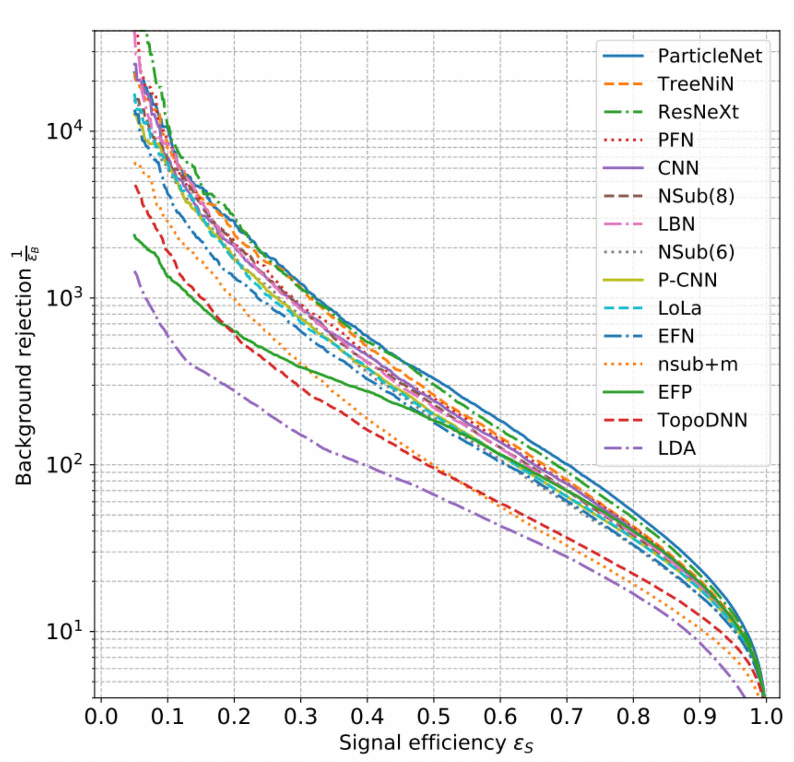
Using Generative AI to Explore the Limits of Jet Tagging

Nishank Gite[1,2], Benjamin Nachman[2], Vinicius Mikuni[2]

ABSTRACT

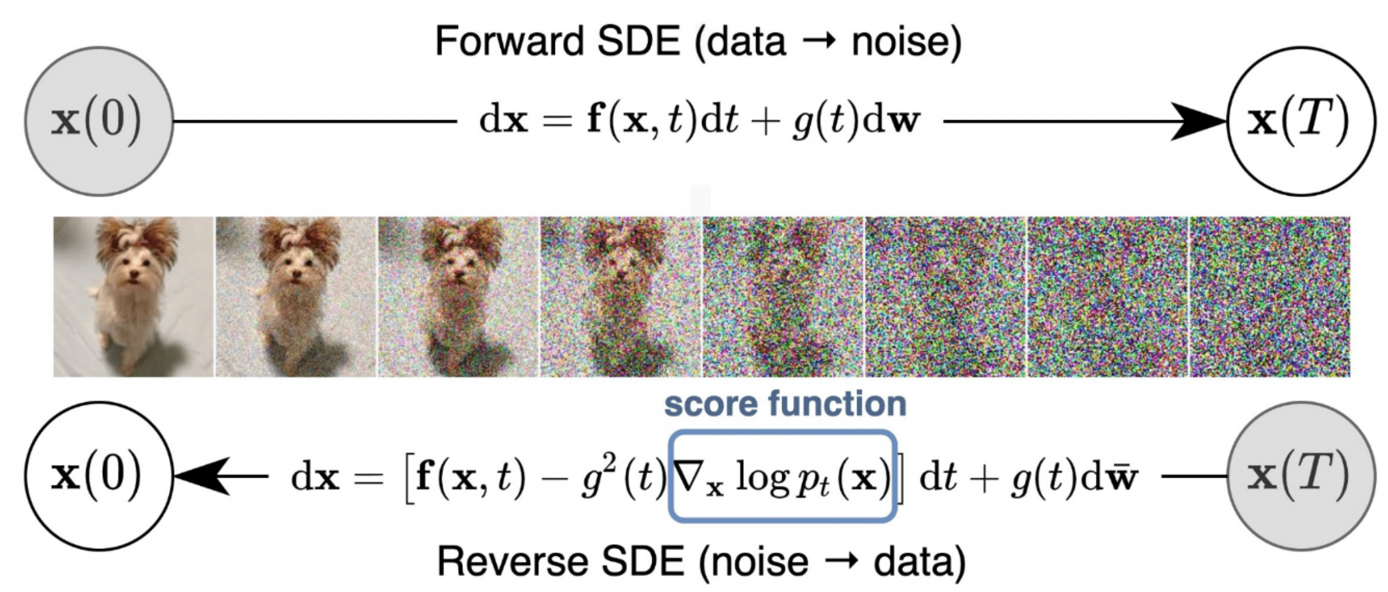
The precise identification of jets originating from high-energy quarks and gluons is paramount for advancing our understanding of fundamental particles and forces. This study introduces a novel deep learning framework designed to probe the limits of jet classifier models by using generative AI. State-of-the-art generative models called diffusion neural networks are used to create synthetic jet data where we simultaneously estimate the probability density by solving a differential equation. The likelihood ratio built from the probability density is the theoretical optimal classifier. Our research goal is to explore how close state-of-the-art classifier models are to this bound. We find that a state of the art transformer model performs very well, noting increases in true positive rates and decreases in false positive rates, but there is still a gap with respect to the optimal classifier.

