THE FLEXIBILITY-RIGIDITY INDEX (FRI): THEORY AND APPLICATIONS

By

Kristopher Opron

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology - Doctor of Philosophy

2016

ABSTRACT

THE FLEXIBILITY-RIGIDITY INDEX (FRI): THEORY AND APPLICATIONS

By

Kristopher Opron

Since the first protein structures were solved in the 1950s, the protein data bank has grown to include over one hundred thousand macromolecular structures ranging in size from small peptides to large viral capsids. These experiments have shown that proteins exhibit a diverse range of structure and function and that these two aspects are closely related. In fact, it is often possible to predict a protein's function from its structure alone. Much of the focus to date has been on the more static regions of proteins for theoretical and practical reasons. However, it is important to note that even well folded proteins experience everlasting fluctuations due to the constant influence from outside forces, which drive motions that are relevant to function such as sidechain fluctuations and conformational shifts. The possible movements that can arise from these fluctuations are determined by a protein's structure. This means flexibility, or the ability to deform from the current conformation under external forces, is an intrinsic property of all proteins, and is closely tied to function. In order to better study protein function in ordered or disordered proteins, we require accurate, efficient, multiscale tools for evaluating flexibility.

This work puts forward a multiscale, multiphysics and multidomain model, the flexibility-rigidity index (FRI), to estimate the flexibility and conformational motions of macromolecular structures. The basic assumption of the present FRI theory is that the geometry or structure of a given protein, together with its specific environment, completely determines the biological function and properties including flexibility and charge. To this end, we utilize monotonically decreasing functions to measure the geometric compactness of a protein and quantify the topological connectivity of atoms or residues in the proteins and nucleic

acids. We define the total rigidity of a molecule by a summation of atomic rigidities. A practical validation of the proposed FRI for flexibility analysis is provided by the prediction of B-factors, or temperature factors of proteins, measured by X-ray crystallography. We employ a test set of 263 structurally distinct proteins to examine the validity and robustness of the proposed FRI method for B-factor estimation or flexibility prediction. The basic FRI algorithm outperforms GNM on this test set by about 20%.

After validation of the basic FRI method we introduce a multikernel-based multiscale FRI (mFRI) strategy to analyze macromolecular flexibility. The essential idea is to employ two or three kernels each parameterized with a different scale to capture the multiple characteristic interaction scales of complex biomolecules. Based on an expanded test set containing 364 proteins, we show that the mFRI method is about 22% more accurate than the GNM method in B-factor prediction. Most importantly, we demonstrate that the present mFRI gives rise to excellent flexibility analysis for many proteins that are difficult cases for GNM and the previously introduced single-scale FRI methods. Finally, for a protein of N residues, we illustrate that the computational complexity of the proposed mFRI is of linear scaling $\mathcal{O}(N)$, in contrast to the order of $\mathcal{O}(N^3)$ for GNM.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

KEY TO ABBREVIATIONS

$aFRI$ - Anisotropic Flexibility-Rigidity Index

ANM - Anisotropic Network Model

$fFRI$ - Fast Flexibility-Rigidity Index

FRI - Flexibility-Rigidity Index

$gANM$ - Generalized Anisotropic Network Model

$gGNM$ - Generalized Gaussian Network Model

GNM - Gaussian Network Model

$mANM$ - Multiscale Anisotropic Network Model

MD - Molecular Dynamics

$mFRI$ - Multiscale Flexibility-Rigidity Index

$mGNM$ - Multiscale Gaussian Network Model

NMA - Normal Modes Analysis

NMR - Nuclear Magnetic Resonance Spectroscopy

$pfFRI$ - Parameter-free Flexibility-Rigidity Index

SVM - Support Vector Machine

CHAPTER I. Summary

Recent technological and methodological advances have dramatically increased the size of the macromolecular structures that can be solved experimentally. This increase in scale has led to challenges in the theoretical description and computer simulation of proteins and nucleic acids. In response to this boom in structure size, there has been an increased interest in multiscale, multiphysics, and/or multidomain models. Such models aim to improve analysis of large macromolecules by reducing the number of degrees of freedom while maintaining modeling accuracy and achieving computational efficiency. To this end, the following work introduces a simple, accurate and efficient multiscale method for analyzing macromolecular flexibility and rigidity in atomic detail, the Flexibility-Rigidity Index or FRI.

The FRI theory is based on the assumption that the most fundamental properties of macromolecules are almost entirely determined by the geometric structure of the protein rather than its sequence, even though the structure is determined primarily by its sequence of amino acids. Simply put, FRI methods use the geometric compactness of a macromolecular structure to determine flexibility and motion at the atomic scale. Unlike the similar and well-known methods based on Normal Modes Analysis, FRI does not require matrix diagonalization for flexibility predictions. In the case of anisotropic calculations, FRI does require some matrix solving, but the FRI method allows for fewer and/or smaller matrices to be solved, thereby cutting down on calculation times drastically. The basic FRI algorithm's computational complexity is approximately $\mathcal{O}(N^2)$, where N is the number of atoms or residues, in contrast to $\mathcal{O}(N^3)$ for methods such as Normal Modes Analysis (NMA) and Gasussian Network Model (GNM) that require solving of a large matrix. In our initial studies, we demonstrate that the proposed FRI gives rise to accurate predictions of protein B-Factors for a set of 263 protein structures taken from X-ray crystallography data in the

Protein Data Bank. We also show that a parameter-free formulation of FRI (pfFRI) is able to achieve about 95% accuracy of the regular FRI algorithm. Furthermore, we compare the accuracy and efficiency of FRI to that of the most popular approaches for flexibility analysis, NMA and GNM. An interpolation algorithm is also introduced in the first work and is used to construct continuous atomic flexibility functions for visualization and use in multiscale multiphysics models.

Beyond the introduction of the basic FRI method, this work introduces various improvements and variations on the FRI method that improve computational efficiency, increase accuracy or add new utility.

The first improvement introduced is the fast FRI (fFRI) algorithm for improving the computational run-time for flexibility analysis. The proposed fFRI further reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ through the implementation of the cell lists method. Intensive validation and comparisons indicate that fFRI is orders of magnitude more efficient and about 10% more accurate overall than some of the most popular methods in the field. The proposed fFRI is able to predict B-factors for α-carbons of the HIV virus capsid (313,236 residues) in less than 30 seconds on a single processor using only one core.

The next major addition we propose is the anisotropic FRI (aFRI) algorithm for the analysis of collective protein dynamics. The aFRI algorithm makes use of adaptive Hessian matrices, ranging from a completely global $3N \times 3N$ matrix to completely local 3×3 matrices. These local 3×3 matrices only describe the motion of a single residue, however, these matrices include global correlation effects, giving rise to predictions that qualitatively match global calculations. The use of adaptive matrices allows for a significant decrease in computational running time due to the advantage of solving many small matrices rather than one large matrix. Eigenvectors obtained through the proposed aFRI algorithms are able to demonstrate collective motions similar to normal modes methods. A large set of proteins is

used to compare the efficiency of the FRI, fFRI, aFRI and GNM methods.

The next step in the development of FRI was the addition of a multiscale concept to the FRI method. Because protein interactions are inherently multiscale and protein flexibility is associated with protein interactions, protein flexibility should have a multiscale characteristic. Existing elastic network models are typically parameterized at a single cut-off distance and therefore may fail to properly predict the thermal fluctuation of the many macromolecules involving multiple characteristic length scales. Therefore we introduce a multiscale flexibility-rigidity index (mFRI) method to resolve this problem. The proposed mFRI utilizes two or three correlation kernels parametrized at different length scales to capture various levels of interactions within and between proteins. It is shown that the mFRI method is about 20% more accurate than the Gaussian Network Model in the B-factor prediction for a set of 364 protein structures. Additionally, we identify multiple instances where mFRI is accurate and GNM is very inaccurate, possibly due to the lack of a multiscale aspect.

In addition to testing on proteins, we test the FRI method for macromolecular complexes that include nucleic acids. Protein-nucleic acid complexes are important for many cellular processes including some of the most essential functions such as transcription and translation. For many protein-nucleic acid complexes, flexibility of both macromolecules is known to be critical for specificity and/or function. Therefore, we have extended FRI flexibility analysis to protein-nucleic acid complexes. We demonstrate by comparison with experimental data that the mFRI multiscale strategy is able to accurately predict the flexibility of protein-nucleic acid complexes. Also, we take advantage of the high accuracy and $\mathcal{O}(N)$ computational complexity of our multiscale FRI method to investigate the flexibility of large ribosomal subunits and an entire ribosome, which is difficult to analyze by alternative approaches due to its size. As a final demonstration of the FRI method for protein-nucleic acid complexes,

we utilize an anisotropic FRI approach, which involves localized Hessian matrices, to study the translocation and active site dynamics of bacterial RNA polymerase.

As GNM and ANM are some of the most popular methods for the study of protein flexibility and related functions, and the FRI method resembles these methods in some ways, it is necessary to clarify the relationship between normal modes-based methods and FRI. To this end, we propose generalized GNM and ANM methods (gGNM and gANM) and show that the GNM Kirchhoff matrix can be built from the ideal low-pass filter, a special case of correlation functions underpinning the FRI method. We propose a unified framework to construct generalized Kirchhoff matrices whose matrix inverse leads to gGNMs, whereas, the direct inverse of its diagonal elements gives rise to FRI method. In addition to exploring this connection, we introduce two new multiscale elastic network models, namely multiscale GNM (mGNM) and multiscale ANM (mANM), which are able to incorporate different scales into generalized Kirchhoff or Hessian matrices.

We illustrate that gGNMs outperform the original GNM method in the B-factor prediction of the set of 364 proteins. We demonstrate that for a given correlation function, FRI and gGNM methods provide essentially identical B-factor predictions when the scale value in the correlation function is sufficiently large. The multiscale aspect of the proposed mGNM and mANM gives rise to a significant improvement, more than 11%, in B-factor predictions over the original GNM and ANM methods. We further demonstrate benefit of our mGNM method in the B-factor predictions of many proteins that the original GNM method fails to accurately predict B-factors for. Also, we show that the present mGNM can be used to analyze protein domain separations and showcase the ability of our mANM for the simulation of protein collective motions.

As an exploration into one of the many applications of FRI, we examined the potential for FRI and aFRI methods in predicting protein hinges. The study of hinges and hinge

motions in proteins has been an important topic and much research has been done in the past.[21,25,26,39,59] Identification of hinge residues is useful for inferring motion and function when molecules are too large for MD simulation on relevant timescales. Other methods such as GNM and NMA have been utilized for this purpose in the past, leading us to the idea that FRI-based methods could place a significant role in hinge analysis. So far we have tried predicting hinge using modes of motion calculated from FRI correlation maps. We have also tried various machine-learning models to predict hinges using a combination of FRI-based metrics and various other residue-level metrics based on solvent accessible surface area, side-chain, hydrophobicity and many other properties. Finally, we show that hinge predictions from FRI modes are at least as accurate as those obtained from other state of the art hinge prediction software.

CHAPTER II. Background and Introduction

2.1 Experimental methods for structural flexibility

The field of structural biology has seen rapid growth in the last few decades. Since the first protein structures were solved in the late 1950s, the protein data bank has grown to include over one hundred thousand macromolecular structures ranging in size from small peptides to large viral capsids. These experiments have shown that proteins exhibit a diverse range of structure and function and that these two aspects are closely related. In fact, it is often possible to predict a protein's function from its structure alone, especially when a homologous protein is available for comparison. Much of the focus to date has been on the more static regions of proteins for theoretical and practical reasons. However, it is important to note that even well folded proteins experience everlasting fluctuations due to the constant influence from outside forces, which drive intrinsic motions such as atomic vibrations and conformational shifts. The possible movements that can arise from these fluctuations are determined by a protein's structure. This means flexibility, or the ability to deform from the current conformation under external forces, is an intrinsic property of all proteins, and is closely tied to function. For instance, protein flexibility can enhance protein-protein and protein-ligand interactions by intermittently offering more favorable binding surfaces through small secondary structure and sidechain fluctuations. Additionally, protein flexibility and motion amplify the probability of barrier crossing in enzymatic reactions. Therefore, the investigation of protein flexibility at multiple scales is vital to the understanding and prediction of protein functions. In fact, even the study of some completely disordered proteins is essential due to their connections to neurodegenerative diseases such as mad cow disease, Alzheimer's disease and Parkinson's disease.[15,67] Therefore, in order to better study protein function in ordered or disordered proteins, we require accurate, efficient, multiscale tools for evaluating flexibility.

Currently, the most important experimental techniques for protein flexibility analysis are X-ray crystallography and Nuclear Magnetic Resonance (NMR). Among the over one hundred thousand structures in the protein data bank (PDB), more than eighty percent were collected by X-ray crystallography. The Debye-Waller factor, or B-factor, is a experimental measure of disorder that can be directly computed from X-ray diffraction data. These B-factors have been observed to correlate with atomic flexibility from MD and NMA experiments, thereby making them an ideal experimental measure of flexibility for comparison with theoretical methods. However, it is important to remember that this is not a perfect correlation because B-factors can be influenced by multiple factors including variations in atomic diffraction cross sections and chemical stability during the diffraction data collection. Therefore, only the B-factors for specific types of atoms, most often C_α atoms, can be directly interpreted as their relative flexibility without corrections. The other major experimental method for accessing protein flexibility is NMR, which often provides structural flexibility information under physiological conditions unlike X-ray diffraction, which requires specific conditions to form suitable crystals. NMR spectroscopy allows the characterization of protein flexibility in diverse spatial dimensions and a large range of time scales. About seven percent of the structures in the PDB are determined by NMR spectroscopy, however, it is unclear how to assign flexibility values to atoms based on NMR spectroscopy data. Therefore we are currently focused on comparing theoretical results to X-ray crystallography results only.

2.2 Computational methods for flexibility and dynamics

The experimental techniques mentioned in the previous section are incredibly powerful for studying protein structure and function, however, they do face some limitations due to technical challenges. For example, some proteins may be extremely difficult or impossible to crystallize and others that do crystallize may do so in forms not relevant to their function.

To address the cases where experimental techniques fail and to increase efficiency of analysis we turn to theoretical approaches. There have been many distinct methods for flexibility and motion analysis proposed over the past few decades. The major examples, and those that can be in some way compared to FRI, are molecular dynamics (MD), NMA, machine-learning models, and multiscale, multiphysics simulations.

Molecular dynamics simulations are at the forefront of computational biochemistry and have contributed significantly to our understanding of the conformational landscapes of proteins, especially conformations that are not directly accessible via experimental techniques due to various technical or practical challenges. These simulations enable us to study proteins that are difficult to study experimentally such as amyloid fibrils, intrinsically disordered proteins, and partially disordered proteins. However, the dynamics of larger macromolecules and systems including multiple molecules typically occur at time scales that are intractable for MD simulations.

A major breakthrough with respect to the scale of protein simulations came with the introduction of normal mode analysis (NMA),[8,29,44,64] a time-independent molecular mechanics method that is related to MD via the time-harmonic approximation.[52] The success of the initial NMA method led to the development of related methods that improve computational running time or add utility. The most notable examples of NMA-related methods are the elastic network model (ENM),[66] GNM,[4,5,27] and anisotropic network model (ANM).[3] The ENM and GNM methods use coarse-grained representations of macromolecules to speed up computation with only a minor loss in accuracy, and the ANM method provides motion predictions. All normal modes-related methods can be used to approximate protein flexibility or B-factors from the first few eigenvectors and eigenvalues of the interaction Hamiltonian in normal modes or the Kirchhoff matrix in ENM and GNM. These quantitative predictions of biomolecular flexibility and their applications are discussed in many review papers.[16,46,61,77]

The lowest-energy eigenvalues from these calculations reflect the protein dynamics through even the longest relevant timescales, something which is typically beyond the reach of MD simulations.[5,8,44,64,66] The normal modes approaches have been improved in many aspects since their introduction, including crystal periodicity corrections[32,40,41,62] and the introduction of the density-cluster rotational-translational blocking[18] to speed up calculations. Still, the computational complexity of these methods is dominated by a computationally inefficient diagonalization of the large N by N matrices used in normal modes. Due to the diagonalization step, these methods have running times that scale according to $\mathcal{O}(N^k)$ time, where N is the matrix dimension and $k \approx 3$. So while normal modes calculations are typically more efficient for calculating long-time dynamics of proteins than MD, the method is not suitable for excessively large macromolecules and macromolecular complexes, e.g. systems with millions of amino acid residues such as those obtained from cryo-EM experiments or theoretical constructs.

An even more recent set of tools for flexibility analysis is that of knowledge based, machine-learning methods. This category includes examples of flexibility prediction by neural networks,[55] support vector regression[81] and two-stage support vector regression.[51] These approaches typically utilize large protein data sets as training data. Therefore, the validity and accuracy and of these methods are dependent on the quality and representativeness of the training data set, qualities that are difficult to prove in the current state of structural biology. Yet another modern approach utilizes graph theory to analyze the bond networks in proteins,[36] employing both geometric and energetic criteria to identify the flexible and rigid regions. Unfortunately, this method relies on normal mode analysis and other costly algorithms which limit it to the same scales as the normal mode tools.

Most recently there has been an increased interest in theoretical methods for flexibility analysis that are developed via multiscale formulations. Multiscale methods combine elastic

mechanics and molecular mechanics to significantly reduce the degrees of freedom of large biomolecular systems.[10] For example, the classical theory of elasticity for DNA loops has been combined with the MD description of protein for protein-DNA interaction complexes.[70] Recently, the continuum elastic modeling of the Canham-Helfrich type of energy functional has been coupled with MD simulations to investigate the complex elastic behavior of Hepatitis B virus capsids.[57] Multiscale based flexibility analysis has a wide range of technical variability. In the best scenario, multiscale methods can take the advantage of each scale to achieve excellent modeling accuracy and computational efficiency. However, multiscale methods are typically technically demanding and computationally complex. A major issue in the field is how to go beyond the phonological domain and make these approaches quantitative and predictive. Reliable analysis and validation with experimental data are indispensable procedures. For these reasons, there is a need to further develop and validate innovative approaches for the flexibility analysis of biomolecular systems.

2.3 The Flexibility-Rigidity Index

This work aims to solve the a major issue with the aforementioned methods, especially MD and NMA, which is the issue of poor scaling. In addition to providing improved scaling for predicting flexibility and long-time scale dynamics it is vital to also match or improve upon the level of accuracy and utility offered by these other methods. To address these issues propose an efficient, accurate method for protein B-factor prediction and flexibility analysis called the Flexibility-Rigidity Index or FRI. The FRI method is based on some simplifying assumptions about macromolecules including that protein dynamics are determined entirely by structure and that side chain effects can be ignored. The FRI algorithm is based on measurements of geometric compactness and topological connectivity of a protein structure at each residue. It is assumed that nearby atoms have stronger interactions and tend to confer stability to other nearby atoms in macromolecules and that this stabilizing effect decays

with increasing distance. Physical interaction potentials are not directly used to represent the interactions in this method and are instead replaced by a monotonically decaying kernel or kernels that is parametrized empirically. In practice, this method gives rise to accurate predictions of protein flexibility or B-factors based on geometric compactness alone.

We noted after the publication of our earliest work on FRI[75] that the name of "flexibility index" was proposed independently by von der Lieth et al.[71] and Jacobs et al.[36] for two different quantities to describe bond strengths. Both of these flexibility indices are distinct from the proposed FRI method. The FRI algorithm is solely structural based and it does not reconstruct any protein interaction Hamiltonian. Only elementary arithmetic is needed in the FRI method for proteins. In particular, the FRI prediction of protein B-factors does not require a stringently minimized structure or a time consuming matrix diagonalization or matrix decomposition step, nor does it involve any training procedure.

2.3.1 fast FRI and anisotropic FRI

Another objective of the present work is to introduce a fast FRI (fFRI) algorithm by using appropriate data structures because computational efficiency is critical for analyzing larger structures. The computational complexity of the proposed fFRI is of $\mathcal{O}(N)$, compared to that of $\mathcal{O}(N^2)$ for the original FRI algorithm and of $\mathcal{O}(N^3)$ for the GNM, where N is the number of atoms. We use the cell lists approach[2] to achieve this reduction in the computational complexity with negligible loss in accuracy.. Another objective is to introduce anisotropic FRI (aFRI) algorithms for the motion analysis of biomolecules. Unlike ANM,[3,52] which is completely global and has $3N \times 3N$ elements in its Hessian matrix, the proposed aFRI algorithms utilize adaptive Hessian matrices, which vary from completely global to completely local. Even in the most local formulation of aFRI, there are collective motions predicted by three sets of eigenvectors. These three modes of motion turn out to correspond with the lowest energy, most dominant modes of global aFRI and ANM in most test cases.

We demonstrate this utility of aFRI on multiple proteins and protein-nucleic acid system.

It was noticed early in the development of FRI that there are a small number of structures for which FRI performs very poorly in flexibility prediction. Furthermore, those structures which cause problems for FRI are likely to be difficult for NMA and GNM as well. One such structure is pictured in Figure 2.1: where the GNM method fails to predict the high flexibility of a hinge region in calmodulin. There are a number of possible reasons for this and similar failures, which are highlighted in this work. The crystal environment, solvent type, co-factors, data collection conditions, and structural refinement procedures are all well-known effects[32,40,41,62] that can interfere with flexibility estimations from X-ray experiments. However, there is one more important cause that has not been discussed in the literature to our best knowledge, namely, multiple characteristic length scales in a single protein structure. Contrary to very small molecules, macromolecules have a wide variety of characteristic length scales, from the small scale of intramolecular bonding to large scale effects observed in protein-nucleic and protein-protein interactions. Therefore, it is reasonable to regard large proteins as multiscale molecules. When a GNM or FRI algorithm is parametrized at a single given cutoff or scale parameter, it captures only a subset of the characteristic length scales and inevitably misses other characteristic length scales of the protein. Consequently, neither method is able to provide accurate B-factor predictions for all macromolecules using a single characteristic length scale.

Therefore one of the objectives of the present work is to introduce a multiscale strategy for protein flexibility analysis. The essential idea is to assess protein topological connectivity and packing compactness at multiple scales by combining multiple FRI kernels or correlation functions. As a result, multiscale FRI (mFRI), is able to simultaneously capture protein crucial characteristic length scales and provide improved B-factor predictions.

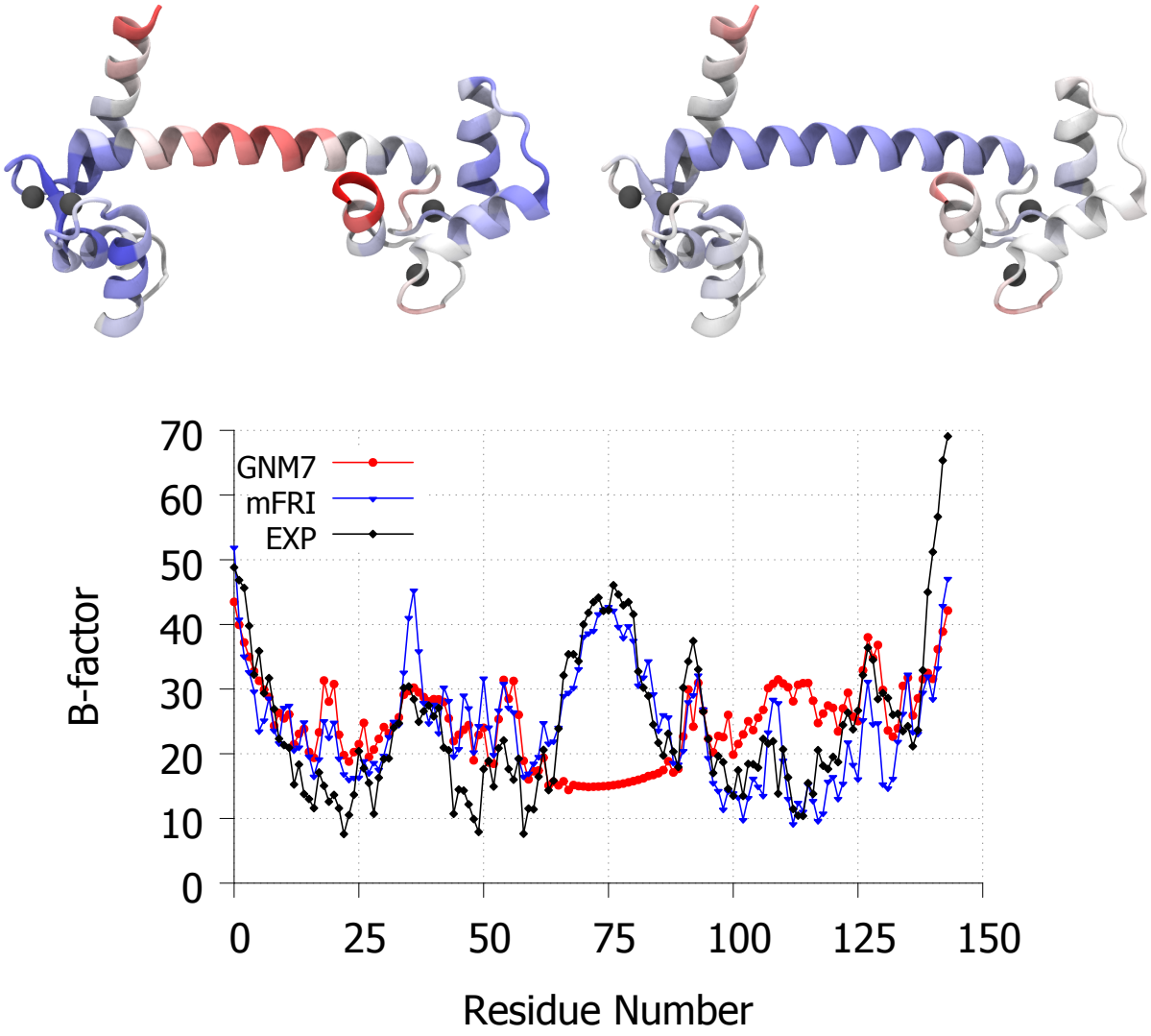


Figure 2.1: The structure of calmodulin (PDB ID: 1CLL) visualized in VMD[34] and colored by experimental B-factors (top left) and GNM predicted B-factors (top right) with red representing the most flexible regions. Bottom, a comparison of predicted B-factor values from mFRI, GNM with a cutoff distance of 7Å , and experimental B-factors taken from the PDB entry.

2.3.2 FRI for Protein-Nucleic Acid Complexes

In addition to proteins, nucleic acids are among the most essential biomolecules for all known forms of life. Nucleic acids often function in association with proteins and play a crucial role in encoding, transmitting and expressing genetic information. Therefore it was necessary to develop FRI methods for nucleic acid chains and protein-nucleic acid complexes. Proteins and nucleic acid chains are dramatically different biomolecules and amino acid residues and nucleotides have different length scales and interaction characteristics. Therefore, a good model should not only allow residues and/or nucleotides to be treated with different length scales, but also adapt a multiscale description of each residue and/or nucleotide. Unlike elastic network models that are parametrized in only one length scale for each particle, mFRI provides a simultaneous multiscale description. Therefore, the present mFRI is able to better capture multiscale collective interactions of protein-nucleic acid complexes. Additionally, many protein-nucleic acid complexes are very large biomolecules and therefore require considerable computational resources to analyze by conventional mode decomposition-based methods. The $\mathcal{O}(N)$ scaling FRI methods provide a more efficient approach to the flexibility analysis of large protein-nucleic acid complexes.

2.3.3 gGNM, mGNM and mANM

Inspired by the improvements that multiple correlation kernels have on the FRI method, we propose a method to incorporate multiscale correlations into the aforementioned mode decomposition-based methods, GNM and ANM. Our approach to address the link between FRI and normal modes methods is twofold. First, we propose a unified framework to construct a generalized GNM (gGNM). We reveal that the GNM Kirchhoff matrix can be constructed from the ideal low-pass filter (ILF), which is the limiting case of admissible FRI correlation functions. We demonstrate that FRI and gGNM are asymptotically equivalent

when the cutoff value in the Kirchhoff matrix or the scale value in the correlation function is sufficiently large. This finding paves the way for understanding the connection between the GNM and FRI methods. To clarify this connection, we introduce a generalized Kirchhoff matrix to provide a unified starting point for the gGNM and FRI methods, which elucidates on the similarity and difference between gGNM and FRI. Based on this new understanding of the gGNM working principle, we propose infinitely many correlation function-based gGNMs. We show that gGNM outperforms the original GNM for the B-factor prediction of a set of 364 proteins. Both gGNM and FRI deliver almost identical results when the scale parameter is sufficiently large. This approach sheds light on the construction of efficient gGNMs.

Additionally, we propose two new methods, multiscale GNM (mGNM) and multiscale ANM (mANM), to account for the multiscale features of biomolecules. The aim is to generalize original GNM and ANM into a multikernel setting so that each kernel can be parametrized at a given characteristic length. This generalization is achieved through the use of a FRI assessment, which predicts the involvement of different scales, followed by an appropriate construction of multikernel GNM or multikernel ANM. This approach works because for a diagonally dominant matrix, the direct inverse of the diagonal element is essentially equivalent to the diagonal element of the inverse matrix. In this work we demonstrate by comparison with experimental data that the proposed mGNM and mANM are able to successfully capture the multiscale properties of protein structures and significantly improve the accuracy of these methods in protein flexibility prediction.

2.3.4 Machine learning and FRI for protein-protein interactions

Support Vector Machine (SVM) is a type of supervised machine learning that has grown in popularity recently due to successful applications across many different fields. Some examples of successful applications of SVM models include drug design,[37] image recognition and text classification,[12,20,38] microarray gene expression data analysis,[9,11,28,30,48,53,80] pro-

tein fold recognition[14,19,43] , protein-protein interaction[7] and protein secondary structure prediction.[33] The basic idea of applying an SVM in this context is to map a set of input into the feature space in which the input data becomes more separable compared to the original input, then construct a maximum-margin hyperplane which separates two classes within the feature space. In this case the two classes being separated are hotspot residues and non-hotspot residues. To test the viability of FRI-derived features for predicting protein-protein interactions, we chose to incorporate FRI-derived metrics in to the KFC2 model,[83] an SVM model that uses residue-scale features to predict protein-protein binding hotspots.

CHAPTER III. Methods

3.1 Flexibility-rigidity index (FRI)

We initially consider only proteins as examples to illustrate the FRI algorithm, although other biomolecules, such as DNA and RNA, can be accommodated with a minor modification of the algorithm. We are particularly interested in a coarse-grained representation. However, methods for a full atom description can be formulated as well.

We seek a structure based algorithm to convert protein geometry into protein topology. To this end, we consider a protein with N C_α atoms. Their locations are represented by $\{\mathbf{r}_j | \mathbf{r}_j \in \mathbb{R}^3, j = 1, 2, \cdots, N\}$. We denote $\|\mathbf{r}_i - \mathbf{r}_j\|$ the Euclidean space distance between ith C_α atom and the jth C_α atom. The distance geometry of protein C_α atoms is utilized to establish the topology connectivity by using monotonically decreasing radial basis functions,

$$C_{ij} = \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}), \tag{3.1}$$

where η_{ij} is a characteristic distance between particles, and $\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij})$ is a correlation function, which is, in general, a real-valued monotonically decreasing function. As a correlation function, it satisfies

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_i\|; \eta_{ii}) \;=\; 1 \tag{3.2}$$

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) \;=\; 0 \quad \text{as} \quad \|\mathbf{r}_i - \mathbf{r}_j\| \to \infty. \tag{3.3}$$

Delta sequences of the positive type discussed in an earlier work[73] are all good choices. For example, one can use generalized exponential functions

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = e^{-(\|\mathbf{r}_i - \mathbf{r}_j\|/\eta_{ij})^\kappa}, \quad \kappa > 0 \tag{3.4}$$

and generalized Lorentz functions

$$\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = \frac{1}{1 + (\|\mathbf{r}_i - \mathbf{r}_j\|/\eta_{ij})^\upsilon}, \quad \upsilon > 0. \tag{3.5}$$

Essentially, the correlation between any two particles should decay according to their distance. Therefore, many other alternatives can be used and so during validation multiple functions are tested.

The correlation map or cross correlation is an important quantity for the GNM. We can define a similar correlation map by setting $\mathbf{C} = \{C_{ij}\}, i, j = 1, 2, \cdots, N$. The correlation map measures the connectivity of C_αs in the protein.

We define an atomic rigidity index μ_i as the summation of topological connectivity

$$\mu_i = \sum_{j=1}^{N} w_{ij} \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}), \quad \forall i = 1, 2, \cdots, N, \tag{3.6}$$

where w_{ij} is a weight function related to the atomic type, The atomic rigidity index μ_i manifests the rigidity or stiffness at the ith atom. In a general sense, the atomic rigidity index reflects the total interaction strength, including both bonded and non-bonded contributions. It is quite straightforward to define the averaged molecular rigidity index as a summation of atomic rigidity indices

$$\bar{\mu}_{\mathrm{MRI}} = \frac{1}{N} \sum_{i=1}^{N} \mu_i. \tag{3.7}$$

The averaged molecular rigidity index can be used to predict molecular thermal stability, bulk modulus, density (compactness), boiling points of isomers, the ratio of surface area over volume, surface tension, etc. A detailed investigation of these aspects is beyond the scope of the present work.

We are now ready to define a position dependent shear modulus

$$\mu(\mathbf{r}) \quad = \quad \sum_{j=1}^{N} w_j(\mathbf{r}) \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_{ij}), \quad \mathbf{r} \in \Omega_E, \tag{3.8}$$

where $w_j(\mathbf{r})$ is a weight function, $\mathbf{r}$ is in the proximity of $\mathbf{r}_i$ and Ω_E is the macromolecular domain. In order to determine $w_j(\mathbf{r})$, we define an average rigidity (or averaged rigidity

index function) by

$$\bar{\mu} = \frac{1}{V}\int \mu(\mathbf{r})d\mathbf{r}, \tag{3.9}$$

where V is the volume of the macromolecule. If $w_j(\mathbf{r})$ is a constant, its value can be uniquely determined by a comparison of $\bar{\mu}$ with experimental shear modulus[58] for a given macromolecule and correlation function.

We also define an atomic flexibility index as

$$f_i = \frac{1}{\mu_i}, \quad \forall i = 1, 2, \cdots, N. \tag{3.10}$$

Since the flexibility at each atom is proportional to its temperature fluctuation, we can express B-factors as

$$B_i^t = af_i + b, \quad \forall i = 1, 2, \cdots, N \tag{3.11}$$

where $\{B_i^t\}$ are theoretically predicted B-factors, and a and b are two constants to be determined by a simple linear regression.

We can also define the averaged molecular flexibility index (MFI) as a summation of atomic flexibility indices

$$\bar{f}_{\text{MFI}} = \frac{1}{N}\sum_{i=1}^{N} f_i. \tag{3.12}$$

MFI should correlate with molecular stability and energy.

For the purpose of visualization, we define a continuous atomic flexibility function as

$$F(\mathbf{r}) = \sum_{j=1}^{N} B_i^t \Psi(\|\mathbf{r} - \mathbf{r}_j\|), \quad \mathbf{r} \in \Omega_E. \tag{3.13}$$

where $\Psi(\|\mathbf{r} - \mathbf{r}_j\|)$ is a general interpolation function for scattered data. Wavelets, spline functions, and modified Shepard's method[56,65] can be employed for the interpolation. One

can map $f(\mathbf{r})$ to the molecular surface to visualize the protein flexibility.[75] Alternatively, one can compute the continuous atomic flexibility function by

$$F(\mathbf{r}) = \frac{1}{\sum_{j=1}^{N} w_j(\mathbf{r})\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta_{ij})}, \quad \mathbf{r}\in\Omega_E. \tag{3.14}$$

Similarly, we can also construct a continuous multiscale flexibility function,

$$f(\mathbf{r}) = b + \sum_{n=1} \frac{a^n}{\sum_{j=1}^{N} w_j^n \Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta^n)}. \tag{3.15}$$

One can map this continuous multiscale flexibility function onto a molecular surface to analyze the flexibility of the molecule.

3.2 FRI correlation maps or matrices

Similar to the cross correlations of GNM and other methods, FRI correlation maps computed using Eq. (3.1) qualitatively reflect the three-dimensional structure of a protein. As a consequence, distinct secondary structures such as α helices and β-sheets exhibit characteristic patterns. After some studying of the patterns it is possible to approximate a proteins secondary and tertiary structure from the patterns of the correlation map alone. However, unlike the cross correlations of GNM, FRI correlation maps are able to offer more quantitative structural information. In fact, since the kernel used to generate the map is known, the distances between all atoms can be calculated and the three-dimensional structure can be reconstructed from the correlation map. Figure 3.1: displays four examples of correlation maps next to their corresponding three-dimensional structure. The scale-bars of the correlation maps include distance values to emphasize the preservation of the 3D structural information.

As stated previously, each secondary structure exhibits a distinct pattern in the correlation maps. The pattern for an α helix is shown in the first row of Fig. 3.1:. The α helix creates a band of high correlation extending about 4 amino acids in either direction from the diagonal. The correlation has a local maximum at the third neighbor residue, due

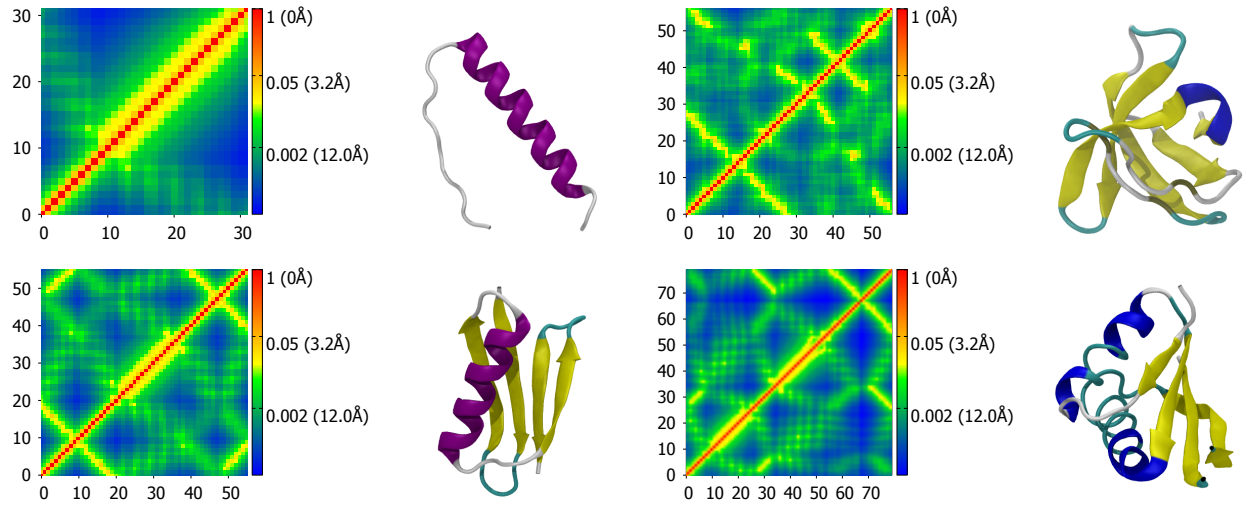


Figure 3.1: Correlation maps and secondary structure representations for four protein structures. Structures used include the alpha-spectrin SH3 domain, the tetramerization domain of the p53 tumor supressor, the B1 immunoglobulin-binding domain of streptococcal protein G and a DNA binding protein from Methanococcus jannaschii, from left to right, top to bottom. Correlation maps are generated using Eq. (3.5) with υ=2.5 and η=1.0Å . Secondary structure visualizations are generated with VMD.[34] Colors represent distance and correlation values for each pair of atoms. The residue numbers for each Cα are listed along the x- and y-axes. The protein are displayed in VMD's "new cartoon" representation and colored by secondary structure determined by STRIDE. The color scheme for secondary structure is: Purple - α helix, blue - 3(10) helix, yellow - β-sheet, cyan - turn, white - coil.

to the structure of the α helix (3.6 amino acid residues per turn). Therefore, the peak at the third residue serves as another signature of an α helix in the FRI correlation map. An increase in correlation between two such neighboring atoms compared to other neighboring pairs indicates the interaction of the α helix and another component. For example, in the third row of Fig. 3.1:, the correlation strength between 29th C_α and 32th C_α is higher, due to interaction of 29th C_α with the third and fourth beta sheets. This is an example of how this type of correlation kernels reflects tertiary structure information.

Other folds such as β-sheets are also easily identified by distinct patterns. One can easily distinguish parallel β-sheets from anti-parallel β-sheets by their patterns with this method. The second row of Fig. 3.1: is a good example of the pattern generated by anti-parallel β-sheets. Anti-parallel β-sheets appear as lines that are perpendicular to the diagonal of the map and the intersection of the two lines of high correlation are the turns between each β strand. Parallel β-sheets appear as lines parallel to the diagonal. In the third row of Fig. 3.1:, an anti-parallel β-sheet is formed by the first and last ten amino acids resulting in a line in the top left and bottom right of the correlation matrix.

The last two rows of Fig. 3.1: both display complex patterns which reflect not only secondary structure information but also the three dimensional arrangement of the secondary structure features. Clearly from the last correlation map, the first β-sheet interacts strongly with the first α helix and the second β-sheet in a parallel manner. It also interacts to a lesser degree with the second α helix and with the last β-sheet in an anti-parallel manner. These patterns and the stabilizing forces from the interactions they represent are lost if one uses a contact or Kirchoff matrix based method instead of a monotonically decreasing radial basis function based correlation map.

3.3 Fast flexibility-rigidity index (fFRI)

As discussed in our earlier work,[75] the original FRI algorithm has the computational complexity of $\mathcal{O}(N^2)$, mainly due to the construction of the correlation matrix. In the present work, we propose a fast FRI (fFRI) algorithm, which computes only the significant elements of the correlation matrix and at the same time maintains the accuracy of the method. As a result, the computational complexity of the fFRI algorithm is of $\mathcal{O}(N)$.

The essential idea is to partition the residues in a protein into cubic boxes according to their spatial locations. For each residue in a given box, we only compute its correlation matrix elements with all residues within the given box and with all residues in the adjacent 26 boxes. The accuracy and efficiency of this approach are determined by the box dimension. We select a box size of R such that

$$\Phi(R;\eta) \leq \varepsilon \tag{3.16}$$

where $\varepsilon > 0$ is a given truncation error. Therefore, for generalized exponential functions (3.4), we have

$$R \geq \eta \left(\ln \frac{1}{\varepsilon} \right)^{\frac{1}{\kappa}}. \tag{3.17}$$

If we set $\varepsilon = 10^{-2}$, we have $R \approx 4.6\eta$ for $\kappa = 1$ and $R \approx 2.15\eta$ for $\kappa = 2$. Note that different κ values have different optimal η values. The higher the κ value is, the larger the optimal η is.

Similarly, for generalized Lorentz functions (3.5), we choose the box size

$$R \geq \eta \left(\frac{1-\varepsilon}{\varepsilon} \right)^{\frac{1}{\upsilon}}. \tag{3.18}$$

Again, if we set $\varepsilon = 10^{-2}$, we have $R \approx 10\eta$ for $\upsilon = 2$ and $R \approx 4.6\eta$ for $\upsilon = 3$.

An optimal R should balance accuracy and efficiency. In Section 4.2, it is found that the selection of $R = 12$Å is near optimal for both exponential and Lorentz functions. In Algorithm , we present a pseudocode to illustrate the truncation algorithm of the fFRI.

