Physics Institute III A

RWTH AACHEN UNIVERSITY

Research Training Group Physics of the Heaviest Particles at the LHC

RWTH AACHEN UNIVERSITY

KAN we improve on HEP classification tasks?

Kolmogorov-Arnold Networks applied to an LHC physics example

arXiv: 2408.02743

Johannes Erdmann, Florian Mausolf, Jan Lukas Späh

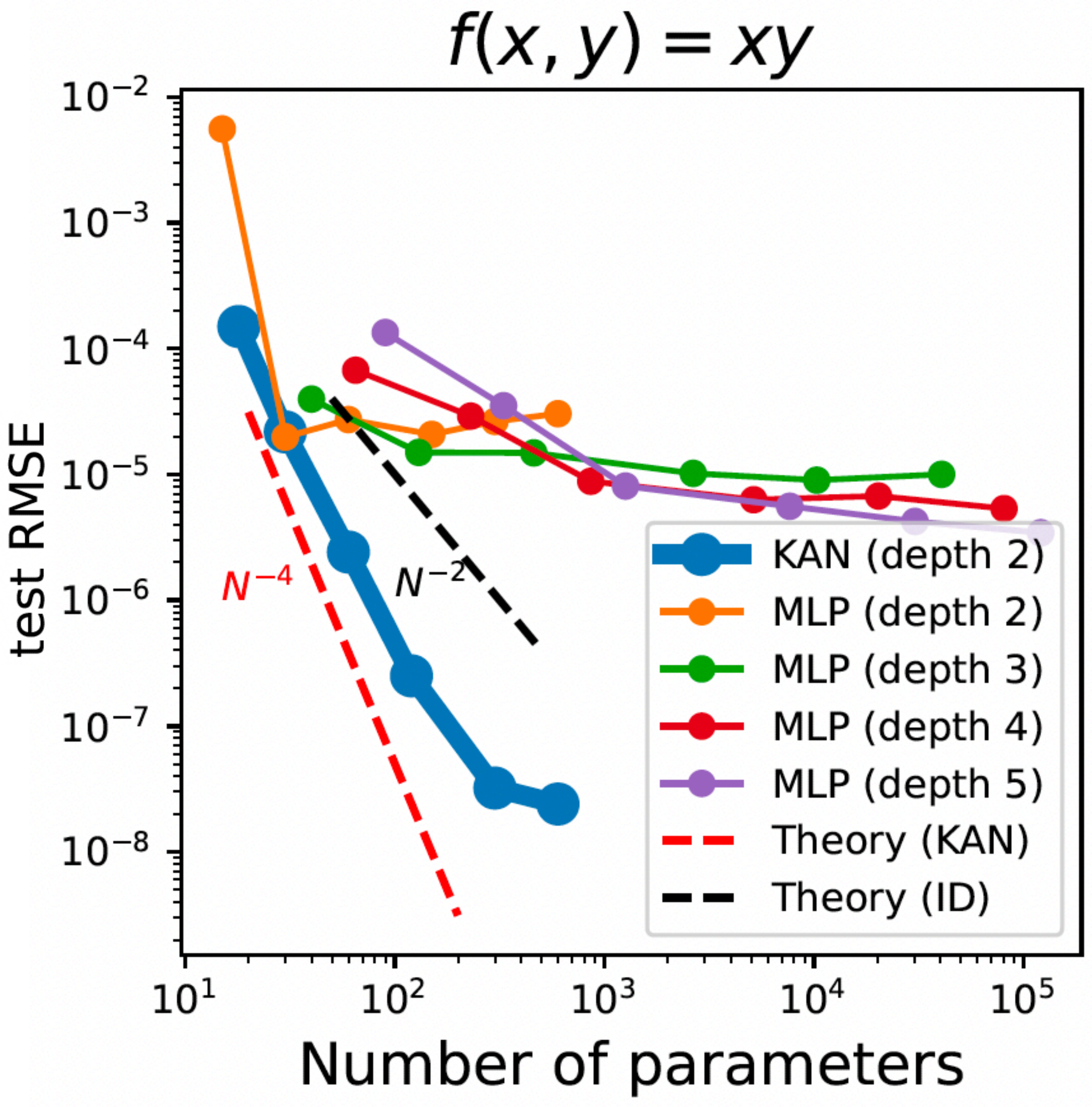
III. Physikalisches Institut A, RWTH Aachen University

