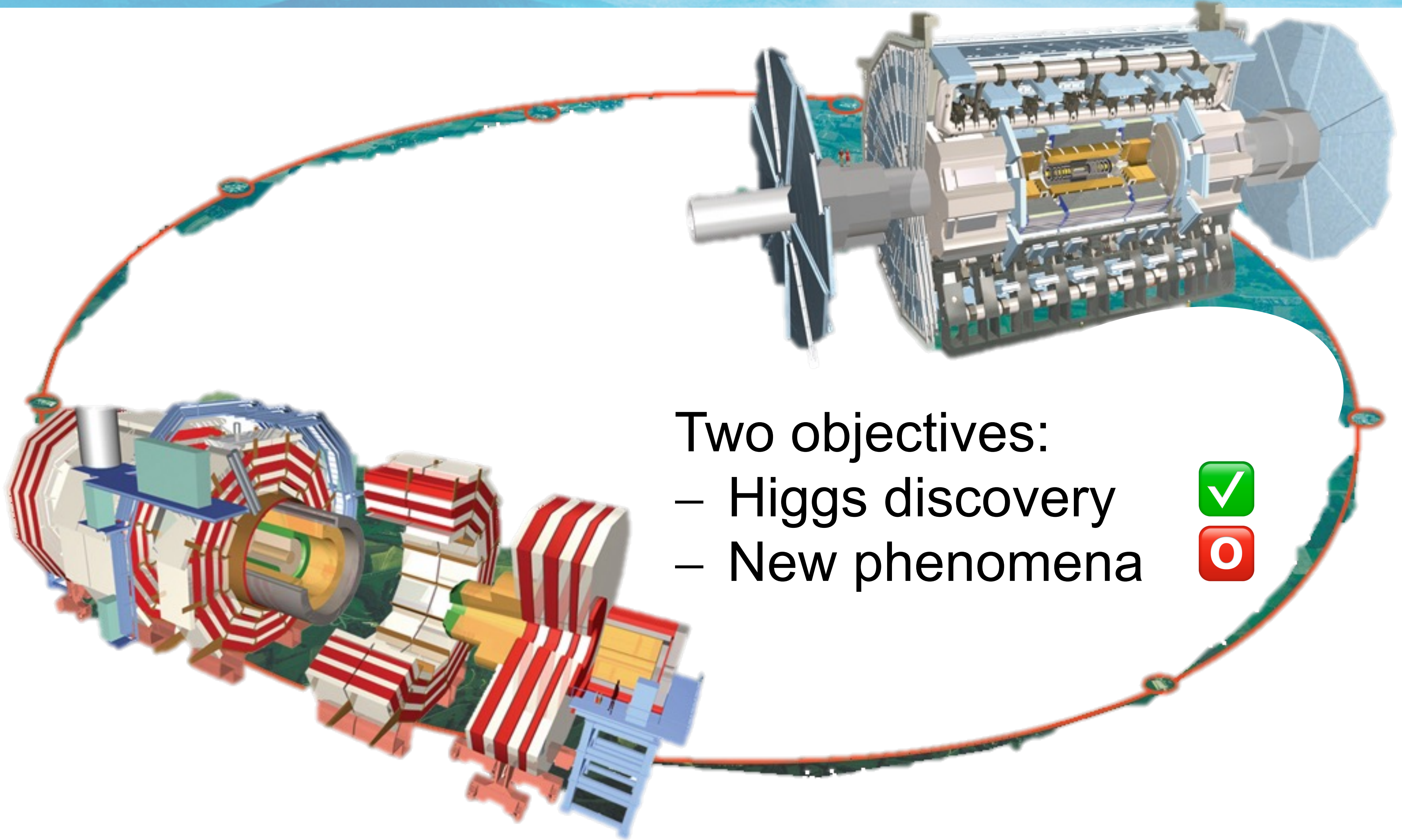
Generative models - opportunities in particle physics

Tobias Golling

UNIVERSITÉ DE GENÈVE

FACULTY OF SCIENCE

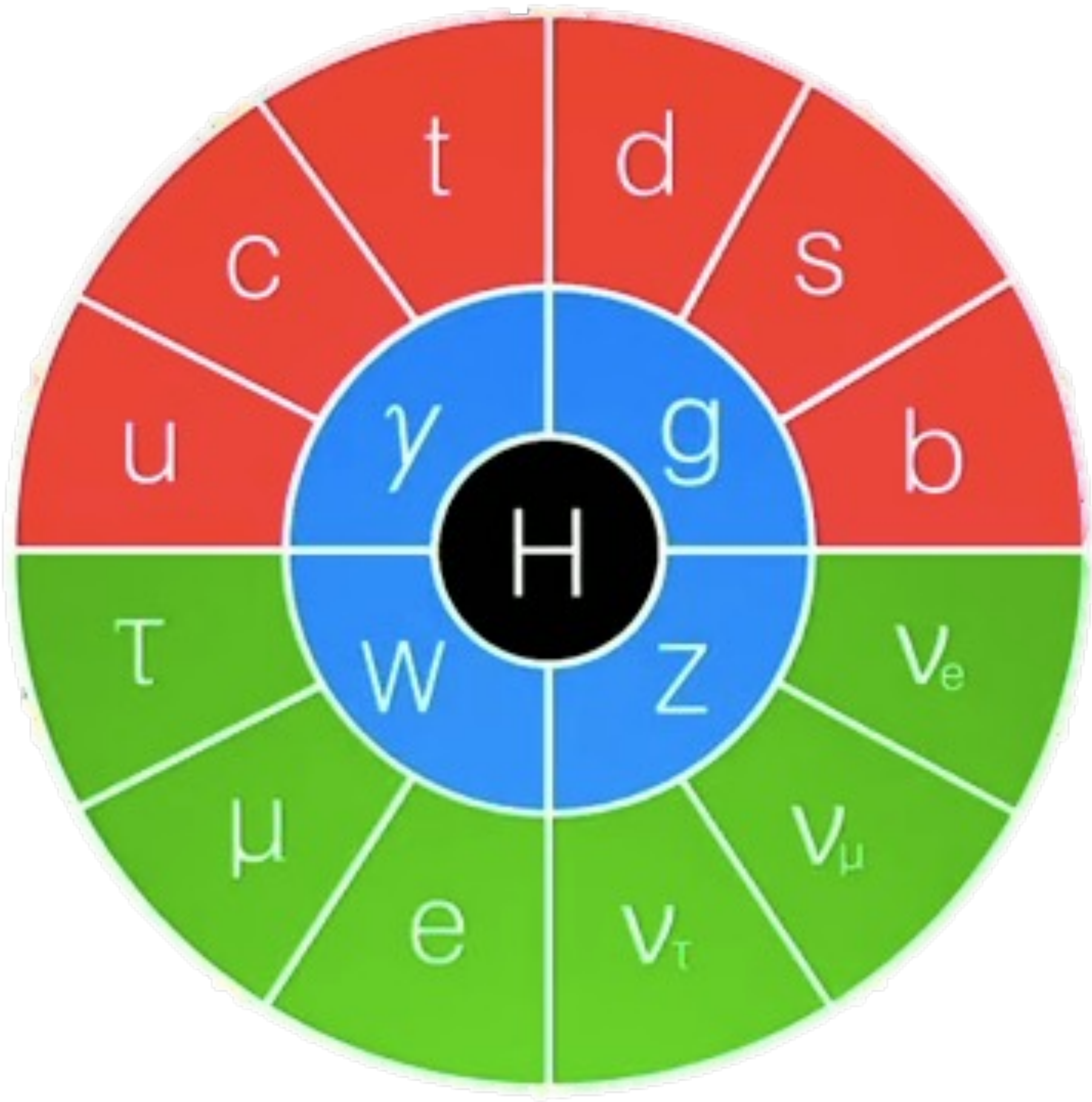
The Large Hadron Collider (LHC)



Two objectives:
- Higgs discovery ✅
- New phenomena 🅾️

The SM* is complete

Why keep going?



*the Standard Model of particle physics

Open mysteries remain

Unexplained observed phenomena

- Dark matter
- Dark energy
- Matter-antimatter asymmetry

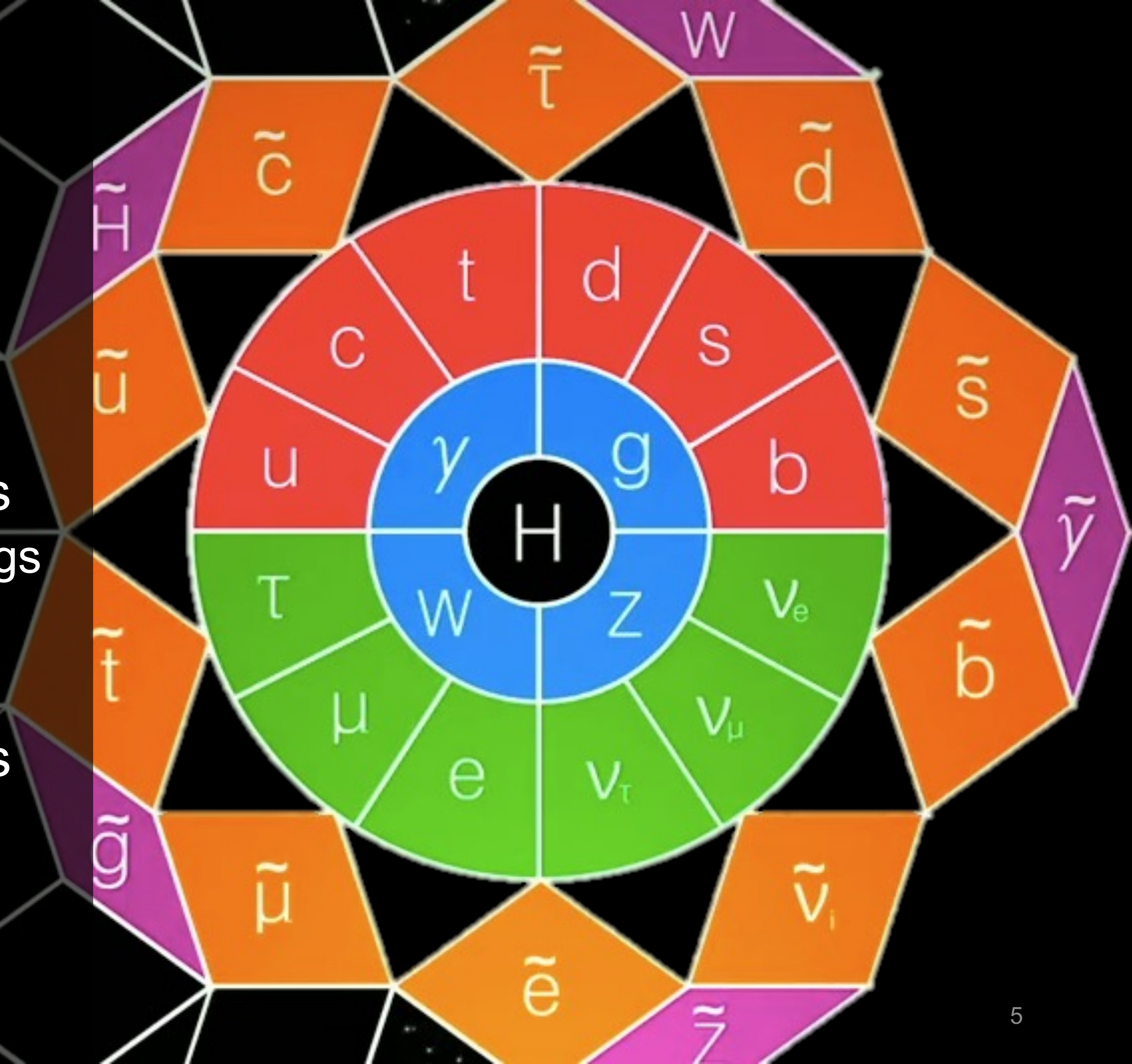
Unsatisfactory SM

- Quantum gravity, naturalness,…

The theory guidance

Hypothesize SM extensions
- Addressing SM shortcomings
- → *Testable* predictions

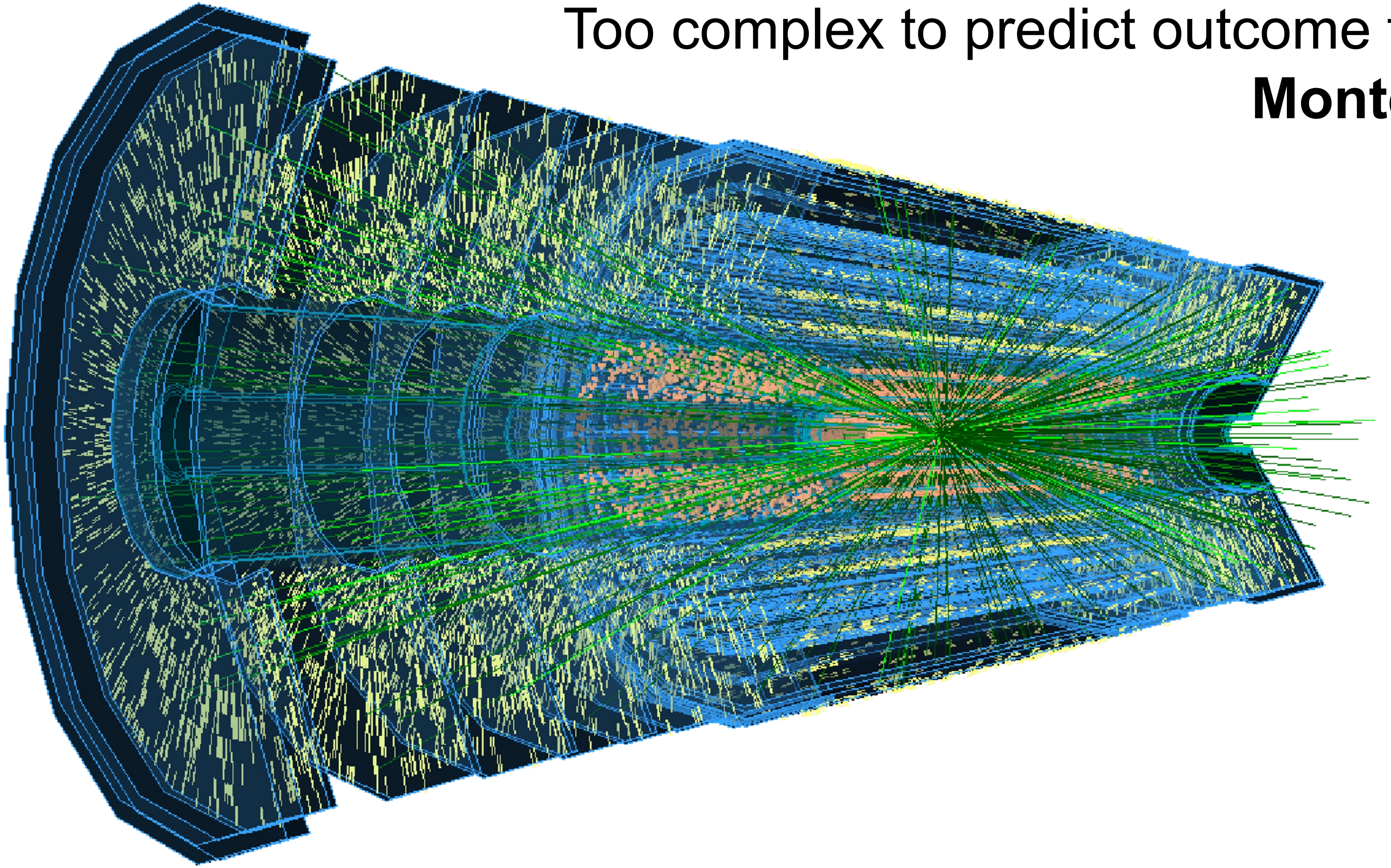
Plethora of BSM extensions



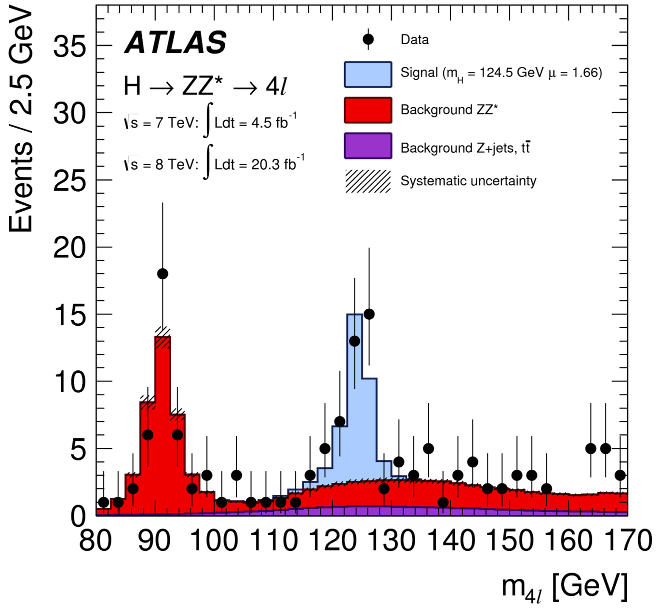
[Beyond-the-SM physics = BSM]

The need for synthetic data

Too complex to predict outcome from first principles:
Monte Carlo simulation



[Phys. Rev. D. 90 (2014) 052004]



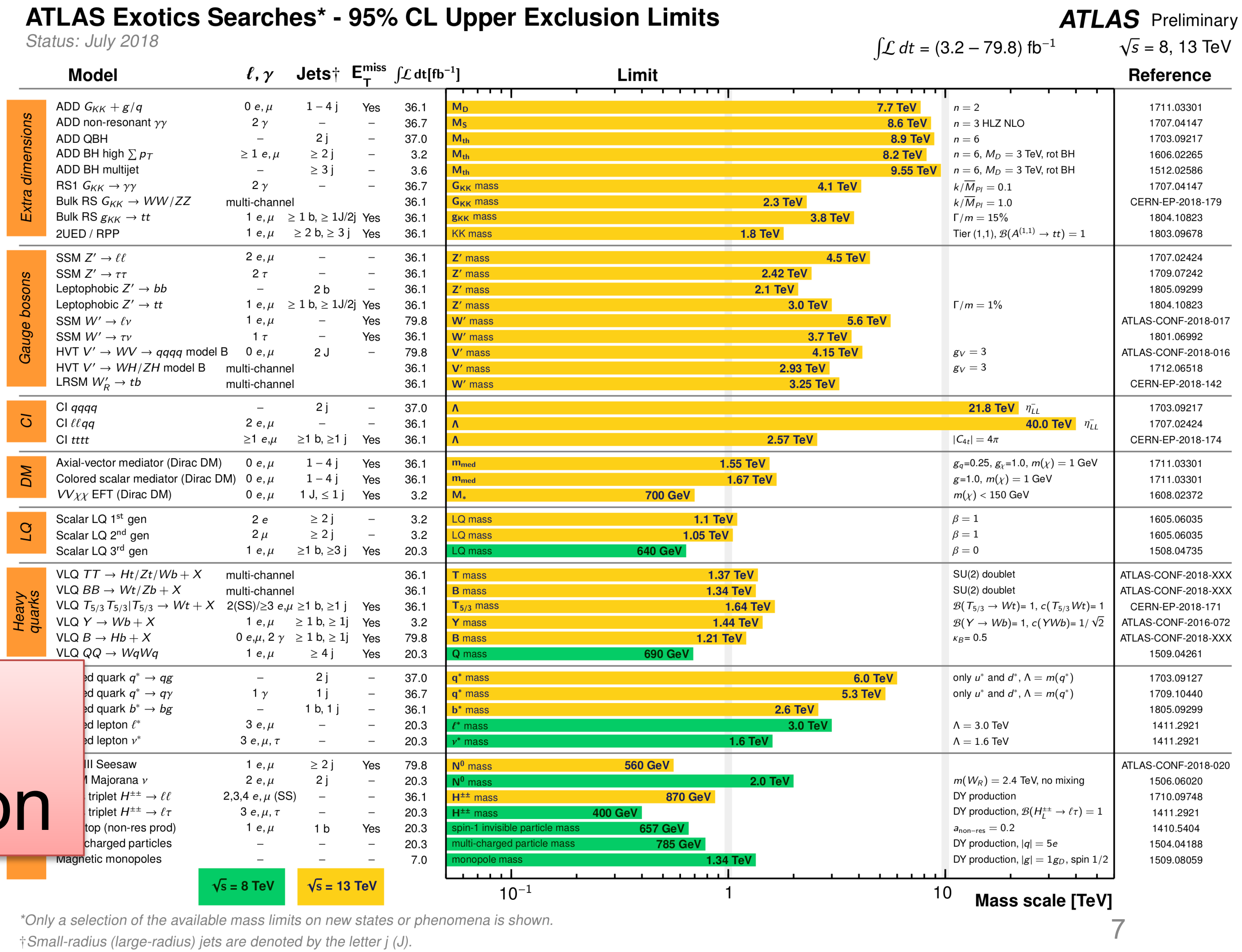
$$p(\text{data}\,|\,\text{theory})$$

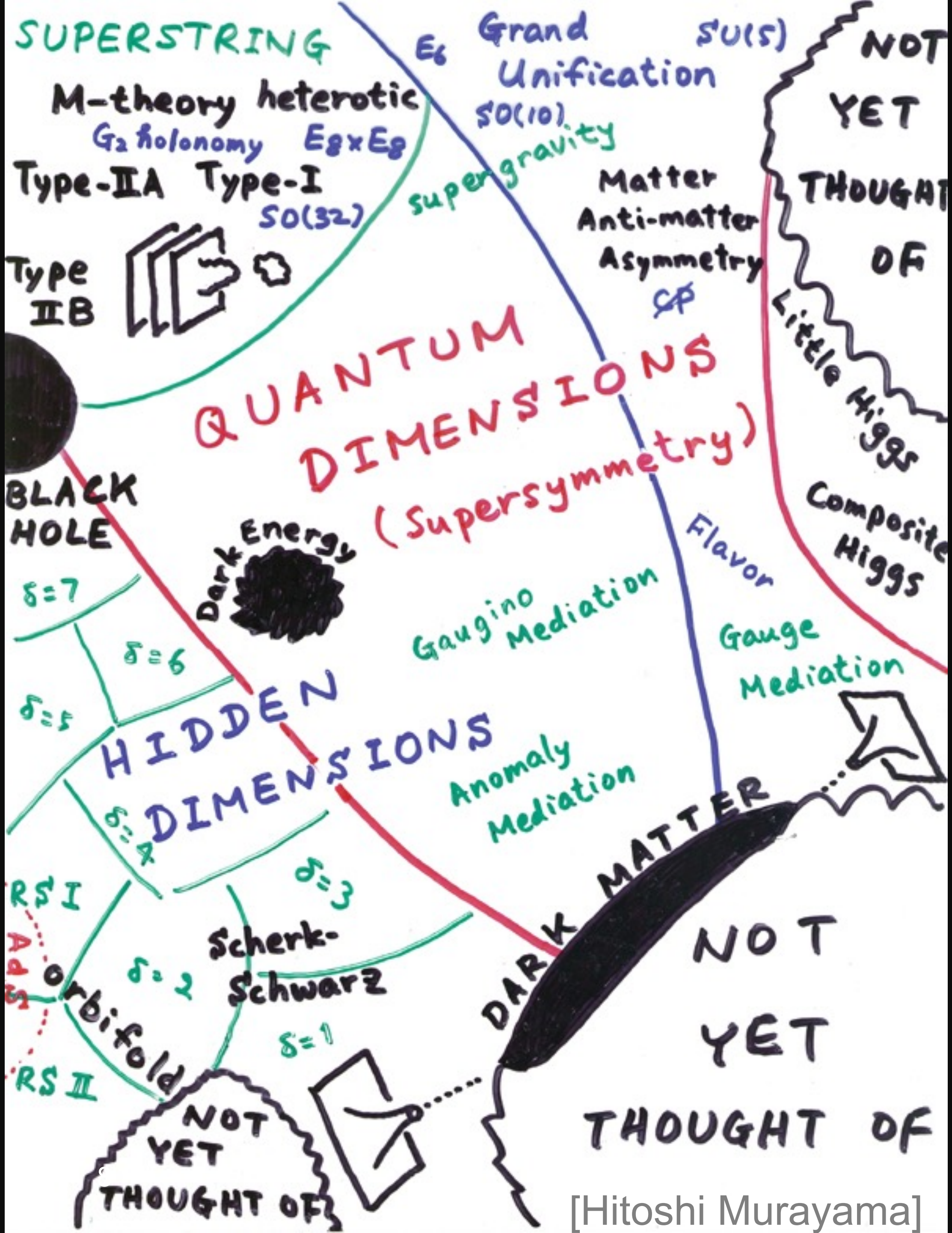
LHC interim evaluation

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

Slow-growth era of LHC: energy & luminosity

Opportunity !
Turning crank → innovation





[Hitoshi Murayama]