Algorithm 1: fFRI algorithm

Input: $atoms(N)$ ▷ XYZ coordinates from PDB file

$mincoor \leftarrow minval(atoms)$ ▷ Compute dimensions of bounding box
$maxcoor \leftarrow maxval(atoms)$
$R \leftarrow boxsize$ ▷ Set size of grid
$Nbox \leftarrow ceiling((maxcoor - mincoor)/R)$ ▷ Compute number of boxes in each direction

for $ii \leftarrow 1, Natoms$ **do**
  $i, j, k \leftarrow ceiling((atoms(ii) - mincoor/R))$ ▷ Count the number of atoms in each box
  $Natoms(i, j, k) \leftarrow Natoms(i, j, k) + 1$
end for

for $k \leftarrow 1, Nbox[3]$ **do**
  for $j \leftarrow 1, Nbox[2]$ **do**
    for $i \leftarrow 1, Nbox[1]$ **do**
      $allocate(box(i, j, k))$ ▷ Allocate space for each box
    end for
  end for
end for

for $ii \leftarrow 1, Natoms$ **do** ▷ Copy coordinates to appropriate box based on 3D coordinates
  $i, j, k \leftarrow ceiling((atoms(ii) - mincoor)/R)$
  $box(i, j, k) \leftarrow atoms(ii)$
end for

for $k \leftarrow 1, Nbox[3]$ **do** ▷ Iterate over boxes
  for $j \leftarrow 1, Nbox[2]$ **do**
    for $i \leftarrow 1, Nbox[1]$ **do**

      for $n_a \leftarrow 1, Natoms(i, j, k)$ **do** ▷ Iterate over atoms in current box

        for $n \leftarrow k - 1, k + 1$ **do** ▷ Iterate over adjacent boxes
          for $m \leftarrow j - 1, j + 1$ **do**
            for $l \leftarrow i - 1, i + 1$ **do**

              for $n_b \leftarrow 1, Natoms(l, m, n)$ **do** ▷ Iterate over atoms in adjacent boxes

                $dist \leftarrow distance(box(i, j, k)(n_a), box(l, m, n)(n_b))$
                $FRI(n_a) \leftarrow kernel(dist)$

              end for

            end for
          end for
        end for

      end for

    end for
  end for
end for

3.4 Multiscale flexibility-rigidity index (mFRI)

The basic idea of multiscale FRI (mFRI) is quite simple. Since macromolecules are inherently multiscale in nature, we utilize multiple correlation kernels that are parameterized at multiple scales to characterize the multiscale flexibility of macromolecules

$$f_i^n = \frac{1}{\sum_{j=1}^N w_j^n \Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_j^n)}, \tag{3.19}$$

where w_j^n, $\Phi^n(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_j^n)$ and η_j^n are the corresponding quantities associated with the nth kernel. We seek the minimization of the form

$$\mathrm{Min}_{a^n,b} \left\{ \sum_i \left| \sum_n a^n f_i^n + b - B_i^e \right|^2 \right\} \tag{3.20}$$

where $\{B_i^e\}$ are the experimental B-factors. In principle, all parameters can be optimized. For simplicity and computational efficiency, we only determine $\{a^n\}$ and b in the above minimization process. For each kernel Φ^n, w_j^n and η_j^n will be selected according to the type of particles.

Specifically, for a simple C_α network, we can set $w_j^n = 1$ and choose a single kernel function parametrized at different scales. The predicted B-factors can be expressed as

$$B_i^{\mathrm{mFRI}} = b + \sum_{n=1} \frac{a^n}{\sum_{j=1}^N \Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta^n)}. \tag{3.21}$$

The difference between Eqs. (3.19) and (3.21) is that, in Eqs. (3.19), both the kernel and the scale can be changed for difefrent n. In contrast, in Eq. (3.21), only the scale is changed. One can use a given kernel, such as

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta^n) = \frac{1}{1 + (\|\mathbf{r} - \mathbf{r}_j\|/\eta^n)^3}, \tag{3.22}$$

to achieve good multiscale predictions.

3.5 Anisotropic flexibility-rigidity index (aFRI)

In this section, we propose a new anisotropic model based on the FRI method. In existing anisotropic methods, the Hessian matrix is always global so the matrix contains all

the $3N \times 3N$ elements for N particles in molecule. In the aFRI model, the Hessian matrix is inherently local and adaptive. Its size may vary from 3×3 for a completely local aFRI to $3N \times 3N$ for a complete global aFRI, depending on the need of a physical problem or computational resources. More local Hessian matrices are smaller and can be solved much faster due to the poor scaling of matrix solving algorithms.

To build the adaptive matrices of aFRI, partition all the N particles in a molecule into a total of M clusters $\{c_1, c_2, \cdots, c_k, \cdots, c_M\}$. Cluster c_k has N_k particles or atoms so that $N = \sum_{k=1}^{M} N_k$. A cluster may be a region of physical interest in a molecule such as an alpha helix, a domain, or a binding site of a protein. One of two extreme cases is that there is only one particle in each cluster. In that case there are N clusters. The other case is that there is only one cluster containing the entire molecule. The result is a Hessian matrix for any size cluster that can be solved individually and retains some information about other cluster properties in the values of the diagonal. For example, if we are interested in the thermal fluctuation of a particular cluster c_k with N_k particles or atoms, we can find $3N_k$ eigenvectors for the cluster. Let us keep in mind that each position vector in $\mathbb{R}^3$ has three components, $\mathbf{r} = (x, y, z)$. We denote

$$\Phi_{uv}^{ij} = \frac{\partial}{\partial u_i}\frac{\partial}{\partial v_j}\Phi(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}), \quad u, v = x, y, z; i, j = 1, 2, \cdots, N. \tag{3.23}$$

Note that for each given ij, we define $\Phi^{ij} = (\Phi_{uv}^{ij})$ as a local anisotropic matrix

$$\Phi^{ij} = \begin{pmatrix} \Phi_{xx}^{ij} & \Phi_{xy}^{ij} & \Phi_{xz}^{ij} \\ \Phi_{yx}^{ij} & \Phi_{yy}^{ij} & \Phi_{yz}^{ij} \\ \Phi_{zx}^{ij} & \Phi_{zy}^{ij} & \Phi_{zz}^{ij} \end{pmatrix}. \tag{3.24}$$

Since rigidity and flexibility can be both anisotropic, it is natural to propose two different aFRI algorithms based on a rigidity Hessian matrix and a flexibility Hessian matrix.

3.5.1 Anisotropic rigidity

Anisotropic rigidity is defined by a rigidity Hessian matrix for an arbitrary cluster c_k. Let us denote $(\mu_{uv}^{ij}(c_k))$ a rigidity Hessian matrix for cluster c_k. Its elements are chosen as

$$\mu_{uv}^{ij}(c_k) = -w_{ij}\Phi_{uv}^{ij}, \qquad i,j \in c_k; i \neq j; u,v = x,y,z \tag{3.25}$$

$$\mu_{uv}^{ii}(c_k) = \sum_{j=1}^{N} w_{ij}\Phi_{uv}^{ij}, \qquad i \in c_k; u,v = x,y,z \tag{3.26}$$

$$\mu_{uv}^{ij}(c_k) = 0, \qquad i,j \notin c_k; u,v = x,y,z. \tag{3.27}$$

The Hessian matrix $(\mu_{uv}^{ij}(c_k))$ is of $3N_k \times 3N_k$ dimensions. Note that the diagonal part, $\mu_{uv}^{ii}(c_k)$, has built in information from all the particles in the system, even if the cluster is completely localized, $N_k = 1,\ \forall k$.

A test of the anisotropic rigidity method is to check if it works for B-factor prediction. To predict B-factors with anisotropic rgidity we collect the diagonal terms of the rigidity Hessian matrix

$$\mu_{\text{diag}}^{i} = \text{Tr}\left(\mu_{uv}^{i}\right) \tag{3.28}$$

$$= \sum_{j=1}^{N} w_{ij}\left[\Phi_{xx}^{ij} + \Phi_{yy}^{ij} + \Phi_{zz}^{ij}\right]. \tag{3.29}$$

We then define a set of anisotropic rigidity (AR) based flexibility indices by

$$f_i^{\text{AR}} = \frac{1}{\mu_{\text{diag}}^{i}}. \tag{3.30}$$

B-factors can be predicted with a set of $\{f_i^{\text{AR}}\}$ by using the linear regression in Eq. (3.11).

3.5.2 Anisotropic flexibility

To analyze biomolecular anisotropic motions in parallel to ANM, we need to examine their anisotropic flexibility. To this end, we further define a flexibility Hessian matrix $\mathbf{F}(c_k)$

for cluster c_k as

$$\mathbf{F}^{ij}(c_k) = -\frac{1}{w_{ij}}(\Phi^{ij})^{-1}, \qquad i,j \in c_k; i \neq j; u,v = x,y,z \tag{3.31}$$

$$\mathbf{F}^{ii}(c_k) = \sum_{j=1}^{N} \frac{1}{w_{ij}}(\Phi^{ij})^{-1}, \qquad i \in c_k; u,v = x,y,z \tag{3.32}$$

$$\mathbf{F}^{ij}(c_k) = 0, \qquad i,j \notin c_k; u,v = x,y,z. \tag{3.33}$$

where $(\Phi^{ij})^{-1}$ denote the unscaled inverse of matrix Φ^{ij} such that $\Phi^{ij}(\Phi^{ij})^{-1} = |\Phi^{ij}|$. Similar to anisotropic rigidity, the diagonal part $\mathbf{F}^{ii}(c_k)$ has built-in information from all particles in the system. Therefore, even if the partition of clusters is completely localized (N clusters), correlation among atomic motions is retained. By diagonalizing $\mathbf{F}(c_k)$, we obtain $3N_k$ eigenvectors for the N_k particles in cluster c_k. Since the selection of c_k is arbitrary, eigenvectors of all other clusters can be attained using the same procedure.

To obtain the B-factor prediction from this anisotropic flexibility, we define a set of anisotropic flexibility (AF) based flexibility indices by

$$f_i^{\mathrm{AF}} = \mathrm{Tr}\left(\mathbf{F}(c_k)\right)^{ii}, \tag{3.34}$$

$$= (\mathbf{F}(c_k))_{xx}^{ii} + (\mathbf{F}(c_k))_{yy}^{ii} + (\mathbf{F}(c_k))_{zz}^{ii}. \tag{3.35}$$

Then Eq. (3.11) is employed to obtain B-factor predictions.

In this work, we only consider the coarse-grained model in which each residue is represented by its C_α. To further simply the model, the differences between residues are ignored. The parameter w_{ij} is assumed to be 1 and η_{ij} is set to a constant η.

3.6 Generalized Gaussian network models (gGNMs)

To establish notation and facilitate new development, let us present a brief review of the GNM and FRI methods. Consider an N-particle coarse-grained representation of a biomolecule. We denote $\{\mathbf{r}_i | \mathbf{r}_i \in \mathbb{R}^3, i = 1, 2, \cdots, N\}$ the coordinates of these particles and $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$ the Euclidean space distance between ith and jth particles. In a nutshell, the

GNM prediction of the ith B-factor of the biomolecule can be expressed as[4,5]

$$B_i^{\mathrm{GNM}} = a\left(\Gamma^{-1}\right)_{ii}, \forall i = 1, 2, \cdots, N, \tag{3.36}$$

where a is a fitting parameter that can be related to the thermal energy and $(\Gamma^{-1})_{ii}$ is the ith diagonal element of the matrix inverse of the Kirchhoff matrix,

$$\Gamma_{ij} = \begin{cases} -1, & i \neq j \text{ and } r_{ij} \leq r_c \\ 0, & i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{j,j\neq i}^{N} \Gamma_{ij}, & i = j \end{cases}, \tag{3.37}$$

where r_c is a cutoff distance. The GNM theory evaluates the matrix inverse by $(\Gamma^{-1})_{ii} = \sum_{k=2}^{N} \lambda_k^{-1} \left[\mathbf{u}_k \mathbf{u}_k^T\right]_{ii}$, where T is the transpose and λ_k and $\mathbf{u}_k$ are the kth eigenvalue and eigenvector of Γ, respectively. The summation omits the first eignmode whose eigenvalue is zero.

The FRI prediction of the ith B-factor of the biomolecule can be given by[49,75]

$$B_i^{\mathrm{FRI}} = a\frac{1}{\sum_{j,j\neq i}^{N} w_j \Phi(r_{ij};\eta)} + b, \forall i = 1, 2, \cdots, N, \tag{3.38}$$

where a and b are fitting parameters, $f_i = \frac{1}{\sum_{j,j\neq i}^{N} w_j \Phi(r_{ij};\eta)}$ is the ith flexibility index and $\mu_i = \sum_{j,j\neq i}^{N} w_j \Phi(r_{ij};\eta)$ is the ith rigidity index. Here, w_j is an atomic number depended weight function that can be set to $w_j = 1$ for a C_α network, and $\Phi(r_{ij};\eta)$ is a real-valued monotonically decreasing correlation function satisfying the following admissibility conditions

$$\Phi(r_{ij};\eta) = 1 \quad \text{as} \quad r_{ij} \to 0 \tag{3.39}$$

$$\Phi(r_{ij};\eta) = 0 \quad \text{as} \quad r_{ij} \to \infty, \tag{3.40}$$

where η is a scale parameter. Delta sequences of the positive type[73] are good choices. Many radial basis functions are also admissible.[49,75] Commonly used FRI correlation functions

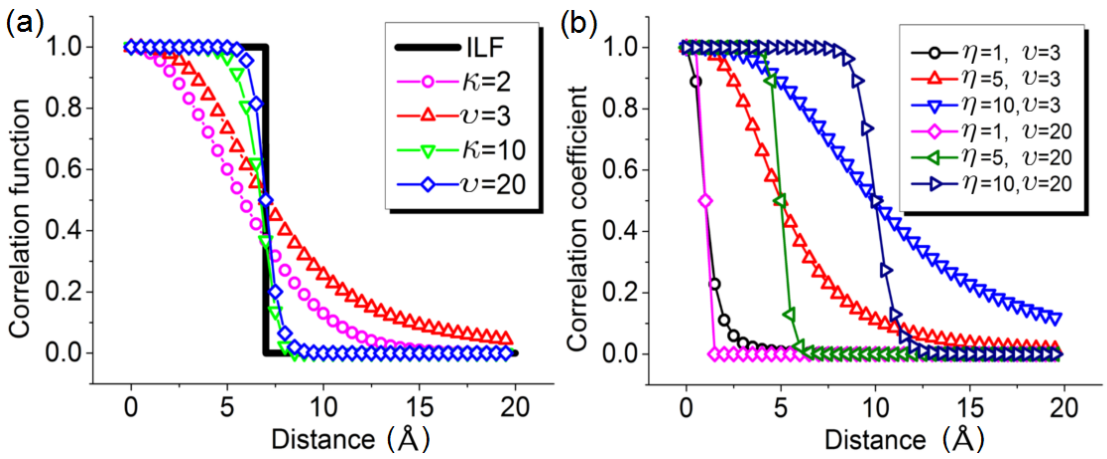


Figure 3.2: Illustration of admissible correlation functions. (a) Correlation functions approach the ILF as $\kappa \to \infty$ or $\upsilon \to \infty$ at $\eta = 7$Å. (b) Effects of varying scale value η. Local correlation is obtained with large υ and small η values. Whereas, nonlocal correlation is generated by small υ and large η values.

include the generalized exponential functions

$$\Phi(r_{ij};\eta,\kappa) = e^{-(r_{ij}/\eta)^{\kappa}}, \quad \kappa > 0, \tag{3.41}$$

and generalized Lorentz functions

$$\Phi(r_{ij};\eta,\upsilon) = \frac{1}{1+(r_{ij}/\eta)^{\upsilon}}, \quad \upsilon > 0. \tag{3.42}$$

A major advantage of the FRI method is that it does not resort to mode decomposition and its computational complexity can be reduced to $O(N)$ by means of the cell lists algorithm used in fast FRI (fFRI).[49] In contrast, the mode decomposition of NMA and GNM has the computational complexity of $O(N^3)$.

To further explore the theoretical foundation of GNM, we examine the parameter limits of the generalized exponential functions (3.4) and the generalized Lorentz functions (3.5).

$$e^{-(r_{ij}/\eta)^{\kappa}} \to \Phi(r_{ij};r_c) \quad \text{as} \quad \kappa \to \infty \tag{3.43}$$

$$\frac{1}{1+(r_{ij}/\eta)^{\upsilon}} \to \Phi(r_{ij};r_c) \quad \text{as} \quad \upsilon \to \infty, \tag{3.44}$$

where $r_c = \eta$ and $\Phi(r_{ij}; r_c)$ is the ideal low-pass filter (ILF) used in the GNM Kirchhoff matrix

$$\Phi(r_{ij}; r_c) = \begin{cases} 1, & r_{ij} \leq r_c \\ 0, & r_{ij} > r_c \end{cases} . \tag{3.45}$$

Relations (3.43) and (3.44) connect FRI correlation functions to the GNM Kirchhoff matrix. It is important to note that the ILF used in GNM is an admissible FRI correlation function. Mathematically, the ILF is a special real-valued monotonically decreasing correlation function and also satisfies admissibility conditions (3.2) and (3.3). In fact, all FRI correlation functions are low-pass filters as well. Therefore, both GNM and FRI admit low-pass filters in their constructions. GNM is very special in the sense that there is only one ILF used even though there are infinitely many other low-pass filters. Figure 3.2: illustrates the behavior and relationship of the above low-pass filters or correlation functions. It is shown that the generalized exponential function and generalized Lorentz function may be longer-ranging and the former decays faster than the latter for a given power. The combination of a low power value and a large scale gives rise to non-local correlations. Earlier tests indicate that a parameterization of $\upsilon = 3$ and $\eta = 3$Å for Lorentz kernel FRI provides accurate flexibility predictions on a set of 364 proteins relative to GNM.[49]

To describe the mathematical foundation and relationship between the GNM and FRI methods, we consider a generalized Kirchhoff matrix[76]

$$\Gamma_{ij}(\Phi) = \begin{cases} -\Phi(r_{ij}; \eta), & i \neq j \\ -\sum_{j,j\neq i}^{N} \Gamma_{ij}(\Phi), & i = j \end{cases} , \tag{3.46}$$

where $\Phi(r_{ij}; \eta)$ is an admissible FRI correlation function. The generalized Kirchhoff matrix includes the Kirchhoff matrix as a special case. It is important to note that each diagonal element is an FRI rigidity index: $\mu_i = \Gamma_{ii}(\Phi)$. Therefore, the generalized Kirchhoff matrix

provides a unified starting point for both the FRI and gGNM methods. However, the difference between the gGNM and FRI methods is that to predict B-factors, the gGNM requires the calculation of the inverse of the Kirchhoff matrix (3.37), whereas, the FRI takes the direct inverse of only the diagonal elements of the generalized Kirchhoff matrix (3.46).

3.7 Multiscale Gaussian network model (mGNM)

The idea behind mGNM is to build a multiscale Kirchhoff matrix, which incorporates various scales instead of a single one. Due to the intrinsic relation between FRI and gGNM discussed in Section 3.6, we make use of the coefficients approximated from the FRI algorithm to construct a multiscale Kirchhoff matrix. In this section, we present two types of algorithms to construct an mGNM method.

3.7.1 Type-1 mGNM

First, we assume that the multiscale Kirchhoff matrix takes the form

$$\Gamma = \sum_n a^n \Gamma^n, \tag{3.47}$$

where a^n and $\Gamma^n = \left(\Gamma_{ij}(\Phi^n(r_{ij};\eta_j^n))\right)$ are the fitting coefficient and generalized Kirchhoff matrix associated with the nth kernel $\Phi^n(r_{ij};\eta^n))$ parameterized at an appropriate scale η^n. We use the mFRI method to evaluate coefficients $\{a^n\}$. Basically, we have multiscale rigidity index $\mu_i = \sum_n a^n \Gamma_{ii}^n$. Then, $\{a^n\}$ are determined via the minimization $\text{Min} \sum_i \left| \frac{1}{\mu_i} - B_i^e \right|^2$, which is equivalent to

$$\text{Min}_{a^n} \left\{ \sum_i \left| \sum_n a^n \Gamma_{ii}^n - \frac{1}{B_i^e} \right|^2 \right\}, \tag{3.48}$$

assuming that $B_i^e > 0$. With the multiscale Kirchhoff matrix given in Eq. (3.47), we carry out routine GNM analysis as described in Eq. (3.36).

3.7.2 Type-2 mGNM

Another algorithm for constructing an mGNM method makes use of fitting coefficients from mFRI directly via the relationship between biomolecular local packing density and flexibility. Basically, we choose several kernels parameterized at various scales and evaluate the best fitting coefficients $\{a_n\}$ and b, with the experimental B-factors using Equation (3.48). The resulting multiscale flexibility index is then used to construct the generalized Kirchhoff matrix as follows

$$\sum_n a^n f_i^n + b = \frac{1}{\Gamma_{ii}}, \forall i = 1, 2, \cdots, N. \tag{3.49}$$

With the relation $f_i^n = \frac{1}{\mu_i^n}, \forall i = 1, 2, \cdots, N$, the above expression can be rewritten as,

$$\Gamma_{ii} = \frac{1}{\sum_n \frac{a^n}{\mu_i^n} + b}, \forall i = 1, 2, \cdots, N. \tag{3.50}$$

Usually, we can use two or three kernels parameterized at different scales. For instance, if we use two kernels, we can further rewrite the above expression as,

$$\Gamma_{ii} = \frac{\mu_i^1 \mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b\mu_i^1 \mu_i^2}, \forall i = 1, 2, \cdots, N. \tag{3.51}$$

Now the problem is to determine the non-diagonal terms of a multiscale Kirchhoff matrix. One simple approach is to subdivide either of the two rigidity indices. For example, we can choose to use the rigidity index for the first kernel. Since we have $\mu_i^n = \sum_{j,j\neq i}^N w_j^n \Phi^n(r_{ij}; \eta^n),\ n = 1, 2$, diagonal term of the mGNM matrix can also be expressed as

$$\Gamma_{ii} = \sum_{j,j\neq i} \frac{\{w_j^1 \Phi^1(r_{ij}; \eta^1)\}\mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b\mu_i^1 \mu_i^2}, \forall i = 1, 2, \cdots, N. \tag{3.52}$$

In this way, the full multiscale Kirchhoff matrix can be expressed as

$$\Gamma_{ij} = \begin{cases} -\frac{\{w_j^1 \Phi^1(r_{ij}; \eta^1)\}\mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b\mu_i^1 \mu_i^2}, & i \neq j \\ -\sum_{j,j\neq i}^N \Gamma_{ij}, & i = j \end{cases}. \tag{3.53}$$

The problem with the matrix in Eq. (3.53) is that the resulting multiscale Kirchhoff matrix is not symmetric, which may lead to computational difficulty. To avoid a non-symmetric matrix, we propose an alternative construction to preserve the symmetry of the matrix.

The alternative is to determine the diagonal terms Γ_{ii} from Eq. (3.50) and then on each row, equally distribute the diagonal term into the non-diagonal parts, under the condition that the resulting matrix remains symmetric. This is shown as an iterative scheme in .

Algorithm 2: Type-2 mGNM multiscale Kirchhoff matrix

Input: $\Gamma_{ii}, i = 1, 2, \cdots, N$ ▷ Diagonal terms are calculated from mFRI

for $j \leftarrow 2, N$ **do** ▷ For the first row and first line of multiscale Kirchhoff matrix.
  $\Gamma_{1j} = \frac{\Gamma_{11}}{N-1}$ ▷ We equally distribute the diagonal terms into non-diagonal parts.
  $\Gamma_{j1} = \Gamma_{1j}$ ▷ Use the symmetry property.
end for

for $i \leftarrow 2, N-1$ **do**
  $sum = 0$
  for $k \leftarrow 1, i-1$ **do**
    $k_1 = k$
    $k_2 = k+1$
    $sum = sum + \Gamma_{k_1 k_2}$ ▷ Summarize over terms already determined from previous iterations.
  end for
  for $j \leftarrow i+1, N$ **do**
    $\Gamma_{ij} = \frac{\Gamma_{ii} - sum}{N-i}$ ▷ We equally distribute the diagonal terms into non-diagonal parts.
    $\Gamma_{ji} = \Gamma_{ij}$ ▷ Use the symmetry property.
  end for
end for

In the construction of the Type-2 mGNM, only the diagonal terms are fixed and determined using mFRI. In B-factor prediction, the non-diagonal values can be very flexible as long as they satisfy the network constraint that the summation of their values equals the diagonal term. We believe this is due to the fact that the success of mGNM in B-factor prediction is determined mostly by the packing information stored in the diagonal terms of its Kirchhoff matrix. In the following discussion, we only use the symmetric scheme in Algorithm for the Type-2 mGNM.

3.8 Multiscale anisotropic network model (mANM)

In mANM, the generalized local 3×3 Hessian matrix H_{ij}^n associated with the nth kernel can be written as

(3.54)

$$H_{ij}^n = -\frac{\Phi^n(r_{ij};\eta^n)}{r_{ij}^2}\begin{bmatrix} (x_j-x_i)(x_j-x_i) & (x_j-x_i)(y_j-y_i) & (x_j-x_i)(z_j-z_i) \\ (y_j-y_i)(x_j-x_i) & (y_j-y_i)(y_j-y_i) & (y_j-y_i)(z_j-z_i) \\ (z_j-z_i)(x_j-x_i) & (z_j-z_i)(y_j-y_i) & (z_j-z_i)(z_j-z_i) \end{bmatrix} \quad \forall\ i \neq j.$$

Note that Hinsen[31] has proposed a special case: $\Phi^n(r_{ij};\eta^n) = e^{-\left(\frac{r_{ij}}{\eta^n}\right)^2}$. We further take the diagonal parts as $H_{ii}^n = -\sum_{i\neq j} H_{ij}^n, \forall i = 1, 2, \cdots, N$. Basically, it is the summation of all the non-diagonal local matrices.

The key component of mANM is to construct a multiscale Hessian matrix employing several Hessian matrices parameterized at different scales and determine their coefficients in the final multiscale Hessian matrix by using mFRI. It should be noticed that for B-factor prediction, each 3 diagonal terms from the inverse Hessian matrix are summarized together. Therefore, in the Hessian matrix based mFRI, the rigidity index associated with the nth kernel is constructed as the summation of the diagonal terms,

(3.55)

$$\mu_i^n = \sum_{i\neq j} \frac{\Phi^n(r_{ij};\eta^n)}{r_{ij}^2}[(x_j-x_i)^2 + (y_j-y_i)^2 + (z_j-z_i)^2] = \sum_{i\neq j} \Phi^n(r_{ij};\eta^n), \forall i = 1, 2, \cdots, N.$$

Indeed, the rigidity index of mANM defined above is the same as the mFRI rigidity index. Therefore, as far as B-factor prediction is concerned, the mFRI approach for constructing an mGNM should work for constructing an mANM as well.

We adopt the approach used in the Type-1 mGNM construction to construct an mANM. We propose a multiscale Hessian matrix $H = \sum_n a^n H^n$, for which the coefficients a^n should

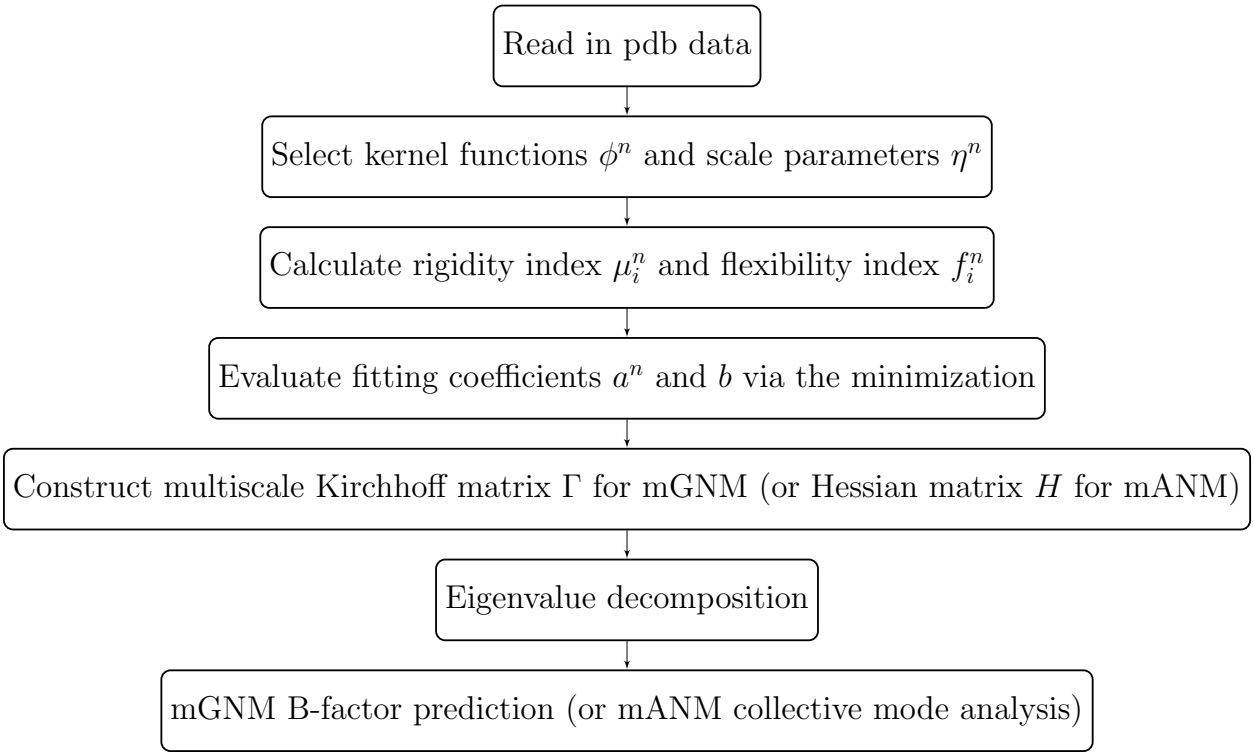


Figure 3.3: Work flow of basic procedure in mGNM and mANM.

be evaluated from

$$\text{Min}_{a^n}\left\{\sum_i \left|\sum_n a^n \mu_i^n - \frac{1}{B_i^e}\right|^2\right\}. \tag{3.56}$$

Again, different matrices $\{H^n\}$ should be parameterized at different scales.

To clarify the proposed multiscale Gaussian network model and multiscale anisotropic network model, we present a flow chart in Fig. 3.3: to illustrate the basic procedure that outlining the methods.

3.9 gGNM mode calculations for predicting hinges

In a plot of gGNM mode values, hinges are often observed where the values of a gGNM mode switch sign and change value drastically. Furthermore there should be a significant number of negative values to one side of the sign switch and positive values on the other side to indicate an actual separation between domains. Otherwise, there may be extraneous

hinge predictions in regions that have many values close to zero and may switch signs many times over the span of just a few residues.

To more easily find potential hinge residues from gGNM modes as described above and in a manner that filters out the minor sign changes, we utilize the cumulative sum of a mode then determine the residue number corresponding to the maximum and/or minimum values of that series of cumulative sums. This technique can identify any number of hinges as long as the number of peaks is identified correctly. For this study, the code detects only the maximum and minimum of the cumulative sum of the first mode which limits detection to two hinges, however this should not affect this study except to possibly limit false positive predictions because the proteins used for testing have at most two major hinge regions. Also, any residues too near either end of the molecule to be separating domains, i.e. residues within 35 residues of the end, are removed from the set of predicted hinges.

The eigenvalues of the relevant modes can be very similar in some cases while the number and/or locations of hinges predicted by those modes is different. In this case we may equally consider all modes with similar eigenvalues. This may lead to additional false positives but is necessary for ensuring maximum sensitivity which is a priority.

3.10 Machine learning and feature selection

The features we test in an attempt to improve on the KFC2 SVM model include various metrics derived from FRI flexibility predictions, FRI-mode calculations, hydrophobicity, sequence based statistics and details of the structure.

Features derived from FRI flexibility predictions include the FRI and mFRI flexibility indices, *FRIf* and *mFRIf*, FRI and mFRI B-factor fitted values and their difference, *FRI, mFRI and dFRI*, and the average mFRI B-factor fitted value within six Angstroms, *avgFRI*.

Features derived from FRI-mode calculations are created from the first two lowest eigenvalued modes of motion. The first lowest eigenvalued mode typically corresponds to the

most important or largest in amplitude motion, therefor we derive most of the features from this mode. The features derived from the first mode include the raw values of the mode residue-by-residue, *Mode 1*, the cumulative sum of the mode values, *cMode 1*, the residue numbers where *Mode 1* is at a local maximum or minimum, *ishinge*, all residues within three of *ishinge*, *ishinge3*, and finally the number of residues to the nearest positive *ishinge* value, *hingedist*. Also included are features derived from mode 2, *Mode 2* and *cMode 2*.

The hydrophobicity of individual residues and regions of a protein is the main driving force behind folding of a protein structure. Therefore, we suspect that hinge residues, which occupy specific positions between domains, are under some pressure to maintain a certain level of hydrophobicity, likely high hydrophobicity due to the sovlent exposed nature of many hinges. The hydrophobicity derived features we test in this study include the hydrophobicity of a single residue, *HP1*, and the average hydrophobicity for residues in a six Angstrom sphere around a residue, *HP6*. Two other metrics are also tested that describe the area within 6 Angstroms of a residue, *RES6* and *ROT6*, the number of residues and the number of rotatable bonds in that region. These features have been used in machine learning models for predicting protein-protein interaction hotspots and showed relatively high predictive worth.

In addition to *RES6* and *ROT6*, we also look at solvent accessible surface area (SASA) based metrics to describe the area around a residue. Using the POPS software package we generate features for hydrophobic SASA, *PhobSASA*, hydrophilic SASA, *PhilSASA*, total SASA, *TotSASA*, number of overlapping atoms, *N(ovrl)*, and the total surface area of the residue, *Surf*.

Finally, we include PSSM features, one feature for each amino acid log likelihood at each position. PSSM derived features are named as *A (PSSM)*, *G (PSSM)*, etc.

A wide range of metrics, derived from various molecular properties, were considered as potentially useful for hinge prediction. These metrics were compared by first filtering by

F-score threshold then ranking the remaining features by the Random Forest importance value.

CHAPTER IV. Validation and Applications

4.1 Basic FRI method

4.1.1 FRI B-factor prediction

To validate the original FRI method, we compare the B-factor predictions obtained from FRI with experimental B-factors from protein X-ray crystallography experiments as shown in Eq. (4.6). A set of 263 proteins was collected from the PDB with preference for high resolution (1.5 Å) protein-only structures that lack structural co-factors. The impact of co-factors on protein stability requires an all atom model and is a topic that will be explored in our future work. The set of 263 proteins was converted to a Cα only format and when atoms have multiple coordinates with occupancy <1.0 the highest occupancy coordinate was kept and all others were discarded. This is a potential source of error in the B-factor predictions. However some proteins with multiple coordinates for atoms were among the highest scoring which suggests that the impact in most cases is small.

The correlation coefficients of B-factor prediction are displayed in Fig. 4.1: for both exponential (expo) and Lorentz kernels. Each protein was tested with both the exponential and Lorentz correlation kernels across a range of parameter values of κ and η for the exponential kernel and υ and η for the Lorentz kernel. Correlation coefficient scores for B-factor predictions below 0.5 account for just 19 of out 263 proteins for the Lorentz kernel based FRI and 14 out of 263 for the exponential kernel based FRI and are not shown in Fig. 4.1:. The reasons for these low scores are the subject of future research and are likely related to the influence of crystal packing effects, structural ligands and side-chain effects that are not approximated well by the Cα course grained model. The accuracy of B-factor prediction is also dependent upon the quality of the experimental data. If multiple coordinates are reported for an atom along with multiple B-factors, then we do not have high confidence in

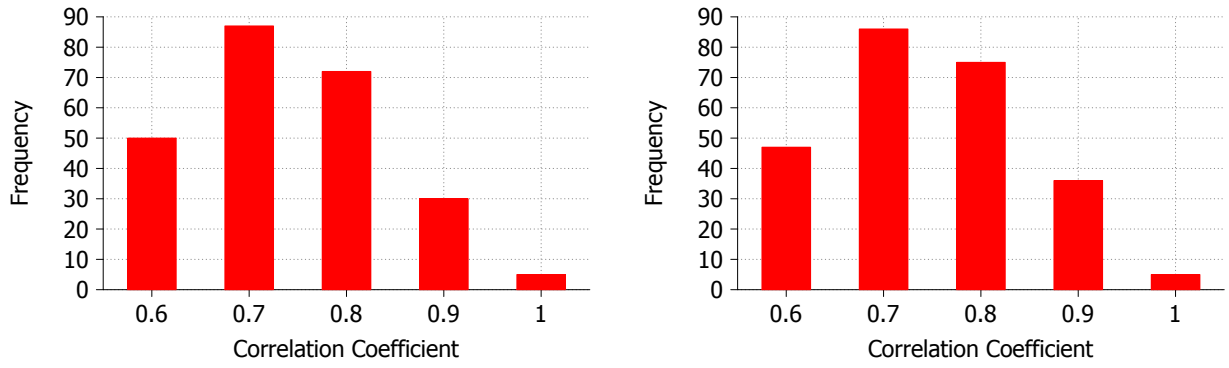


Figure 4.1: Correlation coefficients for experimental vs predicted B-factors using the Lorentz kernel (left) and exponential (right) kernel. The test set consists of 263 C_α only PDB files. Scores below 0.5 are not shown. For the Lorentz kernel, υ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 1.0Å to 40.0Å at an interval of 1.0Å . For the exponential kernel, κ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 0.5Å to 20.0Å at an interval of 0.5Å .

the B-factor and thus the prediction will appear to be less accurate.

A comparison of the experimental vs predicted B-factors for two proteins, 1DF4 and 2Y7L, is shown in Fig. 4.2: to demonstrate the accuracy of our FRI method. These two proteins were in the top five highest correlation coefficients for B-factor predictions using the exponential (2Y7L: 0.928, 1DF4: 0.909) and Lorentz (2Y7L: 0.928, 1DF4: 0.917) kernels. It can be seen from the correlation scores and Figs. 4.1: and 4.2: that both correlation kernels give similar results, especially for these highly accurate predictions.

B-factor prediction was calculated for each protein at a range of parameter values in each kernel. The Lorentz kernel requires parameters, υ and η, while the exponential kernel requires κ and η. The aim is to find values for these parameters that are suitable for most or all proteins so that the method may be made parameter free. The parameters which result in the highest correlation coefficient for each protein are displayed in Fig. 4.3: and Fig. 4.5: for the Lorentz and exponential kernels, respectively.

The optimal value for υ in the Lorentz kernel is found to be near 2.5 for most proteins

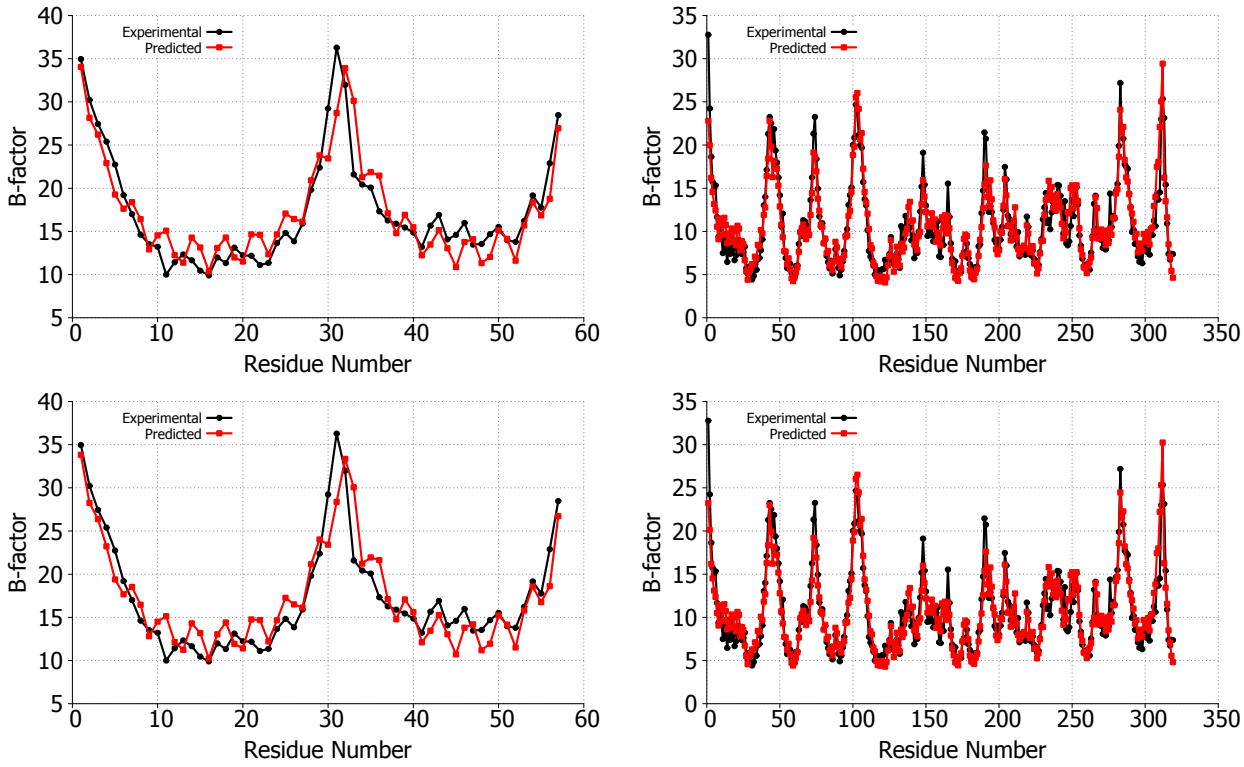


Figure 4.2: Experimental B-factors (black) vs predicted B-factors (red) using the Lorentz (top) and exponential (bottom) correlation kernels. The structures used for comparison are 1DF4 (left) and 2Y7L (right). For these comparisons, the optimal parameters were used for υ, κ and η based on the parameter searches for each correlation kernel. For the Lorentz kernel, υ=1.5 and η=2.0Å are the parameters used for 1DF4 and υ=1.5 and η=19Å are used for 2Y7L. For the exponential kernel, κ=0.5 and η=1.0Å are employed for 1DF4 and κ=0.5 and η=2.5Å for 2Y7L.

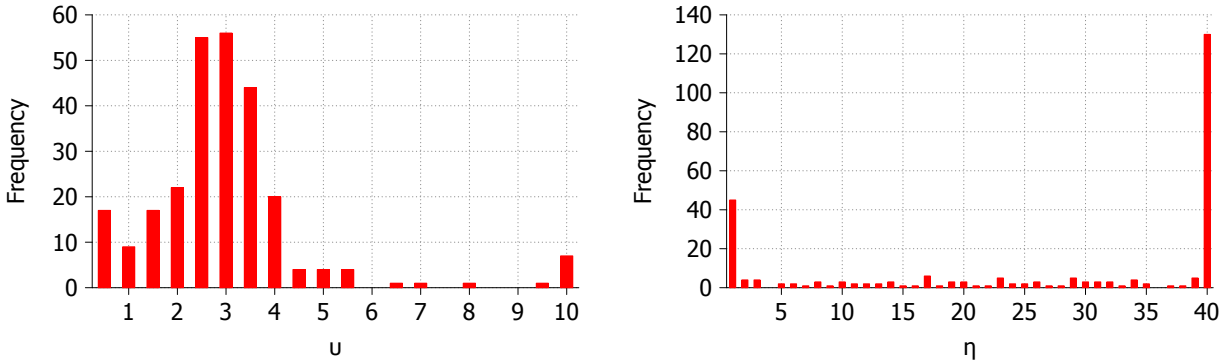


Figure 4.3: Optimal υ parameter value for 263 proteins using the Lorentz correlation kernel. B-factor prediction was calculated for υ values ranging from 0.5 to 10 at an interval of 0.5 and η values ranging from 1.0Å to 40.0Å at an interval of 1.0Å .

