Theory Perspective on Machine Learning for Jets

Jesse Thaler

MIT

Jets and their Substructure from LHC Data, CERN-TH Virtual Workshop — June 3, 2021

The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI) *"eye-phi"*

Advance physics knowledge — from the smallest building blocks of nature to the largest structures in the universe — and galvanize AI research innovation

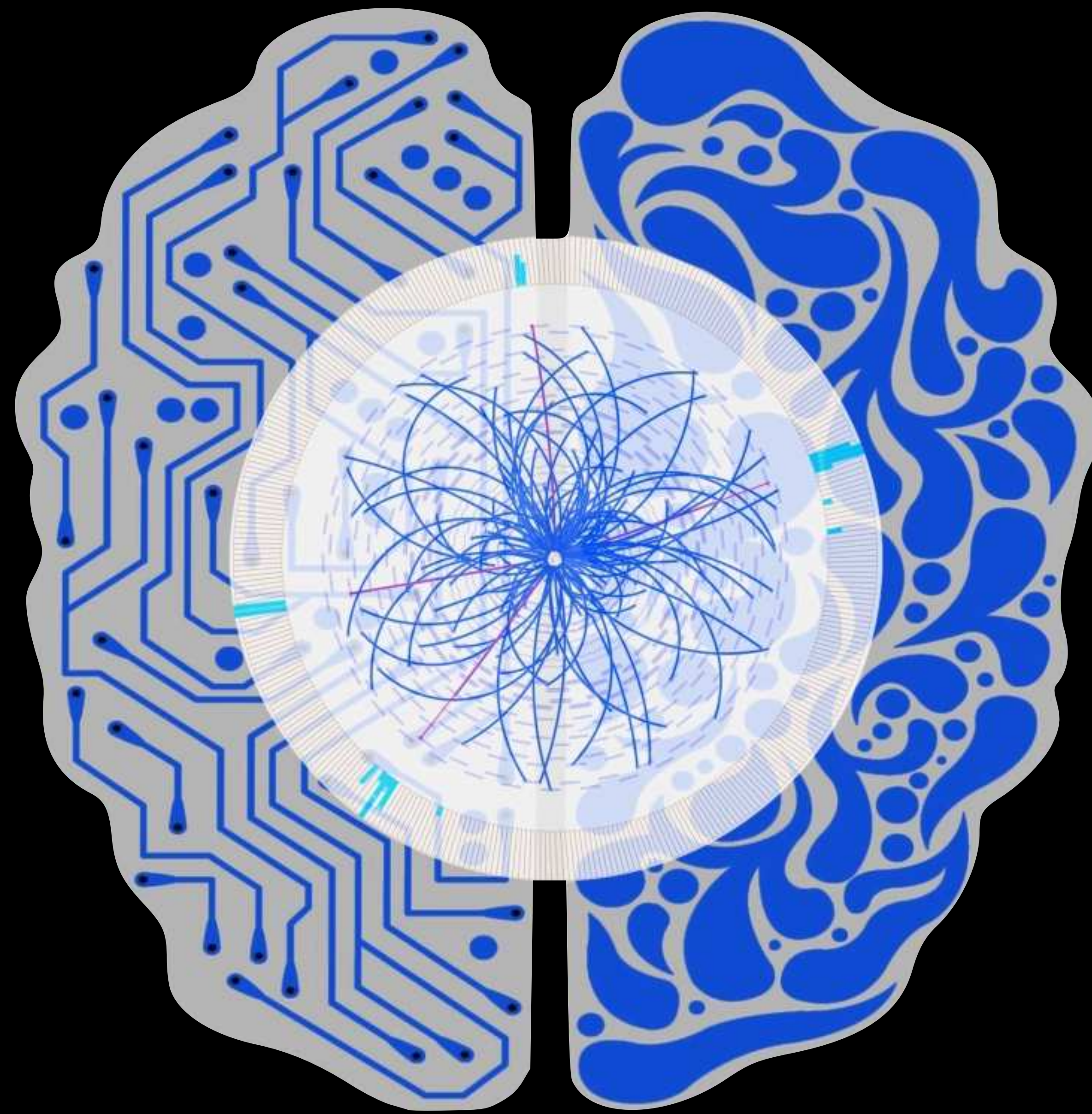
[http://iaifi.org, MIT News Announcement]

AI^2: Ab Initio Artificial Intelligence

Machine learning that incorporates
first principles, best practices, and domain knowledge
from fundamental physics

Symmetries, conservation laws, scaling relations, limiting behaviors, locality, causality, unitarity, gauge invariance, entropy, least action, factorization, unit tests, exactness, systematic uncertainties, reproducibility, verifiability, …

Confronting the Black Box



How do we develop robust machine learning for jet analyses?

Likelihood Ratio Trick

Many QCD/collider/jet problems can be expressed in this form!

Key example of simulation-based inference

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, JDT, arXiv 2021]

Likelihood Ratio Trick

Many QCD/collider/jet 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 theoretical 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, JDT, arXiv 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

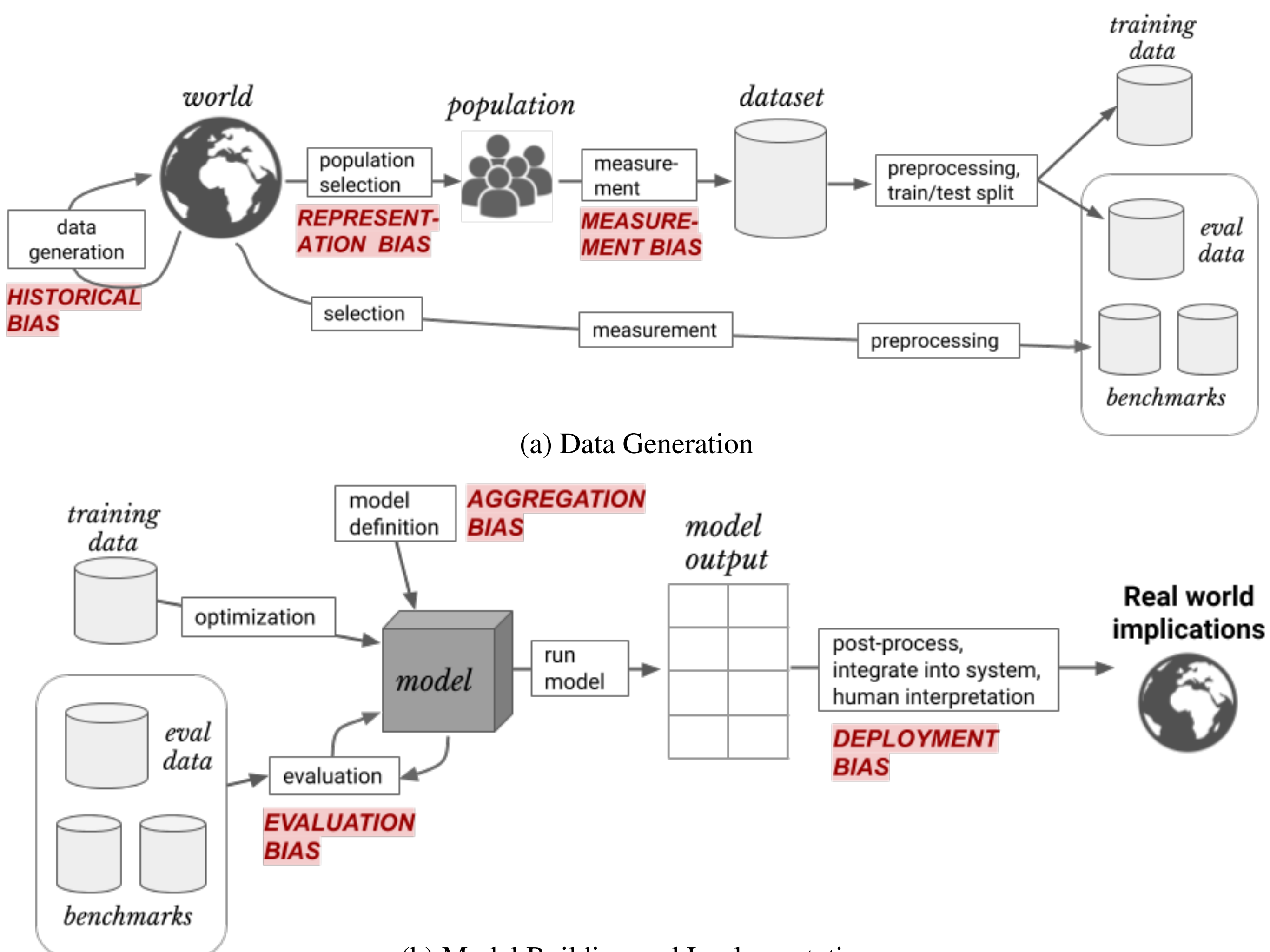
"But where's the physics?!"

In the choice of loss function, data samples, network architecture, and training paradigm

"…"

Many Reasons to be Wary!

"A Framework for Understanding Unintended Consequences of Machine Learning"



(a) Data Generation

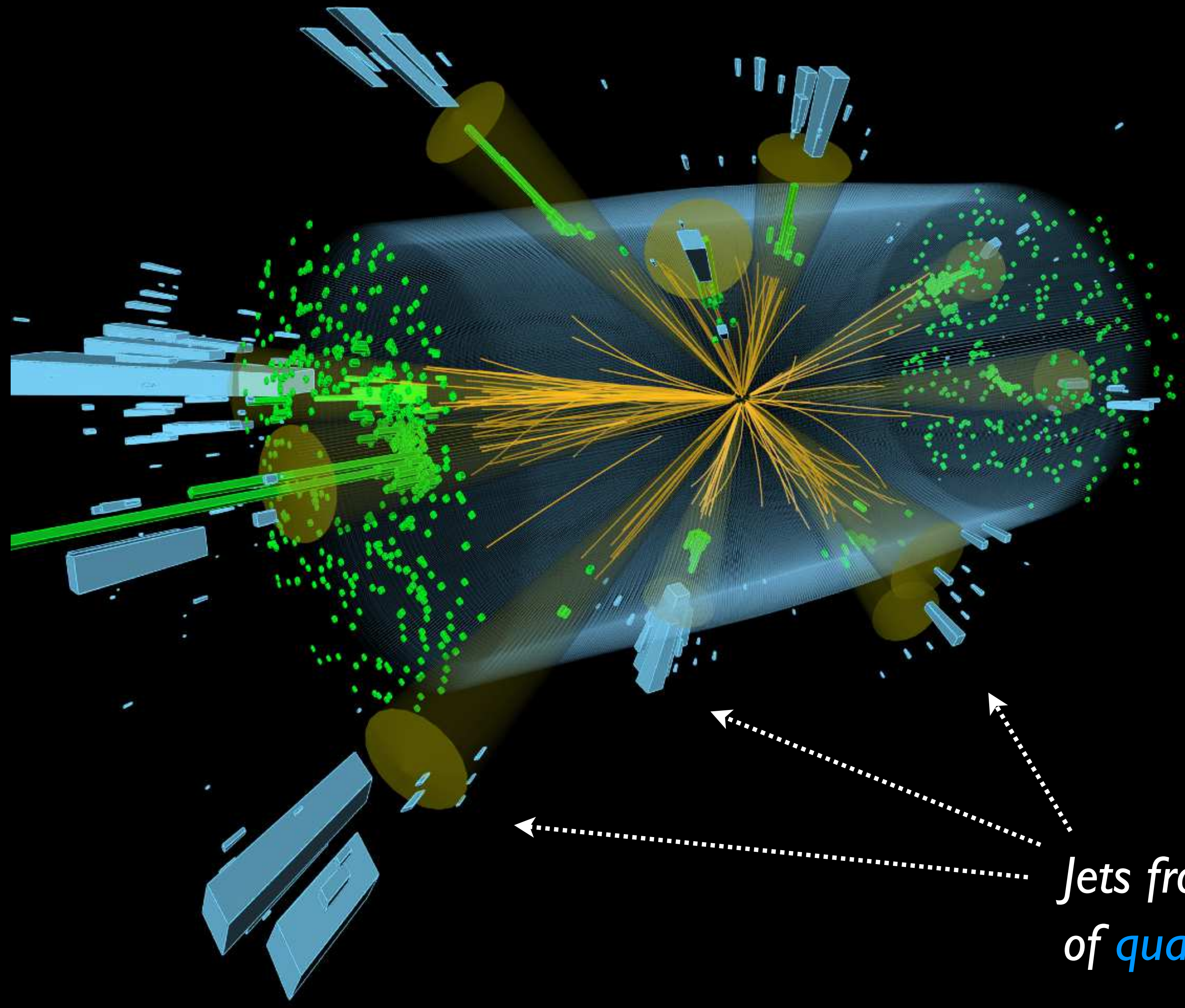
(b) Model Building and Implementation

1. **Historical bias** arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.
2. **Representation bias** arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.
3. **Measurement Bias** arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.
4. **Aggregation bias** arises during model construction, when distinct populations are inappropriately combined. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups.
5. **Evaluation bias** occurs during model iteration and evaluation. It can arise when the testing or external benchmark populations do not equally represent the various parts of the use population. Evaluation bias can also arise from the use of performance metrics that are not appropriate for the way in which the model will be used.
6. **Deployment Bias** occurs after model deployment, when a system is used or interpreted in inapppropriate ways.

For collider physics, "bias" ≈ "systematic uncertainty"

[h/t David Kaiser, MIT SERC; Suresh, Guttag, arXiv 2019]

Machine Learning for Jet Substructure



Jets from fragmentation of quarks and gluons

How can we leverage theory to advance machine learning for jets?

My (Evolving) Perspective

When striving for "interpretable machine learning" we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms

Examples below are representative, not exhaustive; apologies!

