Supplementary Material

1. Energy loss module physics formalism

The initial virtuality of the partons will have a maximum limit set by the preset distribution. These will then be introduced into the MATTER event generator. In MATTER, a single hard parton created at a point r with a forward light-cone momentum $p^+ = (p^0 + \hat{n} \cdot \vec{p}/\sqrt{2})$ where $\hat{n} = \vec{p}/|\vec{p}|$ starts a virtuality-ordered shower.
To ascertain the real virtuality ($t = Q^2$) of the given parton, one may sample a Sudakov form factor,

$$\Delta(t,t_0) = \exp\left[-\int_{t_0}^{t}\frac{dQ^2}{Q^2}\frac{\alpha_s(Q^2)}{2\pi}\int_{t_0/t}^{1-t_0/t} dzP(z)\right.$$
$$\left.\times\left\{1+\int_0^{\zeta+MAX^+} d\zeta^+ \frac{\hat{q}(r+\zeta)}{Q^2(1-z)}\Phi(Q^2,p^+,\zeta^+)\right\}\right], \tag{1}$$

where Φ represents a sum over phase factors that depends on ζ^+, p^+, and Q. The transport coefficient $\hat{q}$ is evaluated at the location of scattering $\vec{r} + \hat{n}\zeta^+$, $P(z)$ is the vacuum splitting function, and $\zeta_M AX^+$ is the maximum length ($1.2\tau_f^+$), which is used to sample the actual splitting time of the given parton with τ_f^+ as the mean light-cone formation time $\tau_f^+ = 2p^+/Q^2$ [10]. After determining Q^2, z can be sampled using the splitting function $P(z)$. The transverse momentum of the created daughter pair can be estimated using the difference in invariant mass between the parent and daughters. This method is repeated until a given parton's Q^2 reaches a specific value for Q_0^2.
Below Q_0^2 the jet might be better characterized by another energy loss module such as LBT, which can evolve according to the linear Boltzmann equation. Q_0 is the virtuality separation scale. For our dataset, the medium-induced gluon spectrum

$$\Gamma^{inel} = \int dx dk_\perp^2 \frac{dN_g}{dx dk_\perp^2 dt}, \tag{2}$$

where the differential spectrum of the radiated gluon is taken from the higher-twist energy loss formalism [24, 32, 40]:

$$\frac{dN_g}{dx dk_\perp^2 dt} = \frac{2\alpha_s C_A \hat{q} P(x) k_\perp^4}{\pi(k_\perp^2 + x^2m^2)^4}\sin^2\left(\frac{t-t_i}{2\tau_f}\right), \tag{3}$$

where x and $k_\perp$ are the fractional energy and transverse momentum of the emitted gluon with respect to its parent parton, α_s is the strong coupling constant, $C_A = N_c$ is the gluon color factor, $P(x)$ is the splitting function, $\hat{q}$ is the transport coefficient, t_i denotes the production time of the given parton, and $\tau_f = 2Ex(1-x)/k_\perp^2 + x^2m^2$ is the formation time of the radiated gluon with E and m as the parton energy and mass, respectively. With these scattering rates, the Monte Carlo method is applied to determine whether scattering happens within a given time step. In this work, we develop an ML model to determine the energy loss model for different values of Q_0 and α_s.

2. Heavy ion collisions

In this section, we show a visualization that depicts the multi-stage approach that is leveraged in the JETSCAPE for jet evolution in Figure 8.

3. Sample events

In this appendix, we provide sample events for configurations one and two of the dataset, depicted in Figures 10 and 11.

4. Calculating accuracy for VGG16 training for 50 epoch Config. #9 - Test Data

One of the methods for assessing classification models is accuracy, which is simply the percentage of correct predictions. For binary classification, accuracy can also be calculated in terms of positives and negatives as in equation (4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

where *TP = True Positives, TN = True Negatives, FP = False Positives,* and *FN = False Negatives*. Table 3 shows an example confusion matrix (VGG16 Model – 50 epoch - Config. #9 – Test data) to calculate model's accuracy. The accuracy is 0.9429, or 94.29% (94 out of 100 instances yielded correct predictions) regarding equation 4.That indicates that our energy loss module classifier is very effective in detecting between *Matter* and *Matter-LBT*.

