Predicting the Sequence-Dependent Backbone Dynamics of Intrinsically Disordered Proteins

Short title: Predictor for IDP sequence-dependent dynamics

Sanbo Qin[1] and Huan-Xiang Zhou[1,2,*]

[1]Department of Chemistry and [2]Department of Physics, University of Illinois Chicago, Chicago, IL 60607, USA

*Corresponding author. E-mail: hzhou43@uic.edu

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.

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_2 have 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_2 whereas 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, 14). 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_2 value 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_2 as 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_2 profile 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_2 measured 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 1; Figure 1). 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 Figure 2, 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 2.

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 1 and Figure 3A). 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 (Figure 3-figure supplement 1A) 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 1 and Figure 3-figure supplement 1B). 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_2 across 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 (Figure 3-figure supplement 1C), 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 (Figure 3-figure supplement 1D), 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_2 data (Table 1 and Figure 3-figure supplement 1E). 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_2 values of 30 IDPs have moderate sequence variations, with σ_{R_2} ranging from 0.5 to 1.5 s^{-1} (Table 1); the histogram of σ_{R_2} calculated for the entire training set peaks around 0.75 s^{-1} (Figure 3A). There is a moderate correlation between σ_{R_2} and $\bar{R}_2$ (Figure 3A, 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

138 temperature. Still, only two of the five IDPs with high $\bar{R}_2$ attributable to lower temperature or
139 presence of glycerol are among the seven IDPs with high sequence variations ($\sigma_{R_2} > 2$ s^{-1}).
140 Therefore the sequence variation of R_2 as captured by σ_{R_2} manifests mostly the intrinsic effect
141 of the IDP sequence, not the influence of external factors such as temperature or magnetic
142 field strength. The mean σ_{R_2} value among the training set is 1.24 s^{-1}.

143 One way to eliminate the influence of external factors is to scale the R_2 values of each
144 IDP by its $\bar{R}_2$; we refer to the results as scaled R_2, or $\mathrm{s}R_2$. We then pooled the $\mathrm{s}R_2$ values for
145 all residues in the training set, and separated them according to amino-acid types. The amino
146 acid type-specific mean $\mathrm{s}R_2$ values, or $\mathrm{ms}R_2$, are displayed in Figure 3B. The seven amino
147 acids with the highest $\mathrm{ms}R_2$ in descending order are Trp, Arg, Tyr, Phe, Ile, His, and Leu.
148 The presence of all the four aromatic amino acids in this "high-end" group immediately
149 suggests π-π stacking as important for raising R_2; the presence of Arg further implicates
150 cation-π interactions. In the other extreme, the seven amino acids with the lowest $\mathrm{ms}R_2$ in
151 ascending order are Gly, Cys, Val, Asp, Ser, Thr, and Asn. Gly is well-known as a flexible
152 residue; it is also interesting that all the four amino acids with short polar sidechains are found
153 in this "low-end" group. Pro has an excessively low $\mathrm{ms}R_2$ [with data from only two IDPs (32,
154 33)], but that is due to the absence of an amide proton.

155

156 **The SeqDYN model and parameters**

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

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

This one-residue model does only minutely better than the null model, with a $\overline{\mathrm{RMSE}}$ 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 $L_{\mathrm{corr}} = b^{-1/2}$. 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 $\mathrm{ms}R_2$.

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_2 data in the entire training set. Others (16, 17, 24) have modeled R_2 as the average of some parameters over a window; a window has an extremely abrupt sequence-distance dependence (1 for $s < L_{\mathrm{corr}}$ and 0 for $s > L_{\mathrm{corr}}$).

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 (ϒ) 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_2 data for a total of 3924 residues. We excluded the 42 Pro residues in the training set because, as already noted, their R_2 values 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 Figure 3B alongside $\mathrm{ms}R_2$. 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 $\mathrm{ms}R_2$, 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 $\mathrm{ms}R_2$. The q values thus also suggest that π-π and cation-π 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 $\mathrm{ms}R_2$ 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 (Figure 4A). 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 $\mathrm{ms}R_2$ and q with amino-acid size. We measure the latter property by the molecular mass (m). As shown in Figure 4B, both $\mathrm{ms}R_2$ 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 $\mathrm{ms}R_2$ or q and bulkiness, with R^2 just below 0.4 (Figure 4C).

The optimized value of b is 3.164×10^{-2}, corresponding to an L_{corr} of 5.6 residues. The resulting optimized $\overline{\mathrm{RMSE}}$ 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 $\overline{\mathrm{RMSE}}$, 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, $\overline{\mathrm{RMSE}}$ rises quickly, reaching 1.22 s^{-1} at L_{corr} = 0. The latter $\overline{\mathrm{RMSE}}$ 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, Υ 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 3.164×10^{-2} from full training as stated above and 3.163×10^{-2} from jackknife training. The close agreement in parameter values between full training and jackknife training suggests no significant bias in the training set.

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 Figure 4-figure supplement 1.

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 Figure 5A. 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} (Figure 3A). 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 the comparison of predicted and measured R_2 values for five IDPs: MKK4, α-synuclein, Mev-P_{NTD}, Sev-NT, and CBP-ID4 (Figure 5B-F). A simple common feature is the falloff of R_2 at 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 Figure 2, with residues colored according to the predicted R_2 values. 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 (Figure 5A). Figure 5B displays the measured and predicted R_2 for MKK4. SeqDYN correctly predicts higher R_2 values 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_2 values for α-synuclein are shown in Figure 5C. Here SeqDYN correctly predicts higher R_2 near 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_2 peaks can be explained by a combination of secondary structure formation (β-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 β-strands and between Asp121 and Lys96). Figure 5D shows that SeqDYN gives excellent R_2 predictions 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_2 profile 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 Figure 5E). A similar situation occurs for CBP-ID4, where the peak in the second long helical region (around Glu113) is underpredicted (green curve in Figure 5F). 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_2 data 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_2 data of CBP-ID4 alone were used to parameterize SeqDYN, the measured R_2 profile is closely reproduced (Figure 5-figure supplement 1). The reversal in R_2 peak 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 β-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 β-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:

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

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 (Figure 5E, 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} (Figure 5F, red curve). The only other IDP for which PsiPred plus

319 filtering predicts a long helix is the N-terminal region of lysyl-tRNA synthetase (KRS-NT).
320 The authors who studied this protein did not report on secondary structure (36), but feeding
321 their reported chemical shifts to the TALOS+ server
322 (https://spin.niddk.nih.gov/bax/nmrserver/talos/) (37) found only short stretches of residues
323 that fall into the helical region of the Ramachandran map. The SeqDYN prediction for KRS-
324 NT is already good [RMSE(-1) = 0.83 s^{-1}]; applying a helix boost would deteriorate the
325 RMSE(-1) to 1.16 s^{-1}.
326
327 **Further test on a set of nine IDPs**
328 We have reserved nine IDPs for testing SeqDYN (parameterized on the training set of 45
329 IDPs). The level of disorder in these test proteins also spans the full range, from absence of
330 secondary structures [ChiZ N-terminal region (12), Pdx1 C-terminal region (11), and TIA-1
331 prion-like domain (17)] to presence of transient short helices [synaptobrevin-2 (7), α-
332 endosulfine (38), YAP (6), angiomotin-like 1 (AMOTL1) (39)] to formation of stable long
333 helices [FtsQ (14) and CAHS-8 (9)]. For eight of the nine test IDPs, the RMSEs of SeqDYN
334 predictions are lower than the experimental σ_{R_2} values, by an average of 0.66 s^{-1}. For the
335 ninth IDP (Pdx1), the SeqDYN RMSE is slightly higher, by 0.06 s^{-1}, than the experimental
336 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
337 1.00 s^{-1} for the training set in the leave-one-out cross-validation.
338 The comparison of predicted and measured R_2 profiles along the sequence is presented
339 in Figure 6A-I. For ChiZ, SeqDYN correctly predicts the major peak around Arg25 and the
340 minor peak around Arg46 (Figure 6A). The R_2 profile of Pdx1 is largely featureless, except
341 for a dip around Gly216, which is correctly predicted by SeqDYN (Figure 6B). Correct
342 prediction is also obtained for the higher R_2 in the first half of TIA-1 prion-like domain than
343 in the second half (Figure 6C). SeqDYN gives an excellent prediction for synaptobrevin-2,
344 including a linear increase up to Arg56 and the major peak around Trp89 (Figure 6D).

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 (Figure 6E). 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 (Figure 6F).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 (Figure 6G). However, whereas the two peaks have approximately equal heights in the measured R_2 profile, 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 (14)] but underestimates the magnitude (RMSE = 2.32 s^{-1}; green curve in Figure 6H). 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 Figure 6H). 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 Figure 6I). 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 Figure 6I).

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 Figure 7A. 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 (Figure 7B).

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_2 corresponded 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_2 profile (Figure 7C). The deepest valley indeed occurs over a Gly/Ala-rich stretch, $\underline{G_{125}ADAQGA_{131}}$, 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 (Figure 3B). 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 β-sheets (major elements of folded ubiquitin) are less resistant to denaturation than α-helices. SeqDYN predicts a relatively flat R_2 profile (Figure 7D), but the reason is that the ubiquitin sequence lacks a contiguous stretch of high-q amino acids.

Discussion

398 We have developed a powerful method, SeqDYN, that predicts the backbone amide
399 transverse relaxation rates (R_2) of IDPs. The method is based on IDP sequences, is extremely
400 fast, and available as a web server at https://zhougroup-uic.github.io/SeqDYNidp/. The
401 excellent performance supports the notion that the ns-dynamics reported by R_2 is coded by the
402 local sequence, comprising up to 6 residues on either side of a given residue. The amino-acid
403 types that contribute the most to coupling within a local sequence are aromatic (Trp, Tyr, Phe,
404 and His), Arg, and long branched aliphatic (Ile and Leu), suggesting the importance of π-π,
405 cation-π, and hydrophobic interactions in raising R_2. These interactions are interrupted by Gly
406 and amino acids with short polar sidechains (Ser, Thr, Asn, and Asp), leading to reduced R_2.
407 Transient short helices produce moderate elevation in R_2, whereas stable long helices result in
408 a big boost in R_2. Tertiary contacts can also raise R_2, but appear to be infrequent in most IDPs
409 (1).

