Identifying the relevant dependencies of the neural network response on characteristics of the input space

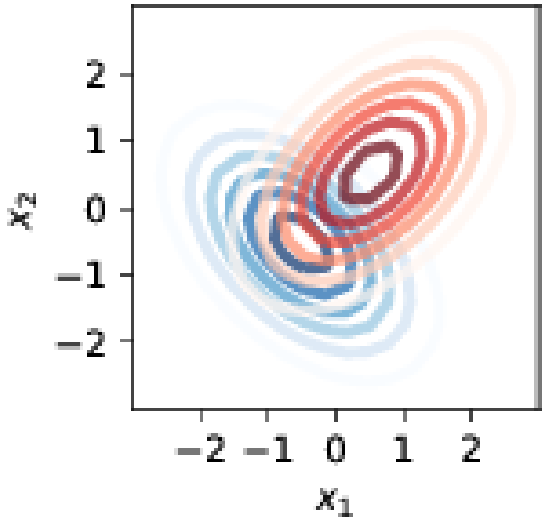
Raphael Friese, Günter Quast, Roger Wolf, Sebastian Wozniewski, Stefan Wunsch
stefan.wunsch@cern.ch

KIT ETP / CERN EP-SFT

Example

Neural network trained as classifier on a dataset with:

- two variables x_1 and x_2
- two processes **signal** and **background**



Is the response of the trained neural network mainly dependent on

- the **marginal distributions** of x_1 and/or x_2?
- the **correlation** of x_1 and x_2?

Motivation

Neural networks gain importance in physics analyses in comparison to cut-based approaches, but pose **new challenges for the estimation of the systematic uncertainties**:

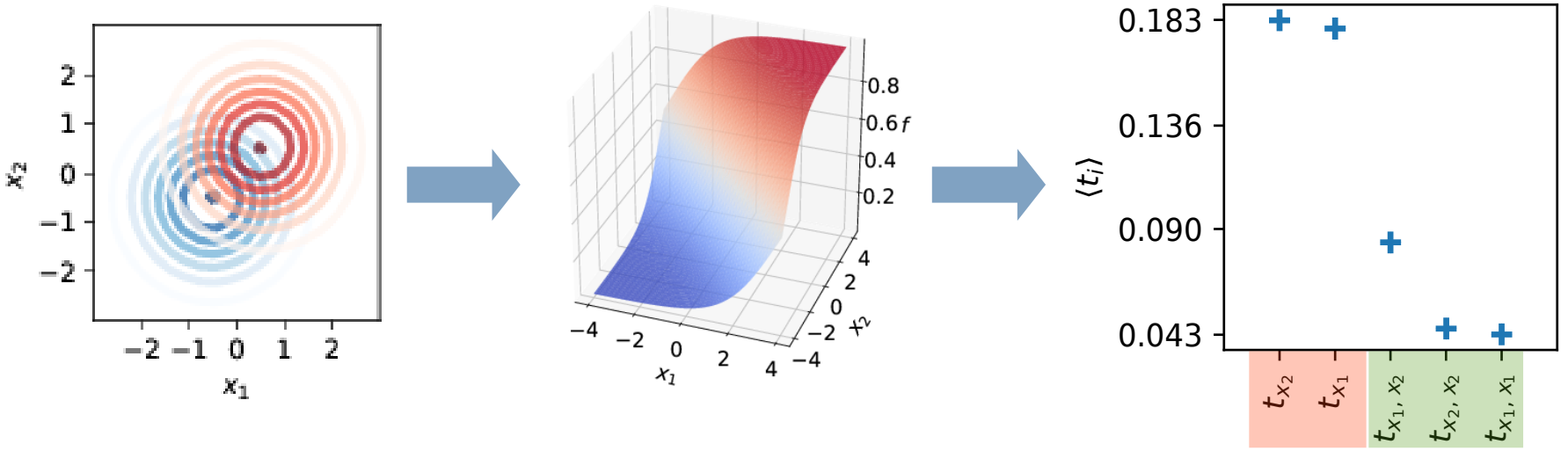
- **Multi-dimensional "cuts"** performed by the neural network are **incomprehensibly** encoded in numerous free parameters of the architecture.
- **Same neural network** architecture may **perform different tasks** based on the training.
- Neural network may **exploit higher-dimensional features** of the inputs, e.g., correlations, which could be **wrongly modeled in the training dataset**.

