Aspen Open Jets: Unlocking LHC Data for Foundation Models in Particle Physic

FERMILAB-PUB-24-0941-AD

arXiv:2412.10504

Aspen Open Jets: Unlocking LHC Data for Foundation Models in Particle Physics

Oz Amram,[1,*] Luca Anzalone,[2,3,†] Joschka Birk,[4,‡] Darius A. Faroughy,[5,§] Anna Hallin,[4,¶] Gregor Kasieczka,[4,6,‖] Michael Krämer,[7,**] Ian Pang,[5,††] Humberto Reyes-Gonzalez,[7,‡‡] and David Shih[5,§§]

[1] *Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[2] *Department of Physics and Astronomy (DIFA), University of Bologna, 40127 Bologna, Italy*
[3] *Istituto Nazionale di Fisica Nucleare (INFN), 40127 Bologna, Italy*
[4] *Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*
[5] *NHETC, Dept. of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, USA*
[6] *Center for Data and Computing in Natural Sciences (CDCS), 22607 Hamburg, Germany*
[7] *Institut für Theoretische Teilchenphysik und Kosmologie, RWTH Aachen University, 52074 Aachen, Germany*

Foundation models are deep learning models pre-trained on large amounts of data which are capable of generalizing to multiple datasets and/or downstream tasks. This work demonstrates how data collected by the CMS experiment at the Large Hadron Collider can be useful in pre-training foundation models for HEP. Specifically, we introduce the AspenOpenJets dataset, consisting of approximately 180M high p_T jets derived from CMS 2016 Open Data. We show how pre-training the OmniJet-α foundation model on AspenOpenJets improves performance on generative tasks with significant domain shift: generating boosted top and QCD jets from the simulated JetClass dataset. In addition to demonstrating the power of pre-training of a jet-based foundation model on actual proton-proton collision data, we provide the ML-ready derived AspenOpenJets dataset for further public use.

I. INTRODUCTION

While particle physics has long used machine learning techniques and is leading the way in adopting modern deep learning methods, the development and application of powerful foundation models – which have transformed natural language processing and computer vision – is still in its early stages in the field. A foundation model is a model that has been pre-trained on a large amount of data for a specific task, and that can be used for different tasks downstream [1]. The pre-training thus serves as a foundation, upon which the downstream tasks can be built. Well-known foundation models include large language models (LLMs) such as GPT-3 [2], BERT [3] and LLaMA [4], image models such as DALL-E 2 [5] and Imagen [6], and mixed-modality models such as CLIP [7] and ALIGN [8].

There are several advantages to the use of foundation models. One example is the potential to boost the performance achievable on small datasets. Having access to a model pre-trained on a larger, similar dataset provides a 'head start' for the downstream task, allowing it to focus on the intricacies of the smaller dataset rather than having to rediscover the general structure of the data. In addition, a foundation model can save both human and computational resources. While pre-training may be a resource-intensive task, the downstream models would require less training, less data, and less time spent on optimization compared to what would be required if these models were built from scratch. Another interesting aspect is that a trained foundation model could enable implicit sharing of data and resources between different parts of the research community, a sharing that would not otherwise be possible.

Foundation models are poised to play a central role in particle physics, especially in the context of experiments at the Large Hadron Collider (LHC). They are able to process the huge amounts of complex and diverse data generated by the LHC and have the potential to integrate multiple modalities - such as images, high-level event features or particle four-vectors - allowing more comprehensive data analyses. Current work on foundation models for particle physics has mainly focused on different event generation and classification tasks [9–14], investigating pre-training strategies, encoding schemes and different architectures, and quantifying the transfer learning and generalization capabilities of these models.

The LHC experiments have made significant amounts of data publicly available [15–19]. Previous open data studies of CMS jets [20–22] have explored the substructure and energy flow of the jets. Foundation models can leverage the LHC open data for training and fine-tuning, potentially integrating different data modalities and data from different experiments. Furthermore, training foundation models on real data will result in a more accurate representation of particle physics processes than training on simulations. Finally, foundation models trained on a large variety of actual LHC data will help to promote transparency and collaboration in particle physics

* oamram@fnal.gov
† luca.anzalone2@unibo.it
‡ joschka.birk@uni-hamburg.de
§ darius.faroughy@rutgers.edu
¶ anna.hallin@uni-hamburg.de
‖ gregor.kasieczka@uni-hamburg.de
** mkraemer@physik.rwth-aachen.de
†† ian.pang@physics.rutgers.edu
‡‡ humberto.reyes@rwth-aachen.de
§§ shih@physics.rutgers.edu

research.

In this paper we present the AspenOpenJets (AOJ) dataset and study the transfer learning of foundation models obtained with this dataset. The AOJ dataset has been constructed from the CMS 2016 open data 'JetHT' [17, 18] and prepared in a form that allows easy use of the dataset for phenomenological studies, especially for training machine learning models. With 180 M jets, AOJ is by far the largest dataset of its kind. We demonstrate an application of this dataset in a foundation model context, by pre-training the OmniJet-α [12] foundation model on AOJ, which is expected to mostly contain QCD jets, and then fine-tuning it to generate jets from the JetClass [23] dataset of simulated jets. We do two such fine-tunings: first for JetClass QCD jets, which occupy a different kinematic region compared to the jets in AOJ, and then for top jets, which differ significantly in several features. We show that pre-training on AOJ indeed benefits the downstream task, making it easier for the model to generate these types of jets with fewer training examples compared to if it had been trained from scratch.

The utility of CMS open data in a pre-training setting has been previously explored in [9], using 1M single-lepton events at the object level which were collected by CMS in 2015. Here we enlarge the scope and complexity by several orders of magnitude to encompass 180M jet events at the full constituent level. Furthermore, the authors of [9] used event selections to shape the open data into a similar topology to the intended target simulated data. In our work, no such shaping is needed.

This paper is organized as follows. Section II describes the CMS open data and the method used to derive AOJ. It also includes a brief description of JetClass, which will be used for the downstream task. Our application example is described in section III, detailing the method and results. Finally, we present our conclusions in section IV. Links to the codebase and the AOJ dataset itself can be found after the conclusions.

