Decoding Spatio-Temporal Patterns in Biomolecular Simulations with Wavelet Analysis and other Time-series Analysis methods

Thesis Report

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

Master of Science

by

Kushwah Amardeepsingh Harishchandra - 19MS165

Supervised by

Prof. Neelanjana Sengupta

Prof. P. K. Panigrahi

IISER KOLKATA

Department of Biological Science

IISER Kolkata

May 2024

Declaration

I, Mr. Kushwah Amardeepsingh Harishchandra, Registration No. 19MS165 dated 31-07-2019, a student of the Department of Biological Sciences of the 5 Year BS-MS Dual Degree Programme Programme of IISER Kolkata, hereby declare that this thesis is my own work and, to the best of my knowledge, it neither contains materials previously published or written by any other person, nor has it been submitted for any degree/diploma or any other academic award anywhere before. I have used the originality checking service to prevent inappropriate copying.

I also declare that all copyrighted material incorporated into this thesis is in compliance with the Indian Copyright Act, 1957 (amended in 2012) and that I have received written permission from the copyright owners for my use of their work.

I hereby grant permission to IISER Kolkata to store the thesis in a database which can be accessed by others.

Amardeepsingh Harishchandra Kushwah
Department of Biological Sciences
Indian Institute of Science Education and Research Kolkata
Mohanpur 741246, West Bengal, India

Certificate

Date: 17th May 2024

This is to certify that the thesis titled ‘Decoding Spatio-Temporal Patterns in Biomolecular Simulations with Wavelet Analysis and Other Time-series Analysis Methods’ submitted by Mr./Ms. Kushwah Amardeepsingh Harishchandra, Registration No. 19MS165 dated 31-07-2019, a student of the Department of Biological Sciences of the 5 Year BS-MS Dual Degree Programme Programme of IISER Kolkata, is based upon his/her own research work under my supervision. I also certify, to the best of my knowledge, that neither the thesis nor any part of it has been submitted for any degree/diploma or any other academic award anywhere before. In my opinion, the thesis fulfills the requirement for the award of the Masters Degree.

Neelanjana Sengupta

Neelanjana Sengupta
Associate Professor
DBS, IISER Kolkata
Mohanpur 741246, West Bengal, India

Prasanta K. Panigrahi

Prasanta K. Panigrahi
Professor
DPS, IISER Kolkata
Mohanpur 741246, West Bengal, India

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me throughout this journey.

First and foremost, I am incredibly grateful to my advisors, Professor Neelanja Sengupta and Professor Prasanta Panigrahi, for their invaluable guidance and support throughout the year. Their expertise and encouragement were instrumental in shaping this thesis.

I would also like to thank Sarbajit Layek, Pallab Dutta, and Hindol Chatterjee for providing the data essential for my analysis and for their continued assistance. Additionally, I am thankful to Manali di, Dheerendra bhaiyya, and Neeraj bhaiyya for their unwavering support and guidance, both academically and personally.

The stimulating discussions with the members of the NS and PKP labs fostered a positive research environment, and I am grateful for their contributions.

Finally, this thesis would not have been possible without the unwavering support of my friends at IISER Kolkata, including Tushar, Parth, Abhiroop, Yusuf, Aritra, Abhay, Saikat, Balwinder, and an endless list of cherished people of IISER K community. Their presence throughout my journey has been a source of immense strength.

I am eternally grateful to Mummy, Papa, and Monu for their love, encouragement, and unwavering belief in me, which has inspired me to pursue my Dreams.

To my loving family

Prologue

This project was not the first idea I explored. It was rather the result of a chain of events/exploration that led to the finalization of the topic for the thesis. I wish readers to know this background to understand the motivation behind pursuing this topic.

The current topic was inspired by the exploration I have been involved in for the last year. It started with applying Wavelet Analysis to study the Sunspot Number time series[21]. This was a crucial learning experience for deciding my initial thesis topic, 'Using surface EEG data to reconstruct real-time 3D reconstruction of the electrical activity of the brain'. Where I planned to use the higher spatio-temporal resolution of the components of the EEG signal at different scales as the initial input signal for the reconstruction, during this time, I was also looking into other properties of the EEG signals like multifractality and presence of Unstable Periodic Orbits, which are known to show usefulness in the detection of changes in brain activity such as an onset of epilepsy [25].

During this time, I learned how the higher time-frequency resolution that Wavelet Analysis provides can be useful in the detection of significant events in a time series [8]. Here, I also got to understand how random processes happening at different scales at the same time can be characterized using the Hurst exponent. This helps in understanding the fractal nature of random processes (persistent and Antipersistant nature of noise) happening at different time scales.[26]

At this point, I would especially like to acknowledge Prof. Neelanjana Sengupta for her valuable insight into how these analyses hold potential in processing molecular dynamics data. Initial analysis showed some interesting trends, which, combined with a lack of resources and progress in my previous endeavor, led to a change in the course of exploration.

With this said, I would like to remark that it would be more appropriate to look at this thesis as an exploration rather than a project with some expected result.

Abstract

With the advent of newer computational technologies, Molecular Dynamics simulations are producing an ever-increasing amount of data. Longer time series, along with the random nature of the fluctuations observed in the data, makes it harder to identify events of interest from the noise. Hence, developing an analysis technique that efficiently identifies these events becomes necessary.

We plan to study relatively recently formalized techniques for analyzing Time Series data - Wavelet Analysis (WA). WA provides an excellent tool for identifying significant motions in the time series. One of the essential aspects of the WA is identifying appropriate background spectra to which the atoms' motion can be compared to determine their significance. This background is generally assumed to be a white or colored noise.

Here, we show that the simple assumption about a fixed fractal dimension of background at all scales may not hold. We show that these time series have a multifractal background and try to understand how this background changes with the simulation method. Hence, it is necessary to consider multifractal noise as a background for this wavelet analysis.

Contents

Chapter 1

Introduction

With the coming of newer technologies, the simulation runs are getting longer and longer, like one by D.E. Shaw research with simulation period in order of milliseconds [13]. We do not know beforehand when the event of interest will occur in this long trajectory. To tackle this problem, techniques such as RMSD (Root Mean Square Deviation) and RMSF (Root Mean Square Fluctuation) of the coordinates relative to the initial or some reference frame have been developed. These methods come with their own drawbacks. RMSD tells us about the deviation of complete structure from the reference frame on each time step, but it does not tell us where exactly these fluctuations are occurring. Similarly, the RMSF gives an idea of how much a region fluctuates throughout the period of simulation, but it does not tell us when these fluctuations occur.

As we will see in the further sections, simply breaking a time series into smaller chunks does not give optimal results due to underlying time-frequency uncertainty. As shown in multiple studiesWavelet analysis could be an effective tool in overcoming the optimization of this problem[23, text][20]. Wavelet analysis has been formalized with notable contributions from Daubechies in the late 1980s [4]. This analysis still lacked an understanding of the significance of each data point in the results. This was overcome in the method proposed by Torrence and Compo in 1998, which suggests a method that gives a quantitative measure of the significance of each data point in the analysis.

One of the important aspects of the WA is identifying appropriate background spectra to which the atoms' motion can be compared to determine their significance. All of these studies consider a single-colored noise as their background.[8, 20, 23, text]

We plan to use relatively recently formalized techniques for analyzing Time Series data - Wavelet Analysis (WA). WA provides an excellent tool for identifying significant motions in the time series. One of the important aspects of the WA is identifying appropriate background spectra to which the atoms' motion can be compared to determine their significance. This background is generally assumed to be a white or colored noise that has a fixed fractal dimension.

Here in this study, we use water molecules as a probe to develop an understanding of this expected background. We show that the simple assumption about a fixed fractal dimension of background at all scales may not hold. We show that these time series have a multifractal background and try to understand how this background changes with the simulation method. Hence, it is more appropriate to consider multifractal noise as a background for this wavelet analysis.

In this chapter we set out to probe how the methods used for Time-Series Analysis (TS Analysis) compare to traditionally available methods in popular Molecular Dynamics (MD) Analysis tools. We will mainly look at the Wavelet Analysis (WA) and its application in finding significant events in long MD Trajectories by decomposing fluctuations into oscillations at different scales. In this chapter, we look at the WA technique and its applications.Then we look at a short review of Molecular Dynamics (MD) simulations and few important component. We will also brief look at the Hurst exponent and its relation to the fractal dimension of the noise.

1.1 Fourier Transform: The Mathematical prism

It is a widely used technique to convert the signal from the time domain to the frequency domain and vice versa. We review this method and then look at its potential shortcomings and how they can be mitigated.

Fourier Transform finds its origins in the idea that any periodic function can be represented as the sum of scaled sines and cosine functions, better known as Fourier series expansion. The Fourier transform is an expansion of this idea to non-periodic functions by assuming the period of the function to be infinite.

The Fourier transform (Tf) for function f(t) is given by

$$(Tf)(\omega) = \frac{1}{\sqrt{2\pi}} \int f(t) e^{-i\omega t} dt \tag{1.1}$$

Where ω is the frequency, t is time and i is $\sqrt{-1}$. This way, we can decompose any function into its constituent frequencies. One downside of this method is that we do not have a time resolution for the frequencies and can only tell which frequencies were present and not when.

