Structural Biology and Molecular Biophysics

Predicting the Sequence-Dependent Backbone Dynamics of Intrinsically Disordered Proteins

Sanbo Qin, Huan-Xiang Zhou

Department of Chemistry • Department of Physics, University of Illinois Chicago, Chicago, IL 60607, USA

Reviewed Preprint
v2 • September 17, 2024
Revised by authors

Reviewed Preprint
v1 • July 18, 2023

Abstract

How the sequences of intrinsically disordered proteins (IDPs) code for functions is still an enigma. Dynamics, in particular residue-specific dynamics, holds crucial clues. Enormous efforts have been spent to characterize residue-specific dynamics of IDPs, mainly through NMR spin relaxation experiments. Here we present a sequence-based method, SeqDYN, for predicting residue-specific backbone dynamics of IDPs. SeqDYN employs a mathematical model with 21 parameters: one is a correlation length and 20 are the contributions of the amino acids to slow dynamics. Training on a set of 45 IDPs reveals aromatic, Arg, and long-branched aliphatic amino acids as the most active in slow dynamics whereas Gly and short polar amino acids as the least active. SeqDYN predictions not only provide an accurate and insightful characterization of sequence-dependent IDP dynamics but may also serve as indicators in a host of biophysical processes, including the propensities of IDP sequences to undergo phase separation.

eLife assessment

In this **useful** study, a **solid** machine learning approach based on a broad set of systems to predict the R2 relaxation rates of residues in intrinsically disordered proteins (IDPs) is described. The ability to predict the patterns of R2 will be helpful to guide experimental studies of IDPs. A potential weakness is that the predicted R2 values may include both fast and slow motions, thus the predictions provide only limited new physical insights into the nature of the underlying protein dynamics, such as the most relevant timescale.

https://doi.org/10.7554/eLife.88958.2.sa3

Introduction

Intrinsically disordered proteins (IDPs) or regions (IDRs) do not have the luxury of a three-dimensional structure to help decipher the relationship between sequence and function. Instead, dynamics has emerged as a crucial link between sequence and function for IDPs (1). Nuclear magnetic resonance (NMR) spin relaxation is a uniquely powerful technique for characterizing IDP dynamics, capable of yielding residue-specific information (2). Backbone ^{15}N relaxation

experiments typically yield three parameters per residue: transverse relaxation rate (R_2), longitudinal relaxation rate (R_1), and steady-state heteronuclear Overhauser enhancement (NOE). While all three parameters depend on ps-ns dynamics, $R_2$is the one most affected by slower dynamics (10s of ns to 1 μs). An increase in either the timescale or the amplitude of slower dynamics results in higher R_2. For IDPs, $R_2$is also the parameter that exhibits the strongest dependence on sequence (1, 2).

$R_2$was noted early on as an important indicator of residual structure in the unfolded state of the structured protein lysozyme (3). This property has since been measured for many IDPs to provide insight into various biophysical processes. Just as the residual structure in the unfolded state biases the folding pathway of lysozyme (3), a nascent α-helix in the free state of Sendai virus nucleoprotein C-terminal domain (Sev-NT), as indicated by highly elevated R_2 (4), biases the coupled binding and folding pathway in the presence of its target phosphoprotein (5). Local secondary structure preformation also facilitates the binding of yes-associated protein (YAP) with its target transcription factor (6). Likewise a correlation has been found between R_2 in the free state and the membrane binding propensity of synaptobrevin-2: residues with elevated R_2have increased propensity for membrane binding (7). R_2 in the free state has also been used to uncover factors that promote liquid-liquid phase separation of IDPs. For example, a nascent α-helix (shown by elevated R_2) is important for the phase separation of the TDP-43 low-complexity domain, as both the deletion of the helical region and a helix-breaking mutation (Ala to Pro) abrogates phase separation (8). Similarly, nascent α-helices in the free state of cytosolic abundant heat-soluble 8 (CAHS-8), upon raising concentration and lowering temperature stabilize to form the core of fibrous gels (9). For the hnRNPA1 low-complexity domain (A1-LCD), aromatic residues giving rise to local peaks in R_2 also mediate phase separation (10).

Both NMR relaxation data and molecular dynamics (MD) simulations have revealed determinants of $R_2$for IDPs. It has been noted that the flexible Gly tends to lower R_2whereas secondary structure and contact formation tend to raise R_2 (11). This conclusion agrees well with recent MD simulations (1, 12–14). These MD studies, using IDP-specific force fields, are able to predict R_2 in quantitative agreement with NMR measurements, without ad hoc reweighting as done in earlier studies. According to MD, most contact clusters are formed by local sequences, within blocks of up to a dozen or so residues (1, 12, 13). Tertiary contacts can also form but are relatively rare; as such their accurate capture requires extremely extensive sampling and still poses a challenge for MD simulations. Contrary to Gly, aromatic residues have been noted as mediators of contact clusters (3, 10).

Schwalbe et al. (15) introduced a mathematical model to describe the R_2 profile along the sequence for lysozyme in the unfolded state. The R_2value of a given residue was expressed as the sum of contributions from this residue and its neighbors. This model yields a mostly flat profile across the sequence, except for falloff at the termini, resulting in an overall bell shape. Klein-Seetharaman et al. (3) then fit peaks above this flat profile as a sum of Gaussians. Cho et al. (16) proposed bulkiness as a qualitative indicator of backbone dynamics. Recently Sekiyama et al. (17) calculated R_2as the geometric mean of “indices of local dynamics”; the latter were parameterized by fitting to the measured R_2 for a single IDP. All these models merely describe the R_2profile of a given IDP, and none of them is predictive.

Here we present a method, SeqDYN, for predicting $R_2$of IDPs. Using a mathematical model introduced by Li et al. (18) to predict propensities for binding nanoparticles and also adapted for predicting propensities for binding membranes (19), we express the R_2 value of a residue as the product of contributing factors from all residues. The contributing factor attenuates as the neighboring residue becomes more distant from the central residue. The model, after training on a set of 45 IDPs, has prediction accuracy that is competitive against that of the recent MD simulations using IDP-specific force fields (1, 12–14). For lysozyme, the SeqDYN prediction agrees remarkably well with R_2measured in the unfolded state.

Results

The data set of IDPs with R_2 rates

We collected R_2 data for a total of 54 nonhomologous IDPs or IDRs (Table S1; Fig. S1). According to indicators from NMR properties, including low or negative NOEs, narrow dispersion in backbone amide proton chemical shifts, and small secondary chemical shifts (SCSs), most of the proteins are disordered with at most transient α-helices. A few are partially folded, including Sev-NT with a well-populated (~80%) long helix (residues 478-491) (20), CREB-binding protein fourth intrinsically disordered linker (CBP-ID4) with > 50% propensities for two long helices (residues 2-25 and 101-128) (21), HOX transcription factor DFD (HOX-DFD) with a well-folded domain comprising three helices (22), and Hahellin (apo form) as a molten globule (23). In **Fig. 1**, we display representative conformations of five IDPs, ranging from fully disordered MAPK kinase 4 (MKK4) (24) and α-synuclein (25) to Measles virus phosphoprotein N-terminal domain (Mev-P_{NTD}) (26) with transient short helices to Sev-NT and CBP-ID4 with stable long helices. The sequences of all the IDPs are listed in Table S2.

We used 45 of the 54 IDPs to train and validate SeqDYN and reserved the remaining 9 for testing. The sequence lengths of the training set range from 39 to 406 residues, with an average of 125.3 residues. Altogether R_2 data are available for 3966 residues. A large majority (35 out of 45) of the 45 IDPs have mean R_2 values ($\bar{R}_2$, calculated among all the residues in a protein) between 2.5 and 5.5 s^{-1} (Table S1 and **Fig. 2A**). This $\bar{R}_2$ range is much lower than that of structured proteins with similar sequence lengths. The low $\bar{R}_2$ values and lack of dependence on sequence length (Fig. S2A) suggest that R_2 of the IDPs is mostly dictated by local sequence instead of tertiary interaction.

The most often used temperature for acquiring the R_2 data was 298 K, but low temperatures (277 to 280 K) were used in a few cases (Table S1 and Fig. S2B). Of the seven IDPs with $\bar{R}_2$ > 6.4 s^{-1}, four can be attributed to low temperatures (27–30), one is due to a relatively low temperature (283 K) as well as the presence of glycerol (20% v/v) (31), and two can be explained by tertiary structure formation [a folded domain (22) or molten globule (23)]. A simple reason for higher R_2 values at lower temperatures is the higher water viscosity, resulting in a slowdown in molecular tumbling; a similar effect is achieved by adding glycerol. In some cases, R_2 was measured at both low and room temperatures (4, 10). To a good approximation, the effect of lowering temperature is a uniform scaling of R_2across the IDP sequence. For Sev-NT, downscaling of the R_2 values at 278 K by a factor of 2.0 brings them into close agreement with those at 298 K (Fig. S2C), with a root-mean-square-deviation (RMSD) of 0.5 s^{-1} among all the residues. Likewise, for A1-LCD, downscaling by a factor of 2.4 brings the R_2 values at 288 K into good match with those at 298 K (Fig. S2D), with an RMSD of 0.4 s^{-1}. Because SeqDYN is concerned with the sequence dependence of R_2, a uniform scaling has no effect on model parameter or prediction; therefore mixing the data from different temperatures is justified. The same can be said about the different magnetic fields in acquiring the R_2data (Table S1 and Fig. S2E). Increasing the magnetic field raises R_2 values, and the effect is also approximated well by a uniform scaling (4, 8, 29).

One measure on the level of sequence dependence of R_2 is the standard deviation, σ_{R_2}, calculated among the residues of an IDP. Among the training set, the R_2values of 30 IDPs have moderate sequence variations, with σ_{R_2} ranging from 0.5 to 1.5 s^{-1} (Table S1); the histogram of σ_{R_2} calculated for the entire training set peaks around 0.75 s^{-1} (**Fig. 2A**). There is a moderate correlation between σ_{R_2} and $\bar{R}_2$ (**Fig. 2A**, inset), reflecting in part the fact that σ_{R_2} can be raised simply by a uniform upscaling, e.g., as a result of lowering temperature. Still, only two of the five IDPs with high $\bar{R}_2$ attributable to lower temperature or presence of glycerol are among the seven IDPs with

Figure 1.

Representative conformations of five IDPs. (A-E) MKK4, α-synuclein, Mev-P_{NTD}, Sev-NT, and CBP-ID4. Conformations were initially generated using TraDES (*http://trades.blueprint.org*) (49), selected to have radius of gyration close to predicted by a scaling function $R_g = 2.54N^{0.522}$ (Å) (50). Conformations for residues predicted as helical by PsiPred plus filtering were replaced by an ideal helix. Finally residues are colored according to a scheme ranging from green for low predicted R_2 to red for high predicted R_2.

