NeuroTD: A Time-Frequency Based Multimodal Learning Approach to Analyze Time Delays in Neural Activities

Xiang Huang[1], Noah Cohen Kalafut[2], Sayali Alatkar[2], Athan Z. Li[2], Qiping Dong[1], Qiang Chang[1,3], and Daifeng Wang[1,2,4]

[1] Waisman Center, University of Wisconsin-Madison, Wisconsin, US
[2] Department of Computer Sciences, University of Wisconsin-Madison, Wisconsin, US
[3] Departments of Medical Genetics and Neurology, University of Wisconsin-Madison, Madison, Wisconsin, US
[4] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Wisconsin, US

Corresponding Author: daifeng.wang@wisc.edu

Abstract. Studying temporal features of neural activities is crucial for understanding the functions of neurons as well as underlying neural circuits. To this end, recent researches employ emerging techniques including calcium imaging, Neuropixels, depth electrodes, and Patch-seq to generate multimodal time-series data that depict the activities of single neurons, groups of neurons, and behaviors. However, challenges persist, including the analysis of noisy, high-sampling-rate neuronal data, and the modeling of temporal dynamics across various modalities. To address these challenges, we developed NeuroTD, a novel deep learning approach to align multimodal time-series datasets and infer cross-modality temporal relationships such as time delays or shifts. Particularly, NeuroTD integrates Siamese neural networks with frequency domain transformations and complex value optimization for inference. We applied NeuroTD to three multimodal datasets to (1) analyze electrophysiological (ephys) time series measured by depth electrodes, identifying time delays among neurons across various positions, (2) investigate neural activity and behavioral time series data derived from Neuropixels and 3D motion captures, establishing causal relationships between neural activities and corresponding behavioral activities, and (3) explore gene expression and ephys data of single neurons from Patch-seq, identifying gene expression signatures highly correlated with time shifts in ephys responses. Finally, NeuroTD is open-source at `https://github.com/daifengwanglab/NeuroTD` for general use.

Keywords: Biologically Inspired Deep Learning · Joint Time-Frequency Domain Alignment · Siamese Neural Network · Multimodal Time-Series Analysis · Neural Time Delays

1 Introduction

Understanding the temporal dynamics of neural activities is essential for decoding neural functions and connectivity. These dynamics often involve specific time delays or shifts, which provide insights into the interactions within neural circuits and the information flow among neurons. For instance, there is a delay in signal transmission during synaptic events between neurons - time takes for a presynaptic neuron to trigger a response in a postsynaptic neuron [1]. This signal transmission typically takes milliseconds and is crucial for understanding synaptic efficiency and the synchronization of neural circuits. Similarly, time delays between neural and behavioral responses occur as signals travel between the brain and body, such as during motor commands or reflexes, where signals pass through the spinal cord to execute movements. Another example is the intrinsic time delays within neurons, such as the intervals between spikes or peaks in electrophysiological responses within a single neuron [2]. These delays reveal variations in neuronal activities and contribute to our understanding of how individual neuron response to stimulus over time.

Biologists have long recognized the biological importance of time delays in neural processes. For instance, owls use auditory time delays to determine target distances during echolocation, where the "best delay" between a pulse and its echo corresponds directly to the distance of prey [3]. Similarly, time delays play a key role in synfire chains, which are neural circuits that propagate synchronous volleys of spikes, critical for computation in the brain [4]. While these studies have provided insights into how time delays function in groups of neurons, they relied on low-resolution methods that could not capture delays at the single-neuron level, and used simple approaches to calculate delays.

Recent advances in measurement techniques, such as calcium imaging [5], depth electrodes [6], and Neuropixels [7], allow us to measure thousands of individual neural activities simultaneously with high temporal resolution, providing unprecedented insight into time delays at the single-neuron level. However, the data generated by these advanced techniques present significant challenges for analyzing time delays. First, high-sampling-rate data from methods like depth electrodes and Neuropixels are often noisy, which complicates the process of extracting meaningful patterns. Second, neural and behavioral activities operate on different time scales, with varying temporal resolutions and distinct dynamics. For instance, electrophysiological signals can be rapid and transient, while behavioral responses may be slower and more continuous. These differences make it difficult to align signals effectively, adding complexity to calculating time delays. Previous studies have proposed multiple approaches to analyze time delays or align time-series data. One widely used method is Dynamic Time Warping (DTW), which aligns two time-series by minimizing the global alignment error through dynamic programming [8]. DTW can handle variable-length time-series and is robust to time shifts, making it useful in contexts like speech recognition and activity monitoring. However, DTW is sensitive to noise, shape, and dynamic differences, and its computational cost is high, though faster approximations exist. Another approach is Canonical Correlation Analysis (CCA), which maximizes the correlation between two sets of signals, reducing dimensionality while preserving meaningful correlations [9]. This makes CCA useful for understanding relationships between multimodal data sources like neural signals and behaviors, but it assumes linear relationships between signals, which limits its effectiveness for nonlinear interactions. The Canonical Time Warping (CTW) method combines DTW and CCA to enable temporal alignment and feature selection for sequences with varying dimensionality [10]. However, CTW struggles with non-linear transformations and complex temporal dynamics, which limit its effectiveness in more intricate alignment tasks. Additionally, its computational intensity makes it less suitable for large-scale applications. Other methods, such as Fourier-based Correlation, leverage the fast Fourier transform (FFT) to align signals based on frequency domain representations [11]. This approach is fast and facilitates noise reduction in the frequency domain, although it assumes linear shifts and may struggle with non-periodic signals. Finally, AlignNet, a deep learning-based method, provides end-to-end learning for aligning audio-visual sequences [12]. However, AlignNet is primarily targeted at video and audio alignment in computer vision, using warping and affinity functions that are not applicable for neural time-series data. It is also computationally expensive, tested on low-sampling-rate video and audio, and can be challenging to integrate with priors, limiting its flexibility for neural data applications.

