

DRAFT VERSION JUNE 3, 2026
Typeset using L^AT_EX preprint style in AASTeX7.0.1

Multimodal Transformer Based Generic Mixture Density Network for Scattering Timescale Estimation of Fast Radio Bursts

BIKASH KHAREL ^{1,2} EMMANUEL FONSECA ^{1,2} SRINJOY DAS ³ MASON NG ^{4,5}
PAUL SCHOLZ ⁶ MAWSON W. SAMMONS ⁷ LORDRICK KAHINGA ^{8,9} AND AFROKK KHAN ^{4,5}

¹*Department of Physics and Astronomy, West Virginia University, PO Box 6315, Morgantown, WV 26506, USA*

²*Center for Gravitational Waves and Cosmology, West Virginia University, Chestnut Ridge Research Building, Morgantown, WV 26505, USA*

³*School of Mathematical and Data Sciences, West Virginia University, PO Box 6310, Morgantown, WV 26506, USA*

⁴*Department of Physics, McGill University, 3600 rue University, Montréal, QC H3A 2T8, Canada*

⁵*Trottier Space Institute, McGill University, 3550 rue University, Montréal, QC H3A 2A7, Canada*

⁶*Department of Physics and Astronomy, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada*

⁷*Laboratoire d'Astrophysique de Marseille, Aix-Marseille Univ., CNRS, CNES, Marseille, France*

⁸*Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95060, USA*

⁹*Department of Physics, College of Natural and Mathematical Sciences, University of Dodoma, 1 Benjamin Mkapa Road, 41218 Iyumbu, Dodoma 259, Tanzania*

ABSTRACT

The discovery rate of fast radio bursts (FRBs) continues to increase with the advent of new radio facilities and yet extracting their astrophysical parameters such as scattering timescale (τ) remains a significant bottleneck. Current τ measurement approaches like fitting analytic template models and scattering aware de-convolution are accurate but slow, sensitive to initialization, limited by low signal to noise and often require manual supervision. These limitations inspired us to explore fast, robust and scalable machine learning methods to estimate the astrophysical parameter value. We present a deep learning approach named Multimodal Transformer Based Generic Mixture Density Network (MT-GMDN) which ingests FRB dynamic spectrum and its corresponding timeseries profile through parallel transformer encoders, fuses their latent representations and predicts the distribution of τ with probabilistic output derived from generic mixture-density formulation. This formulation not only estimates the value of τ but also captures the (zero inflated) nature of FRB populations where a significant fraction of bursts exhibit unresolvable scattering. We trained MT-GMDN on ~ 3500 FRBs from CHIME/FRB Catalog 2 while holding out some fraction of FRBs for validation during training and for testing after the training completes. The model achieves a coefficient of determination (R^2) value of 94% on the expected value of τ for the events with measurable scattering with an excellent recall value of 90% on the test data set. The

Corresponding author: Bikash Kharel
bk0055@mix.wvu.edu

model was also able to incorporate heteroskedastic errors enabling us the construction of a confidence interval for the predictions.

Keywords: Fast radio bursts — Interstellar scattering — Neural networks — Transformers — Mixture density

1. INTRODUCTION

Fast radio bursts (FRBs) are powerful extragalactic astronomical transients of millisecond durations (D. R. Lorimer et al. 2007) and their observed morphology is strongly determined by propagation effects through ionized interstellar and intergalactic media (e.g., J. M. Cordes & M. A. McLaughlin 2003). One of the key propagation signatures is a frequency dependent pulse dispersion and temporal broadening due to the multi-path scattering of the radio signal which is often modeled as a one-sided exponential tail with characteristic scattering timescale (τ) (I. P. Williamson 1974). Figure 1 displays two simulated FRBs. The first (left) shows a sharp Gaussian peak with no scattering while the second (right) exhibits an exponentially decaying tail demonstrating scattering.

Electromagnetic radiation from an FRB source traveling through a turbulent interstellar medium (ISM) with inhomogeneous electron density distribution experiences a frequency dependent phase shift. The “thin screen” model formulates the observed intensity of the scattered radiation as a result of the phase shift to be an impulse response function (IRF) of the form (D. R. Lorimer & M. Kramer 2004):

$$h(t) \propto \exp\left(-\frac{\Delta t}{\tau}\right), \quad (1)$$

where,

- Δt : Time delay between scattered and non-scattered rays and

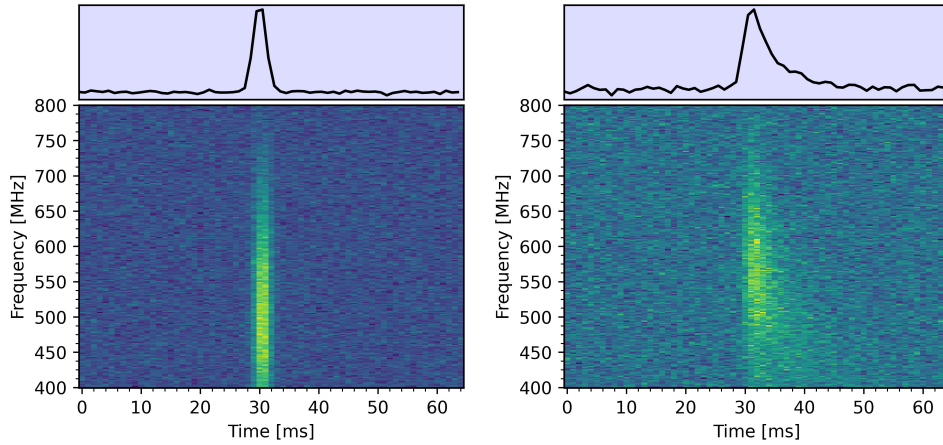


Figure 1. Simulated FRB dynamic spectra with corresponding timeseries at the top of respective dynamic spectrum. The left panel shows an FRB without scattering characterized by a symmetric Gaussian pulse profile, while the right panel shows a scattered FRB with asymmetric pulse profile characterized by exponentially decaying tail.

- τ : Characteristic scattering timescale of the ISM. This scattering timescale exhibits frequency dependence as $\tau \propto f^{-4}$

The resulting observed pulse profile of an FRB signal is thus the convolution of its intrinsic profile with the scattering exponential kernel, i.e.,

$$I_{obs}(t) = I_{int}(t) * h(t) \quad (2)$$

The measurement of τ carries significant astrophysical implications such as probing electron density fluctuations along FRB sight lines which thereby constrain the degree of plasma turbulence within host-galaxy environments (e.g., S. Xu & B. Zhang 2016; K. Shin et al. 2025), diffused halos of intervening galaxies (e.g., S. K. Ocker et al. 2021) and the large scale ionized structure of the Milky Way (e.g., S. Shivraj Patil et al. 2025) and constraining redshift (e.g., J. M. Cordes et al. 2022) and the Hubble constant (e.g., Yang, Tsung-Ching et al. 2025).

There is a rapid rise in the detections of FRBs with commensal real-time surveys on radio telescopes (e.g., CHIME/FRB Collaboration et al. 2018; J.-P. Macquart et al. 2010; J. Jonas & MeerKAT Team 2016; M. Sherman & DSA-110 Collaboration 2023). The current detection rate is $O(10^3 \text{ FRBs yr}^{-1})$, which is much larger compared to other astronomical radio transient phenomena. This data volume and diversity of the FRBs in morphology, observational instrument, and propagation effects pose significant challenges for extracting physical parameters such as τ from the detected FRBs.

Common scattering measurement tools, such as `fitburst` (E. Fonseca et al. 2024), fit parametric templates to the de-dispersed dynamic spectrum using regression methods. While these techniques can often be accurate, they face significant limitations. Parametric template fitting relies heavily on an adequate initial guess of the parameters and is sensitive to both low signal to noise and complex burst morphologies, thereby often failing to converge. Alternative approaches to estimating scattering timescale include Markov Chain Monte Carlo (MCMC) methods (e.g., CHIME/FRB Collaboration et al. 2019) and Fourier Transform based deconvolution methods (e.g., N. D. R. Bhat et al. 2003). However, these methods are computationally slow, require extensive per sample manual intervention and thus are not feasible for analyzing large scale datasets.

This work mainly relies on the data set from CHIME/FRB Catalog 2 (Chime/FRB Collaboration et al. 2026), the largest catalog of unique FRBs till date, which includes 4545 FRBs, comprising 3564 one-off events and 981 repeating bursts from 83 different sources. The CHIME/FRB Catalog 2 data set consists of radiation in the frequency range of 400-800 MHz encoded as dynamic spectrum into 16,384 frequency bins across 160 time samples of temporal resolution 0.983 ms.

The applications of deep learning in FRB science have been primarily focused on classification (e.g., L. Connor & J. van Leeuwen 2018; B. Kharel et al. 2025) and detection (e.g., D. Agarwal et al. 2020). As a novel approach in FRB science, to address the issues with methods of scattering timescale estimation discussed in the above paragraph, we present a deep probabilistic model which estimates the distribution of τ from de-dispersed dynamic spectra and their corresponding pulse profiles. Instead of treating scattering estimation as deterministic regression, we model the conditional distribution $p(\tau|\mathbf{x})$, where $\mathbf{x}$ is the burst's dynamic spectrum. This distribution approximation yields a well-calibrated heteroskedastic¹⁰ uncertainty related to noise associated with the dynamic spectrum containing an FRB signal, allowing one to construct a confidence interval on point estimates.

¹⁰ Uncertainty parameter is data-dependent.

The deep probabilistic framework not only estimates the distribution of τ , but also quantifies the Bernoulli probability of $\tau \geq 0$ as more than 50% of the FRBs have unresolved τ in CHIME/FRB Catalog 2.

We organize our work in this paper as follows. In Section 2 we present the methodology underlying our approach which involves probabilistic formulation of scattering timescale estimation, deep learning architecture used to model the formulated distribution and data sources with preprocessing steps applied. Section 3 presents training details and results which includes evaluation performed on real observations from CHIME/FRB Catalog 2 through comparison with `fitburst`, quantitative validation using simulated data, benchmarking our approach against powerful MCMC based inference and investigation of the robustness of our method under different noise conditions. In Section 4 we discuss the implications of our findings and in Section 5 we outline limitations of our work and directions for future research. Finally, we conclude our work in Section 6.

2. METHODS

2.1. Mixture Density Network Formulation

Based on the probabilistic interpretation of regression tasks, it is usually assumed that a target t (a variable that a regression methods seek to predict) for given data point $\mathbf{x}$ follows a normal distribution

$$p(t|\mathbf{x}) \sim \mathcal{N}(t|\mu(\mathbf{x}), \sigma^2).$$

Where $\mu(\mathbf{x})$ is the mean of the target distribution and usually taken as point estimate while σ is uncertainty parameter or variance of the target. Here the uncertainty in the target is homoskedastic¹¹. The regression is then performed by maximizing the likelihood associated with the training dataset. This is written as

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N \mathcal{N}(t_i|y(x_i, \theta), \sigma^2),$$

where, $y(x_i, \theta)$ is the linear combination of some linear basis functions or neural network with parameters θ . This formulation is overly generic and assumes the distribution of targets to be a normal distribution which may not be always true for a required specific problem domain.

Scattering timescales in FRBs are strictly non-negative (i.e. $\tau \geq 0$) with the vast majority of FRBs exhibiting minimal scattering with a significantly smaller population displaying large scattering tail. Also, the standard error associated with exponential distributions as in Equation 1 is proportional to the measured value. Given the wide range of observed τ for FRBs, it becomes more appropriate to assume the distribution of scattering timescales to follow lognormal distribution instead of normal distribution. Figure 2 demonstrates that the scattering timescales from CHIME/FRB Catalog 2 follow a lognormal distribution. Thus, the distribution of τ for a given dynamic spectrum $\mathbf{x}$ of an FRB event can be expressed as

$$p(\tau|\mathbf{x}) = \text{Lognormal}(\tau|\mu(\mathbf{x}), \sigma^2(\mathbf{x})) \equiv f_{\text{LN}}(\tau|\mu(\mathbf{x}), \sigma^2(\mathbf{x})), \quad (3)$$

where, μ and σ^2 are the location and scale parameters of the lognormal distribution, respectively. Here, the scale of the lognormal distribution is data dependent, i.e., heteroskedastic. This heteroskedastic uncertainty is expected to be larger for dynamic spectrum with low signal to (J. Taylor

¹¹ Variance of the target distribution is data independent

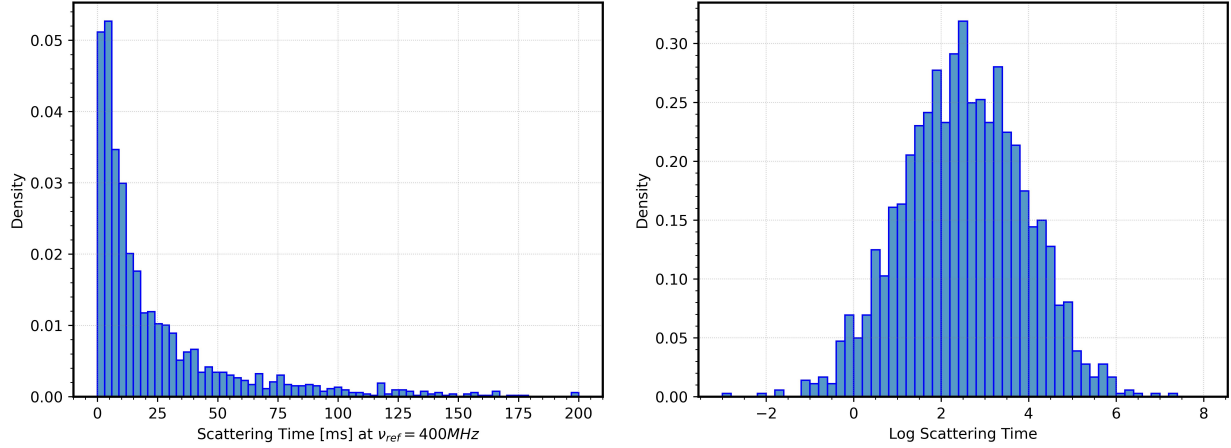


Figure 2. Left: The distribution of τ from CHIME/FRB Catalog 2 (at reference frequency 400 MHz) exhibiting highly skewed heavy tailed profile. Right: Corresponding log transformed ($\ln \tau$) distribution. The exhibition of Gaussian shape in the right panel verify that the scattering timescales follow a lognormal distribution.

