Hiroshima University Doctoral Thesis

Robustness of the elastic network model against chemical or physical fitting of parameters

(パラメタの化学的・物理的最適化に対する弾性ネットワークモデルの頑健性)

2019

Department of Mathematical and Life Sciences,
Graduate School of Science,
Hiroshima University

AMYOT Romain

Table of Contents

Main Thesis

Contents

Appendices **62**

Bibliography **73**

I Introduction

A Biological context

Proteins are the major components of cells taking place in biological processes. The functions of a protein are encoded in its amino-acid sequence determining, in part, its equilibrium structure. This equilibrium structure is altered through the life of the protein as it binds and reacts with other proteins. These conformational changes make it possible a lot of various cellular processes e.g. the family of myosins which perform large motions and are involved in muscle contractions [1, 2], the family of kinesins involved in the cell division [3, 4] or HCV helicase, a virus moving along DNA or RNA, unwinding them and replacing the genetic code by its own one [5].

Other causes of conformational change are mutations. Although only occasionally occurring on a coding part of the DNA, they can completely break down the resulting protein depriving the cell from its expected correct protein type. Mutations can be deleterious e.g. cancers for humans or can be advantageous e.g. the high mutation rate of the HIV virus making it resistant to drugs [6, 7]. All these features make it important the role of the structure and its changes. Structural biology has become a major field in Biology focusing on structural changes, their origins and their effects on the function of proteins. Not only experimentally but also theoretically this topic is of interest. Many works have focused on theoretical coarse-grained models to study the effect of conformational changes or to study the conformational changes originated from perturbations [8].

B Theoretical context

1 Background

The elastic network model (ENM) is a coarse-grained method consisting in modelling interactions via harmonic potentials. The coarse-graining is either at the atomic level by considering all atoms as beads or at the residue level[1] by considering only the alpha-carbon of each residue. Then, according to some rules, beads are connected by Hookean springs with a certain spring constant. The basic method is to connect a pair of beads if their distance is smaller than a certain cut-off distance l_c. The natural length is set as the equilibrium distance and the spring constant is typically the same for all connections. The elastic energy potential between two nodes is:

$$U_{ij}(t) = \frac{1}{2}k(d_{ij}(t) - d_{ij}^0)^2$$

where k is the stiffness constant, d_{ij} is the distance between the two nodes and d_{ij}^0 their initial separation.

This model does not require any energy minimisation prior the analysis since the initial structure is taken as the minimum energy state. Although quite simple, it is able to reproduce, identify and simulate some realistic functional motions. The cyclic motion of the motor HCV helicase has been reproduced using an

[1] This coarse-graining has been chosen in the remaining of this thesis.

ENM [9]. It has also been applied to analyse the behaviours of molecular machines [10] and to test the response to external forces [11].

Currently, two broad categories of ENM are used: the Gaussian Network Model (GNM) [12, 13] and the Anisotropic Network Model (ANM) [14]. The former one assumes residues are subjected to Gaussian isotropic fluctuations while the ANM includes the anisotropy in the fluctuations.

At the equilibrium, assuming small fluctuations, the model can be linearised. Any motion can be described by a combination of independent fundamental ones. That is the Normal Mode Analysis (NMA). It consists in characterising any motion through the frequency and the direction along each degree of freedom (called mode). This is done by solving the eigenvalue problem involving the (linearised) Hessian matrix of potentials. The eigenvalue is related to the frequency while the eigenvector represents the displacement vector (see following section).

Coupled together, ENM and NMA have given quite accurate results in terms of reproducing fluctuations [12, 14]. It has also been shown from systematic studies [15] or from study analysing motor-protein motions [16], that the transition motions between two conformations is often mostly explained by a single slow mode. To be able to explain a transition motion mostly by a single fundamental one is very interesting because we can theoretically visualise, analyse and determine this fundamental motion. In addition, the NMA being relied on the diagonalisation of a matrix, it is faster than classical molecular dynamics simulations. Then, it is a tool of choice when looking at such concerns.

The combination ENM-NMA has then drawn a lot of attention and a lot of studies have introduced the heterogeneity in the network by making the spring constants dependent on the distance and/or on the sequence specificity in order to best fit the biological protein [17, 18, 19, 20, 21, 22]. Indeed, we can fairly think that assuming homogeneity is a strong hypothesis since we may expect a stronger interaction between two residues as their distance is reduced as well as we may expect different affinities between different types of residues (based on the amino acid structure, the polarity, ...). In addition, the cut-off distance is problematic. It has to be chosen according to the protein, not too small to have a network connected enough but not too large otherwise the motions of interest are totally inhibited. A certain value would be suitable for a certain protein but not for another one. The focus on this parameter is the interest of many studies [23, 21, 22, 24]. The aim is to remove it using heterogeneous spring constants. Currently, the community expects to find a model which can be applied to all proteins in a systematic way and with a good accuracy. Despite many efforts, the basic GNM and ANM are still be widely used due to their ease of use and the poor loss of accuracy compared to more elaborated and complicated network.

For reviews about the elastic network model and the normal mode analysis, see [25, 26, 27, 28, 29, 30, 31].

2 Aim of this work

The study of the effects of mutations was our primary goal. Among all alternatives to the classical ENMs, one has attracted our attention. Several years ago, using solution NMR dataset of 1500 proteins, distance and sequence specific models have been determined [22]. Such models are interesting because each pair of amino acids has its own spring constant which is also distance-dependent. We may hope that such spring constants, extracted from a statistical analysis on realistic protein structures in solution, contain a lot of additional valuable information. In particular, it would enable the study on the effect of mutations at a small cost. However, our study on mutations has not given any satisfying results and it appeared that the network is less sensitive to the introduction of this heterogeneity than we expected it.

It turned out that the robustness of networks to the heterogeneity of spring constants becomes our new target. We aim to determine to which extent the tuning of spring constants can improve the modelling and what is the value of the improvement. Most of studies introducing new heterogeneous networks show improvements at the protein level by looking at the overall correlation between experimental and theoretical fluctuations of residues (see below in materials and methods section). However, they do not discuss how valuable are these improvements. The overall correlation can be biased by a small set of residues which greatly influences it. It is especially true for residues which are connected with only a few other residues. Improvement of the estimation of fluctuations of such residue by stiffening their links could result in a damping in the estimation of fluctuations of the other residues. The improvement being more important for these poorly-connected residues than the decline for the other ones, the overall correlation is still improved overlooking this problem. Secondly, how useful are the new information brought by the heterogeneity is generally not discussed. In heterogeneous models, the spring constant values follow a coherent distribution (e.g. values decrease as the distance increases) or determined using experimental data but nothing shows that the improvements come from the cohesiveness of the distribution.

Our aim is then to study sequence-specific models and some other heterogeneous models in a qualitative way to understand whether the improvements are biased or not. In particular, we are interested in understanding to what extent networks are sensitive or not to the tuning of their spring constants and in determining if there really is a meaning to try to improve modelling in this way.

II Materials and methods

A Elastic network model

1 Introduction and limitations

The elastic network model is a model of proteins where interactions between residues or atoms, represented by beads, are approximated by harmonic potentials. From a protein structure which has been determined experimentally, the network is built by coarse-graining the structure taking only the alpha-carbons and by setting a Hookean spring to each pair of residues for which the distance is smaller than a certain cut-off distance l_c. The equilibrium distance is taken as the natural length and the spring constant is the same for all links in the classical ENMs. The equilibrium positions correspond to the conformation with minimal energy by construction of the network avoiding the need of energy minimisation. The total potential energy for the residue i is written as follow:

$$U_i(t) = \frac{1}{2} k \sum_{j=1}^{N} \mathbb{1}_{d_{ij}^0 < l_c} (d_{ij}(t) - d_{ij}^0)^2$$

where k is the spring constant, d_{ij} the distance between residues i and j at time t and d_{ij}^0 the distance at the equilibrium.
The resulting force ($R_i = (x_i, y_i, z_i)$)

$$F_i = -k \sum_{j=1}^{N} \mathbb{1}_{d_{ij}^0 < l_c} (d_{ij}(t) - d_{ij}^0) \frac{R_i - R_j}{d_{ij}(t)}$$

is absolutely not linear in respect to the displacement of the pair. However, assuming small fluctuations ($R_i \sim R_i^0, \forall i \in [\![1, N]\!]$), we can get the linearised equation[2]:

$$\begin{aligned} F_i = -k \sum_{j=1}^{N} \mathbb{1}_{d_{ij}^0 < l_c} \frac{R_i^0 - R_j^0}{(d_{ij}^0)^2} \big[(x_i^0 - x_j^0)(\Delta x_i - \Delta x_j) \\ + (y_i^0 - y_j^0)(\Delta y_i - \Delta y_j) + (z_i^0 - z_j^0)(\Delta z_i - \Delta z_j) \big] \quad (1) \end{aligned}$$

which can be packed into a matrix form in a very convenient way. It should be highlighted that results deduced from an analysis based on this equation, which is the case for the normal mode analysis, is valid only within the hypothesis of small fluctuations around the equilibrium position.

2 Improving the ENM

The ENM considers all interactions as being the same: harmonic potential with the same spring constant. That is a strong hypothesis because in actual proteins, we do not expect that two residues interact the same way disregarding their type and the distance separating them. For example, alanine and glycine are two hydrophobic amino acids. In water, when the protein folds, these amino acids are pushed in the interior of the protein maintaining the two amino acids at close proximity. It is called the hydrophobic force which is a weak force. The electrostatic interactions between two amino acids having opposite charges are attractive while they are repulsive for two amino acids having the same charges. Two consecutive amino acids belonging to the backbone are covalently bonded i.e. the interactions are much more stiff. The strength of a covalent bond in water is much larger than for electrostatic attraction between two charged atoms [32].

An additional problem of the ENM is the cut-off distance because it is arbitrary. Its choice depends on the protein: if too small, the protein suffers from too much free rotations; if too large, the fluctuations can be totally damped because the network is too strongly connected. A certain cut-off distance can be suitable for some proteins but not suitable for some others.

Many studies have tried to fit the spring constants to better modelled proteins as well as to remove the cut-off distance to make a unique model applicable

[2] more details of calculation are available in the relaxation trajectory theory section

to all proteins. The immediate way is to tune the spring constants using a function which decreases when the distance increases [17, 18, 19, 20, 21]. It is based on the intuitive fact that long-range interactions are softer than short-range ones. The cut-off distance being removed, all residues are connected to all others. These methods do not include a special processing for the backbone which is covalently maintained although some of them use different distributions for short-range interactions and for long-range interactions. Moreover, the choice of the distribution does not rely to physical nor biological properties. The inversely squared decreasing function, the exponential decreasing function, etc have been proposed but without justifying them in a biological or physical point of view except the one discussed above.

Some other studies have tried to modify the spring constants according to the types of amino acids, either by inferring them from a large set of experimental data [22] or by directly including all the different bond forces (hydrogen bonds, Van der walls force, etc) [23, 33]. The potential remains harmonic but divided into several categories representing the types of force. However, these models are somehow still cut-off dependent and do not always take into account the distance. In the same aim, inclusion of Lennard-Jones potential for native contacts has been proposed [34, 35].

In a different spirit, the non-linearity of the ENM has been considered to extend the validity of the ENM to large conformational motions [36, 37].

B Normal mode analysis

1 An eigenvalue problem

The normal mode analysis (NMA) holds on the linearisation of the system near the equilibrium.

Let consider a protein composed of N residues. We are working in a $3N$-dimensional space for which the basis $(\chi_i)_{i=1,...,3N}$ is the 3-dimensional coordinates of each residue i.e. $(\chi_1, ..., \chi_{3N}) = (x_1, y_1, z_1, ..., x_N, y_N, z_N)$. We note V the total energy of the system which is a sum of elastic potentials and we note F_i the force along the i^{th} dimension. The latter is obtained by deriving the potential according to the coordinate χ_i:

$$F_i = -\frac{\partial(V(\chi_1, ..., \chi_{3N}))}{\partial \chi_i}$$

Assuming that fluctuations are small around the equilibrium, we can linearise near the equilibrium position $(\chi_1^0, ..., \chi_{3N}^0)$ using Taylor's theorem:

$$F_i = -\left(\frac{\partial^2 V(\chi_1^0, ..., \chi_{3N}^0)}{\partial \chi_1 \partial \chi_i} d\chi_1 + ... + \frac{\partial^2 V(\chi_1^0, ..., \chi_{3N}^0)}{\partial \chi_{3N} \partial \chi_i} d\chi_{3N}\right)$$

Doing the same for all dimensions, we obtain in a matrix form:

$$F = -H\Delta\chi \qquad (2)$$

where H is the matrix of second derivatives of the potential called the Hessian, $F = (F_1, ..., F_{3N})$ and $\Delta\chi = (d\chi_1, ..., d\chi_{3N})$.

The matrix H is symmetric with real positive coefficients and then can be diagonalised

$$F = -H\Delta\chi = -\lambda\Delta\chi$$

with λ the eigenvalue and $\Delta\chi$ the eigenvector. There are $3N$ positive eigenvalues called modes and linked to frequencies of fundamental motions (eigenvectors). Six of them are zero-eigenvalues corresponding to the free rotations and the free translations.

By the second law of Newton:

$$F = m\Delta\ddot{\chi}$$

then,

$$m\Delta\ddot{\chi} = -\lambda\Delta\chi$$

which has for solution

$$\Delta\chi(t) = k\cos\left(\sqrt{\frac{\lambda}{m}}t + \phi\right)V$$

where k and ϕ are constants and V is the normalised eigenvector.

Let $R_i = (x_i, y_i, z_i)$ be the position of the residue i. Its fluctuations around its equilibrium position R_i^0 is a linear combination of eigenvectors $(\Delta\chi_j)_{j=1,\dots,3N}$:

$$\Delta R_i(t) := R_i(t) - R_i^0 = \sum_{j=1}^{3N} a_j \Delta\chi_j(i,t)$$

where the a_j are some constants and $\Delta_j(i,.)$ is the i^{th} coordinate of the j^{th} eigenvector.
Finally,

$$\Delta R_i(t) = \sum_{j=1}^{3N} k_j \cos\left(\sqrt{\frac{\lambda_j}{m}}t + \phi_j\right)V_j(i)$$

where the constants a_j have been included in the k_j.

The normal mode analysis consists in resolving the eigenvalue problem. That would give us the direction and the frequency of each mode characterising any motion carried out by the residues of the given protein.

Remark: The diagonalisation of the Hessian matrix can be done easily with computers in a reasonable time (faster than molecular dynamics simulations) which is one of the reason of the popularity of the normal mode analysis.

We assume the equipartition of energy along each normal mode. Then, for each mode j having $-\lambda_j$ for eigenvalue and $\Delta\chi_j$ for eigenvector, the mechanical energy $E_{m,j}$ is:

$$E_{m,j} = k_B T \qquad (3)$$

The mechanical energy is the sum of the potential energy U_j and the kinetic energy $E_{c,j}$.
The kinetic energy is calculated as:

$$E_{c,j} = \frac{1}{2}m\|\Delta\dot{\chi}_j\|^2 = \frac{1}{2}k_j^2\lambda_j \sin^2\left(\sqrt{\frac{\lambda_j}{m}}t + \phi_j\right)\|V_j\|^2$$

As for the potential energy, we have to integrate the force F_j according to χ_j:

$$U_j = -\int F_j d\chi_j = \lambda_j \int \Delta\chi_j d\chi_j = \lambda_j \frac{(\Delta\chi_j)^2}{2}$$

and then

$$U_j = \frac{1}{2}k_j^2\lambda_j \cos^2\left(\sqrt{\frac{\lambda_j}{m}}t + \phi_j\right)\|V_j\|^2$$

The equation (3) becomes

$$\frac{1}{2}k_j^2\lambda_j = k_B T$$

and we get the pre-factor k_j for each mode j:

$$k_j^2 = 2\frac{k_B T}{\lambda_j}$$

2 The B-factor and the Pearson Correlation Coefficient

The B-factor is a measure of fluctuations of an atom around its equilibrium position. Theoretically, it is estimated as follows:

$$B_i = \frac{8\pi^2}{3}\left\langle(\Delta R_i)^2\right\rangle$$

$$\left\langle(\Delta R_i)^2\right\rangle = \sum_{j=1}^{3N-6}\frac{k_B T}{\lambda_j}\|V_j(i)\|^2$$

where the sum over j represents the sum over the modes, $V_j(i)$ is the vector of the three coordinates corresponding to the residue i of the j^{th} eigenvector. The eigenvectors are orthogonal.

The dimension of eigenvalues λ_j is $[M].[T]^{-2}$, the eigenvectors are dimensionless and the term $k_B T$ being an energy has the dimension of $[M].[L]^2.[T]^{-2}$. Then the unit of B-factors is Å^2. The B-factor resumes to:

$$B_i = \frac{16\pi^2 k_B T}{3}\sum_{j=1}^{3N-6}\frac{\|V_j(i)\|^2}{\lambda_j}$$

The Pearson correlation coefficient (PCC) is a well-known coefficient to compute the correlation between experiment and theory. It is defined as follows:

$$r_b := \frac{< B^{exp} - \bar{B}, B^{the} - \bar{B} >}{\|B^{exp} - \bar{B}\|\|B^{the} - \bar{B}\|} = \frac{\sum_{i=1}^{N}(B_i^{exp} - \bar{B})(B_i^{the} - \bar{B})}{\sqrt{\sum_{i=1}^{N}(B_i^{exp} - \bar{B})^2\sum_{i=1}^{N}(B_i^{the} - \bar{B})^2}}$$

where $B^{exp} = (B_i^{exp})_{i=1,\dots,N}$ are experimental B-factors, $B^{the} = (B_i^{the})_{i=1,\dots,N}$ are theoretical B-factors. The experimental B-factors have been rescaled such as the average $\bar{B}$ is the same for experimental B-factors and theoretical B-factors:

$$\bar{B} = \frac{1}{N}\sum_{i=1}^{N} B_i^{exp} = \frac{1}{N}\sum_{i=1}^{N} B_i^{the}$$

The PCC tells us how experimental B-factors and theoretical B-factors are correlated and thus a measure of accuracy of the theoretical model. The closer the PCC to 1, the more reliable the model (at least for the given experimental data). Figure 1 shows an example of B-factor patterns for adenylate kinase.

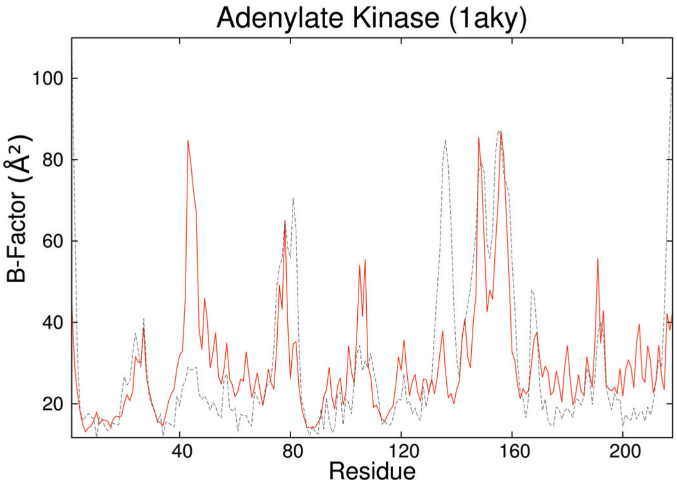


Figure 1: B-factor patterns of adenylate kinase (pdb id: 1aky). The red curve represents the theoretical B-factors obtained with an ANM16 and the black curve represents the experimental B-factors. The PCC is 0.61.

3 The overlap

The overlap between an experimental transition motion from an initial structure to a final structure and the theoretical estimated transition gives the contribution of each mode.
It is calculated as the scalar product between the eigenvector and the structural difference in the two structures :

$$O_i := \frac{| < v_i, \delta_i > |}{\|v_i\|\|\delta_i\|} \in [0,1]$$

where v_i is the eigenvector associated to the mode i and δ_i is the difference between χ_i in the initial structure and χ_i in the final structure.
Mathematically, we make the orthogonal projection along each mode to check its contribution to the motion. The closer to 1 the overlap, the more the mode i contribute to the transition. Figure 2 shows an example of overlap curve for

the transition of HIV-1 protease from its free structure to a structure adopted after a complex with the inhibitor AHA001.

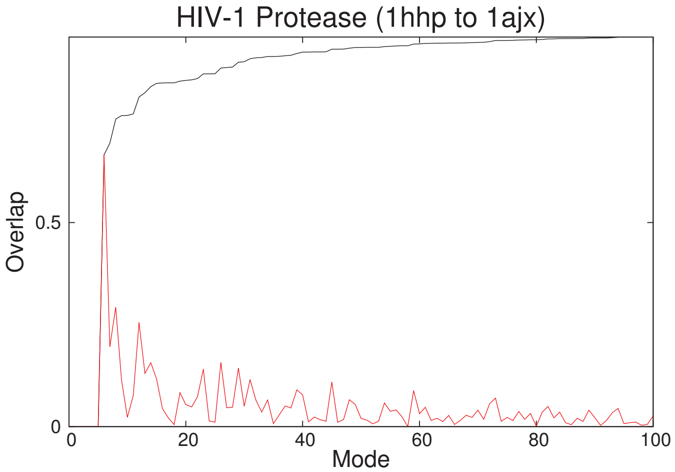


Figure 2: Overlaps of the transition of HIV-1 protease from its free structure (pdb id: 1hhp) to its inhibitor-complexed structure (pdb id: 1ajx). The red curve is the overlap and the black curve is the (squared-)cumulative overlap. The NMA has been carried out for the free structure. Only the first 100 modes over the 297 ones are displayed.

4 The inter-residue variance

This measure introduced recently [22] estimates the fluctuations of the inter-residue distances.
It can be estimated experimentally using the different structures available in NMR pdb files:

$$V_{i,j}^{exp} = \frac{1}{M} \sum_{k=1}^{M} (r_{ij}^{k} - \bar{r}_{i,j})$$

where M is the number of structures in the NMR pdb file, $r_{i,j}^{k}$ is the distance between residue i and residue j in the structure k and $\bar{r}_{i,j}$ is the average distance between i and j over all structures.

As for the theoretical one, it is calculated using eigenvalues and eigenvectors:

$$V_{i,j}^{the} = \sum_{l=1}^{6} \left[\sum_{k=1}^{6} \frac{\partial r_{i,j}}{\partial \chi_k} \langle \chi_k, \chi_l \rangle \right] \frac{\partial r_{i,j}}{\partial \chi_l}$$

where $(\chi_1, \chi_2, \chi_3, \chi_4, \chi_5, \chi_6) = (x_i, y_i, z_i, x_j, y_j, z_j)$ and

$$\langle \chi_k, \chi_l \rangle = 4K_B T \sum_{h=1}^{3n-6} \frac{1}{\lambda_h} V_h(\chi_k) V_h(\chi_l)$$

In the following, we will discuss the rescaled error:

$$E_{i,j} = \frac{V_{i,j}^{the} - V_{i,j}^{exp}}{V_{i,j}^{exp}}$$

C Relaxation trajectory theory

1 Under the assumption of linearity

Theoretically, under the assumption of small fluctuations, the relaxation trajectory can be computed using normal modes. We need the non harmonic form of the equation on the motion of residues in the over-damped limit.
The total energy of the system is:

$$U = \frac{k}{2} \sum_{i=1}^{N} \sum_{j>i}^{N} a_{ij}(d_{ij} - d_{ij}^0)^2$$

with $a_{ij} = 1$ if there is a link between i and j and 0 otherwise.

In the over-damped limit, the temporal displacement along x_i is defined as

$$\frac{dx_i}{dt} := -\Gamma \frac{\partial U}{\partial x_i}$$

where Γ is the mobility having dimension $[T].[M]^{-1}$.

Then,

$$\frac{dx_i}{dt} = -k\Gamma \sum_{j=1}^{N} a_{ij}(d_{ij} - d_{ij}^0) \frac{x_i - x_j}{d_{ij}}$$

If we assume the fluctuations around the equilibrium position are small i.e. $\Delta x_i = x_i - x_i^0 << 1$ and so on, then we can linearise this equation:

$$\frac{x_i - x_j}{d_{ij}} \sim \frac{x_i^0 - x_j^0}{d_{ij}^0}$$

$$\begin{split} d_{ij} - d_{ij}^0 \sim d_{ij}^0 + \frac{x_i^0 - x_j^0}{d_{ij}^0}\Delta x_i - \frac{x_i^0 - x_j^0}{d_{ij}^0}\Delta x_j + \frac{y_i^0 - y_j^0}{d_{ij}^0}\Delta y_i - \frac{y_i^0 - y_j^0}{d_{ij}^0}\Delta y_j \\ + \frac{z_i^0 - z_j^0}{d_{ij}^0}\Delta z_i - \frac{z_i^0 - z_j^0}{d_{ij}^0}\Delta z_j - d_{ij}^0 \end{split} \tag{4}$$

$$d_{ij} - d_{ij}^0 \sim \frac{(x_i^0 - x_j^0)(\Delta x_i - \Delta x_j) + (y_i^0 - y_j^0)(\Delta y_i - \Delta y_j) + (z_i^0 - z_j^0)(\Delta z_i - \Delta z_j)}{d_{ij}^0}$$

following the Taylor's theorem.

We obtain the linearised equation:

$$\frac{d\Delta x_i}{dt} = -k\Gamma \sum_{j=1}^{N} a_{ij} \left(\frac{(x_i^0 - x_j^0)^2}{d_{ij}^0}(\Delta x_i - \Delta x_j) + \frac{(x_i^0 - x_j^0)(y_i^0 - y_j^0)}{d_{ij}^0}(\Delta y_i - \Delta y_j) \right.$$
$$\left. + \frac{(x_i^0 - x_j^0)(z_i^0 - z_j^0)}{d_{ij}^0}(\Delta z_i - \Delta z_j) \right) \quad (5)$$

The three terms in parenthesis correspond to three terms in the Hessian matrix H in (2).
Doing the same calculations for the y-axis and the z-axis, we obtain for $\Delta R_i = (\Delta x_i, \Delta y_i, \Delta z_i)$:

$$\frac{d\Delta R_i}{dt} = -\Gamma \sum_{j=1}^{N} H_{ij} \Delta R_j$$

In the limits of the linearisation, the direction along $\Delta R_i - \Delta R_j$ is approximated by the direction along ΔR_j.

Considering $\Delta R = (\Delta R_1, ..., \Delta R_N)$, the matrix form of the set of equations is:

$$\frac{d\Delta R}{dt} = -\Gamma H \Delta R$$

for which the solution is a sum of exponential decaying according to the eigenvalues:

$$\Delta R(t) = \sum_{\alpha=1}^{3N-6} k_\alpha e^{-\Gamma \lambda_\alpha t} V_\alpha \quad (6)$$

where k_α are the constants, λ_α are the $3N-6$ eigenvalues, the 6 others being zero-eigenvalues and V_α are the corresponding eigenvectors.

In the linearisation limit, any motion is a linear combination of exponentials decreasing according to each eigenvalue along the fundamental motion represented by the eigenvector. At the end of the motion, only the lowest modes remain. This is particularly the moment where we can expect that the linearisation holds and then where the relaxation may match the equation (6).

