LINEAR SYSTEMS ANALYSIS OF MOLECULAR DYNAMICS

BY

STANLEY ANSELM NICHOLSON

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Applied Mathematics
in the Graduate College of the
Illinois Institute of Technology

Approved _____
Adviser

Approved _____
Co-Adviser

Chicago, Illinois
May 2023

ACKNOWLEDGEMENT

This research project would not have been possible without the amazing guidance of Dr. Bob Eisenberg and Dr. David Minh. Their insightful comments and discussions always led me to learn and push through hiccups. I would also like to thank Dr. Chun Liu for interesting and useful discussions. The academic and intellectual environment of Dr. Minh's group has been also invaluable in providing research directions and ideas.

AUTHORSHIP STATEMENT

I, Stanley Anselm Nicholson, attest that the work in this thesis is substantially my own.

In accordance with the disciplinary norm of Applied Mathematics (see IIT Faculty Handbook Appendix S), the following collaborations occurred in the thesis: Dr. David Minh and Dr. Bob Eisenberg provided guidance as is appropriate for MS thesis advisers. Dr. Chun Liu served as my co-adviser for the project and provided advice and useful discussions as is appropriate for a MS thesis adviser.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbol	Definition
$\mathcal{L}\{f(t)\}(s)$	One sided Laplace transform of the function $f(t)$
$\mathcal{F}\{f(t)\}(f)$	One sided Fourier transform of the function $f(t)$
$R_{xx}(\tau)$	Auto correlation of a signal $x(t)$ at a particular lag τ
$R_{xy}(\tau)$	Cross correlation between an input $x(t)$ and output $y(t)$ at a particular lag τ
$G_{xx}(f)$	Auto power of a signal $x(t)$ at a particular frequency f
$G_{xy}(f)$	Cross power between an input $x(t)$ and output $y(t)$ at a particular frequency f
$C_{xy}(f)$	Pairwise coherence between an input $x(t)$ and output $y(t)$ at a particular frequency f
$H(f)$	Frequency function between an input $x(t)$ and output $y(t)$ at a particular frequency f
$\mathbf{C}$	Pairwise coherence matrix where $\mathbf{C}_{ij} = C_{ij}$ is the coherence between atom i and atom j at a predetermined frequency

ABSTRACT

Most proteins reduce the complexity of atomic motion to stable and coherent structures. Molecular dynamics (MD) has provided swaths of trajectory data of proteins. We analyze these trajectories using classical stochastic signal analysis, well established and utilized by engineers. Linear systems analysis operates to uncover linearities given an input and output signal. The coherence function says an input and output are linearly related if and only if the coherence equals one. Analyzing protein motion in the frequency domain allows us to extract a frequency function relating the modes of motion as determined by atomic power spectra. Motivated by biochemistry, we analyze classical interactions like hydrogen bonds and salt bridges and find they act like a linear system, or effective spring. We test our analysis on two protein systems: crambin and the Mu Opioid Receptor (MOR). We extend our results to all pairwise interaction and determine coherent communities of atoms within the MOR. We present various community detection algorithms and demonstrate their validity using common metrics in MD. Identifying rigid and tightly correlated regions of the protein offers great potential in coarse graining protein structure and understanding protein motion.

CHAPTER 1
INTRODUCTION

1.1 Molecular Dynamics

Molecular Dynamics (MD) has been a revolutionary field in developing our understanding of protein motion. The topic of two Nobel prizes, MD is a computational framework to simulate physical systems at the atomic resolution. The motion, vibrations, and interactions of thousands of atoms are computed over the course of millions of time steps. Timescales often span nanoseconds to microseconds of simulation. Intermolecular forces are calculated based on potential terms that attempt to model atomic interactions [1]. This high resolution modelling of the protein offers accuracy and full explanability of the system assuming the model is accurate enough. Biology often uses very few atoms to control proteins with thousands of atoms meaning that such detail is necessary for understanding protein dynamics. Especially when one does not know the crucial atoms, MD becomes invaluable. However, such detail comes at a huge computational cost both in simulation and analysis. Scientifically, we are trying to find a needle set of atoms) in a haystack (protein). The wealth of trajectories MD provides needs dimension reduction to identify which structures are correlated. No where are we guaranteed that our system exhibit linear correlations, for that matter, given the non-linearities associated with the interaction potentials and electrostatics. We want to learn the underlying linear structures that emerge purely from the dynamics. We continue this discussion of protein structure and correlation later on and refer to [1] for detailed discussions of the implementation, philosophy, and applicability of MD.

1.2 Theory of Digital Signal Processing

Digital signal processing (DSP) has a long tradition [2]. We closely follow

the brilliant exposition of its theory and implementation by Bendat and Piersol [3]. The key task of DSP is to twofold: to translate between finite and discrete time to infinite and continuous time and translate the time-domain and frequency-domain. Full technical details are not discussed here but rather a general framework of DSP is presented. Our main goal in this research is to apply the extensively studied and robust electrical engineering tools of power spectra, frequency function, and coherence function to understanding the relationships between the trajectories of atoms.

1.2.1 Frequency Function. The cornerstone of linear systems analysis is the definition of a linear system. In the time-domain, an input $x(t)$ and output $y(t)$ are linearly related if there exists a casual, time-invariant linear ordinary differential equation (ODE) between the two signals $x = x(t)$ and $y = y(t)$

$$a_n\frac{d^n y}{dt^n}+a_{n-1}\frac{d^{n-1}y}{dt^{n-1}}+\cdots+a_1\frac{dy}{dt}+a_0 y = b_m\frac{d^m x}{dt^m}+b_{m-1}\frac{d^{m-1}x}{dt^{m-1}}+\cdots+b_1\frac{dx}{dt}+b_0 x \quad (1.1)$$

with $a_n, \ldots, a_0, b_m, \ldots, b_0 \in \mathbb{R}$. A first course in ODEs will immediately tell us to take the Laplace transform of both sides of Equation 1.1. We solve for $Y(s)/X(s)$ with $X(s) := \mathcal{L}\{x(t)\}$ and $Y(s) := \mathcal{L}\{y(t)\}$. We assume here for simplicity resting initial conditions for all derivatives of x and y since the terms will be absorbed into the constants of the polynomial.

$$\frac{Y(s)}{X(s)} = \frac{b_m s^m + b_{m-1}s^{m-1} + \cdots b_1 s + b_0}{a_n s^n + a_{n-1}s^{m-1} + \cdots a_1 s + a_0} \quad (1.2)$$

We define the ratio of Y and X as H which we call our frequency function (often also called a transfer function). We obtain the simple relation

$$X(s)H(s) = Y(s). \quad (1.3)$$

The Laplace transform is a generalization of the Fourier transform where there is a non-trivial imaginary component of frequency. We will not delve into the technical details regarding one sided transform. Rather, we will move into real frequency

space by evaluating our Laplace transformed signals along $i2\pi f$ where f is the real frequency in units of Hertz (Hz). For brevity we will we denote $X(i2\pi f)$ as $X(f)$ and $Y(i2\pi f)$ as $Y(f)$. The definitions of $X(f)$ and $Y(f)$ correspond to the definition of the Fourier transform of $x(t)$ and $y(t)$ which we will write as $X(f) = \mathcal{F}\{x(t)\}(f)$ and $Y(f) = \mathcal{F}\{y(t)\}(f)$. Mathematically, this looks like

$$X(f) = \int_0^\infty x(t)e^{-i2\pi f}dt. \tag{1.4}$$

The numerical methods to approximate the Fourier transform will be discussed later. We wish to emphasize the paradigmatic approach that the frequency domain offers. Namely, interpreting equation 1.3 in the real frequency domain as an input-output relationship between $x(t)$ and $y(t)$. In the frequency domain, diagram 1.1 represents

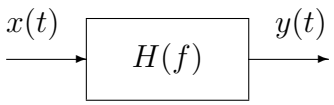


Figure 1.1. Block diagram of input $x(t)$ and output $y(t)$ related by a frequency function $H(f)$.

$\mathcal{F}\{x(t)\}(f)H(f) = \mathcal{F}\{y(t)\}(f)$. In the time domain $x(t) * h(t) = y(t)$ where $h(t) = \mathcal{F}^{-1}\{H(f)\}(t)$ the inverse Fourier transform of H. If we swap the input x and output y with input y and output x our new frequency function is $1/H(f)$.

1.2.1.1 Examples of Linear Systems. Examples of linear systems are plentiful with many mechanical and electrical realizations. A classical model system is that of a damped harmonic oscillator. The output displacement $y(t)$ (measured in meters) of the spring is driven by an input force $x(t)$ (measured in Newtons),

$$M\frac{d^2y}{dt^2} + b\frac{dy}{dt} + ky = x(t) \tag{1.5}$$

with M the mass of the spring in kilograms, b the damping constant given in Newton seconds per meter, and k the spring constant in Newtons per meter. Such an equation

is often written in the interpretable form

$$\frac{d^2y}{dt^2} + 2\beta\frac{dy}{dt} + \omega_0^2 y = x(t) \tag{1.6}$$

with $\omega_0^2 = (2\pi f_0)^2 = k/M$ and $\beta = b/(2M)$. f_0 is the natural resonance frequency of the spring and β is the damping constant. Further discussion of analyzing the spring model 1.6 can be found in Taylor [4].

Performing a similar analysis as described in Section 1.2.1 we obtain a frequency function H relating the input x and output y which is given by

$$H(f) = \frac{1}{-4\pi f^2 + 4\pi\beta i f + 4\pi^2 f_0^2} \tag{1.7}$$

If we set $f_0 = 1$ Hz and vary β we see the expected resonance at 1 Hz and the effect of damping on the magnitude of response at the resonance frequency.

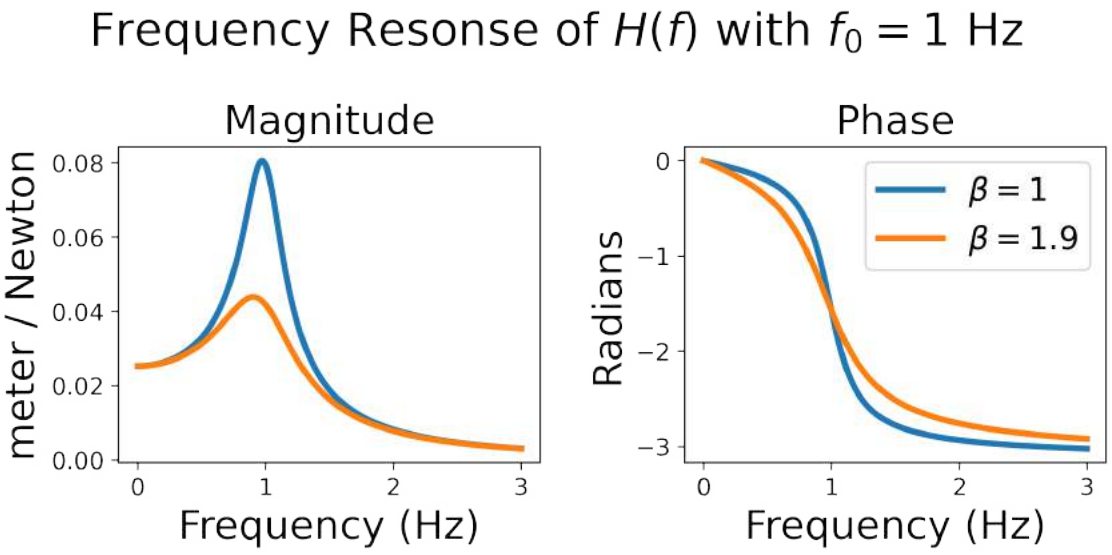


Figure 1.2. Graph of the magnitude and phase of the frequency function H as a function of frequency. We observe the expected resonance at 1 Hz.

1.2.1.2 Motivation for Linear Systems. While rarely appearing in Molecular Dynamics (MD), the frequency function $H(f)$ has a long history in understanding the electrical engineering, materials science, and biochemistry. Measurements of $H(f)$

have been foundational in characterizing material properties for more than a century [5][6] as detailed in [7]. Impedance spectroscopy measurements of the electrical properties of proteins were once used extensively [8] before protein structures were available but are rarely mentioned nowadays.

The following theory of linear systems analysis deals with random signals or stochastic process whose realizations are sampled from a probability distribution. Often reality is contaminated with noise that requires a robust statistical procedure to recover the true underlying signal. Molecular Dynamics (MD) is a physical model contaminated with noise where we are both blessed and cursed with atomic trajectory resolution. We wish to quantify these input-output relationships, thought of as correlations, and extract reduced models from such an inherently high dimensional stochastic system.

1.2.2 The Inverse Problem. In the prior section we introduced the forward problem of linear systems: given a model what is its response curves and behaviors. Engineers are particularly interested in the inverse problem: given data from the dynamics of the system, what is the underlying model? Such a model may take physical interpretation but is not necessary. Whether there are one or two resistors in a black box is not knowable only from the measurement of input and output voltage and current thereby we only create an effective model of the system.

The inverse problem can be stated mathematically. Given discrete and finite time samples of our input and out $\{x(n)\}_{n=1}^{T}$ and $\{y(n)\}_{n=1}^{T}$, what is the model relating the two signals? Are they linearly related and how confidently can we say so? This engineering approach to understanding systems naturally arose from circuits where an input current or voltage is applied and the output response is measured. The question is to identify an effective model of the circuit. Luckily, the inverse problem has been nearly perfectly worked out in the case of linear systems both in identifying

whether such a system *exists* and its *effective* parameters. We will introduce the theory of the inverse problem and numerical approaches that are necessary to solve it.

1.2.2.1 Random Signals. When working in the frequency domain, signals are decomposed into sine waves of different frequency. Random signals or stochastic processes are collections of random variables indexed by time. Examples of random signals include Brownian motion, white noise, $\sin(t + \Phi)$ with $\Phi \sim \text{Unif}[0, 2\pi)$, or trajectories from Molecular Dynamics (MD). Yet the existence of the Fourier transform requires three conditions as described in [9]. The function f has a forward and inverse Fourier transform if

1. $\int_{-\infty}^{\infty} |f(t)|dt < \infty$
2. There are a finite number of discontinuities.
3. f has finite variation.

