INDIAN INSTITUTE OF TECHNOLOGY INDORE

CLUSTERING AND CLASSIFICATION OF GAMMA RAY BURSTS USING MACHINE LEARNING TECHNIQUES

M.Sc. Thesis

By

HARIKRISHNAN R

DEPARTMENT OF ASTRONOMY ASTROPHYSICS AND SPACE ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

CLUSTERING AND CLASSIFICATION OF GAMMA RAY BURSTS USING MACHINE LEARNING TECHNIQUES

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of

Master of Science

by

HARIKRISHNAN R

DEPARTMENT OF ASTRONOMY,ASTROPHYSICS AND SPACE ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled Clustering and Classification of Gamma-Ray Bursts using Machine Learning Techniques in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July-2023 to May-2025 under the supervision of Dr. Amit Shukla, Associate professor, DAASE.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

13-05-2025

Signature of the student with date
Harikrishnan R

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

13-05-2025

Signature of the Supervisor of
M.Sc. thesis (with date)
Dr. Amit Shukla

Harikrishnan R has successfully given his M.Sc. Oral Examination held on **05-05-2025** .

13-05-2025

Signature(s) of Supervisor(s) of MSc thesis
Date:

Programme Coordinator, M.Sc.
Date: 19/05/2025

Convenor, DPGC
Date: 19/05/2025

HoD, DAASE
Date:

ACKNOWLEDGEMENTS

"Gratitude is not only the greatest of virtues but the parent of all others." - Cisero

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Amit Shukla, for his consistent guidance, unwavering support, and immense patience throughout the course of my thesis. His timely insights and encouragement have been instrumental in shaping this work.

I am also deeply thankful to the Department of Astronomy, Astrophysics and Space Engineering, and to IIT Indore for providing an encouraging environment and the resources essential for carrying out this project.

A special word of thanks to my collaborators, Shraddha Mohnani and Dr. Suman Bala. Their involvement, timely suggestions, and valuable advice were crucial in navigating this research in the right direction — this work would not have been possible without them.

I gratefully acknowledge Dr. Sourabh Das for his thoughtful suggestions, which brought new perspectives and helped me move forward with fresh ideas.

To my lab mates — Daisy, Aman, Chandhan, Ayush, and Sushmita — thank you for the stimulating discussions and the camaraderie that made long days of research lighter and more enjoyable.

I would also like to extend my deepest appreciation to my friends — Vishruth, Vijay, Prasad, Parth, Gitaj, Amar, Devesh, Aryan, Annie, Anushka, Ashutosh, Anand, Navanit, Jithu, Keerthi, Nasmi, Leon, Aromal, Jibin, Swaroop, and Eeshan. Your unwavering love, selfless support, and belief in me have meant more than words can express. You've stood by me through every high and low, and your presence made this journey far less daunting.

This thesis — and the journey behind it — would have been significantly more difficult without all of you. Thank you from the bottom of my heart.

Abstract

Gamma-ray bursts are the most energetic explosions in the universe with a staggering amount of energy release greater than the sun within a fraction of seconds. Although half a decade has passed since its discovery, the physics of GRBs remains enigmatic. As a first step towards unraveling the mysteries of this phenomenon, several classification attempts have been made, but to date none of them have been completely successful. Duration-based classification reveals two distinct classes of GRBs based on their progenators, long and short which are considered to be a byproduct of collapse of massive stars and merger events involving compact objects. But more recently, this classification scheme has been facing a lot of road blocks since the discovery of long GRBs along with kilonovae and short GRBs along with supernovae, which contradicts the traditional classification scheme. The use of state-of-the-art machine learning techniques are instrumental in addressing this issue. So the idea is to address this anomaly in classification using machine learning techniques such as dimensionality reduction and clustering. Clustering the high-dimensional light curve of the GRBs in a two-dimensional plane and identifying the key astrophysical significance of the grouping which could unravel the mysteries related to the physics of the burst, emission mechanism etc, is the idea behind this work. However unsupervised learning like dimensionality reduction and clustering being sort of a black box, it is never completely understood what parameters in the feature space make them to produce a particular embedding or different clusters. Identifying the key parameters that influence the clustering and answering the most important question on the validity of the clusters we obtained before attributing astrophysical significance is done by considering different methods of analyzing the data and trying to see if the results are consistent enough with the changes introduced and the influence of noise in the analysis is verified by simulating data of different types of noise and carrying out the clustering analysis.

Contents

List of Figures

List of Tables

ACRONYMS

Acronym	Full Form
GRB	Gamma Ray Burst
BATSE	Burst And Transient Science Experiment
NS	Neutron star
BH	Black hole
UMAP	Uniform Manifold Approximation and Projection
t-SNE	t-distributed Stochastic Neighborhood Embedding
HDBSCAN	Heirarchical Density Based Clustering in Space with Application to Noise
GMM	Gaussian Mixture Models
ICA	Independent Component Analysis
NMF	Non-negative matrix factorisation
Isomap	Isometric Feature Mapping
LLE	Locally Linear Embedding

Chapter 1

Introduction

1.1 Background

Gamma-ray bursts being the most energetic explosions in the universe are a result of catastrophic events involving neutron stars, black holes, and white dwarfs, or they could be a result of the collapse of massive stars [30]. The amount of energy released during this process which is of the order of 10^{46}J, is more than that our sun emits in its lifetime.

The detection and study of these cataclysmic events involve a range of sophisticated detectors. Some of the space-based observatories for the detection of GRBs are FERMI gamma-ray space telescope with Gamma-ray Burst Monitor (GBM) and Large Area Telescope (LAT) on board, SWIFT with BAT (Burst Alert Telescope) and XRT (X-ray telescope), INTEGRAL and KONUS-WIND.

LAT detects gamma rays of energy around 20 Mev to 300 Gev whereas GBM detects the same around the energy 8 Kev to 40 Mev. GBM consists of 12 NaI detectors which work in the energy range 8 Kev to 900 Kev and 2 BGO detectors in the energy range 900 Kev to 40 Mev. This broadband energy covered by both GBM and LAT provides insights into the different emission mechanisms of GRB thereby facilitating the study of their progenitors, different emission mechanisms, energetics, and their variability.

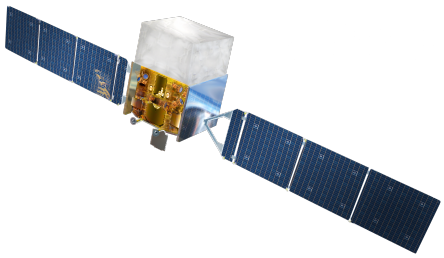


Figure 1.1: Fermi Gamma-Ray Space Telescope

The emission process of GRBs consists of two phases, prompt and afterglow. Prompt emission is the direct emanation of the jet from the central engine whereas afterglow is a result of the interaction of the jet with the external circumburst medium which would drive an external shock and results in the emission process in all the energy bands via synchrotron or self Compton mechanism. The leading framework for the explanation of a GRB model is the Fireball model, which holds the assumption that GRBs are emissions that arise from hot dense plasma that moves at relativistic speeds, formed after a catastrophic event like a compact object merger or a massive star collapse. The internal collisions within this jet result in prompt emission while the external shock produced by the jet with the interaction of external medium results in afterglow emission[15]. Prompt emission could last from a few milliseconds to several hundreds of seconds and is directly correlated with the activity of the central engine whereas the afterglow emission

which ranges from gamma rays to radio waves could last from hours to days. Based on the duration of prompt emission, GRBs can be classified as long and short.

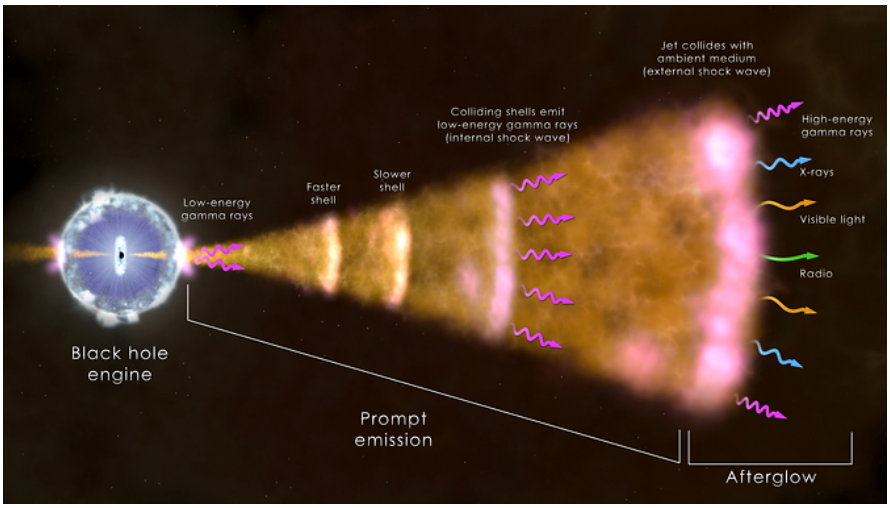


Figure 1.2: GRB Fireball model

The classification of GRBs has been a long-standing problem in the community of astronomy. From the bimodal distribution of the duration (Figure 1.3) obtained from the BATSE on board Compton Gamma-ray Observatory (CGRO), GRBs are traditionally classified as long if the duration which is t90 (time during which 90% of the energy is emitted from the central engine) is greater than 2s and short if the t90 is less than 2s [16]. Moreover, canonically most long GRBs are associated with the collapse of massive stars because of their supernova counterparts [26], whereas short GRBs are attributed to merger events including binary neutron stars, black holes, and white dwarfs[21]. Some of the short GRBs have a close association with kilonova, especially GRB170817A [1] associated with the multi-messenger event GW170817 [3].

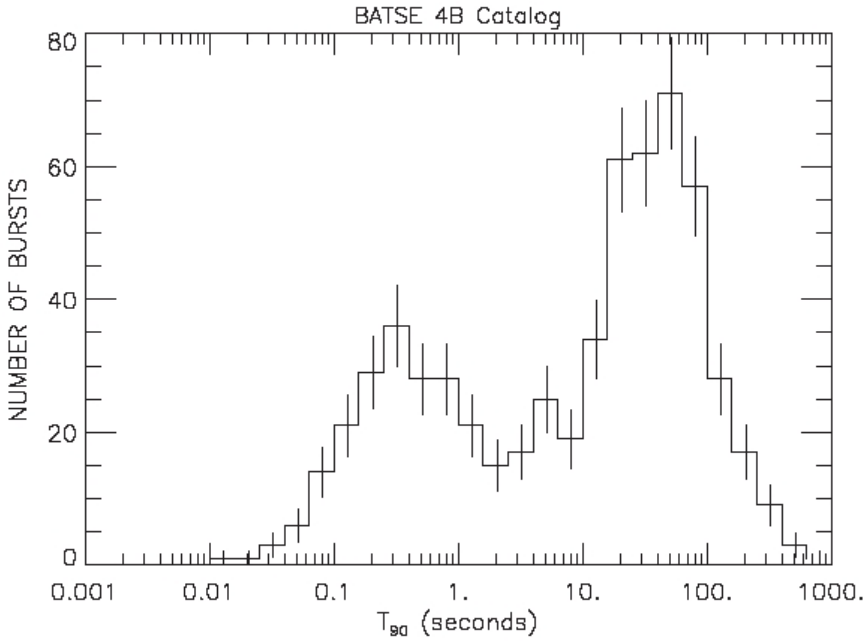


Figure 1.3: Bimodel distribution of GRBs [16]

However, neither all long GRBs have a supernova association nor all short GRBs have a kilonova

association and it is not an easy task to classify the intermediate duration GRBs which fall between 0.5s-3s. That is, along the dividing line of 2s, there is an inherent ambiguity in the classification because of the overlap between the two regions, therefore there is a high chance of misclassifying the shortest long burst and longest short burst. More recently, the classification scenario became pretty complicated with the discovery of long GRBs **GRB211211A and GRB230307A**[22] along with kilonovae which is a hallmark of short GRBs and short **GRB20826A**[32] with a confirmed supernova association.

Therefore we need to find an alternate classification scheme that not only relies on duration but on several other parameters as well. Since there are only a handful of parameters (fluence, redshift, etc) determined for all the detected GRBs, using them for classification may not reflect the actual distinction. One way to tackle this issue is to classify the bursts using the similarities in their light curve. Comparing the light curves of several thousand bursts in different energy ranges can be made possible with state-of-the-art machine-learning techniques. Since the data we have is not labeled, that is the raw light curves are not labeled as a particular burst, we will have to employ unsupervised machine-learning techniques such as clustering methods to identify different populations of GRBs.

There are several caveats associated with the interpretation of the clusters obtained because of the inherent ambiguity involved in these techniques and need for the validation of these clusters since these outcomes from the 'black box' are a result of tuning several data preprocessing steps and the parameters involved in the decision making of grouping the bursts is unknown.

1.2 Objectives

1. To find an unambiguous grouping of GRBs based on their similarities in the light curves, making use of machine learning techniques such as dimensionality reduction and clustering.

2. To perform a comaparative analysis by changing the data preprocessing steps and verify the consistency of results.

3. Check the role of background noise in the light curves on the obtained clusters and verify their astrophysical significance.

1.3 Organization of the Thesis

The structure of the thesis is as follows; the process of light curve extraction, data preprocessing steps, and a detailed view of the machine learning algorithms used are described in the methodology section. Obtained results, caveats in the analysis, and future directions to the work are described in the results, discussion and conclusion part. Several other details regarding the analyses is provided in the Appendix.

Chapter 2

Review of Past Work and Problem Formulation

Recent advances in the field of machine learning have also made significant progress in the field of astronomy. Especially in addressing the classification anomaly, ML based techniques plays a very crucial role [9], [2], [11]. As duration-based classification is limited and poses several challenges whereas a classification based on prompt emission light curves shows that there are several subclasses within the GRB population. Jespersen's [13] work was revolutionary as it was the first paper to address this anomaly using the prompt emission light curves of SWIFT GRBs by taking advantage of the algorithm t-SNE. Results revealed two distinct clustered grouped based on the features of the light curve, which in-turn showed the duration parameter has a strong influence in clustering because the two groups showed the long-short dichotomy as well!, the difference here from the traditional classification scheme is that there are a lot of other parameters involved in the classification inclusive of the duration. Following this work [27] extended this analysis to BATSE [5] and FERMI, and verified the previously obtained results as true.

In a more recent analysis of SWIFT GRBs by Dimple et al[6], five distinct clusters within the GRBs were found, of which two are argued to be associated with kilonovae associated GRBs. Five distinct clusters may point to five different progeneators or different emission mechanisms. A similar analysis was also performed on FERMI-GBM GRBs by Dimple et al [7]. There also five clusters where found and two of the kilonova associated clusters are attributed to the NS-NS mergers and NS-BH mergers.

A similar analysis on the prompt emission light curves of FERMI-GBM GRBs using machine learning techniques and trying to verify the influence of data preprocessing steps in obtained results along with the verification of noise as a parameter in clusterig constitutes the theme of this work. If we could find meaningful clusters it would have profound implications in the field of astronomy, and with the ever-growing multimessenger observations we will be able to validate these findings and it will open more doors within the field with this analysis.

Chapter 3

Methodology

3.1 Data Reduction Pipeline

The data reduction pipeline is developed (by Dr. Suman Bala, collaborator of this work) using the Python package Fermi Gamma-ray Data Tools (GDT provided by Fermi-GBM team). Pipeline preprocessing begins by estimating the least angle / the brightest detector that detected the event. Now TTE (Time Tagged Event) file corresponding to the brightest detector is selected. TTE files of each burst contain the time and count rate information within a header unit of the fits file in an unbinned format. Now using GDT functionalities, data can be binned as we need. A choice of very small binning could capture even the minute variability but the preprocessing would be extremely hard because of the contamination with noise. Whereas a very high binning resolution may give us the exact signal but may miss out essential features of the light curve. Therefore, choice of the binning is really important. For our analysis data/light curves are binned by both 16 ms and 64 ms time resolution. Furthermore, each light curve is extracted into three energy levels, 8-50 Kev, 50-300Kev and 300-900 Kev. Since these light curves are contaminated by noise from other than the GRB source, this background must be removed so as to increase the signal to noise ratio.

3.1.1 Background Fitting

The process begins by selecting the background regions in the light curve before the trigger and after the trigger, now the the source region containing the transient feature is also selected. The very next step is to fit a polynomial of suitable order in the background regions, the polynomial could be of the order between 0 and 4, and is interpolated to the source region. Since each GRB is an exception the selection of source and background regions is different for different GRBs, therefore the process of background fit is carried out individually for each burst. Now the fitted background is removed and light curves are extracted in three different energy ranges. This process of extraction can be easily automated.

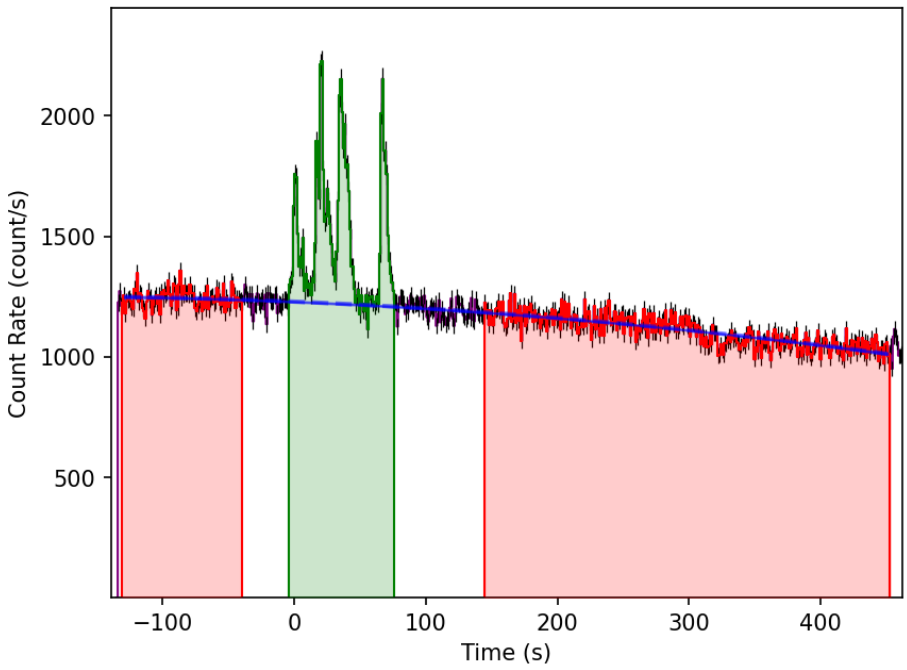


Figure 3.1: Fitting the background of the GRB230814A, the regions in red are the selected background whereas the green region is the source region and the background is fitted with a polynomial of order two.

3.1.2 Wavelet Analysis

The next step in the analysis is to remove the noise from the background fitted light curve. Task is achieved using wavelet analysis [24].

Wavelet analysis is a method parallel to the Fourier transform, the essential difference is that the former offers time and frequency localization and is suitable for nonstationary signals, whereas only either time localization or frequency localization is possible in the latter and cannot be used for non-stationary signals. Wavelet analysis can be used to capture the main transient feature and remove the small random fluctuations in the light curve thereby smoothening it. This can be done by selecting a wavelet from the family of waveforms (Haar, Debauchies (db), symmlet (sym), etc.) that suits well for our purpose. Now, a decomposition level is set so that we can separate our signal into various scales, where typically higher scales(lower coefficients) represent noise; this allows us to separate noise from meaningful features, thereby increasing the signal-to-noise ratio. This is highly efficient in capturing the burst characteristics of light curves.

A wavelet is a localized mathematical object with a high pass filter and low pass filter that convolves with the light curve and separates it into a low-frequency part (approximation coefficients) which gives the overall variability of the light curve and a high-frequency part (detailed coefficients) which contains the small random variations. Now each approximation part is further decomposed into a second layer of low-frequency and high-frequency components. Now this amount of bifurcations is controlled by the decomposition level we set.

Since these high-frequency components contain noise characteristics a universal threshold is applied to the detailed coefficients and the light curve is reconstructed from the remaining coefficients using inverse wavelet transform.

Universal threshold is calculated using the formula,

$$threshold = \sigma \times \sqrt{2ln(N)}$$

where σ is the estimated noise standard deviation and **N** is the length of the signal. We attenuate the coefficients below this threshold, thereby excluding the high-frequency noise. The denoised coefficients are used to reconstruct the light curve, as shown in Figure 3.2.

However, there is a whole zoo of waveforms and decomposition levels we could tweak with, therefore optimal wavelet and decomposition levels is chosen based on the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values by convolving the light curve with different wavelets and decomposition level and the one with low BIC is selected as the optimal one.

AIC and BIC are model selection criterion that can be used to evaluate how well a model fits the data and it penalise a complex model that is the one with a greater number of parameters, in our case the wavelet coefficients.

$$\text{AIC} = 2k + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

$$\text{BIC} = k\ln(n) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

where:

- x_i is the original signal value at time i
- $\hat{x}_i$ is the denoised signal value at time i
- n is the total number of data points
- k is the number of non-zero wavelet coefficients
- σ is the estimated noise standard deviation, computed as $\sigma = \frac{\text{MAD}}{0.6745}$

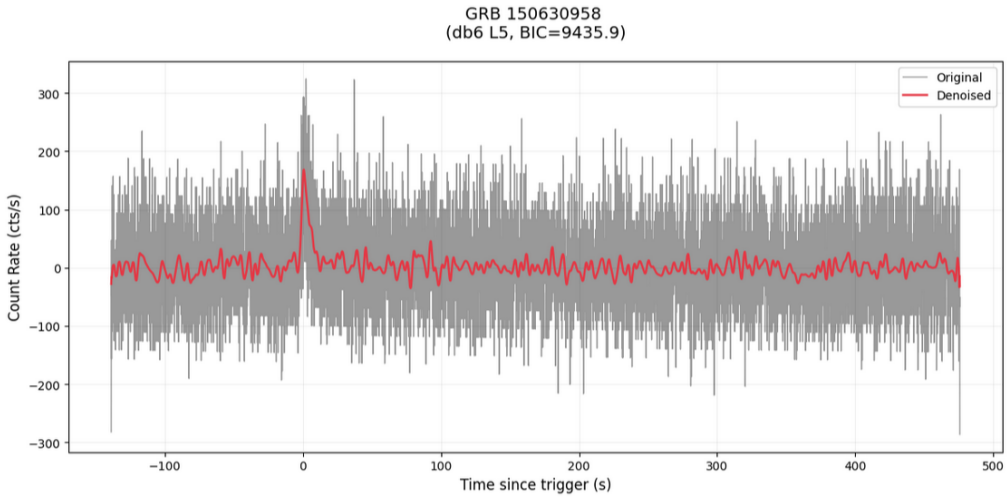


Figure 3.2: Denoising the 16 ms light curve of GRB 150630958 using wavelet analysis with the wavelet db6 and decomposition level 5

Light curve denoising is done for 2000 GRBs in the energy ranges 8-50, 50-300, and 300-900. This process can be automated using the Pywavelet module in Python.

3.1.3 Feature Extraction

Wavelet analysis removes unwanted features that could potentially obscure the dimensionality reduction algorithms in identifying the important features/parameters for clustering or grouping the data. Feature extraction can be done through several methods such as the Fourier transform or more nuansed techniques like the wavelet scattering transform. It is also worth trying to feed the denoised and raw light curves directly to the algorithms without employing their Fourier feature space. The process involves the following steps:

- **Standardization:**Since each light curve has varying lengths it is necessary to homogenize this to avoid possible algorithmic distraction due to different lengths. Therefore, each light curve is either padded with zeros to a fixed length or is truncated to that fixed length. This could help maintain consistent feature lengths across all the GRBs and to have a common starting point. For our analysis, each light curve is extracted from time =0s to time = t90 s. The length of the extracted light curve depends on its t90 value and all the light curves are padded to the light curve with maximum length.

- **Normalization:**It is a known result that mostly long GRBs will have a higher fluence than the short GRBs; therefore in-order to avoid this feature while clustering so that we don't have to render the known result, we will normalize the lightcurves with fluence of that specific energy band. Now these light curves of a single GRB in different energy bands are concatenated together to form a single cohesive time series. Avoiding the normalization procedure would essentially group data based only on the known result. Analysis is carried out for light curves in both three energy band and light curves in single energy band (50-300 Kev).

- **Fourier Transform:**Now the Fourier Transform of the normalized light curves are taken. Now only the Fourier amplitudes are considered while excluding the phase information so as to account for the possible trigger time offset (time at which event happened and time at which event was detected) as mentioned in [13].

  These Fourier amplitude of different light curves stacked together in form of an M $\times$ N matrix forms our Feature vectors, where M is the number of observations or different GRBs where as N is the number of features. The key idea is to use this feature vectors to cluster the GRBs in such a way that GRB light curves of similar amplitudes groups together.

  Arguably, we could use other feature extraction techniques which could absorb the essential features so that it would get clustered based on the astrophysical correlations rather than the artifacts in the data.

3.2 Dimensionality Reduction

One of the main reasons all classification schemes at first converged to considering only a limited number of features, that is, the summary statistics like duration, fluence, Peak energy of the spectrum, etc., is because of the dimensionality curse in addressing the analysis of such a huge volume of data. The clustering process is a cumbersome task in a high-dimensional space because of the amount of redundant features, and due to the task of visualization. Therefore, we need a method that considers only the important features by reducing the dimension which would ease out the computational and visualization problems.

This is where state-of-the-art machine learning techniques come in handy. With the use of dimensionality reduction algorithms, we can reduce the full light curves into points in a two-dimensional plane, which can be further analyzed to find different clusters and sub-clusters.

There are several dimensional reduction techniques from the classic Principal Component Analysis (PCA) to several deep learning architectures. Following are the major classes.

1. **Matrix Factorization Techniques:**
    - Singular Value Decomposition (SVD) [31]
    - Principal Component Analysis (PCA) [14]
    - Kernel PCA [29]
    - Independent Component Analysis (ICA) [25]
    - Non-negative matrix factorisation (NMF) [23]
2. **Non-linear Manifold Learning Methods**
    - Isomap [19]
    - t-distributed stochastic neighbourhood embedding (t-SNE) ([4])
    - Uniform Manifold Approximation and Projection (UMAP) [18]
    - Pairwise Correlated Manifold Approximation and Projection (PaCMAP)
    - Local Linear Embedding (LLE) [12]

- **Uniform Manifold Approximation and Projection(UMAP)** [18] is a highly efficient algorithm which does this job, it is a manifold learning method which is based on topology and graph theory.

  The process begins with the construction of graph with nodes as data points and edges as a similarity measure between the data points, in the high-dimensional space. The parameter defined by the user n-neighbors determines the closest neighbors to be considered while constructing the graph, and min-dist determines the minimum distance between the data points. A similar process is initiated in the low-dimensional space with the graph initialized by PCA(Principal Component Analysis) or by methods such as spectral embedding.

Now, the constructed graphs in both dimensions are matched by optimizing the loss function, which is known as cross-entropy loss.

UMAP is really good at preserving both local structure and global structure of the data compared to t-SNE which is only good at preserving the local structure.

- **Principal component Analysis(PCA)** is a linear dimensionality reduction technique and is the oldest one in the list. It captures the direction of maximum variance and get rid of the redundant features. The process begins by calculating the covariance matrix of the centered data. Eigen values of the covariance matrix are calculated, now eigen vectors, which gives the direction of maximum variance, of the top n eigen values are also calculated. These eigen vectors form the new set of orthogonal axes in which the data instances are projected.

Several other methods exist, including modern approaches like semi-supervision, self-supervision, and transfer and domain adaption algorithms. Neural Network dimensionality reduction methods like AUTOENCODERS are also widely used, where the input layer and output layer of the network will be the same and in between these layers, we will get a reduced representation of the data.

For our analysis we use UMAP with and without initializing with PCA and the results can be compared for light curves in different energy ranges (8-50 Kev, 50-300 Kev and 300-900 Kev) and light curves with different time binning (64 ms and 16 ms).

3.3 Clustering

The very next step in our analysis is to cluster the light curves in the two dimensional embedding obtained. In order to accomplish the task there are several clustering algorithms, like which are centroid based , density based, and a combination of both. The choice of the clustering algorithm is very important. K-means and Gaussian Mixture Model (GMM) are some of the centroid-based algorithms whereas DBSCAN is a density based algorithm and HDBSCAN is a heirarchical density based algorithm

- **K-means** clustering involves three major steps, the first is to choose the value of k or the number of clusters we need in prior, now all data points will be assigned to these k clusters by finding the centroid of these clusters. Data points will be assigned to a particular cluster in a such a manner that each data point will be closer to that centroid of the cluster where the data point belongs than to any other clusters. The very next step is to recalculate the centroids as the mean of all the data points within the cluster. These steps are repeated until the centroid stops moving or until the points stop switching the clusters.

  K-means is the most popular clustering algorithm because of its easy handedness and its efficacy in identifying the spherical clusters if we know the number of clusters beforehand. Otherwise we can estimate the number of clusters using the **elbow method**.

- **GMM** is a more generalization of k-means, but unlike k-means it is not a hard clustering method where a data point is completely assigned to a single cluster. GMM takes the assumption that data points are drawn from a mixture of k different gaussian distributions where k denotes the number

of clusters. Here the method of clustering is soft in the sense that each data point will have a probability of belonging to each of the clusters in such a way that it sums up to one. Parameters of the gaussian distribution are determined using the **Expectation-Maximization** algorithm where the likelihood of the data is estimated using random initialization of the parameter space and the process is iterated until convergence.

Some of the major cons of these centroid based or probabilistic algorithms are that there is an inherent "spherical ball" assumption where we consider the data points are always drawn from gaussian distributions and the number of clusters we need are known in prior. These two issues are addressed in density based algorithms. One such algorithm is DBSCAN.

- **DBSCAN (Density Based Spatial Clustering with Application to Noise) :** first step is to set two hyper-parameters epsilon(ε), which is the maximum distance between two points to be considered as a part of a cluster and minPoints which is the minimum number of points required to form a cluster. The data points with minPoints within a distance of epsilon are considered as core points and these core points are grouped together to form a cluster. Points that are not within an epsilon distance from any core point and are not part of any clusters are considered noise. However, one of the major con of this algorithm is that the parameters are not dynamical which makes it extremely hard to tune them. This is issue is addressed in HDBSCAN.

- **HDBSCAN (Heirarchical Density Based Spatial Clustering with Application to Noise) :** [17] is an updated version of DBSCAN where we consider multiple values of epsilon and make a tree of clusters with each of these values, where at the top each data point will be grouped together to form a single cluster and in the root each data point will individually form a single cluster. Optimal number of clusters are found by eliminating those with very poor densities and most stable clusters are found. Stability is found out by how long a cluster exists as we move through different epsilon values.

The choice of a particular clustering algorithm along with a dimensional reduction algorithm is arbitrary. The analyses have been carried out with different combinations like UMAP-GMM, UMAP-DBSCAN, UMAP-HDBSCAN, etc and it is found that UMAP-HDBSCAN combination is the preferred one [18] since the grouping done by HDBSCAN is density based in a heirarchichal manner which has added advantage in our case as we don't know the number of clusters beforehand and is easier to identify the population outliers and inter-cluster outliers that could be more interesting to study than the clusters itself.

Chapter 4

Results and Discussion

4.1 Analysis with 16 ms Light curves

This section presents the dimensionality reduction and clustering results for the 16 ms binned light curves, both without and with PCA preprocessing. The analysis explores light curves in a single energy band and in all the three energy bands.

4.1.1 Analysis for 50-300 Kev light curves Without PCA Preprocessing

UMAP embedding gives the most crystalline version of our light curves Figure 4.1, with points being close together are those light curves with similar characteristics whereas the greater the neighborhood distance, the greater the dissimilarity. It is a conundrum to answer which all the parameters driving the processing of clustering ie, it is paramount to understand why certain groups are positioned together and why some are clustered apart. As a first step towards this understanding, we need to identify the number of clusters, which here is done using HDBSCAN (Figure 4.2). We can see that HDBSCAN identifies three distinct clusters with two large groups and one smaller group.

To extract meaningful information, the embedding is overlayed with a log t90 duration map of GRBs to see whether it forms any particular pattern (Figure 4.3). As we can see the distinction between long GRBs (colored yellow) and short GRBs (colored violet) in a gradient form from top left to bottom right. This indeed shows the impact of duration as a parameter in driving the process of clustering. However one pertinent question arises, is it the only parameter of influence here? Definitely, it doesn't seem so because we can see that there are two clusters occupied by the long GRBs, cluster ID zero and cluster ID one (hereafter mentioned as long clusters) identified by HDBSCAN. Had it been only duration influencing the process of clustering, there would have been only a long-short dichotomy in the embedding.

To identify if this separation of long clusters is because of the artifacts in the data, power-law indices from the power law fitted on the power spectrum of the light curves are extracted, details are shown in the Appendix. These power-law indices are mapped onto the embedding as shown in the Figure 4.4. It is evident from the figure that the short group of long GRBs are separated because of the difference in their noise characteristics, given by the steeper power index grouping.

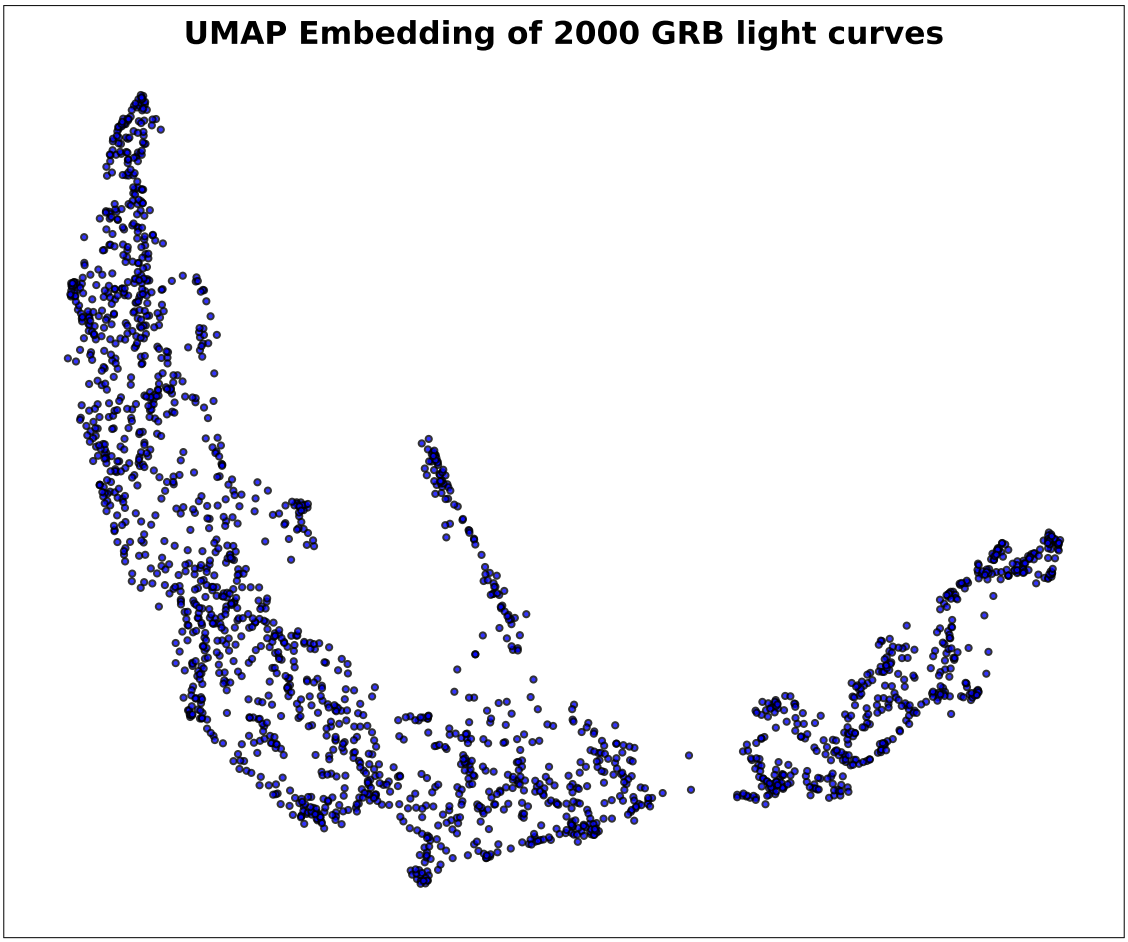


Figure 4.1: UMAP embedding for 50–300 keV light curves. The points which are close to each other are called the neighbors and the points which are far apart are non-neighbors. Neighbors are those GRBs with similar morphological features in their light curve or conversely similar features in their power spectra. Different clusters are identified by the density of grouping using the HDBSCAN algorithm.

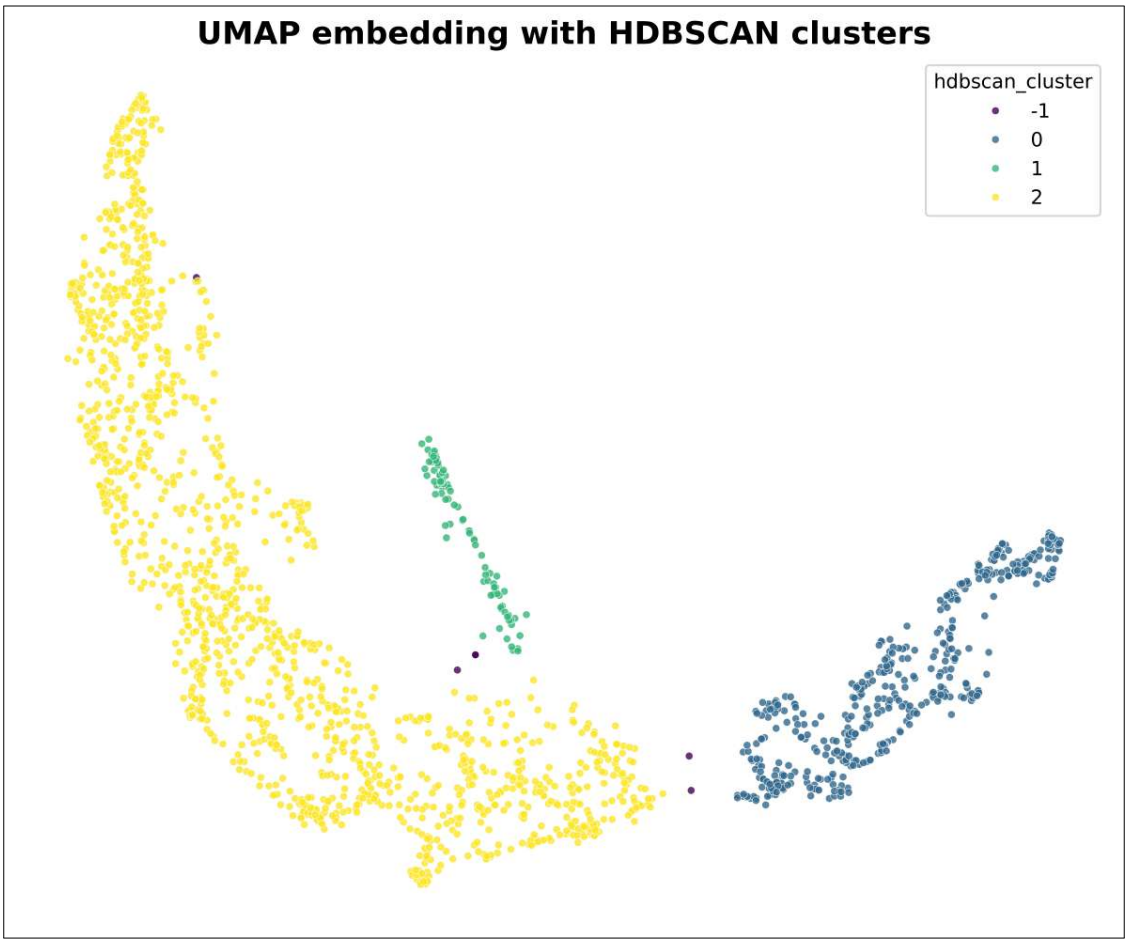


Figure 4.2: HDBSCAN identifies three distinct clusters in with outliers labeled as -1.