Table 3. Confusion Matrix for VGG16 Model – 50 epoch - Config. #9 – Test data

	Predicted	
	MATTER	MATTER-LBT
MATTER	TP: 56192	FP: 3039
MATTER-LBT	FN: 3808	TN: 56961

5. VGG16 training for 30 epochs

In this appendix, we provide the detailed analysis for VGG16 traning for 30 epochs. Table 4 demonstrates the loss and accuracy diagrams and figure 9 demonstrates the accuracy for nine configurations.

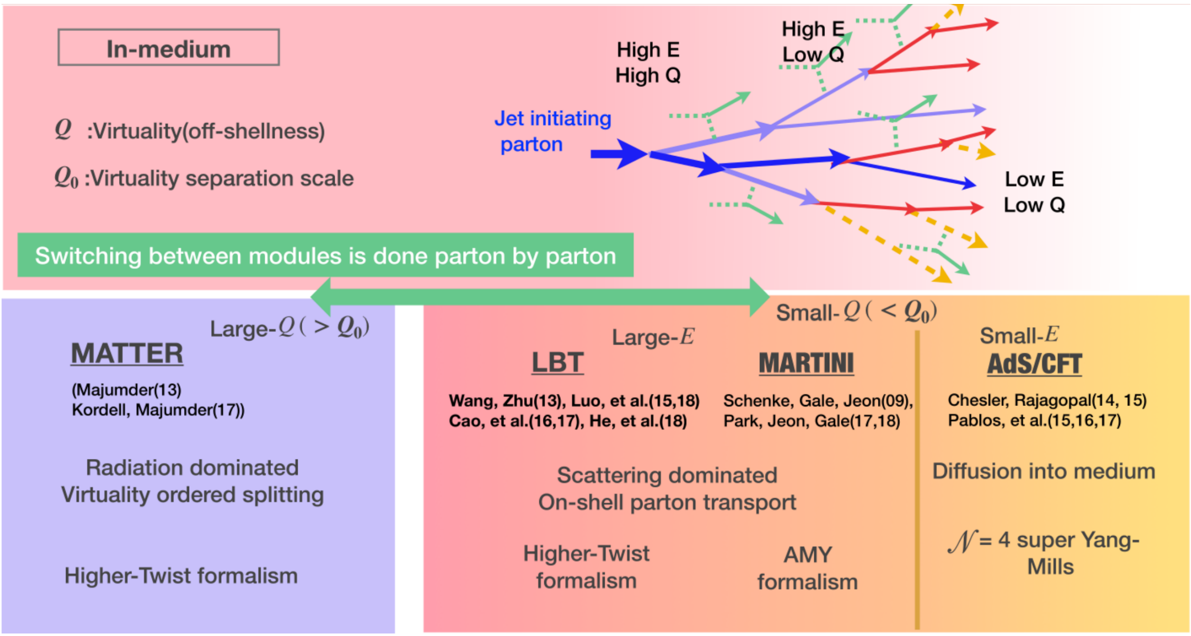

In-medium
Q :Virtuality(off-shellness)
Q_0 :Virtuality separation scale
Switching between modules is done parton by parton
Jet initiating parton
High E High Q
High E Low Q
Low E Low Q
Large-Q (> Q_0)
MATTER
(Majumder(13) Kordell, Majumder(17))
Radiation dominated Virtuality ordered splitting
Higher-Twist formalism
Small-Q (< Q_0)
Large-E
LBT
Wang, Zhu(13), Luo, et al.(15,18) Cao, et al.(16,17), He, et al.(18)
MARTINI
Schenke, Gale, Jeon(09), Park, Jeon, Gale(17,18)
Small-E
AdS/CFT
Chesler, Rajagopal(14, 15) Pablos, et al.(15,16,17)
Scattering dominated On-shell parton transport
Higher-Twist formalism
AMY formalism
Diffusion into medium
$\mathcal{N}$ = 4 super Yang-Mills

Figure 8. Multi-stage approach in heavy-ion collisions, credit to Y. Tachibana et. al. from JETSCAPE collaboration..