ML4Jets 2024

Uncertainties & Interpretability Session

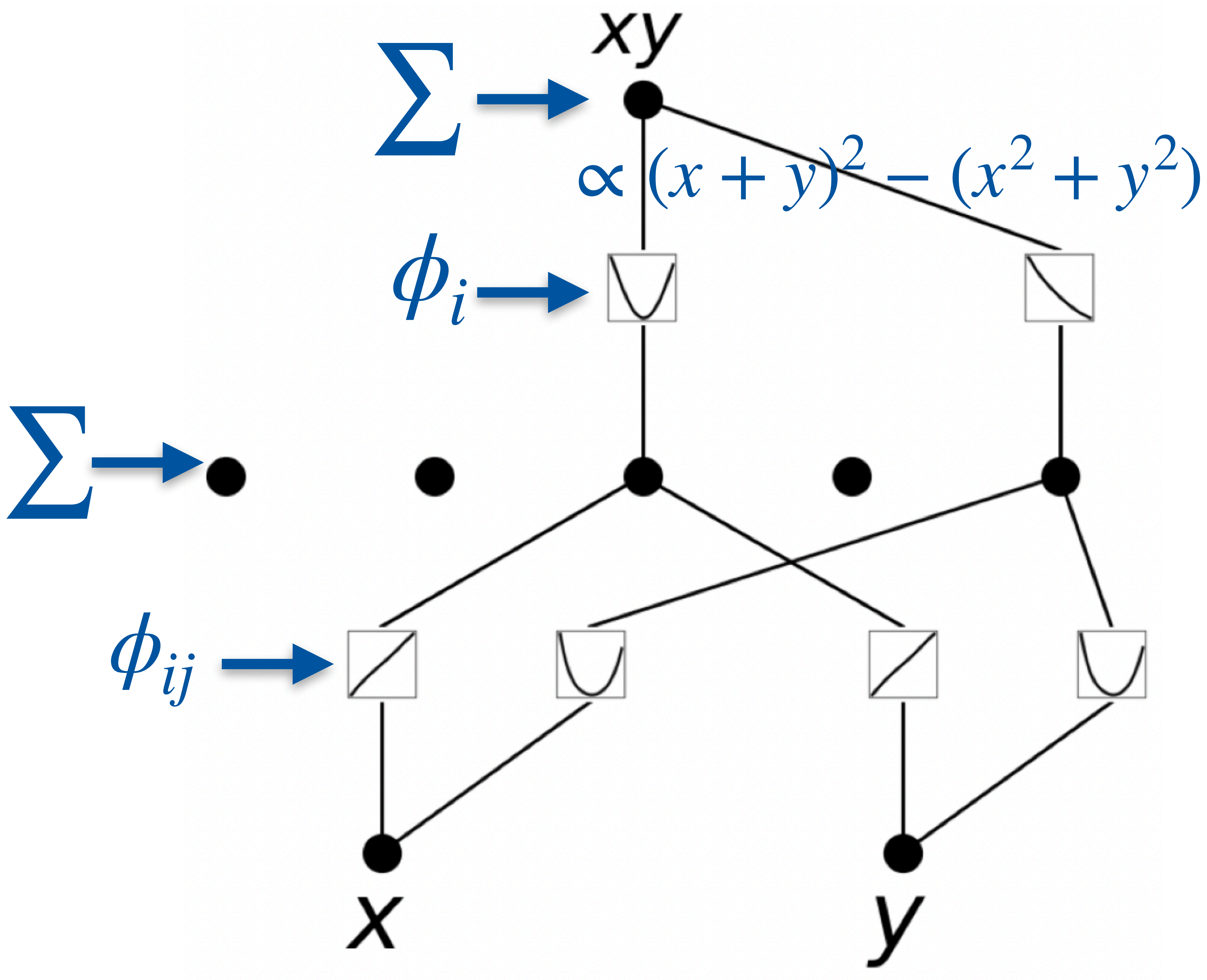
Paris, 7th November 2024

- Kolmogorov-Arnold Networks (KANs) proposed as alternative network architecture
  - Z. Liu et al., 2404.19756
- Advantages over multi-layer perceptrons presented
  - Performance, parameter efficiency and interpretability in multiple tasks
  - Examples provided are rather low dimensional, mathematical datasets
- Efficiency, performance and interpretability are crucial properties in HEP!
  - ➡ Time to explore the potential of KANs here!



Z. Liu et al., 2404.19756

- Inspiration: Kolmogorov-Arnold representation theorem: $f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{2n+1} \phi_i \left(\sum_{j=1}^{n} \phi_{ij}(x_j) \right)$
  - Continuous multivariate functions can be represented as sum of continuous univariate functions
- Motivates network architecture with learnable univariate functions and sum operation on nodes
- Stacking of "KAN layers" with arbitrary number of nodes proposed in 2404.19756



Z. Liu et al., 2404.19756

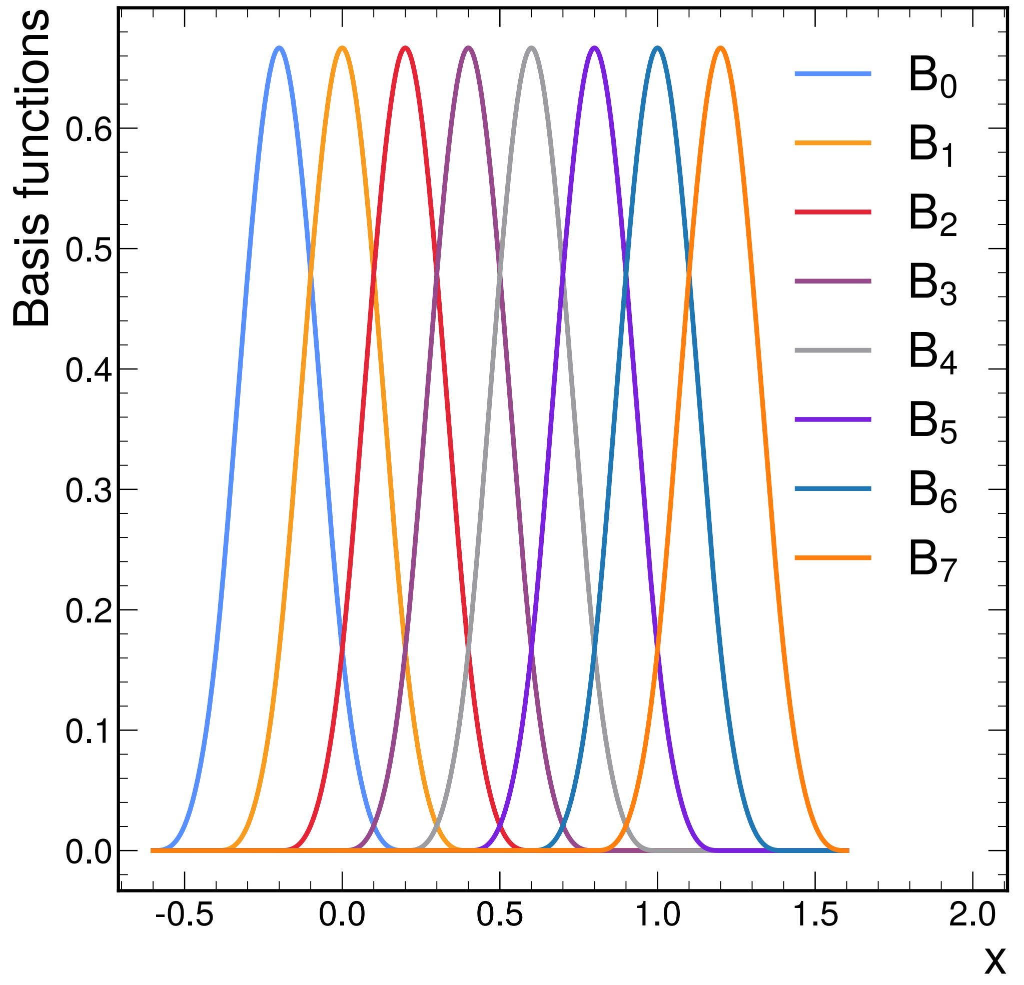
	MLPs	KANs
Edges	Linear weights	Learnable activations
Nodes	Fixed activations	Sum

