NAR Genomics and Bioinformatics, 2024, **6**, lqae082
https://doi.org/10.1093/nargab/lqae082
Advance access publication date: 9 July 2024
Standard Article

Data-driven probabilistic definition of the low energy conformational states of protein residues

Jose Gavalda-Garcia [1,2,†], David Bickel [1,2,†], Joel Roca-Martinez [1,2], Daniele Raimondi [3], Gabriele Orlando [4] and Wim Vranken [1,2,*]

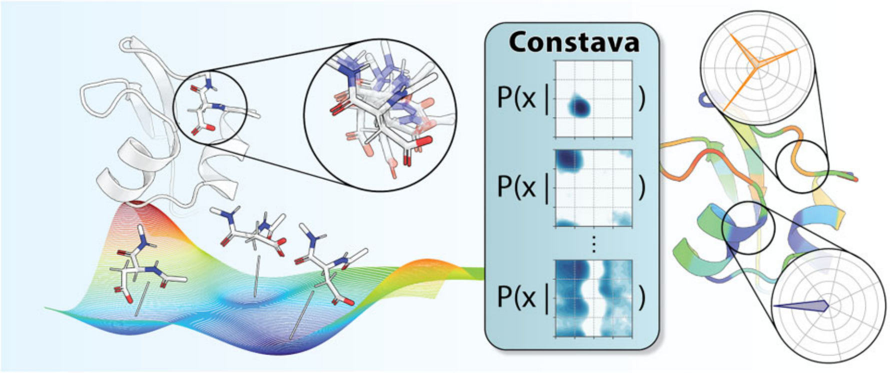
[1] Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, Brussels, Belgium
[2] Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium
[3] ESAT-STADIUS, KU Leuven, Leuven, Belgium
[4] Switch Laboratory, KU Leuven Leuven, Belgium

[*] To whom correspondence should be addressed. Tel: +32 2 6291952; Email: wim.vranken@vub.be
[†] The first two authors should be regarded as Joint First Authors.

Abstract

Protein dynamics and related conformational changes are essential for their function but difficult to characterise and interpret. Amino acids in a protein behave according to their local energy landscape, which is determined by their local structural context and environmental conditions. The lowest energy state for a given residue can correspond to sharply defined conformations, e.g. in a stable helix, or can cover a wide range of conformations, e.g. in intrinsically disordered regions. A good definition of such low energy states is therefore important to describe the behaviour of a residue and how it changes with its environment. We propose a data-driven probabilistic definition of six low energy conformational states typically accessible for amino acid residues in proteins. This definition is based on solution NMR information of 1322 proteins through a combined analysis of structure ensembles with interpreted chemical shifts. We further introduce a conformational state variability parameter that captures, based on an ensemble of protein structures from molecular dynamics or other methods, how often a residue moves between these conformational states. The approach enables a different perspective on the local conformational behaviour of proteins that is complementary to their static interpretation from single structure models.

Graphical abstract



Introduction

The dynamics and related conformational changes of proteins are often essential for their function, but are difficult to characterise and interpret. A key challenge is that this behaviour is typically not encompassed by the available data on protein conformation at atomic resolution, which mostly derive from static representations of the most stable protein conformations in crystalline state(s). The recent AlphaFold2 protein fold predictions are trained on such data and therefore essentially capture those static crystalline states. The dynamic and environmentally determined behaviour that proteins can exhibit in solution is, on the other hand, exemplified by intrinsically disordered regions (IDRs) and proteins (IDPs). These do not have a stable conformation but still perform essential functions in physiological conditions, often by temporarily acquiring a fold, for example when interacting with a binding partner (1). Changes in conformation are also necessary for prevalent allosteric mechanisms in folded

Received: February 13, 2024. **Revised:** June 14, 2024. **Editorial Decision:** June 24, 2024. **Accepted:** June 26, 2024

proteins (2), with even entire fold changes observed for the same protein (3).

To study a protein's conformational variability a variety of techniques can be applied, each with their own advantages and challenges. X-ray crystallography provides high-resolution structures but its ability to capture conformational changes is limited. Even though B-factors can be linked to protein flexibility and rigidity (4), the non-native conditions within crystals do not provide a realistic picture of dynamic protein behaviour, even at room temperature (5). Single particle cryo-electron microscopy, in contrast, allows the characterisation of proteins in near-native states. The resolution of the resulting structures directly correlates to the conformational diversity of the protein, with the possibility to classify heterogenous structural ensembles into different conformations (6). Nuclear magnetic resonance (NMR) provides in-solution information on dynamics of proteins at physiological conditions, from S^2 order parameters determined by relaxation experiments, which capture fast ps-ns timescale movements, to chemical shifts, which cover up to low ms timescale movements (7). The observations in NMR are typically averaged over all the protein molecules present in the sample, however. Thus, if a protein adopts multiple conformations between which it rapidly interchanges, the observed NMR data will show the linear average of a given parameter over the populations of these conformations. Hydrogen-deuterium exchange mass spectrometry (HDX-MS) provides residue- to peptide-resolution information on conformational changes in the ms–s time scale. Similar to NMR, the outcomes are averaged so that individual conformational populations cannot be resolved. Still, the technology is powerful in its ability to identify dynamic regions in proteins (8). Single-molecule Förster resonance energy transfer (smFRET) provides distance information on individual conformations in dynamic mixtures. For FRET experiments the proteins need to be dye labelled. These dyes are typically linked by highly flexible linkers to proteins, lowering the accuracy of the measurements (9). While all of these techniques allow the observation of the conformational diversity of proteins, there is no computational framework available to describe varying propensities towards different experimentally observed conformational states in variable protein regions.

A major challenge in describing the conformational variability of proteins is that their conformational space is extremely high-dimensional. Each amino acid residue in the protein backbone that is not fixed to a single conformation adds to the overall conformational complexity of the protein. Conversely, this provides the opportunity to describe the conformational diversity of a protein in terms of the local conformational freedom of the individual residues. A residue's conformational preferences are determined by its structural context, interactions with other molecules and environmental conditions such as pH or ionic strength, all of which shape its local energy surface. The lowest energy state for a given residue can then correspond to very sharply defined conformations, for example when the residue is part of a stable helix, or can cover a wide range of conformations, for example for residues in intrinsically disordered regions. Amino acid residues that do change conformation under native conditions can therefore, by definition, not be assigned to a specific secondary structure. They should rather be interpreted in probabilistic terms within the secondary structure space where they transition. This requires a definition of low energy conformational states that residues typically adopt in proteins, as well as a description of the conformational variability within these low energy conformational states. To derive such low energy states, ideally, large scale experimental data should be employed to extract statistically relevant behaviours.

Information about the behaviour of amino acids in proteins can be estimated from chemical shifts, the most readily publicly available NMR parameter from the Biological Magnetic Resonance Data Bank (BMRB) (10). This has resulted in methods to estimate the backbone and side-chain dynamics of amino acids (11,12), secondary structure populations (13), and a combined residue-independent metric based on unsupervised machine learning (14). Such parameters summarise the statistically averaged behaviour of residues of proteins in solution, so reflecting their lowest energy state in the employed experimental conditions. NMR structure ensembles are also calculated from NMR observables, notably interatomic distances from the nuclear Overhauser effect (NOE), dihedral angle ranges from dipolar couplings and chemical shifts and bond orientations from residual dipolar couplings (RDCs). Typically, time-averaged NMR data are used to calculate a single NMR model at a time, so resulting in information clashes where mutually incompatible conformations are present in the protein (15). NMR structure ensembles are typically collections of such models, therefore residues that adopt a single well-defined conformation in solution, with all experimental data consistent, will tend to have the same consistent structural definition in all models of the NMR ensemble. For residues that adopt multiple distinct conformations in large enough populations, on the other hand, the observed experimental data will be incompatible within the same structure, so resulting in variable structure definitions in the models of the NMR ensemble. The concept of using such information from NMR ensembles was already successfully employed in for example the Espritz-NMR disorder predictor (16).

