Introduction to ML for particle physics in the precision era

Henning Bahl

UNIVERSITÄT HEIDELBERG
ZUKUNFT
SEIT 1386

CRC Young Scientists Meeting, Heidelberg, 23.7.2025

The goal of particle physics

→ Answer the big fundamental questions!

Nature of EWSB

Neutrino masses

.....

Dark matter

Baryon asymmetry

Naturalness

The goal of particle physics

→ Answer the big fundamental questions!

Nature of EWSB

Neutrino masses

.....

Dark matter

Baryon asymmetry

Naturalness

Can ML find answer these questions for us? **No!**

The goal of particle physics

→ Answer the big fundamental questions!

Nature of EWSB

Dark matter

Baryon asymmetry

Neutrino masses

Naturalness

.....

Can ML find answer these questions for us? **No!**

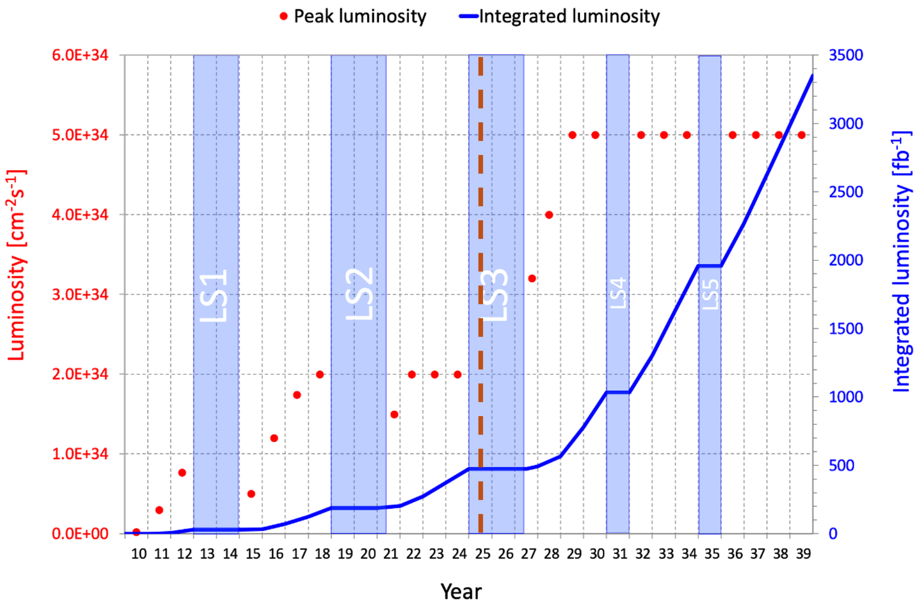
Can it help us with it? **Yes!**

The challenge ahead

- general trend: larger-and-larger experiments collecting more-and-more data

- e.g. LHC: already enormous dataset will be further enlarged by a factor $\sim$ 10

- costs for future experiments increasing

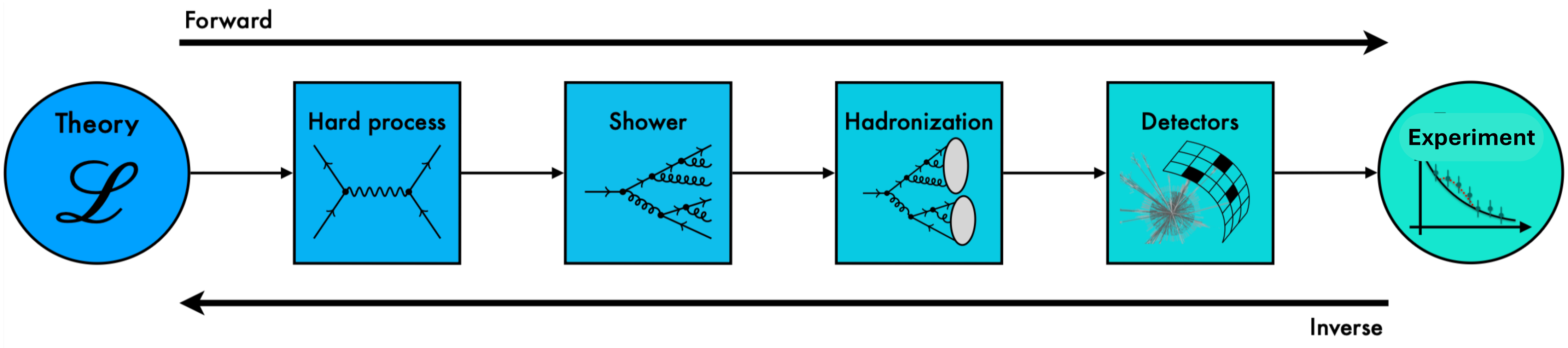


Fully exploit the available data!

- new analysis methods

- theory precision $\simeq$ experimental precision

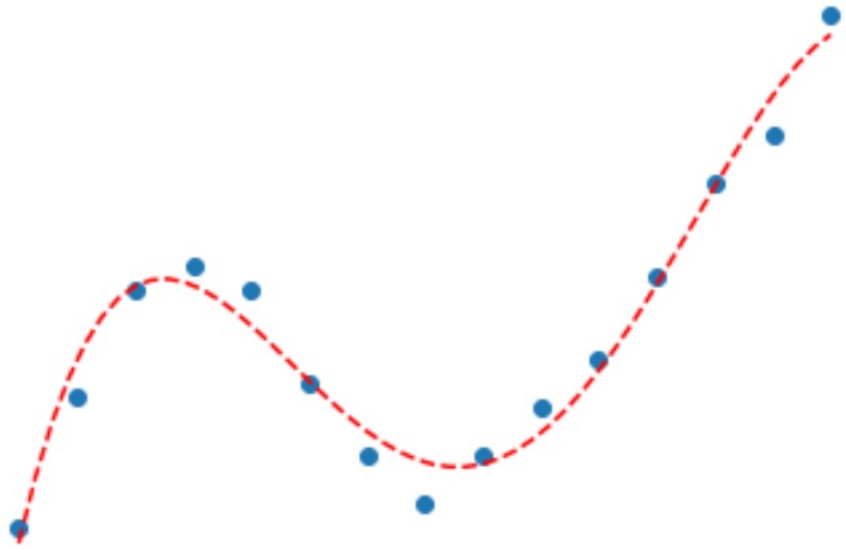
The particle physics workflow



ML can help with each of these steps by increasing

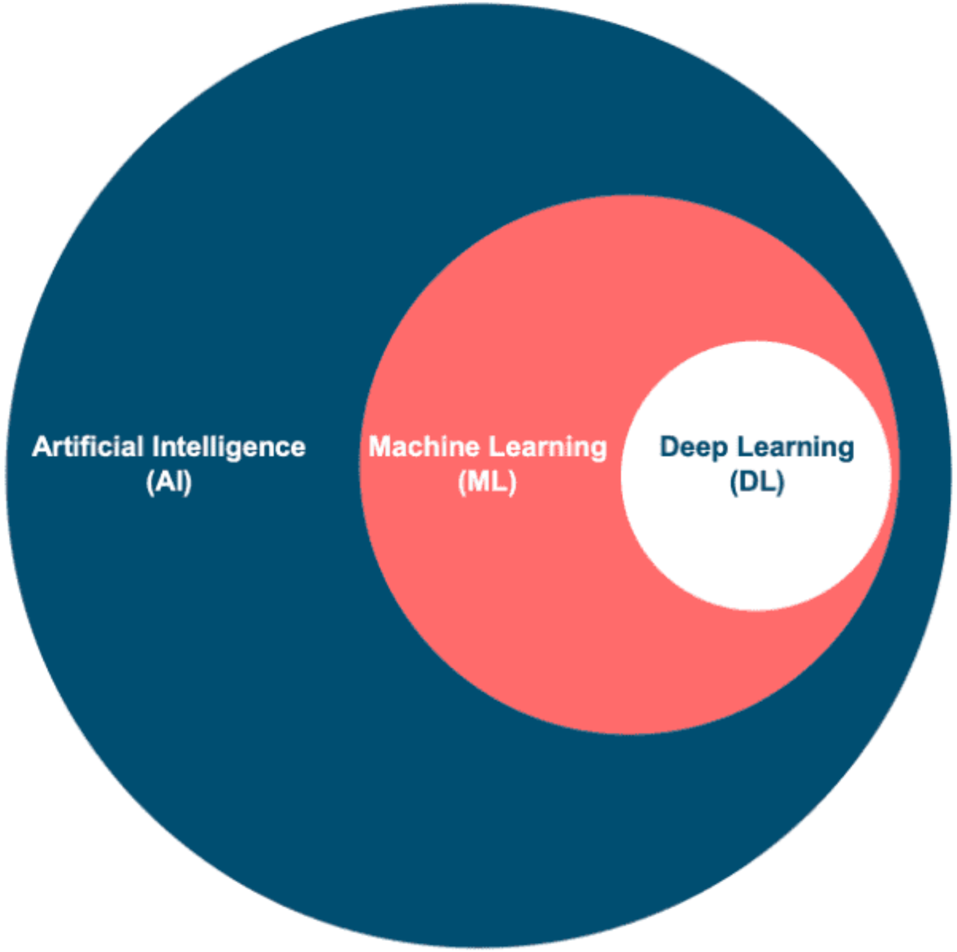
- accuracy/performance and/or
- increase speed

ML in a nutshell



Terminology

- **Artificial Intelligence (AI)**
  - machines performing complex tasks
  - e.g. Feynman diagram generators, …
- **Machine Learning (ML)**
  - subfield of AI where machines learn from data
  - e.g. linear regression, BDTs, …
- **Deep Learning (DL)**
  - subfield of ML using deep neural networks



Types of ML (selection)