II. DATA

A. CMS open data

The CMS experiment [24] is a large general-purpose detector at the LHC. CMS has recently released 16.4 fb^{-1} of data from their 2016 runs (G and H) [17, 18] and a corresponding set of simulations. In contrast to previous data releases [15, 16, 25], this data has been provided in the simplified `MINIAOD` format [26], as well as the even further simplified `NANOAOD` format [27]. In these simplified analysis formats, the central CMS reconstruction steps have been applied, discarding much of the low-level content of the event in favor of a simplified description of the high-level objects (jets, electrons, muons, etc.) commonly used in data analysis. These formats are therefore significantly easier to use for those outside of the collaboration.

The inner layers of CMS consist of a silicon pixel and silicon strip tracker used to measure charged particle trajectories. These are followed by a lead tungstate crystal electromagnetic calorimeter and a brass and scintillator hadronic calorimeter. These detector elements are enclosed in a large solenoid, which produces a magnetic field of 3.8 T. Muon stations are located outside the iron return yoke of the magnet. Events of interest are selected using a two-stage trigger system. The first stage (L1) is implemented in custom hardware to select potentially interesting events within a fixed latency of about 4 μs [28] using information from the calorimeters and muon stations. Events accepted by the L1 system are sent to the high-level trigger (HLT), which consists of a dedicated computing cluster which runs a speed-optimized version of the full reconstruction. The HLT uses this more accurate reconstruction of the event to select events to be saved at a rate of approximately 1 kHz [29].

CMS uses a particle-flow algorithm [30], which combines the information from the different detector systems to reconstruct and identify each individual particle in an event. Particles are classified as either muons, electrons, photons, charged hadrons or neutral hadrons. Jets are clustered from the particle-flow candidates in an event using the anti-k_t jet finding algorithm [31, 32]. The typical large-radius jet size used in CMS for jet substructure searches and measurements is $R = 0.8$, commonly referred to as AK8 jets. For AK8 jets the pileup-per-particle identification algorithm (PUPPI) [33, 34] is used to mitigate the effect of pileup. The PUPPI algorithm assigns a weight to each particle based on the probability that it originates from the primary vertex or from pileup. The momentum of AK8 jets is determined from the weighted vector sum of all particles inside the jet. Additional corrections are applied based on simulation and data-driven measurements to better calibrate the measured jet momentum to the true momentum of the particle from the hard process [35].

The `NANOAOD` format does not natively store information of the jet constituents. However, an extended version (`PFNANO`), which does include this information, can be easily produced from `MINIAOD` using tools provided by CMS [36]. We therefore, for all data and simulations samples used, start with the `MINIAOD` format, process it to `PFNANO` and then apply our selections to select the jets of interest for our dataset.

In order to construct the largest sample of jets possible, we consider all events from the 'JetHT' [17, 18] Run-2 dataset released by CMS to this date. The 'JetHT' dataset includes all events that have been selected by one of the HLT triggers related to jet momenta or total event hadronic activity (HT). Because we do not specify a specific trigger, the sample consists of events from an inclusive 'or' of any of the available triggers of this dataset. We note that this choice works well for building a large dataset for the purposes of foundation model training, but may introduce subtle kinematic biases resulting from

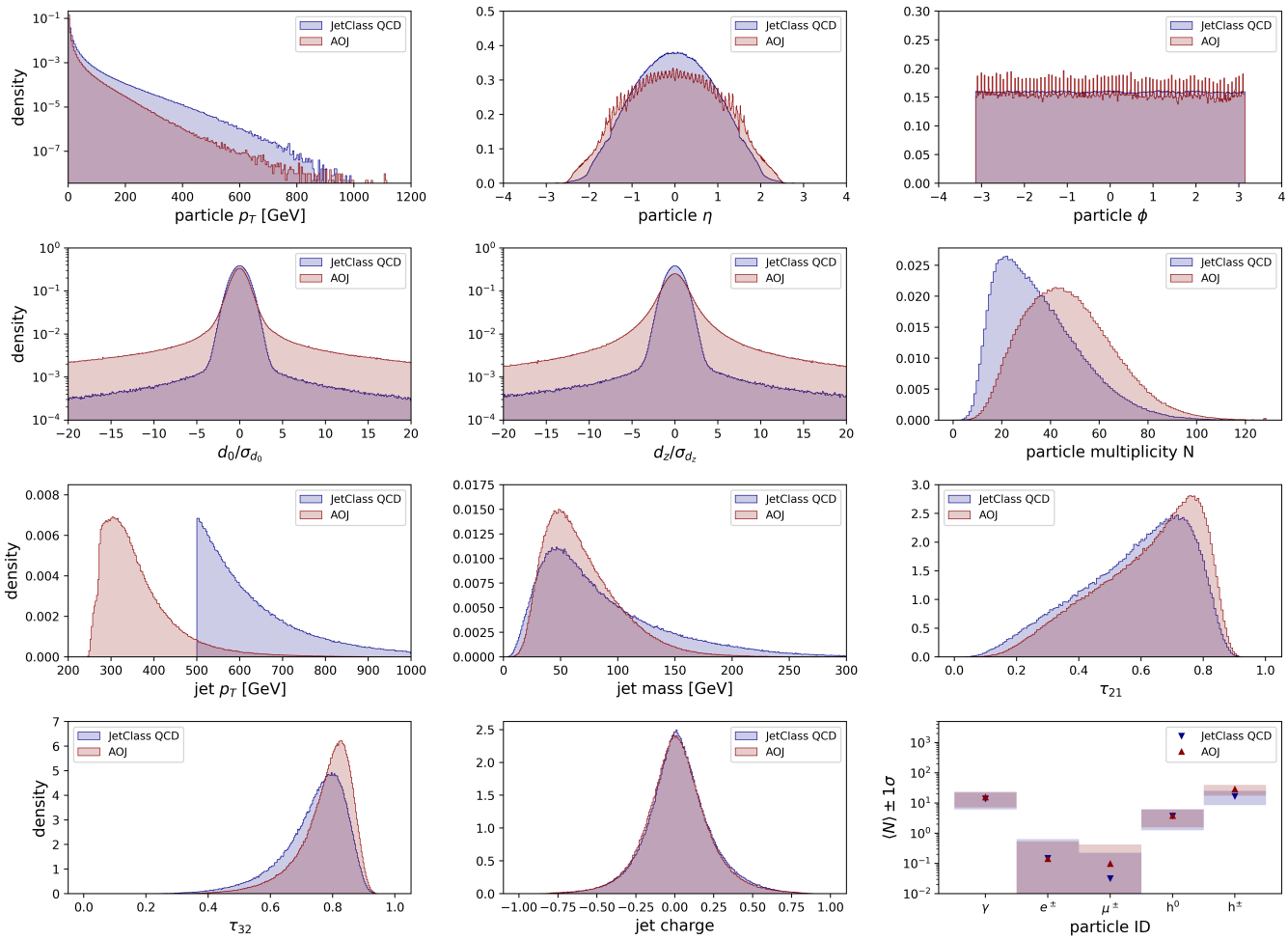

FIG. 1. Comparison of AspenOpenJets (AOJ) and QCD JetClass distributions for particle and jet level observables. The last panel shows the average number of jet constituents for each particle species, accompanied by the corresponding 1σ standard deviation bands.

the combined thresholds of the different triggers, which would complicate its use in a physics-based analysis.

