# A superposition free method for protein conformational ensemble analyses and local clustering based on a backbone differential geometry representation

**Antonio Marinho da Silva Neto[1,2], Samuel Reghim Silva[2], Michele Vendruscolo[3], Carlo Camilloni[4], and Rinaldo Wander Montalvão[5]**

**[1]Centrum Nowych Technologii, Uniwersytet Warszawski, Banacha 2C, 02-097, Warsaw, Poland**
**[2]São Carlos Institute of Physics, University of São Paulo, 13560-970, São Carlos, SP, Brazil.**
**[3]Department of Chemistry, University of Cambridge,Lensfield Road, Cambridge CB2 1EW, United Kingdom**
**[4]Dipartimento di Bioscienze, Università degli Studi di Milano, via Celoria 26, I-20133 Milano, Italy**
**[5]DBC Company, Av. Andaraí 531, 91350-110, Porto Alegre, RS, Brazil**

Corresponding author:
Rinaldo Wander Montalvão[5]

Email address: r.montalvao@xxx.xxx

## ABSTRACT

One of the major challenges of modern structural biology is how to deal with protein flexibility. Beside the experimental difficulties, the lack of a proper mathematical language to represent protein conformational space still remains a problem to be solved. A differential geometry (DG) representation of protein structures can provide a tool to overcome the current limitations of popular representations. Here a DG-based representation of protein backbone is explored on the analyses of protein conformational ensembles. The DG representation consists of representing the protein backbone as a 3D regular curve and describing it by curvature, $\kappa$, and torsion, $\tau$, values per residue. Using this $\kappa/\tau$ metric space as dissimilarity measurement, a protein flexibility measurement based on the maximum $\kappa/\tau$ distance observed, $d_{max}$, were defined and a local clustering method was applied to identify global conformational states. To investigate its efficacy, the proposed methods were applied to two protein test case conformational ensembles: 1) Ubiquitin and 2) c-Myb-KIX binding. Results shows the $\kappa/\tau$ metrics allow to properly judge protein flexibility by avoiding the pitfalls of the superposition problem. The $d_{max}$ measurement presents equally good or superior results when compared with the popular RMSF on the tested systems, specially for the intrinsically unstructured (IUP) protein tested. The clustering approach proposed gives multiple global clustering solutions based on residues local features, therefore can provide insight about residues role on global dynamics. The DG-based backbone representation is an ideal representation of backbone dynamics and the method proposed will be a useful tool for computational structural biology, specially to the analyses of highly flexible proteins (e. g. IUP). The FleXgeo software written for the analyses presented here is freely available for academic usage only at **http://XXX.XXXX.XX/**.

## INTRODUCTION

Accounting for protein flexibility is essential to better predict and understand protein binding features. There are experimental evidence for conformational entropy as one of the main driving forces of protein binding (Frederick et al., 2007; Tzeng and Kalodimos, 2012) and its role on protein evolution (Javier Zea et al., 2013; Parisi et al., 2015; Saldaño et al., 2016). In addition, some estimative suggest flexibility as a prevalent protein property (Lobanov and Galzitskaya, 2015), specially for regulatory proteins (Iakoucheva et al., 2002). Therefore, is not a surprise this topic is relevant for many human diseases (Campbell et al., 2016; Ma and Nussinov, 2016; Whitney et al., 2016). The term "D3" for protein intrinsic disorder in neurodegenerative diseases highlights the role of intrinsically unstructured proteins (IUP) on neurological diseases such as Alzheimer's, Parkinson's and Huntington's disease. (Uversky, 2014). Naturally, this is an important topic for drug design projects. For instance, Gleevec®, a drug used for chronic myeloid leukemia and gastrointestinal stromal tumors treatment, presents high specificity for certain kinds of kinases and this is only explained by accounting for protein flexibility (Agafonov et al., 2015). Gleevec® showcase how crucial accounting for protein flexibility can be to understand protein binding in biomedical and biotechnology projects.

However, understanding the mechanical properties of proteins usually involves obtaining and analyzing conformational ensembles. Those sets of conformations can be obtained by protein crystallography, Nuclear Magnetic Resonance (NMR) and/or molecular dynamics simulations. Despite the conformational sampling be challenging, methods that combine molecular dynamics simulation protocols to reproduce experimental data, such as SAXS (Kimanius et al., 2015), Residual Dipolar Coupling (RDC) (Montalvao et al., 2012), Chemical Shifts(Robustelli et al., 2010) or multiple sources of experimental data(Bonomi et al., 2017), are promising solutions. However, even if one have the perfect protein conformational ensemble, the next challenge is how to get useful insights from this sort of data. The first analyses decision is what coordinate system one will use and the most popular solutions are atomic coordinates, phi-psi angles and/or collective variables (CVs). Atomic coordinates are the most natural way to represent a molecule, but usually there is no interest in translational and rotational motion, which impose the necessity of solving the superposition problem. This can be a problem for very flexible proteins and create some difficulties on dimensionality reduction techniques such as Principal Component Analyses (Sittel et al., 2014). The phi-psi angle avoid the superposition problem, but the periodic nature of angles impose a given topology to the conformational space that can difficult some analyses, such as conformational clustering (Hinsen, 2006). A collective variable can be any set of quantities one can calculate from the structure to represent the main degrees of freedom of the protein. However, it is a system-dependent representation, therefore not scalable to apply to many different proteins, and in practice it is hard to be sure that there is no important degree of freedom neglected on a given set of CVs (Laio and Gervasio, 2008).

A less popular representation is the differential geometry (DG) based representations of protein structure. The main advantages of DG-based representation are intuitive descriptors, because the notion of changes in curvature and torsion are familiar to humans, and the descriptors are absolute in space, avoiding the superposition problem to compare protein structures. It is possible to find applications of DG to protein structure analyses as early as 1978 (Rackovsky and Scheraga, 1978) and among the more recent ones we can mention CHORAL (Montalvão et al., 2005), ARABESQUE (Leung et al., 2012) and POLIPHONY (Pitt et al., 2014). Those three softwares rely on representing protein backbones as 3D regular curves and characterize it by curvature and torsion values per residue. To our knowledge, there is no application of a DG-based representation aimed specifically to protein conformational analyses. By avoiding the superposition problem, DG-based representation can be a better option for backbone protein flexibility analyses and specially advantageous for highly flexible proteins. To investigate this potential advantages, here we explore the application of a DG-based backbone conformational space representation to 1) measure protein flexibility, 2) compare trajectories to a given reference state, and 3) clustering conformations.

