Please do not adjust margins

ARTICLE

Nearest Neighbour Interactions between Amino Acid Residues in Short Peptides and Coil Libraries

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Reinhard Schweitzer-Stenner*[a]

Intrinsically disordered proteins (IDP) or proteins with intrinsically disordered regions (IDR) perform a plethora of functions mostly in a cellular environment. As unfolded proteins, IDPs can adopt molten globule or coil ensembles of conformations. Regarding the latter the question arises whether they are describable as a self-avoiding random coil. Locally, this requires that amino acid residues sample the entire sterically allowed region of the Ramachandran plot with very similar probabilities and independent on the conformational dynamics of their neighbours. However, various lines of experimental and bioinformatic evidence suggest a more restricted, side chain and nearest neighbor dependent conformational space for individual residues. Over the last 25 years short peptides and coil libraries were employed to determine conformational propensities of amino acid residues in unfolded states. The question arises whether conformational ensembles obtained from these two sources are comparable. In this paper, a variety of metrics were used to compare Ramachandran plots of a limited number of GXYG peptides (X,Y: guest residues) with XY dimers in the coil library of Ting et al.(PLOS 6, e1000763, 2010). The results reveal major differences between corresponding plots ,which might in part due to the fact that solely the influence of one of the two neighbours of a given residue is probed by the above coil library while averages were taken over the respective opposite neighbours. The presented results suggest that coil libraries alone might not be a sufficient tool for determining the characteristics of statistical coils of IDPS and IDRs alike.

Introduction

It is now well established that intrinsically disordered proteins (IDPs) and folded proteins with intrinsically disordered regions (IDRs) perform a plethora of functions mostly in a cellular environment. Their very existence shatters the conventional dogma that stipulate the need for a stable folded protein for its function.(1–6) Estimates of the fraction of biologically relevant IDPs vary in the literature.(2,6) A recent study reported 30% of eukaryotic proteins to be at least partially disordered.(7) In many cases IDPs or IDRs of otherwise folded proteins are involved in molecular recognition processes that involve disorder → order transitions.(8–13) In less frequent cases transitions into fuzzy or even another disordered state occur.(10,14,15) Protein disorder was found to dominate in proteins associated with signal transduction, cell-cycle regulation, gene expression and chaperone action.(5,6,16,17) Besides being of pivotal importance for biological processes IDPs can also malfunction by self-assembly into fibrillar structures that have been implicated in several neurological diseases.(2,18–20)

Generally, IDPs exhibit a dynamic behavior that allows them to exist as an ensemble of energetically similar conformations under physiological conditions. Several lines of evidence suggest that in structural terms IDPs can be treated as unfolded proteins so that their conformational ensembles can be classified either as collapsed molten globules (if e.g. water is a poor solvent) and or as self-avoiding random coils (if water is a good solvent). Uversky et al. used UV circular dichroism data to identify pre-molten globules as an additional option.(3) A more nuanced picture emerged from NMR studies using dipolar coupling and secondary chemical shifts, which revealed the occurrence of local residual structure in ensembles which otherwise appear dominantly coiled. The propensities of proteins for these three states depends on their amino acid composition.(21–25) Since IDPs are generally rich in amino acid residues with charged or polar side chains,(26,27) one might suspect that water would function as a good solvent and thus favor the more extended self-avoiding coil state. However, several lines of evidence reported in the literature show that IDPs and homopeptides with polar residues adopt collapsed conformations even in the presence of denaturing co-solvents. The most prominent example is poly-L-glutamine.(28) Even an excess of charges does not guarantee an extended state. Pappu and coworkers showed that a preference for the coiled state requires that the net charge per residue of a polypeptide chain must exceed a threshold value and opposite charges must be well mixed along the backbone chain for moving the molecule above the theta point at a given temperature.(29–33) This notion is confirmed by single molecule experiments on protein segments and predicted scaling exponents of a large number of proteins.(34–36)

Conceptually, both the ideal Flory random coil model and the self-avoiding random walk model are based on the assumption that a polymer can be described as a freely jointed chain of subunits.(37–39) In the case of polypeptides/proteins this would require a nearly unhindered rotation around the NC^{α} and $C^{\alpha}C'$ bonds that connect the peptide groups. Apparently, this idealization is at variance with the observation that steric hindrance and electrostatic interactions between residues restrict the corresponding dihedral backbone angles φ and ψ.(40–43) This restriction is generally considered as largely insignificant for two reasons.(44) First, the sterically allowed regions of the Ramachandran plot of individual amino acid residues are believed to be very similar with the sole exception

[a] Department of Chemistry, Drexel University, Philadelphia, PA 19104, USA.
† Footnotes relating to the title and/or authors should appear here.
Supplementary Information available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

Please do not adjust margins

Please do not adjust margins

of glycine (more conformationally flexible due to the lack of chirality) and proline (more restricted due to the pyrrolidine ring). Hence, individual sequences of polypeptides and proteins would not matter much with regard to experimentally accessible parameters like the radius of gyration and end to end distances. Second, multiple lines of experimental evidence revealed that e.g. the radii of gyration of proteins and IDPs in denaturing solvents exhibit scaling exponents which cluster around 0.6,(45,46) the expectation value for a self-avoiding random walk in a good solvent.(47,48)

Over the last 25 years experimental, computational and bioinformatic evidence has been accumulated which suggests a more restricted and side chain dependent conformational sampling of the Ramachandran space. Spectroscopic experiments on short host-guest residue type peptides such as blocked dipeptides, NH_3^+-GXG-COOH and Ac-GGXGG-NH_2, peptides (X: guest residue) revealed that e.g. alanine exhibits a high preference for a conformation called polyproline II (pPII), while valine slightly favors β-strand over pPII (cf. ref. (49) and references cited therein). Protonated aspartic acid samples turn-supporting conformations to a significant extent. Similar though not identical conformational preferences have been obtained in so-called coil libraries, which were constructed from the dihedral angles adopted in segments of folded proteins which do not depict one of the well-known regular structures (ref. (50) and references cited therein). They are considered as representative of coiled proteins because it is assumed that any non-local interactions and steric constraints are averaged out if one builds Ramachandran distributions with a large number of proteins that possess a significant fraction of loops. Computationally, an increasing number of molecular dynamics (MD) simulations using modified Amber, CHARMM or OPLS forcefield and TIP water models is capable of reproducing the high pPII propensity of alanine to some extent,(51–57) but a reproduction of the side chain specificity has only recently been achieved on a qualitative level.(58)

The discovery of residue specific propensities changed our picture of coiled peptides and proteins in that they suggest that, locally, unfolded peptides and proteins exhibit a statistical coil rather than a (restricted) random coil behaviour. In the case of the latter the Ramachandran plot would be describable by a very broad and shallow minimum of the Gibbs energy landscape in the upper left quadrant of the Ramachandran plot so that all conformations in the region between ϕ=-160° and - 50° as well as between ψ=180° and 90° are practically isoenergetic. Traditionally, it was called the β-strand region.(40,41) A less extended region between ϕ=-160° and - 50° as well as ψ=20° and -70° exhibits similar minimal Gibbs energies as the broad region in the upper left quadrant. It encompasses all types of right-handed helical and type I/II β-turn conformation generally sampled by the i+2 residue. Much smaller regions of higher Gibbs energy are associated with left-handed helical and other turn structures. In the statistical coil case suggested by the short peptide and coil library data, the broad basin in the upper left quadrant splits into two much more confined minima with side chain dependent free energies, which in many cases are not identical. Conformations in the right-handed helical/β-turn I/II region are generally of significantly higher energy. Hence, the free energy landscape is describable as a statistical canonical set of Boltzmann factors. The (restricted) random coil case thus appears as a special case of a statistical coil. In Figure 1 it is approximated by a mixture of nearly isoenergetic Gaussian basins the parameters of which are listed in Table S1 (taken from ref.(59)). Generally, statistical coil ensembles carry significant less conformational entropy then a classical restricted random coil distribution.(59,60)

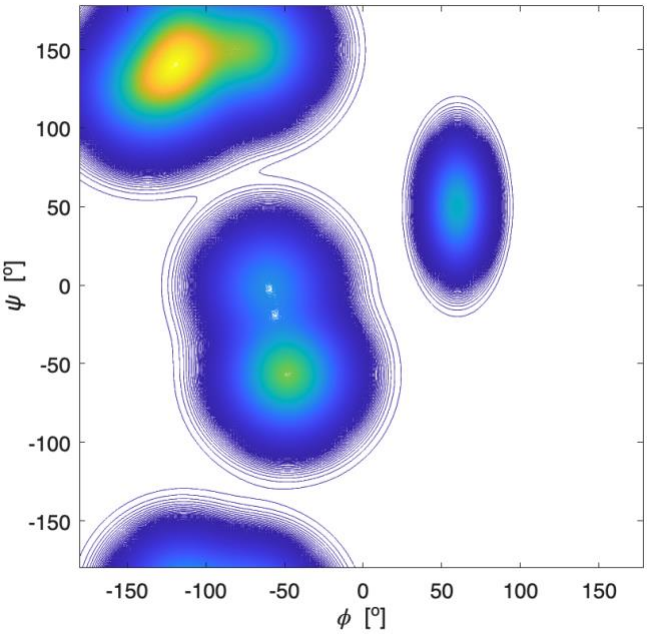

Figure 1. Ramachandran plot of a (local) restricted random coil distribution of conformational ensembles. The respective parameters have been taken from ref.(59). They are listed in Table S1.

Even though the above conformational restrictions of the Ramachandran space seem to have a limited influence on global properties of coils such as the radius of gyration or end to end distance,(46,61,62) their assessment is of utmost importance for producing statistical coil reference systems based on which local transient residual structures in unfolded proteins and IDPs can be identified by nuclear magnetic resonance spectroscopy (ref.(22) and references cited therein). Information about the local properties of coiled peptides and proteins also allow for an accurate estimation of configurational entropies.(60,63) Unfortunately, the task of identifying the coiled state of proteins is complicated by nearest and potentially even second nearest neighbour interactions (NNIs) between residues (ref.(50) and references cited therein). The occurrence of such interactions is well established in general but the underlying mechanisms are not well understood. A very important issue in this regard is whether or not NNIs involve correlated conformational motions between residue, in which case they would depend on the conformation adopted by the neighbours. In other words: NNIs would depend on whether the neighbours sample pPII, β-strand of helical/turn supporting conformations. Several lines of evidence suggest that such structural correlations exist, in which case the isolated pair hypothesis, another pillar of the random coil model, would no longer be valid (ref.(50) and references cited therein). In such a case, configurational entropies and free solvation energies would no longer be additive.