Events are required to pass several data-quality filters commonly used within CMS to remove events that are likely to cause problems in the reconstruction. AK8 jets are required to have transverse momentum greater than 300 GeV[1]. Jets are also required to have an absolute value of pseudorapidity (η) less than 2.5, so that they fall within the range of the CMS tracker. The AK8 jets are also required to pass standard CMS quality criteria used to reject leptons reconstructed as jets and jets that may be poorly reconstructed. We construct our sample out of all AK8 jets that pass these criteria, with no limitation on the number of selected jets per event.

Applying these selections to the CMS open data yields a dataset of 178 million jets, which we call the AspenOpenJets (AOJ). Due to the large cross section of QCD processes, the vast majority of these jets are expected to originate from light quarks and gluons. A small fraction of the jets are expected to originate from the decays of boosted heavy resonances. A rough estimate of the number of boosted fully merged jets produced by W, Z, top, and Higgs bosons in the $b\bar{b}$ decay mode are estimated with Monte Carlo (MC) simulations provided by CMS [37], with the remainder assumed to originate from QCD. The numbers are reported in Table I. More details on the MC samples used are given in Appendix A.

For each jet we store its transverse momentum (p_T), pseudorapidity (η), and azimuthal angular coordinate (ϕ). We also store its mass, groomed with the softdrop algorithm [38] as computed within the CMS reconstruction. Up to 150 constituents of the jet are stored. For each constituent, its 4-momentum is stored in the format (p_x, p_y, p_z, E). We additionally store its transverse im-

[1] The jet transverse momentum (p_T) distribution in fig. 1 extends down to 250 GeV as it is computed directly from vector sum of the jet constituents. Whereas, 300 GeV threshold is applied to the jet after additional corrections have been applied to the jet energy.

Jet type	Approx. Number
QCD	1.8×10^8
W	6.3×10^5
Z	1.8×10^5
top	1.3×10^5
$H \to b\bar{b}$	500

TABLE I. Breakdown of the expected number of jet types in the AOJ dataset from MC simulations.

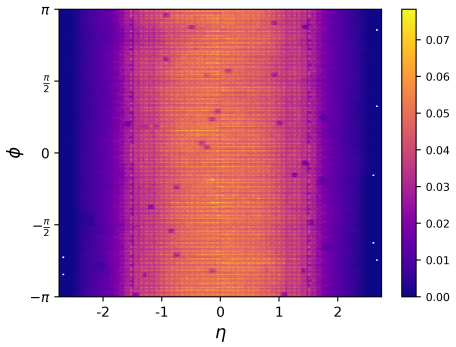

FIG. 2. The particle-level positional distribution of the AspenOpenJets in the $\eta - \phi$ plane, illustrating the granularity of the CMS detector. Note the appearance of vertical features indicating the endcaps at $|\eta| \sim 1.6$ and multiple "dead cells" throughout the detector.

pact parameter (d_0) and longitudinal impact parameter (d_z) with their uncertainties, the charge of the candidate, its particle-ID (PID) in the PDG format[2] [39] and its weight from the PUPPI algorithm. We also include additional jet substructure quantities computed within the CMS reconstruction, including the number of constituents in the jet, N-subjettiness variables [40], various jet-tagging observables from the CMS implementation of ParticleNet [41, 42] and a regression of the jet mass from ParticleNet [42].

In fig. 1 we show the distributions (red histograms) of several particle and jet-level features, based on a subsample of 700 K jets. Interestingly, the effects of detector granularity are seen as spikes in the particle η and ϕ distributions. In fig. 2 we plot a 2D histogram in the $\eta - \phi$ plane. A closer look reveals a grid-like structure corresponding to the granularity of the detector. Furthermore, there are vertical features indicating the endcaps at $|\eta| \sim 1.6$ and several "dead cells" throughout the detector. The distribution of the subjettiness ratio τ_{32} is consistent with the expectation that the AOJ is predominantly composed of QCD jets. The last panel in fig. 1 shows the average number of constituents per jet for each PID, accompanied by the corresponding 1σ standard deviation bands.

B. JetClass

The JetClass [23] dataset is based on simulations and was first introduced in [10]. It contains both particle-level and jet-level information for 125 M simulated jets, distributed over 10 different jet types and divided into a training (100 M), validation (5 M) and test (20 M) set. The simulation is performed with MadGraph5_aMC@NLO [43] for the hard process, followed by Pythia [44] for parton showering and hadronization, and Delphes [45] with the CMS detector [24] card for the detector response. Jets are required to have a transverse momentum of 500 GeV $< p_T <$ 1000 GeV and a pseudorapidity $\eta < 2$, which is more restrictive than the criteria used for AOJ. We overlay the JetClass QCD particle- and jet-level distributions (blue histograms) in fig. 1 to facilitate easy comparison with the corresponding AOJ distributions.

III. EXPERIMENT

In this section, we demonstrate the usefulness of the AOJ dataset in ML-related applications. In particular, we pre-train a foundation model on the full AOJ dataset.

A. Method

The foundation model chosen for this work is OmniJet-α [12], a GPT-style [46] model that predicts the next token in a sequence. The features of the jet constituents[3] are first converted into integer tokens using a VQ-VAE [11, 47–49] model, representing jets as sequences of tokens. The backbone of OmniJet-α is trained to predict the next token in these sequences. After pre-training, the model can generate new sequences autoregressively, which are then decoded by the frozen VQ-VAE back into physical space. Additionally, the pre-trained backbone weights can be loaded and used for other tasks, such as jet classification. In this work, we focus on dataset transfer learning for jet generation and leave the exploration of task switching for future studies.