# METHODS

## Differential geometry representation

The same approach used in POLYPHONY (Pitt et al., 2014) to represent protein backbone as a regular curve was used here. The first step is to represent a protein conformation as a regular 3D curve. The regular curve representation of the protein backbone is obtained by using the C$\alpha$ coordinates as "knots" for a cubic spline interpolation. Therefore, a parametric vector equation, $\vec{r}(t)$, using C$\alpha$ number, $t$, as parameter is obtained,

$$\vec{r}(t) = x(t)\hat{x} + y(t)\hat{y} + z(t)\hat{z} \tag{1}$$

According to the fundamental theorem of space curves, a regular curve can be fully and uniquely characterized by its curvature, $\kappa$, and torsion, $\tau$, values as a function of the arc length. Given the allowable change of parameter it is possible to compute those values as

$$\kappa = \frac{\|\dot{r} \times \ddot{r}\|}{\|\dot{r}\|^3} \tag{2}$$

$$\tau = \frac{(\dot{r} \times \ddot{r})\dddot{r}}{\|\dot{r} \times \ddot{r}\|} \tag{3}$$

The $\kappa$ and $\tau$ attributed to C$\alpha$ coordinates of a residue $t$ are used to represent the residue as a vector $\vec{x}_t$ given by

$$\vec{x}_t = \{\kappa, \tau, t\} \tag{4}$$

Therefore, a protein conformation $c$ is a set of such vectors

$$X_c = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\} \tag{5}$$

where $N$ is the total number of residues. Consequently, a conformational ensemble $E$ will be a set of $X_c$.

## Dissimilarity metrics and flexibility measurement

Using 4 is possible to compare residues position based on a dissimilarity metric, such as the euclidean distance, defined as

$$d^2(\vec{x}_i, \vec{x}_j) = \sqrt{(\kappa_i - \kappa_j)^2 + (\tau_i - \tau_j)^2} \tag{6}$$

The residue number $t$ was not included because only different positions of the same residue were compared. Equation 6 also allows to compare positions observed by a residue and an arbitrary reference state, such as another conformation on the ensemble or a crystallographic structure bound to some ligand. To avoid one of the values dominate the value $d^2(\vec{x}_i, \vec{x}_j)$ due to possible significant differences on scale, the values of $\kappa$ and $\tau$ are normalized to the interval [0,1].

It is convenient to have a single number measurement to quantify the residues flexibility observed on an ensemble, such as the popular Root Mean Square Fluctuation. Given the possibility of a residue present a multimodal distribution of positions, mean values can be misleading. For this reason, the maximum distance observed by a residue, $d_{max,t}$, was calculated. First an histogram of the distribution of $\kappa$ and $\tau$ observed by the residue is calculated; the optimal bin width value is defined by the iterative kernel bandwidth optimization method proposed by Shimazaki and Shinomoto (2010). To remove outliers, the minimum and maximum bin extremes with less than 1% of the total population are not considered on the calculation of $d_{max,t}$. After outliers removal, $d_{max,t}$ is computed by

$$d_{max,t} = \sqrt{(\kappa_{min,t} - \kappa_{max,t})^2 + (\tau_{min,t} - \tau_{max,t})^2} \tag{7}$$

where $\kappa_{min,t}$ and $\kappa_{max,t}$ are the minimum and maximum values of $\kappa$ observed by residue $t$, respectively. $\tau_{min,t}$ and $\tau_{max,t}$ are the minimum and maximum values of torsion observed by residue $t$.

## Local conformational clustering.

Another common problem on conformational ensemble analyses is the identification of conformational states relevant to protein function. This is essentially a clustering problem. Despite global clustering solutions such as the ones provided by Principal Component Analyses be popular, local clustering can provide multiple global solutions and provide insights about each residues role on protein global dynamics. The

DG backbone representation can be used to provide a clustering solution per residue $t$. The clustering algorithm Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campbell et al., 2016), which is available as a Python library (http://hdbscan.readthedocs.io/en/latest/index.html), was applied to $\kappa/\tau$ set of each residue $t$. The only input strictly necessary for HDBSCAN is the minimum cluster size, which is the minimum set of structures a cluster should have for the user to carry about it. The minimum cluster size was set to a default value of 1% of total conformations. The minimum cluster and minimum sample size values were set as 5% and 5 as the default values, respectively.

### Test set

Two conformational ensembles were used to evaluate how the proposed method performs on different scenarios of the protein flexibility spectrum. The Ubiquitin ensemble determined by NMR-restrained metadynamics (Montalvao et al., 2012) contains 300 conformations, available at Protein DataBank (2LJ5), was used as an exemplar of a typical rigid globular protein. The best representation of the other extreme of the spectrum is the IUPs. An interesting case of an IUP binding via templated folding was studied using conformational ensembles determined by NMR-restrained metadynamics(Toto et al., 2016). The authors were able to obtain four different ensembles of the c-myb transactivation domain (c-Myb), an IUP, bound to KIX by mutating different KIX residues(Toto et al., 2016). Based on a free energy relationship analyses, the ensemble of each mutation can be ordered as steps from 'reactant-like' to 'product-like' as 1)'I72V', 2) 'I26V', 3) 'L43A' and 4) 'WT'(Toto et al., 2016). In this context, each ensemble obtained can be interpret as the position on a coordinate reaction(Toto et al., 2016). Although, those conformations are not necessary representatives of the binding steps of c-myb and KIX wild type system, it still allows to evaluate how the method can be used to study an IUPs folding upon binding. The c-Myb conformations of each c-Myb-KIX ensemble were extracted and included on a single conformational set (total of 946 conformations). This c-Myb "mixed" set of conformations was used to evaluate the possibility of identify, explore similarities and get structural insights of ensemble conformational states based only on the DG-based backbone representation.

## RESULTS AND DISCUSSION

### Ubiquitin

**Flexibility measurement.** The Ubq conformational ensembles show an overall rigid structure with the C-terminal tail as the only region with a high degree of flexibility (Figure 1A). Visual inspection of the values of $\kappa$ and $\tau$ observed on the Ubq ensemble reflects a similar general profile (Figure 1B). A local structural variability measurement should reflect this. RMSF captures this feature (Figure 1C), as the vast majority of residues present $RMSF < 1.5$Å, with residue 10 as the most flexible, and the RMSF grows exponentially from residue 74, up to 8.9Å (6 times more flexible than residue 10).

The $d_{max}$ generally agrees with RMSF results (Figure 1D), but indicates a C-terminal tail 2 times more flexible than residue 10. The reason for this is that the C-terminal change its direction in more than one rotational axis at some hinge points, which gives this subjective visual sense of more "disorder" than the rest of the protein. However, the overall C-terminal tail shape changes observed are comparable to the ones observed on residue 10. This became evident after discounting the rigid body motion by aligning only the C-terminal tail conformations (Figure 2). In order to change direction of the C-terminal tail backbone to different planes, the curve needs to change the plane at a hinge region point, consequently changing $\tau$ value. Even though $\kappa$ values on "rigid" regions might change just as much as more flexible regions, it will still be confined to a certain 2D plane if the $\tau$ values do not change. Therefore, the visual subjective notion of a highly flexible region of a protein is related to significant changes in $\tau$.