[see HEPML-LivingReview for extensive bibliography]

When striving for “interpretable machine learning” we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

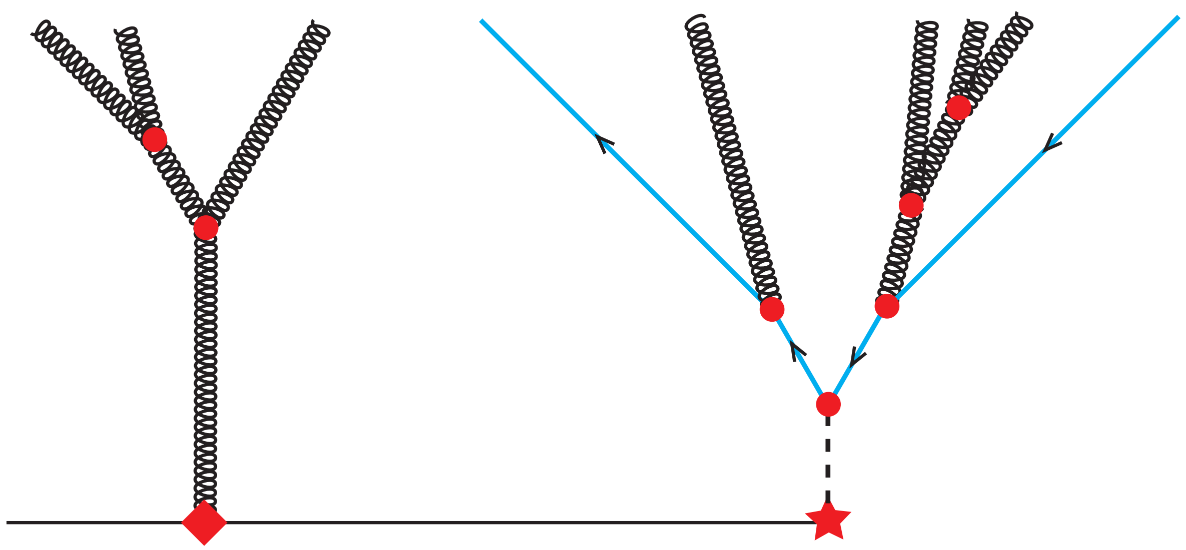
Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms

Compute Likelihood Ratios Directly?

Yes, you can! And if you have enough calculational power, you should!

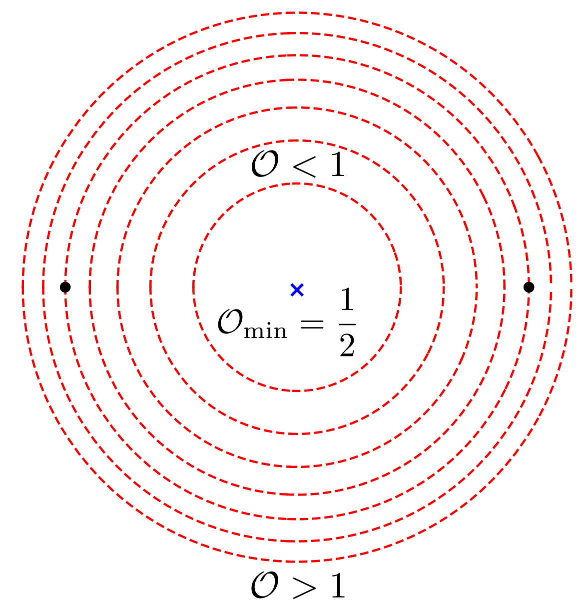
Shower Deconstruction



$$\chi(\{p,t\}_N) = \frac{P(\{p,t\}_N|\mathrm{S})}{P(\{p,t\}_N|\mathrm{B})}$$

[Soper, Spannowsky, PRD 2011]

Color Singlet Identification



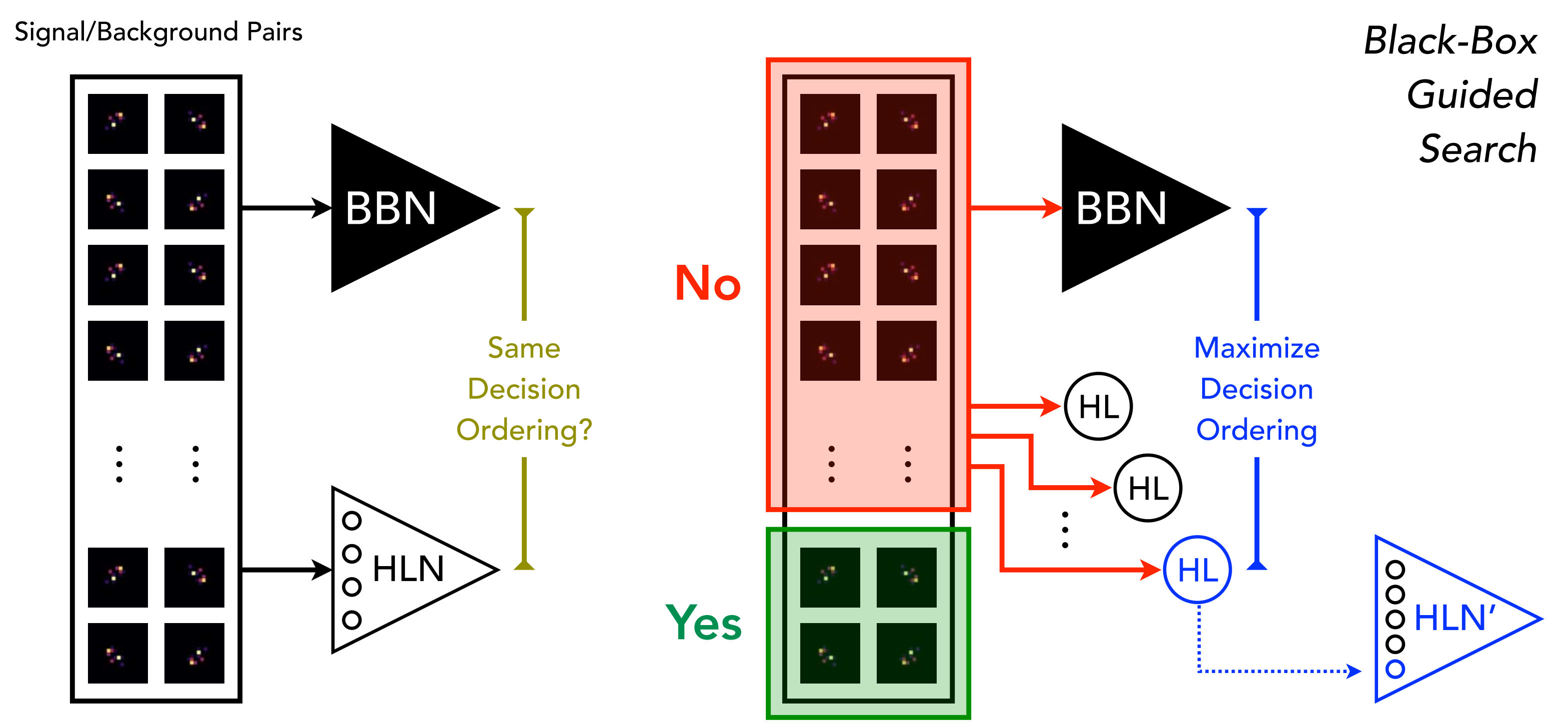
$$\frac{|\mathcal{M}_{\mathcal{B}}|^2}{|\mathcal{M}_{\mathcal{S}}|^2} \simeq \frac{1-\cos\theta_{ak}+1-\cos\theta_{bk}}{1-\cos\theta_{ab}}$$

[Buckley, Callea, Larkoski, Marzani, SciPost 2020]

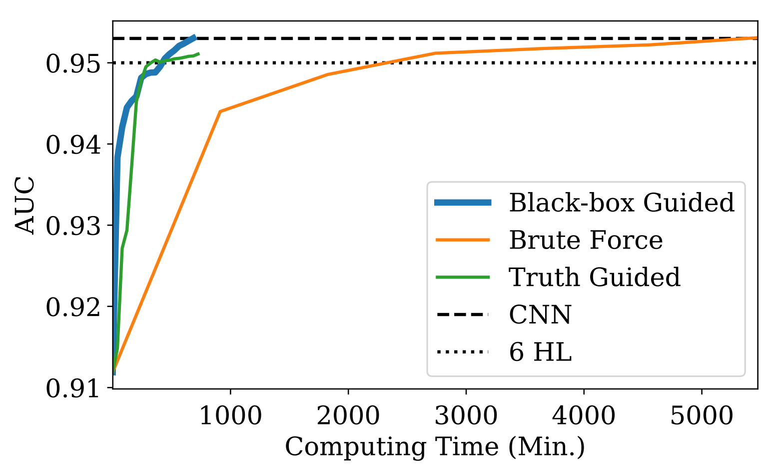
Challenge is that in most cases, best estimate of likelihood ratio comes from complex simulations with no closed form expression

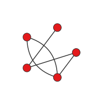
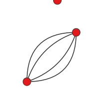
Identifying Novel Jet Observables

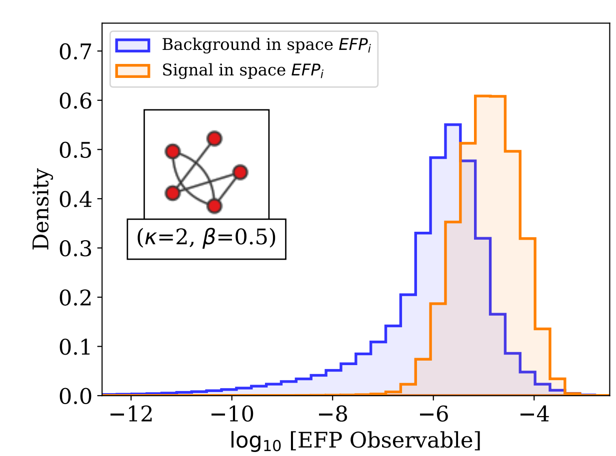
*When **likelihood ratio** is not a function of standard high-level observables*



A glimpse at an alternative history for field of jet substructure



Iteration (n)	EFP	κ	β	Chrom #
0	$M_{\rm jet} + p_{\rm T}$	–	–	–
1		2	$\frac{1}{2}$	2
2		0	2	2
3		0	–	1

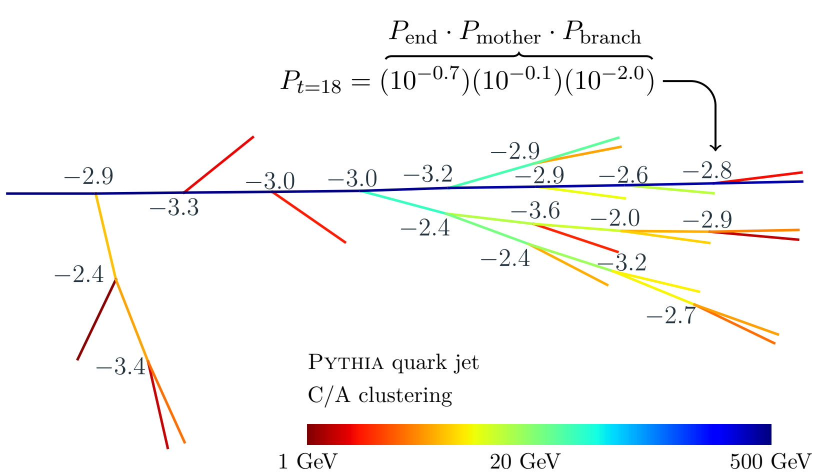


[Faucett, JDT, Whiteson, PRD 2021; using Komiske, Metodiev, JDT, JHEP 2018]

Theory-Inspired Likelihood Parametrizations

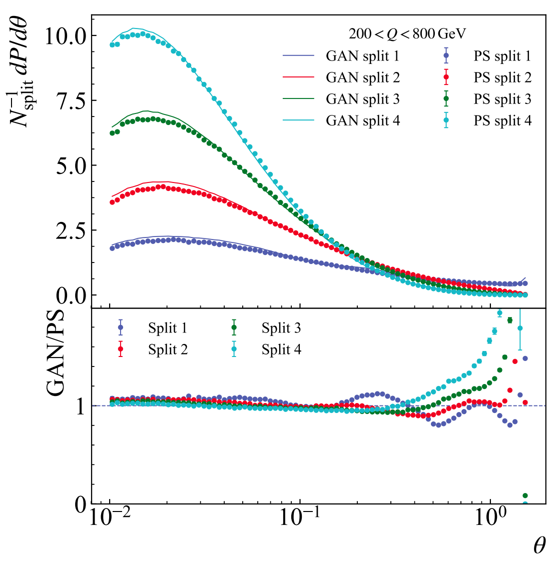
Flexible frameworks for parton-shower-like modeling

JUNIPR



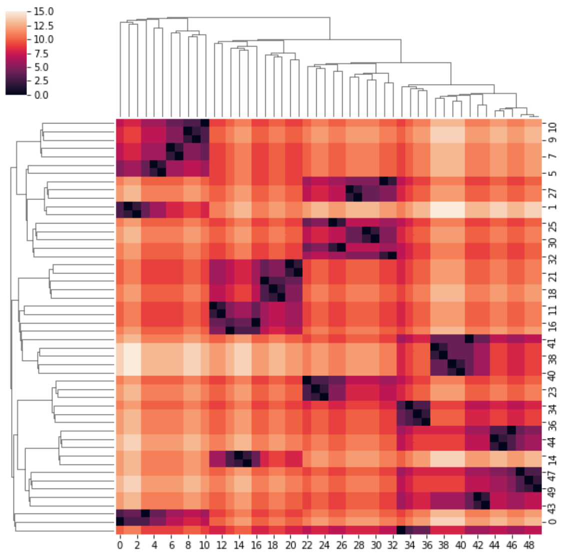
[Andreassen, Feige, Frye, Schwartz, EPJC 2019]

DGLAP White Box



[Lai, Neill, Płoskoń, Ringer, arXiv 2020]

Ginkgo



[Cranmer, Drnevich, Macaluso, Pappadopulo, arXiv 2021]

These methods are based on constrained generative models, hopefully generalizing better than generic methods

Exhibit close relationship between generation and inference

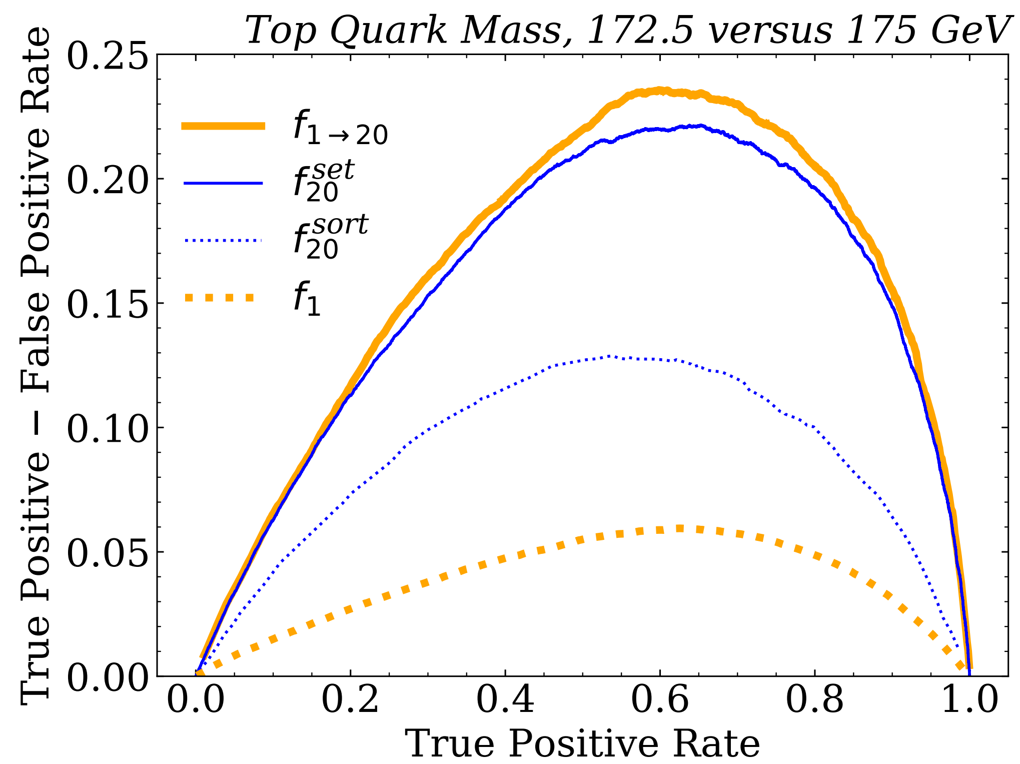
Sidestepping Per-Event Likelihood Ratios?

