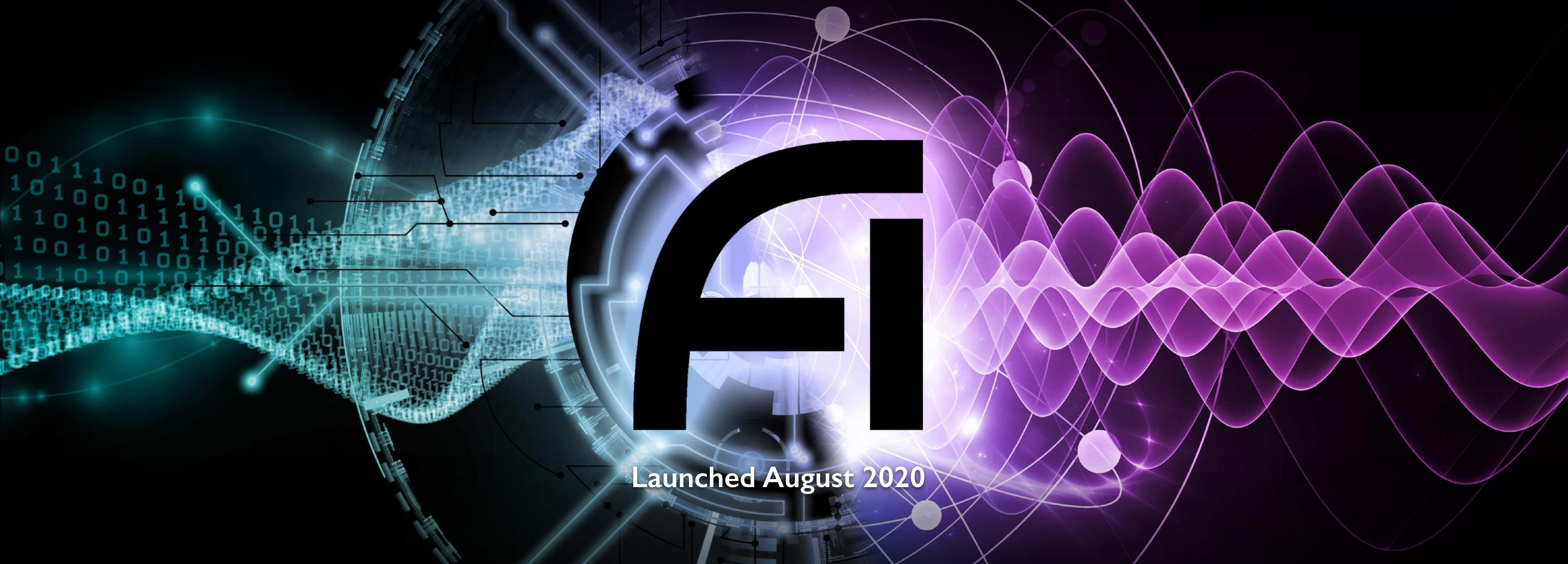
Please put captions in this "black box"

Interpretable Machine Learning for Particle Physics

Jesse Thaler

APS DPF / Pheno 2024, University of Pittsburgh — May 14, 2024

The NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI /aɪ-faɪ/ iaifi.org)



Deep Learning (AI) + Deep Thinking (Physics) = Deeper Understanding

Next Generation of AI + Physics Talent

IAIFI Postdoctoral Fellows

Application deadline typically early October

Albergo

AI and Statistical Physics

Boyda

AI for Lattice QCD

Bright-Thonnney

AI for Particle Physics

Cuesta

AI for Cosmological Observations

Dogra

Mathematical Physics of AI

Gagliano

AI for Time-Domain Astronomy

Golubeva

⇓ AI

Grosso

AI for Collider Physics

Harvey

AI for String Theory

Luo

⇓ UCLA

Micallef

AI for Neutrino Physics

Mishra-Sharma

⇓ BU

Yang

AI Frontiers of Reinforcement Learning

IAIFI Summer School & Workshop

IAIFI

Summer School
August 5–August 9
2024

Summer Workshop: August 12–16, 2024

Machine Learning at DFP-Pheno 2024

Compared to 3 talks at Pheno 2019!

Machine Learning & AI: New Physics

Decision tree autoencoder anomaly detection on FPGA at L1 triggers - take 2 *David Lawrence 105, University of Pittsburgh*	*Tae Min Hong* 14:00 - 14:15
AutoDQM for Anomaly Detection in the CMS Detector *David Lawrence 105, University of Pittsburgh*	*Chosila Sutantawibul* 14:15 - 14:30
Residual ANODE *David Lawrence 105, University of Pittsburgh*	*Ranit Das* 14:30 - 14:45
Exploring Optimal Transport for Event-Level Anomaly Detection at the Large Hadron Collider *David Lawrence 105, University of Pittsburgh*	*Hancheng Li* 14:45 - 15:00
Constraining the SMEFT Higgs Sector with Machine Learning *David Lawrence 105, University of Pittsburgh*	*Radha Mastandrea* 15:00 - 15:15
Probing a GeV-scale Scalar Boson and a TeV-scale Vector-like Quark Associated with $SU(1)_{T3R}$ at the Large Hadron... *Umar Sohail Qureshi*	

Computing, Analysis Tools, and Data Handling

ARCANE Reweighting: A Solution to the Negative Weights Problem in Collider Monte Carlo *David Lawrence 105, University of Pittsburgh*	*Prasanth Shyamsundar* 16:00 - 16:15
A Matrix-Based Approach for Jet-Parton Assignment Leveraging Mass and Momentum Using CMS Open Data *Eric Reinhardt*	
Resolving Combinatorial Problems with Quantum Algorithms *David Lawrence 105, University of Pittsburgh*	*Jacob Scott* 16:30 - 16:45
Multi-vertex jet trigger at ATLAS' upgrade for HL-LHC using Boosted Decision Trees on FPGAs *David Lawrence 105, University of Pittsburgh*	*Santiago Cane* 16:45 - 17:00
Data Quality Monitoring for the HL-LHC Upgrade to the CMS Outer Tracker *David Lawrence 105, University of Pittsburgh*	*Brandi Nicole Skipworth* 17:00 - 17:15
A Herwig7 Underlying Event Tune for Relativistic Heavy Ion Collider Energies at 200 GeV *David Lawrence 105, University of Pittsburgh*	*Umar Sohail Qureshi* 17:15 - 17:30

Machine Learning & AI: Collider Physics

Trackless Jet Vertexing and Timing using ML *Law 109, University of Pittsburgh*	*Wen Han Chiu* 14:00 - 14:15
Towards a data-driven model of hadronization using normalizing flows *Law 109, University of Pittsburgh*	*Ahmed Youssef* 14:15 - 14:30
Search for New Physics in the Merged Diphoton plus Photon final state with the CMS Detector *Law 109, University of Pittsburgh*	*Austin Edwin Townsend* 14:30 - 14:45
The versatility of flow-based fast calorimeter surrogate models *Law 109, University of Pittsburgh*	*Ian Pang* 14:45 - 15:00
Studies into di-Tau mass reconstruction for high mass resonances at the ATLAS experiment *Law 109, University of Pittsburgh*	*Kyle Angelo Granados* 15:00 - 15:15
Deep Learning Based Tagger for Highly Collimated Photons at CMS *Law 109, University of Pittsburgh*	*Kyungmin Park* 15:15 - 15:30

Quantum Field & String Theory: Non-perturbativity and Amplitudes

Machine learning and (large-N) field theory *David Lawrence 104, University of Pittsburgh*	*Zhengkang Zhang* 16:45 - 17:00

QCD & Heavy Ion Physics: Jets and Energy Correlators

Jet Calibration in ATLAS Using Machine Learning Networks *Law 107, University of Pittsburgh*	*Benji Lunday* 15:00 - 15:15

Electroweak & Higgs Physics: Electroweak Physics at the LHC

New W Boson Decay Channel at the LHC *David Lawrence 207, University of Pittsburgh*	*Peiran Li* 17:00 - 17:15

Instrumentation: Neutrinos, Dark Matter, and Scintillation

NuDot, R&D testbed for future large-scale neutrino detectors *Law 111, University of Pittsburgh*	*Masooma Sarfraz* 14:45 - 15:00

Today at Lunch: DOE PI Meeting Computational HEP and AI/ML

Dark Matter: WIMPs, DM Simulation and ML

Sweeping the Dust Away: An unbiased map of the Milky Way's gravitational potential using unsupervised ML *David Lawrence 120, University of Pittsburgh*	*Eric Putney* 16:45 - 17:00

Mini-Symposium: Quantum Instrumentation

Exploring Quantum Machine Learning for High-Energy Physics *University of Pittsburgh / Carnegie Mellon University*	*Jinghong Yang* 15:10 - 15:30

Coordinating Panel for Software and Computing (CPSC) Townhall

Coordinating panel for software and computing townhall *University of Pittsburgh / Carnegie Mellon University*	17:30 - 18:00

Mini-Symposium: Neutrino Science with the DUNE Experiment

Deep-learning at DUNE Far Detector *University of Pittsburgh / Carnegie Mellon University*	*Prof. Jianming Bian* 16:45 - 17:00
Neural Network Based Fast Optical Simulation Method in ProtoDUNE-VD *University of Pittsburgh / Carnegie Mellon University*	*Shuaixiang (Shu) Zhang* 17:00 - 17:15

"…but what is the machine actually learning?"

What does it really mean for ML to be "Interpretable"?

(Or explainable, trustworthy, safe, robust, aligned, helpful, transparent, …)

Obligatory apology that examples below are
heavily drawn from my research in collider physics

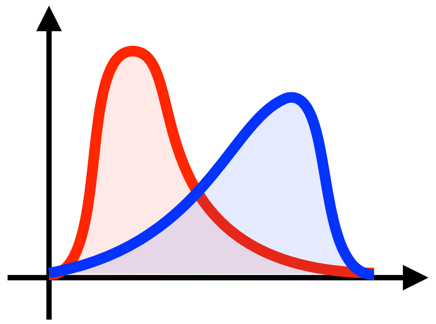
"…but what is the machine actually learning?"

