Interpretable Machine Learning for Particle Physics

Jesse Thaler

PHYSTAT: Statistics meets Machine Learning, Imperial College London — September 10, 2024

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

Launched August 2020

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

Empowering the Next Generation of AI + Physics Talent

IAIFI Postdoctoral Fellows

Application deadline: *October 9, 2024*

Albergo	Boyda	Bright-Thonnney	Cuesta	Dogra	Gagliano	Golubeva	Grosso	Harvey	Luo	Micallef	Mishra-Sharma	Yang
												
AI and Statistical Physics	⇓	AI for Particle Physics	AI for Cosmological Observations	Mathematical Physics of AI	AI for Time-Domain Astronomy	⇓	AI for Collider Physics	AI for String Theory	⇓ UCLA	AI for Neutrino Physics	⇓ BU	AI Frontiers of Reinforcement Learning

IAIFI Summer School & Workshop

IAIFI

Summer School
August 5–August 9
2024

Summer Workshop: August 12–16, 2024

PHYSTAT Workshop Theme: Interpretability

Recalling the PHYSTAT emphasis on the statistical issues involved

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

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

Tuesday Poster Session:

Integrating Explainable AI in Modern High-Energy Physics (the MUCCA Project) — *Joseph Carmignani*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 18:06 - 18:07

Learning Optimal and Interpretable Summary Statistics of Galaxy Catalogs with SBI — *Kai Lehman*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 18:17 - 18:18

Let me know if I missed your poster!

Thursday Morning Talks:

Sorry I won't be there for the discussion!

Interpretability — *Mikael Kuusela*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 10:45 - 11:15

pop-cosmos: an interpretable generative model for the galaxy population over cosmic time — *Hiranya Peiris*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 11:15 - 11:45

Identifying Tau Neutrinos in IceCube — *Philipp Eller*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 11:45 - 12:10

Conditional generation — *Tobias Golling*
Lecture Theatre 2, Blackett Laboratory, Imperial College London — 12:10 - 12:35

Obligatory apology that examples in this talk are heavily drawn from my own research in collider physics

PHYSTAT Workshop Theme: Interpretability

Recalling the PHYSTAT emphasis on the statistical issues involved

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

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

My evolving perspective:

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

We should strive towards **actionable goals** for interpretability, e.g.:

1. Qualitatively assess sources of **systematic uncertainties**
2. Identify **low-rank structures** in high-dimensional datasets

Interpretability Discussion Prompts from Indico (1 of 2)

ChatGPT 4o: "Draw a picture related to this prompt"

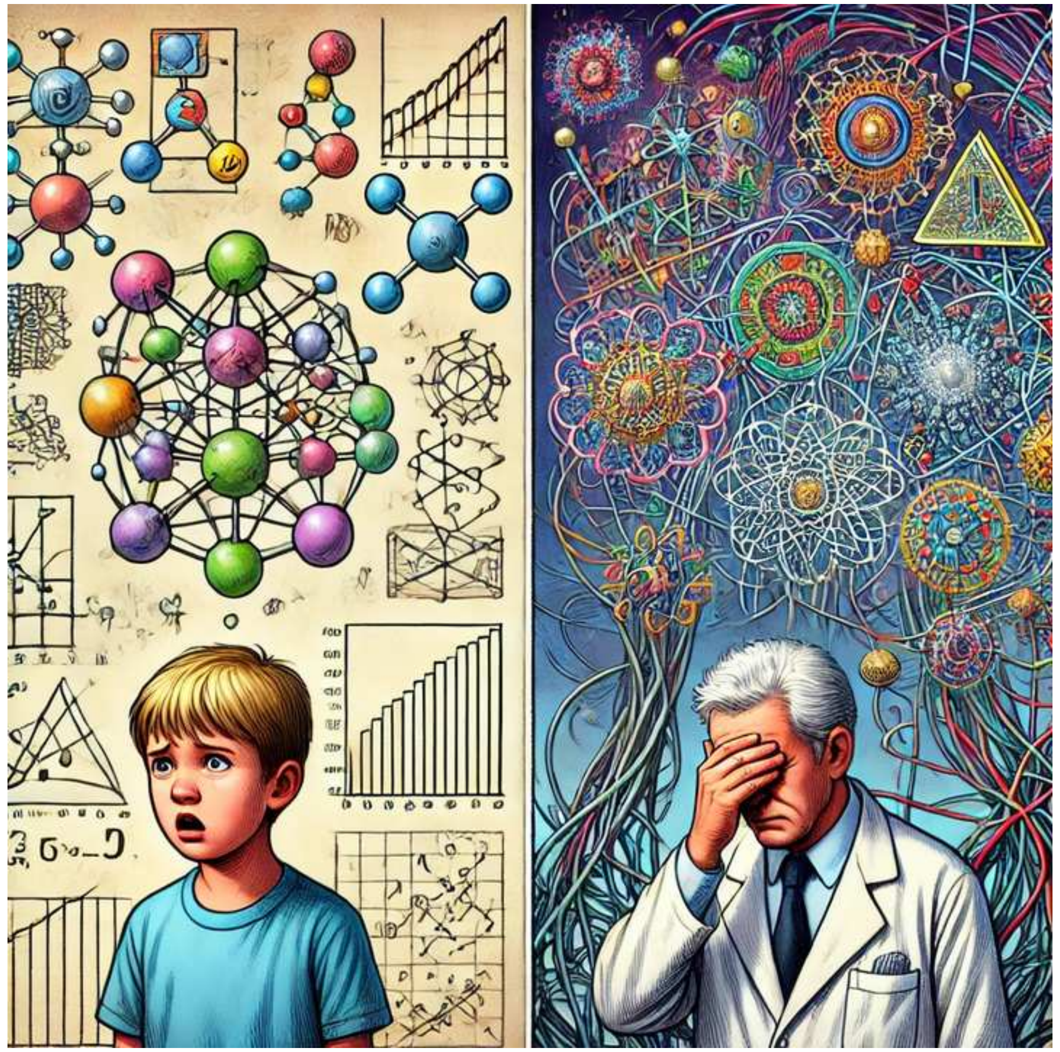
There is the probably apocryphal story of a ML classifier learning to distinguish cats from dogs because in the training sample, all the cats were photographed curled up on living room couches, while the dogs were running outdoors in fields.

How do we ensure that the distinction between, say, signal and background is based on **significant features in the data**, rather than on the particular way that soft particles are simulated?

Can interpretability **help us diagnose** this?

Actionable Goal: Qualitatively Assess Sources of **Systematic Uncertainties**

Interpretability Discussion Prompts from Indico (2 of 2)



ChatGPT 4o: "Draw a picture related to this prompt"

Just as we would not expect a 10-year-old to understand how a single hidden layer NN works, why should a very sophisticated ML procedure be interpretable by a mere human Physicist?

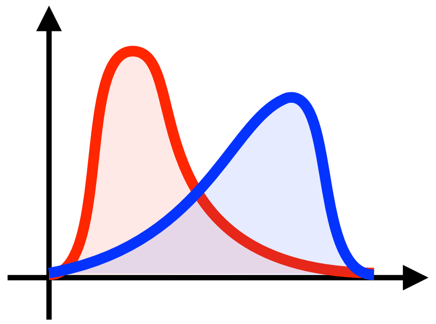
Is it important for our methods to be interpretable, or it is enough just to **check out their properties**?

Is interpretability becoming an **unrealistic goal**?

*(And **what is the point** of fundamental physics anyways?)*

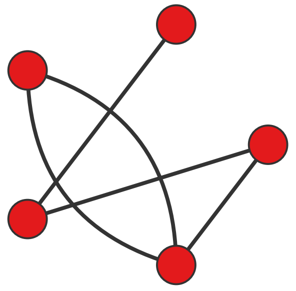
Actionable Goal: Identify **Low-Rank Structures** in High-Dimensional Datasets