The features used for training are the constituent kinematics (p_T, $\Delta\eta$, $\Delta\phi$), where $\Delta\eta = \eta - \eta_{\text{jet}}$ and $\Delta\phi = \phi - \phi_{\text{jet}}$ represent the relative coordinates with respect to the jet axis. Before tokenization, these features were preprocessed using standardization. The AOJ dataset was tokenized using a codebook size of 8192 tokens, employing VQ-VAE hyperparameters as outlined

[2] Note that neutral hadrons (h^0) are assigned the PID = 130 of the neutral kaon K_L^0 while positively/negatively charged hadrons ($h^\pm$) are assigned PID = ±211 of the charged pion.

[3] The jet constituents are p_T ordered and up to the first 128 constituents are considered when training the models.

in [12], with two modifications: the number of attention heads was reduced from 8 to 4 without sacrificing performance, and the AdamW optimizer was replaced with the Ranger optimizer [50]. A larger codebook with ∼32 K tokens was also tested but produced results comparable to the 8 K configuration. Therefore, we only present results for the 8 K tokenization. The backbone model was trained on the full 180 M AOJ dataset with hyperparameters as specified in [12]. The model was trained for 900K steps on 12 GPUs with a batch size of 256 per GPU, and the final state at the end of the training was selected as the backbone model.

To evaluate the transfer learning capabilities of OmniJet-α , we devised a downstream task to test its ability to generate jets of a different type than those it was pre-trained on, and to adapt to datasets with different kinematic cuts. The fine-tuning process involved loading the pre-trained backbone weights and retraining it using jets from the JetClass [23] dataset. To systematically evaluate the performance of the model, we trained it on six datasets with 10^2, 10^3, 10^4, 10^5, 10^6, 10^7 jets to study the effect of dataset size D, and fine-tuned the model separately on QCD jets (q/g) and top-quark jets ($t \to bqq'$) to evaluate its adaptability to different jet substructure types. For each configuration, the model with the lowest validation loss during fine-tuning was selected for jet generation. To provide a baseline for comparison, we used a separate VQ-VAE tokenizer and trained generative models with the same backbone architecture from scratch using randomly initialized weights. This VQ-VAE is the one from Ref. [12] that was trained on all 10 classes of the JetClass dataset. The generative models were trained on the tokenized QCD and top-quark samples. This approach allowed us to quantify the performance improvement achieved by pre-training on the AOJ dataset.

Both training strategies used a batch size of 256[4] per GPU[5] and a patience of 10 epochs. For training samples with sizes up to 100 K jets, the validation set was the same size as the training set. For larger samples, we used validation set sizes of 200 K jets for 1 M training samples and 500 K jets for 10 M training samples.

B. Performance metrics

We consider two complementary evaluation metrics that compare the 1D distribution of a high-level observable $\mathcal{O}$ between the generated jets and the target JetClass reference data:

- Kullback-Leibler divergence:

$$\mathrm{KL}^{\mathcal{O}}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) . \tag{1}$$

  The probability densities $P(x)$ and $Q(x)$ for the observable $\mathcal{O}$ are obtained from the normalized histogram bin counts and $\mathrm{KL}^{\mathcal{O}}(P||Q)$ is computed using `scipy.stats.entropy` [51]. Note that $\mathrm{KL}^{\mathcal{O}}(P||Q)$ is antisymmetric. In our results, we take P to be the generated distribution and Q to be the reference JetClass distribution.

- Wasserstein-1 distance:

$$W_1^{\mathcal{O}}(P,Q) = \min_{\pi\in\Pi} \sum_{x,y} |x-y|\ \pi(x,y) . \tag{2}$$

  Here $\Pi(P,Q)$ is the space of all joint 'coupling' distributions whose marginals are the observable $\mathcal{O}$ distributions P and Q. This quantity is computed with `scipy.stats.wasserstein_distance` [51]. Unlike the $\mathrm{KL}^{\mathcal{O}}(P||Q)$ divergence, $W_1^{\mathcal{O}}(P,Q)$ is a symmetric similarity metric based on optimal transport, i.e. the optimal cost plan that morphs P into Q.

C. Results

A comparison of the two training strategies –"fine-tuned" and "from scratch"– is presented in fig. 3 for QCD and in fig. 4 for top-quark jets. The rows show various high-level distributions, including jet p_T, jet mass, subjettiness ratios τ_{21} and τ_{32}, and particle multiplicity N (i.e. the number of constituents in a jet). The left-most column in each figure displays the distributions generated by the fine-tuned models, while the second column shows the distributions generated by the models trained from scratch. Each plot also includes the reference JetClass distribution (200 K reference jets) as a filled histogram and, for additional context, the AOJ distributions (dotted lines) in the left-most column.

Overall, we find that the pre-trained models are able to achieve a performance gain when using a fraction of the total fine-tuning data. This can be seen in the third and last columns of fig. 3 and fig. 4, where the fine-tuned models (blue lines) typically achieve better[6] metric scores than the models trained from scratch (red lines).

[4] For training samples of 100 jets, the batch size was reduced to 100.
[5] We used a single GPU to train the models with training sample sizes less than or equal to 100 K, and multiple GPUs for models trained on more data.

[6] However, there are instances where models trained from scratch perform equally well on a single metric within uncertainties. For example, this occurs with KL^{p_T} for training sample size $D = 10^3$. This behavior can be attributed to the definition of KL divergence, which penalizes high-probability regions in the generated data where the true probability distribution is low. Therefore, it is crucial to evaluate performance using multiple metrics to gain a comprehensive understanding.