2 Without the assumption of linearity

As for the relaxation without the hypothesis of linearity, we need to use dynamical simulations. It consists in implementing the non-linearised forces acting on each residue. The Euler method has been adopted to numerically solved the equations. At the beginning, in order to deform the current conformation, uniform random forces $f = (f_1, ..., f_N)$ are applied on each residue such as $\|f\|_2 = F_{ini}$ where F_{ini} (in force unit) is the total magnitude. The forces are applied until a certain time T_{ini} (in time unit) after which they are cut and the system goes back to the equilibrium. The relaxation is observed through the distances between three labelled residues. The three labels are chosen according to an automatic way [38]. The first two labels correspond to the pair for which the distance change after applying the slowest mode is the maximal then the last label is chosen as the residue for which the distance change between it and the label 1 is maximal after applying the second slowest mode.

3 Validity of the linearity

The mechanical coordinate Φ is a measure of the deviation of the current structure to the equilibrium one within the linear approximation [39]. It exponentially decreases along each eigenvalue with time. It is defined as the following equation:

$$\frac{d\Phi}{dt} = -\sqrt{-\Gamma\frac{dU(t)}{dt}} \tag{7}$$

where Γ is the mobility.
From equation (7):

$$\Phi(t) = \int_t^{+\infty} \sqrt{-\Gamma\frac{dU(t)}{dt}}\,\mathrm{d}t$$

This equation is solved numerically using the trapezoidal rule.

At the final stage, only the slowest mode remains and Φ has the form:

$$\Phi(t) = A\exp(-\lambda_1\Gamma t)$$

where A is a constant and λ_1 the eigenvalue of the slowest mode.

Then,

$$\frac{d\Phi(t)}{dt} = -\lambda_1\Gamma\Phi(t)$$

From (7), we get

$$\frac{dU(\Phi)}{d\Phi} = -\frac{1}{\Gamma}\frac{d\Phi(t)}{dt}$$

therefore

$$\frac{dU(\Phi)}{d\Phi} = -\lambda_1\Phi$$

Integrating from t to $+\infty$,

$$U(\Phi) = \frac{1}{2}\lambda_1\Phi^2$$

Then, when only the slowest mode remains, the elastic potential U is quadratically dependent on Φ.

This criteria is interesting in the sense that it allows to compare the profile of the elastic energy for a relaxation trajectory (where linearisation is not assumed) with the profile of the elastic energy given by the normal mode analysis (where linearisation is assumed). Particularly, it determines in which extent the normal mode analysis and its linear approximation holds for a given network.

D Large data analysis on crystallographic structures

In the following sections, a study on a large dataset of proteins are presented. The set of proteins has been determined in the following way: from the protein data bank, we have selected crystallographic protein structures for which the resolution is under 2.0Å. From this set, proteins having either a missing atom or a missing residue or an atom with zero occupancy are discarded. Equally, are discarded proteins having non-conventional residue-types (such as “MSE” or “UKN”). Ligands, DNA and RNA are not taken into account in the analysis. Because the experimental B-factors are determined for the crystal asymmetric unit, all chains contained in the pdb file are selected for the analysis. Proteins for which at least one residue has less than 3 connections and proteins showing more than 6 zero-eigenvalues are removed.

In the results, the B-factor value is discussed according to the relative degree. The degree of a residue is the number of connections it has with the other residues belonging to the same protein. The relative degree is the degree rescaled by the maximal degree within the protein the residue belongs to.

Also are discussed the polarity and the secondary structure. The secondary structure in which residues belong to is stated in the pdb file. As for the polarity, it is set according to the residue type regarding the following classification:
- Hydrophobic: ALA, ILE, LEU, MET, PHE, VAL, PRO and GLY
- Polar: GLN, ASN, HIS, SER, THR, TYR, CYS and TRP
- Charged: ASP, GLU, ARG and LYS

E Analysis of solution NMR structures

As compared to crystallographic structures, NMR structures are not well suited for the elastic network model. A lot of them contain intrinsic disordered regions or proteins for which tails fluctuate so much that they break the normal mode analysis down or give non-physical results and bias statistical analysis. The analysis has to get rid of such structures. Furthermore, there is no B-factor in NMR structures and there is no convention to fill this entry. That means all pdb files do not provide the same experimental measure of fluctuations.
Knowing that, the set of NMR structures has been determined using the following way: all solution NMR protein pdb files have been taken. We first discarded the proteins with the same criteria (except resolution) as for the crystallographic dataset, the normal mode analysis being carried out for the first model of the pdb file using ANM10. In addition, we also discarded pdb files having less than 50 models or less than 20 residues. At this stage, the experimental fluctuations are computed using the mean square deviations of each model to the mean model [40, 41, 22]. In order to remove proteins having intrinsic disordered regions, all proteins showing a root mean square displacements of experimental fluctuations over all residues larger than 2Å $\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N} MSD_i} > 2\text{Å}\right)$ were removed from the data set. It leads to 132 structures.

F Experimental methods

The experimental methods used to determine protein structures are important because each of them has its own drawbacks and we have to be aware of that when interpreting the theoretical results based on these protein structures.
We have used structure data coming from X-ray crystallography and from NMR spectroscopy. The former method consists in packing the protein into a crystal and passing beams through this crystal. The protein structure and its features are determined using the diffraction pattern. The drawback of this method is

that the atomic fluctuations may be damped by the packing into the crystal leading to smaller experimental B-factors than what they are in the reality. That would affect essentially the flexible parts. This effect of the crystal on the atomic fluctuations has already been discussed [42] and some studies [20] have tried to take it into account in their modelling. But they model the protein into the crystal and not in their natural environment which is out of aim.
The latter method, NMR spectroscopy, places the protein into a magnetic field and look at its resonance pattern by radio waves. This method allows to determine protein structures in solution and then takes into account all fluctuations due to the solvent when reproducing the structure. However, it does not allow to consider large proteins, only small proteins or domains of large proteins can be probed by this method.

The construction of the spring constant matrix of sequence specific models in [22] has been done using NMR data. Although it contains all information related to the protein in solution, it has been determined using only small proteins which can be a little bit limited: we do not expect the effect of the solvent to be the same for small proteins as for large ones.
Furthermore, in our comparative study, we have used mainly X-ray crystallographic data. The reason is that we have made our study on a set of popular proteins which are large proteins. Also, often, pdb files determined by NMR spectroscopy do not include the B-factor values and we cannot compare the theory to the experiments. We have still carried out a NMR study based on the experimental mean square fluctuations (MSF) calculated from the different models available in pdb files of NMR structures.

G Experimental materials

Follow are the experimental protein structures used for our in-depth studies: Kinesin KIF1A (pdb id: 1i5s and 1i6i) [43], Human kinesin motor domain (pdb id: 1bg2 and 1mkj) [44, 45], Aspartate aminotransferase (pdb id: 9aat, 1ama and 1ivr) [46, 47, 48], Factor H binding protein (pdb id: 2kc0) [49], Maltodextrin binding protein (pdb id: 1jw4, 1anf and 1omp) [50, 51, 52], B1-Type ACP domain (pdb id: 6h0q) [53], Adenylate kinase (pdb id: 1aky, 2ak3 and 4ake) [54, 55, 56], Myosin V (pdb id: 1w7j and 1oe9) [57, 58], Annexin V (pdb id: 1avr and 1avh) [59], Ubiquitin (pdb id: 1xqq) [60], Type IV pilin PILE1 (pdb id: 6i2o) [61], HCV Helicase (pdb id: 1hei) [62], HIV-1 protease (pdb id: 1hhp and 1ajx) [63, 64], Enolase (pdb id: 5enl and 3enl) [65, 66], Thymidylate synthase (pdb id: 2tsc and 3tms) [67, 68], Scallop myosin (pdb id: 1kk8 and 1kk7) [69], F_1-ATPase (pdb id: 1h8h and 1h8e) [70, 71], Penicillin binding protein (pdb id: 1vqq, 4dki and 3zg0) [72, 73, 74], A1-Typr ACP domain (pdb id: 6h0j) [53], VAT-N (pdb id: 1cz4) [75], Acyl carrier protein (pdb id: 5y08), SPH protein (pdb id: 6g7g) [76], Hydrolase (pdb id: 6qeb) [77].

III Sequence specific models

A Models

This section is dedicated to the comparisons of sequence-specific models recently suggested to the classical ones through two main aspects: individual residue fluctuations and collective motions. The aim is to determine where the improvements are located in the protein. Although the authors of these new networks have made a systematic analysis over a large set of proteins, the observed improvement is seen at a protein level and it is still not explained at a residue level. Then we expect some deterministic and systematic improvements specifically localised coming from the local structure, polarity or anything else. A short discussion on our analysis about mutations, which were our original aim, is also available.

Three sequence-specific models have been published in 2013 [22]: two of them have a cut-off distance, 10Å and 13Å respectively. The spring constants are solely sequence-dependent and do not dependent on the distance. Their spring constants range from 0.226 to 2.348. The spring constant of the links forming the backbone are artificially set to 10. They will be designated as sANM10 and sANM13 respectively. The last model is presented as cut-off free but still spring constants are set to 0 if the distance is larger than 16.5Å. This model is both sequence and distance-specific although the distance dependence is not continue but discrete. Its spring constant values range from 0.001 for long-range distance pairs to 13.043. The links of the backbone are artificially set to 43.52. It will be designated as sdANM. In the remain of this thesis, "the sequence-specific models" will refer to these three models.

As for the classical model with a cut-off distance l_c, it will be designated as ANMl_c.

Before starting any comparison, we have to clarify something important. Often in the literature, the introduction of a new heterogeneous model comes with a comparison with classical ANMs. However, generally, when introducing such a model, in addition to change the homogeneity of the ANMs, the connectivity of the theoretical network is also changed. That is a problem because the conclusions hold on the positive effect of the heterogeneity while the method compare the new model with an ANM having a “very-often used” cut-off distance. The results actually test the effect of both the heterogeneity and the change in the connectivity.

We have to keep in mind that to test the effect of the heterogeneity, we have to compare it to the homogeneous model with the corresponding connectivity (i.e. cut-off distance). sANM10 and sANM13 are only sequence-specific and have an explicit cut-off distance, they will be compared to ANM10 and ANM13. As for sdANM, it is both distance and sequence-specific and does not consider interactions larger than 16.5Å, we will then compare it to ANM16.

We have also to keep in mind that we can conclude a heterogeneous parameter-free model performs better than the ANMs if it performs better than any of the ANMs disregarding its cut-off distance.

B Individual fluctuations

1 Study at the residue level

The results of this section was published in [78].

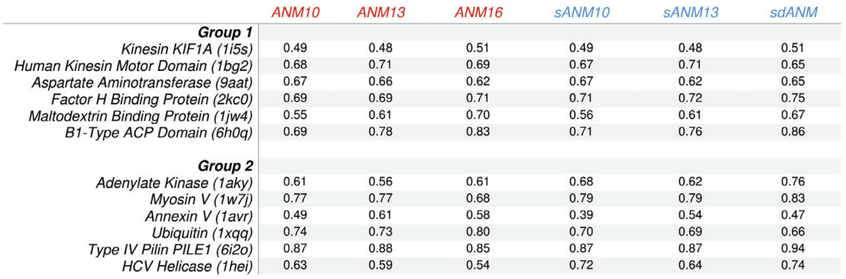
Each residue fluctuates around its equilibrium position. These fluctuations can be determined both experimentally and theoretically for each residue (see Materials and methods section). By testing their correlation, we obtain a criteria of accuracy: the Pearson correlation coefficient. Using this criteria as well as the B-factor patterns, we compare both kinds of model and try to determine where the sequence-specific models perform better. We have carried out an analysis on around 40 proteins. Following our results, two groups emerge: the first one where the ANMs and sANMs perform equally and the second one where an improvement or a decline is observed. We have selected 6 proteins of each group to display their results in this section (see Table 1) while the results on the full set of proteins are available in Table 11 in Appendix.

As for the first group, comparing with the same cut-off distance, the correlation is very robust. The largest difference, holding for Human kinesin motor domain and for factor H binding protein between ANM16 and sdANM, is 0.04. The patterns themselves are also quite robust (see Figure 3 A and B), maltodextrin binding protein (resp. factor H binding protein) displays the same pattern for both ANM10 (resp. ANM13) and sANM10 (resp. sANM13). However, for Human kinesin motor domain, sdANM shows a different pattern as compared to ANM16. Several parts are better estimated by sdANM and some others by ANM16 but as an average criteria, the correlation does not show a large difference (0.69 for ANM16 and 0.65 for sdANM). Two of these parts correspond to unstructured regions and the last one corresponds to the end of a beta-sheet. For the three of them, they are poorly connected in the network as compared to other parts. Interestingly, for this protein, the B-factor pattern of sdANM matches well the one of ANM10.

As for the second group, there are quite large differences: improvements of sdANM over ANM16 of 0.15 for adenylate kinase and myosin V, declines of sdANM of 0.1 (resp. 0.14) for annexin V (resp. ubiquitin), etc. The B-factor patterns of myosin V show a large peak for both ANM16 and sdANM (see Figure 4A). It corresponds to a set of a few residues which form an unstructured region connecting two beta-strands and are poorly connected in the network. The improvement is almost solely due to this part. For the other cut-off distances, the ANMs and the sANMs perform similarly. Adenylate kinase also has such unstructured flexible regions where for some, sdANM performs better and for some others, ANM16 performs better (see Figure 4B). Ubiquitin shows a large drop (0.14) in sdANM. That is due to the over-estimation of the end-tail in sdANM while fluctuations are damped in ANM16 (see Figure 4C). Similarly in annexin V, a part of the tail is over-estimated by both models but more by sANM10 (see Figure 4D).

The case of HCV helicase is interesting. It can perform a large motion involving two of its domains which is well modelled by the ANM8 [9]. However, with a larger cut-off distance, this motion is inhibited. The PCC of HCV helicase, indeed, decreases for classical ANMs as the cut-off distance increases while in sdANM, the PCC is enhanced compared to both sANMs and to classical ANMs (see Table 1). The correlation obtained for ANM8 is 0.79 (not shown) which is very close to the one obtained for sdANM (0.74).

It is worth to note that not only the fluctuations in the tails can be over-estimated but also in some regions in the protein which can not be removed by hand to carry out the analysis. These over-estimations appear on flexible parts which are poorly connected. These parts prevent us to draw coherent and systematic conclusions for how and where the sequence specific models perform better. This problem known in the classical ANMs is not solved in the new sequence-specific models and it is even enhanced because of the softening of long-range interactions. In the other side, this softening brought benefits to HCV helicase for which sdANM has been able to reproduce the thermal fluctuations of ANM8, the network working the best for this protein.
In the next subsection, we have tried a more systematic way by applying the models to a large set of proteins in order to have consistent arguments to relate these over-estimations to flexible parts.



	ANM10	*ANM13*	*ANM16*	*sANM10*	*sANM13*	*sdANM*
Group 1						
Kinesin KIF1A (1i5s)	0.49	0.48	0.51	0.49	0.48	0.51
Human Kinesin Motor Domain (1bg2)	0.68	0.71	0.69	0.67	0.71	0.65
Aspartate Aminotransferase (9aat)	0.67	0.66	0.62	0.67	0.62	0.65
Factor H Binding Protein (2kc0)	0.69	0.69	0.71	0.71	0.72	0.75
Maltodextrin Binding Protein (1jw4)	0.55	0.61	0.70	0.56	0.61	0.67
B1-Type ACP Domain (6h0q)	0.69	0.78	0.83	0.71	0.76	0.86
Group 2						
Adenylate Kinase (1aky)	0.61	0.56	0.61	0.68	0.62	0.76
Myosin V (1w7j)	0.77	0.77	0.68	0.79	0.79	0.83
Annexin V (1avr)	0.49	0.61	0.58	0.39	0.54	0.47
Ubiquitin (1xqq)	0.74	0.73	0.80	0.70	0.69	0.66
Type IV Pilin PILE1 (6i2o)	0.87	0.88	0.85	0.87	0.87	0.94
HCV Helicase (1hei)	0.63	0.59	0.54	0.72	0.64	0.74

Table 1: Table of Pearson correlation coefficients between experimental and theoretical atom fluctuations for a selection of 12 proteins divided into two groups. The first group is composed of proteins for which no changes have been observed in the correlation between the two kinds of model when comparing with the same cut-off distance. The second group gathers the other proteins. A larger table is available in Appendix in table 11. Red colour is related to classical ANM while blue colour is related to sequence-specific models.

2 Systematic study on crystallographic structures

For the systematic study, the protein selection process is described in the materials and methods section. It leads to 2040 structures for ANM16/sdANM, 2038 structures for ANM13/sANM13 and 2009 structures for ANM10/sANM10.
Our aim is to look at the estimation of the B-factor according to the "flexibility" of the residue defined as the relative degree (see materials and methods section). All models mainly over-estimate residues having a small relative degree and then a large flexibility (see Figure 5). There is no major differences between sANMs and their corresponding ANMs. The over-estimations of flexible parts are further enhanced by sdANM compared to ANM16 and the highly connected parts are much more under-estimated. The reason could be that in sdANM, the long-range interactions have been softened while the short-range interactions have been stiffened. Globally, it makes the estimations of fluctuations worse.
We also looked at the averaged estimation errors of the B-factor according to

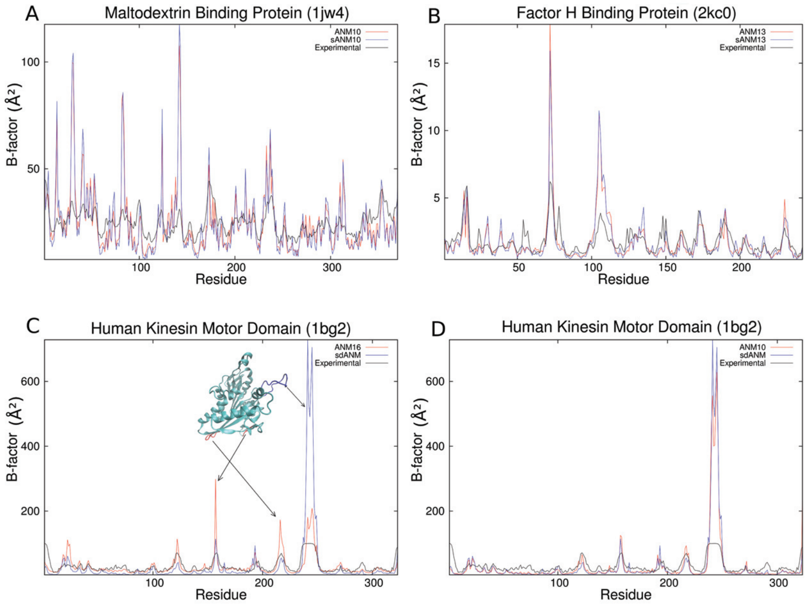


Figure 3: B-factor patterns for some proteins of the first group. Visual representations have been obtained with VMD.

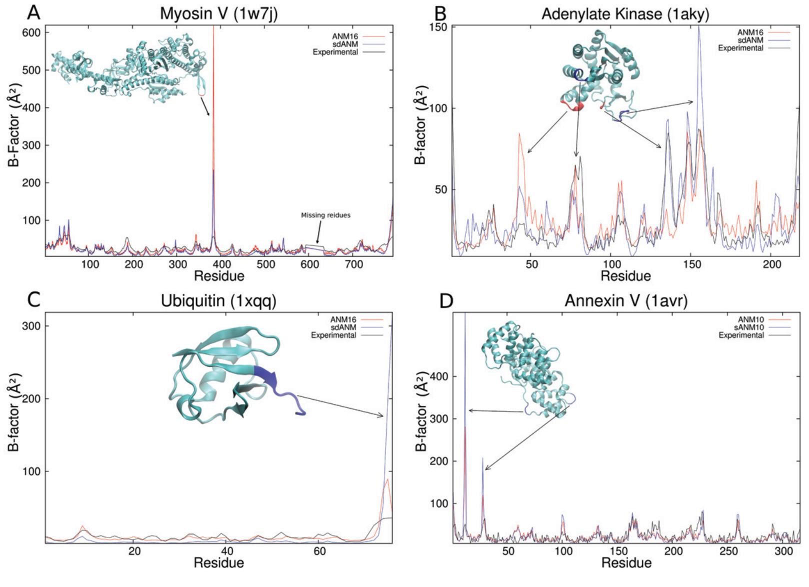


Figure 4: B-factor patterns for some proteins of the second group. Visual representations have been obtained with VMD.

individual residue features which are polarity and secondary structure. This is a way to try to determine if a particular feature enhances or not the estimations when considering the sequence specificity. The error is systematically increased for sequence specific models as compared to their classical ANM counterpart (see Table 2). Although for sANM10 and sANM13, this increase in the error is not so large, sdANM shows the largest errors and in particular its values are closer to the ones given by ANM10 or sANM10 than the ones given by ANM16. The long-range interactions which have been softened in sdANM seem to decline the modelling and to make it somehow resemble a modelling of ANM10. This remark is further supported by the averaged PCC (0.51 for ANM10, 0.52 for sdANM and 0.55 for ANM16) though the values are still similar for the three models.

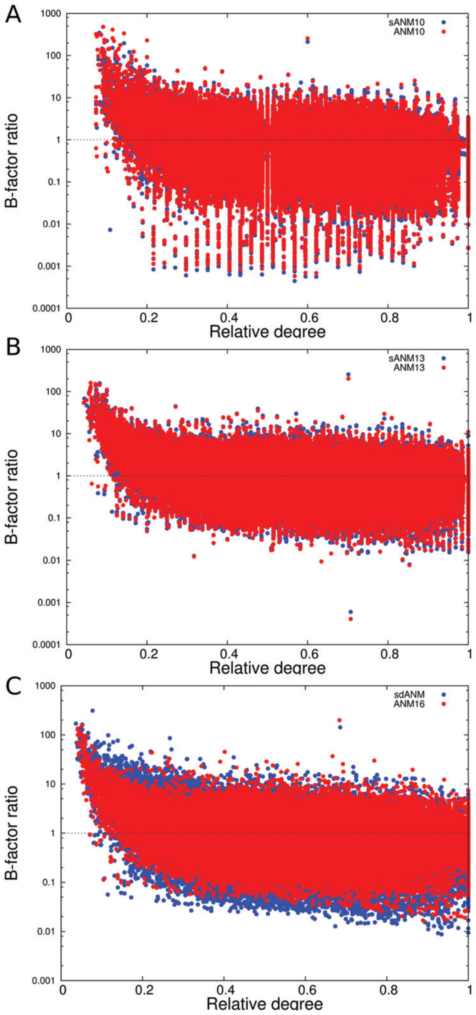


Figure 5: B-factor ratio of the theory over the experiment according to the relative degree (i.e. the degree rescaled by the maximal degree of the protein into which the residue belongs). A: blue points are obtained with sANM10 while red points are obtained with ANM10. B: blue=sANM13 and red=ANM13. C: blue=sdANM and red=ANM16.

	ANM10	ANM13	ANM16	sANM10	sANM13	sdANM
Average PCC	0.51	0.53	0.55	0.50	0.51	0.52
Polarity						
Hydrophobic	8.3	7.0	6.2	9.1	8.7	9.5
Polar	9.0	7.3	6.5	9.1	7.8	9.6
Charged	9.5	8.2	7.3	9.5	8.2	10.5
Secondary Structure						
Alpha Helix	7.2	6.3	5.7	7.5	7.0	7.6
Beta Sheet	6.3	5.4	4.8	6.7	6.3	7.2
Unstructured	11.9	9.7	8.4	12.5	10.9	13.5

Table 2: Summary table of results obtained over 2040 X-ray crystallographic protein structures for sdANM and ANM16, 2038 structures for sANM13 and ANM13 and 2009 structures for sANM10 and ANM10. For the polarities and the secondary structures, the averaged B-factor estimation errors are computed.

3 Analysis of solution NMR structures

Because the sequence specific models have been determined for solution NMR structures, we carried out a study over 132 proteins. The average PCC is indeed improved whatever the cut-off distance we compare. However, the addition of only the sequence does not bring a consequent improvement (see the polarity and secondary structure details in ANM10-sANM10 and in ANM13-sANM13 in Table 3); the errors are reduced, at most, by 0.05Å. As for the sdANM, which includes also a dependence on the distance, there is a real improvement compared to ANM16 (generally, the error is reduced by more than 0.1Å) except for unstructured regions. For these latter regions, the error is quite large, whatever the model. Even for NMR structures, unstructured regions are still over-estimated by the ENM.

	ANM10	ANM13	ANM16	sANM10	sANM13	sdANM
Average PCC	0.64	0.61	0.60	0.67	0.64	0.72
Polarity						
Hydrophobic	0.68	0.73	0.79	0.64	0.69	0.61
Polar	0.70	0.72	0.75	0.68	0.72	0.61
Charged	0.76	0.79	0.80	0.74	0.76	0.66
Secondary Structure						
Alpha Helix	0.36	0.44	0.47	0.31	0.39	0.23
Beta Sheet	0.23	0.27	0.31	0.18	0.23	0.15
Unstructured	1.21	1.20	1.24	1.20	1.21	1.17

Table 3: Summary table of results obtained over 132 solution NMR structure proteins. For the polarities and the secondary structures, the averaged MSF estimation errors are computed.

4 Discussion

We have looked at the thermal fluctuations of individual residues in order to understand and to explain the improvement seen at the protein scale by the newly sequence-specific models. It emerged that unstructured regions, which are known to be badly modelled and to trigger what is called the tip-effect [79, 80], are still miss-predicted by the sequence-specific models. The prediction on these parts are either improved or worsened in a non-systematic way which degrades the value of the overall improvement in the PCC.

As for the structured regions, the addition of the sequence only (sANM10 and sANM13) does not bring a large benefit to the modelling. Indeed, as compared to their ANM counter-part, the errors are slightly decreased for NMR proteins (Table 3) while they are even larger for crystallographic structures (Table 2). When considering also the distance dependence (sdANM), it turns out differently whether considering X-ray crystallographic structures or solution NMR structures. For the former ones, the errors are larger while for the latter ones, they are consistently smaller. The new models, determined using a NMR dataset, do not seem suitable for crystallographic proteins. A possible reason may come from the crystal packing hampering the fluctuations while the sdANM has very soft long-range interactions due to free fluctuations in solution NMR experiments.

We cannot make a direct comparison of error values between X-ray crystallographic data and solution NMR data since the experimental fluctuations are measured differently.

C Collective aspects

1 Transitions

The overlaps between the experimental motion between two known states and its theoretical estimation allow to explain the transition as a decomposition of fundamental motions. Each mode has an overlap between 0 and 1. The closer to 1, the largest the contribution of the mode to the transition motion. It has been shown that, often, information resides in a single mode [15, 16]. This is a strong result since a single fundamental motion (which can be computationally calculated and observed) can explain the transition motion between the two structures. It is then of importance to have a single mode with a large overlap. We have looked at the maximal overlap for the transition of some proteins for all models (see Table 4). sANMs predict similar maximal overlap to their corresponding ANMs, the largest difference being of 0.15 in thymidylate Synthase between sANM10 and ANM10. As for sdANM, the maximal overlap is often larger than the one given by ANM16 (e.g. for myosin V, HIV-1 protease, aspartate aminotransferase, maltodextrin binding protein, etc). Due to its softened interactions, the model is more prone to show large motions. Also, it is interesting to note that, again, the results of sdANM are more closely related to the ones of ANM10 than to the ones of ANM16.

To look deeper into the transition, we have looked at the overlap curves for some examples. The ones for HIV-1 protease for the transition from its free structure (pdb id: 1hhp) to a complex structure with an inhibitor (pdb id: 1ajx) and for scallop myosin from its actin-detached conformation (pdb id: 1kk8) to its near rigor conformation (pdb id: 1kk7), are displayed in Figure 6. In classical models, the modes have swapped according to the cut-off distance. It is still true in the sequence specific models.

