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

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 1

Jesse Thaler

Interpretable Machine Learning for Particle Physics

Please put captions in this "black box"

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

The NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI /aI-faI/ iaifi[.org\)](http://iaifi.org)

Launched August 2020

Next Generation of AI + Physics Talent

AI for **Cosmological Observations**

IAIFI Postdoctoral Fellows

Gagliano Golubeva Grosso

IAIFI Summer School & Workshop

AI for Lattice QCD

AI for Neutrino Physics

Boyda Bright-Thonnney Cuesta Dogra Gagliano Golubeva Grosso Harvey Luo Micallef Mishra-Sharma Yang

AI Frontiers of Reinforcement Learning

AI for Collider Physics

AI for Time-Domain Astronomy

 $\overline{\psi}$

AI and **Statistical** Physics

AI for String Theory

Mathematical Physics of AI

Bright-Thonnney

AI for Particle Physics

㱺

BI

[Summer Workshop](https://iaifi.org/summer-workshop): August 12–16, 2024

Application deadline typically [early October](https://iaifi.org/fellows.html)

Machine Learning at DFP-Pheno 2024

Computing, Analysis Tools, and Data Handling

Machine Learning & AI: New Physics

Mini-Symposium: Neutrino Science with the DUNE Experiment

Jeen-Jearning at DUNF Far Detector

niversity of Pittsburgh / Carnegie Mellon University

Tral Network Based Fast Optical Simulation Method in ProtoDUNE-VD

niversity of Pittsburgh / Carnegie Mellon Universit

Prof. Jianming Bian $16:45 - 17:00$

Shuaixiang (Shu) Zhang $17:00 - 17:15$

Mini-Symposium: Quantum Instrumentation

Exploring Quantum Machine Learning for High-Energy Physics University of Pittsburgh / Carnegie Mellon Universit

Machine Learning & AI: Collider Physics

Quantum Field & String Theory: Non-perturbativity and Amplitudes

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

Jinghong Yang

 $15:10 - 15:30$

QCD & Heavy Ion Physics: Jets and Energy Correlators

Jet Calibration in ATLAS Using Machine Learning Networks **Benji Lunday** Law 107, University of Pittsburgh $15:00 - 15:15$ **Today at Lunch: DOE PI Meeting Computational HEP and AI/ML Coordinating Panel for Software and Computing (CPSC) Townhall** Coordinating panel for software and computing town

 $17:30 - 18:00$

University of Pittsburgh / Camegie Mellon University

Dark Matter: WIMPs, DM Simulation and ML

eeping the Dust Away: An unbiased map of the Milky Way's gravitational potential using unsupervised ML *Eric Putne*j d Lawrence 120. University of Pittsburg

Instrumentation: Neutrinos, Dark Matter, and Scintillation

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

Electroweak & Higgs Physics: Electroweak Physics at the LHC

New W Boson Decay Channel at the LHC David Lawrence 207, University of Pittsburgh

Peiran Li $17:00 - 17:15$

Compared to 3 talks at Pheno 2019!

"…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

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 5

"…but what is the machine actually learning?"

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

My evolving perspective:

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

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

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

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

$\overline{\mathcal{L}}$ The Next Frontier for Interpretability

Case Study in Jet Classification Lase Stud \overline{y} in

Interpretable Machine Learning for Particles Physics

of jet substructure, if combinations like *C*² and *D*² had

And Confronting the Black Box
To benefit from machine learning advances we must ensure

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

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

Foundation models identify generically useful features, which challenge the importance of task alignment ite and the evidence of the ev

$\frac{1}{2}$ The Next Frontier for Interpretability

Case Study in Jet Classification Lase Stud ly in

And Confronting the Black Box
To benefit from machine learning advances we must ensure

Likelihood Ratio Trick

Key example of simulation-based inference

[see e.g. Cranmer, Pavez, Louppe, [arXiv 2015;](https://arxiv.org/abs/1506.02169) D'Agnolo, Wulzer, [PRD 2019](https://arxiv.org/abs/1806.02350); simulation-based inference in Cranmer, Brehmer, Louppe, [PNAS 2020](https://arxiv.org/abs/1911.01429); relation to f-divergences in Nguyen, Wainwright, Jordan, [AoS 2009;](https://arxiv.org/abs/math/0510521) Nachman, Thaler, [PRD 2021\]](https://arxiv.org/abs/2101.07263)

 \setminus *Q*

Learnable Function:

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

Loss Function(al): $\overline{\langle}$ $\log f(x)$ $\rangle_P + \langle f(x) - 1 \rangle$

Many HEP problems can be expressed in this form!