410 It is also possible that R_2 reported by backbone amide ^{15}N relaxation (as is the case for
411 most of the IDPs studied here) may not be particularly sensitive to exchange effects, which
412 likely involve tertiary contact formation. For the D2 domain of p27^{Kip1}, the exchange
413 contributions measured using ^{15}N relaxation were small (< 2.5 s^{-1}) but were as large as 25 s^{-1}
414 when measured by high-power ^{1}H relaxation dispersion (42). This experiment measures the
415 effective transverse relation rate, $R_{2,\mathrm{eff}}$, over a range of effective radiofrequency ω_{eff}. The
416 exchange contribution is maximal for the value $R_{2,\mathrm{eff}}^{\mathrm{low}\,\omega_{\mathrm{eff}}}$ at low ω_{eff} but is largely quenched
417 for the value $R_{2,0}^{\mathrm{app}}$ in the high-ω_{eff} limit. The SeqDYN prediction for this IDP matches much
418 better with $R_{2,0}^{\mathrm{app}}$ than with $R_{2,\mathrm{eff}}^{\mathrm{low}\,\omega_{\mathrm{eff}}}$ (Figure 8). It is not clear whether this IDP is unique in
419 forming persistent tertiary contacts that give rise to substantial exchange contributions or the
420 ^{1}H relaxation dispersion experiment is unique in reporting the exchange contributions. At the
421 minimum, SeqDYN yields the exchange-free portion of the transverse relaxation rate,
422 enabling easy identification of residues that potentially participate in tertiary contacts. For the
423 D2 domain of p27^{Kip1}, SeqDYN correctly predicts the $R_{2,0}^{\mathrm{app}}$ local maxima at W76 and Y88. It
424 is these same two residues that show substantial exchange contributions and putatively

425 participate in tertiary contact (42). Therefore local contacts may seed tertiary contacts. If $R_{2,\mathrm{eff}}$
426 data with substantial exchange contributions become available for more IDPs, SeqDYN may
427 be retrained to make predictions for IDPs forming persistent tertiary contacts.

428 The q parameters, while introduced here to characterize the propensities of amino
429 acids to participate in local interactions, appear to correlate with the tendencies of amino acids
430 to drive liquid-liquid phase separation. Consistent with the rank order of q, Trp, Tyr, and Arg
431 have been reported to be strong drivers of phase separation, Lys is a moderate driver, whereas
432 Gly and Ser suppress phase separation (10, 33, 43). Recent measurements of the threshold
433 concentration produced the following order for the propensity of phase separation by eight
434 nonpolar amino acids in homotetrapeptides of the form XXssXX (ss: backbone disulfide
435 bond): Trp > Phe > Leu > Met > Ile > Val > Ala > Pro (44). This order is the same as that of
436 the q parameters, except that the q values of Ile and Val are in the second and last places,
437 respectively. Threshold concentrations of IDPs are now predicted reasonably well by coarse-
438 grained simulations where each amino acid is modeled by a single bead with a Lennard-Jones
439 diameter d_0 and a stickiness parameter λ (45). Our q parameter shows a good correlation (R^2
440 = 0.59) with the compound parameter $d_0^3\lambda$ (Figure 9). Therefore the q parameter may serve as
441 a predictor for the tendency of an amino acid to drive phase separation. In essence, the same
442 ability of an amino acid, e.g., Trp, to form interactions with neighboring residues of an IDP in
443 the free state also applies when it comes to interactions with residues on neighboring chains in
444 a dense phase.

445 Our method incorporates ideas from a number of previous efforts at describing R_2.
446 The first serious effort was by Schwalbe et al. (15), who accounted for contributions from
447 neighboring residues as additive terms, instead of multiplicative factors as in SeqDYN. Cho et
448 al. (16) and Delaforge et al. (24) used the running average of the bulkiness parameter over a
449 window of five to nine residues as a qualitative indicator of R_2. Here again the calculation
450 was based on an additive model. Sekiyama et al. (17) employed a multiplicative model, with
451 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 (Figure 5-figure supplement 1), 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 3 we compare the RMSEs of SeqDYN predictions with those for R_2 calculations from MD simulations. For five of these IDPs: A1-LCD, Aβ40, 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 β-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_2 than 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.

Methods

Collection of IDPs with measured R_2

Starting from six nonhomologous IDPs in our previous MD study (1), we obtained R_2 data for eight IDPs from the Bimolecular Magnetic Resonance Data Bank (BMRB; https://bmrb.io); data for two other IDPs were from our collaborators (12, 14). Most of the 54 IDPs studied here were from searching the literature. Disorder was judged by dispersion in backbone amide proton chemical shifts, NOE, and SCS. R_2 data that were not available from the authors or BMRB were obtained by digitizing R_2 plots presented in figures of published papers, using WebPlotDigitizer (https://automeris.io/WebPlotDigitizer) (48) and further inspected visually.

Homology of IDPs was checked by sequence alignment using Clustal W (http://www.clustal.org/clustal2) (49), and presented as a clock-like tree using the "ape" package (http://ape-package.ird.fr) (50). IDPs that had discernible homology with the selected training set were removed. Removed IDPs included HOX-SCR and β-synuclein from our previous MD study (1), due to homology with HOX-DFD and α-synuclein, respectively.

Coding for SeqDYN

The training of SeqDYN was coded in python, similar to our previous work for predicting residue-specific membrane association propensities (ReSMAP; https://zhougroup-uic.github.io/ReSMAPidp/) (19). The cost function was the sum of mean-squared-errors for the IDPs in the training set. We used the least_squares function in scipy.optimize, with Trust Region Reflective as the minimization algorithm and all parameters restricted to the positive

range. For the web server (https://zhougroup-uic.github.io/SeqDYNidp/), we rewrote the prediction code javascript.

ACKNOWLEDGEMENTS

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

REFERENCES

1. S. Dey, M. MacAinsh, H. X. Zhou, Sequence-Dependent Backbone Dynamics of Intrinsically Disordered Proteins. *J Chem Theory Comput* **18**, 6310-6323 (2022).
2. A. R. Camacho-Zarco *et al.*, NMR Provides Unique Insight into the Functional Dynamics and Interactions of Intrinsically Disordered Proteins. *Chem Rev* **122**, 9331-9356 (2022).
3. J. Klein-Seetharaman *et al.*, Long-Range Interactions Within a Nonnative Protein. *Science* **295**, 1719-1722 (2002).
4. A. Abyzov *et al.*, Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation. *J Am Chem Soc* **138**, 6240-6251 (2016).
5. R. Schneider *et al.*, Visualizing the Molecular Recognition Trajectory of an Intrinsically Disordered Protein Using Multinuclear Relaxation Dispersion NMR. *J Am Chem Soc* **137**, 1220-1229 (2015).
6. M. Feichtinger *et al.*, Long-range structural preformation in yes-associated protein precedes encounter complex formation with TEAD. *iScience* **25**, 104099 (2022).
7. N.-A. Lakomek, H. Yavuz, R. Jahn, Á. Pérez-Lara, 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 (2019).

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

9. A. Malki *et al.*, Intrinsically Disordered Tardigrade Proteins Self-Assemble into Fibrous Gels in Response to Environmental Stress. *Angew Chem Int Ed Engl* **61**, e202109961 (2022).

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

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

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

13. L. Yu, R. Bruschweiler, Quantitative prediction of ensemble dynamics, shapes and contact propensities of intrinsically disordered proteins. *PLoS Comput Biol* **18**, e1010036 (2022).

14. S. T. Smrt, C. A. Escobar, S. Dey, T. A. Cross, H. X. Zhou, An Arg/Ala-rich helix in the N-terminal region of M. tuberculosis FtsQ is a potential membrane anchor of the Z-ring. *Commun Biol* **6**, 311 (2023).

15. H. Schwalbe *et al.*, 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 (1997).

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

17. N. Sekiyama *et al.*, ALS mutations in the TIA-1 prion-like domain trigger highly condensed pathogenic structures. *Proc Natl Acad Sci U S A* **119**, e2122523119 (2022).

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

19. S. Qin, A. Hicks, S. Dey, R. Prasad, H. X. Zhou, ReSMAP: Web Server for Predicting Residue-Specific Membrane-Association Propensities of Intrinsically Disordered Proteins. *Membranes* **12**, 773 (2022).

20. M. R. Jensen *et al.*, 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 (2008).

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

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

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

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

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

26. S. Milles *et al.*, An ultraweak interaction in the intrinsically disordered replication machinery is essential for measles virus function. *Sci Adv* **4**, eaat7778 (2018).

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

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

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

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

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

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

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

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

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

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

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

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

39. A. Vogel, A. Crawford, A. Nyarko, Multivalent Angiomotin-like 1 and Yes-associated protein form a dynamic complex. *Protein Sci* **31**, e4295 (2022).

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

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

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

43. J. Wang *et al.*, A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688-699 e616 (2018).

44. Y. Zhang, R. Prasad, S. Su, D. Lee, H. X. Zhou, Amino Acid-Dependent Phase Equilibrium and Material Properties of Tetrapeptide Condensates. *Cell Rep Phys Sci* **5**, 102218 (2024).

45. G. Tesei, K. Lindorff-Larsen, Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res Eur* **2**, 94 (2022).

46. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118**, e2016239118 (2021).

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

48. A. Rohatgi, *Webplotdigitizer: Version 4.6* (2022).

49. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

50. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

51. N. Rezaei-Ghaleh, G. Parigi, M. Zweckstetter, Reorientational Dynamics of Amyloid-beta from NMR Spin Relaxation and Molecular Simulation. *J Phys Chem Lett* **10**, 3369-3375 (2019).

52. S. Yao *et al.*, Characterisation of the conformational preference and dynamics of the intrinsically disordered N-terminal region of Beclin 1 by NMR spectroscopy. *Biochim Biophys Acta* **1864**, 1128-1137 (2016).

53. B. Szalaine Agoston, D. Kovacs, P. Tompa, A. Perczel, Full backbone assignment and dynamics of the intrinsically disordered dehydrin ERD14. *Biomol NMR Assign* **5**, 189-193 (2011).

54. Z. Zheng, D. Ma, T. L. Yahr, L. Chen, The transiently ordered regions in intrinsically disordered ExsE are correlated with structural elements involved in chaperone binding. *Biochem Biophys Res Commun* **417**, 129-134 (2012).

55. C. W. Lawrence, S. A. Showalter, Carbon-Detected (15)N NMR Spin Relaxation of an Intrinsically Disordered Protein: FCP1 Dynamics Unbound and in Complex with RAP74. *J Phys Chem Lett* **3**, 1409-1413 (2012).

56. K. A. Burke, A. M. Janke, C. L. Rhine, N. L. Fawzi, Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol Cell* **60**, 231-241 (2015).

57. T. Gruber *et al.*, Macromolecular Crowding Induces a Binding Competent Transient Structure in Intrinsically Disordered Gab1. *J Mol Biol* **434**, 167407 (2022).

58. M. O. Ebert, S. H. Bae, H. J. Dyson, P. E. Wright, NMR relaxation study of the complex formed between CBP and the activation domain of the nuclear hormone receptor coactivator ACTR. *Biochemistry* **47**, 1299-1308 (2008).

59. R. Kiss, D. Kovacs, P. Tompa, A. Perczel, Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain. *Biochemistry* **47**, 6936-6945 (2008).

60. F. C. Lopes *et al.*, Pliable natural biocide: Jaburetox is an intrinsically disordered insecticidal and fungicidal polypeptide derived from jack bean urease. *Febs J* **282**, 1043-1064 (2015).

61. M. De Avila, K. A. Vassall, G. S. Smith, V. V. Bamm, G. Harauz, The proline-rich region of 18.5 kDa myelin basic protein binds to the SH3-domain of Fyn tyrosine kinase with the aid of an upstream segment to form a dynamic complex in vitro. *Biosci Rep* **34**, e00157 (2014).