That is quite typical to what we have found for the other proteins (not shown): either the dominant mode is swapped because of the cut-off distance or there is not such a dominant mode. Only the case of maltodextrin binding protein retained our attention. The dominant mode has swapped in sdANM compared to ANM16 (see Figure 7). The fundamental motion associated to the dominant mode remains the same (see Figure 8, left side). The magnitude of the two eigenvectors are almost the same excepted at one residue (see Figure 8). This residue is located in a region bridging two beta-strands and it is the least connected residue in the theoretical network. This residue contributes to increase the frequency of the dominant mode in ANM16 by inhibiting the overall motion. Mathematically, the coefficient is increased when rescaling the eigenvector. The individual magnitudes are then decreased. It makes the slowest mode swap with the neighbour mode.

	ANM10	*ANM13*	*ANM16*	*sANM10*	*sANM13*	*sdANM*
Myosin V						
1w7j to 1oe9	0.63	0.56	0.54	0.63	0.53	0.63
1oe9 to 1w7j	0.55	0.63	0.62	0.57	0.61	0.56
HIV-1 Protease						
1hhp to 1ajx	0.71	0.67	0.67	0.59	0.77	0.77
1ajx to 1hhp	0.44	0.34	0.31	0.47	0.46	0.50
Aspartate Aminotransferase						
9aat to 1ama	0.57	0.55	0.64	0.64	0.56	0.61
1ama to 9aat	0.55	0.66	0.57	0.60	0.73	0.73
Enolase						
3enl to 5enl	0.23	0.21	0.21	0.22	0.19	0.18
5enl to 3enl	0.23	0.22	0.23	0.24	0.19	0.20
Maltodextrin Binding Protein						
1anf to 1jw4	0.90	0.88	0.80	0.89	0.87	0.87
1jw4 to 1anf	0.78	0.74	0.79	0.84	0.71	0.71
Adenylate Kinase						
1aky to 2ak3	0.38	0.38	0.36	0.38	0.38	0.38
2ak3 to 1aky	0.27	0.31	0.31	0.30	0.33	0.33
Annexin V						
1avr to 1avh	0.41	0.33	0.33	0.40	0.33	0.43
1avh to 1avr	0.33	0.33	0.33	0.35	0.37	0.39
Thymidylate Synthase						
2tsc to 3tms	0.44	0.29	0.24	0.28	0.27	0.37
3tms to 2tsc	0.46	0.37	0.28	0.57	0.34	0.41
Scallop Myosin						
1kk8 to 1kk7	0.85	0.76	0.75	0.85	0.75	0.82
1kk7 to 1kk8	0.70	0.72	0.74	0.70	0.73	0.75
F1-ATPase						
1h8h to 1h8e	0.55	0.51	0.49	0.55	0.47	0.49
1h8e to 1h8h	0.54	0.54	0.53	0.50	0.53	0.51
Kinesin Motor Domain						
1mkj to 1bg2	0.30	0.26	0.24	0.28	0.26	0.42
1bg2 to 1mkj	0.23	0.26	0.18	0.29	0.18	0.25
Penicillin Binding Protein						
1vqq to 4dki	0.33	0.33	0.37	0.33	0.34	0.32
4dki to 1vqq	0.33	0.31	0.30	0.32	0.33	0.32
KIF1A Motor Domain						
1i6i to 1i5s	0.35	0.37	0.24	0.22	0.31	0.27
1i5s to 1i6i	0.28	0.28	0.20	0.24	0.27	0.33

Table 4: Table of the largest overlap value corresponding to the motion contributing the most to the transition for all models and for different proteins.

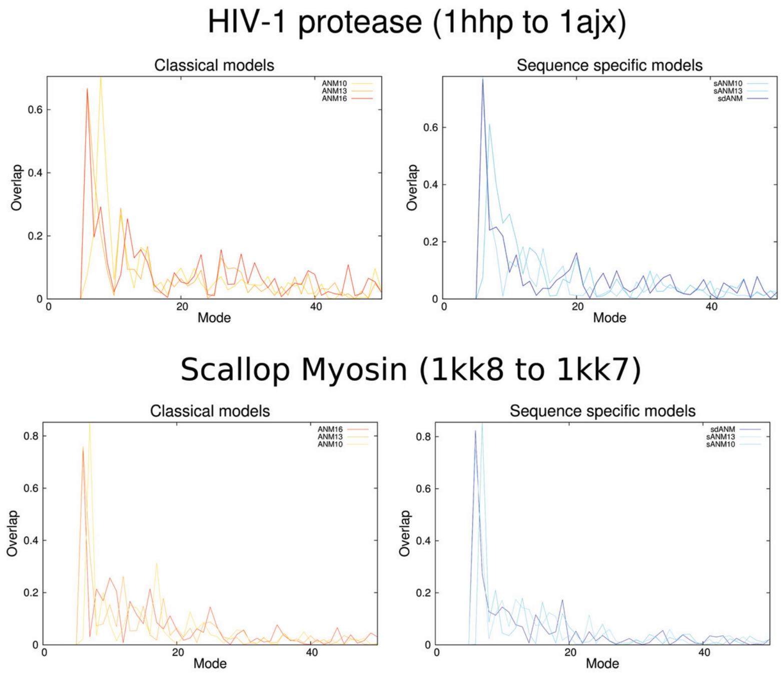


Figure 6: Top panel: Overlaps of the transition of HIV-1 Protease from its free structure (pdb id: 1hhp) to its inhibitor-complexed structure (pdb id: 1ajx). Bottom panel: Overlaps of the transition of scallop myosin from its actin-detached conformation (pdb id: 1kk8) to its near rigor conformation (pdb id: 1kk7). The NMA has been carried out for the first structure. Only the first 50 modes are displayed.

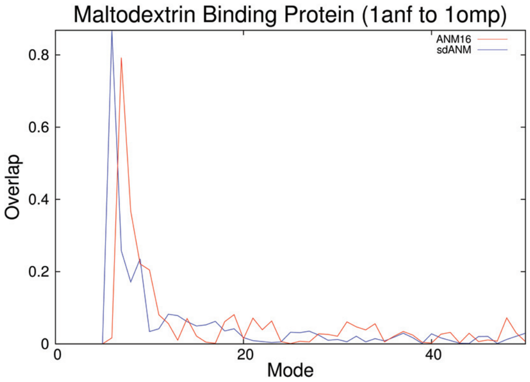


Figure 7: Overlaps of the transition of maltodextrin binding protein from its structure adopted when complexed with maltose (pdb id: 1anf) to its free structure (pdb id: 1omp). Only the first 50 modes are displayed.

2 Inter-residues interactions

Another collective aspect is the inter-residue interaction variations. They have been used to show an improvement of sANMs and sdANM over classical ANMs

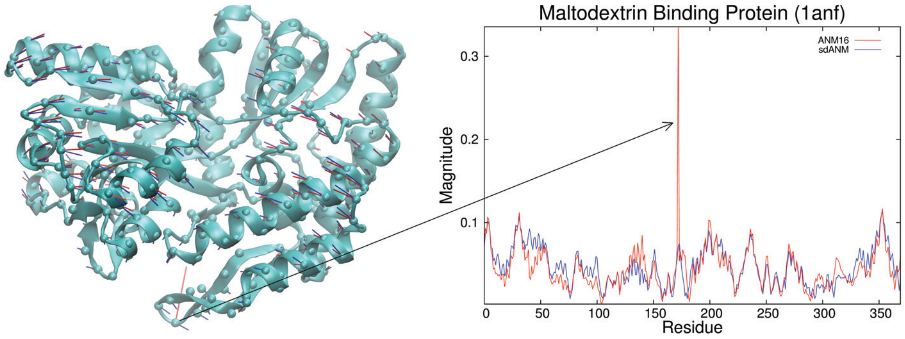


Figure 8: Left: A visual representation of maltodextrin binding protein (pdb id: 1anf) with the eigenvector corresponding to the dominant mode (mode 8 for ANM16, red vectors and mode 7 for sdANM, blue vectors). Right: Magnitude of the eigenvector corresponding to the dominant mode along each residue.

[22]. In this paragraph, we investigate where these improvements are located through three examples, one for each cut-off distance (see Figure 9). The experimental data used to compute the inter-residue variance is available only for NMR structures. We compare directly the ANM with its corresponding sequence specific version by calculating and plotting the difference of the variance error between them (ANM minus its sequence-specific variant) for each residue. A negative value means that ANM performs better than sANM and a positive value means the reverse.

Ubiquitin shows differences between ANM16 and sdANM at the tail where the B-factors are over-estimated (See Figure 9A). The introduction of the sequence specificity brings a decline on these parts for the estimations of inter-residue fluctuations. That is coherent with what happens for individual fluctuations.

In sANM10 and sANM13, the spring constants of the backbone are artificially set to 10 in order to stiffen it. However, it just accentuates even more the error in the fluctuations (see Figure 9 B and C, the diagonal in the heat maps of SPH protein and of factor H binding protein). SPH protein shows a decline in the sANM13 for some pairs of residues corresponding to intereactions between two beads in two different unstructured and flexible regions within the protein where the individual fluctuations are under-estimated. Similarly, factor H binding protein shows an improvement of sANM10 over ANM10 for pairs of residues located in two different regions for which the individual fluctuations are over-estimated.

In our three examples, excepted flexible parts, there is no striking differences between the classical ANM and its sequence specific version in the other parts. Besides they perform very similarly, the only differences between the two kinds of model are located on badly modelled parts of the proteins.

3 Discussion

These studies on the overlaps and on the inter-residue interactions support the study on the B-factors in the way that the sequence specificity does not bring a valuable improvement in the modelling. Worse, it can bring a decline due to its soft large range interactions (see ubiquitin in Figure 9A).

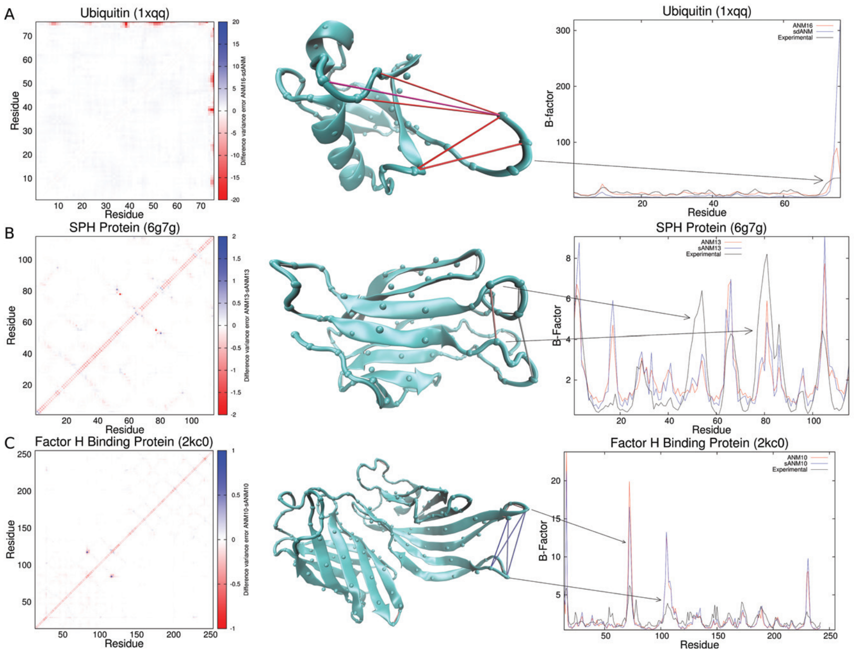


Figure 9: Heat maps show the difference between the inter-residue variance error in ANM minus its corresponding sequence specific version. The more red, the better ANM over sANM. The more blue, the better sANM over ANM. Visual representations show the links having the larger difference seen in the heat maps. The colour code is the same as for heat maps and only links with a difference larger than 15 for A. or 0.5 for B. and C. are displayed. The B-factor patterns are displayed for comparison. A. ANM16 and sdANM are compared for ubiquitin. B. ANM13 and sANM13 are compared for SPH protein. C. ANM10 and sANM10 are compared for factor H binding protein.

As compared to the B-factor where the residues are considered individually, the inter-residue variance considers pairwise distances. It contains additional information about the direction of fluctuations. However, the problems are similar. Indeed, the study on inter-residue fluctuations shows the same problems located at the same places in the protein (i.e. flexible parts) as the study on the B-factors and shows that it is even worse than expected: not only the individual fluctuations are not enhanced but the error in the fluctuations between two flexible parts in the protein can be increased.

D Relaxation trajectories

As determined in the materials and methods section, within the hypothesis of linearity, the over-damped temporal displacement of residues is easily calculated and is a sum of decreasing exponential functions. In the same time, the dynamical simulations allow to look at the relaxation without supposing the linearity. It gives two important pieces of information: how the molecule relaxes and in which extent the hypothesis of linearity is valid. The latter one is answered by a comparison between the two methods. Generally, the linearity holds at the vicinity of the equilibrium state (because small fluctuations are de facto assumed) then the dynamical simulation relaxation would match the relaxation governed by the slowest modes.

The dynamical simulation relaxations are followed via the distances between three residues chosen as labels.

Three motor proteins are considered in this section: HCV helicase (Figure 10), F_1-ATPase (Figure 12) and myosin V (Figure 14). The three of them show large motions and are then well-suited for such analysis. The chosen labels and other details about deformations are stated in appendix. Since the deformed state is random, 100 trajectories are studied for each model for each protein.

In preceding works [38, 39, 9], it has been found that, often for motor proteins, with classical ANMs, the trajectories relax until reaching a deterministic pathway from which trajectories cannot escape and are slowly brought to the equilibrium state. This deterministic pathway corresponds to the relaxation along the slowest mode where linearisation holds.

HCV helicase performs a large motion involving two domains which has been shown to be well modelled by ANM8 but inhibited by ANM with a larger cut-off distance. In the B-factor section, we have seen that the correlation coefficient given by sdANM is quite similar to the one given by ANM8 letting suggest that the motion of interest could be retrieved by sdANM despite of its high connectivity. Unfortunately, the relaxation trajectories do not confirm this fact and suggest the opposite (Figure 11). As expected for ANM, the well-defined pathway is present (Figure 11 B and C). In sdANM, this pathway has been broken down and the trajectories relax in a purely random way to the equilibrium state. That raises two draw-backs: first, the motion of interest is inhibited and second, the linear approximation (and hence the normal mode analysis) holds only when the conformation is very close to the equilibrium position. That is confirmed by the elastic potential energy according to the mechanical coordinate (Figure 11A).

For F_1-ATPase, it has been shown that the linear approximation holds only very closely to the equilibrium state [39] when using ANM10, despite having a kind of valley where trajectories are deterministically driven to the equilibrium. The corresponding sequence specific model, sANM10, shows poor difference and does not show any improvement (Figure 13). Especially, the potential energy within the linear assumption does not match better the one of the dynamical simulations. Neither sANM10 nor sANM13 nor sdANM gives better results (see Figures 36 to 38 in the appendix).

For myosin V, the range of validity of the NMA in ANM10 is much larger than for F_1-ATPase [39]. The range is similar for other cutoff distances 13Å and 16Å (Figures 15B and 43). Again, the addition of the sequence specificity does not bring any enhancement. The performance is rather similar.

Transitions can also be studied with relaxation trajectories. Either the sequence specificity does not bring a valuable difference or the difference deserves it. For HIV-1 protease, from the complexed structure (pdb id: 1ajx), all models go back to the free structure (pdb id: 1hhp) without being trapped by a metastable state. The pattern of relaxation is quasi-similar for all models excepted ANM16 (Figure 16B). As for the relaxation of the elastic potential energy, there is no improvement in the range where the normal mode analysis holds (Figure

16A). Nevertheless, we can note that, again, sdANM resembles more ANM10 than ANM16.
For adenylate kinase, only two models are able to go back to the initial structure (Figure 16D): ANM10 and ANM16. The sequence specific models totally failed and ANM13 is trapped in a metastable state very close to the initial structure. It is even clearer on the elastic energy profile (Figure 16C).

These results are interesting because even though the correlation or the overlap are increased, it does not mean that the model is better upon the linear approximation or that this assumption is true on a larger range of deformations. That is something which has to be kept in mind when evaluating a model using B-factors and the Pearson correlation coefficient in order to make molecular dynamics simulations. What is evaluated by the normal mode analysis is valid only in the linear assumption and it is not guaranteed to work in dynamical simulations where the dynamics are not linear at all.

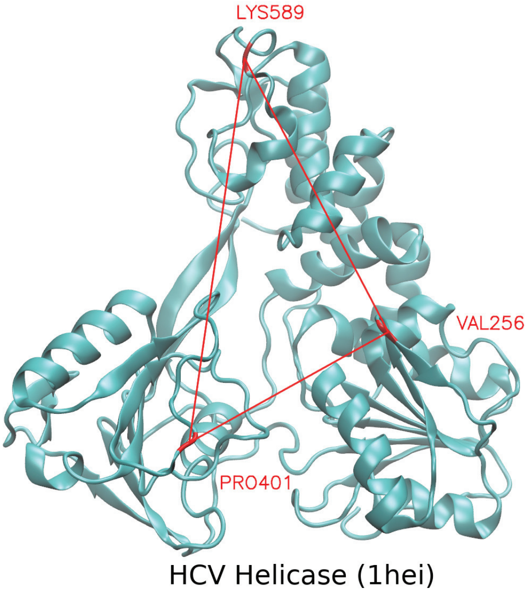


Figure 10: Visual representation of HCV helicase (pdb id: 1hei) with the three labels used.

E Mutations

By using the sequence-specific models, our main goal was to study the effects of mutations. With what we have observed in the preceding subsections, this goal becomes a little bit hopeless. Indeed, if we do not find any difference between the sequence-specific models and the classical ANMs, then it is unlikely we find them if we change the type of residues. We will still have a short look at it and confirm what said above.
In a first step, we wanted to study the effect of a single mutation. For that, we have used a brute-force way where we tried all possible mutations for each residue of a single protein. There are 20 types of residue, so each residue can

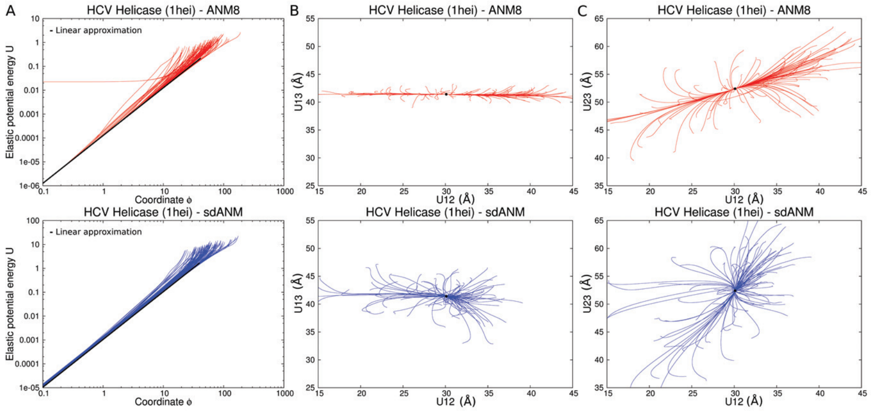


Figure 11: Comparison between ANM8 and sdANM for HCV helicase. A: the elastic potential energy along the mechanical coordinate for 100 simulations (red for ANM8, blue for sdANM) and for the slowest mode (black), B: distance between labels 1 and 3 along the distance between labels 1 and 2 for the 100 simulations, C: distance between labels 2 and 3 along the distance between labels 1 and 2.

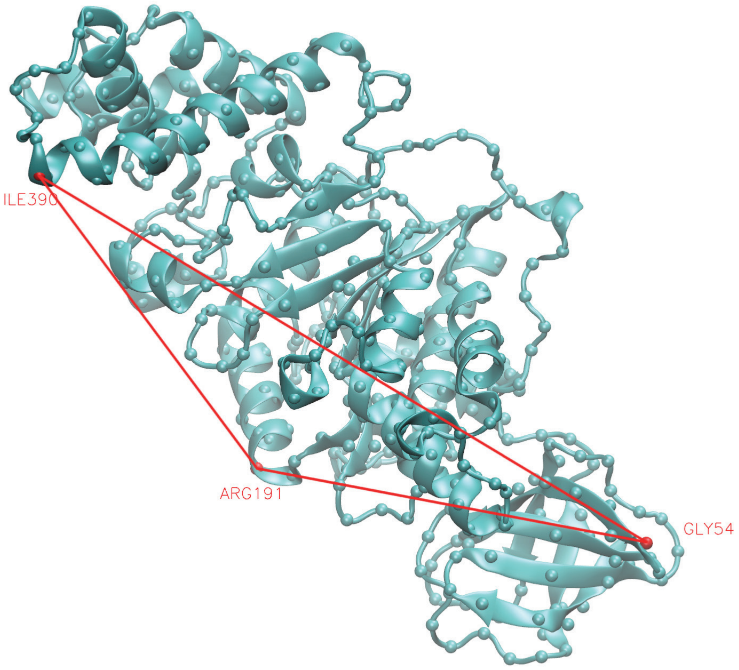


Figure 12: Visual representation of F_1-ATPase (pdb id: 1h8h) with the three labels used.

carry out 19 different mutations. We have chosen the protein HIV-1 protease (pdb id: 1hhp). HIV-1 protease is a dimer of 198 residues. Here, we studied

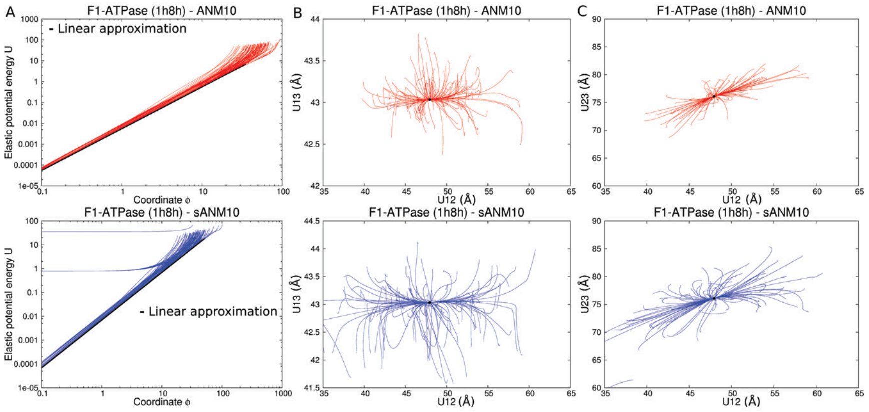


Figure 13: Comparison between ANM10 and sANM10 for F1-ATPase. A: the elastic potential energy along the mechanical coordinate for 100 simulations (red for ANM10, blue for sANM10) and for the slowest mode (black), B: distance between labels 1 and 3 along the distance between labels 1 and 2 for the 100 simulations, C: distance between labels 2 and 3 along the distance between labels 1 and 2.

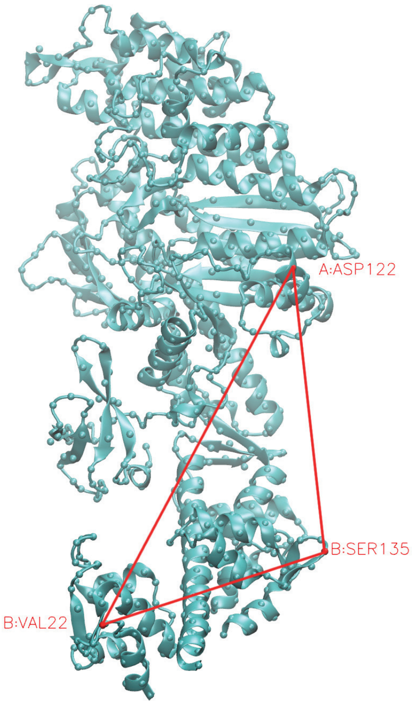


Figure 14: Visual representation of myosin V (pdb id: 1w7j) with the three labels used.

only the monomer (99 residues). We have calculated the overlaps for the transition 1hhp to 1ajx for each mutant (see Figure 17) as well as the root mean

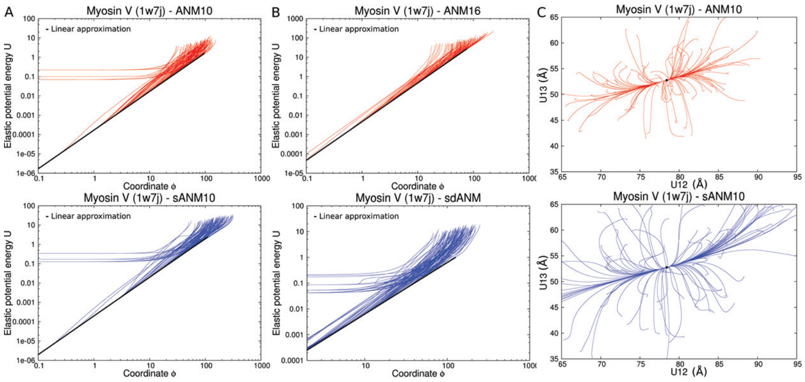


Figure 15: Comparison between ANM10 and sANM10 for myosin V. A: the elastic potential energy along the mechanical coordinate for 100 simulations (red for ANM10, blue for sANM10) and for the slowest mode (black), B: same as A but red is ANM16 and blue is sdANM, C: distance between labels 1 and 3 along the distance between labels 1 and 2.

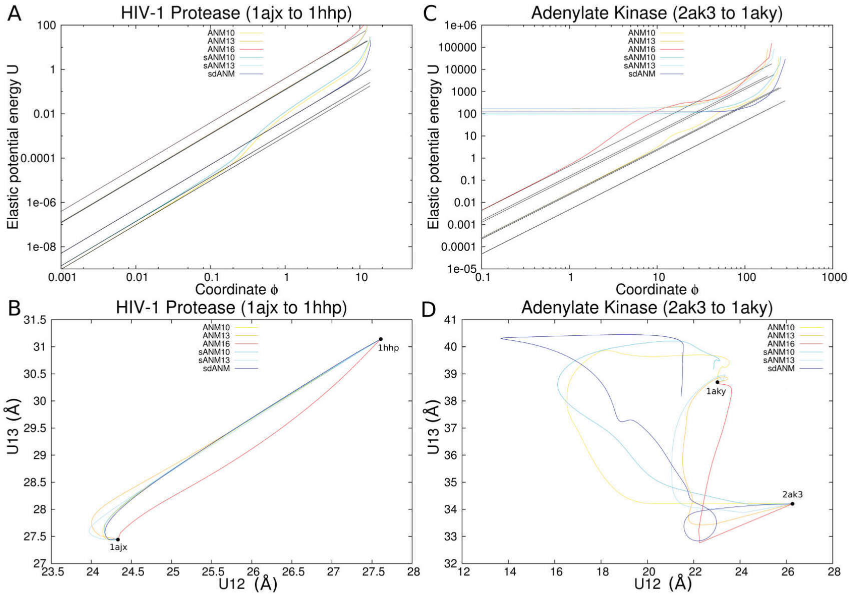


Figure 16: Study of the transition between two known structures. A and B: Elastic potential energy and relaxation trajectories for the transition of HIV-1 Protease from 1ajx to 1hhp. C and D: the same for the transition of adenylate kinase from 2ak3 to 1aky.

square difference (RMSD) between the b-factors of the mutant and the ones of the wild-type (see Figure 18).