Table 4. VGG16 model with 30 epochs: accuracy.

	Accuracy (%)		
	Train	Validation	Test
Config No. 1	89.395	89.4242	89.1383
Config No. 2	91.031	91.0596	91.5408
Config No. 3	84.5407	84.6833	84.4558
Config No. 4	76.0095	76.1054	75.9908
Config No. 5	91.7856	91.8829	91.6892
Config No. 6	94.367	94.3083	94.3483
Config No. 7	86.5311	86.41	86.2825
Config No. 8	93.029	93.0608	93.0133
Config No. 9	94.1714	94.1717	94.0925

Table 5. VGG16 trained models for 50 epochs early stopping and their converged accuracy

Configuration No.	2	3	4	5	7	8
Accuracy (%)	92	93	90	93	89	93

6. Early stopping on VGG15 models

To prevent overfitting early stopping techniques has been applied on the training models. Table 5 shows a detailed accuracy report on each model when it confronted early stopping on VGG16 for 50 epochs.

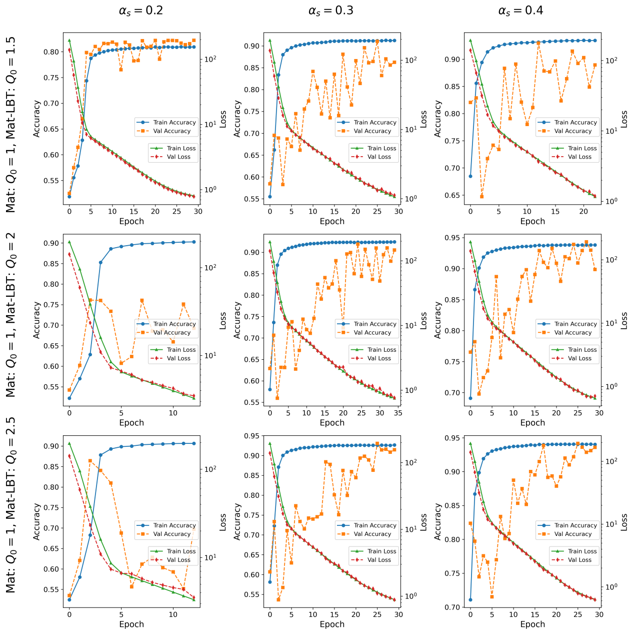
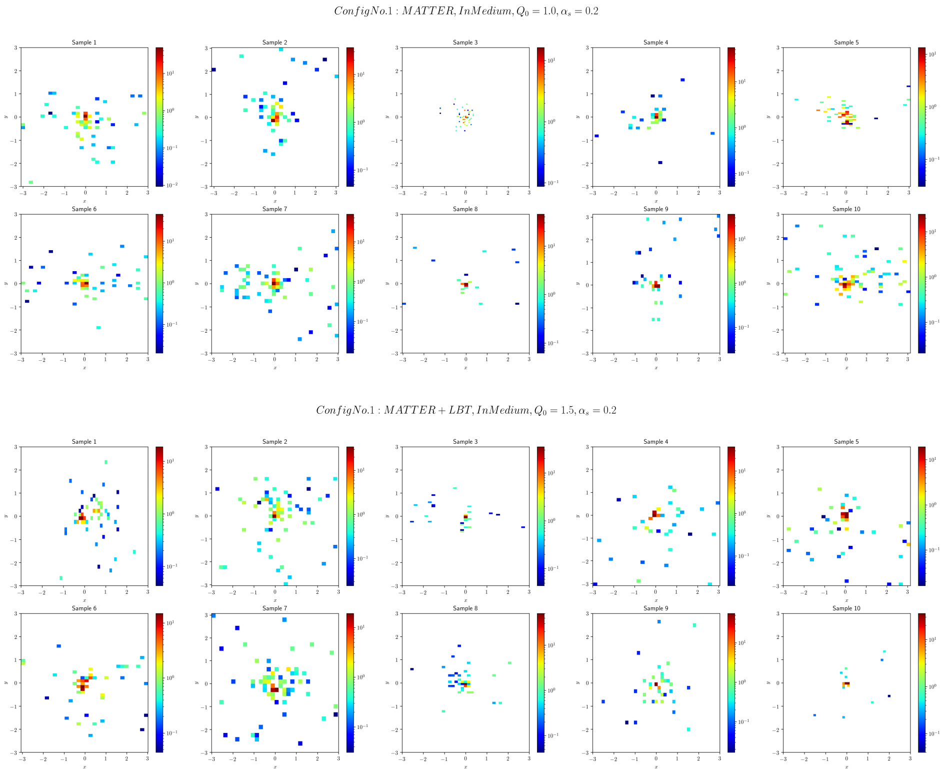