1997) and is also expected to be scaled along with the scattering timescale as standard error of the expected value of exponential tail is directly proportional to the expected value. Therefore, our regression task now becomes maximizing the likelihood associated with the training dataset with the target (τ) distribution being lognormal. This regression task can be written as

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N f_{LN}(\tau_i | \mathbf{x}_i; g_{\theta}(\mathbf{x}_i)), \quad (4)$$

where, $g_{\theta}(\mathbf{x}_i)$ can be any arbitrary function which, in our case, is a neural network that approximates μ_i and σ_i .

We can now design and train a neural network to maximize the likelihood in Equation 4. After training, the network approximates the conditional distribution parameters μ_i and σ_i for any given input $\mathbf{x}_i$. However, a notable challenge arises due to CHIME/FRB Catalog 2 dataset's zero inflation i.e. a large fraction of the population with zero value of scattering timescale. More than 50% of the events in CHIME/FRB Catalog 2 contains FRBs whose scattering timescales are unresolved (i.e. $\tau \approx 0$), from both repeating. This problem demands standard strategies like discarding the data with unresolved τ or two staged neural network cascaded such that first predicts the presence of scattering and the second regressing the distribution of τ . Both of the methods are suboptimal as discarding significant fraction of dataset can introduce selection bias and cascading multistage network is computationally redundant and introduces additional complexity. Instead, we propose a generic mixture density network (GMDN) which can be written by modifying a mixture density network (C. M. Bishop & H. Bishop 2024) as

$$p(\tau | \mathbf{x}) = \sum_{k=1}^K \pi_k(\mathbf{x}) f_k(\tau | \mathbf{x}; g_k(\mathbf{x})), \quad (5)$$

where, $\pi_k(\mathbf{x})$ are mixing coefficients and $g_k(\mathbf{x})$ are some functions or neural networks.

In our case, if $p_0(\mathbf{x})$ denotes the probability of $\tau = 0$ and $\delta(\tau)$ the probability mass function at $\tau = 0$ then

$$\begin{aligned}\pi_1(\mathbf{x}) &= p_0(\mathbf{x}; g_\theta(\mathbf{x})) \quad \text{and} \quad f_1(\mathbf{x}) = \delta(\tau), \\ \pi_2(\mathbf{x}) &= 1 - p_0(\mathbf{x}; g_\theta(\mathbf{x})) \quad \text{and} \quad f_2(\mathbf{x}) = f_{\text{LN}}(\tau|\mathbf{x}; g_\theta(\mathbf{x})).\end{aligned}\tag{6}$$

The neural network $g_\theta(\mathbf{x})$ now predicts the Bernoulli probability of $\tau \geq 0$ as p_0 along with μ and σ parameters for the distribution of τ given an input $\mathbf{x}$.

Finally, we can maximize the likelihood of the observed training data under the distribution described by GMDN in Equation 5, i.e.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N \left\{ \sum_{k=1}^K \pi_k(\mathbf{x}_i) f_k(\tau_i|\mathbf{x}_i; g_k(\mathbf{x}_i)) \right\}.\tag{7}$$

We have thus formulated a GMDN framework that allows a single neural network to model the presence of scattering in an FRB signal along with the distribution of τ . The estimation of the distribution of τ inherently accounts for the heteroskedastic uncertainty as well as enabling a robust confidence interval construction. The choice and architecture of the neural network for maximizing the likelihood in Equation 7 is discussed in the following sections.

2.2. Model Architecture

We experimented with various neural network architectures, starting with a convolutional neural network (CNN) autoencoder for each dynamic spectrum. CNNs impose translational invariance assuming ‘‘spatial stationarity’’, referring to uniform distribution of certain features of the image (Y. LeCun et al. 2015), which is not physically justified for scattering morphology. This is because of the fact that the same scattering morphology at the top of the frequency band corresponds to different scattering timescale than at the bottom of the frequency band. Thus, the CNN based regression failed to give good results with $R^2 < 50\%$ on the regression task. The vision transformer (ViT) (A. Dosovitskiy et al. 2021) has been outperforming CNNs in most tasks but it requires very large training dataset to learn the inductive biases associated with the image data.

Due to the limitations with the aforementioned architectures, we designed a transformer architecture for the representation of FRB dynamic spectra where each time bin in dynamic spectrum acts as a token with each token being d -dimensional frequency vector as shown in Figure 14. Upon training the transformer, we observed that the model was overfitting by performing exceptionally well on training data while not being able to generalize on test data. We then computed a frequency-averaged timeseries by collapsing the dynamic spectrum along the frequency axis. This operation produced a one-dimensional timeseries with the time samples equal to that of original dynamic spectrum. Each time sample is then projected to a d -dimensional vector with a single layer feed forward neural network, thereby creating embedding vectors for each time step. We constructed the same transformer architecture for time series as was developed for the dynamic-spectrum representation described above. However, we observed that the timeseries-based transformer model under-fit the data.

Ultimately, we developed a parallel transformer architecture to process both the dynamic spectrum and time series representations but separately and concatenated the contextual output from both of them. This parallel network achieved a performance superior to all other architectures and methods

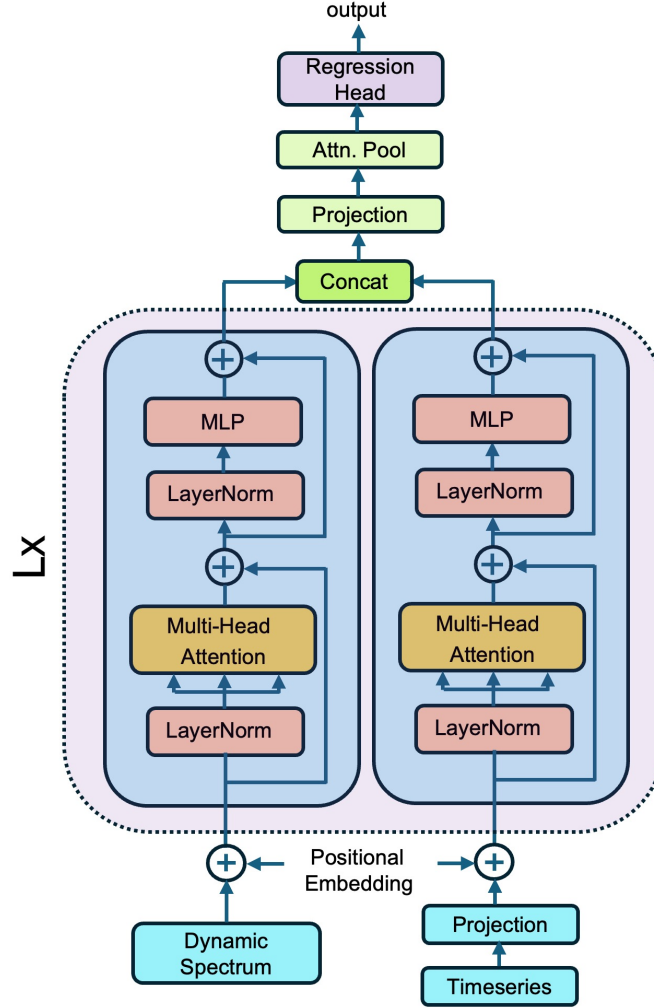


Figure 3. Schematic diagram of the parallel transformer architecture for our regression task. The model processes dynamic spectrum and timeseries representations with two independent parallel transformer branches. The resulting contextual embeddings are then concatenated and passed through a final regression head to predict the output values.

discussed before and was free from both overfitting and underfitting (see Section 3.1). The schematic diagram of the multi modal network architecture is shown in Figure 3. A breakdown of the network’s constituent components are discussed in Appendix A.

2.3. Data and preprocessing

For this work, we used real data acquired with the CHIME telescope presented in CHIME/FRB Catalog 2. The catalog provides de-dispersed dynamic spectra, $I(\nu, t)$, which are cutouts of the total intensity data. This dynamic spectra capture the intensity of FRB sources at different times (t) across different frequency channels (ν). Each dynamic spectrum consists of 16,384 frequency channels spanning the 400-800 MHz band and ~ 160 time samples with a resolution 0.983 ms. This time window is chosen to fully capture both the burst envelope and any scattering tail with the burst peak lying near the center of the window. We applied 64 channel block averaging to downsample frequency channels to 256, which improved the signal to noise while maintaining computational

tractability. Each dynamic spectrum was then independently min-max normalized:

$$I'(\nu, t) = \frac{I(\nu, t) - \min(I(\nu, t))}{\max(I(\nu, t)) - \min(I(\nu, t))}.$$

The normalization ensures all inputs lie in the range $[0, 1]$.

Data for the timeseries transformer can be generated by different dimensionality reduction techniques along the frequency axis as discussed in Appendix A.2. Each timeseries was also independently min-max normalized ensuring all the timeseries points lie in the range $[0, 1]$.

The target τ values for our regression task are taken from CHIME/FRB Catalog 2, which are reported in milliseconds as measured by the modeling tool `fitburst`. The target values are processed in such a way that $\tau' = \ln(\tau)$ for $\tau > 0$ and $\tau' = 0$ otherwise to compute the objective function in Equation B9.

We do not apply any explicit filtering of data samples based on signal to noise ratio, burst width or morphology. This broad inclusion of the data enables the model to learn from the full diversity of FRB population present in the CHIME catalog. Besides CHIME/FRB data, we employ synthetic data generated by `simpulse`¹² for the model evaluation purposes.

3. TRAINING AND RESULTS

The MT-GMDN was trained exclusively on data from CHIME/FRB Catalog 2. The Catalog 2 dataset was partitioned into train (~ 3800), validation (~ 150) and test (~ 150) sets for the events with valid scattering timescale values and the events with *NaN* values of scattering timescale were discarded. Training and validation sets were used while training, and test was used to evaluate the trained model. We randomly sampled the events in the partitions from both repeating and non-repeating sources as the repeating sources also display variability in the scattering (e.g., A. Pandhi et al. 2026; S. K. Ocker et al. 2023). The complete set of hyperparameter values adopted for training the MT-GMDN model are detailed in Appendix C. The final model corresponds to the training step achieving the lowest validation loss which is then used for quantitative evaluation and inferences.

We used the coefficient of determination (R^2) to assess model performance on the regression task, which is defined by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where,

- SS_{res} = Residual sum of squares;
- SS_{tot} = Total sum of squares.

The classification performance was quantified via a confusion matrix displayed in the Figure 7 and the resulting accuracy, recall, precision and F1-scores.

3.1. Quantitative performance on the test set from CHIME/FRB Catalog 2

For the regression task, we computed R^2 between the predicted conditional median scattering timescale value (e^μ) and the Catalog 2 value for the test set. The best value of R^2 was obtained

¹² `simpulse`: C++/python library for simulating pulses in radio astronomy

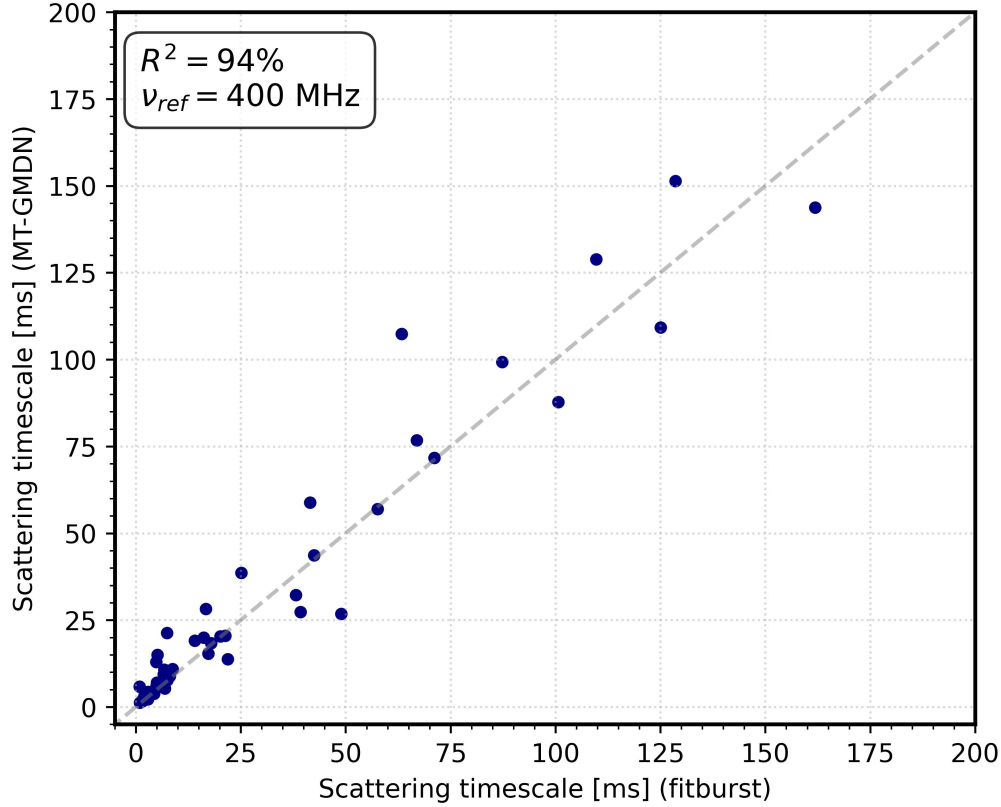


Figure 4. Scatter plot between the MT-GMDN predictions and `fitburst` measured values of scattering timescales at a reference frequency (ν_{ref}) of 400 MHz. The plot consists of events that have resolved scattering in CHIME/FRB Catalog 2 .

to be 94% for the model trained with timeseries computed by spectral averaging discussed in the Section A.2 and only the model performance is discussed in this section. The performances of the models corresponding to timeseries by other methods are discussed in Appendix A.2. Figure 4 shows the comparison between the MT-GMDN predicted scattering timescale values against the `fitburst` measured values in CHIME/FRB Catalog 2 . The result shows a strong linear correlation with most of the points symmetrically distributed along the $y = x$ line.

