Intro to Anomaly Detection in Particle Physics

Oz Amram

May 2024

Contents

1 Introduction

We collect data from many billions of collisions at the LHC and are typically interested in signal processes that occur only once in a billion or less. A simplified picture of how LHC analysis works is :

basic selection → signal enhancing selection → statistical analysis

The basic selection usually defines what type of objects you are going to be looking at. Eg. you might pick out all events with two jets, or all events with two electrons, or one muon and one jet, etc.

The signal enhancing selection is an additional selection applied to remove background events while keeping as many signal events as possible. The selection criteria can be based on the kinematics of the objects (ie angles and momenta), global event variables (ie missing energy) or sub-features of the objects themselves (ie jet substructure). As there are often exist many variables in which signal and background differ, nowadays many analyses use mutlivariate methods like Boosted Decision Trees or Neural Networks to implement the selection. The selection is typically chosen/optimized based on a targeted signal model and the known backgrounds ML-based classifiers are typically trained using simulated signal and background events. While our simulations are imperfect, they typically of high enough quality to learn good classification variables.

However, one trains a classifier to identify a particular signal, you have no guarantee that it will have any sensitivity to any other signal even if they seem to be 'very similar' from a physics perspective.

The final statistical analysis is done to determine whether a signal is present in the selected data or not. Often a fit is performed in a 1D variable in which signal and background are known to have different distributions (eg invariant mass). Different techniques (some using simulation, some data driven) are used to model the background contribution depending on the type of background is present and what selections have been applied. One must always be sure that whatever selection is applied previously will still allow a background estimate to be performed after. In particular, certain selections can bias the background distribution to look similar to the signal in the 1D variable you would like to fit, which makes estimating the background very difficult. So often you want to decorrelate the selection with the final observable you intedn to fit.

What I will refer to as anomaly detection, is attempting to perform the middle step of signal enhancing selection without any reference to a particular signal model. Ie, a model-agnostic reduction of background. The philosophy behind these methods is that we know there is a large space of potential signals that could exist in our data (and likely more that we haven't thought of!) and it is unfeasible to perform a dedicated search for each signal. The hope is that anomaly detection can help us make sure we aren't missing any signal thats hiding in the data. For any particular signal model anomaly detection methods are generally less sensitive than a dedicated search which targets that signal. However because they are designed in a model-agnostic fashion they hopefully are sensitive to a much larger variety of signals. We do not claim that these methods are entirely 'model-independent' such that they have the same sensitivity to all possible signals, rather just that they have some sensitivity to some large-ish class of signals such that we could hope they could discover something previously missed.

In addition to rejecting the background, important consideration when applying such an anomaly cut is to not spoil the final statistical analysis by introducing some bias to the background estimate.

There are an additional set of proposed techniques which attempt to perform the latter two steps at once in a model-agnostic fashion (ie 'New Physics Learning Machine' [1, 2, 3]), essentially by performing a goodness of fit measurement between the data and the SM background in a high dimensional space. These techniques are interesting but so far rely on very high quality simulation of the SM background and so would have a restricted use cases and I will not focus on them here.

The HEP-ML living review [4] is a good resources containing $\sim$ all HEP anomaly detection papers.

2 Classification

This problem of selecting anomalies is a classification problem, just with an *a priori* unknown signal. From the Neyman-Pearson Lemma we know the optimal quantity to distinguish signal versus background is the likelihood ratio:

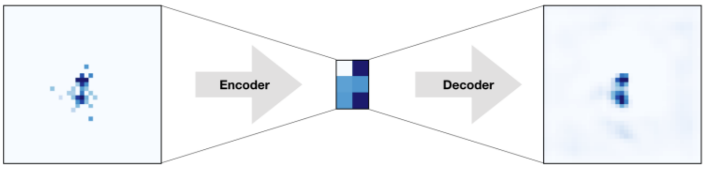

Figure 1: A schematic image of an autoencoder. Taken from [8]

$$L_{S/B}(X) = \frac{P_s(X)}{P_b(X)} \tag{1}$$

Where P_s (P_b) is the probability density of signal (background). The challenge for anomaly detection is that if we don't specify a signal model, we don't know P_s. Generally the set of techniques for anomaly detection can be split into two classes.

