Towards data-driven models of hadronization

Christian Bierlich [1♠] **Phil Ilten** [2†] **Tony Menzo** [2,3,4⋆] **Stephen Mrenna** [2,5✠]

Manuel Szewc [2‖] **Michael K. Wilkinson** [2⊥] **Ahmed Youssef** [2‡] **Jure Zupan** [2,3,4§]

[1] Department of Physics, Lund University, Box 118, SE-221 00 Lund, Sweden
[2] Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221, USA
[3] Berkeley Center for Theoretical Physics, University of California, Berkeley, CA 94720, USA
[4] Theoretical Physics Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5] Scientific Computing Division, Fermilab, Batavia, Illinois 60510, USA

[♠]christian.bierlich@hep.lu.se, [†]philten@cern.ch, [⋆]menzoad@mail.uc.edu, [✠]mrenna@fnal.gov,
[‖]szewcml@ucmail.uc.edu, [⊥]michael.wilkinson@uc.edu, [‡]youssead@ucmail.uc.edu,
[§]zupanje@ucmail.uc.edu,

MLHAD

Abstract

This paper introduces two novel machine learning based approaches to improve hadron-level simulation by integrating experimental observables: **M**icroscopic **A**lterations **G**enerated from **IR** **C**ollections (**MAGIC**), which fine-tunes normalizing flows, pre-trained on simulated data from PYTHIA, on experimental observables, and the **C**ollective **R**eweighting **M**ethod **(CRM)**, which reweights existing fragmentation functions to match experimental observables with a two-step procedure that makes use of a observable-level classifier and hadron-level particle cloud-based regressor. Both methods show a promising direction towards data-driven models for hadronization.

1 Introduction

Monte Carlo Event Generators (MCEG), such as PYTHIA [1], play a vital role in both theoretical and experimental high-energy particle physics research. At collider experiments, MCEGs are used to provide state-of-the-art theoretical predictions that can be directly compared with measured data. A significant challenge within MCEGs is the simulation of how individual quarks and gluons combine into composite objects known as hadrons (*e.g.* protons) which are then observed by detectors. This process is known as *hadronization* and to-date there is still no rigorous theoretical framework that allows for its calculation. Instead, MCEGs rely on physics-inspired phenomenological models such as the Lund string model [2, 3] and the cluster model [4–6]. Although generally successful at describing large amounts of data, these models fail to describe selected subsets of data, motivating the exploration of alternative data-motivated approaches.

Initial efforts to provide a machine learning representation of simplified hadronizing systems employed both (MLHAD) conditional sliced Wasserstein autoencoders (cSWAE) [7] and (HADML) Generative Adversarial Networks (GANs) [8, 9], which were successful in emulating key system attributes.

However, these architectures are limited by their reliance on hadron-level training data that is not available at the experimental level. In experiments, data is accessible via observables, where hadrons

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

are collected as unordered sets, in which the dynamics of generation cannot be directly resolved. This restricts access to hadron-level data from experiments, which in turn limits the ability to train these models on real-world data as they are currently developed.

In this paper, we address this limitation by demonstrating the feasibility of two novel approaches that learn the individual hadron fragmentation dynamics from experimentally available collections of data. Both methods perturb existing hadronization models by the application of statistical reweighting, to produce a new model that better agrees with data.

With the first method, we build on the work in [7], utilizing normalizing flows (NF) [10–12] rather than a sliced Wasserstein autoencoder (SWAE) [13]. The NF architecture offers a number of benefits over the previously analyzed SWAE architecture, both in efficiency and physics modelling. Most notably, it allows for individual microscopic hadron dynamics to be learned using macroscopic experimental observables through a novel training paradigm termed **M**icroscopic **A**lterations **G**enerated from **IR** **C**ollections or **MAGIC**. Here, IR (infrared) collections refer to any ensemble of observables which are sensitive to non-perturbative effects at low energies, *i.e.* in the infrared regime.

In the second method, the **C**ollective **R**eweighting **M**ethod (**CRM**) of the hadronization function, we first encode the likelihood ratio between experimental and simulated data and then translate this to a corrected hadronization model by using the hadronization history for each simulated event. Rather than attempting to learn an arbitrary function, this method builds on existing knowledge, integrating a new physics-inspired approach to optimize the event generation of hadronization models.