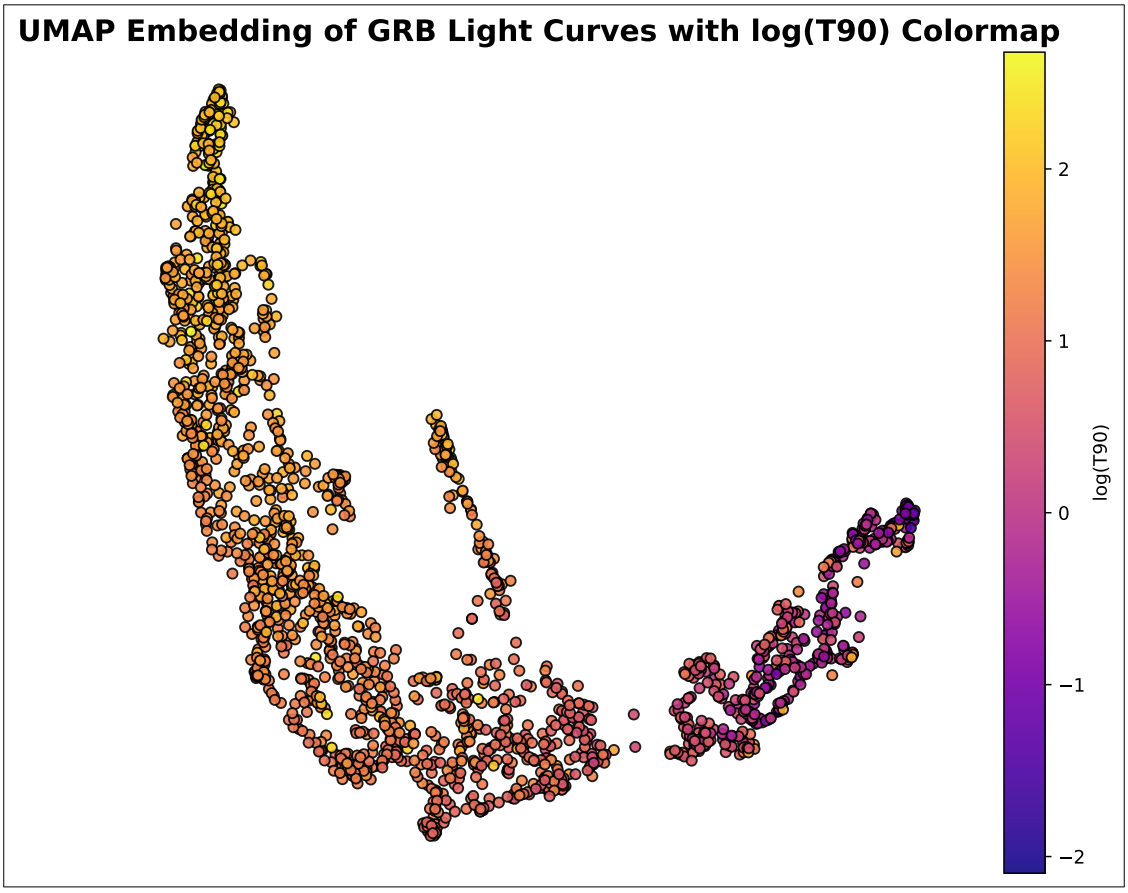


Figure 4.3: UMAP embedding color-coded by duration for 50–300 keV.

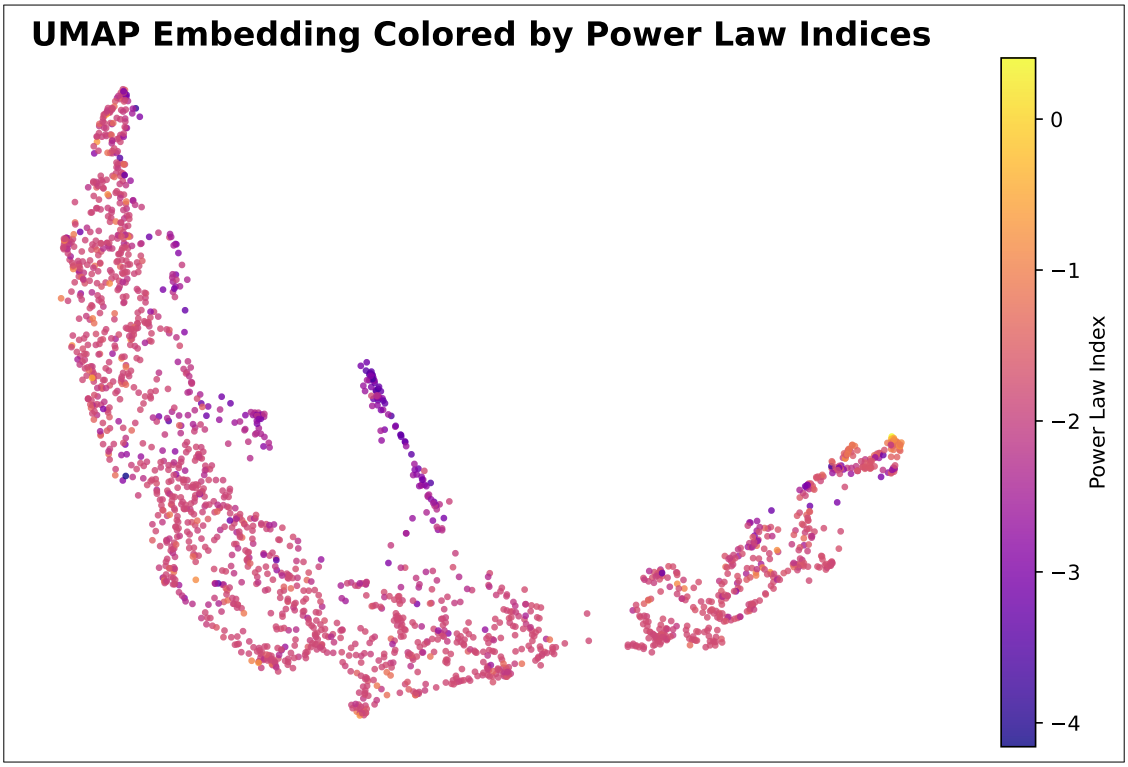


Figure 4.4: UMAP embedding color-coded by the power-law index for 50–300 keV. It can be seen from the figure that power indices vary from zero to a steeper value of four. It can also be noticed that overall embedding has a uniform noise composition with most of the light curves with a value of -2, ie, predominantly red noise. But a closer examination reveals that the cluster marked in green in the HDBSCAN embedding (Figure 4.2) has steeper power indices and this could be the reason it is separated out as a distinct cluster. Also, it could be because of this uniform noise composition in the light curves that they are not forming very distinct clusters. Remarkably short GRBs towards the tip of the right cluster has a white noise composition, it is also grouped together.

4.1.2 Analysis for 50-300 Kev With PCA Pre-processing

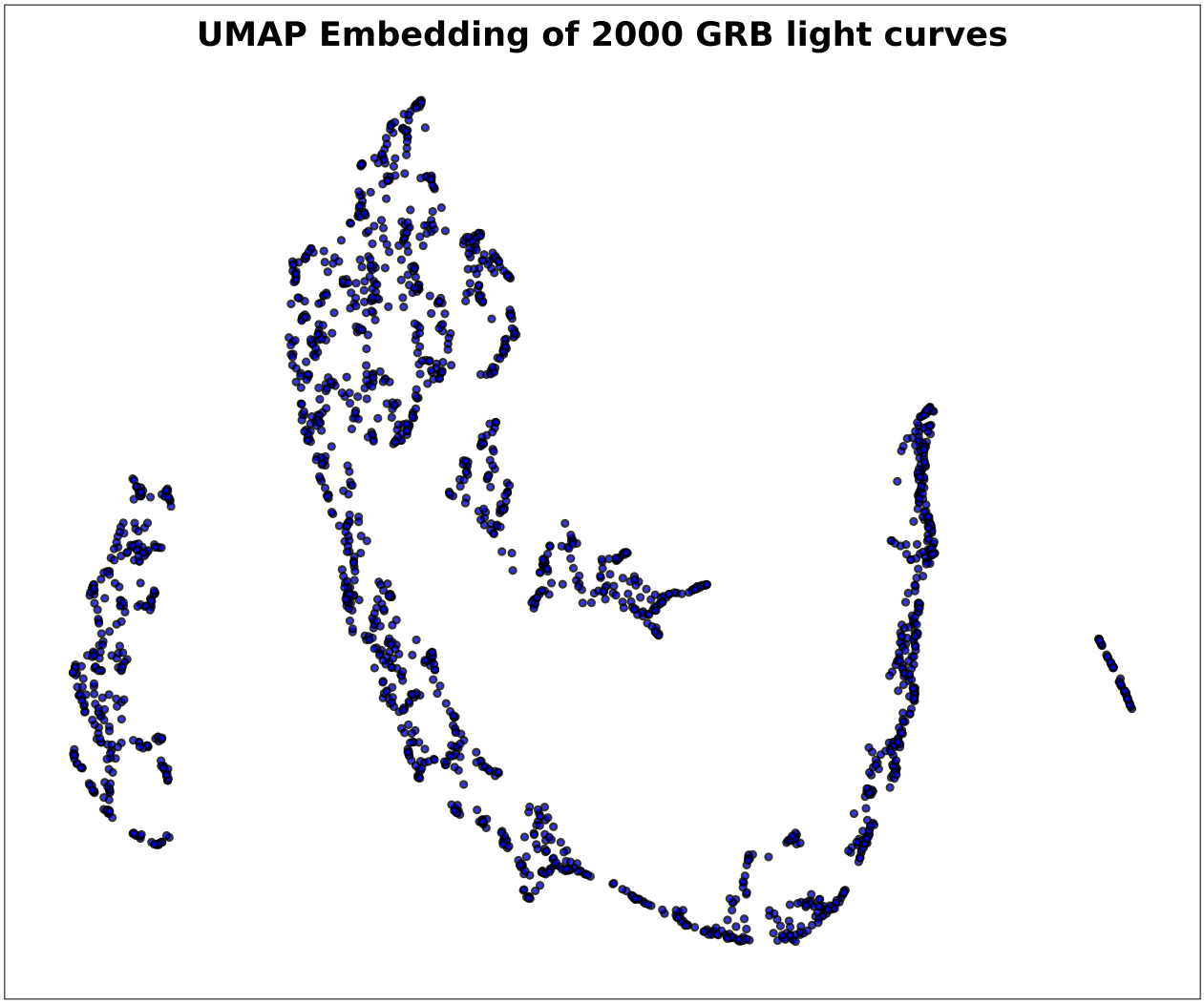


Figure 4.5: UMAP embedding with PCA for 50-300 Kev light curves. As opposed to the previous analyses without PCA, several clusters are separated out here and the distinction between the different clusters are also prominent compared to Figure4.1

To remove the redundancies in the data to minimize this effect of noise in clustering, we can employ PCA ([14],[10]) to select those components in the data with most of the variance and feed those components to UMAP.

It can be seen that the results have drastically changed (Figure 4.5), there are six distinct clusters (Figure 4.6) identified by HDBSCAN as opposed to three earlier. If we look at the duration map of the embedding (Figure 4.7) it is evident that long GRBs as well as short GRBs form two distinct clusters with ultra-long GRBs clustered out distinctly and a certain number of short GRBs occupying as a separate group. Again our null hypothesis here would be that if duration was the sole parameter driving the process of clustering then there should have been only distinction based on duration however it doesn't seem like that here. Now if try to decipher the information from the power-index map (Figure 4.8), it is revealed that the cluster IDs zero, four, and five in the HDBSCAN embedding have a different power index than the rest of the larger group, this could have been the reason that it forms separate clusters.

But what about the ultra-long group with the same power index as the larger cluster and still separated? The reason could be the duration of those GRBs occupying this cluster, the "ultra-long"

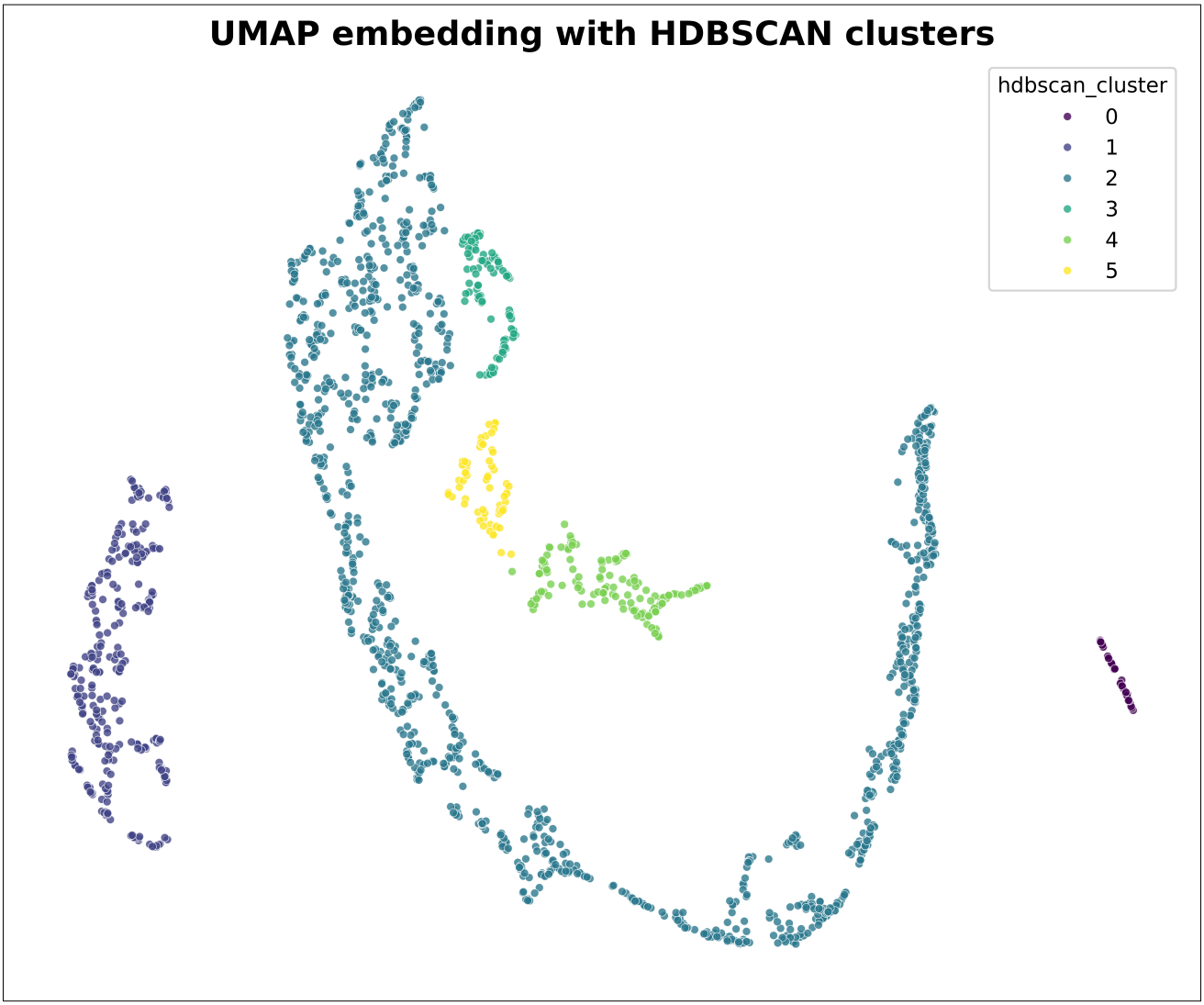


Figure 4.6: HDBSCAN cluster assignment for 16 ms light curves in 50-300 Kev band reveals six distinct clusters. However the cluster blue in color (cluster id-3) could be a sub-cluster of the larger one (cluster id-2). Cluster 4 and cluster 5 are also identified as distinct clusters.

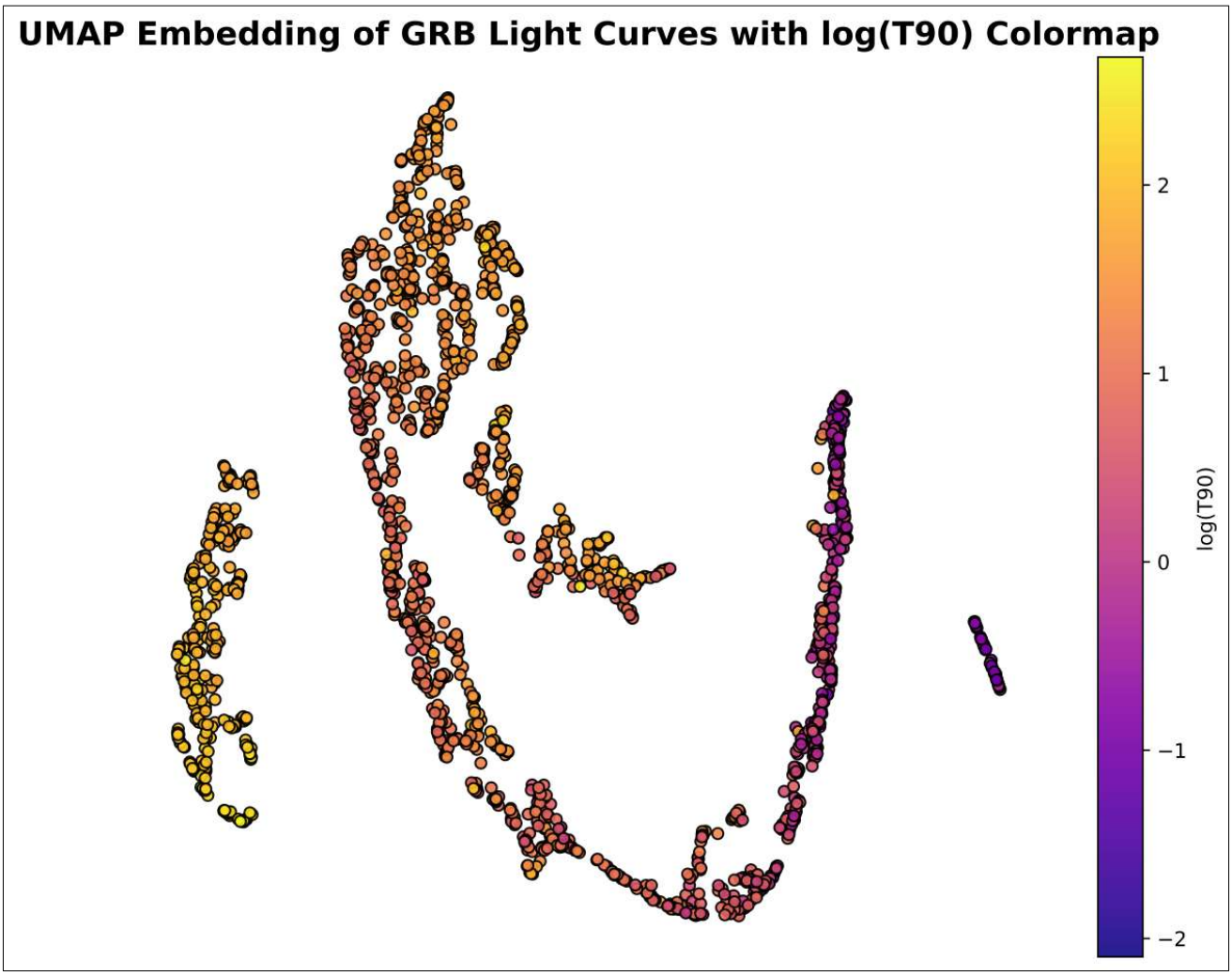


Figure 4.7: UMAP embedding color-coded by duration for 16 ms light curves in 50–300 keV band. A clear gradient in the duration can be seen, with the left side of the figure constituting long GRBs whereas the right side is dominated by the short GRBs. Even within the long GRBs, there are several sub-groups and within Short GRBs, there are two groups. Remarkably, the cluster at the left-most side is occupied by ultra-long GRBs with a log(T90) value of two or above two. It should also be noted that cluster 2 in Figure 4.6 is a combination of long and short GRBs because HDBSCAN identifies the clusters using a notion of distance defined by the user and not based on the physics of the data.

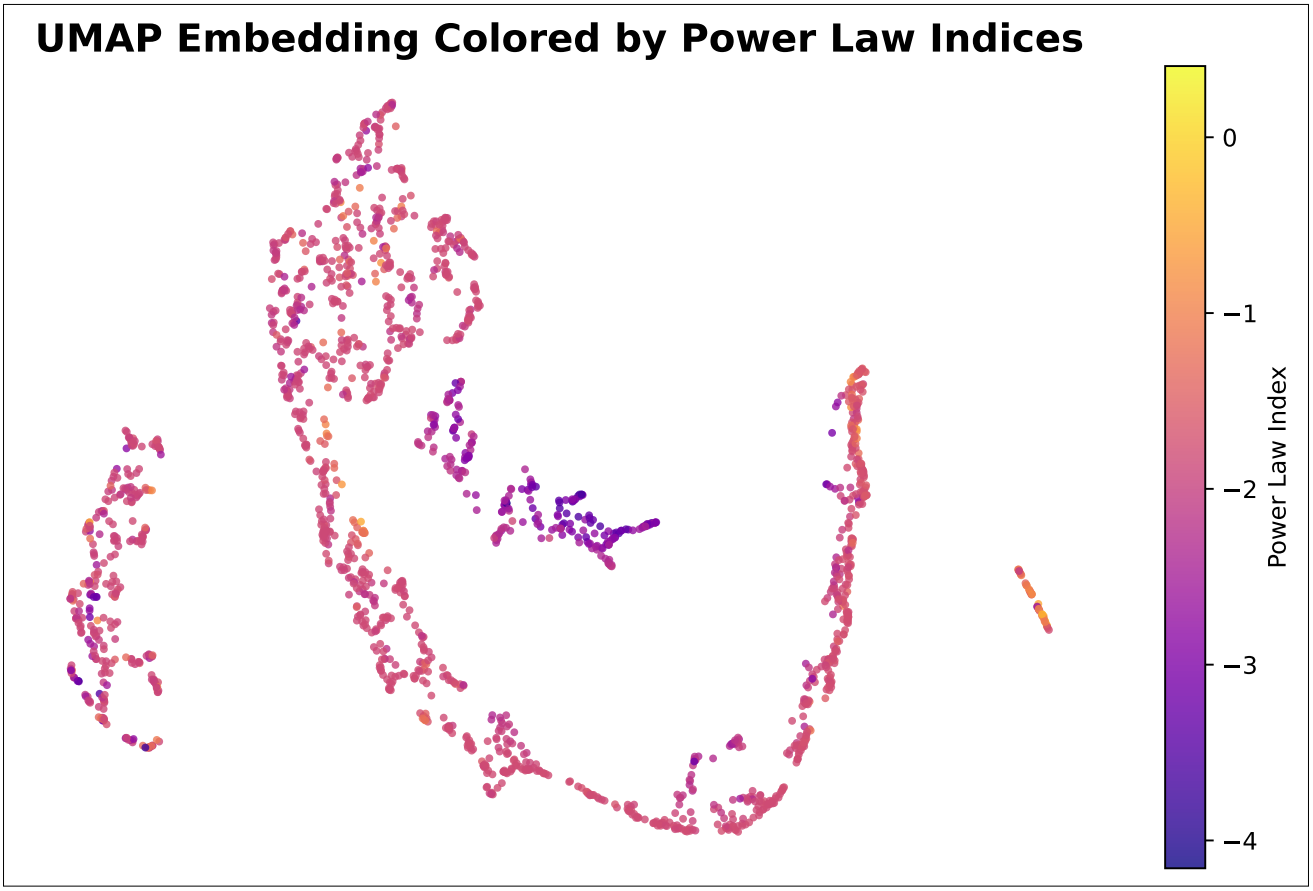


Figure 4.8: UMAP embedding color-coded by power-law index for 16 ms light curves in 50–300 keV band. Similar to Figure 4.4 most of the light curves are dominated by red noise with a power index of two as is the case of cluster two in Figure 4.6. However, clusters 0,1,4 and 5 are telling a different story. The short GRB cluster 0 is distinct from its parent group cluster 2 because it has a white noise (power index of 0) dominance in their light curves. In the case of clusters 4 and 5 (yellow and green), it has consistently a steeper index compared to the rest of the group and this could be the reason that it is separated from the larger group. Fascinatingly cluster 1 with the same noise composition as cluster 2 is still separated and the reason could be their duration. A competition between duration and noise as a parameter for driving the clustering process is evident here.

ones. It seems like there is a competition between duration and noise (manifested in form of power index) as a parameter to drive the process of clustering. In the case of short GRBs noise characteristics are winning this competition whereas for ultra-long GRBs duration take over as the leading parameter.

This hypothesis makes sense because the short GRBs are background-dominated and there is almost zero variability in the denoised signal, only a small Dirac delta peak in the light curve, and everything else is dominated by noise imparted by the background. However in the case of long GRBs in particular for the ultra-long ones, there is enough variability in the signal so that the manifold can be structured based on that information. Some of the light curves for short and long GRBs are given in the Appendix. The next step is to see what all things would change if we included more information in our analysis by concatenating light curves of three different energy bands (8-50 Kev, 50-300 Kev, and 300-900 Kev) and form a cohesive time-series and carry out the same analysis with and without employing PCA.

4.1.3 Analysis for three energy band Without PCA Pre-processing

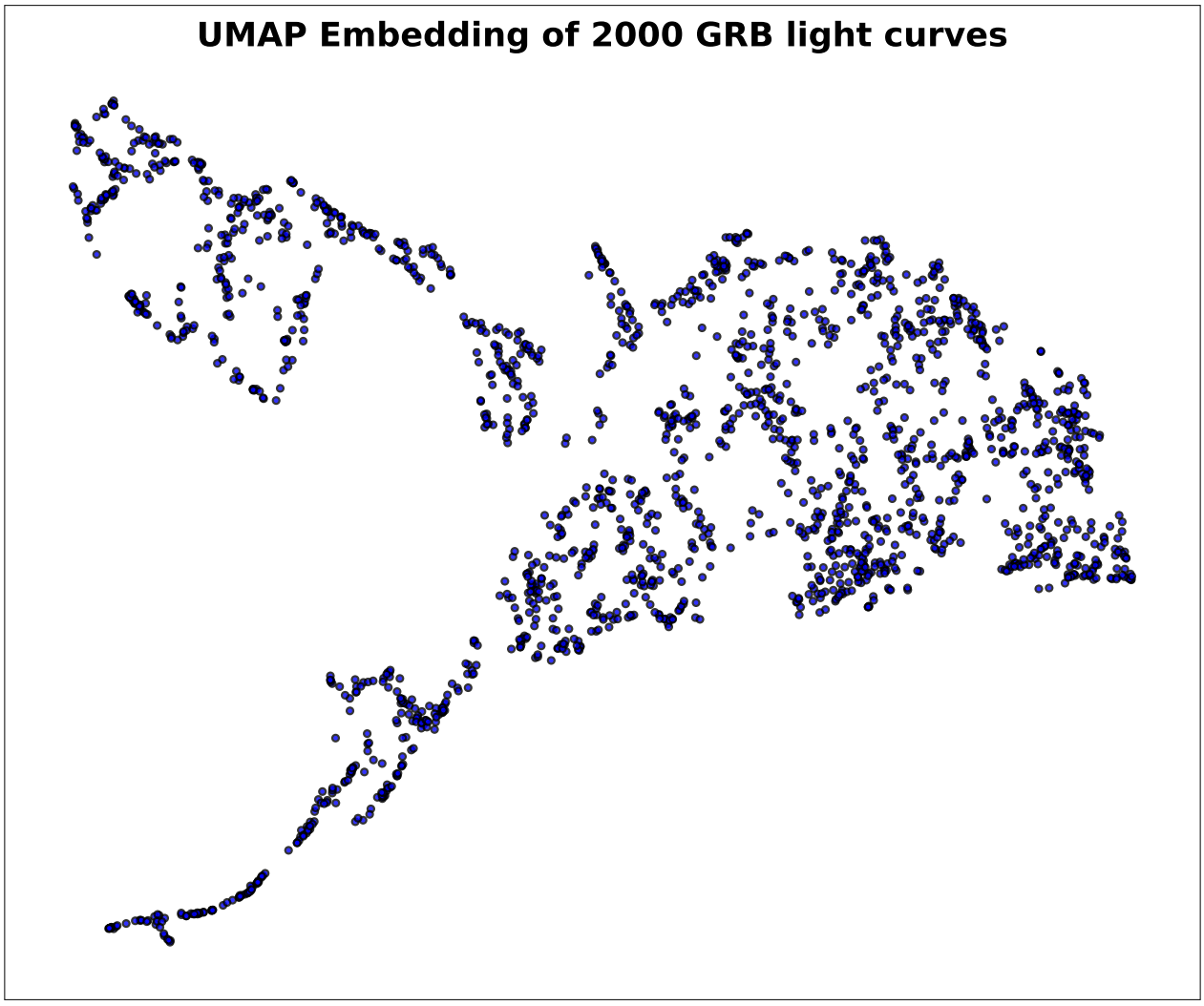


Figure 4.9: UMAP embedding without PCA for 16 ms light curves in 3 energy band.

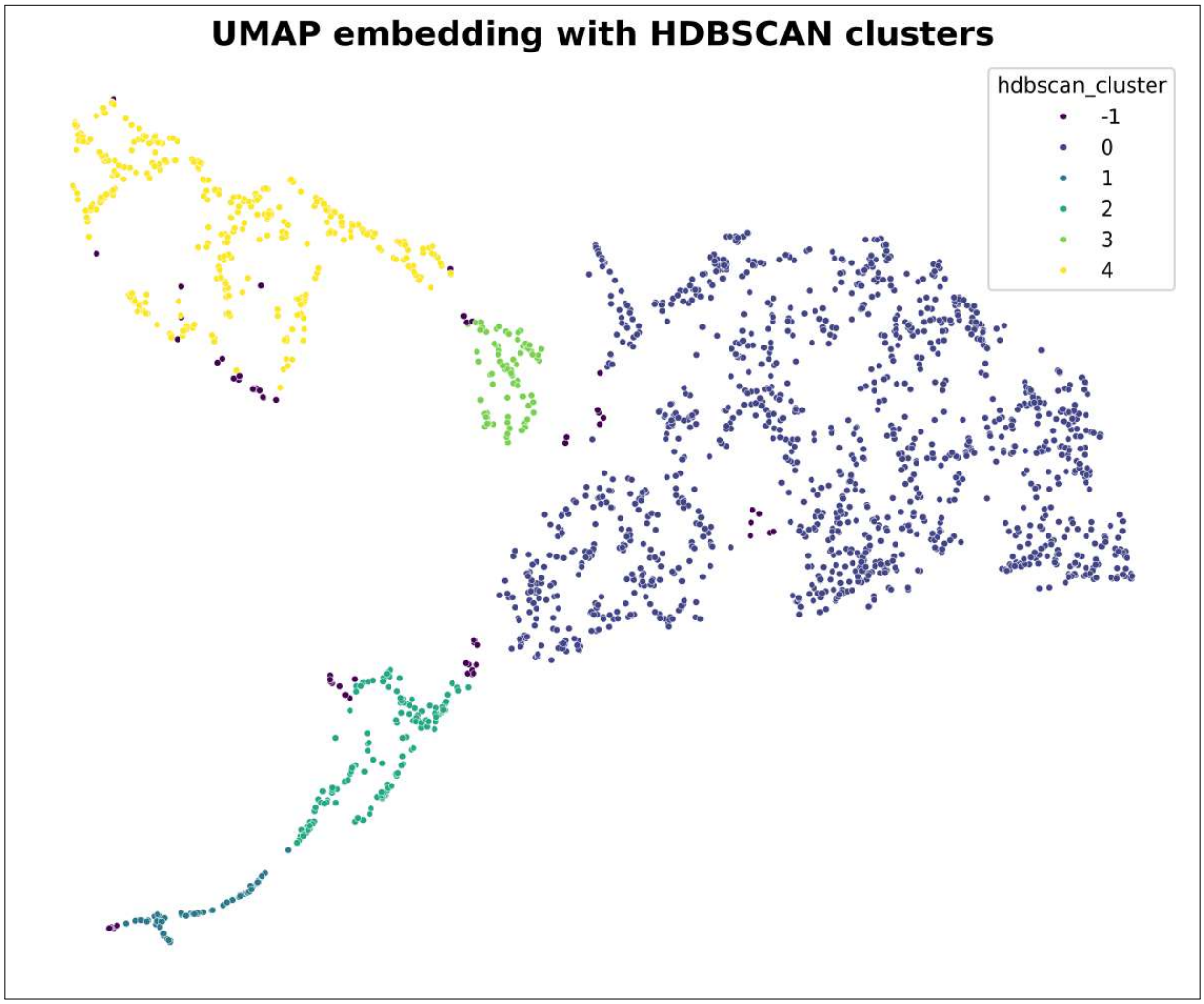


Figure 4.10: HDBSCAN cluster assignment for 16 ms light curves in 3 energy band identifies five distinct clusters. I t can also be seen that there are several sub-clusters within cluster 0 and cluster 4.

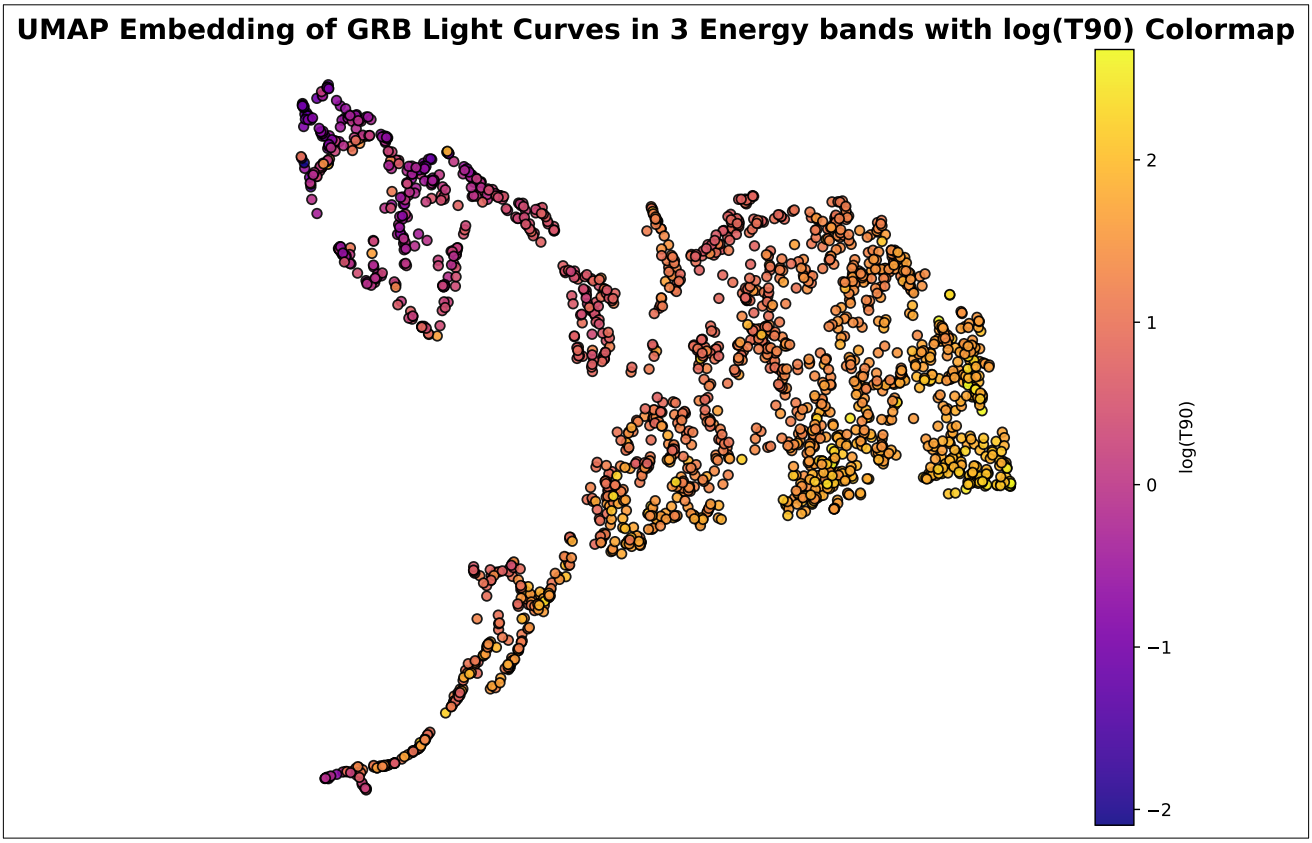


Figure 4.11: UMAP embedding color-coded by duration for 16 ms light curves in 3 energy band. A gradient in duration can be seen from the top left region to right most of the embedding with short GRBs (violet in color) occupying the top left region and the ultra-long GRBs (yellow) occupying the right most region in the embedding. However, a closer examination at the tail of the embedding reveals a small number (<10) of short GRBs positioned there. It could be because of the effect of noise, however, overlaying a map of power index for simultaneous analysis of three different light curves is not meaningful, therefore the reason for this particular positioning of Short GRBs is unknown.

Among the five clusters identified by HDBSCAN (Figure 4.10), a clear distinction can be seen in duration in the left region, predominantly occupied by short GRBs, and the right region, mostly occupied by long GRBs. Additionally, there are several sub-clusters within the long cluster. However, a close examination at the tail of the long cluster reveals the presence of a very small number of short GRBs.

4.1.4 Analysis for three energy band With PCA Pre-processing

A similar analyses is carried out for 16 ms light curves in three energy bands with UMAP initialized by PCA. First step is to generate the scree plot to select the number of eigen vectors/principal components that explains the variance of most of the data. The idea is that PCA removes the low variance noise and retains the high variance features.