Interpretable Machine Learning for Particles Physics



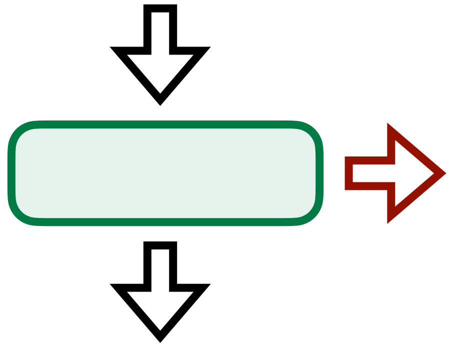
Confronting the Black Box

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



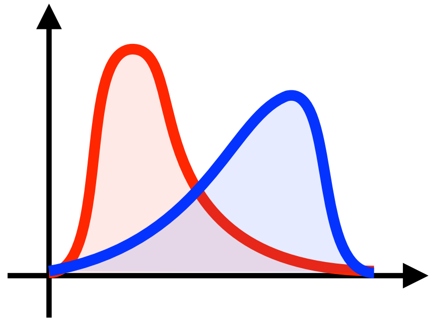
Case Study in Jet Classification

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*

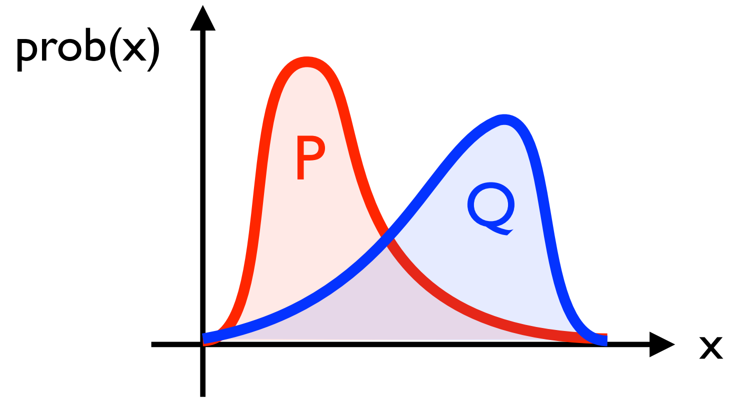
The Next Frontier for Interpretability

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

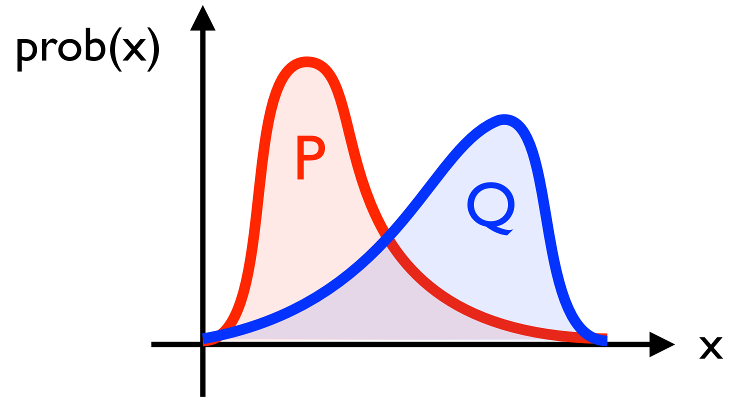
Key example of simulation-based inference

Many HEP problems can be expressed in this form!

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

Training Data: Finite samples P and Q

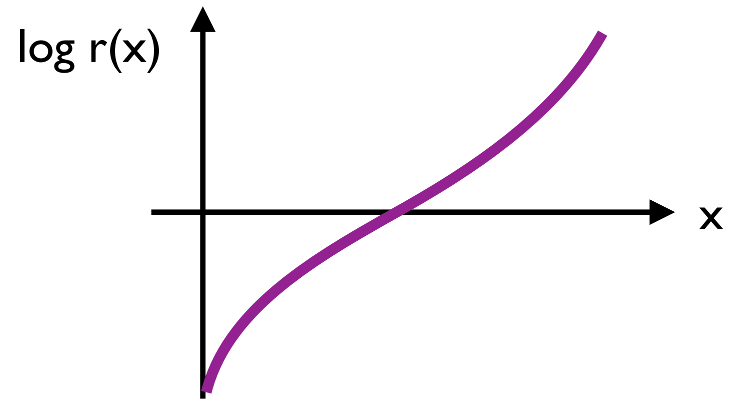
Learnable Function: f(x) parametrized by, e.g., neural networks



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

Asymptotically: $$\arg\min_{f(x)} L = \frac{p(x)}{q(x)}$$ *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

Many HEP problems can be expressed in this form!

Key example of simulation-based inference

Asymptotically, same structure as Lagrangian mechanics!

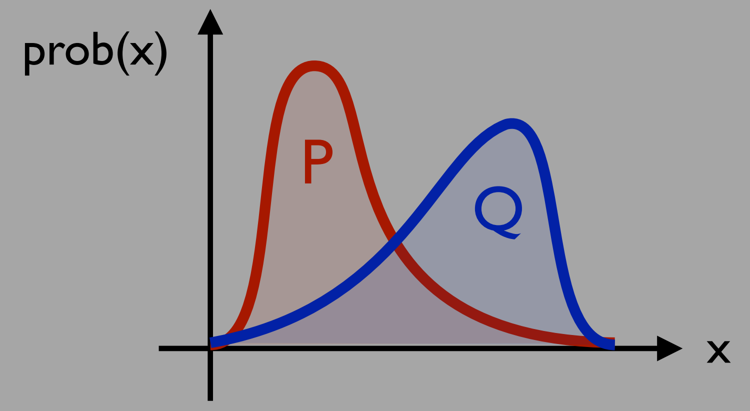
Action: $$L = \int \mathrm{d}x \, \mathcal{L}(x)$$

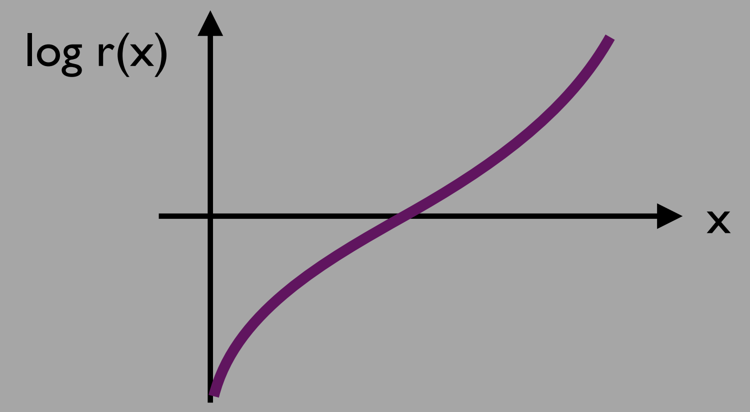
Lagrangian: $$\mathcal{L}(x) = -p(x) \log f(x) + q(x) \big(f(x) - 1 \big)$$

Euler-Lagrange: $$\frac{\partial \mathcal{L}}{\partial f} = 0$$

Solution: $$f(x) = \frac{p(x)}{q(x)}$$

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:

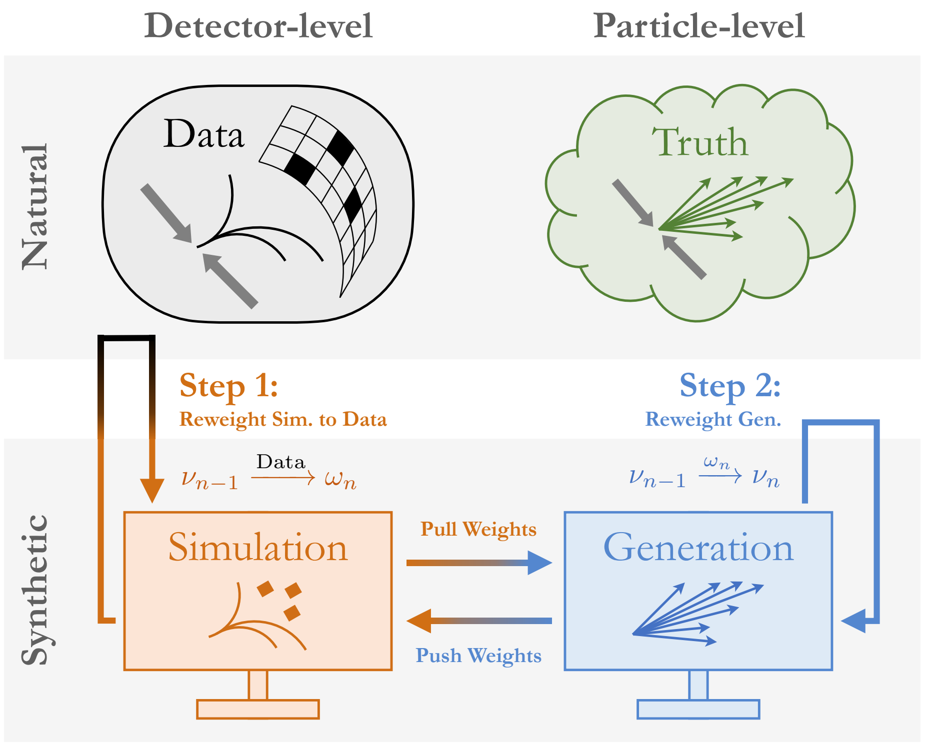
Skeptical of algorithmic/statistical/computational reasoning
Need to explain decisions to external **stakeholders**
Desire to **manage risks** from unforeseen outcomes
…

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

Likelihood Ratio Trick in HEP

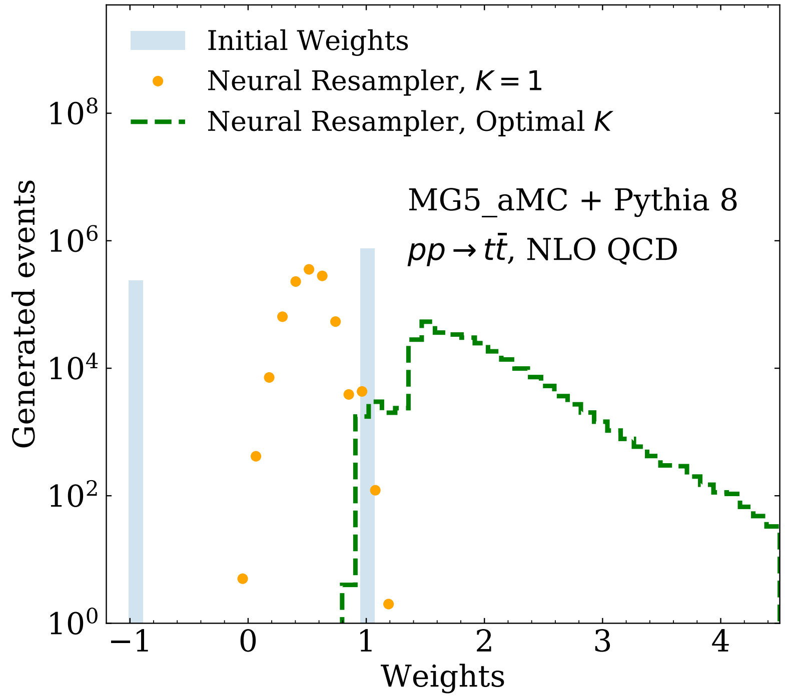
Apologies that examples are all from my own work