Likelihood Ratio Trick

Key example of simulation-based inference

[see e.g. Cranmer, Pavez, Louppe, [arXiv 2015;](https://arxiv.org/abs/1506.02169) D'Agnolo, Wulzer, [PRD 2019](https://arxiv.org/abs/1806.02350); simulation-based inference in Cranmer, Brehmer, Louppe, [PNAS 2020](https://arxiv.org/abs/1911.01429); relation to f-divergences in Nguyen, Wainwright, Jordan, [AoS 2009;](https://arxiv.org/abs/math/0510521) Nachman, Thaler, [PRD 2021\]](https://arxiv.org/abs/2101.07263)

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 10

Learnable Function: Training Data: Finite samples P and Q

Goal: Estimate $p(x) / q(x)$ $f(x)$ parametrized by, e.g., neural networks

 $-\min$
f(*x*) $L =$ $\int dx p(x) \log \frac{p(x)}{2}$ *q*(*x*) Asymptotically: arg min *f*(*x*) $L =$ $p(x)$ *q*(*x*)

Kullback–Leibler divergence

$$
\textsf{Loss Function}(\textsf{al}): \quad L = -\big\langle \log f(x) \big\rangle_P + \big\langle f(x) - 1
$$

Likelihood ratio

Many HEP problems can be expressed in this form!

Many HEP problems can be Likelihood Ratio Trick *expressed in this form!*

Key example of simulation-based inference

[see e.g. Cranmer, Pavez, Louppe, [arXiv 2015;](https://arxiv.org/abs/1506.02169) D'Agnolo, Wulzer, [PRD 2019](https://arxiv.org/abs/1806.02350); simulation-based inference in Cranmer, Brehmer, Louppe, [PNAS 2020](https://arxiv.org/abs/1911.01429); relation to f-divergences in Nguyen, Wainwright, Jordan, [AoS 2009;](https://arxiv.org/abs/math/0510521) Nachman, Thaler, [PRD 2021\]](https://arxiv.org/abs/2101.07263)

Asymptotically, same structure as Lagrangian mechanics!

Action:
$$
L = \int dx \mathcal{L}(x)
$$

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

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

Requires shift in focus from solving problems to specifying problems

"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

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

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

Why might we want ML to be "Interpretable"?

…

Scientific Reasons:

…

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

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

[Andreassen, Komiske, Metodiev, Nachman, JDT, [PRL 2020](https://arxiv.org/abs/1911.09107); + Suresh, [ICLR SimDL 2021\]](https://arxiv.org/abs/2105.04448)

Detector Unfolding Monte Carlo Reweighting Resolution Estimation

[Nachman, JDT, [PRD 2020](https://arxiv.org/abs/2007.11586); inspired by Andersen, Gutschow, Maier, Prestel, EPJC 20201 [Gambhir, Nachman, JDT, [PRL 2022,](https://arxiv.org/abs/2205.03413) [PRD 2022](https://arxiv.org/abs/2205.05084)]

\sim μ μ \sim μ \sim μ \sim μ μ \sim μ genau dargestellt werden sollen. Eine zweite Beziehung zwisehen ~ un4 fl [die ,,Breite" 4er ttyperbel (44)] lielert die Bedingung, die Bindungsenergie $\mathsf S$ $\overline{}$ $\overline{}$ \overline{t} datas $\overline{\mathcal{C}}$ $\overline{\mathbf{I}}$ after a reverse $\overline{\mathbf{l}}$ ensional data \mathbf{r} l ar
ar $\overline{}$ **J** neurall
I the resulting $\overline{}$ sum ofnumber
Danmark
Santa Carolina
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
Danmark
 $\overline{\mathbf{C}}$ out, \boldsymbol{l} \int label of the slice. high-dime numbers validation: (right) \blacksquare *IAentitving IOW-I* ² ⁰*.*⁰⁰⁴⁰⁰ ✓ (26*.*2*^P* ²)(*A*+26*.*2(*^P* ² ²*.*35*NZ*)) *^Z*² ⁺ ¹*.*70◆ error. Zero values were clipped to 10³ for visualization. Error bars are standard depth in the structures in the depth of the standard depth in the standard depth in the standard o with the same training fraction. The same training fraction \mathbf{r} R *a diffuing low-rank structures in high-dimension.* -rank structures in high-dimension *dentifying low-rank structu Identifying low-rank structures in high-dimensional datasets*