Carrying out a quantitative analysis of NNIs is difficult because they are side chain dependent. Hence, even if one focuses on

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

NNIs and thus ignores second neighbour effects, one would have to analyse e.g. 20^3 GXYZG unblocked peptides or blocked tetra-peptides in solution in order to cover all possible combinations. The number reduces to 400, if one assumes that the influence of down- and upstream neighbours are additive. Several lines of evidence suggest that this is too optimistic an assumption.(65–69) Spectroscopic investigations of tetra- and pentapeptides in water clearly revealed significant nearest neighbour interactions, which most likely involve correlation effects.(67,70–73) Even though the number of investigated peptides has been significant, they cannot be considered a representative for the entire manifold of side chain combinations. Therefore, coil libraries have been used as a preferred source for NNI analysis. The understanding of what the term coil library entails has evolved over time. At the beginning it was assumed that just inferring the Ramachandran plot of individual residue types from a large number of secondary structures produces a (local) coil distribution, since all structure forming contributions were assumed to average out. This concept was invalidated by Serrano who showed that an avoidance of amino acid residues in classical secondary structures such as right-handed helices and β-strands leads to completely different Ramachandran plots.(74) For instance, alanine Ramachandran plots derived from unrestricted libraries are dominated by the population of right-handed helical conformations, thus reflecting the high helical propensity of this amino acid.(75) On the contrary, the respective Ramachandran plot of the restricted library is dominated by pPII sampling, very much in agreement with what has been obtained for alanine in short peptides. However, the use of restricted libraries makes it difficult to obtain statistically significant Ramachandran plots of all possible up- and downstream neighbours of a specific amino acid residue (cf. the Ramachandran plots that can be constructed at (76)). As a remedy, Ramachandran plots for individual amino acid residues were constructed by averaging over all neighbours.(77–80) Such a strategy would be valid if individual NNIs with different neighbours were very similar. However, the above spectroscopic studies on tetra- and pentapeptides seem to suggest otherwise. (66,67,72,73)

In this paper, I explore the influence of residue averaging by a comparative analysis of Ramachandran distributions of amino acid residues in tetra-peptides and in the coil-library of Dunbrack and coworkers.(81) Due to my best knowledge, the latter provides the most details about the individual influence of neighbours on Ramachandran distributions of all twenty amino acid residues in that the above averaging is restricted either to the right (downstream) or to the left (upstream) neighbours, respectively. In other words, for all-XY pairs, the Ramachandran plots of X were reported for all Y neighbours, while the influence of the left neighbours were averaged, whereas plots for Y were produced for all X neighbours with average over the right neighbours. Ting et al. used a mathematical algorithm to construct complete Ramachandran plots with a resolution of 5° for the ϕ and ψ coordinate from the obtained data points taken from a restricted coil library (no helices, strands and turns). A limited comparison of these coil library Ramachandran plots with tetra-peptide distributions had been undertaken by Schweitzer-Stenner and Toal based on the use of calculated Hellinger distances (*vide infra*).(71) The current paper presents the results of a by far more comprehensive analysis, which compares respective population of mesostates and calculated J-coupling constants. Revised Hellinger distance calculations are augmented by comparisons of configurational Gibbs entropies for individual residues. In addition to the explicit use of the coil library data of Ting et al. results from the analysis of another coil library of Bax and coworkers,(78) which are based on total averaging over neighbours, are briefly discussed. The results of this analysis reveal that the diversity of individual NNIs is not sufficiently represented by coil library distributions and that averaging over neighbours underestimates the influence of neighbours.

Theoretical background

The basics of the so-called Gaussian model implemented for the analysis of J-coupling constants as well Raman, IR, and vibrational dichroism amide I' profiles of tri-, tetra- and pentapeptides have been described in several earlier papers of my research group.(82,83) Briefly, observables O (i.e. scalar J-coupling constants and amide I' profiles in the above spectra) were calculated as ensemble average of a statistical coil distribution:

$$\langle O \rangle = \frac{\int_{-\pi}^{\pi} O(\phi,\psi) P(\phi,\psi)) d\phi d\psi}{Z} \qquad (1)$$

where O represents the considered spectroscopic observable (vide supra) and Z denotes the partition sum of the ensemble. In the past, the distribution function $P(\varphi,\psi)$ was constructed as a superposition of two-dimensional Gaussian functions:

$$P(\phi,\psi) = \sum_{j=1}^{N} \left(\frac{\chi_j}{2\pi\sqrt{|\hat{V}_j|}}\right) exp(-0.5(\vec{\rho} - \vec{\rho}_{j,0})^T \hat{V}_j^T)(\vec{\rho} - \vec{\rho}_{j,0}) \qquad (2)$$

where the index j represents the jth conformation summed over N different sub-distributions with statistical weights χ_j. The position vector $\vec{\rho}$ represent a coordinate pair (φ,ψ), $\vec{\rho}_{j,0}$ denotes the position of the center of the sub-distribution. $\hat{V}$ is a 2 x 2 matrix, the diagonal of which contains of the standard distributions of the Gaussian sub-distributions with respect to ϕ and ψ. The off-diagonal elements account for correlations between the two coordinates, which in the case of unequal standard deviations for ϕ and ψ lead to a rotation of the axes of the respective ellipse. In the current study, the Gaussian model was solely used for non-linear least square fits of eq. (1) to a set of experimental scalar J-coupling constants of GXYG peptides (i.e. the ϕ-dependent $^3J(H^NH^{C\alpha})$, $^3J(H^NC')$, $^3J(H^{C\alpha}C')$, $^3J(H^NC^\beta)$ and the ψ-dependent $^1J(NC_\alpha)$, which were carried out with a Matlab program (X and Y represent the two guest residues). For the fitting procedure, the statistical weights as well as the positions of the Gaussian sub-distributions were used as free parameters in an iterative procedure described by Milorey et al.(73) The standard deviations of the distributions were slightly varied between 10° and 20° for a fine-tuning of the fits. Scalar coupling constants were calculated by using the Karplus parameters in ref.(84) for the ^3J constants and in ref.(85) for the ^1J constant. Ramachandran plots were calculated with a resolution of 2° for ϕ and ψ. In order to account for the amide I' band profiles, conformations of the entire peptide rather than residue conformations have to be used. The final output of the fitting procedure was used as input for a truncated model applied to reproduce the amide I' profiles. This reduction was necessary to reduce the computational cost of the conformational sampling.

Please do not adjust margins

Please do not adjust margins

A total of 21 conformers represents half of an ellipse around the center pPII and β-strand troughs, respectively. Hence, we substitute the two-dimensional Gaussian distributions by ellipses (not cuboids, as stated in ref.(73)) with axis widths of $2\Delta\varphi_j$ and $2\Delta\psi_j$ in φ and ψ direction, where $\Delta\varphi_j$ and $\Delta\psi_j$ were set equal to the standard deviations of the respective Gaussian distributions. The height of the ellipses is given by the statistical weights of the conformations divided by the total number of representing data points (i.e. 21). Representative plots of these distributions are given in Figure S1. In addition, the less dominant conformations (right-handed helical, inverse γ-turn, asx-turn, etc.) were represented by a single conformation positioned at the respective center of the Gaussian sub-distribution with the statistical weights obtained from the fits of the Gaussian model to J-coupling constants. Several checks revealed that the truncated model reproduced the coupling constants produced by the fitting to the J-coupling data sufficiently well. If the input distribution reproduced the measured amide I' profiles, the analysis was considered completed. If the measured and calculated amide I' profiles differed significantly, the above parameters were slightly modified to reduce the deviations. The newly obtained set of parameters was used a starting point for refitting the J-coupling constants. The procedure was repeated until convergence was obtained. This fitting procedure, which was first implemented by Milorey et al. for their conformational analysis of GRRG and GRRRG peptides,(73) was found to be more reliable and efficient than the procedure that our group implemented in the earlier study of GXYG peptides.(72) Therein, we had used the same restricted sub-space of the Ramachandran plot for manual simulations of the entire data set of J-coupling constants and amide I profiles. The resolution of the conformational sampling was 6°, much lower than the 2° used for the more recent Gaussian analysis of tetra- and penta-peptides.(73) A change of the positions of Gaussian sub-distributions could move part of them outside of the area probed by the respective simulations. Therefore, in order to put all the available tetra-peptide data on an equal footing, I revisited the data of Toal et al.(72) with the above procedure of Milorey et al.(73) It should be noted that the same exciton coupling model was used in both studies to link residue conformations to amide I' band profiles. Details of this model are described in earlier papers.(83)

Different metrics were used to compare the Ramachandran plots of amino acid residues flanked by different neighbours in the investigated tetra-peptides and in the coil library of Ting et al.(81) The Hellinger distance H is a measure of overlap between different distributions. It is defined as follows:

$$H(P_j, P_{j'}) = ||\frac{1}{2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}(\sqrt{P_j(\phi,\psi)} - \sqrt{P_{j'}(\phi,\psi)})^2 d\phi d\psi|| \quad (3)$$

where $P_j(\varphi,\psi)$ and $P_{j'}(\varphi,\psi)$ denote the Ramachandran probability density distributions of the jth residue in tri- and tetrapeptides, respectively. The Hellinger distance can vary between 0 (total agreement) and 1 (completely distinct).† To interpret these numbers, earlier introduced criteria were used.(71) If the Hellinger value lies in the region below 0.1, the corresponding distributions are considered as very similar. If 0.1 < H ≤ 0.25 and 0.25 < H ≤ 0.4, the distributions are moderately similar and moderately dissimilar, respectively. H values larger than 0.4 indicate very dissimilar distributions. The Hellinger distance responds very sensitively to changes of the positions of individual basins (i.e. Gaussian distributions). Sole changes of statistical weights have a lesser influence on H.

An alternative metric is the Gibbs (or Shannon) entropy which for an individual residue is written as:

$$S_x = -R\sum_i P_x(\phi_i,\psi_i) ln P_x(\phi_i,\psi_i) \quad (4)$$

where x is just an indicator of the considered residue and R is the gas constant. The numerical value of S_x depends on the mesh size of the Ramachandran plot (2° for the investigated peptides,5° for the coil library Ramachandran plots).(60,63) Therefore, we just used the differences between the entropies of X with different neighbours as a suitable metric for characterizing NNIs.(59)

Calculating the Gibbs entropies of residues does not provide a complete picture if NNIs involve correlations between conformational motions of neighbours. In such a case only the conformational distributions of entire peptides and even of unfolded and disordered proteins have to be taken into account, because thermodynamic parameters are no longer additive. In order to model such a scenario e.g. for our tetra-peptides, one has to use the Ramachandran plots of GXG peptides as reference and model the NNI induced differences in terms of interaction free energies.(61,66,70,86) The statistical weight of a distinct conformation of a tetra-peptide GXYG where X has adopted conformation L and Y the conformation K is thus written as:

$$\chi_{LK} = \frac{exp[(G_{L,x}+G_{K,y}+\delta G_{LK,yx}+\delta G_{KL,xy})/RT]}{Z} \quad (5)$$

where $G_{L,x}$ and $G_{K,y}$ are the Gibbs free energies of the states L and K of the residues X and Y in GXG/GYG tripeptides. Hence, glycine as neighbour is used as a reference point, for which one would expect minimal nearest neighbour interactions. The interaction free energy between the two conformations in X and Y are denoted as $\delta G_{LK,yx}$ and $\delta G_{KL,xy}$. The latter are obtained by reproducing the differences between the statistical weights of the conformations in GXG/GYG and GxyG.(66,70) Z is the partition sum of the system. It can conveniently be calculated by employing Zimm-Bragg like transfer matrices. Details can be found in ref.(70) The configurational Gibbs entropy of a XY 'dimer' can now be written as:

$$S_{xy} = -R\sum_L(\sum_L \chi_{LK} ln\chi_{LK}) \quad (6)$$

Ramachandran plots of peptide and coil library residues can be compared by calculating the respective population of mesostates. Figure S2 exhibits a graphic representation of the mesostates which were used earlier for the comparison of the Gaussian model with the results of molecular dynamics calculations.(57) Each mesostate is associated with a specific secondary structure. It should be noted that the classical β-strand region is subdivided into areas associated with parallel (pβ), antiparallel (aβ) and transitional β-strand (βt). While pβ and aβ are canonical, βt is a more recent addition introduced by Tran et al.(69) More coarse-grained mesostate models partition this area only into a pPII and a β-strand region. The higher

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

resolution of our model allows for a more nuanced picture of Ramachandran distributions. Since the same mesostates are used for all systems investigated, populations depend on the statistical weight of the respective Gaussian sub-distribution, its width and its position. The population of mesostates are just calculated as mole fraction by integrating the population density over its area in the Ramachandran plot

Results

This section is organized as follows. First, the experimental data reported for GXYG peptides by Toal et al.(72) are revisited using the protocol described in the preceding section. Thus, obtained conformational propensities (statistical weights) and basin positions (of Gaussian sub-distributions) are combined with the results of the more recent tetra-peptide studies of Milorey et al.(67,73) Second, respective Ramachandran plots of GXYG and coil library distributions are compared on a qualitative, visual level. In a third step such comparisons are carried out on a more quantitative level by using calculated J-coupling constants, mesostate populations, Hellinger distances and configurational entropies as metrics.