**Comparison to a reference state.** An useful structural analysis is quantify the difference between the conformations observed on a given ensemble and a state of reference, such as a crystallographic structure. The euclidean distance (equation 6) can be calculated between each residues of each conformation and a reference state DG backbone representation. To illustrate this application, the Ubq ensemble conformations were compared to the structure of Ubq bound to E2 determined by crystallography (3JW0) (Kamadurai et al., 2009). Visual inspection of the structural alignment Ubq-bound state and ensemble conformations points to very similar structures and this is also reflected in the $d^2$ matrix (Figure 3A). Residues 8, 9, 34 and 68-76 (C-terminal tail) present relatively higher divergences from the reference structure (Figure 3B). Assuming the Ubq ensemble is representative of the conformational equilibrium

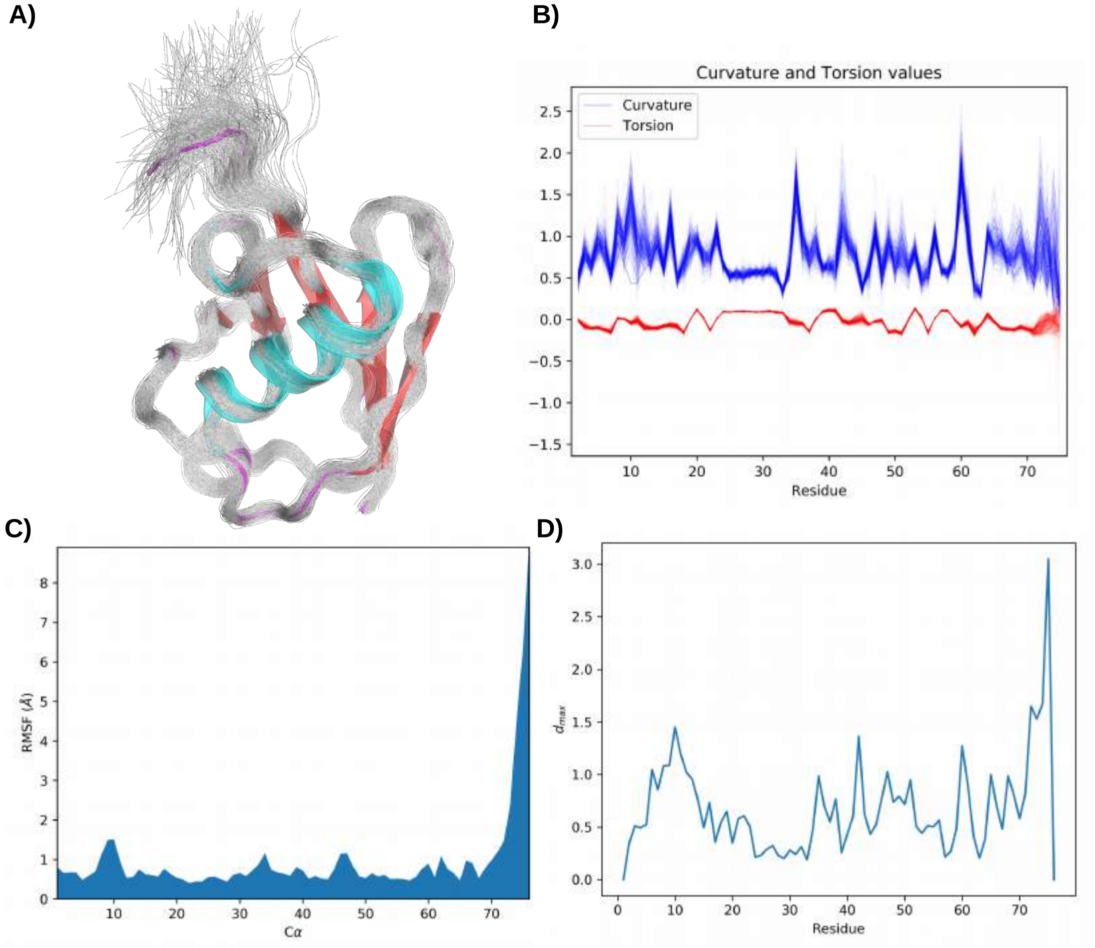

**Figure 1.** A) Conformational ensemble of Ubq (2LJ5). B) $\kappa$ (blue line) and $\tau$ (red line) values per residue for each conformation of the Ubq ensemble. C) Root Mean Square Fluctuation of C$\alpha$ coordinates per residue and D) $d_{max}$ values observed per residues of Ubq ensemble (2LJ5)

in solution, binding to E2 must energetically favor states of those residues not observed in solution. E2 interacts directly with most of those Ubq residues, which suggest some sort of induced fit mechanism involving these residues.

## c-Myb

**Flexibility measurements.** The c-Myb “mixed” ensemble represents a highly flexible protein (Figure 4A). The $\kappa$ and $\tau$ values, given the higher variance of $\tau$ values(Figure 4B), and the RMSF, as the majority of residues presents RMSF $> 4$Å (Figure 4C), reflects this feature. The RMSF results can vary according to the structural alignment solution. Most residues present $d_{max} > 1$, which is similar to the C-terminal tail of Ubq ensemble. Residues 101-112 are the most flexible region according to $d_{max}$. At the first step (’I72V’), the region 88-100 of c-Myb assumes a relatively stable shape promoted by the interaction of c-Myb residues 96-100 with KIXs, while 101-112 region presents a higher degree of flexibility in comparison to the remaining regions of the protein (Figure 5). On the following three steps, those two regions present a similar conformational variability. Note that about $\frac{3}{4}$ of the conformations present a similar shape and this sort of information could be ”diluted” if a mean value of a dissimilarity metric were used, which is one of the advantages of $d_{max}$ over RMSF as a flexibility measurement.

**Comparison to a reference state.** The euclidean distance (equation 6) can be used to compare each residue of each conformation to the final folded stage. To illustrate this application, the heatmap of

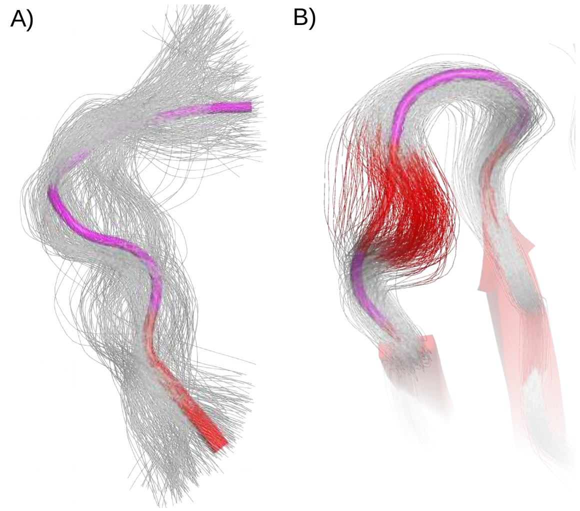

**Figure 2.** A) C-terminal tail of Ubq ensemble (2LJ5) aligned shows a comparable shape variability to B) residue 10 (red), the rigid part of Ubq.

residues euclidean distance matrix between the c-Myb conformations and the conformation of the $\alpha$-helix folded state (Figure 6) was calculated. The heatmap matrix shows the region 98-103 (residue index 10-15) as the first to fold, at the second step ('I26V'). The exception is residue S101 (residue index 13) which presents a larger distance from the final stage at the second step. This residue is in the middle of the helix turn to be formed and residue S101 is the center of a hydrogen bond network involving residue K21 of KIX at this binding step. The S101 position is the one that actually breaks the helix turn, which explain the larger euclidean distance. More about this residue on the following section.