Figure 9. VGG16 model with 30 epochs: accuracy & loss diagrams.

7. Accuracy central tendency and variation analysis of machine learning models

In the pursuit of evaluating the efficacy and applicability of the ML-JET dataset, a series of experiments were conducted employing diverse machine learning methodologies.



$ConfigNo.1 : MATTER, InMedium, Q_0 = 1.0, \alpha_s = 0.2$

$ConfigNo.1 : MATTER + LBT, InMedium, Q_0 = 1.5, \alpha_s = 0.2$

Figure 10. Dataset: Sample events: Config No. 1 Matter and Matter-LBT.

These encompassed logistic regression, decision trees, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) including its linear variant (Linear SVC), and Random Forest, each deployed with various architectures and configurations. Training these models on the ML-JET dataset, we gauged their performance against a held-out test set.

Figure 5 illustrates the binary classification accuracy along with error bars for five distinct machine learning models trained over 5-fold cross-validation and employing four variations in dataset size ranging from 1K to 1000K instances. Our findings underscore the ML-JET dataset's proficiency, particularly in logistic regression models for tasks pertaining to energy loss module classification. These models achieved an average accuracy of approximately 87%, surpassing the performance of other models. However, it's noteworthy that the accuracy of logistic regression models plateaued at around 87% even with an increase in dataset size from 10^5 to 10^6, prompting consideration for alternative approaches within deep learning paradigms.

Linear SVC, Random Forest, KNN, and Decision Tree techniques followed in rankings from 2 to 5 respectively, in terms of their accuracy performance. Similar to logistic regression, Linear SVC exhibited a plateauing trend in accuracy, albeit at around 80% on average. Random Forest displayed a linear increase in accuracy with the expansion of the dataset size. However, extrapolating this trend suggests an immense dataset size requirement of 10^{10} instances to merely attain logistic regression accuracy levels with a dataset size of 10^6. KNN and Random Forest exhibited analogous accuracy trends, showing improvements between 10^3 to 10^4 instances, with marginal gains thereafter, boasting approximately 2-3% better performance.

8. Analysis of point cloud models

Upon scrutinizing the limitations of contemporary machine learning models in terms of computational capacity and accuracy, our focus shifted towards exploring cutting-edge deep neural network methodologies. Specifically, we delved into training PointNet [29] models for addressing the energy loss binary classification problem, employing various settings and configurations.