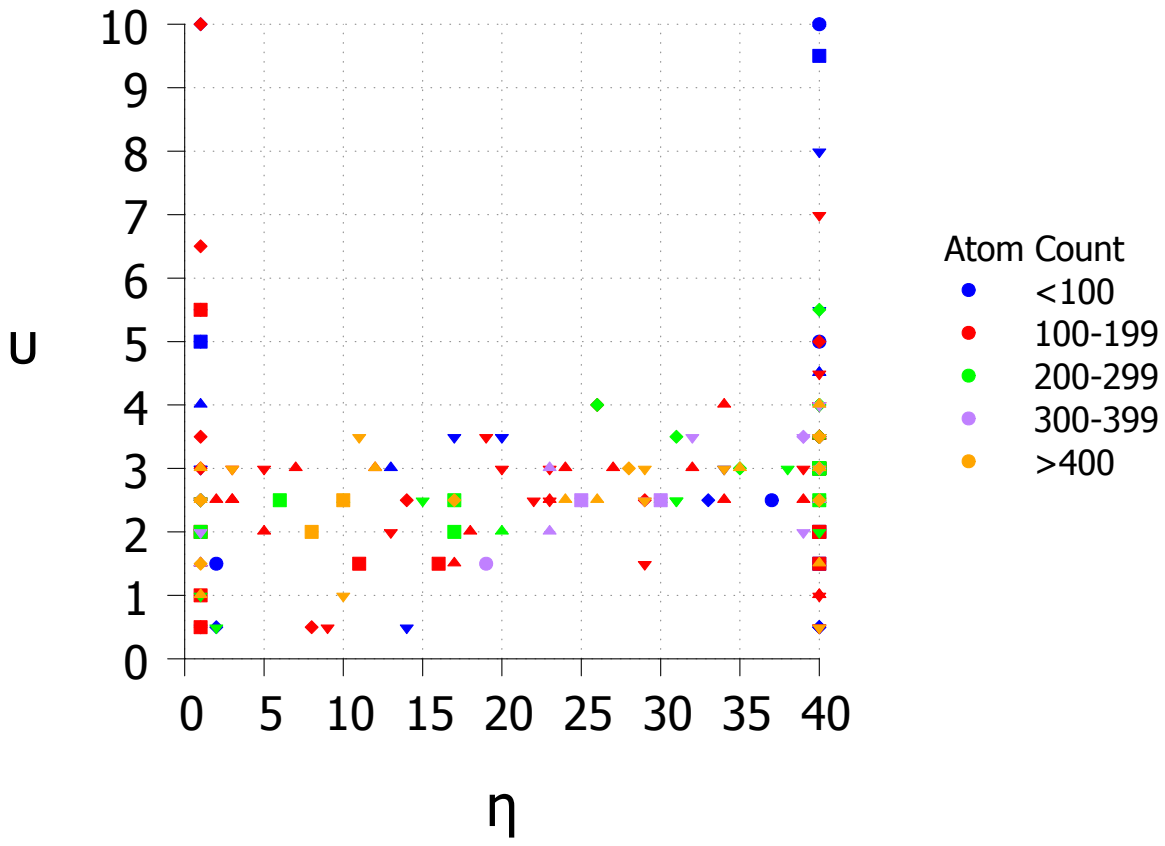


Figure 4.4: Phase diagram for Lorentz kernel optimal parameter values υ and η colored by the size of structure and with shapes corresponding to correlation coefficient. Diamond - 0.5, downward triangle - 0.6, upward triangle - 0.7, square - 0.8 , circle - 0.9. υ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 1.0Å to 40Å at an interval of 1.0Å

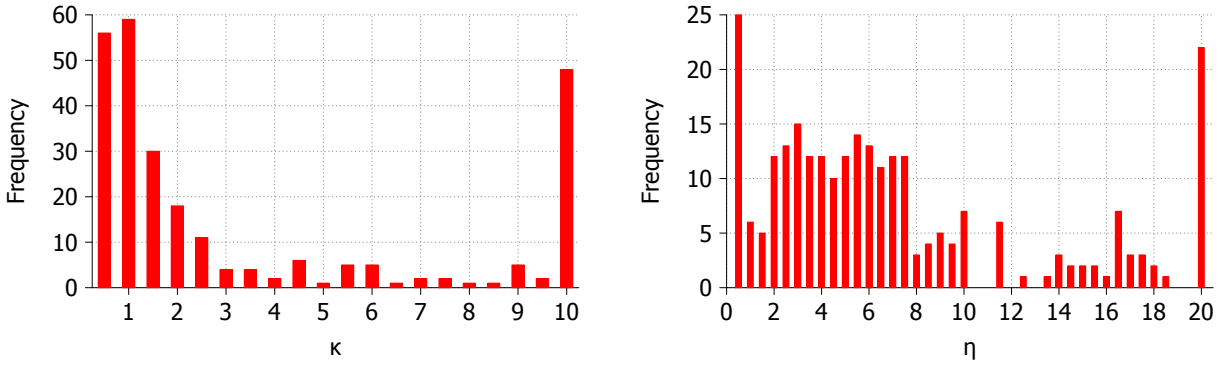


Figure 4.5: Optimal parameters for 263 structures using the exponential correlation kernel. Here κ values range from 0.5 to 10.0 at an interval of 0.5. η values range from 0.5Å to 20.0Å at an interval of 0.5Å.

in the test set. The optimal value for η is typically the highest or lowest tested. The results of the parameter search for υ and η are shown in Fig. 4.3:. This result is a close match to the findings of Yang et al.[79] and their parameter free ENM (pfENM) model. In the pfENM, spring constants are scaled by an inverse power. Yang et al. tested powers 1-10 and found second and third inverse power relationships were the most accurate for B-factor predictions.[79] In our study we also test non-integer powers over the range 0.5 to 10.0 and come to a similar conclusion. The optimal values for η in these tests The optimal value for υ is plotted against the optimal value for η and colored by the size of protein in Fig. 4.4:. There is no clear pattern based on protein size except that some smaller proteins (under 100 atoms) prefer very high values of υ which may be due to a lack of long range interactions.

For the exponential kernel, the optimal κ value for most proteins is between 0.5 and 1 while the optimal η values are more spread out with the majority of proteins having optimal η values from 0.5Å to 8Å . This ambiguity in the optimal parameter value makes the choice of parameters for a parameter free version difficult however the testing of the parameter free exponential kernel method shows that it performs as well as the parameter free Lorentz kernel methods. The optimal values for κ and η for all proteins in the test set are shown

in Fig. 4.5:. Optimal values for κ are 0.5 or 1.0 in most cases with a significant peak at κ=10 which is the highest value tested. Optimal values for η are more varied and there is no clear choice for a parameter free version. There is a large peak at the highest η value tested (η=20Å) as there was for κ however these two peaks do not correspond to the same set of proteins. This point is illustrated in Fig. 4.6: which compares κ and η values. Figure 4.6: also shows that there is no relationship between number of atoms or correlation coefficient and the parameters κ and η. To further inform our choice of parameters for the parameter free exponential method we look at the patterns of correlation scores for every κ and η value combination in Fig. 4.7:. The parameter maps show that for most proteins the choice of κ is most important and that when $\kappa \leq 1$ there are many choices for η that result in very similar correlation coefficients.

To test parameter free versions of the FRI method we chose υ=2.5 and η=1.0Å for the Lorentz kernel and κ=1.5 and η=5.0Å for the exponential kernel. These choices were made based on the parameter searches and limited tests of various parameter values. In Fig. 4.8: we compare the exponential and Lorentz kernel performance based on correlation coefficients from B-factor prediction. The correlation coefficients were highest overall when using the exponential kernel with optimized parameters. The average correlation coefficient of B-factor prediction using the exponential kernel is 0.681 using optimal parameters and 0.627 using the parameter free version. The average correlation coefficient of B-factor prediction using the Lorentz kernel is 0.668 using optimal parameters and 0.627 using the parameter free version. The difference between the exponential and Lorentz kernels is small when using optimized parameters with an average deviation of just 0.0182. The parameter free versions of the kernels also produce very similar correlation coefficients with an average deviation of 0.0365.

The parameter free Lorentz and exponential kernels appear to have similar performance

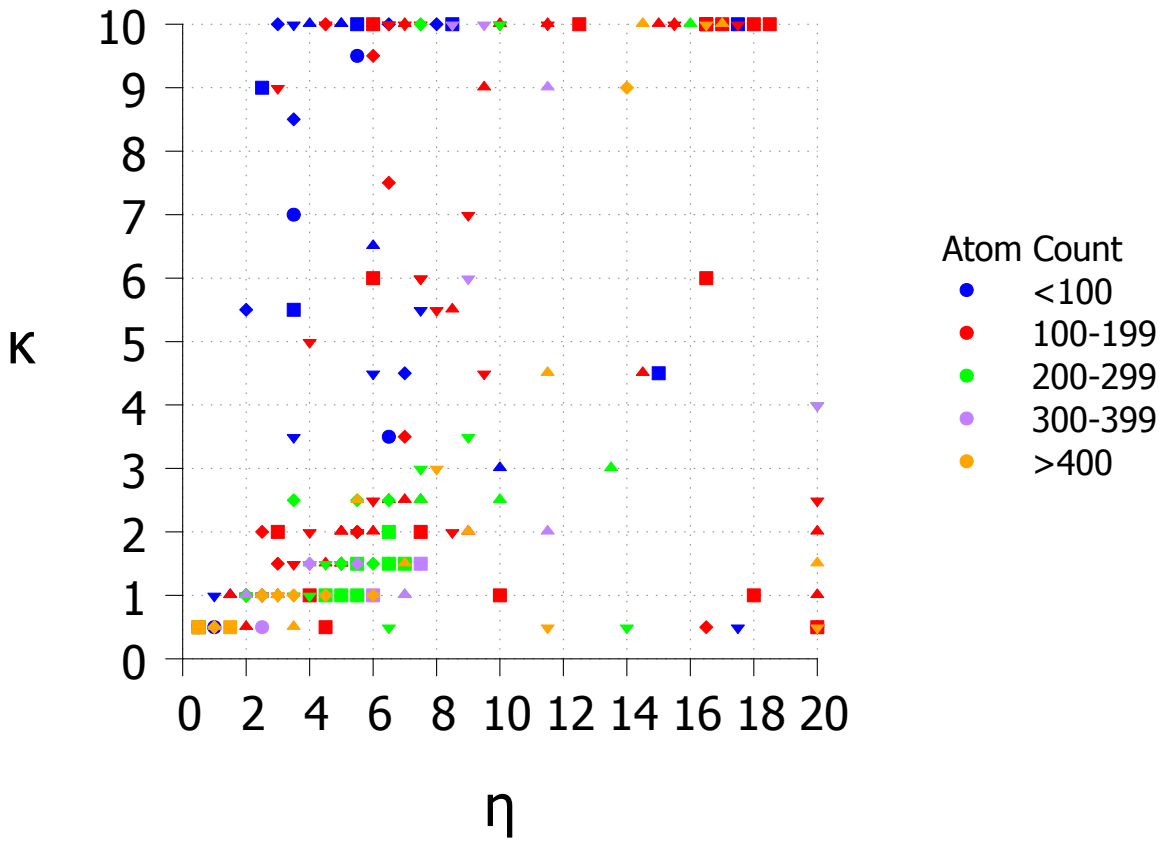


Figure 4.6: Phase diagram for exponential kernel optimal parameter values κ and η colored by the size of structure and with shapes corresponding to correlation coefficient. Diamond - 0.5, downward triangle - 0.6, upward triangle - 0.7, square - 0.8 , circle - 0.9. κ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 0.5Å to 20Å at an interval of 0.5Å .

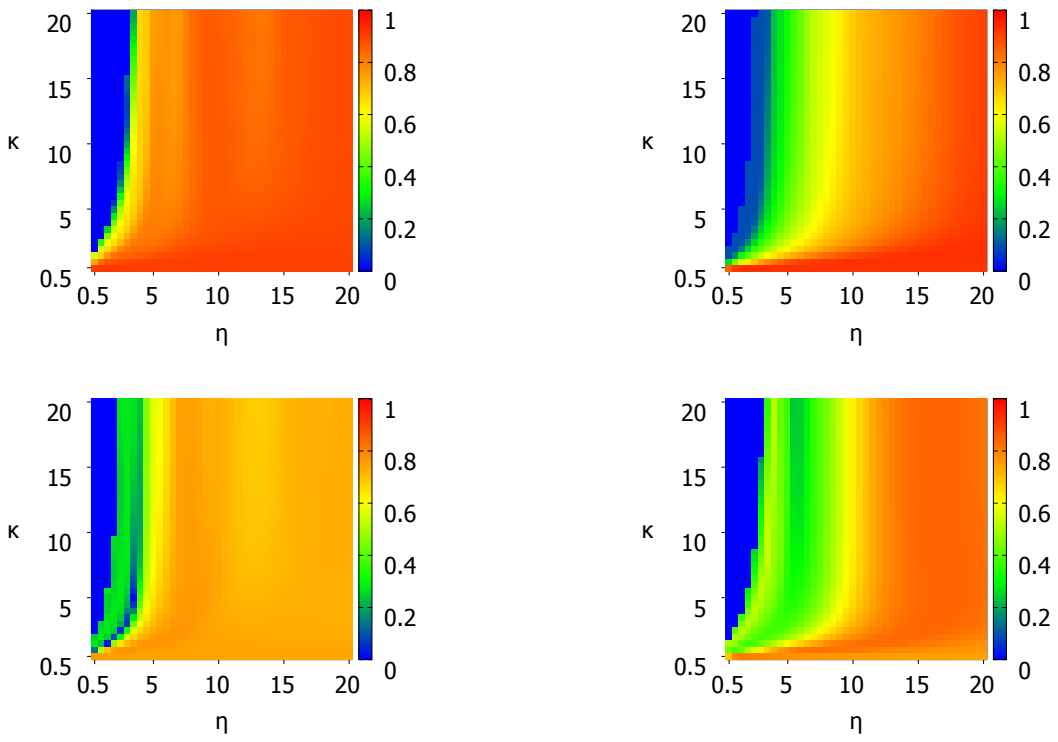


Figure 4.7: Complete results of optimal parameter searches using the exponential correlation kernel for structures 1DF4 (top left), 2Y7L (top right), 2Y9F (bottom left) and 3LAA (bottom right). Structures 1DF4 and 2Y7L (top) represent the high scoring structures, those with scores near 0.9. Structures 2Y7L and 3LAA (bottom) show the typical pattern of correlation scores for the majority of proteins tested. κ values range from 0.5 to 20.0 at an interval of 0.5 and ηvalues range from 0.5Å to 20Å at an interval of 0.5Å .

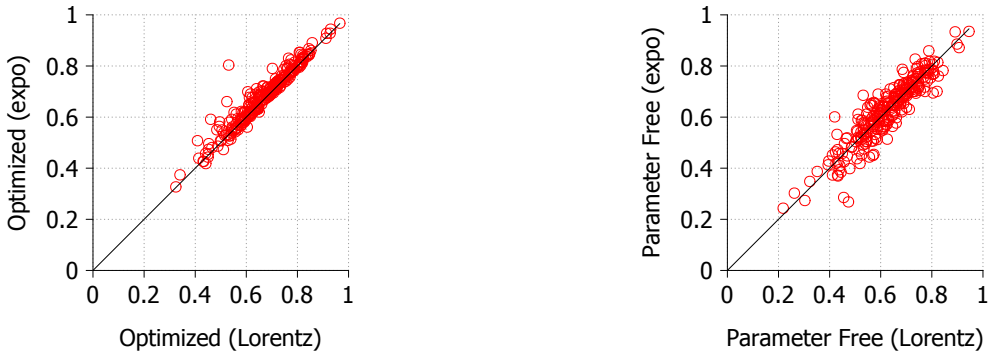


Figure 4.8: Comparison of correlation coefficients calculated using optimal parameters for both Lorentz and exponential correlation kernels. Average deviation = 0.0182 (left) and 0.0365 (right). For the Lorentz kernel optimal parameter search, υ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 1.0Å to 40.0Å at an interval of 1.0Å . For the exponential kernel parameter search, κ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 0.5Å to 20.0Å at an interval of 0.5Å . The parameter free Lorentz kernel uses υ=2.5 and η=1.0Å and the parameter free exponential kernel uses κ=1.5 and η=5.0Å .

and these results do not indicate a clear advantage in using either kernel. In Fig. 4.9: we compare the correlation coefficients from the parameter free and optimized versions of the method for both correlation kernels. In each case the optimized method outperforms the parameter free method no matter which kernel is used. Again this suggests that neither kernel has an advantage over the other for this method. The maximal average deviation among these methods is 0.0549, meaning that the parameter free exponential kernel captures 94% of the best results generated by optimized Lorentz kernel for this set of proteins. Similarly, the parameter free exponential kernel captures 94% of the best results from the optimized exponential kernel. It is worthwhile to note that the parameter free Lorentz kernel (υ=2.5 and η=1.0Å) is able to capture 95% of the best results generated by either the optimized exponential or Lorentz kernel for this set of proteins. Therefore, it appears that the both parameter free kernels are very robust for practical applications.

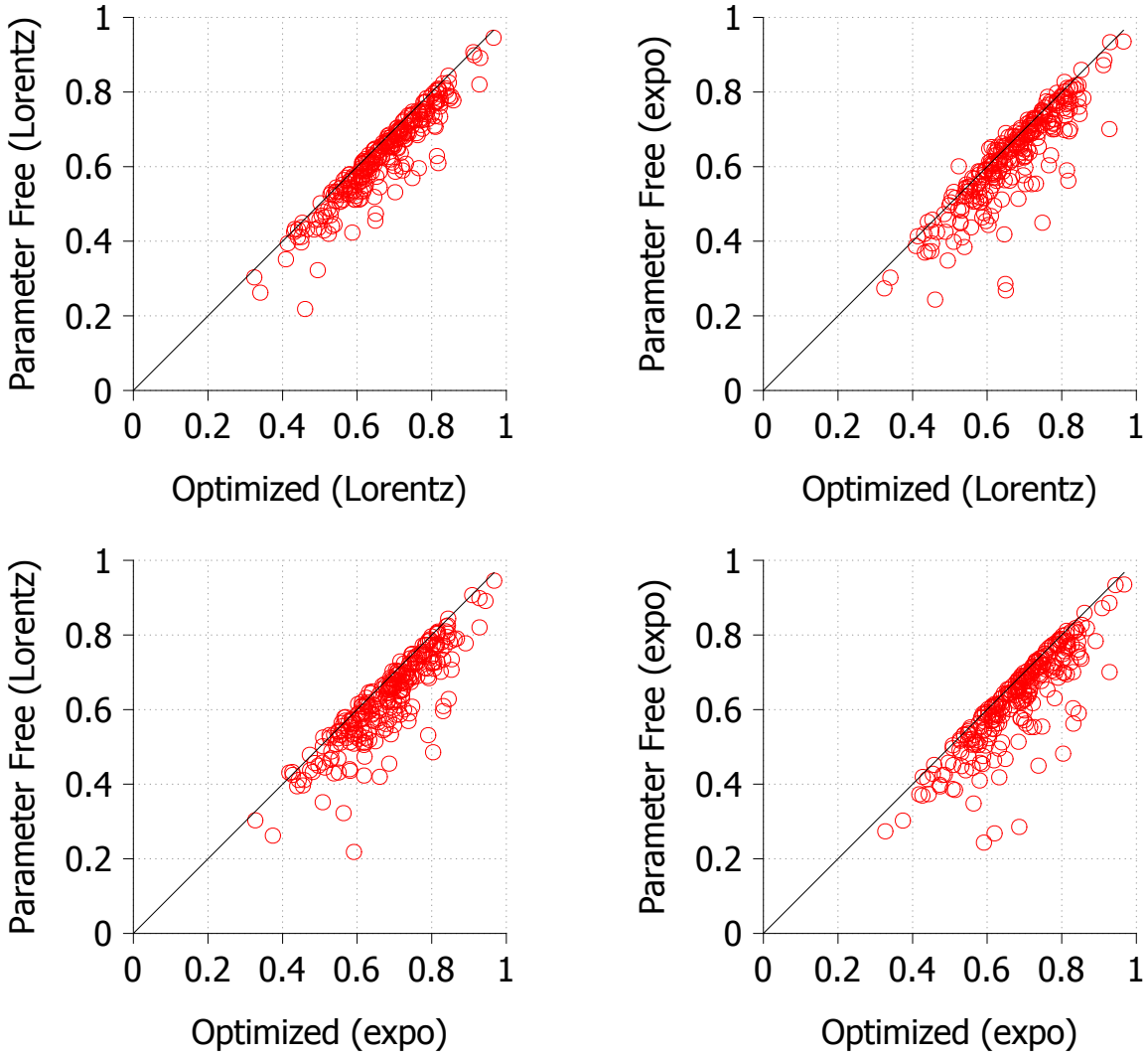


Figure 4.9: Comparison of correlation coefficients calculated using optimal parameters and parameter free versions of the method. The optimized correlation coefficients are the highest scoring from a parameter search. For the Lorentz kernel optimal parameter search, υ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 1.0Å to 40.0Å at an interval of 1.0Å . For the exponential kernel parameter search, κ values range from 0.5 to 10.0 at an interval of 0.5 and η values range from 0.5Å to 20.0Å at an interval of 0.5Å . The parameter free Lorentz kernel uses υ=2.5 and η=1.0Å and the parameter free exponential kernel uses κ=1.5 and η=5.0Å . The line $y = x$ is shown for reference. Points on the line indicate little or no difference between optimized parameters and the parameter free results. Average deviations are 0.0410, 0.0549, 0.0463, and 0.0540 (from left to right and from top to bottom).

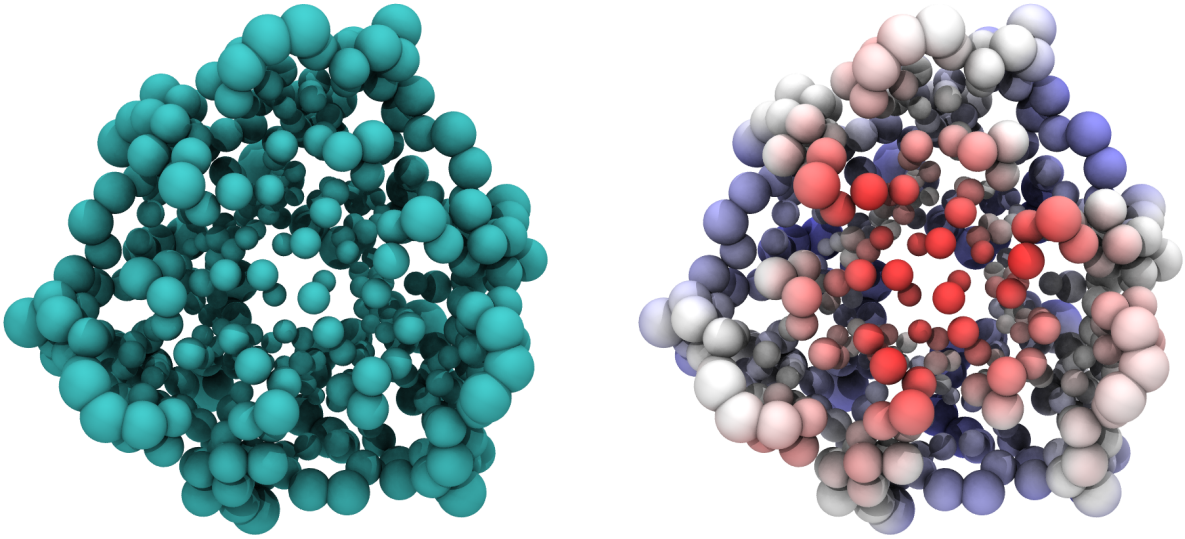


Figure 4.10: C_α atoms of 1QD9 in VDW representation scaled by predicted B-factor (both images) and colored with electrostatics (right). Larger VDW radii represent more flexible atoms such as those near the surface of this soluble protein. Smaller VDW radii represent more rigid atoms such as those in the core of the protein. On the right, atoms are colored by electrotatics revealing two charged domains. First, the flexible outer amino acids have some areas of positive charge that interact with the bulk solvent. Second, a highly negatively charged portion of the protein core is highlighted in red. These charges are stabilized by internal water molecules.

4.1.2 Rigidity and flexibility visualization

From the above analysis, the rigidity and flexibility indices can be obtained at coordinates of C_α atoms in the protein. Such values can be utilized directly for visualization. For the purpose of visualization, it is sufficient to plot either rigidity or flexibility. A large value of the flexibility index can be represented by a large atomic radius in the visualization while a small flexibility index corresponds a small atomic radius. Therefore, we scale atomic van der Waals radii by their flexibility indices as shown in Fig. 4.10: for 1QD9. Clearly, C_αs located near molecular boundary are more flexible.

Additionally, the flexibility index can be visualized together with electrostatic potential.

Specifically, the flexibility is represented by the atomic size while the electrostatics is illustrated by color as shown in the right chart of Fig. 4.10:. There is a correlation between flexibility and partial charge in this structure — charged residues are slightly less flexible. From these figures we see the image of a typical soluble protein with flexible, partially charged residues on the solvent-solute boundary and a less flexible, rigid core. It is well-known that the partially charged flexible outer protein surface is responsible for many protein functions in enzymes, cell signaling and ligand binding. Interestingly this soluble protein has a highly charged core made up of many negatively charged residues interacting with a network of water molecules. This results in a negatively charged, rigid core which is represented by small, red VDW spheres.

Furthermore, in order to study the elastic dynamics, elastostatics, and collective motion of a macromolecule, the continuous atomic rigidity and flexibility functions are required in our multiscale multiphysics multiphysics and multidomain models. The spatially scattered information at each C_α coordinate needs to be interpolated into continuous atomic rigidity and flexibility functions. In this work, we employ the modified Shepard's method to interpolate rigidity and flexibility values at C_α coordinates to build their continuous functions.[56,65] The essence of Shepard's method is to blend local interpolants with locally supported weight functions. For example, the atomic flexibility function can be expressed as

$$F(\mathbf{r}) = \sum_{i=1}^{N} W_i(\mathbf{r}) Q_i(\mathbf{r}), \tag{4.1}$$

where the locally supported weight function is defined as

$$W_i(\mathbf{r}) = \frac{p_i(\parallel \mathbf{r} - \mathbf{r}_i \parallel; R_i)}{\sum_{i=1}^{N} p_i(\parallel \mathbf{r} - \mathbf{r}_i \parallel; R_i)}, \tag{4.2}$$

$$p_i(\parallel \mathbf{r} - \mathbf{r}_i \parallel; R_i) = \begin{cases} \left(\frac{R_i - \|\mathbf{r}-\mathbf{r}_i\|}{R_i \|\mathbf{r}-\mathbf{r}_i\|}\right)^2, & \parallel \mathbf{r} - \mathbf{r}_i \parallel < R_i, \\ 0 & , \parallel \mathbf{r} - \mathbf{r}_i \parallel \geq R_i. \end{cases} \tag{4.3}$$

Here $R_i > 0$ is a constant radius with ith C_α as its center. Its value varies with i so as to include different numbers of points into its influence domain when it is necessary.[65]

Our input data are a set atomic flexibility indices $\{f_i\}$ or the predicted B-factors $\{B_i^t\}$ located at C_αs. We denote $\mathbf{r} = (x, y, z), \mathbf{r} \in S_E$ a general position inside the elastic domain of a macromolecule, and the local interpolant is a nodal function defined as,

$$Q_i(\mathbf{r}) = a_{i1}x^2 + a_{i2}y^2 + a_{i3}z^2 + a_{i4}xy + a_{i5}xz + a_{i6}yz + a_{i7}x + a_{i8}y + a_{i9}z + a_{i10}, \tag{4.4}$$

where a_{ij} are coefficients and $Q_i(\mathbf{r})$ is a quadratic polynomial function which interpolates the predicted B-factors at neighboring set of C_α locations, namely

$$Q_i(\mathbf{r}_j) = B_j^t \delta_{ij} \tag{4.5}$$

where δ_{ij} is the Kronecker delta function. For a given ith C_α, Eq. (4.5) is repeatedly employed on all C_αs within the given sphere of radius R_i and results in a number of algebraic equations. The algebraic equations are solved by using the weighted least square method, which determines coefficients a_{ij}. For sufficiently large data, we can choose 32 surrounding atomic flexibility indices to fit coefficients.[65] Note that the atomic rigidity function ($\mu(\mathbf{r})$) can be constructed in the same manner by replacing B_j^t with μ_j.

In Fig. 4.11: we compare an atomistic and a continuous representation for flexibility of protein 1QD9. The molecular surface on the left is colored by X-ray B-factors, while the molecular surface on the right is colored by the interpolated flexibility values. Overall, the interpolated values mimic the B-factor pattern closely. However, the predicted flexibility at the inner ring of the structure is higher than that given by X-ray B-factors due to the fact water molecules fill part of the inner core in the full structure. The B-factor color map is discontinuous. In contrast, the flexibility map generated with the FRI method has the advantage of being continuous both on the surface and in the interior of the protein. The atomic rigidity function and atomic flexibility function constructed in the present work will

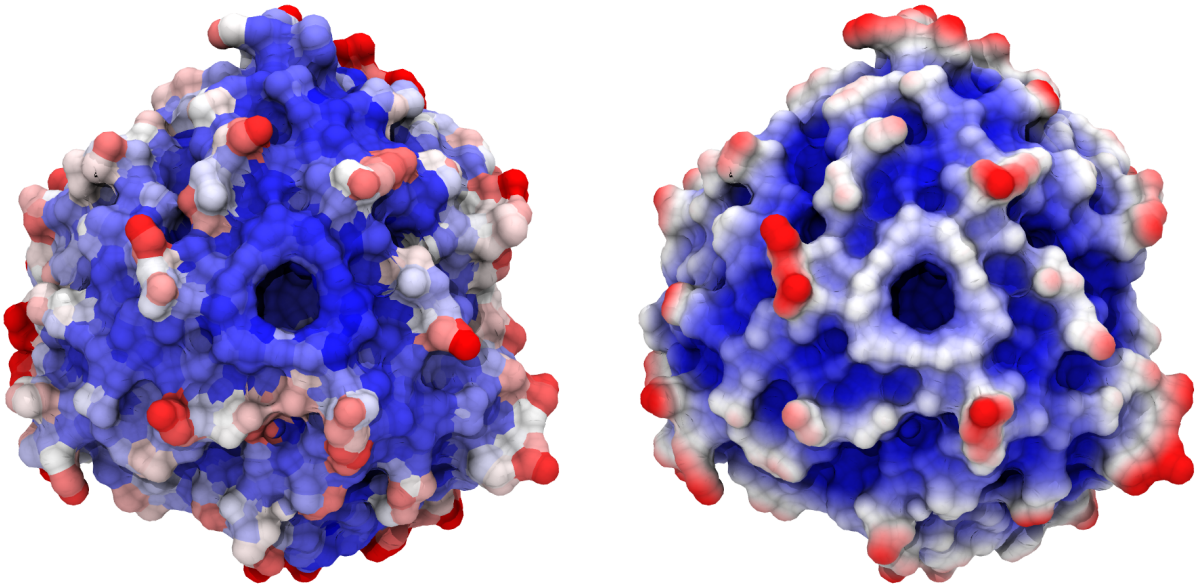


Figure 4.11: The molecular surface of Protein 1QD9 colored by B-factor (left) and continuous FRI representation (right). The flexibility index is calculated using the Lorentz method with υ=2.5 and η=1.0Å . Images generated by VMD using BWR color bar and scale 10 to 50 for B-factors and 0.75 to 0.90 for the flexibility index. In both images, blue regions indicate low flexibility and red regions indicate high flexibility. On the left, B-factor is an atomistic representation of flexibility. On the right, FRI is used to predict flexibility and the continuum representation is mapped to the protein surface. The continuum prediction matches the experimental flexibility pattern closely except for near the core of the protein which contains some structural water not included in our model.

be utilized to study macromolecular elastic dynamics, elastostatics and elastic vibration in our future work.

4.2 Fast FRI method

4.2.1 fFRI parameter testing

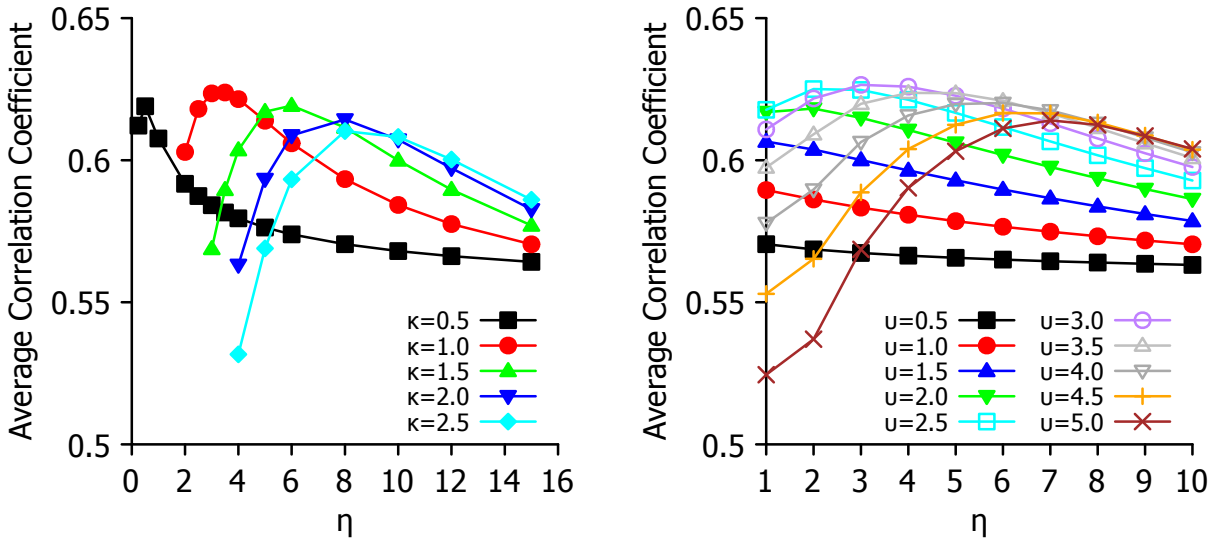


Figure 4.12: Parameter testing for exponential (Left chart) and Lorentz (Right chart) functions. Average correlation coefficient of B-factor predctions of 365 proteins is plot against choice of η for a range of values for κ or υ.

To analyze the best parameter for Lorentz and exponential functions, we study their behavior in Fig. 4.12:, where each function is tested over a range of parameters. For exponential type of functions, $\kappa = 1$ and $\eta = 3$Å give rise to a near optimal parameter-free FRI. Similarly, for Lorentz type of functions, $\upsilon = 3$, and $\eta = 3$Å offer near optimal results. It is seen from Fig. 4.12: that exponential functions are quite sensitive to η values, while Lorentz functions are relatively robust with respect to η. This study provides a basis for the selection of parameter free FRI (pfFRI) schemes.

It is interesting to analyze the performance of the proposed fFRI in terms of accuracy and efficiency. To this end, we first explore the impact of box size to the correlation coefficients

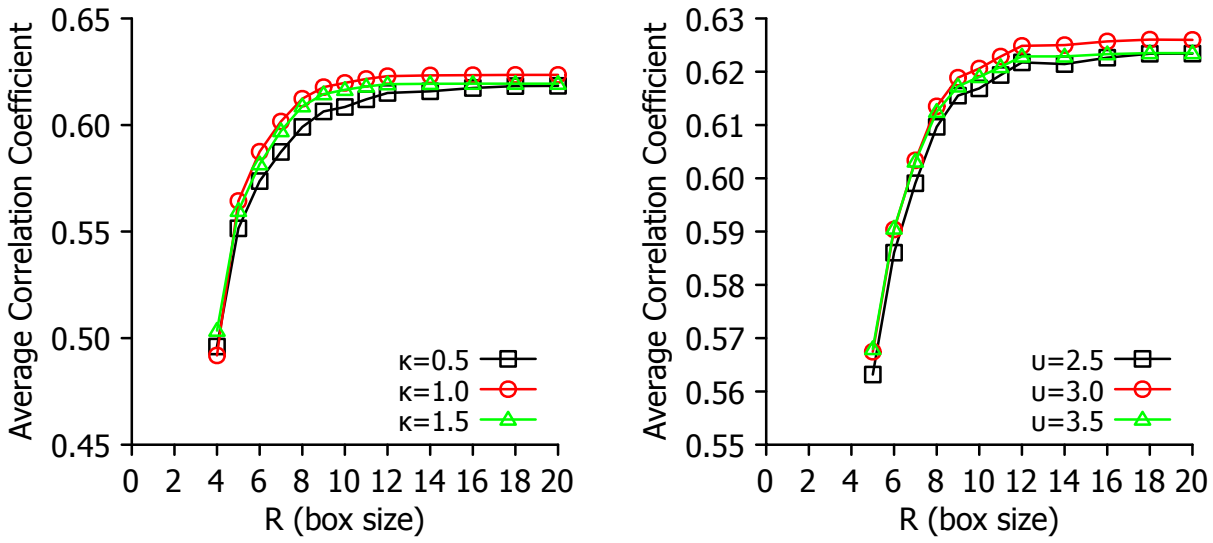


Figure 4.13: The impact of box size to the average correlation coefficient for a set of 365 proteins. The fFRI is examined over a range of values for parameters (κ and υ) to illustrate the relationship between accuracy and choice of box size R.

of a few fFRI schemes in Fig. 4.13:. For each given κ and υ, the best η found in Fig. 4.12: is employed. It is seen from Fig. 4.13: that both exponential and Lorentz types of functions are able to achieve their near optimal performance at $R = 12$Å . Therefore, we recommend $R = 12$Å, $\eta = 3$Å and $\kappa = 1$ for the exponent type of fFRI method. Similarly, $R = 12$Å, $\eta = 3$Å and $\upsilon = 3$ are near optimal for Lorentz type of fFRI methods.

4.2.2 Comparison of B-factor predictions from fFRI, GNM and NMA

4.2.2.1 FRI vs GNM and NMA

In order to compare the FRI and GNM, we re-analyzed the structures from Park et al.[52] with the GNM method with a cutoff value of 7 Å, the same value used by the authors. It was found that some correlation coefficients were artificially low for GNM due to multiple coordinates for some C_α atoms in some PDB data and missing Cα atoms in others. To ensure a fair comparison between the FRI and GNM we re-analyzed the structures using GNM after processing the PDB files to fix these issues. We removed all but the highest

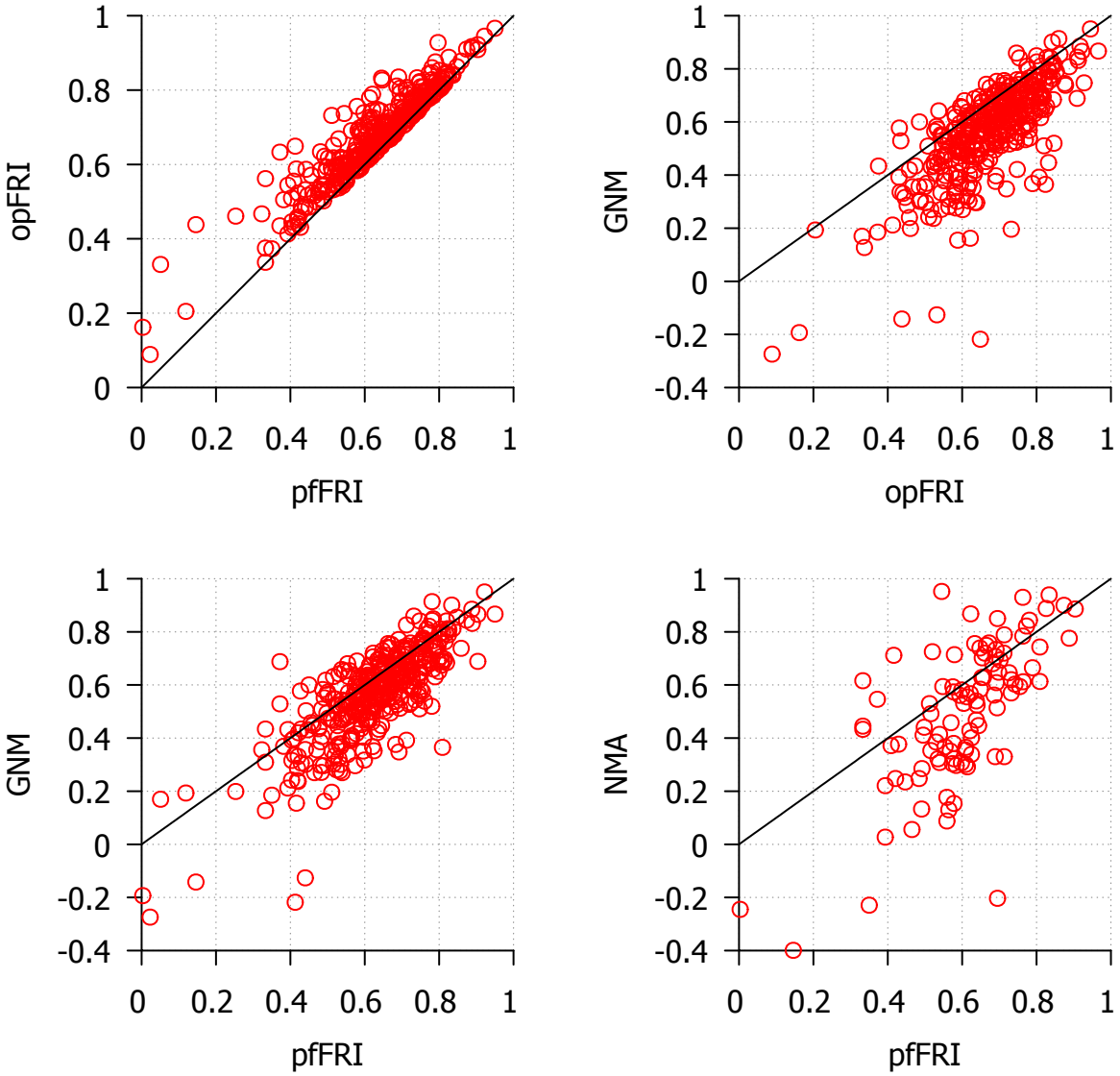


Figure 4.14: Comparison of correlation coefficients from B-factor prediction using GNM, coarse-grained (C_α) NMA and FRI methods. Top left: pfFRI vs opFRI for 365 proteins; Top right: opFRI vs GNM for 365 proteins; Bottom left: pfFRI vs GNM for 365 proteins; Bottom right: pfFRI vs NMA for three sets of proteins used by Park et al.[52] The correlation coefficients for NMA are adopted from Park et al.[52] for three sets of proteins. For optimal FRI, parameter υ is optimized for a range from 0.1 to 10.0. For the parameter free version of the FRI (pfFRI), we set $\upsilon = 3$ and $\eta = 3$Å. The line $y = x$ is included to aid in comparing scores.

occupancy coordinates for each atom and used every Cα atom from the original PDB files to run the GNM B-factor prediction code and calculate corrected correlation coefficients. In Tables 4.2:, 4.3: and 4.4:, optimal and parameter free FRI is compared to the GNM data reported by Park et al.[52] The newly calculated correlation coefficient is shown only if there is a significant improvement using our processed PDB files. On the other hand, Tables 4.5: through 4.9: list all correlation coefficients for GNM from our own tests using our processed PDB files. These correlation coefficients are typically the same as those reported by Park et al.[52] although some have changed. The use of our processed PDB files leads to a slight increase in the average scores for the GNM in our analysis.

Table 4.1: Average correlation coefficients for C_α B-factor prediction with FRI, GNM and NMA for three structure sets from Park et al.[52] and a superset of 365 structures.