**Conformational clustering.** Four ensembles compose the c-Myb ”mixed” conformational set and a good conformational clustering solution should be able to distinquish them. This test scenario can be use to evaluate how good a global clustering can be achieved based on the local clustering solution proposed. In addition, the ensemble is an intrinsically unstructured protein folding into an $\alpha$-helix, so the clustering results also illustrates how this could be used to identify folding steps. The clustering method applied identifies more than one cluster for most of c-Myb residues and most of them discriminates the final folded stage from the remaining steps (data not shown). For instance, here we will focus on the clustering results of residues E96, S101 and K107 to illustrate how the local clustering results can provide useful insights about protein dynamics.

Residue E96 and S101 are able to distinguish the first two steps ('I72V' and 'I26V'), but the third and final step belongs to the same cluster (Figure 7 and 8). On the other hand, residue K107 discriminates between third and final step, and divide the final step in other two individual clusters, but does not discriminate between the first and second steps (Figure 9). Residue E96 main chain interacts with R61 of KIX and this interaction helps to promote a stable positioning of residues 88-96 of c-Myb on the first binding step. At the second step, residues 88-95 present an unstable positioning, while residues 96-107 fold into a helix turn. Residue E96 is the first residue of the stable region and works like a hinge region for region 88-96; this is reflected by the wider $\tau$ variation and a limited range on $\kappa$ (Figure 7B). S101 is part of a hydrogen bond network involving c-Myb residues and K21 of KIX, which contributes to this region stability. This is reflected by the almost constant K107 $\tau$ values and a higher tolerance to changes in $\kappa$ values. Residues E96 and S101 are on the final folded state at the two remaining steps, therefore those state belongs to the same group from a local perspective. The only residue which clearly separates the third and the final binding steps is the K107. The region 107-110 is the last helix turn to fold and the three clusters capture three stages of this helix turn folding (Figure 9A).

## DG-based backbone representation as a new tool for conformational analyses

There are many challenges on dealing with protein flexibility and the mathematical representation is one of them. The popular representations present some limitations and the DG-based backbone avoids most of it. The DG-based representation is a 1) generic representation, therefore it can be applied to every protein,

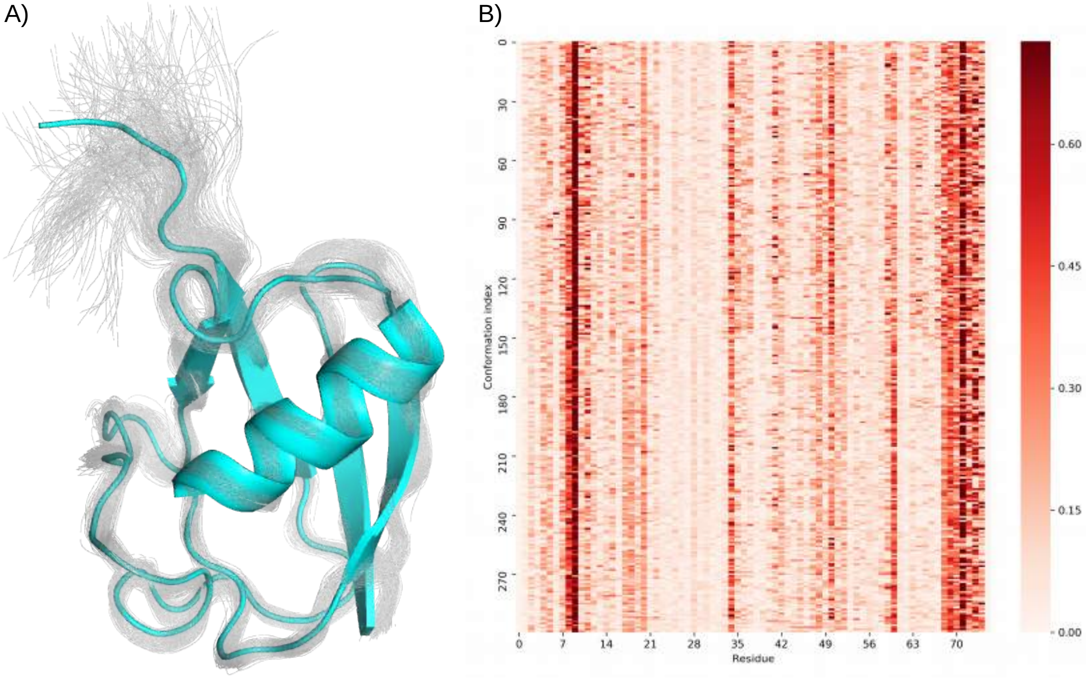

**Figure 3.** A) Ubq conformational ensemble(grey) superimposed with the Ubq bound to E2 structure (3JW0). B) Euclidean distance between each corresponding residue of each Ubq ensemble conformation and Ubq bound to the E2 structure.

2) the descriptors are absolute in space, avoiding the superposition problem, and 3) it provides an uniform metric space, avoiding the problem associated to non-uniform topologies. In addition, the notion of curvature and torsion of a curve are intuitive and there is a low computational cost associated to compute those quantities. The prototype software written for the analyses presented here is not optimized for performance, but all the computation of $\kappa/\tau$ presented took a few seconds on an 2017 average personal notebook (Intel i7-3612QM processor, 8GB RAM, Linux operating system). However, there is no silver bullet for representation of protein conformational space. The DG-based representation main limitation is do not include side chains information. Therefore, it is not a complete representation of protein dynamics and further work is been developed to include side chains representations.

Despite this limitation, the DG-based representation provide a convenient metric space for protein backbone dynamics analyses. The conformations can be compared to another arbitrary reference state by using the euclidean distance as a dissimilarity measurement. The heatmap of the euclidean distance of c-Myb conformations and the final folded state allowed to easily identify protein regions which fold first. This sort of analysis can be useful to identify patterns of similarity/disimilarity on sets of conformations, for instance, analyze protein MD trajectories to identify conformations that get closer or deviate from a known functional conformational state (e. g., ligand-bound or ligand-unbound). The $d_{max}$ as a local backbone flexibility measurement gives equally good or better results than the popular RMSF. If a given backbone region works as a hinge point and all the remaining residues are perfectly rigid, the $d_{max}$ will be 0 for all residues except the ones belonging to the hinge point. In this hypothetical case, the structural superposition would ultimately determinate the RMSF of each region, which is not true for $d_{max}$. Another advantage is that $d_{max}$ reports the maximum $\kappa/\tau$ spread observed by the residue and do not use mean values, giving a better sense of the distribution. Therefore, if a small but relevant set of conformations presents a significant distortion, this information is not lost or shadowed by the averaging procedure.

