Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM)

Kelin Xia,[1] Kristopher Opron,[2] and Guo-Wei Wei[3,a)]

[1]*Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, USA*
[2]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, USA*
[3]*Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio 43210, USA*

Gaussian network model (GNM) and anisotropic network model (ANM) are some of the most popular methods for the study of protein flexibility and related functions. In this work, we propose generalized GNM (gGNM) and ANM methods and show that the GNM Kirchhoff matrix can be built from the ideal low-pass filter, which is a special case of a wide class of correlation functions underpinning the linear scaling flexibility-rigidity index (FRI) method. Based on the mathematical structure of correlation functions, 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. With this connection, we further introduce two multiscale elastic network models, namely, multiscale GNM (mGNM) and multiscale ANM (mANM), which are able to incorporate different scales into the generalized Kirchhoff matrices or generalized Hessian matrices. We validate our new multiscale methods with extensive numerical experiments. We illustrate that gGNMs outperform the original GNM method in the B-factor prediction of a 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. More importantly, we reveal intrinsic multiscale behavior in protein structures. The proposed mGNM and mANM are able to capture this multiscale behavior and thus give rise to a significant improvement of more than 11% in B-factor predictions over the original GNM and ANM methods. We further demonstrate the benefits of our mGNM through the B-factor predictions of many proteins that fail the original GNM method. We show that the proposed mGNM can also be used to analyze protein domain separations. Finally, we showcase the ability of our mANM for the analysis of protein collective motions. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4936132]

I. INTRODUCTION

Under physiological conditions, proteins undergo constant motions, ranging from atomic thermal fluctuation, side-chain rotation, residue swiveling, to domain swirling. Protein motion strongly correlates with protein functions, including molecular docking,[11] drug binding,[1] allosteric signaling,[6] self assembly,[33] and enzyme catalysis.[14] The range of protein motions in a cellular environment depends on the structure's local flexibility, an intrinsic property of a given protein structure. Protein flexibility is reflected by the Debye-Waller factor (B-factor), i.e., the atomic mean-square displacement, obtained in structure determination by x-ray crystallography, NMR, or single-molecule force experiments.[9] However, the B-factor cannot absolutely quantify flexibility—it also depends the crystal environment, solvent type, data collection condition, and structural refinement procedure.[19,24]

The flexibility of a biomolecule can be assessed by molecular dynamics (MD),[34] normal mode analysis (NMA),[5,16,26,45]

graph theory,[21] and elastic network models (ENMs),[2–4,18,28,43] including Gaussian network model (GNM)[3,4] and anisotropic network model (ANM).[2] NMA can be regarded as time-independent MD[37] and diagonalizes the MD potential to obtain a set of eigenvalues and eigenvectors, where the first few eigenvectors predict the collective, global motions, which are potentially relevant to biomolecular functionality. NMA with only the elasticity potential, which is a leading term in the MD potential, was introduced by Tirion[46] and was extended to the network setting in ANM.[2] Here, network refers to the connectivity between particles regardless of their chemical bonds.[13] Yang and Chng[60] demonstrated that due to its network setting, GNM is about one order more efficient than most other flexibility approaches in computational complexity. In terms of B-factor prediction, GNM is typically more accurate than ANM.[35,37] Therefore, GNM has been widely used in the study of biomolecular structure, dynamics, and functions.[7,32,41,60] It has demonstrated its utilities in stability analysis,[31] docking simulation,[15] viral capsids[39,42] and domain motions of hemoglobin,[58] F1 ATPase,[8,63] chaperonin GroEL,[23,62] and the ribosome.[44,49]

In traditional elasticity network models, i.e., ENM, GNM, and ANM, the connectivity is determined by a fixed cutoff

---

a)On leave from Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, USA. Author to whom correspondence should be addressed. Electronic mail: wei@math.msu.edu

distance. All atoms within the cutoff distance are treated equally with no consideration given to effects which scale with distance. In this manner, the behavior of these methods typically depends on the cutoff distance used. Many modified models are proposed to improve the hard cutoff distance practice by the incorporation of distance information.[60] Hinsen has changed the spring constant to a distance-dependent Gaussian function.[18] Riccardi *et al.* have used an inverse 6th power function of distance as the spring coefficient in their elastic network model.[40] A parameter-free ANM has been introduced by using the inverse 2nd power square distance for spring constant.[59]

A common feature of all the aforementioned approaches is that they all depend on the mode decomposition of the potential matrix, which typically has the computational complexity of $O(N^3)$, where N is the order the potential matrix. To bypass the matrix diagonalization, researchers explore the flexibility properties using the local packing density. Many elegant methods and algorithms have been proposed, including the local density model (LDM) by Halle,[17] the local contact model (LCM) by Zhang and Brüschweiler,[61] weighted contact number (WCN) by Lin and others.[20,27,30]

Recently, we have proposed a few mode-decomposition free methods for flexibility analysis, namely, molecular nonlinear dynamics,[54] stochastic dynamics[53] and flexibility-rigidity index (FRI).[35,52] Among them, the FRI method is of $O(N^2)$ in computational complexity and has been accelerated to $O(N)$ by using the cell lists algorithm without loss of accuracy.[35] The essential idea of the FRI method is to evaluate the rigidity index or the compactness of the biomolecular (network) packing by the total correlation, a function of inter-atomic distance. Then, the flexibility index is defined as the inverse of the rigidity index.

Although the original motivation for FRI comes from the design of continuum elasticity with atomic rigidity (CEWAR), FRI shares a similarity with the LDM, LCM, and WCN. To be more specific, all of these methods make use of local packing information and are free from matrix diagonalization. However, significant distinction exists between our FRI methods and other local density based methods and it can be summarized as the following. First, other than the discrete flexibility index and discrete rigidity index, our FRI methods delivers a continuous flexibility function and a continuous rigidity function.[35,52] The continuous rigidity function, which can be regarded as the density distribution function (density estimator) of a biomolecule, plays many important roles beyond the scope of flexibility study.[55] For instance, it can be used to generate biomolecular surface representations,[56,57] which reduce to the Gaussian surface if an appropriate kernel is used. In fact, rigidity function can be applied to decipher the atomic information from the experimental electron density data.[47,51,57] Second, protein multiscale collective motions can be captured by using multiple kernels in our FRI method, called multiscale FRI or multikernel FRI (mFRI).[36] This approach significantly improve the accuracy of FRI B-factor predictions. Third, we proposed an anisotropic FRI (aFRI) model for the evaluation of biomolecular global motions. Different from traditional normal mode analysis or ANM, our aFRI allows adaptive representations, ranging

from a completely global description to a completely local representation.[35]

The objective of the present work is twofold. First, we propose a unified framework to construct generalized GNMs (gGNMs). 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 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. Our 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. Most biomolecules, particularly large macromolecules and protein complexes, have multiple characteristic length scales ranging from covalent bond, residue, secondary structure, and domain dimensions, to protein sizes. Even for small molecules, due to the influence of crystal structure, multiscale effects play a significant role in atomic thermal fluctuations. Consequently, GNM and ANM, which are typically parametrized at a single cutoff distance, often do not work well in characterizing the flexibility of molecules involving multiscale behaviors. Our essential idea 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 constructions of multikernel GNM or multikernel ANM. This approach works because for a diagonally dominant matrix, the direct inverse of the diagonal element can be approximated by the diagonal element of the inverse matrix. We demonstrate that the proposed mGNM and mANM are able to successfully capture the multiscale properties of the protein and significantly improve the accuracy of protein flexibility prediction.

The rest of this paper is organized as the follows. Section II is devoted to methods and algorithms. We first propose a concise formulation of gGNMs using FRI correlation functions in Section II A. We show that there are infinitely many new gGNMs that reduce to the original GNM at appropriate limits of their parameters. To establish notation, we further present a brief review of our mFRI formalism in Section II B. Based on the connection between FRI and GNM, we propose mGNM in Section II C. Specifically, parameters learned from mFRI are used to construct the multiscale Kirchhoff matrix in mGNM. We discuss two types of realizations of mGNMs. As an extension of our mGNM, the mANM method is introduced in Section II D.

We validate the proposed gGNM, mGNM, and mANM by extensive numerical experiments in Section III. We illustrate that the intrinsic multiscale properties of biomolecules are successfully captured in our mGNM and mANM. Finally, in Section IV, we demonstrate the utility of mGNM and mANM for protein flexibility analysis, protein domain separation, and collective motion study. The present work offers a new strategy for the design and construction of accurate, efficient, and robust methods for biomolecular flexibility analysis. This paper ends with a conclusion.

II. METHODS AND ALGORITHMS

A. 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 $\{\mathbf{r}_i | \mathbf{r}_i \in \mathbb{R}^3, i = 1, 2, \ldots, N\}$ 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[3,4]

$$B_i^{\text{GNM}} = a\left(\Gamma^{-1}\right)_{ii}, \forall i = 1, 2, \ldots, N, \qquad (1)$$

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

$$\Gamma_{ij} = \begin{cases} -1, & i \neq j \text{ and } r_{ij} \leq r_c \\ 0, & i \neq j \text{ and } r_{ij} > r_c \\ -\sum_{j, j \neq i}^{N} \Gamma_{ij}, & i = j \end{cases}, \qquad (2)$$

where r_c is a cutoff distance. The GNM theory evaluates the matrix inverse by $\left(\Gamma^{-1}\right)_{ii} = \sum_{k=2}^{N} \lambda_k^{-1} [\mathbf{u}_k \mathbf{u}_k^T]_{ii}$, 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 eigenmode whose eigenvalue is zero.