How to maximize knowledge gain

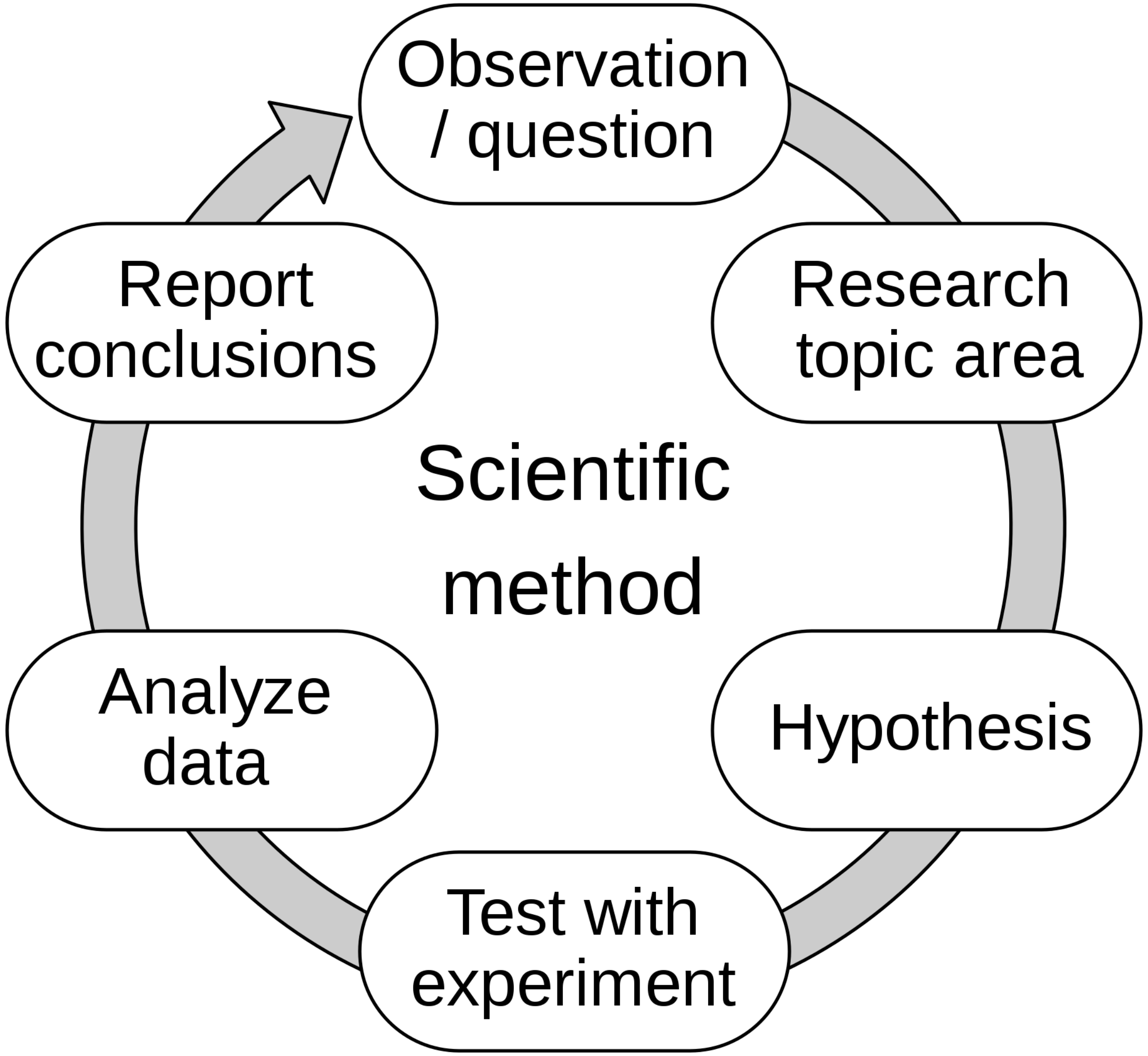
Given person-power, compute, detector, time

How to invest?

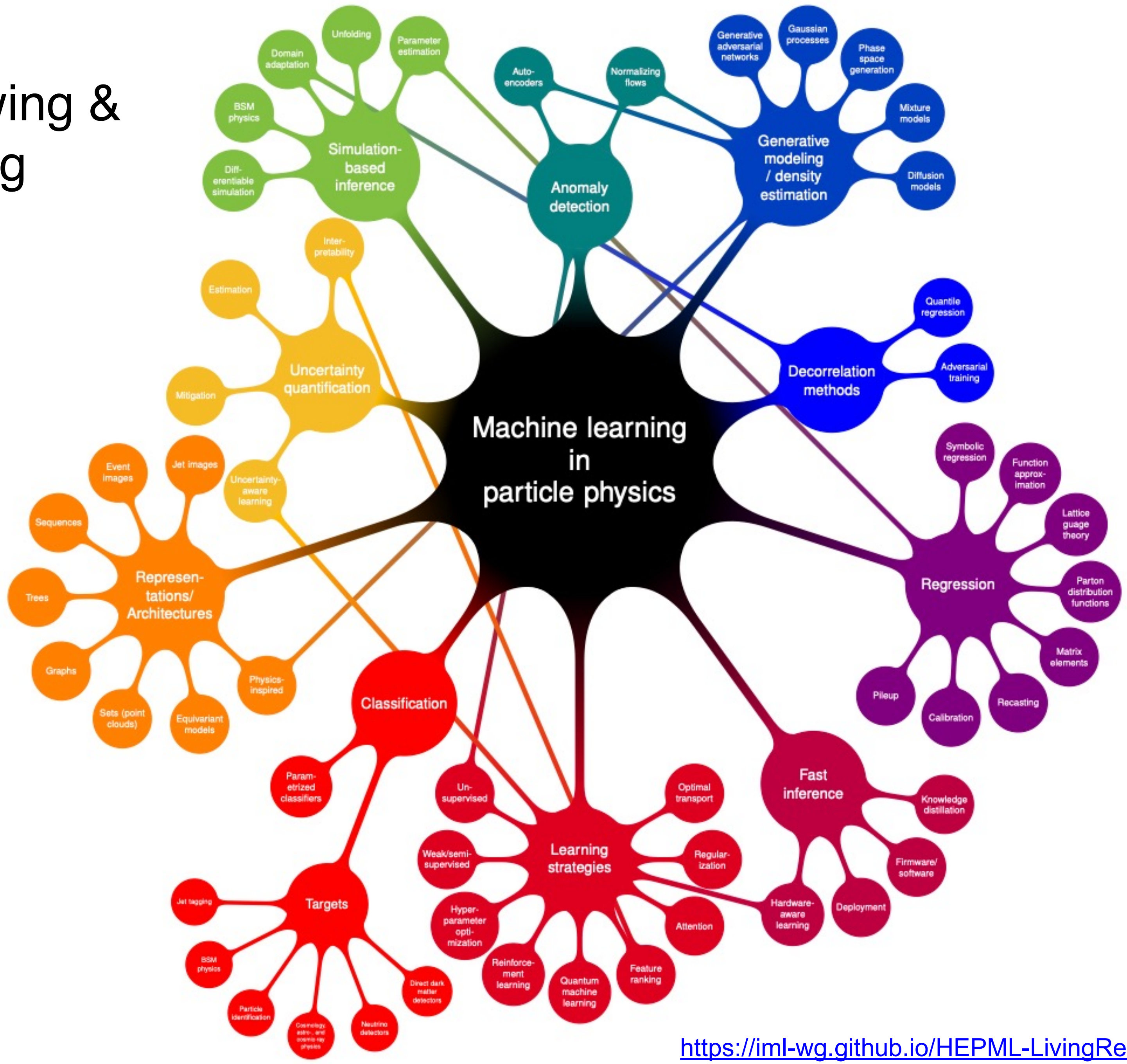
Innovate vs. exploit?

Automating & Accelerating Scientific Discovery

Enabled by generative models



Constantly growing & cross-connecting

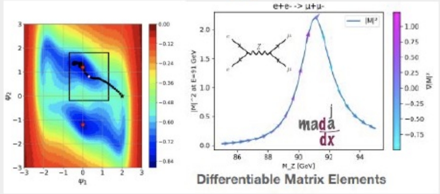


https://iml-wg.github.io/HEPML-LivingReview/
https://github.com/jmduarte/Nomological_Net_ML_Particle_Physics

[Credit: Kazuhiro Terao]

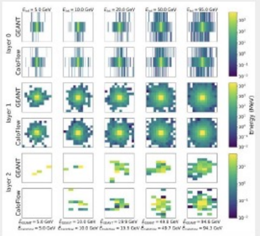
Today: AI/ML everywhere in our workflow

Differentiable Surrogate



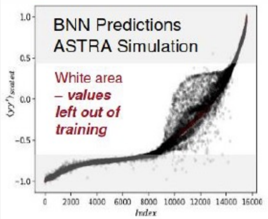
Design Optimization

Generative Models

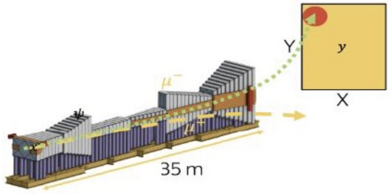


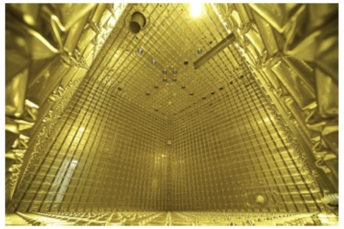
Fast Sim., Stochastic modeling

Ensemble, Bayesian NN, Temperature Scaling



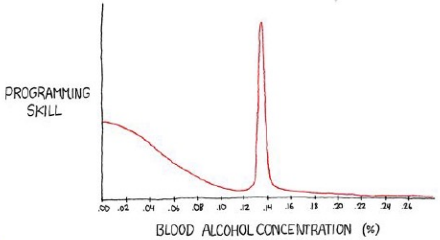
Uncertainty Quantification

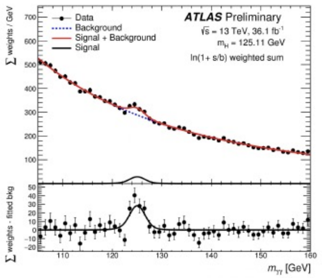




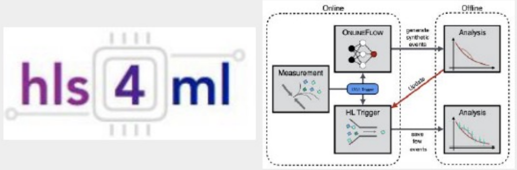
EXPERIMENT DESIGN → BUILDING / INSTALLATION → DATA TAKING / FACILITY OPERATIONS → CALIBRATION PHYSICS INFERENCE → KNOWLEDGE UPDATE → HYPOTHESIS BUILDING → EXPERIMENT DESIGN

SIMULATION



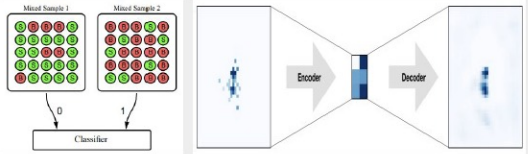


Fast/Edge-ML



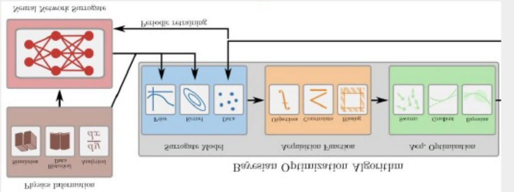
Trigger/Compression

Anomaly Detection



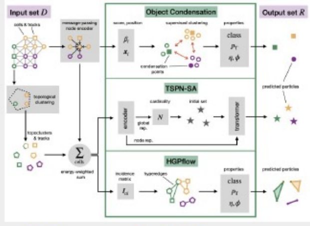
Rare Event, Diagnostics

BO/RL



Control Optimization

CV, Geometric ML



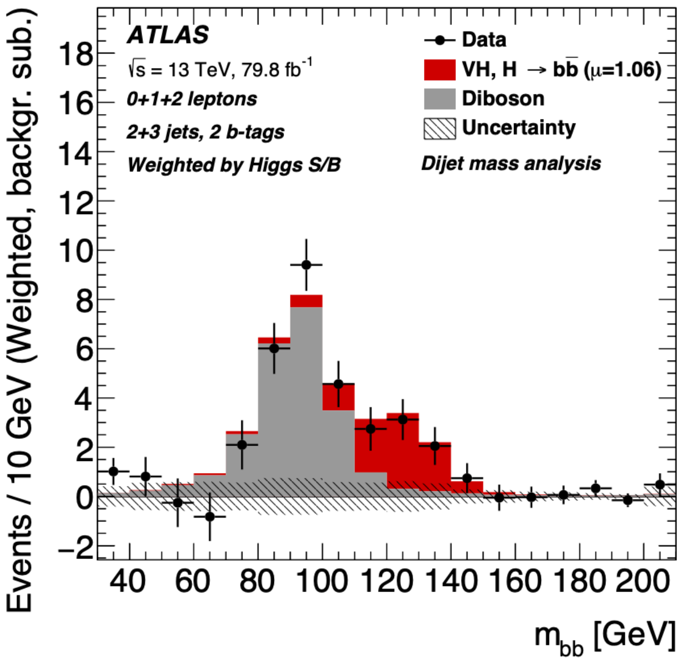
Rare Event, Diagnostics

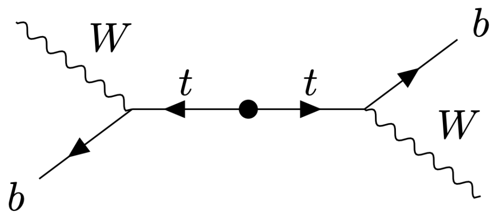
ML@HEP success story

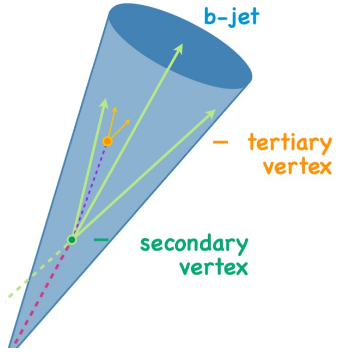
Flagship ML@HEP

Flavor tagging

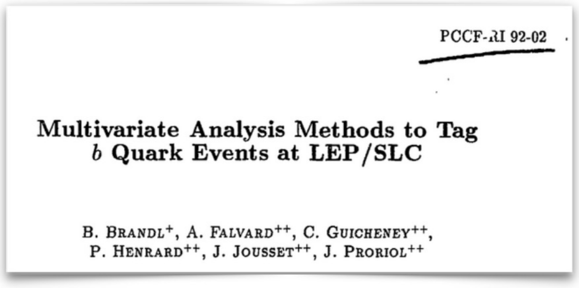
Enabler:
Higgs, top, new phenomena,…





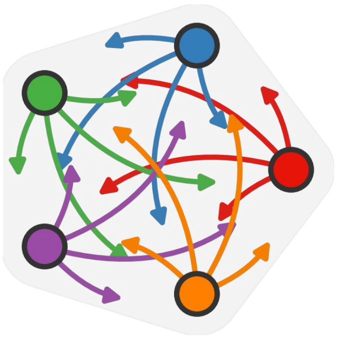


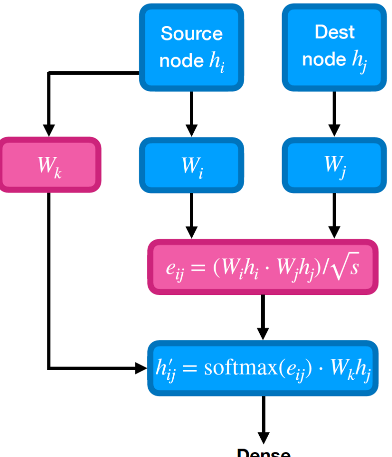
Long history of ML in flavor tagging


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++





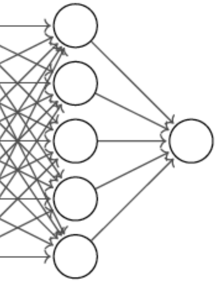
[1706.03762]

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

A lot has been learned:
- Flexible multi-classification
- Hand-designed → end-to-end
- Benefit of auxiliary tasks
- Evolving data representations

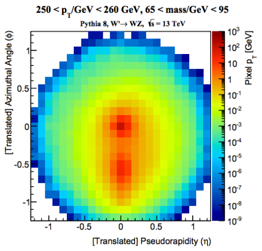
Evolving data representations in HEP

Arbitrary inputs
FF NN



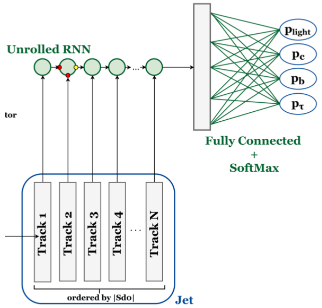
⇒

Images
CNN
[1511.05190]



⇒

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

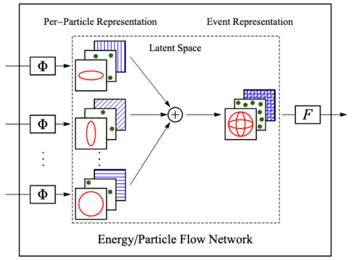


⇒

Point clouds

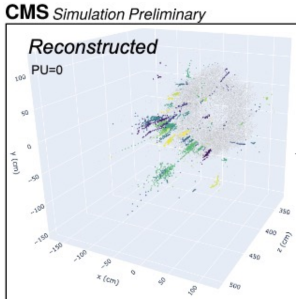
⇒

Deep Sets
[1810.05165]