My evolving perspective:

The desire for human interpretability often arises when we imperfectly specify the task we want to accomplish

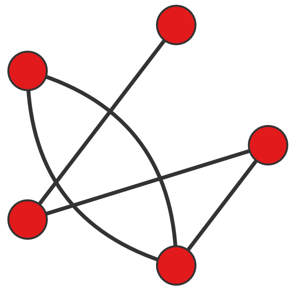
A more actionable definition of interpretability: identifying low-rank structures in high-dimensional datasets

Interpretable Machine Learning for Particles Physics



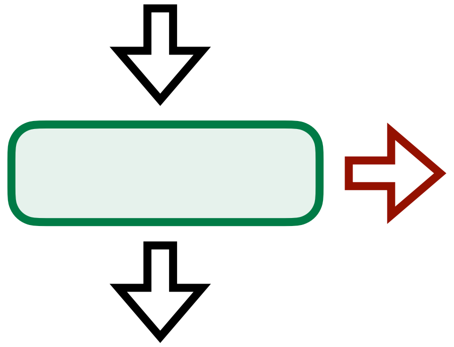
Confronting the Black Box

*To benefit from machine learning advances, we must ensure that our algorithmic choices align with our **scientific goals***



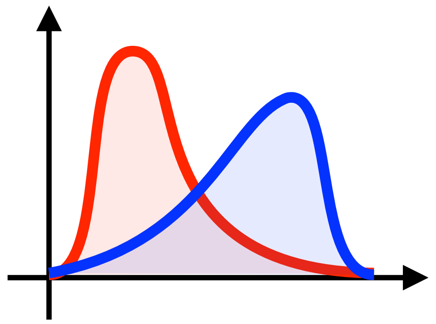
Case Study in Jet Classification

*When possible, pursue **active interpretability**, where you control the network architecture and training paradigm*



The Next Frontier for Interpretability

*Foundation models identify **generically useful features**, which challenge the importance of task alignment*



Confronting the Black Box

To benefit from machine learning advances, we must ensure that our algorithmic choices align with our ***scientific goals***

Case Study in Jet Classification

When possible, pursue ***active interpretability****, where you control the network architecture and training paradigm*

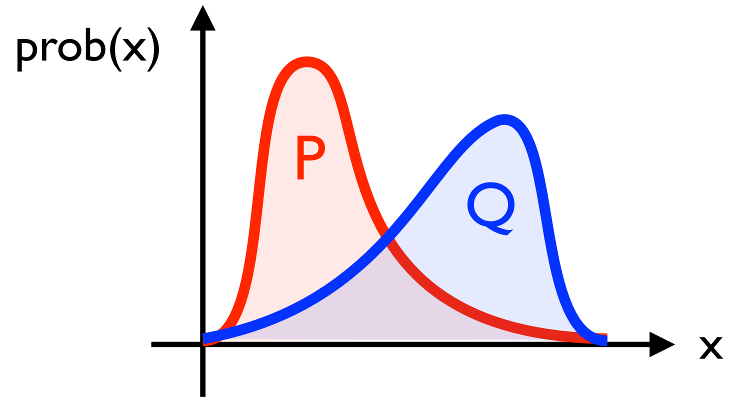
The Next Frontier for Interpretability

Foundation models identify ***generically useful features****, which challenge the importance of task alignment*

Likelihood Ratio Trick

*Key example of **simulation-based inference***

Many HEP problems can be expressed in this form!



Goal:	Estimate p(x) / q(x)
Training Data:	Finite samples P and Q
Learnable Function:	f(x) parametrized by, e.g., neural networks

Loss Function(al): $$L = -\langle \log f(x) \rangle_P + \langle f(x) - 1 \rangle_Q$$

[see e.g. Cranmer, Pavez, Louppe, arXiv 2015; D'Agnolo, Wulzer, PRD 2019; simulation-based inference in Cranmer, Brehmer, Louppe, PNAS 2020; relation to f-divergences in Nguyen, Wainwright, Jordan, AoS 2009; Nachman, Thaler, PRD 2021]

Likelihood Ratio Trick

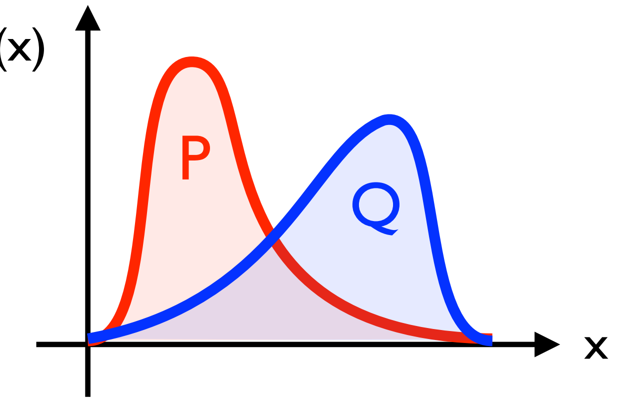
Key example of simulation-based inference

Many HEP problems can be expressed in this form!

Goal: Estimate p(x) / q(x)

Training Data: Finite samples P and Q

Learnable Function: f(x) parametrized by, e.g., neural networks



Loss Function(al):

$$L = -\langle \log f(x) \rangle_P + \langle f(x) - 1 \rangle_Q$$

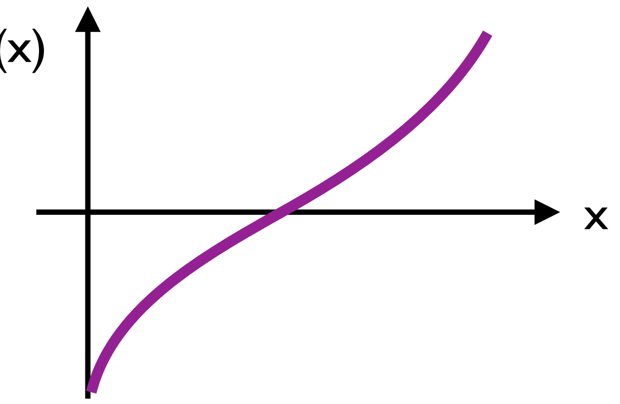
Asymptotically:

$$\underset{f(x)}{\arg\min}\, L = \frac{p(x)}{q(x)}$$

Likelihood ratio

$$-\min_{f(x)} L = \int \mathrm{d}x\, p(x) \log \frac{p(x)}{q(x)}$$

Kullback–Leibler divergence



[see e.g. Cranmer, Pavez, Louppe, arXiv 2015; D'Agnolo, Wulzer, PRD 2019;
simulation-based inference in Cranmer, Brehmer, Louppe, PNAS 2020;
relation to f-divergences in Nguyen, Wainwright, Jordan, AoS 2009; Nachman, Thaler, PRD 2021]

Likelihood Ratio Trick

Key example of simulation-based inference

Many HEP problems can be expressed in this form!

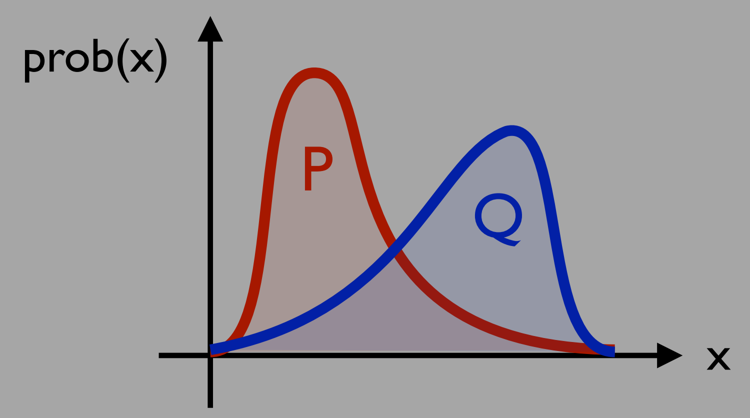
Asymptotically, same structure as Lagrangian mechanics!

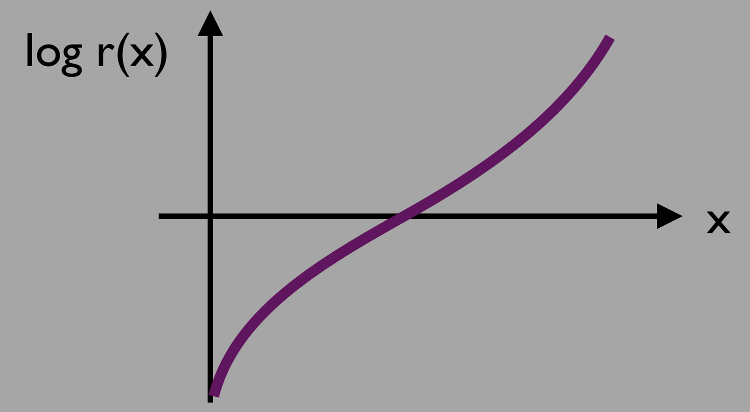
Action: $L = \int \mathrm{d}x \, \mathcal{L}(x)$

Lagrangian: $\mathcal{L}(x) = -p(x) \log f(x) + q(x)\big(f(x) - 1\big)$

Euler-Lagrange: $\dfrac{\partial \mathcal{L}}{\partial f} = 0$ Solution: $f(x) = \dfrac{p(x)}{q(x)}$

Requires shift in focus from solving problems to specifying problems





[see e.g. Cranmer, Pavez, Louppe, arXiv 2015; D'Agnolo, Wulzer, PRD 2019;
simulation-based inference in Cranmer, Brehmer, Louppe, PNAS 2020;
relation to f-divergences in Nguyen, Wainwright, Jordan, AoS 2009; Nachman, Thaler, PRD 2021]

"What is the machine learning?"

For this loss function, an estimate of the likelihood ratio derived from sampled data and regularized by the network architecture and training paradigm

"What is the machine learning?"

For this loss function, an estimate of the likelihood ratio derived from sampled data and regularized by the network architecture and training paradigm

"But I want to understand what it has learned!"

Do you really expect the likelihood ratio to take on a particularly nice functional form?

N.B. QFT calculations often involve special functions that have no elementary representation

"…"

Why might we want ML to be "Interpretable"?

Or explainable, trustworthy, safe, robust, aligned, helpful, transparent, …

Scientific Reasons:

Could be working in **non-asymptotic** regime
Training data might be **biased** in some way
Result could depend on **poorly modeled** features
Limited ability to perform independent **validation**
Need for compact **symbolic** expressions
Desire to **generalize** away from specific context
…

Sociological Reasons:

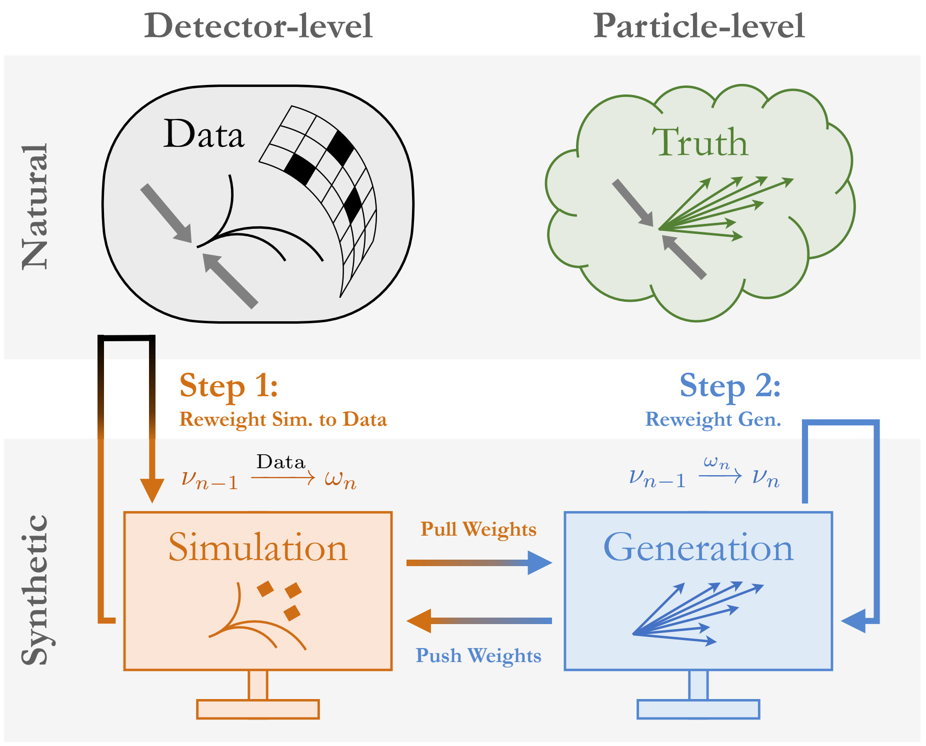
Skeptical of algorithmic/statistical/computational reasoning
Need to explain decisions to external **stakeholders**
Desire to **manage risks** from unforeseen outcomes
…

All valid reasons, but suggest imperfect specification of our initial goals!