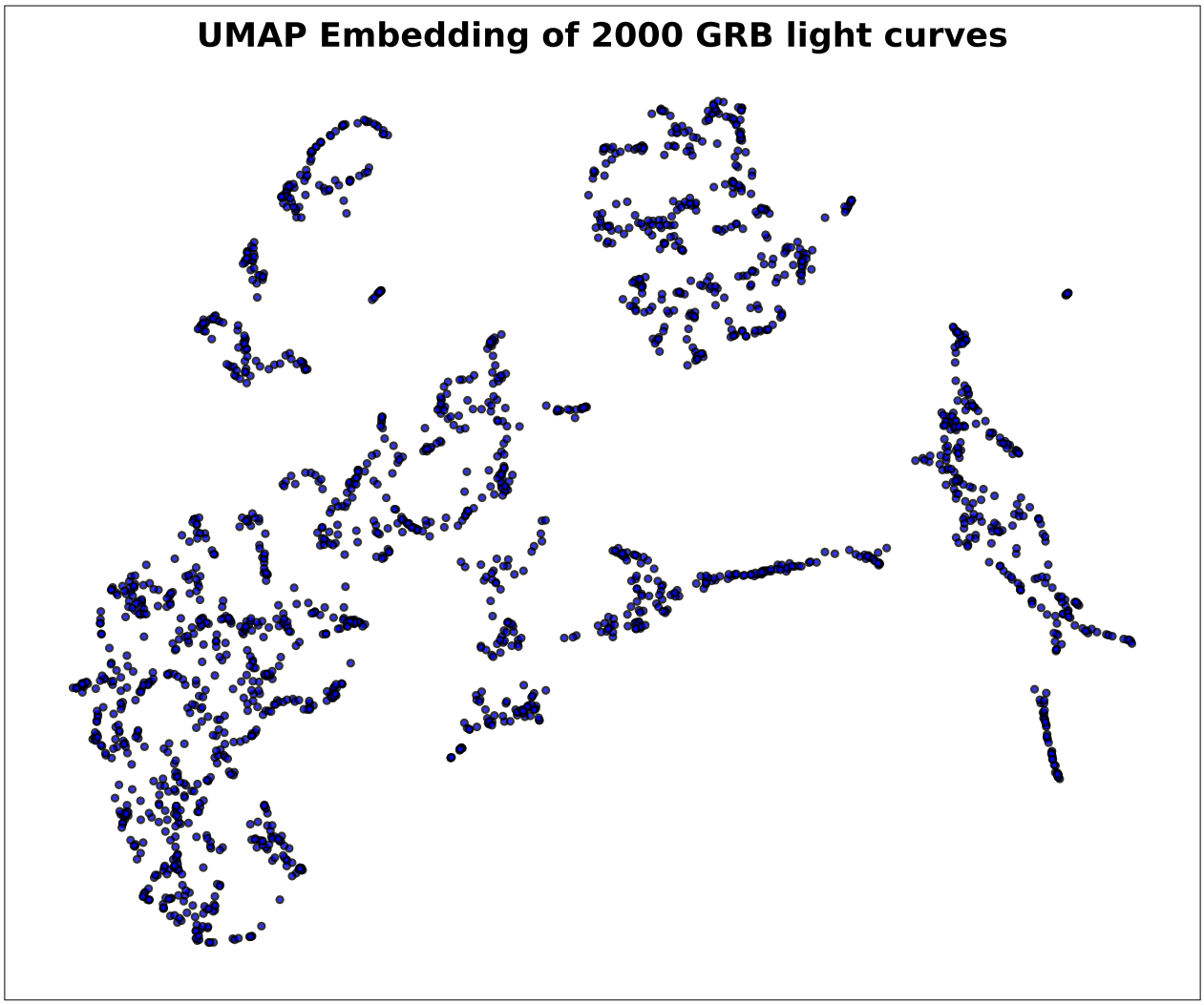


Figure 4.12: UMAP embedding with PCA for 16 ms light curves in 3 energy bands. Here the clusters are far more separated compared to the previous results. The number of outliers (both population outliers and inter-cluster outliers) is larger in this embedding.

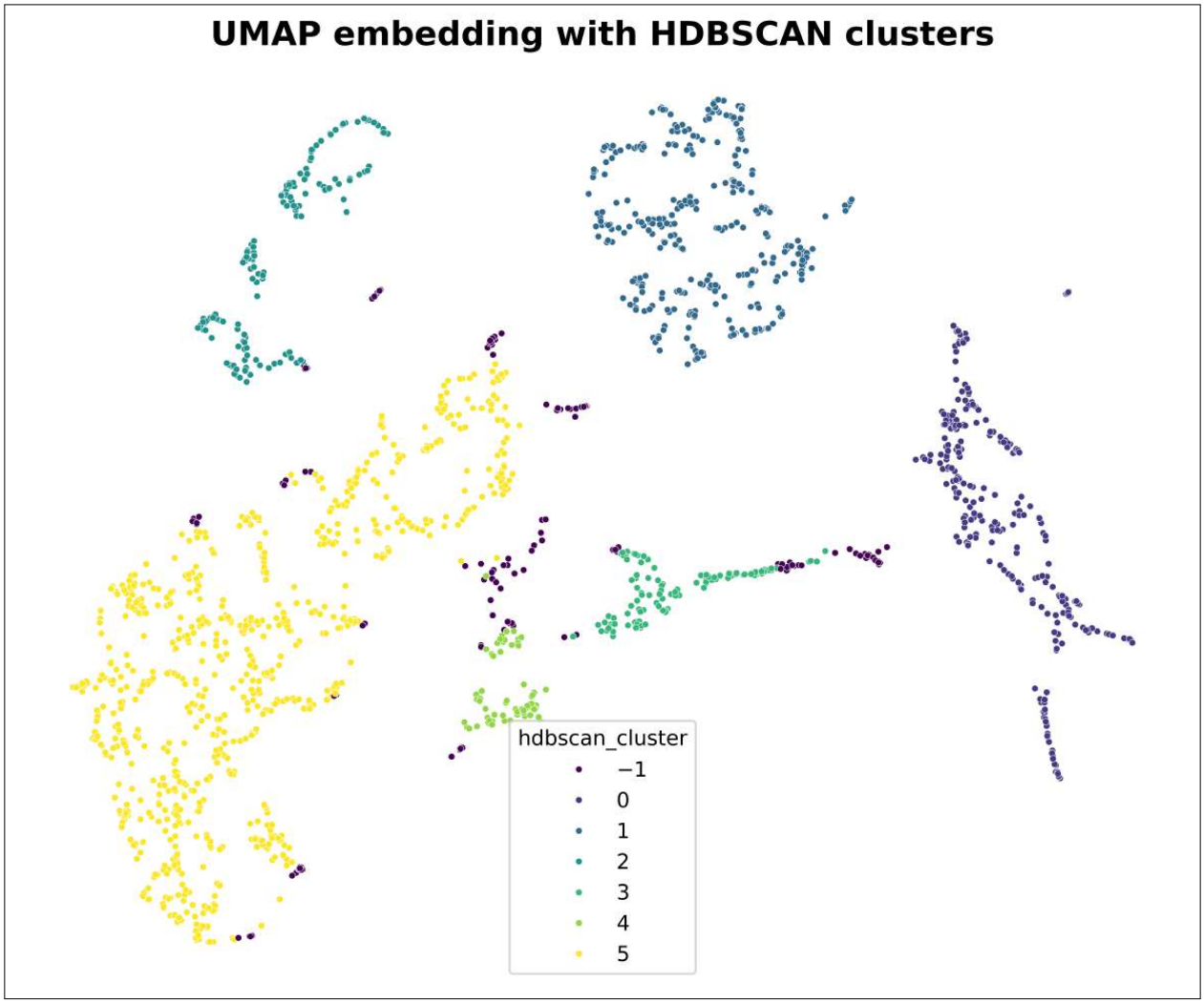


Figure 4.13: HDBSCAN cluster assignment for 16 ms light curves in 3 energy band. There are six clusters identified by HDBSCAN with outliers labeled as noise.

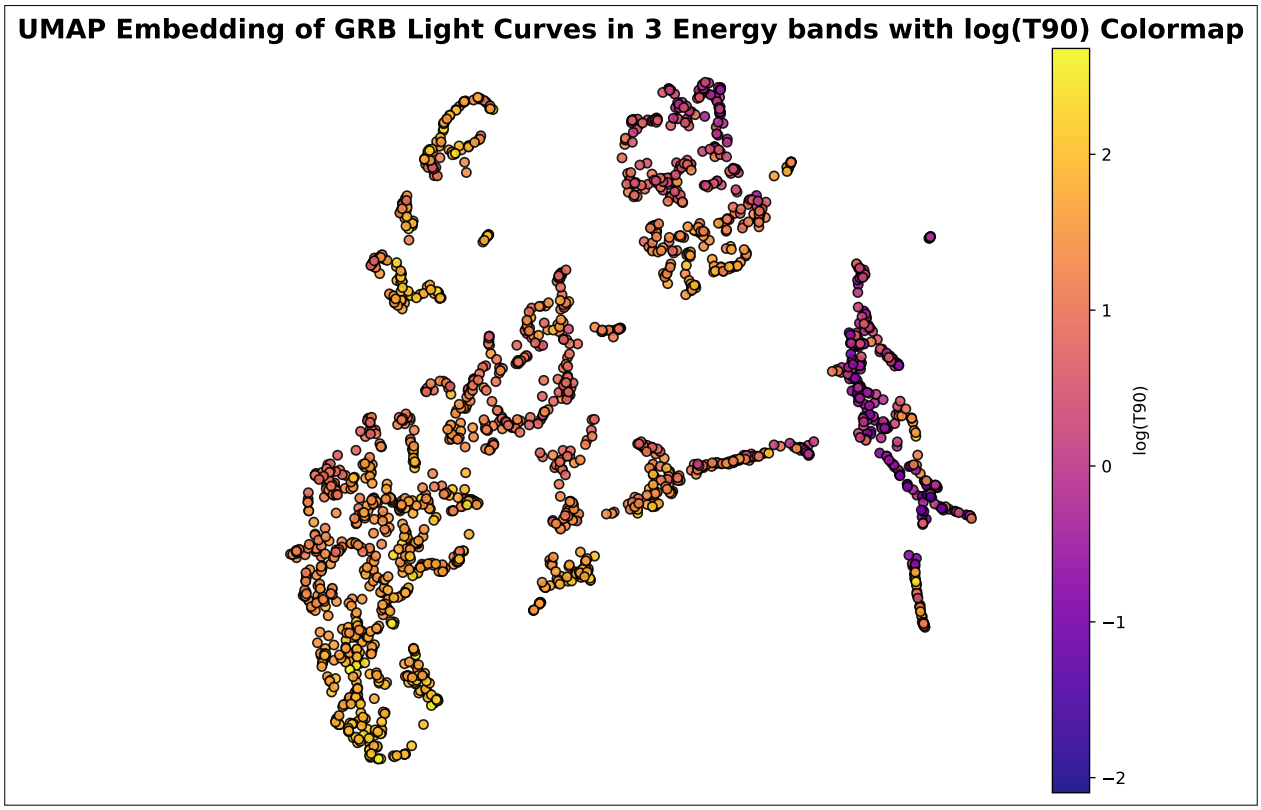


Figure 4.14: UMAP embedding color-coded by duration for 16 ms light curves in 3 energy bands. Out of the six clusters identified by HDBSCAN, two are occupied by short GRBs. However, the presence of long GRBs along with the short GRBs in the cluster is also intriguing. Since there is a mix of long and short GRBs in most of the clusters and their distinction is comparatively not clear, it could be because noise is playing a dominant factor in driving the process of manifold learning.

With the data reduction using PCA before employing UMAP, several sub-clusters are revealed (Figure 4.13) within the short and long GRB groups. One interesting observation is the presence of population outliers and intra-cluster outliers. Even within the two short GRB clusters, a small number of long GRBs occupy a small region, as shown in Figure 4.14.

4.2 16 ms - A comparative analysis

For better clarity of the results obtained, we compare the analysis with and without PCA for both single and three energy band cases.

4.2.1 Analysis with 50–300 keV Light Curves

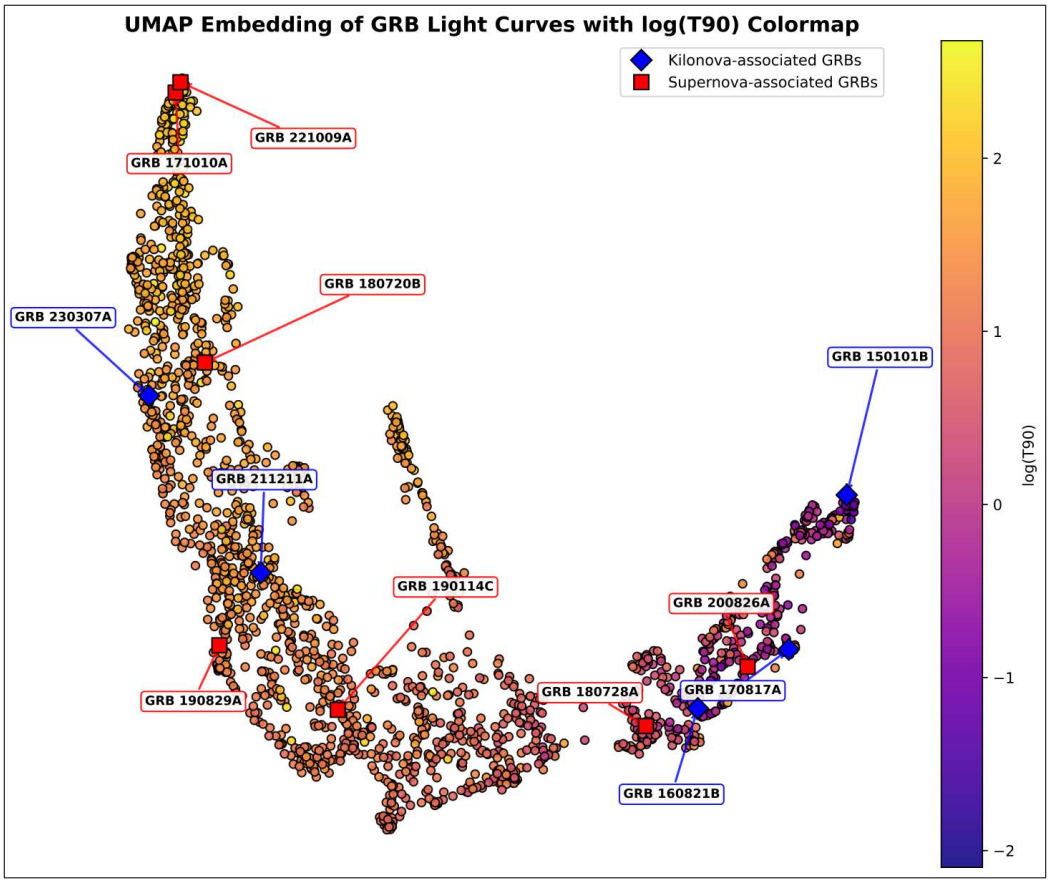


Figure 4.15: UMAP embedding: 50-300 Kev energy band (without PCA) for 16 ms light curves. Supernovae-associated and Kilonovae-associated GRBs are annotated, as red squares and blue diamonds respectively. A comparison can done by tracking the positioning of these annotated triggers in different embeddings. Even if the structure of the embedding is different but the nearest neighbor information is preserved then we can conclude that the results from two different analyses are comparable.

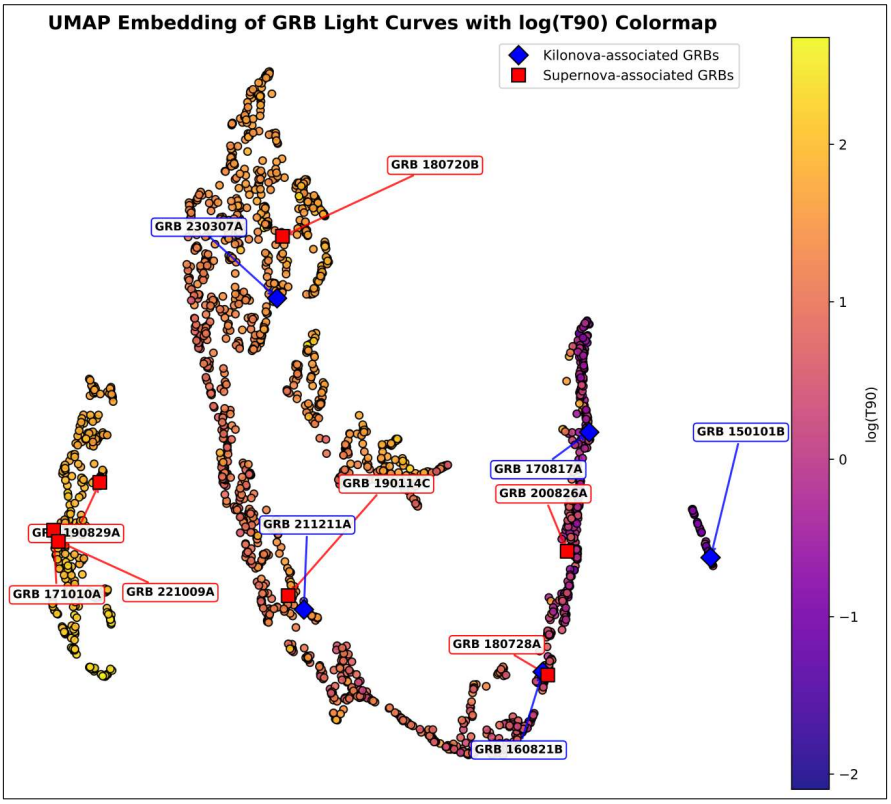
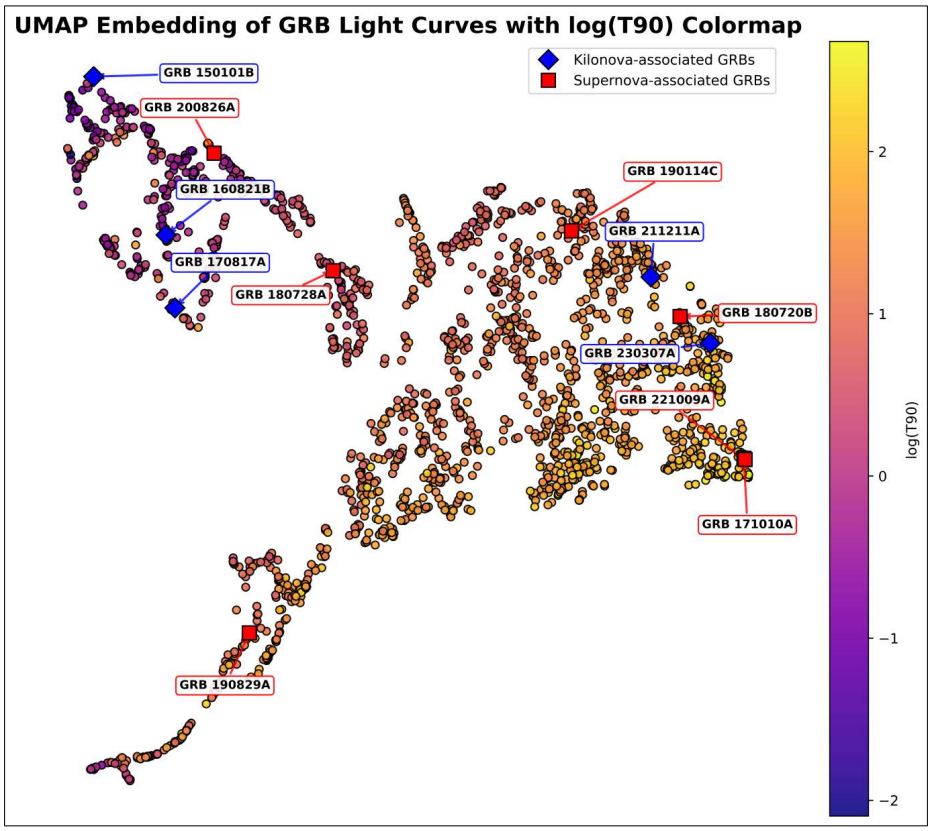


Figure 4.16: UMAP embedding: 50-300 Kev energy band (with PCA) for 16 ms light curves. Supernovae-associated and Kilonovae-associated GRBs are annotated, as red squares and blue diamonds respectively. This can be compared with Figure 4.15 to see whether the positioning of different annotated triggers are same or not. Examination reveals that there are huge dissimilarities in both analyses, with and without PCA initialization, as anticipated.

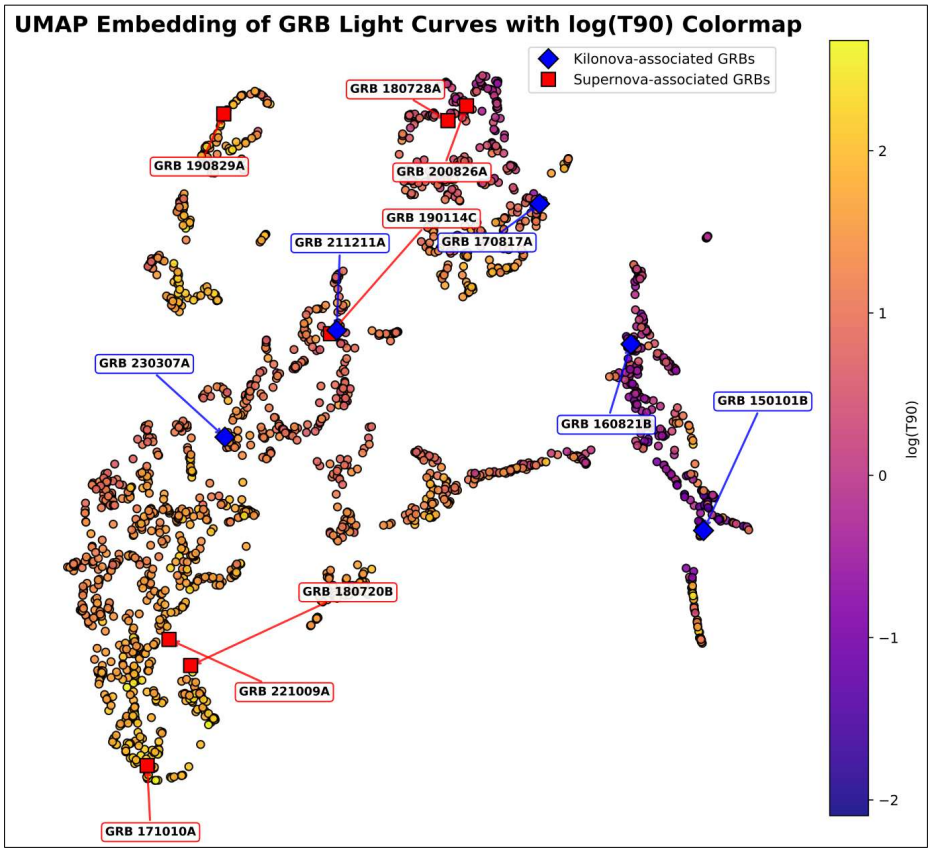
Table 4.1: Comparison of Embedding Characteristics

Similarities	**Differences**
• Separation based on duration. • Separation based on power index is evident in both. • Kilonovae and supernovae associated GRBs do not form distinct groups. • GRBs 171010A and 221009A are immediate neighbors in both.	• Number of clusters are different. • Ultra-long GRBs form a separate cluster when PCA is employed. • GRB 150110B appears isolated in PCA case. • GRB 190829A is not a close neighbor of 171010A and 221009A in the non-PCA case.

4.2.2 Analysis with three energy band Light curves



(a) 16 ms light curves -3 Energy band (without PCA), with annotated triggers.



(b) 16 ms light curves -3 Energy band (with PCA), with annotated triggers.

Figure 4.17: Visual comparison of embeddings under two different preprocessing schemes, three energy band analysis with and without PCA. A number of kilonvae associated and supernovae associated GRBs are annotated.

Table 4.2: Comparison of Embedding Characteristics

Similarities	Differences
• Separation based on duration. • Separation based on power index is evident in both. • Kilonavae and supernovae associated GRBs do not form distinct groups. • GRB 190829A is well separated from other annotated triggers.	• Number of clusters are different. • Short GRBs forming two separate clusters when PCA is employed. • GRBs 150110B and 160821B are distinctively positioned apart on a separate cluster in the analysis with PCA. • Positioning of GRBs 171010A and 221009A are immediate to each other without PCA, whereas with PCA they are positioned with a separation.

4.3 Analysis with 64 ms Light curves

Since a resolution of 16 ms is too good to be true because of the noise dominance, the same analysis can be carried out with light curves of a different binning. We have selected a binning of 64 ms, which is neither too low nor too large. The analysis is carried out for light curves in one energy band and three energy bands with and without PCA initialization.

Contrary to the previous results for the analysis with one energy band without PCA, a certain number of short GRBs are falling separately on the embedding (Figure 4.18 and Figure 4.20) compared to the long GRBs which formed a separate cluster in the same analysis with 16 ms light curves (Figure 4.1). Nevertheless, in both duration maps of the embedding, we can see the gradient going from yellow to violet ie, from long to short duration. This confirms the fact that the process of clustering was indeed based on the duration to a certain extent. If we look at the power index map (Figure 4.21, it is clear that the short GRBs are clustered separately because it has a different power index than the rest of the group. A closer examination reveals the fact that these GRBs are white noise dominated contrary to the long GRBs separated in the 16 ms analysis because of a steeper index (Figure 4.4).

HDBSCAN identifies two clusters in the embedding (Figure 4.19), however there are several subclusters within the large cluster, that are not distinctly separated because of the uniform noise characteristics.

4.3.1 Analysis for 50-300 Kev light curves Without PCA Preprocessing

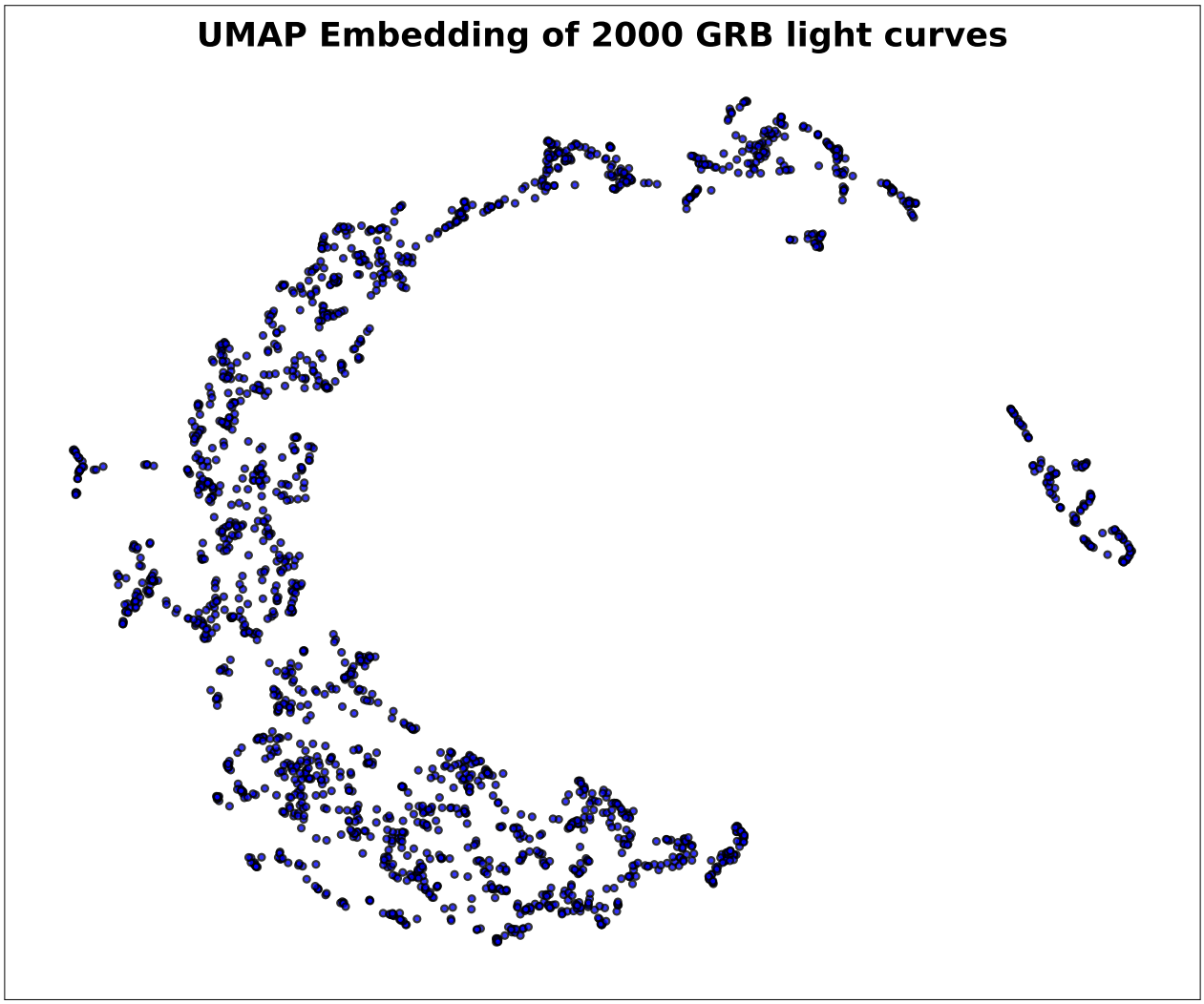


Figure 4.18: UMAP embedding without PCA for 64 ms light curves in 50-300 Kev band. A spiral structure with one large group and a small group is evident. Distinction within the large group is hard to identify. The reason for this kind of separation can be analyzed by overlaying the embedding with duration map and power index map.

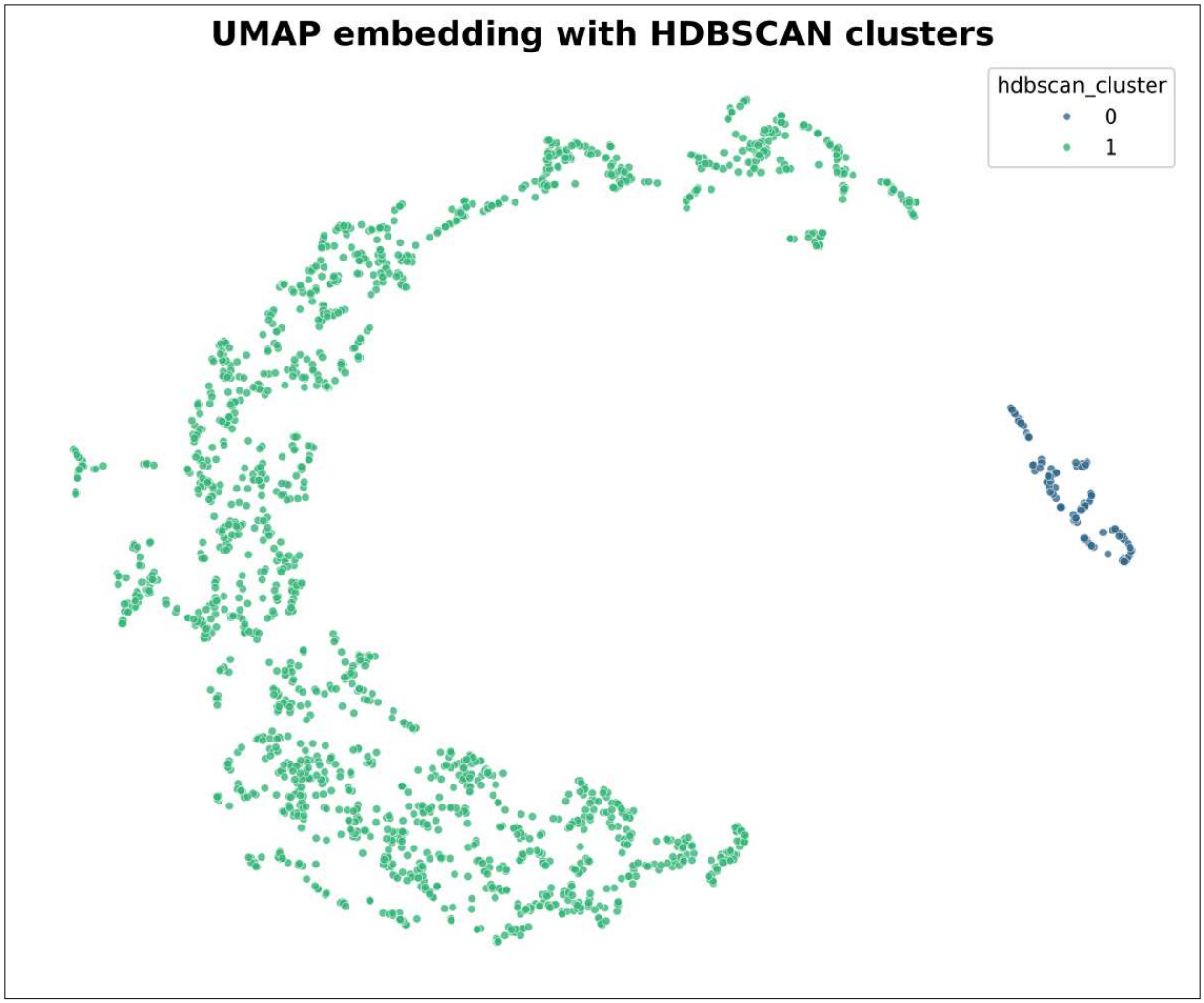


Figure 4.19: HDBSCAN identifies two distinct clusters. Hyperparameter tuning can be done to identify different sub-clusters within the cluster 1.

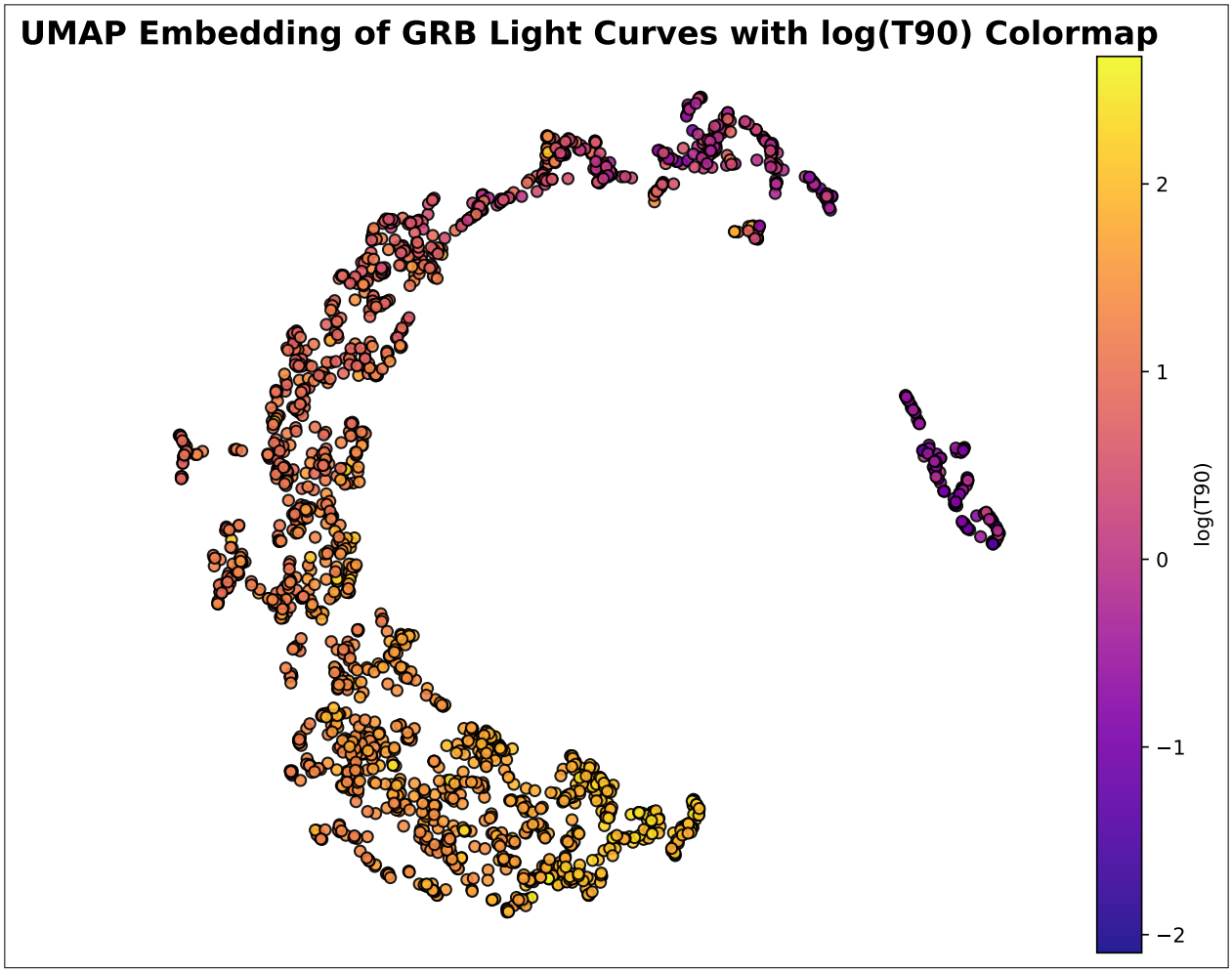


Figure 4.20: UMAP embedding without PCA for 64 ms light curves in 50-300 Kev band, overlayed with duration map. A clear gradient in duration can seen within the spiral structure ranging from yellow to violet. The short cluster is completely occupied by the short GRBs. However, a small number of short GRBs also occupies the large cluster predominantly occupied by the long GRBs. A gradient in duration shows the impact of the manifold structured on the basis of duration. Inorder to decipher the reason for a mix of short and long GRBs, embedding can be over;ayed with the power index map.

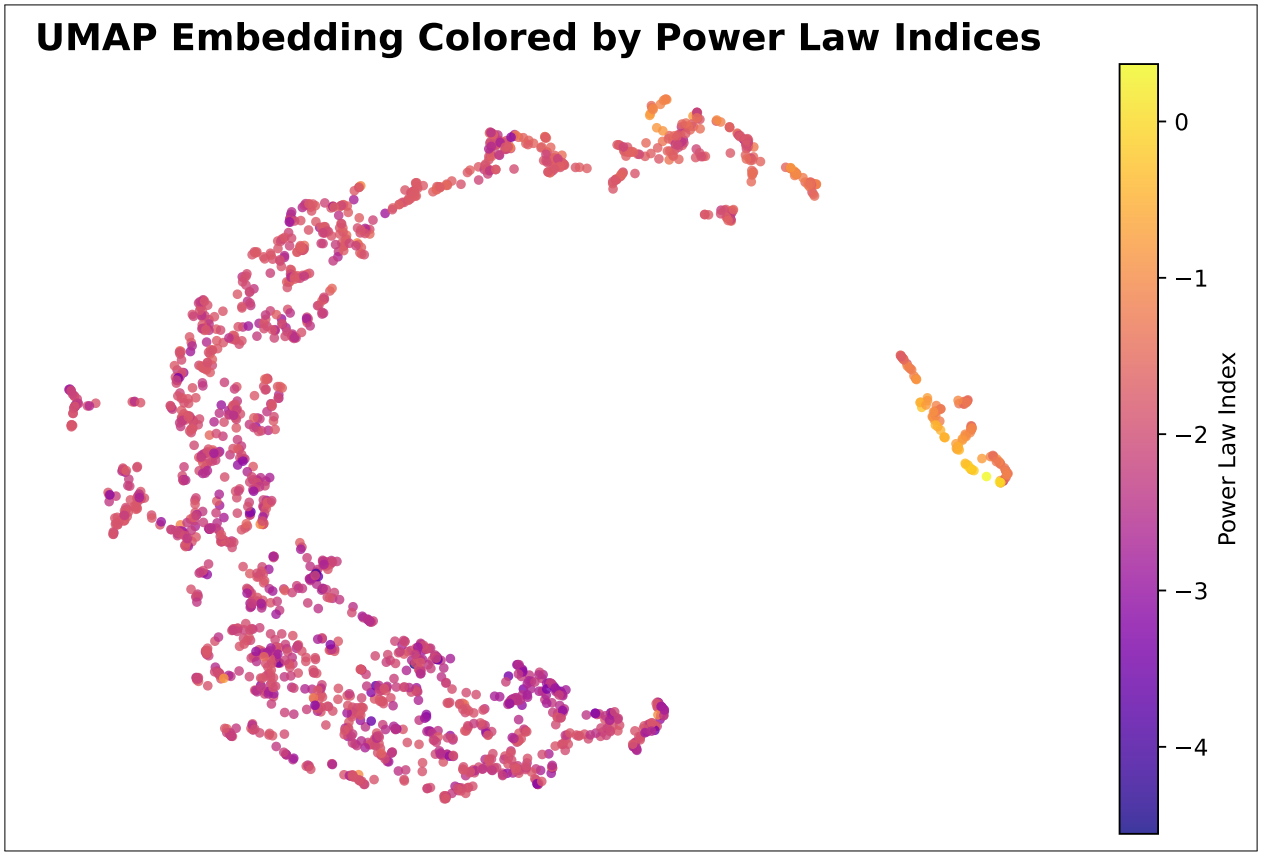


Figure 4.21: UMAP embedding without PCA for 64 ms light curves in 50-300 Kev band, overlayed with power index map. The short cluster which is separated from the large spiral group is predominantly constituted by white noise in their light curves which is evident from the power index of zero of the light curves occupying that group. It is also remarkable that within the large group there is a uniform noise composition with a power index of 2, which could be the possible reason for the non-visibility of sub-clusters within this group.

