Uncertainties in Generative AI

Tilman Plehn

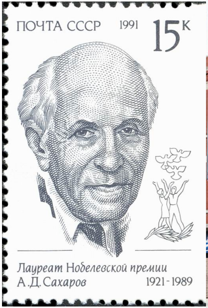
Universität Heidelberg

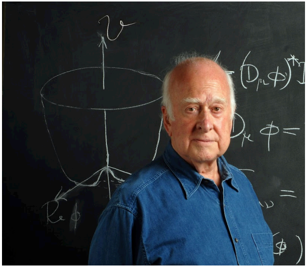
Japan-Heidelberg Workshop, October 2024

Modern LHC physics

Classic motivation

- dark matter?
- matter vs antimatter?
- origin of Higgs boson?





Modern LHC physics

Classic motivation

- dark matter?
- matter vs antimatter?
- origin of Higgs boson?

LHC physics

- fundamental questions
- huge data set
- first-principle, precision simulations
- complete uncertainty control

Successful past

- measurements of total rates
- analyses inspired by simulation
- model-driven Higgs discovery

Modern LHC physics

Classic motivation

- dark matter?
- matter vs antimatter?
- origin of Higgs boson?

LHC physics

- fundamental questions
- huge data set
- first-principle, precision simulations
- complete uncertainty control

Successful past

- measurements of total rates
- analyses inspired by simulation
- model-driven Higgs discovery

First-principle, precision simulations

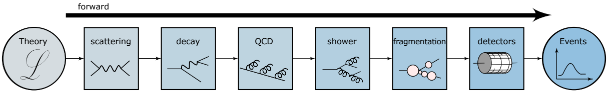
- start with Lagrangian
- calculate scattering using QFT
- simulate collisions
- simulate detectors

$\rightarrow$ LHC collisions in virtual worlds

BSM searches

- compare simulations and data
- understand LHC data systematically
- infer underlying theory [SM or BSM]
- publish useable results

$\rightarrow$ Lots of data science...

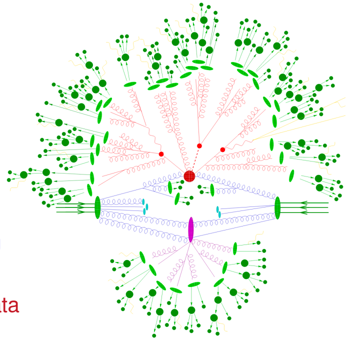


LHC Theory

Turning data into knowledge

- QFT
  start with Lagrangian
  generate Feynman diagrams
- compute hard scattering
  compute decays
  compute jet radiation
- partons inside protons
  hadron-level QCD

$\rightarrow$ First-principle simulations, not modeling



HL-LHC: optimal inference with 10×more data

- statistical improvement $\sqrt{10} > 3$
- rate over phase space to $< 0.1\%$
- SBI starts with Simulation $\leftrightarrow$ theory
- speed the key to precision

$\rightarrow$ MadNIS & Co

Regression with uncertainties

Calibration function for ATLAS calorimeter

- energy measurement for cluster/jet j

$$\langle E \rangle = \int dE \; E \; p(E)$$

- weighted by reproducing training data $p(\theta|T)$

$$p(E) = \int d\theta \; p(E|\theta) \; p(\theta|T)$$

$\to$ θ-distributions defining Bayesian NN

Regression with uncertainties

Calibration function for ATLAS calorimeter

- energy measurement for cluster/jet j

$$\langle E \rangle = \int dE \; E \; p(E)$$

- weighted by reproducing training data $p(\theta|T)$

$$p(E) = \int d\theta \; p(E|\theta) \; p(\theta|T)$$

$\to$ θ-distributions defining Bayesian NN

Variational approximation

- definition of training [think $q(\theta)$ as Gaussian with mean and width]

$$p(E) = \int d\theta \; p(E|\theta) \; p(\theta|T) \approx \int d\theta \; p(E|\theta) \; q(\theta)$$

- similarity through minimal KL-divergence [Bayes' theorem to remove unknown posterior]

$$\begin{aligned} D_\text{KL}[q(\theta), p(\theta|T)] &= \int d\theta \; q(\theta) \; \log \frac{q(\theta)}{p(\theta|T)} \\ &= \int d\theta \; q(\theta) \; \log \frac{q(\theta)p(T)}{p(T|\theta)p(\theta)} \\ &= D_\text{KL}[q(\theta), p(\theta)] - \int d\theta \; q(\theta) \; \log p(T|\theta) + \log p(T) \int d\theta \; q(\theta) \end{aligned}$$