Likelihood Ratio Trick in HEP

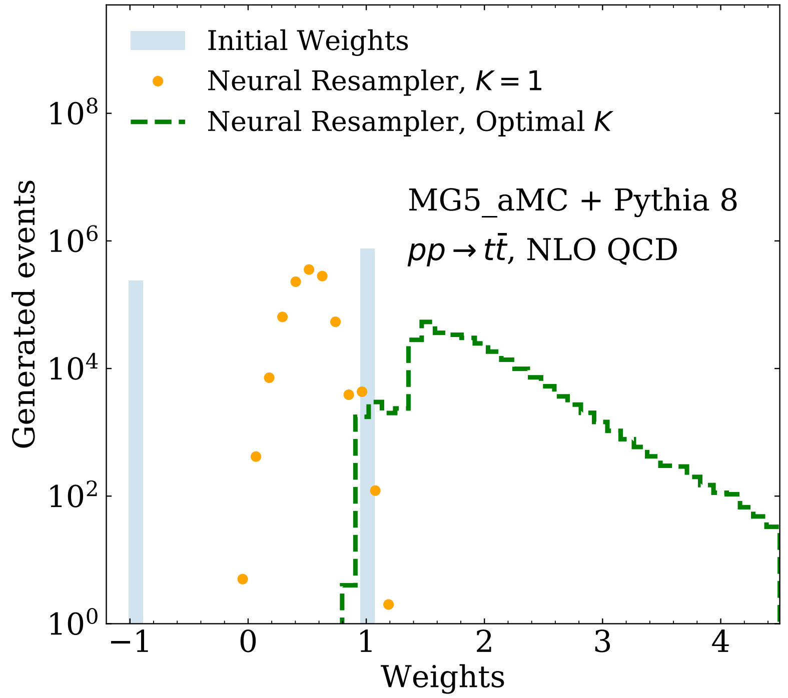
Apologies that examples are all from my own work

Detector Unfolding



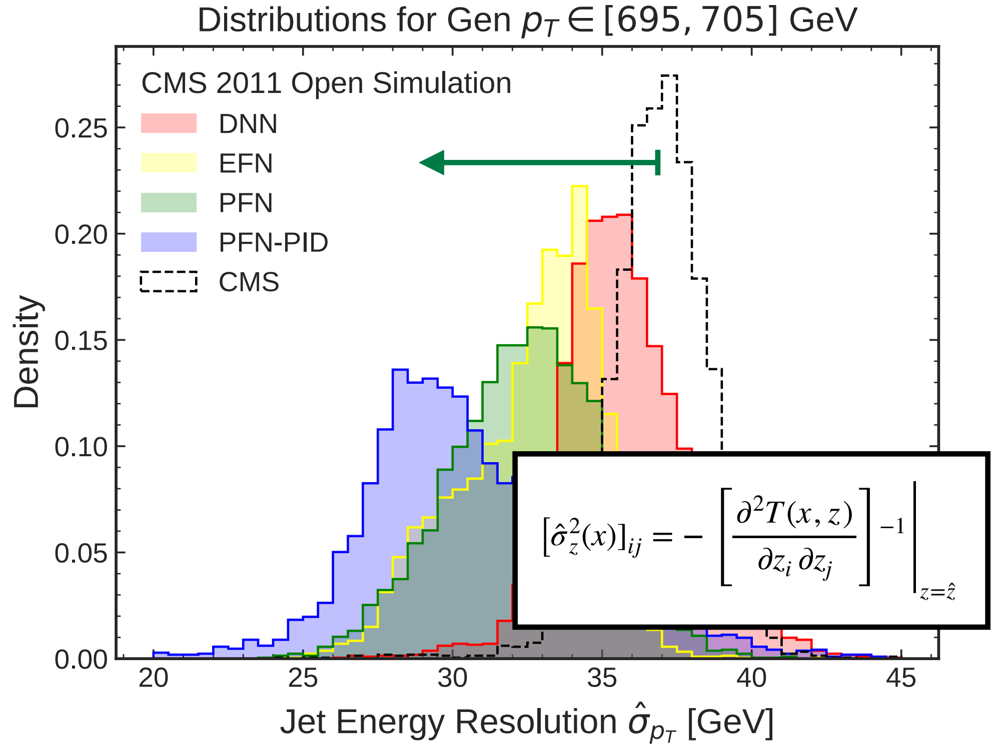
[Andreassen, Komiske, Metodiev, Nachman, JDT, PRL 2020; + Suresh, ICLR SimDL 2021]

Monte Carlo Reweighting



[Nachman, JDT, PRD 2020; inspired by Andersen, Gutschow, Maier, Prestel, EPJC 2020]

Resolution Estimation



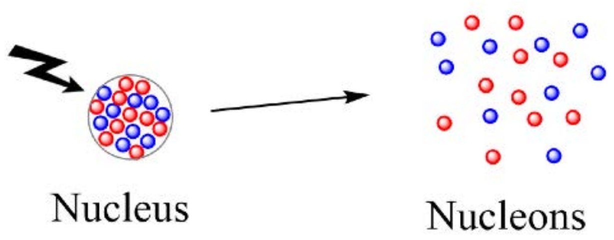
[Gambhir, Nachman, JDT, PRL 2022, PRD 2022]

For these applications, goal is "accuracy" more than "interpretability"

Ask me offline why I think standard methods to assess accuracy, quantify uncertainties, and validate results are incomplete

“Interpretability” as the Primary Goal

E.g. modeling nuclear binding energies



Symbolic Regression

N = neutron number
Z = proton number
A = atomic mass
I = isospin asymmetry
P = Casten factor

Model Order	Obtained Function
1	$\left(\frac{Z}{N} + Z - \frac{1.06N}{Z}\right)\left(I\left(32.4 - \frac{\sqrt[3]{A}N}{Z}\right) + 16.7\right)$
2	$3.42(Z - 14.6)\left(\sqrt[3]{A} - 2.19I - 4.38\right)(I - 0.110\log(A)) + \delta - P + 0.301$
3	$-2.02e^{-0.40\gamma_Z\gamma_N P-(0.040Z)^Z} + 2.99 \cdot 0.867^{(N-Z)^2} - 0.426P(\log(Z) - 3.30) + I$
4	$A^{2/3}e^{-A^{2/3}+Z^{1.10}-Z} I \log\left(\frac{Z}{N}\right) + 0.634e^{A^{2/3}+\sqrt[3]{A}-N} + 0.290\gamma_N\gamma_Z + 0.246$
5	$(0.0000154)^P A^{\frac{2P}{3}}\left(P(N-Z)^2 + N\right)\left(0.0000154(N-1)^2 + P\right)$
6	$\frac{\gamma_N(1-\gamma_Z)}{N} - \exp\left(\left(\sqrt[3]{A} - \frac{Z}{N} - 1.21\right)\left(2(P + 0.108)\left(P^{\sqrt[3]{A}} - P^N\right) - \gamma_Z(1-\gamma_N) - \frac{Z}{N} - 0.426\right)\right)$
7	$1.35I\left((0.324 - I)\left(-\frac{Z}{N} - 1.78\right)\left(\frac{4.30N}{Z} - 0.111\left(A + e^P\right)\right) - P + 1.35I\right)$
8	$(-0.801^{\gamma_N(1-\gamma_Z)} + 0.570P - 2I)(-0.112 + (A - (N-Z)^2 + \frac{AN}{Z})0.801^N)$
9	$9.20 \cdot 10^{-23} \cdot 1330^{A^{\frac{1}{3}}}(-1.97 + \gamma_N(-1 + \gamma_Z) - \gamma_Z + P)(-1330 + N^2 - 2NZ + Z^2)(-670 + N^2 - 2NZ + Z^2)$
10	$3.02\exp\left(-1.91P^{1.94}\left(\frac{Z}{N}\right)^{A^{2/3}} - 0.895e^{-0.227N}N^2 - 0.0268(N-1)^2\right)$

[Munoz, Udrescu, Garcia Ruiz, arXiv 2024; see also
Cranmer, Sanchez-Gonzalez, Battaglia, Xu, Cranmer, Spergel, Ho, NeurIPS 2020]

Cf. Semi-Empirical Mass Formula

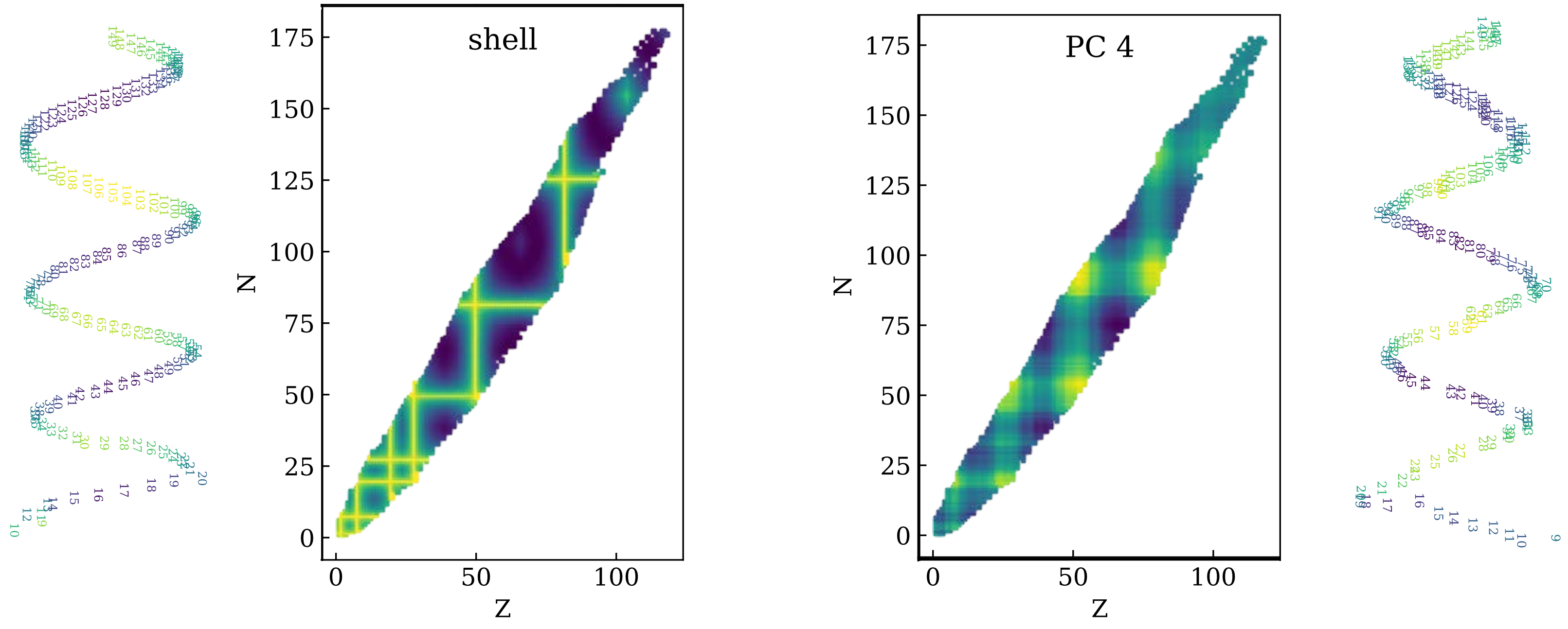
[Weizsäcker, 1935]