Along with the point estimate, we also constructed 95% confidence intervals $e^{\mu \pm 1.96 \times \sigma}$ for the point estimates. Figure 5 presents per sample comparison of MT-GMDN predictions against the `fitburst` measured cataloged values. The vast majority of samples have `fitburst` values within the confidence interval of the MT-GMDN point estimate. This indicates the model’s reliability not only on the point estimates but also on the well calibrated uncertainty bounds. As illustrated in Figure 5, the uncertainty associated with the point estimate scales with τ , which is consistent with the characteristic property of the exponential scattering tail. Thus the model was consistently able to recover τ values along with uncertainty for individual events across varying signal characteristics demonstrating its robustness.

To quantitatively assess the classification performance by the MT-GMDN, we first constructed a receiver operating characteristic (ROC) curve as shown in Figure 6. This ROC curves yields an area under the curve (AUC) value of 0.87 representing an excellent classifier (D. Hosmer et al. 2013).

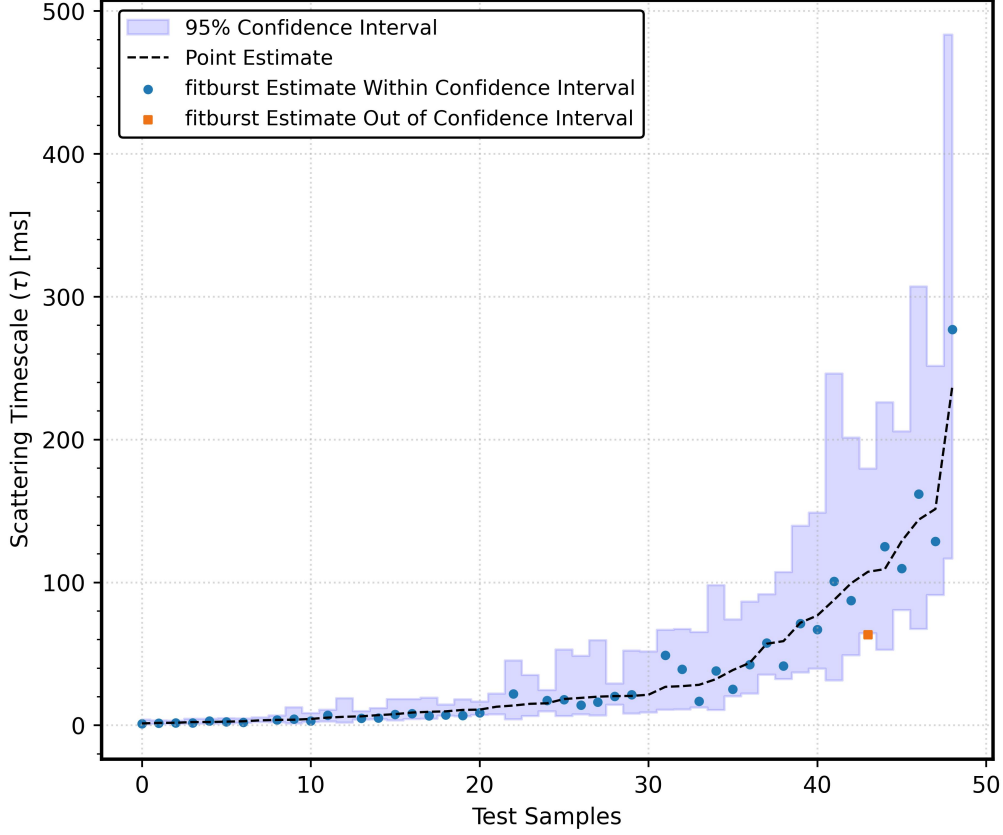


Figure 5. Test set prediction performance with x-axis denoting discrete FRB samples and y-axis scattering timescale at a reference frequency of 400 MHz. `fitburst` measured values are shown as blue dots and MT-GMDN point estimates are represented by black dashed lines while the blue shaded region representing 95% confidence interval. Orange squares denote the FRB samples which are out of the 95% confidence interval.

Figure 7 shows a confusion matrix for the scattering event detection at decision threshold of 0.6 (i.e. if $p_0 \geq 0.6$ then $\tau = 0$ else $\tau > 0$). At our chosen threshold, recall value of 90% was achieved and the overall accuracy was 76%. The precision of 63% suggest that the model is sensitive to scattering, and it predicted scattering in profiles that were unresolved in CHIME/FRB Catalog 2. This conservative approach is critical for ensuring that faint scattering signatures are not overlooked. The decision threshold is an adjustable parameter that can be tuned based on the specific objective of the analysis. For the most balanced performance, we can utilize the ROC curve to determine the optimal value of decision threshold.

3.2. Quantitative performance on simulated data

To analyze the robustness of the MT-GMDN predictions, we evaluated the performance of the model against a synthetic dataset generated with `simpulse` framework. We observed a consistent systematic bias in the recovered τ values, where the model was overestimating the predictions relative to the injected ground truth. The systematic bias was observed when modeling with `fitburst`. This discrepancy likely originates from overestimation of the scattering effect within the `simpulse` simulation backend and/or a fundamental morphological mismatch between the templates used by

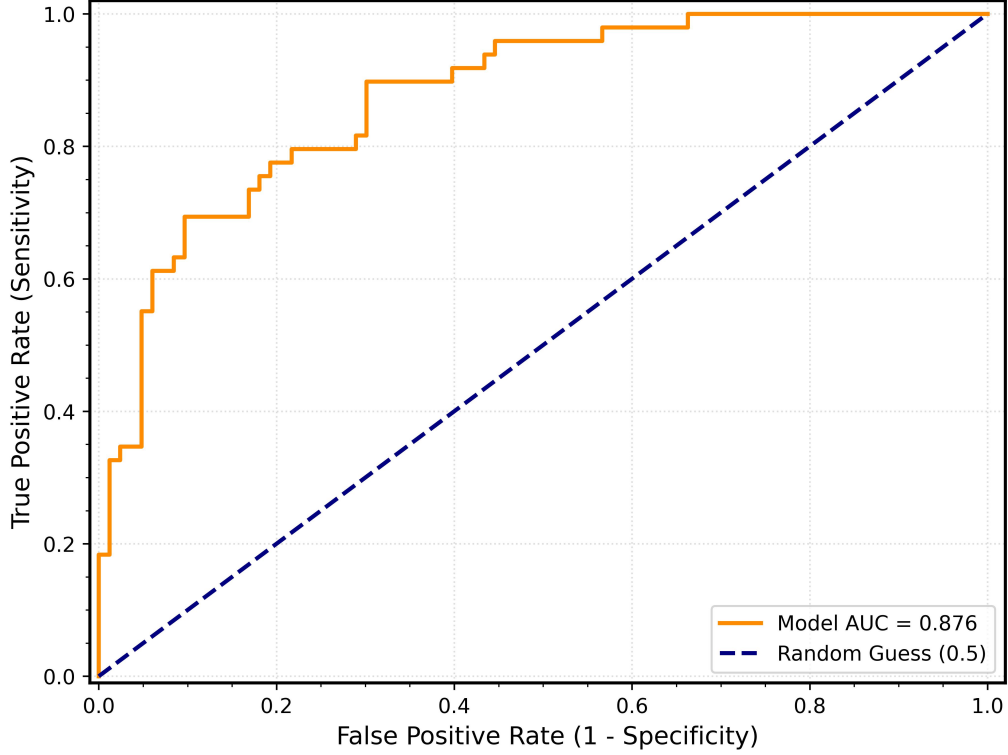


Figure 6. Receiver operating characteristic curve for scattering detection by MT-GMDN at different decision threshold values.

`fitburst` from which our training target labels were derived and numerical methods implemented in the `simpulse` backend.

We accounted this systematic offset by a linear calibration of the model’s point estimate and upper and lower bound of the confidence interval. The linear calibration was obtained by modeling zero-noise dynamic spectra with `fitburst` for the minimum and maximum values of scattering timescales in the simulated data and then calculating the mean offset factor between the measured and injected values. As illustrated in Figure 8, the systematic offset correction aligns the predictions with the ideal line ($y = x$), along with the injected target values falling entirely within the predicted 95% confidence interval. Furthermore, the MT-GMDN captures the measurement uncertainty which are represented by the error bars. These uncertainties scale proportionally with the τ values and is physically expected as higher τ values dominate the pulse morphology and suppresses the signal to noise of the burst. The successful capture of the target values with heteroskedastic uncertainty confirms that the MT-GMDN provides both accurate point estimate and reliable confidence interval for scattering timescale.

3.3. MT-GMDN vs. *fitburst* on simulated data

The process of model fitting with `fitburst` involves two steps. At first, a model without scattering is fitted and the parameters obtained from the first fit is then used to create a model with scattering. Chi-square (χ^2) statistic for each of the fit is computed. These χ^2 statistics are then used to obtain F-statistic

$$F = \frac{\chi_1^2 - \chi_2^2}{\chi_2^2} \frac{N_{\text{dof},2}}{\Delta N_{\text{fit}}},$$

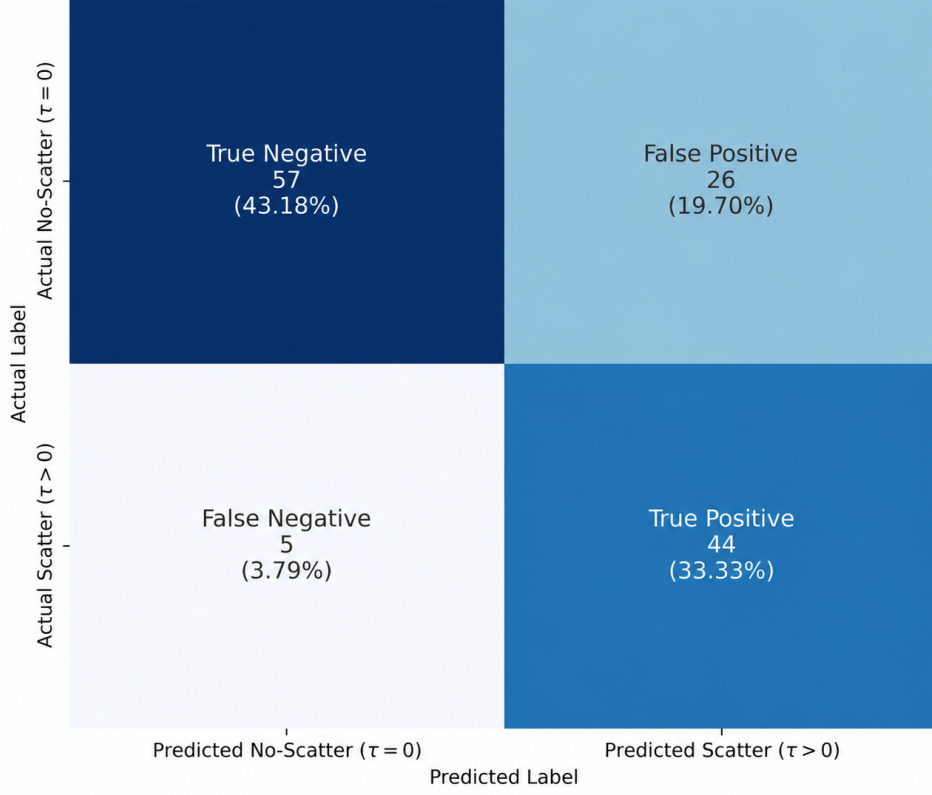


Figure 7. Confusion matrix for the scattering event detection at a detection threshold value of $p_0 = 0.6$. Here no-scatter refers to unresolved scattering rather than physical absence of scattering.

where $N_{\text{dof},2}$ is the number of degrees of freedom in the complex model with scattering and ΔN_{fit} is the difference in the number of free parameters between the two models. Using the F-statistic, p -value is obtained which represents the improvement in the fit quality after the introduction of the scatter broadening parameter in the model.

To compare the robustness of our MT-GMDN against **fitburst**, we created common sets of broadband synthetic events and applied both **fitburst** and MT-GMDN methods to recover the τ values. **fitburst** performed exceptionally well on the data sets where the signal to noise of the events were high and so did MT-GMDN. However, in the low SNR data set where the scattering tails often get mixed up with the background noise, MT-GMDN outperforms **fitburst** which is illustrated in Figure 9. To quantify the performance difference between **fitburst** and MT-GMDN we again employed R^2 score as a metric for comparison. R^2 score of 39% was obtained for **fitburst** measured values while the R^2 score was 68% for MT-GMDN predicted values. Moreover, Figure 9 (Left) for **fitburst** measurement shows significant number of events for which the p -value of F-test are significantly high thereby failing to reject the null hypothesis that the non-scattering model is better fit for data. For some of the events, target values fall out of the 95% confidence interval and uncertainties do not scale well with noise levels and τ values. Figure 9 (Right) shows that all the target values fall within the confidence interval predicted by MT-GMDN with uncertainties scaling well with noise levels and τ values. The 95% confidence intervals are very broad for MT-GMDN method as compared to

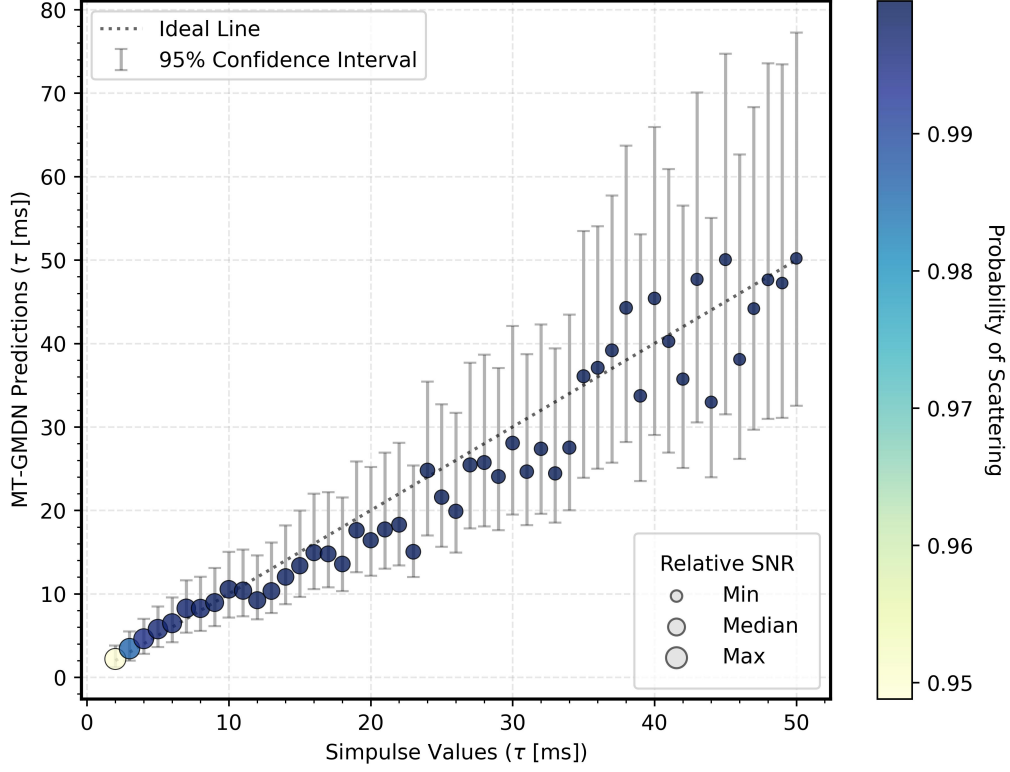


Figure 8. Scattering timescale recovery for synthetic broadband pulses generated by `simpulse` with intrinsic widths of 1 ms. The vertical error bars denote 95% confidence interval with all of the target values within the interval. The MT-GMDN model achieves R^2 score of 92% on the point estimate of the τ values. Measurement uncertainty scales with τ justifying the inherent difficulty in characterizing the morphology of the highly scattered profiles.