Detector Unfolding



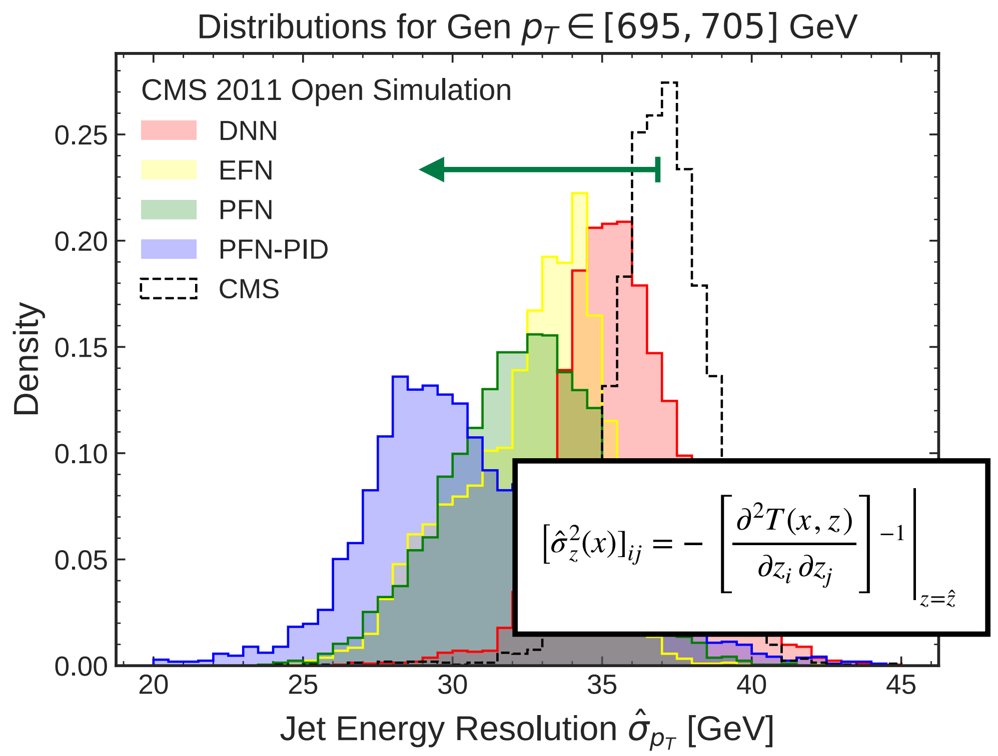
[Andreassen, Komiske, Metodiev, Nachman, JDT, PRL 2020; + Suresh, ICLR SimDL 2021]

Monte Carlo Reweighting



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

Resolution Estimation



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

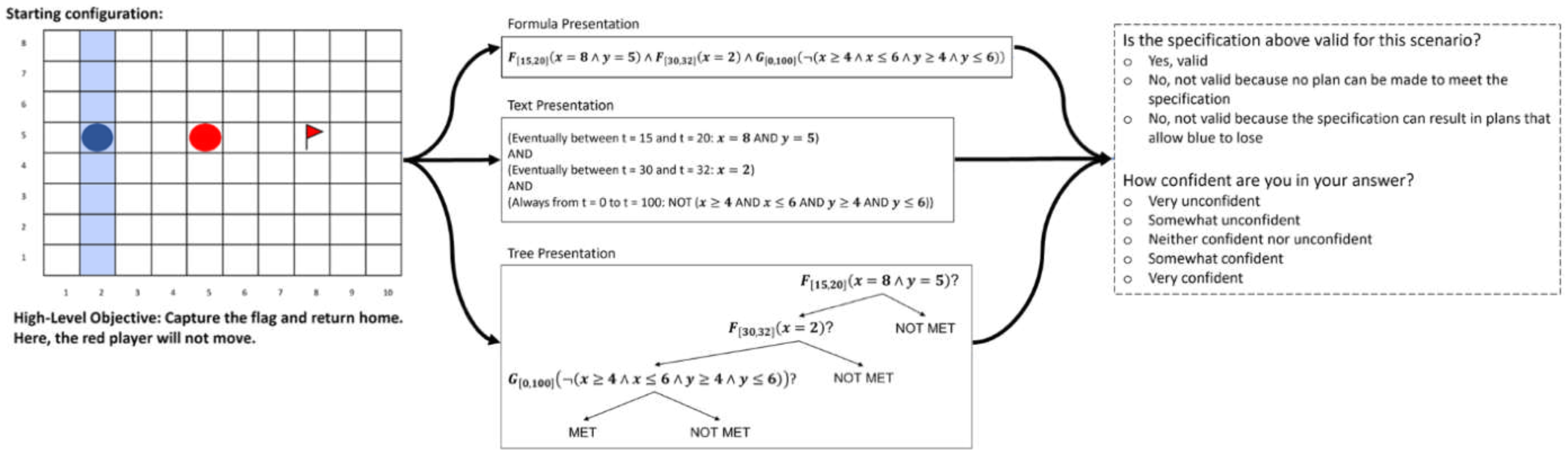
For these applications, goal is "accuracy" more than "interpretability"

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

Are “Formal Specifications” Human Interpretable?

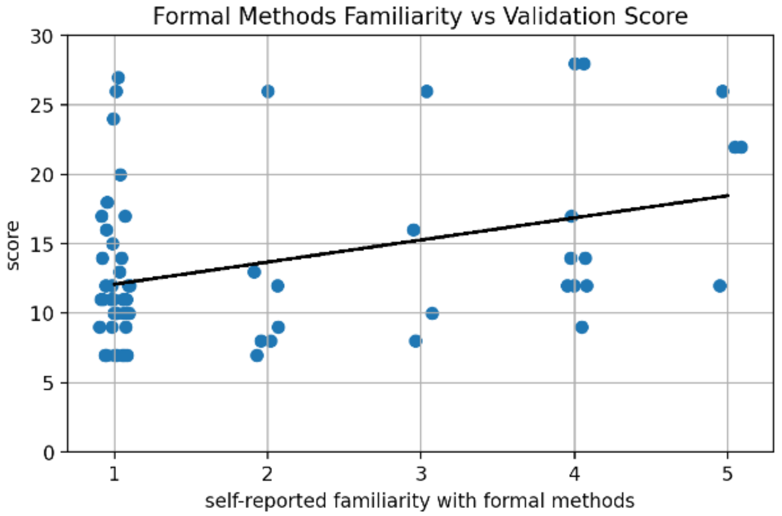
Case study in planning with signal temporal logic

Capture the Flag:



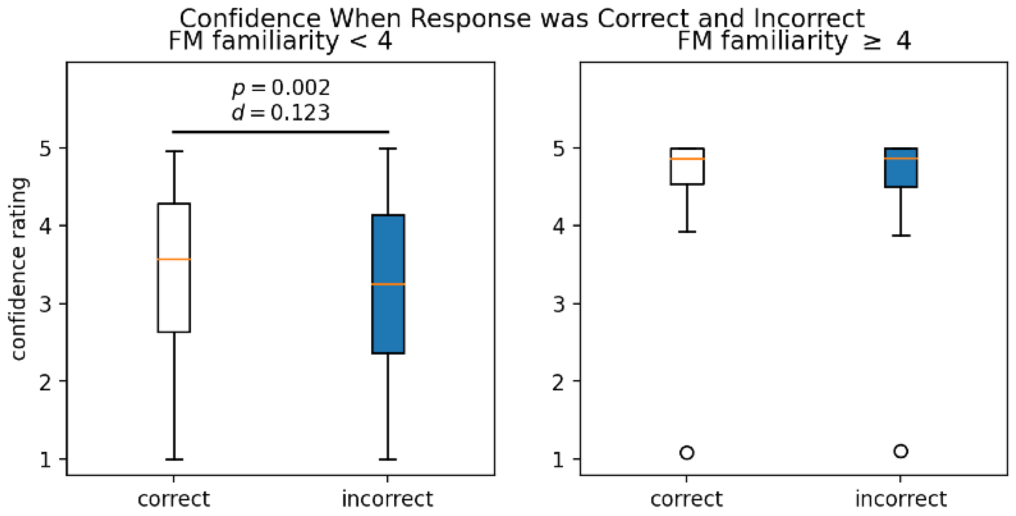
“Is this a valid solution?”

“Are you confident?”



Humans are relatively incompetent…

…and experts are overconfident



Lesson: Formal validity is a distinct goal from human verifiability

[Siu, Leahy, Mann, SERC 2024; h/t Dave Kaiser]

The HEP Definition of "Interpretability"?