Physics Institute III A RWTH AACHEN UNIVERSITY

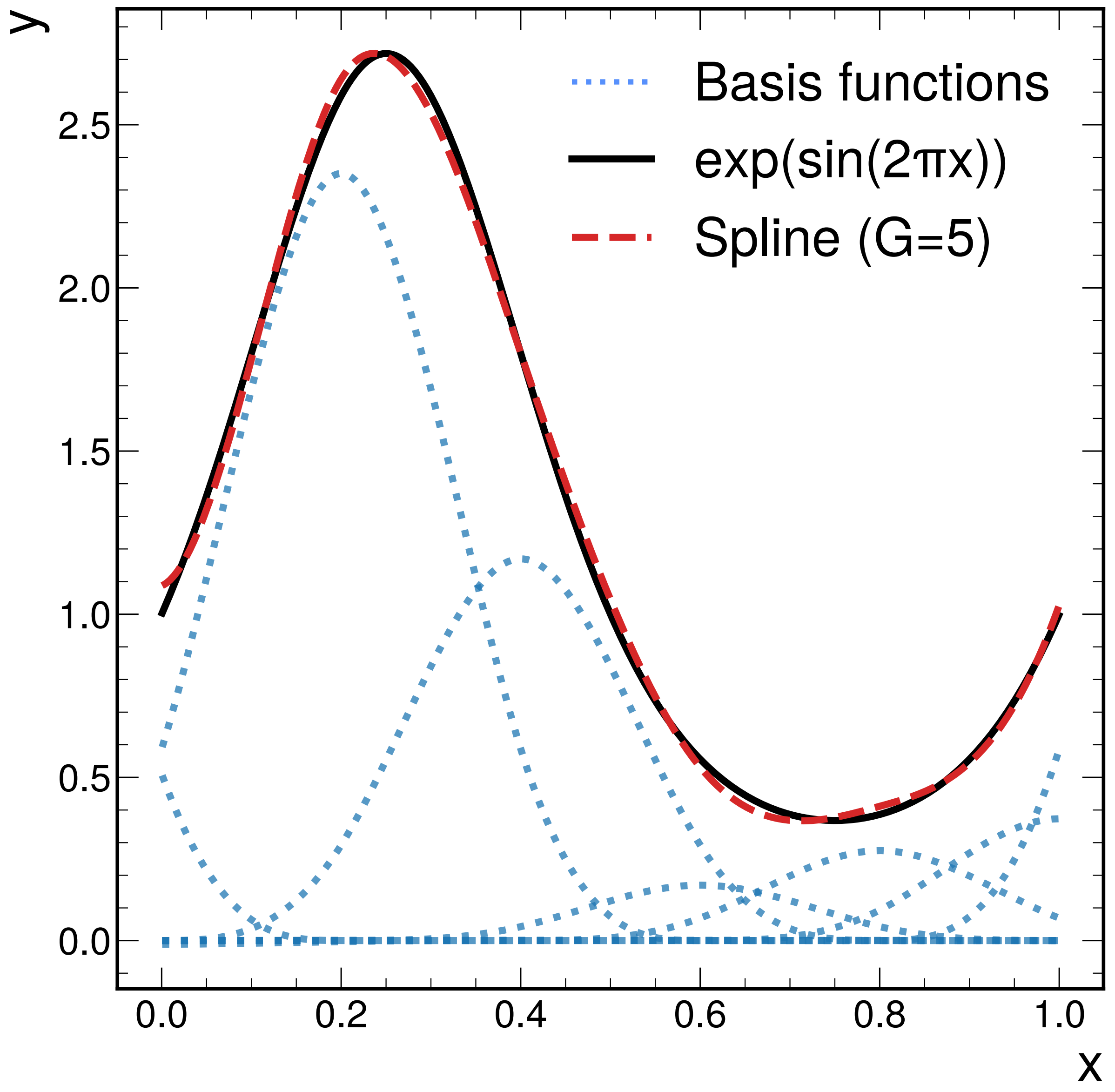
B-splines for learnable activations

- Learnable activation functions can be defined with B-splines
- $\text{activation}(x) = w_1 \cdot \text{SiLU}(x) + w_2 \cdot \sum_{i=0}^{G+k-1} c_i \cdot B_i(x)$
  - w_i, c_i: trainable parameters
  - $B_i(x)$: B-spline basis functions of degree k

