



# The amounts of thermal vibrations and static disorder in protein X-ray crystallographic B-factors

Hyuntae Na, Konrad Hinsén, Guang Song

## ► To cite this version:

Hyuntae Na, Konrad Hinsén, Guang Song. The amounts of thermal vibrations and static disorder in protein X-ray crystallographic B-factors. *Proteins - Structure, Function and Bioinformatics*, 2021, 89 (11), pp.1442. 10.1002/prot.26165 . hal-03358119

**HAL Id: hal-03358119**

**<https://hal.science/hal-03358119>**

Submitted on 1 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The Amounts of Thermal Vibrations and Static Disorder in Protein X-ray Crystallographic B-factors

Hyuntae Na<sup>1</sup>, Konrad Hinsén<sup>2</sup>, and Guang Song<sup>3\*</sup>

<sup>1</sup>Department of Computer Science,

Penn State Harrisburg,

Middletown, PA 17057, USA

<sup>2</sup>Centre de Biophysique Moléculaire,

CNRS Orléans

Synchrotron Soleil - Division Expériences

Saint Aubin - BP 48

91192 Gif sur Yvette Cedex, France

<sup>3</sup>Department of Computer Science,

Program of Bioinformatics and Computational Biology,

Iowa State University, Ames, IA 50011, USA

\*Correspondence to: Guang Song, Tel: +1-515-294-1696;

Fax: 515-294-0258; E-mail: gsong@iastate.edu

Short title: The Amounts of Thermal Vibrations and Static Disorder in B-factors

## Abstract

Crystallographic B-factors provide direct dynamical information on the internal mobility of proteins that is closely linked to function, and are also widely used as a benchmark in assessing elastic network models. A significant question in the field is: what is the exact amount of thermal vibrations in protein crystallographic B-factors? This work sets out to answer this question. First, we carry out a thorough, statistically sound analysis of crystallographic B-factors of over 10,000 structures. Second, by employing a highly accurate all-atom model based on the well-known CHARMM force field, we obtain computationally the magnitudes of thermal vibrations of nearly 1,000 structures. Our key findings are: (i) the magnitude of thermal vibrations, surprisingly, is nearly protein-independent, as a corollary to the universality for the vibrational spectra of globular proteins established earlier; (ii) the magnitude of thermal vibrations is small, less than  $0.1 \text{ \AA}^2$  at 100 K; (iii) the percentage of thermal vibrations in B-factors is the lowest at low resolution and low temperature ( $<10\%$ ) but increases to as high as 60% for structures determined at high resolution and at room temperature. The significance of this work is that it provides for the first time, using an extremely large dataset, a thorough analysis of B-factors and their thermal and static disorder components. The results clearly demonstrate that structures determined at high resolution and at room temperature have the richest dynamics information. Since such structures are relatively rare in the PDB database, the work naturally calls for more such structures to be determined experimentally.

**Keywords:** B-factor; dynamics; static disorder; thermal vibrations; mean-square displacements; X-ray crystallography; NMA; elastic network models

# 1 Introduction

The Debye–Waller factor (DWF), or B-factor as it is called in protein X-ray crystallography, is a factor used to describe for each atom the degree to which its electron density spreads out. If protein molecules in all unit cells had exactly the same conformation, were motionless, and lined up perfectly by having exactly the same position and orientation relative to the centers/axes of their unit cells, then the electron density would have no spread and all the B-factors would be 0.

In reality, none of above is true. First, protein molecules may assume different conformational states in unit cells<sup>1</sup>, which is one kind of static disorder. Second, protein molecules may have different positions and orientations within unit cells, which is another kind of static disorder. This latter kind of static disorder may be further divided into translational disorder (due to positional difference in unit cells) or rotational disorder (due to orientational differences in unit cells)<sup>2</sup>, although in practice the two can only occur together in a densely packed crystal. Third, protein crystals are not perfectly periodic and have defects. This is further aggravated by radiation damage, in particular for structures obtained using intense synchrotron radiation. Radiation damage can be significantly reduced by freezing the crystal and carrying out the experiment at a low cryogenic temperature (at around 100 K or lower)<sup>3</sup>, which is the very reason most structures are determined at 100 K nowadays. Data collection at temperatures below 100 K was even proposed to further reduce radiation damage<sup>4</sup>. At low cryogenic temperature, one possible consequence of radiation damage is that it may increase rotational disorder via “structural rearrangements from radiation-induced breaks in crystal contacts”<sup>5</sup>. The effect of radiation damage on B-factors is complicated. It is dose-dependent and varies from protein to protein. Since it is indistinguishable from static disorder in the analyses we perform, we no longer make the distinction in the rest of the paper. The recent advent of serial synchrotron crystallography presents an alternative

way to overcome radiation damage and creates a renewed hope and interest in carrying out X-ray crystallography at room temperature<sup>6</sup>. Fourth, protein molecules are certainly not motionless: Protein molecules may vibrate within each unit cell or have rigid body motions relative to one another, i.e., vibrations across unit cells. The former is represented by optical waves while the latter acoustic waves<sup>7</sup>. Because of the aforementioned static disorder and dynamics in a protein crystal, if we were to align all the unit cells and observe them over time, we would see that protein molecules within unit cells do not align perfectly, but spread out around a mean conformation. The spread around the mean position of each atom  $i$  is generally modeled as a Gaussian function and the magnitude of the spread is characterized by  $\langle u_i^2 \rangle$ , the mean-square displacement from the mean position.

The isotropic Debye-Waller factor, or B-factor, is related to  $\langle u_i^2 \rangle$  as:

$$B_i = \frac{8\pi^2}{3} \langle u_i^2 \rangle. \quad (1)$$

Or,

$$\langle u_i^2 \rangle = \frac{3}{8\pi^2} B_i. \quad (2)$$

Accounting for the contributions from different sources mentioned above, the mean-square displacement  $\langle u_i^2 \rangle$  can be divided and subdivided as,

$$\langle u_i^2 \rangle = \langle u_i^2 \rangle_d + \langle u_i^2 \rangle_{ld} = \underbrace{\langle u_i^2 \rangle_v + \langle u_i^2 \rangle_{aw}}_{\text{dynamics}} + \underbrace{\langle u_i^2 \rangle_{ld}^c + \langle u_i^2 \rangle_{ld}^r + \langle u_i^2 \rangle_{ld}^t + \langle u_i^2 \rangle_{ld}^f}_{\text{static disorder}}, \quad (3)$$

where the subscripts  $d$  and  $ld$  denote dynamics and lattice disorders, respectively, and the superscripts represent conformational ( $c$ ), rotational ( $r$ ), translational ( $t$ ) disorders, and crystal defects ( $f$ ). The dynamics term  $\langle u_i^2 \rangle_d$  can be subdivided into thermal vibrations within unit cells (denoted by subscript  $v$ ) and vibrations across unit cells, or acoustic waves (denoted by subscript  $aw$ ). According to Hinsen's calculations, the contribution from acoustic waves

$\langle u_i^2 \rangle_{aw}$  is significantly smaller in comparison to thermal vibrations within unit cells  $\langle u_i^2 \rangle_v$ , about 14% for lysozyme<sup>7</sup>. A further decomposition of  $\langle u_i^2 \rangle_v$  into contributions from internal and rigid-body motions can be attempted but is difficult because these two contributions produce very similar patterns<sup>8</sup>. The separation can be made more reliably when covariance information is available in addition to single-atom fluctuations<sup>9</sup>. Note also that  $\langle u_i^2 \rangle_v$  represents the mean-square displacement due to thermal vibrations of protein molecules under the influence of crystal packing. It is not the same as thermal vibrations of a single protein molecule. The influence of crystal packing has been studied by several groups of researchers<sup>7,10,11</sup>. Compared to that of a single molecule, peaks of residues that are near the surface of the protein are significantly reduced due to crystal contacts, while the rest of  $\langle u_i^2 \rangle_v$  is slightly smaller under the influence of crystal packing<sup>7</sup>.

Frauenfelder et al.<sup>1</sup> used the Mössbauer effect on protein crystals to determine  $\langle u_i^2 \rangle_d$  apart from the entanglement with lattice disorder  $\langle u_i^2 \rangle_{ld}$ . Comparison of results from Mössbauer effect with X-ray B-factors suggested a way of determining  $\langle u_i^2 \rangle_{ld}$ . Unfortunately, few nuclei are Mössbauer active<sup>2</sup>. Based on their computation on myoglobin, they concluded that lattice disorder accounted for only 15-30% of B-factors, and that the dominant contribution to B-factors at room temperature was thermal motions<sup>1,2</sup>. The conclusion was drawn from sparse data but the work was influential in the field.

The idea that thermal motions or  $\langle u_i^2 \rangle_d$  is the dominant contribution to B-factors was, however, challenged when it was noted by Kurinov and Harrison that instead of having a much higher B-factor at high temperature, lysozyme had nearly identical B-factors at cryogenic (about 100 K) and at room temperature (about 300 K)<sup>12</sup>. Kurinov and Harrison found that for lysozyme, the contribution from lattice disorder  $\langle u_i^2 \rangle_{ld}$  was actually much higher than  $\langle u_i^2 \rangle_d$  at low temperature and the two became comparable at high temperature. That is, at room temperature, the contribution of thermal fluctuations is the highest, at about 50%. Kurinov and Harrison also found that lattice disorder was nearly residue independent but

increased considerably at low temperature<sup>12</sup>.

A more recent work by Hinsen provided further evidence suggesting that static disorder was the dominant contribution to B-factors<sup>7</sup>. By comparing with experimental compressibility data, Hinsen carried out a careful study to quantify the contribution from dynamics, specifically from vibrational motions of protein molecules within unit cells (optical waves) and rigid body motions of protein molecules relative to one another across unit cells (acoustic waves). The results indicated that the percentage of contribution from dynamics  $\langle u_i^2 \rangle_d$  could be as low as 4-7% in lysozyme, or up to 16% if a higher experimental compressibility value was used. In either case, the percentage of  $\langle u_i^2 \rangle_d$  in  $\langle u_i^2 \rangle$  is very low. However, due to scarcity of experimental data on compressibility, the conclusion is limited only to tetragonal lysozyme crystals.

The extent of dynamic contributions to B-factors was also computed using MD simulations. For example, MD simulations of crambin by Burden and Oakley showed that the magnitude of the dynamical contribution was several times smaller than the crystallographic B-factors. MD simulations were performed also by Hinsen and co-workers<sup>9</sup> on a small number of proteins to compute their atomic displacements and cross correlations. Unfortunately, the magnitudes of atomic displacements were not reported nor compared with crystallographic B-factors. A fundamental problem with computing  $\langle u_i^2 \rangle_d$  from MD simulations is insufficient sampling of conformational space, which leads to a systematic underestimation of conformational fluctuations.

**Contribution of this work.** In this work, using an extremely large dataset containing more than 10,000 proteins, we carry out a thorough, statistically sound analysis of crystallographic B-factors and find that *temperature* and *structural resolution* are the two major determinants of B-factors. Specifically, (i) the magnitude of B-factors of a protein (which is defined as the mass-weighted average of those of individual atoms) is nearly temperature independent, but decreases linearly as resolution increases due to the decrease in lattice dis-

order; (ii) lattice disorder decreases significantly also as temperature increases, and since it is compensated by nearly the same amount of increase in dynamics, the magnitude of B-factors appears nearly unchanged as a result. Furthermore, we use the all-atom sbNMA model<sup>13,14</sup> based on the CHARMM force field<sup>15</sup> to compute the magnitude of thermal vibrations of nearly 1,000 structures and find that it is nearly *protein-independent*, having a narrow distribution that peaks at  $0.093 \text{ \AA}^2$  at 100 K. We show that this universality in the magnitude of thermal vibrations is a direct corollary of the universality of vibrational spectra of globular proteins established earlier<sup>16–18</sup>. Put together, these results show that the percentage of thermal vibrations in B-factors is lowest at low resolution and low temperature (only 8-9% at 100 K and 2.0  $\text{\AA}$ ) and can become as high as 60% at high resolution (1.1  $\text{\AA}$  or higher) and high temperature.

**Significance.** Crystallographic B-factors provide important dynamical information for the internal mobility of proteins that is often linked to function. An insightful understanding of B-factors and the various influences on them has significant implications. First, it will allow researchers to better utilize its dynamic information and carry out meaningful comparisons between B-factors of different structure forms. For example, by comparing crystal structures of a given enzyme at apo and ligand-bound form, one may be able to determine how the binding has modified mobility at the active site<sup>2</sup>. Secondly, understanding B-factors should be significant also to X-ray crystallography and structure refinement. One of the earliest models for interpreting B-factors attributed the observed atomic displacement  $\langle u_i^2 \rangle$  to rigid body motions of molecules in crystals and was carried out by Cruickshank<sup>19</sup>, who interpreted atom displacements in the crystal with a translation tensor and a libration tensor. The work was later expanded by Schomaker and Trueblood<sup>20</sup> in their well-known translation, libration, and screw (TLS) model<sup>20</sup>, which is still widely used today in X-ray structure refinement<sup>21</sup>. Alternatively, normal mode analysis (NMA) for macromolecules<sup>22–24</sup>, developed in the 1980’s, also was used to understand atomic displacement in B-factors and a normal mode refinement



method was developed<sup>25,26</sup>. A good understanding of B-factors may inspire new refinement methods to be developed that may result in better structure quality and in turn, better understanding of protein dynamics. Thirdly, crystallographic B-factors have been widely used as a benchmark in evaluating elastic network models. In most of these evaluations, B-factors were, mistakenly, used interchangeably as the magnitudes of thermal vibrations. A better understanding of B-factors, especially the amounts of thermal vibrations that are contained in them, can allow more accurate and meaningful assessments to be performed. It can provide also insightful guidance on structure dataset selection when performing such an assessment.

