AutoSciDACT: Automated Scientific Discovery through Contrastive Embedding and Hypothesis Testing

Samuel Bright-Thonney[1,2] **Christina Reissel**[1] **Gaia Grosso**[1,2] **Nathaniel Woodward**[1,3] **Katya Govorkova**[1] **Andrzej Novak**[1] **Sang Eon Park**[1] **Eric Moreno**[1] **Philip Harris**[1,2]
[1]Department of Physics, Massachusetts Institute of Technology
[2] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[3] Department of Physics, University of Wisconsin, Madison

Abstract

Novelty detection in large scientific datasets faces two key challenges: the noisy and high-dimensional nature of experimental data, and the necessity of making *statistically robust* statements about any observed outliers. While there is a wealth of literature on anomaly detection via dimensionality reduction, most methods do not produce outputs compatible with quantifiable claims of scientific discovery. In this work we directly address these challenges, presenting the first step towards a unified pipeline for novelty detection adapted for the rigorous statistical demands of science. We introduce AutoSciDACT (Automated Scientific Discovery with Anomalous Contrastive Testing), a general-purpose pipeline for detecting novelty in scientific data. AutoSciDACT begins by creating expressive low-dimensional data representations using a contrastive pre-training, leveraging the abundance of high-quality simulated data in many scientific domains alongside expertise that can guide principled data augmentation strategies. These compact embeddings then enable an extremely sensitive machine learning-based two-sample test using the New Physics Learning Machine (NPLM) framework, which identifies and statistically quantifies deviations in observed data relative to a reference distribution (null hypothesis). We perform experiments across a range of astronomical, physical, biological, image, and synthetic datasets, demonstrating strong sensitivity to small injections of anomalous data across all domains.

1 Introduction

Scientific discovery is often characterized by serendipity: an unexpected observation turns out to have a profound impact on a field, leading to rapid progress or discovery. Today's data-rich scientific landscape is potentially brimming with curious or unexplained observations, but the scale and complexity of available data increasingly obscures genuine novelties behind statistical noise or incidental fluctuations. As scientific datasets continue to grow, so too grows the challenge of uncovering meaningful unexplained phenomena.

The scientific method traditionally provides a structured framework for discovery, encompassing observation, inquiry, research, hypothesis formulation, experimentation, and conclusion (top row of Fig. 1). Effective implementation of this method relies on human intuition and domain expertise to identify relevant observables and devise meaningful experiments. However, given the magnitude of modern datasets, the process would significantly benefit from automated tools to efficiently identify the most promising regions for discovery.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

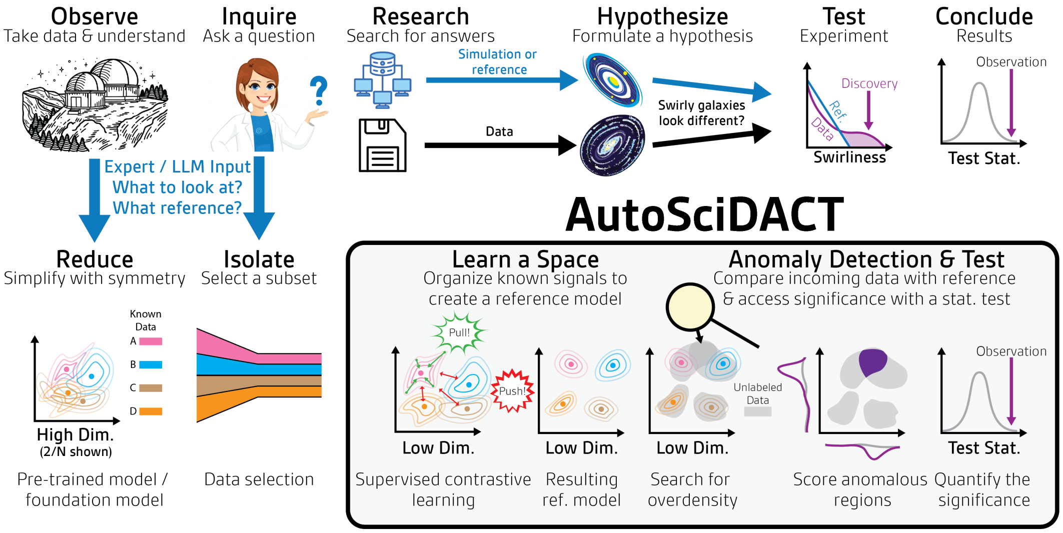


Figure 1: Illustration of the scientific method (top row) and the AutoSciDACT pipeline (bottom row), emphasizing the corresponding methodological steps implemented within AutoSciDACT.

To address this challenge, it is essential to develop methods that intelligently prioritize informative regions—areas where genuine scientific surprises are most likely to emerge. Traditional feature-engineering approaches are human-driven and domain-specific, limiting scalability and generalizability. Recent advances utilizing agentic AI systems and large language models can partially automate aspects of scientific inquiry [1–3], yet still lack integrated frameworks capable of rigorous, automated hypothesis testing and validation.

We introduce AutoSciDACT (Automated Scientific Discovery with Anomalous Contrastive Testing), a pipeline that parallels the scientific method and streamlines scientific inquiry by automating key steps of the scientific discovery process. AutoSciDACT streamlines the phases of data reduction, hypothesis formulation, and statistical testing (bottom row of Fig. 1) by deploying contrastive learning together with the New Physics Learning Machine (NPLM) [4]. Contrastive learning is used to reduce raw, high-dimensional datasets into expressive low-dimensional feature embeddings while NPLM provides a statistically robust mechanism for identifying and quantifying novel structures within this learned embedding space. A key insight with AutoSciDACT is the effective, automated incorporation of domain expertise as a tool to reduce the dimensionality to a small number of well-behaved features, making it possible to construct a robust statistical model. Using NPLM, our pipeline systematically compares incoming data with reference distributions of known (background) data, finding the most anomalous regions and quantifying their statistical significance.

We validate our approach using synthetic benchmarks and real-world datasets from astronomy, physics and biomedical domains. AutoSciDACT reliably detects meaningful novelties while remaining robust against spurious variations. Our results demonstrate the promise of combining structured contrastive learning methods with automated statistical hypothesis testing to accelerate scientific discoveries.

Contributions Our main contributions can be summarized as follows:

- An end-to-end pipeline for novelty discovery in scientific datasets that is readily transferable across domains.
- A principled procedure for incorporating scientific simulations, hand-labeled data, and expert knowledge into a contrastive dimensionality reduction pipeline.
- A statistically rigorous framework for *quantifying the significance* of observed anomalies, beyond simply flagging anomalous datapoints.
- A realistic demonstration of novelty detection in four disparate scientific domains.

2 Related Work

Contrastive Learning Contrastive learning is a powerful tool for learning expressive, low-dimensional data representations. Self-supervised methods such as SimCLR [5], MoCo [6], VICReg [7], and Barlow Twins [8] use data augmentations (e.g., blurring, cropping) to promote semantically meaningful and well-separated embeddings. Supervised contrastive learning [9] uses class labels to define positive pairs, more efficiently capturing semantic relationships between datapoints but requiring labeled datasets to train. Its applications have expanded to structured data [10], multimodal inputs [11], and domain-specific tasks [12], demonstrating broad utility.

Contrastive Anomaly Detection Several existing methods leverage contrastive embeddings to search for out-of-distribution (OOD) data, but primarily focus on identifying *individual* anomalous instances for e.g. industrial applications. In contrast, AutoSciDACT (our method) is tailored for scientific contexts and makes *statistical* statements about the presence of anomalous data, identifying distribution-level deviations relative to a baseline expectation (null hypothesis). Existing approaches include Refs. [13–21], all of which rely primarily on the AUROC for identifying OOD points as a figure of merit. They do not attempt to statistically quantify observations of OOD data, as is required in the scientific context. CADet [13] is nearest to our setup in using a two-sample test, but focuses again on AUROC. In scientific contexts, various combinations of contrastive learning and anomaly detection have been used for domain-specific applications – e.g. in astronomy [22, 23], histology, [24, 25] and particle physics [26, 27] – but, to our knowledge, no unified approaches have been proposed.

Hypothesis Testing for Anomaly Detection Traditional goodness-of-fit (GOF) tests are powerful in univariate settings but struggle beyond that due to the curse of dimensionality. Simple multivariate extensions have limiting factors. The Mahalanobis distance-based test [28], for instance, assumes Gaussianity and is sensitive to the choice of covariance estimation, limiting its applicability in complex data regimes. Recent machine learning-based methods have introduced model-agnostic, data-driven alternatives to enable non-parametric, highly adaptable high-dimensional tests. To quantify distributional differences, Maximum Mean Discrepancy (MMD) [29–31] embeds distributions into a reproducing kernel Hilbert space, while the Classifier Two-Sample Test (C2ST) [32] uses a trained classifier's accuracy. Other methods include density-ratio estimation [33] and generative modeling frameworks that assess sample likelihoods [34]. In this work, we draw from statistical anomaly detection tests developed for high-energy physics, where sensitivity to subtle deviations is crucial. Autoencoders [35–46] and semi-supervised binary classifiers [47] have been widely used to score anomalies, but not statistically test them. The NPLM algorithm [48, 49] was introduced as an end-to-end score-and-test tool, outperforming classic GOF and classifier-based tests [50], showing sensitivity to a wide class of anomalies, and allowing incorporation of systematic (epistemic) uncertainties [51].

3 The AutoSciDACT Pipeline

The AutoSciDACT pipeline consists of two phases: pre-training and discovery. The aim of the pre-training phase is to learn an expressive, low-dimensional representation of a scientific dataset that retains key semantic features while reducing potentially hundreds or thousands of input dimensions to a handful. The discovery phase uses these embeddings in the NPLM anomaly detection and hypothesis testing framework to search for novelty in a scientific dataset. AutoSciDACT is designed for discovering *statistically significant* anomalies, prioritizing detection of distributional shifts (e.g. overdensities, distortions, outlier clusters) with respect to a background-only hypothesis, rather than instance-level anomalies. The power of NPLM (and any statistical test) degrades with data dimensionality, quickly requiring prohibitively large sample sizes to make statistically significant observations of small signals. As such, the reduction in the pre-training phase is critical. The bottom row of Fig. 1 summarizes the key steps and features of AutoSciDACT.

3.1 Pre-Training: Contrastive Embeddings

The backbone of our pipeline is an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ trained with contrastive learning to map raw data from its high-dimensional input space $\mathcal{X}$ to a low-dimensional representation in $\mathbb{R}^d$. Contrastive objectives are designed to maximize alignment between like inputs (positive pairs) while separating unlike inputs (negative pairs) in the learned space. We use the SimCLR framework [5], which trains an encoder f_θ alongside a projection head g_ϕ (typically a small MLP) with the following contrastive loss:

$$\mathcal{L}_{\text{SimCLR}} = -\sum_{i\in\mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \tilde{\mathbf{z}}_i)/\tau)}{\sum_{j\neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}, \tag{1}$$

where $\mathbf{z} = g_\phi(f_\theta(\mathbf{x}))$, $(\mathbf{z}_i, \tilde{\mathbf{z}}_i)$ are a positive pair, $\text{sim}(\cdot,\cdot)$ is the cosine similarity, τ is a configurable temperature, and the sum in the denominator runs over all other pairings in a batch $\mathcal{B}$. Traditionally, positive pairs $(\mathbf{x}_i, \tilde{\mathbf{x}}_i)$ are constructed on-the-fly from inputs $\mathbf{x}_i \in \mathcal{B}$ using random augmentations that preserve semantic meaning. The projection head g_ϕ is discarded after training, with embeddings $\mathbf{h} = f_\theta(\mathbf{x})$ used for downstream tasks.

In AutoSciDACT we use *supervised* contrastive learning (SupCon) [9], which leverages labeled training data to create positive pairs from the same class and negative pairs from different classes. The training objective is a simple generalization of the SimCLR loss:

$$\mathcal{L}_{\text{SupCon}} = -\sum_{i\in\mathcal{B}} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p)/\tau)}{\sum_{j\neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}, \tag{2}$$

where $P(i)$ is the set of all positive (same-class) pairs of input i in the batch. Using labels to define $P(i)$ encodes a much richer notion of similarity from the full spectrum of a given input class, rather than having to indirectly learn (or fail to learn) important features from views of individual inputs. This also avoids the ill-defined question of identifying the "best" augmentations that promote expressive learned features. Practitioners in many scientific domains have ready access to large quantities of labeled training data from high-quality simulations or expert-labeled databases, so requiring labels is not often a significant bottleneck. When augmentations are desirable to encourage learning scientifically relevant meta-features (e.g. Lorentz invariance for particle physics datasets), or if class labels are unavailable, SupCon can easily incorporate augmented views in the positive set $P(i)$. In conjunction with labels, this offers a way to inject additional scientific domain knowledge via tailored augmentations.

In addition to the contrastive objective $\mathcal{L}_{\text{SupCon}}$ we include an optional supervised cross-entropy loss $\mathcal{L}_{\text{CE}}$, which we found beneficial for learning embeddings with a more regular structure and class separation. Our full loss function is thus

$$\mathcal{L} = \mathcal{L}_{\text{SupCon}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}, \tag{3}$$

where $\lambda_{\text{CE}} \sim 0.1$ - 0.5 is set to make the classification objective sub-dominant.

3.2 Discovery: Anomaly Detection & Hypothesis Testing

In the discovery phase we use the embedding f_θ to process unseen datasets and search for anomalous clusters, overdensities, or outliers in the low-dimensional space. The search process is a classic scientific hypothesis test: a **reference** dataset $\mathcal{R}$ composed of known backgrounds is compared to an **observed** dataset $\mathcal{D}$ of unknown composition, and we seek to accept or reject the null hypothesis that $\mathcal{R}$ and $\mathcal{D}$ are identically distributed (i.e. there are no new phenomena in the observed data). We implement this test with NPLM, which in conjunction with the expressive learned embeddings enables extraordinary sensitivity to new signals.

