# Autoencoders for Symbolic Distillation of Black-Boxes

**ML4Jets 2026** | September 16, 2026 | Vienna

Alan Gu
Independent Researcher & High Schooler

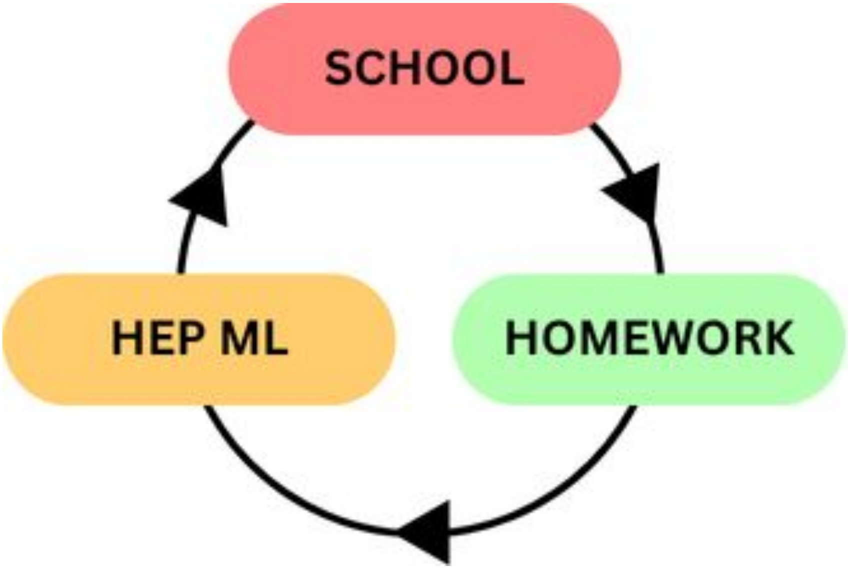
alangu612@gmail.com

# An October of Tabular Data

**Building shallow neural networks for the 2014 ATLAS Higgs Machine Learning Challenge on Kaggle**

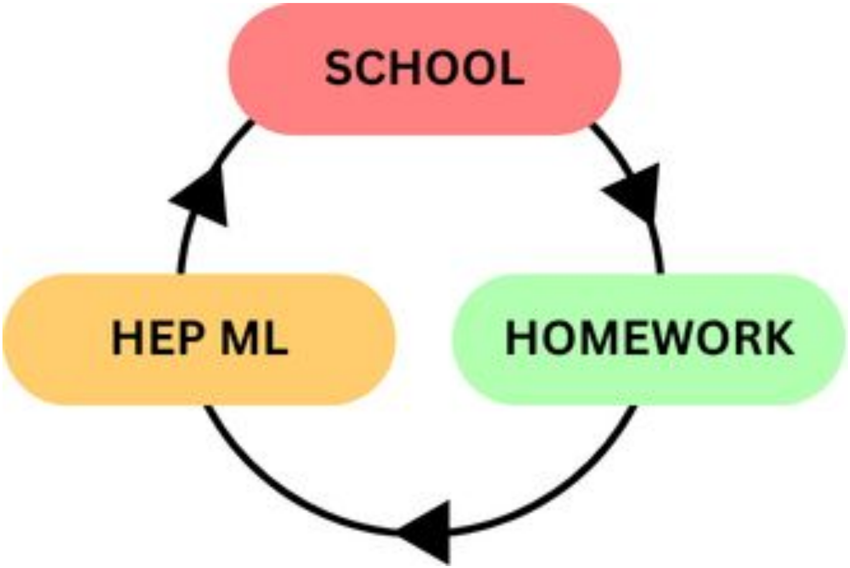
# An October of Tabular Data

**Building shallow neural networks for the 2014 ATLAS Higgs Machine Learning Challenge on Kaggle**



# An October of Tabular Data

**Building shallow neural networks for the 2014 ATLAS Higgs Machine Learning Challenge on Kaggle**

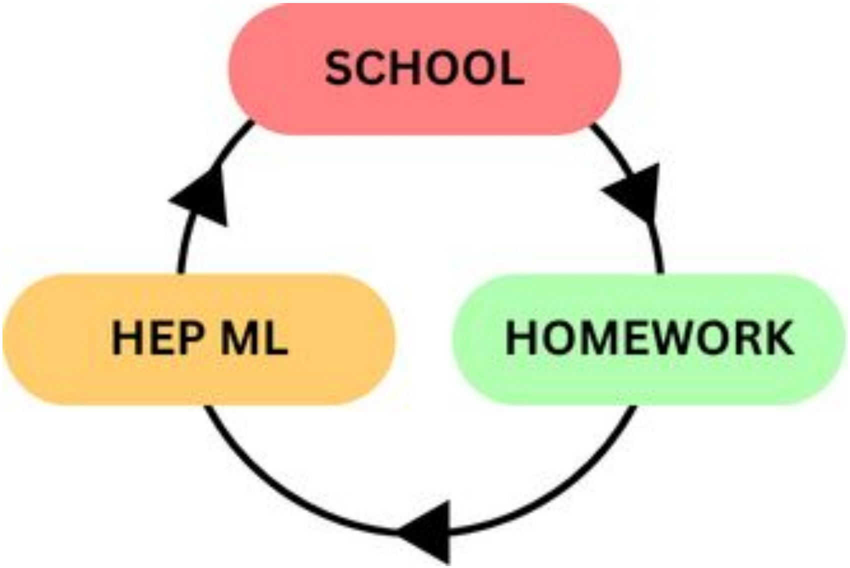



**Process:**

- Start with high-level features
- Feed as tabular inputs to the neural network

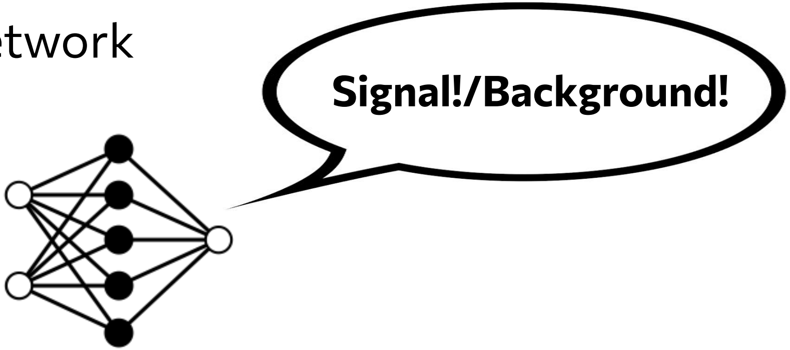
# An October of Tabular Data

**Building shallow neural networks for the 2014 ATLAS Higgs Machine Learning Challenge on Kaggle**




**Process:**

- Start with high-level features
- Feed as tabular inputs to the neural network



# Taking a Look @ SOTA

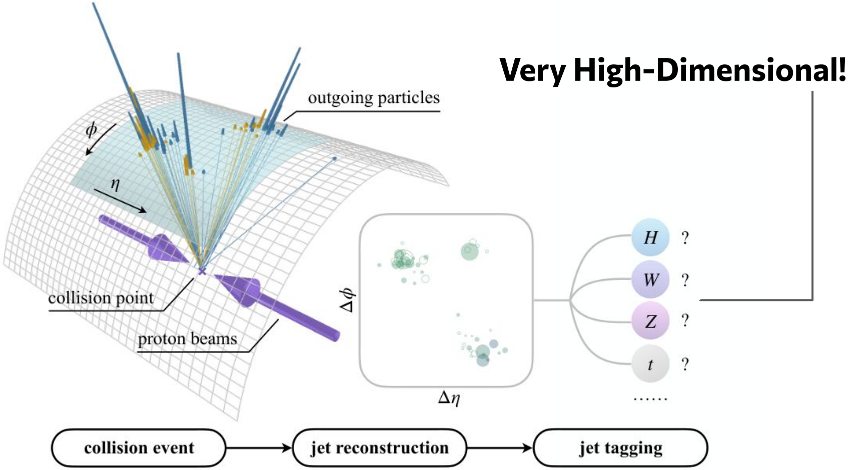
- Jet inputs are now represented as unordered sets of particles to feed into deep neural networks (DNNs)
  - JetClass Dataset [1]
  - Top Tagging Dataset [2]
- GNNs/Transformers/Deep Sets
  - ParticleNet [3]
  - Particle Transformer [1]
  - Energy Flow Networks [4]

