# THE DEVELOPMENT OF A MACHINE LEARNING-BASED HADRONIZATION MODEL

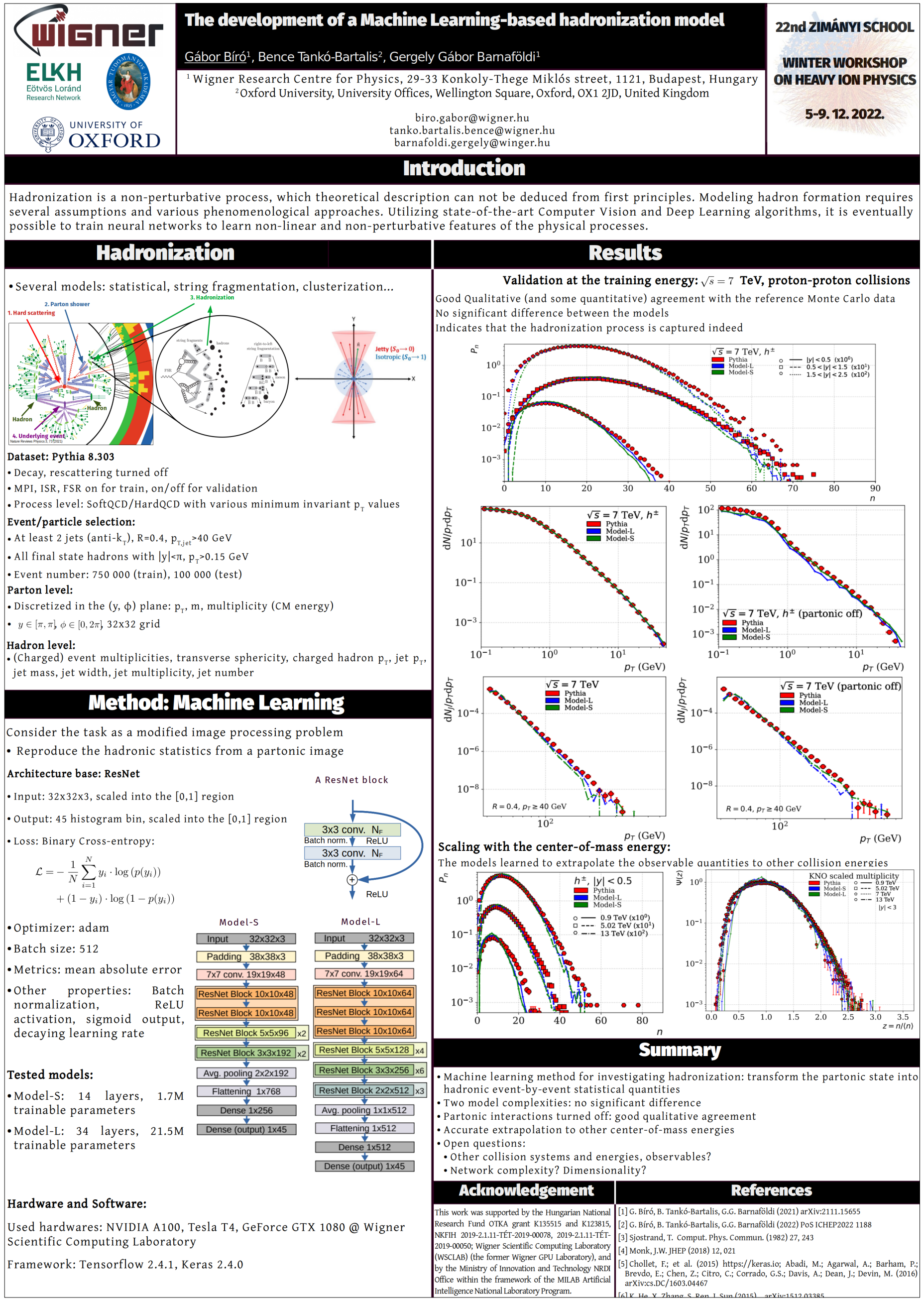
22nd ZIMÁNYI SCHOOL

WINTER WORKSHOP
ON HEAVY ION PHYSICS
5-9. 12. 2022.