The NPLM algorithm NPLM builds on the classical likelihood ratio test introduced by Neyman et al. [52], using a test statistic defined as:

$$t(\mathcal{D}) = 2\max_{\boldsymbol{w}} \sum_{x\in\mathcal{D}} \log \frac{\mathcal{L}(x|\mathcal{H}_{\boldsymbol{w}})}{\mathcal{L}(x|\mathcal{H}_{\mathbf{0}})}\,. \tag{4}$$

A trainable model $f_{\boldsymbol{w}}(x)$ parametrizes a family of alternative hypotheses $\mathcal{H}_{\boldsymbol{w}}$ with respect to the null $\mathcal{H}_{\mathbf{0}}$ on inputs $x \in \mathbb{R}^d$, with a corresponding alternative density of the form:

$$p(x|\mathcal{H}_{\boldsymbol{w}}) = p(x|\mathcal{H}_0)\,\exp[f_{\boldsymbol{w}}(x)]\,. \tag{5}$$

This formulation enables a signal-agnostic approach: instead of specifying a particular signal model, the algorithm learns the deviation directly from data by solving a maximum likelihood problem reframed as a machine learning task. We follow the model introduced in [4], where the problem is solved as a binary classification between the data of interest $\mathcal{D}$, labeled $y = 1$, and the reference sample $\mathcal{R}$, labeled $y = 0$. The model is a Nyström approximated kernel method

$$f_{\boldsymbol{w}} = \sum_{i=1}^{\mathrm{M}} w_i k_i(x) \tag{6}$$

with $\mathrm{M} \sim \sqrt{|\mathcal{D}| + |\mathcal{R}|}$ Gaussian kernels k_i and trainable mixture coefficients $\{w_i \in \mathbb{R}\}_{i=1}^{\mathrm{M}}$, and minimizing a regularized weighted binary cross-entropy

$$\mathcal{L}_{\mathrm{NPLM}}[f_{\boldsymbol{w}}] = \sum_{(x,y)} \left[w_{\mathcal{R}}(1-y)\log\left(1 + e^{f_{\boldsymbol{w}}}\right) + y\log\left(1 + e^{-f_{\boldsymbol{w}}}\right)\right] + \lambda \sum_{i,j} w_i w_j k_i(x_j) \tag{7}$$

To ensure robustness, the size of the reference sample $|\mathcal{R}|$ is chosen to be substantially larger than the data $|\mathcal{D}|$ and is reweighted so that the expected yield under $\mathcal{H}_0$ matches the expected experimental one, which may differ from the observed size $|\mathcal{D}|$. This design choice makes the test sensitive to both shape and normalization deviations.[1]

Once the training is complete, the test statistic is estimated from the solution $f_{\hat{\boldsymbol{w}}}$ as

$$t_{\mathrm{NP}}(\mathcal{D}) = -2\left(\sum_{(x,y)} w_{\mathcal{R}}(1-y)\left(e^{f_{\hat{\boldsymbol{w}}}(x)} - 1\right) - y f_{\hat{\boldsymbol{w}}}(x)\right) \tag{8}$$

To calibrate the test, we estimate the distribution $p(t_{\mathrm{NP}}|\mathcal{H}_0)$ via pseudo-experiments ("toys"), generating (i.e. sampling from a larger pool) datasets under the null hypothesis and computing the corresponding test statistics to form an empirical distribution $\mathcal{T}_0$. The p-value is then evaluated empirically as:

$$p = \frac{1}{|\mathcal{T}_0|}\sum_{t\in\mathcal{T}_0} \mathbb{I}[t > t(\mathcal{D})]\,. \tag{9}$$

It can also be estimated asymptotically from the distribution of $\mathcal{T}_0$, which can be fit to a suitable χ^2 distribution [4]. The asymptotic estimate is useful in cases where the deviation of $\mathcal{D}$ with respect to $\mathcal{R}$ is large (e.g. $Z = 5\sigma$), in which case the number of toys required for the empirical estimate would be prohibitively large.

The power of the NPLM test strongly depends on the choice of the kernels' width, as it determines the scale of distortions the model is sensitive to. To mitigate this feature and make the model more robust, we adopt an extended version of the algorithm introduced in [53], where multiple widths are considered and combined to obtain a final p-value. The authors of [53] explore several options for combining the tests based on "local" p-values, but in this work we choose the average of p-values as a rule. The average score is typically less powerful than the single "optimal" kernel, which can be considered a kind of "look-elsewhere" effect accounting for various kernel hypotheses.

We consider six different kernel widths for our experiments, with their precise numerical values chosen according to the distribution of pairwise distances between data points in the embedding space. More precisely, the first five values are the 1st, 25th, 50th, 75th, and 99th percentiles of the empirical pairwise distance distribution (computed with a subset data points from the training set); the last value is twice the 99th percentile, and it ensures sensitivity to out-of-distribution anomalies. This choice means the numerical kernel widths vary among datasets, so we denote them by their corresponding quantiles: $\sigma_{\mathrm{ker}} \in \{q_1, q_{25}, q_{50}, q_{75}, q_{99}, 2q_{99}\}$.

The NPLM procedure provides a flexible, multivariate, unbinned likelihood-ratio test that is agnostic to the source of the anomaly, making it well-suited for unsupervised anomaly detection tasks. Comparisons with alternative GoF approaches presented in [50] show the impressive sensitivity of the method to subtle distortions of the data density distribution. As for any GoF approach relying on density estimation from empirical samples, limitations arise when scaling the data dimensionality. In this work, we target the curse of dimensionality by compressing high-dimensional raw data with contrastive embeddings.

[1] This sensitivity is important in contexts where data collection windows determine $|\mathcal{D}|$, and deviations in event rates may signal anomalies.

4 Datasets

We demonstrate AutoSciDACT on a diverse collection of five synthetic, image, and scientific datasets. Each dataset contains a large collection of data from "background" (i.e. well-understood) classes that are used in the contrastive pre-training phase, along with a set of anomalous "signal" data. In the discovery phase we construct datasets $\mathcal{D}$ with small signal injections and use NPLM to detect the novel component. We briefly describe each dataset below, with additional details available in App. A. Due to space constraints, we defer two further studies to the appendix: a genomics task identifying novel butterfly hybrids from wing images (App. A.6), and searching for four-lepton decays of the Higgs boson in real LHC data (App. E).

Synthetic Data The synthetic dataset is designed to illustrate the core functionality of AutoSciDACT independent of details specific to scientific datasets. It consists of points $\mathcal{X} \subset \mathbb{R}^{D+M}$ with D meaningful dimensions and M noisy dimensions. The noisy dimensions are sampled from $\mathcal{U}(0, 1)$, and the meaningful dimensions are populated by N Gaussian clusters $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ $(i = 1, \ldots, N)$ with means $\boldsymbol{\mu}_i \sim \mathcal{U}(0, 1)$ and covariances $\boldsymbol{\Sigma}_i \sim \mathcal{U}(0, 0.5)$. The pairwise distances of the Gaussian clusters are adjusted such that a 1% injection of one cluster on top of 10k samples of another yields a deviation of 3.5 standard deviations. The full $D + M$-dimensional space is then randomly rotated to obscure the discriminating variables. The contrastive embedding is trained on $N - 1$ clusters with one held out as a signal, with the backbone architecture f_θ being a simple MLP.

Astronomy For an astronomical baseline we choose gravitational wave data recorded by the Laser Interferometer Gravitational-Wave Observatories (LIGO) in Hanford, WA and Livingston, LA [54]. Although gravitational waves from compact binary systems have been detected [55], many hypothetical sources remain unobserved, making the challenge particularly intriguing for anomaly detection methods. The data consist of 50 ms time-series signals from two channels - one for each interferometer - sampled at 4096 Hz (200 measurements per channel) [56, 57]. The different classes of data consist of pure background ($\sim$ gaussian noise), "glitches" (periods of short-duration transient instrumental noise), and six observed or hypothetical sources of astrophysical signals. A seventh signal class with "white noise burst" (WNB) waveforms is held out from pre-training and is injected as an anomaly in the discovery phase. The encoder architecture is a one-dimensional ResNet, following the technical setup explored in [58] for identification of binary black holes.

Particle Physics Our particle physics baseline is JetClass [59, 60], a large dataset consisting of simulated *jets*: energetic, collimated streams of $\mathcal{O}(100)$ particles that are produced in proton-proton collisions at the Large Hadron Collider (LHC). We use a subset of JetClass consisting of jets from quantum chromodynamics (QCD) processes (quark/gluon), top quark decays ($t \to bqq'$), and W/Z vector boson decays ($V \to qq'$). We hold out signal jets from boosted Higgs boson decays to bottom quarks ($H \to b\bar{b}$), inspired by recent measurements of $H \to b\bar{b}$ in the combined gluon fusion and vector boson production modes by the CMS experiment [61]. We use the Particle Transformer (ParT) architecture [60] – a variant of the Transformer architecture [62] adapted for particle physics – as the contrastive encoder.

Histology As an example from life sciences, we aim to identify abnormal tissue in histopathological images. The abundance of healthy tissue data and the difficulty in collecting samples with various abnormalities render histology particularly well-suited to anomaly-detection tasks. We use publicly available optical microscope images from stained tissue samples [63]. Our reference sample contains seven classes of tissue from mice (brain, heart, kidney, liver, lung, pancreas, spleen) and one class of normal liver tissue from rats. We aim to detect anomalous mouse liver tissue caused by non-alcoholic fatty liver disease (NAFLD). Inputs are 256x256 pixel (0.44µm/pixel) resolution tissue tiles extracted from the whole slide image with Masson's trichome staining. As a backbone, we train the best performing architecture from Ref. [25], EfficientNet-B0 [64].

Images We use the CIFAR-10 dataset [65], arbitrarily holding out class 1 as the anomaly and pre-training on the other nine classes. In the discovery phase of the pipeline, we use images from CIFAR-5m [66] to supplement the CIFAR-10 test set and expand the number of data points available for hypothesis testing.[2] We use a ResNet-50 encoder backbone with pre-trained weights [67], swapping out only the final fully connected layer with a slightly larger MLP and fine-tuning it on the CIFAR contrastive embedding task.

[2] CIFAR-5m was introduced in [66] and consists of images generated by a diffusion model trained on CIFAR-10, which were found to be nearly indistinguishable from CIFAR-10 by a pre-trained classifier.

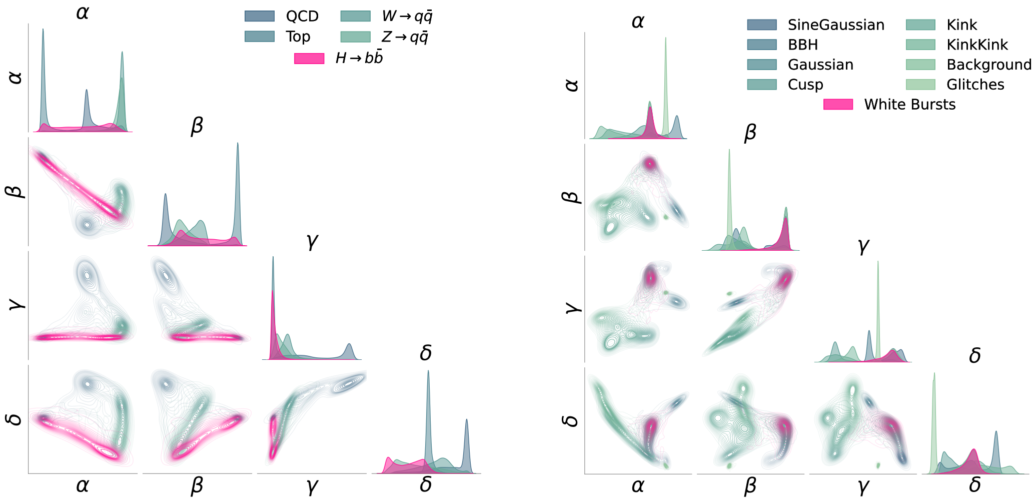


Figure 2: Contrastive embeddings for *Particle physics* (left) and *Astronomy* (right) datasets. The high-dimensional input is projected down to four dimensions $(\alpha, \beta, \gamma, \delta)$. Background classes are shown in hues of blue and green, while the anomaly is overlaid in hot pink.

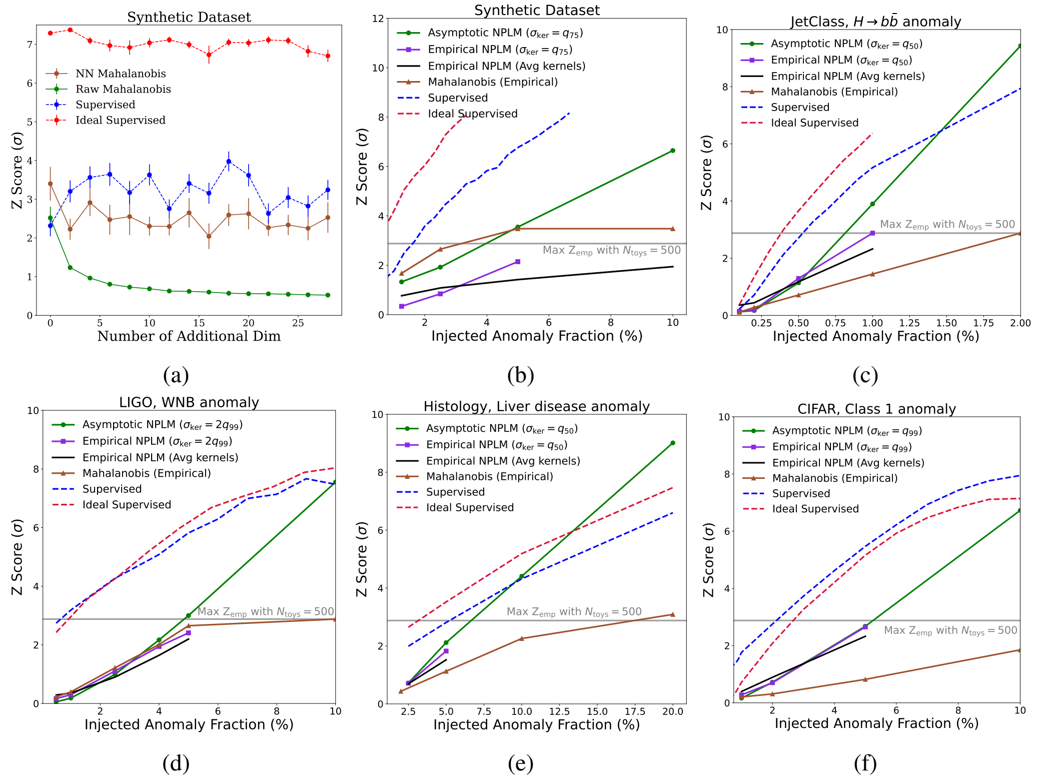
5 Experiments

Embedding We train and evaluate the AutoSciDACT pipeline on each dataset in the same manner, making only small adjustments in pre-training to adapt to the specifics of each dataset (see App. A for full details). We fix the embedding dimension to $d = 4$ for all encoders to put each on equal footing for NPLM, whose performance varies with input dimensionality. The choice of a low embedding dimension is made to ensure that statistical tests remain tractable, and to demonstrate that it is possible to obtain strong anomaly detection performance with a very compact representation (see App. B.3 for a study of larger embedding dimensions). In Fig. 2 we visualize the learned contrastive embeddings for JetClass and LIGO, with embeddings of the anomalous class - which was not included in the training - indicated in pink. The anomalous cluster in JetClass manifests as an extended and distinct cluster, while in LIGO it is an overdensity near a background-populated region. Flagging the latter would be challenging for traditional per-datapoint anomaly detection methods, but we will demonstrate that NPLM detects it as an overdensity.