 τ k C \sim solle etwas geringer sein als die empirische Bindungsenergie kalender Der vergleich der termen Energien enter der inben der inben est.
In the leads of the leads o
Alternatives of the leads of the
 m of inter $of i$ \mathbf{b} n
n ϵ the network th around aph₎ sumC
C the network orob
 information of \overline{a} π *n* This is an actionable definition of interpretability, which may or may not be relevant to modular*I* his is an actionable defi where α , you are *defined* to the average processing, which may interpretability which may ar may not be relevant to the physics where α ² di_e average pairwise **L2-distance** between α where *d*_{at} *x* average commercies of most processing, which may of the the relevant to the physics proble This is an actionable definition of interpretability, which may or may not be relevant to the physics problem of interest

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 16 elements in **x** and **y.** This is the ratio of the average of the average of the set of the average of the average of the set of the set of the set of the set ϵ interpretable mathine Learning for randeler hysics elements in *x* and *y*. The *x* and *y* and *y* and **y** and *y* and **y** and **y** and **y** average the average of the second second second second $\frac{1}{2}$ esse maier (min, $\frac{1}{2}$ min) $\frac{1}{2}$ miter pretable machine Learning to embeddings to extract values to extract values to extract values of \mathcal{E} elements in *x* and *y* average the *n* ϵ pretable raditime Learning for randeler hysics

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

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

Foundation models identify generically useful features, which challenge the importance of task alignment $\frac{1}{2}$ The Next Frontier for Interpretability ite and the evidence of the ev

Case Study in Jet Classification Lase Stud \overline{y} in

of jet substructure, if combinations like *C*² and *D*² had not been previously identified. Distributions of the EFPs

and **After Confronting the Black Box**
To benefit from machine learning advances, we must ensure

The More Things Change…

Jet classification, from my talk at Pheno 2019

Application of Likelihood Ratio Trick Interpretability in Machine Learning

Trying to plot 256 dimensional latent space

See Pheno 2019 talk for insights

Three Lessons since Pheno 2019
 Apologies that exame whose bounded gradients ensure bounded sensitivity to iree Lessons since Pheno Zu with constituent momenta *p*1*, p*2*,...,pM*, an EFN com-Theory Lossans since P **built on the standard EFR of a standard EFR in the standard E**

Highlighting the power of active interpretability Highlighting the powe **In this paper is proventy of a set of a number of the preceding pools** \mathcal{F} n lighting the power of act Highlighting the hower of active intert *p*₂, p₂, p₂

> If there are simple operations like multiplication and It there are simple operations like multiplication and $\langle\Phi^{a_1}\Phi^{a_2}\rangle_{\cal D}$ sums that don't really require "interpretation", you can italisative machine learning architecture bake those into your machine learning architecture $\langle \Phi^{a_1} \Phi^{a_2} \rangle_{\mathcal{P}}$ sums that don't re $\langle \Phi^{a_1} \Phi^{a_2} \rangle_{\mathcal{P}}$ sums that don't really req $\langle \mathbf{\Phi}^{a_1} \mathbf{\Phi}^{a_2} \rangle$ If there are simple operations like multiplication and $\langle \mathbf{\Phi}^{a_1} \mathbf{\Phi}^{a_2} \rangle$ $f(x) = \frac{f(x)}{f(x)}$ is safilly the output of the observation that the original text is more interpretation to $f(x)$ for $f(x)$ Γ is a natural permutationfunction *^F* : ^R` ! ^R*^d*out maps the latent representation structure of an EFN guarantees and an EFN guarantees a naturally permutation-
The contract of an EFN guarantees and an EFN guarantees and an EFN guarantees and an EFN guarantees and an EFN

T_{max} Pooling operation, as given by Eq. (2), as summation step over the latent representation *^a* in Deep

Apologies that examples are all from my own work Apologies that examples
are all from my even work Pooling is capable of reducing the latent dimension of

rever inputs
where **Figures**

of weighted Deep Sets, to produce Moment EFNs in

If there is a property you want your network to have, $\|\Phi(\hat{p}_1) - \Phi(\hat{p}_2)\|$ it there is a property you want your network to have,
 ζ is a sure to impose algorithmic guardrails, otherwise $\leq L \| p_1 - p_2 \|$ the machine might pursue undesirable optimization If there is a proporty vou Sets architectures, generalized to weighted sums in EFNs, your network to have,
is suerdrails otherwise iic guaruralis, other wise
esirable optimization space representation \mathbf{P} $T_{\rm eff}$ is e $T_{\rm eff}$ and the gradients of the gradients of the gradients of these functions of these functions of these functions of the gradients of th An *L*-EFN extends the EFN setup by constraining

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

\nthe point \mathbf{r} is a matrix.

