Conditional VampPrior β-Variational AutoEncoders and Physics-Informed Neural Networks for the simulation and detection of the Higgs Boson

JOSUE VELASCO

February 2025

1 Abstract

In this work, I present a physics-informed framework for Higgs boson identification that integrates deep generative modeling with physics-based constraints. Synthetic data, augmented by a Conditional VampPrior β-Variational Autoencoder that generates both detector features and labels that serves as the basis for training a Physics-Informed Neural Network (PINN). The PINN is designed to not only perform signal classification but also enforce key physical constraints derived from the four-momentum of particle decays. In particular, my loss function incorporates three physics regularization terms: an invariant mass constraint that penalizes deviations of the reconstructed mass from the expected 125GeV, an angular correlation constraint that enforces the expected transverse angular separation between decay products, and an energy balance constraint that promotes consistency with energy–momentum conservation. These constraints are combined with a binary cross-entropy classification loss, guiding the model toward physically plausible predictions. Experimental results indicate that this integrated approach improves the extraction of Higgs boson signals in noisy collider environments. Future work will explore further refinement of the constraint weights and alternative formulations to enhance the model's fidelity to real-world physics.

2 Introduction

The discovery and precise measurement of the Higgs boson represent monumental achievements in high-energy physics, confirming the mechanism of electroweak symmetry breaking as predicted by the Standard Model. Experiments at the Large Hadron Collider (LHC), notably within the ATLAS experiment [2], generate vast and complex datasets, where rare signal events are deeply embedded in overwhelming backgrounds. Traditional statistical methods have provided robust tools for data analysis, yet the increasing volume and complexity of data call for more sophisticated, data-driven approaches.

Recent advances in deep learning have introduced generative models such as Variational Autoencoders (VAEs) [5] [3] as powerful tools for modeling high-dimensional data distributions. VAEs can learn compact latent representations of complex data, allowing for the generation of realistic synthetic datasets that mimic real detector responses. Enhancements such as the Variational Mixture of Posteriors Prior (VampPrior) [7] further improve these models by learning a more flexible prior, thereby capturing the underlying data diversity more effectively.

In parallel, Physics-Informed Neural Networks (PINNs) have emerged as a promising methodology to embed known physical laws directly into the training process [6]. By incorporating theoretical constraints (such as the invariant mass of Higgs decay products which is expected to peak near 125GeV), PINNs can improve the reliability and interpretability of model predictions. This dual approach of generative modeling and physics-informed learning offers a compelling strategy for both simulating realistic high-energy physics data and enhancing signal extraction performance.

In this work, I propose an integrated framework that leverages a conditional VampPrior-enhanced conditional VAE for synthetic data generation and a PINN for signal extraction. The conditional VAE is trained to generate detector data conditioned on event labels (signal vs. background), while the PINN is designed to predict key physical observables, such as the reconstructed invariant mass, under the constraint that the predictions adhere to established physics.

This paper is organized as follows:

- Section 3 and 4: reviews related work in generative modeling and physics-informed learning for high-energy physics.
- Section 5: presents experimental results and evaluations.
- Section 6: concludes with discussions and future work directions.

3 Conditional VampPrior β-VAE

Variational Autoencoders (VAEs) are unsupervised generative models that learn a latent representation of the input data by minimizing the reconstruction error via the Kullback–Leibler (KL) divergence between the approximate posterior and a prior distribution. The training objective of the loss function, given by the Evidence Lower Bound (ELBO) [4], is defined as:

$$\mathcal{L}(x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \beta D_{\mathrm{KL}} \left(q_\phi(z|x) \| p_\theta(z)\right) \tag{1}$$

where:

- $\mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right]$ =Reconstruction term, composed by:
  - $\mathbb{E}_{q_\phi(z|x)}$ = Expected posterior given a latent prior z
  - $\log p_\theta(x|z)$ = decoder that reconstructs the data from a latent representation z obtained form a posterior given x

- $D_{\mathrm{KL}} \left(q_\phi(z|x) \| p_\theta(z)\right)$ = The actual KL divergence that checks how much the posterior $q(z|x)$ deviates form the prior p(z) and keep it as close as possible to the prior

- β = regularization term that controls the impact of the KL Divergence

To enhance the flexibility of the latent representation, I employ a VampPrior (a learned mixture of variational posteriors derived from pseudo-inputs)

$$p(z) = \frac{1}{K} \sum_{k=1}^{K} q_\phi(z|u_k)$$

where u_k are the pseudo-inputs, this way, the prior learns to more accurately reflect the true underlying distribution of the data.