To address those challenges, we develop NeuroTD, a deep learning approach to analyze time delays in neural activities by combining a Siamese neural network with frequency domain transformation. By mapping

data into a shared latent space using deep learning and applying frequency transformations to convert time delays into linear phase shifts, it ensures the efficient computation of noise-resilient time delays with high accuracy. We applied NeuroTD to three multimodal datasets: analyzing electrophysiological time series from depth electrodes to identify time delays across neurons, exploring neural activity and behavioral time series from Neuropixels and 3D motion capture to establish causal relationships between neural and behavioral activities, and correlating gene expression with electrophysiological data from Patch-seq to identify gene expression signatures linked to time shifts in intrinsic neuron responses.

2 Methods

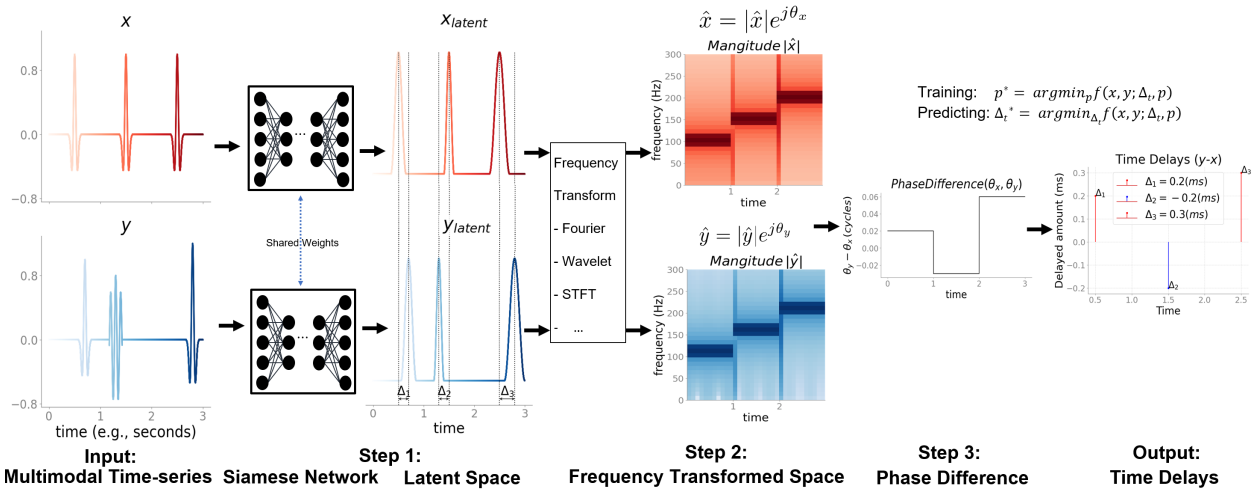

Fig. 1. NeuroTD: A Time-Frequency Based Multimodal Learning Approach to Analyze Time Delays in Neural Activities. Input - Multimodal Time-series Data: We begin with two time-series datasets, $\mathbf{x}$ (red) and $\mathbf{y}$ (blue), which can represent different modalities, such as neural activities and behavioral data. In this example, both series have three corresponding events, each with different shapes and magnitudes that we aim to align. **Step 1 - Infer Common Time-Domain Latent Space by Siamese Network:** We employ a deep Siamese neural network, composed of two identical subnetworks with shared weights. Each subnetwork processes one of the time-series inputs, transforming them into a latent space representation. In this space, the transformed representations, x_{latent} and y_{latent}, capture the core features of the original time-series in a unified form. **Step 2 - Transform to Frequency Space:** These latent space representations are then converted into the frequency domain using techniques such as the Fourier Transform, Wavelet Transform, or Short-Time Fourier Transform. The transformed signals are represented as $\hat{\mathbf{x}} = |\hat{\mathbf{x}}|e^{j\theta_x}$ and $\hat{\mathbf{y}} = |\hat{\mathbf{y}}|e^{j\theta_y}$, where $|\hat{\mathbf{x}}|$ and $|\hat{\mathbf{y}}|$ indicate magnitudes, and θ_x and θ_y represent the phases of the signals. **Step 3 - Calculate Phase Difference:** The phase difference between $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ is calculated to determine the time delays between the two modalities. **Output - Time Delays:** The time delays, Δ_t, are inferred from these phase differences. For example, in the figure, there are three time delays: Δ_1, Δ_2, and Δ_3, corresponding to three events (visible in the inputs $\mathbf{x}$ and $\mathbf{y}$). For instance, $\Delta_1 = 0.2 millisecond > 0$ indicates that the first event in $\mathbf{x}$ precedes the first event in $\mathbf{y}$ by 0.2 millisecond, while $\Delta_2 = -0.2 millisecond < 0$ means that the second event in $\mathbf{x}$ lags behind the second event in $\mathbf{y}$ by 0.2 millisecond. The entire process is trained end-to-end, minimizing the error between the predicted and actual time delays. During training, optimal Siamese network parameters p are learned by minimizing the difference between predicted and actual time delays: $p^* = \arg\min_p f(x, y; \Delta_t, p)$. During prediction, the time delay is estimated as: $\Delta t^* = \arg\min_{\Delta_t} f(x, y; \Delta t, p)$ for fixed parameters.

