Supporting Information: Pseudo-Improper-Dihedral Model for Intrinsically Disordered Proteins

Łukasz Mioduszewski[1], Bartosz Różycki[1], Marek Cieplak[1]

[1]Institute of Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland

Contents

1 Cosine function approximation

We replaced the cosine function by its algebraic approximation to speed the computations up. The approximation is [1]:

$$0.5 \cdot \cos \upsilon + 0.5 \approx \frac{(\upsilon/\pi)^2 - 2 \cdot |\upsilon/\pi| + 1}{2 \cdot (\upsilon/\pi)^2 - 2 \cdot |\upsilon/\pi| + 1} \tag{1}$$

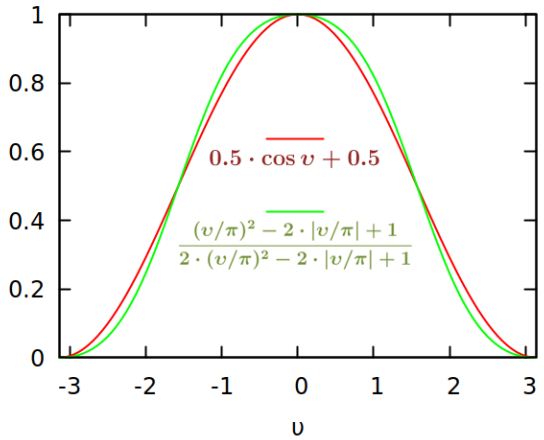


Figure S1: Comparison of the cosine function and its algebraic approximation.

2 Distance and PID angle distributions

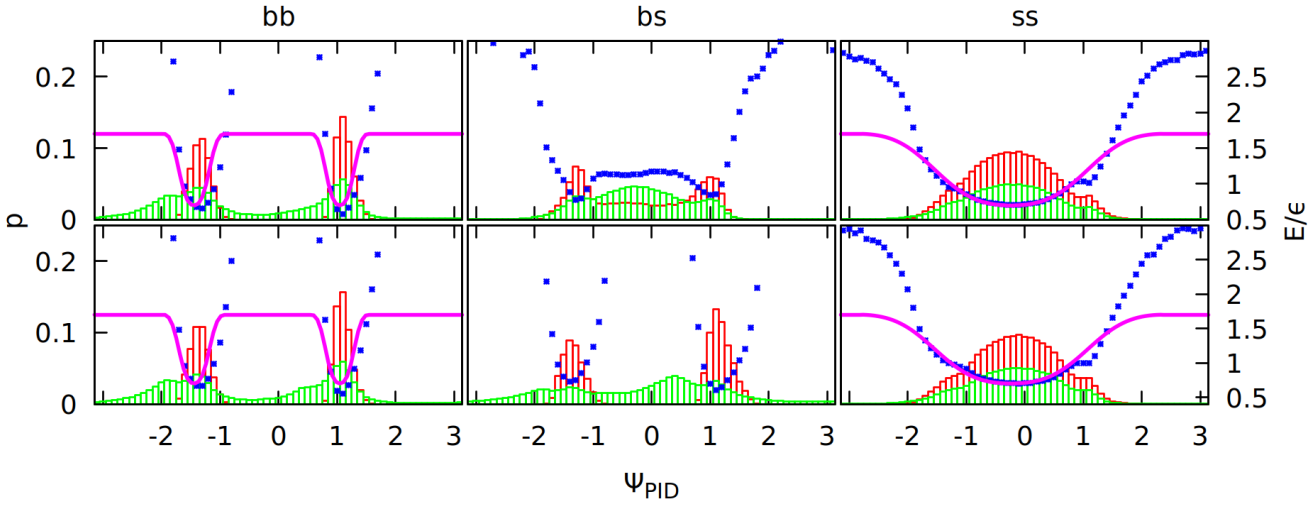
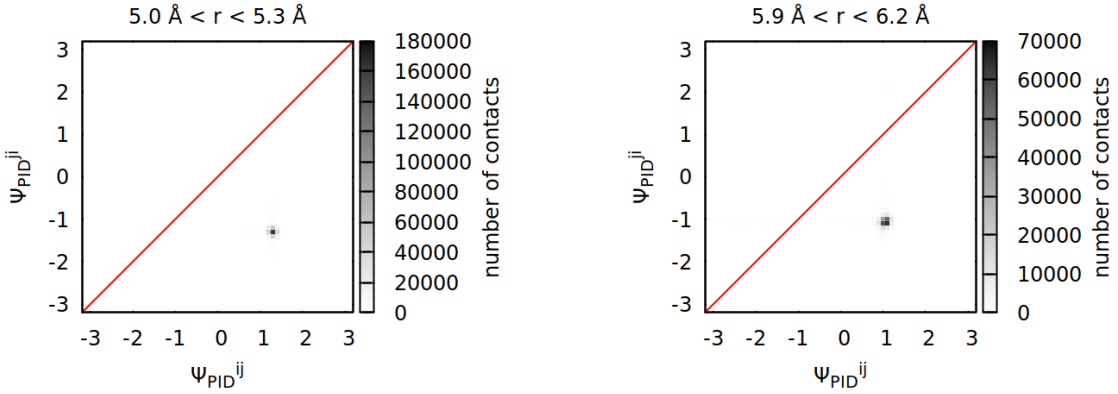


Figure S2: Distributions of the PID angles (in radians) in the contacts from the PDB survey that include a given type (green histograms). Local $i, i+3$ and $i, i+4$ contacts are excluded. Each contact has two angles. Distribution of the first is on the top panels, of the second on the bottom panels. Subdistributions made from contacts that obey the directional criteria defined in [2] are shown as red histograms. The potential resulting from Boltzmann Inversion procedure (blue dots, unit of energy $\epsilon \approx 1.5$ kcal/mol) was fitted to an analytical function (purple line).

2.1 Right-handedness

Distributions of PID angles are different for the first and second angle for $i, i+3$ and $i, i+4$ contacts due to the right-handedness of most of α helices (and in our statistics the first PID angle is for the residue earlier in the sequence). This is visible in Fig. S3. Distributions for $i, i+5$ and more nonlocal contacts are mostly symmetric for the operation of exchanging the first and the second PID angles in a contact, with the exception of backbone-sidechain cases (see the subsection about backbone-sidechain contacts).



$i, i+3$ contacts $\qquad$ $i, i+4$ contacts

Figure S3: Two-dimensional distribution of backbone-backbone contacts, where the first PID angle in a contact is on one axis and the second PID angle (counting from the N to C terminal) is on the other axis. The assymetry probably comes from the right-handedness of an α-helix. The distribution is taken for C_α-C_α distances r in a 0.3 Å window to avoid noise from other distances. Note that for those distances the distribution is extremely narrow (each dot has size 0.1 rad x 0.1 rad).

2.2 Distance distributions

r^{ss}_{min}	Gln	Cys	Ala	Ser	Val	Thr	Ile	Leu	Asn	Asp	Lys	Glu	Met	His	Phe	Arg	Tyr	Trp
Gln	8.63																	
Cys	7.72	7.56																
Ala	7.39	6.97	6.42															
Ser	7.64	6.97	6.53	6.65														
Val	7.81	7.56	7.06	7.17	7.65													
Thr	7.77	7.40	6.94	6.97	7.54	7.30												
Ile	8.24	7.95	7.45	7.52	8.06	7.93	8.53											
Leu	8.44	8.07	7.65	7.68	8.29	8.12	8.77	8.93										
Asn	8.19	7.49	7.02	7.18	7.54	7.46	7.96	8.14	7.74									
Asp	8.15	7.18	6.73	6.99	7.22	7.19	7.65	7.86	7.50									
Lys	8.69	7.83	7.26	7.73	7.69	7.79	8.16	8.39	8.11	8.59								
Glu	8.41	7.45	7.04	7.41	7.50	7.51	7.97	8.20	8.00		8.90							
Met	8.84	8.29	7.91	7.94	8.48	8.33	8.95	9.14	8.49	8.15	8.80	8.61	9.29					
His	8.64	8.17	7.50	7.88	7.92	7.98	8.37	8.57	8.36	8.50	8.58	8.84	8.93	8.83				
Phe	8.95	8.50	8.17	8.24	8.69	8.58	9.11	9.34	8.65	8.51	8.79	8.75	9.55	8.98	9.73			
Arg	9.26	8.24	7.99	8.27	8.31	8.50	8.76	8.98	8.87	9.12		9.52	9.27	9.23	9.26			
Tyr	9.27	8.26	8.02	8.36	8.39	8.58	8.78	9.02	8.96	9.35	9.04	9.48	9.28	9.38	9.56	9.51	9.34	
Trp	9.58	8.95	8.65	8.75	9.22	9.14	9.57	9.79	9.11	9.10	9.21	9.48	10.02	9.66	10.17	9.82	10.08	10.85

Table S1: Average distances, in Å, for the ss contacts as derived from the CATH database, first defined in [2]. Empty boxes indicate same-charged residues that cannot form ss contacts (as well as GLY and PRO). $r^{bb+}_{min} = 5.6$ Å, $r^{bb-}_{min} = 6.2$ Å.