Regression with uncertainties

Calibration function for ATLAS calorimeter

- energy measurement for cluster/jet j

$$\langle E\rangle = \int dE \; E \; p(E)$$

- weighted by reproducing training data $p(\theta|T)$

$$p(E) = \int d\theta \; p(E|\theta) \; p(\theta|T)$$

$\to$ θ-distributions defining Bayesian NN

Variational approximation

- definition of training [think $q(\theta)$ as Gaussian with mean and width]

$$p(E) = \int d\theta \; p(E|\theta) \; p(\theta|T) \approx \int d\theta \; p(E|\theta) \; q(\theta)$$

- similarity through minimal KL-divergence [Bayes' theorem to remove unknown posterior]

$$\begin{aligned} D_{\text{KL}}[q(\theta), p(\theta|T)] &= \int d\theta \; q(\theta) \; \log \frac{q(\theta)}{p(\theta|T)} \\ &= \int d\theta \; q(\theta) \; \log \frac{q(\theta)p(T)}{p(T|\theta)p(\theta)} \\ &\approx D_{\text{KL}}[q(\theta), p(\theta)] - \int d\theta \; q(\theta) \; \log p(T|\theta) \equiv \mathcal{L} \end{aligned}$$

$\to$ Two-term loss: likelihood + prior

Relation to deterministic networks

Regularization

- BNN loss

$$\mathcal{L} = -\int d\theta \; q(\theta) \; \log p(T|\theta) + D_{\text{KL}}[q(\theta), p(\theta)]$$

$$= -\int d\theta \; q(\theta) \; \log p(T|\theta) + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \cdots$$

- deterministic network

$$q(\theta) = \delta(\theta - \theta_0) \quad \Rightarrow \quad \mathcal{L} \approx -\log p(T|\theta_0) + \frac{(\theta_0 - \mu_p)^2}{2\sigma_p^2}$$

$\to$ Likelihood with L2-regularization

Relation to deterministic networks

Regularization

- BNN loss

$$\mathcal{L} = -\int d\theta\; q(\theta)\; \log p(T|\theta) + D_{\text{KL}}[q(\theta), p(\theta)]$$

$$= -\int d\theta\; q(\theta)\; \log p(T|\theta) + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \dots$$

- deterministic network

$$q(\theta) = \delta(\theta - \theta_0) \quad \Rightarrow \quad \mathcal{L} \approx -\log p(T|\theta_0) + \frac{(\theta_0 - \mu_p)^2}{2\sigma_p^2}$$

$\to$ Likelihood with L2-regularization

Dropout

- Bernoulli weights

$$q(\theta) \to q(x) = \rho^x (1-\rho)^{1-x} \Big|_{x=0,1} \qquad \text{with} \quad \theta = x\theta_0$$

$\to$ Regularized likelihood with dropout

Statistics vs systematics

Network evaluation

- expectation value using trained network $q(\theta)$

$$\langle E \rangle = \int dE d\theta \; E \; p(E|\theta) \; q(\theta) \equiv \int d\theta \; q(\theta) \overline{E}(\theta) \qquad \text{with} \qquad \overline{E}(\theta) = \int dE \; E \; p(E|\theta)$$

- corresponding variance

$$\begin{aligned} \sigma_{\text{tot}}^2 &= \int dE d\theta \; (E - \langle E \rangle)^2 \; p(E|\theta) \; q(\theta) \\ &= \int d\theta \; q(\theta) \left[\overline{E^2}(\theta) - 2 \langle E \rangle \overline{E}(\theta) + \langle E \rangle^2 \right] \\ &= \int d\theta \; q(\theta) \left[\overline{E^2}(\theta) - \overline{E}(\theta)^2 + \left(\overline{E}(\theta) - \langle E \rangle \right)^2 \right] \equiv \sigma_{\text{syst}}^2 + \sigma_{\text{stat}}^2 \end{aligned}$$

Two uncertainties

- statistical — vanishing for perfect training: $q(\theta) \to \delta(\theta - \theta_0)$

$$\sigma_{\text{stat}}^2 = \int d\theta \; q(\theta) \left[\overline{E}(\theta) - \langle E \rangle \right]^2 = \left[\overline{E}(\theta_0) - \langle E \rangle \right]^2$$