Another successful application of the DG-based representation is conformational clustering. In principle, there is no right or wrong on a given clustering solution; what matters is if it captures a relevant feature of the dataset or not. Protein conformational dynamics is often complex and is unreasonable to

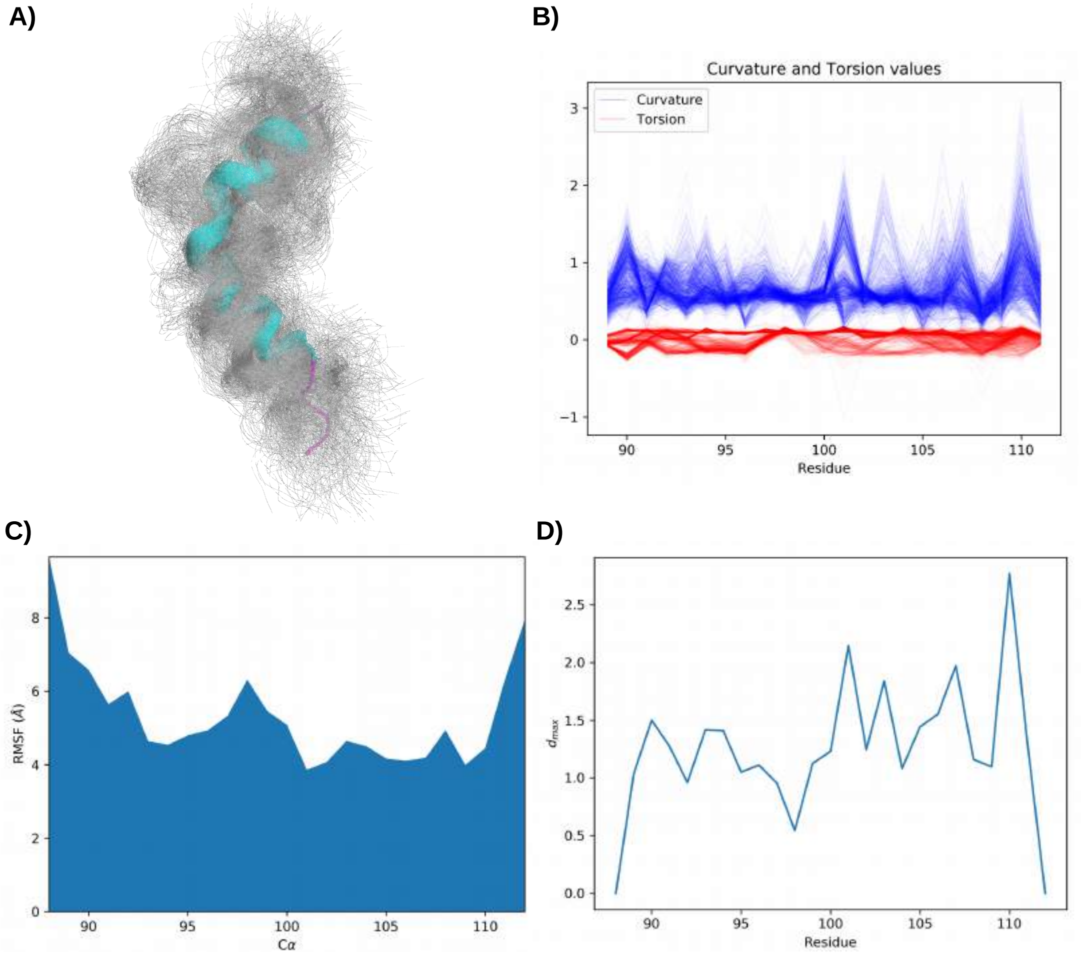

**Figure 4.** A) The c-Myb "mixed-step" set of conformations, which is composed by individual ensembles of c-Myb binding to KIX steps (Toto et al., 2016). B) $\kappa$ (blue line) and $\tau$ (red line) values for each conformation of the cmyb ensemble. C) Root Mean Square Deviation of C$\alpha$ coordinates and D) $d_{max}$ values observed per residues of c-Myb "mixed steps" set of conformations.

expect a single clustering solution to capture all relevant dynamics features. The method proposed here provide a per residue global clustering solution. The advantage of this approach is to provide multiple global clustering solutions based on local features. As illustrated by c-Myb residues E96, S101 and K107, each residue clustering solution can provide insights about the residues role at a global level. The proposed approach is unique, as it identifies global dynamics directly related to local features. This allows an easy and intuitive bottom-up investigation of individual residues role on protein dynamics. To our knowledge, there is no similar approach to protein conformational clustering available today and this can be a useful new tool for structural biologists working with protein conformational ensembles of any nature.

As a future work, we devise a combination of this method with the CamSurf (to be published) approach to generate surfaces from a conformational ensemble. CamSurf represents the surface of each conformation as projections on a base of spherical harmonics polynomials which allow to detect similarity among conformations in terms of similarity of projection coefficients. This surface representation together with the DG based backbone clustering solution can provide a way to optimize ensemble docking protocols.

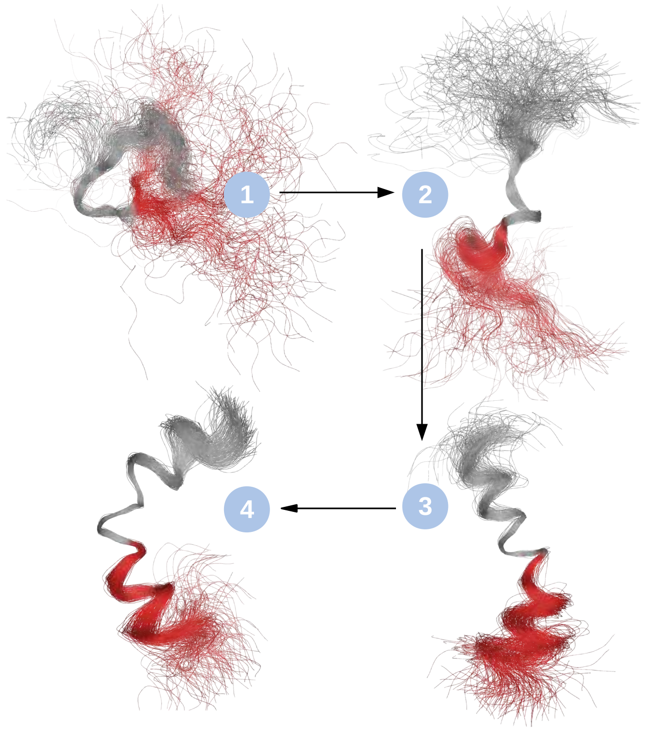

**Figure 5.** Individual cmyb-KIX binding steps. The region 101-112 (red) presents a significantly higher flexibility (as shown by $d_{max}$), but only at the first step. The structural alignment was made considering only the region 96-100, which forms stable interactions with KIX residues at the first step (Toto et al., 2016). Therefore, the structural alignments presented highlight the protein dynamics using this region as a reference point.