MOTIVATIONS

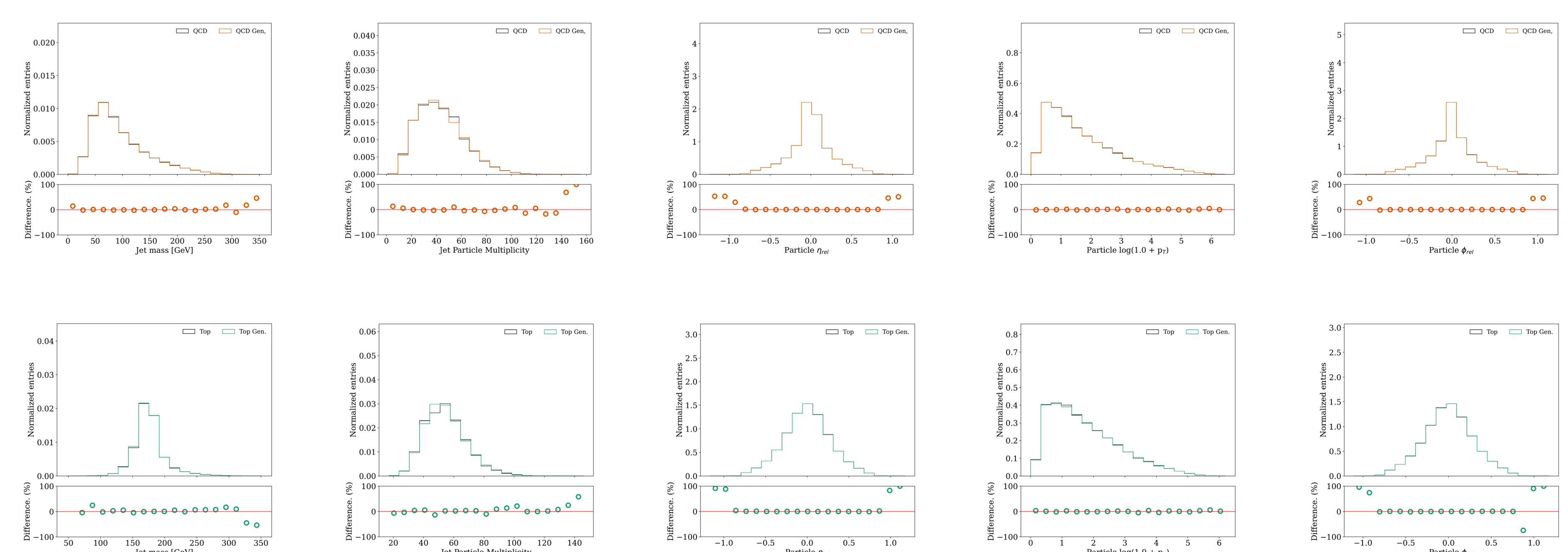


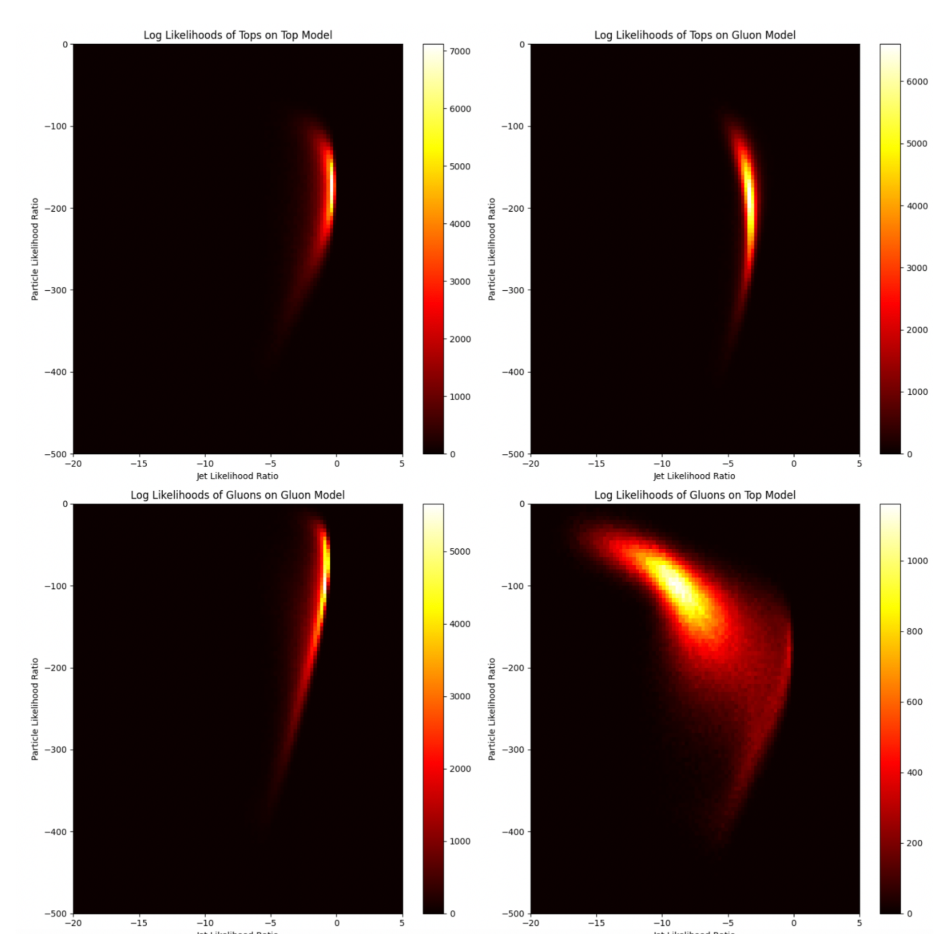
The image shows the development of hadronic jet tagging through comparison of various machine learning models through their ROC curves indicating their ability to distinguish between top quarks and gluons. Accurate identification of the origin of jets is crucial for particle physics experiments, as it helps in the search for new phenomena and precision measurements of standard model parameters. Previous studies, as depicted by the various models in the table, have made significant advancements in this field, however, the goal of this study is not just to add another incremental improvement but to conceptualize the "best" possible classifier. By creating a surrogate model that can generate a synthetic likelihood ratio, this study aims to understand the theoretical maximum performance a classifier could achieve.

METHODS



Diffusion models are a class of generative models that works in two directions; the forward stochastic differential equation adds noise over time T, mimicking a gradual increase in entropy and transforming the initial data into pure noise. The reverse SDE does the opposite; it starts with the noise and applies a log-likelihood score function to guide the noise back to a plausible and coherent data structure. Training the model to understand this transition, the reverse process can be used to generate new synthetic samples of jet data by sampling from the noise distribution and then 'denoising' to create data samples that are indistinguishable from real gluon and top jet events. Given the challenges in computing the likelihood ratio directly from raw data, we leveraged this model, applying the "likelihood ratio trick." Through learning the score of the probability density, we perform a monotonic transformation on the model output to approximate the probability density of observing some event given the origin jet as either top or gluon, then using this we determined the likelihood ratio, serving as a proxy for the optimal classifier performance.



