Scaling laws for amplitude surrogates

Joaquín Iturriza Ramirez

In collaboration with Anja Butter, Bertrand Laforge, Víctor Bresó Pla and Henning Bahl

IRN Terascale @ Montpellier 2025

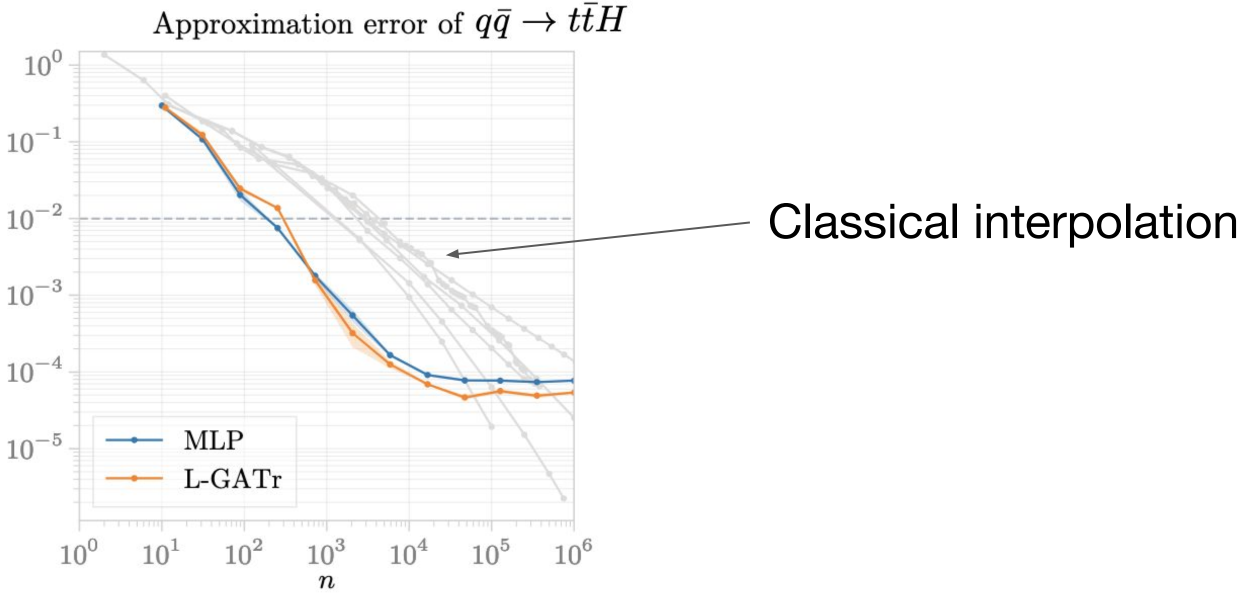
Co-funded by the European Union

Motivation

Scattering amplitudes are at the heart of MC event generators

Calculating higher order scattering amplitudes is **expensive**

ML surrogate models are excellent interpolates



How accurate can we become and what is the price?

arXiv:2412.09534

Motivation

Similar scaling behaviours have been observed in many applications of deep learning

Large Language Modeling[1]

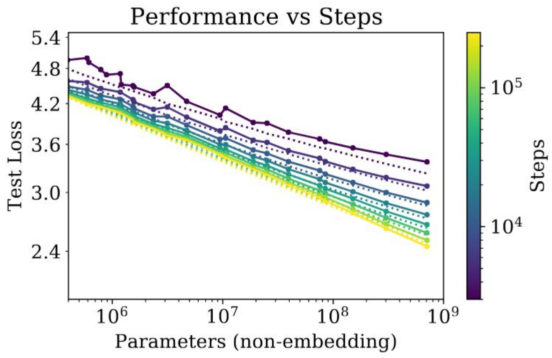
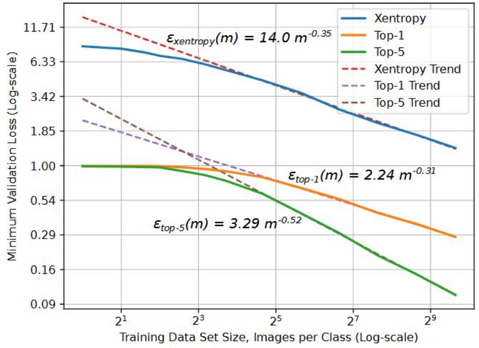
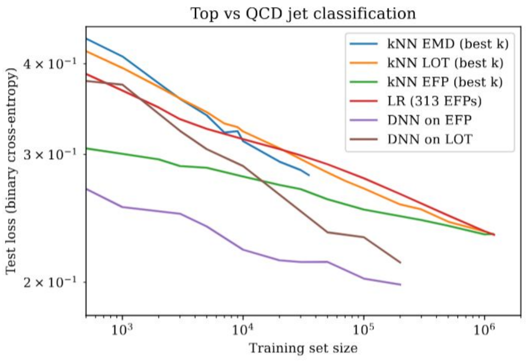


Image Classification[2]



Jet Classification[3]



Performance improves predictably as a power law with **training dataset size** D**, computing resources** C and **number of parameters** N

$$\boxed{L(X,Y,Z) = (X_c/X)^{\alpha_X} + K(Y,Z)}$$

Exponent of power law could be related to **intrinsic dimensionality**[4]

[1] arXiv:2001.08361, [2] arXiv:1712.00409, [3] arXiv:2312.02264, [4] arXiv:2102.06701

Can we determine scaling laws for scattering amplitude surrogates?

Compare

- Many different processes:

  $q\bar{q} \to t\bar{t}H,\ q\bar{q} \to Z + ng,\ q\bar{q} \to WZ + ng,\ q\bar{q} \to WWZ,\ gg \to \gamma\gamma + ng$
- Scalings in D, C and N
- Different loss functions: **MSE** vs **Heteroscedastic Loss** for uncertainties estimation
- Different architectures: **MLP** vs **LloCa-Transformer**

If scaling laws are universal $\longrightarrow$ **We predict desired accuracy for given resources**

We know the **degrees of freedom**: Test relation between scaling and intrinsic dimensionality

Data

4-momentum of particles in the process $\longrightarrow$ **Amplitude**

All the data is generated with MadGraph, all amplitudes calculated at **lowest order**

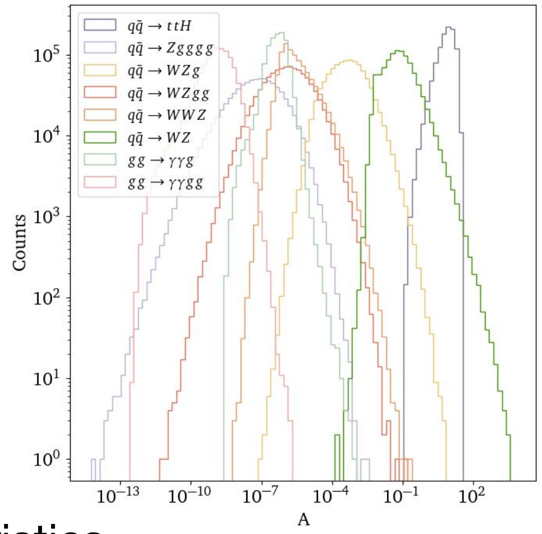
Main case study $q\bar{q} \rightarrow t\bar{t}H$

Jet-associated Z production: $q\bar{q} \rightarrow Z + ng$

Jet-associated W Z production: $q\bar{q} \rightarrow WZ,\ q\bar{q} \rightarrow WZg,\ q\bar{q} \rightarrow WZgg$

W W Z production: $q\bar{q} \rightarrow WWZ$

Jet-associated di-photon production: $gg \rightarrow \gamma\gamma g,\ gg \rightarrow \gamma\gamma gg$