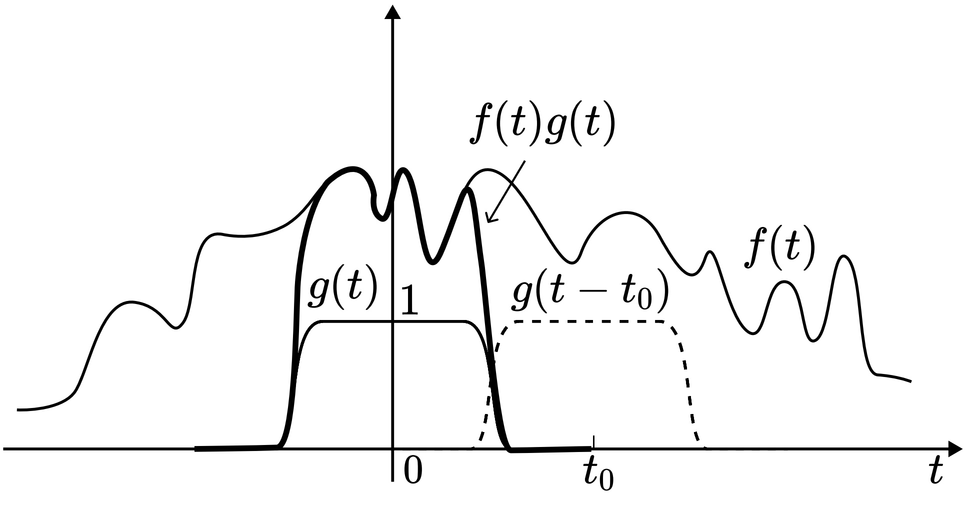

Figure 1.1: The function $f(t)$ is convoluted with the window function $g(t)$ then the Fourier coefficients of $f(t) \cdot g(t)$ is calculated. This procedure is repeated for translated versions of the window function $g(t-t_0)$, $g(t-2t_0)\ldots$

One of the methods for achieving better time resolution is windowing the signal and then applying the Fourier transform. So the function $f(t)$ is convoluted with the window function $g(t)$, which results in time localized slice $f(t) \cdot g(t)$. Then, we apply the Fourier transform to get frequency information for this slice. This is known as Windowed Fourier Transform or WFT in short.

$$(T^{win}f)(\omega, t') = \frac{1}{\sqrt{2\pi}} \int dt f(t) \cdot g(t - t')e^{-i\omega t} \tag{1.2}$$

A widely used discretized version of WFT uses regularly spaced t and ω with values $t = nt_0$, $\omega = m\omega_0$, where $m, n \in \mathbb{Z}$ and fixed $\omega_0, t_0 > 0$. Then we get the discretized version of equation 1.2 as:

$$T^{win}_{m,n}(f) = \frac{1}{\sqrt{2\pi}} \int dt f(t) \cdot g(t - nt_0)e^{-im\omega t} \tag{1.3}$$

The window function $g(t)$ can be chosen with a reasonable set of properties so as to preserve the properties of the original function. A popular choice for $g(t)$ is a Gaussian function, which is well concentrated in both time and frequency domain. Additionally, if the window function is well concentrated around zero then the $(T^{win}f)(\omega, t')$ can be interpreted as the content of $f(t)$ near the time t' as per 1.2. Here we translate the window by changing the value of n to span the signal in its entirety as seen schematically from the figure 1.1.

Now, let's look at an example where the frequency content of a signal varies over time figure: 1.2. The Fourier Transform will only tell us about the overall frequency content for the entire duration. Whereas in the WFT, we can see how the frequency content is shifting from lower frequencies to higher frequency. The figure produced here is called spectrogram which tells us about the spectral content of the signal.

Here we should note that we cannot have precise both time and frequency content of the signal simultaneously. This is due to underlying time-frequency uncertainty in nature $\Delta t \cdot \Delta\omega \geq \frac{1}{2}$ [5, text]. Here, the lower bound of the frequency depends on the size of the window used and the upper bound is given by *Nyquist limit.* Hence, the more time resolution we try to achieve by using a smaller time window the more we lose resolution at lower frequencies. This point can be seen in action in figure 1.3

This method is not ideal for nonstationary signals due to the absence of one fixed *"ideal"* window size. In the next section, we shall take a look at a

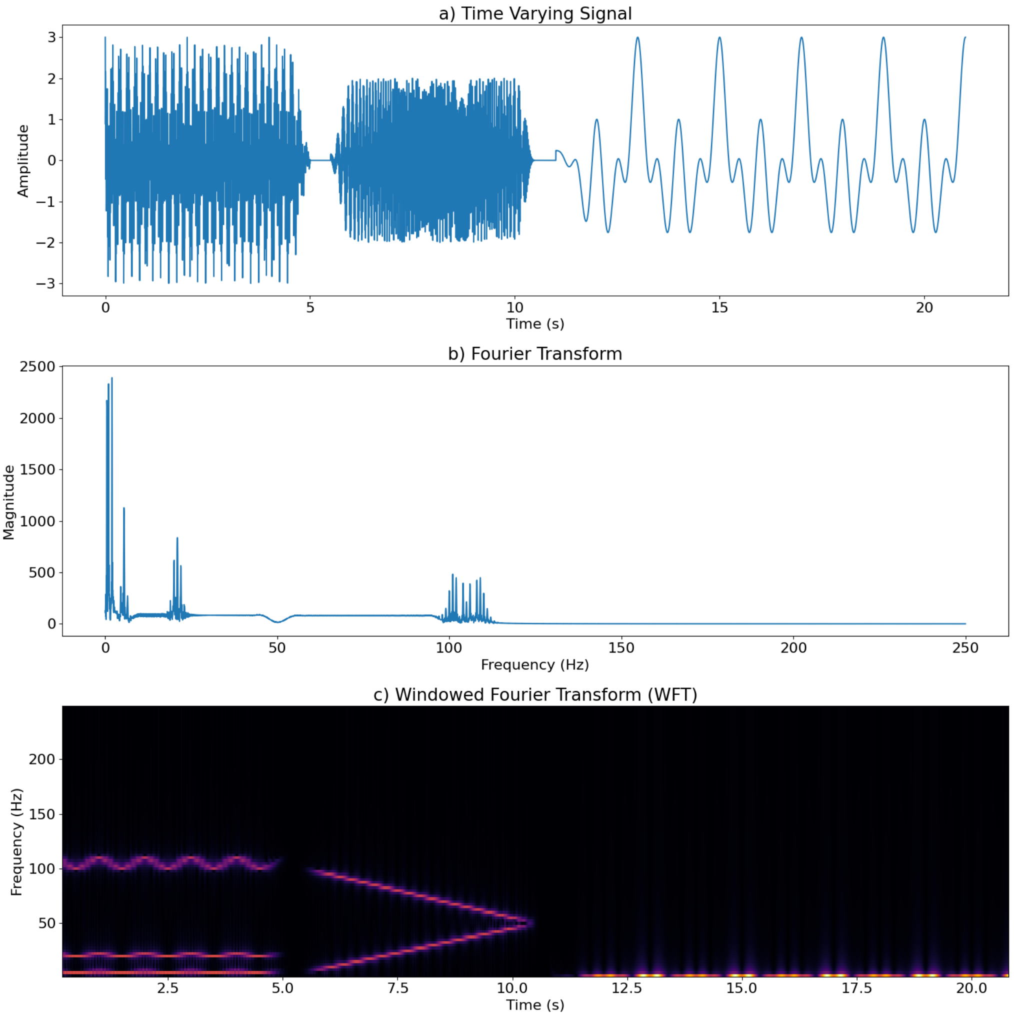

Figure 1.2: *a*) A signal with non-stationary frequency content with amplitude in arbitrary units *b*) Here, we can see the Fourier transform of a non-stationary signal (*a*) tells us only about the frequency content that was ever present during the entire duration. *c*) We use a rectangular window function for WFT where we can clearly see the intricate fashion in which the frequency content of the signal varies.

method that can overcome this uncertainty by using appropriate scaling in different regions of the time-frequency domain.