62. S. Mokhtarzada, C. Yu, A. Brickenden, W. Y. Choy, Structural characterization of partially disordered human Chibby: insights into its function in the Wnt-signaling pathway. *Biochemistry* **50**, 715-726 (2011).

63. M. Schiavina *et al.*, Ensemble description of the intrinsically disordered N-terminal domain of the Nipah virus P/V protein from combined NMR and SAXS. *Sci Rep* **10**, 19574 (2020).

64. J. L. Neira *et al.*, Dynamics of the intrinsically disordered protein NUPR1 in isolation and in its fuzzy complexes with DNA and prothymosin alpha. *Biochim Biophys Acta Proteins Proteom* **1867**, 140252 (2019).

65. B. Mateos *et al.*, The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators. *J Mol Biol* **432**, 3093-3111 (2020).

66. M. Xie, D. W. Li, J. Yuan, A. L. Hansen, R. Bruschweiler, Quantitative Binding Behavior of Intrinsically Disordered Proteins to Nanoparticle Surfaces at Individual Residue Level. *Chemistry* **24**, 16997-17001 (2018).

67. J. Song *et al.*, Intrinsically disordered gamma-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure. *Proc Natl Acad Sci U S A* **105**, 1505-1510 (2008).

68. C. Olivieri *et al.*, Multi-state recognition pathway of the intrinsically disordered protein kinase inhibitor by protein kinase A. *Elife* **9** (2020).

69. A. Borgia *et al.*, Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61-66 (2018).

70. M. A. Ahmed, V. V. Bamm, G. Harauz, V. Ladizhansky, The BG21 isoform of Golli myelin basic protein is intrinsically disordered with a highly flexible amino-terminal domain. *Biochemistry* **46**, 9700-9712 (2007).

71. V. Csizmok, I. C. Felli, P. Tompa, L. Banci, I. Bertini, Structural and dynamic characterization of intrinsically disordered human securin by NMR spectroscopy. *J Am Chem Soc* **130**, 16873-16879 (2008).

72. T. Mittag *et al.*, Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494-506 (2010).

73. X. Wang *et al.*, A large intrinsically disordered region in SKIP and its disorder-order transition induced by PPIL1 binding revealed by NMR. *J Biol Chem* **285**, 4951-4963 (2010).

74. R. Thapar, G. A. Mueller, W. F. Marzluff, The N-terminal domain of the Drosophila histone mRNA binding protein, SLBP, is intrinsically disordered with nascent helical structure. *Biochemistry* **43**, 9390-9400 (2004).

75. N. Rezaei-Ghaleh *et al.*, Local and global dynamics in intrinsically disordered synuclein. *Angew Chem Int Ed Engl* **57**, 15262-15266 (2018).

76. I. R. Chandrashekaran *et al.*, Structure and Functional Characterization of the Conserved JAK Interaction Region in the Intrinsically Disordered N-Terminus of SOCS5. *Biochemistry* **54**, 4672-4682 (2015).

77. P. Barre, D. Eliezer, Structural transitions in tau k18 on micelle binding suggest a hierarchy in the efficacy of individual microtubule-binding repeats in filament nucleation. *Protein Sci* **22**, 1037-1048 (2013).

78. E. A. Cino, M. Karttunen, W. Y. Choy, Effects of molecular crowding on the dynamics of intrinsically disordered proteins. *PLoS One* **7**, e49876 (2012).

79. J. Harris, M. Shadrina, C. Oliver, J. Vogel, A. Mittermaier, Concerted millisecond timescale dynamics in the intrinsically disordered carboxyl terminus of gamma-tubulin induced by mutation of a conserved tyrosine residue. *Protein Sci* **27**, 531-545 (2018).

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

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

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

Table 1. Experimental conditions, means and standard deviations of measured R_2, and SeqDYN prediction RMSEs

Protein name	# of res	Temp (K)	B_0 (MHz)	$\bar{R}_2$ (s^{-1})	σ_{R_2} (s^{-1})	RMSE (s^{-1})	PMID(ref)
Training set (45)[a]							
A1-LCD	131	298	800	2.68	0.46	0.60	32029630(10)
Aβ40	40	278	600	3.40	0.92	0.38	31181936(51)
Ash1	83	278	800	9.80	1.40	1.41	27807972(28)
Beclin1	165	288	800	5.37	1.03	1.14	27288992(52)
CAPRIN1	103	303	600	5.34	0.88	0.72	31898464(33)
CBP-ID4	207	283	700	5.45	2.55	2.01;1.90[b]	29790640(32)
GbnD4-DHD	91	280	700	6.81	1.55	1.28	29309054(30)
ERD14	185	288	600	3.96	0.87	0.54	21336827(53)
ExsE	88	298	600	3.18	0.88	0.76	22138394(54)
FCP1	85	298	500	2.94	0.54	0.43	26286791(55)
FUS	163	298	850	3.48	0.51	0.54	26455390(56)
GAb1	82	298	500	3.99	0.88	0.89	34929201(57)
hACTR	69	304	600	3.26	0.47	0.49	18177052(58)
Hahellin	92	298	800	9.94	2.69	2.85	24671380(23)
hCSD1	141	298	500	3.56	0.93	0.99	18537264(59)
HOX-DFD	90	298	600	6.98	3.15	1.99	30802457(22)
hZIP4-ICL2	100	283	800	9.54	2.37	1.58	30793391(31)
Jaburetox	94	298	800	6.01	2.30	2.27	25605001(60)
KRS-NT	72	303	600	3.26	0.93	0.83	24983501(36)
MBP-xα2	70	295	600	3.83	0.60	0.54	25343306(61)
MKK4	86	278	850	4.49	1.42	0.63	29276882(24)
N-Cby	63	298		4.19	1.20	1.25	21182262(62)
Niv-P_{NTD}	406	288	700	5.41	1.82	1.66	33177626(63)
NS5A-D2D3	268	278	800	8.62	3.85	2.14	26445449(27)
NUPR1	93	298	600	2.98	0.82	0.76	31325636(64)
OPN	220	310	800	2.59	0.82	0.54	31794728(65)
p53TAD	73	298	850	2.72	0.66	0.33	30240067(66)
PDEγ	87	298		3.96	1.05	0.71	18230733(67)
PKIα	75	300	900	3.41	0.87	0.52	32338601(68)
Mev-P_{NTD}	304	298	950	2.92	0.59	0.48	30140745(26)
ProTα	113	283	800	3.40	0.56	0.43	29466338(69)
Pup	64	298	850	2.66	0.51	0.43	30240067(66)
rmBG21	199	300	600	4.06	0.90	0.63	17676872(70)
RPB1	201	277	850	6.48	1.74	1.33	28945358(29)
securin	202	283	500	5.49	1.13	1.08	19053469(71)
Sev-NT	124	298	600	3.20	1.42	0.76;0.38[b]	27112095(4)
Sic1	92	278	500	3.34	0.59	0.48	20399186(72)
SKIPN	71	298		5.64	1.05	1.46	20007319(73)
SLBP-NT	113	298	600	3.96	1.40	1.61	15260482(74)
α-synuclein	140	298	600	2.96	0.53	0.44	30184304(75)
SOCS5-JIR	70	303	800	4.32	2.36	1.91	26173083(76)
tau K18	129	283	700	4.12	0.95	0.83	23740819(77)
TC1	106	298	600	4.65	1.61	1.24	23189168(78)
TDP-43	151	283	500	4.07	1.51	0.96	27545621(8)
γ-tubulin-CT	39	288	500	2.23	0.35	0.27	29127738(79)
Test set (9)							
AMOTL1	207	283	800	8.45	2.55	2.04	35481651(39)
CAHS-8	233	303	850	4.43	3.25	2.36;1.92[b]	34750927(9)
ChiZ	64	298	800	4.33	0.89	0.74	32585849(12)

α-endosulfine	121	298	800	3.21	0.81	0.48	34346186(38)
FtsQ	99	305	800	6.44	3.78	2.32;1.71[b]	36959324(14)
Pdx1	83	298	500	2.98	0.70	0.76	30525611(11)
synaptobrevin-2	96	278	600	5.54	1.80	0.72	30975750(7)
TIA-1	91	310	800	4.01	0.89	0.55	36112647(17)
YAP	122	298	800	3.19	1.44	1.23	35378854(6)

[a]For training set, RMSE is calculated for prediction based on leave-one-out training (using 44 IDPs).

[b]First number is for SeqDYN prediction; second number is after applying a helix boost.

Table 2. Sequences of 54 IDPs[a]

Training set (45)