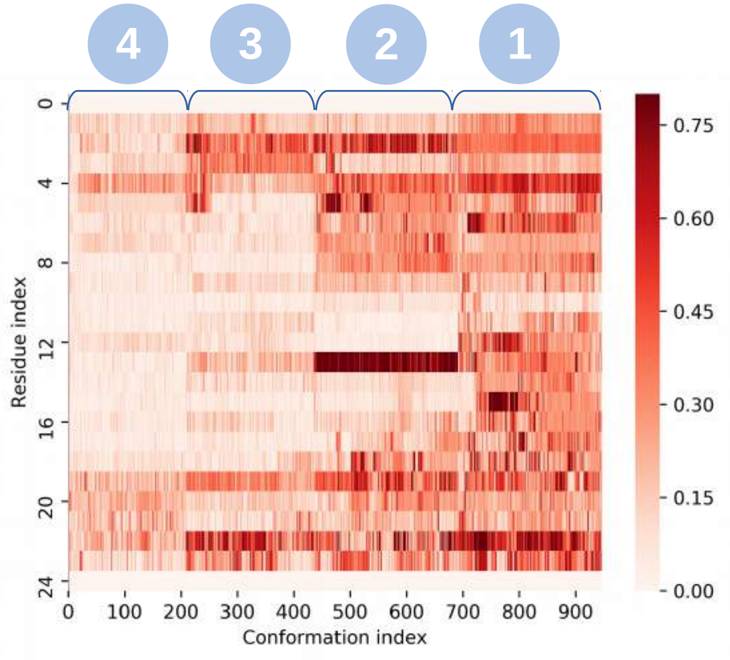

**Figure 6.** Heatmap of residues euclidean distance matrix of each conformation of the "mixed steps" ensembles and the c-Myb folded state. The first and last residues always present $\kappa = \tau = 0$

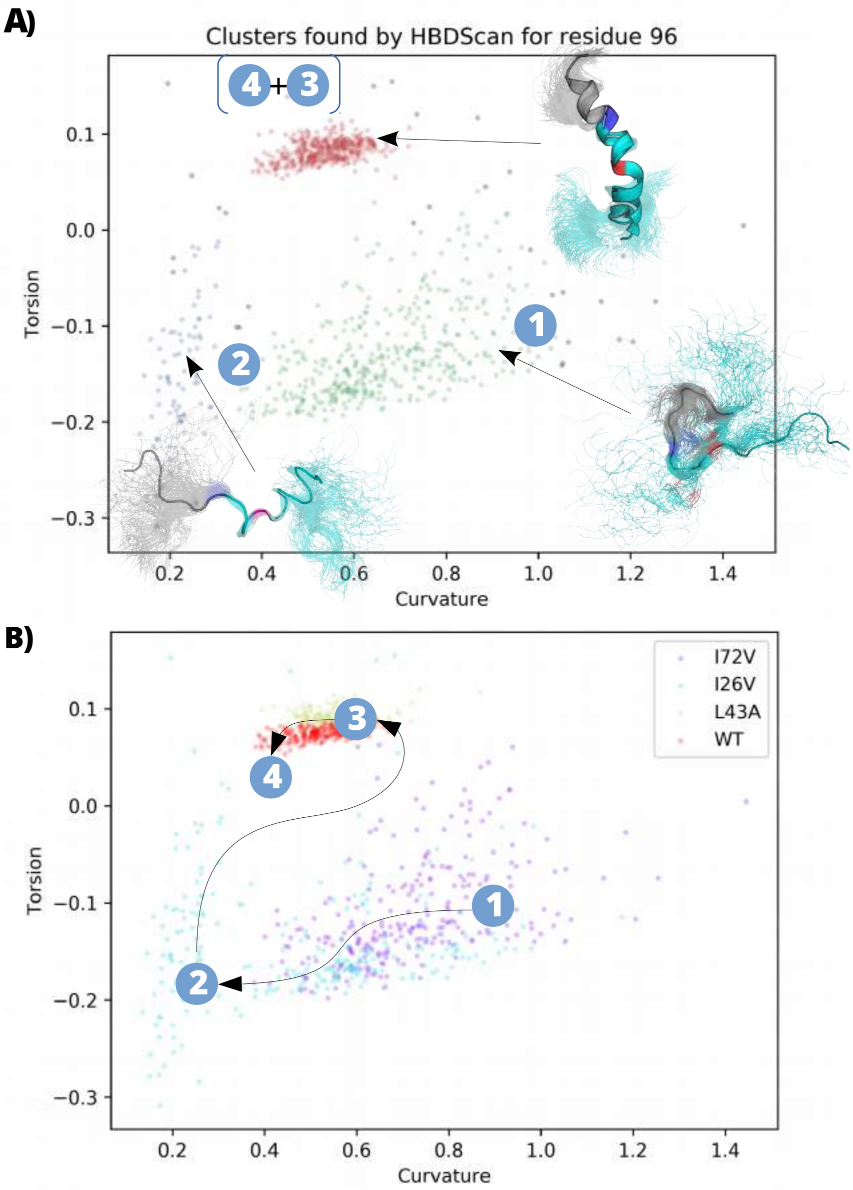

**Figure 7.** A) HDBScan clustering results of residue E96 $\kappa$ and $\tau$ values. Three clusters were identified, the first and second clusters correspond to conformations of the first and second binding steps, the third cluster correspond to the third and final step. The clusters conformations of each group are presented. Residues 88-95 (grey), E96 (blue), and S101 (red) of c-Myb are highlighted on the clusters representation. B) Residue 96 $\kappa$ and $\tau$ values of each individual c-Myb-KIX binding steps, indicated by the KIX mutation, namely I72V (first step), I26V (second step), L43A (third step) and WT (Fourth step).