⇒

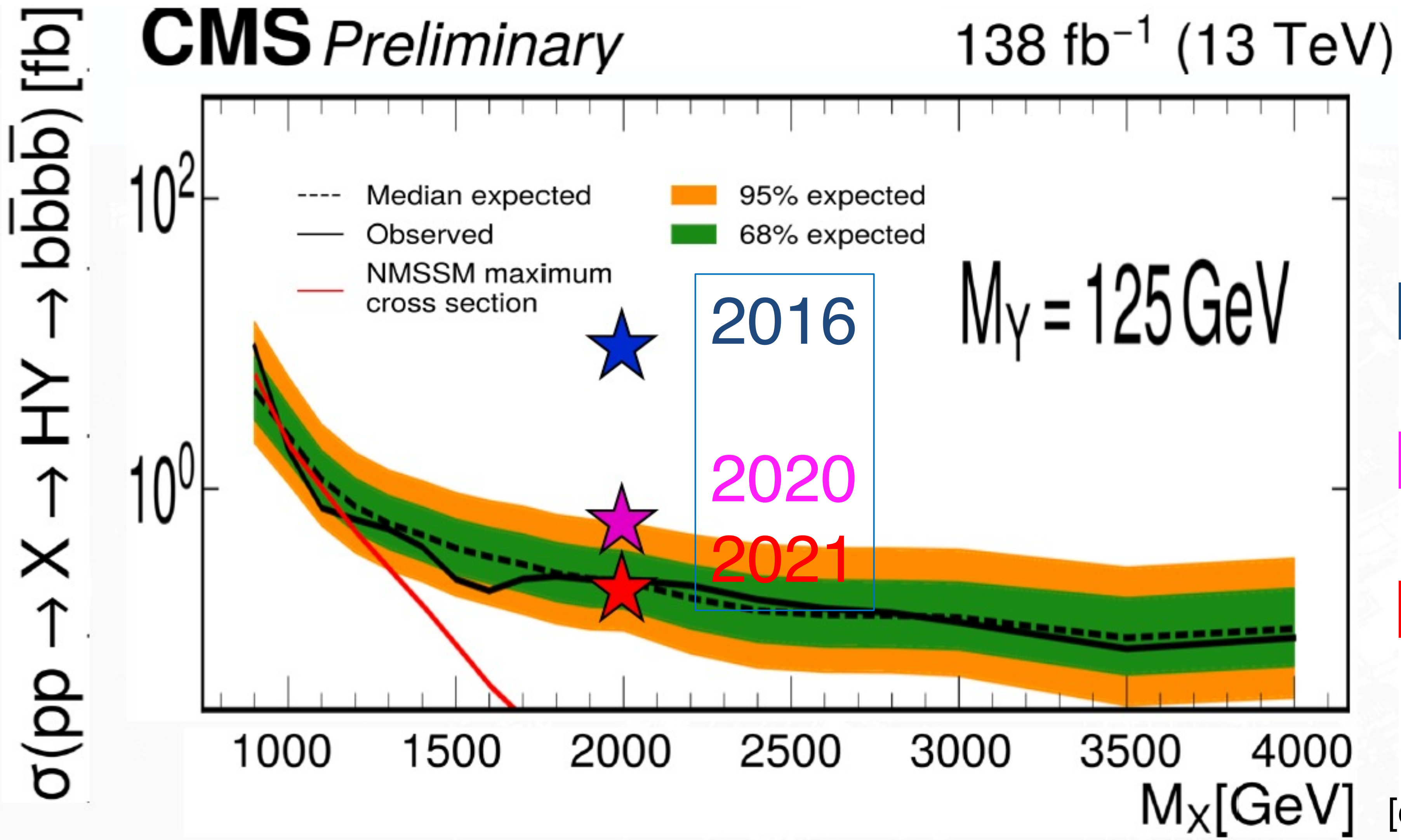
GNN / Transformer
[2203.01189]



⇒

…

Impact on physics



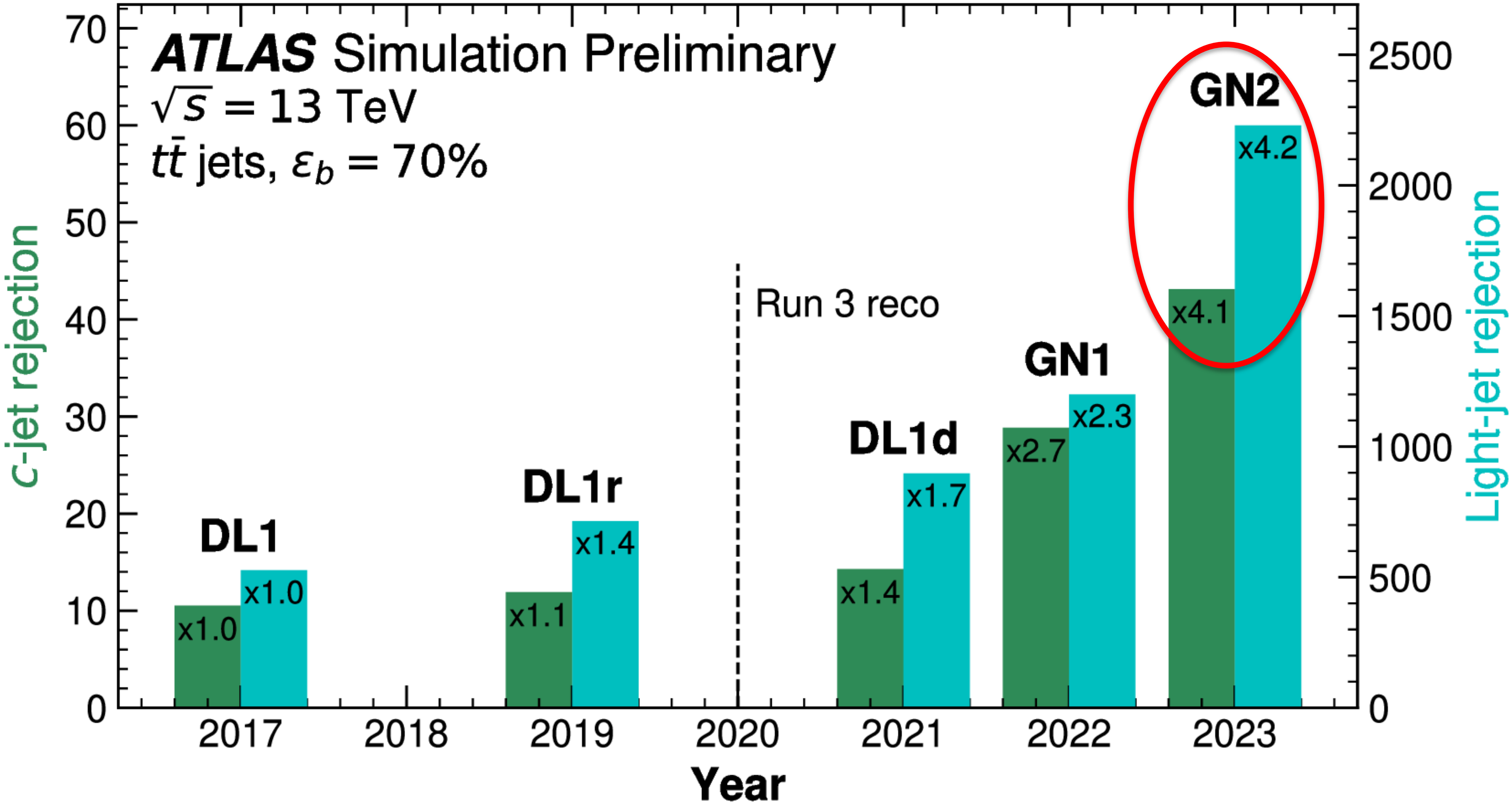
[1710.04960]

[CMS-B2G-20-004]

[2204.12413]

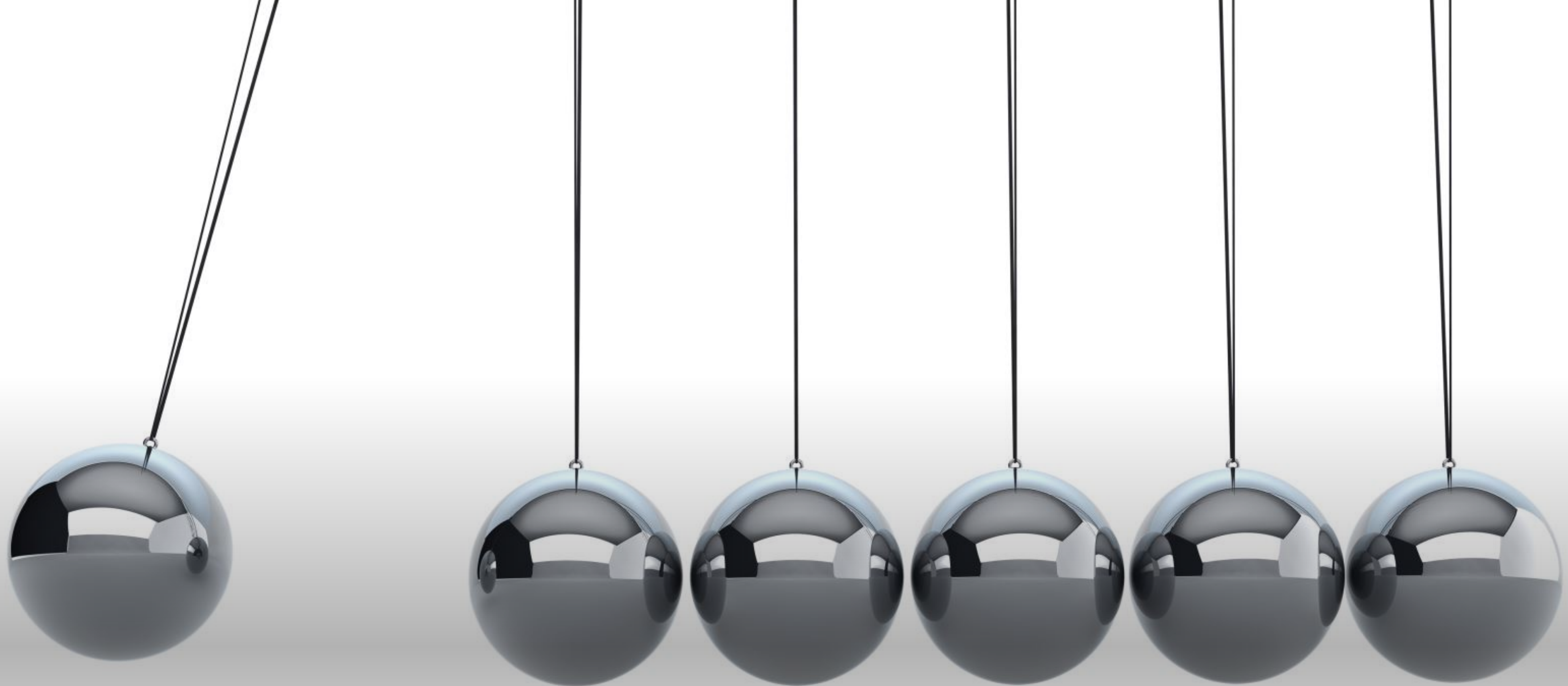
[credit: Loukas Gouskos]

When are we reaching a plateau?



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

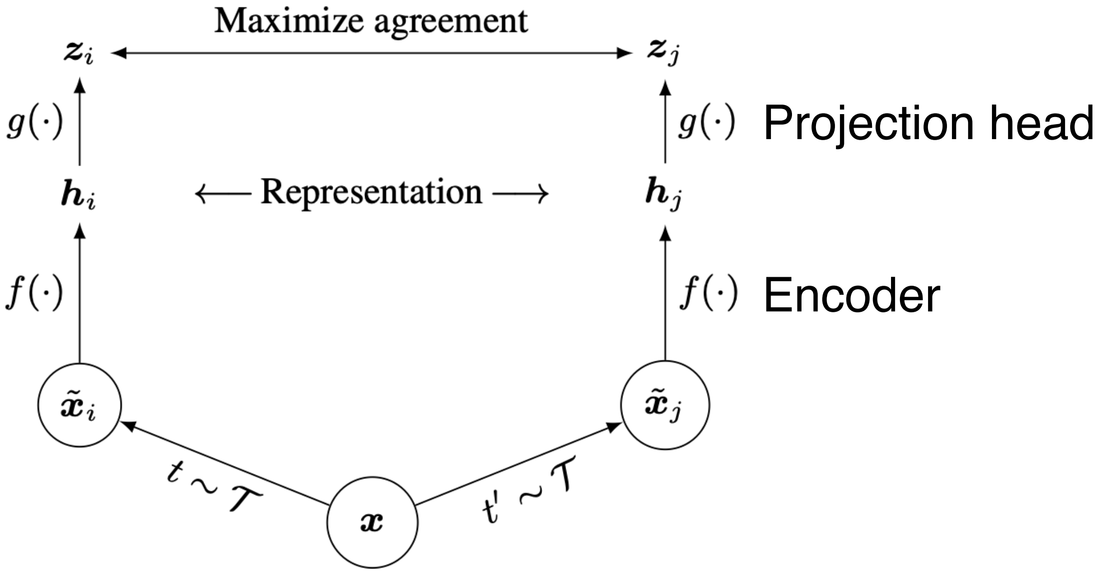
Physics-aware AI
[the edge of science]



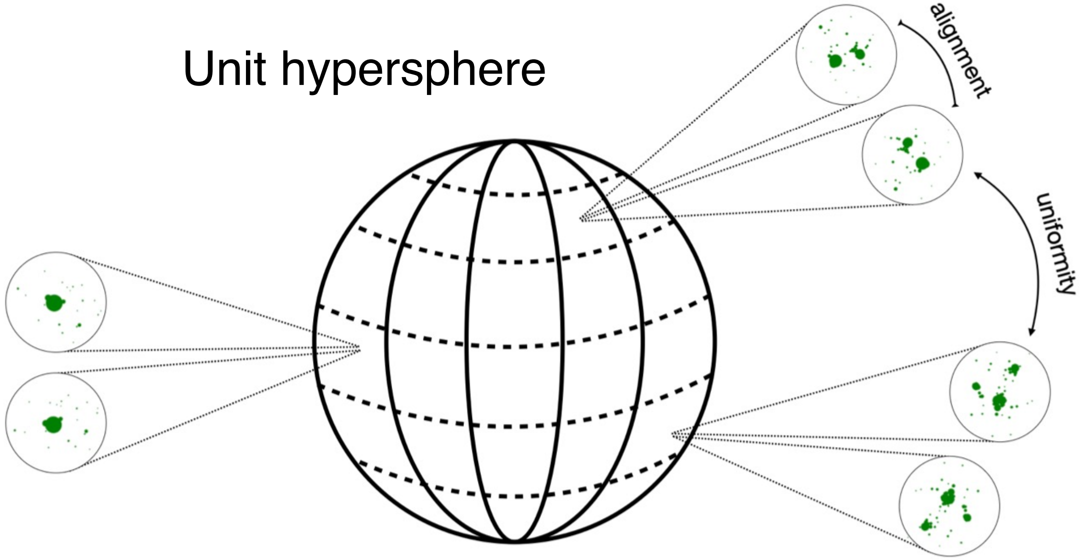
The difference between language models & PP?

We have a model

Invariance to transformation: contrastive learning



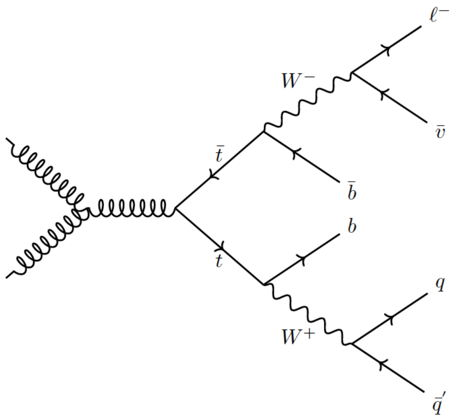
[JetCLR [2108.04253] (based on SimCLR Hinton et al.)]



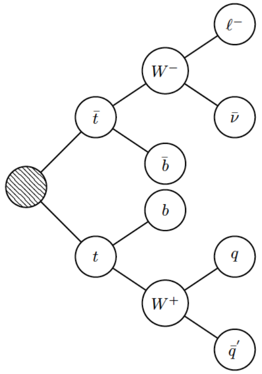
$$s(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|} = \cos\theta_{ij}$$

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

Encode physics into a GNN

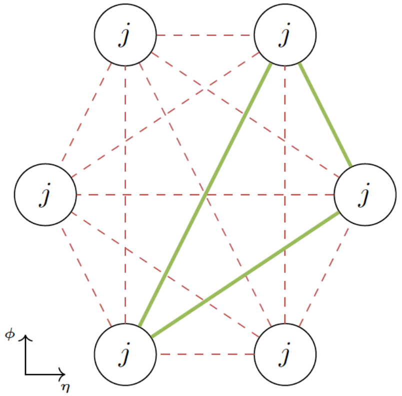


(a) Feynman diagram

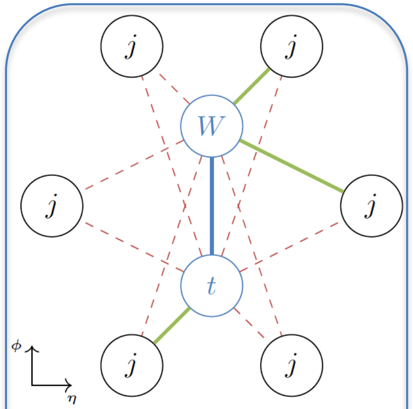


(b) Node and edge graph

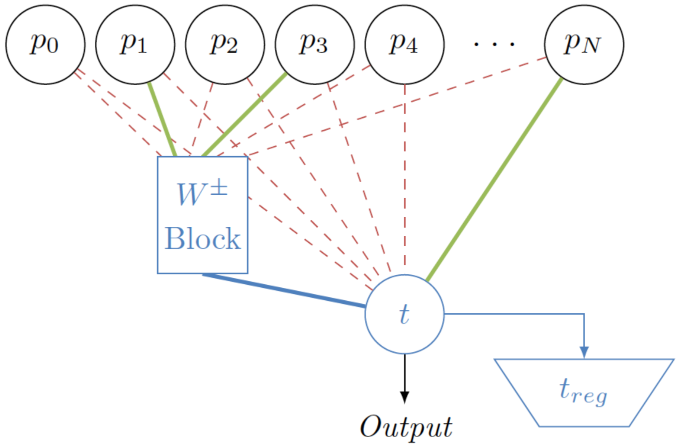
Encode information by leaving out edges

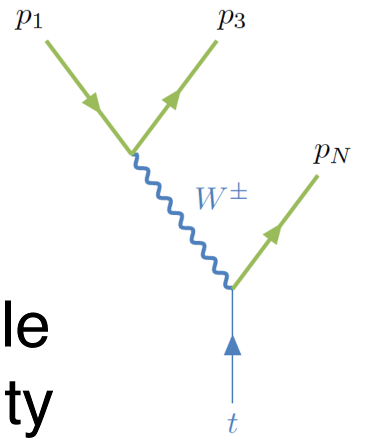


(a) Fully connected graph



(b) Topograph



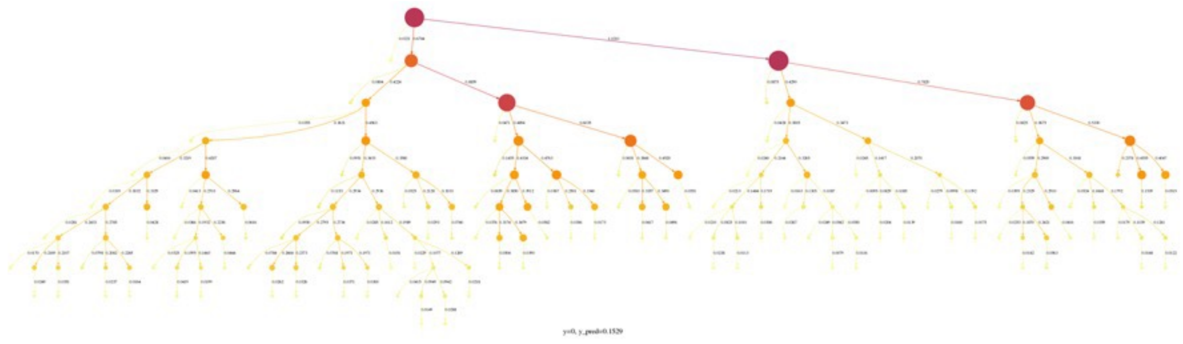


Modular
Generalizable
Interpretability
Combinatorics solving
Downstream tasks

[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]

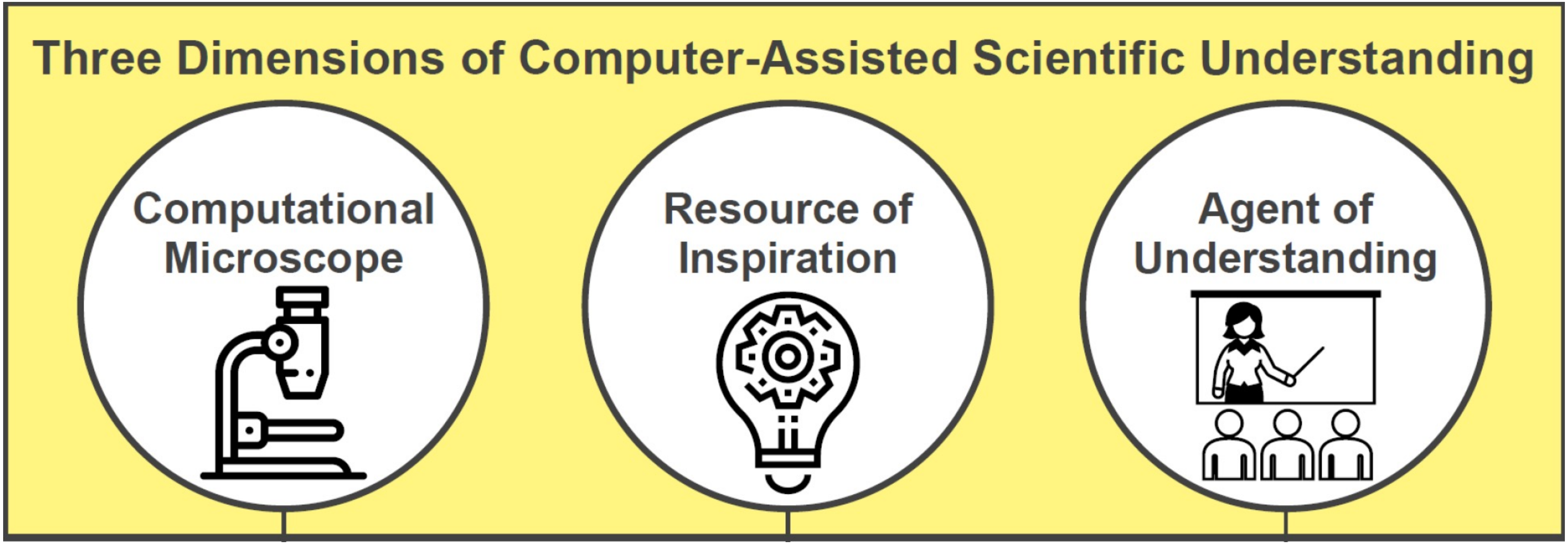
All models are wrong, but some are are useful.