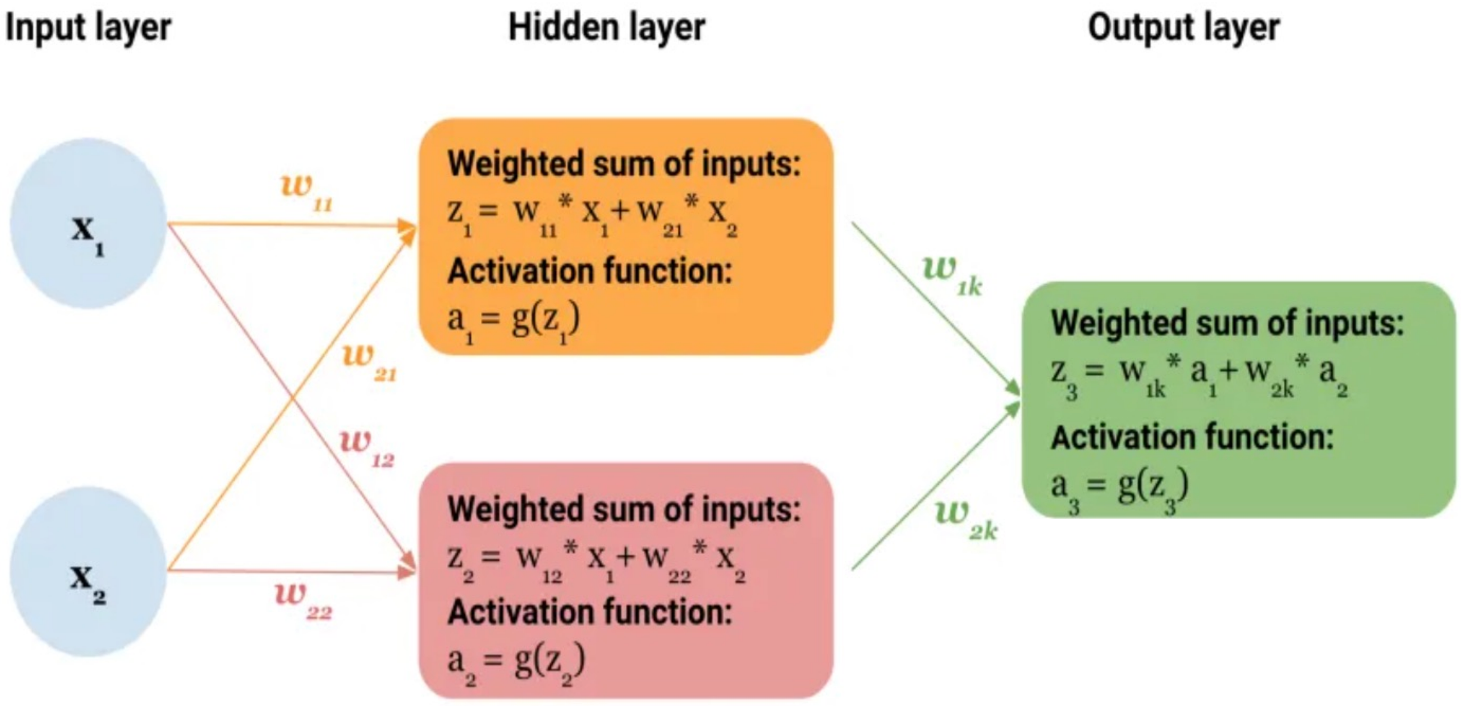
Tasks

- regression (e.g. calorimeter calibration)
- classification (e.g. jet tagging)
- generation (e.g. event generation)

Learning types

- supervised (e.g. amplitude regression)
- unsupervised (e.g. data clustering)
- semi-supervised (e.g. anomaly detection)

Neural networks



- activation introduces non-linearity (e.g. $g(x) = \max(0, x)$)
- adjust weights by minimizing loss
- large enough network can in principle approximate any function

ML workflow

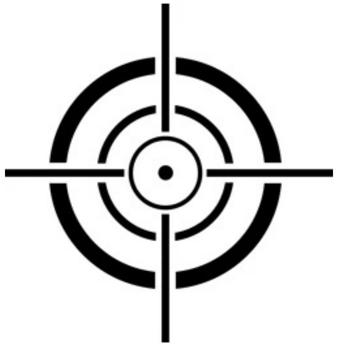
1. define the problem
2. collect and preprocess the dataset
3. define your ML model
4. training
5. evaluation

ML workflow

1. define the problem
2. collect and preprocess the dataset
3. define your ML model
4. training
5. evaluation

- ML strategy — *multiple ways to approach problem*
- loss — *what objective do I want to optimize?*
- architecture — *what is the best structure for my NN?*
- encode physics knowledge — *symmetries, …*

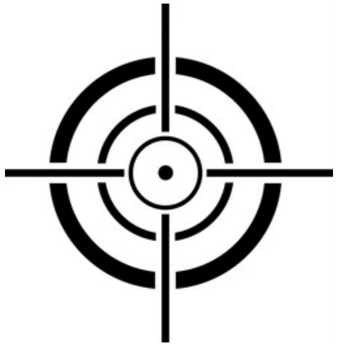
ML for particle physics



precision

precision

ML for particle physics

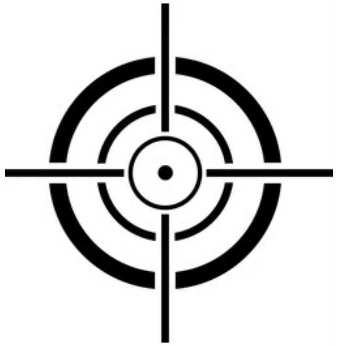


precision

precision

control

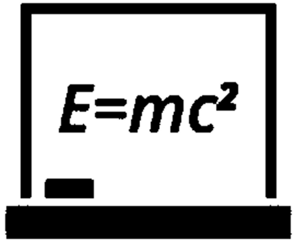
ML for particle physics



precision

precision

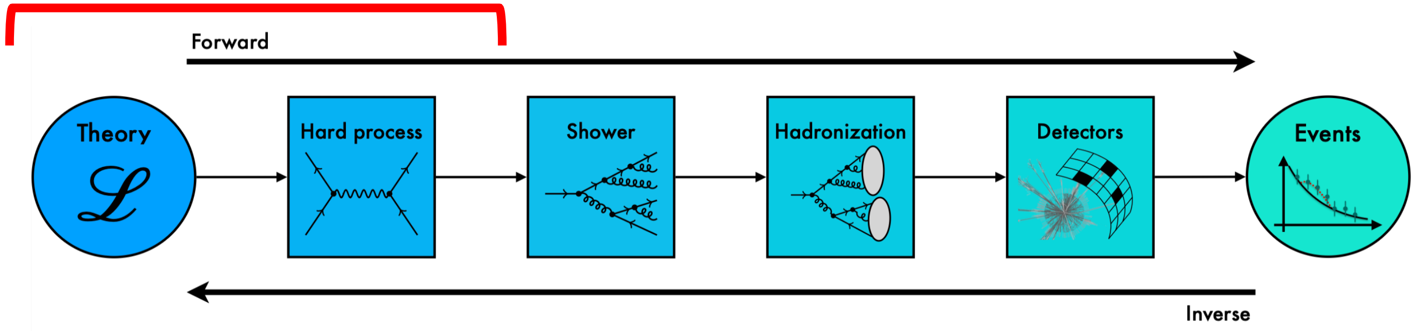
control



physics

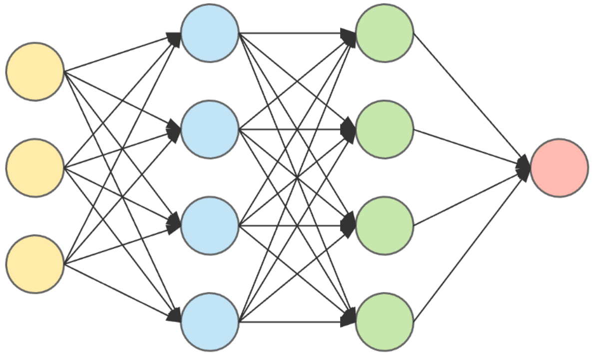
Amplitude surrogates

fast high-precision event generation



Amplitude surrogates

- evaluating analytic expressions for amplitudes $|\mathcal{M}|^2$ can be very expensive due to
  - higher-order corrections
  - large final-state multiplicities

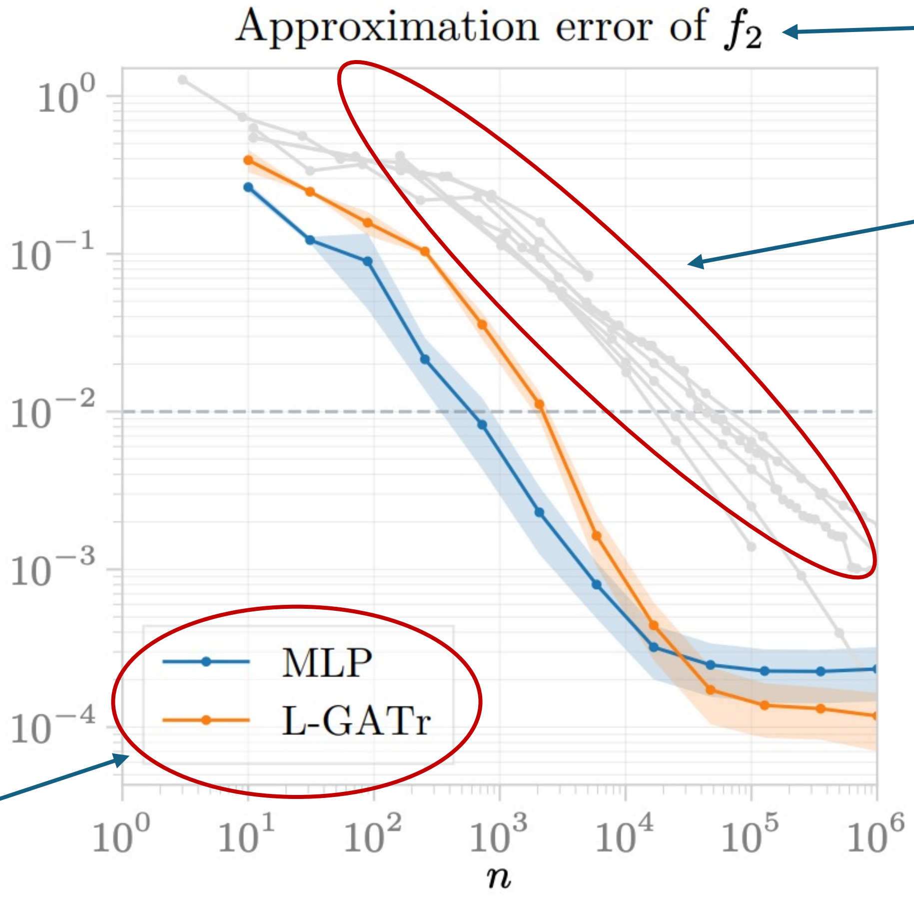
$|\mathcal{M}|^2 \approx$ 

- idea:
  - generate small training sample using full analytic expression
  - train a NN to approximate $|\mathcal{M}|^2$
  - generate events using NN surrogate, which is much faster to evaluate

→ fast high-precision event generation

Comparison to classical interpolation

[Bresó, Heinrich, et al., 2412.09534]