The first class of techniques is based on the idea of **outlier detection**. Usually we are dominated by background events in our data sample. So even if one knows nothing about the signal, one can learn about the background distribution. Then we can call events which seem very unlikely under the background distribution as anomalies. Essentially this defines $\frac{1}{P_b(X)}$ as an anomaly score.

The second set of techniques essentially leverage the fact that in HEP our signals are often **overdense** in a particular localized region. This fact is then used to learn $\frac{P_s(X)}{P_b(X)}$ from the data itself. These techniques rely on domain knowledge to determine regions where a signal may be localized and control regions which would not have signal but do have very similar backgrounds. These techniques have the nice property that their anomaly score converges to s_{opt} in the limit of large statistics/large signals.

These two paradigms have different pro's and con's and may be applied in different scenarios.

There are hybrid approaches which attempt to live somewhere in between pure model agnostic approaches and traditional analysis. These usually use some representative signal models as a loose prior [5, 6]. These methods are interesting but I will not discuss them here.

3 Outlier Detection Methods

We generally do not know the full likelihood function of the Standard Model, especially for complicated objects like jets. So often we train a network on background dominated data such that we can learn a proxy for P_b.

Most of the work along these lines have used autoencoders (first proposed in [7, 8]) as a the proxy for learning P_b. Autoencoders are a type of neural network that takes input data of some dimension $X \sim \mathbb{R}^d$ and then 'encodes' (E) it into some latent space of smaller dimension $Z \sim \mathbb{R}^k$ for $k < d$. A decoder network D then takes Z and tries to recovery the original data. The network is therefore defined by $E(X) = Z$ and $D(Z) = X'$. It is usually trained with an L2 loss $\mathcal{L} = ||X - D(E(X))||^2$. This is shown graphically in Figure 1

The idea is that if this autoencoder is trained on a sample of background events, it will learn how to do perform this compression/decompression pretty well for background events. An 'anomalous' event coming from some signal, will then essentially by 'out of distribution' for the networks training sample, and therefore it will perform poorly at this compression/decompression task. We can therefore use the same L2 loss evaluated on new events as an anomaly score. We hope that it encodes something $\sim \frac{1}{P_B(X)}$ (but we are not really guaranteed this).

An alternate strategy to autoencoders is to learn $P_b(X)$ directly either by using a variational autoencoder and evaluating the density in the latent space, or by using a something like a normalizing flow. This can similarly this can be trained on a sample of background events and then applied to other events to find outliers.

Directly learning the probability distribution of extremely high dimensional data can be difficult. So most applications of this method have so far used human-chosen high-level features of dimensionality O(10) that we know to be good for detecting a variety of signals. This is in contrast the autoencoder based methods which generally take as input low-level representations of the object to both reduce human bias and have sufficiently large dimensionality for compression to make sense. VAE's have not seen to give a significant boost in performance compared to autoencoders. With recent advances in generative modeling like diffusion, there is perhaps more room to explore directly learning $P_b(X)$ based on lower level features.

3.1 Challenges and Future Directions

One inherent limitation of all outlier detection / P_b-based methods is that probability densities are not invariant under coordinate transformations (first highlighted in HEP anomaly detection context in [9]). If one transform the data as $y = f(x)$, then the probability density changes to

$$p_y(y) = p_x(f^{-1}(y))|\frac{d}{dy}|f^{-1}(y)| \tag{2}$$

where the last term is the Jacobian of the transformation. For non-trivial transformations, this Jacobian can radically alter the location of high/low density regions of the probability density in x vs y. For example, if $f(x)$ is the cumulative distribution function of x, then in y all points will have uniform density and no point will be rarer than any other.

In practice this means that the choice of data representation and pre-processing define significant inductive biases that control what kind of anomalies the method will be sensitive to. Eg evaluating the density on x or $log(x)$ can have non-trivial impacts on the resulting anomaly score.

Note that for ratios of probability densities, like $L_{S/B}(x)$, the Jacobian in the numerator and denominator cancel and therefore the classification score is invariant.