The FRI prediction of the ith B-factor of the biomolecule can be given by[35,52]

$$B_i^{\text{FRI}} = a \frac{1}{\sum_{j, j \neq i}^{N} w_j \Phi(r_{ij}; \eta)} + b, \forall i = 1, 2, \ldots, N, \qquad (3)$$

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:

$$\Phi(r_{ij}; \eta) = 1 \quad \text{as} \quad r_{ij} \to 0, \qquad (4)$$

$$\Phi(r_{ij}; \eta) = 0 \quad \text{as} \quad r_{ij} \to \infty, \qquad (5)$$

where η is a scale parameter. Delta sequences of the positive type[50] are good choices. Many radial basis functions are also admissible.[35,52] Commonly used FRI correlation functions include the generalized exponential functions

$$\Phi(r_{ij}; \eta, \kappa) = e^{-(r_{ij}/\eta)^\kappa} \quad \kappa > 0, \qquad (6)$$

and generalized Lorentz functions

$$\Phi(r_{ij}; \eta, \upsilon) = \frac{1}{1 + (r_{ij}/\eta)^\upsilon}, \quad \upsilon > 0. \qquad (7)$$

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 our fast FRI (fFRI).[35] 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, let us examine the parameter limits of generalized exponential functions (6) and generalized Lorentz functions (7),

$$e^{-(r_{ij}/\eta)^\kappa} \to \Phi(r_{ij}; r_c) \quad \text{as } \kappa \to \infty, \qquad (8)$$

$$\frac{1}{1 + (r_{ij}/\eta)^\upsilon} \to \Phi(r_{ij}; r_c) \text{ as } \upsilon \to \infty, \qquad (9)$$

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

$$\Phi(r_{ij}; r_c) = \begin{cases} 1, & r_{ij} \leq r_c \\ 0, & r_{ij} > r_c \end{cases}. \qquad (10)$$

Relations (8) and (9) unequivocally connect FRI correlation functions to the GNM Kirchhoff matrix. It is important to examine whether the ILF is still a FRI correlation function. Mathematically, the ILF is a special real-valued monotonically decreasing correlation function and also satisfies admissibility conditions (4) and (5). In fact, all FRI correlation functions are low-pass filters as well. Therefore, both GNM and FRI admit low-pass filters in their constructions. Indeed, GNM is very special in the sense that there is only one unique ILF, while, there are infinitely many other low-pass filters. Figure 1 illustrates the behavior and relation of the above low-pass filters or correlation functions. Clearly, the ILF is completely
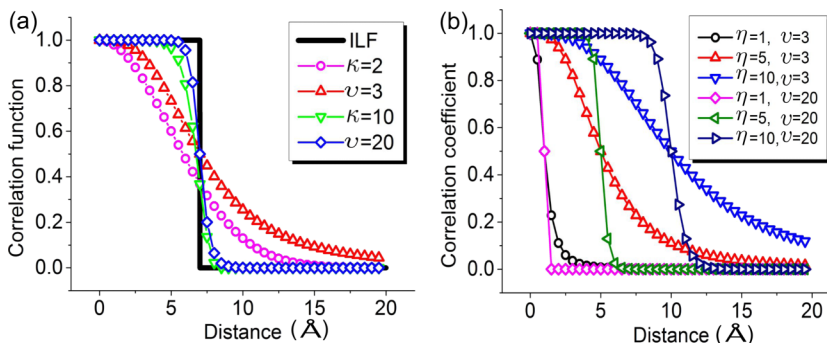

FIG. 1. Illustration of admissible correlation functions. (a) Correlation functions approach the ILF as $\kappa \to \infty$ or $\upsilon \to \infty$ at $\eta = 7$ Å. (b) Effects of varying scale value η. Local correlation is obtained with large υ and small η values. Whereas, nonlocal correlation is generated by small υ and large η values.

localized for any given cutoff value. In general, generalized exponential function and generalized Lorentz function are delocalized 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 nonlocal correlations. Our earlier test indicates that $\upsilon = 3$ and $\eta = 3$ Å provide a good flexibility analysis for a set of 364 proteins.[35]

To further bring to light the mathematical foundation of the GNM and FRI methods, we consider a generalized Kirchhoff matrix[53,54]

$$
\Gamma_{ij}(\Phi) = \begin{cases} -\Phi(r_{ij};\eta), & i \neq j \\ -\sum_{j,j\neq i}^{N} \Gamma_{ij}(\Phi), & i = j \end{cases}, \tag{11}
$$

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 a 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 striking difference between the gGNM and FRI methods is that to predict B-factors, the gGNM seeks a matrix inverse of Kirchhoff matrix (2), whereas, the FRI takes the direct inverse of the diagonal elements of generalized Kirchhoff matrix (11).

B. mFRI

Due to the widely existed multiscale in biomolecules, especially large macromolecules and protein complexes, the mFRI method is proposed to better capture nonlocal collective motions.[36] In this approach, two or three correlation kernels that are parametrized at multiple scales are employed simultaneously to characterize protein multiscale properties. The flexibility index can be expressed as,

$$
f_i^n = \frac{1}{\sum_{j,j\neq i}^{N} w_j^n \Phi^n(r_{ij};\eta^n)}, \forall i = 1,2,\ldots,N, \tag{12}
$$

where w_j^n, $\Phi^n(r_{ij};\eta^n)$, and η^n are the corresponding quantities associated with the nth kernel. The essence of the mFRI is to minimize of the following form:

$$
\text{Min}_{a^n,b} \left\{ \sum_i \left| \sum_n a^n f_i^n + b - B_i^e \right|^2 \right\}, \tag{13}
$$

where B_i^e are the experimental B-factors for the ith particle. We choose kernels with various scale parameters and obtain the optimized fitting coefficients.

Specifically, for a coarse-grained network model with only C_α atoms, we can set $w_j^n = 1$ and choose a single type of kernel function parametrized at different scales. The predicted B-factors can be expressed as

$$
B_i^{\text{mFRI}} = b + \sum_{n=1} \frac{a^n}{\sum_{j,j\neq i}^{N} \Phi(r_{ij};\eta^n)}, \forall i = 1,2,\ldots,N. \tag{14}
$$

Unlike the scheme in Eq. (12) where various types of kernels can be chosen, we only select one type of kernels in Eq. (14). In this way, the multiscale nature in biomolecules can be clearly demonstrated. Some commonly used kernel functions

include generalized Lorentz kernel,

$$
\Phi(\|\mathbf{r} - \mathbf{r}_j\|;\eta^n) = \frac{1}{1 + (\|\mathbf{r} - \mathbf{r}_j\|/\eta^n)^3}, \tag{15}
$$

and the generalized exponential kernel,

$$
\Phi(\|\mathbf{r} - \mathbf{r}_j\|;\eta^n) = e^{-\frac{\|\mathbf{r}-\mathbf{r}_j\|}{\eta^n}}. \tag{16}
$$

These kernels define a continuous multiscale rigidity function by using the fitting coefficients from the minimization process as following:

$$
\mu^n(\mathbf{r}) = \sum_{j=1}^{N} w_j^n \Phi(\|\mathbf{r} - \mathbf{r}_j\|;\eta^n). \tag{17}
$$

This expression can be used to construct new protein surfaces. Similarly, we can also construct a continuous multiscale flexibility function,

$$
f(\mathbf{r}) = b + \sum_{n=1} \frac{a^n}{\sum_{j=1}^{N} w_j^n \Phi(\|\mathbf{r} - \mathbf{r}_j\|;\eta^n)}. \tag{18}
$$

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

C. mGNM

The essential component for our 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 II A, we make use of the coefficients approximated from our FRI to construct a multiscale Kirchhoff matrix. In this section, we present two types of algorithms to construct mGNM.

1. Type-1 mGNM

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

$$
\Gamma = \sum_n a^n \Gamma^n, \tag{19}
$$

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 $\Phi^n(r_{ij};\eta^n)$) parametrized at an appropriate scale η^n. We use our mFRI 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

$$
\text{Min}_{a^n} \left\{ \sum_i \left| \sum_n a^n \Gamma_{ii}^n - \frac{1}{B_i^e} \right|^2 \right\}, \tag{20}
$$

assuming that $B_i^e > 0$. With the multiscale Kirchhoff matrix given in Eq. (19), we carry our routine GNM analysis as described in Eq. (1).

2. Type-2 mGNM

Another algorithm for constructing mGNM is to make use of fitting coefficients from mFRI directly via the relation

ALGORITHM 1. Type-2 mGNM multiscale Kirchhoff matrix.