NLO $q\bar{q} \to t\bar{t}H$ amplitude

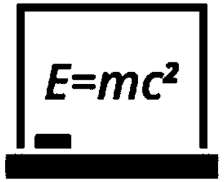
classical interpolation techniques

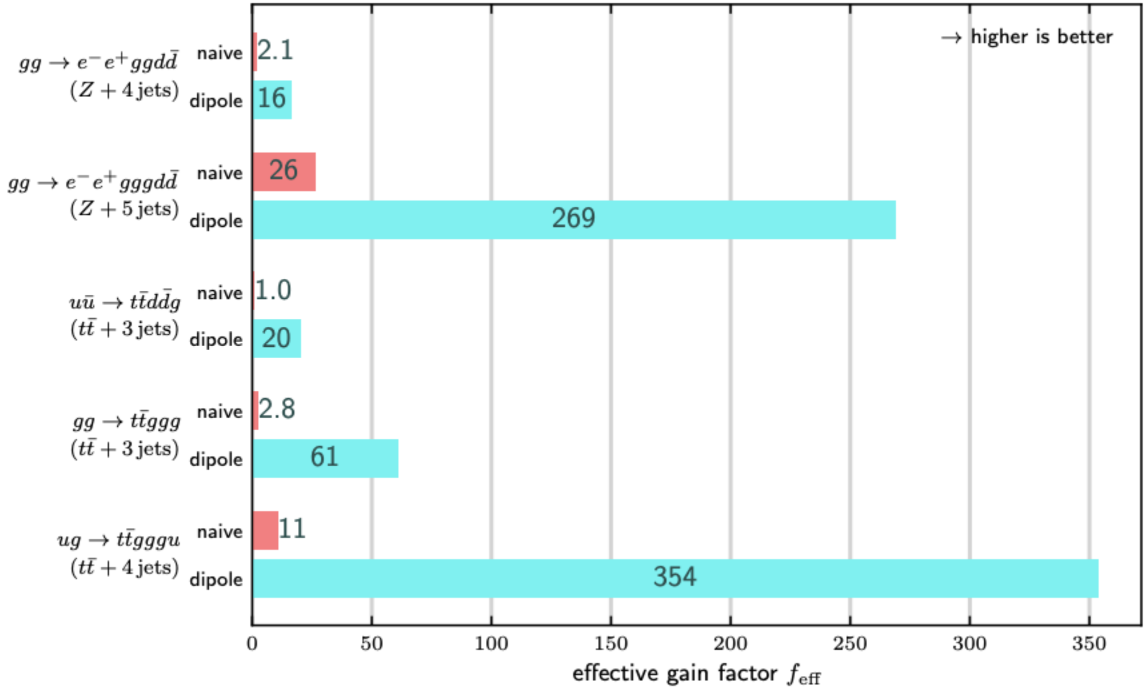
ML surrogates

ML surrogates outperform classical interpolation techniques

Speed comparison

[Janßen et al.,2301.13562]





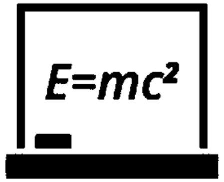
	Sherpa default			with dipole-model surrogate				
Process	t_{ME}[ms]	t_{PS}[ms]	ϵ_{full}	t_{surr}[ms]	x_{max}	$\epsilon_{\text{1st,surr}}$	$\epsilon_{\text{2nd,surr}}$	f_{eff}
$gg \to e^- e^+ ggd\bar{d}$	54	0.40	1.411 %	0.14	2.6	1.418 %	39 %	16
$gg \to e^- e^+ gggd\bar{d}$	16 216	5.70	0.076 %	0.20	3.6	0.085 %	29 %	269

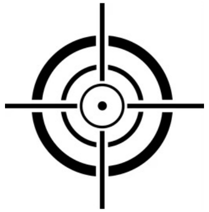
$$f_{\text{eff}} = \frac{T_{\text{standard}}}{T_{\text{surrogate}}}$$

dipole vs naïve:
encode singularity structure of amplitudes

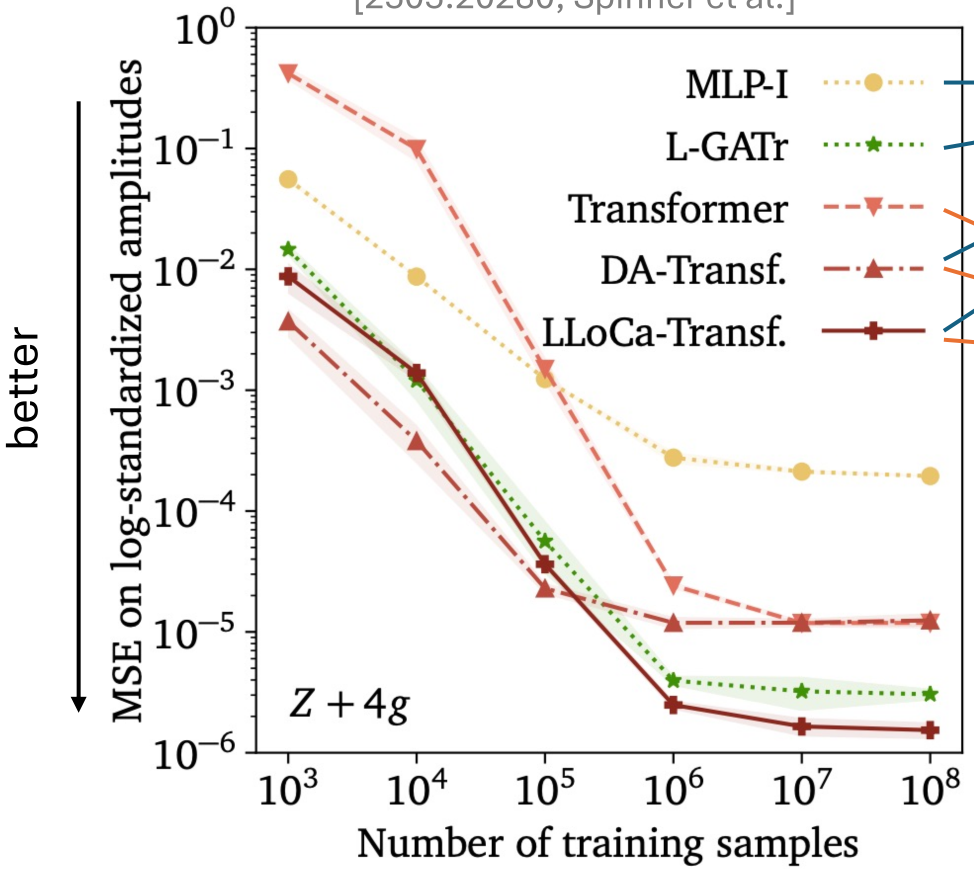
Large speed-ups possible!

More physics knowledge





[2505.20280, Spinner et al.]



Lorentz invariance

permutation invariance ($g_i \leftrightarrow g_j$)

Exploiting symmetries drastically improves performance!

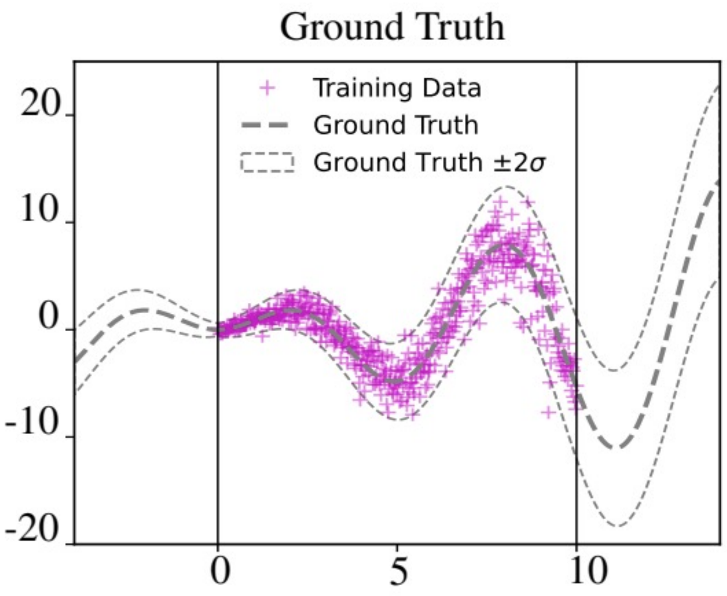
Monitoring uncertainties

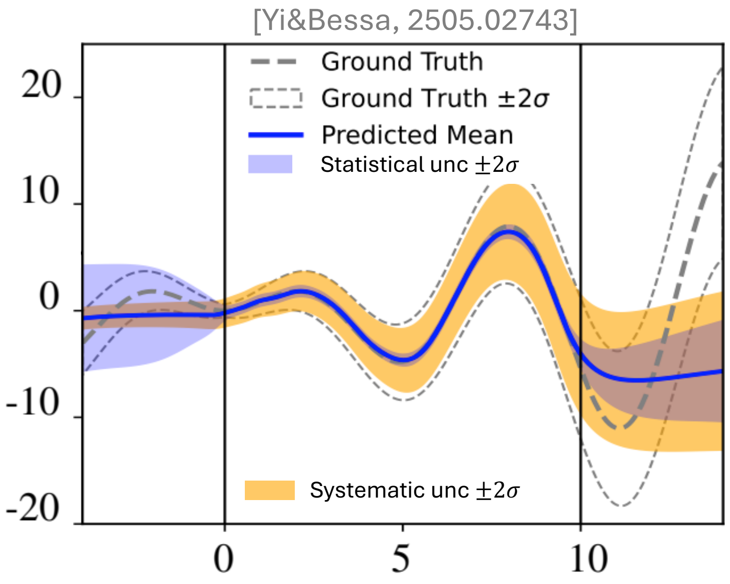
"All models are wrong, but some — those that know when they can be trusted — are useful!"