Anomaly detection & hypothesis testing We follow a standardized procedure for signal injection, anomaly detection, and hypothesis testing for each dataset. As described in Sec. 3.2, we compile a reference sample $\mathcal{R}$ from the test set composed entirely of the known classes used in training. We then construct a "observed data" set $\mathcal{D}$, also from the known classes and in the same relative proportions as $\mathcal{R}$. We mimic the presence of novelty in $\mathcal{D}$ by injecting some number $N_S = f_S|\mathcal{D}|$ of anomalous signal datapoints from the held out class, where typically $f_S \lesssim 0.1$. For each injection rate f_S, we run 500 NPLM pseudo-experiments to populate a distribution of test statistics $t(\mathcal{D}; f_S)$, re-sampling $\mathcal{D}$ and signal injections each time.[3] To calibrate the test, we run 500 additional pseudo-experiments with $f_S = 0$ to populate a reference distribution of $t(\mathcal{D}|f_S = 0)$. The empirical and asymptotic p-value and Z-score are computed. For each dataset, we scan f_S across a range of injection fractions and plot the resulting Z-scores in Fig. 3. The size and composition of $\mathcal{R}$ and $\mathcal{D}$ are fixed by the practical limitations of each dataset (i.e. the test set size), and f_S is varied in a range where NPLM starts to become sensitive to the injected signal. This information is summarized in Table 1 in App. A. Each panel of Fig. 3 also includes results from three baseline statistical tests to compare with NPLM: two supervised tests that incorporate explicit knowledge of the signal, and a test based on the Mahalanobis distance [28].

Supervised baselines We use two fully-supervised baselines as an estimate of best-case anomaly detection performance. We denote them "supervised" and "ideal supervised", distinguishing the extent to which knowledge of the true signal is utilized. For the "supervised" baseline we train an MLP to identify the desired signal in the contrastive embedding space, while for "ideal supervised"

[3] In cases where the test datasets are large enough, we re-sample $\mathcal{R}$ as well. Full details are in Appendix A.



(a) (b) (c)

(d) (e) (f)

Figure 3: Statistical significances (Z scores) of NPLM and other baseline methods for detecting various fractions of anomalous signals injected into background-dominated samples in (a) scanning additional random variables to the same Synthetic toy at a fixed fraction of 0.6% with 10k Background, and then for the fraction for (b) Synthetic (2k), (c) particle physics, (d) astronomy, (e) histology, and (f) image datasets. In all cases, NPLM is able to discover very small signals with high confidence. The upper limit of the empirical Z scores is indicated by a gray line at roughly 2.88σ and is set by the fixed number of pseudo-experiments (500), so empirical numbers are not quoted beyond this point. The Asymptotic NPLM approximates the large pseudo-experiment limit at large Z scores.

baseline we do the same but first *re-train* the contrastive embedding with the true signal added to the training set (the encoder has explicit knowledge of the signal). The reference $\mathcal{R}$ and observed $\mathcal{D}$ datasets are constructed the same way as NPLM, but the hypothesis test relies on more typical statistical methodology. We construct one-dimensional distributions of classifier scores s (normalized to the range [0,1]), and use the points in $\mathcal{R}$ to construct a background-only shape template $f_R(s)$ and signal points to construct a signal template $f_S(s)$. We then perform a binned maximum likelihood fit to the classifier scores in $\mathcal{D}$ under the null ($H_0 : \mathcal{D} \sim a_1 p_{\mathcal{R}}(s)$) and alternative ($H_1 : \mathcal{D} \sim a_1 p_{\mathcal{R}}(s) + a_2 p_S(s)$) hypotheses and compute the test-statistic $\Delta\chi^2 = \chi^2_{H_0} - \chi^2_{H_1}$. We compute empirical and asymptotic[4] p-values and Z-scores over many pseudo-experiments [68].

Mahalanobis baseline As a comparison with analytic anomaly score, we use the Mahalanobis distance metric [28]. For each pseudo-experiment we compute the mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ for the embeddings of each background class i in $\mathcal{R}$. The Mahalanobis distance is then $d_{\mathrm{Maha}}(\mathbf{x}, \mathcal{R}) = \min_i (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$, and we define the test statistic as $t_{\mathrm{Maha}}(\mathcal{D}) = \sum_{\mathbf{x} \in \mathcal{D}} d_{\mathrm{Maha}}(\mathbf{x}, \mathcal{D})$. Empirical p-values and Z-scores are computed as before.

[4] $\Delta\chi^2$ is asymptotically χ^2 distributed with one degree of freedom.

6 Discussion

6.1 Results

The results in Fig. 3 clearly demonstrate the power of the AutoSciDACT pipeline, with NPLM flagging highly statistically significant deviations ($Z \gtrsim 3$ or $p \lesssim 10^{-3}$) with signal fractions as low as 1%. The two supervised baselines provide a reasonable upper limit on the sensitivity to the signal given full knowledge of its distribution in the embedded space, and in some cases, NPLM performs near this limit. Beyond roughly 5σ, some trends break down, but at this level of significance ($p \sim 10^{-7}$), discovery is extremely clear.

In all but the synthetic datasets, NPLM significantly outperforms the Mahalanobis baseline due to the flexible range of distortions and overdensities it is capable of modeling in the input space. The Mahalanobis test is best suited to cases where each background cluster is roughly normally distributed, and by construction is not sensitive to overdensities near the bulk of any given cluster. Since the synthetic dataset is constructed from Gaussian clusters, Mahalanobis is quite effective in this case. In Fig. 3(a), we leverage the computational efficiency of Mahalanobis distance by running over 100 toy synthetic datasets per point, comparing performance on raw versus embedded inputs. As is clear with the raw performance, additional random variables quickly destroy the sensitivity to hidden signals. Sensitivity is preserved across all numbers of random variables using the fixed-dimensional embeddings as inputs, as the large number of noisy dimensions has little impact on the quality of the embedding.

For both the LIGO and JetClass dataset, we approach the supervised limit at a Z-score of 3, which rivals or exceeds all anomaly detection algorithms within their respective domains [69–74]. While astronomy and particle physics have long leveraged statistically rigorous anomaly-detection techniques, their application to histology illustrates a successful transfer of methods across scientific disciplines. The results on the histology datasets align with the findings reported in [25], which demonstrate that embedding spaces constructed with label information outperform those based solely on data augmentations. With AutoSciDACT, we introduce a new method capable of detecting localized abnormalities that may be present in only a small fraction of tissue, a capability that is essential both for early detection of disease and for guiding pathologists' judgments on toxic compounds.

6.2 Limitations

Domain knowledge Since AutoSciDACT relies exclusively on domain knowledge in the label information, its performance is highly correlated with the label quality. Although labeling is easy and accurate in some domains (e.g. simulations, or organ labels for histological patches), labeling large training subsets can be laborious or impossible. For all baseline results, we also assume equal distributions from all background classes in the reference distribution for both pre-training and discovery. However, the actual composition of the reference sample during discovery needs to resemble the one in the observed data, and may require additional input from domain experts. These are problems routinely solved by scientists, so they do not pose a major obstacle to implementing the pipeline.

Embedding dimensionality Embedding into a small space ($d = 4$) limits expressivity, though the features learned in contrastive pre-training will typically be more useful than handpicked variables. This is most evident in the LIGO and CIFAR-10 results in Fig. 3, where the "ideal supervised" benchmark falls short of the supervised one when it should in principle do better. This is due to the density of a large number of classes, which struggle to be perfectly separated in the four-dimensional space. The "ideal" scenario, including an additional class in the learned space, exacerbates this problem. The embedding dimension can be reasonably scaled up (see App. B.3), but beyond a certain point (e.g. hundreds or thousands of dimensions) NPLM's sensitivity will degrade substantially due to the sparsity of the data.

Domain shift and uncertainties In all experiments, we assume that the reference dataset correctly resembles the background distribution of the data. While this is an exact assumption in cases where it is possible to label subsets of data, the reference sample might contain domain shifts if it is constructed from data recorded under different conditions or from simulation. The impact of domain shift on contrastive embeddings has been studied in [75], and the inclusion of epistemic uncertainties within both NPLM and the embeddings is possible [50, 76]. Extensions of AutoSciDACT, including domain shifts and estimation of the associated epistemic uncertainties, are left for future work.

6.3 Conclusion & Future Work

In summary, we have presented what is, to our knowledge, the first end-to-end scientific pipeline for novelty discovery in arbitrary datasets with a rigorous statistical foundation based on hypothesis testing. We show that using AutoSciDACT, we discover anomalous signals with high statistical significance ($\geq 3\sigma$) even when the data contains only a percent-level signal fraction and the dimensionality of the raw data is large. By applying AutoSciDACT to five different datasets from four different scientific domains, we prove the methods' universality and transferability, enabled by the strict decorrelation of expert knowledge encapsulated in label information from the actual analysis pipeline. For a comprehensive scientific outcome, incorporating potential domain shifts along with their associated uncertainties is essential. We plan to further extend AutoSciDACT through known extensions of our methods. More generally, by abstracting the scientific method, our approach presents a framework that automates scientific discovery, leading to the possibility of rapid, comprehensive, and rigorous scientific analysis on all data.

7 Acknowledgments

This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). CR acknowledges support under a Postdoc.Mobility fellowship from the Swiss National Science Foundation (SNF) (Grant No. 222340). Computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Following the publication in NeurIPS. It was pointed out that reference [12] was linked to an incorrect, AI-hallucinated article. We acknowledge this mistake, which resulted from an error in preparing the BibTeX citation using LLMs, prompting the LLM to generate the `BibTeX` citation with the correct first author and the first word of the title. The updated draft now includes the intended reference.

References

[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[2] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

[3] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, and Ryutaro Tanno. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[4] Marco Letizia, Gianvito Losapio, Marco Rando, Gaia Grosso, Andrea Wulzer, Maurizio Pierini, Marco Zanetti, and Lorenzo Rosasco. Learning new physics efficiently with nonparametric methods. *Eur. Phys. J. C*, 82(10):879, 2022. doi: 10.1140/epjc/s10052-022-10830-y.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[9] 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.

[10] Beliz Gunel, Cheng Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Alec Radford, Jong Wook Kim, Jack Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Miles Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[12] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021.

[13] Charles Guille-Escuret, Pau Rodriguez, David Vazquez, Ioannis Mitliagkas, and Joao Monteiro. Cadet: Fully self-supervised out-of-distribution detection with contrastive learning. *Advances in Neural Information Processing Systems*, 36:7361–7376, 2023.

[14] Puck de Haan and Sindy Löwe. Contrastive predictive coding for anomaly detection. *arXiv preprint arXiv:2107.07820*, 2021.

[15] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018.

[16] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

[17] Hao Liu and Pieter Abbeel. Hybrid discriminative-generative training via contrastive learning. *arXiv preprint arXiv:2007.09070*, 2020.

[18] Umar Khalid, Ashkan Esmaeili, Nazmul Karim, and Nazanin Rahnavard. Rodd: A self-supervised approach for robust out-of-distribution detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 163–170. IEEE, 2022.

[19] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[20] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5216–5223, Apr. 2020. doi: 10.1609/aaai.v34i04.5966. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5966`.

[21] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[22] Nicolas Baron Perez, Marcus Brüggen, Gregor Kasieczka, and Luisa Lucie-Smith. Classification of radio sources through self-supervised learning, 2025. URL `https://arxiv.org/abs/2503.19111`.

[23] Sara Webb, Michelle Lochner, Daniel Muthukrishna, Jeff Cooke, Chris Flynn, Ashish Mahabal, Simon Goode, Igor Andreoni, Tyler Pritchard, and Timothy M C Abbott. Unsupervised machine learning for transient discovery in deeper, wider, faster light curves. *Monthly Notices of the Royal Astronomical Society*, 498(3):3077–3094, 09 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2395. URL https://doi.org/10.1093/mnras/staa2395.

[24] Benjamin Voigt, Oliver Fischer, Bruno Schilling, Christian Krumnow, and Christian Herta. Investigation of semi- and self-supervised learning methods in the histopathological domain. *Journal of Pathology Informatics*, 14:100305, 2023. ISSN 2153-3539. doi: https://doi.org/10.1016/j.jpi.2023.100305. URL https://www.sciencedirect.com/science/article/pii/S2153353923001190.

[25] Igor Zingman, Birgit Stierstorfer, Charlotte Lempp, and Fabian Heinemann. Learning image representations for anomaly detection: Application to discovery of histological alterations in drug development. *Medical Image Analysis*, 92:103067, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.103067. URL https://www.sciencedirect.com/science/article/pii/S1361841523003274.

[26] Barry M. Dillon, Radha Mastandrea, and Benjamin Nachman. Self-supervised anomaly detection for new physics. *Phys. Rev. D*, 106:056005, Sep 2022. doi: 10.1103/PhysRevD.106.056005. URL https://link.aps.org/doi/10.1103/PhysRevD.106.056005.

[27] 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. 2 2025.

[28] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[29] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[30] Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: learning and combining kernels for two-sample testing without data splitting. NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[31] Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.

[32] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

[33] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[34] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations (ICLR)*, 2019.

[35] Olmo Cerri, Thong Q. Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Variational Autoencoders for New Physics Mining at the Large Hadron Collider. *JHEP*, 05: 036, 2019. doi: 10.1007/JHEP05(2019)036.

[36] Pratik Jawahar, Thea Aarrestad, Nadezda Chernyavskaya, Maurizio Pierini, Kinga A. Wozniak, Jennifer Ngadiuba, Javier Duarte, and Steven Tsan. Improving variational autoencoders for new physics detection at the lhc with normalizing flows. *Frontiers in Big Data*, Volume 5 - 2022, 2022. ISSN 2624-909X. doi: 10.3389/fdata.2022.803685. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.803685.

[37] Ekaterina Govorkova et al. Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider. *Nature Mach. Intell.*, 4:154–161, 2022. doi: 10.1038/s42256-022-00441-3.

[38] Thorben Finke, Michael Krämer, Alessandro Morandini, Alexander Mück, and Ivan Oleksiyuk. Autoencoders for unsupervised anomaly detection in high energy physics. *JHEP*, 06:161, 2021. doi: 10.1007/JHEP06(2021)161.

[39] Abhijith Gandrakota. Realtime Anomaly Detection at the L1 Trigger of CMS Experiment. *PoS*, ICHEP2024:1025, 2025. doi: 10.22323/1.476.1025.

[40] Vasilis Belis, Patrick Odagiu, and Thea Klaeboe Aarrestad. Machine learning for anomaly detection in particle physics. *Reviews in Physics*, 12:100091, 2024. ISSN 2405-4283. doi: https://doi.org/10.1016/j.revip.2024.100091. URL `https://www.sciencedirect.com/science/article/pii/S2405428324000017`.

[41] D. Abadjiev et al. Autoencoder-Based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter. *Comput. Softw. Big Sci.*, 8(1):11, 2024. doi: 10.1007/s41781-024-00118-z.

[42] Oliver Knapp, Olmo Cerri, Guenther Dissertori, Thong Q. Nguyen, Maurizio Pierini, and Jean-Roch Vlimant. Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark. *Eur. Phys. J. Plus*, 136(2):236, 2021. doi: 10.1140/epjp/s13360-021-01109-4.

[43] Sang Eon Park, Dylan Rankin, Silviu-Marian Udrescu, Mikaeel Yunus, and Philip Harris. Quasi Anomalous Knowledge: Searching for new physics with embedded knowledge. *JHEP*, 21:030, 2020. doi: 10.1007/JHEP06(2021)030.