– GEORGE BOX

Useful in what sense?

What is scientific understanding?

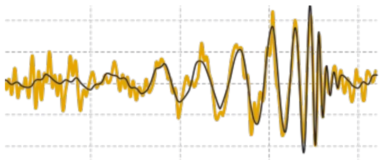
[We want more than an AI oracle]



[2204.01467]

ML interpretability for science

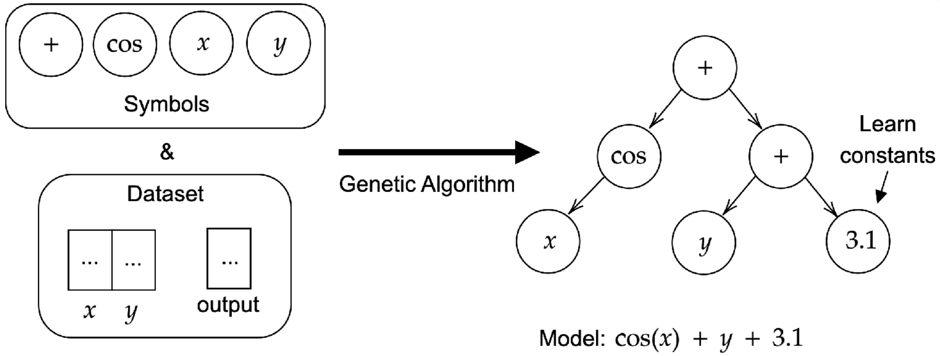
Science



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

Computer vision

???

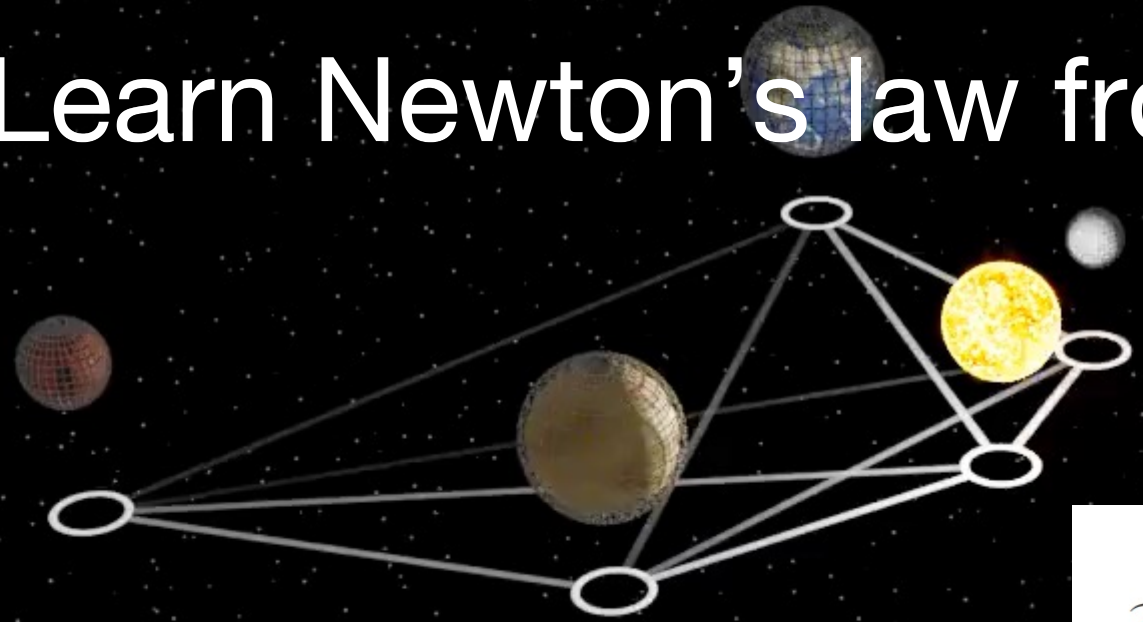


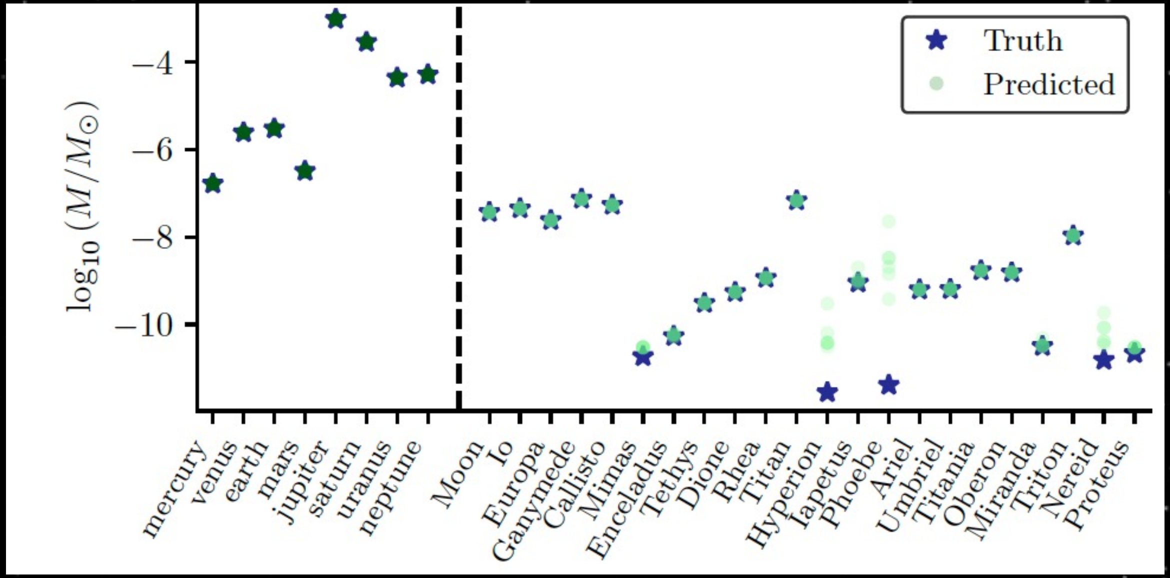
Analytic models often generalize better than NN
Symbolic regression as inductive bias

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

[Miles Cranmer – Hammers & Nails 2022]

Learn Newton's law from solar system



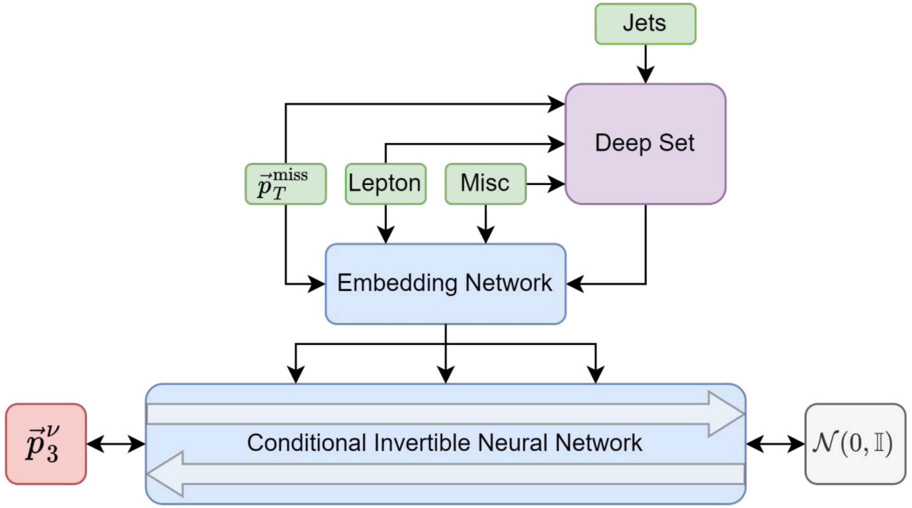


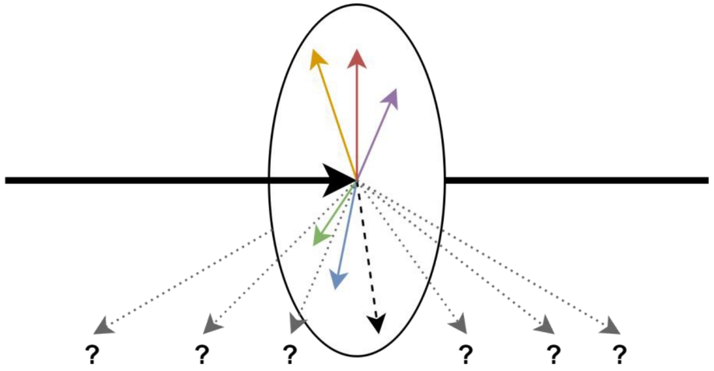
GNN → PySR → Learn masses + dynamics

2202.02306

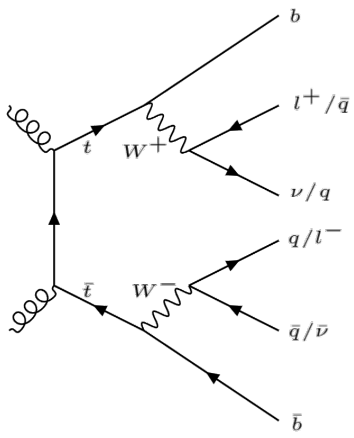
Surrogate modeling

ν-Flows: Conditional Neutrino Regression

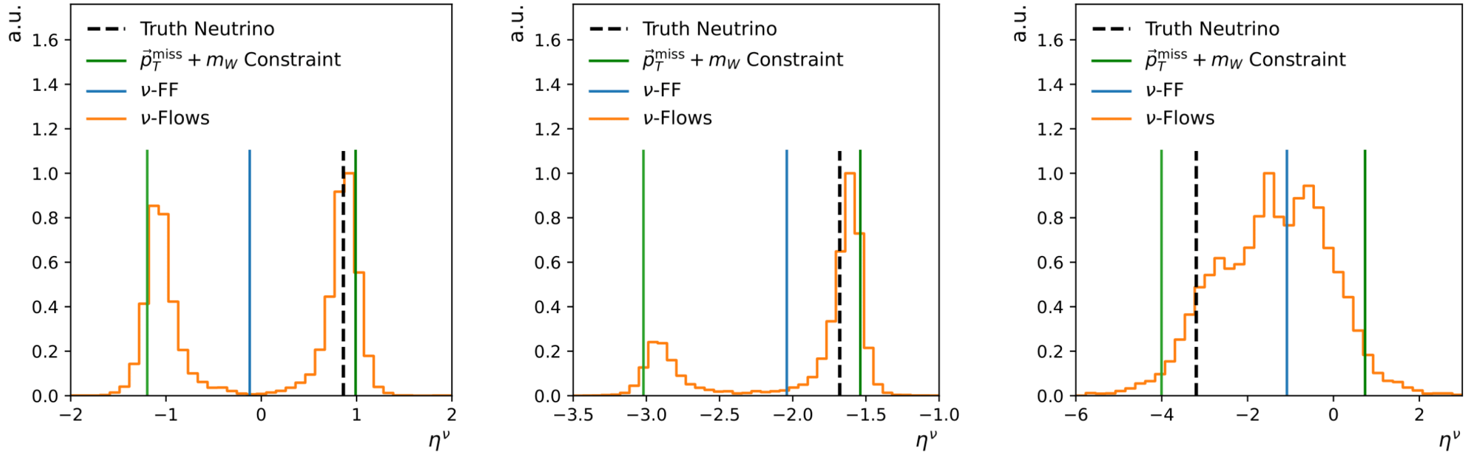




Learn ν **3**-momentum



Cherry picked representative examples:



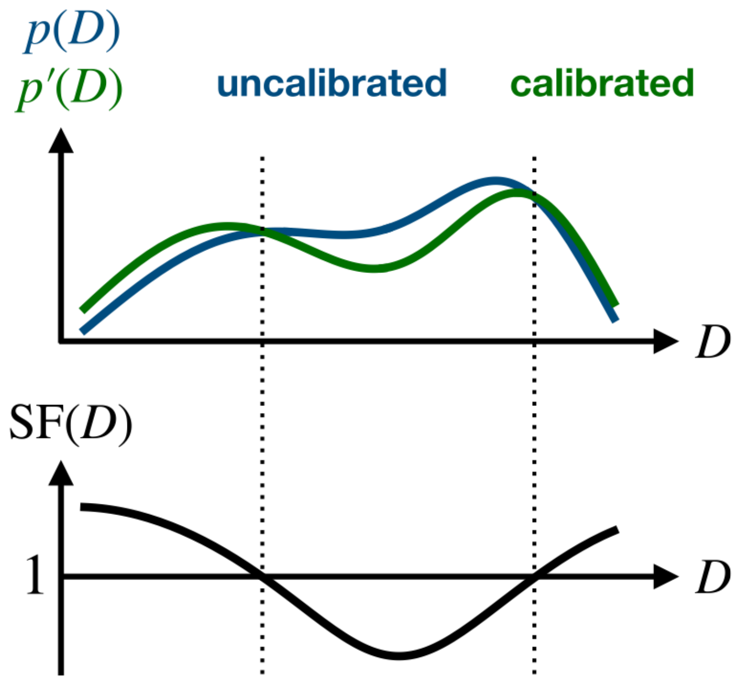
Conditional normalizing flow: learn **conditional likelihood** over neutrino momenta assuming an underlying process (inductive bias)

Improve over traditional method

[2207.00664]

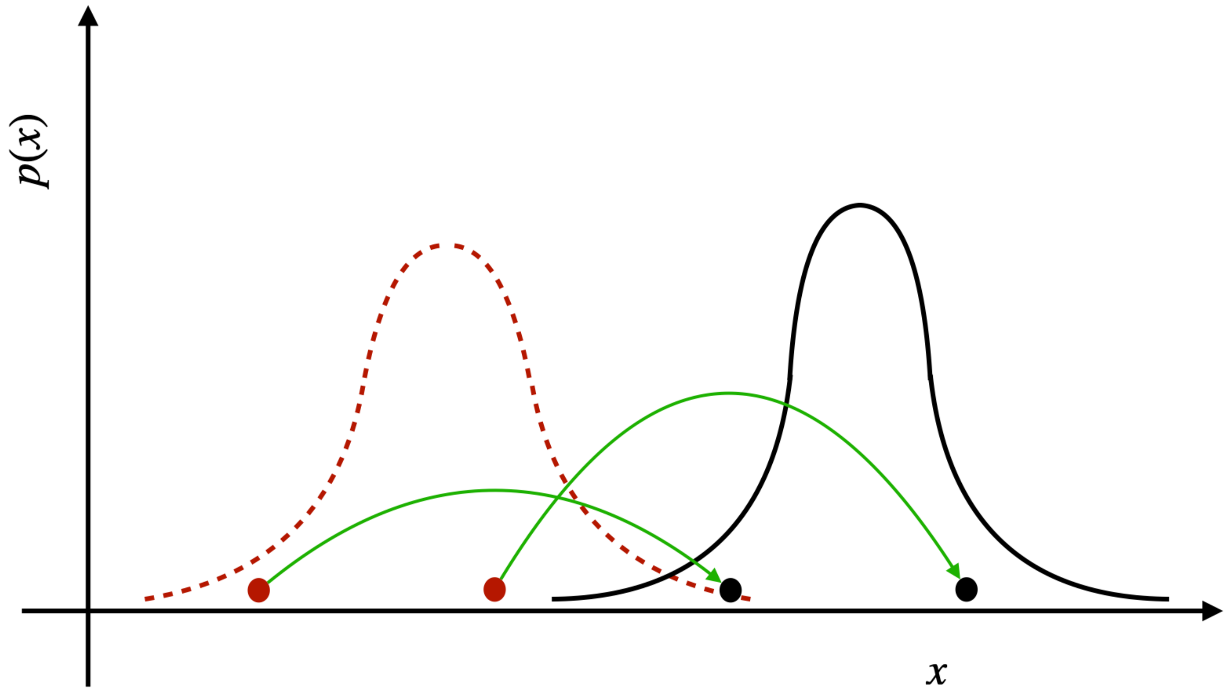
Domain shift: calibrate synthetic to real data

1. Reweighting
- Tricks to battle curse of dimensionality
- Non-overlapping support



(b)

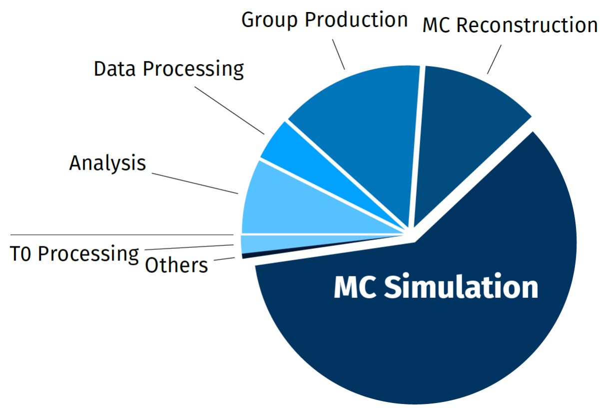
2. “Transport your problems away”



[2107.08648]

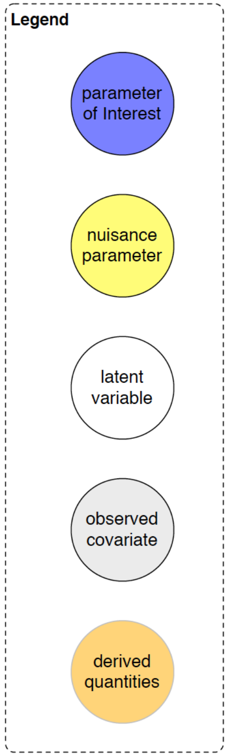
Full Sim & Reco

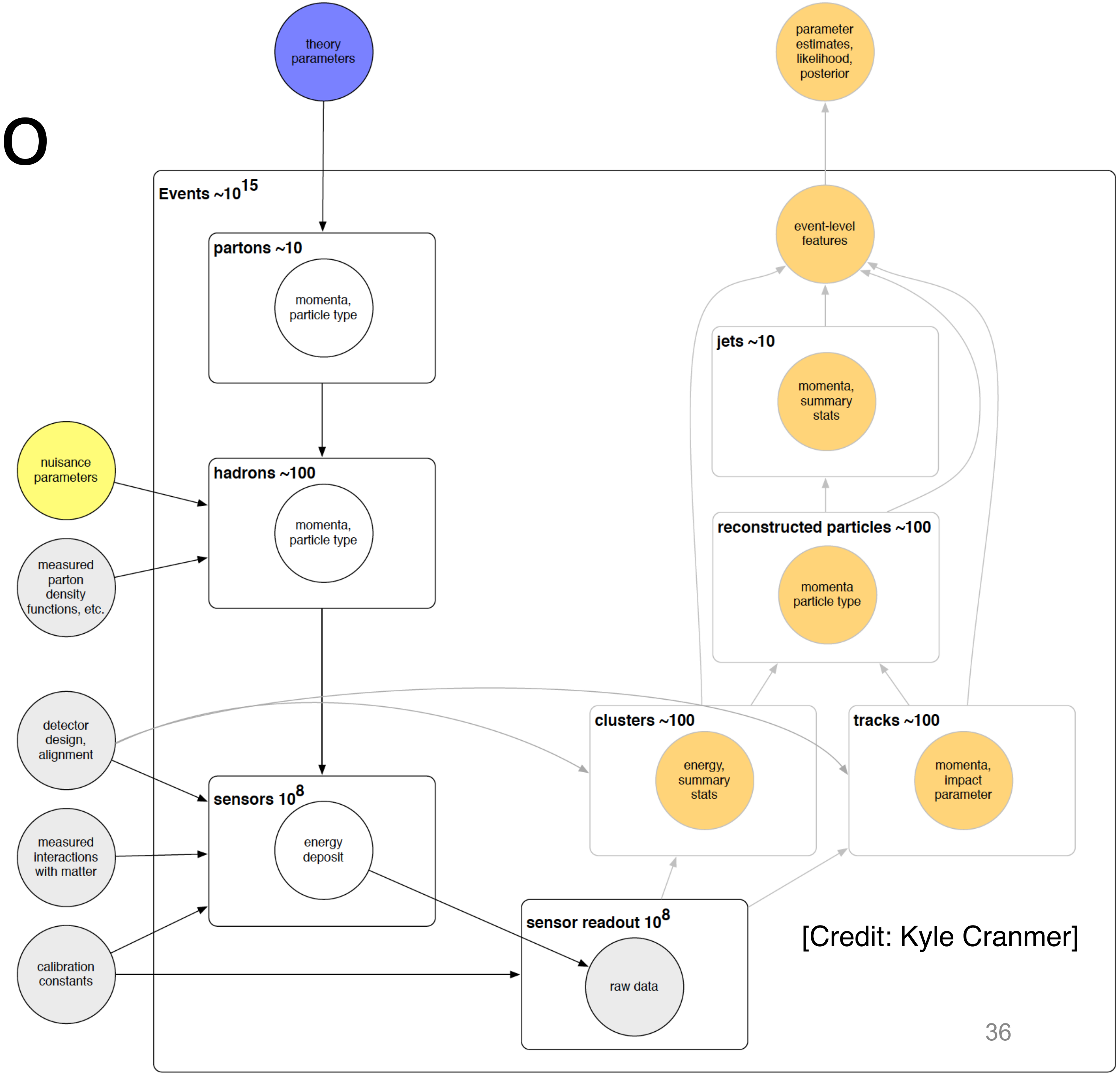
Bottleneck: computing budget



Up to 10 min / event

[LHCC-2022-005]