However, capturing the inherently probabilistic information from protein structure models that encompass multiple conformations is difficult. To date, no method is available that can describe varying propensities towards different conformational states in highly variable regions. We here propose a data-driven probabilistic definition of six low energy conformational states that are typically accessible for amino acid residues in proteins. This definition is based on solution NMR information for 1322 proteins, through a combined analysis of structure ensembles with interpreted chemical shifts. Each conformational state corresponds to a specific region in the backbone dihedral space (Figure 1). Within these regions, residues have the freedom to adopt various conformations without encountering significant energy barriers. We further introduce a conformational state variability parameter that captures how often a residue moves between these conformational states. Such transitions would indicate changes in the local environment of the residues, so resulting in a change of their low energy state. The approach therefore enables a different perspective on the conformational behaviour of proteins that is complementary to their static interpretation from single structure models. To demonstrate the effectiveness of our method, we apply it to a set of molecular dynamics (MD) simulations consisting of 113 proteins that exhibit diverse behaviours ranging from structurally rigid to highly dynamic disordered proteins. Additionally, we apply our method to large ensembles of selected intrinsically disordered proteins

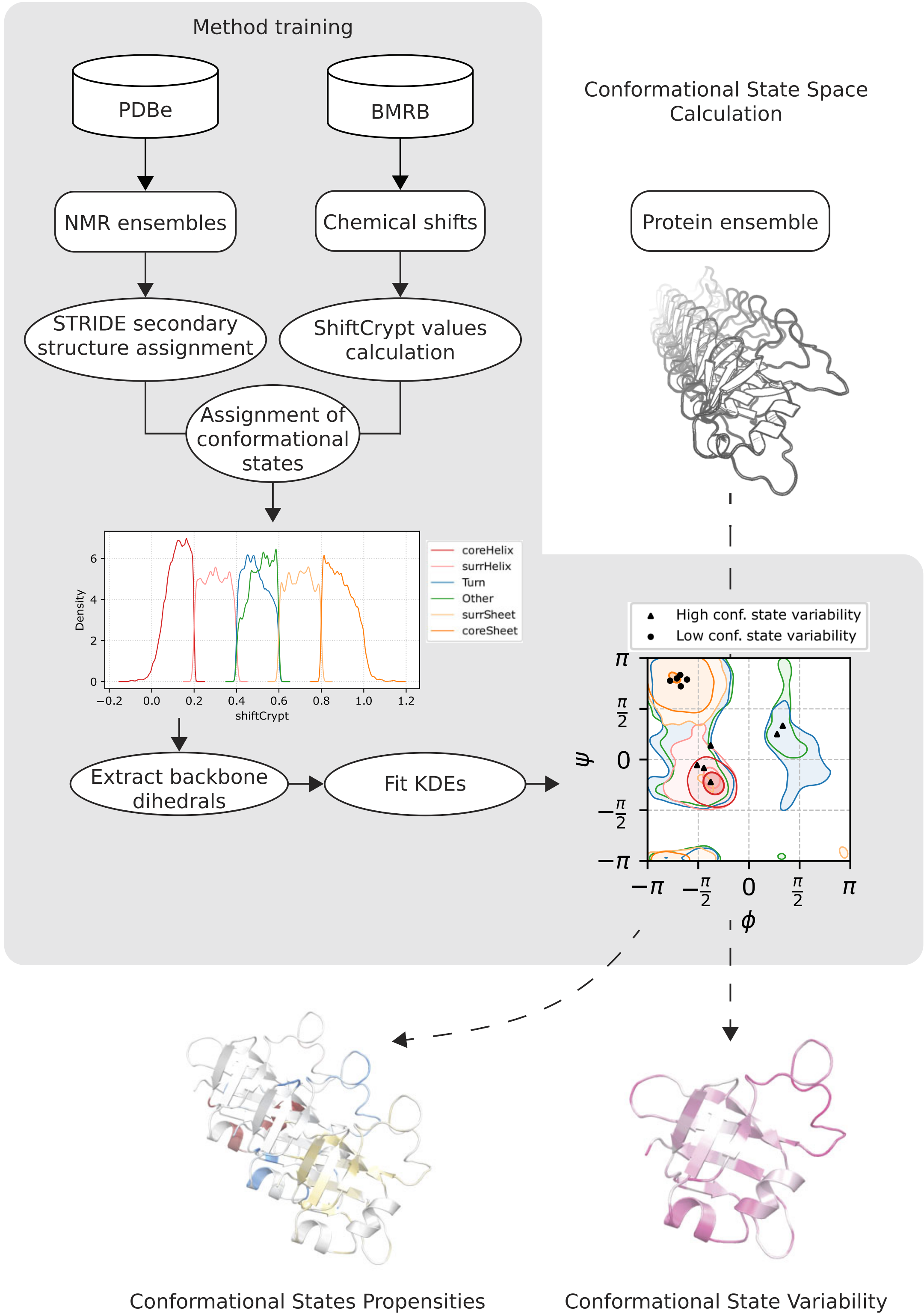

Method training
PDBe
BMRB
NMR ensembles
Chemical shifts
STRIDE secondary structure assignment
ShiftCrypt values calculation
Assignment of conformational states
Density
shiftCrypt
coreHelix
surrHelix
Turn
Other
surrSheet
coreSheet
Extract backbone dihedrals
Fit KDEs
Conformational State Space Calculation
Protein ensemble
High conf. state variability
Low conf. state variability
ψ
ϕ
Conformational States Propensities
Conformational State Variability

Figure 1. Visual representation of the KDE training workflow and the calculation of conformational state propensities and conformational state variability. The training of the models (grey background region) employs NMR ensembles from the PDBe and chemical shift information from the BMRB. The NMR ensembles are used to obtain secondary structure assignments of each residue. The chemical shifts are employed to calculate the ShiftCrypt values. Both data sources are combined to assign the residues to a conformational state. Once the conformational state assignment of these residues is completed, their backbone dihedrals are extracted and used to fit six KDEs, one per conformational state. With these trained KDEs the conformational states probabilities as well as the conformational state variability can be calculated (dashed arrows).

(IDPs) to showcase its ability to identify regions of relative order within otherwise disordered proteins.

We implemented the approach as a python package, named Constava, that is available on PyPI: https://pypi.org/project/constava/.

Materials and methods

An overview of the connection between the Methods sections below is provided in Figure 1.

Generation of probabilistic model of conformational states

Data collection

Chemical shift data were collected from the Biological Magnetic Resonance Data Bank (BMRB) (10), using previously described criteria (17). Summarised, only entries between reported pH 5–7, temperature 293–313 K were selected for which chemical shift data were available for ^{1}H, ^{13}C and ^{15}N for at least half the residues in the sequence. Entries with samples containing agents that strongly influence protein behaviour (e.g. TFE, SDS, etc.; for complete list consult Supplementary Table S1 in (17)) were excluded, and chemical shift re-referencing was performed with VASCO (18). Only BMRB entries with corresponding NMR ensembles of at least 10 models were retained, resulting in 1414 entries totaling 147 116 residues. STRIDE was applied to all individual models of the NMR ensemble (19,20). STRIDE uses structural information to assign secondary structure labels to each residue and thereby provides a categorical description of the conformation of that residue (for each model in the NMR ensemble). ShiftCrypt (14) was applied to the chemical shifts to obtain a single value per residue in the range 0 to 1 that relates to that residue's behaviour in relation to secondary structure propensity and backbone dynamics. ShiftCrypt values near the extremes point towards residues that adopt rigid helix (near 0) or sheet (near 1) conformations, while deviations towards the center indicate increasingly dynamic residues with decreasing degrees of helix/sheet propensity. The data are available in Zenodo: https://doi.org/10.5281/zenodo.10371447.

Conformational state definition

Based on the ShiftCrypt values and STRIDE (19,20) secondary structure assignments of residues in NMR ensembles, we define six conformational states. We then annotated all residues in the NMR ensembles according to these conformational states, resulting in 62 125 annotated residues from 1322 protein ensembles. To obtain structural insights on conformational space that residues in each state cover, we extracted the backbone dihedral angles ϕ and ψ for all residues in a given category. To avoid the over-/underrepresentation of individual structures based on the number of models deposited in their NMR ensembles, we randomly selected 5 models for each structural ensemble assigned to a given BMRB ID. This yielded a set of (ϕ, ψ)-pairs for each of the conformational states.

- *Core helix*: Residues that exclusively adopt helical conformation in all models of their associated NMR ensemble with shiftCrypt values ≤ 0.2; $N = 93\,957$ residues
- *Surrounding helix*: Residues that adopt helical conformation in the majority of models of their associated NMR ensemble with shiftCrypt values in the range (0.2, 0.4]; $N = 8180$ residues
- *Core sheet*: Residues that exclusively adopt extended conformation in all models of their associated NMR ensemble with shiftCrypt values ≥ 0.8; $N = 47\,280$ residues
- *Surrounding sheet*: Residues that adopt extended conformation in the majority of models of their associated NMR ensemble with shiftCrypt values in the range [0.6, 0.8); $N = 11\,280$ residues
- *Turn*: Residues that adopt turn conformation in the majority of models of their associated NMR ensemble with shiftCrypt values in the range (0.4, 0.6); $N = 75\,377$ residues
- *Other*: Residues that adopt coil conformation in the majority of models of their associated NMR ensemble with shiftCrypt values in the range (0.4, 0.6); $N = 74\,542$ residues

Gaussian kernel density estimators were then built for each conformational state using the scikit-learn python package with bandwidth for the Gaussians equal to 0.13 radians. They represent how likely it is that a residue in a conformational state (as defined above) occupies a certain region in the backbone dihedral space as observed in the NMR ensemble. The robustness of these probabilistic models is assessed in Supplementary Figure S1.

