Covariant Contrastive Learning for Uncertainty-Aware Anomaly Detection

Shelley Tong[1,2] **Gaia Grosso**[1,2] **Philip Harris**[1,2]
[1]Massachusetts Institute of Technology
[2]NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI)
Cambridge, MA
{tshelley, gaiag795, pcharris}@mit.edu

Abstract

Machine learning (ML) based anomaly detection offers a promising, model-agnostic alternative to traditional analyses of LHC data, with supervised contrastive learning providing powerful low-dimensional neural embeddings of high-dimensional physics observables. However, the propagation of systematic uncertainties through learned representations remains poorly understood, severely limiting the application of these methods to real analyses. We address this gap with a likelihood-based regularization that fine-tunes the structure of contrastive latent spaces to covary predictably with systematic distortions, returning a model of such distortions that can be directly exploited in a downstream profile-likelihood test for anomaly detection. Using an emulated CMS Level-1 trigger data, we demonstrate that our method successfully captures the effect of jet momentum scale uncertainty via a linear latent model while retaining sensitivity across signal benchmarks, improving interpretability, robustness, and scalability of the anomaly detection task.

1 Introduction

The search for new physics at the Large Hadron Collider (LHC) increasingly involves *anomaly detection* (AD) strategies: signal-agnostic approaches that identify statistically significant deviations from a well-characterized Standard Model (SM) background without committing to a specific signal hypothesis ([1–5]). Supervised contrastive learning with physics process labels has shown remarkable success in producing organized neural embeddings for AD, achieving order-of-magnitude improvements in detection power over direct feature-space methods [6].

Despite this promise, deploying such methods in real LHC analyses faces a fundamental challenge: *systematic uncertainties*. Every physics measurement must account for imperfect detector calibration and theory modeling via nuisance parameters [7], yet the propagation of systematic distortions through ML algorithms is generally uncontrolled, making a calibrated downstream statistical test an open problem.

Systematic uncertainties fall into two broad categories. *Discrete* uncertainties, that are difficult to characterize, such as the choice of parton shower generator, can be handled with *invariance*: treating same-process events from different generator configurations as positive pairs in the contrastive loss makes the embedding insensitive to the generator choice [8]. *Continuous* uncertainties, parameterized by real-valued nuisance parameters ν such as jet energy scale or luminosity, instead require the latent space to *covary* with ν in a controlled way so that ν can be profiled.

This paper proposes a method to impose this covariance structure via a fine-tuning objective that organizes the latent space so that the effect of a systematic distortion on the log-density-ratio is linear

Conference on Physics and AI at Stanford University (PAI 2026).

in ν. The resulting space admits a simple, tractable parametric model of the systematic, enabling a likelihood-ratio AD test [9] at scale. Sections 2–5 cover related work, the method, experiments on CMS Level-1 trigger simulation, and conclusions.

2 Related work

Supervised contrastive embeddings for AD. [6, 10] established that supervised contrastive losses with physics process labels produce latent spaces well-suited for signal-agnostic AD at the LHC, with domain-knowledge-informed label encoding being the key ingredient for statistical discovery performance. That work did not address systematic uncertainties, which is the gap filled here.

Equivariant contrastive learning. [11] showed that augmenting invariance-based SSL with an auxiliary head that predicts the applied transformation—encouraging non-trivial *equivariance*—improves representation quality. We adapt this idea to the physics setting, replacing geometric transformations with systematic distortions parameterized by nuisance parameters.

AD in presence of systematic uncertainties. The "New Physics Learning Machine" [12, 13], extended in [9] to incorporate continuous systematics, provides the statistical backbone of this work. Related approaches to handle systematics uncertainties include parameterized classifiers [14] and uncertainty-aware training [15]; others approach the problem imposing invariance via adversarial decorrelation [16, 17]. Our contribution is to enable [9] to operate in the latent space of a contrastive encoder, by ensuring the background model there has a tractable linear dependence on ν.