## 2 Methods

### 2.1 The sbNMA model

The spring-based NMA, or sbNMA, was developed by Na and Song<sup>13</sup> in 2014. It is an all-atom model based on the CHARMM force field<sup>15</sup>, though it is extendable to other force fields as well. It is designed to maintain the high accuracy of the classical normal mode analysis (cNMA) using all-atom force fields and at the same time circumvent its cumbersome step of energy minimization, which makes cNMA inconvenient to use. The most significant contribution of sbNMA thus is that it maintains mostly the accuracy of cNMA while not requiring energy minimization. It can be applied directly to experimental structures like elastic network models<sup>27-29</sup>. Extensive studies have shown that i) it can faithfully reproduce the vibrational spectrum of globular proteins<sup>18</sup> and capsids<sup>30</sup>, a feat that very few models are able to accomplish; ii) its mean square fluctuation predictions correlate nearly perfectly with cNMA, reaching a correlation of nearly 90%<sup>13</sup>; iii) it proves to be a better benchmark to evaluate elastic network models than B-factors<sup>14</sup>.

There are a few key components in sbNMA, which are described below in detail.

**The force field.** sbNMA uses the CHARMM force field<sup>15</sup>. The force field parameters are given in a parameter file that includes parameters for bond stretching, bond bending, torsional terms, improper terms, and van der Waals interactions. All the parameters of these five types are extracted and saved in arrays. The remaining CHARMM parameters are currently not used in sbNMA. These include primarily parameters for electrostatic interactions, as well as CMAP. It is difficult to include electrostatic interactions in sbNMA since they yield large negative spring constants<sup>13</sup>. van der Waals interactions may introduce a negative spring constant too but that occurs only at large separation distances when it becomes very weak and thus negligible (see Fig. 1 in Ref. 13). Additionally, electrostatic interactions were found to contribute much less than van der Waals interactions<sup>13</sup>. A possible alternative is to include an explicit term for hydrogen bonds, which may compensate somewhat for the absence of electrostatic interactions. This may get included in a future release of sbNMA. The combined contribution of van der Waals interactions and electrostatics was found to be well approximated by a harmonic term with a distance-dependent force constant<sup>31</sup>.

**Actual interactions of a protein.** To perform normal mode analysis on a given protein system and to compute its Hessian matrix, one needs to first determine all the interactions in the system. While this part is greatly simplified in contact-based elastic network models, it is much more involved in sbNMA, since one needs to determine all the 2-body, 3-body, and 4-body interactions. To determine these terms, sbNMA uses the psfgen (or autopsf) script from VMD<sup>32</sup>. Given a pdb file, psfgen/autopsf is able to add missing atoms and guess their coordinates, add disulphide patches automatically, and generate both a pdb and a protein structure file (or a psf file). For example, given a pdb file 1ubq.pdb, the psf file generated by VMD contains all the bond stretching, bond bending, torsional, and improper interactions. These interaction pairs, triplets, quadruplets are extracted from the psf file and used in sbNMA computations. The van der Waals interactions are computed directly based on pairwise distances among the atoms.

**Building the Hessian matrix.** Once we have the interaction parameters and the lists of interactions, we can construct the Hessian matrix and the mass matrix. The bond stretching term and van der Waals term are 2-body interactions and their contributions to the Hessian matrix are easy to implement. The bond bending term (3-body) as well as torsional and improper terms (4-body) are implemented in a similar fashion to what was described in the work by Blondel and Karplus<sup>33</sup>.

The entire sbNMA code as well as the datasets used in this work is made publicly available. The sbNMA code can be found at <https://github.com/htna/sbNMA-Matlab>. The locations of the datasets are given later in the paper where they are discussed.

## 2.2 Mean-square displacement

The mean-square displacement of the  $i^{th}$  atom in a protein, or  $\langle u_i^2 \rangle$ , can be expressed as the summation of two key components, using a notation similar to Kurinov and Harrison’s<sup>12</sup>,  $\langle u_i^2 \rangle_d$  and  $\langle u_i^2 \rangle_{ld}$ , with the former representing the displacement due to dynamics and the latter the displacement due to lattice disorder. i.e.,

$$\langle u_i^2 \rangle = \langle u_i^2 \rangle_d + \langle u_i^2 \rangle_{ld}. \quad (4)$$

Lattice disorder happens because the protein molecules in a crystal’s unit cells may not have exactly the same conformation, nor identical position and orientation relative to the unit cell’s origin and cell axes<sup>1,2</sup>.

Lattice disorder was thought to be temperature independent<sup>2</sup>. However, later work found that lattice disorder increased significantly as temperature decreased<sup>12</sup>. Kurinov and Harrison attributed this decrease to shock freezing<sup>12</sup>. Lattice disorder also correlates strongly with structural resolution. Structural resolution is a sensitive outward indicator of the underlying static disorder, since crystals with low disorder will diffract well and reach high resolution

while crystals with high disorder will diffract poorly and end up with a low resolution.

$\langle u_i^2 \rangle_d$  represents the contribution from protein dynamics in the crystal and can be divided into contributions from thermal vibrations within unit cells and acoustic waves across unit cells<sup>7,11</sup> as shown in the Introduction section, i.e.,

$$\langle u_i^2 \rangle_d = \langle u_i^2 \rangle_v + \langle u_i^2 \rangle_{aw}. \quad (5)$$

$\langle u_i^2 \rangle_d$  is often denoted by  $\langle \Delta \mathbf{R}_i^2 \rangle$  in the normal mode analysis literature, i.e.,

$$\langle u_i^2 \rangle_d \equiv \langle \Delta \mathbf{R}_i^2 \rangle. \quad (6)$$

In this work, we use  $\langle u_i^2 \rangle_d$  and  $\langle \Delta \mathbf{R}_i^2 \rangle$  nearly interchangeably to denote the contribution from thermal vibrations: however,  $\langle u_i^2 \rangle_d$  is used mostly to represent the contribution to B-factors that is due to dynamics, while  $\langle \Delta \mathbf{R}_i^2 \rangle$  is used to represent calculated mean-square displacements. Note that in order to have the fairest comparison, acoustic waves, crystal packing as well as solvent contribution also should be considered when computing  $\langle \Delta \mathbf{R}_i^2 \rangle$ . Neither  $\langle u_i^2 \rangle_d$  nor  $\langle \Delta \mathbf{R}_i^2 \rangle$  should be confused with  $\langle u_i^2 \rangle$  itself (Eq. (1)), which represents the total mean-square displacements of atoms in protein crystals.

According to normal mode theory,  $\langle \Delta \mathbf{R}_i^2 \rangle$  can be computed from the inverse of the Hessian matrix. Most normal mode analysis work done in the past ignored the lattice disorder term  $\langle u_i^2 \rangle_{ld}$  in Eq. (4) altogether (as well as  $\langle u_i^2 \rangle_{aw}$ ) and thus linked the predicted  $\langle \Delta \mathbf{R}_i^2 \rangle$  value to the crystallographic B-factor as:

$$B_i = \frac{8\pi^2}{3} \langle \Delta \mathbf{R}_i^2 \rangle. \quad (7)$$

However, Eq. (7) holds true only if the lattice disorder term  $\langle u_i^2 \rangle_{ld}$  in Eq. (4) is negligible. In this work, along with earlier observations by others<sup>7,12</sup>, we show that  $\langle u_i^2 \rangle_{ld}$  should not be neglected since it is actually the dominant contribution to B-factors.

## 2.3 The magnitudes of mean-square displacement and thermal vibrations of a protein

While  $\langle u_i^2 \rangle$  and  $\langle \Delta \mathbf{R}_i^2 \rangle$  are used to represent the magnitudes of mean-square displacement and thermal vibrations, respectively, of individual atoms, in the rest of this paper we use  $\langle u^2 \rangle$  and  $\langle \Delta \mathbf{R}^2 \rangle$  to denote the magnitudes of mean-square displacement and thermal vibrations, respectively, of a whole protein, which are defined as the *mass-weighted* averages of  $\langle u_i^2 \rangle$  and  $\langle \Delta \mathbf{R}_i^2 \rangle$  over all the atoms, i.e.,

$$\langle \Delta \mathbf{R}^2 \rangle = \frac{\sum_i m_i \langle \Delta \mathbf{R}_i^2 \rangle}{\sum_i m_i}, \quad (8)$$

$$\langle u^2 \rangle = \frac{\sum_i m_i \langle u_i^2 \rangle}{\sum_i m_i}, \quad (9)$$

where  $i$  is the index of atoms. In the rest of the paper,  $i$  will be used systematically as the index of atoms while  $k$  the index of modes.

Alternatively, the magnitude of thermal vibrations can be computed using a subset of atoms such as all the  $\text{C}^\alpha$  atoms, as is commonly done. In such a case, a subscript is used and the magnitude is denoted by  $\langle \Delta \mathbf{R}^2 \rangle_{\text{C}^\alpha}$  or  $\langle u^2 \rangle_{\text{C}^\alpha}$ , to distinguish it from the magnitude computed using all the atoms.

## 3 Results

### 3.1 Protein Dataset

PDB\_SELECT<sup>34,35</sup> provides lists of representative protein chains with low mutual sequence identity selected from the protein data bank (PDB)<sup>36</sup>. For this work, we use a list of 1,522 proteins with 1.3 Å resolution or higher and R-factor 0.19 or better, and a longer list of 12,600 proteins with 2.0 Å resolution or higher and R-factor 0.20 or better. Both lists can be

found at <https://swift.cmbi.umcn.nl/gv/select/HTML/20191101>, maintained by Gert Vriend. For convenience and easy reference in the rest of the paper, we name the first list/dataset *pdb1522* and the second *pdb12600*, according to the numbers of proteins in the lists. All pdb files are then downloaded from PDB website (<https://www.rcsb.org/>)<sup>36</sup>. We choose structures whose resolution is 2.0 Å or higher and R-factor of 0.20 or better as it was found that a resolution of 2.0 Å is the lowest at which systematic errors in B-factors are insignificant<sup>2</sup>.

The vast majority of the crystal structures were determined at around 100 K<sup>37</sup>. 100 K is the typical cryogenic temperature in the present-day structure determination by crystallography to reduce radiation damage. A gaseous nitrogen stream held at 90–110 K is used to cool the crystals during the data collection<sup>37</sup>. Fig. 1 shows the histogram of temperatures at which the 1522 structures in *pdb1522* were determined. Not surprisingly, most of these crystal structures were determined at a temperature around 100 K. There are a small number of structures whose pdb files contain NULL values instead of temperature information. These structures are assigned a temperature of 0 K in the histogram plot in Fig. 1(A) (the leftmost bar). Only 32 (out of 1522, or 2%) structures are determined at a temperature higher than 270 K. Fig. 1(B) shows the distribution of chain lengths of the structures in the data set, ranging from a few dozen to over 700 residues. Fig. 1(C) shows the distribution of resolution, ranging from 0.48 Å to 1.29 Å.

To include a larger number of room temperature structures in our study, we also use the larger dataset *pdb12600* (whose list of proteins can be found at <https://swift.cmbi.umcn.nl/gv/select/HTML/20191101/PDB.LIS-20191101-2.0-0.20>). In this dataset, there are 330 structures that were determined at temperature 270 K or higher. There are 10,574 structures determined at 100 K. Some structures in the *pdb12600* dataset, e.g. 2q4o.pdb, contain more than one model. In these cases, only the first model is used. Fig. 2 shows statistics for the *pdb12600* dataset. Whereas the distribution of resolutions extends to significantly lower

resolutions compared to the pdb1522 dataset (Fig. 1), the distribution of chain lengths in both datasets are very similar.

### 3.2 Statistical Analysis of Crystallographic B-factors

Based on dataset *pdb12600*, Fig. 3 shows the magnitude of the mean square displacement  $\langle u^2 \rangle$  computed from crystallographic B-factors, as a function of structural resolution, for two subsets of structures: a large subset of 10,574 structures determined at 100 K, and a small subset of 330 structures determined at room temperature, between 270 K and 300 K. The structures in each subset were first sorted by resolution and then grouped into 10 equal-size bins (in the increasing order of resolution). Within each bin, the median  $\langle u^2 \rangle$  and the median resolution were computed and used to plot Fig. 3.

Remarkably, the magnitudes of mean square displacement  $\langle u^2 \rangle$  at the two temperatures are nearly the same, especially at higher resolution, in the range of 1.1-1.6 Å. What is also remarkable is that at both temperatures,  $\langle u^2 \rangle$  decreases nearly linearly as resolution increases.  $\langle u^2 \rangle$  is over 1.0 Å<sup>2</sup> at resolution 1.95 Å and becomes 0.5 Å<sup>2</sup> at resolution 1.1 Å, reduced by half.

