Journal of Computational Chemistry

WILEY

Journal of COMPUTATIONAL CHEMISTRY

RESEARCH ARTICLE OPEN ACCESS

Multiscale Differential Geometry Learning for Protein Flexibility Analysis

Hongsong Feng[1] | Jeffrey Y. Zhao[2] | Guo-Wei Wei[1,3,4]

[1]Department of Mathematics, Michigan State University, East Lansing, Michigan, USA | [2]Vestavia Hills High School, Vestavia Hills, Alabama, USA | [3]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, USA | [4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

Correspondence: Guo-Wei Wei (wei@math.msu.edu)

Received: 5 November 2024 | **Revised:** 6 February 2025 | **Accepted:** 13 February 2025

Funding: This work was supported by Bristol-Myers Squibb (65109), National Science Foundation (DMS-2052983, IIS-1900473), Michigan State University Foundation (100424), and National Institutes of Health (R01AI164266, R35GM148196).

Keywords: blind prediction | multiscale differential geometry | protein flexibility

ABSTRACT

Protein structural fluctuations, measured by Debye-Waller factors or B-factors, are known to be closely associated with protein flexibility and function. Theoretical approaches have also been developed to predict B-factor values, which reflect protein flexibility. Previous models have made significant strides in analyzing B-factors by fitting experimental data. In this study, we propose a novel approach for B-factor prediction using differential geometry theory, based on the assumption that the intrinsic properties of proteins reside on a family of low-dimensional manifolds embedded within the high-dimensional space of protein structures. By analyzing the mean and Gaussian curvatures of a set of low-dimensional manifolds defined by kernel functions, we develop effective and robust multiscale differential geometry (mDG) models. Our mDG model demonstrates a 27% increase in accuracy compared to the classical Gaussian network model (GNM) in predicting B-factors for a dataset of 364 proteins. Additionally, by incorporating both global and local protein features, we construct a highly effective machine-learning model for the blind prediction of B-factors. Extensive least-squares approximations and machine learning-based blind predictions validate the effectiveness of the mDG modeling approach for B-factor predictions.

1 | Introduction

Proteins are polypeptide structures made up of one or more long chains of amino acid residues. According to the well-established sequence structure function dogma [1], protein structures dictate their functions in various biological processes, including DNA replication, molecular transport, and providing structural support to cells. However, protein structures are not static; they exhibit fluctuations and thermodynamic movements under physiological conditions. These movements arise as proteins respond to external stimuli, and protein flexibility, which measures a protein's capacity to deform from its equilibrium state under external force, represents an intrinsic property of the protein structure.

Protein flexibility can be assessed through experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR), and single-molecule force experiments. In X-ray crystallography, the Debye-Waller factor or beta factor (B-factor) characterizes protein flexibility by describing the attenuation of X-ray scattering due to atomic displacements. Displacement of atoms from their mean position in a crystal structure diminishes the scattered X-ray intensity. B-factor reflects the degree to which an atom's position deviates from its average location.

The relationship between the B-factor and the mean square displacement $< u^2 >$ of an atom is given by [2]

$$B = 8\pi^2 < u^2 >$$

which indicates that a higher B-factor corresponds to a greater atomic displacement, reflecting increased flexibility. The "temperature factor" field in the ATOM records of the PDB files normally contains the B-factor. Theoretically, the fluctuation amplitude of an atom in a protein correlates with its B-factor reported during the structure determination from X-ray diffraction data. However, reported B-factors may not fully account for variations in atomic diffraction cross-sections and chemical stability during data collection. NMR serves as another crucial technique for analyzing protein flexibility, allowing for the study of flexibility across different time scales and under physiological conditions.

Given that protein flexibility is associated with significant conformational variations, reactivity, and enzymatic functions [3–5], analyzing protein flexibility is essential for understanding protein structure, function, and dynamics [6]. Understanding protein flexibility is also important for docking [7] and computational drug design [8, 9]. This necessity has driven the development of numerous theoretical approaches to predict B-factors for specific protein structures, including molecular dynamics (MD) [10], normal mode analysis (NMA) [11–14], and elastic network models (ENM) [15–20], as well as theories such as the Gaussian network model (GNM) [16, 17] and anisotropic network model (ANM) [15] over the past few decades. Although MD simulations can provide comprehensive conformational landscapes of proteins using an all-atom representation, they require extensive computational resources for long-time integrations involving a significant number of degrees of freedom. To mitigate this issue, time-independent approaches are often utilized, functioning as time-harmonic approximations of Newton's equations [21].

As one of the first time-independent B-factor prediction methods, the NMA [11–14] employs Hooke's law to create an elastic mass-and-spring network for alpha carbons (C_α). In this model, each C_α is represented as a node, with edges connecting the nodes if their Euclidean distance is below a predefined threshold. This network effectively captures local covalent and noncovalent interactions between atoms and their neighbors and can be represented by a Hamiltonian interaction matrix. The eigenvalue analysis of this matrix yields the characteristic frequencies of the protein and predicts B-factors. The performance of elastic network computations has been further enhanced by various elastic network models (ENM) [15–20]. Notably, the anisotropic network model (ANM) [15] functions as an NMA incorporating only the leading elasticity/MD potential and establishes connections between particles regardless of their chemical bonds [22]. By disregarding anisotropic motion in the ANM network, the Gaussian network model (GNM) [16, 17] accelerates B-factor analysis, being approximately one order of magnitude more efficient than most other network approaches [23]. However, eigenvalue analysis in these network approaches requires matrix diagonalization, which has a computational complexity of $\mathcal{O}(N^3)$, where N is the number of atoms in the network. This becomes computationally expensive for larger proteins, highlighting the need for more efficient flexibility analysis methods. The graph method was reported for protein flexibility analysis [24]. Sequence-based predictions and analysis of protein flexibility were also proposed [25–27]. Various machine learning approaches have also been developed for protein flexibility predictions [27–29]. AlphaFold2 and deep learning were utilized to elucidate enzyme conformational flexibility [30]. Recently, Xu et al. have proposed a method to employ both sequence information and structure information to predict protein B-factors [31].

The flexibility and rigidity index (FRI) methods [32, 33] have emerged as a matrix-decomposition-free approach for B-factor predictions. The fundamental assumption behind FRI methods is that protein flexibility and rigidity can be fully determined from the protein structure alone, eliminating the need to refer back to the protein interaction Hamiltonian, thereby bypassing costly matrix diagonalization. As a geometric graph approach, FRI methods construct a distance matrix using radial basis functions to non-linearly scale the distance from atom to atom [34]. This allows for the evaluation of the rigidity index of each atom, which reflects the compactness of biomolecular packing. Consequently, the inverse of the rigidity index yields the flexibility index, correlating to the B-factor for each C_α atom [32, 33]. The original FRI [32] has a computational complexity of $\mathcal{O}(N^2)$, while the fast FRI (fFRI) [33] can be accelerated to $\mathcal{O}(N)$ using a cell list algorithm. Parameter-free fFRI has demonstrated approximately 10% greater accuracy than GNM for a set of 365 proteins, while being orders of magnitude faster [33]. To capture multiscale atomic interactions, a multiscale flexibility rigidity index (mFRI) method [35] has been developed, employing multiple radial basis functions or kernels with varying parameterizations, resulting in significant improvements in the accuracy of FRI B-factor predictions. In addition, the anisotropic FRI (aFRI) model has been introduced for the analysis of high-quality anisotropic motions of biomolecules [33].

Topological data analysis (TDA) methods have also been applied to protein flexibility analysis. In [36], atom-specific persistent homology was constructed as a local atomic-level representation of a molecule using a global topological tool. The resulting atom-specific topological features were integrated with machine learning algorithms for B-factor predictions. Furthermore, evolutionary homology (EH) was introduced in [37] based on time evolution-based filtration and topological persistence. By coupling with dynamical systems or chaotic oscillators, the corresponding EH captures the time-dependent topological invariants of macromolecules, which makes it applicable to protein flexibility analysis. Pun et al. reported machine learning-based prediction of RNA flexibility using weighted persistent homology [38].

Relevant to the present work are differential geometry (DG) approaches for the multiscale modeling of biomolecular systems [39]. Differential geometry, a branch of mathematics, employs advanced techniques from calculus to investigate curves and surfaces. Early biomolecular applications of DG methods focused on solvation analysis, as molecular surface modeling is crucial to understanding the geometric interactions between a solute protein and its surrounding solvent environment. Several differential geometry-based solvation models have been developed for molecular surface construction and solvation analysis [40, 41]. In particular, the first variational solute-solvent

interface, known as the minimal molecular surface (MMS), was proposed in 2006 based on the Laplace-Beltrami flow [41]. Consequently, this was coupled with Poisson-Boltzmann (PB) and Poisson-Nernst-Planck (PNP) models to create a family of differential geometry-based multiscale models for predicting solvation free energies [42, 43] and ion channel transport [44, 45].

Recently, advancement in DG approaches for geometric learning of biomolecular properties has been achieved in [46], where critical chemical, physical, and biological information is encoded in element interactive manifolds extracted from a high-dimensional structural data space through a multiscale discrete-to-continuum density mapping. Low-dimensional DG representations, such as element interactive curvatures, can then be combined with robust machine learning algorithms for biomolecular modeling, leading to accurate predictions of molecular solvation-free energy, protein-ligand binding affinity, and drug toxicity [46]. In [47], molecular surface representations of element interactive manifolds were constructed as low-dimensional surface-based descriptors, resulting in a significant dimensional reduction for geometric learning. The resulting element interactive surface area (EISA) score facilitates an accurate machine learning algorithm to predict protein-ligand binding affinity. Machine learning models based on Ricci curvature (FPRC) have been proposed for protein-ligand binding affinity predictions [48]. More recently, multiscale differential geometry learning has been effectively applied to single-cell RNA sequencing data analysis [49]. The curvature-based approach was developed for the analysis of cell states in single-cell transcriptomic data [50].

The goal of this work is to introduce a multiscale differential geometry (mDG) model for protein flexibility analysis. In the mDG model, a correlation function will be constructed as in the FRI methods [32, 33, 35], based on atomic interactions of C_α atoms, which has a complexity of $\mathcal{O}(N^2)$ or $\mathcal{O}(N)$ when the fast FRI algorithm [33] is adopted. However, instead of directly calculating the rigidity or flexibility index, in this work, the correction function will be treated as a high-dimensional manifold, which embeds information on all atomic interactions. Based on the principle of DG, low-dimensional atom-atom interactive manifolds will be extracted by using curvature analysis. Since curvatures are localized, they are able to capture the atomic properties of proteins, such as protein B-factors, protein-ligand binding, and protein-protein interactions. Moreover, multiscale modeling will be carried out so that the resulting family of Riemannian manifolds can capture atom-atom interactions at different scales. By analytically calculating the curvatures at C_α atom centers, a set of mDG features is generated. Following the convention of protein flexibility analysis, the mDG features are integrated with a regression algorithm for B-factor predictions, which ensures a fair comparison of other prediction approaches.