Inference of conformational state propensities

Given a conformational ensemble, the kernel density estimates can then be used to infer the conformational state propensities for each residue. To do so, (ϕ, ψ)-angle pairs are extracted from the conformational ensemble and divided in M samples of N (ϕ, ψ)-pairs. This can be done by bootstrapping N data points M times, or by using a sliding window of size N. For each of these samples the evidence is calculated by evaluating the probability density functions (PDF) at the given (ϕ, ψ)-coordinates (equation (1)). Then, the likelihood for any state is its evidence relative to all other evidence (equation (2)). We generally report the average over all M samples obtained from the ensemble.

$$\text{Evidence}_m = \prod_{n=0}^{n<N} \text{PDF}_{\text{state}}(\phi_n, \psi_n) \tag{1}$$

$$\text{Likelihood}_m = \frac{\text{Evidence}_m}{\sum^{\text{all states}} \text{Evidence}} \tag{2}$$

Notably, the sample size N chosen to calculate the evidence strongly impacts the certainty with which conformational states are inferred. For small sample sizes ($N \leq 5$), the method captures likelihoods that reflect the local backbone dynamics and conformational ambiguity, while for larger sample sizes ($N > 20$) the likelihoods result in a categorical assignment of a preferred conformational state for most residues (see Supplementary Figures S2 and S3). In general, we suggest a value of $N = 3$ to obtain information of the conformational state variability (see below), and a value of $N = 25$ to extract preferred conformational states.

Definition of conformational state variability

In a conformational ensemble a residue may not only undergo conformational changes (i.e. variation of its ϕ- and ψ-angles captured by the conformational state KDEs), but may change its conformational preference altogether. We label this as conformational state variability, expressing the ability of a residue to switch back and forth between conformational states. Note that this does not directly relate to changes of the ϕ- and ψ-angles but captures the ability of a residue to exist in multiple conformational states, taking into account the structural variation that would be expected for any residue in that given conformational state.

To calculate the conformational state variability of a given residue in an ensemble we calculate the root mean square distance from the average state across all M samples (equation (3)). Here, L_m is the likelihood of one sample to belong to a conformational state as described in equation (2) and $\langle L_m \rangle$ is the average likelihood across all samples.

$$\text{Conf. State Var.} = \sqrt{\frac{1}{M}\sum_{m}^{M}\sum_{i}^{\substack{\text{Conf.}\\\text{states}}}\left(L_{m,i} - \langle L_m \rangle\right)^2} \qquad (3)$$

Molecular dynamics simulations

Reference molecular dynamics (MD) simulations were performed for 113 proteins in order to validate the method proposed in this work and to identify parameters that allow the study of different aspects of conformational state variability. The simulated proteins cover a wide range of proteins, from disordered to partially ordered to fully ordered. The disordered and partially ordered entries were selected from MobiDB (21) based on their coil percentage, using as criteria >90% coil (disordered) and approximately 50% coil (partially ordered), respectively. Only entries with an annotation coverage higher than 90% were included. A diverse set of rigid structures was also defined by selecting proteins with different architectures and topologies based on their CATH code (22). All the selected proteins had structures available that were experimentally determined by NMR and had an associated entry on BMRB with extensive NMR chemical shift data available (^{13}C, ^{15}N and ^{1}H). The full protein list with the selected PDB codes that were used as starting conformations is available as supplementary information (see Supplementary Data 1).

The MD simulations were prepared and performed with GROMACS (23) using the CHARMM36m force field (24). All 113 proteins were subjected to the same protocol to yield an internally consistent data set. Starting from the PDB structure (1st model of an NMR ensemble), water molecules and heteroatoms were removed. The systems were solvated using TIP3P water (25) in a rhombic dodecahedron simulation box and Na$^+$ or Cl$^-$ ions were added to neutralise the systems. Then, the systems were minimised, followed by two equilibration steps. In the first step, the temperature was equilibrated at 300 K over 1 ns of simulation under NVT ensemble. The Bussi thermostat was used with separate heat bath couplings for solute and solvent (26). The second equilibration step was performed over 1 ns under NPT ensemble to keep the pressure at 1 bar employing the Parrinello-Rahman barostat (27).

We assessed the impact of sampling frequency and overall simulation length on the sampling of protein backbone conformations on five different proteins (PDB entries 1CLL, 1D3Z, 1QM3, 5PTI, and 2KUY, Supplementary Figure S4). All simulations were performed for 1000 ns and coordinates were recorded every 1 ns to limit the correlation between subsequent sets of coordinates. To assess whether the sampling time was sufficient to capture experimentally observed conformational states, we compared the conformational diversity in the MD trajectories against structures annotated in the CoDNaS database (28). For the 22 proteins represented in the database, we found that the MD sampling either exceeds the conformational variability recorded in CoDNaS or produces similar results (Supplementary Figure S5).

To calculate the conformational state propensities and variability for each residue, the backbone (ϕ, ψ)-angles were extracted for all trajectories using GROMACS' chi module. Then, conformational state propensities and variability were inferred as described above; samples of size 3 and 25 were drawn by bootstrapping for 10 000 times from the trajectory. Bootstrapping was chosen as with long sampling intervals independence of subsequent samples is assumed. Sampling size 3 was used to analyse local dynamics while sample size 25 informs on the conformational state propensity.

For comparison with traditional secondary structure assignments, we applied DSSP (29,30) and STRIDE (19,20) to all conformations extracted from the trajectories. For DSSP we used GROMACS' do_dssp module, while for STRIDE all frames of the MD trajectories were extracted as individual PDBs to be analysed. We further calculated DSSP and STRIDE propensities as the fraction of frames per residue that are assigned to a given secondary structure category (H, E, ...).

Analysis of α- and β-synuclein

We obtained structural ensembles for both α-synuclein (αSyn) and β-synuclein (βSyn) from the Protein Ensemble Database (αSyn: PED00024, βSyn: PED00003) (31–33). The backbone (ϕ, ψ)-angles were extracted from the conformational ensembles. To calculate the conformational state propensities and variability for each residue as described above, samples of size 25 were drawn by bootstrapping for 500 times.

Results

In this section, we discuss application cases for our probabilistic approach to define protein conformation and dynamics. We will describe the relation between the metrics and established measures of conformation, and finally apply the method to structural ensembles derived from MD simulations and NMR experiments.

Probabilistic definition of conformational states

The definition of the conformational states is based on the analysis of NMR ensembles, using the assumption that residues with a consistent structural definition in all models of the NMR ensemble adopt a single well-defined conformation, whereas residues with variable structure definitions adopt a wider range of conformations. This definition is further refined based on a ShiftCrypt-based interpretation of chemical shift information, which reflects the in-solution behaviour of residues. In total 1322 NMR structural ensembles (62 125 residues) were so analysed and their residues subdivided into six conformational state categories (see Table 1): *Core helix*, *Surrounding helix*, *Core sheet*, *Surround-*

Table 1. Description of conformational states and conformational state variability

Parameter	Abbreviation	Definition	Traditional interpretation
Conformational state variability	-	The change in conformational states probability vectors based on the dihedral angles of models in a protein structure ensemble. Residues with changing conformational states propensities within the different structures of the ensembles feature higher values.	The capability of a residue to adopt diverse conformational states.
Core helix	coreHelix	Residues with minimal backbone dynamics that exclusively exhibit α-helical conformation.	Residues in very stable rigid α-helices.
Surrounding helix	surrHelix	Residues with significant backbone dynamics that mostly exhibit α-helical conformation, but undergo conformational changes.	Residues found in kink regions of α-helices, but also encompassing transient helices, flexible helix termini, π-, and 3_{10}-helices.
Core sheet	coreSheet	Residues with minimal backbone dynamics that exclusively exhibit β-sheet conformation.	Residues in stable rigid β-sheets but also in fully extended regions.
Surrounding sheet	surrSheet	Residues with significant backbone dynamics that mostly exhibit β-sheet conformation, but undergo conformational changes.	Residues found in kink regions and in the flexible termini of β-sheets as well as residues that preferably adopt the ppII state.
Turn	Turn	Residues with high backbone dynamics labeled as mostly Turn conformation by STRIDE.	Flexible residues found in turn-like conformations.
Other	Other	Residues with high backbone dynamics labeled as mostly Coil conformation by STRIDE.	Flexible residues that cannot be properly described by any of the other classes, typically found in disordered regions.

ing sheet, *Turn* and *Other*. To obtain probabilistic models of these states in the backbone dihedral space, as extracted from NMR ensembles, Gaussian kernel density estimators (KDE) were trained. The conformational state propensities obtained with the dihedrals and each of the KDEs are a quantitative measure of the likelihood of a residue to assume this conformational state at any time of observation, given its backbone dihedral angles. These conformational states so reflect typical experimentally observed low energy states of residues (Figure 2). The corresponding backbone dihedral space distribution indicates the expected range of backbone dihedrals that are available to a residue in that conformational state without having to overcome significant energy barriers. In particular the *Core helix* and *Core sheet* states exhibit deep wells in their energy landscapes, that in turn reflect the restrained backbone dynamics of such residues. In contrast, energy wells in *Turn* and *Other* states are more shallow, allowing for higher conformational freedom. A residue's conformational state, thus, captures the preferential conformation of residues, as traditional secondary structure classification does, as well as dynamical properties of the local backbone.