**GÁBOR BÍRÓ**
biro.gabor@wigner.hu

Gergely Gábor Barnaföldi
Bence Tankó-Bartalis


arXiv:2111.15655
arXiv:2210.10548



## The development of a Machine Learning-based hadronization model


Gábor Bíró[1], Bence Tankó-Bartalis[2], Gergely Gábor Barnaföldi[1]

[1] Wigner Research Centre for Physics, 29-33 Konkoly-Thege Miklós street, 1121, Budapest, Hungary
[2] Oxford University, University Offices, Wellington Square, Oxford, OX1 2JD, United Kingdom

biro.gabor@wigner.hu
tanko.bartalis.bence@wigner.hu
barnafoldi.gergely@winger.hu


22nd ZIMÁNYI SCHOOL
WINTER WORKSHOP ON HEAVY ION PHYSICS
5-9. 12. 2022.

### Introduction

Hadronization is a non-perturbative process, which theoretical description can not be deduced from first principles. Modeling hadron formation requires several assumptions and various phenomenological approaches. Utilizing state-of-the-art Computer Vision and Deep Learning algorithms, it is eventually possible to train neural networks to learn non-linear and non-perturbative features of the physical processes.

### Hadronization

- Several models: statistical, string fragmentation, clusterization...



**Dataset: Pythia 8.303**
- Decay, rescattering turned off
- MPI, ISR, FSR on for train, on/off for validation
- Process level: SoftQCD/HardQCD with various minimum invariant $p_T$ values

**Event/particle selection:**
- At least 2 jets (anti-$k_T$), R=0.4, $p_{T,jet}$>40 GeV
- All final state hadrons with $|y|<\pi$, $p_T$>0.15 GeV
- Event number: 750 000 (train), 100 000 (test)

**Parton level:**
- Discretized in the (y, ϕ) plane: $p_T$, m, multiplicity (CM energy)
- $y \in [\pi, \pi]$, $\phi \in [0, 2\pi]$, 32x32 grid

**Hadron level:**
- (Charged) event multiplicities, transverse sphericity, charged hadron $p_T$, jet $p_T$, jet mass, jet width, jet multiplicity, jet number

### Method: Machine Learning

Consider the task as a modified image processing problem
- Reproduce the hadronic statistics from a partonic image

**Architecture base: ResNet**
- Input: 32x32x3, scaled into the [0,1] region
- Output: 45 histogram bin, scaled into the [0,1] region
- Loss: Binary Cross-entropy:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i)\cdot\log(1-p(y_i))$$

- Optimizer: adam
- Batch size: 512
- Metrics: mean absolute error
- Other properties: Batch normalization, ReLU activation, sigmoid output, decaying learning rate



**Tested models:**
- Model-S: 14 layers, 1.7M trainable parameters
- Model-L: 34 layers, 21.5M trainable parameters

**Hardware and Software:**

Used hardwares: NVIDIA A100, Tesla T4, GeForce GTX 1080 @ Wigner Scientific Computing Laboratory

Framework: Tensorflow 2.4.1, Keras 2.4.0

### Results

**Validation at the training energy: $\sqrt{s}=7$ TeV, proton-proton collisions**

Good Qualitative (and some quantitative) agreement with the reference Monte Carlo data
No significant difference between the models
Indicates that the hadronization process is captured indeed



**Scaling with the center-of-mass energy:**

The models learned to extrapolate the observable quantities to other collision energies



### Summary

- Machine learning method for investigating hadronization: transform the partonic state into hadronic event-by-event statistical quantities
- Two model complexities: no significant difference
- Partonic interactions turned off: good qualitative agreement
- Accurate extrapolation to other center-of-mass energies
- Open questions:
  - Other collision systems and energies, observables?
  - Network complexity? Dimensionality?