Key to proper estimation of **systematic uncertainties**:
Precise understanding of the trained neural network and the **relevant dependencies of the neural network response on the inputs**.

Approach

1) **Taylor expansion of the trained neural network**

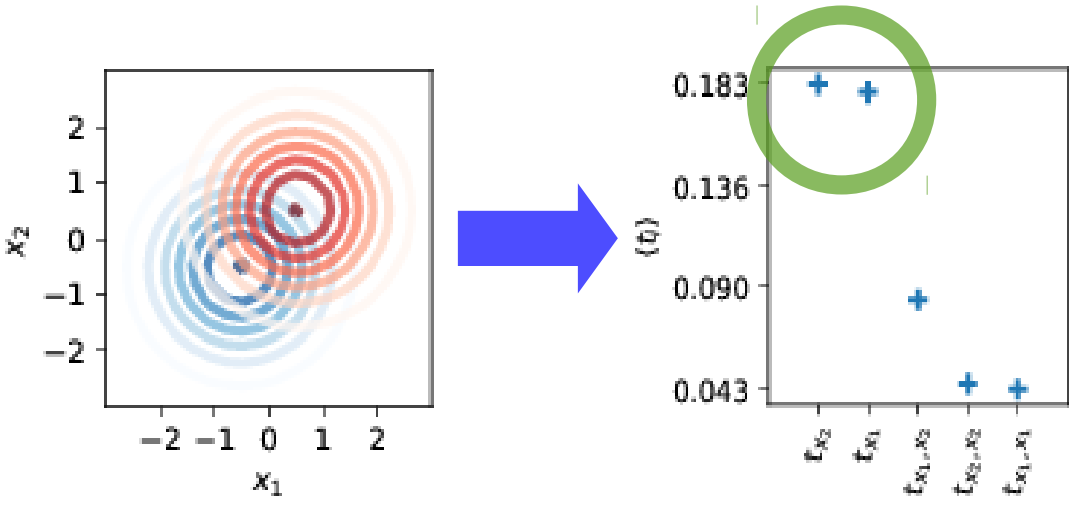
2) Identification of the **Taylor coefficients** with **features** of the inputs



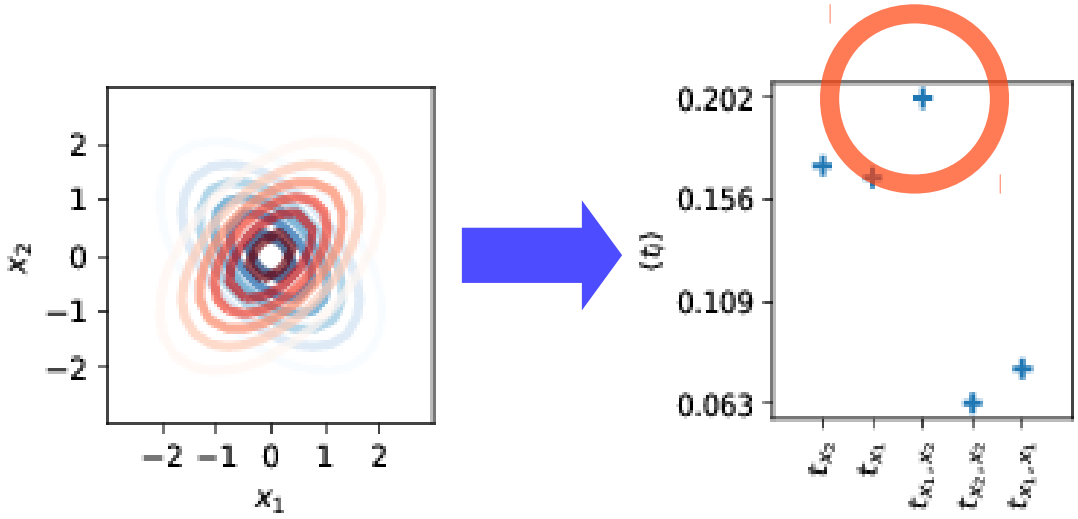
$$T(x,y) = f(a,b) + (x-a)f_x(a,b) + (y-b)f_y(a,b) + \frac{1}{2!}\Big((x-a)^2 f_{xx}(a,b) + 2(x-a)(y-b)f_{xy}(a,b) + (y-b)^2 f_{yy}(a,b)\Big) + \cdots$$