[44] Filip Morawski, Michał Bejger, Elena Cuoco, and Luigia Petre. Anomaly detection in gravitational waves data using convolutional autoencoders. *Machine Learning: Science and Technology*, 2(4):045014, jul 2021. doi: 10.1088/2632-2153/abf3d0. URL `https://dx.doi.org/10.1088/2632-2153/abf3d0`.

[45] Eric A. Moreno, Bartlomiej Borzyszkowski, Maurizio Pierini, Jean-Roch Vlimant, and Maria Spiropulu. Source-agnostic gravitational-wave detection with recurrent autoencoders. *Mach. Learn. Sci. Tech.*, 3(2):025001, 2022. doi: 10.1088/2632-2153/ac5435.

[46] Ryan Raikman et al. GWAK: gravitational-wave anomalous knowledge with recurrent autoencoders. *Mach. Learn. Sci. Tech.*, 5(2):025020, 2024. doi: 10.1088/2632-2153/ad3a31.

[47] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP*, 10:174, 2017. doi: 10.1007/JHEP10(2017)174.

[48] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning New Physics from a Machine. *Phys. Rev. D*, 99(1):015014, 2019. doi: 10.1103/PhysRevD.99.015014.

[49] 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.

[50] Gaia Grosso, Marco Letizia, Maurizio Pierini, and Andrea Wulzer. Goodness of fit by Neyman-Pearson testing. 5 2023.

[51] 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.

[52] Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933. doi: 10.1098/rsta.1933.0009. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009`.

[53] Gaia Grosso and Marco Letizia. Multiple testing for signal-agnostic searches for new physics with machine learning. *Eur. Phys. J. C*, 85(1):4, 2025. doi: 10.1140/epjc/s10052-024-13722-5.

[54] J. Aasi et al. Advanced LIGO. *Class. Quant. Grav.*, 32:074001, 2015. doi: 10.1088/0264-9381/32/7/074001.

[55] R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Phys. Rev. X*, 13(4):041039, 2023. doi: 10.1103/PhysRevX.13.041039.

[56] LIGO Scientific. The O3a Data Release, April 2021. URL `https://doi.org/10.7935/nfnt-hm34`.

[57] LIGO Scientific. The O3b Data Release, November 2021. URL `https://doi.org/10.7935/pr1e-j706`.

[58] Ethan Marx et al. Machine-learning pipeline for real-time detection of gravitational waves from compact binary coalescences. *Phys. Rev. D*, 111(4):042010, 2025. doi: 10.1103/PhysRevD.111.042010.

[59] Huilin Qu, Congqiao Li, and Sitian Qian. Jetclass: A large-scale dataset for deep learning in jet physics, June 2022. URL `https://doi.org/10.5281/zenodo.6619768`.

[60] Huilin Qu, Congqiao Li, and Sitian Qian. Particle Transformer for Jet Tagging. 2 2022.

[61] Armen Tumasyan et al. A portrait of the Higgs boson by the CMS experiment ten years after the discovery. *Nature*, 607(7917):60–68, 2022. doi: 10.1038/s41586-022-04892-x. [Erratum: Nature 623, (2023)].

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[63] I. Zingman, B. Stierstofer, and F. Heinemann. Nafld pathology and healthy tissue samples, 2022. URL `https://osf.io/gqutd/`.

[64] M. Tan and Q.V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. pages 6105–6114. International Conference on Machine Learning, ICML, 2019.

[65] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[66] Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. *arXiv preprint arXiv:2010.08127*, 2020.

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[68] Herman Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573–578, 1954.

[69] Gregor Kasieczka et al. The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics. *Rept. Prog. Phys.*, 84(12):124201, 2021. doi: 10.1088/1361-6633/ac36b9.

[70] Ryan Raikman et al. A Neural Network-Based Search for Unmodeled Transients in LIGO-Virgo-KAGRA's Third Observing Run. 12 2024.

[71] Vasileios Skliris, Michael R. K. Norman, and Patrick J. Sutton. Toward real-time detection of unmodeled gravitational wave transients using convolutional neural networks. *Phys. Rev. D*, 110(10):104034, 2024. doi: 10.1103/PhysRevD.110.104034.

[72] Georges Aad et al. Weakly supervised anomaly detection for resonant new physics in the dijet final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2 2025.

[73] 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.

[74] Vladimir Chekhovsky et al. Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s}$ = 13 TeV. 12 2024.

[75] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning, 2022. URL `https://arxiv.org/abs/2203.02261`.

[76] Philip Harris, Jeffrey Krupa, Michael Kagan, Benedikt Maier, and Nathaniel Woodward. Resimulation-based self-supervised learning for pretraining physics foundation models. *Phys. Rev. D*, 111(3):032010, 2025. doi: 10.1103/PhysRevD.111.032010.

[77] Giacomo Meanti, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Efficient hyperparameter tuning for large scale kernel ridge regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

[78] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. In *Advances in Neural Information Processing Systems 32*, 2020.

[79] Sascha Husa, Sebastian Khan, Mark Hannam, Michael Pürrer, Frank Ohme, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. i. new numerical waveforms and anatomy of the signal. *Phys. Rev. D*, 93:044006, Feb 2016. doi: 10.1103/PhysRevD.93.044006. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.044006`.

[80] Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. ii. a phenomenological model for the advanced detector era. *Phys. Rev. D*, 93:044007, Feb 2016. doi: 10.1103/PhysRevD.93.044007. URL `https://link.aps.org/doi/10.1103/PhysRevD.93.044007`.

[81] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration. LVK Algorithm Library - LALSuite. Free software (GPL), 2018.

[82] Thibault Damour and Alexander Vilenkin. Gravitational wave bursts from cosmic strings. *Phys. Rev. Lett.*, 85:3761–3764, 2000. doi: 10.1103/PhysRevLett.85.3761.

[83] Thibault Damour and Alexander Vilenkin. Gravitational wave bursts from cusps and kinks on cosmic strings. *Phys. Rev. D*, 64:064008, 2001. doi: 10.1103/PhysRevD.64.064008.

[84] Yuka Matsui and Sachiko Kuroyanagi. Gravitational wave background from kink-kink collisions on infinite cosmic strings. *Phys. Rev. D*, 100(12):123515, 2019. doi: 10.1103/PhysRevD.100.123515.

[85] B. Abbott et al. Search for gravitational-wave bursts in LIGO data from the fourth science run. *Class. Quant. Grav.*, 24:5343–5370, 2007. doi: 10.1088/0264-9381/25/3/039801. [Erratum: Class.Quant.Grav. 25, 039801 (2008)].

[86] Florent Robinet, Nicolas Arnaud, Nicolas Leroy, Andrew Lundgren, Duncan Macleod, and Jessica McIver. Omicron: a tool to characterize transient noise in gravitational-wave detectors. *SoftwareX*, 12:100620, 2020. doi: 10.1016/j.softx.2020.100620.

[87] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

[88] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[89] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[90] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[91] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[92] Anniina Mattila, Chris Jiggins, and Ian Warren. University of Helsinki butterfly collection - Anniina Mattila bred specimens, February 2019. URL `https://doi.org/10.5281/zenodo.2555086`.

[93] Camilo Salazar, Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, and Imogen Gavins. Camilo Salazar and Cambridge butterfly wing collection batch 1, May 2019. URL `https://doi.org/10.5281/zenodo.2735056`.

[94] Ian Warren and Chris Jiggins. Miscellaneous Heliconius wing photographs (2001-2019) Part 2, February 2019. URL `https://doi.org/10.5281/zenodo.2553501`.

[95] Gabriela Montejo-Kovacevich, Quentin Paynter, and Amin Ghane. Heliconius erato cyrbia, Cook Islands (New Zealand) 2016, 2019, 2021, September 2021. URL `https://doi.org/10.5281/zenodo.5526257`.

[96] Gabriela Montejo-Kovacevich, Eva van der Heijden, and Chris Jiggins. Cambridge butterfly collection - GMK Broods Ikiam 2018, November 2020. URL `https://doi.org/10.5281/zenodo.4291095`.

[97] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, Eva Wiltshire, and Imogen Gavins. Cambridge butterfly wing collection batch 9, May 2019. URL `https://doi.org/10.5281/zenodo.2714333`.

[98] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, and Eva Wiltshire. Cambridge butterfly wing collection batch 8, May 2019. URL `https://doi.org/10.5281/zenodo.2707828`.

[99] Erika Pinheiro de Castro, Christopher Jiggins, Karina Lucas da Silva-Brandŏ00e3o, Andre Victor Lucci Freitas, Marcio Zikan Cardoso, Eva Van Der Heijden, Joana Meier, and Ian Warren. Brazilian Butterflies Collected December 2020 to January 2021, February 2022. URL `https://zenodo.org/records/5561246`.

[100] Patricio Salazar, Gabriela Montejo-Kovacevich, Ian Warren, and Chris Jiggins. Cambridge butterfly wing collection - Patricio Salazar PhD wild and bred specimens batch 2, January 2019. URL `https://doi.org/10.5281/zenodo.2548678`.

[101] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, and Eva Wiltshire. Cambridge butterfly wing collection batch 7, May 2019. URL `https://doi.org/10.5281/zenodo.2702457`.

[102] Patricio Salazar, Gabriela Montejo-Kovacevich, Ian Warren, and Chris Jiggins. Cambridge butterfly wing collection - Patricio Salazar PhD wild and bred specimens batch 1, December 2018. URL `https://doi.org/10.5281/zenodo.1748277`.

[103] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, Camilo Salazar, Marianne Elias, Imogen Gavins, Eva Wiltshire, Stephen Montgomery, and Owen McMillan. Cambridge and collaborators butterfly wing collection batch 10, May 2019. URL `https://doi.org/10.5281/zenodo.2813153`.

[104] Gabriela Montejo-Kovacevich, Chris Jiggins, and Ian Warren. Cambridge butterfly wing collection batch 1- version 2, May 2019. URL `https://doi.org/10.5281/zenodo.3082688`.

[105] Joana I. Meier, Patricio Salazar, Gabriela Montejo-Kovacevich, Ian Warren, and Chris Jggins. Cambridge butterfly wing collection - Patricio Salazar PhD wild specimens batch 3, October 2020. URL `https://doi.org/10.5281/zenodo.4153502`.

[106] Chris Jiggins and Ian Warren. Cambridge butterfly wing collection - Chris Jiggins 2001/2 broods batch 2, January 2019. URL `https://doi.org/10.5281/zenodo.2550097`.

[107] Chris Jiggins and Ian Warren. Cambridge butterfly wing collection - Chris Jiggins 2001/2 broods batch 1, January 2019. URL `https://doi.org/10.5281/zenodo.2549524`.

[108] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, and Eva Wiltshire. Cambridge butterfly wing collection batch 6, May 2019. URL `https://doi.org/10.5281/zenodo.2686762`.

[109] Ian Warren and Chris Jiggins. Miscellaneous Heliconius wing photographs (2001-2019) Part 3, February 2019. URL `https://doi.org/10.5281/zenodo.2553977`.

[110] Ian Warren and Chris Jiggins. Miscellaneous Heliconius wing photographs (2001-2019) Part 1, February 2019. URL `https://doi.org/10.5281/zenodo.2552371`.

[111] Gabriela Montejo-Kovacevich, Chris Jiggins, Ian Warren, and Eva Wiltshire. Cambridge butterfly wing collection batch 5, May 2019. URL `https://doi.org/10.5281/zenodo.2684906`.

[112] Chris Jiggins, Gabriela Montejo-Kovacevich, Ian Warren, and Eva Wiltshire. Cambridge butterfly wing collection batch 3, May 2019. URL `https://doi.org/10.5281/zenodo.2682458`.

[113] Gabriela Montejo-Kovacevich, Chris Jiggins, and Ian Warren. Cambridge butterfly wing collection batch 2, May 2019. URL `https://doi.org/10.5281/zenodo.2677821`.

[114] Patricio A. Salazar, Nicola Nadeau, Gabriela Montejo-Kovacevich, and Chris Jiggins. Sheffield butterfly wing collection - Patricio Salazar, Nicola Nadeau, Ikiam broods batch 1 and 2, November 2020. URL `https://doi.org/10.5281/zenodo.4288311`.

[115] Gabriela Montejo-Kovacevich, Eva van der Heijden, Nicola Nadeau, and Chris Jiggins. Cambridge butterfly wing collection batch 10, November 2020. URL `https://doi.org/10.5281/zenodo.4289223`.

[116] Samuel Stevens, Jiaman Wu, Matthew J. Thompson, Elizabeth G.Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M. Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 19412-19424, 2024.

[117] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[118] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[119] Antoine Chatalic, Marco Letizia, Nicolas Schreuder, and Lorenzo Rosasco. An efficient permutation-based kernel two-sample test. *arXiv preprint arXiv:2502.13570*, 2025.

[120] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[121] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

[122] Serguei Chatrchyan et al. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

[123] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[124] Cms open data. URL `https://opendata.cern.ch/`.

NeurIPS 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: Our abstract and introduction claim that we have introduced a new pipeline for novelty detection in scientific datasets, which is described in Sec. 3. We claim strong sensitivity to small injections of novel signals into background-only datasets, which is supported by the experiments (Sec. 5).

   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: We directly address limitations in the conclusion, Sec. 6. We discuss how our method is better suited or finding *distributional* anomalies via GoF, rather than finding single point anomalies. We also discuss how domain shift could be tackled in future work.

   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: We do not derive any novel theoretical results. All statistical theory we discuss related to NPLM is established in prior work, which is cited.

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: In sections 3, 4 and 5 we describe in detail the contrastive learning setup we use, the statistical hypothesis testing framework, and the experiments we run. Most datasets we use are open source, with the only exception being a synthetic dataset that we describe how to generate. We describe the datasets, model architectures, and training setups in detail in App. A.

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: Our repository is publicly hosted on GitHub, and we will write detailed instructions about running code and reproducing experiments in the README. We will include the GitHub link in the final version of this paper, after double-blind review.

   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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These are discussed for each dataset in Appendix A.

   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: The foundation of our results are carefully considered and statistically rigorous, as this paper's main focus is scientific hypothesis testing. This is discussed in great detail in Secs. 3 and 5.

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: This is discussed in App. A

   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: [Yes]

   Justification: Our work is focused on hypothesis testing and discovery in large scientific datasets. We do not anticipate any potential harms caused by our research or as a result of it.

   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: Our paper is about scientific hypothesis testing and discovery in existing datasets. We do not anticipate any immediate societal impact of finding outliers in large datasets from e.g. physics and astronomy, aside from extremely broad-scope and minor potential impacts related to uncovering new understanding of the universe. In a bio-sciences context, this could potentially lead to some new medical understanding or discovery, but it is nearly impossible to accurately forecast the occurrence or impact of such discoveries.

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: We use standard, publicly available datasets and machine learning architectures. Our pipeline is designed to be run on scientific datasets. We anticipate no irresponsible use of our code or data.