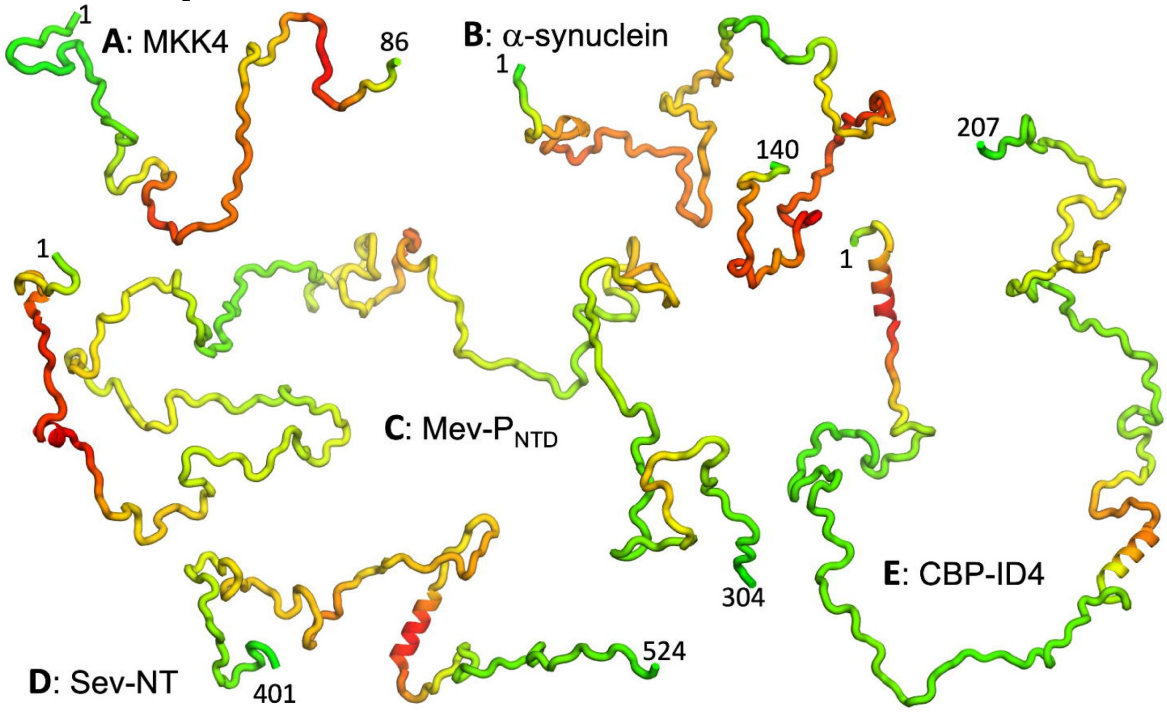
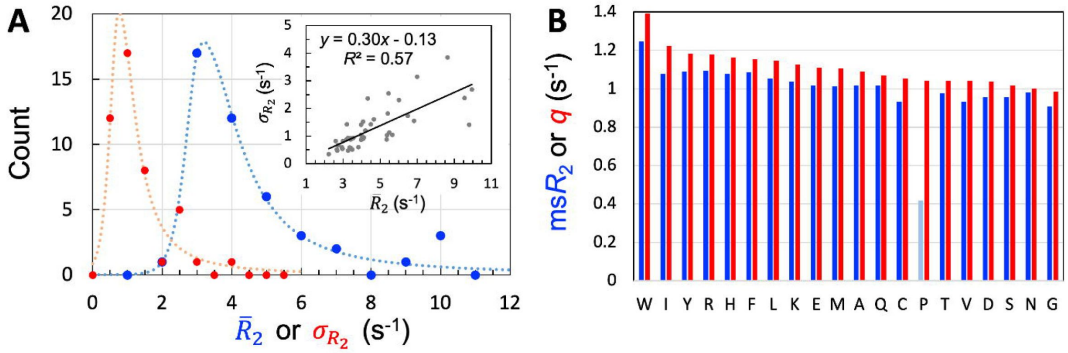


Figure 2.

Properties of the 45 IDPs in the training set. (A) Histograms of means and standard deviations, calculated for individual proteins. Curves are drawn to guide the eye. Inset: correlation between $\bar{R}_2$ and σ_{R_2}. (B) Experimental mean scaled R_2 (msR_2) and SeqDYN q parameters, for the 20 types of amino acids. Note that Pro residues have low msR_2 for the lack of backbone amide proton. Amino acids are in descending order of q.



high sequence variations (σ_{R_2} > 2 s^{-1}). Therefore the sequence variation of R_2as captured by σ_{R_2} manifests mostly the intrinsic effect of the IDP sequence, not the influence of external factors such as temperature or magnetic field strength. The mean σ_{R_2} value among the training set is 1.24 s^{-1}.

One way to eliminate the influence of external factors is to scale the R_2 values of each IDP by its $\bar{R}_2$; we refer to the results as scaled R_2, or sR_2. We then pooled the sR_2 values for all residues in the training set, and separated them according to amino-acid types. The amino acid type-specific mean sR_2 values, or msR_2, are displayed in **Fig. 2B**. The seven amino acids with the highest msR_2 in descending order are Trp, Arg, Tyr, Phe, Ile, His, and Leu. The presence of all the four aromatic amino acids in this "high-end" group immediately suggests π-π stacking as important for raising R_2; the presence of Arg further implicates cation-π interactions. In the other extreme, the seven amino acids with the lowest msR_2 in ascending order are Gly, Cys, Val, Asp, Ser, Thr, and Asn. Gly is well-known as a flexible residue; it is also interesting that all the four amino acids with short polar sidechains are found in this "low-end" group. Pro has an excessively low msR_2[with data from only two IDPs (32, 33)], but that is due to the absence of an amide proton.

The SeqDYN model and parameters

The null model is to assume a uniform R_2 for all the residues in an IDP. The root-mean-square-error (RMSE) of the null model is equal to the standard deviation, σ_{R_2}, of the measured R_2 values. The mean RMSE, $\overline{\text{RMSE}}$, of the null model, equal to 1.24 s^{-1} for the training set, serves as the upper bound for evaluating the errors of R_2predictors. The next improvement is a one-residue predictor, where first each residue (with index n) assumes its amino acid-specific mean sR_2 (msR_2) and then a uniform scaling factor Y is applied:

$$R_2(n) = \Upsilon \cdot \text{ms}R_2(n) \tag{1}$$

This one-residue model does only minutely better than the null model, with a $\overline{\text{RMSE}}$ of 1.22 s¯1.

In SeqDYN, we account for the influence of neighboring residues. Specifically, each residue i contributes a factor $f(i;n)$ to the R_2 value of residue n. Therefore

$$R_2(n) = \Upsilon \prod_{i=1}^{N} f(i;n) \tag{2a}$$

where N is the total number of residues in the IDP. The contributing factor depends on the sequence distance $s = |i - n|$ and the amino-acid type of residue i:

$$f(i;n) = 1 + \frac{q(i) - 1}{1 + bs^2} \tag{2b}$$

There are 21 global parameters. The first 20 are the q values, one for each of the 20 types of amino acids; the last parameter is b, appearing in the Lorentzian form of the sequence-distance dependence. We define the correlation length, L_{corr}, as the sequence distance at which the contributing factor is midway between the values at s = 0 and ∞. It is easy to verify that $L_{corr} = b^{-1/2}$. Note that the single-residue model can be seen as a special case of SeqDYN, with L_{corr} set to 0 and q set to msR_2.

The functional forms of Eqs (2a,b) were adapted from Li et al. (18); we also used them for predicting residue-specific membrane association propensities of IDPs (19). In these previous applications, a linear term was also present in the denominator of Eq (2b). In our initial training of SeqDYN, the coefficient of the linear term always converged to near zero. We thus eliminated the linear term. In addition to the Lorentzian form, we also tested a Gaussian form for the sequence-distance dependence and found somewhat worse performance. The more gradual attenuation of the Lorentzian form with increasing sequence distance evidently provides an overall better model

for the R_2data in the entire training set. Others (16, 17, 24) have modeled R_2as the average of some parameters over a window; a window has an extremely abrupt sequence-distance dependence (1 for $s < L_{corr}$ and 0 for $s > L_{corr}$).

We parametrized the SeqDYN model represented by Eqs (2a,b) on the training set of 45 IDPs. In addition to the 21 global parameters noted above, there are also 45 local parameters, namely one uniform scaling factor (Y) per IDP. The parameter values were selected to minimize the sum of the mean-square-errors for the IDPs in the training set, calculated on R_2data for a total of 3924 residues. We excluded the 42 Pro residues in the training set because, as already noted, their R_2values are lower for chemical reasons. We will present validation and test results below, but first let us look at the parameter values.

The q values are displayed in **Fig. 2B** alongside msR_2. In descending order, the seven amino acids with the highest q values are Trp, Ile, Tyr, Arg, His, Phe, and Leu. These are exactly the same amino acids in the high-end group for msR_2, though their order there is somewhat different. In ascending order, the seven amino acids (excluding Pro) with the lowest q values are Gly, Asn, Ser, Asp, Val, Thr, and Cys. The composition of the low-end group is also identical to that for msR_2. The q values thus also suggest that ρ-ρ and cation-ρ interactions in local sequences may raise R_2, whereas Gly and short-polar residues may lower R_2.

Given the common amino acids at both the high and low ends for msR_2 and q, it is not surprising that these two properties exhibit a strong correlation, with a coefficient of determination (R^2; excluding Pro) at 0.92 (**Fig. 3A**). Also, because the high-end group contains the largest amino acids (e.g., Trp and Tyr) whereas the low-end group contains the smallest amino acids (e.g., Gly and Ser), we anticipated some correlation of ms$R_2$and q with amino-acid size. We measure the latter property by the molecular mass (m). As shown in **Fig. 3B**, both msR_2 and q indeed show a medium correlation with m, with R^2 = 0.67 (excluding Pro) and 0.61, respectively. A bulkiness parameter was proposed as an indicator of sequence-dependent backbone dynamics of IDPs (16, 24). Bulkiness was defined as the sidechain volume-to-length ratio, and identified amino acids with aromatic or branched aliphatic sidechains as bulky (34). We found only modest correlations between either msR_2 or q and bulkiness, with R^2 just below 0.4 (**Fig. 3C**).

The optimized value of b is 3.164 ξ 10^{-2}, corresponding to an L_{corr} of 5.6 residues. The resulting optimized $\overline{\text{RMSE}}$ is 0.95 s^{-1}, a clear improvement over the value 1.24 s^{-1} of the null model. To check the sensitivity of prediction accuracy to b, we set b to values corresponding to L_{corr}= 0, 1, 2, .., and retrained SeqDYN for b fixed at each value. Note that the null-model $\overline{\text{RMSE}}$, 1.24 s^{-1}, sets an upper bound. This upper bound is gradually reached when L_{corr} is increased from the optimal value. In the opposite direction, when $L_{corr}$is decreased from the optimal value, $\overline{\text{RMSE}}$ rises quickly, reaching 1.22 s^{-1} at L_{corr} = 0. The latter $\overline{\text{RMSE}}$ is the same as that of the single-residue model. Lastly we note that there is a strong correlation between the uniform scaling factors and $\bar{R}_2$ values among the 45 IDPs (R^2 = 0.77), as to be expected. For 39 of the 45 IDPs, Y values fall in the range of 0.8 to 2.0 s^{-1}.

