Using normal mode analysis on protein structural models. How far can we go on our predictions?

Nuria Cirauqui Diaz, Elisa Frezza, Juliette Martin

Short informative title: Normal Mode Analysis on protein models.

Using Normal Mode Analysis on protein structural models. How far can we go on our predictions?

Nuria Cirauqui Diaz[a], Elisa Frezza[b], Juliette Martin[a*]

[a] Université de Lyon, CNRS, UMR 5086 Molecular Microbiology and Structural Biochemistry, IBCP, 7 passage du Vercors, F-69367 Lyon, France

[b]Université de Paris, CiTCoM, CNRS, F-75006 Paris, France

* Corresponding author and lead contact: juliette.martin@ibcp.fr

ACKNOWLEDGMENTS

This project/research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2). We would like to thank Guillaume Launay for useful discussions during the course of this study.

ABSTRACT

Normal Mode Analysis is a fast and inexpensive approach that is largely used to gain insight into functional protein motions, and more recently to create conformations for further computational studies. However, when the protein structure is unknown, the use of computational models is necessary. Here, we analyze the capacity of normal mode analysis in internal coordinate space to predict protein motion, its intrinsic flexibility and atomic displacements, using protein models instead of native structures, and the possibility to use it for model refinement. Our results show that normal mode analysis is quite insensitive to modelling errors, but that calculations are strictly reliable only for very accurate models. Our study also suggests that internal normal mode analysis is a more suitable tool for the improvement of structural models, and for integrating them with experimental data or in other computational techniques, such as protein docking or more refined molecular dynamics simulations.

Keywords

Normal mode analysis; internal coordinates; coarse-grained models; protein flexibility; protein motions; conformational changes; model refinement; structural models.

1. INTRODUCTION

Normal Mode Analysis (NMA) is a computationally inexpensive method extensively used to predict large amplitude motions in proteins, which often relate to biologically relevant functions [1–4]. In many cases, a few low-energy normal modes account for most of the structural differences between two conformational states [5–8]. Moreover, the atomic fluctuations obtained by normal modes match experimental B-factors [9, 10], as well as those calculated based on molecular dynamics simulations [3, 11–14]. For these reasons, NMA is widely applied to study important transition pathways of biomolecules [15, 16], so as to get insights into allosteric pathways [17–19]. It can also be applied to generate input conformations for molecular docking [15, 20–24], and even for structural refinement of X-ray diffraction data [25, 26], small angle X-ray scattering (SAXS) [27] or cryo-electron microscopy (cryo-EM) [28].

NMA can be performed following different approaches and algorithms [29]. There are two main strategies for computing normal modes, depending on the description of the degrees of freedom: Cartesian Coordinate Space (CCS) and Internal Coordinate Space (ICS). The former is the most popular approach and is computationally simpler [30–33]. In this approach, the Cartesian coordinates of all the atoms or a subset of them are used as variables. However, several works showed the advantages of computing NMA in ICS, usually using dihedral angles as variables, though any other internal property (i.e. bonds and valence angles) could be also considered [33–39]. First, normal mode analysis in ICS allows to extend the validity of the harmonic approximation of the conformational energy hypersurface [40]. Second, this approach offers another advantage as it includes a reduced number of variables, which results in a reduced number of modes. In this

way, while for NMA in CCS most of the relevant protein motions are located within the first non-trivial 10-20 modes [32], this number can be reduced to 5 in NMA in ICS [36] and fewer modes contribute to the conformational change [35, 36, 39]. Finally, this method provides a third, intrinsic advantage: the atom connectivity is conserved, allowing to model larger conformational changes. Indeed, our previous studies showed the better performance of NMA in ICS in generating protein conformations of the bound state, when starting from the unbound structure [36].

NMA is typically used in conjunction with coarse-grained protein models in which the pseudoatoms are connected by springs if their distance is closer than a chosen cutoff distance. This strategy allows reducing the computational cost and using the starting structure as a reference for calculations [33, 41]. Coarse-grained models represent proteins in a broad range of particle size, from a single particle per protein [42, 43], to several particles per residue [35, 44–47]; but the most common representation considers merely the alpha carbons [33]. Nonetheless, all-atom calculations are also being applied, which often allow for more accurate frequency calculation [29]. It has already been proven that both the protein representation and the variable system influence the number of low-energy modes needed to describe a relevant conformational change [29, 35].

Several studies have demonstrated the conservation of large amplitude motions (and therefore low-energy modes) among homologous proteins [48–52], and even within a set of proteins that possess the same fold despite low sequence identity [48, 53, 54]. Low energy modes have thus proven to be robust to sequence variations [14, 53, 55]. Along with this observation comes the question of whether or not normal mode analysis could be applied on protein structural models. Due to the lack of experimental data, it often happens that no structure is available for the protein of interest, and molecular modeling is the only available tool for predicting it. When working

with molecular dynamics simulations, protein models usually deviate from the starting configuration, and therefore from the native structure, partly due to the limited accuracy of force fields [56, 57]. On the other hand, normal mode calculations were presented as more robust [53, 58], although no extensive analysis has been done in this regard. Hollup and co-worker showed, using computer-generated models of three proteins, that intrinsic dynamics are maintained when the main protein architecture is conserved [53]. Nonetheless, to our knowledge, no study at large scale has been conducted till date to assess the reliability of mode calculation when we are working with models instead of with the native structure.

To fill this gap, we herein thoroughly analyze the reliability of normal mode calculations performed on computational models. Due to its proven better performance, we used NMA in internal coordinate space to carry out our study, using the same protein representation as in our previous works [35, 36]. Nonetheless, to compare with standard approaches, we also computed NMA in CCS considering only Cα atoms. We performed this work on a large scale, taking advantage of the available models of the Critical Assessment of protein Structure Prediction (CASP) experiment [59, 60], selecting 99 native structures and 420 models with values of Root Mean Square Deviation (RMSD) to the native structure up to 5 Å. We assessed the accuracy of normal modes by analyzing the mode overlap, the global flexibility and residue fluctuations. Moreover, we checked whether or not we could use NMA as a tool to improve protein models. Our results show that, for models with initial RMSD to the native structure smaller than 3 Å, results are reliable in terms of both motion prediction and flexibility, while the reliability decreases for less accurate models. Nonetheless, up to 5 Å, mode prediction could be still used to get a qualitative idea of the native motions, with most Root Weighted Square Inner Product (RWSIP) values

native/model higher than 0.5. The prediction of flexible regions is accurate for models until 4 Å RMSD. Finally, we observed that NMA in internal coordinates was able to better refine models than NMA in CCS, while preserving structure valence.

2. METHODS

2.1 Data set

The two last experiments available when this work started, CASP11 and CASP12, were used as a source of structural models [60, 77]. In CASP, proteins are divided by domains, defined as "structurally compact evolutionary modules in proteins that serve as the basic units of folding" [59]. On multi-domain structures, the lowest modes usually correspond to motions between domains. We thus excluded multi-domain targets from our analysis, in order to focus on intra-domain motions. Incomplete models (i.e. not presenting all residues resolved in the native structure) were discarded. We selected only models with Cα RMSD to the native structure lower than 5 Å, because above that value comparisons will tend not to be meaningful. After those filters, we finished with a very unequal number of models per target. In order not to enrich our results by a particular protein; we selected a maximum of 5 models per target. For that, we performed clustering with the gromos method of the *g_cluster* program of the gromacs v5.1 package [78], so as to obtain 5 clusters at the maximum and select the center of each cluster. The final benchmark comprises 99 native structures and a total of 420 models. They are here named according to their code in the respective CASP experiment.

2.1.1 Classification of native structures by properties

The classification by oligomerization state was based on the biological subunit of the PDB entry, indicated on the CASP database. When this entry was not noted in CASP, we looked for it based on the protein sequence, and checked that both structures were similar, though some minimal differences in loop refinement were found. Secondary structure was assigned using DSSP [79, 80]. Residues with no DSSP assignment were classified loops. Proteins with more loop residues than the 3rd quartile of the distribution were classified as loop-rich. Proteins with more than 5 times α-helix residues than β-sheet residues were classified as α-rich. Proteins with more than 5 times β-sheet residues than α-helix residues were classified as β-rich.

2.2 Internal Normal Mode Calculation

Normal mode analysis (NMA) is a technique to investigate the vibrational motion of a harmonic oscillating system around a minimum energy conformation. NMA are usually computed using Cartesian coordinate space [81, 82]. In our study, we computed NMA in internal coordinate space (iNMA) since this approach allows to extend the validity of the harmonic approximation of the conformational energy hypersurface [40]. Under a harmonic approximation in iNMA, Hessian and kinetic energy matrices can be diagonalized, allowing the analytic solution of the Lagrangian equations of motion. The eigenvectors of this matrix are the normal modes, and the eigenvalues

are related to the squares of the associated frequencies. The protein main conformational changes are usually represented by a combination of some of the lowest frequency modes, but the amplitude of this motion is not defined because it depends on several conditions such as temperature.

2.2.1 Calculation protocol

The calculation of internal normal modes was performed using a home-made program, which has been described and evaluated in our previous studies [35, 36]. Briefly, an anisotropic elastic network model was used, with a modified Zacharias representation of the proteins [35, 36, 47]. In this coarse-grained model, N, C′, and Cα are considered explicitly (rather than using a Cα pseudoatom), whereas the side chains (with the exception of Gly) are represented by either one or two pseudoatoms (SC1, SC2), each representing groups of atoms within the side chain [36]. As internal variables we considered merely the backbone dihedral angles φ and ψ, as they usually change significantly more than the peptide group dihedral ω. NMA in Cartesian coordinates was performed using the Bio3D program [83], with Cα coordinates as variables (cNMA). For sake of comparison, we used similar parameters as those used for internal NMA (iNMA). In other words, an anisotropic elastic network model (ENM) [41] was built considering a fixed force of 0.6 kcal mol^{-1} for atoms under a cutoff of 15 Å.

2.2.2 Conversion from internal coordinates to Cartesian ones

To allow a better comparison with cNMA, normal modes calculated in iNMA were converted to the Cartesian space. This conversion was performed considering all modes, using the Taylor expansion of the Cartesian coordinates to calculate atom displacements [35]. Afterwards, the orthogonalization of the normal modes was achieved with the Gram-Schmidt process [84], via the gramSchmidt function of the pracma v1.9.9 R package (https://www.rdocumentation.org/packages/pracma).

2.3 Evaluation of similarity of Normal Mode calculations

2.3.1 Prediction of protein flexibility

The intrinsic protein flexibility (IFlex) was predicted using the following equation [15]:

$$IFlex = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{w_i^2}} \quad (1)$$

where N is the number of non-trivial modes ($3n$-6 for cNMA and $2n$-1 for iNMA, for a protein of n residues) and w_i is the eigenvalue associated to mode i.

The weighted difference in IFlex values between a structural model and the corresponding native structure (ΔIFlex) was calculated using the following equation:

$$\Delta IFlex = \left(\frac{IFlex_i - IFlex_j}{IFlex_j} \right) * 100 \quad (2)$$

where $IFlex_i$ and $IFlex_j$ are the IFlex values of model i and native structure j, respectively. Frequency units are expressed in 10THz.

2.3.2 Root Mean Square Inner Product (RMSIP) and Root Weighted Square Inner Product (RWSIP)

The RMSIP [63] is a measure of the similarity between two sets of normal modes or principal components of the covariance matrix, defined as:

$$RMSIP = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(U_i \bullet V_j\right)^2}{N}} \quad (3)$$

where U_i and V_j are the set of eigenvectors of the NMA of model (i) and native structure (j), respectively.