Wide **variety** of processes covering many different characteristics

Neural Networks

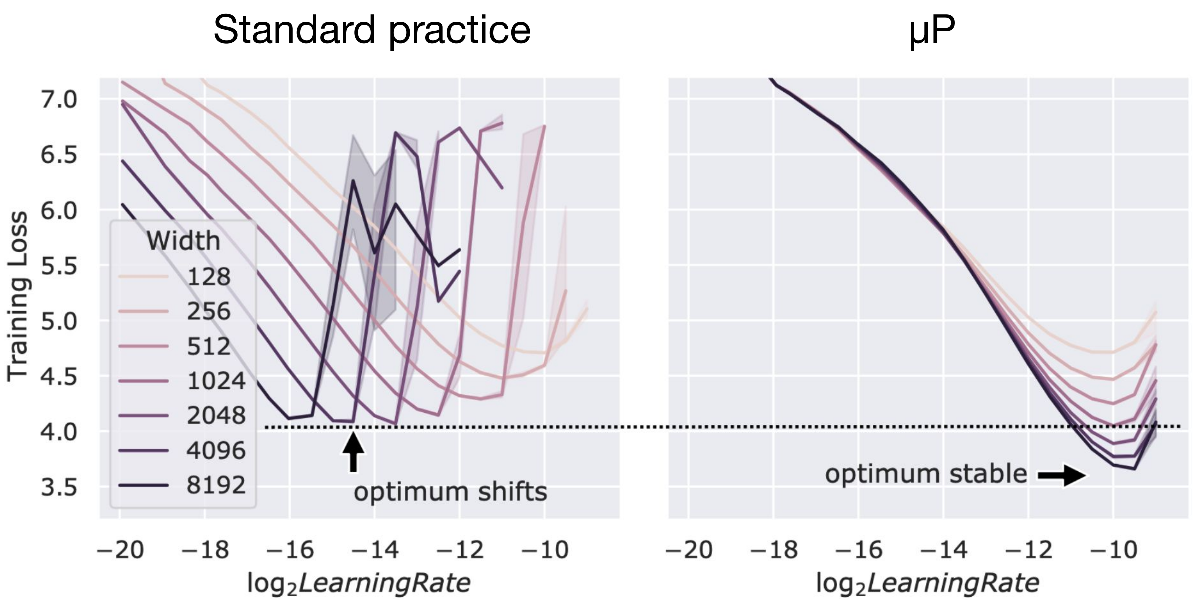
MLP set up:

- Preprocessing:
  - Inputs: **Invariants** derived from 4-vectors
  - Amplitudes: **Log-standardized** $\hat{a} = \frac{\log(a) - \overline{\log(a)}}{\sigma_{\log(a)}}$ (except $q\bar{q} \to t\bar{t}H$)
- Unless explicitly changed:
  - 500 Hidden neurons, 4 hidden layers, **~10^6 parameters**
  - **10^6 Training dataset size**
  - **10^4 Epochs**
- **GELU** non-linearities
- Batchsize = 256
- **Cosine Annealing** scheduler
- **Hyperparameter Transfer**

Hyperparameter Transfer[1]

Optimal HP remain consistent across widths

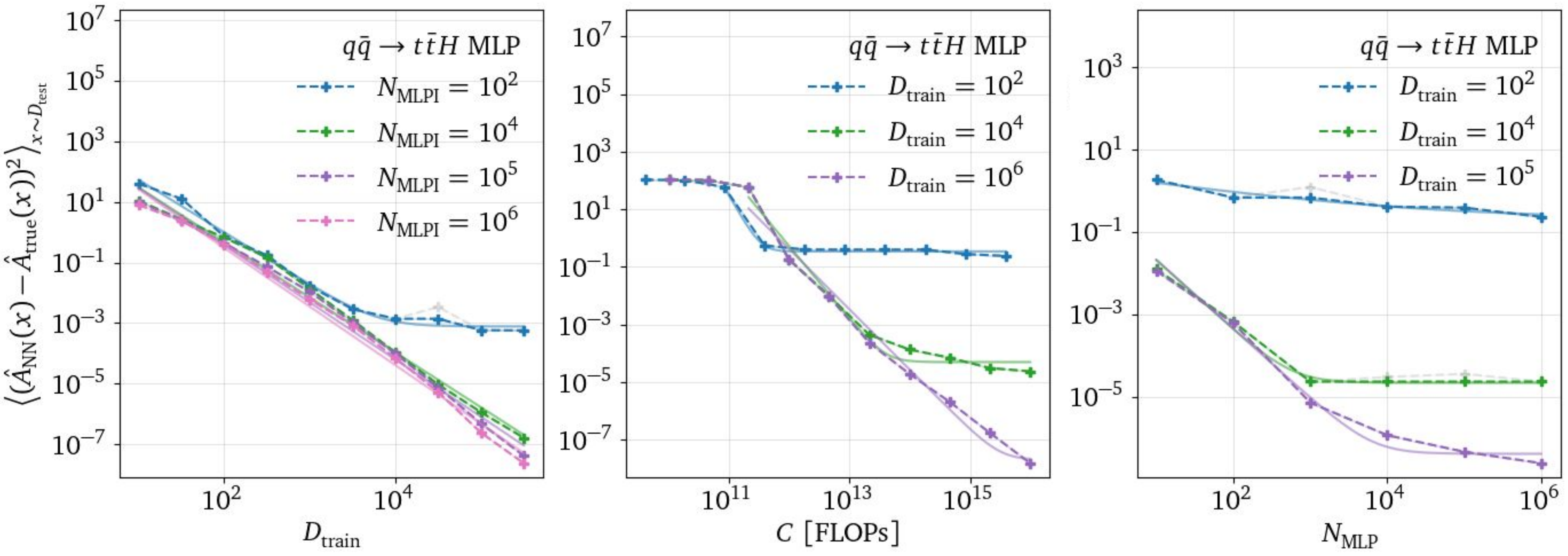
- Better initialization
- Better optimizer
- Optimizer treats layers differently
- Results are comparable or better that standard practice



Allows to optimize HP on a smaller (cheaper) model **once** and use them for different network sizes

[1] arXiv:2203.03466

Results: MLP, $q\bar{q} \to t\bar{t}H$ MSE loss



Very clean power laws, consistent slopes

Losses

MSE: $$\boxed{\mathcal{L}_{\text{MSE}} = \left\langle (A_{\text{true}}(x) - A_{\text{NN}}(x))^2 \right\rangle_{x \sim D_{\text{train}}}}$$

Heteroscedastic Loss:

Assume amplitude regression follows a normal distribution $p(A|x) = \mathcal{N}(A|\bar{A}(x), \sigma^2(x))$

Minimize negative log-likelihood:

$$\mathcal{L} = -\left\langle \log p(A|x) \right\rangle_{x \sim D_{\text{train}}} \longrightarrow \boxed{\mathcal{L}_{\text{het}} = \left\langle \frac{(A_{\text{true}}(x) - \bar{A}(x))^2}{2\sigma^2(x)} + \log \sigma(x) \right\rangle_{x \sim D_{\text{train}}}}$$

The NN predicts 2 outputs: $\bar{A}(x)$ and $\sigma(x)$

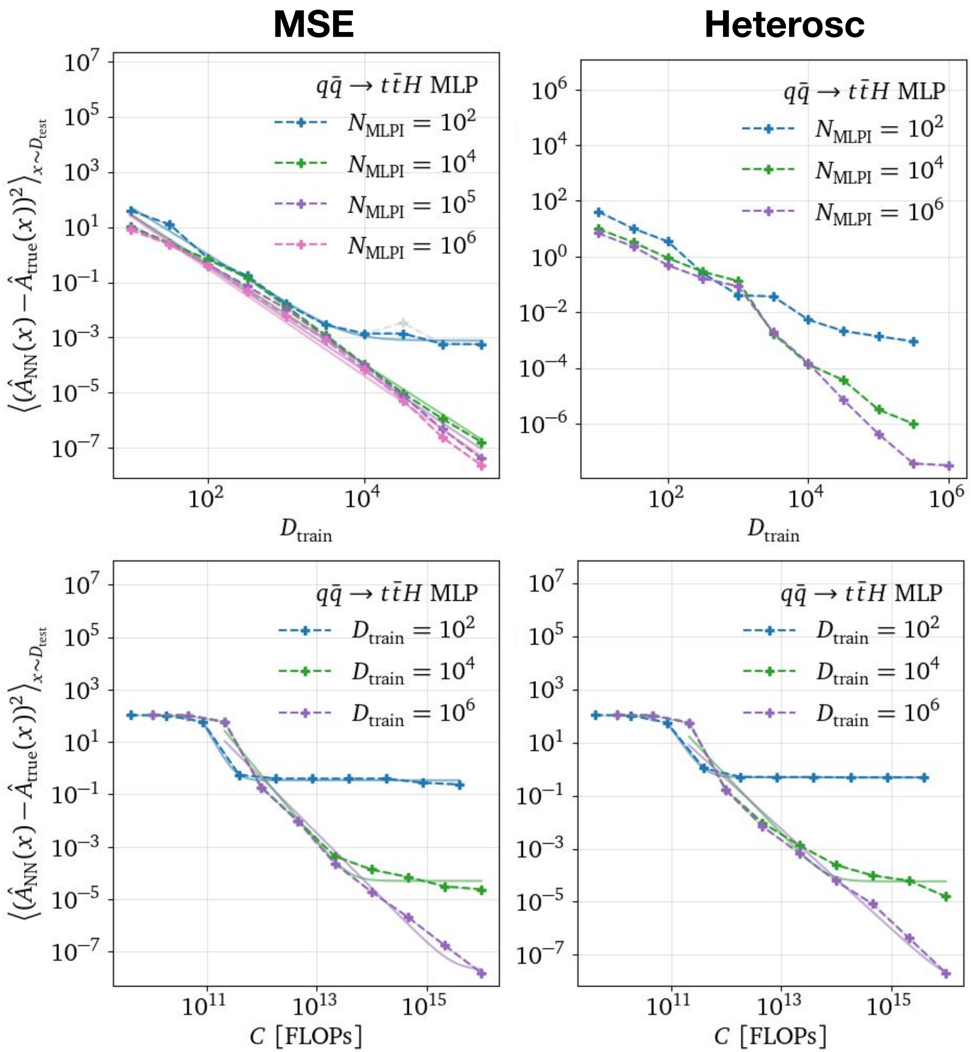
If it's well calibrated $t(x) = \frac{A_{\text{NN}}(x) - A_{\text{true}}(x)}{\sigma(x)}$ should follow a $\mathcal{N}(0, 1)$