This approximately linear decrease is not totally surprising as the well-known theory behind the Wilson plot<sup>38</sup> implies a decrease of the average B-factors with increasing crystal resolution for a model of an unstructured unit cell. The Wilson plot shows

$$\ln \frac{\langle I(s) \rangle}{\sum_k f_k(s)^2}, \quad (10)$$

as a function of  $s^2$ , where  $s$  is the wavenumber of the Bragg peaks.  $I(s)$  is the measured intensity, averaged over the peaks in a small  $s$  interval. The denominator depends only on the system's atomic scattering factors  $f_k(s)$  and can be considered a normalization factor. The Wilson plot is thus a histogram of logarithmic normalized intensities. For an unstructured

system, in which the atoms in the unit cell are distributed randomly, the intensities are given by

$$I(s) = I_0 \cdot e^{-2Bs^2}, \quad (11)$$

where  $I_0$  is proportional to the incoming beam intensity, and the Wilson plot is a straight line with slope  $-2B$ . Deviations from a straight line indicate deviations from a random distribution of atoms and thus an interesting structure. Many features in these deviations can be interpreted directly, without requiring a refined structure<sup>39</sup>.  $B$  is often considered an “average” B factor, though it is not a simple average over the atomic Debye-Waller factors, nor in fact solely attributable to thermal fluctuations.

The resolution  $d$  of the crystal is defined as  $d = 1/s_{\max}$ , where  $s_{\max}$  is the largest wavenumber at which there are detectable Bragg peaks. They correspond to the smallest detectable intensity  $I_{\min}$ , which, like  $I_0$ , is a parameter of the experimental setup. This means that  $I_0 \cdot e^{-2Bs_{\max}^2} = I_{\min}$  or

$$B = A \cdot d^2, \quad (12)$$

where  $A = -\frac{1}{2} \ln(I_{\min}/I_0)$  is a positive number collecting all dependencies on the experimental setup. Assuming that the variation of  $A$  among crystallographic experiments is not strongly correlated with the resolution of the collected data, one should thus expect B factors to decrease with increasing crystal resolution. The decrease appears more linear than quadratic in Fig. 3 for two reasons: (1) we are looking at a very small range of  $d$  and (2) there are important deviations from Wilson statistics in this particular range of  $d$ , which corresponds to the length scales of chemical bonds in organic molecules<sup>39</sup>.

Lastly, Fig. 3 shows that if only C $^{\alpha}$  atoms are considered, the magnitudes of mean-square displacements are significantly smaller, by as much as 0.07-0.08 Å<sup>2</sup>, as shown by the thin lines in Fig. 3. Table 1 shows further the magnitudes of mean square displacements of other subsets of atoms. It is evident from the table that side chain atoms have a significantly



higher mean-square displacement than backbone atoms. Since backbone and side chain atoms should experience a similar magnitude in most of the static disorder, the difference is most likely due to dynamics (or thermal vibrations). This will be tested in a later section, where the magnitudes of thermal vibrations are computed.

So what does Fig. 3 tell us? Recall that B-factors have two key components  $\langle u^2 \rangle_v$  and  $\langle u^2 \rangle_{ld}$ , representing contributions from thermal vibrations and lattice disorder respectively. Since  $\langle u^2 \rangle$  is nearly independent of temperature, its fast decrease with increasing resolution seen in Fig. 3 has to come from static lattice disorder. This implies that lattice disorder must be the dominant contribution at low resolution.

Moreover, the temperature independence of  $\langle u^2 \rangle$  also implies that lattice disorder has to be the dominant contribution to B-factors at low temperature (100 K), irrespective of the resolution. Consider the percentage  $p_T$  of the vibrational contribution to the total mean-square displacement at temperature  $T$ . Experimental evidence on the glass transition in proteins<sup>40</sup> shows that  $\langle u^2 \rangle_v$  grows superlinearly with  $T$  above  $T_g$ , the glass transition temperature of proteins that is 200-220 K<sup>40</sup>. It is thus at least three times as high at 300 K as it is at 100 K, whereas  $\langle u^2 \rangle$  is almost the same at both temperatures. This implies

$$p_{100} \leq \frac{1}{3} p_{300}. \quad (13)$$

Since  $p_{300}$  cannot exceed 100%, the upper limit for  $p_{100}$  is 33.3%, and probably both percentages are much lower than their theoretical maximum. This is consistent with observations made by Kurinov and Harrison<sup>12</sup> and by Hinsen<sup>7</sup> on a small sample of proteins, that at  $T=100$  K static lattice disorder  $\langle u^2 \rangle_{ld}$  is several times higher than thermal vibrations  $\langle u^2 \rangle_v$ . As temperature increases,  $\langle u^2 \rangle_v$  increases and takes on a higher percentage of B-factor. Our statistical analysis of a large set of protein structures shows that this phenomenon is prevalent throughout the PDB database.

Fig. 4 shows the box plots (with notches) of B-factors in each resolution range, with the median resolutions of each range given at the bottom of the figure. The median  $\langle u^2 \rangle$  values for the 100 K case (Fig. 4(A)) are statistically significantly different at different resolutions, since the notches on the box plots are narrow and hardly overlap. For the 300 K case (Fig. 4(B)), some of the notches between neighboring resolutions overlap, probably due to the fact that the number of data points, or structures determined at 300 K, is much smaller. Nevertheless, the overall trend is the same as that in Fig. 4(A).

A curious question naturally arises: what percent of B-factors is thermal vibrations? This is addressed in the next section.

### 3.3 The magnitude of thermal vibrations in proteins

In order to answer this question, we obviously need to find a different way to obtain the magnitude of thermal vibrations than crystallographic B-factors, since the latter are tainted with static disorder.

Thankfully, accurate force fields such as CHARMM<sup>15</sup>, Amber<sup>41</sup>, GROMOS<sup>42</sup>, etc. have been developed for molecular modeling, MD simulations, and normal mode computations. Parameters in these force fields are calibrated by fitting to experimental data for small molecules, such as vibrational spectra, and also to computational results from quantum chemistry. Therefore, one can compute the magnitude of thermal vibrations in proteins using these force fields, using either MD simulations<sup>9</sup> or NMA<sup>14</sup>. The drawback of running MD simulations is that it requires fairly tedious preparation steps such as solvation, minimization, and equilibration. Therefore, it is difficult to apply this technique to a large set of proteins. Moreover, MD simulations for macromolecules notoriously suffer from insufficient sampling for proteins that have flexible parts.

Since our goal is to compute just the magnitude of thermal vibrations, another approach better suited for this task is classical normal mode analysis<sup>22-24</sup>, or cNMA. cNMA starts

from the same force fields as MD, but constructs a harmonic approximation to the potential energy surface around a local minimum. In general, this harmonic approximation leads to a systematic underestimation of thermal vibration amplitudes. For protein crystals at cryogenic temperatures, which are clearly below the protein glass transition that occurs around 200 K, it can however be expected to be valid<sup>40</sup>. At such low temperatures, quantum zero-point fluctuations are another candidate for introducing non-linearities into the temperature-fluctuation dependence. We are not aware of any estimations of their impact in protein crystals, but for simpler inorganic crystals, they were shown to be negligible at 100 K<sup>43</sup>.

To run cNMA on a protein, a structure taken from PDB has to be first energetically minimized. The process is cumbersome and, unfortunately, causes significant deviations from the original PDB structures (for a recent study on this, see Ref. 13). To overcome this, Na and Song<sup>13</sup> developed a spring-based all-atom NMA model called sbNMA that maintains most of the accuracy of cNMA and yet can be applied directly to PDB structures, without the need for energy minimization<sup>13</sup>.

sbNMA<sup>13</sup> (see Methods section) almost perfectly correlates with cNMA<sup>13</sup> in predicting thermal vibration amplitudes. Additionally, the vibrational spectrum it predicted matches nearly perfectly with that of cNMA and with experimental results<sup>18</sup>. For these reasons, sbNMA seems to be the best model for computing the magnitude of thermal vibrations in proteins. In the following, we first present a novel theoretical result on the universality of the magnitude of thermal vibrations among globular proteins, followed by computations from sbNMA.

### 3.3.1 The magnitude of thermal vibrations is universally the same among all globular proteins: a theoretical result

One significant theoretical finding in the present work is that the magnitude of thermal vibrations is universally the same among all globular proteins. This universality can be proven as a direct corollary of the universality of vibrational spectra of globular proteins that was established in earlier work<sup>16,18,30</sup>.

Let  $g(\omega)$  denote the normalized density of vibrational modes. The universality of the vibrational spectrum means that all globular proteins have the same  $g(\omega)$ . Fig. 5 shows the vibrational spectrum of ubiquitin (1ubq.pdb) as computed from sbNMA.

A significant corollary of this universality is another universality: the magnitude of thermal vibrations of all globular proteins is also universally the same.

*Proof.* First, because  $g(\omega)$  is the same for all proteins, the following mean over the range of  $\omega$  is a constant, which we denote by  $\frac{1}{\omega_p^2}$ ,

$$\left\langle \frac{1}{\omega^2} \right\rangle = \int \frac{1}{\omega^2} g(\omega) d\omega = \text{const} = \frac{1}{\omega_p^2}. \quad (14)$$

We call  $\omega_p$  the *characteristic frequency of proteins*. Note that the integral is well-defined because  $g(\omega) = 0$  for  $\omega < \omega_{\min}$ , where  $\omega_{\min} > 0$  is the frequency of the slowest vibrational mode of the protein.

Next, let  $\mathbf{H}$  and  $\mathbf{M}$  be the Hessian and mass matrices, respectively, and  $\mathbf{v}_k$  and  $\lambda_k$  the  $k^{\text{th}}$  mode and eigenvalue. Let  $\mathbf{H}_m$  be the mass-weighted Hessian matrix and  $\mathbf{q}_k$  its orthonormal eigenvectors. We have:

$$\mathbf{H}_m = \mathbf{M}^{-1/2} \mathbf{H} \mathbf{M}^{-1/2}, \quad (15)$$

$$\mathbf{H}_m \mathbf{q}_k = \lambda_k \mathbf{q}_k, \quad (16)$$

$$\mathbf{q}_k = \mathbf{M}^{1/2} \mathbf{v}_k. \quad (17)$$

Now since each mode's mean potential energy amounts to  $\frac{1}{2}k_B T$ , the amplitude  $A_k$  of mode  $\mathbf{v}_k$  can be determined from:

$$\frac{1}{2}A_k^2 \mathbf{v}_k^T \mathbf{H} \mathbf{v}_k = \frac{1}{2}k_B T, \quad (18)$$

where  $k_B$  is Boltzmann constant and  $T$  is temperature. Thus, we have,

$$\frac{1}{2}A_k^2 \lambda_k = \frac{1}{2}k_B T, \quad (19)$$

which leads to  $A_k^2 = \frac{k_B T}{\lambda_k}$ .

Lastly, since

$$\langle \Delta \mathbf{R}_i^2 \rangle = \sum_k A_k^2 \mathbf{v}_{i,k}^2 = \sum_k \frac{k_B T}{\lambda_k} \mathbf{v}_{i,k}^2, \quad (20)$$

combined with Eq. (8) and the fact that  $\sum_i m_i \mathbf{v}_{i,k}^2 = 1$ , the magnitude of thermal vibrations is:

$$\langle \Delta \mathbf{R}^2 \rangle = \frac{\sum_i m_i \langle \Delta \mathbf{R}_i^2 \rangle}{\sum_i m_i} = \frac{k_B T \sum_k \frac{1}{\lambda_k} \sum_i m_i \mathbf{v}_{i,k}^2}{\sum_i m_i} = k_B T \frac{\sum_k \frac{1}{\lambda_k}}{\sum_i m_i} = k_B T \frac{\sum_k \frac{1}{\omega_k^2}}{\sum_i m_i}. \quad (21)$$

The last summation on the numerator can be approximated with an integration. Recalling that there are  $3N-6$  non-zero modes, we have,

$$\langle \Delta \mathbf{R}^2 \rangle = \frac{(3N-6)k_B T}{\sum_i m_i} \int \frac{1}{\omega^2} g(\omega) d\omega = \frac{3k_B T}{m_p \omega_p^2}, \quad (22)$$

where  $m_p$  is defined as the *characteristic average mass* of atoms in a protein and is,

$$m_p = \frac{\sum_i m_i}{N} \cdot \frac{N}{N-2} \approx \frac{\sum_i m_i}{N}. \quad (23)$$

Our computation (see details in next section) shows  $m_p$  is nearly the same among all globular

proteins and has a median value of 7.15 g/mol, which is reasonable since about half of the atoms in proteins are hydrogens.

In conclusion, Eq. (22) shows that the magnitude of thermal vibrations is universally the same for all globular proteins. The magnitude is dictated by the characteristic average mass  $m_p$  and characteristic frequency of proteins  $\omega_p$ , and is linearly proportional to temperature within the limits of validity of the harmonic potential approximation.  $\square$

Readers familiar with the theory of liquids may recognize the similarity of the first equality in Eq. (22) with the Green-Kubo relation linking the diffusion constant to the velocity autocorrelation function and its Fourier transform, the density of states<sup>44</sup>. In fact, it can be seen as a variant of that relation for systems confined to a finite phase space volume, which is the case for the internal dynamics of a protein constrained by its covalent bonds. The use of normal modes in the above derivation is only a convenience, not a necessity, and the relation is therefore not limited to the harmonic potential approximation.