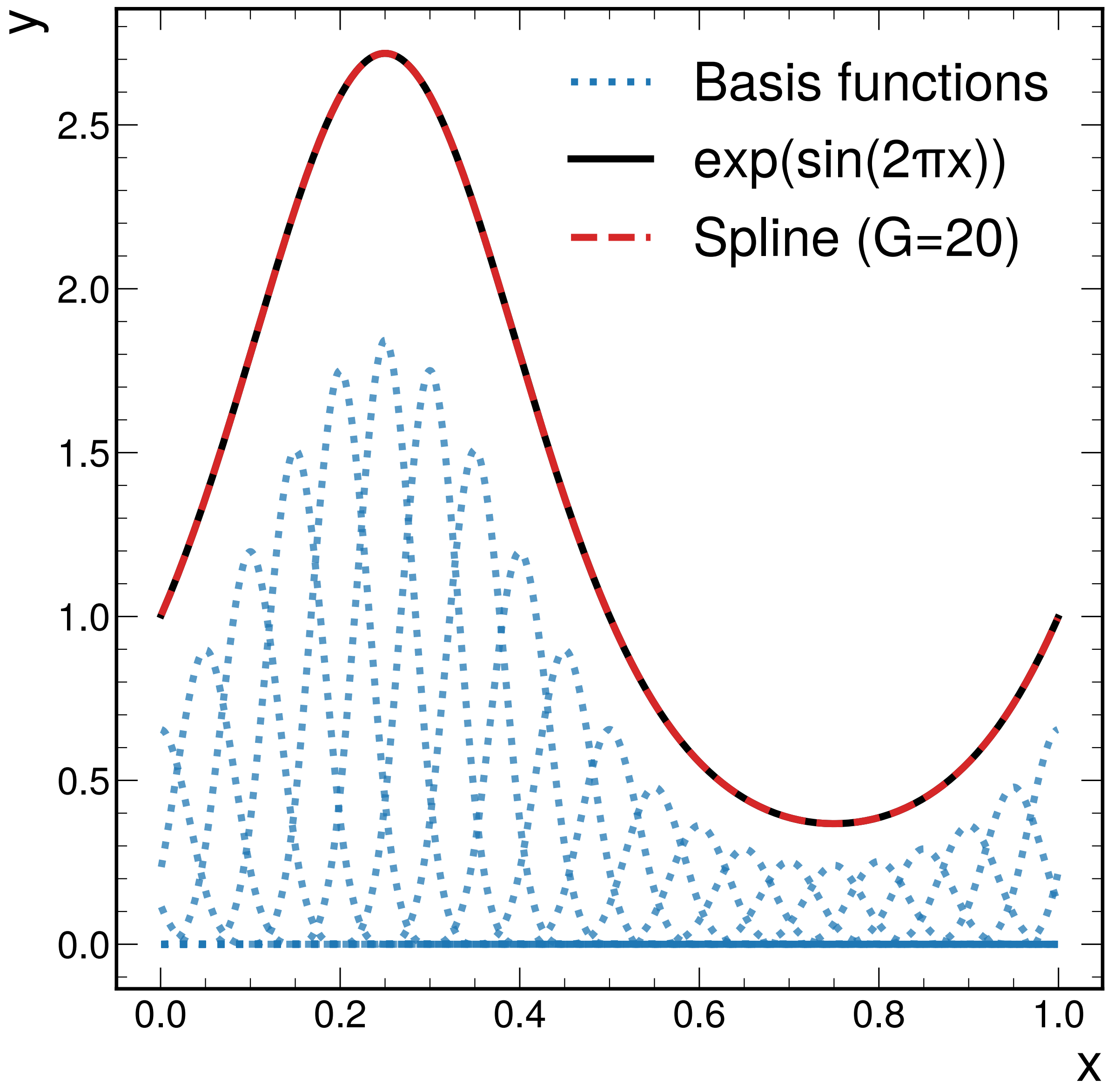
$G = 5,\ k = 3,$
grid range $= (0,1)$



Adapt coefficients c_i
to fit function



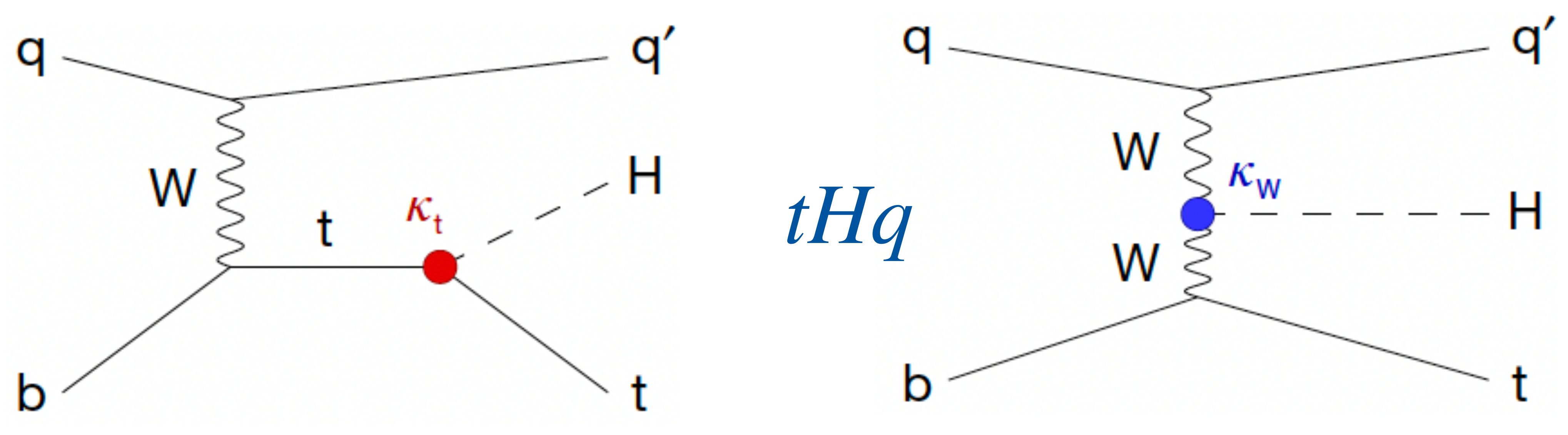
Increase G for
more flexibility



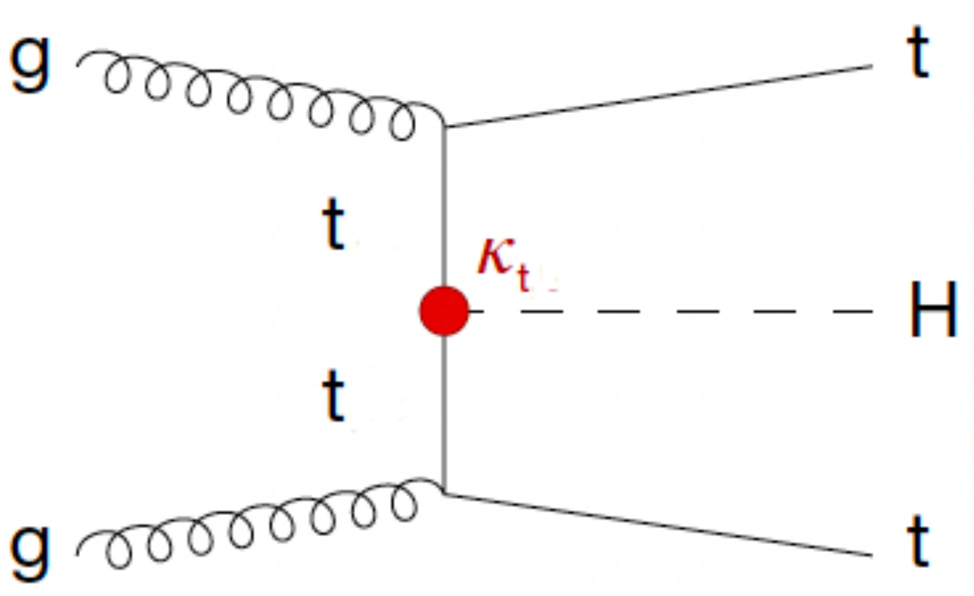
- Grid parameter G regulates number of basis functions, $G + k$ functions used