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: We use and cite open-source datasets and code from open source machine learning libraries, and plan to make all of our code public.

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: [NA]

Justification: We release no new datasets.

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: We use no crowdsourcing or research with human subjects.

Guidelines:

- 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: We do not do any studies inolving human subjects.

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: [NA]

Justification: Our LLM usage is limited to editing the written words in this draft.

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.

Dataset	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{D}\rvert$	f_S range (%)
Synthetic	10,000	2,000	0.5 - 10
Astronomy	20,000	4,000	0.5 - 10
Particle Physics	50,000	10,000	0.1 - 2
Histology	6,296	500	2.5 - 20
Images (CIFAR-10)	9,000	1,800	1 - 10
Genomics	1,764	100	5 - 20

Table 1: The reference dataset size, observed dataset size, and range of injected anomaly fractions (relative to the background component in $\mathcal{D}$) for each dataset considered in our study. These parameters are used when running NPLM pseudo-experiments across a range of signal fractions to produce the results shown in Fig. 3.

A Dataset, Training, and Evaluation Details

All of the experiments presented in this paper were run on an academic computing cluster. The contrastive trainings were run on a single NVIDIA A100 GPU in all cases, and none took more than a few hours to compute. The kernel-based NPLM tests used the GPU-accelerated Falkon package [77, 78] and also ran on a single GPU, with a typical set of 100 toys with $|\mathcal{R}| = 10,000$ and $|\mathcal{D}| = 2,000$ taking 10-20 minutes. All other metrics such as the Mahalanobis test were computed on CPU nodes and were not a significant computational bottleneck.

Table 1 summarizes the reference dataset sizes $|\mathcal{R}|$, observed dataset sizes $|\mathcal{D}|$ and signal fractions f_S used in the experiments presented in Sec. 5. These numbers are typically limited by the test set sizes for each dataset, and by requirement that $|\mathcal{R}|$ be significantly larger than $|\mathcal{D}|$ for NPLM. We fix the ratio at $|\mathcal{R}| = 5|\mathcal{D}|$, except for histology and genomics where the datasets are very small.

As mentioned in Sec. 3.2, the kernel size σ is a configurable hyper-parameter of NPLM, and the performance varies somewhat as the kernel width changes. In practice, all NPLM pseudo-experiments are run with six different variations of the kernel width $\sigma = [0.1, 1.5, 2.6, 3.6, 4.9, 9.8]$. The four dimensional input data are standardized according to the mean and standard deviation of the reference sample $\mathcal{R}$, so these widths refer to a common scale. The smallest-width kernels are best at adapting to small, local features and distortions in the data, while the widest ones can capture excesses or outlier far in the tails away from the bulk background distribution. In Fig. 3 of the main text we present the asymptotic and empirical NPLM Z scores corresponding to the best-performing kernel width, but we also show average empirical Z of all six kernels in black. We plot full results for all kernel widths, both asymptotic and empirical Z scores, in App. B.

A.1 Synthetic Data

The synthetic dataset aims to broadly look at challenging datasets that are largely overlapping and high-dimensional. As part of that, we insisted on a series of core elements to ensure a robust construction. Namely, the chosen mixture of Gaussians

- signals are fully reproducible,
- the minimum pairwise optimized significance between clusters was 3.5 standard deviations; this was computed through the computation of a cumulative distribution about the mean of the Gaussian, assuming a scenario of 1 percent background in a sample size of 10000 events.
- All discriminating variables were randomly mixed among non-discriminating variables
- Gaussian means and sigmas are bounded in ranges of $[0, 1]$ and $[0.02, 0.5]$, respectively.

For each dataset, we generated 10k events in each of the data classes. Datasets ranging from 3 separate classes to 20 were generated, along with additional random variables ranging from 0 to 30 variables. Additionally, to address the variation of models, we generated roughly 100 separate random models for each point. The total number of models utilized is 3870.

With each model, two trainings are performed, using a simple 4-layer MLP with a separate MLP classifier with no output activation and a projector to compute the contrastive loss; there are a total

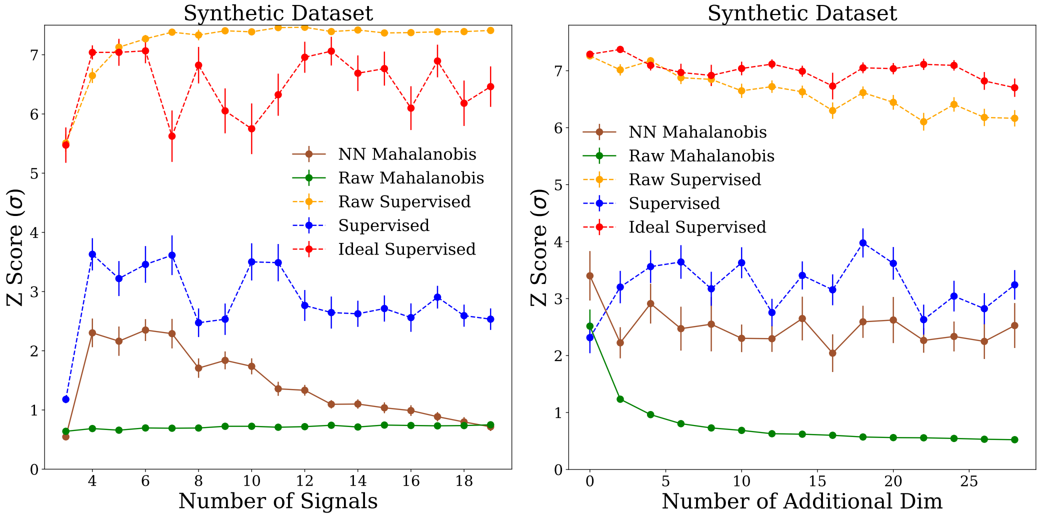


Figure 4: Z-score detection statistic for an anomaly using Mahalanobis distance, and supervised learning on the raw and learned embedded space. (a) We observe the impact of additional signals on sensitivity to find an anomaly. (b) We observe the change in sensitivity vs the number of additional random variables obtained from the average of 100 toys, adding an additional signal model and discriminating variables with each increment on the x-axis. The green corresponds to the Mahalanobis distance on raw inputs, the orange corresponds to a fully supervised algorithm on raw inputs, and the red corresponds to a supervised algorithm trained on the ideal contrastive space, where the signal is known. The blue shows the result of a supervised training (knowing signal) on a contrastive space where the signal is unknown; the brown shows the result of the Mahalanobis distance on the same space.

of 12k trainable parameters. The loss function used was SimCLR with $\tau = 0.01$ and $\lambda_{\mathrm{CE}} = 0.5$, a learning rate of 0.001 with a batch size of 1000 is used along with a cosine annealing. Trainings are performed over 50 epochs and take roughly 5 minutes on a CPU. Similarly, for the supervised algorithms, a 4-layer MLP of roughly the same size was utilized.

Figure 4(a) presents the result of scanning over the toy models, computing the asymptotic Mahalanobis distance for the embedded space, the raw space, and applying a supervised algorithm on both, and with an additional supervised algorithm on the embedded space trained with the hidden signal. The left plot adds an additional signal and an additional discriminating dimension with each variable. Here, we observe that at least four signals are needed to span the space, and high sensitivity is observed, which gradually goes down as the confusion and density from so many signals make it hard for a specific point to separate itself. The right plot shows the impact of additional random dimensions on the data. We observed that either embedding or a supervised algorithm is sufficient to overcome a loss of sensitivity present from just adding random, fluctuating dimensions.

A.2 Astronomy

We utilize the AutoSciDACT pipeline to identify anomalous gravitational-wave sources. Our dataset comprises time-series recorded by the two advanced LIGO detectors [54] - Hanford, Washington, and Livingston, Louisiana - spanning the third observing run (O3) from April 2019 to March 2020; this data is publicly available [56, 57]. Our data preparation and labeling process closely follows the setup described in [46, 70]. Class-balanced reference sets are constructed by injecting simulated signals into real background; the background class is simply segments with no injections. We inject compact-binary coalescences including binary black hole mergers from phenomenological model `IMRPhenomPv2` [79, 80] , (sine-) Gaussian signals [81], signals from cusps [82], kinks [83], double kink events [84], and signals from band-limited white noise bursts (WNBs) [85]. Moreover, a dedicated “glitch” class is obtained using Omicron’s veto criteria [86].

Each input consists of two sequences, one for each detector, with each sequence containing 200 points, corresponding to a 50 ms long time series sampled at 4096 Hz. Exemplary time-series can be

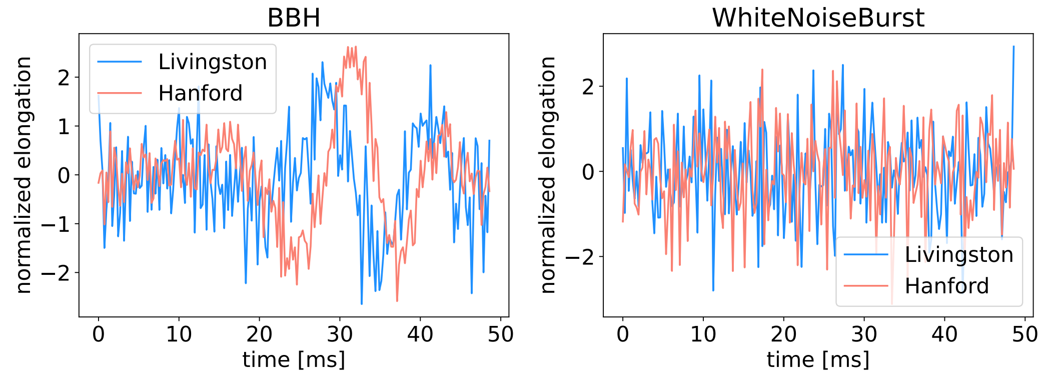


Figure 5: Example LIGO signal waveforms: Signals from binary black holes (left) and white noise burst (right).

found in Fig. 5. The dataset comprises a total of about 530,000 samples, with an near-uniform class balance in the pre-training. To enhance data processing and support network training, the data are normalized to have a unit standard deviation on a per-sample basis. Of the nine total classes, the WNB class is withheld from pre-training and treated as the held-out anomaly for discovery. White-noise bursts are deliberately model-agnostic: they represent correlated, band-limited stochastic fluctuations whose spectra are flat over the analysis band. As such they emulate the “worst-case” burst—lacking distinctive phase evolution or chirp structure, providing a stringent test of the pipeline’s capacity to disentangle subtle, structure-poor signals from detector noise.

The available data is split into a training, validation and test dataset with the test dataset not only utilized for testing the pre-training performance, but also for constructing the reference $\mathcal{R}$ and data distribution $\mathcal{D}$ for the hypothesis test.

The optimization of the backbone encoder f_θ - a one-dimensional ResNet with about 7.2M trainable weights - uses the combined loss objective (SimCLR temperature $\tau = 0.5$, $\lambda_{CE} = 0.5$) and the AdamW optimizer [87] with an initial learning rate of 0.001 and 350 batch size. To facilitate improved convergence and generalization, a cosine annealing learning rate schedule is employed. The training is set up for a maximum of 25 epochs, with early stopping in case the validation loss does not decrease for more than five epochs.

A.3 Particle Physics

JetClass[5] is an open-source particle physics dataset introduced in [60]. The training set consists of 100M jets from 10 classes (10M per class), of which we use five for a total of 50M training samples: QCD (quark/gluon), top quark ($t \to bqq'$), W boson ($W \to qq'$), Z boson ($Z \to q\bar{q}$) and Higgs ($H \to b\bar{b}$). The validation set consists of 500,000 jets per class and is used during training to monitor performance. The test set consists of 2M jets per class and is used to construct reference and data samples for all NPLM hypothesis tests presented in the main text.

For the encoder, we use the particle transformer architecture nearly exactly as described in [60], using a particle embedding of dimension 128, eight self-attention layers with eight heads, and two class-attention layers with the final 4-dimensional embedding derived from the final class token plus a fully connected layer. We use the same 17 per-particle input features described in [60], including information on the particle’s energy/momentum, trajectory, and particle type. We cap the input size at 64 particles per jet. We train the encoder with a SimCLR temperature $\tau = 0.1$ and a classifier strength of $\lambda_{\text{CE}} = 0.1$, running for 100 epochs with an initial learning rate of 5×10^{-4} annealed to 10^{-5} on a cosine schedule and using the AdamW optimizer [88]. We use a batch size of 512.

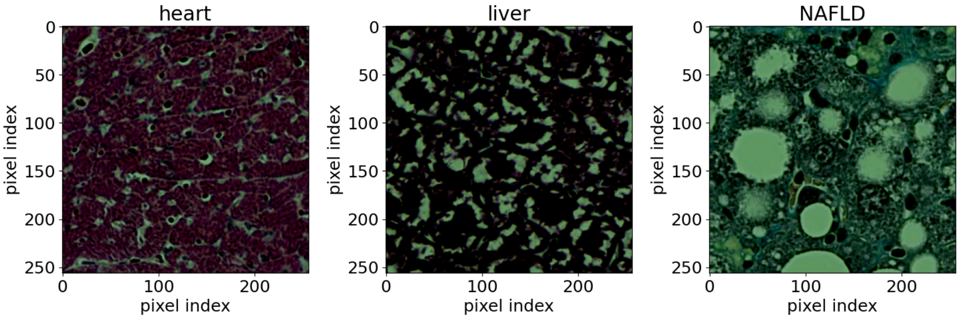


Figure 6: From left to right: exemplary image patches of heart tissue, liver tissue and liver tissue with a non-alcoholic fat Non-Alcoholic Fatty Liver Disease (NAFLD) [63].

A.4 Histology

Exemplary image patches from the publicly available histological dataset [63] are shown in Fig. 6. The training dataset is balanced and includes samples from mouse tissues of the brain, heart, kidney, liver, lung, pancreas, and spleen and a separate class containing normal liver tissue samples from rats. Each class comprises approximately 6,300 samples. The dataset is split into training, validation, and testing sets with a ratio of 70%/20%/10%. In addition, an independent test set consists of approximately 2,300 samples of normal mouse liver tissue and an equal number of samples with non-alcoholic fatty liver disease (NAFLD).

Due to statistical constraints, the reference distribution $\mathcal{R}$ is constructed from the training data, with prior validation to ensure the absence of overfitting. The data distribution $\mathcal{D}$ for normal tissue is derived from both the training set and the independent dataset containing only healthy mouse liver samples.

The backbone encoder f_θ, EfficientNet-B0 [64] with approximately 4.8M trainable parameters, is optimized using a combined loss objective, incorporating SimCLR contrastive loss (temperature $\tau = 0.5$) and cross-entropy loss with weighting $\lambda_{\text{CE}} = 0.5$. Optimization is performed with the AdamW optimizer, using an initial learning rate of 0.001 and a batch size of 32. To promote stable convergence and improved generalization, a cosine annealing learning rate schedule is employed. Training is conducted for a maximum of 25 epochs, with early stopping triggered if the validation loss fails to improve for more than five consecutive epochs. Each training run takes between one and two hours on a single NVIDIA A100 GPU.