Categorization from Living Review of ML for Particle Physics

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 [DOI]
- Statistical divergences in high-dimensional hypothesis testing and a modern technique for estimating them
- Interpretable machine learning approach for electron antineutrino selection in a large liquid scintillator detector
- Explainable AI classification for parton density theory

Would authors of these papers agree that this is a goal of their methods?

Do these methods provide quantitative or qualitative assessment of uncertainties?

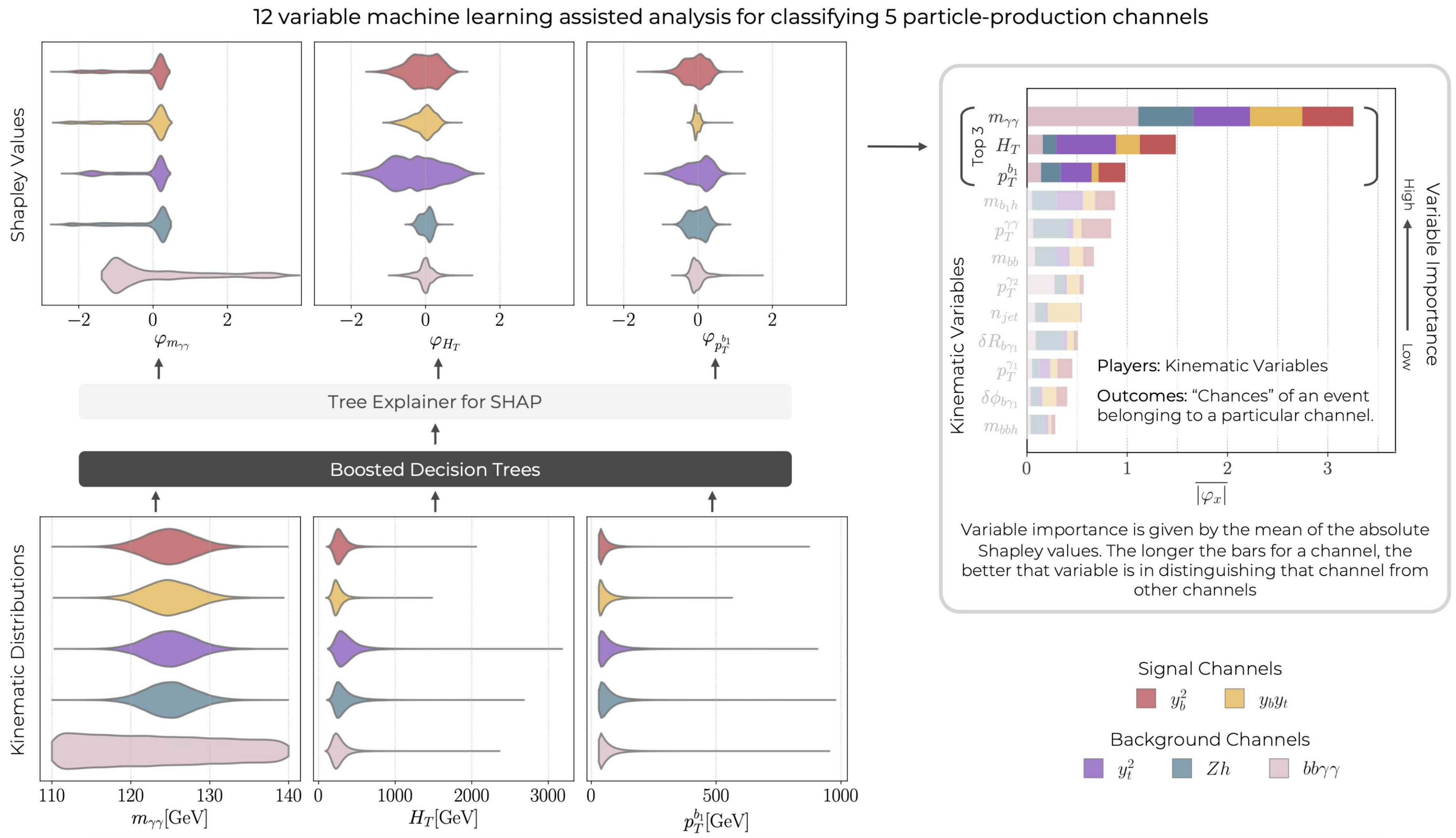
[HEPML-LivingReview, moderated by Nachman, Feickert, Krause, Winterhalder]

For fundamental physics, what actionable goals do we want to achieve through interpretability…?

…and are those goals statistically sound?

Interpretability as Uncertainty Quantification

E.g. SHapley Additive exPlanations (SHAP)



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

Explicit Goal:

Identify features driving decisions about classification

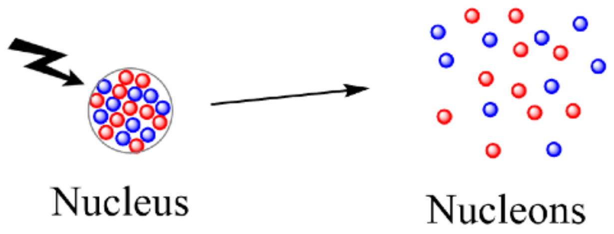
Implicit Goal:

Verify that these features are physically relevant

Actionable Goal: Qualitatively Assess Sources of Systematic Uncertainties

Interpretability as Knowledge Distillation

E.g. modeling nuclear binding energies



Symbolic Regression

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	$3.02\exp\left(-1.91P^{1.94}\left(\frac{Z}{N}\right)^{A^{2/3}} - 0.895e^{-0.227N}N^2 - 0.0268(N-1)^2\right)$

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

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

Cf. Semi-Empirical Mass Formula

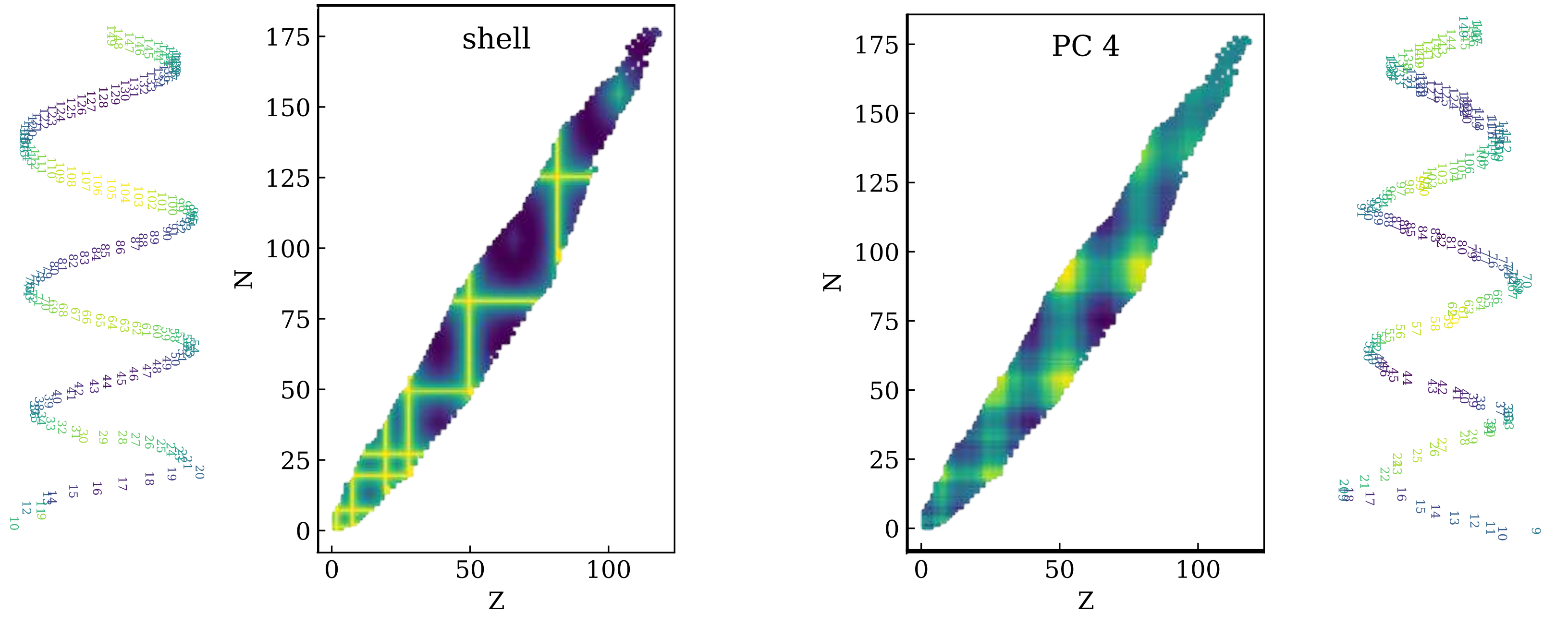
[Weizsäcker, 1935]

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

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

Latent Space Topography

Human vs. Machine

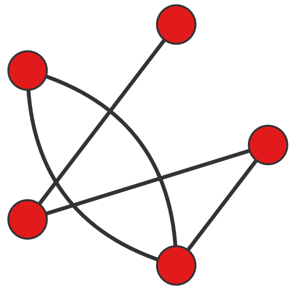


[Kitouni, Nolte, Trifinopoulos, Kantamneni, Williams, ICML 2023;
[Kitouni, Nolte, Pérez-Díaz, Trifinopoulos, Williams, ICML 2024]