It can be easily seen that white noise and Brownian motion violate at least the first and third conditions. *However*, we can still make sense of these signals from a frequency approach by defining the cross (auto) correlation function $R_{xy}(\tau) = \mathbb{E}[x(t)y(t + \tau)]$. Properties of the autocorrelation function can be found in [3]. Intuitively, we measure the correlation of two signals with one delayed by a time step τ. Such time correlation functions are useful in identifying an underlying periodic structure within either signal. One of the least periodic signals is that of a white noise $w(t)$ with mean 0 and variance 1. Its autocorrelation function is

$$R_{ww}(\tau) = \mathbb{E}[w(t)w(t + \tau)] = \begin{cases} 1, & \text{if } \tau = 0 \\ 0, & \text{else} \end{cases} = \delta(\tau). \tag{1.8}$$

For generalizing our applicability of the Fourier transform to stochastic signals, we see that R_{ww} satisfies the conditions of outlined in the above list. Thus, random signals have well defined Fourier transforms in the context of their autocorrelation function through the signal's power spectrum.

1.2.2.2 Power Spectra. The one-sided *power spectrum* of a signal $x(t)$ is defined as

$$G_{xx}(f) = \lim_{T\to\infty} \mathbb{E}\left| \int_0^T x(t) e^{-i2\pi f} dt \right|^2 \tag{1.9}$$

We interpret the power spectra of a signal as a measure of the amount of energy at a specific frequency. The power of a signal at a frequency f is given by $G_{xx}(f)$. Sine waves have delta functions as their power spectra since all their power is concentrated at the frequency of motion. White noise has constant power at every frequency. In order to formulate the existence of Fourier transforms for stochastic signals we build on our discussion in Section 1.2.2.1. The Weiner-Khinchin theorem relates the autocorrelation function to the power spectra [10],

$$G_{xx}(f) = \mathcal{F}\{R_{xx}(\tau)\}(f) = \int_0^\infty R_{xx}(\tau) e^{-i2\pi\tau} d\tau. \tag{1.10}$$

where $R_{xx}(\tau) := \mathbb{E}(x(t)x(t+\tau))$ is the auto-correlation function. There are details of convergence in the aperiodic case that require care. The utility of this formulation is that it allows us to extend the notion of frequency function to aperiodic functions [10]. Many textbooks define the power spectrum simply using the auto-correlation function but we argue that 1.9 offers more realistic interpretation.

One useful perspective for viewing the power of our random process is

$$\mathbb{E}(|x(t)|^2) = R_{xx}(0) = \int_0^\infty G_{xx}(f) df. \tag{1.11}$$

which is the variance of the stochastic process given it has mean 0 [11].

We define the cross power spectra similar to equation 1.9 and also present its

cross-correlation formulation (by the Weiner-Khinchin theorem),

$$G_{xy}(f) = \lim_{T\to\infty} \mathbb{E}\left(\int_0^T x(t)e^{-i2\pi f}dt\right)\left(\int_0^T x(t)e^{-i2\pi f}dt\right)^* = \mathcal{F}\{R_{xy}(\tau)\}. \quad (1.12)$$

Here $\cdot^*$ denotes the complex conjugate since Fourier transforms may return complex numbers.

We highlight the important guiding principle in our frequency analysis of trajectories given by Parseval's theorem or thought of as energy conservation. Namely,

$$\int_0^\infty |x(t)|^2 dt = \int_0^\infty \left|\int_0^\infty x(t)e^{-i2\pi f}\right|^2 df \quad (1.13)$$

where the inner integral on the left hand side is the Fourier transform of our signal $x(t)$. Parseval's theorem tells us that the total energy of a signal is the same whether in the time or frequency domain. This reinterpretation of traditionally time domain signals offers insight that can be rigorously guaranteed to represent the system analytically.

1.2.2.3 Power Spectra and the Frequency Function. As demonstrated in [11], the frequency function preserves power. That is if $x(t)$ and $y(t)$ are linearly related through $H(f)$ and if $G_{xx}(f)$ is the autopower of $x(t)$, then

$$G_{xy}(f) = H(f)G_{xx}(f) \quad (1.14)$$

where $G_{xy}(f)$ is the cross power. Furthermore,

$$G_{yy}(f) = |H(f)|^2 G_{xx}(f). \quad (1.15)$$

We see that the frequency function H provides a systematic way to capture the information of how signals are related at each frequency of motion. Power at individual frequencies is scaled by some function of that frequency. Interpreting the form of frequency functions provides the *effective* model of the underlying system and how an input and output are related.

1.2.2.4 Coherence. The coherence function can be thought of as the measure of how linear (in the sense of a frequency function) two signals $x(t)$ and $y(t)$ are. It is defined follows

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \tag{1.16}$$

Notice that coherence is symmetric about input and output, namely $C_{xy} = C_{yx}$. If two signals $x(t)$ and $y(t)$ are linearly related then,

$$\begin{aligned} |C_{xy}(f)|^2 &= \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \\ &= \frac{|H(f)G_{xx}(f)|^2}{G_{xx}(f)|H(f)|^2 G_{xx}} \\ &= \frac{|G_{xx}|^2}{G_{xx}^2} = 1. \end{aligned} \tag{1.17}$$

Therefore, we see that two linearly related signals have a coherence of 1 for all frequencies. Although we may have frequency dependent linear systems where $C_{xy}(f^*) = 1$ for some frequency f^*. This either may result from the linear system only being probed at the frequency f^* or that we can model the correlation as an effective linear system. If we do not *observe* a particular frequency of motion, then we will never be able to determine whether a linear system exists at such a frequency. There is no power at the frequency that we may try to estimate. In impedance spectroscopy, we probe our linear system with a white noise input to observe the linear system across the entire frequency spectrum. It should be noted that this white noise is sent through a lowpass filter because true white noise has infinite power. The applicability arises naturally from different frequencies being orthogonal and uncorrelated. Instead in the time domain, because of continuity of motion, there is arbitrarily perfect correlation locally within time. However, the frequency domain offers this special property of understanding the system in terms of different frequencies. This discussion offers two paradigms towards utilizing coherence as either a measurement either treated as linear systems identifier between input-output or correlation metric. *However*,

these two approaches are fundamentally intertwined and must not be separated when analyzing coherence results. We offer the correlation approach as a familiar framework to understand coherence in the context of modes of motion.

1.2.3 Coherence as Correlation. In the prior section, we discussed the importance of probing a linear system at all frequencies to observe the frequency response. Here we consider a particular frequency of motion at 1 Hz where we demonstrate the utility of coherence as a correlation metric in frequency. The toy system is that of noisy, out-of-phase coupled harmonic oscillators. The noise in the example is a contamination of our system across the entire measured frequency spectrum.

$$\begin{aligned} x(t) &= \sin(2\pi t) + \epsilon\eta_1(t) \\ y(t) &= \sin(2\pi t + \pi/2) + \epsilon\eta_2(t). \end{aligned} \tag{1.18}$$

where η_1, η_2 are independent white noises with mean 0 and variance 1 and ϵ is a small positive number. We employ the formulae for the one-sided power spectra.

$$G_{xx}(f) = \frac{1}{2}\delta(f-1) + 2\epsilon, \;\; G_{yy}(f) = \frac{1}{2}\delta(f-1) + 2\epsilon. \tag{1.19}$$

The cross power spectrum is given by

$$G_{xy}(f) = \frac{1}{2}\delta(f-1)e^{i\pi/2}. \tag{1.20}$$

Therefore, our coherence function between x and y is

$$C_{xy}(f) = \frac{\left|\frac{1}{2}\delta(f-1)e^{-i\pi/2}\right|^2}{\left(\frac{1}{2}\delta(f-1)+2\epsilon\right)\left(\frac{1}{2}\delta(f-1)+2\epsilon\right)} = \frac{\frac{1}{4}\delta(f-1)}{\frac{1}{4}\delta(f-1)+2\epsilon\delta(f-1)+4\epsilon^2}. \tag{1.21}$$

We observe that if $\epsilon \to 0$ that $C_{xy}(f) = \delta(f-1)$ implying that there is a linear system at the frequency 1 Hz. When $\epsilon > 0$, we will clearly have a sub-unity coherence, but this is expected since independent white noise contaminates our signal. Therefore, in our numerical estimations we should expect that we never fully reach unitary coherence. Furthermore, if we require that coherence be at least 0.9 between x and

y, then we would need a signal to noise ratio of around 10. These toy problems allow us to understand a model system for interpreting our numerical estimates of linear systems we find in Molecular Dynamics simulations.

1.3 Estimation

Numerical algorithms for approximating functions has been the hallmark of numerical analysis. Many developments in the field have revolutionized the frequency domain approach. Particularly, the Discrete Fourier Transform (DFT) and its accelerated counter-part the Fast Fourier Transform (FFT) made computationally tractable the estimation of many frequency domain quantities. We emphasize power spectral estimation it enables the immediate estimation of the frequency and coherence functions.

1.3.1 Signal Theory. We think of a digital data as being sampled at a constant sampling frequency from some "ground-truth" continuous signal since we can only do discrete computations. The sampling frequency f_s is measured in Hertz (Hz) or samples per second. Due to the time scale of MD simulations we work in Giga-hertz (GHz) or samples per nanosecond. A consequence in order to properly analyze or "reconstruct" our signal is that we must only look at phenomena with a frequency less than $f_s/2$ also known as the Nyquist frequency (look at Shannon's sampling theorem) [3]. This is important to understand because all frequency domain estimations given by MATLAB or Python will be in the frequency range $[0, f_s/2)$.

1.3.2 Power Estimation. In the dawn of digital signal processing, power estimation was a breakthrough in the development of numerical algorithms. We avoid direct estimation of power spectra through Discrete Fourier Transforms (DFTs) because of their plentiful numerical artifacts that must be carefully attended to. We follow in one of the most popular power estimation methods known as the periodogram which was

improved by Welch [12]. A detailed discussion can be found in Bendat and Piersol [3]. At a high level, the method splits the time-series signal into disjoint bins (also called segments), estimates the power spectra within each bin, and then overlaps across the bins to smooth the power estimation and remove artifacts. Probability distributions of the estimators have been worked out by [13] in cases of non-overlapping windowing but are intractable analytically in the case of overlapping. Although particular estimation details differ across programming languages and packages, we emphasize the main approach of Welch and his periodogram method. Pioneered in 1967, this method remains one of the most popular method for estimating power spectra still taught as the primary method of power spectra estimation in DSP courses [14].

Once we have obtained estimates of $\hat{G}_{xx}(f_n), \hat{G}_{yy}(f_n), \hat{G}_{xy}(f_n)$ for $f_n \in [0, f_s)$, we can estimate the coherence and frequency function

$$\hat{C}_{xy}(f_n) = \frac{|\hat{G}_{xy}(f_n)|^2}{\hat{G}_{xx}(f_n)\hat{G}_{yy}(f_n)}, \quad \hat{H}(f_n) = \frac{\hat{G}_{xy}(f_n)}{\hat{G}_{xx}(f_n)}. \tag{1.22}$$

1.3.3 Welch's Periodogram Method. Welch's method for estimating power spectra is outlined and follows his exposition in [12].

Given a sequence time-domain signal $x(t)$ with $t = 0, \ldots, N-1$. We take L, potentially overlapping, segments with each being D time units apart. Then

$$\begin{aligned} x_1(t) &= x(t), \quad t = 0, \ldots, L-1 \\ x_2(t) &= x(t+D), \quad t = 0, \ldots, L-1 \\ &\vdots \\ x_K(t) &= x(t+(K-1)D), \quad t = 0, \ldots, L-1 \end{aligned} \tag{1.23}$$

such that $(K-1)D + L = N$ so that these x_i for $i = 1, \ldots, K$ covers the entire time signal x. Then we take the DFT of each x_i multiplied by some windowing function,

in our case Hanning [3].

$$A_k(n) = \frac{1}{L}\sum_{t=0}^{L-1} x_k(t)w(t)e^{-2kitn/L} \tag{1.24}$$

taking the magnitude squared and scaling A_k, we obtain $I_k(f_n) = \frac{L}{U}|A_k(n)|^2$ where $f_n = n/L$ and $U = \frac{1}{L}\sum_{t=0}^{L-1} w^2(t)$. Then we obtain the estimate for the auto power $\hat{G}_{xx}$ for x,

$$\hat{G}_{xx}(f_n) = \frac{1}{K}\sum_{i=1}^{K} I_k(f_n). \tag{1.25}$$

a similar method is utilized for the cross power $\hat{G}_{xy}$ except the I_k are computed for by $A_k^x(n)(A_k^y(n))^*$ where $\cdot^*$ denotes the complex conjugate.

1.3.3.1 Implementation of Power Estimation. Below we outline the functions used in power estimation that utilize the classical Welch's method. We explain the parameter choices given in MATLAB's `pwelch` [15]. Similar parameters exist in SciPy's `welch` [16]. We have found very minimal differences in the power and coherence estimation between these two packages, something that is important to ensure when dealing with the artifacts of the DFT.

There are 5 different relevant parameters in the `pwelch` power estimation function which implements "Welch's Periodogram method" [12]:

```
[Pxx f] = pwelch(x, win, Noverlap, Nfft, fs)
```
(1.26)

- `x` is a signal of length `N`
- `win` is the window for each segment with length given by L in Section 1.3.3. We use `hanning(Nfft)`
- `Noverlap` is the number or fraction of overlaps between each segment. We use `Noverlap = Nfft / 2` to give 50% overlap between disjoint segments

- `Nfft`, size of disjoint segments. We use `Nfft = N / 32` to give 32 disjoint segments. Here $K = 32$ where K is the same as in Section 1.3.3. Having `Nfft` be a power of 2 improves computation by the FFT.

- `fs` is the sampling frequency. This varies depending on the sampling rate of the Molecular Dynamics simulation.

1.3.4 DFT Size Choices. The DFT is rife with numerical artifacts if one is not careful. Usually the DFT is based on a small number of samples, on the order of 2048 in many cases, and is often as small as (the recommended default value of) 256 in the software we use. In that case, artifacts produced by the limited number of samples can be very serious, both discretization artifacts and truncation artifacts. Choosing the number of DFT points depends on the desired frequency resolution: [number of DFT points] equals [sampling rate] multiplied by [frequency resolution]. The best frequency resolution is a trade-off between noisy and biased estimation. Further discussion on the statistics of estimation is available in Section 9.2 of Bendat and Piersol [3].