$$E(Z,N) = \left(-\sqrt{\alpha^2+\beta^2} + \sqrt{\alpha^2+\beta^2\frac{(Z-N)^2}{(Z+N)^2}}\right)\left[(Z+N-1) - \gamma(Z+N-1)^{2/3}\right] + \frac{3e^2}{r_0(Z+N)^{1/3}}\left(1 - \delta\frac{|Z-N|}{Z+N}\right)\left[\frac{Z^2}{5} - \left(\frac{Z}{2}\right)^{4/3}\right]. \quad (51)$$

Die Konstanten α, β, γ, δ, r_0 wurden nun auf zwei Wegen bestimmt.

Latent Space Topography

Human vs. Machine



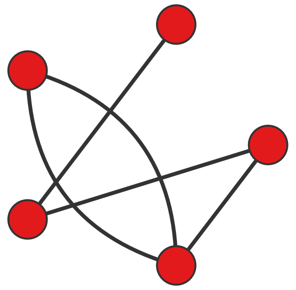
[Kitouni, Nolte, Trifinopoulos, Kantamneni, Williams, ICML 2023;
+ Pérez-Díaz, to appear ICML 2024]

Identifying low-rank structures in high-dimensional datasets

This is an actionable definition of interpretability, which may or may not be relevant to the physics problem of interest

Confronting the Black Box

*To benefit from machine learning advances, we must ensure that our algorithmic choices align with our **scientific goals***



Case Study in Jet Classification

*When possible, pursue **active interpretability**, where you control the network architecture and training paradigm*

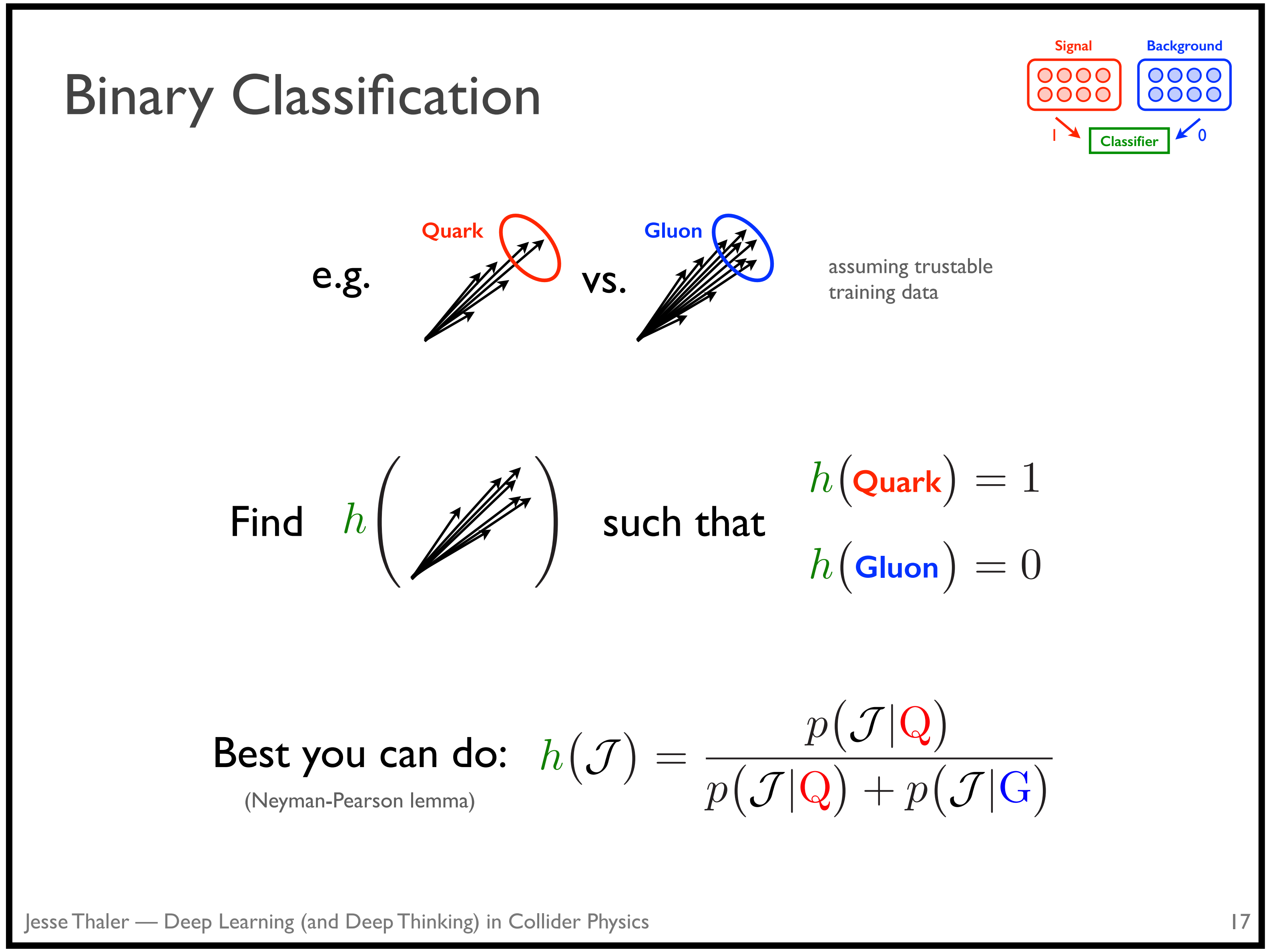
The Next Frontier for Interpretability

*Foundation models identify **generically useful features**, which challenge the importance of task alignment*