Another significant challenge of the autoencoder based methods is their so called **'complexity bias'**. Because the anomaly score is based on this compression task, more complex X's (of a higher intrinsic dimension) generally have higher anomaly scores regardless of their presence in the training sample. One manifestation of this bias is that an autoencoder trained on only QCD jets will successfully identify top jets as anomalous, but an autoencoder trained on top jets struggles to identify QCD jets as anomalous [10]. Normalized autoencoders [11] attempt to remedy this issue by turning autoencoders into an energy based probabilistic model. The network then gets penalized for reconstructing well out of distribution data, which makes the network a better representation of P_b than an standard autoencoder. Training this kind of model is often challenging in practice because of stability issues.

A third challenge is in decorrelating the anomaly score of these methods with some chosen feature. Often we would like a selection of the anomaly score not to bias a certain distribution so we can later use it in background estimation. Often these distributions (such as the mass of jet) are falling, make events in the tail generally more anomalous. In some cases, where the distribution we care about is strongly correlated with the other features it has been difficult to achieve this decorrelation.

4 Overdensity-locating methods

Given a set of unlabeled data events it seem impossible to train a classifier to distinguish signal vs background. Suppose however, someone has magically provided you two mixed samples M_1 and M_2 which are composed of a mixture of signal and background events and has told you that M_1 has a larger fraction of signal events in it (f_1) than M_2 does (f_2). Then training a classifier to distinguish between from M_1 and M_2 will converge to the optimal signal versus vs background classifier:

$$L_{M1/M2}(X) = \frac{P_{M1}(X)}{P_{M2}(X)} = \frac{f_1 P_s(X) + (1 - f_1)P_b(X)}{f_2 P_s(X) + (1 - f_2)P_b(X)} = \frac{f_1 L_{S/B} + (1 - f_1)}{f_2 L_{S/B} + (1 - f_2)} \tag{3}$$

One can check that for $f_1 > f_2$ this is just a monotonic rescaling of $L_{S/B}$ and therefore defines an equivalent classifier. In practice we often have $f_1 << 1$ and $f_2 \sim 0$, that is one sample has an O(1%) signal fraction and the other is nearly background pure. For $f_2 = 0, L_{M1/M2}$ becomes:

$$L_{M1/M2}(X) = \frac{f_1 P_s(X) + (1 - f_1)P_b(X)}{P_b(X)} = f_1 + (1 - f_1)L_{S/B} \tag{4}$$

which makes the scaling with $L_{S/B}$ more obvious.

This training setup shown graphically in Figure 2.

The key assumption here is that the background events in the two samples are sampled from the same underlying distribution. If this is true, the only way to distinguish the two samples is the difference in relative signal fractions between the two samples so the classifier will learn to distinguish signal versus background. If there is any bias such that the background events from the two samples do not come from the same distribution, then this will typically dominate the loss (because $f_1 << 1$) such that the network will learn this background bias rather than signal vs background discrimination.

This type of training is called 'weak supervision' or Classification Without Labels (CWoLa) in the literature (first proposed in [12]).

The question then becomes how can we define mixed samples in our data that have these properties. The first, and most-studied, case is when our signal

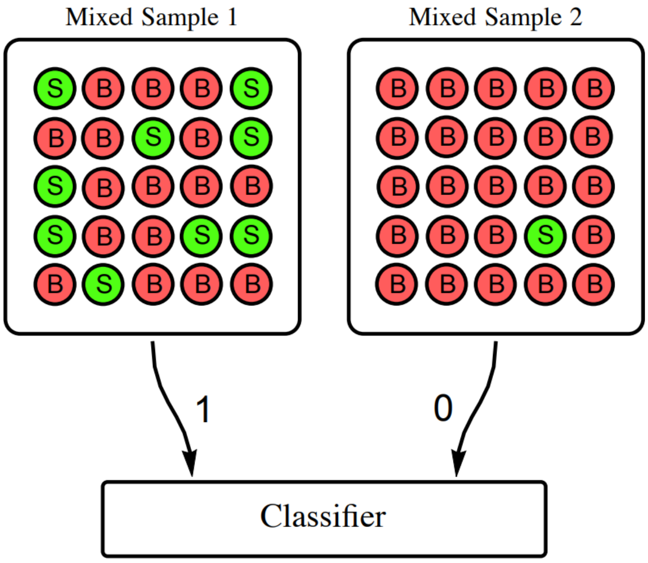

Figure 2: An illustration of weakly supervised training. A classifier is trained to distinguish between two mixed samples of signal and background events. Taken from [12].

is a narrow resonance. In a resonance search, signals manifest as a relatively narrow peak in some invariant mass distribution on top of a larger background that typically has a falling distribution. One can therefore guess a window in this distribution where one hopes that the signal is localized to.