Input: $\Gamma_{ii}, i = 1, 2, \ldots, N$ ▷ 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 $j \leftarrow i+1, N$
       $\Gamma_{ij} = \frac{\Gamma_{ii} - sum}{N-i}$ ▷ We equally distribute the diagonal terms into non-diagonal parts.
       $\Gamma_{ji} = \Gamma_{ij}$ ▷ Use the symmetry property.
    end for
end for

---

between biomolecular local packing density and its flexibility. Basically, we choose several kernels parametrized at various scales and evaluate the best fitting coefficients, i.e., $\{a_n\}$ and b, with the experimental B-factors using Equation (13). The resulting multiscale flexibility index is then used to construct the generalized Kirchhoff matrix as following:

$$\sum_n a^n f_i^n + b = \frac{1}{\Gamma_{ii}}, \forall i = 1, 2, \ldots, N. \qquad (21)$$

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

$$\Gamma_{ii} = \frac{1}{\sum_n \frac{a^n}{\mu_i^n} + b}, \forall i = 1, 2, \ldots, N. \qquad (22)$$

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

$$\Gamma_{ii} = \frac{\mu_i^1 \mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b \mu_i^1 \mu_i^2}, \forall i = 1, 2, \ldots, N. \qquad (23)$$

Now the problem is to determine the non-diagonal terms of our 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 our mGNM matrix can also be expressed as

$$\Gamma_{ii} = \sum_{j,j\neq i} \frac{\{w_j^1 \Phi^1(r_{ij}; \eta^1)\} \mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b \mu_i^1 \mu_i^2}, \forall i = 1, 2, \ldots, N. \qquad (24)$$

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

$$\Gamma_{ij} = \begin{cases} -\dfrac{\{w_j^1 \Phi^1(r_{ij}; \eta^1)\} \mu_i^2}{a^1 \mu_i^2 + a^2 \mu_i^1 + b \mu_i^1 \mu_i^2}, & i \neq j \\ -\sum_{j,j\neq i}^N \Gamma_{ij}, & i = j \end{cases}. \qquad (25)$$

The problem with the matrix in Eq. (25) is that the resulting multiscale Kirchhoff matrix is not symmetric, which may lead to computational difficulty. To avoid non-symmetric matrix, we further propose an alternative construction to preserve the symmetry of the matrix.

Our basic idea is to determine the diagonal terms Γ_{ii} from Eq. (22) and then on each row, equally distribute the diagonal term into the non-diagonal parts, under condition that the resulting matrix remains symmetric. To this end, we propose an iterative scheme as shown in Algorithm 1.

It also should be noticed that in the construction of our Type-2 mGNM, only the diagonal terms are fixed and determined from the 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 to 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 1 as our Type-2 mGNM.

D. mANM

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

$$H_{ij}^n = -\frac{\Phi^n(r_{ij}; \eta^n)}{r_{ij}^2} \begin{bmatrix} (x_j - x_i)(x_j - x_i) & (x_j - x_i)(y_j - y_i) & (x_j - x_i)(z_j - z_i) \\ (y_j - y_i)(x_j - x_i) & (y_j - y_i)(y_j - y_i) & (y_j - y_i)(z_j - z_i) \\ (z_j - z_i)(x_j - x_i) & (z_j - z_i)(y_j - y_i) & (z_j - z_i)(z_j - z_i) \end{bmatrix} \forall i \neq j. \qquad (26)$$
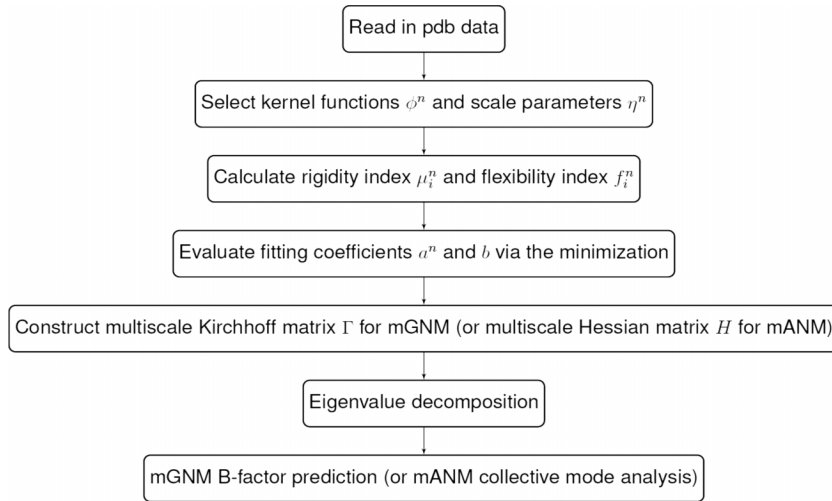
FIG. 2. Work flow of basic procedure in mGNM and mANM.

Note that Hinsen[18] 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 $= -\sum_{i \neq j} H_{ij}^n, \forall i = 1, 2, \ldots, N$. Basically, it is the summation of all the non-diagonal local matrices.

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

$$\mu_i^n = \sum_{i \neq j} \frac{\Phi^n(r_{ij}; \eta^n)}{r_{ij}^2}[(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2]$$
$$= \sum_{i \neq j} \Phi^n(r_{ij}; \eta^n), \forall i = 1, 2, \ldots, N. \quad (27)$$

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

We adopt the approach used in Type-1 mGNM to construct mANM. We propose a multiscale Hessian matrix as $H = \sum_n a^n H^n$, and the coefficients a^n should be evaluated from

$$\text{Min}_{a^n} \left\{ \sum_i \left| \sum_n a^n \mu_i^n - \frac{1}{B_i^e} \right|^2 \right\}. \quad (28)$$

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

To clarify the proposed multiscale Gaussian network model and multiscale anisotropic network model, we present a flow chart to illustrate their basic procedure in Fig. 2.

III. VALIDATIONS

A. The performance of generalized Gaussian network models

1. Comparison between gGNM and FRI

Based on the analysis in Section II A, it is straightforward to construct correlation function based gGNMs via the matrix inverse of generalized Kirchhoff matrix (11), which leads to infinitely many new gGNMs, including the original GNM as a special limiting case. It is also possible to construct the FRI by using the Kirchhoff matrix, which gives rise to a unique FRI. Questions arise as what are the relative performance of these correlation function based gGNM and FRI methods. Another question is whether there is any further relation between these two distinguished approaches. Specifically, what is the relation between the diagonal elements of the gGNM matrix inverse and the FRI direct inverse of the diagonal elements, for a given generalized Kirchhoff matrix? To answer these questions, we select two representative correlation functions, i.e., the Lorentz ($\upsilon = 3$) and ILF functions to construct generalized Kirchhoff matrix (11). The Lorentz function is a typical example for many correlation functions studied in our earlier work.[35] In contrast, the ILF function is an extreme case of FRI correlation functions. Resulting two generalized Kirchhoff matrices (11) can be used for calculating the gGNM matrix inverse or the inverse diagonal elements of the FRI matrix. This results in possible combinations or methods, namely, FRI-Lorentz, FRI-ILF, GNM-Lorentz, and GNM-ILF. Performances of these methods are carefully analyzed.

To answer the above mentioned questions, we first employ a protein from pathogenic fungus Candida albicans (Protein Data Bank ID: 2Y7L) with 319 residues as shown in Fig. 3(a) to explore the aforementioned four methods. We consider the coarse-grained C_α representation of protein 2Y7L. We denote $B^{GNM-ILF}$, $B^{FRI-ILF}$, $B^{GNM-Lorentz}$, and $B^{FRI-Lorentz}$, 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 employed for a comparison. The Pearson product-moment correlation coefficient (PCC)
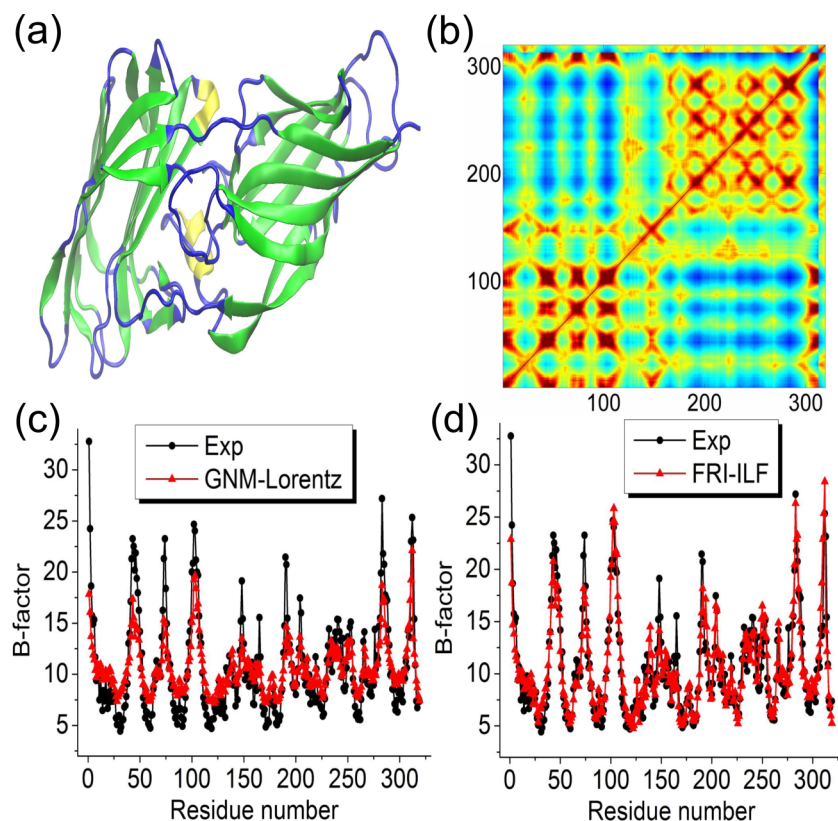
FIG. 3. Illustration of protein 2Y7L. (a) Structure of protein 2Y7L having two domains; (b) correlation map generated by using GNM-Lorentz indicating two domains; (c) comparison of experimental B-factors and those predicted by GNM-Lorentz ($\eta = 16$ Å); (d) comparison of experimental B-factors and those predicted by FRI-ILF ($r_c = 24$ Å).

is used to measure the strength of the linear relationship or dependence between each two sets of B-factors. To evaluate the performance of four methods, we compute the PCCs between predicted B-factors and experimental B-factors. Since performance of these methods depends on their parameters, i.e., the cutoff distance (r_c) in the ILF or the scale value (η) in the Lorentz function, the theoretical B-factors are computed over a wide range of r_c and η values.