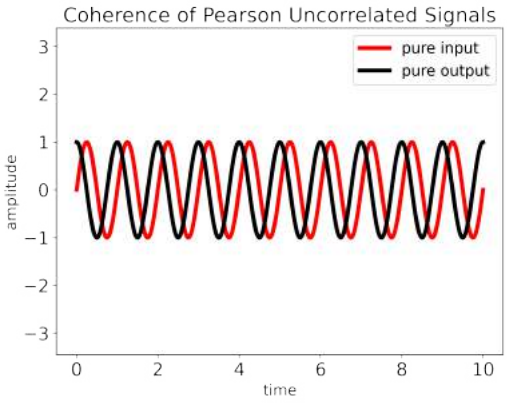
1.3.5 Toy Problem Revisited. We recall our toy problem introduced in Section 1.2.3 where we have two signals x and y,

$$\begin{aligned} x(t) &= \sin(2\pi t) + \epsilon\eta_1(t) \\ y(t) &= \sin(2\pi t + \pi/2) + \epsilon\eta_2(t) \end{aligned} \tag{1.27}$$

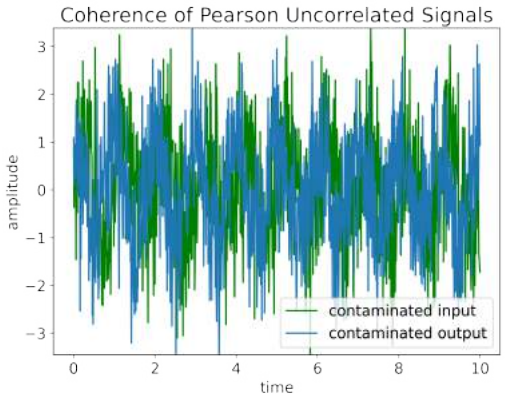
where η_1, η_2 are independent white noises and $\epsilon > 0$. We found that $C_{xy} \approx 1$ at 1 Hz. Here we validate our analytic result with computing the coherence numerically.

We sample our signals at a rate of $f_s = 100$ Hz for 10 seconds so that we observe 10 periods of both signals. Having enough periods or information of the signal ensures we can robustly estimate its power. Here we plot our signals in time with and without

the independent noise that we added to it. Notice that signal to noise ratio is 1 to 1. Since we want a frequency resolution of 0.1 Hz to make sure the 1 Hz frequency



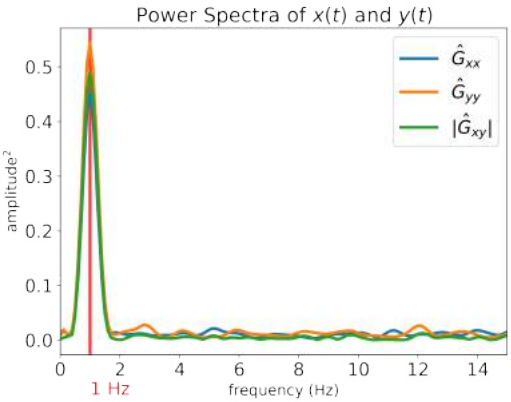
(a) The pure sine wave signals without contamination from independent white noise sources.



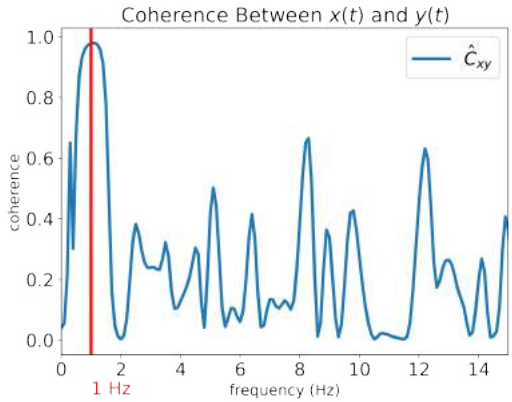
(b) Sine waves with the independent white noises added to each. The coherence estimation algorithm only sees these signals.

Figure 1.3. Sine waves without and with white noises sampled at a frequency of 100 Hz over 10 seconds.

is precise, we use $L = 1000$ DFT points and $K = 256$ number of segments. We use the Hanning window to average over each segment. There is also a 50% overlap between each segment. The choice for these parameters is standard what is used for when we apply these algorithms to MD data. We obtain the numerical estimates of the auto and cross powers G_{xx}, G_{yy}, G_{xy} and the coherence C_{xy}: One important observation is the analytic power spectra and coherence functions are expectations over the white noises. Because of finite time and finite sampling, we obtain non-trivial values – away from our analytic solutions – for both the power and coherence at frequencies different from 1. We say two signals are coherent if their coherence is at least 0.9 which matches our numerical experiments. If the pairwise coherence is less than 0.9, we say the two signals are *not* linearly related and must be careful not to be swayed into thinking they are. In Figure 1.4, we see that the coherence



(a) The auto and cross power of x and y as functions of frequency. Notice is peak at 1 Hz corresponding to our estimates that the auto and cross powers were given by $\delta(f-1)$.



(b) The coherence function between x and y as a function of frequency. We find it takes the form of a delta function at 1 Hz.

Figure 1.4. Estimated power spectra and coherence of and between the two sine waves contaminated with noise $x(t)$ and $y(t)$.

reaches 0.6 near a frequency of 8 Hz but our setup says indicates there is no linear system at 8 Hz. The power of coherence only arises when it is sufficiently close to unity as also demonstrated in the proof outlined in equation 1.17. If we increased both our sampling frequency and sampling time, our coherence function and power spectra would converge to their analytically determined forms.

It should be noted that there are many complicated biases that enter into the estimation via the windows, the overlap between bins, and many more. For example, the broad peak seen at 1 Hz is a consequence of the Hanning windowing and overlap. There is a great deal of study involved in identifying which windowing techniques, parameters, etc. should be used to estimate power spectra and coherence. However, coherence is less sensitive to these parameters because many biases are divided away after the power spectra are estimated. We find that our coherence results from the context of molecular dynamics are robust to changes in the windowing,

window size, number of Fourier transform points, and overlap leading us to believe in our unitary coherence results. Amazingly, the error of coherence estimation, in particular numerical settings, is actually inversely proportional to the true coherence! Further insightful discussions about the intricacies of estimation can be found in Bendat and Piersol [3].

CHAPTER 2

APPLICATION OF COHERENCE TO PROTEIN DYNAMICS

2.1 Simulation and Assumptions

Molecular Dynamics (MD) provides stochastic trajectory data of protein motion.[1] As we described in Section 1.1, MD – in essence – solves Newton's equations of motion given additional electrostatic and torsion angle constraints.[2] One of the principle assumptions about the data from MD is ergodicity that is required to properly estimate ensemble thermodynamic quantities. We know that ergodicity implies stationarity which is all that is required for a well-defined coherence estimation. This weakening of the assumption regarding trajectories makes coherence analysis appealing where ergodicity may not necessarily hold for all simulations. Future work revolves around studying non-stationary signals using wavelet coherence that can identify transient linear systems. Although in the rest of the paper we deal with stationary signals that naturally arise from our MD simulations.

2.1.1 Protein Systems. We analyze two protein systems: crambin and the Mu Opioid Receptor (MOR). Their biological context provides insight into what coherent structure we may uncover but the estimation does not make such assumptions. Crambin is a small globular protein found in Abyssinian cabbage. We chose it for its high resolution crystal structure, its size, and its rigidity [18].

The MOR is a G-protein coupled receptor (GPCR) that has become an important drug target with the growing opioid epidemic in the United States. As a GPCR, it has a complicated signal transduction pathway whose mechanism still is

[1] Chapter 2 follows similarly to our paper [17] but introduces additional ideas.

[2] Various integration methods such as Verlet or Langevin are sympletic numerical schemes that prevent the simulation from "drifting" or diverging from the true solution.

not entirely understood. The big picture is that different opioids induce different signals resulting in varying severity of adverse effects. This problem is also known as biased agonism. General clinical side effects include peripheral effects (constipation, urinary retention, and hives) and central effects (nausea, sedation, respiratory depression, and hypotension) [19]. Opioids derived from the kratom plant have been shown to have reduced respiratory depression compared to lofentanil, a dangerously potent opioid [20]. These results suggest structural nuances due to drug binding are crucial for minimizing side effects. The MOR, along with G-proteins bound to it (Protein Data Bank 6DDF), has a crystal structure resolution of 3.50 Angstroms, limiting crucial detail [21]. Although another structure (Protein Data Bank 5C1M) with a 2.07 Angstrom resolution has just the receptor [22].

The mechanism for MOR and opioid binding is still unclear, but recent developments have led to new directions. Initially, it was understood that opioid binding to MOR activates the G-protein and recruits beta-arrestins [19]. The G-proteins induce analgesia, and the beta-arrestins create unfavorable side effects. Although later, it was shown that beta-arrestin-2 knock-out mice that side effects occur nonetheless [23]. These data suggest the need for a more specific classification of the underlying mechanism. The problem is also exacerbated by the existence of two other opioid receptors: delta and kappa. A holistic understanding of the opioid receptors must be made in order to classify the effects of individual opioids fully. While making the problem of structure-based drug development difficult, the role of computational scientists is still necessary for understanding the structural effects of opioid binding. We study the MOR from the angle of the coherence function to understand the underlying structures involved in drug binding and the associated protein dynamics.

2.1.2 Molecular Dynamics Setup. We used code developed in the Minh group (https://github.com/swillow/pdb2amber) to prepare a model of the system. A crys-

tallographic structure of crambin (Protein Data Bank 1CRN) was protonated using PROPKA [24] at pH 7. The protonated protein was inserted into a cubic box of water with 0.1 M NaCl. The AMBER ff14SB force field [25] was used for the protein, OPC3 parameters [26] for water, and Joung and Cheatam TIP4P/EW parameters for ions [27]. The length of each side of the water box was 75 Å, much larger than crambin. Preliminary calculations with a smaller box of 45 Å showed that crossing the boundary affects the coherence. To avoid the ambiguity of unwrapping, we only report results from the system in the larger box in which the crambin molecule never crossed the periodic boundary.

MD simulation was performed with OpenMM version 7.4.2 [26]. First, the system was minimized. Isothermal MD was performed using the Langevin integrator at temperature T = 300 K with a time step of 2 fs for 100 ns. (Shorter simulations performed with the deterministic Verlet integrator gave indistinguishable results.) Samples were stored every 0.01 ns (equivalent to a rate of 100 GHz).

2.2 Coherence Setup

We take three-dimensional trajectories from the Molecular Dynamics (MD) simulation as our signals. Here we work in the Cartesian coordinate system where each atom i has a three-dimensional position $\mathbf{r}_i(t) = (x_i(t), y_i(t), z_i(t))$ where i ranges from 1 to the number of atoms. Because coherence analysis takes one dimensional input we define the centered displacement of atom i as

$$d_i(t) = ||\mathbf{r}_i(t)|| - \mathbb{E}||\mathbf{r}_i(t)|| = \sqrt{x_i^2(t) + y_i^2(t) + z_i^2(t)} - \mathbb{E}\sqrt{x_i^2(t) + y_i^2(t) + z_i^2(t)}. \quad (2.1)$$

Often in time-series analysis we "center" or "detrend" the signal by subtracting its mean from itself. That is, the expectation is taken over time. We then define powers and consequently coherence with respect to the centered displacements $d_i(t)$ and $d_j(t)$

of the atoms i and j.

$$C_{ij}(f) = \frac{|G_{d_i d_j}(f)|^2}{G_{d_i d_i}(f) G_{d_j d_j}(f)} \tag{2.2}$$

where we are performing the power estimation between input $d_i(t)$ and output $d_j(t)$. We do not make any assumptions about how these signals are distributed and simply estimate the power spectra and coherence function.

2.3 Chemical Interactions

Biochemists and structural biologists have identified foundational and recurrent interactions between atoms. Covalent bonds are a staple of organic chemistry but many important atomic interactions hold together and create the structure of the protein. Within Molecular Dynamics (MD), force fields dictate the type of interactions between atoms whether they include electrostatic, mechanical, hydrophobic (Lennard-Jones), and other inter-molecular interaction terms [1]. Common atom-atom interactions are found across proteins that define and dictate higher order structures in proteins. Of particular interest to us is the hydrogen bond. Loosely speaking, the hydrogen bond exists between an electron donor and acceptor held together by weak electrostatic forces. A particular pattern of hydrogen bonding can result in secondary structures known as alpha helices and beta strands. These structures are found throughout nearly all proteins and form a cornerstone of our understanding of structural biology. Another related interaction of interest is the salt bridge loosely defined as the attraction between negatively and positively charged atoms such as between carboxylate and guanidinium functional groups. Such interactions are dominated by electrostatics and their tight connection may suggest linearity.

Structural motifs in biology are reproducible and robust, occurring in proteins across kingdoms and chemical, physical, and physiological conditions. Particular residues found in particular subsequences form alpha helices that determine protein folding. For example, the hydrogen bonding of the nitrogens and oxygens between

the leaves of the helix induce alpha helices. Although it does not necessarily hold

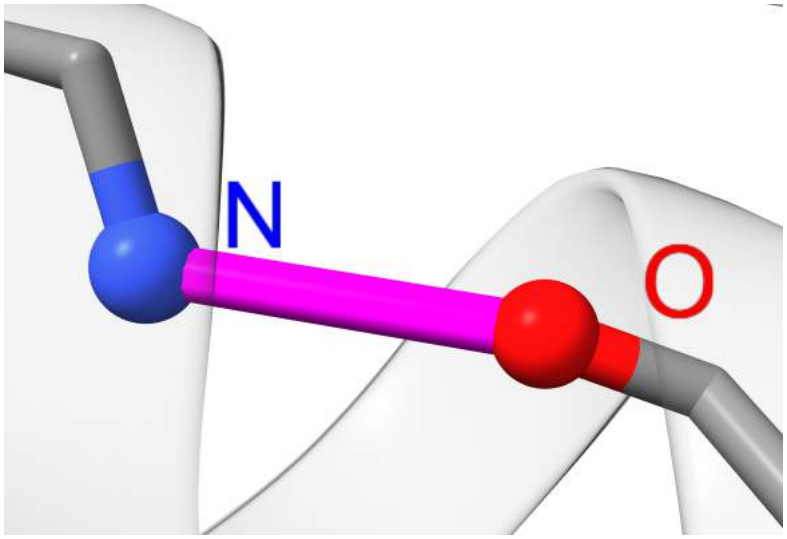


Figure 2.1. Example of hydrogen bond in the main alpha helix of crambin. The nitrogen-oxygen hydrogen bond is treated as an input-output relationship.

in general, linearity is a sufficient condition for robustness. If interactions are linear, it provides the opportunity to learn protein structure by identifying chains of linear interactions. Coherence provides an objective criterion to accept or reject our hypothesis by providing a binary answer – unitary coherence if and only if there exists a linear system. We will see that according to the coherence function, the heavy atoms nitrogen and oxygen (shown in Figure 2.1 involved in hydrogen bonding are linear and that we can exploit this information to learn protein structure.