>A1-LCD (Uniprot ID P04256, residues 186-320; deletion of 258-263)
GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGG
NYNNQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
>Aβ40 (Uniprot ID Q28053, residues 7-46)
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
>Ash1 (Uniprot ID P34233, residues 420-500)
GASASSSPSPSTPTKSGKMRSRSSSPVRPKAYTPSPRSPNYHRFALDSPPQSPRRSSNSSITKKGSRRSSGSSPT
RHTTRVCV
>Beclin1 (Uniprot ID Q14457, residues 1-150)
MGSSHHHHHHSQDPMEGSKTSNNSTMQVSFVSQRSSQPLKLDTSFKILDRVTIQELTAPLLTTAQAKPGETQEEE
TNSGEEPFIETPRQDGVSRRFIPPARMMSTESANSFTLIGEASDGGTMENLSRRLKVTGDLFDIMSGQTDVDHPL
SEESTDTLLDQLDTY
>CAPRIN1 (Uniprot ID Q14444, residues 607-709)
SRGVSRGGSRGARGLMNGYRGPANGFRGGYDGYRPSFSNTPNSGYTQSQFSAPRDYSGYQRDGYQQNFKRGSGQS
GPRGAPRGRGGPPRPNRGMPQMNTQQVN
>CBP-ID4 (Uniprot ID Q92793, residues 1852-2057)
MQQQIQHRLQQAQLMRRRMATMNTRNVPQQSLPSPTSAPPGTPTQQPSTPQTPQPPAQPQPSPVSMSPAGFPSVA
RTQPPTTVSTGKPTSQVPAPPPPAQPPPAAVEAARQIEREAQQQQHLYRVNINNSMPPGRTGMGTPGSQMAPVSL
NVPRPNQVSGPVMPSMPPGQWQQAPLPQQQPMPGLPRPVISMQAQAAVAGPRMPSVQ
>GbnD4-DHD (Uniprot ID A0A808VWJ6, residues 412-482)
MKHHHHHHHHGGLVPRGSHGSDEGVPDALRADTVPRAGPVRYARRRYWIGEARSDALAPAAPLEREPLPAEAMGA
YFAIRRTDADDTVAAH
>ERD14 (Uniprot ID P42763, residues 1-185)
MAEEIKNVPEQEVPKVATEESSAEVTDRGLFDFLGKKKDETKPEETPIASEFEQKVHISEPEPEVKHESLLEKLH
RSDSSSSSSSEEEGSDGEKRKKKKEKKKPTTEVEVKEEEKKGFMEKLKEKLPGHKKPEDGSAVAAAPVVVPPPVE
EAHPVEKKGILEKIKEKLPGYHPKTTVEEEKKDKE
>ExsE (Uniprot ID Q9I322, residues 1-81)
MKIESIPPVQPSQDAGAEAVGHFEGRSVTRAAVRGDDRSSVAGLARWLARNVAGDPRSEQALQRLADGDGTPLEA
RTVRRREFLEGSS
>FCP1 (Uniprot ID Q9Y5B0, residues 879-961)
PGPEEQEEEPQPRKPGTRRERTLGAPASSERSAAGGRGPRGHKRKLNEEDAASESSRESSNEDEGSSSEADEMAK
ALEAELNDLM
>FUS (Uniprot ID P35637, residues 1-163)
MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSGYSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQGY
GSTGGYGSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQPQSGSYSQQPSYGGQQQSYGQQQSYN
PPQGYGQQNQYNS
>GAb1 (Uniprot ID B7Z3B9, residues 510-591)
SSPMIKPKGDKQVEYLDLDLDSGKSTPPRKQKSSGSGSSVADERVDYVVVDQQKTLALKSTREAWTDGRQSTESE
TPAKSVK
>hACTR (Uniprot ID Q9Y6Q9, residues 1023-1091)
GTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEIDRALGIPELVNQGQALEPK
>Hahellin (Uniprot ID Q2SHN6, residues 162-252)
MGEKTVKLYEDTHFKGYSVELPVGDYNLSSLISRGALNDDLSSARVPSGLRLEVFQHNNFKGVRDFYTSDAAELS
RDNDASSVRVSKMETTN
>hCSD1 (Uniprot ID P20810, residues 137-277)
AVPVESKPDKPSGKSGMDAALDDLIDTLGGPEETEEENTTYTGPEVSDPMSSTYIEELGKREVTIPPKYRELLAK
KEGITGPPADSSKPIGPDDAIDALSSDFTCGSPTAAGKKTEKEESTEVLKAQSAGTVRSAAPPQEK
>HOX-DFD (Uniprot ID P07548, residues 337-426)
TDGERIIYPWMKKIHVAGVANGSYQPGMEPKRQRTAYTRHQILELEKEFHYNRYLTRRRRIEIAHTLVLSERQIK
IWFQNRRMKWKKDNK
>hZIP4-ICL2 (Uniprot ID Q6P5W5, residues 424-498)
GDRGPEFELGTLPRDPEDLEDGPCGHSSHSHGGHSHGVSLQLAPSELRQPKPPHEGSRADLVAEESPELLNPEPR
RLSPELRLLPYGHGLSAWSHPQFEK
>Jaburetox (Uniprot ID I1K3K3, residues 230-320)
MGPVNEANCKAAMEIVCRREFGHKEEEDASEGVTTGDPDCPFTKAIPREEYANKYGPTIGDKIRLGDTDLIAEIE
KDFALYGDESVFGGGKVIH
>KRS-NT (Uniprot ID Q15046, residues 1-72)
MAAVQAAEVKVDGSEPKLSKNELKRRLKAEKKVAEKEAKQKELSEKQLSQATAAATNHTTDNGVGPEEESVD
>MBP-xα2 (Uniprot ID P04370, residues 172-237)

SIGRFFSGDRGAPKRGSGKDSHTRTTHYGSLPQKSQHGRTQDENPVVHFFKNIVTPRTPPPSQGKGRGLS
>MKK4 (Uniprot ID P45985, residues 1-86)
MAAPSPSGGGGSGGGSGSGTPGPVGSPAPGHPAVSSMQGKRKALKLNFANPPFKSTARFTLNPNPTGVQNPHIER
LRTHSIESSGK
>N-Cby (Uniprot ID B0QY54, residues 1-63)
MPFFGNTFSPKKTPPRKSASLSNLHSLDRSTREVELGLEYGSPTMNLAGQSLKFENGQWIAET
>Niv-P_{NTD} (Uniprot ID P0C1C7, residues 1-406)
MDKLELVNDGLNIIDFIQKNQKEIQKTYGRSSIQQPSIKDQTKAWEDFLQCTSGESEQVEGGMSKDDGDVERRNL
EDLSSTSPTDGTIGKRVSNTRDWAEGSDDIQLDPVVTDVVYHDHGGECTGYGFTSSPERGWSDYTSGANNGNVCL
VSDAKMLSYAPEIAVSKEDRETDLVHLENKLSTTGLNPTAVPFTLRNLSDPAKDSPVIAEHYYGLGVKEQNVGPQ
TSRNVNLDSIKLYTSDDEEADQLEFEDEFAGSSSEVIVGISPEDEEPSSVGGKPNESIGRTIEGQSIRDNLQAKD
NKSTDVPGAGPKDSAVKEEPPQKRLPMLAEEFECSGSEDPIIRELLKENSLINCQQGKDAQPPYHWSIERSISPD
KTEIVNGAVQTADRQRPGTPMPKSRGIPIKK
>NS5A-D2D3 (Uniprot ID O92972, residues 2163-2419)
GHMASGSLRGGEPEPDVTVLTSMLTDPSHITAETAKRRLARGSPPSLASSSASQLSAPSLKATCTTHHDSPDADL
IEANLLWRQEMGGNITRVESENKVVILDSFEPLHADGDEREISVAAEILRKSRKFPSALPIWARPDYNPPLLESW
KDPDYVPPVVHGCPLPPTKAPPIPPPRRKRTVVLTESNVSSALAELATKTFGSSGSSAVDSGTATALPDQASDDG
DKGSDVESYSSMPPLEGEPGDPDLSDGSWSTVSEEASEDVVCC
>NUPR1 (Uniprot ID O60356, residues 2-82)
MRGSHHHHHHGSATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAANTNRPSPGGHER
KLVTKLQNSERKKRGARR
>OPN (Uniprot ID F1NSM8, residues 46-264)
MHQDHVDSQSQEHLQQTQNDLASLQQTHYSSEENADVPEQPDFPDVPSKSQETVDDDDDDDNDSNDTDESDEVFT
DFPTEAPVAPFNRGDNAGRGDSVAYGFRAKAHVVKASKIRKAARKLIEDDATTEDGDSQPAGLWWPKESREQNSR
ELPQHQSVENDSRPKFDSREVDGGDSKASAGVDSRESQGSVPAVDASNQTLESAEDAEDRHSIENNEVTR
>p53TAD (Uniprot ID P04637, residues 1-71)
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPRV
>PDEγ (Uniprot ID P61248, residues 1-87)
MNLEPPKAEIRSATRVMGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDITVIAPWEAFNH
LELHELAQYGII
>PKIα (Uniprot ID P61925, residues 2-76)
TDVETTYADFIASGRTGRRNAIHDILVSSASGNSNELALKLAGLDINKTEGEEDAQRSSTEQSGEAQGEAAKSES
>Mev-P_{NTD} (Uniprot ID P03422, residues 1-304)
MAEEQARHVKNGLECIRALKAEPIGSLAIEEAMAAWSEISDNPGQERATCREEKAGSSGLSKPCLSAIGSTEGGA
PRIRGQGPGESDDDAETLGIPPRNLQASSTGLQCYYVYDHSGEAVKGIQDADSIMVQSGLDGDSTLSGGDNESEN
SDVDIGEPDTEGYAITDRGSAPISMGFRASDVETAEGGEIHELLRLQSRGNNFPKLGKTLNVPPPPDPGRASTSG
TPIKKGTERRLASFGTEIASLLTGGATQCARKSPSEPSGPGAPAGNVPECVSNAALIQEWTPESGTTISPRSQNN
EEGG
>ProTα (Uniprot ID P06454, residues 1-111)
GPMSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEEEEEGGEEEEEEEEGDGE
EEDGDEDEEAESATGKRAAEDDEDDDVDTKKQKTDEDD
>Pup (Uniprot ID P9WHN4, residues 1-64)
MAQEQTKRGGGGGDDDDIAGSTAAGQERREKLTEETDDLLDEIDDVLEENAEDFVRAYVQKGGQ
>rmBG21 (Uniprot ID P04370, residues 2-190)
GNHSGKRELSAEKASKDGEIHRGEAGKKRSVGKLSQTASEDSDVFGEADAIQNNGTSAEDTAVTDSKHTADPKNN
WQGAHPADPGNRPHLIRLFSRDAPGREDNTFKDRPSESDELQTIQEDPTAASGGLDVMASQKRPSQRSKYLATAS
TMDHARHGFLPRHRDTGILDSIGRFFSGDRGAPKRGSGKVSLEHHHHHH
>RPB1 (Uniprot ID P24928, residues 1773-1970)
GHMSPNYTPTSPNYSPTSPSYSPTSPSYSPTSPSYSPSSPRYTPQSPTYTPSSPSYSPSSPSYSPASPKYTPTSP
SYSPSSPEYTPTSPKYSPTSPKYSPTSPKYSPTSPTYSPTTPKYSPTSPTYSPTSPVYTPTSPKYSPTSPTYSPT
SPKYSPTSPTYSPTSPKGSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN
>securin (Uniprot ID O95997, residues 1-202)
MATLIYVDKENGEPGTRVVAKDGLKLGSGPSIKALDGRSQVSTPRFGKTFDAPPALPKATRKALGTVNRATEKSV
KTKGPLKQKQPSFSAKKMTEKTVKAKSSVPASDDAYPEIEKFFPFNPLDFESFDLPEEHQIAHLPLSGVPLMILD
EERELEKLFQLGPPSPVKMPSPPWESNLLQSPSSILSTLDVELPPVCCDIDI
>Sev-NT (Uniprot ID Q07097, residues 401-524)
LSGGDGAYHEPTGGGAIEVALDNADIDLETEAHADQDARGWGGESGERWARQVSGGHFVTLHGAERLEEETNDED
VSDIERRIAMRLAERRQEDSATHGDEGRNNGVDHDEDDDAAAVAGIGGI
>Sic1 (Uniprot ID P38634, residues 1-90)
GSMTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAPPNSNMGMTSPFNG
LTSPQRSPFPKSSVKRT
>SKIPN (Uniprot ID G3V5R3, residues 59-129)
GDGGAFPEIHVAQYPLDMGRKKKMSNALAIQVDSEGKIKYDAIARQGQSKDKVIYSKYTDLVPKEVMNADD
>SLBP-NT (Uniprot ID Q9VAN6, residues 17-108)