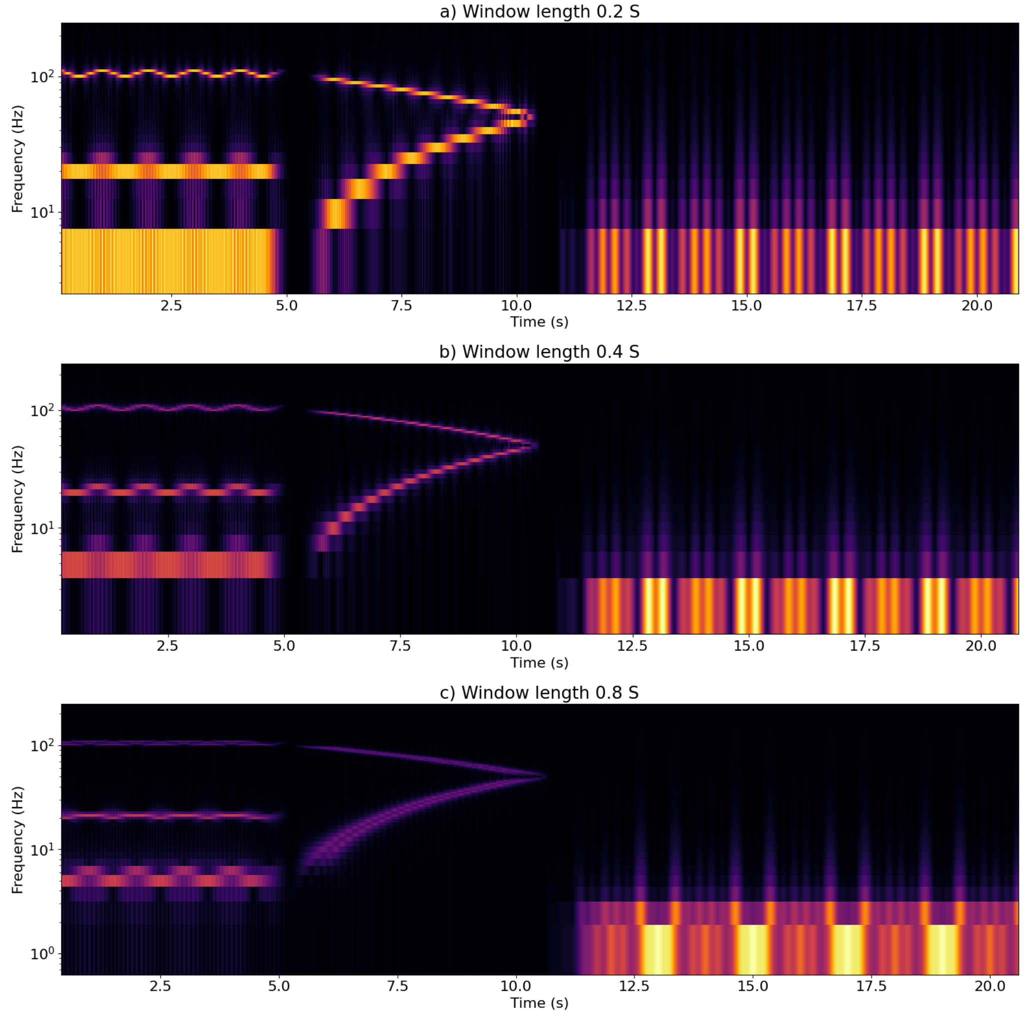

Figure 1.3: Here we use three different window sizes 0.2S, 0.4S, and 0.8S, to illustrate how the time-frequency uncertainty manifests in WFT. We use a log scale on the frequency axis to emphasize resolution in frequency domain *a*) Here, we obtain better resolution in time than the other two, owing to the small window size - this can be seen by comparing the bottom right portion of the graph *b*) This shows a tradeoff in time-frequency uncertainty - the frequency resolution becomes better as we increase the window size. *c*) This shows another end of the tradeoff where a gain in frequency resolution is apparent from the arrow-like section.

1.2 Wavelet Transform: The Mathematical Microscope

The idea of wavelets has been present since the early 20th century, which can be traced back to the works of Alfred Haar. It was not until the near end of the 20th century that these concepts were formalized and became widely popular. Since then, it has been used in a wide array of applications in fields like Seismology, Electrical engineering, and studying brain EEG signals, to name a few [29].

Wavelet Transform decomposes the signal into the time-localized orthogonal basis (Mother Wavelets), which tells us about the time-localized contribution of each component at different scales. There are two versions of WA techniques: Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT). The Wavelet analysis is a vast topic in itself and we won't be able to do justice to it in a single section of a chapter. Hence we will look at it from an application point of view. In this section, we will first introduce what Continuous Wavelet Transform (CWT) is. Subsequently, we will take a brief look at how to select Wavelet for a given application.

1.2.1 Continuous Wavelet Transform

Continuous Wavelet Transform (CWT) of a function $f(t)$ with respect to basis function ψ is given by:

$$(T^{wav}f)(a,b) = \int_{-\infty}^{\infty} f(t)\overline{\psi_{a,b}(t)}dt \tag{1.4}$$

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{1.5}$$

$$\int dt\psi(t) = 0 \tag{1.6}$$

Here, a is the scale parameter controlling the dilation/compression of the wavelet. b is the position parameter that controls the translation of the wavelet along the time axis. The functions $\psi_{a,b}$ are called 'wavelets', ψ is called the 'mother wavelet'. The wavelet must fulfill certain conditions, which we will look at in detail in the section 1.2.3.

What exactly do we mean by translation and dilation of the wavelet? When we change b we shift the center of the wavelet from 0 to b, thus translating the wavelet. The a here is used to change the width of mother wavelet. It stretches the wavelet along the time axis and thus changing the 'scale' or effectively 'frequency'. The equation (1.4) refers to a continuous version of the CWT. We can vary the a & b continuously to obtain CWT coefficients at each scale (determined by a) and at each point in time (determined by b). As discussed in the section 1.2.3, the energy of the wavelet at all the scales needs to be fixed (preferably to 1). We use the scaling factor $\frac{1}{\sqrt{a}}$ to scale it's amplitude and stretch it along time [8]. We shall look at it in detail in the section 1.2.3.

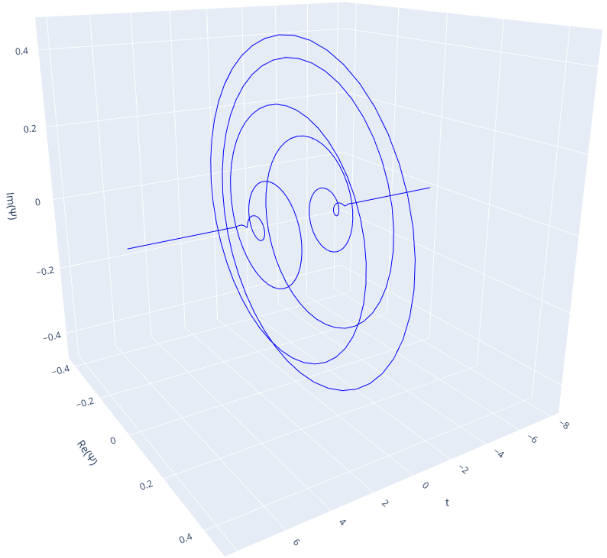

Figure 1.4: The complex Morlet Wavelet

For this body of work, we have used the complex Morlet Wavelet to analyze data. The reason for this choice is explained in the next section. The Complex Morlet Wavelet is constructed in two parts: the oscillatory part and the envelope. The oscillatory part is a point moving on a unit circle in the argand plane with constant angular speed. The envelope is a Gaussian (window) - so the complete function is given by:

$$\psi(t) = \frac{1}{\sqrt{\pi f_b}} e^{2\pi i f_c t} e^{-t^2/f_b} \tag{1.7}$$

Where f_c is the central frequency of the Mother Wavelet, and f_b is the scaling parameter. We need to fine-tune two parameters to yield the best results for the given signal.

The key difference to note here is that we use scales instead of frequency. Scale and Frequency are inversely proportional, and the proportionality constant depends on the ψ we use. Also, we can independently choose arbitrary scales to find the CWT coefficients.

1.2.2 Revisiting uncertainity

The scaling of the wavelets over time ensures the scaling of the 'window' equivalent at all levels. This leads to an optimum time-frequency tradeoff at every scale, leading to better resolution.

Let's look at an example of CWT fig 1.5 where we compare it with windowed Fourier transform. As we know from the figure 1.2, the choice of window size can affect the time and frequency resolution. But here, we observe that the CWT can have both time and frequency resolution by optimizing the tradeoff independently for each scale. As a result, we are able to resolve the lower frequencies without losing the time resolution at any scale.

1.2.3 Mother Wavelet

The 'Mother Wavelet' essentially is the probe we choose to analyze the signal. This is the basis on which we represent the time series. However, unlike the Fourier transform, it does not need to be a sine or cosine function. We can construct a function (Mother Wavelet) that best captures the properties of the time series, but this function needs to meet certain conditions for it to be a valid wavelet. We will briefly take a look at these conditions.

Admissibility Condition: A function $\psi(t)$ qualifies as a mother wavelet only if it satisfies the admissibility condition:

$$\int_{-\infty}^{\infty} \frac{\left|\hat{\psi}(\omega)\right|^2}{\omega} d\omega < +\infty \tag{1.8}$$