A.5 Images

We use the CIFAR-10 [65] dataset, applying the standard resize to 232×232 with interpolation, crop to 224×224, and standardizing using the ImageNet [89] mean and standard deviation. For the encoder, we use the Pytorch-provided pre-trained ResNet-50 [67] weights and replace the final fully connected layer with an MLP of hidden dimensions $[512, 256, 128]$ and an output dimension of 4. Only these final MLP weights are floated during training.

We use the 50,000 CIFAR-10 training images to pre-train the encoder, using only 45,000 in practice because class 1 is held out as the anomaly. When evaluating with NPLM we introduce 100,000 images from CIFAR-5m [66] in order to boost the number of points available for demonstrating our method. The approximately 5 million images in CIFAR-5m were generated by an unconditional denoising diffusion probabilistic model (DDPM) [90] trained on CIFAR-10, then labeled by the 98.5% accurate Big-Transfer model [91]. Trainings took about 1 hour on a single A100 GPU, and ran for 50 epochs with a learning rate of 10^{-3} annealed to 10^{-5} on a cosine schedule and with a batch size of 512. The SimCLR temperature is set to $\tau = 0.1$ and the cross-entropy strength to $\lambda_{\text{CE}} = 0.5$.

[5] https://zenodo.org/records/6619768

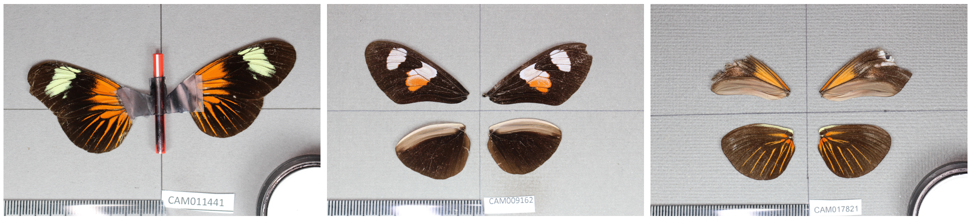


Figure 7: From left to right: Exemplary images of two different butterfly subspecies (left, center) and a hybrid offspring of those species (right).

A.6 Genomics

As an example from genomics, we evaluate our approach on the dataset provided in the "Butterfly Hybrid Detection" challenge[6]. Butterflies come in different subspecies characterized by visual differences in color and pattern on the wings and can sometimes mix and produce offspring. We aim to detect these anomalies, so called hybrids, with AutoSciDACT.

The dataset consists of 1991 colour images of size 224×224 pixels, annotated to belong to one of 14 classes, with a highly imbalanced class distribution (some classes contain almost 100 times more instances than others) [92–115]. Example images of two different species and a hyrbid are shown in Fig.7. We adopt a fixed split of 80% training, 10% validation and 10% test sets. Our model uses a backbone encoder based on the BioClip architecture [116], with approximately 430k trainable weights. Training employs the AutoSciDACT combined loss: a contrastive self-supervised term following the SimCLR formulation with temperature parameter $\tau = 0.5$, and a supervised cross-entropy term, weighted with $\lambda_{\text{CE}} = 0.5$. Optimization is performed using the AdamW optimizer at a learning rate of 0.001 and batch size of 32. We apply cosine annealing learning-rate scheduling over 100 epochs and minimal value 0.0001. Training is terminated by early stopping when no improvement is observed for five consecutive epochs; in our experiments this is the case after 43 epochs.

The contrastive embedding resulting from the projection of image data into a four-dimensional latent space is shown in Fig. 8 (left). Among all experiments, the anomaly in the genomics case shows the clearest separation from the background classes. The performance of AutoSciDACT in this setting is presented in Fig. 8 (right), which displays the statistical significance (Z-scores) obtained by NPLM and various baseline methods for detecting varying proportions of hybrid butterfly species images injected in background-dominated samples. We omit asymptotic results, since the underlying assumption that the test statistic is χ^2-distributed, is invalid in light of the limited statistics of the dataset. The limited statistics of this dataset is also the reason why it is not included in the main results. Notably, the empirical Mahalanobis distance baseline achieves the strongest performance due to the excellent separation of the anomalous class in the contrastive embedding space. Conversely, the Maximum Mean Discrepancy (MMD) test yields the weakest performance, potentially due to suboptimal kernel width choice (see Appendix B.2 for further discussion).

B Additional Experiments

B.1 Impact of NPLM kernel width

The choice of NPLM kernel width has a significant impact on the sensitivity of the test, and there is no *a priori* choice that can optimize sensitivity to potentially anomalous features. The results in Fig. 3 included Z-scores from the best-performing kernel and the average over all kernel widths considered. For completeness, we plot Z-scores from all kernel width settings in Figures 9 (asymptotic) and 10 (empirical). These figures exactly mirror Fig. 3, displaying results for each of the five datasets in the five panels (a)-(e). In Fig. 10 we also show the kernel-averaged Z-score in black. In general, intermediate to larger kernels do the best, with only fairly small variations among the top performers. The empirical Z-scores cannot exceed roughly 2.88σ due to the number of pseudo-experiments (500).

[6] https://www.codabench.org/competitions/3764/

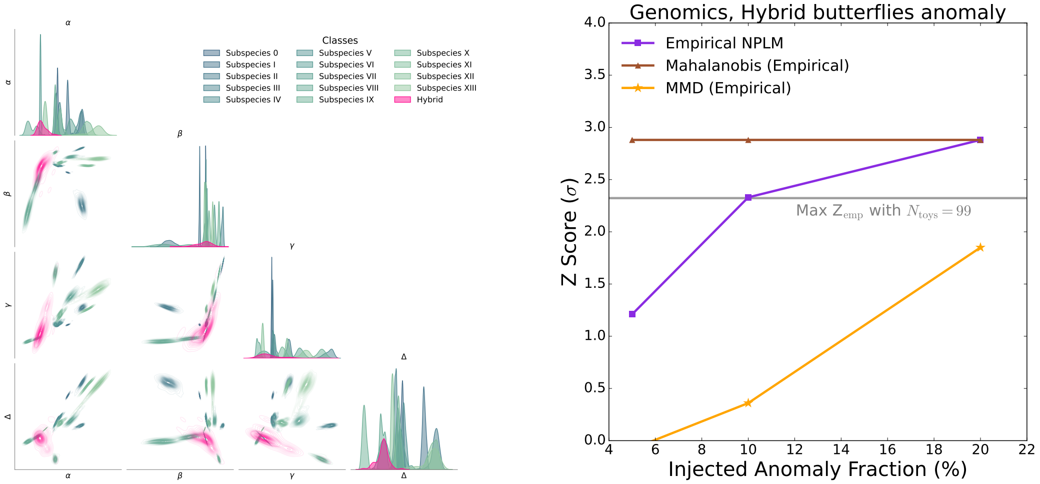


Figure 8: Results for the genomics experiment: Contrastive embeddings in four dimensions (left). Background classes are shown in hues of blue and green, while the anomaly is overlaid in hot pink. Results of the statistical tests (right).

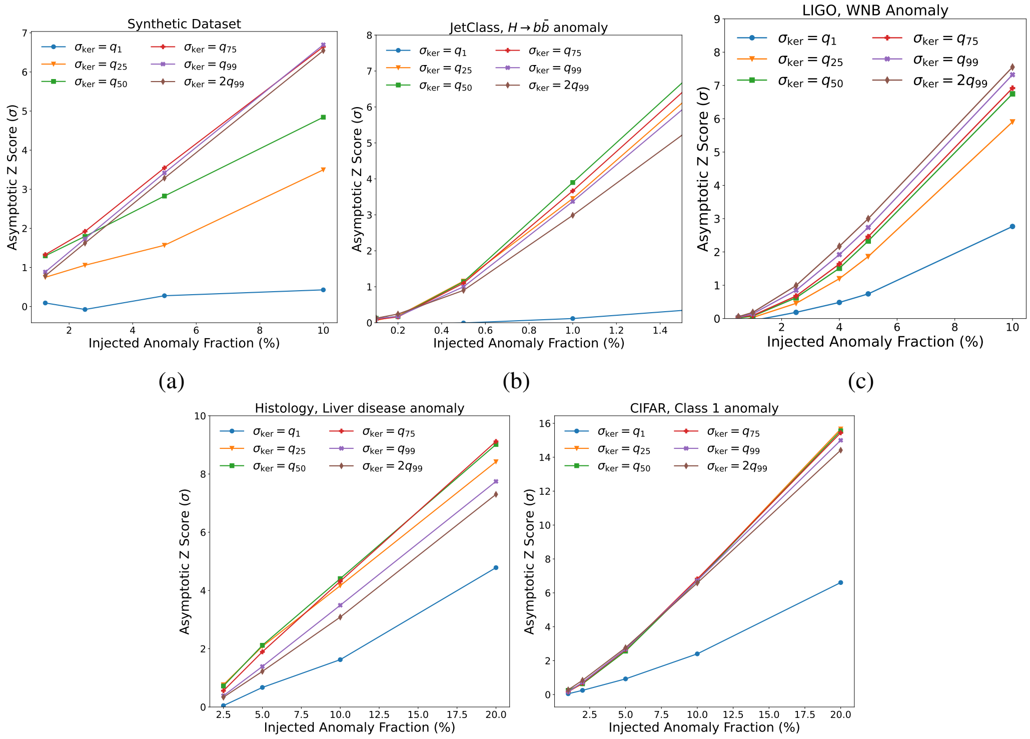
B.2 Comparisons with additional baselines

We further contextualize AutoSciDACT's performance by comparing with two additional anomaly detection baselines: Maximum Mean Discrepancy (MMD) [29] and Fréchet Inception Distance (FID) [117, 118].

Nyström approximated Maximum Mean Discrepancy (N-MMD). The Maximum Mean Discrepancy (MMD) is a kernel-based statistical test used to determine whether two datasets come from the same distribution [29]. It works by mapping the data into a high-dimensional reproducing kernel Hilbert space (RKHS) via a chosen kernel (e.g. Gaussian) and computing the distance between the mean embeddings of the two distributions in that space. MMD has been shown to be sensitive to subtle differences between distributions, making it particularly effective in high-dimensional settings. However, its computational cost, quadratic in the number of examples, limits its applicability to small sample sizes or low-dimensional problems. To address this, [119] introduced a scalable variant of the MMD test based on a Nyström approximation of the kernel matrix, enabling its use on larger datasets while maintaining statistical power. Since the NPLM test employed in this work is also built upon the Nyström approximation, we perform a direct comparison between Nyström-MMD and NPLM under matched settings, using the same number of centroids and same kernel bandwidth.

Fréchet distance (FD). The Fréchet distance has emerged as a popular metric for comparing probability distributions, particularly in the context of generative modeling [117, 118]. Under the assumption that both distributions are Gaussian, it admits a closed-form expression involving only their means and covariances, making it computationally efficient and interpretable. However, this assumption can limit its effectiveness when the underlying distributions exhibit significant non-Gaussian behavior, such as heavy tails or multimodality. Despite this, the Fréchet distance remains widely used due to its robustness in high-dimensional settings and its ability to capture both mean and covariance differences.

In Fig 11 we reproduce Fig. 3 with empirical MMD and FID Z-scores included for the particle physics, astronomy, histology, and image datasets. NPLM outperforms FID/MMD for the particle physics and image datasets, approximately matches them for astronomy, and, surprisingly, underperforms in histology. This wide range of outcomes makes it difficult to draw unambiguous conclusions, but broadly suggests that the "best" anomaly detection method depends strongly on the structure of the embedding space and the size of the dataset at hand. NPLM's sensitivity scales with available sample size, and the histology dataset has by far the smallest available test datasets among all experiments in the main body of the paper (see Table 1). In such data-constrained cases, other methods may perform better as they do not require likelihood ratio estimation. In other cases, the structure of the data embeddings may also confer an advantage (e.g. if clusters are approximately Gaussian). However,



(a) (b) (c)

(d) (e)

Figure 9: Asymptotic NPLM Z scores as a function of injected signal yield for all six kernel choices $\sigma = [0.1, 1.5, 2.6, 3.6, 4.9, 9.8]$ used in pseudo experiments. We show results for our five benchmark datasets: Synthetic (a), JetClass (b), LIGO (c), Histology (d), and CIFAR-10 (e).

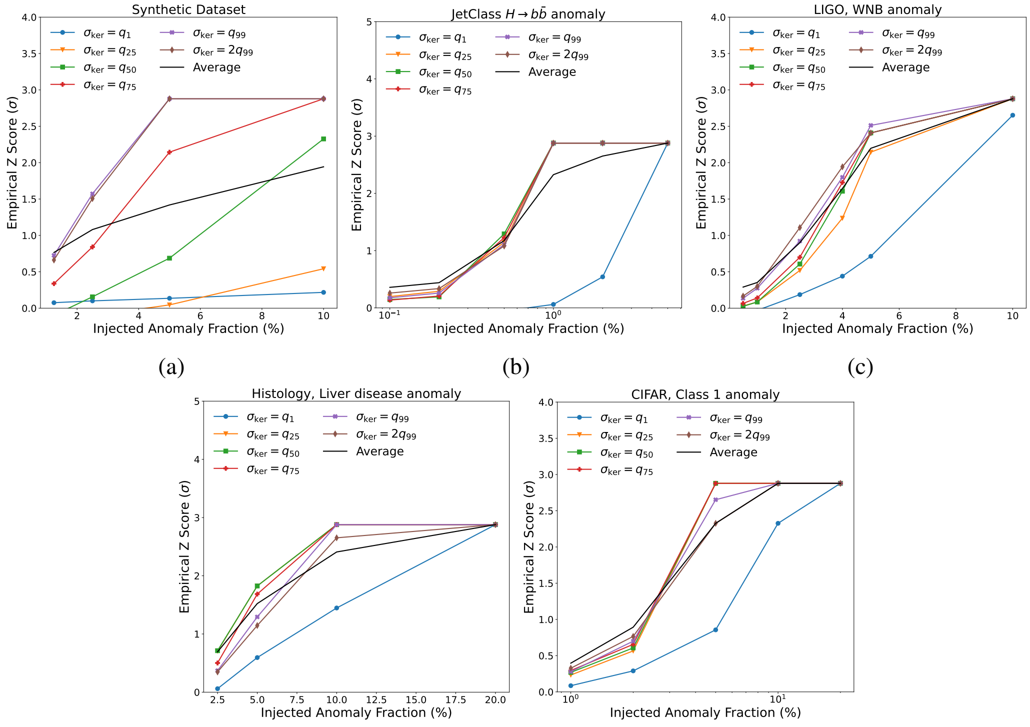
NPLM's strong performance in all but the most data-constrained cases positions it as a reliable and competitive choice.