Conformational propensities and basin positions.

In a first step the experimental data which Toal et al. reported for a series of cationic GXYG peptides in aqueous solution were revisited by employing the protocol of Milorey et al. (*vide supra*). The choice of cationic peptides was motivated by the need to obtain sufficiently high intensities of amide proton signals. The influence of terminal charges on the conformational distributions of central residues in AAA, GAG and GVG have been found to be negligible while it is moderate for GDG.(56,67) These finding suggests that even in tripeptides hydration of terminal groups cause a significant shielding and thus a reduction of electrostatic interactions. The results of the GXYG analysis are presented in Tables S1 and S2 as well as in Figures S3-S7. Table S1 compares the experimental J coupling constants with the ones resulting from the fitting procedure described in the preceding section. The corresponding calculated coupling constants reported by Toal et al. are listed for comparison. Table S2 lists the obtained statistical weights and peak coordinates of the Gaussian sub-distributions for the two non-terminal X,Y residues of the investigated GXYG. Corresponding values reported by Toal et al. are listed in parenthesis for comparison. Figure S3 shows a representative set of fits to the amide I' profiles in the IR, isotropic Raman, anisotropic Raman and VCD spectrum of the investigated GXYG peptides. The fits to corresponding band profiles were obtained with the same spectral parameters (i.e. positions and halfwidths of three Gaussian bands related to transitions into the first excited excitonic state of coupled oscillators) used by Toal et al. It should be noted that owing to the influence of nearest neighbour and non-nearest neighbour coupling, amide I' profiles reflect the conformational states of the peptide rather than the ones of individual residues, which are specifically probed by the utilized J-coupling constants. Figures S4-S7 depict the statistical weight of pPII, β-strand and of all helical and turn-supporting conformations for a specific residue in different GXYG pe1ptides (e.g. of alanine in GD**A**G, GS**A**G, G**A**LG, G**A**VG, G**A**FG and GF**A**G, target residue in bold). In addition to the revised statistical weights of the GXYG peptides investigated by Toal et al.(72), recent results obtained for GAFG, GFAG, GDFG, GFDG, GDDG and GKKG have been added to the graphs.(66,67,73) Representative Ramachandran plots of GXYG peptides are shown below where they are compared with corresponding coil library plots.

Similarities and discrepancies between the new and old parameter values of the Gaussian model are briefly delineated in the Supporting Information. Here, it can just be stated that the main observations of Toal et al. remain valid, namely the rather pronounced decrease of the pPII propensity of alanine in the presence of unlike neighbours, particularly valine. Concomitantly, the β-strand propensity is increased. Aspartic acid ranks second in terms of nearest neighbour sensitivity, but in this case pPII increases in the presence of unlike neighbours. As reported earlier, aspartic acid itself as neighbour causes significant changes and rather position dependent changes. Mole fraction variations of lysine, leucine and valine by unlike neighbours are generally modest, in line with Toal et al. Like neighbours (i.e. in G**K**KG, GK**K**G and GL**L**G produce much more pronounced changes). Similar observations have been made for aspartic acid and arginine containing GXYG peptides. However, the more recently investigated phenylalanine containing peptides are subject to significant mutual NNIs with unlike neighbors.

In order to facilitate a thermodynamic understanding of the plots in Figures S4-S7, Figure S8 depicts the Gibbs free energy difference between pPII and β-strand basins for different populations of the upper left quadrant of the Ramachandran plot. A comparison of the later with the former reveals that a significant number of NNI induced population changes are associated with Gibbs energy changes of several kJ/mol. The destabilization of pPII by serine and valine are associated with Gibbs energy changes of ca. 4 and 3 kJ/mol respectively.

It should be noted that observed changes of statistical weights convey only a part of the picture. As one can infer from Table S2, noticeable changes of the ϕ-position and to a more limited extend of the ψ-position were obtained for all investigated series, including the leucine and serine series for which the propensity changes are moderate. As a consequence, related J-coupling constants are different (Table S1). To illustrate this point, let us look on the propensities and $^{3}J(H^{N}H^{C\alpha})$ value of some leucine residues: The statistical weights for pPII of GA**L**G, G**L**LG, GV**L**G and GD**L**G are 0.42, 0.43, 0.4 and 0.49. Apparently, they are very similar. The first three values are identical within their statistical uncertainty which generally does not exceed ± 0.5.(82) The corresponding $^{3}J(H^{N}H^{C\alpha})$ values are 7.4, 7.62, 8.15 and 6.78 Hz. These variations lie clearly outside of the statistical error limits (< ±0.2 Hz). As shown below, changes of basin positions lead to rather large values of the Hellinger distance.

Please do not adjust margins

Please do not adjust margins

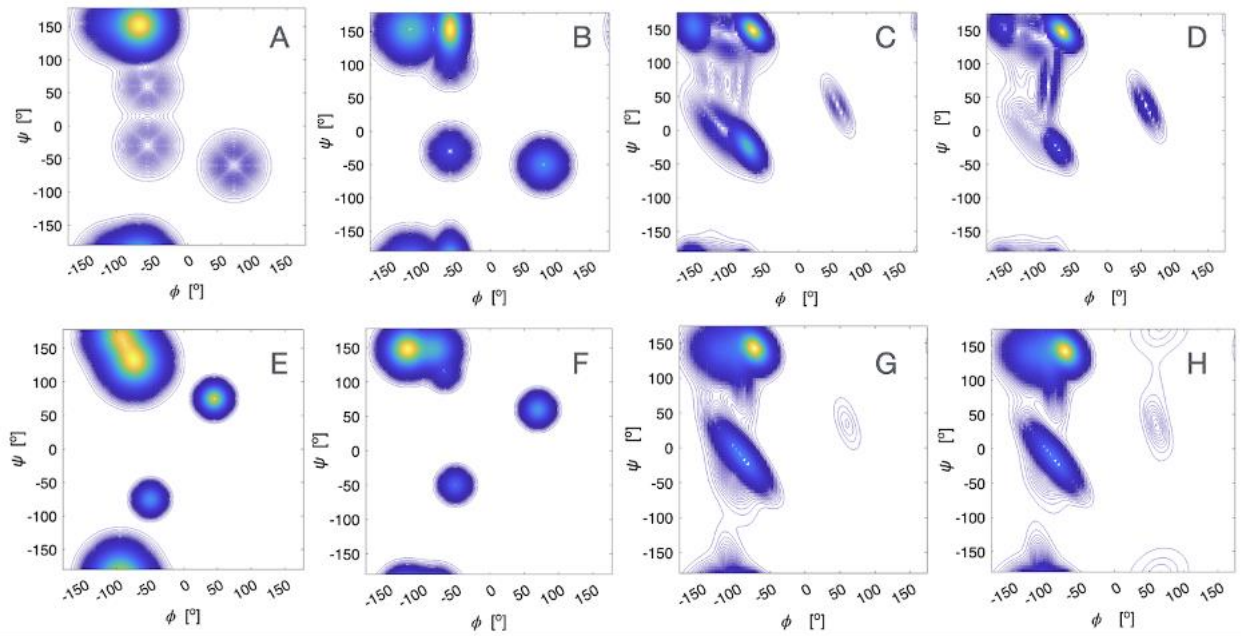

Figure 2: Ramachandran plot of alanine residues in cationic GAG (A), cationic G**A**VG (B), in the coil library segments all-**A**G and all-**A**V (C and D), in cationic GLG (E), cationic GL**L**G (F) and in the coil library segments G**L**-all and L**L**-all (G and H). The tripeptide Ramachandran plots were created with the Gaussian parameters reported in ref.(87) The plots for G**A**VG and GL**L**G were produced for this study by an analysis of experimental data reported by Toal et al.(72), as described in the main text. The plots for A and L in the above coil library segments were created with files obtained from the website of the Dunbrack group with the kind permission of Dr. Dunbrack.(81)

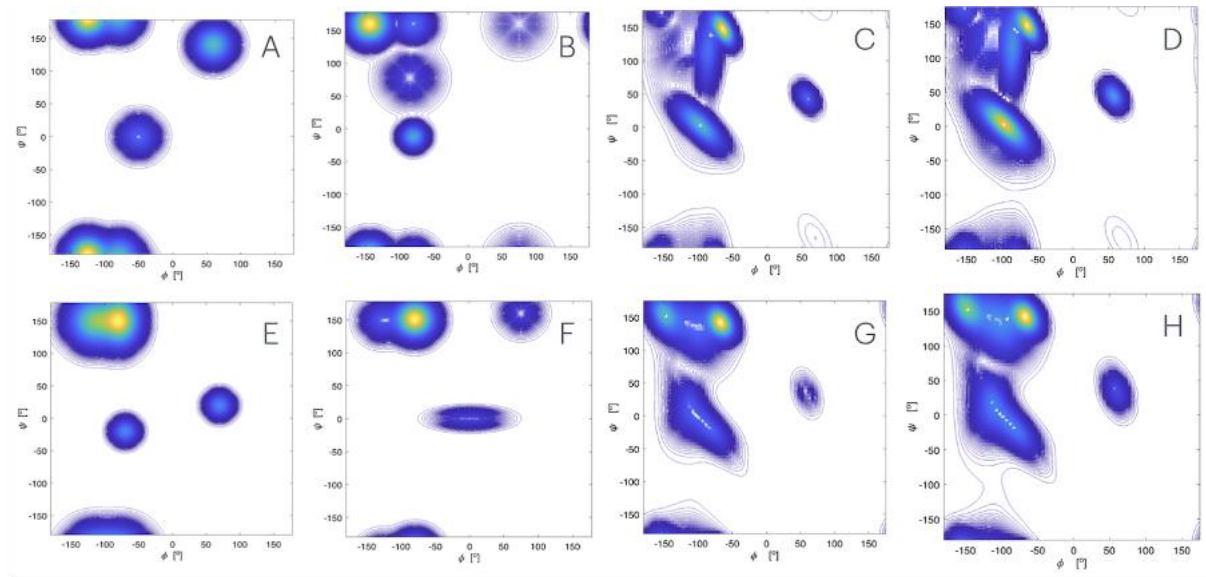

Figure 3: Ramachandran plot of the aspartic acid residue in cationic GDG (A), cationic GD**D**G (B), in the coil library segment G**D**-all and D**D**-all (C and D), in cationic GFG (E), cationic GD**F**G (F) and in the coil library segment G**F**-all and D**F**-all (F). The tripeptide Ramachandran plots were created with the Gaussian parameters reported in refs.(87,88), with slight modifications of the GDG mole fractions.(67) The plots for GD**D**G and GD**F**G were created by using the respective Gaussian parameters reported in earlier papers.(67) The plots for D and F in the above coil library segments were created with files obtained from the website of the Dunbrack group with the kind permission of Dr. Dunbrack.(81)

Comparison of peptide and coil library Ramachandran plots.