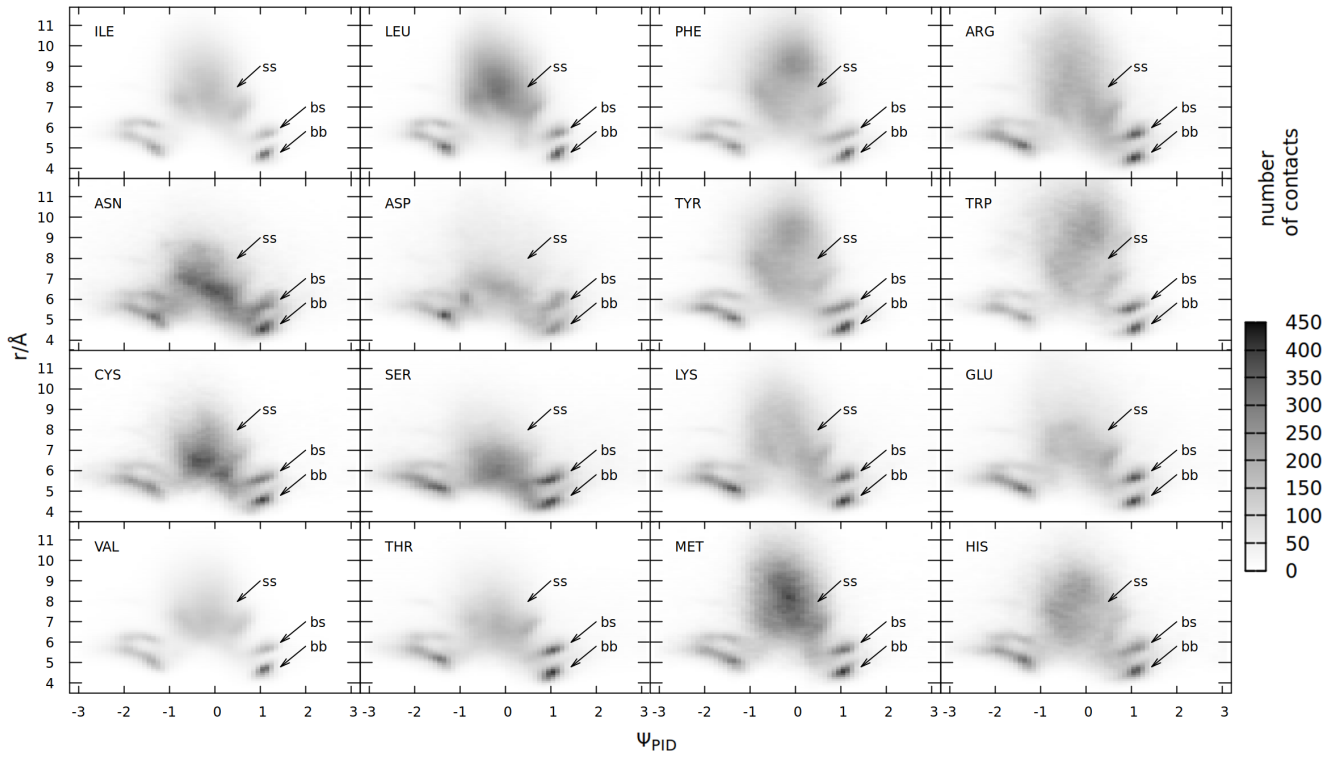
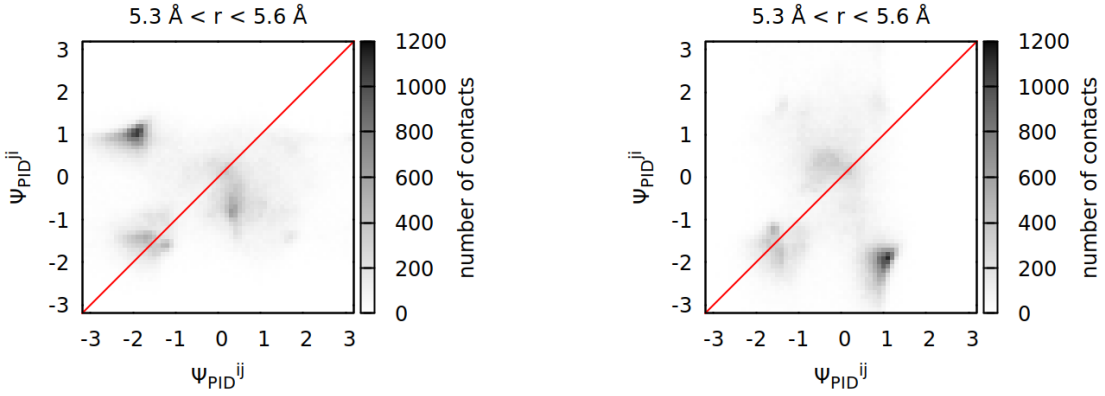


Figure S4: Two-dimensional distributions of contacts, where PID angle (in radians) for a given amino acid is on one axis and C_α-C_α distance is on the other axis. $i, i+3$ and $i, i+4$ contacts are excluded.

2.3 Backbone-sidechain contacts

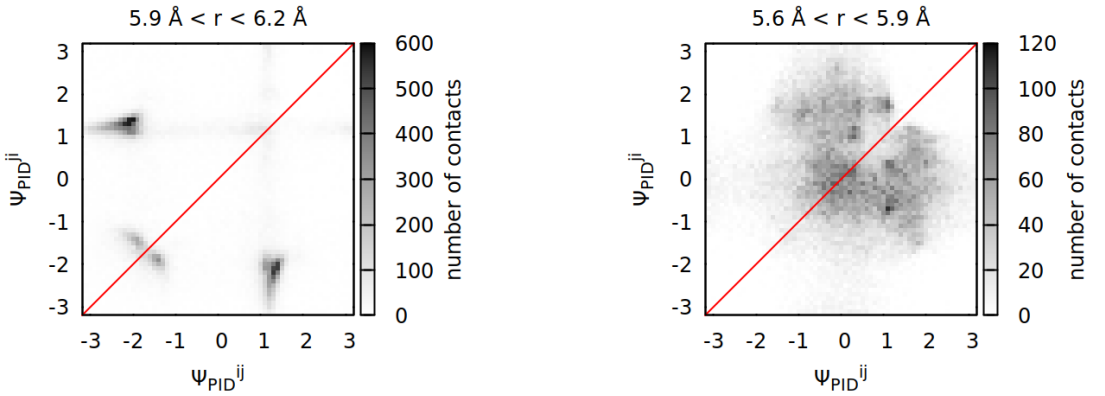
Backbone-sidechain contacts have very broad distributions of PID angles and C_α-C_α distances. We tried to distinguish the case where the first residue is the backbone and the second is the sidechain (bs) from the opposite case (sb) in the contact distributions that can be of more than one type. The main difference occurs for C_α-C_α distance 5.3 Å $< r <$ 5.6 Å, where a peak seems to be associated with a given type of contact (Fig. S5). However, a closer analysis of structures in VMD revealed that this peak corresponds to backbone-backbone contacts, and the sidechain is just a steric hindrance that makes only one combination of PID angles possible. This was confirmed by plotting distributions of contacts belonging to only one overlap type: distribution containing only bb contacts had both peaks (left part of Fig. S6). Most of the bs contacts are also of the bb type (see Table 1 in the main article), so we decided to indicate those mixed bs contacts by arrows in Fig. S4 and in Fig. 3 in the main article.

The only bs (containing also sb) distribution (right part of Fig. S6) was very broad and showed a combination of two cases: $\Psi^{ij}_{PID} \approx 0$ rad and $\Psi^{ji}_{PID} \approx 1.5$ rad or vice versa: $\Psi^{ji}_{PID} \approx 0$ rad and $\Psi^{ij}_{PID} \approx 1$ rad. This is consistent with the results in Fig. S2, as one residue donors a backbone, and the second residue donors a sidechain. These distributions are, however, very broad and range for distances 5.5 Å$<$ r $<$ 7 Å (and become even broader for larger distances, see Fig. S7).