Could be helpful, though per-event information is usually complete

Collider events are independent and identically distributed…

$$\prod_{i=1}^{N} \frac{p(x_i|\theta_A)}{p(x_i|\theta_B)} = \frac{p(\{x_1,\ldots,x_N\}|\theta_A)}{p(\{x_1,\ldots,x_N\}|\theta_B)}$$

…therefore, (parametrized) per-event binary classifiers can be used to construct asymptotically optimal per-ensemble inference tools



[Nachman, JDT, arXiv 2021; see mixed sample discussion in Metodiev, Nachman, JDT, JHEP 2017]

When striving for "interpretable machine learning" we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

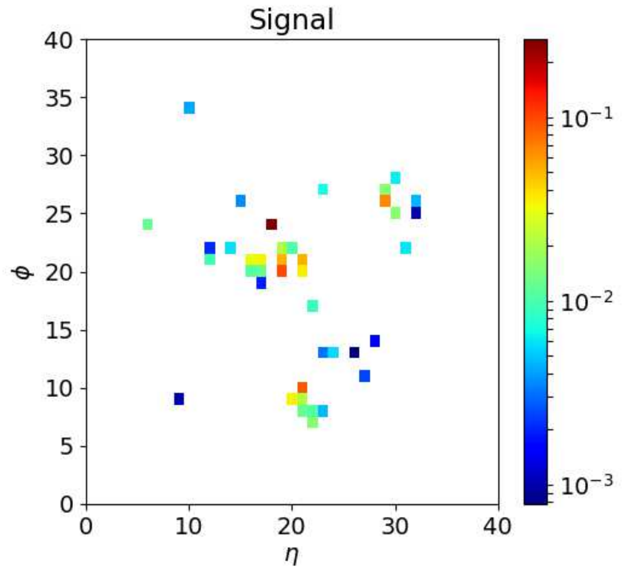
Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms

Jet Representations

Pixelized Image

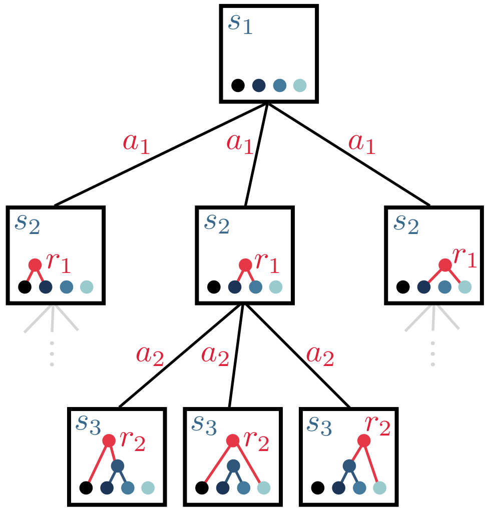
Calorimetry



[review in Kagan, arXiv 2020]

Hierarchical Tree

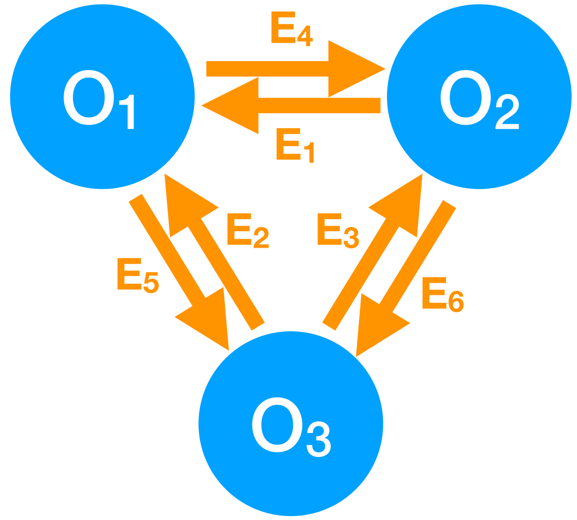
Binary Splittings



[e.g. Brehmer, Macaluso, Pappadopulo, Cranmer, NeurIPS 2020]

Graphs

Pairwise Interactions

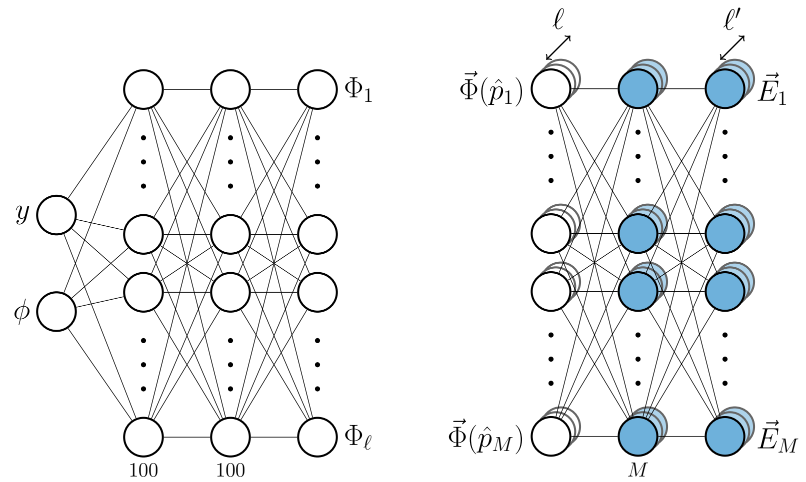


[e.g. Moreno, Cerri, Duarte, Newman, Nguyen, Periwal, Pierini, Serikova, Spiropulu, Vlimant, EPJC 2020]

Imposes implicit theoretical prior (typically a good thing!)
Influences choice of network architecture

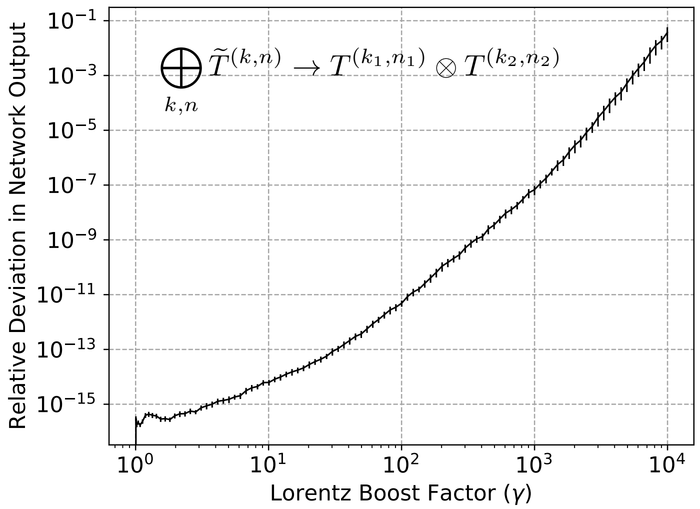
From Principles to Network Architectures

Permutation Equivariance



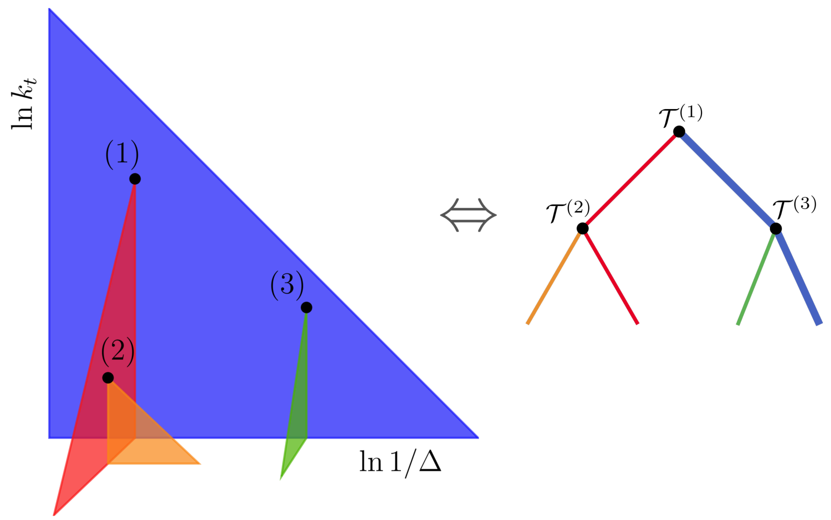
[Dolan, Ore, PRD 2021]

Lorentz Equivariance



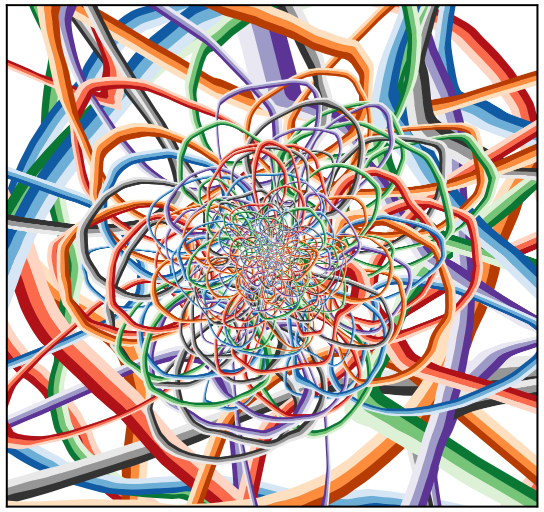
[Bogatskiy, Anderson, Offermann, Roussi, Miller, Kondor, arXiv 2020]

Lund Plane Emissions



[Dreyer, Qu, JHEP 2021]

Infrared and Collinear Safety



[Komiske, Metodiev, JDT, JHEP 2019]

...

Energy Flow Networks

Architecture designed around symmetries and interpretability

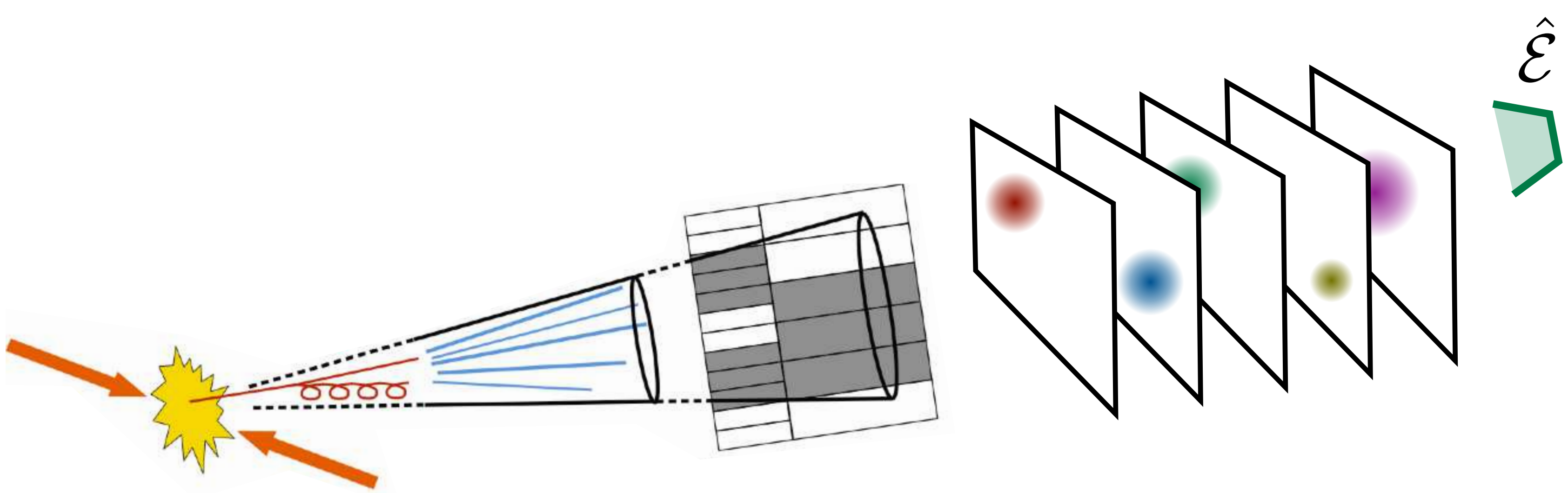
Latent space of dim ℓ

Permutation invariant

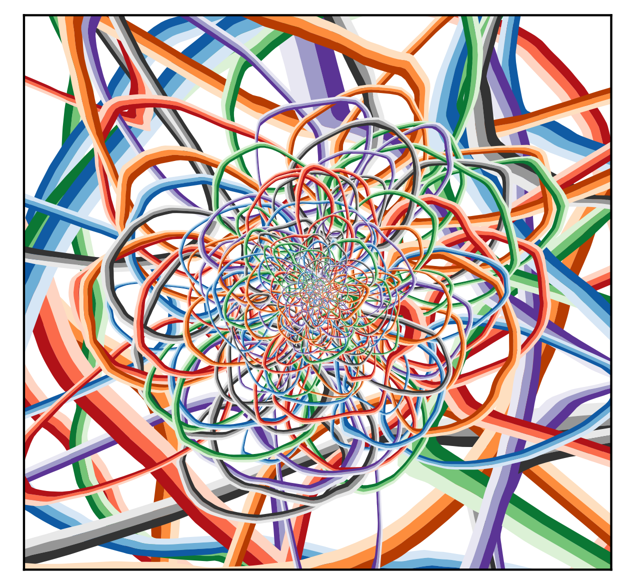
Linear weights (IRC safe)

$$S(\mathcal{J}) = F(V_1, V_2, \ldots, V_\ell) \qquad V_a(\mathcal{J}) = \sum_{i \in \mathcal{J}} E_i \, \Phi_a(\hat{n}_i)$$

Easy to plot!