Figures 2 and 3 show the Ramachandran plots of a selected number of investigated residue pairs in GXYG and all-XY/XY-all segments of the coil library (all indicates the averaging over all 20 neighbours). The plots of the remaining tetra peptides and coil library segments investigated in this paper are shown in Figures S9-S16. In addition to the tetra-peptide plots, GXG plots for the respective target residues are shown, e.g. GAG for G**A**VG and all-**A**G for all-**A**V, where bold characters indicate the target residue.

A first superficial look at all the displayed Ramachandran plot reveals the following qualitative picture. In the investigated tetra-peptides, nearest neighbour interactions vary the pP2 <-> β equilibrium, though to a different extent (vide supra). For quite a substantial number of peptides, they increase the

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

distance between the two basins along the ϕ-coordinate (GS**A**G, GD**L**G, GS**L**G, G**L**LG, GV**L**G, G**S**AG, G**S**KG, and G**S**VG). These changes do not correlate with population changes of the pPII and β-strand. Turn supporting conformations, which are significant in the Ramachandran plots of D and S containing peptides, are significantly varied by their neighbours. Generally, interactions between like residues in homo-dimers (G**D**DG, GD**D**G, G**R**RG and GR**R**G) seem to be more pronounced than the ones in hetero-dimers. The coil dimer Ramachandran plots do not convey a similar story. While they clearly differ for different amino acid residues (just compare the all-XG plots), NNI induced changes seem to be rather limited. Most interestingly and also surprisingly, the data suggest a dominant population of the pPII basin for all residues with the exception of valine, for which the β-strand is more significantly populated, while the pPII basin has moved into a region which is generally associated with the ith residue of type II β-turns. The Ramachandran plots of most coil library dimers exhibit a significant distance between pPII and β-basins with the latter positioned outside of the canonical positions of residues in antiparallel and parallel β-sheets. This suggests a non-Arrhenius Kramers-type structure of the transition state.

The above description of peptide and coil Ramachandran is pretty much based on visual inspection. In what follows I use metrics introduced in the preceding section to provide a more quantitative comparison of Ramachandran plots.

Comparison of J-coupling constants

While the above discussed Ramachandran plots suggest significant differences between conformational distributions obtained from a Gaussian analysis of short peptides and those derived from coil library data of Ting et al., these discrepancies themselves do not per se show that peptide and coil library distributions are different, since the Ramachandran plots obtained for the former are model dependent. Fortunately, the Ramachandran plots of the coil library contain enough data points to allow for the calculation of ensemble averaged J-coupling constants. Figures 4 and 5 display the experimental and calculated J-coupling constants for alanine and aspartic acid containing tri- and tetra-peptides. For the remaining peptides, the corresponding figures are shown in Figures S17 to S22. Plots shown in the left column compare experimental values of the investigated tri- and tetra-peptides with corresponding coil library values. The right columns display plots compares Gaussian J-coupling constants with the ones derived from coil library distributions. In addition to the J-coupling constants of the coil library of Ting et al., the last bar in some of the panels visualizes corresponding J-coupling values Shen et al. derived from Ramachandran plots of an extensive coil library, which does not contain any residues involved in turn stabilizing hydrogen bonding.(78) Contrary to Ting et al. they solely reported backbone distributions for which the influence of both neighbours was averaged.

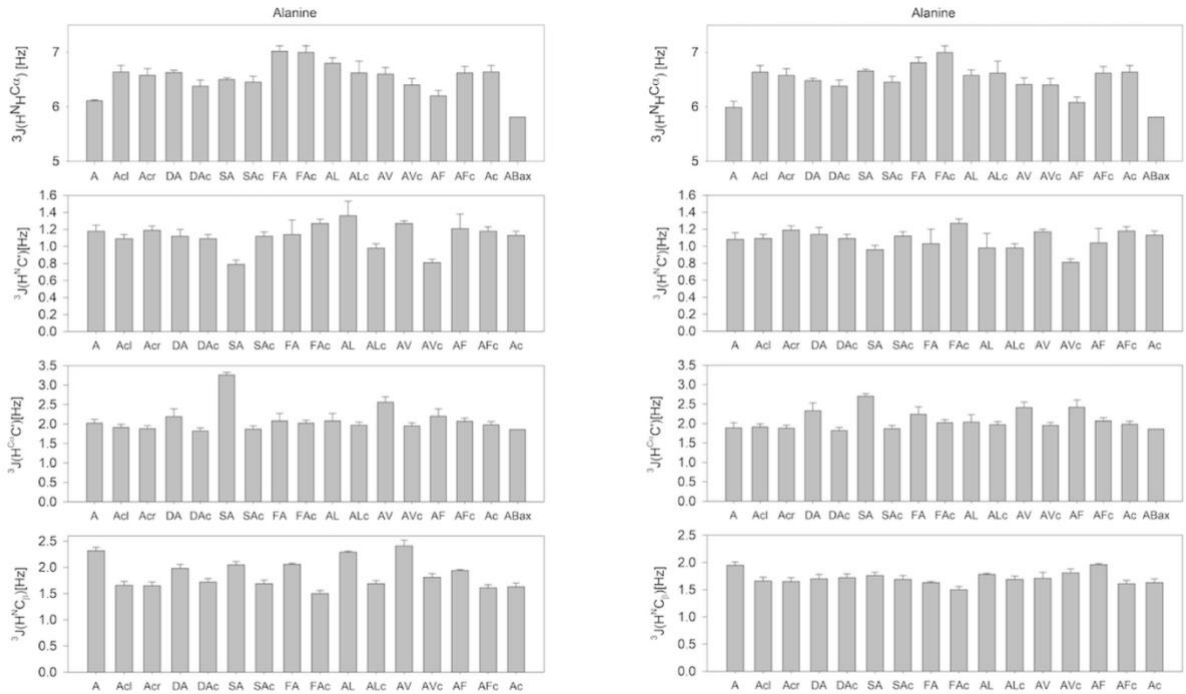

Figure 4: Comparison of experimental, Gaussian modelling and coil library J-coupling constants of alanine residues in different contexts. (Left): Comparison of experimental coupling constants in GAG (denoted as A), GD**A**G (denoted as DA), GS**A**G (denoted as SA), GF**A**G (denoted as FA), G**A**LG (denoted AL), G**A**VG (denoted AV) and G**A**FG (denoted as AF) with corresponding coil library constants (lc (left coil): all-**A**G, rc (right coil): G**A**-all, c(coil): all-**A**Y, and X**A**-all depending on the corresponding tetra-peptide sequence. Ac values were obtained by averaging over both neighbours. ABax indicates the coupling parameters inferred from a coil library of Bax and coworkers.(78) The respective Ramachandran plots were derived by averaging over all neighbouring residues. Experimental values were taken from refs.(66,72) (Right): Comparison of J-coupling constants obtained from Gaussian modelling of the indicated tri- and tetra-peptides and corresponding coil library constants. The notation is the same used for the figures in the left column.

Please do not adjust margins

Please do not adjust margins

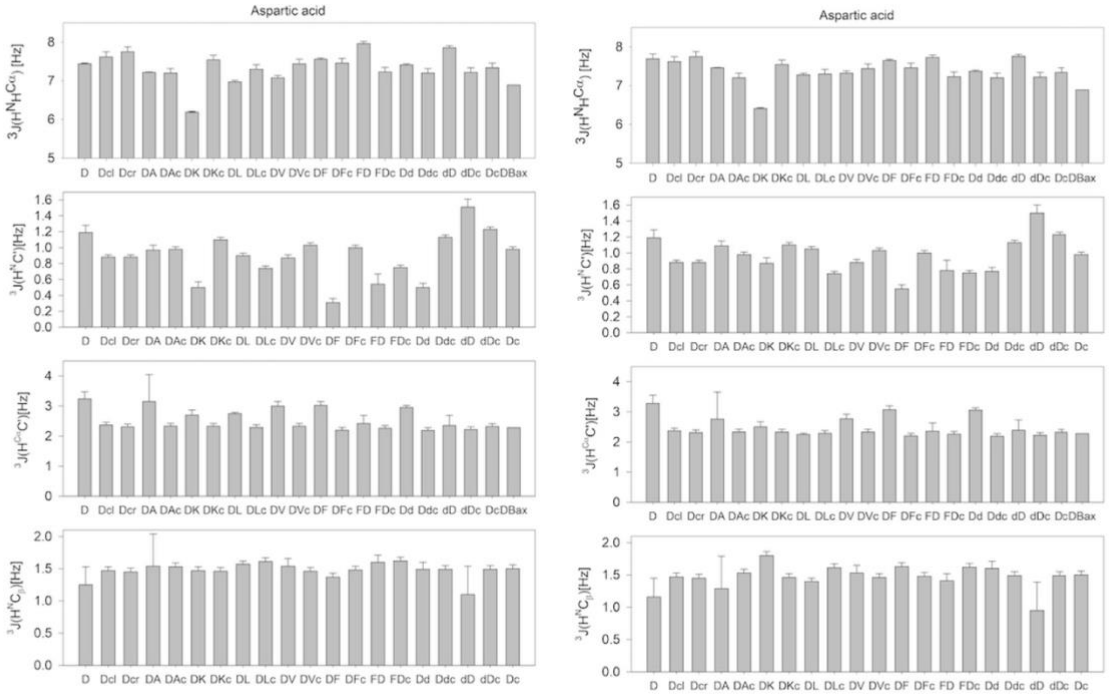

Figure 5. Comparison of experimental, Gaussian modelling and coil library J-coupling constants of aspartic acid residues in different contexts. (Left): Comparison of experimental coupling constants in GDG (denoted as D), G**D**AG (denoted as DA), G**D**KG (denoted as DK), G**D**LG (denoted as DL), G**D**VG (denoted DV), G**D**FG (denoted DF), GF**D**G (denoted as FD), G**D**DG (denoted as Dd) and GD**D**G (denoted a dD) with corresponding coil library constants (lc (left coil): all-**D**G, rc (right coil): G**D**-all, c(coil): all-**D**Y and X**D**-all, depending on the corresponding tetra-peptide sequence. D_c values were obtained by averaging over both neighbours.(81) DBax indicates the coupling parameters inferred from a coil library of Bax and coworkers.(78) The respective Ramachandran plots were derived by averaging over all neighbouring residues. Experimental values were taken from refs.(66,72) (Right): Comparison of J-coupling constants obtained from Gaussian modelling of the indicated tri- and tetra-peptides and corresponding coil library constants. The notation is the same used for the figures in the left column.

Let us first focus on the coupling constants of the Ting et al. library. For alanine, the most pronounced differences between experimental and coil library values are obtained for **A**F, S**A** and **A**V. For aspartic acid, profound difference between respective $^3J(H^N,H^{C\alpha})$, $^3J(H^N,C')$ and $^3J(H^{C\alpha},C')$ values are revealed for **D**K, F**D**, **D**F, **D**D and D**D**. Regarding the scalar coupling parameters depicted in the Supporting Information discrepancies between respective $^3J(H^N,C')$ values are noteworthy. In most cases the coil library values are significantly larger than the experimental and Gaussian values (**S**A, **S**L, **S**V, **S**K, K**L**, V**L**, S**V**, D**V** and **V**L). This reflects the very negative ϕ values of the β-strand basin in the respective coil library Ramachandran distributions. Generally, β-strand and pPII basins are in closer proximity in the plots created by Gaussian modelling. Overall, the coupling constants obtained from Gaussian modelling are closer to the experimental values than corresponding coil library values. This finding indicates that the distributions of residues in short peptides and in the Ting et al. library are different.