As for the overlaps, only the first 5 first non-zero modes are displayed. Excepted for some cases, the overlap values remain robust (Figure 17, top-panel). Only one case retained our attention, the mutant 1830 (Figure 17, bottom-panel). In this case, the dominant mode has been replaced by the following mode although the two have still similar overlap values. leucine 97 has mutated to a glycine residue. We have not investigated more this case as it is not so interesting.

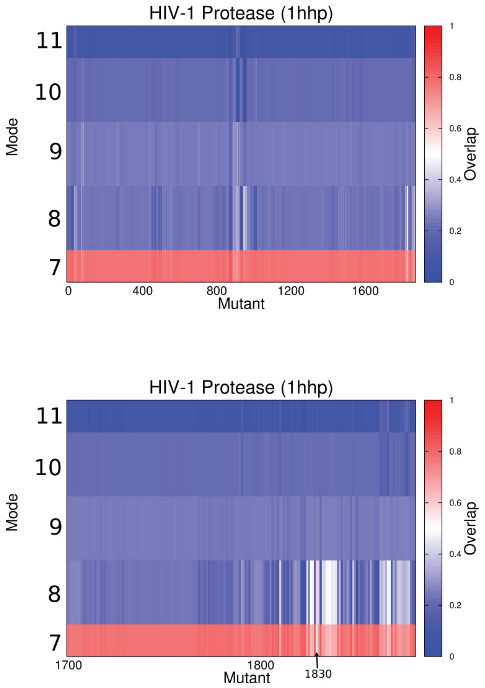


Figure 17: Overlap value for the 5 first non-zero modes for each possible single-mutation mutant for the protein HIV-1 protease (pdb id: 1hhp). The top-panel shows all mutants while the bottom-panel is focused on the last ones.

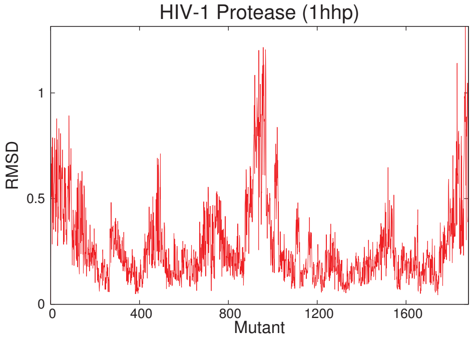


Figure 18: RMSD between the B-factors of the mutant and the ones of the wild-type.

However, as for the RMSD of the B-factors, there is a little change, the maximum being around 1.2.

Interestingly, we can visualise which residue is prone to a lot of change due to a mutation in the protein and which one is prone to induce an overall change in the protein in terms of individual fluctuations (B-factors).

For the latter, for each residue i, we have calculated the average RMSD between B-factors of the wild-type and the ones of the mutant, induced by a mutation on this residue i:

$$I_i := \frac{1}{19}\sum_{k=1}^{19} RMSD(WT, M_k^i)$$

where WT means Wild-Type, M_k^i means the mutant with the k^{th} mutation on residue i and the RMSD between B-factors is

$$RMSD(WT, M_k^i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(B_{WT}(j) - B_{M_k^i}(j))^2}$$

with N being the number of residues.

For the former, we have calculated the RMSD over all possible mutants between the B-factor in the wild-type and the one in the mutant:

$$S_i := \sqrt{\frac{1}{N_M}\sum_{j=1}^{N_M}(B_{WT}(i) - B_{M_j}(i))^2}$$

where N_M is the number of possible mutants (for 1hhp, 99×19) and M_j is the j^{th} mutant.

For each residue, we obtained two numbers and we have visualised HIV-1 protease, 1hhp, with residues coloured according to these numbers. According to

I_i in Figure 19 and according to S_i in Figure 20.
Figure 19 tells us that mutations inducing the most overall changes are located in the isolated loop and at the extremities. In both parts, residues are less connected than the others in the theoretical network. Figure 20 tells us that the residues the most prone to change after a mutation where-ever in the protein are also located in the isolated loop and at the extremities. If we combine both Figures 19 and 20, we can see that mutations on flexible parts induce the larger changes in the protein but these changes are undergone by these same flexible parts.

Even if the sdANM is not suited enough to link a theoretical study on mutations to experimental works, it is still interesting to see that the flexible parts again emerge from our study. Here, the mutation is modelled as changing the stiffness constant of springs related to a single residue (the one which has mutated), then we look at the deviation of the obtained mutant to the wild-type. Finally, it is like comparing two models as we did it in the preceding subsections. Thus, the problem of flexible parts is not surprising. Using single-mutations, we can directly see where are the most sensitive parts and that supports what have been found above. The poor sensitivity against single-mutations is not incoherent since deleterious mutations are pretty rare in cells. And if we think the tuning of spring constants as mutations, it is also coherent that the overall protein is robust against it. Proteins had to adopt robust shapes across the evolution to be stable enough otherwise they would not be able to carry out its current functions. This robustness is theoretically reflected in the network connectivity. However, some mutations have been shown to be deleterious because the structure of the protein is disrupted and cannot works correctly. Then, we can again suspect flexible parts to be prone to deleterious mutations.

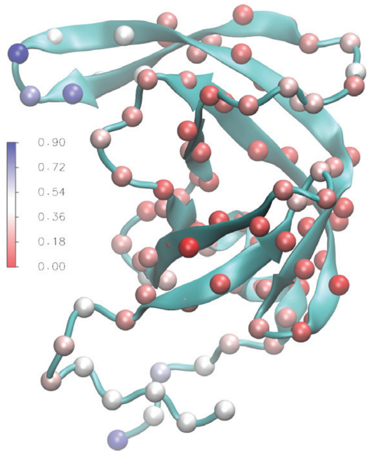


Figure 19: A Visual representation of HIV-1 protease 1hhp where each residue is coloured according to the average RMSD induced by a mutation on it.

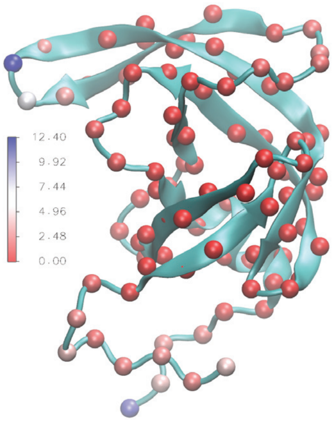


Figure 20: A Visual representation of HIV-1 protease 1hhp where residues are coloured according to the root mean square deviation of the B-factor deviation from the wild-type B-factor across all possible single-mutation mutants.

F Discussion

In this study, we tried to understand how the addition of the sequence specificity to the modelling improves the accuracy of the model. In particular, we have investigated individual proteins and looked at the individual features of their residues. Because in preceding studies the improvements were seen at a protein scale using a correlation coefficient, the origin of these improvements were not clear. It appears that the new sequence specific models are closer to classical models than we expected it. Whatever the measuring method we use (B-factor, overlaps, etc), the results are quite robust against the addition of the sequence specificity. When there is a difference, it involves some poorly connected residues which bias the overall interpretation we could make if looking only at protein-scale measures. Indeed, the estimations of individual fluctuations of such residues are incoherently enhanced or inhibited and are still badly modelled. Although it can improve the correlation coefficient, the improvement is not qualitative and especially not systematic. We looked at the averaged estimation error of individual fluctuations according to some individual features and have not found any systematic improvement of sequence-specific models for any of the considered feature. However, the flexible parts are very often over-estimated whatever the model. Then, this is already a major drawback of classical ANMs which are not resolved in sequence specific models, it is even further enhanced by sdANM due to the softening of long-range interactions.
As explained in materials and methods, the X-ray crystallography method inhibits the fluctuations of residues located in flexible part like the surface. Although mainly used in this study and widely available on the Protein Data Bank, X-ray crystallography structures suffer from the drawback of over-estimation and are not the best suitable structures to look at the individual fluctuations. However, it is not enough to explain the over-estimations since they have been also observed in NMR structures like ubiquitin (pdb id: 1xqq). In our examples, both the amplitude and the number of over-estimations are globally smaller for NMR structures than for X-ray crystallographic structures but it should be

noted that NMR can be used only for small proteins. Then, because NMR structures are generally smaller, a cut-off distance of at least 10Å is already enough to damp the flexible parts where it is not for large proteins.

IV Parameter-free heterogeneous networks

In the preceding section, we have seen that the addition of the sequence specificity in the network has not brought valuable improvements. That leads us to think that the tuning of spring constants may be not as well as we would expect it. However, several studies have developed parameter-free models where the spring constants are determined according to a decreasing function of the distance of the pair [21, 22, 17, 19]. The model is parameter free in the sense that the cut-off is either imposed [22] or totally removed by putting a spring to every pair of residues with a fitted constant [21]. Besides the improvement shown is generally not very large (~ 0.1 of enhancement for the average PCC), the new model is compared to ANM having a cut-off distance that the community "often uses". At first view, it makes sense since the ANM with large cut-off distances are discarded in the community[3]. However, when comparing a heterogeneous all-connected model or a heterogeneous model with a large cut-off distance to the ANM10 or to the ANM13[4], we intrinsically test both the effect of the heterogeneity and of the connectivity. Indeed, the network structure itself is changed and the improvement or any differences observed can be due to the connectivity instead of the heterogeneity. Then, we cannot draw conclusion for the effect of the heterogeneity from such a study. Saying that, it finally makes no sense to discard heterogeneous ANM with large cut-off distances or "all-connected" because they are "never used" by the community. If the new heterogeneous model works as large-cut-off-distance or all-connected homogeneous ANMs, then the conclusion is not that the heterogeneity improves the modelling but increasing the cut-off distance improves the modelling. That is exactly what is shown for the sequence-specific models in Table 2. What happens if we take cut-off distances even larger than 16Å ? Should we expect improvements of the correlations ?

Indeed, fluctuations in the flexible parts are over-estimated by the previous considered models. Then, when increasing the connectivity of the network, these parts are stiffened which increases the overall correlation (see Figure 21, myosin V and scallop myosin). For proteins which do not have such over-estimated parts, the network becomes too connected and the motions, even the small individual fluctuations, are damped (see Figure 21, maltodextrin binding protein and HCV helicase). In particular, the transition could not be studied.

That is something important because parameter-free networks have such a high connectivity and potentially the same draw-back. As in the sANMs and the sdANM, the difference may reside on flexible parts.

A Models

We will compare some well known parameter-free models to the classical homogeneous ANM with and without a cut-off distance. The parameter-free homo-

[3] for reasons we will see below

[4] which are homogeneous since all springs have the same constant

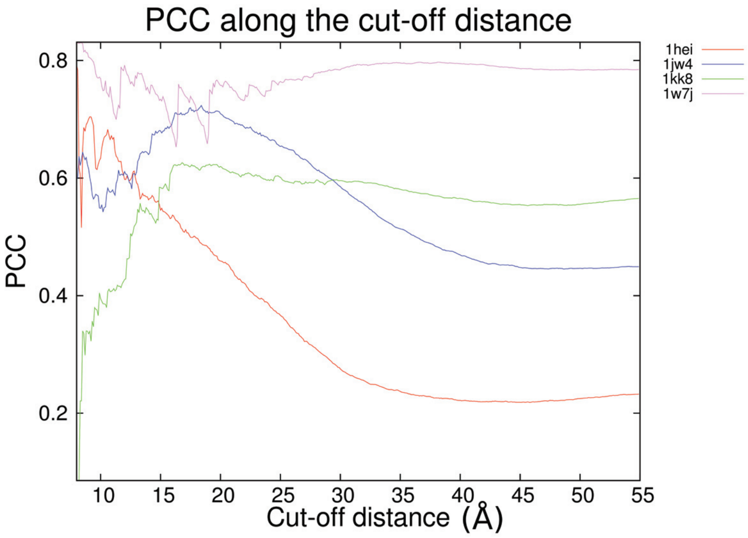


Figure 21: The Pearson correlation coefficient plotted according to the cut-off distance for four proteins: HCV helicase (1hei), maltodextrin binding protein (1jw4), scallop myosin (1kk8) and myosin V (1w7j).

geneous model will be called ANM-AC[5].
There are three major free parameter distance dependent models:

- the first is called dANM, the spring constants are inversely proportional to the square of the distance [21]:

$$k(i,j) = \frac{1}{d_{i,j}^2}$$

where $d_{i,j}$ is the distance between residues i and j in the equilibrium structure.

- the second is called ExpANM, the spring constants decay exponentially as the distance is large [17]:

$$k(i,j) = exp\left(-\frac{d_{i,j}^2}{c^2}\right)$$

where c is a coefficient shown to have the best agreement for $c = 3$Å [17] or for $c = 7$Å [19]. In the doubt, we will take both.

- the last one is called hANM, the spring constants follow an inverse of a power 6 [19]:

$$k(i,j) = (205.5 \times d_{i,j} - 571.2)\mathbb{1}_{d_{i,j}\leqslant 4} + 305.9.10^3 \frac{1}{d_{i,j}^6}\mathbb{1}_{d_{i,j}>4}$$

[5] AC for All Connected

Typically, the short range interactions are stiffened.

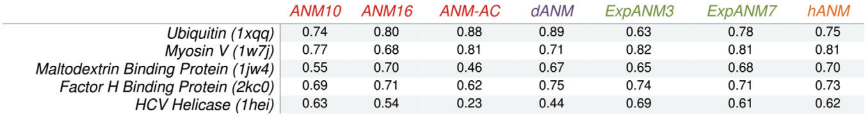
B B-factor patterns

We first look at some representative protein examples. The aim is to determine where the improvement are located and especially to determine whether the improvements are coherent or biased by flexible parts. A part of the results are available in Table 5 and the full part in Table 12 in appendix.

As for the overall correlation, for some proteins, it is improved by parameter-free heterogeneous models, for some proteins not. Sometimes, dANM performs better than ANM16 while hANM performs worse (e.g. ubiquitin), sometimes, it is the reverse (e.g. HCV helicase) and so on. The heterogeneous networks work completely differently although it is the same kind of heterogeneity i.e. a decreasing dependency on the distance. If the adding of this distance-dependency were valuable, then all these models would work at least similarly and improve or decline the ANM16 (or other ANMs) for the same proteins.

As for individual proteins, ubiquitin is improved by dANM (0.80 for ANM16 against 0.89 for dANM) but also by ANM-AC (0.88) suggesting an improvement due to the connectivity. Indeed, the tail becomes damped by both dANM and ANM-AC (see Figure 22A). ExpANM7 and hANM still over-estimate it (even more than ANM16). Interestingly, the "ranking" of models at the tail (hANM>ExpANM7>ANM16>dANM>ANM-AC) is the inverse of the one on the rest of the protein (ANM-AC>dANM>ANM16>ExpANM7>hANM). In particular, the B-factor pattern curves corresponding to ANM-AC and dANM tend to be flat. That suggests a loose of accuracy in enough-connected parts in ANM-AC and in dANM compensated by the damping of the tail. The case of myosin V is similar to the one of ubiquitin (see Figure 22B). The poorly connected part is not located at the tail but within the protein. As for maltodextrin binding protein and Factor H binding protein, both are badly modelled by ANM-AC but none of the heterogeneous parameter-free model shows a large improvement to ANM16 (see Table 6). The B-factor patterns of maltodextrin are globally similar (Figure 22C). The ones of factor H differ a little bit for over-estimated part essentially (Figure 22D). dANM performs better on them at the expense of some other parts. The improvement of dANM on flexible parts is then, although not fully, compensated by the decline in some other parts compared to ANM16. HCV helicase is the typical example for which the strong connectivity inhibits its motions. The cut-off distance of 8Å has been shown to be the best for this protein (PCC is 0.79); above this value, the network is too strongly connected. Heterogeneous networks still inhibit the correlation compared to ANM8 and do not seem able to recover a similar behaviour to ANM8 for HCV helicase.

Again, flexible parts look suspicious and seem to play the largest role in the improvements seen in the overall correlation for parameter-free heterogeneous models. In particular, the improvements do not come from the coherent and physical property-based tuning of spring constants. Ubiquitin is a great example of that. dANM having a spring constant distribution narrower than the other heterogeneous networks, it can damp more efficiently the flexible parts. But inevitably, it has consequences on the other parts. The improvement resumes to the damping on the flexible part disregarding the physical meaning brought by spring constants. ANM-AC which is able to damp efficiently the flexible parts, work as well although the spring constants have not any physical meaning.



	ANM10	ANM16	ANM-AC	dANM	ExpANM3	ExpANM7	hANM
Ubiquitin (1xqq)	0.74	0.80	0.88	0.89	0.63	0.78	0.75
Myosin V (1w7j)	0.77	0.68	0.81	0.71	0.82	0.81	0.81
Maltodextrin Binding Protein (1jw4)	0.55	0.70	0.46	0.67	0.65	0.68	0.70
Factor H Binding Protein (2kc0)	0.69	0.71	0.62	0.75	0.74	0.71	0.73
HCV Helicase (1hei)	0.63	0.54	0.23	0.44	0.69	0.61	0.62

Table 5: Pearson correlation coefficient for 5 proteins obtained with classical homogeneous ANM and with heterogeneous ANM.

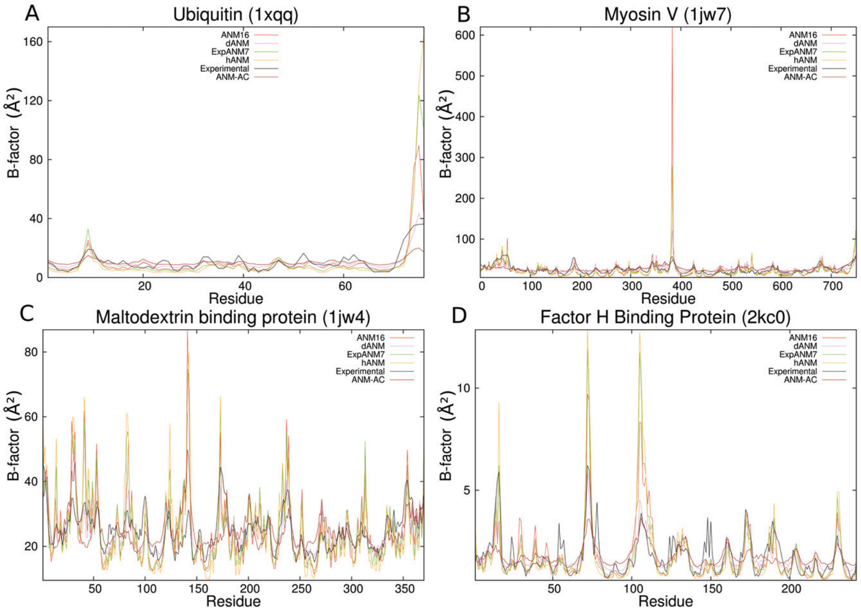


Figure 22: B-factor patterns for some proteins. The red curve is obtained with ANM16, the violet one with dANM, the green one with ExpANM7, the orange one with hANM and the dark red one with ANM-AC. The black curve represents the experimental B-factors.

C Statistical analysis

To enforce the observations of the preceding subsection, we analyse the B-factors over a lot of protein structures. For the parameter-free models, we have taken the same structures accepted by the algorithm for ANM16 since we mainly make comparisons with this model.

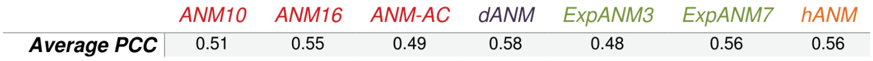
As for the averaged Pearson correlation coefficient, the statistical data indicates an improvement of dANM, ExpANM7 and hANM compared to ANM-AC (see Table 6). Then, ANM-AC does not seem robust against spring constant tuning and its performances can be improved. This is quite different from what we have observed for cut-off dependent ANMs. However, the heterogeneous models do not show a great improvement compared to ANM16 (0.01 for ExpANM7 and hANM, 0.03 for dANM).

The most surprising things come from the B-factor ratio (predicted B-factor over experimental B-factor) of the proteins of the systematic study (Figure 23). All distributions are distributed around 1 (1 being the ideal case) with a more or less wide dispersion. As expected, ANM-AC reduces the over-estimation of poorly connected parts and globally the dispersion is smaller as compared to ANM16 (Figure 23A). As drawbacks, flexible parts are more readily under-estimated and highly connected parts are slightly over-estimated. What is surprising and not observed in the overall correlation, is the quasi-identical results given by dANM (Figure 23B). The flexible parts are slightly less damped but globally, the features of the distribution are the same as compared to ANM-AC.

Also striking is the similarity between ExpANM7, hANM and ANM16. ExpANM7 performs exactly the same way as ANM16 (Figure 23C). The distribution of B-factors of hANM is slightly narrower than the one of ANM16 but the shape is the same (Figure 23D).

Several things emerge from this statistical analysis. First, the averaged correlation coefficient is not improved so much by heterogeneous models as compared to ANM16, at least, not as we would expect it from a model for which spring constants have been coherently tuned. As suspected, for dANM, which is the model performing the best in regard of the correlation, it is a matter of flexible parts which biases the correlation. As for proof, it has the same features and then drawbacks as the ANM-AC: gain of accuracy in flexible parts, loss of accuracy in non-flexible parts. The other heterogeneous models do not show any striking differences.

Last observation is that adding heterogeneity in ANM-AC changes the results. In that sense, ANM-AC is sensitive to the tuning of spring constants. However, the heterogeneous models, although having an all-connected network, seem to behave like a cut-off dependent homogeneous model.



	ANM10	ANM16	ANM-AC	dANM	ExpANM3	ExpANM7	hANM
Average PCC	0.51	0.55	0.49	0.58	0.48	0.56	0.56

Table 6: Summary table of results obtained over 2040 X-ray crystallographic protein structures for each model.

D Little value brought by the heterogeneity

In the preceding subsection, we have observed that the heterogeneous parameter-free models do not work as the homogeneous version. In that sense, the ANM-AC is not robust against spring constant tuning. On the other hand, these heterogeneous models, although parameter-free, seem to behave like a homogeneous ANM with a certain cut-off distance. It seems the parameter-free models have an imposed cut-off distance because of the very small values of long-range interactions. The function determining the values of spring constants (the inverse square function for dANM, the decreasing exponential for ExpANMs, etc) depends solely on the distance of the pair like the Heaviside function used for

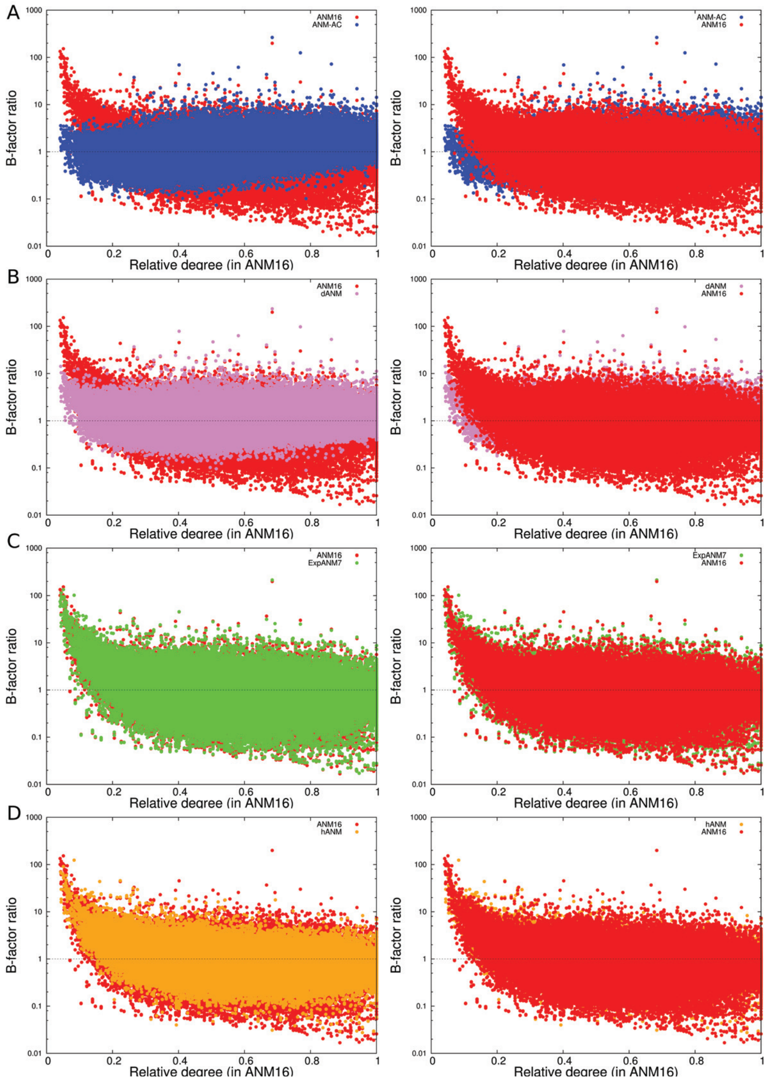


Figure 23: B-factor ratio (predicted B-factor over experimental B-factor) for parameter-free homogeneous (ANM-AC) and heterogeneous (dANM, ExpANM7 and hANM) models alongside the ANM16 according to the relative degree defined in the preceding subsection for ANM16. Note that for parameter-free models, the relative degree is the same for all residues. For every graph: Red points represent ANM16 and the y-axis is log-scaled. A: blue points represent ANM-AC, B: violet points represent dANM, C: green points represent ExpANM7, D: orange points represent hANM. Because the two models displayed on the same graph overlap, the right graphics are the same as the ones but the order of plotting has been shifted.

the homogeneous ANM (what we call “cut-off”). From a certain distance, this function gives values small enough that the resulting link has a weak effect. This distance is the same for all proteins since the function is protein-independent.

Then a kind of cut-off distance is imposed in the heterogeneous parameter-free models which makes the analysis similar to the ones with a classical ANM with the same cut-off distance.

That could be the reason why we observed robustness when adding the sequence specificity but not when considering all-connected heterogeneous networks. The heterogeneity is such that it makes the model works as having a cut-off distance. It would then be more judicious to compare the parameter-free heterogeneous models with homogeneous cut-off dependent ANM rather than to the ANM-AC. For dANM and ExpANM7, imposing a cut-off distance at 50Å and at 16Å respectively, they fit almost perfectly the models without cut-off distance (see Figure 24 A and B, left panel) highlighting the very poor effect of the links for which the distance is larger than this imposed cut-off distance. The ANM with the same cut-off (Figure 24 right panel) shows quite well agreement to the parameter-free version of dANM, ExpANM7 and hANM. The ANM is still slightly less efficient but it is interesting to see that the homogeneous model without any logical way of setting the spring constants can perform almost as well as heterogeneous models for which the spring constants have been chosen coherently.

The hANM is a little bit different from the two other cases. We have not found a cut-off distance for which the cut-off dependent hANM and the cut-off dependent ANM matches perfectly the parameter-free hANM. The cut-off 20Å gives good enough approximations (Figure 24C).

These cut-off distances have been chosen intuitively. We tried to determine them in a coherent way using the sequence-specific model examples. For these models, the range of spring constant values has a dispersion of five orders of magnitude (from 0.001 to 43.52). So, we determined former cut-off distances such as the range of spring constant values has a five-order of magnitude dispersion. However, it does not lead to satisfying results, the cut-off distances being too large.

E A short discussion on the NMR case

It is also interesting to see what happens for NMR structures. For that, we take the same 132 solution NMR protein structures selected in the preceding section. The averaged errors are larger for dANM and ExpANM7 than for ANM10 for almost all chemical properties and structures (see Table 7). As for hANM, there are some slight improvements.