`fitburst` at very low signal to noise values of the synthetic bursts (<8 for bursts with `simpulse` injected value of > 15 ms), however the point estimates and lower bounds are robust as compared to `fitburst`.

Besides the performance on recovering τ values from data, MT-GMDN also has computational advantage. `fitburst` relies on non-linear least squares optimization for each event which is an iterative process and the algorithm time complexity scales up with the number of frequency channels (N_ν), time samples (N_t) and number of iterations to converge. After the convergence, `fitburst` computes exact Hessian and inverts it to obtain the covariance matrix and hence uncertainty. This operation also scales with N_ν and N_t . However, once trained, MT-GMDN has the time complexity of $\mathcal{O}(1)$ regardless of the complexity of a burst. It has been observed that MT-GMDN achieves inference speed of approximately four orders of magnitude (10^4) faster than `fitburst` on the retrieval of τ value and the corresponding uncertainty, enabling real time parameter estimation and uncertainty quantification.

3.4. MT-GMDN vs. MCMC on simulated data

MCMC can be used to sample from the posterior distribution in Bayesian approach of parameter estimation when marginal likelihood (evidence) over all the possible values of parameters in high dimensional parameter space becomes intractable. We used the same mathematical model as used

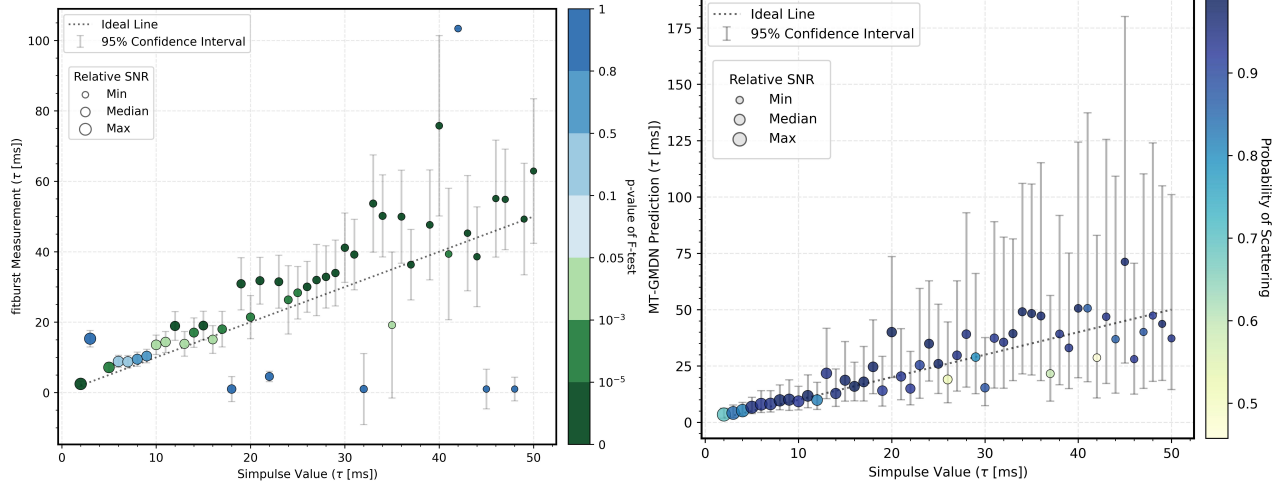


Figure 9. Comparison of `fitburst` measured values of τ on synthetic data against MT-GMDN predictions. The synthetic data set consists of broadband pulses with maximum SNR for the event with minimum scattering. Fluence for each of the event is held constant such that the SNR gets lower for the events with higher scattering due to the redistribution of radiation energy across time in exponentially decaying tail. Left: `fitburst` measurement with the vertical bars representing 95% confidence interval, the size of the markers corresponding to relative SNR and the colorbar represents p -value of F-test for scattering vs non-scattering model fit. Right: MT-GMDN predictions with vertical bars capturing 95% confidence interval with the marker size corresponding to relative SNR while the marker color representing the probability of scattering. A linear correction factor was applied to both `fitburst` and MT-GMDN methods. The maximum signal to noise value of the synthetic data was 20 corresponding to the left most point in the plots and minimum was 4 corresponding to the right most point in the plots.

by `fitburst` to model the scatter broadening in pulse profile (M. M. McKinnon 2014) given as

$$T_{kn} = \left(\frac{\nu_k}{\nu_r} \right)^{-\delta} \exp \left[\frac{\sigma^2}{2\tau_k^2} - \frac{(t_{kn} - t_0)}{\tau_k} \right] \left[1 + \operatorname{erf} \left(\frac{t_{kn} - (t_0 + \sigma^2/\tau_k)}{\sigma\sqrt{2}} \right) \right]. \quad (8)$$

Here, T_{kn} is the pulse profile value at k^{th} channel and n^{th} time bin. σ is the Gaussian width of the pulse, t_0 is arrival time of the pulse, τ_k is scattering timescale and δ is the index of scattering accounting for frequency dependence of scattering and was set to 4 which is the empirical standard value for FRBs. ν_r is the reference frequency which was fixed to 400 MHz for our analysis.

The spectrum was modeled by using the running power law model given as

$$F_k = \left(\frac{\nu_k}{\nu_r} \right)^{\gamma + \beta \ln(\frac{\nu_k}{\nu_r})}. \quad (9)$$

Here, γ is spectral index and β is spectral running.

Finally, a two dimensional dynamic spectrum D_{kn} was then modeled by combining Equations 8 and 9 as

$$M_{kn} = AF_k T_{kn}, \quad (10)$$

with $A = 10^\alpha$, the amplitude of overall wave packet characterized by global amplitude parameter α .

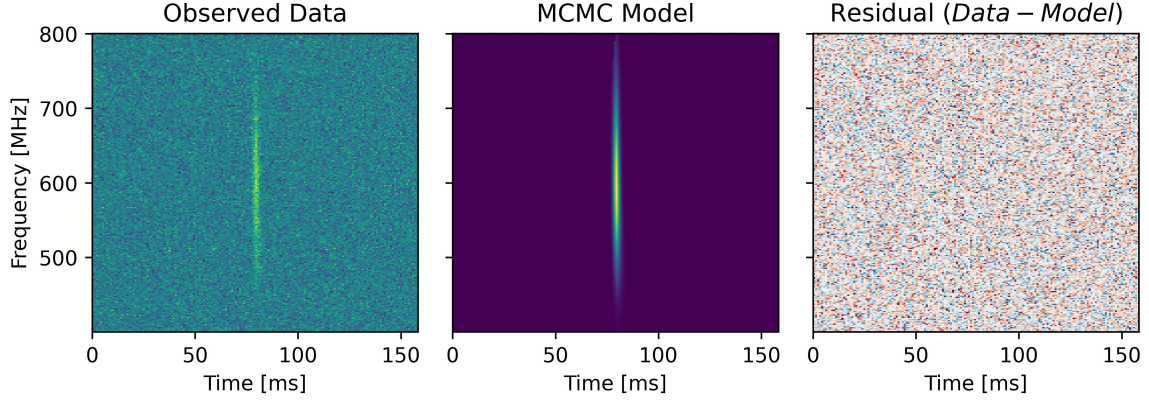


Figure 10. Dynamic spectrum of a synthetic FRB with modeled by MCMC method. The injected pulse had an intrinsic width of 1 ms and scattering timescale of 2 ms.

To evaluate the MT-GMDN performance against MCMC on estimating the scattering timescale, we needed to create complete model of each dynamic spectrum in this evaluation. An example of a model generated by MCMC method is shown in Figure 10. Figure 11 illustrates the parameter distributions along with their corresponding 95% credible intervals for the simulated FRB previously displayed in Figure 10.

We modeled 50 simulated FRBs with MCMC using lognormal distribution as prior for the τ and extracted the posterior distribution for τ by combining all MCMC chains. The heteroskedasticity was incorporated in the MCMC modeling by passing the noise value of a dynamic spectrum directly to the Gaussian likelihood of the MCMC. The sampler used was “No-U-Turn Sampler” (M. D. Hoffman & A. Gelman 2011) with a burn-in period of 2000 and 4 parallel chains. The comparison plot for MCMC method and MT-GMDN methods is shown in Figure 12.

Some of the key metrics comparing the performance of MCMC versus MT-GMDN on both point estimate and uncertainty are summarized in the Table 1. MT-GMDN performs slightly better on root mean square error (RMSE) and R^2 score indicating improved predictive accuracy and better explained variance than MCMC. Both of the approaches also demonstrate negligible bias in the point estimation with MT-GMDN slightly better than MCMC. The significant difference between MT-GMDN and MCMC methods is on the the uncertainty modeling behaviors. MT-GMDN exhibits over-coverage, potentially due to conservative estimation of variance in the mixture density framework. However, under-coverage in uncertainty by the MCMC method indicates narrow predictive intervals reflecting its limitations in capturing the full variability of the data. Despite MCMC producing narrower uncertainty intervals, its under-coverage indicates that these intervals are not reliable while MT-GMDN prioritize more conservative but statistically consistent uncertainty estimates at the expense of sharpness. MT-GMDN thus demonstrates competitive or superior point-estimates and lower bounds of the uncertainty interval.

3.5. Heteroskedastic uncertainty modeling by MT-GMDN

To assess the uncertainty estimation nature of MT-GMDN based on the amount of noise in a dynamic spectrum, we analyze the correlation between the predicted uncertainty parameter (σ) and SNR of some simulated bursts. Higher SNR samples are expected to yield more precise estimates

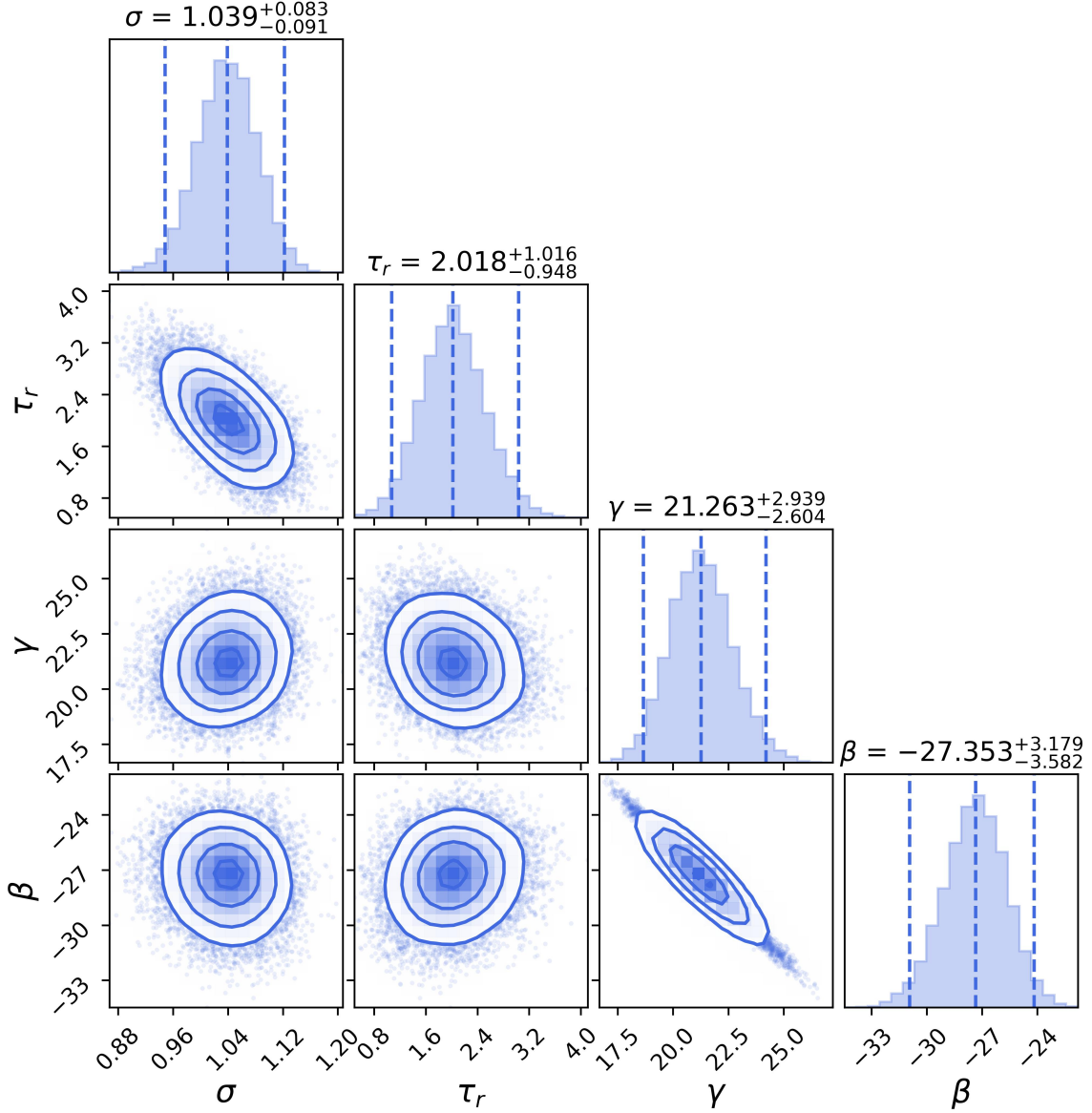


Figure 11. Posterior distribution of some of the model parameters in Equation 10 generated by MCMC method for a simulated FRB shown in the Figure 10. The central vertical line in the corner plots represent median or the point estimate while the first and third vertical lines denote the 95% credible interval. The posterior distributions are obtained by combining the samples from all MCMC sampler chains. The simulated burst had width of 1 ms and scattering time of 2 ms and MCMC method successfully recovered those values as seen in the corner plots for τ_r as scattering time and σ as the width.

while lower SNR events are expected to produce broader uncertainty interval. This implies a negative correlation between predictive uncertainty and SNR of FRB events.