4.3.2 Analysis for 50-300 Kev With PCA Pre-processing

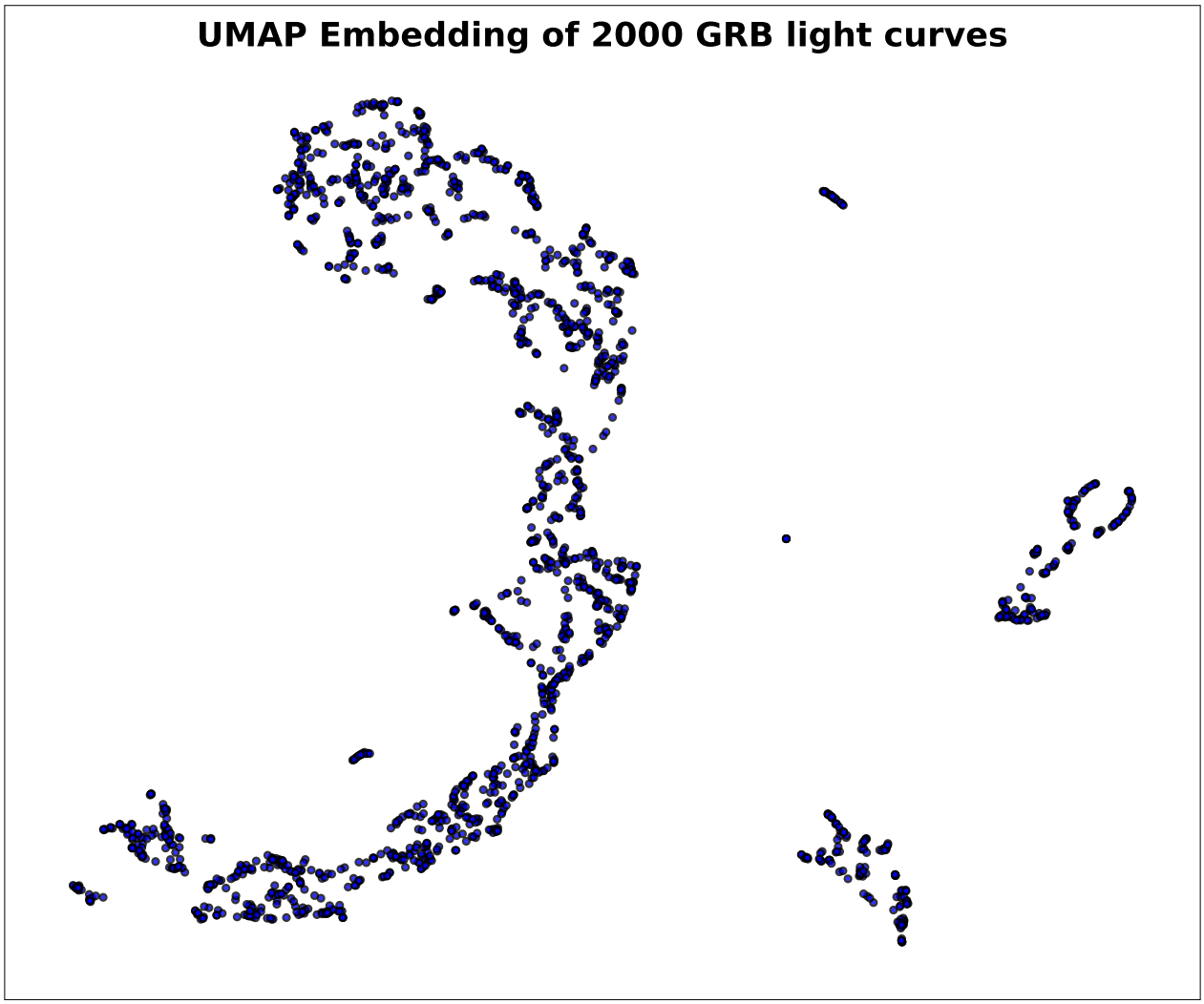


Figure 4.22: UMAP embedding with PCA for 64 ms light curves in 50-300 Kev band. There is one large group and two small groups distinctly separated. Several outliers are also identifiable. Removing the low variance redundancies using PCA made the clusters more distinct from each other.

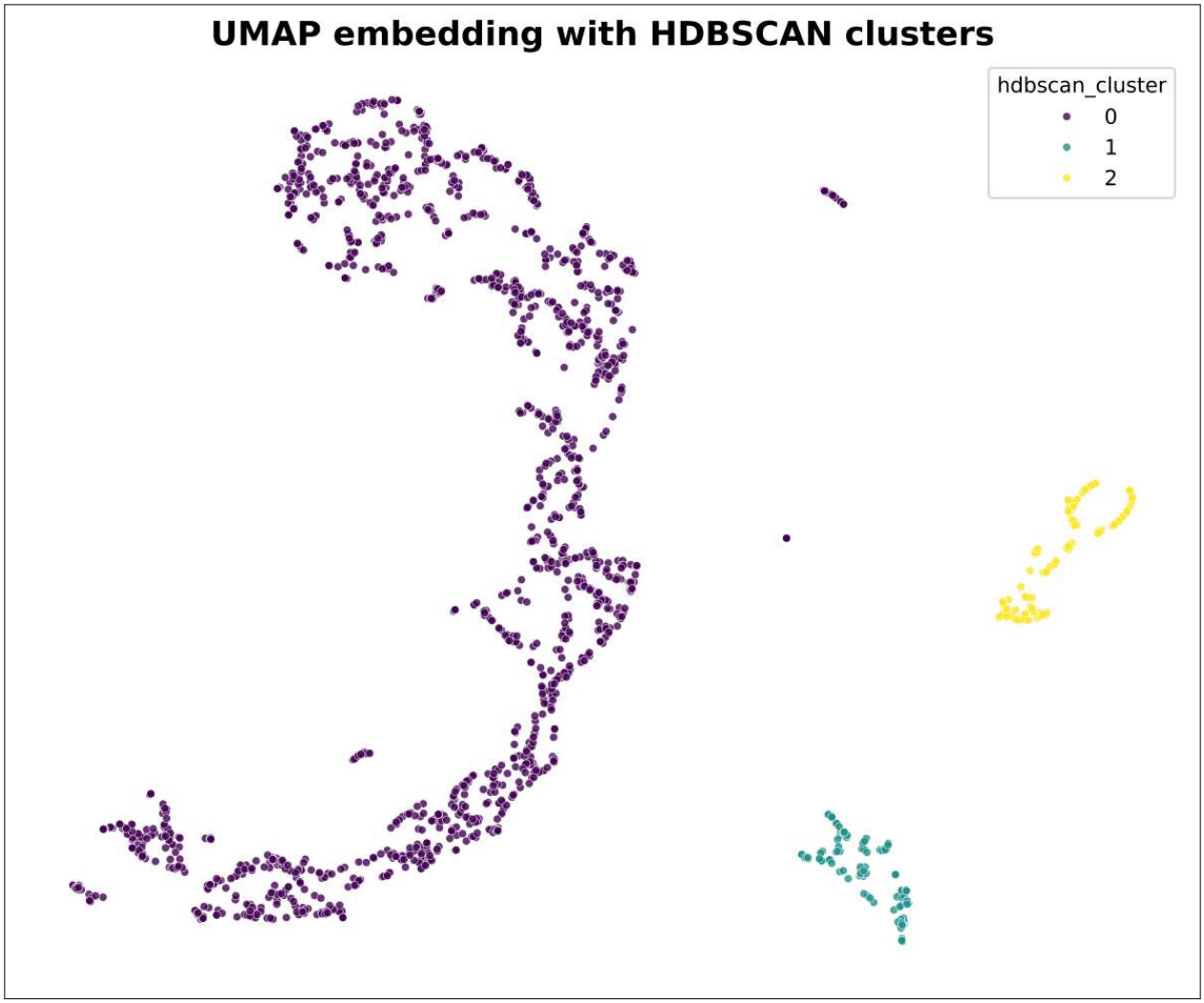


Figure 4.23: HDBSCAN identifies three distinct clusters. However, the outliers which exist as only a handful of data points are not labeled as noise here, in contrast, it is identified as a part of cluster 0.

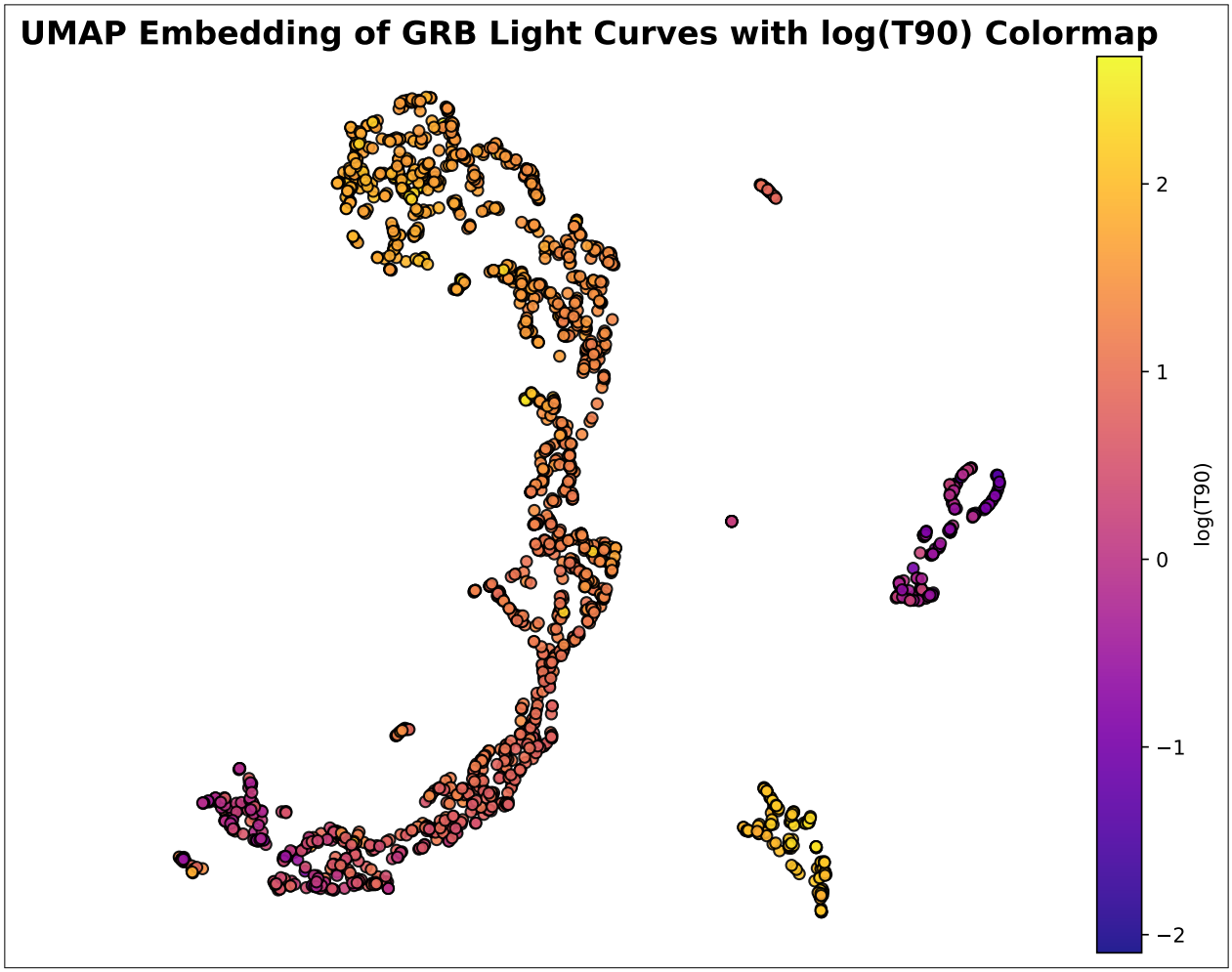


Figure 4.24: UMAP embedding with PCA for 64 ms light curves in 50-300 Kev band, overlayed with duration map. Similar to the previous findings, a clear gradient in the duration is visible. The large group (cluster 0) is a mixture of long and short GRBs whereas cluster 1 is dominated by the ultra-long GRBs and cluster 2 is completely occupied by short GRBs, rendering a similar result to that of 16 ms light curves in 50-300 Kev with PCA initialization Figure 4.7, however, the number of ultra-long GRBs within the group seems to be different. The reason for a mixture of short and long GRBs can be found from the power law map Figure 4.25

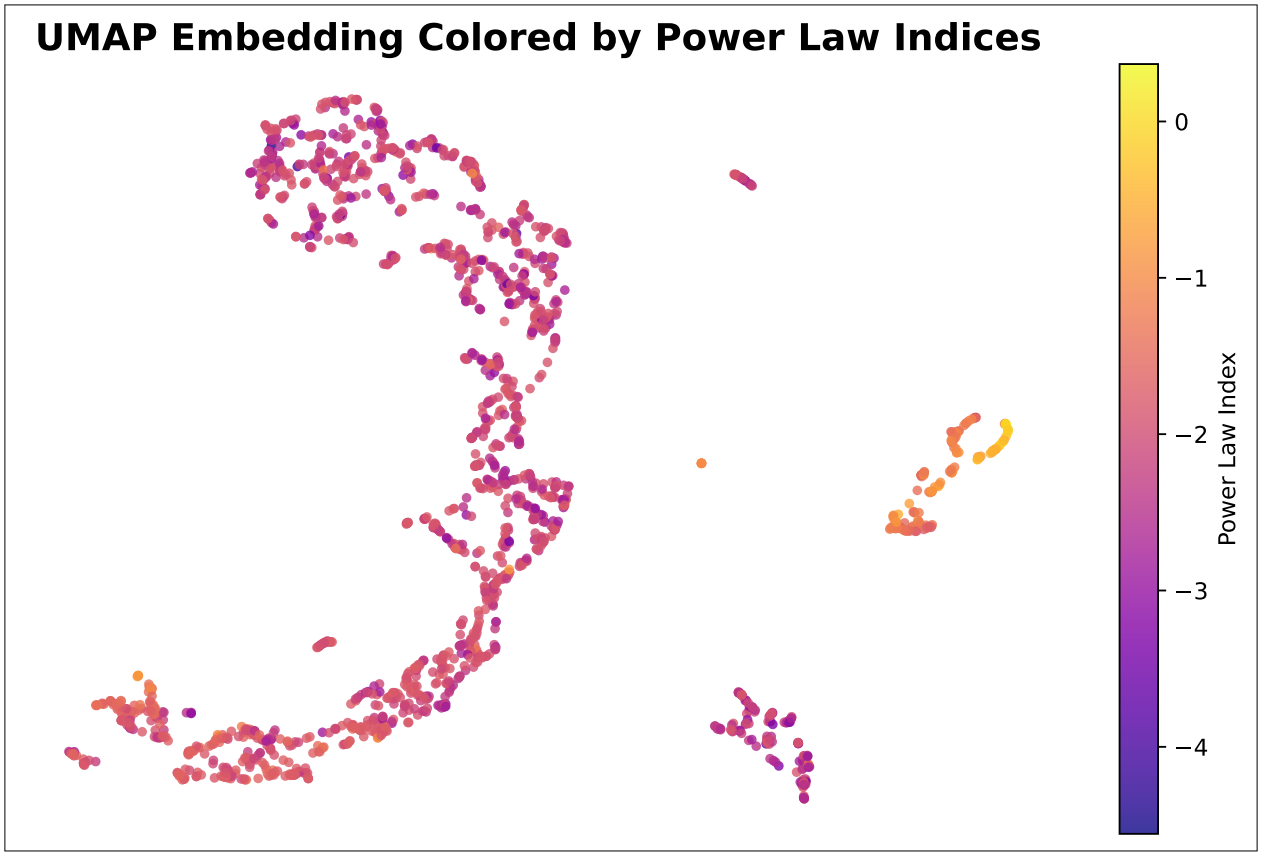


Figure 4.25: UMAP embedding with PCA for 64 ms light curves in 50-300 Kev band, overlayed with power index map. It is evident that the short GRBs are separated because of the white noise dominance in their light curves and the ultra-long group is separated because of their duration, which is clear from their reluctance to join the larger group even though it has the same red noise characteristics. The reason for the mix in the large group is because of the uniformity in their power indices.

We can see the same paradigm persisting here as of the analysis of 16 ms light curves in one energy band with PCA initialization, one set of ultra-long GRBs, and one set of short GRBs clustered out separately based on the duration map overlayed onto the embedding. Again the power index map is telling the same story here, short GRBs that form a separate group have consistently different noise characteristics (white noise dominated) than the rest of the group which is part of the large cluster. Now in the case of the "long-outlier" group, the separation is not based on noise because it is evident that the larger group also has the same power indices, therefore, the grouping here is because of the duration properties of the light curves, the same scenario we encountered in case of 16 ms light curve analysis, the tug-of-war between duration (signal) properties and noise (background) properties.

4.3.3 Analysis for three energy band Without PCA Pre-processing

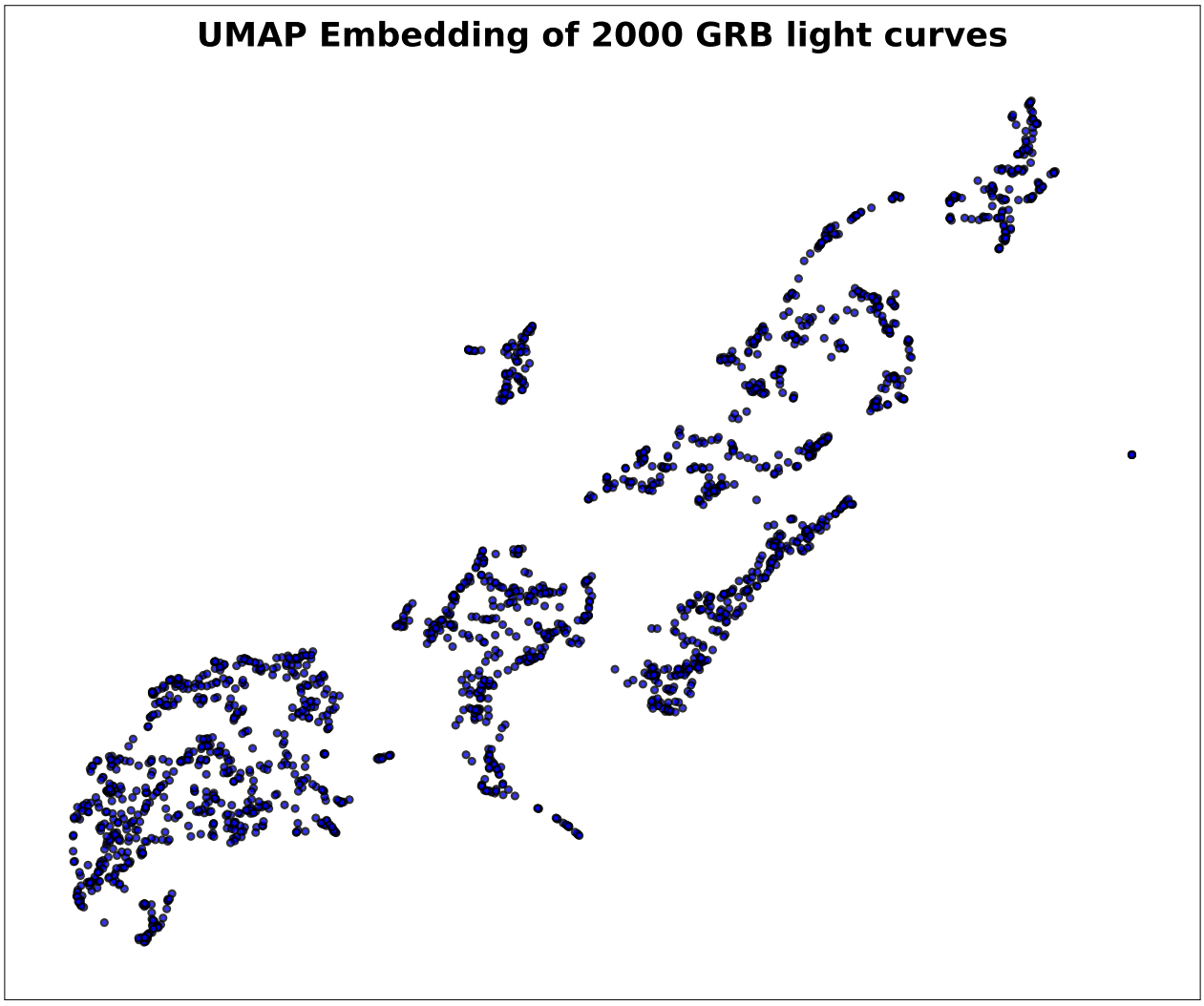


Figure 4.26: UMAP embedding of 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands without PCA initialization. Analysis reveals several distinct clusters. The positioning of outliers are also fascinating like the one in the very mid-right of the embedding.

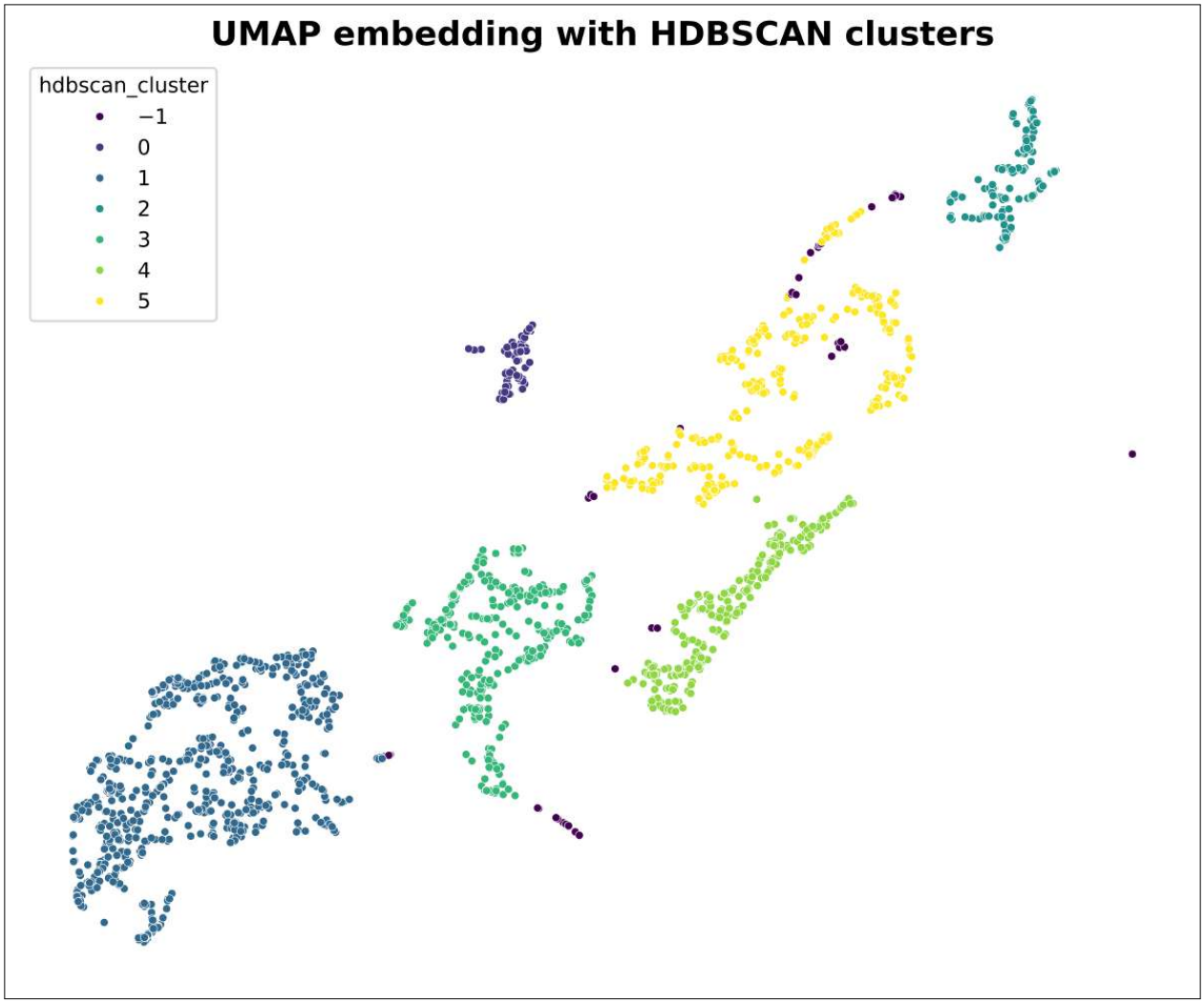


Figure 4.27: HDBSCAN identifies six distinct clusters with outliers labeled as -1.

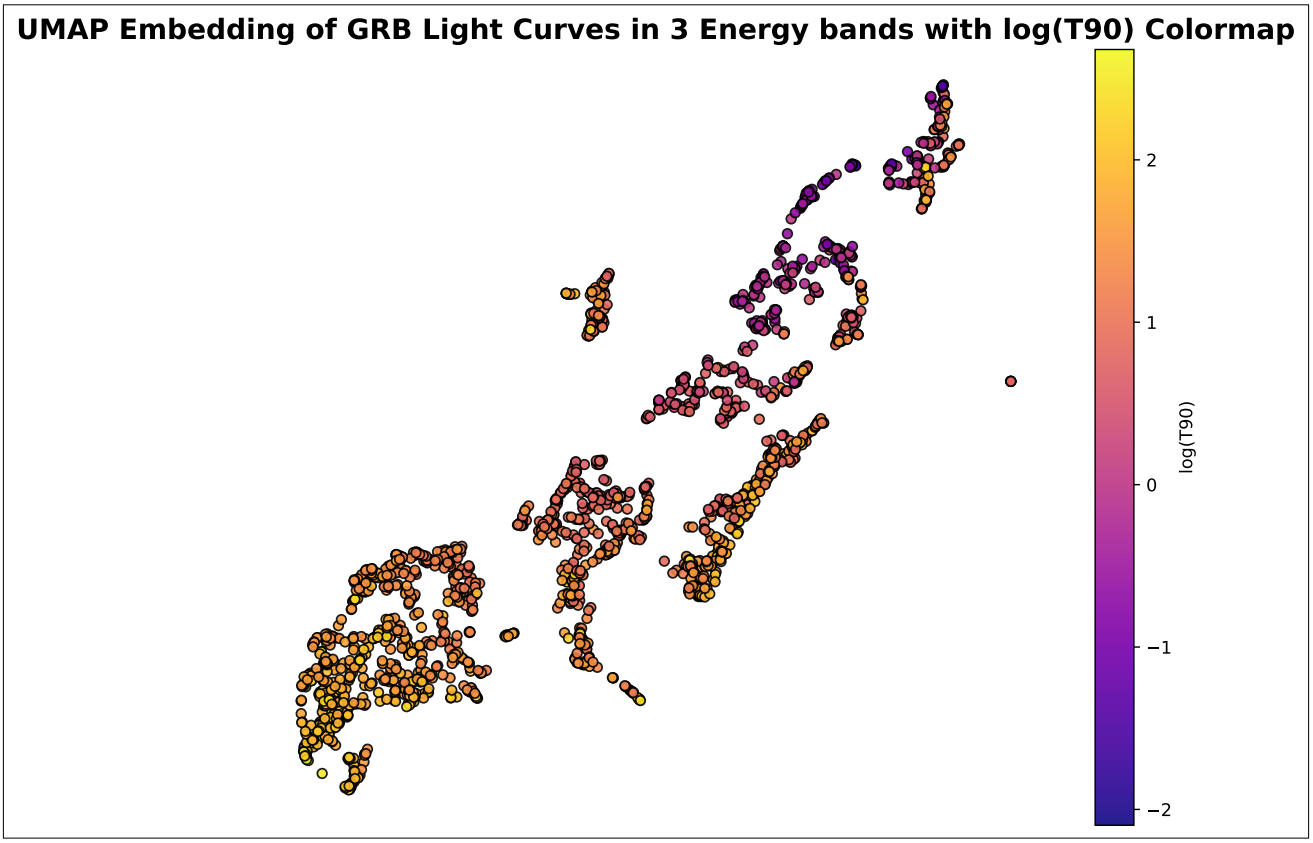


Figure 4.28: UMAP embedding with PCA for 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands, overlayed with duration map. A clear gradient of duration ranging from yellow (long-GRBs) to violet (short GRBs) is visible and this signifies the impact of duration as a parameter for clustering the light curves. However, within the short GRB groups, their positioning right next to some of the long GRBs is also intriguing and this effect could be because of the interplay of noise in structuring the manifold.

Again with three energy bands, the distinction based on duration is clear even though it is spread over different clusters identified by HDBSCAN. However, with three energy bands an analysis of how the power index affects the clustering is pretty complicated as there is no single way to compare the power spectrum of light curves in different energy bands. Nonetheless, an average power spectral comparison is amendable however, it is not done. The reasons for the sub-cluster groups could be a combination of duration, noise, and several other factors that are yet to be unraveled.

4.3.4 Analysis for three energy band With PCA Pre-processing

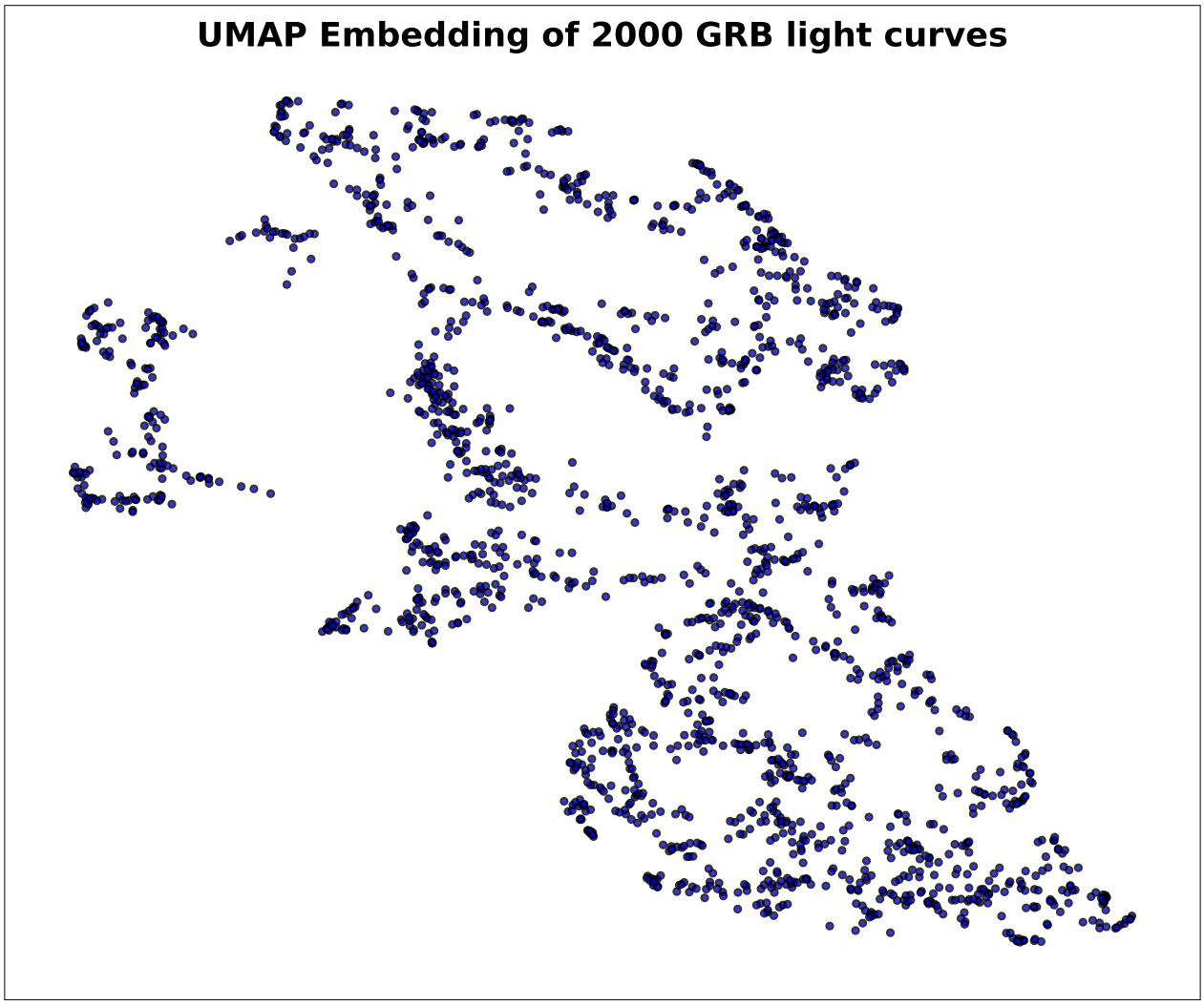


Figure 4.29: UMAP embedding of 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands with PCA initialization. When initialized with PCA, data points got dispersed in the embedding compared to Figure 4.26.Identification of population outliers could be daunting task in this embedding.

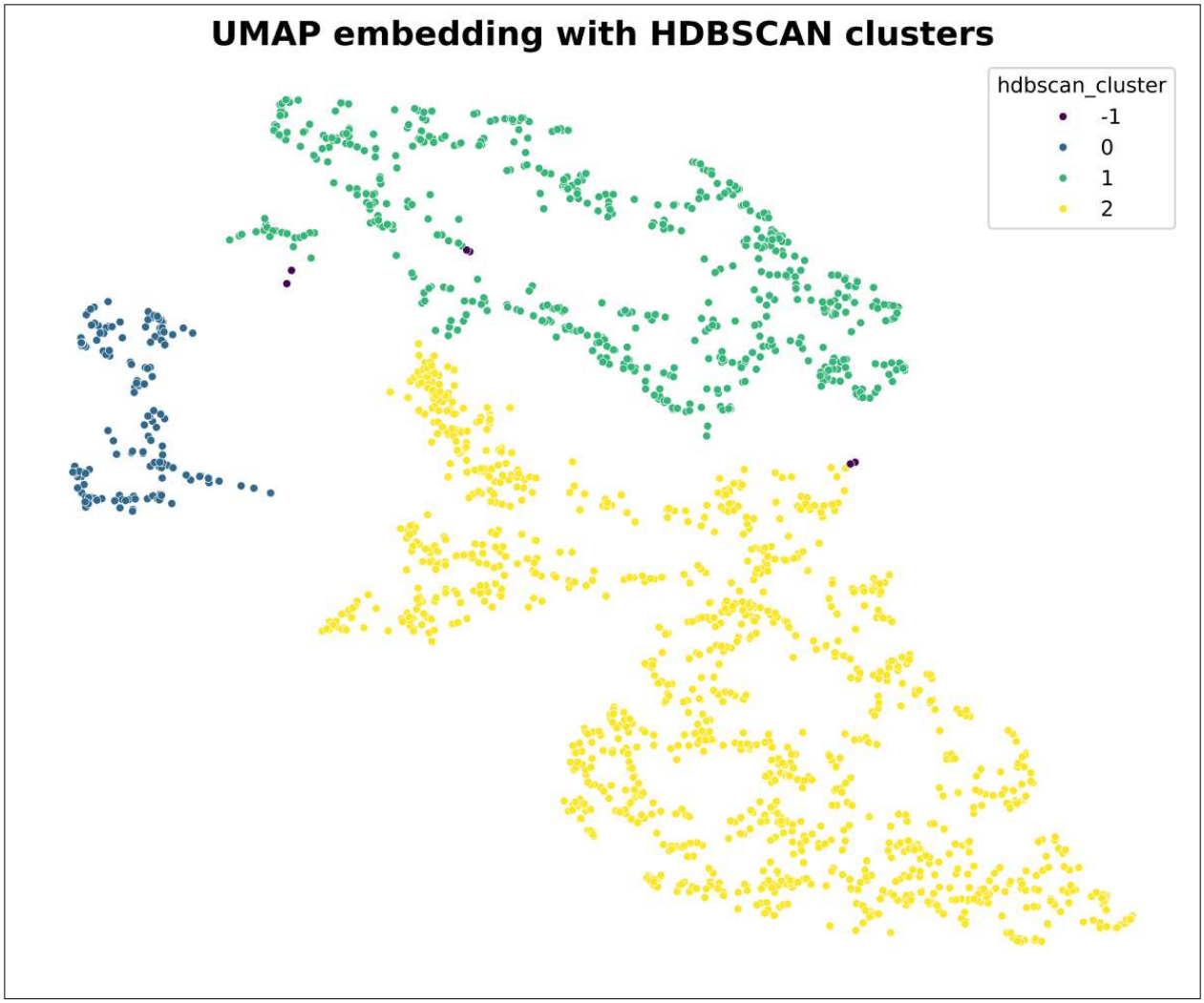


Figure 4.30: HDBSCAN identifies three separate clusters with outliers labeled as -1. With further hyperparameter tuning it will be possible to identify several sub-clusters within the larger ones. More information can be extracted by overlaying the embedding with the duration map and power index map.

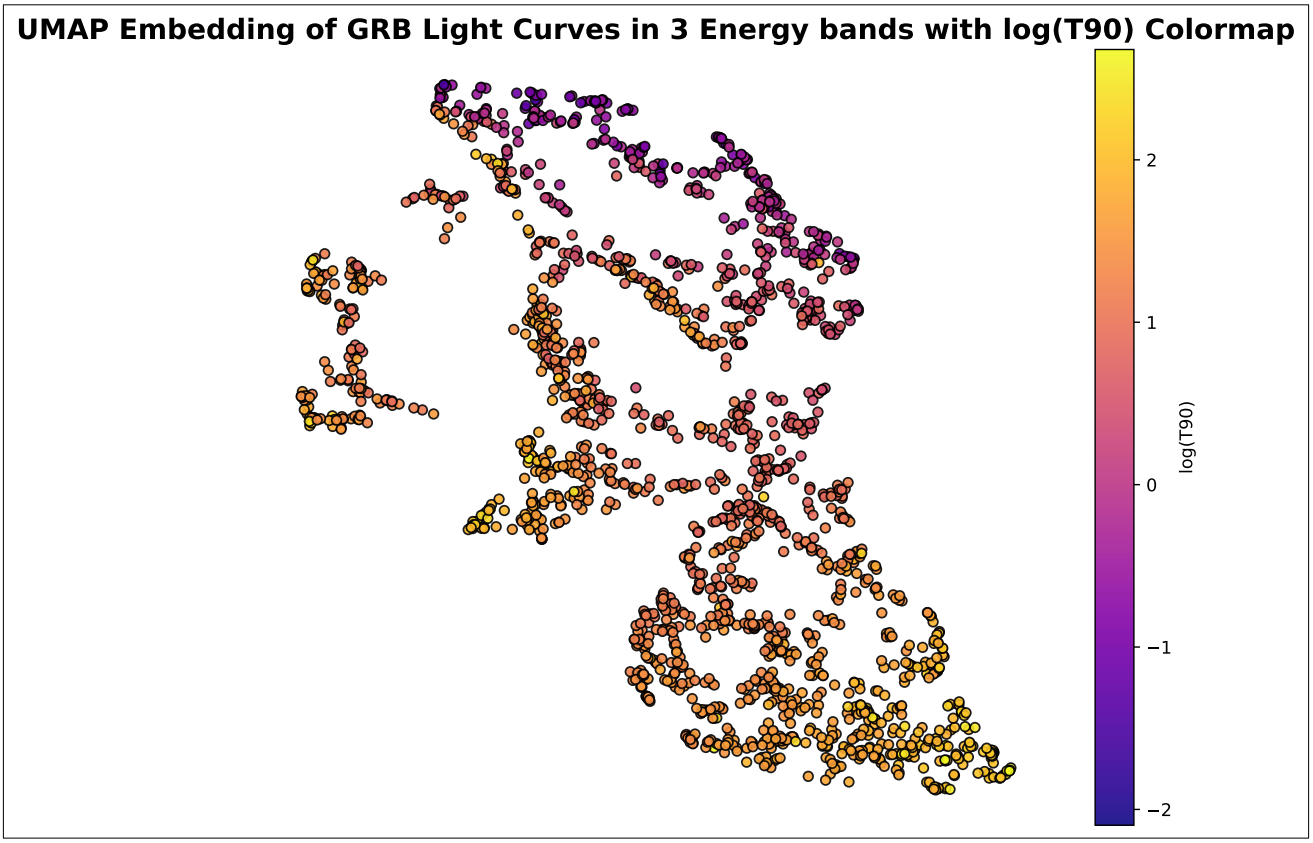


Figure 4.31: UMAP embedding with PCA for 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands, overlayed with duration map. A clear gradient of duration ranging from bottom right to top and this signifies the impact of duration as a parameter for clustering the light curves. One of the remarkable findings is that the short GRBs are not falling into different clusters as is the case with 16 ms light curves Figure 4.14. The presence of the ultra-long group at the tail of the embedding is also intriguing. A supreme distinction in clusters, ie, the boundaries between different clusters are not well defined and this could be because of the part played by noise in the light curves in shaping the background.