Our **Fig. 1** illustrates the NeuroTD approach. By combining a Siamese neural network with frequency domain transformation, NeuroTD enhances time series alignment with higher accuracy and efficiency. Starting with two time-series datasets, like neural activity and behavior data, NeuroTD handles cases where these datasets contain similar events but differing temporal patterns, complicating direct comparison. We input these datasets into a Siamese neural network, which maps them into a shared latent space, capturing their

core features. In this latent space, we apply a frequency transformation, such as the Short-Time Fourier Transform (Fourier Transform that uses a window function), to convert the signals into the frequency domain. This step leverages the Fourier shift theorem, which translates time delays into linear phase shifts, allowing for quick and robust computation of time delays. It uses Fast Fourier Transform to enhances computational speed. Additionally, the frequency domain makes it easier to incorporate biological constraints, like frequency response limits, than the time domain. Finally, by examining the phase differences in the transformed signals, NeuroTD computes time delays with resilience to noise, achieving a fast and accurate analysis of neural dynamics.

Overall, we use the following loss function to measure the shifts between $\mathbf{x}$ and $\mathbf{y}$, given the Siamese network $s(\cdot)$ and the shift Δ_t:

$$L(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\text{DFT}\{s(\mathbf{x}; \mathbf{p})\} - e^{j\boldsymbol{\delta}\Delta_t} \odot \text{DFT}\{s(\mathbf{y}; \mathbf{p})\}\|^2, \tag{1}$$

where $\boldsymbol{\delta} = \frac{2\pi}{N}[0, 1, ..., N-1]^T \in \mathbb{R}^N$, $\odot$ denotes element-wise multiplication, and DFT is the Discrete Fourier Transform. During training, we learn the optimal Siamese network parameters p by minimizing $\mathbf{p}^* = \arg\min_{\mathbf{p}} L(\mathbf{x}, \mathbf{y})$. During prediction, we estimate the time delay as $\Delta t^* = \arg\min_{\Delta t} L(\mathbf{x}, \mathbf{y})$ for fixed parameters $\mathbf{p}$.

In the next sections, we describe the details of each step in the NeuroTD approach.

2.1 Input - Multimodal Time-series Data

We consider two time-series datasets, $\mathbf{x}$ and $\mathbf{y}$, which may originate from the same or different modalities, such as neural activities and behavioral data. If these datasets have different lengths (i.e., number of sample points) or sampling rates, we preprocess the data by the following frequency adjustment.

Multimodal Frequency Adjustment: Let the two time-series $\mathbf{x}$ and $\mathbf{y}$ have different sampling rates f_x and f_y respectively, for example $f_x = 4000$ Hz using Neuropixels and $f_y = 120$ Hz using a motion camera, we need to preprocess them to a comparable sampling rate. Without loss of generality, we assume $f_x > f_y$ and upsample $\mathbf{y}$ to match the high sampling rate f_x: $y'[n] = \text{resample}(y[n], f_y, f_x)$. Instead of explicitly resampling the signals in the time domain, we use the following frequency-domain adjustment method. We compute the DFT and IDFT (inverse DFT) with zero-padding in the frequency domain to upsample $\mathbf{y}$ to

$$y'[n] = \text{real}(\text{IDFT}\{\text{pad}(\text{DFT}\{y\} \cdot \frac{f_x}{f_y})\}), \tag{2}$$

where $\frac{f_x}{f_y}$ is the scale factor to adjust the magnitudes of $\mathbf{y}$ to ensure it is comparable after the upsampling, *real* is the real part of a complex number, pad is zero-padding from N_y points to N_x points, and DFT and $IDFT$ are the Discrete Fourier Transform and Inverse Discrete Fourier Transform using N_y points and N_x points after zero-padding, respectively:

$$\text{DFT}\{y\} = \frac{1}{\sqrt{N_y}} \sum_{n=0}^{N_y-1} y[n] e^{-j2\pi kn/N_y}, \tag{3}$$

$$\text{IDFT}\{\hat{y}\} = \frac{1}{\sqrt{N_x}} \sum_{k=0}^{N_x-1} \hat{y}[k] e^{j2\pi kn/N_x}. \tag{4}$$

After the optional frequency adjustment, the two time-series $\mathbf{x}$ and $\mathbf{y}$ have the same length N_x, which we denote by N for simplicity.

2.2 Step 1 - Infer Common Time-Domain Latent Space by Siamese Network

We use Siamese Network to map the two time-series respectively $\mathbf{x}_{\text{latent}}$ and $\mathbf{y}_{\text{latent}}$ to as their latent representations $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ respectively:

$$\mathbf{x}_{\text{latent}} = s(\mathbf{x}; \mathbf{p}), \tag{5}$$

$$\mathbf{y}_{\text{latent}} = s(\mathbf{y}; \mathbf{p}). \tag{6}$$

The Siamese networks consist of two subnetworks with shared weights, where $s(\cdot)$ is the non-linear latent mapping using the Siamese neural network with parameters $\mathbf{p}$. Each subnetwork consists of the six fully connected layers, with ReLU activation functions applied after each hidden layer for non-linearity. The input time-series are transformed through shared weights to ensure a unified latent space, capturing the underlying similarities for subsequent analysis.

2.3 Step 2 - Transform to Frequency Space

We denote by $\hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathbb{C}^N$ the Discrete Fourier Transforms (DFT) of $\mathbf{x}_{\text{latent}}$ and $\mathbf{y}_{\text{latent}}$, respectively. We have

$$\hat{\mathbf{x}} = \text{DFT}\{s(\mathbf{x}; \mathbf{p})\}, \tag{7}$$

$$\hat{\mathbf{y}} = \text{DFT}\{s(\mathbf{y}; \mathbf{p})\}, \tag{8}$$

where $\text{DFT}\{\cdot\}$ maps any sequence $\mathbf{z} \in \mathbb{R}^N$ to its Fourier representation

$$\text{DFT}\{\mathbf{z}\} = \sum_{n=0}^{N-1} z[n] e^{-j2\pi kn/N}. \tag{9}$$

According to the Fourier Shift Theorem, if $\mathbf{x}$ is circularly shifted to the left by $t^\star$ samples, the DFT of the shifted sequence is $e^{j\boldsymbol{\delta}\Delta_t} \odot \hat{\mathbf{x}}$, where j is the imaginary unit such that $j^2 = -1$, vector $\boldsymbol{\delta} = \frac{2\pi}{N}[0, 1, ..., N-1]^T \in \mathbb{R}^N$, scalar $\Delta_t \in \mathbb{R}$ denotes a shift amount, and operator $\odot$ denotes element-wise multiplication.