If a signal is present, the mixed sample defined by the events in this window will contain some non-zero signal fraction, while events outside of the window will not. Events in this window can then serve as the potentially signal-rich mixed sample. Events outside this region will have very similar background events to those inside the window, but in general the do not exactly match. This is illustrated in Figure 3

Different approaches have therefore been taken to construct the background rich mixed sample. The original proposal [14, 15] was just to use a weighted sample of the events from the sidebands adjacent to the signal-window. Others have improved upon this approach by training some sort of generative-like model from the sidebands and then interpolating it into the signal region. Samples are then drawn from this generative model to construct the background-rich sample. The first of these methods was CATHODE [13, 16] which used a normalizing flow trained in the sidebands. Other methods include SALAD [17] which uses simulation to help with the interpolation, CURTAINS [18, 19] which 'transports' events from the sidebands, , and FETA [20] which uses transport and simulation. These 4 methods seem to perform roughly similarly [21].

If there is no signal present in the region, the two mixed samples will both contain only background events. As long as there is no bias in the background samples, the classifier will likely pick out some random statistical fluctuation in one sample versus the other, which will not cause an bias/false-signals in the final statistical analysis. The procedure needs to be repeated for different

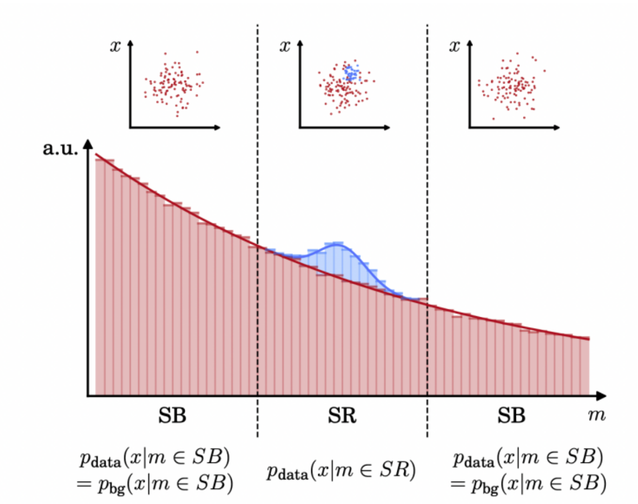

Figure 3: An illustration of the resonance-based overdensity scenario. The signal is localized in particular region of the resonant variable. Within that region there is an overdensity in the feature space from the signal events that is not present in the sidebands. Figure taken from [13].

choices of the signal mass window because we do not know where the signal would manifest *a priori.*

4.1 Challenges and Future Directions

One challenge with these methods is that the weakly supervised training is very 'noisy'. Most of the network's loss is dominated by trying to perform an impossible task of separating identical background distributions. The performance can therefore change dramatically as a function of the amount of signal present in the data. We would like to be sensitive to small signal fractions and also include as large a feature space as possible to be sensitive to many different possible anomalies. But these two goals have been found to be somewhat in tension, the larger the input feature space the easier it is for the network to overfit the background differences and thus you lose sensitivity to small signal fractions (though the general approach can still work [22]). This means that in general we are restricted to using a set of hand-picked features rather than low-level ones. Recent work [23, 24] has also showed that using large ensembles of Boosted decision trees rather than neural networks allows you to use a larger feature space without as much loss in performance.

Another direction is to find additional strategies to define these mixed samples such that they can be applied to more use cases rather than just resonances (or additional 'tricks' to learn the likelihood ratio [25] in data). One approach is to factorize your signal into two parts which are uncorrelated for background and use an outlier detector one side to construct mixed samples for the other [26]. Other ideas are based on finding control regions which have identical backgrounds to the signal region for specific analysis cases [27, 28]. One could hope to look for anomalies in the tails of distributions, which requires extrapolation of your background model, which is difficult to control for ML models working in high dimensions. One approach tried extrapolating from two directions to better control this and found some success [29] I think there is more to explore here.

