Normalizing Flows for Physics Data Analyses

Jan Gavranovič (jan.gavranovic@ijs.si) [1,2]   Borut Paul Kerševan (borut.kersevan@ijs.si) [1,2]

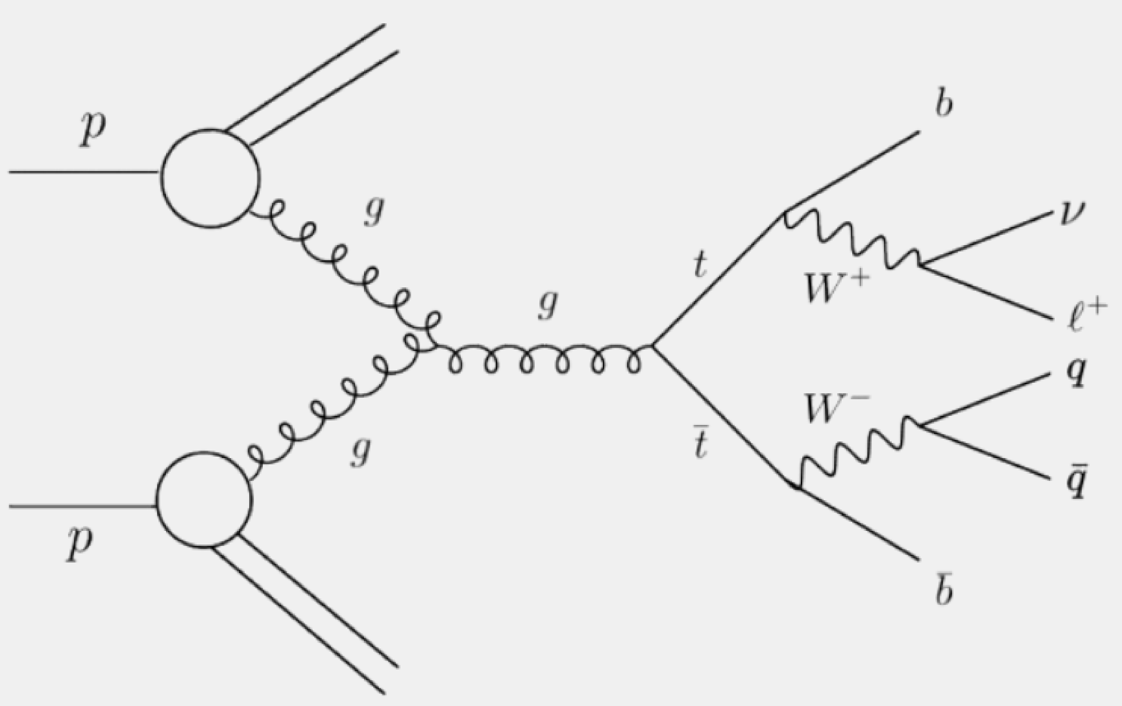
[1]Jožef Stefan Institute, Ljubljana, Slovenia   [2]Faculty of Mathematics and Physics, University of Ljubljana, Slovenia

Introduction

- LHC produces **big data** $\Rightarrow$ MC and analysis need to follow
- Can generative models be used to support physics modeling?
- Problem: do not know the true generating data distribution
- Objective: approximate $p_{\text{data}}(\boldsymbol{x})$ to enable **infinite sampling**
- *Learn* true $p_{\text{data}}(\boldsymbol{x})$ from $\boldsymbol{x} \in \mathbb{R}^D$ using approximate $p_{\text{model},\boldsymbol{\theta}}(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$
- Focus on LHC *analysis-specific* distributions of *final* analysis variables

Higgs Benchmark Dataset

- **Publicly available dataset** with 11M events and 28 variables
- Binary classification problem: signal (BSM) vs. background ($t\bar{t}$)
- Use as test for LHC *final* event simulation with normalizing flows



- 21 low-level and 7 high-level variables
- Data preprocessing (*feature scaling*) is a crucial step in training
- **Task: train ML model to generate *new background events***

Normalizing Flows (Invertible Neural Networks)

- Two pieces:
  1. base distribution $p_u(\boldsymbol{u})$, typically $\mathcal{N}(\boldsymbol{u}|0, \mathbf{I})$
  2. differentiable transformation $\boldsymbol{x} = T(\boldsymbol{u})$ with an inverse $\boldsymbol{u} = T^{-1}(\boldsymbol{x})$
- Construct **a flow** by composing together **many transformations** T:

$$T = T_K \circ \ldots \circ T_1 \quad \text{and} \quad T^{-1} = T_1^{-1} \circ \ldots \circ T_K^{-1}$$

- Transformations T are (invertible) neural networks with parameters ϕ
- Generative process:

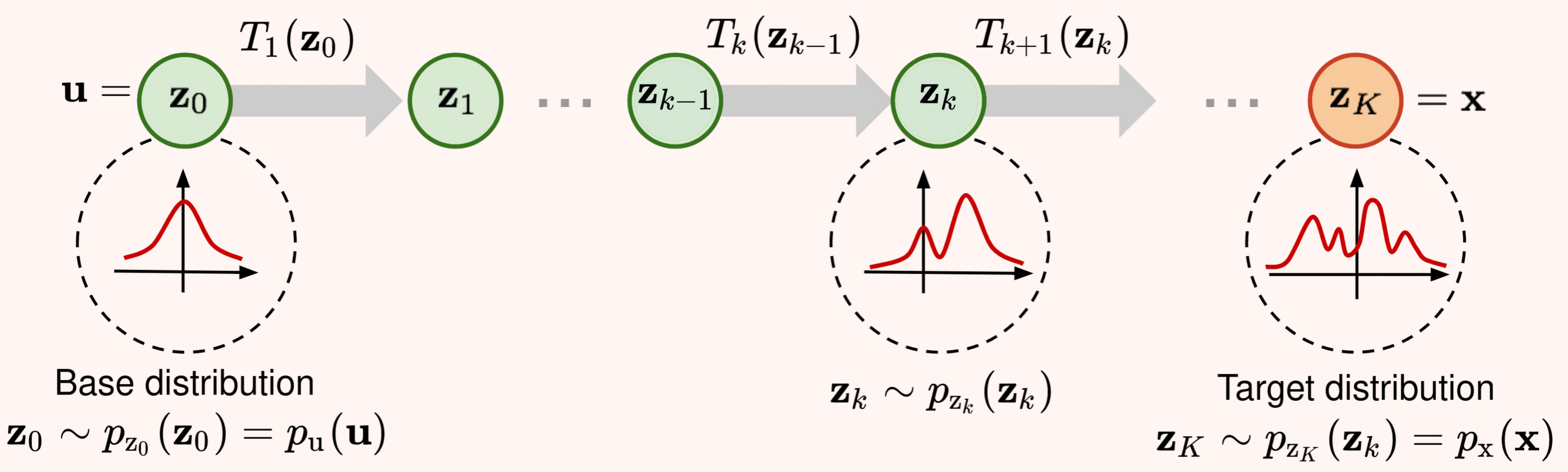
$$\boldsymbol{x} = T(\boldsymbol{u}) \approx p_x(\boldsymbol{x}) \quad \text{with sampling} \quad \boldsymbol{u} \sim p_u(\boldsymbol{u})$$

- Density evaluation using change of variables formula:

$$p_x(\boldsymbol{x}) = p_u(T^{-1}(\boldsymbol{x})) \left| \det \frac{\partial T^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right|$$

Forward and Inverse Directions

- **Forward direction:** $\boldsymbol{z}_k = T_k(\boldsymbol{z}_{k-1})$ for $k = 1, \ldots, K$ with $\boldsymbol{z}_0 = \boldsymbol{u}$ (infer)
- **Inverse direction:** $\boldsymbol{z}_{k-1} = T_k^{-1}(\boldsymbol{z}_k)$ for $k = K, \ldots, 1$ with $\boldsymbol{z}_K = \boldsymbol{x}$ (train)



- Similar to autoencoder: forward mode $\Leftrightarrow$ decoder, backward mode $\Leftrightarrow$ encoder
- Loss function has two terms (**log-likelihood + log-determinant**):