- systematic — vanishing for perfect data: $p(E|\theta) \to \delta(E - E_0)$

$$\sigma_{\text{syst}}^2 = \int d\theta \; q(\theta) \left[\overline{E^2}(\theta) - \overline{E}(\theta)^2 \right]$$

Transforming LHC analyses

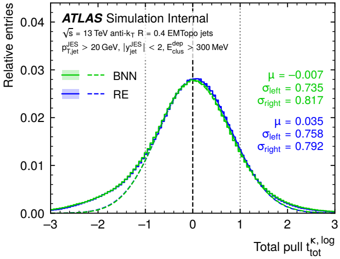
Calibration with uncertainties [Vogel, Loch, TP,...]

- interpretable topo-cluster phase space x
- learned calibration

$$\mathcal{R}^{\text{BNN}}(x) = \mathcal{R}(x) = \frac{E^{\text{EM}}(x)}{E^{\text{dep}}(x)}$$

- learned uncertainties $\Delta\mathcal{R}(x)$ [Nina Elmer's poster]
  Bayesian neural networks vs repulsive ensembles

$\to$ error vs data spread checked by pull



Transforming LHC analyses

Calibration with uncertainties [Vogel, Loch, TP,...]

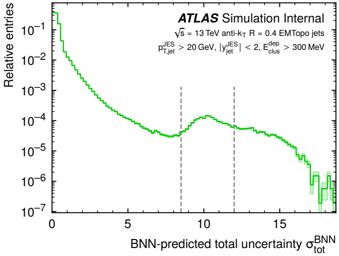
- interpretable topo-cluster phase space x
- learned calibration

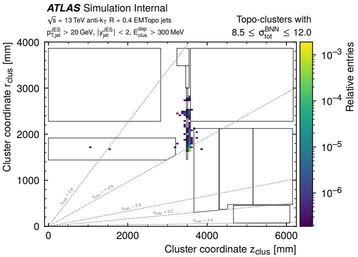
$$\mathcal{R}^{\text{BNN}}(x) = \mathcal{R}(x) = \frac{E^{\text{EM}}(x)}{E^{\text{dep}}(x)}$$

- learned uncertainties $\Delta\mathcal{R}(x)$ [Nina Elmer's poster]
  Bayesian neural networks vs repulsive ensembles

$\to$ error vs data spread checked by pull

$\to$ Understand data using uncertainties



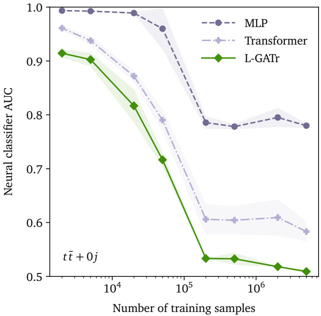


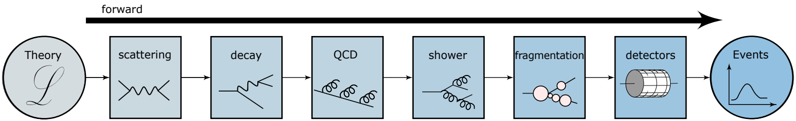
Generative AI

Forward simulations

- learn phase space density
  sample Gaussian $\to$ phase space
- Variational Autoencoder
  $\to$ low-dimensional physics
- Generative Adversarial Network
  $\to$ generator trained by classifier
- Normalizing Flow/Diffusion
  $\to$ (bijective) mapping [INN]
- JetGPT, ViT
  $\to$ non-local structures
- Equivariant L-GATr
  $\to$ guarantee Lorentz symmetry

$\to$ Combinations: equivariant transformer CFM...