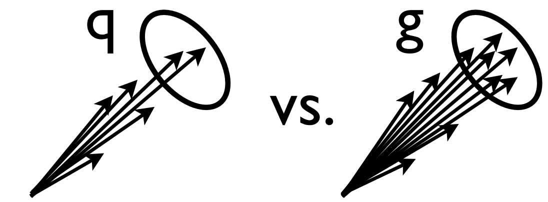


Learned ℓ = 256 latent space for quark/gluon discrimination:

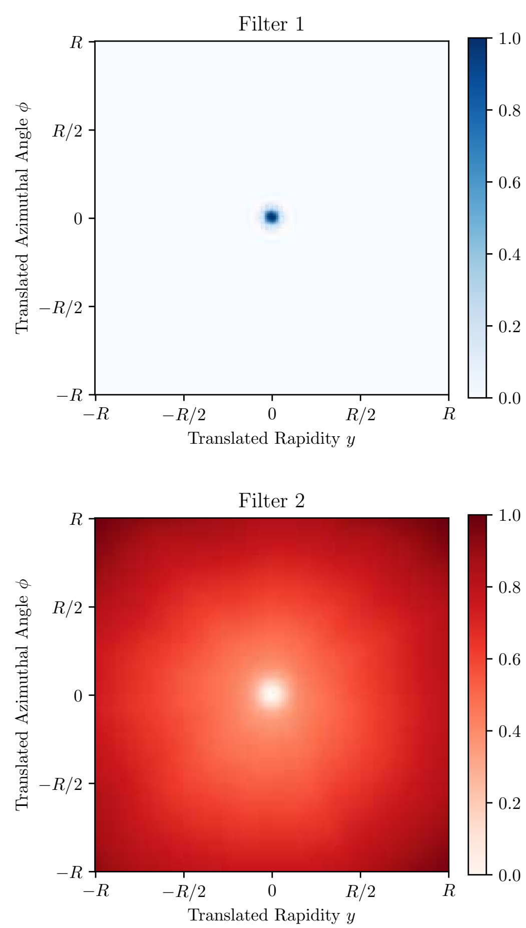


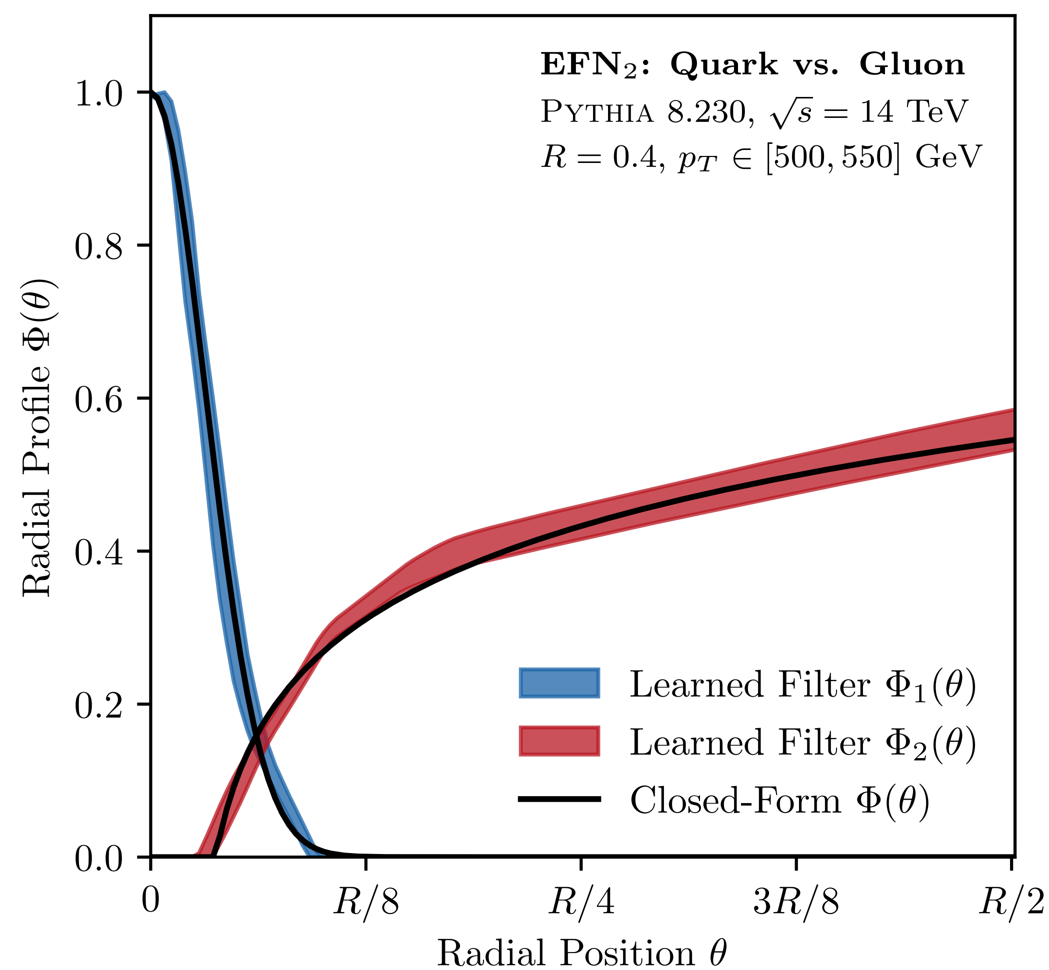
[Komiske, Metodiev, JDT, JHEP 2019; see also Komiske, Metodiev, JDT, JHEP 2018; code at energyflow.network; special case of Zaheer, Kottur, Ravanbakhsh, Poczos, Salakhutdinov, Smola, NIPS 2017; histogram pooling in Cranmer, Kreisch, Pisani, Villaescusa-Navarro, Spergel, Ho, ICLR SimDL 2021]

Learning from the Machine



For ℓ = 2, EFN learns radial moments: $\sum_{i \in \text{jet}} z_i f(\theta_i)$ cf. Angularities: $f(\theta) = \theta^\beta$

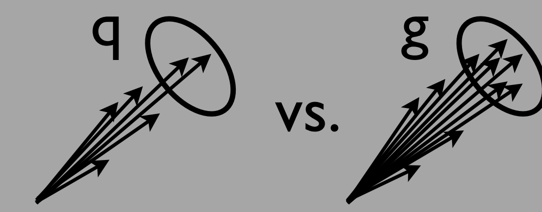




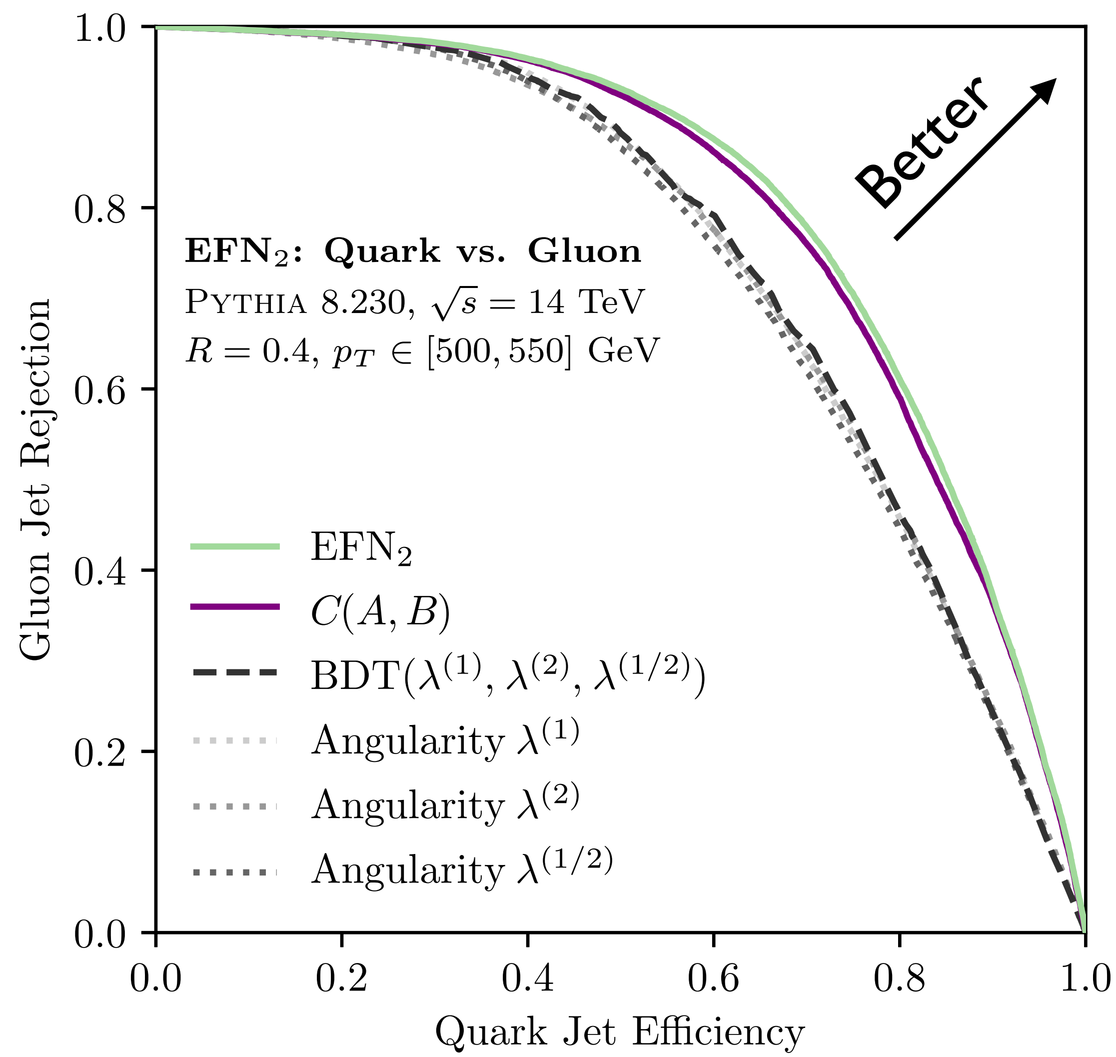
Traditional QCD observables emphasize homogeneous angular scaling
But EFN reveals that likelihood ratio exhibits collinear/wide-angle separation

[Komiske, Metodiev, JDT, JHEP 2019;
cf. Larkoski, JDT, Waalewijn, JHEP 2014; using Berger, Kucs, Sterman, PRD 2003; Ellis, Vermilion, Walsh, Hornig, Lee, JHEP 2010]

Learning from the Machine



For ℓ = 2, EFN learns radial moments: $\sum_{i \in \text{jet}} z_i f(\theta_i)$ cf. Angularities: f(θ) = θ^β



EFN outperformed a domain expert (i.e. me)

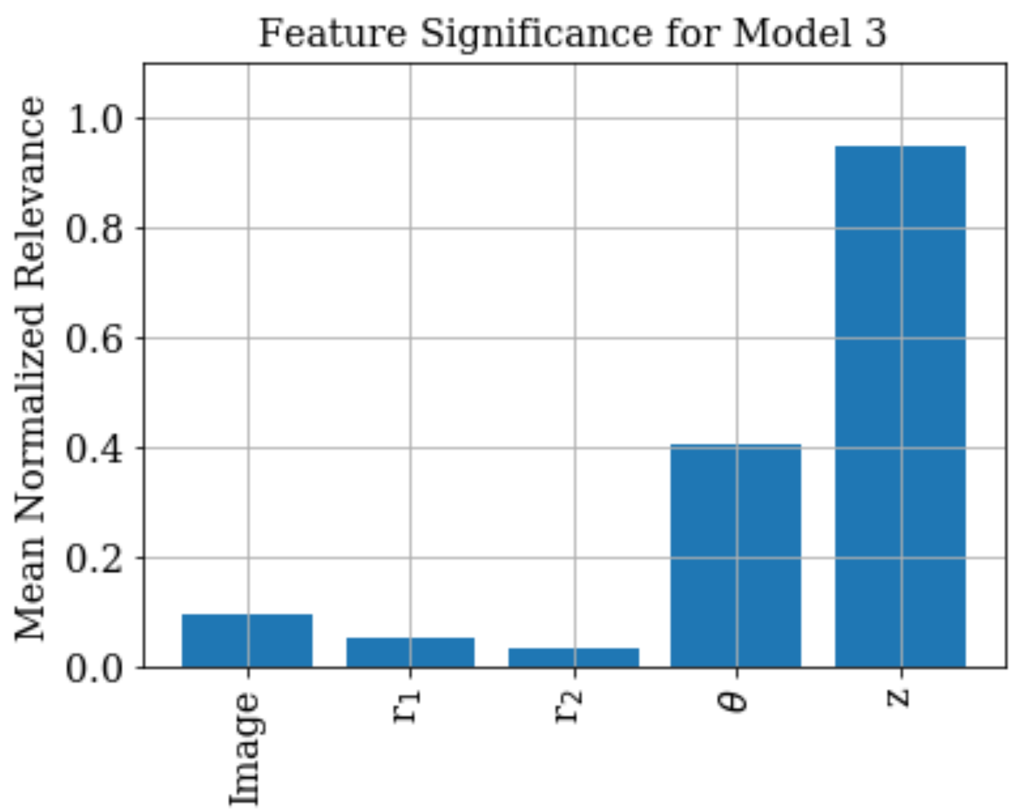
But we reverse engineered the machine (and learned something about QCD)

[Komiske, Metodiev, JDT, JHEP 2019;
cf. Larkoski, JDT, Waalewijn, JHEP 2014; using Berger, Kucs, Sterman, PRD 2003; Ellis, Vermilion, Walsh, Hornig, Lee, JHEP 2010]

More Network Architectures for Interpretability

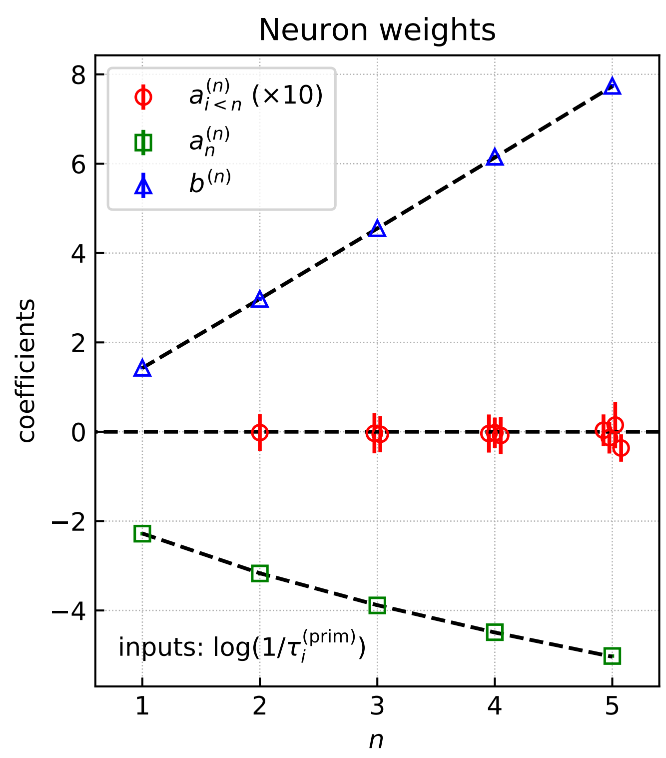
Rendering the black box more transparent