PDB set	opFRI	pfFRI	GNM	NMA
Small	0.667	0.594	0.541	0.480
Medium	0.664	0.605	0.550	0.482
Large	0.636	0.591	0.529	0.494
Superset	0.673	0.626	0.565	NA

To directly compare the FRI with GNM and NMA, we calculated the correlation coefficient of C_α B-factor predictions for the three structure sets taken from Park et al.[52] To further compare the FRI and GNM, we also calculated the accuracy of these two methods on a superset of 365 structures. Two versions of the FRI are used for these tests. The first, optimal FRI (opFRI), searches a wide range of parameters for the highest scoring parameter and the second, parameter free FRI (pfFRI), uses $\upsilon = 3$ and $\eta = 3$Å in all cases. The correlation coefficients for three sets proposed by Park et al. are reported in Tables 4.2:, 4.3: and 4.4: for FRI, GNM and NMA. The results of the B-factor predictions for the superset are shown in Fig. 4.14:. Using the top left chart as an example, both axes are correlation coefficients. For each circle, its x-coordinate is its correlation coefficient for pfFRI, while its

y-coordinate is its correlation coefficient for opFRI. Since all circles are located above the diagonal line, opFRI always outperform pfFRI. The average correlation scores for optimal FRI, parameter free FRI, GNM and NMA for each set of structures are listed in Table 4.1:. As shown in Table 4.1: and Fig. 4.14:, opFRI outperforms pfFRI in many cases although the majority of structures have little difference in their score for each method. Both optimal and parameter free FRI methods outperform GNM and NMA for most structures. B-factor prediction with the FRI is most accurate for smaller structures (¡70 residues). All three methods tend to perform worse as the structures get larger except in the case of NMA where the medium-sized structures scored slightly lower than the large-sized structures. This behavior is expected because as proteins get larger their structures become more complex and may include structural co-factors and more amino acid side chain interactions that contribute to the protein's stability. The coarse-grained C_α representation used in these methods is unable to capture these kinds of details. The average increase in correlation coefficients when using the FRI over GNM on the superset of 365 proteins is 0.096 for opFRI and 0.059 for pfFRI. Additionally, opFRI and pfFRI are more accurate on average than GNM and NMA for all three sets of structures used by Park et al.[52] From these results we conclude that both FRI and pfFRI are more accurate on average than either GNM or NMA.

4.2.2.2 fFRI vs GNM

Table 4.10: lists the average correlation coefficients of B-factor prediction for 365 proteins using fFRI schemes at a given truncation ($R = 12$Å). It is seen that the proposed fFRI schemes implemented in either exponential ($\eta = 3$Å and $\kappa = 1$) or Lorentz ($\eta = 3$Å and $\upsilon = 3$) are at least 10% more accurate than the GNM.

4.3 Multikernel multiscale FRI method

In this section, we implement and validate the proposed mFRI for B-factor prediction. An immediate concern is the accuracy of multi-kernel FRI method which is tested by the B-

Table 4.2: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for small-size structures. †GNM and NMA values are taken from the coarse-grained (Cα) GNM and NMA results reported in Park et al.[52] except where starred (*). Starred values indicate correlation coefficients, from our own test of GNM, that have significantly increased compared to the values reported by Park et al.[52]

PDB ID	N	opFRI	pfFRI	GNM †	NMA †
1AIE	31	0.588	0.416	0.155	0.712
1AKG	16	0.373	0.350	0.185	-0.229
1BX7	51	0.726	0.623	0.706	0.868
1ETL	12	0.710	0.609	0.628	0.355
1ETM	12	0.544	0.393	0.432	0.027
1ETN	12	0.089	0.023	-0.274	-0.537
1FF4	65	0.718	0.613	0.674	0.555
1GK7	39	0.845	0.773	0.821	0.822
1GVD	52	0.781	0.732	0.591	0.570
1HJE	13	0.811	0.686	0.616	0.562
1KYC	15	0.796	0.763	0.754	0.784
1NOT	13	0.746	0.622	0.523	0.567
1O06	20	0.910	0.874	0.844	0.900
1OB4	16	0.776	0.763	0.750*	0.930
1OB7	16	0.737	0.545	0.652*	0.952
1P9I	29	0.754	0.742	0.625	0.603
1PEF	18	0.888	0.826	0.808	0.888
1PEN	16	0.516	0.465	0.270	0.056
1Q9B	43	0.746	0.726	0.656	0.646
1RJU	36	0.517	0.447	0.431	0.235
1U06	55	0.474	0.429	0.434	0.377
1UOY	64	0.713	0.653	0.671	0.628
1USE	40	0.438	0.146	-0.142	-0.399
1VRZ	21	0.792	0.695	0.677*	-0.203
1XY2	8	0.619	0.570	0.562	0.458
1YJO	6	0.375	0.333	0.434	0.445
1YZM	46	0.842	0.834	0.901	0.939
2DSX	52	0.337	0.333	0.127	0.433
2JKU	35	0.805	0.695	0.656	0.850
2NLS	36	0.605	0.559	0.530	0.088
2OL9	6	0.909	0.904	0.689	0.886
2OLX	4	0.917	0.888	0.885	0.776
6RXN	45	0.614	0.574	0.594	0.304

Table 4.3: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for medium-size structures. †GNM and NMA values are taken from the coarse-grained (Cα) GNM and NMA results reported in Park et al.[52] except where starred (*). Starred values indicate correlation coefficients, from our own test of GNM, that have significantly increased compared to the values reported by Park et al.[52]

PDB ID	N	opFRI	pfFRI	GNM †	NMA †
1ABA	87	0.727	0.698	0.613	0.057
1CYO	88	0.751	0.702	0.741	0.774
1FK5	93	0.590	0.568	0.485	0.362
1GXU	88	0.748	0.634	0.421	0.581
1I71	83	0.549	0.516	0.549	0.380
1LR7	73	0.679	0.657	0.620	0.795
1N7E	95	0.651	0.609	0.497	0.385
1NNX	93	0.795	0.789	0.631	0.517
1NOA	113	0.622	0.604	0.615	0.485
1OPD	85	0.555	0.409	0.398	0.796
1QAU	112	0.678	0.672	0.620	0.533
1R7J	90	0.789	0.621	0.368	0.078
1UHA	83	0.726	0.665	0.638*	0.308
1ULR	87	0.639	0.594	0.495	0.223
1USM	77	0.832	0.809	0.798	0.780
1V05	96	0.629	0.599	0.632	0.389
1W2L	97	0.691	0.564	0.397	0.432
1X3O	80	0.600	0.559	0.654	0.453
1Z21	96	0.662	0.638	0.433	0.289
1ZVA	75	0.756	0.579	0.690	0.579
2BF9	36	0.606	0.554	0.680*	0.521
2BRF	100	0.795	0.764	0.710	0.535
2CE0	99	0.706	0.598	0.529	0.628
2E3H	81	0.692	0.682	0.605	0.632
2EAQ	89	0.753	0.690	0.695	0.688
2EHS	75	0.720	0.713	0.747	0.565
2FQ3	85	0.719	0.692	0.348	0.508
2IP6	87	0.654	0.578	0.572	0.826
2MCM	113	0.789	0.713	0.639	0.643
2NUH	104	0.835	0.691	0.771	0.685
2PKT	93	0.162	0.003	-0.193*	-0.165
2PLT	99	0.508	0.484	0.509*	0.187
2QJL	99	0.594	0.584	0.594	0.497
2RB8	93	0.727	0.614	0.517	0.485
3BZQ	99	0.532	0.516	0.466	0.351
5CYT	103	0.441	0.421	0.331	0.102

Table 4.4: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for large-size structures. †GNM and NMA values are taken from the coarse-grained (Cα) GNM and NMA results reported in Park et al.[52] except where starred (*). Starred values indicate correlation coefficients, from our own test of GNM, that have significantly increased compared to the values reported by Park et al.[52]

PDB ID	N	opFRI	pfFRI	GNM †	NMA †
1AHO	64	0.698	0.625	0.562	0.339
1ATG	231	0.613	0.578	0.497	0.154
1BYI	224	0.543	0.491	0.552	0.133
1CCR	111	0.580	0.512	0.351	0.530
1E5K	188	0.746	0.732	0.859	0.620
1EW4	106	0.650	0.644	0.547	0.447
1IFR	113	0.697	0.689	0.637	0.330
1NKO	122	0.619	0.535	0.368	0.322
1NLS	238	0.669	0.530	0.523*	0.385
1O08	221	0.562	0.333	0.309	0.616
1PMY	123	0.671	0.654	0.685	0.702
1PZ4	114	0.828	0.781	0.843	0.844
1QTO	122	0.543	0.520	0.334	0.725
1RRO	112	0.435	0.372	0.529	0.546
1UKU	102	0.665	0.661	0.742	0.720
1V70	105	0.622	0.492	0.162	0.285
1WBE	204	0.591	0.577	0.549	0.574
1WHI	122	0.601	0.539	0.270	0.414
1WPA	107	0.634	0.577	0.417	0.380
2AGK	233	0.705	0.694	0.512	0.514
2C71	205	0.658	0.649	0.560	0.584
2CG7	90	0.551	0.539	0.379	0.308
2CWS	227	0.647	0.640	0.696	0.524
2HQK	213	0.824	0.809	0.365	0.743
2HYK	238	0.585	0.575	0.510	0.593
2I24	113	0.593	0.498	0.494	0.441
2IMF	203	0.652	0.625	0.514	0.401
2PPN	107	0.677	0.638	0.668	0.468
2R16	176	0.582	0.495	0.618*	0.411
2V9V	135	0.555	0.548	0.528	0.594
2VIM	104	0.413	0.393	0.212	0.221
2VPA	204	0.763	0.755	0.576	0.594
2VYO	210	0.675	0.648	0.729	0.739
3SEB	238	0.801	0.712	0.826	0.720
3VUB	101	0.625	0.610	0.607	0.365

Table 4.5: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for a set of 365 proteins. GNM scores reported here are the result of our tests as described in Section 4.1.1

PDB ID	N	opFRI	pfFRI	GNM	PDB ID	N	opFRI	pfFRI	GNM
1ABA	87	0.727	0.698	0.613	1PEF	18	0.888	0.826	0.808
1AGN	1492	0.331	0.051	0.170	1PEN	16	0.516	0.465	0.270
1AHO	64	0.698	0.625	0.562	1PMY	123	0.671	0.654	0.685
1AIE	31	0.588	0.416	0.155	1PZ4	114	0.828	0.781	0.843
1AKG	16	0.373	0.350	0.185	1Q9B	43	0.746	0.726	0.656
1ATG	231	0.613	0.578	0.497	1QAU	112	0.678	0.672	0.620
1BGF	124	0.603	0.539	0.543	1QKI	3912	0.809	0.751	0.645
1BX7	51	0.726	0.623	0.706	1QTO	122	0.543	0.520	0.334
1BYI	224	0.543	0.491	0.552	1R29	122	0.650	0.631	0.556
1CCR	111	0.580	0.512	0.351	1R7J	90	0.789	0.621	0.368
1CYO	88	0.751	0.702	0.741	1RJU	36	0.517	0.447	0.431
1DF4	57	0.912	0.889	0.832	1RRO	112	0.435	0.372	0.529
1E5K	188	0.746	0.732	0.859	1SAU	114	0.742	0.671	0.596
1ES5	260	0.653	0.638	0.677	1TGR	104	0.720	0.711	0.714
1ETL	12	0.710	0.609	0.628	1TZV	141	0.837	0.820	0.841
1ETM	12	0.544	0.393	0.432	1U06	55	0.474	0.429	0.434
1ETN	12	0.089	0.023	-0.274	1U7I	267	0.778	0.762	0.691
1EW4	106	0.650	0.644	0.547	1U9C	221	0.600	0.577	0.522
1F8R	1932	0.878	0.859	0.738	1UHA	83	0.726	0.665	0.638
1FF4	65	0.718	0.613	0.674	1UKU	102	0.665	0.661	0.742
1FK5	93	0.590	0.568	0.485	1ULR	87	0.639	0.594	0.495
1GCO	1044	0.766	0.693	0.646	1UOY	64	0.713	0.653	0.671
1GK7	39	0.845	0.773	0.821	1USE	40	0.438	0.146	-0.142
1GVD	52	0.781	0.732	0.591	1USM	77	0.832	0.809	0.798
1GXU	88	0.748	0.634	0.421	1UTG	70	0.691	0.610	0.538
1H6V	2927	0.488	0.429	0.306	1V05	96	0.629	0.599	0.632
1HJE	13	0.811	0.686	0.616	1V70	105	0.622	0.492	0.162
1I71	83	0.549	0.516	0.549	1VRZ	21	0.792	0.695	0.677
1IDP	441	0.735	0.715	0.690	1W2L	97	0.691	0.564	0.397
1IFR	113	0.697	0.689	0.637	1WBE	204	0.591	0.577	0.549
1K8U	89	0.553	0.531	0.378	1WHI	122	0.601	0.539	0.270
1KMM	1499	0.749	0.744	0.558	1WLY	322	0.695	0.679	0.666
1KNG	144	0.547	0.536	0.512	1WPA	107	0.634	0.577	0.417
1KR4	110	0.635	0.612	0.466	1X3O	80	0.600	0.559	0.654
1KYC	15	0.796	0.763	0.754	1XY1	18	0.832	0.645	0.447
1LR7	73	0.679	0.657	0.620	1XY2	8	0.619	0.570	0.562
1MF7	194	0.687	0.681	0.700	1Y6X	87	0.596	0.524	0.366

Table 4.6: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for a set of 365 proteins. GNM scores reported here are the result of our tests as described in Section 4.1.1

PDB ID	N	opFRI	pfFRI	GNM	PDB ID	N	opFRI	pfFRI	GNM
1N7E	95	0.651	0.609	0.497	1YJO	6	0.375	0.333	0.434
1NKD	59	0.750	0.703	0.631	1YZM	46	0.842	0.834	0.901
1NKO	122	0.619	0.535	0.368	1Z21	96	0.662	0.638	0.433
1NLS	238	0.669	0.530	0.523	1ZCE	146	0.808	0.757	0.770
1NNX	93	0.795	0.789	0.631	1ZVA	75	0.756	0.579	0.690
1NOA	113	0.622	0.604	0.615	2A50	457	0.564	0.524	0.281
1NOT	13	0.746	0.622	0.523	2AGK	233	0.705	0.694	0.512
1O06	20	0.910	0.874	0.844	2AH1	939	0.684	0.593	0.521
1O08	221	0.562	0.333	0.309	2B0A	186	0.639	0.603	0.467
1OB4	16	0.776	0.763	0.750	2BCM	413	0.555	0.551	0.477
1OB7	16	0.737	0.545	0.652	2BF9	36	0.606	0.554	0.680
1OPD	85	0.555	0.409	0.398	2BRF	100	0.795	0.764	0.710
1P9I	29	0.754	0.742	0.625	2C71	205	0.658	0.649	0.560
2CE0	99	0.706	0.598	0.529	2OLX	4	0.917	0.888	0.885
2CG7	90	0.551	0.539	0.379	2PKT	93	0.162	0.003	-0.193
2COV	534	0.846	0.823	0.812	2PLT	99	0.508	0.484	0.509
2CWS	227	0.647	0.640	0.696	2PMR	76	0.693	0.682	0.619
2D5W	1214	0.689	0.682	0.681	2POF	440	0.682	0.651	0.589
2DKO	253	0.816	0.812	0.690	2PPN	107	0.677	0.638	0.668
2DPL	565	0.596	0.538	0.658	2PSF	608	0.526	0.500	0.565
2DSX	52	0.337	0.333	0.127	2PTH	193	0.822	0.784	0.767
2E10	439	0.798	0.796	0.692	2Q4N	153	0.711	0.667	0.740
2E3H	81	0.692	0.682	0.605	2Q52	412	0.756	0.748	0.621
2EAQ	89	0.753	0.690	0.695	2QJL	99	0.594	0.584	0.594
2EHP	248	0.804	0.804	0.773	2R16	176	0.582	0.495	0.618
2EHS	75	0.720	0.713	0.747	2R6Q	138	0.603	0.540	0.529
2ERW	53	0.461	0.253	0.199	2RB8	93	0.727	0.614	0.517
2ETX	389	0.580	0.556	0.632	2RE2	238	0.652	0.613	0.673
2FB6	116	0.791	0.786	0.740	2RFR	154	0.693	0.671	0.753
2FG1	157	0.620	0.617	0.584	2V9V	135	0.555	0.548	0.528
2FN9	560	0.607	0.595	0.611	2VE8	515	0.744	0.643	0.616
2FQ3	85	0.719	0.692	0.348	2VH7	94	0.775	0.726	0.596
2G69	99	0.622	0.590	0.436	2VIM	104	0.413	0.393	0.212
2G7O	68	0.785	0.784	0.660	2VPA	204	0.763	0.755	0.576
2G7S	190	0.670	0.644	0.649	2VQ4	106	0.680	0.679	0.555
2GKG	122	0.688	0.646	0.711	2VY8	149	0.770	0.724	0.533
2GOM	121	0.586	0.584	0.491	2VYO	210	0.675	0.648	0.729

Table 4.7: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for a set of 365 proteins. GNM scores reported here are the result of our tests as described in Section 4.1.1

PDB ID	N	opFRI	pfFRI	GNM	PDB ID	N	opFRI	pfFRI	GNM
2GXG	140	0.847	0.780	0.520	2W1V	548	0.680	0.680	0.571
2GZQ	191	0.505	0.382	0.369	2W2A	350	0.706	0.638	0.589
2HQK	213	0.824	0.809	0.365	2W6A	117	0.823	0.748	0.647
2HYK	238	0.585	0.575	0.510	2WJ5	96	0.484	0.440	0.357
2I24	113	0.593	0.498	0.494	2WUJ	100	0.739	0.598	0.598
2I49	398	0.714	0.683	0.601	2WW7	150	0.499	0.471	0.356
2IBL	108	0.629	0.625	0.352	2WWE	111	0.692	0.582	0.628
2IGD	61	0.585	0.481	0.386	2X1Q	240	0.534	0.478	0.443
2IMF	203	0.652	0.625	0.514	2X25	168	0.632	0.598	0.403
2IP6	87	0.654	0.578	0.572	2X3M	166	0.744	0.717	0.655
2IVY	88	0.544	0.483	0.271	2X5Y	171	0.718	0.705	0.694
2J32	244	0.863	0.848	0.855	2X9Z	262	0.583	0.578	0.574
2J9W	200	0.716	0.705	0.662	2XHF	310	0.606	0.591	0.569
2JKU	35	0.805	0.695	0.656	2Y0T	101	0.778	0.774	0.798
2JLI	100	0.779	0.613	0.622	2Y72	170	0.780	0.754	0.766
2JLJ	115	0.741	0.720	0.527	2Y7L	319	0.928	0.797	0.747
2MCM	113	0.789	0.713	0.639	2Y9F	149	0.771	0.762	0.664
2NLS	36	0.605	0.559	0.530	2YLB	400	0.807	0.807	0.675
2NR7	194	0.803	0.785	0.727	2YNY	315	0.813	0.804	0.706
2NUH	104	0.835	0.691	0.771	2ZCM	357	0.458	0.422	0.420
2O6X	306	0.814	0.799	0.651	2ZU1	360	0.689	0.672	0.653
2OA2	132	0.571	0.456	0.458	3A0M	148	0.807	0.712	0.392
2OCT	192	0.567	0.550	0.540	3A7L	128	0.713	0.663	0.756
2OHW	256	0.614	0.539	0.475	3AMC	614	0.675	0.669	0.581
2OKT	342	0.433	0.411	0.336	3AUB	116	0.614	0.608	0.637
2OL9	6	0.909	0.904	0.689	3B5O	230	0.644	0.629	0.601
3BA1	312	0.661	0.624	0.621	3MD4	12	0.860	0.781	0.914
3BED	261	0.845	0.820	0.684	3MD5	12	0.649	0.413	-0.218
3BQX	139	0.634	0.481	0.297	3MEA	166	0.669	0.669	0.600
3BZQ	99	0.532	0.516	0.466	3MGN	348	0.205	0.119	0.193
3BZZ	100	0.485	0.450	0.600	3MRE	383	0.661	0.641	0.567
3DRF	547	0.559	0.549	0.488	3N11	325	0.614	0.583	0.517
3DWV	325	0.707	0.661	0.547	3NE0	208	0.706	0.645	0.659
3E5T	228	0.502	0.489	0.296	3NGG	94	0.696	0.689	0.719
3E7R	40	0.706	0.687	0.642	3NPV	495	0.702	0.653	0.677
3EUR	140	0.431	0.427	0.577	3NVG	6	0.721	0.617	0.597

Table 4.8: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for a set of 365 proteins. GNM scores reported here are the result of our tests as described in Section 4.1.1

PDB ID	N	opFRI	pfFRI	GNM	PDB ID	N	opFRI	pfFRI	GNM
3F2Z	149	0.824	0.792	0.740	3NZL	73	0.627	0.583	0.506
3F7E	254	0.812	0.803	0.811	3O0P	194	0.727	0.706	0.734
3FCN	158	0.640	0.606	0.632	3O5P	128	0.734	0.698	0.630
3FE7	91	0.583	0.533	0.276	3OBQ	150	0.649	0.645	0.655
3FKE	250	0.525	0.476	0.435	3OQY	234	0.698	0.686	0.637
3FMY	66	0.701	0.655	0.556	3P6J	125	0.774	0.767	0.810
3FOD	48	0.532	0.440	-0.126	3PD7	188	0.770	0.723	0.589
3FSO	221	0.831	0.817	0.793	3PES	165	0.697	0.642	0.683
3FTD	240	0.722	0.713	0.634	3PID	387	0.537	0.531	0.642
3FVA	6	0.835	0.825	0.789	3PIW	154	0.758	0.744	0.717
3G1S	418	0.771	0.700	0.630	3PKV	221	0.625	0.597	0.568
3GBW	161	0.820	0.747	0.510	3PSM	94	0.876	0.790	0.745
3GHJ	116	0.732	0.511	0.196	3PTL	289	0.543	0.541	0.468
3HFO	197	0.691	0.670	0.518	3PVE	347	0.718	0.667	0.568
3HHP	1234	0.720	0.716	0.683	3PZ9	357	0.709	0.709	0.678
3HNY	156	0.793	0.723	0.758	3PZZ	12	0.945	0.922	0.950
3HP4	183	0.534	0.500	0.573	3Q2X	6	0.922	0.904	0.866
3HWU	144	0.754	0.748	0.841	3Q6L	131	0.622	0.577	0.605
3HYD	7	0.966	0.950	0.867	3QDS	284	0.780	0.745	0.568
3HZ8	192	0.617	0.502	0.475	3QPA	197	0.587	0.442	0.503
3I2V	124	0.486	0.441	0.301	3R6D	221	0.688	0.669	0.495
3I2Z	138	0.613	0.599	0.317	3R87	132	0.452	0.419	0.286
3I4O	135	0.735	0.714	0.738	3RQ9	162	0.510	0.403	0.242
3I7M	134	0.667	0.635	0.695	3RY0	128	0.616	0.606	0.470
3IHS	169	0.586	0.565	0.409	3RZY	139	0.800	0.784	0.849
3IVV	149	0.817	0.797	0.693	3S0A	119	0.562	0.524	0.526
3K6Y	227	0.586	0.535	0.301	3SD2	86	0.523	0.421	0.237
3KBE	140	0.705	0.704	0.611	3SEB	238	0.801	0.712	0.826
3KGK	190	0.784	0.775	0.680	3SED	124	0.709	0.658	0.712
3KZD	85	0.647	0.611	0.475	3SO6	150	0.675	0.666	0.630
3L41	220	0.718	0.716	0.669	3SR3	637	0.619	0.611	0.624
3LAA	169	0.827	0.647	0.659	3SUK	248	0.644	0.633	0.567
3LAX	106	0.734	0.730	0.584	3SZH	697	0.817	0.815	0.697
3LG3	833	0.658	0.614	0.589	3T0H	208	0.808	0.775	0.694
3LJI	272	0.612	0.608	0.551	3T3K	122	0.796	0.748	0.735
3LJI	272	0.612	0.608	0.551	3T3K	122	0.796	0.748	0.735
3M3P	249	0.584	0.554	0.338	3T47	141	0.592	0.527	0.447

Table 4.9: Correlation coefficients for B-factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for a set of 365 proteins. GNM scores reported here are the result of our tests as described in Section 4.1.1

PDB ID	N	opFRI	pfFRI	GNM	PDB ID	N	opFRI	pfFRI	GNM
3M8J	178	0.730	0.728	0.628	3TDN	357	0.458	0.419	0.240
3M9J	210	0.639	0.574	0.296	3TOW	152	0.578	0.556	0.571
3M9Q	176	0.591	0.510	0.471	3TUA	210	0.665	0.658	0.588
3MAB	173	0.664	0.591	0.451	3TYS	75	0.853	0.800	0.791
3U6G	248	0.635	0.632	0.526	4DT4	160	0.776	0.738	0.716
3U97	77	0.753	0.736	0.712	4EK3	287	0.680	0.680	0.674
3UCI	72	0.589	0.526	0.495	4ERY	318	0.740	0.701	0.688
3UR8	637	0.666	0.652	0.597	4ES1	95	0.648	0.625	0.551
3US6	148	0.698	0.586	0.553	4EUG	225	0.570	0.529	0.405
3V1A	48	0.531	0.487	0.583	4F01	448	0.633	0.372	0.688
3V75	285	0.604	0.596	0.491	4F3J	143	0.617	0.598	0.551
3VN0	193	0.840	0.837	0.812	4FR9	141	0.671	0.655	0.501
3VOR	182	0.602	0.557	0.484	4G14	15	0.467	0.323	0.356
3VUB	101	0.625	0.610	0.607	4G2E	151	0.760	0.755	0.758
3VVV	108	0.833	0.741	0.753	4G5X	550	0.786	0.754	0.743
3VZ9	163	0.785	0.749	0.695	4G6C	658	0.591	0.590	0.528
3W4Q	773	0.737	0.725	0.649	4G7X	194	0.688	0.587	0.624
3ZBD	213	0.651	0.516	0.632	4GA2	144	0.528	0.485	0.406
3ZIT	152	0.430	0.404	0.392	4GMQ	92	0.678	0.628	0.550
3ZRX	221	0.590	0.562	0.391	4GS3	90	0.544	0.522	0.547
3ZSL	138	0.691	0.687	0.526	4H4J	236	0.810	0.806	0.689
3ZZP	74	0.524	0.460	0.448	4H89	168	0.682	0.588	0.596
3ZZY	226	0.746	0.709	0.728	4HDE	168	0.745	0.728	0.615
4A02	166	0.618	0.516	0.303	4HJP	281	0.703	0.649	0.510
4ACJ	167	0.748	0.746	0.759	4HWM	117	0.638	0.622	0.499
4AE7	186	0.724	0.717	0.717	4IL7	85	0.446	0.404	0.316
4AM1	345	0.674	0.619	0.460	4J11	357	0.620	0.562	0.401
4ANN	176	0.551	0.536	0.470	4J5O	220	0.793	0.757	0.777
4AVR	188	0.680	0.605	0.650	4J5Q	146	0.742	0.742	0.689
4AXY	54	0.700	0.623	0.720	4J78	305	0.658	0.648	0.608
4B6G	558	0.765	0.756	0.669	4JG2	185	0.746	0.736	0.543
4B9G	292	0.844	0.816	0.763	4JVU	207	0.723	0.697	0.553
4DD5	387	0.615	0.596	0.351	4JYP	534	0.688	0.682	0.538
4DKN	423	0.781	0.761	0.539	4KEF	133	0.580	0.530	0.324
4DND	95	0.763	0.750	0.582	5CYT	103	0.441	0.421	0.331
4DPZ	109	0.730	0.726	0.651	6RXN	45	0.614	0.574	0.594
4DQ7	328	0.690	0.683	0.376					

Table 4.10: Average correlation coefficients (CC) of B-factor prediction for a set of 365 proteins using fFRI ($R = 12$). The improvements of the fFRI over the GNM prediction (0.565) are given in parentheses.

Exponential parameters	Avg. CC	Lorentz parameters	Avg. CC
κ=0.5, η=0.5	0.615 (8.8%)	υ=2.5, η=2.0	0.622 (10.1%)
κ=1.0, η=3.0	0.623 (10.3%)	υ=3.0, η=3.0	0.626 (10.8%)
κ=1.5, η=6.0	0.619 (9.6%)	υ=3.5, η=4.0	0.623 (10.3%)

factor prediction of a set of 364 protein structures.[49] Another concern is the parameterization of mFRI and how the choice of the η parameter affects B-factor prediction for structures of different sizes. Finally, we examine whether the proposed mFRI is as computationally efficient as the original fFRI.

4.3.1 mFRI B-factor prediction

To test the accuracy of mFRI on protein structures we use a test set containing 364 protein structures. This is the same test set used in our previous FRI paper where the Protein Data Bank (PDB) identities are listed[49] and it contains test sets used in GNM studies.[52] This test set omits one structure present in previous FRI studies (PDB ID: 1AGN) due to unrealistic B-factor data.

To quantitatively assess the performance of the proposed multikernel based mFRI method, we consider the correlation coefficient

$$C_c = \frac{\eta_{i=1}^{N} \left(B_i^e - \bar{B}^e\right) \left(B_i^t - \bar{B}^t\right)}{\left[\eta_{i=1}^{N} (B_i^e - \bar{B}^e)^2 \eta_{i=1}^{N} (B_i^t - \bar{B}^t)^2\right]^{1/2}}, \tag{4.6}$$

where $\{B_i^t, i = 1, 2, \cdots, N\}$ are a set of predicted B-factors by using the proposed method and $\{B_i^e, i = 1, 2, \cdots, N\}$ are a set of experimental B-factors extracted from the PDB file. Here $\bar{B}^t$ and $\bar{B}^e$ the statistical averages of theoretical and experimental B-factors, respectively.

4.3.1.1 Multiscale correlations of macroproteins

To illustrate the multiscale behavior of flexibility analysis, we need to construct correlation functions with sharp kernel response similar to that of a Heaviside step function. To

this end, we set $\kappa = 25$ for the exponential type of correlation kernels. In this case, the one-kernel FRI method behaves like the GNM method. The best performance for one-kernel FRI is obtained at $\eta^1 = 7$Å and the associated averaged correlation coefficient for the 364 test set is 0.540, which is similar to that obtained by using GNM.[49] Obviously, the cutoff type of kernel behavior obtained at $\kappa = 25$ does not recognize any large-scale correlation beyond 7Å in macromolecules. To capture large-scale correlations, we employ the second exponential kernel with its scale ($\eta^2 > \eta^1$) varying over a range of values as shown in Table 4.11:.

N	η^2=9Å	η^2=12Å	η^2=15Å	η^2=17Å	η^2=20Å	η^2=25Å
0-99	0.055	0.083	0.100	0.102	0.097	0.083
100-199	0.061	0.093	0.101	0.100	0.099	0.093
200-299	0.051	0.087	0.097	0.097	0.095	0.087
300-399	0.069	0.108	0.115	0.119	0.123	0.108
400-499	0.079	0.126	0.148	0.157	0.155	0.126
500+	0.064	0.107	0.136	0.143	0.140	0.107
Overall	0.060	0.094	0.106	0.108	0.106	0.094

Table 4.11: Improvements in averaged correlation coefficients for the B-factor prediction of a set of 364 proteins due to the introduction of an additional kernel parameterized at a large scale (η^2). Two exponential kernels with $\kappa = 25$ are employed. The first kernel's scale value is set to $\eta^1 = 7.0$Å in all cases. The second kernel's scale value (η^2) is varied and listed on the top of the table. Results are organized and split by the size of the structures based on the number of amino acids in order to show the impact of different η^2 values on different sizes of proteins.

To analyze the scale behavior due to protein size, we classify 364 proteins into 6 groups. The improvements of averaged correlation coefficients due to the introduction of an additional kernel are listed in Table 4.11: for a number of large scale values η^2. First, the B-factor predictions from various size classes are significantly benefited from the introduction of the large-scale kernel. Additionally, at the scale value of $\eta^2 = 17$Å, the averaged correlation coefficient is 0.648 and the associated improvement to the original FRI or GNM methods is 20% for the set of 364 proteins. Note that this multiscale improvement cannot be easily achieved by GNM, NMA, or any other mode decomposition based methods. Moreover, the

large-scale kernel leads to the most significant improvement in the B-factor prediction for relatively large proteins, proteins with 400-499 residues, which indicates that large proteins have more significant multiscale correlations than small proteins do. Finally, the improvement in the B-factor prediction for proteins with more than 500 residues is not as much as that for proteins with 400-499 residues, which indicates that two scales are not enough to capture all the multiscale correlations in proteins with more than 500 residues. This observation suggests that three kernels or multikernels are needed for the B-factor prediction of excessively large proteins.

4.3.1.2 parameterization of two-kernel based mFRI

To further understand the two-kernel based mFRI method, we consider the combination of two types of kernels. Previous tests of single kernel FRI indicate that the Lorentz type and exponential type of correlation kernels are the two most accurate single kernel types. This leads us to try the combination of these two types of kernels. The same set of 364 proteins is employed to test our method. To simplify the parameter searches, we set the κ parameter of the exponential kernel to 1.0 and set the υ parameter of the Lorentz kernel to 3.0, which are the optimal values from single kernel tests.[49] Our results are depicted in Fig. 4.15:. As expected, the addition of a second correlation kernel results in an overall increase in accuracy for B-factor predictions. For single kernel FRI, the average correlation coefficient for B-factor prediction on the set of 364 structures is 0.626. By switching to the two-kernel FRI the averaged correlation coefficient for this set increased up to 0.663. The improvement in the B-factor prediction accuracy is six percent over previous single kernel FRI methods. This averaged correlation coefficient is also better than the value of 0.640 achieved with two sharp response exponential kernels ($\kappa = 25$).

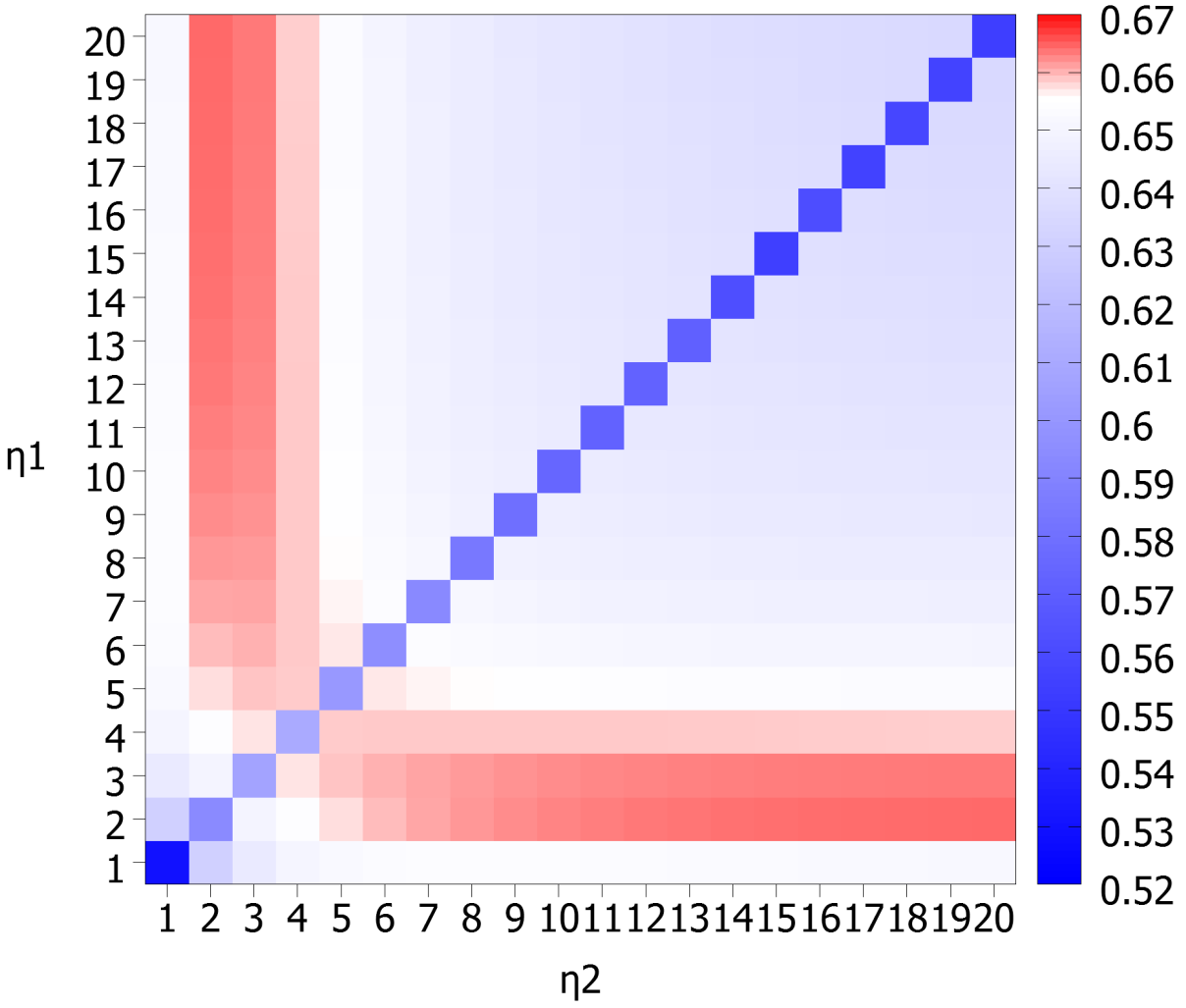


Figure 4.15: Parameter testing for a two-kernel based mFRI method. Values for η are varied for each kernel, both Lorentz kernels. Here η values for either kernel are listed along the axises. The averaged correlation coefficient for B-factor prediction on a set of 364 proteins is shown in each cell of the matrix and color coded for convenience with red representing the highest correlation coefficients and green the lowest. Obvious, the combination of a relatively small-scale kernel and a relatively large-scale kernel delivers best prediction, which shows the importance of incorporating multiscale in protein flexibility analysis.

Figure 4.15: indicates that the best results are attained either from the combination

of a relatively small-scale exponential kernel and a relatively large-scale Lorentz kernel, or from the combination of a relatively small-scale Lorentz kernel and a relatively large-scale exponential kernel. The combination of two small-scale kernels or the combination of two large-scale kernels does not offer much improvement to the original single kernel FRI method. This behavior proves again the importance of incorporating multiscale in the flexibility analysis of macromolecules.

4.3.1.3 Three Kernel based mFRI

The latest multikernel FRI method combines three kernels. After some testing we have decided upon using one kernel of exponential decay ($\kappa = 1$) and two kernels of Lorentz type ($\upsilon = 3$) with different scale (η) parameter values. The choice of kernels and parameters is driven by the idea that each kernel should capture interactions of different ranges, e.g., short-, medium- and long-range interactions each being represented by a different kernel. The exponential kernel is chosen to represent the slowest decaying forces with $\eta^3 = 15$Å and $\kappa = 1$ while the two Lorentz type of kernels capture relative short- and medium-range interactions with parameters $\upsilon = 3$, $\eta^1 = 3.0$Å and $\eta^2 = 7$Å, respectively. The associated averaged correlation coefficient for the 364 test set is 0.689, which is about 22% better than what obtained by using the GNM method.[49] Other combinations of kernel parameters were tried in which the exponential kernel exhibited the quickest decay, however, they did not perform as well in B-factor prediction tests. The fast decaying Lorentz kernel, $\eta^1 = 3$Å and $\upsilon = 3$, may be well suited to capture the effect of chemical bonds due to its particular shape of decay which highly favors interactions below 3.0 Å.

4.3.2 Computational complexity of mFRI

It has been previously been demonstrated that the computational complexity of the single kernel FRI method is asymptotically of $\mathcal{O}(N^2)$. By making use of the cell lists algorithm,

fFRI achieves a computational complexity of $\mathcal{O}(N)$. The addition of multiple kernels to the FRI method does not affect this aspect of scaling, however, the running time for B-factor prediction does increase with each additional kernel slightly. Indeed, the multi-kernel regression requires to optimize one more parameter with the addition of each new kernel. The impact of these changes on the running time of FRI-based B-factor prediction is shown in Figure 4.16:. We employ the same data sets and test conditions as those described in our earlier paper[49] for the present test. The data used for testing mFRI and fFRI are the same as those used in testing the fFRI in Table VIII of Ref.[49] In testing the GNM, the same data set as that listed in Table VIII of Ref.[49] is employed.

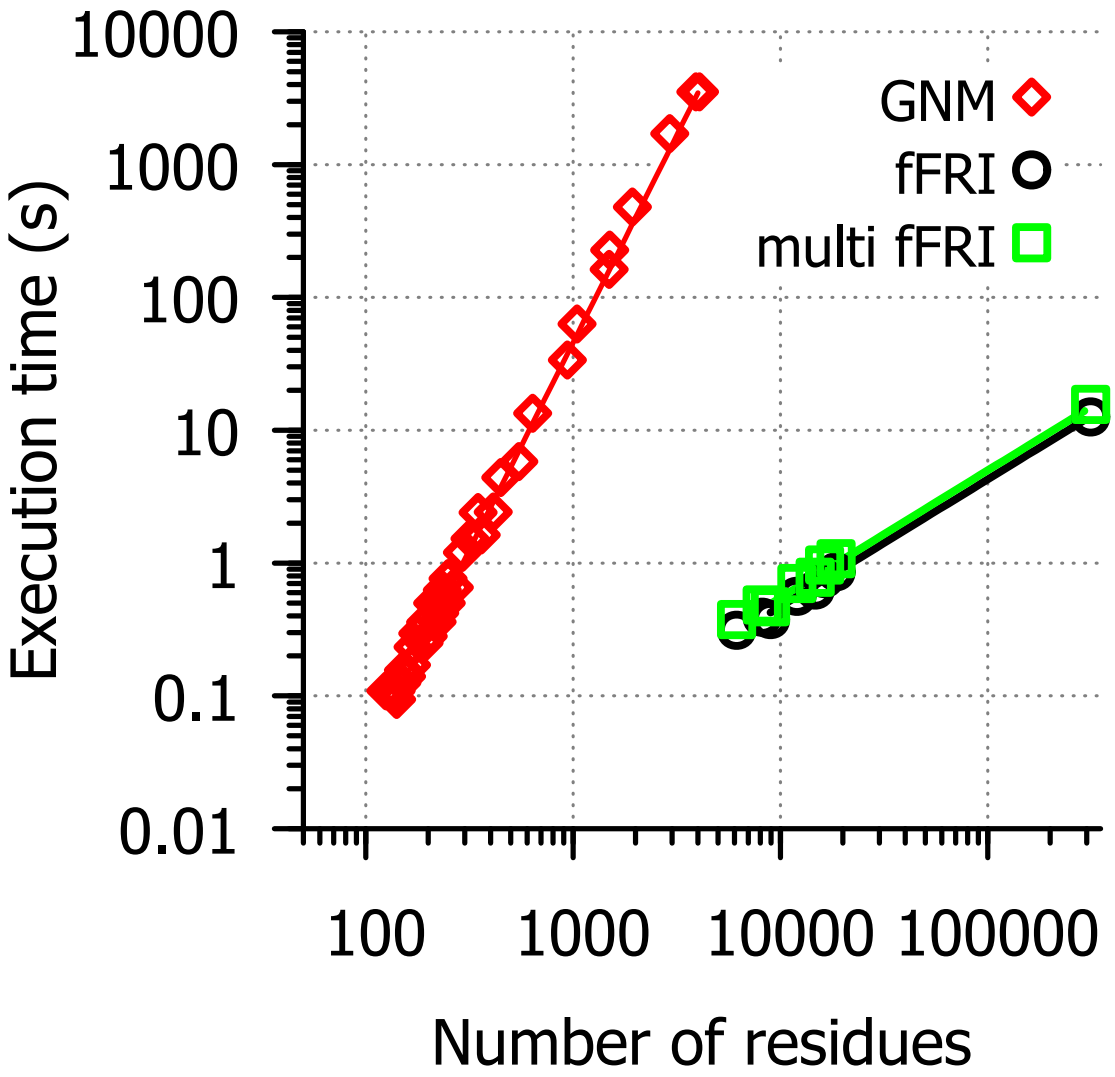


Figure 4.16: Computational efficiency of multikernel fast FRI (multi fFRI) relative to single kernel fast FRI (fFRI) and GNM. The data sets used for the present efficiency study are the same as those listed in Table VIII of Ref.[49]

Clearly the impact of extra kernels does not affect the essentially linear scaling of fFRI with lines of fit for fFRI and multikernel fast FRI (multi fFRI) being $t = 7 * 10^{-6} * N^{0.957}$ and $t = 8 * 10^{-6} * N^{0.959}$ respectively. The increase in computation time is minor especially for molecules with smaller numbers of atoms. In contrast, the line of fit for the GNM is $t = 4 * 10^{-8} * N^{3.09}$.[49] Note that each increase in one additional kernel leads to only one more fitting parameter, for which the fitting time is negligibly small. Only in extreme cases,

with systems far larger than those currently studied atomistically, might single kernel FRI be preferred. Therefore, it is preferable to use multikernel based mFRI over single kernel FRI provided there is a significant increase in accuracy and reliability, as was demonstrated previously. Note that the largest test molecule is an HIV virus capsid, which has than 313236 amino acid residues. It would take the GNM more than 120 years to finish the prediction if the computer memory is not a problem. In contrast, the proposed mFRI does the job in about 30 seconds or less on a single workstation depending on the processing power.