The RWSIP [63] is a similar measure, which also takes into account the eigenvalues. It is defined as:

$$RWSIP = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i v_j \left(U_i \bullet V_j\right)^2}{\sum_{i=1}^{N} u_i v_j}} \quad (4)$$

where u_i and v_i are the eigenvalues of the covariance matrix corresponding to the eigenvectors U_i and V_j, for model (*i*) and native structure (*j*), respectively.

Unless otherwise stated, according to common practice, we considered the first 10 non-trivial modes (N=10) for RMSIP calculation. For iNMA, the RMSIP/RWSIP was computed after conversion to the Cartesian Coordinate Space.

2.3.3 Overlap

The overlap (O_i) between a given eigenvector *i* and the difference vector between a model and its corresponding native structure was calculated using the inner product:

$$O_i = V \cdot U_i \quad (5)$$

where V is the difference vector between the initial structures native/model, and U_i is a particular eigenvector of the model (*i*).

2.3.4 Root Mean Square Fluctuation (RMSF)

The mean square fluctuation of the atomic positions (RMSF) along the internal normal modes was performed as described previously, after conversion to the Cartesian Coordinate space [35]. For cNMA, the calculation was performed with the *fluct.nma* tool of the Bio3D package [83]. From

RMSF calculations, temperature factors (B-factors) can be obtained by the following equation and compared with experimental ones [85]:

$$B = \frac{8\pi^2}{3} RMSF^2 \quad (6)$$

In X-ray experiments, B-factors (B) are usually defined as a measure of the spatial fluctuations of atoms around their average position, and their motion is described as an isotropic Gaussian distribution of displacements about this average position [86].

2.4 Generation of modified conformations by NMA

We generated modified conformations of native structures and structural models by applying a given amplitude and phase to them using normal modes. In order to study larger conformational changes, we used mode amplitudes beyond those corresponding to room temperature (termed scaled amplitudes), as already performed by others and us [36, 87]. To note, the use of scaled amplitudes has shown to relate better with experimental B-factors [87].

In iNMA, we used separately the 5 first lowest frequency modes in each direction for generating sets of modified conformations [36]. By increasing the amplitude, we obtained conformations differing from the initial structure with increasing RMSD values (0.5 to 5 Å) (see Figure S4). For cNMA, we used the Bio3D procedure (magnitude value of 50, step of 1) [83]. Structures with steric clashes (as defined later on) were removed. For the analysis of model refinement using NMA, we selected, from the pool of generated conformations, the one closest to the native structure.

2.4.1 Clash parameter

The increase of the modified amplitude may, at some point, result in the creation of clashes, or steric overlaps, on the generated conformations. The clash parameter between a pair of pseudoatoms *i* and *j* is defined as [36]:

$$clash_{ij} = r_{p,i} + r_{p,j} - d_{ij} - c \quad (7)$$

where d_{ij} is the distance between atoms *i* and *j*, $r_{p,i}$ and $r_{p,j}$ are the radii for the pseudoatoms *i* and *j*, respectively, and *c* is a cutoff for the Van der Waals overlap. For iNMA, backbone atoms Cα, N and C were considered, and a value of 1.5 Å was used for *c*. For cNMA, as we had only Cα coordinates, *c* was set to a value of 1 Å. We selected these values to still allow for some soft clashes but ensuring models that could be corrected by minimization. Positive values indicate the presence of strong steric overlaps, and therefore conformations presenting positive $clash_{ij}$ values were discarded.

2.4.2 Structural deformation by bond stretching

As it is well-known that cNMA deforms structures, for example by an overstretching of backbone bonds, we calculated the RMSD of the virtual $C\alpha_i$-$C\alpha_{i+1}$ bonds as [36]:

$$RMSD_{C\alpha_i - C\alpha_{i+1}} = \sqrt{\frac{\sum_{i=1}^{N_d} (d_{C\alpha_i - C\alpha_{i+1}} - d^0_{C\alpha_i - C\alpha_{i+1}})^2}{N_d}} \quad (8)$$

where N_d represents the number of $C\alpha_i$-$C\alpha_{i+1}$ virtual bonds and d and d^0 are, respectively, the length of each virtual bond in the generated conformation by NMA and in the initial structure.

2.4.3 Root Mean Square Deviation (RMSD) calculation

All structural fit and all RMSD calculations between two conformations were performed using the McLachlan algorithm [88] as implemented in the program ProFit v3.1 [89], considering only the backbone atoms Cα, C and N. For modified conformations generated by cNMA, only Cα atoms were used.

2.5 Evaluation of Model Quality

As an alternative to RMSD, which requires the knowledge of the native structure, model quality was assessed with QMEAN [90], a well-established approach to estimate model quality in the absence of the native structure, based on a linear combination of six structural descriptors. Four descriptors are statistical and derived from known structures (solvent accessibility, backbone geometry, inter-atomic packing), and two descriptors evaluate the consistency of structural features with sequence-based predictions. Qmean6 score was computed by interactive access to the QMEAN server [91]. The score ranges from 0 (low quality) to 1 (high quality).

2.6 Statistical approach

All statistical treatment was performed using the R statistical environment [92]. The Spearman correlation coefficient (ρ) was used for data with non gaussian distribution [93]. Distributions were compared using the Wilcoxon test (two samples)[94] and Kruskal-Wallis test (more than two samples) [95]. The area under the ROC curve (AUC) was computed with the pROC package [96]. The AUC were used for (i) quantifying the separation of IFlex distributions between two groups of proteins with different characteristics and (ii) quantifying the quality of the prediction of flexible regions using NMA-derived RMSF. In the latter case, in each native structure, residues with high experimental B-factor (value higher than the average + 1.5 standard deviation) were classified as flexible residues and NMA-derived RMSF were used to predict these flexible residues. AUC values vary between 0 (wrong prediction for all residues) and 1 (perfect prediction), where 0.5 denotes a random prediction. The rationale to use AUC here instead of Spearman correlation is that the goal is to identify peaks in flexibility profiles, whereas correlation is influenced by the precise ranking of all residues, within and outside of the peaks.

3. RESULTS

To build up our dataset, we selected 99 targets of the Critical Assessment of protein Structure Prediction (CASP) experiment, with its 99 native structures and a total of 420 models (see section Methods for more details) [61]. To clarify the naming scheme, a target means a particular

protein, the native structure (NS) is the experimental structure, and the model is a prediction of its tridimensional structure. CASP models have the advantage to cover a broad range of applied methodologies, from *de novo* to comparative modeling, from server to human generation. Moreover, the native structure is available and the quality of the models has been evaluated using standardized measurements [62]. Also, CASP targets tend to be selected due to their difficulty, mostly in terms of absence of near homologues, making it an interesting source for studying normal mode applications outside straightforward cases [59]. Finally, they include all kinds of proteins, with different secondary structures, either in a monomeric form or in complex with another biomolecule, representing an unbiased set for very general applications. We did not consider models with initial RMSD to their native structure higher than 5 Å, as we wanted to focus in cases with no strong mistakes in the modeling. We recall that NMA was computed using internal coordinates (hereby called iNMA), which preserves protein internal geometry, and in Cartesian coordinates using a Cα model (hereby called cNMA) for comparison, as this is still the most popular approach.

3.1 Motions predicted by NMA are not strongly affected by modeling errors

According to Fuglebakk and co-workers, the Root Mean Square Inner Product (RMSIP) of the normal modes, defined in Equation (3) in subsection 2.3.2, so as the similar Root Weighted Square Inner Product (RWSIP), defined in Equation (4) in subsection 2.3.2, are some of the best known metrics to compare sets of protein motions [63]. Here, we used these quantities to evaluate whether or not motions calculated from structural models resembled those predicted for their

native structures. It is important to stress that, in the case of iNMA, as several sets of torsions can lead to the same final conformation, and, moreover, modes are not strictly orthogonal in the internal space, it is mandatory to convert results from the internal to the Cartesian space [35] before calculating the RMSIP/RWSIP. In Figure 1, we present the results for RWSIP defined as the sum of the overlaps between the first 10 modes, weighted by the eigenvalues (see the section Methods for more details). We used this number of modes since it is well known that they capture most of the relevant motions in both iNMA and cNMA [36]. Globally, NMA is able to capture the main protein motions using structural models, in both ICS and CCS: average RWSIP is equal to 0.69 for iNMA and 0.73 for cNMA.

As shown in Figure 1, there is a clear trend of decreasing RWSIP values as RMSD increases. In other words, very accurate models, meaning those having less than 2 Å RMSD to their native structure, always presented high RWSIP values (RWSIP > 0.65 and > 0.80 for iNMA and cNMA respectively). This trend is less clear for iNMA (Figure 1A) than for cNMA (Figure 1B), suggesting that an accurate model in terms of RMSD to the native does not ensure a correct (i.e. similar to the native) calculation of protein motions in the internal space. These results indicate that iNMA can be more sensitive to local details, as already observed in our previous studies [35, 36]. Interestingly, for models of moderate quality (RMSD greater than 2.5 Å), iNMA produces fewer cases with low RWSIP values (<0.5) than cMNA. This suggests that iNMA could be more tolerant than cNMA to model inaccuracy for medium quality models.

It is well-known that for iNMA only few modes are sufficient to model protein motions with respect to cNMA, as already observed for the prediction of large conformational changes in our previous studies [35, 36]. In fact, when only the first 5 lowest modes are considered, a more

pronounced decrease in RMSIP/RWSIP values is observed for cNMA than iNMA, with a larger number of outliers (see Figure S1). Finally, our results were not influenced by the existence of gaps in the native structure and in models (*i.e.* missing residue(s) inside a given chain of the structure), which occured in 25% of the proteins under investigation (see Figure S2).

RWSIP results indicate that protein motions predicted from models are similar to those predicted from the native structures. We wanted to check if this means that they are sampling a similar conformational space, that is, if we could obtain similar conformations by applying normal modes on either a native structure or a model. Indeed, for rigid proteins which at physiological conditions would not sample more than 1 Å RMSD, we cannot expect to obtain similar and meaningful conformations working with models with initial RMSD to the native higher than this value.

Then, we selected four of the most flexible proteins in the dataset (based on the IFlex value as defined in the section Methods and analyzed below) and created normal mode modified conformations for native structures and models with RMSD about 3 Å, as explained in the section Methods. Figure 2 shows these examples of conformational sampling along the first normal mode obtained by iNMA and cNMA. For these proteins, both iNMA and cNMA predict similar motions either from the native or from the model. These results encourage using protein models in the case of flexible proteins for generating conformations, for example for protein docking. This is important since the conformational space of very flexible proteins is more difficult to explore by other approaches, such as standard molecular dynamics. We would like to stress that unrealistic overstretching of the flexible protein regions was observed for cNMA (for example targets T0783-D1 and T0805-D1 in Figure 2).

3.2 Prediction of protein intrinsic flexibility is not affected by NMA approaches

Among the several methods published to estimate protein flexibility from NMA [15, 64, 65], Dobbins and co-workers proposed a function to predict the intrinsic flexibility of a protein based on the eigenvalues (herein termed IFlex, see Equation (1) in the subsection 2.3.1), and they observed an agreement with the observed RMSD between two functional protein states (bound-unbound) [15]. Here, we used the same approach to compare the prediction of protein intrinsic flexibility obtained using iNMA ($IFlex^i$) and cNMA ($IFlex^c$) for native structures (NS). Figure 3 summarizes our results. Proteins are similarly ranked by predicted flexibility independently of the approach used for NMA calculation (Spearman correlation coefficient ρ=0.86 for IFlex values calculated with either iNMA or cNMA).