A visual inspection of the Ramachandran plots in Figures 1,2, and S6-S16 reveals that differences between corresponding coil library distributions are by far less significant than the ones between the Gaussian plots. This impression is put on a more quantitative basis by calculating the average experimental J coupling constants for the GXYG peptides with a common residue (e.g. G**F**AG, GA**F**G, G**F**DG and GD**F**G) and the corresponding coil library dimers (e.g. all-**F**A, A**F**-all, all-**F**D and D**F**-all). The respective values are listed in Table 1. First, a comparison with corresponding GXG values reveal residue specific differences for both data sets. Second, while differences between average coupling constant values are noteworthy, they are not as pronounced as the differences between the standard deviations, which for most of the considered sequences are larger for the experimental than for the coil library-based coupling constants. This result strongly suggests that compared with GXYG the side chain specificity of nearest neighbour interactions is significantly reduced in the Ting et al. coil library.

Please do not adjust margins

Please do not adjust margins

	$<^3J(H^NH^{C\alpha})>_e$ [Hz]	$<3J(H^NHC\alpha)>_c$ [Hz]	$<^3J(H^NC')>_e$ [Hz]	$<^3J(H^NC')>_c$ [Hz]	$<^3J(H^{C\alpha}C')>_e$ [Hz]	$<^3J(H^{C\alpha}C')>_c$ [Hz]	$<^3J(H^NC_\beta)>_e$ [Hz]	$<^3J(H^NC_\beta)>_c$ [Hz]
A	7.03±0.51 (6.11)	7.07±0.54	1.15±0.181 (1.18)	1.07±0.15	2.39±0.42 (2.02)	1.95±0.09	2.12±0.17 (2.32)	1.67±0.1
D	7.2±0.48 (7.44)	7.32±0.12	0.8±0.38 (1.19)	0.98±0.16	2.79±0.27 (3.24)	2.01±0.14	1.43±0.15 (1.25)	1.52±0.06
S	6.82±0.14 (6.99)	7.29±0.15	0.79±0.22 (0.87)	1.27±0.07	3.26±0.31 (2.33)	2.09±0.07	1.78±0.16 (1.71)	1.42±0.08
K	6.82±0.34 (6.63)	7.39±0.2	0.98±0.02	0.96±0.12	2.37±0.59	1.95±0.08	1.67±0.13	1.51±0.11
L	7.63±0.27 6.78	7.5±0.11	0.74±0.15 (0.84)	0.71±0.04	2.72±0.49 (2.45)	2.1±0.04	1.64±0.07 (1.75)	1.59±0.04
V	7.94±0.55 (7.46)	8.0±0.17	0.89±0.17 (0.91)	1.05±0.08	2.67±0.5 (2.33)	2.21±0.07	1.45±0.11 (1.59)	1.31±0.08
F	7.33±0.46 (7.45)	7.62±0.11	0.82±0.32 (0.74)	1.45±0.09	1.96±0.5 (2.2)	2.2±0.1	1.59±0.42 (1.79)	1.48±0.19

Table 1. Listing of average experimental (e) (66,67,72,73) and calculated J-coupling constants of the Ting et al. coil library (c).(81) The average was taken over the entire set of residue dimers in GXYG and coil library segments containing the indicated residue type. Coupling parameters of these residues in GXG peptides are listed in parenthesis for comparison.(87–89)

Compared with the Gaussian and Ting et al. coil library values, the all-average coupling constants of Shen et al. in Figures 3, 4, and S17-S22 exhibit common trends. The $^3J(H^NH^{C\alpha})$ values are generally closer to the respective peptide coupling constants, while the $^3J(H^{C\alpha}C')$ values are closer to the respective coil library set. The few Ramachandran plots reported in their paper exhibit much broader basins than the ones in the Gaussian and coil library plots presented in this paper.

Comparisons of mesostate populations

In the Gaussian model conformational propensities are expressed in terms of statistical weights (mole fractions) of the utilized Gaussian sub-distributions. Within the context of the model, this makes sense thermodynamically, since these Gaussian distributions correspond to a classical harmonic approximation of the respective potential energy minimum. However, this parametrization does not allow a quantitative comparison with Ramachandran plots from molecular dynamics simulations and coil libraries. As shown earlier for the use of the former, calculating the population of mesostates (Figure S2) provides a handy solution.(57)

The mesostate populations of the investigated residues in peptides and the Ting et al. coil library are shown in Figure S23. The populations of right-handed helices and β-turn I/IIi+2 conformations as well as the ones of antiparallel and parallel β-strand mesostates have been added up. For alanine, pPII dominates all distributions. For G**A**VG, the pPII population is significantly reduced, but it is still slightly larger than the combined β-strand confirmations, while the respective mole fractions are practically identical (Figure S4). The pPII mesostate of aspartic acid is comparatively weakly populated in GDG. Both, tetra-peptide and coil library data indicate that nearest neighbour interactions significantly increase the pPII population at the expense of asx-turns and βap. pPII is dominantly populated in the distributions of lysine and leucine; some exceptions notwithstanding (GSKG, GVKG, GLLG, GALG and GVLG). Most of the serine distributions exhibit comparable sampling of pPII and βap while most of the valine distributions are dominated by βt and βap (exceptions are DV-all, all-VL and KV-all). The sampling of the combined mesostate of right-handed helices and β-turn I/IIi+2 in the coil library distributions is significantly more pronounced than it is in the investigated peptides. Discrepancies between mesostate populations and the mole fractions displayed in Figures S4-S7 result from the fact that the Gaussian sub-distributions overlap and may contribute to more than one mesostate. This is certainly true for many pPII

Please do not adjust margins

Please do not adjust margins

and β-strand Gaussians with relatively high ψ-values, which overlap with the ε-mesostate. Generally, pPII sub-distributions may contribute to the βt mesostate. Since some ψ-values of residues decrease in the presence of non-glycine neighbours, the pPII sampling can increase even if the statistical weight of the sub-distribution decreases. Alternatively, a narrowing of e.g. the pPII sub-distribution reduces the overlap with other mesostates thus increasing the population of its own mesostate. While the statistical weights of Gaussian sub-distributions are thermodynamically meaningful, the population of mesostates appears to be a more informative metric for the characterization and comparison of conformational distributions since their positions in the Ramachandran plot do not vary, whereas they are adjustable parameters in the Gaussian analysis.

Figure 6 compare the changes of mesostate populations in the investigated tetra-peptides and corresponding coil library sequences caused by upstream and downstream neighbours. For the calculation of these changes, GXG peptides were used as a reference for the tetra-peptides, while all-XG and GY-all were used for all-XY and XY-all.

The bar diagrams in Figure 6 reveal a complex picture, from which I distil only the most significant differences and similarities between peptide and coil distributions. For alanine in tetra-peptides all pPII populations decrease, in particular in the presence of serine and valine. This observation is in line with the behaviour of the corresponding mole fractions (Figure S4), but it is in contrast to the pPII changes in the coil library plots of **A**F-all, **A**L-all, and all-**A**V. Similar discrepancies exist at least partially for the series of serine, lysine and valine. However, for aspartic acid and phenylalanine, pPII changes occur in the same direction. The changes obtained for aspartic acid are peculiar, in that nearly all pPII changes are positive, with all-DA as sole exception. For phenylalanine, the positive pPII changes are pronounced for FA and FD. In the coil library, the changes move in same direction, but much less pronounced.

As one can infer from Figure S22, the combined population of the right-handed helical and type II/I β-turn mesostates is much more pronounced in nearly all coil library Ramachandran plots even though it is generally reduced by nearest neighbour interactions in the presence of non-glycine neighbours. Its conformational sampling remains significant, though. In quite a few cases the differences between (Gaussian) peptide and coil library distributions reflect a stabilization of pPII in the later (cf. Figure 5).

It should be noted that the displayed changes do not add up to zero for the following reasons. First, the Ramachandran plots of GxG peptides exhibit substantial sampling of the so-called ε-region, which lies right above the lower boundary of the lower left quadrant of the Ramachandran plot. The population of this region is generally reduced in tetra-peptides. Second, the Ramachandran plots of coil library residues are more inhomogeneous than the Gaussian plots. While they display relatively sharp maxima at the basin positions (sometimes even sharper than the corresponding Gaussian maxima), the transition regions are more populated in the coil library distributions. Changes therein are not probed by the mesostates used for the plots in Figure S2.

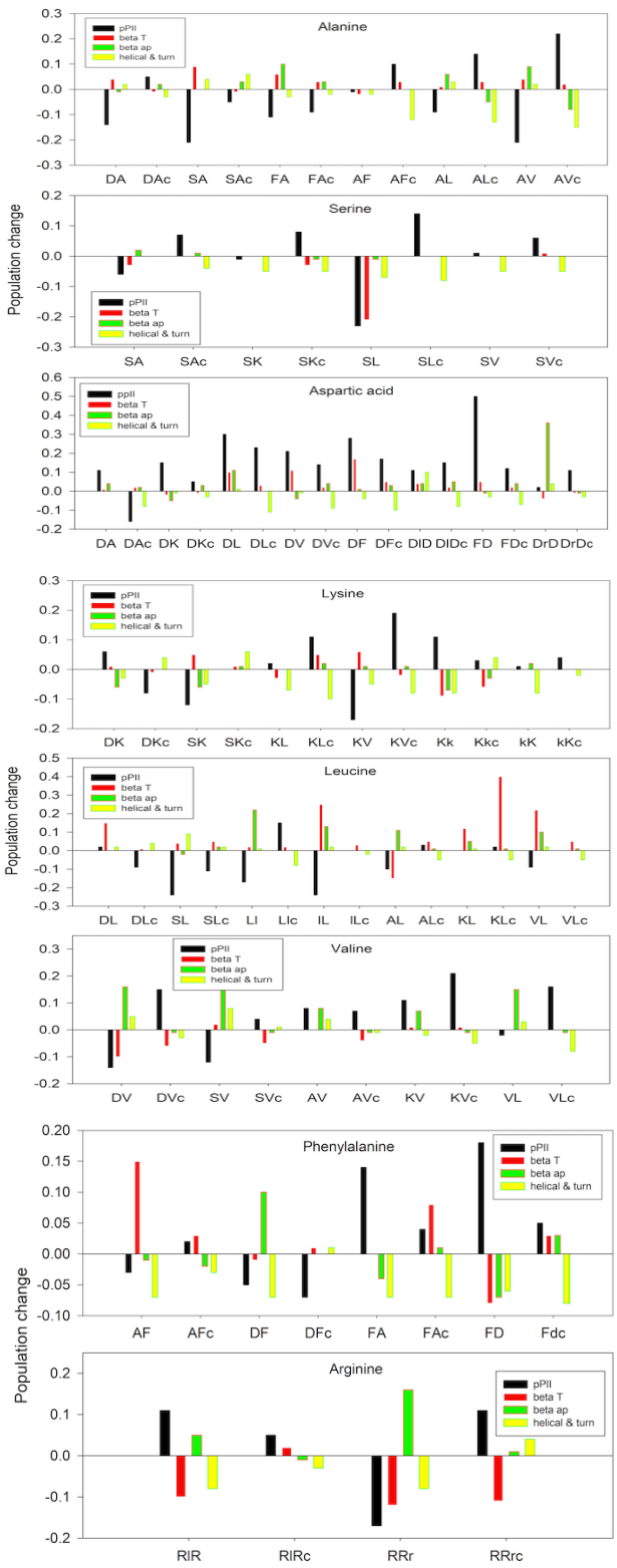

Figure 6 Plot of differences between the population of selected mesostates in GXYG and GXG (GYG) peptides and between populations in all-XY and all-XG as well as between XY-all and GY-all coil library segments. Note that the populations of mesostates associated with the parallel and antiparallel β-sheet structures as well as with helical and turn conformations have been added up, respectively.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

