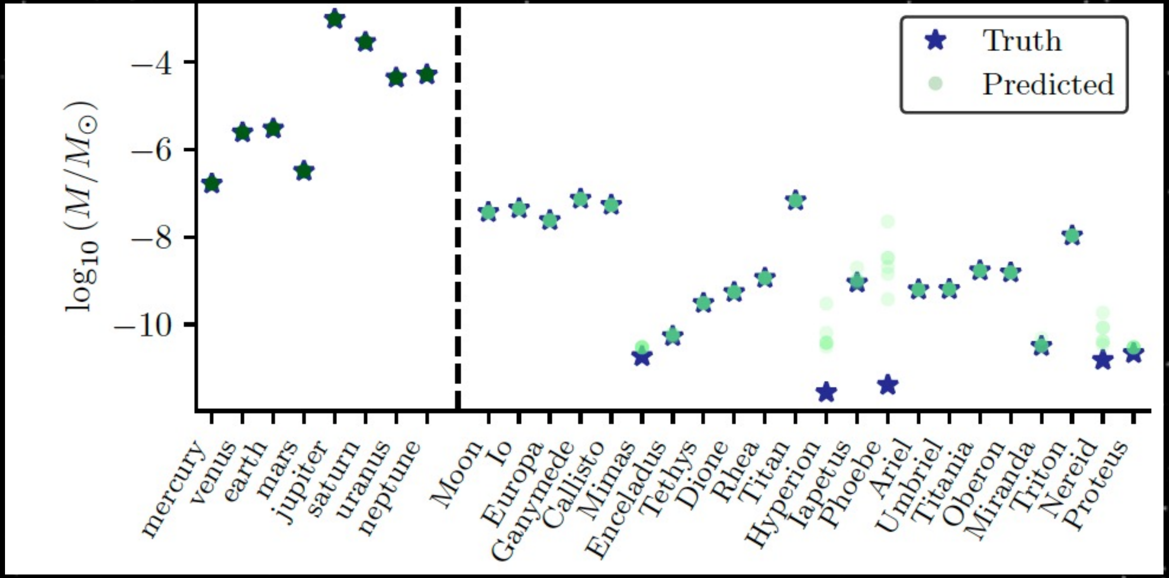
Analytic models often generalize better than NN
Symbolic regression as inductive bias

NN weights [black box] → PySR → **Analytic expression** [insights]

[Miles Cranmer – Hammers & Nails 2022]

Learn Newton’s law from solar system





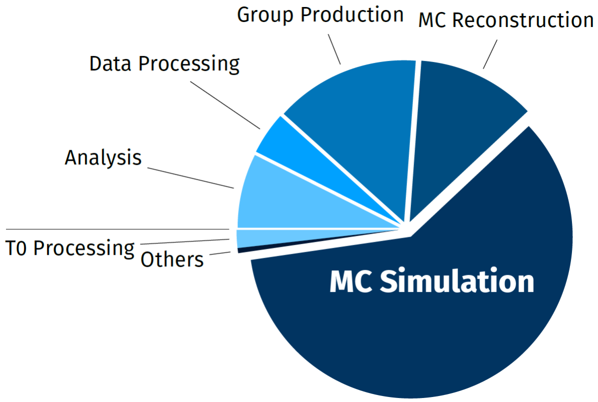
GNN → PySR → Learn masses + dynamics

[2202.02306](https://arxiv.org/abs/2202.02306)

Surrogate modeling

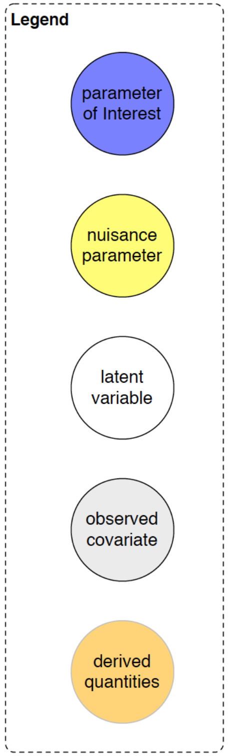
Full Sim & Reco

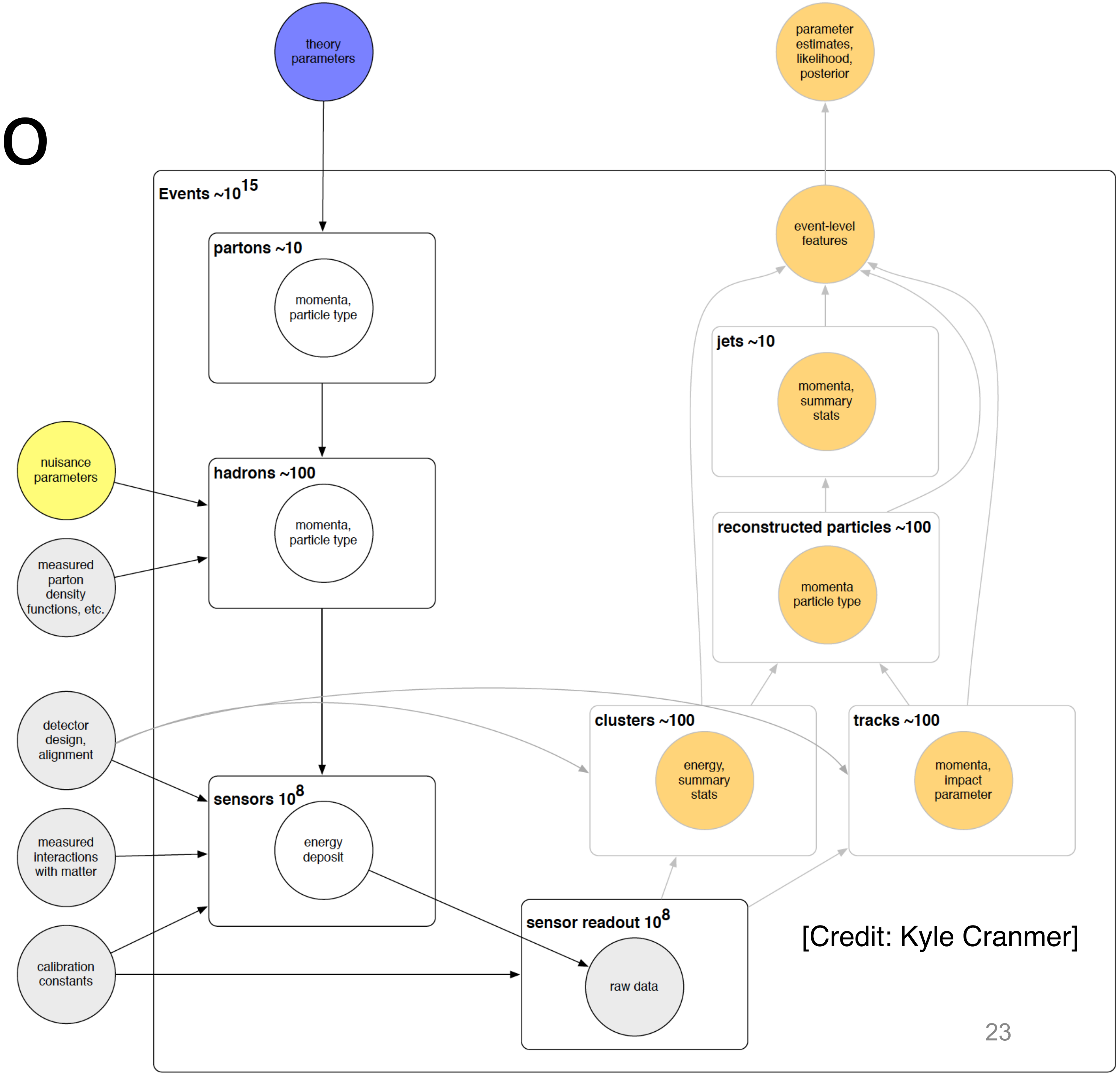
Bottleneck:
computing budget



Up to 10 min / event

[LHCC-2022-005]