MGSSHHHHHHSSGLVPRGSHMGSGSLNSSASSISIDVKPTMQSWAQEVRAEFGHSDEASSSLNSSAASCGSLAKK
ETADGNLESKDGEGREMAFEFLDGVNEVKFERLVKEEK
>α-synuclein (Uniprot ID P37840, residues 1-140)
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVT
AVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
>SOCS5-JIR (Uniprot ID A0A5E4BAI0, residues 12-81)
RSLRQRLQDTVGLCFPMRTYSKQSKPLFSNKRKIHLSELMLEKCPFPAGSDLAQKWHLIKQHTAPVSPHS
>tau K18 (Uniprot ID Q9MYX8, residues 186-314)
QTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQIVYKPVDLSKV
TSKAGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE
>TC1 (Uniprot ID Q9NR00, residues 1-106)
MKAKRSHQAIIMSTSLRVSPSIHGYHFDTASRKKAVGNIFENTDQESLERLFRNSGDKKAEERAKIIFAIDQDVE
EKTRALMALKKRTKDKLFQFLKLRKYSIKVH
>TDP-43 (Uniprot ID Q13148, residues 267-414)
GHMNRQLERSGRFGGNPGGFGNQGGFGNSRGGGAGLGNNQGSNMGGGMNFGAFSINPAMMAAAQAALQSSWGMMG
MLASQQNQSGPSGNNQNQGNMQREPNQAFGSGNNSYSGSNSGAAIGWGSASNAGSGSGFNGGFGSSMDSKSSGWG
M
>γ-tubulin-CT (Uniprot ID P53378, residues 439-473)
LLRGAAEQDSYLDDVLVDDENMVGELEEDLDADGDHKLV

Test set (9)

>AMOTL1 (Uniprot ID Q8IY63, residues 178-384)
STQPQQNNEELPTYEEAKAQSQFFRGQQQQQQQQGAVGHGYYMAGGTSQKSRTEGRPTVNRANSGQAHKDEALKE
LKQGHVRSLSERIMQLSLERNGAKQHLPGSGNGKGFKVGGGPSPAQPAGKVLDPRGPPPEYPFKTKQMMSPVSKT
QEHGLFYGDQHPGMLHEMVKPYPAPQPVRTDVAVLRYQPPPEYGVTSRPCQLPFPST
>CAHS-8 (Uniprot ID P0CU50, residues 1-227)
MSGRNVESHMERNEKVVVNNSGHADVKKQQQQVEHTEFTHTEVKAPLIHPAPPIISTGAAGLAEEIVGQGFTASA
ARISGGTAEVHLQPSAAMTEEARRDQERYRQEQESIAKQQEREMEKKTEAYRKTAEAEAEKIRKELEKQHARDVE
FRKDLIESTIDRQKREVDLEAKMAKRELDREGQLAKEALERSRLATNVEVNFDSAAGHTVSGGTTVSTSDKMEIK
RNENLYFQ
>ChiZ (Uniprot ID I6YA32, residues 1-64)
MTPVRPPHTPDPLNLRGPLDGPRWRRAEPAQSRRPGRSRPGGAPLRYHRTGVGMSRTGHGSRPV
>α-endosulfine (Uniprot ID O43768, residues 1-121)
MSQKQEEENPAEETGEEKQDTQEKEGILPERAEEAKLKAKYPSLGQKPGGSDFLMKRLQKGQKYFDSGDYNMAKA
KMKNKQLPSAGPDKNLVTGDHIPTPQDLPQRKSSLVTSKLAGGQVE
>FtsQ (Uniprot ID Q8IY63, residues 1-99)
MTEHNEDPQIERVADDAADEEAVTEPLATESKDEPAEHPEFEGPRRRARRERAERRAAQARATAIEQARRAAKRR
ARGQIVSEQNPAKPAARGVVRGLK
>Pdx1 (Uniprot ID P52945, residues 204-283)
GPGEEDKKRGGGTAVGGGGVAEPEQDCAVTSGEELLALPPPPPPGGAVPPAAPVAAREGRLPPGLSASPQPSSVA
PRRPQEPR
>synaptobrevin-2 (Uniprot ID P63027, residues 1-96)
MSATAATAPPAAPAGEGGPPAPPPNLTSNRRLQQTQAQVDEVVDIMRVNVDKVLERDQKLSELDDRADALQAGAS
QFETSAAKLKRKYWWKNLKMM
>TIA-1 (Uniprot ID P31483, residues 320-386)
MGSSHHHHHHHHHHHHSENLYFQGGQYVPNGWQVPAYGVYGQPWSQQGFNQTQSSAPWMGPNYSVPPPQGQNGSM
LPSQPAGYRVAGYETQ
>YAP (Uniprot ID P46937, residues 50-171)
AGHQIVHVRGDSETDLEALFNAVMNPKTANVPQTVPMRLRKLPDSFFKPPEPKSHSRQASTDAGTAGALTPQHVR
AHSSPASLQLGAVSPGTLTPTGVVSGPAATPTAQHLRQSSFEIPDDV

[a]Terminal tags and other insertions are underlined.

907 Table 3. RMSEs (s^{-1}) of R_2 predictions by SeqDYN and MD for 10 IDPs

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]

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

909 [b]Helix boost applied.

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

911 [d]From ref (1).

912 [e]RMSE is scaled down by a factor of 2.39, to correct for the effect of temperature (MD at 288
913 K; see Figure 3-figure supplement 1C).

914 [f]From ref (13).

915 [g]RMSE is scaled down by a factor of 2.99, to correct for the effects of temperature and
916 magnetic field (MD at 274 K and 850 MHz; see Figure 3-figure supplement 1B).

917 [h]Originally calculated in ref (12) with correction in ref (80).

918 [i]From ref (14).

919

920 **FIGURE CAPTIONS**

921 Figure 1. Clock-like tree plot showing lack of homology among the 45 IDPs.

922 The level of homology between two sequences is measured by the distance from their
923 convergence point to the center of the clock. The highest level of apparent identity is between
924 A1-LCD and TDP-43, at 25%, but these two proteins differ in both secondary structure
925 formation and R_2 characteristics. There is, however, a 20-residue overlap between the N-
926 terminus of MBP-xα2 and the C-terminus of rmBG21.

927 Figure 2. Representative conformations of five IDPs. (A-E) MKK4, α-synuclein, Mev-P_{NTD},
928 Sev-NT, and CBP-ID4. Conformations were initially generated using TraDES
929 (http://trades.blueprint.org) (81), selected to have radius of gyration close to predicted by a
930 scaling function $R_g = 2.54N^{0.522}$ (Å) (82). Conformations for residues predicted as helical by
931 PsiPred plus filtering were replaced by an ideal helix. Finally residues are colored according
932 to a scheme ranging from green for low predicted R_2 to red for high predicted R_2.

933 Figure 3. Properties of the 45 IDPs in the training set. (A) Histograms of means and standard
934 deviations, calculated for individual proteins. Curves are drawn to guide the eye. Inset:
935 correlation between $\bar{R}_2$ and σ_{R_2}. (B) Experimental mean scaled R_2 (msR_2) and SeqDYN q
936 parameters, for the 20 types of amino acids. Note that Pro residues have low msR_2 for the
937 lack of backbone amide proton. Amino acids are in descending order of q.

938 Figure 3-figure supplement 1. Possible effects of sequence length, temperature, and magnetic
939 field on R_2. (A) Lack of dependence of $\bar{R}_2$ on sequence length. (B) Counts of IDPs with R_2
940 measured at various temperatures. (C) Matching of R_2 profiles of Sev-NT measured at two
941 temperatures after uniform scaling. (D) Matching of R_2 profiles of A1-LCD measured at two
942 temperatures after uniform scaling. (E) Counts of IDPs with R_2 measured at various magnetic
943 fields.

944 Figure 4. SeqDYN model parameters. (A) Correlation between msR_2 and q. The values are
945 also displayed as bars in Figure 3B. (B) Correlation of msR_2 and q with amino-acid
946 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.

Figure 4-figure supplement 1. T-test of on the q parameters for pairs of amino acids. q parameters were obtained from five-fold cross-validation training, resulting in five independent values for each q parameter. Mean presented as red bars; standard deviation presented as error bar. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ns, not significant. q parameters for all neighboring pairs not explicitly indicated are not significantly different.

Figure 5. 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.

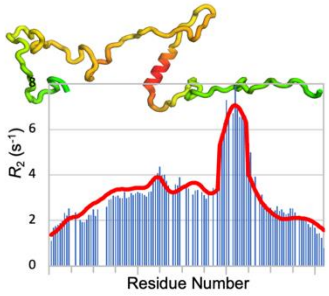
Figure 5-figure supplement 1. Close reproduction (curve) of the measured R_2 profile (bars) of CBP-ID4 when that set of data alone was used to parameterize SeqDYN. The resulting model has no value for predicting R_2 for other proteins.

Figure 6. Measured (bars) and predicted (curves) R_2 profiles 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_2 does 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 7. 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).

Figure 8. Comparison between SeqDYN prediction (curves) and effective transverse relaxation rate (bars) from ^{1}H dispersion relaxation experiment. (A) $R_{2,\mathrm{eff}}$ in the high-ω_{eff} limit. (B) $R_{2,\mathrm{eff}}$ at low ω_{eff}.

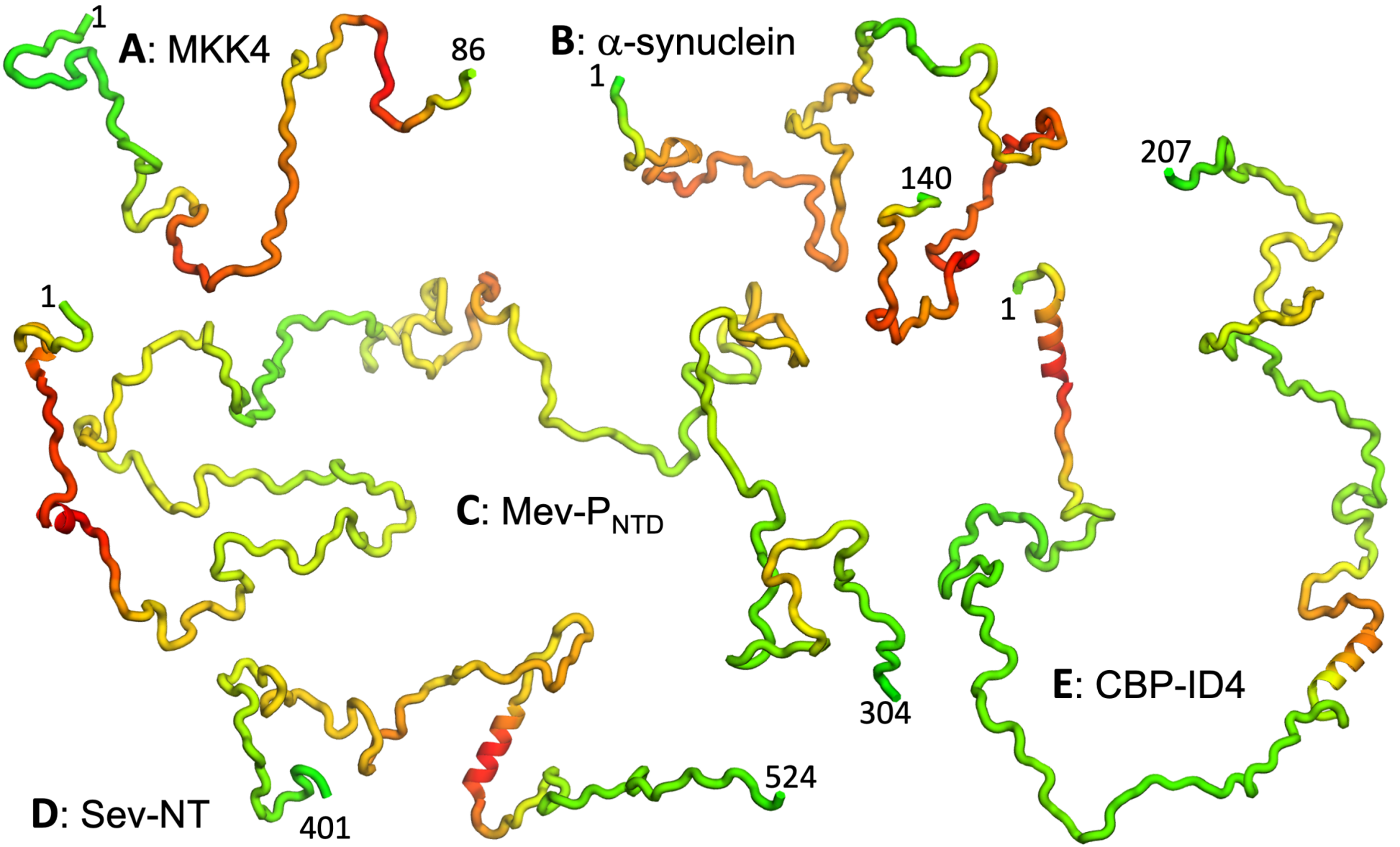
Figure 9. Correlation between the stickiness parameters (λ) and the NMR relaxation parameters (q). The regression line is shown as dashes.