3 Method

Our statistical analysis workflow combines three main steps: (1) building a compressed neural embedding with supervised contrastive learning; (2) modeling the effect of continuous systematic uncertainties in the embedded space via a likelihood-based regularization; and (3) performing a profile-likelihood statistical test for anomaly detection. We describe these steps in turn.

(1) Supervised contrastive learning for neural embedding. We train a transformer-based neural encoder $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ ($k \ll d$) using the supervised contrastive loss [18]:

$$\mathcal{L}_{\text{supCL}} = -\sum_i \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\mathbf{z}_i \cdot \mathbf{z}_p / \tau}}{\sum_{a \neq i} e^{\mathbf{z}_i \cdot \mathbf{z}_a / \tau}}, \tag{1}$$

where $\mathbf{z}_i = f_\theta(\mathbf{x}_i)/\|f_\theta(\mathbf{x}_i)\|$ is the ℓ_2-normalized embedding, $P(i)$ indexes events produced according to the same background physics process in the batch, and τ is a temperature hyperparameter. This organizes the latent space by background physics process (data and labels come from simulation), clustering SM events together and naturally isolating potential BSM anomalies.

(2) Systematic uncertainties as a covariance problem. A continuous systematic uncertainty is parameterized by a nuisance parameter ν that induces a deterministic distortion $\mathbf{x}(\nu_0) \mapsto \mathbf{x}(\nu)$ of the observables. After encoding, this produces a transformation of the latent distribution: $\mathbf{z}(\nu_0) \mapsto \mathbf{z}(\nu)$. Without structural constraints on f_θ, this transformation can be arbitrarily nonlinear in ν, making it intractable to construct a parametric model of the log-density-ratio $r(\mathbf{z}|\nu) = {}^{n(\mathbf{z}|R_\nu)}/_{n(\mathbf{z}|R_{\nu_0})}$ as required for the statistical test described at step (4).

We therefore propose to impose a *linear covariance* condition on the latent space: the encoder should be organized so that the log-density-ratio is linear in ν,

$$\log r(\mathbf{z}|\nu) = \log \frac{n(\mathbf{z}|R_\nu)}{n(\mathbf{z}|R_{\nu_0})} = g_\phi(\mathbf{z}) \cdot \nu, \tag{2}$$

where $g_\phi(\mathbf{z})$ is a function of the learned embedding only. Under this condition, ν is straightforwardly estimable via maximum likelihood directly in the latent space, making nuisance profiling tractable.

(3) Likelihood-based fine-tuning. We enforce the linear covariance condition (2) by fine-tuning the encoder jointly with a model g_ϕ of the log-density-ratio. We generate labeled pairs $(\mathbf{x}(\nu_0), \mathbf{x}(\nu))$ for a discrete grid of ν values spanning the expected uncertainty range $[-\sigma, +\sigma]$, and train g_ϕ as a parameterized classifier between the nominal and shifted distributions following [9]. Specifically, we

minimize the quadratics loss

$$\mathcal{L}_{\text{cov}} = -\frac{1}{N_\nu} \sum_\nu \left[\sum_{\mathbf{x}\sim R_\nu} \sigma(g_\phi(f_\theta(\mathbf{x})) \cdot \nu)^2 + \sum_{\mathbf{x}\sim R_{\nu_0}} [1 - \sigma(g_\phi(f_\theta(\mathbf{x})) \cdot \nu)]^2 \right], \tag{3}$$

where σ denotes the sigmoid function and the sum runs over the grid of nuisance values. By the likelihood-ratio argument of [9], the minimizer of this loss satisfies $g_\phi(\mathbf{z}) \cdot \nu \to \log r(\mathbf{z}|\nu)$, so training g_ϕ directly implements the approximation in (2). The encoder and classifier are then jointly optimized with the combined objective $\mathcal{L} = \mathcal{L}_{\text{supCL}} + \alpha\, \mathcal{L}_{\text{cov}}$, where $\alpha > 0$ controls the relative weight of the two terms. Backpropagating through g_ϕ into f_θ incentivizes the encoder to concentrate the systematic effect into latent directions that are linearly aligned with ν, yielding a tractable and interpretable distortion model. After the joint optimization has converged the embedding is frozen and the linear model is further fine-tuned to reach higher levels of accuracy, requried for step (4).