As presented next, we evaluate the performance of SeqDYN by leave-one-out cross validation, where each IDP in turn was left out of the training set and the model was trained on the remaining 44 IDPs to predict R_2 for the IDP that was left out. The parameters from the leave-one-out (also known as jackknife) training sessions allow us to assess the potential bias of the training set. For this purpose, we compare the values of the 21 global parameters, either from the full training set or from taking the averages of the jackknife training sessions. For each of the q parameters, the values from these two methods differ only in the fourth digit; e.g., for Leu, they are both 1.1447 from full training and from jackknife training. The values for b are 3.164 ξ 10^{-2} from full training as stated above and 3.163 ξ 10^{-2} from jackknife training. The close agreement in parameter values between full training and jackknife training suggests no significant bias in the training set.

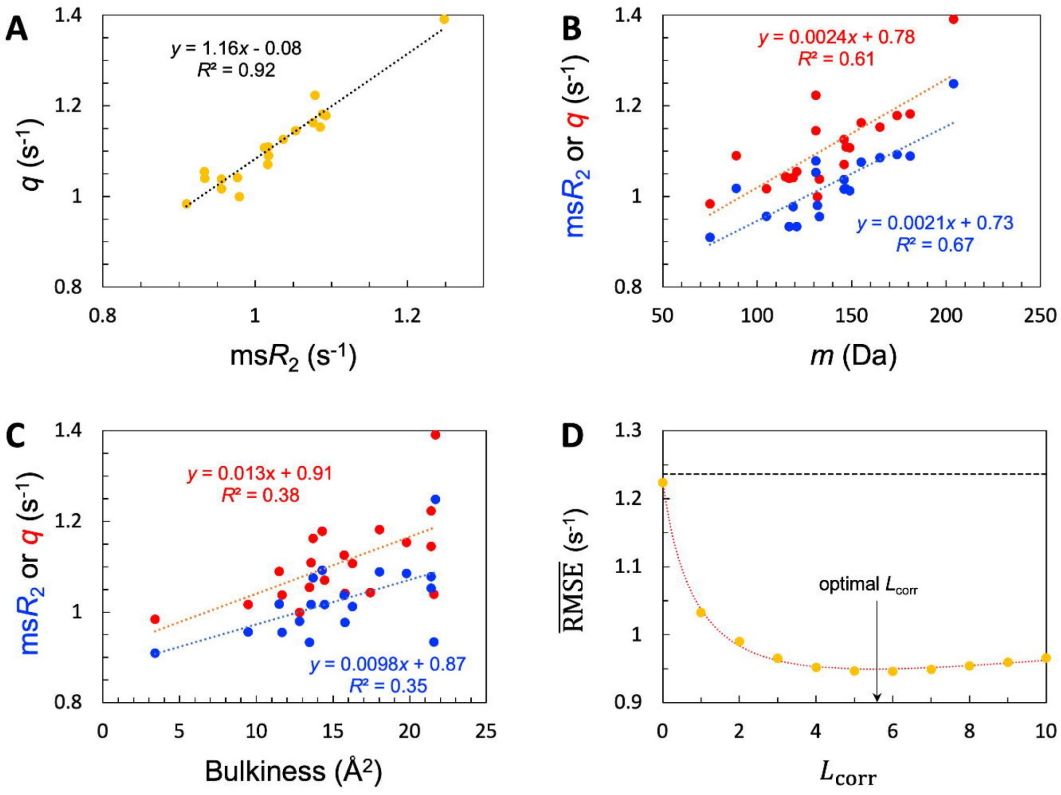


Figure 3.

SeqDYN model parameters. (A) Correlation between msR_2 and q. The values are also displayed as bars in **Fig. 2B**. (B) Correlation of msR_2 and q with amino-acid molecular mass. (C) Correlation of msR_2 and q with bulkiness. (D) The optimal correlation length and deterioration of SeqDYN prediction as the correlation length is moved away from the optimal value.

Another question of interest is whether the difference between the q parameters of two amino acids is statistically significant. To answer this question, we carried out five-fold cross-validation training, resulting in five independent estimates for each parameter. For example, the mean ± standard deviation of the q parameter is 1.1405 ± 0.0066 for Leu and 1.2174 ± 0.0211 for Ile. A t-test shows that their difference is extremely statistically significant ($P < 0.0001$). In contrast, the difference between Leu and Phe (q = 1.1552 ± 0.0304) is not significant. t-test results for other pairs of amino acids are found in Fig. S3.

Validation of SeqDYN predictions

We now present leave-one-out cross-validation results. We denote the RMSE of the R_2 prediction for the left-out IDP as RMSE(−1). As expected, RMSE(−1) is higher than the RMSE obtained with the IDP kept in the training set, but the increases are generally slight. Specifically, all but eight of the IDPs have increases < 0.1 s^{-1}; the largest increase is 0.35 s^{-1}, for CBP-ID4. The mean RMSE(−1), or $\overline{\mathrm{RMSE}}$ (−1) for the 45 IDPs is increased by 0.05 s^{-1}over $\overline{\mathrm{RMSE}}$, to 1.00 s^{-1}. The latter value is still a distinct improvement over the mean RMSE 1.24 s^{-1} of the null model. The histogram of RMSE(−1) for the 45 IDPs is shown in **Fig. 4A**. It peaks at 0.5 s^{-1}, which is a substantial downshift from the corresponding peak at 0.75 s^{-1} for σ_{R_2} (**Fig. 2A**). Thirty-four of the 45 IDPs have RMSE(−1) values lower than the corresponding σ_{R_2}.

To further illustrate the performance of SeqDYN, we present comparison of predicted and measured R_2 values for five IDPs: MKK4, α-synuclein, Mev-P_{NTD}, Sev-NT, and CBP-ID4 (**Fig. 4B-F**). A simple common feature is the falloff of R_2at the N-and C-termini, resulting from missing upstream or downstream residues that otherwise would be coupled to the terminal residues, as first recognized by Schwalbe et al. (15). Representative conformations of the five IDPs are displayed in **Fig. 1**, with residues colored according to the predicted R_2values. For four of these IDPs, the RMSE(−1) values range from 0.44 to 0.76 s^{-1} and are scattered around the peak of the histogram, while the RMSE(−1) for the fifth IDP, namely CBP-ID4, the RMSE(−1) value is 2.01 s^{-1} and falls on the tail of the histogram (**Fig. 4A**). **Figure 4B** displays the measured and predicted R_2 for MKK4. SeqDYN correctly predicts higher R_2values in the second half of the sequence than in the first half. It even correctly predicts the peak around residue Arg75. The sequence in this region is H_{72}IERLRTH$_{79}$; six of these eight residues belong to the high-end group. In contrast, the lowest R_2 values occur in the sequence S_7GGGGSGGGSGSG$_{19}$, comprising entirely of two amino acids in the low-end group.

R_2values for α-synuclein are shown in **Fig. 4C**. Here SeqDYN correctly predicts higher R_2near the C-terminus and a dip around Gly68. However, it misses the R_2 peaks around Tyr39 and Asp121. MD simulations (1) have found that these R_2peaks can be explained by a combination of secondary structure formation (ý-sheet around Tyr39 and polyproline II helix around Asp121) and local (between Tyr39 and Ser42) or long-range (between Asp121 and Lys96) interactions. SeqDYN cannot account for long-range interactions (e.g., between ý-strands and between Asp121 and Lys96). **Figure 4D** shows that SeqDYN gives excellent R_2predictions for Mev-P_{NTD}. It correctly predicts the high peaks around Arg17, Glu31, Leu193, and lower peaks around Arg235 and Trp285, but does underpredict the narrow peak around Tyr113.

The overall R_2profile of Sev-NT is predicted well by SeqDYN, but the peak in the long helical region (residues 478-491) is severely underestimated (green curve in **Fig. 4E**). A similar situation occurs for CBP-ID4, where the peak in the second long helical region (around Glu113) is underpredicted (green curve in **Fig. 4F**). While the measured R_2 exhibits a higher peak in the second helical region than in the first helical region (around Arg16), the opposite is predicted by SeqDYN. When the R_2data were included in the training set (i.e., full training), the second peak is higher than the first one, but that is not a real prediction because the R_2 data themselves were used for training the model. It merely means that the SeqDYN functions can be parameterized to produce any prescribed R_2 profile along the sequence. Indeed, when the R_2data of CBP-ID4 alone were used to

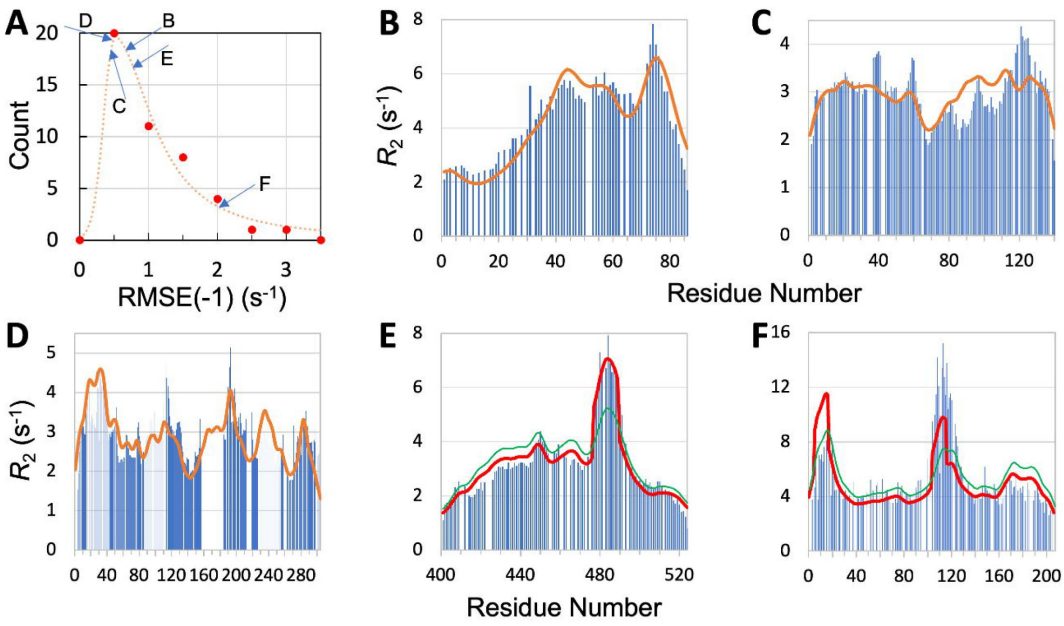


Figure 4.