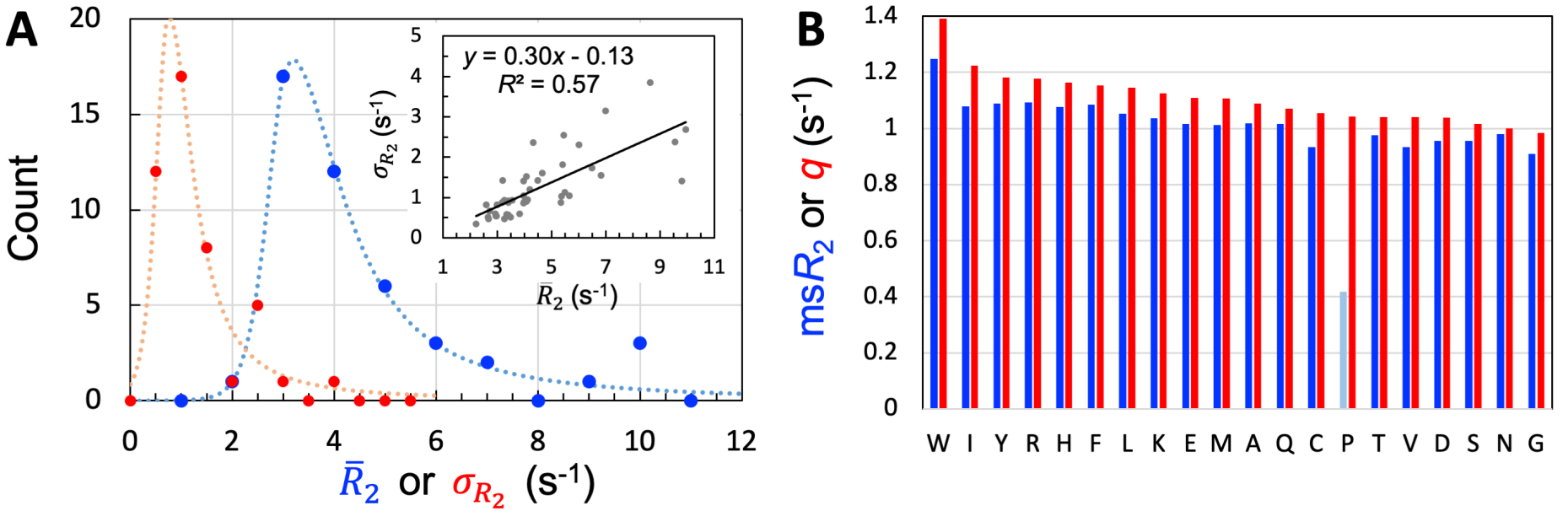


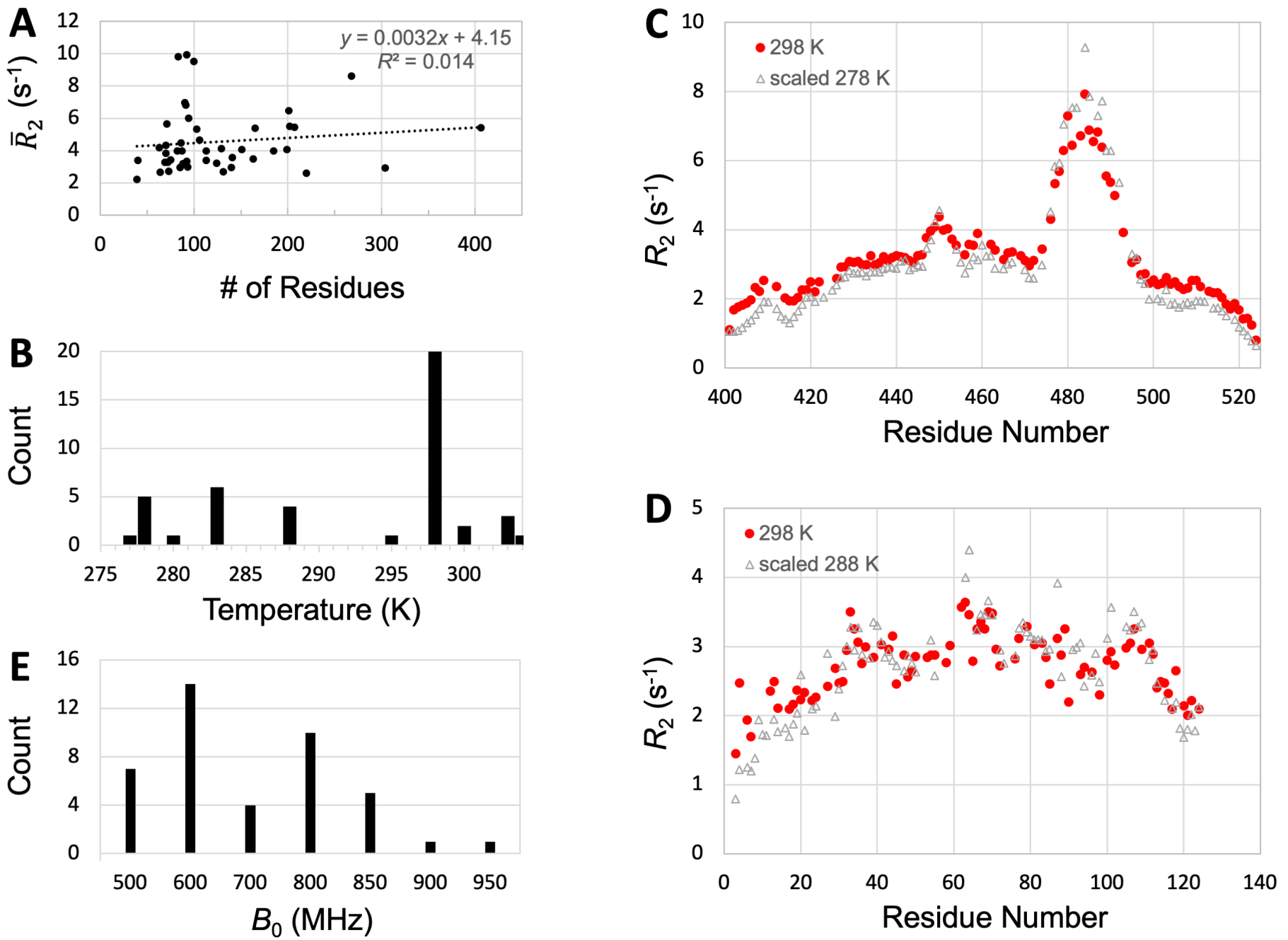
Dynamics is a crucial link between sequence and function for intrinsically disordered proteins (IDPs). Qin and Zhou developed a sequence-based method to predict IDP dynamics. The predictions provide deep physicochemical insight and can also serve as indicators of properties including phase-separation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

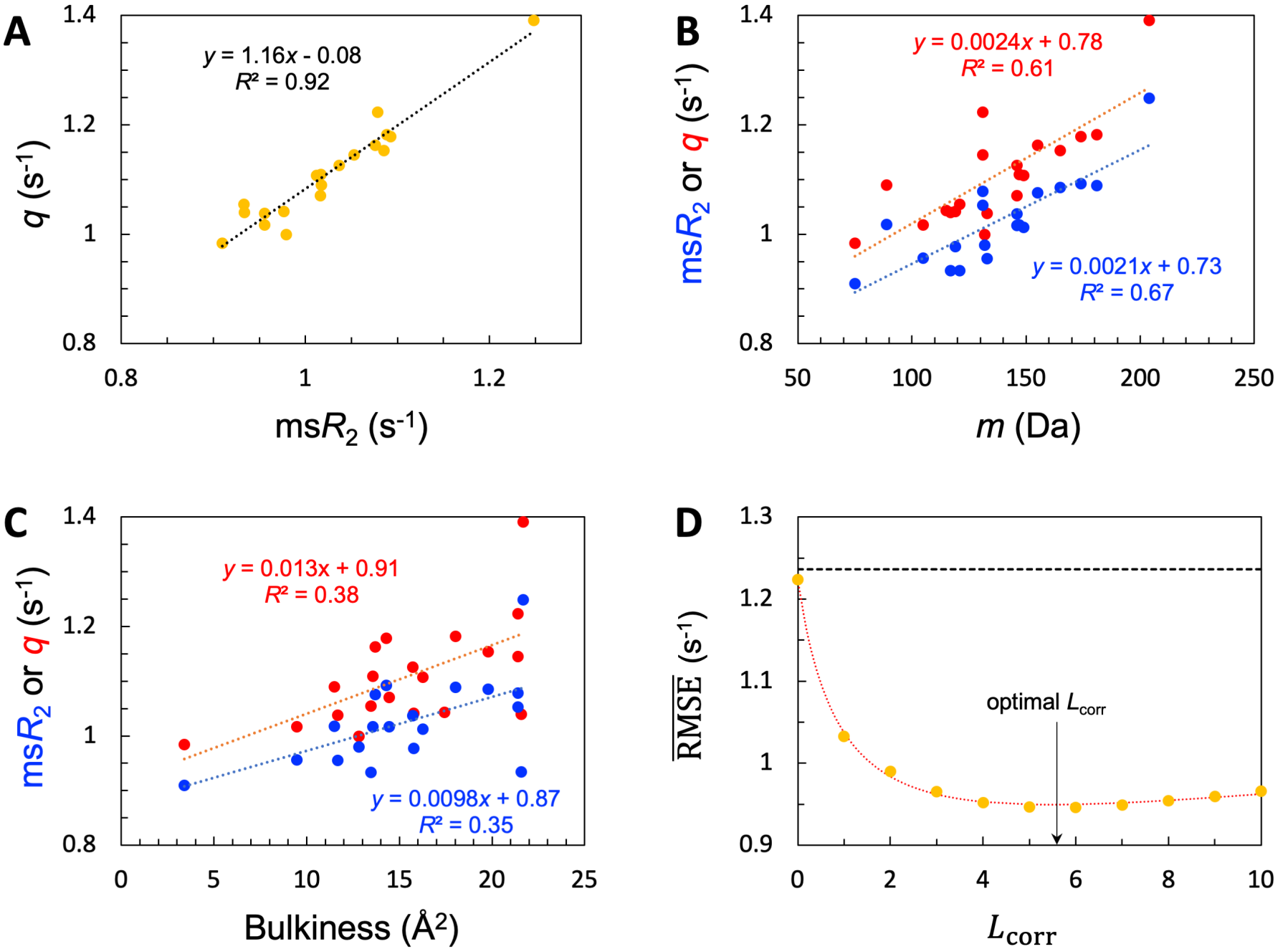
Keywords: NMR spectroscopy; phase separation; intrinsically disordered proteins; backbone dynamics