Figure 4 depicts PCCs between various B-factors for protein 2Y7L. As shown in Fig. 4(a), the cutoff distance r_c of the ILF is varied from 5 Å to 64 Å. The PCCs between $B^{\text{GNM-ILF}}$ and B^{Exp}, and between $B^{\text{FRI-ILF}}$ and B^{Exp}, indicate that both GNM-ILF and FRI-ILF are able to provide accurate predictions of the experimental B-factors. Their best predictions are attained around $r_c = 24$ Å, which is

significantly larger than the commonly used GNM cutoff distance of 7-9 Å.

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 20 Å. 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 $B^{\text{GNM-ILF}}$ and $B^{\text{FRI-ILF}}$ over the same range of cutoff distances. As shown in Fig. 4(a), there is a strong linear dependence between $B^{\text{GNM-ILF}}$ and $B^{\text{FRI-ILF}}$ for $r_c \geq 10$ Å. To understand
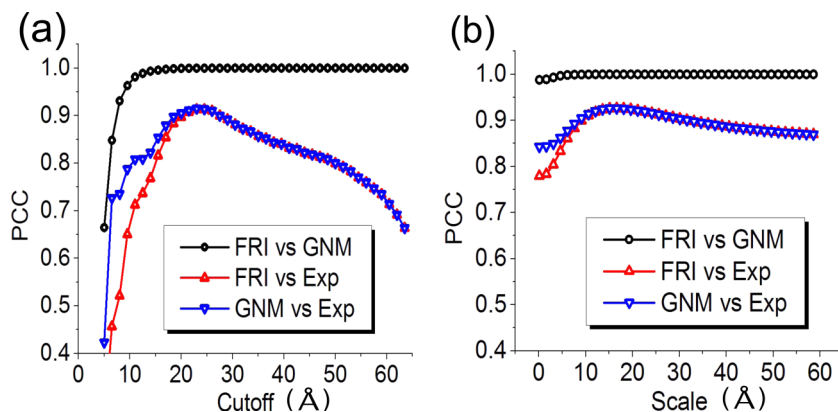

FIG. 4. PCCs between various B-factors for protein 2Y7L. (a) Correlations between $B^{\text{GNM-ILF}}$ and B^{Exp}, between $B^{\text{FRI-ILF}}$ and B^{Exp}, and between $B^{\text{GNM-ILF}}$ and $B^{\text{FRI-ILF}}$; (b) correlations between $B^{\text{GNM-Lorentz}}$ and B^{Exp}, between $B^{\text{FRI-Lorentz}}$ and B^{Exp}, and between $B^{\text{GNM-Lorentz}}$ and $B^{\text{FRI-Lorentz}}$.
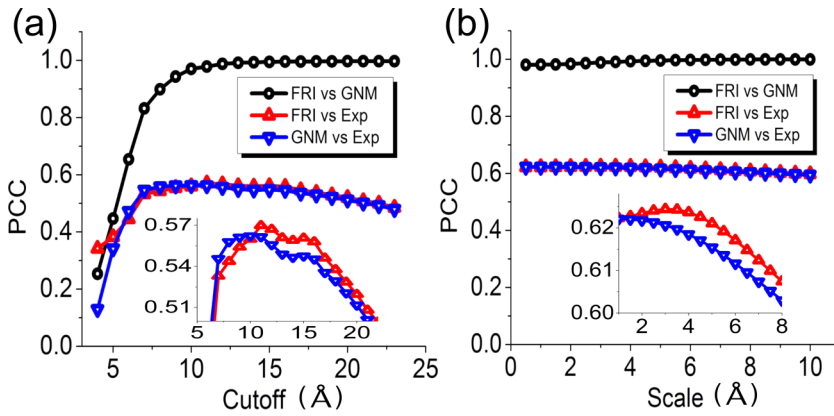
FIG. 5. PCCs between various B-factors averaged over 364 proteins. (a) Correlations between $B^{\text{GNM-ILF}}$ and B^{Exp}, between $B^{\text{FRI-ILF}}$ and B^{Exp}, and between $B^{\text{GNM-ILF}}$ and $B^{\text{FRI-ILF}}$; (b) correlations between $B^{\text{GNM-Lorentz}}$ and B^{Exp}, between $B^{\text{FRI-Lorentz}}$ and B^{Exp}, and between $B^{\text{GNM-Lorentz}}$ and $B^{\text{FRI-Lorentz}}$.

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

$$\left(\Gamma^{-1}(\Phi(r_{ij}; r_c \to \infty))\right)_{ii} = \frac{N-1}{N^2}, \qquad (29)$$

and the FRI inverse of the ith diagonal element

$$\frac{1}{\sum_{j,j\neq i}^{N} \Phi(r_{ij}; r_c \to \infty)} = \frac{1}{N-1}. \qquad (30)$$

These results elucidate the strong asymptotic correlation between $B^{\text{GNM-ILF}}$ and $B^{\text{FRI-ILF}}$ in Fig. 4(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, i.e., 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 are illustrated in Fig. 4(b) for the scale value η from 0.5 Å to 64 Å. First, it is seen that the GNM-Lorentz is a successful new approach. In fact, it outperforms the original GNM for the peak PCCs. A comparison of the predicted B-factors and the experimental B-factors is plotted in Figs. 3(c) and 3(d) for GNM-Lorentz and FRI-ILF, respectively. It is seen that $B^{\text{FRI-ILF}}$ more closely matches the experimental B-factors than $B^{\text{GNM-Lorentz}}$ does due to the different fitting schemes employed by two methods as shown in Eqs. (1) and (3), respectively.

As shown in Fig. 4(b), the predictions from GNM-Lorentz and FRI-Lorentz become identical as $\eta \geq 5$ Å. A strong correlation between $B^{\text{GNM-Lorentz}}$ and $B^{\text{FRI-Lorentz}}$ is revealed at an even smaller scale value. This behavior leads us to speculate a general relation

$$\left(\Gamma^{-1}(\Phi(r_{ij}; \eta))\right)_{ii} \longrightarrow \frac{c}{\sum_{j,j\neq i}^{N} \Phi(r_{ij}; \eta)}, \quad \eta \to \infty, \quad (31)$$

where c is a constant. Relation (31) 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. (29) and (30).

Relation (31) is a very interesting and powerful result not only for sake of understanding GNM and FRI methods but also for the design of accurate and efficient new methods.

It should be noticed that our findings are consistent with the previous finding[38] 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.

3. Validation of gGNM with extensive experimental data

It remains to prove that the above findings from a single protein are translatable and verifiable to a large class of biomolecules. To this end, we consider a set of 364 proteins, which is a subset of the 365 proteins utilized and documented in our earlier work.[35] 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 5 illustrates our results. Figure 5(a) plots the results of the ILF implemented in both GNM and FRI methods with the cutoff distance varied from 4 Å to 23 Å. Figure 5(b) depicts similar results obtained by using the Lorentz function implemented in two methods. The scale value is explored over the range of 0.5 Å–10 Å. We summarize these results from several aspects as following.

First, the proposed new method, GNM-Lorentz, is very accurate for the B-factor prediction of 364 proteins as shown in Fig. 5(b). The best GNM-Lorentz prediction is about 10.7% better than that of the original GNM shown in Fig. 5(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 practically valuable. Similarly, FRI-Lorentz is also about 10% more accurate than FRI-ILF in the 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.
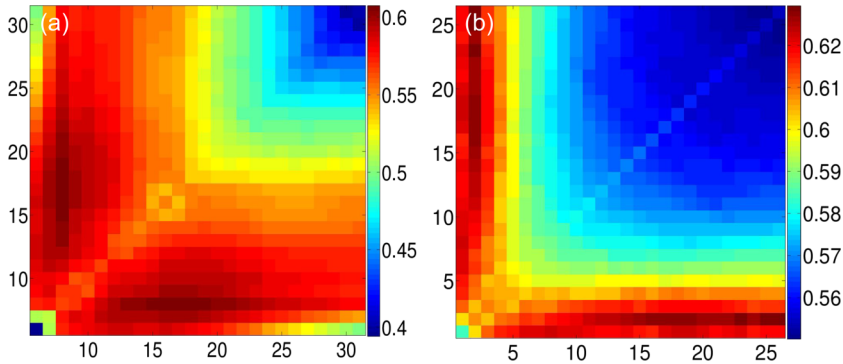
FIG. 6. The average PCCs over 362 proteins for Type-1 mGNM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two exponential kernels and their scales η are systematically varied in the range of [1 Å, 26 Å].
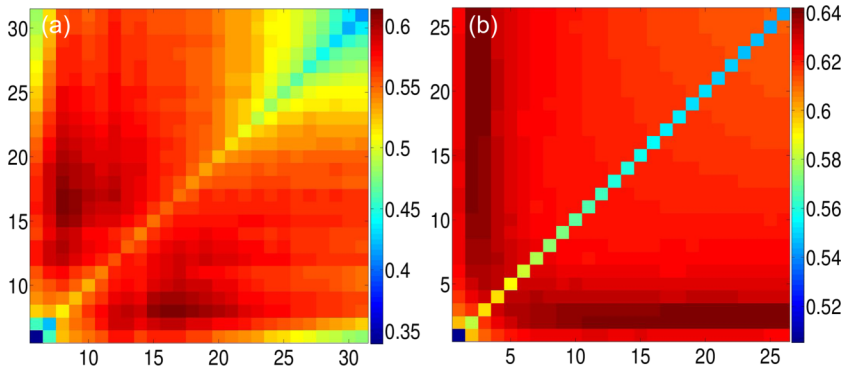

FIG. 7. The average PCCs over 362 proteins for Type-2 mGNM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two exponential kernels and their scales η are systematically varied in the range of [1 Å, 26 Å].