The remainder of the paper is organized as follows. In Section 2, the proposed mDG model will be presented. Details on manifold extraction and curvature evaluation will be offered. Section 3 is dedicated to the numerical results to demonstrate the performance of the proposed algorithm. A comparison with several existing prediction methods will also be considered. A summary and future plan will be discussed at the end of this paper.

2 | Theory and Algorithm

Our approximation of protein flexibility, or B-factors, is based on the protein's 3D structure. Protein structures encompass a wide range of characteristic length scales associated with various molecular interactions, including covalent bonds, hydrogen bonds, van der Waals forces, alpha helices, and beta sheets, among others [35]. These molecular interactions are closely related to protein flexibility, making it essential to account for these multiscale interactions in our mathematical modeling.

In this study, we propose multiscale differential geometry (mDG) modeling to capture the multiscale collective motions of macromolecules. To characterize atomic interactions at varying distances, we employ multiple correlation kernels with different scaling parameters. The process begins with the introduction of atomic interaction manifolds, followed by the derivation of multiscale atomic interaction curvatures. These curvatures are used in the present study for the B-factor approximation. The performance of our mDG methods in predicting B-factors is demonstrated in the Experiments section.

2.1 | Atomic Interactive Manifolds

We propose using differential geometry modeling to capture the atomic interactions underlying atomic displacements or thermal motions. Our approach assumes that the intrinsic properties of proteins reside on a family of low-dimensional manifolds embedded within the high-dimensional space of protein structures. To achieve this, we converted discrete point cloud data (the atoms in a protein) into a continuous density distribution through a discrete-to-continuum mapping. The resulting density functions are then used to construct a series of low-dimensional manifolds that encapsulate these intrinsic atomic properties. In the following, we present the construction of an atomic interaction manifold.

We follow the coarse-grained approach for the B-factor approximation by only considering C_α atoms in a protein. Assume a protein with the number of C_α atoms equal to M. Let $\mathcal{X} = \{\mathbf{r}_1, \ldots, \mathbf{r}_M\}$ and vector $\mathbf{r}_i$ represent the 3D coordinate of ith C_α atom. Denote $\|\mathbf{r} - \mathbf{r}_i\|$ as the Euclidean distance between a point $\mathbf{r} \in \mathbb{R}^3$ and the atom $\mathbf{r}_i$. The unnormalized atomic interaction density can be given by a discrete to continuum mapping [32, 33, 35]

$$\rho(\mathbf{r}) = \sum_{j=1}^{M} w_j \Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) \tag{1}$$

where parameter w_j is the density weight for jth atom, and parameter η_j is a characteristic distance between a point in the 3D space with the jth atom. As we solely consider C_α atoms in this study, w_j is uniformly set as 1. Here, Φ is a correlation function with C^2 continuity and is chosen to have following admissibility properties:

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) \to 0, \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to \infty \tag{2}$$

$$\Phi(\|\mathbf{r} - \mathbf{r}_j\|; \eta_j) \to 1, \quad \text{as} \quad \|\mathbf{r} - \mathbf{r}_j\| \to 0 \tag{3}$$

Commonly used monotonically decaying kernel functions, such as radial basis functions, follow this pattern. Our previous work [33] has shown that the generalized exponential function,

$$\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta_j) = e^{-(\|\mathbf{r}-\mathbf{r}_j\|/\eta_j)^{\kappa}}, \kappa > 0 \tag{4}$$

and generalized Lorentz function

$$\Phi(\|\mathbf{r}-\mathbf{r}_j\|;\eta_j) = \frac{1}{1+(\|\mathbf{r}-\mathbf{r}_j\|/\eta_j)^{\nu}}, \nu > 0 \tag{5}$$

not only meet the admissibility assumptions, but also are excellent choices for protein flexibility analysis. For simplicity, we will consider only the generalized exponential function in this work.

The kernel function (4), which uses various resolution parameters η and κ, characterizes the geometric and topological compactness of atomic interactions. A multiscale representation can be appropriately designed by selecting suitable values for these two resolution parameters. Our approach adheres to FRI theory [35] for describing multiscale atomic interactions within molecules or for characterizing biomolecular interactions. Generally, a larger value of η indicates a lower resolution and a slower decay, which has an equivalent effect to smaller power values of κ. In this work, we set $w_j^n = 1$ as we focus our B-factor analysis on C_α atoms.

The correlation function ρ in Equation (1) for atomic interactions is governed by the scale parameter η and the decay parameter κ. By selecting multiple values for η and κ, we achieve a multiscale characterization of various atomic interactions. Figure 1 illustrates the isosurfaces generated from a correlation function based on a single kernel function at different isovalues, constructed from a set of C_α atoms. Consequently, utilizing various kernels with different scaling configurations in the correlation function (1) improves the embedding of different atomic interactions.

2.2 | Multiscale Differential Geometry of Differential Manifolds

In the FRI methods [32, 33, 35], the correlation function (1) is used to directly define the rigidity and flexibility indices for B-factor predictions. In this work, a different usage of this function will be explored. Mathematically, this function can be regarded as a manifold that encapsulates atomic interactions, making it feasible to use differential geometry to interpret these interactions. To this end, we first review the calculus on differentiable manifolds.

Let $\mathbf{M}: U \to \mathbb{R}^{n+1}$ be a C^2 mapping, where $U \subset \mathbb{R}^n$ is an open set with compact closure [41]. The mapping $\mathbf{M}(\mathbf{u}) = (M_1(\mathbf{u}), \ldots, M_n(\mathbf{u}), M_{n+1}(\mathbf{u}))$ represents a position vector on a hypersurface, where $\mathbf{u} = (u_1, \ldots, u_n) \in U$. The tangent or directional vectors of $\mathbf{M}$ are defined as $V_i = \frac{\partial \mathbf{M}}{\partial u_i}$ for $i = 1, \ldots, n$. The Jacobian matrix of $\mathbf{M}$ is given by $D\mathbf{M} = (V_1, V_2, \ldots, V_n)$. Using the notation $\langle \cdot \rangle$ for the Euclidean inner product in $\mathbb{R}^{n+1}$, the first fundamental form I is defined as:

$$I(V_i, V_j) := \langle V_i, V_j \rangle$$

for any pair of tangent vectors $V_i, V_j \in T_{\mathbf{u}}\mathbf{M}$, where $T_{\mathbf{u}}\mathbf{M}$ denotes the tangent hyperplane at $\mathbf{M}(\mathbf{u})$. In the coordinates $\mathbf{M}(\mathbf{u})$, the first fundamental form can be expressed as a symmetric and positive definite matrix $(g_{ij}) = (I(V_i, V_j))$.

Let $\mathbf{N}(\mathbf{u})$ denote the unit normal vector defined by the Gauss map $\mathbf{N}: U \to \mathbb{R}^{n+1}$:

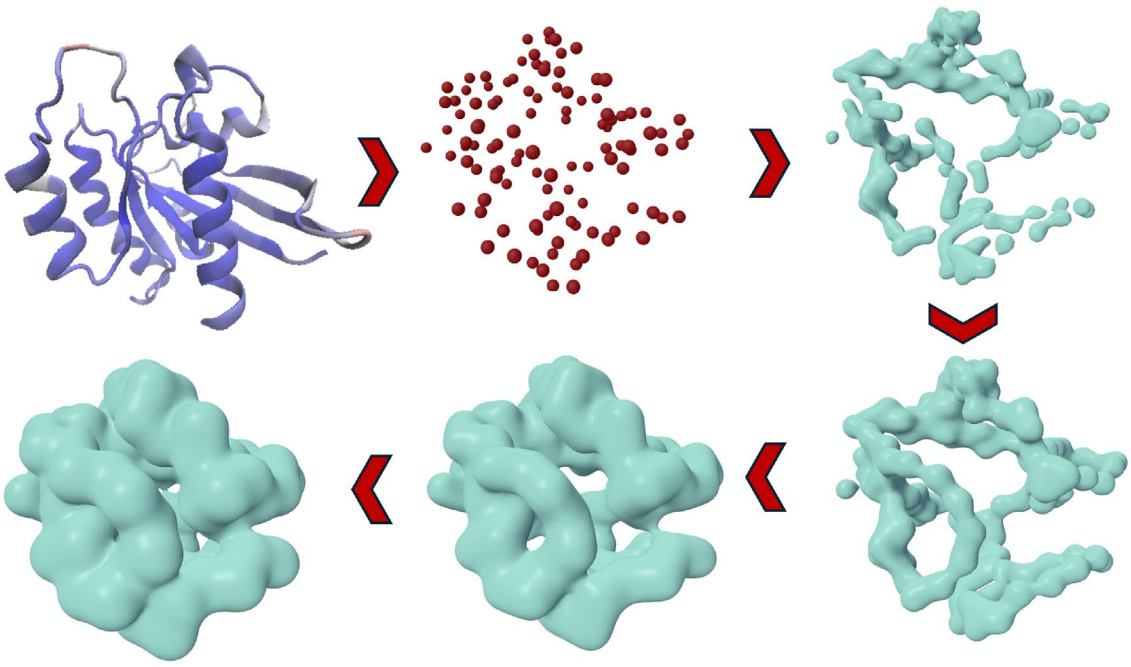

FIGURE 1 | A series of manifold for the C_α atoms of protein 1CRR at 4 different isovalues with level set function (1). The density function is determined by kernel function (4) with parameters $\tau = 3$ and $\kappa = 5$. The red arrow indicates the isosurfaces generated with decreasing isovalues.

$$\mathbf{N}(u_1,\ldots,u_n)=\frac{V_1\times V_2\times\cdots\times V_n}{\|V_1\times V_2\times\cdots\times V_n\|}\in\perp_{\mathbf{u}}\mathbf{M}$$

where $\times$ represents the cross product in $\mathbb{R}^{n+1}$ and $\perp_{\mathbf{u}}\mathbf{M}$ is the normal space of $\mathbf{M}$ at the point $\mathbf{p}=\mathbf{M}(\mathbf{u})$. The normal vector $\mathbf{N}$ is orthogonal to the tangent hyperplane $T_{\mathbf{u}}\mathbf{M}$ at $\mathbf{M}(\mathbf{u})$. Using $\mathbf{N}$ and the tangent vectors V_i, the second fundamental form is defined as:

$$II(V_i,V_j)=(h_{ij})_{i,j=1,\ldots,n}=\left(\left\langle-\frac{\partial\mathbf{N}}{\partial u_i},V_j\right\rangle\right)_{ij}$$

The mean curvature H is computed as $H=h_{ij}g^{ji}$, following the Einstein summation convention, with $g^{ji}=(g_{ij})^{-1}$. Additionally, the Gaussian curvature K is given by:

$$K=\frac{\mathrm{Det}(h_{ij})}{\mathrm{Det}(g_{ij})}$$

2.3 | Multiscale Atomic Interactive Curvatures

For the present study, it is sufficient to limit our discussion to the three-dimensional (3D) space $\mathbf{r}=(x,y,z)$, instead of the general $\mathbb{R}^n$. Based on the kernel function ρ, different manifolds can be generated by considering different isovalues ρ_0 in the level set representation $\rho(x,y,z)=\rho_0$. Here, $\rho(x,y,z)$ can be assumed to be non-degenerate, i.e., the norm of its gradient is non-zero when it is equal to ρ_0. In the following discussion, we assume that its projection onto z is non-zero. Then, a point on the iso-surface $\rho(x,y,z)=\rho_0$ has its z coordinate represented as $z=d(x,y)$, so that the iso-surface takes the form $\rho(x,y,d(x,y))=\rho_0$. Note that if the projection onto z is zero, we can carry out a similar process in the x or y direction, and a similar conclusion holds.