Window-based Time-varying Alignment: To find the time-variant shifts, we break the time-series signal $\mathbf{x}$ into frames by applying a weighted window w of length M and get the mth windowed frame as:

$$x_m[n] \coloneqq x[n]w[n - mH], \tag{10}$$

where H denotes the hop size which is the number of samples by which each successive window advances, the overlap of successive windowed frames can be computed as $M - H$, and w is a window function such as Hanning:

$$w[n] = 0.5 - 0.5 \cos\left(\frac{2\pi n}{M-1}\right). \tag{11}$$

Applying the Short Time Fourier Transform (STFT) to $\mathbf{x}$, i.e., applying the Fourier transform to each windowed frame $x_m[n]$, we can obtain the complex-valued spectrograms:

$$\hat{x}[k, m] = \sum_{n=0}^{N-1} x[n]w[n - mH]e^{-j2\pi kn/N}, \tag{12}$$

where k represents the frequency bin, m represents the time frame

Let $x[n]$ and $y[n]$ be the two discrete time-series with the same length (N) and sampling rate (f_s), we use the following steps to compute the time-variant shifts between them.

Calculate the STFT of both time-series:

$$\hat{x}[k,m] = \sum_{n=0}^{N-1} x[n]w[n-mH]e^{-j2\pi kn/N}, \tag{13a}$$

$$\hat{y}[k,m] = \sum_{n=0}^{N-1} y[n]w[n-mH]e^{-j2\pi kn/N}. \tag{13b}$$

2.4 Step 3 - Calculate Phase Difference

Extract the phase difference between the two time-series to find the shifts. Instead of directly to compute the phase difference between the two time-series at each frequency bin k and time frame m: $\theta_1[k,m] = \arg(\hat{x}[k,m]), \theta_2[k,m] = \arg(\hat{y}[k,m]), \Delta\theta[k,m] = \theta_2[k,m] - \theta_1[k,m]$, we compute the phase of the conjugate multiplication

$$\Delta\theta[k,m] = \arg(\hat{c}_{1,2}[k,m]), \tag{14}$$

where

$$\hat{c}_{1,2}[k,m] = \hat{x}[k,m]\hat{y}[k,m]^* \tag{15}$$

is the conjugate multiplication of the two STFT.

2.5 Output - Time Delays

We can use the phase difference to compute the time shift between the two time-series for each frequency component k and time frame m:

$$n_{\text{shift}}[k,m] = \frac{\Delta\theta[k,m]}{2\pi k/N}. \tag{16}$$

We perform a weighted least square fitting to compute a time shift to aggregate all the frequency components at each time frame m, where the error function is defined as:

$$E(\Delta t[m]) = \sum_{k=1}^{K} w[k]\left(\Delta\theta[k,m] - 2\pi k\frac{\Delta t[m]}{N}\right)^2, \tag{17}$$

where $E(\Delta t[m])$ is the error function $w[k]$ is the weight applied to each frequency component which can be set as inversely proportional to noise or variance at that frequency.

To find the optimal $\Delta t[m]$, we differentiate $E(\Delta t[m])$ with respect to $\Delta t[m]$ and set it to zero:

$$\frac{\partial E(\Delta t[m])}{\partial \Delta t[m]} = -2\sum_{k=1}^{K} w[k]\cdot\left(\Delta\theta[k,m] - 2\pi k\frac{\Delta t[m]}{N}\right)\cdot\frac{2\pi k}{N} = 0. \tag{18}$$

The time delays at time frame m can be computed as:

$$\Delta t[m] = \frac{\sum_{k=1}^{K} w[k]\cdot\Delta\theta[k,m]\cdot k}{\sum_{k=1}^{K} w[k]\cdot\left(\frac{2\pi k}{N}.\right)^2} \tag{19}$$

2.6 Evaluation

To evaluate the performance of NeuroTD, we use the root-mean-square error (RMSE) as the primary metric to measures the difference between the predicted time delay Δt^* and actual time delay Δt: RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta t_i - \Delta t_i^*)^2}$, where N is the number of samples in the dataset. A lower RMSE indicates a better alignment between the time-series signals and more accurate time delay estimation. We also provide a

confidence score of the computed time-shift, which is the root-mean-square error of the two time-series signals after alignment using the computed shifts. A lower confidence score indicates a more reliable time-shift estimation. This can be computed using in the time domain representation of the two signals:

$$\text{Confidence} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |x[n] - y[n - \Delta t^*]|^2}. \tag{20}$$

However, in case of subsample accuracy where Δt^* involves fractional of sample intervals, we can use the following formula, we can use the frequency domain representation of the two signals:

$$\text{Confidence} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\hat{x}[n] - \hat{y}[n]e^{-j2\pi n \Delta t^*/N}|^2}. \tag{21}$$

2.7 Implementation and Hyperparameters

The NeuroTD algorithm is implemented in Python using the PyTorch [13]. The Siamese neural network is constructed using fully connected layers with ReLU activation functions. The frequency domain transformation is performed using the SciPy FFT module. The algorithm is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The training process involves minimizing the loss function defined in Eq. (1) to learn the optimal parameters of the Siamese network. The algorithm is trained for 100 epochs on the simulated dataset and 500 epochs on the real-world datasets to ensure convergence. The time delays are computed using the phase differences between the frequency domain representations of the signals, as described in Section 2.3.