Some proteins have much higher values of IFlex compared to others. The most outlier target, T0865-D1, corresponds to the C-terminal coiled-coil domain of the SH3 domain-containing kinase-binding protein 1 from human (PDB code 2N64) [66]. Therefore, its high IFlex value is expected, as coiled-coils are known for their high flexibility. Protein target T0805-D1 is a bacterial nitroreductase, likely oligomeric (see Table S1). Protein T0865-D1 is also oligomeric (see Table S1). As we performed NMA on a single protomer, and not on the oligomer, the missing interchain interactions and spatial constraints of the second protomer can easily justify the predicted high flexibility. Target T0820-D2 is an uncharacterized protein identified in marine metagenomic data, whose oligomer status is unknown.

To better explore the obtained differences in predicted flexibility, we classified our proteins based on some selected characteristics (presence of loops, missing regions, secondary structure and oligomeric state of the native structure; Table S1). No differences were observed in IFlex values related to the presence of gaps (i.e. missing regions) or secondary structure content (Table S2). However, a significant increase in IFlex values was obtained for oligomeric proteins and, when the calculations are done with iNMA, also for loop-rich proteins. To evaluate the separation between groups, we used AUC (area under the ROC curve) values, as explained in the section Methods. In all cases, the separation among groups was more marked for iNMA. The same differences among subgroups were obtained when the flexibility was predicted using the models instead of the native structures (Table S3), with greater significance in the case of iNMA. Therefore, iNMA is able to better capture flexibility differences related to structural aspects, both from the native structure and from protein models.

3.3 Protein intrinsic flexibility can be well captured by structural models

To evaluate our capability to predict protein intrinsic flexibility using structural models, we used IFlex to analyze the differences between intrinsic flexibility predicted for a given native structure and its models. To do so, we computed the weighted difference in IFlex values, calculated as the difference between IFlex computed for each model and their native structure divided by the value obtained for the native structure and multiplied by 100 (ΔIFlex) (see Equation (2) in the subsection 2.3.1). Positive values of ΔIFlex denote models which are more flexible than the corresponding native structure, while negative values denote the opposite. The results are shown

in Figure 4 as a function of the initial RMSD model/native. The median values of ΔIFlex are 0.86 for iNMA and 0.31 for cNMA, and the means are 3.70 and 1.87 respectively. Those positive values indicate a slight tendency to create models with higher flexibility than the native protein structure. Outliers with negative values (i.e. under-estimated flexibility) correspond to very flexible proteins highlighted in Figure 3.

Even more relevant is the fact that for models presenting up to 3 Å RMSD to the native structure, 90% of the points are within the interval [-10,10] of ΔIFlex with both approaches, meaning an error lower than 10% compared to the native structure IFlex (see Figure 4). This shows that the calculation of IFlex is reliable in the case of small modelling errors. Even for models with RMSD as high as 5 Å, 90% of the data points are within the [-19,26] interval of ΔIFlex for iNMA and [-20,38] for cNMA, showing that the calculation on structural models approximates well the flexibility of the native structure. To note, for models with RMSD higher than 3 Å, we observe that the ΔIFlex values are less spread (i.e. narrower empirical 90% interval) in iNMA than cNMA, suggesting that iNMA is more tolerant to modelling errors in the case of models with medium quality.

Noteworthy, there is a good correlation between values of ΔIFlex calculated either by cNMA or iNMA ($\rho = 0.77$, data not shown). In addition, there is also a good correlation between IFlex predicted from native structures and IFlex predicted from models, with ρ equal to 0.78 for iNMA and 0.80 for cNMA (see Figure S3). The above results suggest that structural errors have a similar impact on the flexibility prediction in spite of the NMA approach used.

3.4 Flexible regions can be predicted from models

The atomic displacements predicted by NMA have been shown to well correlate with experimental B-factors [41, 67]. In this study, we also verified if this correlation is still valid for protein models. To do so, we computed the Root Mean Square Fluctuation (RMSF) along normal modes for each model, and compared it to the experimental B-factors of its respective native structure. We used AUC computation, as explained in the section Methods, to predict regions with high experimental B-factors using NMA-derived RMSF.

The results are shown in Figure 5. Panel 5A illustrates the three comparisons that have been carried out: B-factors versus RMSF of native structures (dashed red line), B-factors versus RMSF of models (dashed blue line), and RMSF of native structures versus RMSF of models (dashed green line). As shown in Figure 5A, flexible regions are predicted with high accuracy from NMA-derived RMSF computed on native structures: average AUC is equal to 0.83 for both iNMA and cNMA. The small discrepancy observed could moreover be explained by the fact that B-factors are heavily influenced by non-thermal contributions [68, 69] and do not reflect only the structural flexibility of proteins [63]. When structural models are used instead of native structures, the prediction remains accurate: average AUC is equal to 0.79 for iNMA and 0.80 for cNMA. If we now compare the RMSF values between native structures and models, we observe very high prediction success: average AUC is equal to 0.92 for iNMA and 0.93 for cNMA.

As expected, the prediction of flexible regions from models (dashed blue line in Figure 5A) is impacted by the model quality, see Figure 5B. We observe a clear decrease in prediction accuracy for models with RMSD greater than 4 Å, using both NMA approaches. Panel 5C shows an example of a pair native/model, where we can observe how flexible regions are well identified

using NMA, despite the medium accuracy of the model (RMSD to the native greater than 4 Å). Our results suggest that NMA on structural models is capable of capturing the flexible regions of the native structure independently of the approach used.

3.5 iNMA allows model refinement preserving the protein valence structure

Although NMA has been mostly used for getting insights into the main protein motions, it is becoming more and more relevant in the generation of different conformations of a protein [70], and it has also been used for improving protein models [71, 72]. Here, we aimed to analyze how close we could move a model to its native structure by applying normal modes. To this aim, for each model, we generated a set of conformations by applying a subset of normal modes computed by iNMA and cNMA, using different amplitudes (see the section Methods for more details). The modified conformation presenting the minimum RMSD to the native was selected ($RMSD_{NMA}$), and compared to the initial RMSD between the model and its native structure (see Figure 6A-B). In Figure 6, to highlight the change in protein valence structure, the dots are colored based on the RMSD of $C\alpha_i$-$C\alpha_{i+1}$ virtual bonds between the new conformation and the initial model (see the section Methods for details). For very accurate models (low RMSD values), the improvement in terms of RMSD to the native is very small for both iNMA and cNMA. On the contrary, as the initial RMSD model/native increases, iNMA outperforms cNMA, with lower $RMSD_{NMA}$ values while preserving protein connectivity.

To quantify the best improvement in terms of native-likeness using a statistical approach, for each model we also computed the fractional RMSD, ΔRMSD%, defined as ΔRMSD% =

100(RMSD-$RMSD_{NMA}$)/RMSD (Figure 6C-D). Large values of fractional RMSD mean that the model approaches the native structure by applying a specific normal mode and amplitude. By comparing the histograms in Figures 6C and 6D, one can see that iNMA represents a promising approach for model refinement. In fact, the mean ΔRMSD% value is 9.4 and 5.6 for iNMA and cNMA, respectively while the median is equal to 7.0 and 3.9, respectively. Thus, iNMA is capable of refining models by reducing the initial RMSD value to the native by 10% on an average (red line in Figure 6C). It is worth noticing that these results were obtained by considering only one mode, and therefore expected to improve when two or more modes are combined [36].

Not only does iNMA improve the predictions compared to cNMA, but the latter can dramatically overextend flexible regions [36], by increasing $C\alpha_i$–$C\alpha_{i+1}$ virtual bonds to unrealistic values. It is consequently understandable that when structures derived from cNMA modes are used in other computational approaches or in combination with experimental data [73–75], further structural refinement steps are necessary that can also lose most of the conformational change obtained by cNMA. On the contrary, iNMA allows to preserve the protein valence structure and it is a better candidate for integrative modelling and for improving computational models.

4. DISCUSSION

Internal normal mode analysis is a fast computational approach to predict large conformational changes without modifying protein connectivity. In this study, we investigated the capability of iNMA to predict protein motion, its intrinsic flexibility, atomic displacements

and conformational changes using protein models instead of native structures. This step is necessary if we would like to use structural models in computational studies when its native structure is unknown. To assess this approach, we compared the results obtained with either iNMA or cNMA.

Our results show that, when working with models of medium quality (i.e. expected RMSD to the native structure lower than 5 Å), normal mode predictions do not diverge substantially with those of the native structure. The predicted motions do resemble (RWSIP above 0.5) and normal modes are able to locate the most flexible regions of the protein. Therefore, in most cases, we could use NMA to generate modified conformations of computational models. Nonetheless, calculations are only strictly reliable for very accurate models (RMSD to the native up to 2-3 Å). Above that cutoff, the biggest errors can be obtained in the calculation of intrinsic flexibility, mostly for very flexible proteins.

As expected, we observed that iNMA is more sensitive to structural characteristics related to flexibility, ranking as more flexible those proteins whose native structure corresponds to an oligomeric state, and also those presenting very long loops. Moreover, it is shown as a better approach for creating modified conformations and for refining protein models while preserving the protein valence structure. Also, when applied for model refinement, it was able to get closer to the native structure than cNMA using only a single mode. These results are very promising and we expect that modelling refinement will be further improved by combining at least 2 modes, as already observed in our previous study on protein conformational changes [36].

Another important point is how to evaluate the expected accuracy of a model. There are several ways to assess it. For example, sequence identity template/target for homology

modelling, or confidence scores in *de novo* modeling, which depend on the significance of threading template alignments and the convergence parameters. Moreover, some approaches are being developed with the aim of predicting model accuracy based on quality assessment tools, such as QMEANDisCo [76]. While those methods are promising, they are still under active development. Here, we observed no clear relationship between the models' QMEAN scores and RWSIP values obtained from the comparison of NMA computed on models and native structures (see the section Methods for the details and Figure S5). This means that a good score for the nativeness of a model does not ensure a native-like calculation of normal modes. Nonetheless, taking into account the expected RMSD of a model and the results presented here, it is possible to get an idea of the reliability of NMA performed on that model, and therefore the usage that we can make of any information coming from them.

From another perspective, our study also contributes to a better understanding of protein dynamics. Earlier studies based on NMA computed on either proteins with similar fold or protein models (Tiwari and Reuter 2018) helped to better understand how protein intrinsic dynamics is guided by its global architecture, but it is also fine-tuned for a specific function. Our results also support this picture: small modelling errors (up to 2-3 Å) do not strongly affect either protein flexibility or its dynamics, although some differences can be observed, which might be relevant in structures differing before the application of NMA by more than 3 Å.

In summary, the outcome of our study suggests that iNMA is a more suitable tool than cNMA for improving structural models, and for integrating them with either experimental data (i.e. SAXs or Cryo-EM) or other computational techniques, such as protein docking. Nonetheless, doing so without prior knowledge of the native structure and/or its expected

conformational changes is still challenging. Some advances have been made in this field, such as the reduction in the number of modes that needs to be considered (Frezza and Lavery, 2019). To further enhance iNMA as a predictive tool, we plan to conduct an analysis on a range of biological and physicochemical properties, searching for potential indicators pointing to relevant modes and movements.

REFERENCES

1. Bahar I, Lezon TR, Yang L-W, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys 2010;39:23–42.
2. Kim MH, Lee BH, Kim MK. Robust elastic network model: A general modeling for precise understanding of protein dynamics. J Struct Biol 2015;190(3):338–347.
3. Rueda M, Chacón P, Orozco M. Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. Structure 2007;15(5):565–575.
4. Yang L, Song G, Jernigan RL. How well can we understand large-scale protein motions using normal modes of elastic network models? Biophys J 2007;93(3):920–929.
5. Hinsen K, Thomas A, Field MJ. Analysis of domain motions in large proteins. Proteins 1999;34(3):369–382.
6. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 2002;48(4):682–695.
7. Marques O, Sanejouand YH. Hinge-bending motion in citrate synthase arising from normal mode calculations. Proteins 1995;23(4):557–560.
8. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng 2001;14(1):1–6.
9. Bahar I, Atilgan AR, Demirel MC, Erman B. Vibrational Dynamics of Folded Proteins: Significance of Slow

and Fast Motions in Relation to Function and Stability. Phys Rev Lett 1998;80(12):2733–2736.