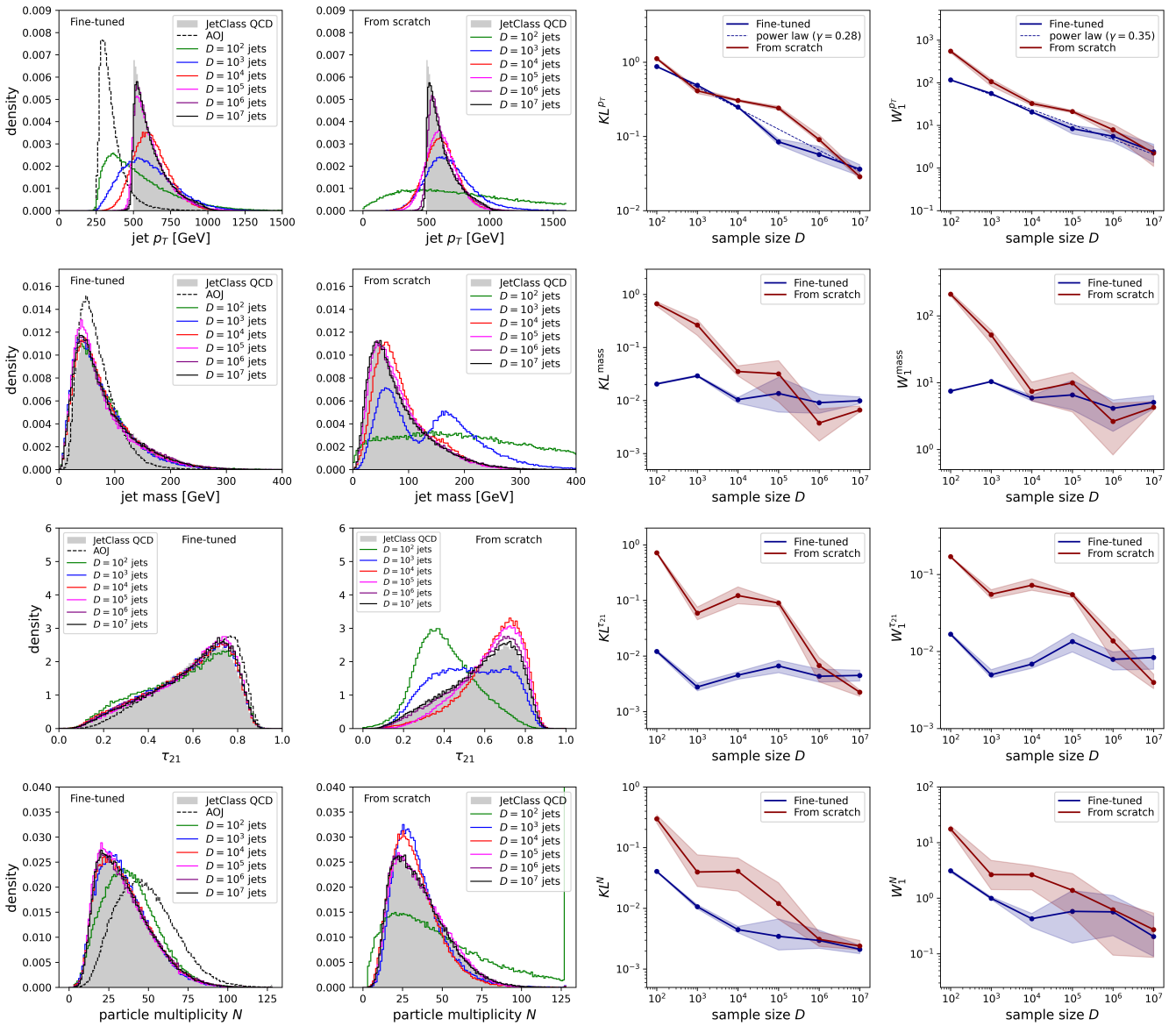

FIG. 3. A comparison of the generation quality of models trained on QCD jets from JetClass, for different training sample sizes D. The performance of a foundation model pre-trained on AOJ (far left) is compared to a model trained from scratch (center left). The generation quality is compared across several high level features of the jets. Two quantitative metrics, the Kullback-Leibler divergence (center right) and Wasserstein-1 distance (far right), are computed as a function of the training sample size D to compare how well the generated jets from each model matches the target distribution from JetClass. We also report the mean value and the envelopes over 5 trainings with different random seeds. When applicable, we also show power-law fits $\propto D^{-\gamma}$ to the metrics.

This is apparent for the KL divergence and Wasserstein-1 metrics across different training sample sizes below 1 M jets. However, as the available training data exceeds 1 M jets, we find that the fine-tuning begins to saturate, and the fine-tuned and from-scratch models perform similarly within uncertainties. In some features, the fine-tuned models even exhibit a slight decrease in performance compared to those trained entirely from scratch, indicating a diminishing return from pre-training at larger dataset scales. The reasons for this are unclear and would require further study. One possibility is that the AOJ tokenization is suboptimal for describing JetClass and this is exposed with enough JetClass training data. Another possibility is that the pre-training itself can harm model performance by locking in the weights to a suboptimal part of parameter space, a phenomenon that has been observed previously in the ML literature [52] and was termed "ossification".

We now examine in more detail the transfer learning from AOJ to JetClass for QCD jets, as shown in fig. 3.