We further define the conformational state variability parameter as the ability of a residue to adopt multiple conformational states in a given protein in a specific environment. This variability does not reflect actual dynamics, but rather indicates whether that residue is switching between different low energy conformational states. For example, a residue in states *Turn* or *Other* may exhibit high backbone dynamics (its (ϕ, ψ)-angles change constantly) while still having a low conformational state variability (its overall behaviour is best described by one conformational state). This would be a typical conformational state adopted by residues in disordered regions of proteins. A residue in a well-defined stable helix, on the other hand, will also have low conformational state variability as well as very low backbone dynamics.

Constava captures transient conformational states from MD ensembles

The new metric was tested on a diverse set of conformational ensembles extracted from MD simulations of 1000 ns length for 113 proteins exhibiting varying degrees of structural order/disorder. From the trajectories (ϕ, ψ)-angles were extracted at intervals of 1 ns, to minimise the correlation between subsequent snapshots. The conformational state propensities were then inferred from each individual (ϕ, ψ)-angle pair, or using subsets of (ϕ, ψ)-angle pairs of varying sizes (Figure 3A and B).

To investigate the impact of the number of (ϕ, ψ)-angle pairs in the inference, a bootstrapping procedure was first employed to sample N (ϕ, ψ)-angle pairs from random time points from the MD simulations. Increasing the number of N, where more angle pairs are considered at the same time, leads to residues being assigned more preferentially to specific conformational states. This is due to the conformational states sharing overlapping regions in the (ϕ, ψ)-space, so that a single angle pair data point will correspond to multiple regions (Figure 2). When increasing the sample size N, however, the information available to the model to assign a specific conformational state increases and results in more unique assignments of the conformational state. Further increases of N also then start to bias unique conformational states, with lower populated states that the protein might transition through disregarded. Thus, to assess a protein's conformational state variability lower values of N are more informative, whilst higher values reflect the most preferred conformational state(s). Based on our analysis of the impact of the sample size N on the results, we propose two sampling sizes also used throughout our work: $N = 3$ to study conformational state variability, as it provides the best compromise between a linear distribution values and a reasonable data range (Supplementary Figures S2 and S3), while $N = 25$ is well suited to obtain preferred conformational

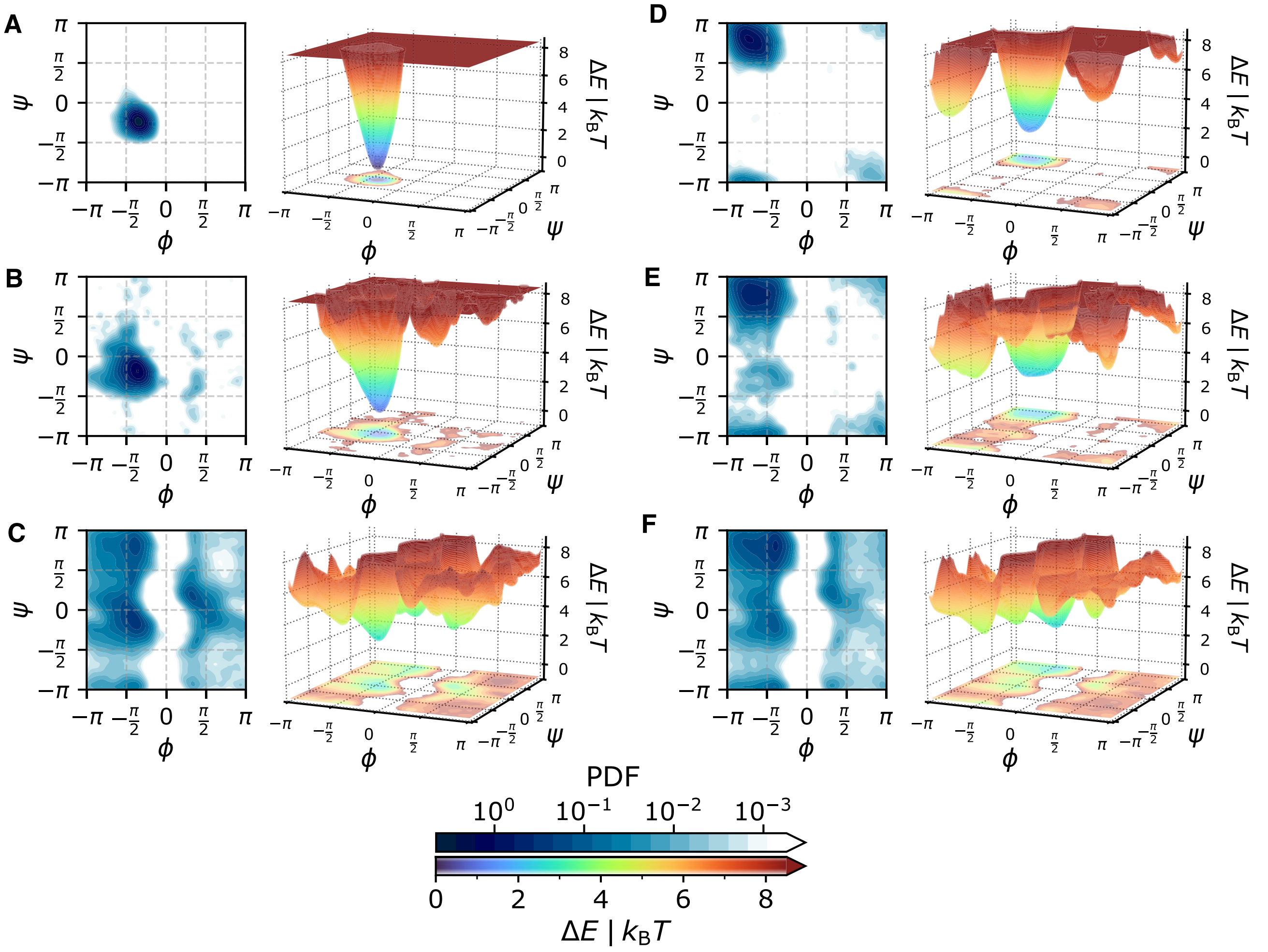


Figure 2. Definition of the six conformational states in (ϕ, ψ)-space. Each panel shows a conformational state represented by a continuous probability density function on the left, and the derived potential energy surfaces in the (ϕ, ψ)-space on the right. The conformational states are shown for: (**A**) *Core helix*, (**B**) *Surrounding helix*, (**C**) *Turn*, (**D**) *Core sheet*, (**E**) *Surrounding sheet* and (**F**) *Other*. The potential energy surfaces illustrate how, e.g. a *Core helix* residue is conformationally restricted by high energy barriers, while *Turn* residues can adopt a wide range of backbone conformations without having to overcome such high energy barriers. This is further exemplified by the 2D projection in Supplementary Figure S12.

states without losing information on less populated states (see Supplementary Figure S6).

The random sampling of the bootstrapping procedure, however, loses any information about connections between models, such as the timeline in an MD simulation. A sliding-window approach was therefore implemented, which instead of randomly sampling from the whole population chooses blocks of adjacent (time-related) samples. This resulted in a much more sensitive detection of transient conformational states than regular bootstrapping (Figure 3C and D). Moreover, by gradually extending the window size of the sliding window, information about the time intervals in which these states are transitioning between themselves can be obtained. Small window sizes pick up more transient states of shorter duration, while longer window sizes only detect those states that are stable for a longer time (Figure 3C and E). Supplementary Figure S15 shows an overview over the state transitions observed in our conformational dataset. Due to the higher sensitivity, using a sliding window is the preferred subsampling method for data obtained from time series (e.g. MD trajectories). For ensembles where no relation between adjacent ensemble members can be assumed (e.g. NMR ensembles), bootstrapping is the method of choice, as there is no direct connection between the sampled models.

Relation between NMR-defined conformational states and traditional secondary structures assignments

DSSP (29,30) and STRIDE (19,20) are two of the most common tools to assign static secondary structure on protein structures. By applying them on every structure of an ensemble, conformational variable regions would receive changing assignments. To compare how such serial secondary structure assignments differ from our probabilistic method, we applied Constava, DSSP and STRIDE to 113 MD ensembles. For the narrowly defined *Core helix* and *Core sheet* conformational states, near linear relations to H (α-helix) and E (β-sheet), respectively, emerged (Supplementary Figure S7A, C and S8A, C). This demonstrates that high propensities of *Core helix* and *Core sheet* indicate rigid residues that adopt stable secondary structure conformations, which are equally captured by DSSP and STRIDE. For the more dynamic conformational states, relationships with traditional secondary struc-