[1] H. Qu, C. Li, and S. Qian, Particle transformer for jet tagging (2024), arXiv:2202.03772 [hep-ph].
[2] G. Kasieczka et al., The machine learning landscape of top taggers, SciPost Physics 7, 10.21468/scipostphys.7.1.014 (2019).
[3] H. Qu and L. Gouskos, Jet tagging via particle clouds, Physical Review D 101, 10.1103/physrevd.101.056019 (2020).
[4] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: deep sets for particle jets, Journal of High Energy Physics 2019, 10.1007/jhep01(2019)121 (2019).

# Taking a Look @ SOTA

- Jet inputs are now represented as unordered sets of particles to feed into deep neural networks (DNNs)
  - JetClass Dataset [1]
  - Top Tagging Dataset [2]
- GNNs/Transformers/Deep Sets
  - ParticleNet [3]
  - Particle Transformer [1]
  - Energy Flow Networks [4]

**High-Level Observables**





[1] H. Qu, C. Li, and S. Qian, Particle transformer for jet tagging (2024), arXiv:2202.03772 [hep-ph].
[2] G. Kasieczka et al., The machine learning landscape of top taggers, SciPost Physics 7, 10.21468/scipostphys.7.1.014 (2019).
[3] H. Qu and L. Gouskos, Jet tagging via particle clouds, Physical Review D 101, 10.1103/physrevd.101.056019 (2020).
[4] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: deep sets for particle jets, Journal of High Energy Physics 2019, 10.1007/jhep01(2019)121 (2019).

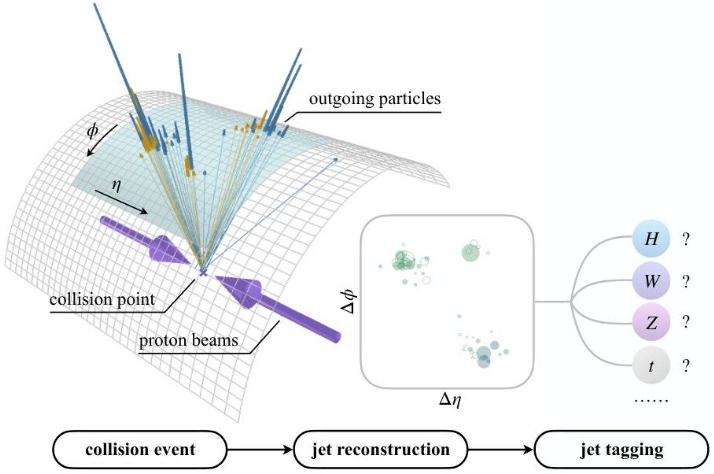
# THE ISSUE

As particle cloud jet tagging becomes more ubiquitous, models tend to lose their simplicity & interpretability.

# The Goal

**Goal:** Assess the extent to which a lower-dimensional set of features can approximate the outputs of a black-box jet tagger that uses particle cloud inputs

**High-Level Observables**



# Three Works, Three Different Fields

1. Rediscovering orbital mechanics with machine learning, P. Lemos et al. (2022).
    a. Applied symbolic distillation to obtain a closed-form mathematical expression supervised by a GNN trained on orbital mechanics data
2. Discovering interpretable models of scientific image data with deep learning, C. J. Soelistyo and A. R. Lowe (2024).
    a. Uses the latent variables outputted by a variational autoencoder for hard-label classification using symbolic regression in medical imaging
3. The physics behind ML-based quark-gluon taggers, S. Vent et al. (2026).
    a. Correlates ParticleNet's latent features with known physics observables

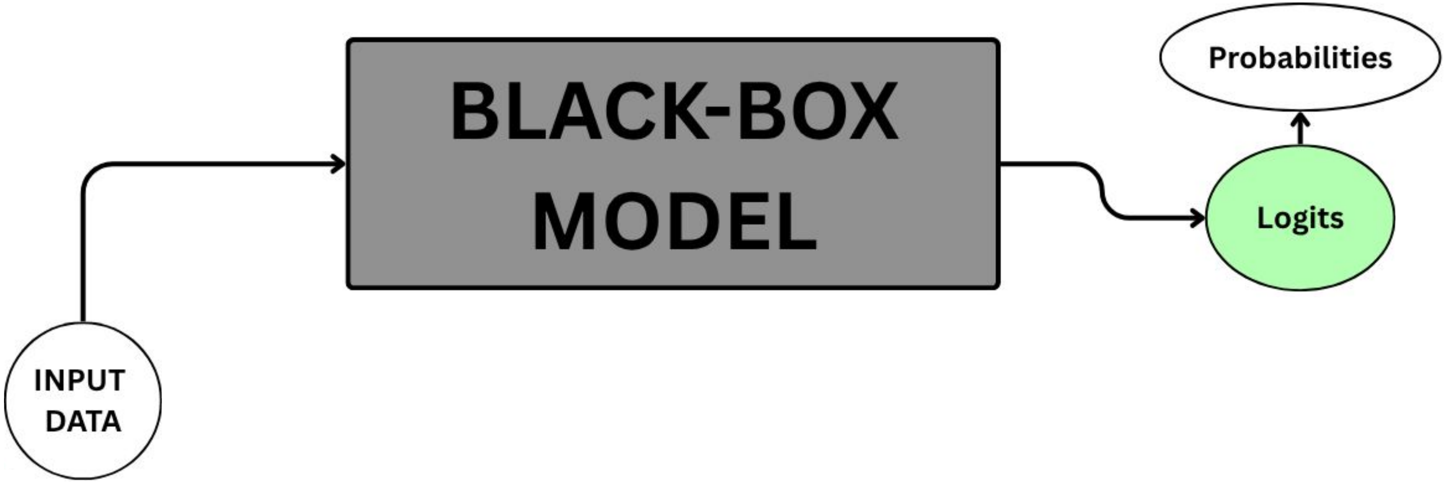
# Three Works, Three Different Fields

1. Rediscovering orbital mechanics with machine learning, P. Lemos et al. (2022).
    a. Applied symbolic distillation to obtain a closed-form mathematical expression supervised by a GNN trained on orbital mechanics data
2. Discovering interpretable models of scientific image data with deep learning, C. J. Soelistyo and A. R. Lowe (2024).
    a. Uses the latent variables outputted by a variational autoencoder for hard-label classification using symbolic regression in medical imaging
3. The physics behind ML-based quark-gluon taggers, S. Vent et al. (2026).
    a. Correlates ParticleNet's latent features with known physics observables