The distance-dependent models are typically all connected. However, proteins which can be determined by NMR are generally small. In our dataset, the mean number of residues is 107 which is quite small. That means, a too much connected network could inhibit the individual fluctuations. This is what we see for ANM-AC and in a smaller extend for the distance-dependent models for which some long-range interactions are softer than in the ANM-AC.

F Maximal overlap

Another possible drawback of a strong connectivity is the inhibition of motions of the protein. For that, we can look at the overlaps, in particular, the maximum one for a given transition. The maximum overlap is clearly decreased for all protein transitions using ANM-AC (Table 8). As for the heterogeneous networks,

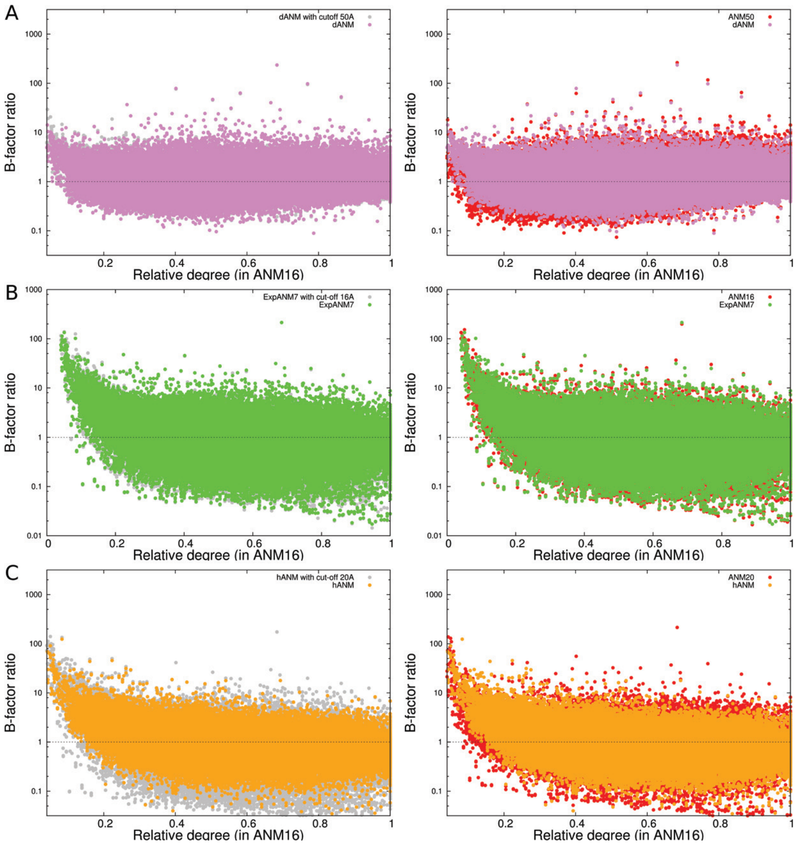


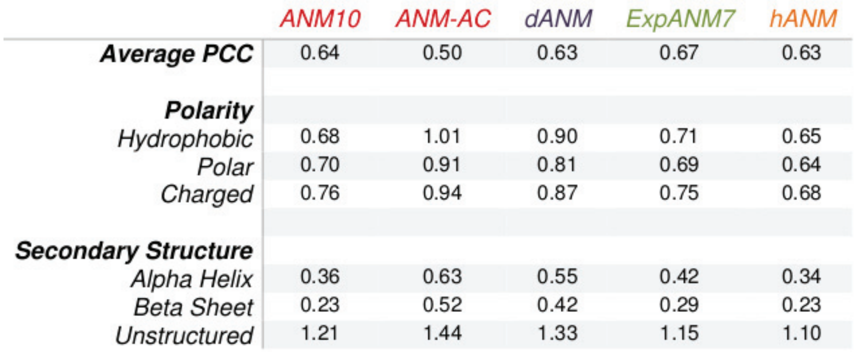
Figure 24: Left panel: B-factor ratio (predicted B-factor over experimental B-factor) for parameter-free heterogeneous dANM, ExpANM7 and hANM models alongside their version with a cut-off distance (50Å for dANM, 16Å for ExpANM7 and 20Å for hANM). Right panel: the cut-off version is replaced by the corresponding ANM. The relative degrees of ANM16 have been conserved to have a comparison. The y-axis is log-scaled.

they show similar results to ANMs with a cut-off distance except dANM which shows slightly smaller values. For the latter model, it is confirmed by a study made several years ago [21] done over 170 proteins. In this study, dANM shows globally smaller max overlaps than ANM13. Indeed, the percentage of overlaps larger than 0.6 has dropped from 40% in ANM13 to 20% in dANM while the percentage of overlaps smaller than 0.3 has increased from 34% in ANM13 to 49% in dANM. That shows a tendency of the overlaps to drop in dANM.

These results support what has been found above: the heterogeneous parameter-free models behave similarly to cut-off dependent homogeneous models.

G The limits of linearity

Finally, it is interesting to have a look at relaxation patterns which give us valuable information about the validity of the linear assumption (Figure 25).



	ANM10	ANM-AC	dANM	ExpANM7	hANM
Average PCC	0.64	0.50	0.63	0.67	0.63
Polarity					
Hydrophobic	0.68	1.01	0.90	0.71	0.65
Polar	0.70	0.91	0.81	0.69	0.64
Charged	0.76	0.94	0.87	0.75	0.68
Secondary Structure					
Alpha Helix	0.36	0.63	0.55	0.42	0.34
Beta Sheet	0.23	0.52	0.42	0.29	0.23
Unstructured	1.21	1.44	1.33	1.15	1.10

Table 7: Summary table of results obtained over 132 solution NMR structure proteins. For the polarities and the secondary structures, the averaged MSF estimation errors are computed.

Myosin V was chosen as an example which is a motor protein for which the linearity is valid for large motions. If the ExpANM7 and hANM reproduce well the relaxations found for ANM16 (although not improved), dANM fails to do it. Again, the strong connectivity of dANM plays tricks on it and reduces the range on which the linear approximation can be applied. The large motions of myosin V cannot be retrieved by dANM although the correlation of individual fluctuations is better than ANM16. That is a quite strong drawback because although it is supposed to be better under the assumption of linearity, it is valid only at very close to the equilibrium conformation.

H Discussion

The Pearson correlation coefficient is a tool to estimate the accuracy of a model. However, it is a tool at the protein-level disregarding the details of the protein. The heterogeneous parameter-free models give better correlation than ANM with widely used cut-off distances. However, looking at the residue-level, the improvements suffer from the same biases as the ones observed for sequence-specific models: the flexible parts. But more than that, the information which has been added (the distance dependence) is not reflected in the B-factor patterns. Indeed, the improvement consists in stiffening the flexible parts at the expenses of the non-flexible ones. This is what happened for ubiquitin: the patterns shown by ANM-AC and dANM are almost flat reflecting no coherence in the prediction of residue fluctuations except the one of damping the poorly connected residues. Despite this flat pattern, the correlation is better because the flexible part is damped. This is even more striking when looking at the individual B-factor prediction over more than 2,000 proteins. Either there is almost no difference as compared to ANM16 or the flexible parts are damped at the expenses of non-flexible ones. Even more, we can find cut-off distances for which the results of the homogeneous models match well the results of heterogeneous parameter-free models. That is a drawback of these latter models in two ways: first, it means that the tuning of the spring constants has a very poor effect and second, some of them resemble homogeneous ANM having a large cut-off distance, larger than the ones often used within the community,

	ANM10	*ANM16*	*ANM-AC*	*dANM*	*ExpANM3*	*ExpANM7*	*hANM*
Myosin V							
1w7j to 1oe9	0.63	0.54	0.16	0.47	0.61	0.60	0.64
1oe9 to 1w7j	0.55	0.62	0.16	0.47	0.55	0.60	0.61
HIV-1 Protease							
1hhp to 1ajx	0.71	0.67	0.20	0.44	0.55	0.67	0.71
1ajx to 1hhp	0.44	0.31	0.25	0.26	0.51	0.46	0.42
Aspartate Aminotransferase							
9aat to 1ama	0.57	0.64	0.15	0.36	0.51	0.51	0.51
1ama to 9aat	0.55	0.57	0.26	0.38	0.76	0.67	0.63
Enolase							
3enl to 5enl	0.23	0.21	0.10	0.18	0.22	0.22	0.24
5enl to 3enl	0.23	0.23	0.10	0.18	0.24	0.22	0.27
Maltodextrin Binding Protein							
1anf to 1jw4	0.90	0.80	0.34	0.70	0.87	0.89	0.79
1jw4 to 1anf	0.78	0.79	0.33	0.72	0.79	0.74	0.87
Adenylate Kinase							
1aky to 2ak3	0.38	0.36	0.28	0.31	0.32	0.40	0.37
2ak3 to 1aky	0.27	0.31	0.14	0.22	0.31	0.31	0.29
Annexin V							
1avr to 1avh	0.41	0.33	0.24	0.27	0.40	0.33	0.33
1avh to 1avr	0.33	0.33	0.12	0.27	0.48	0.38	0.41
Thymidylate Synthase							
2tsc to 3tms	0.44	0.24	0.30	0.21	0.40	0.30	0.37
3tms to 2tsc	0.46	0.28	0.28	0.20	0.63	0.39	0.40
Scallop Myosin							
1kk8 to 1kk7	0.85	0.75	0.37	0.56	0.86	0.80	0.73
1kk7 to 1kk8	0.70	0.74	0.38	0.53	0.74	0.73	0.72
F1-ATPase							
1h8h to 1h8e	0.55	0.49	0.38	0.66	0.52	0.51	0.50
1h8e to 1h8h	0.54	0.53	0.35	0.63	0.57	0.53	0.50
Kinesin Motor Domain							
1mkj to 1bg2	0.30	0.24	0.14	0.26	0.27	0.25	0.35
1bg2 to 1mkj	0.23	0.18	0.15	0.19	0.21	0.18	0.21
Penicillin Binding Protein							
1vqq to 4dki	0.33	0.37	0.14	0.24	0.32	0.37	0.39
4dki to 1vqq	0.33	0.30	0.16	0.24	0.32	0.36	0.37
KIF1A Motor Domain							
1i6i to 1i5s	0.35	0.24	0.15	0.29	0.24	0.37	0.25
1i5s to 1i6i	0.28	0.20	0.16	0.29	0.29	0.28	0.28

Table 8: Maximum overlap for parameter-free models alongside with classical ANMs with a cut-off distance.

and then have drawbacks that the community wants to avoid. Among these drawbacks, there are especially the inhibitions of motions and the narrowing of the range of validity of the linear assumption.
The distance dependence of spring constants, associated to an all-connected network essentially imposed a kind of cut-off distance and decrease the connectivity.

V Randomisation of spring constants

A Method

In the two preceding sections, we have seen that the modelling is quite poorly sensitive to the addition of valuable biologically and chemically related information through the value of spring constants. As an extreme and brute-force

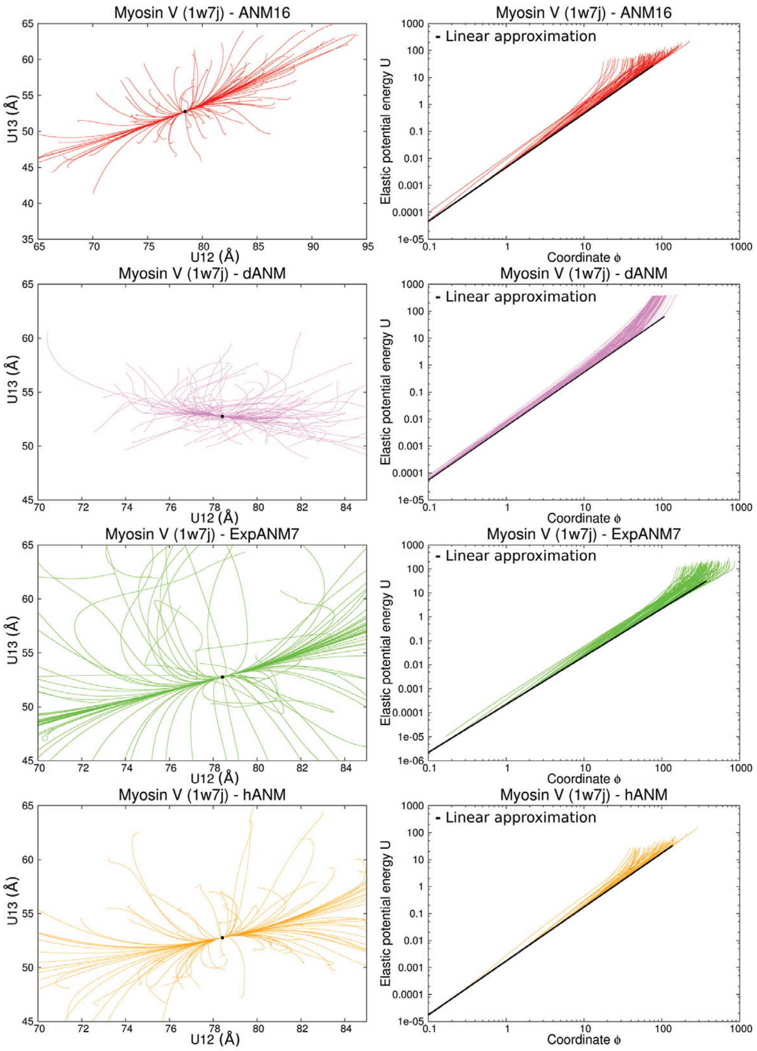


Figure 25: Relaxation trajectories for myosin V (pdb id: 1w7j) with ANM16, dANM, ExpANM7 and hANM. The labels are the same as the one used above. For ANM16: $F_{ini} = 50$ and $T_{ini} = 1000$ while $F_{ini} = 25$ and $T_{ini} = 1000$ for the other models.

way to test to which extent it is true, we randomised the spring constant values. The network connectivity is determined using a cut-off distance then all spring constant values are chosen randomly using the uniform distribution between 0.001 and 1 ($\mathcal{U}([0.001, 1])$). This range is large enough to allow a lot of diverse values but not too large to make consider a link as vanishing. There is no other rule followed by the distribution. It means that the random network generated by this method is not biologically or physically relevant (other than the shape) and can have very stiff long-range interactions while having very soft short-range interactions. The obtained random networks will be compared to the corresponding homogeneous ANM with the same cut-off distance.

There are two aims for this study. The first one is to determine to which extent the network is not sensitive to the changes in the spring constant values. The second one is to determine the features the best performing networks could have. Indeed, producing thousands of random network for the same protein with the same connectivity, some networks would perform better than some others. We want to know why and where they perform better.

B Sensitivity to randomisation

Because this randomisation does not follow any rules and generally lead to non-coherent situations, we do not expect an accurate value when averaging the Pearson correlation coefficient over thousands of random networks. A priori, it is easier to construct bad models than good ones when randomly determined.
Results obtained over either 10,000 or 100,000 random networks are displayed in part in Table 9 and in full part in Table 13 in the Appendix. They astonishingly resemble the ones for the classical ANM. Even when these last ones perform very well, the randomisation is able to reproduce such a good performance in average (see ubiquitin, pdb id: 1xqq) although the PCC is always slightly smaller.
In parenthesis, is displayed the root mean square deviation (RMSD) of the PCCs across the random networks i.e.

$$RMSD = \sqrt{\frac{1}{\sum\limits_{i=1}^{M}(M-i)} \sum_{i=1}^{M}\sum_{j>i}^{M}(PCC_i - PCC_j)^2}$$

where M is the number of random networks.

Most cases show a RMSD value of around 0.10 meaning that taking two of the random networks, in average, the difference in their PCCs is around 0.10 which highlights a kind of poor sensitivity of the theoretical network to spring constant values. We may note that this difference may reside solely on flexible parts or even to one or two residues which are badly modelled. For some cases, the RMSD is quite large (of the order of 0.3) highlighting a sensibility to spring constant values for the protein and the cut-off considered. Essentially small proteins show such large RMSD (for examples: A1-Type ACP domain, VAT-N, factor H binding protein). Interestingly, the sensitivity of spring constants is not specifically related to the protein but rather to the network itself. Indeed, considering the same protein, changing the connectivity of the network can drastically increase or decrease the sensitivity. For example, the RMSD for Random10 of VAT-N is 0.06 and increases to 0.21 for Random16; the RMSD for Random10 of factor H binding protein is 0.16 and decreases to 0.05 for Random16. To have a visualisation of the dispersion, we have plotted the distribution of the PCC over the 100,000 random networks for the NMR protein structures using the Random16 model. The distributions are almost Gaussian. Some are quite thin (for examples, 2kc0, 6qeb, ...) and some others are slightly wider (6h0j, 1cz4, ...). Generally, the left-tail is wider than the right-tail. That makes sense since it is easier to decrease the accuracy rather than to increase it when choosing the spring constants randomly without any coherent rules.
These results suggest that the connectivity itself of the network contributes for

most of the accuracy, then the tuning of spring constants ”deviates” the accuracy toward the improvement or the decline within a limited range. These results raise an interesting question: Is there a limit to the improvement by coherently choosing spring constants?
To try to answer this question, in the next subsection, we have investigated the 1% of networks giving the best PCCs.

	ANM10	ANM13	ANM16	Random10 (RMSD)	Random13 (RMSD)	Random16 (RMSD)
Myosin V (1w7j)	0.77	0.77	0.68	0.74 (0.07)	0.73 (0.07)	0.64 (0.09)
Adenylate Kinase (1aky)	0.61	0.56	0.61	0.58 (0.12)	0.54 (0.1)	0.60 (0.08)
Scallop Myosin (1kk7)	0.14	0.31	0.38	0.14 (0.05)	0.28 (0.09)	0.36 (0.04)
Human Kinesin Motor Domain (1mkj)	0.41	0.69	0.70	0.39 (0.09)	0.68 (0.01)	0.69 (0.02)
HCV Helicase (1hei)	0.63	0.59	0.54	0.61 (0.04)	0.58 (0.02)	0.53 (0.02)
Ubiquitin (1xqq)	0.74	0.73	0.80	0.71 (0.18)	0.72 (0.18)	0.78 (0.14)
Type IV Pilin PILE1 (6i2o)	0.87	0.88	0.85	0.84 (0.15)	0.86 (0.12)	0.84 (0.09)
A1-Type ACP Domain (6h0j)	0.80	0.58	0.57	0.76 (0.29)	0.56 (0.26)	0.55 (0.19)
VAT-N (1cz4)	0.37	0.32	0.60	0.36 (0.06)	0.32 (0.05)	0.59 (0.21)
Factor H Binding Protein (2kc0)	0.69	0.69	0.71	0.66 (0.16)	0.68 (0.08)	0.69 (0.05)

Table 9: Table of Pearson correlation coefficient for selected proteins. The table of the full set of proteins is available in appendix (Table 13). For the random models, the PCC has been averaged over either 10,000 or 100,000 networks. The RMSD through the networks is calculated in parenthesis.

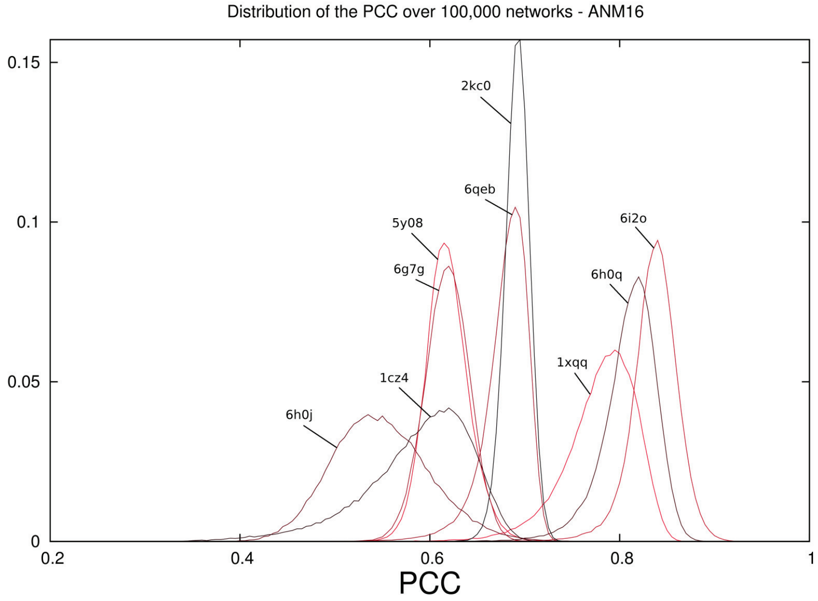


Figure 26: Distribution of the Pearson correlation coefficient over 100,000 networks for the 9 NMR protein structures we have considered in this study.

C Heterogeneity of the best networks

In this subsection, we have investigated some cases in depth. In particular, we have looked at the top 1% of networks which gives the best correlations. We would like to know where are the improvements and how coherent are the distributions of spring constants on this top 1%.

The first example is ubiquitin. It has a tail for which fluctuations are over-estimated by the ANMs. The top 1% networks have PCCs ranging from 0.846 to 0.874 for Random16 which is much larger than for ANM16 (0.80). The heat map of spring constants averaged over the top 1% shows that the large majority of links has roughly the same spring constant (white color in the heat map in Figure 27A). The links having larger or smaller spring constants involve the last 5 residues i.e. the tail. The B-factor pattern is improved almost only at this location (Figure 27A) which is even more improved by the best network (the one with the maximum correlation). The distribution of spring constants do not have any physical coherence, it just stiffens or weakens some links involving the tail and improves its modelling.

The heat map of the top 1% of adenylate kinase for Random10 seems quite homogeneous (Figure 27B). Only two links, each located at one tail, are significantly heterogeneous. The B-factor pattern of the top 1%, as well as the one of the best network, shows improvement essentially at the tails. Otherwise, there are some small improvements for flexible parts but the top 1% performs mainly like ANM10. Again, the distribution of spring constants of the top 1% best networks does not follow any logic but specifically improves the modelling of tails. The correlations of the top 1% ranges from 0.647 to 0.704.

The average correlation for HIV-1 protease over 100,000 networks having a connectivity determined by a cut-off distance of 13Å is 0.18 (against 0.20 for ANM13). The top 1% has correlations ranging from 0.285 to 0.336. As for the two preceding cases, the heterogeneity is mainly at both tails and at the flexible part representing the flap (Figure 27C). Interestingly, the B-factor patterns of the top 1% and the best network are not as good as expected compared to the improvement in term of Pearson correlation coefficient. Indeed, the stiffness of spring constants in the flap-structure and one of the tails are weakened resulting in larger over-estimations which are compensated by the stiffening of the second tail. Despite the improvement of the correlation, the B-factors themselves are not improved so much.

As for the other proteins, the situation is generally similar: the average spring constants of links of the top 1% are mainly the same except for some links involving either the tails or some flexible parts inducing some improvements in the parts concerned. However, the B-factor pattern is not always as much improved as the correlation is and situations like the one of HIV-1 protease arise.

An additional situation emerged from the analysis: the situation where all spring constants have different values without any coherent tendency. That is especially true for large proteins with a large cut-off like maltodextrin binding protein for random networks with cut-off distance 50Å (Figure 28A) or scallop myosin for random networks where everything is connected (Figure 28B).

D Discussion

In this section, we randomised the spring constants. The randomisation is such that the spring constants are not coherently distributed. Their value is just uniformly chosen between 0.001 and 1. This ensures that all links have a large enough spring constant value in order to be not negligible. Under 0.001, we consider that the link would be negligible and these values are not taken into

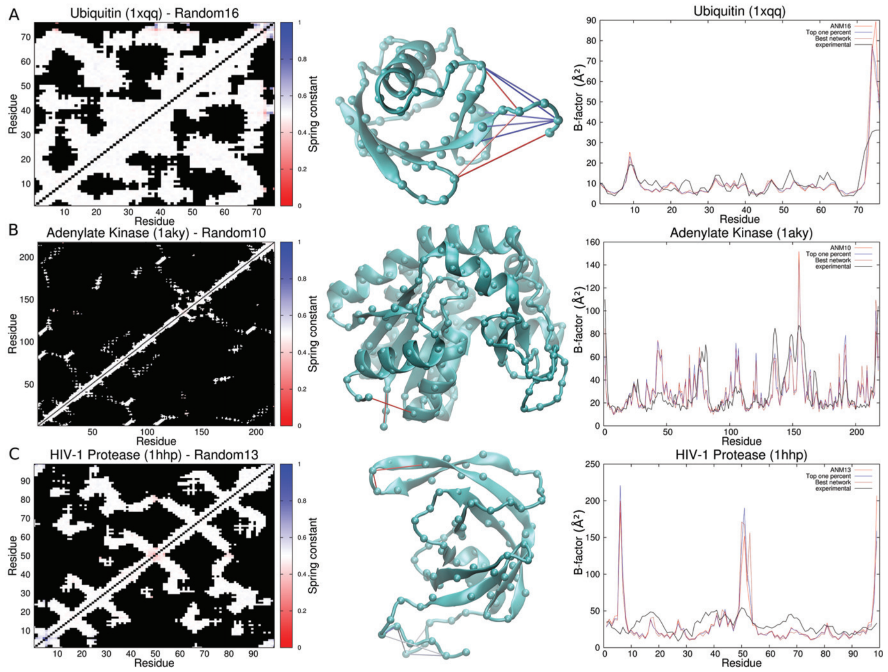


Figure 27: Left: Heat map of the values of spring constants averaged over the 1% networks showing the best correlations. Middle: visual representation of the protein where links with spring constant values showing a difference of more than 25% compared to the averaged one are displayed. Right: B-factor patterns of ANM, the top 1% where the averaged spring constant values have been used, the best networks and the experiments. A: Ubiquitin (pdb id: 1xqq). B: Adenylate kinase (pdb id: 1aky). C: HIV-1 Protease (pdb id: 1hhp).

account because we wanted to test the robustness of defined networks[6] and then wanted to discard the cases where some links would vanish due to a too small spring constant value. So what we did is to test the robustness of particular networks for which the connectivity has been determined by a cut-off distance and all their links are randomly set. Note that, if we use a larger upper bound than 1, we should increase the lower bound as well since the negligibility is relative to the size of the range.

Averaging the correlation between experiment and theory over 10,000 or 100,000 networks, we found results very close to the homogeneous ANM (Tables 9 and 13 in appendix). Mathematically, it is not trivial. The expected value of the uniform distribution $\mathcal{U}([a,b])$ is $\dfrac{a+b}{2}$ and all spring constants follow the exactly same distribution so the resulting network obtained by the averaging over N random networks converges toward the homogeneous ANM as N increases. However, nothing guarantees that the average of Pearson correlation coefficients converges toward the one of the homogeneous ANM. Indeed, although the forces are linear[7], the diagonalisation process is not, and the eigenvalues of a sum of

[6] defined by the cut-off distance

[7] They have been linearised to carry out the NMA

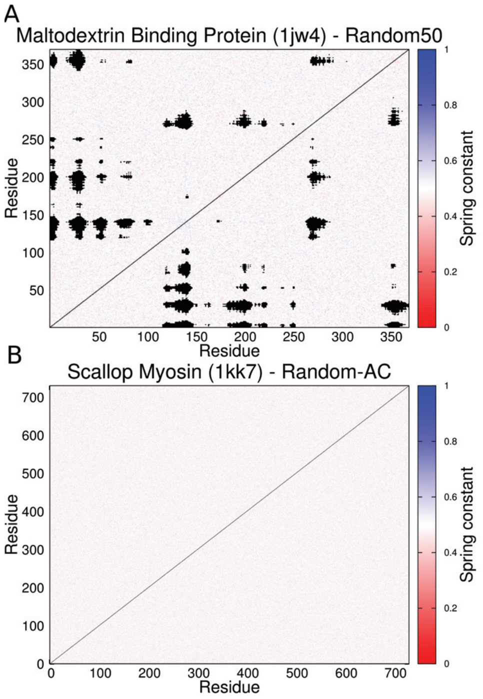


Figure 28: Heat map of the values of spring constants averaged over the 1% networks showing the best correlations. A: Maltodextrin binding protein (pdb id: 1jw4). B: Scallop myosin (pdb id: 1kk7).

matrices do not correspond to the sums of eigenvalues of the matrices. The average correlation itself is not a measure of robustness of the network. We then calculated the RMSD of the correlations across the random networks. Most of the networks we studied show a quite small RMSD (less than 0.1) leading to a thin normal distribution of the correlation (Figure 26). That suggests a poor sensitivity of these networks against spring constant values.