Figure 13 shows the variation of uncertainty parameter σ with SNR for two representative cases of scattering with scattering timescales of 5 ms (left) and 10 ms (right). As the scatter plots suggest non-linear relationship between σ and SNR, we implemented Spearman correlation method to quantify the correlation in terms of correlation coefficient ρ . We found very strong negative correlations between σ and SNR for both cases with statistical significance test yielding p -values of $p = 9.8 \times 10^{-18}$ for the

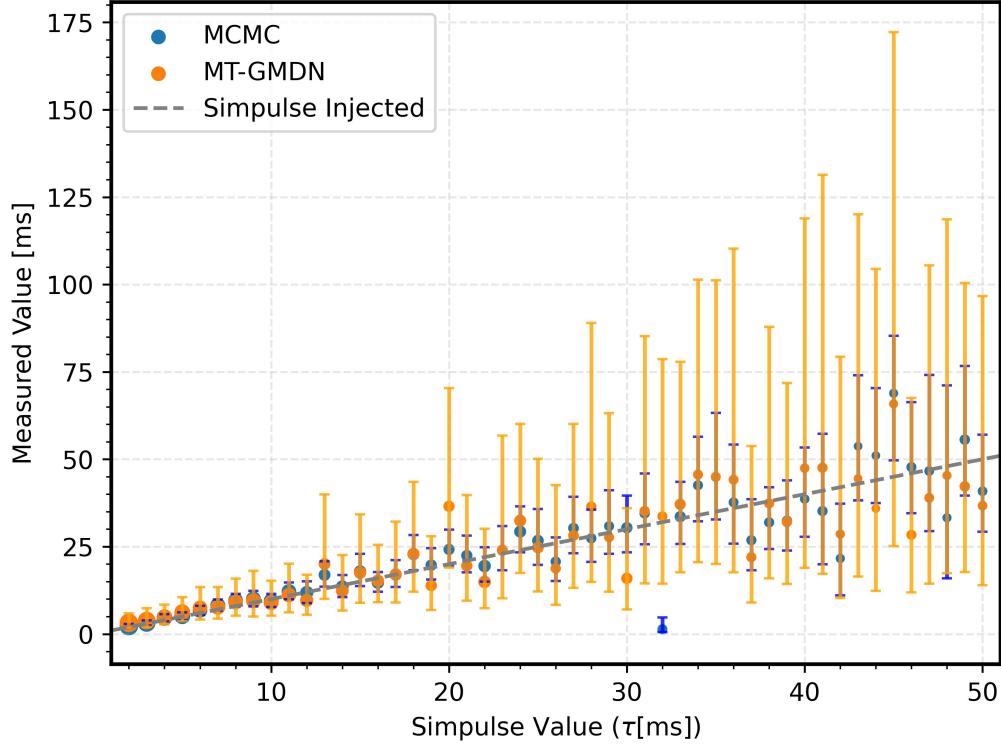


Figure 12. Comparison of MCMC vs MT-GMDN performance on the simulated dataset. The size of the markers denote the SNR of corresponding dynamic spectrum and it ranges from 20 to 4 with larger size denoting higher SNR. The error bars for MT-GMDN predictions represent 95% confidence interval and the error bars for MCMC estimates represent 95% credible interval. The pulse width of all the bursts were set to 1 ms. Both the MCMC and MT-GMDN estimations were systematic bias corrected as introduced by simpulse.

Table 1. MT-GMDN vs. MCMC Comparison Metrics

Metric	Equation	MT-GMDN	MCMC	Comments
RMSE	$\left[\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \right]^{1/2}$	7.62	7.83	Lower is better
R^2	$1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$	0.71	0.69	Higher is better
Bias	$\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$	-0.003	-0.068	Ideal ≈ 0
95% Coverage	$\frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i \in [l_i, u_i])$	1	0.89	Ideal ≈ 0.95
MIW	$\frac{1}{N} \sum_{i=1}^N (u_i - l_i)$	19.69	2.49	Lower is better if calibrated

NOTE—RMSE: Root mean square error, MIW: Mean interval width, $\hat{y}_i$: Point estimate, y_i : Target, u_i : Upper bound estimate, l_i : Lower bound estimate

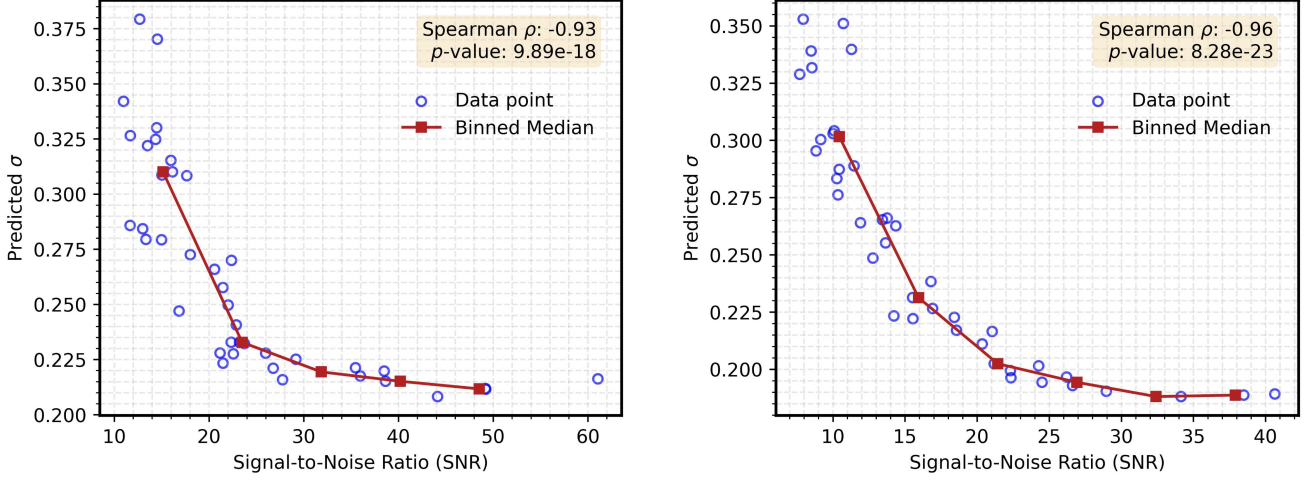


Figure 13. Relationship between predicted uncertainty parameter (σ) and SNR for two representative cases. Left plot displays the variation of σ with SNR for simulated events with width of 1 ms and τ of 5 ms while the right plot is for simulated events with width of 1 ms and τ of 10 ms.

case with $\tau = 5$ ms and $p = 8.28 \times 10^{-23}$ for the case with $\tau = 10$ ms. This confirms that observed correlations are highly unlikely by chance. This statistically significant correlation result confirms that MT-GMDN produces heteroscedastic uncertainty estimates consistent with observational noise properties.

4. DISCUSSIONS

This work demonstrates that FRB morphology can be directly used by a deep learning framework to efficiently and precisely estimate astrophysical parameters such as τ without relying on explicit pulse-shape fitting. The proposed MT-GMDN framework models the full conditional distribution $p(\tau|\mathbf{x})$ capturing both resolved scattering and heteroskedastic uncertainty in a statistically principled manner.

One of the key results is that the model predicts significant fraction of false positives on the test data set from CHIME/FRB Catalog 2 as seen in Figure 7. This property of the model should not be interpreted as purely an error because the catalog reported $\tau = 0$ values are only the lower limits but not physically exact zeros. Figure 9 and Appendix E clearly demonstrate that for low SNR and very subtle scattering events the traditional scattering measurement methods might fail. However, MT-GMDN appears to be more sensitive to subtle temporal asymmetries due to low level scattering signatures which are not captured by traditional fitting approaches when operating near the instrumental resolution limit or low SNRs.

The choice of conditional median as the primary point estimate of τ instead of mean has been motivated by various factors. One of the main reasons being its robustness to a heavy-tailed lognormal distribution. While the mean is strongly influenced by the large scattering tails, the median often faithfully represents typical scattering timescale imposed by the data. Also, the median of the lognormal corresponds to the exponentiated mean of the underlying normal distribution which allows a convenient construction of the confidence intervals on the point estimation of τ .

5. LIMITATIONS AND FUTURE WORKS

Despite our MT-GMDN yielding promising results, there are still several limitations that motivates future work. The first limitation is that the model was trained on a very small size of training data set and there were modest number of samples with high τ values (> 100 ms). This may hinder the model's performance on accurately characterizing highly scattered events. Furthermore, the model was trained exclusively on total intensity data from CHIME/FRB Catalog 2, generalization to high resolution CHIME/FRB data (baseband data) and also cross instrumental generalization remains to be investigated. Another notable constraint is the model's maximum temporal window (context length) of 162 time samples. This limits the characterization of events requiring larger temporal extent.

To enhance the model performance, future work will involve training the model by augmenting the CHIME/FRB catalog dataset with synthetic bursts injected into real telescope noise. We also plan to incorporate high resolution data sets in the training process allowing the model to better capture subtle burst structures. Additionally, we plan to extend the model architecture to estimate other morphological characteristics beyond the current implementation.

6. CONCLUSION

In this work, we have presented a Multimodal Transformer Based Generic Mixture Density Network to infer an astrophysical parameter τ from burst morphology encoded in a dynamic spectrum. The model jointly estimates the probability of unresolved scattering and the conditional distribution of τ and thus yielding robust point estimates alongside well-calibrated uncertainty bounds.

The introduction of multimodal framework that leverages complementary information from both temporal domain and spectral domain solves the fundamental challenge of limited dataset to train the transformer based architecture. Our probabilistic mixture density formulation provides a robust solution to handle zero-inflated data set during training thus allowing accurate modeling of both scattered and unresolved scattered signals. The successful application of the model on real CHIME/FRB test data set and the synthetic test data set demonstrates the robustness and scientific utility of our framework.

The proposed framework offers a non-parametric alternative to classical scattering pipelines and can be implemented to automate the large-scale scattering inference. The classical methods are still useful for verification purposes.

ACKNOWLEDGMENTS

We acknowledge that CHIME is located on the traditional, ancestral, and unceded territory of the Syilx/Okanagan people.

We are grateful to the staff of the Dominion Radio Astrophysical Observatory, which is operated by the National Research Council of Canada. CHIME is funded by a grant from the Canada Foundation for Innovation (CFI) 2012 Leading Edge Fund (Project 31170) and by contributions from the provinces of British Columbia, Québec and Ontario. The CHIME/FRB project is funded by a grant from the CFI 2015 Innovation Fund (Project 33213) and by contributions from the provinces of British Columbia and Québec, and by the Dunlap Institute for Astronomy and Astrophysics at the University of Toronto. Additional support is provided by the Canadian Institute for Advanced Research (CIFAR), McGill University and the McGill Space Institute thanks to the Trottier Family Foundation, and the University of British Columbia. The CHIME/Pulsar instrument hardware is

funded by the Natural Sciences and Engineering Research Council (NSERC) Research Tools and Instruments (RTI-1) grant EQPEQ 458893-2014.

M.N. is a Fonds de Recherche du Quebec - Nature et Technologies (FRQNT) postdoctoral fellow. M.W.S acknowledges support from the French government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université - AMIDEX (AMX-23-CEI- 088). P.S. acknowledges the support of an NSERC Discovery Grant (RGPIN-2024-06266).

Software: Numpy (C. R. Harris et al. 2020), Matplotlib (J. D. Hunter 2007), PyTorch (A. Paszke et al. 2019), torchvision (TorchVision maintainers and contributors 2016), pandas (T. pandas development team 2020), seaborn (M. L. Waskom 2021), fitburst (E. Fonseca et al. 2024), simpulse (K. Smith 2005)

DATA AND SOFTWARE AVAILABILITY

The source code used in this work is publicly available on GitHub¹³.