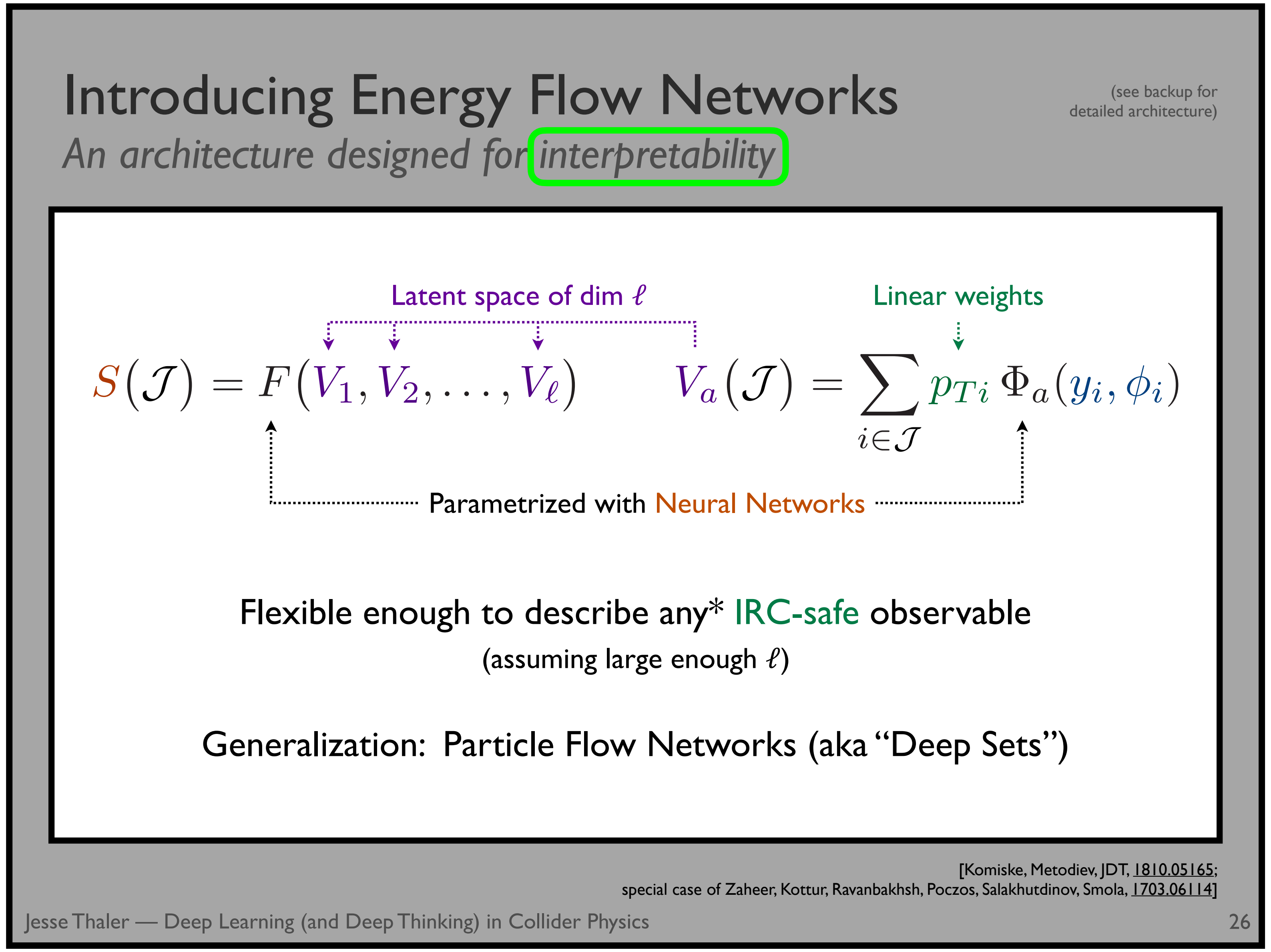
The More Things Change…

Jet classification, from my talk at Pheno 2019

Application of Likelihood Ratio Trick

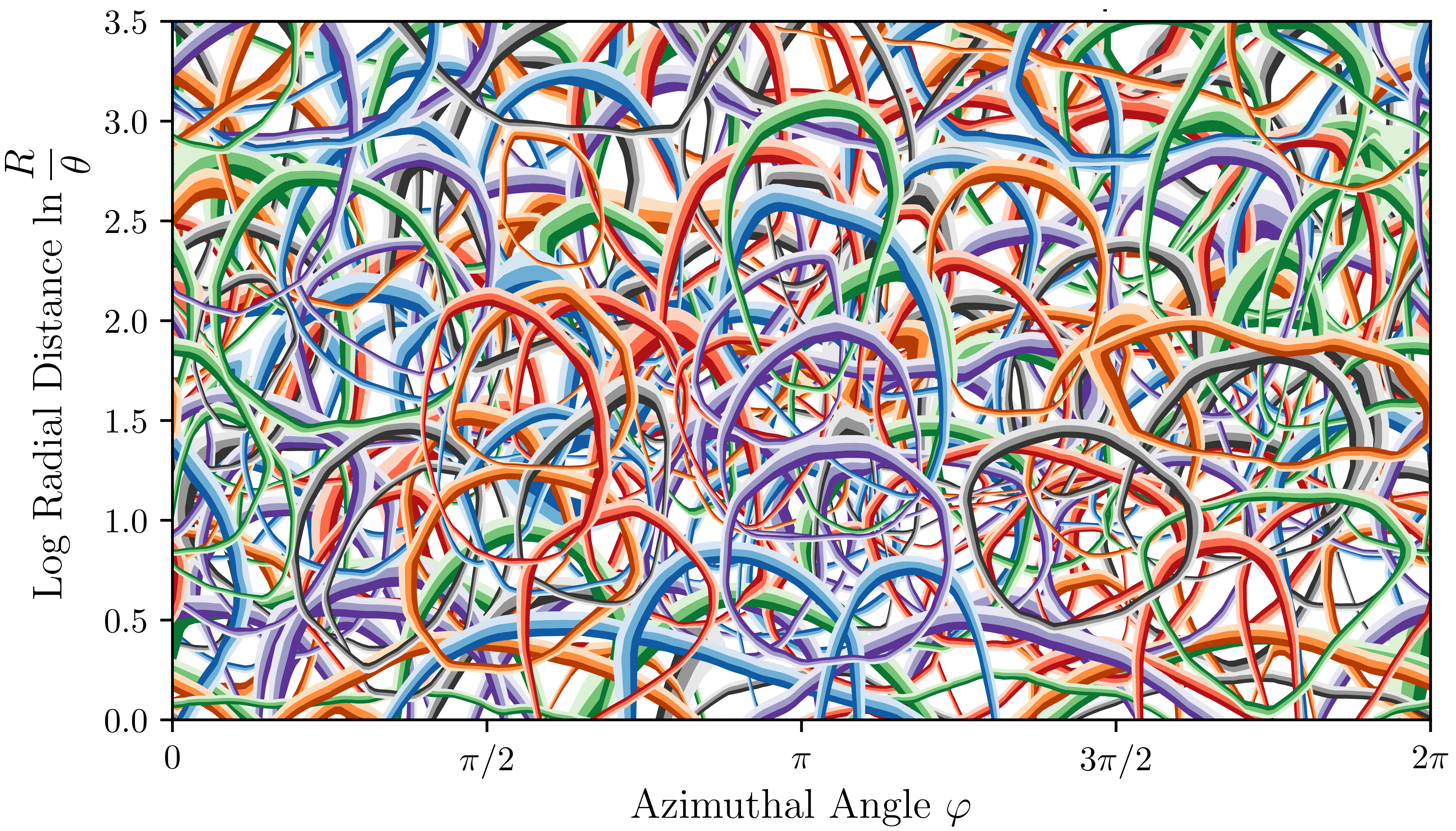


Interpretability in Machine Learning



Does this Really Count as "Interpretable"?

Visualizing Energy Flow Networks



Trying to plot 256 dimensional latent space

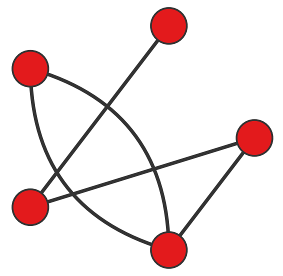
See Pheno 2019 talk for insights at L = 2

[Komiske, Metodiev, JDT, JHEP 2019]

Three Lessons since Pheno 2019

Highlighting the power of active interpretability

Apologies that examples are all from my own work



If you have a catalog of trusted observables, you can translate a black-box algorithm on low-level inputs into a simple classifier on high-level features

$$\langle \Phi^{a_1} \Phi^{a_2} \rangle_{\mathcal{P}}$$

If there are simple operations like multiplication and sums that don't really require "interpretation", you can bake those into your machine learning architecture

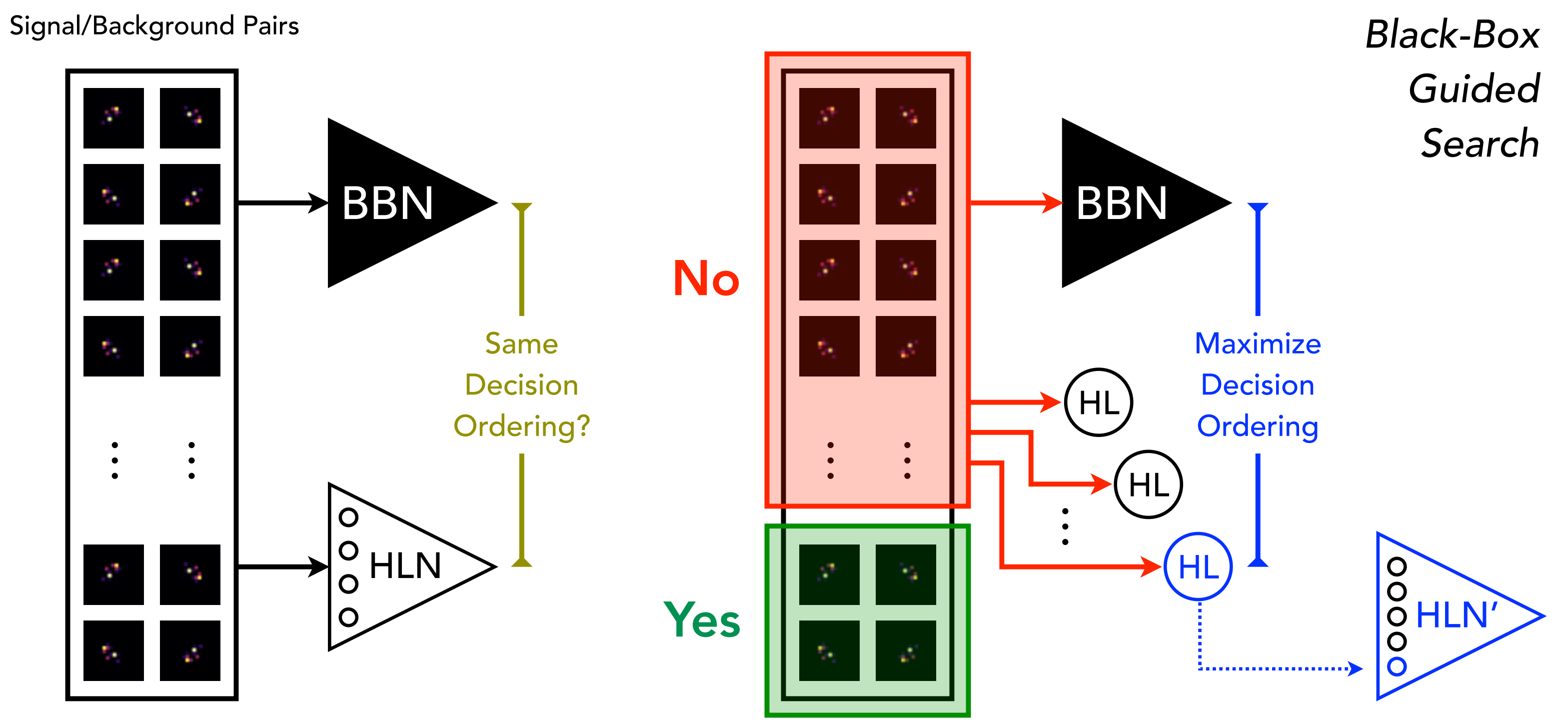
$$\begin{aligned} &\|\Phi(\hat{p}_1) - \Phi(\hat{p}_2)\| \\ &\leq \ L\|\hat{p}_1 - \hat{p}_2\| \end{aligned}$$

If there is a property you want your network to have, make sure to impose algorithmic guardrails, otherwise the machine might pursue undesirable optimization

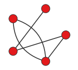
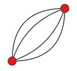
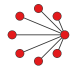
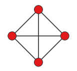
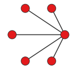
Translating the Black Box

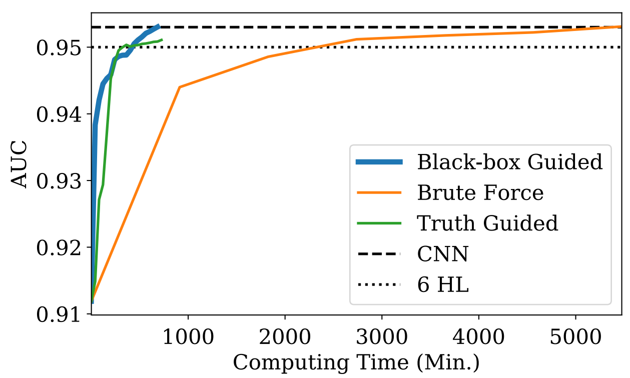
Selecting Energy Flow Polynomials that mimic CNN decisions

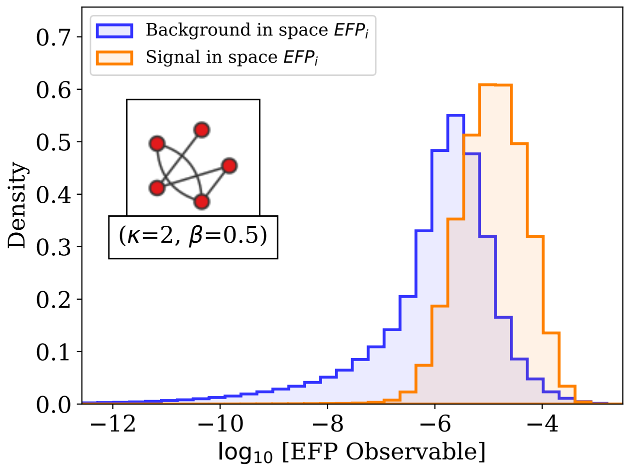
Iteratively building likelihood ratio estimate from catalog of high-level observables



A glimpse at an alternative history for field of jet substructure

	Iteration (n)	EFP	κ	β	Chrom #
	0	$M_{\text{jet}} + p_{\text{T}}$	–	–	–
	1		2	$\frac{1}{2}$	2
	2		0	2	2
ubiquitous	3		0	–	1
	4		1	$\frac{1}{2}$	2
	5		-1	–	1
used in C_3	6		1	$\frac{1}{2}$	4
	7		-1	$\frac{1}{2}$	2