It is also interesting to ask what domains outside of HEP have anomalies that manifest as similar over-densities to which these techniques could be applied.

5 Results using these techniques

There have so far been a limited number of experimental results using these techniques. The first LHC anomaly detection result as a search for dijet resonances using weak supervision[30] from ATLAS. However as this was a first foray into these techniques it used a very limited 2D feature space to look for anomalies (just the invariant masses of the two jets). The first search using outlier detection methods was a search for resonance decaying to Higgs plus an anomalous jet [31] This search utilized a form of an autoencoder to tag the anomalous jet. Another ATLAS search [32] looked for resonances in various two-body invariant mass distributions (lepton-jet, jet-jet, etc). Event-level characteristics like the number, momentum and angles of all reconstructed particles are encoded into a matrix that is used as input to an autoencoder to select anomalous events.

A very recent CMS search looked for dijet resonances with anomalous jet substructure [33] and employed multiple different anomaly detection methods as complementary methods. There was one outlier-detection detection based on a variational autoencoder, three methods using weak supervision and a hybrid method that used both outlier detection and some signal priors.

The first application of these techniques outside HEP has been to the automated detection of stellar streams [34, 35, 36, 37].

Hopefully the number of experimental results using these techniques will grow in the coming years.

References

[1] Raffaele Tito D'Agnolo and Andrea Wulzer. Learning New Physics from a Machine. *Phys. Rev.*, D99(1):015014, 2019.

[2] Raffaele Tito D'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning Multivariate New Physics. 2019.

[3] Raffaele Tito d'Agnolo, Gaia Grosso, Maurizio Pierini, Andrea Wulzer, and Marco Zanetti. Learning New Physics from an Imperfect Machine. *Eur.Phys.J.C*, 82:275, 11 2021.

[4] HEP ML Community. A Living Review of Machine Learning for Particle Physics.

[5] 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.

[6] Chi Lung Cheng, Gurpreet Singh, and Benjamin Nachman. Incorporating Physical Priors into Weakly-Supervised Anomaly Detection. 5 2024.

[7] Theo Heimel, Gregor Kasieczka, Tilman Plehn, and Jennifer M. Thompson. QCD or What? *SciPost Phys.*, 6(3):030, 2019.

[8] Marco Farina, Yuichiro Nakai, and David Shih. Searching for New Physics with Deep Autoencoders. 2018.

[9] Gregor Kasieczka, Radha Mastandrea, Vinicius Mikuni, Benjamin Nachman, Mariel Pettee, and David Shih. Anomaly Detection under Coordinate Transformations. *Phys.Rev.D*, 107:015009, 9 2022.

[10] Thorsten Buss, Barry M. Dillon, Thorben Finke, Michael Krämer, Alessandro Morandini, Alexander Mück, Ivan Oleksiyuk, and Tilman Plehn. What's Anomalous in LHC Jets? *SciPost Phys.*, 15:168, 2 2022.

[11] Barry M. Dillon, Luigi Favaro, Tilman Plehn, Peter Sorrenson, and Michael Krämer. A Normalized Autoencoder for LHC Triggers. *SciPost Phys.Core*, 6:074, 6 2022.

[12] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP*, 10:174, 2017.

[13] Anna Hallin, Joshua Isaacson, Gregor Kasieczka, Claudius Krause, Benjamin Nachman, Tobias Quadfasel, Matthias Schlaffer, David Shih, and Manuel Sommerhalder. Classifying Anomalies THrough Outer Density Estimation (CATHODE). *Phys.Rev.D*, 106:055006, 9 2021.

[14] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Anomaly Detection for Resonant New Physics with Machine Learning. *Phys. Rev. Lett.*, 121(24):241803, 2018.

[15] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Extending the search for new resonances with machine learning. *Phys. Rev.*, D99(1):014038, 2019.

[16] Anna Hallin, Gregor Kasieczka, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. Resonant anomaly detection without background sculpting. *Phys.Rev.D*, 107:114012, 10 2022.