Taking partial derivatives of $\rho(x,y,d(x,y))=\rho_0$ with respect to x and y gives

$$\rho_x(x,y,d(x,y))+\rho_z(x,y,d(x,y))d_x=0 \tag{6}$$

$$\rho_y(x,y,d(x,y))+\rho_z(x,y,d(x,y))d_y=0 \tag{7}$$

Consequently, we have $d_x=-\frac{\rho_x}{\rho_z}$ and $d_y=-\frac{\rho_y}{\rho_z}$. The coefficients in the first and second fundamental forms can then be defined as

$$\begin{gathered}E(x,y,d(x,y))=\langle\rho_x,\rho_x\rangle,\quad F(x,y,d(x,y))=\langle\rho_x,\rho_y\rangle\\ G(x,y,d(x,y))=\langle\rho_y,\rho_y\rangle\end{gathered} \tag{8}$$

$$\begin{gathered}L(x,y,d(x,y))=\langle\rho_{xx},\mathbf{n}\rangle,\quad M(x,y,d(x,y))=\langle\rho_{xy},\mathbf{n}\rangle\\ N(x,y,d(x,y))=\langle\rho_{yy},\mathbf{n}\rangle\end{gathered} \tag{9}$$

where $\langle a,b\rangle$ denotes the inner product of a and b, and $\mathbf{n}$ is the out normal direction of the iso-surface $\rho(x,y,d(x,y))=\rho_0$.

In terms of these coefficients, the 3D Gaussian curvature K and mean curvature H can be computed as

$$K=\frac{LN-M^2}{EG-F^2}\quad H=\frac{1}{2}\frac{LG-2MF+NE}{EG-F^2} \tag{10}$$

Substituting E, F, G, L, M, N into Equation (10), the Gaussian curvature K can be given as [51]

$$\begin{aligned}K=&\frac{2\rho_x\rho_y\rho_{xz}\rho_{yz}+2\rho_x\rho_z\rho_{xy}\rho_{yz}+2\rho_y\rho_z\rho_{xy}\rho_{xz}}{g^2}\\&-\frac{2\rho_x\rho_z\rho_{xz}\rho_{yy}+2\rho_y\rho_z\rho_{xx}\rho_{yz}+2\rho_x\rho_y\rho_{xy}\rho_{zz}}{g^2}\\&+\frac{\rho_z^2\rho_{xx}\rho_{yy}+\rho_x^2\rho_{yy}\rho_{zz}+\rho_y^2\rho_{xx}\rho_{zz}}{g^2}\\&-\frac{\rho_x^2\rho_{yz}^2+\rho_y^2\rho_{xz}^2+\rho_z^2\rho_{xy}^2}{g^2}\end{aligned} \tag{11}$$

where $g=\rho_x^2+\rho_y^2+\rho_z^2$. The mean curvature, which represents the average second derivative in the normal direction, is given by:

$$\begin{aligned}H\quad=&-\frac{1}{2g^{\frac{3}{2}}}\left[2\rho_x\rho_y\rho_{xy}+2\rho_x\rho_z\rho_{xz}+2\rho_y\rho_z\rho_{yz}\right.\\&\left.-(\rho_y^2+\rho_z^2)\rho_{xx}-(\rho_x^2+\rho_z^2)\rho_{yy}-(\rho_x^2+\rho_y^2)\rho_{zz}\right]\end{aligned} \tag{12}$$

Additionally, the minimum curvature $\mu_{\min}$ and maximum curvature $\mu_{\max}$ can be determined as:

$$\mu_{\min}=H-\sqrt{H^2-K}$$
$$\mu_{\max}=H+\sqrt{H^2-K}$$

In differential geometry, various curvature measures describe the deviation of a geometric object from flatness, which is applicable to curves, surfaces, and higher-dimensional manifolds. Gaussian curvature and mean curvature are particularly useful for characterizing atomic interactions.

We note that the Gaussian curvature and mean curvature computed in (11) and (12) are actually functions of the 3D space, i.e., $K(x,y,z)$ and $H(x,y,z)$ for any $(x,y,z)\in\mathbb{R}^3$, although they are derived from iso-surface representation. Moreover, given the density function ρ, both the Gaussian curvature and the mean curvature are continuous and can be computed analytically. This ensures that their expressions are free from numerical errors, making them well-suited for modeling atomic interactions. Additionally, the computational cost is relatively low since the density function includes only the C_α atoms, which are limited in number within the given datasets. In previous work [33], fast algorithms were developed to take advantage of the rapid decay of the kernel effect within a narrow manifold band. This approach effectively addresses challenges related to the high-computational demands for larger proteins in practical applications.

2.4 | Multiscale Differential Geometry for Protein B-Factor Modeling

Based on the interactive manifold described in (1) and our differential geometry analysis, it is reasonable to use Gaussian and mean curvatures to provide a quantitative measure (K,H) of the interaction of an atom with others. To this end, we obtain a collection of Gaussian and mean curvatures as atomic features by varying the values of η and κ in the correlation

function (1). Specifically, we consider a set of η values, η_i for $i = 1, 2, \ldots, p$, and a set of κ values, κ_j for $j = 1, 2, \ldots, q$. For the mth atom among M atoms in a given protein molecule, we define a curvature vector:

$$C_m = \{(K_m^{i,j}, H_m^{i,j}) | i = 1, 2, \ldots, p; j = 1, 2, \ldots, q\} \tag{13}$$

The above vector serves as a set of local features for a C_α atom, representing its interactions with other atoms in the protein. Although we focus on B-factor predictions for C_α atoms in proteins, the approach presented in this work provides a general framework that can be used to predict B-factors for any atom in a protein.

In the current study, we consider two types of B-factor predictions. The first type follows the convention of protein flexibility analysis. We use the above mDG features to fit B-factors within a given protein using a least squares minimization:

$$\min_{a^{i,j}, b^{i,j}} \left\{ \sum_m \left| \sum_{i,j} a^{i,j} K_m^{i,j} + \sum_{i,j} b^{i,j} H_m^{i,j} + c - B_m^e \right|^2 \right\} \tag{14}$$

where B_m^e are the experimental B-factors. The parameters $a^{i,j}$, $b^{i,j}$, and c must be determined through the optimization problem stated in Equation (14). Note that the curvatures $K_m^{i,j}$ and $H_m^{i,j}$ are associated with the parameters η_i and κ_j, which are preset in our multiscale modeling. Specifically, we use $\kappa = 2$ and 5, and η is set to 5, 9, 13, 17, 21, 25, and 29. The least square minimization yields a multiple linear regression model. This way, we can have B factor approximations for a set of atoms in a given protein and use certain metrics to evaluate the approximation accuracy. For example, we use the Pearson correlation coefficient as detailed in the following section.

2.5 | Additional Features for Machine Learning

The second type of B-factor prediction we considered is a blind prediction for protein B-factors. Blind prediction involves using a training set containing sufficient information about C_α atoms and their corresponding B-factor values as labels to predict the B-factor values for C_α atoms in the test set. We utilize mDG features as local descriptors of protein structures. These mDG features are combined with additional global and local protein features to build machine-learning models. Each PDB structure includes a set of global features, such as the R-value, protein resolution, and the number of heavy atoms, which are provided in the PDB files. The local features of each protein include packing density, amino acid type, occupancy, and secondary structure information generated by STRIDE software [52]. STRIDE provides detailed secondary structure information for a protein based on its atomic coordinates from a PDB file, classifying each atom into categories such as alpha helix, 3-10 helix, π-helix, extended conformation, isolated bridge, turn, or coil. Furthermore, STRIDE provides ϕ and ψ angles and residue solvent-accessible area, resulting in a total of 12 features. The packing density of each C_α atom in a protein is determined by the density of surrounding atoms. We defined short, medium, and long-range packing density features for each C_α atom. The packing density of the ith C_α atom is defined as

$$p_i^d = \frac{N_d}{N} \tag{15}$$

where d represents the specified cutoff distance in Angstroms, N_d denotes the number of atoms within the Euclidean distance d from the ith atom, and N is the total number of heavy atoms in the protein. The packing density cutoff values used in this study are provided in Table 1.

TABLE 1 | Packing density parameter in distance dÅ.

Short	Medium	Long
$d < 3$	$3 \geq d < 5$	$5 \leq d$

TABLE 2 | Details about the mDG features, as well as the additional global and local features utilized in this study, are provided.

mDG features (28)	$K(\kappa, \eta)$, $H(\kappa, \eta)$, $\kappa = 2, 5$, $\eta = 5, 9, 13, 17, 21, 25, 29$
Global features (3)	R-value, protein resolution, # of heavy atoms
Local features (9)	Packing density value (short, medium, and long-ranged)
	amino acid type, occupancy, secondary structure type each atom belongs to
	ϕ angle, ψ angle, residue solvent-accessible area

Our mDG features, combined with the global and local features inherent in each PDB file, provide a comprehensive set of features for each C_α atom in the protein. The details about these features are provided in Table 2.

For blind predictions, we integrate these features with machine learning algorithms to build regression models. To demonstrate the performance of our machine learning model for blind predictions, we perform two validation tasks: 10-fold cross-validation and leave-one-(protein)-out validation. The modeling and predictions focus on the B-factors of C_α atoms. Details and results are presented in the following section. The hyperparameters used in the two types of algorithms are given in Table 3.

3 | Results

3.1 | Data Sets

In this work, we use two datasets, one from Refs. [33, 35] and the other from [21]. The first contains 364 proteins [33, 35], and the second [21] contains three sets of proteins with small, medium, and large sizes, which are subsets of the 364 proteins. The number of proteins in each dataset is as indicated in Table 4.