2.8 Datasets and Data Processing

We use four types of datasets. First, a simulated dataset with time-varying delays was generated to benchmark NeuroTD against existing methods, ensuring model validation under controlled conditions. The remaining three are real-world multimodal datasets: (1) a neural circuits dataset that includes electrophysiological (ephys) recordings from 32 neurons in the human hippocampus, measured using depth electrodes, with their spatial positions recorded for mapping time delays and signal transmission pathways [6]; (2) a neuron-behavior dataset from the rat neocortex, where neural activities were captured via Neuropixels and behavioral data, specifically head movement, was recorded through 3D motion capture [14]; and (3) an ephys and genes dataset, which contains electrophysiological recordings and gene expression data from neurons in the mouse motor cortex collected using Patch-seq [15]. Before applying NeuroTD, we preprocess all datasets by aligning the time-series signals and normalizing them to have zero mean and unit variance.

3 Results

3.1 Simulation Study

In this simulation study, we generate two signals with dynamically changing time delays to evaluate NeuroTD's effectiveness. **Fig. 2A** illustrates simulated signals with increasing time delays, starting at 1.56 millisecond and extending to 3.32 millisecond by the end. To replicate realistic experimental conditions, we add Gaussian noise at ten levels, with standard deviations (std) ranging from 0 to 0.20 (linearly spaced), simulating challenges typical in neural recordings. We repeat simulation for each noise level 100 times to ensure statistical robustness.

The results in **Fig. 2B** and **Fig. 2C** demonstrate that NeuroTD consistently provides accurate alignment and time-delay estimation, even as noise levels increase. Specifically, in **Fig. 2B**, when tested at a fixed noise level (std=0.20), NeuroTD achieves a root-mean-square error (RMSE) of 0.16 millisecond, demonstrating high precision in alignment with the ground truth. In contrast, the Discrete Time Warping (DTW) method

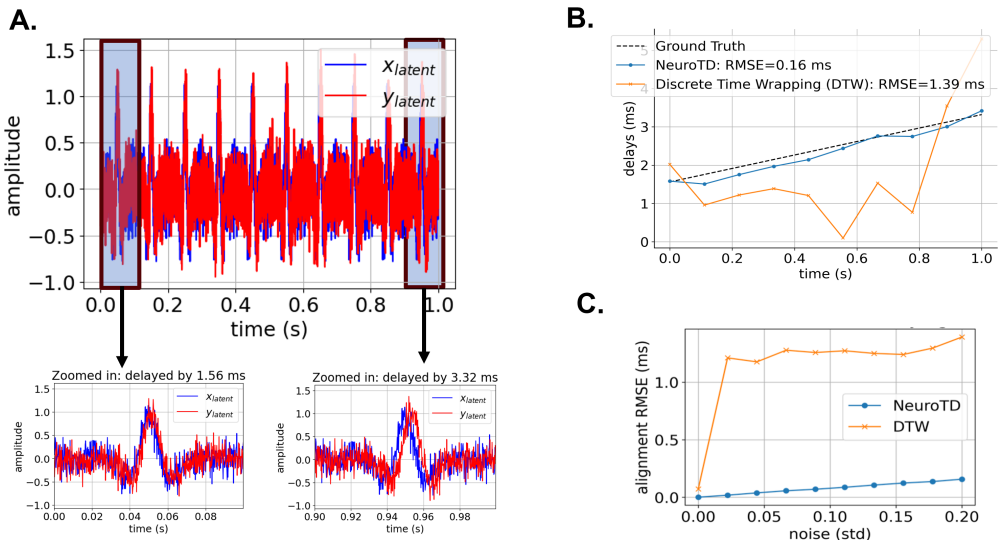

Fig. 2. Simulation Study. A. Simulated Time-Varying Delays: We simulate two signals, x_{latent} (blue) and y_{latent} (red), where the time delay changes dynamically throughout the duration. The zoomed-in sections illustrate specific delays of 1.56 millisecond at the start and 3.32 millisecond at the end, respectively. We add Gaussian noise at various levels to replicate real-world measurement conditions. **B. Performance Comparison at a Fixed Noise Level:** We compare the performance of NeuroTD and Discrete Time Warping (DTW) under a given noise level. The x-axis shows the time (s), and the y-axis shows the delay over time. NeuroTD (blue) closely aligns with the ground truth (black) across all time points, achieving an RMSE of 0.16 millisecond, while DTW (orange) displays significantly higher errors, with an RMSE of 1.39 millisecond. **C. Performance Under Various Noise Levels:** We evaluate NeuroTD and DTW across increasing noise levels, with NeuroTD consistently maintaining a low RMSE, demonstrating its robustness.

shows significantly larger errors, with an RMSE of 1.39 millisecond. In **Fig. 2C**, further evaluation under escalating noise levels confirms that NeuroTD outperforms DTW, maintaining a lower RMSE across conditions and demonstrating superior robustness in handling noisy neural data.

3.2 Application I: Time Delays of Neural Activities associated with 3D Positions of Human Hippocampal Neurons

In this application, we analyze dataset from Boran et al.[6] that records electrophysiological (ephys) activities in the human hippocampus using depth electrodes. The data includes time-series measurements from 32 neurons individually and simultaneously measured over hours. Each neuron is recorded at distinct spatial positions, allowing for spatial mapping of neural activities and time delays between the recorded signals.