What is the biological significance of the characteristic frequency  $\omega_p$ ? This is a very interesting question. The characteristic frequency  $\omega_p$  is clearly different from the frequencies of the slowest vibrational modes that often correspond to functional motions. It is, however, proportional to the overall magnitude of thermal vibrations (Eq. 22), and thus characterizes the overall flexibility of proteins. The fact that all proteins share a similar characteristic frequency  $\omega_p$  suggests that the material properties of proteins are similar, for example, having a similar elastic modulus<sup>45</sup>. This is a consequence of the fact that all proteins share the same elements of primary structure (amino acid residues) and secondary structure (helices, sheets, and loops).

### 3.3.2 The magnitude of thermal vibrations from sbNMA computations

In practice,  $\omega_p$  varies somewhat from protein to protein. We apply sbNMA to all structures in the *pdb1522* dataset (which is smaller than *pdb12600* and has much higher resolution) that

were determined at 80-120K and without broken chains. There are 1,197 such structures. For some of these structures, autopsf from VMD<sup>32</sup> is not able not to generate protein structure files (psf files). In the end, there are 984 structures with protein structure files successfully generated. We apply sbNMA to all these structures. A list of PDB-ids of these 984 structures and the magnitudes of their thermal vibrations computed from sbNMA<sup>46</sup>, as well as their autopsf-generated pdb and psf files are publicly available at <https://github.com/htna/Bfactors>.

Our computation shows  $\omega_p$  (see Fig. 6) of most proteins falls in the range of [75-125]  $\text{cm}^{-1}$  and has a median value of 102.7  $\text{cm}^{-1}$ , which is close to the first peak in the vibrational spectrum (see Fig. 5) that is at about 80  $\text{cm}^{-1}$ .  $\omega_p$  is also found to be independent of protein size (Fig. S3). Fig. 6 also shows the distribution of  $m_p$ , with a median value of 7.15 g/mol.

Using the characteristic frequency of proteins  $\omega_p$ , which is 102.7  $\text{cm}^{-1}$ , and the characteristic average mass  $m_p$ , which is 7.15 g/mol, the expected magnitude of thermal vibrations at 100 K is,

$$\begin{aligned}
\langle \Delta \mathbf{R}^2 \rangle &= \frac{3k_B T}{m_p \omega_p^2} = \frac{0.5961 \text{ kcal/mol}}{7.15 \text{ g/mol} \times (102.7 \text{ cm}^{-1} \times 2\pi c)^2} \\
&= \frac{0.5961 \times 4,184 \text{ Joule/mol}}{7.15 \times 10^{-3} \text{ kg/mol} \times (102.7 \text{ cm}^{-1} \times 2\pi c)^2} \\
&= 0.093 \times 10^{-20} \text{ m}^2 = 0.093 \text{ \AA}^2,
\end{aligned} \tag{24}$$

where  $c$  is the speed of light and is  $2.997925 \times 10^{10}$  cm/sec.

Fig. 7 shows the histogram of the magnitude of thermal vibrations  $\langle \Delta \mathbf{R}^2 \rangle$  for these 984 structures. The distribution has a sharp peak at around 0.093  $\text{\AA}^2$ . This magnitude of thermal vibrations is independent of the resolution of the structures selected (see Figs. S1 and S2 in Supplemental Information). To put this magnitude into perspective, 0.01  $\text{\AA}^2$  is the zero-point vibration limit<sup>47</sup>. The variance displayed in the range is possibly due to variations in packing density. Note that in our computation using sbNMA we consider only contributions

of normal modes of a free protein molecule. Neither crystal packing nor the acoustic modes in crystal are taken into account. Including crystal packing would reduce the magnitude of the predicted  $\langle u^2 \rangle_d$  marginally, while the inclusion of the acoustic modes would increase it (by 14% for lysozyme<sup>7</sup>). Therefore it is reasonable to conclude that the actual magnitude of  $\langle u^2 \rangle_d$  is still around  $0.093 \text{ \AA}^2$  at 100 K.

Table 2 shows the magnitudes of thermal vibrations of different atom types computed from sbNMA. Note that the magnitude of thermal vibrations is similar among backbone atoms but is about twice as high at side chains. This is consistent with Table 1, which shows the magnitudes of mean-square displacements of the same groups of atoms from crystallographic B-factors. Both tables indicate that side chain atoms are significantly more mobile than backbone atoms, but the effect is much more pronounced for the thermal vibrations, with  $\langle \Delta \mathbf{R}^2 \rangle_{SC} / \langle \Delta \mathbf{R}^2 \rangle_{C^\alpha}$  being almost 2, compared to  $\langle u^2 \rangle_{SC} / \langle u^2 \rangle_{C^\alpha}$  around 1.2. Notice that the magnitude of thermal vibrations shown in Table 2 is about an order smaller than B-factors (Table 1), implying that the dominant contribution to B-factors is not thermal vibrations, but static disorder. One implication of this is that one should not limit oneself to  $C^\alpha$  atoms when estimating the magnitude of thermal vibrations in a crystal. Calculations using  $C^\alpha$  atoms only would underestimate the magnitude of thermal vibration by 30-40% (compared to using all atoms). Lastly, we note there are statistical uncertainties regarding the reported magnitude of  $0.093 \text{ \AA}^2$  at 100 K. As shown in Table 2, the range between the 25th and 75th percentiles is  $[0.077, 0.128] \text{ \AA}^2$ .

Considering that a unit of  $k_B T$  at 100 K is about 0.2 kcal/mol, we have:

$$\frac{1}{2} k_{eff} \langle \Delta \mathbf{R}^2 \rangle = \frac{3}{2} k_B T \quad (25)$$

with

$$k_{eff} = \frac{3k_B T}{\langle \Delta \mathbf{R}^2 \rangle} = \frac{3 \times 0.2}{0.093} = 6.4 \text{ kcal/mol/\AA}^2. \quad (26)$$



This implies that the average effective spring constant constraining each atom is about 6.4 kcal/mol/Å<sup>2</sup>.

$k_{eff}$  gives us a sense of how much each atom should be constrained in a given model. For example, if a model has a force constant in the order of 1 kcal/mol/Å<sup>2</sup>, in order for each atom to receive on average a total constraint of about 6.4 kcal/mol/Å<sup>2</sup>, the model would need to choose an interaction range that gives each residue/atom effectively 6-7 neighbors on average.

### 3.4 The percentages of thermal vibrations and lattice disorder in B-factors

Using the magnitude of thermal vibrations computed from sbNMA, which is about 0.093 Å<sup>2</sup> at 100 K for most proteins (Fig. 7), Fig. 8 shows the breakdown of  $\langle u^2 \rangle$  (solid lines) into its two key components: thermal vibrations  $\langle u^2 \rangle_v$  (dotted-dashed lines) and lattice disorder  $\langle u^2 \rangle_{ld}$  (dashed lines) at 100 K (in blue) and 300 K (in red). At 300 K, the lower bound (since  $\langle u^2 \rangle_v$  grows superlinearly with temperature) of the magnitude of thermal vibrations  $\langle u^2 \rangle_v$  is 0.279 Å<sup>2</sup>, thrice as high as the magnitude of thermal vibrations at 100 K, whereas our estimate of  $\langle u^2 \rangle_{ld}$  at 300 K is an upper bound.

There are a number of observations from Fig. 8 that are interesting.

*Lattice disorder decreases as resolution increases.* The data in Fig. 8 shows convincingly that static lattice disorder decreases, nearly perfectly linearly, as the resolution increases, decreasing so much that when compared with that at resolution 1.95 Å, it is reduced by more than half at resolution 1.1 Å. If this trend remains true and we do a linear extrapolation to the right, we expect that lattice disorder becomes 0 at resolution around 0.5 Å (more on this later). This resolution dependence was noted by Petsko and Ringe, that “protein crystals that diffract to very high resolution should have small lattice disorder contributions.”<sup>2</sup>.

*Lattice disorder decreases as temperature increases.* If the above resolution-dependence is expected, what comes as a little surprising is that lattice disorder is not temperature-independent as commonly thought<sup>2</sup>, but decreases also as the temperature increases from 100 K to 300 K, by an amount of 0.12 Å<sup>2</sup> or more. The amount of decrease is consistent for all resolutions. This observation is not totally new but noted by Kurinov and Harrison in their study of lysozyme<sup>12</sup>, who suggested that a strong increase in lattice disorder should happen at low temperature due to shock freezing.

*Other potential contributions to lattice disorder.* Another key question is whether or not static lattice disorder depends on crystal forms. This is probably true, and possibly on many other factors. Fig. 4 shows that there is still some sizable variation of static disorder within a given resolution bin (i.e., even when resolution and temperature are fixed.) The variations may be due to crystal form, experimental condition, proteins themselves, etc.

*The magnitude of thermal vibrations  $\langle u^2 \rangle_v$  increases with temperature but is independent of resolution.* As an intrinsic property of proteins, the magnitude of thermal vibrations clearly should not depend on resolution. On the other hand, it is linearly proportional to temperature (see Eq. (20)) between the Debye temperature  $T_D$ , below which  $\langle u^2 \rangle_v$  is temperature independent and equal to the zero-point vibration<sup>2</sup>, and the glass transition temperature of about 200 K. The amplitude of the thermal vibrations at least triples as temperature increases from 100 K to near 300 K.

*The gap between contributions from  $\langle u^2 \rangle_v$  and  $\langle u^2 \rangle_{ld}$  narrows as resolution and temperature increase.* Fig. 8 shows that the gap between contributions from lattice disorder (dashed line) and dynamic vibrations (dotted-dashed line) narrows as temperature and/or resolution increase. Consequently, at high temperature (300 K)  $\langle u^2 \rangle_v$  overtakes  $\langle u^2 \rangle_{ld}$  at high resolution (1.1 Å). The overtake may take place at a lower resolution than 1.1 Å considering the superlinear growth of  $\langle u^2 \rangle_v$  with  $T$  above  $T_g$ , the glass transition temperature. This is in agreement with what Kurinov and Harrison concluded for lysozyme after comparing

B-factors of lysozyme with results from Mössbauer spectroscopy<sup>12</sup>.

*Thermal fluctuations  $\langle u^2 \rangle_v$  become the dominant contribution to B-factors at sub-1.1 Å resolution at 300 K.* At 300 K and at 1.1 Å resolution, the amount of thermal fluctuations  $\langle u^2 \rangle_v$  is at least on par with that of lattice disorder  $\langle u^2 \rangle_{ld}$ . By linear extrapolation (to the right), it is expected that the contribution of lattice disorder should become even lower at higher resolutions than 1.1 Å. The amount of thermal fluctuations should be higher than lattice disorder at sub-1.1 Å resolution at 300 K, while an even higher resolution (around 0.5 Å, which is the current technological limit of X-ray crystallography) is required for this to happen for structures determined at 100 K. As a result, thermal vibrations will become dominant in B-factors. However, structures with such resolution and temperature are rare in the PDB. Fig. 9 shows the amount of total mean-square displacement  $\langle u^2 \rangle$  at sub 1.1 Å resolution. Since  $\langle u^2 \rangle_v$  remains unchanged at 0.093 Å<sup>2</sup>, it shows that indeed that  $\langle u^2 \rangle_{ld}$  decreases further down to be about the same as  $\langle u^2 \rangle_v$  at 0.5-0.6 Å resolution. The lower bound of  $\langle u^2 \rangle$  of about 0.13 Å<sup>2</sup> at 100 K sets the upper limit for  $\langle u^2 \rangle_v$  at the same temperature. This upper limit is consistent with the magnitude of  $\langle u^2 \rangle_v$  predicted by sbNMA (Fig. 7), providing an additional support to the reliability of the sbNMA computations.

*The magnitude of B-factors is nearly temperature independent.* Lastly, it is remarkable that the magnitude of B-factors is nearly the same at the two temperatures (the solid red and blue lines in Fig. 8). The increase in thermal fluctuations at higher temperature is compensated by reduction in static disorder, by nearly exactly the same amount. A similar phenomenon was observed by Kurinov and Harrison in their analysis of B-factors of lysozyme<sup>12</sup>. It is worth noting here that only the magnitude of B-factors is nearly temperature independent, not B-factors themselves.

*Summary.* So what are the percentages of thermal fluctuations and lattice disorder in B-factors? It depends both on the resolution and the temperature at which a structure is determined. The percentage of thermal vibrations in B-factors is at its lowest at low

temperature and low resolution (as surprisingly low as 8-9%), and is at its highest at high temperature and high resolution (to nearly 60% or perhaps even higher). The finding that the percentage of  $\langle u^2 \rangle_v$  at low temperature (100K) and at low resolution (about 2.0 Å) is only 8-9% may seem surprising, but it agrees perfectly with what Hinsen found for lysozyme from his HCA model<sup>48</sup> after calibrating it with experimental compressibility data<sup>7</sup>.