(4) Statistical anomaly detection in the regulated latent space. The fine-tuned encoder and distortion model g_ϕ integrate directly into the profile-likelihood AD framework of [9]. Following that strategy, we compute the standard profile-likelihood ratio test

$$t(\mathcal{D}) = 2 \log \frac{\max_{\nu,\,\zeta} \left[\mathcal{L}(H_{\nu,\zeta} \,|\, \mathcal{D}) \cdot \mathcal{L}_{\text{aux}}(\nu)\right]}{\max_{\nu'} \left[\mathcal{L}(R_{\nu'} \,|\, \mathcal{D}) \cdot \mathcal{L}_{\text{aux}}(\nu')\right]}, \tag{4}$$

where $\mathcal{L}_{\text{aux}}(\nu)$ is an auxiliary Gaussian likelihood term constraining the nuisance parameter. The reference SM hypothesis, R_ν, is modeled according to Eq. 2: $n(x|R_\nu) = n(x|R_{\nu_0}) \cdot e^{g_\phi(\mathbf{z})\cdot\nu}$. The alternative hypothesis $H_{\nu,w}$ is modeled with an additional neural network h_ζ $n(x|H_{\nu,\zeta}) = n(x|R_{\nu_0}) \cdot e^{g_\phi(\mathbf{z})\cdot\nu + h_\zeta(\mathbf{z})}$.

4 Application

Dataset. We validate the method on simulated CMS Level-1 trigger data [19] available on Zenodo [20–24]. Each event is represented by up to 19 reconstructed objects: 4 electrons/photons, 4 muons, 10 jets, and one missing transverse energy (MET) slot. Each object is described by (p_T, η, ϕ), yielding 57 input features per event, with zero padding applied when fewer objects are present. The background sample consists of W (59.2%), QCD multijet (33.8%), Z (6.7%), and $t\bar{t}$ (0.3%) events, while the signal benchmarks are LQ $\to b\tau$, $A \to 4\ell$, $h^0 \to \tau\tau$, and $h^\pm \to \tau^\pm\nu$. To model systematic effects, we introduce a nuisance parameter ν that rescales the jet transverse momentum as $p_T(\nu) = e^\nu p_T(0)$, with nominal value $\nu_0 = 0$, and we set the standard uncertainty $\sigma = 2.5\%$, a typical percent-level size of jet energy scale uncertainties in LHC analyses [25].

Embedding Model. Following Ref. [6], we use a Transformer-based model to map 57-dimensional input into a 4-dimensional embedding, which gave the strongest AD performance among the tested dimensions. We leave the systematic study of nuisance uncertainty in higher dimensional embeddings to future work. Nominal SM background events are used for the contrastive task with physics labels. The model is trained in a 5-fold co-training setup with a linear model (2.1×10^6 events per fold per ν dataset, 60 epochs, batch size 1024, learning rate 2×10^{-6}). Although co-training takes longer than standard contrastive training, it is only trained once and frozen in NPLM tests. The covariance loss plateaus in training, while linear nuisance model benefits from longer fine-tuning.

Linear Model. We co-train a downstream linear model on the learned embeddings. The model is implemented as a fully connected network with architecture $(4, 512, 256, 128, 64, 32, 1)$, ReLU activations in the hidden layers, and a linear output. It is trained jointly with the supervised contrastive embedding model using a learning rate of 2×10^{-6}, L_2 regularization of 10^{-6}, and co-training regularization $\alpha = 0.1$, then further fine-tuned on 1.6×10^6 events per nuisance parameter. Both nominal and nuisance-shifted samples at $\pm 1\sigma$ and $\pm 2\sigma$ are used for this task.