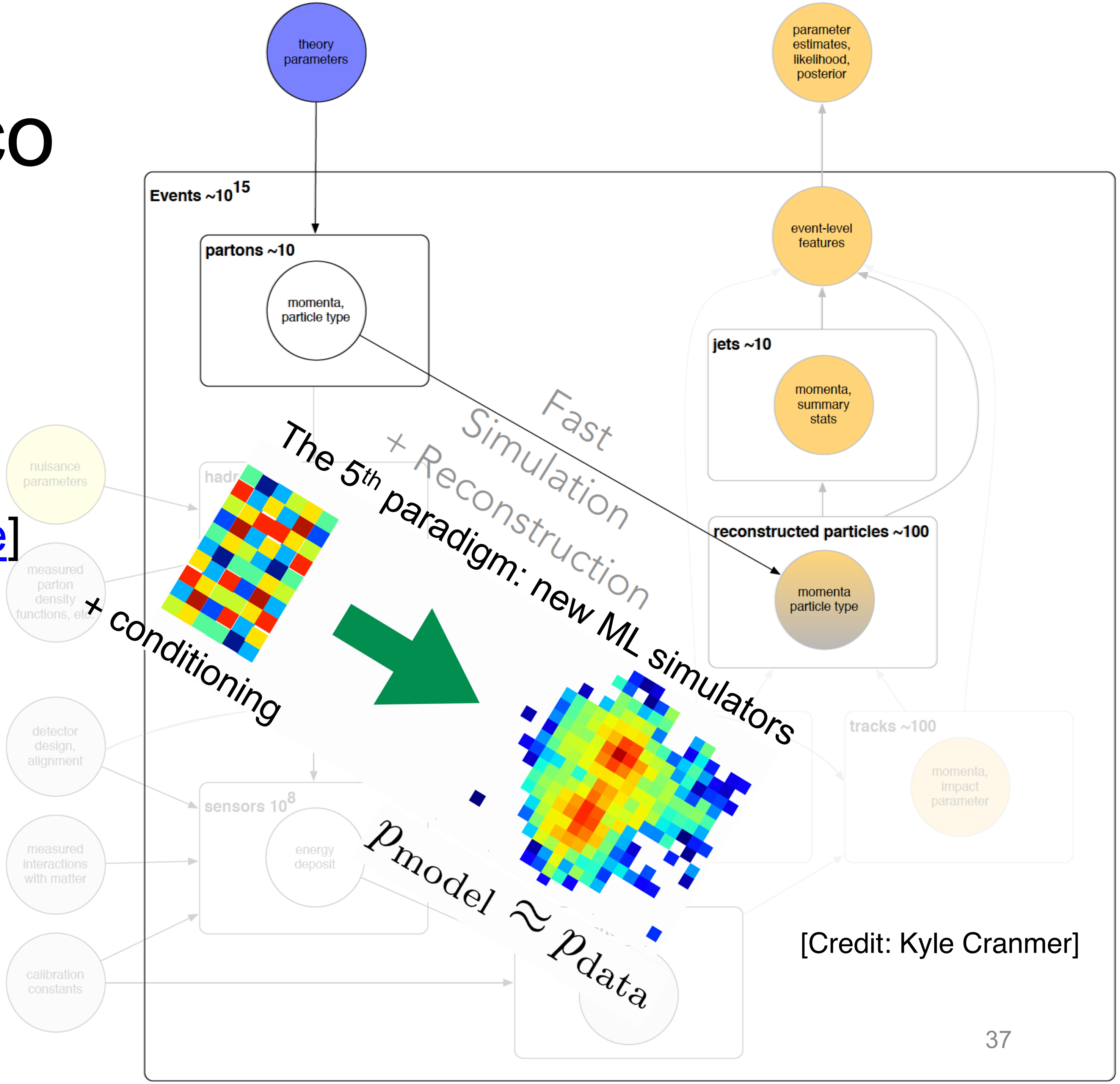


[Credit: Kyle Cranmer]

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

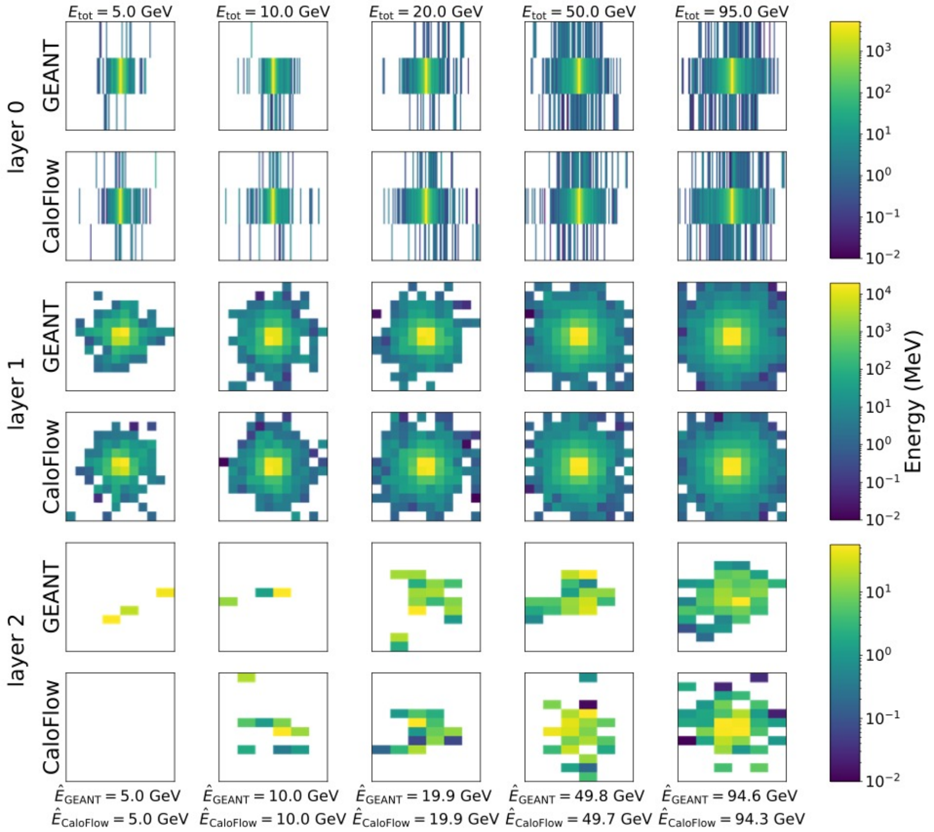
[Differentiable & fast]

Faces

[Karras et al., 2018]

VAEs, GANs, Flows, Diffusion,…

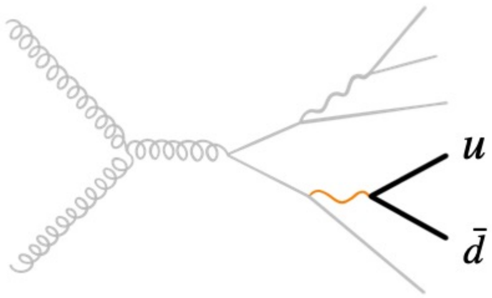
Images of calo showers



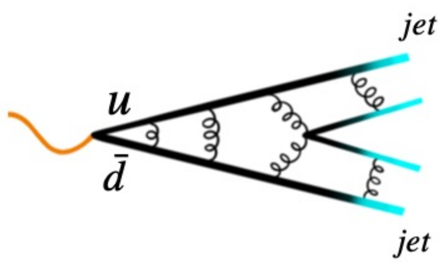
[CaloFlow]

Generation from noise

Parton Interactions
$\mathcal{O}(10)$



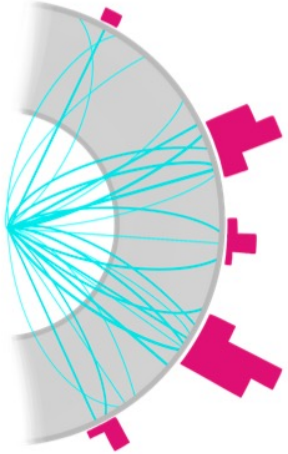
Showering
$\mathcal{O}(100)$



NOISE

[1907.03764,…]

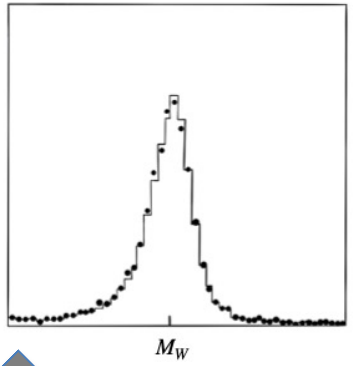
Detection
$\mathcal{O}(10^6)$



NOISE

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

Reconstruction
$\mathcal{O}(10)$

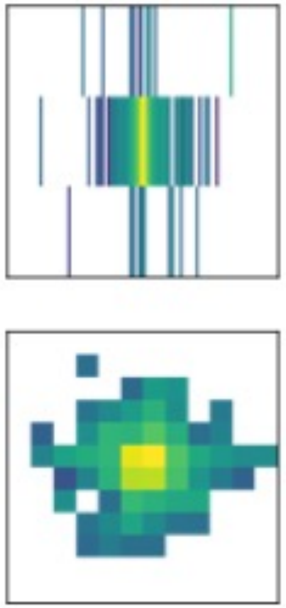


NOISE

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

→ Conditioning

Images → Point cloud



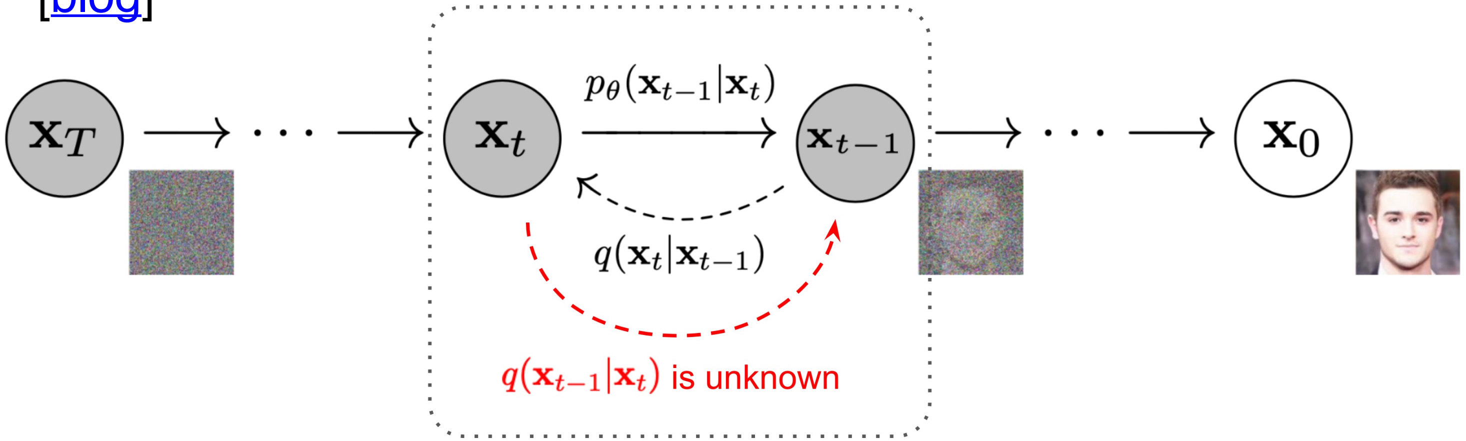
Decouple modeling
from detector geometry

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

New on the market: point cloud diffusion

[PC-JeDi]

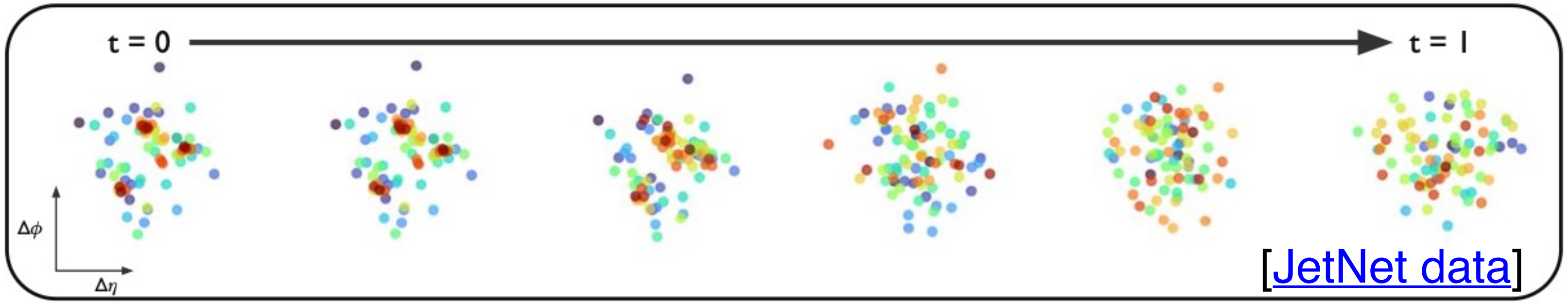
[blog]



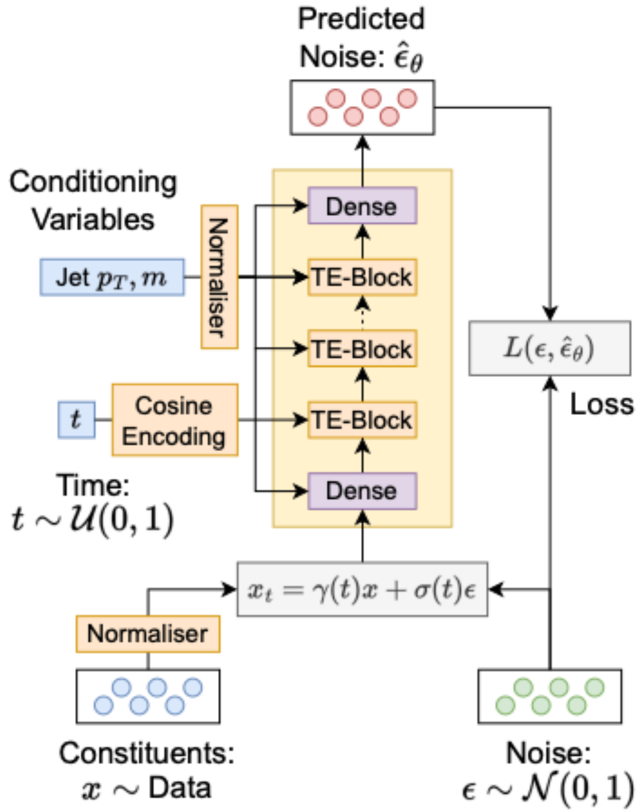
Gradually add Gaussian noise (right-to-left=forward)

Reverse "learn the noise"

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



[JetNet data]

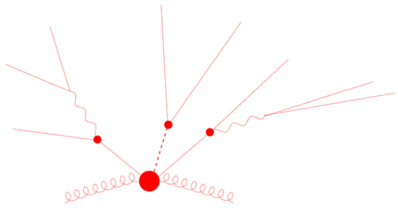


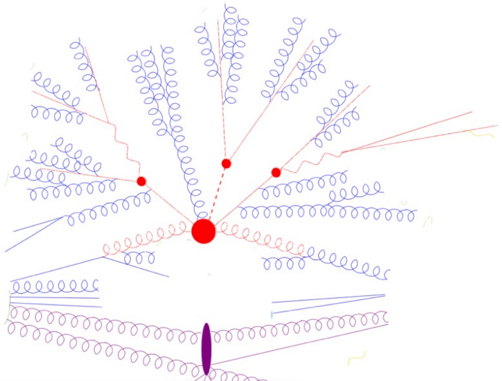
Transformer Encoder (TE) Block

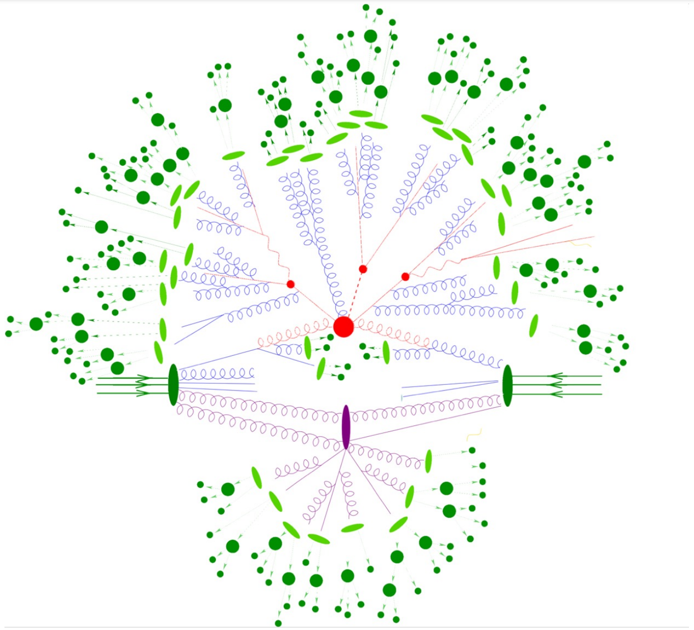
[See also 2206.11898]

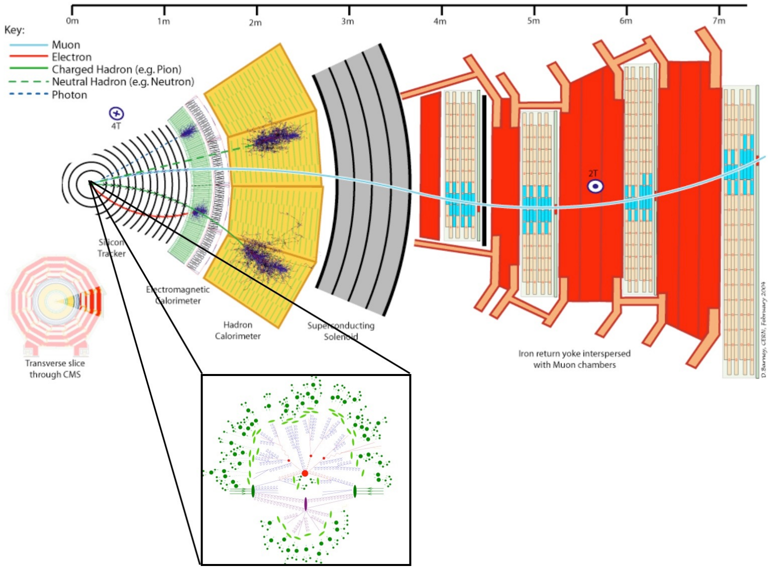
Invertible surrogates to solve inverse problem