### 3.5 A case study on lysozyme

In the aforementioned study on lysozyme crystals, Hinsen<sup>7</sup> found that thermal vibrations, especially at low temperature, contribute only a tiny fraction to the B-factors. The two lysozyme crystal structures used in his work are 1IEE<sup>49</sup> and 2LYM<sup>50</sup>. The properties of these two structures are summarized in Table 3. By fitting to the compressibility of lysozyme crystals obtained from experiments, Hinsen found that the total amount of thermal fluctuations, including both optical and acoustic modes, was surprisingly low, accounting for only 4.3% (1IEE) and 7.2% (2LYM) of the B-factors. Choosing the highest published value on compressibility increased their prediction on the contribution of thermal vibrations to B-factors to 9.7% for 1IEE and 16.3% for 2LYM<sup>7</sup>. A general caveat is that the compressibilities were measured for dried lysozyme crystals in order to improve their stability in the course of the experiment. It is possible that drying modifies the compressibility itself.

On the other hand, Kurinov and Harrison<sup>12</sup> estimated, by comparison with Mössbauer spectroscopy, that thermal vibrations in tetragonal lysozyme should account for 50% of the B-factors at 300 K. There is thus a significant difference regarding the percentage of dynamics in B-factor between Hinsen’s prediction based on compressibility and Kurinov and Harrison’s estimation based on Mössbauer spectroscopy. It should be noted that Mössbauer active nuclei in most proteins are scarce and there is inherent difficulty in interpreting the Mössbauer effect in protein crystal<sup>2</sup>, and consequently, Kurinov and Harrison’s result based on comparison with Mössbauer spectroscopy is debatable.

Here we compute the amount of thermal vibrations in 1IEE and 2LYM using sbNMA (see Methods section). Fig. 10 shows the predicted  $\langle u_i^2 \rangle_v$  (or  $\langle \Delta \mathbf{R}_i^2 \rangle$ ) for each residue as well as the mean-square displacement  $\langle u_i^2 \rangle$  from B-factors. Both  $\langle u_i^2 \rangle$  and  $\langle \Delta \mathbf{R}_i^2 \rangle$  represent mass-weighted averages over all the atoms within each residue. Taking the ratio between the magnitude of predicted thermal vibrations  $\langle \Delta \mathbf{R}^2 \rangle$  and mean-square displacement  $\langle u^2 \rangle$  (Eqs. (8) and (9)) from the B-factors of these two structures, we find that the percentage of thermal vibrations in B-factors is 17% in 1IEE (110 K) and 32% in 2LYM (300 K), with the latter being significantly less than the 50% estimated by Kurinov and Harrison<sup>2</sup> for lysozyme at 300 K. The difference, however, can be explained by the superlinear growth of thermal fluctuations above  $T_g$ <sup>40</sup>. If only C $^\alpha$  atoms are used, in order to allow a direct comparison with Hinsen’s method<sup>7</sup>, we find 12% and 26% for 1IEE and 2LYM respectively, which are larger than upper limits of 9.7% and 16.3% cited above.

Fig. 11 shows the differences between 1IEE and 2LYM in mean-square displacement,  $\langle u_i^2 \rangle$ , and the magnitude of thermal vibrations,  $\langle \Delta \mathbf{R}_i^2 \rangle$ . The figure reveals that part of the difference in B-factors between the two structures is attributed to increased harmonic vibrations due to the temperature increase (what is beneath the black line) and part of it must be due to anharmonic motions and static disorder (between the black line and the green line). 2LYM was determined at a significantly lower resolution than 1IEE and at a higher temperature (see Table 3). It is thus not surprising that its B-factors contain a greater amount of static disorder as well as anharmonic motions.

## 4 Discussion

In this work, using a very large dataset that contains over 10,000 sequence-nonredundant proteins, we carry out a thorough, statistically sound analysis of crystallographic B-factors. Combined with computations from all-atom sbNMA, our study reveals the following fun

facts about B-factors, including the amounts of dynamics and static disorder in them and their dependence on temperature and resolution. Specifically,

1. *Temperature-independence.* The magnitude of B-factors is nearly temperature independent. The magnitude of thermal vibrations, or  $\langle u^2 \rangle_v$ , increases with temperature, and since it is compensated by nearly the same amount of decrease in  $\langle u^2 \rangle_{ld}$ , the magnitude of B-factors appears nearly unchanged as a result (Fig. 3).
2. *Linearity.* The magnitude of B-factors decreases linearly as resolution increases (Figs. 3 and 9).
3. *The range of magnitude.* The magnitude of mean-square displacement, or  $\langle u^2 \rangle$ , vary greatly, from 0.1 to 1.1 Å<sup>2</sup> in the resolution range 0.5-2.0 Å, displaying a 10-fold increase (Figs. 3 and 9).
4. *Universality.* The magnitude of thermal vibrations, or  $\langle u^2 \rangle_v$ , in proteins is nearly universal, having a narrow distribution that peaks at 0.093 Å<sup>2</sup> at 100 K (Fig. 7).
5. *Percentage of dynamics in B-factors.* Consequently, the percentage of  $\langle u^2 \rangle_v$  in B-factor ranges from 8-9% (low resolution) to nearly 70% (high resolution) at 100 K, the temperature at which most structures were determined, again displaying a nearly 10-fold increase.

The temperature-independence and linearity of B-factor on resolution ( $d$ ) implies that the magnitude of mean-square displacement, or  $\langle u^2 \rangle$ , can be estimated by the following formula:

$$\langle u^2 \rangle_{estimate} = 0.1 + \frac{1.1 - 0.1}{2.0 - 0.5}(d - 0.5) = 0.1 + \frac{2}{3}(d - 0.5), \quad (27)$$

where the numerator and denominator reflects the ranges of changes in  $\langle u^2 \rangle$  and  $d$  respectively.

Accordingly, the percentage of dynamics in B-factors at 100 K, or  $p_{100}$  as defined earlier, can be estimated by the following formula:

$$p_{100} = \frac{\langle u^2 \rangle_v}{\langle u^2 \rangle_{estimate}} = \frac{0.093}{\langle u^2 \rangle_{estimate}} = \frac{0.093}{0.1 + \frac{2}{3}(d - 0.5)}. \quad (28)$$

Similarly, a lower bound for the percentage of dynamics in B-factor at 300 K, or  $p_{300}$ , can be estimated by:

$$p_{300} \geq \frac{0.279}{\langle u^2 \rangle_{estimate}} = \frac{0.279}{0.1 + \frac{2}{3}(d - 0.5)}, \quad (29)$$

which utilizes also the finding that  $\langle u^2 \rangle$  is temperature-independent.

Fig. 12, as a graphical rerendering of Fig. 8, well captures this dependence of B-factor and its two components  $\langle u^2 \rangle_{ld}$  and  $\langle u^2 \rangle_v$  on both temperature and resolution. The figure shows that the gap between  $\langle u^2 \rangle_{ld}$  and  $\langle u^2 \rangle_v$  narrows both with the increase in temperature (white arrows) and with the increase in resolution (green arrow). At low resolution and low temperature (marked on the leftmost end of Fig. 12),  $\langle u^2 \rangle_v$  accounts for only a tiny fraction of B-factor, while at high resolution and at high temperature, the contributions of  $\langle u^2 \rangle_{ld}$  and  $\langle u^2 \rangle_v$  become similar.

B-factors, being easily accessible and conveniently available, have been an appealing benchmark for assessing elastic network models and are used widely. The fact that lattice disorder was found to be the dominant contribution to B-factors prompted some authors to recommend against using B-factors as a benchmark. Our present work does not disagree that lattice disorder is the dominant contribution to B-factors at low temperature (or cryogenic temperature, about 100 K), but shows that the dynamics component can become at least as important at high resolution and at high temperature. This important realization brings back hope that B-factors of structures determined at *both high temperature and high resolution* still contain *rich* experimental data on protein dynamics that should be utilized.

One significant implication of this is that when evaluating elastic network models using

crystallographic B-factors, one should choose structures not only with high resolution, which is intuitively practiced mostly, but also determined at high temperature, which is not realized nor practiced yet. This finding prompts us to propose the following hypothesis: *correlation with crystallographic B-factors should see a statistically significant improvement when using structures determined at high resolution (1.1 Å or better) and at high (room) temperature.* Unfortunately, such structures are relatively rare in the PDB, as they are experimentally difficult to obtain because of radiation damage. It is our hope that our work here will renew interest in room-temperature crystallography, motivating experimental innovations (such as serial synchrotron crystallography<sup>6</sup>) and efforts to determine many such structures in the near future. Should this hypothesis prove true, it would help the NMA/ENM community identify the best dataset for evaluating elastic network models.

**The magnitude of thermal vibrations is nearly protein-independent.** Another interesting finding from this work is that the magnitude of thermal vibrations, or  $\langle u^2 \rangle_v$ , is nearly *protein-independent* with a universal value of 0.093 Å<sup>2</sup> at 100 K. This universality must reflect the packing density and the nature of interactions shared by crystallizable proteins. And we further prove that this universality is a direct corollary of the universality of the vibrational spectrum of globular proteins established earlier<sup>16,18,30</sup>. A number of evidences from crystallographic B-factors indirectly support that these predicted magnitudes for thermal vibrations are in the right range: (i) Out of the 10,574 structures determined at 100 K in the pdb12600 dataset, the first 3 smallest mean-square displacements are 0.134, 0.147, and 0.148 Å<sup>2</sup> (see also Fig. 9); (ii) Out of the 330 structures determined at 270-300 K in the pdb12600 dataset, the first 3 smallest mean-square displacement are 0.213, 0.262, and 0.294 Å<sup>2</sup>. These mean-square displacement values from crystallographic B-factors mark the experimental upper bound for the magnitude of thermal vibrations at these temperatures.

**Solvent effect.** In a protein crystal, there is plenty of space between protein molecules that is occupied by crystallization solvent. The amount of solvent in crystal is about 40-



60%<sup>51</sup>. Solvent effect has been considered in normal mode computations<sup>48,52–55</sup>. A recent work by Na and co-workers<sup>56</sup> provides a feasible approach to include the effect of solvent in normal mode computations. As part of future work, we plan to examine how much the magnitude of thermal vibrations is dampened by the solvent using Na and coworkers’ approach.

## References

- [1] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou. Temperature-dependent x-ray-diffraction as a probe of protein structural dynamics. *Nature*, 280(5723):558–563, 1979.
- [2] G. A. Petsko and D. Ringe. Fluctuations in protein structure from x-ray diffraction. *Annu Rev Biophys Bioeng*, 13:331–71, 1984.
- [3] David W. Rodgers. Cryocrystallography. *Structure*, 2(12):1135–1140, Dec 1994.
- [4] A. Meents, S. Gutmann, A. Wagner, and C. Schulze-Bries. Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proc Natl Acad Sci U S A*, 107(3):1094–9, 2010.
- [5] D. H. Juers and M. Weik. Similarities and differences in radiation damage at 100 k versus 160 k in a crystal of thermolysin. *J Synchrotron Radiat*, 18(Pt 3):329–37, 2011.
- [6] E. de la Mora, N. Coquelle, C. S. Bury, M. Rosenthal, J. M. Holton, I. Carmichael, E. F. Garman, M. Burghammer, J. P. Colletier, and M. Weik. Radiation damage and dose limits in serial synchrotron crystallography at cryo- and room temperatures. *Proc Natl Acad Sci U S A*, 117(8):4142–4151, 2020.
- [7] K. Hinsén. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24(4):521–8, 2008.

- [8] Y. Dehouck and U. Bastolla. The maximum penalty criterion for ridge regression: Application to the calibration of the force constant in elastic network models. *Integrative Biology*, 9(7):627–641, July 2017.
- [9] E. Fuglebakk, N. Reuter, and K. Hinsén. Evaluation of protein elastic network models based on an analysis of collective motions. *J Chem Theory Comput*, 9(12):5618–28, 2013.
- [10] S. Kundu, J.S. Melton, D.C. Sorensen, and Jr. Phillips, G.N. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys J*, 83(2):723–732, 2002.
- [11] D. Riccardi, Q. Cui, and Jr. Phillips, G. N. Application of elastic network models to proteins in the crystalline state. *Biophys J*, 96(2):464–75, 2009.
- [12] I. V. Kurinov and R. W. Harrison. The influence of temperature on lysozyme crystals. structure and dynamics of protein and water. *Acta Crystallogr D Biol Crystallogr*, 51(Pt 1):98–109, 1995.
- [13] H. Na and G. Song. Bridging between normal mode analysis and elastic network models. *Proteins*, 82:2157–2168, September 2014.
- [14] H. Na and G. Song. Conventional nma as a better standard for evaluating elastic network models. *Proteins*, 83(2):259–67, 2015.
- [15] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, April 1998.

- [16] D. ben Avraham. Vibrational normal-mode spectrum of globular proteins. *Phys Rev B Condens Matter*, 47(21):14559–14560, 1993.
- [17] K. Hinsen and G. R. Kneller. Projection methods for the analysis of complex motions in macromolecules. *Mol. Sim.*, 23:275–292, 2000.
- [18] H. Na, G. Song, and D. ben Avraham. Universality of vibrational spectra of globular proteins. *Phys Biol*, 13(1):016008, 2016.
- [19] D. W. J. Cruickshank. The analysis of the anisotropic thermal motion of molecules in crystals. *Acta Crystallographica*, 9(8):754–756, 1956.
- [20] V. Schomaker and K. N. Trueblood. On the rigid-body motion of molecules in crystals. *Acta Crystallographica Section B*, 24(1):63–76, Jan 1968.
- [21] B. Stec, R. Zhou, and M. M. Teeter. Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. *Acta Crystallogr D Biol Crystallogr*, 51(Pt 5):663–81, 1995.
- [22] M. Levitt, C. Sander, and P. S. Stern. The normal modes of a protein: Native bovine pancreatic trypsin inhibitor. *Int. J. Quant. Chem.*, 10:181–199, 1983.
- [23] N. Go, T. Noguti, and T. Nishikawa. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA*, 80(12):3696–3700, June 1983.
- [24] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA*, 80(21):6571–6575, November 1983.
- [25] A. Kidera and N. Go. Refinement of protein dynamic structure - normal mode refinement. *Proceedings of the National Academy of Sciences of the United States of America*, 87(10):3718–3722, 1990.