2 Method and Results

We use the Lund string model prescription of hadronization to demonstrate the performance of our two strategies. Within the Lund string model, hadronizing quark/anti-quark pairs are connected via QCD flux strings, whose energy linearly grows with increasing quark/anti-quark separation. For large enough separations, it becomes energetically favorable to produce additional quark/anti-quark pairs along the string, causing the string to undergo iterative breaking. Each break emits a hadron h, while conserving energy, momentum, and flavor. A comprehensive architecture and illustration of this process can be found in [7]. The Lund string model has been tested extensively, and its parameters tuned to the relevant experimental observable distributions. Although generally in agreement with data, there are still instances where the model outputs deviate from experimental data. For instance, comparison of data from proton-proton and ion-ion collision with PYTHIA show discrepancies at the level of $O(20\%)$ to $O(50\%)$ [14]. These deficiencies underscore a need for more refined experimental observables and more sophisticated hadronization models. Given a lack of clear theoretical direction, considering data-driven hadronization models is essential not only to produce better descriptions of data, but also to help to understand the underlying physics.

2.1 MAGIC

MAGIC Method: MAGIC has two training phases. First, an NF, referred to as the *base–model*, is constructed using the FREIA [15] software library and trained on simulated single hadron emission kinematics $\boldsymbol{x} = \{p_z, p_T\}$ similar to what was done in [7]. The second fine-tuning phase of training requires a three-component training dataset: (1) hadron-level kinematics $\boldsymbol{X}$, simulated from the base-model (2) simulated observables $\boldsymbol{Y}^{\text{sim}}$, obtained from the simulated hadron-level kinematics, and (3) real target observables from experiment $\boldsymbol{Y}^{\text{exp}}$. As a proof of principle, we utilize only a single observable, namely, the total number of hadrons in an event, known as the hadron multiplicity. The objective of the fine-tuning phase is to modify the base-model's likelihood such that the statistical distance between the updated model observable $\boldsymbol{Y}^{\text{sim}'}$ and $\boldsymbol{Y}^{\text{exp}}$ is minimized. This is implemented in practice by utilizing the Wasserstein or earth mover's distance (EMD) [16–19] as the loss function.

The primary challenge of fine-tuning the base model on experimental observables lies in the large number of events needed to compute the observable. Without access to the model likelihood, the events would need to be simulated again after every training iteration to obtain the simulated observables in the context of the updated model likelihood. Simulating events after each model update is not computationally feasible. To avoid this, we utilize the access to the model likelihood provided by NFs, and assign probabilistic weights for each emission in each event of a given training

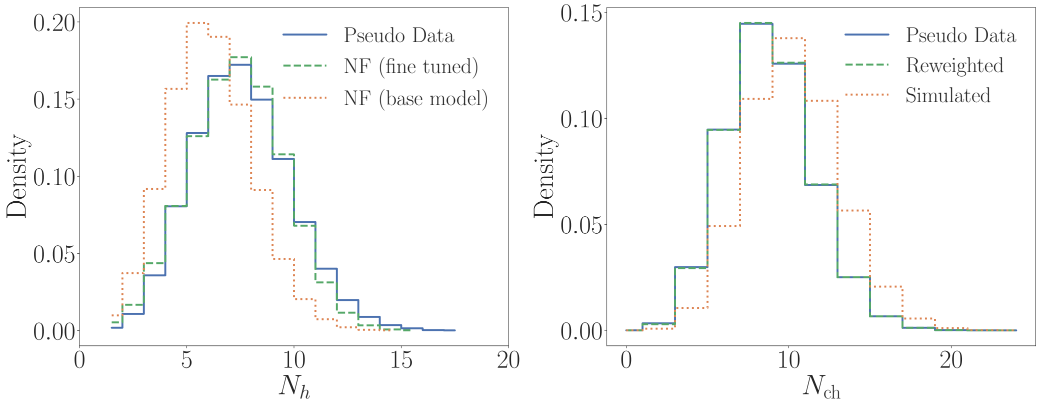

Figure 1: **Left using MAGIC:** Comparison of hadron multiplicity N_h among base, fine-tuned, and target "experimental" distributions, which is donated as pseudo data. **Right using CMR:** Comparison of the charged multiplicity N_{ch} between PYTHIA simulated, experimental (pseudo) data, and hadron-level reweighted PYTHIA simulation.