10. Keskin O, Jernigan RL, Bahar I. Proteins with similar architecture exhibit similar large-scale dynamic behavior. Biophys J 2000;78(4):2093–2106.
11. Micheletti C, Carloni P, Maritan A. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. Proteins 2004;55(3):635–645.
12. Pontiggia F, Colombo G, Micheletti C, Orland H. Anharmonicity and self-similarity of the free energy landscape of protein G. Phys Rev Lett 2007;98(4):048102.
13. Skjaerven L, Martinez A, Reuter N. Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. Proteins 2011;79(1):232–243.
14. Yang L, Song G, Carriquiry A, Jernigan RL. Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. Struct Lond Engl 1993 2008;16(2):321–330.
15. Dobbins SE, Lesk VI, Sternberg MJE. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. Proc Natl Acad Sci 2008;105(30):10390–10395.
16. Zheng W, Brooks BR, Hummer G. Protein conformational transitions explored by mixed elastic network models. Proteins 2007;69(1):43–57.
17. Fanelli F, Felline A, Raimondi F, Seeber M. Structure network analysis to gain insights into GPCR function. Biochem Soc Trans 2016;44(2):613–618.
18. Rodgers TL, Townsend PD, Burnell D, Jones ML, Richards SA, McLeish TCB, Pohl E, Wilson MR, Cann MJ. Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors. PLoS Biol 2013;11(9):e1001651.
19. Seckler JM, Leioatts N, Miao H, Grossfield A. The interplay of structure and dynamics: insights from a survey of HIV-1 reverse transcriptase crystal structures. Proteins 2013;81(10):1792–1801.
20. Cavasotto CN. Normal mode-based approaches in receptor ensemble docking. Methods Mol Biol Clifton NJ 2012;819:157–168.
21. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant

normal modes. J Am Chem Soc 2005;127(26):9632–9640.

22. Dietzen M, Zotenko E, Hildebrandt A, Lengauer T. On the applicability of elastic network normal modes in small-molecule docking. J Chem Inf Model 2012;52(3):844–856.
23. Jiménez-García B, Roel-Touris J, Romero-Durana M, Vidal M, Jiménez-González D, Fernández-Recio J. LightDock: a new multi-scale approach to protein–protein docking. Bioinformatics 2018;34(1):49–55.
24. Lindahl E, Delarue M. Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. Nucleic Acids Res 2005;33(14):4496–4506.
25. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J 2002;83(2):723–732.
26. Poon BK, Chen X, Lu M, Vyas NK, Quiocho FA, Wang Q, Ma J. Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-A crystallographic resolution. Proc Natl Acad Sci 2007;104(19):7869–7874.
27. Panjkovich A, I. Svergun D. Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis. Phys Chem Chem Phys 2016;18(8):5707–5719.
28. Brink J, Ludtke SJ, Kong Y, Wakil SJ, Ma J, Chiu W. Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. Struct Lond Engl 1993 2004;12(2):185–191.
29. Dykeman EC, Sankey OF. Normal mode analysis and applications in biological physics. J Phys Condens Matter Inst Phys J 2010;22(42):423202.
30. Bahar I, Rader AJ. Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 2005;15(5):586–592.
31. Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev 2010;110(3):1463–1497.
32. Stein A, Rueda M, Panjkovich A, Orozco M, Aloy P. A Systematic Study of the Energetics Involved in Structural Changes upon Association and Connectivity in Protein Interaction Networks. Structure 2011;19(6):881–889.
33. Wako H, Endo S. Normal mode analysis as a method to derive protein dynamics information from the Protein

Data Bank. Biophys Rev 2017;9(6):877–893.

34. Bray JK, Weiss DR, Levitt M. Optimized torsion-angle normal modes reproduce conformational changes more accurately than cartesian modes. Biophys J 2011;101(12):2966–2969.

35. Frezza E, Lavery R. Internal Normal Mode Analysis (iNMA) Applied to Protein Conformational Flexibility. J Chem Theory Comput 2015;11(11):5503–5512.

36. Frezza E, Lavery R. Internal Coordinate Normal Mode Analysis: A Strategy To Predict Protein Conformational Transitions. J Phys Chem B 2019;123(6):1294–1301.

37. Hoffmann A, Grudinin S. NOLB: Nonlinear Rigid Block Normal-Mode Analysis Method. J Chem Theory Comput 2017;13(5):2123–2134.

38. López-Blanco JR, Aliaga JI, Quintana-Ortí ES, Chacón P. iMODS: internal coordinates normal mode analysis server. Nucleic Acids Res 2014;42(Web Server issue):W271-276.

39. Mendez R, Bastolla U. Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. Phys Rev Lett 2010;104(22):228103.

40. Kitao A, Hayward S, G N. Comparison of normal mode analyses on a small globular protein in dihedral angle space and Cartesian coordinate space. Biophys Chem 1994;52(2):107–114.

41. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 2001;80(1):505–515.

42. Hills RD, Brooks CL. Insights from Coarse-Grained Gō Models for Protein Folding and Dynamics. Int J Mol Sci 2009;10(3):889–905.

43. Peeters K, Taormina A. Group theory of icosahedral virus capsid vibrations: a top-down approach. J Theor Biol 2009;256(4):607–624.

44. Gopal SM, Mukherjee S, Cheng Y-M, Feig M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. Proteins Struct Funct Bioinforma 2010;78(5):1266–1281.

45. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. J Chem Theory Comput 2008;4(5):819–834.

46. Sterpone F, Melchionna S, Tuffery P, Pasquali S, Mousseau N, Cragnolini T, Chebaro Y, St-Pierre J-F, Kalimeri M, Barducci A, Laurin Y, Tek A, Baaden M, Nguyen PH, Derreumaux P. The OPEP protein model:

from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. Chem Soc Rev 2014;43(13):4871–4893.

47. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci Publ Protein Soc 2003;12(6):1271–1282.

48. Cirauqui N, Abriata LA, Goot FG van der, Dal Peraro M. Structural, physicochemical and dynamic features conserved within the aerolysin pore-forming toxin family. Sci Rep 2017;7(1):13932.

49. Kolan D, Fonar G, Samson AO. Elastic network normal mode dynamics reveal the GPCR activation mechanism. Proteins 2014;82(4):579–586.

50. Laberge M, Yonetani T. Common dynamics of globin family proteins. IUBMB Life 2007;59(8–9):528–534.

51. Lai J, Jin J, Kubelka J, Liberles DA. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. J Mol Biol 2012;422(3):442–459.

52. Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophys J 2005;88(2):1291–1299.

53. Hollup SM, Fuglebakk E, Taylor WR, Reuter N. Exploring the factors determining the dynamics of different protein folds. Protein Sci Publ Protein Soc 2011;20(1):197–209.

54. Zheng W, Doniach S. A comparative study of motor-protein motions by using a simple elastic-network model. Proc Natl Acad Sci U S A 2003;100(23):13253–13258.

55. Zheng W, Brooks BR, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci U S A 2006;103(20):7664–7669.

56. Chen J, Brooks CL. Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 2007;67(4):922–930.

57. Heo L, Feig M. What makes it difficult to refine protein models further via molecular dynamics simulations? Proteins 2018;86 Suppl 1:177–188.

58. Tiwari SP, Reuter N. Conservation of intrinsic dynamics in proteins — what have computational models taught us? Curr Opin Struct Biol 2018;50:75–81.

59. Kinch LN, Li W, Schaeffer RD, Dunbrack RL, Monastyrskyy B, Kryshtafovych A, Grishin NV. CASP 11

target classification. Proteins 2016;84 Suppl 1:20–33.

60. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins 2018;86 Suppl 1:7–15.

61. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23(3):ii–v.

62. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. Proteins 2016;84 Suppl 1:15–19.

63. Fuglebakk E, Echave J, Reuter N. Measuring and comparing structural fluctuation patterns in large protein datasets. Bioinforma Oxf Engl 2012;28(19):2431–2440.

64. Camps J, Carrillo O, Emperador A, Orellana L, Hospital A, Rueda M, Cicin-Sain D, D'Abramo M, Gelpí JL, Orozco M. FlexServ: an integrated tool for the analysis of protein flexibility. Bioinformatics 2009;25(13):1709–1710.

65. Kovacs JA, Chacón P, Abagyan R. Predictions of protein flexibility: First-order measures. Proteins Struct Funct Bioinforma 2004;56(4):661–668.

66. Kühn J, Wong LE, Pirkuliyeva S, Schulz K, Schwiegk C, Fünfgeld KG, Keppler S, Batista FD, Urlaub H, Habeck M, Becker S, Griesinger C, Wienands J. The adaptor protein CIN85 assembles intracellular signaling clusters for B cell activation. Sci Signal 2016;9(434):ra66.

67. Yang L, Song G, Jernigan RL. Comparisons of experimental and computed protein anisotropic temperature factors. Proteins Struct Funct Bioinforma 2009;76(1):164–175.

68. Hinsen K. Structural flexibility in proteins: impact of the crystal environment. Bioinforma Oxf Engl 2008;24(4):521–528.

69. Soheilifard R, Makarov DE, Rodin GJ. Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. Phys Biol 2008;5(2):026008.

70. Zhang Z, Ehmann U, Zacharias M. Monte Carlo replica-exchange based ensemble docking of protein conformations. Proteins Struct Funct Bioinforma 2017;85(5):924–937.

71. Rai BK, Tawa GJ, Katz AH, Humblet C. Modeling G protein-coupled receptors for structure-based drug discovery using low-frequency normal modes for refinement of homology models: Application to H3

antagonists. Proteins Struct Funct Bioinforma 2010;78(2):457–473.

72. Warszycki D, Rueda M, Mordalski S, Kristiansen K, Satała G, Rataj K, Chilmonczyk Z, Sylte I, Abagyan R, Bojarski AJ. From Homology Models to a Set of Predictive Binding Pockets–a 5-HT1A Receptor Case Study. J Chem Inf Model 2017;57(2):311–321.
73. Gorba C, Miyashita O, Tama F. Normal-Mode Flexible Fitting of High-Resolution Structure of Biological Molecules toward One-Dimensional Low-Resolution Data. Biophys J 2008;94(5):1589–1599.
74. Pérard J, Leyrat C, Baudin F, Drouet E, Jamin M. Structure of the full-length HCV IRES in solution. Nat Commun 2013;4:1612.
75. Schindler CEM, Beauchêne IC de, Vries SJ de, Zacharias M. Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI. Proteins Struct Funct Bioinforma 2017;85(3):391–398.
76. Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics 2020;36(6):1765–1771.
77. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins 2016;84 Suppl 1:4–14.
78. Berendsen HJC, Spoel D van der, Drunen R van. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun 1995;91(1–3):43–56.
79. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.
80. Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, Vriend G. A series of PDB-related databanks for everyday needs. Nucleic Acids Res 2015;43(D1):D364–D368.
81. Goldstein H, Poole CP, Safko JL. Classical mechanics. San Francisco: Addison Wesley; 2002.
82. Ma J. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 2005;13(3):373–380.
83. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22(21):2695–2696.
84. Arfken GB. Gram-Schmidt Orthogonalization. In: Mathematical methods for physicists. 3rd ed. Orlando, FL: Academic press; 1985. p 516–520.

85. Kuzmanic A, Zagrovic B. Determination of Ensemble-Average Pairwise Root Mean-Square Deviation from Experimental B-Factors. Biophys J 2010;98(5):861–871.