[Faucett, JDT, Whiteson, PRD 2021;
using Komiske, Metodiev, JDT, JHEP 2018; C_3 from Larkoski, Salam, JDT, JHEP 2013]

[redacted]

Moments of Clarity

Alternative pooling operations for streamlined latent spaces

Combining per-particle features through multiplication and summation

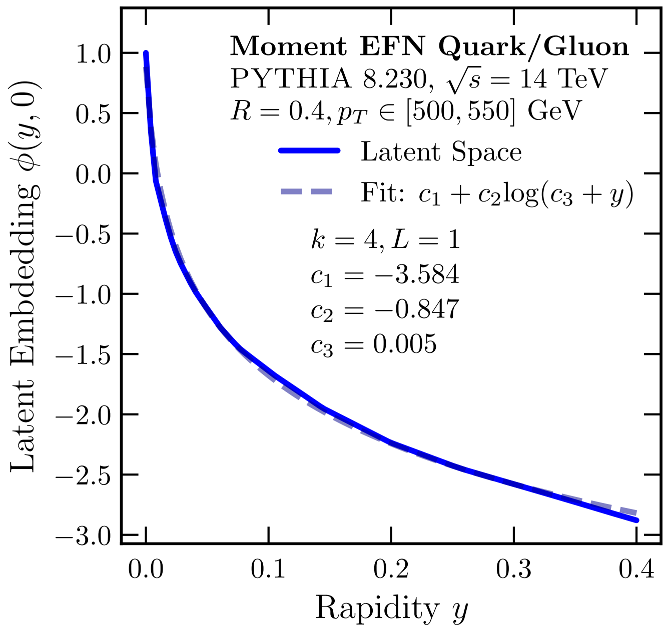
$$\mathcal{O}_k(\mathcal{P}) \equiv F\Big(\langle \Phi^{a} \rangle_{\mathcal{P}} \, , \langle \Phi^{a_1} \Phi^{a_2} \rangle_{\mathcal{P}} \, , \dots \langle \Phi^{a_1} \dots \Phi^{a_k} \rangle_{\mathcal{P}} \Big)$$

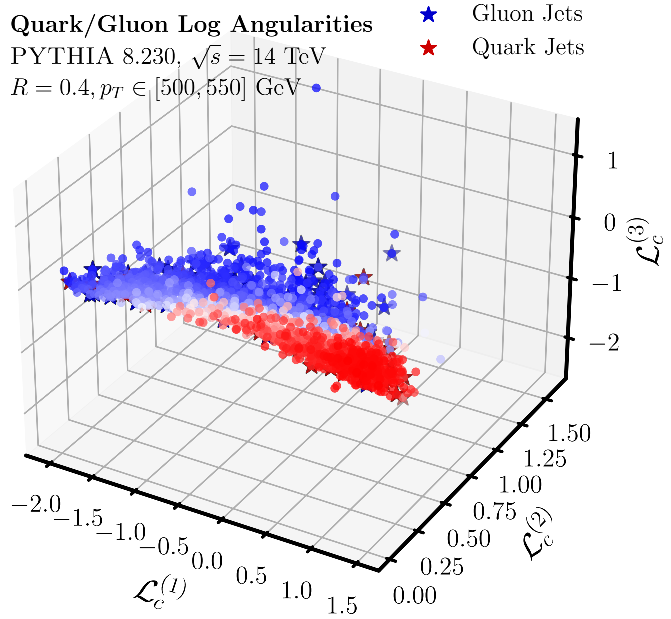
$$\sum_{i \in \mathcal{P}} z_i \Phi^{a}(x_i) \qquad\qquad \sum_{i \in \mathcal{P}} z_i \Phi^{a_1}(x_i) \Phi^{a_2}(x_i)$$

Sum Pooling (Deep Sets, EFN, k=1) | Moment Pooling (k = 2)

Same philosophy (and scaling) as Energy Flow Networks, just new permutation-invariant pooling operations

Single learned feature with k = 4 mimics four separate learned features





Log Angularity through Symbolic ReGression: $\Phi_{\mathcal{L}}(r) = c_1 + c_2 \log(c_3 + r)$

[Gambhir, Osathapan, JDT, arXiv 2024; building off Komiske, Metodiev, JDT, JHEP 2019; see also Cranmer, Kreisch, Pisani, Villaescusa-Navarro, Spergel, Ho, ICLR 2021 SimDL]

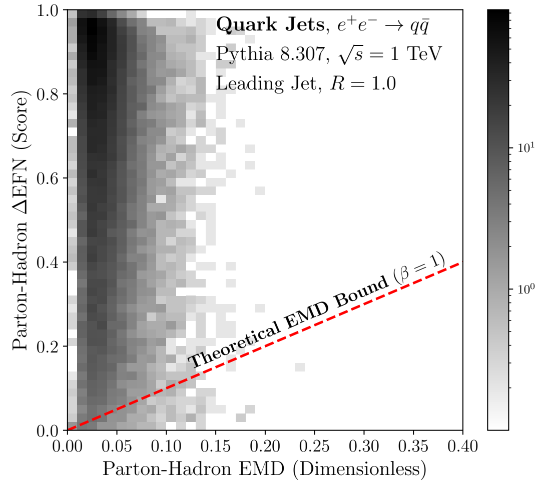
Safe but Incalculable

Formal IRC safety doesn't immediately ensure small non-perturbative corrections

Regularizing learned features to ensure controlled behavior of per-particle representations

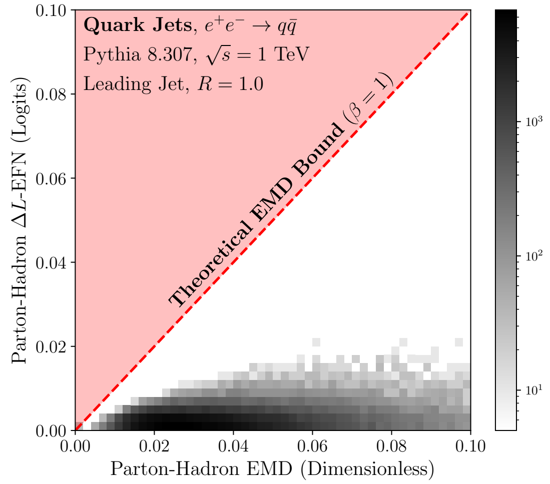
Energy Flow Networks

$$\mathrm{EFN}(\{p_1,\ldots,p_M\}) = F\left(\sum_{i=1}^{M} z_i \Phi(\hat{p}_i)\right)$$

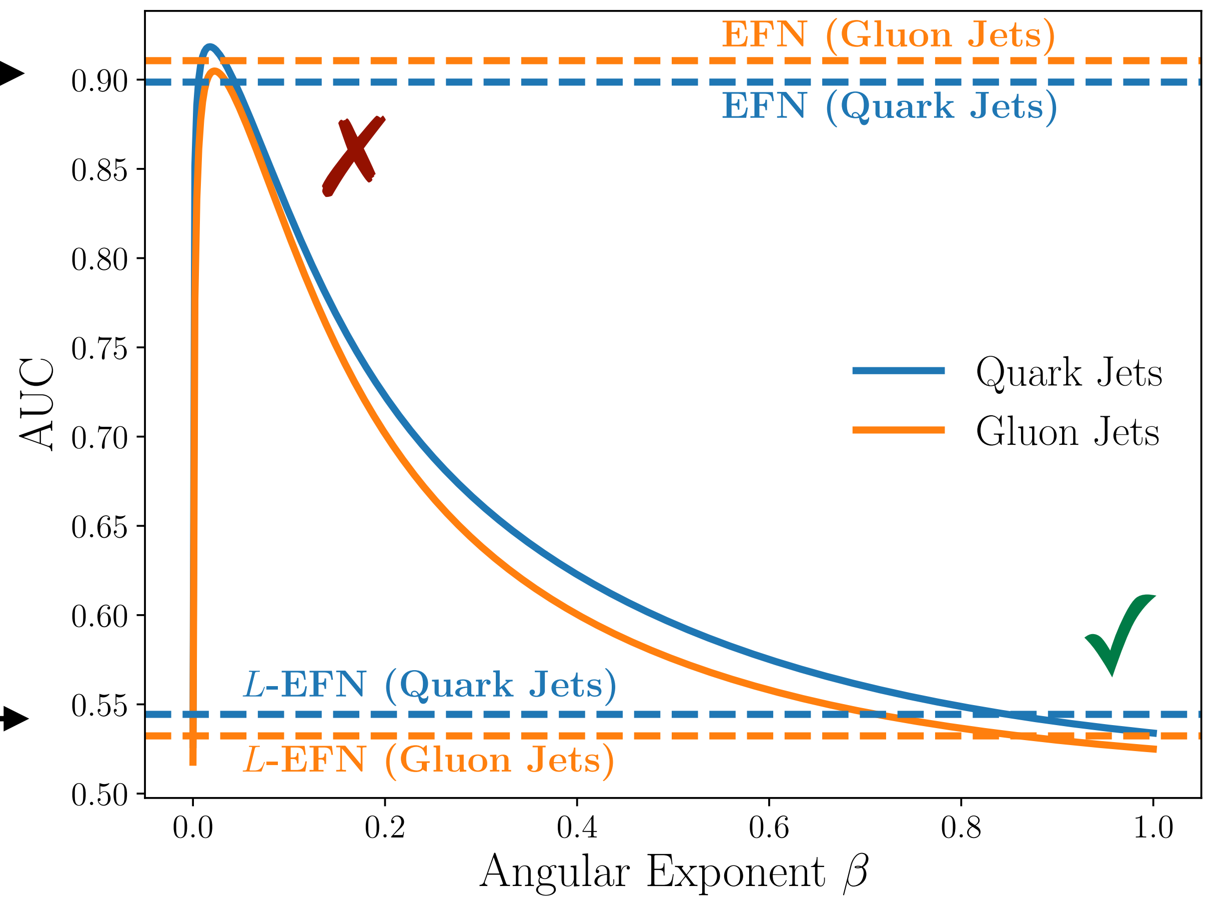


Lipschitz Energy Flow Networks

$$\|\Phi(\hat{p}_1) - \Phi(\hat{p}_2)\| \;\leq\; L\|\hat{p}_1 - \hat{p}_2\|$$



Parton vs. Hadron Sensitivity



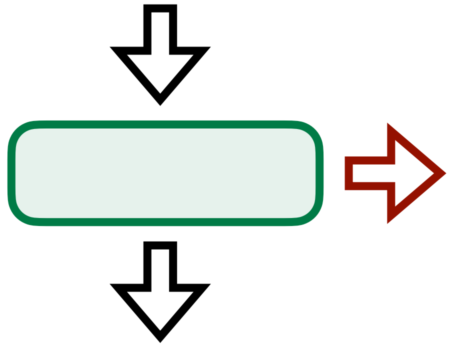
[Bright-Thonney, Nachman, JDT, arXiv 2023;
see also Komiske, Metodiev, JDT, PRL 2019; Kitouni, Nolte, Williams, MLST 2023]

Confronting the Black Box

*To benefit from machine learning advances, we must ensure that our algorithmic choices align with our **scientific goals***

Case Study in Jet Classification

*When possible, pursue **active interpretability**, where you control the network architecture and training paradigm*



The Next Frontier for Interpretability

*Foundation models identify **generically useful features**, which challenge the importance of task alignment*

From the Living Review of ML for Particle Physics

Fascinating categorization!