Results: MLPI, $q\bar{q} \to t\bar{t}H$, Heteroscedastic loss

Slightly worse performance for scaling in dataset size

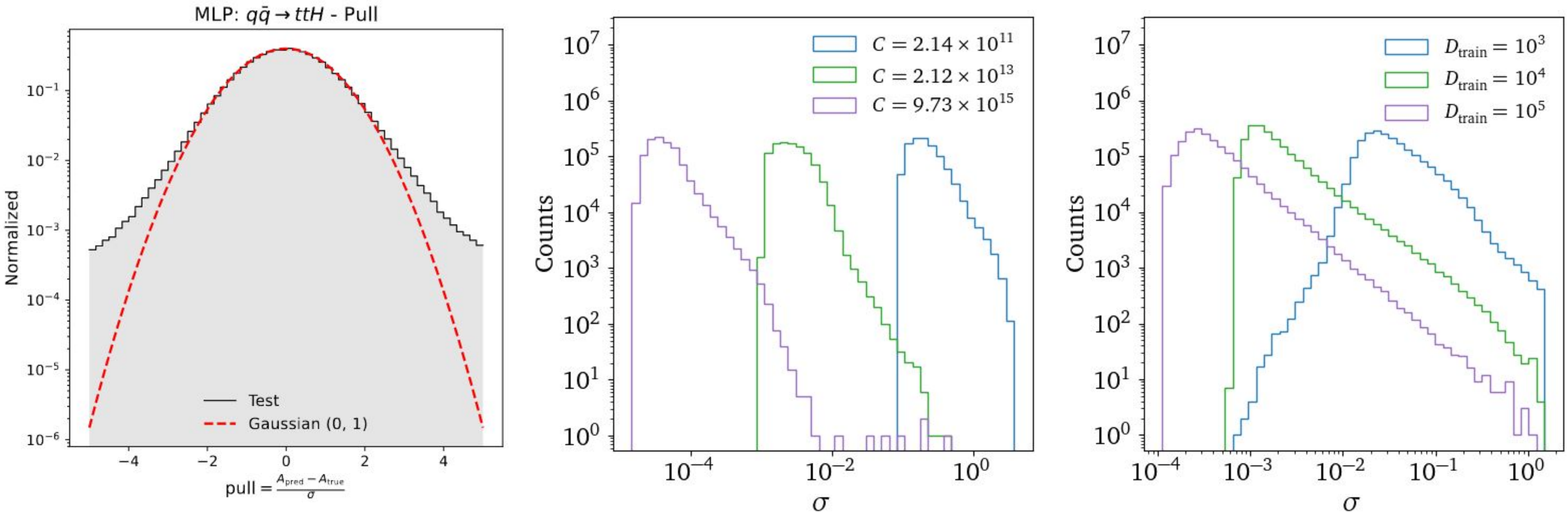
Very similar performance for scaling in computing resources