Dataset

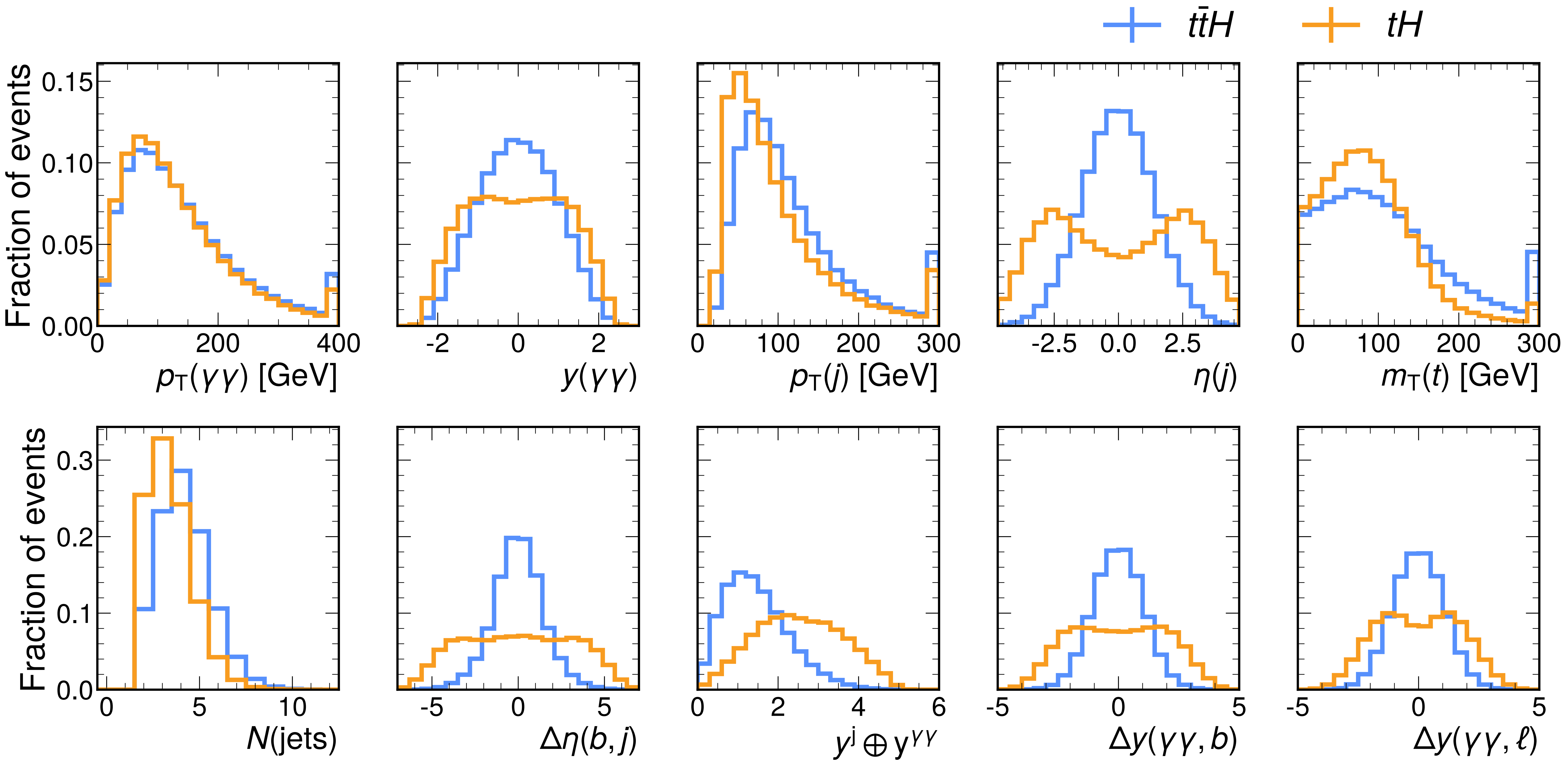
- $t\bar{t}H$ vs. tHq classification in $H \to \gamma\gamma$ decay channel
- MadGraph (LO) + Pythia + Delphes with CMS card
- Typical $H \to \gamma\gamma$ event selection for leptonic channel
- 22 kinematic features constructed