Distinction based on duration is clear here as well. It seems that out of the three clusters identified by HDBSCAN, one is predominantly occupied by the short GRBs and the rest by the long GRBs, however, within the long cluster there could be several sub-clusters owing to several reasons inclusive of noise characteristics. Here again, a direct noise comparison is impractical because we have light curves of three different energy bands.

4.4 Some Interesting Comparisons

4.4.1 Comparison of 64 ms light curves

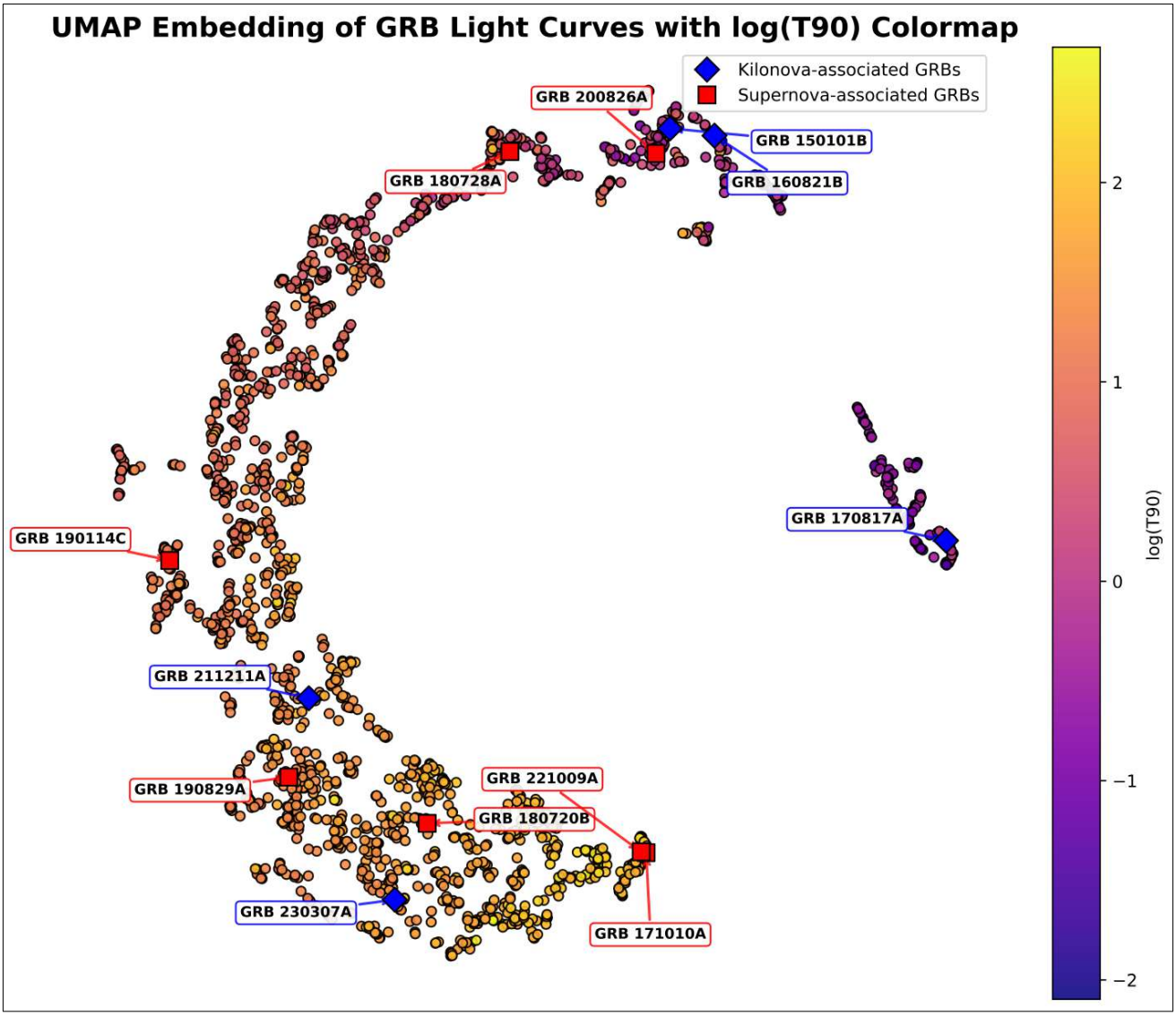


Figure 4.32: UMAP embedding without PCA for 64 ms light curves in 50-300 Kev band with duration color map overlayed and some of the special GRBs are annotated. GRBs associated with Supernova are annotated using red squares and GRBs associated with Kilonovae are annotated using blue diamonds. The positioning of the ultra-long GRBs 221009A and 171010A on top of each other is notable. Also, the positioning of GRB 170817A at a distinct cluster is intriguing.

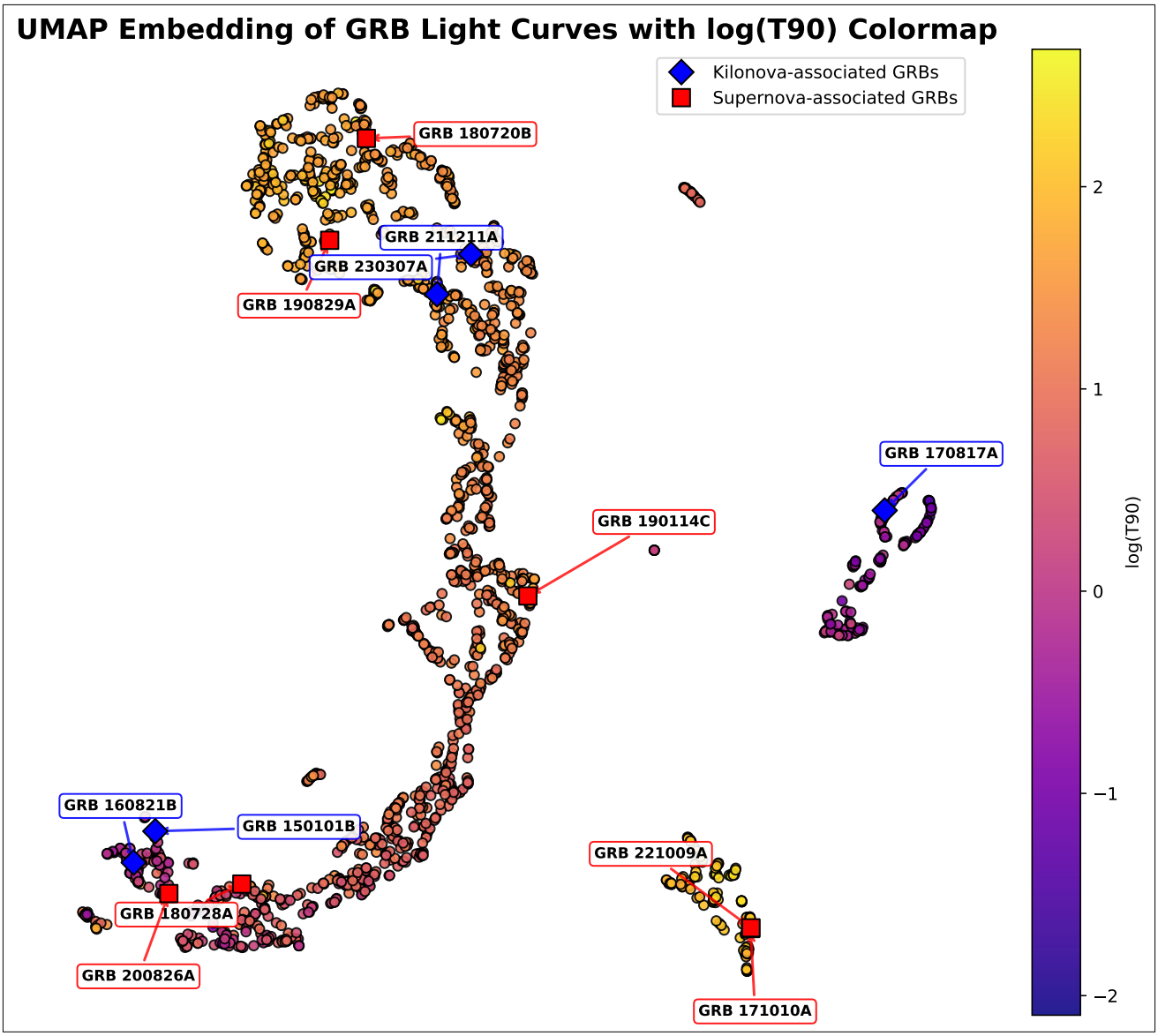


Figure 4.33: UMAP embedding with PCA for 64 ms light curves in 50-300 Kev band with duration color map overlayed with the annotation of Supernova associated GRBs and Kilonovae associated GRBs. The ultra-long GRBs 221009A and 171010A are now positioned at a different cluster. Similar to the analyses without PCA, GRB 170817A is positioned at a distinct cluster. The proximity of GRBs 211211A and 230307A is also interesting.

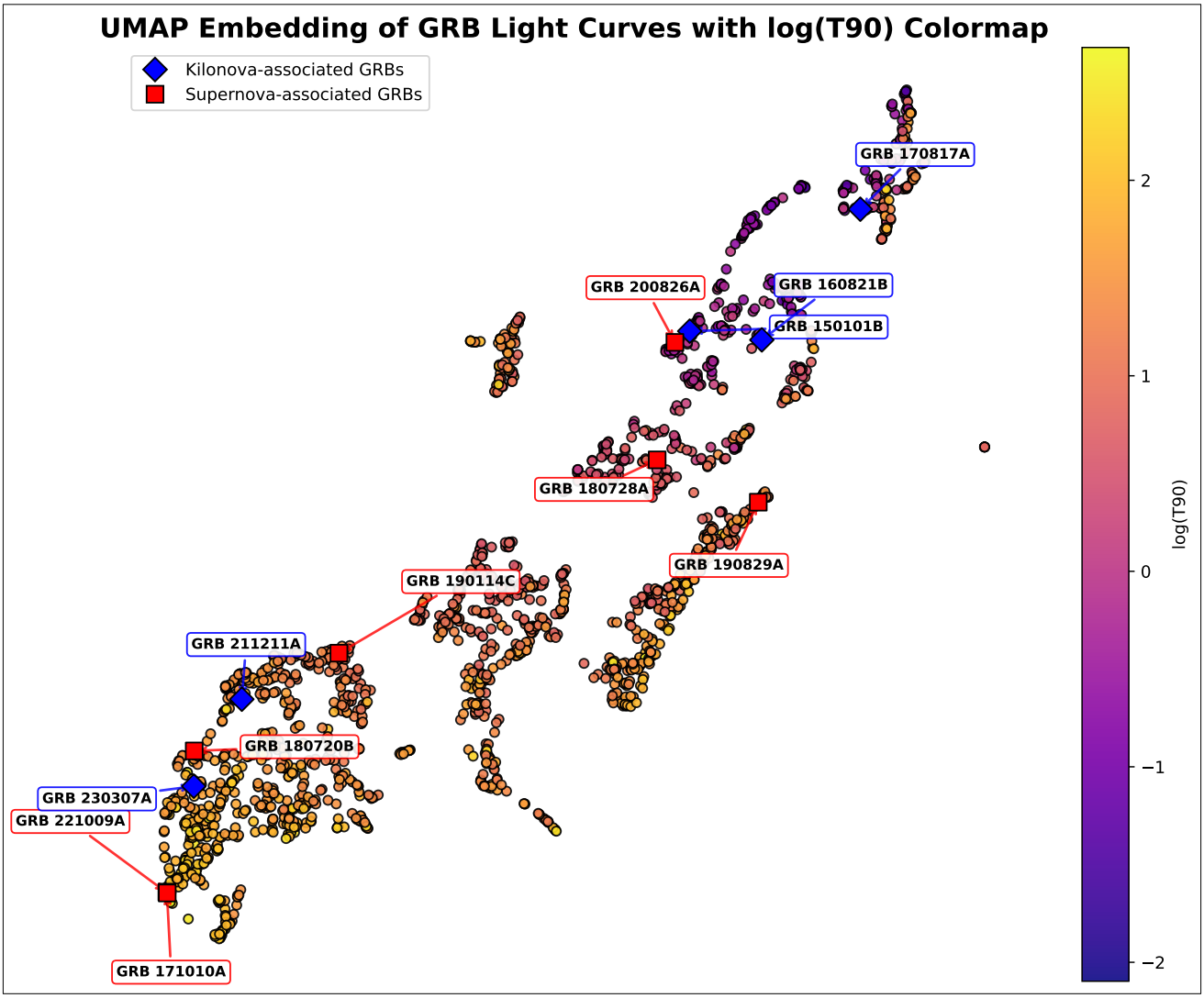


Figure 4.34: UMAP embedding without PCA for 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands with supernova associated and kilonovae associated GRBs annotated in the duration color map. The positioning of GRBs 221009A and 171010A are same that of previous two renderings. However, here GRB 170817A is not occupying a distinct cluster.

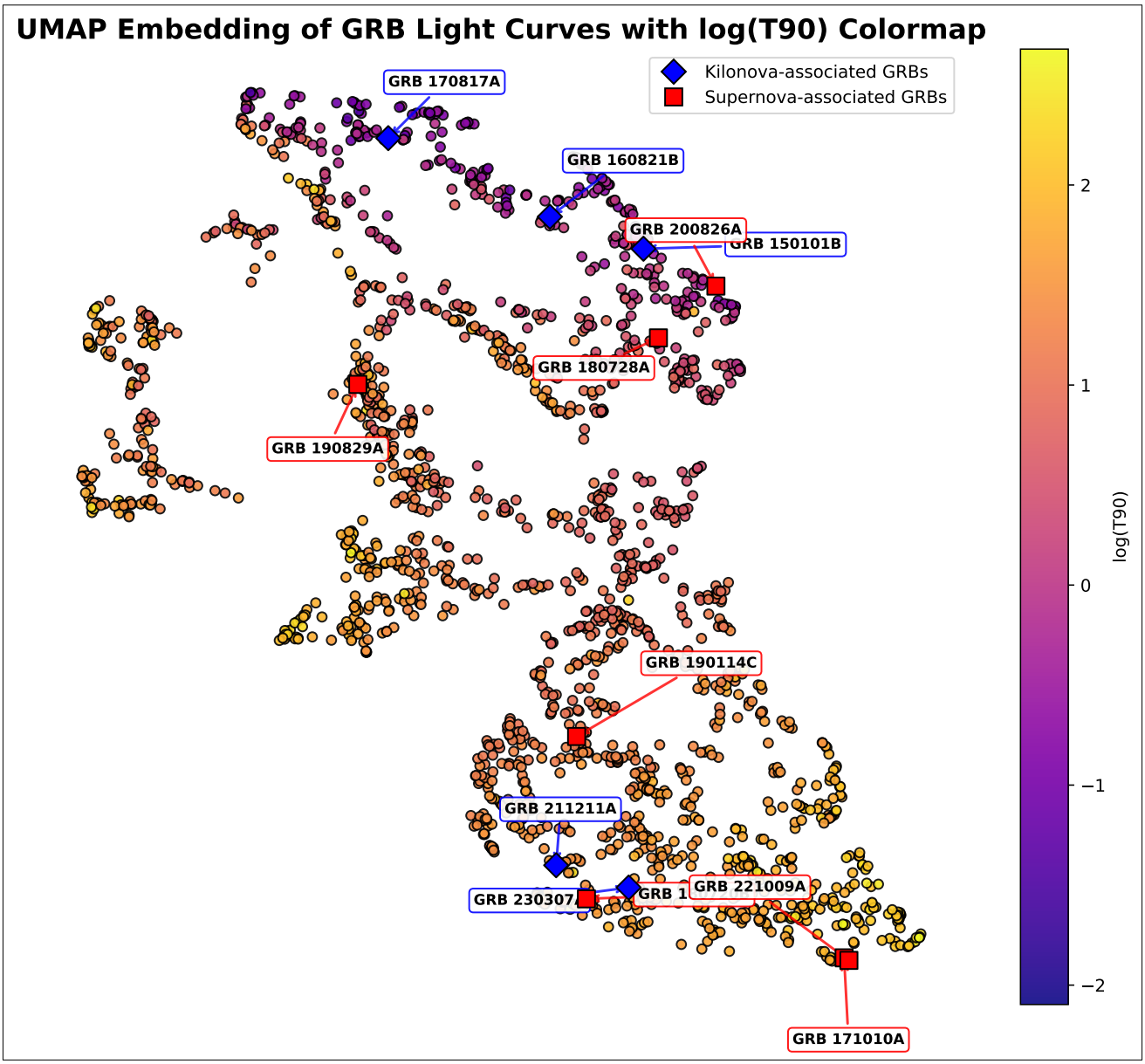


Figure 4.35: UMAP embedding with PCA for 64 ms light curves in 8-50 Kev, 50-300 Kev, and 300-900 Kev bands with supernova-associated and kilonovae-associated GRBs annotated in the duration color map. Even after initializing the embedding with PCA, GRBs 221009A and 171010A are following the same trend. Same as the previous result GRB 170817A is positioned within the same cluster with other short GRBs. In both single-energy band and three-energy band analyses, with PCA initialization, the proximity of GRBs 211211A and 230307A is consistent.

As the adage goes, "A picture speaks a thousand words", we can compare how including and excluding certain information can alter our results drastically. In Figure 4.32 and Figure 4.33, i.e., the analysis for one energy band with and without PCA processing, it is evident that the number of identifiable clusters is different—two in Figure 4.32 and three in Figure 4.33, with an extra ultra-long group gifted by the PCA. However, the presence of two separate short GRB groups is evident in both figures, and from the results of the previous section using the power index, it is clear that it is due to the dominance of white noise in their light curves.

Now for Figure 4.33 and Figure 4.34, even though the rendering of the overall embedding looks drastically different, a closer examination reveals that the neighborhood information in both analyses is consistent with each other.

- Short GRBs and long GRBs are well separated in the embedding.
- GRBs 211211A and 230307A are close neighbors (true except for Figure 4.32).
- GRBs 221009A and 171010A fall on top of each other (an interesting remark is that this is true for all four embeddings!).
- Remarkable separation of GRB 170817A, which is more evident in Figures 4.32 and 4.33 where it occupies a distinct group.
- Positioning of GRB 190829A is also fascinating.

As mentioned, despite the differences in the structure of their embeddings, the analysis of three energy bands with and without PCA (Figure 4.34 and Figure 4.35) is quite similar. However, this is only true for the light curves with 64 ms binning. So what is so special about 64 ms light curves that the analysis with and without PCA is giving the same results? The answer could lie in the noise characteristics of their light curves.

16 ms light curves, being predominant with noise, allow PCA to remove redundancies while the number of components explaining most of the variance is selected from the scree plot. However, in the case of 64 ms light curves with three energy band analyses, PCA becomes redundant because of the dominance of signal over noise in their light curves.

4.4.2 Comparison of 16 ms and 64 ms analysis

It is really important to understand the differences and similarities between the results obtained from 16 ms and 64 ms light curve analyses as the differences contribute to a different interpretation of the clusters.

As a thumb rule, if we are able to decode the same information from two different analyses, that would be something to ponder since we will be attributing astrophysical significance to the clusters if they are found valid. However, the data we are feeding is different in both cases; the 16 ms light curves will have more information regarding both the pulse variability and background variability, therefore the results are expected to be different.

First of all, we can compare the analyses for 64 ms and 16 ms light curves in three bands without PCA (Figure 4.34 and Figure 4.36) and admire the similarities in their results. The long-short dichotomy

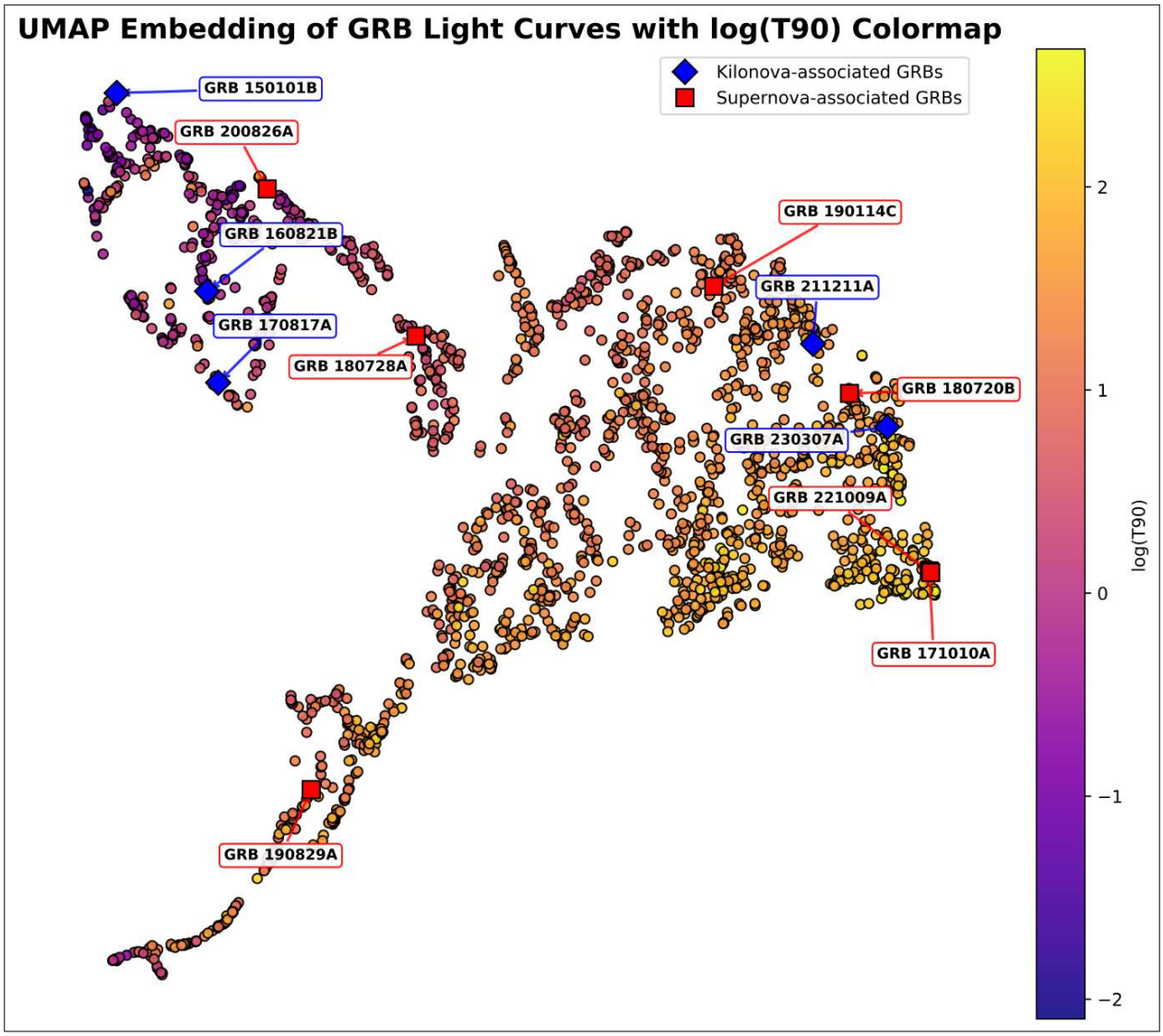


Figure 4.36: UMAP embedding with the color map overlayed for the analyses of 16 ms light curves in three energy bands without PCA. Supernova-associated GRBs are annotated with red squares and Kilonovae-associated GRBs are annotated with blue diamonds. The positioning of GRBs 221009A and 171010A on top of each other is consistent with the previous results.

is not surprising at all; however, the positioning of the triggers within the clusters is also very similar, which is a fact to wonder. Starting from the proximities of GRBs 221009A and 171010A, while we move on further toward their neighbors, astonishingly both the embeddings serve the same information. Think of it like this: identifying the districts within an Indian map, but here with two disfigured maps sharing the same information.

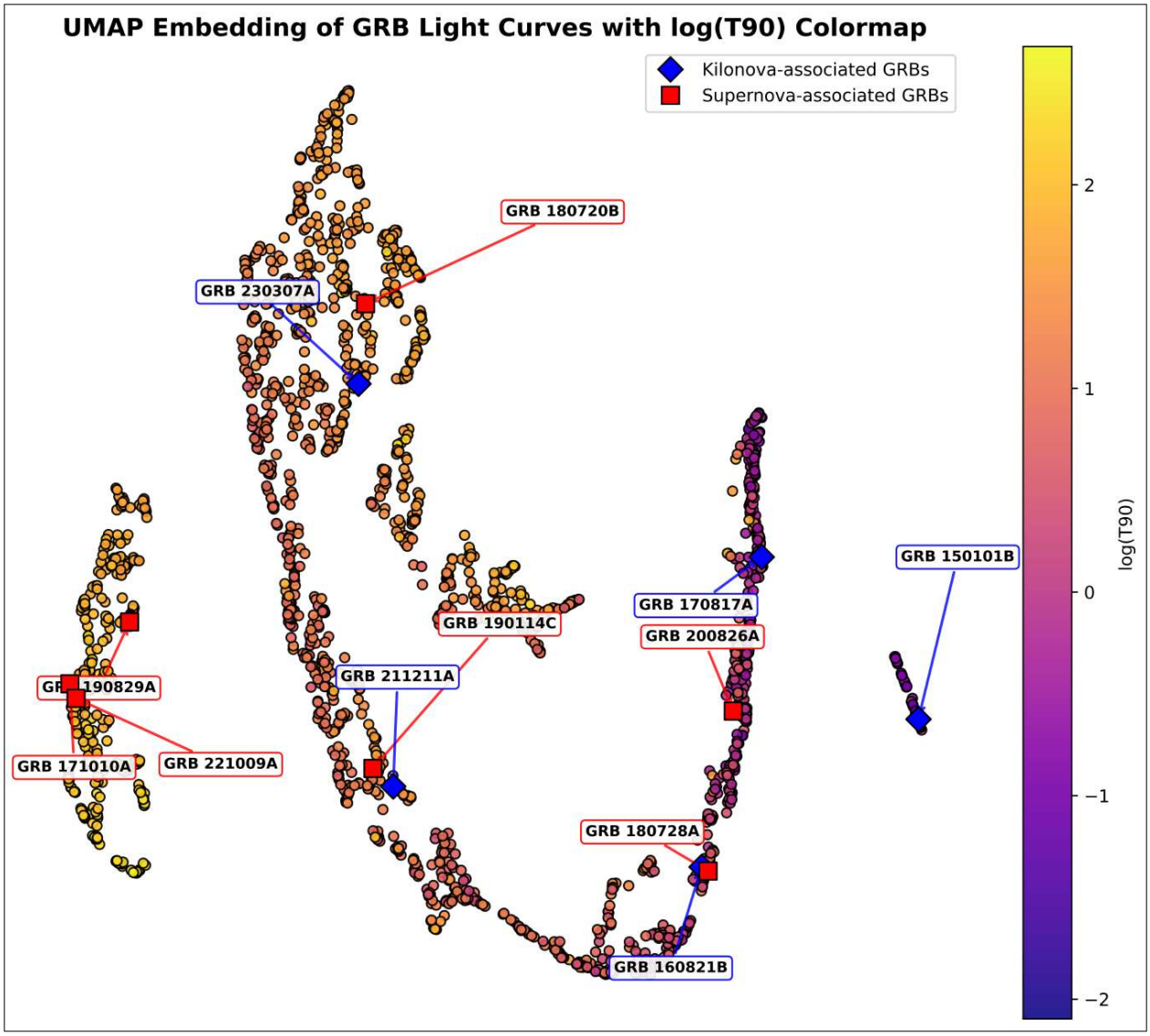


Figure 4.37: UMAP embedding with the color map overlayed for the analyses of 16 ms light curves in 50-300 Kev band with PCA. Supernova-associated GRBs are annotated with red squares and Kilonovae-associated GRBs are annotated with blue diamonds.

In the case of analyses in the 50–300 keV band with PCA, as shown in Figure 4.33 and Figure 4.37, there are stark dissimilarities—even though the striking similarity in forming ultra-long and short GRB clusters separate from the large group persists. As discussed previously, these above-mentioned clusters are a result of the duration characteristics and noise characteristics of the light curves, respectively.

However, the information contained within these clusters is slightly different. Consider the positioning of GRB 170817A in Figure 4.33; whereas in Figure 4.37, it is GRB 150101B that occupies a distinct cluster. Within the ultra-long group in Figure 4.37, GRBs 221009A and 171010A are positioned close to GRB 190829A, whereas in Figure 4.33, GRB 190829A is not a neighbor of these GRBs. GRBs 211211A and 230307A are not neighbors in Figure 4.37, however, this is not the case in Figure 4.33.

Therefore, there is a drastic change in results if we compare the same analyses with PCA, and the changes are to an ignorable extent in the case without PCA.

Despite all these effects, one thing we consistently deduce from the plots is the influence of noise and duration as parameters driving the clustering process. In order to support this claim, we have simulated light curves of different noise characteristics and re-ran the analysis. The results are shown in the next section.

4.5 Light curve simulation and Clustering

To prove that noise in the light curves is a significant parameter driving the structuring of the embedding, purely noise-dominated light curves are simulated and clustered to see if noise of different types falls into distinct groups. Simulations are achieved using the Emmanoulopoulos Light Curve Simulation algorithm [8], where we will input the power spectral density and flux distribution parameters of the light curves and then it will simulate light curves with input parameters. Emmanoupoulos is an updated version of [28], where the flux distribution can be modeled to our needs other than the default Gaussian distribution.

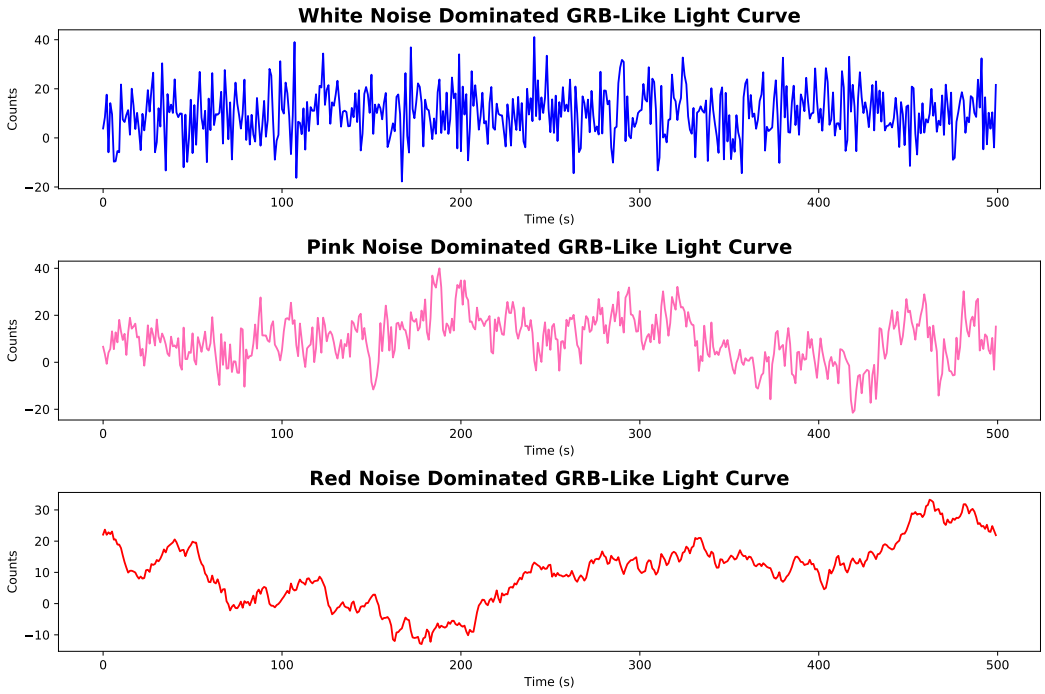
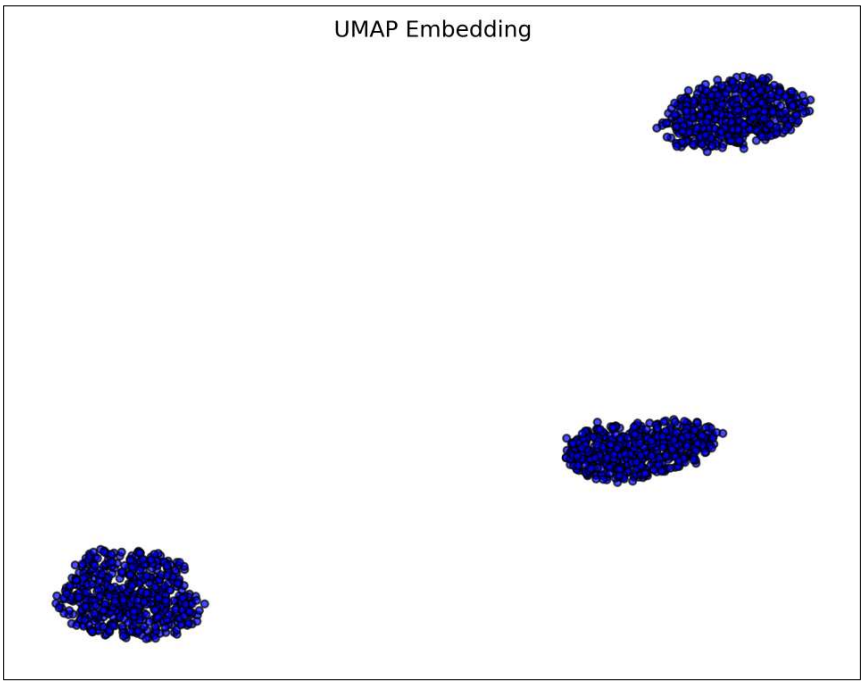


Figure 4.38: Simulated light curves of three different noise characteristics, white noise, pink noise and red noise respectively.

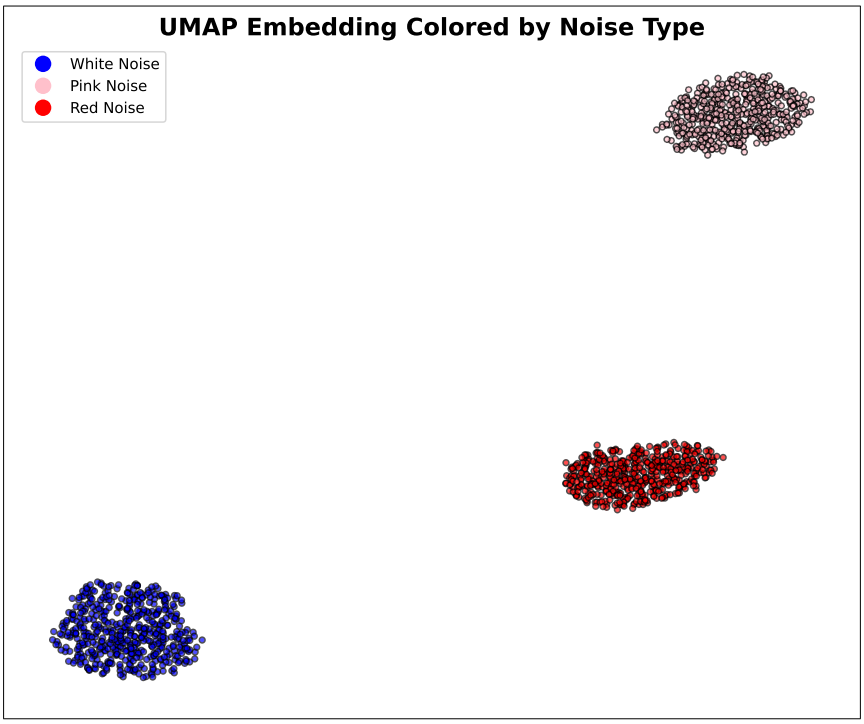
Light curves of power index zero, one, and two corresponding to the white noise (uncorrelated variations), pink noise (intermediate correlations), and red noise (highly correlated) are simulated, say five hundred light curves each. Now these light curves undergo the same preprocessing steps like wavelet denoising, normalization, padding, etc, and are reduced dimensionally and are then clustered.

Our null hypothesis would be noise does not influence the process of clustering and if that is the case we would expect an embedding with a mishmash of data with no clear separation. However,

if noise does have a significant influence, we would be able to see a clear distinction in the clusters thereby rejecting the null hypothesis. The results of this analysis are shown in the plot below.



(a) UMAP embedding obtained after analysing the simulated light curves.



(b) UMAP embedding mapped with actual noise type of light curve

Figure 4.39: Embedding obtained after the analysis of simulated light curves.

Since our data is labeled, this forms a supervised clustering problem where we can map the exact noise property of the light curves as shown in Figure 4.39b. This proves the impact of noise characteristics in the clustering process. Now moving one step closer to reality, we can take different combinations of the noise and perform the analysis again as real-world light curves are a mix of different combinations of noise. The results are shown below.