Important to note: Trained on Heterosc loss, showing MSE



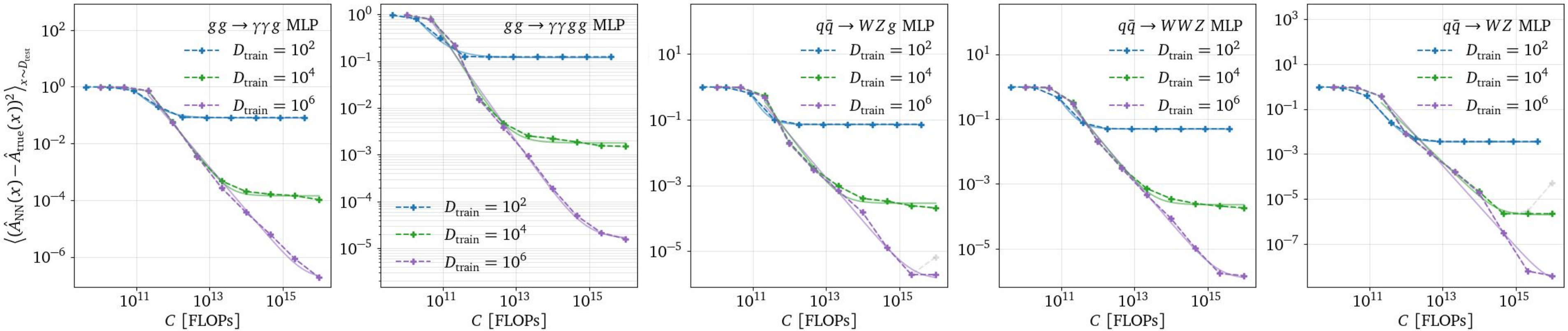
Uncertainties

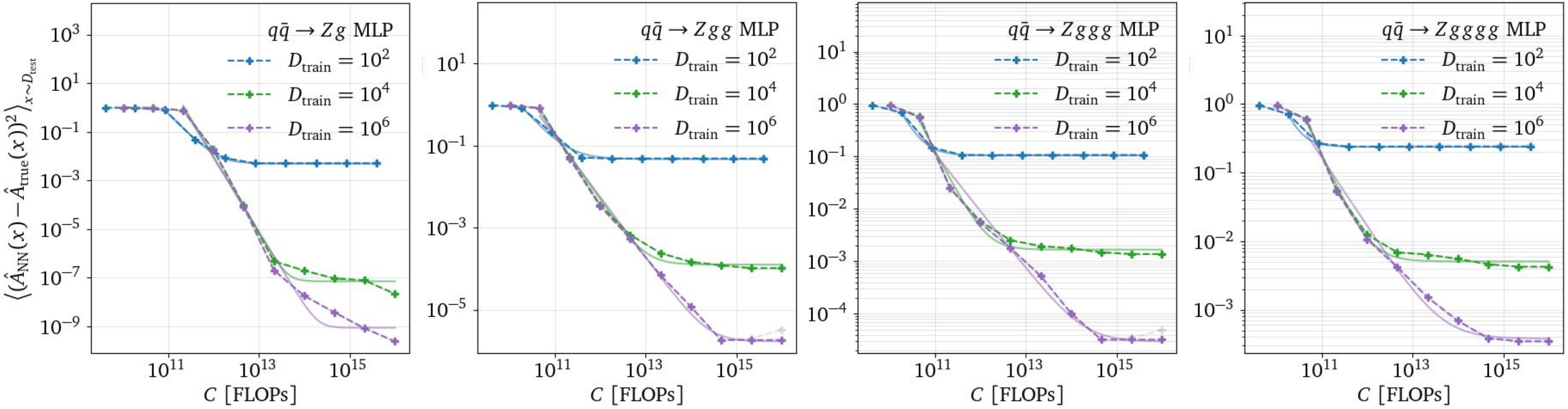
Calibration | N=10^6 | N=10^4



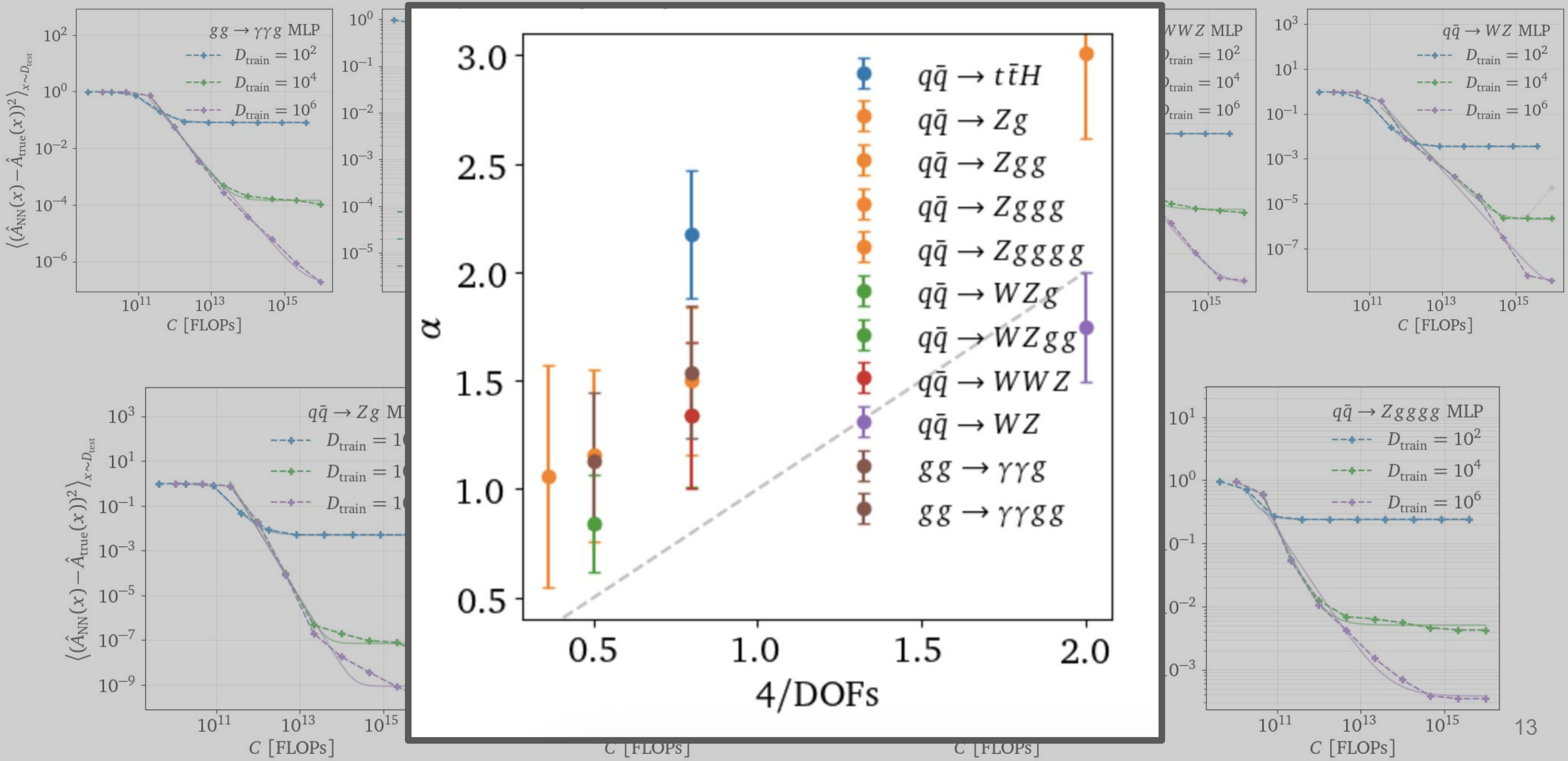
Well calibrated uncertainties for large enough dataset sizes