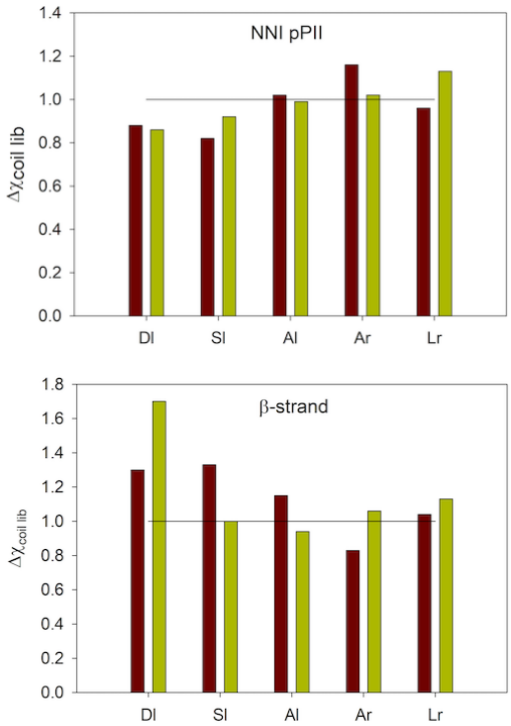

Figure 7. Graphic representation of the influence of the indicated amino acid residues on the average pPII (left) and β-strand (right) propensity as upstream (l) and downstream (r) neighbour. The dark red bars represent GXYG values, the green bars visualize the corresponding number of the coil library of Shen et al.(78)

In this context, it is illustrative to also compare the above GXYG peptides with the coil library of Shen et al.(78) The authors listed amino acid residue propensities only listed in relative terms. Their approach was guided by the assumption that an averaging over all down- and upstream neighbours in the utilized library would yield propensities that represent the coiled state of an IDP. They reported reduced propensity values calculated as:

$$\chi_{i,c}^{r} = \frac{\chi_{i,c}}{\langle \chi_{i,c} \rangle_i} \quad (7)$$

The numerator is the average value for a given amino acid residue *i* adopting a conformation *c.* The average in the denominator of eq. (7) was taken over all residue types. Hence, a value of 1 implies that a propensity for a distinct conformation of an amino acid type corresponds to its average value of the coil library. Figure S24 depicts the differences between respective reduced propensity values for pPII and β-strand. In line with the corresponding Gaussian and the Ting et al. distributions alanine shows the by highest pPII propensity; the reduced value for the Shen et al. library is only slightly lower than the corresponding GXG value. Interestingly, the Shen et al. coil library also exhibits higher reduced pPII values for valine, aspartic acid, glutamic acid, isoleucine, lysine and leucine. This seems to be counterintuitive for the aliphatic residues of this list, but it is at least in qualitative agreement with propensities in the library of Ting et al. (Figure 6).

In order to compare nearest neighbour effects in GXYG and the Shen et al. library, the influence of D, S, A and K on the reduced propensities of their downstream neighbours as well as of A and L on their upstream neighbours in GXYG was calculated as the ratio of the average propensity in the presence and the absence of the considered neighbour. The individual red bars were calculated as ratio of the calculated averages in the presence (G**X**YG and GX**Y**G) and the absence of interacting neighbours (GXG). The green bars in Figure 7 represent corresponding changes of the average pPII and β-strand propensity of all residues in the coil library of Shen et al. The pPII values are very similar for Dl, Sl and Al (for the notation, see the figure caption). Downstream alanine produces more pPII in GXYG than in it does in the coil, while it is the other way round if leucine is the downstream neighbour (*vide supra*). Differences between β-strand values are more pronounced (significantly more favored in the coil library for Dl, and Ar), while more favored in GXYG for Sl and Al.

Taken together, a comparison of conformational mesostate occupations in GXYG and in coil libraries reveal major differences. Overall, the nearest neighbour induced changes in the later seem to be less pronounced. Interestingly, however, they are also sometimes qualitatively different. This notion particularly applies to pPII the population of which is enhanced by aliphatic neighbours in coil libraries (V and L), whereas the same type of neighbour decreases the pPII population in GXYG. Right-handed helical and type I/II β-turn i+2 conformations is generally more populated in the Ting et al. library. The comparison with the Shen et al. library provides some interesting insights but lacks specificity since propensities are only expressed in relative terms after averaging over all neighbours.

Comparison of Hellinger distances and configurational entropies

Eqs. (3) was used to calculate the Hellinger distance between the Ramachandran plots of GXG (GYG) and GXYG/GXYG for all investigated amino acid residue targets. In addition, Hellinger distances between plots of all-XY/XY-all and all-XG/GY-all were calculated. As clearly indicated by the bar plots in Figure 8 the Hellinger values obtained for the Ramachandran plots of the tetra-peptides exceed by far those of the Ting et al. library. With the exception of the serine series, all Hellinger values of tetra-peptide residues indicate a moderate or large dissimilarity between corresponding Ramachandran plots of tetra- and tripeptides. On the contrary, the overwhelming majority of the values obtained for the coil library sequences lie in the moderate similarity region.[‡] As indicated above, Hellinger distances are very sensitive to changes of basin positions and respond only moderately to changes of basin populations. Thus, the data plotted in Figure 8 suggest minimal changes of basins in the coil library, in line with what one would infer from a visual inspection of Figures 1, 2 and S9-S16. Altogether, the plots in Figure 8 clearly suggest that nearest neighbour induced changes are much less pronounced in the Ting et al. library than in the investigated GXYG peptides.

Please do not adjust margins

Please do not adjust margins

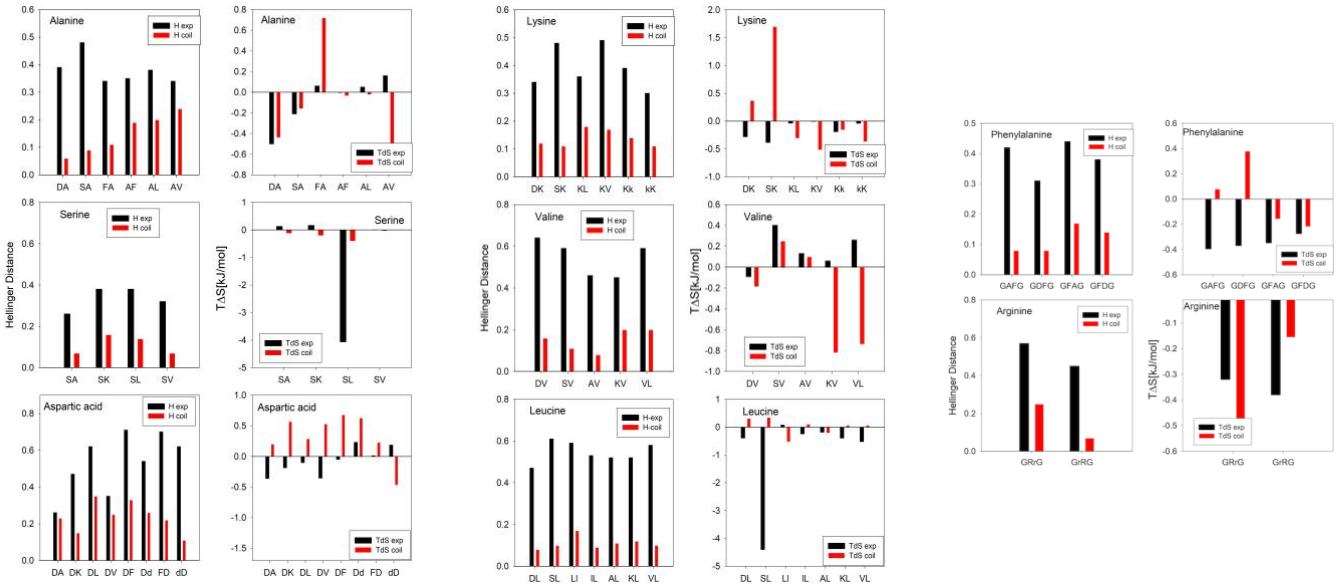

Figure 8: Hellinger distances and configurational entropy contribution to Gibbs free energy changes at room temperature due to nearest neighbour interactions for the indicated amino acid residues in GXYG peptides (black) and corresponding all-XY and XY-all segments of the Ting et al. library (red).

Since nearest neighbour interactions affect Ramachandran distributions, the configurational entropy changes as well. The corresponding changes of the Gibbs energy contribution are also plotted in Figure 7. Note that the depicted values were calculated for the indicated residue, not for the entire dimeric XY segments. Negative values for a distinct residue X(Y) imply that its Ramachandran plots is less entropic in the presence of the neighbour Y. The displayed TΔS values suggest a by far more complex picture than the Hellinger distances. While entropy changes in the tetra-peptides are dominant for some residues, it is the other way round for others (**A**F, **A**V, **V**L, **K**V, K**V**, **V**L, **R**R and quite a few D containing peptides). Even more surprising is the observation that entropy changes carry opposite signs for all serine neighbours and nearly all aspartic acid neighbours (**S**L and **D**D being an exception). Opposite behavior is also observed for e.g. K**V**, **V**L, S**K**, and D**F**. These findings suggest that relatively short side chains with hydrogen bonding capacity such as aspartic acid and serine behave differently as neighbours and as targets in short peptides and in coil libraries.

For most GXYG peptides, nearest neighbour interactions decrease the entropic contribution to the Gibbs energy of their neighbours by several hundreds of Joule/mol. G**S**LG and GS**L**G stand out in that the respective entropic Gibbs energy changes are in the kJ/mol regime for both non-terminal residues. This is quite astonishing. A closer look at the Ramachandran plots in Figures S9 and S11 informs us about the respective reason. In the case of GS**L**G a very broad basin composed of overlapping pPII and β-strand populations in GLG is replaced by two clearly separated basins giving rise to much narrower sub-distributions. Compared with GSG, basin separation in the G**S**LG plot could again be the source of the entropy reduction.

Thus, these data suggest that a coil state of a proteins might carry less configurational entropy than assumed based on the (local) random coil model. However, the mixture of entropy increases and decreases obtained for the Ting et al. coil library does not allow for such a clear assessment. A more detailed discussion of this issue is given in the subsequent paragraph.

Discussion

Nearly thirty years ago an analysis of $^{3}J(H^{N}H^{C\alpha})$ coupling constants of unfolded proteins lead to the conclusion that Flory's isolated pair hypothesis is not a valid supposition.(90) Subsequently, the influence of nearest and non-nearest neighbours on the conformational propensity was mostly inferred from the analysis of coil libraries(60,61,63,78,80,81,86,91,92) and from computational work on model peptides(69,92–94). The general view that emerged from these analyses suggested that aliphatic and aromatic neighbours decrease a residue's pPII propensity. For alanine, Avbelji and Baldwin attributed this to conformational dependent changes of the solvation free energy.(93) Computational work of the Sosnick group revealed that the observed nearest neighbour interactions involve correlations between the dynamics of interacting residues, which implies that these interactions depend on the adopted conformation of the respective neighbours.(92) This is an important information since a violation of the isolated pair hypothesis requires such correlation effects. A conformational detuning that depends solely on the side chain properties of neighbours would not eliminate the additivity of conformational and solvation free energies and would therefore not lead to a violation of the isolated pair hypothesis.