[17] Mayee F. Chen, Benjamin Nachman, and Frederic Sala. Resonant anomaly detection with multiple reference datasets. *JHEP*, 07:188, 2023.

[18] John Andrew Raine, Samuel Klein, Debajyoti Sengupta, and Tobias Golling. CURTAINs for your Sliding Window: Constructing Unobserved Regions by Transforming Adjacent Intervals. *Front.Big Data*, 6:899345, 3 2022.

[19] Debajyoti Sengupta, Samuel Klein, John Andrew Raine, and Tobias Golling. CURTAINs Flows For Flows: Constructing Unobserved Regions with Maximum Likelihood Estimation. 5 2023.

[20] Tobias Golling, Samuel Klein, Radha Mastandrea, and Benjamin Nachman. Flow-enhanced transportation for anomaly detection. *Phys. Rev. D*, 107(9):096025, 2023.

[21] Tobias Golling, Gregor Kasieczka, Claudius Krause, Radha Mastandrea, Benjamin Nachman, John Andrew Raine, Debajyoti Sengupta, David Shih, and Manuel Sommerhalder. The interplay of machine learning-based resonant anomaly detection methods. *Eur. Phys. J. C*, 84(3):241, 2024.

[22] Erik Buhmann, Cedric Ewen, Gregor Kasieczka, Vinicius Mikuni, Benjamin Nachman, and David Shih. Full phase space resonant anomaly detection. *Phys. Rev. D*, 109(5):055015, 2024.

[23] Thorben Finke, Marie Hein, Gregor Kasieczka, Michael Krämer, Alexander Mück, Parada Prangchaikul, Tobias Quadfasel, David Shih, and Manuel Sommerhalder. Back To The Roots: Tree-Based Algorithms for Weakly Supervised Anomaly Detection. *Phys.Rev.D*, 109:034033, 9 2023.

[24] Marat Freytsis, Maxim Perelstein, and Yik Chuen San. Anomaly Detection in Presence of Irrelevant Features. *JHEP*, 02:220, 10 2023.

[25] Eric M. Metodiev, Jesse Thaler, and Raymond Wynne. Anomaly Detection in Collider Physics via Factorized Observables. 11 2023.

[26] Oz Amram and Cristina Mantilla Suarez. Tag N' Train: A Technique to Train Improved Classifiers on Unlabeled Data. 2 2020.

[27] Thorben Finke, Michael Krämer, Maximilian Lipp, and Alexander Mück. Boosting mono-jet searches with model-agnostic machine learning. *JHEP*, 08:015, 2022.

[28] Gregor Kasieczka, John Andrew Raine, David Shih, and Aman Upadhyay. Complete Optimal Non-Resonant Anomaly Detection. 4 2024.

[29] Kehang Bai, Radha Mastandrea, and Benjamin Nachman. Non-resonant anomaly detection with background extrapolation. *JHEP*, 04:059, 2024.

[30] ATLAS Collaboration. Dijet resonance search with weak supervision using 13 TeV pp collisions in the ATLAS detector. 2020.

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

[32] Georges Aad et al. Search for New Phenomena in Two-Body Invariant Mass Distributions Using Unsupervised Machine Learning for Anomaly Detection at s=13 TeV with the ATLAS Detector. *Phys. Rev. Lett.*, 132(8):081801, 2024.

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

[34] David Shih, Matthew R. Buckley, Lina Necib, and John Tamanas. Via Machinae: Searching for Stellar Streams using Unsupervised Machine Learning. *Mon.Not.Roy.Astron.Soc.*, 509:5992, 4 2021.

[35] David Shih, Matthew R. Buckley, and Lina Necib. Via machinae 2.0: Full-sky, model-agnostic search for stellar streams in gaia dr2, 2023.

[36] Mariel Pettee, Sowmya Thanvantri, Benjamin Nachman, David Shih, Matthew R. Buckley, and Jack H. Collins. Weakly-supervised anomaly detection in the milky way, 2023.

[37] Debajyoti Sengupta, Stephen Mulligan, David Shih, John Andrew Raine, and Tobias Golling. Skycurtains: Model agnostic search for stellar streams with gaia data, 2024.