Results scaling on computing resources

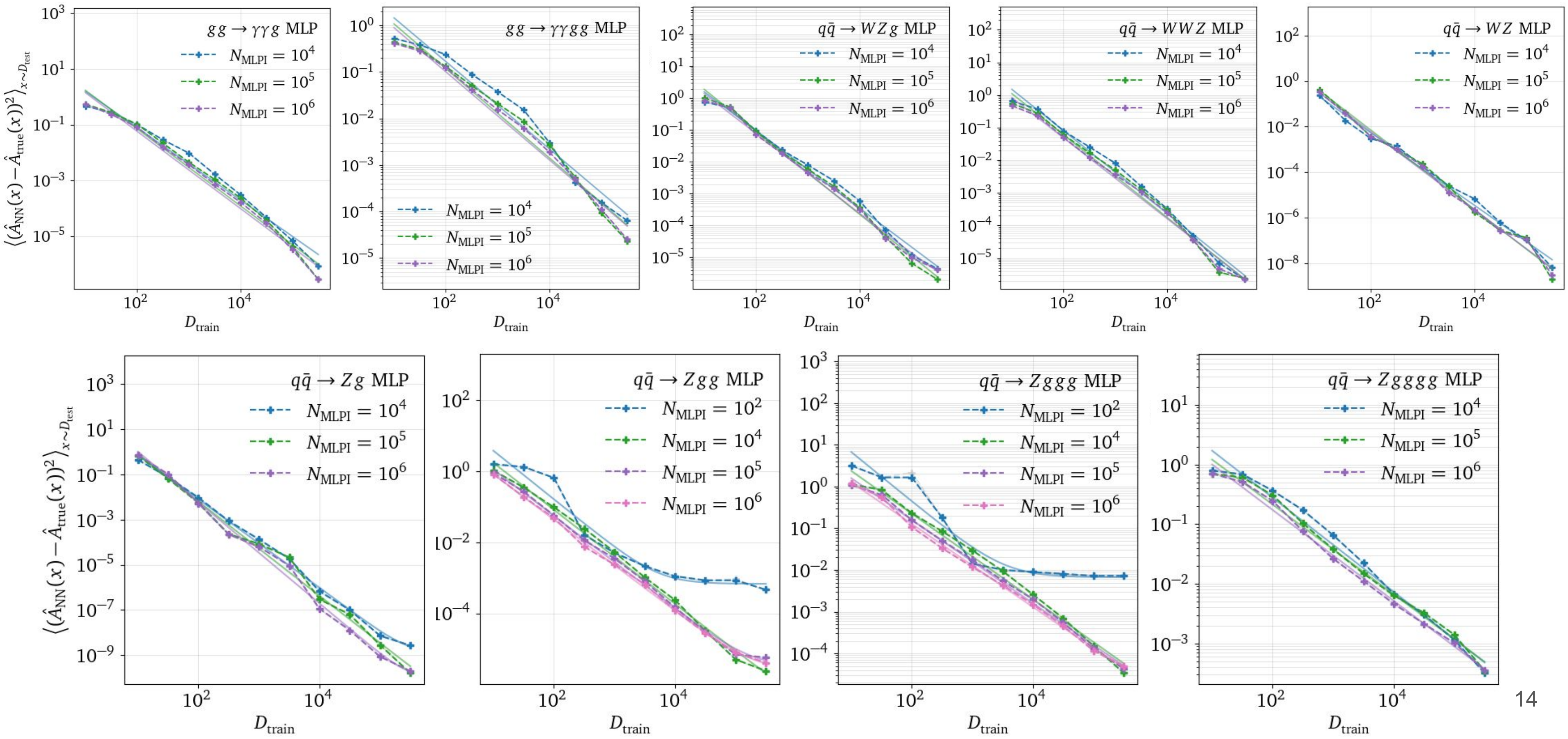




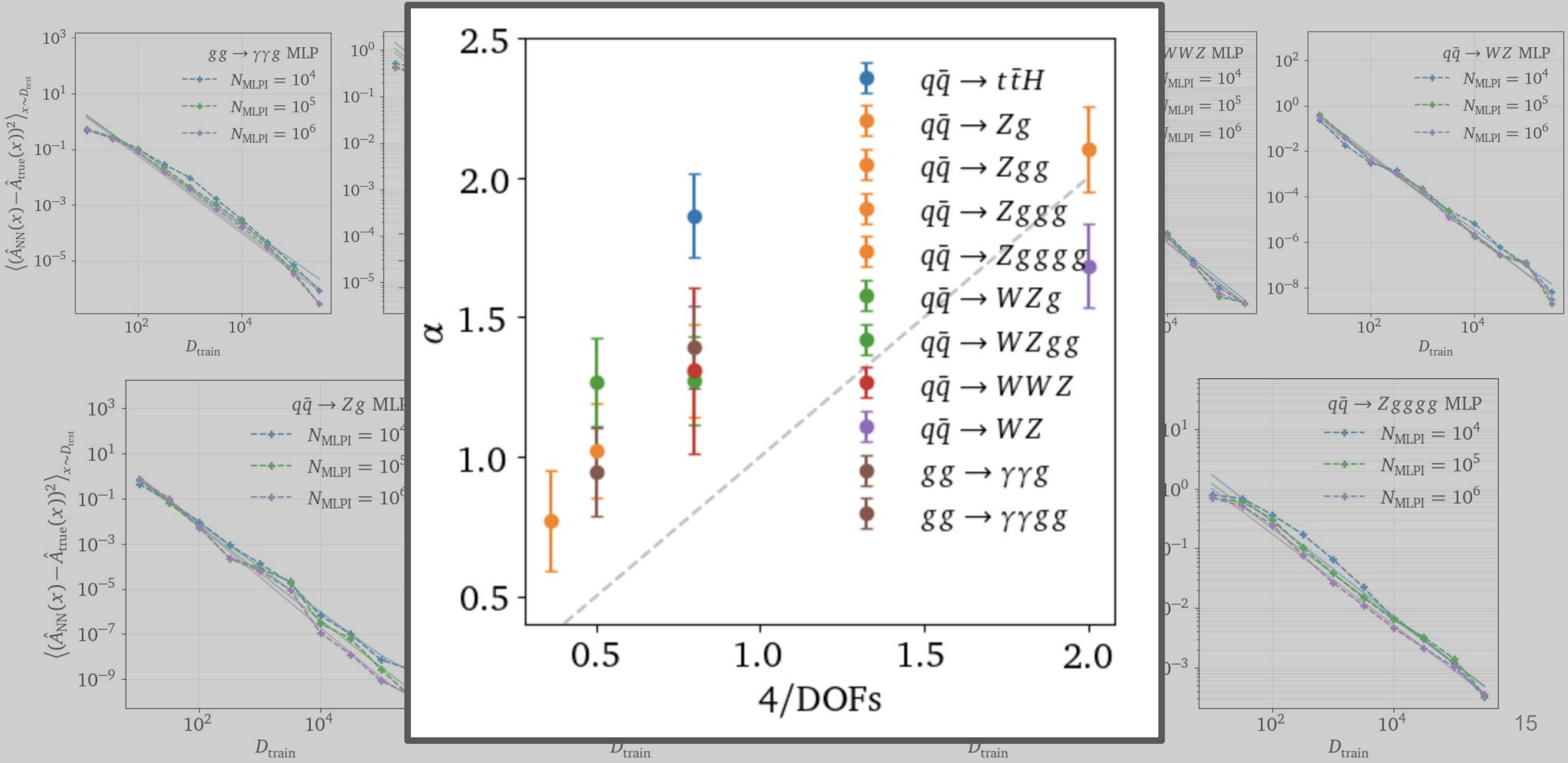
Results scaling on computing resources



Results scaling on dataset size

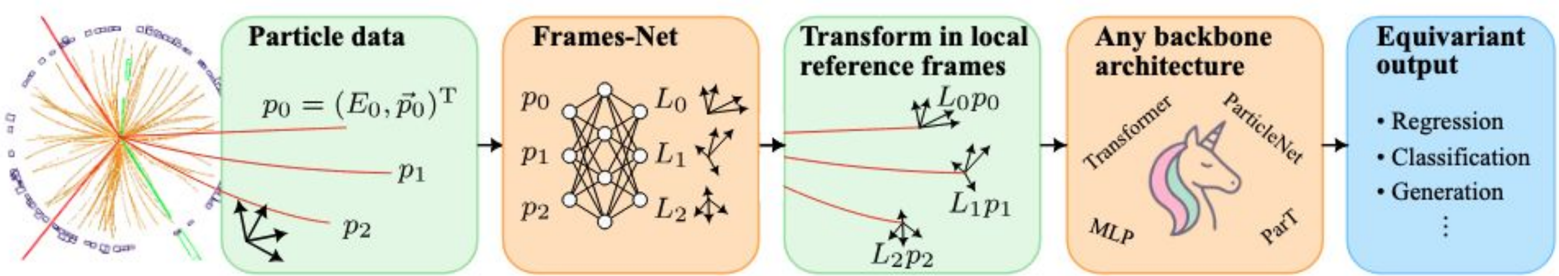


Results scaling on dataset size



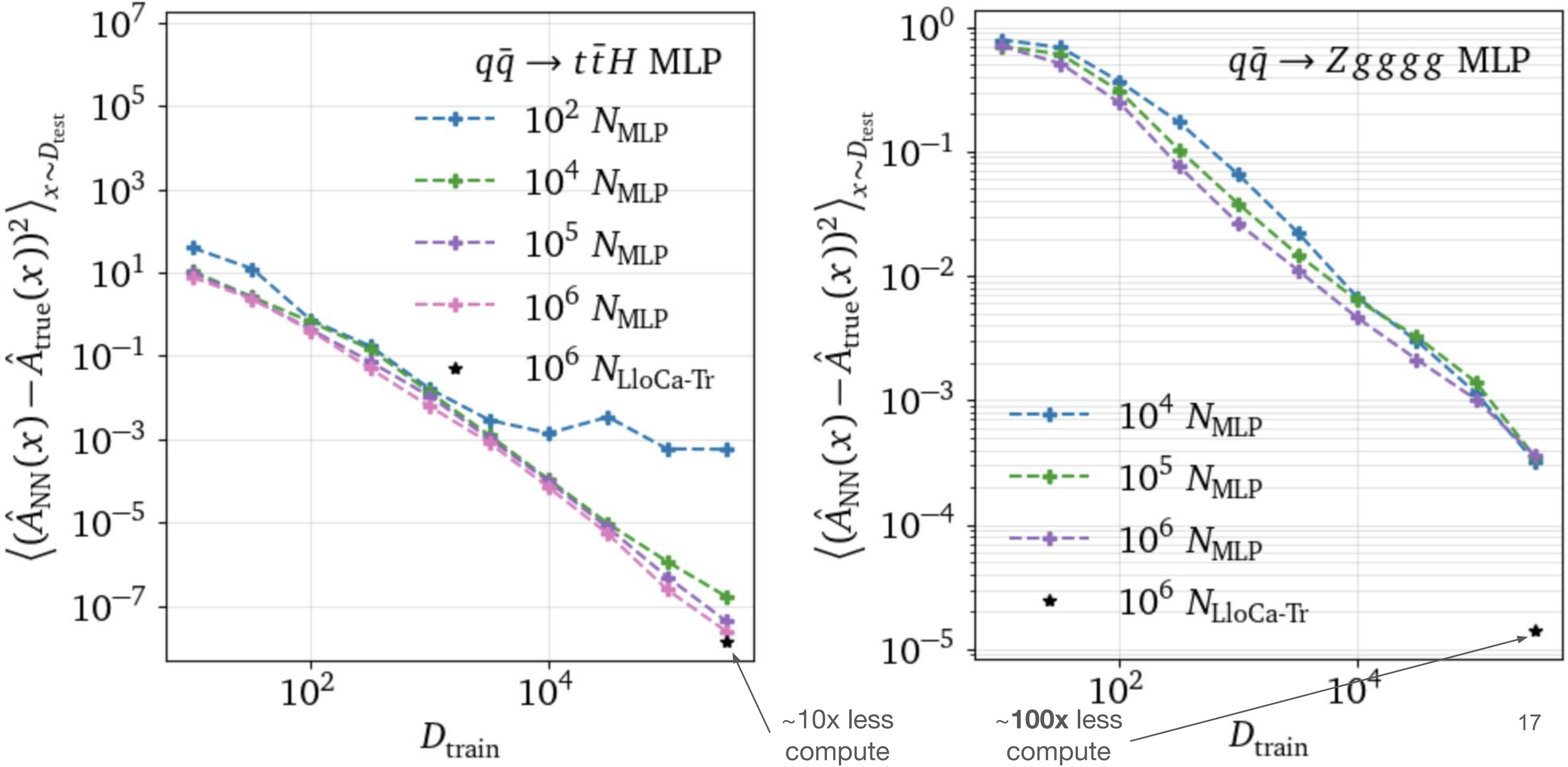
Work in progress: LloCa[1]