- [26] A. Kidera and N. Go. Normal mode refinement - crystallographic refinement of protein dynamic structure .1. theory and test by simulated diffraction data. *Journal of Molecular Biology*, 225(2):457–475, 1992.
- [27] M. M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.
- [28] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design*, 2(3):173–181, 1997.
- [29] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33(3):417–429, March 1998.
- [30] H. Na and G. Song. Fast normal mode computations of capsid dynamics inspired by resonance. *Phys Biol*, 15(4):046003, 2018.
- [31] K. Hinsen and G. R. Kneller. A simplified force field for describing vibrational protein dynamics over the whole frequency range. *Journal of Chemical Physics*, 111(24):10766–10769, 1999.
- [32] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J Mol. Graphics*, 14:33–38, 1996.
- [33] A. Blondel and M. Karplus. New formulation for derivatives of torsion angles and improper torsion angles in molecular mechanics: Elimination of singularities. *Journal of Computational Chemistry*, 17(9):1132–1141, 1996.
- [34] R. W. W. Hooft, C. Sander, and G. Vriend. Verification of protein structures: Side-chain planarity. *Journal of Applied Crystallography*, 29:714–716, 1996.

- [35] W. G. Touw, C. Baakman, J. Black, T. A. H. te Beek, E. Krieger, R. P. Joosten, and G. Vriend. A series of pdb-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368, 2015.
- [36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–42, 2000.
- [37] E. Garman. ‘cool’ crystals: macromolecular cryocrystallography and radiation damage. *Curr Opin Struct Biol*, 13(5):545–51, 2003.
- [38] A. J. C. Wilson. Determination of Absolute from Relative X-Ray Intensity Data. *Nature*, 150(3796):152–152, August 1942.
- [39] R. J. Morris, E. Blanc, and G. Bricogne. On the interpretation and use of  $\langle |E^2| \rangle (d^*)$  profiles. *Acta Crystallographica Section D Biological Crystallography*, 60(2):227–240, February 2004.
- [40] D. Ringe and G. A. Petsko. The ‘glass transition’ in protein dynamics: what it is, why it occurs, and how to exploit it. *Biophys Chem*, 105(2-3):667–80, 2003.
- [41] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, September 2000.
- [42] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the gromos force-field parameter sets 53a5 and 53a6. *J Comput Chem*, 25(13):1656–76, 2004.
- [43] C. Malica and A. Dal Corso. Temperature-dependent atomic  $B$  factor: An *ab initio* calculation. *Acta Crystallogr A Found Adv*, 75(4):624–632, July 2019.

- [44] Jean Pierre Hansen and Ian R McDonald. *Theory of Simple Liquids*. Elsevier, Amsterdam, 2006.
- [45] G. Song. Bridging between material properties of proteins and the underlying molecular interactions. *PLoS One*, 16(5):e0247147, 2021.
- [46] H. Na and G. Song. Bridging between normal mode analysis and elastic network models. *Proteins*, 82(9):2157–68, 2014.
- [47] B. T. M. Willis and A. W. Pryor. *Thermal vibrations in crystallography*. Cambridge University Press, London ; New York, 1975.
- [48] K. Hinsén, A. J. Petrescu, S. Dellerue, M. C. Bellissent-Funel, and G. R. Kneller. Harmonicity in slow protein dynamics. *Chemical Physics*, 261(1-2):25–37, 2000.
- [49] C. Sauter, F. Otalora, J. A. Gavira, O. Vidal, R. Giege, and J. M. Garcia-Ruiz. Structure of tetragonal hen egg-white lysozyme at 0.94 angstrom from crystals grown by the counter-diffusion method. *Acta Crystallographica Section D-Structural Biology*, 57:1119–1126, 2001.
- [50] C. E. Kundrot and F. M. Richards. Crystal-structure of hen egg-white lysozyme at a hydrostatic-pressure of 1000 atmospheres. *Journal of Molecular Biology*, 193(1):157–170, 1987.
- [51] H. Neurath and R. L. Hill. *The proteins : composition, structure and function*. Academic Press, New York, 3rd edition, 1975.
- [52] G. Lamm and A. Szabo. Langevin modes of macromolecules. *Journal of Chemical Physics*, 85(12):7334–7348, 1986.
- [53] J. Kottalam and D. A. Case. Langevin modes of macromolecules - applications to crambin and dna hexamers. *Biopolymers*, 29(10-11):1409–1421, 1990.

- [54] S. Hayward, A. Kitao, F. Hirata, and N. Go. Effect of solvent on collective motions in globular protein. *Journal of Molecular Biology*, 234(4):1207–1217, 1993.
- [55] A. Ansari. Langevin modes analysis of myoglobin. *Journal of Chemical Physics*, 110(3):1774–1780, 1999.
- [56] A. Bose Majumdar, I. J. Kim, and H. Na. Effect of solvent on protein structure and dynamics. *Phys Biol*, 17(3):036006, 2020.

## Figure Legends

### Figure 1.

Statistics of the pdb1522 dataset. (A) The distribution of crystallographic temperatures at which crystal structures in the dataset were determined. (B) The distribution of the lengths of protein chains in the dataset. (C) the histogram of resolution: the range is from 0.48 Å to 1.29 Å.

### Figure 2.

Statistics of the pdb12600 dataset. (A) The distribution of the lengths of protein chains. (B) The distribution of the resolutions of proteins.

### Figure 3.

The magnitude of mean-square displacement  $\langle u^2 \rangle$  versus structural resolution, at 100 K (in blue) and 300 K (in red), using structures from dataset pdb12600 that were determined at these temperatures. Of all the structures in pdb12600, those determined at 100 K (10,574 structures) and at near 300 K (330 structures) are sorted by resolution and then grouped into 10 equal-size bins. Within each bin, the median  $\langle u^2 \rangle$  and the median resolution are selected and plotted. The magnitude is smaller if instead of all heavy atoms, only C $^\alpha$  atoms are used, by as much as 0.07-0.08 Å $^2$ , as shown by the thin lines.

### Figure 4.

The magnitude of B-factors as a function of resolution: (A) at 100 K, and (B) at 300 K.



### Figure 5.

The vibrational spectrum of ubiquitin. The vibrational spectra of all globular proteins follows the same curve.

### Figure 6.

(A) The characteristic frequency of globular proteins  $\omega_p$  and (B) the characteristic average mass of atoms in globular proteins  $m_p$ .

### Figure 7.

The magnitude of thermal vibrations of proteins  $\langle \Delta \mathbf{R}^2 \rangle$  at 100 K. All proteins are from the pdb1522 data, having a resolution of 1.3 Å or better.

### Figure 8.

The amount of dynamics (thermal vibrations)  $\langle u^2 \rangle_v$  and lattice disorder  $\langle u^2 \rangle_{ld}$  in B-factors as a function of structural resolution at different temperatures.

### Figure 9.

The magnitude of mean-square displacement  $\langle u^2 \rangle$  versus resolution at sub 1.1 Å resolution and at 100 K. All structures (marked by crosses) are from the pdb12600 dataset. The blue line represents the moving average of the data with a window size of 10.

### Figure 10.

Mean-square displacement  $\langle u_i^2 \rangle$  from crystallographic B-factors of structures 1IEE and 2LYM as well as thermal fluctuations  $\langle \Delta \mathbf{R}_i^2 \rangle$  computed from sbNMA. For each residue,  $\langle u_i^2 \rangle$  and

$\langle \Delta \mathbf{R}_i^2 \rangle$  represent the mass-weighted averages over all the atoms within the residue.

### Figure 11.

Difference between structures 1IEE and 2LYM in mean-square displacement  $\langle u_i^2 \rangle$  from crystallographic B-factors as well as thermal fluctuations  $\langle \Delta \mathbf{R}_i^2 \rangle$  computed from sbNMA.

### Figure 12.

The temperature and resolution dependence of B-factors and their two components  $\langle u^2 \rangle_{ld}$  and  $\langle u^2 \rangle_v$ . The upper band, stretching from blue to red, represents how the contribution of  $\langle u^2 \rangle_{ld}$  decreases at temperature increases (marked also by the white arrow), while the lower band represents  $\langle u^2 \rangle_v$ . The difference between  $\langle u^2 \rangle_{ld}$  and  $\langle u^2 \rangle_v$ , initially very large (marked out on the leftmost end), narrows both at the increase in temperature (white arrows) and at the increase in resolution (green arrow).

Table 1: The magnitudes of mean-square displacements, or  $\langle u^2 \rangle$ , as computed using all the heavy atoms (column 1), or using subsets of atoms on the backbone (denoted by subscripts  $N$ ,  $C^\alpha$ ,  $C$ , or  $O$ ) or those on side chains (denoted by subscript  $SC$ ). The numbers in the table represent the median magnitudes. All units are in  $\text{\AA}^2$ . The magnitudes are the same at 100 K and 300 K.

|       | $\langle u^2 \rangle$ | $\langle u^2 \rangle_N$ | $\langle u^2 \rangle_{C^\alpha}$ | $\langle u^2 \rangle_C$ | $\langle u^2 \rangle_O$ | $\langle u^2 \rangle_{SC}$ |
|-------|-----------------------|-------------------------|----------------------------------|-------------------------|-------------------------|----------------------------|
| 100 K | 0.79                  | 0.69                    | 0.71                             | 0.72                    | 0.74                    | 0.86                       |
| 300 K | 0.79                  | 0.69                    | 0.71                             | 0.72                    | 0.74                    | 0.86                       |

Table 2: The magnitudes of thermal vibrations at 100 K as computed using all the atoms (column 1, using Eq. (8)), or using subsets of atoms on the backbone (denoted by subscripts  $N$ ,  $C^\alpha$ ,  $C$ , or  $O$ ) or those on side chains (denoted by subscript  $SC$ ). The numbers in the first row represent the median magnitudes. All units are in  $\text{\AA}^2$ . The numbers in the second row represent the ranges between 25 percentile and 75 percentile.

| $\langle \Delta \mathbf{R}^2 \rangle$ | $\langle \Delta \mathbf{R}^2 \rangle_N$ | $\langle \Delta \mathbf{R}^2 \rangle_{C^\alpha}$ | $\langle \Delta \mathbf{R}^2 \rangle_C$ | $\langle \Delta \mathbf{R}^2 \rangle_O$ | $\langle \Delta \mathbf{R}^2 \rangle_{SC}$ |
|---------------------------------------|---|--|---|---|--|
| 0.093                                 | 0.059                                   | 0.061  | 0.057                                   | 0.081                                   | 0.118                                      |
| [0.077, 0.128]                        | [0.047, 0.084]                          | [0.049, 0.090]                                   | [0.046, 0.084]                          | [0.068, 0.119]                          | [0.097, 0.157]                             |

Table 3: A summary of the properties of two lysozyme crystal structures.

| PDB-id | Temperature [Kelvin] | Resolution [ $\text{\AA}$ ] | R-factor |
|--------|----------------------|-----------------------------|----------|
| 1IEE   | 100                  | 0.94                        | 0.12     |
| 2LYM   | 300                  | 2.0                         | 0.15     |