Additionally, the FRI-Lorentz method is able to attain the best average prediction for 364 proteins among four methods as shown in the zoomed in parts in Fig. 5(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. 5. This finding offers a solid confirmation of Eq. (31). Therefore, correlation function based gGNMs, including the original GNM as a special case, is 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,[35] 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 s, respectively, for the FRI-ILF, FRI-Lorentz, GNM-ILF, and GNM-Lorentz. The test is performed on a cluster with 8 Intel Xeon 2.50 GHz CPUs and 128 GB memory. In fact, gGNM methods are very fast for small proteins as well. Most of the accumulated gGNM CPU times are due to the computation of three largest proteins (i.e., 1F8R, 1H6V, and 1QKI) in the test set.

Finally, it is worth mentioning that the earlier FRI rigidity index includes the contribution from the self correlation, i.e., the diagonal term.[35,52] The present findings do not change if the summation in generalized Kirchhoff matrix (11) is modified to include the diagonal term and then the calculation of gGNM matrix inverse is modified to include the contribution from first eigen mode, i.e., $\left(\Gamma^{-1}\right)_{ii} = \sum_{k=1}^{N} \lambda^{-1}\left[\mathbf{u}_k\mathbf{u}_k^T\right]_{ii}$. In fact,

this modification makes the generalized Kirchhoff matrix less singular and faster converging.

B. The performance of multiscale Gaussian network models

1. Type-1 mGNM

We validate our two types of mGNM with various parameter values over a set of 362 proteins. Two largest proteins, i.e., 1H6V and 1QKI, are removed from our earlier data set of 364 proteins[35] due to the limited computational resources. Two kinds of kernels, i.e., 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 ILF kernel based test, the cutoff distances in both kernels vary from 5 Å to 31 Å. For exponential kernel based test, we set $\kappa = 1$ and vary η in both kernels within the range of [1 Å, 26 Å]. The PCCs with experimental B-factors are averaged over 362 proteins. The results for Type-1 mGNM are demonstrated in Figures 6(a) and 6(b). When two ILF kernels are used in Figure 6(a), we can seen that the largest average PCCs are concentrated around the

TABLE I. 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

TABLE II. 64 large-sized proteins in the 364-protein data set[35] 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						

region where two kernels have dramatically different cutoff distances, i.e., one 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 parametrized at different cutoff distances. Moreover, the best PCC is distributed around cutoff distance 7 Å, which is consistent with the optimal cutoff distance (7 Å) recommended for the traditional GNM method. Similar multiscale behavior can also be observed for exponential kernel based mGNM as demonstrated in Figure 6(b).

2. Type-2 mGNM

The results of Type-2 mGNM with ILF kernels and exponential kernels are demonstrated in Figures 7(a) and 7(b), respectively. The multiscale property is observed 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 further achieves the best average PCC of 0.614. Similar results are observed in exponential kernel models. For generalized GNM, the best average PCC is about 0.608. This has been enhanced to 0.629 in Type-1 mGNM and further improved to 0.642 in Type-2 mGNM. Detailed comparisons are summarized in Table I.

C. The performance of multiscale anisotropic network models

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 Table II. It should be noticed that the Hessian matrix used in mANM is $3N \times 3N$, which is 9 times larger than the correspondent 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. The results are plotted in Figure 8. First of all, one can still see the multiscale effect in this set of proteins. The best average PCC values for mANM are achieved at the combination of a relatively small cutoff distance (7 Å) and a relatively large cutoff distance. These values are much higher than those on the diagonal, which are the average PCC values of the traditional (single kernel) ANM. For 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 I. 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.

IV. 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 of 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
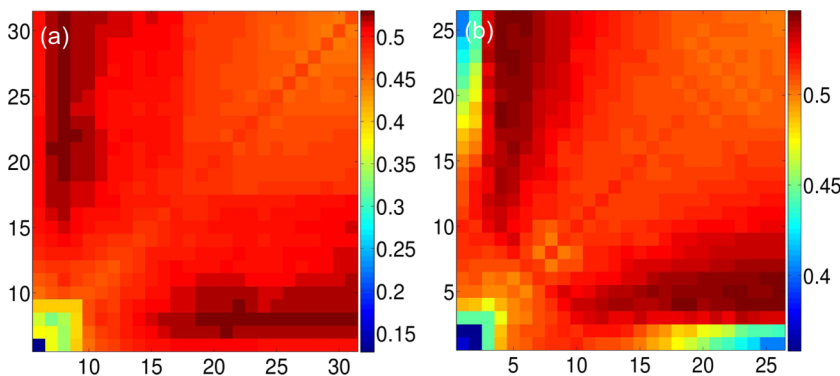

FIG. 8. The average PCCs over 300 proteins for mANM. (a) Two ILF kernels and their cutoff distances are systematically changed from 5 Å to 31 Å. (b) Two Gaussian kernels ($\kappa = 2$) and their scales η are systematically varied in the range of [1 Å, 26 Å].
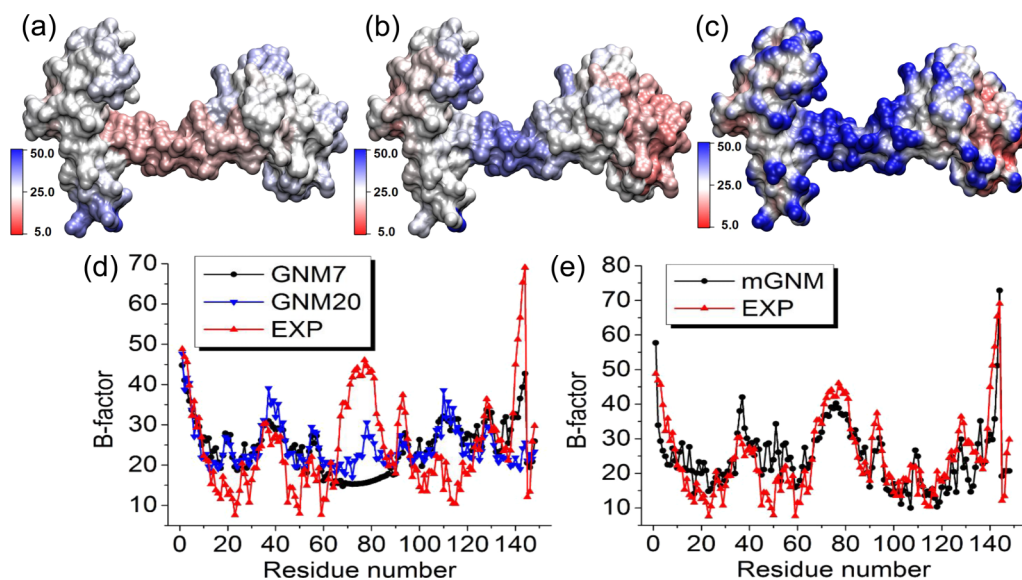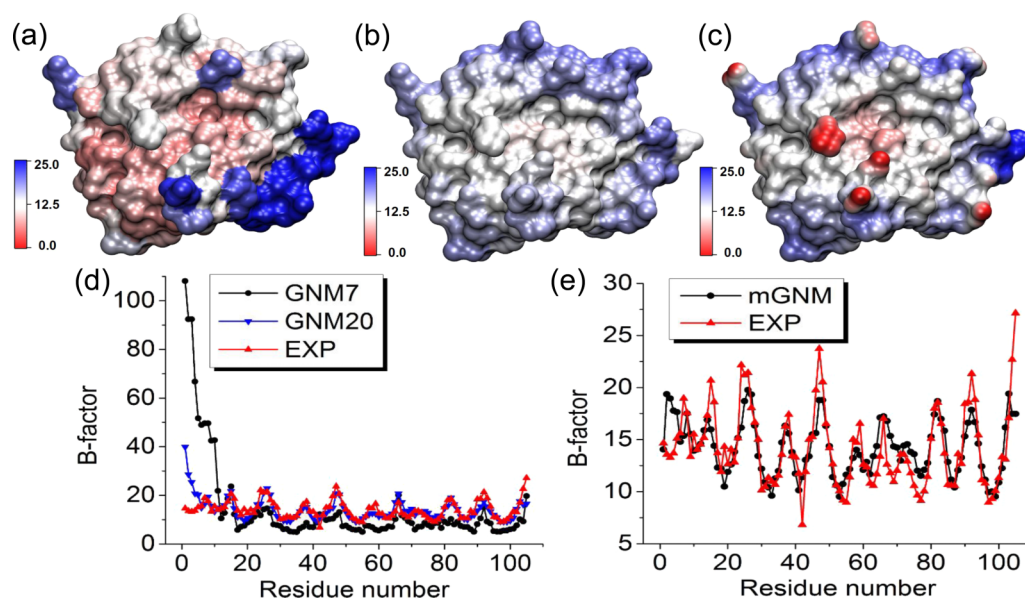
FIG. 9. The comparison between Type-2 mGNM with exponential kernel and traditional GNM for the B-factor prediction of protein 1CLL. Two scales, i.e., $\eta^1 = 3$ Å and $\eta^2 = 25$ Å, are employed in mGNM. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by our Type-2 mGNM. (c) Molecular surface colored by multiscale flexibility function in Equation (18). (d) B-factors predicted by traditional GNM with cutoff distances 7 Å (GNM7) and 20 Å (GNM20). (e) B-factors predicted by mGNM.