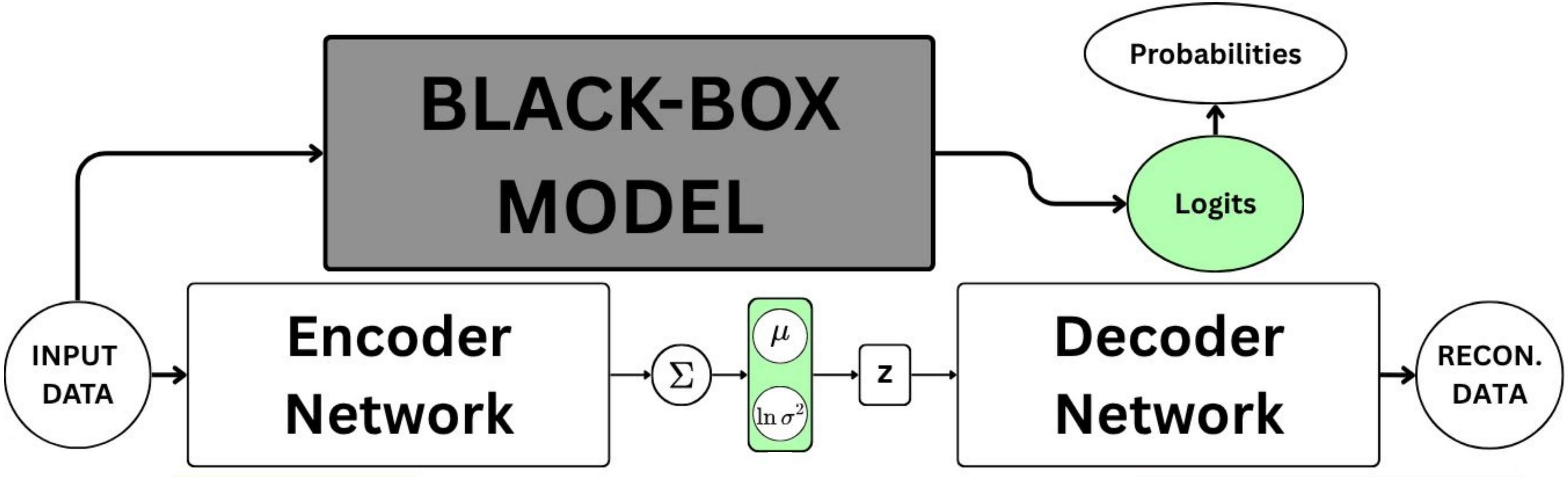
# ~~The~~ A Solution

# ~~The~~ A Solution

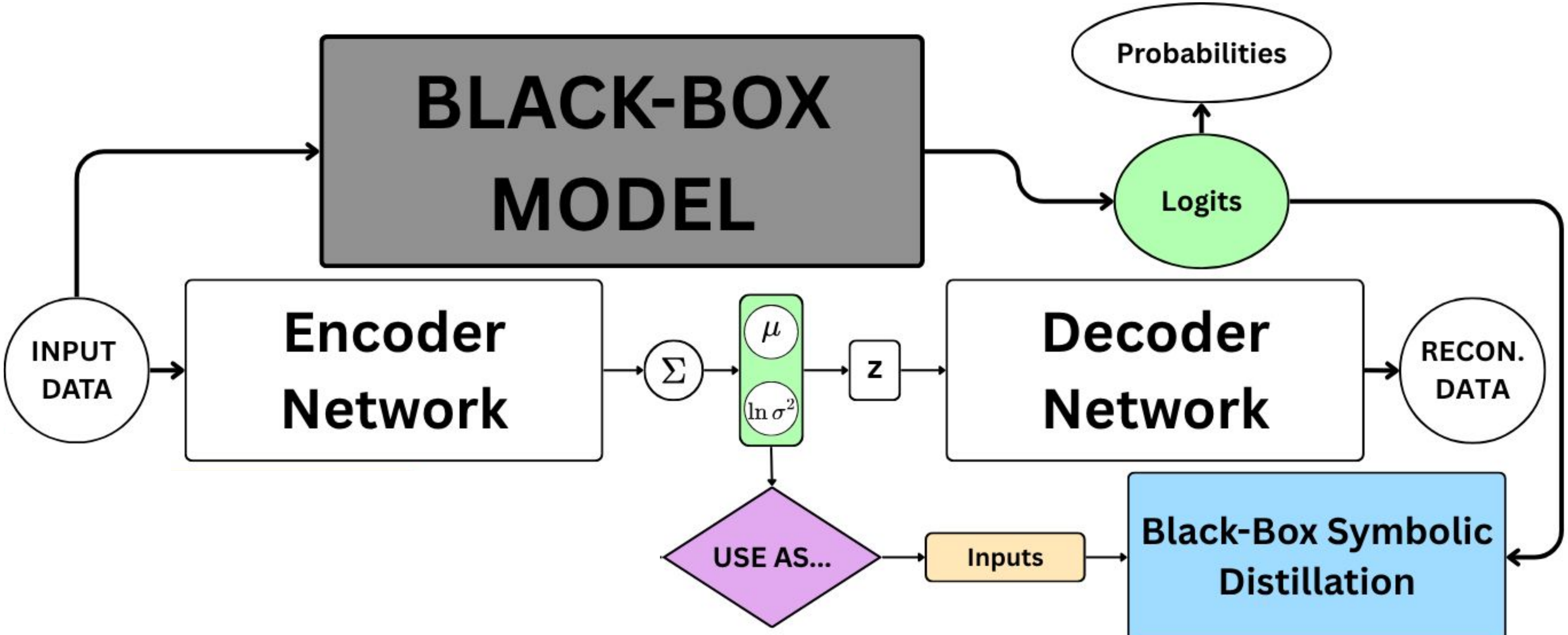
# ~~The~~ A Solution



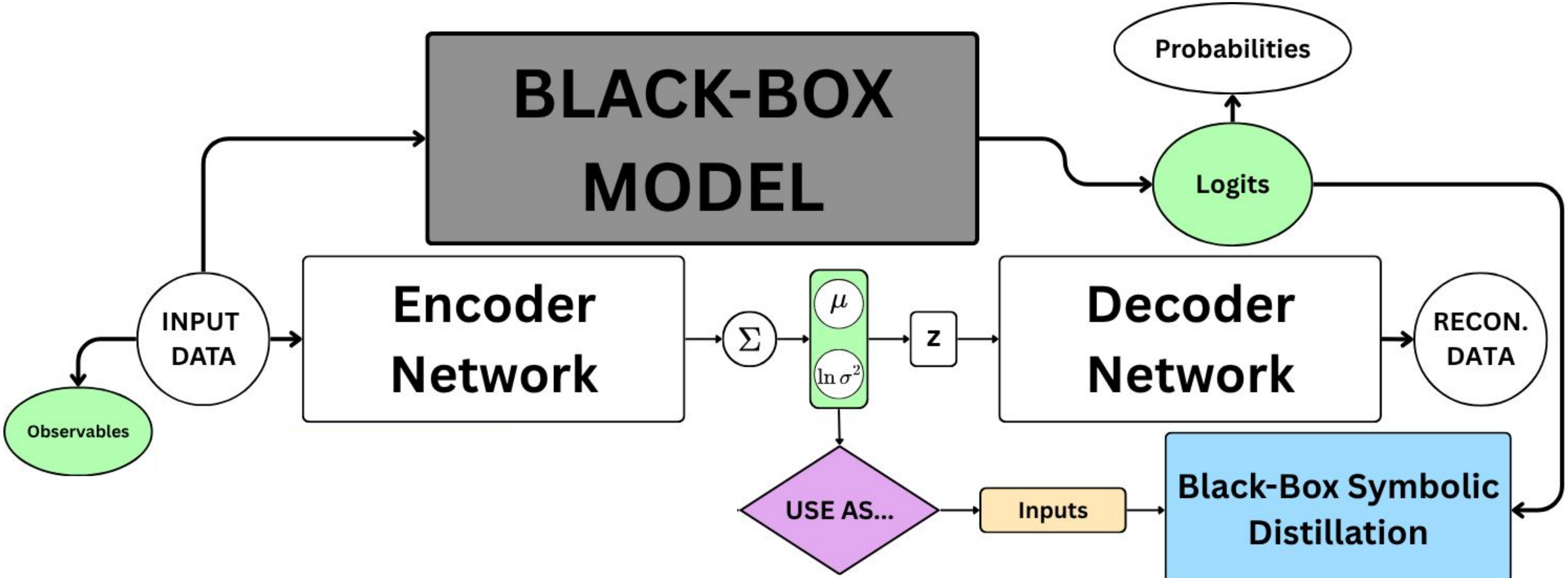
# ~~The~~ A Solution



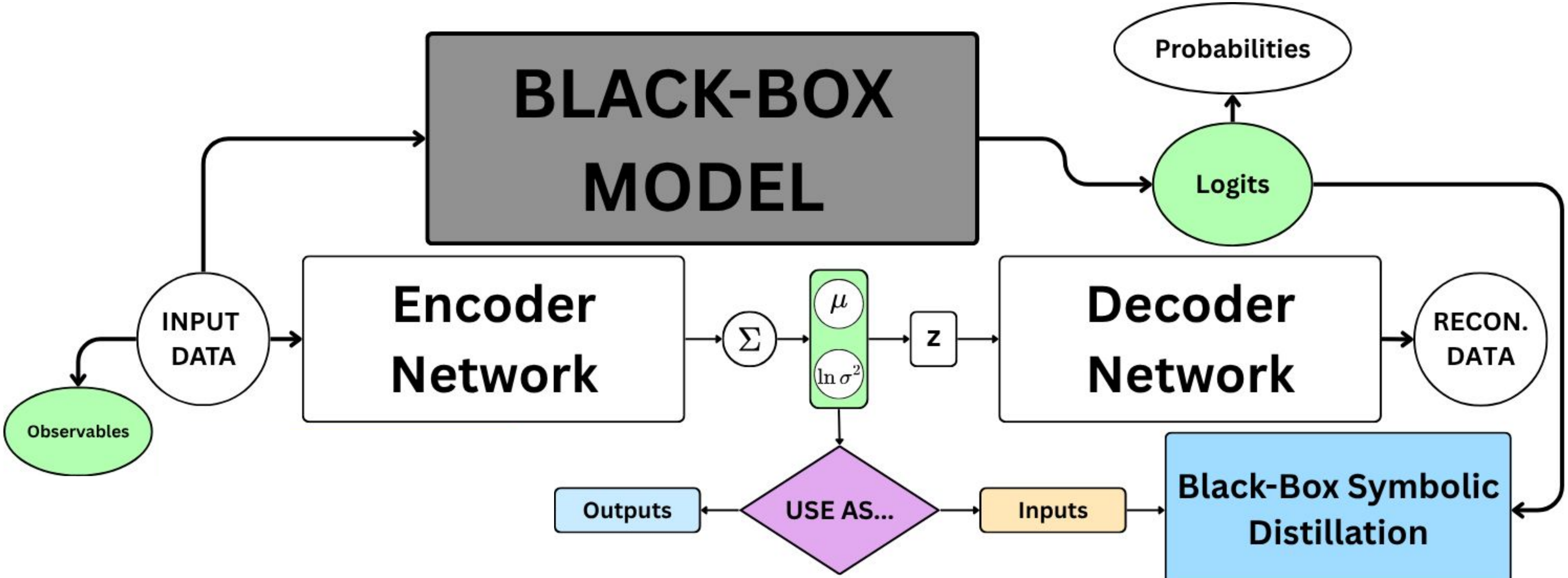
# ~~The~~ A Solution



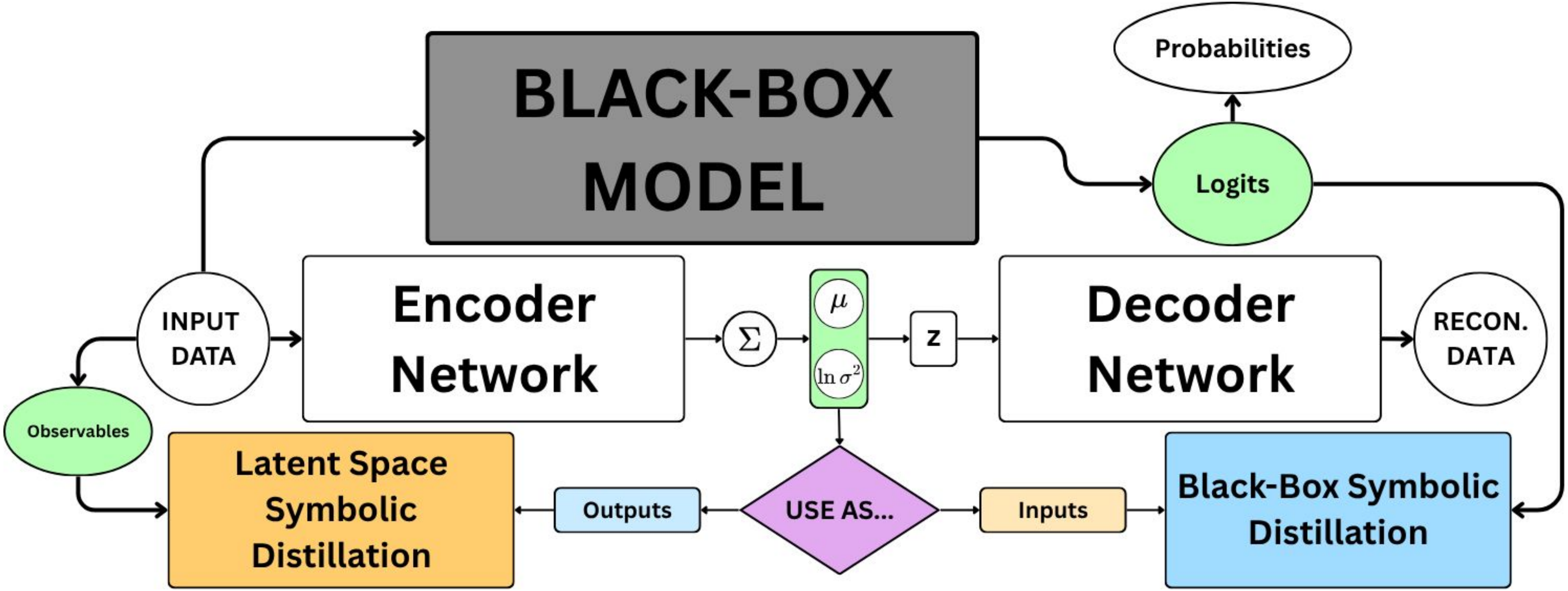
# ~~The~~ A Solution



# ~~The~~ A Solution



# ~~The~~ A Solution



# Selected Jet Tagging Tasks

- Three signal events from the JetClass Dataset [1]:
  - $t \to bqq'$
  - $W \to qq$
  - $H \to gg$
- $Z \to \nu\nu + \mathrm{jets}$ was selected as the primary background class
- All DL models were trained on a 10% subset of the dataset. SR models were trained on a 1% subset.