Secondly, we investigated, over the random networks, the ones giving the best correlations. We wanted to determine which trend these networks have and if the same trend is found for many cases. As for the sequence specific models and parameter-free models, the differences are mainly located at the tails and flexible parts. Averaging the spring constants over the top 1% of networks, we see that most of links have the same spring constants which means that these links do not influence the correlation and does not account for the improvement otherwise a tendency would emerge. A tendency emerges for some links involving residues located at the tails or on flexible parts: either they are stiffened or weakened compared to the other links. For some cases, that leads to a better modelling of these parts essentially. For some other cases, it is a little bit more complex: some of these parts are better modelled but some other are worse modelled (see HIV-1 Protease in Figure 27C). That would suggest that the improvement of a part A and of the overall correlation requires the worsening of a part B. We should not forget that these tendencies has been observed for an average over many random networks (1,000 for HIV-1 Protease) and not only

for one single network. It means that the improvement of part A and the worsening of part B are systematic when the networks show a better correlation for a specific protein with a particular cut-off distance.
On the other side, the investigation of the top 1% of some proteins does not show any tendency for any link (Figure 28). Most links seem to play a role in improving the overall correlation otherwise their value would be around the expected value of the uniform distribution. However, more investigations are needed to determine the coherence (if there is) of their value in the improvement of the network.

Mathematically, the current randomisation can be viewed as a perturbation from the homogeneous ANM. In particular, due to the linearity of the force in respect to the spring constant, the Hessian matrix can be decomposed such as:

$$H_{random} = H_{ANM} + P$$

where H_{random} is the Hessian of the randomised matrix, H_{ANM} is the Hessian of the ANM and P the matrix of perturbations associated. A field of mathematics, eigenvalue perturbation theory, studies the eigenvalues of such matrices. In particular, we can give boundaries to the eigenvalues of H_{random} in respect of those of H_{ANM} and P. First, using the Wielandt-Hoffman theorem [81, 82]:

$$\sum_{i=1}^{N} (\lambda_i(H_{random}) - \lambda_i(H_{ANM}))^2 \leq tr(PP^T)$$

where $\lambda_i(H)$ is the i^{th} eigenvalue of H.

An alternative is the weyl's inequality [83]:

$$\forall k \in [\![1, n]\!], \lambda_n(P) \leq \lambda_k(H_{random}) - \lambda_k(H_{ANM}) \leq \lambda_1(P)$$

where eigenvalues are re-ordered in descending order.

The last equality requires to know the eigenvalues of the matrix of perturbations. Although they can be computationally determined, it is intractable theoretically.
The former inequality, although more interesting, does not help us so much. The matrix P being symmetric, the matrix product PP^T reduces to the matrix product of P with itself. Then the trace of such a product corresponds to the sum of all squared elements of the matrix P i.e.

$$\sum_{i=1}^{N} (\lambda_i(H_{random}) - \lambda_i(H_{ANM}))^2 \leq \sum_{i,j=1}^{N} p_{i,j}^2 \tag{8}$$

What we know about the elements of P is that the more links the network has, the more filled the matrix P is. However, it does not guarantee that smaller networks would have a smaller bound in equation (8) since an element $p_{i,j}$ is itself a sum of elastic-force-derivative-like terms which can be either positive or negative. These terms depend on the inverse squared of the distance between two residues.

From the side of random network theory, works are mainly focused on randomising the setting of the connectivity itself rather than the weights of a defined network. That would be also a nice future work to randomise the links themselves.

VI Interface

To carry out all simulations above, we have implemented ourselves the code and developed a user-friendly interface easy to use (see Figure 29). It is divided into two tabs: the first one is dedicated to the normal mode analysis while the second one carries out relaxation trajectories.

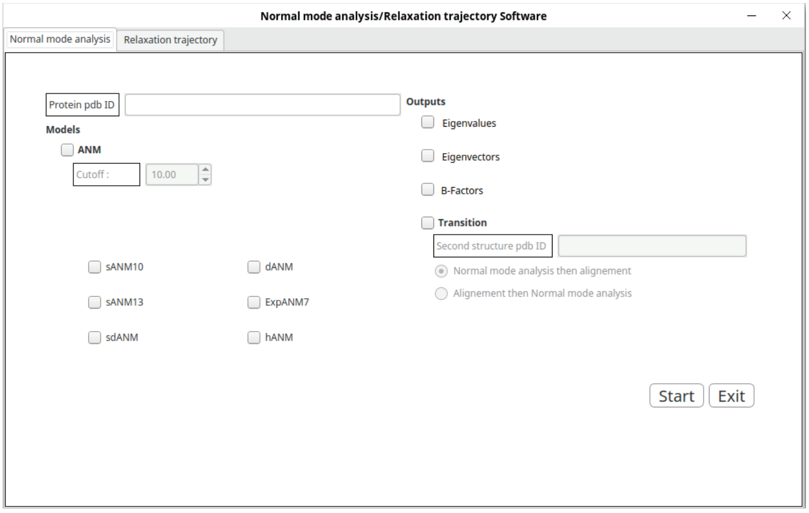


Figure 29: Interface of our software.

A Normal mode analysis

First, the pdb ID of the protein has to be entered in the appropriate box. The algorithm first check the current folder then, if it has not found the file, goes to the protein data bank website. All alpha-carbons are extracted disregarding the number of chains. Then, the file has to be pre-processed by hand in the desired way before. Files other than pdb files are not compatible.
Then we can choose the models:

- ANM is classical model for which the cut-off distance as well as the spring constant can be chosen

- sANM10, sANM13 and sdANM are the sequence specific models

- dANM, ExpANM7 and hANM being the distance heterogeneous models described in section III

It is possible to choose several models at the same time, the code being able to independently carry out the normal mode analysis for each model and gives model-specific outputs.

As for the outputs:

- Eigenvalues: output the eigenvalues in the ascending order. The six first ones should be 0.

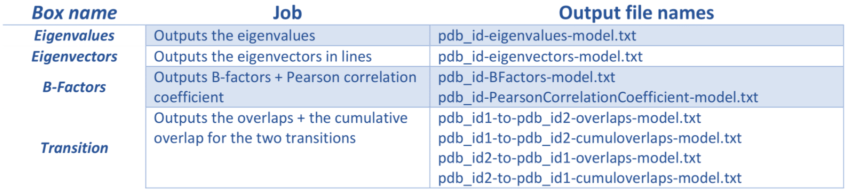
- Eigenvectors: output the eigenvectors, each line corresponds to one eigenvector.

- B-Factors: output the B-factors and the Pearson correlation coefficient in two separate files. In the B-factors file, each line corresponds to one residue, the first column is the index of residues, the second column is the sequence number of the residue, the third column is the theoretical B-factor calculated as explained in the materials and methods section and the fourth column is the experimental B-factor extracted from the pdb file.

- Transition: output the overlaps and the cumulative overlap of the transitions between the main pdb id and another one in separate files. A second structure has to be provided following the same rules as stated above. There are two ways of doing: either the NMA is carried out before taking the common set of residues and proceeding to the alignment or the inverse. The overlaps are calculated for the two transitions: ”pdb_id1-to-pdb_id2” means that the overlaps are calculated using the spectrum of pdb_id1. The alignment is automatically done.

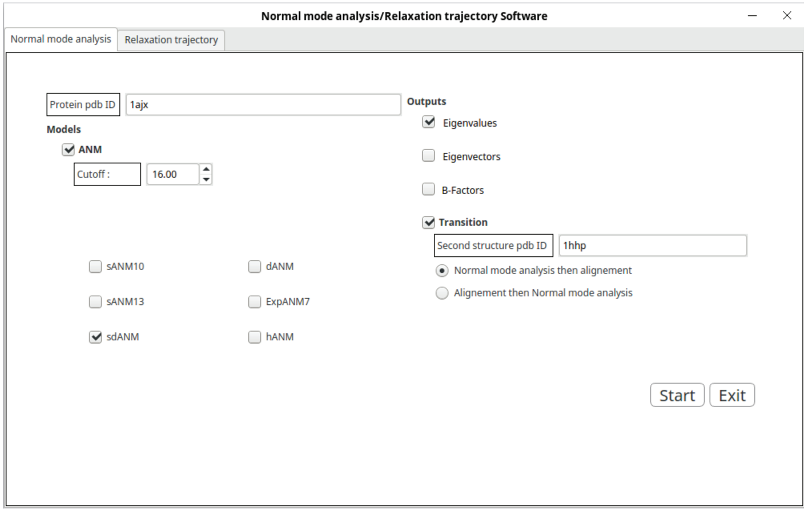
Table 10 summarises the outputs and how are generated their names. ”pdb_id” should be replaced by the provided pdb id and ”model” by the chosen model. Several outputs can be chosen at the same time.

The simulation starts when clicking the button ”Start”, a window will pop-up and the simulation can be stopped by clicking ”Cancel” which will kill the software process. An example of use is illustrated in Figure 30.



Box name	Job	Output file names
Eigenvalues	Outputs the eigenvalues	pdb_id-eigenvalues-model.txt
Eigenvectors	Outputs the eigenvectors in lines	pdb_id-eigenvectors-model.txt
B-Factors	Outputs B-factors + Pearson correlation coefficient	pdb_id-BFactors-model.txt pdb_id-PearsonCorrelationCoefficient-model.txt
Transition	Outputs the overlaps + the cumulative overlap for the two transitions	pdb_id1-to-pdb_id2-overlaps-model.txt pdb_id1-to-pdb_id2-cumuloverlaps-model.txt pdb_id2-to-pdb_id1-overlaps-model.txt pdb_id2-to-pdb_id1-cumuloverlaps-model.txt

Table 10: Summary of the outputs and their file names for the normal mode analysis.



Protein pdb ID: 1ajx

Models

☑ ANM

Cutoff : 16.00

☐ sANM10 ☐ dANM

☐ sANM13 ☐ ExpANM7

☑ sdANM ☐ hANM

Outputs

☑ Eigenvalues

☐ Eigenvectors

☐ B-Factors

☑ Transition

Second structure pdb ID: 1hhp

● Normal mode analysis then alignement

○ Alignement then Normal mode analysis

Figure 30: Example of use for the normal mode analysis. The pdb id of HIV-1 protease 1ajx is provided, we want to use the normal mode analysis with both ANM16 and sdANM to output the eigenvalues of 1ajx and to study the transition from and to 1hhp via overlaps. The pdb files of 1ajx and 1hhp have to be in the folder as well as the ”sdANM.txt” file for sdANM.

B Relaxation trajectory

As for the second tab, the software carries out dynamics simulations. The selections of the protein and of the models are the same as for the first tab.

The dynamical simulation consists, here, in deforming the initial structure and to look at how it goes back to the equilibrium. For the deformations, there are two possibilities which can be coupled. The first one is to randomly apply deformations. The initial force refers to the total magnitude. We apply uniform random forces to each residue which are rescaled such as the sum over all residues equals the total magnitude. The initial time refers to the time at which random forces are cut and the system starts to go back to the equilibrium. The user can choose to apply these forces either to all residues or to only a part of them in which case, the user has to provide a pdb-like file where the selected residues are recorded.

The second possibility is to choose a second structure which plays the role of the deformed state. We, then, look at the transition to the main structure.

The providing of the time step is not mandatory and is set to the inverse of the largest eigenvalue if not chosen by the user.

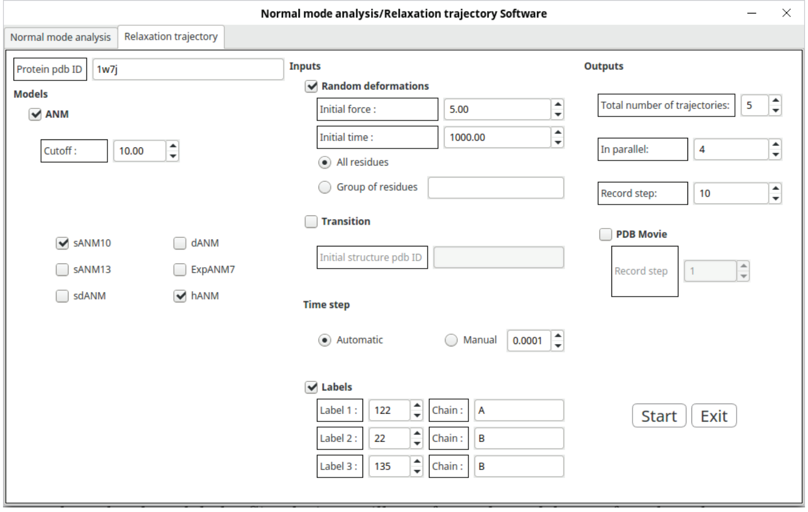
Three residues are taken as labels to follow the relaxation process. We can choose them manually by providing the sequence number of the residue and the chain letter to which the residue belongs to. If the box remains unchecked, the labels are automatically chosen[8] [38].

As for the outputs, it is possible to run several simulations using the same parameters. It is particularly useful in the case of random deformations to obtain valuable statistical data. These simulations can be run in parallel by setting a

[8] The method is explained in the appendix about relaxation trajectories.

number of desired threads. When a simulation ends in a thread, another one is started until reaching the number of simulations.
The distances between labels, their normalised distances and the mechanical coordinates are outputted for each simulation. The rate of recording can be changed via the interface. When choosing to run multiple simulations, a file compiling the distances between labels over all simulations is outputted. The size of this file is reduced by discarding cases where the distances have not changed during a time step, they are removed. Details of the content for these files are available in the next section.
Finally, a pdb movie can be selected as an output. It is readable by VMD. The positions of residues are recorded at a rate chosen by the user.

An illustration of use is available in Figure 31.



Normal mode analysis/Relaxation trajectory Software

Protein pdb ID: 1w7j

Models

☑ ANM

Cutoff : 10.00

☑ sANM10 □ dANM
□ sANM13 □ ExpANM7
□ sdANM ☑ hANM

Inputs

☑ Random deformations

Initial force : 5.00

Initial time : 1000.00

● All residues

○ Group of residues

□ Transition

Initial structure pdb ID

Time step

● Automatic ○ Manual 0.0001

☑ Labels

Label 1 : 122 Chain : A

Label 2 : 22 Chain : B

Label 3 : 135 Chain : B

Outputs

Total number of trajectories: 5

In parallel: 4

Record step: 10

□ PDB Movie

Record step 1

Figure 31: Example of use for the relaxation trajectories. The dynamical simulation will be carried out for myosin V (pdb id: 1w7j) for the models ANM10, sANM10 and hANM. The deformations are random and we have chosen ourselves the three labels. Simulations will run for each model one after the other. For each model, 5 simulations will run with 4 threads running in parallel.

C Details of output files

There are three main output text files for each simulation. In the following, ”model” refers to the model chosen and ”i” refers to the i^{th} simulation.

- pdb_id-Relaxation-model-i.txt (or pdb_id_ini-to-pdb_id_final-Relaxation-model-i.txt if an initial structure is used). This file contains the distances between the three labels starting from the deformed state. The first column is the time, then the distances 1-2, 1-3 and 2-3. The three last columns are the normalised shift of the current distances from the equilibrium ones ($\frac{d(t)-d^0}{d^0}$).

- pdb_id-Relaxation-model-CompiledTrajectories.txt compiles the above relaxation file for all simulations i in one file. Two blank lines separate two simulations. In this file, if, for one simulation, the distances at time $t + dt$ have not changed as compared to the ones at time t, then they are discarded. It allows to reduce the file size in order to be more easily readable by rendering software like Gnuplot.

- pdb_id-MechanicalCoordinate-model-%d.txt contains the elastic potential energy according to the mechanical coordinate. The first column is the mechanical coordinate Φ, the second is the elastic potential energy U and the last one is the potential energy along the slowest mode (first non-zero mode).

In addition to these three files, two other informative files are outputted. A file which displays the labels used in the simulations (convenient if the labels are automatically chosen) and a log file summarising the simulation parameters.

VII Future directions: Molecular dynamics simulations

An even more helpful way of getting insights about the motions and conformational changes, is to couple the ANM to molecular dynamics simulation (MD simulations). In such simulations, after coarse-graining (or not) the protein of interest, each residue or atom follows the Newton's equation of motion which allows to track spatially and timely the structural changes. Brownian dynamics simulations are such MD simulations based on the Brownian motion and the Langevin equation.
The Brownian motion originated from the observations of Robert Brown in 1827 on the chaotic motions of pollen particles in water [84]. Then, in 1905, Einstein described theoretically the Brownian motion linking it to the diffusion equation [85]. A few years latter, Langevin gave a different but equivalent theoretical description of the Brownian motion using a stochastic equation now known as the Langevin equation [86]. By this article, he also laid the foundation of the Langevin dynamics and *a fortiori* of the Brownian dynamics.
The Brownian dynamics itself is beyond the ENM and can be applied to any kind of network and is not limited to proteins modelled with harmonic potentials. The Brownian dynamics is the overdamped limit of the Langevin dynamics. Each residue follows an equation of its motion having two parts: a deterministic one representing forces between atoms and a stochastic one allowing the diffusion. Many biological aspects can be included in the deterministic part such as the exclusion volume effect consisting in avoiding the overlapping of two residues.
Typically, for a residue i in a protein having N residues, its Langevin equation is determined as follow:

Starting from the Newton law,

$$F_h^i(t,r) + F_d^i(t,r) + F_s^i(t,r) = 0 \tag{9}$$

where $r = (r_1, ..., r_N)$ is the vector of positions of residues.

The forces acting on the bead i are :

- The (viscous friction) hydrodynamic force

$$F_h^i = -\epsilon \dot{r_i}$$

where ϵ is the drag coefficient

-F_d^i being the sum of all deterministic forces:

- Harmonic forces (in the case of the ENM)

$$U = \frac{\kappa}{2}\sum_{j=1}^{N} a_{ij}(d_{ij} - d_{ij}^0)^2$$

with $a_{ij} = 1$ if there is a link between i and j and 0 otherwise.

- Lennard-Jones Potential force to model the Van der Walls force and the exclusion volume between two residues [87, 88]

$$F_{LJ}^i = -4\zeta\left(-12\frac{\sigma^{12}}{d^{13}} + 6\frac{\sigma^6}{d^7}\right)$$

ζ is the depth of the potential well, σ is the distance at which the potential

- The stochastic force F_s^i: it represents the collisions of the bead with the molecules of solvent. It is modelled as a Brownian motion of mean 0 $\left(\left\langle F_s^i \right\rangle = 0\right)$ due to the isotropy and a second moment $\left\langle F_s^i(t), F_s^j(t+dt)\right\rangle = \dfrac{2K_BT\epsilon}{dt}$ counteracting the dissipative forces [89]. K_B is the Boltzmann constant and T the temperature of the system.

By rearranging the equation (9), we obtain the Langevin equation on the position of the residue i :

$$dr_i = \frac{1}{\epsilon}F_d^i(t,r)dt + \sqrt{\frac{2K_BT}{\epsilon}}dW_i \tag{10}$$

where dW_i is a Gaussian process of mean 0 and variance dt.

In addition or in replacement of harmonic forces, any kind of forces can be used like rigid rods [90, 91], attractive-repulsive forces [92], bending forces [93], etc.
An interesting aspect of such simulations is that reactions and bindings can be included. It is then possible to probe the effect of a binding or a reaction to the conformation of proteins. Now, simulation techniques have been developed for the Brownian dynamics with reactions [94, 95, 96].

However, such details and freedoms of modelling come with drawbacks. The main one being the time consumption. Indeed, the time step should be smaller than a typical motion order and the equations have to be updated every time step. In particular, when considering reactions and bindings, the system has to reach the steady state before being able to collect valuable statistical data. Many studies have proposed techniques to reduce this time by jumping from event to event [97, 98, 99]. However, such methods are may be not suited if we are interested in the variation of the conformations of proteins since they overlook them. Another way would be to coarse-grain even more to reduce the number of residues in the system in the expense of some degree of accuracy.

VIII Conclusion

The first study of this thesis has focused on three recent sequence-specific models. They have been compared to the classical ANM where spring constants are homogeneous. The addition of the sequence-specificity has not brought any valuable improvement into the modelling. Although some proteins are better modelled by the sequence-specific models, the improvement does not reflect any kind of chemical properties. Furthermore, such improvements are not systematic. This weak sensitivity is unexpected since the spring constants have been determined using a NMR data set via statistical tools. Thus, a lot of information has been added when inferring the spring constants: polarity, neighbourhood, effects of the solvent, etc. However, the connectivity itself seems to drive the performance or, at least, mostly.

Observing the insensitivity to the spring constant fitting in the first study, the second one focused on some distance-dependent models which showed improvements in the Pearson correlation coefficient. The particularity of these models is that the cut-off distance is removed and the spring constants are fit accordingly. In that sense, these networks are fully connected i.e. all residues are connected with all others. These distance-dependent models definitely improve the homogeneous version of the full-connected network. However, the fitting is such that the long-rang interactions are very weak and poorly affect the modelling. This, inevitably, imposes a kind of cut-off distance. Indeed, the correlation coefficients are greatly influenced by the same flexible parts as in the first study. Then comparing these parameter-free heterogeneous models to non-parameter-free homogeneous models, we observed weak sensitivity. Again, this is unexpected since the fitting of spring constants relies on an intuitive argument which is the short-range interactions are stronger than long-range interactions. The addition of the distance in the modelling is not reflected in the improvement seen at the protein level.

The two first studies, both, support the same conclusion: the ANM is insensitive to the fitting of its spring constants and additions of chemical- or physical-related properties, other than the structure itself, do not bring any valuable information to the modelling. Then, we focused our last study on random fitting of spring constants. Particularly, the spring constant values are incoherently randomised while the connectivity remains the same. In average, performances are similar to the ones of a classical ANM and the distribution of PCCs around the average one is generally not wide highlighting a true robustness against the fitting of spring constants. The investigations of the networks with the best performances reveal interesting information for proteins for which the performances of the preceding networks are biased by flexible parts. The large PCC is solely due to the better modelling of flexible parts. Indeed, only spring constants of links related to these parts are systematically stronger or weaker than the mean value. Then, enhancements do not reflect any kind of physical or chemical properties. This last study shows in a large extent the robustness of the ANM against the fitting of its spring constant and it highlights the non coherence of the improvements when they are.

Our three studies show that fitting the spring constants in order to improve the modelling is a not a good direction. The properties intended to be included in the model are not reflected in the performances preventing from any kind of valuable study on these properties (e.g. a study on the effect of mutations). The major drawback of the ANM is the inability to correctly predict the fluctuations of flexible parts. This is still true for all versions of the ANM excepted for those which completely damp the fluctuations. This drawback for ANM has been already reported [79, 80, 100], it is called the tip effect. Many studies focused on suppressing this effect by bond angle restrictions [79, 80] or by taking into account the crystal packing for crystallographic proteins [20]. The strategy of both of them consists in the stiffening of the flexible parts. It is in agreement with our randomisation study.

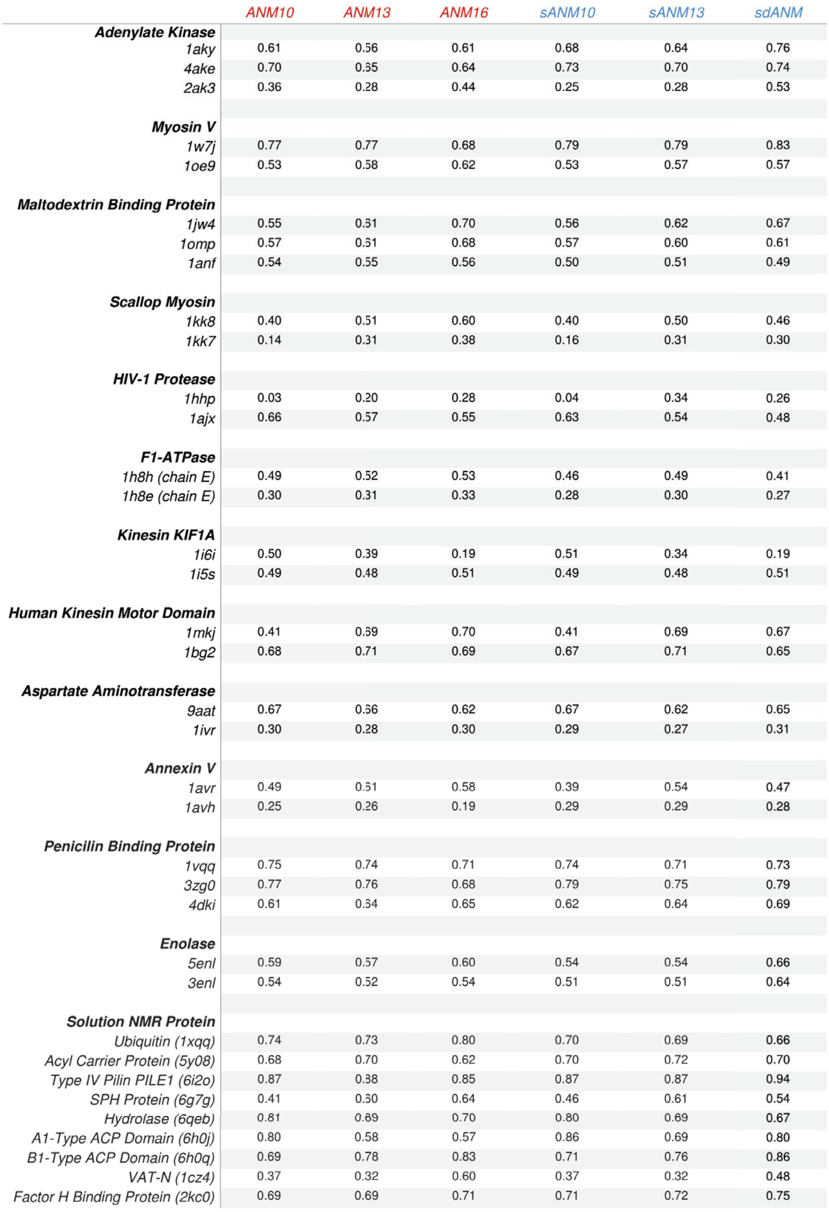
A future direction of this work would be to look at the design of the model itself. It is determined in an empirical manner where the only physical feature is the relative positions of residues. Otherwise, the choice of connections is not related to any chemical or physical arguments. Its simplicity already gives useful and accurate results but it would be difficult to valuably include chemical details in the modelling if the network itself is not related to any chemical background. Then, including physical and chemical details at the level of the setting of the connectivity of the network should be the next steps toward the improvement of the ANM. More or less in this spirit, some studies has already focused on considering several type of interactions by making the spring constants interaction-type dependent [23, 33]. However, these models include additional parameters making a systematic study even harder. Up to date, it is missing.