Quality of SeqDYN predictions. (A) Histogram of RMSE(−1). Letters indicate RMSE(−1) values of the IDPs to be presented in panels (B-F). (B-F) Measured (bars) and predicted (curves) R_2 profiles for MKK4, α-synuclein, Mev-P_{NTD}, Sev-NT, and CBP-ID4. In (E) and (F), green curves are SeqDYN predictions and red curves are obtained after a helix boost.

parameterize SeqDYN, the measured R_2 profile is closely reproduced (Fig. S3). The reversal in R_2peak heights between the two helical regions is the reason for the aforementioned unusual increase in RMSE when CBP-ID4 was left out of the training set.

R2 boost in long helical regions

It is apparent that SeqDYN underestimates the R_2 of stable long helices. Transient short helices does not seem to be a problem, since these are present, e.g., in Mev-P_{NTD}, where transient helix formation in the first 37 residues and between residues 189-198 (26) coincides with R_2 peaks that are correctly predicted by SeqDYN. SeqDYN can treat coupling between residues within the correlation length of 5.6 residues, but a much longer helix would tumble more slowly than implied by an $L_{corr}$of 5.6, and thus it makes sense that SeqDYN would underestimate R_2 in that case.

Our solution then is to apply a boost factor to the long helical region. To do so, we have to know whether an IDP does form long helices and if so what the constituent residues are. Secondary structure predictors tend to overpredict α-helices and ý-strands for IDPs, as they are trained on structured proteins. One way to counter that tendency is to make the criteria for α-helices and ý-strands stricter. We found that, by filtering PsiPred (*http://bioinf.cs.ucl.ac.uk/psipred*) (35) helix propensity scores (p_{Hlx}) with a very high cutoff of 0.99, the surviving helix predictions usually correspond well with residues identified by NMR as having high helix propensities. For example, for Mev-P_{NTD}, PsiPred plus filtering predicts residues 14-17, 28-33, and 191-193 as helical; all of them are in regions that form transient helices according to chemical shifts (26). Likewise long helices are also correctly predicted for Sev-NT (residues 477-489) and CBP-ID4 (residues 6-17 and 105-116) (20, 21).

We apply a boost factor, B_{Hlx}, to helices with a threshold length of 12:

$$B_{\mathrm{Hlx}} = 1 + \alpha p_{\mathrm{Hlx}} \Theta(p_{\mathrm{Hlx}} \geq 0.99; L_{\mathrm{Hlx}} \geq 12) \quad (3)$$

The Θ function is 1 if the helix propensity score is above the filtering cutoff and the helix length (L_{Hlx}) is above the threshold, and 0 otherwise. With a boost amplitude α at 0.5, the boosted SeqDYN prediction for Sev-NT reaches excellent agreement with the measured R_2 (**Fig. 4E**, red curve). The RMSE(−1) is reduced from 0.76 s^{-1} to 0.38 s^{-1} upon boosting.

Applying the same helix boost to CBP-ID4 also results in a modest reduction in RMSE(−1), from 2.01 to 1.90 s^{-1} (**Fig. 4F**, red curve). The only other IDP for which PsiPred plus filtering predicts a long helix is the N-terminal region of lysyl-tRNA synthetase (KRS-NT). The authors who studied this protein did not report on secondary structure (36), but feeding their reported chemical shifts to the TALOS+ server (*https://spin.niddk.nih.gov/bax/nmrserver/talos/*) (37) found only short stretches of residues that fall into the helical region of the Ramachandran map. The SeqDYN prediction for KRS-NT is already good [RMSE(−1) = 0.83 s^{-1}]; applying a helix boost would deteriorate the RMSE(−1) to 1.16 s^{-1}.

Further test on a set of nine IDPs

We have reserved nine IDPs for testing SeqDYN (parameterized on the training set of 45 IDPs). The level of disorder in these test proteins also spans the full range, from absence of secondary structures [ChiZ N-terminal region (12), Pdx1 C-terminal region (11), and TIA-1 prion-like domain (17)] to presence of transient short helices [synaptobrevin-2 (7), α-endosulfine (38), YAP (6), angiomotin-like 1 (AMOTL1) (39)] to formation of stable long helices [FtsQ (13) and CAHS-8 (9)]. For eight of the nine test IDPs, the RMSEs of SeqDYN predictions are lower than the experimental σ_{R_2} values, by an average of 0.66 s^{-1}. For the ninth IDP (Pdx1), the SeqDYN RMSE is slightly higher, by 0.06 s^{-1}, than the experimental σ_{R_2}. Together, the nine test IDPs have a mean RMSE of 1.13 s^{-1}, close to the $\overline{\mathrm{RMSE}}$ (−1) of 1.00 s^{-1} for the training set in the leave-one-out cross-validation.

The comparison of predicted and measured R_2 profiles along the sequence is presented in **Fig. 5A-I**. For ChiZ, SeqDYN correctly predicts the major peak around Arg25 and the minor peak around Arg46 (**Fig. 5A**). The R_2 profile of Pdx1 is largely featureless, except for a dip around Gly216, which is correctly predicted by SeqDYN (**Fig. 5B**). Correct prediction is also obtained for the higher R_2 in the first half of TIA-1 prion-like domain than in the second half (**Fig. 5C**). SeqDYN gives an excellent prediction for synaptobrevin-2, including a linear increase up to Arg56 and the major peak around Trp89 (**Fig. 5D**).

The prediction is also very good for α-endosulfine, including elevated R_2 around Glu34, which coincides with the presence of a transient helix, and depressed R_2 in the last 40 residues (**Fig. 5E**). The only miss is an underprediction for the peak around Lys74. SeqDYN also predicts well the overall shape of the R_2 profile for YAP, including peaks around Asn70, Leu91, Arg124, and Arg161, but severely underestimates the peak height around Asn70 (**Fig. 5F**).NOE signals indicate contacts between Met86, Leu91, Fhe95, and Fhe96 (6); evidently this type of local contacts is captured well by SeqDYN. The R_2 elevation around Asn70 is mostly due to helix formation: residues 61-74 have helix propensities up to 40% (6). PsiPred predicts helix for residues 62-73, but only residues 65-68 survive the filtering that we impose, resulting in a helix that is too short to apply a helix boost. The prediction for AMOTL1 is mostly satisfactory, including peaks around Phe200 and Arg264 and a significant dip around Gly292 (**Fig. 5G**). However, whereas the two peaks have approximately equal heights in the measured R_2profile, the predicted peak height around Phe200 is too low. SCSs indicate helix propensity around both R_2 peaks (39). PsiPred also predicts helix in both regions, but only five and two residues, respectively, survive after filtering, and are too short for applying a helix boost.

For FtsQ, SeqDYN correctly predicts elevated R_2 for the long helix [residues 46-74 (13)] but underestimates the magnitude (RMSE = 2.32 s^{-1}; green curve in **Fig. 5H**). PsiPred plus filtering predicts a long helix formed by residues 47-73. Applying the helix boost substantially improves the agreement with the measured R_2, with RMSE reducing to 1.71 s^{-1} (red curve in **Fig. 5H**). SeqDYN also gives a qualitatively correct R_2 profile for CAHS-8, with higher $R_2$for the middle section (residues 95-190) (RMSE = 2.36 s^{-1}; green curve in **Fig. 5I**). However, it misses the extra elevation in $R_2$for the first half of the middle section (residues 95-145). According to SCS, the first and second halves have helix propensities of 60% and 30%, respectively (9). PsiPred plus filtering predicts helices for residues 96-121, 124-141, 169-173, and 179-189. Only the first two helices, both in the first half of the middle section, are considered long according to our threshold. Once again, applying the helix boost leads to marked improvement in the predicted in R_2, with RMSE reducing to 1.92 s^{-1} (red curve in **Fig. 5I**).

Inputting the sequences of structured proteins predicts R_2 in the unfolded state

SeqDYN is trained on IDPs, what if we feed it with the sequence of a structured protein? The prediction using the sequence of hen egg white lysozyme, a well-studied single-domain protein, is displayed in **Fig. 6A**. It shows remarkable agreement with the R_2 profile measured by Klein-Seetharaman et al. (3) in the unfolded state (denatured by 8 M urea at pH 2 and reduced to break disulfide bridges), including a major peak around Trp62, a second peak around Trp111, and a third peak around Trp123. Klein-Seetharaman et al. mutated Trp62 to Gly and the major peak all but disappeared. This result is also precisely predicted by SeqDYN with the mutant sequence (**Fig. 6B**).

SeqDYN also predicts well the R_2 profiles of other proteins in the unfolded state. For unfolded apomyoglobin (8 M urea; pH 2.3), Schwarzinger et al. (40) claimed that depressed R_2corresponded to stretches of small amino acids (Gly and Ala) whereas elevated corresponded to local hydrophobic interactions. SeqDYN reproduces all the observed peaks and valleys in the R_2profile (**Fig. 6C**). The deepest valley indeed occurs over a Gly/Ala-rich stretch,

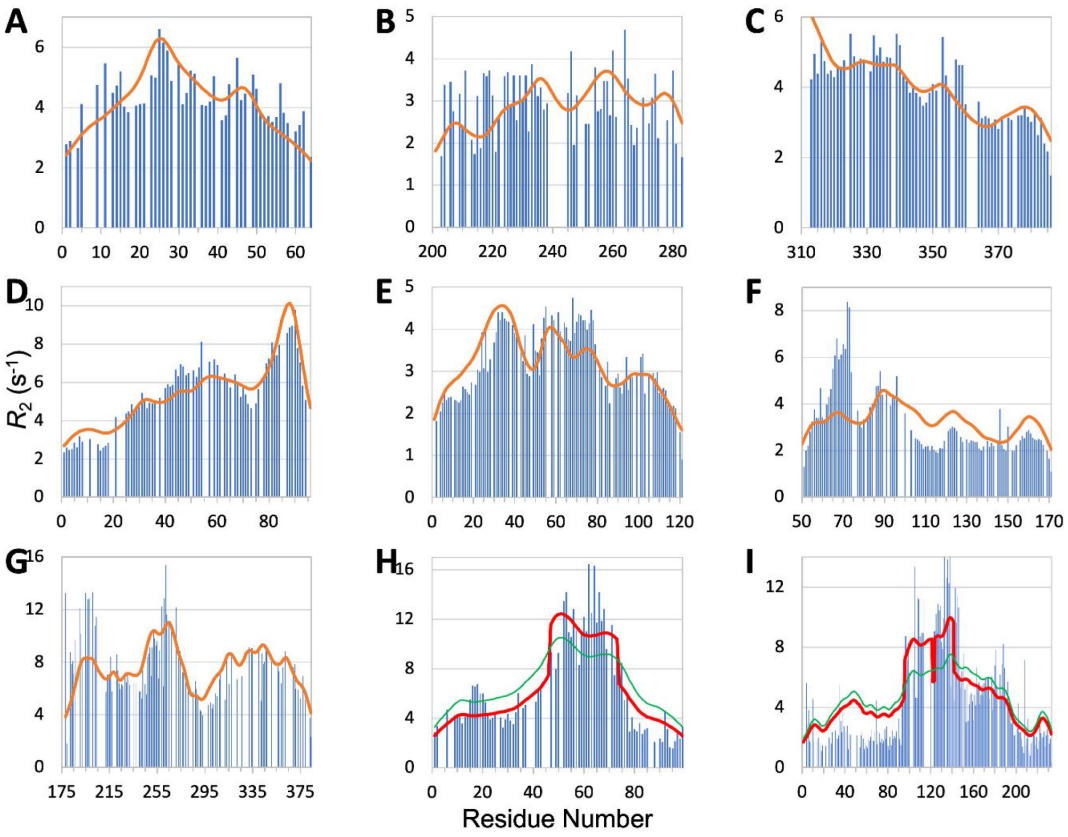


Figure 5.