FIG. 10. The comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 1V70 B-factor prediction. Two scales, i.e., η^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 (18). (d) B-factors predicted by traditional GNM with cutoff distances 7 Å (GNM7) and 20 Å (GNM20). (e) B-factors predicted by mGNM.

motions of a protein near its equilibrium. These issues are investigated in this section.

A. 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.[36,37] Park *et al.* have shown that GNM PCCs with experimental B-factors can be negative.[37] In this work, we demonstrate that the present mGNM is able to deliver good B-factor predictions by capturing multiscale features. To this end, we consider four proteins, i.e., 1CLL, 1V70, 2HQK,

and 1WHI. The Type-2 mGNM with two exponential kernels is utilized in our study. As depicted in Figure 7(b), there is a wide range of scale parameters that deliver accurate B-factor predictions. We simply choose $\kappa = 1, \eta^1 = 3$ Å and $\kappa = 1, \eta^2 = 25$ Å in our studies. To draw a comparison, the traditional GNM, i.e., GNM-ILF, is employed with different cutoff distances, namely, 7 and 20 Å, which are denoted as GNM7 and GNM20, respectively.

Figures 9–12 illustrate our results. In each figure, protein surfaces are colored by B-factor values predicted by GNM7, mGNM, and the flexibility function in Eq. (18), respectively, in insets (a)–(c). The comparisons of B-factors predicted
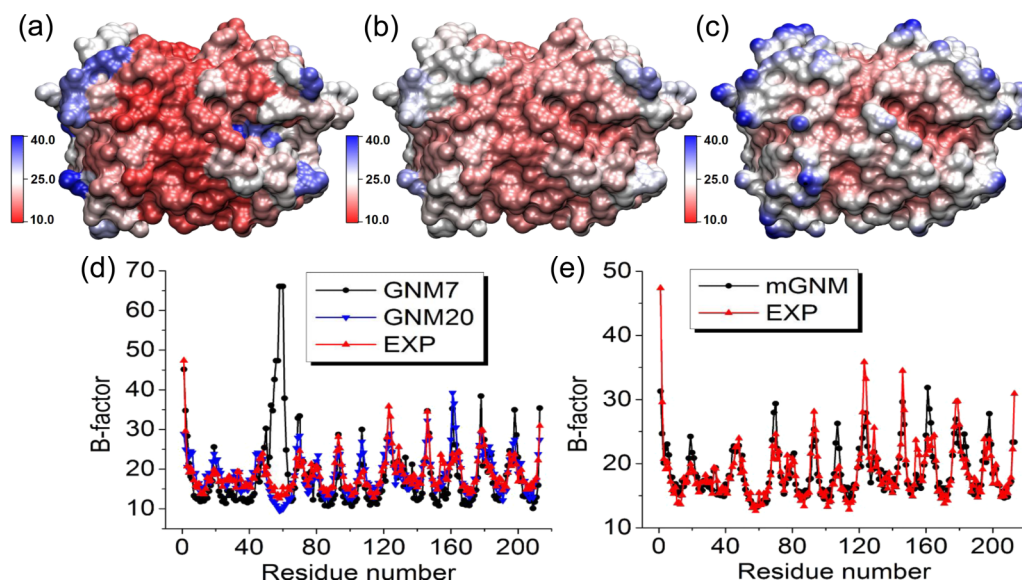
FIG. 11. The comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 2HQK B-factor prediction. Two scales, i.e., η^1 = 3 Å and η^2 = 25 Å, are used in mGNM. (a) Molecular surface colored by B-factors predicted by GNM with cutoff distance 7 Å. (b) Molecular surface colored by B-factors evaluated by our Type-2 mGNM. (c) Molecular surface is colored by multiscale flexibility function in Equation (18). (d) B-factors predicted by traditional GNM with cutoff distances 7 Å (GNM7) and 20 Å (GNM20). (e) B-factors predicted by mGNM.

by GNM7 and GNM20 with those of experiments are demonstrated in inset (d). Similarly, the comparisons of the predicted B-factors by mGNM with those of experiments are plotted in inset (e). A summary of related PCC values are listed in Table I.

Flexible hinges are important to protein function but may not be easily detected by GNM type of methods.[12,22] As shown in Figure 9, the original GNM parametrized 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, our two-kernel mGNM is able to capture the hinge behavior.

Protein 1V70 shown in Figure 10 is another difficult case for the traditional GNM method. At cutoff distance 7 Å, 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, our two-kernel mGNM provides a very good prediction.

Figure 11 illustrates one more interesting situation. The tradition GNM with cutoff distance 7 Å over-predicts the B-factors for residues near number 58. However, at a large cutoff distance of 20 Å, it is able to offer accurate results. In this case, our mGNM is able to further improve the accuracy.
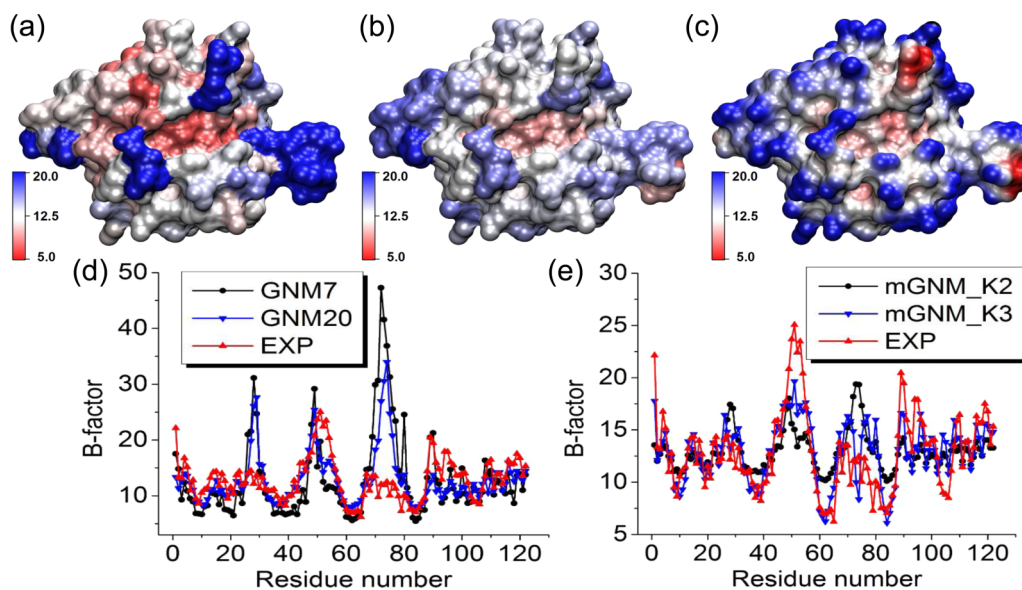

FIG. 12. The comparison between Type-2 mGNM with exponential kernel and traditional GNM for protein 1WHI B-factor prediction. Two mGNMs are used. The first one, mGNM_K2, has two exponential kernels with $\kappa = 1$, $\eta^1 = 3$ Å, and $\eta^2 = 25$ Å. The second mGNM, mGNM_K3, has an extra exponential 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 our Type-2 mGNM. (c) Molecular surface is colored by multiscale flexibility function in Equation (18). (d) B-factors predicted by traditional GNM with cutoff distances 7 Å (GNM7) and 20 Å (GNM20). (e) B-factors predicted by two mGNMs, i.e., mGNM_K2 and mGNM_K3.

TABLE III. Case study of B-factor prediction for four proteins in three different schemes, i.e., 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)

The case of 1WHI given in Figure 12 is difficult. The GNM with two different parameterizations does not work well. However, our two-kernel mGNM does not work well either. Its PCC of 0.484 is just a minor improvement of GNM values 0.270 (obtained at $r_c = 7$ Å) and 0.370 (obtained at $r_c = 20$ Å). It should be noticed that our 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 III and Figure 12, our three-kernel mGNM is able to deliver a good PCC of 0.766.