Since the task requires generating synthetic labels (the target variable), I extend the VampPrior β-VAE into a conditional model, conditioning both the encoder and decoder on one-hot encoded labels. In this way, the model not only learns the distribution of the detector features but also how these features correlate with the signal and background classes, resulting in an updated loss function:

$$\mathcal{L}(x, c) = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z, c)] - \beta D_{KL}(q_\phi(z|x, c) \| p(z|c)) \tag{2}$$

In this formulation, $q_\phi(z|x, c)$ and $p_\theta(x|z, c)$ denote the conditional encoder and decoder, respectively, and $p(z|c)$ represents the conditional prior.

Furthermore, to ensure that the synthetic data mimics the real data distribution, I incorporate a Maximum Mean Discrepancy (MMD) term into the loss function

[8], weighted by a regularization parameter λ. This additional term helps to align the generated synthetic distribution with that of the real data.

$$\text{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j) \quad (3)$$

where $k(\cdot, \cdot)$ is a kernel function (typically Gaussian). By minimizing this term along with the reconstruction and KL divergence losses, the overall loss function becomes:

$$\mathcal{L}_{total}(x, c) = L(x, c) + \lambda \text{MMD}(p_{\text{data}}(x), p_{\text{gen}}(x)) \quad (4)$$

4 Physics-Informed Neural Network

Following the generation of synthetic data, it was integrated with the actual dataset to enhance the foundation on which the model is trained. Particle identification in high-energy physics is essentially predicated upon its four-momentum, which includes its mass, momentum, and energy p_x, p_y, p_z, E [1]. Crucial observables for Higgs boson identification consist of invariant mass, transverse momentum, and energy balance. To steer the Physics-Informed Neural Network (PINN) towards predictions that adhere to physical plausibility, three physics-based regularization terms were incorporated into the loss function. These terms ensure that the network's output remains consistent with established physical constraints.

4.1 Invariant Mass Constraint

For a true Higgs boson event, the invariant mass of the decay products, reconstructed using, for example, the `DER_mass_MMC` variable, should cluster around 125GeV. I penalize deviations from this target using the following loss term:

$$\mathcal{L}_{\text{mass}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i > 0.5) \cdot \left(m_{\text{est}}^{(i)} - 125.0 \right)^2 \quad (5)$$

where:

- $\hat{y}_i > 0.5$ = the model's prediction for signal for the i-th element
- $m_{\text{est}}^{(i)}$ = the estimated mass (`DER_mass_MMC`) for the i-th element

This quadratic penalty ensures that the reconstruction of the mass is strongly enforced around the known Higgs mass.

4.2 Angular Correlation Constraint

Due to limitations in measuring momentum along the beam (z) axis, analyses in the ATLAS experiment focus on the transverse plane (x–y plane). In this plane, the azimuthal angle ϕ becomes crucial. For Higgs boson decays into a lepton and a hadronic tau, the relative azimuthal angle between these particles can serve as a proxy for the event topology. I define the angular constraint as:

$$\mathcal{L}_{\text{angle}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i > 0.5) \cdot \left((\phi_\tau^{(i)} - \phi_{\text{lep}}^{(i)}) - \frac{\pi}{2}\right)^2 \tag{6}$$

Here, the term $(\phi_\tau^{(i)} - \phi_{\text{lep}}^{(i)})$ is used as an approximation of the reconstructed azimuthal angle of the Higgs boson, ϕ_H. Subtracting $\pi/2$ sets a target value corresponding to the transverse (x–y) configuration (e.g. when the decay products are ideally balanced in the plane), and squaring the difference imposes a quadratic penalty for deviations.

4.3 Energy Balance Constraint

Energy conservation in a decay process implies that the sum of the energies (or the four-momenta) of the decay products must match the energy of the parent particle. For the Higgs boson, the invariant mass is computed via the relation:

$E = \sqrt{p^2 + m^2}$

where E is the energy, p is the momentum, and m is the rest mass. In the detector, the Missing Transverse Energy (MET) is an important observable that accounts for energy carried away by neutrinos. My energy balance constraint is formulated heuristically as:

$$\mathcal{L}_{\text{energy}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i > 0.5) \cdot \left([(p_\tau^{(i)} + p_{lep}^{(i)}) + m_{tr}^{(i)}] - E_{tr}^{(i)}\right)^2 \tag{7}$$

where:

- $(p_\tau^{(i)} + p_{lep}^{(i)})$ = the total transverse momentum given by the sum of the transeverse momentum $\sqrt{p_x^2 + p_y^2}$ of the hadronic tau and the transverse momentum $\sqrt{p_x^2 + p_y^2}$ of the lepton of the i-th element
- m_{tr} = The transverse mass $m_{tr}(\vec{a}, \vec{b}) = \sqrt{(\sqrt{a_x^2 + a_y^2} + \sqrt{b_x^2 + b_y^2})^2 - (a_x + b_x)^2 - (a_y + b_y)^2}$ between the missing transverse energy and the lepton of the i-th element
- $E_{tr}^{(i)}$ = the total transverse energy of the i-th element

This formulation is intended to enforce energy balance by penalizing discrepancies between the combined transverse energy (including a component of the reconstructed invariant mass) and the measured transverse energy. While heuristic, it is motivated by the underlying physics: conservation of energy and momentum should ensure that the sum of the decay product energies approximates the parent particle's energy. The precise form (including the division by 2 in some formulations) should be empirically validated and might require refinement based on detector calibration and resolution studies.

The above physics regularization terms are designed to guide the model's predictions to be physically consistent with the expected kinematics of Higgs boson decays:

- The Invariant Mass Constraint ensures that the reconstructed mass clusters around 125GeV.
- The Angular Correlation Constraint enforces a target angular configuration in the transverse plane by penalizing deviations from an ideal $\pi/2$ difference.
- The Energy Balance Constraint aims to enforce energy conservation by comparing the combined transverse energy of decay products with the measured transverse energy.

These formulations are inspired by established methods in high-energy physics and the guidelines provided in the Kaggle Higgs Boson Challenge documentation [1]. However, since the exact implementation of physics constraints can vary between analyses, it is critical to validate and potentially refine these terms based on detailed experimental studies and domain expertise.

Therefore, the total loss function incorporates both supervised classification loss and physics violation penalties of Higgs decay kinematics. Then, the total loss is calculated by combining binary cross-entropy (data classification) and physics violation penalties with a regularier λ to control the influence in the total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classification}} + \lambda_{\text{mass}}\mathcal{L}_{\text{mass}} + \lambda_{\text{angle}}\mathcal{L}_{\text{angle}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}} \tag{8}$$

5 Results

It is worth mentioning that other generative methods were attempted such a Generative Adversarial Network (GAN), however, a preliminary exploration did not return the best results as seen in Figure 1.

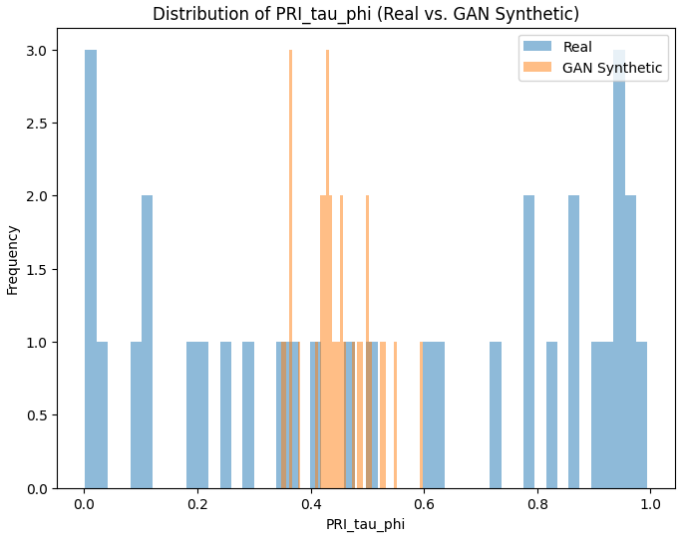

Figure 1: GAN model

Yet again, the preliminary results of the VAE weren't promising either as shown in Figure 2.