Folding









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

Hard scatter

Radiation

Hadronization

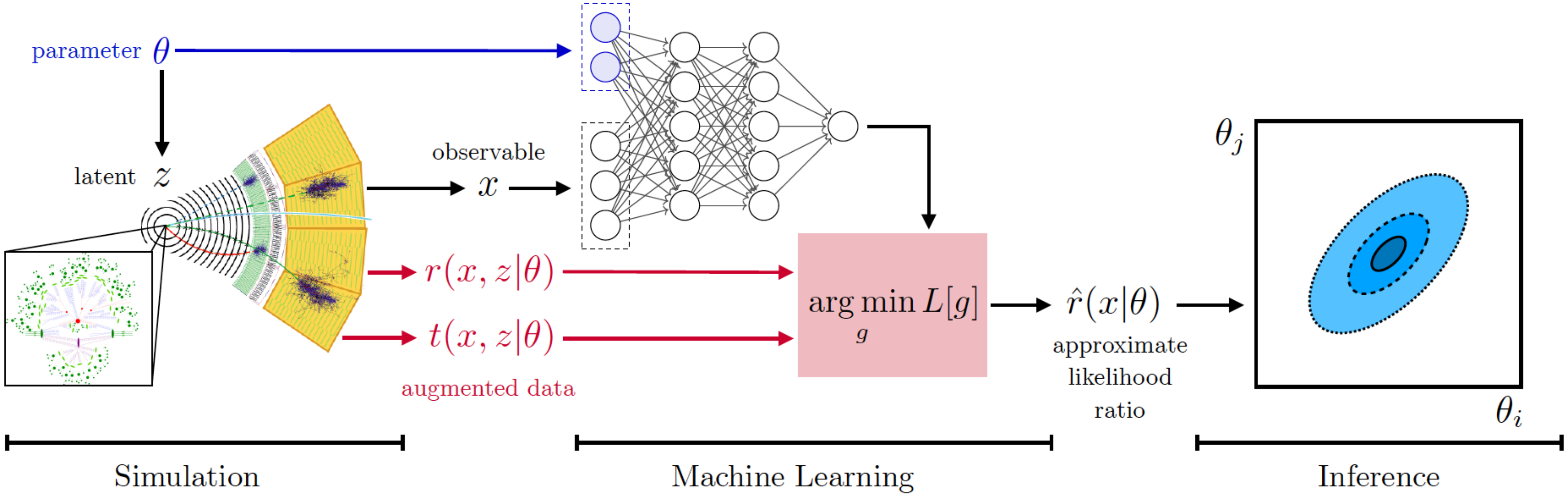
Detector

Unfolding

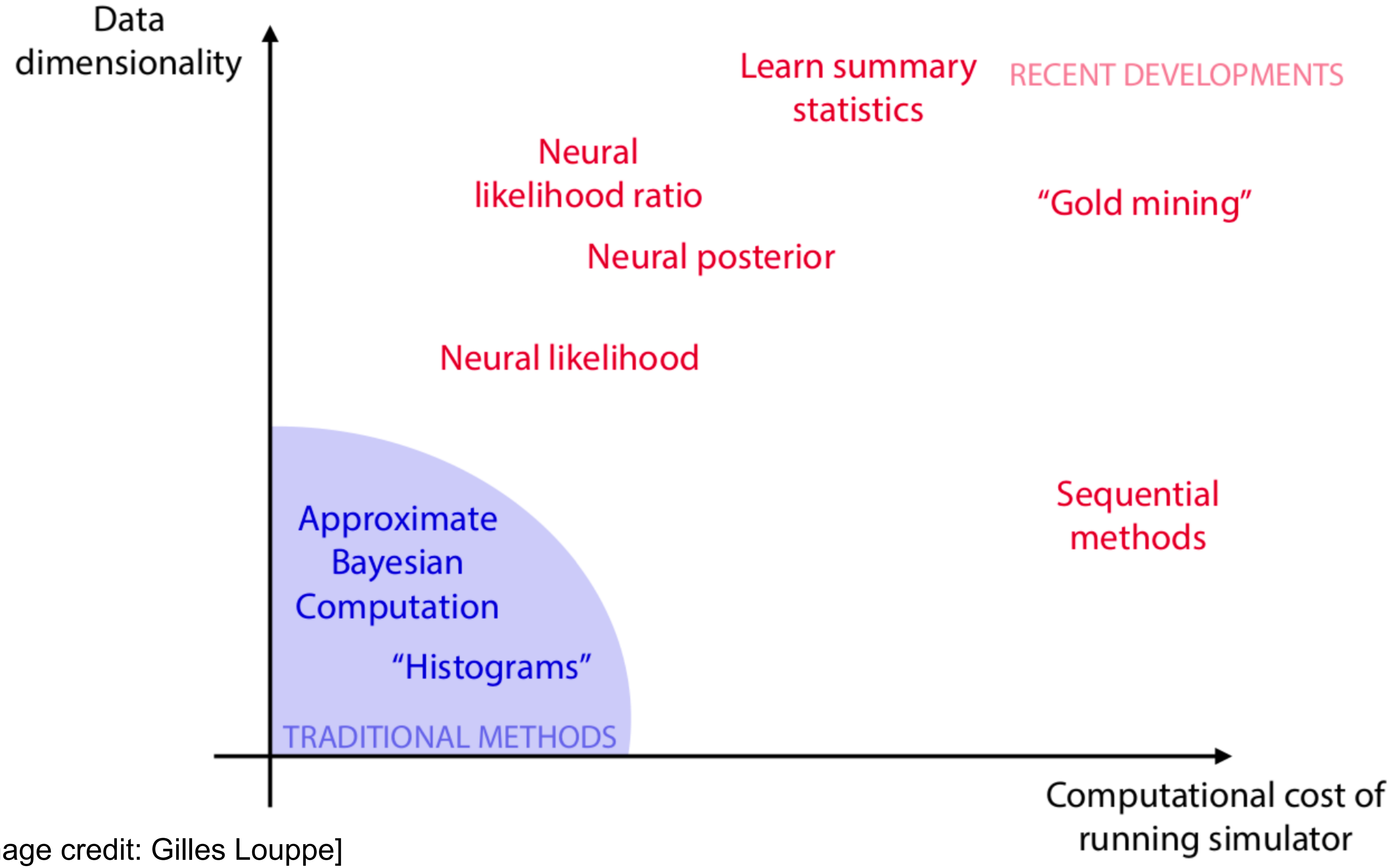
Simulation-based inference: learn p(θ|x)

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]



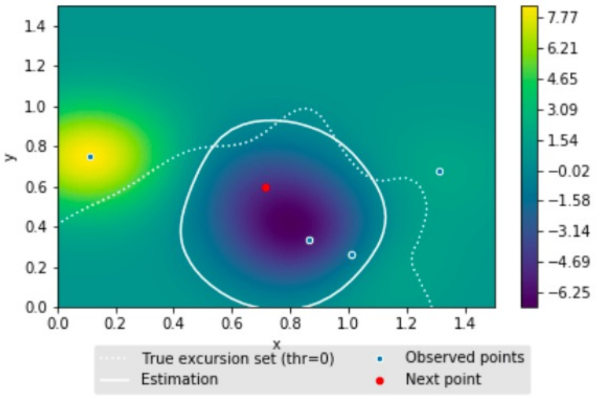
[Image credit: Gilles Louppe]

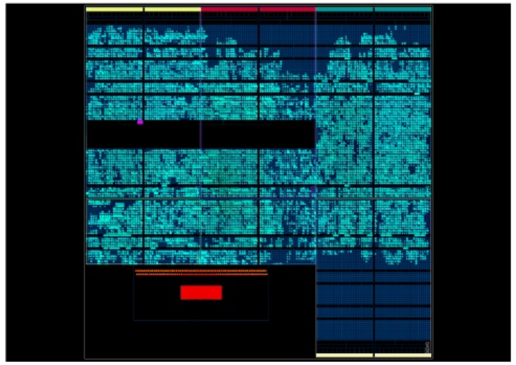
$$p(\text{theory} \mid \text{data}) = \frac{p(\text{data} \mid \text{theory})\, p(\text{theory})}{p(\text{data})}$$

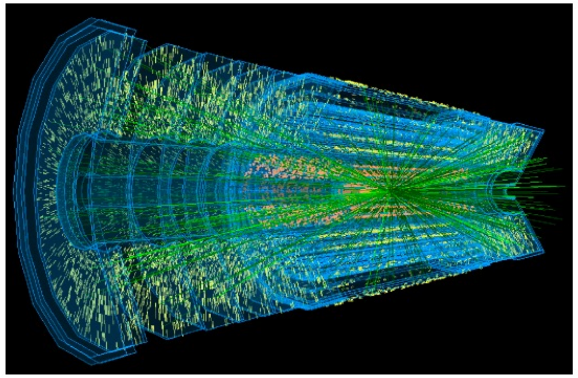
[Lukas Heinrich - Detector design using differential programing]

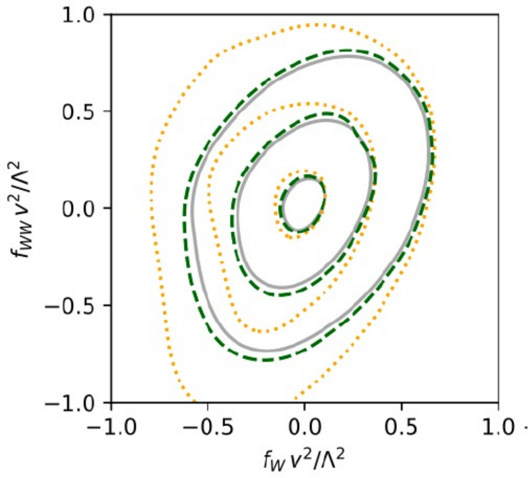
Ultimate goal:
Learning about Nature

Optimizing the science output









Optimal Theory Exploration

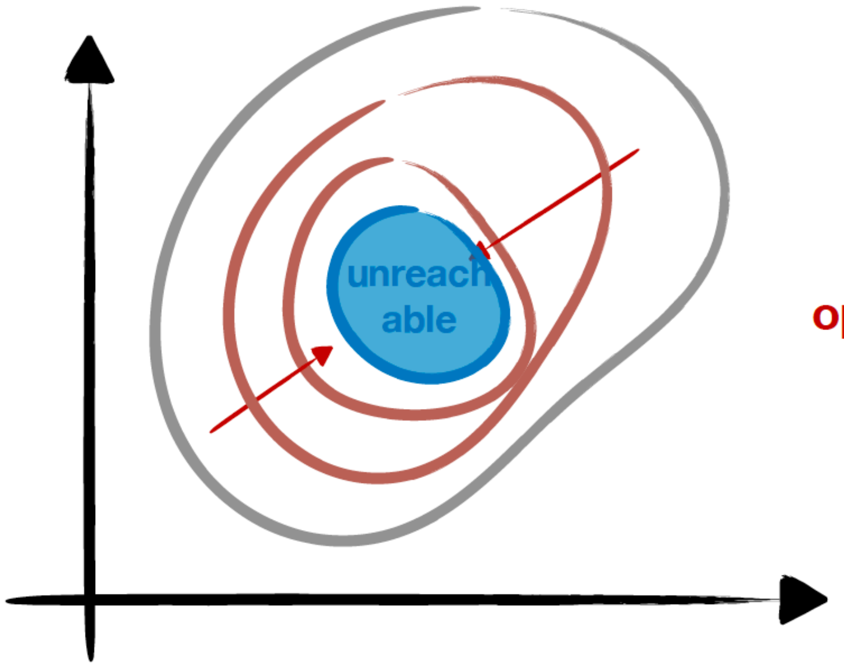
Optimal Data Taking / Experiment Operations

Optimal Reconstruction

Optimal Analysis

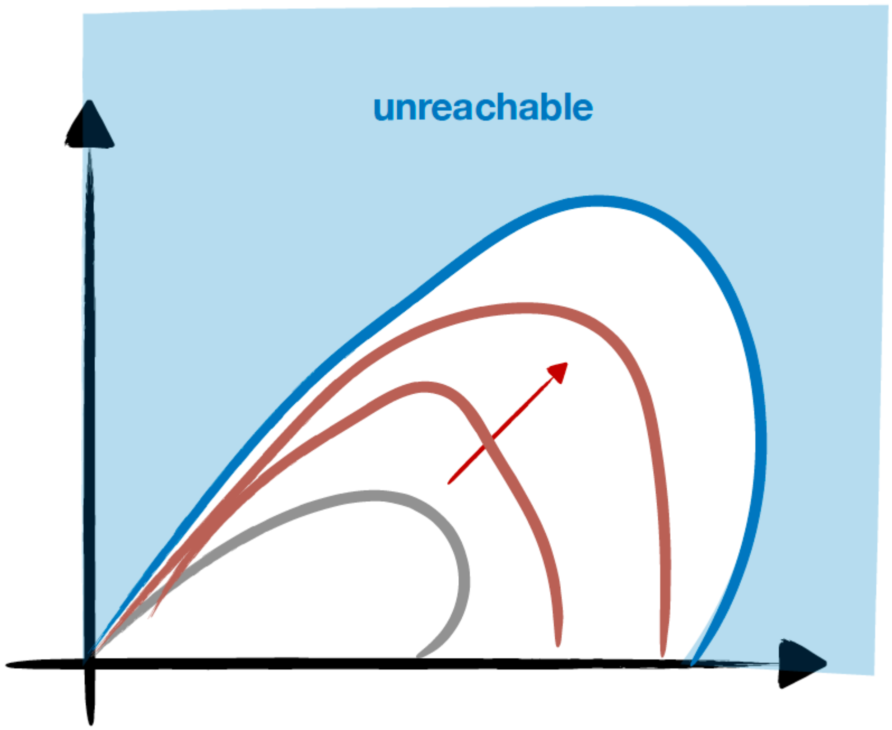
[Lukas Heinrich - Detector design using differential programing]

Natural limit: true posterior *p*(theory | data)



Measurements
(e.g. Higgs Couplings)

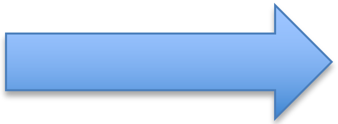
unoptimized
optimized (e.g. w/ ML)

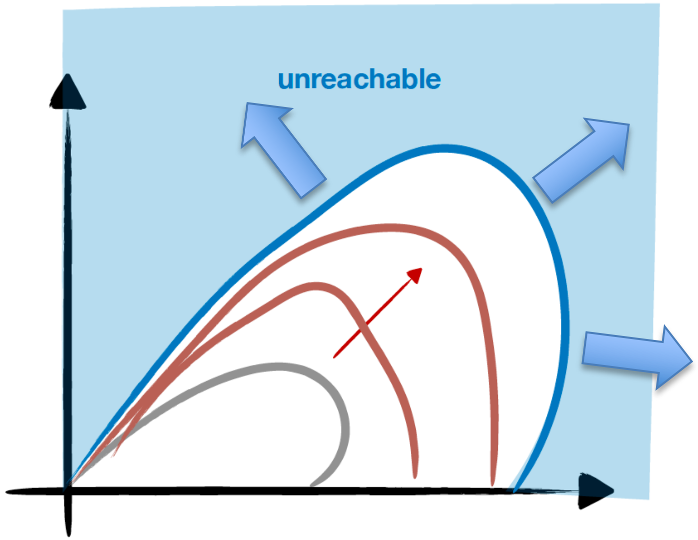


Searches
(e.g. Supersymmetry)

[Lukas Heinrich - Detector design using differential programing]

Opportunity: new ***optimal*** detector





Goal: optimize
p(theory | data)

[Lukas Heinrich - Detector design using differential programing]

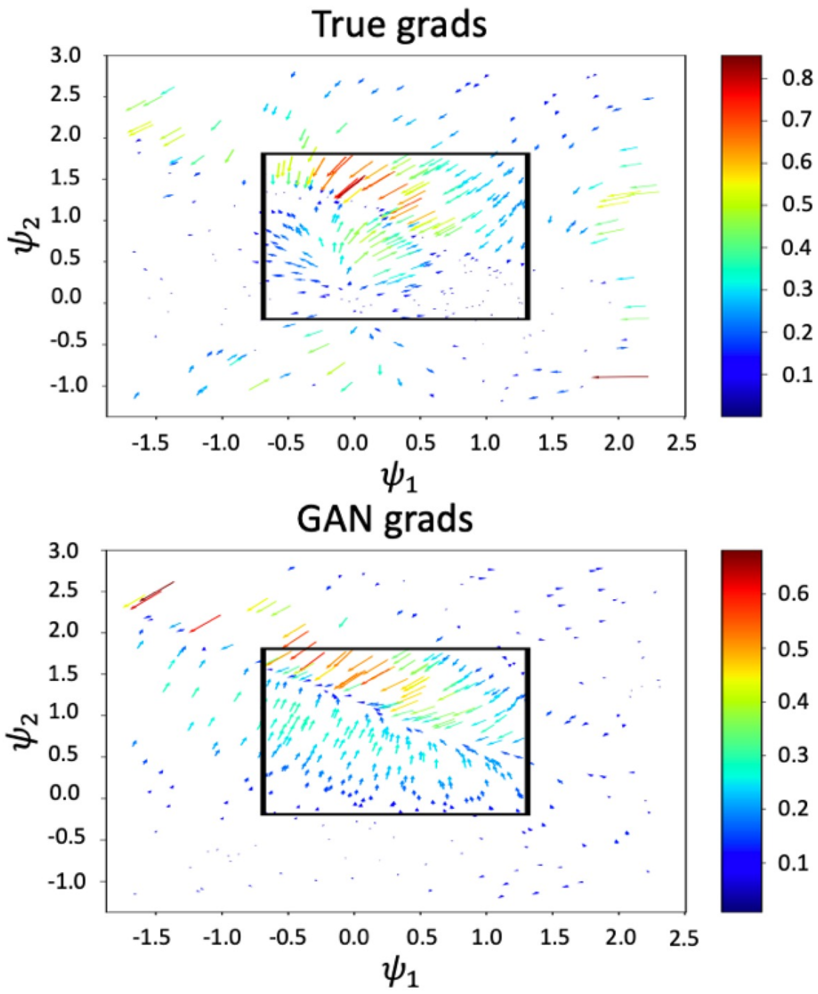
Need **design-conditional** model $p(x \mid \theta, \mathbf{D})$

Approximate $p(x \mid \theta, D)$ using **generative model**
→ **Fast**
→ **Differentiable**

Challenge:
$p(x \mid \mathbf{D})$ without already exploring all design space **D**

Solution:
train local models as you optimize [2002.04632]

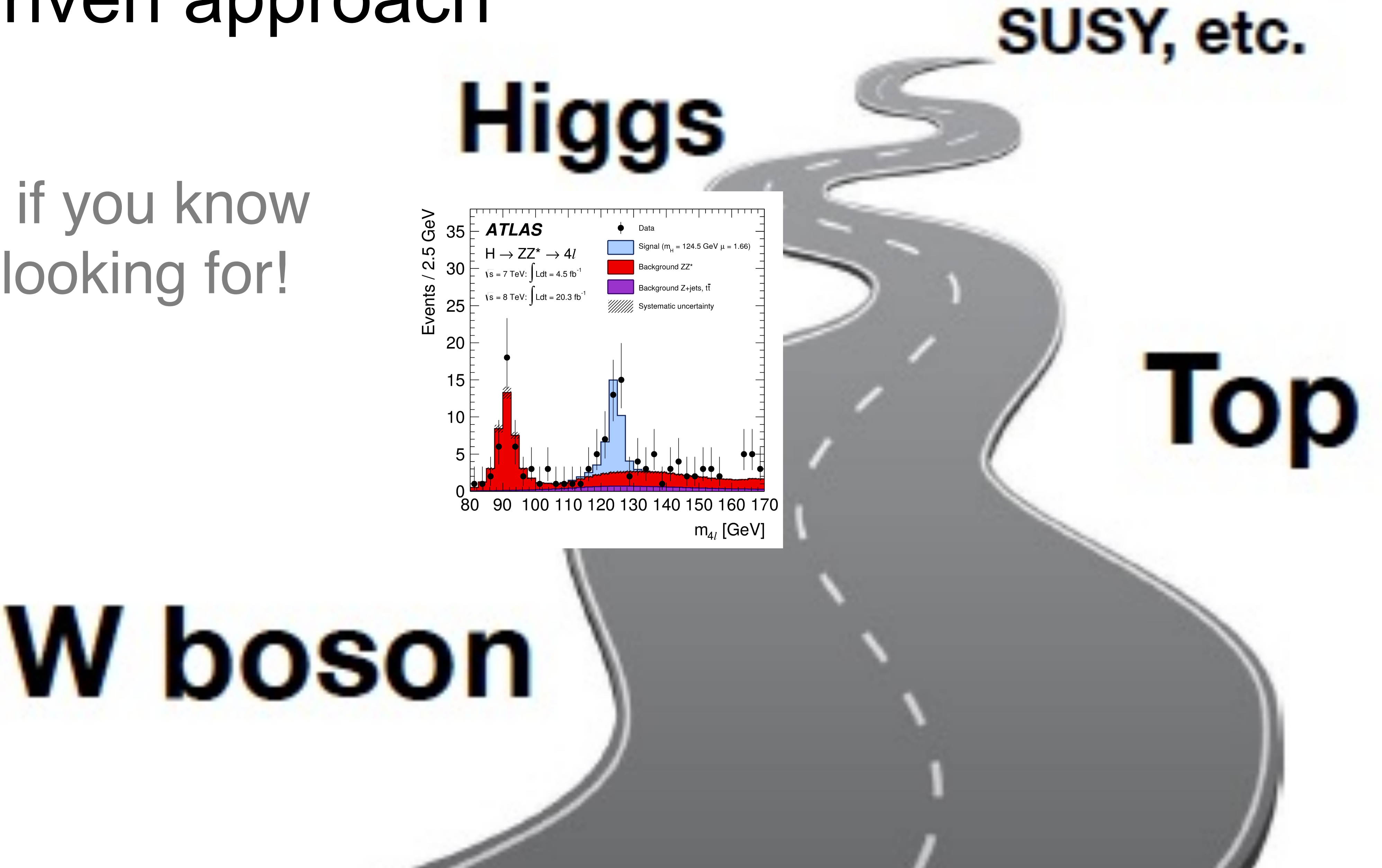
Detector design is a challenging frontier in ML@HEP
Fine-tune human design → discovery of novel designs



Search for the Unknown

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$_1$				BSM $\to$ SM$_1$ $\times$ SM$_2$				BSM $\to$ complex			
										q/g	γ/π^0's	b	$\cdots$	tZ/H	bH		$\cdots$	$\tau qq'$	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

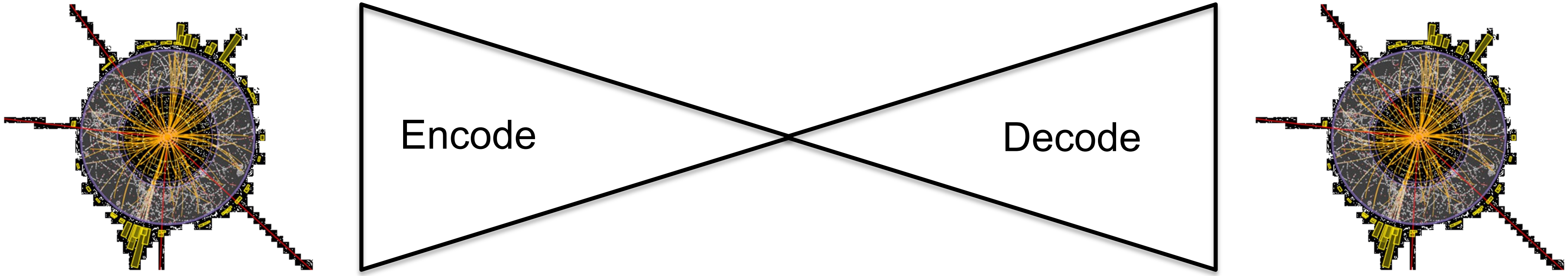
Model-agnostic search portfolio

1. Unsupervised autoencoder-style outlier detection

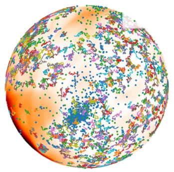
2. Semi-supervised in-situ background modeling

Fabulous idea: outlier detection with autoencoders

Train on *normal* (=SM)