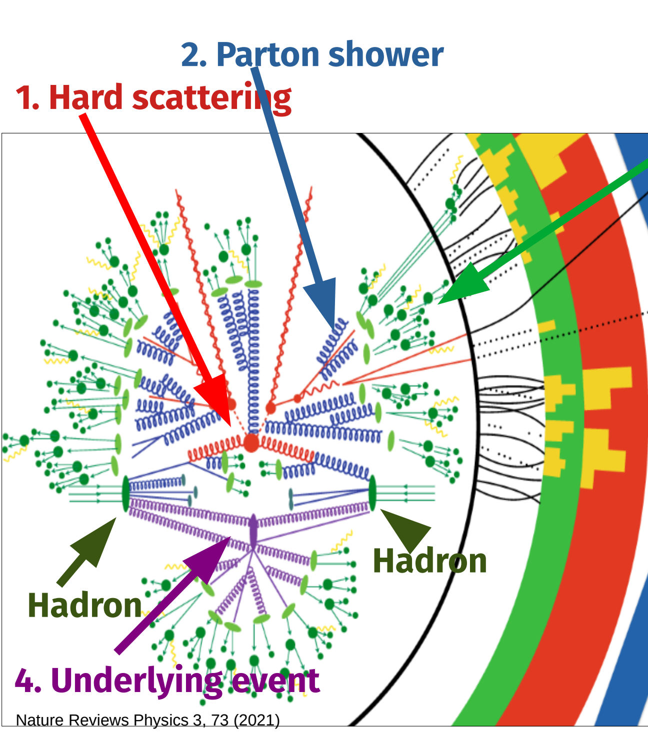
### Acknowledgement


This work was supported by the Hungarian National Research Fund OTKA grant K135515 and K123815, NKFIH 2019-2.1.11-TÉT-2019-00078, 2019-2.1.11-TÉT-2019-00050; Wigner Scientific Computing Laboratory (WSCLAB) (the former Wigner GPU Laboratory), and by the Ministry of Innovation and Technology NRDI Office within the framework of the MILAB Artificial Intelligence National Laboratory Program.

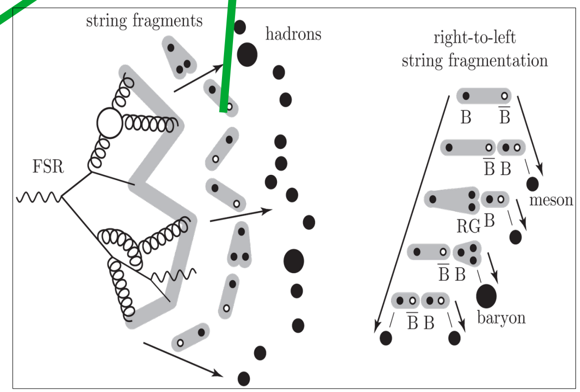

### References


[1] G. Bíró, B. Tankó-Bartalis, G.G. Barnaföldi (2021) arXiv:2111.15655
[2] G. Bíró, B. Tankó-Bartalis, G.G. Barnaföldi (2022) PoS ICHEP2022 1188
[3] Sjostrand, T. Comput. Phys. Commun. (1982) 27, 243
[4] Monk, J.W. JHEP (2018) 12, 021
[5] Chollet, F.; et al. (2015) https://keras.io; Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. (2016) arXiv:cs.DC/1603.04467
[6] K. He, X. Zhang, S. Ren, J. Sun (2015) arXiv:1512.03385

# Hadronization with Machine Learning




Nature Reviews Physics 3, 73 (2021)




## Input:

Parton level

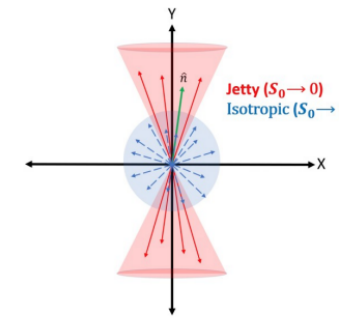
Discretized in the $(y, \phi)$ plane: $p_T$, m, multiplicity

$y \in [\pi, \pi]$ 32 bins

$\phi \in [0, 2\pi]$ 32 bins

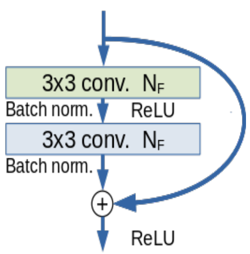
## Hadron level output:

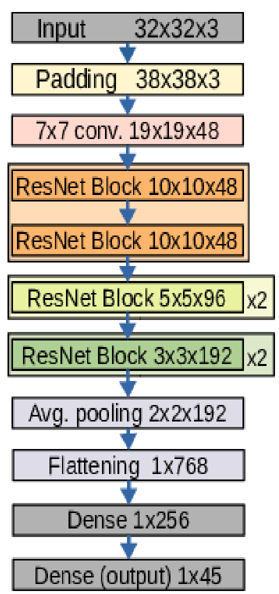
(Charged) event multiplicity, (tr-)sphericity, mean jet $p_T$, -mass, -width, -multiplicity

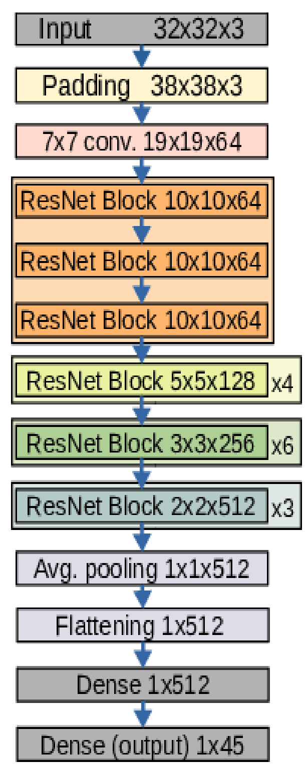



| | Model S | Model L |
|---|---|---|
| Trainable parameters | 1.7 M | 20 M |

## ResNet blocks:




Used hardwares: Nvidia Tesla T4, GeForce GTX 1080 @ Wigner Scientific Computing Laboratory

Framework: Tensorflow 2.4.1, Keras 2.4.0

## Monte Carlo data: Pythia 8.303

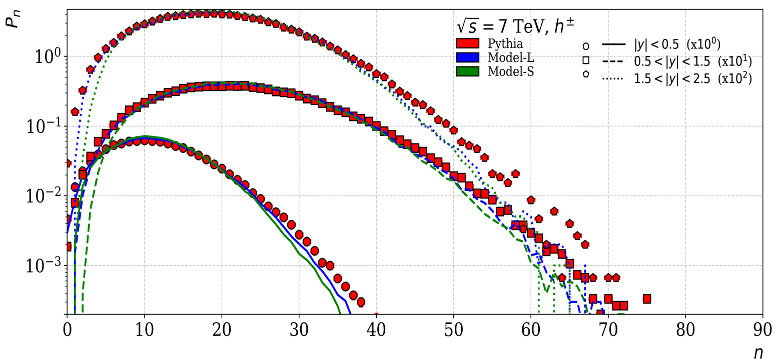
Monash tune
Rescattering and decays turned off
ISR, FSR, MPI: turned on
Selection:

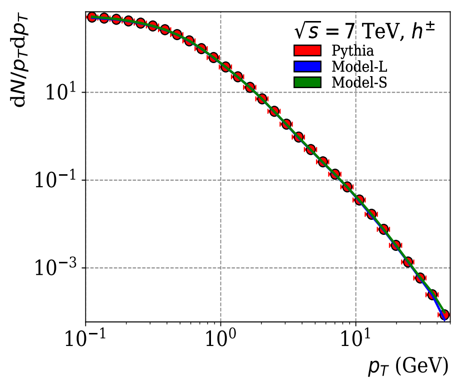
- All final particles with $|y| < \pi$
- At least 2 jets
  - Anti-$k_T$
  - R=0.4
  - $p_T$>40 GeV

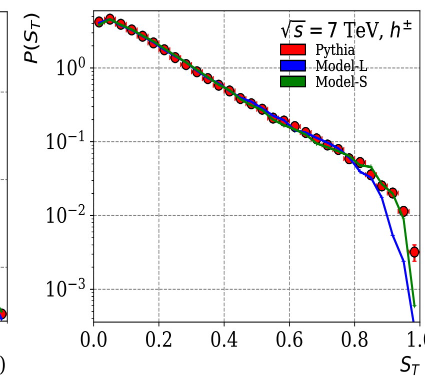
Event number:

- Train: 750 000, $\sqrt{s}$ = 7 TeV
- Validation and test: 100 000
- ~20 GB raw data