86. Willis BTM, Pryor AW. Thermal Vibrations in Crystallography. London ; New York: Cambridge University Press; 1975.

87. Suhre K, Sanejouand Y-H. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 2004;32(Web Server):W610–W614.

88. McLachlan AD. Rapid comparison of protein structures. Acta Crystallogr A 1982;38(6):871–873.

89. Martin ACR, Porter CT. http://www.bioinf.org.uk/software/profit/. .

90. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2011;27(3):343–350.

91. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res 2009;37(suppl_2):W510–W514.

92. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.

93. Rees DG. Essential Statistics, Fourth Edition. : CRC Press; 2000.

94. Wilcoxon F. Individual Comparisons by Ranking Methods. Biom Bull 1945;1(6):80–83.

95. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. J Am Stat Assoc 1952;47(260):583–621.

96. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.

FIGURE LEGENDS

Figure 1. Motions predicted by NMA are not strongly affected by modeling errors. (A) RWSIP values for the comparison of the first 10 lowest modes computed on either a model or its respective native structure using iNMA. Results are presented as a function of the RMSD between this pair of structures. On the margin, the same values are shown as histograms. The horizontal grey line indicates RWSIP equal to 0.5 and the dashed broken red line indicates the empirical interval containing 90% of the data, for the intervals of RMSD native/model: 0-1, 1-1.5, 1.5-2, 2-2.5, 2.5-3, 3-3.5, 3.5-4, 4-4.5, 4.5-5. (B) Same as (A) but for modes calculated with cNMA.

Figure 2. **iNMA allows native-like conformational sampling of protein models with structure preservation.** Left column: superimposition of native structures (NS, red) and structural models (blue), with corresponding RMSD. Marked by two dashed rectangles (gold for iNMA and purple for cNMA) are the comparison of the conformations generated by NMA. From up to down, the models shown are: T0783TS169_2-D1, T0805TS277_1-D1, T0865TS407_2-D1, T0903TS467_1-D1, with their respective native structure. The black arrows show the direction of the motion. The thickness of the ribbon represents the extent of RMSD (Å) of $C\alpha_i$-$C\alpha_{i+1}$ virtual bonds between the new conformation and the starting structure. The scale is the same for all figures, from 0 to 3.6 Å.

Figure 3. Prediction of protein intrinsic flexibility is not affected by NMA approaches. In the main image, we show the comparison of IFlex values predicted for the native structures (NS) using either iNMA ($IFlex^{I}_{NS}$) or cNMA ($IFlex^{C}_{NS}$), with the value of the Spearman correlation coefficient, ρ. The structures of the three most outliers are shown (1: T0865-D1, 2: T0820-D2, 3: T0805-D1). In the lookup, we show only targets with predicted $IFlex^{I}_{NS}$ lower than 2. On the margins, the same $IFlex_{NS}$ values are shown as histograms (purple for cNMA and yellow for iNMA).

Figure 4. Protein intrinsic flexibility is well captured by structural models. The weighted difference between IFlex values calculated for a native structure and each model (ΔIFlex) is here compared to the RMSD between this pair of structures, and colored according to the IFlex value of the native structure ($IFlex_{NS}$). The same ΔIFlex values are shown as histograms on the margin. Horizontal grey lines indicate ΔIFlex equal to -10 and 10; dashed broken black lines indicate the empirical interval containing 90% of the data, for the same intervals of RMSD native/model used in Figure 1. IFlex values are expressed in units of 10^{13} s. (A) Predictions performed with iNMA. (B) Predictions performed with cNMA.

Figure 5. **Flexible regions can be predicted from models**. (A) Overview of the comparison scheme. (B) Comparison between experimental B-factors of native structures and NMA-derived RMSF computed on structural models (dashed blue line in panel A), using AUC values, as a function of the RMSD between this pair of structures. On the margin, the same values are shown as histograms. Dashed lines indicate the empirical interval containing 90% of the data, for the

same intervals of RMSD native/model used in Figure 1. (C) Illustrative example of target T0786-D1 and model T0786TS436_1-D1 (RMSD=4.6 Å), with the superposition of B-factor profiles (black) and iNMA-derived RMSF profiles (red for native structure, blue for model). The structures of both native and model are colored according to either B-factors or squared RMSF. The region around index 230, where we observe low B-factors but high RMSF, is shadowed in grey in the plot, and shown by a dashed grey circle on the structure. This region is part of the oligomerization interface (PDB structure 4QVU is a tetramer), justifying the discrepancy.

Figure 6. iNMA allows model refinement preserving the protein valence structure. (A) Minimum RMSD value to the native structure obtained by applying NMA on each structural model, compared to the original RMSD between this pair of structures. The color scale corresponds to the amount of bond stretching on the generated structures, calculated as the RMSD (Å) of $C\alpha_i$-$C\alpha_{i+1}$ virtual bonds between the new conformation and the initial model. Red dots represent values of RMSD (Å) of $C\alpha_i$-$C\alpha_{i+1}$ virtual bonds equal to or larger than 1 Å. In this panel, results using iNMA. (B) Same as (A), but using cNMA. (C) Fractional RMSD (ΔRMSD%= 100(RMSD-$RMSD_{NMA}$)/RMSD) as a function of the original RMSD native/model. On the margins, the same values are shown as histograms. The same color scale of the previous panels is used. The red line represents the mean value of ΔRMSD% for the same intervals of RMSD native/model used in Figure 1. (D) Same as (C), but using cNMA.

FIGURES

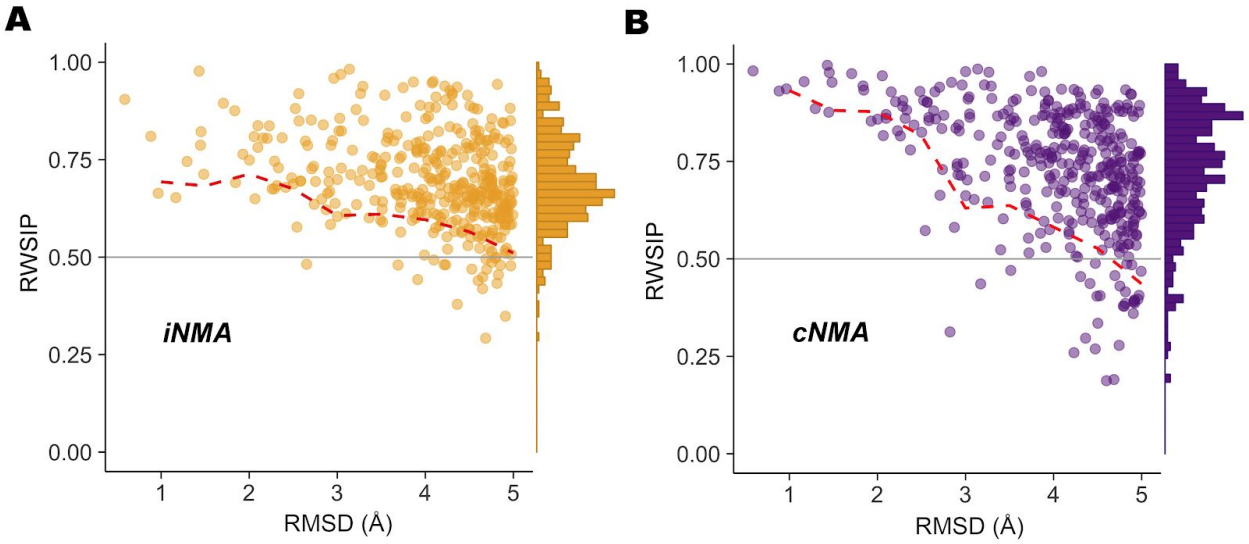

Figure 1. Motions predicted by NMA are not strongly affected by modeling errors.

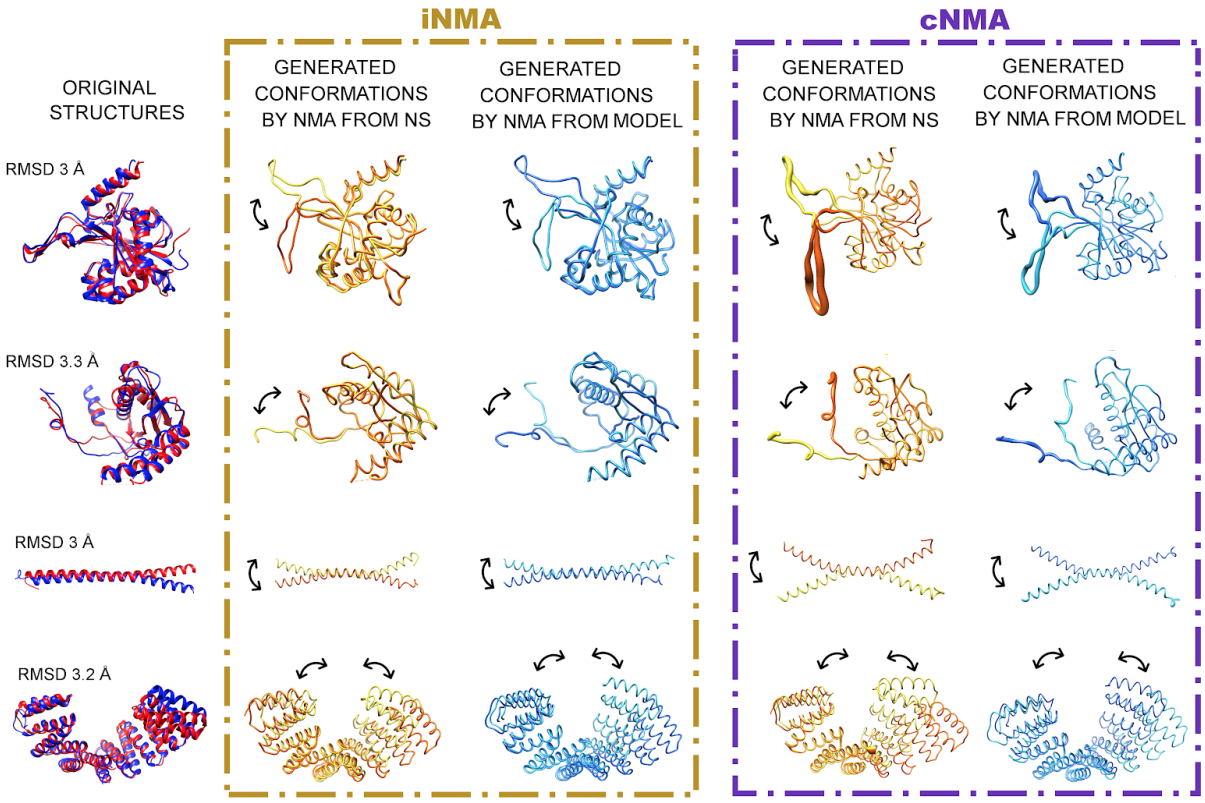

Figure 2. iNMA allows native-like conformational sampling of protein models with structure preservation.