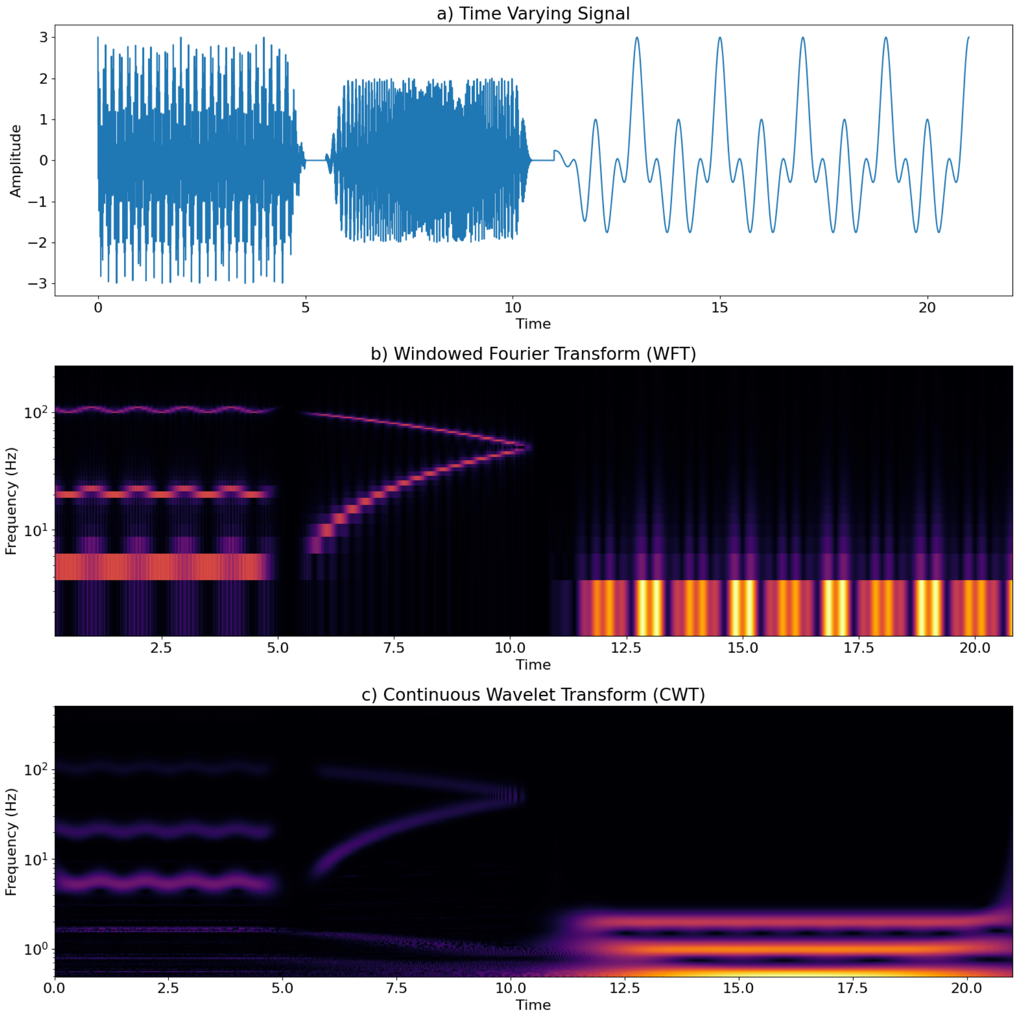

Figure 1.5: *a*) The non-stationary signal. *b*) WFT with window size 0.4s *c*) CWT using complex Morlet Wavelet with appropriately tuned wavelet parameters. We can observe from this example how CWT can be tuned to obtain Optimum Time-Frequency resolution. Unlike in WFT, we do not have to trade resolution in time everywhere to obtain better frequency resolution in the lower frequency band.

Where $\hat{\psi}(\omega)$ is the fourier transfrom of $\hat{\psi}(t)$. This can be guaranteed if the following two conditions are fulfilled.[5, 14]

Zero Mean: The mother wavelet has a zero mean, meaning its integral over the entire domain equals zero:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{1.9}$$

Second: The $\hat{\psi}$ is continuously diffrentiable and

$$\hat{\psi(\omega)} = 0 \quad \text{for } \omega = 0 \tag{1.10}$$

We also need to take into account several other factors while choosing a wavelet-like:

Complex or Real: The Complex wavelet returns amplitude and phase information and is ideal for studying oscillatory behavior. On the other hand, the Real wavelet can be used to study discontinuities and sudden jumps and can be utilized to isolate peaks.

Width: The width of the wavelet is the e-folding time or the decaying time. This is what decides the time-frequency tradeoff for a given central frequency and helps in optimizing this tradeoff.

Shape: The chosen wavelet should align closely with the features of the time series under consideration for best match. However, it should be noted while analyzing the spectra that this choice matters less as we can still capture the qualitative behavior regardless.

1.2.4 Statistical test of significance

When 1D time series diffuses into 2d time-frequency series, we don't have a very clear idea of the significance of each data point we look at. Hence, a statistical test to determine the confidence value for each data point becomes necessary. This method was proposed by C. Torrence and G.P. Compo in 1998. We will briefly review the significance test proposed by Torrence and Compo in the context of El Nino Southern Oscillation.

In order to determine the significance level of any given spectra, we first need to select appropriate background spectra. Then, it is assumed that the spectra of physical phenomena are distributed about this mean background. This spectrum is then used to establish a null hypothesis to determine the significance of a peak in power wavelet power spectra.

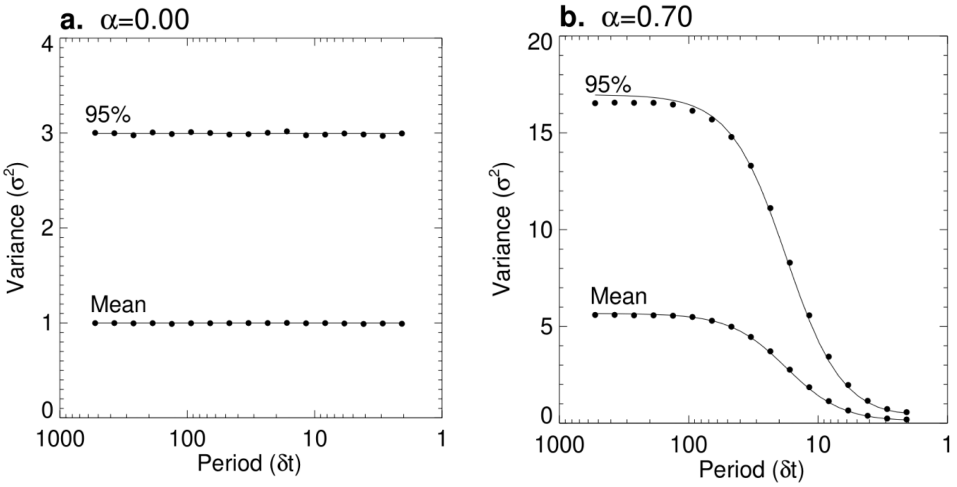

Figure 1.6: *a*) The figure shows the theoretical wavelet spectrum of white noise ($\alpha = 0.0$) (thin black line) and calculated average spectrum of 100000 Monte Carlo simulations of white noise ($\alpha = 0.00$) (black dots). The theoretical spectra is considered as mean to define the top thin line with 95% confidence level, which for a χ_2^2 distribution is $\chi_2^2(95\%)$ times the mean spectrum similarly black dots are the 95% values calculated from Montecarlo run *b*) similar process as **a** but for red noised ($\alpha = 0.70$) **©American Meteorological Society. Used with permission.**

In the original paper, the author argues the background of processes with geophysical properties could be modelled using white or red noise. This noise is used to produce the expected power spectra of the process, and then this spectrum is assumed as a mean of the distribution of the Fourier spectrum, which is then used to define the distribution of expected Fourier/wavelet coefficients. When the observed coefficients are above the expected value (decided by p-value) then the null hypothesis is that there is no significant difference in the expected background and the observation does not hold. Hence, it is termed "significant".[8]

It has been shown that the smoothed Fourier spectra of any process approach wavelet spectra are the same.[6, 7] Hence, it was argued that the Fourier spectra of the appropriate noise can be taken as the background of the process.

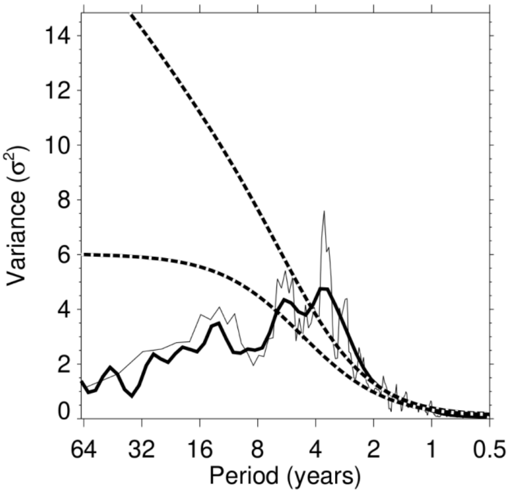

Figure 1.7: The figure shows an example of this method by the original authors. Here, the background is assumed to be a red noise process with $\alpha = 0.72$. The power spectra of this process is taken as mean (lower dashed line) relative to which 95% confidence interval is defined (upper dashed line). The thick solid line is the global wavelet spectrum for the Niño3 SST and the thin line is a smoothed five-point running average of Fourier spectra. **©American Meteorological Society. Used with permission.**

1.3 Molecular Dynamics: a brief look

Molecular Dynamics simulation is a widely used technique across various fields such as biology, chemistry, physics and material science to study the properties of molecular systems. This method calculates the time-dependent behavior of molecular systems, providing insights that are unattainable through experimental methods. Here, we look at a short review of MD simulations and its components.

Molecular Dynamics uses Newton's law of motion to simulate the motion of atoms. The atoms interact with each other through potential energy functions derived using quantum mechanics and empirical measurements. The force on each atom is calculated using these functions, and their position and velocities are updated iteratively at each time step. This process requires the use of several key components to simulate the dynamic behavior accurately:

Force Field: It is a Mathematical representation of the potential energy landscpe of the system which dictates the interaction between atoms in the system. These functions are a combination of equations describing specific interactions (like bond stretching, angle bending, and torsions) and parameters derived from quantum mechanics calculations or fitting experimental data. These force fields also take in account non bonded interactions like electrostatic and Van der Waals forces. There are many force field available each optimized for specific applications like AMBER, CHARMM, and GROMOS.

Integration Algorithm: There are many Integration algorithm available for performing the velocity and position calculation using forces calculated by Force Fields. Verlet is a widely used and computationally efficient algorithm. It utilizes positions and velocities from previous time steps to calculate new positions and velocities.

Ensemble Conditions: MD simulations can be run under various statistical ensembles such as NVT (constant number of particles, volume, and temperature) and NPT (constant number of particles, pressure, and temperature) to match experimental conditions. This is achived with the help of Thermostat and Barostat.

1.3.1 Thermostat

Thermostat is a computational tool that helps to regulate temperature of the system. MD simulation tracks the positions and velocities of atoms or molecules over time. The thermostat continuously monitors the average kinetic energy of these particles, which is a measure of their temperature. If the temperature deviates beyond a certain range, the thermostat mimics the exchange of energy with an external heat bath. We will take a brief look at a few of the widely use schemes.