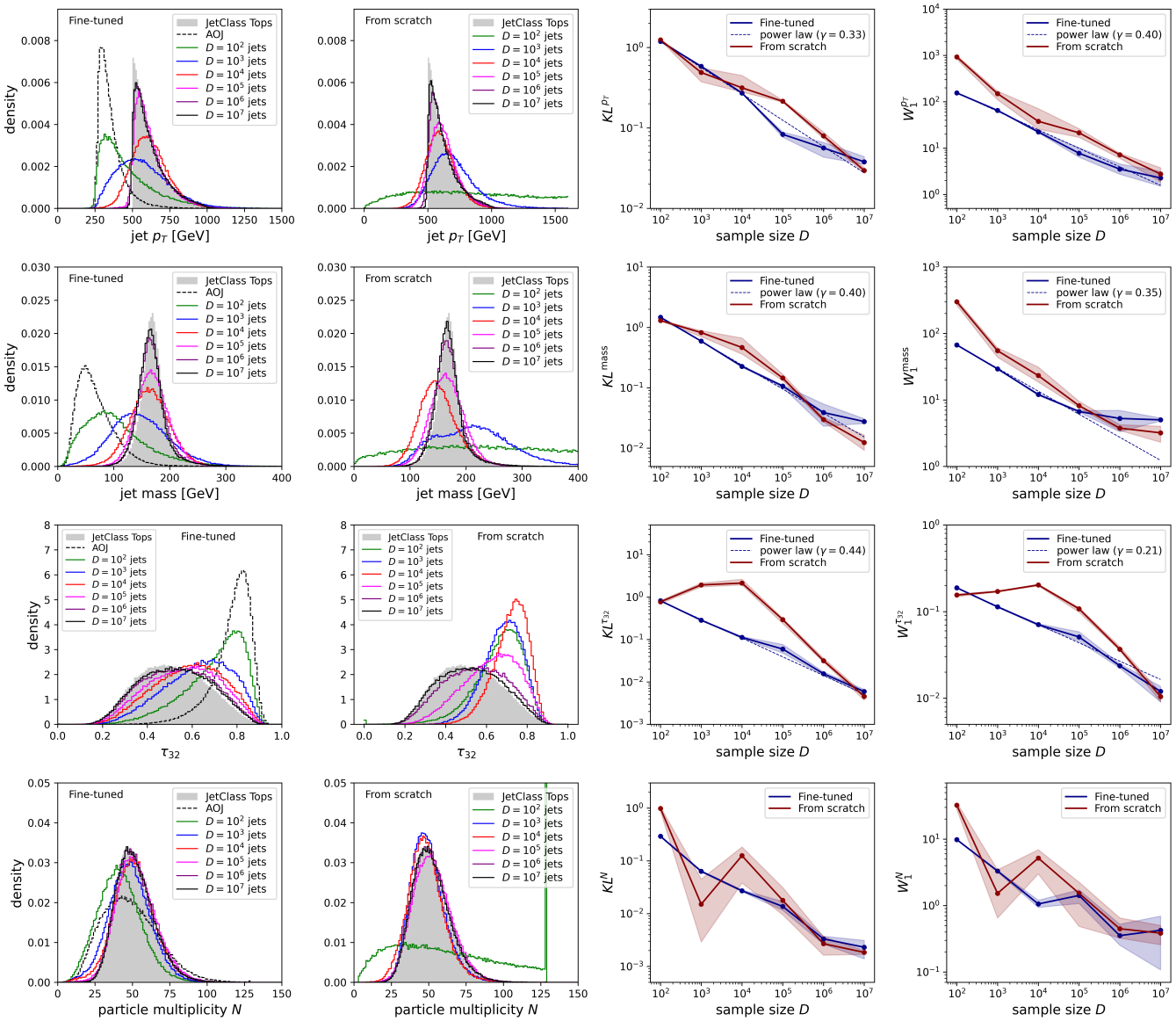

FIG. 4. A comparison of the generation quality of models trained on top-quark jets from JetClass, for different training sample sizes D. The performance of a foundation model pre-trained on AOJ (far left) is compared to a model trained from scratch (center left). The generation quality is compared across several high level features of the jets. Two quantitative metrics, the Kullback-Leibler divergence (center right) and Wasserstein-1 distance (far right), are computed as a function of the training sample size D to compare how well the generated jets from each model matches the target distribution from JetClass. We also report the mean value and the envelopes over 5 trainings with different random seeds. When applicable, we also show power-law fits $\propto D^{-\gamma}$ to the metrics.

The primary challenge in fine-tuning arises from the differences in the domains of the jet p_T distributions and particle multiplicities between the two datasets. JetClass imposes a narrower p_T window between 500 GeV and 1000 GeV and tends to have around ~ 30 particles per jet on average, while AOJ only applies a lower cut around 250 GeV and has around ~ 50 particles per jet on average. When fine-tuning, the pre-trained model must adapt to these differences and learn to generate jets with fewer particles and the new kinematic cuts. The results for the p_T distributions and particle multiplicities show how the fine-tuned model smoothly interpolates between the AOJ and JetClass as a function of the training sample size D. Furthermore, the generation quality improves with increasing training sample size, achieving almost optimal results with a reduced sample size of 100 K jets for the p_T distribution and 10 K for the multiplicities. In contrast, the randomly initialized models trained from scratch require a significantly larger dataset, approximately an order of magnitude more jets, to achieve optimal perfor-

mance. For the jet mass and the substructure observable τ_{21}, the fine-tuning performance saturates with a remarkably small number of training samples, resulting in an almost flat dependence of the metrics on D. Specifically, fine-tuning requires only about 1000 jets to reach optimal performance, whereas training from scratch requires significantly more data to achieve comparable results. This is not too surprising, as the transfer from AOJ to JetClass for these features does not require much effort, given the similarity in the shapes of these distributions.

Next, we examine the results in fig. 4, which correspond to top-quark jet generation. Since the AOJ dataset is completely dominated by QCD jets, the transfer learning task between AOJ and JetClass becomes more challenging due to the change in jet type. For instance, fine-tuning must map the low-mass Sudakov peak characteristic of QCD jets to the top-quark resonance peak at $m_t = 175$ GeV. Additionally, the model must transition from generating 1-prong QCD jets to 3-pronged top-quark jets, as reflected in the horizontal gap between the peaks in the τ_{32} substructure distributions. Despite these obstacles, we find that transfer learning consistently outperforms training from scratch when using smaller datasets. For example, the top-quark mass, represented by the mode in the jet mass distribution, is learned with good precision by the fine-tuned model using a training dataset of only 10 K jets. In contrast, the model trained from scratch requires an order of magnitude more data to achieve comparable performance. A similar trend is observed for the τ_{32} distribution, where transfer learning morphs the AOJ jets into JetClass top-quark jets, achieving competitive results with only 100 K training jets. In comparison, models trained from scratch fail to generate jets with a reliable substructure for training sample sizes below 1 M jets.

D. Power-law scaling

Finally, we observe that the performance metrics for the fine-tuned models, when not in the saturated regime, follow a power-law scaling of the form $D^{-\gamma}$ with respect to the training sample size D. Neural scaling laws, such as these, have attracted broad interest and have been analyzed extensively in various settings in recent years, see [53] for a nice overview of the ML literature and original references. In particular, they have been observed in the context of jet classification [54]. We find that the exponents for a wide range of high-level observables are consistently within the range $0.2 < \gamma < 0.45$ (see dashed lines in fig. 3 and 4). This scaling behavior holds for both the KL and Wasserstein-1 metrics and applies to both QCD and top-quark, suggesting a form of *universality*. The only exception to this scaling is the particle multiplicity distribution, which shows a more irregular dependence on the training sample size. Overall, this demonstrates that the generative performance scales predictably and uniformly with the training sample size across most high-level observables, enabling a smooth and consistent transfer learning between the AOJ and JetClass datasets.