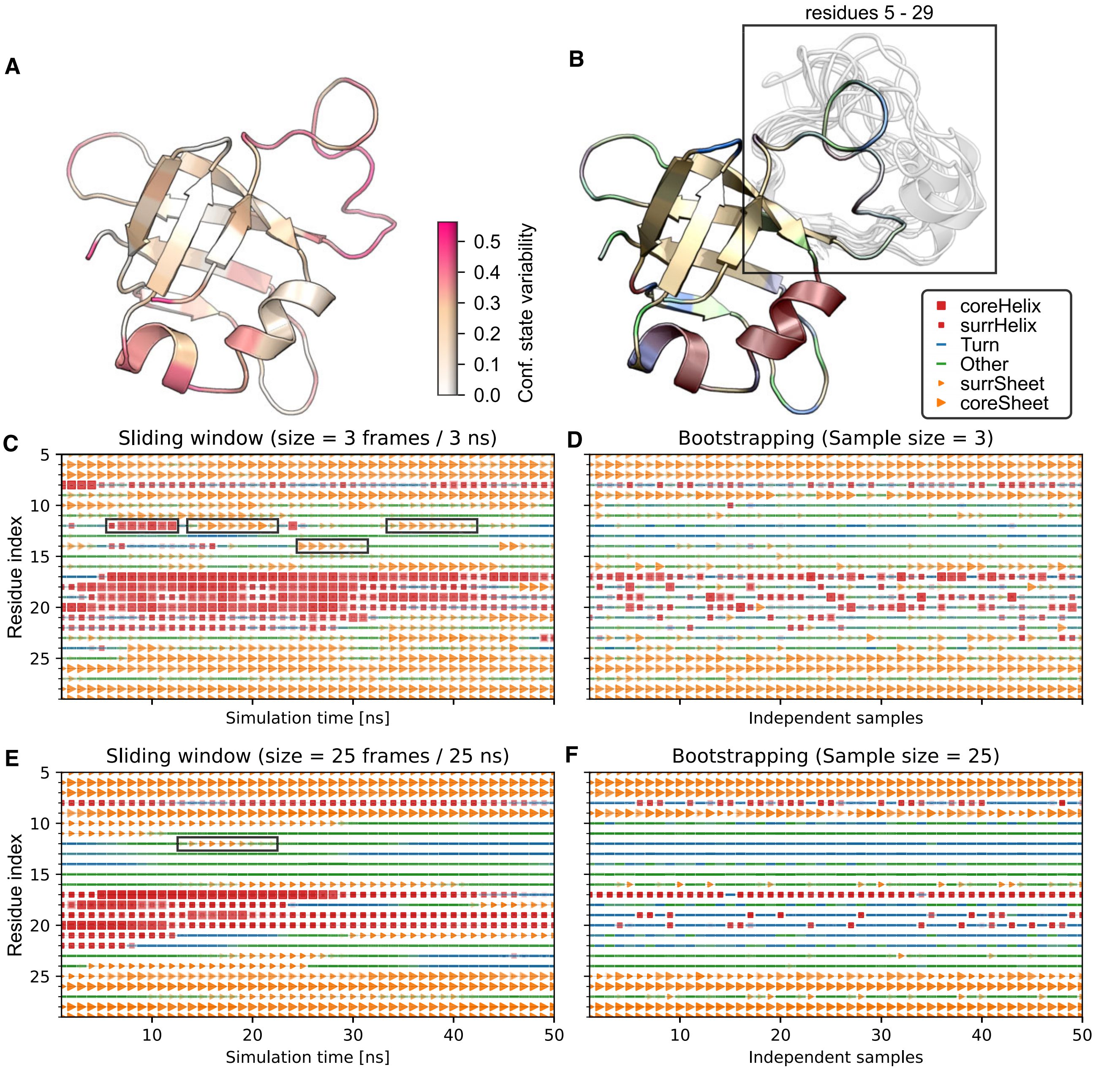


Figure 3. Analysis of MD simulation ensembles with Constava. The method was applied on the conformational ensemble derived from 100 ns of MD simulation on *E. coli* ribosomal protein L25 (PDB ID: 1b75). (**A**) Conformational state variability mapped on the structure. Regions of high variability are the loop region as well as the C-terminal end of the helical region, which switches between helical and turn-like states. (**B**) Conformational states propensities mapped on the structure. The rectangle highlights residues 5–29, where the largest conformational changes occur. (**C**) Conformational states as a time series along the simulation, using a sliding window with $N = 3$ (here 3 ns). Black boxes highlight transient conformational states apparent at this time-resolution. (**D**) Conformational states obtained from bootstrapping ($N = 3$). Adjacent samples are not related. Transient states are sometimes detected, but are underrepresented in comparison to the sliding-window method. (**E**) Conformational states as a time series along the simulation, using a sliding window with $N = 25$ (here 25 ns). Black boxes highlight transient conformational states apparent at this time-resolution. Notably, fewer transient states are detected compared to the smaller window size (panel C). (**F**) Conformational states obtained from bootstrapping ($N = 25$). With increasing sample size the likelihood to detect low populated transient conformational states further diminishes, and the detected conformational states increasingly converge on unique solutions.

tures contributions are more complex, without clear correlations (Supplementary Figure S7B, D–F and S8B, D–F).

However, the probabilistic nature of the conformational states makes Constava more robust in detecting secondary structure propensities from ensemble data than approaches that only consider static protein structures like DSSP and STRIDE. When analysing conformational ensembles from e.g. MD simulations, Constava has the potential to detect residues moving towards a secondary structure state, well before it is recognised by DSSP or STRIDE (Figure 4 and Supplementary Figure S9). In this example, Constava picks up on the propensities of multiple consecutive residues to form a helix long before the helix, as defined by DSSP and STRIDE, emerges. In the study of IDPs that exhibit high conformational freedom, so that relevant transient conformations cannot be efficiently sampled, this ability of Constava could be extremely

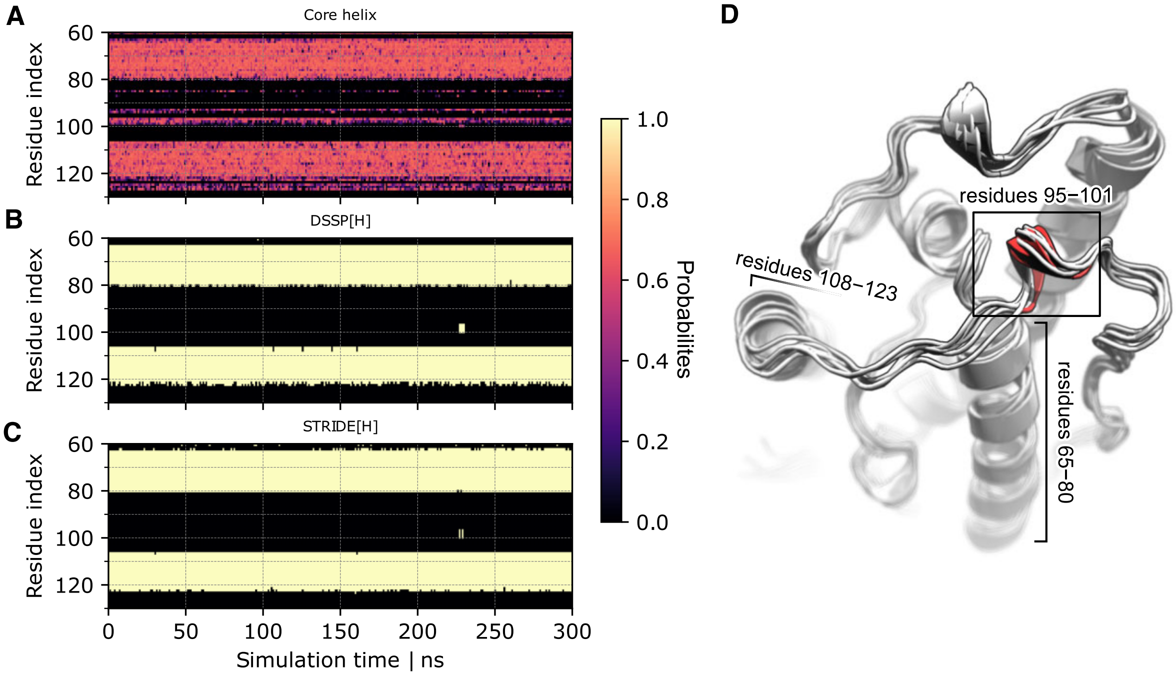


Figure 4. Comparison of conformational state propensities with traditional secondary structure assignments. The plot shows residues 60–130 of Endonuclease V (PDB ID: 2end). (**A**) Assignments of *Core helix* propensities for the first 300 ns of the simulation. Notably, Constava continuously detects *Core helix* propensities ∼0.6 for residues 95–100. (**B**) DSSP assignment of H (α-helix) for the first 300 ns of the simulation. As DSSP performs a classification, the propensities for H are 0 or 1. The transient helix for residues 95–100 only appears shortly after more than 200 ns of simulation. (**C**) STRIDE assignment of H (α-helix) for the first 300 ns of the simulation. As STRIDE performs a classification, the propensities for H are 0 or 1. The transient helix for residues 95-100 only appears shortly after more than 200 ns of simulation. (**D**) Structure of Endonuclease V with regions shown in panels A, B and C labeled.

useful in picking up pre-sampled conformations that are for example relevant for forming interactions, reducing the need for unachievable exhaustive sampling.