In blind predictions, proteins 1ob4, 1ob7, 2oxl, and 3md5 are excluded from the data set because the STRIDE software cannot provide features for these proteins. We exclude protein 1agn because of the known problems with this protein data. Proteins 1nko, 2oct, and 3fva are also excluded because these proteins have residues with B-factors reported as zero, which is unphysical. The following proteins were also excluded due to inconsistent protein data processed with STRIDE compared to original PDB data: 3dwv, 3mgn, 4dpz, 2j32, 3mea, 3a0m, 3ivv, 3w4q, 3p6j, and 2dko. A total of 346 proteins were used for blind predictions. The data is available in our GitHub repository.

TABLE 3 | The hyperparameters of random forest (RF) and gradient boosting decision tree (GBDT) used for the B-factor blind predictions.

RF parameters	GBDT parameters
	$n\_estimators = 1,000$
$n\_estimators = 1,000$	$max\_depth = 7$
$max\_depth = 8$	$min\_samples\_split = 5$
$min\_samples\_split = 4$	$subsample = 0.8$
$min\_samples\_leaf = 0.8$	$learning\_rate = 0.002$
	$max\_features = "sqrt"$

TABLE 4 | Details of the datasets utilized in this study.

Dataset	Total
Set-364 [33, 35]	364
B-factor (small) [21]	30
B-factor (medium) [21]	36
B-factor (large) [21]	34

3.2 | Evaluation Metrics

To quantitatively assess our method for B-factor prediction, we use the Pearson correlation coefficient (PCC):

$$\mathrm{PCC}(\mathbf{x},\mathbf{y}) = \frac{\sum_{m=1}^{M}(B_m^e - \overline{B}^e)(B_m^t - \overline{B}^t)}{\sqrt{\sum_{m=1}^{M}(B_m - \overline{B}^e)^2 \sum_{m=1}^{M}(B_m^t - \overline{B}^t)^2}}$$

where $B_m^t, m = 1, 2, \ldots, M$ are the predicted B-factors using the and $B_m^e, m = 1, 2, \ldots, M$ are the experimental B-factors from the PDB file. M indicates the number of C_α atoms. Here $\overline{B}^e$ and $\overline{B}^t$ are the averaged B-factors.

3.3 | Experiments

3.3.1 | Least Square Approximations

For the least square approximations in B-factor modeling, we consider four B-factor datasets above. With the same benchmark datasets, there exist other advanced models in the literature, including Gauss network model (GNM) [16, 17], flexibility rigidity index-based approaches such as pfFRI [33] and opFRI [33], topology-based methods such as atom-specific persistent homology (ASPH) [36] and evolutionary homology (EH) [37]. Figure 2a gives the comparison between our mDG model and these approaches to model the superset. Our model gives the highest average PCC value of 0.715 for the 364 proteins, surpassing the average PCC values of 0.698 for EH, 0.673 for opFRI, 0.65 for ASPH, 0.626 for pfFRI, and 0.565 for GNM, respectively. This represents a significant improvement of 2.4%, 6.24%, 10%, 14.2%, and 26.5%, respectively. Table A1 presents the detailed comparative results between our mDG method and other approaches in terms of PCC value. Remarkably, mDG achieves PCC values higher than all these previous methods on 209 of the 364 proteins.

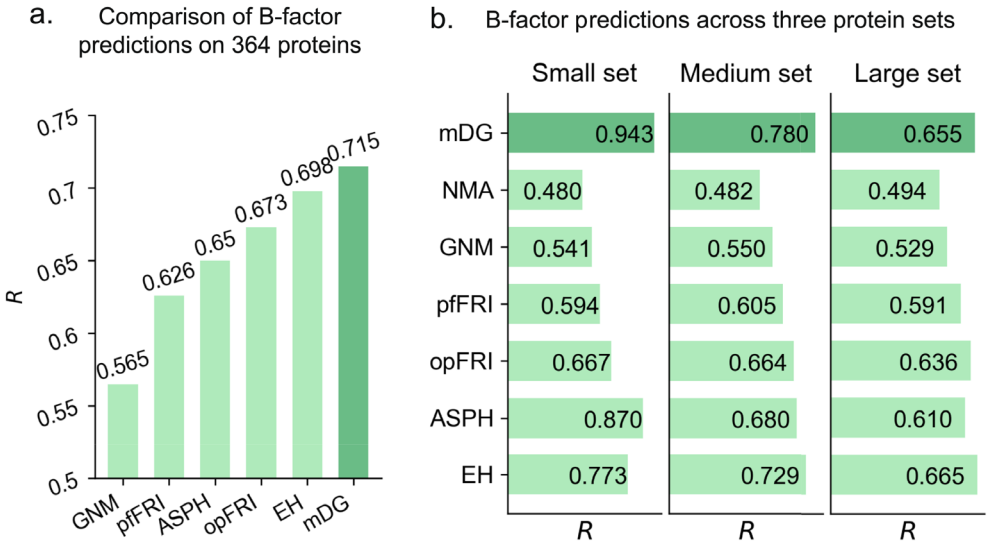

FIGURE 2 | The average pearson correlation coefficient (PCC) values using various advanced models for four B-factor prediction datasets are illustrated. (a) compares the average Pearson correlation coefficients between our mDG model and previous B-factor prediction models for 364 proteins, and (b) compares the average PCC between our mDG model and previous B-factor prediction models across small, medium, and large protein datasets.

The other three datasets consist of proteins of small, medium, and large sizes. We use these datasets to further validate the performance of our model. Figure 2b compares the average PCC values of several models for each dataset. Normal mode analysis (NMA) [11] is another B-factor prediction model. mDG achieved mean correlation coefficients of 0.943, 0.780, and 0.655 for the small, medium, and large protein sets, respectively. Our mDG model outperforms the previous methods, showing improvements of 22% and 7% on the small and medium protein sets, respectively. Our mDG has an average PCC value of 0.655 for the large protein dataset, which is only slightly below the previous state-of-the-art method EH [37] with PCC of 0.665. These comprehensive comparisons demonstrate the robustness of the mDG model for B-factor prediction across different protein sizes. In fact, the performance of the mDG model for the dataset of large proteins can be improved by increasing the number of kernels, particularly by employing additional exponential kernel functions with larger η values. A large protein has more atomic interactions. Embedding such kernel functions is beneficial in capturing those atomic interactions far away from the given atom.

The performance of GNM and NMA is poor in these four datasets with average PCC values less than 0.6. Overall, our mDG model significantly outperforms the two well-known models. Previous studies [35] have found that GNM and NMA do not provide reliable B-factor predictions on some proteins with a hinge region or other special protein structures, as shown in the following case studies. There are various reasons for their low performance, partially due to the cutoff distance in building their models. Atomic interactions outside a cutoff distance are not taken into account. Our multiscale differential geometry approach uses kernel functions to capture all atomic interactions within a molecule, while various scale distances, η, in the kernel functions allow one to capture multi-resolution atomic interactions. The multiscale different geometry modeling can properly address those challenges GNM and NMA face.

3.3.2 | Machine Learning Blind Predictions

For blind predictions, our mDG features, along with other global and local features, are integrated with two types of machine learning algorithms: Gradient-boosting decision trees (GBDT) and random forest trees (RF). We conducted several experiments, the first of which involved leave-one-(protein)-out prediction using the four datasets mentioned above. We trained the models ten times independently with different random seeds and calculated the average Pearson correlation coefficients from the ten sets of modeling predictions. The performance of the two types of machine learning models is shown in Table 5, where the GBDT-based models produce better predictions than the RF-based models.

Least squares approximations were used to show the effectiveness of our mDG-based model. The overall performance of our model mDG is better than those in the literature. Here, we present several case studies of relatively complex protein structures, which demonstrate the effectiveness of mDG modeling over others.

TABLE 5 | Average pearson correlation coefficients (PCC) of leave-one (protein)-out predictions for the four B-factor datasets. The PCC results with random forest tree and gradient-boosting decision tree modeling are compared.

Protein set	RF	GBDT
Small	0.460	0.509
Medium	0.513	0.582
Large	0.475	0.557
Superset	0.526	0.587

TABLE 6 | Average pearson correlation coefficient (PCC) from protein-level 10-fold cross-validation predictions with the collected 346 proteins. The B-factor values of C_α atoms in each protein are predicted. The average PCC value is calculated from ten independent tests. The PCC results with random forest and gradient-boosting decision tree modeling are compared.

Protein set	RF	GBDT
Superset	0.400	0.407

TABLE 7 | Average Pearson correlation coefficient (PCC) from C_α-level 10-fold cross-validation predictions with all C_α atoms in the collected 346 proteins. The average PCC value is calculated from ten independent tests. The PCC results with random forest tree and gradient-boosting decision tree modeling are compared.

Protein set	RF	GBDT
Superset	0.842	0.859

We also performed a 10-fold cross-validation in our modeling. We used nine out of ten subsets of the 346 proteins to train our model, while the remaining subset is used for testing. Specifically, the characteristics of the C_α atoms in the training proteins are combined and used to train the models. Those in the remaining proteins are used for testing. Ten different splits were performed. Table 6 gives the average PCC values for two types of machine learning models. GBDT modeling gives superior predictions than RF-based modeling.

We also performed a 10-fold cross-validation alternative for our modeling. The dataset consists of 346 proteins, which contain more than 74,000 C_α atoms in total. In each of the ten independent models, nine out of ten subsets of atoms are used to train the models, while the remaining subset is used for testing. Table 7 gives the PCC values of two types of machine learning models with different algorithms. GBDT modeling yields slightly better predictions than RF-based modeling.

3.3.3 | Case Studies of Mdg Modeling

The first example is protein calmodulin (PDBID: 1CLL) with 144 residues, which plays a key role in the signal transduction of calcium by modulating its interactions with various target

proteins, such as kinases and phosphatases. Calmodulin's remarkable structural flexibility enables it to recognize a wide variety of target proteins. Proteins with hinge structures, such as calmodulin, can undergo significant conformational changes, making it an excellent example for this type of analysis. The upper section of Figure 3 displays the proteins colored by the experimental or predicted B-factor values. The central region of calmodulin, as shown in Figure 3 is a long α-helix, which has a high-degree of flexibility based on B-factor values from PDB 1CLL. Comparisons of B-factor predictions between the mDG and GNM models can be observed. It is clear that the B-factor values predicted by mDG are very close to the experimental values, whereas those predicted by GNM are less accurate. The lower section of Figure 3 presents a detailed numerical comparison, highlighting that the GNM method exhibits significant errors around residues 65–85. In contrast, the mDG method provides accurate B-factor predictions for these residues. GNM7 and GNM8 denote predictions made using the GNM model with the cutoff distances of 7Å and 8Å, respectively. Adjusting the cutoff distance in the GNM model does not resolve the inaccuracies observed in the hinge region.

The second example is a potential antibiotic synthesis protein (PDBID: 1V70) with 105 residues. Comparisons of predicted B-factors using the mDG and GNM models are shown in Figure 4. The coloration of the B-factor values of the mDG predictions aligns closely with the experimental B-factor values, whereas the coloration of the GNM method shows discrepancies. The problematic portion for B-factor predictions is at one end of a protein chain. In this case, there is an overestimation of flexibility for residues 1–10 when using GNM. Again, the GNM8 model gives marginally better predictions. However, neither is capable of reaching the accuracy of mDG. As in the last case for 1CLL protein, 1V70 is a moderate-size protein. mDG exhibits excellent B-factor predictions.