Input Feature Relevance



[Agarwal, Hay, Iashvili, Mannix, McLean, Morris, Rappoccio, Schubert, JHEP 2021]

Analytic Calculations



[Kasieczka, Marzani, Soyez, Stagnitto, JHEP 2020]

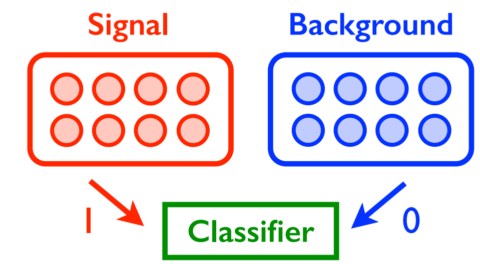
Imposing specific theoretical structures might reduce performance but might also yield better robustness/generalizability

When striving for “interpretable machine learning” we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

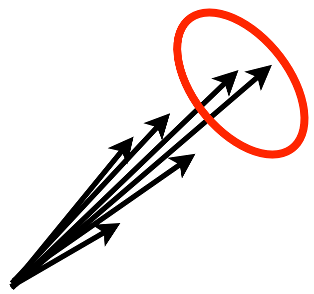
Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms



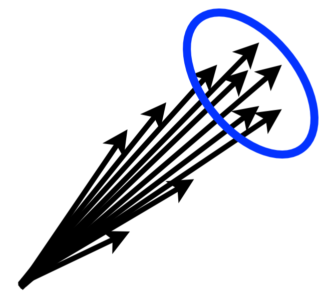
Quark/Gluon Classification

"Hello, World!" of Jet Physics

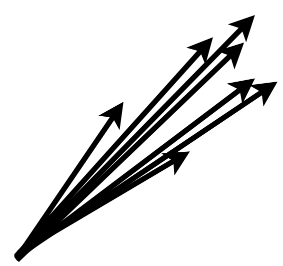


Quark
$C_q = 4/3$

vs.



Gluon
$C_g = 3 = 9/3$

Find h() such that

$$h(\text{Quark}) = 1$$

$$h(\text{Gluon}) = 0$$

Best you can do: $$h(\mathcal{J}) = \left(1 + \frac{p(\mathcal{J}|\text{G})}{p(\mathcal{J}|\text{Q})}\right)^{-1}$$

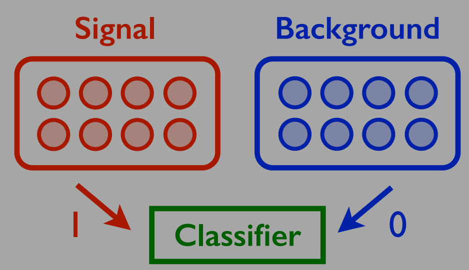
(Neyman-Pearson lemma)

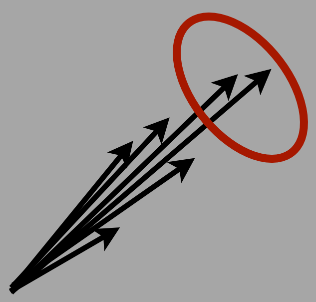
Likelihood ratio yields optimal binary classifier (and vice versa)

[see e.g. Gras, Höche, Kar, Larkoski, Lönnblad, Plätzer, Siódmok, Skands, Soyez, JDT, JHEP 2017; Komiske, Metodiev, Schwartz, JHEP 2017; Komiske, Metodiev, JDT, JHEP 2018]

Quark/Gluon Classification

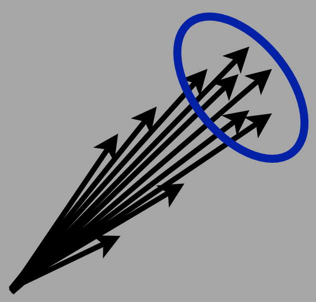
"Hello, World!" of Jet Physics





Quark
$C_q = 4/3$

vs.



Gluon
$C_g = 3 = 9/3$

What do you mean by "quark" and "gluon"?

Jets are clusters of colorless hadrons!

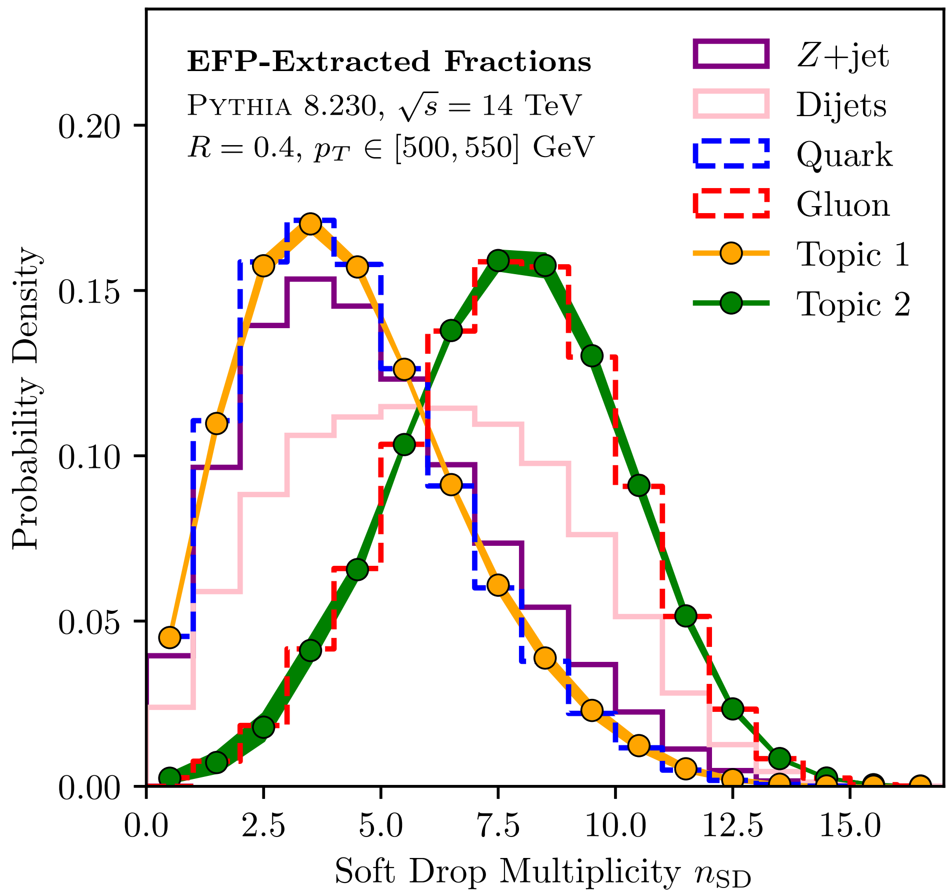
Parton shower "truth" is but a (useful) fiction!

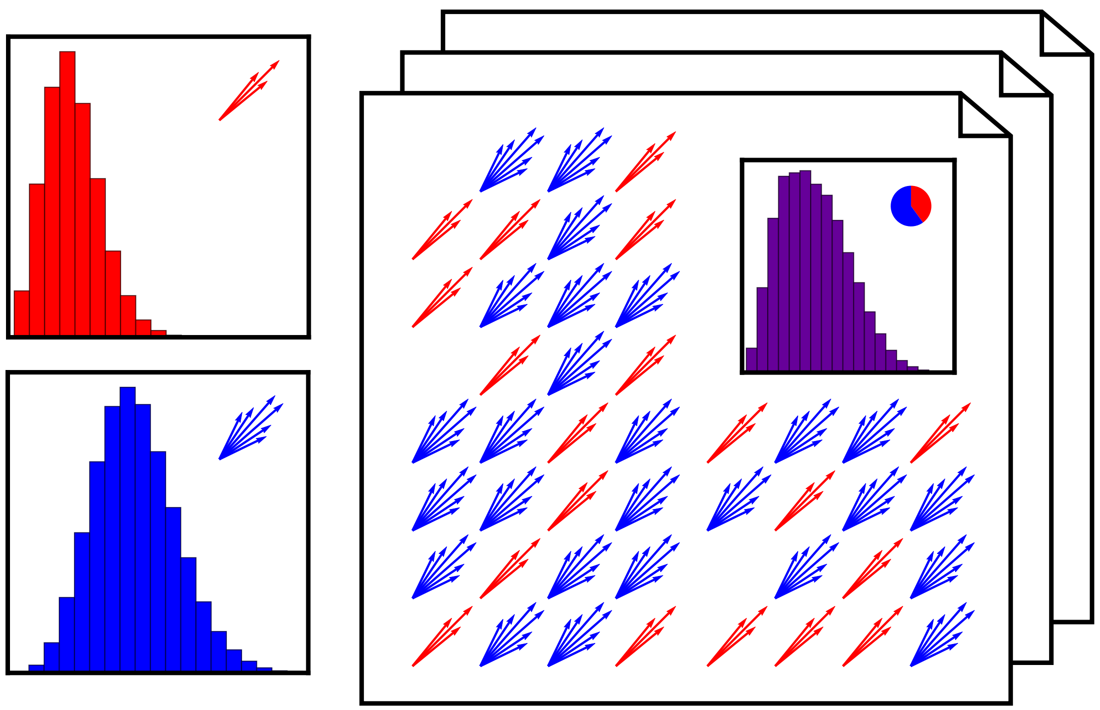
Likelihood ratio yields optimal binary classifier (and vice versa)

[see e.g. Gras, Höche, Kar, Larkoski, Lönnblad, Plätzer, Siódmok, Skands, Soyez, JDT, JHEP 2017; Komiske, Metodiev, Schwartz, JHEP 2017; Komiske, Metodiev, JDT, JHEP 2018]

Topic Modeling to Disentangle Data Samples

While you can't unambiguously label individual jets, you can extract quark and gluon distributions from hadron-level measurements





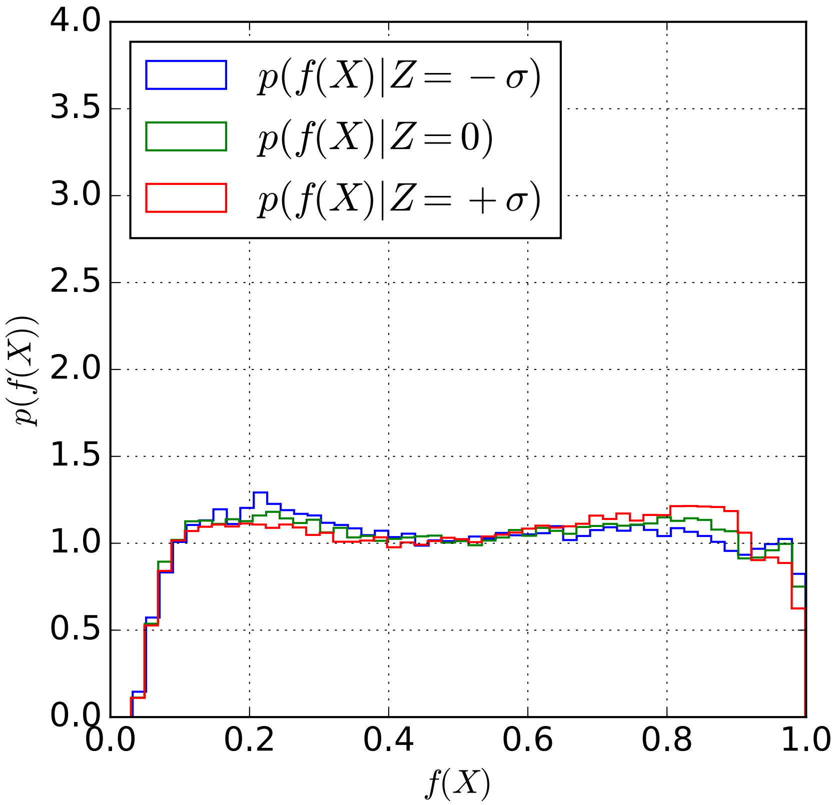
Key concept from natural language processing: "anchor words"

[Komiske, Metodiev, JDT, JHEP 2018; cf. ATLAS, PRD 2019; using Metodiev, Nachman, JDT, JHEP 2017; Metodiev, JDT, PRL 2018]
see also Blanchard, Flaska, Handy, Pozzi, Scott, PLMR 2013; Katz-Samuels, Blanchard, Scott, JMLR 2016]

Parameterized Data Samples

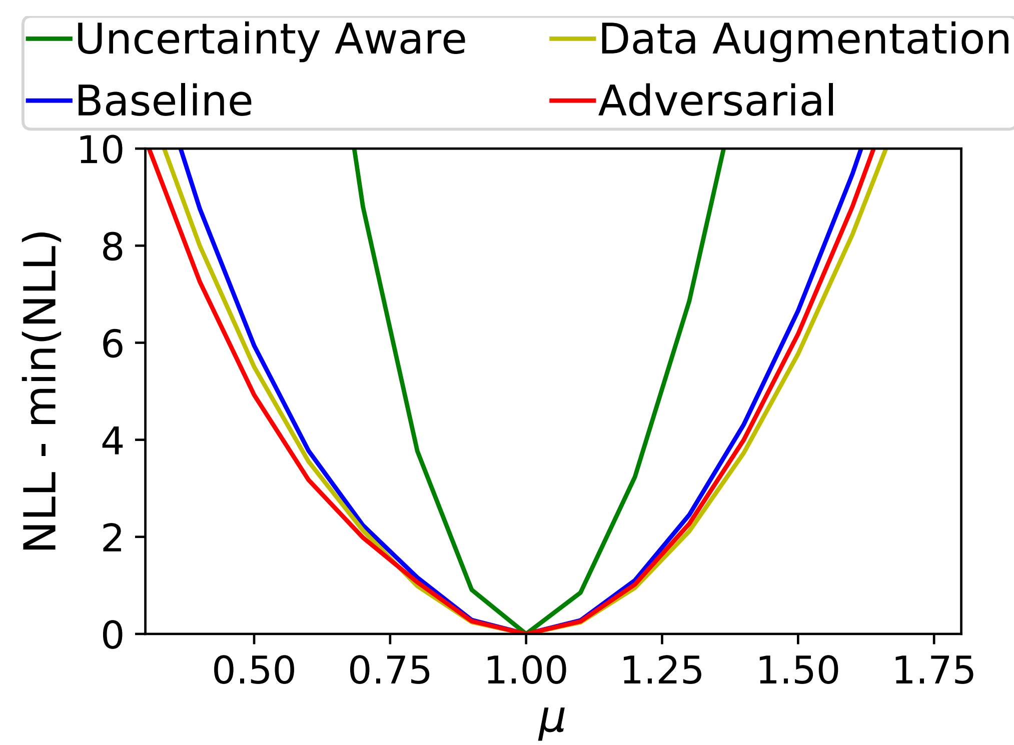
Incorporating nuisance parameters into training

Learn to Pivot…



[Louppe, Kagan, Cranmer, NeurIPS 2017]

…or Learn to Profile?