[1] H. Qu, C. Li, and S. Qian, Particle transformer for jet tagging (2024), arXiv:2202.03772 [hep-ph].

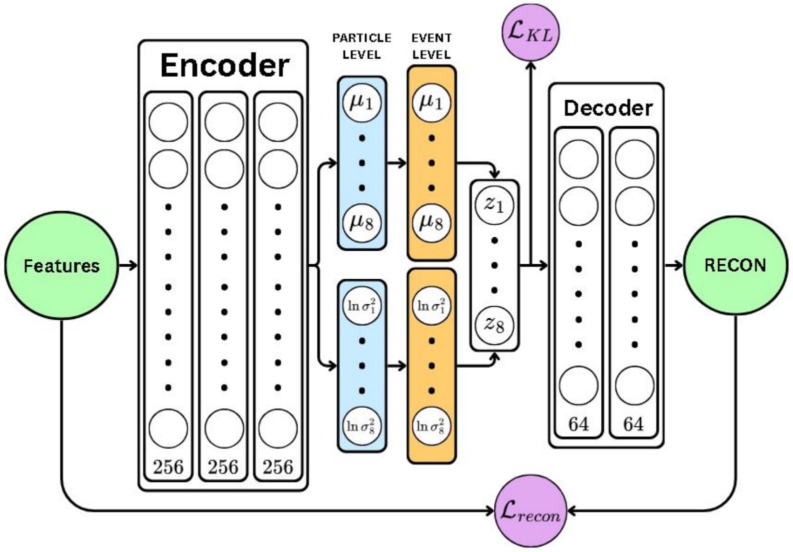
# Black-Box Architectures

- Chosen black-box jet taggers to distill:
  - Particle Transformer [1]
  - ParticleNet [3]
  - ResNet [5]
- All DL models are implemented in PyTorch
- Using the Ranger optimizer [6], models were trained for 5 epochs on the training set with a batch size of 64, with validation set performance evaluated in between.
- Post-training, black-box models outputs for the training set were saved to produce targets for SR

[1] H. Qu and L. Gouskos, Jet tagging via particle clouds, Physical Review D 101, 10.1103/physrevd.101.056019 (2020).
[3] H. Qu, C. Li, and S. Qian, Particle transformer for jet tagging (2024), arXiv:2202.03772 [hep-ph].
[5] J. S. H. Lee, I. Park, I. J. Watson, and S. Yang, Quark-Gluon Jet Discrimination Using Convolutional Neural Networks, J. Korean Phys. Soc. 74, 219 (2019), arXiv:2012.02531 [hep-ex].
[6] L. Wright and N. Demeure, Ranger21: a synergistic deep learning optimizer (2021), arXiv:2106.13731 [cs.LG].

# Deep Sets Autoencoder

- Deep Sets autoencoder architecture adapted from [B. Ostdiek, SciPost Physics 12, 10.21468/scipostphys.12.1.045 (2022)].
- Particle-level representations are first constructed, before being summed into event-level representations
- 8 event-level latent means, along with 8 corresponding 8 log-variances



# Two Types of Symbolic Distillation

## Black-Box

- Inputs: Latent variables of the autoencoder
- Targets: Signal-Background logit differences of the black-box jet tagger
- Unary operators: (tanh, sqrt, sin)

$$\mathcal{L}(r) = c\ell(f(X)) + \begin{cases} \frac{r^2}{2} & \text{if} |r| \leq \alpha \\ \alpha|r| - \frac{\alpha^2}{2} & \text{otherwise} \end{cases}$$

Where $\alpha = 1.35$ & $c = 0.01$

## Latent Space

- Inputs:
  - Basic jet kinematic properties ($p_T$, $\eta$, $\phi$, m, E)
  - Soft Drop groomed jet mass ($m_{SD}$)
  - 4 N-Subjettiness variables ($\tau_1$, $\tau_2$, $\tau_3$, $\tau_4$)
  - Set of 36 Energy Flow Polynomials (EFPs) up to degree $d \leq 4$
  - Jet Multiplicity
- Targets: Autoencoder latent means
- Unary operators: (log, sqrt, exp)
- Same objective function as black-box

**SR performed in PySR [7]**

[7] M. Cranmer, Interpretable machine learning for science with pysr and symbolicregression.jl (2023), arXiv:2305.01582 [astro-ph.IM].

# Classification Performance

- Ranges of surrogate AUC % drops vs. full models
  - $t \rightarrow bqq'$: 6.25%-7.18%
  - $W \rightarrow qq$: 11.87%-12.68%
  - $H \rightarrow gg$: 6.25%-10.5%
- Ranges of $\frac{Rej_{DL}}{Rej_{Surrogate}}$
  - $t \rightarrow bqq'$: 11.26-21.45
  - $W \rightarrow qq$: 6.28-8.96
  - $H \rightarrow gg$: 3.28-6.30

| Model | $t \rightarrow bqq'$ AUC | $W \rightarrow qq$ AUC | $H \rightarrow gg$ AUC |
|---|---|---|---|
| ParT | **0.9855** | **0.9615** | **0.9569** |
| ParT-S | 0.9147 | 0.8396 | 0.8699 |
| ResNet | 0.9796 | 0.9527 | 0.9522 |
| ResNet-S | 0.9184 | 0.8396 | 0.8927 |
| ParticleNet | 0.9809 | 0.9566 | 0.9568 |
| ParticleNet-S | 0.9147 | 0.8422 | 0.8564 |

| Model | $t \rightarrow bqq'$ $\mathrm{Rej}_{50\%}$ | $W \rightarrow qq$ $\mathrm{Rej}_{50\%}$ | $H \rightarrow gg$ $\mathrm{Rej}_{50\%}$ |
|---|---|---|---|
| ParT | **285.3** | **151.4** | **55.6** |
| ParT-S | 13.3 | 16.9 | 10.1 |
| ResNet | 159.4 | 106.2 | 47.2 |
| ResNet-S | 14.0 | 16.9 | 14.4 |
| ParticleNet | 149.8 | 113.8 | 55.4 |
| ParticleNet-S | 13.3 | 14.98 | 8.8 |

*Surrogates highlighted in yellow*

# Trends Across Equations

| Model | Equation $t \to bqq'$ |
|---|---|
| ParT-S | $\Delta z = -9.51\mu_4 - 3.148$ |
| ResNet-S | $\Delta z = \frac{-1.93^{ln \sigma_0^2} - \mu_4}{0.113}$ |
| ParticleNet-S | $\Delta z = -8.07\mu_4 - 2.56$ |

| Model | Equation $W \to qq$ |
|---|---|
| ParT-S | $\Delta z = \frac{\mu_7 \left(\ln \sigma_7^2 + \mu_2 - \mu_3\right)}{0.319}$ |
| ResNet-S | $\Delta z = \frac{\mu_7 \left(\ln \sigma_7^2 + \mu_2 - \mu_3\right)}{0.353}$ |
| ParticleNet-S | $\Delta z = 5.59 \tanh\left(\mu_7 \left(\mu_2 - \mu_3 - 0.690\right)\right)$ |

| Model | Equation $H \to gg$ |
|---|---|
| ParT-S | $\Delta z = -\mu_6 - 5.14 \left(\ln \sigma_7^2 - \mu_7 + 1.61\right)$ |
| ResNet-S | $\Delta z = \left(\mu_5 + 2.48\right)\left(-\ln \sigma_6^2 - \ln \sigma_7^2 - \mu_6 + \mu_7 - 3.24\right)$ |
| ParticleNet-S | $\Delta z = -5.48 \left(\ln \sigma_7^2 - \mu_7 + 1.51\right)$ |