In contrast, the randomly initialized models trained from scratch do not satisfy such universal scaling across most physical observables. The reason behind this may be that these operate in the small-data regime [55], where their performance is limited by the lack of sufficient training data. In this regime, models can only perform marginally better than random guessing, as each additional sample provides limited new information for learning. The breaking of scaling is particularly evident for the jet substructure features, such as τ_{21} for QCD and τ_{32} for tops, and to a lesser extent for the jet mass, where the performance metrics exhibit irregular and dataset-dependent scaling as the training sample size increases. Only after the model has seen sufficient training data, approximately around 1 M jets, does the training stabilize, resulting in reliable jet substructure and jet mass modeling. This discrepancy highlights the advantage of pre-training in achieving a more robust and systematic improvement in generative performance.

IV. CONCLUSION

We are releasing the AspenOpenJets (AOJ) dataset, which comprises 180 million jets from CMS open data, in a standard machine learning-friendly format. This dataset provides an excellent resource for exploring foundation models and other machine learning techniques that use large, unlabelled datasets for pre-training.

In addition, we present the first jet-based foundation model in high-energy physics trained on real collider data from the AOJ dataset. By using this rich dataset to pre-train a foundation model, we show that fine-tuning for specific tasks - such as generating jets of different types in different kinematic regions - yields significant performance improvements over models trained from scratch. In particular, the generative performance of our foundation model scales predictably with training sample size, facilitating effective transfer learning. This demonstrates the power of foundation models to provide improved performance even over significant domain shifts.

We hope that this study will encourage further engagement with open data from LHC experiments, and promote advances that bridge phenomenological studies and deployment in actual experiments.

ACKNOWLEDGEMENTS

This work was initiated at the Aspen Center for Physics, supported by National Science Foundation grant PHY-2210452. We would like to thank Alexander Mück for discussions. The research of MK and HR-G is supported by the Deutsche Forschungsgemeinschaft DFG under grant 396021762 – TRR 257: Particle physics phenomenology after the Higgs discovery. JB, AH and GK

Process	CMS Dataset Name	Cross Section (pb)
data	/JetHT/Run2016G-UL2016_MiniAODv2-v2/MINIAOD	-
	/JetHT/Run2016H-UL2016_MiniAODv2-v2/MINIAOD	-
QCD	/QCD_Pt_300to470_TuneCP5_13TeV_pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	7823
	/QCD_Pt_470to600_TuneCP5_13TeV_pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	648.2
	/QCD_Pt_600to800_TuneCP5_13TeV_pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	186.9
	/QCD_Pt_800to1000_TuneCP5_13TeV_pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	32.3
$t\bar{t}$	/TTToHadronic_TuneCP5_13TeV-powheg-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	378.5
	/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v1/MINIAODSIM	365.2
W + jets	/WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	315.6
	/WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	68.6
	/WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	34.9
Z+ jets	/ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	145.4
	/ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	34.0
	/ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	18.7
H→ $b\bar{b}$	/GluGluHToBB_Pt-200ToInf_M-125_TuneCP5_MINLO_13TeV-powheg-pythia8/RunIISummer20UL16MiniAODv2-106X_mcRun2_asymptotic_v17-v2/MINIAODSIM	0.27

TABLE II. A list of the datasets used to construct AOJ and the Monte Carlo samples used to estimate its composition.

are supported by the DFG under the German Excellence Initiative – EXC 2121 Quantum Universe – 390833306, and by PUNCH4NFDI – project number 460248186. DAF, IP, and DS are supported by DOE grant DOE-SC0010008. OA is supported by 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. LA is supported by the University of Bologna. Additionally, we acknowledge support from the Maxwell computational resources at Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany, and computing resources provided by RWTH Aachen University under project rwth0934. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award HEP-ERCAP0027491.

CODE

The code that was used to create AspenOpenJets from CMS open data can be found at https://github.com/OzAmram/AOJProcessing, and the code used for the OmniJet-α model and its training can be found at https://github.com/uhh-pd-ml/omnijet_alpha.

DATASET

The AspenOpenJets dataset can be found at http://doi.org/10.25592/uhhfdm.16505.

Appendix A: CMS Samples

MC samples provided by CMS were used to estimate the composition of the AOJ sample. CMS uses a variety of generators for the simulation different physics processes, including MADGRAPH5_aMC@NLO [43], PYTHIA [44], and POWHEG [56–58]. Each sample uses PYTHIA with the underlying event tune CP5 [59] to simulate the parton shower and hadronization. All samples use the NNLO NNPDF 3.1 parton distribution functions [60–62]. Some physics processes are split into several samples covering different kinematic regions. Each sample is normalized to the appropriate cross section.

To estimate the number of fully merged W, Z, top, and H→ $b\bar{b}$ jets the corresponding MC samples were used. A matching criteria was applied based on truth-level information from the generator to ensure the estimate reflected the number of fully-merged jets of each type. For each heavy resonance, generator-level information was used to check that all the quarks from the heavy resonance decay were within $\Delta R < 0.8$ of the reconstructed AK8 jet found by our selection. The remaining 99.5% of jets not estimated to be heavy resonance decays are assumed to be QCD. The resulting numbers should be considered rough estimates, as the MC modeling of these signals is imperfect, and the estimates do not incorporate corrections to the simulated trigger and reconstruction efficiency typically employed in physics analyses by CMS.

Table II lists the CMS data and MC samples used and their corresponding cross sections.

[1] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," (2022), arXiv:2108.07258 [cs.LG].
[2] T. B. Brown *et al.*, "Language models are few-shot learners," (2020), arXiv:2005.14165 [cs.CL].
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers

for Language Understanding," (2019), arXiv:1810.04805 [cs.CL].
[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," (2023), arXiv:2302.13971 [cs.CL].
[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," (2022), arXiv:2204.06125 [cs.CV].
[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," (2022), arXiv:2205.11487 [cs.CV].
[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," (2021), arXiv:2103.00020 [cs.CV].
[8] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, CoRR **abs/2102.05918** (2021), 2102.05918.
[9] T. Kishimoto, M. Morinaga, M. Saito, and J. Tanaka, in *37th Conference on Neural Information Processing Systems* (2023) arXiv:2312.06909 [hep-ex].
[10] H. Qu, C. Li, and S. Qian, (2022), arXiv:2202.03772 [hep-ph].
[11] L. Heinrich, T. Golling, M. Kagan, S. Klein, M. Leigh, M. Osadchy, and J. A. Raine, (2024), arXiv:2401.13537 [hep-ph].
[12] J. Birk, A. Hallin, and G. Kasieczka, Mach. Learn. Sci. Tech. **5**, 035031 (2024), arXiv:2403.05618 [hep-ph].
[13] P. Harris, M. Kagan, J. Krupa, B. Maier, and N. Woodward, "Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models," (2024), arXiv:2403.07066 [hep-ph].
[14] V. Mikuni and B. Nachman, (2024), arXiv:2404.16091 [hep-ph].
[15] CMS Collaboration, "Multijet primary dataset in AOD format from RunA of 2011 (/MultiJet/Run2012A-22Jan2013-v1/AOD)," (2022).
[16] CMS Collaboration, "Multijet primary dataset in AOD format from RunA of 2012 (/MultiJet/Run2012A-22Jan2013-v1/AOD)," (2022).
[17] CMS Collaboration, "JetHT primary dataset in MINIAOD format from RunG of 2016 (/JetHT/Run2016G-UL2016_MiniAODv2-v2/MINIAOD)," (2024).
[18] CMS Collaboration, "JetHT primary dataset in MINIAOD format from RunH of 2016 (/JetHT/Run2016H-UL2016_MiniAODv2-v2/MINIAOD)," (2024).
[19] ATLAS Collaboration, "DAOD_PHYSLITE format 2015-2016 Open Data for Research from the ATLAS experiment," (2024).
[20] A. Larkoski, S. Marzani, J. Thaler, A. Tripathee, and W. Xue, Phys. Rev. Lett. **119**, 132003 (2017), arXiv:1704.05066 [hep-ph].
[21] A. Tripathee, W. Xue, A. Larkoski, S. Marzani, and J. Thaler, Phys. Rev. D **96**, 074003 (2017), arXiv:1704.05842 [hep-ph].
[22] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, Phys. Rev. D **101**, 034009 (2020), arXiv:1908.08542 [hep-ph].
[23] H. Qu, C. Li, and S. Qian, "JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics," (2022).
[24] S. Chatrchyan *et al.* (CMS), JINST **3**, S08004 (2008).
[25] CMS Collaboration, "Jet primary dataset in AOD format from RunB of 2010 (/Jet/Run2010B-Apr21ReReco-v1/AOD)," (2014).
[26] G. Petrucciani, A. Rizzi, and C. Vuosalo (CMS), J. Phys. Conf. Ser. **664**, 7 (2015), arXiv:1702.04685 [physics.ins-det].
[27] M. Peruzzi, G. Petrucciani, A. Rizzi, and for the CMS Collaboration, Journal of Physics: Conference Series **1525**, 012038 (2020).
[28] A. M. Sirunyan *et al.* (CMS), JINST **15**, P10017 (2020), arXiv:2006.10165 [hep-ex].
[29] V. Khachatryan *et al.* (CMS), JINST **12**, P01020 (2017), arXiv:1609.02366 [physics.ins-det].
[30] A. M. Sirunyan *et al.* (CMS), JINST **12**, P10003 (2017), arXiv:1706.04965 [physics.ins-det].
[31] M. Cacciari, G. P. Salam, and G. Soyez, JHEP **04**, 063 (2008), arXiv:0802.1189 [hep-ex].
[32] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. **C72**, 1896 (2012), arXiv:1111.6097 [hep-ph].
[33] A. M. Sirunyan *et al.* (CMS), JINST **15**, P09018 (2020), arXiv:2003.00503 [hep-ex].
[34] D. Bertolini, P. Harris, M. Low, and N. Tran, JHEP **10**, 059 (2014), arXiv:1407.6013 [hep-ph].
[35] V. Khachatryan *et al.* (CMS), JINST **12**, P02014 (2017), arXiv:1607.03663 [hep-ex].
[36] C. Collaboration, "PFNano producer tool for CMS 2016," (2024), software available from `https://opendata.cern.ch/record/12504`.
[37] CMS Collaboration, "Monte Carlo simulations of 2016 data," (2024).
[38] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, JHEP **05**, 146 (2014), arXiv:1402.2657 [hep-ph].
[39] S. Navas *et al.* (Particle Data Group), Phys. Rev. D **110**, 030001 (2024).
[40] J. Thaler and K. Van Tilburg, JHEP **03**, 015 (2011), arXiv:1011.2268 [hep-ph].
[41] H. Qu and L. Gouskos, Phys. Rev. D **101**, 056019 (2020), arXiv:1902.08570 [hep-ph].
[42] (2020).
[43] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, JHEP **07**, 079 (2014), arXiv:1405.0301 [hep-ph].
[44] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, Comput. Phys. Commun. **191**, 159 (2015), arXiv:1410.3012 [hep-ph].
[45] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaitre, A. Mertens, and M. Selvaggi (DELPHES 3), JHEP **02**, 057 (2014), arXiv:1307.6346 [hep-ex].
[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever (2018).
[47] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," (2018), arXiv:1711.00937 [cs.LG].
[48] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," (2022), arXiv:2106.08254 [cs.CV].

[49] M. Huh, B. Cheung, P. Agrawal, and P. Isola, "Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks," (2023), arXiv:2305.08842 [cs.LG].

[50] L. Wright, "Ranger - a synergistic optimizer." https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer (2019).

[51] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, Nature Methods **17**, 261 (2020).

[52] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, CoRR **abs/2102.01293** (2021), 2102.01293.

[53] P. Villalobos, "Scaling laws literature review," (2023), accessed: 2024-12-13.

[54] J. Batson and Y. Kahn, (2023), arXiv:2312.02264 [hep-ph].

[55] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, arXiv preprint arXiv:1712.00409 (2017).

[56] P. Nason, JHEP **11**, 040 (2004), arXiv:hep-ph/0409146 [hep-ph].

[57] S. Frixione, P. Nason, and G. Ridolfi, JHEP **09**, 126 (2007), arXiv:0707.3088 [hep-ph].

[58] S. Alioli, P. Nason, C. Oleari, and E. Re, JHEP **06**, 043 (2010), arXiv:1002.2581 [hep-ph].

[59] A. M. Sirunyan *et al.* (CMS), Eur. Phys. J. C **80**, 4 (2020), arXiv:1903.12179 [hep-ex].

[60] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, Nucl. Phys. B **838**, 136 (2010), arXiv:1002.4407 [hep-ph].

[61] R. D. Ball *et al.* (NNPDF), JHEP **04**, 040 (2015), arXiv:1410.8849 [hep-ph].

[62] R. D. Ball *et al.* (NNPDF), Eur. Phys. J. C **77**, 663 (2017), arXiv:1706.00428 [hep-ph].