4.4 Multiscale FRI applications

The improvement in the averaged correlation coefficient for B-factor prediction on a set of 364 proteins discussed in the last section obscures the fact that some structures show much larger improvements than just ten percent. In this section, we highlight some examples where the improvement is up to three times more accurate than GNM and also single-kernel FRI. The proposed mFRI provides excellent B-factor predictions for many cases that the previous single-scale based theories and algorithms do not work at all.

We explore some of the advantages of using FRI for flexibility analysis by exploring applications of the multikernel FRI method. First, we explore the improvement in representing hinges in protein structures that comes from using one, two and three kernel methods. Then, we briefly highlight some other instances of flexibility prediction where the multiple kernel approach is clearly superior to to singly parameterized methods such as GNM and the original FRI formulation. In all cases, GNM is used to predict b-factors using the suggested parameterization of 7 Angstroms. An attempt was made to find a more optimal parameter for GNM for each structure in this section, at the values of 5, 6 and 8 Angstroms. If a significantly better parameterization was found, the results of b-factor prediction using that parameter are included in the figure and indicated by the name of the data series. GNM7 represents a GNM with a cutoff of 7 angstroms and other parameterizations are named sim-

ilarly with the value changed to reflect the new parameter. For example, a parameterization of 5 Angstroms would be labeled as GNM5. In all cases, the mFRI b-factor predictions have a higher correlation to the experimental data than any parameterization of GNM.

4.4.1 Fitting flexible hinge regions

Protein hinge regions have been shown to be correlated with active sites and catalysis in enzymes. Flexibility has a major role in specificity of binding of a protein to other proteins, nucleic acids or other molecules. An active site or docking region that is more flexible will accommodate more varied substrates or partners while more rigid domains are more specific. Protein hinges are also found separating large domains of proteins. In this context, the hinges can be very important for protein conformational changes. The protein featured in this section, calmodulin, is a good example of a hinge that affects both structure and function.

The central region of calmodulin shown in Figure 2.1: is a long α-helix which is unwound or kinked at the middle when no calcium is bound to the two distal metal coordinating domains. In both forms, with or without calcium bound, this helix retains a large degree of flexibility based on B-factor values from the PDB files (1CLL and 1CFD).

Many tools exist for the prediction and analysis of hinges in proteins using bioinformatics,[26] graph theory[21,39,59] and energetics.[25] The proposed mFRI has capabilities similar to those in these tools. The mFRI can be used to predict hinge regions by regions of high FRI values or predicted B-values. Many tools exist for the prediction and analysis of hinges in proteins using bioinformatics,[26] graph theory[21,39,59] and energetics.[25] The proposed mFRI has capabilities similar to those in these tools. The mFRI can be used to predict hinge regions by regions of high FRI values or predicted B-values.

A comparison of various pfFRI methods and GNM for the B-factor prediction of calcium-bound calmodulin is displayed in Figure 4.17:. B-factor prediction by single kernel FRI and

GNM is unable to accurately predict the hinge region in the middle of the protein with any parameter. Two- and three-kernel based mFRI methods, on the other hand, are much more accurate in the hinge region. As more kernels are added, the accuracy can be seen to grow but sufficient accuracy is achieved at three kernels.

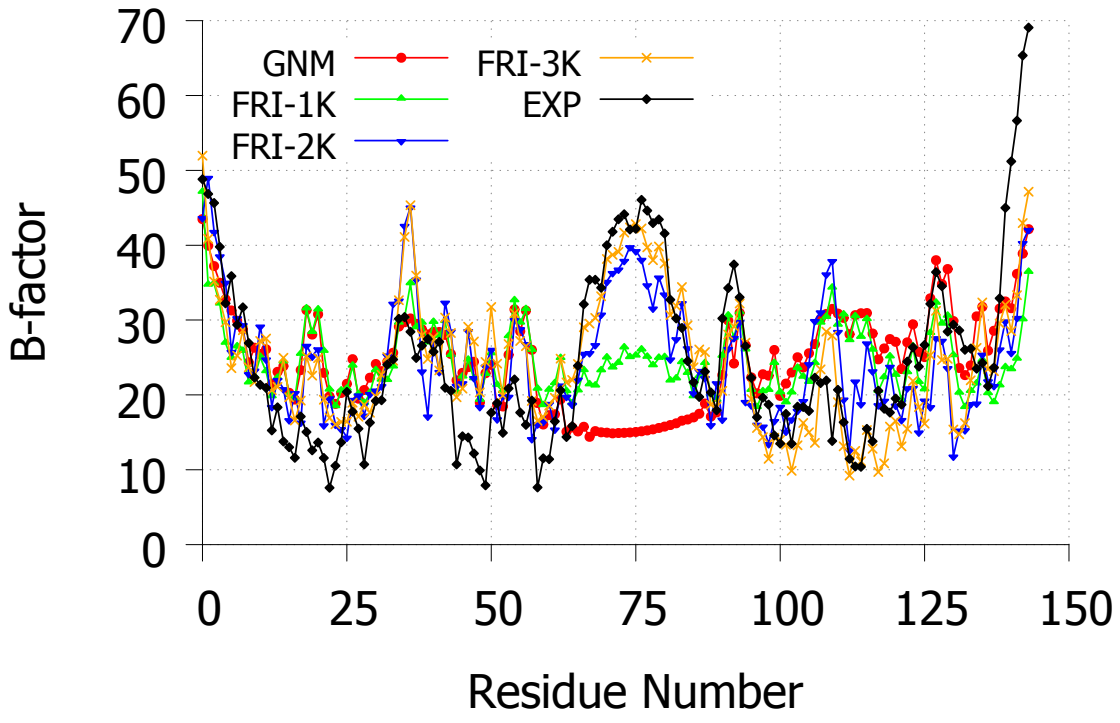


Figure 4.17: Comparison of B-factor predictions of calmodulin (PDB ID: 1CLL) using the GNM (cutoff distance is 7Å) and FRI methods. Experimental B-factors show a flexible hinge region in the middle as shown in Figure 2.1:. One-kernel FRI (FRI-1K) is parameterized at $\upsilon = 3$, $\eta = 3.0$. Two-kernel FRI (FRI-2K) is parameterized at $\kappa^1 = 1, \eta^1 = 3$Å, $\upsilon^2 = 3$, and $\eta^2 = 10$Å. Three-kernel FRI (FRI-3K) is parameterized at $\upsilon^1 = 3$, $\eta^1 = 3$Å, $\upsilon^2 = 3$, $\eta^2 = 7$Å, $\kappa^3 = 1$, and $\eta^3 = 15$Å. The three kernel based mFRI delivers the best B-factor prediction for the flexible hinge region.

4.4.2 Other proteins that benefit from mFRI

In this section we look at four specific cases to demonstrate why a multiscale approach is required to capture the complexity of interactions or correlations. In each case, we have used both three-kernel based mFRI and GNM to predict B-factors for the structures. When

GNM performed poorly, different parameters were tried to see if there is a more ideal parameterization. The results of B-factor prediction are mapped on to the residues for visual comparison and shown plotted against the experimental values for more detail.

4.4.2.1 Cyan fluorescent protein

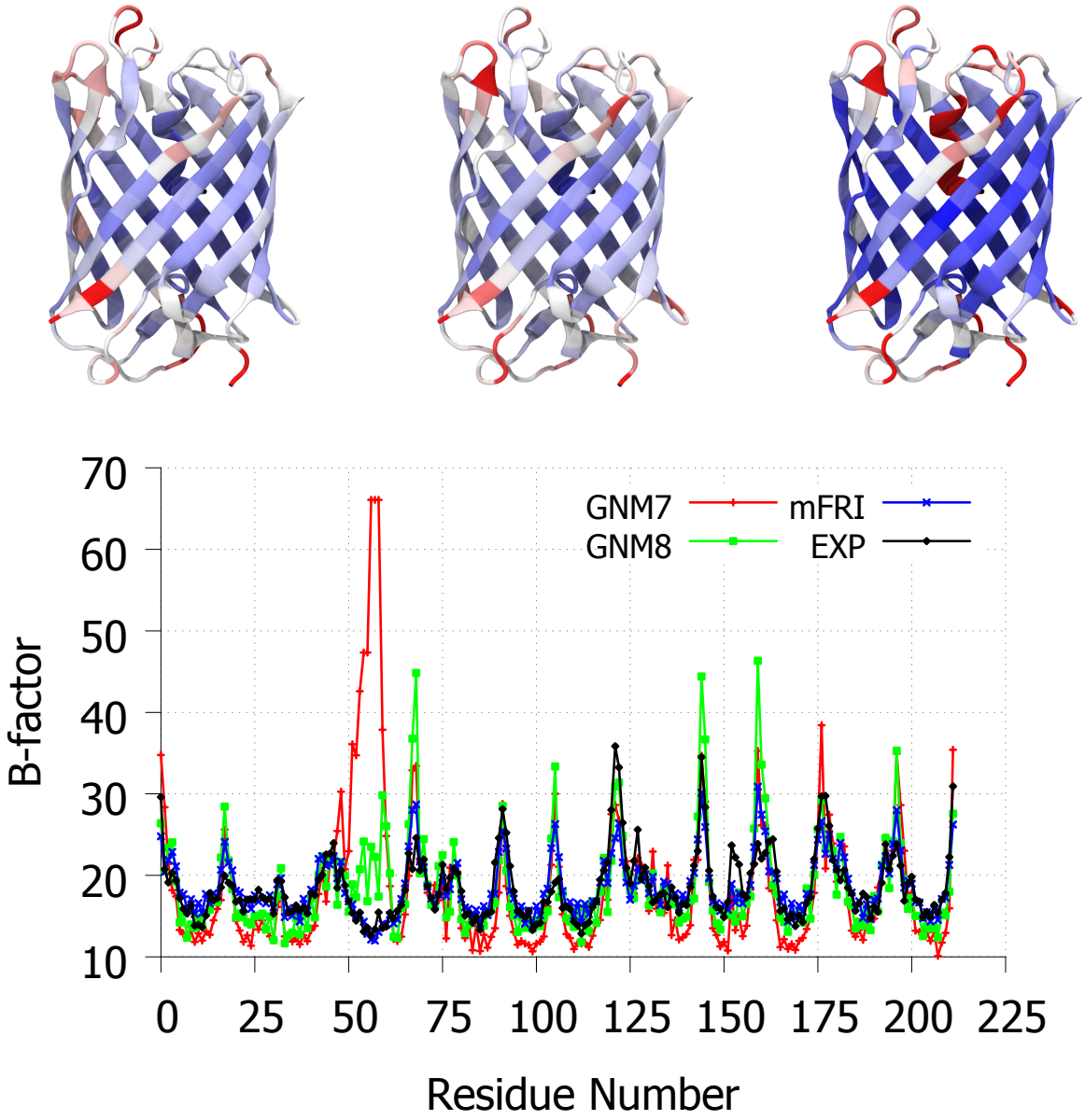


Figure 4.18: Top, a visual comparison of experimental B-factors (left), FRI predicted B-factors (midlle) and GNM predicted B-factors (right) for the engineered teal fluorescent protein, mTFP1 (PDB ID:2HQK). Bottom, The experimental and predicted B-factor values plotted per residue. The GNM naming convention indicated the cutoff used for the GNM method in angstroms, for example, GNM7 is the GNM method with a cutoff of 7Å.

Cyan fluorescent protein (CFP), shown in Figure 4.18:, is a homolog of the famous green fluorescent protein (GFP). Isolated from the crystal jellyfish in the 1990s,[60] GFP

enabled a revolution in biochemistry by allowing the tagging and tracking of a wide range of molecules. CFP was found later in Anthozoa species which have turned out to be a good source of fluorescent proteins with varied emission spectra.[47] In this example we examine the flexibility of an engineered CFP[1] (PDB ID: 2HQK), mTFP1. It is clear in Figure 4.18: that GNM B-factor predictions contain a large error around residues 50-60 which is very pronounced at the recommended cutoff of 7 Angstroms and is still somewhat problematic when the cutoff is changed to 8 Angstroms. mFRI on the other hand has no issue with this particular region. Upon further inspection, it is clear that the offending region is the small, alpha-helical region suspended in the center of the beta-barrel. It is not surprising that this sort of configuration would be highly cutoff dependent in a scheme such as GNM, which has hard cutoffs for connectivity. It would appear that this structure is dominated by short-range interactions but the region of residues 50-60 is affected to a large degree by mid-range interactions, therefore there are at least two important scales of interaction in this case. It follows then that mFRI, which has kernels to capture short- and mid-range interactions, would perform better than GNM7 or GNM8 methods alone in B-factor predictions which is exactly what we see from the results in Figure 4.18:.

4.4.2.2 Antibiotic synthesis protein from Thermus thermophilus

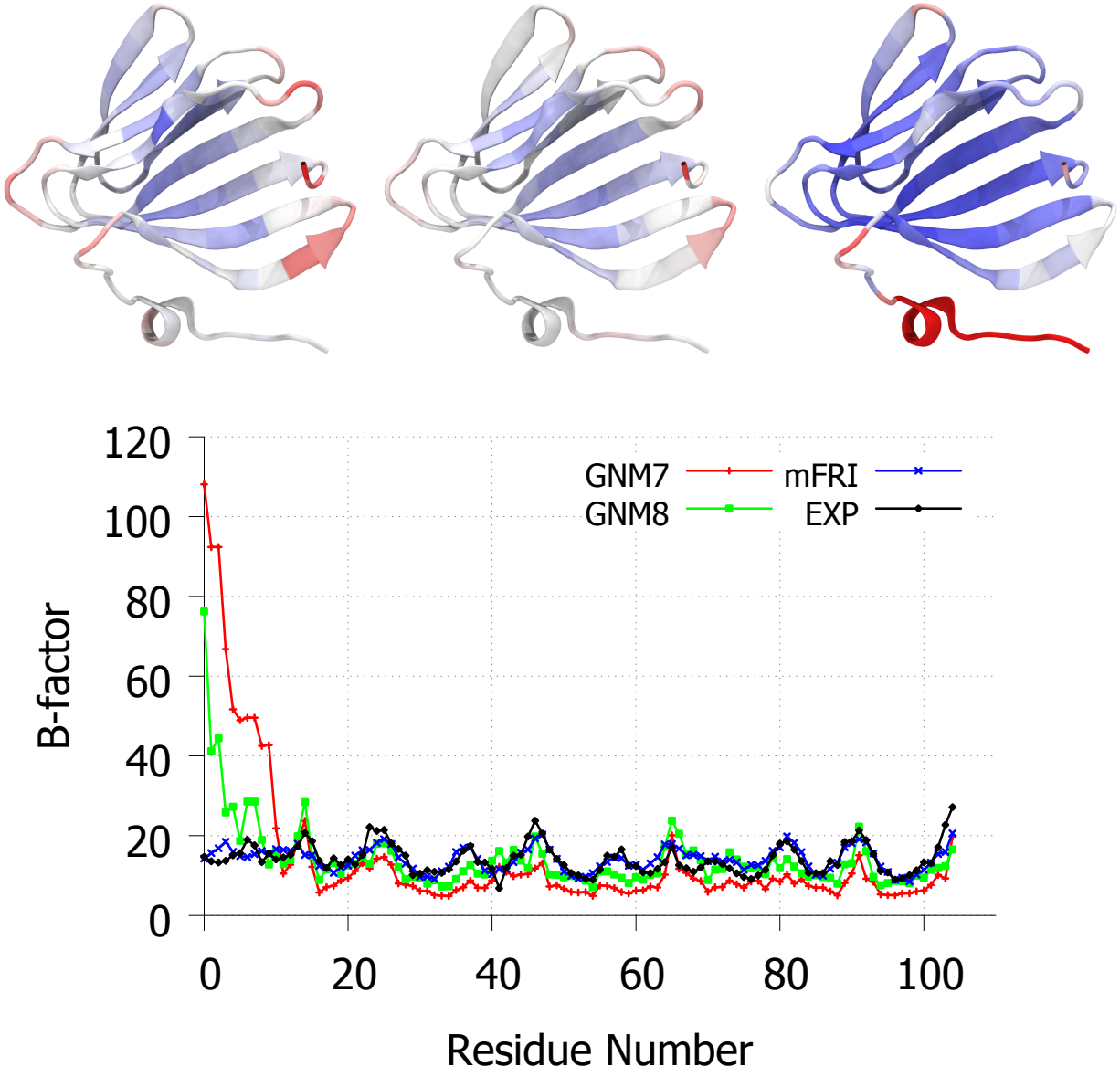


Figure 4.19: Top, a visual comparison of experimental B-factors (left), FRI predicted B-factors (midlle) and GNM predicted B-factors (right) for the engineered teal flourescent protein, mTFP1 (PDB ID:1V70). Bottom, The experimental and predicted B-factor values plotted per residue.

A similar situation exists with the structure 1V70, a probable antibiotic synthesis protein, which is shown in Figure 4.19:. As in the last example, the problematic portion for

B-factor prediction comes at the end of a protein chain. In this case there is an overestimation of flexibility for residues 1-10 when using GNM. Again, varying parameters from the recommended 7Å results in marginally better results, however no parameterization is able to reach the accuracy of mFRI.

4.4.2.3 Ribosomal subunit L14

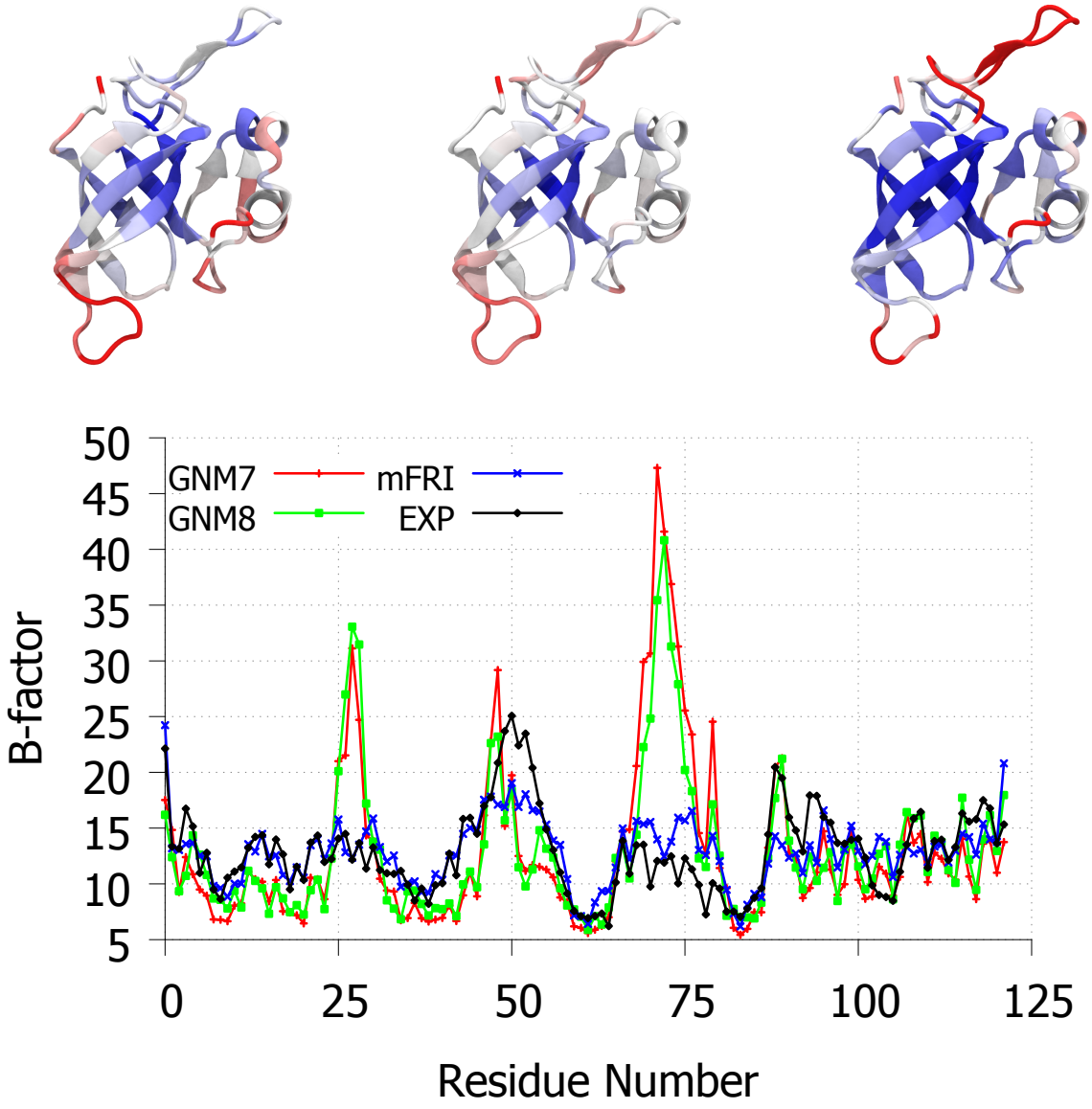


Figure 4.20: Top, a visual comparison of experimental B-factors (left), FRI predicted B-factors (midlle) and GNM predicted B-factors (right) for the ribosomal protein L14 (PDB ID:1WHI). Bottom, The experimental and predicted B-factor values plotted per residue.

The third example is a biologically important molecule, ribosomal subunit L14, a component of the 60S ribosomal subunit.[17] Depicted in Figure 4.20:, L14 is a structurally diverse protein containing regions of alpha helix, beta-barrel, parallel beta strands and a beta-hairpin

motif. The pattern of flexibility predicted by GNM for this structure is shown to be over exaggerated as the rigid areas are predicted to be more rigid than they actually are and vice verse. This pattern exists in most GNM results due to the use of a hard cutoff in the Kirchhoff matrix. Such a hard cutoff will inevitably lead to the overestimation of bond importance near the edge of the cutoff, therefore, if a large number of interactions exist for a particular atom near the cutoff point, there is likely to be a large error in the estimation of flexibility for that atom. This is likely what is happening with the errors in GNM calculation of the proteins in Figures 4.18:, 4.19: and 4.20:, the protein at the end of the chain may be near the edge of the cutoff distance for many interactions with the bulk of the proteins. While adjusting GNM's cutoff distance may temper the error being introduced, it cannot eliminate it completely unless they change to a soft decaying kernel method such as FRI.

4.4.2.4 Marine snail toxin

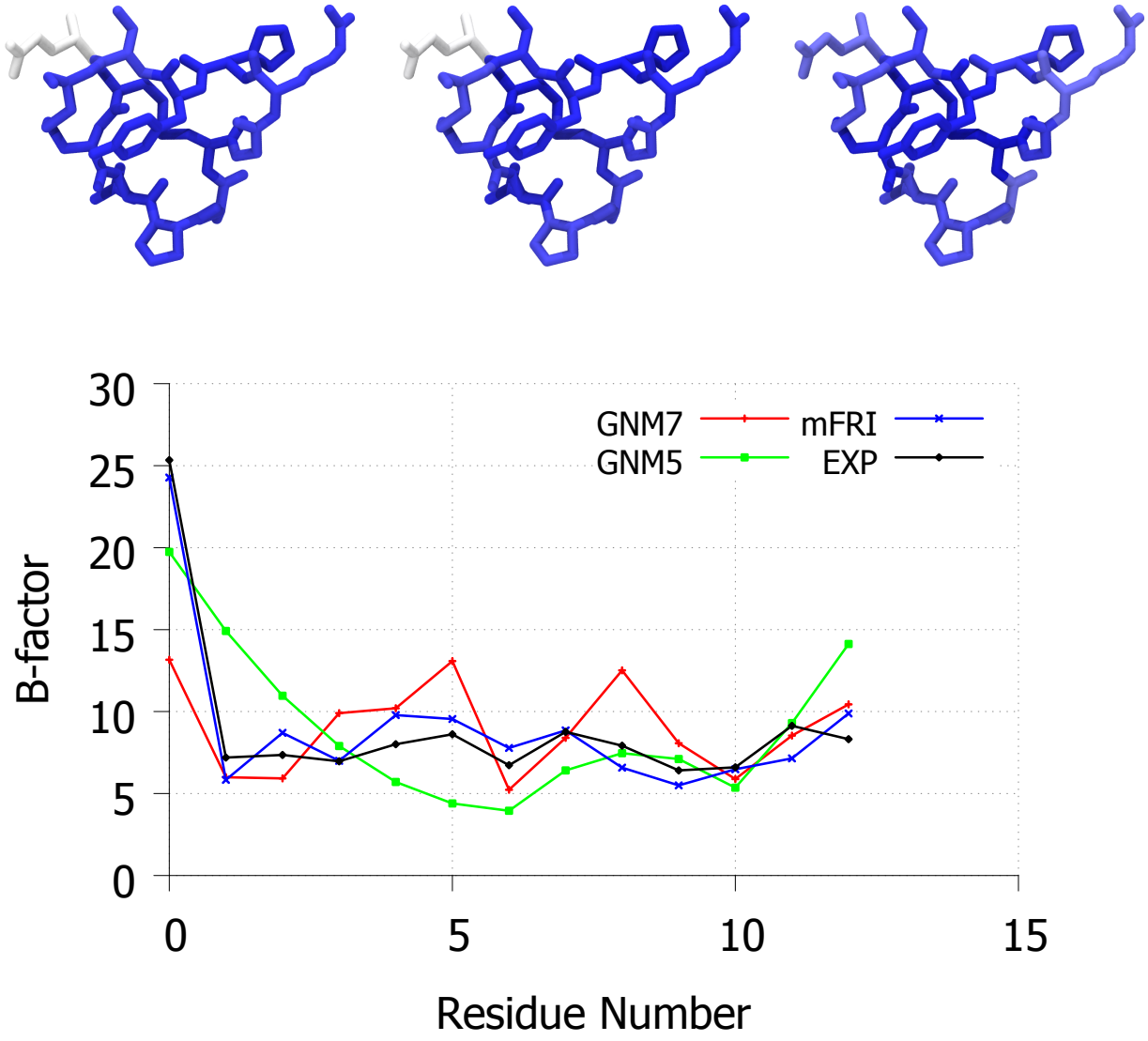


Figure 4.21: Top, a visual comparison of atomic experimental B-factors (far left), C-alpha experimental B-factors (left), FRI predicted B-factors (right) and GNM predicted B-factors (far right) for the marine snail conotoxin (PDB ID:1NOT). Bottom, The experimental and predicted B-factor values plotted per residue.

The final example is not a protein but a peptide molecule, a predatory marine snail toxin, shown in Figure 4.21:. This peptide adopts a cyclical secondary structure which is made up

of two connected loops created by two disulfide bonds. In this structure there happens to be a particular residue at the beginning of the chain which is much more flexible than the others. This is a difficult case for flexibility prediction, especially coarse-grained predictions, as there may be side-chain interactions making large contributions to the flexibility of some atoms and there are two disulfide bonds that link. Nevertheless, mFRI is able to accurately reproduce the high flexibility of the first residue. GNM on the other hand is unable to recreate the pattern of flexibility at any parameterization. This is again due to the use of a hard cutoff in the GNM method and the use of a single kernel. The differences in distances between residues in this structure are too subtle to be captured by a method that treats distance with a hard cutoff. The kernels used in FRI are sensitive enough to detect the difference in distances between atoms in this structure which leads to finding the single stand-out residue.

4.5 FRI for protein-nucleic acid complexes

In this section, we parameterize and test the previously described mFRI on protein-nucleic acid structures. A immediate concern is whether the proposed mFRI is as efficient on protein-nucleic structures as it is on protein-only structures as shown in a previous study.[74] The accuracy of the mFRI method is tested by the B-factor prediction of two sets of protein-nucleic acid structures, including a set of 64 molecules used in a recent GNM study[78] and a set of 203 molecules for more accurate parameterization of mFRI.

4.5.1 Coarse-grained representations of protein-nucleic acid complexes

In this section, we consider flexibility analysis of protein-nucleic acid complexes. To this end, we need coarse-grained representations. We consider three coarse-grained representation of nucleic acids to be used in conjugation with the Cα-only representation used for proteins. These three models are identical to those used by Yang et al.[78] and are named M1, M2 and

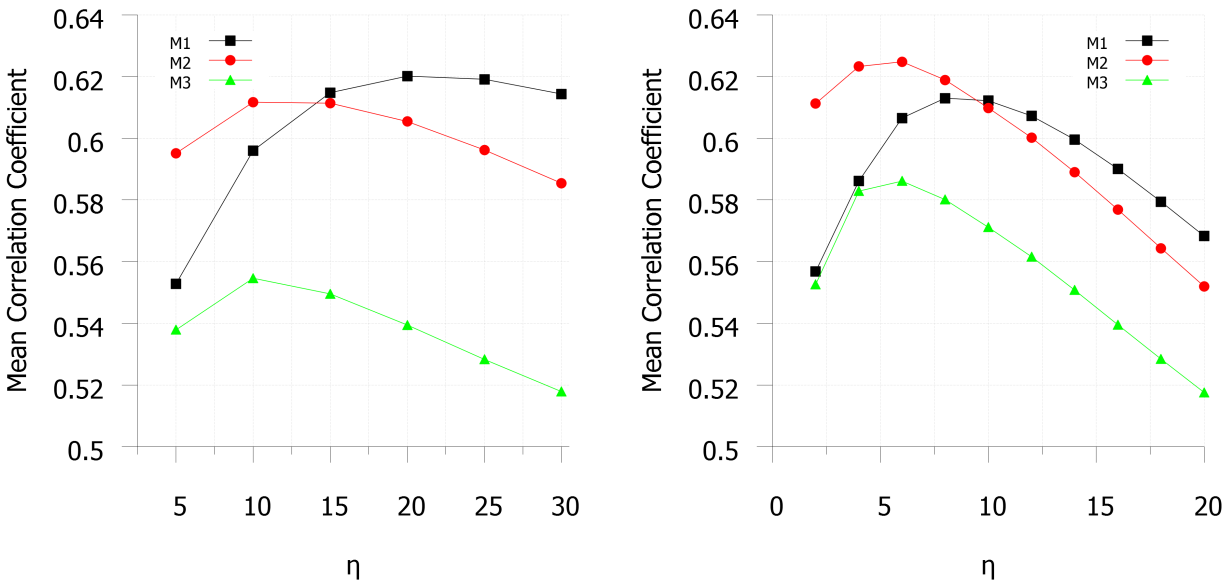


Figure 4.22: MCCs for single kernel parameter test using the M1 (squares), M2 (circles) and M3 (triangles) representations. Lorentz kernel with $\upsilon = 3$ is used. The parameter η is varied to find the maximum MCC on the test set of structures. The results for a set of 64 protein-nucleic structures (PDB IDs listed in Table 4.12:) are shown on the left, while results for a separate set of 203 structures (PDB IDs listed in Table 4.13:) is shown on the right for more general selections.

Figure 4.23: Illustration highlighting atoms used for coarse-grained representations in protein-nucleic acid complexes for FRI and GNM. In addition to protein Cα atoms, Model M1 considers the backbone P atoms for nucleotides. Model M2 includes M1 atoms and adds the sugar O4' atoms for nucleotides. Model M3 includes M1 atoms and adds the sugar C4' atoms and the base C2 atoms for nucleotides.

M3. Model M1 consists of the backbone P atoms and protein Cαatoms. Model M2 contains the same atoms as M1 but also includes sugar O4' atoms. Model M3 includes atoms from M1 and adds the sugar C4' atoms and base C2 atoms, see Fig. 4.23:.

Model M1 is similar to protein Cα representations because they are both backbone-only representations. The atoms in M1 are 6 bonds apart while Cα atoms are 3 bonds apart. Model M2 includes P atoms and adds the O4' atoms located on the ribose portion of the nucleotide. Finally, model M3 includes atoms of P, C4' and C2, a carbon from the base portion of the nucleotide, see Fig. 4.23:. As point out by Yang et al.,[78] nucleotides are approximately three times more massive than amino acids and so model M3 with three nodes per nucleotide is consistent in this sense with using Cα atoms for the protein representation.

4.5.2 mFRI B-factor precictions for protein-nucleic acid structures

To parameterize and test the accuracy of multikernel fFRI on protein-nucleic acid structures, we use a dataset from Yang et al.[78] containing 64 structures. In addition, we construct a larger database of 203 high resolution structures. This expanded protein-nucleic structure set was obtained by searching the Protein Data Bank (PDB) for structures that contain both

Protein and DNA and structure which have an X-ray resolution between 0.0 and 1.75 Å. All PDB files are processed by removing low occupancy atomic coordinates for structures having residues with multiple possible coordinates. The PDB IDs of the 64 and 203 structures can be found in Table 4.12: and Table 4.13:, respectively.

To quantitatively assess the performance of the proposed multikernel FRI method, we consider the correlation coefficient (CC)

$$\mathrm{CC} = \frac{\sum_{i=1}^{N} \left(B_i^e - \bar{B}^e\right)\left(B_i^t - \bar{B}^t\right)}{\left[\sum_{i=1}^{N}(B_i^e - \bar{B}^e)^2 \sum_{i=1}^{N}(B_i^t - \bar{B}^t)^2\right]^{1/2}}, \tag{4.7}$$

where $\{B_i^t, i = 1, 2, \cdots, N\}$ are a set of predicted B-factors by using the proposed method and $\{B_i^e, i = 1, 2, \cdots, N\}$ are a set of experimental B-factors read from the PDB file. Here $\bar{B}^t$ and $\bar{B}^e$ the statistical averages of theoretical and experimental B-factors, respectively.

4.5.2.1 Multikernel FRI testing on protein-nucleic structures

Previous tests of single kernel FRI indicate that the Lorentz type and exponential type correlation kernels are the two most accurate kernel types. This leads us to try the combination of these two types of kernels. The resulting multikernel FRI method requires four parameters, namely, κ and η for the exponential kernel and υ and η for the Lorentz kernel.

4.5.2.2 Single kernel FRI testing

In order to compare FRI and GNM methods for protein-nucleic acid structures, we test our single kernel FRI at a range of η values. For this test we use the Lorentz kernel with $\upsilon = 3$ for B-factor prediction on both structures sets and all three representations (M1, M2 and M3). The results are shown in Figure 4.22:. For the 64 structure set, single kernel FRI has a maximum mean correlation coefficient (MCC) to experimental B-factors for M1, M2 and M3 representations of 0.620, 0.612 and 0.555. Comparatively, GNM had a MCC of approximately 0.59, 0.58 and 0.55 for M1, M2 and M3 for the same data set.[78] The maximum MCCs for FRI on the larger data set for M1, M2 and M3 are 0.613, 0.625 and 0.586, respectively. The

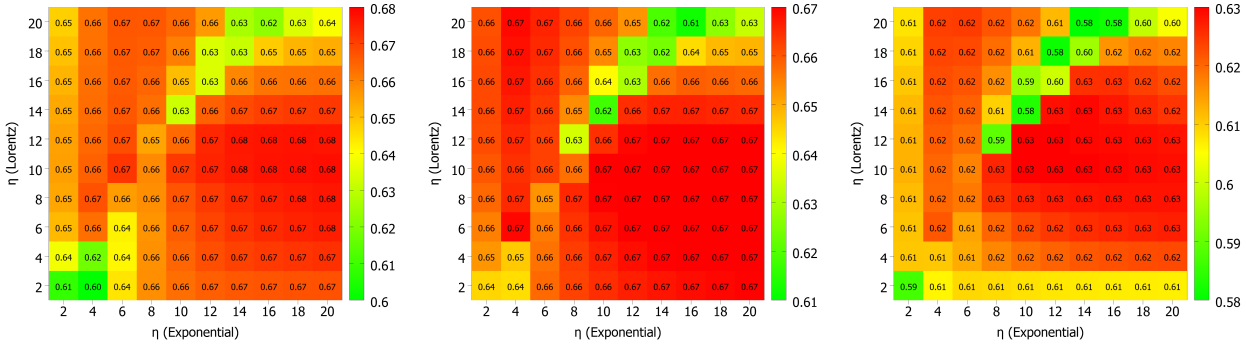


Figure 4.24: Mean correlation coefficients (MCCs) for two-kernel FRI models on a set of 203 protein-nucleic structures. From left to right, MCC values are shown for M1, M2 and M3 representations. We use one Lorentz kernel with $\upsilon = 3.0$ and one exponential kernel with $\kappa = 1.0$. The values of parameter η for both kernels are varied from 2 to 20 Å.

M1 and M2 representations perform better than the M3 representation.

4.5.2.3 Parameter-free multikernel FRI

As with protein-only structures, we develop multikernel FRIs with multiple kernels to improve accuracy of prediction on protein-nucleic acid structures. In order to simplify the FRI method, we try to develop an accurate parameter-free version for a two-kernel mFRI. We use a combination of one Lorentz and one exponential kernel. Values for parameters υ and κ are set to 3.0 and 1.0 respectively based on the results of previous FRI studies.[49] The optimal values for η in both kernels are determined by testing a range of possible values from 2 to 20 Å. All three representations (M1, M2 and M3) described previously are considered. The results of these tests on the set of 203 protein-nucleic acid structures are shown in Figure 4.24:.

As expected, the addition of another kernels results in an overall increase in accuracy for the 203 complex set. For two-kernel mFRI, the MCCs increase up to 0.68 for M1, 0.67 for M2 and 0.63 for M3. The choice of η turns out to be very robust based on results shown in Figure 4.24:.

Table 4.12: Correlation coefficients (CCs) between predicted and experimental B-factors for the set of 64 protein-nucleic structures.[78] Here N1, N2 and N3 values represent the number of atoms used for the M1, M2 or M3 representations for each structure. We use the parameter-free two-kernel mFRI model with one exponential kernel ($\kappa = 1$ and $\eta = 18$ Å) and one Lorentz kernels ($\upsilon = 3$, $\eta = 18$ Å. PDB IDs marked with an asterisk (*) indicate structure containing only nucleic-acid residues.

	M1		**M2**		**M3**	
PDB ID	CC	N1	CC	N2	CC	N3
1asy	0.647	1114	0.645	1248	0.631	1382
1b23	0.751	471	0.774	537	0.714	603
1c0a	0.763	653	0.704	721	0.598	789
1CX0	0.821	162	0.763	234	0.627	306
1drz	0.846	162	0.754	234	0.585	306
1efw	0.537	1286	0.647	1412	0.660	1538
1egk*	0.273	104	0.298	212	0.267	320
1ehz*	0.623	62	0.706	124	0.722	186
1evv*	0.710	62	0.769	124	0.770	186
1f7u	0.577	670	0.588	734	0.603	798
1ffk	0.759	6482	0.793	9310	0.809	12138
1ffy	0.520	991	0.549	1066	0.568	1141
1fg0*	0.720	498	0.723	996	0.721	1494
1fir*	0.687	61	0.576	122	0.439	183
1fjg	0.461	3915	0.585	5428	0.600	6941
1gid*	0.649	316	0.643	632	0.583	948
1gtr	0.724	603	0.747	677	0.645	751
1h3e	0.717	507	0.724	586	0.645	663
1h4s	0.671	1011	0.704	1076	0.626	1141
1hr2*	0.599	313	0.589	628	0.585	943
1i94	0.489	3923	0.615	5437	0.652	6951
1i9v*	0.615	73	0.631	147	0.642	220
1j1u	0.730	372	0.671	446	0.456	520
1j2b	0.686	1300	0.712	1448	0.672	1596
1j5a	0.532	3158	0.548	5932	0.510	8706
1j5e	0.427	3909	0.546	5422	0.553	6935
1jj2	0.799	6567	0.839	9443	0.836	12319
1jzx	0.586	3158	0.600	5932	0.561	8706
1l8v*	0.700	312	0.688	626	0.672	940
1l9a	0.849	211	0.789	336	0.675	461
1lng	0.780	183	0.595	280	0.405	377

Table 4.11, continued: Correlation coefficients (CCs) between predicted and experimental B-factors for the set of 64 protein-nucleic structures.[78] Here N1, N2 and N3 values represent the number of atoms used for the M1, M2 or M3 representations for each structure. We use the parameter-free two-kernel mFRI model with one exponential kernel ($\kappa = 1$ and $\eta = 18$ Å) and one Lorentz kernels ($\upsilon = 3$, $\eta = 18$ Å. PDB IDs marked with an asterisk (*) indicate structure containing only nucleic-acid residues.

	M1		**M2**		**M3**	
PDB ID	CC	N1	CC	N2	CC	N3
1m5k	0.904	402	0.841	622	0.760	842
1m5o	0.921	405	0.872	629	0.810	853
1mfq	0.773	341	0.688	468	0.543	595
1mms	0.507	317	0.548	433	0.646	549
1n32	0.388	3916	0.494	5447	0.517	6978
1nbs*	0.547	270	0.566	540	0.573	810
1o0c	0.766	602	0.758	676	0.636	750
1qf6	0.608	710	0.578	779	0.540	848
1qrs	0.671	603	0.672	677	0.586	751
1qtq	0.620	602	0.640	676	0.596	750
1qu2	0.520	991	0.549	1066	0.568	1141
1qu3	0.579	954	0.599	1029	0.613	1104
1rc7	0.599	256	0.566	296	0.470	336
1s72	0.823	6636	0.839	9507	0.831	12378
1ser	0.748	855	0.743	917	0.657	978
1sj3	0.880	167	0.805	240	0.614	313
1tn2*	0.686	62	0.712	124	0.676	186
1tra*	0.624	62	0.670	124	0.660	186
1ttt	0.578	1401	0.564	1587	0.515	1773
1u0b	0.757	535	0.754	609	0.621	683
1u6b	0.476	312	0.490	531	0.506	750
1u9s*	0.446	155	0.432	310	0.419	465
1vby	0.877	167	0.792	240	0.587	313
1vc0	0.878	167	0.804	240	0.611	313
1vc5	0.861	164	0.840	234	0.685	304
1y0q*	0.491	230	0.484	463	0.472	696
1y26*	0.677	70	0.697	141	0.709	212
1yfg*	0.565	64	0.600	128	0.623	192
1yhq	0.835	6636	0.840	9507	0.831	12378
1yij	0.836	6636	0.851	9507	0.842	12378
2tra*	0.614	65	0.614	130	0.613	195
3tra*	0.645	64	0.615	128	0.620	192
4tra*	0.679	62	0.715	124	0.694	186

Table 4.13: The PDB IDs of the 203 high resolution protein-nucleic structures used in our single-kernel FRI parameter test. IDs marked with an asterisk indicate those containing only nucleic acids residues.

PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID	PDB ID
1A1H	1A1I	1AAY	1AZP	1BF4	1C8C	1D02	1D2I	1DC1	1DFM
1DP7	1DSZ	1EGW	1EON	1F0V	1FIU	1H6F	1I3W	1JK2	1JX4
1K3W	1K3X	1L1Z	1L3L	1L3S	1L3T	1L3V	1LLM	1MNN	1NJX
1NK0	1NK4	1OJ8	1ORN	1PFE	1QUM	1R2Z	1RFF	1RH6	1SX5
1T9I	1U4B	1VTG	1WTO	1WTQ	1WTV	1XJV	1XVK	1XVN	1XVR
1XYI	1ZS4	2ADW	2AXY	2BCQ	2BCR	2BOP	2C62	2C7P	2EA0
2ETW	2EUW	2EUX	2EUZ	2EVF	2EVG	2FMP	2GB7	2HAX	2HEO
2HHV	2IBT	2IH2	2ITL	2NQ9	2O4A	2OAA	2ODI	2P2R	2PY5
2Q10	2R1J	2VLA	2VOA	2WBS	2XHI	2Z70	2ZKD	3BIE	3BKZ
3BM3	3BS1	3D2W	3EY1	3EYI	3FC3	3FDE	3FDQ	3FSI	3FYL
3G00	3G9M	3G9O	3G9P	3GO3	3GOX	3GPU	3GQ4	3HPO	3HT3
3HTS	3I0W	3I2O	3I3M	3I49	3I8D	3IGK	3JR5	3JX7	3JXB
3JXY	3JXZ	3KDE	3KXT	3M4A	3MR3	3MXM	3NDH	3O1M	3O1P
3O1S	3O1T	3O1U	3OQG	3PV8	3PVI	3PX0	3PX4	3PX6	3PY8
3QEX	3RKQ	3RZG	3S57	3S5A	3SAU	3SJM	3TAN	3TAP	3TAQ
3TAR	3THV	3TI0	3U6E	3U6P	3V9W	3ZDA	3ZDB	3ZDC	3ZDD
4A75	4B21	4B9S	4DFK	4DQI	4DQP	4DQQ	4DS4	4DS5	4DSE
4DSF	4E0D	4ECQ	4ECV	4ECX	4ED0	4ED2	4ED7	4ED8	4EZ6
4F1H	4F2R	4F2S	4F3O	4F4K	4F8R	4FPV	4GZ1	4GZN	4HC9
4HIK	4HIM	4HLY	4HTU	4HUE	4HUF	4HUG	4IBU	4IX7	4KLG
4KLI	4KLM	4KMF							

Table 4.14: MCCs of Gaussian network model (GNM),[78] single kernel flexibility-rigidity index (FRI) and two-kernel mFRI for three coarse-grained representations (M1, M2,and M3). A set of 64 protein-nucleic acid structures[78] is used.

	GNM[78]	**FRI**	**Two-kernel mFRI**
M1	0.59	0.620	0.666
M2	0.58	0.612	0.668
M3	0.55	0.555	0.620

We have also carried out a similar test of two-kernel mFRI ($\upsilon = 3.0$ and $\kappa = 1.0$) for the set of 64 protein-nucleic acid structures. Note that this has many large complexes. The MCCs for M1, M2 and M3 models are 0.668, 0.666 and 0.620, respectively, which are similar to what we have found for the set of 203 structures. The set of 64 structures includes 19 structures composed of nucleic acids and no amino acids. The MCCs for this nucleic acid-only subset 0.608, 0.617 and 0.603 for M1, M2 and M3 models. The correlation coefficients for all 64 individual molecular complexes are listed in Table 4.12:.

To summarize the performance of Gaussian network model, single kernel FRI, and two-kernel mFRI, we list their MCCs for the 64 protein-nucleic acid structures in Table 4.14:. It can be seen that, the FRI outperforms GNM in all three representations, and two-kernel mFRI further significantly improves the accuracy of our method and achieves up to 15% improvement compared with GNM.[78] Based on our earlier test,[50] we believe that our three-kernel mFRI can deliver a better prediction.

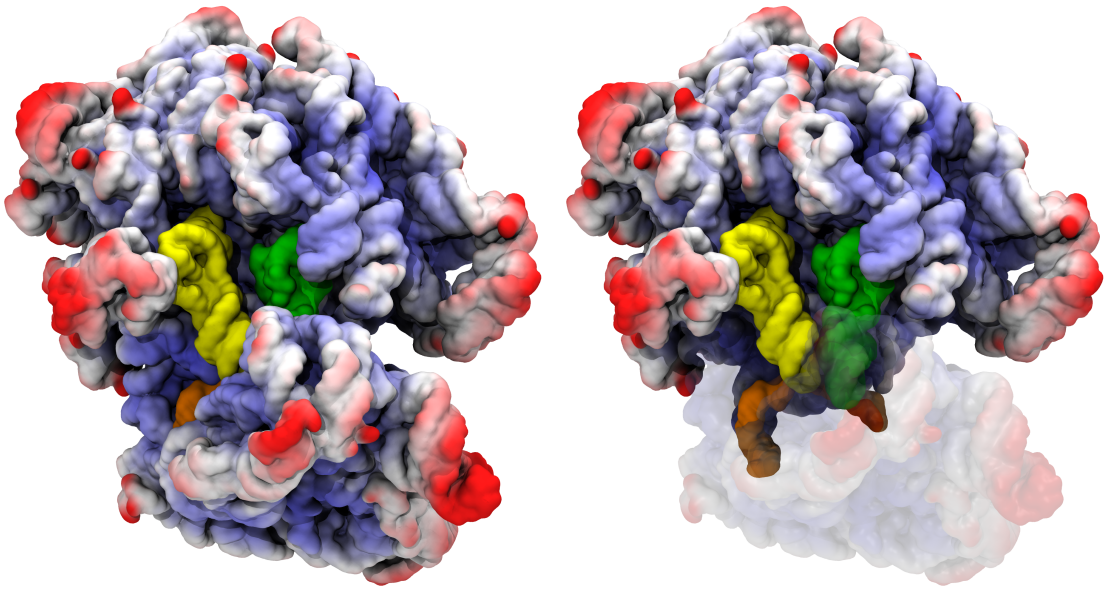
4.6 Protein-nucleic acid structure applications

In this section we briefly explore the applications of the mFRI and aFRI methods to large protein-nucleic acid complexes. We highlight a few particular examples where mFRI improves upon previous FRI methods, in particular, for the flexibility prediction of ribosomes. Further, we show how aFRI is well suited for the study of the dynamics of large macromolecular complexes using the bacterial RNA polymerase active site as an example.

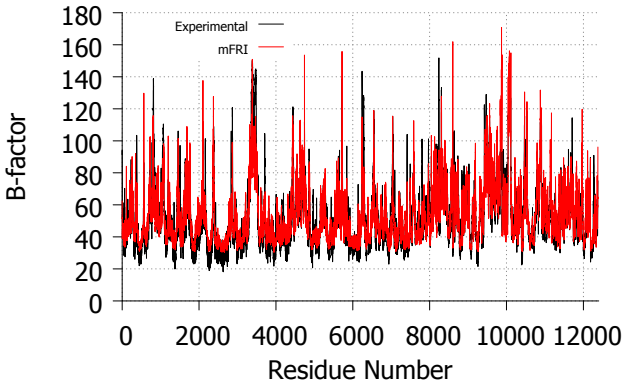
4.6.1 mFRI flexibility prediction for ribosomes

Some of the largest and most biologically important structures that contain both protein and nucleic acids are ribosomes. Ribosomes are the protein synthesizers of the cell and connect amino acid into polymer chains. In ribosomes, proteins and RNA interact through intermolecular effects, such as electrostatic interactions, hydrogen bonding, hydrophobic interactions, base stacking and base pairing. RNA tertiary structures can significantly influence protein-RNA interactions. Ribosomes are primarily composed of RNA with many smaller associated proteins as shown in Fig. 4.25:. The top of Fig. 4.25: shows the 50S subunit of the ribosome (PDB ID: 1YIJ) with the nucleic acids in a smooth surface representation with the protein subunits bound and shown in a secondary structure representation. The set of 64 structures used in our tests contains a number of ribosomal subunits. Due to their multiscale nature, these structures also happen to be among those that benefit the most from using multikernel FRI over single kernel FRI or GNM. For example, in the case of ribosome 50S subunit structure (PDB ID:1YIJ), B-factor prediction with three-kernel FRI yields a CC value of 0.85, while that of single kernel FRI is only around 0.3. GNM does not provide a good B-factor prediction for this structure either. The three-kernel mFRI model we used is one exponential kernel ($\kappa = 1$ and $\eta = 15$ Å) and two Lorentz kernels ($\upsilon = 3$, $\eta = 3$ Å and $\upsilon = 3$, $\eta = 7$Å). The comparison between mFRI-predicted and experimental B-factors for ribosome 50S subunit structure is demonstrated in Fig. 4.25:.

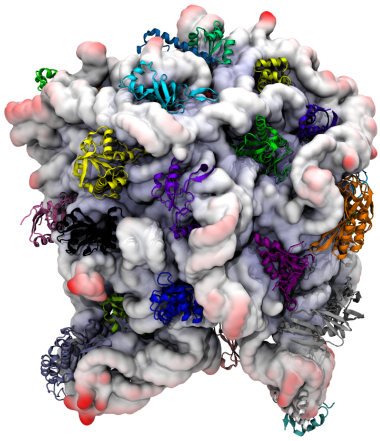
By using the fitting coefficients from the above 50S subunit (1YIJ) flexibility analysis, we have obtained flexibility predictions for the entire ribosome (PDB ID:4V4J) as well as many protein subunits and other RNAs that associate with it, see Fig. 4.25:. To avoid confusion, the B-factors for 4V4J are uniquely determined by using not only the same three-kernel mFRI model from the case 1YIJ, and also its fitting parameters $a^1 =, a^2 =, a^3$, and b. Again, the FRI values are mapped by color to the smooth surface of the nucleic acids, however,



(a) Complete ribosome with bound tRNAs PDB ID: 4V4J.



(b) Ribosome 50S subunit PDB ID: 1YIJ B factors



(c) Ribosome 50S subunit PDB ID: 1YIJ

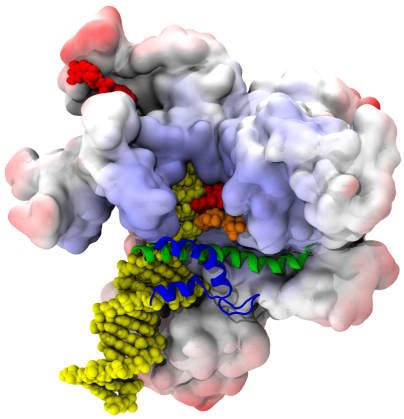
Figure 4.25: Complete ribosome with bound tRNAs (yellow (A site) and green (P site)) and mRNA Shine-Delgarno sequence (orange) PDB ID: 4V4J. The same correlation coefficients and fitting parameters from mFRI model of protein 1YIJ are used. A comparison of predicted and experimental B-factor data for Ribosome 50S subunit PDB ID: 1YIJ. The CC value is 0.85 using the parameter free three-kernel mFRI model. Nucleic acids are shown as a smooth surface colored by FRI flexibility values (red for more flexible regions) while bound protein subunits are colored randomly and shown in a secondary structure representation. We achieve a CC value up to 0.85 using parameter free three-kernel mFRI model with one exponential kernel ($\kappa = 1$ and $\eta = 15$ Å) and two Lorentz kernels ($\upsilon = 3$, $\eta = 3$ Å and $\upsilon = 3$, $\eta = 7$Å).

in these bottom figures the protein subunits are omitted to draw attention instead to the various types of RNA involved in this structure.

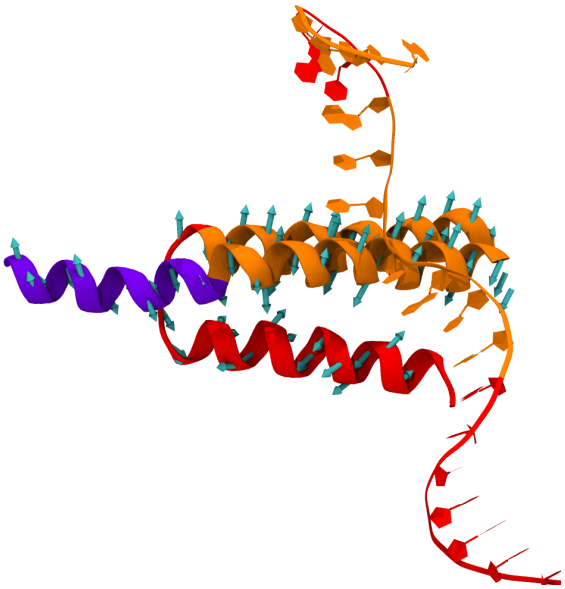
4.6.2 aFRI conformational motion prediction on an RNA polymerase structure

RNA polymerase is one of the essential enzymes for all life on Earth as we know it today and possibly from the very beginning of life.[13,35] Despite this importance, the mechanisms for many of the polymerase's functions are still not well understood on the atomic level. Considerable effort has been spent both experimentally and computationally to understand RNAP polymerase function in more detail but many questions remain. The study of RNA polymerase experimentally or computationally is difficult and often expensive due to the size of the system and variety of molecules involved. The minimal required elements for a bacterial or eukaryotic RNA polymerase include multiple protein subunits, a double stranded DNA molecule, a single stranded RNA molecule, free nucleotides, various ions (Mg^{2+}, Zn^{2+}, Na^{+} etc.) and solvent. A typical setup for this system in all-atom molecular dynamics includes 300,000 atoms when solvated. With this number of atoms and current computer power, it is often not feasible to simulate these molecules on biologically relevant timescales using MD. Perhaps the most popular tool for studying long time dynamics of biomolecules is normal mode analysis (NMA) and its related methods such as the anisotropic network model (ANM). These methods have been successfully used to study protein dynamics for many proteins, however, at their maximum accuracy, their computational complexity is of $\mathcal{O}(N^3)$, where N is the number of atoms. This is a problem because many cellular functions involve a large number of macromolecules with many thousands to millions of residues to consider. Therefore, future computational studies of biomolecules beyond the protein scale will require methods with better scaling properties such as FRI and aFRI.

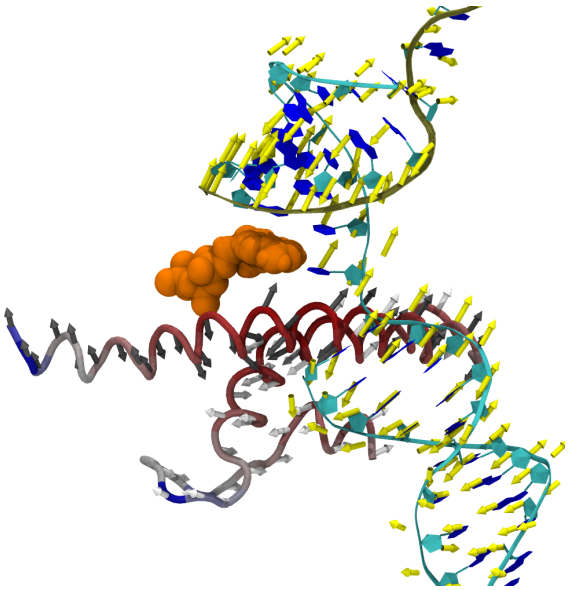
In this example, we use completely local anisotropic FRI to examine correlated motions in regions near the active site of bacterial RNA polymerase, including the bridge helix,



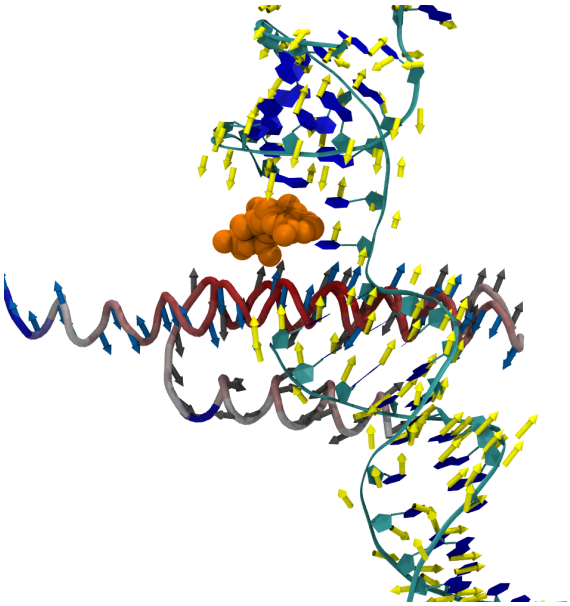
(a) RNA Polymerase with closed trigger loop



(b) Correlated motion near active site



(c) aFRI mode 1 - Open TL



(d) aFRI mode 1 - Closed TL

Figure 4.26: The first RNAP local aFRI mode for the bridge helix, trigger loop and nucleic acids from both open (PDB ID: 2PPB) and closed (PDB ID: 2O5J) configurations. Arrows represent the direction and relative magnitude of atomic fluctuations. Arrows for the bridge helix, trigger loop and nucleic acids are pictured as blue, white and yellow, respectively.

trigger loop and nucleic acid chains. We examine the relationship between these components' motions and their contributions to critical functions such as catalysis and translocation. We use the anisotropic rigidity form in section 3.5 with the Lorentz kernel ($\upsilon = 2$ and $\eta = 3$ Å). Figure 4.26:a is a simplified representation of RNA polymerase (PDB ID 2PPB) that shows these important features which are buried in the core of the largest protein subunits, β and β'. The bridge helix and trigger loop, shown in green and blue respectively, are parts of the protein that have been implicated in most of the essential functions of the polymerase. Mutational studies of these regions result in modulation of the polymerase speed and accuracy, both positively and negatively, indicating the regions are important for normal functioning of the enzyme. How these regions aid these functions and how they interact remains an open question. With this demonstration of local aFRI analysis we hope to shed some light on how these essential parts of RNA polymerase work together.

Local aFRI, as described in earlier work, is much less computationally costly than global aFRI or NMA and has been shown to have qualitatively similar results for small to large size single proteins. To further validate the local aFRI method we compare the conclusions from a local aFRI study of RNAP to those of NMA based studies. The RNA polymerase elongation complex is a relatively large system but it is still tenable for NMA methods. NMA has been applied to both bacterial and eukaryotic RNA polymerase in the past[23,69] which provides us with a point of comparison for our results.

Local aFRI produces three modes of motion sorted from lowest to highest frequency vibration according to eigenvalue as in NMA. In Figure 4.26: we present findings from the lowest frequency mode effectively focusing on the most dominant motion of each conformation. Two major conformations of RNA polymerase are considered, those with open and closed trigger loop regions (Figures 4.26:c and 4.26:d.) A closed trigger loop is one that is completely folded into two parallel alpha helices while an open trigger loop has a region of

disordered loop between two shorter helices and is slightly bent away from the bridge helix . The closing or folding of the trigger loop into the closed conformation is assumed to follow binding of an NTP in the active site and to precede catalysis. After catalysis, it is suspected that the trigger loop opens or unfolds to facilitate translocation and permit new NTPs to enter the active site.

The results of aFRI analysis on the effect of trigger loop closing reveal a distinct change in correlated motions in open and closed trigger loop conformations. These changes involve interactions between the bridge helix, the trigger loop and the nucleic acid regions. In Figure 4.26:b, regions of high correlation are color coded which reveals that the bridge helix is composed of two highly self correlated portions suggesting the presence of a hinge in the bridge helix. In fact, the central portion of the bridge helix has been observed as a kinked or bent helix in a yeast RNAP structure.[72] Additionally, it is observed that a portion of the bridge helix and the N-terminal helix of the trigger loop are highly correlated in the closed trigger loop structure only. This set of two helices is situated directly next to the active site and could provide stability to aid catalysis after trigger loop closing.

Additionally, correlation between nucleic acids and protein shows marked differences from the open trigger loop to closed trigger loop structures. The motions indicated in Figures 4.26:c and 4.26:d show that the open trigger loop structure is primed to translocate based on the direction of highly correlated motions of the upstream and downstream nucleic acids. By contrast, the closed trigger loop nucleic acid motions are considerably less correlated and not in the direction of translocation. This is the expected relationship as it matches the results from previous biological and NMA studies of RNA polymerase.[23]

These differences between a closed trigger loop and open trigger loop structure reveal potentially important structural changes that arise as the RNA polymerase switches between open and closed trigger loop conformations during the transition between translocation and

catalysis. Specifically, the results for the closed trigger loop conformation suggest the presence of a stabilized catalytic area which is made of the N-terminal helix of the trigger loop and the bridge helix. The results for the open trigger loop conformation show no such coordination of the active site helices and instead indicates a less defined hinge and coordinated motion in the direction of translocation. Taken together these results provide a potential explanation for how trigger loop opening and closing is correlated with translocation and catalysis respectively.

4.7 Generalized GNM, multiscale GNM and multiscale ANM methods

4.7.1 Generalized Gaussian network model

4.7.1.1 Comparison between gGNM and FRI

Based on the analysis in Section 3.6, it is straightforward to construct correlation function-based gGNMs via the matrix inverse of the generalized Kirchhoff matrix (3.46), which leads to infinitely many new gGNMs including the original GNM as a special limiting case. Also, it is possible to construct a FRI method using the Kirchhoff matrix of GNM. In light of these observations it is necessary to directly compare the performance of the related methods and to explore whether there is any further relationship between these two approaches, specifically the diagonal elements of the gGNM matrix inverse and the direct inverse of the diagonal elements of a generalized Kirchhoff matrix. To address this question, we select two representative correlation functions, the Lorentz ($\upsilon = 3$) and ILF functions, to construct the generalized Kirchhoff matrix (3.46). The Lorentz function is a frequently used correlation function in our earlier work.[49] In contrast, the ILF function, while typical in GNM, is an extreme case of FRI correlation function not previously considered in work on FRI. The resulting two generalized Kirchhoff matrices (3.46) can be used for calculating the gGNM matrix inverse or the inverse diagonal elements of the FRI matrix. This results in possible combinations of methods, namely, FRI-Lorentz, FRI-ILF, GNM-Lorentz and GNM-ILF.

To test the methods described above, we analyze the flexibility of a protein from pathogenic fungus Candida albicans (Protein Data Bank ID: 2Y7L) with 319 residues as shown in Fig. 4.27:(a). We consider the coarse-grained C_α representation of protein 2Y7L. We denote $B^{\text{GNM}-\text{ILF}}$, $B^{\text{FRI}-\text{ILF}}$, $B^{\text{GNM}-\text{Lorentz}}$ and $B^{\text{FRI}-\text{Lorentz}}$ respectively the predicted B-factors of GNM-ILF, FRI-ILF, GNM-Lorentz and FRI-Lorentz methods. The experimental B-factors from X-ray diffraction, B^{Exp}, are also displayed for comparison. The Pearson product-moment correlation coefficient (PCC) is used to measure the strength of the linear relationship or dependence between each sets of predicted or experimental B-factors. Since performance of these methods depends on their parameters, the cutoff distance (r_c) in the ILF and the scale value (η) in the Lorentz function, the theoretical B-factors are computed over a wide range of r_c and η values to find the parameters that work best for each method.

Figure 4.28: depicts PCCs between various sets of B-factors for protein 2Y7L. As shown in Fig. 4.28: (a), the cutoff distance r_c of the ILF is varied from 5Å to 64Å. The PCCs between $B^{\text{GNM}-\text{ILF}}$ and B^{Exp}, and between $B^{\text{FRI}-\text{ILF}}$ and B^{Exp}, indicate that both GNM-ILF and FRI-ILF are able to provide accurate predictions of flexibility compared the experimental B-factors. The best predictions are attained around $r_c = 24$Å, which is significantly larger than the commonly used GNM cutoff distance of 7-9Å.

4.7.1.2 Intrinsic behavior of gGNM at large cutoff distance

It is interesting to observe that GNM-ILF and FRI-ILF provide essentially identical predictions when the cutoff distance is equal to or larger than 20Å. This phenomenon indicates that when the cutoff is sufficiently large, the diagonal elements of the gGNM inverse matrix and the direct inverse of the diagonal elements of the FRI correlation matrix become linearly dependent. To examine the relation between GNM-ILF and FRI-ILF, we compute PCCs between $B^{\text{GNM}-\text{ILF}}$ and $B^{\text{FRI}-\text{ILF}}$ over the same range of cutoff distances. As shown in Fig. 4.28:(a), there is a strong linear dependence between $B^{\text{GNM}-\text{ILF}}$ and $B^{\text{FRI}-\text{ILF}}$ for $r_c \geq 10$Å.

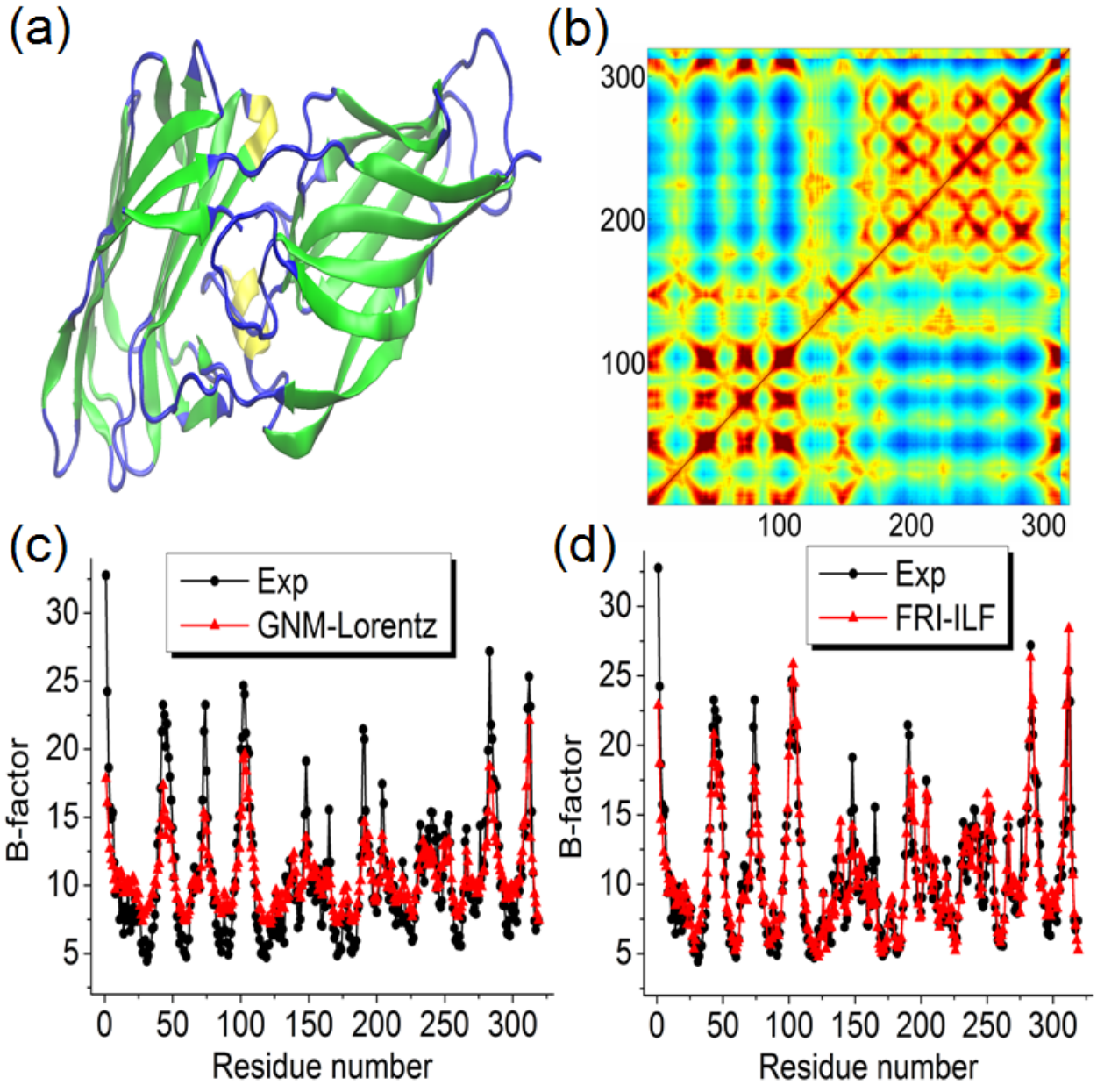


Figure 4.27: Illustration of protein 2Y7L. (a) Structure of protein 2Y7L having two domains; (b) Correlation map generated by using GNM-Lorentz indicating two domains; (c) Comparison of experimental B-factors and those predicted by GNM-Lorentz ($\eta = 16$Å); (d) Comparison of experimental B-factors and those predicted by FRI-ILF ($r_c = 24$Å).

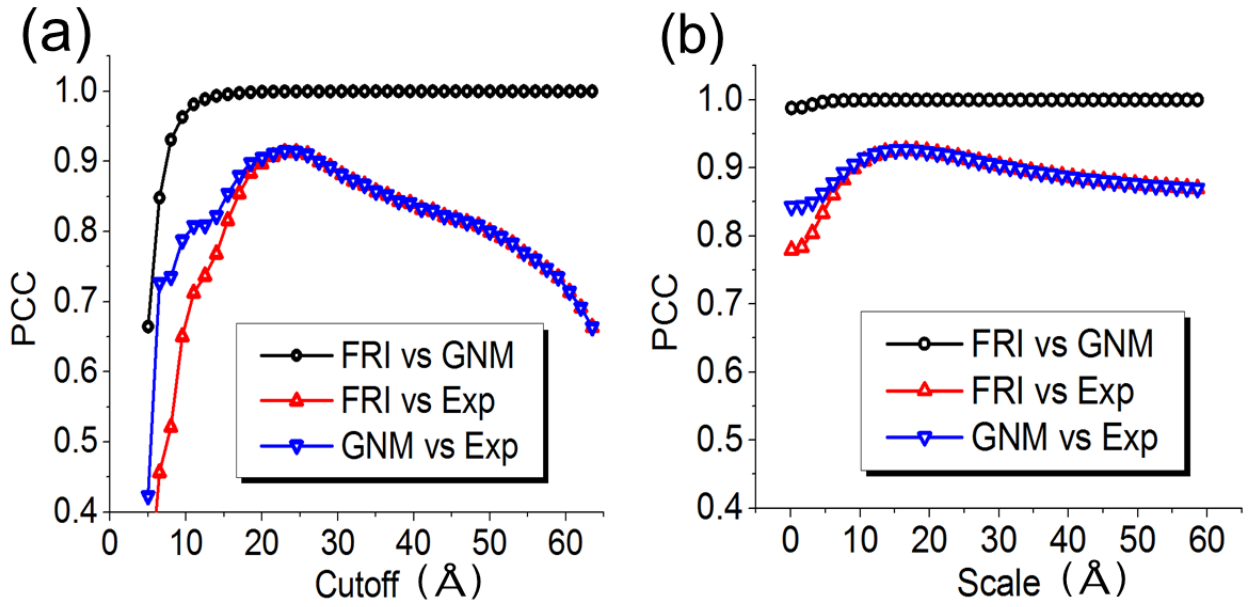


Figure 4.28: PCCs between various B-factors for protein 2Y7L. (a) Correlations between $B^{\mathrm{GNM-ILF}}$ and B^{Exp}, between $B^{\mathrm{FRI-ILF}}$ and B^{Exp}, and between $B^{\mathrm{GNM-ILF}}$ and $B^{\mathrm{FRI-ILF}}$; (b) Correlations between $B^{\mathrm{GNM-Lorentz}}$ and B^{Exp}, between $B^{\mathrm{FRI-Lorentz}}$ and B^{Exp}, and between $B^{\mathrm{GNM-Lorentz}}$ and $B^{\mathrm{FRI-Lorentz}}$.

To understand this dependence at large cutoff distance, we consider an extreme case when the cutoff distance is equal to or even larger than the protein size, so all the particles within the network are fully connected. In this situation, we can analytically calculate ith diagonal element of the GNM inverse matrix

$$\left(\Gamma^{-1}(\Phi(r_{ij}; r_c \to \infty))\right)_{ii} = \frac{N-1}{N^2}, \tag{4.8}$$

and the FRI inverse of the ith diagonal element

$$\frac{1}{\sum_{j,j\neq i}^{N} \Phi(r_{ij}; r_c \to \infty)} = \frac{1}{N-1}. \tag{4.9}$$

These results show a strong asymptotic correlation between $B^{\mathrm{GNM-ILF}}$ and $B^{\mathrm{FRI-ILF}}$ in Fig. 4.28:(a). They also explain why predictions of the original GNM and FRI-ILF deteriorate as r_c is sufficiently large because all the predicted B-factors become identical, either $\frac{N-1}{N^2}$ or $\frac{1}{N-1}$. And two methods deliver very similar results, especially when the total number is very

large, as we have $\frac{\frac{N-1}{N^2}}{\frac{1}{N-1}} \to 1$ when $N \to \infty$.

The performance and comparison between GNM-Lorentz and FRI-Lorentz is illustrated in Fig. 4.28:(b) where the scale parameter η ranges from 0.5Å to 64Å. It is seen from these results that the GNM-Lorentz method is a successful new approach. In fact, it outperforms the original GNM. A comparison of the predicted B-factors and the experimental B-factors is plotted in Figs. 4.27:(c) and 4.27:(d) for GNM-Lorentz and FRI-ILF, respectively. It is seen that $B^{\text{FRI}-\text{ILF}}$ more closely matches the experimental B-factors than $B^{\text{GNM}-\text{Lorentz}}$ does due to the different fitting schemes employed by two methods as shown in Eqs. (3.36) and (3.38), respectively.

As shown in Fig. 4.28:(b), the predictions from GNM-Lorentz and FRI-Lorentz become identical as $\eta \geq 5$Å. A strong correlation between $B^{\text{GNM}-\text{Lorentz}}$ and $B^{\text{FRI}-\text{Lorentz}}$ is revealed at an even smaller scale value. This behavior leads to a general relation

$$\left(\Gamma^{-1}(\Phi(r_{ij};\eta))\right)_{ii} \longrightarrow \frac{c}{\sum_{j,j\neq i}^{N}\Phi(r_{ij};\eta)},\ \eta \to \infty, \tag{4.10}$$

where c is a constant. Relation (4.10) means that the correlation function based gGNM is equivalent to the FRI for a given admissible correlation function when the scale parameter is sufficiently large. This relation is certainly true for the ILF as analytically proved in Eqs. (4.8) and (4.9). Relation (4.10) is a very interesting and powerful result not only for the sake of understanding the GNM and FRI methods and their relationship, but also for the design of accurate and efficient new methods.

It should be noticed that our findings are consistent with the previous finding[54] that, the local packing density described by the direct inverse of the diagonal terms represents only the leading order but not the entire set of the dynamics described by gGNM. Our results reveal an interesting connection between FRI and gGNM when the characteristic distance is sufficiently large.

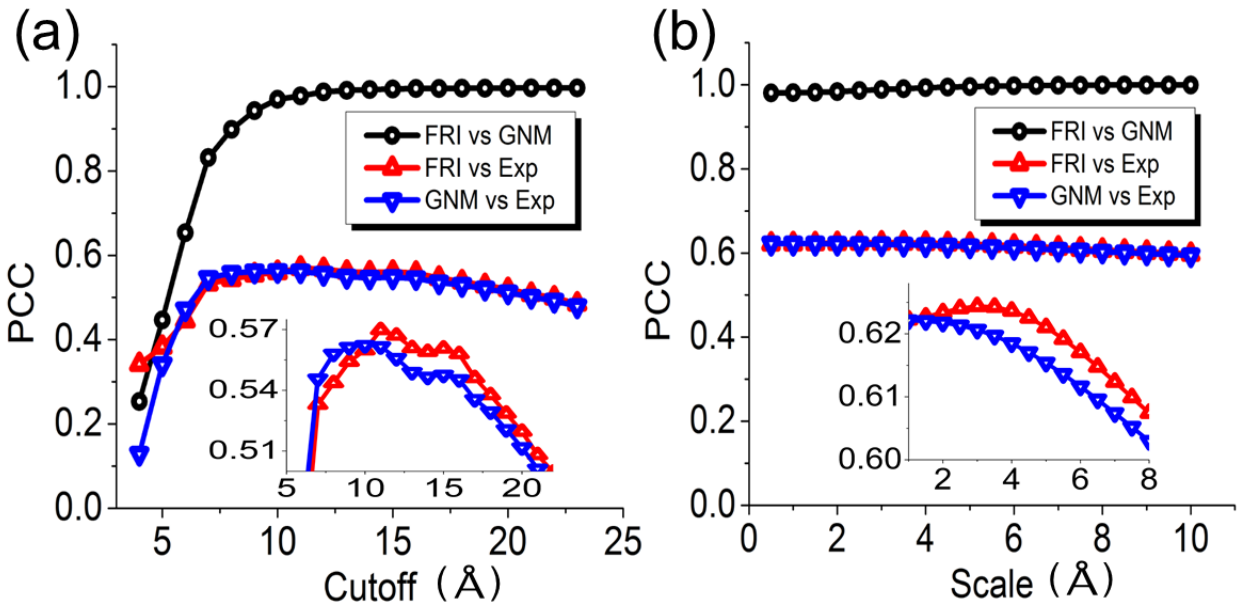


Figure 4.29: PCCs between various B-factors averaged over 364 proteins. (a) Correlations between $B^{\mathrm{GNM-ILF}}$ and B^{Exp}, between $B^{\mathrm{FRI-ILF}}$ and B^{Exp}, and between $B^{\mathrm{GNM-ILF}}$ and $B^{\mathrm{FRI-ILF}}$; (b) Correlations between $B^{\mathrm{GNM-Lorentz}}$ and B^{Exp}, between $B^{\mathrm{FRI-Lorentz}}$ and B^{Exp}, and between $B^{\mathrm{GNM-Lorentz}}$ and $B^{\mathrm{FRI-Lorentz}}$.

4.7.1.3 Validation of gGNM with extensive experimental data

It remains to be proven that the above findings from a single protein are translatable and verifiable on a large class of biomolecules. To this end, we consider a set of 364 proteins, a subset of the 365 proteins utilized and documented in our earlier work.[49] The omitted protein is 1AGN, which has been found to have unrealistic experimental B-factors. We carry out systematic studies of four methods over a range of cutoff distances or scale values. For each given r_c or η, the PCCs between two sets of B-factors are averaged over 364 proteins. Figure 4.29: illustrates our results. Figure 4.29:(a) plots the results of the ILF implemented in both GNM and FRI methods with the cutoff distance varied from 4Å to 23Å. Figure 4.29:(b) depicts similar results obtained by using the Lorentz function implemented in two methods. The scale value is varied over the range of 0.5Å to 10Å.

First, it is evident that the proposed new method, GNM-Lorentz, is very accurate for

the B-factor prediction of 364 proteins as shown in Fig. 4.29:(b). The best GNM-Lorentz prediction is about 10.7% better than that of the original GNM shown in Fig. 4.29:(a). In fact, GNM-Lorentz outperforms the original GNM over a wide range of parameters for this set of proteins, which indicates that the proposed generalization is valuable. Similarly, FRI-Lorentz is also about 10% more accurate than FRI-ILF in B-factor prediction. Since the ILF is a special case and there are infinitely many FRI correlation functions, there is a wide variety of correlation function based gGNMs that are expected to deliver more accurate flexibility analysis than the original GNM does.

Additionally, the FRI-Lorentz method is able to attain the best average prediction for 364 proteins among the four methods as shown in the zoomed in parts in Fig. 4.29:(b). However, for a given correlation function, the difference between FRI and gGNM predictions is very small.

Moreover, for a given admissible FRI function, gGNM and FRI B-factor predictions are strongly linearly correlated and reach near 100% correlation when $r_c > 9$Å or $\eta > 0.5$Å for 364 proteins as demonstrated in Fig. 4.29:. This finding offers a solid confirmation of Eq. (4.10). Therefore, correlation function based gGNMs, including the original GNM as a special case, are indeed equivalent to the corresponding FRI methods in the flexibility analysis for a wide range of commonly used scale values.

Furthermore, it has been shown that the fast FRI is a linear scaling method,[49] while gGNM scales as $O(N^3)$ due to their matrix inverse procedure. As a result, the accumulated CPU times for the B-factor predictions of 364 proteins at $r_c = 7$ or $\eta = 3$ are 0.88, 1.57, 5071.32 and 4934.79 seconds respectively for the FRI-ILF, FRI-Lorentz, GNM-ILF and GNM-Lorentz. The test is performed on a cluster with 8 Intel Xeon 2.50GHz CPUs and 128GB memory. gGNM methods are very fast for small proteins and most of the accumulated gGNM CPU time is due to the computation of three largest proteins (1F8R, 1H6V

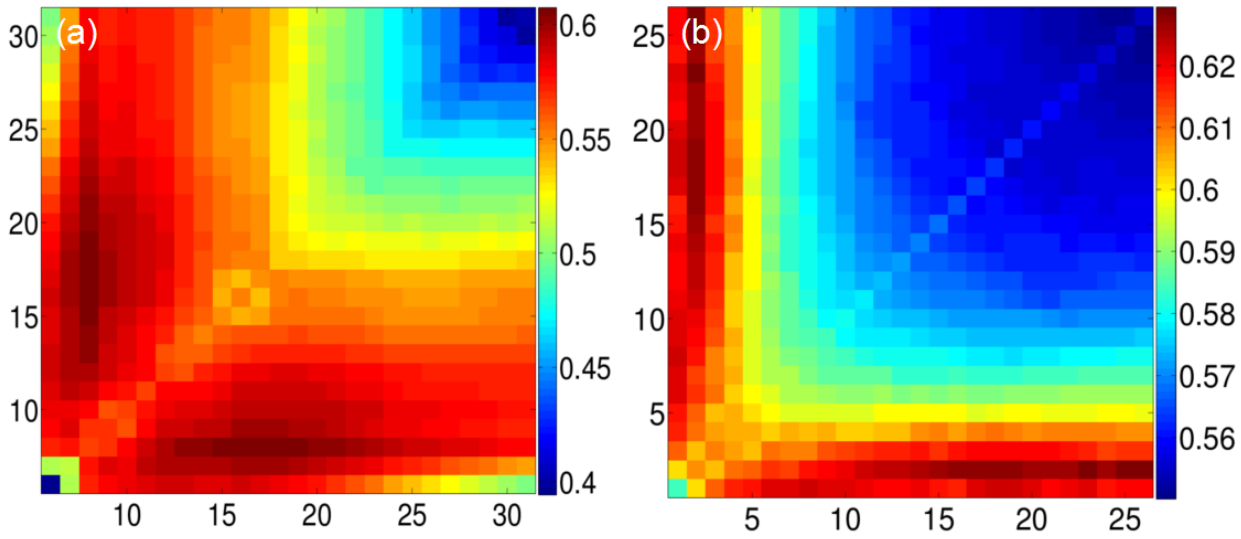


Figure 4.30: The average PCCs over 362 proteins for Type-1 mGNM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two exponential kernels and their scales η are systematically varied in the range of [1Å, 26Å].

and 1QKI) in the test set.

Finally, it is worth mentioning that the earlier FRI rigidity index includes the contribution from the self correlation.[49,75] The present findings do not change if the summation in the generalized Kirchhoff matrix (3.46) is modified to include the diagonal term and then the calculation of gGNM matrix inverse is modified to include the contribution from first eigenmode, $(\Gamma^{-1})_{ii} = \sum_{k=1}^{N} \lambda^{-1} \left[\mathbf{u}_k \mathbf{u}_k^T\right]_{ii}$. In fact, this modification makes the generalized Kirchhoff matrix less singular and faster converging.