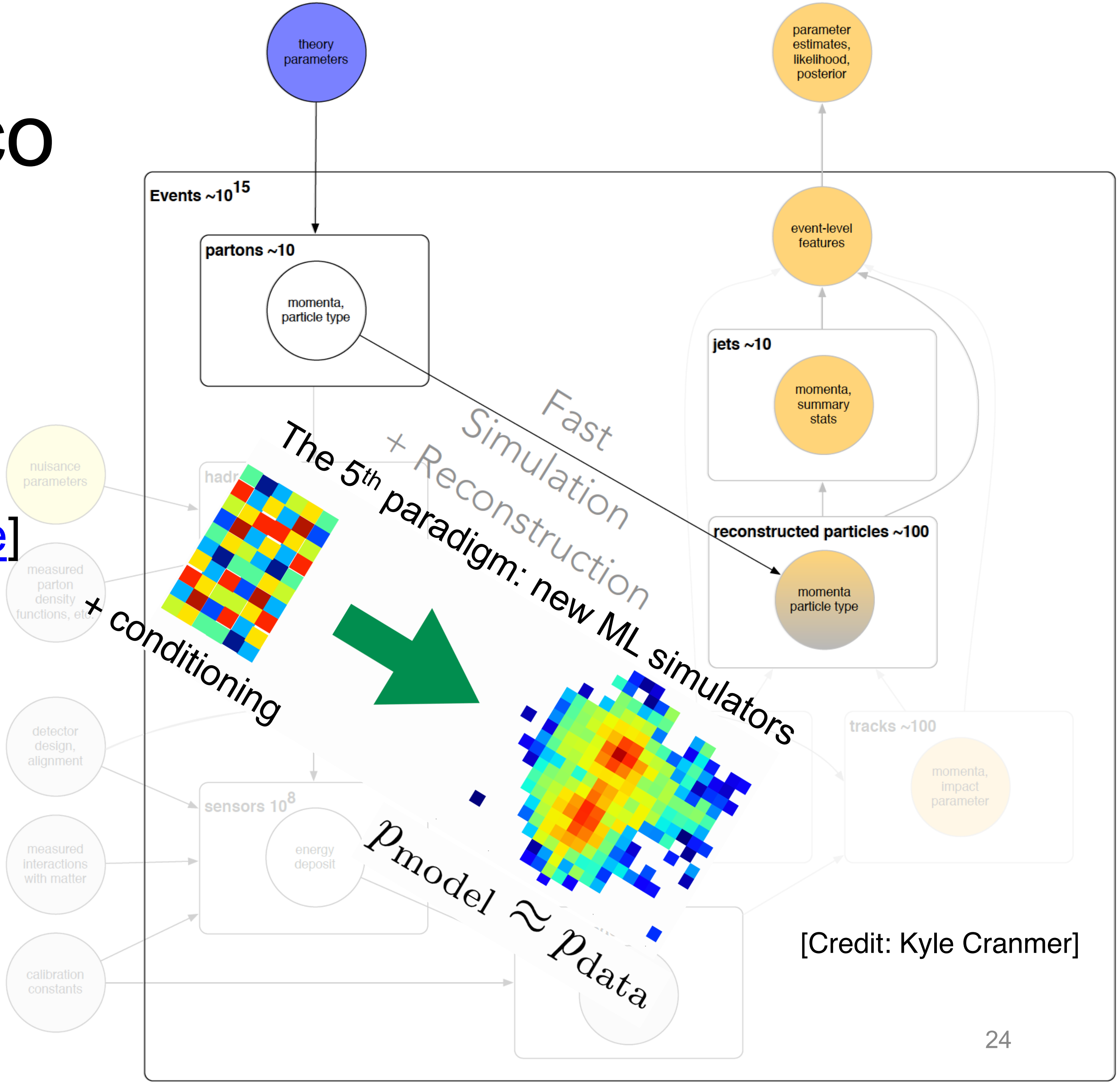


[Credit: Kyle Cranmer]

Fast Sim & Reco

Challenges:

- Fidelity, flexibility, portability
- Non-uniform geometry [FastCaloGAN, Geometry-aware]
- Sparse data
- Large dynamic range: tails
- Validation [2211.10295]
- Uncertainty
- Understanding inductive bias [GANplification]



[Credit: Kyle Cranmer]

Toolbox: generative models

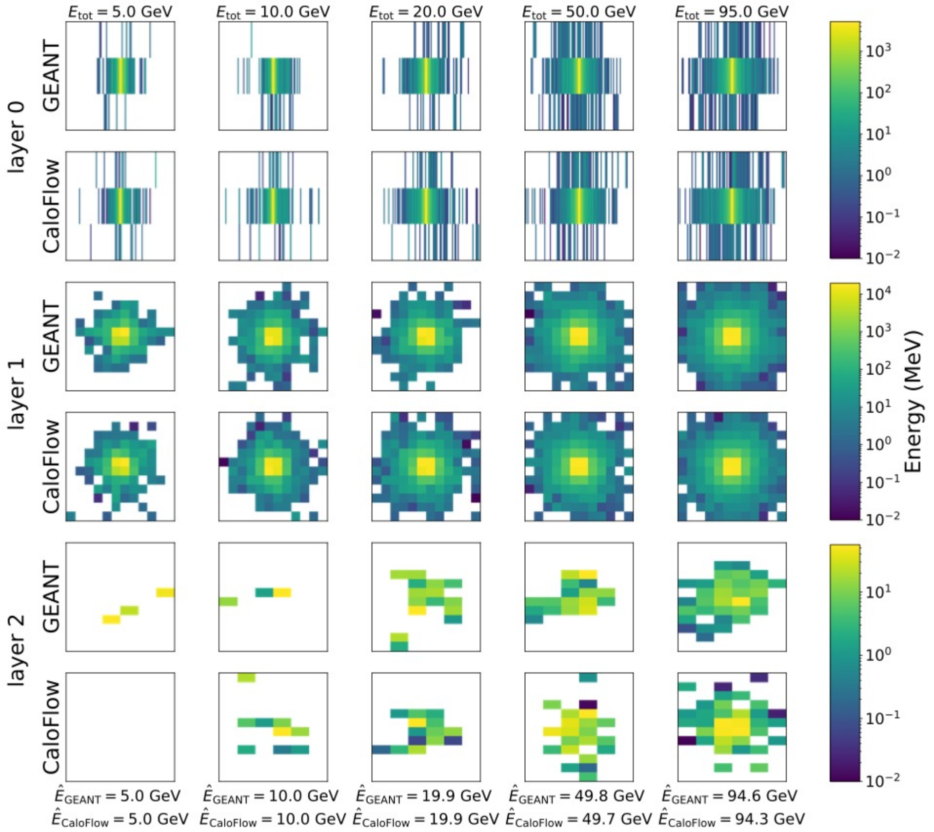
[Differentiable & fast]

Faces

[Karras et al., 2018]

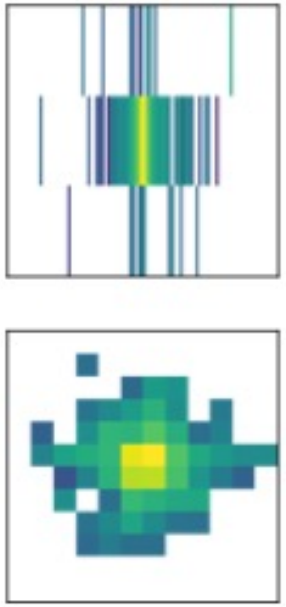
VAEs, GANs, Flows, Diffusion,…

Images of calo showers



[CaloFlow]

Images → Point cloud

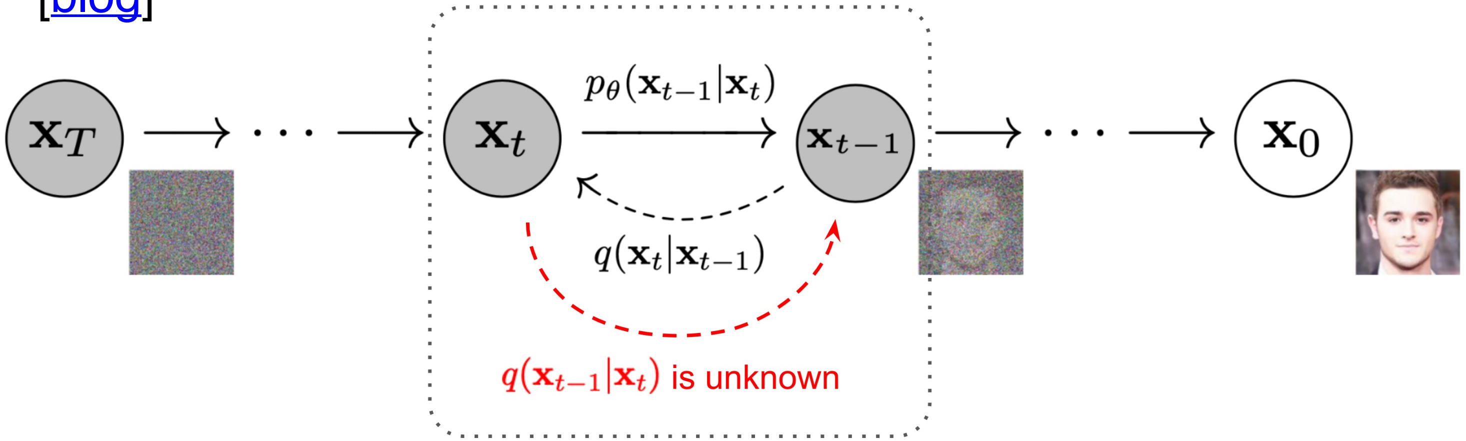


Decouple modeling
from detector geometry

- Addresses sparsity issue
- Promotes portable solutions
- Encode symmetries (inductive bias)

New on the market: point-cloud diffusion

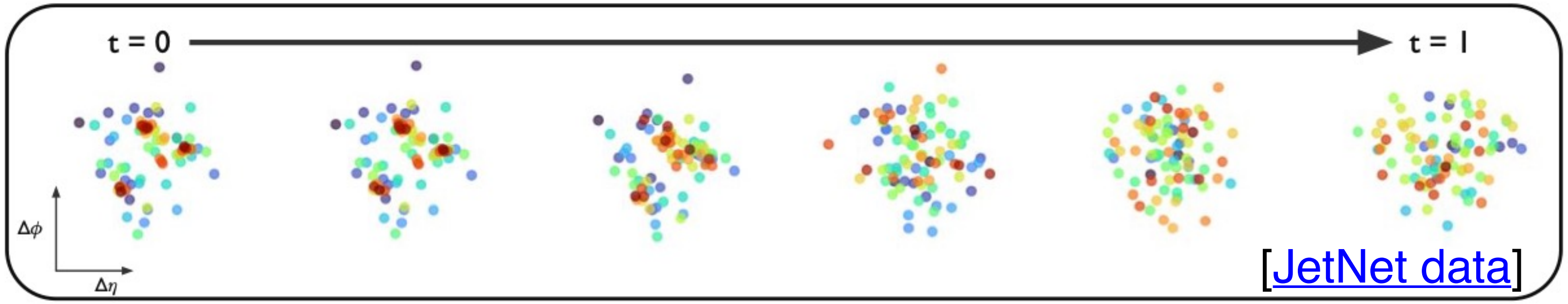
[blog]