Uncertainty Quantification

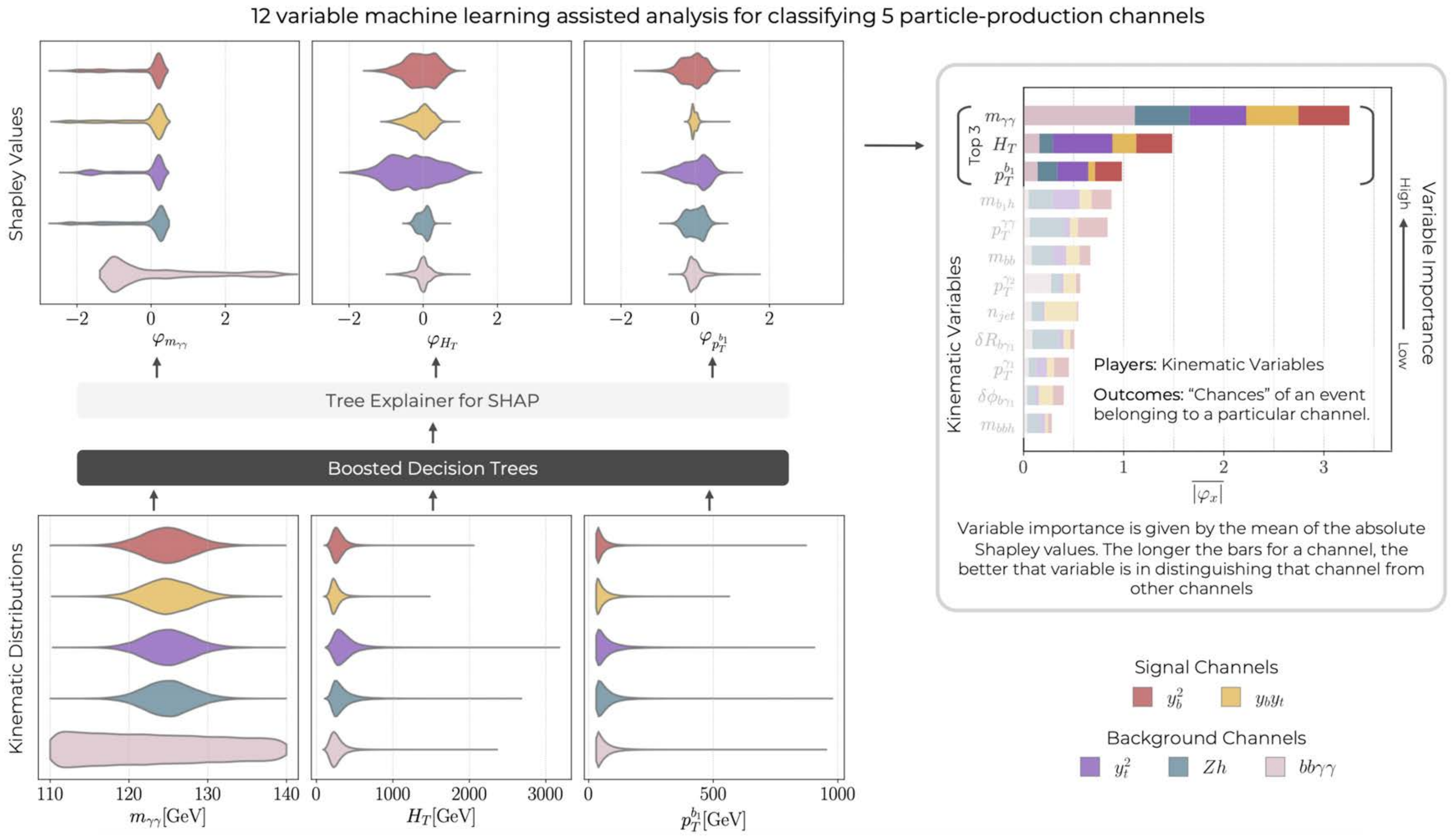
Interpretability

- Jet-images — deep learning edition [DOI]
- What is the Machine Learning? [DOI]
- CapsNets Continuing the Convolutional Quest [DOI]
- Explainable AI for ML jet taggers using expert variables and layerwise relevance propagation [DOI]
- Resurrecting $b\bar{b}h$ with kinematic shapes [DOI]
- Safety of Quark/Gluon Jet Classification
- An Exploration of Learnt Representations of W Jets
- Explaining machine-learned particle-flow reconstruction
- Creating Simple, Interpretable Anomaly Detectors for New Physics in Jet Substructure [DOI]
- Improving Parametric Neural Networks for High-Energy Physics (and Beyond) [DOI]
- Lessons on interpretable machine learning from particle physics [DOI]
- A Detailed Study of Interpretability of Deep Neural Network based Top Taggers [DOI]
- Interpretability of an Interaction Network for identifying $H \to b\bar{b}$ jets [DOI]
- Interpretable Machine Learning Methods Applied to Jet Background Subtraction in Heavy Ion Collisions [DOI]
- Interpretable deep learning models for the inference and classification of LHC data

Alternative answer to: “What is the goal of interpretable ML?”

[HEPML-LivingReview, moderated by Nachman, Feickert, Krause, Winterhalder]

E.g. SHapley Additive exPlanations (SHAP)



Goal:
Identify features driving decisions about classification

Quite similar to goal of identifying low-rank features

[Grojean, Paul, Qian, Strümke, Nature Reviews Physics 2022]

“…but what is the machine actually learning?”

To the extent that “interpretability” is about identifying features…

"…but what is the machine actually learning?"

To the extent that "interpretability" is about identifying features…

The Next Frontier: Foundation Models

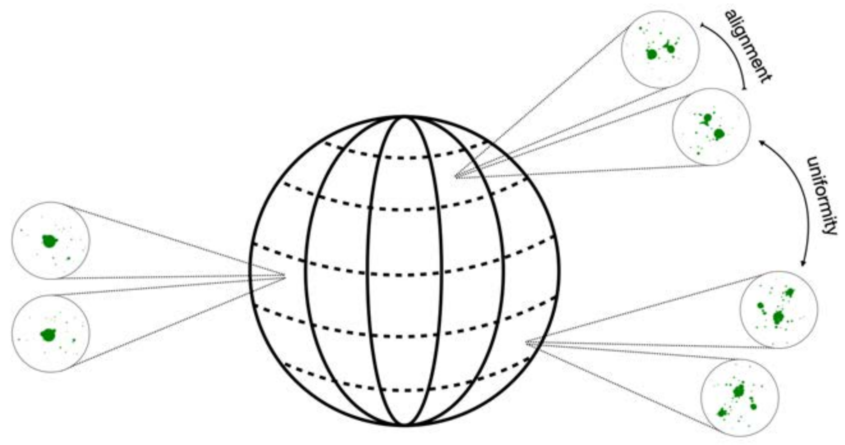
Identify features useful for generic tasks, which get reused for specialized applications

Purposeful misalignment between initial and downstream goals

Foundation Models for HEP

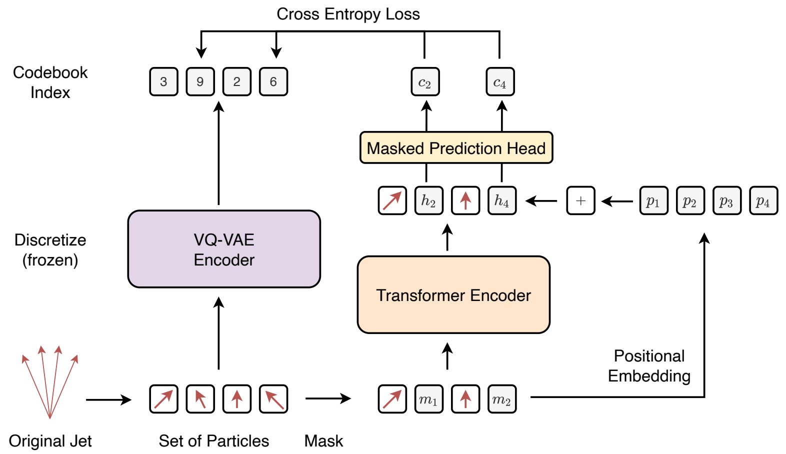
The natural evolution of transfer learning

Symmetry Augmentation



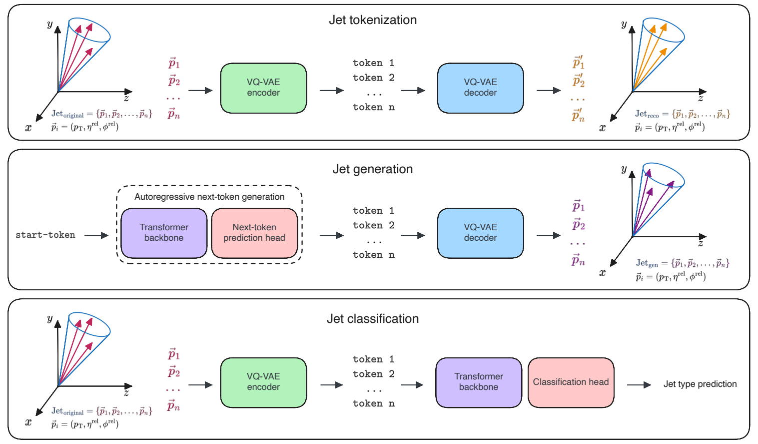
[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, SciPost 2021]

Masked Particle Modeling



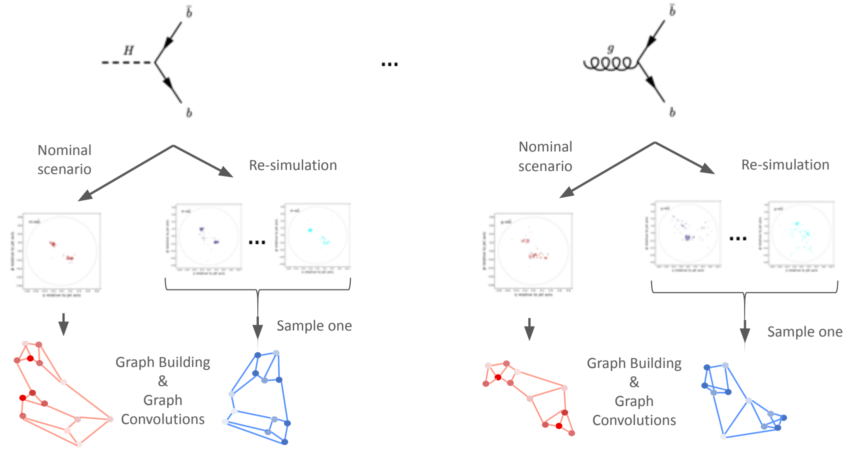
[Heinrich, Golling, Kagan, Klein, Leigh, Osadchy, Raine, arXiv 2024]