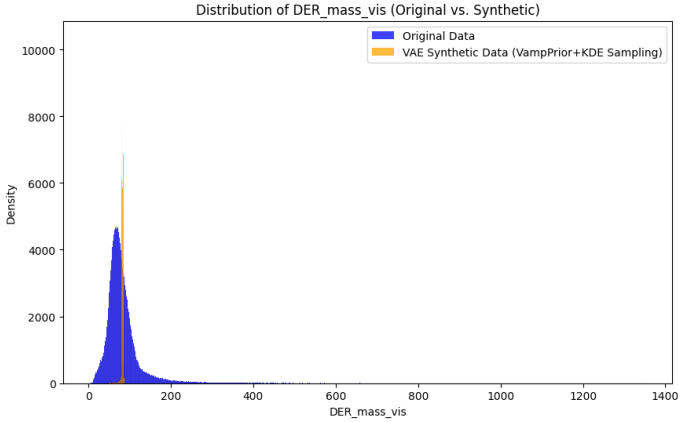

Figure 2: Standard VAE

Nevertheless, after comparing both with the real dataset, the VAE seemed the most aligned as seen in Figure 3, therefore selecting it as the go-to-model for data generation, but the question was on how to improve it when the distribuiton of each feature varied enormously from each other, which after some research the use of a VampPrior came out as a plausible solution, which after experimenting with it, turned out to be an ideal one as seen in Figure 4.

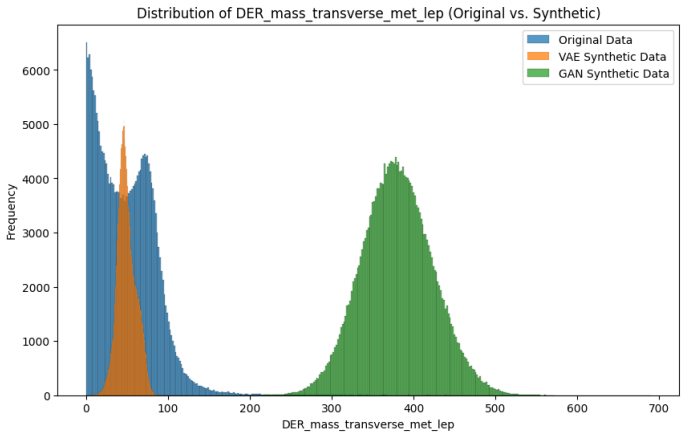

Figure 3: GAN vs VAE

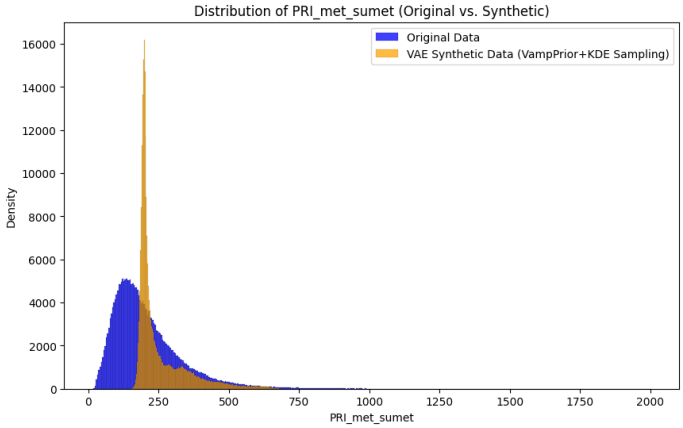

Figure 4: VampPrior and KL Divergence

But it was still not the best fit for the great diversity between features, so further research was necessary, which the MMD showed as a good option to combine it as an extra methodology with the KL Diverge to better align the posterior of

the decoder with the prior of the real dataset which was the perfect option as seen in Figure 5.

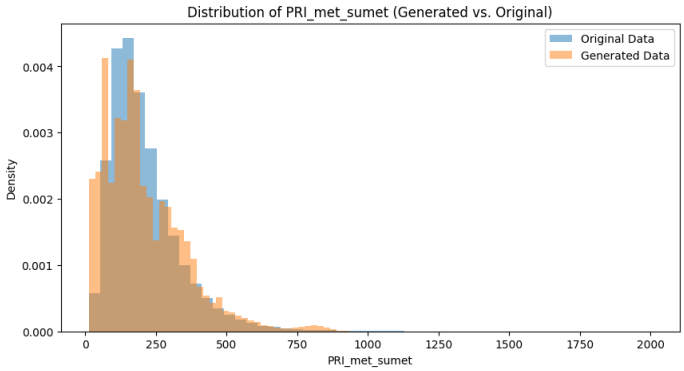

Figure 5: VampPrior, KL Divergence and MMD

Once the conflict of the correct data generation process was fixed, I was able to combine both datasets (original and synthetic) to train the PINN model. I said that the physics constraints were also regularized by a λ parameter to control the influence of each one, and after some rule-of-thumb hyperparameter tunning based on the individual losses to control those who affected the model, I found an ideal configuration over several iterations which resulted in a validation accuracy of 75% as seen in Figure 6.