APPENDIX

A. BREAKDOWN OF MODEL ARCHITECTURE

A.1. *Dynamic Spectrum Transformer*

The dynamic spectrum transformer ingests dynamic spectrum which is tokenized along n -dimensional time axis with each token being d -dimensional frequency vector:

$$\mathbf{x}_{DS} \in \mathbb{R}^{n \times d}.$$

This choice of dynamic spectrum input preserves the frequency details for each time stamp and allows

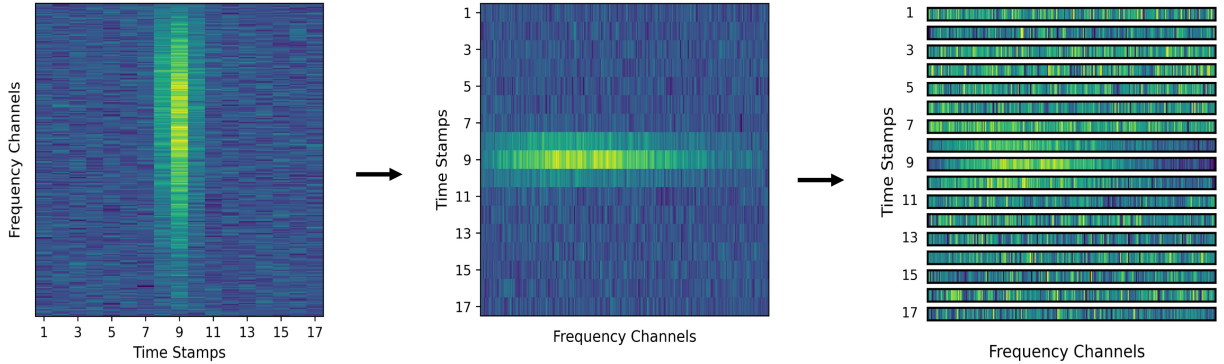


Figure 14. Illustration of the tokenization scheme for a dynamic spectrum. Each time stamp acts as a distinct token and the frequency intensities vector at that time stamp serves as token embedding referred to as d -dimensional frequency vector.

the self attention mechanism to model the long range temporal dependencies along with frequency dependent temporal smearing.

¹³ <https://github.com/kharelb/Scattering-Timescale-Estimation-With-Transformers>

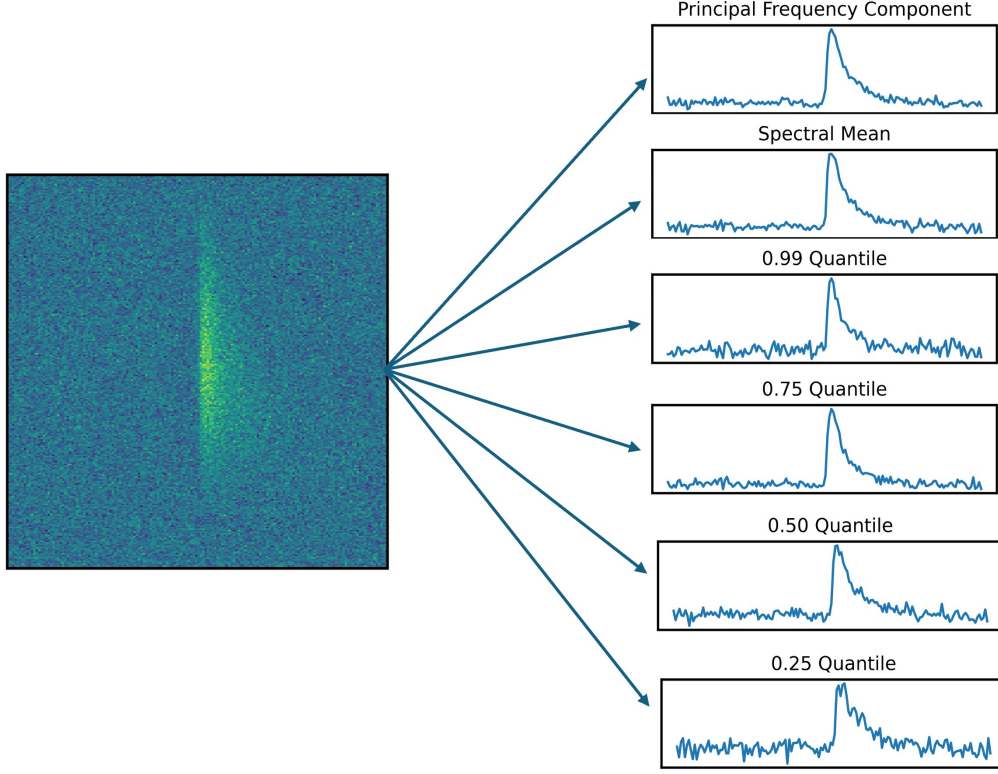


Figure 15. Timeseries (pulse profile) created by different dimensionality reduction techniques along the frequency dimension.

The dynamic spectrum transformer is similar to the standard transformer encoder (A. Vaswani et al. 2023) consisting of two encoder layers with each encoder layer containing 4 self-attention heads which performs full self-attention. We experimented with higher number of heads (8, 16) but there was no significant change in the performance of the model yielding the similar values of R^2 as given by 4 heads.

A.2. Timeseries Transformer

We can compute timeseries from a dynamic spectrum by performing dimensionality reduction along the frequency axis. Some of the timeseries computation methods include simple spectral averaging, projection onto principal frequency components using principal component analysis (PCA), and quantile computation across frequency channels. Timeseries obtained by different approaches as discussed are shown in Figure 15. Timeseries data thus can be represented as

$$\mathbf{x}_{\text{TS}} \in \mathbb{R}^n.$$

The timeseries representation of the dynamic spectrum captures global temporal asymmetry and scattering tails while suppressing frequency-dependent structure and thus provides a complementary information to the dynamic spectrum.

The projection layer in timeseries transformer projects each scalar timeseries value to d -dimensional vector with a simple feed forward network. Each projected vector corresponding to each time sample in timeseries serves as token embedding which is illustrated in Figure 16.

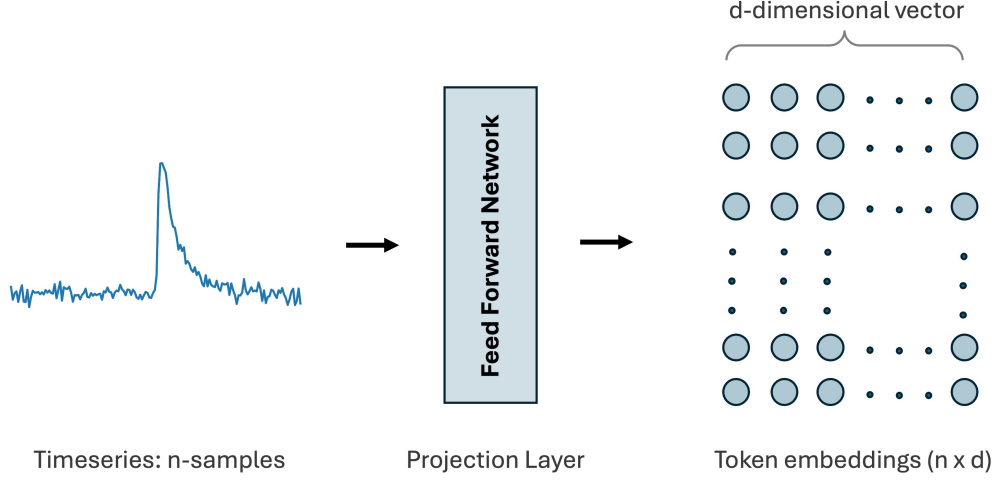


Figure 16. Illustration of the projection layer where each time sample is mapped to d -dimensional learnable embedding vector by a simple feed forward neural network.

The timeseries transformer has the same number of encoder layers and attention heads as the dynamic spectrum transformer. This transformer thus focuses on capturing pulse asymmetry, tail decay structure, temporal sub-components and burst duration variations.

A.3. Positional encoding

To make both of the transformers aware of temporal position, positional embedding vectors were added to both dynamic spectrum representation token embeddings and timeseries representation token embeddings. The same positional embeddings were used for both transformers.

$$\begin{aligned}\tilde{\mathbf{x}}_{\text{DS}} &= \mathbf{x}_{\text{DS}} + \mathbf{r} \\ \tilde{\mathbf{x}}_{\text{TS}} &= \mathbf{x}_{\text{TS}} + \mathbf{r}\end{aligned}$$

Upon experimenting, learnable positional embedding yielded better result in our case than sinusoidal positional embedding introduced by [A. Vaswani et al. \(2023\)](#).

A.4. Latent representation and fusion

Each transformer generates a sequence of context rich embeddings of the same shape as input:

$$\mathbf{z}_{\text{DS}} \in \mathbb{R}^{n \times d}, \quad \mathbf{z}_{\text{TS}} \in \mathbb{R}^{n \times d}.$$

These two feature rich latent representation outputs from the encoders are fused by concatenation:

$$\mathbf{z} = [\mathbf{z}_{\text{DS}}; \mathbf{z}_{\text{TS}}] \in \mathbb{R}^{n \times 2d}$$

This strategy preserves the modality specific information and also prevents interference between the temporal and spectral features extracted.

A.5. Projection and attention pooling

The fused encoder outputs are then passed through linear projection layer to produce output

$$\tilde{\mathbf{z}} = \mathbf{W}^T \mathbf{z} + b, \tag{A1}$$

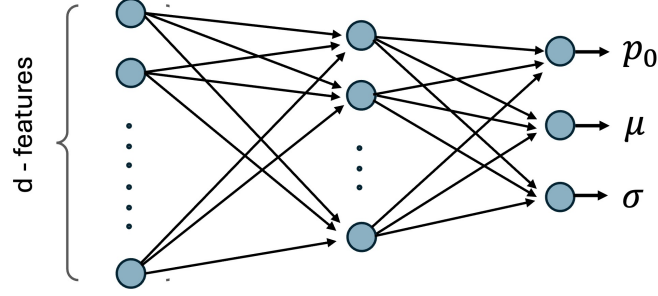


Figure 17. A schematic diagram of regression head. The d -dimensional output from the preceding attention pooling layer is passed through a linear layer to obtain the parameters of the mixture density as outputs.

where $\tilde{\mathbf{z}} \in \mathbb{R}^{n \times d}$ and $\mathbf{W}$ and $\mathbf{b}$ are weights and bias associated with the linear transformation.

We utilize an attention based pooling mechanism to aggregate the output from the last projection layer to a single vector for regression head. We first project each vector from $\tilde{\mathbf{z}}$ corresponding to time step to a scalar score using learnable linear transformation

$$\mathbf{e} = \mathbf{W}^T \tilde{\mathbf{z}} + b,$$

where $\mathbf{W}$ and b are learnable parameters. The learned scores are then normalized across the temporal dimension using the softmax function to obtain attention weights α_t :

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t=1}^n \exp(e_t)}$$

Finally, a weighted sum is performed to produce a global representation

$$h = \sum_{t=1}^n \alpha_t \tilde{\mathbf{z}}_t.$$

This approach allows the model to weight different time stamps dynamically based on their relevance.

A.6. Regression head

The regression head accepts a latent vector $h \in \mathbb{R}^d$ from the preceding attention pooling layer. The latent vector is then passed through a linear layer to give an output $\tilde{h}$. It is then fed into three parallel linear heads as shown in Figure 17 which are responsible for distinct distribution parameters of the Bernoulli and lognormal distributions in the mixture density. The three outputs from the regression head are further discussed below.

Bernoulli head (f_{p_0}): It predicts the probability of zero scattering timescale i.e. the Bernoulli probability of the mixture density.

$$p_0 = \sigma(f_{p_0}(\tilde{h}))$$

Where σ is standard sigmoid function

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

Location head (f_μ): It predicts the location parameter of the lognormal distribution in the mixture density.

$$\mu = f_\mu(\tilde{h})$$

Scale head (f_σ): It predicts the scale parameter of the lognormal in the mixture density.

$$\sigma = \text{softplus}(f_\sigma(\tilde{h}))$$

Where,

$$\text{softplus}(x) = \ln(1 + e^x)$$

B. OPTIMIZATION FUNCTION

The optimization task in our case is maximizing the likelihood with respect to the mixture density given by Equation 5. So our objective (error) function as seen in Equation 7 is

$$E(\boldsymbol{\theta}) = \prod_{i=1}^N \left\{ \sum_{k=1}^K \pi_k(\mathbf{x}_i) f_k(\tau_i | \mathbf{x}_i; g_k(\mathbf{x}_i)) \right\}. \quad (\text{B2})$$

Where, $\boldsymbol{\theta}$ are the weights of the neural network. Using Equations 6, Equation B2 becomes

$$E(\boldsymbol{\theta}) = \prod_{i=1}^N \{ p_0(\mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \delta(\tau_i) + (1 - p_0(\mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))) f_{\text{LN}}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \}. \quad (\text{B3})$$

If $p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))$ is the probability of observing scattering timescale τ_i for a given sample $\mathbf{x}_i$ then there arises two cases:

Case I: $\tau_i = 0$:

$$p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) = p_0(\mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))$$

Case II: $\tau_i > 0$:

$$p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) = (1 - p_0(\mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))) f_{\text{LN}}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))$$

Thus the objective function is:

$$E(\boldsymbol{\theta}) = \prod_{i=1}^N p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \quad (\text{B4})$$

Instead of maximizing the above expression directly, it becomes computationally efficient to maximize the log transformed value of the above expression. Since, the logarithm is a monotonically increasing function, maximizing above expression is equivalent to maximizing the log transformed value. Thus our objective function can be written as

$$E(\boldsymbol{\theta}) = \sum_{i=1}^N \ln \{ p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \}. \quad (\text{B5})$$

Maximizing above equation is equivalent to minimizing the negative of it, so the objective function becomes:

$$E(\boldsymbol{\theta}) = - \sum_{i=1}^N \ln \{ p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \}. \quad (\text{B6})$$

Now, let us derive the logarithm of probabilities for both cases $\tau_i = 0$ and $\tau_i > 0$

Case I: $\tau_i = 0$: If probability under the model:

$$p_0(\tau_i = 0 | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) = p_{0,i}$$

Then,

$$\ln\{p_0(\mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))\} = \ln p_{0,i} \quad (\text{B7})$$

Case II: $\tau_i > 0$: In this case:

$$\begin{aligned} p_0(\tau_i > 0 | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) &= (1 - p_{0,i}) f_{LN}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \\ \implies \ln p_0(\tau_i > 0 | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) &= \ln(1 - p_{0,i}) + \ln f_{LN}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) \end{aligned}$$

But since,

$$\begin{aligned} f_{LN}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) &= \frac{1}{\tau_i \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln \tau_i - \mu_i)^2}{2\sigma_i^2}\right) \\ \implies \ln f_{LN}(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) &= -\ln(\tau_i) - \ln(\sigma_i) - \frac{1}{2} \ln(2\pi) - \frac{(\ln \tau_i - \mu_i)^2}{2\sigma_i^2} \end{aligned}$$

So,

$$\ln p_0(\tau_i > 0 | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i)) = \ln(1 - p_{0,i}) - \ln(\tau_i) - \ln(\sigma_i) - \frac{1}{2} \ln(2\pi) - \frac{(\ln \tau_i - \mu_i)^2}{2\sigma_i^2} \quad (\text{B8})$$

Finally, our error function given by Equation B6 normalized for data size becomes

$$E(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^N \ln \{p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))\} \quad (\text{B9})$$

With,

$$\ln \{p(\tau_i | \mathbf{x}_i; g_{\boldsymbol{\theta}}(\mathbf{x}_i))\} = \begin{cases} \ln p_{0,i}, & \tau_i = 0, \\ \ln(1 - p_{0,i}) - \ln(\tau_i) - \ln(\sigma_i) - \frac{1}{2} \ln(2\pi) - \frac{(\ln \tau_i - \mu_i)^2}{2\sigma_i^2}, & \tau_i > 0 \end{cases} \quad (\text{B10})$$

Thus the Equation B9 is our final objective function which is minimized during training by gradient descent method.