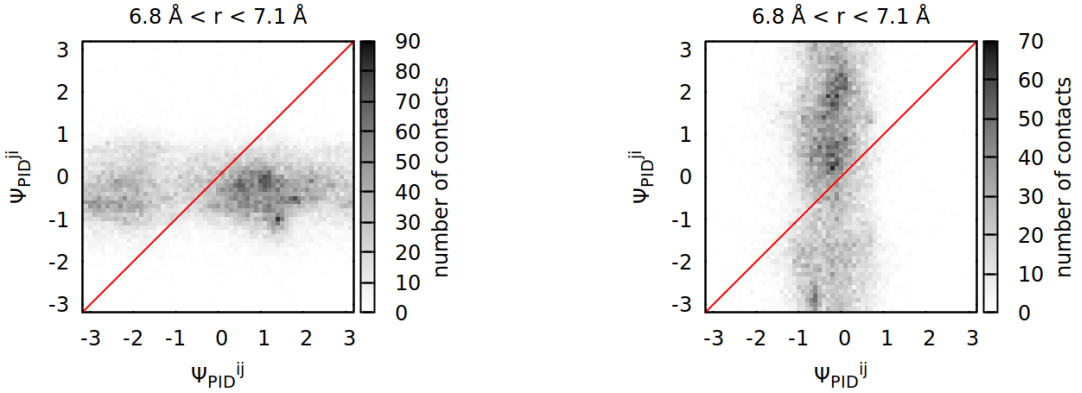
backbone-sidechain (bs) sidechain-backbone (sb)

Figure S5: Two-dimensional distribution of contacts, where the first PID angle (in radians) in a contact is on one axis and the second angle is on the other axis. Those contacts could also include other types of overlaps (bb or ss). C_α-C_α distance r is within the range given on top of the graphs.



only backbone-backbone (bb) only sidechain with backbone (bs and sb)

Figure S6: Two-dimensional distribution of contacts, where the first PID angle (in radians) in a contact is on one axis and the second angle is on the other axis. Contacts that contain only bb overlaps (left panel) or only bs and/or sb overlaps (right) were included. C_α-C_α distance r is within the range given on top of the graphs.



only backbone-sidechain (bs) only sidechain-backbone (sb)

Figure S7: Two-dimensional distribution of contacts, where the first PID angle (in radians) in a contact is on one axis and the second angle is on the other axis. Contacts that contain only bs overlaps (left panel) or only sb overlaps (right) were included. C_α-C_α distance r is within the range given on top of the graphs. Note that the number of contacts is smaller than in all other distributions.

2.4 Statistics without dividing into bb, bs and ss types

The bs contacts were hard to quantify, so total distributions were also plotted (Fig. S8). It turns out that bb and ss contacts have pretty distinguishable peaks even without using the information about the contact type from the overlaps. The top panels of Fig. S8 show that bb contacts have two peaks corresponding to $\psi_0^{bb+} = 1.05$ rad and $\psi_0^{bb-} = -1.44$ rad, while ss contacts correspond to one broader $\psi_0^{ss} = -0.23$ rad peak (shown on the bottom panel). Those are the same peaks as in Fig. S2.

It is interesting to note that $\psi_{PID} \approx +1$ rad is common for smaller C_α-C_α distance than those for $\psi_{PID} \approx -1$ rad, which is reflected in our potential ($r_{min}^{bb+} = r_{min}^{bb-} - 0.6$ Å).

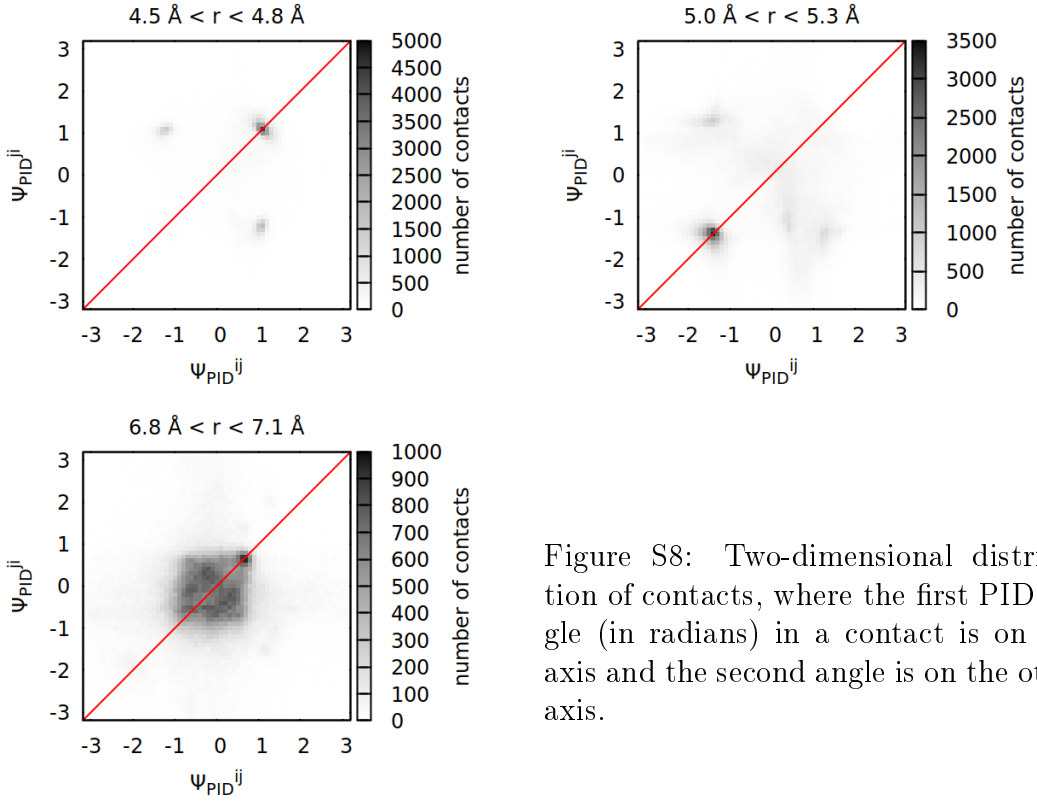


Figure S8: Two-dimensional distribution of contacts, where the first PID angle (in radians) in a contact is on one axis and the second angle is on the other axis.

3 The parameterization

3.1 Proteins used for the parameterization

We used a set of 23 intrinsically disordered proteins with different lengths, occupying different areas of the Das-Pappu diagram (see Fig. S9) and the Uversky plot (see Fig. S10) [6]. Full Das-Pappu diagram S11 shows the whole state space with unused regions (proteins usually do not have that many charged residues [7]).

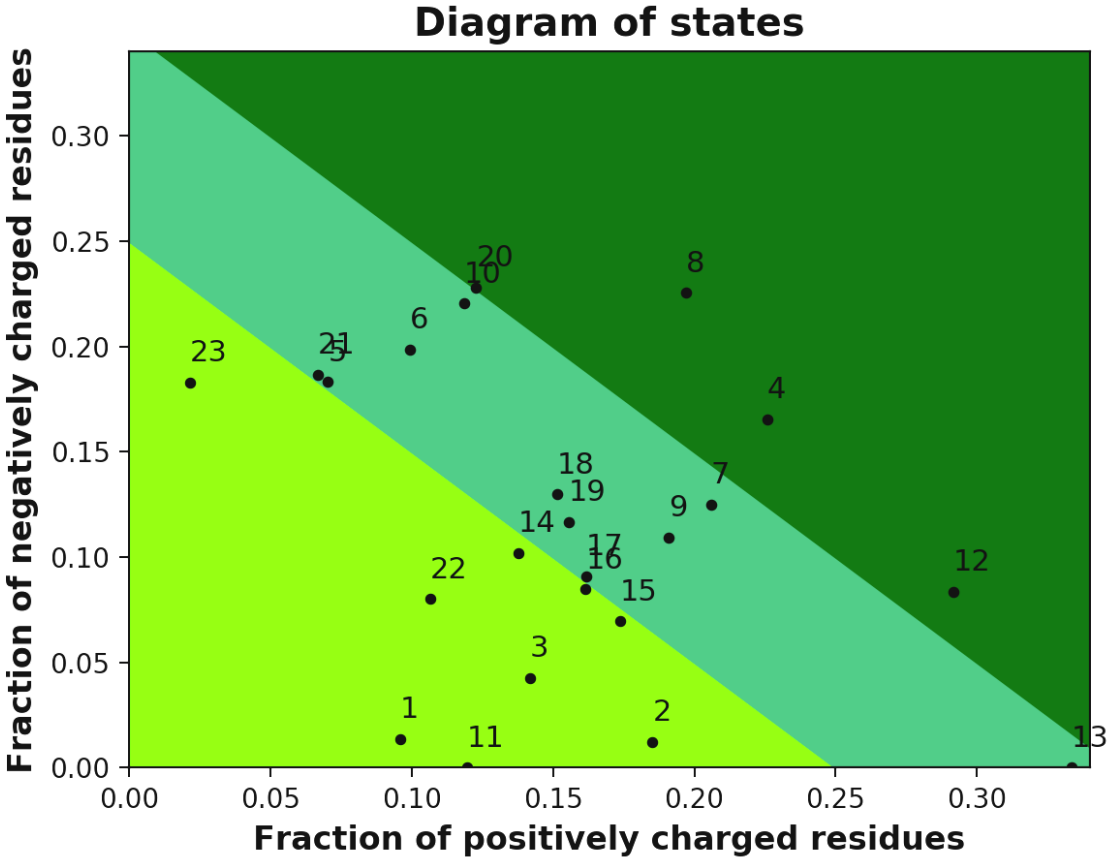


Figure S9: Das-Pappu diagram of the 23 IDPs used for the parameterization, labeled by their number in Table S2.

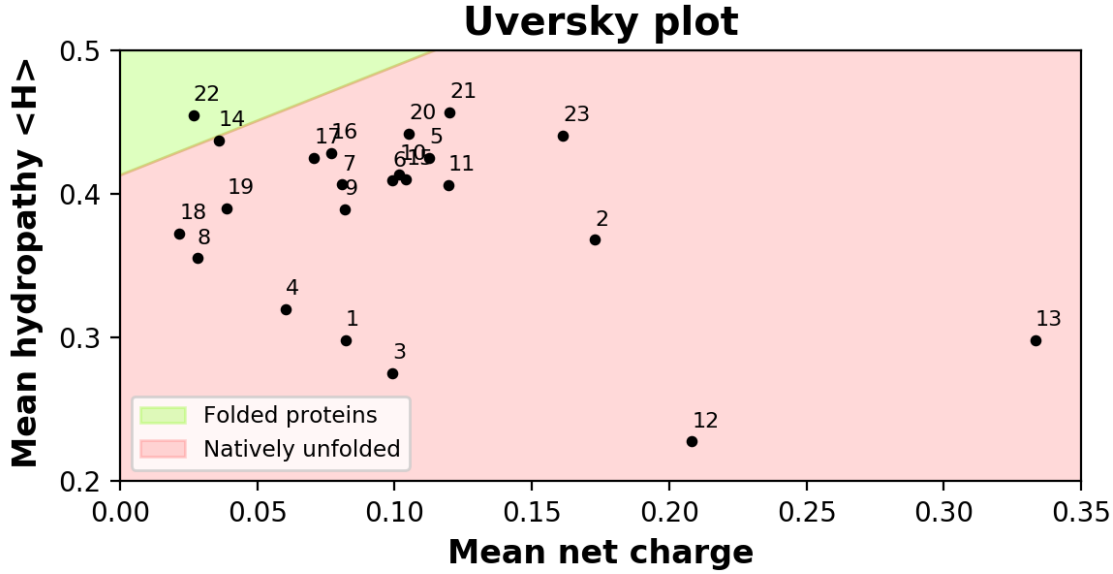


Figure S10: Uversky plot of the 23 IDPs used for the parameterization, labeled by their number in Table S2. The vertical axis uses Kyte-Doolittle hydropathy score [8] rescaled to 0-1 range.

It is interesting to note that the value of the screening length used significantly affects the results. Using uniform $s = 10$ Å resulted in worse agreement with experiment (we were unable to find a model better than Gaussian chain with $b = 6.7$ Å, results not shown).

nr	id	n	R_g/Å	s/Å	source
1	IB5	73	27.9 ± 1.0	13.6	[9]
2	Ash1	81	28.4 ± 3.4	7.9	[9]
3	I11ng	141	41.1 ± 1.0	13.6	[9]
4	RNaseE	248	52.6 ± 0.3	7.85	[9]
5	ACTR	71	25.1 ± 1.3	6.8	[10]
6	NHE1	131	36.3 ± 1.8	6.8	[10]
7	sNase	136	21.2 ± 1.0	23.3	[10]
8	5AAA	142	22.15 ± 0.87	4.3	[11]
9	6AAA	110	28.1 ± 0.1	7.7	[11]
10	8AAC	59	14.6 ± 0.5	13.6	[11]
11	9AAA	92	29.9 ± 0.3	7.55	[11]
12	his5	24	13.6 ± 0.2	7.85	[12]
13	RS	24	12.62 ± 0.07	6.7	[13]
14	tauK10	168	40.0 ± 1.0	7.4	[14, 19]
15	tauK17	144	36.0 ± 2.0	7.4	[14, 19]
16	tauK18	130	38.0 ± 3.0	7.4	[14, 19]
17	tauK19	99	35.0 ± 1.0	7.4	[14, 19]
18	tauK25	185	41.0 ± 2.0	7.4	[14, 19]
19	tauK44	283	52.0 ± 2.0	7.4	[14, 19]
20	RNF4	57	25.8 ± 3.9	8.5	[15, 19]
21	NRG1	75	26.8 ± 1.1	7.4	[16, 19]
22	PIR	75	26.5 ± 0.5	7.8	[17, 19]
23	p53	93	28.7 ± 0.3	6.6	[18, 19]

Table S2: Properties of the 23 IDPs used for the parameterization: id used for identification, number of residues n, experimental R_g, screening length s and source of information.

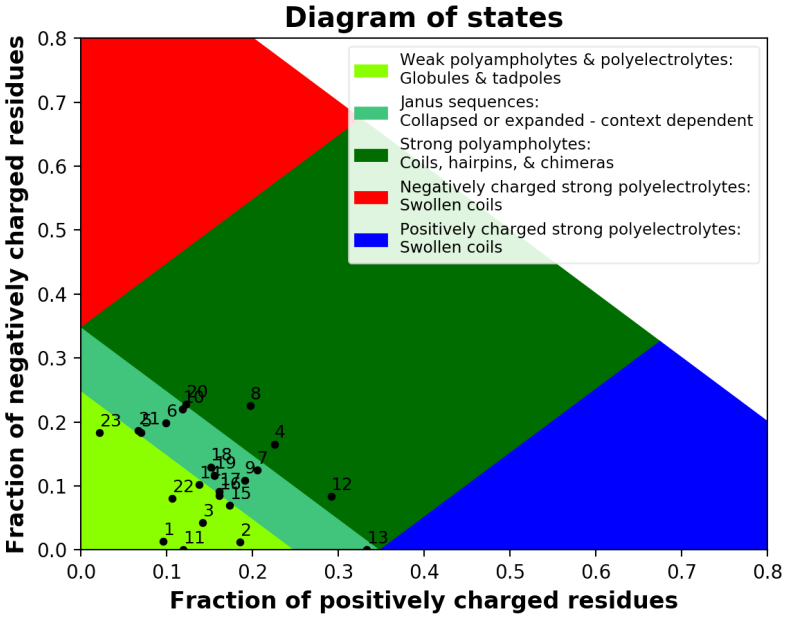


Figure S11: Das-Pappu diagram of the 23 IDPs, labeled by their number in Table S2. The shape of Janus sequences depends on factors like the ionic force. In our simulations we always assumed the same ionic force as in SAXS experiments.

IB5	SARSPPGKPQGPPQQEGNKPQGPPPPGKPQGPPPAGGNPQQPQAPPAGKP QGPPPPPQGGRPPRPAQGQQPPQ
Ash1	GASASSSPSPSTPTKSGKMRSSSPVRPKAYTPSPRSPNYHRFALDSPPQS PRRSSNSSITKKGSRRSSGSSPTRHTTRVCV
II1ng	ISGKPVGRRPQGGNQPQRPPPPPGKPQGPPPQGGWQSQGPPPPPGKPEGR PPQGRNQSQGPPPHPGKPERPPPQGGSQGTPPPPGKPERPPPQGGNQSHR PPPPPGKPERPPPQGGNQSRGPPPHRGKPEGPPPQEGNKSR
RNaseE	ERQQDRRKPRQNNRRDRNERRDTRSERTEGSDNREENRRNRRQAQQQTAE TRESRQQAEKARIADTADEQQAPRRERSRRRNDDKRQAQQEAKALNVEEQ SVQETEQEERVRPVQPRRKQRQLNQKVRYEQSVAEEAVVAPVVEETVAAE PIVQEAPAPRTELVKVPLPVVAQTAPEQQEENNADNRDNGGMPRRSRRSP RHLRVSGQRRRRYRDCRYPIQSPMPLTVACASPELASGKVWIRVPIVR
ACTR	GTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEI DRALGIPELVNQGQALEPKQD
NHE1	MVPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVN EELKGKVLGLSRDPAKVAEEDEDDDGGIMMRSKETSSPGTDDVFTPAPSD SPSSQRIQRCLSDPGPHPEPGEGEPFFPKGQ
sNase	ATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLVDTPETKHPKKG VEKYGPEASAFTKKMVENAKKIEVEFDKGQRTDKYGRGLAYIYADGKMVN EALVRQGLAKVAYVYKPNNTHEQHLRKSEAQAKKEK
5AAA	MDYKDDDDKNRALSPMVSEFETIEQENSYNEWLRAKVATSLADPRPAIPH DEVERRMAERFAKMRKERSKQMDYKDDDDKNRALSPMVSEFETIEQENSY NEWLRAKVATSLADPRPAIPHDEVERRMAERFAKMRKERSKQ
6AAA	VRTKADSVPGTYRKVVAARAPRKVLGSSTSATNSTSVSSRKAENKYAGGN PVCVRPTPKWQKGIGEFFRLSPKDSEKENQIPEEAGSSGLGKAKRKACPL QPDHTNDEKE
8AAC	MEAIAKHDFSATADDELSFRKTQILKILNMEDDSNWTRAELDGKEGLIPS NYIEMKNHD
9AAA	GSMTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPST TKSFKNAPLLAPPNSNMGMTSPFNGLTSPQRSPFPKSSVKRT
his5	DSHAKRHHGYKRKFHEKHHSHRGY
RS	GAMGPSYGRSRSRSRSRSRSRSRS
tauK10	QTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGS LGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETH KLTFRENAKAKTDHGAEIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQ LATLADEVSASLAKQGL
tauK17	SSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAP VPMPDLKNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNI HHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE
tauK18	MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCG SKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSE KLDFKDRVQSKIGSLDNITHVPGGGNKKIE
tauK19	MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCG SLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE
tauK25	MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGI GDTPSLEDEAAGHVTQARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAA PPGQKGQANATRIPAKTPPAPKTPPSSGEPPKSGDRSGYSSPGSPGTPGS RSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRL
tauK44	MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGI GDTPSLEDEAAGHVTQARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAA PPGQKGQANATRIPAKTPPAPKTPPSSGEPPKSGDRSGYSSPGSPGTPGS RSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKS KIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEV KSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE
RNF4	GHMGSWEAEPIELVETAGDEIVDLTCESLEPVVVDLTHNDSVVIVDERRR PRRNARR
NRG1	MEIYSPDMSEVAAERSSSPSTQLSADPSLDGLPAAEDMPEPQTEDGRTPG LVGLAVPCCACLEAERLRGCLNSEK
PIR	QSVSPMRSVSENSLVAMDFSGQKTRVIDNPTEALSVAVEEGLAWRKKGCL RLGNHGSPTAPSQSSAVNMALHRSQ
p53	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDI EQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPL

Table S3: Sequences of the 23 IDPs used for the parameterization.

3.2 Model ranking

We considered 246 different versions of our model. All symbols used in their names are explained in the legend (Table S4). A few of them were not defined in the main text: we considered versions (indicated by letter W) where the PID angle potential was fitted to the Boltzmann inversion potential based on all overlap contacts (green histograms in Fig. 2 in the main text).

We also checked L-J potential with a bump (indicated by a subscript $_B$), designed to mimic the potential barrier resulting from a hydration shell made by water molecules surrounding the residues.

In charged residues the charge is located near the end of the sidechain, so we tried using the PID potential for the electrostatic interactions (denoted by letter D) to take account of that directionality.

The best fit to the experimental data for the Kuhn length of the Gaussian chain is $b = 6.7$ Å, but we also included other values of b for comparison.

In the Table S5 we list the models sorted by their Pearson coefficient P (see the main text). We also included $\chi^2 = \frac{1}{N}\sum_{p=1}^{N}\left(\frac{(R_g^{exp}-R_g^{sim})^2}{\sigma_{sim}^2+\sigma_{exp}^2}\right)_p$, where σ_{sim} is the uncertainty of the simulation result defined by the jacknife resampling method and σ_{exp} is the experimental uncertainty from Table S2. Uncertainties are much smaller than $R_g^{exp} - R_g^{sim}$, so the values of χ^2 are large (values bigger than 1000 are indicated as „>1000").

For each protein and each version of the model we made 20 trajectories. For the quasi-adiabatic models (letter A) the equilibration time was 75 000 τ and the total simulation time was 150 000 τ. For PID models (letters P and W) both times were 10x smaller (based on the times needed for R_g to stop depending on the initial conditions). This difference may indicate that the timescales in the PID model are longer than in the previous one, so Fig. 6 in the main text may underestimate the efficiency of the PID model.

P	PID potential (fit to red histograms in Fig. 2)
A	Quasi-adiabatic potential
W	PID potential, wide version (fit to green histograms in Fig. 2)
superscript $^+$	$i, i+4$ attractive contacts on
superscript $^-$	$i, i+4$ attractive contacts off
L	Standard Lenard-Jones potential
F	Lenard-Jones potential with flat region between r_{min}^{bb} and r_{min}^{ss} for ss contacts
L_B	potential defined by equation $\phi(r) = \epsilon^{LJ}\left[\left(\frac{r_{min}}{r}\right)^{12} - \frac{9}{2^{1/6}}\left(\frac{r_{min}}{r}\right)^{7} + \frac{13}{2}\left(\frac{r_{min}}{r}\right)^{6}\right]$
F_B	Same as L_B, but with flat region between r_{min}^{bb} and r_{min}^{ss} for ss contacts
ME	Matrix of interactions where each contact has the same amplitude (default 1 ϵ)
MJ	Miyazawa-Jernigan matrix of interactions [3]
MD	MDCG matrix of interactions [4]
Subscript $_{0-1}$	scaling factor for the matrix (from $_0$ to $_1$)
C	Classic Debye-Hueckel electrostatics with permittivity 80
T	Debye-Hueckel electrostatics with permittivity 4 Å/r [5]
R	Electrostatics only for residues with the same sign (oppositely charged residues interact via the ss contacts as uncharged residues)
D	Directional electrostatics with PID potential $\lambda_A(\psi_A)\lambda_B(\psi_B)V_{D-H}(r)$, where $V_{D-H}(r) = \frac{\pm e^2 \exp(-r/s)}{4\pi\varepsilon\varepsilon_0 r}$ is the classic Debye-Hueckel potential
GC	Gaussian chain model with Kuhn length b

Table S4: Legend for the names of the model variants.

P^- F $MD_{0.1}$ C	0.814	13.3	P^+ F_B $MD_{0.1}$ C	0.726	40.1	P^- F_B $MD_{0.3}$ C	0.682	720.9
P^- F $MD_{0.4}$ C	0.813	14.0	A^- L $MD_{0.1}$ C	0.725	99.8	W^- F_B $MD_{0.2}$ C	0.681	138.5
P^+ F $MD_{0.1}$ C	0.812	13.7	A^- L $MD_{0.4}$ C	0.725	99.5	W^+ F_B $MD_{0.1}$ C	0.679	156.6
P^- F $ME_{0.0}$ C	0.812	14.4	A^- L_B $MD_{0.4}$ C	0.725	99.5	P^+ F_B MD_1 T	0.679	>1000
P^+ F $MD_{0.2}$ C	0.811	14.3	P^+ F $MD_{0.1}$ R	0.723	21.2	W^+ F_B $MD_{0.1}$ D	0.677	147.3
P^- F $ME_{0.0}$ T	0.809	11.6	W^+ F_B $MD_{0.1}$ R	0.723	130.0	P^+ F_B $MD_{0.1}$ C	0.677	32.9
P^- F $MD_{0.2}$ C	0.809	13.9	A^+ L $MD_{0.1}$ C	0.721	137.0	P^+ F_B $MD_{0.2}$ C	0.675	32.2
P^- F $MD_{0.3}$ R	0.808	29.6	A^+ L $MD_{0.4}$ C	0.721	136.7	W^+ F_B $MD_{0.3}$ R	0.674	173.5
P^+ F $MD_{0.4}$ C	0.806	16.7	A^+ L_B $MD_{0.4}$ C	0.721	136.7	A^+ F $MD_{0.2}$ R	0.673	197.3
P^+ F $MD_{0.4}$ D	0.803	13.6	A^+ L $ME_{0.5}$ R	0.721	80.7	A^+ F_B $MD_{0.2}$ R	0.673	197.3
P^- F $MD_{0.4}$ D	0.802	10.7	P^+ F $MD_{0.3}$ C	0.720	113.1	A^+ F $MD_{0.1}$ R	0.673	197.5
P^+ F $MD_{0.1}$ D	0.799	13.0	P^+ F_B $MD_{0.3}$ C	0.720	27.0	A^+ F_B $MD_{0.1}$ R	0.673	197.5
P^+ F $MD_{0.2}$ D	0.790	13.1	W^+ F_B $MD_{0.2}$ R	0.718	136.4	A^+ F $MD_{0.4}$ R	0.673	202.1
GC, $b = 6.7$ Å	0.788	89.1	P^- F $MD_{0.5}$ R	0.718	386.9	A^+ F_B $MD_{0.4}$ R	0.673	202.1
P^- F_B $MD_{0.5}$ C	0.788	17.9	P^- F $MD_{0.1}$ R	0.718	386.9	P^+ F_B MD_1 C	0.673	>1000
P^- F_B $MD_{0.1}$ C	0.788	17.9	P^- F_B $MJ_{0.1}$ R	0.716	24.2	W^- F_B $MD_{0.4}$ D	0.673	125.2
P^- F_B $MJ_{0.1}$ C	0.786	17.3	P^+ F_B MD_1 R	0.713	525.2	P^+ F_B $MD_{0.2}$ D	0.672	29.4
P^- F $ME_{0.0}$ D	0.786	13.6	P^+ F_B $MD_{0.1}$ D	0.713	29.0	P^+ F_B $MD_{0.1}$ D	0.672	27.7
P^- F $MD_{0.1}$ D	0.786	13.4	P^- F_B $MD_{0.5}$ R	0.712	24.5	W^+ F_B $MD_{0.2}$ D	0.670	146.0
P^- F_B $ME_{0.1}$ C	0.785	61.6	P^- F_B $MD_{0.1}$ R	0.712	24.5	W^- F_B $MD_{0.1}$ D	0.669	123.0
P^- F $MD_{0.2}$ D	0.783	12.5	P^- F $MD_{0.1}$ R	0.707	25.2	P^+ F_B $MD_{0.4}$ C	0.669	34.0
P^+ F $MD_{0.3}$ R	0.783	43.1	P^- F $MJ_{0.1}$ R	0.707	333.2	W^- F_B $MD_{0.2}$ D	0.668	127.7
A^- L $MD_{0.4}$ R	0.765	62.8	P^- F $ME_{0.0}$ R	0.704	27.6	P^- F_B $MD_{0.1}$ C	0.666	34.4
A^- L $MD_{0.1}$ R	0.764	62.5	P^- F $MD_{0.2}$ R	0.702	25.9	P^- F $ME_{0.1}$ R	0.666	677.3
A^+ L $MD_{0.4}$ R	0.752	94.2	GC, $b = 5.2$ Å	0.700	513.4	P^+ F_B $MD_{0.3}$ R	0.665	35.1
A^+ L_B $MD_{0.4}$ R	0.752	95.0	W^+ F_B $MD_{0.4}$ C	0.698	144.3	P^+ F_B $MD_{0.4}$ D	0.663	29.8
A^+ L $MD_{0.1}$ R	0.751	95.3	P^+ F $MD_{0.5}$ R	0.698	211.5	P^- F_B $MD_{0.1}$ D	0.663	33.8
P^- F_B $ME_{0.1}$ R	0.744	19.8	P^+ F $MD_{0.1}$ R	0.698	211.5	P^- F_B $MD_{0.4}$ C	0.659	35.9
W^- F_B $MD_{0.4}$ R	0.744	97.0	W^- F_B $MD_{0.4}$ C	0.698	118.2	P^- F_B $MD_{0.2}$ C	0.658	35.1
P^- F $MD_{0.3}$ C	0.738	214.6	P^+ F_B $MD_{0.5}$ R	0.692	34.1	P^- F_B $MD_{0.3}$ R	0.656	36.9
P^- F_B $MD_{0.3}$ C	0.733	24.1	P^+ F_B $MD_{0.1}$ R	0.692	34.1	P^- F $MD_{0.5}$ C	0.655	629.9
W^- F_B $MD_{0.2}$ R	0.730	101.2	W^- F_B $MD_{0.3}$ R	0.688	127.8	P^- F $MD_{0.1}$ C	0.655	629.9
W^- F_B $MD_{0.1}$ R	0.730	79.6	P^+ F_B MD_1 D	0.688	689.2	P^- F_B $MD_{0.2}$ D	0.653	34.1
P^- F $MD_{0.4}$ R	0.729	21.2	P^+ F $MJ_{0.1}$ R	0.687	254.1	P^- F_B $MD_{0.4}$ D	0.651	35.8
P^- F_B $MD_{0.5}$ R	0.728	483.6	W^- F_B $MD_{0.1}$ C	0.686	99.0	A^- F $MD_{0.1}$ R	0.650	167.4
P^- F_B $MD_{0.3}$ R	0.728	483.6	W^+ F_B $MD_{0.4}$ D	0.686	148.9	A^- F $MD_{0.4}$ R	0.649	167.3
P^+ F_B $MD_{0.1}$ T	0.727	29.9	A^+ L $ME_{0.5}$ C	0.685	100.2	A^- F_B $MD_{0.4}$ R	0.649	167.3
P^+ F $MD_{0.4}$ R	0.727	22.4	A^+ L_B $MD_{0.2}$ C	0.685	100.2	A^+ F $ME_{0.5}$ R	0.647	158.5
P^+ F $MD_{0.2}$ R	0.727	22.3	A^+ L_B $MD_{0.1}$ C	0.685	100.2	P^- F $MJ_{0.1}$ C	0.643	527.3
W^+ F_B $MD_{0.4}$ R	0.726	127.2	W^+ F_B $MD_{0.2}$ C	0.682	156.9	A^+ F $MD_{0.2}$ C	0.643	247.4
P^+ F_B $MD_{0.5}$ C	0.726	40.1	P^- F_B $MD_{0.5}$ C	0.682	720.9	A^+ F_B $MD_{0.2}$ C	0.643	247.4

Table S5: Pearson coefficients and χ^2 values for all model variants, part 1.

W^- F_B $MD_{0.5}$ R	0.643	179.7
A^+ F $MD_{0.1}$ C	0.642	247.5
A^+ F_B $MD_{0.1}$ C	0.642	247.5
P^+ F $MD_{0.1}$ D	0.642	603.0
A^+ F $MD_{0.4}$ C	0.642	252.2
A^+ F_B $MD_{0.4}$ C	0.642	252.2
P^+ F $MD_{0.1}$ T	0.641	682.1
P^+ F $MD_{0.5}$ C	0.634	439.5
P^+ F $MD_{0.1}$ C	0.634	439.5
A^- F $MD_{0.4}$ C	0.634	211.5
A^- F $MD_{0.1}$ C	0.633	211.9
A^- F_B $MD_{0.4}$ C	0.633	211.5
P^+ F $ME_{0.1}$ R	0.632	639.0
W^- F_B $MD_{0.3}$ C	0.631	176.3
P^+ F_B $MD_{0.1}$ R	0.631	42.8
W^+ F_B $MD_{0.5}$ R	0.630	230.0
W^+ F_B $MD_{0.1}$ R	0.630	230.0
P^+ F $MJ_{0.1}$ D	0.629	333.9
P^+ F_B $MD_{0.2}$ R	0.628	41.8
A^+ F ME_1 T	0.625	282.3
P^+ F $MJ_{0.1}$ T	0.623	407.8
P^+ F $MJ_{0.1}$ C	0.621	512.9
P^+ F_B $MD_{0.4}$ R	0.620	43.1
P^- F_B $MD_{0.2}$ R	0.619	43.2
W^+ F_B $MD_{0.3}$ C	0.618	220.8
P^- F_B $MD_{0.1}$ R	0.618	45.7
A^+ L ME_1 C	0.617	299.9
A^+ F $ME_{0.5}$ C	0.612	183.8
P^- F_B $MD_{0.4}$ R	0.610	44.7
P^- F $ME_{0.1}$ C	0.602	>1000
A^+ L ME_1 T	0.601	334.4
P^+ F $ME_{0.1}$ D	0.596	923.2
W^- F_B $MD_{0.5}$ C	0.595	286.2
W^+ F_B $MD_{0.5}$ C	0.586	399.6
W^+ F_B $MD_{0.1}$ C	0.586	399.6
W^+ F_B $MD_{0.1}$ D	0.584	466.4
P^+ F $ME_{0.1}$ C	0.582	>1000
P^- F_B $MJ_{0.3}$ R	0.581	>1000
W^+ F_B $MD_{0.1}$ T	0.579	380.2
P^+ F $ME_{0.1}$ T	0.577	>1000
GC, $b = 4.1$ Å	0.570	>1000
P^- F_B $MJ_{0.3}$ C	0.558	>1000
P^- F $MD_{0.5}$ R	0.534	>1000
P^- F $MD_{0.3}$ R	0.534	>1000
W^+ F_B $MD_{0.5}$ R	0.531	933.3
GC, $b = 3.8$ Å	0.531	>1000
P^- F_B $ME_{0.3}$ R	0.527	>1000
P^- F $MD_{0.5}$ C	0.512	>1000
P^- F $MD_{0.3}$ C	0.512	>1000
P^+ F MD_1 R	0.511	>1000
W^+ F_B $MD_{0.5}$ D	0.510	>1000
W^+ F_B $MD_{0.5}$ C	0.504	>1000
W^+ F_B $MD_{0.5}$ T	0.500	>1000
P^- F_B $MJ_{0.5}$ R	0.499	>1000
P^- F_B $ME_{0.3}$ C	0.498	>1000
P^+ F MD_1 D	0.497	>1000
P^+ F MD_1 C	0.492	>1000
P^+ F $MD_{0.2}$ C	0.492	>1000
W^+ F $MD_{0.1}$ R	0.488	599.1
P^+ F MD_1 T	0.486	>1000
P^+ L ME_1 C	0.485	>1000
W^+ F $MD_{0.2}$ R	0.485	548.7
W^- F $MD_{0.1}$ R	0.485	482.5
W^- F $MD_{0.4}$ R	0.482	432.4
W^+ F $MD_{0.4}$ R	0.481	531.5
W^- F $MD_{0.2}$ R	0.481	606.3
P^- F_B $MJ_{0.5}$ C	0.480	>1000
W^+ F $MD_{0.2}$ C	0.477	550.4
W^+ F $MD_{0.1}$ C	0.475	740.2
W^+ F $MD_{0.4}$ C	0.474	678.3
W^+ F $MD_{0.1}$ D	0.472	602.1
W^+ F $MD_{0.4}$ D	0.472	536.2
W^- F $MD_{0.2}$ C	0.471	532.9
W^- F $MD_{0.1}$ C	0.471	513.4
W^+ F $MD_{0.2}$ D	0.471	726.9
W^- F $MD_{0.4}$ C	0.470	576.1
W^- F $MD_{0.4}$ D	0.468	606.8
W^- F $MD_{0.2}$ D	0.468	597.7
W^+ F $MD_{0.3}$ R	0.465	770.1
W^- F $MD_{0.1}$ D	0.465	536.0
W^- F $MD_{0.3}$ R	0.464	546.1
P^+ L ME_1 T	0.459	>1000
W^+ F $MD_{0.3}$ C	0.458	712.7
W^- F $MD_{0.3}$ C	0.455	743.2
W^+ F $MD_{0.5}$ R	0.450	901.0
W^+ F $MD_{0.1}$ R	0.450	901.0
P^- F $MJ_{0.3}$ R	0.450	>1000
W^- F $MD_{0.5}$ R	0.448	638.4
W^+ F $MD_{0.1}$ D	0.444	961.5
P^- F $MJ_{0.3}$ C	0.444	>1000
W^+ F $MD_{0.5}$ C	0.443	970.3
W^+ F $MD_{0.1}$ C	0.443	970.3
P^+ F $MJ_{0.3}$ R	0.440	>1000
W^+ F $MD_{0.1}$ T	0.439	873.8
W^- F $MD_{0.5}$ C	0.438	830.8
P^+ F $MJ_{0.3}$ D	0.435	>1000
P^+ F $MJ_{0.3}$ C	0.434	>1000
P^+ F $MJ_{0.3}$ T	0.431	>1000
P^- F_B $ME_{0.5}$ R	0.427	>1000
P^- F $ME_{0.3}$ R	0.421	>1000
P^- F_B $ME_{0.5}$ C	0.420	>1000
P^- F $ME_{0.3}$ C	0.416	>1000
W^+ F $MD_{0.5}$ R	0.414	>1000
P^- F $MJ_{0.5}$ R	0.413	>1000
P^+ F $ME_{0.3}$ R	0.412	>1000
P^+ F $ME_{0.3}$ D	0.412	>1000
W^+ F $MD_{0.5}$ C	0.412	>1000
P^- F $MJ_{0.5}$ C	0.410	>1000
W^+ F $MD_{0.5}$ D	0.409	>1000
P^+ F $ME_{0.3}$ C	0.408	>1000
W^+ F $MD_{0.5}$ T	0.408	>1000
P^+ F $MJ_{0.5}$ R	0.406	>1000
P^+ F $ME_{0.3}$ T	0.406	>1000
P^+ F $MJ_{0.5}$ D	0.403	>1000
P^+ F $MJ_{0.5}$ C	0.402	>1000
P^+ F $MJ_{0.5}$ T	0.401	>1000
P^- F $ME_{0.5}$ R	0.389	>1000
P^- F $ME_{0.5}$ C	0.386	>1000
P^+ F $ME_{0.5}$ D	0.384	>1000
P^+ F $ME_{0.5}$ R	0.384	>1000
GC, $b = 2.7$ Å	0.382	>1000
P^+ F $ME_{0.5}$ C	0.381	>1000
P^+ F $ME_{0.5}$ T	0.381	>1000

Table S5: Pearson coefficients and χ^2 values for all model variants, part 2.

3.3 SAXS profiles

Fig. S12 shows a direct comparison of experimental SAXS data (magenta) with results of CG simulations of protein 6AAA (red). The best agreement is seen in the region of lowest scattering angles, with $q < 0.15$ Å^{-1}, which contains most of information on the overall shape and size of the protein.

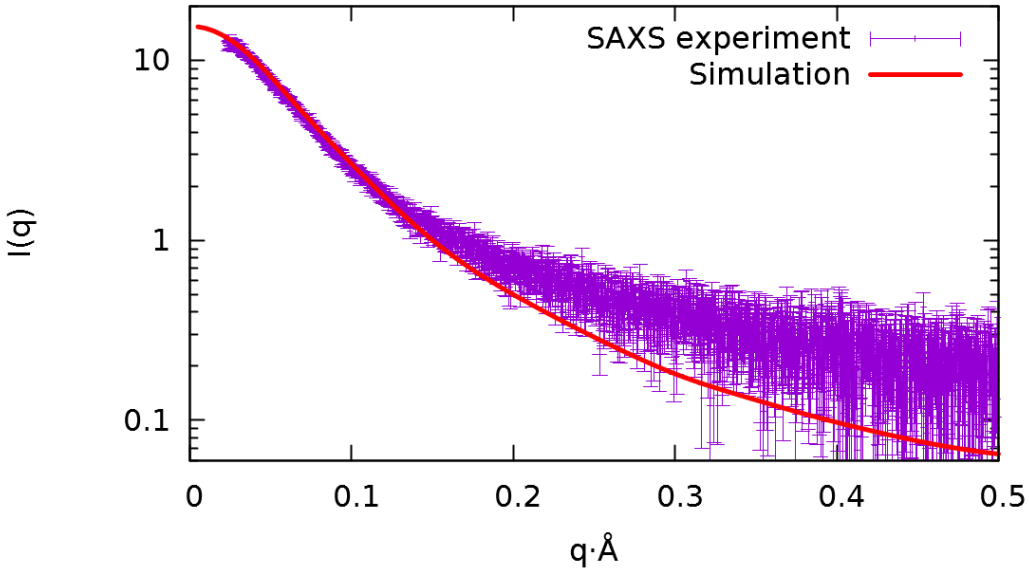


Figure S12: SAXS intensity I as a function of momentum transfer q measured experimentally for protein 6AAA [20] (magenta) and compared to simulation results (red). The simulation was carried out with the P$^-$ F MD$_{0.1}$ C variant of the model. The scattering profile was computed for the 6AAA simulation trajectory using an algorithm co-developed with the EROS method [21]. Default parameters of the hydration shell on the protein surface were used. Intensity (in arbitrary units) was rescaled to match the experiment.