batch. This weight is computed using the likelihood ratio between the base and updated model and stored in an event weight array with a length equal to that of the training batch:

$$\boldsymbol{w} = \begin{pmatrix} \prod_{i=1}^{N_1} w_i \\ \prod_{j=1}^{N_2} w_j \\ \vdots \\ \prod_{k=1}^{N_n} w_k \end{pmatrix} , \quad \text{with} \quad w_i = \frac{\mathcal{P}_X^{F'}(p_z^{h_i}, p_T^{h_i})}{\mathcal{P}_X^{F}(p_z^{h_i}, p_T^{h_i})} , \tag{1}$$

where $\mathcal{P}_X^{F'}$ and $\mathcal{P}_X^{F}$ represent the updated and base model likelihood functions, respectively. In this way, for each training batch, we obtain a multiplicity sample from the updated model $\boldsymbol{Y}^{\text{sim}'}$, by reweighting each multiplicity value in $\boldsymbol{Y}^{\text{sim}}$.

MAGIC Training and Results: For the "experimental" hadron multiplicity samples $\boldsymbol{Y}^{\text{exp}}$ we create pseudo data by generating $N = 10^5$ PYTHIA hadronization events with a non-default Lund a parameter[1] value of 1.0. The baseline model undergoes fine-tuning over 20 epochs using a fixed learning rate of $\delta = 10^{-4}$.

A comparison between the hadron multiplicity distributions obtained from the base and fine-tuned models against the targeted experimental distribution is shown in fig. 1. The results reveal that the kinematics obtained from the **MAGIC** fine-tuned model produce a hadron multiplicity distribution in good agreement with that of the target experimental distribution, demonstrating the efficacy of the **MAGIC** method in enhancing model predictions.

2.2 Collective Reweighting Method of the hadronization function

In this section, we introduce the Collective Reweighting Method **(CRM)**, a novel approach for enhancing hadron-level simulation using experimental observables. In this approach, we train a classifier to distinguish between simulated and experimental observables and a regressor that translates the classifier-derived event weights into an individual hadron emission reweighting. This can in turn be used to derive a data-driven fragmentation function that better aligns with experimental data.

Classifier and Likelihood Ratio: A classifier is trained to distinguish between a set of simulated events $\boldsymbol{Y}^{\text{sim}}$ and a set of experimental events $\boldsymbol{Y}^{\text{exp}}$. For a well-trained classifier, its output $g(\vec{y})$ for an event $\vec{y}$, converges to a monotonic function from which we can extract the likelihood ratio between experimental data and simulation, expressed in terms of measured observables. This function can

[1] Within PYTHIA, the Lund a parameter is referred to as `StringZ:aLund` with a default value of 0.68.

thus be utilized to match simulations to data by reweighting, with the event weight being

$$w(\vec{y}) = \frac{g(\vec{y})}{1 - g(\vec{y})} = \frac{\mathcal{P}(\vec{y}|\text{exp})}{\mathcal{P}(\vec{y}|\text{sim})} , \tag{2}$$

where the second equality is the likelihood ratio. Assuming discrepancies between data and simulation arise from hadronization mismodelling, the learned event weight expressed in terms of measured observables $w(\vec{y})$ can be re-expressed in terms of the collection of hadrons contained in each event $\{\vec{h}_k, k = 1, ..., K\}$:

$$w(\vec{y}) \equiv w(\{h_k\}) = \frac{\mathcal{P}(\{h_k\}|\text{exp})}{\mathcal{P}(\{h_k\}|\text{sim})} , \tag{3}$$