[Ghosh, Nachman, Whiteson, arXiv 2021]

Regardless of the approach, inspires renewed theoretical focus on uncertainty modeling for Monte Carlo generation

When striving for "interpretable machine learning" we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

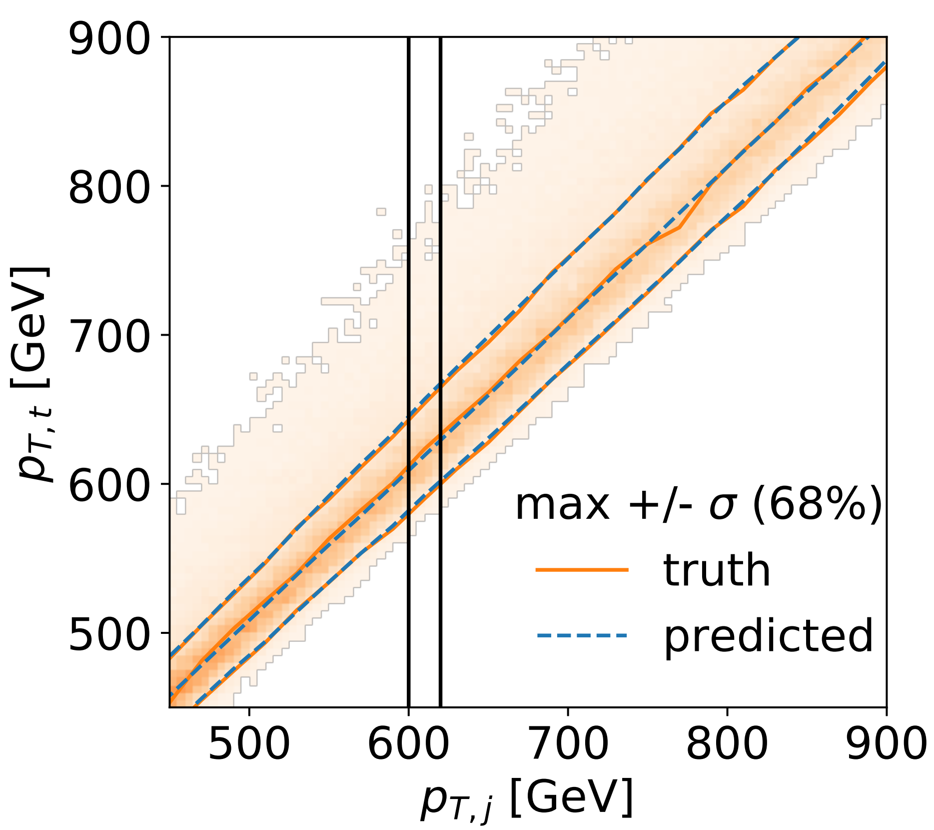
Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms

Loss Function for Bayesian Analyses

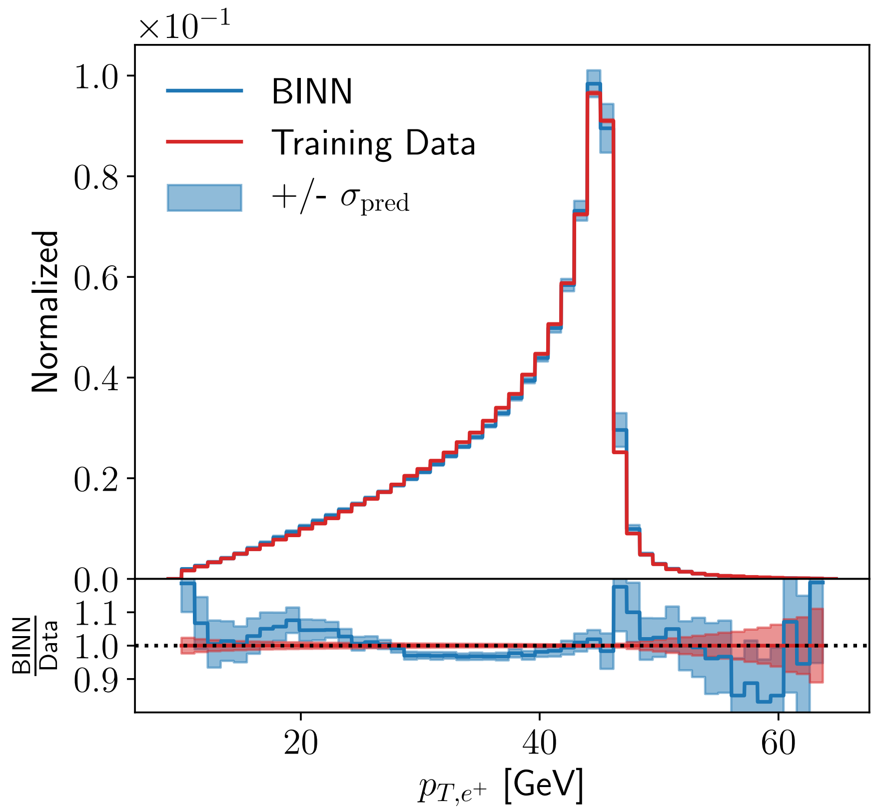
Treating network parameters as having a prior distribution

Jet Energy Scale



[Kasieczka, Luchmann, Otterpohl, Plehn, SciPost 2020]

Event Generation



[Bellagente, Haußmann, Luchmann, Plehn, arXiv 2021]

Use ELBO loss to capture both statistical and systematic uncertainties
Worth developing a frequentist version of this approach?

Loss Function for Bayesian Analyses

Treating network parameters as having a prior distribution

Nothing intrinsically wrong with Bayesian analyses,
but have to be aware of cases with strong prior dependence

$$\int d\theta\, \theta\, p(\theta|\text{data}) \neq \max_{\theta} p(\text{data}|\theta)$$

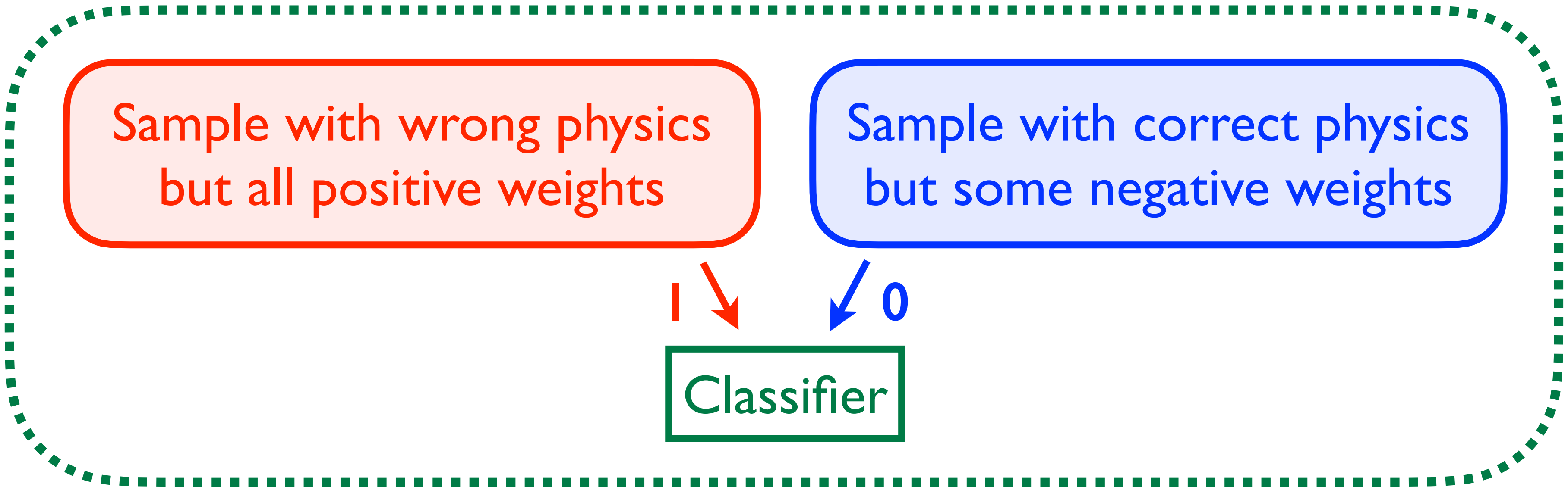
If needed, can treat neural networks as static objects
and calibrate them as if they were ordinary observables

[see further discussion in Cranmer, Pavez, Louppe, arXiv 2015; Nachman, SciPost 2020]

Use ELBO loss to capture both statistical and systematic uncertainties
Worth developing a frequentist version of this approach?

Training Paradigm for Preserved Uncertainties

When the goal is to maintain statistical properties



$$\Rightarrow w(x)$$

Plain reweighting yields all positive weights with correct asymptotic probability density

$$p_{\text{wrong}}(x) \times w(x) = p_{\text{correct}}(x)$$

Improved resampling through auxiliary neural network yields correct statistical uncertainties

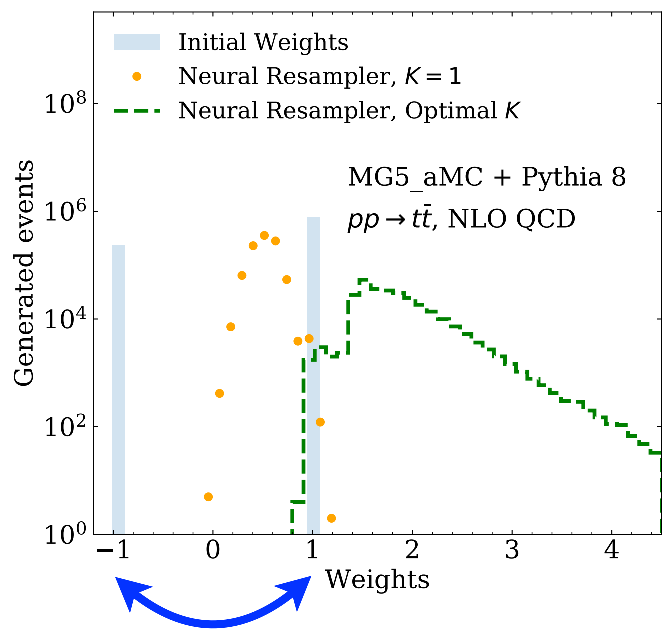
$$\left(\frac{\delta p}{p}\right)^2 = \frac{\langle w^2 \rangle}{\langle w \rangle^2}$$

[Nachman, JDT, PRD 2020; building on Andersen, Gutschow, Maier, Prestel, EPJC 2020]

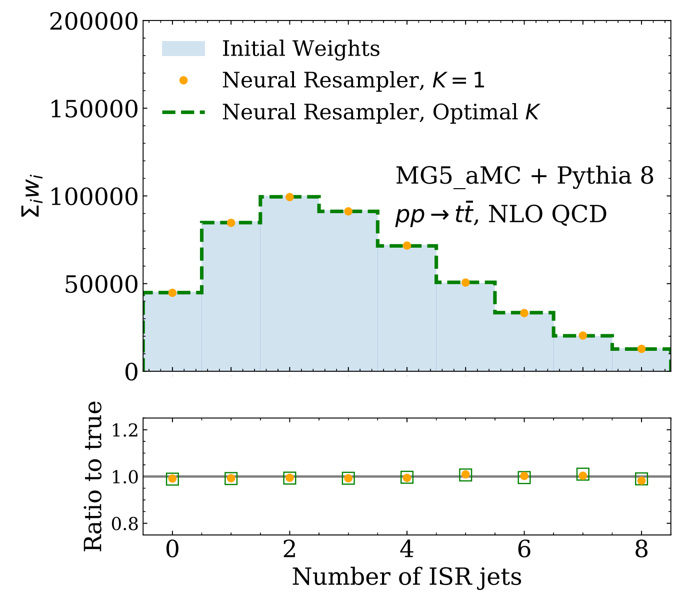
Training Paradigm for Preserved Uncertainties

Case Study in Jet Physics at Large Hadron Collider

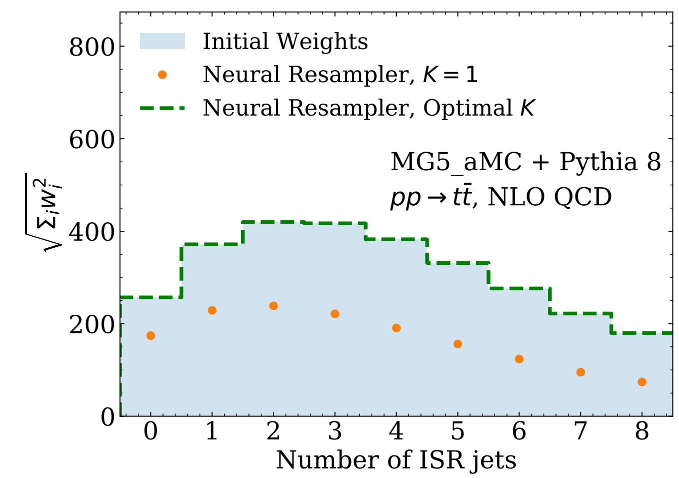
Original sample: large weight cancellations



Reweighting recovers desired distribution



Resampling recovers desired uncertainties



Improved resampling through auxiliary neural network yields correct statistical uncertainties

$$\left(\frac{\delta p}{p}\right)^2 = \frac{\langle w^2 \rangle}{\langle w \rangle^2}$$

[Nachman, JDT, PRD 2020; building on Andersen, Gutschow, Maier, Prestel, EPJC 2020]

Theory Perspective on ML for Jets

When striving for "interpretable machine learning" we are essentially hoping that likelihood ratios can be approximated via theoretically well-motivated forms

We can impose theoretical priors by judicious choice of network architecture that captures the underlying structures and symmetries of our problem

Machine learning methods can only be as robust and reliable as the data samples used for training

We are making progress towards uncertainty quantification, using more elaborate loss functions and training paradigms

Looking forward to your thoughts and discussion!