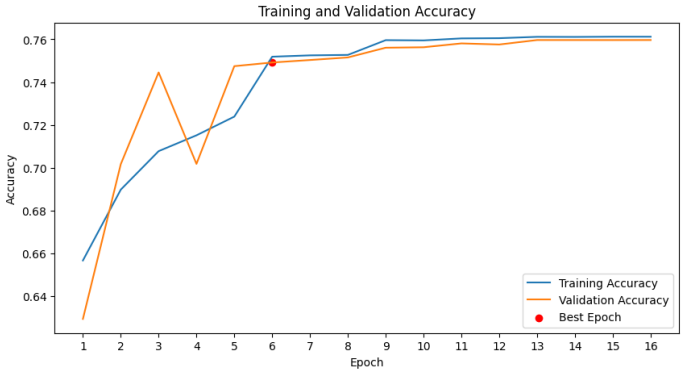

Figure 6: PINN's accuracy

When evaluated in conjunction with other metrics, the model exhibits a commendable performance, especially given its simplicity, as it comprises only three layers and a limited number of neurons. This is evident from the metrics, as the

PINN demonstrates a Receiver Operating Characteristic (ROC) curve score of 81 (Figure 7).

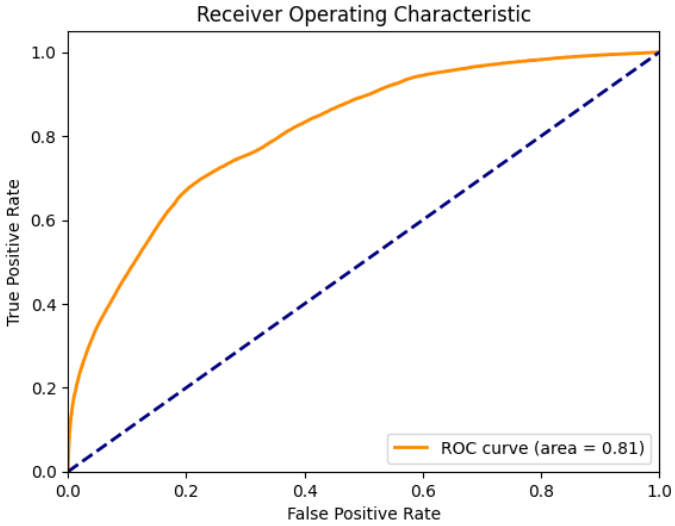

Figure 7: Area under the ROC curve

Nonetheless, there remains potential for enhancement, as corroborated by other metrics; specifically, the low recall rate of 42% indicates inefficacy in identifying the Higgs boson, despite the model achieving a commendable precision level of 74%, culminating in a moderate F1-score of 53%. This is further substantiated by the confusion matrix (Figure 8), which reveals a substantial number of instances erroneously classified as noise (nearly 20,000), which, in actuality, should be recognized as signals of the Higgs boson, in contrast to the 14,000 samples that were correctly identified. These findings, particularly in terms of precision, underscore the significant role of physics constraints in enabling the model to distinctly recognize the Higgs boson, ensuring that the predictions are congruent with the expected behavior of this particle.

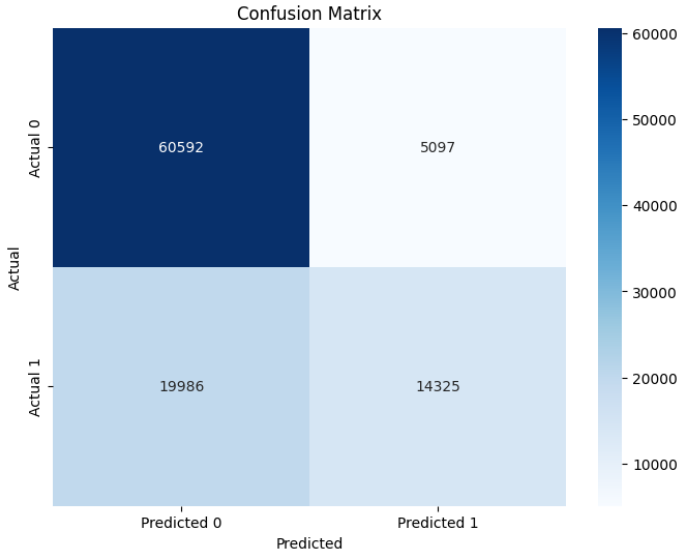

Figure 8: Confusion Matrix

6 Discussion and Future work

My work demonstrates the viability of integrating physics-informed regularization into a deep learning framework for Higgs boson signal extraction. By incorporating domain-specific constraints such as the invariant mass, angular correlation, and energy balance, I enforce key physical principles within the loss function, thereby guiding the network to produce more physically plausible predictions. The conditional VampPrior β-VAE model successfully generates synthetic data that captures the joint distribution of features and labels, while the PINN leverages physics-based regularizers to refine signal identification.