Figure 7a presents a comprehensive overview of the binary classification accuracy along with error bars for five distinct machine learning models, trained over 10 folds, 32 epochs, and with dataset sizes ranging from 1K to 1000K instances. The results obtained are highly encouraging. Notably, a linear correlation is observed between dataset size and average accuracy. Furthermore, as the dataset size in-

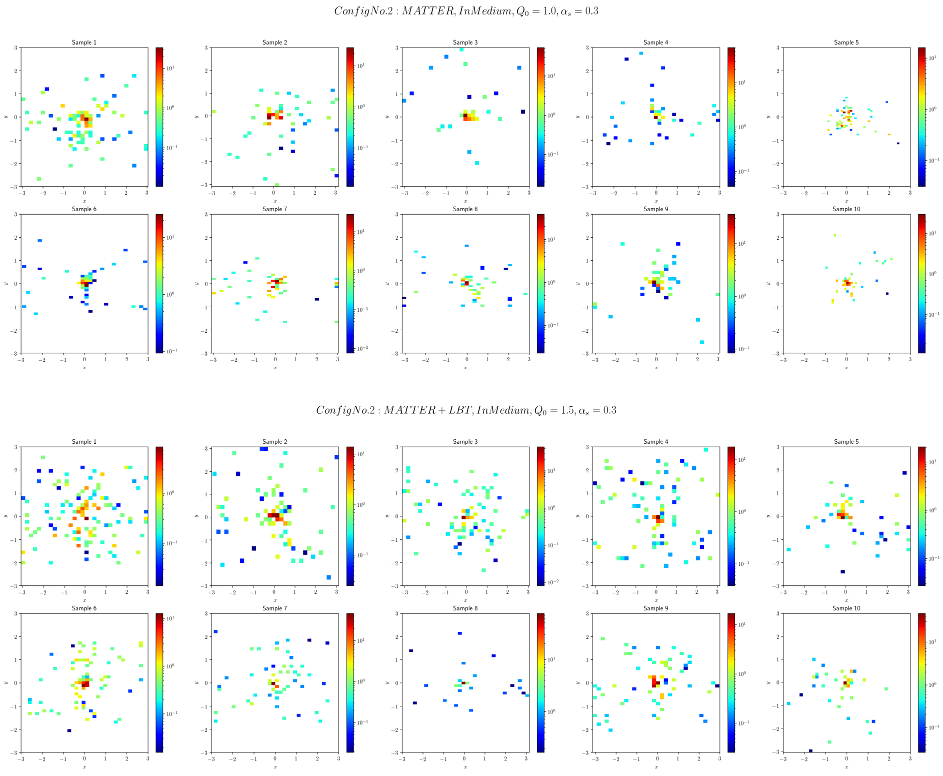


Figure 11. dataset: Sample events: Config No. 2 Matter and Matter-LBT.

creases, the standard deviation of accuracy diminishes, indicating improved stability in accuracy metrics. Notably, point clouds achieve an average accuracy of approximately 88% with a dataset size of 10^5. Remarkably, this outperforms logistic regression on a dataset size of 10^6, showcasing the consistent and linear progress achieved by PointNet models.

Additionally, Figure 7b illustrates the trajectory of training loss across epochs, demonstrating a consistent decrease, indicating effective learning from the training data. Conversely, the validation loss exhibits an initial decrease but later manifests fluctuations, suggestive of potential overfitting as the training progresses. Towards the latter stages of training, a slight increase in validation loss further corroborates the presence of overfitting tendencies.

The training accuracy steadily ascends with each epoch, as anticipated due to the model's learning process. However, the validation accuracy showcases a plateauing trend after a certain epoch, indicating limited improvement in performance on unseen data beyond a certain point.

The widening chasm between training and validation loss serves as a telltale sign of overfitting, wherein the model excels on the training data but struggles to generalize to unseen instances. Despite these challenges, the final validation accuracy hovers around 86-87%, a commendable achievement within the realm of heavy ion physics and its specific requirements.

9. MNIST Net accuracy analysis

After training the MNIST model for 50 epochs, it results in 82.23% average accuracy on the test data over all nine configurations as shown in Figure 3 and Table 2. One can try tweaking this model with different settings to get a better score. An obvious tweak is increasing the epochs, which improves accuracy at the expense of time, although modest improvement is expected given the sophisticated features of the energy loss module.