Experimentally, the issue of nearest neighbour interactions was early recognized by the NMR community. Chemical shifts and to a lesser extent $^{3}J(H^{N}H^{C\alpha})$ coupling constants in short glycine-based host-guest peptides were considered as ideal references for a random coil conformation. However, the context

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

dependence of these parameters in proteins led to the conclusion that their dependence on the choice of the neighbour has to be taken into account.(95,96) All attempts to resolve this issue without the need of investigating a total of at least 400 amino acid pairs were based on the assumption that nearest neighbour effects depend solely on the neighbour, not on the targeted residue itself. This notion implies for instance that valine as neighbour has the same influence on alanine and aspartic acid. Along this way, Schwarzinger et al. investigated the influence of the X-residue on the adjacent glycine residues in Ac-GGXGG-NH2.(97) The use of glycine as representative target was later challenged by Kjaergaard and Poulsen, who introduced glutamine as a more useful alternative.(98)

The first experimental work aimed at exploring in more detail the residue specific nearest neighbour influence on chemical shifts and $^3J(H^NH^{C\alpha})$ in short peptides was carried out by Jung et al.(99) These authors undertook the incredible task of measuring the $^3J(H^NH^{C\alpha})$ coupling constants and H^N and $H^{C\alpha}$ chemical shifts of 361 Ac-XY-NH_2 blocked dipeptides. They subjected their data solely to a statistical analysis in that they calculated the average coupling constant and the corresponding standard deviation for a given target amino acid residue. Only coupling constants that were outside of this standard deviation were considered as reflecting nearest neighbour coupling. However, using the standard deviation as a criterion for statistical significance is problematic in this context unless it reflects the statistical errors of the experimental values, which are generally small.(72) Nevertheless, it should be noted that neighbour induced changes of the J-coupling constant within a given target series is less pronounced in the dipeptide data set than it is in GXYG. Significant differences between corresponding coupling constants are noteworthy for valine and leucine, mostly at the C-terminal position (A**L**, A**V**, **V**L, K**V**, with L**L** as an exception; cf. Figure S25). One might be tempted to rationalize such differences in terms of end effects, i.e. different influences of the terminal NH^{3+} and COOH groups of GXYG and the blocking groups the dimers used by Jung et al., but such an explanation seems to be at variance with the experimentally based notion that the conformational propensity of blocked dipeptides with a single aliphatic residues and two peptide groups resemble the one of the respective guest residue in GxG.(56) However, one has to note a peculiarity of the peptides used by Jung et al. Generally, dipeptides contain a single amino acid flanked by two peptide groups with methyl groups attached to the carbonyl carbon and the amide nitrogens. They can be regarded as suitable mimics of glycine. A corresponding system with two residues would actually be a blocked tripeptide with three peptide groups and terminal methyl groups. On the contrary, the C-terminal residue of the peptides used by Jung et al. are linked to a primary amide which differ from secondary amides in multiple respects. First, their electronic structures can be expected to be different. In water, the amide I wavenumber (predominantly a C=O stretching mode) of a primary amide is significantly blue-shifted by approximately 20 cm^{-1} with respect to its position in secondary amide spectra (acetamide versus N-methylacetamide).(100) This suggests a stabilisation of the C=O double bond over C=N in the peptide's resonance structure. Second, the organisation of water in the hydration shells of primary and secondary amides should be different, since the terminal NH_2-group can donate two hydrogen bonds to water, which owing to the connectivity of hydration water of peptides can be expected to influence the conformational propensity of the C-terminal residue. It should be noted in this context that backbone and side chain hydration has been proposed as the key determinant of individual conformational propensities as well as of nearest neighbour interactions.(51,52,93,101–106)

In this context it must be emphasised that a structural analysis of short peptides just based on $^3J(H^NH^{C\alpha})$ is highly problematic, since its value depends on the mole fractions of basins, their width and their position along the ϕ-axis. Shi et al., who used this coupling constant to analyse the pPII - β-strand equilibrium of a guest residue in AcGGXGG-NH_2, tried to remedy this conundrum by using the basin positions in a rather unrestricted coil library produced by Avbelj and Baldwin.(107) The comparison of Ramachandran plots of the Ting library with the ones of tetra-peptides in this paper cast some doubt on the suitability of this a strategy.

The study of Jung et al. was followed by the above-described work of Toal et al.(72) on GXYG peptides dealt within this paper. Compared with the former, it is rather limited in scope (15 peptide, later augmented by another 8 GXYG and 2 GXYZG peptides by Milorey et al.(66,67,73)), but the data basis for each peptide is much broader (four ϕ- and one ψ-dependent scalar coupling constants and four different amide I' profiles in different vibrational spectra). This large data set allowed for the construction of Ramachandran plots in terms of a superposition of two-dimensional Gaussian sub-distributions. While certainly still a simplification, the modelling of GxG peptides with the Gaussian model yielded a far better agreement with experimental data than molecular dynamics simulations with rather different force fields and water models.(57,104) Hence, it is justified to use the obtained GXYG Ramachandran distributions as a suitable approximation for a comparison with data sets obtained from restricted coil libraries. The use of the latter is based on the assumption that a large ensemble of residues located in ordered but not regularly structured segments of proteins produce conformational distributions representative of the coiled state of unfolded and disordered peptides and proteins. Restrictions to which even residues in structured loops of proteins are subjected are being thought to average out.

One of the main problems one encounters with the use of coil libraries is that even the enormous number of protein structures used for assembling the latest version of coil libraries are insufficient to even explore the total of 8000 XYZ-segments which would account for exploring nearest neighbour interactions between all twenty amino acid residues. In order to address this issue, researchers average over all down- and upstream neighbours to obtain a sufficient number of entries to their Ramachandran plots. Bax and colleagues used this strategy to obtain a set of secondary chemical shifts and scalar coupling constants considered as a good approximation for the coiled state of peptides and proteins.(77,78) Ting et al. dealt with the limitations of their data set by an interpolation procedure called the hierarchical Dirichlet process.(81,108) Thus, they could produce Ramachandran plots with a 5° resolution in ϕ and ψ direction. Even with this approach they could not explore the full set of XYZ combinations, but they could restrict residue averaging to either the upstream or downstream neighbour of a selected target residue, which lead to Ramachandran plots for X**Y**-all and all-**X**Y.

Please do not adjust margins

Please do not adjust margins

The comparison of GXYG and coil library distributions revealed similarities, but also significant differences. In line with results on alanine containing peptides, the coil library confirms the comparatively high propensity of this residue type for pPII. Qualitatively, the coil library distribution also exhibits the relatively low pPII and the high turn (right-handed helical and type II/II i+2) propensities of aspartic acid. The relatively high β-strand propensity of valine is also reproduced. However, these similarities are more than outweighed by differences due to nearest neighbour interactions. They are discussed in detail in the following.

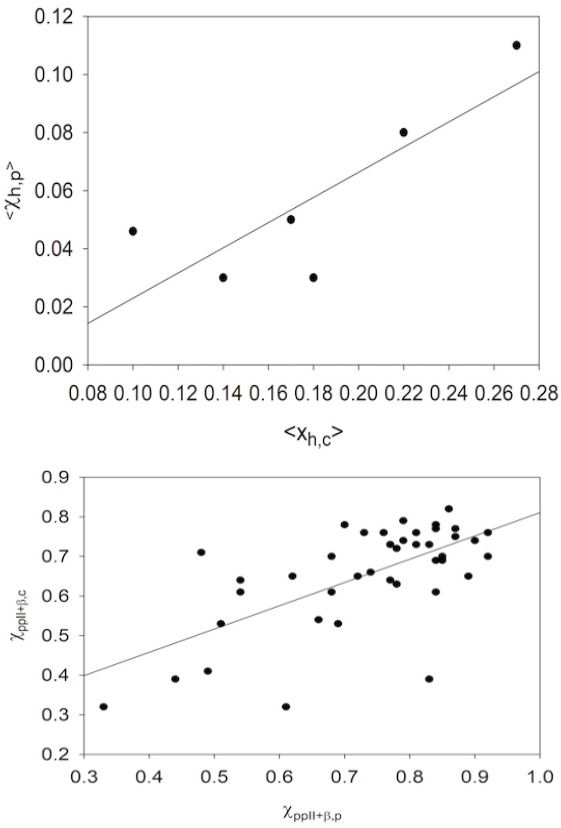

Figure 9. Correlation plots comparing mesostate populations in GxyG and the Ting et al. coil library.(81) Upper panel: Comparison of the combined population of right-handed α-helix and type I/II β-turn i+2 mesostates in GXYG (p) and the coil library (c) averaged over all target neighbours of A, V, S, D, L and F. Lower panel: Comparison of the combined population of pPII and all β-strand mesostates in GXYG (p) and the coil library (c). The solid lines result from a regression analysis described in the text.

In order to shed some light on the differences between tetra-peptide and coil library distributions, the average population of the combined mesostates of right-handed helical and the very similar type I/II β-conformation in GXYG and the Ting et al. library were compared to each other. Since the respective numbers are very low for the former the averaging was carried out over all neighbours in the investigated GXYG set. The result of this analysis is visualized by the left plot in Figure 9. The regression analysis yielded a slope of 0.433±0.12 and a Pearson correlation coefficient value of 0.81. This value is indicative of a strong correlation, but one should keep in mind that the data basis of the plot is limited. The obtained correlation nevertheless indicates that the population of the right-handed helical/type I/II β region, which are merged in many Ramachandran plots, is by about a factor of two higher in the coil library, in spite of the relative to glycine negative influence of neighbours. Such a population enhancement is not surprising. One of the limitations of the use of short peptides as model systems for a statistical coil state of a protein is that the population of the right-handed helical region is underestimated compared with the situation in polypeptides and proteins. All theories of helix ⇔ coil transitions invoke a cooperative process where the helix formation goes through a nucleation process. The mole fraction of right-handed helices in GXG peptides or blocked dipeptides should be closely related to the nucleation parameters σ and v of the Zimm-Bragg and Lifson-Roig parameters, respectively.(87) With an increasing length of an oligo-peptide, the formation of short transient helices becomes more likely. The total fraction would depend on the respective values of the helix propensity parameters s and w. For oligo-alanines, the increase of the helical fraction with the number of residues has been predicted by MD simulations,(51,109) while experimental results are somewhat conflicting.(110,111) Generally, it is permissible to state that the increased population of right-handed helices and very similar turn-supporting structures in the coil library of Ting et al. is something to be expected. The above scaling factor of 2.3 should be taken as an average subject to the above indicated significant standard deviation of ±0.6.

If one ignores other less populated β- and γ-turn structures as well as asx-turns and left-handed helices the result of the above analysis suggests most of the differences between nearest neighbour interactions in GXYG and coil library segments involve redistributions between pPII and β-strand. If this supposition is correct, the total fraction of extended mesostates in the two systems should exhibit some correlation. The right plot in Figure 9 compares the two fractions for all the pairs investigated experimentally. The Pearson coefficient of 0.64 still indicates a medium level correlation. However, it should be noted that the data basis for this correlation plot is much stronger than the one used for the comparison of right-handed helical populations. The slope of the correlation plot for the extended structures is 0.58±0.15, which suggests a reduced variability of the extended structure population in the coil library.