B.3 Varying embedding dimension

We use a very low contrastive embedding dimension ($d = 4$) for the main results of this paper. This choice was made to ensure the tractability of our statistical anomaly detection technique (NPLM), which relies on statistical hypothesis testing that can quickly lose sensitivity in high dimensions. In this section we explore the impact of increasing the embedding dimension, extending our experiments up to $d = 32$[7]. Figure 12 shows the Z-score as a function of d for the CIFAR, JetClass, and LIGO datasets, where the best-performing kernel widths are used and the signal injection fractions are set such that NPLM has good but not fully-saturated sensitivity at the default setting $d = 4$.

There are no unambiguous trends for any of the anomaly detection methods, with all methods performing relatively stably up to $d = 32$. NPLM's sensitivity declines very modestly in the CIFAR and JetClass examples, but slightly *improves* in the LIGO example. The MMD and Fréchet metrics are similarly stable. Interestingly, the Mahalanobis distance appears to almost always benefit from a larger dimensionality. This could be explained by class-specific clusters having more "room" to spread out and condense under the contrastive objective in a higher-dimensional space. As Mahalanobis distance is sensitive to the distribution of these clusters, this would likely improve performance (assuming that the anomalous cluster is separated from the rest, i.e. not an overdensity in an existing cluster). The absence of clear trends for NPLM is encouraging, suggesting AutoSciDACT could be applied to problems requiring higher-dimensional latent spaces. Where possible, however, we vouch for keeping the dimensionality low as this is beneficial for classical statistical analysis and uncertainty quantification.

[7] This is still modest relative to standard embedding sizes in e.g. computer vision or natural language, but a reasonable choice for many scientific applications.



(a) (b) (c)

(d) (e)

Figure 10: Empirical NPLM Z scores as a function of injected signal yield for all six kernel choices $\sigma = [0.1, 1.5, 2.6, 3.6, 4.9, 9.8]$ used in pseudo experiments. We show results for our five benchmark datasets: Synthetic (a), JetClass (b), LIGO (c), Histology (d), and CIFAR-10 (e). We also show the average empirical Z score from all six kernels in black.

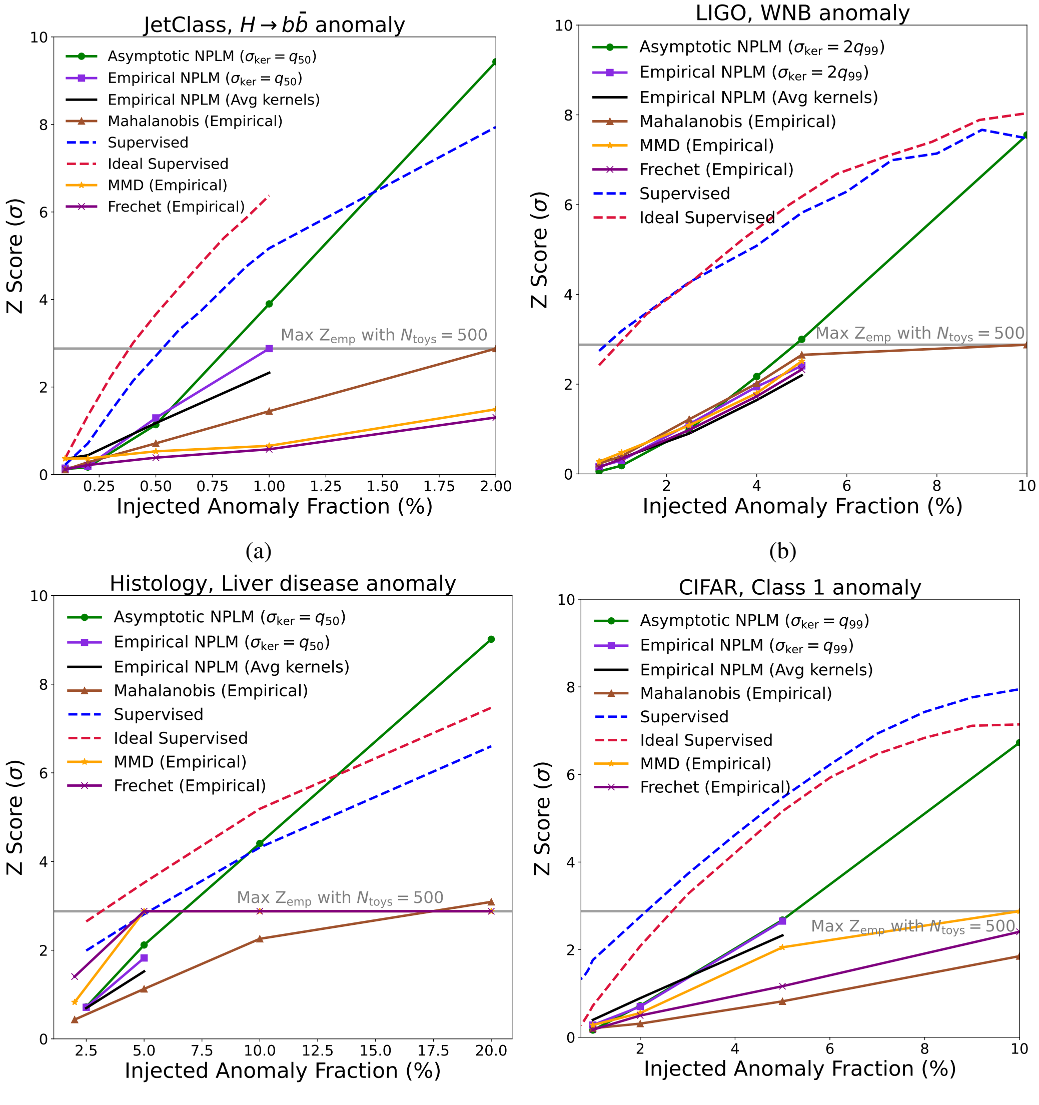
B.4 Noisy labels

Many real-world scientific datasets are labeled by hand or by algorithms with known error rates, both of which result in some fraction of mislabeled training data. To assess our sensitivity to such cases, we train variants of our JetClass contrastive embedding model with 1, 2, 5, and 10% label noise, meaning $x\%$ of samples the training datasets are randomly mislabeled. The results are shown in Fig. 13, where again the best-performing kernel width is chosen and the injected signal fraction is set such that NPLM has good performance at zero noise. As expected, sensitivity drops as label noise is increased, falling from 4σ to about 2σ at 10% noise.

C Comparison to previous results from histology

Our setup for the histology study makes use of the same dataset as in [25]. In [25], the authors proposed a contrastive learning method for anomaly detection. The learned embedding returns a 320-dimensional representation, and a standard one-class SVM with the Radial Basis Function (RBF) kernel and margin error $\nu = 0.1$ is then trained on anomaly-free data to construct an anomaly score. Applying a threshold that ensures zero false positives on the training set, the tiles constituting a data sample are tagged as either standard or anomalous. The final anomaly metric used for comparisons is an average across the tags of all tiles in the sample. This statistical test assumes the anomaly is out-of-distribution and the distribution of the background class in the chosen representation is well clustered. While these assumptions are met in [25], out-of-distribution is not ensured for an arbitrary representation space agnostic to the anomaly source, which motivates our choice for using NPLM as a universal statistical test.

Since [25] does not present results in terms of p-value, we implement the one-class SVM trained on our four-dimensional embedding space for comparison purposes. We find that the performance



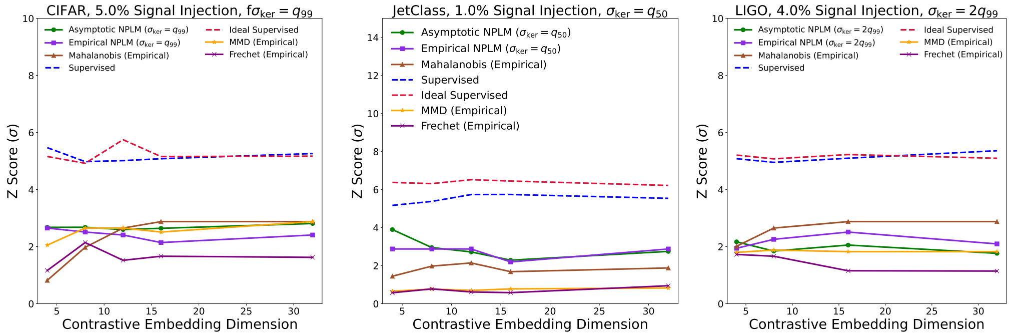
(a)

(b)

(c)

(d)

Figure 11: A reproduction of Fig. 3 including the Maximum Mean Discrepancy (MMD) and Fréchet Inception Distance (FID) baselines for the JetClass (a), LIGO (b), Histology (c), and CIFAR-10 (d) datasets.



(a) (b) (c)

Figure 12: Z score as a function of embedding dimension for NPLM and baseline anomaly detection methods in the CIFAR-10 (a), JetClass (b), and LIGO (c) datasets. The best-performing kernel widths are chosen for presentation, and signal injection fractions are set such that NPLM has good but not fully-saturated sensitivity at the $d = 4$ baseline.

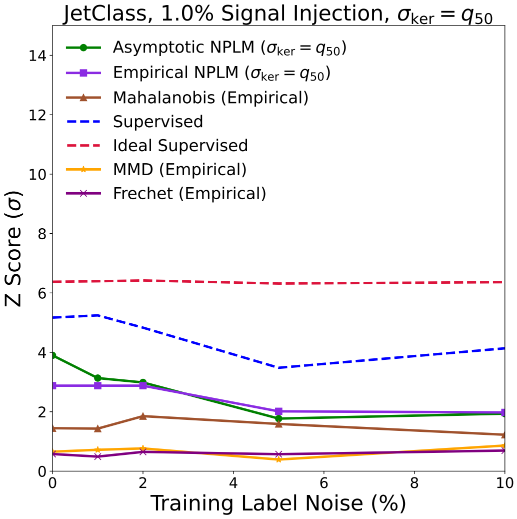


Figure 13: Z score as a function of label noise (%) dimension for NPLM and baseline anomaly detection methods for the JetClass dataset. The best-performing kernel widths are chosen for presentation, and signal injection fractions are set such that NPLM has good but not fully-saturated sensitivity at the zero-noise baseline.

of the test highly depends on the threshold set on the one-class output score. With a threshold that allows for 10% false positive rate, the one-class SVM performs comparably to the NPLM version. The one-class SVM has no discriminative power when the threshold is set to 1% false positive rate, corresponding to a more subtle anomalous contribution. The reason for such variance is the fact that in the latent representation, the anomaly does not necessarily lie outside the distribution. Additional tests run on the astronomy data (where the signal highly overlaps with one of the background classes) show an even more striking example of detection failure.

In conclusion, these studies reassure our strategy of using the NPLM to compute a two-sample test rather than focusing on out-of-distribution detection because two-sample tests are sensitive to a wider range of anomalous behavior. This is particularly relevant for the histology data since more than one anomalous tile is expected per sample, and capturing collective behaviors enables the detection of more subtle anomalies, e.g., to detect diseases manifesting in tissue earlier.

D Discriminating CIFAR-10.1 and CIFAR-5m

To further demonstrate the capabilities of AutoSciDACT for detecting distributional shifts, we apply it to the problem of quantifying the difference between CIFAR-10 and the synthetic CIFAR-5m dataset [66], as well as the independently curated CIFAR-10.1 dataset [120]. CIFAR-5m was discussed in the main text and App. A, and consists of approximately 5 million images generated by a diffusion model trained on CIFAR-10. CIFAR-10.1 was introduced in [120], where it was specifically collected and curated to be as close to CIFAR-10 as possible (e.g. drawing from similar sources of images). The idea was to test whether classifiers trained on CIFAR-10 readily generalized to data outside of CIFAR-10 but in principle distributionally identical. It was found that performance dropped substantially when evaluating on CIFAR-10, indicating some non-trivial shift in the dataeset. Recent work in Liu et. al. [121] and Guille-Escuret et. al. [13] has applied some versions of two-sample tests to distinguish CIFAR-10 and CIFAR-10.1, both showing strong evidence of a distributional shift between the two.

In Fig. 14 we tackle this question with AutoSciDACT. We train a four-dimensional contrastive embedding on the full CIFAR-10 training dataset without holding out any classes, as individual classes are not considered anomalous in this context. We use the same architecture and training procedure as for the CIFAR-10 results in the main text. We use this encoder to embed the CIFAR-10m test set, the CIFAR-10.1 set, and 100,000 randomly selected images from the CIFAR-5m set with the same class proportions as CIFAR-10 test and CIFAR-10.1. We run 500 NPLM pseudo experiments for each of the six kernel widths $\sigma = [0.1, 1.5, 2.6, 3.6, 4.9, 9.8]$ to produce the following distributions of test statistics:

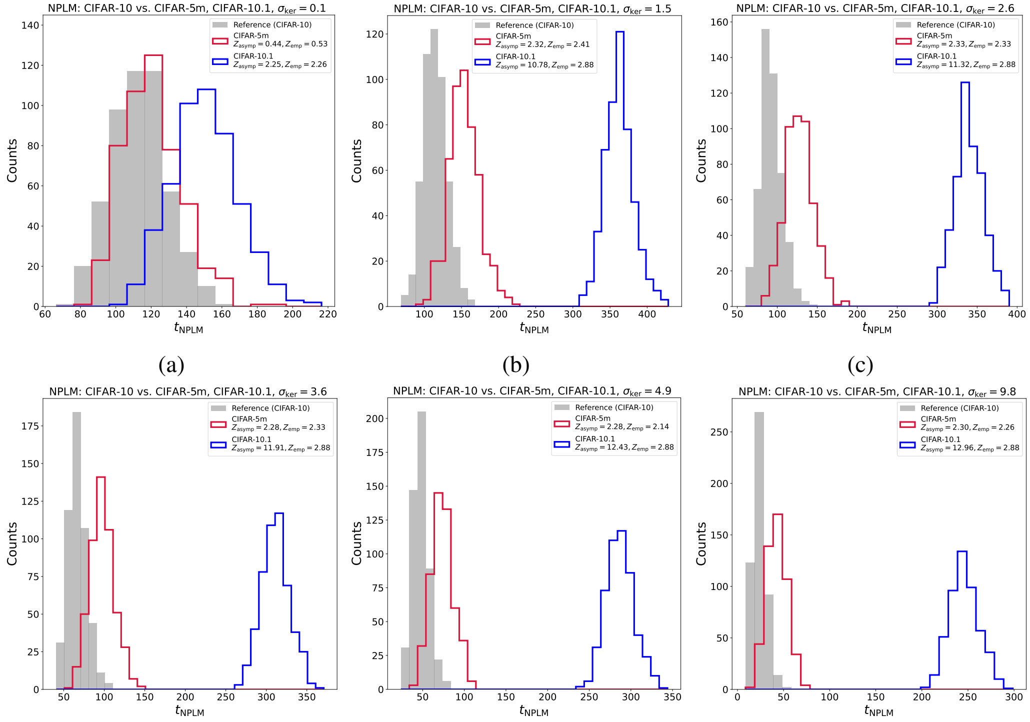
1. **Null hypothesis:** in each experiment $\mathcal{R}$ is composed of 8500 randomly sampled CIFAR-10 test set images, and $\mathcal{D}$ from the remaining 1500.
2. **CIFAR-10.1:** in each experiment $\mathcal{R}$ is composed of 8500 randomly sampled CIFAR-10 test set images, and $\mathcal{D}$ from 1500 randomly sampled CIFAR-10.1 images.
3. **CIFAR-5m:** in each experiment $\mathcal{R}$ is composed of 8500 randomly sampled CIFAR-10 test set images, and $\mathcal{D}$ from 1500 randomly sampled CIFAR-5m images.