Permutation invariance and equivariance yield better results



[1] arXiv:2505.20280

Work in progress: LloCa



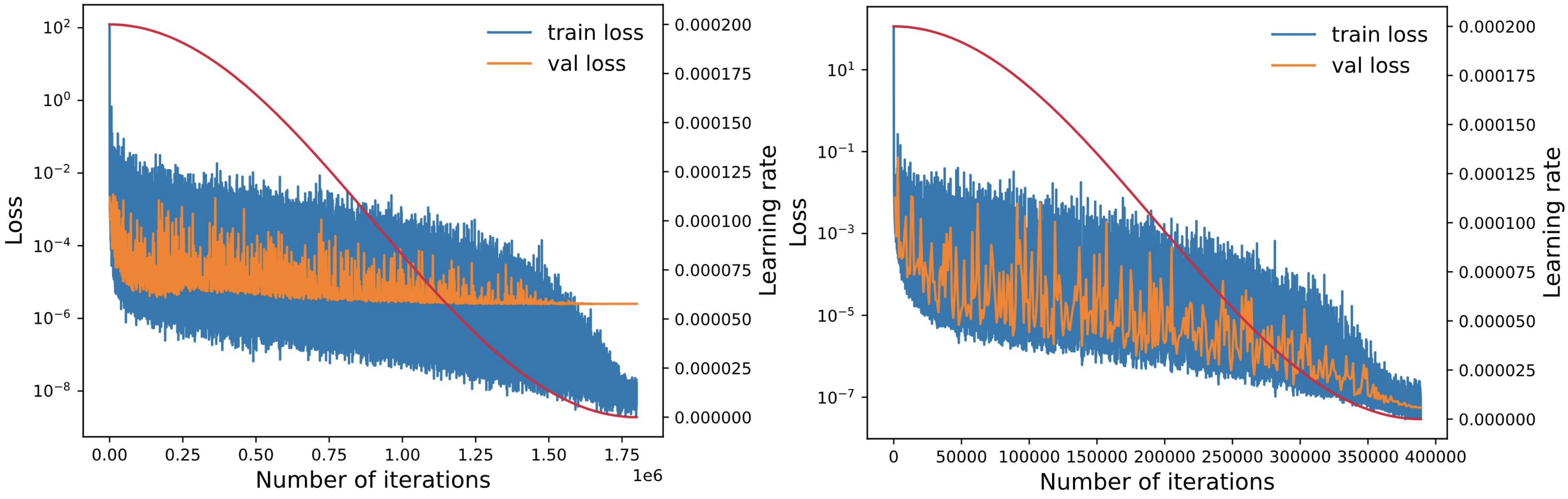
Conclusions

- Observed very clean and predictable scaling laws for amplitude surrogates
- Dataset size bottlenecks are the most important
- Scaling with MSE and Heterosc loss
- Well calibrated uncertainties across many orders of magnitude
- Observed relationship between scaling and DOFs
- Promising results with equivariant NNs

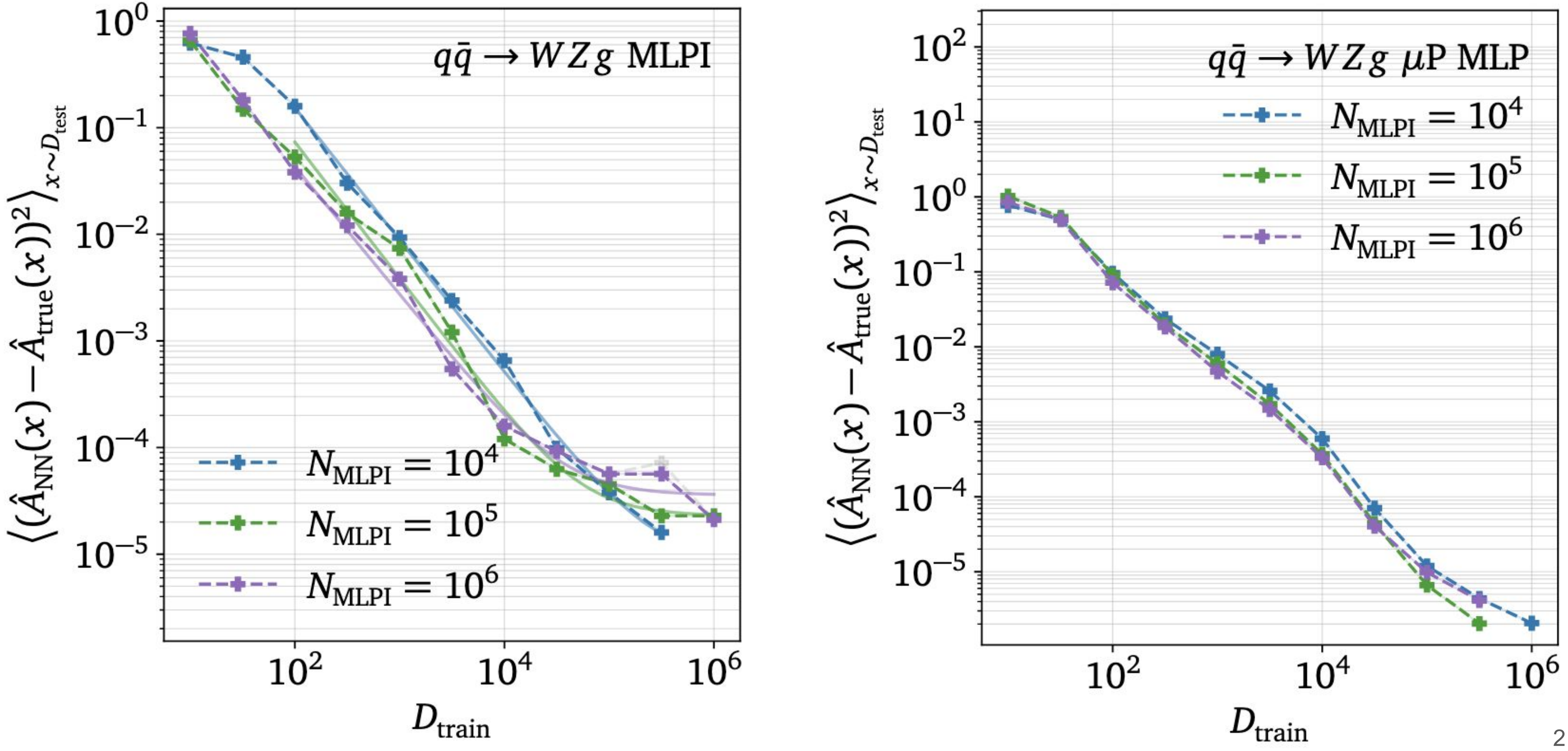
Thanks!