George Box (adapted)

→ our NN should not only give as a prediction but also tell us how certain it is

Regression with uncertainties





- statistical uncertainty ≙ lack of training data
- systematic uncertainty ≙ noise in the data, lack in model expressivity

Modelling the systematic uncertainty

- log-likelihood loss:

sum over training dataset

true amplitudes

$$\mathcal{L} = -\sum_{x_i, A_i \in D_\text{train}} \log p(A_\text{true}(x_i) | x_i, \theta)$$

NN parameters

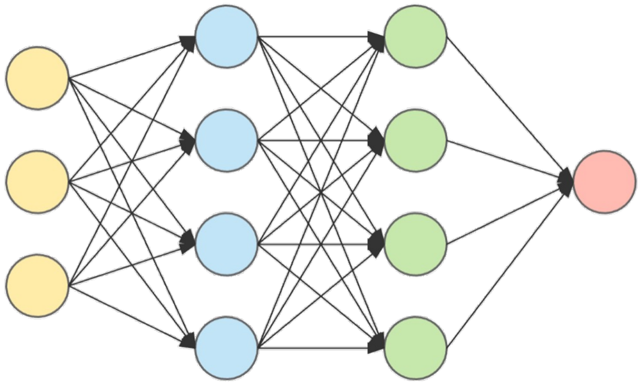
phase-space point

- assume Gaussian likelihood: $p(A|x,\theta) = \mathcal{N}(\overline{A}(x), \sigma_\text{syst}^2(x))$
- NN learns both: $\overline{A}(x)$ and $\sigma_\text{syst}(x)$

$$\Rightarrow \text{heteroskedastic loss: } \mathcal{L} = \sum_i \left[\frac{\left(\overline{A}(x_i) - A_\text{true}(x_i)\right)^2}{2\sigma_\text{syst}^2(x_i)} + \log\left(\sigma_\text{syst}(x_i)\right) \right]$$

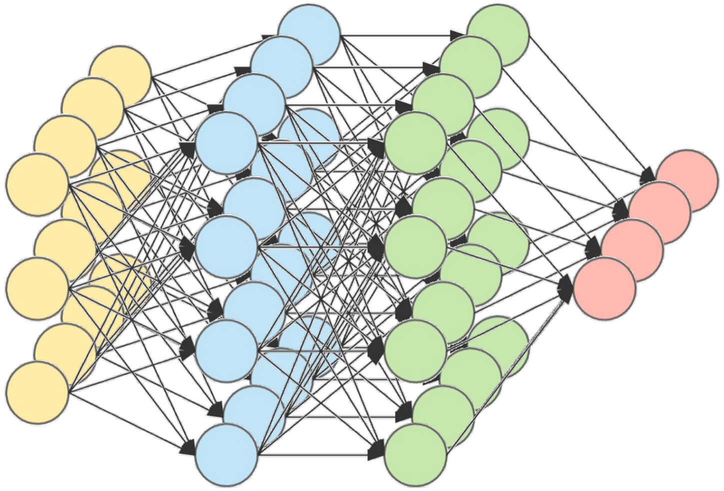
- constant σ_syst → MSE loss

Modelling the statistical uncertainty



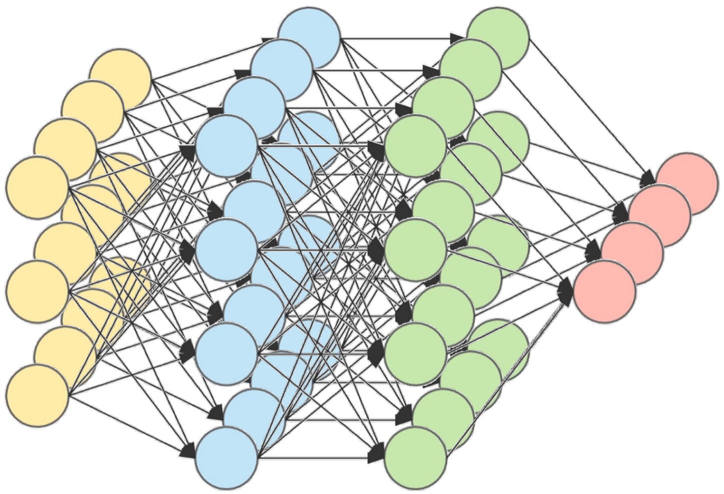
- train ensemble of networks
- each networks leads to slightly different result
- spread of network predictions ~ statistical uncertainty
- less data → higher spread

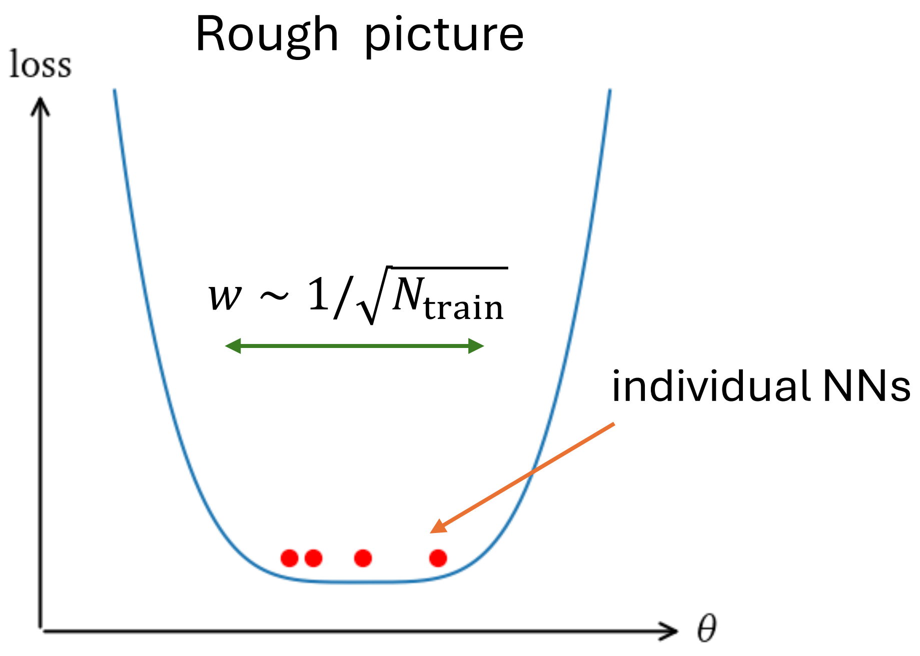
Modelling the statistical uncertainty



- train ensemble of networks
- each networks leads to slightly different result
- spread of network predictions $\sim$ statistical uncertainty
- less data $\rightarrow$ higher spread

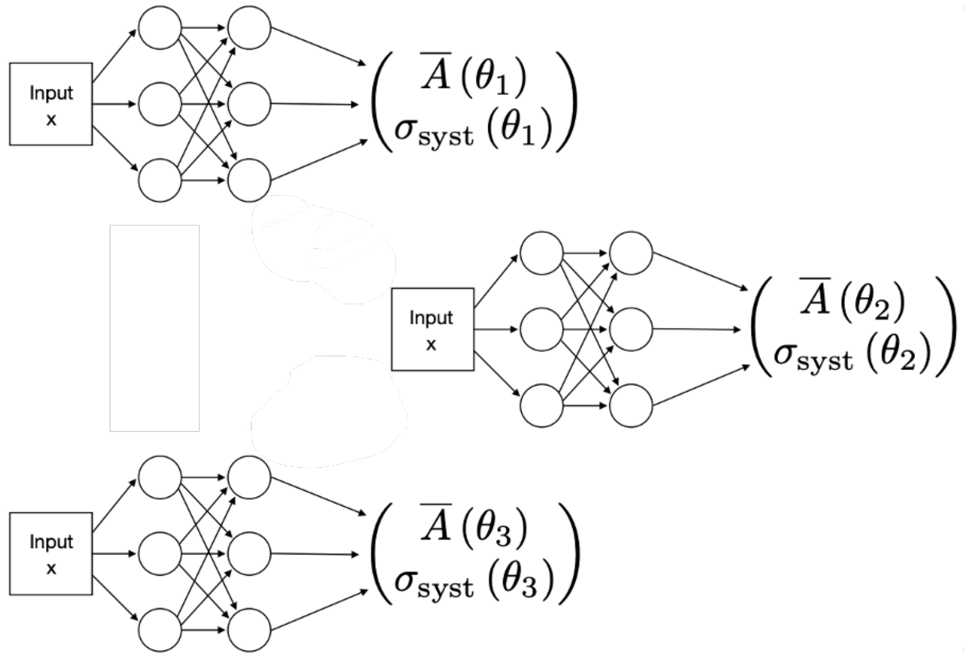
Modelling the statistical uncertainty





- train ensemble of networks
- each networks leads to slightly different result
- spread of network predictions ~ statistical uncertainty
- less data → higher spread

Bringing it all together

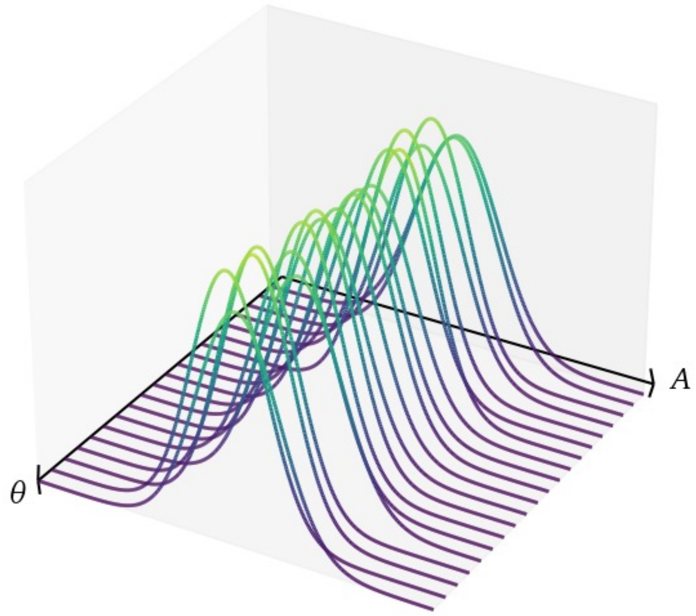


$$\langle A \rangle = \frac{1}{N} \sum_i^N \overline{A}(\theta_i)$$

$$\sigma_{\text{syst}}^2 = \frac{1}{N} \sum_i^N \sigma_{\text{syst}}^2(\theta_i)$$

$$\sigma_{\text{stat}}^2 = \frac{1}{N} \sum_i^N \left(\langle A \rangle - \overline{A}(\theta_i) \right)^2$$

$$\sigma_{\text{tot}}^2 = \sigma_{\text{syst}}^2 + \sigma_{\text{stat}}^2$$

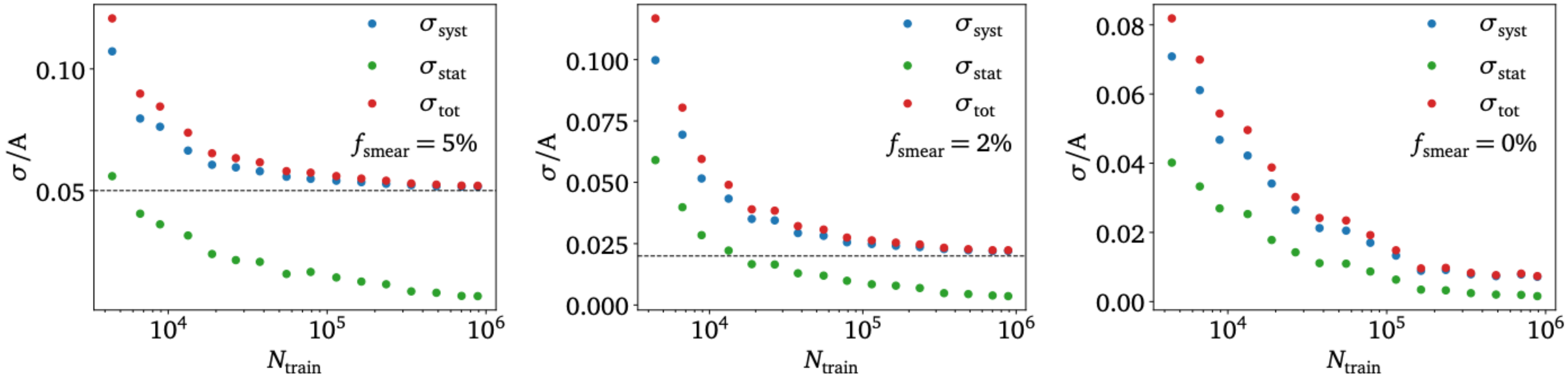


Combined learnable modelling of systematic and statistical uncertainties!