Measured (bars) and predicted (curves) R_2profiles for ChiZ N-terminal region, TIA1 prion-like domain, Pdx1 C-terminal region, synaptobrevin-2, α-endosulfine, YAP, AMOTL1, FtsQ, and CAHS-8. In (C), R_2does not fall off at the N-terminus because the sequence is preceded by an expression tag MGSSHHHHHHHHHHHHS. In (H) and (I), green curves are SeqDYN predictions and red curves are obtained after a helix boost.

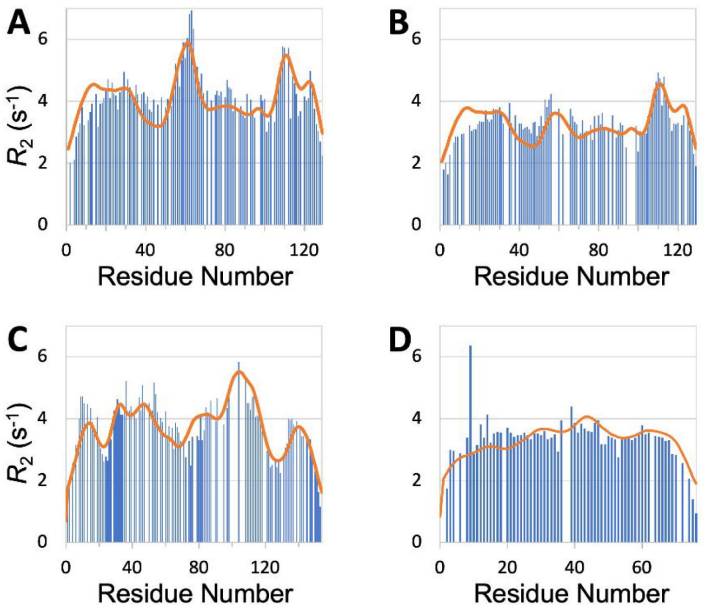


Figure 6.

R_2 profiles predicted (curves) by SeqDYN show close agreement with those measured (bars) on structured proteins in the unfolded state. (A) Wild-type lysozyme (8 M urea; pH 2; cysteine-methylated). (B) Lysozyme with Trp62 to Gly mutation (pH 2). Methylated cysteines were treated as Ala in the SeqDYN predictions. (C) Apomyoglogin (8 M urea; pH 2.3). (D) Ubiquitin (8 M urea; pH 2).

G_{125}ADAQGA$_{131}$, but the highest peak occurs over a stretch, I_{102}KYLEFI$_{108}$, that contains both hydrophobic and charged residues, all of which are on the high end of the *q* parameters (**Fig. 2B**). The R_2 profile of unfolded ubiquitin (8 M urea; pH 2) is relatively flat, which Wirmer et al. (41) attributed to lack of residual secondary structure, based on the assumption that ý-sheets (major elements of folded ubiquitin) are less resistant to denaturation than α-helices. SeqDYN predicts a relatively flat R_2profile (**Fig. 6D**), but the reason is that the ubiquitin sequence lacks a contiguous stretch of high-*q* amino acids.

Discussion

We have developed a powerful method, SeqDYN, that predicts the backbone amide transverse relaxation rates (R_2) of IDPs. The method is based on IDP sequences, is extremely fast, and available as a web server at *https://zhougroup-uic.github.io/SeqDYNidp/*. The excellent performance supports the notion that the ns-dynamics reported by $R_2$is coded by the local sequence, comprising up to 6 residues on either side of a given residue. The amino-acid types that contribute the most to coupling within a local sequence are aromatic (Trp, Tyr, Phe, and His), Arg, and long branched aliphatic (Ile and Leu), suggesting the importance of ρ-ρ, cation-ρ, and hydrophobic interactions in raising R_2. These interactions are interrupted by Gly and amino acids with short polar sidechains (Ser, Thr, Asn, and Asp), leading to reduced R_2. Transient short helices produce moderate elevation in R_2, whereas stable long helices result in a big boost in R_2. Tertiary contacts can also raise R_2, but appear to be infrequent in most IDPs (1).

It is also possible that R_2reported by backbone amide ^{15}N relaxation (as is the case for most of the IDPs studied here) may not be particularly sensitive to exchange effects, which likely involve tertiary contact formation. For the D2 domain of p27^{Kip1}, the exchange contributions measured using ^{15}N relaxation were small (< 2.5 s^{-1}) but were as large as 25 s^{-1} when measured by high-power ^{1}H relaxation dispersion (42). This experiment measures the effective transverse relation rate, $R_{2,eff}$, over a range of effective radiofrequency w_{eff}. The exchange contribution is maximal for the value $R_{2,eff}{}^{low\ weff}$ at low w_{eff} but is largely quenched for the value $R_{2,0}{}^{app}$ in the high-w_{eff}limit. The SeqDYN prediction for this IDP matches much better with $R_{2,0}{}^{app}$ than with $R_{2,eff}{}^{low\ weff}$ (Fig. S5). It is not clear whether this IDP is unique in forming persistent tertiary contacts that give rise to substantial exchange contributions or the ^{1}H relaxation dispersion experiment is unique in reporting the exchange contributions. At the minimum, SeqDYN yields the exchange-free portion of the transverse relaxation rate, enabling easy identification of residues that potentially participate in tertiary contacts. For the D2 domain of p27^{Kip1}, SeqDYN correctly predicts the $R_{2,0}{}^{app}$ local maxima at W76 and Y88. It is these same two residues that show substantial exchange contributions and putatively participate in tertiary contact (42). Therefore local contacts may seed tertiary contacts. If $R_{2,eff}$ data with substantial exchange contributions become available for more IDPs, SeqDYN may be retrained to make predictions for IDPs forming persistent tertiary contacts.

The *q* parameters, while introduced here to characterize the propensities of amino acids to participate in local interactions, appear to correlate with the tendencies of amino acids to drive liquid-liquid phase separation. Consistent with the rank order of *q*, Trp, Tyr, and Arg have been reported to be strong drivers of phase separation, Lys is a moderate driver, whereas Gly and Ser suppress phase separation (10, 33, 43). Recent measurements of the threshold concentration produced the following order for the propensity of phase separation by eight nonpolar amino acids in homotetrapeptides of the form XXssXX (ss: backbone disulfide bond): Trp > Phe > Leu > Met > Ile > Val > Ala > Pro (44). This order is the same as that of the *q* parameters, except that the *q* values of Ile and Val are in the second and last places, respectively. Threshold concentrations of IDPs are now predicted reasonably well by coarse-grained simulations where each amino acid is modeled by a single bead with a Lennard-Jones diameter d_0 and a stickiness parameter λ (45). Our *q* parameter shows good correlation (R^2 = 0.59) with the compound

parameter $d_0{}^3\lambda$ (Fig. S6). Therefore the q parameter may serve as a predictor for the tendency of an amino acid to drive phase separation. In essence, the same ability of an amino acid, e.g., Trp, to form interactions with neighboring residues of an IDP in the free state also applies when it comes to interactions with residues on neighboring chains in a dense phase.

Our method incorporates ideas from a number of previous efforts at describing R_2. The first serious effort was by Schwalbe et al. (15), who accounted for contributions from neighboring residues as additive terms, instead of multiplicative factors as in SeqDYN. Cho et al. (16) and Delaforge et al. (24) used the running average of the bulkiness parameter over a window of five to nine residues as a qualitative indicator of R_2. Here again the calculation was based on an additive model. Sekiyama et al. (17) employed a multiplicative model, with R_2 calculated as a geometric mean of "indices of local dynamics" over a five-residue window.

These indices, akin to our q parameters, were trained on a single IDR (TIA-1 prion-like domain) and used to reproduce the measured $R_2$for the same IDR. As we have illustrated on CBP-ID4 (Fig. S3), training on a single protein merely biases the parameters to that model and has little value in predicting $R_2$for other proteins. In comparison, SeqDYN is trained on 45 IDPs and its predictions are robust and achieve quantitative agreement with measured R_2.

Ten of the IDPs tested here have been studied recently by MD simulations using IDP-specific force fields (1, 12–14). In **Table 1** we compare the RMSEs of SeqDYN predictions with those for R_2 calculations from MD simulations. For five of these IDPs: A1-LCD, Aφ340, α-synuclein, tau K18, and FtsQ, RMSEs of SeqDYN and MD are remarkably similar. Four of these IDPs lack significant population of α-helices or φ3-sheets, but FtsQ forms a stable long helix. For one other IDP, namely HOX-DFD, MD, by explicitly modeling its folded domain, does a much better job in predicting R_2than SeqDYN (RMSEs of 1.40 s^{-1} vs 1.99 s^{-1}).

However, for the four remaining IDPs: p53TAD, Pup, Sev-NT, and ChiZ, SeqDYN significantly outperforms MD, with RMSEs averaging only 0.47 s^{-1}, compared to the MD counterpart of 1.14 s^{-1}. Overall, SeqDYN is very competitive against MD in predicting R_2, but without the significant computational cost. While MD simulations can reveal details of local interactions, as noted for α-synuclein, and capture tertiary interactions if they occur, they still suffer from perennial problems of force-field imperfection and inadequate sampling. SeqDYN provides an accurate description of IDP dynamics at a "mean-field" level, but could miss idiosyncratic behaviors of specific local sequences.

Deep-learning models have become very powerful, but they usually have millions of parameters and require millions of protein sequences for training (46). In contrast, SeqDYN employs a mathematical model with dozens of parameters and requires only dozens of proteins for training. Reduced models (by collapsing amino acids into a small number of distinct types) have even been trained on < 10 IDPs to predict propensities for binding nanoparticles (18) or membranes (19). The mathematical model-based approach may be useful in other applications where data, similar to R_2, are limited, including predictions of IDP secondary chemical shifts or residues that bind drug molecules (47) or protein targets, or even in protein design, e.g., for recognizing an antigenic site or a specific DNA site.

Acknowledgements

This work was supported by Grant GM118091 from the National Institutes of Health.

IDP name	SeqDYN	MD
A1-LCD	0.60[a]	0.59[d, e]
Aβ40	0.38[a]	0.38[d]
HOX-DFD	1.99[a]	1.40[d]
α-synuclein	0.44[a]	0.50[d]
p53TAD	0.33[a]	1.04[f]
Pup	0.43[a]	1.00[f]
Sev-NT	0.38[a, b]	1.10[d, g]
tau K18	0.83[a]	0.80[d]
ChiZ	0.74[c]	1.40[h]
FtsQ	1.71[c, b]	1.70[i]

[a]Based on leave-one-out training (using 44 IDPs).