Gradually add Gaussian noise (right-to-left=forward)

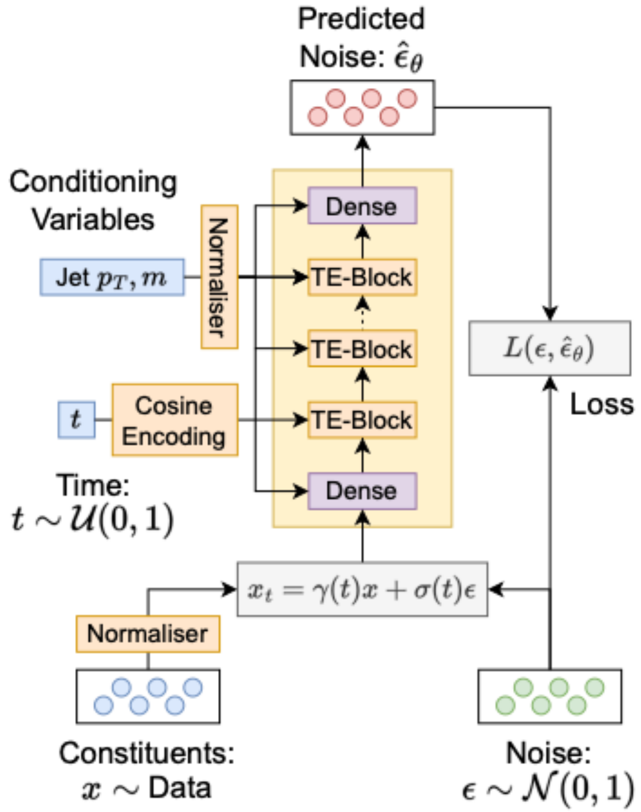
Reverse "learn the noise"

1000 → 100 → ~20 steps (over last ~year)



[JetNet data]

[PC-JeDi]



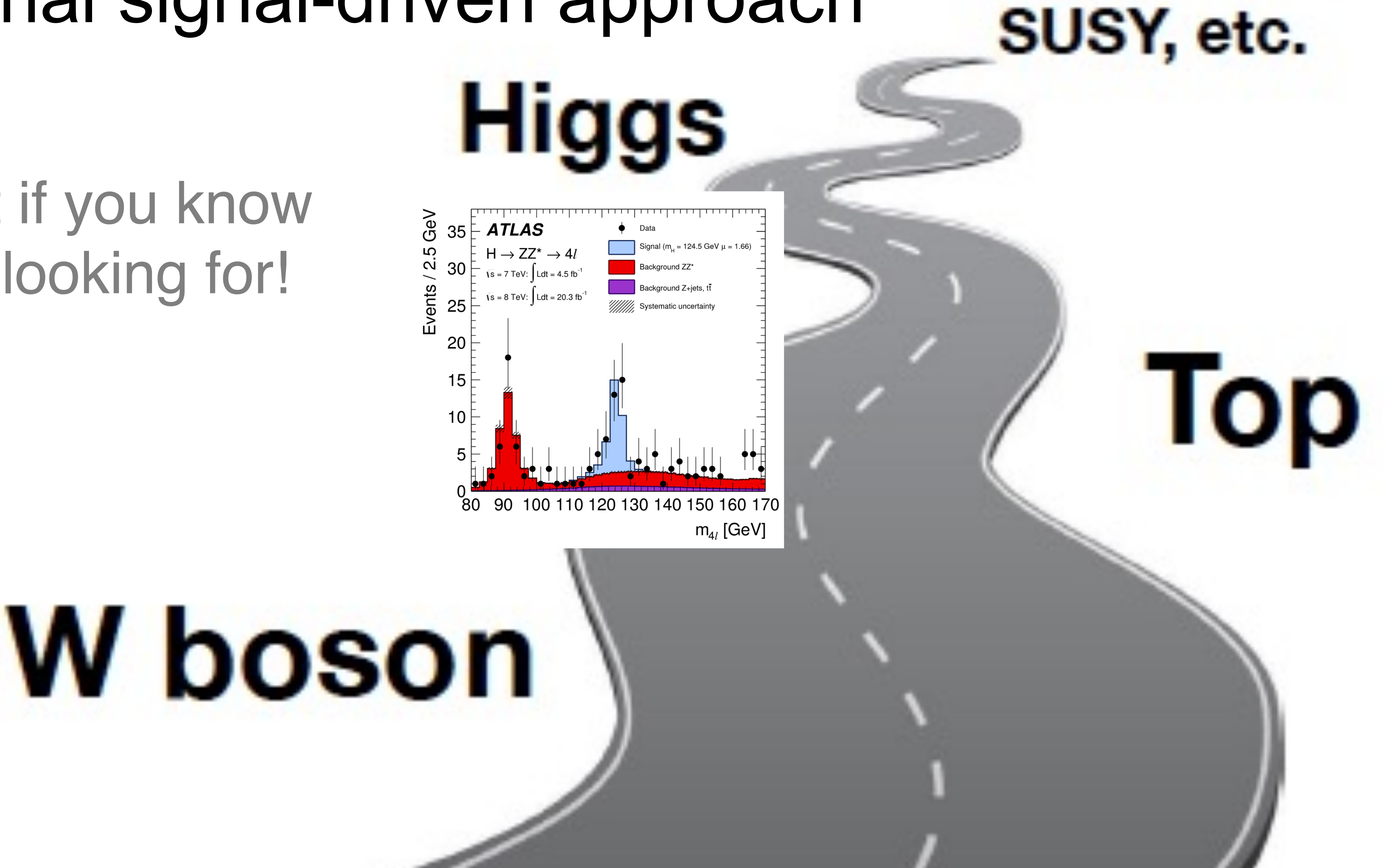
Transformer Encoder (TE) Block

[See also 2206.11898]

Search for the Unknown

Traditional signal-driven approach

Works great if you know what you're looking for!



Strategy breaks down as confidence in model decreases

Playing the lottery

How to maximize the discovery potential

Current approach is
inefficient & incomplete

	e	μ	τ	q/g	b	t	γ	Z/W	H	BSM $\to SM_1 \times SM_1$: q/g	γ/π^0's	b	$\cdots$	BSM $\to SM_1 \times SM_2$: tZ/H	bH	$\cdots$	BSM $\to$ complex: $\tau qq'$	eqq'	$\mu qq'$	$\cdots$
e	[37, 38]	[39, 40]	[39]	∅	∅	∅	[41]	[42]	∅	∅	∅	∅		∅	∅	∅	∅	[43, 44]	∅	
μ		[37, 38]	[39]	∅	∅	∅	[41]	[42]	∅	∅	∅	∅		∅	∅	∅	∅	∅	[43, 44]	
τ			[45, 46]	∅	[47]	∅	∅	∅	∅	∅	∅	∅		∅	∅	∅	[48, 49]	∅	∅	
q/g				[29, 30, 50, 51]	[52]	∅	[53, 54]	[55]	∅	∅	∅	∅		∅	∅	∅	∅	∅	∅	
b					[29, 52, 56]	[57]	[54]	[58]	[59]	∅	∅	∅		[60]	∅	∅	∅	∅	∅	
t						[61]	∅	[62]	[63]	∅	∅	∅		[64]	[60]	∅	∅	∅	∅	
γ							[65, 66]	[67–69]	[68, 70]	∅	∅	∅		∅	∅	∅	∅	∅	∅	
Z/W								[71]	[71]	∅	∅	∅		∅	∅	∅	∅	∅	∅	
H									[72, 73]	[74]	∅	∅		∅	∅	∅	∅	∅	∅	
BSM $\to SM_1 \times SM_1$: q/g										∅	∅	∅		∅	∅	∅	∅	∅	∅	
γ/π^0's											[75]	∅		∅	∅	∅	∅	∅	∅	
b												[76, 77]		∅	∅	∅	∅	∅	∅	
$\vdots$																				
$\vdots$																				

[1907.06659]

Vast signature space **un**explored

Rephrasing the problem:

Look for deviations from SM in model agnostic way

Cast a wide web

Inform future searches

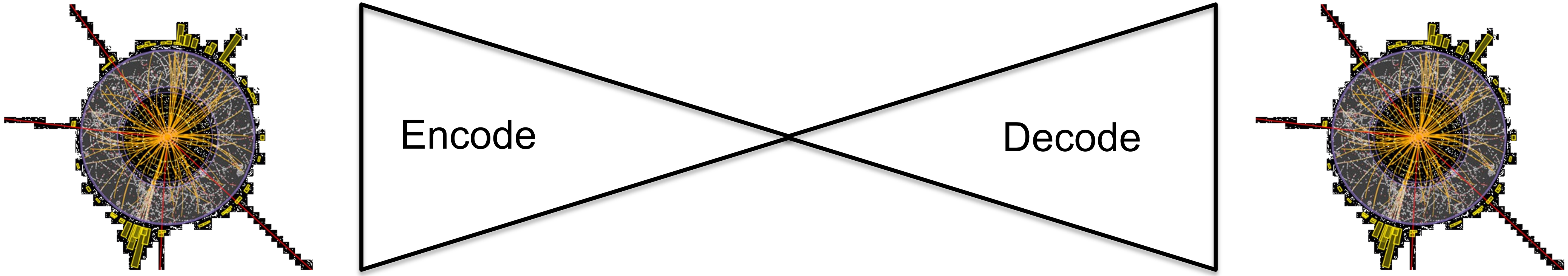
Model-agnostic search portfolio

1. Unsupervised autoencoder-style outlier detection

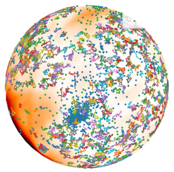
2. Semi-supervised in-situ background modeling