Scheduler

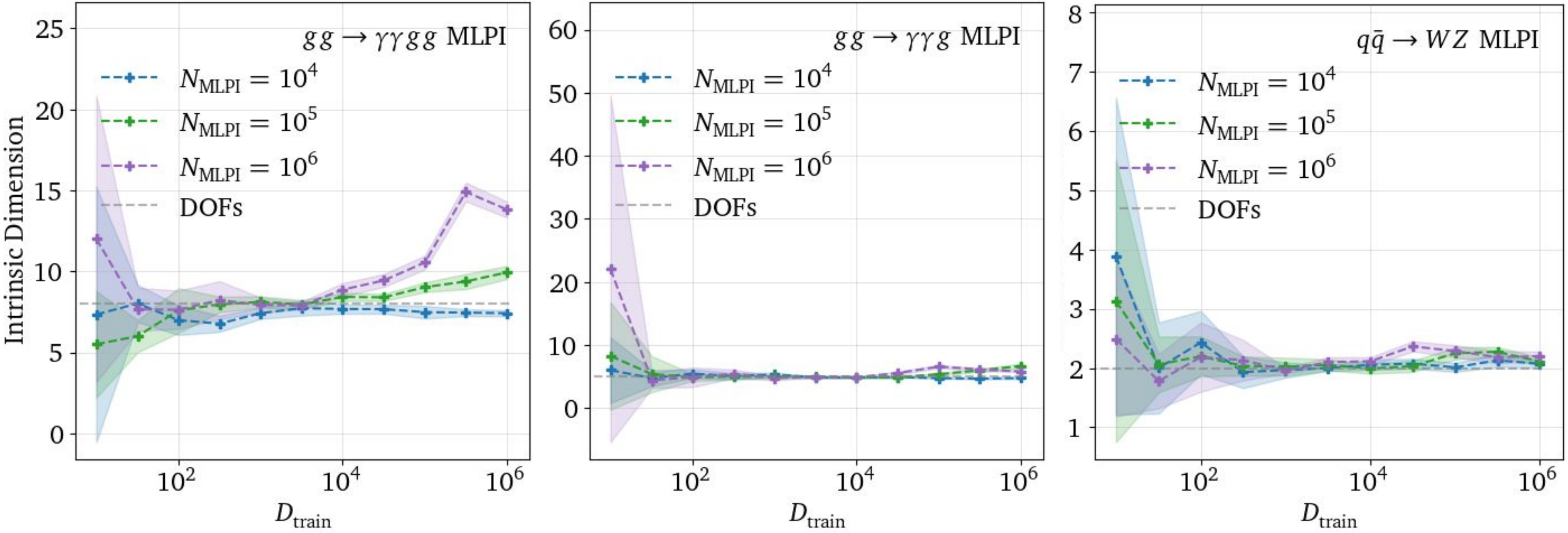
Scheduler has to go to 0 **just at the right time**