- The complexities (number of operators + terms) of surrogate equations ranged from 4-13
- Outputted logit differences tended to follow linear relationships with respect to most latent variables
- The same few latent variables tended to follow similar trends across equations

# Examples of Latent Mean Interpretations

$t \to bqq'$

$W \to qq$

$H \to gg$

**Disclaimer: These are POST-HOC interpretations!**

# Examples of Latent Mean Interpretations

$$t \to bqq'$$

$$\mu_4 \approx -4.92\tau_1 + 0.737$$

$$\mu_0 \approx 6.87\sqrt{\tau_3} - 1.43$$

$$W \to qq$$

$$H \to gg$$

**Disclaimer: These are POST-HOC interpretations!**

# Examples of Latent Mean Interpretations

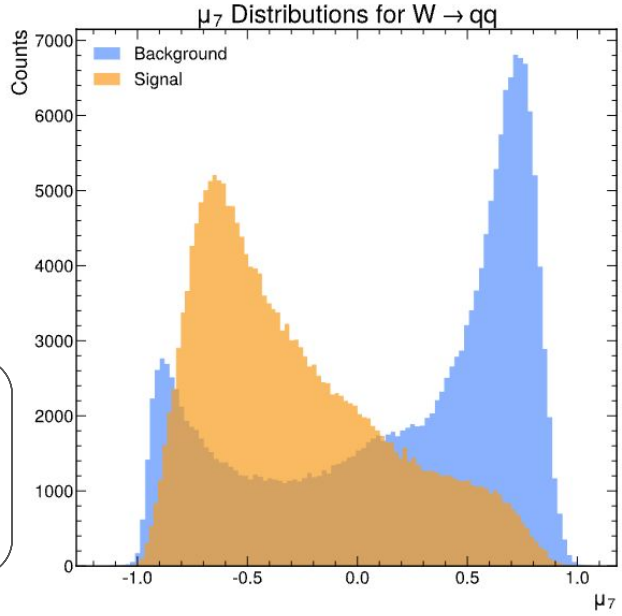
$$t \to bqq'$$

$$\mu_4 \approx -4.92\tau_1 + 0.737$$

$$\mu_0 \approx 6.87\sqrt{\tau_3} - 1.43$$

*Interesting case of signal/background separation in a latent variable!*

$$W \to qq$$




$$\mu_7 \approx \frac{-0.452\tau_1 - 0.452\left(-0.373 + \frac{m_{SD}}{E_{jet}}\right)\left(\sqrt{\tau_2} + 0.0899\right)}{EFP_{21} + 0.0308}$$

$$H \to gg$$

**Disclaimer: These are POST-HOC interpretations!**

# Examples of Latent Mean Interpretations

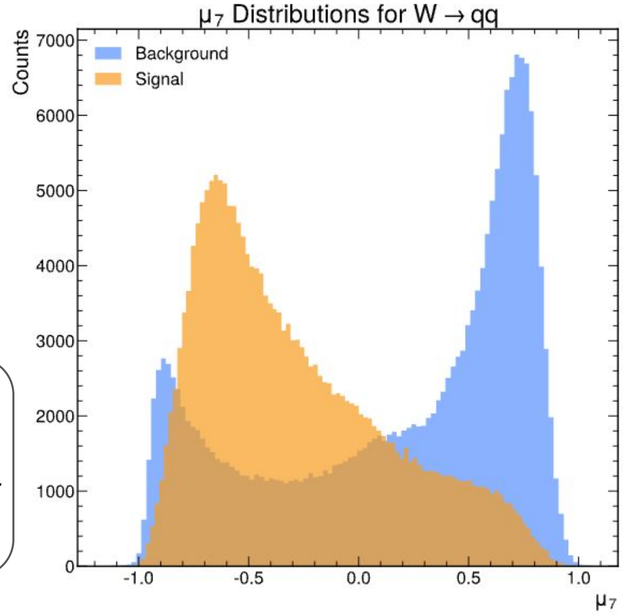
## $t \to bqq'$

$$\mu_4 \approx -4.92\tau_1 + 0.737$$

$$\mu_0 \approx 6.87\sqrt{\tau_3} - 1.43$$

*Interesting case of signal/background separation in a latent variable!*

## $W \to qq$




$$\mu_7 \approx \frac{-0.452\tau_1 - 0.452\left(-0.373 + \frac{m_{SD}}{E_{jet}}\right)\left(\sqrt{\tau_2} + 0.0899\right)}{EFP_{21} + 0.0308}$$

## $H \to gg$

$$\mu_7 \approx \left(0.130\eta^2 - 4.77\tau_1\right) \ln\left(0.945^{m_{SD}} + EFP_1\right) - 0.705$$

$$\mu_5 \approx \frac{EFP_{21} + 0.0653}{\frac{EFP_{18}}{0.154} + 0.0323} - 1.42$$

**Disclaimer: These are POST-HOC interpretations!**

# “WHY SHOULD WE CARE?”

# Takeaways

# Takeaways

- Surrogates are vastly more interpretable and simplistic compared to black-box jet taggers using particle cloud inputs

# Takeaways

- Surrogates are vastly more interpretable and simplistic compared to black-box jet taggers using particle cloud inputs
- However, surrogates fall behind in general classification performance metrics, though this drop is consistent across architectures

# Takeaways

- Surrogates are vastly more interpretable and simplistic compared to black-box jet taggers using particle cloud inputs
- However, surrogates fall behind in general classification performance metrics, though this drop is consistent across architectures
- Surrogates produced in these experiments still do not perform strongly enough to replace existing SOTA models in analytical jet tagging tasks

# Future Directions

# Future Directions

- If the latent variables themselves were supervised (e.g. KLD loss with a downstream classification branch), surrogates can potentially be used as black-box interpretations

# Future Directions

- If the latent variables themselves were supervised (e.g. KLD loss with a downstream classification branch), surrogates can potentially be used as black-box interpretations
- A fuller picture of latent variable & SR trends can be constructed by running multiple random-seed experiments (w/ varying architecture sizes)

# Future Directions

- If the latent variables themselves were supervised (e.g. KLD loss with a downstream classification branch), surrogates can potentially be used as black-box interpretations
- A fuller picture of latent variable & SR trends can be constructed by running multiple random-seed experiments (w/ varying architecture sizes)
- Constructing surrogates w/ varying objective functions (e.g. MSE instead of Huber)

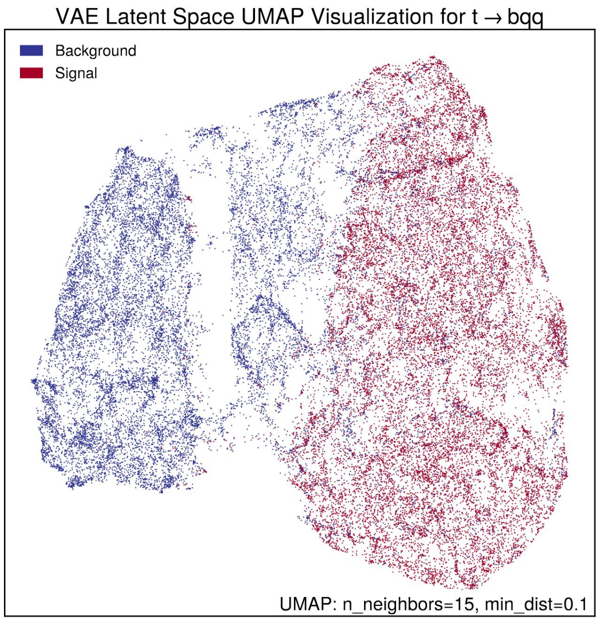
# Thank you for your attention! Questions?