4.7.2 Multiscale Gaussian network model

4.7.2.1 Type-1 mGNM

We validate our two types of mGNM with various parameter values over a set of 362 proteins. Two largest proteins, 1H6V and 1QKI, are removed from our earlier data set of 364 proteins[49] due to the limited computational resources. Two kinds of kernels, ILF and exponential, are employed. To explore the multiscale behavior, we use two kernels of the same type but with different characteristic distances in our mGNM schemes. For the ILF kernel

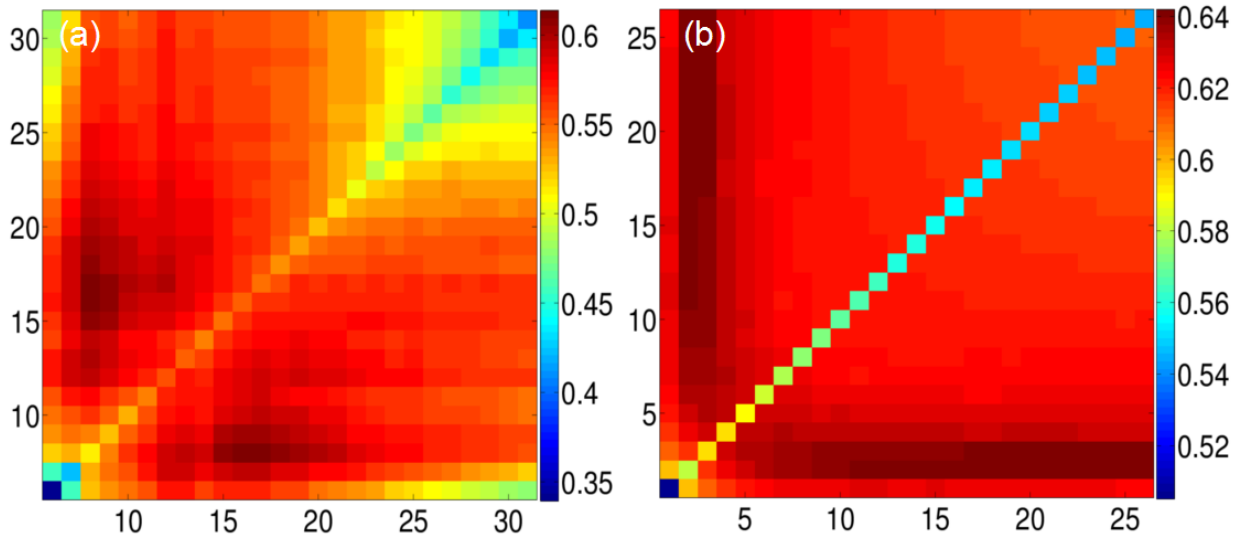


Figure 4.31: The average PCCs over 362 proteins for Type-2 mGNM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two exponential kernels and their scales η are systematically varied in the range of [1Å, 26Å].

based test, the cutoff distances in both kernels vary from 5Å to 31Å. For the exponential kernel based test, we set $\kappa = 1$ and vary η in both kernels within the range of [1Å, 26Å]. The PCCs with experimental B-factors are averaged over 362 proteins. The results for the Type-1 mGNM are demonstrated in Figures 4.30: (a) and (b). When two ILF kernels are used in Figure 4.30: (a), we can seen that the largest average PCCs are concentrated around the region where two kernels have dramatically different cutoff distances with one cutoff being around 7 Å and the other ranging from 14 to 20 Å. Our results indicate that in this set of proteins there is a multiscale property that is better described by mGNM parameterized at different cutoff distances. Moreover, the best PCC is distributed around cutoff distance 7Å, which is consistent with the optimal cutoff distance (7Å) recommended for the traditional GNM method. Similar multiscale behavior can also be observed for an exponential kernel based mGNM as demonstrated in Figure 4.30: (b).

4.7.2.2 Type-2 mGNM

The results of Type-2 mGNMs with ILF kernels and exponential kernels are demonstrated in Figures 4.31: (a) and (b), respectively. The multiscale property is observed to improve predictions for both cases. Compared with Type-1 mGNM, Type-2 mGNM is able to achieve better average PCCs with respect to experimental B-factors. For two ILF kernels, the best average PCC for traditional GNM is 0.567. Type-1 mGNM has significantly improved it to 0.607. Additionally, Type-2 mGNM achieves the best average PCC of 0.614. Similar results are observed in exponential kernel models. For the generalized GNM, the best average PCC is about 0.608. This has been improved to 0.629 in Type-1 mGNM and further improved to 0.642 in Type-2 mGNM. Detailed comparisons are summarized in Table 4.15:.

Table 4.15: The best average PCCs with experimental B-factors. Results for GNM and mGNM are averaged over 362 proteins. Results for ANM and mANM are averaged over 300 proteins.

Kernel	GNM	Type-1 mGNM	Type-2 mGNM	Kernel	ANM	mANM
ILF	0.567	0.607	0.614	ILF	0.490	0.531
Exponential	0.608	0.629	0.642	Gaussian	0.518	0.546

4.7.3 Multiscale anisotropic network models

Table 4.16: 64 Large-sized proteins in the 364-protein data set[49] but not included in our mANM test due to limited computational resource.

1F8R	1GCO	1H6V	1IDP	1KMM	1QKI	1WLY	2A50	2AH1	2BCM
2COV	2D5W	2DPL	2E10	2ETX	2FN9	2I49	2O6X	2OKT	2POF
2PSF	2Q52	2VE8	2W1V	2W2A	2XHF	2Y7L	2YLB	2YNY	2ZCM
2ZU1	3AMC	3BA1	3DRF	3DWV	3G1S	3HHP	3LG3	3MGN	3MRE
3N11	3NPV	3PID	3PTL	3PVE	3PZ9	3SRS	3SZH	3TDN	3UR8
3W4Q	4AM1	4B6G	4B9G	4DD5	4DKN	4DQ7	4ERY	4F01	4G5X
4G6C	4J11	4J78	4JYP						

To study the performance of the multiscale anisotropic network model, we use 300 proteins obtained from the dataset with 364 proteins by removing the largest 64 proteins listed in

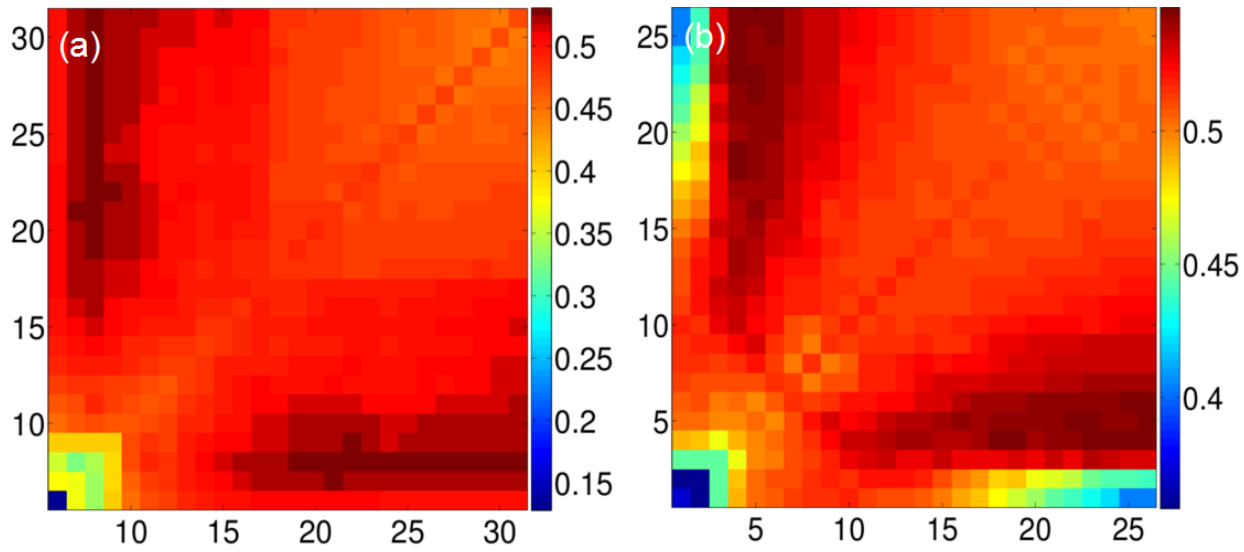


Figure 4.32: The average PCCs over 300 proteins for mANM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two Gaussian kernels ($\kappa = 2$) and their scales η are systematically varied in the range of [1Å, 26Å].

Table 4.16:. The Hessian matrix used in mANM is $3N \times 3N$, which is 9 times larger than the corresponding Kirchhoff matrix in gGNM. This poses more challenges as the computational time grows exponentially with the size of the Hessian matrix.

We consider ILF kernel and Gaussian kernel ($\kappa = 2$) based mANM methods in our test study. Our results are plotted in Figure 4.32:. First, one can still see the multiscale effect in this set of proteins as the best average PCC values of mANM are achieved at the combination of a relatively small cutoff distance (7Å) and a relatively large cutoff distance. These values are much higher than those on the diagonal, which represent the average PCC values of the traditional (single kernel) ANM. For the Gaussian kernel based mANM, we see a similar pattern. However, it achieves better predictions than those of the ILF kernel based mANM. This results are also listed in Table 4.15:. Although the ANM methods are not as accurate as the GNM methods, they are able to offer unique collective motions that otherwise cannot be obtained by the GNM methods.

4.8 mGNM and mANM applications

Having demonstrated the ability of mGNM and mANM for capturing protein multiscale behavior and improving B-factor predictions, we consider a few applications to showcase the proposed methods. First, we take on a set of proteins that fail the original GNM in various ways. This analysis might shed light on why the proposed mGNM works better than the original GNM. Additionally, GNM and ANM can provide domain information for a protein structure. It is well known that GNM eigenvectors can be used to indicate the possible divisions of domains and domain-domain interactions. Finally, ANM eigenvectors are widely used to predict the collective motions of a protein near its equilibrium.

4.8.1 B-factor prediction of difficult cases using mGNM

It is well known that the traditional GNM does not work well in the B-factor prediction for certain proteins for various reasons.[50,52] Park et al. have shown that GNM PCCs with experimental B-factors can be negative.[52] In this work, we demonstrate that the mGNM method is able to deliver more satisfactory B-factor predictions by capturing multiscale features. To demonstrate this we consider four proteins, 1CLL, 1V70, 2HQK and 1WHI. The Type-2 mGNM with two exponential kernels is used for these applications. As depicted in Figure 4.31:(b), there is a wide range of scale parameters that deliver accurate B-factor predictions. We choose $\kappa = 1, \eta^1 = 3$ Å and $\kappa = 1, \eta^2 = 25$Å to use in this test. For comparisons to the original method, the traditional GNM, or GNM-ILF, is employed with different cutoff distances, namely 7 Åand 20 Å, which are denoted as GNM7 and GNM20, respectively.

Figures 4.33:, 4.34:, 4.35: and 4.36: illustrate the results. In each figure, protein surfaces are colored by B-factor values predicted by GNM7, mGNM and the flexibility function in Eq. (3.15), respectively in subfigures (a), (b) and (c). The comparisons of B-factors predicted by GNM7 and GNM20 with those of experiments are demonstrated in subfigures (d). Similarly,

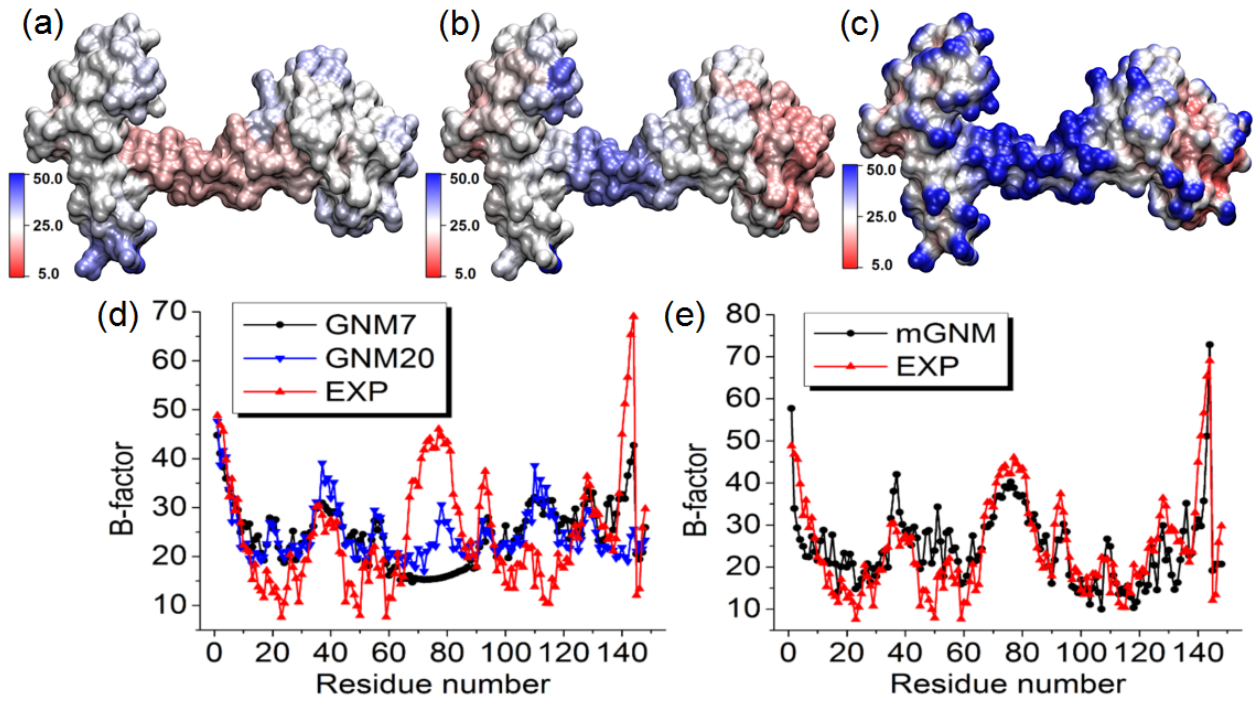


Figure 4.33: Comparison between Type-2 mGNM with exponential kernel and traditional GNM for the B-factor prediction of protein 1CLL. Two scales, $\eta^1 = 3$Å and $\eta^2 = 25$Å, are employed in mGNM. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by our Type-2 mGNM. (c) Molecular surface colored by multiscale flexibility function in Equation (3.15). (d) B-factors predicted by traditional GNM with cutoff distances 7Å (GNM7) and 20Å (GNM20). (e) B-factors predicted by mGNM.

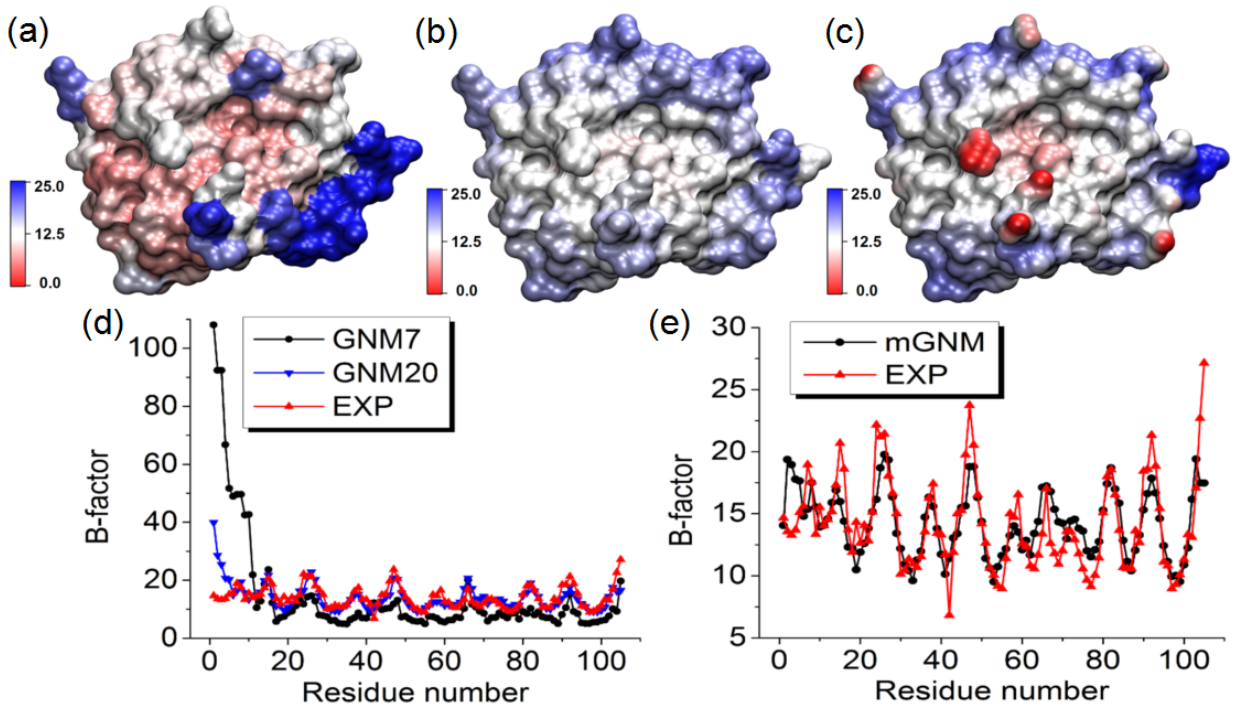


Figure 4.34: Comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 1V70 B-factor prediction. Two scales, $\eta^1 = 3$Å and $\eta^2 = 25$Å, are employed in mGNM. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by our Type-2 mGNM. (c) Molecular surface is colored by multiscale flexibility function in Equation (3.15). (d) B-factors predicted by traditional GNM with cutoff distances 7Å (GNM7) and 20Å (GNM20). (e) B-factors predicted by mGNM.

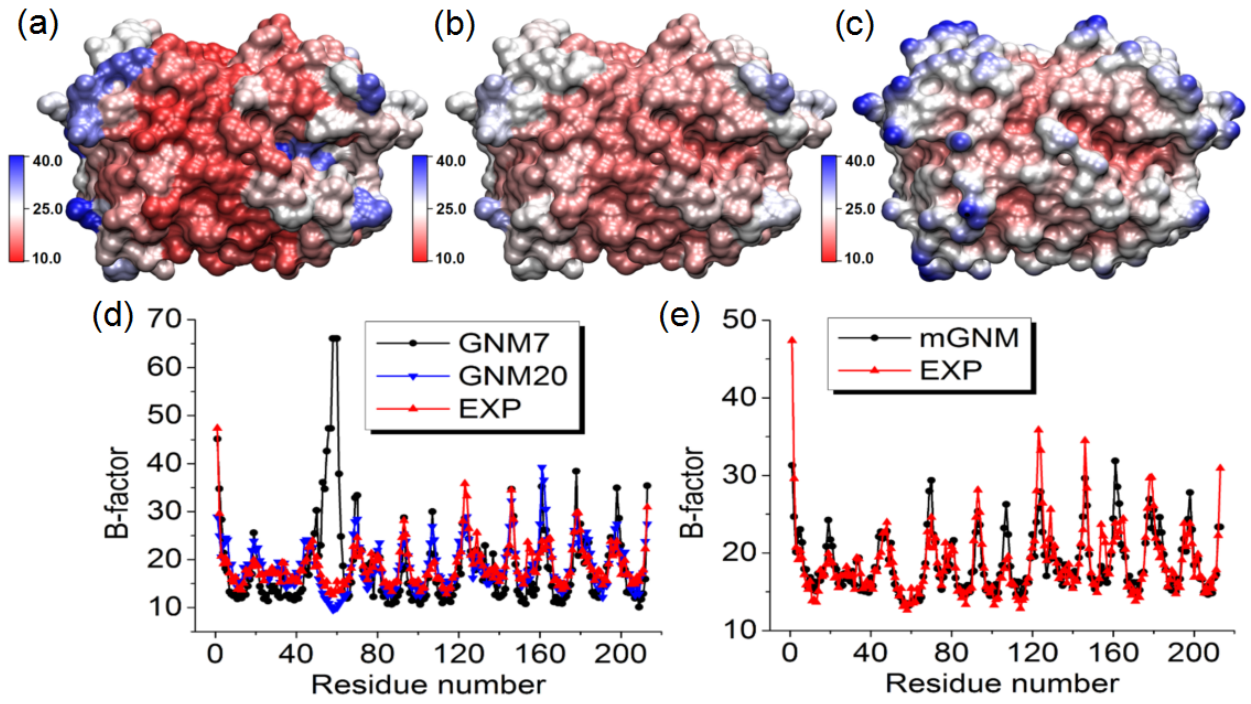


Figure 4.35: Comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 2HQK B-factor prediction. Two scales, $\eta^1 = 3$Å and $\eta^2 = 25$Å, are used for mGNM. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by the Type-2 mGNM. (c) Molecular surface is colored by multiscale flexibility function in Equation (3.15). (d) B-factors predicted by traditional GNM with cutoff distances 7Å (GNM7) and 20Å (GNM20). (e) B-factors predicted by mGNM.

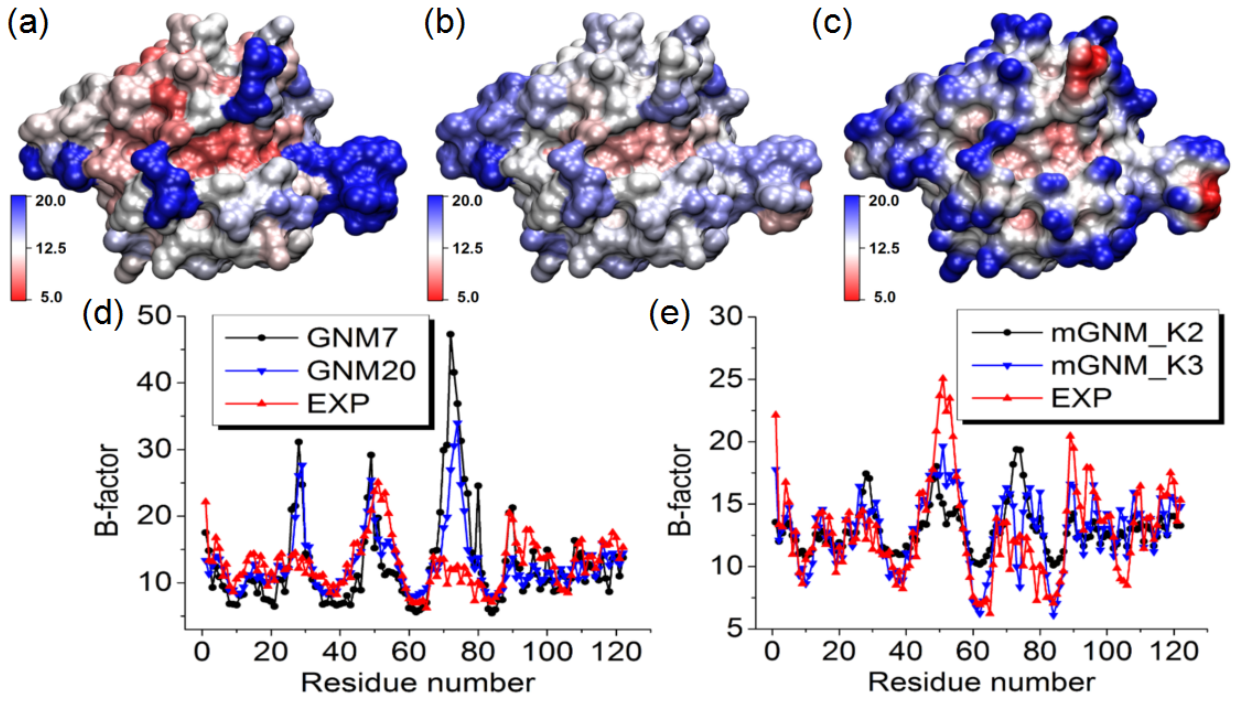


Figure 4.36: Comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 1WHI B-factor prediction. Two mGNMs are used. The first one, mGNM_K2, has two exponential kernels with $\kappa = 1$, $\eta^1 = 3$Å and $\eta^2 = 25$Å. The second mGNM, mGNM_K3, has an extra exponential kernel with $\kappa = 1$ and $\eta^3 = 10$ Å. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by a Type-2 mGNM. (c) Molecular surface is colored by multiscale flexibility function in Equation (3.15). (d) B-factors predicted by traditional GNM with cutoff distances 7Å (GNM7) and 20Å (GNM20). (e) B-factors predicted by two mGNMs, mGNM_K2 and mGNM_K3.

the comparisons of the predicted B-factors by mGNM with those of experiments are plotted in subfigures (e). A summary of related PCC values are listed in Table 4.15:.

Table 4.17: Case study of B-factor prediction for four proteins in three different schemes: GNM7, GNM20 and mGNM. In the case of 1WHI, we use mGNM with two kernels and three kernels (value in parentheses).

PDB ID	GNM7	GNM20	mGNM
1CLL	0.261	0.235	0.763
1V70	0.162	0.548	0.750
2HQK	0.365	0.781	0.833
1WHI	0.270	0.370	0.484(0.766)

Flexible hinges are sometimes important to protein functions but they are not always easily detected by GNM type methods.[25,39] As shown in Figure 4.33:, the original GNM parameterized at cutoff distance 7 or 20 Å does not work well for the hinge located around residues 65-85. In fact, the GNM method cannot predict the flexible hinge at any given cutoff distance. Whereas, the two-kernel mGNM is able to capture the hinge behavior.

Protein 1V70 shown in Figure 4.34: is another difficult case for the traditional GNM method. At cutoff distance 7Å, it severely over-predicts the B-factors of the first 12 residues. However, its prediction improves if a larger cutoff distance is used. In contrast, the two-kernel mGNM provides a very good prediction.

Figure 4.35: illustrates one more interesting situation. The tradition GNM with cutoff distance 7Å over-predicts the B-factors for residues near number 58. However, at a large cutoff distance of 20Å, it is able to offer accurate results. In this case, mGNM is able to further improve the accuracy.

The case of 1WHI given in Figure 4.36: is difficult for both methods tested. GNMs with two different parameterization do not work well and two-kernel mGNM, while more accurate, still does not reach a PCC greater than 0.5. Its PCC of 0.484 is just a minor improvement of GNM PCCs, 0.270 (obtained at $r_c = 7$Å) and 0.370 (obtained at $r_c = 20$Å). It should be

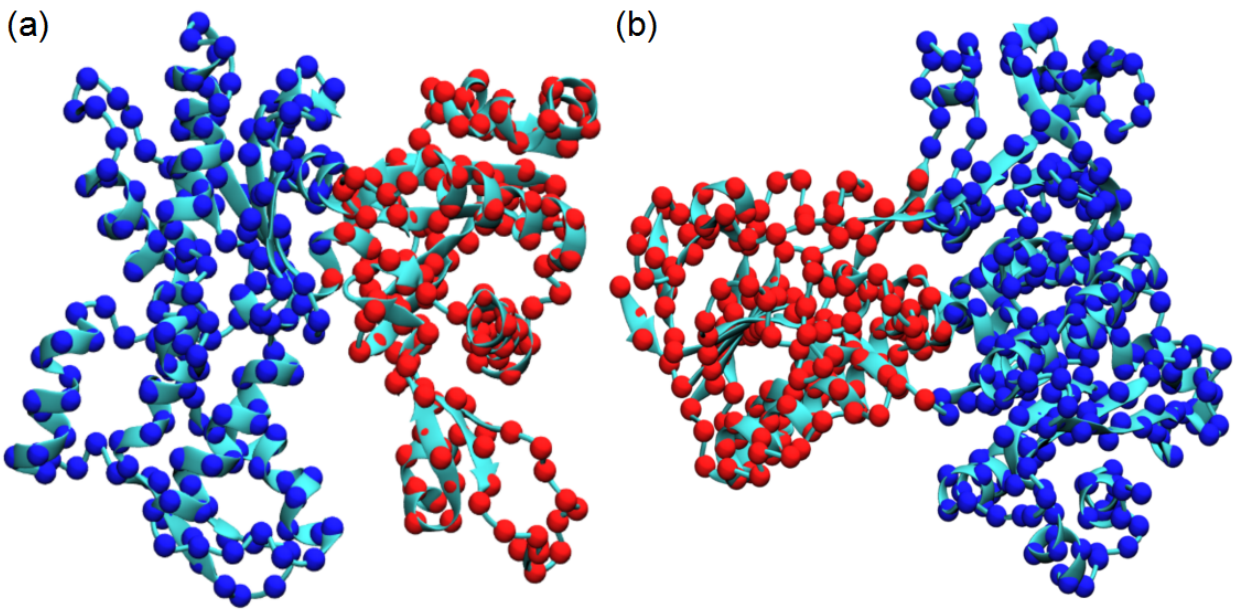


Figure 4.37: Protein domain decomposition with Type-1 mGNM. The first eigenvector (Fiedler vector) is used to decompose the protein into two domains. (a) protein 1ATN (chain A); (b) protein 3GRS.

noticed that mGNM can simultaneously incorporate several scales, therefore, we employ an extra kernel with $\kappa = 1, \eta^3 = 10$ Å to deal with this protein. As shown in Table 4.17: and Figure 4.36:, the three-kernel mGNM is able to deliver a good PCC of 0.766.

4.8.2 Domain decomposition using mGNM

Mathematically, the first smallest nonzero eigenvalue is called algebraic connectivity or Fiedler value and the related eigenvector is called Fiedler vector. It is known that the Fiedler vector can be used to decompose a protein into two domains. Each particle in the protein is assigned with a value (element) from the Fiedler vector and these particles are grouped according to their positive or negative signs. The particles with zero values can be classified into either group as they are usually in a linking region between two domains.

To test the performance of the mGNM methods, we utilize two test proteins, 1ATN (chain A) and 3GRS, which are also used by Kundu, et al.[42] We compare the performance of two

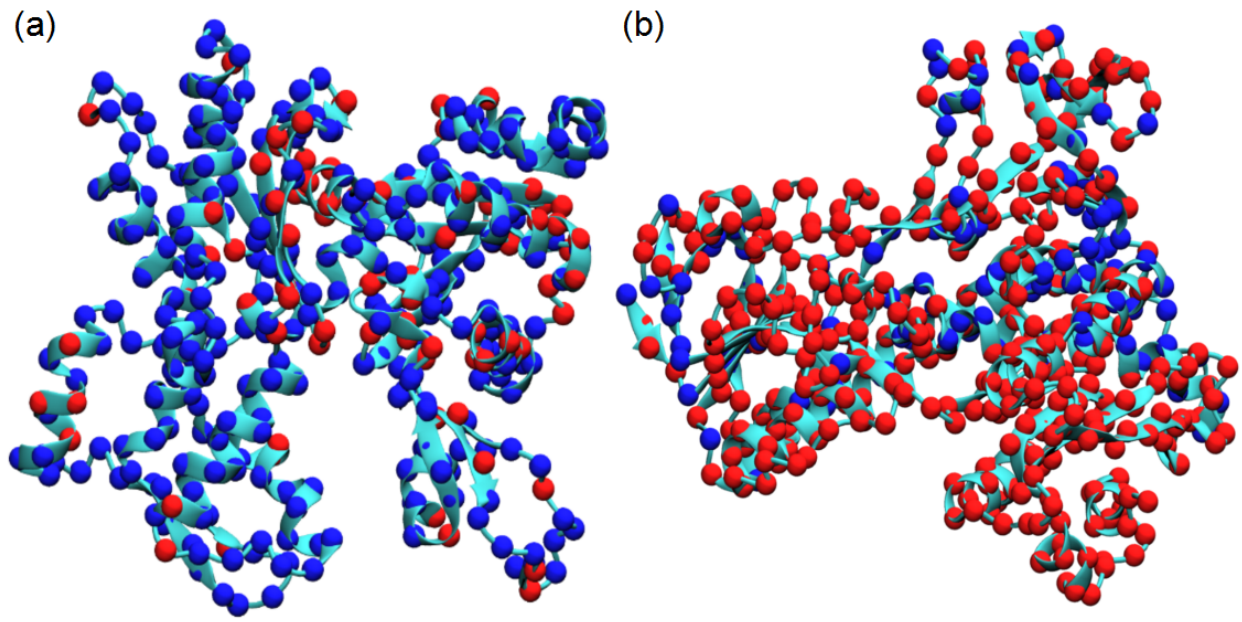


Figure 4.38: Protein domain decomposition with Type-2 mGNM. The first eigenvector (Fiedler vector) is used to decompose the protein into two domains. (a) protein 1ATN (chain A); (b) protein 3GRS. It can be seen that Type 2 mGNM fails in protein domain decomposition.

types of mGNMs. In Type-1 mGNM, we use the exponential kernels with $\kappa = 1, \eta^1 = 3$ Å and $\kappa = 1, \eta^2 = 25$Å. In Type-2 mGNM, we use three exponential kernels with the same two kernels as Type-1 mGNM with an extra kernel parameterized as $\kappa = 1, \eta^3 = 10$ Å. The results are depicted in Figures 4.37: and 4.38:, respectively. It can be seen that Type-1 mGNM delivers a great decomposition, which is also consistent with the prediction from traditional GNM.[42] However, the Type-2 mGNM does not produce a reasonable result. This is due to the fact that Algorithm is designed to construct the symmetric Kirchhoff matrix with required diagonal elements and its non-diagonal elements do not properly reflect the protein connectivity.

However, the PCCs of Type-1 mGNM for 1ATN and 3GRS are 0.460 and 0.658. Whereas, the PCCs of Type-2 mGNM for 1ATN and 3GRS are 0.660 and 0.666. These results indicate that the B-factor values are mainly dictated by the diagonal matrix elements while the

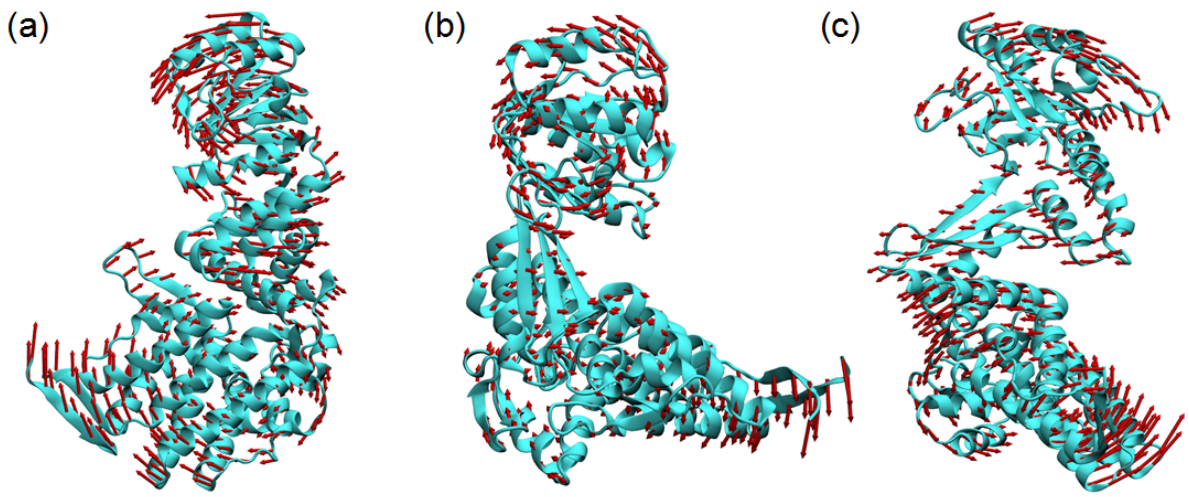


Figure 4.39: The collective motions of protein 1GRU (chain A). The seventh, eighth and ninth modes calculated from mANM are demonstrated in (a), (b) and (c), respectively.

domain separation is determined by non-diagonal matrix elements.

4.8.3 Collective motion simulation using mANM

GNM is an isotropic model which quantifies the general atomic fluctuations in a molecule. In contrast, ANM is designed to describe the anisotropic properties, such as collective motions of a molecule near equilibrium. Typically, the first six modes, corresponding to six zero (or near zero) eigenvalues, represent the trivial translational and rotational modes of a complex biomolecule. Global modes that are unique to the biomolecular structure are described by eigenvectors associated with the next smallest nonzero eigenvalues. Due to its simplicity, accuracy and availability, ANM is widely used to study the dynamics of biomolecules.

In the present work, we have designed an mANM method to maintain the aforementioned properties. To validate mANM for anisotropic mode analysis, we use two test proteins, 1GRU (chain A) and 1URP (chain A). The protein 1GRU is chaperonin GroEL, a benchmark test for ANM.[68,82] We employ mANM with two Gaussian kernels ($\kappa = 2$) with $\eta = 5$Å and $\eta = 20$Å. We compute eigenvectors associated with the first three nonzero eigenvalues. As

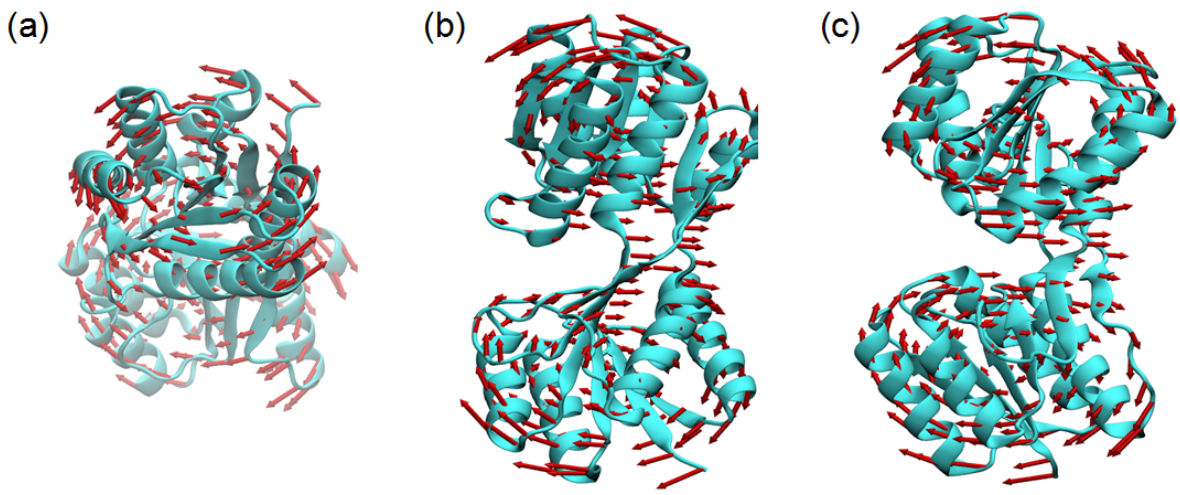


Figure 4.40: The collective motions of protein 1URP (chain A). The seventh, eighth and ninth modes calculated from mANM are demonstrated in (a), (b) and (c), respectively.

illustrated in Figure 4.39:, the mANM results are in an excellent agreement with those of ANM for chaperoin GroEL.[68,82]

To further validate the mANM method, we examine another test case, 1URP. This molecules is a ribose-binding protein and its anisotropic motions have been studied previously.[45] We utilize the same set of parameters described above. Figure 4.40: demonstrates the mANM results for this structure and again the results are in close agreement with the traditional ANM analysis.[45]

4.9 FRI-based hinge prediction validation with known hinging proteins

In this section we take an in-depth look at the hinges predicted by gGNM modes on a high quality set of test cases borrowed from the StoneHinge[39] study and from earlier studies by Flores, et al. This set includes 32 structures, open and closed conformations of 16 different proteins. Each protein in the set has a known hinge type motion mentioned in the source literature.

4.9.1 gGNM mode-based hinge prediction

The FRI-based hinge predictions for 19 out of the 32 structures studied were clear matches while another 11 cases are partial hits, where there is at least one true positive and one false positive or false negative. Together there are 30 of the 32 structures for which the predictions are at least partially accurate. Complete results for gGNM mode-based hinge prediction are shown in Tables 4.18:, 4.19: and 4.20:. Prediction accuracy is determined under the *loose criterion* used by earlier comprehensive hinge studies, a 14 residue window around the predicted hinge point. A case where the literature hinges and gGNM mode hinge predictions are in perfect agreement, open and and closed forms of ovotransferrin, is shown in Figure 4.41:. The previously identified hinge residues are residues 333 and 342. Visual inspection of the region shows the residues from 333 to 342 are all random coil, therefore we consider this range to be a hinge region. gGNM mode hinge prediction places the center of the hinge at residue 344 in the closed conformation and 339 in the open conformation, in very close agreement with the 333 to 342 range. Therefore we count this among the full hits.

Next we examine the proteins on which the automatic gGNM-based mode hinge prediction does not agree with the known hinge residues, the case for 2 of 32 structures. It is important to note that in two of these cases, the second gGNM mode provides accurate predictions. Failure to predict hinges accurately can happen because gGNM modes are not accurate for a particular structure (due to a variety of reasons including missing ligands, unaccounted for crystal effects, etc.), because the hinging motions of consequence are not the hinges between the largest domains, or because there are multiple modes with very similar, low eigenvalues. One example where the important hinge is not the one dividing domains is lactoferrin, shown in Figure 4.43:. The case of lactoferrin has been difficult for many hinge prediction methods such as StoneHinge and TLSMD while others such as FlexOracle can readily and accurately identify the biologically important hinges. This is likely due to the

fact that the most biologically interesting hinging motion is not at a domain separation, but rather a smaller fluctuation unique to one domain. Some earlier hinge studies have considered this as the only hinge to predict, however there is some evidence from crystallographic studies that lactoferrin does in fact have a hinging motion between its two largest domains. This hinging between domains is what is suggested by the first gGNM mode. Furthermore, the second gGNM mode identifies the smaller, biologically-relevant hinging. Therefore, this may be a case of where the initial literature-based identification of hinge residues was wrong instead of a failure of gGNM-based hinge prediction.

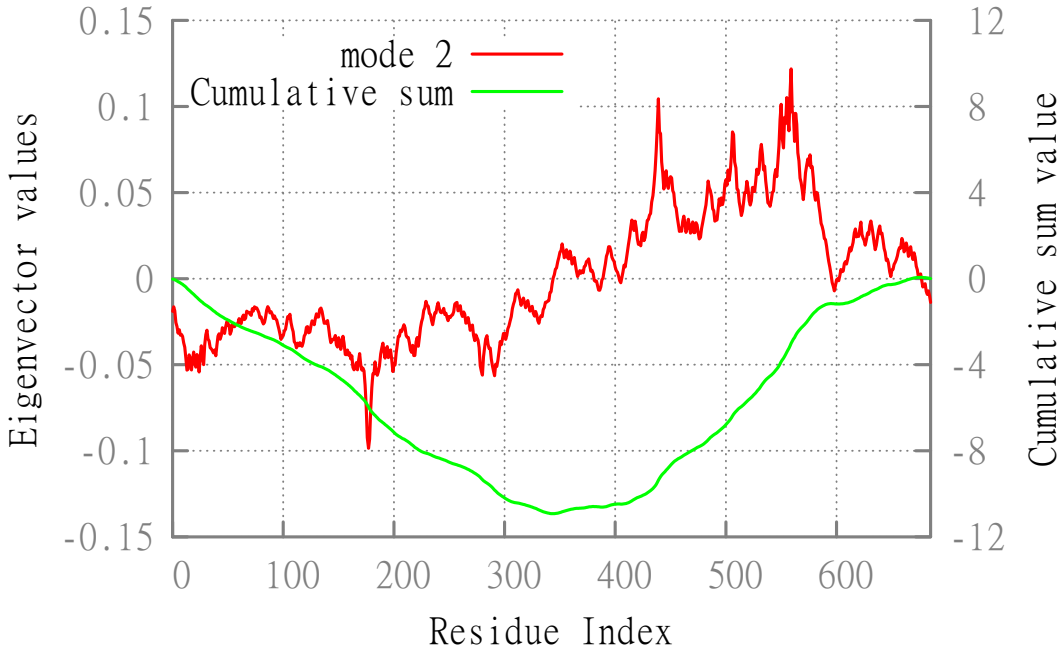
Another example where the second mode contains important hinge predictions is ribose binding protein open conformation, Figure 4.42:. The closed conformation prediction for ribose binding protein is a perfect hit, however, the open conformation gives the correct hinge predictions when considering the second gGNM mode. Based on this result we suggest always considering the hinge predictions from the second mode of gGNM particularly when the hinge(s) predicted by the first mode do not appear to be at a hinging region upon visual inspection of the structure.