where $\mathcal{P}(\{h_k\}|\text{sim})$ is known, if not analytically then numerically through PYTHIA simulation. We frame the problem of learning $w(\vec{y}) = f(\{h_k\})$ as a regression problem, where f is a learned function. To do this, we represent each collection of emissions as a directed graph, where the node features and edges are given by the hadronization history produced from PYTHIA, and we make use of graph neural networks (GNNs) to regress $w(\{h_k\})$. For the simplified quark anti-quark initial string studied here, the hadronization history implemented by PYTHIA is a Markov process where each emission is dependent only on the previous emission. With this factorization structure, we can parameterize the logarithm of the event weight as the sum of the logarithms for individual conditional probabilities

$$\log w(\{h_k\}) = \sum_{k=1}^{K} f(h_k|h_{k-1}) , \tag{4}$$

where h_0 is a null node and f is a learned function which is applied repeatedly across the event and is parameterized in terms of a message passing neural network [20]. The method thus makes use of the fact that in simulation we have access to underlying information which is not available in data. By translating differences at the observable level to a reweighting function at the hadron level and combining these weights with our knowledge of $\mathcal{P}(h_k|\text{sim})$, we obtain a data-driven fragmentation function $\mathcal{P}(h_k|h_{k-1}, \text{exp}) = f(h_k|h_{k-1})\mathcal{P}(h_k|h_{k-1}, \text{sim})$ that better aligns the simulation with the experimental data at the observable event level, leading to a more accurate simulation.

CRM training and results. To illustrate the method, we simulate electron-positron collisions with a center-of-mass energy at the measured Z-boson mass, with the Z decaying into a light quark/anti-quark pair, and allow only hadronization to pions without decaying. To keep the simple $q\bar{q}$ structure, we do not include parton showers before hadronization. The simulated dataset is obtained with PYTHIA by setting the a, b and σ_{p_T} parameters to their Monash tune [21] values $a = 0.68$, $b = 0.98$ and $\sigma_{p_T} = 0.335$. For the measurement simulation, we also use PYTHIA but change a to 0.3 and fix the other parameters. For each event, we consider nine high-level observables that are accessible in real data, including charge multiplicity, and for the simulated dataset we store nine node features for each simulated hadron.

For the first step, a gradient boosting classifier implemented in XGBOOST [22] was trained on 10^6 events with a learning rate of 0.5 and $\text{max}_{\text{depth}}$ of 5. The hyperparameters are tuned to ensure optimal calibration. For the second step, we utilize a POINTNET architecture [23] to perform point cloud regression, which was trained on 10^5 events, each consisting of a variable length set of unordered emissions. For the POINTNET architecture, the message function is a multilayer perceptron with three layers, with the first two layers containing 256 neurons each and a ReLU activation function, and the last layer consisting of a linear layer with a single neuron. Thus, the message passing function is by construction the f function shown in eq. (4).

In fig. 1 (right) we show the results for the charge multiplicity (N_{ch}), which show a notable improvement in matching the experimental data through event reweighting. This is then translated to a data-driven fragmentation function which shows good agreement with the one used to generate the experimental pseudo data.

3 Discussion and Outlook

In this paper, we introduced two novel approaches on how to train hadron-level events on experiment-accessible observables. While the capability of **MAGIC** was demonstrated on pseudo-experimental data, it showed promising results by only employing only one observable. **CRM** on the other hand was applied to more observables and showed the capability to tune hadronization models to pseudo data, albeit with a higher bias deriving from its larger reliance on the Lund string model. A more detailed description of **MAGIC** can be found in [24]. Future work will apply both of these method to more experimentally accessible observables, using more realistic string topologies, and a detailed study of both methods will be published in the near future.

Broader Impact

A data-driven hadronization model will significantly impact a large range of collider experiments, allowing for more accurate theoretical predictions while also providing checks on theoretical assumptions such as factorization and universality. Beyond its use in particle physics, the learned techniques can be generalized to any problems where empirical simulators with known probabilistic models need to be improved upon to match real-world data. With its compromise between model bias and data-driven focus, the proposed methods can be a robust alternative to fully data-driven generative models.