Fig. 3A shows a 10 millisecond segment from 16 of the neurons, providing a detailed zoomed-in view of neural activities over this time window. The heatmap in **Fig. 3B** shows the time delays between all pairs of 32 neurons, with delays ranging from -20 milliseconds to 20 milliseconds. We computed these delays using NeuroTD, which allowed for precise detection even under complex scenarios. In **Fig. 3C**, we selected eight neurons with lower alignment errors to visualize their spatial configuration and the corresponding time delays. The 3D plot displays these neurons' positions within the hippocampus, where the lines connecting the neurons represent the delay intervals. The arrows show the inferred directionality of signal transmission, suggesting a potential signal transmission pathway where signals propagate from the initial set of neurons (1 and 2) to the intermediate set (3, 4, 5) and eventually to the final set (6, 7, 8). These results not only provide insights into localized signal propagation within a specific neural network but also demonstrate the potential of scaling this analysis to include more neurons. Extending this approach to larger networks could yield a detailed map of neural signal transmission across extensive brain circuits, helping researchers better understand how neural pathways and circuits function within the brain's architecture.

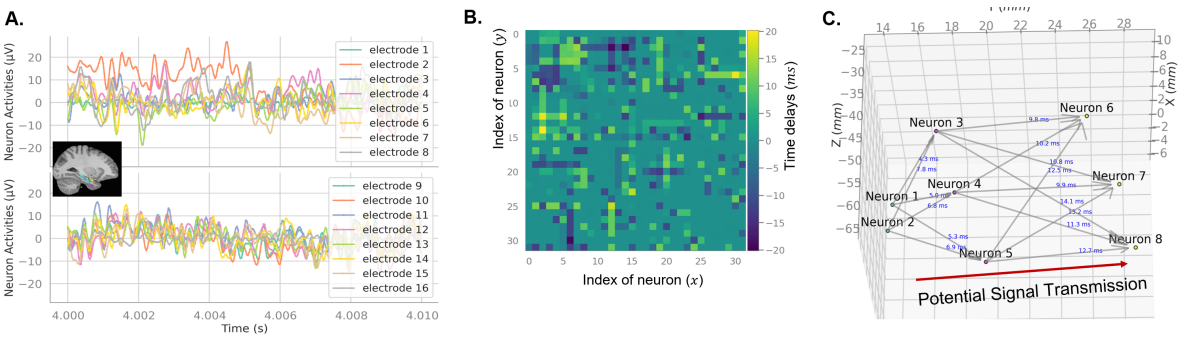

Fig. 3. Application I: Time Delays of Neural Activities associated with 3D Positions of Human Hippocampal Neurons. A. Activities of 16 out of 32 neurons by depth electrodes [6]: The time series data recorded by depth electrodes in the human hippocampus. We only show part of the data, 10 millisecond of recording for 16 neurons at different positions, though the recording have 32 neurons and can span several hours. **B. Time delays among 32 neurons**: Heatmap showing the time delays between each pair of 32 neurons. The intensity of the colors indicates the magnitude of the time delays—computed by NeuroTD. **C. Time delays vs. electrode positions across selected neurons**: From the previous heatmap of 32 neuron pairs, we selected 8 neurons with lower alignment errors. This 3D plot shows the spatial positions of these neurons recorded by depth electrodes. Each neuron is labeled, and the lines connecting them represent the time delays between this pair, with arrows showing the direction of signal transmission.

3.3 Application II: Time Delays Suggest Causality between Neurons and Behaviors in Rat Neocortex

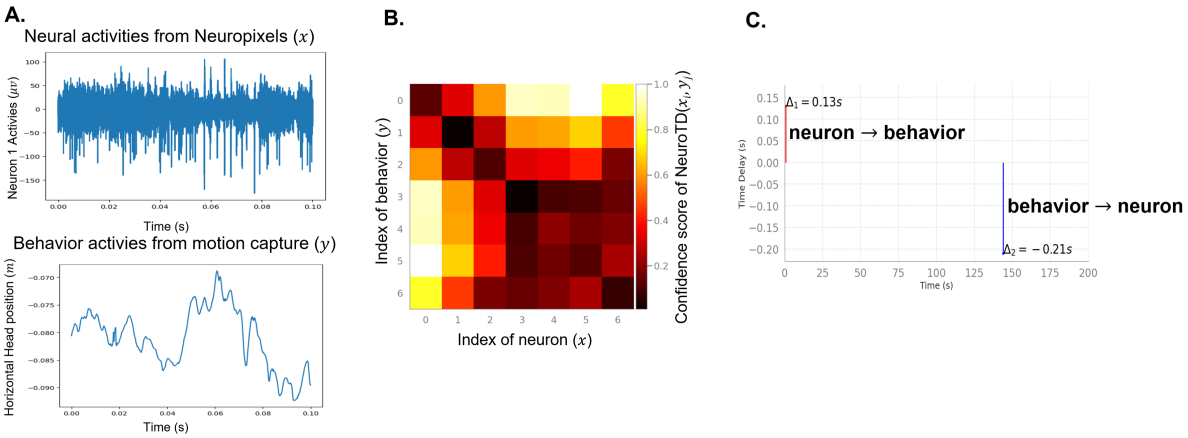

Fig. 4. Application II: Time Delays Suggest Causality between Neurons and Behaviors in Rat Neocortex. A. Neural and behavioral activities in rat neocortex: The neural activity is recorded from Neuropixels (top plot), showing a time series of neuron spikes over time (x), while the behavior is captured through motion capture, showing head position movements over time (y) [14]. **B. Time delays confidence map**: Heatmap displaying the confidence scores of time delays between each neural activity (index x) and behavior activity (index y), computed using NeuroTD. **C. Causality from selected neuron and behavior**: The plot shows the time delays for a selected highly confident neuron and behavior pair, with positive values indicating that the neuron activity leads behavior (e.g., $\Delta_1 = 0.13$ seconds), and negative values showing that behavior leads neuron activity (e.g., $\Delta_2 = -0.21$ seconds).