Actionable Goal: Identify **Low-Rank Structures** in High-Dimensional Datasets

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

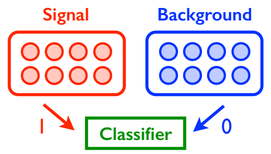
*Foundation models identify **generically useful features**, which challenge the importance of task alignment*

The More Things Change…

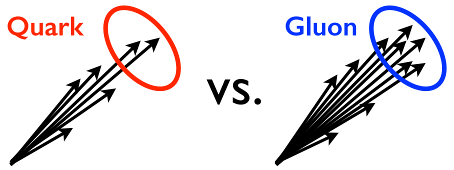
Jet classification, from talks I was giving in 2019

Application of Likelihood Ratio Trick

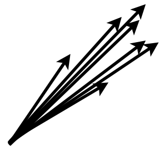
Binary Classification



e.g.



assuming trustable training data

Find $h\Big($  $\Big)$ such that $h(\text{Quark}) = 1$, $h(\text{Gluon}) = 0$

Best you can do: $h(\mathcal{J}) = \dfrac{p(\mathcal{J}|\mathrm{Q})}{p(\mathcal{J}|\mathrm{Q}) + p(\mathcal{J}|\mathrm{G})}$

(Neyman-Pearson lemma)

Jesse Thaler — Deep Learning (and Deep Thinking) in Collider Physics 17

Interpretability in Machine Learning

Introducing Energy Flow Networks

An architecture designed for interpretability

(see backup for detailed architecture)

Latent space of dim ℓ

Linear weights

$$S(\mathcal{J}) = F(V_1, V_2, \ldots, V_\ell) \qquad V_a(\mathcal{J}) = \sum_{i\in\mathcal{J}} p_{Ti}\, \Phi_a(y_i, \phi_i)$$

Parametrized with Neural Networks

Flexible enough to describe any* IRC-safe observable

(assuming large enough ℓ)

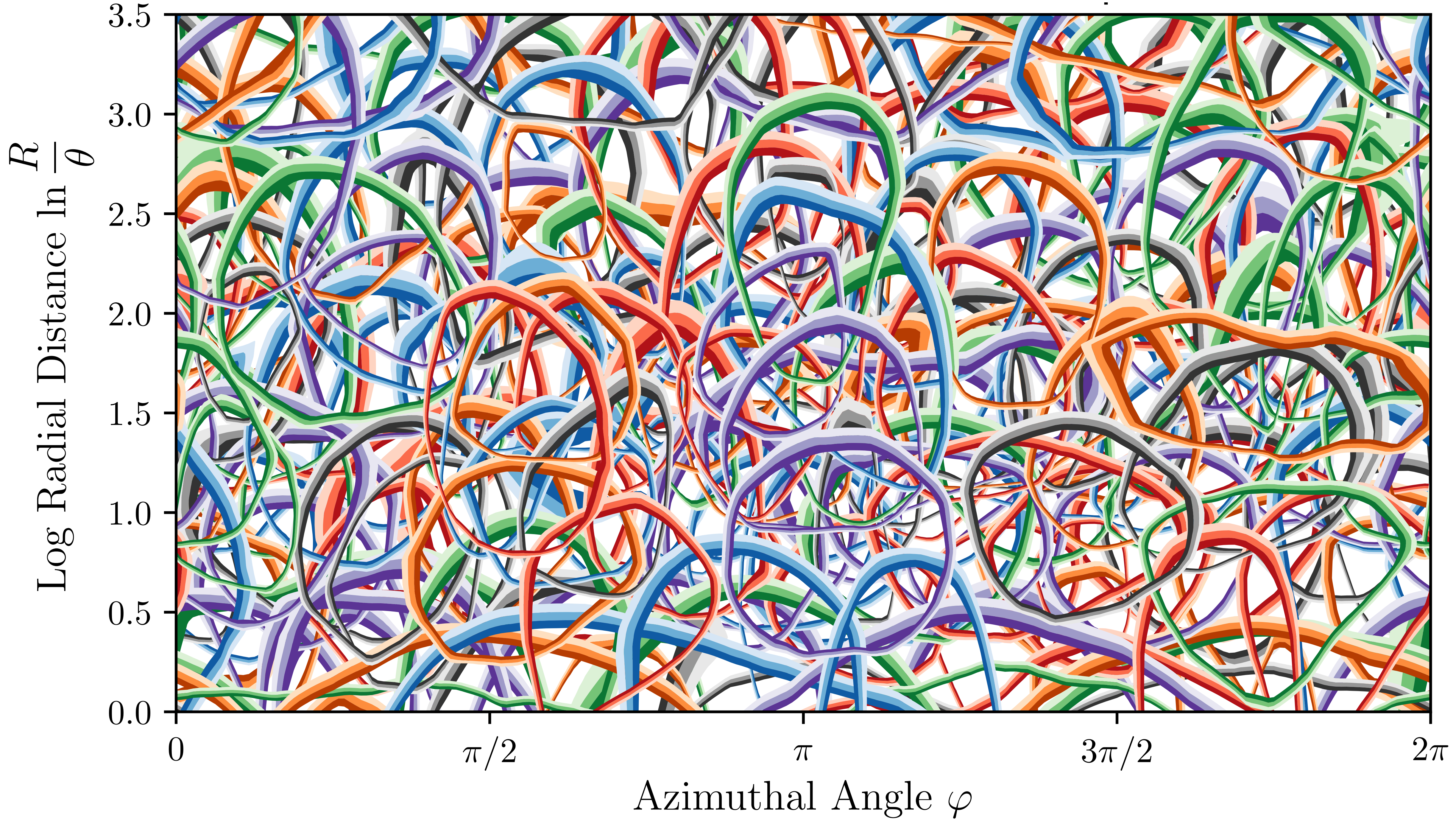
Generalization: Particle Flow Networks (aka "Deep Sets")

[Komiske, Metodiev, JDT, 1810.05165; special case of Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, Smola, 1703.06114]

Jesse Thaler — Deep Learning (and Deep Thinking) in Collider Physics 26

Does this Really Count as “Interpretable”?

Visualizing Energy Flow Networks



Trying to plot 256 dimensional latent space

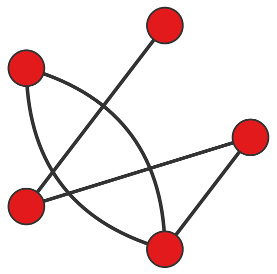
See paper for genuine insights at L = 2

[Komiske, Metodiev, JDT, JHEP 2019]

Three Lessons since 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

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

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

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

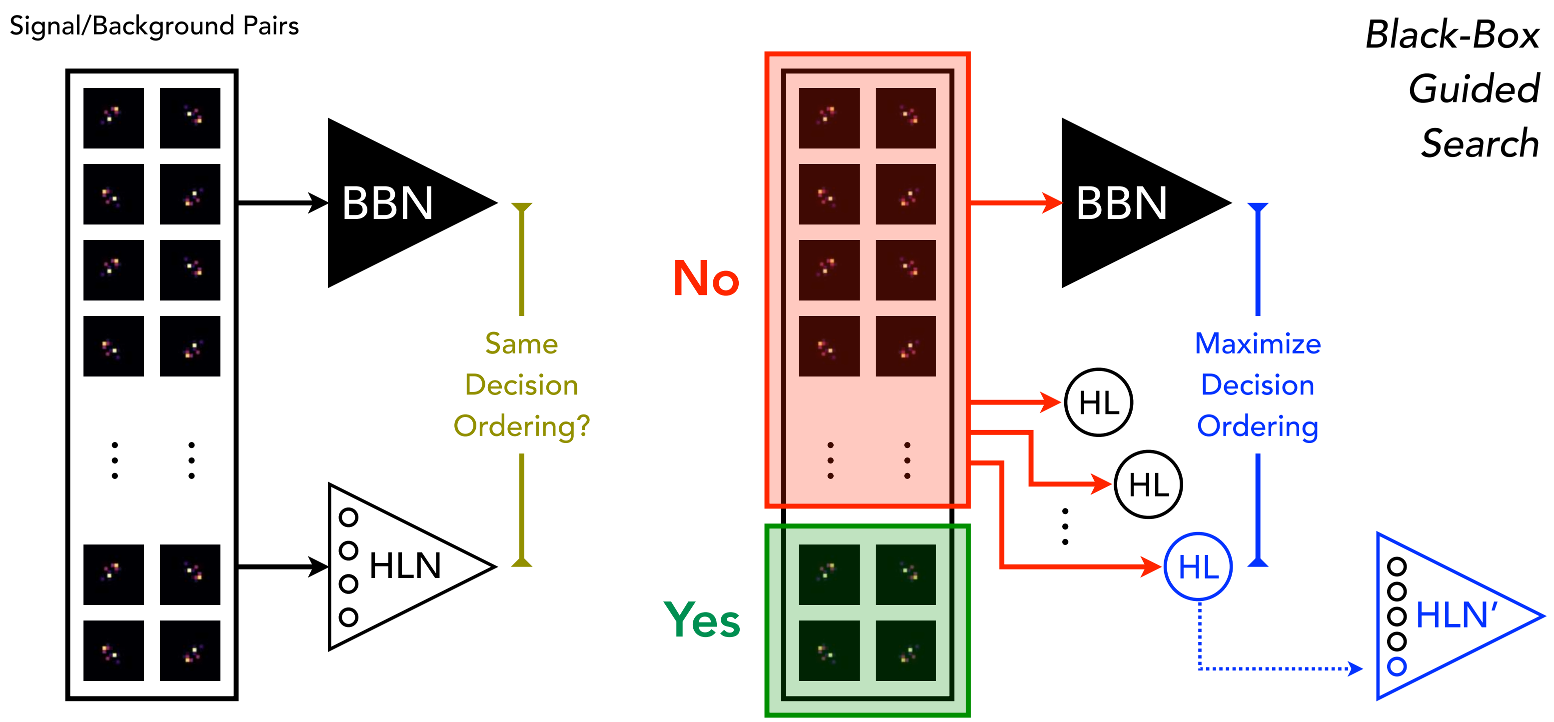
If there is a property you want your network to have, make sure to impose **algorithmic guardrails**, otherwise the machine might pursue undesirable optimization