NPLM Model. h_ζ is implemented as a fully connected neural network with architecture $(4, 4, 4, 1)$ (45 trainable parameters), sigmoid activation in the hidden layers and linear output. As in [13], a weight clipping regularizer is tuned using 50 pseudo-experiments drawn from the nominal reference sample R_0 to produce an empirical test statistic distribution matching a χ^2_{45}. We train using 52,000 reference events and 10,000 background events (D) per replica. The optimal weight clipping is 1.96.

NPLM Validation Tests. We validate the calibrated model on nuisance-shifted SM samples $D \sim R_\nu$ using the same event counts. The feature-wise comparison between the nuisance-shifted samples and

the fitted nuisance model is shown in Figure 1. We validate the nuisance dependence of the embedding distributions by comparing the $\nu = \pm 1\sigma$ and nominal R_0 background-only distributions using a KS test. We use 200 replicas per ν value. As shown in Table 1, all p-values are greater than 0.5, indicating good agreement. This shows that the linear nuisance model captures the nuisance-induced deformation of the embedding space, rather than misidentifying it as new-physics-like structure.

Signal Injection. We assess the sensitivity of the method by injecting a 1% signal component (100 events) into the SM background sample and testing for departures from the null expectation. We compute the resulting z-score with respect to the empirical test-statistic distribution under R_0, estimated from 200 replica. A large deviation from the null indicates excess structure not accounted for by systematic effects alone, consistent with a detectable new-physics-like signal. The resulting z-scores for each injected signal benchmark are summarized in Table 1. The sensitivity of the test is independent of the ν setting.

Compute Resources. Embedding models were trained on 48 GB GPU, with a typical training time of 4 hours 45 minutes. Linear-model fine-tuning was performed on 48 GB GPU, with a typical training time of 3 hours. The NPLM model was trained on CPUs, with a typical training time of 50 minutes.

Code Availability. The code and detailed documentation to reproduce the experiments can be found at `https://github.com/tshelley200/cl4ad_uncertainty`.

5 Conclusion

We introduced a method to impose linear covariance of continuous systematic uncertainties in contrastive latent spaces, enabling rigorous profile-likelihood anomaly detection at the LHC. By fine-tuning the encoder with an auxiliary objective that enforces the log-density-ratio to be linear in the nuisance parameter, the method integrates directly with the AD framework of [9], yielding a scalable and statistically well-founded workflow for anomaly detection in the presence of continuous systematics. Future work will extend the method to multiple simultaneous nuisance parameters, bringing contrastive AD pipelines closer to deployment in full-scale analyses.

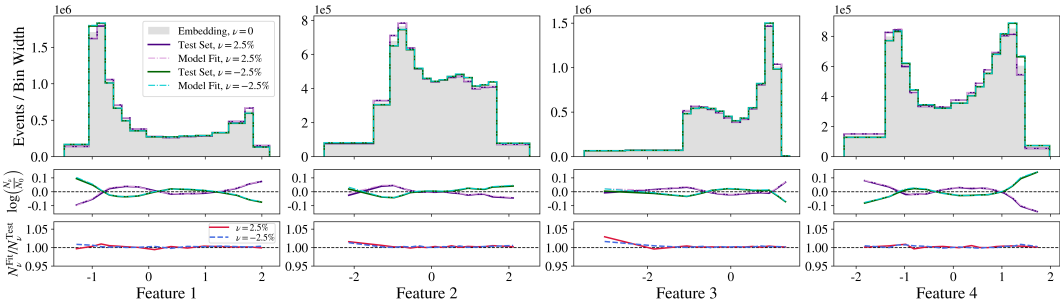

Figure 1: **Validation of the linear nuisance model in the latent space.** Top: feature distributions for the reference sample at $\nu = 0$ (gray fill), nuisance-shifted test samples (solid lines), and fitted model predictions at $\nu = \pm 1\sigma$ (dash-dotted lines). Middle: corresponding $\log(N_\nu/N_0)$ ratios. Bottom: ratios of fitted model to test sample, $N_\nu^{\text{fit}}/N_\nu^{\text{test}}$.