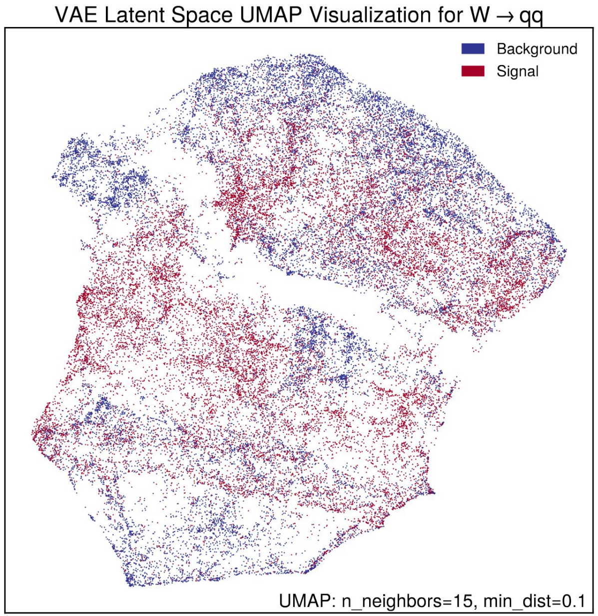
I am also looking for future research opportunities. If you have any ideas for me, feel free to reach out!

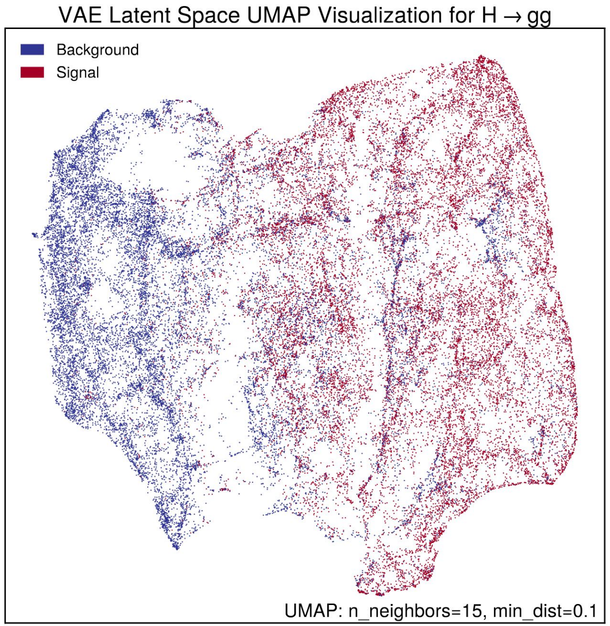
alangu612@gmail.com

# Backup Slides

# UMAP Latent Space Visualizations



# Post-Hoc Interpretability

- Post-hoc interpretability studies analyze how an already-trained black-box model came to a given decision
- These methods can be either be model-agnostic (black-box) or model-specific
- Black-Box Techniques
  - SHAP [5]
  - LIME [6]
- Model-Specific Techniques
  - Attention Heat Maps [7]
  - Layerwise Relevance Propagation (LRP) [8]

[5] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions (Curran Associates, Inc., 2017)
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier (2016), arXiv:1602.04938 [cs.LG].
[7] P. Bhatnagar, J. Duarte, A. Gandrakota, E. E. Khoda, J. Ngadiuba, V. G. Sahu, and A. Wang (CMS), Interpreting and Accelerating Transformers for Jet Tagging 10.2172/2474973
[8] G. Agarwal, L. Hay, I. Iashvili, B. Mannix, C. McLean, M. Morris, S. Rappoccio, and U. Schubert, Explainable ai for ml jet taggers using expert variables and layerwise relevance propagation, arXiv: Data Analysis, Statistics and Probability (2020

# Summary of Methods

## Framework:

1. Construct a set of latent variables discovered by an independently trained Deep Sets variational autoencoder (VAE) as a compressed representation of the particle cloud data
2. Symbolically regress the logit differences of a black-box jet tagger with the latent variables as inputs, creating a symbolic surrogate
3. Interpret the meaning of each latent variable via symbolic regression, with a defined set of KNOWN jet observables as inputs

## Methods

- Surrogates were trained to approximate 3 deep learning architectures across 3 jet tagging tasks
- Surrogates were compared with their DL ROC AUCs & background rejection rates
- The symbolic approximation of variables were used as post-hoc interpretations

# Black-Box Symbolic Distillation

- The PySR package [10] was used as the primary symbolic regression package
- Configurations:
  - Basic arithmetic binary operators (+, −, ×, ÷, ^) used
  - Nonlinear unary operators (tanh, sqrt, sin)
  - Maximum complexity (sum of the counts of operators, variables, weights) of 40
  - Population size 27, 48 populations, 4560 iterations, 1520 cycles per iteration
- Latent means & log variances were taken as inputs, and fit to the signal - background logit differences of the DL models
- Huber Loss objective function with threshold of 1.35 & parsimony of 0.01

$$\mathcal{L}(r) = c\ell(f(X)) + \begin{cases} \frac{r^2}{2} & \text{if} |r| \leq \alpha \\ \alpha|r| - \frac{\alpha}{2} & \text{otherwise} \end{cases}$$

[10] M. Cranmer, Interpretable machine learning for science with pysr and symbolicregression.jl (2023), arXiv:2305.01582 [astro-ph.IM].

# Selected Jet Observables

- IRC-safe observables
  - Basic jet kinematic properties ($p_T$ , $\eta$, $\phi$, m, E)
  - Soft Drop groomed jet mass (mSD)
  - 4 N-Subjettiness variables ($\tau_1$, $\tau_2$, $\tau_3$, $\tau_4$)
  - Set of 36 Energy Flow Polynomials (EFPs) up to degree $d \leq 4$
- IRC-unsafe observables
  - Jet Multiplicity
- Using PySR, jet observables were fit to the latent means with identical configurations to black-box symbolic distillation with the following changes:
  - Max equation complexity of 30
  - Unary operators log, sqrt, and exp

# Partial Dependence Plots (PDPs)

- The most common surrogate-wide latent variables were identified
- The average trends between latent variable values and surrogate signal-background were assessed via PDPs
- Each surrogate PDP line was centered on their respective values at the minimum latent variable value