Poor reconstruction = *anomaly*



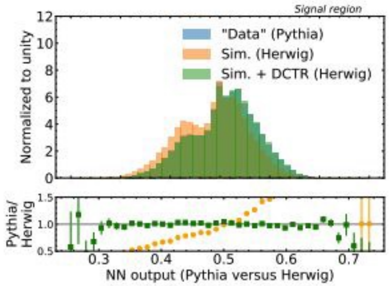
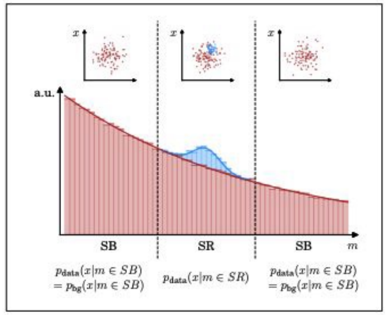
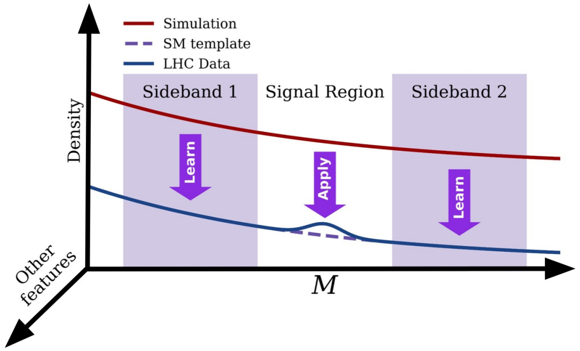
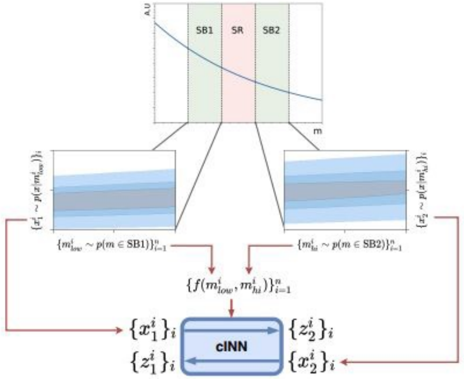
[NAE]

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*

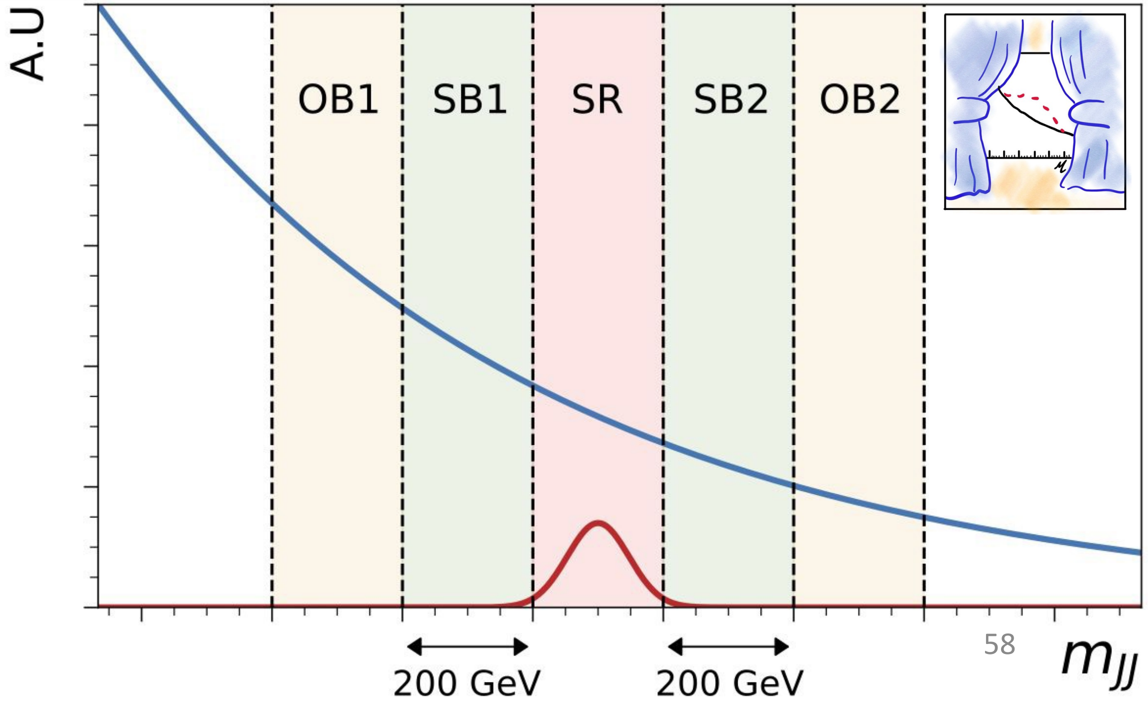
	Learn from simulation	Learn from data (SB)
Modeling the likelihood ratio	SALAD 	CATHODE*  [*see also LaCATHODE & ANODE]
Morphing the features	FETA 	CURTAINs & Flow4Flows 

[*Fidelity of simulation alone insufficient]

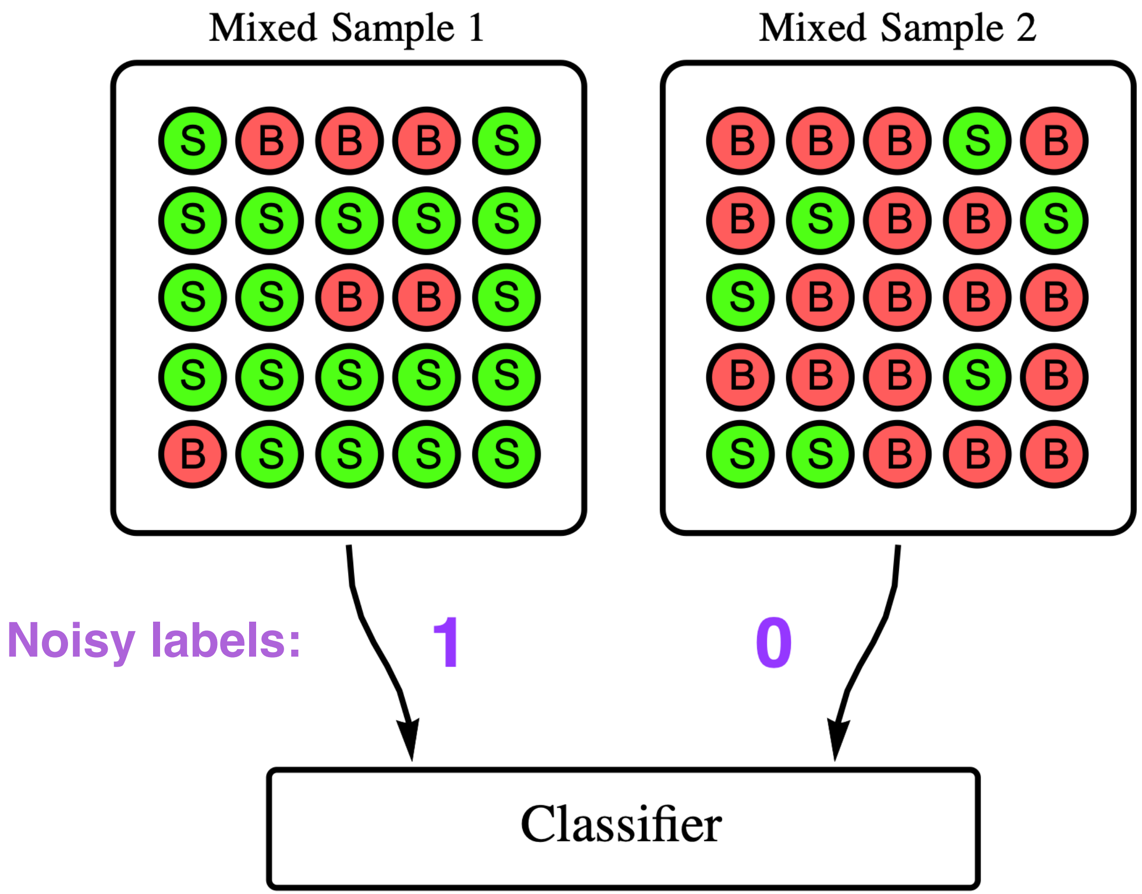
In-situ background modeling for bump hunt

[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

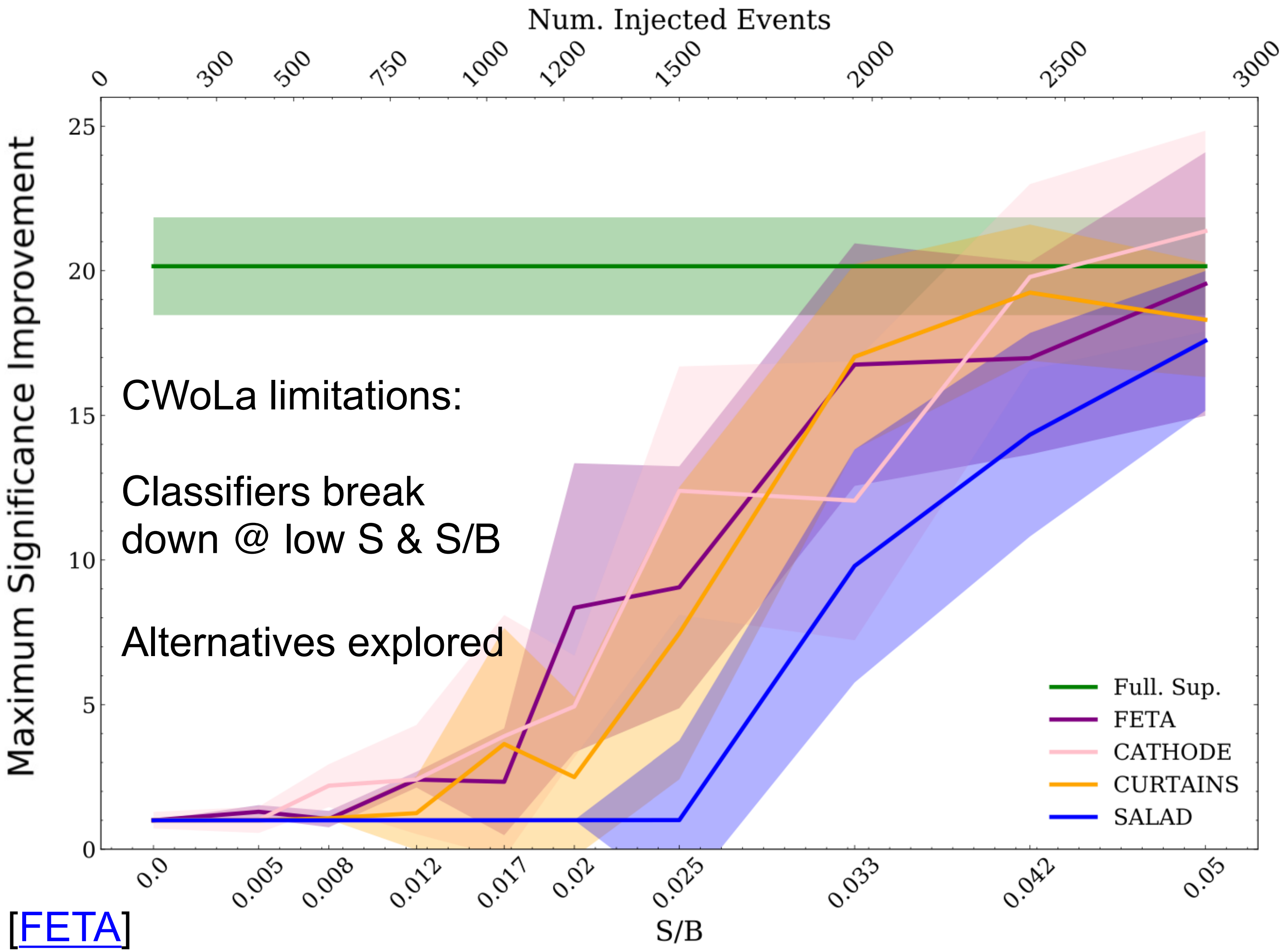
Abandon notion of *event label*

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

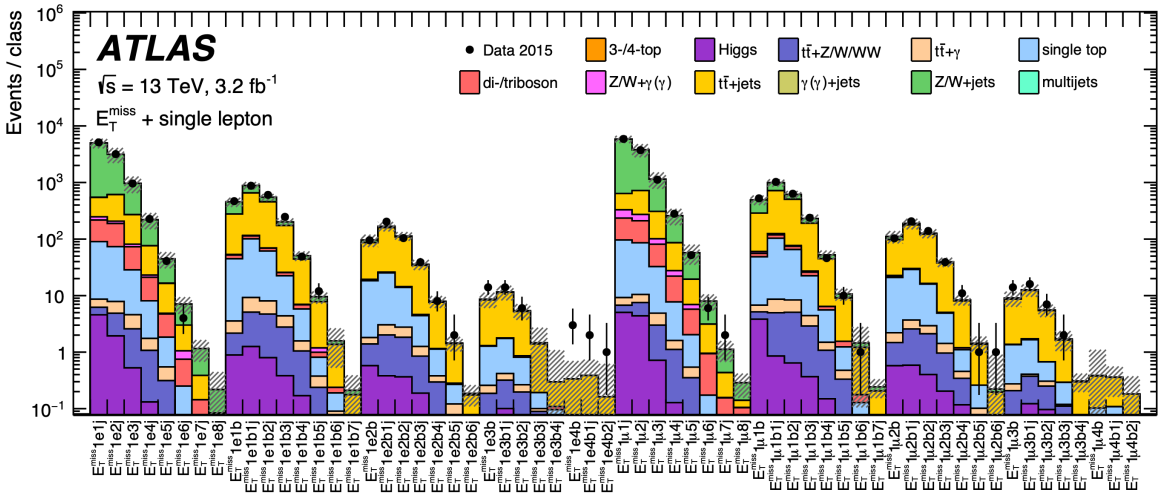


[FETA]

Similar performance of methods

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

Questions beyond in-situ modeling + CWoLa



10^5 signal region [1807.07447]

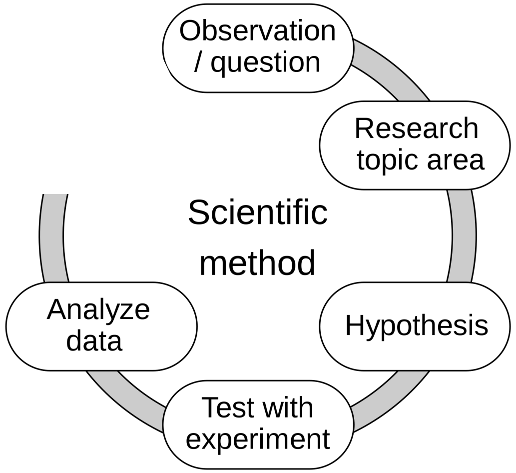
Automation

The choice of feature space

Data slicing & #tests [look elsewhere effect]

Dial up/down the physics prior

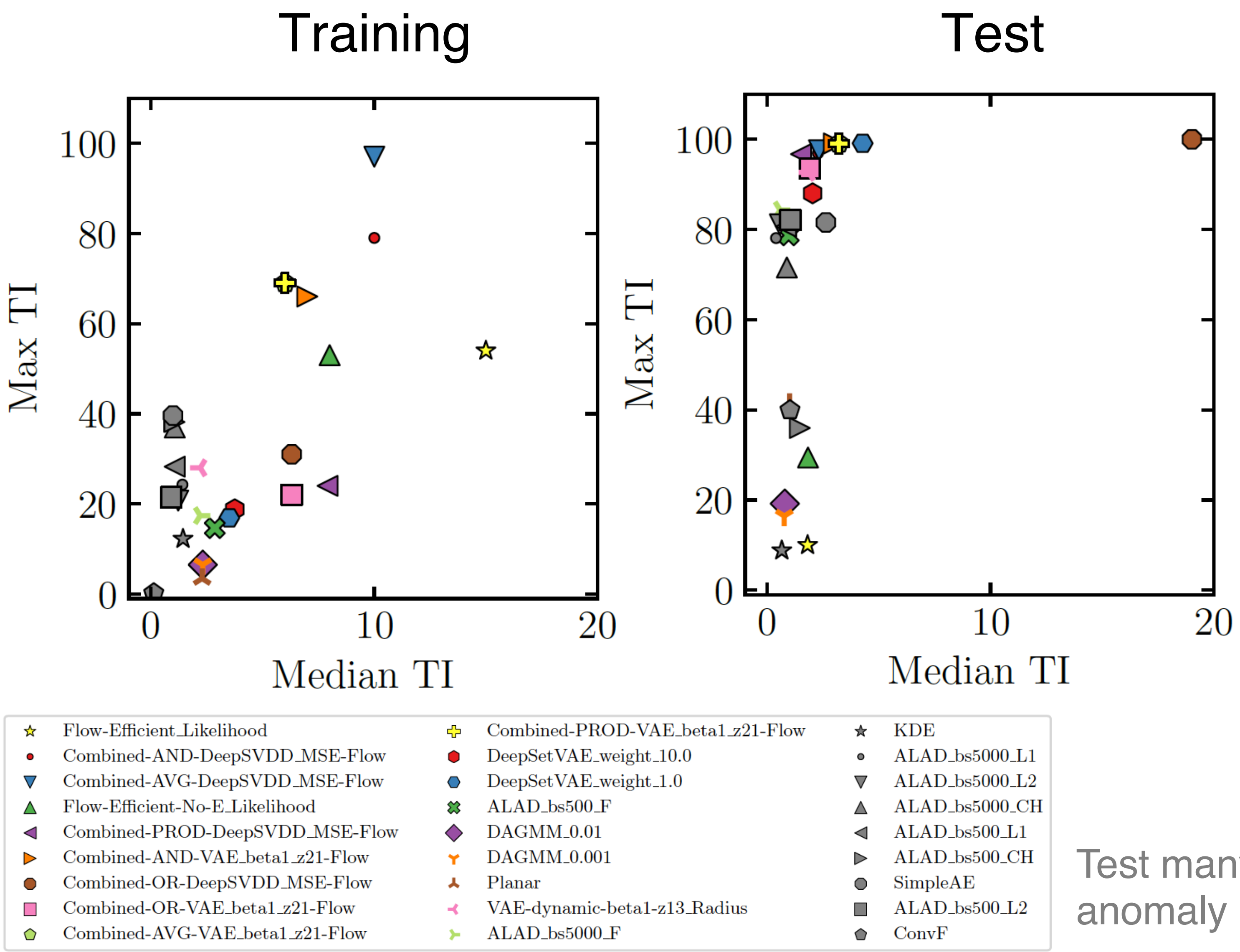
Interpretation w/o benchmarks



How to interpret null results?