Nuisance setting	KS p-value	z-score for injected signal benchmark			
		LQ $\to b\tau$	$A \to 4\ell$	$h^0 \to \tau\tau$	$h^\pm \to \tau^\pm\nu$
$\nu = 0$	—	$1.31^{+0.29}_{-0.23}$	≥ 2.58	$0.68^{+0.21}_{-0.18}$	≥ 2.58
$\nu = +0.025$	$0.79^{+0.1}_{-0.01}$	$1.26^{+0.28}_{-0.22}$	≥ 2.58	$0.68^{+0.21}_{-0.18}$	≥ 2.58
$\nu = -0.025$	$0.79^{+0.1}_{-0.01}$	$1.26^{+0.28}_{-0.22}$	≥ 2.58	$0.57^{+0.20}_{-0.18}$	≥ 2.58

Table 1: **Validation and signal-injection results.** The second column gives the KS-test p-value for the SM-only validation test, comparing the empirical test-statistic distribution at $\nu = \pm 1\sigma$ to the $\nu = 0$ background-only reference. The remaining columns give the median observed z-scores for each injected signal benchmark. Errors represent the Clopper-Pearson 68% confidence intervals.

References

[1] Vladimir Chekhovsky et al. Model-agnostic search for dijet resonances with anomalous jet substructure in proton–proton collisions at $\sqrt{s}$ = 13 TeV. *Rept. Prog. Phys.*, 88(6):067802, 2025. doi: 10.1088/1361-6633/add762.

[2] Aram Hayrapetyan et al. Machine-learning techniques for model-independent searches in dijet final states. 12 2025. doi: 10.5281/zenodo.16656501.

[3] Georges Aad et al. Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. *Phys. Rev. Lett.*, 125(13):131801, 2020. doi: 10.1103/PhysRevLett.125.131801.

[4] Georges Aad et al. Weakly supervised anomaly detection for resonant new physics in the dijet final state using proton-proton collisions at s=13 TeV with the ATLAS detector. *Phys. Rev. D*, 112(7):072009, 2025. doi: 10.1103/2yq5-vj59.

[5] Georges Aad et al. Anomaly detection search for new resonances decaying into a Higgs boson and a generic new particle X in hadronic final states using $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector. *Phys. Rev. D*, 108:052009, 2023. doi: 10.1103/PhysRevD.108.052009.

[6] Kyle Metzger, Lana Xu, Mia Sodini, Thea K. Arrestad, Katya Govorkova, Gaia Grosso, and Philip Harris. Anomaly-preserving contrastive neural embeddings for end-to-end model-independent searches at the LHC. *Phys. Rev. D*, 112(7):072011, 2025. doi: 10.1103/5n77-ynsp.

[7] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, 71:1554, 2011. doi: 10.1140/epjc/s10052-011-1554-0. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

[8] Philip Harris, Michael Kagan, Jeffrey Krupa, Benedikt Maier, and Nathaniel Woodward. Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models. 3 2024.

[9] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning new physics from an imperfect machine. *Eur. Phys. J. C*, 82(3):275, 2022. doi: 10.1140/epjc/s10052-022-10226-y.

[10] Samuel Bright-Thonney, Christina Reissel, Gaia Grosso, Nathaniel Woodward, Katya Govorkova, Andrzej Novak, Sang Eon Park, Eric Moreno, and Philip Harris. Autoscidact: Automated scientific discovery through contrastive embedding and hypothesis testing. *arXiv preprint arXiv:2510.21935*, 2025.

[11] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021.

[12] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning new physics from a machine. *Physical Review D*, 99(1):015014, 2019.

[13] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning multivariate new physics. *Eur. Phys. J. C*, 81(1):89, 2021. doi: 10.1140/epjc/s10052-021-08853-y.

[14] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5):235, 2016.