2.3.1 Coherence of Chemical Interactions. We pursue understanding coherence in terms of biochemically well-defined atomic interactions found throughout crambin. Hydrogen bonds and salt bridges are the predominant "bonds" we will consider in our pairwise coherence analysis.

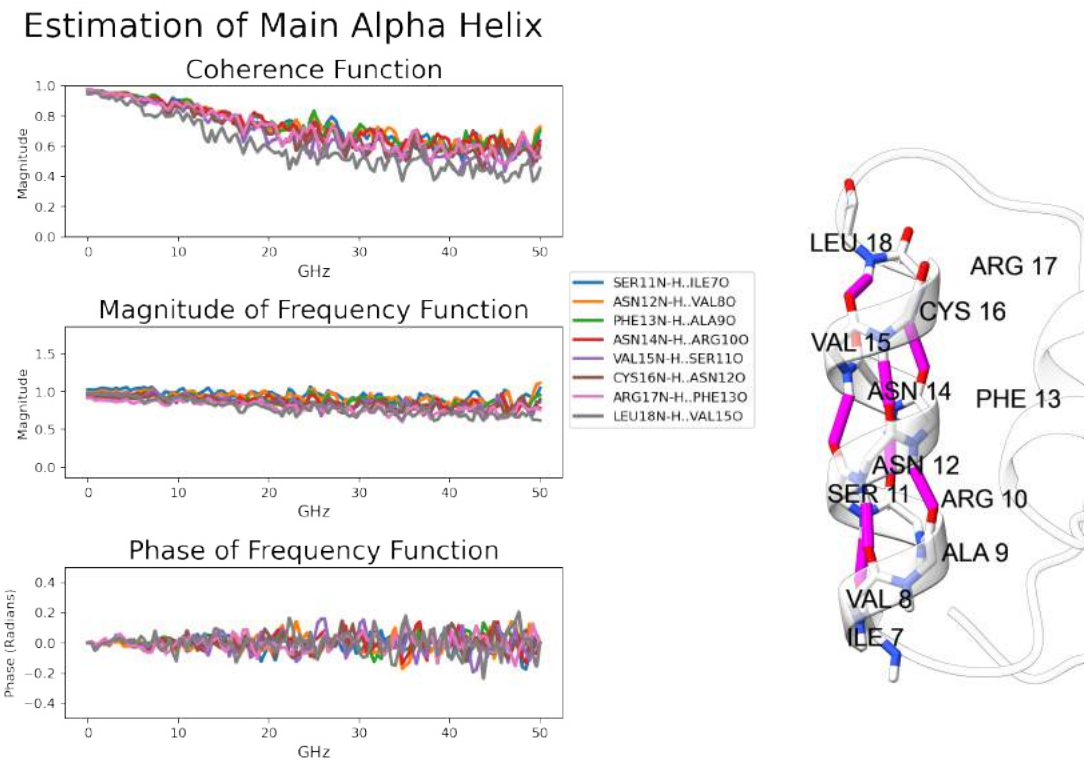
2.3.1.1 Coherence of Structure Hydrogen Bonds: the Alpha Helix. The hydrogen bond defines secondary structures in proteins by definition. These foundational chemical interactions make up much of protein rigidity and shape. Chemically

difficult to define, we provide a context-dependent quantitative measure for hydrogen bonds. We find such interactions to be linear at low frequencies suggesting its importance as a staple that holds together backbone atoms.

We consider the centered displacements of the backbone nitrogens and oxygens in the alpha helices of crambin. We look at d_{N_i} and d_{O_i} where N_i is the ith nitrogen and O_i is the ith oxygen such that N_i and O_i are hydrogen bonded as determined by ChimeraX [28]. We consider their pairwise coherence between each nitrogen and oxygen that are hydrogen bound and graph the results in Figure 2.2 and display them in Table 2.1. On the left hand side of Figure 2.2 we see the hydrogen bonds (show in magenta) that hold together the main alpha helix of crambin. We amazingly find that they have a coherence of near one for low frequencies! Their frequency function estimates having little frequency dependence and near constant magnitude and phase suggest they are rigidly held together. The whole helix being held together by such interactions is not a coincidence when considering how rigid the helix and protein is. The frequency response function is consistent with this H-bond behaving as an

Table 2.1. Mean Coherence H-Bond Results in Crambin's Alpha Helices

Donor (Input)	Acceptor (Output)	Mean Coherence	Mean Magnitude	Mean Phase
Main α-helix				
SER11N	ILE7O	0.959	1.038	-0.001
ASN12N	VAL8O	0.957	0.992	-0.004
PHE13N	ALA9O	0.953	0.955	0.006
ASN14N	ARG10O	0.947	0.966	0.010
VAL15N	SER11O	0.944	0.964	0.010
CYS16N	ASN12O	0.948	0.920	0.008
ARG17N	PHE13O	0.950	0.884	0.012
LEU18N	VAL15O	0.919	0.964	0.001
Small α-helix				
ALA27N	GLU23O	0.938	1.000	-0.006
THR28N	ALA24O	0.915	1.012	-0.016
TYR29N	ILE25O	0.937	0.970	-0.007



(a) Frequency dependence of the coherence and frequency functions. The legend labels each hydrogen bond as the residue name, its residue ID, and the input nitrogen and output oxygen.

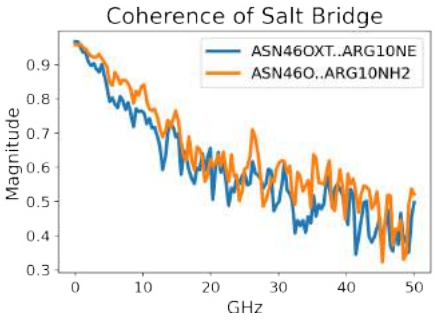
(b) In magenta are the hydrogen bonds we estimated the coherence and frequency functions for.

Figure 2.2. On the left we have the pairwise coherences of the eight hydrogen bonds that hold together the main alpha helix of crambin. The right side depicts the helix itself.

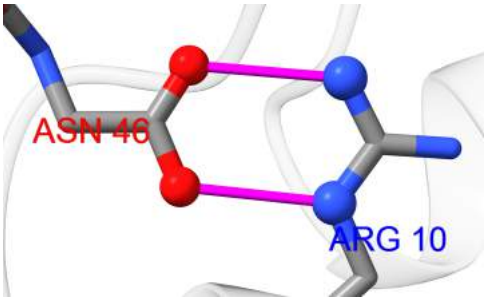
effective rigid spring. We can only be confident of this approximation when the coherence is near 1. We see this situation occur between 0.391 GHz and 5.08 GHz. We can only say this is an effective model because we do not extract the forcing from the MD simulations. Future work on identifying the effective spring constants and appropriate linear system is necessary to fully elucidate the meaning of coherence. The range of unit coherence dictates either the true linear model or provides a mode of motion that can be approximated as linear. The behavior at higher frequencies is less clear given that coherence is sub unity; both the magnitude and phase of the

frequency function are consequently noisier. Because the coherence is lower and the system is not linear at these frequencies, the estimated frequency response function is no longer meaningful and uninterpretable in terms of an analytical model.

2.3.1.2 Coherence of the Salt Bridge Bond. We applied the same analysis we did in the hydrogen bonds of the alpha helix to a crucial salt bridge found in crambin. It is understood that Asn46 and Arg10 salt bridge together holding together the protein tightly and ensuring a well-defined protein structure [29]. We find a similar result for such an interaction with coherence near unity for a low frequency range (a subset of the hydrogen bond coherent frequency range). The result may depend on whether such salt bridging persists throughout the entire simulation as opposed to a subset of the trajectory. Nonetheless, the coherence of unity assures us that we can build an effective model of this interaction for low frequency motion.



(a) Frequency dependence of the coherence function between the two salt bridges depicted on the right side.

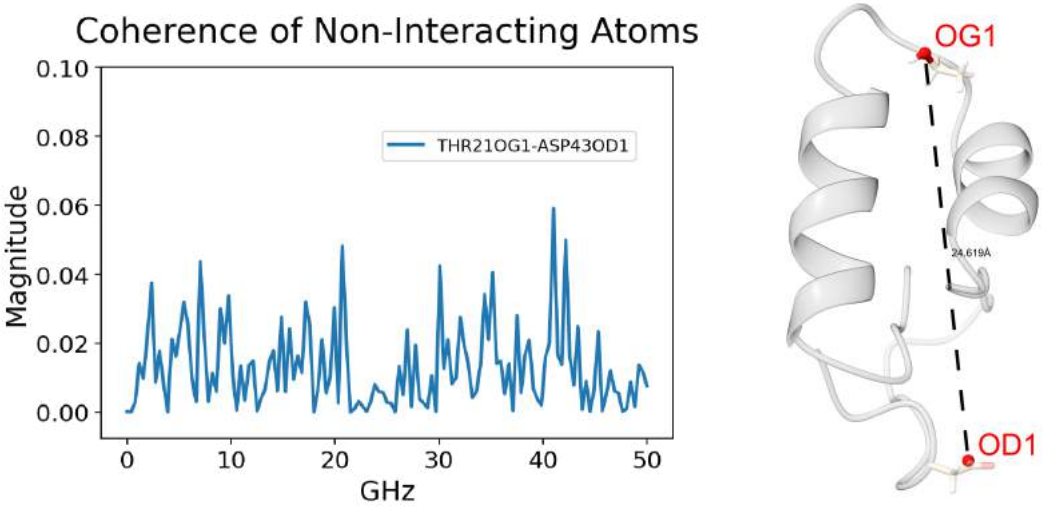


(b) In magenta are the two salt bridges between Asp46 and Arg10. The top magenta line is between Asn46 OXT and Arg10 NE and the bottom line is between atoms Asn46 O.

Figure 2.3. We depict the coherence between salt bridging atoms and the salt bridge itself. We observe a very similar coherence regime to hydrogen bonds.

2.3.2 Non-Coherence of Atoms. As a control for the coherence we study the pairwise interaction of two atoms far away from each other. Because of the

inherent non-linearities in protein movement we expect that such atoms have a low pairwise coherence. We take two atoms that are greater than 24 angstroms apart and compute their pairwise coherence. Their motions are too non-linear for there to exist a frequency function between them evidenced by coherence near ˜0.02.



(a) Near zero coherence for two pairs of non-interacting atoms gauruntees that we have no linear system or effective model between the atoms.

(b) These two atoms are non-interacting given their greater than 24 Angstrom distance from one another.

Figure 2.4. An example of non-coherence between two far-apart atoms chosen at randomly that do not no interact under MD simulations.

2.4 Coherence of Structures

We perform the same analysis we have in the last section but instead average the positions of the atoms in each leaf, or turn, of the main alpha helix. Namely, for each leaf L_k $k = 1, \ldots, 6$ we average $\mathbf{r}_i$ of each backbone atom i in leaf k

$$\mathbf{l}_k(t) = \frac{1}{|L_k|} \sum_{i=1}^{|L_k|} \mathbf{r}_i(t). \tag{2.3}$$

We then calculate the coherence between $d_{\mathbf{l}_k}(t)$ and $d_{\mathbf{l}_{k+1}}(t)$ for $k = 1, \ldots, 5$. Because the coherence function has a similar shape to coherence in H-bonds (see Figure 2.2),

we display the average coherence between 0.391 GHz and 5.08 GHz in Table 2.2. The average position of leaf k, given by $\mathbf{l}_k$, is displayed in Figure 2.5. We observe a similar story to the fundamental chemical interactions (H-bonds and salt bridges) between the two atoms, an effective rigid spring model between the adjacent leaves of the helix. The specific underlying spring model has yet to be determined but we know that H-bonds, salt bridges, and adjacent leaves of alpha helices behave and motion – with respect to input and output – nearly identically. This approach allows us to develop coarse grained models of intermolecular and interstructural interactions. We expect that this results for the leaves is a consequence of the rigidity of the protein crambin and the way that multiscale protein motion is correlated across its structures. More work is necessary to understanding the significance, if at all, of the magnitude and phase of the estimated frequency function. Identifying these larger

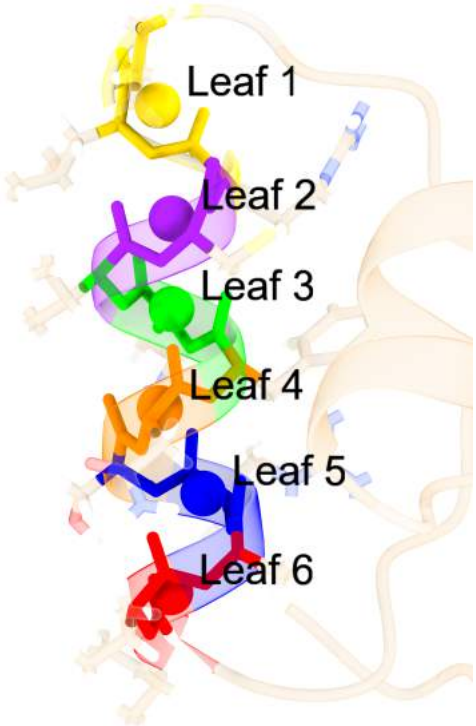


Figure 2.5. Averaged positions of each leaf of the alpha helix.

groups of proteins is crucial to providing a quantitative approach to understanding what objects are defined from protein motion. From this perspective, we need to study all pairwise coherences and identify *communities* of atoms that behave *together* as linear systems. Because of the non-linear correspondence between protein primary and tertiary structure, we need to study the coherence matrix.

Table 2.2. Mean Coherence Results of Leaves of Main Alpha Helix

Input and Output	Mean Coherence	Mean Magnitude	Mean Phase
leaf 1 and 2	0.948	0.865	0.002
leaf 2 and 3	0.938	1.044	0.003
leaf 3 and 4	0.929	0.938	0.006
leaf 4 and 5	0.936	1.070	0.010
leaf 5 and 6	0.918	1.058	0.001

CHAPTER 3

THE COHERENCE MATRIX AND COHERENT COMMUNITIES

Coherence identifies correlations between input and output atoms. We are motivated to consider all pairwise coherent interactions, define a coherence matrix, and understand how to uncover bodies or communities of atoms with a high density of coherent interactions. We interpret these results physically in the context of the proof given in Equation (1.17).