Generative Uncertainties

Unsupervised Bayesian networks

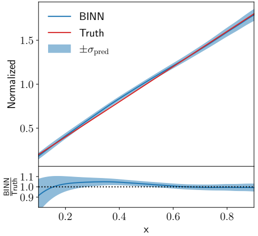
- data: event sample [points in 2D space]
  learn phase space density
  standard distribution in latent space [Gaussian]
  sample from latent space
- Bayesian version
  allow weight distributions
  learn uncertainty map
- 2D wedge ramp

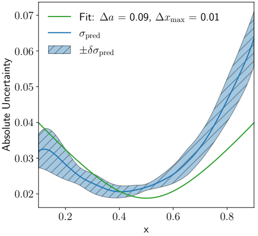
$$p(x) = ax + b = ax + \frac{1 - \frac{a}{2}(x_{\max}^2 - x_{\min}^2)}{x_{\max} - x_{\min}}$$

$$(\Delta p)^2 = \left(x - \frac{1}{2}\right)^2 (\Delta a)^2 + \left(1 + \frac{a}{2}\right)^2 (\Delta x_{\max})^2 + \left(1 - \frac{a}{2}\right)^2 (\Delta x_{\min})^2$$

explaining minimum in $\sigma(x)$

$\to$ INNs, diffusion just (non-parametric) fits



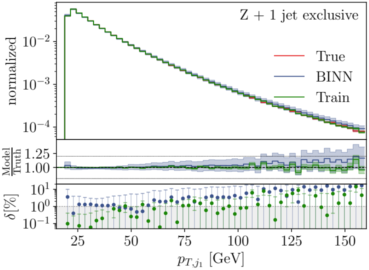


Events with uncertainties

Bayesian network generator

- network with weight distributions [Gal (2016)]
  sample weights [defining error bar]
  frequentist: efficient ensembling

$\Rightarrow$ Training-related error bars



Events with uncertainties

Bayesian network generator

- network with weight distributions [Gal (2016)]
  sample weights [defining error bar]
  frequentist: efficient ensembling

$\Rightarrow$ Training-related error bars

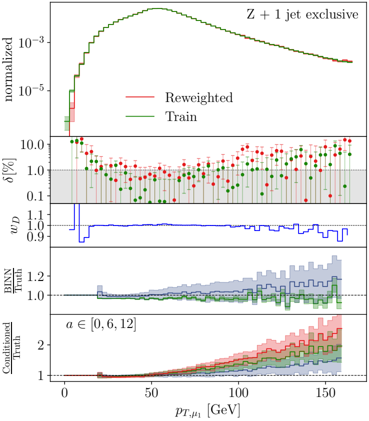
Theory uncertainties

- BNN regression/classification:
  systematics from data augmentation
- systematic uncertainties in tails

$$w = 1 + a \left(\frac{p_{T,j_1} - 15 \ \text{GeV}}{100 \ \text{GeV}} \right)^2$$

- augment training data [$a = 0 \dots 30$]
- train conditionally on a
  error bar from sampling a

$\Rightarrow$ Systematic/theory error bars



Controlling generative networks

Compare generated with training data [Das, Favaro, Heimel, Krause, TP, Shih]

- easy for regression $\Delta = (E_\text{data} - E_\theta)/E_\text{data}$
- harder for generation, unsupervised density
  classify training vs generated events learned density ratio

$$w(x_i) = \frac{D(x_i)}{1 - D(x_i)} = \frac{p_\text{data}(x_i)}{p_\text{model}(x_i)}$$

$\to$ Weight ratio over interpretable phase space

Controlling generative networks

Compare generated with training data [Das, Favaro, Heimel, Krause, TP, Shih]

- easy for regression $\Delta = (E_\text{data} - E_\theta)/E_\text{data}$
- harder for generation, unsupervised density
  classify training vs generated events learned density ratio

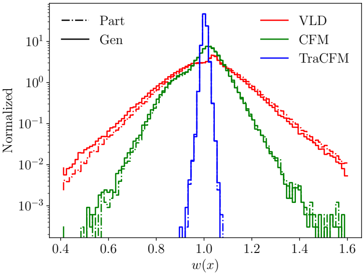
$$w(x_i) = \frac{D(x_i)}{1 - D(x_i)} = \frac{p_\text{data}(x_i)}{p_\text{model}(x_i)}$$

$\rightarrow$ Weight ratio over interpretable phase space

Progress in NN-generators

- train to generate events
- compare different architectures
- performance from width of distribution
- accuracy of learned density over phase spa

$\rightarrow$ Systematic performance test



Controlling generative networks

Compare generated with training data [Das, Favaro, Heimel, Krause, TP, Shih]

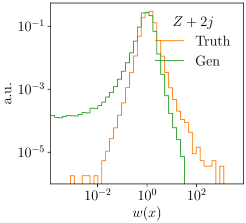
- easy for regression $\Delta = (E_\text{data} - E_\theta)/E_\text{data}$
- harder for generation, unsupervised density
  classify training vs generated events learned density ratio