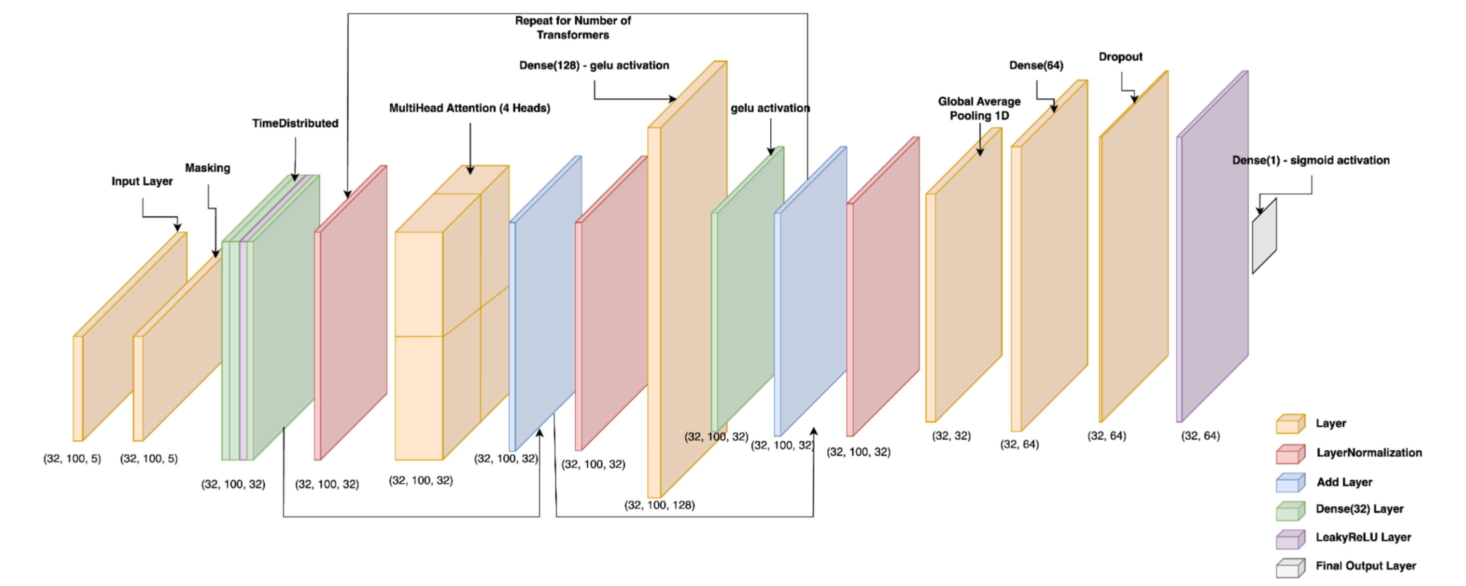


$$L(x) = \frac{p(x|\theta_1)}{p(x|\theta_0)}$$

$$\log L(x) = \log p(x|\theta_0) - \log p(x|\theta_1)$$

Our first set of findings validate our diffusion model's capability to generate synthetic data. By analyzing residuals, we determined that the generated jet data features closely mirrors the statistical properties of real jet data. This visually confirms that our approximated score function must strongly resemble the true score function for the data, with an error bound of .0001, which allows us to extend this formalism to the likelihoods. The Heatmaps displayed represent the log-likelihood ratios obtained from a diffusion model trained to differentiate between top jets and gluon jets. When evaluating gluon data with a gluon model and top data with a top model, we observe a peak at or near zero in the log-likelihood ratio with probabilities approaching one. Conversely, when top data is assessed with a gluon model and vice versa, the peaks shift away from zero, becoming more negative. This indicates a lower probability, moving towards zero. The sharpness of the peak and its location on the likelihood axis tells us about the model's confidence. Peaks close to zero on the correct model signify strong agreement between the model's output and the actual data, while peaks far from zero on the incorrect model affirm that the model is effectively identifying inconsistencies.