10. Deep models precision, recall, and F1-score analysis

In the realm of machine learning and statistical analysis, precision, recall, and F1-score are fundamental metrics used to evaluate the performance of classification models. Precision refers to the accuracy of positive predictions made by the model, measuring the proportion of true positive predictions among all positive predictions. Recall, on the other hand, assesses the model's ability to identify all relevant instances, representing the proportion of true positive predictions among all actual positive instances in the dataset.

Table 6. MNIST Net model evaluation: precision, recall, F1-score

Config No.	Precision (%)		Recall (%)		F1-Score (%)	
	MATTER	LBT	MATTER	LBT	MATTER	LBT
1	87	92	92	86	89	89
2	95	89	88	95	91	92
3	92	79	75	94	83	86
4	73	79	82	70	76	76
5	95	89	88	95	91	92
6	93	96	96	93	94	94
7	82	92	93	79	87	85
8	94	92	92	94	93	93
9	95	93	93	95	94	94

F1-score, often considered the harmonic mean of precision and recall, offers a balanced measure that combines both metrics into a single value, providing a comprehensive assessment of the model's predictive performance.

For the MNIST Net model, we notice variations in performance metrics across different configurations in Table 6. Generally, the MATTER configuration tends to exhibit higher precision compared to the LBT configuration, indicating its ability to classify positive instances more accurately. However, the LBT configuration demonstrates higher recall values, suggesting its effectiveness in capturing a higher proportion of actual positive instances. This trade-off between precision and recall is reflected in the F1-score, where both configurations achieve similar scores, balancing the precision-recall trade-off. Notably, Config No. 6 stands out with exceptionally high precision, recall, and F1-score values, indicating its superior performance across both configurations.

Similarly, for the VGG16 Net model, we observe variations in performance metrics across different configurations in Table 7. Like the MNIST Net model, the MATTER configuration tends to exhibit higher precision, while the LBT configuration demonstrates higher recall values. Again, the F1-score values indicate a balance between precision and recall for both configurations. Interestingly, Config No. 6 appears to perform exceptionally well across both configurations, achieving high precision, recall, and F1-score values. These observations suggest that certain configurations may be more suitable for specific tasks within the energy loss module binary classification problem.

In Figure 12, we demonstrated the PointNet's mean error bar across dataset sizes for precision, recall, and F1-score. It exhibits improving trends, with precision ranging from 0.55 to 0.8121, recall from 0.635 to 0.81005, and F1-score from 0.55 to 0.8121. Smaller datasets tend to have larger error bars, with precision showing slightly higher variability than recall or F1-score. However, as dataset size increases, error bars decrease, indicating more precise performance estimates.

Table 7. VGG16 Net model for 50 epochs evaluation: precision, recall, F1-score.

Config No.	Precision (%)		Recall (%)		F1-Score (%)	
	MATTER	LBT	MATTER	LBT	MATTER	LBT
1	87	92	92	86	89	89
2	95	89	88	95	91	92
3	92	79	75	94	83	86
4	73	79	82	70	76	76
5	95	89	88	95	91	92
6	93	96	96	93	94	94
7	82	92	93	79	87	85
8	94	92	92	94	93	93
9	95	93	93	95	94	94

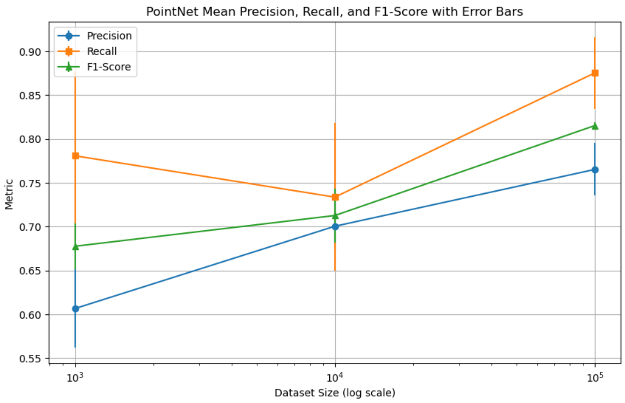


Figure 12. PointNet models precision, recall, and F1-score mean and error bar.