Nonetheless, as previously indicated, there is scope for further refinement through the adjustment of hyperparameters, such as the regularization (λ) associated with the physics constraints, which modulate its influence on the total loss function. For example, the penalty imposed by the angular correlation constraint, currently formulated as:

$$\mathcal{L}_{\text{angle}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i > 0.5) \cdot \left((\phi_\tau^{(i)} - \phi_{\text{lep}}^{(i)}) - \frac{\pi}{2} \right)^2 \tag{9}$$

could be further tuned or modified to better match empirical angular distributions. Similarly, the energy balance constraint, which compares the reconstructed energy components, would benefit from a more rigorous calibration,

potentially by incorporating full four-momentum conservation instead of the heuristic formulation currently used.

Additionally, the integration of supplementary constraints or alterations to the model architecture could be considered. For instance, incorporating constraints derived from a full kinematic reconstruction (using the relation $E^2 = p^2 + m^2$) or adopting hybrid architectures that blend convolutional and recurrent layers might better capture the complex spatiotemporal correlations in collider data.

Furthermore, a similar methodological approach could be applied to the cVamp-Prior β-VAE model, aimed at enhancing the generation of synthetic data to more accurately mirror the characteristics of the real data as some features don't finish to reflect the real distribution as seen in Figure 9.

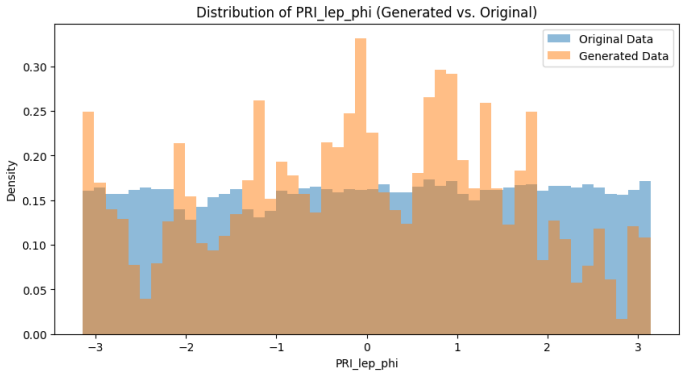

Figure 9: Future fine-tunning work

In my current approach, I have combined synthetic labels with the real dataset, bypassing separate fine-tuning; however, future studies could explore different training regimes to determine the optimal strategy. In fact, another option that is worth exploring is pre-training the generative model on real data, followed by fine-tuning with physics-informed losses, which may yield higher-quality synthetic data for subsequent training of the PINN.

In summary, while my current framework effectively marries data-driven learning with domain-specific physical constraints, significant potential remains for further improvement. Future work will focus on refining the hyperparameters, exploring alternative formulations of the physics constraints, and experimenting with advanced network architectures, all with the goal of enhancing the fidelity and robustness of Higgs boson signal extraction.

References

[1] Claire Adam-Bourdariosa, Glen Cowanb, C´ecile Germainc, Isabelle Guyond, Balazs Kegl, and David Rousseaua. Learning to discover: the higgs boson machine learning challenge. 2014.

[2] ATLAS Collaboration. Atlas experiment, cern. `https://atlas.cern/`, 2025.

[3] Universidad de Sevilla. Variational autoencoder.

[4] Hao-Zhe Feng, Kezhi Kong, Minghao Chen, Tianye Zhang, Minfeng Zhu, and Wei Chen. Shot-vae: Semi-supervised deep generative models with label-aware elbo approximations.

[5] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2013.

[6] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

[7] J. M. Tomczak and M. Welling. Vampprior: Variational mixture of posteriors prior for variational autoencoders. *arXiv preprint arXiv:1705.07120*, 2018.

[8] Onur Tunali. Maximum mean discrepancy (mmd) in machine learning. `https://www.kaggle.com/code/onurtunali/maximum-mean-discrepancy#references`, 2025.