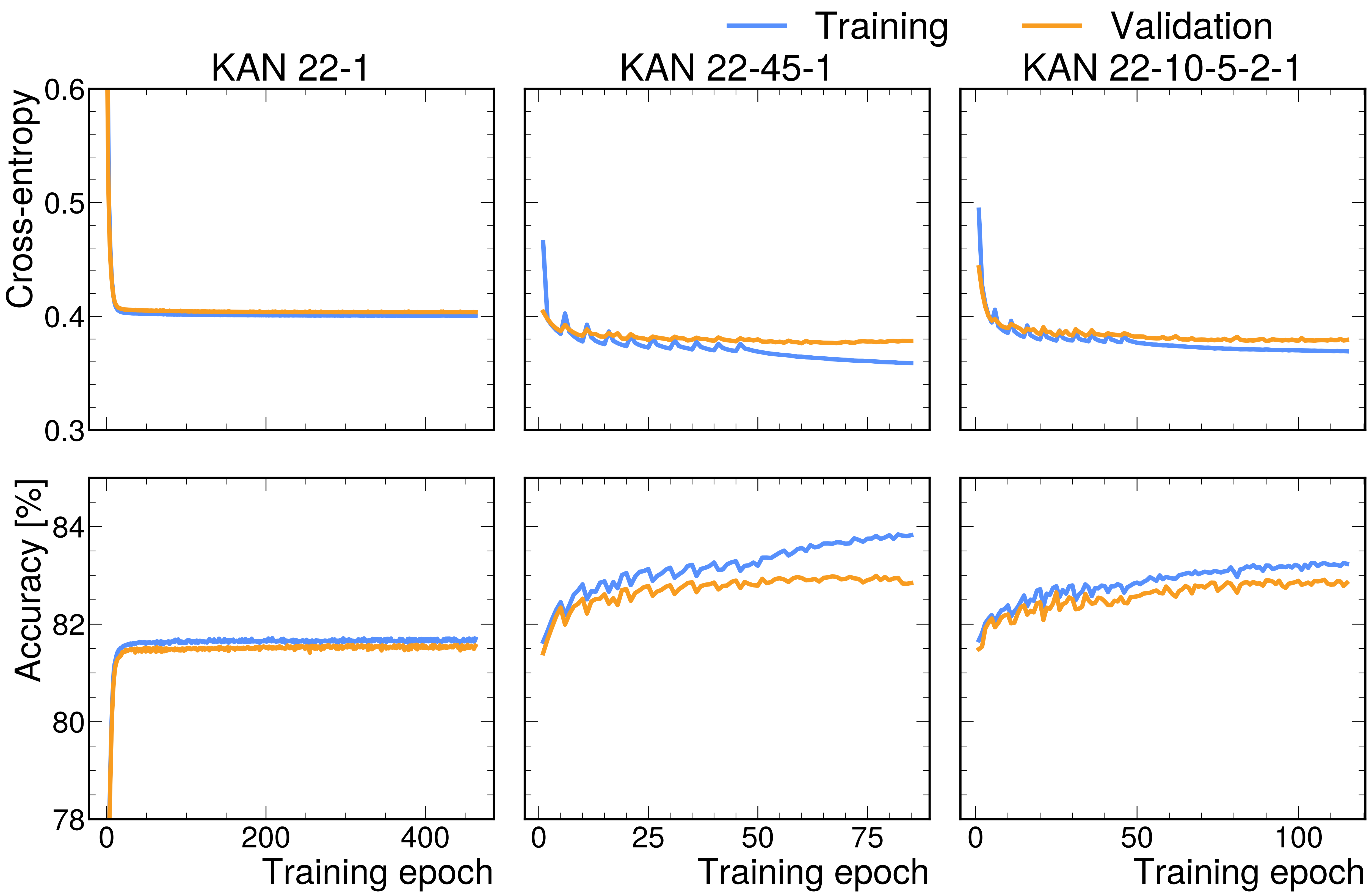
tHq



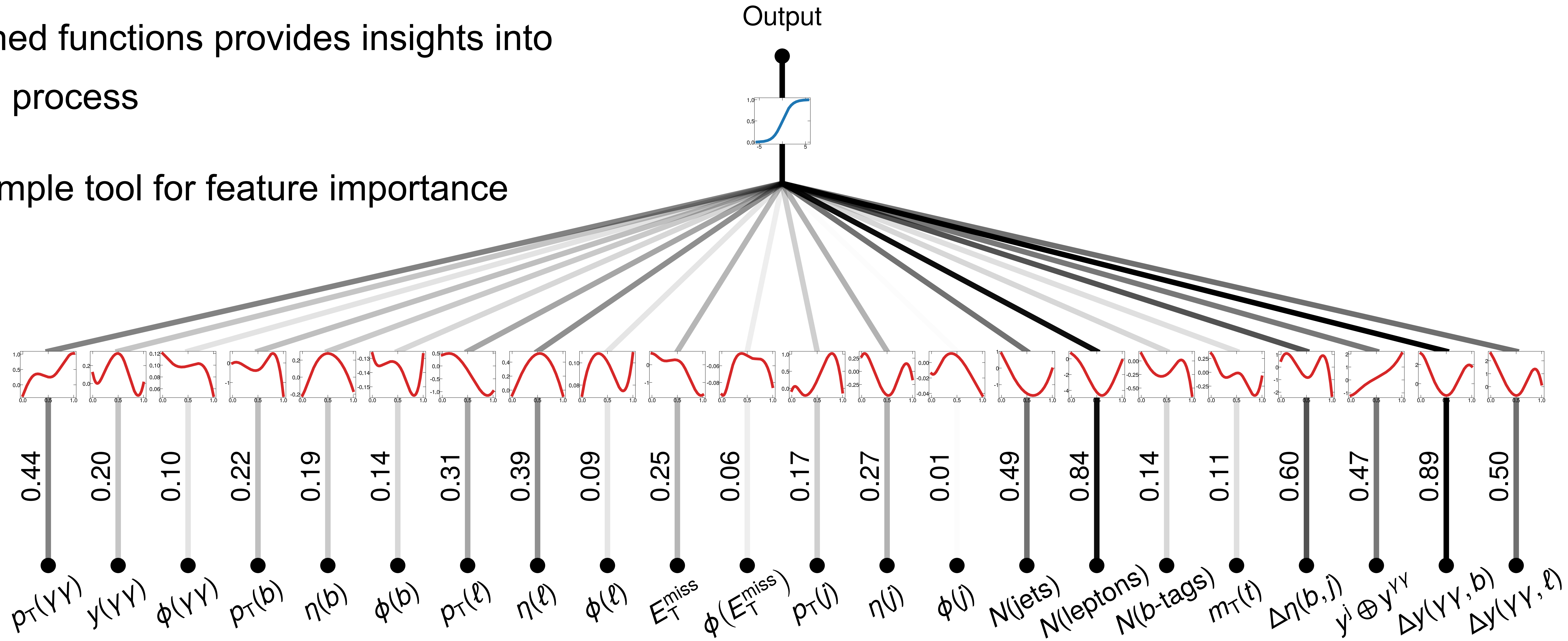
$t\bar{t}H$



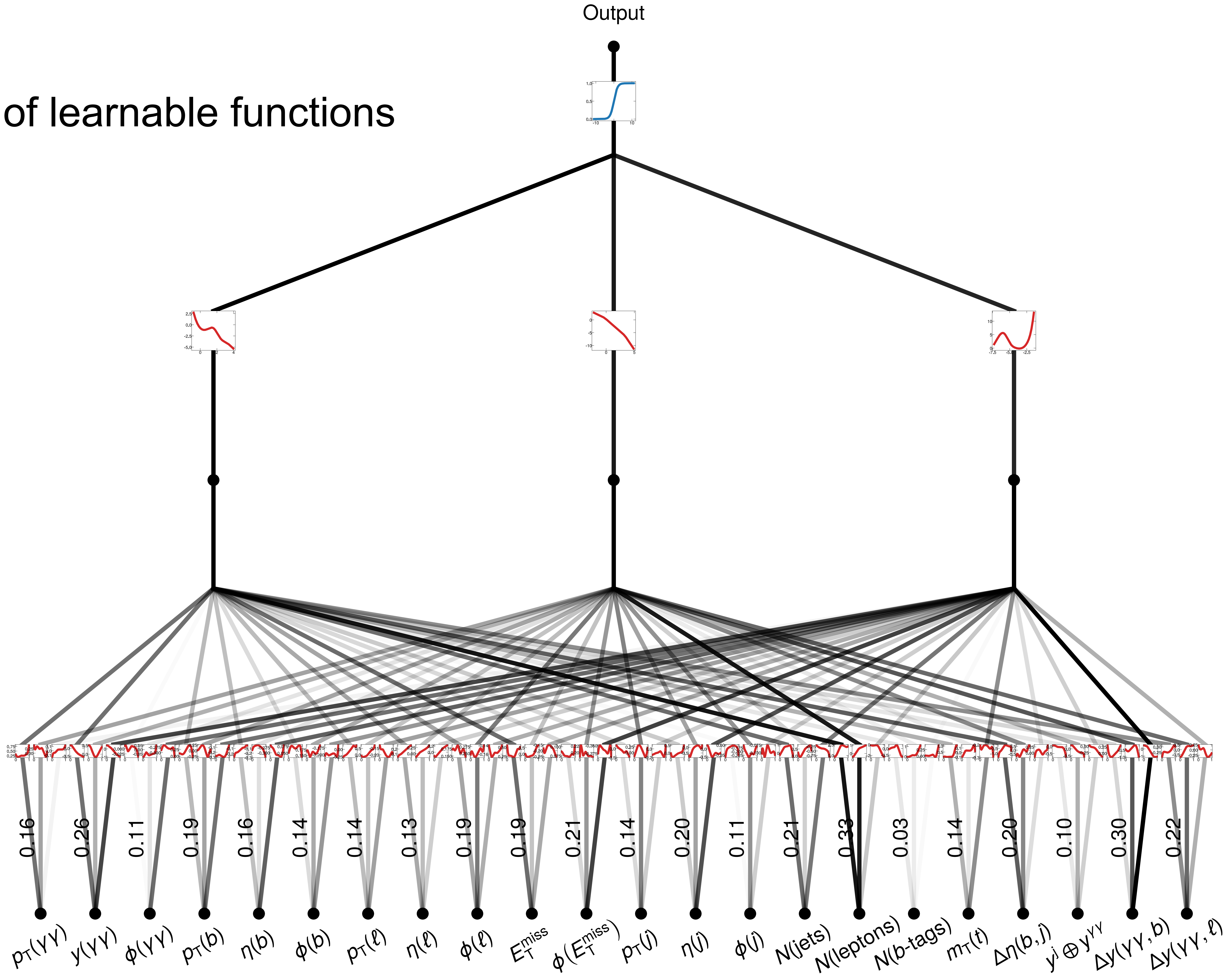
- Using Pykan package for training
- Basic setup for classification task
  - Sigmoid activation of output
  - Cross-entropy loss
  - Adam optimizer
- Stable trainings without much fine-tuning found
- Even single-layer network reaches >81.5% accuracy, two layers needed for better performance (~83%)



- Small models: moderate number of spline functions
  - Here: single-layer KAN (22–1)
- Analysing learned functions provides insights into model decision process
- L_1-norms as simple tool for feature importance