Relation of conformational state variability to traditional dynamics metrics

The MD simulations enable a comparison of the conformational state variability with established metrics. The root mean square fluctuations (RMSF) and circular variance (CV), as calculated from the MD simulations, and the S^2 order parameter derived from the NMR chemical shifts using the random coil index (S^2_{RCI}) are shown in relation to the conformational state variability, calculated with the bootstrap method for 3 samples repeated 10 000 times, in Figure 5 and Supplementary Figure S10. There is a very weak correlation between RMSF and conformational state variability (see Figure 5A), depicting a tendency for residues with low RMSF to also show low conformational state variability (e.g. in rigid stable helix). Whereas RMSF quantifies all backbone displacements, even if they do not affect the conformational state, the conformational state variability only measures switches between conformational states at the residue level.

There is a more pronounced correlation between CV and conformational state variability (Figure 5B). CV represents the variation in the backbone dihedral angle space, with no incorporation of their conformational meaning. In contrast, conformational state variability represents transitions between conformational states that do not directly correlate with changes in the backbone dihedrals, as each conformational state accounts for varying degrees of backbone dynamics. Therefore, higher circular variance does not necessarily lead to high conformational state variability. For instance, a highly flexible coil would consistently occupy the *Other* state (low conformational state variability), while displaying high backbone dynamics (high CV). For such a residue to show high conformational state variability as well, it would need to switch conformational states, e.g. by partaking in the formation of a transient helix.

The S^2_{RCI}, a measure of backbone rigidity derived from chemical shifts for the protein in solution, demonstrates a negative correlation to the conformational state variability (see Figure 5C). Yet, the relationship is not directly linear. The cluster of residues with low conformational state variability and high S^2_{RCI} values (bottom right cell) represents residues with neither backbone movement nor changes in the conformational state space, such as in rigid stable helix. The region with S^2_{RCI} values below 0.69 predominantly features high conformational state variability values (top left cell), indicating highly dynamic residues that do change their conformational state. Note that there are also residues that have low conformational state variability, indicating they adopt a unique state (*Other* or *Turn*).

Finally, many residues have S^2_{RCI} higher than 0.69, indicating residues where the NMR chemical shift values point to low backbone dynamics, but that still exhibit high conformational state variability. These represent residues that transition through the conformational state space but where these changes are not captured by the NMR chemical shift values (top middle and right cells). Since NMR chemical shift values

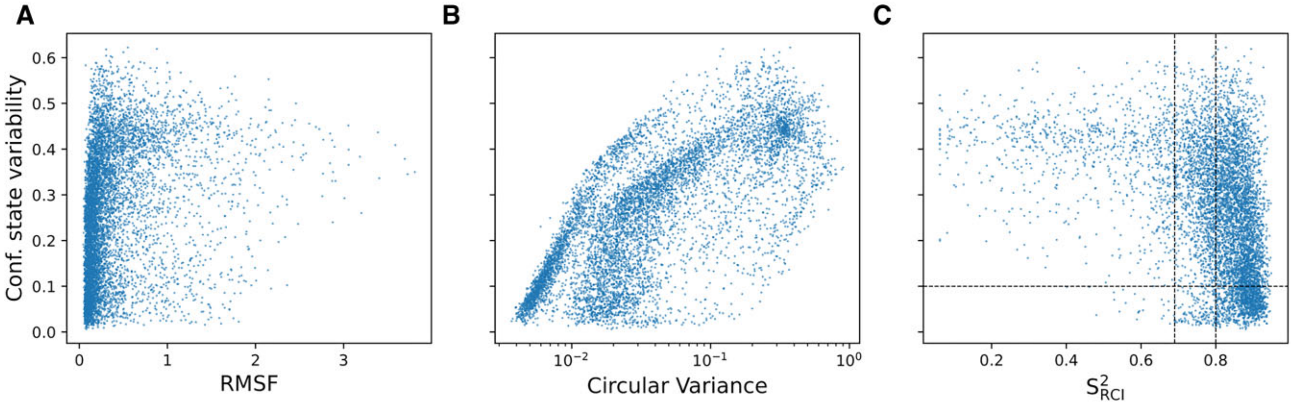


Figure 5. Conformational state variability (bootstrap sample size 3, 10 000 samples) versus traditional metrics. (**A**) Root mean square fluctuations (RMSF) per residue calculated for all residues of proteins in the MD data set, Pearson's $r = 0.27$ ($p = 0.00$). (**B**) Circular variance (CV) calculated for all residues of proteins in the MD data set, Pearson's $r = 0.56$ ($p < 0.001$). (**C**) S^2_{RCI} from residues in 62 proteins from the MD data set for which S^2_{RCI} values were available (Supplementary Data 2), Pearson's $r = -0.41$ ($p < 0.001$). The vertical dotted lines indicate the border between likely disordered, context-dependent and ordered residues (left to right) as defined in (17). The horizontal dotted line is a visual guide to distinguish between residues with very low and higher conformational state variability.

are averaged over all conformations adopted by the protein in solution, this is possible when the chemical environment of the atoms in the residue remains relatively constant (e.g. switches between core and *Surrounding helix*) or where a rigid stable conformation dominates, but is not exclusive (e.g. a residue that is *Core sheet* in 50% of the conformations but adopts other conformational states the rest of the time).

These findings indicate that conformational state variability offers complementary insights that are not directly inferable from existing dynamics metrics.

Conformational state variability for amino acids per order preference class

Amino acids have intrinsic properties that predispose them to facilitate the formation of ordered or disordered regions in proteins. Based on a previously used classification (17) the conformational state variability was examined within three classes: preferentially ordered (cysteine, phenylalanine, isoleucine, leucine, valine, tryptophan, and tyrosine), neutral (alanine, glutamic acid, lysine, methionine, glutamine, arginine, threonine, and histidine) and preferentially disordered (aspartic acid, glycine, asparagine, proline, and serine) residues. As expected, amino acids with preference for order exhibit lower conformational state variability than those with preference for disorder (Figure 6). Notable exceptions in the disordered category are glycine and proline. Both show low conformational state variabilities, which in this case reflects their tendency to exclusively adopt *Turn* or *Other* conformational states, without switching to other conformational states. Both these states represent mostly flat energy landscapes with low energy barriers, thus allowing for random movement or disorder (Figure 2). Note that aspartic acid, asparagine, and serine have higher conformational state variability, indicating they are more capable of switching states. This highlights an interesting feature of the conformational state variability, residues that remain in the same low energy conformational state, even if that conformational state is effectively dynamic, exhibit low conformational state variability. This is also reflected in *Other* conformational states in Supplementary Figure S13.

Conformational states capture regions of partial order in α- and β-synuclein

To assess the ability of our method to obtain relevant features from structural ensembles, we applied the method to two the Protein Ensemble Database (PED) ensembles of two well studied IDPs: α-synuclein (αSyn) and β-synuclein (βSyn). αSyn is an intrinsically disordered peptide whose aggregation plays a central role in the pathogenesis of Parkinson's disease, while its βSyn family member is less prone to aggregation.

Both proteins have been extensively studied by NMR, which indicated that the N-terminal part of both proteins (αSyn: aa 1-98, βSyn: aa 1–76) adopts random coil conformation. In contrast, the C-terminal region of both proteins exhibited extended conformations, with αSyn primarily being in β-sheet conformation (32) and βSyn adopting a polyproline II (ppII) state (33,34). Constava here illustrates its ability to extract these key features from a structural ensemble. While the N-terminal parts in both proteins primarily show *Other* with intermittent *Turn* corresponding to random coil state, the C-terminal propensities are shifted towards extended conformations, i.e. *Surrounding sheet* and *Core sheet* (Figure 7). Notably, the *Core sheet* region, which is exclusively populated in αSyn is strictly limited to β-sheet conformations, while *Surrounding sheet* which is prevalent in βSyn covers all extended states including the ppII region. Whilst these ensembles were generated using the NMR data and so should reflect experimental observations, Constava here shows the ability to directly pick up these relevant structural determinants with straightforward visualisation and interpretation possible.