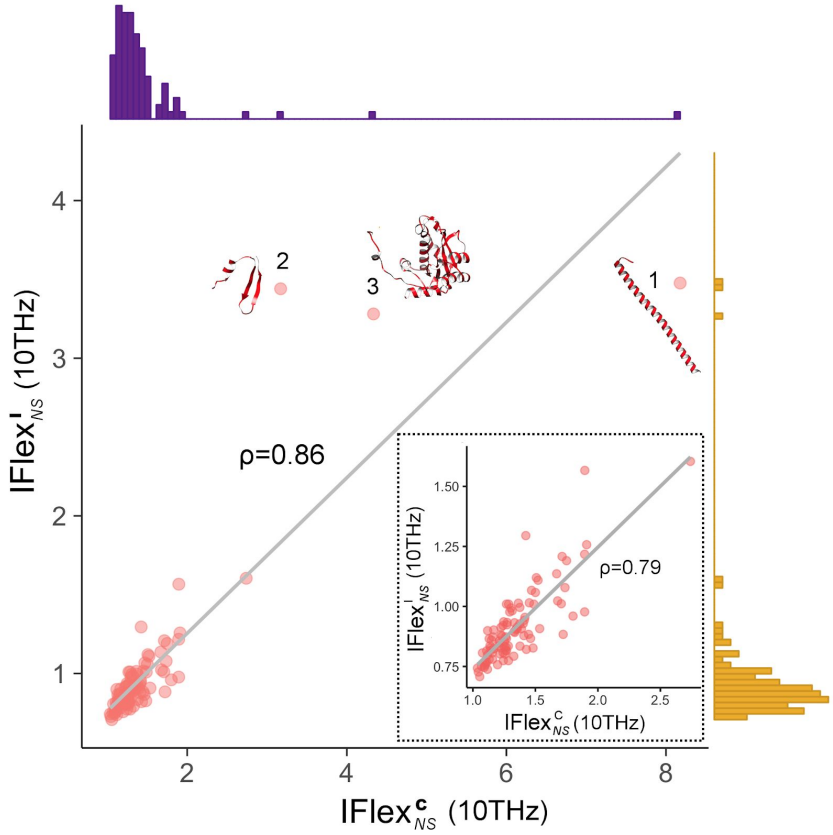

Figure 3. Prediction of protein intrinsic flexibility is not affected by NMA approaches.

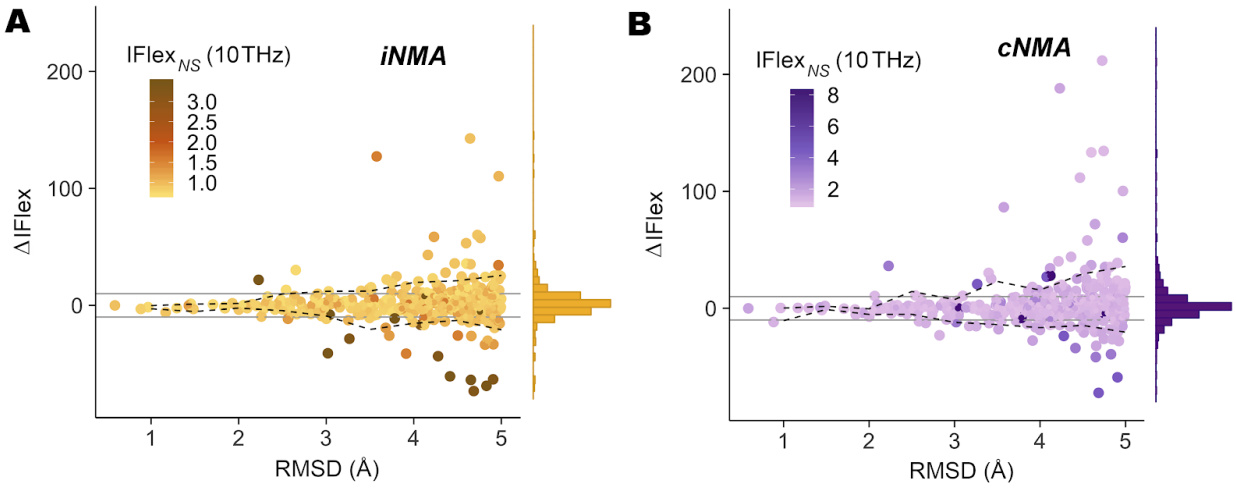

Figure 4. Protein intrinsic flexibility is well captured by structural models.

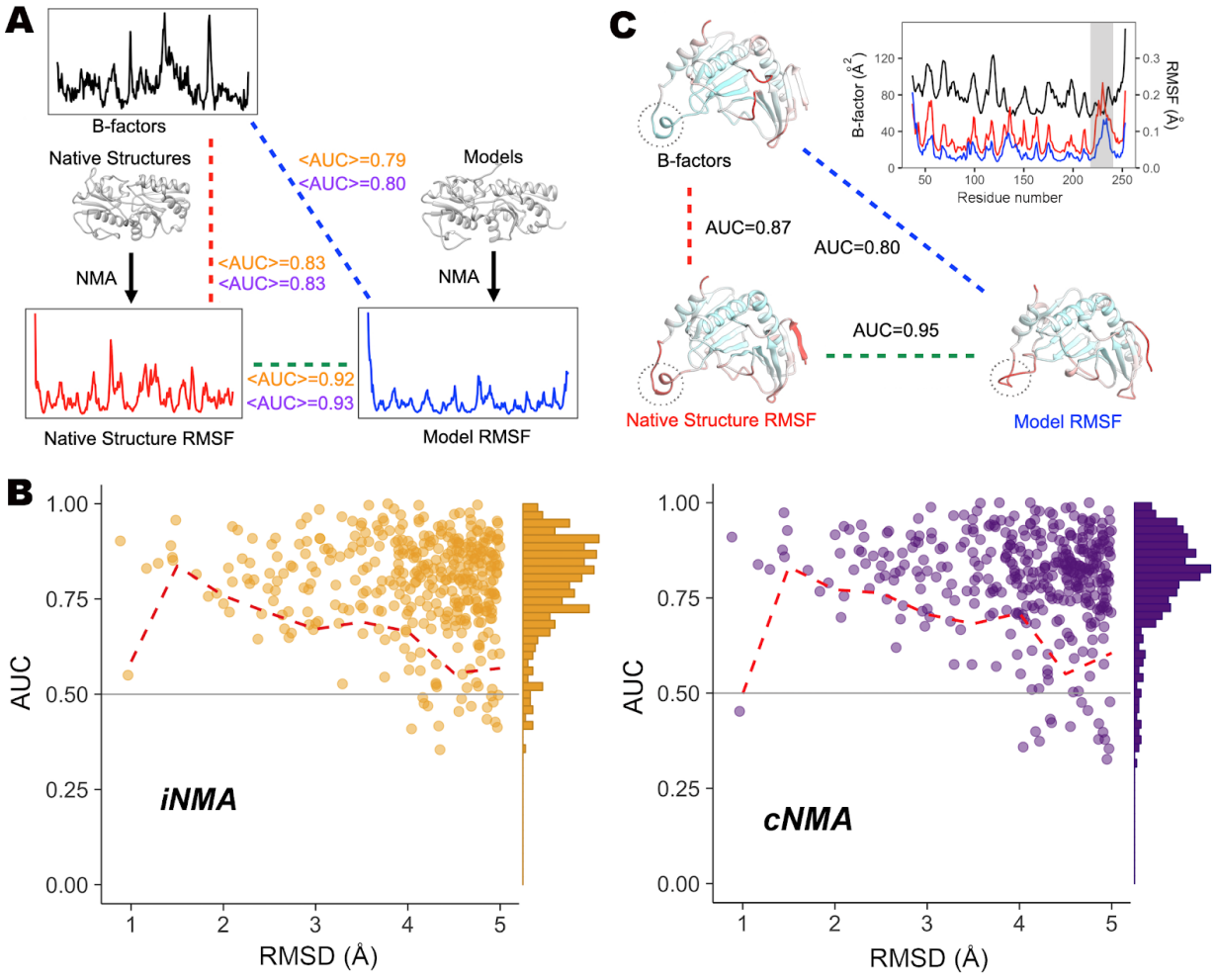

Figure 5. **Flexible regions can be predicted from models**.

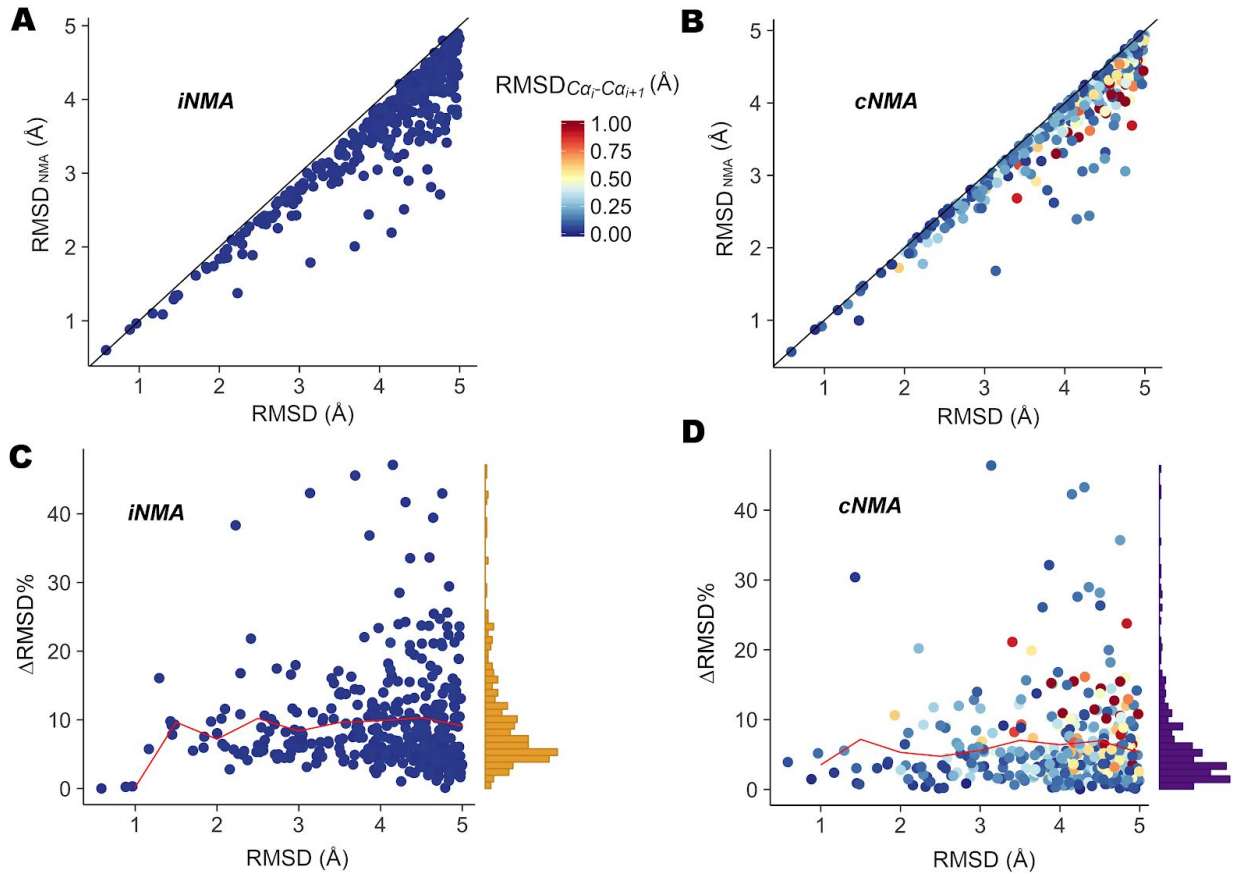

Figure 6. iNMA allows model refinement preserving the protein valence structure.