Berendsen Thermostat: Imagine all the particles (atoms/molecules) in the simulation box as tiny billiard balls. The Berendsen thermostat works by occasionally adjusting the velocities of these ”balls” proportional to temperature difference to bring their average kinetic energy closer to a desired temperature. It’s good for equilibration but doesn’t perfectly maintain a true ”canonical ensemble” as these adjustments introduce some randomness, causing the temperature to fluctuate slightly.

Nosé-Hoover Thermostat: This approach introduces a sort of ghost particle into the simulation box alongside the real particles. This ghost particle interacts with the real ones, exchanging kinetic energy to maintain the desired temperature. This ensures the overall ”motion energy” stays constant. This is a popular thermostat that closely mimics a true canonical ensemble.

Langevin Thermostat: This method focuses on simulating the effect of a solvents on biological molecules. It applies random forces to the particles, along with a damping or viscous force, mimicking the interactions with surrounding solvent molecules. This helps maintain a constant temperature. This helps mimicking particles experiencing random bumps from water molecules, which is quite realistic for biological systems.

Velocity Rescaling Thermostat: In this method the thermostat periodically adjusts the velocities of all particles proportionally to maintain a constant temperature. This method is similar to Berendsen thermostat, but the adjustments happen more frequently. While effective for some systems, it might not perfectly preserve the ”phase space distribution”.

1.3.2 Barostat

The barostat maintains the pressure of the system by adjusting the volume of the box by modifying the velocities and positions of the particles according to certain schemes. Here we will take a brief look at few of the widely used Barostats.

Berendsen Barostat: The Berendsen barostat is conceptually similar to a thermostat. It scales the volume of the simulation box to adjust the pressure. The scaling factor is determined by the difference in pressure between the bath and the system. This barostat is not recommended for production runs as it does not correctly sample from the NPT ensemble.

Parrinello-Rahman Barostat: The Parrinello-Rahman barostat is similar to the Andersen barostat but it allows to change the shape of the box, as well as the size. This is called anisotropic deformation. This method is well-suited for studying how molecules behave under different pressures, especially when their structure might be affected by pressure. This is the most commonly used barostat algorithm in MD simulations.

Andersen Barostat: The Andersen barostat is an extended system method. The system is coupled to a pressure bath by including an additional degree of freedom in the equations of motion. This mathematical artifact mimics the action of a piston scaling the volume of the system. The Andersen barostat does sample from the NPT ensemble and can be used for production runs. However, it only allows the isotropic deformation of the unit cell.

1.4 Understanding Noise

Generally, the term noise refers to unwanted interference in a signal. Mathematically, noise is a stochastic function of space and time. The white noise behaves like independent and identically distributed random variables, which is a normal distribution with 0 mean. Here, the autocorrelation function is a delta function with a zero time lag.

The white noise carries equal power at all frequencies, which can be seen through its power spectral density (PSD) plot fig1.8. The power spectrum of the noise can also be dependent on frequency, which gives rise to what is known as colored noise. For colored noise, the power spectral density $S(f)$ can be given as:

$$S(f) \propto \frac{1}{f^{\alpha}} \tag{1.11}$$

Where $0 < \alpha < 2$ which is related to fractal dimension of the noise(D) as $D = \frac{3-\alpha}{2}$ [1]. On a log-log PSD plot the slope of the power spectrum is $-\alpha$, presence of different slopes in different regions indicates presence of multifractality.

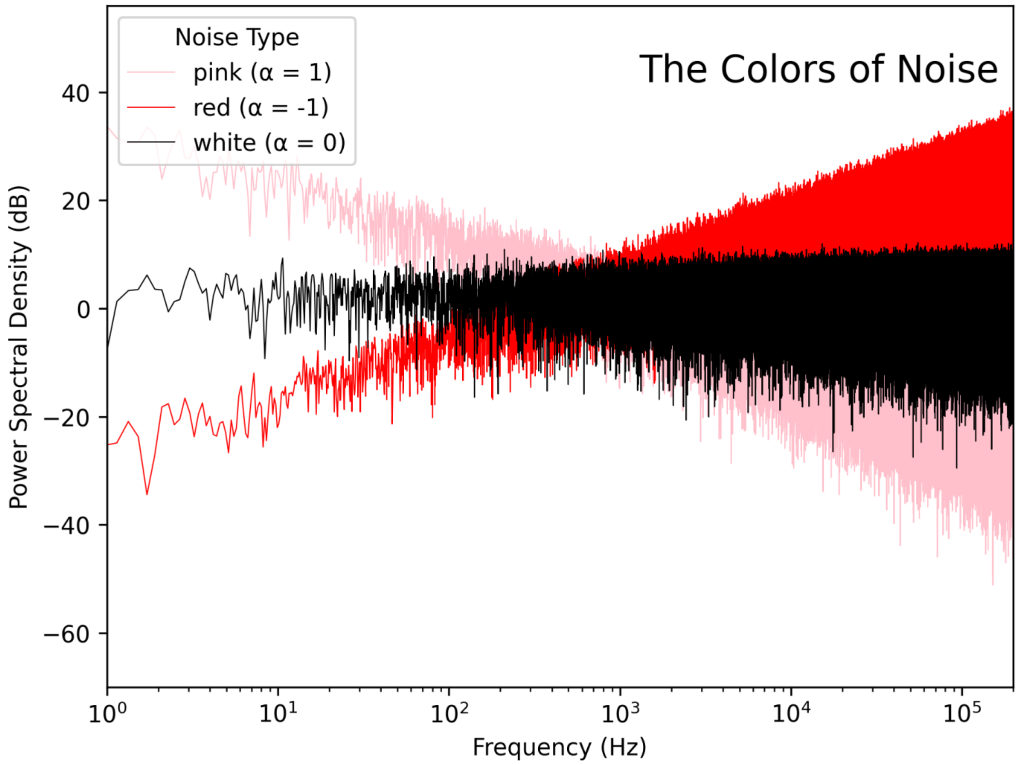

Figure 1.8: The PSD plot of simulated colored noise with different α values. We can observe the slope of the resultant PSD plot approximately $-\alpha$.

1.4.1 Brief look at the randomness in MD data

The atoms' trajectories in MD simulations combine nonlinear and random processes. The nonlinear part comes from many-body interaction, and the random part is introduced by the inherent stochasticity of the Thermostat and Barostat used in the simulation. We looked at whether this resultant can be modeled as noise. If so, then what should be the fractal dimension of the process? Is this dimension a fixed value at all scales?

Chapter 2

Methods

2.1 Standardization of data

We need to consider the differences introduced due to simulation conditions to understand the nature of the expected background present in the MD simulation data. For example, a change in simulation temperature changes the velocity distribution of the atoms in the system. This change makes it difficult to compare the background due to differences in the amplitudes of both velocity and position time series. We rescale the magnitudes of the time series to make them comparable. Once these become comparable, we can develop an understanding of the nature of changes that these differences can have on the background. We use Standardization as a method to bring all data to the same scale. The standardized version of a time series $\mathbf{X}$ is given by:

$$\mathbf{X}_{std} = \frac{(\mathbf{X} - \mu)}{\sigma} \tag{2.1}$$

Here:
μ is the mean of data
σ is the std deviation

The resulting data has a mean of 0 and a standard deviation of 1. This method requires data to follow a normal distribution.

2.2 Power Spectral Density

The Power Spectral Density using windowed Fourier transform of a series is given by:

$$PSD = \frac{Amplitude^2}{\Delta t \times F_s} \tag{2.2}$$

Where Amplitude is the absolute value of the Fourier transform, Δt is the window size, F_s is the sampling frequency.

Similarly, the PSD for wavelet transform is defined as the average energy distribution across different scales (frequencies) obtained by wavelet transform. Hence, it is simply the square of the amplitude of wavelet coefficients.

$$PSD = Amplitude^2 \tag{2.3}$$

2.3 Generation of data

We have three systems in consideration for this study. Here is a brief description of their simulation parameters:

Sytem 1:Water box simulation with Tip3p water model at 250K and 300K.

- Time step 1fs
- saving freq 20fs
- total run 50ps
- Thermostat: Velocity rescale
- Barostat: Berendsen.

We consider trajectories of 100 water molecules to perform analysis and take a sample average. We consider velocity data for the analysis.

Sytem 2:Water box simulation with Tip3p water model at 182K (Temperature of maximum density)[12], 250K and 300K.

- Time step 2fs
- saving freq 20fs
- 2ns Equilibration followed
- total run 100ps.
- Thermostat: Nose-hoover
- Barostat: Parrinello-Rahman
- Columbtype PME

We consider trajectories of 100 water molecules to perform analysis and take sample average. We consider velocity and position data for the analysis.

2.4 The methodology used in this work

When looking at the background for wavelet analysis, we look at how each particle will behave naturally in this situation. Ideally, a Monoatomic molecule (like an ideal gas) should be considered; however, water molecules are the most widely used solvent in molecular dynamics studies. These molecules then interact with all the other molecules of interest, like proteins and lipids in the system under consideration. This contributes the most to the background in wavelet analysis when studying these systems, making them an ideal particle to look at for practical purposes. We use water molecules as probes to capture the background in each system.

We simulate Bulk water (plain water without Solute) at different temperatures for the two systems mentioned above.

The coordinate and velocity of the Oxygen atom are used to determine the position and velocity of a water molecule. When analyzing the position data of **system 1** and **system 2** we first translate the origin to the position of the oxygen atom in the first frame. This is to say, every trajectory starts at the origin, and the position of the water molecule at any frame is given relative to this point. For velocity data, we consider the raw series for analysis.