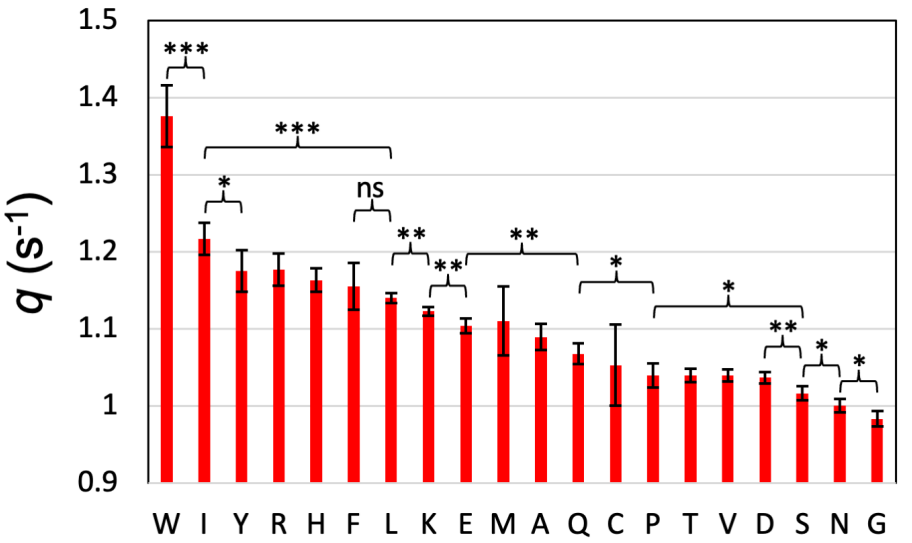
GbnD4-DHD
NUPR1
SLBP-NT
Beclin1
MKK4
FUS
RPB1
Ash1
CAPRIN1
TDP-43
A1-LCD
rmBG21
MBP-xα2
γtubulin-CT
ProTα
α-synuclein
KRS-NT
Jaburetox
Sev-NT
Pup
Mev-PNTD
ExsE
tauK18
PDEγ
SOCS5-JIR
OPN
hCSD1
FCP1
ERD14
Hahellin
PKIα
HOX-DFD
TC1
SKIPN
Aβ40
N-Cby
hACTR
NS5A-D2D3
GAb1
securin
hZIP4-ICL2
Sic1
CBP-ID4
p53TAD
Niv-PNTD

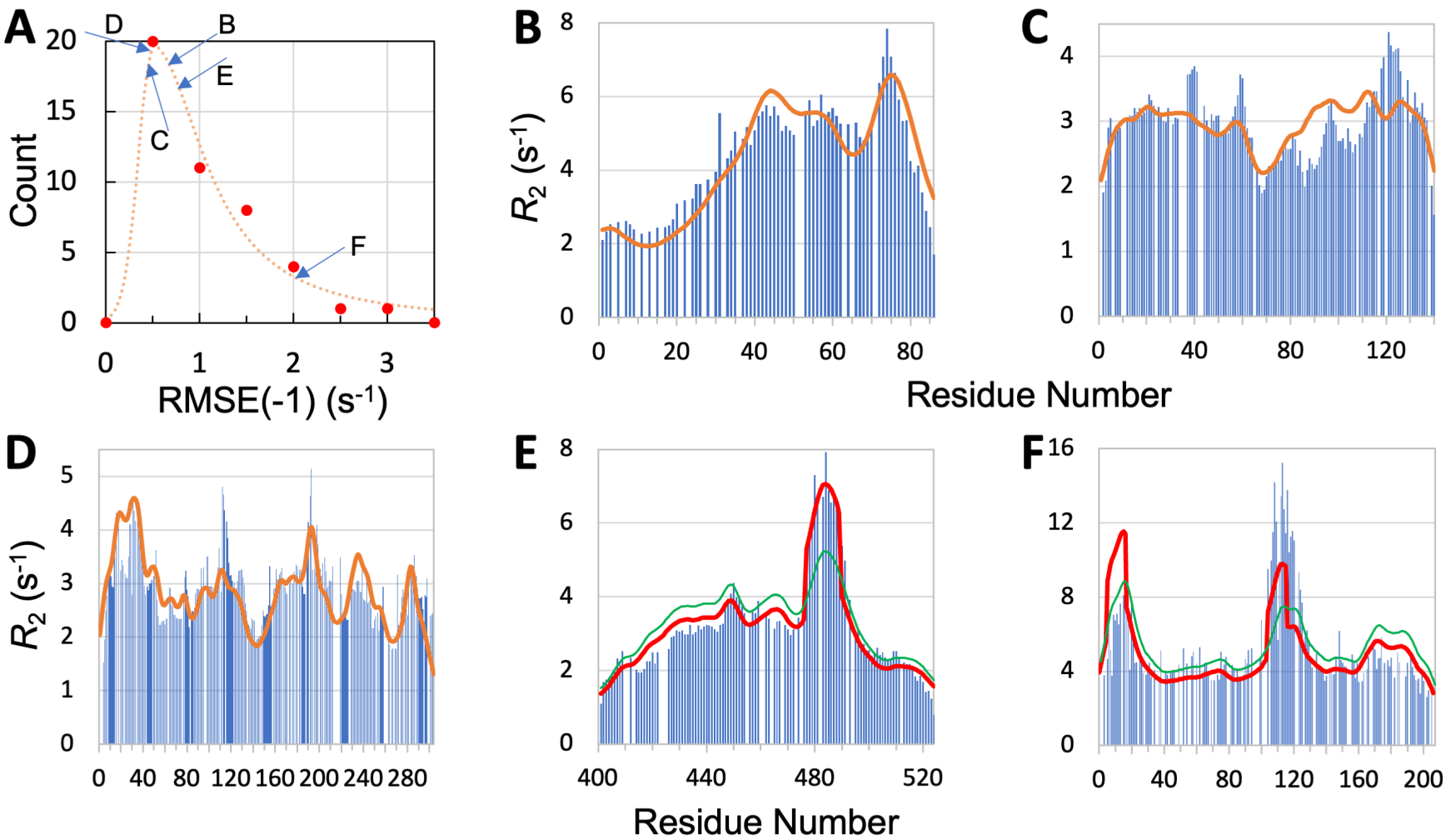

1
A: MKK4
86
B: α-synuclein
1
140
207
1
1
C: Mev-P$_{NTD}$
304
E: CBP-ID4
524
D: Sev-NT
401

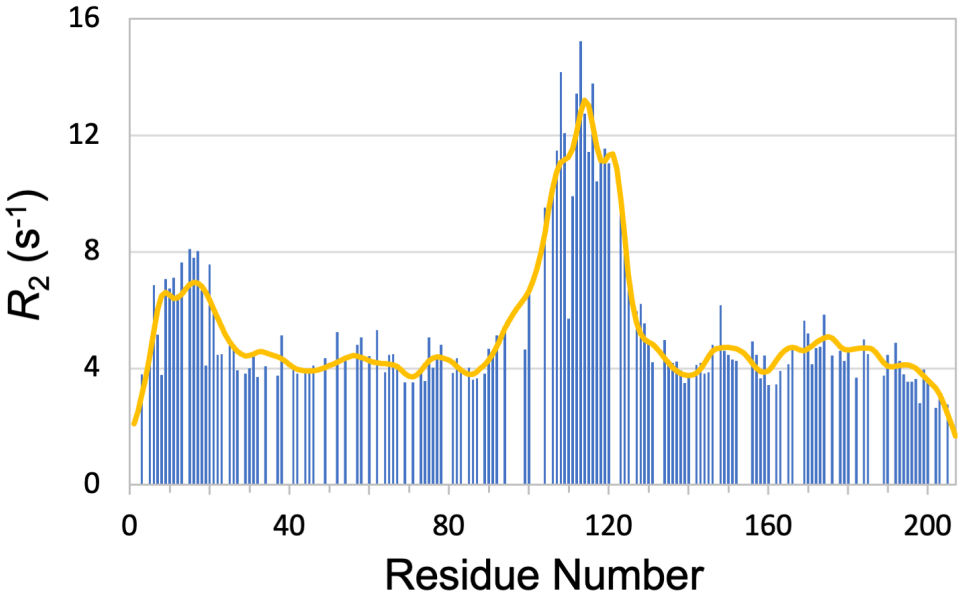

A
Count
20
15
10
5
0
0
2
4
6
8
10
12
$\bar{R}_2$ or σ_{R_2} (s^{-1})
y = 0.30x - 0.13
R^2 = 0.57
σ_{R_2} (s^{-1})
$\bar{R}_2$ (s^{-1})
B
msR_2 or q (s^{-1})
1.4
1.2
1
0.8
0.6
0.4
0.2
0
W I Y R H F L K E M A Q C P T V D S N G