3.1 Introduction

Previous analyses focused on relating coherence to our intuition regarding what interactions biochemists would expect to be linear. We take a data-driven perspective on analyzing our coherence as a global metric of protein motion. Namely, we calculate all the pairwise coherences and define our coherence matrix $\mathbf{C} = [C_{ij}]$ where

$$C_{ij}(f) = \frac{|G_{ij}(f)|^2}{G_{ii}(f)G_{jj}(f)}. \tag{3.1}$$

where i and j are indices of atoms i and j in our protein. We will ignore the frequency dependence, for now, and fix a particular frequency or range of frequencies and discuss implications in Section 3.4.1. Notice that $\mathbf{C}$ is symmetric and positive semi-definite.

3.2 Clustering the Coherence Matrix

We intuitively think about coherent clusters in proteins as rigid bodies. Because motion is highly linear in a region dense with coherent pairs we expect some type of rigidity in the group's motion. Various clustering algorithms are employed and explored in the following sections. The goal is not to perfectly identify which atoms are in which clusters but to develop reproducible methods that can be compared based on metrics commonly associated with rigid bodies or other well-defined concepts of protein structure.

3.2.1 Setup. We think of atoms as nodes in a network with their pairwise coherence as a weighted edge. From this perspective $\mathbf{C}$ is a real-valued adjacency matrix for a weighted undirected graph. For particular community detection algorithms, we consider an adjacency matrix representing an unweighted undirected graph

$$\overline{\mathbf{C}}^{\kappa}_{ij} := \begin{cases} 1 & C_{ij} \geq \kappa \\ 0 & \text{else} \end{cases} \tag{3.2}$$

where we only consider edges whose weight (pairwise coherence) is at least κ. We demonstrate the dependence of body identification on κ. We say a threshold κ of 0.9 is adequate in confidently determining the existence of a linear system between input atom i and output atom j referring back to our discussion in 1.2.3.

3.3 Identifying Communities of Atoms

There are two paradigmatic approaches to interpreting $\mathbf{C}$ and $\overline{\mathbf{C}}$. From a statistical perspective, $\mathbf{C}$ represents a frequency domain correlation function. Namely when $C_{ij} = 1$ we know there exists a linear relationship (frequency function) between atom i and j. On the other hand, an algorithmic perspective views $\overline{\mathbf{C}}$ as an undirected graph of atoms subject to commonly used community detection graph algorithms. We coincidentally followed similar ideas outlined in [30] but expand their results by considering the coherence function as a correlation. These two fundamentally different perspectives offer insight into the underlying linear structures within proteins. We view agreement between these two methods as a validation of an effective model for reduced rigid body motions.

For the statistical viewpoint, we take standard approaches to analyzing pairwise correlation matrices such as through PCA. From a graph perspective, we draw from a great depth of literature on community detection in social networks. Communities of atoms are understood to form rigid bodies which are connected by linear

or non-linear mechanisms to other communities. We first provide various approaches to identifying these communities and then provide a metric and ideas to evaluate identified communities.

When proposing various approaches to identifying bodies we must admit that there are no unique definitions of rigid bodies. We are limited by multiple factors. The lack of sampling from the MD simulation limits is always primary. We are not guaranteed for the protein to preserve correlations across conformation changes. The way energy is moved throughout the protein may be non-linear and non-homogeneous. Correlations offer a way to probe the *linear* transfer of energy throughout the protein. From this perspective, correlations must depend on the protein's conformation and dynamics that constrain and are informed by the way energy is moved within the protein. General protein structure often linearizes the mechanism of energy transfer. For example, secondary structures like alpha helices have measurable spring constants and displacement laws. We have also shown that H-bonds, salt bridges, and leaves of the alpha helix behave linearly.

Coherence as a correlation provides a physically interpretable metric of uncovering how energy is transferred in a structured system. Electrical circuits have components that create flow arising from boundary conditions. Coherence can identify whether such flow is due to linear components like resistors or capacitors. General linear systems analysis provides an estimation of the aggregated interaction between the components by estimating impedance curves. If we ran current through a solution of salt water (a system with no structure), we likely expect there to be no coherence because energy transfer to widely non-linear in such systems. *However*, biology provides us hope. Biology is a hierarchy of structures that linearizes the flow of energy. For example, muscle acts like a resistor and cellular membranes act like capacitors [31]. These amazingly accurate models only can arise from biology's use of

hierarchical structures from proteins to protein complexes to cells to tissues. There is a perspective that these structures *must* be linear because they are robust (see our discussion in 2.3). We now try to find sufficient conditions for identifying when robust structures *imply* linearity. Secondary structure is often conserved across various physiological, chemical, and physical conditions. Intrinsically disordered proteins like p53 have fluid structures that only become alpha helices upon interaction with a target protein [32]. Robust structures are important to carrying out function. In general, we are not necessarily guaranteed to have coherent structures preserved under a similarly large range of conditions. Although, if coherence coincides with secondary structure in some instances we can expect a corresponding robustness. This discussion motivates how we choose our object detection algorithms and how to evaluate the identified coherent structures. Our definition of a structure relies on our hope to utilize it in understand protein dynamics and allostery. For example, we may require all atoms within a body to be coherent with one another. Such a condition may be too discerning and we may find no such regions with large amounts of atoms. We are also motivated by the definition of secondary structures where a local definition of, for example, hydrogen bonding every fourth residue defines a leaf that defines a full alpha helix. Non-adjacent leaves are not hydrogen-bound to one another but still are part of the larger alpha helix.

This discussion motivates considering another order or protein structure based off of correlations measured as coherence. Quantitative approaches to identifying chemical interactions, protein structure, and allostery have been limited. Many correlation metrics do not offer the same physical interpretation as coherence. Linear systems analysis is a well-established methodology in engineering that leads to interpretable models of the system one is probing.

3.4 Results for the Mu Opioid Receptor

We performed coherence analysis to the Mu Opioid Receptor (MOR) as we did to crambin. We obtain a coherence tensor of dimensions $4960 \times 4960 \times 129$ where 4960 is the number of atoms in the receptor and ligand and 129 represents the number of frequency points from 0 to 5/3 GHz (determined by the sampling rate of our simulations). Coherence matrices were obtained for the three different simulations of the MOR with lofentanil 3R4S. We plot these coherence matrices evaluated at 0.13 GHz as seen in Figure 3.1 and use these matrices for any community detection algorithm. Clearly coherence depends on the conformations observed in the simulation and thus depends on initial conditions. Figure 3.2 demonstrates one approach to these variations in coherence. We take the minimum of the coherence across the three runs to only consider coherent interactions preserved under different initial conditions. We expect these minimum coherent interactions to be robust. It should be acknowledged that many statistics of proteins vary across time and simulation depending on where the protein is in the energy landscape. H-bonds are broken and reformed many times throughout simulation but we still consider there to be a well-defined object that is called a H-bond.

3.4.1 Frequency Dependence in Coherence. We also visualize the frequency dependence of the coherence matrix by considering a subset of the pairwise interactions. For many interactions, coherence decays similarly as it does shown in Figure 2.2. It is not entirely clear why high coherence observed in the MOR occurs at lower frequencies (between ˜0.13 GHz and ˜0.6 GHz) versus the crambin's coherent range (˜0.39 GHz and ˜5.1 GHz). One possible rationale is MOR has 350 amino acids versus (˜38.5 kDa) crambin's 46 amino acids (4.74 kDa). Aggregate motion in larger proteins should be slower than in smaller proteins which corresponds to a smaller frequency. As a very rough calculation, the MOR is roughly 8 times heavier than

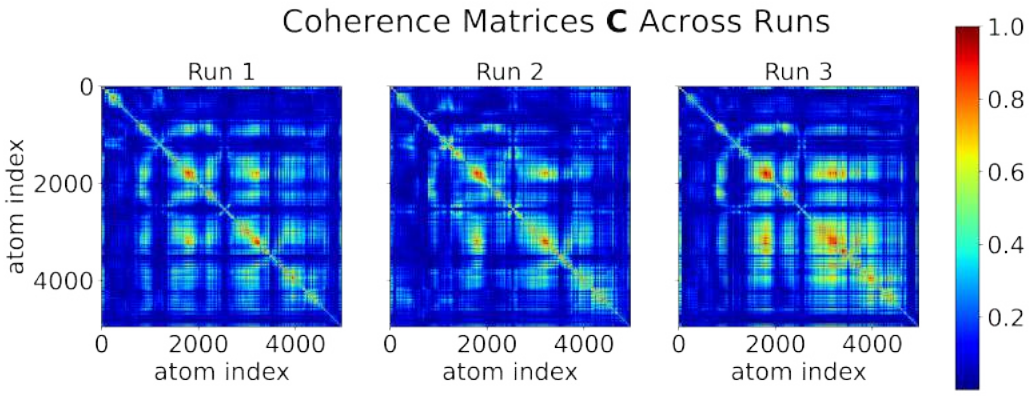


Figure 3.1. Coherence matrices **C** of three independent runs of the MOR with lofentanil 3R4S. We see very similar coherence matrices from the first two simulations and a more strongly coherent structure in the third run.

crambin and linearly scaling the coherent frequency range of the MOR by this factor gives 1.1 GHz to 4.8 GHz. Whether this coherent frequency range scaling exists across other proteins has not been checked. There is no immediate reason why these frequency ranges should be linear in mass and is something that should be pursued further.

There are some particular frequencies where coherence for many interactions drops. Because the MOR is not a globular protein, energy is transferred in many more non-linear ways than in crambin. Some interactions have a slightly lower coherence for low frequencies where it rises for higher frequencies than roughly 5 GHz. It is very difficult to interpret coherence, frequency functions, and power spectra at individual frequencies and is future work. For the rest of the discussion, we generally do not consider frequency dependence in our results yet it is necessary to future physical interpretations. Understanding this work in the context of vibrational modes, wavenumbers, and the field of spectroscopy offers another fascinating avenue for the

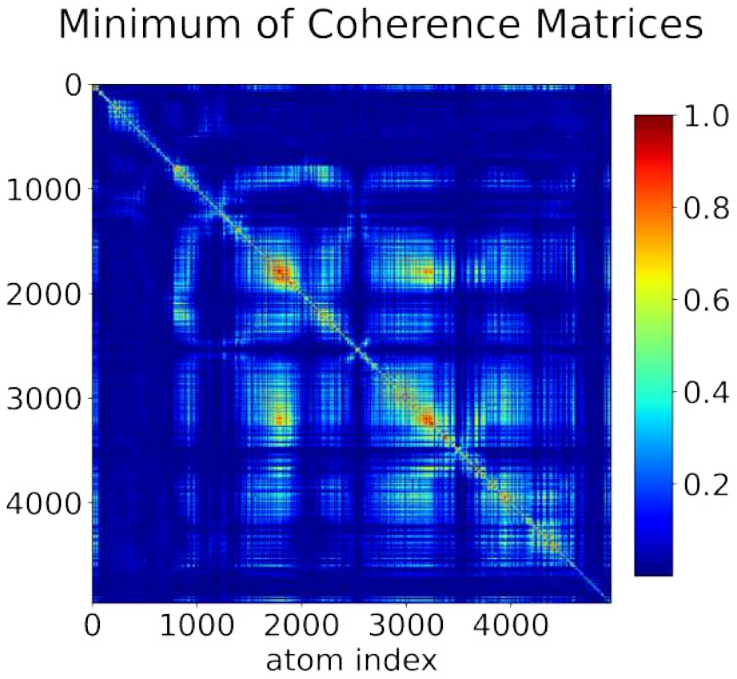


Figure 3.2. We take the minimum coherence of a pair across the three runs. In other words, the coherence matrix $\tilde{\mathbf{C}}_{ij} = \min(\mathbf{C}^1_{ij}, \mathbf{C}^2_{ij}, \mathbf{C}^3_{ij})$

study of coherence [3].

3.4.2 Evaluation Metric. There are many non-unique ways to define and measure the quality of a community of atoms. The particular metric depends on the context of the question, its interpretation, and level of abstraction. Because we interpret coherence as identifying linear systems that can be modelled by effective springs, we take the view that communities identified are rigid bodies. They can either be rigid internally, as defined by looking at the variation of the pairwise distances of particles, or externally by the variation of their movement with respect to the whole protein. We primarily opt for the latter interpretation as we view coherent communities having tightly correlated motions with each other rather than with the protein.

A common metric for understanding group motions of atoms is the root mean

square deviation (RMSD). Suppose we have a finite set of atoms S_t at time t in the simulation with L number of atoms. We define the $\text{RMSD}(t, t')$ as

$$\text{RMSD}(t, t') = \sqrt{\frac{1}{L}\sum_{i=1}^{L} ||S_t^i - S_{t'}^i||^2} \tag{3.3}$$

where $S_t^i = (x_t^i, y_t^i, z_t^i)$ the 3D coordinate position of atom i at time t. For brevity, we write $\text{RMSD}(0, t)$ as $\text{RMSD}(t)$.

In order to evaluate a community of atoms, we align frame t to frame t' by the protein's the backbone atoms using `MDAnalysis` [33, 34]. If there are no backbone atoms in our community, we do not consider it for the sake of reproducibility. We then compute the pairwise time $\text{RMSD}_m(t, t')$ matrix for each community m in $\{S_t^m\}_{m=1}^M$ where M is the number of communities of atoms. A pairwise RMSD average across the simulation debiases the results from being aligned to a single frame [35]. The time average of community m's RMSD (defined below) gives an reduced metric as a function of some threshold for our community detection algorithm.

$$\overline{\text{RMSD}}_m^T = \frac{1}{T^2}\sum_{t,t'=1}^{T} \text{RMSD}_m(t, t'). \tag{3.4}$$

We vary a threshold parameter that identifies groups in our algorithm so that we can understand the effects of including or excluding atoms in groups on $\overline{\text{RMSD}}_m^T$. We expect that if our algorithm becomes more stringent in requiring a higher density of coherence per community that $\overline{\text{RMSD}}_m^T$ decreases. Often this will take the form of the time averaged RMSD decreasing monotonically as a threshold parameter κ increases. We must also be conscious that RMSD will decrease naturally as fewer atoms are in the community because of a more stringent threshold. We also consider the size of the communities we identify and expect that we can uncover truly rigid communities of atoms that maintain a core group of atoms while maintaining a low RMSD.