C. IMPLEMENTATION DETAILS AND HYPERPARAMETERS

In this section, we provide the specific configuration and hyperparameter settings used during training to ensure the reproducibility of our results. The model architecture was implemented with deep learning framework PyTorch (A. Paszke et al. 2019) and was trained on NVIDIA A100 Tensor Core GPU provided by Dolly Sods¹⁴. All the model architectural parameters were set through manual tuning on the validation dataset. The checkpoint that achieved the minimum validation loss was the best model with robust configuration for generalization presented in Table 2.

¹⁴ A high performance computing cluster at West Virginia University.

Table 2. MT-GMDN Hyperparameter Configuration

Category	Hyperparameter	Value	Description
Architecture	Transformer Heads	4	Multi-head attention count
	Encoder Layers	3	Number of stacked transformer blocks
	Embedding Dim (d_{model})	256	Latent vector size
	FFNN Dimension	2048	Feed-forward hidden layer size
	Dropout Rate	0.1	Regularization to prevent overfitting
	MDN Hidden Units	[256, 256]	Neurons in the mixture density head
Training	Batch Size	32	Number of samples per iteration
	Learning Rate	1×10^{-4}	Learning rate for optimizer
	Weight Decay	0.09	L_2 regularization penalty
	Max Epochs	400	Total training iterations
	Optimizer	AdamW	Optimization algorithm
Data	Temporal Window	162	Maximum input time samples
	Frequency Channels	256	Spectral channels of dynamic spectra

NOTE—The hyperparameters for both dynamic spectrum and timeseries transformers are same.

D. MODEL PERFORMANCE WITH DIFFERENT TIMESERIES APPROACHES

The estimation of scattering timescale with MT-GMDN relies on reduction of two dimensional dynamic spectrum into a corresponding one dimensional timeseries. While there are different statistical approaches to generate timeseries from a dynamic spectrum, in this work we utilized spectral mean, frequency quantiles, and first principal frequency component using principal component analysis (PCA) and evaluated their performance after training. The best performing model was achieved on spectral mean approach and is explained in Section 3.

Figure 18 presents the evaluation of different alternative approaches other than spectral mean on the real test data set. The results suggests that these methods perform reliably at low scattering timescales but they struggle for pulses exhibiting significant scattering tails.

E. MODEL COMPARISON WITH SOME INDIVIDUAL SYNTHETIC BURSTS

Here, we present a comparison between `simpulse` generated pulses, `fitburst` measured values and corresponding p-values of F-test for scattering detections and the estimates from the MT-GMDN.

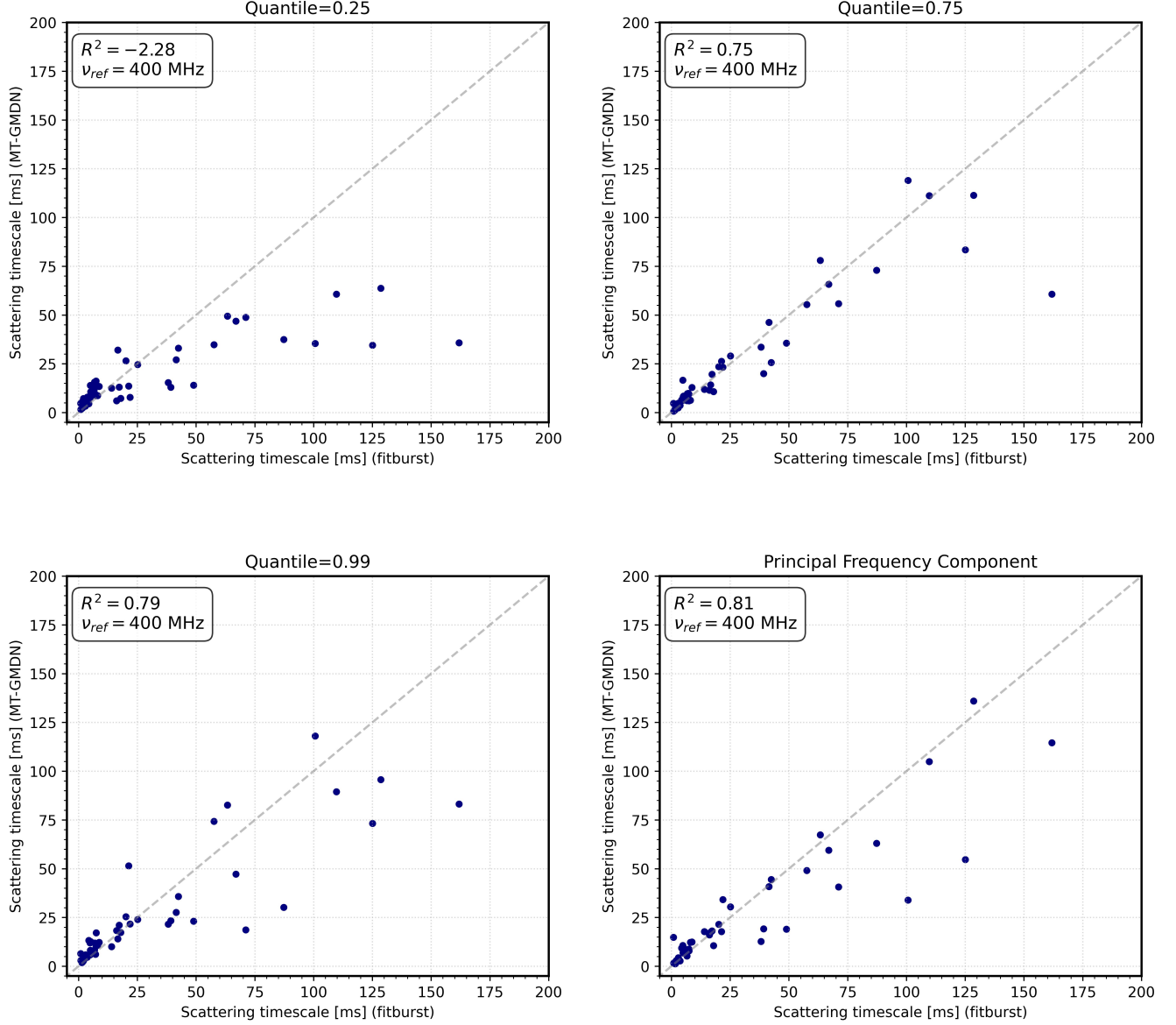


Figure 18. Evaluation of MT-GMDN performance across four different timeseries extraction methods.

All of the examples in the Figure 19 have intrinsic width of 1 ms and same fluence values while noise levels vary between the bursts. p-values from `fitburst` measurements dictates the existence of scattering in a burst while prediction probability of scattering dictates the presence of scattering by MT-GMDN approach. Examining these individual bursts with aforementioned methods provides an insight on how MT-GMDN performs across different burst profiles. We can observe that MT-GMDN maintains prediction accuracy of τ values in various regimes where `fitburst` returns non significant p-values ($p > 0.05$) or even when `fitburst` fails to detect a burst in a highly scattered low SNR profile.

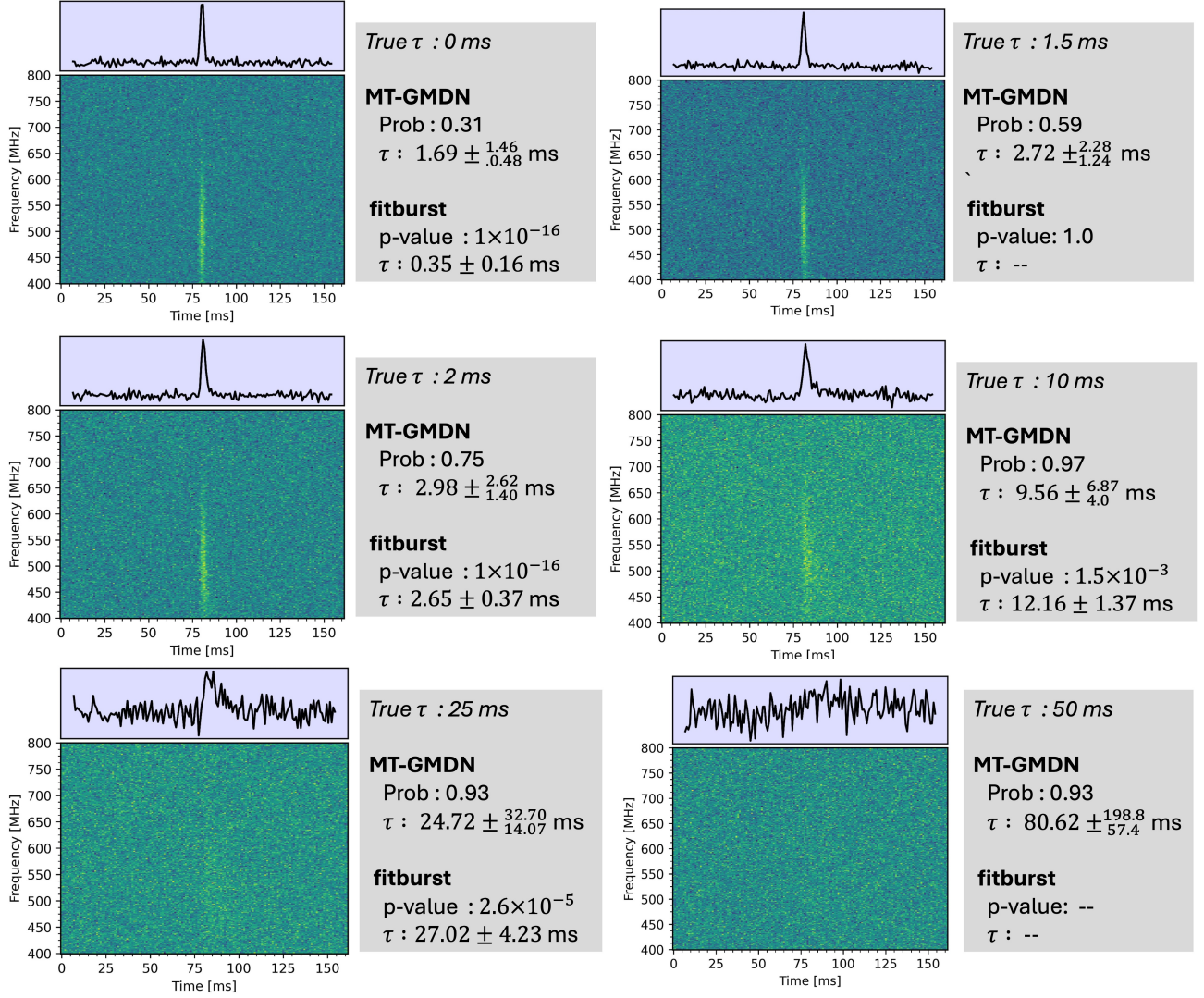


Figure 19. Each panel includes two sub-panels with the first sub-panel containing a simulated dynamic spectrum and the corresponding timeseries. While the second sub-panel describes real injected τ value, the MT-GMDN point estimate with 95% confidence interval, and the `fitburst` fit statistics. The uncertainty provided for the `fitburst` measurement here corresponds to the 1σ limits. All the τ values are referenced at 400 MHz.

F. SOME FRBS WITH STRONG NOISE AND RADIO FREQUENCY INTERFERENCE

To demonstrate significance of MT-GMDN, we present a set of representative FRBs shown in the Figure 20 from the CHIME/FRB Catalog 2 that are affected by noise and radio frequency interference (RFI). Due to the strong noise and heavy RFI, extracting any physical information such as τ is very challenging by traditional methods discussed before. Despite the challenges, MT-GMDN produces visibly stable predictions with increased uncertainty to lower quality observations as displayed in Figure 20.