Behavior of uncertainties

[HB et al.,2412.12069]

$$A_{\text{train}} \sim \mathcal{N}(A_{\text{true}}, \sigma_{\text{train}}^2)$$
$$\sigma_{\text{train}} = f_{\text{smear}} A_{\text{true}}$$



Test: apply different levels of Gaussian noise to amplitudes

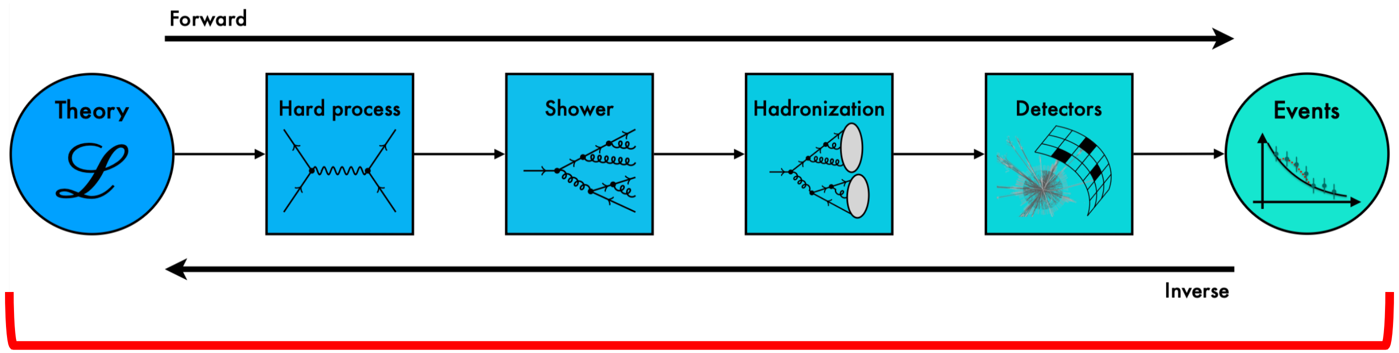
- statistical unc. decreases with more training data
- systematic unc. converges to level of applied noise

→ reliable uncertainty estimate

Same techniques also applicable to all kind of other problems!

Simulation-Based Inference

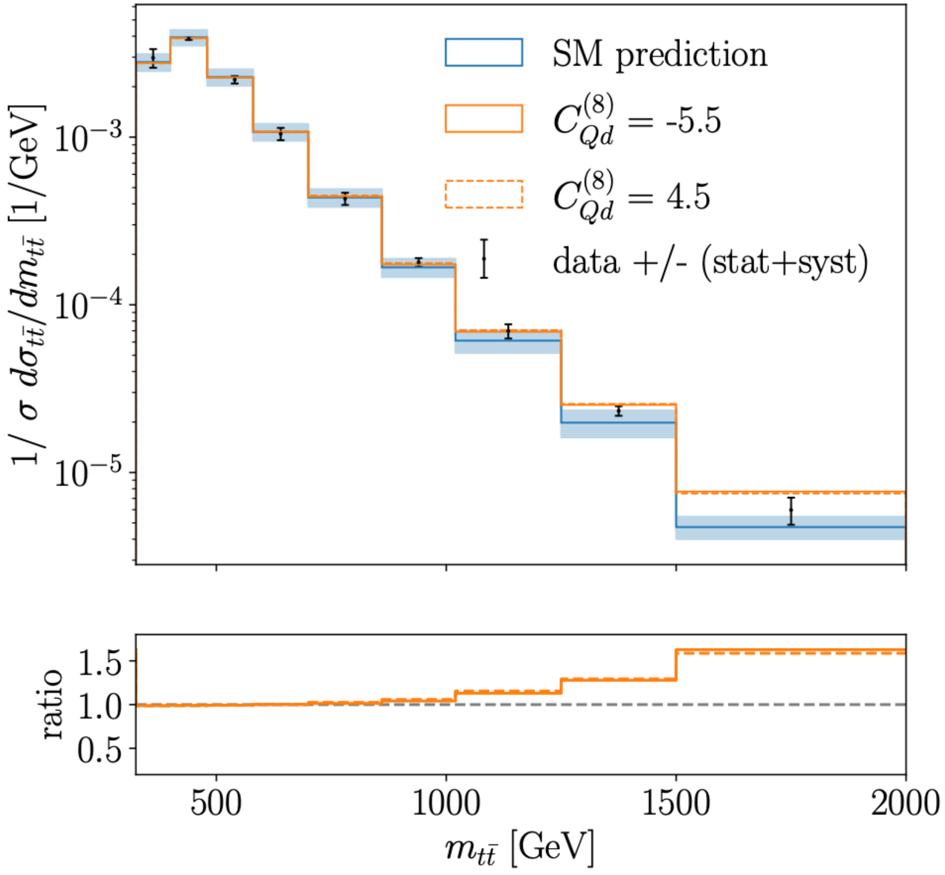
fully exploiting high-dimensional data



Classical parameter inference

- reduce dimension of phase space summary statistics →
- bin summary statistics
- compare resulting histogram to SM/BSM predictions

[Elmer et al.,2312.12502]



Advantage: humanly digestible plots

Disadvantage: loss of information

Full likelihood

phase space point

theory parameters

- Monte-Carlo simulation chain allows us to sample full likelihood $p(x|\theta)$. But cannot directly compute it.

- Neyman-Pearson lemma: likelihood ratio $r(x|\theta,\theta_0) \equiv \frac{p(x|\theta)}{p(x|\theta_0)}$ is most powerful statistical test

- but we can regress to reco-level $r(x|\theta,\theta_0)$ using known parton-level $r(z_p|\theta,\theta_0)$:

$$\mathcal{L} = \left\langle \left[r(z_p|\theta,\theta_0) - \underbrace{r_\varphi(x|\theta,\theta_0)}_{\text{NN}} \right]^2 \right\rangle_{\underbrace{x,z_p \sim p(x|z_p)p(z_p|\theta);\theta\sim q(\theta)}_{\text{average over event sample}}}$$

→ unbinned multi-dimensional inference without information loss

Encoding amplitude structure

[Schöfbeck et al., 2107.10859, 2205.12976]

Theory structure for e.g. SMEFT:

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_i}{\Lambda^2} O_i \equiv \mathcal{L}_{\text{SM}} + \sum_i \theta_i O_i$$

$$|\mathcal{M}(z_p|\theta)|^2 = |\mathcal{M}_{\text{SM}}(z_p)|^2 + \theta_i |\mathcal{M}_i(z_p)|^2 + \theta_i \theta_j |\mathcal{M}_{ij}(z_p)|^2$$

encode into likelihood

$$R(x|\theta,\theta_0) \equiv \frac{d\sigma(x|\theta)/dx}{d\sigma(x|\theta_0)/dx} = \frac{\sigma(\theta)p(x|\theta)}{\sigma(\theta_0)p(x|\theta_0)}$$

$$R(x|\theta,\theta_0) = 1 + (\theta-\theta_0)_i R_i(x) + (\theta-\theta_0)_i(\theta-\theta_0)_j R_{ij}(x)$$

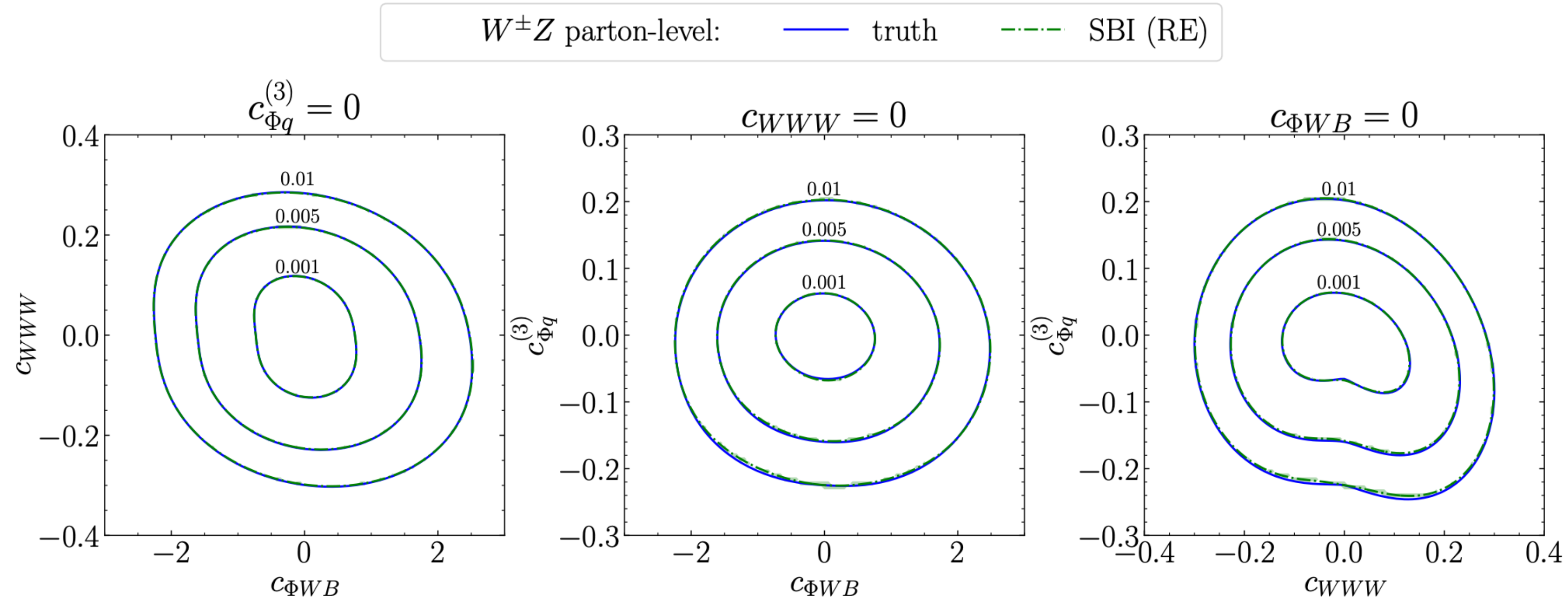
$$R_i(z_p) \equiv \frac{\partial}{\partial\theta_i} \frac{d\sigma(z_p|\theta)/dz_p}{d\sigma(z_p|\theta_0)/dz_p}\bigg|_{\theta=\theta_0} = \frac{\partial_{\theta_i}|\mathcal{M}(z_p|\theta)|^2}{|\mathcal{M}(z_p|\theta_0)|^2}\bigg|_{\theta_0}$$

$$R_{ij}(z_p) \equiv \frac{\partial^2}{\partial\theta_i\partial\theta_j} \frac{d\sigma(z_p|\theta)/dz_p}{d\sigma(z_p|\theta_0)/dz_p}\bigg|_{\theta=\theta_0} = \frac{\partial_{\theta_i}\partial_{\theta_j}|\mathcal{M}(z_p|\theta)|^2}{|\mathcal{M}(z_p|\theta)|^2}\bigg|_{\theta_0}$$

learn coefficients $R_{i,ij}$ separately → theory parameter dependence fully factored out