Next Token Prediction



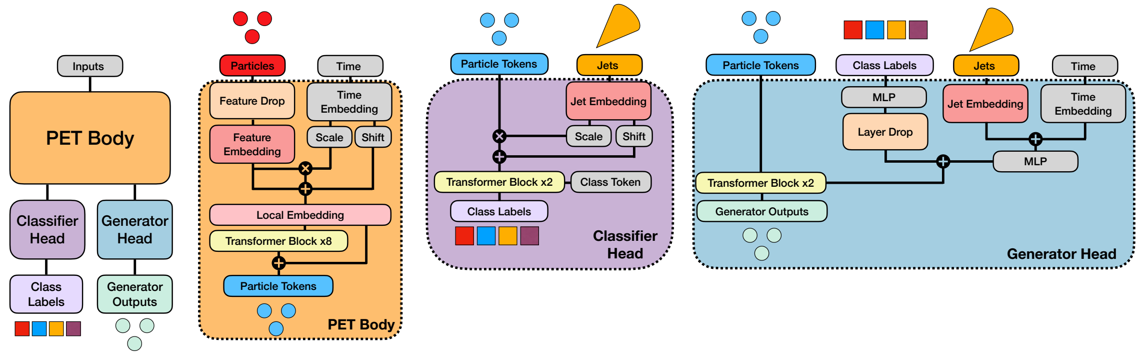
[Birk, Hallin, Kasieczka, arXiv 2024]

Re-Simulation Similarity



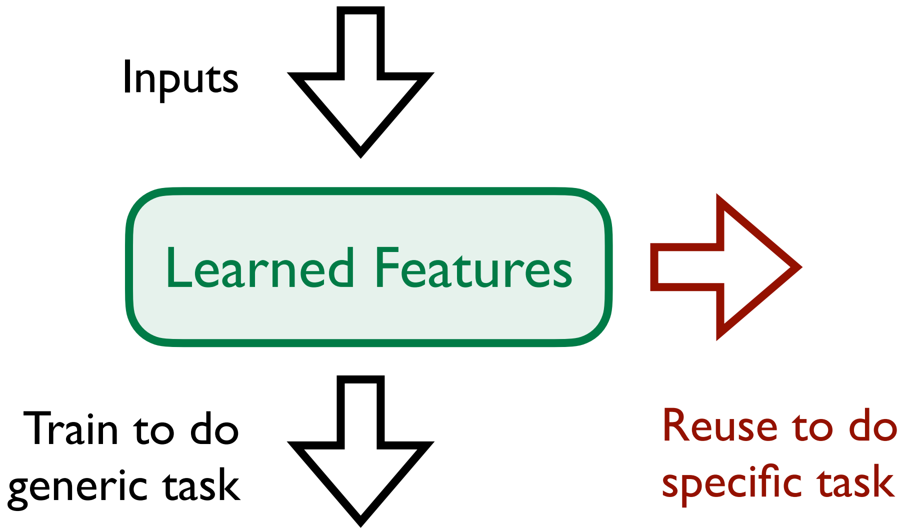
[Harris, Kagan, Krupa, Maier, Woodward, arXiv 2024]

Multi-Category Classification



[Mikuni, Nachman, arXiv 2024]

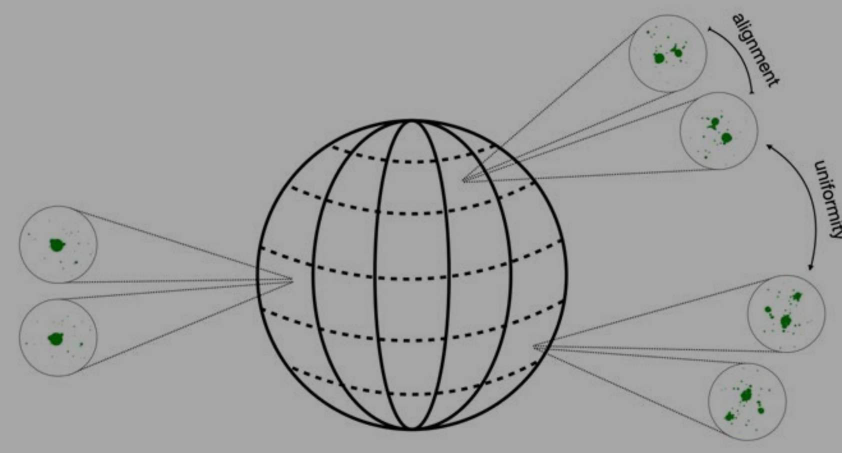
Your Next Paper



Foundation Models for HEP

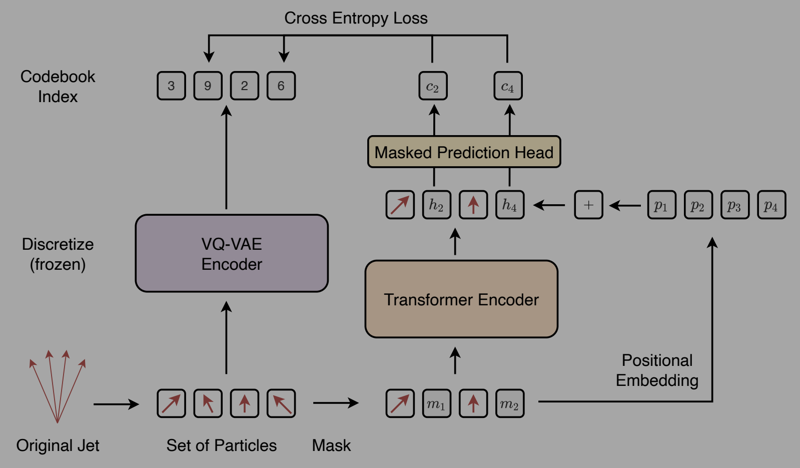
The natural evolution of transfer learning

Symmetry Augmentation



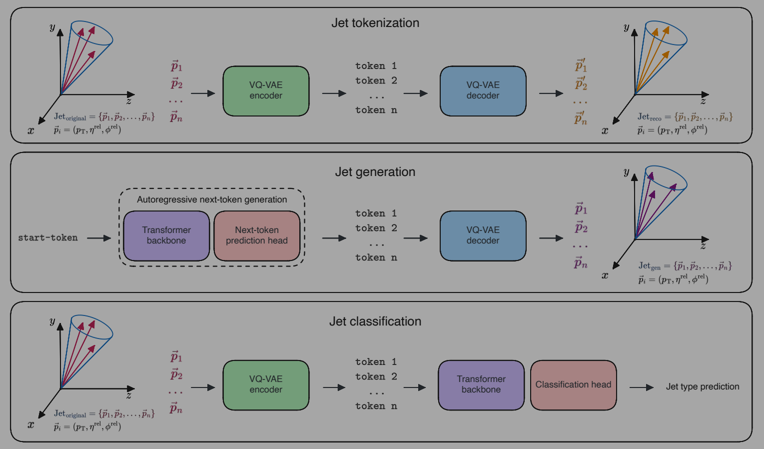
[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, SciPost 2021]

Masked Particle Modeling



[Heinrich, Golling, Kagan, Klein, Leigh, Osadchy, Raine, arXiv 2024]

Next Token Prediction

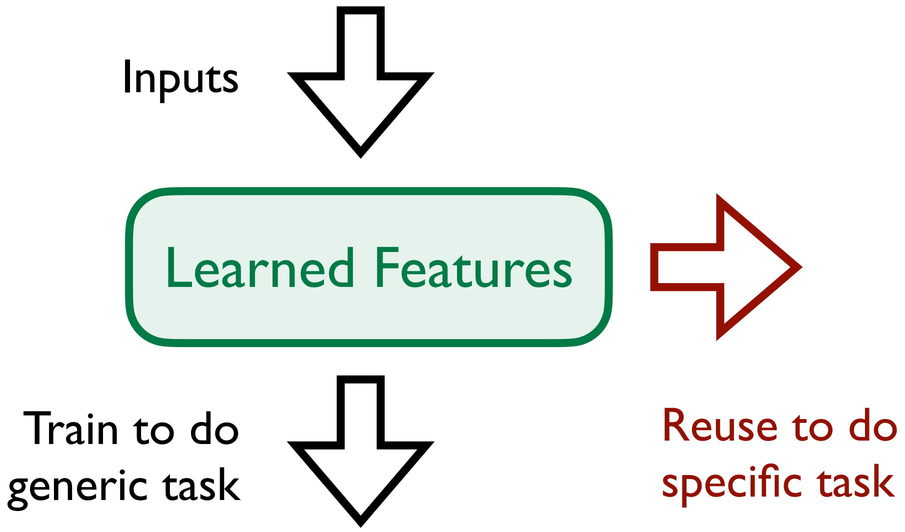


[Birk, Hallin, Kasieczka, arXiv 2024]

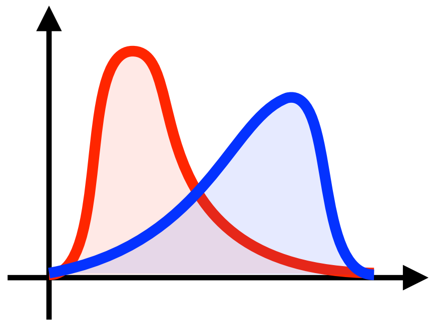
Asymptotically, pre-training cannot yield improved performance, but very effective in practice

"What is the machine learning?!"

Your Next Paper

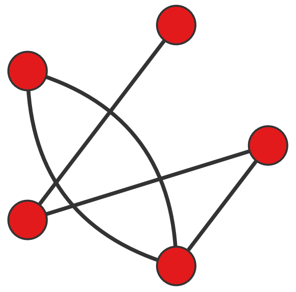


Interpretable Machine Learning for Particles Physics



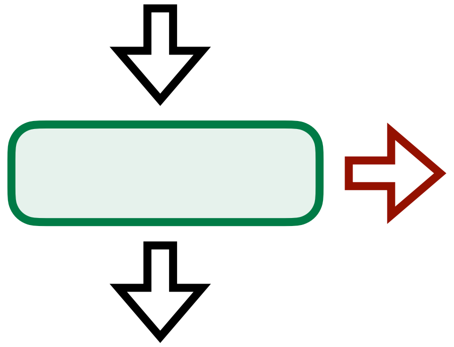
Confronting the Black Box

To benefit from machine learning advances, we must ensure that our algorithmic choices align with our ***scientific goals***



Case Study in Jet Classification

When possible, pursue ***active interpretability****, where you control the network architecture and training paradigm*



The Next Frontier for Interpretability

Foundation models identify ***generically useful features****, which challenge the importance of task alignment*

The NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI /aɪ-faɪ/ iaifi.org)

Artificial intelligence as a pathway to scientific insight

Physics intelligence as a pathway to AI innovation

Progress driven by early career talent with interdisciplinary expertise

Consider applying to IAIFI Postdoctoral Fellowship this Fall!