Fabulous idea: outlier detection with autoencoders

Train on *normal* (=SM)



Poor reconstruction = *anomaly*



[NAE]

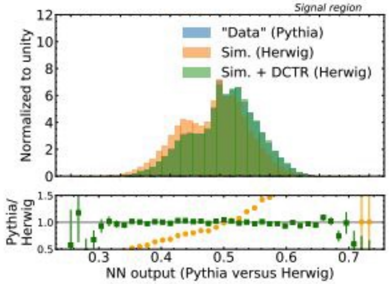
Challenges:

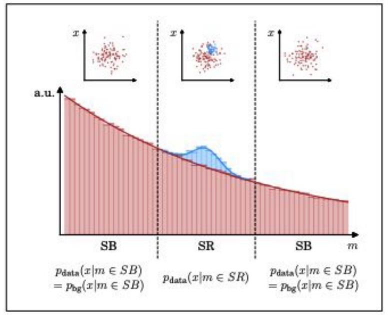
- Outlier in high-dimensional space
- Performance (e.g. anomaly metric dominated by mass)
- Add physics priors without becoming supervised

Jet level [1808.08979, 1808.08992, 2007.01850, 2301.04660…]
Event level [1806.02350, 2105.14027…]

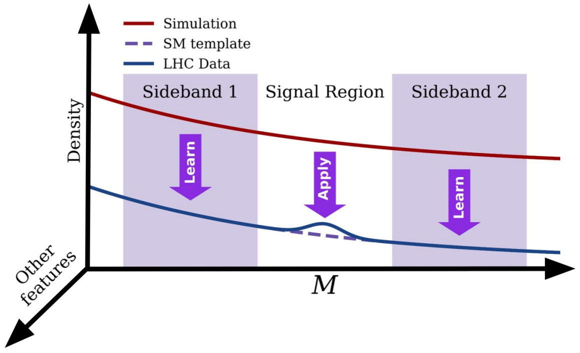
Learning high-D background templates*

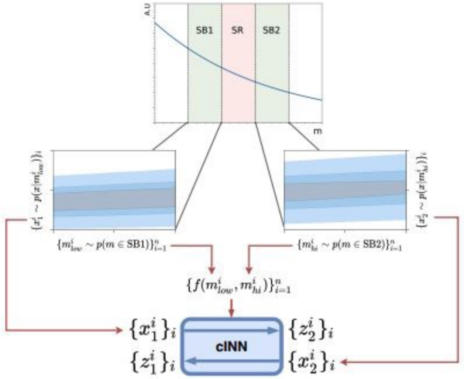
<table>
<tr><th></th><th>Learn from simulation</th><th>Learn from data (SB)</th></tr>
<tr><td>Modeling the likelihood ratio</td><td>SALAD


</td><td>CATHODE*



[*see also LaCATHODE & ANODE]</td></tr>
<tr><td>Morphing the features</td><td>FETA


</td><td>CURTAINs & Flow4Flows

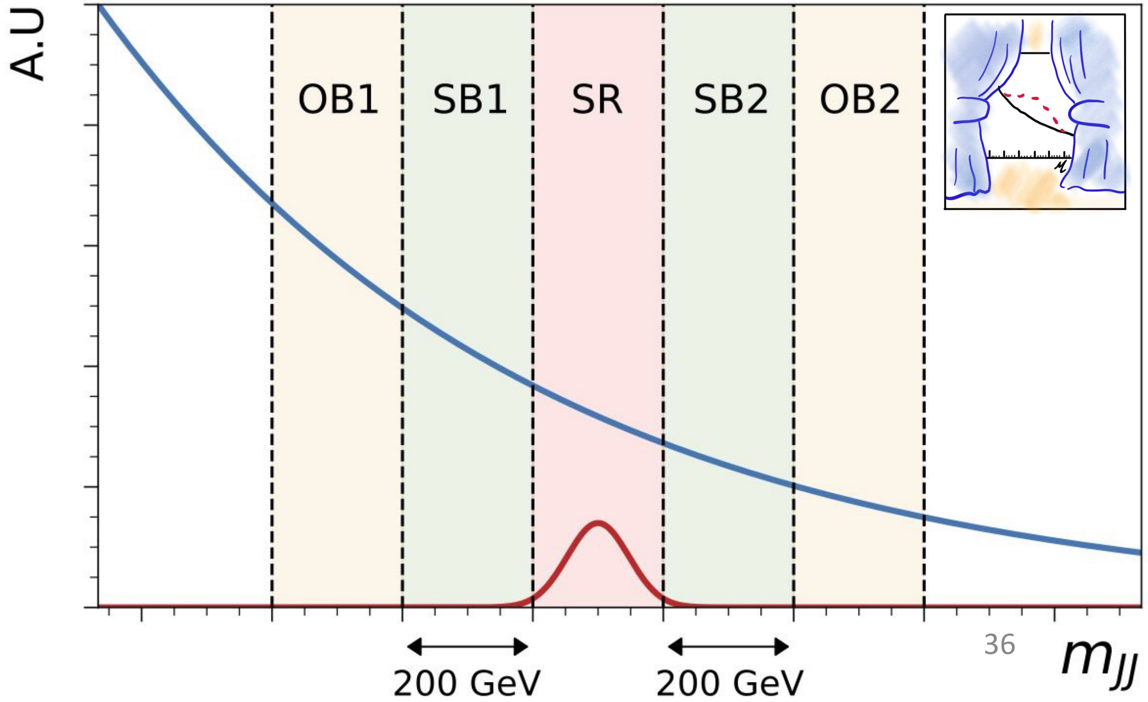

</td></tr>
</table>

[*Fidelity of simulation alone insufficient]

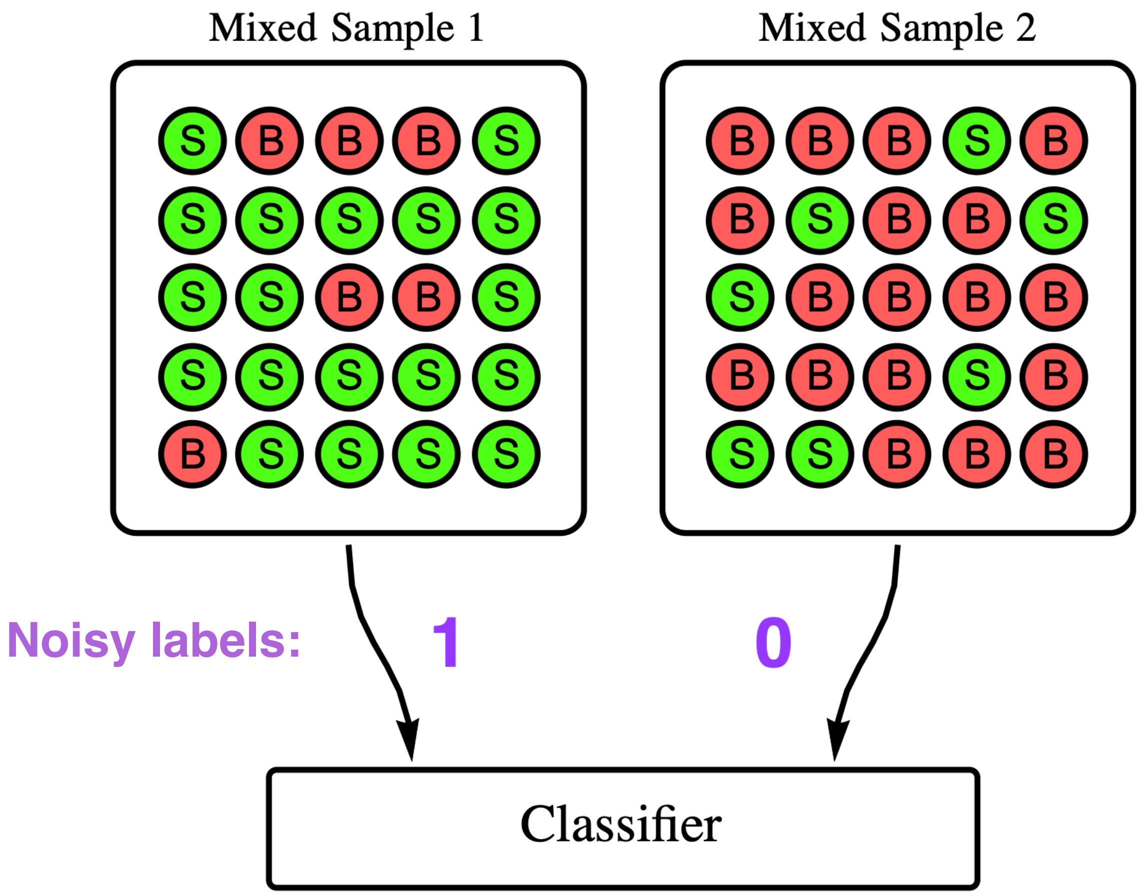
In-situ background modeling for bump hunt

[Predicted by stable-diffusion-animation]

What would a **SB background datapoint [apple tree]** look like if it had a **SR mass [age]** value?