ARCHITECTURE

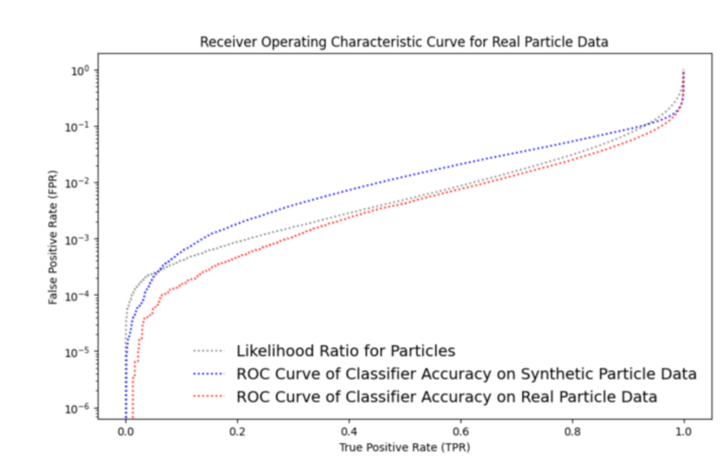


The architecture is designed to handle the high-dimensional and structured data associated with hadronic jet tagging due to several key features:

Masking Layer: The Masking layer ensures that any irrelevant or padding data does not affect the model, which is crucial for dealing with variable-length particle jets.
TimeDistributed Layers: These layers apply a layer to every temporal slice of an input, as each particle's data needs to be processed independently before being combined, mirroring the physical process of particles in a jet contributing independently to the jet's overall characteristics.
LeakyReLU Layers: By having some neurons remain slightly active, it can help in learning fine details of non-linear and complex patterns, allowing the model to continue learning even if some nodes start outputting negative values.
Add Layer: Allows layers to "skip" if they do not contribute towards reducing the loss, which helps avoid the Vanishing Gradient Problem.

MultiHead Attention: The model uses MultiHead Attention with 4 heads, allowing the network to focus on different parts of the particle sequence simultaneously, enabling the model to capture complex relationships between the particles in a jet.
Layer Normalization and GELU Activation: Normalization stabilizes learning and GELU allows the model to capture non-linear relationships, which are expected due to the complex nature of jet data.
Global Average Pooling 1D: Consolidates information across all particles using global average pooling, reducing the feature dimension while retaining essential information.
Dense and Output Layers: Combined with dropout this prevents overfitting, ending in a single-node output layer with a sigmoid activation function for binary classification.

RESULTS



As shown, this is the **preliminary** ROC curve of the classifier trained on synthetic particle data, and we see that the blue curve closely approaches the curve derived from the likelihood ratios. However, it has not fully converged, indicating that while the model captures the essence of the data, refinement is needed to enhance accuracy. We concentrate on particle data in this analysis, which inherently encompasses broader jet characteristics within its feature set. Although these are strong results, we can improve performance through various methods. By investigating the impact of adding more layers to the neural network we can potentially capture more complex relationships in the data. Exploring different architectures (e.g., convolutional layers for spatial patterns or recurrent layers for sequential data) can also offer significant improvements. We can also adjust hyperparameters such as learning rate, batch size, or the number of neurons in each layer and utilize techniques like grid search or Bayesian optimization to systematically find the best hyperparameters settings. Finally, exploiting ensemble methods that combine the predictions from multiple models using various methods like Random Forests or Gradient Boosting can improve overall performance.

ACKNOWLEDGEMENTS

[1] University of California, Berkeley
[2] Lawrence Berkeley National Laboratory
[3] AI Summer. Everything you need to know about Diffusion Models in Deep Learning. 2022.
[4] G. Kasieczka, T. Plehn, A. Butter, K. Cranmer, D. Debnath, B. M. Dillon, M. Fairbairn,D. A. Faroughy, W. Fedorko, C. Gay, et al. (2019), arXiv:1902.09914.