## CONCLUSION

Here we have proposed a method to evaluate protein flexibility and identify conformational state based on a DG backbone representation. The flexibility measurement proposed show equally good or better results than the popular RMSF. The local conformational clustering provides a per residue global solution which gives insights about the role of residues on global dynamics. The main advantage of DG-based representation is avoiding the superposition problem, which is ideal to analyses of highly flexible proteins conformational ensembles, such as IUPs or protein folding trajectories. In addition, the method is specially sensitive to identifying helices and, consequently, evaluate helix stability. The FleXgeo software used

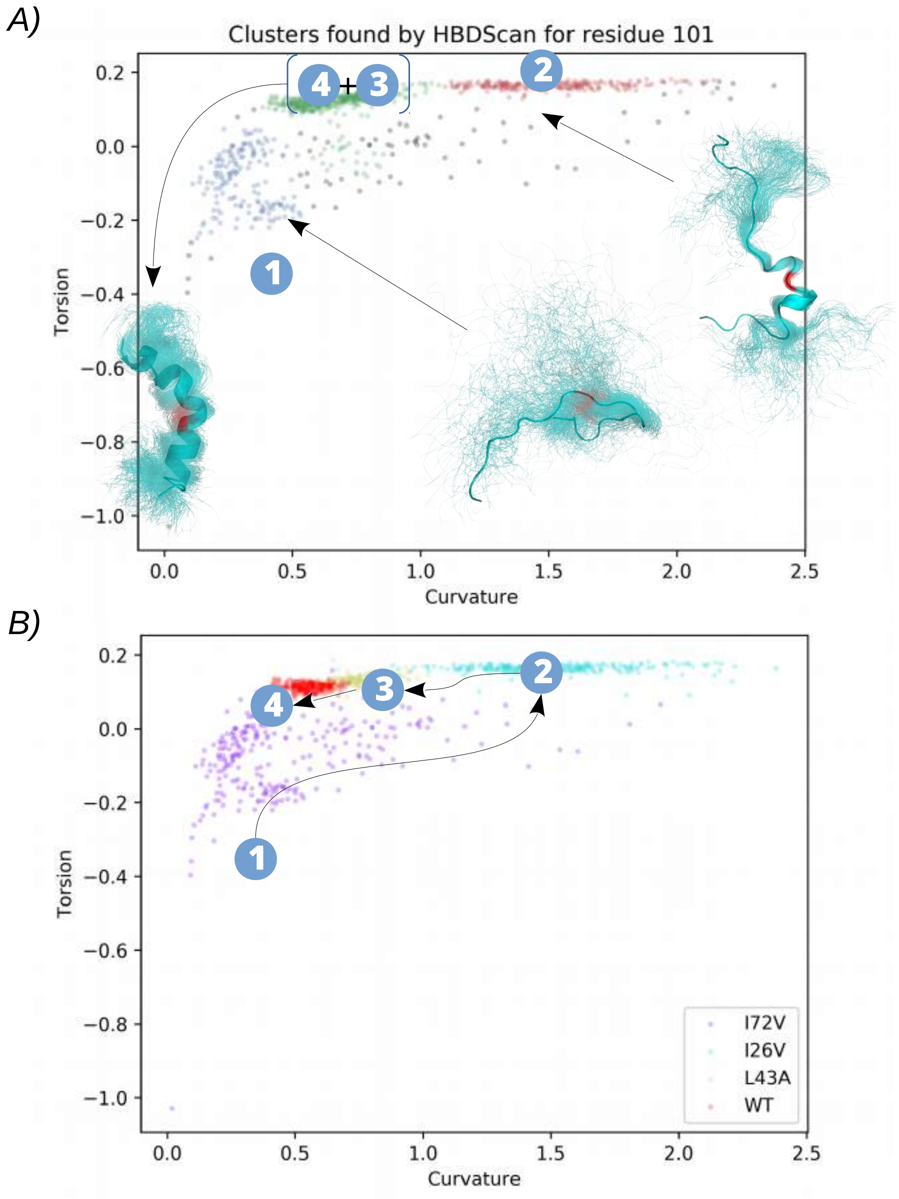

**Figure 8.** A) HDBScan clustering results of residue S101 $\kappa$ and $\tau$ values. Just as for E96, three clusters were identified, the first and second clusters correspond to conformations of the first and second binding steps, the third cluster correspond to the third and final step. The clusters conformations of each group are presented. Residue S101 (red) is highlighted on the clusters representation. B) Distribution of residue S101 $\kappa$ and $\tau$ values of each individual c-Myb-KIX binding steps, indicated by the KIX mutation, namely I72V (first step), I26V (second step), L43A (third step) and WT (Fourth step).

to obtain the results presented here in this work is publicly available at **http://xxx.xxx.xx/** for academic usage only.

As a future work, we devise a combination of this method with the CamSurf approach to generate surfaces from a conformational ensemble. CamSurf (to be published) represents the surface of each conformation as projections on a base of spherical harmonics polynomials in order to provide a robust analytic description of molecular surfaces. It also detects similarity among conformations in terms of projection coefficients to create a more compact representation where similar structural information is common to all surfaces, while retaining specific details of each one. Since the DG-based method groups similar conformations, the clustering technique presented here can be applied in CamSurf to separate