Classification without labeling (CWoLa)



Maximize sensitivity to signal

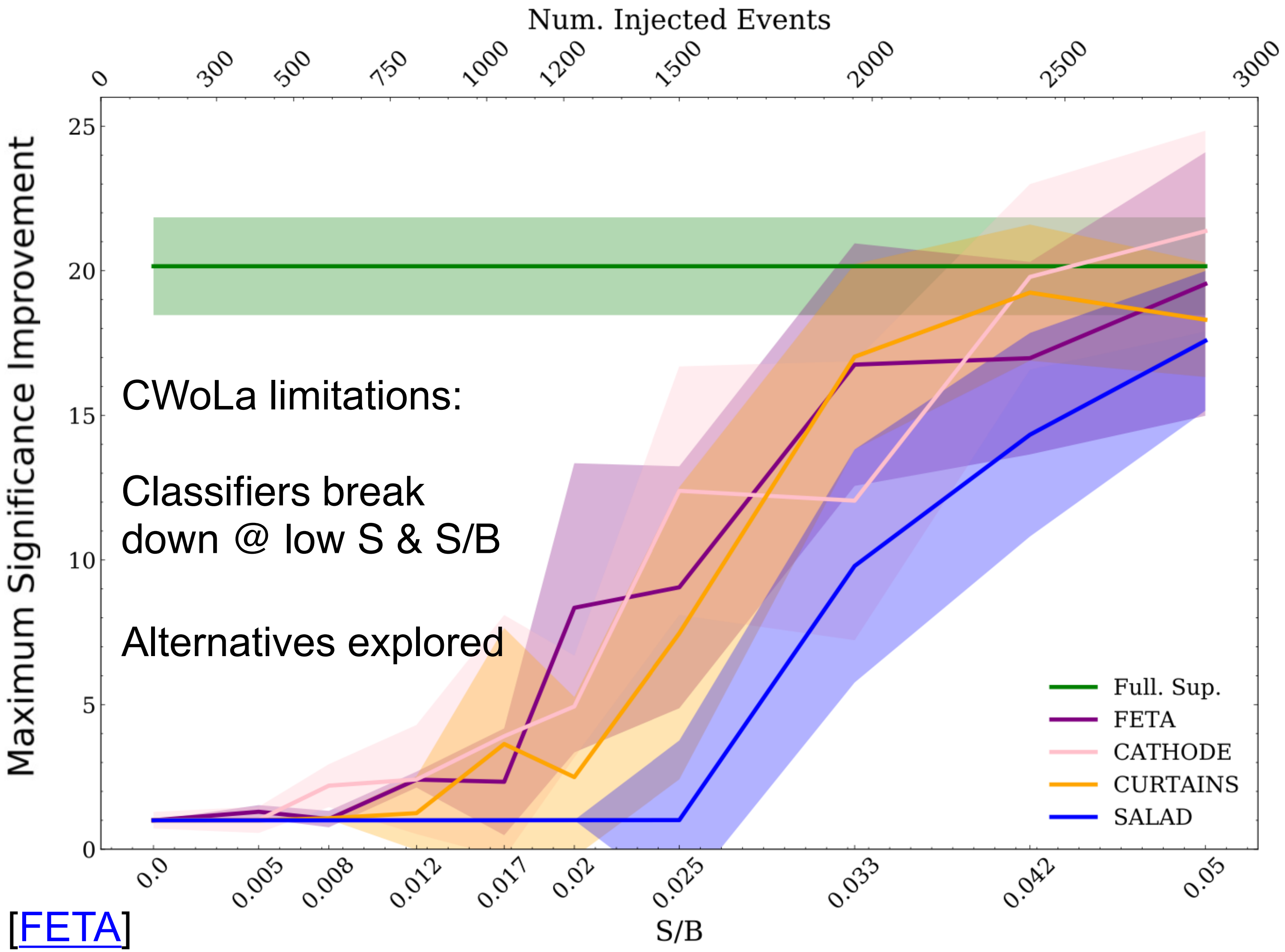
Abandon notion of *event label*

Noisy labels to be S or B

Bump hunt [1902.02634]
ATLAS analysis [2005.02983]

Beyond resonances
e.g. symmetries [2203.07529]

Comparison of methods



CWoLa limitations:

Classifiers break down @ low S & S/B

Alternatives explored

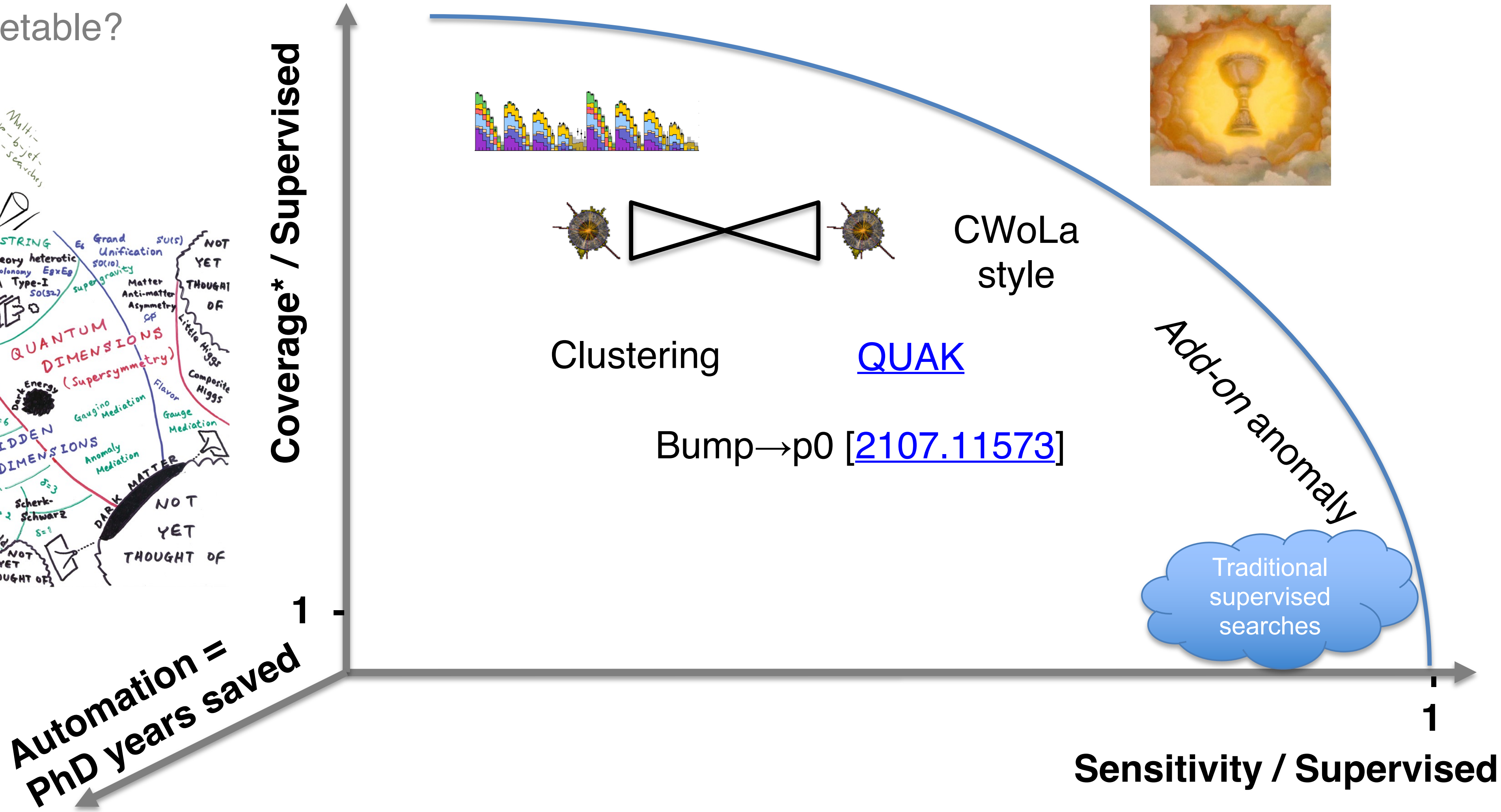
[FETA]

Similar performance of methods

Study complementarity & sensitivity to # & *noisiness* of features

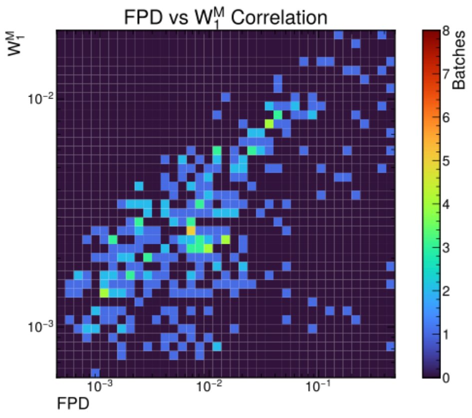
Quantifying search capability

*Volume in embedded space,
adjusted ROC: 2208.05484
Human-interpretable?