Features: **Marginal distributions** **Correlations**

Application on toy scenarios

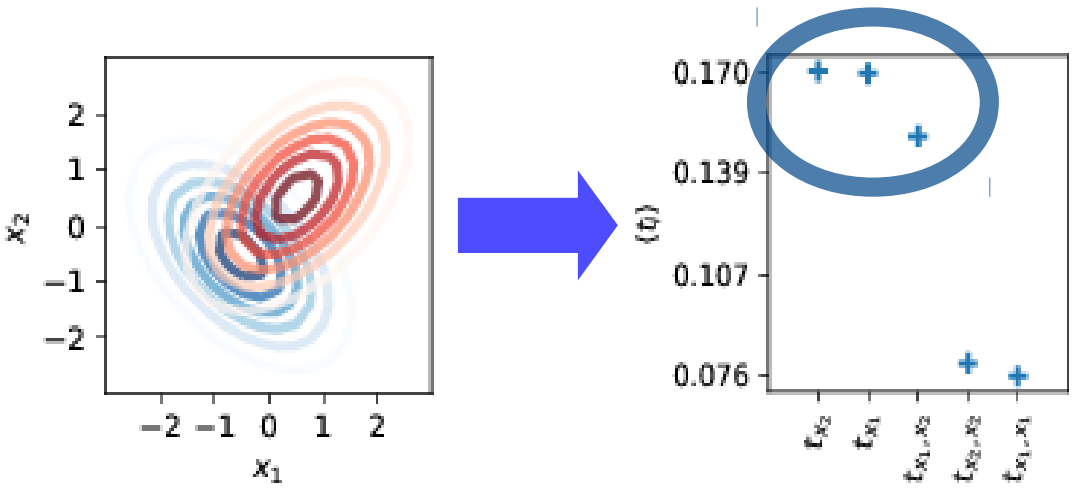


Separation by **marginal distributions** visible by **<t_{x1}>** and **<t_{x2}>**.