# pp @ LHC, Training, validation and  predictions



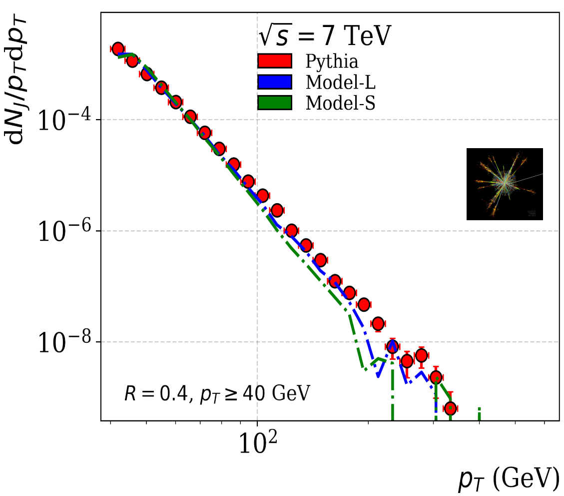



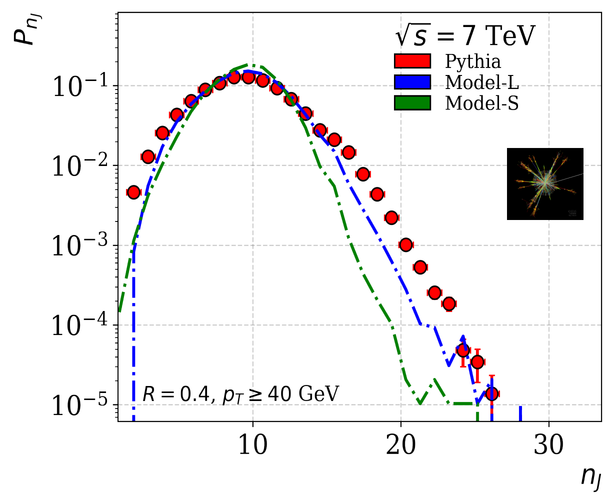

Charged hadron multiplicity at various rapidity windows

**Good agreement for both models**

Charged hadron transverse momentum

0.1 GeV ≤ pT ≤ 50 GeV



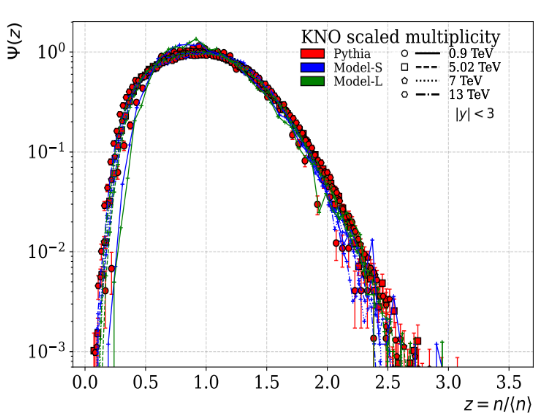



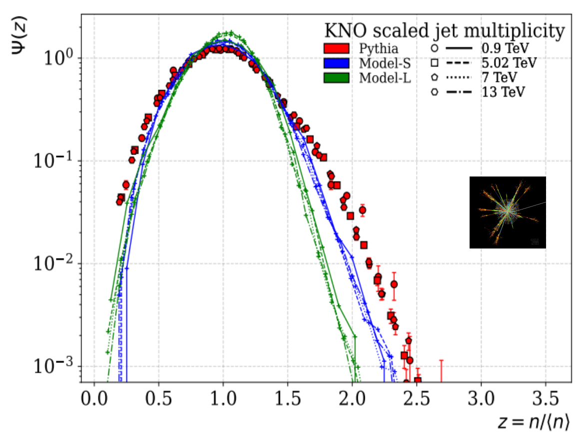

Jets:

- Mean $p_T \leq 400$ GeV
- Mean multiplicity

**The smaller model performs better**

**Training only at a single c.m. energy, predictions at other energies**




Scaling function for multiplicities at various energies:

$$P_n = \frac{1}{\langle n \rangle} \Psi \left( \frac{n}{\langle n \rangle} \right)$$

Charged hadron multiplicities in **jetty** events: good overlap and agreement at **all LHC energies**

Mean jet multiplicities: different scaling for the models