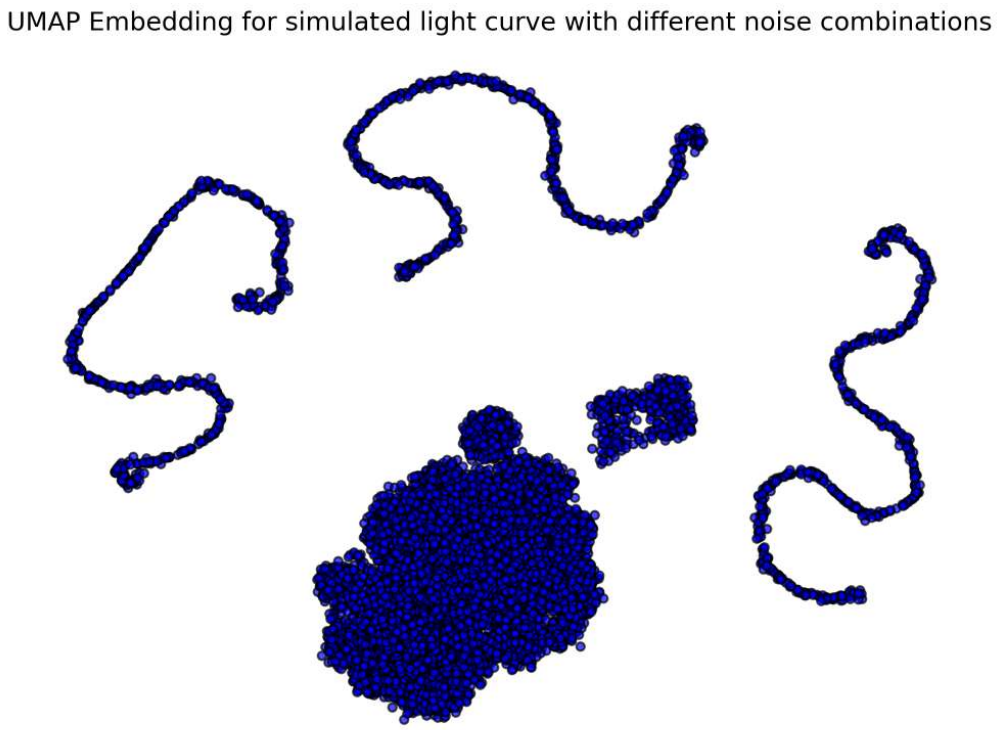


Figure 4.40: UMAP embedding obtained after considering different combinations of noise

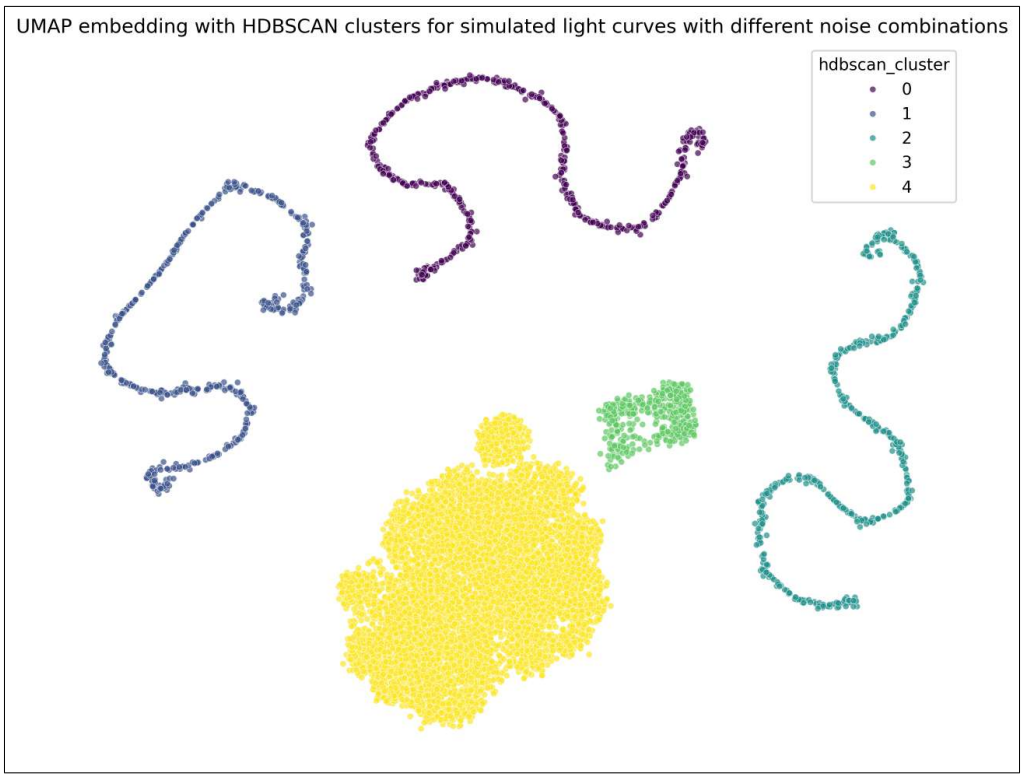


Figure 4.41: UMAP embedding clustered with HDBSCAN.

We can see that several distinct clusters are formed in the UMAP embedding. Our next step is to identify these clusters using HDBSCAN and map the real combination of duration of each data point.

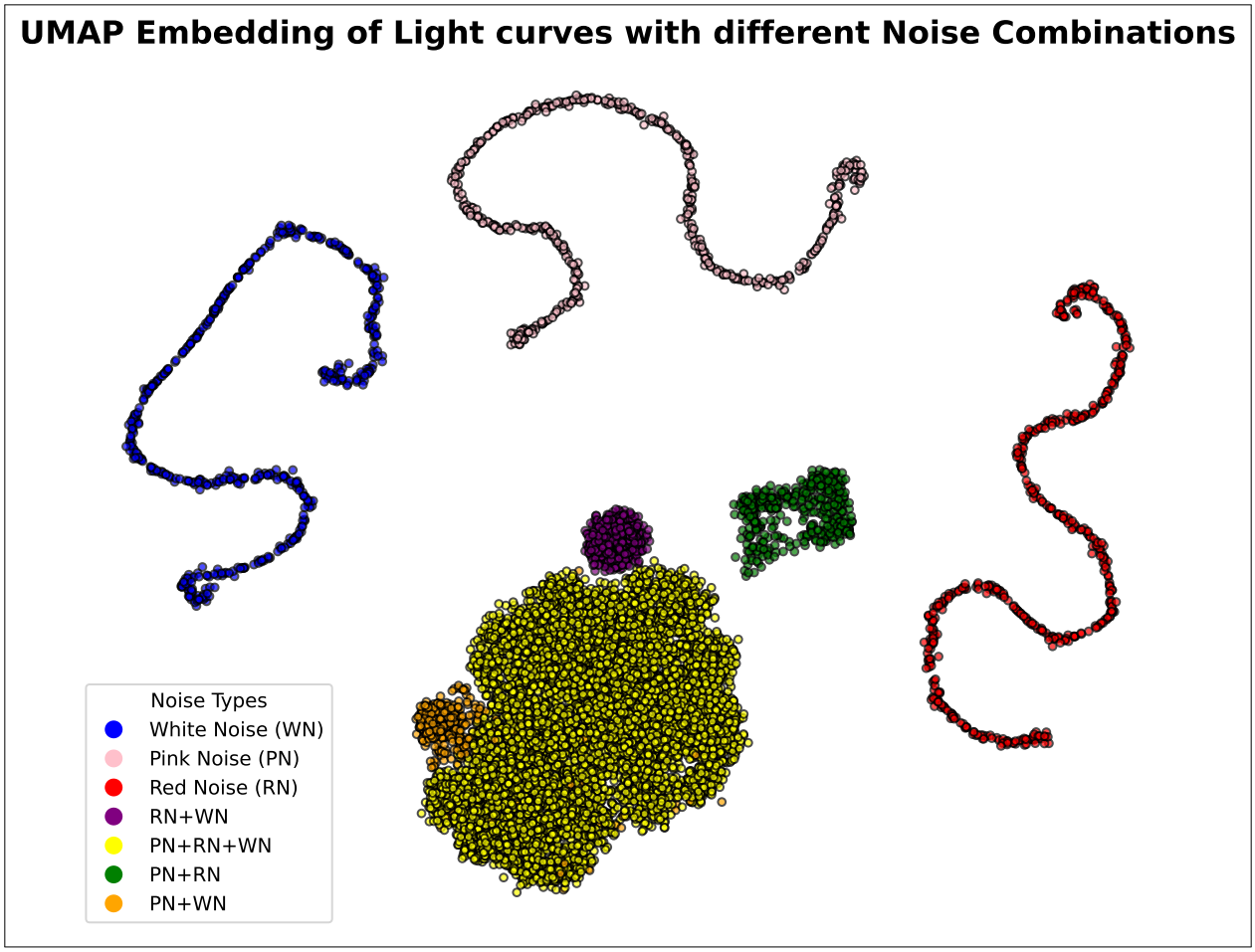


Figure 4.42: UMAP embedding colored with actual type of noise combinations in the light curve

It didn't come as a surprise that the distinct clusters in the UMAP formed a unique combination of noise along with pure noise-dominated light curves. It is fascinating to see how even the structures are different for different combinations. All the pure noise-dominated light curves form a worm-like structure, which stands out from the rest of the group. One interesting thing to note here is that HDBSCAN is identifying only five clusters despite in reality it being seven. Even if we were to select the number of clusters manually the answer, of five would have been no different, ie, had the data have been not labeled there is no way to know that there would be seven clusters.

It is because of the close similarity of the combinations, White noise + Pink noise + Red noise, Pink noise + White noise, and to an extent, Red noise + White noise. The first two are nearly indistinguishable!

The results of the simulation prove our hypothesis that noise characteristics of the light curve, say white, pink, red, etc play a crucial role in forming the structure of the manifold. It is also important to note that duration also has a significant impact but only in the case of ultra-long GRBs.

This calls for caution while interpreting the clusters because the noise in our analysis is coming from the background of the signal and not from the variability in its transient, spiky pulse. Therefore, it has almost zero physical significance. To make meaningful conclusions from the clusters, first, we have to mitigate the impact of background in the analysis thereby removing the noise as a parameter driving this process.

Chapter 5

Conclusions

It was seemingly impossible to make data-driven decisions from the big data, serving the definition of three 'V' pillars, Volume, Velocity, and Veracity, in astronomy in the past decade[20]. However, with the exponential growth in the field of data science, analyzing a large amount of data is no longer a dare. The growth of Machine learning out of the computer science classrooms served a great deal of help in untangling some of the hardest conundrums in astronomy. One of them is the classification of GRBs, which is a long-standing problem.

Even though analyzing the data has become easier, the trade-off during this period of evolution was done with the interpretation of results. Conclusions drawn from the analysis, especially from an unsupervised method like clustering are prone to ambiguity. The results are sensitive to hyperparameter tuning and as there is no optimal way to perform this task, all the results obtained are both meaningful and incomplete at the same time. That is, each UMAP embedding with a different hyperparameter will have a different part of the same story to be untold.

In our case, it is of foremost importance to know which parameters have a significant impact on driving the process of clustering and how sensitive the data preprocessing steps like wavelet denoising and feature extraction are to the results drawn.

From our analysis taking into account several edge cases, like considering different energy bands, initializing UMAP with and without PCA, and carrying out the analysis with light curves of different binning, it has been able to prove that even the slightest change in the preprocessing steps would impart a huge impact on the number of clusters formed and the results drawn from those clusters, ie it is like a butterfly-effect. However in all the cases two things were common, separation of clusters based on duration and noise properties of the light curve. In the case of ultra-long GRBs, duration was the sole parameter driving the manifold learning whereas in the case of short GRBs, noise played a significant role. This is because short GRBs are dominated by background noise especially when we are extracting the light curves from zero to t90 and are making the lengths equal by padding it with zeroes. It is this padding that is influencing a certain number of short GRBs to position at a different cluster.

Figure 5.1 shows the light curves extracted in the range from zero to t90 s and then padded to a fixed length and their corresponding |FFT|. It is clear from the figure how the amplitudes vary for light curves with different t90 values. For the light curves with t90 in the milli-second range, light curves will look even more 'disfigured' with a single pulse, and the rest of everything will be a series

of zero counts, which contributes to the background in our analyses. However hard we try, it is nearly impossible to input just the variability in the pulse excluding the noise as we need the data to form a matrix to feed into the algorithm and the positioning and length of the pulse variability are different for different bursts.

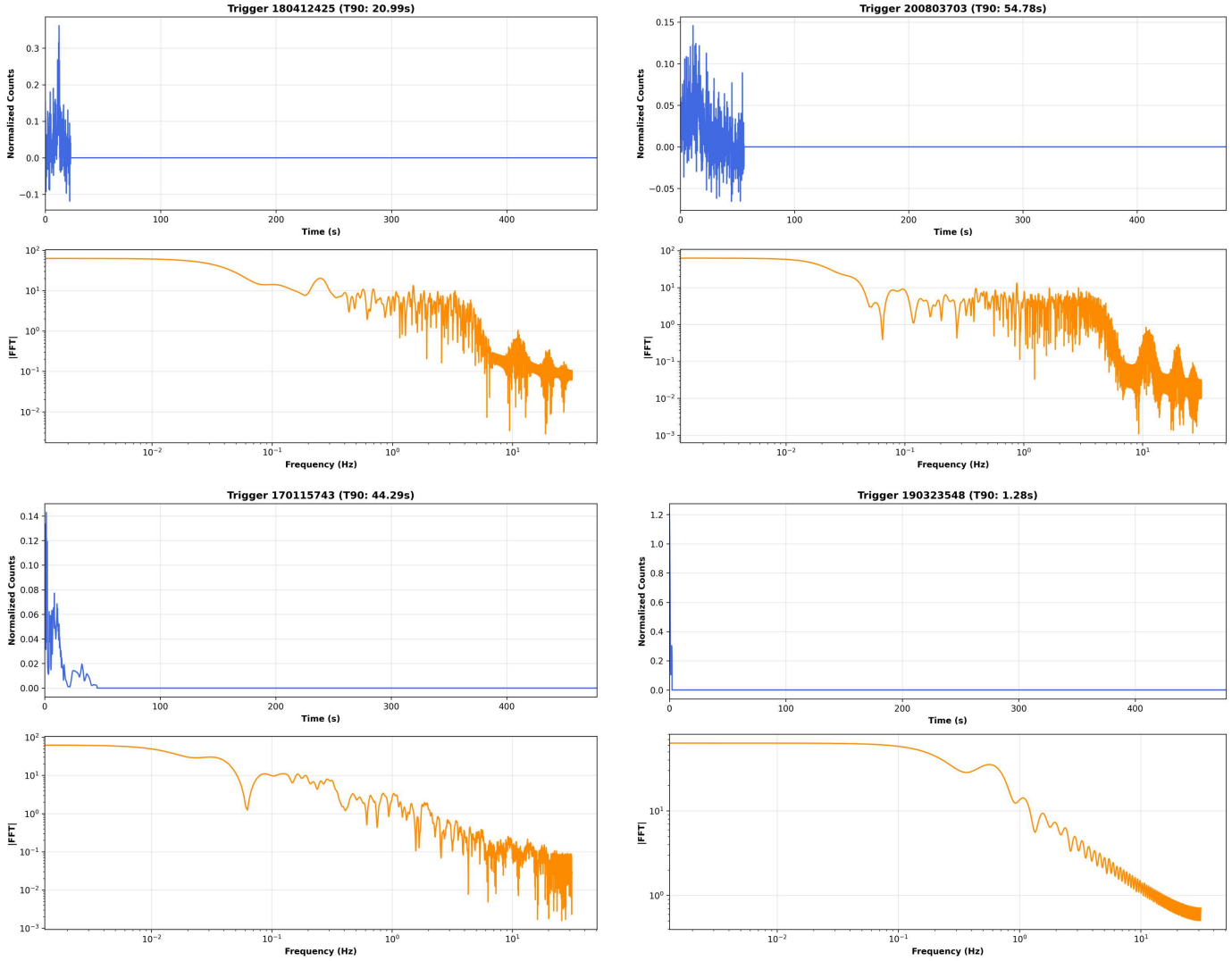


Figure 5.1: Light curves extracted in the range zero to t90 s and their corresponding power spectra

From our analysis as well as from the simulations it turns out that noise from the background of the light curve plays a crucial role in grouping the data points. This issues an alert while interpreting the clusters, ie, they may not have any physical significance as we have thought earlier. This is true at least in the case of short GRBs which are noise-dominated. Therefore we could misattribute physical significance to something that may be merely a statistical artifact.

Moving forward with this work, our main goal is to mitigate the impact of the background of the light curves in our analyses and draw meaningful information from the clusters. Increasing the sample size is another aspect of this work. At present, there are 2000 GRB light curves preprocessed (fitted the background and denoised) which we are planning to extend to around 3,800 in number. The results obtained from FERMI-GBM can be compared with data from other instruments like SWIFT-BAT, INTEGRAL, etc.

Chapter A

Light curves and power spectra within each cluster

Consider the UMAP embedding of 16 ms light curves in 50-300 Kev band without PCA initialization.

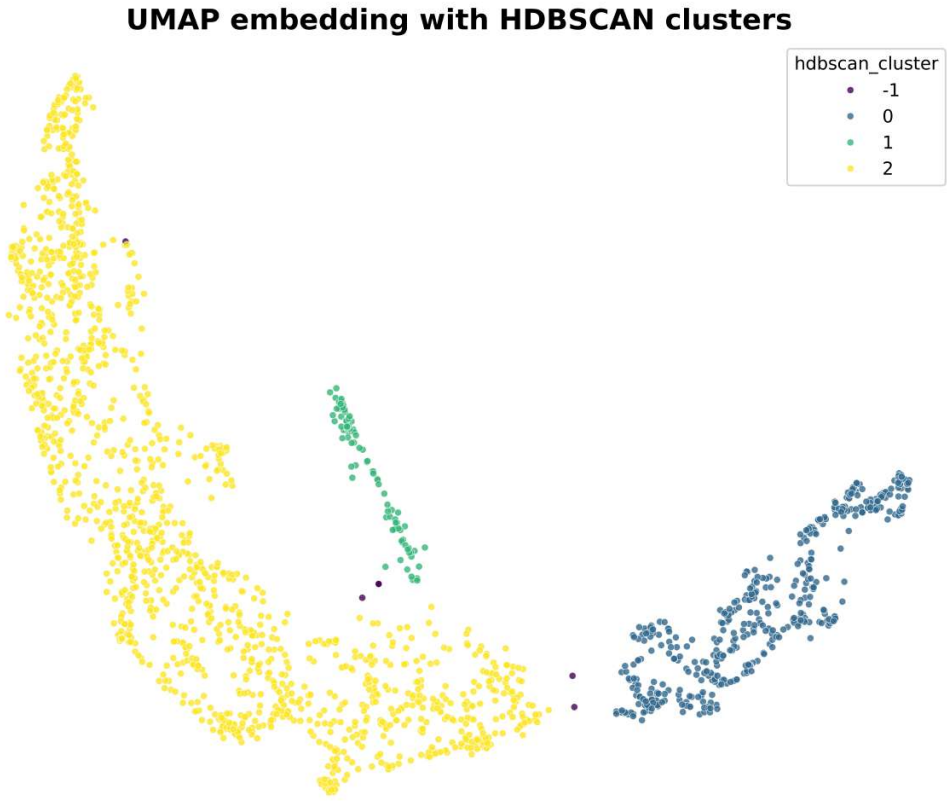
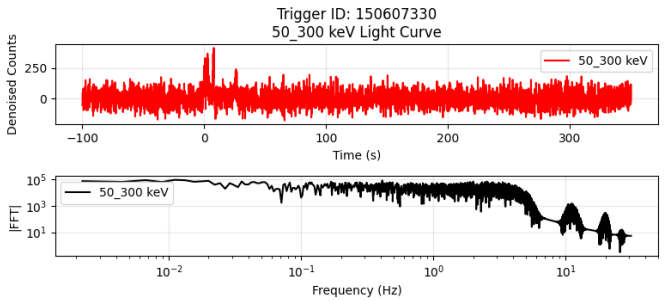


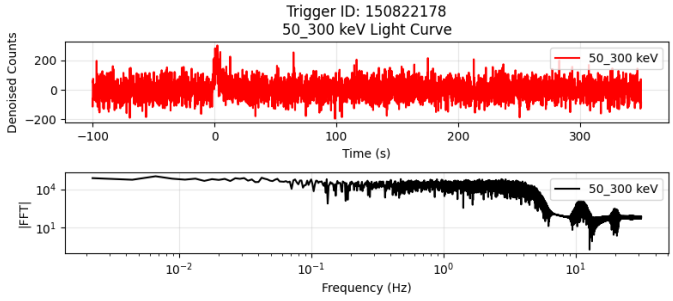
Figure A.1: HDBSCAN embedding of 16 ms light curves in 50-300 Kev band without PCA initialization

We can take a look at how the light curves look in each cluster by taking a few representative samples.

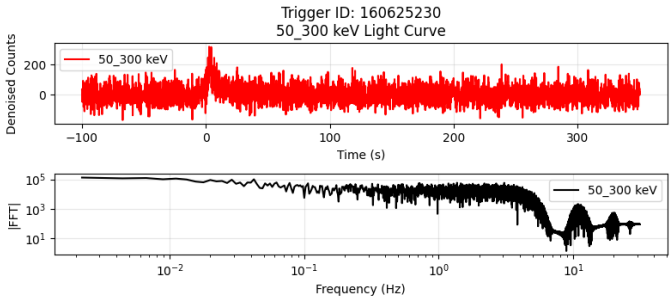
A.0.1 Light curves in the cluster 1 (Green in color)



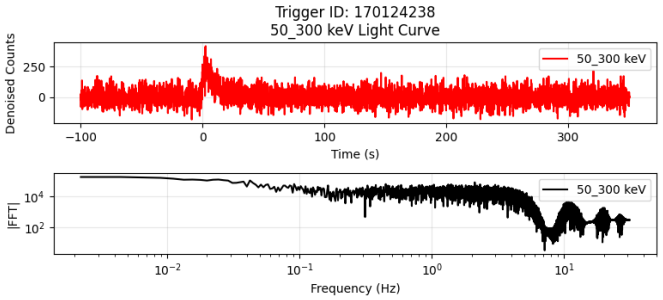
Trigger ID: 150607330



Trigger ID: 150613995



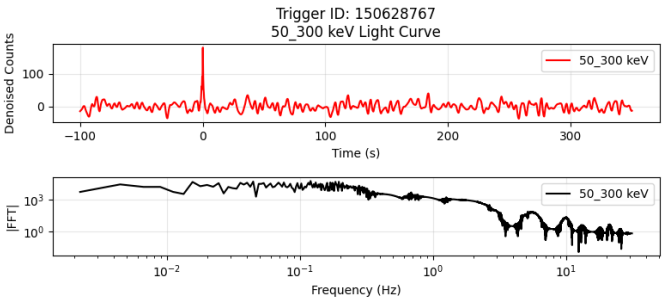
Trigger ID: 150822178



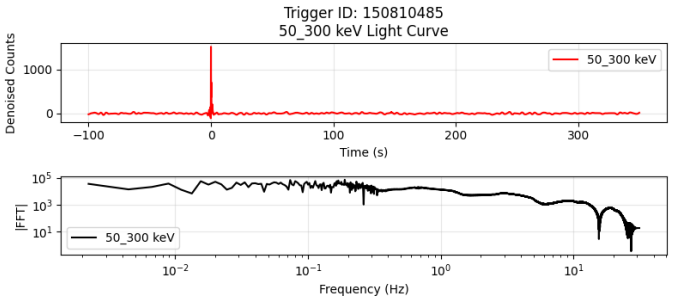
Trigger ID: 151023104

Table A.1: Light curves and power spectra for selected GRB triggers in the cluster 1 (green colored cluster). Each plot shows the denoised light curve and its corresponding Fourier power spectrum.

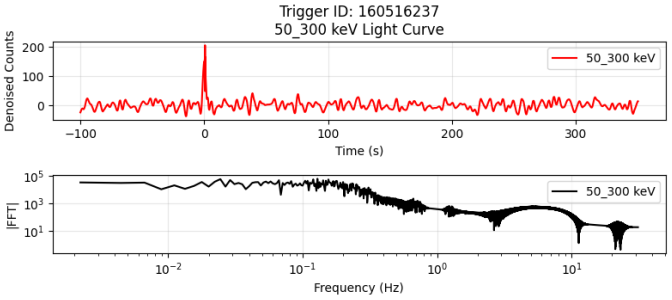
A.0.2 Light curves in the cluster 0 (Blue in color)



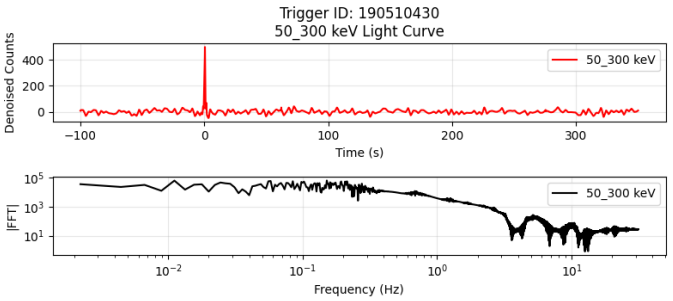
Trigger ID: 150607330



Trigger ID: 150613995



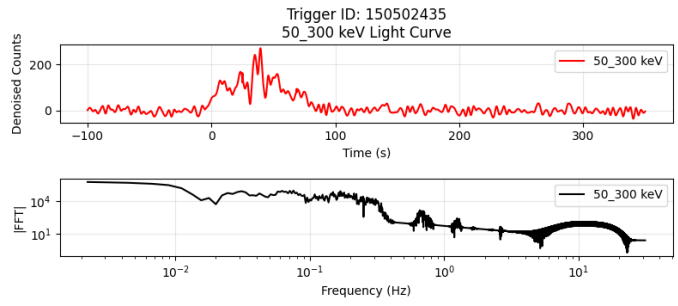
Trigger ID: 150822178



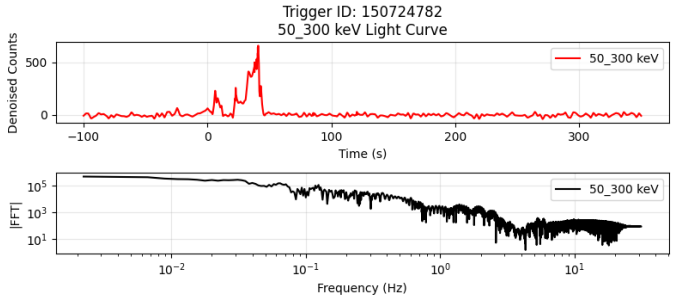
Trigger ID: 151023104

Table A.2: Light curves and power spectra for selected GRB triggers in the cluster 0 (blue colored cluster). Each plot shows the denoised light curve and its corresponding Fourier power spectrum.

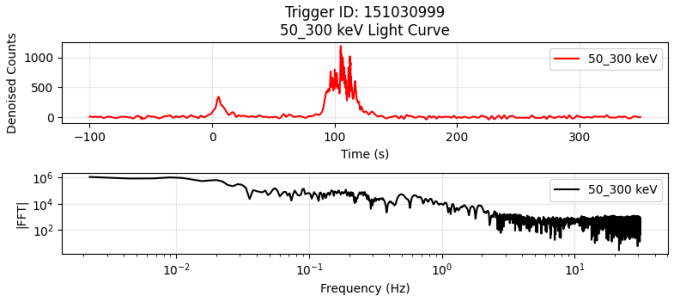
A.0.3 Light curves in the cluster 2 (Yellow in color)



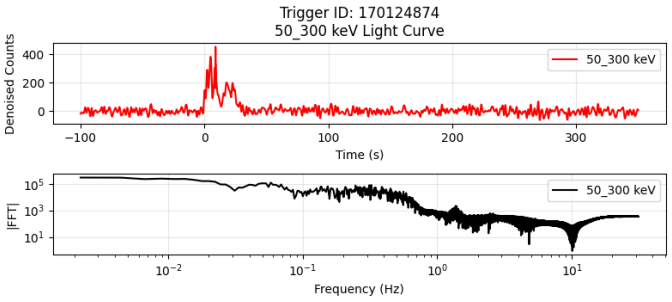
Trigger ID: 150607330



Trigger ID: 150613995



Trigger ID: 150822178



Trigger ID: 151023104

Table A.3: Light curves and power spectra for selected GRB triggers in the cluster 2 (yellow colored cluster). Each plot shows the denoised light curve and its corresponding Fourier power spectrum.

It is evident that within the clusters light curves are more or less similar and in different clusters, light curves are significantly different as expected. We can see the same trend for the Fourier amplitudes of the light curves as well. However in Table A.1, the power spectra show periodic behavior at high frequencies, this behavior is manifested from the background of the light curve. Even though the fluctuations seem to be random, it is appearing as a periodic behavior in the light curve. However in our analysis, light curves are extracted only in the range from zero to t90 s, and all the light curves are padded with zeroes to the maximum length of the particular light curve, which in our case is 483 s.

Therefore the light curves we are feeding to the algorithm looks like that in the Figure 5.1. However the same analyses with the 'real' light curves, ie light curves being not extracted in a range from zero to t90, but from say -50s to 350s is also done as shown in the Chapter C of the appendix. One interesting thing is that there are only ignorable differences in both the analyses except for that of three energy bands.

Chapter B

Analysis with 16 ms Light curves in 8-50 Kev and 50-300 Kev bands

B.1 With PCA initialization

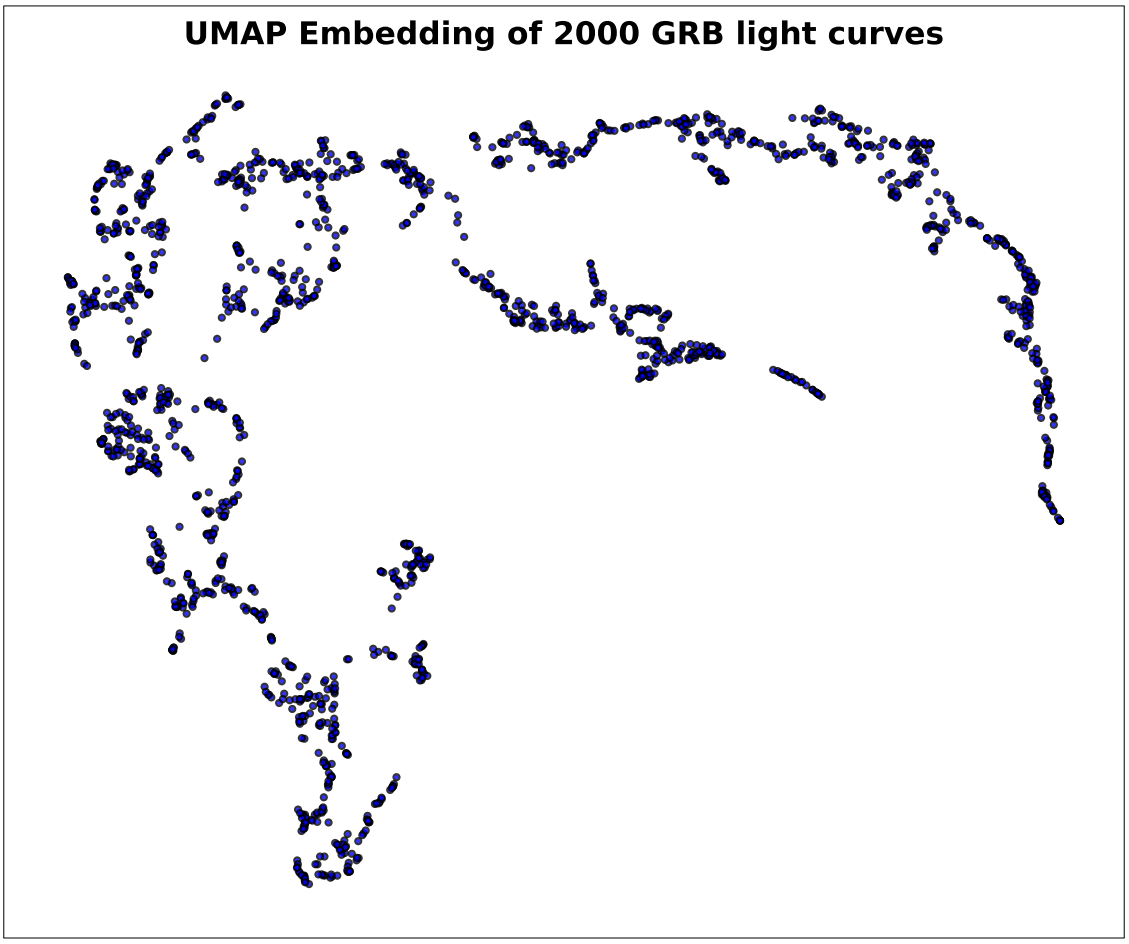


Figure B.1: UMAP embedding of 16 ms light curves in 8-50 Kev and 50-300 Kev bands with PCA initialization.

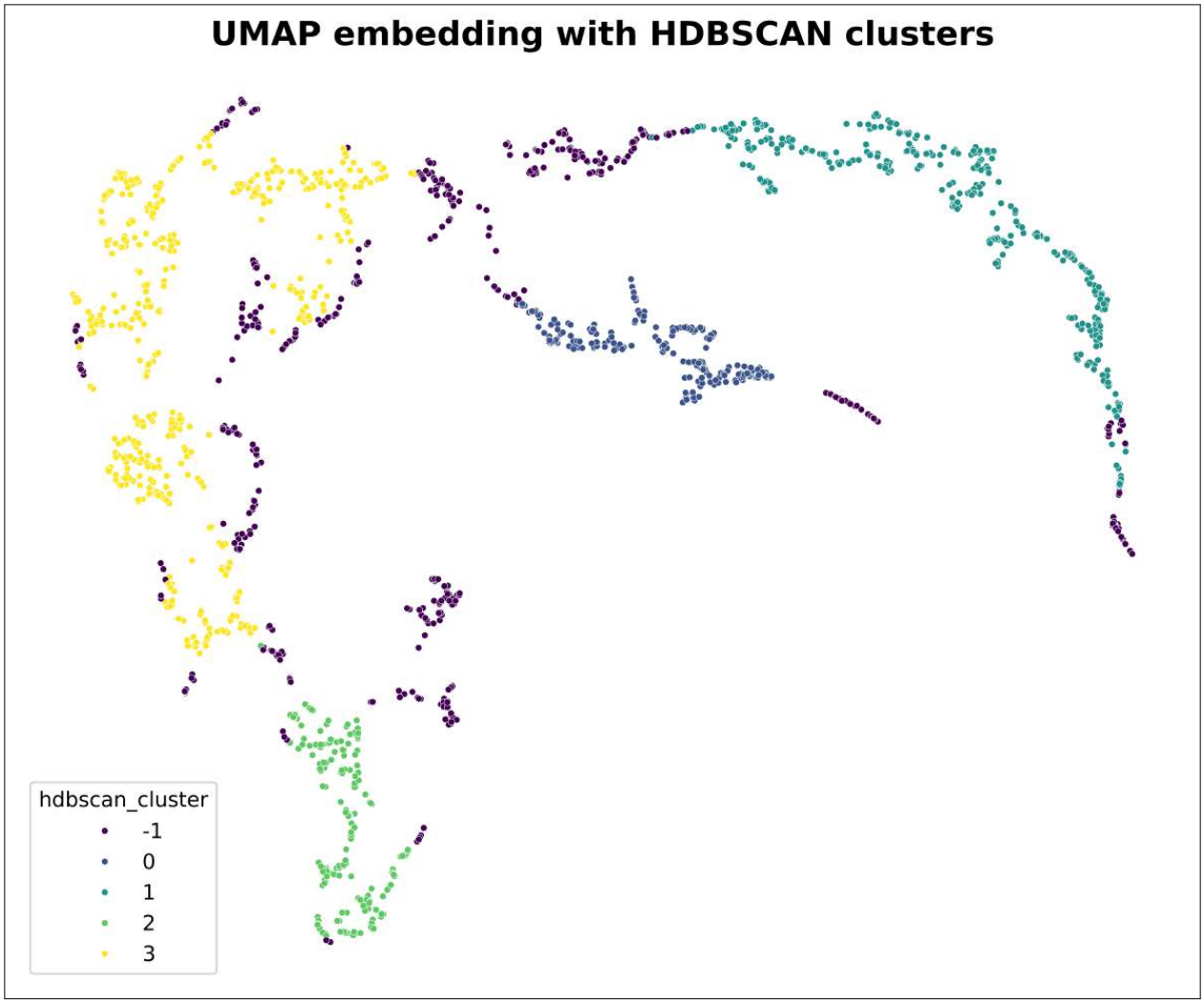


Figure B.2: There are four clusters identified by HDBSCAN with outliers labeled as -1

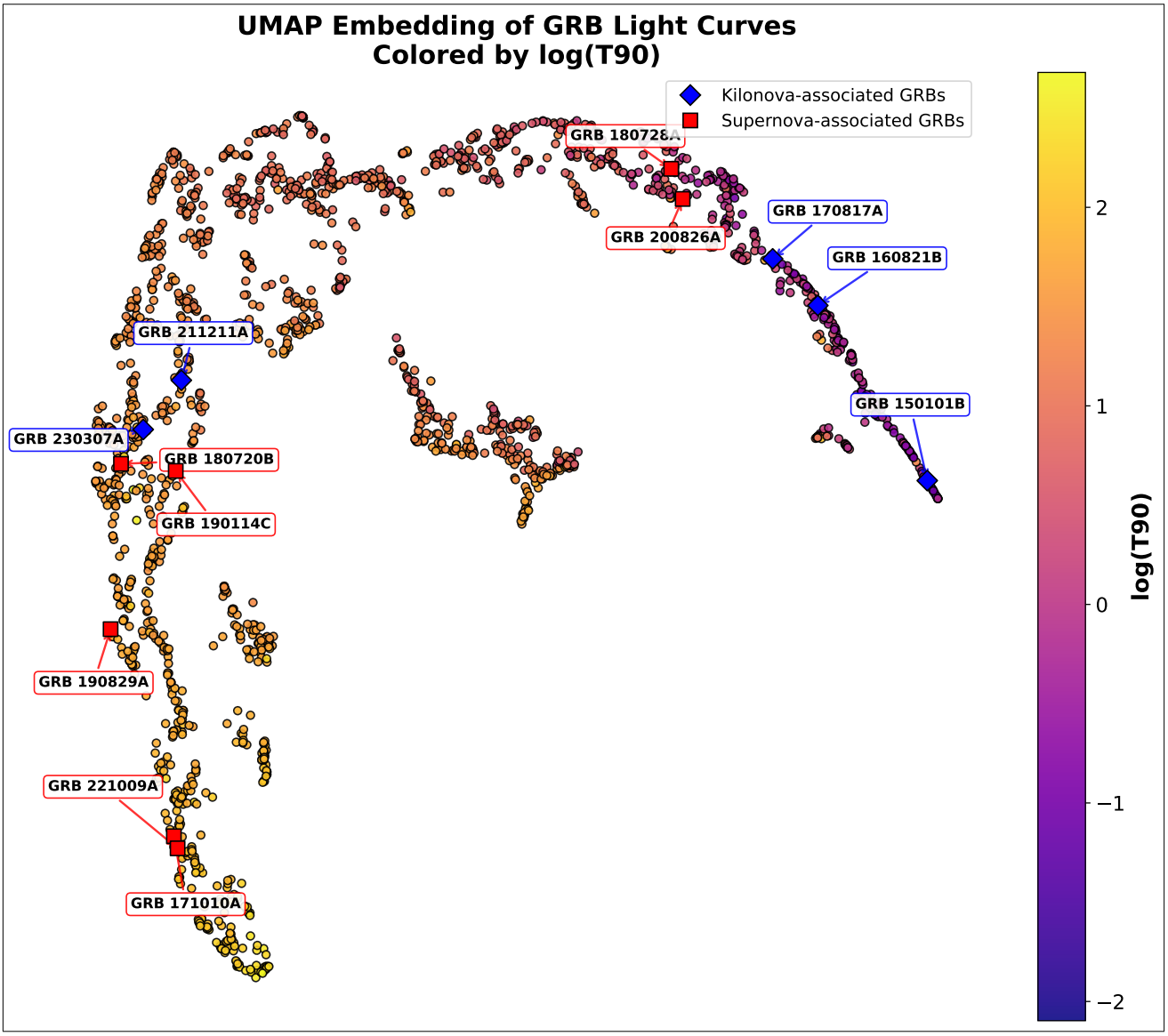


Figure B.3: UMAP embedding with the color map overlayed for the analyses of 16 ms light curves in three energy bands without PCA. Supernova-associated GRBs are annotated with red squares and Kilonovae-associated GRBs are annotated with blue diamonds. Consistent with the previous results, GRBs 171010A and 221009A are very close to each other. GRBs 230307A and 211211A are also nearest neighbors. Positioning of Supernova associated GRBs 180728A and 200826A together with the short GRBs is intriguing.