Distributions of the NPLM test statistic for each scenario and each kernel width are plotted in Fig. 14, with the corresponding asymptotic and empirical Z scores for CIFAR-10.1 and CIFAR-5m relative to the CIFAR-10-only null hypothesis indicated in the legends. Even the smallest and worst-performing kernel $\sigma = 0.1$ distinguishes CIFAR-10.1 from CIFAR-10 at the 2.2σ level, while the remaining larger kernels distinguish it extremely easy beyond even the 10σ level. This indicates a clear distributional shift between CIFAR-10 and CIFAR-10.1, and presents one of the first (to our knowledge) statistically rigorous quantifications of this discrepancy.

The discrepancy between CIFAR-10 and CIFAR-5m is notably *much* smaller, saturating near 2.3σ for the best performing kernels. This is exactly in line with what one would expect, given that CIFAR-5m is generated from a diffusion model trained on CIFAR-10. A well-trained diffusion model is able to model its training distribution exceedingly accurately, and the relatively small deviation we observe here underscores this fact. More interestingly, this hints at an unexpected but fascinating potential use case for AutoSciDACT as a method for evaluating the quality of generative models.

E Searching for the Higgs boson in LHC Data

To demonstrate how AutoSciDACT might be used in a more realistic setting, we use it to search for evidence of the Higgs boson (H) in a dataset of real proton-proton collision data collected by the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC). We specifically target the four-lepton final-state, where the Higgs boson decays to four electrons, four muons, or two electrons and two muons ($pp \to H \to e^+e^-e^+e^-$, $\mu^+\mu^-\mu^+\mu^-$, or $\mu^+\mu^-e^+e^-$). The Higgs boson was discovered through the observation of an excess in predominantly these final states and the di-photon final state [122, 123], with Higgs-like events isolated using hand-tuned selections on physics-motivated variables that were reconstructed from observed data. Here, we replace the majority of this selection with the AutoSciDACT pipeline, keeping only a loose "pre-selection" of events designed to suppress large, well-known background processes.



(a) (b) (c)

(d) (e) (f)

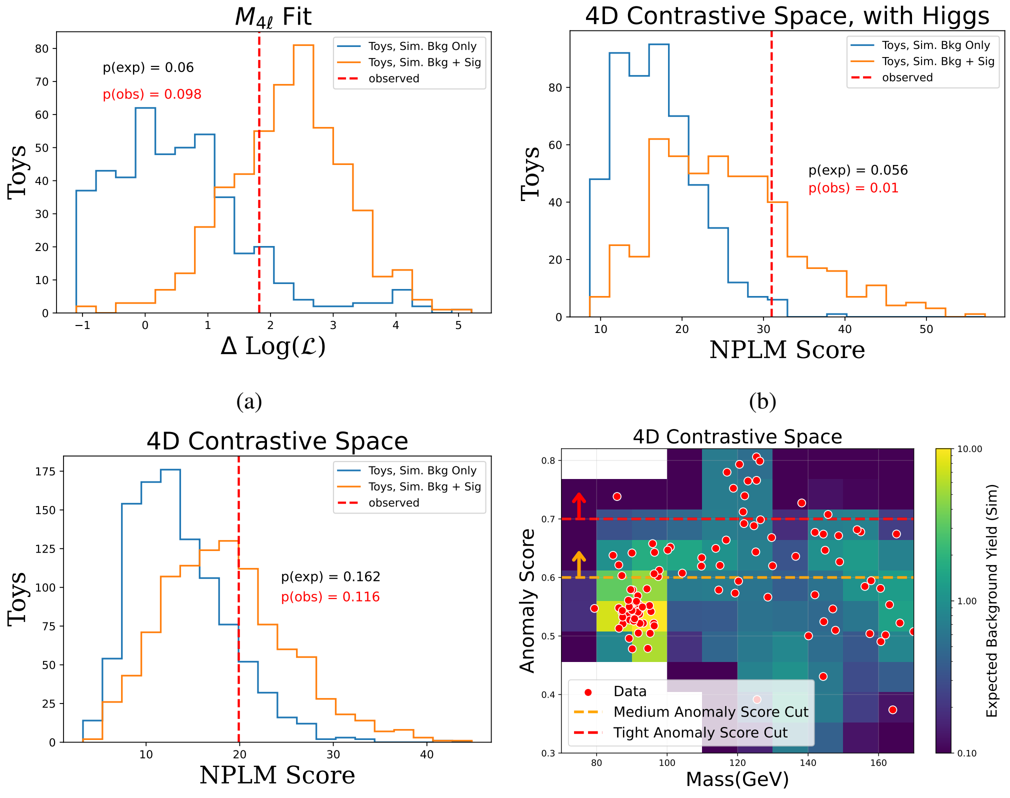
Figure 14: NPLM tests comparing CIFAR-10 to CIFAR-5m and CIFAR-10.1 using a four dimensional contrastive embedding. The reference is built from the CIFAR-10 test set, and test stastics for the null hypothesis are shown in gray. Test statistics comparing against CIFAR-5m are shown in red, while those comparing against CIFAR-10.1 are shown in blue. Panels (a)-(f) corresponding to the six different NPLM kernel widths $\sigma = [0.1, 1.5, 2.6, 3.6, 4.9, 9.8]$.

We run AutoSciDACT on both real and simulated events from the CMS Open Data [124] for 2011 and 2012. We impose a similar data pre-selection to the standard CMS search, requiring at least 4 well-identified and loosely isolated electrons or muons with transverse momenta (p_T) greater than 5 GeV, the most energetic of which ("leading lepton") having $p_T > 20$ GeV. Additionally, we require one opposite-sign pair of electrons or muons with invariant mass ($m_{\ell^+\ell^-}$) greater than 12 GeV. In the first stage of AutoSciDACT, we pre-train the contrastive encoder using simulated events from the dominant background processes (Z boson pair production with decays to the $4e$, 4μ, and $2e2\mu$ final states). The encoder is a small MLP that takes as input the full kinematic information (p_x, p_y, p_z, E) and particle ID (electron or muon) of the four leptons in the event, for a total of 20 inputs. As in the main paper, we train a four-dimensional embedding space.

In the search phase, we follow the standard procedure for CMS data analysis and compute expected p-values based on simulated data, then measure the observed p-value in real data. Expected p-values are computed empirically by running many toys for the background-only and background + Higgs hypotheses to obtain distributions of a test statistic (either the NPLM test statistic or one obtained from a direct fit to the four-lepton invariant mass, see below). For each NPLM toy, we sample the reference ($\mathcal{R}$) and observed ($\mathcal{D}$) datasets from simulated background events, adding additional simulated Higgs events to $\mathcal{D}$ for toys under the background + signal hypothesis. Finally, a single test statistic is computed using the *true* data (i.e. observed in CMS), and the observed p-value is computed relative to the test statistic for background-only toys.

We perform three different tests using the statistical procedure described above:

- **Baseline Discovery** We assume full knowledge of the Higgs boson, and perform a one-dimensional hypothesis test using the single most discriminating variable: the four-lepton invariant mass $m_{4\ell}$. The Higgs boson signal should manifest itself as a localized "bump" or peak in a histogram of observed $m_{4\ell}$ at $m_{4\ell} = m_H$ (approximately 125 GeV). We construct



(a)

(b)

(c)

(d)

Figure 15: Top row: test-statistic distributions for simulated background-only (blue) (orange) simulated background + Higgs signal (orange), along with observed test statistic (red), for a direct fit to $m_{4\ell}$ (left) and AutoSciDACT with Higgs included in training (right). Bottom left: Expected and observed test statistics/p-values for vanilla AutoSciDACT trained without knowledge of the Higgs. Bottom right: Distributions of $m_{4\ell}$ versus AutoSciDACT NPLM anomaly score in simulated backgrounds (filled) and observed data (points), where the most anomalous observed data lies near the true Higgs mass (125 GeV).

background-only (B) and signal (S) $m_{4\ell}$ histogram shape templates using simulated events, then fit the observed $m_{4\ell}$ histogram to the background-only (B) or signal + background ($S + B$) hypotheses. The statistical significance of an observation is then computed from the delta log-likelihood test statistic obtained from the two fits: $\Delta \log \mathcal{L} = \log \mathcal{L}_{S+B} - \log \mathcal{L}_B$. We view this as the baseline sensitivity for discovery.

- **AutoSciDACT NPLM** We run NPLM on the four-dimensional embedding space obtained from the contrastive encoder. We expect the presence of domain shift between the simulated and observed data (i.e. mis-modeling) to somewhat weaken the sensitivity of discovery.
- **Supervised AutoSciDACT NPLM** We re-train our contrastive encoder with simulated Higgs events included in the training set, so that it explicitly learns about the signal of interest. The Higgs signal should be well-separated from backgrounds in the learned embedding space, so we expect sensitivity in this case to be on par with the baseline obtained from the $m_{4\ell}$ fit.

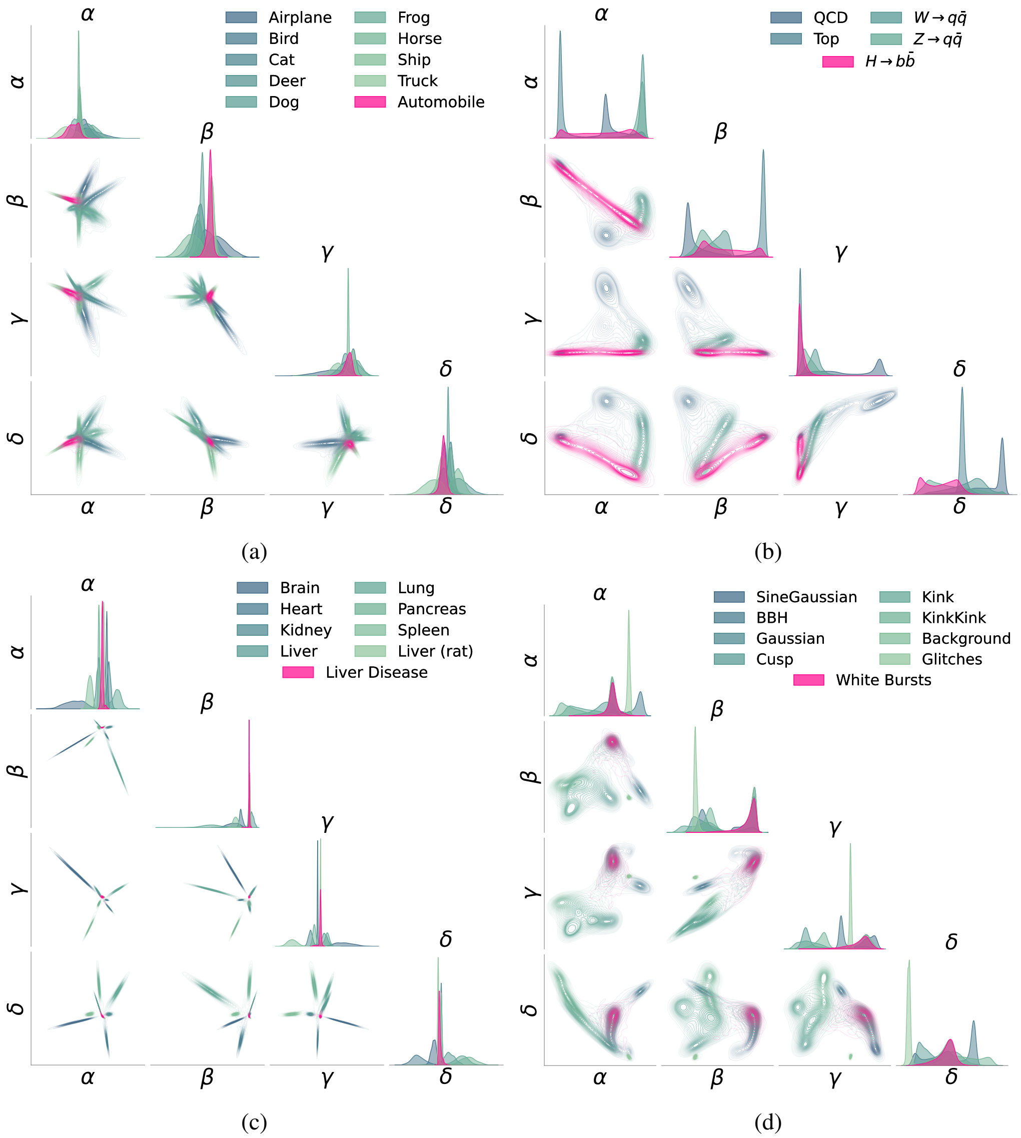
We present results from our analysis in Fig. 15. The top row shows expected test statistic distributions and expected/observed p-values for the baseline (left) and *supervised* AutoSciDACT (right) methods. Both methods achieve similar expected performance, with the observed p-values differing somewhat in either direction due to known domain shifts between simulated and real LHC data. These shifts lead to a larger spread in observed p-values, and given that we are not properly accounting for systematic uncertainties in our analysis, they have a noticeable effect. The expected p-values are thus more informative, as they provide a clear impression of how effectively each method finds a Higgs signal without the impact of domain shift.

Figure 15(c) shows results using *standard* AutoSciDACT (i.e. a contrastive space trained without knowledge of the Higgs). As expected, it is somewhat less sensitive than the baseline methods. However, even the baselines do not achieve highly significant p-values, largely due to the small dataset size and the fact that we only consider one Higgs decay mode[8] To better interpret these results, in Fig. 15(d) we plot two-dimensional distributions of $m_{4\ell}$ versus NPLM anomaly score for simulated backgrounds (filled) and observed data (points). We observe that the most anomalous points in *real data* are clustered near 120-130 GeV, near the known mass of the Higgs (125 GeV). We draw two horizontal lines depicting example medium/tight thresholds on the NPLM score, showing how these Higgs-like events could be isolated in a dataset by selecting on anomaly score. Note that a full analysis using this selection would likely increase sensitivity, since it would use both the 4-lepton mass and the anomaly score.

F Embedding Space Visualizations

For reference, we include visualizations of the four-dimensional contrastive embedding spaces for the CIFAR-10, JetClass, LIGO, and histology datasets in Figure 16.

[8] The expected significance in this channel was only $\sim 2\sigma$ in the original CMS paper [122], with discovery claimed only by combining results from several channels.



(a)

(b)

(c)

(d)

Figure 16: Corner plots showing the four-dimensional contrastive embedding spaces for CIFAR-10 (a), JetClass (b), histology (c), and LIGO (d). The turqoise clusters correspond to the classes used in training the encoder, and the pink cluster shows the distribution of "anomalous" signal in the learned space.