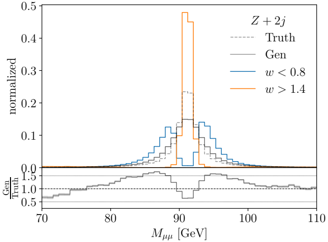
$$w(x_i) = \frac{D(x_i)}{1 - D(x_i)} = \frac{p_\text{data}(x_i)}{p_\text{model}(x_i)}$$

$\to$ Weight ratio over interpretable phase space

Weights over event phase space [same for jets, calorimeter showers]

- $w(x_i) \gg 1$ missing feature
  $w(x_i) \ll 1$ missing cut
- shifted weights indicating poor resolution





Controlling generative networks

Compare generated with training data [Das, Favaro, Heimel, Krause, TP, Shih]

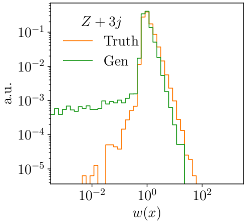
- easy for regression $\Delta = (E_\text{data} - E_\theta)/E_\text{data}$
- harder for generation, unsupervised density
  classify training vs generated events learned density ratio

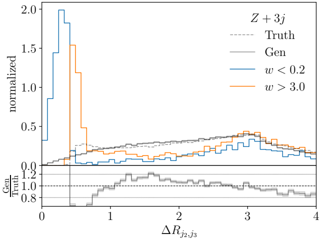
$$w(x_i) = \frac{D(x_i)}{1 - D(x_i)} = \frac{p_\text{data}(x_i)}{p_\text{model}(x_i)}$$

$\to$ Weight ratio over interpretable phase space

Weights over event phase space [same for jets, calorimeter showers]

- $w(x_i) \gg 1$ missing feature
  $w(x_i) \ll 1$ missing cut
- small weights indicating missing feature





Transforming LHC physics

Number of searches

- SBI: signal and background simulations
- CPU-limitation for many signals?

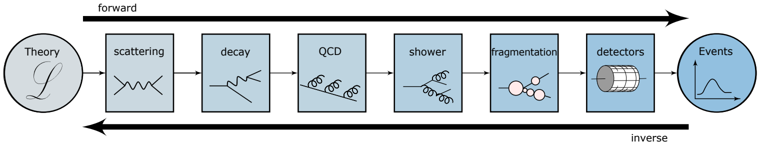
Optimal analyses

- theory limiting many analyses, but continuous progress
- allow for analyses to be updated?

Public LHC data

- common lore:
  LHC data too complicated for amateurs
- in truth:
  hard scattering and decay simulations public
  BSM physics not in hadronization and detector

$\rightarrow$ Unfold to suitable level [EFT?]



ML for LHC Theory

Developing ML for the best science

- just another numerical tool for a numerical field
- transformative new common language
- driven by money from data science and medical research
- be 10000 Einsteins,

  ...improving established tools
  ...developing new tools for established tasks
  ...transforming through new ideas

→ Nobel prize given — mature field!

:2211.01421v2 [hep-ph] 17 Mar 2024

Modern Machine Learning for LHC Physicists

Tilman Plehn[a*], Anja Butter[a,b], Barry Dillon[a],
Theo Heimel[a], Claudius Krause[c], and Ramon Winterhalder[d]

[a] Institut für Theoretische Physik, Universität Heidelberg, Germany
[b] LPNHE, Sorbonne Université, Université Paris Cité, CNRS/IN2P3, Paris, France
[c] HEPHY, Austrian Academy of Sciences. Vienna, Austria
[d] CP3, Université catholique de Louvain, Louvain-la-Neuve, Belgium

March 19, 2024

Abstract

Modern machine learning is transforming particle physics fast, bullying its way into our numerical tool box. For young researchers it is crucial to stay on top of this development, which means applying cutting-edge methods and tools to the full range of LHC physics problems. These lecture notes lead students with basic knowledge of particle physics and significant enthusiasm for machine learning to relevant applications. They start with an LHC-specific motivation and a non-standard introduction to neural networks and then cover classification, unsupervised classification, generative networks, and inverse problems. Two themes defining much of the discussion are well-defined loss functions and uncertainty-aware networks. As part of the applications, the notes include some aspects of theoretical LHC physics. All examples are chosen from particle physics publications of the last few years.[1]