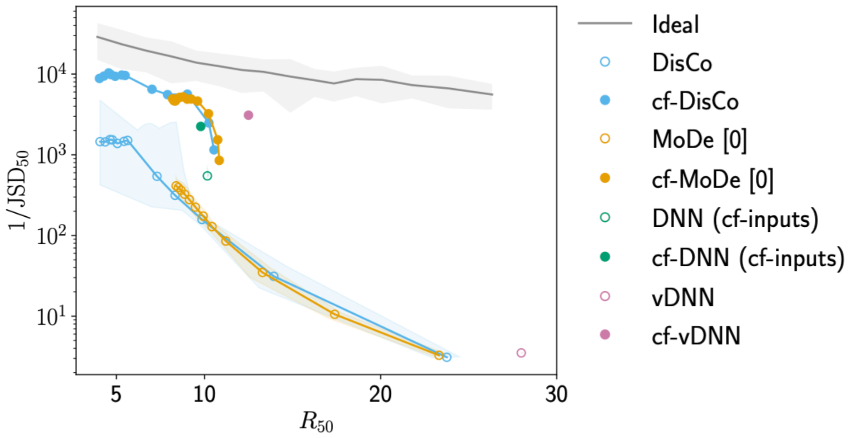
Towards a discussion

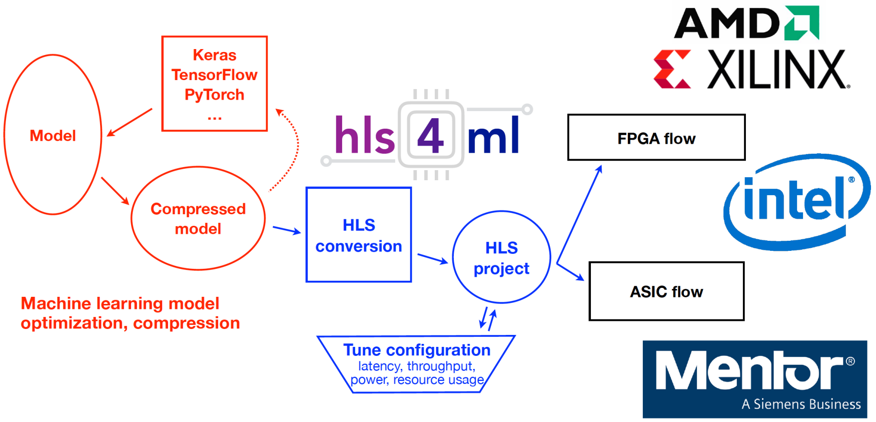
Many more challenges



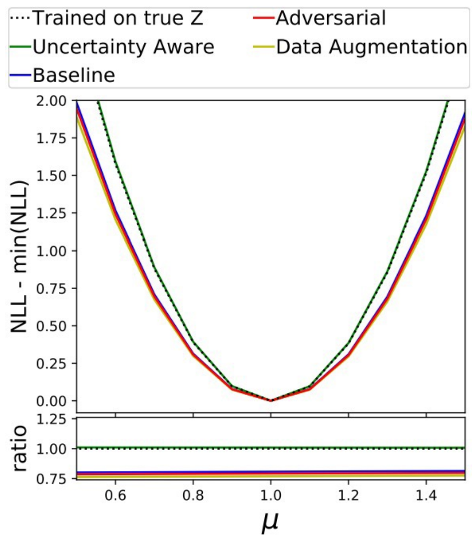
Evaluation of gen. models
[Compare metrics 2211.10295]

Decorrelation
[*Ethical AI* in Science]
[e.g. 2211.02486]





Offline → online
[On-the-edge, 1804.06913, hls4ml]



Making scientific decisions in the presence of **uncertainties**
[e.g. 2105.08742]

& Social challenges

Fast-moving ML $\leftrightarrow$ Slow Experiment time scale

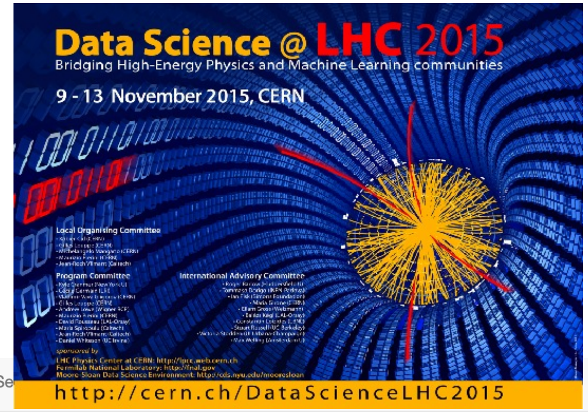
ML@HEP competitive $\leftrightarrow$ *Open Science* @ Experiment

Need faster concept-to-production cycle

& Opportunities

- AI as a *muse* to science
  - ML to suggest new theories [active learning]
- Human-in-the-loop AI
  - Optimal detector design assisted by AI
- Differentiable programming → differentiable physics
- Data analysis in theory space [simulation-based inference]
- Diverse AI-assisted search portfolio [rigor/bias/automation]
- More use of GNNs & Transformers
- Impact of diffusion & foundation models – relevance of *language* aspect? [Feynman diagrams?]
- …

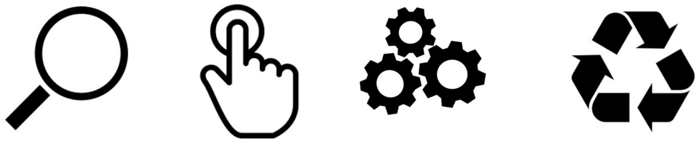
The HEP-AI ecosystem



- Workshops & long-term collaborations (with industry)
  - Synergies & cross-pollination
  - Catalyst for R&D
  - Evaluate & compare
  - Community consensus

- Common benchmarks & metrics
  - Top-tagging reference data
  - CaloChallenge
  - Anomaly challenges
  - JetNet

Findable Accessible Interoperable Reusable



Journal of Brief Ideas   Home   New idea   Trending ideas   All ideas   About   Se

Create standalone simulation tools to facilitate collaboration between HEP and machine learning community

By Kyle Cranmer, Tim Head, jean-roch vlimant, Vladimir Gligorov, Maurizio Pierini, Gilles Louppe, Andrey Ustyuzhanin, Balázs Kégl, Peter Elmer, Juan Pavez, Amir Farbin, Sergei Gleyzer, Steven Schramm, Lukas Heinrich, Michael Williams, Christian Lorenz Müller, Daniel Whiteson, Peter Sadowski, Pierre Baldi

dslhc   machinelearning   datascience   open data   simulation

Discussions at recent workshops have made it clear that one of the key barriers to collaboration between high energy physics and the machine learning community is access to training data. Recent successes in data sharing through the HiggsML and Flavours of Physics Kaggle challenges have borne much fruit, but required significant effort to coordinate.

While static simulated datasets are useful for challenges, in the course of investigating new machine learning techniques it is advantageous to be able to generate training data on demand (e.g. Refs. 1, 2, 3).
Therefore we recommend efforts be made to produce the ingredients required to facilitate such collaboration:

- Specific challenges for HEP experiments should be fully specified such that minimal domain-specific knowledge is required to attack them.
- Stand-alone simulators should be made open source. They should be developed to be easy to use without domain-specific expertise, while still being representative of real experimental challenges. Such a simulation will permit non-HEP re[searchers to generate realistic HEP datasets for training and testing. These simulators could]
  range from truth-le[vel]
  sensor arrays.
- Performance metri[c]
  solutions.

Sign in with ORCID

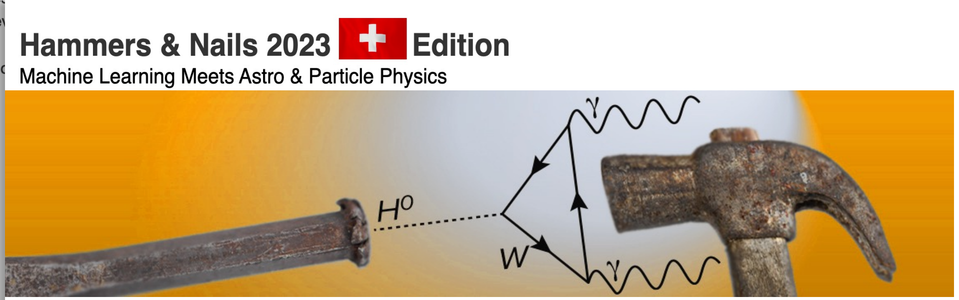
Authors

Kyle Cranmer, Tim Head, jean-roch vlimant, Vladimir Gligorov, Maurizio Pierini, Gilles Louppe, Andrey Ustyuzhanin, Balázs Kégl, Peter Elmer, Juan Pavez, Amir Farbin, Sergei Gleyzer, Steven Schramm, Lukas Heinrich, Michael Williams, Christian Lorenz Müller, Daniel Whiteson, Peter Sadowski, Pierre Baldi

Metadata

DOI 10.5281/zenodo.46864

Published: 26 Feb, 2016



Summary

- Continued success stories [e.g. object tagging]
- Transformative: automation & acceleration
  - Surrogate modeling to efficiently model complex systems
- Inject physics into AI $\Leftrightarrow$ Interpretability
- Innovation $\rightarrow$ Exploitation

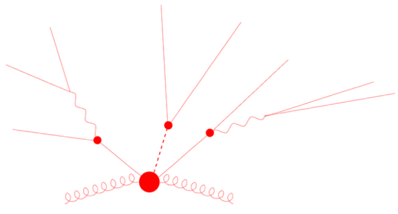
Outlook:

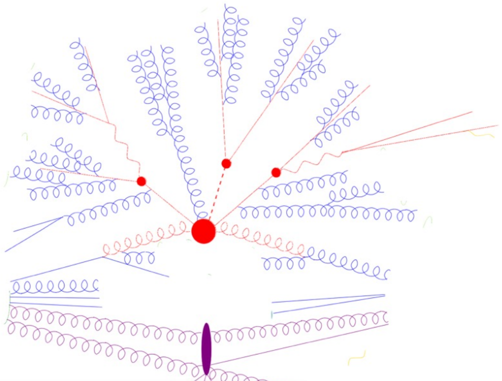
- Tackle problems which were considered unsolvable