We perform our analysis on this transformed position and raw velocity time series data. To compare how the temperature affects the fractal dimension of the expected background, we perform the same analysis on Standardized versions of these series. Standardization is used to bring all amplitudes of the time series to the same range. This ensures that Fourier and wavelet coefficient amplitudes are directly comparable (at different temperatures).

We look at the power spectral density of each series in all cases as the wavelet coefficients are complex.

For wavelet analysis, we use complex Morlet wavelets to calculate wavelet coefficients. Complex Morlet wavelets are used to study both time-localized features and phase information of the given series. We use f_b as 1 and f_c as 1.5 (see equation 1.7) for constructing the mother wavelet, which we will use to perform wavelet analysis of these time series.

Chapter 3

Results, Discussion and some additional Observations

3.1 Presence of multifractality

3.1.1 Velocity data

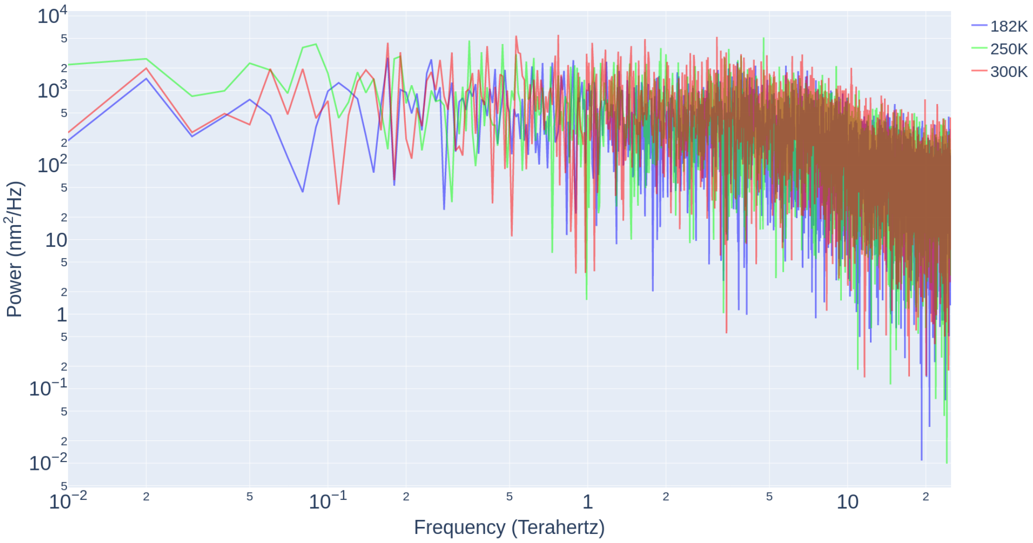

Figure 3.1: Fourier Power Spectral Density of Velocity data for single water molecule of **Sytem 2**

We begin by looking at Fourier Power Spectral Density of velocity time series of **system 2**; comparing this to figure 1.8, we can see the presence of two different parts with regions with different slopes.

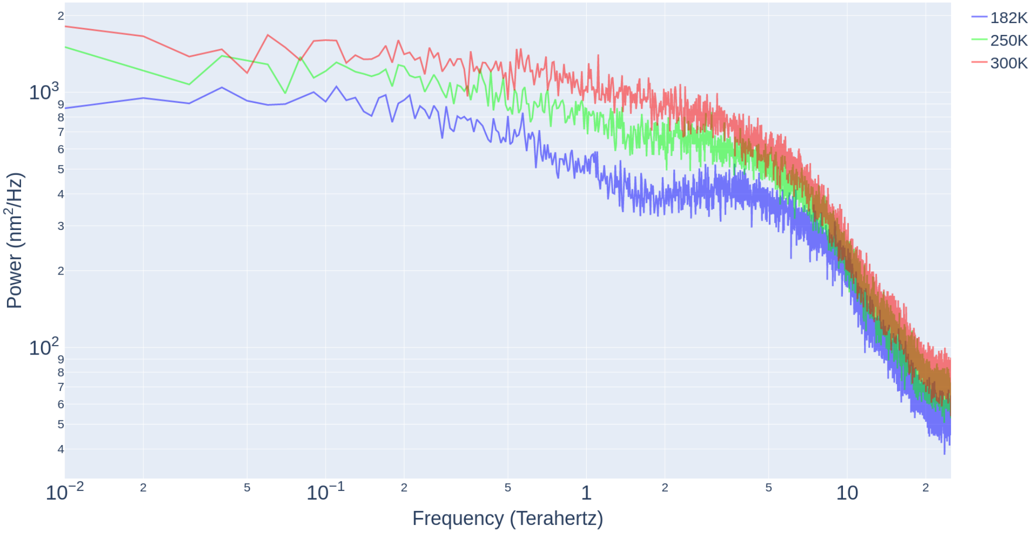

Figure 3.2: Average Fourier Power Spectral Density of Velocity data for 100 water molecules of **Sytem 2**

This Difference becomes more evident if we take the average Fourier PSD of 100 molecules (figure 3.2). Moreover, it can be approximated as roughly the running average of the averaged PSD. Just like we saw in figure 1.6, this average should approach the theoretical background of the MD simulation (figure 3.3).

Here, we can clearly see the presence of at least two different slopes at two different scales, with the change of slope occurring at roughly 3 Terahertz. The 182K curve can roughly be broken down into three different slopes at three different scales.

Similarly, in the case of **system 1**, we can see at least two different slopes at two different regions. Also, we can see the similarity in the power spectrum, which is reflective of the fact that we are using a velocity rescale thermostat for this simulation.

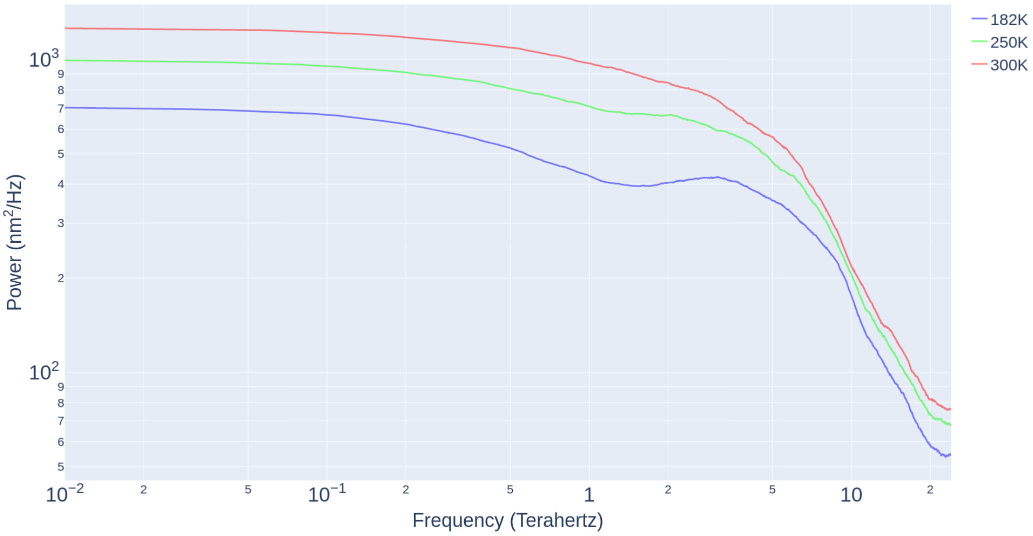

Figure 3.3: Average Fourier Power Spectral Density of Velocity data for 100 water molecules of **Sytem 2** with a running average of 100

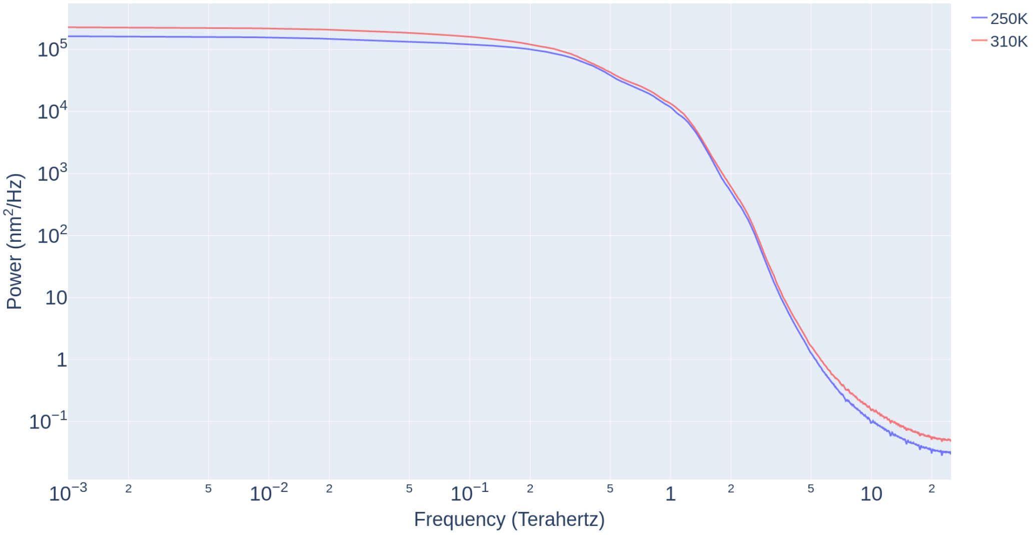

Figure 3.4: Average Fourier Power Spectral Density of Velocity data for 100 water molecules of **Sytem 1** with a running average of 100