Parton-level cross-check: $W^{\pm}Z$ production

- consider effects of three SMEFT operators

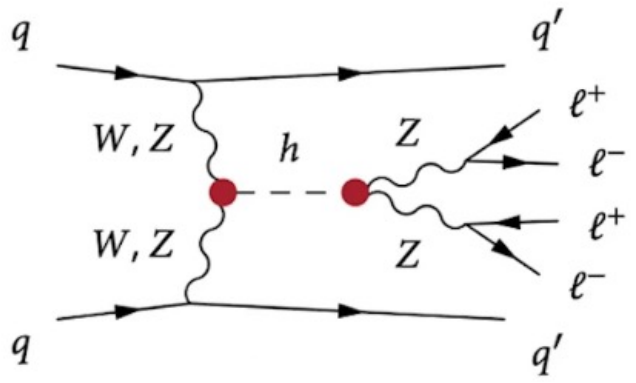


[HB et al.,2410.07315]

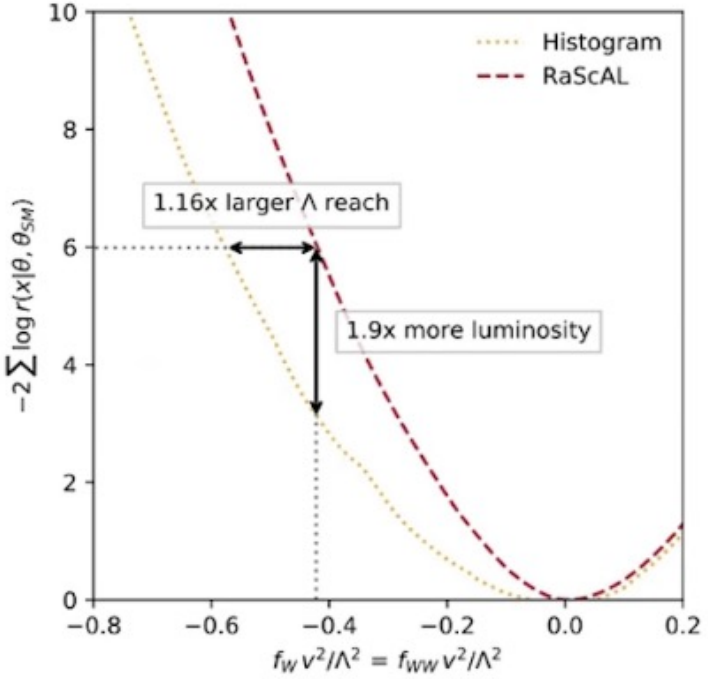
almost perfectly learns high-dimensional likelihood

Reco-level: VBF with $H \to 4\ell$

[Brehmer et al., 1805.00013]



$$\mathcal{L} = \mathcal{L}_{\rm SM} + \boxed{\frac{f_W}{\Lambda^2}} \underbrace{\frac{\mathrm{i}g}{2} (D^\mu \phi)^\dagger \sigma^a D^\nu \phi \, W^a_{\mu\nu}}_{\mathcal{O}_W} - \boxed{\frac{f_{WW}}{\Lambda^2}} \underbrace{\frac{g^2}{4} (\phi^\dagger \phi) \, W^a_{\mu\nu} W^{\mu\nu\, a}}_{\mathcal{O}_{WW}}$$



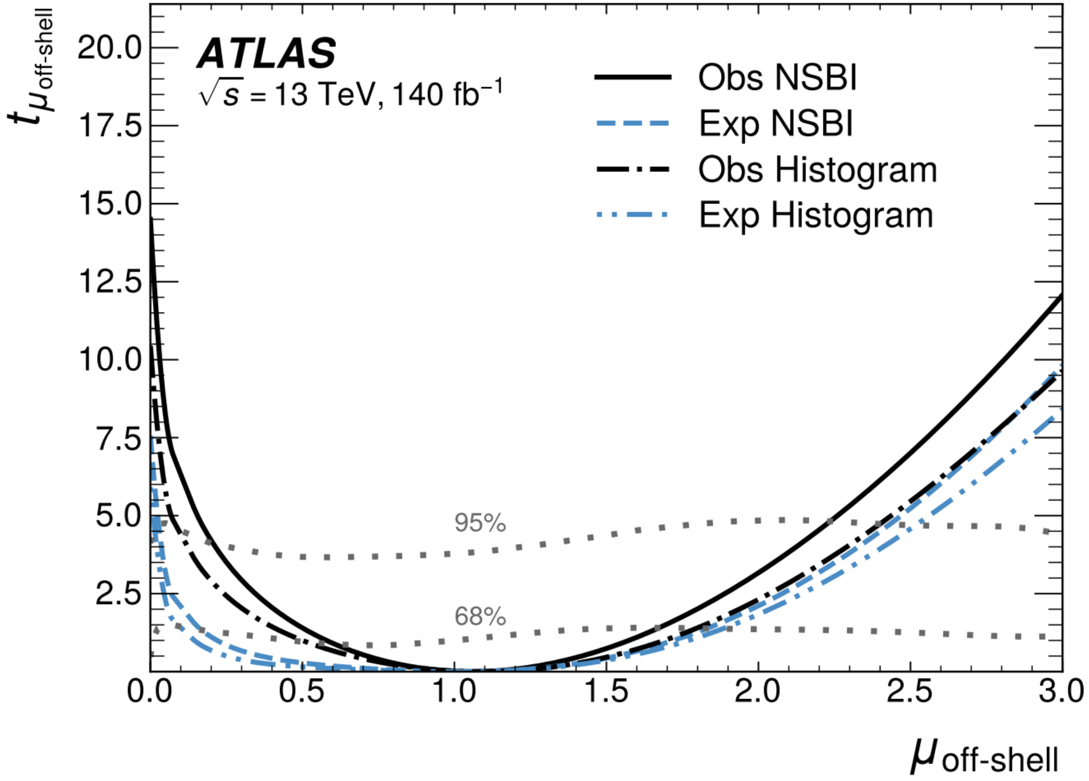
➡ Huge potential to improve sensitivity of a wide variety of measurements/searches

But is SBI also viable in a realistic analysis including uncertainties etc.?

1st experimental SBI analysis

[ATLAS-CONF-2024-016]

- goal: measure off-shell signal strength in $H \to ZZ$ channel
- full treatment of statistical and systematic uncertainties
- large sensitivity improvement for low $\mu_{\text{off-shell}}$



proves potential of SBI for full experimental analysis

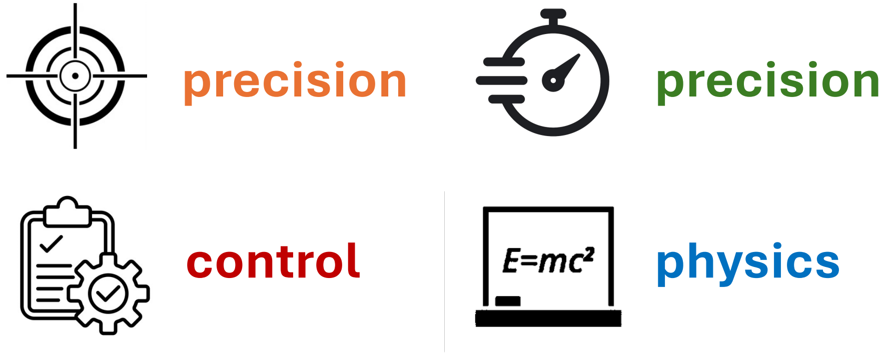
Conclusions

Conclusions

- particle physics is in the precision era
  → large amounts of multidimensional data

- ML methods excel in such an environment

- huge potential for increasing



  - speed → e.g., amplitude surrogates
  - performance → e.g., simulation-based inference

- uncertainty-aware NNs allow for controlled modelling

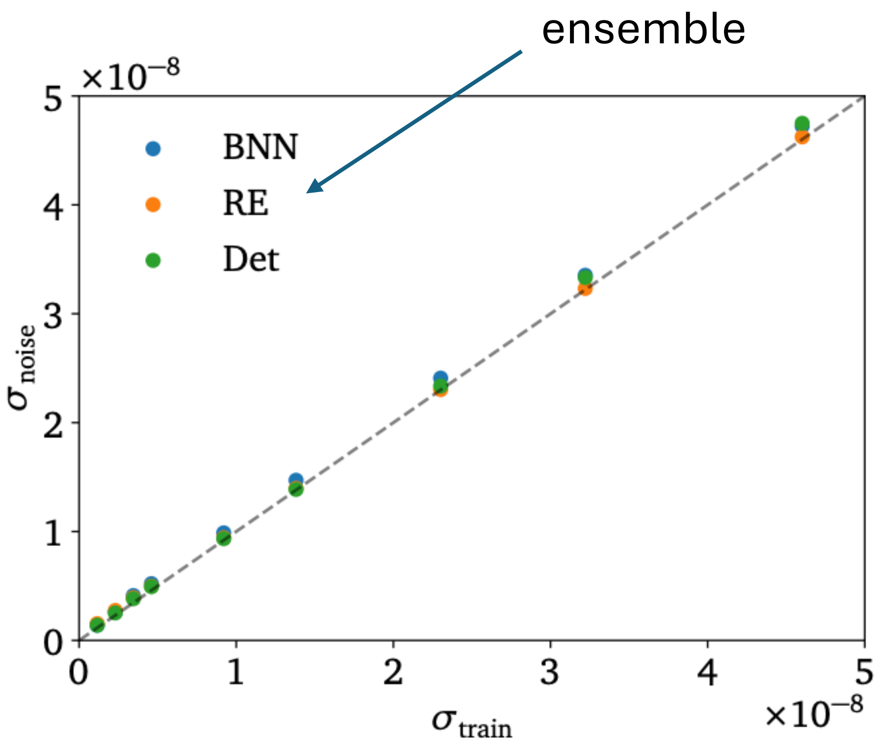
- encoding physics knowledge → large performance boosts

→ ML methods will be indispensable for the future of particle physics

Appendix

Behavior of uncertainties

[HB et al.,2412.12069]



NNs can reliably extract noise level!