If you have a catalog of trusted observables, you can let us the latent of latent dimensions of trusted observables, you can translate a black-box algorithm on low-level inputs into a simple classifier on high-level features =2*,*= *mandate a black-box algorithm on low-level in the a simple classifiers an high lovel footures* **Pooling operation into a simple classifier on** where *zⁱ* = *pT ,i/pT ,*jet is the constituent momentum or en and α is the particle of α relative to the set of the into a simple c

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

$$
\begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix}
$$

 $T_{\rm tot}$ selected during each iteration of the black-box guiding strategy beginning strategy beginning from HLN0, which uses $T_{\rm tot}$

[Faucett, JDT, Whiteson, [PRD 2021](https://arxiv.org/abs/2010.11998); using Komiske, Metodiev, JDT, <u>JHEP 2018</u>; C₃ from Larkoski, Salam, JDT, <u>JHEP 2013]</u> Faucett, JD I, VYhitesor
The the points of correctly of the points Faucett, JDT, Whiteson, <u>P.</u>
Ing Komiske, Metodiev, JDT, <u>JHEP 2018</u>; C₃ from Larkoski, Salam, JDT, <u>JH</u> *p*2 T **Example of Sec. INCONSTRUCT INCORPORATION:**

A glimpse at an alternative history for field of jet substructure

Translating the Black Box

Selecting Energy Flow Polynomials that mimic CNN decisions

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

Alternative pooling operations for streamlined latent spaces In this paper, we introduce \mathcal{L} is the moment Pooling, a natural Alternative pooling operations for streamlined fatent spaces Alternative pooling operations for streamlin is a generalization of Deep Sets-style architectures. The \mathcal{L} native pooling operations for streamlined latent spaces

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 22 \ddot{x}

Moments of Clarity \mathbf{A} EFN architecture, show how it naturally arises as a generalization of \mathbb{R}^n $emerts of Clarity$ of Eq. (2) simply extends *F* from being a function of

> Γ and Γ in Γ or Γ and Γ is performance. The Moment of Γ Combining per-particle reatures through multiplication and summation Combining per-particle features thr multiplication and summation k α par particle footures through puris puris puris de formation de la partie d
de la partie de la Combining per-particle features through

context, where EFN (2) define an order **EFN**, $\mathbf{r} = \mathbf{r} \times \mathbf{r}$ and $\mathbf{r} =$ (Deep sets, LTTV, K=T) $(K - 2)$ where *^zⁱ* are weights and *^xⁱ* ² ^R*^d*. In the collider $(P^e(P)$ decs, $L^f(P, R-1)$ and (R^e) Sum Pooling (Deep Sets, EFN, k=1)

Ok(*P*) ⌘ *F* ^h*^a*i*^P ,*h*^a*1*^a*² ⁱ*^P , ...*h*^a*¹ *...^a^k* ⁱ*^P* where *k* is the highest order moment considered. This procedure is inspired by histogram pooling [38], in which the are histograms binned in *x*. We focus primarily on applying Moment Pooling to EFNs in the collider physics dimensional random variable *^a*(*x*) defined over a base space ^R*^d*, taken over *^P*: ^h*^a*i*^P* ⌘ ^X *i*2*P zi^a*(*xi*)*,* (3) *^O*2(*P*) = *^F*(h*^a*i*^P ,*h*^a*1*^a*² ⁱ*^P*)*,* (5) where ^h*^a*1*^a*² ⁱ is the second moment of , which is: ^h*^a*1*^a*² ⁱ*^P* ⁼ ^X *i*2*P zi^a*¹ (*xi*)*^a*² (*xi*)*.* (6)

r ² = $c_1 + c_2$ ¹ where *^zⁱ* are weights and *^xⁱ* ² ^R*^d*. In the collider ing *k* = 4*, L* = 1 Moment EFN, as a function of rapidity (*y*). Log Angularity through _{The} $\mathcal{L}_{\text{unab}}$ shown as a data below for $\mathcal{L}_{\text{unab}}$ **SYITDUIL NEUTESSIUI** $1 - C_1 + C_2 \log(C_3 + T)$ $\Phi_{\mathcal{L}}(r) = c_1 + c_2 \log(c_3 + r)$, sion: Symbolic ReGression:

 ϵ are discussed in Refs. (39, 40). The discussed in Refs. and discussed in Refs. (39, 40). Same prinosophy (and scaling) as Energy Flow Networks, sume or better performance of the process of classical permitted on p Same philosophy (and scaling) as Energy Flow Networks, Log sume prinosophy (and scaning) as Litergy riow by flow permatation-liveriant pooling operation processes *just new permutation-invariant pooling operations*

> to be constructed using fewer base functions. With fewer base functions. With fewer base functions. We can see Exambhir, Osathapan, JDT, <u>ar Xiv 2024</u>; building off Komiske, Metodiev, JDT, <u>JHEP 2019; Andreal State of Band</u>
And State Common Kenisch, Bisani Villagesuse, Neueuro, Sacreal He JCLB 2021, Sim DL1 ranmer, Kreisch, Pisani, Villaescusa-Navarro, Spergel, Ho, <u>ICLR 2021 SimDLJ, Andreplace the neural network in th</u>
Moment EFN: We can replace the new replace the network in the new replace the new replace the new replace t $T_{\rm s}$ is the same as $T_{\rm s}$ in the closed-form expressions defined in $T_{\rm s}$ in the trainable parameters of $T_{\rm s}$ see also Cranmer, Kreisch, Pisani, Villaescusa-Navarro, Spergel, Ho, <u>[ICLR 2021 SimDL](https://simdl.github.io/files/40.pdf)]</u>
And Latent fit to the late

Moment Pooling

 $(k = 2)$

 -3.0

 -2.5

 -2.0

Formal IRC safety doesn't immediately ensure small non-perturbative corrections with constituent momenta *p*1*, p*2*,...,pM*, an EFN com-

we show how to train an IRC-safe neural network to be

 $m_{\tilde{t}}$

thereby constructing an observable that is "safe but in-

calculable".³ As a step towards restoring calculablity,

we introduce Lipschitz Energy Flow Networks (*L*-EFNs),

whose bounded gradients ensure bounded sensitivity to

 $T_{\rm eff}$

tions with large non-perturbative corrections is not new,

even if it may not be widely appreciated. The stan-

A. Lipschitz Energy Flow Networks

Safe but Incalculable The but Indalculable built on top of a standard EFN, which provides a generic standard EFN, which provides a generic standard provi

putes a function of the form:

zi(ˆ*pi*)

Re<mark>gularizing learned fea</mark>tures to ensure controlled behavior of per-particle representations

, (1)

will be small. Using EFNs as a representative example, we show how to train an IRC-safe neural network to be maximally sensitive to non-perturbative hadronization, thereby constructing an observable that is "safe but incalculable".³ As a step towards restoring calculablity, II. METHODOLOGIES

The *L*-EFN architecture we propose in this work is

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

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

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

$\overline{\mathcal{L}}$ The Next Frontier for Interpretability

Case Study in Jet Classification Lase Stud ly in

of jet substructure, if combinations like *C*² and *D*² had not been previously identified. Distributions of the EFPs

and **After Confronting the Black Box**
To benefit from machine learning advances, we must ensure

 \checkmark

[[HEPML-LivingReview,](https://iml-wg.github.io/HEPML-LivingReview/) moderated by Nachman, Feickert, Krause, Winterhalder]

From the Living Review of ML for Particle Physics

Fascinating categorization!

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

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

\blacksquare curvature as well for the system under system E.g. SHapley Additive exPlanations (SHAP) interprets machine learning models using Shapley values.

12 variable machine learning assisted analysis for classifying 5 particle-production channels

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 26 *driving the different channels*

 Shapley value distributions, $**8.67**$ **Tree and SHAP. This leads to the higherarchy of variable importance,** $**9.1**$ [Grojean, Paul, Qian, Strümke, [Nature Reviews Physics 2022\]](https://arxiv.org/abs/2203.08021)

Goal: Identify features driving decisions about classification

Quite similar to goal of identifying low-rank features

"…but what is the machine actually learning?"

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

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

"…but what is the machine actually learning?"