SUPPORTING INFORMATION

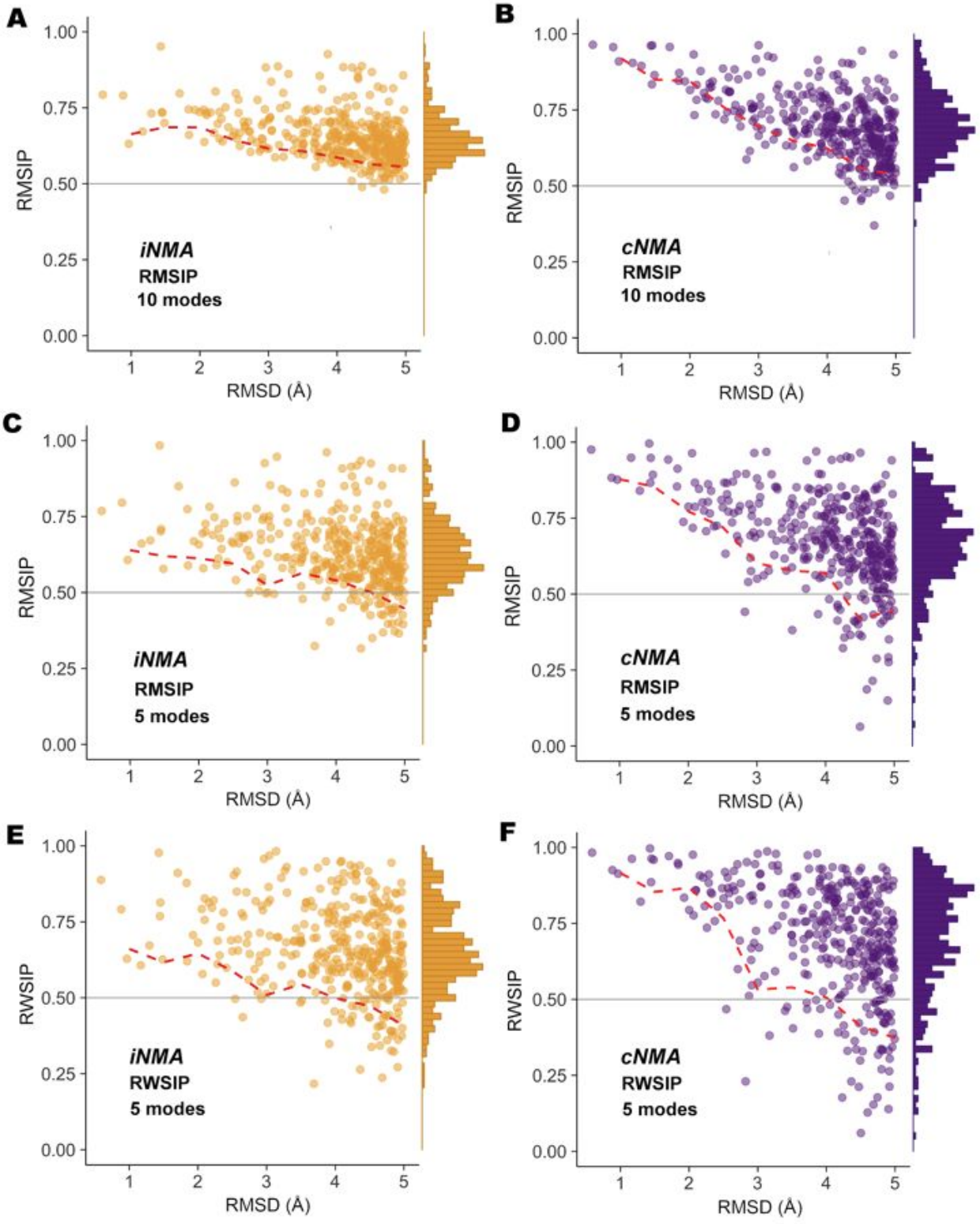
A
B
C
D
E
F
RMSIP
RWSIP
RMSD (Å)
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
iNMA
RMSIP
10 modes
cNMA
RMSIP
10 modes
iNMA
RMSIP
5 modes
cNMA
RMSIP
5 modes
iNMA
RWSIP
5 modes
cNMA
RWSIP
5 modes

Figure S1. Results of motion comparison are not impacted by considering either RMSIP or RWSIP, or by the number of modes taken into account. (A) RMSIP values for the comparison of the first 10 modes calculated from either a model or its respective native structure. Results are presented as a function of the RMSD between this pair of structures. On the margin, the same values are shown as histograms. The horizontal grey line indicates RMSIP equal to 0.5 and the dashed broken red line indicates the empirical interval containing 90% of the data, for the intervals of RMSD native/model: 0-1, 1-1.5, 1.5-2, 2-2.5, 2.5-3, 3-3.5, 3.5-4, 4-4.5, 4.5-5. (B) Same as (A) but for modes calculated with cNMA. (C) Similar to (A) but RMSIP values for the comparison of the first 5 modes calculated from either a model or its respective native structure.(D) Same as (C) but for modes calculated with cNMA. (E) Similar to (A) but RWSIP values for the comparison of the first 5 modes calculated from either a model or its respective native structure.(F) Same as (E) but for modes calculated with cNMA.

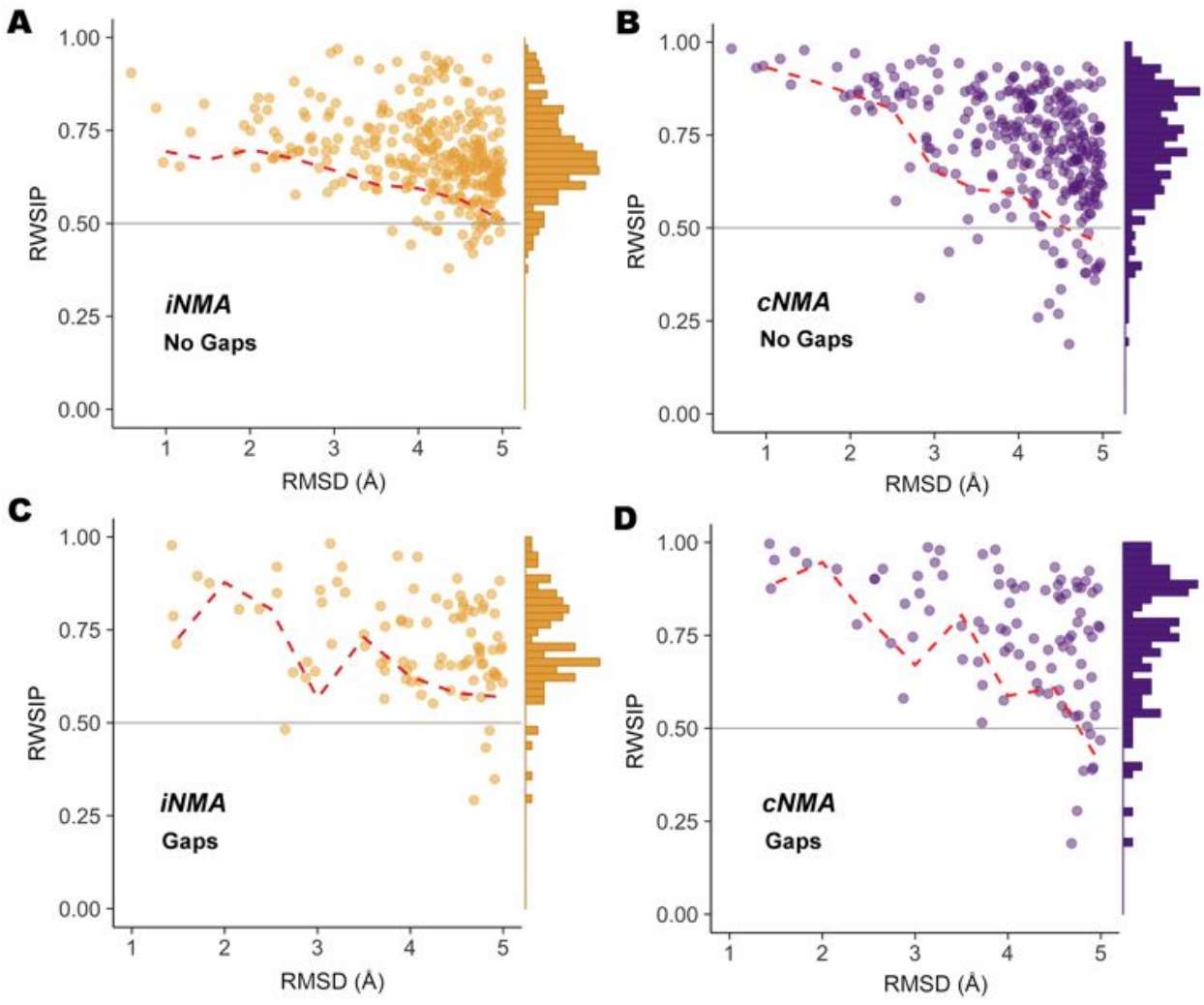

Figure S2. Gaps (i.e. missing regions) in structures have no impact on the mode reproduction. (A) For structures without gaps (71 of 99 targets, see Table S1), RWSIP values for the comparison of the first 10 modes calculated from either a model or its respective native structure. Results are presented as a function of the RMSD between this pair of structures. On the margin, the same values are shown as histograms. The horizontal grey line indicates RWSIP equal to 0.5 and the dashed broken red line indicates the empirical interval containing 90% of the data, for the same intervals of RMSD native/model as Figure S1. (B) Same as (A) but for modes calculated with cNMA. (C) Similar to (A) but for structures with gaps (28 of 99 targets, see Table S1). (D) Sames as (C) but for modes calculated with cNMA.

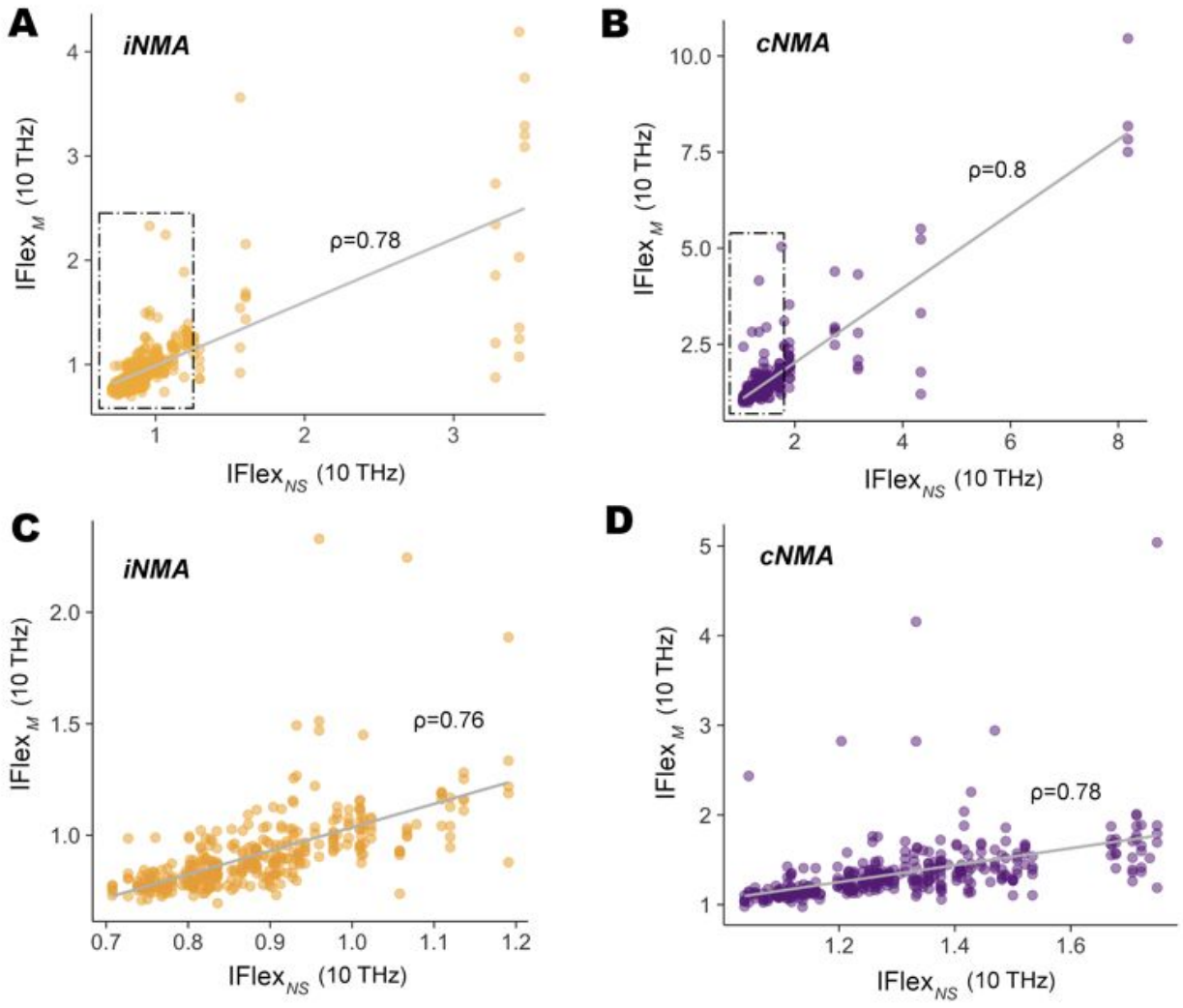

Figure S3. There is a good correlation between IFlex predicted from native structures and IFlex predicted from models. (A) Comparison between the IFlex values predicted for each model ($IFlex_M$, y-axis) and its native structure ($IFlex_{NS}$, x-axis). The value of the Spearman ρ is shown. Values of native structure IFlex under the 90th percentile are shown by a dashed box. (B) Same as (A) but for modes calculated with cNMA. In (C) and (D), the comparison was performed only for the pairs whose native structure presented IFlex value under the 90th percentile (dashed boxes in (A) and (B)).