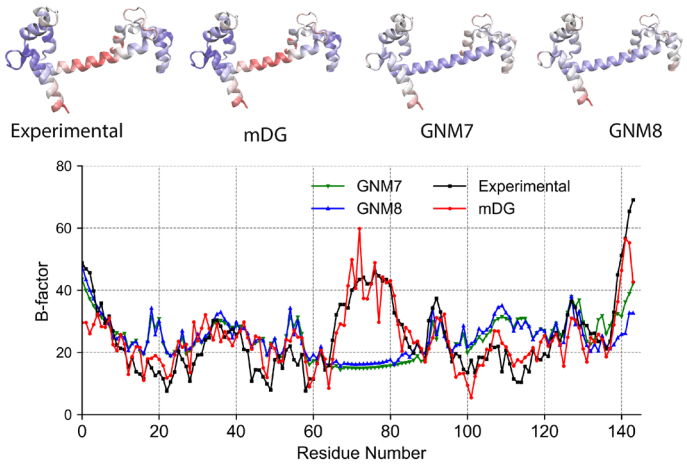

FIGURE 3 | The upper section displays the 1CLL protein colored according to B-factor values from the experimental method, the mDG model, and the GNM model. The lower panel presents a detailed comparison of predicted B-factor values from various models alongside the experimental B-factor values. GNM7 and GNM8 refer to GNM modeling with cutoff distances of 7 Å and 8 Å, respectively.

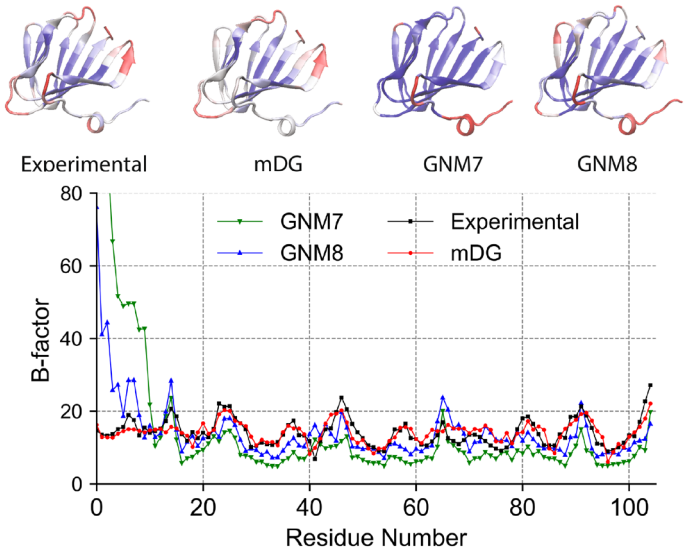

FIGURE 4 | The upper section displays the 1V70 protein colored according to B-factor values from the experimental method, the mDG model, and the GNM model. The lower panel presents a detailed comparison of predicted B-factor values from various models alongside the experimental B-factor values. GNM7 and GNM8 refer to GNM modeling with cutoff distances of 7 Å and 8 Å, respectively.

In the third example, we examine the flexibility prediction for the protein 2hqk, which has 232 residues. The lower section of Figure 5 clearly shows that the GNM model exhibits significantly poor B-factor predictions around residues 50–60, with particularly pronounced errors at the recommended cutoff distance of 7 Å. Even with an adjusted cutoff distance of 8 Å, GNM8 does not resolve this issue. In contrast, the mDG model does not exhibit these problems. The protein color based on predictions of mDG, shown in the upper section of Figure 5, closely resembles the experimental B-factors. Further inspection reveals that the problematic region of residues 50–60 corresponds to a small alpha-helical segment within the beta-barrel. This example highlights the sensitivity of the GNM model to cutoff distances and underscores how protein flexibility can be influenced by atomic interactions across various ranges. The mDG model, utilizing kernel functions with multiple scales, effectively captures these diverse molecular interactions, demonstrating its superiority over the GNM both theoretically and experimentally. Protein 2HQK has a larger structure size than those in the above two cases. The numerical comparison validates the robustness of mDG in the B-factor prediction of a large-size protein.

As the last example, we consider protein 2GZQ, which has relatively low protein flexibility with B-factor values smaller than 15 except at one end of the protein chain. Overall, the predicted B-factor values using the GNM model are slightly higher than those using the mDG model. The predictions from our mDG model are closer to the experimental values than those of GNM. Figure 6 gives a detailed comparison between our mDG and GNM models. Protein 2GZQ also has a moderately large size with 203 residues. Our mDG model still remains effective and yields accurate predictions of B-factor values for such a large-sized protein.

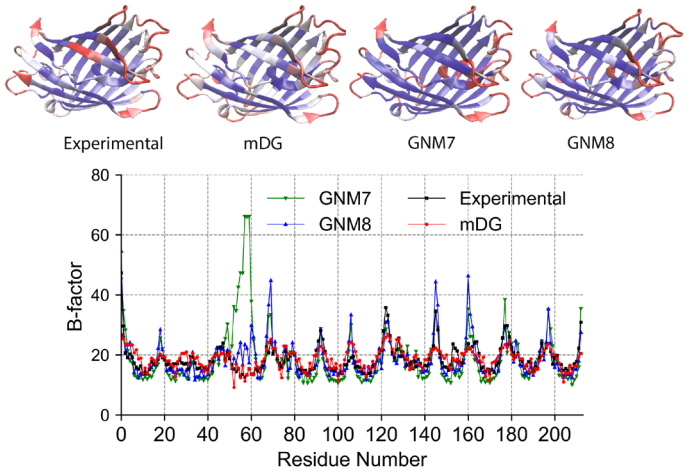

FIGURE 5 | The upper section displays the 2HQK protein colored according to B-factor values from the experimental method, the mDG model, and the GNM model. The lower panel presents a detailed comparison of predicted B-factor values from various models alongside the experimental B-factor values. GNM7 and GNM8 refer to GNM modeling with cutoff distances of 7 Å and 8 Å, respectively.

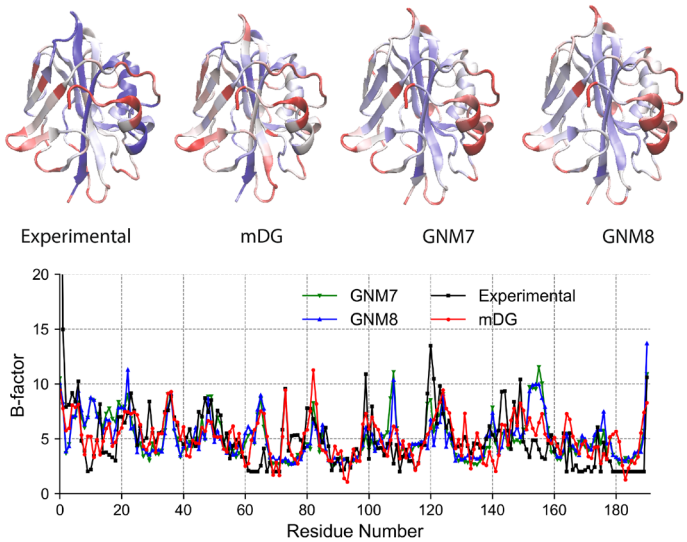

FIGURE 6 | The upper section displays the 2GZQ protein colored according to B-factor values from the experimental method, the mDG model, and the GNM model. The lower panel presents a detailed comparison of predicted B-factor values from various models alongside the experimental B-factor values. GNM7 and GNM8 refer to GNM modeling with cutoff distances of 7 Å and 8 Å, respectively.

4 | Concluding Remarks

Protein flexibility is crucial to protein functions and its prediction is important for us to understand the protein properties. The intrinsic structural complexity hinders the understanding of protein flexibility. Effective computational approaches have been designed to predict B-factor values that reflect protein flexibility, such as GNM [16, 17], pfFRI [33], ASPH [36], opFRI [33], EH [37], and NMA [11]. Our multiscale differential geometry model is a novel approach in this regard. Its effectiveness for B-factor predictions has been demonstrated with least square approximation in comparison with these available methods. With the assumption that atomic properties are sampled on low-dimensional manifolds in the high-dimensional protein structures, we construct a series of density-defined manifolds using soft-decaying kernel functions. The mean and Gauss curvatures from differential geometry are proper tools for analyzing atomic interactions. Different scales used in these density functions are beneficial for capturing atomic interactions in different distance ranges, which contributes to effective multiscale modeling. In this sense, it is superior to the hard cutoff strategy in other methods, such as the GNM method, which may overlook the atomic interactions away from the cutoff distances. Our mDG method does not have such an issue, especially by employing the multiscale strategy.

The integration of mDG features with additional global and local features intrinsic to protein structures and structure determination conditions gives rise to useful machine learning models for blind predictions. Such blind-prediction models are useful to assess the B factor values or protein flexibility when the experimental B factors are unavailable. The extensive experiments with leave-one (protein)-out and 10-fold cross-validation confirm the effectiveness and robustness of our machine learning models, especially the gradient-boosting decision tree model.

Acknowledgments

This work was supported in part by NIH grants R01AI164266 and R35GM148196, NSF grants DMS-2052983 and IIS-1900473, MSU Research Foundation, and Bristol-Myers Squibb 65109.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/fenghon1/MDG_bfactor.

References

1. C. B. Anfinsen, "Principles That Govern the Folding of Protein Chains," *Science* 181, no. 4096 (1973): 223–230.

2. K. N. Trueblood, H.-B. Bürgi, H. Burzlaff, et al., "Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature," *Acta Crystallographica Section A: Foundations of Crystallography* 52, no. 5 (1996): 770–781.

3. K. Teilum, J. G. Olsen, and B. B. Kragelund, "Functional Aspects of Protein Flexibility," *Cellular and Molecular Life Sciences* 66 (2009): 2231–2247.

4. K. Teilum, J. G. Olsen, and B. B. Kragelund, "Protein Stability, Flexibility and Function," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1814, no. 8 (2011): 969–976.

5. J. Ma, "Usefulness and Limitations of Normal Mode Analysis in Modeling Dynamics of Biomolecular Complexes," *Structure* 13, no. 3 (2005): 373–380.

6. H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The Energy Landscapes and Motions of Proteins," *Science* 254, no. 5038 (1991): 1598–1603.

7. B. R. Chandrika, J. Subramanian, and S. D. Sharma, "Managing Protein Flexibility in Docking and Its Applications," *Drug Discovery Today* 14, no. 7–8 (2009): 394–400.

8. H. A. Carlson and J. A. McCammon, "Accommodating Protein Flexibility in Computational Drug Design," *Molecular Pharmacology* 57, no. 2 (2000): 213–218.