Appendices

A B-factor and PCC studies

A.1 For sequence specific models

Table 11 shows the Pearson correlation coefficient for the full set of proteins used as individual study for sequence specific models.



	ANM10	ANM13	ANM16	sANM10	sANM13	sdANM
Adenylate Kinase						
1aky	0.61	0.56	0.61	0.68	0.64	0.76
4ake	0.70	0.65	0.64	0.73	0.70	0.74
2ak3	0.36	0.28	0.44	0.25	0.28	0.53
Myosin V						
1w7j	0.77	0.77	0.68	0.79	0.79	0.83
1oe9	0.53	0.58	0.62	0.53	0.57	0.57
Maltodextrin Binding Protein						
1jw4	0.55	0.61	0.70	0.56	0.62	0.67
1omp	0.57	0.61	0.68	0.57	0.60	0.61
1anf	0.54	0.55	0.56	0.50	0.51	0.49
Scallop Myosin						
1kk8	0.40	0.51	0.60	0.40	0.50	0.46
1kk7	0.14	0.31	0.38	0.16	0.31	0.30
HIV-1 Protease						
1hhp	0.03	0.20	0.28	0.04	0.34	0.26
1ajx	0.66	0.57	0.55	0.63	0.54	0.48
F1-ATPase						
1h8h (chain E)	0.49	0.52	0.53	0.46	0.49	0.41
1h8e (chain E)	0.30	0.31	0.33	0.28	0.30	0.27
Kinesin KIF1A						
1i6i	0.50	0.39	0.19	0.51	0.34	0.19
1i5s	0.49	0.48	0.51	0.49	0.48	0.51
Human Kinesin Motor Domain						
1mkj	0.41	0.69	0.70	0.41	0.69	0.67
1bg2	0.68	0.71	0.69	0.67	0.71	0.65
Aspartate Aminotransferase						
9aat	0.67	0.66	0.62	0.67	0.62	0.65
1ivr	0.30	0.28	0.30	0.29	0.27	0.31
Annexin V						
1avr	0.49	0.61	0.58	0.39	0.54	0.47
1avh	0.25	0.26	0.19	0.29	0.29	0.28
Penicilin Binding Protein						
1vqq	0.75	0.74	0.71	0.74	0.71	0.73
3zg0	0.77	0.76	0.68	0.79	0.75	0.79
4dki	0.61	0.64	0.65	0.62	0.64	0.69
Enolase						
5enl	0.59	0.57	0.60	0.54	0.54	0.66
3enl	0.54	0.52	0.54	0.51	0.51	0.64
Solution NMR Protein						
Ubiquitin (1xqq)	0.74	0.73	0.80	0.70	0.69	0.66
Acyl Carrier Protein (5y08)	0.68	0.70	0.62	0.70	0.72	0.70
Type IV Pilin PILE1 (6i2o)	0.87	0.88	0.85	0.87	0.87	0.94
SPH Protein (6g7g)	0.41	0.60	0.64	0.46	0.61	0.54
Hydrolase (6qeb)	0.81	0.69	0.70	0.80	0.69	0.67
A1-Type ACP Domain (6h0j)	0.80	0.58	0.57	0.86	0.69	0.80
B1-Type ACP Domain (6h0q)	0.69	0.78	0.83	0.71	0.76	0.86
VAT-N (1cz4)	0.37	0.32	0.60	0.37	0.32	0.48
Factor H Binding Protein (2kc0)	0.69	0.69	0.71	0.71	0.72	0.75

Table 11: Table of Pearson correlation coefficients. Carried out for some popular proteins with the 6 models we have considered, red being related to classical ANM and blue being related to sequence-specific models.

A.2 Heterogeneous parameter-free models

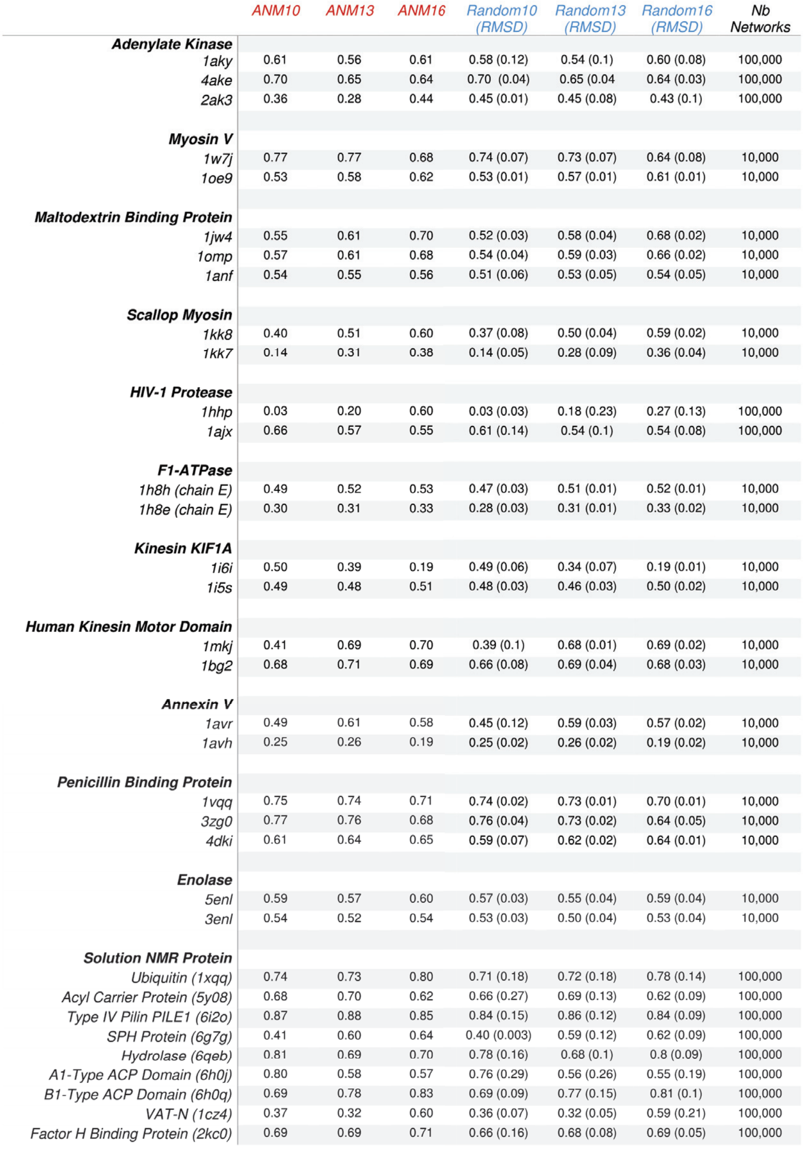
Table 12 shows the Pearson correlation coefficient for the full set of proteins used as individual study for parameter-free heterogeneous models.

	ANM10	*ANM16*	*ANM-AC*	*dANM*	*ExpANM3*	*ExpANM7*	*hANM*
Adenylate Kinase							
1aky	0.61	0.61	0.62	0.65	0.76	0.70	0.73
4ake	0.70	0.64	0.50	0.55	0.76	0.68	0.69
2ak3	0.36	0.44	0.40	0.67	0.48	0.49	0.54
Myosin V							
1w7j	0.77	0.68	0.81	0.71	0.82	0.81	0.81
1oe9	0.53	0.62	0.61	0.68	0.58	0.59	0.61
Maltodextrin Binding Protein							
1jw4	0.55	0.70	0.46	0.67	0.65	0.68	0.70
1omp	0.57	0.68	0.49	0.66	0.61	0.67	0.67
1anf	0.54	0.56	0.65	0.71	0.50	0.57	0.56
Scallop Myosin							
1kk8	0.40	0.60	0.61	0.66	0.30	0.54	0.55
1kk7	0.14	0.38	0.45	0.56	0.11	0.38	0.40
HIV-1 Protease							
1hhp	0.03	0.60	0.35	0.28	0.04	0.23	0.22
1ajx	0.66	0.55	0.59	0.58	0.56	0.61	0.59
F1-ATPase							
1h8h (chain E)	0.49	0.53	0.55	0.53	0.43	0.51	0.50
1h8e (chain E)	0.30	0.33	0.36	0.37	0.25	0.32	0.31
Kinesin KIF1A							
1i6i	0.50	0.19	0.59	0.66	0.19	0.29	0.38
1i5s	0.49	0.51	0.56	0.63	0.21	0.50	0.52
Human Kinesin Motor Domain							
1mkj	0.41	0.70	0.71	0.77	0.62	0.71	0.71
1bg2	0.68	0.69	0.69	0.74	0.60	0.74	0.74
Aspartate Aminotransferase							
9aat	0.67	0.62	0.50	0.60	0.68	0.68	0.68
1ivr	0.30	0.30	0.44	0.49	0.33	0.29	0.36
Annexin V							
1avr	0.49	0.58	0.34	0.59	0.32	0.29	0.56
1avh	0.25	0.19	0.34	0.56	0.25	0.19	0.28
Penicillin Binding Protein							
1vqq	0.75	0.71	0.30	0.63	0.71	0.75	0.75
3zg0	0.77	0.68	0.52	0.71	0.76	0.78	0.79
4dki	0.61	0.65	0.55	0.64	0.64	0.66	0.67
Enolase							
5enl	0.59	0.60	0.58	0.69	0.64	0.65	0.68
3enl	0.54	0.54	0.49	0.60	0.64	0.60	0.64
Solution NMR Protein							
Ubiquitin (1xqq)	0.74	0.80	0.88	0.89	0.63	0.78	0.75
Acyl Carrier Protein (5y08)	0.68	0.62	0.38	0.51	0.50	0.72	0.79
Type IV Pilin PILE1 (6i2o)	0.87	0.85	0.60	0.76	0.93	0.88	0.91
SPH Protein (6g7g)	0.41	0.64	0.66	0.62	0.36	0.61	0.59
Hydrolase (6qeb)	0.81	0.70	0.95	0.94	0.62	0.72	0.74
A1-Type ACP Domain (6h0j)	0.80	0.57	0.53	0.60	0.77	0.66	0.78
B1-Type ACP Domain (6h0q)	0.69	0.83	0.57	0.79	0.72	0.86	0.89
VAT-N (1cz4)	0.37	0.60	0.49	0.68	0.37	0.51	0.53
Factor H Binding Protein (2kc0)	0.69	0.71	0.62	0.75	0.74	0.71	0.73

Table 12: Pearson correlation coefficient for distance-dependent models alongside with classical ANMs with a cut-off distance.

A.3 For the randomised spring constant models

Table 13 shows the averaged Pearson correlation coefficient among 10,000 or 100,000 random networks for the full set of proteins.



	ANM10	ANM13	ANM16	Random10 (RMSD)	Random13 (RMSD)	Random16 (RMSD)	Nb Networks
Adenylate Kinase							
1aky	0.61	0.56	0.61	0.58 (0.12)	0.54 (0.1)	0.60 (0.08)	100,000
4ake	0.70	0.65	0.64	0.70 (0.04)	0.65 (0.04	0.64 (0.03)	100,000
2ak3	0.36	0.28	0.44	0.45 (0.01)	0.45 (0.08)	0.43 (0.1)	100,000
Myosin V							
1w7j	0.77	0.77	0.68	0.74 (0.07)	0.73 (0.07)	0.64 (0.08)	10,000
1oe9	0.53	0.58	0.62	0.53 (0.01)	0.57 (0.01)	0.61 (0.01)	10,000
Maltodextrin Binding Protein							
1jw4	0.55	0.61	0.70	0.52 (0.03)	0.58 (0.04)	0.68 (0.02)	10,000
1omp	0.57	0.61	0.68	0.54 (0.04)	0.59 (0.03)	0.66 (0.02)	10,000
1anf	0.54	0.55	0.56	0.51 (0.06)	0.53 (0.05)	0.54 (0.05)	10,000
Scallop Myosin							
1kk8	0.40	0.51	0.60	0.37 (0.08)	0.50 (0.04)	0.59 (0.02)	10,000
1kk7	0.14	0.31	0.38	0.14 (0.05)	0.28 (0.09)	0.36 (0.04)	10,000
HIV-1 Protease							
1hhp	0.03	0.20	0.60	0.03 (0.03)	0.18 (0.23)	0.27 (0.13)	100,000
1ajx	0.66	0.57	0.55	0.61 (0.14)	0.54 (0.1)	0.54 (0.08)	100,000
F1-ATPase							
1h8h (chain E)	0.49	0.52	0.53	0.47 (0.03)	0.51 (0.01)	0.52 (0.01)	10,000
1h8e (chain E)	0.30	0.31	0.33	0.28 (0.03)	0.31 (0.01)	0.33 (0.02)	10,000
Kinesin KIF1A							
1i6i	0.50	0.39	0.19	0.49 (0.06)	0.34 (0.07)	0.19 (0.01)	10,000
1i5s	0.49	0.48	0.51	0.48 (0.03)	0.46 (0.03)	0.50 (0.02)	10,000
Human Kinesin Motor Domain							
1mkj	0.41	0.69	0.70	0.39 (0.1)	0.68 (0.01)	0.69 (0.02)	10,000
1bg2	0.68	0.71	0.69	0.66 (0.08)	0.69 (0.04)	0.68 (0.03)	10,000
Annexin V							
1avr	0.49	0.61	0.58	0.45 (0.12)	0.59 (0.03)	0.57 (0.02)	10,000
1avh	0.25	0.26	0.19	0.25 (0.02)	0.26 (0.02)	0.19 (0.02)	10,000
Penicillin Binding Protein							
1vqq	0.75	0.74	0.71	0.74 (0.02)	0.73 (0.01)	0.70 (0.01)	10,000
3zg0	0.77	0.76	0.68	0.76 (0.04)	0.73 (0.02)	0.64 (0.05)	10,000
4dki	0.61	0.64	0.65	0.59 (0.07)	0.62 (0.02)	0.64 (0.01)	10,000
Enolase							
5enl	0.59	0.57	0.60	0.57 (0.03)	0.55 (0.04)	0.59 (0.04)	10,000
3enl	0.54	0.52	0.54	0.53 (0.03)	0.50 (0.04)	0.53 (0.04)	10,000
Solution NMR Protein							
Ubiquitin (1xqq)	0.74	0.73	0.80	0.71 (0.18)	0.72 (0.18)	0.78 (0.14)	100,000
Acyl Carrier Protein (5y08)	0.68	0.70	0.62	0.66 (0.27)	0.69 (0.13)	0.62 (0.09)	100,000
Type IV Pilin PILE1 (6i2o)	0.87	0.88	0.85	0.84 (0.15)	0.86 (0.12)	0.84 (0.09)	100,000
SPH Protein (6g7g)	0.41	0.60	0.64	0.40 (0.003)	0.59 (0.12)	0.62 (0.09)	100,000
Hydrolase (6qeb)	0.81	0.69	0.70	0.78 (0.16)	0.68 (0.1)	0.8 (0.09)	100,000
A1-Type ACP Domain (6h0j)	0.80	0.58	0.57	0.76 (0.29)	0.56 (0.26)	0.55 (0.19)	100,000
B1-Type ACP Domain (6h0q)	0.69	0.78	0.83	0.69 (0.09)	0.77 (0.15)	0.81 (0.1)	100,000
VAT-N (1cz4)	0.37	0.32	0.60	0.36 (0.07)	0.32 (0.05)	0.59 (0.21)	100,000
Factor H Binding Protein (2kc0)	0.69	0.69	0.71	0.66 (0.16)	0.68 (0.08)	0.69 (0.05)	100,000

Table 13: Table of Pearson correlation coefficient for the full set of proteins. For the random models, the PCC has been averaged over either 10,000 or 100,000 networks. The RMSD through the networks is calculated in parenthesis.

B Relaxation trajectory studies

B.1 For the sequence specific models

1 Method

The three labels are chosen according to an automatic way [38]. For HCV helicase, the normal mode analysis has been carried out for ANM8. The two first labels correspond to the pair for which the distance change after applying the slowest mode is the maximal then the last label is chosen as the residue for which the distance change between it and the label 1 is maximal after applying the second slowest mode. The same three labels have been saved and used for the other models (ANM10, sdANM, ...). For F_1-ATPase and myosin V, the procedure is the same but the labels have been determined for ANM10.

- HCV Helicase
  - Label 1: Chain A Residue 257
  - Label 2: Chain A Residue 402
  - Label 3: Chain A Residue 589

- F1-ATPase
  - Label 1: Chain E Residue 191
  - Label 2: Chain E Residue 390
  - Label 3: Chain E Residue 54

- Myosin V
  - Label 1: Chain A Residue 122
  - Label 2: Chain B Residue 22
  - Label 3: Chain B Residue 135

As for the deformations, uniform random forces $f = (f_1, ..., f_N)$ are applied on each residue such as $\|f\|_2 = F_{ini}$ where F_{ini} (in force unit) is the total magnitude. The forces are applied until a certain time T_{ini} (in time unit) after which they are cut and the system goes back to the equilibrium. The total magnitude and the initial time at which the relaxation start depends on the protein and the model:

- HCV helicase
  - ANM8: $F_{ini} = 1$, $T_{ini} = 5000$
  - ANM10: $F_{ini} = 40$, $T_{ini} = 200$
  - ANM13: $F_{ini} = 100$, $T_{ini} = 200$
  - ANM16: $F_{ini} = 500$, $T_{ini} = 1000$
  - sANM10: $F_{ini} = 20$, $T_{ini} = 1000$
  - sANM13: $F_{ini} = 200$, $T_{ini} = 1000$
  - sdANM: $F_{ini} = 5$, $T_{ini} = 1000$

- F1-ATPase
  - ANM10: $F_{ini} = 10$, $T_{ini} = 30000$
  - ANM13: $F_{ini} = 50$, $T_{ini} = 100$
  - ANM16: $F_{ini} = 500$, $T_{ini} = 1000$

- sANM10: $F_{ini} = 20$, $T_{ini} = 1000$
- sANM13: $F_{ini} = 100$, $T_{ini} = 1000$
- sdANM: $F_{ini} = 10$, $T_{ini} = 1000$

- Myosin V
  - ANM10: $F_{ini} = 5$, $T_{ini} = 1000$
  - ANM13: $F_{ini} = 10$, $T_{ini} = 1000$
  - ANM16: $F_{ini} = 50$, $T_{ini} = 1000$
  - sANM10: $F_{ini} = 10$, $T_{ini} = 1000$
  - sANM13: $F_{ini} = 10$, $T_{ini} = 1000$
  - sdANM: $F_{ini} = 5$, $T_{ini} = 1000$

2 Results

All results about the relaxation trajectories along the three label distances for HCV helicase, F_1-ATPase and myosin V are displayed in Figures 32 to 42.

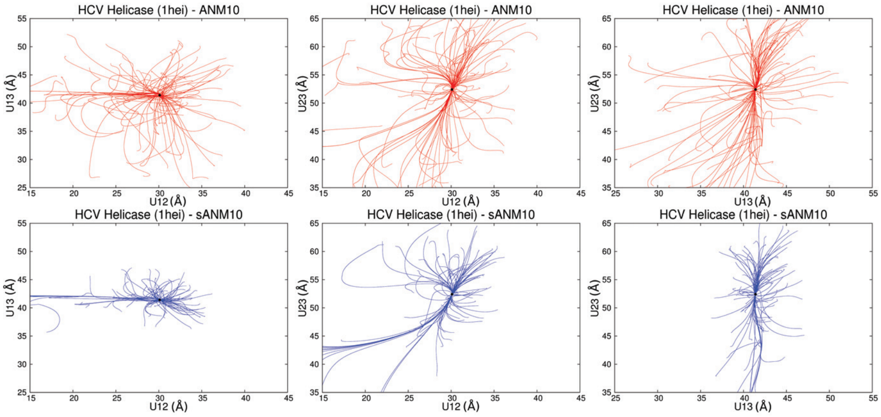


Figure 32: 100 relaxation trajectories from random deformations for HCV helicase with ANM10 (red) and sANM10 (blue).

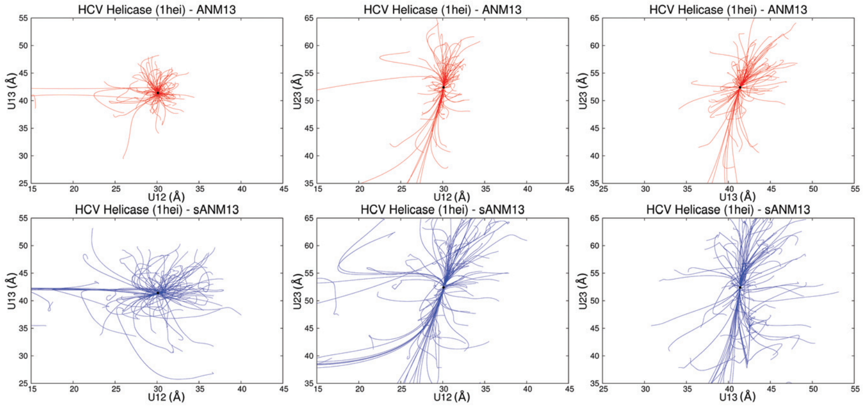


Figure 33: 100 relaxation trajectories from random deformations for HCV helicase with ANM13 (red) and sANM13 (blue).

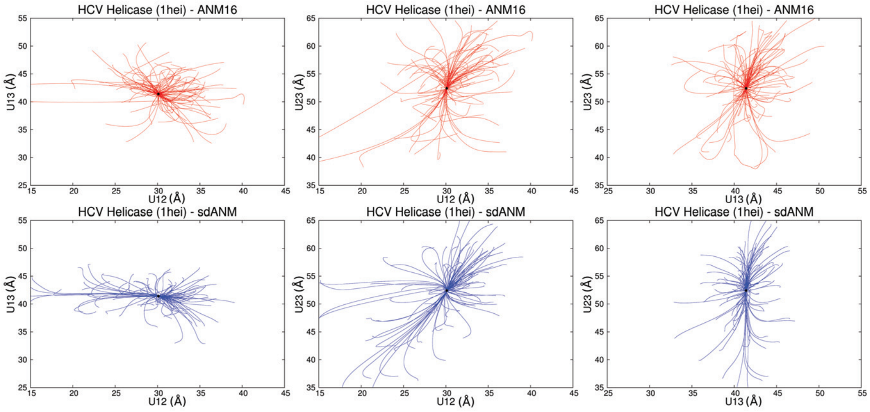


Figure 34: 100 relaxation trajectories from random deformations for HCV helicase with ANM16 (red) and sdANM (blue).

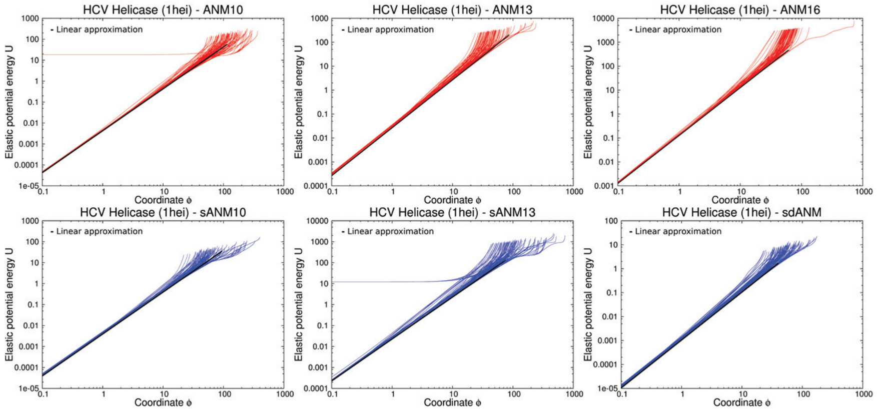


Figure 35: Elastic energy relaxations corresponding to 100 relaxation trajectories starting from random deformations for HCV helicase (pdb id: 1hei). The black line corresponds to the energy relaxation of the lowest frequency mode.

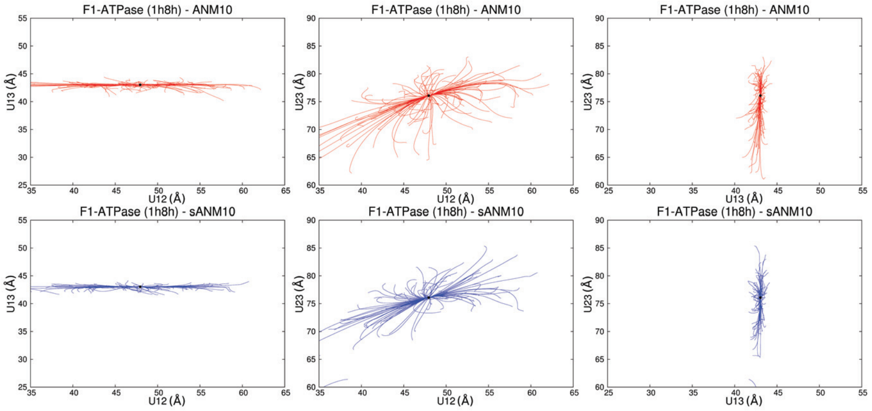


Figure 36: 100 relaxation trajectories from random deformations for F_1-ATPase with ANM10 (red) and sANM10 (blue).

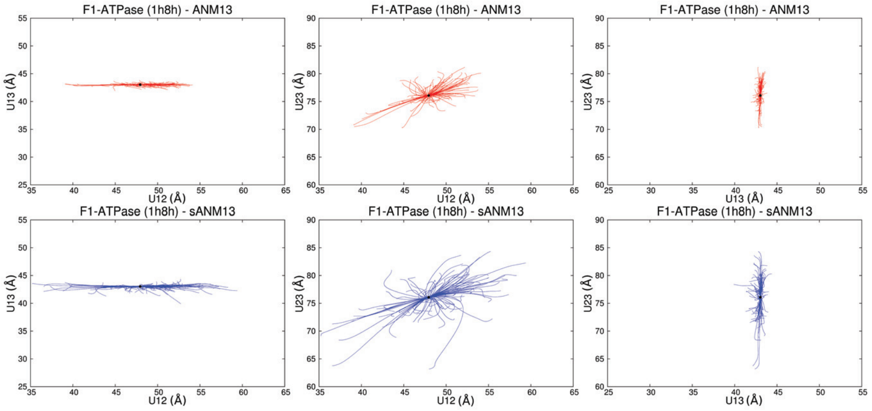


Figure 37: 100 relaxation trajectories from random deformations for F_1-ATPase with ANM13 (red) and sANM13 (blue).

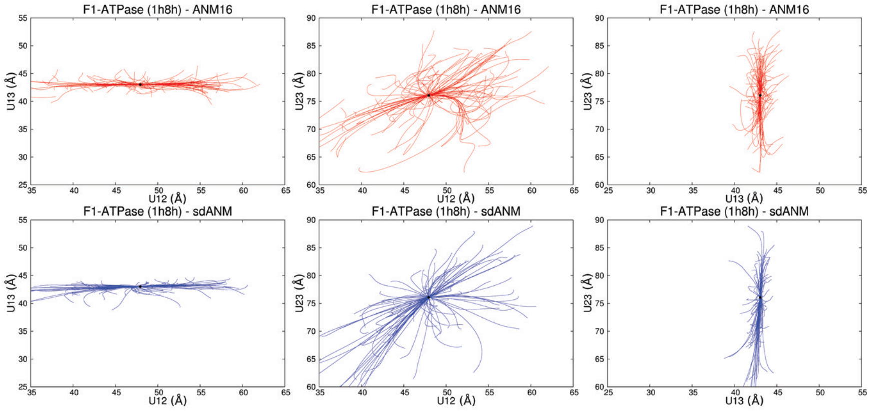


Figure 38: 100 relaxation trajectories from random deformations for F_1-ATPase with ANM16 (red) and sdANM (blue).