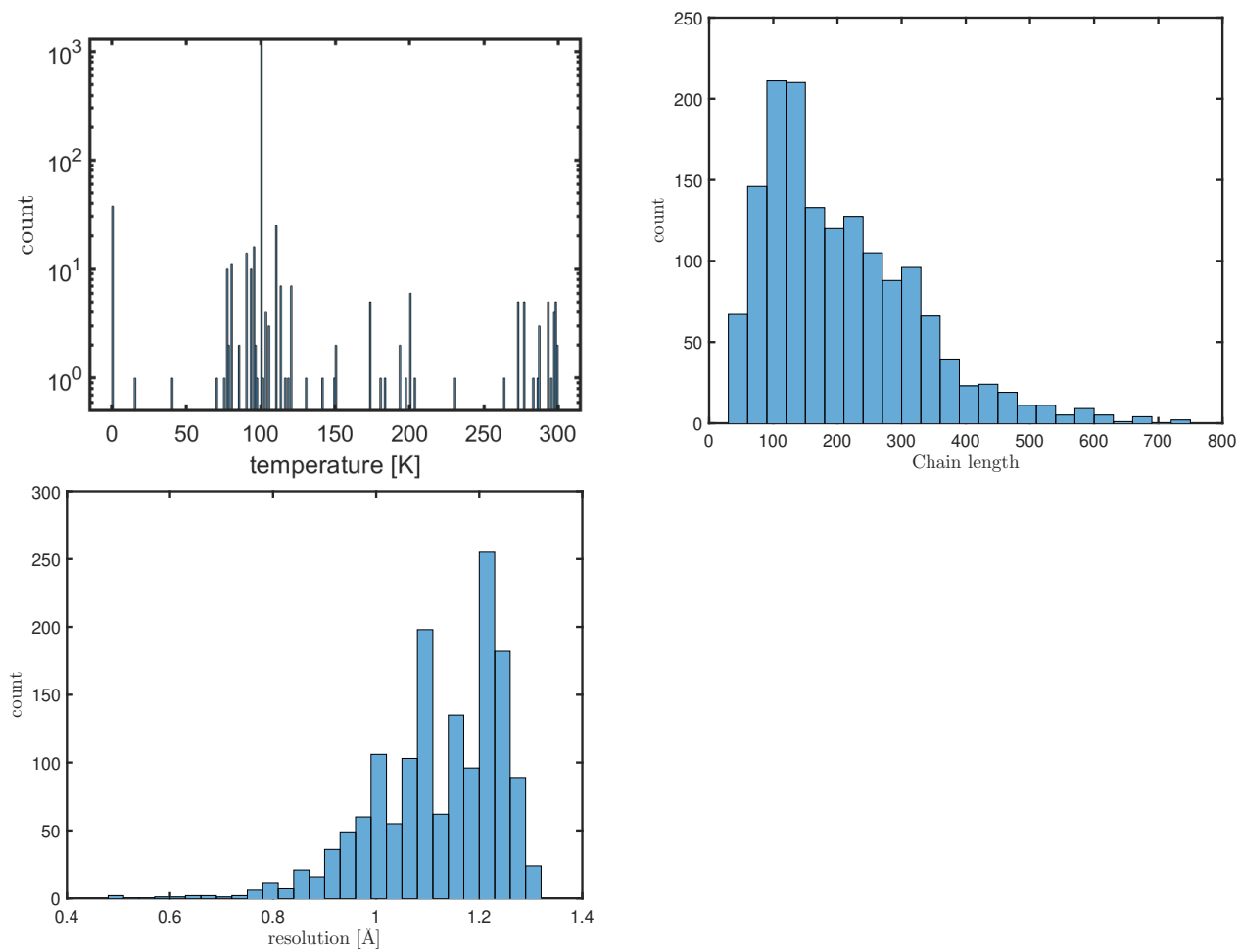


Figure 1:

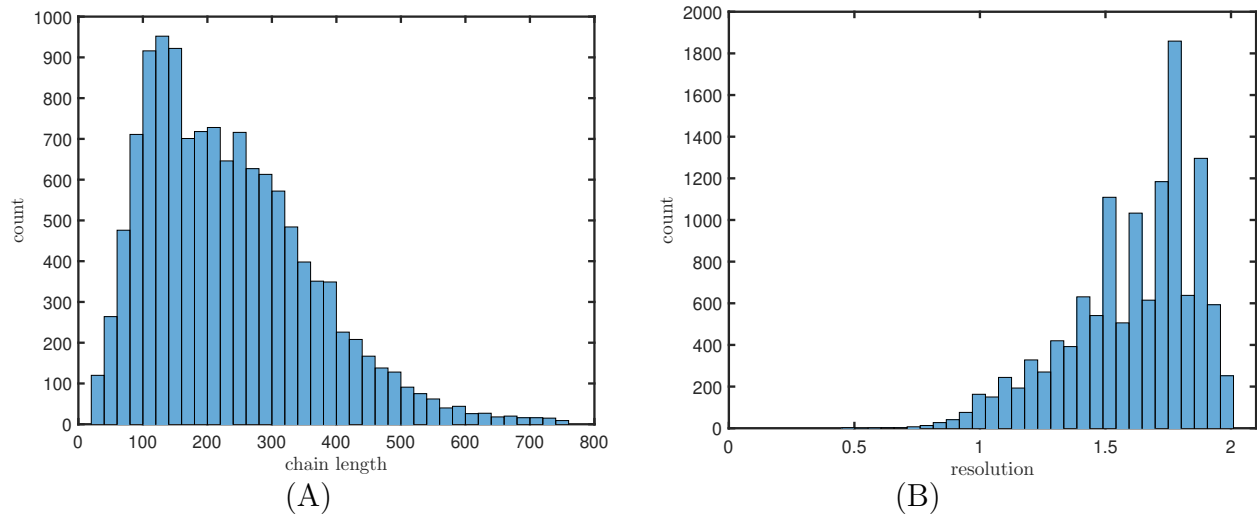


Figure 2:

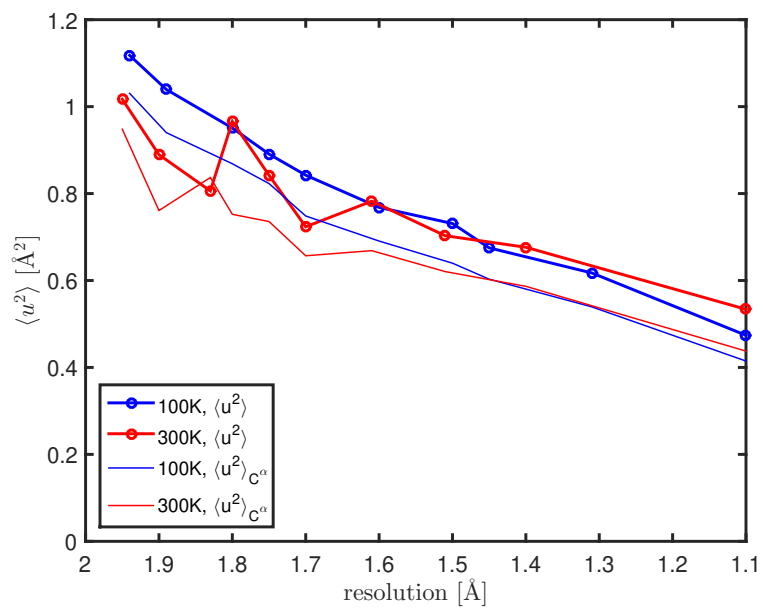


Figure 3:

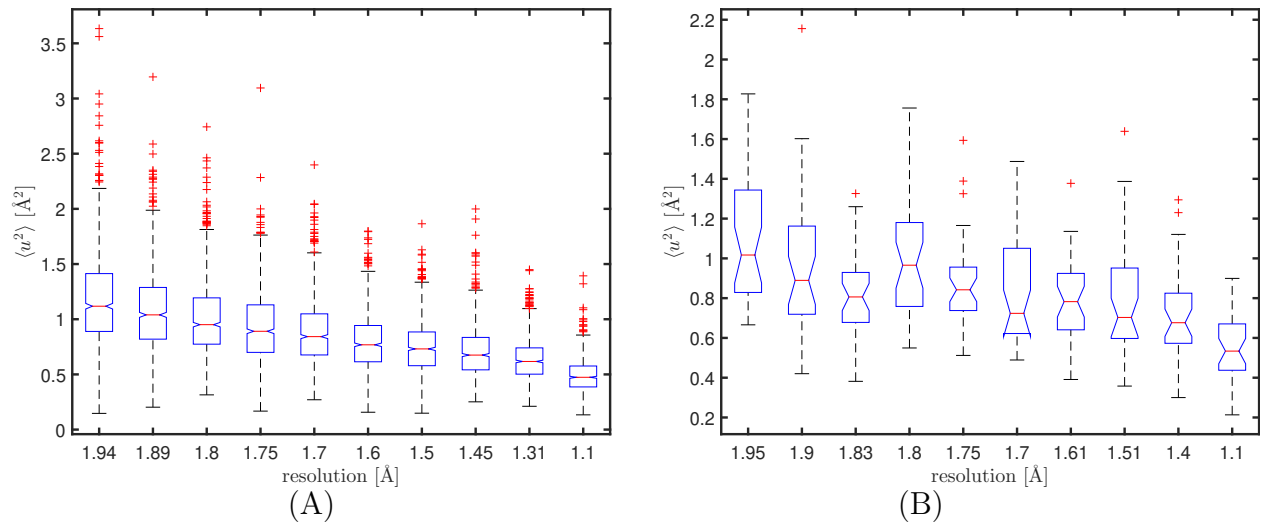


Figure 4:

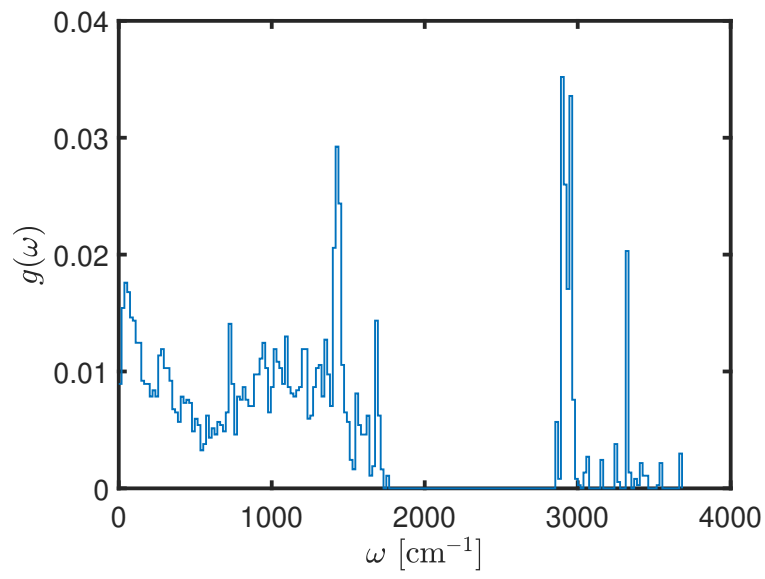


Figure 5:

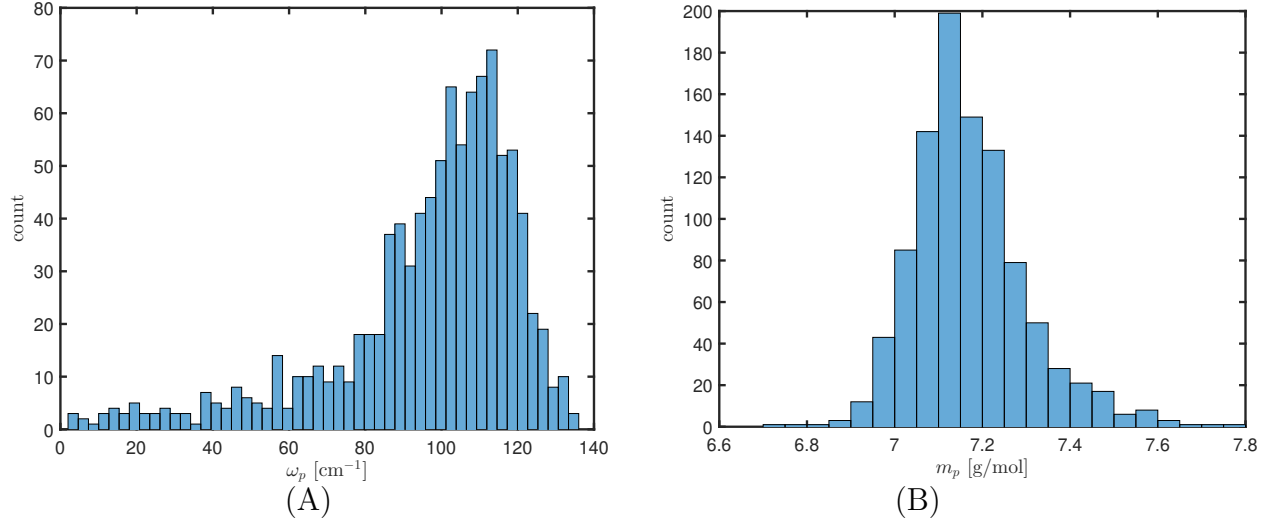


Figure 6:

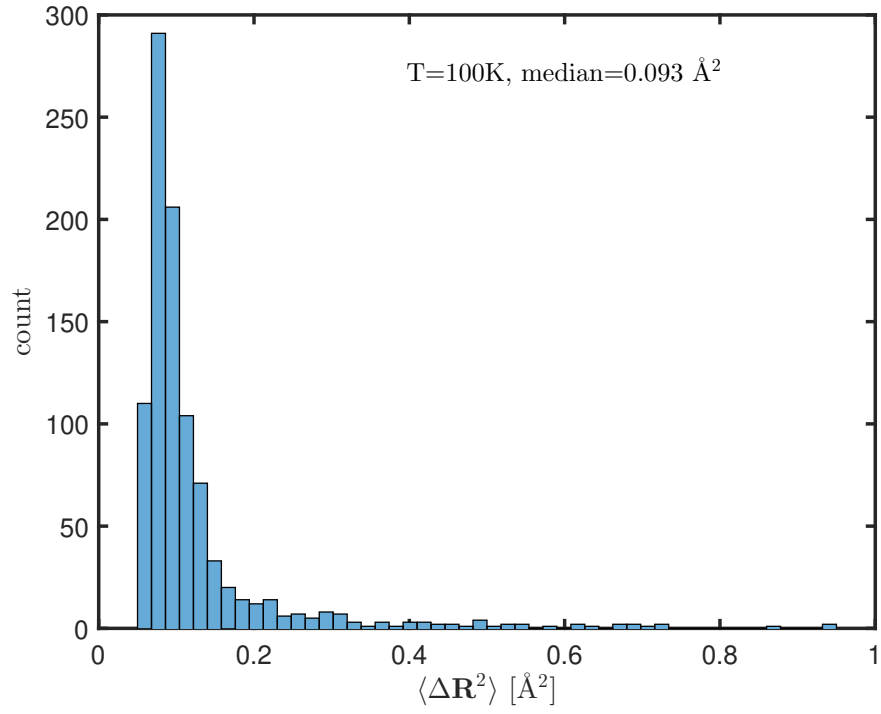


Figure 7:

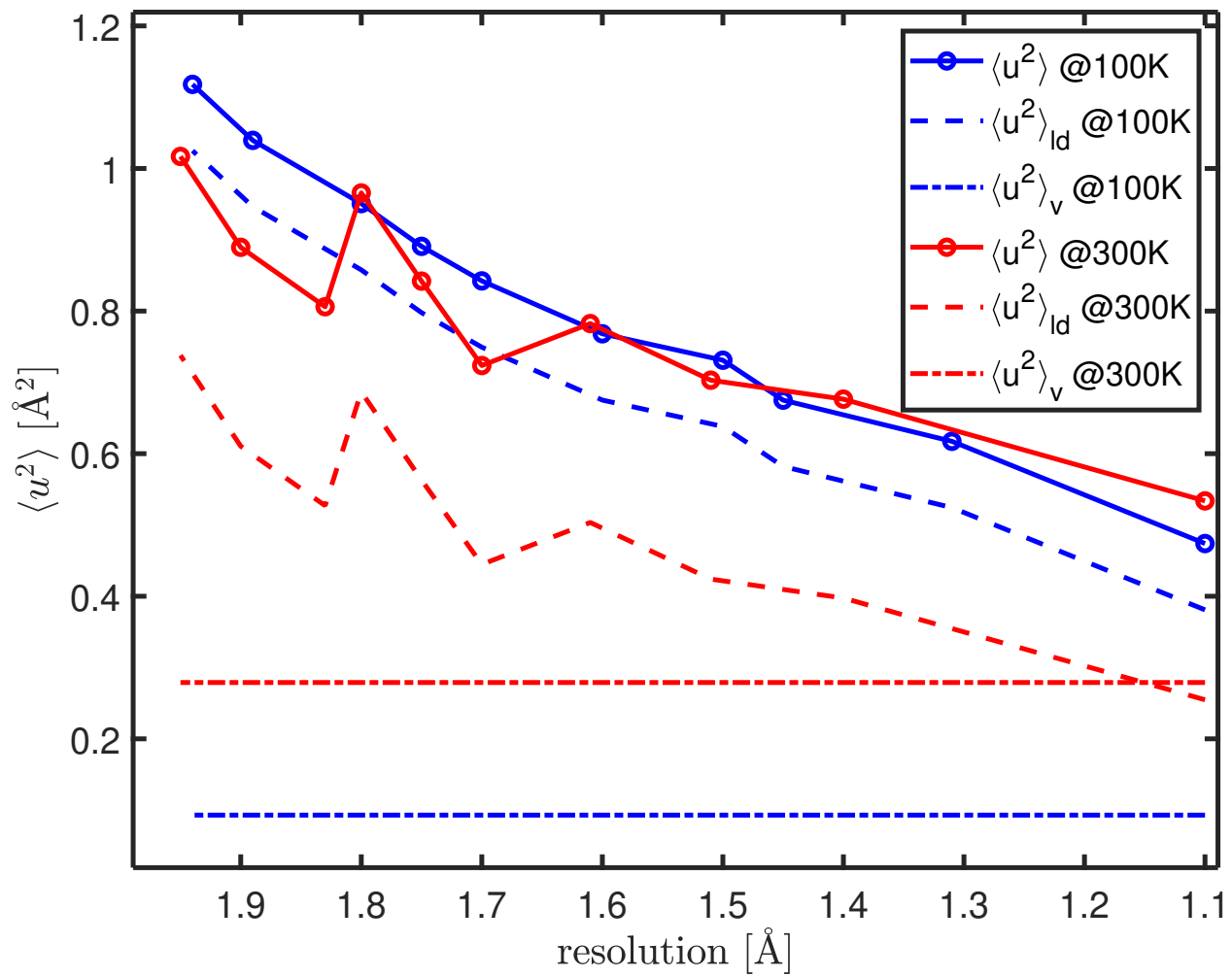


Figure 8:



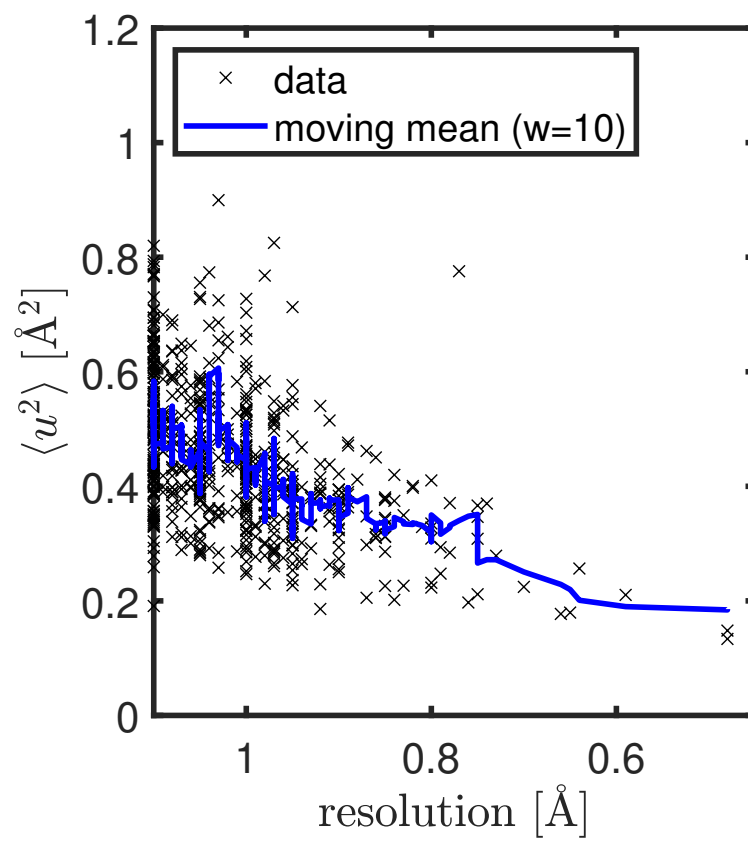


Figure 9:

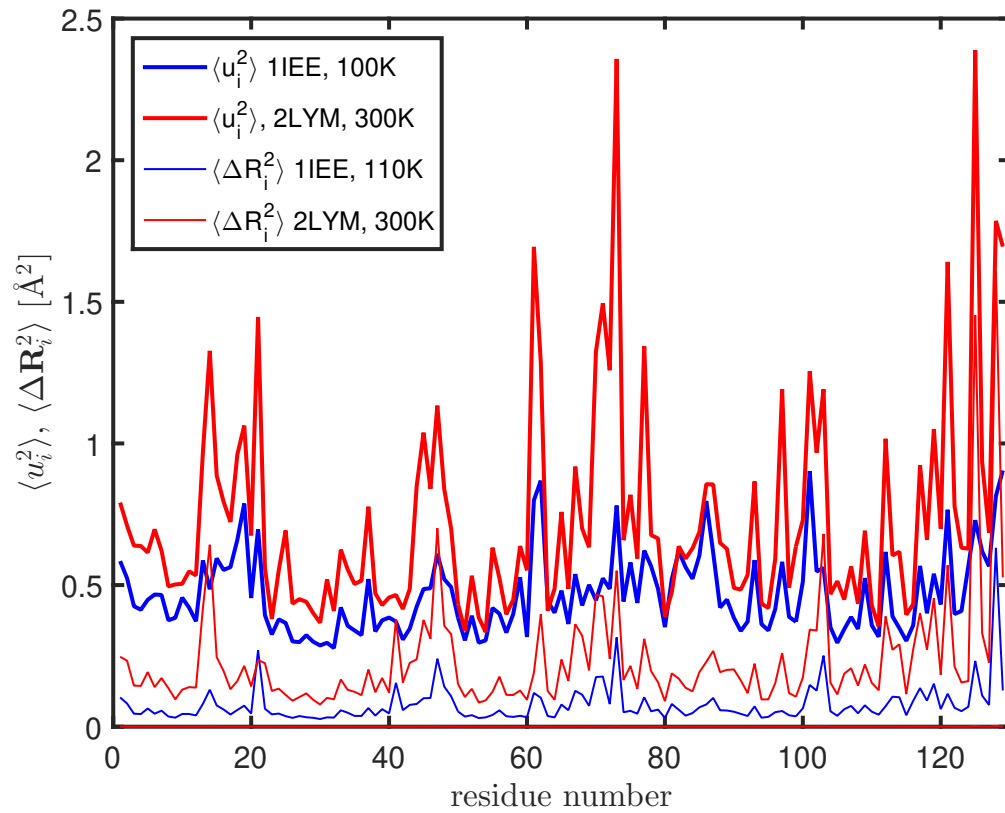


Figure 10:

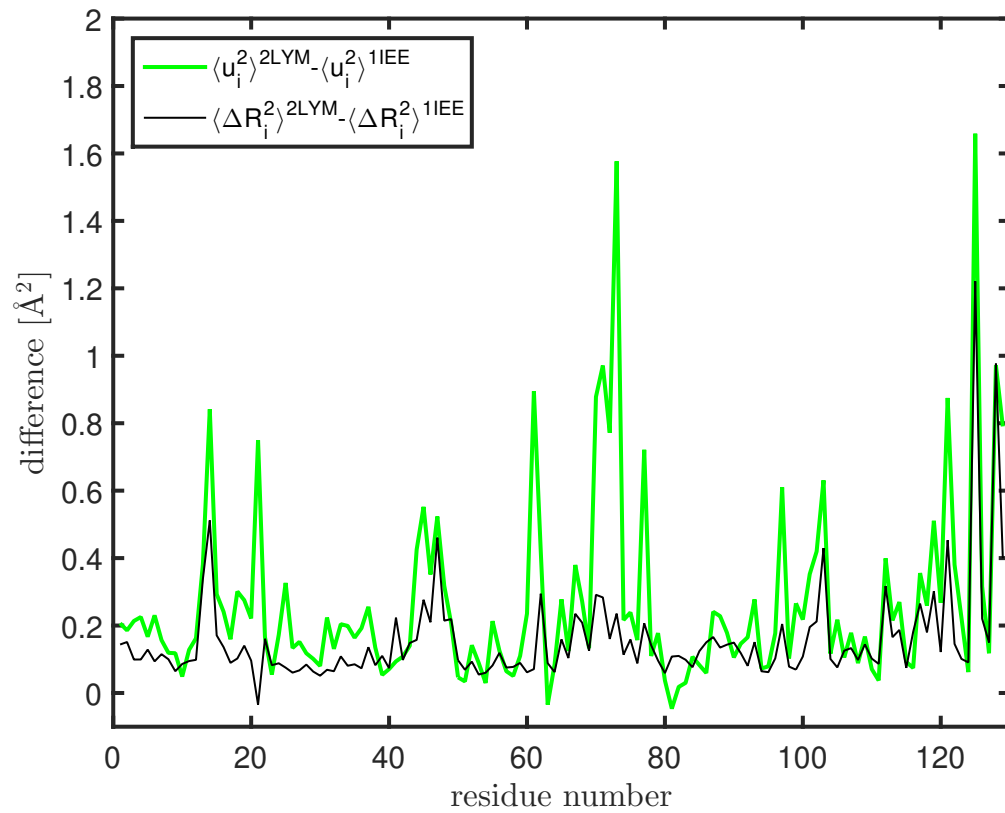


Figure 11:

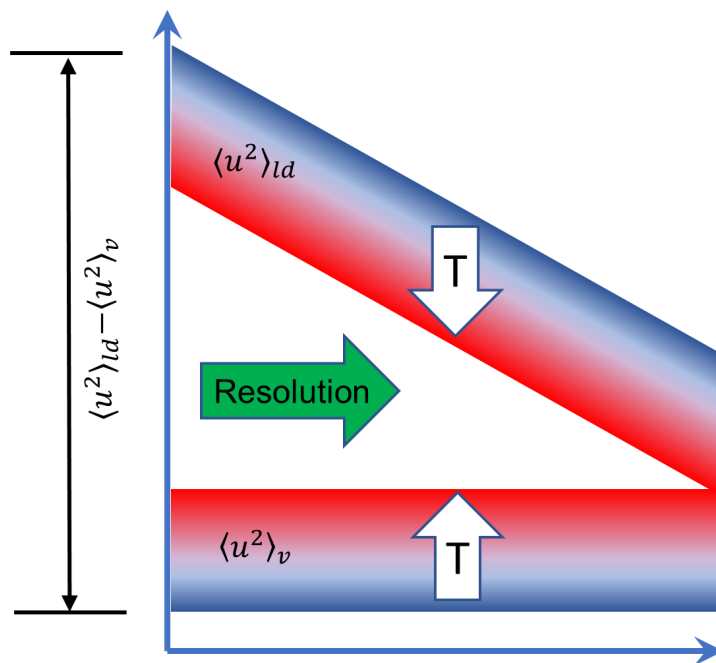


Figure 12: