Biophysical Journal, Volume *113*

Supplemental Information

An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations

Mads Nygaard, Birthe B. Kragelund, Elena Papaleo, and Kresten Lindorff-Larsen

Supplementary Information

As starting points for deriving a relationship between R_g and R_h we take the scaling laws

$$R_{xs} = R_{0xs}N^{\nu_{xs}} \quad \text{(S1)}$$

and the phenomenological linear relationship

$$\frac{R_g}{R_h} = a_N R_g + b_N \quad \text{(S2)}$$

In Eq. S1 x={g,h} determines whether the relationship refers to R_g or R_h, and s={*folded*, *unfolded*, *IDP*} refers to which of these states the scaling law is meant to describe.

We proceed by making the assumption that we approximately can take the scaling laws for the folded state (F) and disordered state (U) to represent the compact and expanded regions of the R_g/R_h ratio, and thus obtain the following expression for the slope:

$$\begin{aligned} a_N &= \frac{\left(\frac{R_{gU}}{R_{hU}}\right) - \left(\frac{R_{gF}}{R_{hF}}\right)}{R_{gU} - R_{gF}} \\ &= \frac{\frac{R_{0gU}}{R_{0hU}}N^{(\nu_{gU}-\nu_{hU})} - \frac{R_{0gF}}{R_{0hF}}N^{(\nu_{gF}-\nu_{hF})}}{R_{0gU}N^{\nu_{gU}} - R_{0gF}N^{\nu_{gF}}} \\ &\approx \frac{\frac{R_{0gU}}{R_{0hU}} - \frac{R_{0gF}}{R_{0hF}}}{R_{0gU}N^{\nu_{gU}} - R_{0gF}N^{\nu_{gF}}} \end{aligned} \quad \text{(S3)}$$

where the approximation is justified by the experimental observation that the scaling exponents are similar for R_g and R_h, and depend mostly on whether the protein is compact or disordered.

Taking this expression and substituting in the values for compact states into Eq. S2, we obtain an expression for the intercept:

$$\begin{aligned} b_N &= \left(\frac{R_{gF}}{R_{hF}}\right) - a_N R_{gF} \\ &\approx \frac{R_{0gF}}{R_{0hF}}N^{(\nu_{gF}-\nu_{hF})} \\ &\quad - \frac{\left(\frac{R_{0gU}}{R_{0hU}} - \frac{R_{0gF}}{R_{0hF}}\right)R_{0gF}N^{\nu_{gF}}}{R_{0gU}N^{\nu_{gU}} - R_{0gF}N^{\nu_{gF}}} \\ &\approx \frac{R_{0gF}}{R_{0hF}} \\ &\quad - \frac{\left(\frac{R_{0gU}}{R_{0hU}} - \frac{R_{0gF}}{R_{0hF}}\right)R_{0gF}N^{\nu_{gF}}}{R_{0gU}N^{\nu_{gU}} - R_{0gF}N^{\nu_{gF}}} \end{aligned} \quad \text{(S4)}$$

Putting everything together we end up with:

$$\begin{aligned} \frac{R_g}{R_h} &\approx \frac{\left(\frac{R_{0gU}}{R_{0hU}} - \frac{R_{0gF}}{R_{0hF}}\right)\left(R_g - R_{0gF}N^{\nu_{gF}}\right)}{R_{0gU}N^{\nu_{gU}} - R_{0gF}N^{\nu_{gF}}} \\ &\quad + \frac{R_{0gF}}{R_{0hF}} \end{aligned} \quad \text{(S5)}$$

This expression is based on the assumption of a linear relationship (Eq. S2) and that the scaling exponents are the same for R_g and R_h. Furthermore, we note that the scaling laws (Eq. S1), in particular for the highly heterogeneous disordered state, are not meant to apply for individual structures, adding also to the approximate nature of the expression. For the same reason, we do not expect that the experimentally determined values for the scaling factors (R_{0xs}) will be optimal for describing properties of individual structures, and we instead treat these values as fitting parameters. Keeping the scaling exponents constant ($\nu_{xF} = 0.33$ and $\nu_{xU} = 0.6$) there are four parameters to be determined in Eq. S5.

Initial attempts to fit these parameters revealed strong correlations between the parameters and a large uncertainty in particular for R_{0gU}. This observation may also be related to empirical observation that all lines in Fig. 2A appear to cross near a single point. Assuming, however, that $R_{0gU} = R_{0gF}$ we obtained a much more robust fit of almost the same quality. With this further approximation we have:

$$\begin{aligned} \frac{R_g}{R_h} &\approx \frac{\left(\frac{R_{0gF}}{R_{0hU}} - \frac{R_{0gF}}{R_{0hF}}\right)\left(R_g - R_{0gF}N^{\nu_{gF}}\right)}{R_{0gF}(N^{\nu_{gU}} - N^{\nu_{gF}})} \\ &\quad + \frac{R_{0gF}}{R_{0hF}} \\ &= \frac{\alpha_1\left(R_g - \alpha_2 N^{0.33}\right)}{N^{0.60} - N^{0.33}} + \alpha_3 \end{aligned} \quad \text{(S6)}$$

where α_1, α_2, and α_3 are the three fitting parameters. As described in the main text, non-linear least-squares regression resulted in α_1=(0.216 ± 0.001)Å^{-1}, α_2=(4.06 ± 0.02)Å, and α_3=(0.821 ± 0.002).

Finally, We note here that a leading-order correction term to the scaling laws for the ensemble averaged values of R_h and R_g have also been shown to give rise to an explicit chain-length dependency of the R_g/R_h-ratio for disordered polymers[1].

[1] Dünweg, B., Reith, D., Steinhauser, M., & Kremer, K. (2002). Corrections to scaling in the hydrodynamic properties of dilute polymer solutions. The Journal of Chemical Physics, 117(2), 914–924. http://doi.org/10.1021/ma001499k

N	Sequence
20	DSNRPERCRGGAGVKIKMAR
30	RQQVRGPLYHLESSAPRVARAESSAAAAEV
40	YNGQLVTQAGAGINGGDDLVPAPKPPQKSRIEGQQIIQNP
80	VQSRYYEGKAYRHNANKMPSLIIVLEGPKVTDEILGAQILNKIANSSEQVKYTTTMSIVGVYDANVRRNLKPIVSPAEDE
100	ENEPNKAAPLEQSQAESEPIHQIDVSWGDKPSSAEPVSRQTTVASTVSRPGNPEPVRWQYCLGTLTAPDLELRKLHPEKSSHGP HPVMQYEHDTSSSVLF
200	DEKSGYSDLDMGVQSAKVIQTPETADAESGEMFPFLKNATHAELGHAEVPRTISDHSEFEARDNTQDVSVRGILEDFSVDNPSR GVKSEWENEKGYYVSFSFVFPDGDPLVKKTKVPVLAKGYKPETGVQDGNIELSGVGAGEASLEGLEDEETSNVMTKDSPIEYES ISPRRPATTHKGGYTVGGENRAQRELETAEIS
300	SMEKLAEDGIINDPHALSQPKIATKRGRGIHDEGDLLVLADEYIAEVKQKRDAVLSDQASPNSKTDPGESGPSALPPAKGEPSG RSTVQSGQGMQHMARETQQAMRVIRKKRGGKAKSDPKNCRDRANEAPLTRKVVQVDSMSPLSCDAEKDQLGQTGDTKAGKNSGP PRGSVEKYSSDTFRKSVAGNVVTAKNADKMPLQEATLNRSAQRSVNTSMFDLQSGVATRRQILEDSPDGHEGDQPRMRVILAIL GSGQENTLAPSLRKFACKVVQQAFSPTEKEDPLVGHTHDPGLEAYIES
400	RQAQSVRWGFQGKSLHSSMWSRLPNSGTGHVRPSRQLPPADGTAGMEEKPLYSGDPVEEHPLQTQDYGVGRIAREANSQQEYNT LTRQGEEDDELNGMKQVAAVDSRPSIGAQAPDGIKIDQRQIKDEKSVGPEKTDPGPVQSGKYSGGEFLGSKLKSLPDTYHLDKP EETNSKEKTVRGFAGSVPADTYRKSAPQHESIMPFHTVPASETKEEEGGMGCRHVEADNKAAGLEPELTAFAEPRGVSDKVTTE AAPNLNPSNSGGDDKYCKKMVAASSSWGQPPFGPNLTVALYSSQENGPPTRSDSKAVKDDLQETKEQAKIIYSFLEAEYGKSKR ESQTNGASKFDLLDDNDVAGGGPPEESELKVFHAFEESEDPTRLLSIDDAPGLQLFGAANPQDN
450	RSYDDANPSQAKDKPMYTPLSGLKWVSGQHSKVQISIPLNKDIEQYASGPAAPHWDFTQDVGGRKQLSGAIYVQMGLHGEGETR VNEPPVQRPALSLKNVAKTTCGEFASGAESLTVGAYSESADEELEVVKYVKRKIGSLPLVRARADVEGGVDLYRLEALEEPPPQ KAKPEKAADRIEKDSIEGRENLEPVNLDLLVEDQATNQENEEAQEPLSGPLESQPVLNPGKPINMDPVERLGAHPDLEAMCASE ELGGGEDEGGTTKGVDETEKFMSDSDGHRKENKKMEHPPERGQSLAVTQDISYGEPSLSNVSQLESRVEEGIAEGPRAGRSRDM ESPKALQLTAEQVVYGQDASFDAGLSNIVQGVNEGHTDGLYAAKRTTKILPDPQVEQAYSFSFAIQQDEAFDALNEILMGAAHN IFHLHVPEESKSKGRPLEHDESTGMSQGGK

Table S1. Sequences of scrambled peptides with an IDP-like amino acid composition.

Name (N)	Sequence
A-beta (40)	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
SBD (61)	GSMMSASSQSPNPNNPAEYCSTIPPLEYCSTIPPLQQAQASGALSSPPPTVMVPVGVLKHP
CTL9-I98A (92)	AAEELANAKKLKEQLEKLTVTIPAKAGEGGRLFGSITSKQAAESLQAQHGLKLDKRKIELADAIRALGYTNVPVKL HPEVTATLKVHVTEQK
Hdm2-ADB (95)	SSSSESTGTPSNPDLDAGVSEHSGDWLDQDSVSDQFSVEFEVESLDSEDYSLSEEGQELSDEDDEVYQVTVYQAGE SDTDSFEEDPEISLADYWK
Sml (104)	MQNSQDYFYAQNRCQQQQAPSTLRTVTMAEFRRVPLPPMAEVPMLSTQNSMGSSASASASSLEMWEKDLEERLNSI DHDMNNNKFGSGELKSMFNQGKVEEMDF
Prothymosin alpha (110)	MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNANEENGEQEADNEVDEEEEEGGEEEEEEEEGDGEEEDG DEDEEAESATGKRAAEDDEDDDVDTKKQKTDEDD
TC1 (112)	HHHHHHMKAKRSHQAIIMSTSLRVSPSIHGYHFDTASRKKAVGNIFENTDQESLERLFRNSGDKKAEERAKIIFAI DQDVEEKTRALMALKKRTKDKLFQFLKLRKYSIKVH
Alpha synuclein (140)	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVTA VAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
CFTR R region (189)	GAMESAERRNSILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGEKRKNSILNPINSIRKFSIVQKTPLQMNGIE EDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVLNLMTHSVNQGQNIHRKTTASTRKVSLAPQAN LTELDIYSRRLSQETGLEISEEINEEDLKECLFDDME
Tau K45 (198)	MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIGSTENLKHQPGG GKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKD RVQSKIGSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVY
RYBP (234)	HHHHHHMTMGDKKSPTRPKRQAKPAADEGFWDCSVCTFRNSAEAFKCSICDVRKGTSTRKPRINSQLVAQQVAQQY ATPPPPKKEKKEKVEKQDKEKPEKDKEISPSVTKKNTNKKTKPKSDILKDPPSEANSIQSANATTKTSETNHTSRP RLKNVDRSTAQQLAVTVGNVTVIITDFKEKTRSSSTSSSTVTSSAGSEQQNQSSSGSESTDKGSSRSSTPKGDMSA VNDESF
3D7 6H MSP2 (237)	MIKNESKYSNTFINNAYNMSIRRSMAESKPSTGAGGSAGGSAGGSAGGSAGGSAGGSAGSGDGNGADAEGSSSTPA TTTTTKTTTTTTTTNDAEASTSTSSENPNHKNAETNPKGKGEVQEPNQANKETQNNSNVQQDSQTKSNVPPTQDAD TKSPTAQPEQAENSAPTAEQTESPELQSAPENKGTGQHGHMHGSRNNHPQNTSDSQKECTDGNKENCGAATSLLNN SSNHHHHHH

Table S2. Sequences of IDPs used to generate conformational ensembles.

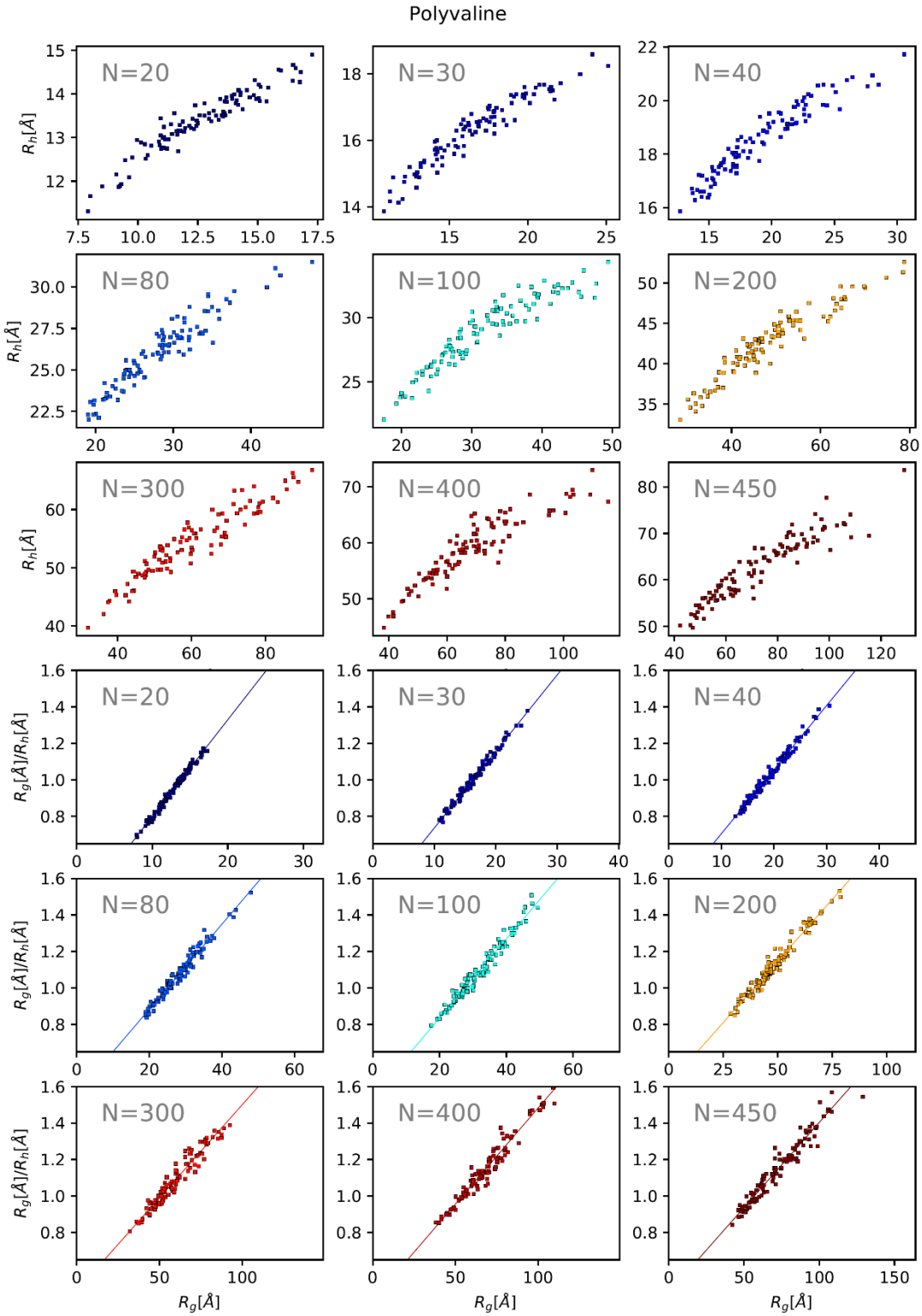

Figure S1. Details of the level of expansion as quantified by R_g and R_h for the individual peptides. The nine upper panels show R_h and the nine lower panels show the R_g/R_h ratio, in each case plotted against R_g. This figure shows the data for the poly-valine peptides, and each subpanel corresponds to a specific chain length.

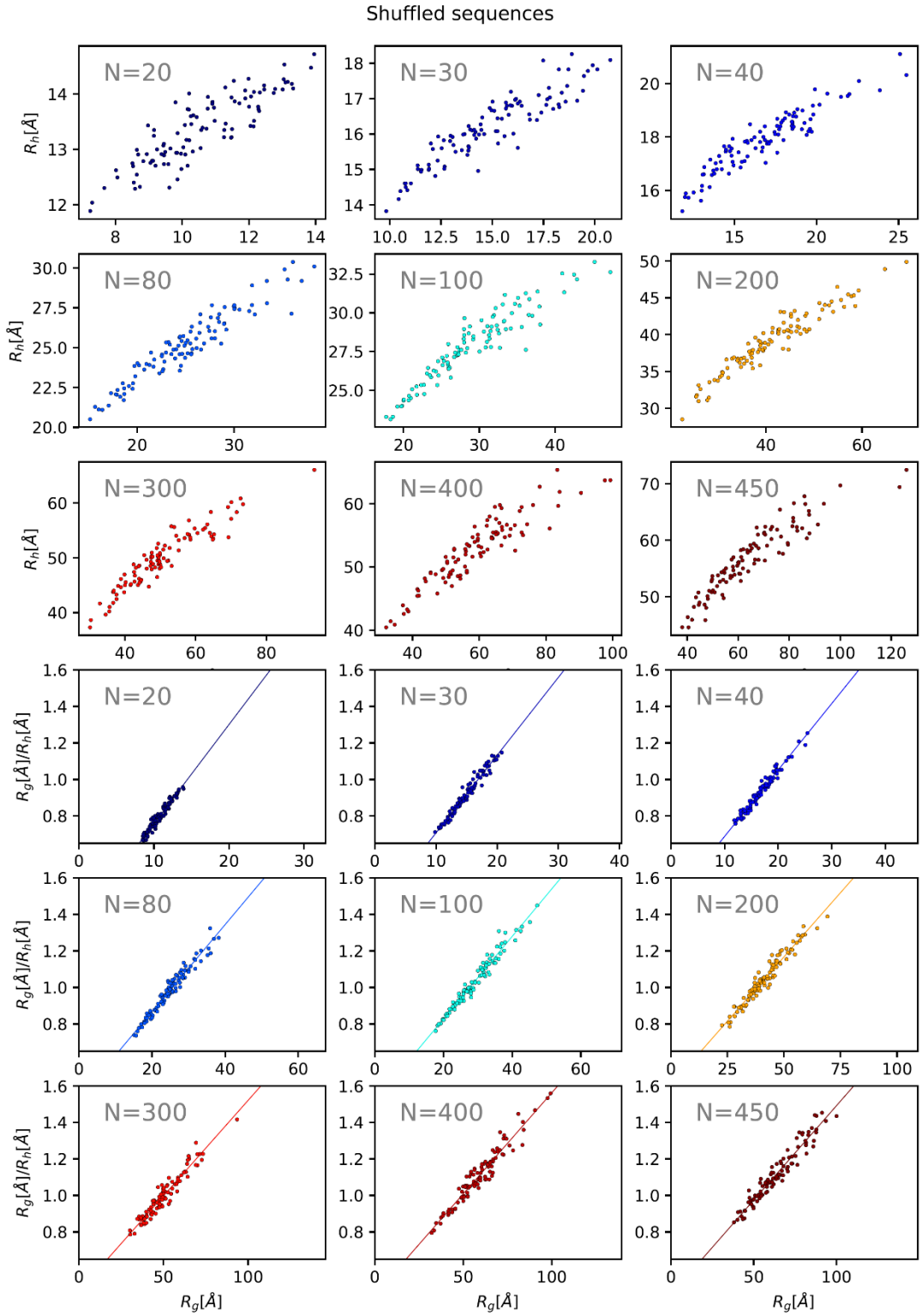

Figure S2. Details of the level of expansion as quantified by R_g and R_h for the individual peptides. The nine upper panels show R_h and the nine lower panels show the R_g/R_h ratio, in each case plotted against R_g. This figure shows the data for the peptides with IDP-like sequences, and each subpanel corresponds to a specific chain length.

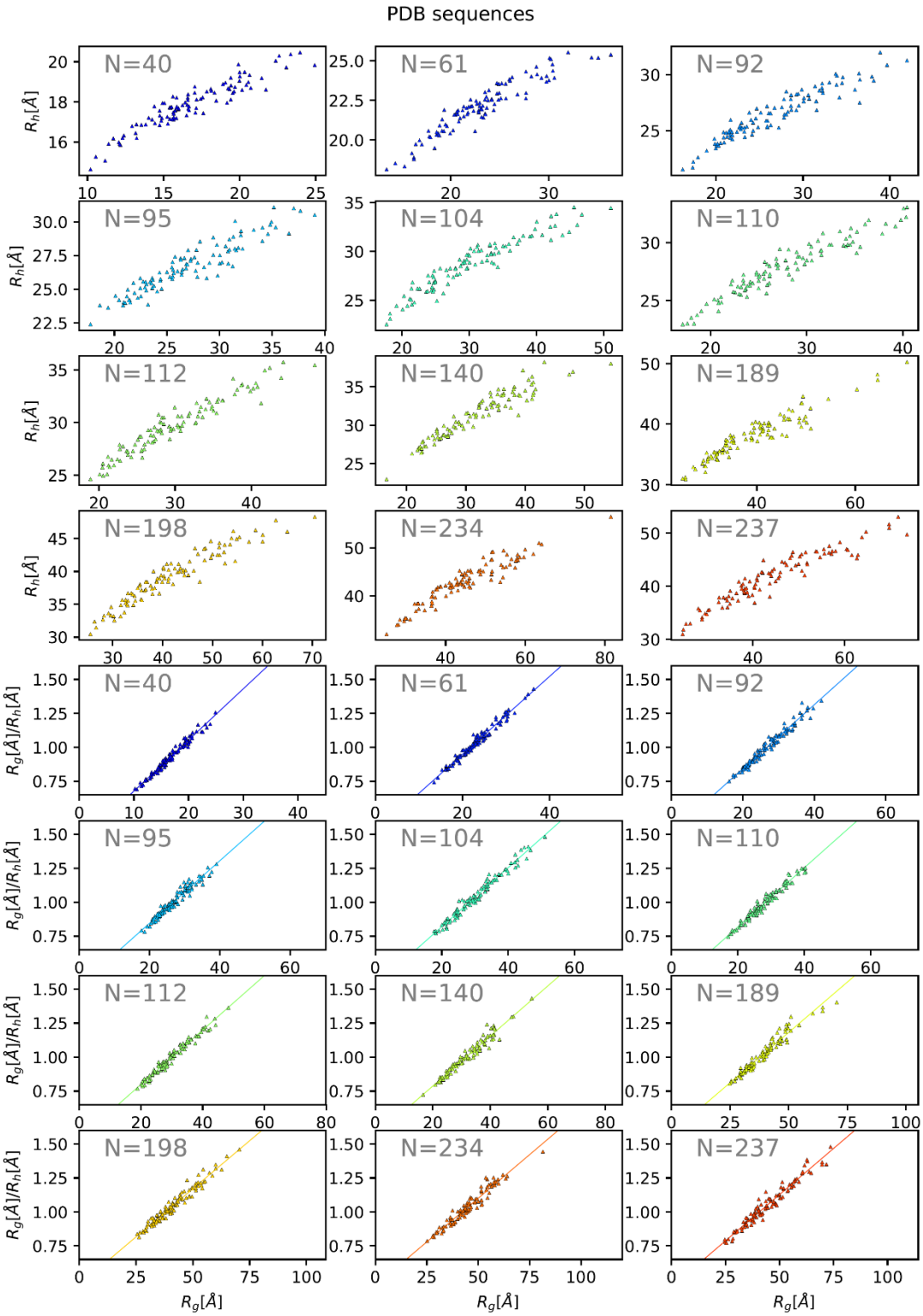

Figure S3. Details of the level of expansion as quantified by R_g and R_h for the individual peptides. The nine upper panels show R_h and the nine lower panels show the R_g/R_h ratio, in each case plotted against R_g. This figure shows the data for IDPs, and each subpanel corresponds to a specific chain length.

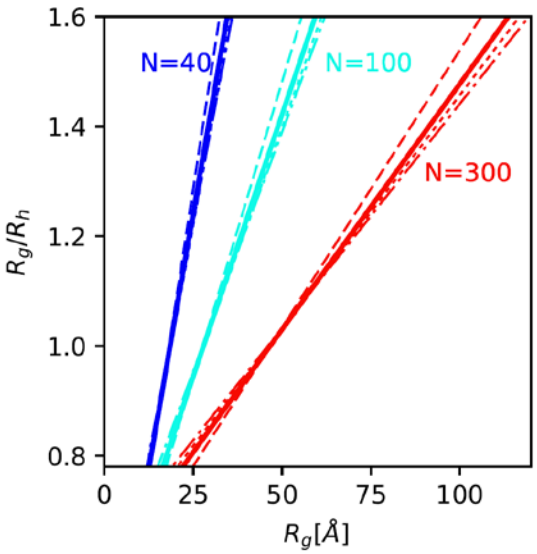

Figure S4. Evaluating the effect of the input sequences on the resulting model. We repeated the fitting described in the main text using each of the three peptide sets individually. As a method for comparison, the figure shows the resulting relationship obtained from these fits and compares it to the fit obtained using the full data. We show the results from three representative chain lengths *N*=40 (blue), *N*=100 (cyan), and N=300 (red). For each chain length, the four lines correspond to the final model obtained using either the full data (full line), only the poly-valine data (dash-dotted line), peptides with IDP-like sequences (dots) and IDPs (dashes). Overall, the results show that the fits are very robust to the input data used. The biggest discrepancies are observed for the fit to the IDP sequences only and for the longest chain length (*N*=300), though this is likely explained by the fact that the longest protein in this data set is only 237 residues long, thus under-restraining the fit at longer chain lengths.

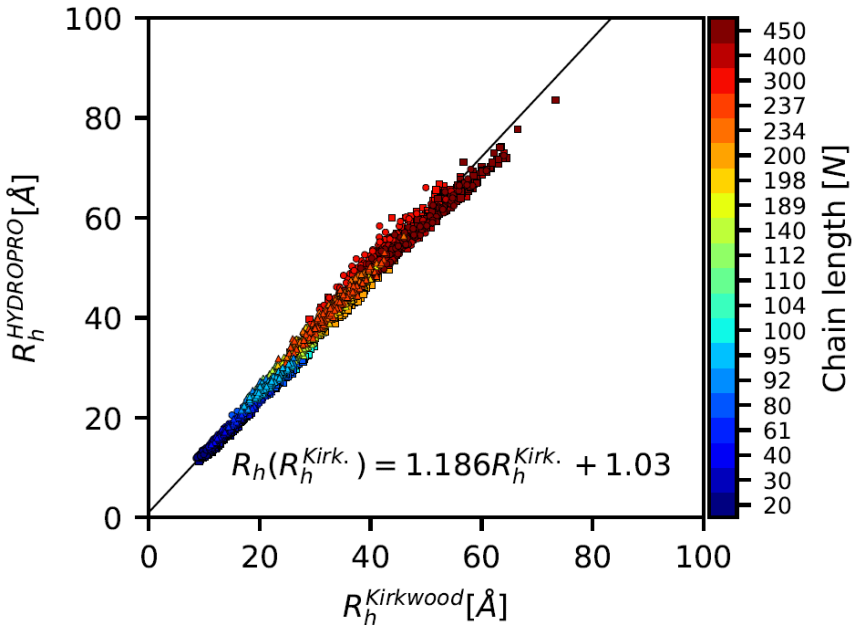

Figure S5. We compared the value of R_h calculated using the standard Kirkwood formula (main text Eq. 2) with the results obtained using HYDROPRO. As expected the values of R_h obtained using only the pairwise distances between protein atoms are smaller than those obtained using the full HYDROPRO calculations. Nevertheless, the two are strongly correlated, suggesting that it is also possible to estimate R_h using Eq. 2 and the linear fit. The colours correspond to the chain length (see right bar).

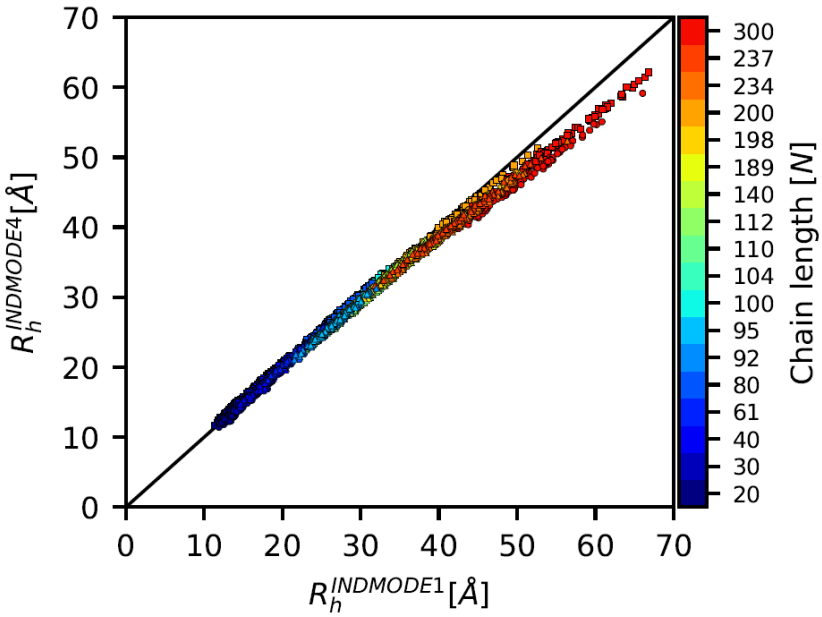

Figure S6. To test for any potential systematic errors caused by the two models used by HYDROPRO, we calculated R_h with the course grained INDMODE 4 for all peptides with N≤300 and compared them to the datasets with the R_h calculated using the finer grained INDMODE 1. For peptides up to length ~200 residues the two methods give very similar results (line corresponds to the diagonal), but for the longest peptides the coarser-grained model underestimates slightly the value of R_h compared to the atom-based model.

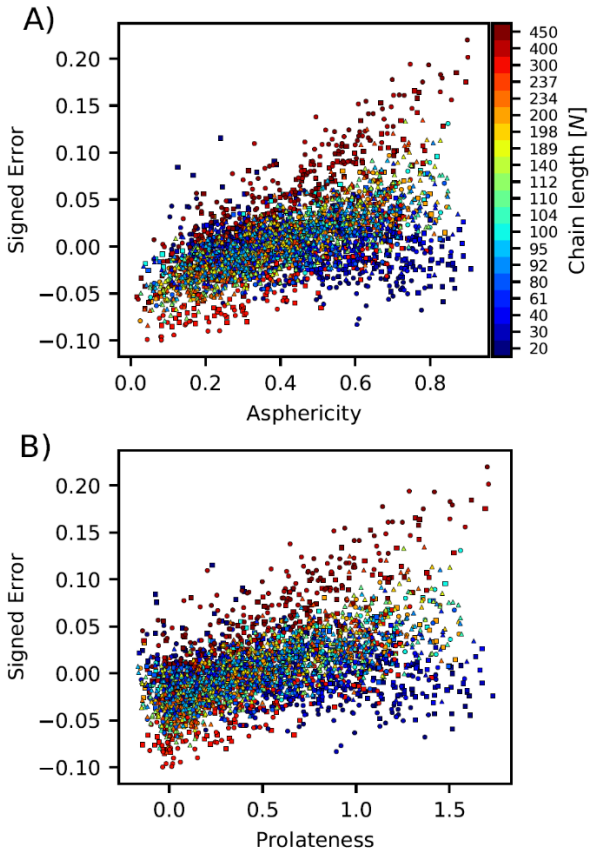

Figure S7. We investigated to what extent the prediction error depends on the shape of the conformers, and thus calculated the asphericity and prolateness of each conformer. The calculated asphericity (A) and prolateness (B) was plotted against the signed error (difference between prediction based on Eq. 7 and the value obtained by HYDROPRO). We found a weak correlation between the error (r^2 ~0.30 and r^2 ~0.28 in panels A and B, respectively). Asphericity values greater than 0 gives an indication of the anisotropy of the molecule. Negative values of prolateness correspond to oblate shapes whereas positive values correspond to prolate shapes. For perfect spheres both prolateness and asphericity is 0.

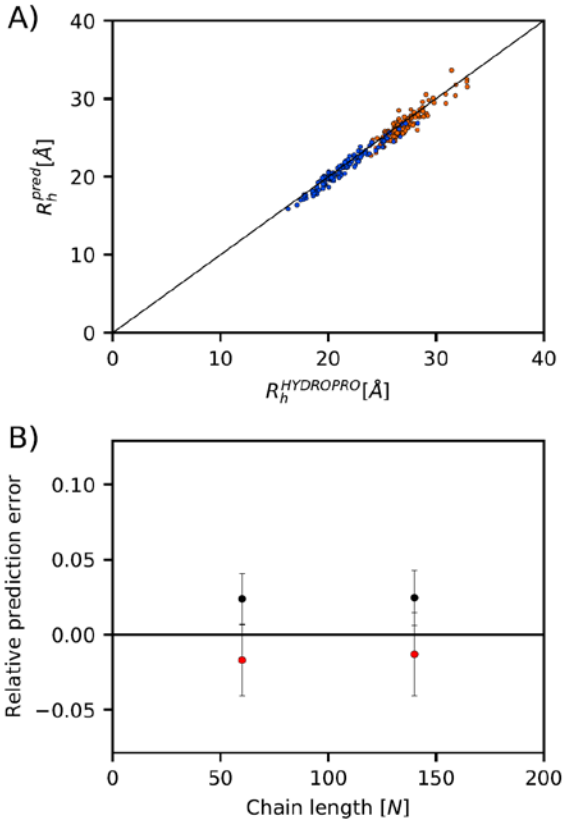

Figure S8. We examined whether the model that we derived also provides accurate results for conformations generated e.g. by all-atom molecular dynamics (MD) simulations. As described in the main text we thus calculated R_h using Eq. 7 and compared the results to those obtained using HYDROPRO for two sets of conformations generated by MD. In particular, we performed R_h calculations for ~100 conformations of a domain from the HIV1-integrase (N=60) and of α-synuclein (N=140). A: We find a strong correlation between the values predicted from Eq. 7 and those obtained directly by HYDROPRO for both the domain from HIV1-integrase (blue) and α-synuclein (orange). B: We also calculated the mean unsigned (black) and signed (red) error for the two proteins, and found values comparable to those obtained from the Flexible-Meccano structures.