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 28

Identify features useful for generic tasks, which get reused for specialized applications Purposeful misalignment between initial and downstream goals

The Next Frontier: Foundation Models

Foundation Models for HEP

Symmetry Augmentation

[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, <u>SciPost 2021]</u>

Re-Simulation Similarity *^s*(*zi, z^j*) = *^zⁱ · ^z^j |zi||z^j |* = cos ✓*ij ,* (6)

Masked Particle Modeling Next Token Prediction 2

[Heinrich, Golling, Kagan, Klein, Leigh, Osadchy, Raine, [arXiv 2024\]](https://arxiv.org/abs/2401.13537) $\frac{1}{2}$ ieimich, Obimig, inagan, iniem, Leigh, Osauch), name, <u>al XIV-VO-V</u>

Multi-Category Classification iilti (Gatagar UILI LALEYUL \mathbf{r} in the domains of language and vision. We refer readers explored the discretization of continuous particle features \blacksquare In parallel to the present effort on self-supervised foundation models, investigations are ongoing on the potential

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 29 \overline{a} are the theory specific neural net-

[Birk, Hallin, Kasieczka, <u>arXiv 2024]</u>

[Mikuni, Nachman, [arXiv 2024](https://arxiv.org/abs/2404.16091)]

The natural evolution of transfer learning

Foundation Models for HEP

Symmetry Augmentation

[Dillon, Kasieczka, Olischlager, Plehn, Sorrenson, Vogel, <u>SciPost 2021]</u>

cause the positive pairs appear in the numerator, while the negative pairs contribute to the denomination, the loss decreases when the distance between positive pairs becomes smaller pairs becomes smaller and when the distance between \mathbf{w} is referred to as the temperature and controls the relative influence of positive \mathbf{r} negative pairs. The cosine similarity in Eq.(6) is not a proper distance metric, but we can various downstream tasks. Thursday explored in Refs. [18–20], largely focusing on contrastive pre-training using augmentations of jets. Supervised pretraining strategies have also been explored in Ref. [21]. unordered sets, in contrast to the sequential nature of sentences, we develop a masked prediction scheme which is a property of \mathbf{f} challenge stems from the continuous nature of particle features, in contrast to the discrete dictionary found in language models but similar to the challenges of masking "What is the machine learning?!"

On this sphere we define the similarity between two jets as [36]

and the total loss is given by the sum over all positive pairs in the batch, *^L* ⁼ ^P

Masked Particle Modeling Next Token Prediction 2

define an angular distance as *d*(*zi, z^j*) = ✓*ij/*⇡ = 0 *...* 1, such that it satisfies the triangle

inequality.

Uniformity vs alignment

II. RELATED WORK

<u>ij being angle being the angle between the strastive loss in a positive loss for a positive loss for a positiv</u> jets is definition in the terms of the terms *i*,*z*₀*p*_e*r*^{*i*}/*p*e*r*^{*for*} *p*erformance, but v to the recent recent recent recent recent real \mathcal{L}_1 for an overview. Most closely rela connat via masking and prediction of missing words as a pre-training σ and the state σ language modeling method to images by masking and predicting patches of input images. On masked modeling $s = \frac{1}{2} \int_{0}^{1} \frac{1}{2} \int_{0}^{1} \frac{1}{2} \frac{1}{2$ the impact of removing positional information in masked image modeling was examined in Ref. [11], and using posipretext tasks followed by fine-tuning in a hierarchical \blacksquare Asymptotically, pre-training cannot yield improved The proposed model and training scheme is summarized \blacksquare ing. In line with the MLM framework employed by \blacksquare \bf{J} , at the \bf{C} performance, but very effective in practice

developed (for example, see Refs. [24–31]). Transformers

model (Birk, Hallin, Kasieczka, <u>arXiv 2024</u>)

The natural evolution of transfer learning

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

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

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

$\overline{\mathcal{L}}$ The Next Frontier for Interpretability

Case Study in Jet Classification Lase Stud \overline{y} in

Interpretable Machine Learning for Particles Physics

And Confronting the Black Box
To benefit from machine learning advances we must ensure

The NSF Institute for Artificial Intelligence and Fundamental Interactions (IAIFI /aI-faI/ iaifi[.org\)](http://iaifi.org)

Jesse Thaler (MIT, IAIFI) — Interpretable Machine Learning for Particle Physics 32

Arti cial intelligence as a pathway to scientific insight

> *Progress driven by early career talent with interdisciplinary expertise Consider applying to IAIFI Postdoctoral Fellowship this Fall!*

Physics intelligence as a pathway to AI innovation