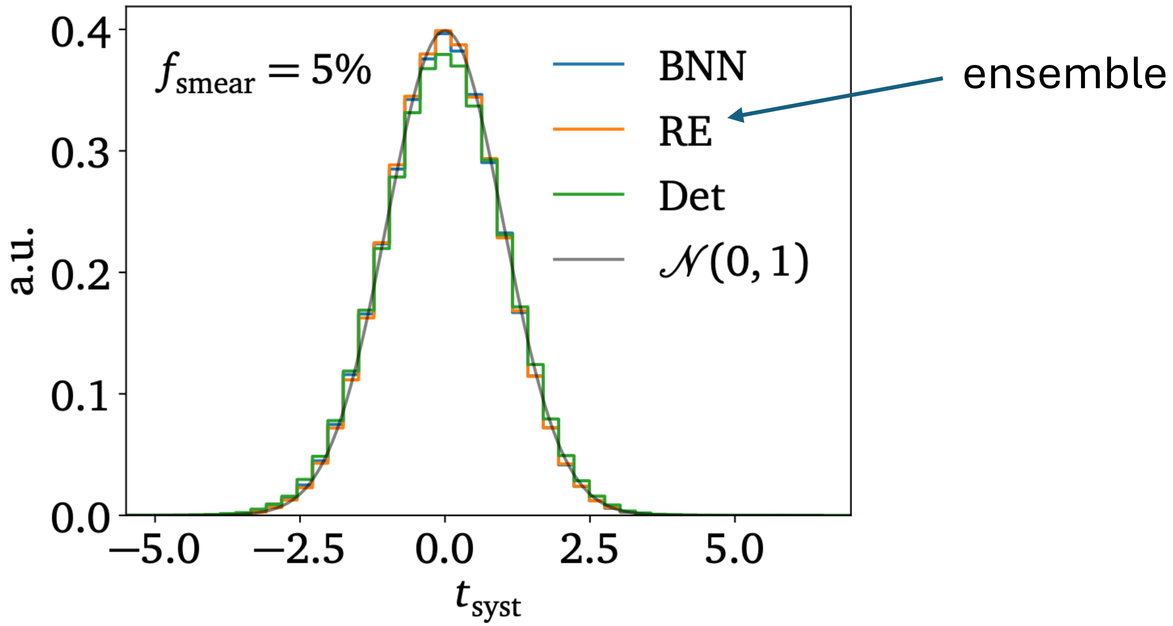
→ Are these uncertainties calibrated?

Calibration of uncertainties

- statistical uncertainties negligible for our application
- define systematic pull:

$$t_{\text{syst}} = \frac{\langle A\rangle(x) - A_{\text{train}}(x)}{\sigma_{\text{syst}}(x)}$$

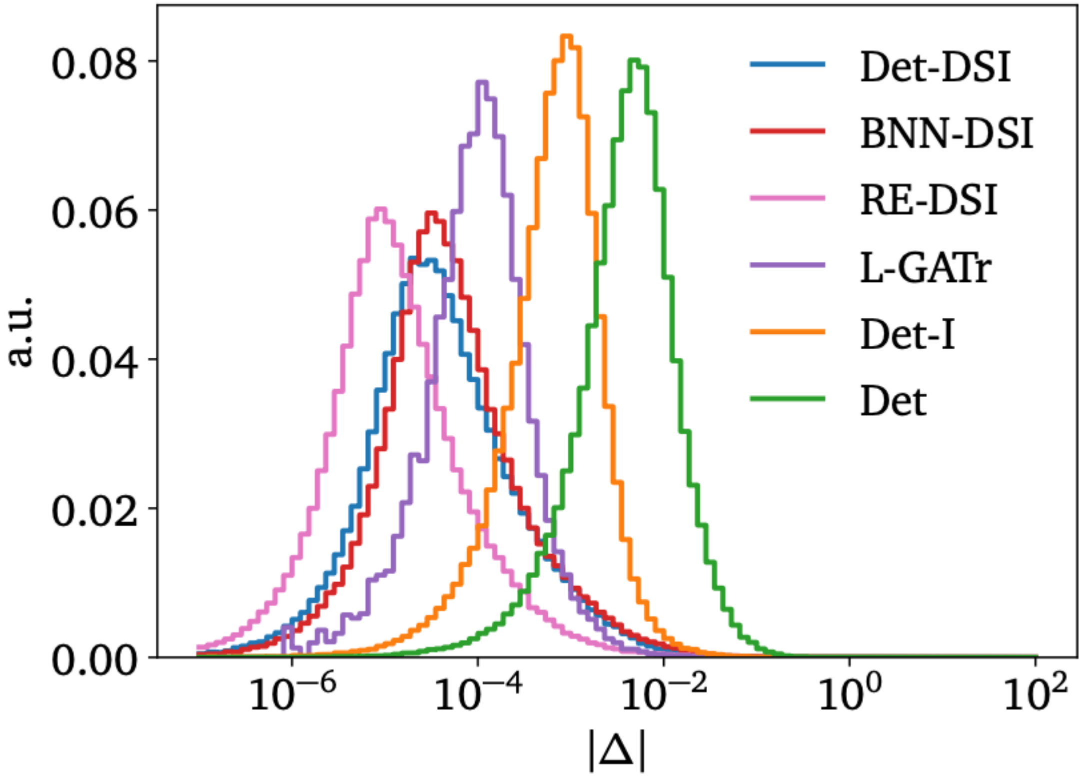
- if calibrated, t_{syst} distribution should follow $\mathcal{N}(0,1)$



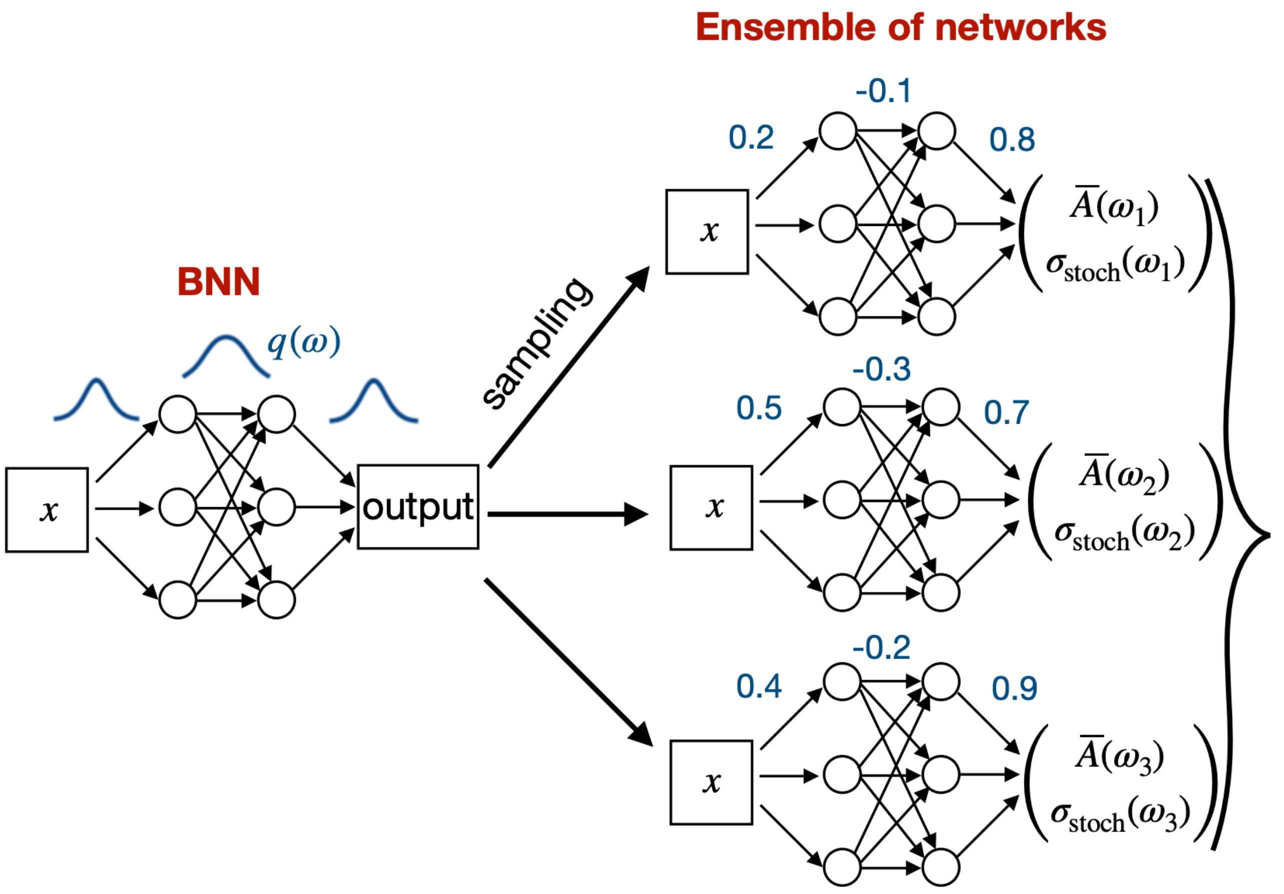
Almost perfectly calibration → reliable uncertainty estimate

Same techniques also applicable to all kind of other problems!