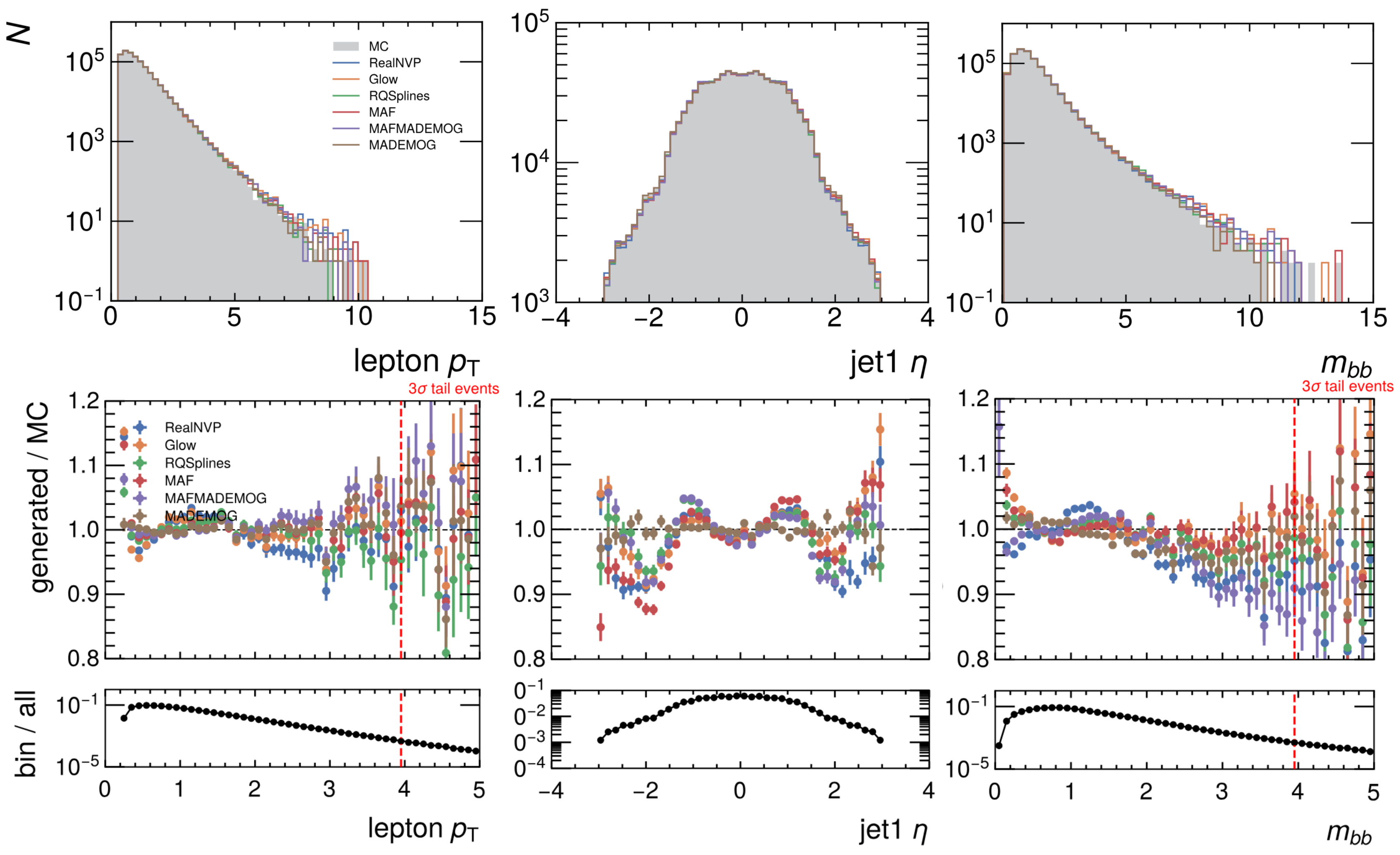
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \left[\log p_u \left(T^{-1}(\boldsymbol{x}_n; \boldsymbol{\phi}); \boldsymbol{\psi} \right) + \log \left| \det J_{T^{-1}}(\boldsymbol{x}_n; \boldsymbol{\phi}) \right| \right]$$

- Use *gradient descent* to get the best parameters:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \mathcal{L}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \equiv \{\boldsymbol{\phi}, \boldsymbol{\psi}\}$$