9. S. J. Teague, "Implications of Protein Flexibility for Drug Discovery," *Nature Reviews Drug Discovery* 2, no. 7 (2003): 527–541.

10. J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of Folded Proteins," *Nature* 267, no. 5612 (1977): 585–590.

11. 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," *Journal of Computational Chemistry* 4, no. 2 (1983): 187–217.

12. N. Go, T. Noguti, and T. Nishikawa, "Dynamics of a Small Globular Protein in Terms of Low-Frequency Vibrational Modes," *Proceedings of the National Academy of Sciences* 80, no. 12 (1983): 3696–3700.

13. M. Levitt, C. Sander, and P. S. Stern, "Protein Normal-Mode Dynamics: Trypsin Inhibitor, Crambin, Ribonuclease and Lysozyme," *Journal of Molecular Biology* 181, no. 3 (1985): 423–447.

14. M. Tasumi, H. Takeuchi, S. Ataka, A. M. Dwivedi, and S. Krimm, "Normal Vibrations of Proteins: Glucagon," *Biopolymers* 21, no. 3 (1982): 711–714.

15. A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of Fluctuation Dynamics of Proteins With an Elastic Network Model," *Biophysical Journal* 80, no. 1 (2001): 505–515.

16. I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, "Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability," *Physical Review Letters* 80 (1998): 2733–2736.

17. 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, no. 3 (1997): 173–181.

18. K. Hinsen, "Analysis of Domain Motions by Approximate Normal Mode Calculations," *Proteins: Structure, Function, and Bioinformatics* 33, no. 3 (1998): 417–429.

19. G. Li and Q. Cui, "A Coarse-Grained Normal Mode Approach for Macromolecules:An Efficient Implementation and Application to Ca^{2+}-Atpase," *Biophysical Journal* 83, no. 5 (2002): 2457–2474.

20. F. Tama and Y.-H. Sanejouand, "Conformational Change of Proteins Arising From Normal Mode Calculations," *Protein Engineering, Design and Selection* 14, no. 1 (2001): 1–6.

21. J.-K. Park, R. Jernigan, and W. Zhijun, "Coarse Grained Normal Mode Analysis vs. Refined Gaussian Network Model for Protein Residue-Level Structural Fluctuations," *Bulletin of Mathematical Biology* 75, no. 1 (2013): 124–160.

22. P. J. Flory, "Statistical Thermodynamics of Random Networks," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 351, no. 1666 (1976): 351–380.

23. 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:BBI.S460 (2008): 25–45.

24. D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, "Protein Flexibility Predictions Using Graph Theory," *Proteins: Structure, Function, and Bioinformatics* 44, no. 2 (2001): 150–165.

25. A. Schlessinger and B. Rost, "Protein Flexibility and Rigidity Predicted From Sequence," *Proteins: Structure, Function, and Bioinformatics* 61, no. 1 (2005): 115–126.

26. A. G. de Brevern, A. Bornot, P. Craveur, C. Etchebest, and J.-C. Gelly, "Predyflexy: Flexibility and Local Structure Prediction From Sequence," *Nucleic Acids Research* 40, no. W1 (2012): W317–W322.

27. Y. V. Meersche, G. Cretin, A. G. de Brevern, J.-C. Gelly, and T. Galochkina, "Medusa: Prediction of Protein Flexibility From Sequence," *Journal of Molecular Biology* 433, no. 11 (2021): 166882.

28. M. R. Masters, A. H. Mahmoud, Y. Wei, and M. A. Lill, "Deep Learning Model for Efficient Protein–Ligand Docking With Implicit Side-Chain Flexibility," *Journal of Chemical Information and Modeling* 63, no. 6 (2023): 1695–1707.

29. X. Song, L. Bao, C. Feng, et al., "Accurate Prediction of Protein Structural Flexibility by Deep Learning Integrating Intricate Atomic Structures and Cryo-Em Density Information," *Nature Communications* 15, no. 1 (2024): 5538.

30. G. Casadevall, C. Duran, and S. Osuna, "Alphafold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design," *JACS Au* 3, no. 6 (2023): 1554–1562.

31. G. Xu, Y. Yang, Y. Lv, Z. Luo, Q. Wang, and J. Ma, "Opus-Bfactor: Predicting Protein b-Factor With Sequence and Structure Information," 2024.bioRxiv.

32. K. Xia, K. Opron, and G.-W. Wei, "Multiscale Multiphysics and Multidomain Models–Flexibility and Rigidity," *Journal of Chemical Physics* 139, no. 19 (2013): 194109.

33. K. Opron, K. Xia, and G.-W. Wei, "Fast and Anisotropic Flexibility-Rigidity Index for Protein Flexibility and Fluctuation Analysis," *Journal of Chemical Physics* 140, no. 23 (2014): 234105.

34. K. Xia and G.-W. Wei, "Stochastic Model for Protein Flexibility Analysis," *Physical Review E* 88 (2013): 062709.

35. K. Opron, K. Xia, and G.-W. Wei, "Communication: Capturing Protein Multiscale Thermal Fluctuations," *Journal of Chemical Physics* 142, no. 21 (2015): 211101.

36. D. Bramer and G.-W. Wei, "Atom-Specific Persistent Homology and Its Application to Protein Flexibility Analysis," *Computational and Mathematical Biophysics* 8, no. 1 (2020): 1–35.

37. Z. Cang, E. Munch, and G.-W. Wei, "Evolutionary Homology on Coupled Dynamical Systems With Applications to Protein Flexibility Analysis," *Journal of Applied and Computational Topology* 4, no. 4 (2020): 481–507.

38. C. S. Pun, B. Y. S. Yong, and K. Xia, "Weighted-Persistent-Homology-Based Machine Learning for Rna Flexibility Analysis," *PLoS One* 15, no. 8 (2020): e0237747.

39. G.-W. Wei, "Differential Geometry Based Multiscale Models," *Bulletin of Mathematical Biology* 72 (2010): 1562–1622.

40. G. W. Wei, Y. Sun, Y. C. Zhou, and M. Feig, "Molecular Multiresolution Surfaces," 2005.arXiv preprint math-ph/0511001.

41. P. W. Bates, G.-W. Wei, and S. Zhao, "Minimal Molecular Surfaces and Their Applications," *Journal of Computational Chemistry* 29, no. 3 (2008): 380–391.

42. Z. Chen, N. A. Baker, and G. W. Wei, "Differential Geometry Based Solvation Model i: Eulerian Formulation," *Journal of Computational Physics* 229, no. 22 (2010): 8231–8258.

43. Z. Chen, N. A. Baker, and G. W. Wei, "Differential Geometry Based Solvation Model Ii: Lagrangian Formulation," *Journal of Mathematical Biology* 63, no. 6 (2011): 1139–1200.

44. D. Chen, Z. Chen, and G.-W. Wei, "Quantum Dynamics in Continuum for Proton Transport Ii: Variational Solvent–Solute Interface," *International Journal for Numerical Methods in Biomedical Engineering* 28, no. 1 (2012): 25–51.

45. G.-W. Wei, Q. Zheng, Z. Chen, and K. Xia, "Variational Multiscale Models for Charge Transport," *SIAM Review* 54, no. 4 (2012): 699–754.

46. D. D. Nguyen and G.-W. Wei, "Dg-Gl: Differential Geometry-Based Geometric Learning of Molecular Datasets," *International Journal for Numerical Methods in Biomedical Engineering* 35, no. 3 (2019): e3179.

47. M. M. Rana and D. D. Nguyen, "Eisa-Score: Element Interactive Surface Area Score for Protein-Ligand Binding Affinity Prediction," *Journal of Chemical Information and Modeling* 62, no. 18 (2022): 4329–4341.

48. J. J. Wee and K. Xia, "Forman Persistent Ricci Curvature (Fprc)-Based Machine Learning Models for Protein–Ligand Binding Affinity Prediction," *Briefings in Bioinformatics* 22, no. 6 (2021): bbab136.

49. H. Feng, S. Cottrell, Y. Hozumi, and G.-W. Wei, "Multiscale Differential Geometry Learning of Networks With Applications to Single-Cell Rna Sequencing Data," *Computers in Biology and Medicine* 171 (2024): 108211.

50. T. Huynh and Z. Cang, "Topological and Geometric Analysis of Cell States in Single-Cell Transcriptomic Data," *Briefings in Bioinformatics* 25, no. 3 (2024): bbae176.

51. K. Xia, X. Feng, Z. Chen, Y. Tong, and G.-W. Wei, "Multiscale Geometric Modeling of Macromolecules i: Cartesian Representation," *Journal of Computational Physics* 257 (2014): 912–936.

52. M. Heinig and D. Frishman, "Stride: A Web Server for Secondary Structure Assignment From Known Atomic Coordinates of Proteins," *Nucleic Acids Research* 32, no. 2 (2004): W500–W502.

Appendix A

Table A1 presents the comparisons of B-factor predictions from our mDG model and other literature models [21, 33]. The best predictions are highlighted in bold.

TABLE A1 | The Comparison of Correlation Coefficients of Mdg With Previous Methods Including Opfri, Prfri, and GNM. N Refers to the Number of Residues in the Protein. the Best Value for Each Protein is Marked in Bold.