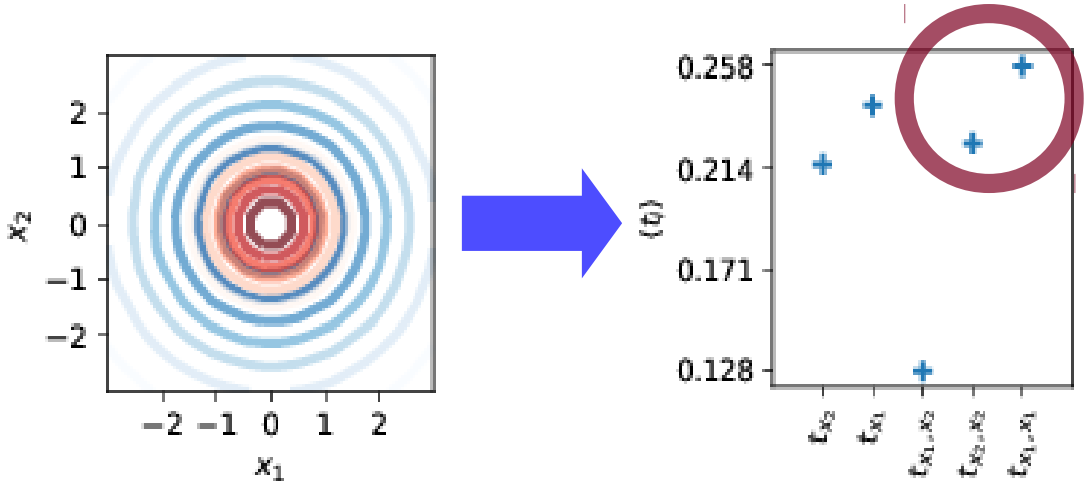


Separation by **correlation** visible by **<$t_{x1,x2}$>**.

Application on toy scenarios



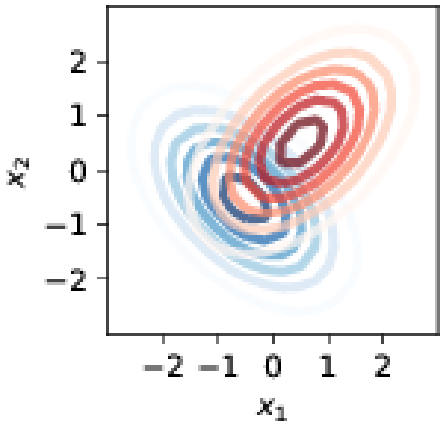
Scenarios with a **mixture of features** can be identified.



Separation due to width of distribution visible by $\langle t_{x1,x1} \rangle$ and $\langle t_{x2,x2} \rangle$.