Encoding our physics knowledge



Bayesian neural networks



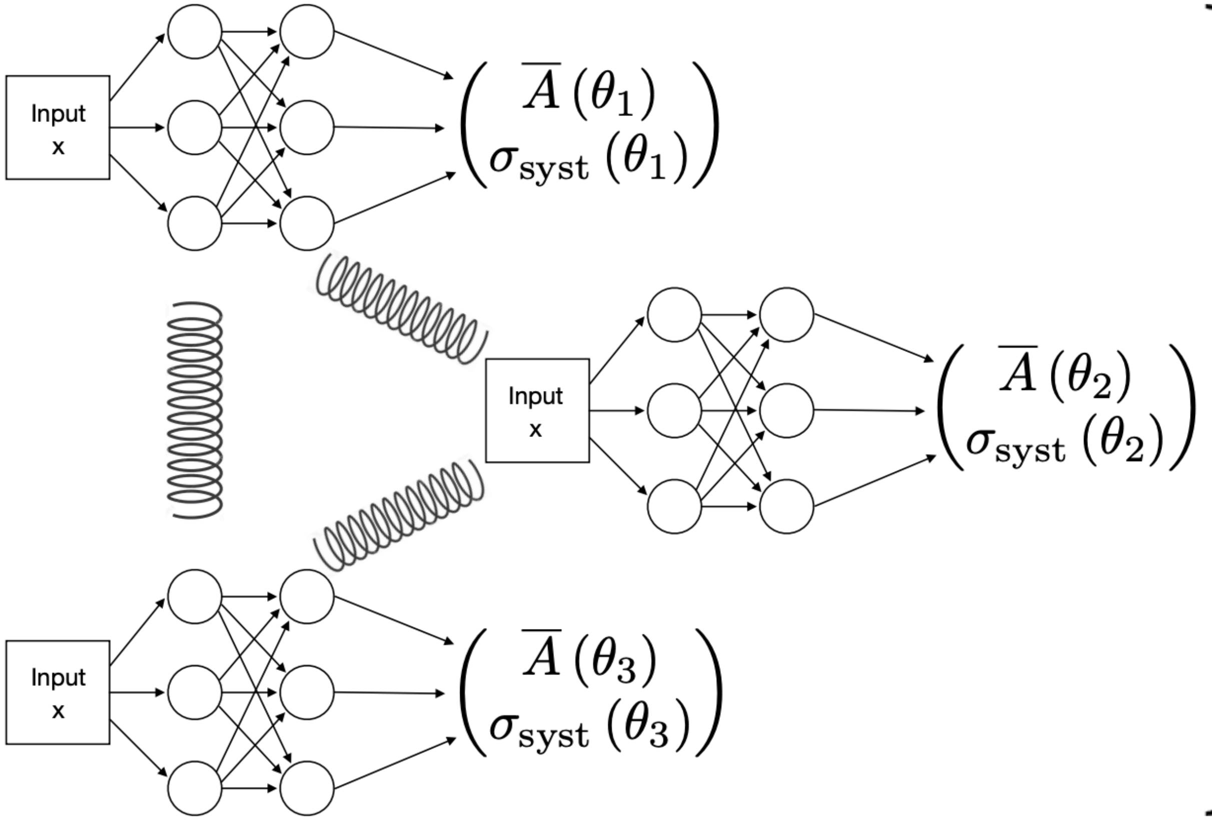
Output

$$\langle A \rangle = \frac{1}{N} \sum_i^N \bar{A}(\omega_i)$$

$$\sigma_{\text{stoch}}^2 = \frac{1}{N} \sum_i^N \sigma_{\text{stoch}}^2(\omega_i)$$

$$\sigma_{\text{pred}}^2 = \frac{1}{N} \sum_i^N (\langle A \rangle - \bar{A}(\omega_i))^2$$

Repulsive ensembles



$$\langle A \rangle = \frac{1}{N} \sum_i^N \overline{A}(\theta_i)$$

$$\sigma_{\text{syst}}^2 = \frac{1}{N} \sum_i^N \sigma_{\text{syst}}^2(\theta_i)$$

$$\sigma_{\text{stat}}^2 = \frac{1}{N} \sum_i^N \left(\langle A \rangle - \overline{A}(\theta_i) \right)^2$$

Full likelihood

phase space point

theory parameters

- Monte-Carlo simulation chain allows us to sample full likelihood $p(x|\theta)$. But cannot directly compute it.

- train classifier D to distinguish BSM sample ($\sim p(x|\theta)$) and SM sample ($\sim p(x|\theta_0)$) :

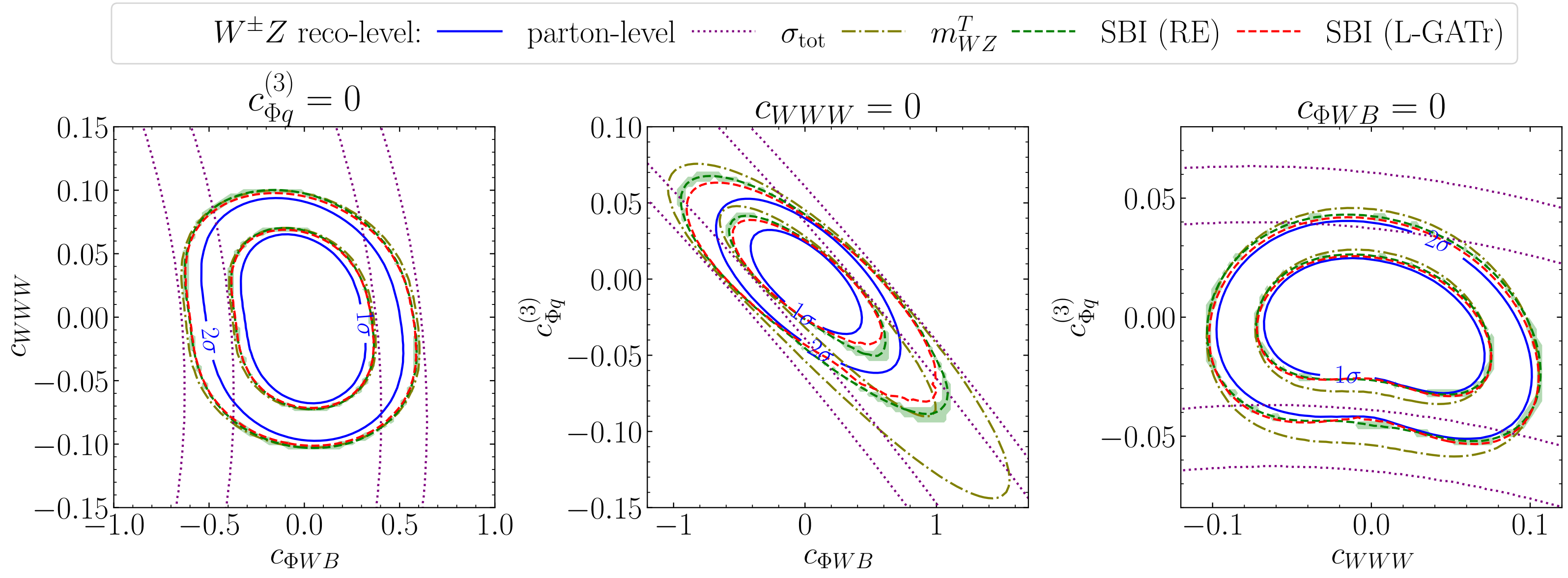
$$D_{\text{opt}}(x|\theta,\theta_0) = \frac{p(x|\theta_0)}{p(x|\theta) + p(x|\theta_0)} \rightarrow \text{likelihood ratio } \frac{p(x|\theta)}{p(x|\theta_0)} = \frac{1 - D_{\text{opt}}}{D_{\text{opt}}}$$

- Neyman-Pearson lemma: likelihood ratio is most powerful statistical test

Unbinned multi-dimensional inference without information loss

Advanced SBI tools

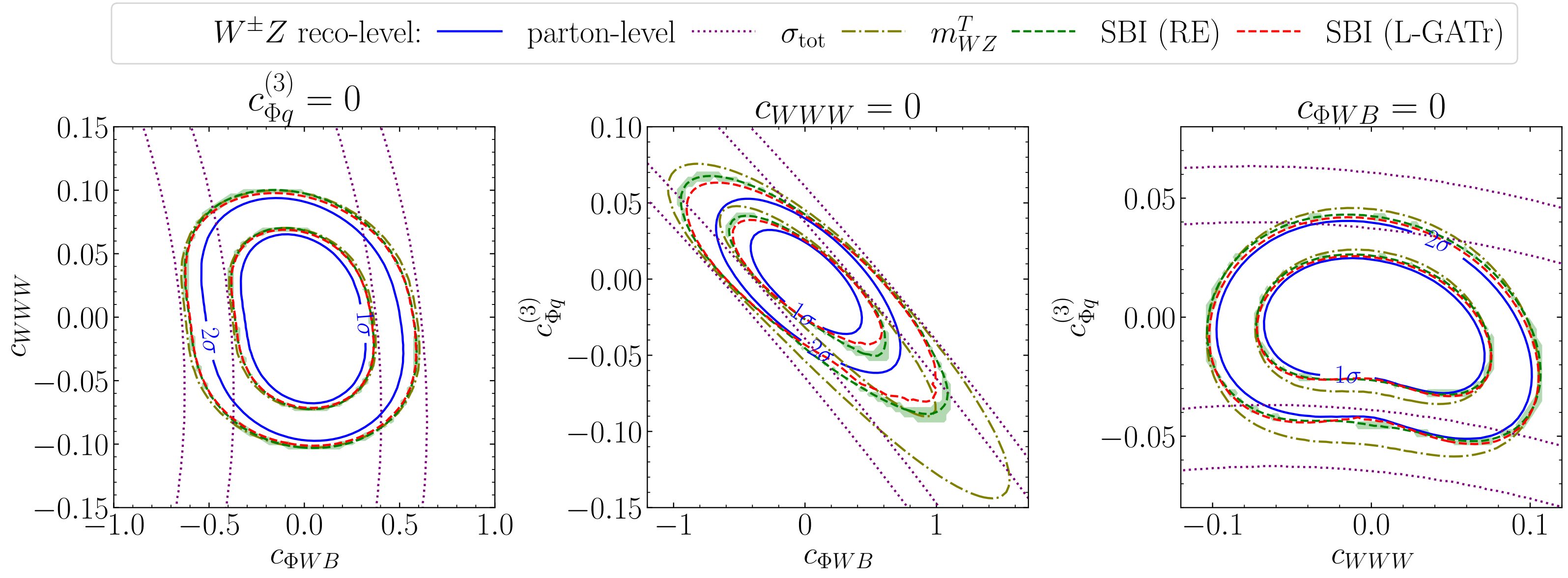
[HB et al.,2410.07315]



- target: SMEFT operators in $W^{\pm}Z$ production
- numerically stable results
- significantly better bounds than for histogram
- variety of cross-checks allows validating results

Advanced SBI tools

[HB et al.,2410.07315]



- target: SMEFT operators in $W^{\pm}Z$ production
- numerically stable results
- significantly better bounds than for histogram
- variety of cross-checks allows validating results

Future directions:

- application to masses, NLO corrections
- more pheno studies
- work towards real data application