[n.b.: According to HEPML-LivingReview, these papers are categorized respectively as “feature ranking”, “point clouds”, and “equivariance”]

Translating the Black Box

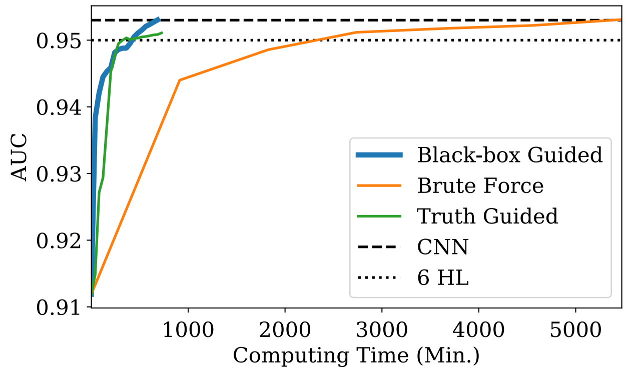
Selecting Energy Flow Polynomials that mimic CNN decisions

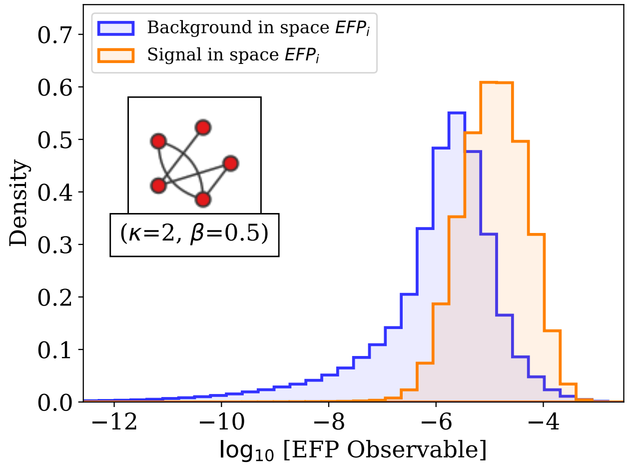
Iteratively building likelihood ratio estimate from catalog of high-level observables



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]

Moments of Clarity

Alternative pooling operations for streamlined latent spaces

Combining per-particle features through multiplication and summation

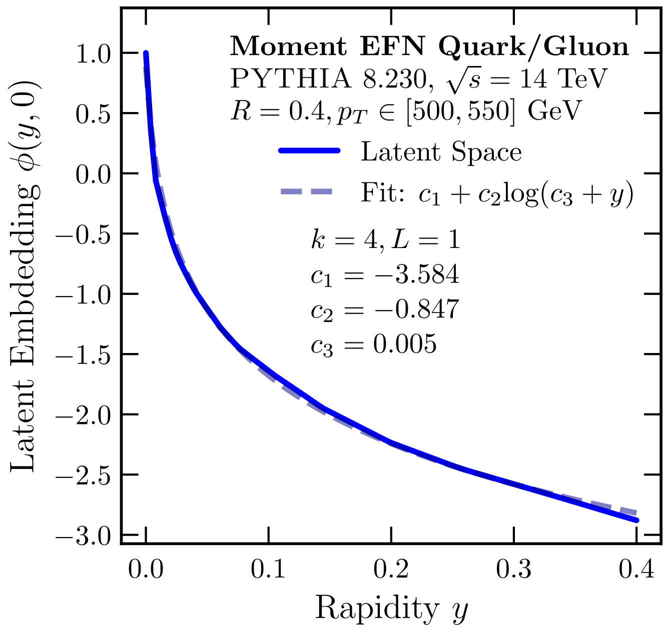
$$\mathcal{O}_k(\mathcal{P}) \equiv F\left(\langle \Phi^a \rangle_\mathcal{P} \, , \langle \Phi^{a_1} \Phi^{a_2} \rangle_\mathcal{P} \, , \ldots \langle \Phi^{a_1} \ldots \Phi^{a_k} \rangle_\mathcal{P} \right)$$

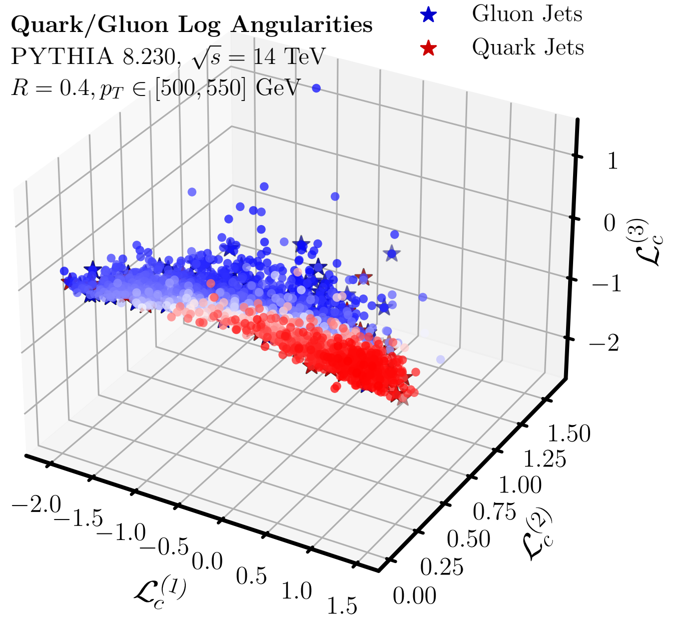
$$\sum_{i \in \mathcal{P}} z_i \Phi^a(x_i) \qquad \sum_{i \in \mathcal{P}} z_i \Phi^{a_1}(x_i) \Phi^{a_2}(x_i)$$

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]