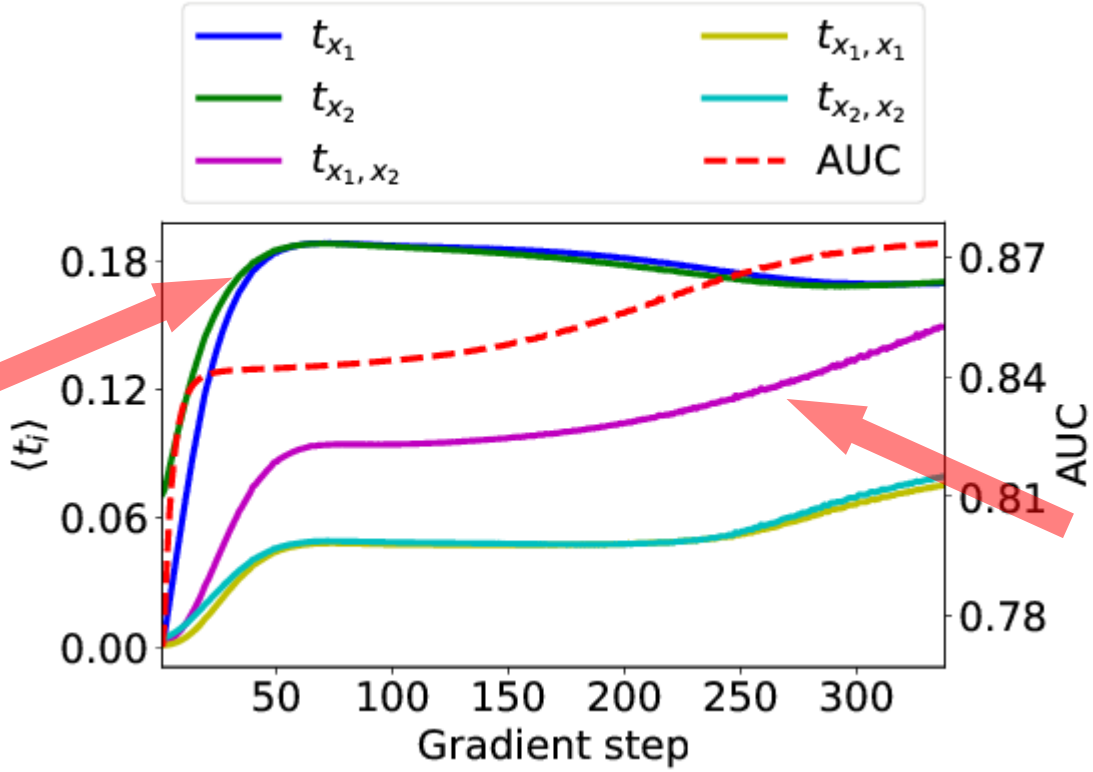
Visualization of the learning progress

Analyzed scenario with mixed features:



Performed analysis of Taylor coefficients **after each gradient step**:

First:
Learned separation by **marginal distributions**



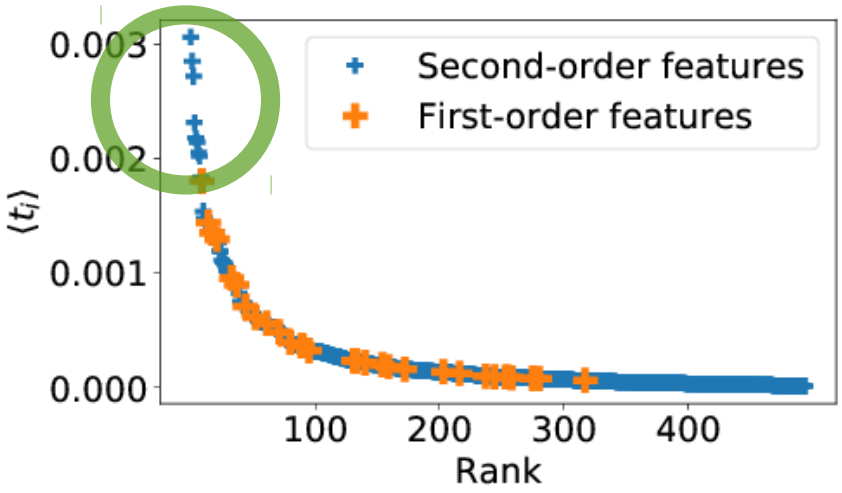
Second:
Learned separation by **correlation**

Application on a physics dataset

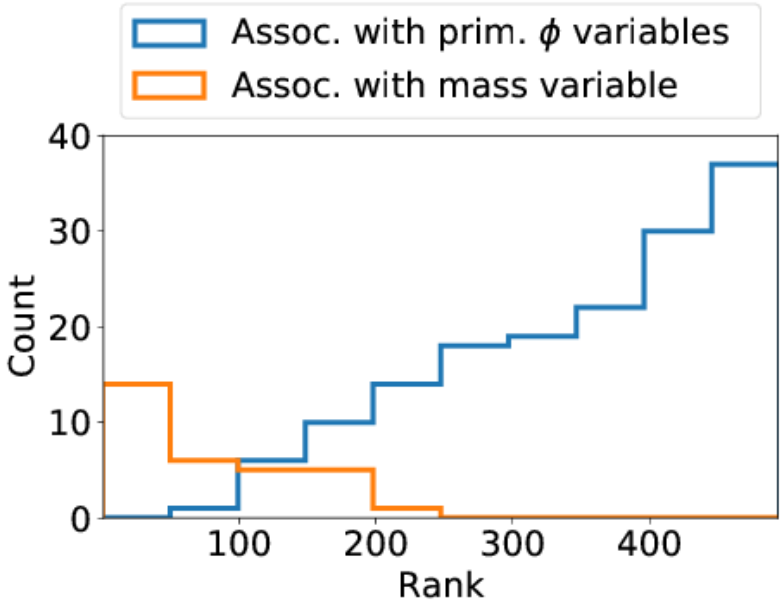
- Use **dataset from Higgs boson machine learning challenge** launched by the ATLAS collaboration in 2014
- Simulated **$H \rightarrow \tau\tau$ events** and events from background processes with similar topologies
- **Binary classification** task
- Dataset consists of **30 variables** (21 low-level and 9 high-level variables)
- **Calculate metric of relevance** for all features up to 2nd order