[b]Helix boost applied.

[c]Based on training by the full training set (45 IDPs).

[d]From ref (1).

[e]RMSE is scaled down by a factor of 2.39, to correct for the effect of temperature (MD at 288 K; see Fig. S2C).

[f]From ref (14).

[g]RMSE is scaled down by a factor of 2.99, to correct for the effects of temperature and magnetic field (MD at 274 K and 850 MHz; see Fig. S2B).

[h]Originally calculated in ref (12) with correction in ref (48).

[i]From ref (13).

Table 1.

RMSEs (s^{-1}) of R_2 predictions by SeqDYN and MD for 10 IDPs

Dynamics is a crucial link between sequence and function for intrinsically disordered proteins (IDPs). Qin and Zhou developed a sequence-based method to predict IDP dynamics. The predictions provide deep physicochemical insight and can also serve as indicators of properties including phase-separation propensities of IDP sequences.

References

1. Dey S., MacAinsh M., Zhou H. X. (2022) **Sequence-Dependent Backbone Dynamics of Intrinsically Disordered Proteins** *J Chem Theory Comput* **18**:6310–6323

2. Camacho-Zarco A. R., et al. (2022) **NMR Provides Unique Insight into the Functional Dynamics and Interactions of Intrinsically Disordered Proteins** *Chem Rev* **122**:9331–9356

3. Klein-Seetharaman J., et al. (2002) **Long-Range Interactions Within a Nonnative Protein** *Science* **295**:1719–1722

4. Abyzov A., et al. (2016) **Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation** *J Am Chem Soc* **138**:6240–6251

5. Schneider R., et al. (2015) **Visualizing the Molecular Recognition Trajectory of an Intrinsically Disordered Protein Using Multinuclear Relaxation Dispersion NMR** *J Am Chem Soc* **137**:1220–1229

6. Feichtinger M., et al. (2022) **Long-range structural preformation in yes-associated protein precedes encounter complex formation with TEAD** *iScience* **25**

7. Lakomek N.-A., Yavuz H., Jahn R., Pérez-Lara Á. (2019) **Structural dynamics and transient lipid binding of synaptobrevin-2 tune SNARE assembly and membrane fusion** *Proc Natl Acad Sci U S A* **116**:8699–8708

8. Conicella A. E., Zerze G. H., Mittal J., Fawzi N. L. (2016) **ALS Mutations Disrupt Phase Separation Mediated by alpha-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain** *Structure* **24**:1537–1549

9. Malki A., et al. (2022) **Intrinsically Disordered Tardigrade Proteins Self-Assemble into Fibrous Gels in Response to Environmental Stress** *Angew Chem Int Ed Engl* **61**

10. Martin E. W., et al. (2020) **Valence and patterning of aromatic residues determine the phase behavior of prion-like domains** *Science* **367**:694–699

11. Cook E. C., Sahu D., Bastidas M., Showalter S. A. (2019) **Solution Ensemble of the C-Terminal Domain from the Transcription Factor Pdx1 Resembles an Excluded Volume Polymer** *J Phys Chem B* **123**:106–116

12. Hicks A., Escobar C. A., Cross T. A., Zhou H. X. (2020) **Sequence-dependent correlated segments in the intrinsically disordered region of ChiZ** *Biomolecules* **10**:1–23

13. Smrt S. T., Escobar C. A., Dey S., Cross T. A., Zhou H. X. (2023) **An Arg / Ala-Rich Helix in the N-Terminal Region of M . tuberculosis FtsQ Anchors FtsZ to Membranes** *Communications Biology* https://doi.org/10.1038/s42003-023-04686-5

14. Yu L., Bruschweiler R. (2022) **Quantitative prediction of ensemble dynamics, shapes and contact propensities of intrinsically disordered proteins** *PLoS Comput Biol* **18**

15. Schwalbe H., et al. (1997) **Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea** *Biochemistry* **36**:8977–8991

16. Cho M. K., et al. (2007) **Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau** *J Am Chem Soc* **129**:3032–3033

17. Sekiyama N., et al. (2022) **ALS mutations in the TIA-1 prion-like domain trigger highly condensed pathogenic structures** *Proc Natl Acad Sci U S A* **119**

18. Li D. W., Xie M., Bruschweiler R. (2020) **Quantitative cooperative binding model for intrinsically disordered proteins interacting with nanomaterials** *J Am Chem Soc* **142**:10730–10738

19. Qin S., Hicks A., Dey S., Prasad R., Zhou H. X. (2022) **ReSMAP: Web Server for Predicting Residue-Specific Membrane-Association Propensities of Intrinsically Disordered Proteins** *Membranes* **12**

20. Jensen M. R., et al. (2008) **Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein** *J Am Chem Soc* **130**:8055–8061

21. Piai A., et al. (2016) **Just a Flexible Linker? The Structural and Dynamic Properties of CBP-ID4 Revealed by NMR Spectroscopy** *Biophys J* **110**:372–381

22. Maiti S., et al. (2019) **Dynamic Studies on Intrinsically Disordered Regions of Two Paralogous Transcription Factors Reveal Rigid Segments with Important Biological Functions** *J Mol Biol* **431**:1353–1369

23. Patel S., Ramanujam V., Srivastava A. K., Chary K. V. (2014) **Conformational propensities and dynamics of a betagamma-crystallin, an intrinsically disordered protein** *Phys Chem Chem Phys* **16**:12703–12718

24. Delaforge E., et al. (2018) **Deciphering the Dynamic Interaction Profile of an Intrinsically Disordered Protein by NMR Exchange Spectroscopy** *J Am Chem Soc* **140**:1148–1158

25. Sung Y. H., Eliezer D. (2007) **Residual structure, backbone dynamics, and interactions within the synuclein family** *J Mol Biol* **372**:689–707

26. Milles S., et al. (2018) **An ultraweak interaction in the intrinsically disordered replication machinery is essential for measles virus function** *Sci Adv* **4**

27. Sólyom Z., et al. (2015) **The Disordered Region of the HCV Protein NS5A: Conformational Dynamics, SH3 Binding, and Phosphorylation** *Biophys J* **109**:1483–1496

28. Martin E. W., et al. (2016) **Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation** *J Am Chem Soc* **138**:15323–15335

29. Janke A. M., et al. (2018) **Lysines in the RNA Polymerase II C-Terminal Domain Contribute to TAF15 Fibril Recruitment** *Biochemistry* **57**:2549–2563

30. Jenner M., et al. (2018) **Mechanism of intersubunit ketosynthase–dehydratase interaction in polyketide synthases** *Nat Chem Biol* **14**:270–275

31. Bafaro E. M., Maciejewski M. W., Hoch J. C., Dempski R. E. (2019) **Concomitant disorder and high-affinity zinc binding in the human zinc- and iron-regulated transport protein 4 intracellular loop** *Protein Sci* **28**:868–880

32. Murrali M. G., Piai A., Bermel W., Felli I. C., Pierattelli R. (2018) **Proline Fingerprint in Intrinsically Disordered Proteins** *ChemBioChem* **19**:1625–1629

33. Wong L. E., Kim T. H., Muhandiram D. R., Forman-Kay J. D., Kay L. E. (2020) **NMR Experiments for Studies of Dilute and Condensed Protein Phases: Application to the Phase-Separating Protein CAPRIN1** *J Am Chem Soc* **142**:2471–2489

34. Zimmerman J. M., Eliezer N., Simha R. (1968) **The characterization of amino acid sequences in proteins by statistical methods** *J Theor Biol* **21**:170–201

35. McGuffin L. J., Bryson K., Jones D. T. (2000) **The PSIPRED protein structure prediction server** *Bioinformatics* **16**:404–405

36. Cho H. Y., et al. (2014) **Characterization of the interaction between lysyl-tRNA synthetase and laminin receptor by NMR** *FEBS Lett* **588**:2851–2858

37. Shen Y., Delaglio F., Cornilescu G., Bax A. (2009) **TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts** *J Biomol NMR* **44**:213–223

38. Thapa C., Roivas P., Haataja T., Permi P., Pentikainen U. (2022) **Interaction mechanism of endogenous PP2A inhibitor protein ENSA with PP2A** *FEBS J* **289**:519–534

39. Vogel A., Crawford A., Nyarko A. (2022) **Multivalent Angiomotin-like 1 and Yes-associated protein form a dynamic complex** *Protein Sci* **31**

40. Schwarzinger S., Wright P. E., Dyson H. J. (2002) **Molecular hinges in protein folding: the urea-denatured state of apomyoglobin** *Biochemistry* **41**:12681–12686

41. Wirmer J., Peti W., Schwalbe H. (2006) **Motional properties of unfolded ubiquitin: a model for a random coil protein** *J Biomol NMR* **35**:175–186

42. Ban D., Iconaru L. I., Ramanathan A., Zuo J., Kriwacki R. W. (2017) **A Small Molecule Causes a Population Shift in the Conformational Landscape of an Intrinsically Disordered Protein** *J Am Chem Soc* **139**:13692–13700

43. Wang J., et al. (2018) **A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins** *Cell* **174**:688–699

44. Zhang Y., Prasad R., Su S., Lee D., Zhou H. X. (2024) **Amino Acid-Dependent Material Properties of Tetrapeptide Condensates** *bioRxiv* https://doi.org/10.1101/2024.05.14.594233

45. Tesei G., Lindorff-Larsen K. (2022) **Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range** *Open Res Eur* **2**

46. Rives A., et al. (2021) **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences** *Proc Natl Acad Sci U S A* **118**

47. Robustelli P., et al. (2022) **Molecular Basis of Small-Molecule Binding to alpha-Synuclein** *J Am Chem Soc* **144**:2501–2510

48. Hicks A., MacAinsh M., Zhou H. X. (2021) **Removing thermostat distortions of protein dynamics in constant-temperature molecular dynamics simulations** *J Chem Theory Comput* **17**:5920–5932

49. Feldman H. J., Hogue C. W. (2002) **Probabilistic sampling of protein conformations: new hope for brute force?** *Proteins* **46**:8–23

50. Bernado P., Blackledge M. (2009) **A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering** *Biophys J* **97**:2839–2845

Editors

Reviewing Editor
Qiang Cui
Boston University, Boston, United States of America

Senior Editor
Qiang Cui
Boston University, Boston, United States of America

Reviewer #2 (Public review):

Qin, Sanbo and Zhou, Huan-Xiang created a model, SeqDYN, to predict nuclear magnetic resonance (NMR) spin relaxation spectra of intrinsically disordered proteins (IDPs), based primarily on amino acid sequence. To fit NMR data, SeqDYN uses 21 parameters, 20 that correspond to each amino acid, and a sequence correlation length for interactions. The model demonstrates that local sequence features impact the dynamics of the IDP, as SeqDYN performs better than a one residue predictor, despite having similar numbers of parameters. SeqDYN is trained using 45 IDP sequences and is retrained using both leave-one-out cross validation and five-fold cross validation, ensuring the model's robustness. While SeqDYN can provide reasonably accurate predictions in many cases, the authors note that improvements can be made by incorporating secondary structure predictions, especially for alpha-helices that exceed the correlation length of the model. The authors apply SeqDYN to study nine IDPs and a denatured ordered protein, demonstrating its predictive power. The model can be easily accessed via the website mentioned in the text.