Discussion

The advent of AlphaFold2 has revolutionised the field of structural bioinformatics by rapidly and accurately providing the structure of protein regions that adopt a stable well-defined fold. The focus is now shifting to predicting multiple conformational states for proteins, with as extreme case providing structure ensembles for intrinsically disordered regions or proteins. However, describing such conformational ensembles requires probabilistic metrics that capture (transient) features of such ensembles, as single models cannot represent the overall

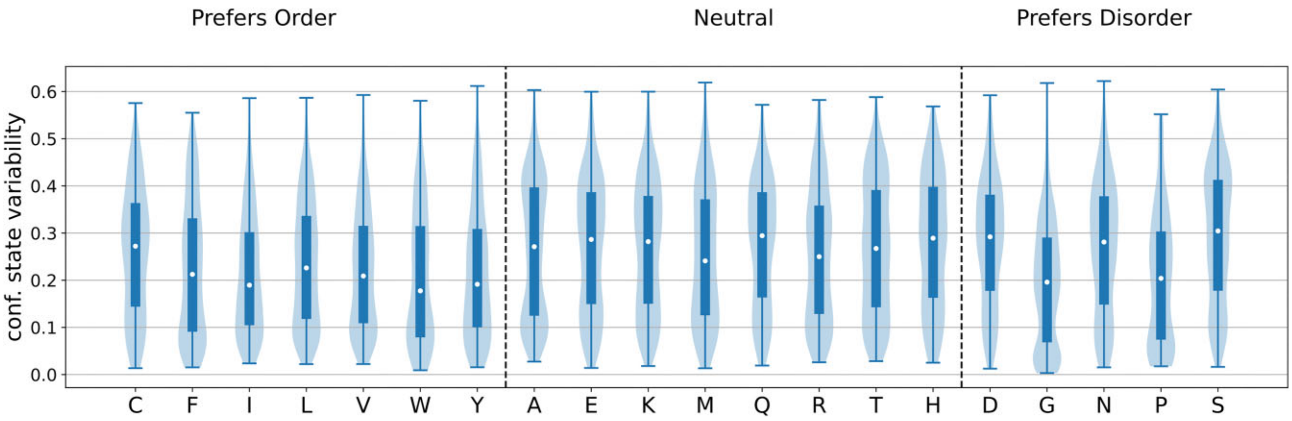


Figure 6. Conformational state variability values per amino acid type grouped according to their order preference. The 5 C- and N-terminal residues of each protein were excluded to remove end-of-chain bias towards disorder (e.g. methionine is often found as the first residue in a sequence). Amino acids which prefer order generally have the lowest conformational state variability. Glycine and proline have lower conformational state variability values as they tend to adopt respectively *Turn* and *Other* states. The raw data plot is available in Supplementary Figure S11. The conformational state variability (Variability) displayed in this figure was inferred with bootstrap sample size 3 (10 000 samples). Each violin represents the density of residues along the range conformational state variability values. Encased within each violin, a blue bar delineates the inter-quartile range (IQR), extending from the first quartile (Q1) to the third quartile (Q3), thus, encompassing the middle 50% of the data points, and it contains a white dot which marks the median. From the ends of this bar, thin lines stretch out to the extremes, capped at the minimum and maximum values observed in the data set.

behaviour of a protein, for example even if the protein does form transient helix turns these will traditionally only be identified in the models if stable hydrogen bonds are present. This work introduces a probabilistic data-driven definition of likely low-energy conformational states of amino acid residues in proteins based on the analysis of NMR ensembles and associated chemical shifts. We defined a total of six conformational states: *Core helix*, *Surrounding helix*, *Core sheet*, *Surrounding sheet*, *Other*, and *Turn*. We further introduced the concept of conformational state variability, a parameter that describes how often a residue switches between these conformational states. Both concepts are encompassed in the Constava software, which enables users to infer conformational states and their variability from an ensemble of structure models produced by molecular dynamics or other approaches. Constava describes the observed behaviour of each individual residue in such a conformational ensemble in terms of the relative propensities of the conformational states it adopts. Previous work demonstrated how such probabilistic descriptions can be used to efficiently sample realistic protein conformations (35).

When applied to molecular dynamics simulations, Constava is able to capture preferred and transient conformational states adopted by residues. Small sample sizes detect low populated transient states and provide a description of the conformational state dynamics of the residue. In contrast, larger sample sizes will detect only preferential conformational states and disregard transient states. The method can therefore detect different conformational features depending on the parameters used. When the subsampling directly uses a sliding window over the time series information in MD trajectories, it is even more sensitive towards transient conformational states that residues adopt. Since the sample size in that case directly relates to time intervals in the simulation, the method can be fine tuned to capture conformational states that persist specific time intervals. This distinguishes Constava from predictors of disordered regions or aggregation regions, such as IUPred (36), DisoMine (37) or TANGO (38). It provides values representing the likelihood for the defined conformational states, which also capture the protein's conformational dynamics, while through the conformational state variability it identifies the way residues change between conformational states. Moreover, as these values are derived from statistical potentials, the Constava conformational states in principle allow for a thermodynamical interpretation (Figure 2).

The conformational state variability definition is non-linear and is not necessarily directly related to absolute changes in backbone dihedral angle space. The *Other* state, for example, covers a wide range of dihedral space, so if a residue samples this range in dihedral space it will be exclusively assigned the *Other* state and have low conformational state variability - it adopts the same low-energy conformational state. Further motions within this same conformational state then also do not lead to any changes in its conformational state. Residues that exhibit high conformational state variability on the other hand do switch between different typical low-energy states. They therefore indicate regions that experience 'conformational state switches' in the protein. Such an interpretation differs significantly from metrics based purely on structural descriptors like root mean square fluctuations (RMSF) or circular variance (CV), which non-discriminatively interpret variations in respectively inter-atomic cartesian distances or dihedral angles. Consequently, it is not possible to infer the conformational state variability directly from these other metrics (Figure 5).

The conformational state variability does relate in different ways to the NMR chemical shift derived RCI-derived S^2 order parameter (S^2_{RCI}). Three categories of residues can be discerned: Highly ordered residues with low potential for conformational states changes, disordered residues which continuously change between conformational states, and residues capable of adopting stable conformational states, but with the capability of transitioning among different states. The entropy of the conformational states in the (ϕ, ψ)-space is a measure of how sensitive the conformational state variability is to changes in the backbone dihedrals (Supplementary Figure S14). Remarkably, the last category is comprised of residues with relatively high S^2_{RCI}, indicating little backbone dynam-

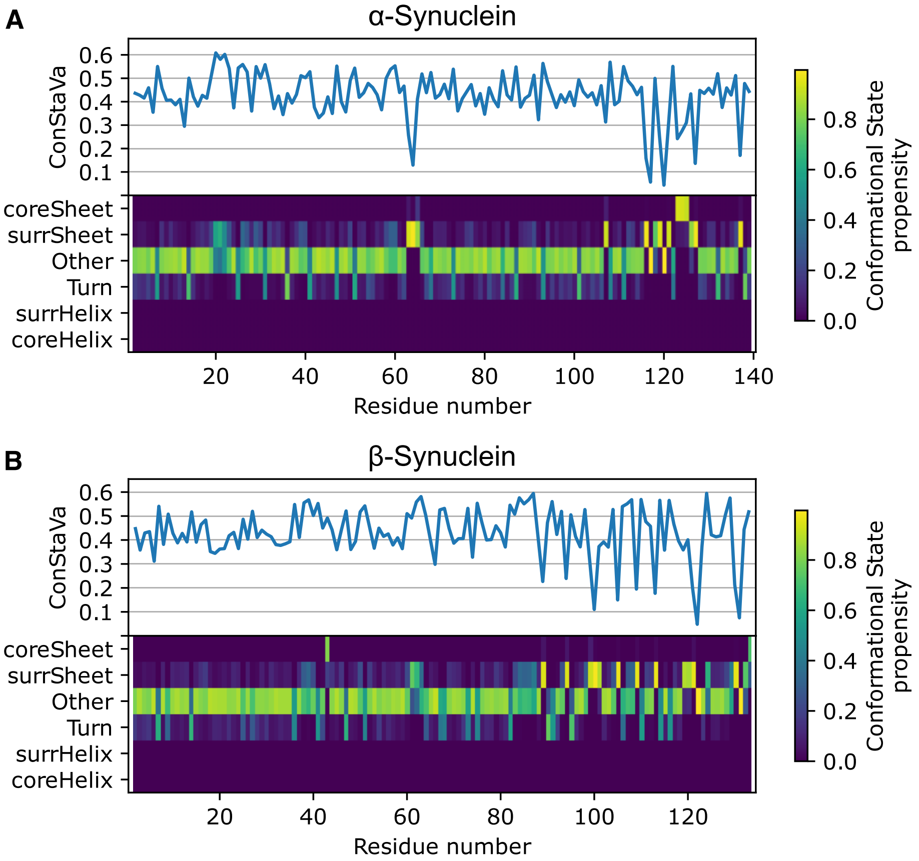


Figure 7. Results of Constava for α- and β-synuclein. For each protein the per-residue conformational state variability is shown (top) as well as the conformational state propensities for all the six conformational states (bottom) as calculated from the PED ensemble. (**A**) In α-synuclein the N-terminal region (aa 1-98) is mostly in *Other* with intermittent *Turn* and localised *Surrounding sheet* conformational states around residues 20 and 63. In the C-terminal part of the protein *Surrounding sheet* and *Core sheet* become more prominent, indicating an increased preference for extended structures, with reduced conformational state variability. (**B**) In β-synuclein the N-terminal region (aa 1–76) is mostly in *Other* with again intermittent *Turn* conformational states and localised *Surrounding sheet* and one outlier *Core sheet* residue. The C-terminal part of the protein shows *Surrounding sheet* but no *Core sheet*, suggesting a prevalence of ppII-like conformations rather than actual β-sheets.

ics, that still transition between states. This showcases how Constava can detect conformational state changes that are not picked up by NMR chemical shift values.