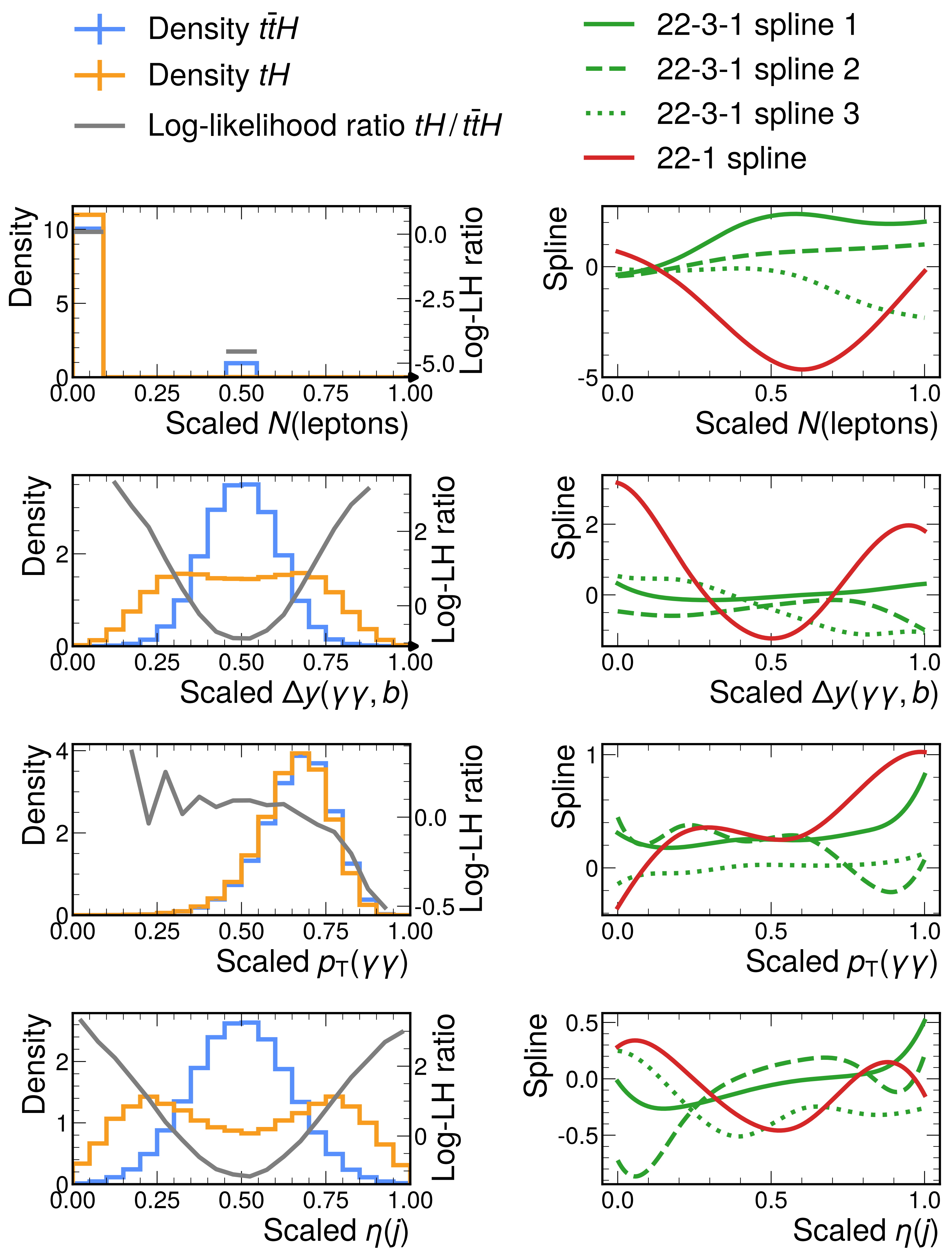


- Deeper / wider models can contain large number of learnable functions
  - E.g. 22–45–1 KAN: > 1000 splines
  - Here: 22–3–1 KAN: 69 splines
- Understanding the reasoning of larger models becomes very difficult

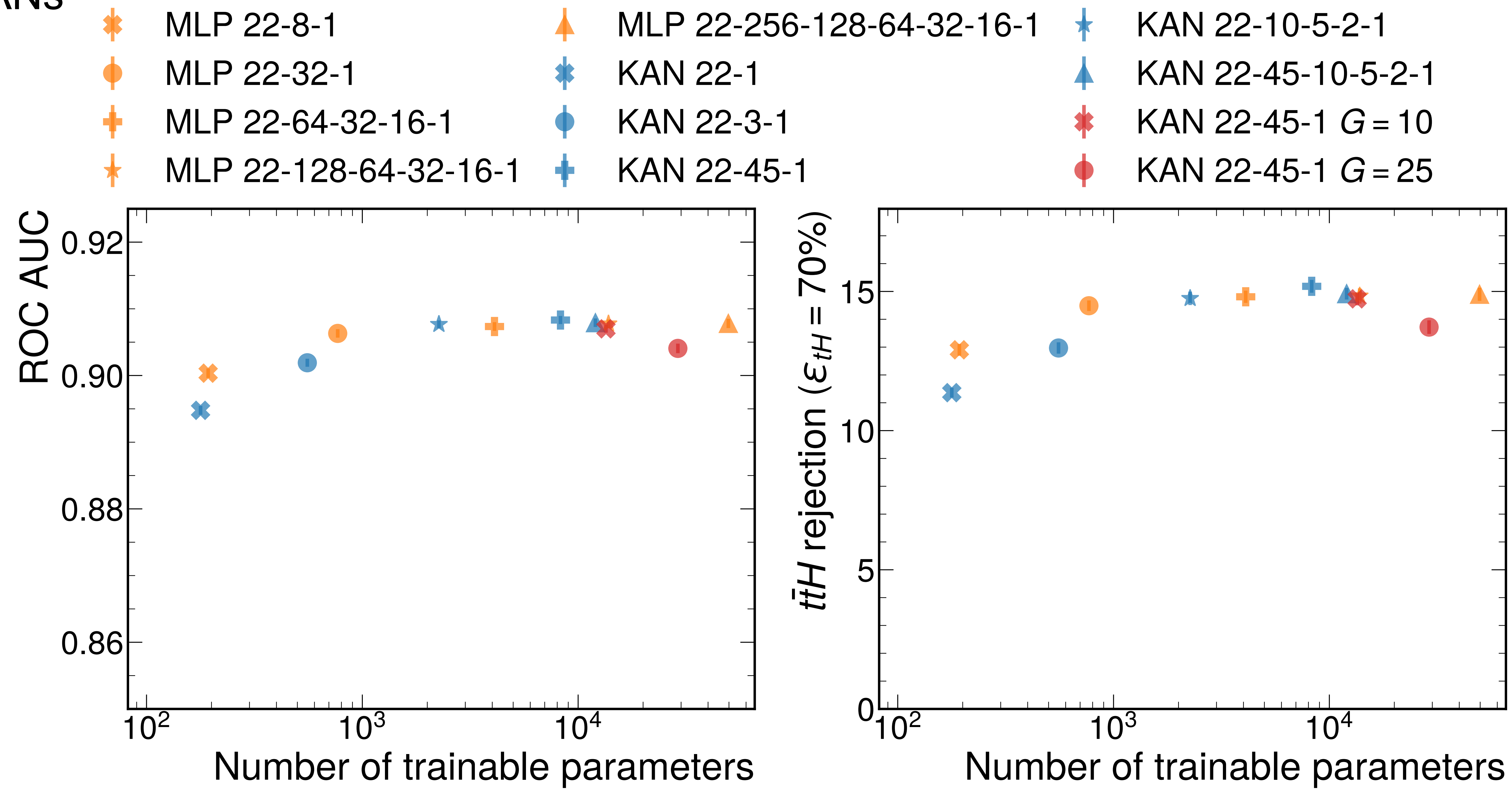


Interpretability: splines

- Patterns can be observed in small models
- Single-layer KAN:
  - One function transforms each input feature
  - Splines often resemble log-likelihood ratio of input features
- Deeper and wider KANs:
  - Many functions act on one input feature
  - More complex representations of input data