Application on a physics dataset

30 input variables result in 495 features:
- 30 marginal distributions
- 465 pairs of variables

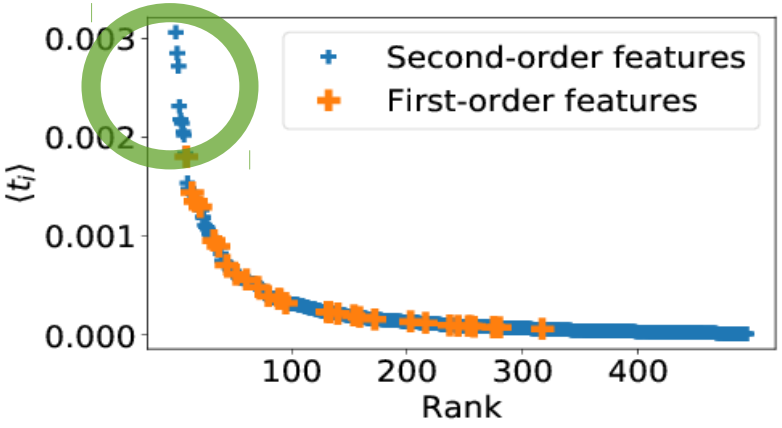


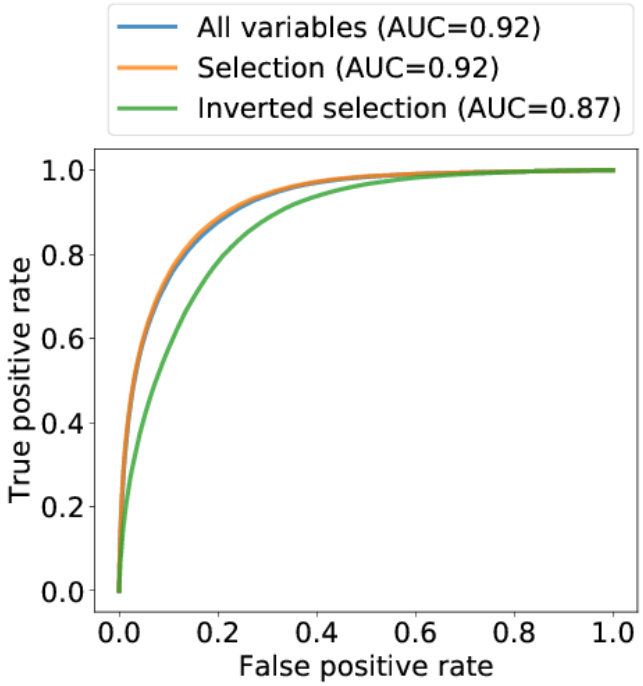
Only a **few features** are identified as **influential**.
→ **This knowledge simplifies the estimation of systematics greatly.**



Mass variables are identified as **highly important** while **Φ variables** are rated as **less significant**.
→ Matches the expectation from physics.