The conformational state variability also provides an interesting perspective on amino acid behaviour. Although the general categories of order-neutral-disorder inducing amino acids are as expected represented by low to high conformational variability, glycine and proline frequently feature lower values than the other amino acids in the disordered category. This indicates their preference to adopt exclusively the *Turn* or *Other* conformations as their low-energy states, with wide and flat energy landscapes for their dihedral space that allow them to continuously adopt a wide range of backbone conformations (Figure 2). Only rarely do they adopt multiple conformational states, in contrast to serine for example, which easily switches between them and can therefore accommodate multiple conformational states. Constava can, in this light, also identify regions with higher degrees of order in structure ensembles of intrinsically disordered proteins, represented by core and surrounding states. In the PED ensembles of α-synuclein (αSyn) and β-synuclein (βSyn), Constava directly identifies the β-sheet prone regions in αSyn as well as the ppII prone regions in the C-terminus of βSyn.

These examples showcase how the probabilistic definitions of conformational states allows for the reliable extraction of high-level information about protein behaviour in the study of (highly) dynamic proteins. Where large conformational ensembles are available, Constava does offer a novel way to identify the conformational states of residues, as well as how likely residues are to switch between conformational states. This is particularly useful in describing and classifying proteins with highly variable structures, such as IDPs (1) or fold-switching proteins (39), for which static definitions of conformation are not appropriate. We hope this approach will help in redefining our interpretation of proteins as dynamic, not static, entities with variable behaviour. The method is available as a PyPI package and all data used in this is freely available in the public domain.

Data availability

All data used in this work have been compiled in a single Zenodo repository (https://doi.org/10.5281/zenodo.10371447). The Constava software is available as a PyPI package (https://pypi.org/project/constava/). The source code is accessible on GitHub (https://github.com/Bio2Byte/constava) and may be used and redistributed under GPLv3 license; version 1.0.0 in preparation of this manuscript has been deposited in Zenodo (https://doi.org/10.5281/zenodo.10649793). An interactive Colab notebook is available with examples of the execution of the software: https://colab.research.google.com/github/Bio2Byte/public_notebooks/blob/main/constava_examples.ipynb.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank Adrian Diaz for the invaluable help in the distribution of this software.

Author contributions: W.V., D.R. and G.O. conceptualised the study. G.O. developed the initial methodology. W.V., D.R., G.O. and D.B. provided supervision. W.V. provided the NMR data and directed the project. J.G.-G. and D.B. developed, implemented and validated the method. J.R.-M. performed the MD simulations. All authors contributed to the writing of the manuscript.

Funding

European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie [813239 to J.R.-M., J.G.-G.]; Research Foundation Flanders (FWO) [G.032816N to G.O., G.028821N to D.B.]; Research Foundation Flanders (FWO) International Research Infrastructure [I000323N to W.V.]; COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology); the resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government.

Conflict of interest statement

None declared.

References

1. Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, **37**, 509–516.
2. Guo,J. and Zhou,H.-X. (2016) Protein Allostery and Conformational Dynamics. *Chem. Rev.*, **116**, 6503–6515.
3. Pirchi,M., Ziv,G., Riven,I., Cohen,S.S., Zohar,N., Barak,Y. and Haran,G. (2011) Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.*, **2**, 493.
4. Sun,Z., Liu,Q., Qu,G., Feng,Y. and Reetz,M.T. (2019) Utility of B-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem. Rev.*, **119**, 1626–1665.
5. Fenwick,R.B., van den Bedem,H., Fraser,J.S. and Wright,P.E. (2014) Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E445–E454.
6. Cheng,Y. (2018) Single-particle cryo-EM—How did it get here and where will it go. *Science*, **361**, 876–880.
7. Sekhar,A. and Kay,L.E. (2019) An NMR View of Protein Dynamics in Health and Disease. *Annu. Rev. Biophys.*, **48**, 297–319.
8. Beveridge,R. and Calabrese,A.N. (2021) Structural Proteomics Methods to Interrogate the Conformations and Dynamics of Intrinsically Disordered Proteins. *Front. Chem.*, **9**, 603639.
9. Agam,G., Gebhardt,C., Popara,M., Mächtel,R., Folz,J., Ambrose,B., Chamachi,N., Chung,S.Y., Craggs,T.D., De Boer,M., *et al.* (2023) Reliability and accuracy of single-molecule FRET studies for characterization of structural dynamics and distances in proteins. *Nat. Methods*, **20**, 523–535.
10. Hoch,J.C., Baskaran,K., Burr,H., Chin,J., Eghbalnia,H., Fujiwara,T., Gryk,M., Iwata,T., Kojima,C., Kurisu,G., *et al.* (2022) Biological Magnetic Resonance Data Bank. *Nucleic Acids Res.*, **51**, D368–D376.
11. Berjanskii,M.V. and Wishart,D.S. (2005) A simple method to predict protein flexibility using secondary chemical shifts. *J. Am. Chem. Soc.*, **127**, 14970–14971.
12. Berjanskii,M.V. and Wishart,D.S. (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. *J. Am. Chem. Soc.*, **135**, 14536–14539.
13. Camilloni,C., De Simone,A., Vranken,W.F. and Vendruscolo,M. (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, **51**, 2224–2231.
14. Orlando,G., Raimondi,D. F. and Vranken,W. (2019) Auto-encoding NMR chemical shifts from their native vector space to a residue-level biophysical index. *Nat. Commun.*, **10**, 2511
15. Vranken,W.F. (2014) NMR structure validation in relation to dynamics and structure determination. *Prog. Nucl. Mag. Reson. Spect.*, **82**, 27–38.
16. Walsh,I., Martin,A. J.M., Di Domenico,T. and Tosatto,S. C.E. (2012) ESpritz: accurate and fast prediction of protein disorder.. *Bioinformatics*, **28**, 503–509.
17. Cilia,E., Pancsa,R., Tompa,P., Lenaerts,T. and Vranken,W.F. (2013) From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun.*, **4**, 2741.
18. Rieping,W. and Vranken,W.F. (2010) Validation of archived chemical shifts through atomic coordinates. *Proteins*, **78**, 2482–2489.
19. Lovell,S.C., Davis,I.W., Arendall,W.B. III, de Bakker,P.I.W., Word,J.M., Prisant,M.G., Richardson,J.S. and Richardson,D.C. (2003) Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins*, **50**, 437–450.
20. Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.
21. Piovesan,D., Necci,M., Escobedo,N., Monzon,A.M., Hatos,A., Mičetić,I., Quaglia,F., Paladin,L., Ramasamy,P., Dosztányi,Z., *et al.* (2020) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.
22. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.* (2021) CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
23. Van Der Spoel,D., Lindahl,E., Hess,B., Groenhof,G., Mark,A.E. and Berendsen,H.J.C. (2005) GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, **26**, 1701–1718.
24. Huang,J., Rauscher,S., Nawrocki,G., Ran,T., Feig,M., de Groot,B.L., Grubmüller,H. and MacKerell,A.D. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*, **14**, 71–73.
25. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
26. Bussi,G., Donadio,D. and Parrinello,M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.*, **126**, 014101.
27. Parrinello,M. and Rahman,A. (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.*, **52**, 7182–7190.
28. Monzon,A.M., Rohr,C.O., Fornasari,M.S. and Parisi,G. (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. *Database*, **2016**, baw038.
29. Touw,W.G., Baakman,C., Black,J., te Beek,T.A., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
30. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

31. Lazar,T., Martínez-Pérez,E., Quaglia,F., Hatos,A., Chemes,L., Iserte,J.A., Méndez,N.A., Garrone,N.A., Saldaño,T., Marchetti,J., *et al.* (2020) PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, **49**, D404–D411.
32. Allison,J.R., Rivers,R.C., Christodoulou,J.C., Vendruscolo,M. and Dobson,C.M. (2014) A relationship between the transient structure in the monomeric state and the aggregation propensities of α-synuclein and β-synuclein. *Biochemistry*, **53**, 7170–7183.
33. Allison,J.R., Varnai,P., Dobson,C.M. and Vendruscolo,M. (2009) Determination of the free energy landscape of α-synuclein using spin label nuclear magnetic resonance measurements. *J. Am. Chem. Soc.*, **131**, 18314–18326.
34. Bertoncini,C.W., Rasia,R.M., Lamberto,G.R., Binolfi,A., Zweckstetter,M., Griesinger,C. and Fernandez,C.O. (2007) Structural characterization of the intrinsically unfolded protein beta-synuclein, a natural negative regulator of alpha-synuclein aggregation. *J. Mol. Biol.*, **372**, 708–722.
35. Hamelryck,T., Kent,J.T. and Krogh,A. (2006) Sampling realistic protein conformations using local structural bias. *PLoS Comput. Biol.*, **2**, e131
36. Erdős,G., Pajkos,M. and Dosztányi,Z. (2021) IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.*, **49**, W297–W303.
37. Orlando,G., Raimondi,D., Codicè,F., Tabaro,F. and Vranken,W. (2022) Prediction of disordered regions in proteins with recurrent neural networks and protein dynamics. *J. Mol. Biol.*, **434**, 167579.
38. Fernandez-Escamilla,A.-M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
39. Porter,L.L. and Looger,L.L. (2018) Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 5968–5973.

Received: February 13, 2024. **Revised:** June 14, 2024. **Editorial Decision:** June 24, 2024. **Accepted:** June 26, 2024