muP vs normal MLP



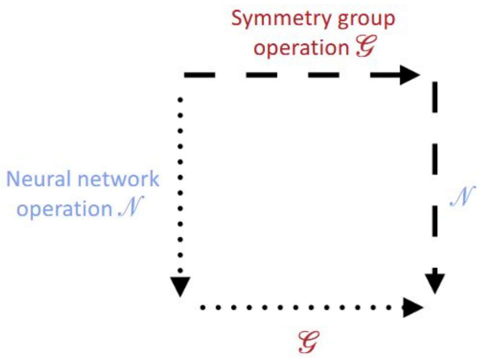
Internal Intrinsic Dimension

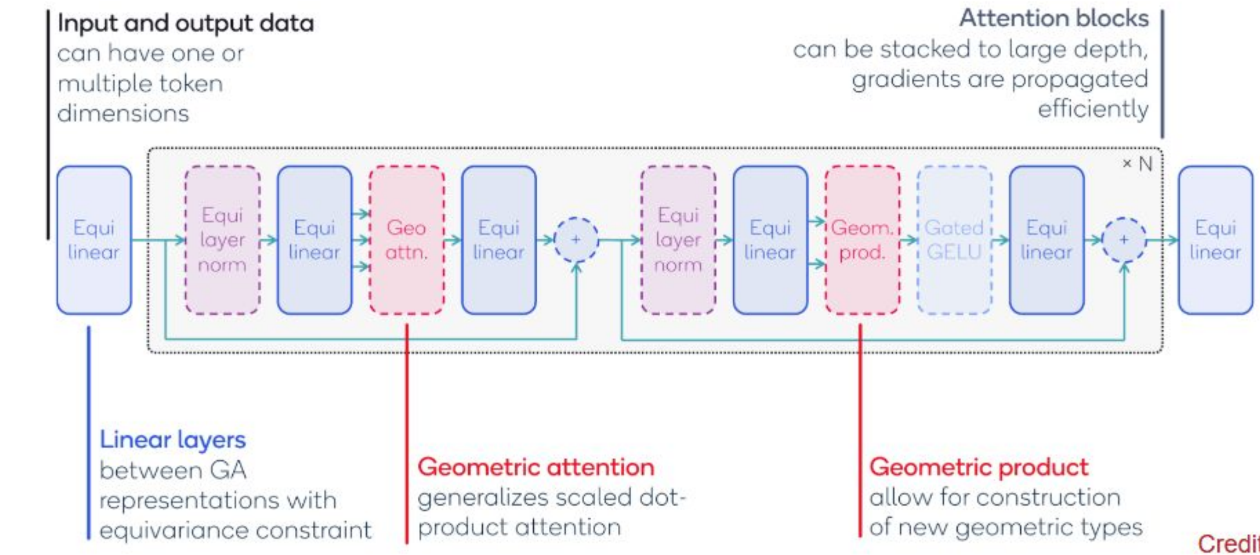


L-GATr

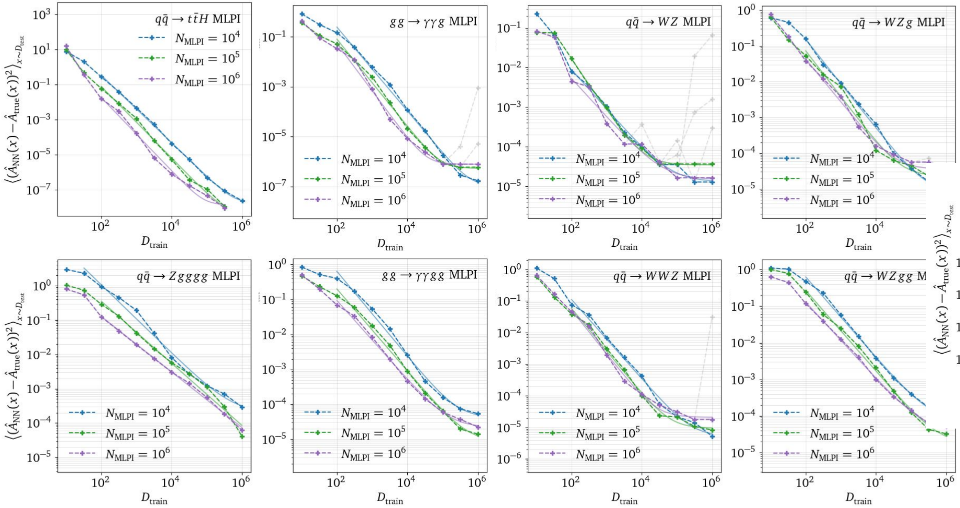
L-GATr = Equivariance + Transformer

$$\mathscr{G}(\mathscr{N}(x)) = \mathscr{N}(\mathscr{G}(x))$$

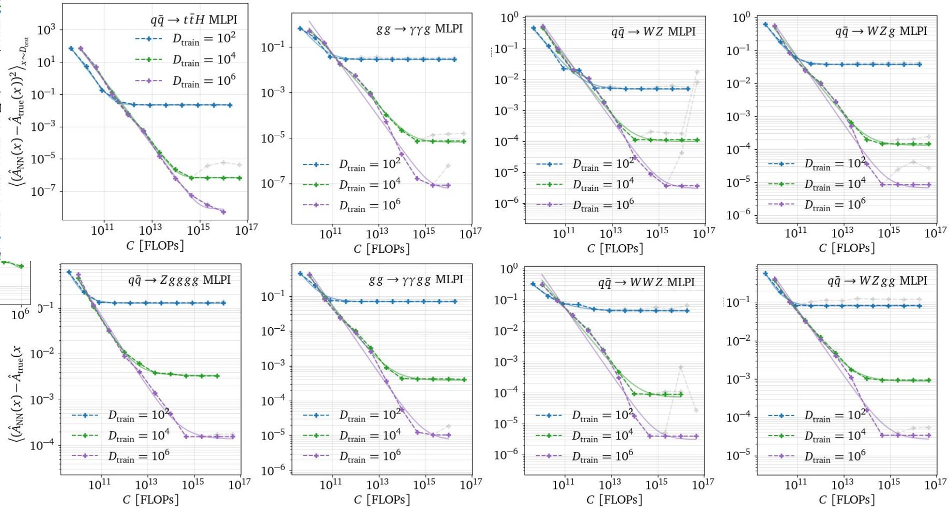




Credits to Johann Brehmer

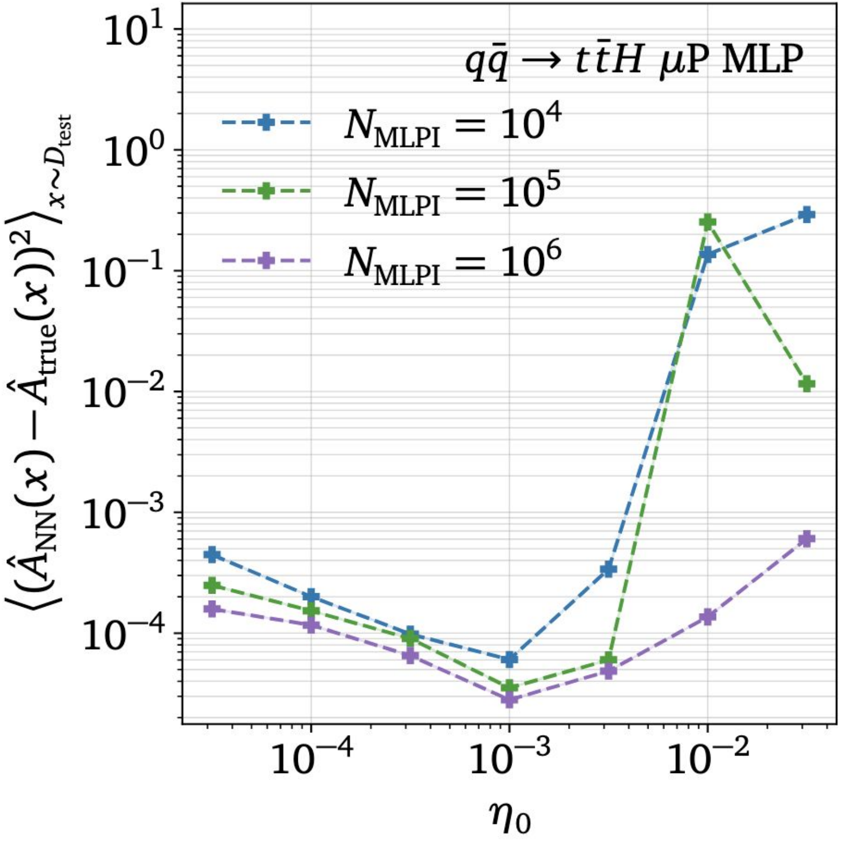


$q\bar{q} \to t\bar{t}H$ MLPI
$gg \to \gamma\gamma g$ MLPI
$q\bar{q} \to WZ$ MLPI
$q\bar{q} \to WZg$ MLPI
$q\bar{q} \to Zgggg$ MLPI
$gg \to \gamma\gamma gg$ MLPI
$q\bar{q} \to WWZ$ MLPI
$N_{\mathrm{MLPI}} = 10^4$
$N_{\mathrm{MLPI}} = 10^5$
$N_{\mathrm{MLPI}} = 10^6$
D_{train}
$\langle (\hat{A}_{\mathrm{NN}}(x) - \hat{A}_{\mathrm{true}}(x))^2 \rangle_{x\sim D_{\mathrm{test}}}$



$q\bar{q} \to t\bar{t}H$ MLPI
$gg \to \gamma\gamma g$ MLPI
$q\bar{q} \to WZ$ MLPI
$q\bar{q} \to WZg$ MLPI
$q\bar{q} \to Zgggg$ MLPI
$gg \to \gamma\gamma gg$ MLPI
$q\bar{q} \to WWZ$ MLPI
$q\bar{q} \to WZgg$ MLPI
$D_{\mathrm{train}} = 10^2$
$D_{\mathrm{train}} = 10^4$
$D_{\mathrm{train}} = 10^6$
C [FLOPs]
$\langle (\hat{A}_{\mathrm{NN}}(x) - \hat{A}_{\mathrm{true}}(x))^2 \rangle_{x\sim D_{\mathrm{test}}}$

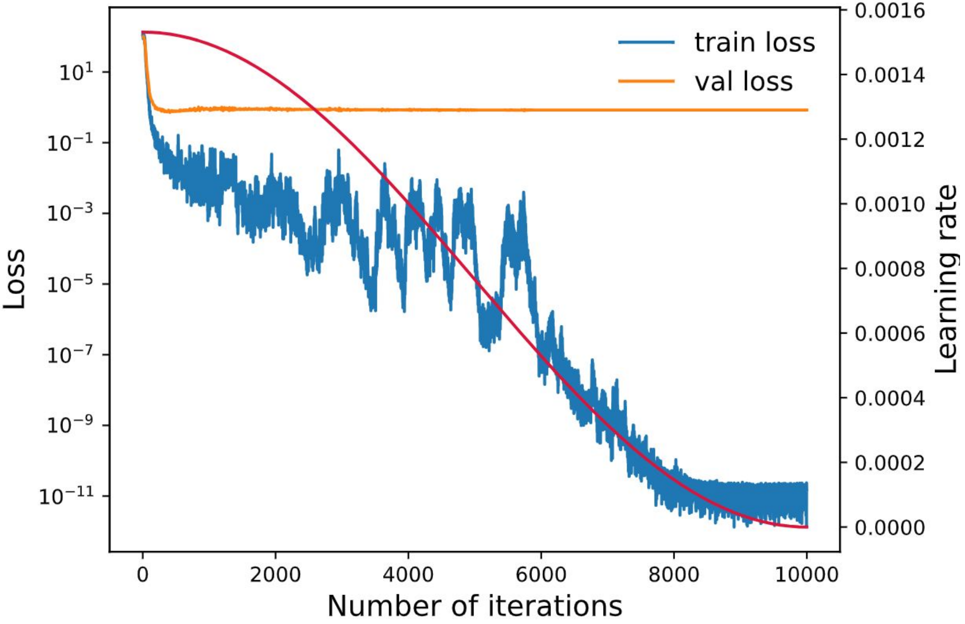
muP

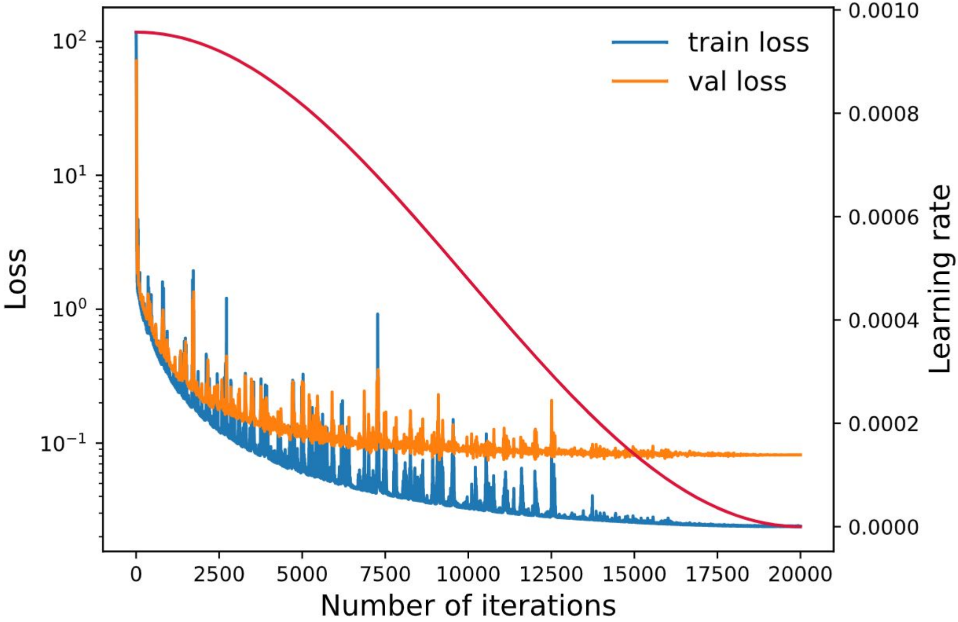


$q\bar{q} \to t\bar{t}H$ μP MLP
$N_{\mathrm{MLPI}} = 10^4$
$N_{\mathrm{MLPI}} = 10^5$
$N_{\mathrm{MLPI}} = 10^6$
$\left\langle (\hat{A}_{\mathrm{NN}}(x) - \hat{A}_{\mathrm{true}}(x))^2 \right\rangle_{x \sim D_{\mathrm{test}}}$
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-4}
10^{-3}
10^{-2}
η_0

muP overfitting for small datasets

100 training points





Overfitted Het loss