3.1.2 Position data

The understanding of PSD of positional data is important for studying structural changes. Here, we don't see very clear difference in slope across

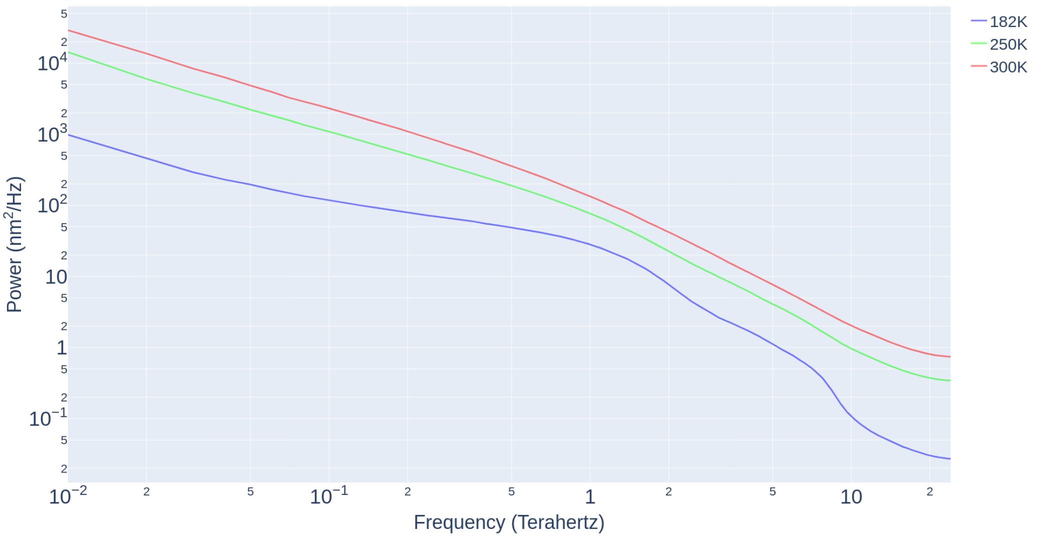

Figure 3.5: Average Fourier Power Spectral Density of Velocity data for 100 water molecules of **Sytem 2** with a running average of 100

scales for 300K and 250K systems. But, for 182K, we observe roughly four regions with different slopes. which makes this system (182K) an interesting system for further investigations, as it has some peculiar differences in both the velocity Fourier PSD and position Fourier PSD.

3.2 Effect of temperature on expected background

To further probe how the fractal nature of the systems changes with temperature and make the data ”scale” free (not related to frequency [term used for multifractal analysis]), we Standardize the data before looking at Fourier PSD of these processes.

3.2.1 Velocity data

The Fourier PSD of Standardized velocity data (Rescaled data) of **Sytem 2** shows interesting crossover PSD values. This trend inversion signifies that the fraction of the total energy of a given system at higher frequencies is

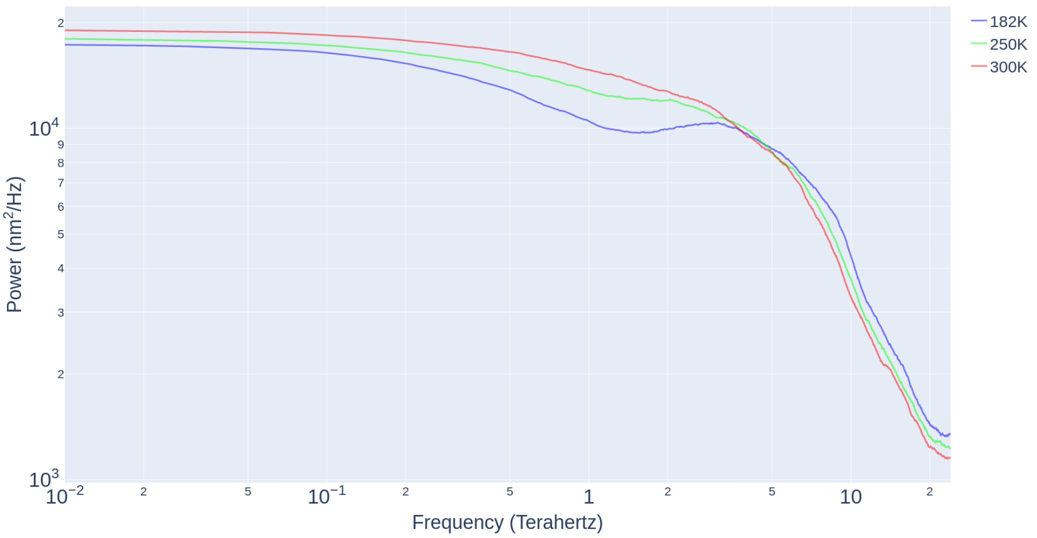

Figure 3.6: Average Fourier Power Spectral Density of standardized Velocity data for 100 water molecules of **Sytem 2** with a running average of 100

higher for low temperatures than for high temperatures. Also, the fraction of total energy of the system of the system at lower frequencies is exactly in reverse order. This indicates an inversion of trend and the presence of some central temperature around which this inversion of trend happens. In the section 3.3, we will discuss the interpretation of this crossover.

The curves for two temperatures in **system 1** (figure 3.7) show a remarkably close resemblance. However, we can still see the presence of this tendency for inversion.

3.2.2 Position data

In the Fourier PSD of the standardized position, the crossover occurs at a very low frequency, and factors like only 100 water molecules were considered for the calculation of this data. This is a log-log plot, making the presence of this crossover questionable. However, there is no denying that the 182K plot shows some peculiar behavior. This can potentially be a topic of further exploration.

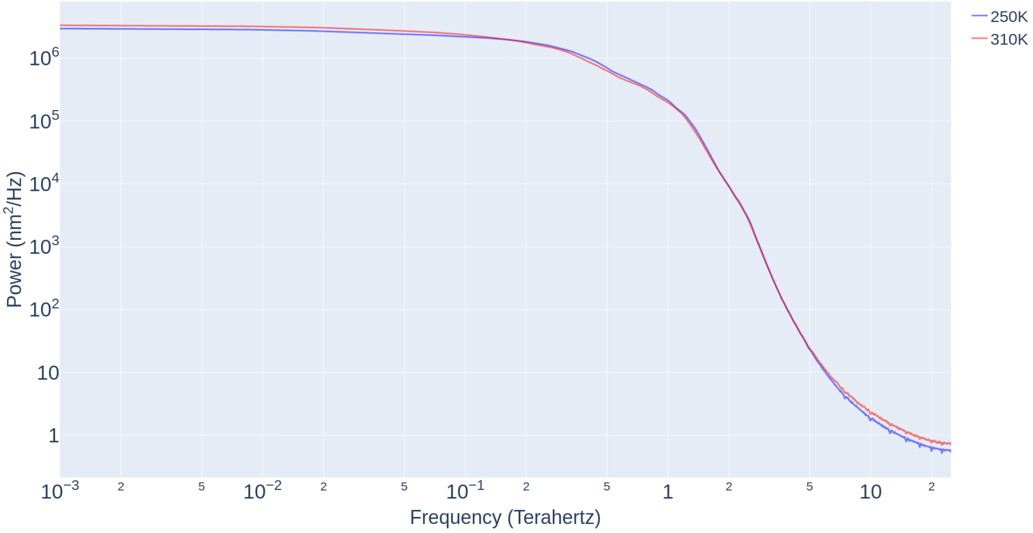

Figure 3.7: Average Fourier Power Spectral Density of standardized Velocity data for 100 water molecules of **Sytem 1** with a running average of 100

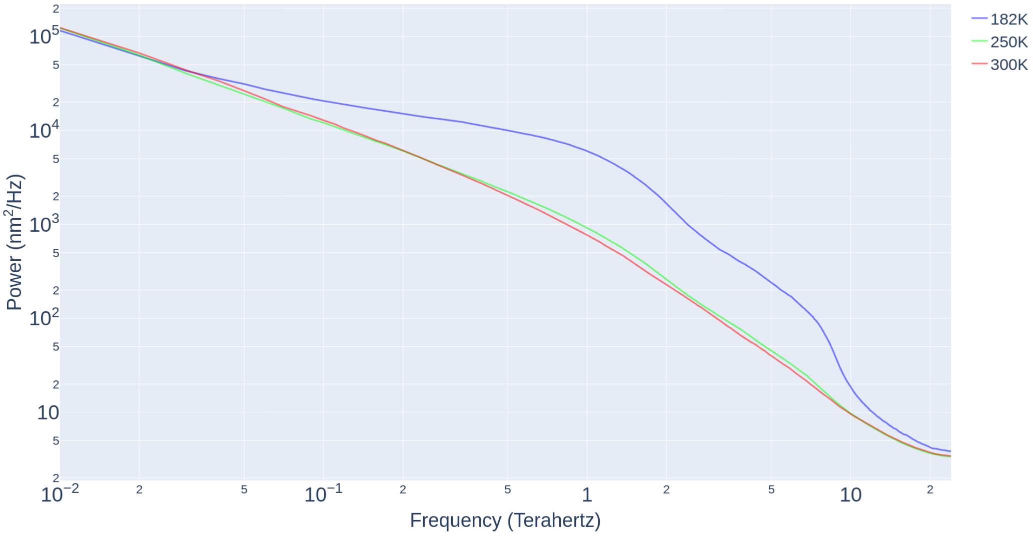

Figure 3.8: Average Fourier Power Spectral Density of standardized Velocity data for 100 water molecules of **Sytem 2** with a running average of 100

3.3 Some additional Discussion

Chapter 4

Conclusion and Future Outlook

The results obtained in the last chapter make it amply clear that the assumption that wavelet background can be modeled using monofractal noise when analyzing molecular dynamics data is simply not true. In this study, we show the definitive presence of a multifractal background. More analysis is required to understand how each simulation parameter affects the fractal nature of this background. Moreover, further studies are needed to understand the effect of changing constituents of the solvent.