We do not know what it is that we have not found

Poor man's assessment: *benchmarking*



TI = Total Improvement:
significance improvement over many signal benchmarks

- Max TI
- Median TI
- (Min TI)

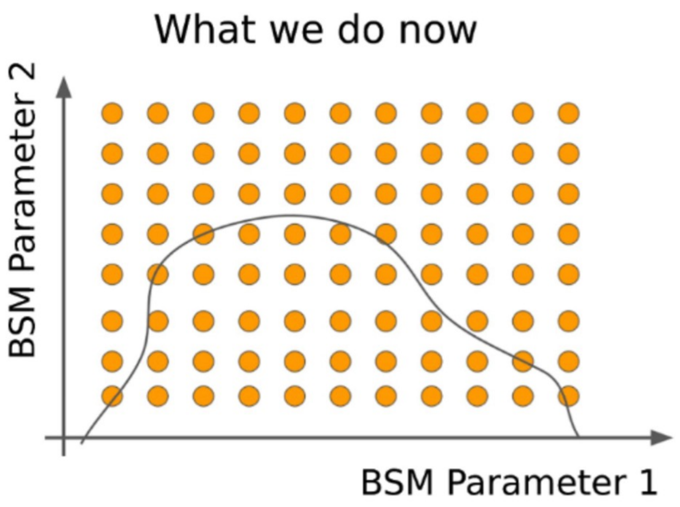
Objective:
Make *reinterpretation* possible

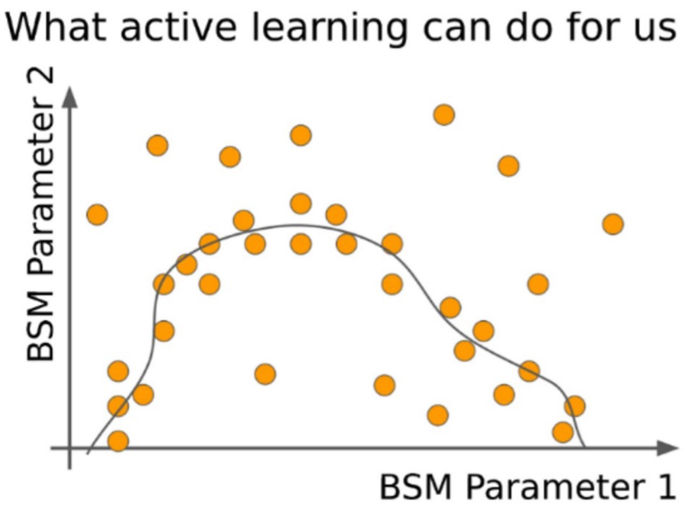
Test many anomaly models

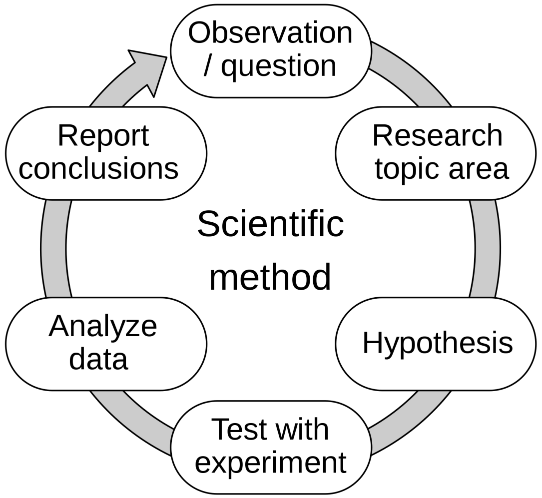
[2105.14027]

Recastability to close the loop

Smart sampling with active learning: **simulate on demand**





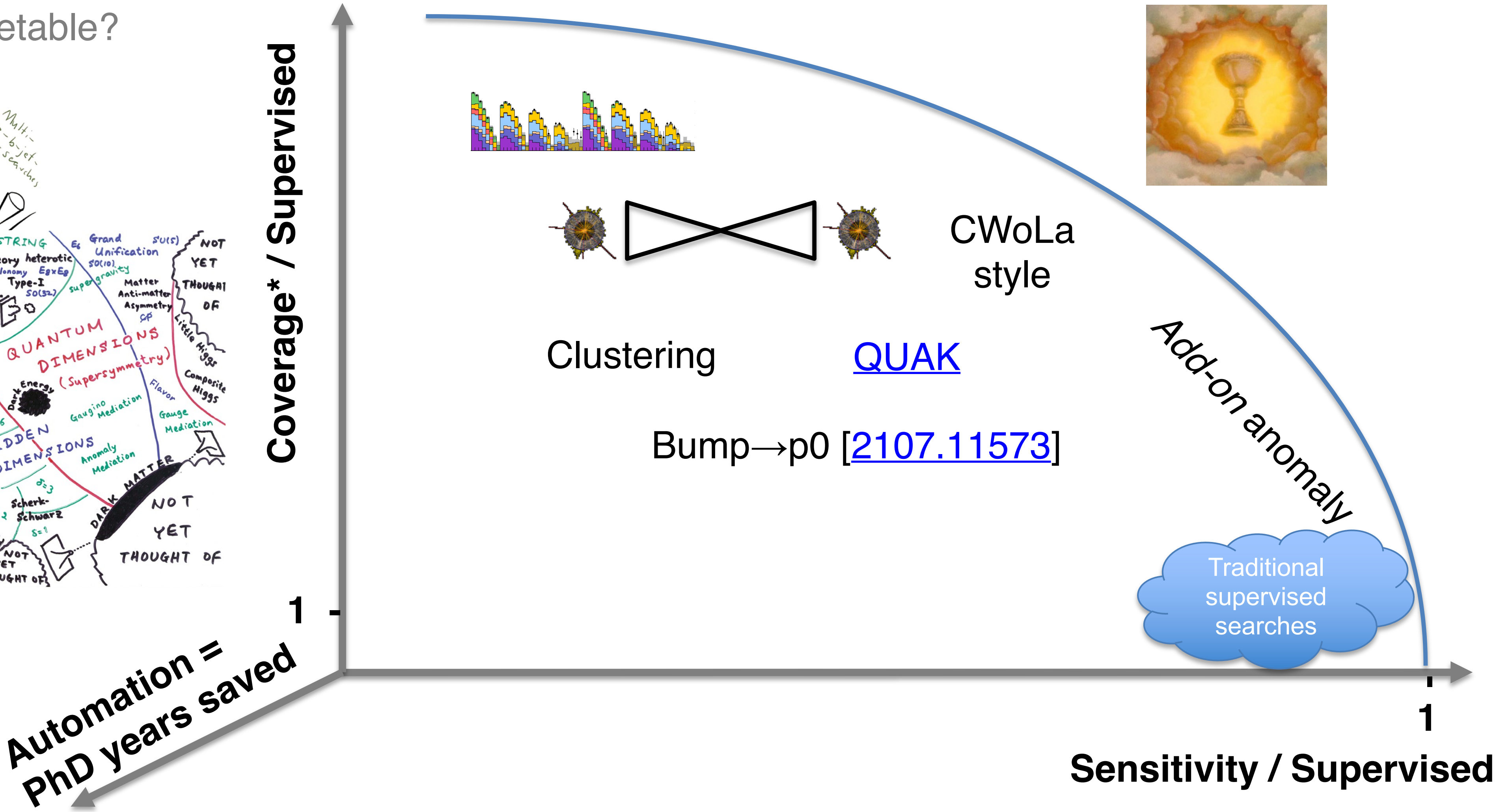


Thrives on high-dimensional space

[Regress on upper limit with GP for ATLAS mono-H(bb) search, ACAT22, ATL-PHYS-PUB-2022-045]

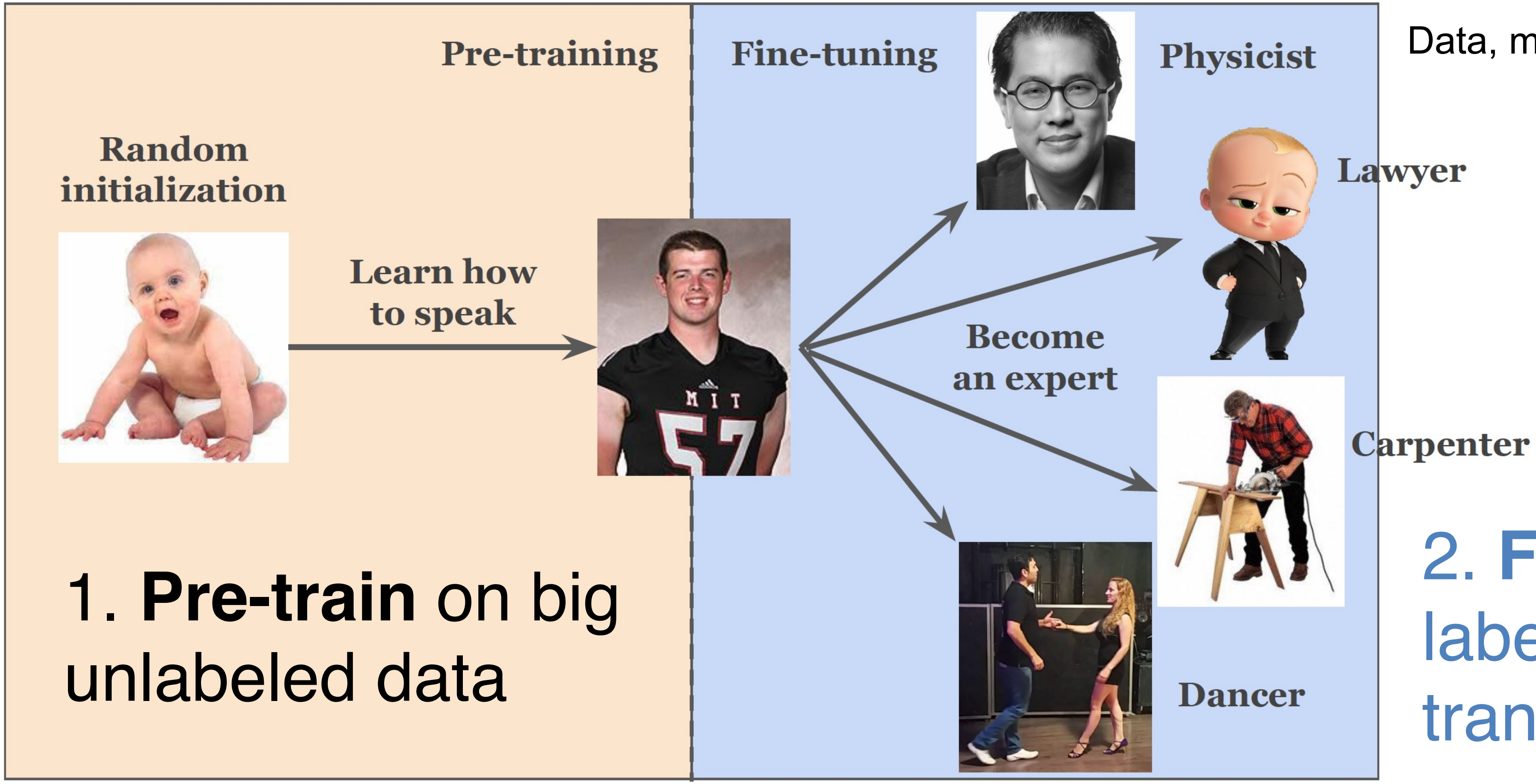
Quantifying search capability

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



Looking ahead [speculative]

The power of foundation models [LLM]



1. **Pre-train** on big unlabeled data

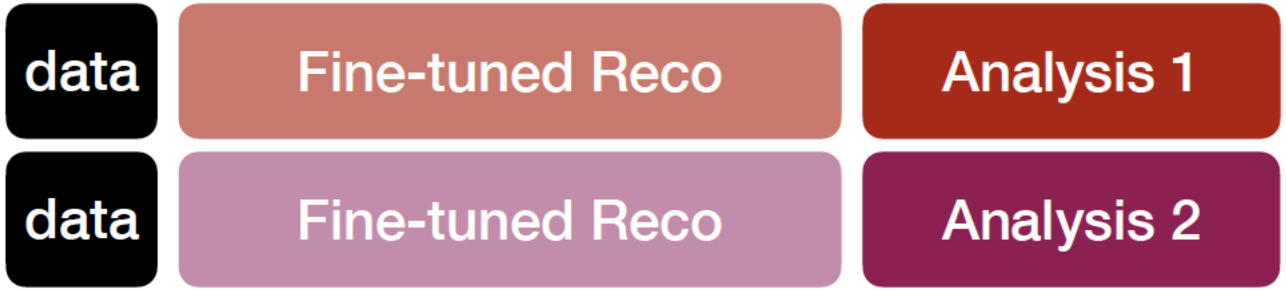
Data, memory & compute hungry

2. **Fine-tune** on labeled data + transfer learning

[Image credit: Kazuhiro Terao]

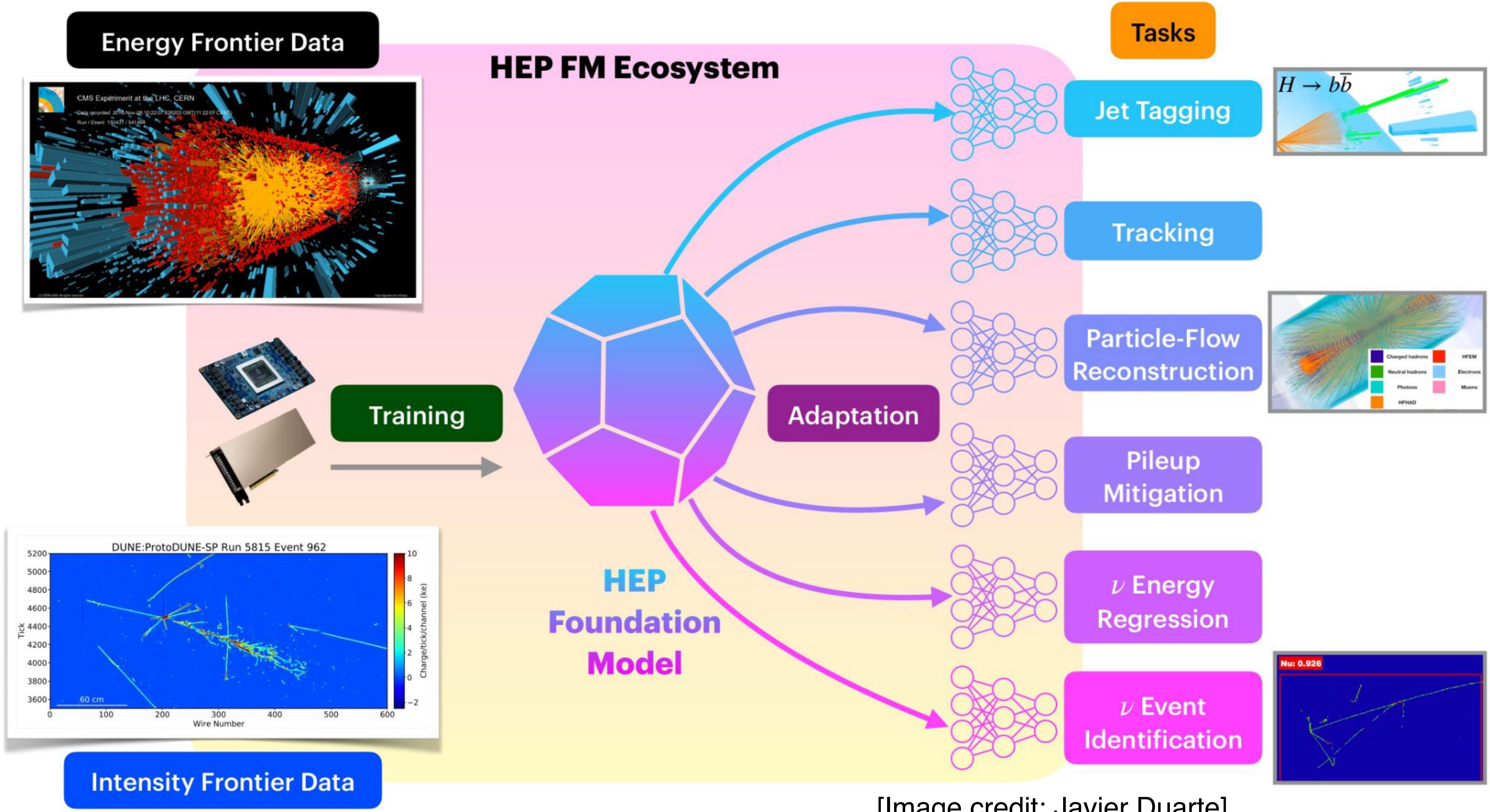
ChatGPT for HEP? – *Maximalist* ML

Big common pre-trained feature extractor:
Low-level features → Truth (e.g. *Higgs score)*



Open question:
One backbone > $\sum$ backbones per object ?

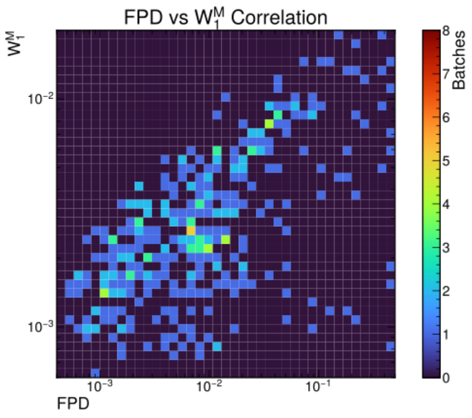
[Lukas Heinrich, H&Ns 2022]



[Image credit: Javier Duarte]

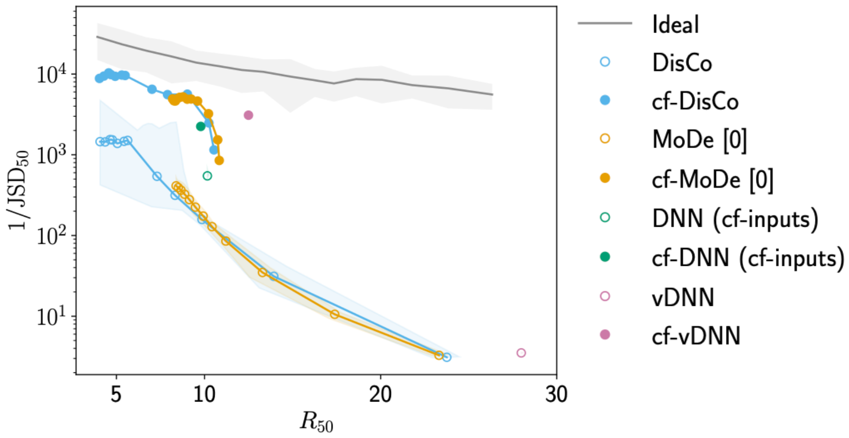
Towards a discussion

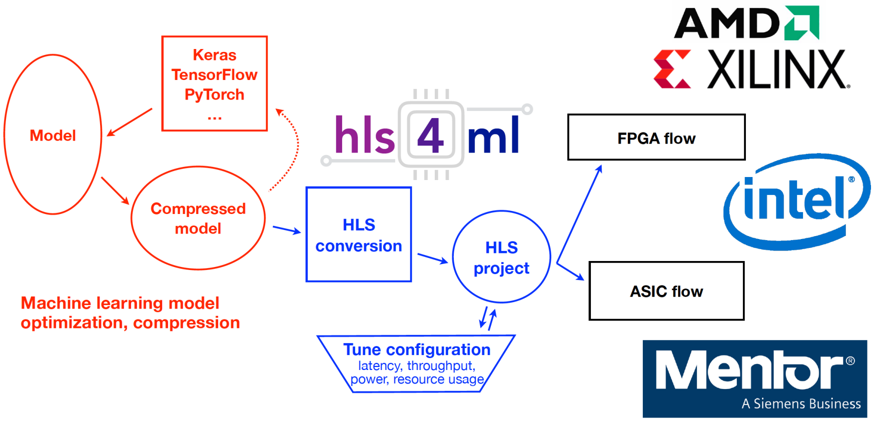
Many more challenges



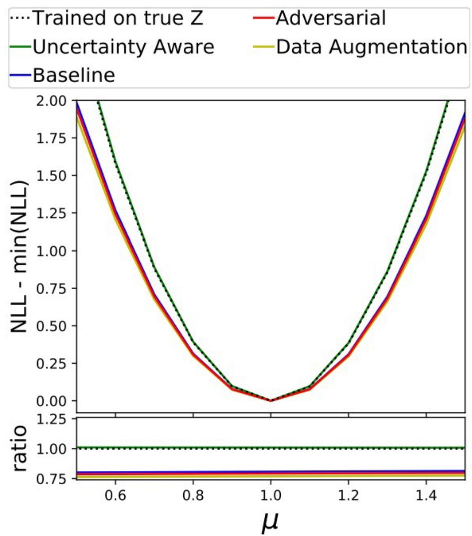
Evaluation of gen. models
[Compare metrics 2211.10295]

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 ↔ **Slow** Experiment time scale

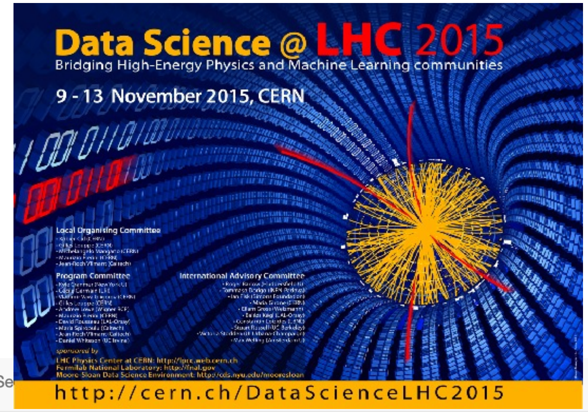
ML@HEP **competitive** ↔ ***Open Science*** @ Experiment

Need faster **concept-to-production** cycle

& Opportunities

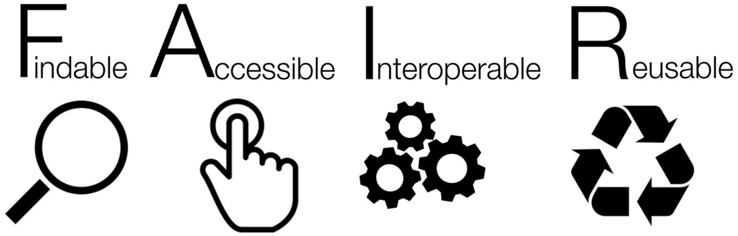
- AI as a *muse* to science
  - ML to suggest new theories [active learning]
- Human-in-the-loop AI
  - Optimal detector design assisted by AI
- Differentiable programming → differentiable physics
- Data analysis in theory space [simulation-based inference]
- Diverse AI-assisted search portfolio [rigor/bias/automation]
- 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



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 researchers to generate realistic HEP datasets for training and testing. These simulators could range from truth-le[…]
sensor arrays.
- Performance metric[…]
solutions.

Sign in with ORCID

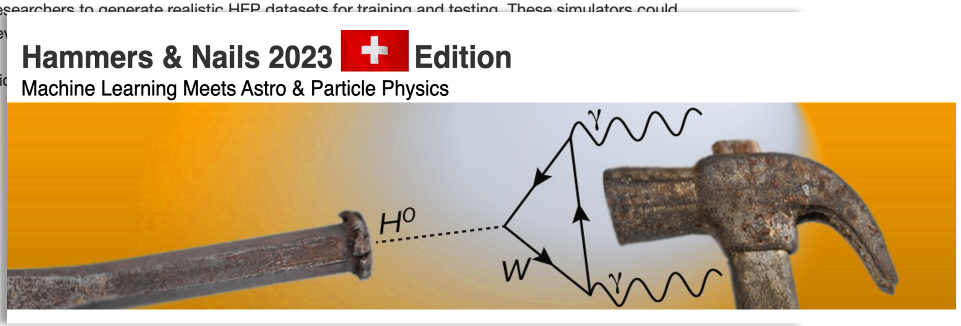
Authors

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

<u>Outlook</u>:

Attack problems which were considered unsolvable

ATLAS EXPERIMENT

syn·er·gy | 'sinərjē |

UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE

PIs


TG


Slava Voloshynovskiy


François Fleuret

PhD students


Tomke Schröer


Malte Algren


Lukas Ehrke


Matthew Leigh


Debajyoti Sengupta


Sam Klein


Guillaume Quétant


Mariia Drozdova


Ivan Oleksiyuk


Bálint Máté


Atul Kumar Sinha


Daniele Paliotta

postdocs


Knut Zoch


Kinga Wozniak


Johnny Raine

This could be you !


Olga Taran

Full spectrum: physics - CS

FONDATION ERNEST BONINCHI

Alexander von Humboldt Stiftung / Foundation

Swiss National Science Foundation

Sinergia

Swiss Government Excellence Scholarships
STUDY IN SWITZERLAND