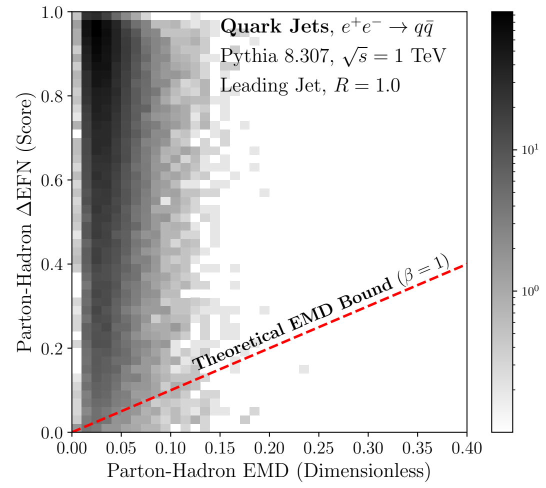
Safe but Incalculable

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

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

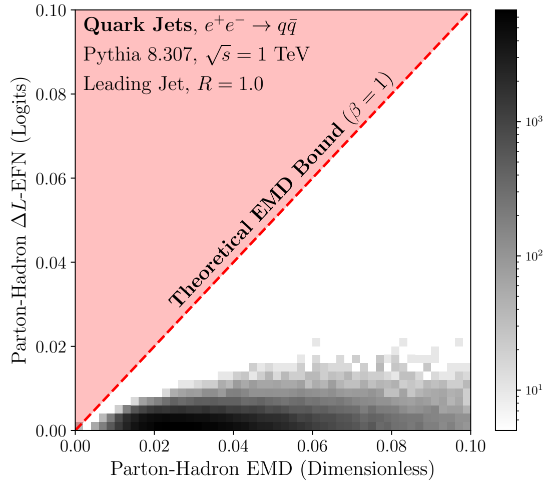
Energy Flow Networks

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

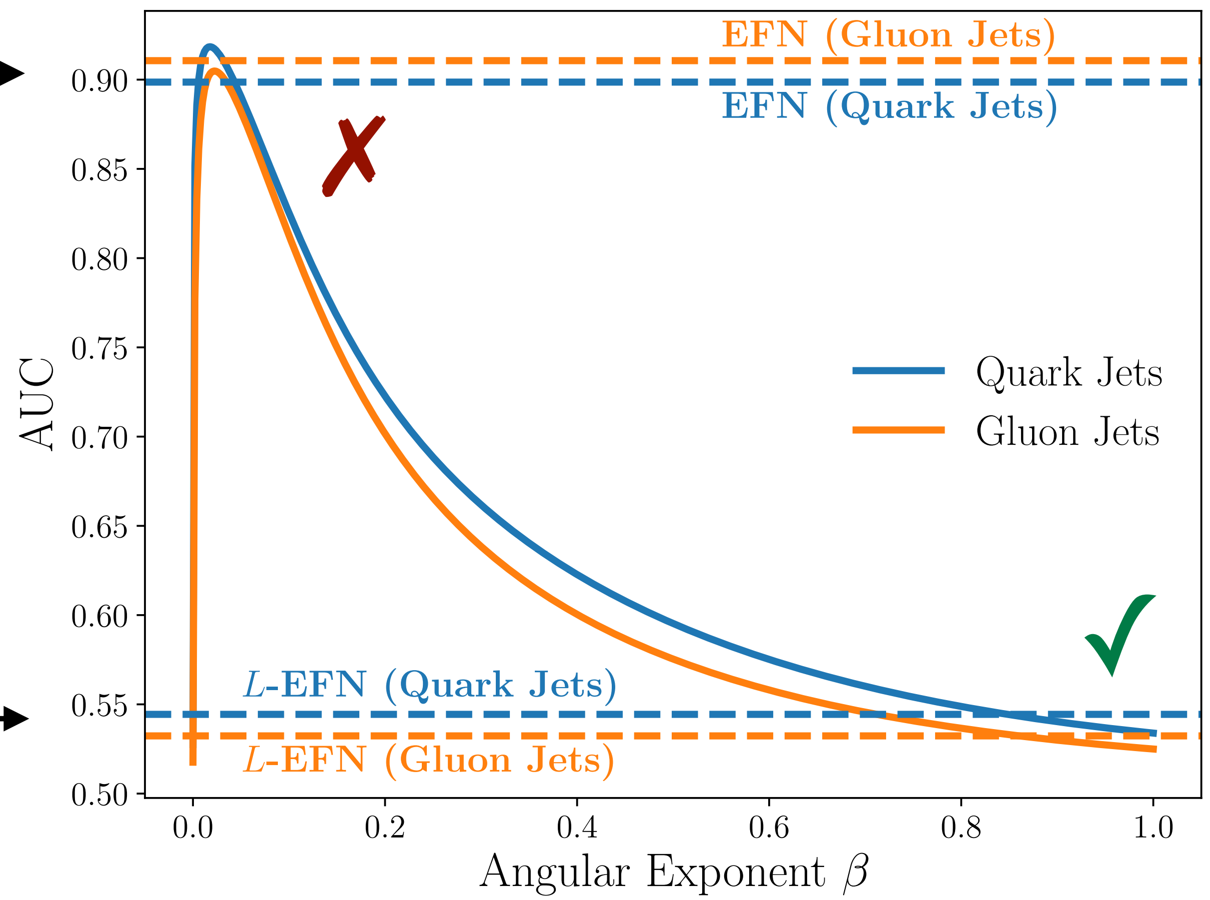


Lipschitz Energy Flow Networks

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



Parton vs. Hadron Sensitivity



[Bright-Thonney, Nachman, JDT, PRD 2024;
see also Komiske, Metodiev, JDT, PRL 2019; Kitouni, Nolte, Williams, MLST 2023]

Whether or not these techniques count as "interpretable",
they are designed to be more robust to systematic effects…

Actionable Goal: **Qualitatively Assess** Sources of **Systematic Uncertainties**

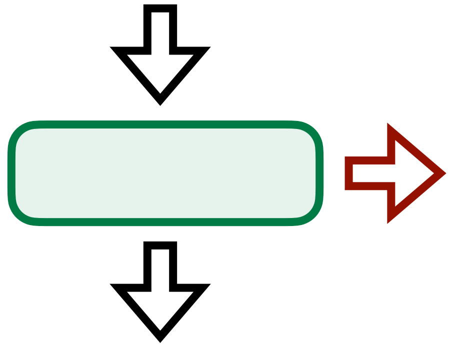
…though it is unclear how to quantify the level of
improvement without additional dedicated studies

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*

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

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

The Next Frontier: Foundation Models

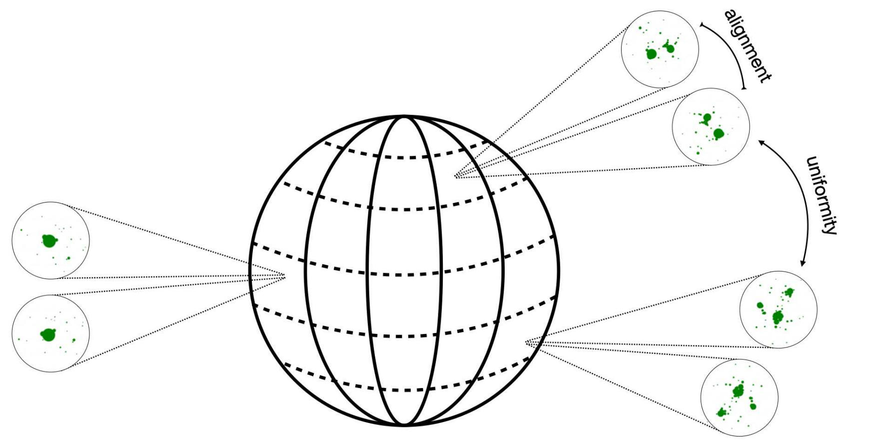
Identify features useful for generic tasks on large datasets, which get reused/refined for specialized applications on small datasets

Purposeful misalignment between initial and downstream goals

Foundation Models for HEP

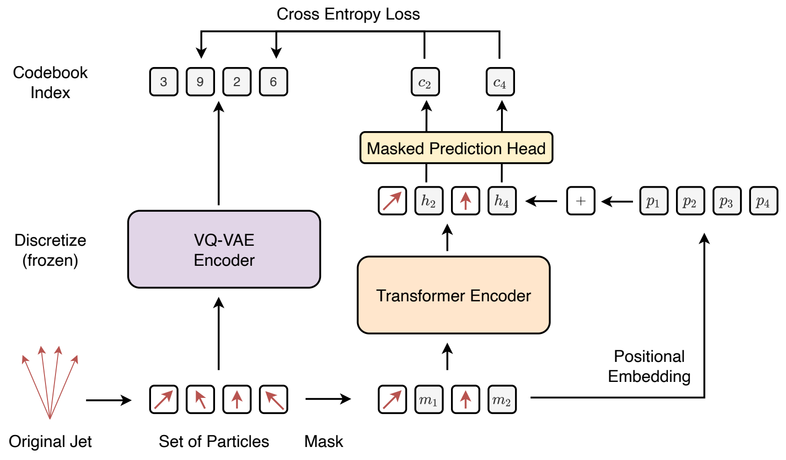
The natural evolution of transfer learning

Symmetry Augmentation



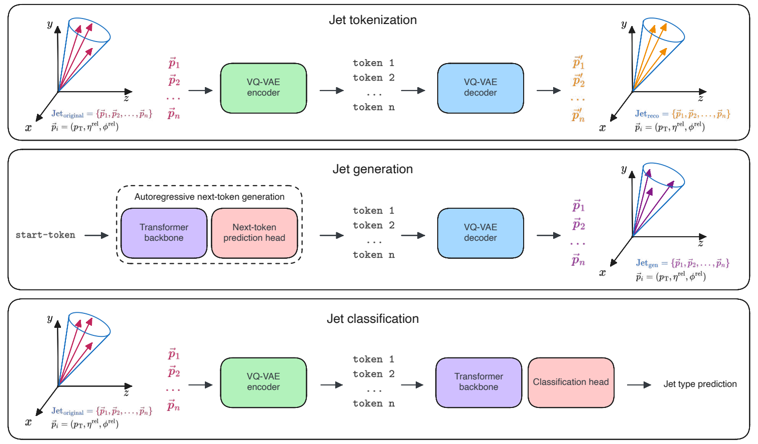
[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, SciPost 2021]

Masked Particle Modeling



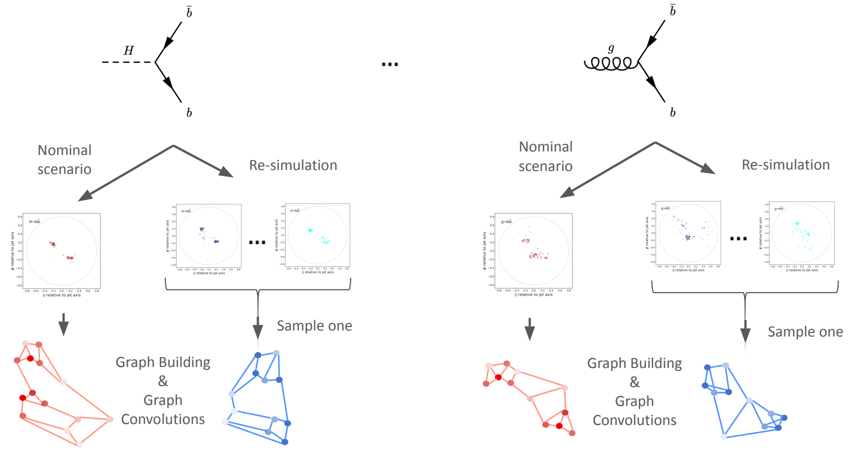
[Heinrich, Golling, Kagan, Klein, Leigh, Osadchy, Raine, arXiv 2024]