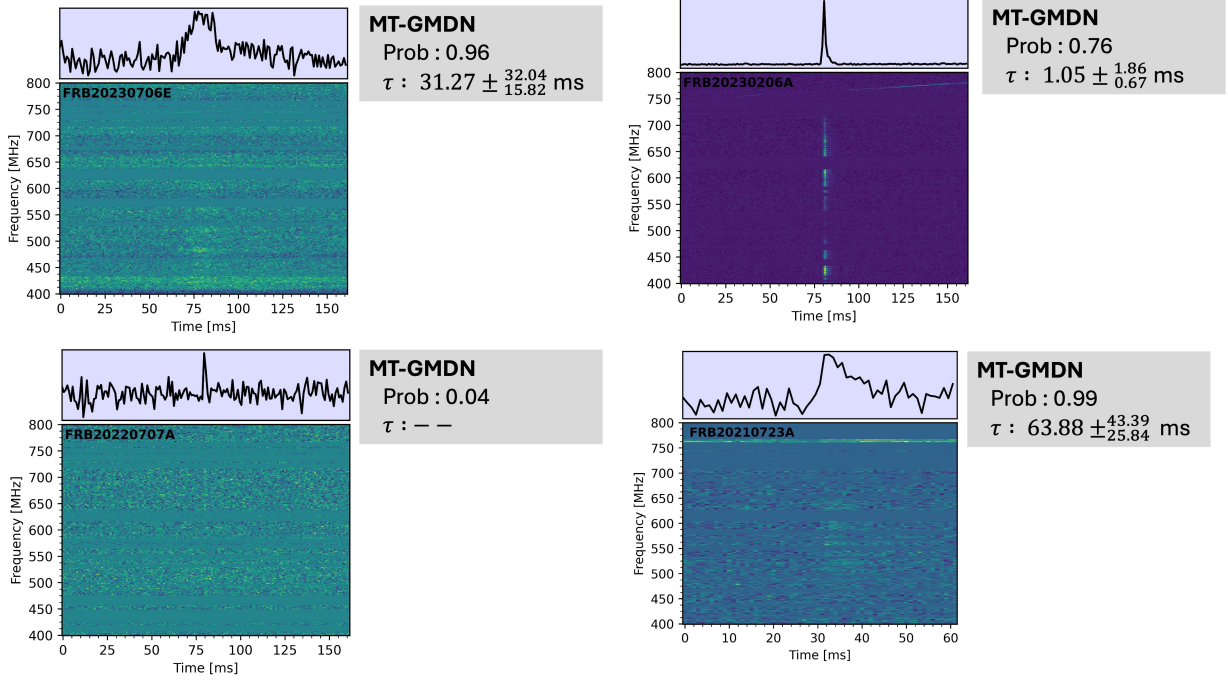


Figure 20. Representative FRBs from the CHIME/FRB Catalog 2 where traditional methods failed to extract physical parameters due to strong noise and RFI. Each panel includes two sub-panels with the first sub-panel (left) containing a dynamic spectrum and the corresponding timeseries, while the second sub-panel (right) contains MT-GMDN estimates. The estimates include probability of scattering, point estimate for τ value and its 95% predictive interval.

G. ALTERNATIVE GAMMA MIXTURE DENSITY FORMULATION

As seen above from results, the lognormal formulation can produce wide upper confidence bounds for events with higher scattering and/or lower signal to noise. This is because, the variance is computed in the log scale and then transformed back to linear scale. A very small variation of σ in log scale can lead to large variation in the linear scale. One alternative to resolve the inflated confidence interval upper bound can be the use of distribution in the mixture density formulation with lighter tail than lognormal. Also the distribution should be strictly positive and operating in linear scale.

Gamma distribution satisfies all of the properties of alternative distribution discussed above. In this section, we examine the effect of distribution on confidence interval width by replacing the lognormal with the gamma distribution in the mixture density formulation in Equation 5. Gamma distribution for a random variable x is given by

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, \quad (\text{G11})$$

where α is the shape parameter, β is rate parameter and $\Gamma(\alpha)$ is the Gamma function. Following Section B, we can easily compute the error function by replacing $f_2(\mathbf{x})$ in Equation 6 by the Gamma distribution. The logarithm term in Equation B10 thus gets modified to

$$\log p(\tau_i | \mathbf{x}_i) = \begin{cases} \log p_{0,i}, & \tau_i = 0, \\ \log(1 - p_{0,i}) + \alpha_i \log \beta_i - \log \Gamma(\alpha_i) + (\alpha_i - 1) \log \tau_i - \beta_i \tau_i, & \tau_i > 0. \end{cases} \quad (\text{G12})$$

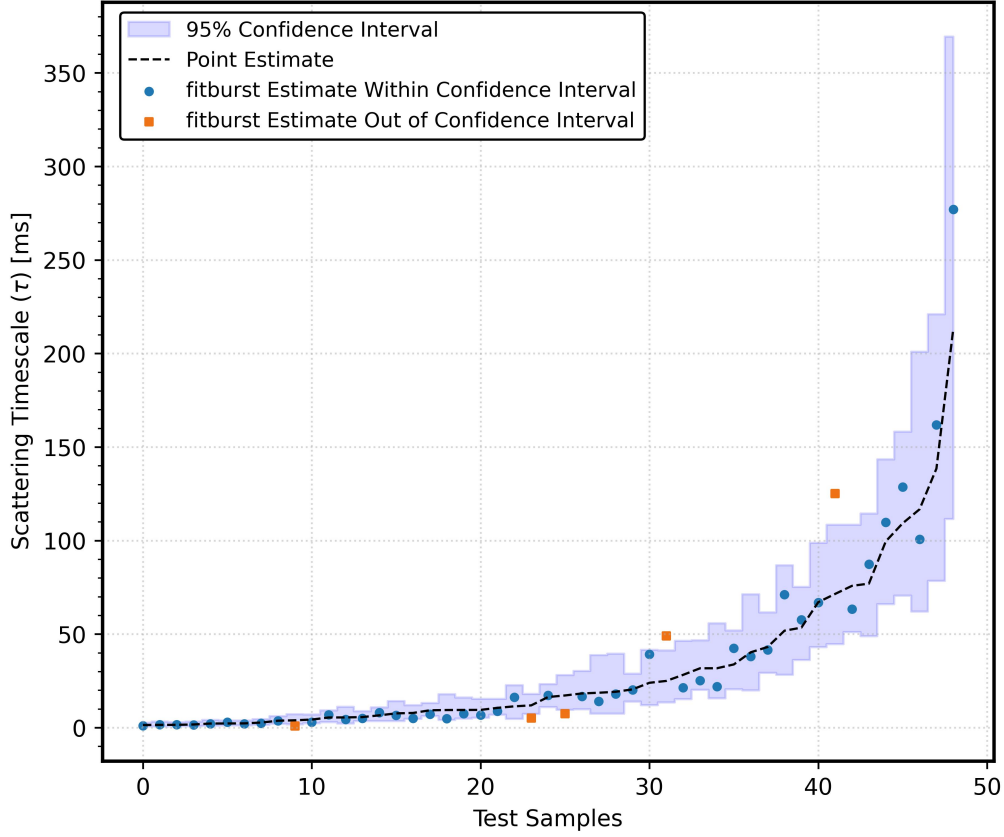


Figure 21. Test set prediction performance with x-axis denoting discrete FRB samples and y-axis scattering timescale at a reference frequency of 400 MHz. The mixture density formulation here implements the Gamma distribution instead of the Lognormal distribution. `fitburst` measured values are shown as blue dots and MT-GMDN point estimates are represented by black dashed lines while the blue shaded region representing 95% confidence interval. Orange squares denote the FRB samples which are out of the 95% confidence interval.

The neural network was then trained to estimate the parameters of the Gamma distribution and the Bernoulli's parameter in the mixture density formulation. Point estimate was computed by the median of the target distribution with the upper and the lower bounds defined by 0.025 and 0.975 quantiles respectively. Figure 21 shows per sample comparison of MT-GMDN predictions against the `fitburst` measured values after replacing the Lognormal distribution by the Gamma distribution in mixture density formulation for the test data set. R^2 -score of 88% was achieved which represents slight degradation in the point estimation with the Gamma distribution method. However, it can be clearly seen from Figures 5 and 21 that the width of the confidence interval in Gamma distribution method produces much more narrower confidence intervals than that of Lognormal distribution method. This can be quantified by the MIW defined in the Table 1. The MIW with the Gamma distribution is 32.74 while with the Lognormal distribution is 56.38. The classification performances with both the Gamma and Lognormal distributions in the mixture density formulation remained almost identical since Bernoulli's distribution was used in both of the cases for classification.

Thus the Gamma component can significantly reduce the inflated confidence intervals for the high scattering and/or low signal to noise dynamic spectrum at the expense of slightly degraded R^2 -score in the point estimate.

REFERENCES

- Agarwal, D., Aggarwal, K., Burke-Spolaor, S., Lorimer, D. R., & Garver-Daniels, N. 2020, Monthly Notices of the Royal Astronomical Society, 497, 1661, doi: [10.1093/mnras/staa1856](https://doi.org/10.1093/mnras/staa1856)
- Bhat, N. D. R., Cordes, J. M., & Chatterjee, S. 2003, ApJ, 584, 782, doi: [10.1086/345775](https://doi.org/10.1086/345775)
- Bishop, C. M., & Bishop, H. 2024, Deep Learning: Foundations and Concepts (Springer), doi: [10.1007/978-3-031-45468-4](https://doi.org/10.1007/978-3-031-45468-4)
- CHIME/FRB Collaboration, Amiri, M., Bandura, K., et al. 2018, ApJ, 863, 48, doi: [10.3847/1538-4357/aad188](https://doi.org/10.3847/1538-4357/aad188)
- CHIME/FRB Collaboration, Amiri, M., Bandura, K., et al. 2019, Nature, 566, 230, doi: [10.1038/s41586-018-0867-7](https://doi.org/10.1038/s41586-018-0867-7)
- Chime/FRB Collaboration, Abbott, T., Andersen, B. C., et al. 2026, ApJS, 283, 34, doi: [10.3847/1538-4365/ae3828](https://doi.org/10.3847/1538-4365/ae3828)
- Connor, L., & van Leeuwen, J. 2018, The Astronomical Journal, 156, 256, doi: [10.3847/1538-3881/aae649](https://doi.org/10.3847/1538-3881/aae649)
- Cordes, J. M., & McLaughlin, M. A. 2003, ApJ, 596, 1142, doi: [10.1086/378231](https://doi.org/10.1086/378231)
- Cordes, J. M., Ocker, S. K., & Chatterjee, S. 2022, Astrophys. J., 931, 88, doi: [10.3847/1538-4357/ac6873](https://doi.org/10.3847/1538-4357/ac6873)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2021, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, <https://arxiv.org/abs/2010.11929>
- Fonseca, E., Pleunis, Z., Andersen, B. C., et al. 2024, The Astrophysical Journal Supplement Series, 272, 7, doi: [10.3847/1538-4365/ad27d6](https://doi.org/10.3847/1538-4365/ad27d6)
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357, doi: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2)
- Hoffman, M. D., & Gelman, A. 2011, arXiv e-prints, arXiv:1111.4246, doi: [10.48550/arXiv.1111.4246](https://doi.org/10.48550/arXiv.1111.4246)
- Hosmer, D., Lemeshow, S., & Sturdivant, R. 2013, Applied Logistic Regression: Third Edition (wiley), doi: [10.1002/9781118548387](https://doi.org/10.1002/9781118548387)
- Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90, doi: [10.1109/MCSE.2007.55](https://doi.org/10.1109/MCSE.2007.55)
- Jonas, J., & MeerKAT Team. 2016, in MeerKAT Science: On the Pathway to the SKA, 1, doi: [10.22323/1.277.0001](https://doi.org/10.22323/1.277.0001)
- Kharel, B., Fonseca, E., Brar, C., et al. 2025, Repeating vs. Non-Repeating FRBs: A Deep Learning Approach To Morphological Characterization, <https://arxiv.org/abs/2509.06208>
- LeCun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436
- Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J., & Crawford, F. 2007, *Science*, 318, 777, doi: [10.1126/science.1147532](https://doi.org/10.1126/science.1147532)
- Lorimer, D. R., & Kramer, M. 2004, Handbook of Pulsar Astronomy, Vol. 4 (Cambridge, UK ; New York: Cambridge University Press)
- Macquart, J.-P., Bailes, M., Bhat, N. D. R., et al. 2010, PASA, 27, 272, doi: [10.1071/AS09082](https://doi.org/10.1071/AS09082)
- McKinnon, M. M. 2014, Publications of the Astronomical Society of the Pacific, 126, 476, doi: [10.1086/676975](https://doi.org/10.1086/676975)
- Ocker, S. K., Cordes, J. M., & Chatterjee, S. 2021, ApJ, 911, 102, doi: [10.3847/1538-4357/abeb6e](https://doi.org/10.3847/1538-4357/abeb6e)
- Ocker, S. K., Cordes, J. M., Chatterjee, S., et al. 2023, MNRAS, 519, 821, doi: [10.1093/mnras/stac3547](https://doi.org/10.1093/mnras/stac3547)
- pandas development team, T. 2020, pandas-dev/pandas: Pandas, latest Zenodo, doi: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134)
- Pandhi, A., Nimmo, K., Andrew, S., et al. 2026, ApJL, 1000, L53, doi: [10.3847/2041-8213/ae52f8](https://doi.org/10.3847/2041-8213/ae52f8)
- Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32 (Curran Associates, Inc.), 8024–8035
- Sherman, M., & DSA-110 Collaboration. 2023, in American Astronomical Society Meeting Abstracts, Vol. 241, American Astronomical Society Meeting Abstracts #241, 239.04
- Shin, K., Leung, C., Simha, S., et al. 2025, Astrophys. J., 993, 208

- Shivraj Patil, S., Main, R. A., Fonseca, E., et al. 2025, arXiv e-prints, arXiv:2509.06721, doi: [10.48550/arXiv.2509.06721](https://arxiv.org/abs/2509.06721)
- Smith, K. 2005, simpulse: C++/python library for simulating FRBs and pulsars,, <https://github.com/kmsmith137/simpulse>
- Taylor, J. 1997, Introduction to Error Analysis, the Study of Uncertainties in Physical Measurements, 2nd Edition
- TorchVision maintainers and contributors. 2016, TorchVision: PyTorch’s Computer Vision library,, <https://github.com/pytorch/vision> GitHub
- Vaswani, A., Shazeer, N., Parmar, N., et al. 2023, Attention Is All You Need, <https://arxiv.org/abs/1706.03762>
- Waskom, M. L. 2021, Journal of Open Source Software, 6, 3021, doi: [10.21105/joss.03021](https://doi.org/10.21105/joss.03021)
- Williamson, I. P. 1974, MNRAS, 166, 499, doi: [10.1093/mnras/166.3.499](https://doi.org/10.1093/mnras/166.3.499)
- Xu, S., & Zhang, B. 2016, ApJ, 832, 199, doi: [10.3847/0004-637X/832/2/199](https://doi.org/10.3847/0004-637X/832/2/199)
- Yang, Tsung-Ching, Hashimoto, Tetsuya, Hsu, Tzu-Yin, et al. 2025, A&A, 693, A85, doi: [10.1051/0004-6361/202450823](https://doi.org/10.1051/0004-6361/202450823)