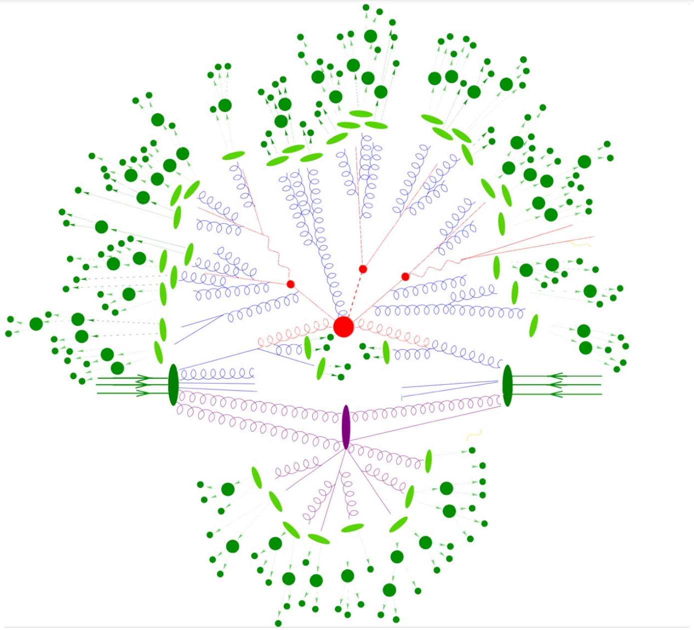
Backup

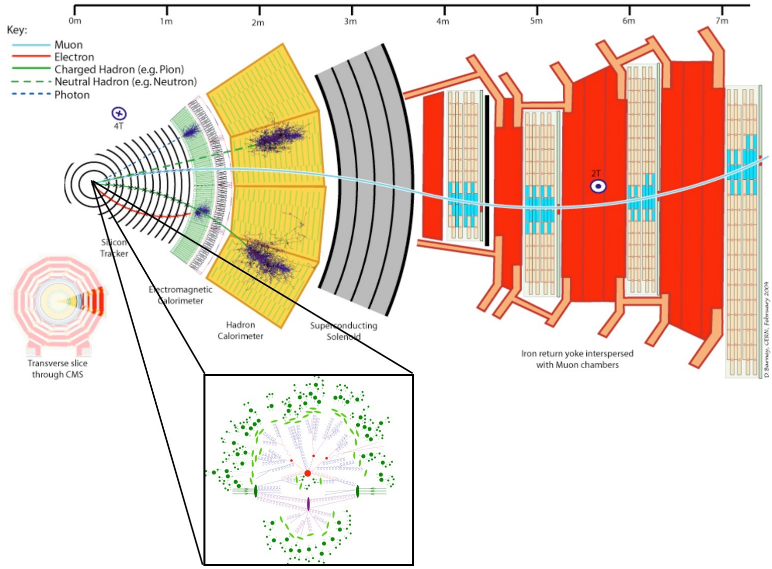
Invertible surrogates to solve inverse problem

Folding









Unfolding allows to
- Compare at theory level
- Compare between experiments
- More useful data

Hard scatter

Radiation

Hadronization

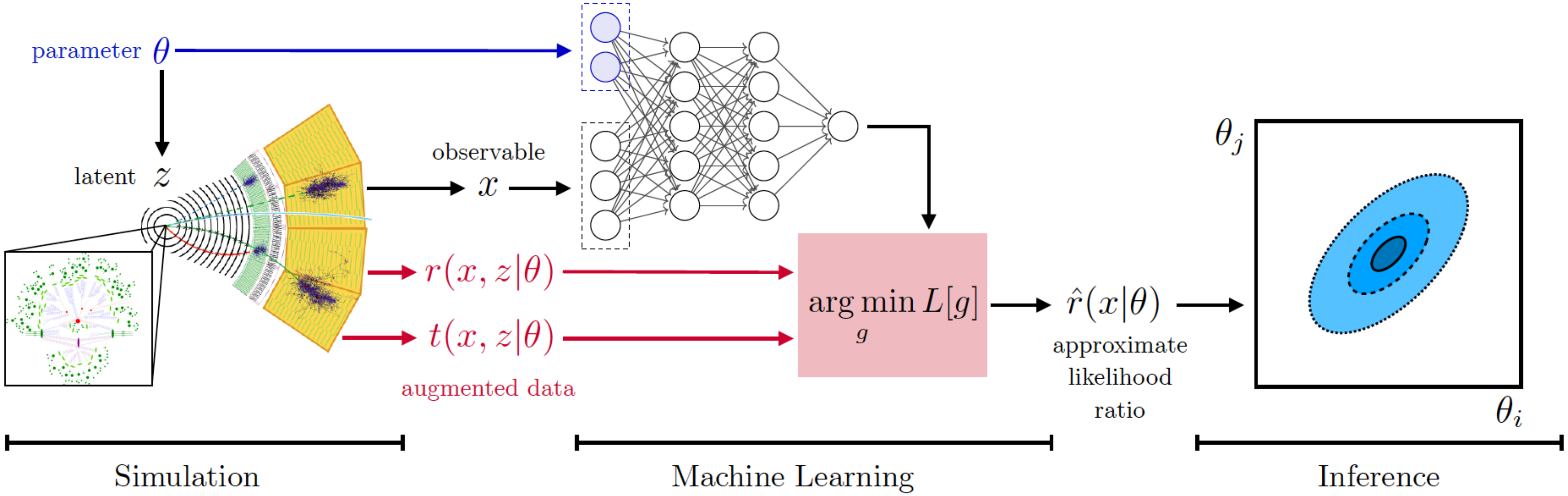
Detector

Unfolding

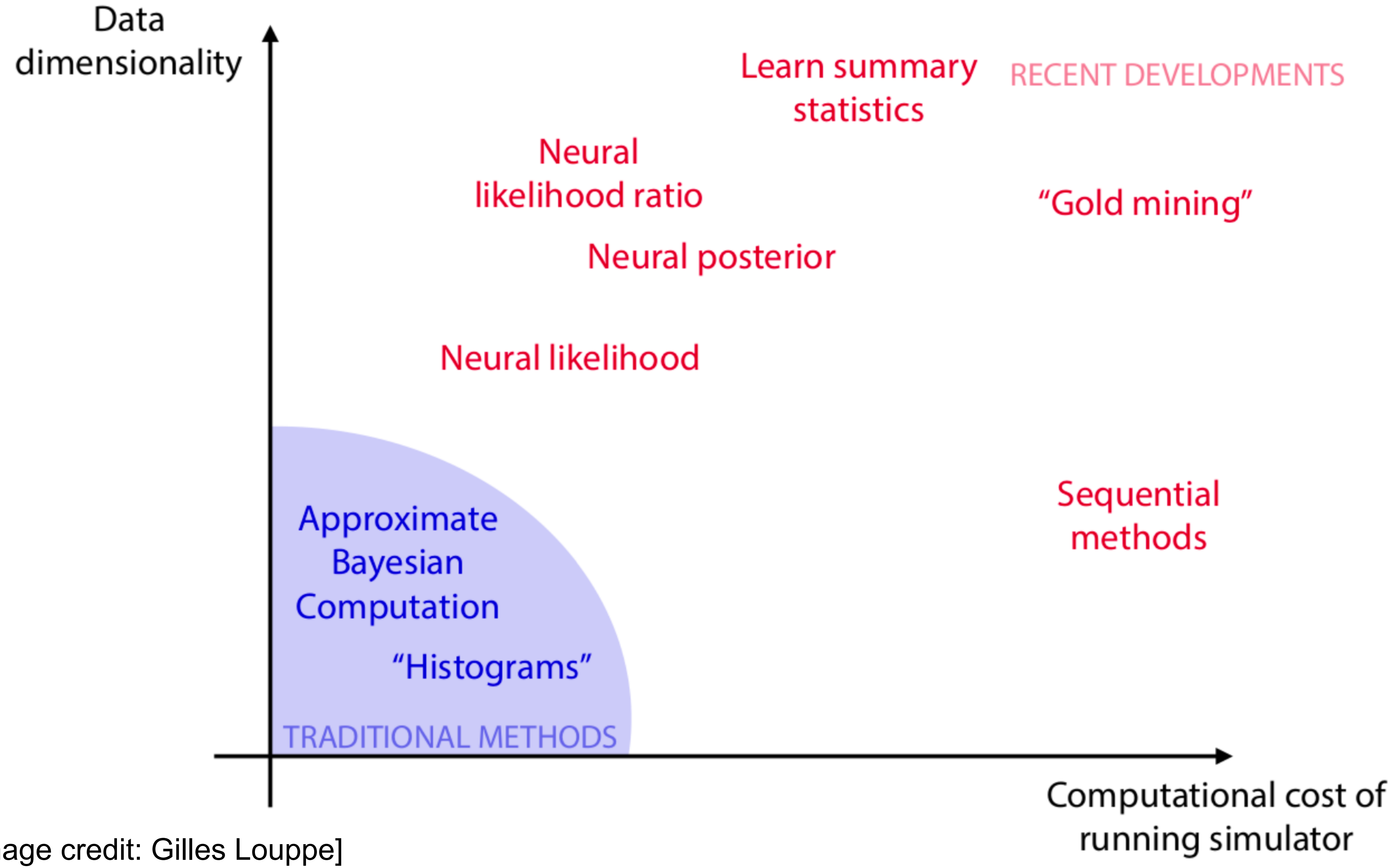
Simulation-based inference: learn p(θ|x)

accounting for latent variables [parton shower, detector effects,…]

Replace **computationally expensive numerical integrals** (MEM, NNLO event weights etc.) with a **regression** phase (ML)