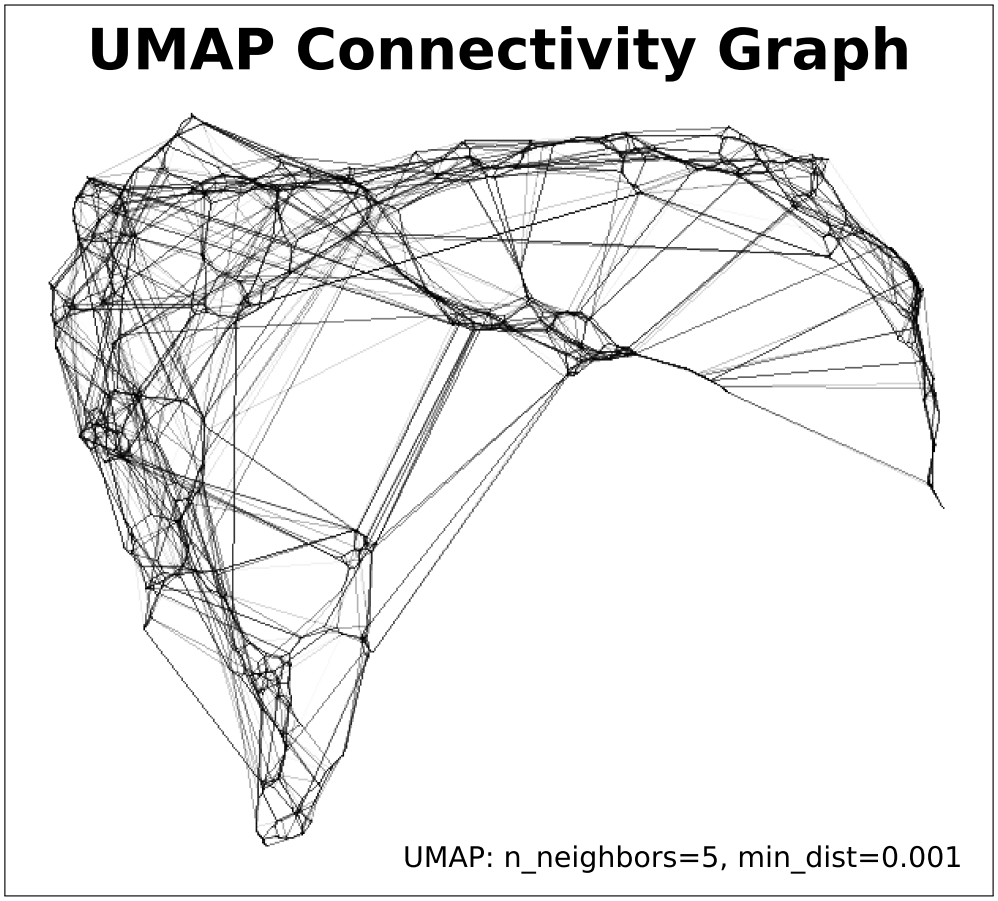


Figure B.4: Connectivity graph of UMAP, a representation of whom is near whom before dimensionality reduction distorts the distance in lower dimension

Same analyses is carried out for light curves in two energy bands, 8-50 Kev and 50-300 Kev. It can be seen from the Figure B.3 that here again the manifold is structured based on the duration and is consistent with our previous results. Now in order to understand how each points are connected to each other, we can plot a connectivity graph with nodes as data points and the edges connecting them as a measure of similarity Figure B.4.

Connectivity graph gives us an idea of how data points were close in high dimensional space before we projected on to the lower dimension. Closer points in the higher dimension are connected with a higher edge weight (darker in color), where as if the points are not neighbors then it will have low edge weights (lighter in color). Densely connected groups in the graphs represent the potential clusters in the embedding and if two groups are weakly connected it represents separate groups. Isolated nodes represent the outliers. HDBSCAN performs well on UMAP because it works directly on this graph layout.

One important thing to keep in mind while interpreting the connectivity graph is that it does not directly reflect the two dimensional embedding, it is the high dimensional structure UMAP constructs so as to lay things down easily on to the lower dimension.

The same analyses for 16 ms light curves in two energy bands are carried out without PCA initialization. Results are shown in the Figure **??** and the connectivity graph is shown in the Figure B.8.

B.2 Without PCA initialization

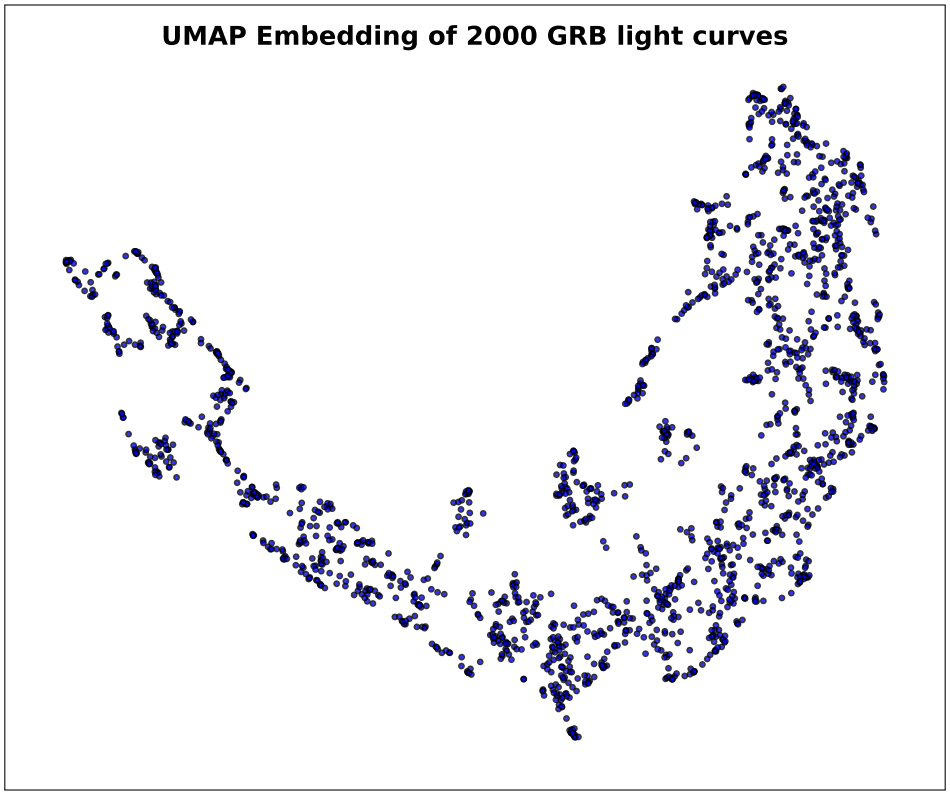


Figure B.5: UMAP embedding of light curves in two energy bands without PCA

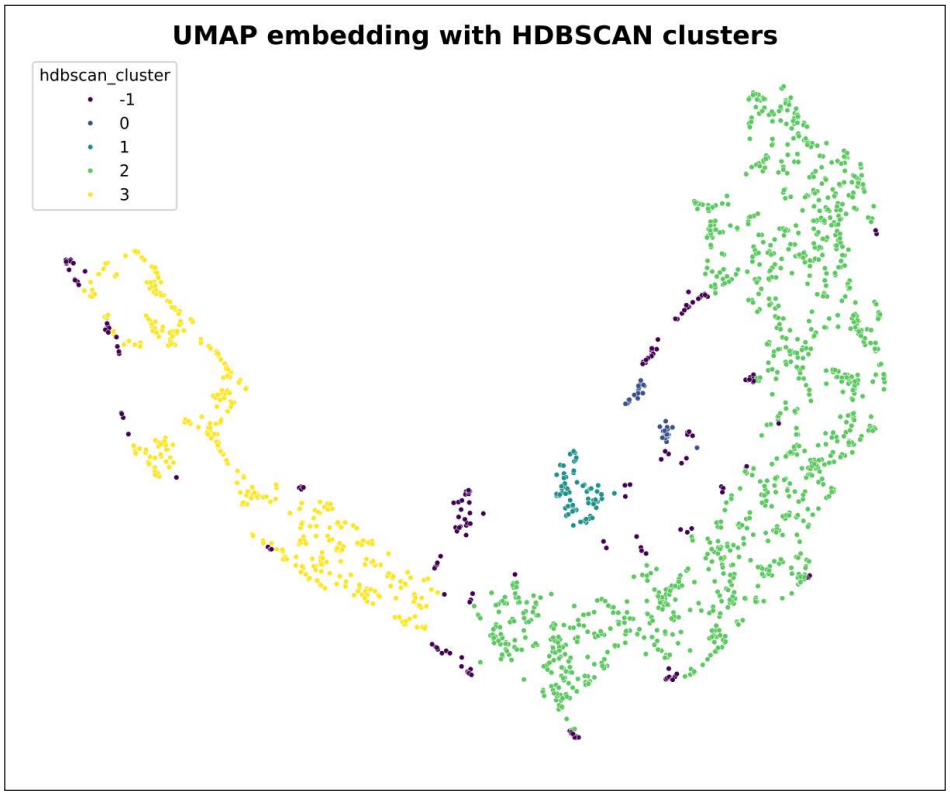


Figure B.6: Clusters identified by HDBSCAN with outliers labeled as noise

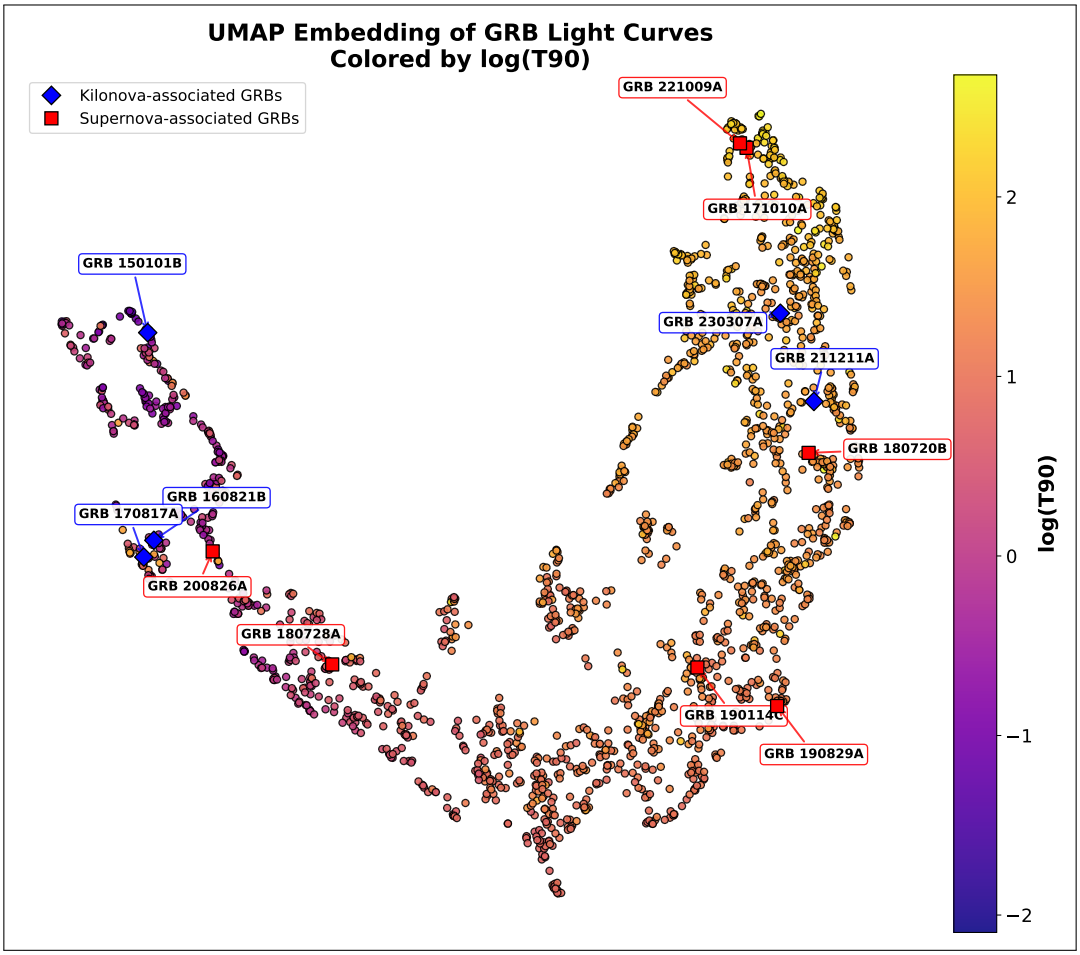


Figure B.7: UMAP embedding with the color map overlayed for the analyses of 16 ms light curves in two energy bands without PCA. Supernova-associated GRBs are annotated with red squares and Kilonovae-associated GRBs are annotated with blue diamonds. Consistent with the previous results, GRBs 171010A and 221009A are very close to each other. GRBs 230307A and 211211A are also nearest neighbors. Positioning of Supernova associated GRBs 180728A and 200826A together with the short GRBs is intriguing.

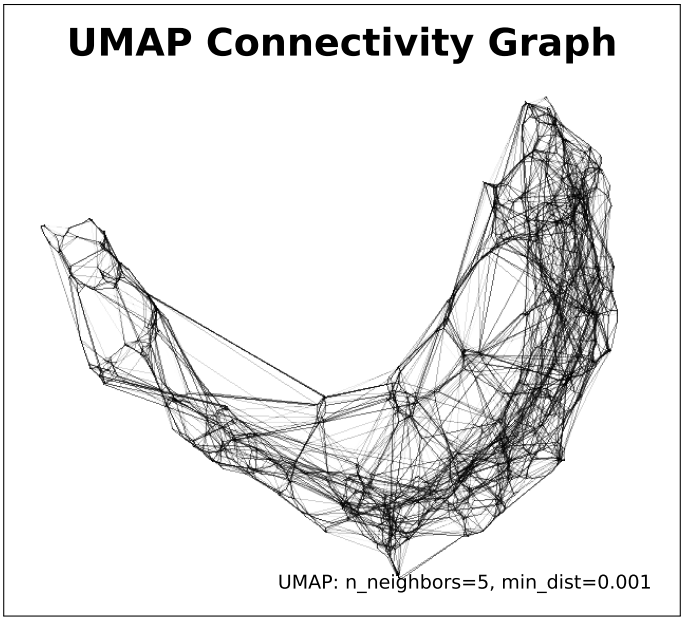


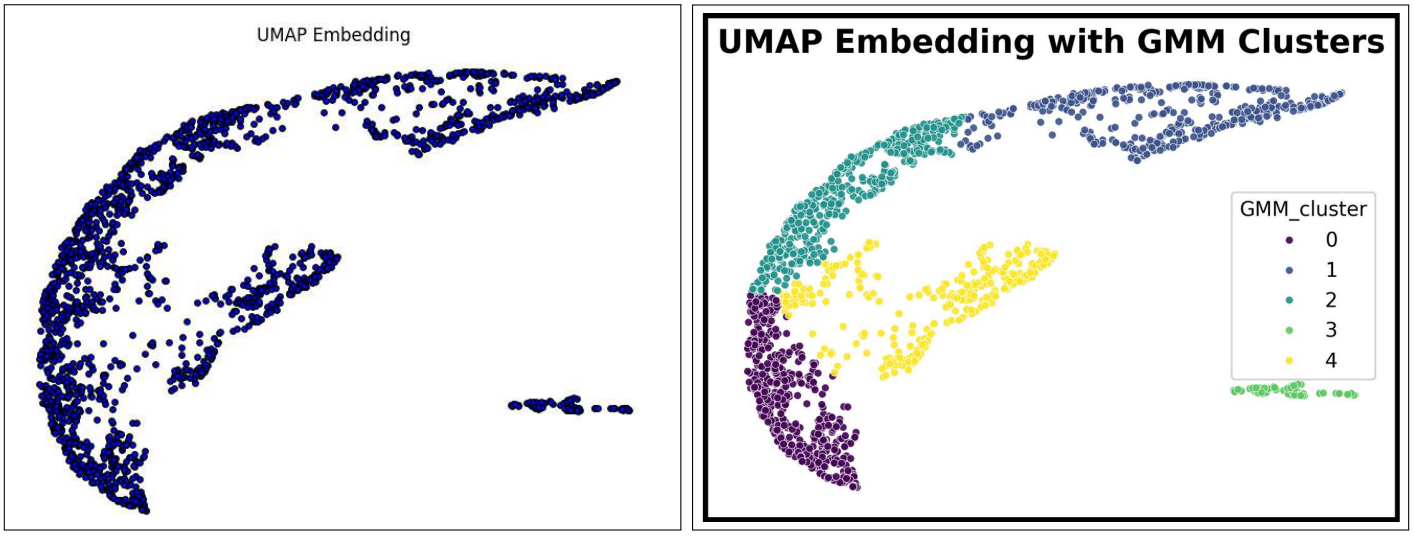
Figure B.8: Connectivity graph of UMAP for 16 ms light curves in two energy bands without PCA

Chapter C

Analysis with a different length of light curve

All the analyses discussed in the previous section were done with light curves extracted in the range zero to t90 s, it would be an interesting thing to know how the results will dance around with a different length of light curves. The following analyses are carried out with 64 ms light curves extracted from 50s before the trigger to 250s after the trigger.

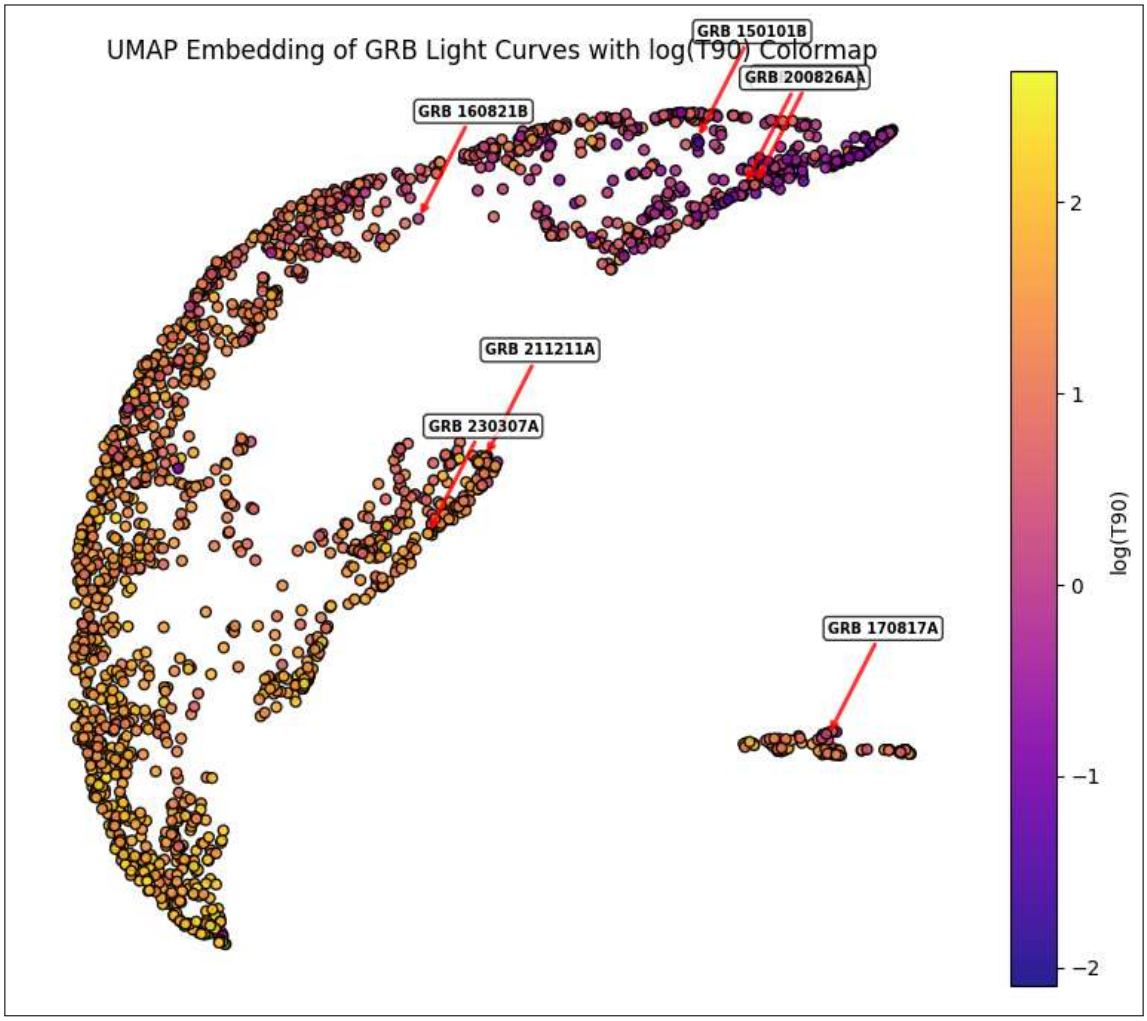
C.0.1 With PCA initialization for 50–300 Kev band



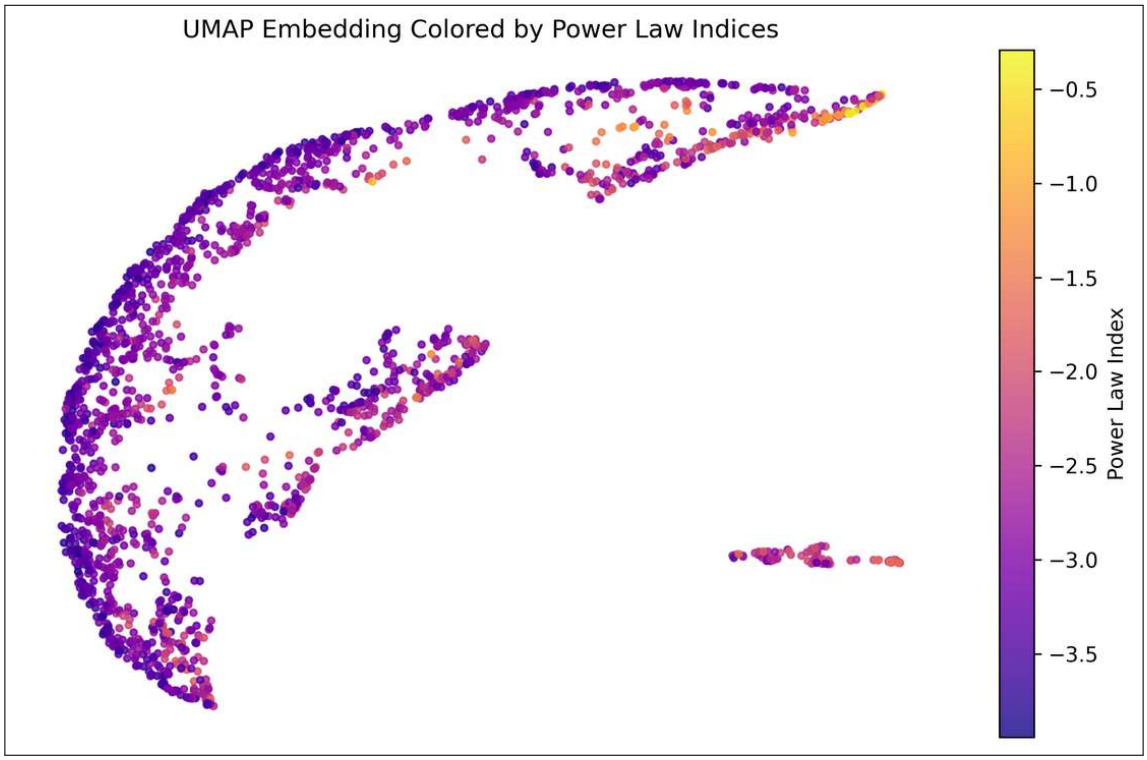
(a) UMAP embedding

(b) GMM clusters

Figure C.1: Analyses for 64 ms light curve in 50–300 Kev band with PCA initialization: UMAP embedding and the corresponding HDBSCAN embedding.



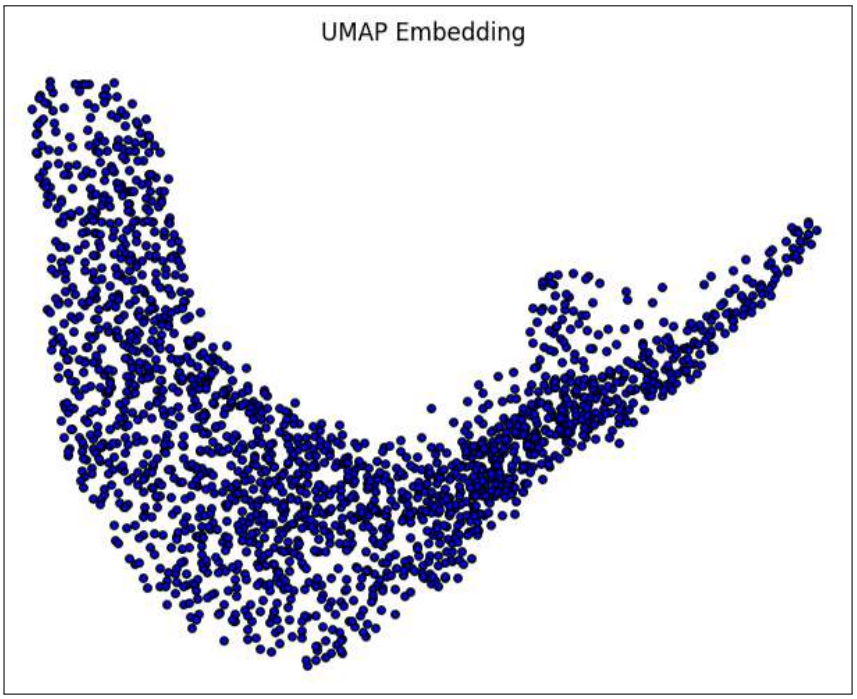
(c) Duration map of UMAP



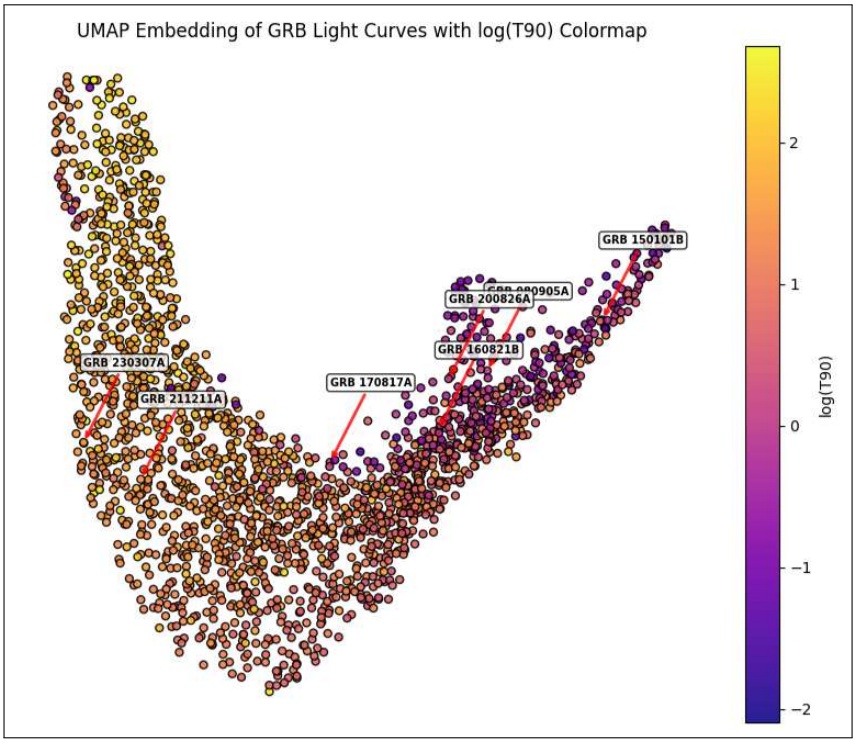
(d) Power-index map

Figure C.2: Duration map and power index map of embedding

C.0.2 Without PCA initialization for 50–300 Kev band

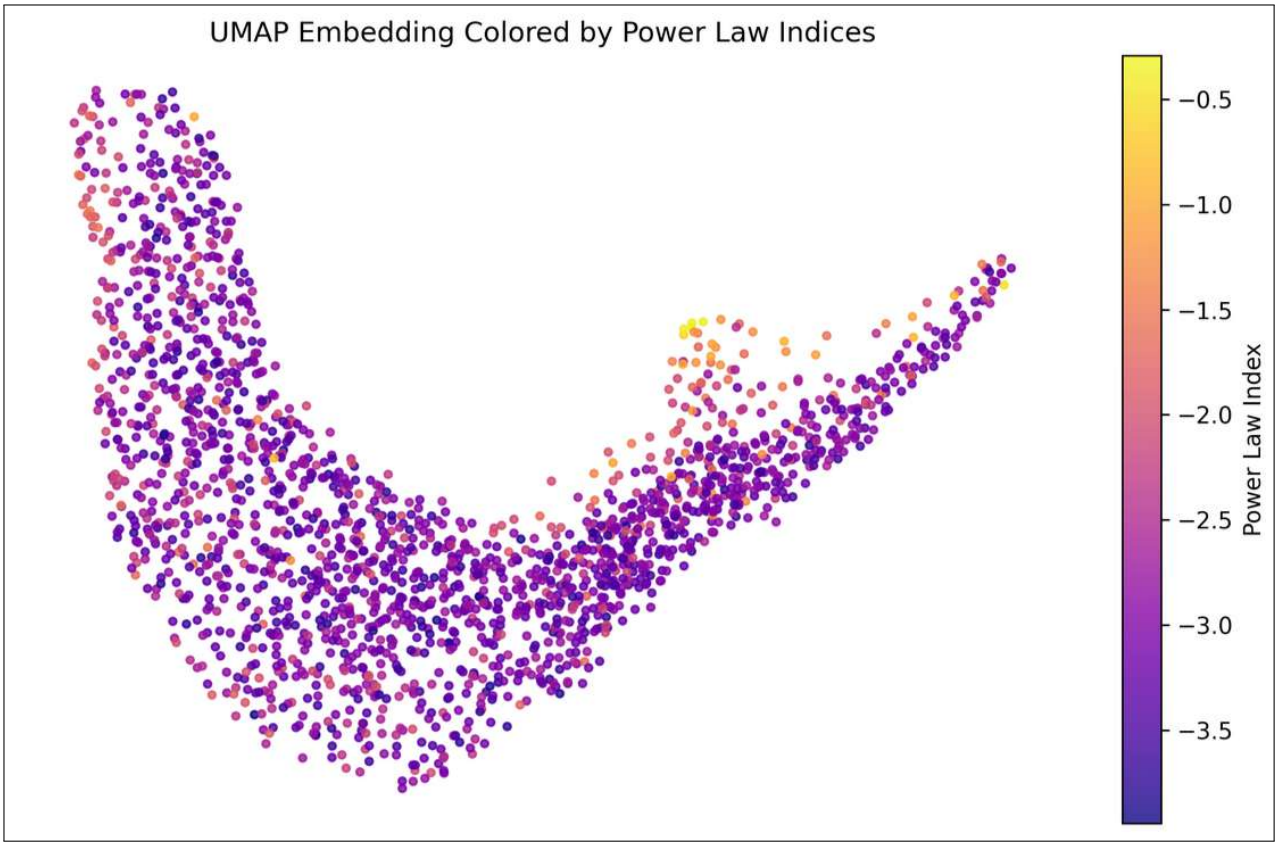


(a) UMAP embedding



(b) Duration map of UMAP embedding

Figure C.3: UMAP and duration map for 64 ms light curve in 50–300 Kev band without PCA initialization



(a) Power-index map of embedding

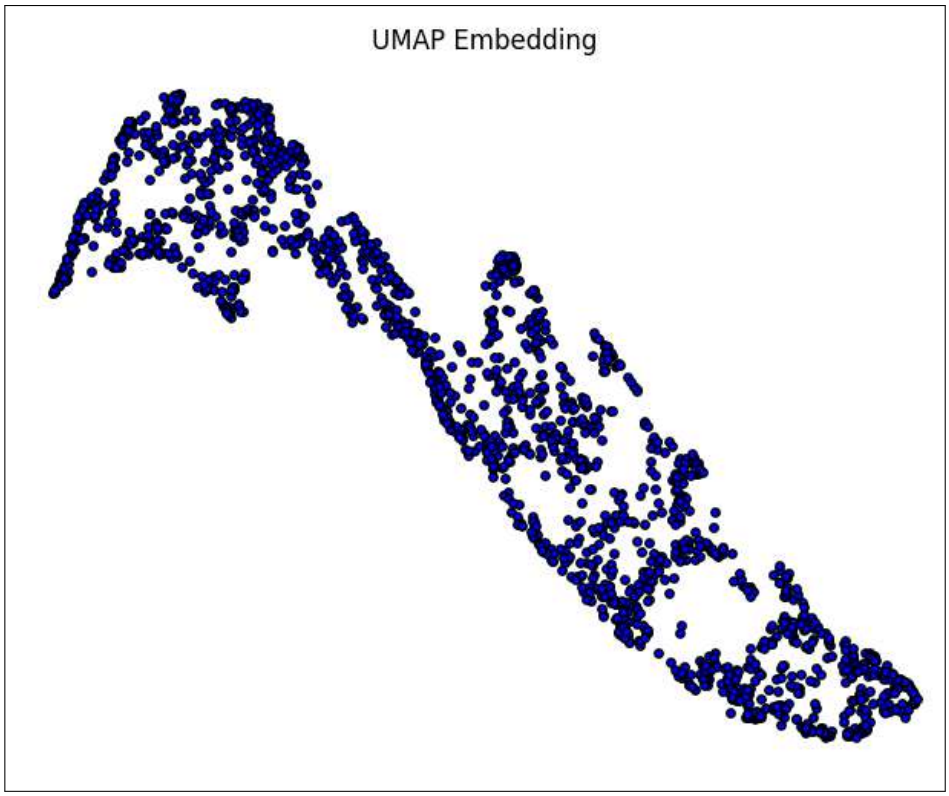
Figure C.4: Power-index map for 64 ms light curve in 50–300 Kev band without PCA initialization

From Figure C.3 and Figure C.4, it is evident that for the analyses of 64 ms light curves in the energy band 50–300 Kev, with and without PCA initialization, the process of clustering is heavily dependent upon duration and noise parameters.

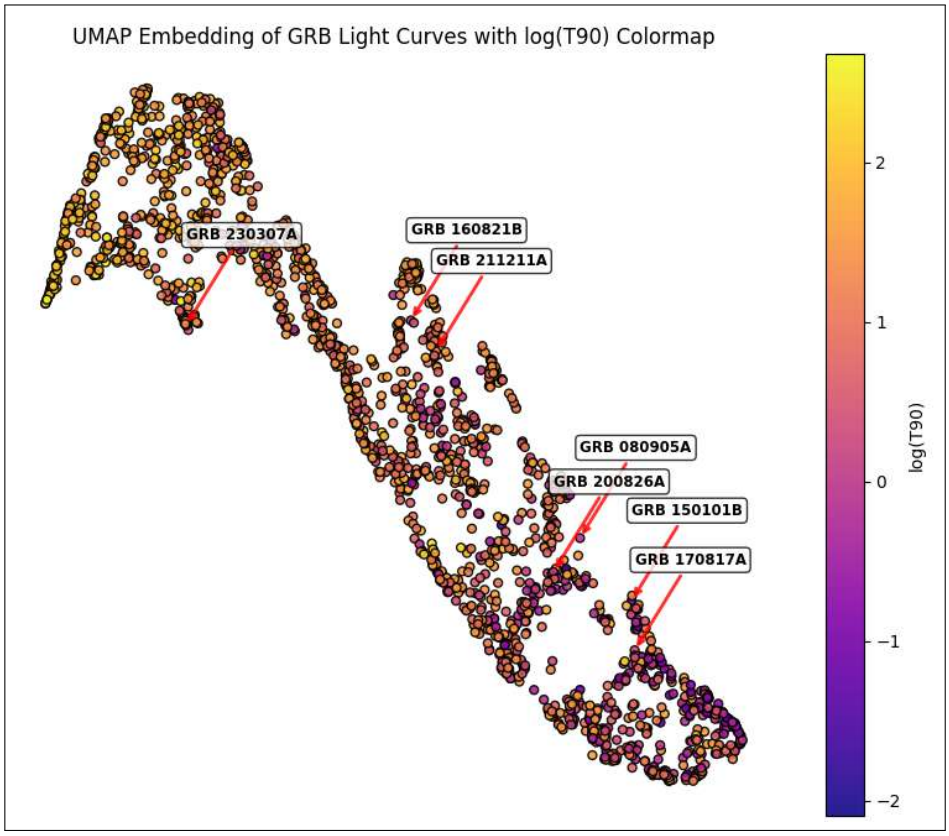
There is a clear gradient in duration seen in both the duration maps, Figure C.3 and Figure C.2. In both the duration maps, short GRBs are distinctly forming a separate group. The reason for this particular positioning of short GRBs can be found from the power-index map of these embeddings. It is clear from Figures C.2 and Figure C.4 that the distinction is due to the dominance of white noise present in the background of the short GRB light curves.

The results here are consistent with the previous ones using both 16 ms and 64 ms light curves extracted to a particular length based on their t90 values.