3.4.3 PCA. We consider principal component analysis (PCA) of $\mathbf{C}$ to understand which atoms account for most of the variation in the principal components of $\mathbf{C}$.

This is a common approach in understanding relationships among variables given their pairwise correlation or covariance matrix [36].

In particular, we perform Singular Value Decomposition (SVD) to the matrix $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ with $\mathbf{U}$ and $\mathbf{V}$ unitary matrices. PCA enables us to identify particular linear combinations of atoms that contribute the greatest to the total protein coherence. These combinations of atoms are called the principal components (PC) of $\mathbf{C}$. Atoms that contribute non-trivially to a specific PC are said to be part of group because they are correlated with other atoms in said PC. That is, if the i and j-th component of, say, the first principal component (PC1) p_i and p_j are "large" then we think that these atoms are correlated. Atoms i and j contribute more mass to the linear combination corresponding to the principal component.

Another interesting interpretation of the PCA of $\mathbf{C}$ is that principal components are orthogonal. In essence, these PCs capture different groups of atoms that have coherent motions. Different PCs offer better insight into the communities that exist within the protein.

3.4.3.1 Principal Component Explanation Power. We perform PCA on the coherence matrix $\mathbf{C}$. We decided to look at the first ten principal components (PCs) given the Skree plot (Figure 3.3) of the explainability of the variance for each principal component. This is a heuristic often used in PCA given that the first four PCs capture more than 80% of the variance in the coherence. We perform our community identification algorithm on the first ten PCs because in total they account for nearly 90% of the total variance in the coherence matrix.

3.4.3.2 Community Identification Algorithm. We are motivated to find atoms who contribute relatively large weights to the lowest PCs because we interpret these atoms as forming strongly correlated communities. We design the algorithm as fol-

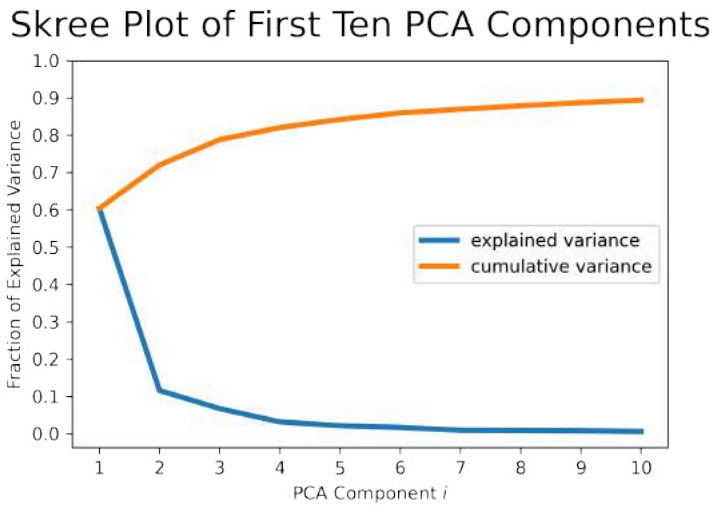


Figure 3.3. We plot the first ten PC's explainability ratio of the variance and see that we capture nearly 90% of the variance in $\mathbf{C}$ using the first ten PCs.

lows:

1. Compute the first M PCs of our $\mathbf{C}$ matrix which is $n \times n$.

2. We let $k = M, M-1, \ldots, 1$. For the kth PC (denoted by PC^k), look at the distribution of components of the PC vector $\{\mathrm{PC}_i^k\}_{i=1}^n$.

3. Choose the pth percentile for a chosen p of the distribution $\{\mathrm{PC}_i^k\}_{i=1}^n$ and calculate $F^{-1}(p)$ where F is the quantile function.

4. We say atom j is in cluster k if $|\mathrm{PC}_j^k| \geq F^{-1}(p)$. If atom j is in a cluster already, it now belongs to cluster k.

In words, we simply look for atoms that have a non-trivial eigenvector component (PC component) according to some threshold. The last step ensures that atoms are always in the lowest index PC they can be. In order to empirically validate our object detection we compute the averaged RMSD described in Section 3.4.2. We also consider functions of PC_j^k such as its absolute value in step 4 of our algorithm because non-trivial components indicates a contribution to the correlation of a PC.

3.4.3.3 Results. We color the first PC red, the second PC orange, and continuing on for all ten PCs according to the colors of the rainbow interpolating for the extra number of PCs. We observe that the structures identified by PCA, for the first three PCs – roughly – are spatially concentrated and trace out two alpha helices. However, higher PCs contain atoms across the entire protein. The disparate nature of the blue atoms, which correspond to PC7, is due to PC7 encompassing a large portion of the protein but the identified atoms are part of lower order PCs like the first one. We have simply identified a mode of correlation whose atoms span across the entire protein. If we vary the percentile described in the PCA algorithm we obtain three structures

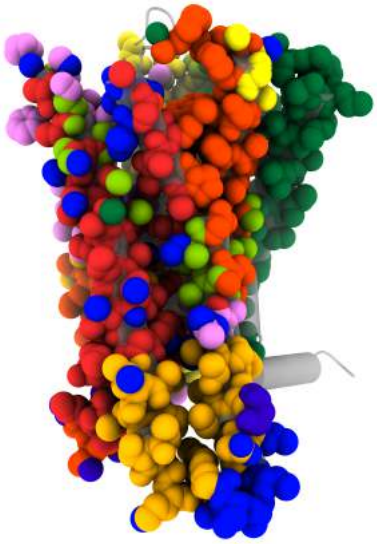


Figure 3.4. The communities identified by PCA are not spatially concentrated necessary but rather trace out some structures while accounting for PC correlations across the protein.

corresponding to percentiles of 95, 90, and 85. We observe more separation in the communities as we increase the percentile. Atoms with higher PC component values may be spatially correlated because pairwise coherence is inversely proportional to the pairwise distance. Much of the change in the identified communities arises in roughly the first three PCs where fewer atoms are identified with these PCs. The atoms that are lost are not associated with the remaining PCs.

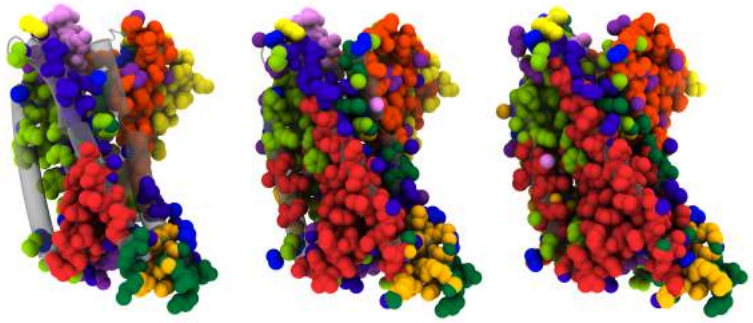


Figure 3.5. Three runs of the PCA algorithm using a threshold of 95, 90, and 85 from left to right.

3.4.4 Girvan Newman Algorithm. In their seminal paper on the algorithm, Girvan and Newman established a graph algorithm to identify communities on graphs [37]. The foundational metric is that of edge centrality defined similarly to node betweenness centrality. Given a graph (V, E) with vertices V and edges E, an edge's $e \in E$ centrality is the number of shortest paths that run through the edge e given two nodes (often normalized by the number of shortest paths that run through e). Intuitively, wish to remove highly central edges two communities exist between these two nodes.

The Girvan Newman algorithm encapsulates the theme of locally structured regions connected by linear or non-linear mechanisms discussed in section 3.3. Inherently the algorithm identifies communities connected by at least one linear interaction and removes it to delineate regions of the protein. Densely coherent regions will be uncovered as separate communities.

We treat the coherence network as an adjacency matrix where each atom is a node and an edge exsits only if two atoms have a coherence of at least κ. Our most important value of κ is 0.9 but we also consider 0.85 and 0.95. We use `networkx`'s implementation of the algorithm called `girvan_newman` [38].

3.4.4.1 Results. As we can see in Figure 3.6, the identified clusters trace out

secondary structures and even identify regions that connect two secondary structures. The entire protein is partitioned into coherent regions. We study some of the quantitative properties of the groups below. The communities coincide partly with secondary structure but interestingly include suprastructures that involve the interaction of multiple structures and motifs. Notice the atoms of two neighboring alpha helices colored gold in the middle protein structure of Figure 3.6. It is also interest-

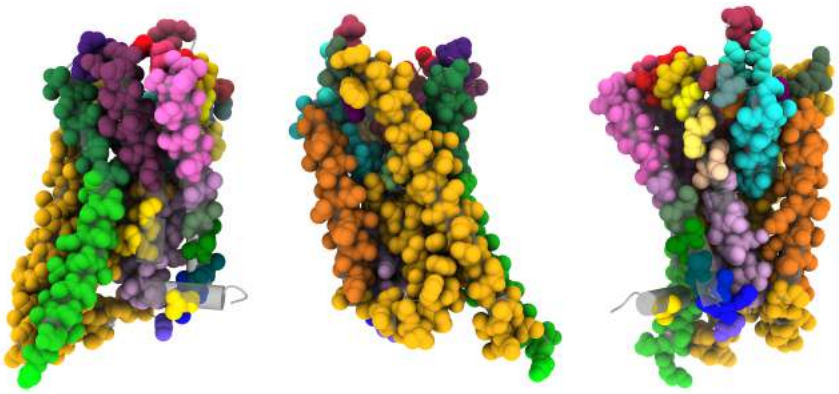


Figure 3.6. Three captures of the MOR each 120° rotated clockwise where the communities are identified by the Girvan Newman algorithm.

ing to change our κ threshold parameter from 0.95, 0.90, and 0.85 where we see that different resolutions of structures are identified, as seen in Figure 3.7. For example, the blue central helix in the middle protein is identified in the regime of $\kappa = 0.90$ and 0.85 but is split up into two separate helices (cyan and orange) in the $\kappa = 0.95$ regime. Different hierarchical orders of structures are identified between thresholds. In the left most protein in Figure 3.7, the purple, magenta, and red communities represent individual leaves of the helix they trace out whereas for other thresholds it is identified as a single helix. Relating coherence value thresholds to various structural motifs would provide a quantitative measure could identify the boundaries of structures purely from their simulation. The individual leaves of the helix are tightly coherent within themselves but less coherent with their adjacent leaves.

3.4.4.2 Robustness. We compared the Girvan Newman algorithm results on an

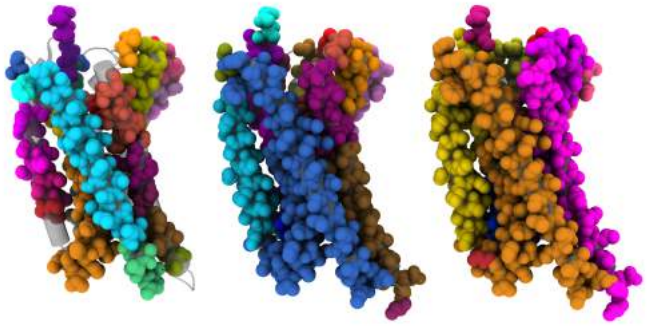
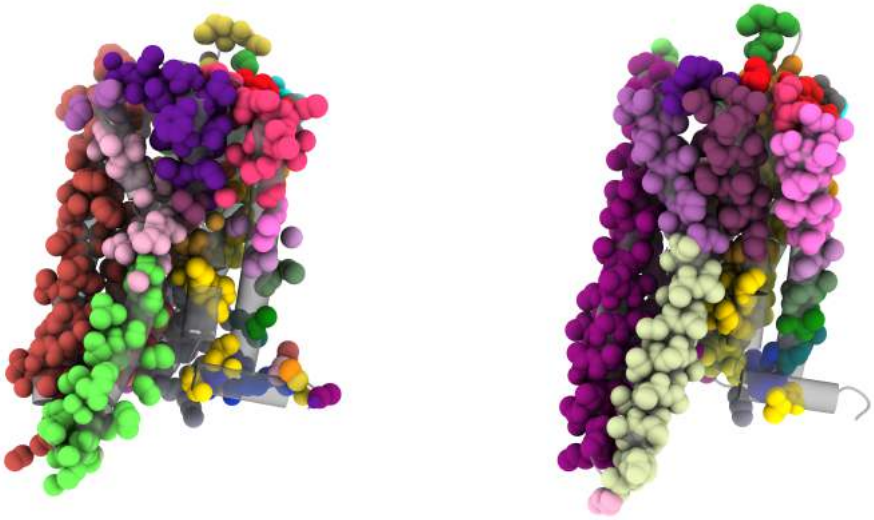


Figure 3.7. Three runs of the Girvan Newman algorithm using a threshold of 0.95, 0.90, and 0.85 from left to right.

coherent adjacency matrix both including and excluding the opioid ligand lofentanil 3R4S. Our motivation is to understand the role identified communities play in protein structure. We, in effect, try performing a type of deletion experiment by either removing the ligand or a residue and reidentify the communities. This may be an approach towards performing computational site directed mutagensis and determining how the coherence protein structures would modify in the absence of a ligand or residue. The interpretation of the results is still opaque and requires further investigation but we propose this methodology as a an exciting potential direction for understanding coherence.

Algorithmically, we redefine our smaller coherence matrix $\tilde{\mathbf{C}}$ by again only considering pairwise coherences of at least κ but also excluding atoms in the ligand lofentanil 3R4S. We perform the Girvan Newman algorithm on this smaller matrix. Interestingly, the identified communities trace out similar atoms except with Figure 3.8(a) showing sparser groups. It is surprising that the 63 atoms of lofentanil being removed from the protein creates a noticeable different in the coherence communities.



(a) Girvan Newman algorithm applied to the MOR *without* the ligand lofentanil 3R4S bound at a threshold of 0.9.

(b) Girvan Newman algorithm applied to the MOR *with* the ligand lofentanil 3R4S bound at a threshold of 0.9.

Figure 3.8. The Girvan Newman algorithm ran on the same coherence matrix of the MOR except with the algorithm excluding the ligand atoms in the left figure.

3.5 Coherence and Pearson Correlation

A commonly used metric for analyzing correlations between signals is the Pearson correlation coefficient ρ [39]. Because we interpret coherence as a type of frequency domain correlation, it is insightful to compare and contrast situations where either is more useful than the other.