[15] Pablo De Castro and Tommaso Dorigo. Inferno: inference-aware neural optimisation. *Computer Physics Communications*, 244:170–179, 2019.

[16] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Advances in neural information processing systems*, 30, 2017.

[17] Gregor Kasieczka and David Shih. Disco fever: Robust networks through distance correlation. *arXiv preprint arXiv:2001.05310*, 2020.

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[19] Ekaterina Govorkova, Ema Puljak, Thea Aarrestad, Maurizio Pierini, Kinga Anna Woźniak, and Jennifer Ngadiuba. Lhc physics dataset for unsupervised new physics detection at 40 mhz. *Scientific Data*, 9(1):118, 2022.

[20] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised new physics detection at 40 mhz: Training dataset, June 2021. URL `https://doi.org/10.5281/zenodo.5046389`.

[21] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised new physics detection at 40 mhz: A -> 4 leptons signal benchmark dataset, June 2021. URL `https://doi.org/10.5281/zenodo.5046446`.

[22] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised new physics detection at 40 mhz: Lq -> b tau signal benchmark dataset, July 2021. URL `https://doi.org/10.5281/zenodo.5055454`.

[23] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised new physics detection at 40 mhz: h0 -> tau tau signal benchmark dataset, July 2021. URL `https://doi.org/10.5281/zenodo.5061633`.

[24] Thea Aarrestad, Ekaterina Govorkova, Jennifer Ngadiuba, Ema Puljak, Maurizio Pierini, and Kinga Anna Wozniak. Unsupervised new physics detection at 40 mhz: h+ -> tau nu signal benchmark dataset, July 2021. URL `https://doi.org/10.5281/zenodo.5061688`.

[25] ATLAS Collaboration. A precise measurement of the jet energy scale derived from single-particle measurements and in situ techniques in proton-proton collisions at 13 TeV with the ATLAS detector. 2024.

Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contribution—a likelihood-based regularization that imposes linear covariance of continuous systematics in contrastive latent spaces—and the scope is accurately described as a proof-of-concept on emulated CMS Level-1 trigger data with a single nuisance parameter (jet momentum scale). The experimental results in Section 4 directly support these claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The conclusion acknowledges that the method has been validated for a single continuous nuisance parameter and that extension to multiple simultaneous nuisance parameters is left for future work. The experiments are conducted on a single benchmark dataset with four signal benchmarks, and the paper does not claim generalization beyond this setting.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain formal theorems or proofs. The theoretical motivation for the loss function (Eq. 3) relies on the likelihood-ratio training argument of [9], which is cited and described qualitatively.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The dataset [19] is publicly available. Architecture details, hyperparameters, training setup, and evaluation procedure are fully specified in Section 4 for all three models (embedding, linear model, NPLM). The nuisance parameter definition and the validation protocol are described in detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset is publicly available [19]. Code is available at an anonymized repository linked in the paper and will be made public upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 reports dataset composition, input representation, nuisance parameter definition, model architectures, optimizer settings, learning rates, regularization strengths, batch sizes, number of epochs, co-training setup, and evaluation procedure for all models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 reports KS p-values with Clopper-Pearson 68% confidence intervals, and z-scores are computed with respect to empirical test-statistic distributions estimated from 200 toy samples. The statistical procedure is described in the NPLM Validation Tests paragraph. Also in this case Clopper-Pearson 68% confidence intervals are reported.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are reported in Section 4 under the paragraph "compute resources"

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: This work concerns fundamental physics research with no foreseeable negative societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is foundational research in particle physics methodology. It has no direct path to negative societal applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models and dataset involved in this work pose no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset [19] is cited. All software packages and prior methods used are cited in the bibliography.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is provided in an anonymized repository linked in the paper, with documentation. The repository will be made fully public upon acceptance.

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper involves only simulated physics data; no human subjects are involved.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects are involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer:

    Answer: [NA]

    Justification: LLMs were not used as part of the core methodology. They were used only for writing assistance, which does not require declaration under NeurIPS policy.

    Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.