B. 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. The way to subdivide a protein is quite natural. As each particle in the protein is assigned with a value (element) from the Fiedler vector, one simply groups these particles according to their positive or negative signs. More specifically, all atoms with positive values are in the same group and those with negative values are in other group. The ones with zero values can be classified into either groups as their are usually the link region between two domains.

To test the performance of our mGNM schemes, we adopt two test proteins, i.e., 1ATN (chain A) and 3GRS, used by Kundu et al.[25] We compare the performance of our two types of mGNM. In Type-1 mGNM, we use the two 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 an extra kernel parametrized as $\kappa = 1, \eta^3 = 10$ Å. Our results are depicted in Figures 13 and 14, respectively. It can be seen that Type-1 mGNM delivers a great decomposition, which is also consistent with the prediction from traditional GNM.[25] However, the Type-2 mGNM does not produce a reasonable result. This is due to the fact that Algorithm 1 is designed to construct the symmetric Kirchhoff matrix with the required diagonal elements. Its non-diagonal elements do not properly reflect the protein connectivity.

However, we should notice that the PCCs of Type-1 mGNM for 1ATN and 3GRS are 0.460 and 0.658, respectively. Whereas, the PCCs of Type-2 mGNM for 1ATN and 3GRS are 0.660 and 0.666, respectively. These results indicate that the B-factor values are mainly dictated by the diagonal matrix elements, while the domain separation is determined by non-diagonal matrix elements, which reflect the protein connectivity in Type-1 mGNM, but have little to do with the packing geometry in Type-2 mGNM.
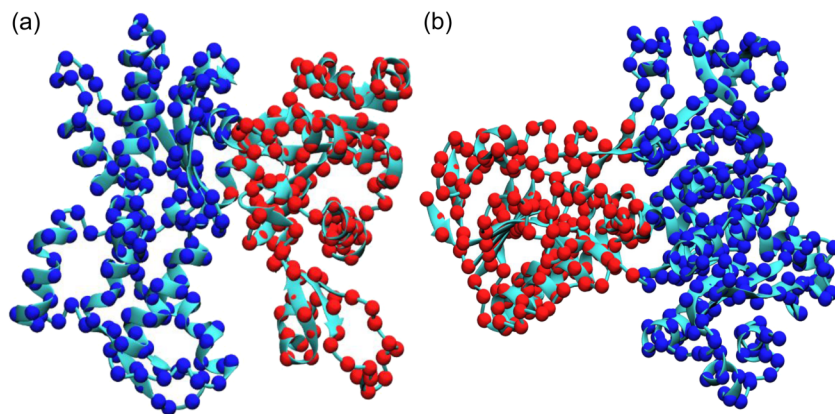

FIG. 13. Protein domain decomposition with Type-1 mGNM. The first eigenvector (Fiedler vector) is used to decompose the protein into two domains. (a) Protein 1ATN (chain A); (b) protein 3GRS.
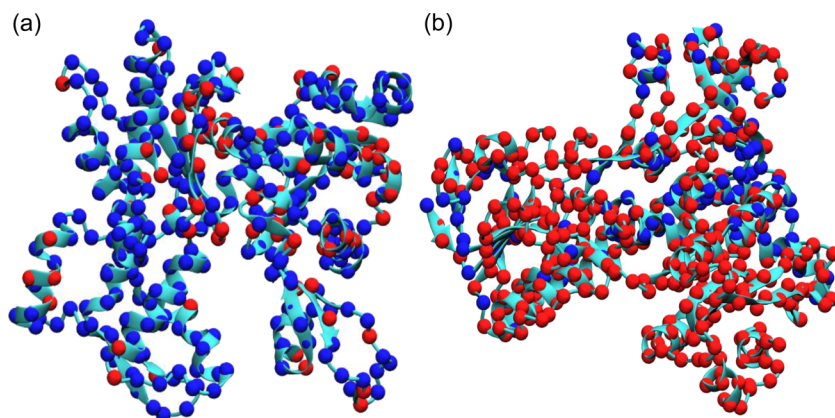

FIG. 14. Protein domain decomposition with Type-2 mGNM. The first eigenvector (Fiedler vector) is used to decompose the protein into two domains. (a) Protein 1ATN (chain A); (b) protein 3GRS. It can be seen that Type 2 mGNM fails in protein domain decomposition.
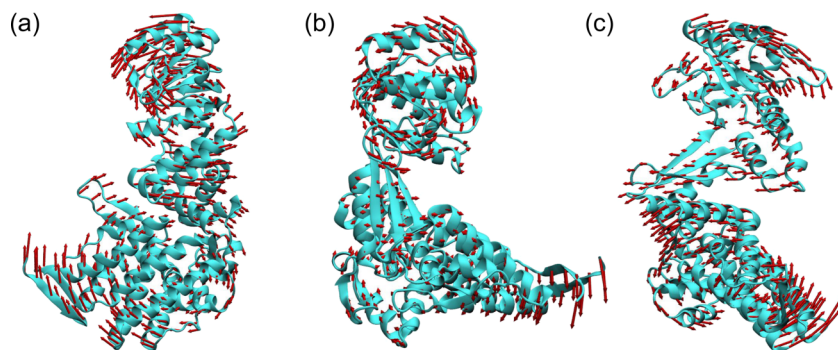
FIG. 15. The collective motions of protein 1GRU (chain A). The seventh, eighth, and ninth modes calculated from our mANM are demonstrated in (a)–(c), respectively.
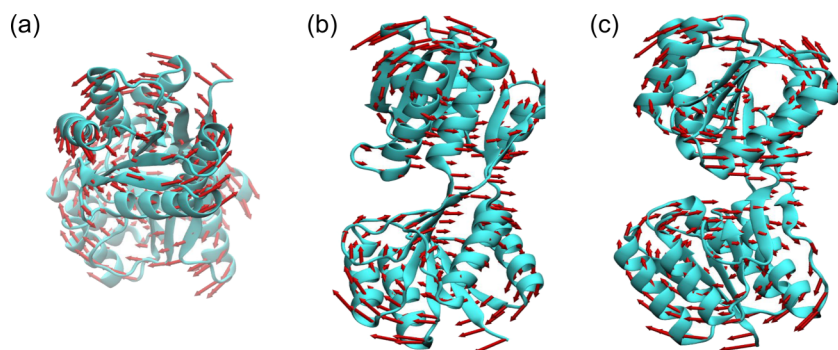

FIG. 16. The collective motions of protein 1URP (chain A). The seventh, eighth, and ninth modes calculated from our mANM are demonstrated in (a)–(c), respectively.

C. Collective motion simulation using mANM

GNM is an isotropic model which quantifies the atomic scalar fluctuations in molecule. In contrast, ANM is designed to describe the anisotropic properties, such as collective motions of a molecule near the 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 nonzero (next smallest) eigenvalues. Due to its simplicity, ANM is widely used to study the dynamics of biomolecules.

In the present work, we have designed our mANM to maintain the aforementioned properties. To validate our mANM for anisotropic mode analysis, we use two test proteins, i.e., 1GRU (chain A) and 1URP (chain A). The protein 1GRU is chaperonin GroEL, a benchmark test for ANM.[48,62] We employ our 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 illustrated in Figure 15, our mANM results are in an excellent agreement with those of ANM for chaperonin GroEL.[48,62]

To further validate our mANM, we examine another test case, 1URP. It is a ribose-binding protein and its anisotropic motions have been studied in the literature.[29] We utilize the same set of parameters described above. Figure 16 demonstrates mANM results. Our results are in a close agreement with the traditional ANM analysis.[29]

V. CONCLUSION

GNM and ANM are popular methods for macromolecular flexibility analysis. Alternative methods, FRI[35,52] and aFRI,[35]

have been introduced to achieve better accuracy and more adaptivity in our recent work. Most recently, we have further proposed mFRI[36] to capture the multiscale behavior in macromolecules. Our mFRI utilizes multiple kernels which are parametrized at different scales to describe macromolecular multiscale connectivity. We have shown that mFRI is about 20% more accurate than GNM in the B-factor prediction of a set of 364 protein.[36] Motivated by these achievements, we propose a few FRI based generalizations of GNM and ANM in this work.

First, we construct a series of gGNMs. We show that the original Kirchhoff matrix used in GNM can be constructed by using the 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 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 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,[35,37]

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 protein due to limited computer resource. 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 collective motions. Our mANM is found to offer similar results as those of the original ANM method.

It is worth pointing out that our mGNM and mANM methods are not unique. How to design optimal 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[10] and conformational changes.[43]

ACKNOWLEDGMENTS

[1]Alvarez-Garcia, D. and Barril, X., "Relationship between protein flexibility and binding: Lessons for structure-based drug design," J. Chem. Theory Comput. **10**(6), 2608–2614 (2014).

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

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

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

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

[6]Bu, Z. and Callaway, D. J., "Proteins move! Protein dynamics and long-range allostery in cell signaling," Adv. Protein Chem. Struct. Biol. **83**, 163–221 (2011).

[7]Cui, Q. and Bahar, I., *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall/CRC, 2010).

[8]Cui, Q., Li, G. J., Ma, J., and Karplus, M., "A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase," J. Mol. Biol. **340**(2), 345–372 (2004).

[9]Dudko, O. K., Hummer, G., and Szabo, A., "Intrinsic rates and activation free energies from single-molecule pulling experiments," Phys. Rev. Lett. **96**, 108101 (2006).