PDB ID	N	opFRI	pfFRI	GNM	mDG	PDB ID	N	opFRI	pfFRI	GNM	mDG
1ABA	87	0.727	0.698	0.613	**0.868**	1AHO	64	0.698	0.625	0.562	**0.847**
1AIE	31	0.588	0.416	0.155	**0.976**	1AKG	16	0.373	0.350	0.185	**1.000**
1ATG	231	**0.613**	0.578	0.497	0.574	1BGF	124	**0.603**	0.539	0.543	0.520
1BX7	51	0.726	0.623	0.706	**0.802**	1BYI	224	0.543	0.491	0.552	**0.568**
1CCR	111	0.580	0.512	0.351	**0.767**	1CYO	88	0.751	0.702	0.741	**0.823**
1DF4	57	**0.912**	0.889	0.832	0.883	1E5K	188	0.746	0.732	**0.859**	0.659
1ES5	260	0.653	0.638	**0.677**	0.661	1ETL	12	0.710	0.609	0.628	**1.000**
1ETM	12	0.544	0.393	0.432	**1.000**	1ETN	12	0.089	0.023	−0.274	**1.000**
1EW4	106	0.650	0.644	0.547	**0.688**	1F8R	1932	**0.878**	0.859	0.738	0.635
1FF4	65	0.718	0.613	0.674	**0.862**	1FK5	93	0.590	0.568	0.485	**0.661**
1GCO	1044	**0.766**	0.693	0.646	0.553	1GK7	39	0.845	0.773	0.821	**0.935**
1GVD	52	0.781	0.732	0.591	**0.875**	1GXU	88	0.748	0.634	0.421	**0.833**
1H6V	2927	**0.488**	0.429	0.306	0.239	1HJE	13	0.811	0.686	0.616	**1.000**
1I71	83	0.549	0.516	0.549	**0.773**	1IDP	441	**0.735**	0.715	0.690	0.667
1IFR	113	0.697	0.689	0.637	**0.812**	1K8U	89	0.553	0.531	0.378	**0.857**
1KMM	1499	**0.749**	0.744	0.558	0.488	1KNG	144	0.547	0.536	0.512	**0.652**
1KR4	110	0.635	0.612	0.466	**0.789**	1KYC	15	0.796	0.763	0.754	**1.000**
1LR7	73	0.679	0.657	0.620	**0.783**	1MF7	194	0.687	0.681	**0.700**	0.694
1N7E	95	0.651	0.609	0.497	**0.794**	1NKD	59	0.750	0.703	0.631	**0.805**
1NKO	122	0.619	0.535	0.368	**0.759**	1NLS	238	**0.669**	0.530	0.523	0.577
1NNX	93	0.795	0.789	0.631	**0.864**	1NOA	113	0.622	0.604	0.615	**0.682**
1NOT	13	0.746	0.622	0.523	**1.000**	1O06	20	0.910	0.874	0.844	**1.000**
1O08	221	**0.562**	0.333	0.309	0.385	1OB4	16	0.776	0.763	0.750	**1.000**
1OB7	16	0.737	0.545	0.652	**1.000**	1OPD	85	0.555	0.409	0.398	**0.639**
1P9I	29	0.754	0.742	0.625	**1.000**	2CE0	99	0.706	0.598	0.529	**0.871**
2CG7	90	0.551	0.539	0.379	**0.662**	2COV	534	0.846	0.823	0.812	**0.850**
2CWS	227	0.647	0.640	**0.696**	0.537	2D5W	1214	**0.689**	0.682	0.681	0.414
2DKO	253	**0.816**	0.812	0.690	0.672	2DPL	565	0.596	0.538	**0.658**	0.443
2DSX	52	0.337	0.333	0.127	**0.699**	2E10	439	**0.798**	0.796	0.692	0.617
2E3H	81	**0.692**	0.682	0.605	0.671	2EAQ	89	0.753	0.690	0.695	**0.866**
2EHP	248	**0.804**	**0.804**	0.773	0.711	2EHS	75	0.720	0.713	0.747	**0.777**
2ERW	53	0.461	0.253	0.199	**0.899**	2ETX	389	0.580	0.556	0.632	**0.653**
2FB6	116	0.791	0.786	0.740	**0.795**	2FG1	157	0.620	0.617	0.584	**0.719**
2FN9	560	0.607	0.595	**0.611**	0.547	2FQ3	85	0.719	0.692	0.348	**0.876**
2G69	99	0.622	0.590	0.436	**0.813**	2G7O	68	0.785	0.784	0.660	**0.844**
2G7S	190	**0.670**	0.644	0.649	0.601	2GKG	122	0.688	0.646	0.711	**0.776**

(Continues)

TABLE A1 | (Continued)

PDB ID	N	opFRI	pfFRI	GNM	mDG	PDB ID	N	opFRI	pfFRI	GNM	mDG
2GOM	121	0.586	0.584	0.491	**0.710**	2GXG	140	**0.847**	0.780	0.520	0.818
2GZQ	191	**0.505**	0.382	0.369	0.480	2HQK	213	**0.824**	0.809	0.365	0.738
2HYK	238	0.585	0.575	0.510	**0.619**	2I24	113	0.593	0.498	0.494	**0.614**
2I49	398	**0.714**	0.683	0.601	0.671	2IBL	108	0.629	0.625	0.352	**0.700**
2IGD	61	0.585	0.481	0.386	**0.824**	2IMF	203	**0.652**	0.625	0.514	0.574
2IP6	87	0.654	0.578	0.572	**0.858**	2IVY	88	0.544	0.483	0.271	**0.774**
2J32	244	**0.863**	0.848	0.855	0.693	2J9W	200	**0.716**	0.705	0.662	0.674
2JKU	35	0.805	0.695	0.656	**0.948**	2JLI	100	0.779	0.613	0.622	**0.828**
2JLJ	115	**0.741**	0.720	0.527	0.635	2MCM	113	**0.789**	0.713	0.639	0.649
2NLS	36	0.605	0.559	0.530	**0.900**	2NR7	194	**0.803**	0.785	0.727	0.708
2NUH	104	0.835	0.691	0.771	**0.872**	2O6X	306	**0.814**	0.799	0.651	0.661
2OA2	132	0.571	0.456	0.458	**0.582**	2OCT	192	0.567	0.550	0.540	**0.569**
2OHW	256	0.614	0.539	0.475	**0.754**	2OKT	342	0.433	0.411	0.336	**0.500**
2OL9	6	0.909	0.904	0.689	**1.000**	3BA1	312	**0.661**	0.624	0.621	0.516
3BED	261	**0.845**	0.820	0.684	0.806	3BQX	139	0.634	0.481	0.297	**0.730**
3BZQ	99	0.532	0.516	0.466	**0.751**	3BZZ	100	0.485	0.450	0.600	**0.804**
3DRF	547	**0.559**	0.549	0.488	0.475	3DWV	325	**0.707**	0.661	0.547	0.682
3E5T	228	0.502	0.489	0.296	**0.549**	3E7R	40	0.706	0.687	0.642	**0.937**
3EUR	140	0.431	0.427	**0.577**	0.556	3F2Z	149	**0.824**	0.792	0.740	0.786
3F7E	254	**0.812**	0.803	0.811	0.750	3FCN	158	**0.640**	0.606	0.632	0.436
3FE7	91	0.583	0.533	0.276	**0.755**	3FKE	250	0.525	0.476	0.435	**0.672**
3FMY	66	0.701	0.655	0.556	**0.885**	3FOD	48	0.532	0.440	−0.126	**0.887**
3FSO	221	**0.831**	0.817	0.793	0.553	3FTD	240	**0.722**	0.713	0.634	0.605
3FVA	6	0.835	0.825	0.789	**1.000**	3G1S	418	0.771	0.700	0.630	**0.793**
3GBW	161	0.820	0.747	0.510	**0.829**	3GHJ	116	0.732	0.511	0.196	**0.828**
3HFO	197	**0.691**	0.670	0.518	0.569	3HHP	1234	**0.720**	0.716	0.683	0.492
3HNY	156	**0.793**	0.723	0.758	0.768	3HP4	183	0.534	0.500	0.573	**0.653**
3HWU	144	0.754	0.748	**0.841**	0.675	3HYD	7	0.966	0.950	0.867	**1.000**
3HZ8	192	0.617	0.502	0.475	**0.729**	3I2V	124	0.486	0.441	0.301	**0.642**
3I2Z	138	0.613	0.599	0.317	**0.642**	3I4O	135	0.735	0.714	0.738	**0.760**
3I7M	134	0.667	0.635	0.695	**0.762**	3IHS	169	0.586	0.565	0.409	**0.757**
3IVV	149	**0.817**	0.797	0.693	0.743	3K6Y	227	0.586	0.535	0.301	**0.695**
3KBE	140	0.705	0.704	0.611	**0.773**	3KGK	190	**0.784**	0.775	0.680	0.754
3KZD	85	0.647	0.611	0.475	**0.811**	3L41	220	**0.718**	0.716	0.669	0.636
3LAA	169	**0.827**	0.647	0.659	0.575	3LAX	106	0.734	0.730	0.584	**0.757**
3LG3	833	**0.658**	0.614	0.589	0.406	3LJI	272	**0.612**	0.608	0.551	0.530
3M3P	249	**0.584**	0.554	0.338	0.543	3M8J	178	**0.730**	0.728	0.628	0.696
3M9J	210	0.639	0.574	0.296	**0.696**	3M9Q	176	0.591	0.510	0.471	**0.625**
3MAB	173	**0.664**	0.591	0.451	0.610	3U6G	248	0.635	0.632	0.526	**0.656**
3U97	77	0.753	0.736	0.712	**0.762**	3UCI	72	0.589	0.526	0.495	**0.624**

(Continues)

TABLE A1 | (Continued)

PDB ID	N	opFRI	pfFRI	GNM	mDG	PDB ID	N	opFRI	pfFRI	GNM	mDG
3UR8	637	**0.666**	0.652	0.597	0.530	3US6	148	**0.698**	0.586	0.553	0.538
3V1A	48	0.531	0.487	0.583	**0.807**	3V75	285	0.604	0.596	0.491	**0.613**
3VN0	193	**0.840**	0.837	0.812	0.836	3VOR	182	0.602	0.557	0.484	**0.690**
3VUB	101	0.625	0.610	0.607	**0.739**	3VVV	108	**0.833**	0.741	0.753	0.736
3VZ9	163	0.785	0.749	0.695	**0.799**	3W4Q	773	**0.737**	0.725	0.649	0.593
3ZBD	213	**0.651**	0.516	0.632	0.649	3ZIT	152	0.430	0.404	0.392	**0.528**
3ZRX	221	**0.590**	0.562	0.391	0.588	3ZSL	138	0.691	0.687	0.526	**0.711**
3ZZP	74	0.524	0.460	0.448	**0.779**	3ZZY	226	**0.746**	0.709	0.728	0.542
4A02	166	0.618	0.516	0.303	**0.743**	4ACJ	167	0.748	0.746	**0.759**	0.726
4AE7	186	0.724	0.717	0.717	**0.767**	4AM1	345	**0.674**	0.619	0.460	0.653
4ANN	176	**0.551**	0.536	0.470	0.518	4AVR	188	**0.680**	0.605	0.650	0.599
4AXY	54	0.700	0.623	0.720	**0.881**	4B6G	558	**0.765**	0.756	0.669	0.567
4B9G	292	**0.844**	0.816	0.763	0.590	4DD5	387	**0.615**	0.596	0.351	0.604
4DKN	423	**0.781**	0.761	0.539	0.654	4DND	95	0.763	0.750	0.582	**0.765**
4DPZ	109	0.730	0.726	0.651	**0.767**	4DQ7	328	**0.690**	0.683	0.376	0.683
1PEF	18	0.888	0.826	0.808	**1.000**	1PEN	16	0.516	0.465	0.270	**1.000**
1PMY	123	0.671	0.654	0.685	**0.745**	1PZ4	114	0.828	0.781	**0.843**	0.818
1Q9B	43	0.746	0.726	0.656	**0.917**	1QAU	112	0.678	0.672	0.620	**0.848**
1QKI	3912	**0.809**	0.751	0.645	0.547	1QTO	122	0.543	0.520	0.334	**0.671**
1R29	122	0.650	0.631	0.556	**0.709**	1R7J	90	0.789	0.621	0.368	**0.806**
1RJU	36	0.517	0.447	0.431	**0.955**	1RRO	112	0.435	0.372	**0.529**	0.503
1SAU	114	0.742	0.671	0.596	**0.774**	1TGR	104	0.720	0.711	0.714	**0.843**
1TZV	141	0.837	0.820	**0.841**	0.711	1U06	55	0.474	0.429	0.434	**0.871**
1U7I	267	**0.778**	0.762	0.691	0.728	1U9C	221	0.600	0.577	0.522	**0.725**
1UHA	83	0.726	0.665	0.638	**0.738**	1UKU	102	0.665	0.661	**0.742**	0.721
1ULR	87	0.639	0.594	0.495	**0.665**	1UOY	64	0.713	0.653	0.671	**0.838**
1USE	40	0.438	0.146	−0.142	**0.936**	1USM	77	**0.832**	0.809	0.798	0.815
1UTG	70	0.691	0.610	0.538	**0.782**	1V05	96	0.629	0.599	0.632	**0.728**
1V70	105	0.622	0.492	0.162	**0.788**	1VRZ	21	0.792	0.695	0.677	**1.000**
1W2L	97	0.691	0.564	0.397	**0.735**	1WBE	204	0.591	0.577	0.549	**0.598**
1WHI	122	0.601	0.539	0.270	**0.734**	1WLY	322	**0.695**	0.679	0.666	0.597
1WPA	107	**0.634**	0.577	0.417	0.587	1X3O	80	0.600	0.559	0.654	**0.844**
1XY1	18	0.832	0.645	0.447	**1.000**	1XY2	8	0.619	0.570	0.562	**1.000**
1Y6X	87	0.596	0.524	0.366	**0.882**	1YJO	6	0.375	0.333	0.434	**1.000**
1YZM	46	0.842	0.834	**0.901**	0.884	1Z21	96	0.662	0.638	0.433	**0.718**
1ZCE	146	**0.808**	0.757	0.770	0.764	1ZVA	75	0.756	0.579	0.690	**0.796**
2A50	457	**0.564**	0.524	0.281	0.561	2AGK	233	**0.705**	0.694	0.512	**0.705**
2AH1	939	**0.684**	0.593	0.521	0.429	2B0A	186	0.639	0.603	0.467	**0.694**
2BCM	413	**0.555**	0.551	0.477	0.479	2BF9	36	0.606	0.554	0.680	**0.951**
2BRF	100	0.795	0.764	0.710	**0.877**	2C71	205	**0.658**	0.649	0.560	0.585