The authors have adequately addressed the majority of my previous concerns. However, I still wonder if an attempt to fit the individual protein fitting parameter based on temperature and magnetic field strength would be possible. The authors would have 45 data points on which to fit such a parameter, which would only depend on two variables.

https://doi.org/10.7554/eLife.88958.2.sa2

Reviewer #3 (Public review):

The revised manuscript adds some new relevant analyses. It still, however, is unclear which timescales of motions the method refers to and there is confusion about whether the model can predict "slower motions". While the authors answer some of my points, others are left unanswered. That is of course the authors' prerogative, and readers will in any case be able to read the reviewer comments. I am not sure it is productive to add further comments at this point.

Below are my comments from the first round of review:

The manuscript by Qin and Zhou presents an approach to predict dynamical properties of an intrinsically disordered protein (IDP) from sequence alone. In particular, the authors train a simple (but useful) machine learning model to predict (rescaled) NMR R2 values from sequence. Although these R2 rates only probe some aspects of IDR dynamics and the method does not provide insight into the molecular aspects of processes that lead to perturbed dynamics, the method can be useful to guide experiments.

A strength of the work is that the authors train their model on an observable that directly relates to protein dynamics. They also analyse a relatively broad set of proteins which means that one can see actual variation in accuracy across the proteins.

A weakness of the work is that it is not always clear what the measured R2 rates mean. In some cases, these may include both fast and slow motions (intrinsic R2 rates and exchange contributions). This in turn means that it is actually not clear what the authors are predicting. The work would also be strengthened by making the code available (in addition to the webservice), and by making it easier to compare the accuracy on the training and testing data.

https://doi.org/10.7554/eLife.88958.2.sa1

Author response:

The following is the authors' response to the original reviews.

> *In this useful study, a solid machine learning approach based on a broad set of systems to predict the R2 relaxation rates of residues in intrinsically disordered proteins (IDPs) is described. The ability to predict the patterns of R2 will be helpful to guide experimental studies of IDPs. A potential weakness is that the predicted R2 values may include both fast and slow motions, thus the predictions provide only limited new physical insights into the nature of the relevant protein dynamics.*

Fast motions are less sequence-dependent (e.g., as shown by R1). Hence the sequence-dependent part of R2 singles out slow motion.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Solution state 15N backbone NMR relaxation from proteins reports on the reorientational properties of the N-H bonds distributed throughout the peptide chain. This information is crucial to understanding the motions of intrinsically disordered proteins and as such has focussed the attention of many researchers over the last 20-30 years, both experimentally, analytically and using numerical simulation.*
>
> *This manuscript proposes an empirical approach to the prediction of transverse 15N relaxation rates, using a simple formula that is parameterised against a set of 45 proteins. Relaxation rates measured under a wide range of experimental conditions are combined to optimize residuespecific parameters such that they reproduce the overall shape of the relaxation profile. The purely empirical study essentially ignores NMR relaxation theory, which is unfortunate, because it is likely that more insight could have been derived if theoretical aspects had been considered at any level of detail.*

NMR relaxation theory is very valuable in particular regarding motions on different timescales. However, it has very little to say about the sequence dependence of slow motions, which is the focus of our work.

> *Despite some novel aspects, in particular the diversity of the relaxation data sets, the residuespecific parameters do not provide much new insight beyond earlier work that has also noted that sidechain bulkiness correlated with the profile of R2 in disordered proteins.*

The novel insight from our work is that R2 can mostly be predicted based on the local sequence.

> *Nevertheless, the manuscript provides an interesting statistical analysis of a diverse set of deposited transverse relaxation rates that could be useful to the community.*

Thank you!

> *Crucially, and somewhat in contradiction to the authors stated aims in the introduction, I do not feel that the article delivers real insight into the nature of IDP dynamics. Related to this, I have difficulty understanding how an approximate prediction of the overall trend of expected transverse relaxation rates will be of further use to scientists working on IDPs. We already know where the secondary structural elements are (from 13C chemical shifts which are essential for backbone assignment) and the necessary 'scaling' of the profile to match experimental data actually contains a lot of the information that researchers seek.*

Again, the novel insight is that slow motions that dictate the sequence dependence of R2 can mostly be predicted based on the local sequence. The scaling factor may contain useful information but does not tell us anything about the sequence dependence of IDP dynamics.

This reviewer brings up a lot of valuable points, clearly from an NMR spectroscopist's perspective. The emphasis of our paper is somewhat different from that perspective. For example, we were interested in whether tertiary contacts make significant contributions to R2, as sometimes claimed. Our results show that, in general, they do not; instead local contacts dominate the sequence dependence of R2.

> *(1) The introduction is confusing, mixing different contributions to R2 as if they emanated from the same physics, which is not necessarily true. 15N transverse relaxation is said to report on 'slower' dynamics from 10s of nanoseconds up to 1 microsecond. Semi-classical Redfield theory shows that transverse relaxation is sensitive to both adiabatic and non-adiabatic terms, due to spin state transitions induced by stochastic motions, and dephasing of coherence due to local field changes, again induced by stochastic motions. These are faster than the relaxation limit dictated by the angular correlation function. Beyond this, exchange effects can also contribute to measured R2. The extent and timescale limit of this contribution depends on the particular pulse sequence used to measure the relaxation. The differences in the pulse sequences used could be presented, and the implications of these differences for the accuracy of the predictive algorithm discussed.*

Indeed pulse sequences affect the measured R2 values. We make the modest assumption that such experimental idiosyncrasy would not corrupt the sequence dependence of IDP dynamics. As for exchange effects, our expectation is that the current SeqDYN may not do well for R2s where slow exchange plays a dominant role in generating sequence dependence,

as tertiary contacts would be prominent in those cases; we now present one such case (new Fig. S5).

> *(2) Previous authors have noted the correlation between observed transverse relaxation rates and amino acid sidechain bulkiness. Apart from repeating this observation and optimizing an apparently bulkiness-related parameter on the basis of R2 profiles, I am not clear what more we learn, or what can be derived from such an analysis. If one can possibly identify a motif of secondary structure because raised R2 values in a helix, for example, are missed from the prediction, surely the authors would know about the helix anyway, because they will have assigned the 13C backbone resonances, from which helical propensity can be readily calculated.*

We think that a sequence-based method that is demonstrated to predict well R2 values from expensive NMR experiments is significant. That pi-pi and cation-pi interactions are prominent features of local contacts and may seed tertiary contacts and mediate inter-chain contacts that drive phase separation is a valuable insight.

> *(3) Transverse relaxation rates in IDPs are often measured to a precision of 0.1s-1 or less. This level of precision is achieved because the line-shapes of the resonances are very narrow and high resolution and sensitivity are commonly measurable. The predictions of relaxation rates, even when applying uniform scaling to optimize best-agreement, is often different to experimental measurement by 10 or 20 times the measured accuracy. There are no experimental errors in the figures. These are essential and should be shown for ease of comparison between experiment and prediction.*

Again, our focus is not the precision of the absolute R2 values, but rather the sequence dependence of R2.

> *(4) The impact of structured elements on the dynamic properties of IDPs tethered to them is very well studied in the literature. Slower motions are also increased when, for example the unfolded domain binds a partner, because of the increased slow correlation time. The ad hoc 'helical boosting' proposed by the authors seems to have the opposite effect. When the helical rates are higher, the other rates are significantly reduced. I guess that this is simply a scaling problem. This highlights the limitation of scaling the rates in the secondary structural element by the same value as the rest of the protein, because the timescales of the motion are very different in these regions. In fact the scaling applied by the authors contains very important information. It is also not correct to compare the RMSD of the proposed method with MD, when MD has not applied a 'scaling'. This scaling contains all the information about relative importance of different components to the motion and their timescales, and here it is simply applied and not further analysed.*

Actually, applying the boost factor achieves the effect of a different scaling factor for the secondary structure element than for the rest of the protein.

Regarding comparing RMSEs of SeqDYN and MD, it is true that SeqDYN applies a scaling factor whereas MD does not. However, even if we apply scaling to MD results it will not change the basic conclusion that “SeqDYN is very competitive against MD in predicting _R_2, but without the significant computational cost.”

> *(5) Generally, the uniform scaling of all values by the same number is serious oversimplification. Motions are happening on all timescales they are giving rise to different transverse relaxation. It is not possible to describe IDP relaxation in terms of one single motion. Detailed studies over more than 30 years, have demonstrated that*

> *more than one component to the autocorrelation function is essential in order to account for motions on different timescales in denatured, partially disordered or intrinsically unfolded states. If one could 'scale' everything by the same number, this would imply that only one timescale of motion were important and that all others could be neglected, and this at every site in the protein. This is not expected to be the case, and in fact in the examples shown by the authors it is also never the case. There are always regions where the predicted rates are very different from experiment (with respect to experimental error), presumably because local dynamics are occurring on different timescales to the majority of the molecule. These observations contain useful information, and the observation that a single scaling works quite well probably tells us that one component of the motion is dominant, but not universally. This could be discussed.*

The reviewer appears to equate a single scaling factor with a single type of motion -- this is not correct. A single scaling factor just means that we factor out effects (e.g., temperature or magnetic field) that are uniform across the IDP sequence.

> *(6) With respect to the accuracy of the prediction, discussion about molecular detail such as pi-pi interactions and phase separation propensity is possibly a little speculative.*

It is speculative; we now add more support to this speculation (p. 18 and new Fig. S6).

> *(7) The authors often declare that the prediction reproduces the experimental data. The comparisons with experimental data need to be presented in terms of the chi2 per residue, using the experimentally measured precision which as mentioned, is often very high.*

Again, our interest is the sequence dependence of R2, not the absolute R2 value and its measurement precision.