Application on a physics dataset



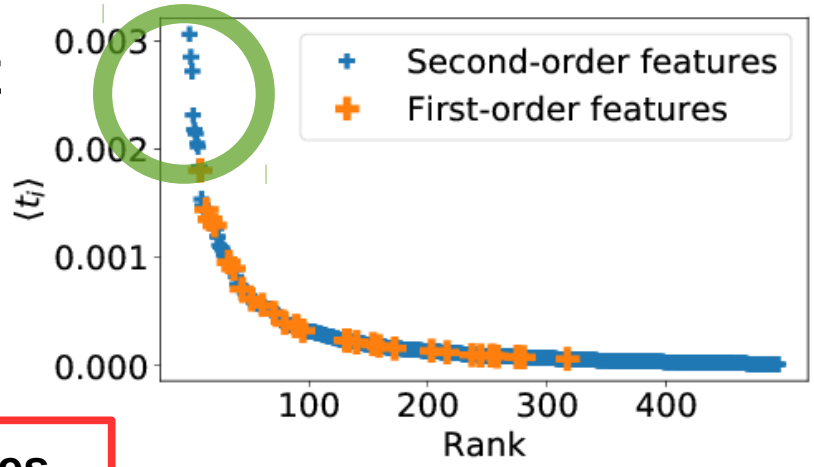


Comparison of performance for:

- Training on **30 variables** of **full dataset**
- Training on **9 variables** contributing to upper **5% of most important features**
- Training on **21 variables** of **inverted selection**

Immense reduction of dimensionality without loss of performance.

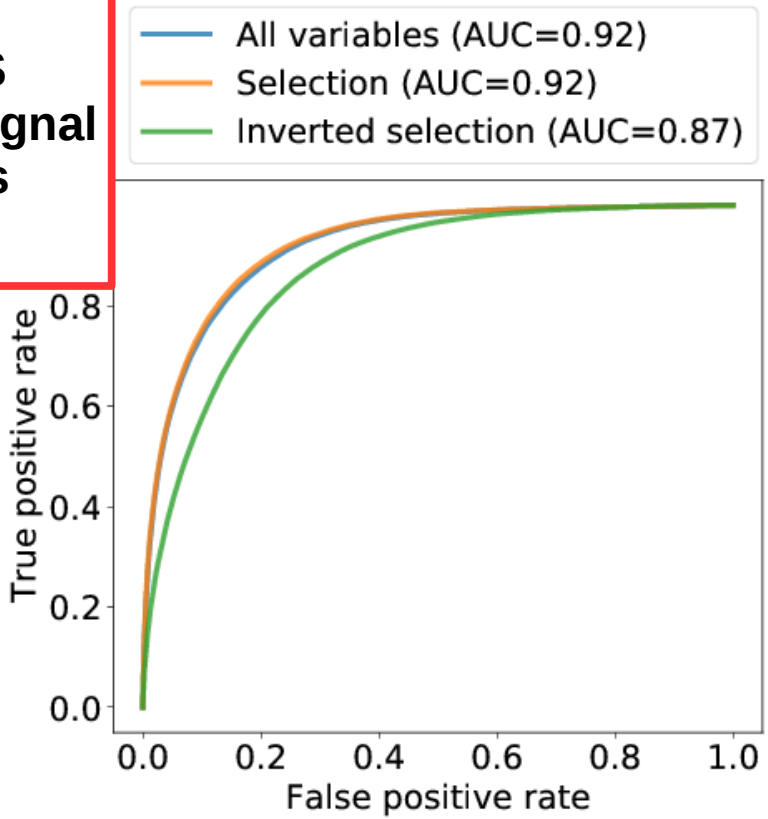
Application on a physics dataset



9 variables contributing to upper **5% of most important features**:

- **DER_mass_MMC**
- **DER_mass_vis**
- **DER_mass_jet_jet**
- DER_deltar_tau_lep
- DER_pt_ratio_lep_tau
- DER_mass_transverse_met_lep
- PRI_lep_pt
- PRI_tau_pt
- PRI_jet_all_pt

Identified variables used in physics analyses by CMS and ATLAS for signal discrimination as most important.



Summary

- Proposed usage of **Taylor expansion of neural network function** to identify relevant **dependencies of the neural network response on characteristics of the inputs**.

- Toy studies presented the application of the approach in well-defined scenarios.

- Application of the approach on a physics dataset shows the **usability in physics analyses** supporting the **in-depth understanding of the trained neural network** to facilitate the **estimation of systematic uncertainties**.

- Paper with all details is already submitted and available as pre-print on arXiv **here**.