In this application, we used a dataset from Mimica et al. [14] that involves detailed neural and behavioral recordings from rats implanted with Neuropixels probes (version 1.0, IMEC, Belgium). The dataset contains electrophysiological (ephys) recordings sampled at 32 kHz from the rat neocortex, targeting primary sensory and motor cortices. In addition to the neural data, behavioral activity, specifically head position and

movement, was captured using motion capture technology with a sampling rate of 120 Hz. This multimodal dataset allows us to investigate the relationship between neural activities and physical behaviors, providing insights into the causal interactions between the brain and body movements. Because the two modalities have different lengths sampling rates, we preprocess the data by the multimodal frequency adjustment method in Section 2.1.

Figure **Fig. 4A** provides a segment of the time-series data, illustrating how both neural spikes and head position movements evolve over time. NeuroTD then computes time delays between these two modalities, as shown in the heatmap in **Fig. 4B**. The color intensity in this heatmap correlates with the confidence score of each time delay, highlighting the strongest and most reliable causal relationships. Quantitatively, we identified that highly confident neuron-behavior pairs often demonstrated time delays with specific values. For instance, **Fig. 4C** illustrates a case where neuron activity leads behavior with a delay of $\Delta_1 = 0.13$ seconds. In contrast, instances where behavior precedes neuron activity show delays like $\Delta_2 = -0.21$ seconds. These results indicate that specific neural activities may either precede or follow physical actions, suggesting potential causative mechanisms where certain neural spikes might initiate or respond to particular behavioral changes. This analysis provides a quantitative understanding of how neural dynamics and behaviors are interconnected, emphasizing the brain's role in controlling and responding to physical actions in a real-time context.

3.4 Application III: Correlating Genes with Electrophysiological Time-delay Features in Mouse Motor Cortex

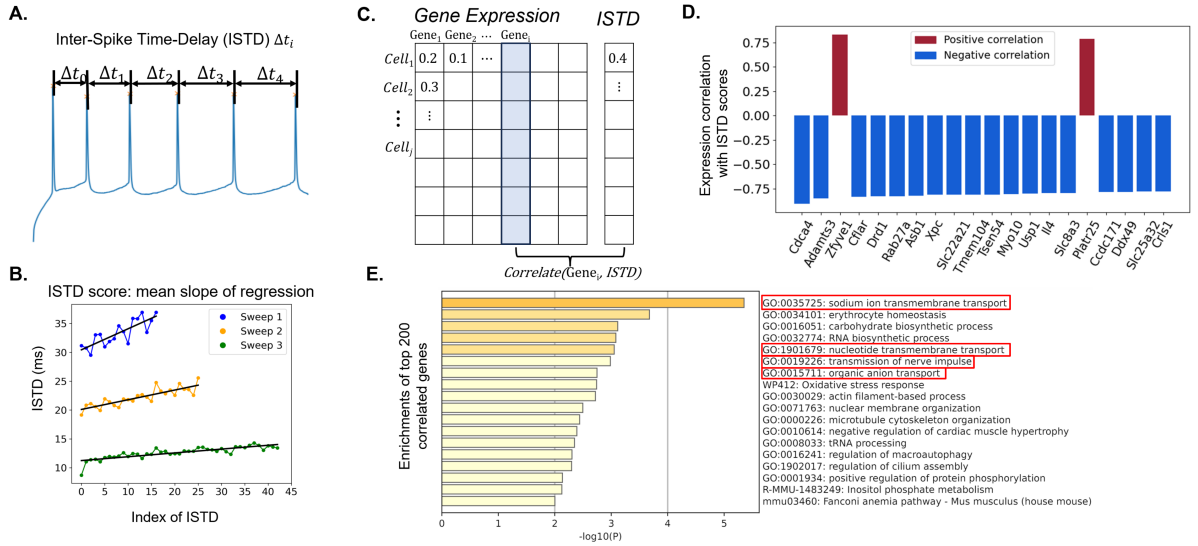

Fig. 5. Application III: Correlating Genes with Electrophysiological Time-delay Features in Mouse Motor Cortex. A. Inter-Spike Time-Delay (ISTD): The ephys signal during a sweep, showing varying time gaps between consecutive spikes (Δt_i) over time. **B. ISTD Score: Mean Slope of Regression:** Mean slopes of the changing ISTD for each neuron across multiple sweeps. Three sweeps are plotted, each with a different change rate of time-delay. **C. Gene Expression versus Time-delay Features:** Heatmap showing gene expression levels across different cells (top) and their corresponding ISTD scores (bottom). Each column corresponds to a specific cell. **D. Correlation between Gene Expression and ISTD Scores:** Bar chart showing the top 20 genes most strongly correlated with ISTD scores. Red bars represent positive correlations, while blue bars represent negative correlations. **E. Enrichments of Top 200 Correlated Genes:** Enrichment analysis of the top 200 genes (100 positive, 100 negative correlations) showing the most significantly enriched Gene Ontology (GO) terms. The key terms associated with electrophysiological activity are highlighted, including sodium ion transmembrane transport, nucleotide transmembrane transport, transmission of impulse transmission, and organic anion transport.