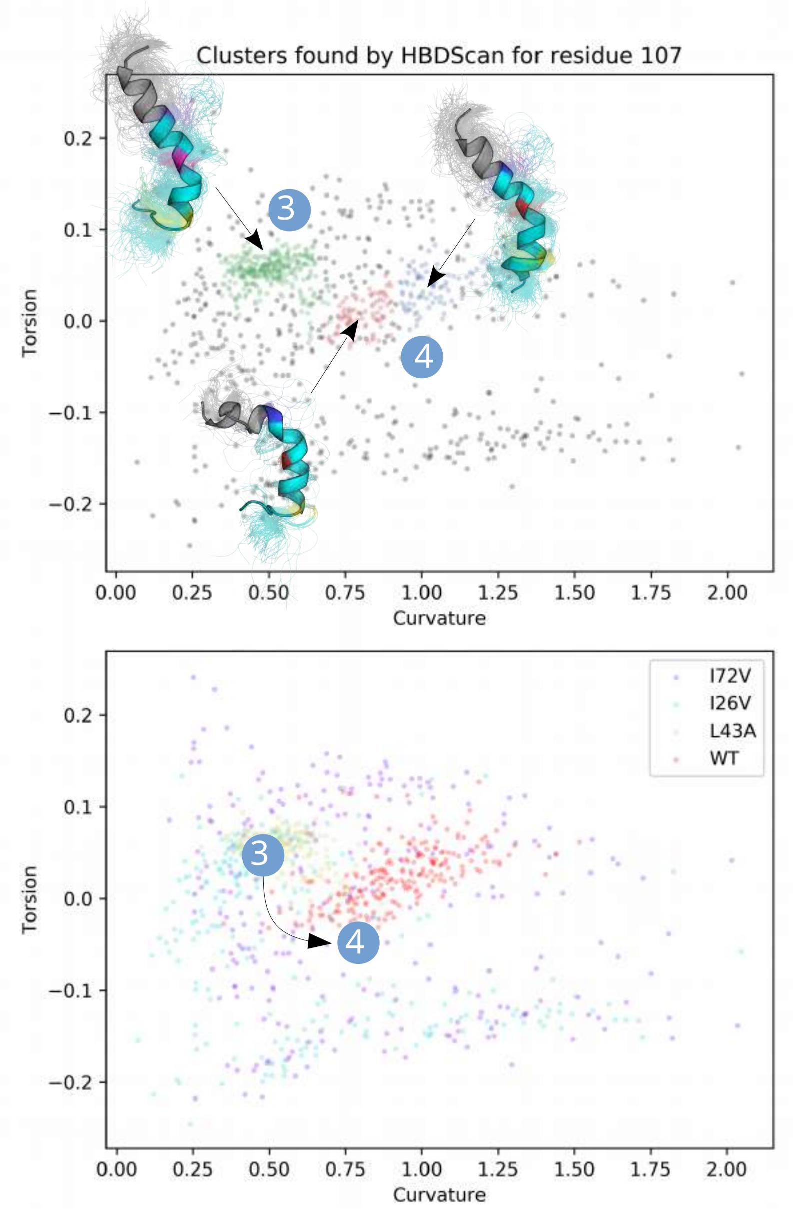

**Figure 9.** A) HDBScan clustering results of residue K107 $\kappa$ and $\tau$ values. Three clusters were identified, the first cluster contains most of the conformations of third binding step. The remaining clusters corresponds to the final binding step but divided in two groups. This clustering results captures the three different states of the last helix turn to fold. The clusters conformations of each group is presented. Residue 88-95 (grey), E96 (blue), S101 (red) and K107 (yellow) are highlighted on the clusters representation. B) Distribution of residue S101 $\kappa$ and $\tau$ values of each individual c-Myb-KIX binding steps, indicated by the KIX mutation, namely I72V (first step), I26V (second step), L43A (third step) and WT (Fourth step).

common structural information in more similar groups, therefore attaining higher reconstruction quality for a given level of commonality. As can be seen in figure **??**, CamSurf reconstructions considering the ensemble of cmyBEns as a whole have higher distortions, as indicated by higher average RMSD, than reconstructions of two separate clusters, thus allowing higher percentage of commonality among the projection coefficients and more pronounced memory space savings.

# ACKNOWLEDGMENTS

This work was supported by São Paulo Research Foundation (FAPESP) grant numbers 2013/18398-0 and 2013/20929-4.

# REFERENCES

Agafonov, R. V., Wilson, C., and Kern, D. (2015). Evolution and intelligent design in drug development. *Frontiers in Molecular Biosciences*, 2(May):27.

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.*, 42:106–116.

Campbell, E., Kaltenbach, M., Correy, G. J., Carr, P. D., Porebski, B. T., Livingstone, E. K., Afriat-Jurnou, L., Buckle, A. M., Weik, M., Hollfelder, F., Tokuriki, N., and Jackson, C. J. (2016). The role of protein dynamics in the evolution of new enzyme function. *Nature Chemical Biology*, 12(11):944–950.

Frederick, K. K., Marlow, M. S., Valentine, K. G., and Wand, a. J. (2007). Conformational entropy in molecular recognition by proteins. *Nature*, 448(7151):325–9.

Hinsen, K. (2006). Comment on: "Energy landscape of a small peptide revealed by dihedral angle principal component analysis". *Proteins: Structure, Function, and Bioinformatics*, 64(3):795–797.

Iakoucheva, L. M., Brown, C. J., Lawson, J., Obradović, Z., and Dunker, A. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology*, 323(3):573–584.

Javier Zea, D., Miguel Monzon, A., Fornasari, M. S., Marino-Buslje, C., and Parisi, G. (2013). Protein Conformational Diversity Correlates with Evolutionary Rate. *Molecular Biology and Evolution*, 30(7):1500–1503.

Kamadurai, H. B., Souphron, J., Scott, D. C., Duda, D. M., Miller, D. J., Stringer, D., Piper, R. C., and Schulman, B. A. (2009). Insights into Ubiquitin Transfer Cascades from a Structure of a UbcH5B-Ubiquitin-HECTNEDD4L Complex. *Molecular Cell*, 36(6):1095–1102.

Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E., and Andersson, M. (2015). SAXS-Guided Metadynamics. *Journal of Chemical Theory and Computation*, 11(7):3491–3498.

Laio, A. and Gervasio, F. L. (2008). Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science.

Leung, H., Montaño, B., Blundell, T. L., Vendruscolo, M., and Montalvão, R. W. (2012). ARABESQUE: A TOOL FOR PROTEIN STRUCTURAL COMPARISON USING DIFFERENTIAL GEOMETRY AND KNOT THEORY. *World Research Journal of Peptide and Protein*, 1(1):33–40.

Lobanov, M. Y. and Galzitskaya, O. V. (2015). How common is disorder? Occurrence of disordered residues in four domains of life. *International Journal of Molecular Sciences*, 16(8):19490–19507.

Ma, B. and Nussinov, R. (2016). Protein dynamics: Conformational footprints. *Nature Chemical Biology*, 12(11):890–891.

Montalvao, R. W., De Simone, A., and Vendruscolo, M. (2012). Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings. *Journal of biomolecular NMR*, 53(4):281–92.

Montalvão, R. W., Smith, R. E., Lovell, S. C., and Blundell, T. L. (2005). CHORAL: a differential geometry approach to the prediction of the cores of protein structures. *Bioinformatics (Oxford, England)*, 21(19):3719–25.

Parisi, G., Zea, D. J., Monzon, A. M., and Marino-Buslje, C. (2015). Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology*, 32:58–65.

Pitt, W. R., Montalvão, R. W., and Blundell, T. L. (2014). Polyphony: superposition independent methods for ensemble-based drug discovery. *BMC bioinformatics*, 15(1):324.

Rackovsky, S. and Scheraga, H. A. (1978). Differential Geometry and Polymer Conformation. 1. Comparison of Protein Conformations 1a,b. *Macromolecules*, 11(6):1168–1174.

Robustelli, P., Kohlhoff, K., Cavalli, A., and Vendruscolo, M. (2010). Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure (London, England : 1993)*, 18(8):923–33.

Saldaño, T. E., Monzon, A. M., Parisi, G., and Fernandez-Alberti, S. (2016). Evolutionary Conserved Positions Define Protein Conformational Diversity. *PLOS Computational Biology*, 12(3):e1004775.

Shimazaki, H. and Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *Journal of Computational Neuroscience*, 29(1-2):171–182.

Sittel, F., Jain, A., and Stock, G. (2014). Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *Journal of Chemical Physics*, 141(1).

Toto, A., Camilloni, C., Giri, R., Brunori, M., Vendruscolo, M., and Gianni, S. (2016). Molecular Recognition by Templated Folding of an Intrinsically Disordered Protein. *Scientific Reports*, 6(February):21994.

Tzeng, S.-R. and Kalodimos, C. G. (2012). Protein activity regulation by conformational entropy. *Nature*, 488(7410):236–40.

Uversky, V. N. (2014). The triple power of D3: protein intrinsic disorder in degenerative diseases. *Frontiers in bioscience (Landmark edition)*, 19:181–258.

Whitney, D. S., Volkman, B. F., and Prehoda, K. E. (2016). Evolution of a Protein Interaction Domain Family by Tuning Conformational Flexibility. *Journal of the American Chemical Society*, page jacs.6b05954.