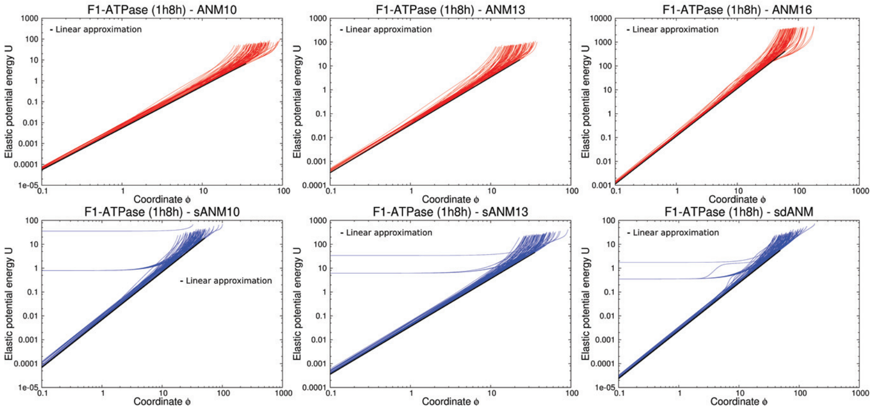


Figure 39: Elastic energy relaxations corresponding to 100 relaxation trajectories starting from random deformations for F_1-ATPase (pdb id: 1h8h). The black line corresponds to the energy relaxation of the lowest frequency mode.

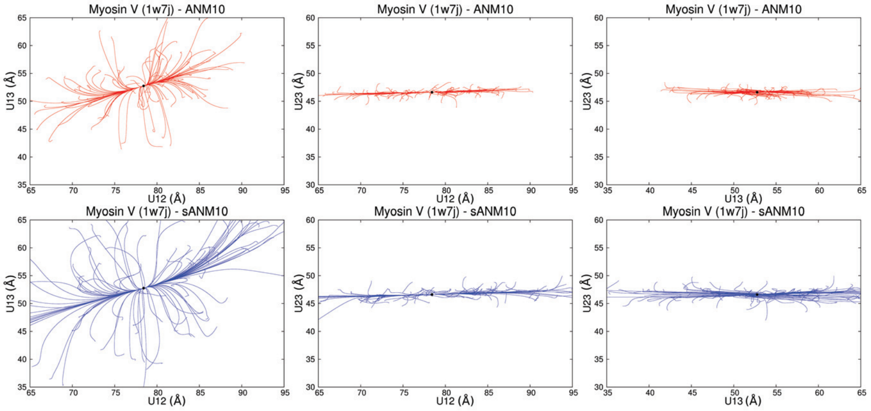


Figure 40: 100 relaxation trajectories from random deformations for myosin V with ANM10 (red) and sANM10 (blue).

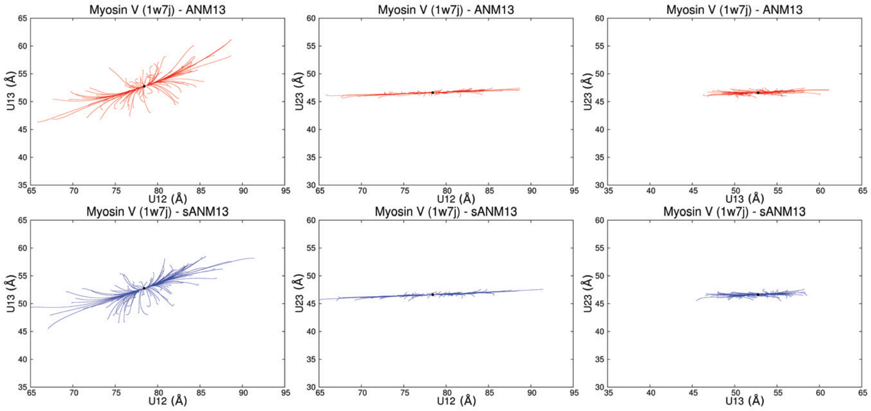


Figure 41: 100 relaxation trajectories from random deformations for myosin V with ANM13 (red) and sANM13 (blue).

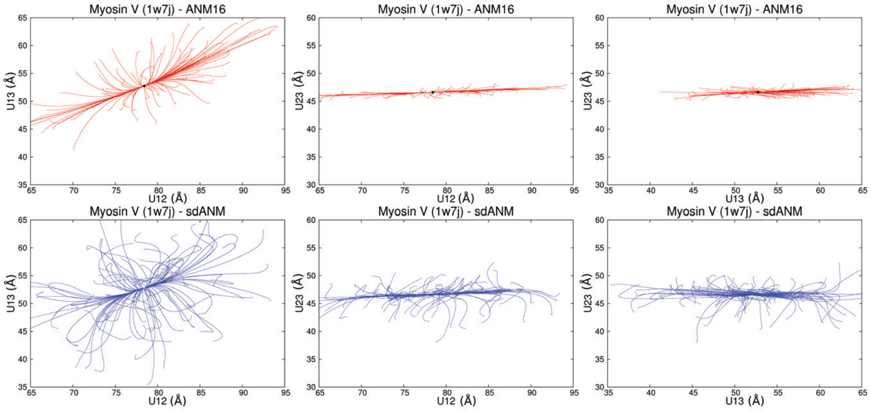


Figure 42: 100 relaxation trajectories from random deformations for myosin V with ANM16 (red) and sdANM (blue).

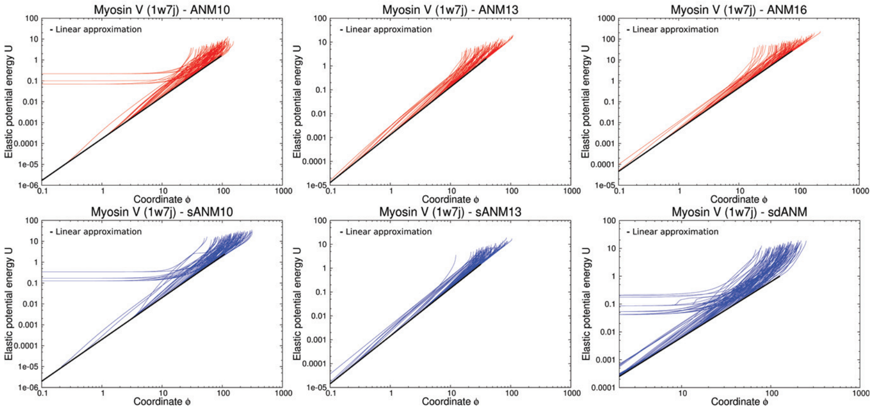


Figure 43: Elastic energy relaxations corresponding to 100 relaxation trajectories starting from random deformations for myosin V (pdb id: 1w7j). The black line corresponds to the energy relaxation of the lowest frequency mode.

B.2 For heterogeneous parameter-free models

Only myosin V was used for this study. The labels are the same as above:
- Label 1: Chain A Residue 122
- Label 2: Chain B Residue 22
- Label 3: Chain B Residue 135

As for the total magnitude and the initial time:
- ANM16: $F_{ini} = 50$, $T_{ini} = 1000$
- dANM: $F_{ini} = 25$, $T_{ini} = 1000$
- ExpANM7: $F_{ini} = 25$, $T_{ini} = 1000$
- hANM: $F_{ini} = 25$, $T_{ini} = 1000$

Bibliography

[1] Andrew F Huxley and R Niedergerke. Structural changes in muscle during contraction: interference microscopy of living muscle fibres. *Nature*, 173(4412):971, 1954.

[2] Hugh Huxley and Jean Hanson. Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. *Acta Physiol. Scani*, 6:123, 1954.

[3] Nobutaka Hirokawa, Yasuko Noda, and Yasushi Okada. Kinesin and dynein superfamily proteins in organelle transport and cell division. *Current opinion in cell biology*, 10(1):60–73, 1998.

[4] Harukata Miki, Yasushi Okada, and Nobutaka Hirokawa. Analysis of the kinesin superfamily: insights into structure and function. *Trends in cell biology*, 15(9):467–476, 2005.

[5] David N Frick. The hepatitis c virus ns3 protein: a model rna helicase and potential drug target. *Current issues in molecular biology*, 9(1):1, 2007.

[6] Terri Watkins, Wolfgang Resch, David Irlbeck, and Ronald Swanstrom. Selection of high-level resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrobial agents and chemotherapy*, 47(2):759–769, 2003.

[7] Ashraf Brik and Chi-Huey Wong. Hiv-1 protease: mechanism and drug discovery. *Organic & biomolecular chemistry*, 1(1):5–14, 2003.

[8] Holger Flechsig. Design of elastic networks with evolutionary optimized long-range communication as mechanical models of allosteric proteins. *Biophysical journal*, 113(3):558–571, 2017.

[9] Holger Flechsig and Alexander S Mikhailov. Tracing entire operation cycles of molecular motor hepatitis c virus helicase in structurally resolved dynamical simulations. *Proceedings of the National Academy of Sciences*, 107(49):20875–20880, 2010.

[10] Yuichi Togashi and Kunihiko Kaneko. Switching dynamics in reaction networks induced by molecular discreteness. *Journal of Physics: Condensed Matter*, 19(6):065150, 2007.

[11] Eran Eyal and Ivet Bahar. Toward a molecular understanding of the anisotropic response of proteins to external forces: insights from elastic network models. *Biophysical Journal*, 94(9):3424–3435, 2008.

[12] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, 2(3):173–181, 1997.

[13] Turkan Haliloglu, Ivet Bahar, and Burak Erman. Gaussian dynamics of folded proteins. *Physical review letters*, 79(16):3090, 1997.

[14] Ali Rana Atilgan, SR Durell, Robert L Jernigan, MC Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[15] Florence Tama and Y-H Sanejouand. Conformational change of proteins arising from normal mode calculations. *Protein engineering*, 14(1):1–6, 2001.

[16] Wenjun Zheng and Sebastian Doniach. A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences*, 100(23):13253–13258, 2003.

[17] Konrad Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Bioinformatics*, 33(3):417–429, 1998.

[18] Konrad Hinsen and Gerald R Kneller. A simplified force field for describing vibrational protein dynamics over the whole frequency range. *The Journal of Chemical Physics*, 111(24):10766–10769, 1999.

[19] Konrad Hinsen, Andrei-Jose Petrescu, Serge Dellerue, Marie-Claire Bellissent-Funel, and Gerald R Kneller. Harmonicity in slow protein dynamics. *Chemical Physics*, 261(1-2):25–37, 2000.

[20] Demian Riccardi, Qiang Cui, and George N Phillips Jr. Application of elastic network models to proteins in the crystalline state. *Biophysical journal*, 96(2):464–475, 2009.

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

[22] Yves Dehouck and Alexander S Mikhailov. Effective harmonic potentials: insights into the internal cooperativity and sequence-specificity of protein dynamics. *PLoS computational biology*, 9(8):e1003209, 2013.

[23] Dmitry A Kondrashov, Qiang Cui, and George N Phillips Jr. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophysical journal*, 91(8):2760–2767, 2006.

[24] Kei Moritsugu and Jeremy C Smith. Coarse-grained biomolecular simulation with reach: realistic extension algorithm via covariance hessian. *Biophysical journal*, 93(10):3460–3469, 2007.

[25] Yuichi Togashi and Holger Flechsig. Coarse-grained protein dynamics studies using elastic network models. *International journal of molecular sciences*, 19(12):3899, 2018.

[26] Ivet Bahar and AJ Rader. Coarse-grained normal mode analysis in structural biology. *Current opinion in structural biology*, 15(5):586–592, 2005.

[27] Jianpeng Ma. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure*, 13(3):373–380, 2005.

[28] Qiang Cui and Ivet Bahar. *Normal mode analysis: theory and applications to biological and chemical systems.* CRC press, 2005.

[29] Ivet Bahar, Timothy R Lezon, Lee-Wei Yang, and Eran Eyal. Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, 39:23–42, 2010.

[30] José Ramón López-Blanco and Pablo Chacón. New generation of elastic network models. *Current opinion in structural biology*, 37:46–53, 2016.

[31] Holger Flechsig and Yuichi Togashi. Designed elastic networks: Models of complex protein machinery. *International journal of molecular sciences*, 19(10):3152, 2018.

[32] B. Alberts, A.D. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell: Sixth International Student Edition.* 500 Tips. Garland Science, Taylor and Francis Group, 2015.

[33] Jay I Jeong, Yunho Jang, and Moon K Kim. A connection rule for α-carbon coarse-grained elastic network models using chemical bond information. *Journal of Molecular Graphics and Modelling*, 24(4):296–306, 2006.

[34] Junichi Higo and Hideaki Umeyama. Protein dynamics determined by backbone conformation and atom packing. *Protein engineering*, 10(4):373–380, 1997.

[35] Adolfo B Poma, Mai Suan Li, and Panagiotis E Theodorakis. Generalization of the elastic network model for the study of large conformational changes in biomolecules. *Physical Chemistry Chemical Physics*, 20(25):17020–17028, 2018.

[36] Osamu Miyashita, José Nelson Onuchic, and Peter G Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proceedings of the National Academy of Sciences*, 100(22):12570–12575, 2003.

[37] Brice Juanico, Y-H Sanejouand, Francesco Piazza, and Paolo De Los Rios. Discrete breathers in nonlinear network models of proteins. *Physical review letters*, 99(23):238104, 2007.

[38] Yuichi Togashi and Alexander S Mikhailov. Nonlinear relaxation dynamics in elastic networks and design principles of molecular machines. *Proceedings of the National Academy of Sciences*, 104(21):8697–8702, 2007.

[39] Yuichi Togashi, Toshio Yanagida, and Alexander S Mikhailov. Nonlinearity of mechanochemical motions in motor proteins. *PLoS computational biology*, 6(6):e1000814, 2010.

[40] Lee-Wei Yang, Eran Eyal, Chakra Chennubhotla, JunGoo Jee, Angela M Gronenborn, and Ivet Bahar. Insights into equilibrium dynamics of proteins from comparison of nmr and x-ray data with computational predictions. *Structure*, 15(6):741–749, 2007.

[41] Edvin Fuglebakk, Julián Echave, and Nathalie Reuter. Measuring and comparing structural fluctuation patterns in large protein datasets. *Bioinformatics*, 28(19):2431–2440, 2012.

[42] Konrad Hinsen. Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics*, 24(4):521–528, 2007.

[43] Masahide Kikkawa, Elena P Sablin, Yasushi Okada, Hiroaki Yajima, Robert J Fletterick, and Nobutaka Hirokawa. Switch-based mechanism of kinesin motors. *Nature*, 411(6836):439, 2001.

[44] F Jon Kull, Elena P Sablin, Rebecca Lau, Robert J Fletterick, and Ronald D Vale. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature*, 380(6574):550, 1996.

[45] Charles V Sindelar, Mary Jane Budny, Sarah Rice, Nariman Naber, Robert Fletterick, and Roger Cooke. Two conformations in the human kinesin power stroke defined by x-ray crystallography and epr spectroscopy. *Nature Structural & Molecular Biology*, 9(11):844, 2002.

[46] Catherine A McPhalen, Michael G Vincent, and Johan N Jansonius. X-ray structure refinement and comparison of three forms of mitochondrial aspartate aminotransferase. *Journal of molecular biology*, 225(2):495–517, 1992.

[47] Catherine A McPhalen, Michael G Vincent, Daniel Picot, Johan N Jansonius, Arthur M Lesk, and Cyrus Chothia. Domain closure in mitochondrial aspartate aminotransferase. *Journal of molecular biology*, 227(1):197–213, 1992.

[48] Andreas Graf von Stosch. Aspartate aminotransferase complexed with erythro-β-hydroxyaspartate: Crystallographic and spectroscopic identification of the carbinolamine intermediate. *Biochemistry*, 35(48):15260–15268, 1996.

[49] Francesca Cantini, Daniele Veggi, Sara Dragonetti, Silvana Savino, Maria Scarselli, Giacomo Romagnoli, Mariagrazia Pizza, Lucia Banci, and Rino Rappuoli. Solution structure of the factor h-binding protein, a survival factor and protective antigen of neisseria meningitidis. *Journal of Biological Chemistry*, 284(14):9022–9026, 2009.

[50] Xiaoqun Duan and Florante A Quiocho. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry*, 41(3):706–712, 2002.

[51] Florante A Quiocho, John C Spurlino, and Lynn E Rodseth. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure*, 5(8):997–1015, 1997.

[52] Andrew J Sharff, Lynn E Rodseth, John C Spurlino, and Florante A Quiocho. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry*, 31(44):10657–10663, 1992.

[53] Luisa Moretto, Rachel Heylen, Natalie Holroyd, Steven Vance, and R William Broadhurst. Modular type i polyketide synthase acyl carrier protein domains share a common n-terminally extended fold. *Scientific reports*, 9(1):2325, 2019.

[54] U Abele and GE Schulz. High-resolution structures of adenylate kinase from yeast ligated with inhibitor ap5a, showing the pathway of phosphoryl transfer. *Protein Science*, 4(7):1262–1271, 1995.

[55] Kay Diederichs and Georg E Schulz. The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate amp at 1.85 å resolution. *Journal of molecular biology*, 217(3):541–549, 1991.

[56] CW Müller, GJ Schlauderer, Jochen Reinstein, and Georg E Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4(2):147–156, 1996.

[57] Pierre-Damien Coureux, H Lee Sweeney, and Anne Houdusse. Three myosin v structures delineate essential features of chemo-mechanical transduction. *The EMBO journal*, 23(23):4527–4537, 2004.

[58] Pierre-Damien Coureux, Amber L Wells, Julie Ménétrey, Christopher M Yengo, Carl A Morris, H Lee Sweeney, and Anne Houdusse. A structural state of the myosin v motor without bound nucleotide. *Nature*, 425(6956):419, 2003.

[59] Robert Huber, Robert Berendes, Alexander Burger, Monika Schneider, Andrej Karshikov, Hartmut Luecke, Jürgen Römisch, and Eric Paques. Crystal and molecular structure of human annexin v after refinement: implications for structure, membrane binding and ion channel formation of the annexin family of proteins. *Journal of molecular biology*, 223(3):683–704, 1992.

[60] Kresten Lindorff-Larsen, Robert B Best, Mark A DePristo, Christopher M Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128, 2005.

[61] Jamie-Lee Berry, Ishwori Gurung, Jan Haug Anonsen, Ingrid Spielman, Elliot Harper, Alexander MJ Hall, Vivianne J Goosens, Claire Raynaud, Michael Koomey, Nicolas Biais, et al. Global biochemical and structural analysis of the type iv pilus from the gram-positive bacterium streptococcus sanguinis. *Journal of Biological Chemistry*, 294(17):6796–6808, 2019.

[62] Nanhua Yao, T Hesson, M Cable, Z Hong, AD Kwong, HV Le, and Patricia C Weber. Structure of the hepatitis c virus rna helicase domain. *Nature structural biology*, 4(6):463, 1997.

[63] S Spinelli, QZ Liu, PM Alzari, PH Hirel, and RJ Poljak. The three-dimensional structure of the aspartyl protease from the hiv-1 isolate bru. *Biochimie*, 73(11):1391–1396, 1991.

[64] Kristina Bäckbro, Seved Löwgren, Katrin Österlund, Johnson Atepo, Torsten Unge, Johan Hultén, Nicholas M Bonham, Wesley Schaal, Anders Karlén, and Anders Hallberg. Unexpected binding mode of a cyclic sulfamide hiv-1 protease inhibitor. *Journal of medicinal chemistry*, 40(6):898–902, 1997.

[65] Lukasz Lebioda, Boguslaw Stec, John M Brewer, and Ewa Tykarska. Inhibition of enolase: the crystal structures of enolase-calcium (2+)-2-phosphoglycerate and enolase-zinc (2+)-phosphoglycolate complexes at 2.2-. ang. resolution. *Biochemistry*, 30(11):2823–2827, 1991.

[66] Boguslaw Stec and Lukasz Lebioda. Refined structure of yeast apo-enolase at 2.25 å resolution. *Journal of molecular biology*, 211(1):235–248, 1990.

[67] William R Montfort, Kathy M Perry, Eric B Fauman, Janet S Finer-Moore, Gladys F Maley, Larry Hardy, Frank Maley, and Robert M Stroud. Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dump and an anti-folate. *Biochemistry*, 29(30):6964–6977, 1990.

[68] Kathy M Perry, Eric B Fauman, Janet S Finer-Moore, William R Montfort, Gladys F Maley, Frank Maley, and Robert M Stroud. Plastic adaptation toward mutations in proteins: structural comparison of thymidylate synthases. *Proteins: Structure, Function, and Bioinformatics*, 8(4):315–333, 1990.

[69] Daniel M Himmel, S Gourinath, L Reshetnikova, Y Shen, Andreq G Szent-Györgyi, and Carolyn Cohen. Crystallographic findings on the internally uncoupled and near-rigor states of myosin: further insights into the mechanics of the motor. *Proceedings of the National Academy of Sciences*, 99(20):12645–12650, 2002.

[70] R Ian Menz, Andrew GW Leslie, and John E Walker. The structure and nucleotide occupancy of bovine mitochondrial f1-atpase are not influenced by crystallisation at high concentrations of nucleotide. *FEBS letters*, 494(1-2):11–14, 2001.

[71] R Ian Menz, John E Walker, and Andrew GW Leslie. Structure of bovine mitochondrial f1-atpase with nucleotide bound to all three catalytic sites: implications for the mechanism of rotary catalysis. *Cell*, 106(3):331–341, 2001.

[72] Daniel Lim and Natalie CJ Strynadka. Structural basis for the β lactam resistance of pbp2a from methicillin-resistant staphylococcus aureus. *Nature Structural & Molecular Biology*, 9(11):870, 2002.

[73] Andrew L Lovering, Michael C Gretes, Susan S Safadi, Franck Danel, Liza De Castro, Malcolm GP Page, and Natalie CJ Strynadka. Structural insights into the anti-methicillin-resistant staphylococcus aureus (mrsa)

activity of ceftobiprole. *Journal of Biological Chemistry*, 287(38):32096–32102, 2012.

[74] Lisandro H Otero, Alzoray Rojas-Altuve, Leticia I Llarrull, Cesar Carrasco-López, Malika Kumarasiri, Elena Lastochkin, Jennifer Fishovitz, Matthew Dawley, Dusan Hesek, Mijoon Lee, et al. How allosteric control of staphylococcus aureus penicillin binding protein 2a enables methicillin resistance and physiological function. *Proceedings of the National Academy of Sciences*, 110(42):16808–16813, 2013.

[75] M Coles, T Diercks, J Liermann, A Gröger, B Rockel, W Baumeister, KK Koretke, A Lupas, J Peters, and H Kessler. The solution structure of vat-n reveals a 'missing link'in the evolution of complex enzymes from a simple $\beta\alpha\beta\beta$ element. *Current Biology*, 9(20):1158–1168, 1999.

[76] Karthik V Rajasekar, Shuangxi Ji, Rachel J Coulthard, Jon P Ride, Gillian L Reynolds, Peter J Winn, Michael J Wheeler, Eva I Hyde, and Lorna J Smith. Structure of sph (self-incompatibility protein homologue) proteins: a widespread family of small, highly stable, secreted proteins. *Biochemical Journal*, 476(5):809–826, 2019.

[77] Suresh K Vasa, Himanshu Singh, Kristof Grohe, and Rasmus Linser. Assessment of a large enzyme–drug complex by proton-detected solid-state nmr spectroscopy without deuteration. *Angewandte Chemie International Edition*, 58(17):5758–5762, 2019.

[78] Romain Amyot, Yuichi Togashi, and Holger Flechsig. Analyzing fluctuation properties in protein elastic networks with sequence-specific and distance-dependent interactions. *Biomolecules*, 9(10), 2019.

[79] Mingyang Lu, Billy Poon, and Jianpeng Ma. A new method for coarse-grained elastic normal-mode analysis. *Journal of chemical theory and computation*, 2(3):464–471, 2006.

[80] Billy K Poon, Xiaorui Chen, Mingyang Lu, Nand K Vyas, Florante A Quiocho, Qinghua Wang, and Jianpeng Ma. Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-å crystallographic resolution. *Proceedings of the National Academy of Sciences*, 104(19):7869–7874, 2007.

[81] Theodor Wieland and Helmut Fritz. Stufenweiser abbau von peptiden mit hilfe der lossenschen reaktion. *Chemische Berichte*, 86(9):1186–1198, 1953.

[82] Alan J Hoffman and Helmut W Wielandt. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 118–120. World Scientific, 2003.

[83] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, Dec 1912.

[84] Robert Brown. *The miscellaneous botanical works of Robert Brown*, volume 27. Ray society, 1866.

[85] Albert Einstein. On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat. *Annalen der physik*, 17:549–560, 1905.

[86] Paul Langevin. Sur la théorie du mouvement brownien. *CR Acad. Sci. Paris*, 146(530-533):530, 1908.

[87] John Edward Jones. On the determination of molecular fields. ii. from the equation of state of a gas. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 106, pages 463–477. The Royal Society, 1924.

[88] John E Lennard-Jones. Cohesion. *Proceedings of the Physical Society*, 43(5):461, 1931.

[89] Hans C Öttinger. *Stochastic processes in polymeric fluids: tools and examples for developing simulation algorithms*. Springer Science & Business Media, 2012.

[90] Tony W Liu. Flexible polymer chain dynamics and rheological properties in steady flows. *The Journal of Chemical Physics*, 90(10):5826–5842, 1989.

[91] Madan Somasi, Bamin Khomami, Nathanael J Woo, Joe S Hur, and Eric SG Shaqfeh. Brownian dynamics simulations of bead-rod and bead-spring chains: numerical algorithms and coarse-graining issues. *Journal of non-newtonian fluid mechanics*, 108(1-3):227–255, 2002.

[92] Hiroshi Taketomi, Yuzo Ueda, and Nobuhiro Gō. Studies on protein folding, unfolding and fluctuations by computer simulation. *Chemical Biology & Drug Design*, 7(6):445–459, 1975.

[93] Matteo Pasquali and David C Morse. An efficient algorithm for metric correction forces in simulations of linear polymers with constrained bond lengths. *The Journal of chemical physics*, 116(5):1834–1838, 2002.

[94] Donald L Ermak and JA McCammon. Brownian dynamics with hydrodynamic interactions. *The Journal of chemical physics*, 69(4):1352–1360, 1978.

[95] MP Allen. Brownian dynamics simulation of a chemical reaction in solution. *Molecular Physics*, 40(5):1073–1087, 1980.

[96] Scott H Northrup, Stuart A Allison, and J Andrew McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics*, 80(4):1517–1524, 1984.

[97] Jeroen S van Zon and Pieter Rein Ten Wolde. Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space. *The Journal of chemical physics*, 123(23):234910, 2005.

[98] Jeroen S van Zon and Pieter Rein Ten Wolde. Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics. *Physical review letters*, 94(12):128103, 2005.

[99] Steven S Andrews and Dennis Bray. Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical biology*, 1(3):137, 2004.

[100] Nicholas Leioatts, Tod D Romo, and Alan Grossfield. Elastic network models are robust to variations in formalism. *Journal of chemical theory and computation*, 8(7):2424–2434, 2012.

Articles

(1) Analyzing Fluctuation Properties in Protein Elastic Networks with Sequence-Specific and Distance-Dependent Interactions.
Romain Amyot, Yuichi Togashi and Holger Flechsig
Biomolecules 9 (2019), 549.