Backup Slides

Using All *Five* LHC Interaction Points

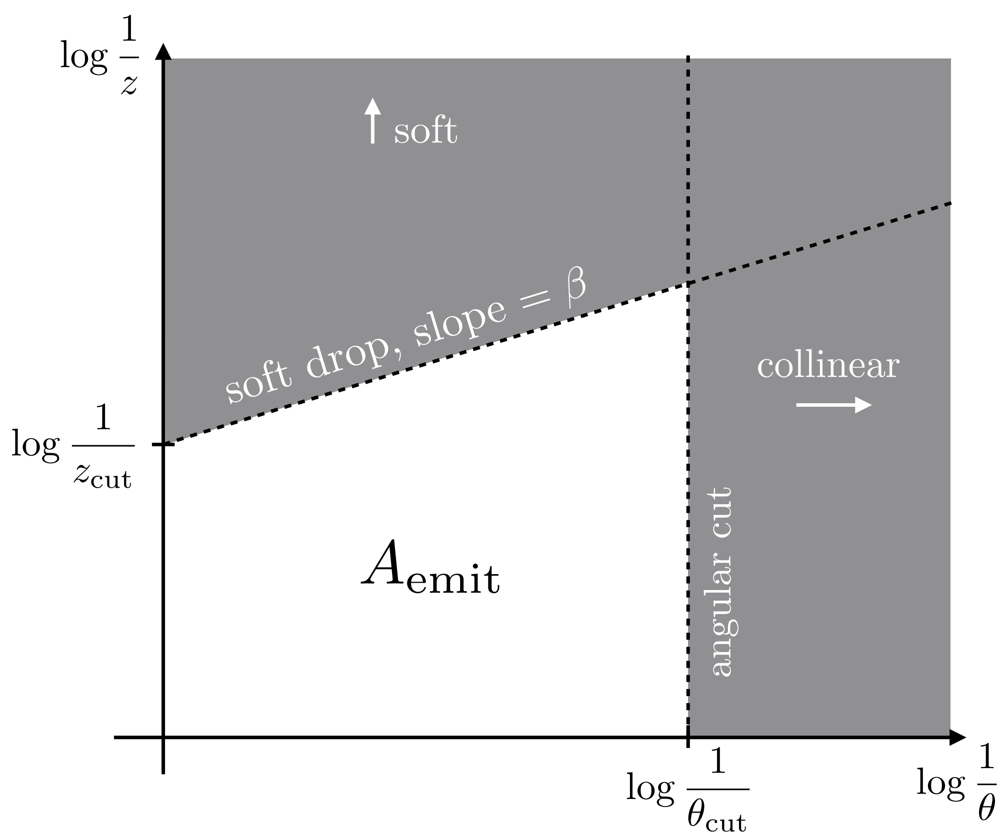
P1 P2 P5 P8 R1

[from my talk at ML4Jets 2018, with apologies]

E.g. Quark/Gluon Classification

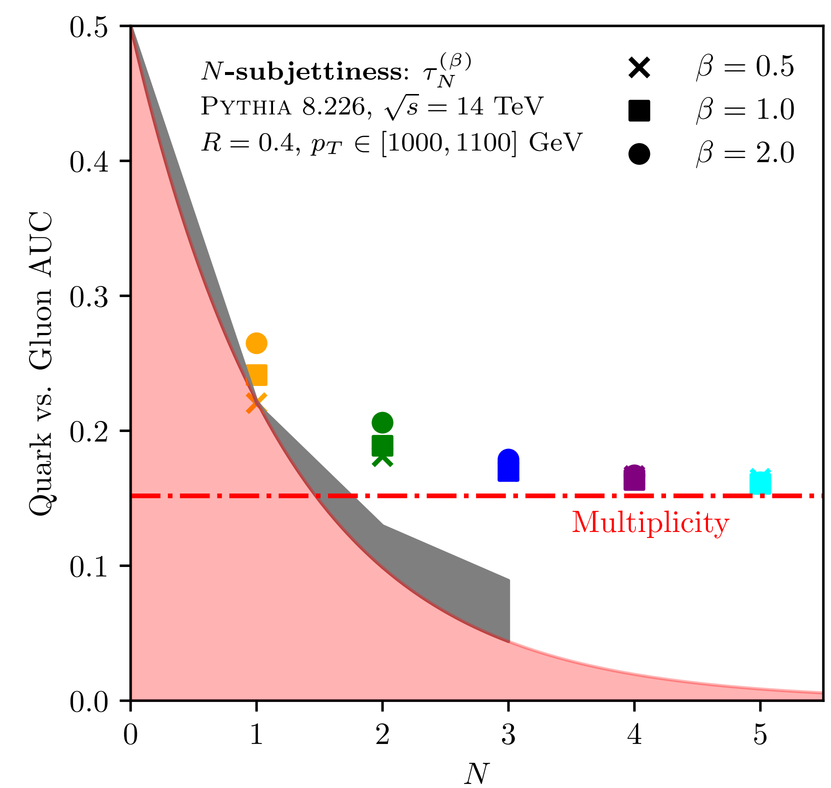
In various limits, likelihood ratio is monotonically related to…

Soft-dropped Multiplicity



[Frye, Larkoski, JDT, Zhou, JHEP 2017]

IRC Safe Multiplicity



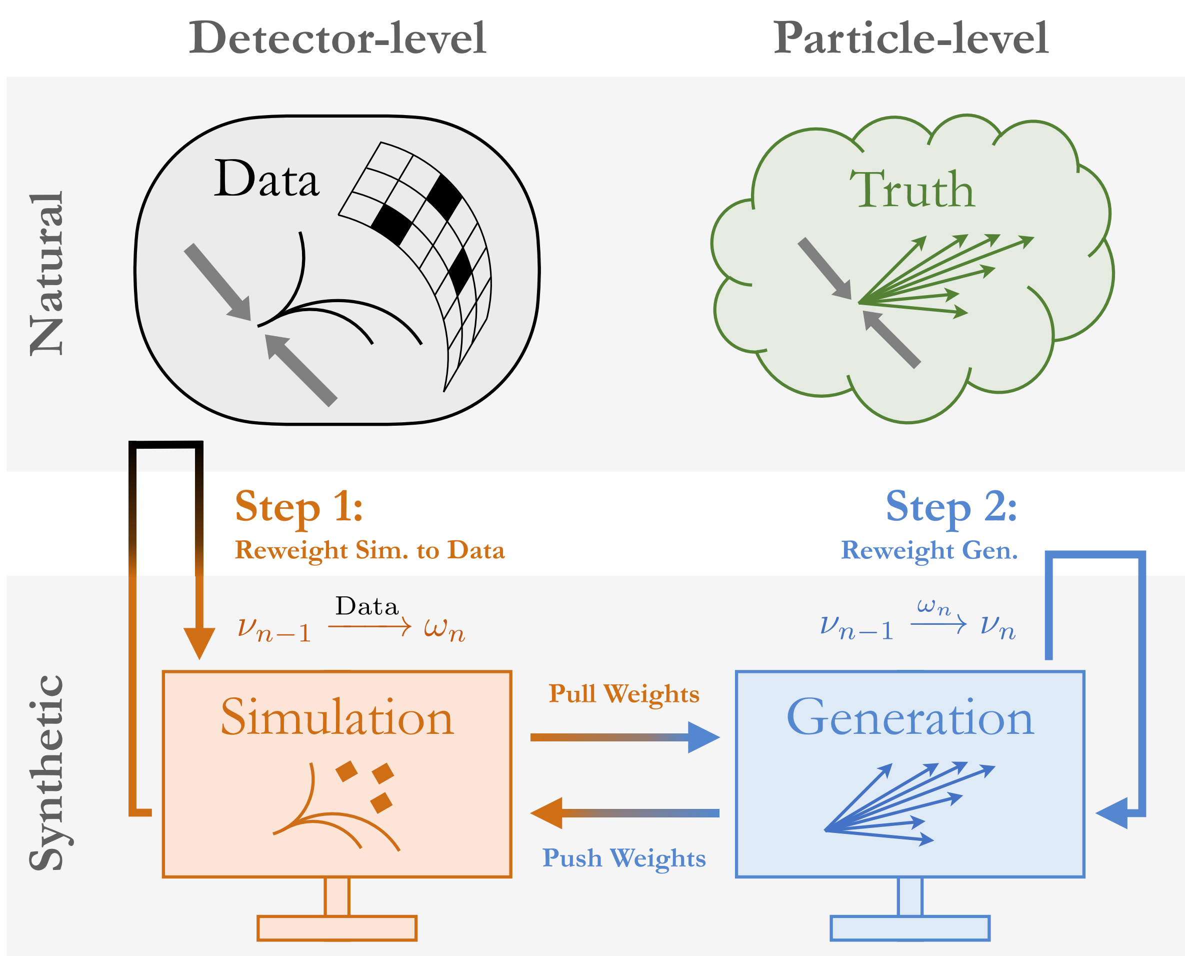
[Larkoski, Metodiev, JHEP 2019]

Away from these limits, the likelihood ratio is
not typically simple, elegant, or interpretable (but we can hope!)

E.g. Detector Unfolding

OmniFold

Multi-dimensional unbinned *detector corrections*
via iterated application of *likelihood ratio trick*



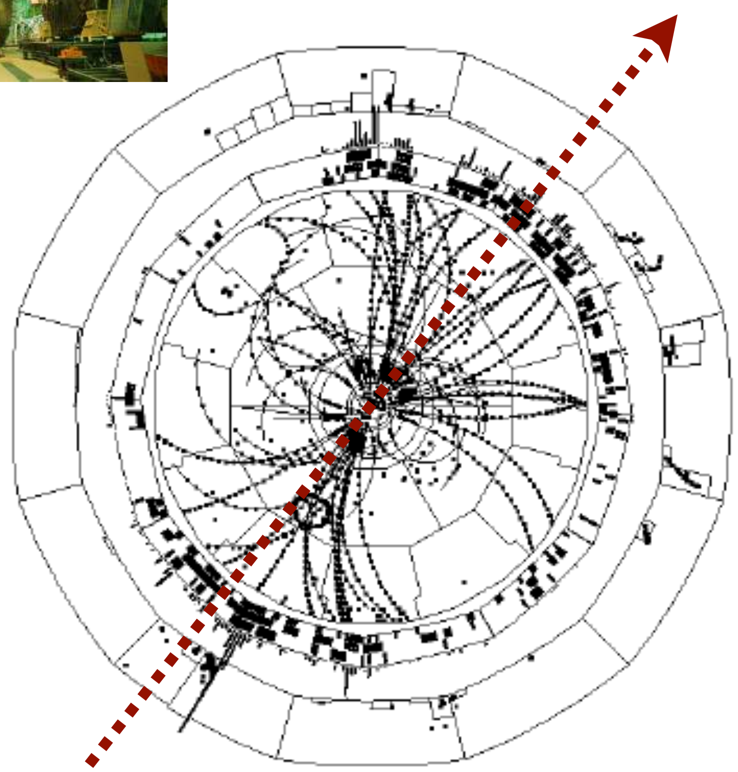
Use ML to compute reweighting factors

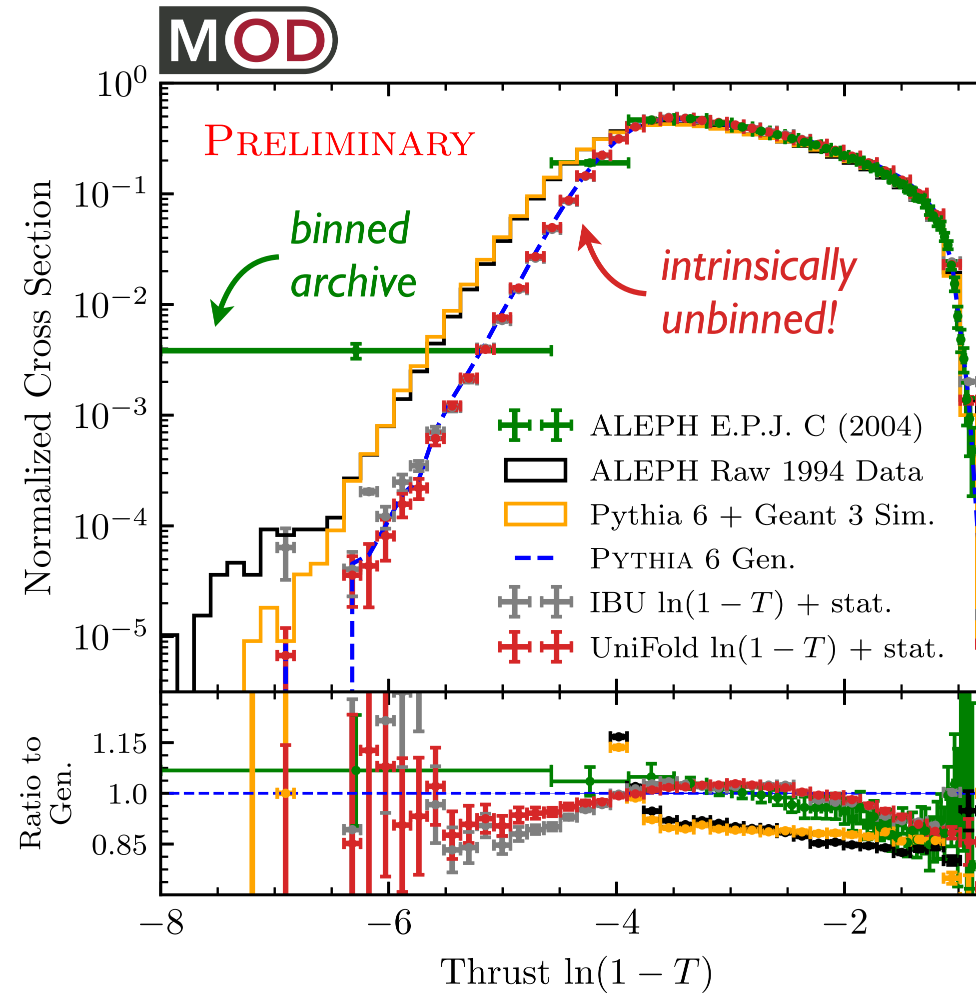
[Andreassen, Komiske, Metodiev, Nachman, JDT, PRL 2020; + Suresh, ICLR SimDL 2021;
Komiske, McCormack, Nachman, arXiv 2021; see unfolding comparison in Petr Baron, arXiv 2021]