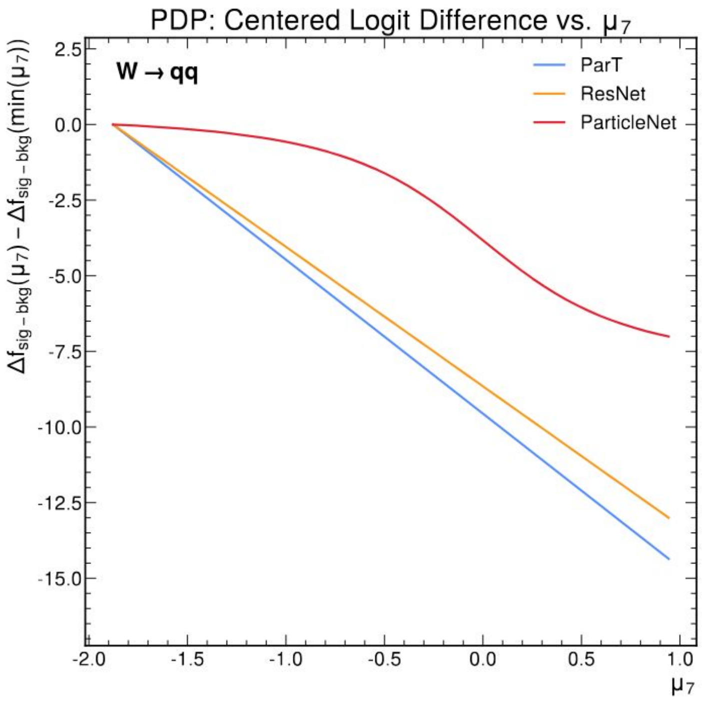



FIG. 5: Centered Partial Dependence Plots of surrogate logit differences vs. $\mu_7$

# Trends in Boosted t → bqq Tagging

- Surrogates tend to each find a simple, positive relationship between the logit difference and the 1-subjettiness variable
  - For background-like jets, the $\tau 1$ term approaches 0, contributing to a lower logit difference value

| Model | Equation<br>$t \to bqq$ |
|---|---|
| ParT-S | $\Delta z = -9.51\mu_4 - 3.148$ |
| ResNet-S | $\Delta z = \frac{-1.93^{ln\sigma_0^2} - \mu_4}{0.113}$ |
| ParticleNet-S | $\Delta z = -8.07\mu_4 - 2.56$ |

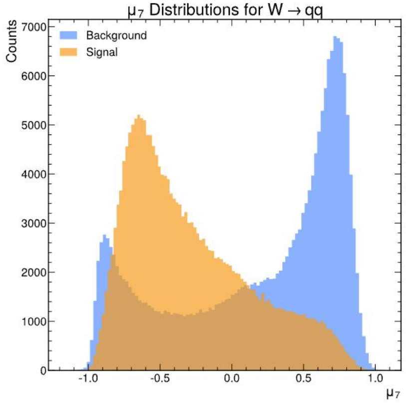
$$\mu_4 \approx -4.92\tau_1 + 0.737$$

# Trends in Boosted W → qq Tagging

- PDPs highlight that μ2, μ3, and ln σ27 have negligible impact on logit difference values
- Each surrogate finds a downward monotonic relationship with respect to μ7
- While μ7 does not have a simple mapping to jet observables, histograms reveal that it serves as a signal-background discriminating term

| Model | Equation |
|---|---|
| | $W \to qq$ |
| ParT-S | $\Delta z = \frac{\mu_7\left(\ln \sigma_7^2 + \mu_2 - \mu_3\right)}{0.319}$ |
| ResNet-S | $\Delta z = \frac{\mu_7\left(\ln \sigma_7^2 + \mu_2 - \mu_3\right)}{0.353}$ |
| ParticleNet-S | $\Delta z = 5.59 \tanh\left(\mu_7\left(\mu_2 - \mu_3 - 0.690\right)\right)$ |

$$\mu_7 \approx \frac{-0.452\tau_1 - 0.452\left(-0.373 + \frac{m_{SD}}{E_{jet}}\right)\left(\sqrt{\tau_2} + 0.0899\right)}{EFP_{21} + 0.0308}$$



# Trends in Boosted H → gg Tagging

- Surrogates tend to discover a positive monotonic relationship with $\mu_7$ which approaches lower magnitudes for background jets (τ1 ≈ 0)

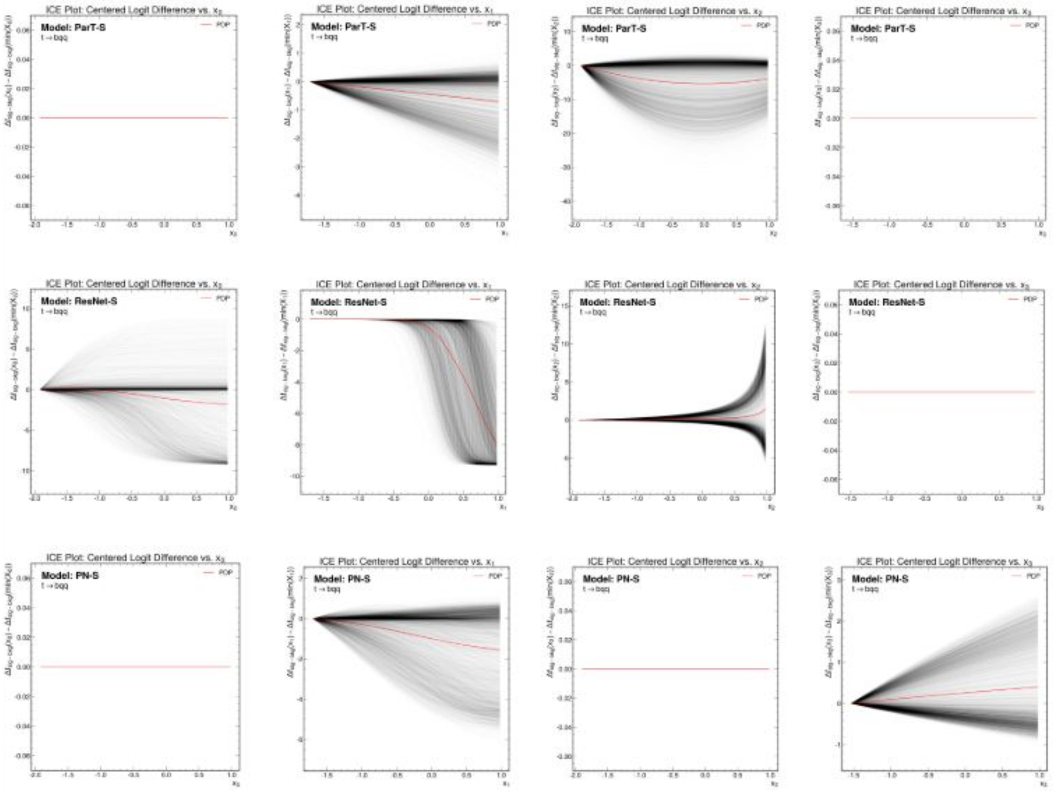
| Model | Equation |
|---|---|
| | $H \to gg$ |
| ParT-S | $\Delta z = -\mu_6 - 5.14\left(\ln\sigma_7^2 - \mu_7 + 1.61\right)$ |
| ResNet-S | $\Delta z = (\mu_5 + 2.48)\left(-\ln\sigma_6^2 - \ln\sigma_7^2 - \mu_6 + \mu_7 - 3.24\right)$ |
| ParticleNet-S | $\Delta z = -5.48\left(\ln\sigma_7^2 - \mu_7 + 1.51\right)$ |

$$\mu_7 \approx \left(0.130\eta^2 - 4.77\tau_1\right)\ln\left(0.945^{m_{SD}} + EFP_1\right) - 0.705$$

# Notes on Interpretability

- Since the VAE latent variables were independently discovered, surrogates cannot be taken as interpretations of the black-box jet taggers, but rather independent approximations of their final outputs
- The $\ln \sigma_i^2$ terms within the symbolic equations represent the log-widths of the posterior, and there has been no clean mapping found to jet observables
- The latent variables used within the final symbolic equations are still produced by the black-box encoder, whereas the observable-level equations are post-hoc interpretations

# ICE Plot Examples




**Disclaimer: These were test runs on a 0.1% subset of the data!**

# LinkedIn