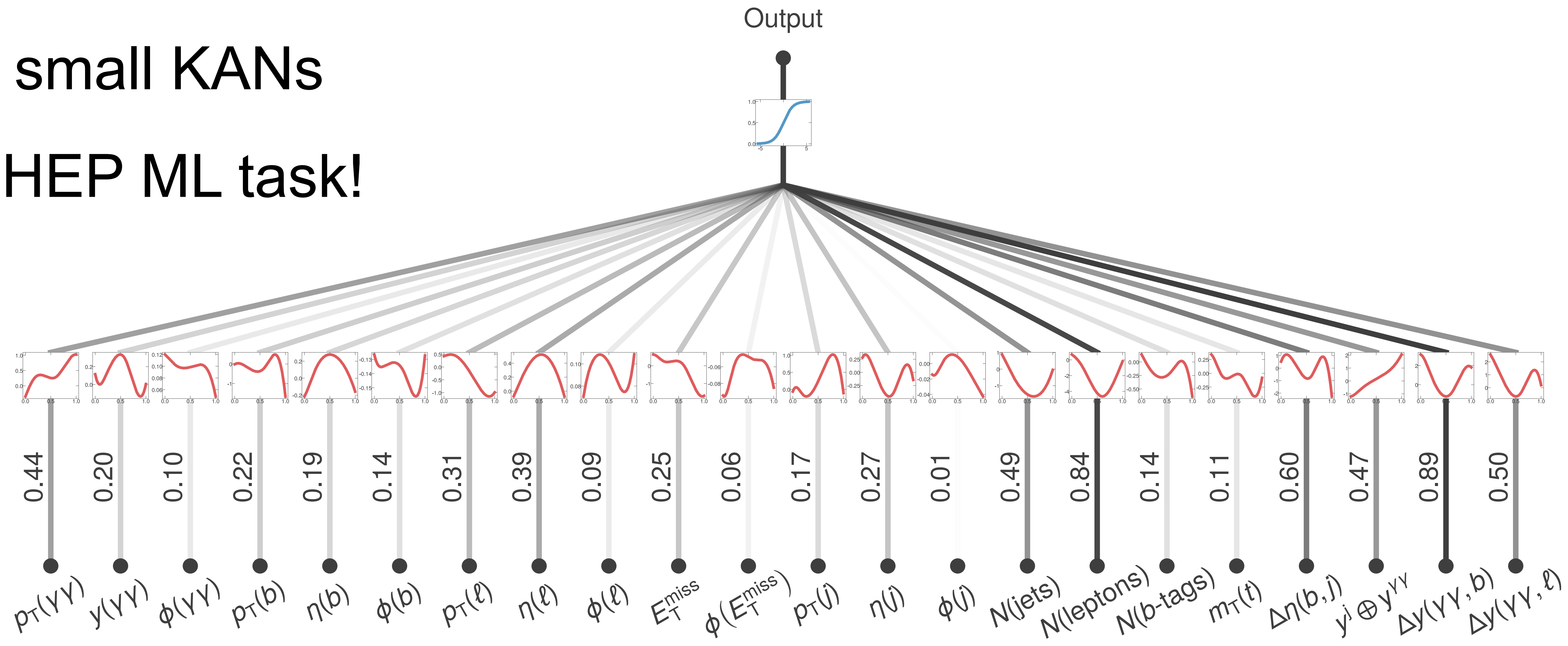


- Studied MLPs and KANs of different architectures for this task
- <1000 trainable parameters: MLPs beat KANs
    - Interpretability of small KANs comes at a cost in performance
- Similar performance for models with more parameters
    - Best tested model was a KAN!



Conclusions

- Applied Kolmogorov-Arnold Networks to high-energy physics for the first time
- Classification of $t\bar{t}H$ & tH events in $H \rightarrow \gamma\gamma$ decay channel
  - Example of typical complexity for binary HEP event classification
- Observed that KANs can achieve similar performance as MLPs, but don't appear more parameter efficient on our dataset
- Advantages in interpretability over MLPs exist for small KANs
  - May be worth to consider KANs for your next HEP ML task!
- Pre-print about our study: [2408.02743]

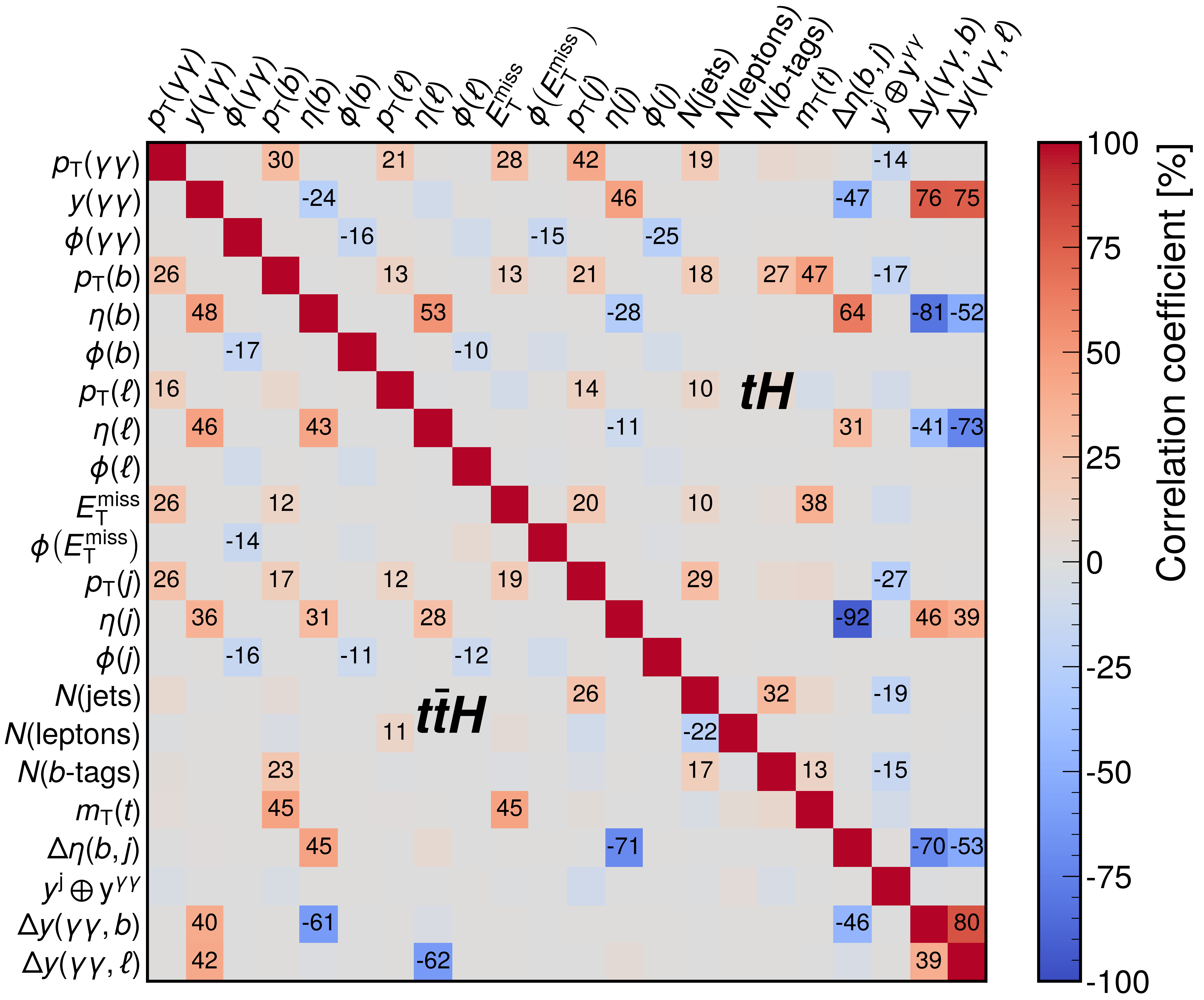


Backup

Physics Institute III A
RWTH AACHEN UNIVERSITY

Correlations of features

- Non-trivial correlation pattern
- Significant differences between the two classes



Physics Institute III A

RWTH AACHEN UNIVERSITY

B-spline basis functions

- $\text{activation}(x) = w_1 \cdot \text{SiLU}(x) + w_2 \cdot \sum_{i=0}^{G+k-1} c_i \cdot B_i(x)$
  - w_i, c_i: trainable parameters
  - $B_i(x)$: B-spline basis functions of degree k

➡ Recursive definition with Cox–de Boor formula

- $B_i^0(x) = 1$ if $t_i \leq x < t_{i+1}$; $B_i^0(x) = 0$ otherwise
- For $k > 0$: $B_i^k(x) = \frac{x - t_i}{t_{i+k} - t_i} B_i^{k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1}^{k-1}(x)$.