4 Structured proteins

We did not check how all 246 variants of our model perform for structured proteins, because the variants of the model best for structured proteins turned out to be very different from the top variants for IDPs - in the Table S6 we see that the smallest RMSD is achieved for interaction matrices multiplied by the factor 1, whereas for IDPs this factor is smaller than 0.5 for the top models.

We measured Root Mean Square Deviation (RMSD) from the native structure of 3 proteins (1L2Y: 20 residues, 1ERY: 39 residues, 1UBQ: 76 residues) for simulations starting from a self-avoiding random walk or from the native state. In both cases the simulations lasted 150 000 τ.

1L2Y			1ERY			1UBQ		
Model id	F_{RMSD}	N_{RMSD}	Model id	F_{RMSD}	N_{RMSD}	Model id	F_{RMSD}	N_{RMSD}
P$^+$F$_B$ME$_1$T	6.0 Å	6.0 Å	P$^-$F$_B$ME$_1$T	7.0 Å	6.3 Å	P$^+$F MJ$_1$T	11.1 Å	7.0 Å
P$^-$F$_B$ME$_1$T	6.0 Å	6.0 Å	W$^+$F$_B$ME$_1$T	7.0 Å	6.3 Å	P$^+$F$_B$ME$_1$T	11.3 Å	7.7 Å
P$^-$F ME$_1$T	6.1 Å	6.1 Å	P$^+$F$_B$ME$_1$T	7.0 Å	6.6 Å	P$^-$F MJ$_1$T	11.4 Å	7.3 Å
P$^-$F MD$_{0.1}$C	6.2 Å	6.1 Å	W$^+$L ME$_1$T	7.0 Å	6.7 Å	W$^-$F ME$_1$T	11.4 Å	7.3 Å
P$^-$F MD$_{0.4}$C	6.2 Å	6.1 Å	W$^-$F ME$_1$T	7.3 Å	6.7 Å	W$^+$F$_B$MJ$_{0.5}$T	11.4 Å	8.3 Å
P$^-$F$_B$MJ$_1$T	6.2 Å	6.2 Å	W$^-$F MJ$_1$T	7.3 Å	6.8 Å	P$^-$F$_B$ME$_1$T	11.5 Å	8.1 Å
P$^-$F MJ$_1$T	6.3 Å	6.3 Å	W$^-$F$_B$MJ$_1$T	7.5 Å	6.0 Å	P$^+$F MD$_1$T	12.2 Å	9.3 Å
P$^+$F$_B$MJ$_1$T	6.4 Å	6.4 Å	P$^+$L ME$_1$T	7.5 Å	6.8 Å	W$^-$L ME$_1$T	12.6 Å	-
P$^+$F MJ$_1$T	6.4 Å	6.5 Å	P$^-$F$_B$MJ$_1$T	7.6 Å	6.1 Å	P$^-$F MD$_{0.1}$C	12.7 Å	7.8 Å
W$^-$L$_B$MD$_1$T	8.6 Å	8.8 Å	W$^+$F$_B$MD$_{0.5}$T	8.0 Å	7.4 Å	P$^-$L ME$_1$T	12.8 Å	-

Table S6: RMSD for 3 structured proteins (1L2Y, 1ERY and 1UBQ) for simulations starting from a self-avoiding walk (folding, F_{RMSD}) and from the native structure (N_{RMSD}).

References

[1] The function is made from two mirrored algebraic sigmoid functions, retrieved from: https://blob.pureandapplied.com.au/sigmoid-a-post-about-an-algebraic-function-im-having-too-much-fun/ (accessed January 15, 2020).

[2] Mioduszewski, Ł.; Cieplak, M. Disordered peptide chains in an α-c-based coarse-grained model. *Phys. Chem. Chem. Phys.* **2018**, *20*, 19057-19070.

[3] Miyazawa, S.; Jernigan, R. L. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*(3), 623-644.

[4] Betancourt, M. R.; Omovie, S. J. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.* **2009**, *130*(19), 195103.

[5] Tozzini, V.; Trylska, J.; Chang, C. E.; McCammon, J. A. Flap opening dynamics in hiv-1 protease explored with a coarse-grained model. *J. Struct. Biol.* **2007**, *157*(3), 606-615.

[6] Holehouse, A. S.; Das, R. K.; Ahad, J. N.; Richardson M. O.G.; Pappu R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **2017**,*112*(1), 16-21.

[7] Requião, R. D.; Fernandes, L.; de Souza, H.; Rossetto, S.; Domitrovic, T.; Palhano, F. L. Protein charge distribution in proteomes and its impact on translation. *PLOS comp. biol.* **2017**, *13*(5), e1005549.

[8] Kyte, J.; Doolittle, R. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105-132.

[9] Cragnell, C.; Rieloff, E.; Skepö, M. Utilizing coarse-grained modelling and monte carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. *J. Mol. Biol.* **2018**, *430*(16), 2478-2492.

[10] Dignon, G.; Zheng, W.; Kim, Y.; Best, R.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Comp. Biol.* **2018**, *14*(1), e1005941. [Table S2]

[11] Varadi, M.; Kosol, S.; Lebrun, P.; Valentini, E.; Blackledge, M.; Dunker, A. K.; Felli, I. C.; Forman-Kay, J. D.; Kriwacki, R. W.; Pierattelli, R.; Sussman, J.; Svergun, D. I.; Uversky, V. N.; Vendruscolo, M.; Wishart, D.; Wright, P. E.; Tompa, P. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucl. Acids Res.* **2014**, *42*(D1), D326-D335.

[12] Cragnell, C.; Durand, D.; Cabane, B.; Skepo, M. Coarse-grained modeling of the intrinsically disordered protein histatin 5 in solution: Monte carlo simulations in combination with SAXS. *Proteins* **2016**, *84*, 777-791.

[13] Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; De Groot, B. L.; Grubmüller, H. Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513-5524.

[14] Mylonas, E.; Hascher, A.; Bernado, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochem.* **2008**, *47*(39), 10345–10353.

[15] Kung, C. C.; Naik, M. T.; Wang, S.; Shih, H.; Chang, C.; Lin, L.; Chen, C.; Ma, C.; Chang, C.; Huang, T. Structural analysis of poly-sumo chain recognition by the rnf4-sims domain. *Biochem. J.* **2014**, *462*, 53-65.

[16] Chukhlieb, M.; Raasakka, A.; Ruskamo, S. Kursula, P. The N-terminal cytoplasmic domain of neuregulin 1 type III is intrinsically disordered. *Amino Acids* **2015**, *47*(8), 1567-1577.

[17] Moncoq, K.; Broutin, I.; Larue, V.; Perdereau, D.; Cailliau, K.; Browaeys-Poly, E.; Burnol, A.-F.; Ducruix, A. The pir domain of grb14 is an intrinsically unstructured protein: implication in insulin signaling. *FEBS Lett.* **2003**, *554(3)*, 240-246.

[18] Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 240-246.

[19] Cordeiro, T. N.; Herranz-Trillo, F.; Urbanek, A. N.; Estaña, A. N.; Cortés, J.; Sibille, N.; Bernadó, P. N. Structural Characterization of Highly Flexible Proteins by Small-Angle Scattering. *Adv. Exper. Med. Biol.* **2017**, *1009*, 107-129.

[20] De Biasio, A.; Ibáñez de Opakua, A.; Cordeiro, T. N.; Villate, M.; Merino, N.; Sibille, N.; Lelli, M.; Diercks, T.; Bernadó, P.; Blanco, F. J. p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. *Biophys. J.* **2014**, *106*(4), 865–874. SAXS data was retrieved from `https://web.archive.org/web/20160911130751/http://pedb.vib.be/accession.php?ID=PED6AAA` (accessed May 12, 2020).

[21] Różycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19*(1), 109-116.