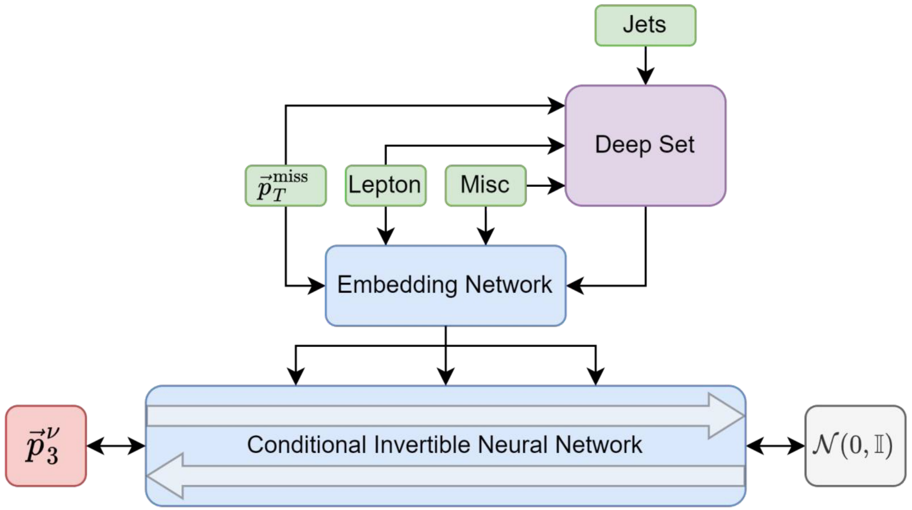


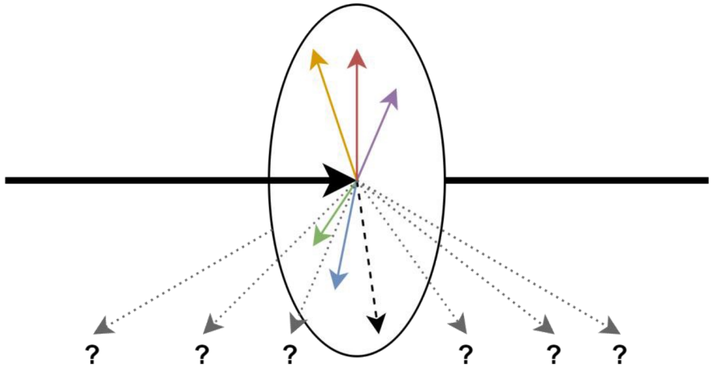
[1805.00013]

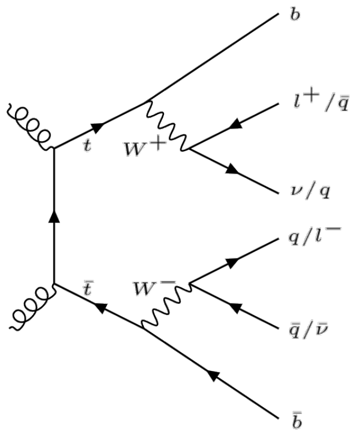


[Image credit: Gilles Louppe]

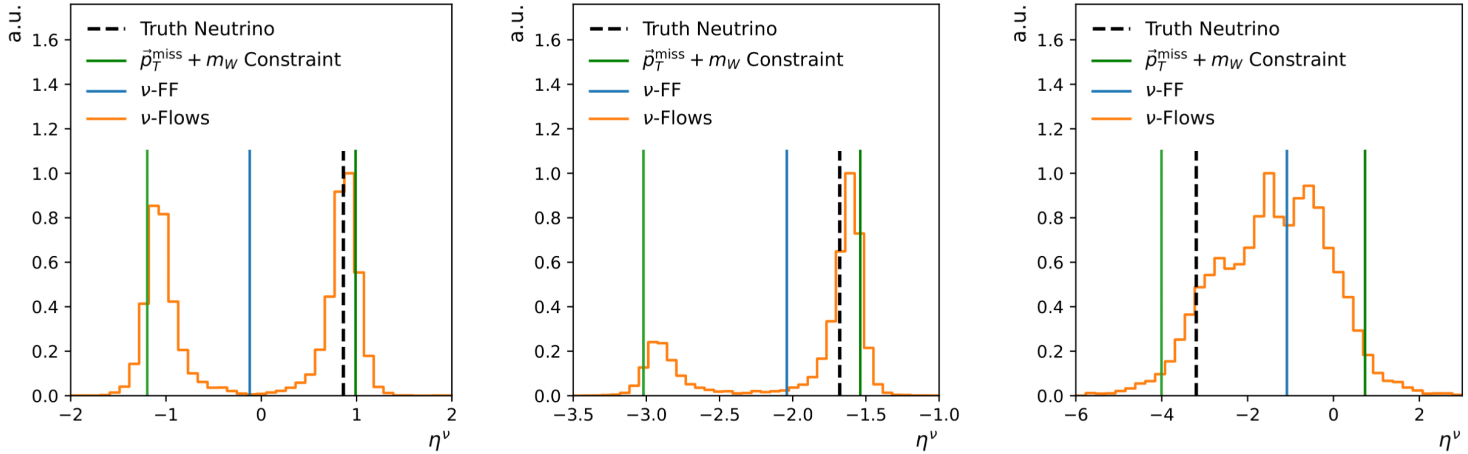
ν-Flows: Conditional Neutrino Regression







Cherry picked representative examples:

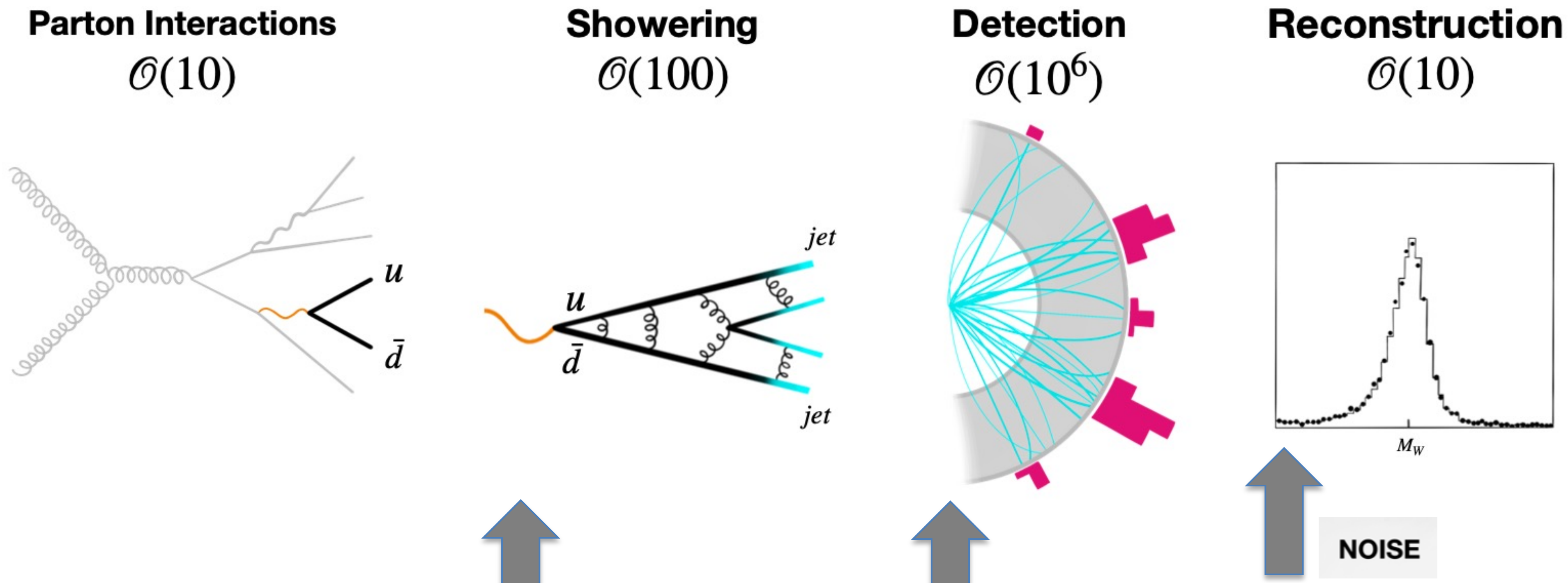


Conditional normalizing flow: learn **conditional likelihood** over neutrino momenta assuming an underlying process (inductive bias)

Improve over traditional method

[2207.00664]

Generation from noise



Create 4-vectors at analysis level
[1901.00875,1901.05282,…]

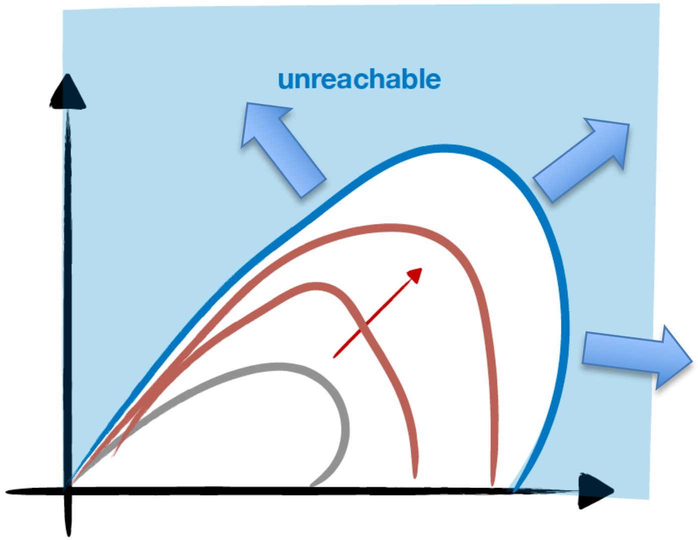
→ Conditioning

[1907.03764,…]

[1701.05927,1712.10321,2005.05334,
EPIC-GAN,…]

Opportunity: ***optimal*** detector design

Paradigm-based



Goal: optimize
p(theory | data)

[Lukas Heinrich - Detector design using differential programing]

Need **design-conditional** model *p*(x | θ,**D**)

Approximate *p*(x | θ,D) using **generative model**
→ **Fast**
→ **Differentiable**

Challenge:
p(x | **D**) without already exploring all design space **D**

Solution:
train local models as you optimize [2002.04632]

Detector design is a challenging frontier in ML@HEP
Fine-tune human design → discovery of novel designs