3.5.1 Analytical Comparison. In our demonstration of coherence as a correlation metric in Section 1.2.3, we showed that the coherence between $x(t) = \sin(2\pi t)$ and $y(t) = \sin(2\pi t + \pi/2)$ is given by $C_{xy}(f) = \delta(f-1)$. Coherence clearly demonstrates the perfect correlation between the two signals. The frequency function $H(f) = \delta(f-1)e^{i\pi/2}$ between the two signals can even be numerically estimated. Note that H only exists at 1 Hz and is not uniquely defined at any other frequency.

The Pearson ρ has a similar structure to the coherence with covariance (cross power) in the numerator and variances (auto powers) in the denominator.

$$\rho = \frac{\mathrm{Cov}(x(t), y(t))}{\mathrm{Var}(x(t))\mathrm{Var}(y(t))} = \frac{\int_0^\infty \{x(t) - \mathbb{E}_t[x(t)]\}\{y(t) - \mathbb{E}_t[y(t)]\}dt}{\sqrt{\int_0^\infty \{x(t) - \mathbb{E}_t[x(t)]\}^2 dt \int_0^\infty \{y(t) - \mathbb{E}_t[y(t)]\}^2 dt}} \tag{3.5}$$

where the expectation is taken over time. The Pearson correlation identifies the correlation between x and y as zero.

$$\rho = \frac{\int_0^\infty \sin(2\pi t)\sin(2\pi(t+\pi/2))dt}{\int_0^\infty \sin^2(2\pi t)dt \int_0^\infty \sin^2(2\pi(t+\pi/2))dt} = 0 \tag{3.6}$$

The denominator is clearly non-zero and the numerator is zero implying the Pearson correlation is zero.

Here we see that the Pearson correlation and coherence give drastically different results and resolutions. The problem of time lag is fundamental to the Pearson ρ. Any type of periodic structure within signals will be missed by the Pearson correlation and consequently missing important interactions. Signals with identical frequency and small phase delay will necessary have sub unity Pearson correlation. From these perspective, coherence offers a greater resolution in particular signals that have some periodicity in them. Pearson is agnostic to any notion of periodicity and may offer broader applicability in this sense. See [40, 41] for a discussion on robustness issues of estimating the Pearson coefficient and alternatives tried in MD. We wish to experimentally determine whether there are situations where they are similar or different and thus present our metrics for evaluating communities and compare the community detection algorithms using both Pearson and coherence.

3.5.2 Comparison in Community Identification. Pearson correlation is one of the most common ways of identifying correlations within a protein's movement. We compare the Pearson correlation to the coherence function by comparing the communities both metrics identify using algorithms outlined in Section 3.3.

3.5.2.1 PCA. We apply the PCA algorithm directly to the estimated Pearson correlation matrix described earlier. As explained in Appendix A, the Pearson correlation PC distributions are much wider. This means more atoms are considered part of any one PC and thus higher percentiles are necessary to identify communities of comparable size to the coherence matrix. The major difference between Pearson and coherence in the PCA algorithm is the PC ordering of similar communities. The left most helix is identified, in Figure 3.9, as part of PCs 2 and 10. However in the PCA algorithm applied to the coherence matrix, shown in Figure 3.5, does not identify the left most helix. These differences arise from the distributions of PC components. It should be noted that PCs with small explanability are less sensitive than lower order PCs [42]. We should consider the robustness of PCA and Pearson estimation when comparing results between these two metrics.

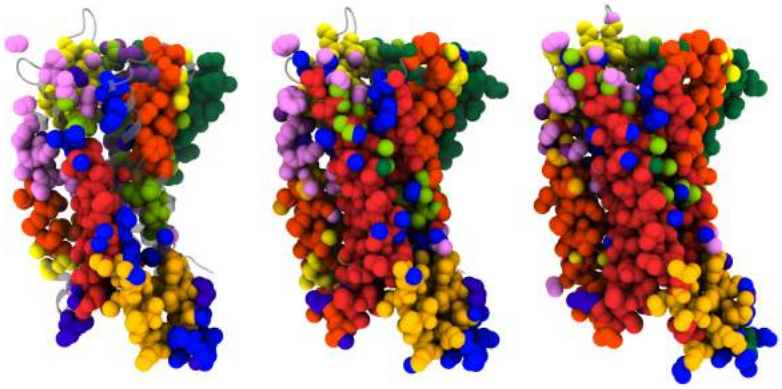
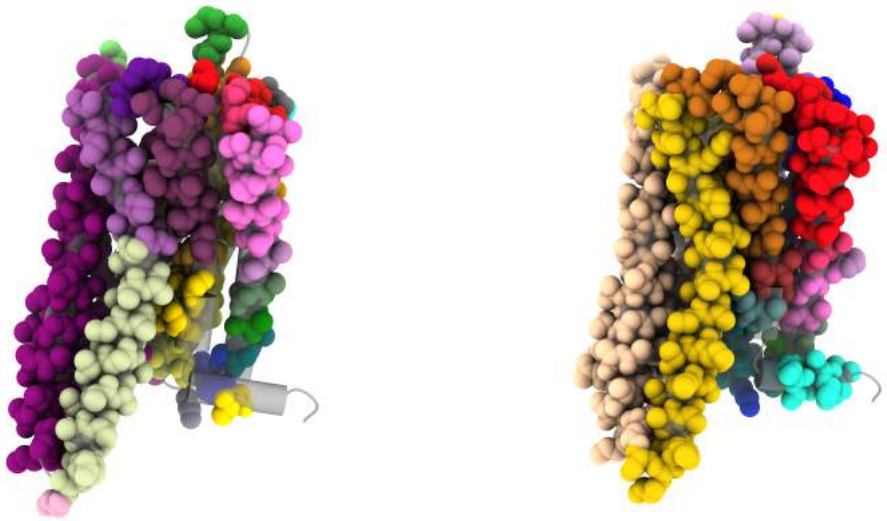


Figure 3.9. Three runs of the PCA algorithm applied to the Pearson correlation matrix using a threshold of 95th, 90th, and 85th from left to right.

3.5.2.2 Girvan Newman. We apply the Girvan Newman algorithm to the absolute value of the Pearson matrix with a threshold of 0.9, the same as when analyzing the coherence matrix. In Figure 3.10 we demonstrate the differences in the identified communities from both correlation metrics. Atoms with identical colors are in the same community. Strikingly, the coherence and Pearson correlation identify similar bodies as seen, for example, in the left hand side of both proteins (purple

in 3.10(a) and peach in 3.10(b)). However, the kinked helix is identified as two separate regions (light purple and green in 3.10(a)) by coherence and as one region by Pearson (gold in 3.10(b)). We also plot, similarly to Figure 3.7, the Girvan Newman



(a) Results using the coherence matrix with a threshold of 0.9.

(b) Results using the absolute value of the Pearson matrix with a threshold of 0.9.

Figure 3.10. Each group of colored atoms defines a community identified by the Girvan Newman algorithm.

algorithm applied to the regular Pearson correlation. Given that there are more identified Pearson correlations than otherwise, we find that it identifies much larger communities atoms. The right most protein (threshold of 0.85) is a majority of the same community because of so many extra correlations. This means having a higher threshold is important to identifying truly tightly correlated regions, but Pearson is not a robust estimator. Consequently, this makes Pearson a complicated metric to use in an interpretable way.

3.5.3 Evaluating Communities. We present more detailed results of the community identification algorithm in the context of comparing the coherence and Pearson correlation metrics. We see a great deal of correspondence qualitatively and some-

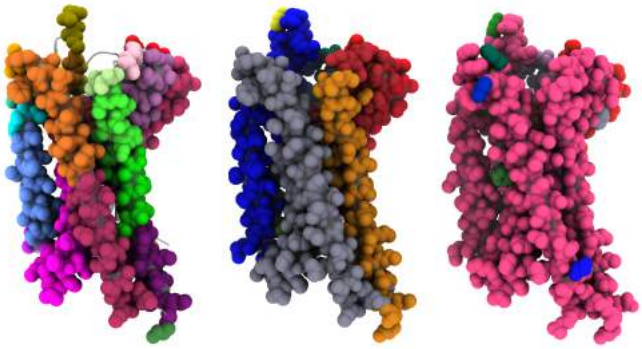


Figure 3.11. Three runs of the Girvan Newman algorithm applied to the Pearson correlation matrix using a threshold of 0.95, 0.90, and 0.85 from left to right.

times quantitatively. The RMSD based metric defined in Section 3.4.2 is computed for each community, or cluster, of atoms identified. A smaller RMSD measure indicates the community is more rigid.

3.5.3.1 PCA. We consider the absolute value of the principal component (PC) vector components using the coherence and regular Pearson matrices. The results are shown in Figure 3.12. There is a great deal of correspondence the communities identified by both coherence and Pearson measures. As the threshold parameter is decreased, larger communities of atoms are identified by each PC accompanied by a slight increase in the RMSD of each community. Interestingly, the first three PCs converge to having the same number of atoms while maintaining a relatively unchanged RMSD. Further investigations should be done into the effect of only considering the first three PCs in the estimation.

3.5.3.2 Girvan Newman. Here we consider the Girvan Newman algorithm applied to the coherence matrix and the regular Pearson correlation matrix. The results are shown in Figure 3.13. We observe similar trends to the PCA results. Oddly, the number of communities identified for the Pearson correlation increases with the threshold indicating that the larger structures may be broken up into smaller ones.

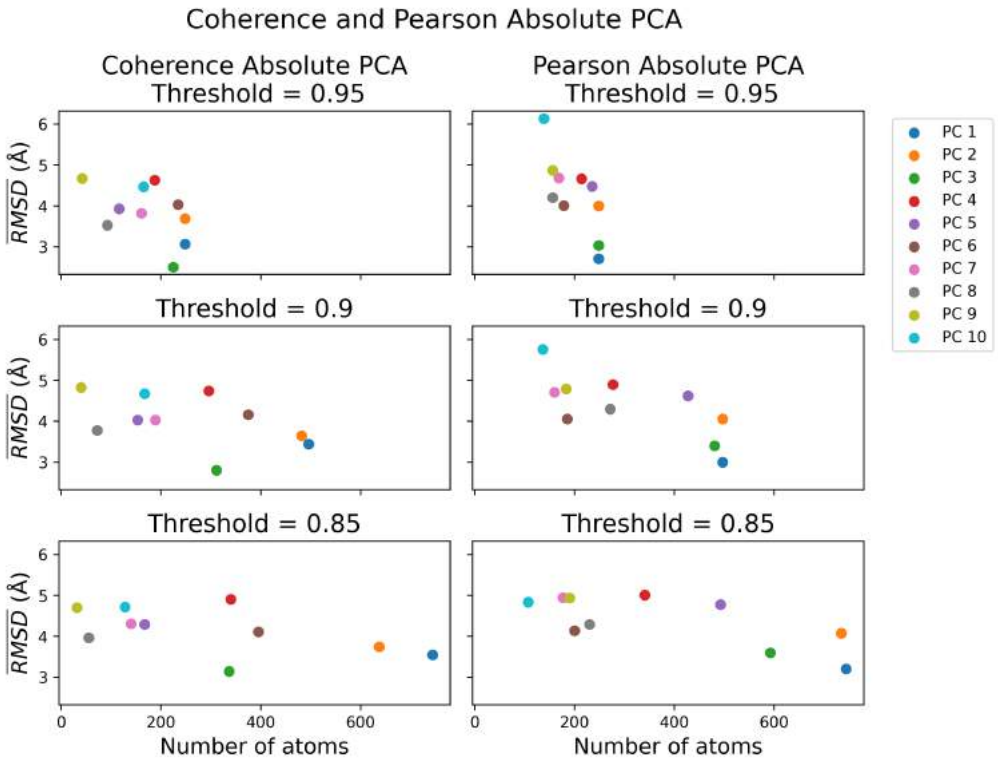


Figure 3.12. The evaluation of communities defined by the PCA algorithm (explained in Section 3.4.3.2) applied to the coherence and Pearson matrices. The communities are compared by the RMSD metric defined in Section 3.4.2.

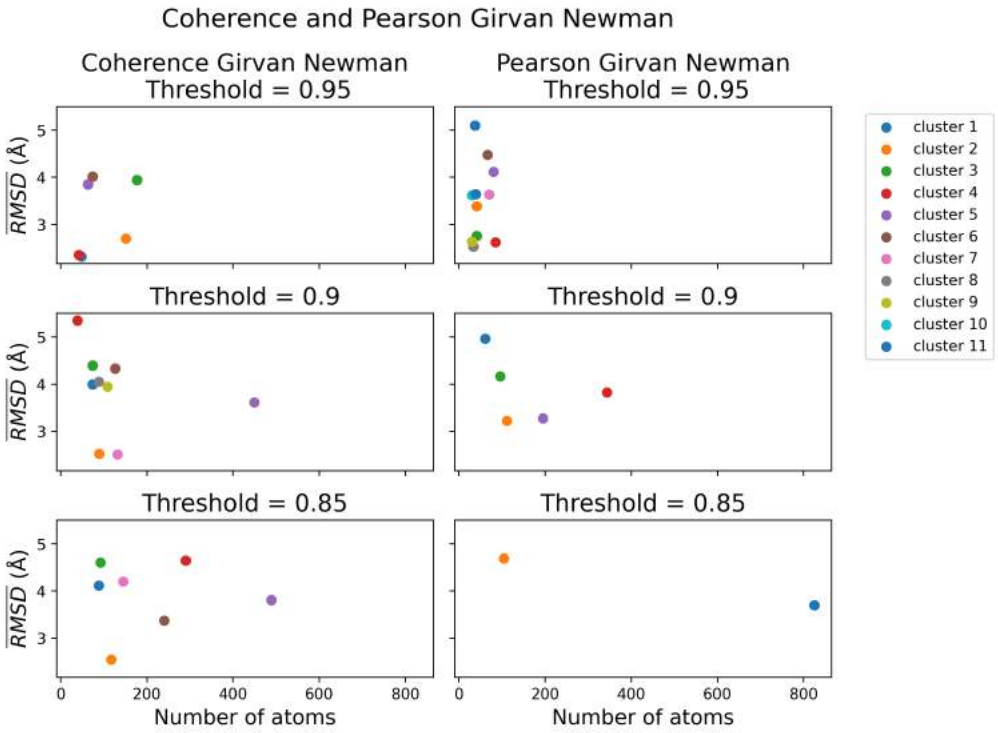


Figure 3.13. The evaluation of communities defined by the Girvan Newman algorithm (explained in Section 3.4.4) applied to the coherence and Pearson matrices. The communities are compared by the RMSD metric defined in Section 3.4.2.