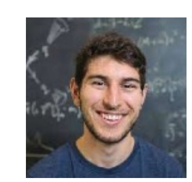



E.g. Detector Unfolding

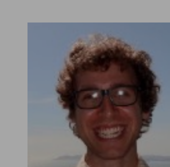
Back to the Future with ALEPH Archival Data







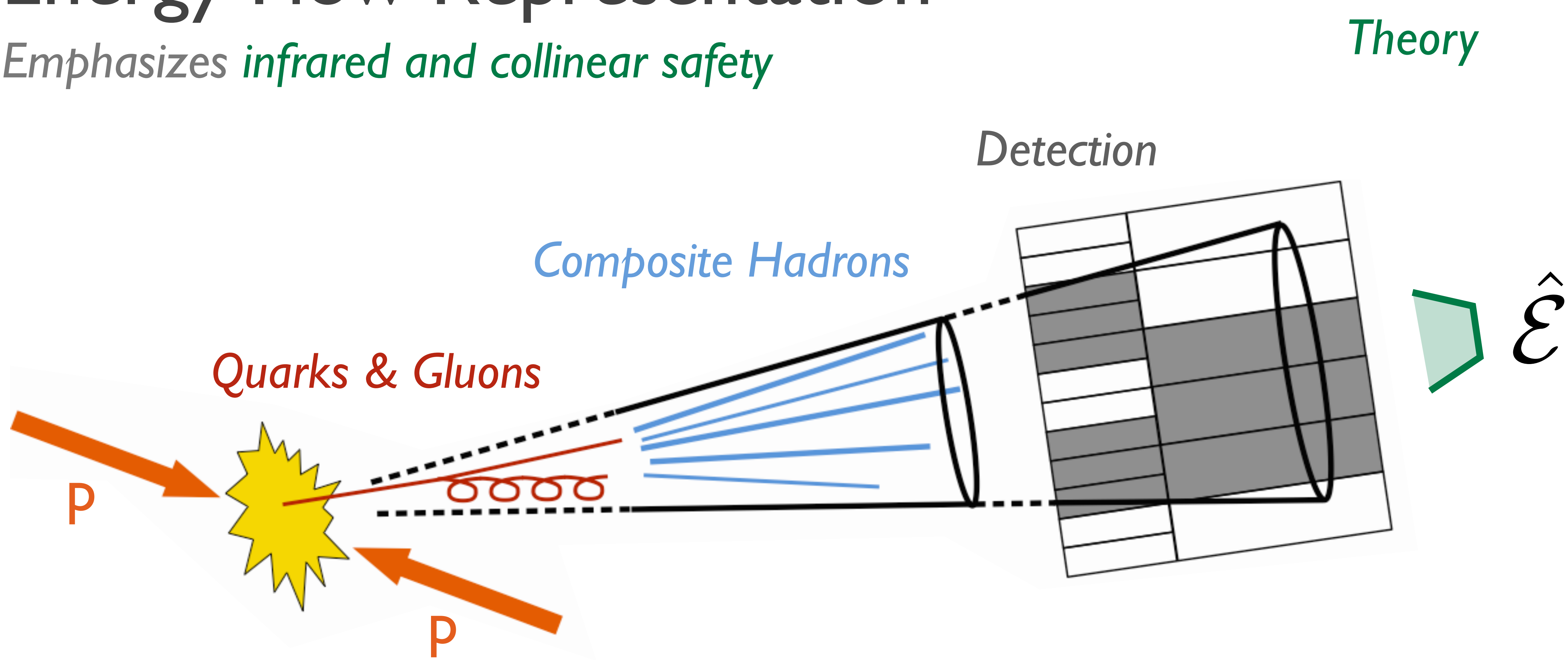
[talk by Badea, ICHEP 2020; cf. ALEPH, EPJC 2004]
[see also Badea, Baty, Chang, Innocenti, Maggi, McGinn, Peters, Sheng, JDT, Lee, PRL 2019; H1, DIS2021]

[Andreassen, Komiske, Metodiev, Nachman, JDT, PRL 2020; + Suresh, ICLR SimDL 2021;
Komiske, McCormack, Nachman, arXiv 2021; see unfolding comparison in Petr Baron, arXiv 2021]

Energy Flow Representation

Emphasizes infrared and collinear safety



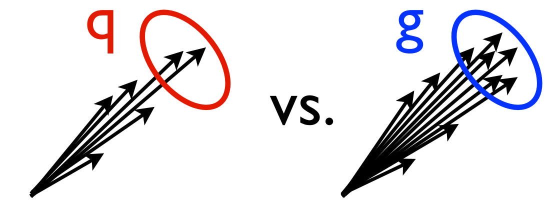
Energy Flow:

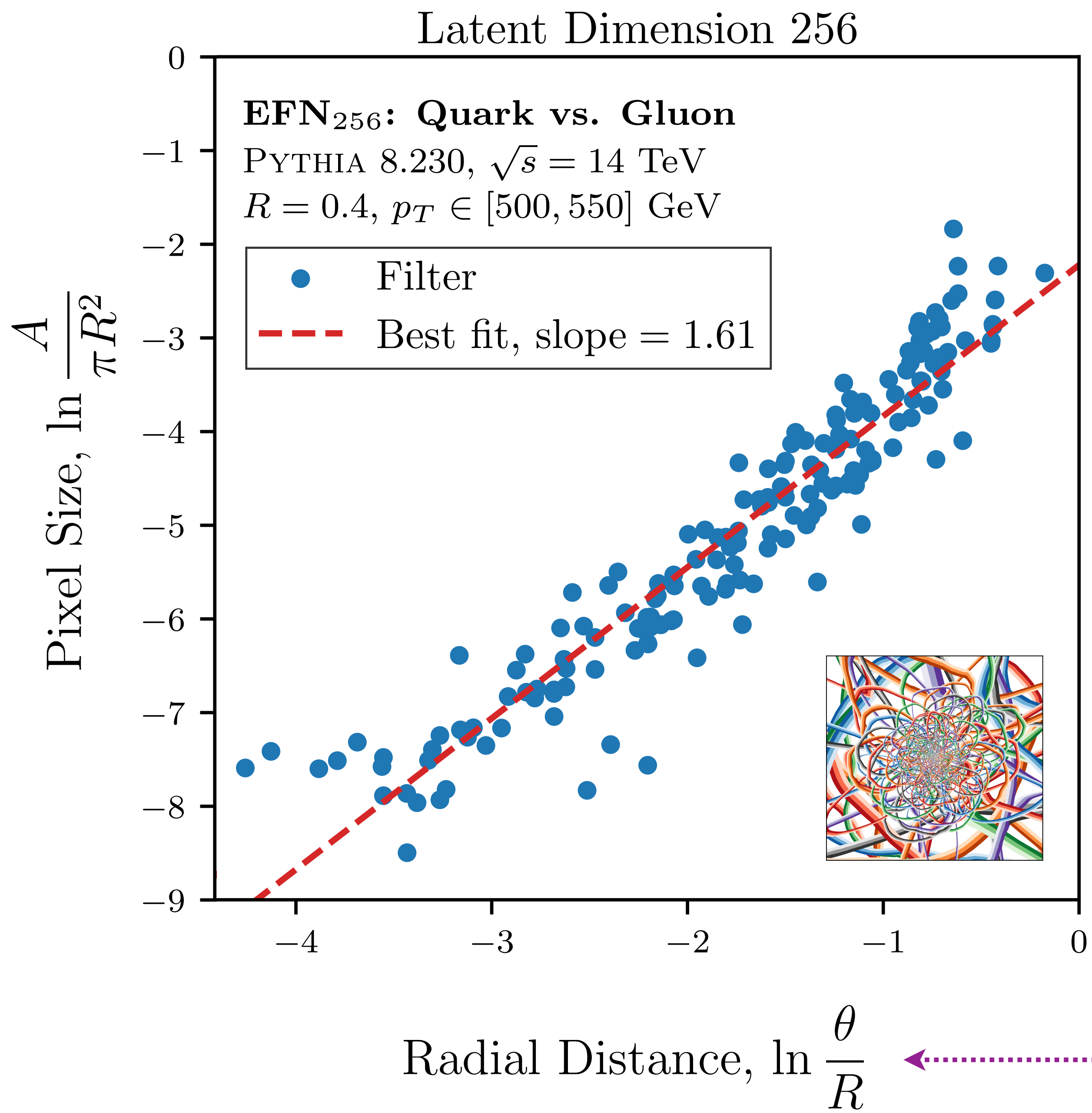
Robust to hadronization and detector effects
Well-defined for massless gauge theories

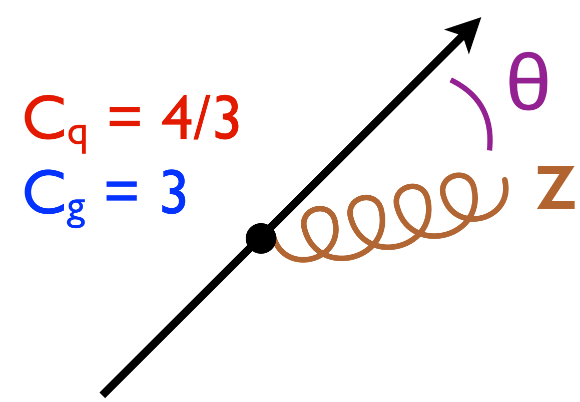
$$\hat{\mathcal{E}} \simeq \lim_{t\to\infty} \hat{n}_i T^{0i}(t, vt\hat{n})$$

[see e.g. Sveshnikov, Tkachov, PLB 1996; Hofman, Maldacena, JHEP 2008; Mateu, Stewart, JDT, PRD 2013; Belitsky, Hohenegger, Korchemsky, Sokatchev, Zhiboedov, PRL 2014; Chen, Moult, Zhang, Zhu, PRD 2020]
[complementary perspective on IRC unsafe information in Chakraborty, Lim, Nojiri, Takeuchi, JHEP 2020]

Machine Learning Collinear QCD





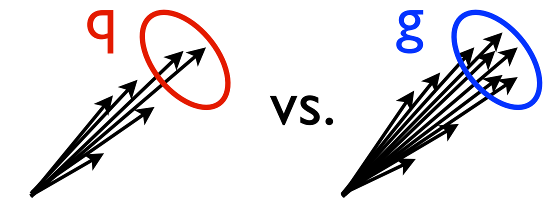


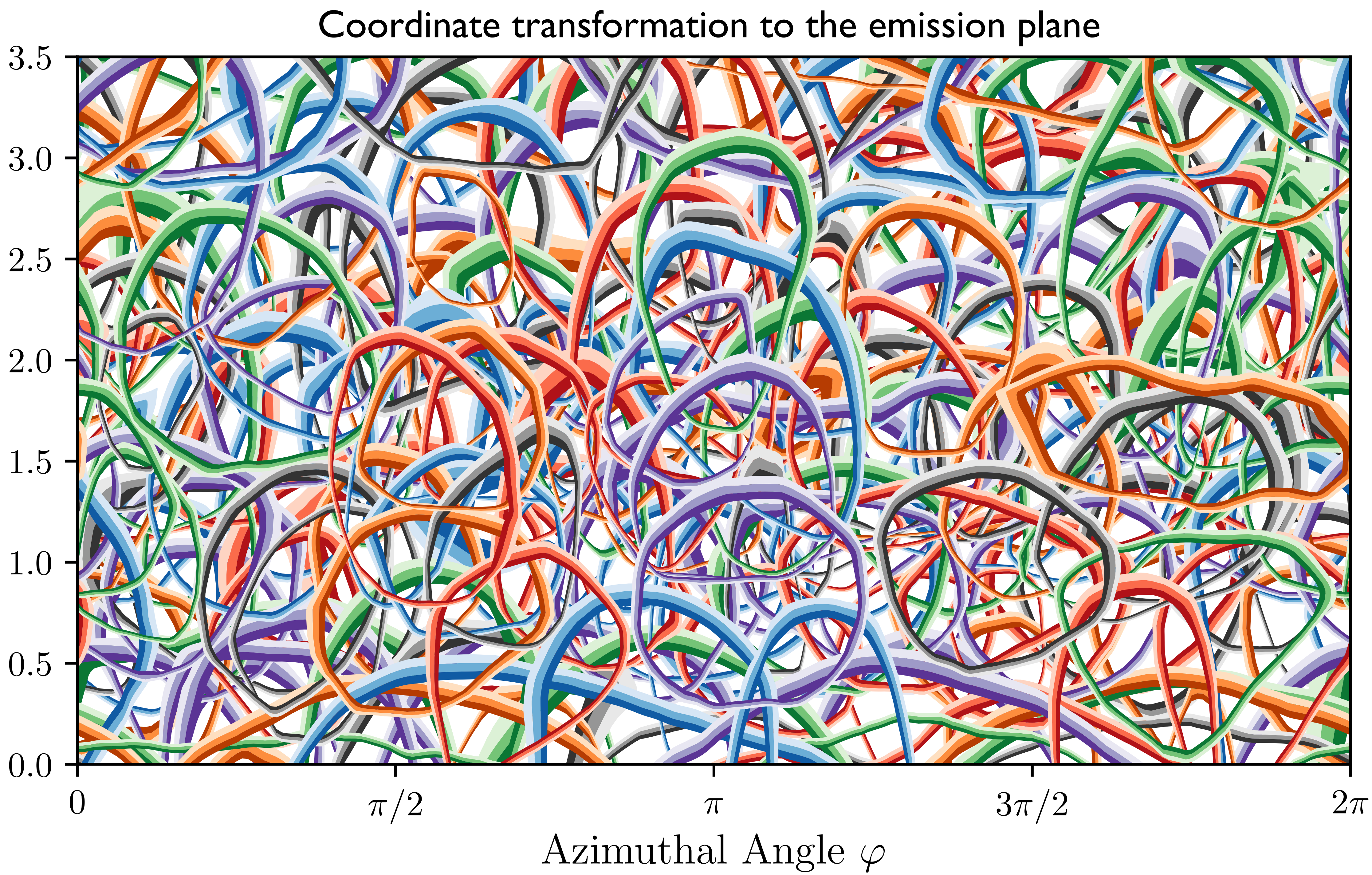
$$\mathrm{d}P_{i \to ig} \simeq \frac{2\alpha_s}{\pi} C_i \frac{\mathrm{d}\theta}{\theta} \frac{\mathrm{d}z}{z}$$

Collinear Soft

[Komiske, Metodiev, JDT, JHEP 2019]

En Route to the Lund Plane





[Komiske, Metodiev, JDT, JHEP 2019; see also Dreyer, Salam, Soyez, JHEP 2018]