C.0.3 With and without PCA initialization for light curves in three energy bands

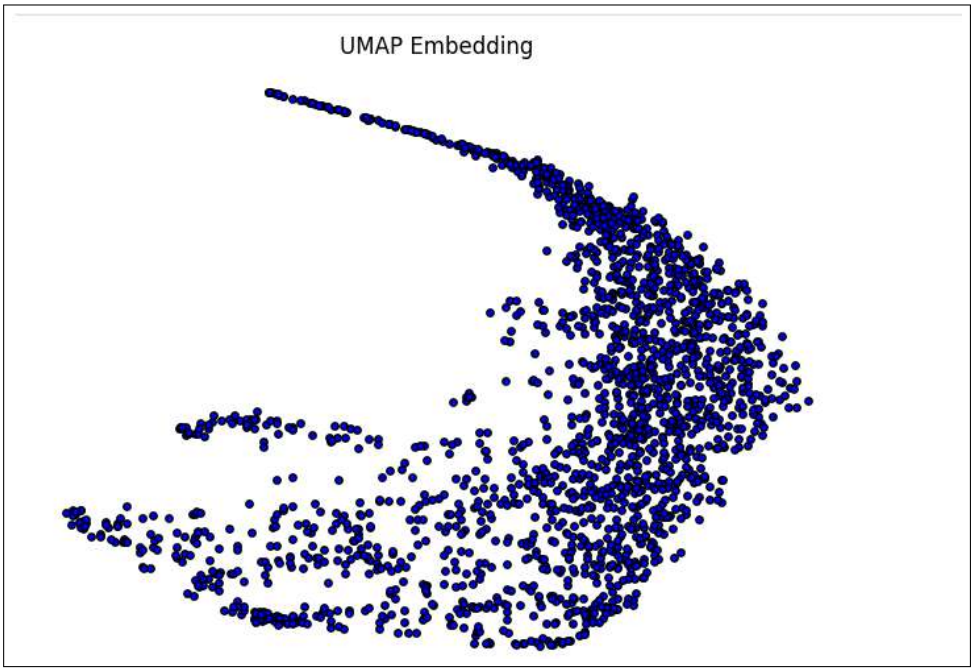


(a) 64 ms light curves in three energy bands with PCA

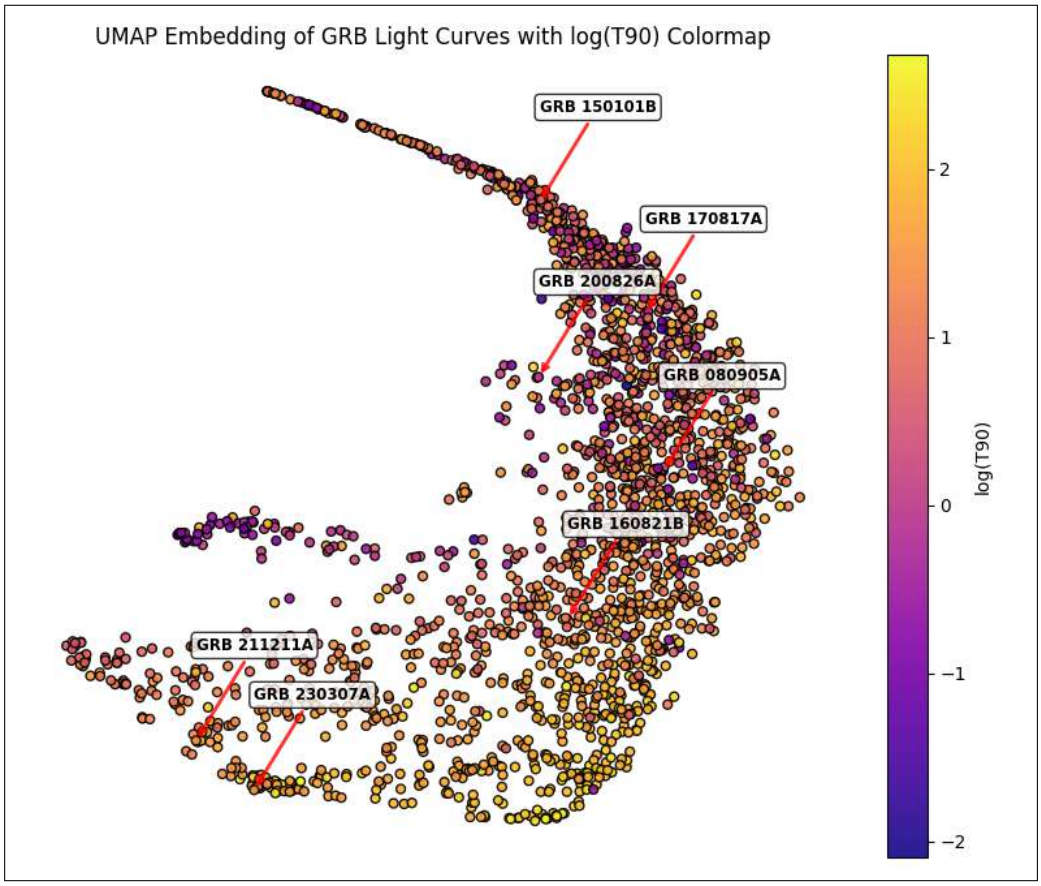


(b) With duration color mapping (PCA)

Figure C.5: UMAP embedding and duration map of 64 ms light curves in three energy bands with PCA initialization



(a) 64 ms light curves in three energy bands without PCA



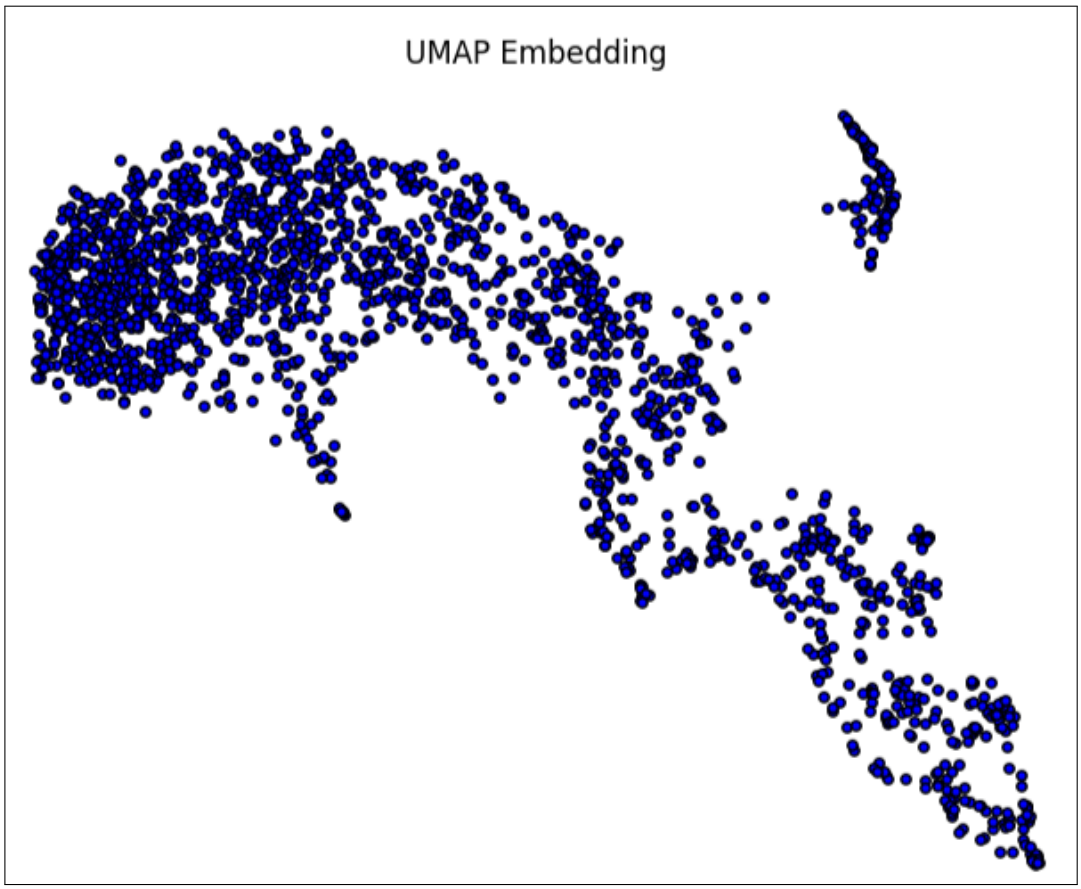
(b) With duration color mapping (No PCA)

Figure C.6: UMAP embedding and duration color map of 64 ms light curves in three energy bands without PCA initialization

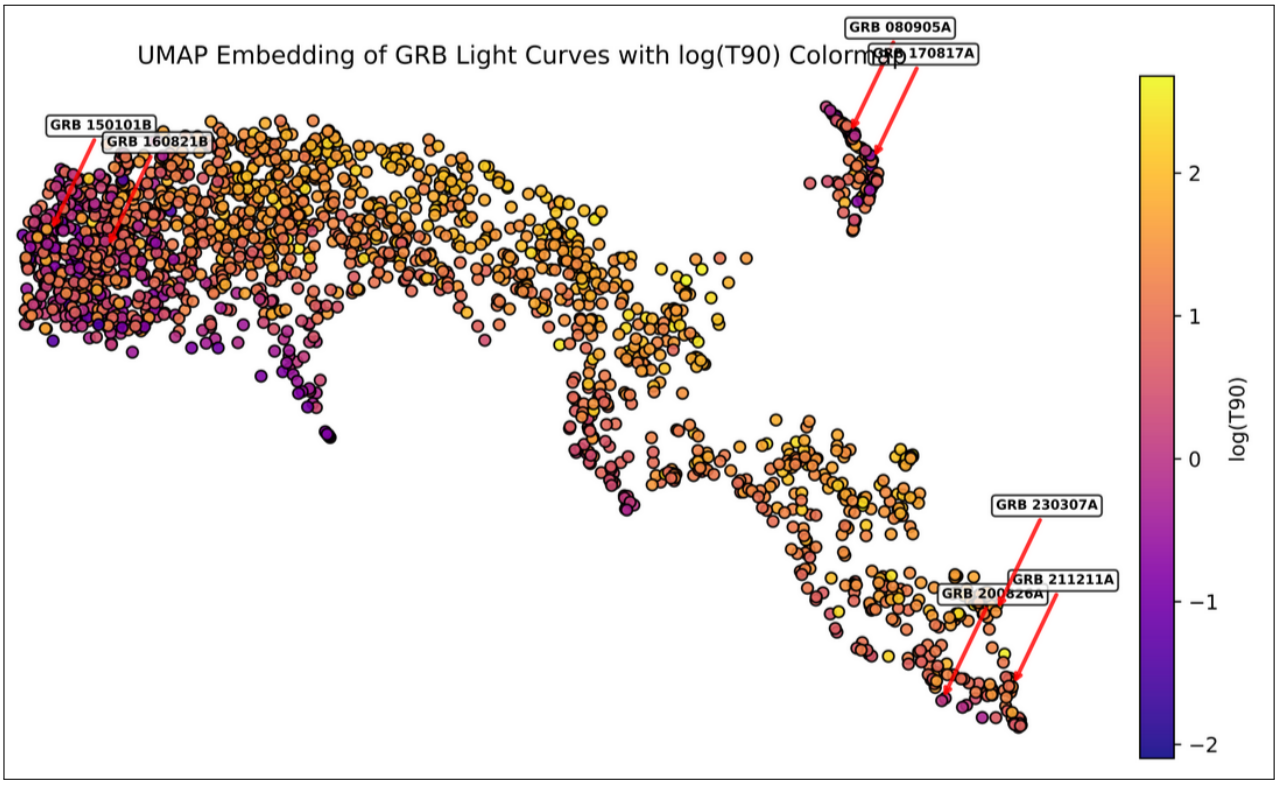
With the concatenation of three energy bands, the distinction based on duration is less evident, although it is still faintly present.

Now we can see how far the results could shift if we do not denoise our light curves.

C.0.4 Analyses with raw light curves which are not denoised

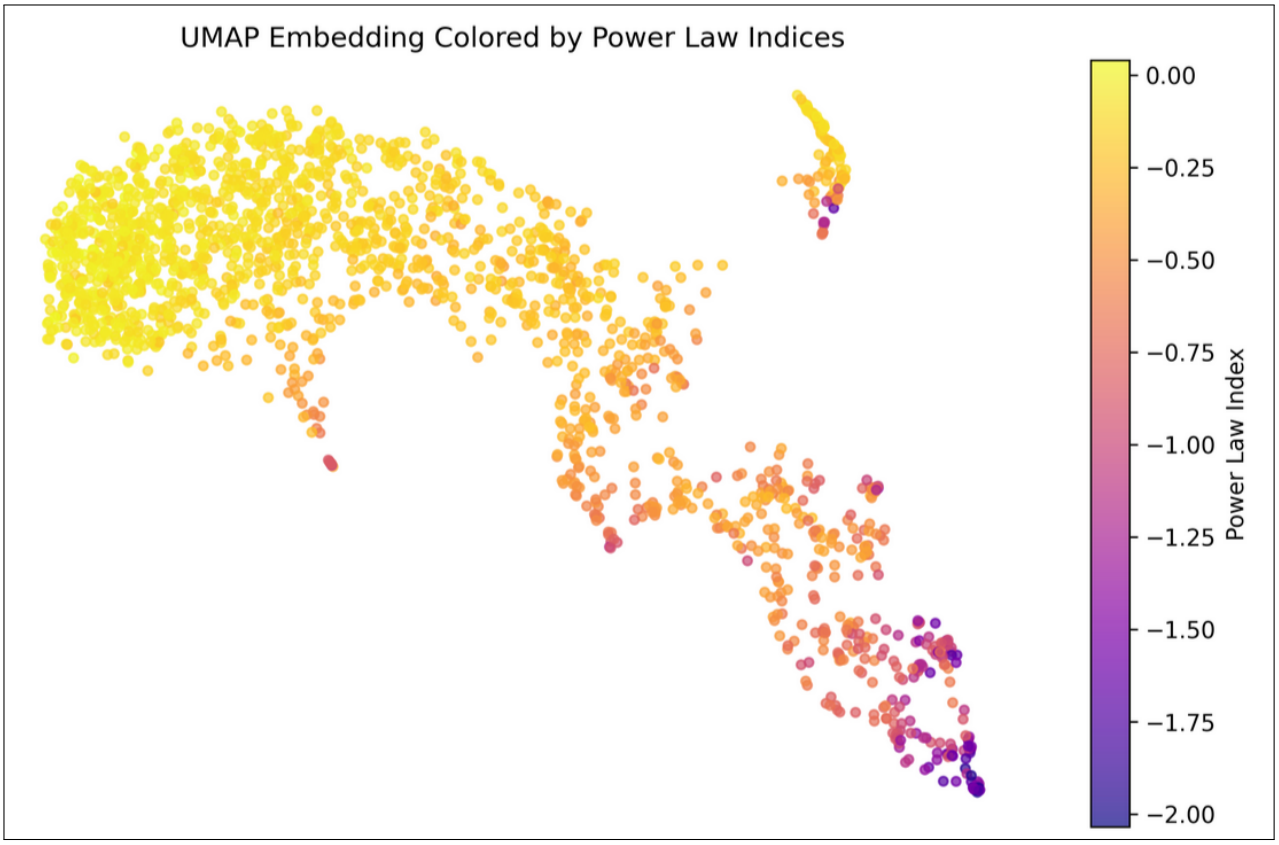


(a) UMAP embedding of light curves which are not denoised



(b) Duration map of UMAP embedding

Figure C.7: UMAP and duration map of 16 ms light curves without denoising



(a) Power-index map of embedding showing dominance of white noise (yellow).

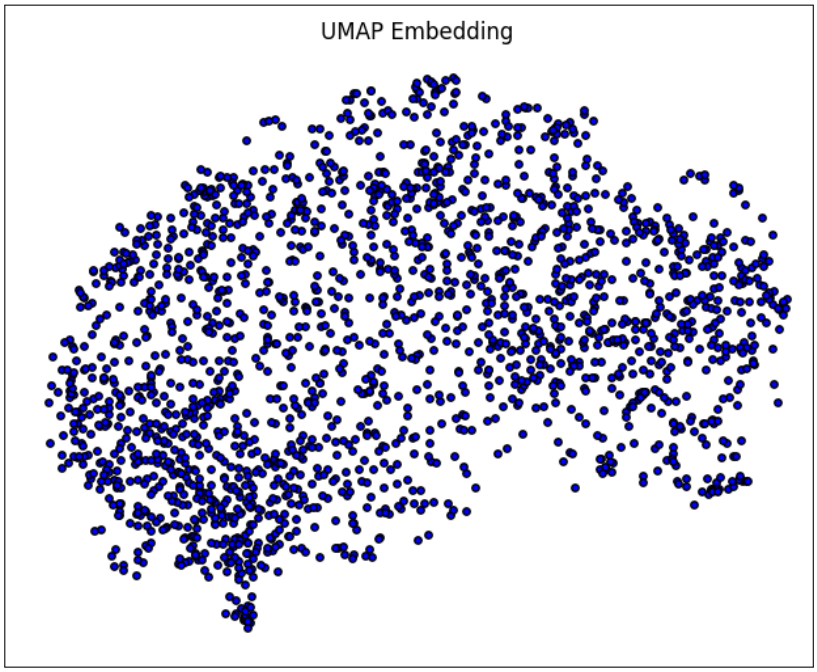
Figure C.8: Power-index map for 16 ms light curves without denoising

It is evident that without the application of wavelet denoising, light curves are dominated by white noise and the process clustering depends on this noise parameter while structuring the embedding. Short GRBs in the embedding occupy two different regions, but the interesting thing is both these locations are predominantly white noise only. Therefore the reason for the separation of short GRBs is unknown here!, which contradicts our previous results which consistently rendered the positioning of short GRBs as due to the white noise.

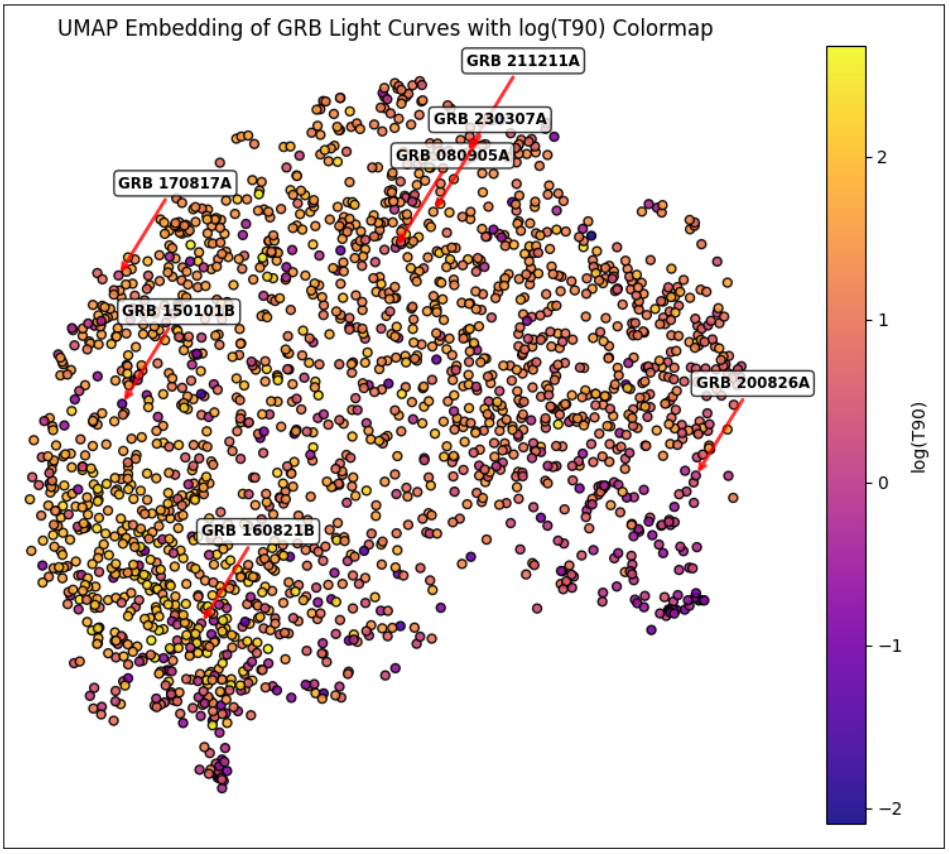
Another interesting thing is some of the intermediate-duration GRBs are dominated by red noise in their light curves and as a result their grouping can be seen at the bottom right.

Another striking thing here is that the noise in the light curves imparted by the wavelet denoising is of no longer gaussian, on the other side it is of steeper power indices ranging from 2 to 4, and this is the noise component which drives the clustering to some extend, especially in case of short GRBs.

C.0.5 Analysis with Light curves as direct input vectors and not Fourier Amplitudes of the light curves

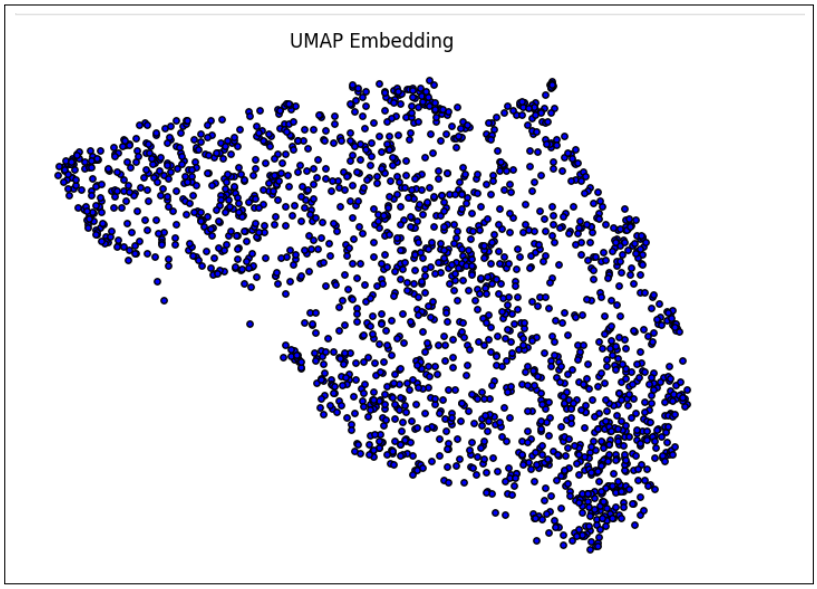


(a) 64 ms light curves in three energy bands without PCA using light curves as input vectors

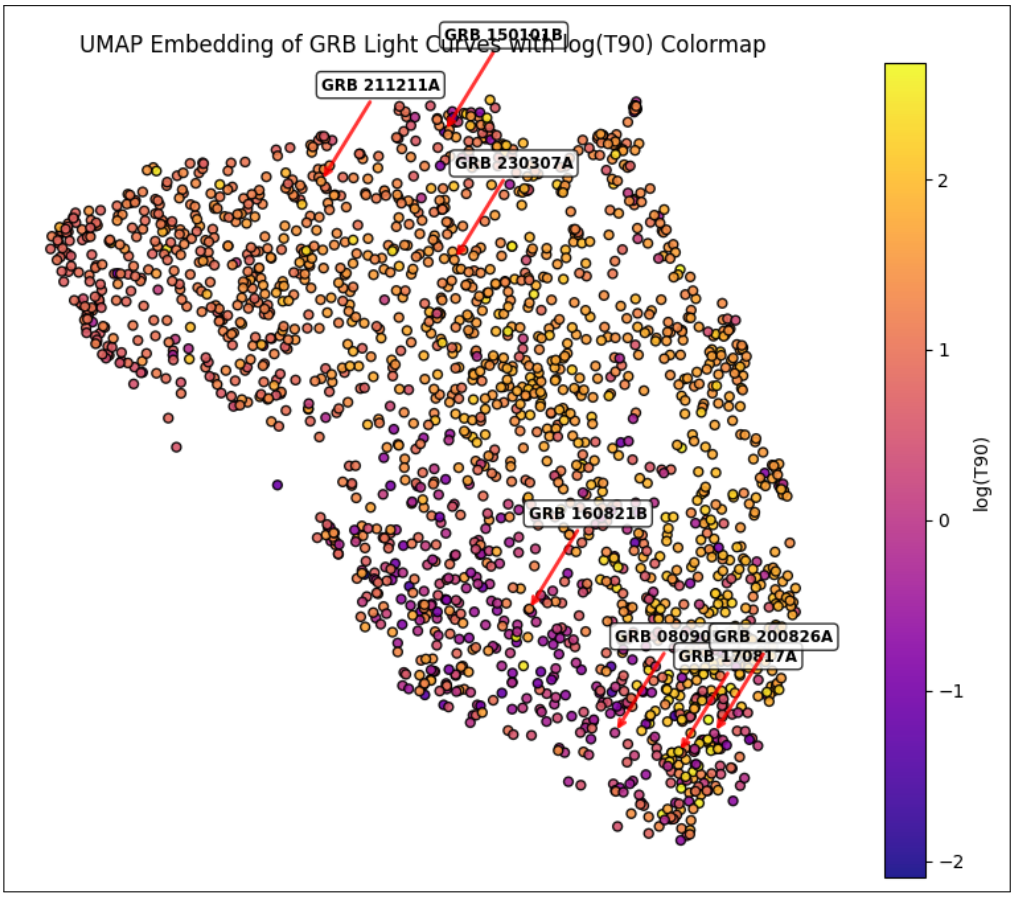


(b) Duration color mapping for above embedding

Figure C.9: UMAP embedding and color map of 64 ms light curves in three energy bands without PCA, using light curves as input vectors.



(a) 64 ms light curves in three energy bands with PCA using light curves as input vectors

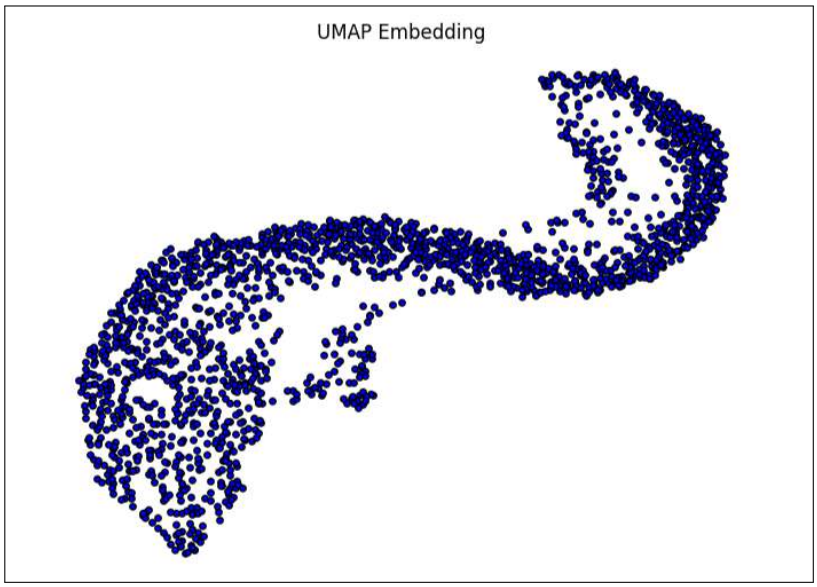


(b) Duration color mapping for above embedding

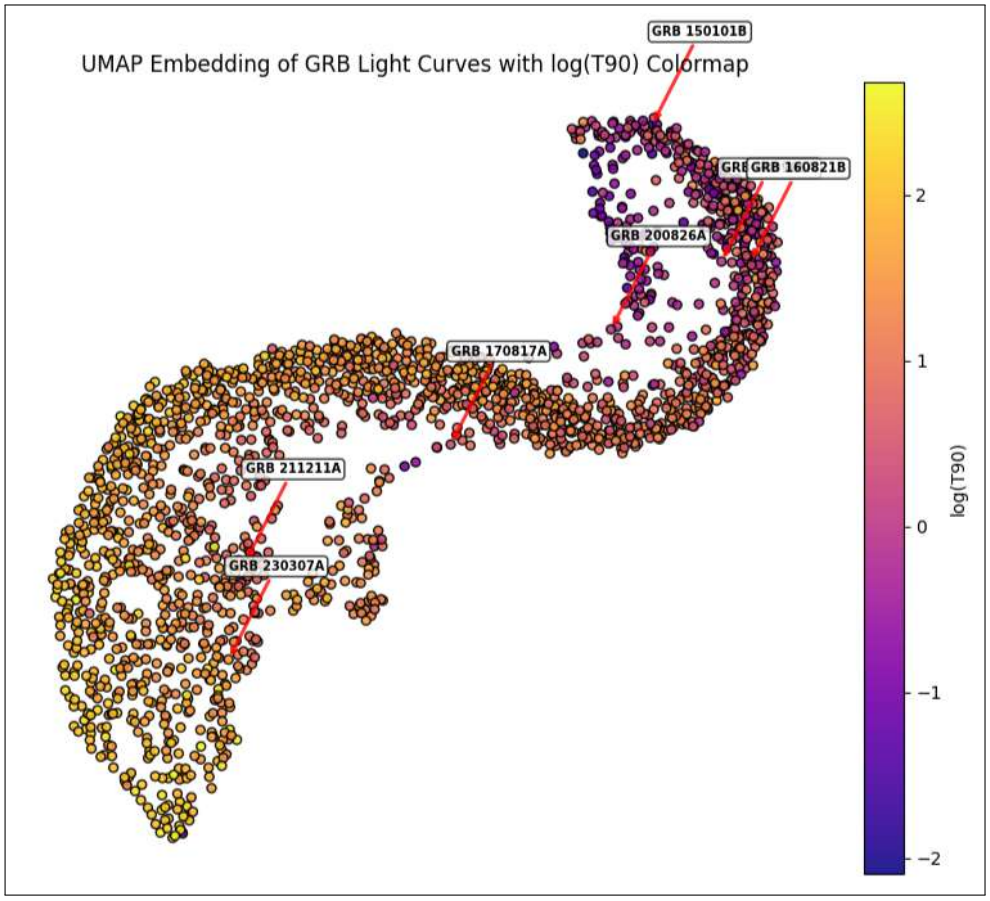
Figure C.10: UMAP embeddings of 64 ms light curves in three energy bands with PCA, using light curves as input vectors. Even though a distinction between short GRBs and long GRBs is visible, clusters are inseparable which makes the use of a clustering algorithm redundant here.

C.1 16 ms light curve analyses

C.1.1 With PCA initialization for 50–300 KeV band

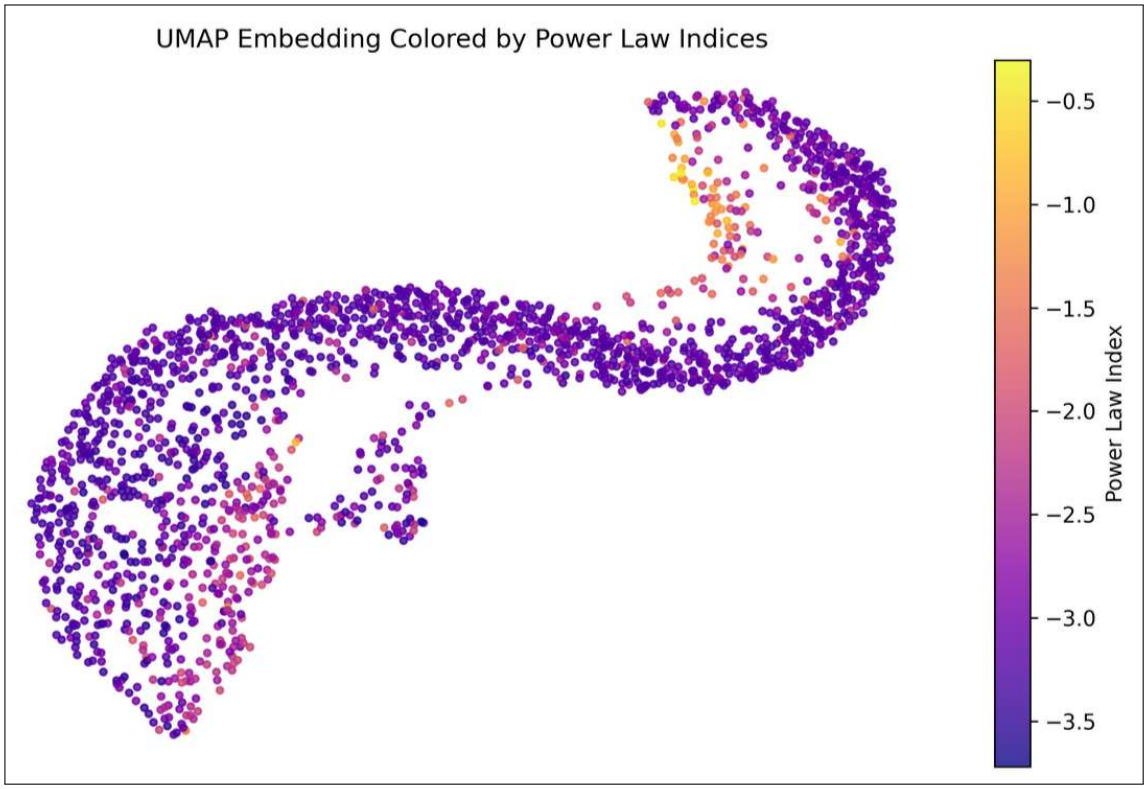


(a) UMAP embedding



(b) Duration map of UMAP embedding showing a clear distinction of short and long GRBs.

Figure C.11: UMAP embedding and duration map for 16 ms light curves in 50-300 KeV band with PCA initialization for a different length of light curve.

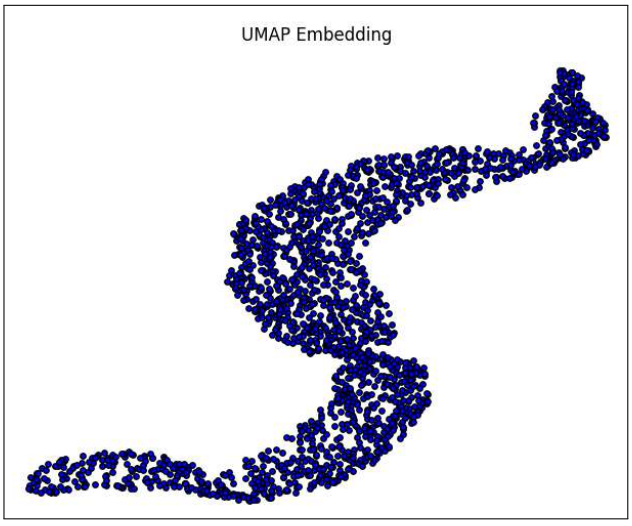


(a) Power-index map of embedding

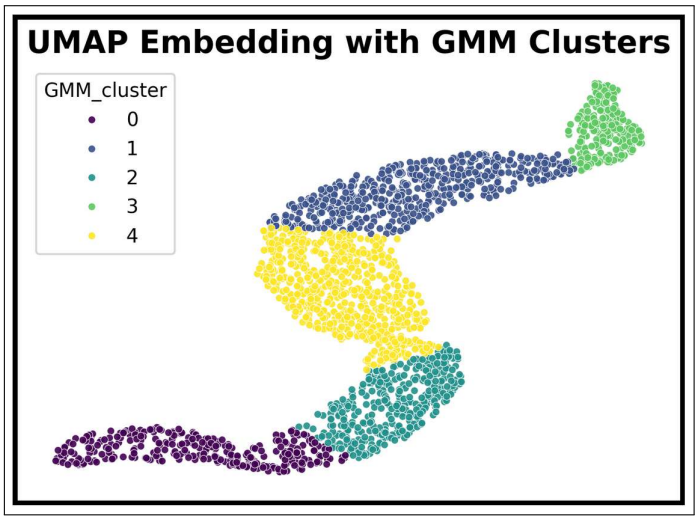
Figure C.12: Power-index map for 16 ms light curves in 50-300 KeV band with PCA initialization

Separation based on the duration is clearly evident from the gradient of long GRBs to short GRBs in Figure C.11b. Separation of a certain group of short and long GRBs because of different power-index is also visible in Figure C.12a.

C.1.2 With PCA initialization for Light curves in three energy bands

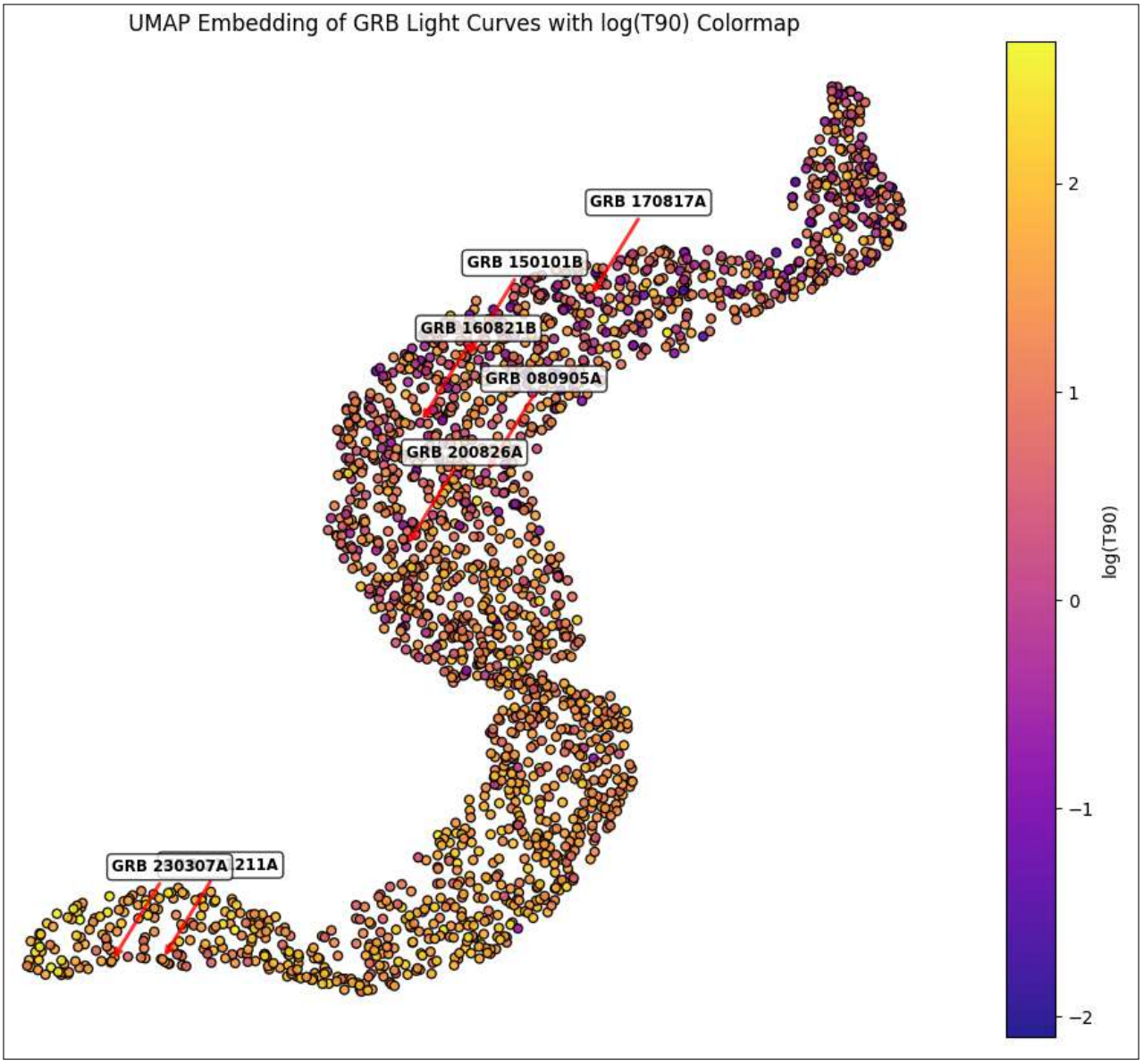


(a) UMAP embedding



(b) Clusters identified by GMM

Figure C.13: UMAP embedding and GMM clustering for 16 ms light curves in three energy bands with PCA initialization for a different length of light curves



(a) Duration map of embedding

Figure C.14: Duration map of UMAP embedding for 64 ms light curves in three energy bands with PCA initialization for a different light curve length. Some of the special GRBs are annotated. There is a clear distinction for GRBs in case of ultra-long ones. However, short GRBs are not distinctly separated, clusters being not well defined, the top right region is a mix of long and short GRBs

Here the embedding is clustered using GMM because HDBSCAN was not able to perform well on the embedding due to the indistinct nature of the clusters. This could be a result of the uniformity in their noise characteristics. However, the positioning of GRBs is clearly based on the duration, as is evident from the duration map (Figure C.14a).

Bibliography

[1] Abbott B. P., et al., 2017, , 848, L13

[2] Acuner Z., Ryde F., 2017, Monthly Notices of the Royal Astronomical Society, 475, 1708

[3] Belczynski K., et al., 2018, arXiv e-prints, p. arXiv:1812.10065

[4] Cai T. T., Ma R., 2021, arXiv e-prints, p. arXiv:2105.07536

[5] Chattopadhyay S., Maitra R., 2017, Monthly Notices of the Royal Astronomical Society, 469, 3374

[6] Dimple Misra K., Arun K. G., 2023, Astrophysical Journal Letters, 949, L22

[7] Dimple Misra K., Arun K. G., 2024, , 974, 55

[8] Emmanoulopoulos D., McHardy I. M., Papadakis I. E., 2013, , 433, 907

[9] Fotopoulou S., 2024, Astronomy and Computing, 48, 100851

[10] Gao J., Hu W., Chen Y., 2024, arXiv e-prints, p. arXiv:2412.19423

[11] Garcia-Cifuentes K., Becerra R. L., De Colle F., Cabrera J. I., Del Burgo C., 2023, , 951, 4

[12] Ghojogh B., Ghodsi A., Karray F., Crowley M., 2020, arXiv e-prints, p. arXiv:2011.10925

[13] Jespersen C. K., Severin J. B., Steinhardt C. L., Vinther J., Fynbo J. P. U., Selsing J., Watson D., 2020, , 896, L20

[14] Jolliffe I. T., Cadima J., 2016, Philosophical Transactions of the Royal Society of London Series A, 374, 20150202

[15] Katz J. I., 1994, , 432, L107

[16] Kouveliotou C., Meegan C. A., Fishman G. J., et al., 1993, Astrophysical Journal Letters, 413, L101

[17] Malzer C., Baum M., 2019, arXiv e-prints, p. arXiv:1911.02282

[18] McInnes L., Healy J., Melville J., 2018, arXiv e-prints, p. arXiv:1802.03426

[19] Mehrbani E., Kahaei M. H., 2021, arXiv e-prints, p. arXiv:2103.04060

[20] Mickaelian A. M., 2020, Communications of the Byurakan Astrophysical Observatory, 67, 159

[21] Paczynski B., 1986, , 308, L43

[22] Rastinejad J. C., et al., 2022, , 612, 223

[23] Ren B., Pueyo L., Zhu G. B., Debes J., Duchêne G., 2018, , 852, 104

[24] Sajadian S., Fatheddin H., 2023, , 166, 252

[25] Shlens J., 2014, arXiv e-prints, p. arXiv:1404.2986

[26] Stanek K. Z., et al., 2003, , 591, L17

[27] Steinhardt C. L., Mann W. J., Rusakov V., Jespersen C. K., 2023, , 945, 67

[28] Timmer J., König M., 1995, , 300, 707

[29] Wang Q., 2012, arXiv e-prints, p. arXiv:1207.3538

[30] Woosley S. E., 1993, , 405, 273

[31] YongchangWang Zhu L., 2017, in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). pp 471–475, doi:10.1109/ICIS.2017.7960038

[32] Zhang B.-B., et al., 2021, Nature Astronomy, 5, 911