(Continues)

TABLE A1 | (Continued)

PDB ID	N	opFRI	pfFRI	GNM	mDG	PDB ID	N	opFRI	pfFRI	GNM	mDG
2OLX	4	0.917	0.888	0.885	**1.000**	2PKT	93	0.162	0.003	0.193	**0.744**
2PLT	99	0.508	0.484	0.509	**0.708**	2PMR	76	0.693	0.682	0.619	**0.801**
2POF	440	0.682	0.651	0.589	**0.692**	2PPN	107	0.677	0.638	0.668	**0.763**
2PSF	608	0.526	0.500	**0.565**	0.317	2PTH	193	**0.822**	0.784	0.767	0.778
2Q4N	153	0.711	0.667	**0.740**	0.698	2Q52	412	**0.756**	0.748	0.621	0.666
2QJL	99	0.594	0.584	0.594	**0.770**	2R16	176	0.582	0.495	0.618	**0.646**
2R6Q	138	0.603	0.540	0.529	**0.659**	2RB8	93	0.727	0.614	0.517	**0.794**
2RE2	238	0.652	0.613	**0.673**	0.498	2RFR	154	0.693	0.671	**0.753**	0.727
2V9V	135	0.555	0.548	0.528	**0.623**	2VE8	515	**0.744**	0.643	0.616	0.637
2VH7	94	0.775	0.726	0.596	**0.845**	2VIM	104	0.413	0.393	0.212	**0.599**
2VPA	204	**0.763**	0.755	0.576	0.610	2VQ4	106	0.680	0.679	0.555	**0.754**
2VY8	149	**0.770**	0.724	0.533	0.702	2VYO	210	0.675	0.648	**0.729**	0.633
2W1V	548	**0.680**	**0.680**	0.571	0.562	2W2A	350	**0.706**	0.638	0.589	0.519
2W6A	117	**0.823**	0.748	0.647	0.634	2WJ5	96	0.484	0.440	0.357	**0.698**
2WUJ	100	0.739	0.598	0.598	**0.804**	2WW7	150	0.499	0.471	0.356	**0.569**
2WWE	111	0.692	0.582	0.628	**0.883**	2X1Q	240	**0.534**	0.478	0.443	0.516
2X25	168	0.632	0.598	0.403	**0.643**	2X3M	166	**0.744**	0.717	0.655	0.690
2X5Y	171	0.718	0.705	0.694	**0.754**	2X9Z	262	**0.583**	0.578	0.574	0.492
2XHF	310	**0.606**	0.591	0.569	0.517	2Y0T	101	0.778	0.774	0.798	**0.799**
2Y72	170	**0.780**	0.754	0.766	0.712	2Y7L	319	**0.928**	0.797	0.747	0.671
2Y9F	149	0.771	0.762	0.664	**0.787**	2YLB	400	**0.807**	**0.807**	0.675	0.629
2YNY	315	**0.813**	0.804	0.706	0.721	2ZCM	357	0.458	0.422	0.420	**0.525**
2ZU1	360	**0.689**	0.672	0.653	0.600	3A0M	148	**0.807**	0.712	0.392	0.710
3A7L	128	0.713	0.663	**0.756**	0.639	3AMC	614	**0.675**	0.669	0.581	0.670
3AUB	116	0.614	0.608	0.637	**0.735**	3B5O	230	0.644	0.629	0.601	**0.725**
3MD4	12	0.860	0.781	0.914	**1.000**	3MD5	12	0.649	0.413	−0.218	**1.000**
3MEA	166	0.669	0.669	0.600	**0.737**	3MGN	348	0.205	0.119	0.193	**0.704**
3MRE	383	**0.661**	0.641	0.567	0.487	3N11	325	**0.614**	0.583	0.517	0.613
3NE0	208	**0.706**	0.645	0.659	0.627	3NGG	94	0.696	0.689	**0.719**	0.674
3NPV	495	**0.702**	0.653	0.677	0.476	3NVG	6	0.721	0.617	0.597	**1.000**
3NZL	73	0.627	0.583	0.506	**0.812**	3O0P	194	0.727	0.706	**0.734**	0.629
3O5P	128	**0.734**	0.698	0.630	0.642	3OBQ	150	0.649	0.645	**0.655**	0.576
3OQY	234	**0.698**	0.686	0.637	0.619	3P6J	125	0.774	0.767	0.810	**0.838**
3PD7	188	**0.770**	0.723	0.589	0.670	3PES	165	0.697	0.642	0.683	**0.736**
3PID	387	0.537	0.531	**0.642**	0.378	3PIW	154	**0.758**	0.744	0.717	0.615
3PKV	221	**0.625**	0.597	0.568	0.614	3PSM	94	**0.876**	0.790	0.745	0.870
3PTL	289	0.543	0.541	0.468	**0.596**	3PVE	347	**0.718**	0.667	0.568	0.635
3PZ9	357	**0.709**	**0.709**	0.678	0.602	3PZZ	12	0.945	0.922	0.950	**1.000**
3Q2X	6	0.922	0.904	0.866	**1.000**	3Q6L	131	0.622	0.577	0.605	**0.728**
3QDS	284	**0.780**	0.745	0.568	0.656	3QPA	197	**0.587**	0.442	0.503	0.469

(Continues)

TABLE A1 | (Continued)

PDB ID	N	opFRI	pfFRI	GNM	mDG	PDB ID	N	opFRI	pfFRI	GNM	mDG
3R6D	221	**0.688**	0.669	0.495	0.681	3R87	132	0.452	0.419	0.286	**0.589**
3RQ9	162	0.510	0.403	0.242	**0.675**	3RY0	128	0.616	0.606	0.470	**0.761**
3RZY	139	0.800	0.784	**0.849**	0.828	3S0A	119	0.562	0.524	0.526	**0.657**
3SD2	86	0.523	0.421	0.237	**0.760**	3SEB	238	0.801	0.712	**0.826**	0.583
3SED	124	0.709	0.658	0.712	**0.819**	3SO6	150	0.675	0.666	0.630	**0.756**
3SR3	637	0.619	0.611	**0.624**	0.395	3SUK	248	**0.644**	0.633	0.567	0.618
3SZH	697	**0.817**	0.815	0.697	0.657	3T0H	208	**0.808**	0.775	0.694	**0.793**
3T3K	122	**0.796**	0.748	0.735	0.685	3T47	141	0.592	0.527	0.447	**0.740**
3TDN	357	0.458	0.419	0.240	**0.627**	3TOW	152	0.578	0.556	0.571	**0.782**
3TUA	210	0.665	0.658	0.588	**0.706**	3TYS	75	0.853	0.800	0.791	**0.896**
4DT4	160	**0.776**	0.738	0.716	0.708	4EK3	287	**0.680**	**0.680**	0.674	0.608
4ERY	318	**0.740**	0.701	0.688	0.642	4ES1	95	0.648	0.625	0.551	**0.790**
4EUG	225	0.570	0.529	0.405	**0.612**	4F01	448	0.633	0.372	**0.688**	0.640
4F3J	143	0.617	0.598	0.551	**0.734**	4FR9	141	0.671	0.655	0.501	**0.673**
4G14	15	0.467	0.323	0.356	**1.000**	4G2E	151	0.760	0.755	0.758	**0.767**
4G5X	550	**0.786**	0.754	0.743	0.551	4G6C	658	**0.591**	0.590	0.528	0.497
4G7X	194	**0.688**	0.587	0.624	0.660	4GA2	144	**0.528**	0.485	0.406	0.513
4GMQ	92	0.678	0.628	0.550	**0.762**	4GS3	90	0.544	0.522	0.547	**0.821**
4H4J	236	**0.810**	0.806	0.689	0.705	4H89	168	**0.682**	0.588	0.596	0.523
4HDE	168	**0.745**	0.728	0.615	0.585	4HJP	281	**0.703**	0.649	0.510	0.620
4HWM	117	0.638	0.622	0.499	**0.801**	4IL7	85	0.446	0.404	0.316	**0.732**
4J11	357	**0.620**	0.562	0.401	0.587	4J5O	220	**0.793**	0.757	0.777	0.580
4J5Q	146	0.742	0.742	0.689	**0.859**	4J78	305	0.658	0.648	0.608	**0.733**
4JG2	185	**0.746**	0.736	0.543	0.661	4JVU	207	0.723	0.697	0.553	**0.736**
4JYP	534	**0.688**	0.682	0.538	0.573	4KEF	133	0.580	0.530	0.324	**0.705**
5CYT	103	0.441	0.421	0.331	**0.641**	6RXN	45	0.614	0.574	0.594	**0.889**