[10]Eyal, E., Chennubhotla, C., Yang, L. W., and Bahar, I., "Anisotropic fluctuations of amino acids in protein structures: Insights from x-ray crystallography and elastic network models," Bioinformatics **23**(13), i175-i184 (2007).

[11]Fischer, M., Coleman, R. G., Fraser, J. S., and Shoichet, B. K., "Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery," Nat. Chem. **6**, 575–583 (2014).

[12]Flores, S. and Gerstein, M., "FlexOracle: Predicting flexible hinges by identification of stable domains," BMC Bioinf. **8**(1), 215 (2007).

[13]Flory, P. J., "Statistical thermodynamics of random networks," Proc. R. Soc. London, Ser. A **351**, 351–378 (1976).

[14]Fraser, M. W., Clarkson, J. S., Degnan, S. C., Erion, R., Kern, D., and Alber, T., "Hidden alternative structures of proline isomerase essential for catalysis," Nature **462**, 669–673 (2009).

[15]Gerek, Z. N. and Ozkan, S. B., "A flexible docking scheme to explore the binding selectivity of PDZ domains," Protein Sci. **19**, 914–928 (2010).

[16]Go, N., Noguti, T., and Nishikawa, T., "Dynamics of a small globular protein in terms of low-frequency vibrational modes," Proc. Natl. Acad. Sci. U. S. A. **80**, 3696–3700 (1983).

[17]Halle, B., "Flexibility and packing in proteins," Proc. Natl. Acad. Sci. U. S. A. **99**, 1274–1279 (2002).

[18]Hinsen, K., "Analysis of domain motions by approximate normal mode calculations," Proteins **33**, 417–429 (1998).

[19]Hinsen, K., "Structural flexibility in proteins: Impact of the crystal environment," Bioinformatics **24**, 521–528 (2008).

[20]Huang, S. W., Shih, C. H., Lin, C. P., and Hwang, J. K., "Prediction of NMR order parameters in proteins using weighted protein contact-number model," Theor. Chem. Acc. **121**(3-4), 197–200 (2008).

[21]Jacobs, D. J., Rader, A. J., Kuhn, L. A., and Thorpe, M. F., "Protein flexibility predictions using graph theory," Proteins: Struct., Funct., Genet. **44**(2), 150–165 (2001).

[22]Keating, K. S., Flores, S. C., Gerstein, M. B., and Kuhn, L. A., "StoneHinge: Hinge prediction by network analysis of individual protein structures," Protein Sci. **18**(2), 359–371 (2009).

[23]Keskin, O., Bahar, I., Flatow, D., Covell, D. G., and Jernigan, R. L., "Molecular mechanisms of chaperonin GroEL-GroES function," Biochemistry **41**, 491–501 (2002).

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

[25]Kundu, S., Sorensen, D. C., and Phillips, G. N. J., "Automatic domain decomposition of proteins by a Gaussian network model," Proteins: Struct., Funct., Bioinf. **57**(4), 725–733 (2004).

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

[27]Li, D. W. and Brüschweiler, R., "All-atom contact model for understanding protein dynamics from crystallographic b-factors," Biophys. J. **96**(8), 3074–3081 (2009).

[28]Li, G. H. and Cui, Q., "A coarse-grained normal mode approach for macromolecules: An efficient implementation and application to Ca(2+)-ATPase," Bipophys. J. **83**, 2457–2474 (2002).

[29]Li, H. Y., Cao, Z. X., Zhao, L. L., and Wang, J. H., "Analysis of conformational motions and residue fluctuations for escherichia coli ribose-binding protein revealed with elastic network models," Int. J. Mol. Sci. **14**(5), 10552–10569 (2013).

[30]Lin, C. P., Huang, S. W., Lai, Y. L., Yen, S. C., Shih, C. H., Lu, C. H., Huang, C. C., and Hwang, J. K., "Deriving protein dynamical properties from weighted protein contact number," Proteins: Struct., Funct., Bioinf. **72**(3), 929–935 (2008).

[31]Livesay, D. R., Dallakyan, S., Wood, G. G., and Jacobs, D. J., "A flexible approach for understanding protein stability," FEBS Lett. **576**, 468–476 (2004).

[32]Ma, J. P., "Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes," Structure **13**(3), 373–380 (2005).

[33]Marsh, J. A. and Teichmann, S. A., "Protein flexibility facilitates quaternary structure assembly and evolution," PLoS Biol. **12**(5), e1001870 (2014).

[34]McCammon, J. A., Gelin, B. R., and Karplus, M., "Dynamics of folded proteins," Nature **267**, 585–590 (1977).

[35]Opron, K., Xia, K. L., and Wei, G. W., "Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis," J. Chem. Phys. **140**, 234105 (2014).

[36]Opron, K., Xia, K. L., and Wei, G. W., "Communication: Capturing protein multiscale thermal fluctuations," J. Chem. Phys. **142**, 211101 (2015).

[37]Park, J. K., Jernigan, R., and Wu, Z., "Coarse grained normal mode analysis vs. refined Gaussian network model for protein residue-level structural fluctuations," Bull. Math. Biol. **75**, 124–160 (2013).

[38]Rader, A. J., Chennubhotla, C., Yang, L. W., Bahar, I., and Cui, Q., "The Gaussian network model: Theory and applications," in *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall/CRC, 2006), Vol. 9, pp. 41–64.

[39]Rader, A. J., Vlad, D. H., and Bahar, I., "Maturation dynamics of bacteriophage HK97 capsid," Structure **13**, 413–421 (2005).

[40]Riccardi, D., Cui, Q., and Phillips, G. N., "Application of elastic network models to proteins in the crystalline state," Biophys. J. **96**(2), 464–475 (2009).

[41]Skjaerven, L., Hollup, S. M., and Reuter, N., "Normal mode analysis for proteins," J. Mol. Struct.: THEOCHEM **898**, 42–48 (2009).

[42]Tama, F. and Brooks III, C. K., "Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis," J. Mol. Biol. **345**, 299–314 (2005).

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

[44]Tama, F., Valle, M., Frank, J., and Brooks III, C. K., "Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy," Proc. Natl. Acad. Sci. U. S. A. **100**(16), 9319–9323 (2003).

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

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

[47]Topf, M., Lasker, K., Webb, B., Wolfson, H., and Sali, W. C. A., "Protein structure fitting and refinement guided by cryo-EM density," Structure **16**, 295–307 (2008).

[48]Uyar, A., Kantarci-Carsibasi, N., Haliloglu, T., and Doruker, P., "Features of large hinge-bending conformational transitions. Prediction of closed structure from open state," Biophys. J. **106**(12), 2656–2666 (2014).

[49]Wang, Y., Rader, A. J., Bahar, I., and Jernigan, R. L., "Global ribosome motions revealed with elastic network model," J. Struct. Biol. **147**, 302–314 (2004).

[50]Wei, G. W., "Wavelets generated by using discrete singular convolution kernels," J. Phys. A: Math. Gen. **33**, 8577–8596 (2000).

[51]Wriggers, W., Milligan, R. A., and McCammon, J. A., "Situs: A package for docking crystal structures into low-resolution maps from electron microscopy," J. Struct. Biol. **125**, 185–195 (1999).

[52]Xia, K. L., Opron, K., and Wei, G. W., "Multiscale multiphysics and multidomain models—Flexibility and rigidity," J. Chem. Phys. **139**, 194109 (2013).

[53]Xia, K. L. and Wei, G. W., "A stochastic model for protein flexibility analysis," Phys. Rev. E **88**, 062709 (2013).

[54]Xia, K. L. and Wei, G. W., "A Galerkin formulation of the MIB method for three dimensional elliptic interface problems," Comp. Math. Appl. **68**, 719–745 (2014).

[55]Xia, K. L. and Wei, G. W., "Persistent homology analysis of protein structure, flexibility and folding," Int. J. Numer. Methods Biomed. Eng. **30**, 814–844 (2014).

[56]Xia, K. L. and Wei, G. W., "Multidimensional persistence in biomolecular data," J. Comput. Chem. **36**, 1502–1520 (2015).

[57]Xia, K. L. and Wei, G. W., "Persistent topology for cryo-EM data analysis," Int. J. Numer. Methods Biomed. Eng. **31**, e02719 (2015).

[58]Xu, C., Tobi, D., and Bahar, I., "Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T ↔ R2 transition," J. Mol. Biol. **333**, 153–168 (2003).

[59]Yang, L., Song, G., and Jernigan, R. L., "Protein elastic network models and the ranges of cooperativity," Proc. Natl. Acad. Sci. U. S. A. **106**(30), 12347–12352 (2009).

[60]Yang, L. W. and Chng, C. P., "Coarse-grained models reveal functional dynamics–I. Elastic network models–theories, comparisons and perspectives," Bioinform. Biol. Insights **2**, 25–45 (2008).

[61]Zhang, F. L. and Brüschweiler, R., "Contact model for the prediction of NMR N-H order parameters in globular proteins," J. Am. Chem. Soc. **124**(43), 12654–12655 (2002).

[62]Zheng, W., Brooks, B. R., and Thirumalai, D., "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).

[63]Zheng, W. J. and Doniach, S., "A comparative study of motor-protein motions by using a simple elastic-network model," Proc. Natl. Acad. Sci. U. S. A. **100**(23), 13253–13258 (2003).