In this application, we examine the relationship between time delays in electrophysiological responses and gene expression in the mouse motor cortex using data from Scala et al. [15]. **Fig. 5A** illustrates the inter-spike time delays (ISTDs), or the time gaps between consecutive spikes, recorded during neural activity sweeps. These gaps change in response to neural stimuli and indicate how neurons adjust their firing patterns. **Fig. 5B** quantifies these changes by computing the slope of the ISTD shifts over multiple sweeps. The slopes reflect how quickly or slowly the time delays evolve, offering insights into the dynamics of neuronal firing patterns. The ISTD score for each neuron is computed as the mean slope across sweeps, capturing robust neural dynamics. As illustrated in **Fig. 5C**, we correlate the ISTD scores with gene expression data across many cells. **Fig. 5D** reveals the top 20 genes that show the strongest correlations—either positive or negative—with the time-delay features. This correlation helps identify genes potentially associated with rapid changes in neural response, providing a genetic link to neural firing dynamics. We performed an enrichment analysis of the top 200 correlated genes (**Fig. 5E**), which highlighted several key biological processes relevant to electrophysiological signal transmission. The following Gene Ontology (GO) terms are most closely related to neural electrophysiological activities: GO:0035725: sodium ion transmembrane transport (p-value $10^{-5.7}$) – Sodium ion channels are essential for generating action potentials, crucial for nerve signal transmission. GO:1901679: nucleotide transmembrane transport (p-value $10^{-3.5}$) – This process is involved in moving nucleotides across membranes, indirectly influencing cellular energetics and signaling in neurons. GO:0019226: transmission of nerve impulse (p-value $10^{-2.9}$) – Directly related to the propagation of electrical signals through neurons, fundamental to neural communication. GO:0015711: organic anion transport (p-value $10^{-2.5}$) – Organic anions like chloride are vital for maintaining membrane potential and neural excitability. These processes are critical to neuronal excitability and signal transmission, underscoring the link between gene expression and electrophysiological behavior.

4 Discussion

In this work, we introduced NeuroTD, a time-frenquency based multimodal deep learning approach designed to analyze time delays in neural activities across various scales. The approach integrates Siamese neural networks with frequency domain transformations to infer temporal relationships. Our results highlight several key findings: NeuroTD effectively captures time delays between neural signals, identifies neuron-behavior causalities, and correlates gene expression with electrophysiological timing, offering significant insights into the neural circuits and their functional dynamics.

While NeuroTD demonstrates its value across multiple modalities, certain limitations remain. The current approach has yet to be validated extensively through experimental data across different conditions. Future iterations of NeuroTD may improve prediction accuracy by incorporating biological constraints, such as frequency band limits, into a Bayesian framework. Another improvement is to calculate delays use all neurons simultaneously, rather than just one pair at a time, enhancing accuracy and robustness. Additionally, we can improve computational efficiency. Most neural recordings contain long periods of inactivity. Optimizing NeuroTD by prefiltering inactive time periods highly likely reduce computation time. Finally, the approach can be expanded to align multidimensional datasets such as EEG (electroencephalography) and fMRI.

5 Acknowledgments, Funding, Data and Code Availability

This work is supported by National Institutes of Health grants, RF1MH128695, R01AG067025, R21NS128761 , R21NS127432, P50HD105353, and National Science Foundation Career Award 2144475. All data and code are available at `https://github.com/daifengwanglab/NeuroTD`.

References

[1] BL Sabatini and WG Regehr. "Timing of synaptic transmission". In: *Annual review of physiology* 61.1 (1999), pp. 521–542.

[2] Marcela Lipovsek, Cedric Bardy, Cathryn R Cadwell, Kristen Hadley, Dmitry Kobak, and Shreejoy J Tripathy. "Patch-seq: Past, present, and future". In: *Journal of Neuroscience* 41.5 (2021), pp. 937–946.

[3] Liqun Luo. *Principles of neurobiology*. Garland Science, 2020.

[4] Sven Schrader, Sonja Grun, Markus Diesmann, and George L Gerstein. "Detecting synfire chain activity using massively parallel spike train recording". In: *Journal of neurophysiology* 100.4 (2008), pp. 2165–2176.

[5] Christine Grienberger and Arthur Konnerth. "Imaging calcium in neurons". In: *Neuron* 73.5 (2012), pp. 862–885.

[6] Ece Boran, Tommaso Fedele, Peter Klaver, Peter Hilfiker, Lennart Stieglitz, Thomas Grunwald, and Johannes Sarnthein. "Persistent hippocampal neural firing and hippocampal-cortical coupling predict verbal working memory load". In: *Science advances* 5.3 (2019), eaav3687.

[7] Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings". In: *Science* 372.6539 (2021), eabf4588.

[8] Toni Giorgino. "Computing and visualizing dynamic time warping alignments in R: the dtw package". In: *Journal of statistical Software* 31 (2009), pp. 1–24.

[9] Xiaowei Zhuang, Zhengshi Yang, and Dietmar Cordes. "A technical review of canonical correlation analysis for neuroscience applications". In: *Human brain mapping* 41.13 (2020), pp. 3807–3833.

[10] Feng Zhou and Fernando Torre. "Canonical time warping for alignment of human behavior". In: *Advances in neural information processing systems* 22 (2009).

[11] Xiaohua Tong, Zhen Ye, Yusheng Xu, Sa Gao, Huan Xie, Qian Du, Shijie Liu, Xiong Xu, Sicong Liu, Kuifeng Luan, et al. "Image registration with Fourier-based image correlation: A comprehensive review of developments and applications". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.10 (2019), pp. 4062–4081.

[12] Jianren Wang, Zhaoyuan Fang, and Hang Zhao. "Alignnet: A unifying approach to audio-visual alignment". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3309–3317.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[14] Bartul Mimica, Tuçe Tombaz, Claudia Battistin, Jingyi Guo Fuglstad, Benjamin A Dunn, and Jonathan R Whitlock. "Behavioral decomposition reveals rich encoding structure employed across neocortex in rats". In: *Nature Communications* 14.1 (2023), p. 3947.

[15] Federico Scala, Dmitry Kobak, Matteo Bernabucci, Yves Bernaerts, Cathryn René Cadwell, Jesus Ramon Castro, Leonard Hartmanis, Xiaolong Jiang, Sophie Laturnus, Elanine Miranda, et al. "Phenotypic variation of transcriptomic cell types in mouse motor cortex". In: *Nature* 598.7879 (2021), pp. 144–150.