CHAPTER 4

CONCLUSION

Coherence as a metric has been unexplored in the context of molecular dynamics (MD). Its robust estimation and physical interpretation make it an appealing metric to compute and analyze within various biological systems. We have demonstrated its insightful application to well-defined chemical interactions such as H-bonds and salt bridges. These coherent interaction define effective linear models that interact with one another through linear energy transfer at the coherent frequencies. In crambin, we observe that nearly all well-defined chemical bonds have coherent interactions. In the MOR, we find that – due to its non-globular nature – only a subset of these weaker interactions exhibit coherence (excluding interations like covalent bonds). When such interactions are observed, we can properly identify the groups of atoms that underly these highly coherent regions. From this perspective, we demonstrated that such communities can be identified and loosely correspond to rigid systems in some particular cases. There is much biological context that is necessary to interpret the communities themselves and their relative rigidities. For example, the frequency dependence of coherence and frequency functions is an important step towards interpreting which protein motions are linear. We find correlations across pairwise coherence, especially those localized in space, at particular frequencies. Do low frequencies correspond to alpha helical or whole protein movement? Is medium frequency ranges a subset of these higher order structures? From this perspective, the Pearson correlation is too coarse of a metric to parse various protein motions. Incorporating frequency dependence into correlation offers many interesting insights into elucidating structure and function. Many proteins have well-defined functions that provide direction into interpreting which structures exist and interact with each other and the extra-protein environment. PCA and the Girvan Newman algorithm are only two approaches that elucidate these linear structures but do not necessarily

indicate anything about their interactions with each other. These are only a subset of the avenues when considering the nascent application of coherence to MD.

We characterize our future interests into two directions: correlation and input-output. Interpreting coherence as a correlation metric has not been its primary application in engineering. Yet, when we think about it as a frequency-analogue of Pearson, we expand the toolkit of algorithms that may be applicable to coherence. PCA is often applied to correlation or covariance matrices and its application to coherence has proven intriguing. Furthermore, cross power represents a type of frequency-analogue of covariance which we have only pursued slightly. We also wish to compare coherence to Pearson in both analytic and biological contexts. We have identified one clear situation where coherence is triumphant but there may be signals where Pearson is more useful. Although, it is important to only consider trajectories that are relevant in the context of MD which involves testing both metrics on various protein systems. We aim to focus on proteins that have well-defined allosteric networks or non-obvious protein substructures that coherence may then be tested on. Communities detected by coherence are necessarily non-unique based on the context in which we wish to analyze them. Incorporating various evaluation metrics of communities would provide physical insight into the behavior of these atoms. Direct future work involves taking the MD-informed chosen communities and looking at the same averaged RMSD metric on various cryo-EM structures of the MOR. Learning protein structure only through computation to inform physical experiment offers exciting connections and interpretations between MD, coherence, and protein structure and function.

Alternatively to the correlation perspective on coherence, we can study protein signals other than displacements of atoms to derive further physical interpretation of protein motion. In the spring example illustrated earlier, we worked with an input force and output displacement. From the coherence and frequency functions,

we could determine whether such a spring existed and its effective parameters. Very similarly, extracting the forcing from the MD simulation, identifying the correct coordinates, and performing the same coherence analysis will let us extract intermolecular strengths in the form of effective spring constants. Other signals to consider are the velocities of atoms because the power of the velocity is the kinetic energy of the atom! These signals may be much more transient that the displacements we observe making it necessary to study the wavelet coherence. Instead of only the frequency domain, we consider a combination of frequency and time that allows for transient linear systems, see Bendat and Piersol for further discussion [3]. Other interesting coherence approaches involve studying multiply input and output systems where we find atoms that together create a linear system for either a single or multiple atoms.

We have only begun to apply the coherence function to MD. Its novel application provides an exciting opportunity to learn linear systems and protein structure purely from simulation. Without any assumptions about the system itself we learn non-apparent communities of atoms within the MOR. Providing a mathematical and biological treatment of coherence as a tool in MD offers exciting research directions.

APPENDIX A

PCA DISTRIBUTIONS

The distribution of components for a fixed principal component (PC) may offer some insight into the structure of the PC and the coherence matrix. Their interpretation is unclear and simply presented for completion in Figure A.1. We also plot the same distributions for the PCA of the Pearson matrix in Figure A.2. One speculation about the difference in the distributions between Pearson and coherence is that Pearson identifies spurious correlations. These drastic differences in the shapes of these distributions indicates the underlying structure of these matrices differ. Although the bodies identified may be similar in the Girvan Newman algorithm, the fundamental structure of the correlations determined by both Pearson and coherence are different.

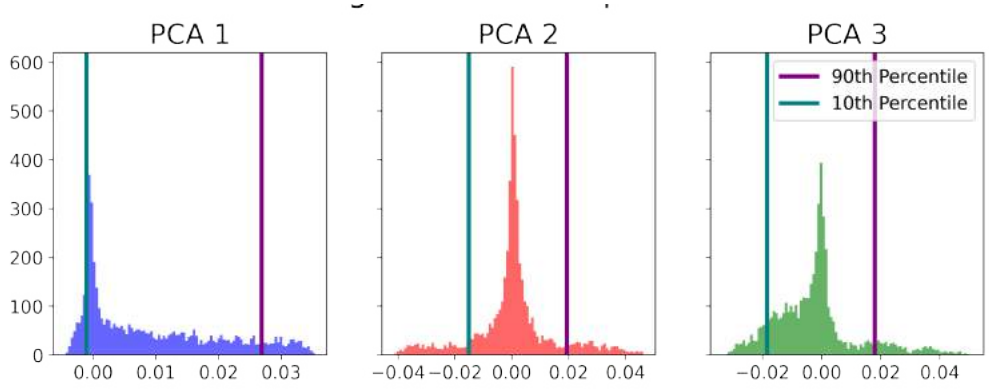


Figure A.1. Distribution of the eigenvector components that correspond to the the first three principal components of the coherence matrix PCA.

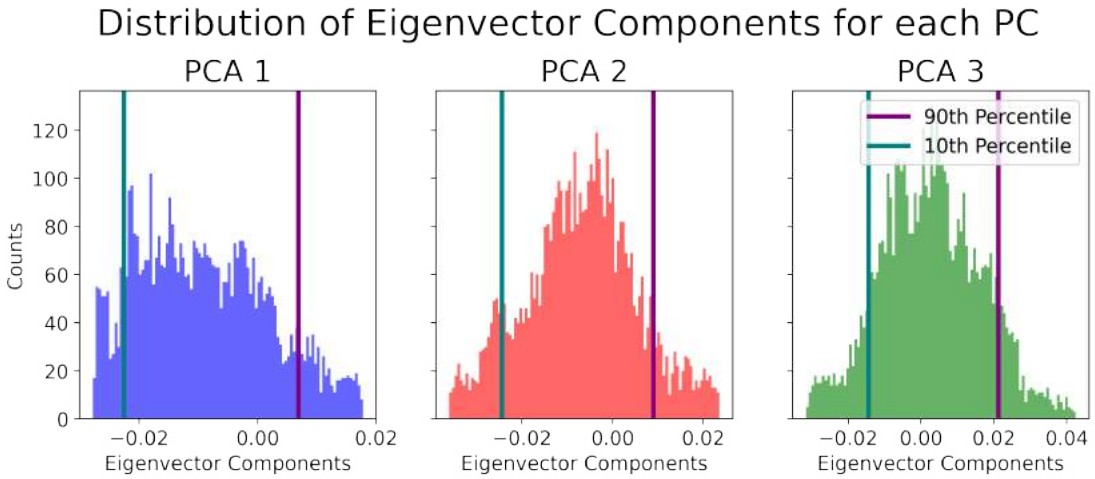


Figure A.2. Distribution of the eigenvector components that correspond to the the first three principal components of the Pearson matrix PCA.

BIBLIOGRAPHY

[1] J. D. Durrant and J. A. McCammon, "Molecular dynamics simulations and drug discovery," *BMC Biology*, vol. 9, p. 71, Oct. 2011.

[2] "Fifty Years of Signal Processing: The IEEE Signal Processing Society and its Technologies 1948-1998," *IEEE Signal Processing Society*, Dec. 2015.

[3] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures, 4th Edition — Wiley*.

[4] J. R. Taylor, *Classical Mechanics*. Sausalito, Calif: University Science Books, null edition ed., Jan. 2005.

[5] R. Eisenberg, "Electrical Structure of Biological Cells and Tissues: Impedance spectroscopy, stereology, and singular perturbation theory," Nov. 2015.

[6] E. Barsoukov and R. Macdonald, *Impedance Spectroscopy — Wiley Online Books*. John Wiley & Sons, Inc., second ed., Jan. 2005.

[7] V. A. Parsegian, *Van Der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists*. Cambridge University Press, Nov. 2005.

[8] J. L. Oncley, J. D. Ferry, and J. Shack, "The Measurement of Dielectric Properties of Protein Solutions; a Discussion of Methods and Interpretation," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 6, pp. 21–23, Jan. 1938.

[9] E. W. Weisstein, "Fourier Transform." https://mathworld.wolfram.com/.

[10] S. Engelberg, *Random Signals and Noise: A Mathematical Introduction*. CRC Press, Oct. 2018.

[11] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, 2000.

[12] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, June 1967.

[13] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, pp. 337–344, Aug. 1973.

[14] J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Hoboken, NJ: Pearson, 5th edition ed., Feb. 2021.

[15] "Welch's power spectral density estimate - MATLAB pwelch." https://www.mathworks.com/help/signal/ref/pwelch.html.

[16] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt,

SciPy 1.0 Contributors, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.

[17] S. Nicholson, D. D. L. Minh, and R. Eisenberg, "H-Bonds in Crambin: Coherence in an α-Helix," *ACS Omega*, vol. 8, pp. 13920–13934, Apr. 2023.

[18] M. M. Teeter, "Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin," *Proceedings of the National Academy of Sciences*, vol. 81, pp. 6014–6018, Oct. 1984.

[19] R. Al-Hasani and M. R. Bruchas, "Molecular Mechanisms of Opioid Receptor-dependent Signaling and Behavior," *Anesthesiology*, vol. 115, pp. 1363–1381, Dec. 2011.

[20] Q. Qu, W. Huang, D. Aydin, J. M. Paggi, A. B. Seven, H. Wang, S. Chakraborty, T. Che, J. F. DiBerto, M. J. Robertson, A. Inoue, B. L. Roth, S. Majumdar, R. O. Dror, B. K. Kobilka, and G. Skiniotis, "Structural insights into distinct signaling profiles of the μOR activated by diverse agonists," Dec. 2021.

[21] A. Koehl, H. Hu, S. Maeda, Y. Zhang, Q. Qu, J. M. Paggi, N. R. Latorraca, D. Hilger, R. Dawson, H. Matile, G. F. X. Schertler, S. Granier, W. I. Weis, R. O. Dror, A. Manglik, G. Skiniotis, and B. K. Kobilka, "Structure of the M-opioid receptor–Gi protein complex," *Nature*, vol. 558, pp. 547–552, June 2018.

[22] W. Huang, A. Manglik, A. J. Venkatakrishnan, T. Laeremans, E. N. Feinberg, A. L. Sanborn, H. E. Kato, K. E. Livingston, T. S. Thorsen, R. C. Kling, S. Granier, P. Gmeiner, S. M. Husbands, J. R. Traynor, W. I. Weis, J. Steyaert, R. O. Dror, and B. K. Kobilka, "Structural insights into M-opioid receptor activation," *Nature*, vol. 524, pp. 315–321, Aug. 2015.

[23] A. Gillis, A. Kliewer, E. Kelly, G. Henderson, M. J. Christie, S. Schulz, and M. Canals, "Critical Assessment of G Protein-Biased Agonism at the μ-Opioid Receptor," *Trends in Pharmacological Sciences*, vol. 41, pp. 947–959, Dec. 2020.

[24] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, and J. H. Jensen, "Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values," *Journal of Chemical Theory and Computation*, vol. 7, pp. 2284–2295, July 2011.

[25] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *Journal of Chemical Theory and Computation*, vol. 11, pp. 3696–3713, Aug. 2015.

[26] S. Izadi and A. V. Onufriev, "Accuracy limit of rigid 3-point water models," *The Journal of Chemical Physics*, vol. 145, p. 074501, Aug. 2016.

[27] T. E. Cheatham III and S. Joung, "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations — The Journal of Physical Chemistry B,"

[28] E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, and T. E. Ferrin, "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein Science: A Publication of the Protein Society*, vol. 30, pp. 70–82, Jan. 2021.

[29] D. Bang, V. Tereshko, A. A. Kossiakoff, and S. B. H. Kent, "Role of a salt bridge in the model protein crambin explored by chemical protein synthesis: X-ray structure of a unique protein analogue, [V15A]crambin-α-carboxamide," *Molecular BioSystems*, vol. 5, pp. 750–756, June 2009.

[30] S. Bowerman and J. Wereszczynski, "Detecting Allosteric Networks Using Molecular Dynamics Simulation," *Methods in Enzymology*, vol. 578, pp. 429–447, 2016.

[31] B. Eisenberg, "Asking biological questions of physical systems: The device approach to emergent properties," *Journal of Molecular Liquids*, vol. 270, pp. 212–217, Nov. 2018.

[32] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, "Classification of Intrinsically Disordered Regions and Proteins," *Chemical Reviews*, vol. 114, pp. 6589–6631, July 2014.

[33] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.

[34] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, J. Domański, D. L. Dotson, S. Buchoux, I. M. Kenney, and O. Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," *Proceedings of the 15th Python in Science Conference*, pp. 98–105, 2016.

[35] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, and D. M. Zuckerman, "Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]," *Living Journal of Computational Molecular Science*, vol. 1, no. 1, pp. 5067–5067, 2019.

[36] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, p. 20150202, Apr. 2016.

[37] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, June 2002.

[38] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11–15, 2008.

[39] T. Ichiye and M. Karplus, "Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations," *Proteins*, vol. 11, no. 3, pp. 205–217, 1991.

[40] E. Saccenti, M. H. W. B. Hendriks, and A. K. Smilde, "Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models," *Scientific Reports*, vol. 10, p. 438, Jan. 2020.

[41] H. Kamberaj and A. van der Vaart, "Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations," *Biophysical Journal*, vol. 97, pp. 1747–1755, Sept. 2009.

[42] W. J. Krzanowski, "Sensitivity of Principal Components," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 3, pp. 558–563, 1984.