Acknowledgments. AY, JZ, MS, and TM acknowledge support in part by the DOE grant DE-SC1019775, and the NSF grant OAC-2103889. JZ acknowledges support in part by the Miller Institute for Basic Research in Science, University of California Berkeley. SM is supported by the Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. CB acknowledges support from the Knut and Alice Wallenberg foundation, contract number 2017.0036. PI is supported by NSF grant OAC-2103889 and NSF-PHY-2209769. TM acknowledges support in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research (SCGSR) program. The SCGSR program is administered by the Oak Ridge Institute for Science and Education for the DOE under contract number DE-SC0014664.

References

[1] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. doi: 10.1016/j.cpc.2015.01.024.

[2] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys. Rept.*, 97:31–145, 1983. doi: 10.1016/0370-1573(83)90080-7.

[3] Bo Andersson. The Lund model. *Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol.*, 7:1–471, 1997.

[4] Richard D. Field and Stephen Wolfram. A QCD Model for e+ e- Annihilation. *Nucl. Phys. B*, 213:65–84, 1983. doi: 10.1016/0550-3213(83)90175-X.

[5] Thomas D. Gottschalk. An Improved Description of Hadronization in the {QCD} Cluster Model for e^+e^- Annihilation. *Nucl. Phys. B*, 239:349–381, 1984. doi: 10.1016/0550-3213(84)90253-0.

[6] B.R. Webber. A QCD Model for Jet Fragmentation Including Soft Gluon Interference. *Nucl. Phys. B*, 238:492–528, 1984. doi: 10.1016/0550-3213(84)90333-X.

[7] Phil Ilten, Tony Menzo, Ahmed Youssef, and Jure Zupan. Modeling hadronization using machine learning. *SciPost Phys.*, 14:027, 2023. doi: 10.21468/SciPostPhys.14.3.027. URL `https://scipost.org/10.21468/SciPostPhys.14.3.027`.

[8] Aishik Ghosh, Xiangyang Ju, Benjamin Nachman, and Andrzej Siodmok. Towards a deep learning model for hadronization. *Phys. Rev. D*, 106(9):096020, 2022. doi: 10.1103/PhysRevD.106.096020.

[9] Jay Chan, Xiangyang Ju, Adam Kania, Benjamin Nachman, Vishnu Sangli, and Andrzej Siodmok. Fitting a Deep Generative Hadronization Model. 5 2023.

[10] Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, nov 2021. doi: 10.1109/tpami.2020.2992934. URL `https://doi.org/10.1109%2Ftpami.2020.2992934`.

[11] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/rezende15.html`.

[12] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1410.8516`.

[13] Soheil Kolouri, Charles E. Martin, and Gustavo K. Rohde. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *CoRR*, abs/1804.01947, 2018. URL `http://arxiv.org/abs/1804.01947`.

[14] Nadine Fischer and Torbjörn Sjöstrand. Thermodynamical string fragmentation. *Journal of High Energy Physics*, 2017(1):140, 01 2017. ISSN 1029-8479. doi: 10.1007/JHEP01(2017)140. URL `https://doi.org/10.1007/JHEP01(2017)140`.

[15] Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. Framework for Easily Invertible Architectures (FrEIA), 2018-2022. URL `https://github.com/vislearn/FrEIA`.

[16] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. The Hidden Geometry of Particle Collisions. *JHEP*, 07:006, 2020. doi: 10.1007/JHEP07(2020)006.

[17] C´ edric Villani. *Optimal transport, old and new*. Springer, Berlin, 2008.

[18] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs and Modeling*. 2015. URL `https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf`.

[19] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[21] Peter Skands, Stefano Carrazza, and Juan Rojo. Tuning PYTHIA 8.1: the Monash 2013 Tune. *Eur. Phys. J. C*, 74(8):3024, 2014. doi: 10.1140/epjc/s10052-014-3024-y.

[22] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145%2F2939672.2939785`.

[23] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017.

[24] Christian Bierlich, Phil Ilten, Tony Menzo, Stephen Mrenna, Manuel Szewc, Michael K. Wilkinson, Ahmed Youssef, and Jure Zupan. Towards a data-driven model of hadronization using normalizing flows, 2023.