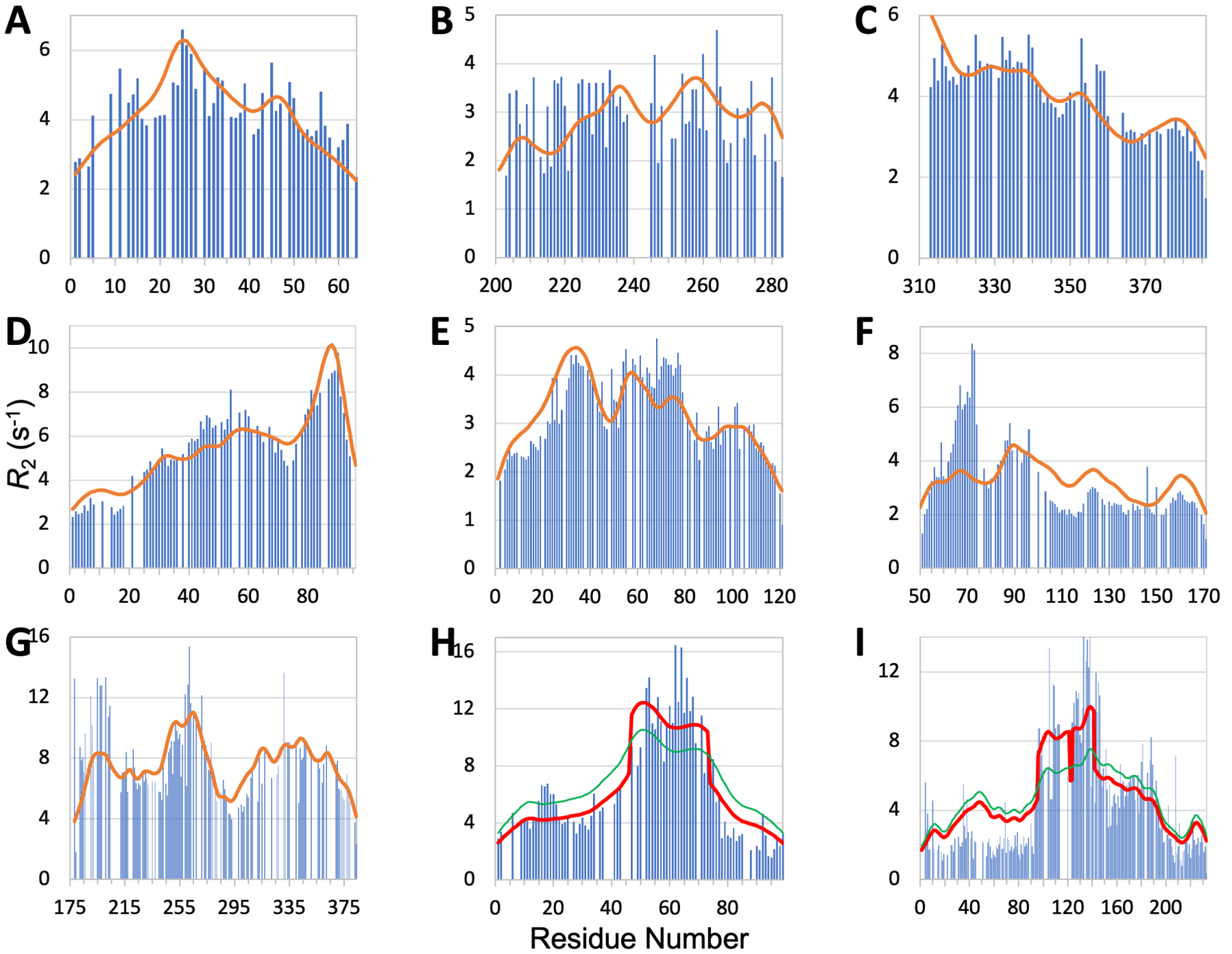

A
y = 0.0032x + 4.15
R² = 0.014
$\bar{R}_2$ (s^{-1})
# of Residues
B
Count
Temperature (K)
C
298 K
scaled 278 K
R_2 (s^{-1})
Residue Number
D
298 K
scaled 288 K
R_2 (s^{-1})
Residue Number
E
Count
B_0 (MHz)

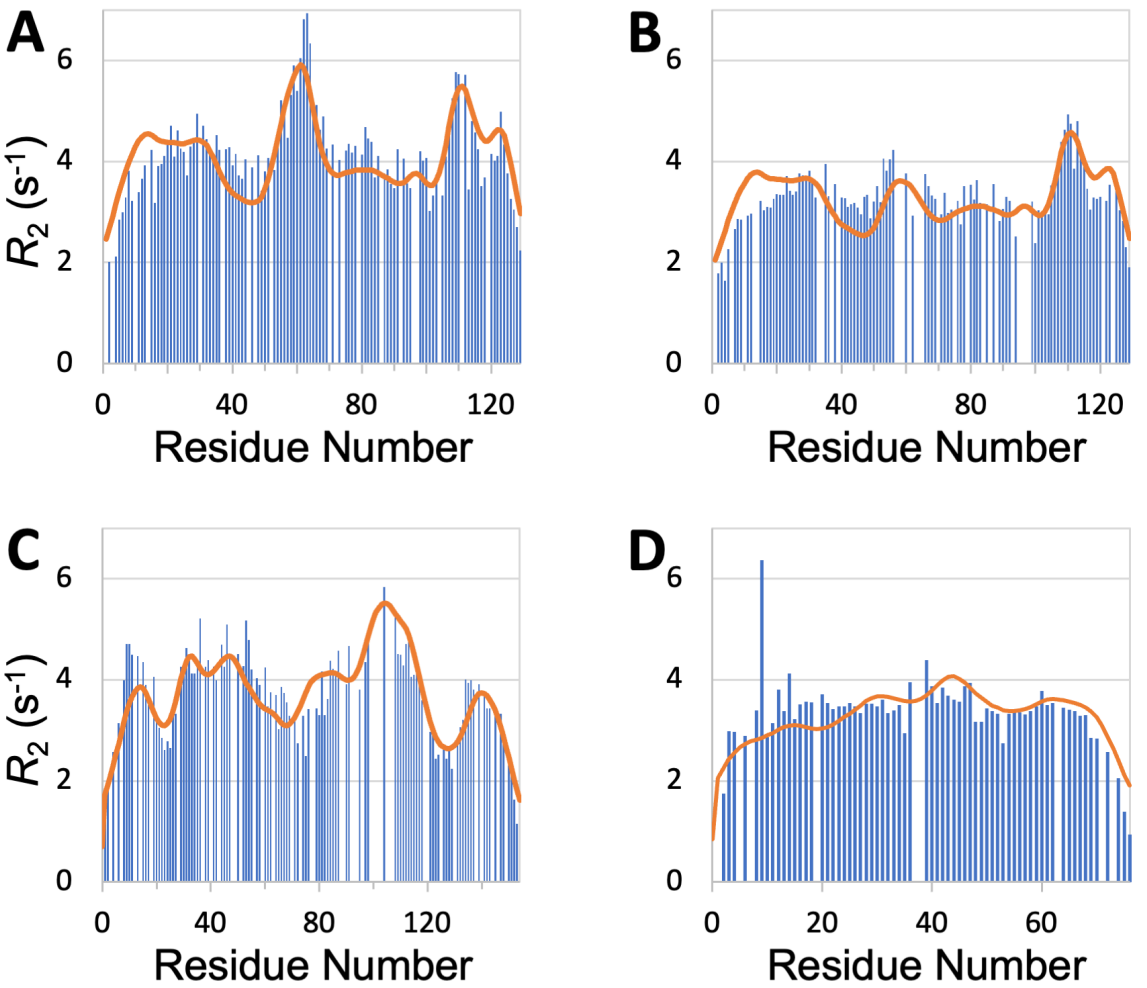

A
y = 1.16x - 0.08
R² = 0.92
q (s-1)
msR2 (s-1)
B
y = 0.0024x + 0.78
R² = 0.61
y = 0.0021x + 0.73
R² = 0.67
msR2 or q (s-1)
m (Da)
C
y = 0.013x + 0.91
R² = 0.38
y = 0.0098x + 0.87
R² = 0.35
msR2 or q (s-1)
Bulkiness (Å2)
D
optimal Lcorr
RMSE (s-1)
Lcorr

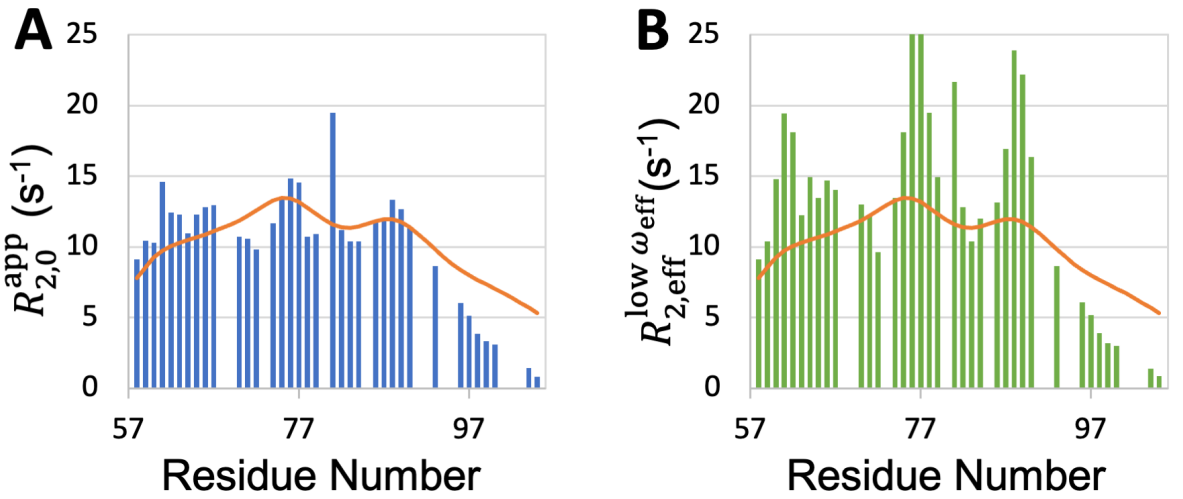

1.5
1.4
1.3
1.2
1.1
1
0.9
q (s-1)
***
***
*
ns
**
**
**
*
*
**
*
*
W I Y R H F L K E M A Q C P T V D S N G

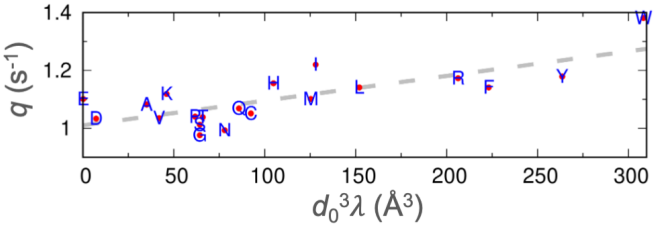

A
B
C
D
E
F
Count
RMSE(-1) (s-1)
R2 (s-1)
Residue Number


16
12
8
4
0
R_2 (s^{-1})
0
40
80
120
160
200
Residue Number


A
B
C
D
E
F
G
H
I
R_2 (s^{-1})
Residue Number


A
B
C
D
R_2 (s^{-1})
0
2
4
6
0
40
80
120
20
60
Residue Number


A
25
20
15
10
5
0
$R_{2,0}^{app}$ (s^{-1})
57
77
97
Residue Number
B
25
20
15
10
5
0
$R_{2,eff}^{low\ \omega_{eff}}$(s^{-1})
57
77
97
Residue Number



1.4
1.2
1
q (s-1)
0
50
100
150
200
250
300
d03λ (Å3)
E
D
A
K
V
Pt
S
G
N
O
C
H
I
M
L
R
F
Y
W