> ***Reviewer #2 (Public Review):***
>
> *Qin, Sanbo and Zhou, Huan-Xiang created a model, SeqDYN, to predict nuclear magnetic resonance (NMR) spin relaxation spectra of intrinsically disordered proteins (IDPs), based primarily on amino acid sequence. To fit NMR data, SeqDYN uses 21 parameters, 20 that correspond to each amino acid, and a sequence correlation length for interactions. The model demonstrates that local sequence features impact the dynamics of the IDP, as SeqDYN performs better than a one residue predictor, despite having similar numbers of parameters. SeqDYN is trained using 45 IDP sequences and is retrained using both leave-one-out cross validation and five-fold cross validation, ensuring the model's robustness. While SeqDYN can provide reasonably accurate predictions in many cases, the authors note that improvements can be made by incorporating secondary structure predictions, especially for alpha-helices that exceed the correlation length of the model. The authors apply SeqDYN to study nine IDPs and a denatured ordered protein, demonstrating its predictive power. The model can be easily accessed via the website mentioned in the text.*
>
> *While the conclusions of the paper are primarily supported by the data, there are some points that could be extended or clarified.*
>
> *(1) The authors state that the model includes 21 parameters. However, they exclude a free parameter that acts as a scaling factor and is necessary to fit the experimental data (lambda). As a result, SeqDYN does not predict the spectrum from the sequence de-novo, but requires a one parameter fitting. The authors mention that this factor is necessary due to non-sequence dependent factors such as the temperature and magnetic field strength used in the experiment.*

> *Given these considerations, would it be possible to predict what this scaling factor should be based on such factors?*

There are still too few data to make such a prediction.

> *(2) The authors mention that the Lorentzian functional form fits the data better than a Gaussian functional form, but do not present these results.*

We tested the different functional forms at the early stage of the method development. The improvement of the Lorentzian over the Gaussian was slight and we simply decided on the Lorentzian and did not go back and do a systematic analysis.

> *(3) The authors mention that they conducted five-fold cross validation to determine if differences between amino acid parameters are statistically significant. While two pairs are mentioned in the text, there are 190 possible pairs, and it would be informative to more rigorously examine the differences between all such pairs.*

We now present t-test results for other pairs in new Fig. S3.

> ***Reviewer #3 (Public Review):***
>
> *The manuscript by Qin and Zhou presents an approach to predict dynamical properties of an intrinsically disordered protein (IDP) from sequence alone. In particular, the authors train a simple (but useful) machine learning model to predict (rescaled) NMR R2 values from sequence. Although these R2 rates only probe some aspects of IDR dynamics and the method does not provide insight into the molecular aspects of processes that lead to perturbed dynamics, the method can be useful to guide experiments.*
>
> *A strength of the work is that the authors train their model on an observable that directly relates to protein dynamics. They also analyse a relatively broad set of proteins which means that one can see actual variation in accuracy across the proteins.*
>
> *A weakness of the work is that it is not always clear what the measured R2 rates mean. In some cases, these may include both fast and slow motions (intrinsic R2 rates and exchange contributions). This in turn means that it is actually not clear what the authors are predicting. The work would also be strengthened by making the code available (in addition to the webservice), and by making it easier to compare the accuracy on the training and testing data.*

Our method predicts the sequence dependence of R2, which is dominated by slower dynamics.

> ***Recommendations for the authors:***
>
> ***Reviewer #2 (Recommendations For The Authors):***
>
> *(1) Should make sure to define abbreviations such as NMR and SeqDYN.*

We now spell out NMR at first use. SeqDYN is the name of our method and is not an abbreviation.

> *(2) The authors do not mention how the curves in Figure 2A are calculated.*

As we stated in the figure caption, these curves are drawn to guide the eye.

> *(3) May be interesting to explore how the model parameters (q) correlate with different measures of hydrophobicity (especially those derived for IDPs like Urry). This may point to a relationship between amino acid interactions and amino acid dynamics*

We now present the correlation between q and a stickiness parameter refined by Tesei et al. (new ref 45) and used for predicting phase separation equilibrium (new Fig. S6).

> *(4) The authors demonstrate that secondary structure cannot be fully accounted for by their model. They make a correction for extended alpha-helices, but the strength of this correction seems to only be based on one sequence. Would a more rigorous secondary structure correction further improve the model and perhaps allow its transferability to ordered proteins?*

We have five 4 test cases (Figs. 4E, F and 5H, I). However, we doubt that the SeqDYN method will be transferable to ordered proteins.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *Changes that could strengthen the manuscript substantially.*
>
> *(1) The authors do not really define what they mean by dynamics, but given that they train and benchmark on R2 measurements, the directly probe whatever goes into the measured R2. Using a direct measurement is a strength since it makes it clear what they are predicting. It also, however, makes it difficult to interpret. This is made clear in the text when the authors, for example write "R2 is the one most affected by slower dynamics (10s of ns to 1 μs and beyond)." First, with the "and beyond" it could literally mean anything. Second, the "normal" R2 rate is limited up to motions up to the (local) "tumbling/reorganization" time (which is much faster), so any slow motions that go into R2 would be what one would normally call "exchange". The authors should thus make it clearer what exactly it is they are probing. In the end, this also depends on the origin of the experimental data, and whether the "R2" measurements are exchange-free or not. This may be a mixture, which hampers interpretations and which may also explain some of the rescaling that needs to be done.*

We now remove “and beyond”, and also raise the possibility that R2 measurements based on 15N relaxation may have relatively small exchange contributions (p. 17).

> *(2) Related to the above, the authors might consider comparing their predictions to the relaxation experiments from Kriwacki and colleagues on a fragment of p27. In that work, the authors used dispersion experiments to probe the dynamics on different timescales. The authors would here be able to compare both to the intrinsic R2 rates (when slow motions are pulsed away) as well as the effective R2 rates (which would be the most common measurement). This would help shed light on (at least in one case) which type of R2 the prediction model captures. https://doi.org/10.1021/jacs.7b01380*

We now report this comparison in new Fig. S5 and discuss its implications (p. 17-18).

> *(3) In some cases, disagreement between prediction and experiments is suggested to be due to differences in temperature, and hence is used as an argument for the rescaling done. Here, the authors use a factor of 2.0 to explain a difference between 278K and 298K, and a factor of 2.4 to explain the difference between 288K and 298K. It would be surprising if the temperature effect from 288K->298K is larger than from 278K->298K. Does this not suggest that the differences come as much from other sources?*

Note that the scaling factors 2.0 and 2.4 were obtained on two different IDPs. It is most likely that different IDPs have different scaling factors for temperature change. As a simple model, the tumbling time for a spherical particle scales with viscosity and the particle volume; correspondingly the scaling factor for temperature change should be greater for a larger particle than for a smaller particle.

> *(4) The authors find (as have others before) aromatic residues to be common at/near R2 peaks. They suggest this to be indicative for Pi-Pi interactions. Could this not be other types of interactions since these residues are also "just" more hydrophobic? Also, can the authors rule out that the increased R2 rates near aromatic residues is not due to increased dynamics, but simply due to increased Rex-terms due to greater fluctuations in the chemical shifts near these residues (due to the large ring current effects).*

We noted both pi-pi and cation-pi as possible interactions that raise R2. There can be other interactions involving aromatic residues, but it's unlikely to be only hydrophobic as Arg is also in the high-*q* end. For the same reason, a ring-current based explanation would be inadequate.

> *(5) The authors write: "We found that, by filtering PsiPred (http://bioinf.cs.ucl.ac.uk /psipred) (35) helix propensity scores (p,-.) with a very high cutoff of 0.99, the surviving helix predictions usually correspond well with residues identified by NMR as having high helix propensities." It would be good to show the evidence for this in the paper, and quantify this statement.*

The cases of most interest are the ones with long predicted helices, of which there are only 3 in the training set. For Sev-NT and CBP-ID4, we already summarize the NMR data for helix identification in the first paragraph of Results; the third case is KRS-NT, which we elaborate in p. 14.

> *(6) When analysing the nine test proteins, it would be very useful for the reader to get a number for the average accuracy on the nine proteins and a corresponding number for the training proteins. The numbers are maybe there, but hard to find/compare. This would be important so that one can understand how well the model works on the training vs testing data.*

We now present the mean RMSE comparison in p. 14.

> *(7) The authors write: "The $_q$ parameters, while introduced here to characterize the propensities of amino acids to participate in local interactions, appear to correlate with the tendencies of amino acids to drive liquid-liquid phase separation." It would be good to show this data and quantify this.*

We now list supporting data in p. 18 and present new Fig. S6 for further support.

> *(8) It is great that the authors have made a webservice available for easy access to the work. They should in my opinion also make the training code and data available, as well as the final trained model. Here it would also be useful to show the results from the use of a Gaussian that was also tested, and also state whether this model was discarded before or after examining the testing data.*

We have listed the IDP characteristics and sequences in Tables S1 and S2. We're unsure whether we can disseminate the experimental R2 data without the permission of the original

authors. As for the Gaussian function, as stated above, it was abandoned at an early state, before examining the testing data.

> *Changes that would also be useful*
>
> *(1) The authors should make it clearer what they predict and what they don't. They mention transient helix formation and various contacts, but there isn't a one-to-one relationship between these structural features and R2 rates. Hence, they should make it clearer that they don't predict secondary structure and that an increased R2 rate may be indicative of many different structural/dynamical features on many different time scales.*

We clearly state that we apply a helix boost after the regular SeqDYN prediction.

> *(2) The authors write "Instead, dynamics has emerged as a crucial link between sequence and function for IDPs" and cite their own work (reference 1) as reference for this statement. As far as I can see, that work does not study function of IDPs. Maybe the authors could cite additional work showing that the dynamics (time scales) affects function of IDPs beyond "just" structure? Otherwise, the functional consequences are not clear. Maybe the authors mean that R2 rates are indicative of (residual) structure, but that is not quite the same. Also, even in that case, there are likely more appropriate references.*

Ref. 1 summarized a number of scenarios where dynamics is related to function.

> *(3) The authors might want to look at some of the older literature on interpreting NMR relaxation rates and consider whether some of it is worth citing.*
>
> *Fitting/understanding R2 profiles https://doi.org/10.1021/bi020381o https://doi.org/10.1007/s10858-006-9026-9*
>
> *MD simulations and comparisons to R2 rates without ad hoc reweighting (in addition to the papers from the authors themselves). https://doi.org/10.1021/ja710366c https://doi.org/10.1021/ja209931w*

The R2 data for the two unfolded proteins are very helpful! We now present the comparison of these data to SeqDYN prediction in Fig. 6C, D. The MD papers are superseded by more recent studies (e.g., refs. 1 and 14).

> *There are more like these.*
>
> *(4) In the analysis of unfolded lysozyme, I assume that the authors are treating the methylated cysteines (which are used in the experiments) simply as cysteine. If that is the case, the authors should ideally mention this specifically.*

Treatment of methylated cysteines is now stated in the Fig. 6 caption.

> *(5) The authors write "Pro has an excessively low msR2 [with data from only two IDPs (32, 33)], but that is due to the absence of an amide proton." It would be useful with an explanation why lacking a proton gives rise to low 15N R2 rates.*

That assertion originated from ref. 32.

> *(6) When applying the model, the authors predict msR2 and then compare to experimental R2 by rescaling with a factor gamma. It would be good to make it clearer*

> *whether this parameter is always fitted to the experiments in all the comparisons. It would be useful to list the fitted gamma values for all the proteins (e.g. in Table S1).*

We already give a summary of the scaling factors (“For 39 of the 45 IDPs, Y values fall in the range of 0.8 to 2.0 s–1”, p. 10).

> *(7) p. 14 "nineth" -> "ninth"*

Corrected

https://doi.org/10.7554/eLife.88958.2.sa0