Table 4.18: gGNM-based hinge predictions for 32 protein structures compared with consensus hinge residues determined from literature and other hinge studies.[39]

Protein Name	Closed	Prediction	Consensus	Open	Prediction	Consensus
Ovotransferrin	1aiv	344	333, 342	1ovt	339	333, 342
Adenylate kinase	1ake	110, 168	124-126, 161-163	2ak3	113, 173	124-126, 161-163
CAPK	1atp	126	119-126	1ctp	126	119-126
Biotin carboxylase	1bnc	114, 208	130-131, 203-204	1dv2	107, 210	130-131, 203-204
DNA polymerase beta	1bpd	97	79-83, 91-93	2bpg	143 (86, 259 m2)	79-83, 91-93
Calmodulin	1cll	80	76-80	1cfd	78	76-80
Elastase	1ezm	142	132-135	1u4g	142	132-135
GluR2	1fto	108, 218	214-215	1ftm	108, 218	214-215
Lir-1	1g0x	95	95-96	1p7q	96	95-96
Bence-Jones Protein	4bjl	115	108-116	4bjl	112	108-116
Inorganic pyrophosphatase	1k20	187	188-192	1k23	188	188-192
Phosphoglycerate kinase	1kf0	194	201-205, 402-404	1hdi	194	201-205, 402-404
Lactoferrin	1lfh	343	90, 250 (340*)	1lfg	340 (91, 250 m2)	90, 250 (340*)
LAO binding protein	1lst	89, 192	89-91, 182-194	2lao	89, 191	89-91, 182-194
Glutamine blinding protien	1wdn	87, 182	85-90, 178-185	1ggg	87, 182	85-90, 178-185
Ribose binding protein	2dri	103, 235	103-104, 235-236	1urp	151 (103, 235 m2)	103-104, 235-236

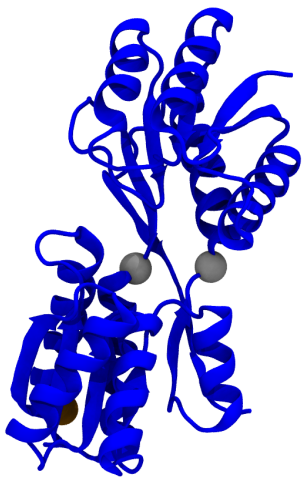
(a) 1AIV - view 1

(b) 1AIV - view 2

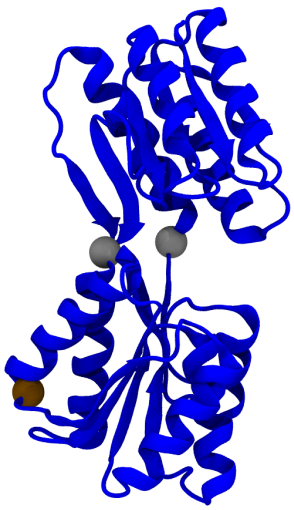


(c) 1AIV Mode 1

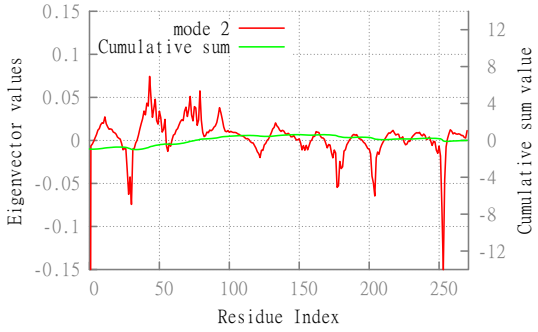
Figure 4.41: Top, secondary structure representation of ovotransferrin with hinge residues highlited by VdW representations of their C-alpha atoms. Bottom, values by residue for modes 1 and 2 (left y-axis) with cumulative sum (right y-axis). The maximum and minimum values of the cumulative sum correspond to hinge points



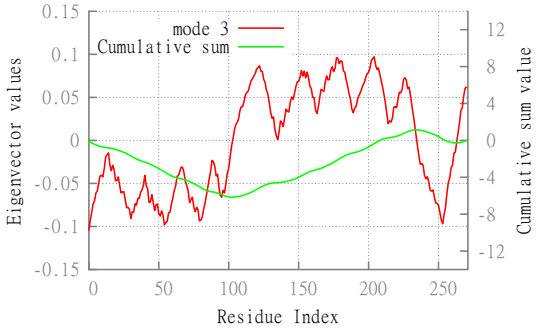
(a) 1URP - view 1



(b) 1URP - view 2



(c) 1URP Mode 1

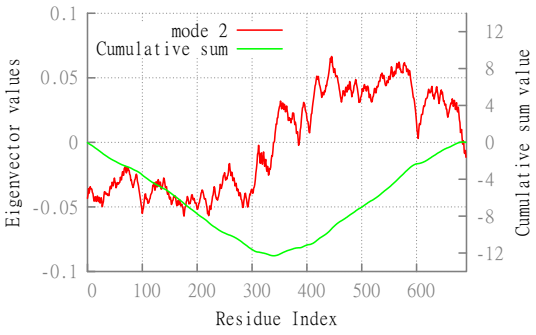


(d) 1URP Mode 2

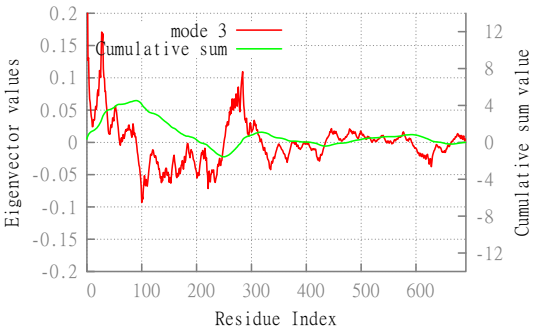
Figure 4.42: Top, secondary structure representation of ribose binding protein with hinge residues highlited by VdW representations of their C-alpha atoms. Bottom, values by residue for modes 1 and 2 (left y-axis) with cumulative sum (right y-axis). The maximum and minimum values of the cumulative sum correspond to hinge points

(a) 1LFG - view 1

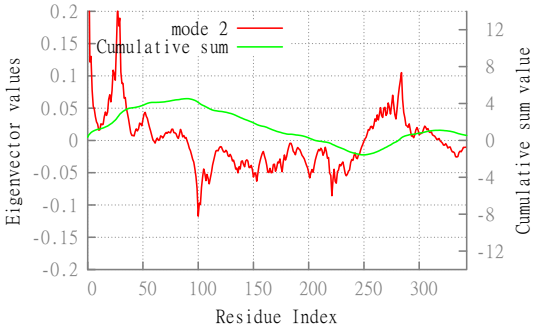
(b) 1LFG - view 2



(c) 1LFG Mode 1



(d) 1LFG Mode 2



(e) 1LFG Mode 1 - Domain only calculation

Figure 4.43: Top, secondary structure representation of lactoferrin with hinge residues highlited by VdW representations of their C-alpha atoms. Bottom, values by residue for modes 1 and 2 (left y-axis) with cumulative sum (right y-axis). The maximum and minimum values of the cumulative sum correspond to hinge points.

Table 4.19: gGNM-based hinge predictions for 32 protein structures compared with consensus hinge residues determined from literature and other hinge studies.[39] Y - The hinge(s) are completely and uniquely identified, P - A predicted hinge is off from a true hinge position by less than 5 amino acids or there is a false positive or negative, N - Failure to identify any major hinges.

Protein Name	Closed	Prediction	Consensus	Open	Prediction	Consensus
Ovotransferrin	1aiv	Y	333, 342	1ovt	Y	333, 342
Adenylate kinase	1ake	P	124-126, 161-163	2ak3	P	124-126, 161-163
CAPK	1atp	Y	119-126	1ctp	Y	119-126
Biotin carboxylase	1bnc	P	130-131, 203-204	1dv2	P	130-131, 203-204
DNA polymerase beta	1bpd	P	79-83, 91-93	2bpg	N (P mode 2)	79-83, 91-93
Calmodulin	1cll	Y	76-80	1cfd	Y	76-80
Elastase	1ezm	Y	132-135	1u4g	Y	132-135
GluR2	1fto	P	214-215	1ftm	P	214-215
Lir-1	1g0x	Y	95-96	1p7q	Y	95-96
Bence-Jones Protein	4bjl	Y	108-116	4bjl	Y	108-116
Inorganic pyrophosphatase	1k20	Y	188-192	1k23	Y	188-192
Phosphoglycerate kinase	1kf0	P	201-205, 402-404	1hdi	P	201-205, 402-404
Lactoferrin	1lfh	P	90, 250 (340*)	1lfg	P	90, 250 (340*)
LAO binding protein	1lst	Y	89-91, 182-194	2lao	Y	89-91, 182-194
Glutamine blinding protien	1wdn	Y	85-90, 178-185	1ggg	Y	85-90, 178-185
Ribose binding protein	2dri	Y	235-236	1urp	N (Y mode 2)	235-236

Table 4.20: Summary of hits for gGNM-based predictions of hinges for 32 PDBs. Full - The hinge(s) are completely and uniquely identified, Partial - A predicted hinge is off from a true hinge position by less than 5 amino acids or there is a false positive or negative, None - Failure to identify any major hinges.

Yes	19
Partial	11
No	2

4.9.2 Machine learning feature ranking

The first round of feature ranking, calculating the F-score, included all 55 considered features and the complete results are shown in Table 4.21:. The F-score serves as a useful first filter because it is quickly calculated and the scores are independent of the other variables tested, which is advantageous because it does not incorrectly weight features due to the presence of many correlated features, a weakness of random forest. The disadvantage of this approach is that there is no indication of which variables are highly correlated and therefore typically should not be used in the same model. A second round of feature ranking was done on the top features based on F-score and these results are displayed in Table 4.22:

The top four features by F-score are all derived from gGNM mode 1 calculations. These features include *ishinge3*, *ishinge*, *cMode 1* and *hingedist*. The F-scores for the latter three features are very similar while the first, *ishinge3*, is more than double the F-score of those three. This discrepancy is probably due to the fact that many hinges in this data set span multiple residues and *ishinge3* essentially marks seven residue regions as hinges based on gGNM mode 1 values. Therefore, we suspect these features are all essentially predicting the same thing and this emphasizes the importance of trying multiple slightly different feature formulations to find one with the most value as a predictor.

Interestingly the feature scores fall off sharply after the gGNM mode derived features. The features rounding out the top ten, in order, are *isH*, *HP6*, *RES6*, *isC*, *insec*, *ROT6* and *P (PSSM)*. Two of these features are secondary structure related, three features describe the local environment within 6 Angstroms and one feature is sequence related. While the F-scores of these are considerably lower than the top four, these features in the top ten deserve some future, in-depth analysis to see if other features related to these could be created that are more useful for predictions.

After the top ten features there are the mode 2 derived features and FRI flexibility index

related features. As mentioned earlier, sometimes it is hard to distinguish between the most important mode when two modes have very similar eigenvalues. Therefore it may be that mode 2 features are only useful in those cases. With this is mind we can refine the model further by checking for similar eigenvalues for the lowest modes of each structure and, if they are found to be sufficiently close, we combine the hinge predictions from both modes when creating features. Under this scheme, all 32 structures' hinges are at least partially predicted correctly by this method.

4.9.3 SVM model prediction results

This section provides an example of what is achievable using support vector machine modeling with gGNM mode features. Listed in Tables 4.23: and 4.24: are the hinge residues predicted by two of the more accurate SVM models we are able to create with the features tested. The features used in the first model include *FRIf*, *dFRI*, *FRIf*, *ishinge*, *ishinge3*, *hingedist*, *HP6*, *RES6* and *ROT6*. The features used in the second model include *ishinge*, *ishinge3*, *hingedist*, *Mode1* and *cMode1*. Many combinations of features were tried and, as expected from the feature ranking results, the only essential features are those derived from mode 1 of mGNM.

The results from the SVM models are very similar to those from the mGNM mode predictions however the SVM model tends to predict fewer hinge residues. In some cases this serves to remove false positives but it also causes false negatives. In the end, it would be just as good to use the mGNM predictions rather than take the extra step to use the SVM model.

Table 4.21: Feature importance rankings by F-score. F-scores are calculated using the LIBSVM software.

F-score rank	F-score	Feature name	F-score rank	F-score	Feature name
1	0.056996	ishinge3	31	0.000135	D (PSSM)
2	0.017780	ishinge	32	0.000099	Phil/A^2
3	0.016408	cMode 1	33	0.000085	prop1
4	0.010333	hingedist	34	0.000083	isB
5	0.003040	isH	35	0.000082	isT
6	0.002399	HP6	36	0.000081	C (PSSM)
7	0.002279	RES6	37	0.000070	H (PSSM)
8	0.001591	isC	38	0.000064	Phob/A^2
9	0.001551	insec	39	0.000047	N (PSSM)
10	0.001375	ROT6	40	0.000043	A (PSSM)
11	0.001301	P (PSSM)	41	0.000036	Surf/A^2
12	0.001283	Mode 2	42	0.000035	E (PSSM)
13	0.001184	cMode 2	43	0.000034	HP1
14	0.000525	N(overl)	44	0.000025	Q (PSSM)
15	0.000495	FRI f-index	45	0.000014	S (PSSM)
16	0.000473	Avg. mFRI within 6A	46	0.000012	L (PSSM)
17	0.000461	Mode 1	47	0.000010	K (PSSM)
18	0.000422	Difference of mFRI and FRI	48	0.000010	Y (PSSM)
19	0.000409	isE	49	0.000009	I (PSSM)
20	0.000409	G (PSSM)	50	0.000003	Total/A^2
21	0.000400	mFRI B-factor fit	51	0.000002	V (PSSM)
22	0.000400	prop5	52	0.000002	%SASA
23	0.000379	T (PSSM)	53	0.000001	R (PSSM)
24	0.000324	mFRI f-index	54	0.000000	isI
25	0.000302	prop3	55	0.000000	nosec
26	0.000275	M (PSSM)			
27	0.000241	isG			
28	0.000209	FRI B-factor fit			
29	0.000151	W (PSSM)			
30	0.000143	F (PSSM)			

Table 4.22: Feature importance rankings by random forest method. Importance values calculated using the R package caret comman, varImp.

Rank	Importance	Importance (Scaled)	Feature Name
1	0.79	100.00	cMode1
2	0.73	85.78	ishinge5
3	0.61	53.53	RES6
4	0.60	51.52	HP6
5	0.59	48.30	PSSM(P)
6	0.58	45.94	cMode2
7	0.58	44.81	ROT6
8	0.57	44.07	insec
9	0.56	41.97	isC
10	0.55	37.96	N(overlap)
11	0.40	0.00	isH

Table 4.23: SVM results for a model with eight of the top ranked features, *FRIf*, *dFRI*, *FRIf*, *ishinge*, *ishinge3*, *hingedist*, *HP6*, *RES6* and *ROT6*.

Closed	Prediction	Consensus	Open	Prediction	Consensus
1aiv	342	333, 342	1ovt	341-342	333, 342
1ake		124-126, 161-163	2ak3		124-126, 161-163
1atp	123-126	119-126	1ctp	123-126	119-126
1bnc	115-116, 205-206	130-131, 203-204	1dv2	104-108	130-131, 203-204
1bpd		79-83, 91-93	2bpg	92-93	79-83, 91-93
1cll	76-80	76-80	1cfd	76-80	76-80
1ezm	142	132-135	1u4g	141	132-135
1fto	105-108, 215-221	214-215	1ftm	105-108, 215-221	214-215
1g0x	92-98	95-96	1p7q	93-99	95-96
4bjl	114-117	108-116	4bjl	114-117	108-116
1k20	190	188-192	1k23	188-191	188-192
1kf0		201-205, 402-404	1hdi		201-205, 402-404
1lfh	341-342	90, 250 (340*)	1lfg	338-343	90, 250 (340*)
1lst	89-92, 189-195	89-91, 182-194	2lao	87-92, 189	89-91, 182-194
1wdn	85-88, 179-185	85-90, 178-185	1ggg	84-88, 179-185	85-90, 178-185
2dri	102-104, 235-236	235-236	1urp		235-236

Table 4.24: SVM results for a model with five mGNM-based features, *ishinge*, *ishinge3*, *hingedist*, *Mode1* and *cMode1*.

Closed	Prediction	Consensus	Open	Prediction	Consensus
1aiv		333, 342	1ovt	341	333, 342
1ake	124-126, 161-163	124-126, 161-163	2ak3		124-126, 161-163
1atp	122-126	119-126	1ctp	123-126	119-126
1bnc	205-206	130-131, 203-204	1dv2	104-108	130-131, 203-204
1bpd	81	79-83, 91-93	2bpg	79-82	79-83, 91-93
1cll	77-83	76-80	1cfd	76-80	76-80
1ezm	139	132-135	1u4g	139-141	132-135
1fto	109-111, 214	214-215	1ftm	105-108, 215-221	214-215
1g0x	92-94	95-96	1p7q		95-96
4bjl	116-117	108-116	4bjl	114-117	108-116
1k20	190	188-192	1k23	188-191	188-192
1kf0	195-197, 253	201-205, 402-404	1hdi		201-205, 402-404
1lfh		90, 250 (340*)	1lfg	343	90, 250 (340*)
1lst	90-92	89-91, 182-194	2lao	90, 186-190	89-91, 182-194
1wdn	85-88, 179-185	85-90, 178-185	1ggg	84-88, 179-185	85-90, 178-185
2dri	102, 235	235-236	1urp		235-236

CHAPTER V. Conclusions and Future Directions

5.1 Conclusions

In living organisms, proteins and nucleic acids carry out a vast variety of functions including providing structural support, catalyzing chemical reactions, replicating DNA, or responding to stimuli. Many of these functions are performed through synergistic interactions or correlations over multiple length scales, including atomic, van der Waals, residue, alpha-beta complex, domain-domain and protein-protein interactions. Popular existing flexibility methods such as Gaussian network model do not directly account for the multiscale nature of macromolecular interactions and fail to predict Debye-Waller factors or B-factors for many proteins that involve multiple length characteristics.

This work puts forward a multiscale, multiphysics and multidomain model, the flexibility-rigidity index (FRI), to estimate the static property of macromolecules. A basic assumption of the present FRI theory is that the geometry or structure of a given protein together with its specific environment, namely, solvent, assembly or crystal lattice, completely determines the biological function and properties including flexibility, rigidity and energy. As such, the present approach bypasses the construction of the Hamiltonian and interaction potentials. A possible drawback of the present method is that the full geometric and topological information of a protein complex is usually not available, which contributes to modeling errors.

We utilize monotonically decreasing functions to measure the geometric compactness of a protein and quantify the topological connectivity of atoms or residues in the proteins and nucleic acids. Physically, FRI characterizes the total interaction strength at each atom or residue and thus it reflects the atomic rigidity and flexibility. Additionally, we define the total rigidity of a molecule by a summation of atomic rigidities. A practical validation of the proposed FRI for flexibility analysis is provided by the prediction of B-factors, or

temperature factors of proteins, measured by X-ray crystallography. We employ a set of 263 proteins to examine the validity, explore the reliability and demonstrate the robustness of the proposed FRI method for B-factor and/or flexibility prediction. We analyze the performance of two classes of correlation kernels, specifically the exponential type and the Lorentz type, for B-factor prediction. The exponential type of correlation kernel involves two parameters, exponential order and characteristic length. The Lorentz type of correlation kernel also involves two parameters, power order and characteristic length. By searching the parameter space for optimal predictions, parameter-free correlation kernels are obtained. It is found that the parameter-free correlation kernel of the Lorentz type is able to retain about 95% accuracy compared to the optimized results.

After validation of the basic FRI method we introduced a multikernel-based multiscale FRI (mFRI) strategy to analyze macromolecular flexibility. The essential idea is to employ two or three kernels each parameterized with a different scale to capture the multiple characteristic interaction scales of complex biomolecules. Based on an expanded test set containing 364 proteins, we show that the mFRI method is about 20% more accurate than the GNM method in B-factor prediction. Additionally, we demonstrate that the present mFRI gives rise to excellent flexibility analysis for many proteins that are difficult cases for GNM and the previously introduced single-scale FRI methods. Finally, for a protein of N residues, we illustrate that the computational complexity of the proposed mFRI is of linear scaling $\mathcal{O}(N)$, in contrast to the order of $\mathcal{O}(N^3)$ for GNM.

An increased interest in large systems of macromolecular complexes is what requires and inspires the latest advances in the FRI methods. FRI has proven to be well suited to make calculations on scales relevant to current biochemical and biophysical research. In particular, fFRI boasts a computational complexity on the scale of $\mathcal{O}(N)$, meaning that it far outpaces alternative tools such as GNM. Additionally, FRI has been previously demonstrated

to maintain superior accuracy to previous methods even at such efficient computational complexity. Now FRI's utility has been extended to the nucleic acid domain-enabling study of many important biological systems such as the RNA polymerase example featured in this paper. Due to the unique formulation of FRI and aFRI we were able to analyze a complex system for biologically relevant details that cannot be accessed by global methods or time-dependent methods

A contributing factor for FRI's increased efficiency compared to existing methods is that GNM and NMA are essentially global methods in a sense that they rely on the solution of the global eigenvalue problem to predict local atomic properties, e.g., B-factors. In contrast, FRI is a local method and utilizes the local geometric information to predict local atomic properties. In parallel, there are (global) band theory of solids and (local) atomic orbital model of solids. The former is good for describing many global physical properties such as electrical conductivity and thermal lattice motions in terms of excitations, while the latter is more powerful for explaining localized chemical reactivity and catalysis of solids.

One of the major drawbacks of GNM is the poor scaling with the number of residues or atoms in the system. The matrix diagonalization of normal modes methods is of $\mathcal{O}(N^3)$ computational complexity, where N is the number of residues. The computational complexity of the original FRI algorithm is of $\mathcal{O}(N^2)$. In the present work, we propose a fast FRI (fFRI) algorithm, which further reduces the computational complexity of FRI to $\mathcal{O}(N)$. Both FRI and fFRI do not involve the time consuming matrix decomposition. As a result, it takes less than 30 seconds for the fFRI method to predict the B-Factors of an HIV virus structure with more than three hundred thousands of residues, which would require many years for GNM to compute. Additionally, both the exponential-based parameter-free fFRI and the Lorentz-based parameter-free fFRI are about 10% more accurate than the GNM in the B-factor prediction of 364 proteins.

Anisotropic motions between protein domains are known to correlate with protein functions. To describe protein anisotropic fluctuations, we also introduce anisotropic FRI (aFRI) algorithms. We introduce an adaptive aFRI method that partitions the molecule into many clusters with variable sizes. We specifically examine two extreme cases, a one-cluster partition and an N-cluster partition, which result in a single completely global $3N \times 3N$ Hessian matrix and N completely localized 3×3 Hessian matrices, respectively. The computational complexity of aFRI varies from $\mathcal{O}(N^3)$ to $\mathcal{O}(N)$. Although aFRI Hessian matrices can be completely local, they still contain much non-location correlation. As such, all of three protein modes predicted by the completely local aFRI exhibit highly collective global motions. The eigenmodes obtained from the completely global aFRI closely resemble those of the anisotropic network model (ANM).[3,6] However, modes constructed from the completely local aFRI show different collective motion patterns. Since there is no analytical solution for collective motions, it is not possible to judge whose collective motions are more correct. In general, the eigenmodes of ANM and the completely global aFRI exhibit a slightly better synergistic effect than modes generated by using the completely local aFRI.

In addition to the quantitative aspects, the proposed FRI has a few visual applications. First, the correlation maps of the FRI are capable of revealing both short- and long-distance interactions or connectivity. Since correlation map elements are directly related to the original distances by a known radial basis function, the distances can be labeled on the map as well. Additionally, the predicted B-factors can be plotted as the radii of residues to visualize the amplitude of thermal fluctuations. This plot becomes even more interesting when atomic spheres are colored with the electrostatics.[75] The close correlation between flexibility and large electrostatic potentials can be unveiled, which sheds light on intrinsic protein structural properties. Moreover, the predicted B-factors can be plotted with secondary structures to have an overall picture of structural flexibility. Finally, as continuous functions, the atomic

rigidity function and atomic flexibility function can be projected onto protein molecular surfaces or other surface representations to analyze flexibility.

Another application of FRI and aFRI is the analysis of protein domains. Existing methods, such as GNM and ANM, are well known for domain analysis. The present FRI provides a clear correlation map for domain identifications. It is found that aFRI gives rise to highly collective domain motion patterns, although not all parts of a domain move uniformly in aFRI modes of motion.

Protein-nucleic acid complexes are essential to all living organisms. The function of these complexes depends crucially on their flexibility, an intrinsic property of a macromolecule. However, for many large protein-nucleic acid complexes, such as ribosomes and RNA polymerases, the present flexibility analysis approaches can be problematic due to their computational complexity scaling of $\mathcal{O}(N^3)$ and neglecting multiscale effects.

Therefore we also introduce the Flexibility-rigidity index (FRI) methods parameterized[49,50,75] for the flexibility analysis of protein-nucleic acid structures. We show that a multiscale FRI (mFRI) realized by multiple kernels parameterized at multiple length scales is able to significantly outperform the Gaussian network model (GNM) for the B-factor prediction of a set of 64 protein-nucleic acid complexes.[78] The FRI methods are not only accurate, but also efficient, as their computational complexity scales as $\mathcal{O}(N)$. Additionally, anisotropic FRI (aFRI), which has cluster Hessian matrices, offers collective motion analysis for any given cluster, i.e, subunit or domain in a biomolecular complex.

We can apply FRI methods to a large ribosomal subunit (1YIJ) with multiple subunits. We note that both original single-scale FRI and GNM do not work well for this structure. It is found that the multiscale strategy is crucial for the flexibility analysis of multi-subunit structures. The correlation coefficients between FRI predictions and experimental B-factors for 1YIJ improve from 0.3 for single-scale FRI to 0.85 for multiscale FRI. We further use the

fitting coefficients obtained from 1YIJ to predict the flexibility of an entire ribosome, 4V4J. We found that mFRI has an advantage for analyzing large biomolecular complexes due to both higher speeds and accuracy.

We have also demonstrated the utility of the anisotropic FRI (aFRI) for analyzing the translocation of an RNA polymerase, which involves protein, DNA, RNA, nucleotide substrates and various ions. Both experimental and computational studies of RNA polymerases are difficult and expensive due to the size and complexity of the biomolecular complex. The molecular mechanism of RNA polymerase translocation is an interesting, open research topic. The present work makes use of localized aFRI to elucidate the synergistic local motions of a bacterial RNA polymerase. These findings are consistent with those from much more expensive molecular dynamics simulations and normal mode analysis.[23,24]

Also, to clarify the relationship between normal modes methods and FRI, we construct a series of generalized Gaussian network models (gGNMs). We show that the original Kirchhoff matrix used in GNM can be constructed by using the ideal low-pass filter (ILF), which is a special case of a family of admissible correlation kernels (or functions) used in FRI. Based on this connection, we propose a unified framework to construct generalized Kirchhoff matrices for both GNM and FRI. More specifically, the inverse of the generalized Kirchhoff matrices leads to infinitely many gGNMs and the direct inverse of the diagonal terms gives rise to FRI. We reveal the identical behavior between gGNM and FRI at a large cutoff distance or characteristic scale for B-factor protein predictions. Additionally, we propose multiscale Gaussian network models (mGNMs) based on the relationship of GNM and FRI. Essentially, we develop a two-step procedure to construct mGNMs. In the first step, we utilize mFRI to come up with an optimal combination of multiscale kernels. In the second step, we try to implement the same combination of multiscale kernels in the generalized Kirchhoff matrices for mGNMs. However, this step is not unique because for a given Kirchhoff matrix,

GNM and FRI are connected only through diagonal elements. Two types schemes, Type-1 mGNM and Type-2 mGNM, are proposed in this work. Moreover, we propose multiscale anisotropic network models (mANMs) based on the similarity between ANM and GNM and the connection between GNM and FRI. Since ANM is typically less accurate than GNM in B-factor prediction,[49,52] its main utility is for collective motion analysis. We therefore have developed mANMs to maintain the physical connectivity of protein atoms in the Kirchhoff matrix.

We have carried out intensive numerical experiments to validate the proposed gGNM, mGNM and mANM methods for B-factor predictions. The gGNM method is examined over a set of 364 proteins. It is found that the proposed gGNM is about 10% more accurate than GNM in B-factor prediction. For mGNM, we use only a set of 362 proteins due to limited computer resources. We show that mGNM can achieve about 13% improvement over GNM. Similarly, the proposed mANM is about 11% more accurate than its counterpart, ANM, in B-factor prediction over a set of 300 proteins. Further, we consider three types of applications of the proposed mGNM and mANM methods. One type of application is to analyze the flexibility of proteins that fail the original GNM method in various ways. We employ four proteins to demonstrate the advantage of the proposed mGNM in flexibility analysis. Another application is the study of protein domain separations. The first nontrivial eigenmode of the multiscale Kirchhoff matrix is used. We found from the analysis of two proteins that Type-1 mGNM does a good job in domain analysis while Type-2 mGNM does not work for this purpose. The other application concerns the protein's collective motions. mANM is found to offer similar results to those of the original ANM method.

It is important to note that the mGNM and mANM methods are not limited to the examples shown in this work. The design of new mGNM and mANM methods is still an open problem. Essentially, we hope these new methods are efficient, accurate and robust.

More specifically, high accuracy in B-factor prediction is a main criterion. Additionally, having the ability to provide correct protein domain analysis is a desirable property as well. For mANM, the capability of offering correct motion analysis is a major requirement. The quality of both domain and motion analyses depends on how to design non-diagonal matrix elements so as to properly reflect the physical connectivity among particles. In the future, we will carefully consider the present mANM for other interesting applications, namely anisotropic B-factors[22] and conformational changes.[63]

The study of hinges has been an important topic and much research has been done in the past.[21,25,26,39,59] Identification of hinge residues is useful for inferring motion and function when molecules are too large for MD simulation on relevant timescales. Other methods, such as GNM and NMA have been utilized. FRI-based methods could place a significant role in hinge analysis. In tests so far, gGNM mode-based hinge predictions are at least partially correct for all of the structures analyzed using the automatic, simple analysis method we propose. Furthermore, with some human interpretation along with consideration of the second mode, it is possible to positively identify almost every single hinge in the test set. Feature ranking results demonstrate that many of the molecular characteristics that are useful in other machine learning models, such as SVM-based hot spot predictors, are not useful predictors of hinging. Features based on flexibility, sequence and secondary structure have little correlation to hinging residues based on F-score. The best SVM model that could be produced uses gGNM mode-based predictions as the only feature, and the predictions from these models show only minor differences in sensitivity and specificity compared to the predictions taken directly from observing gGNM mode-based features alone. Therefore, unless other features are found that are more predictive, gGNM mode-based predictions are just as useful as any SVM model built upon them. In this study we have laid out a framework for a machine learning model for hinge detection. Unfortunately, none of the

features we have tried so far improve the model significantly beyond a model based purely on hinge prediction from gGNM modes. Nevertheless, there is still the possibility for other features we have not yet tested to improve this model.

5.2 Future directions

One of the most important future goals for FRI is to make the software easy to access and easy to use. A major step toward completing this goal is the development of a web server for the various FRI tools. Development has begun on a web-based tool that allows users to run FRI tools for flexibility, hinge and anisotropic motion predictions. The web tool accommodates any standard format PDB file of proteins and/or nucleic acids. In addition to the web server version of the FRI tools, we plan to host various executable files for FRI tools for the Windows and Linux platforms as well as the source code. Finally, we plan to create plug-ins for the PyMol and Visual Molecular Dynamics programs to enable quick FRI tool access within these popular tools. Hopefully, with increased accessibility, FRI methods will completely replace normal modes methods as the most popular tool for calculating flexibility and long-time dynamics of macromolecules.

Anisotropic B-factors provide a possible opportunity for further validation of the anisotropic FRI method. Unfortunately, the number of structures with anisotropic B-factors is much lower than structures with isotropic B-factors. Additionally, there are very few tools that are designed to read anisotropic B-factors from PDB format files. In the near future we plan to test aFRI against all of the PDB structures that include anisotropic B-factor values.

We also plan to further pursue implementation of FRI based features in machine learning models. Although this initial attempt at improving a machine learning model with FRI was not met with much success, we believe there are other applications where FRI may be of use. In particular, flexibility is known to play a role in small molecule binding and protein-nucleic acid binding. Therefore we are attempting to use FRI flexibility and gGNM mode

calculations to improve machine learning models for these applications. Work has begun on a protein-nucleic binding model inspired by the DBSI model from Julie Mitchell at the University of Wisconsin. We aim to include FRI-based features and to improve the accuracy of the electrostatics calculations in such models to improve the model's overall prediction accuracy.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Hui-wang Ai, J Henderson, S Remington, and R Campbell. Directed evolution of a monomeric, bright and photostable version of clavularia cyan fluorescent protein: structural characterization and applications in fluorescence imaging. *Biochem. J*, 400:531–540, 2006.

[2] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford: Clarendon Press, 1987.

[3] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505 – 515, 2001.

[4] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman. Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability. *Phys. Rev. Lett*, 80:2733 – 2736, 1998.

[5] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2:173 – 181, 1997.

[6] A Bakan, L. M. Meireles, and I. Bahar. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27:1575 –1577, 2011.

[7] Joel R Bock and David A Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

[8] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[9] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262267, Apr 2000.

[10] Michael F. Brown. Curvature Forces in Membrane Lipid-Protein Interactions. *Biochemistry*, 51(49):9782–9795, DEC 11 2012.

[11] Michael PS Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Manuel Ares, and David Haussler. Support vector machine classification of microarray gene expression data. *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09*, 1999.

[12] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26(1):514, 2001.

[13] Zachary F Burton. The old and new testaments of gene regulation: Evolution of multisubunit rna polymerases and co-evolution of eukaryote complexity with the rnap ii ctd. *Transcription*, 5(3), 2014.

[14] Yu-Dong Cai, Xiao-Jun Liu, Xue-biao Xu, and Guo-Ping Zhou. Support vector machines for predicting protein structural class. *BMC bioinformatics*, 2(1):1, 2001.

[15] F. Chiti and C. M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333 – 366, 2006.

[16] Q. Cui and I. Bahar. *Normal mode analysis: theory and applications to biological and chemical systems*. Chapman and Hall/CRC, 2010.

[17] Christopher Davies, Stephen W White, and V Ramakrishnan. The crystal structure of ribosomal protein l14 reveals an important organizational component of the translational apparatus. *Structure*, 4(1):55–66, 1996.

[18] Omar N. A. Demerdash and Julie C. Mitchell. Density-cluster NMA: A new protein decomposition technique for coarse-grained normal mode analysis. *Proteins:Structure Function and Bioinformatics*, 80(7):1766–1779, JUL 2012.

[19] Chris HQ Ding and Inna Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.

[20] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM, 1998.

[21] Ugur Emekli, Schneidman-Duhovny, Dina, Haim Wolfson, Ruth Nussinov, and Turkan Haliloglu. HingeProt: automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227, 2008.

[22] Eran Eyal, Chakra Chennubhotla, Lee-Wei Yang, and Ivet Bahar. Anisotropic fluctuations of amino acids in protein structures: insights from x-ray crystallography and elastic network models. *Bioinformatics*, 23(13):i175–i184, 2007.

[23] Michael Feig and Zachary F Burton. Rna polymerase ii flexibility during translocation from normal mode analysis. *Proteins: Structure, Function, and Bioinformatics*, 78(2):434–446, 2010.

[24] Michael Feig and Zachary F Burton. Rna polymerase ii with open and closed trigger loops: active site dynamics and nucleic acid translocation. *Biophysical journal*, 99(8):2577–2586, 2010.

[25] Samuel Flores and Mark Gerstein. FlexOracle: predicting flexible hinges by identification of stable domains. *BMC bioinformatics*, 8(1), 2007.

[26] Samuel Flores, Long Lu, Julie Yang, Nicholas Carriero, and Mark Gerstein. Hinge atlas: relating protein sequence to sites of structural flexibility. *BMC bioinformatics*, 8, 2007.

[27] P. J. Flory. Statistical thermodynamics of random networks. *Proc. Roy. Soc. Lond. A,*, 351:351 – 378, 1976.

[28] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[29] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci.*, 80:3696 – 3700, 1983.

[30] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[31] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417 – 429, 1998.

[32] K. Hinsen. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24:521 – 528, 2008.

[33] Sujun Hua and Zhirong Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of molecular biology*, 308(2):397–407, 2001.

[34] W. Humphrey, A. Dalke, and K. Schulten. VMD – visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[35] Lakshminarayan M Iyer, Eugene V Koonin, and L Aravind. Evolutionary connection between the catalytic subunits of dna-dependent rna polymerases and eukaryotic rna-dependent rna polymerases and the origin of rna polymerases. *BMC structural biology*, 3(1):1, 2003.

[36] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins-Structure, Function, and Genetics*, 44(2):150–165, AUG 1 2001.

[37] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98 Lecture Notes in Computer Science*, page 137142, 1998.

[38] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[39] Kevin S Keating, Samuel C Flores, Mark B Gerstein, and Leslie A Kuhn. StoneHinge: hinge prediction by network analysis of individual protein structures. *Protein Science*, 18(2):359–371, 2009.

[40] D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and Jr. G. N. Phillips. Protein structural variation in computational models and crystallographic data. *Structure*, 15:169 – 177, 2007.

[41] S. Kundu, J. S. Melton, D. C. Sorensen, and Jr. G. N. Phillips. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.*, 83:723 – 732, 2002.

[42] S. Kundu, D. C. Sorensen, and G. N. Jr. Phillips. Automatic domain decomposition of proteins by a Gaussian network model. *Proteins: Structure, Function, and Bioinformatics*, 57(4):725–733, 2004.

[43] Christina S Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific symposium on biocomputing*, volume 7, pages 566–575, 2002.

[44] M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181(3):423 – 447, 1985.

[45] H. Y. Li, Z. X. Cao, L. L. Zhao, and J. H. Wang. Analysis of conformational motions and residue fluctuations for escherichia coli ribose-binding protein revealed with elastic network models. *International Journal of Molecular Sciences*, 14(5):10552–10569, 2013.

[46] J. P. Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13:373 – 180, 2005.

[47] Mikhail V Matz, Arkady F Fradkov, Yulii A Labas, Aleksandr P Savitsky, Andrey G Zaraisky, Mikhail L Markelov, and Sergey A Lukyanov. Fluorescent proteins from nonbioluminescent anthozoa species. *Nature biotechnology*, 17(10):969–973, 1999.

[48] Sayan Mukherjee, P Tamayo, D Slonim, A Verri, T Golub, J Mesirov, and T Poggio. Support vector machine classification of microarray data. 1999.

[49] K. Opron, K. L. Xia, and G. W. Wei. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *Journal of Chemical Physics*, 140:234105, 2014.

[50] Kristopher Opron, K. L. Xia, and G. W. Wei. Communication: Capturing protein multiscale thermal fluctuations. *Journal of Chemical Physics*, 142(211101), 2015.

[51] Xiao-Yong Pan and Hong-Bin Shen. Robust Prediction of B-Factor Profile from Sequence Using Two-Stage SVR Based on Random Forest Feature Selection. *Protein and Peptide Letters*, 16(12):1447–1454, 2009.

[52] J. K. Park, Robert Jernigan, and Zhijun Wu. Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations. *Bulletin of Mathematical Biology*, 75:124 –160, 2013.

[53] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. *Proceedings of the fifth annual international conference on Computational biology - RECOMB '01*, 2001.

[54] A. J. Rader, C. Chennubhotla, L. W. Yang, I. Bahar, and Q. Cui. The Gaussian network model: Theory and applications. *Normal mode analysis: Theory and applications to biological and chemical systems*, 9:41–64, 2006.

[55] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker. Protein flexibility and intrinsic disorder. *Protein Sci.*, 13:71 – 80, 2004.

[56] R. J. Renka. Multivariate interpolation of large sets of scattered data. *ACM Transactions on Mathematical Software*, 14(2):139–148, JUN 1988.

[57] Wouter H Roos, Melissa M Gibbons, Anton Arkhipov, Charlotte Uetrecht, NR Watts, PT Wingfield, Alasdair C Steven, Albert JR Heck, Klaus Schulten, William S Klug, and Gijs JL Wuite. Squeezing protein shells: How continuum elastic models, molecular dynamics simulations, and experiments coalesce at the nanoscale. *Biophysical Journal*, 99:1175– 1181, 2010.

[58] David Sept and Fred C. MacKintosh. Microtubule Elasticity: Connecting All-Atom Simulations with Continuum Mechanics. *Physical Review Letters*, 104(1), Jan 8 2010.

[59] Maxim Shatsky, Ruth Nussinov, and Haim J Wolfson. FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *Journal of Computational Biology*, 11(1):83–8106, 2004.

[60] Osamu Shimomura, Frank H Johnson, and Yo Saiga. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea. *Journal of cellular and comparative physiology*, 59(3):223–239, 1962.

[61] L. Skjaerven, S. M. Hollup, and N. Reuter. Normal mode analysis for proteins. *Journal of Molecular Structure: Theochem.*, 898:42 – 48, 2009.

[62] G. Song and R. L. Jernigan. vgnm: a better model for understanding the dynamics of proteins in crystals. *J. Mol. Biol.*, 369(3):880 – 893, 2007.

[63] F. Tama and Y. H. Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, 14:1 – 6, 2001.

[64] M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwidedi, and S. Krimm. Normal vibrations of proteins: Glucagon. *Biopolymers*, 21:711 – 714, 1982.

[65] William I. Thacker, Jingwei Zhang, Layne T. Watson, Jeffrey B. Birch, Manjula A. Iyer, and Michael W. Berry. Algorithm 905: SHEPPACK: Modified Shepard Algorithm for Interpolation of Scattered Multivariate Data. *ACM Transactions on Mathematical Software*, 37(3), SEP 2010.

[66] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905 – 1908, 1996.

[67] V. Uversky and A. K. Dunker. Controlled chaos. *Sceince*, 322:1340 – 1341, 2008.

[68] A. Uyar, N. Kantarci-Carsibasi, T. Haliloglu, and P. Doruker. Features of large hinge-bending conformational transitions. prediction of closed structure from open state. *Biophysical Journal*, 106(12):2656–2666, 2014.

[69] Adam Van Wynsberghe, Guohui Li, and Qiang Cui. Normal-mode analysis suggests protein flexibility modulation throughout rna polymerase's functional cycle. *Biochemistry*, 43(41):13083–13096, 2004.

[70] E Villa, A Balaeff, L Mahadevan, and K Schulten. Multiscale method for simulating protein-DNA complexes. *Multiscale Modeling & Simulation*, 2(4):527–553, 2004.

[71] C. W. von der Lieth, K. Stumpf-Nothof, and U. Prior. A bond flexibility index derived from the constitution of molecules. *Journal of Chemical Information and Computer Science*, 36:711–716, 1996.

[72] Dong Wang, David A Bushnell, Kenneth D Westover, Craig D Kaplan, and Roger D Kornberg. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell*, 127(5):941–954, 2006.

[73] G. W. Wei. Wavelets generated by using discrete singular convolution kernels. *Journal of Physics A: Mathematical and General*, 33:8577 – 8596, 2000.

[74] K. L. Xia, X. Feng, Y. Y. Tong, and G. W. Wei. Persistent homology for the quantitative prediction of fullerene stability. *Journal of Computational Chemsitry*, 36:408–422, 2015.

[75] K. L. Xia, K. Opron, and G. W. Wei. Multiscale multiphysics and multidomain models — Flexibility and rigidity. *Journal of Chemical Physics*, 139:194109, 2013.

[76] K. L. Xia and G. W. Wei. A stochastic model for protein flexibility analysis. *Physical Review E*, 88:062709, 2013.

[77] L. W. Yang and C. P. Chng. Coarse-grained models reveal functional dynamics–I. elastic network models–theories, comparisons and perspectives. *Bioinformatics and Biology Insights*, 2:25 – 45, 2008.

[78] Lee-Wei Yang, A Rader, Xiong Liu, Cristopher Jursa, Shann Chen, Hassan Karimi, and Ivet Bahar. oGNM: online computation of structural dynamics using the gaussian network model. *Nucleic acids research*, 34(Web Server issue):W24–W31, 2006.

[79] Lei Yang, Guang Song, and Robert L. Jernigan. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America* , 106(30):12347–12352, JUL 28 2009.

[80] Chen-Hsiang Yeang, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Ryan M Rifkin, Michael Angelo, Michael Reich, Eric Lander, Jill Mesirov, and Todd Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17(suppl 1):S316–S322, 2001.

[81] Z Yuan, TL Bailey, and RD Teasdale. Prediction of protein B-factor profiles. *Proteins-Structure Function and Bioinformatics*, 58(4):905–912, MAR 1 2005.

[82] W. Zheng, B. R. Brooks, and D. Thirumalai. Allosteric transitions in the chaperonin groel are captured by a dominant normal mode that is most robust to sequence variations. *Biophys. J.*, 93:2289 – 2299, 2007.

[83] Xiaolei Zhu and Julie C Mitchell. Kfc2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. *Proteins: Structure, Function, and Bioinformatics*, 79(9):2671–2683, 2011.