Next Token Prediction



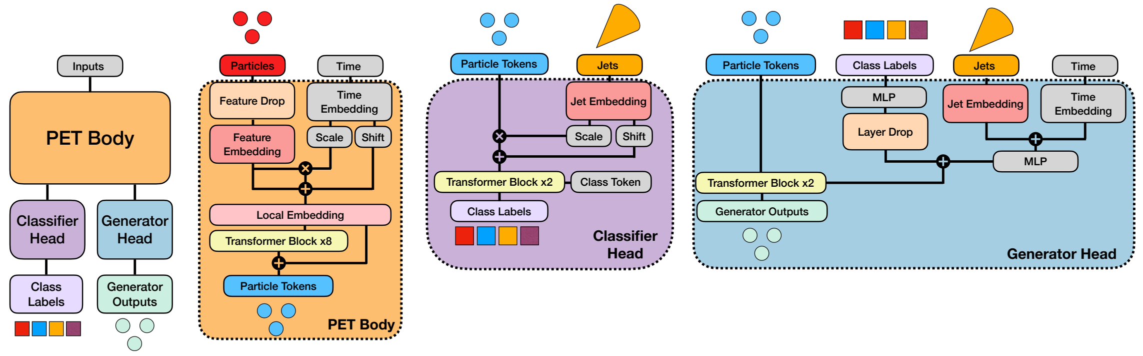
[Birk, Hallin, Kasieczka, arXiv 2024]

Re-Simulation Similarity



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

Multi-Category Classification



[Mikuni, Nachman, arXiv 2024]

Your Next Paper

Inputs

Learned Features

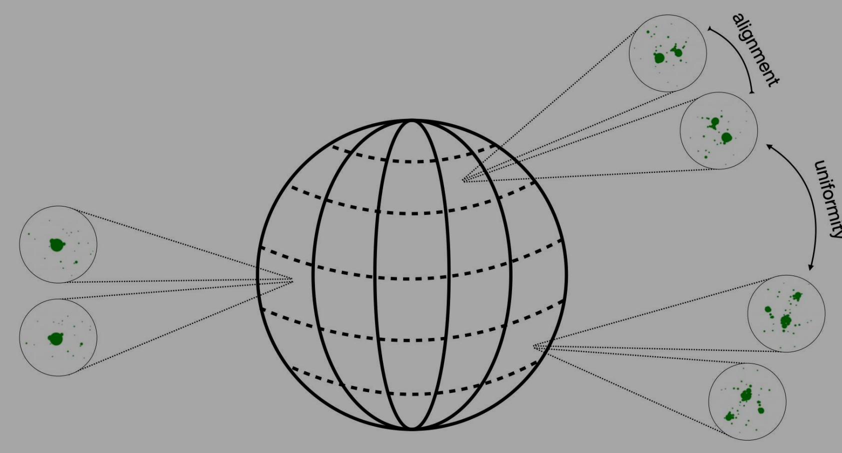
Train to do generic task

Reuse to do specific task

Foundation Models for HEP

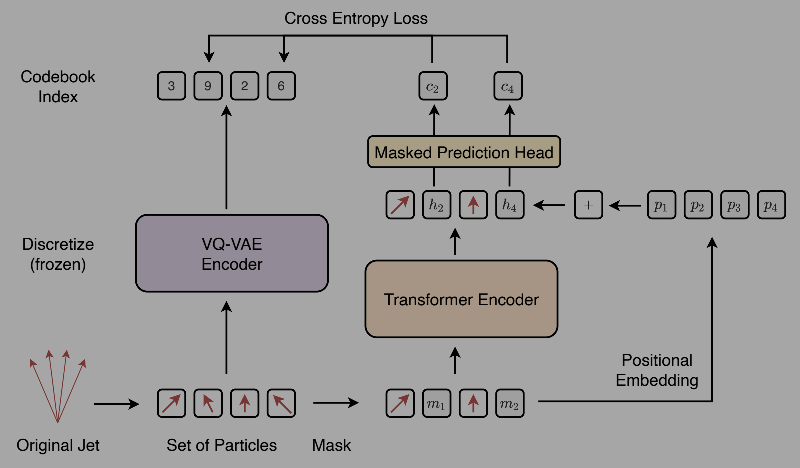
The natural evolution of transfer learning

Symmetry Augmentation



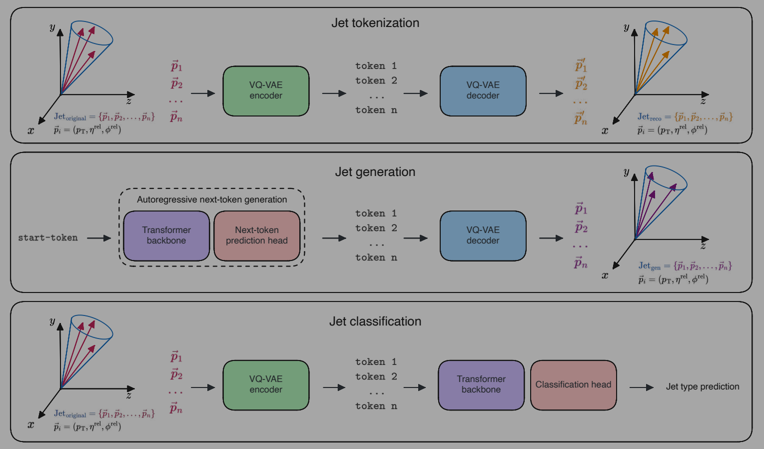
[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, SciPost 2021]

Masked Particle Modeling



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

Next Token Prediction

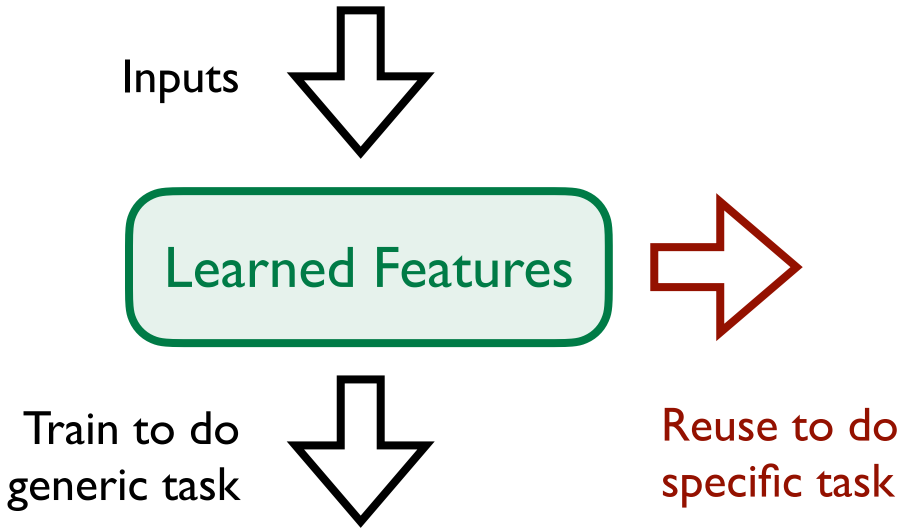


[Birk, Hallin, Kasieczka, arXiv 2024]

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

"What is the machine learning?!"

Your Next Paper



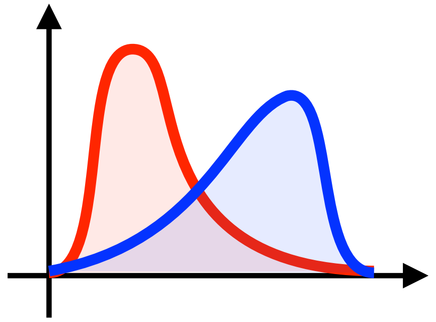
If you have access to a large ancillary data set,
pre-training is a powerful way to learn useful features…

???

Actionable Goal: Identify **~~Low-Rank~~ Structures** in **High-Dimensional Datasets**

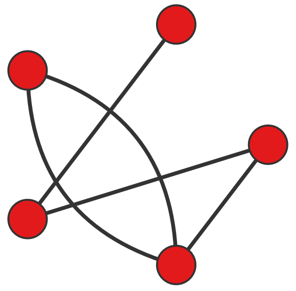
…though I am unsure of the statistical implications
of leveraging information gained from auxiliary tasks

Interpretable Machine Learning for Particles Physics



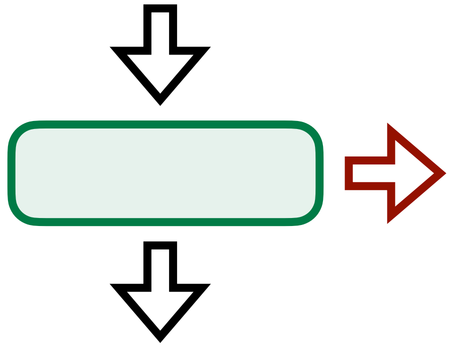
Confronting the Black Box

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*

PHYSTAT Workshop Theme: Interpretability

ChatGPT 4o: "Draw a picture related to this prompt"

My evolving perspective (open to changing my mind!):

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

We should strive towards **actionable goals** for interpretability:

1. Qualitatively assess sources of **systematic uncertainties**
2. Identify **low-rank structures** in high-dimensional datasets
3. [Your ideas here!]

Actionable Goal: Start a **Vibrant Discussion** of Interpretability at PHYSTAT!