Currently available studies like [20, 23, text] use the monofractal background for designing significance tests for wavelet analysis. We plan to develop an analysis method along similar lines that considers the presence of multifractal backgrounds in MD simulation data. We also wish to understand how this background changes with a change in solvent.

The velocity autocorrelation function (VACF) is regularly used to understand the underlying nature of dynamical processes in a molecular system. The Wiener–Khinchin theorem links the Fourier transform of VACF to PSD of velocity [24, text], making wavelet analysis an indispensable tool for gaining better spatio-temporal resolution. As applications of this method, we initially planned to analyze how water dynamics in the hydration layer of the proteins and confined water in a nanotube compare to bulk water.

References

1. Mandelbrot, B. B. Self-Affine Fractals and Fractal Dimension. *Physica Scripta* **32,** 257. ISSN: 1402-4896. https://dx.doi.org/10.1088/0031-8949/32/4/001 (2024) (Oct. 1985).
2. Daubechies, I. in *Ten Lectures on Wavelets* 1–16 (Society for Industrial and Applied Mathematics, Jan. 1992). ISBN: 978-0-89871-274-2. https://epubs.siam.org/doi/10.1137/1.9781611970104.ch1 (2024).
3. Daubechies, I. in *Ten Lectures on Wavelets* 17–52 (Society for Industrial and Applied Mathematics, Jan. 1992). ISBN: 978-0-89871-274-2. https://epubs.siam.org/doi/10.1137/1.9781611970104.ch2 (2024).
4. Daubechies, I. *Ten Lectures on Wavelets* 369 pp. ISBN: 978-0-89871-274-2. https://epubs.siam.org/doi/book/10.1137/1.9781611970104 (2024) (Society for Industrial and Applied Mathematics, Jan. 1992).
5. Farge, M. WAVELET TRANSFORMS AND THEIR APPLICATIONS TO TURBULENCE. *Annual Review of Fluid Mechanics* **24.** Publisher: Annual Reviews, 395–458. ISSN: 0066-4189, 1545-4479. https://www.annualreviews.org/content/journals/10.1146/annurev.fl.24.010192.002143 (2024) (Volume 24, 1992 Jan. 1, 1992).
6. PERCIVAL, D. P. On estimation of the wavelet variance. *Biometrika* **82,** 619–631. ISSN: 0006-3444. https://doi.org/10.1093/biomet/82.3.619 (2024) (Sept. 1, 1995).
7. Perrier, V., Philipovitch, T. & Basdevant, C. Wavelet spectra compared to Fourier spectra. *Journal of Mathematical Physics* **36,** 1506–1519. ISSN: 0022-2488, 1089-7658. https://pubs.aip.org/jmp/article/

36/3/1506/229933/Wavelet-spectra-compared-to-Fourier-spectra (2024) (Mar. 1, 1995).

8. Torrence, C. & Compo, G. P. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society* **79.** Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 61–78. ISSN: 0003-0007, 1520-0477. https://journals.ametsoc.org/view/journals/bams/79/1/1520-0477_1998_079_0061_apgtwa_2_0_co_2.xml (2024) (Jan. 1, 1998).

9. Modi, J. K., Nanavati, S. P., Phadke, A. S. & Panigrahi, P. K. Wavelet transforms: Application to data analysis - I. *Resonance* **9,** 10–22. ISSN: 0973-712X. https://doi.org/10.1007/BF02834969 (2024) (Nov. 1, 2004).

10. Modi, J. K., Nanavati, S. P., Phadke, A. S. & Panigrahi, P. K. Wavelet transforms: Application to data analysis - II. *Resonance* **9,** 8–13. ISSN: 0973-712X. https://doi.org/10.1007/BF02834302 (2024) (Dec. 1, 2004).

11. Hünenberger, P. H. in *Advanced Computer Simulation: Approaches for Soft Matter Sciences I* (eds Dr. Holm, C. & Prof. Dr. Kremer, K.) 105–149 (Springer, Berlin, Heidelberg, 2005). ISBN: 978-3-540-31558-2. https://doi.org/10.1007/b99427 (2024).

12. Vega, C. & Abascal, J. L. F. Relation between the melting temperature and the temperature of maximum density for the most common models of water. *The Journal of Chemical Physics* **123,** 144504. ISSN: 0021-9606 (Oct. 8, 2005).

13. Shaw, D. E. *et al. Millisecond-scale molecular dynamics simulations on Anton* in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (Association for Computing Machinery, New York, NY, USA, Nov. 14, 2009), 1–11. ISBN: 978-1-60558-744-8. https://doi.org/10.1145/1654059.1654126 (2024).

14. Stéphane, M. in *A Wavelet Tour of Signal Processing (Third Edition)* (ed Stéphane, M.) 1–31 (Academic Press, Boston, Jan. 1, 2009). ISBN: 978-0-12-374370-1. `https://www.sciencedirect.com/science/article/pii/B9780123743701000057` (2024).

15. Kaiser, G. *A Friendly Guide to Wavelets* ISBN: 978-0-8176-8110-4 978-0-8176-8111-1. `https://link.springer.com/10.1007/978-0-8176-8111-1` (2024) (Birkhäuser, Boston, 2011).

16. Kaiser, G. in *A Friendly Guide to Wavelets* (ed Kaiser, G.) 60–77 (Birkhäuser, Boston, 2011). ISBN: 978-0-8176-8111-1. `https://doi.org/10.1007/978-0-8176-8111-1_3` (2024).

17. Kaiser, G. in *A Friendly Guide to Wavelets* (ed Kaiser, G.) 44–59 (Birkhäuser, Boston, 2011). ISBN: 978-0-8176-8111-1. `https://doi.org/10.1007/978-0-8176-8111-1_2` (2024).

18. Pagliai, M., Muniz-Miranda, F., Cardini, G., Righini, R. & Schettino, V. Spectroscopic properties with a combined approach of *ab initio* molecular dynamics and wavelet analysis. *Journal of Molecular Structure. MOLECULAR SPECTROSCOPY AND MOLECULAR STRUCTURE 2010* **993,** 438–442. ISSN: 0022-2860. `https://www.sciencedirect.com/science/article/pii/S0022286011001281` (2024) (May 3, 2011).

19. Benson, N. C. & Daggett, V. Wavelet Analysis of Protein Motion. *International Journal of Wavelets, Multiresolution and Information Processing* **10,** 1250040. ISSN: 0219-6913 (July 2012).

20. Benson, N. C. & Daggett, V. Wavelet analysis of protein motion. *International Journal of Wavelets, Multiresolution and Information Processing* **10.** Publisher: World Scientific Publishing Co., 1250040. ISSN: 0219-6913. `https://www.worldscientific.com/doi/abs/10.1142/S0219691312500403` (2024) (July 2012).

21. Mukhopadhyay, S., Dash, D., Mitra, A. & Panigrahi, P. K. *Application of Fourier and Wavelet Transform for analysing 300 years Sunspot numbers to Explain the Solar Cycles* Dec. 10, 2013. arXiv: `1310.6876[cs]`. `http://arxiv.org/abs/1310.6876` (2024).

22. Ngui, W. K., Leong, M. S., Hee, L. M. & Abdelrhman, A. M. Wavelet Analysis: Mother Wavelet Selection Methods. *Applied Mechanics and Materials* **393.** Publisher: Trans Tech Publications Ltd, 953–958. ISSN: 1662-7482. https://www.scientific.net/AMM.393.953 (2024) (2013).

23. Heidari, Z., Roe, D. R., Galindo-Murillo, R., Ghasemi, J. B. & Cheatham, T. E. Using Wavelet Analysis To Assist in Identification of Significant Events in Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **56,** 1282–1291. ISSN: 1549-960X (July 25, 2016).

24. Balakrishnan, V. in *Elements of Nonequilibrium Statistical Mechanics* (ed Balakrishnan, V.) 223–242 (Springer International Publishing, Cham, 2021). ISBN: 978-3-030-62233-6. https://doi.org/10.1007/978-3-030-62233-6_16 (2024).

25. Pal, M., Bhattacherjee, S. & Panigrahi, P. K. *Unstable periodic orbits are faithful biomarker for the onset of epileptic seizure* Pages: 2021.09.03.21263098. Sept. 10, 2021. https://www.medrxiv.org/content/10.1101/2021.09.03.21263098v1 (2024).

26. *Multifractal Detrended Analysis Method and Its Application in Financial Markets — SpringerLink* https://link.springer.com/book/10.1007/978-981-10-7916-0 (2024).

27. *Polymers — Free Full-Text — Effect of the Water Model in Simulations of Protein–Protein Recognition and Association* https://www.mdpi.com/2073-4360/13/2/176 (2024).

28. *Resonance Journal of Science Education — Indian Academy of Sciences* https://www.ias.ac.in/describe/article/reso/009/03/0050-0064 (2024).

29. *Wavelets in Geophysics, Volume 4 - 1st Edition — Elsevier Shop* https://shop.elsevier.com/books/wavelets-in-geophysics/foufoula-georgiou/978-0-08-052087-2 (2024).