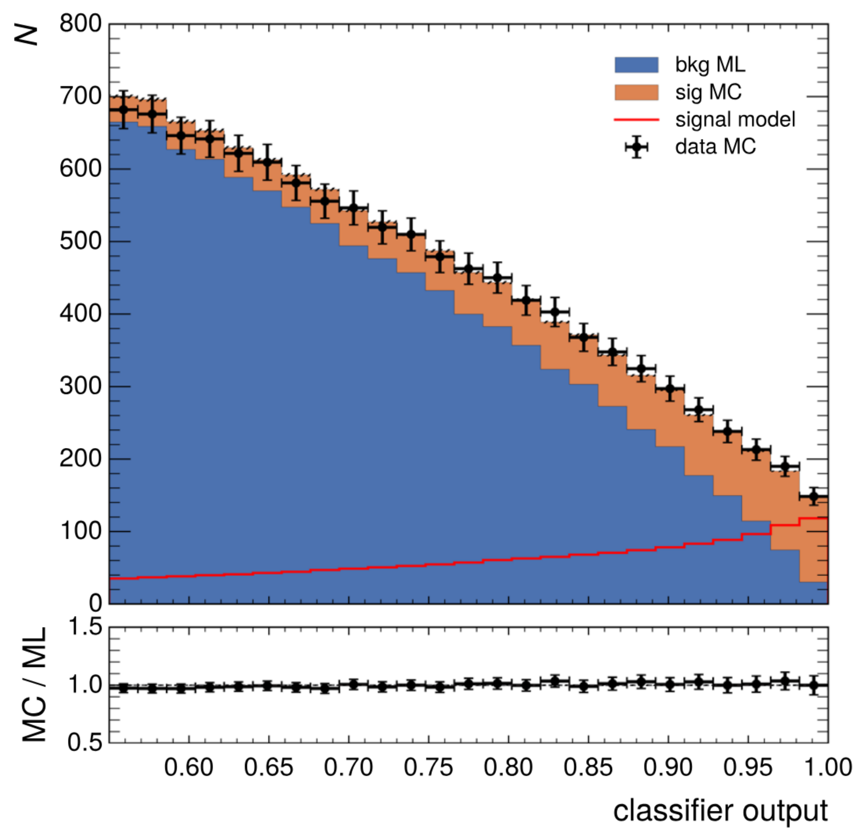
Performance Evaluation

- Comparison of ML *generated* and MC *simulated* distributions
- *Best* model was selected for the final analysis
- Performance was measured using **statistical distances** and **classifier two sample testing**

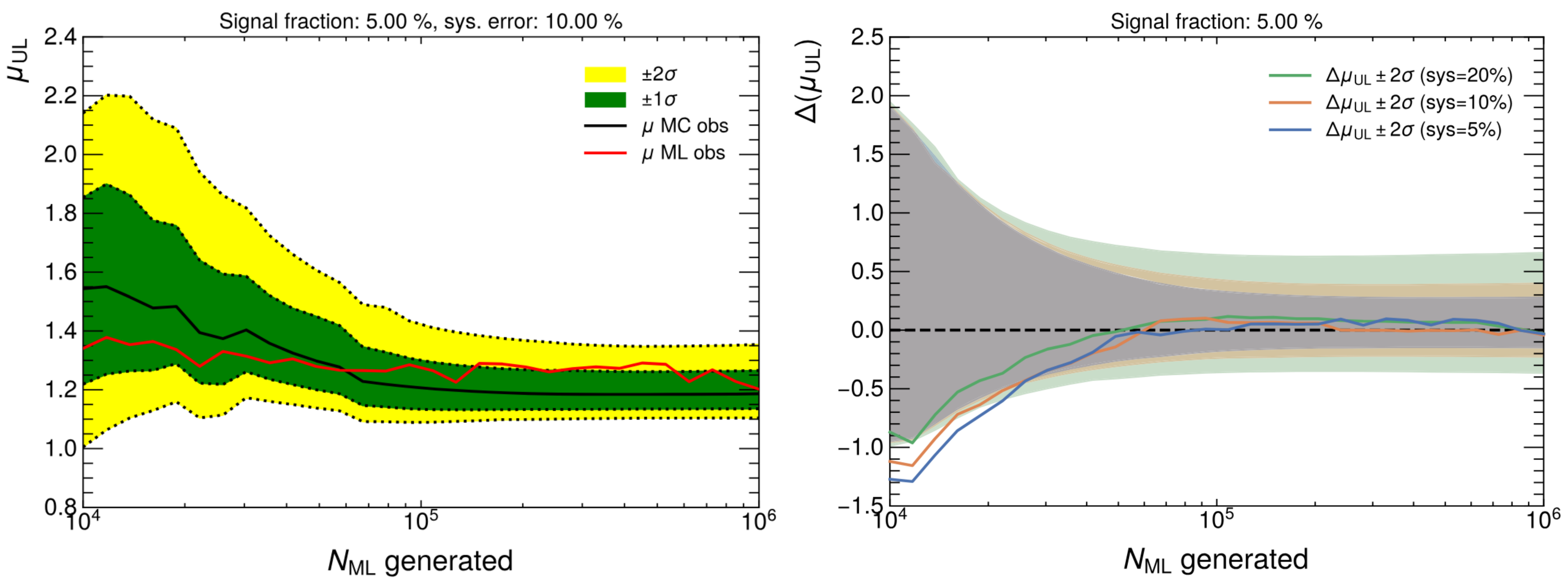


Physics Analysis

- A simplified analysis was performed, involving preselection (baseline cuts) and a **NN-based classifer final selection**



- **Upper limits on the signal strength** μ for the likelihood fit to the classifier score distribution as a function of ML-generated events were calculated



Summary

- Generative modeling is a **promising new tool** for physics data analyses
- Needs careful performance evaluation and validation for physics use cases

References

[1] J. Gavranovič and B. P. Kerševan.
Systematic evaluation of generative machine learning capability to simulate distributions of observables at the large hadron collider.
Eur. Phys. J. C, 84(9):911, 2024.

Jožef Stefan Institute
Ljubljana, Slovenia
FMF UNIVERSITY OF LJUBLJANA
Faculty of Mathematics and Physics
October 19 - 25, 2024
CHEP 2024
Conference on Computing in High Energy and Nuclear Physics