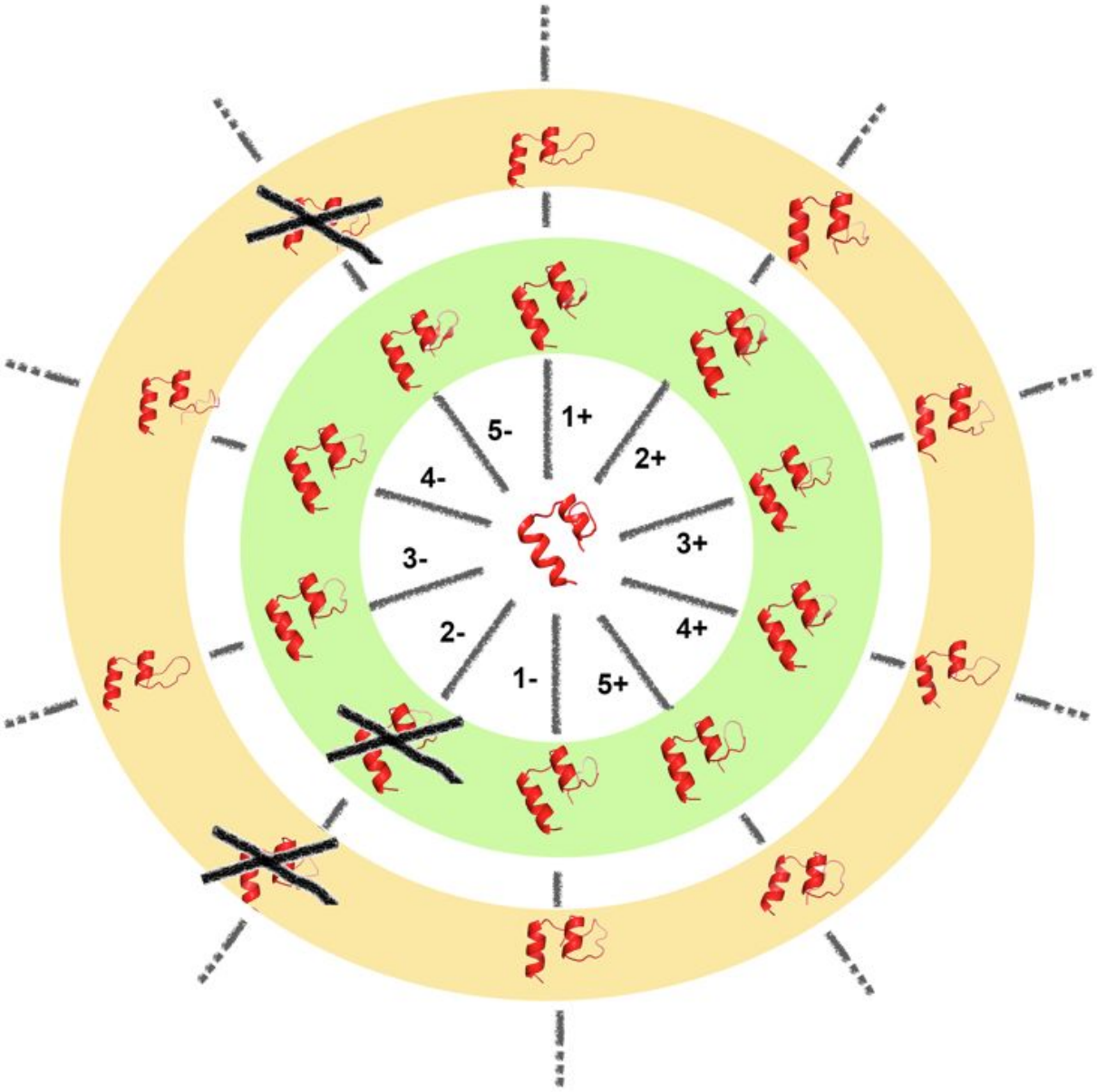

Figure S4. Scheme of the methodology for creation of generated conformations by iNMA. For each original structure, 10 conformations were created along the two directions of the first 5 normal modes, till an RMSD of 0.5 Å between the new conformation and the original one was reached (set shadowed in green). Then, the amplitude was increased to generate similarly 10 conformations but with an RMSD of 1 Å to the original structure (set shadowed in orange), and subsequently till an RMSD of 5 Å was obtained (not represented). The modes are indicated by the

numbers 1-5, and the directions by a sign (+,-). The conformations with clashes, which were removed, are cancelled by a cross.

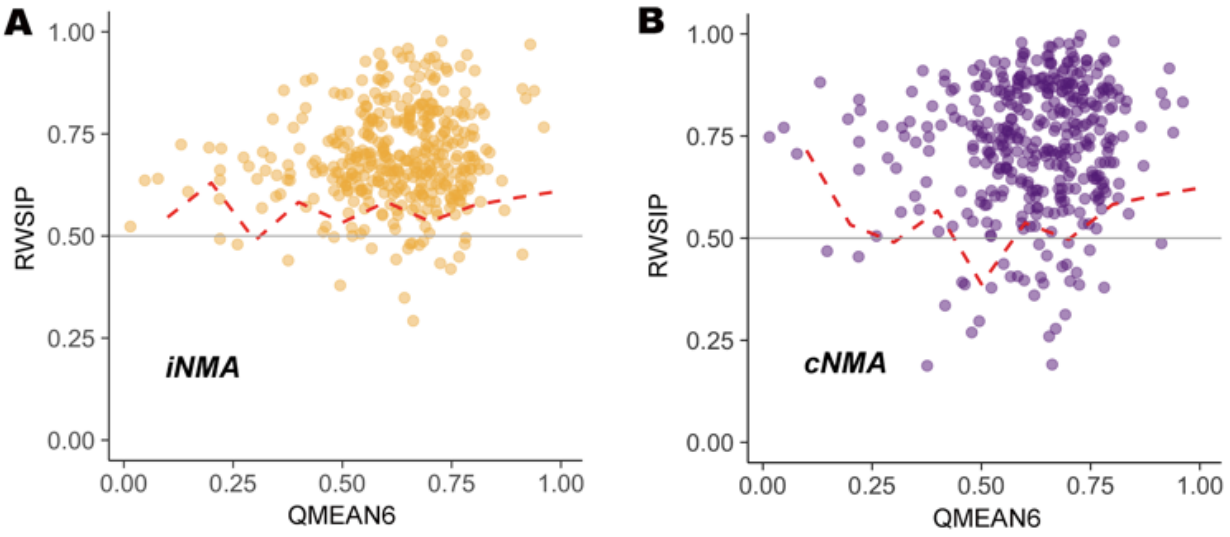

Figure S5. There is no clear relationship between RWSIP value model/native and the model's quality QMEAN score. (A) The RWSIP values for the comparison of the first 10 modes calculated from either a model or its respective native structure, presented in Figure 1, is here shown as a function of the QMEAN6 score of each model. The horizontal grey line indicates RWSIP equal to 0.5 and the dashed broken red line indicates the empirical interval containing 90% of the data, for bins of QMEAN6 score of 0.1. (B) Same as (A) but for modes calculated with cNMA. Compared to Figure 1, where a clear trend of increasing RWSIP values was observed as the RMSD model/native decreased, here this trend is absent, showing no relationship between QMEAN quality scores and normal modes similarity model/native.

Table S2. Analysis of the predicted flexibility for different subgroups of native structures, classified based on structural features (related to Table S1). Results for iNMA ($IFlex^I$) and cNMA ($IFlex^C$). Bold fonts indicate significant differences at the 5% threshold (p-value of the Mann-Whitney or Wilcoxon test).

Feature	Classification	Number of structures in dataset	Mean $IFlex^I$ value (± standard deviation)	Mean $IFlex^C$ value (± standard deviation)
Secondary structure	Both	56	0.97 ± 0.48	1.39 ± 0.51
	Beta	27	0.93 ± 0.15	1.35 ± 0.22
	Alpha	16	1.15 ± 0.66	1.92 ± 1.71
p-value			0.0846300	0.1314000
AUC			0.611	0.582
Oligomerization	Monomer	56	**0.89 ± 0.14**	**1.3 ± 0.21**
	Oligomer	43	**1.12 ± 0.65**	**1.67 ± 1.17**
p-value			**0.0040640**	**0.0109300**
AUC			0.670	0.650
Loops	No	74	**0.96 ± 0.43**	1.47 ± 0.90
	Yes	25	**1.06 ± 0.52**	1.45 ± 0.41

p-value			**0.0099590**	0.0979100
AUC			0.673	0.611
Gaps	No	71	0.98 ± 0.45	1.48 ± 0.88
	Yes	28	1.00 ± 0.47	1.43 ± 0.60
p-value			0.5471000	0.8551000
AUC			0.539	0.512

Table S3. Analysis of the predicted flexibility for different subgroups of models, classified based on the structural features or their native structures (related to Tables S1 and S2). Results for iNMA ($IFlex^I$) and cNMA ($IFlex^C$). Bold fonts indicate significant differences at the 5% threshold (p-value of the Mann-Whitney or Wilcoxon test).

Feature	Classification	Number of models in dataset	Mean $IFlex^I$ value (± standard deviation)	Mean $IFlex^C$ value (± standard deviation)
Secondary structure	Both	241	**0.96 ± 0.34**	**1.44 ± 0.62**
	Beta	108	**0.95 ± 0.28**	**1.41 ± 0.35**
	Alpha	70	**1.16 ± 0.60**	**1.95 ± 1.73**
p-value			**0.000984**	**0.001655**
AUC			0.600	0.594
Oligomerization	Monomer	234	**0.91 ± 0.24**	**1.37 ± 0.35**
	Oligomer	185	**1.08 ± 0.51**	**1.71 ± 1.25**
p-value			**5e-07**	**0.0015520**
AUC			0.643	0.590
Loops	No	316	**2.21 ± 1.39**	**1.52 ± 0.98**

	Yes	103	**0.96 ± 0.37**	**1.50 ± 0.48**
p-value			**2.32e-05**	**0.0010890**
AUC			0.639	0.607
Gaps	No	305	0.99 ± 0.46	1.48 ± 0.85
	Yes	114	1.01 ± 0.51	1.44 ± 0.65
p-value			0.8722000	0.6983000
AUC			0.495	0.512