The plots in Figure 6 reveal that nearest neighbour interactions in the coil library enhance the pPII population more than they do in the investigated GXYGs. The **A**V dimer in GXYG and all-AV exemplifies this difference, in that valine substantially reduces the pPII population of alanine in G**A**VG, while it enhances it in all-**A**V. In order to check whether an upstream valine would have the same influence on alanine we analysed the mesostate population of alanine in V**A**-all. The obtained populations of A are nearly identical with the ones of G**A**-all. This result suggests that at least for alanine the apparent influence of upstream neighbours is nearly eliminated by the averaging over all downstream neighbours. I checked this supposition by calculating the mesostate populations for all-X-all for all X-residues investigated in this study. For most of the later, the differences between all-X-all and GX-all were very small (< 5%). Only X=D showed a slight enhancement of pPII at the expense of type I/II β-turn.

In order to illustrate the possible influence of averaging over downstream neighbours I performed the following very simple

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

calculation. It was assumed that the target residue in XY-all can adopt only pPII and β-strand conformations. The respective mole factions were calculated as follows:

$$\chi_\beta = \sum_j [\frac{e^{-\Delta G_y + dG_{xy} + dG_{yj}}}{NZ_j}] \quad (8a)$$

$$\chi_{pPII} = 1 - \chi_\beta \quad (8b)$$

where the partition sum Z_j is written as:

$$Z_j = 1 + e^{-\Delta G_y + dG_{xy} + dG_{yj}} \quad (9)$$

ΔG_y denotes the intrinsic Gibbs free energy difference between pPII and β-strand of residue Y, dG_{xy} is the free energy change caused by interaction with the upstream residue X and dG_{xj} represents the free energy change caused by interaction with the j*th* downstream residue. N=20 is the number of residues. The distribution of interaction energies dG_{xj} was heuristically described by a Gaussian function:

$$dG_{xj} = dG_{xj}^0 e^{-(j-j_0)^2/\sigma_j^2} \quad (10)$$

j_0 is the position of the maximum the choice of which is arbitrary, σ_j is the standard deviation. All free energies are written in units of RT. Two scenarios, one with a high (0.67) and the other one with a low intrinsic pPII propensity (0.4) were considered. The correspond free energy differences are 0.7 and -0.4. Next, a dG_{xy} value of 0.3 was added to account for the general tendency of downstream neighbours. It reduced the corresponding pPII population to 0.59 and 0.33 (*vide infra*). In a third step, the Gaussian term in eq. (10) was added. A variation of the free energy term and of the standard deviation revealed that dG_{xy}^0 =-0.8, j_0=5 and σ_j=5 reproduced the assumed intrinsic propensities for both rather different scenarios. The results of this very heuristic calculation show that the averaging of Ramachandran plots of a distinct Y-residue over all natural downstream neighbours can practically neutralize the weaker nearest neighbour interactions with the upstream residue X. This would not happen if only nearest neighbour interactions with individual upstream residues would be considered. The reader should note that the rather large standard deviation of the utilized Gaussian distribution, which reflects an averaging over a relatively wide range of negative free energies.

The results of the above analysis suggest an asymmetry between nearest neighbour interactions in the coil library in that the interactions with upstream neighbours dominate over the ones with downstream neighbours. This notion is corroborated further by the observation that out of 26 residue pairs for which the coil library data indicate a nearest neighbour induced increase of the pPII population, 20 are of the all-**X**Y type, while 6 out of 7 neighbours that reduce the pPII population of the target residue are at the upstream position (X**Y**-all type).

The above unbalance between interactions with down- and upstream neighbours is still present in GXYG peptides but to a lesser extent. In most peptides, nearest neighbour interactions reduce the pPII population. In 12 out of 19 cases this involves the respective upstream neighbours. pPII is enhanced in 14 cases, from which 9 involve downstream neighbours.

Another peculiarity of the coil library distributions is the rather large separation between the visible pPII and β-strand peaks in nearly all distributions. While the respective distances are often larger in GXYG than in respective GXG/GYG plots, they are not really close to the ones in the coil library. Figure S26 illustrates this point by depicting the ϕ dependence of the probability density for the serine residue in all-**S**L at ψ=150°. The plot reveals actually revealed the existence of an additional very shallow maximum at 110° and of course the aforementioned dominance of pPII which is closer to the one in the corresponding GSLG plot (Figure S11). Similar tendencies can be seen in Ramachandran plots of the coil library used by the Sosnick group.(76) Interestingly, a very detailed coil library investigation of Jiang et al. revealed that the distance between the two basins depends on the χ_1 angle of the side chain, where the g^- rotamer seems to produce the largest distance (their plots also represent an average over all neighbours in the library).(112) Hence, the discrepancy between coil library and peptide basin position noted in this study might actually reflect different populations of side chain rotamers. Needless to say, that this would further complicate the picture. Sosnick and coworkers actually provided computational evidence for coupling between side chain and backbone dynamics.(63)

In view of the differences between the neighbour dependence of residues in GXYG peptides and respective coil library dimers the question arises which of the two data sets are more useful for the identification of coil states. Being able to do so is necessary for the identification of residual structures in unfolded and disordered proteins.(77,78,113,114) The availability of a large data set for all natural amino acid residues in the coil library of Ting et al. is a clear advantage over the limited number of GXYG peptides that have thus far been investigated. The comparison of coil library and GXYG distributions has thus far revealed that averaging over neighbours might artificially diminish nearest neighbour effects and of course eliminates the side chain specificity of such interactions for a target residue. It is frequently believed that the differences caused by the latter are statistically insignificant. To check whether such a notion has merits I had a closer look at the reported $^3J(H^NH^{C\alpha})$ constants of two IDPs that are generally thought to be in an ideal coil state, namely the amyloid peptide $A\beta_{1-40}$(115) and α-synuclein (115). Table 2 list the J-coupling constants of valine, phenylalanine, glutamic acid and alanine. The listed coupling constants of α-synuclein were taken from the C-terminal segment between residue 90 and 140, which does not exhibit any transient secondary structures. Note that experimental uncertainties of $^3J(H^NH^{C\alpha})$ are very small and do rarely exceed 0.1 Hz.(72,87,89) Since the values for α-synuclein had to be taken from a graphic representation, the corresponding error might be slightly larger. The data listed in Table 2 show that the coupling constant of valine varies within a range of 1.2 Hz. The corresponding intervals of alanine, glutamic acid and phenylalanine are 0.8, 1.4 and 0.7 Hz. Interestingly, while coupling constant values of valine scatter around the corresponding GVG value, the coupling constants for A and E are mostly lower, in line with an earlier observation for the tau protein.(71,114) This might reflect a higher degree of sampling of right-handed helical conformations. The Ting et

Please do not adjust margins

Please do not adjust margins

al. library value for the corresponding all-F-all distributions is close to the average value of F (7.46 Hz). For all-A-all and all-V-all the Ting library values are significantly larger than the average (5.2 and 7.55 Hz, respectively). Interestingly, the Shen et al. J-coupling values are close to the respective average for valine, glutamic acid and phenylalanine, while the corresponding alanine value still significantly exceeds the average of the alanine values in Table 2. These findings might suggest that the Shen et al. library does indeed represent the average behaviour of amino acid residues in the coil state. However, the data set in Table 2 is insufficient to arrive at definite conclusion at the present time. What can be stated is that the residue specificity of nearest neighbour interactions is too pronounced for the employed averaging to yield coupling constants that are representative of a coiled state of an IDP.

Valine	Phenylalanine	Glutamic acid	Alanine
6.78 (E**V**H)[1]	7.02 (E**F**R)[1]	6.3 (**E**F)[1]	5.6 (F**A**E)[1]
8.3 (L**V**F)[1]	7.7 (V**F**F)[1]	6.29 (Y**E**V)[1]	5.5 (G**A**I)[1]
7.49 (M**V**G)[1]	7.74 (F**F**A)[1]	5.95 (A**E**D)[1]	5.1 (S**A**A)[2]
7.94 (V**V**F)[1]		6.8 (N**E**E)[2]	5.0 (A**A**A)[2]
7.94 (G**V**V)[1]		6.6 (E**E**G)[2]	5.0 (A**A**T)[2]
8.02 (V**V**F)[1]		6.2 (Q**E**G)[2]	5.9 (E**A**T)[2]
7.2 (M**V**D)[2]		7.4 (Y**E**M)[2]	
7.46 (G**V**G)[3]	7.45 (G**F**G)[3]	6.99 (G**E**G)[3]	6.1 (G**A**G)
8.1 (all-**V**-all, Ting et al.)[4]	7.5 (all-**F**-all, Ting et al.)[4]		6.64 (all-**A**-all, Ting et al.)[4]
7.73 (all-**V**-all, Shen et al.)[5]	7.23 (all-**F**-all, Shen et al.)[5]	6.55 (all-E-all, Shen et al.)[5]	5.81(all-**A**-all, Shen et al.)[5]

Table 2. List of $^3J(H^NH^{C\alpha})$ constants in units of Hz of the indicated amino acid residue in [1]A$\beta_{1\text{-}40}$,(116) [2]α-synuclein (residues 89-140),(115) [3] GXG peptides(87–89) and the coil libraries of [4]Ting et al.(81) and [5]Shen et al.(78) Values for the latter were obtained by averaging over both neighbours.

Summary and Outlook

This paper presents a comparison of the influence of nearest neighbours on a selected number of amino acid residues in cationic tetra-peptides and the coil library of Ting et al. The way how the latter has been obtained allowed us to directly compare amino acid pairs and the different influence of upstream and downstream neighbours. However, while these pairs are flanked by glycines in the tetra-peptides, averages were taken over all neighbours of the consider pairs. As a consequence, either the influence of the upstream or of the downstream neighbour is explicitly taken into account. Ramachandran plots were compared by means of different metrics; i.e. mesostate populations, Hellinger distances, configurational entropy and several scalar J-coupling constants. Differences between GXYG and coil library distributions are noteworthy. Generally, nearest neighbour interaction induced changes are more pronounced in GXYG. For quite a noticeable number of pairs, the neighbours reduce the pPII propensity of amino acids in GXYG, while they increase it in the coil library. The averaging over either upstream or downstream neighbours reduce the specific influence of the opposite neighbours. These findings and a comparison with a limited number of $^3J(H^NH^{C\alpha})$ constants of coiled IDPs led to the conclusion that the residue specificities of nearest neighbour interactions should not be ignored. As a corollary of this conclusion, one has to recognize that identifying the coil state of a protein and/or an IDP is a complicated task. If one took short peptides as a starting point, one would have to investigate a total of 8000 GXYZG peptides just for a thorough assessment of nearest neighbour interactions. The exploration of non-nearest neighbour interactions would require an even larger number of hexa- and even hepta-peptides. At present, this would not be a realistic perspective. With regard to nearest neighbour interactions, progress might be achievable by subjecting the complete data set of the Ting et al. library to a global analysis, which would

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

have to be based on the assumption that the interaction energies associated with the two neighbours are additive. We know that this is not the case, but the results of such an analysis might at least allow the construction of suitable approximations of coil state distributions. A possible experiment-based algorithm might use reported correlations between pPII and β-strand populations of neighbours,(70) but they would have to be augmented by similar respective relationships between these and right-handed helical conformations. In principle, conducting molecular dynamics simulations would be a more promising venue. However, moving along this path requires combinations of force field and water models, which can reproduce intrinsic propensities as well as nearest neighbour interactions. Particularly regarding the latter, current approaches are still in their infancy.(57,58,65,73,117) The GXYG peptides used for this study should be used as benchmarks for the necessary force field development.

Why is it necessary to gain a comprehensive picture of nearest neighbour interactions? First, as already indicated above, it would be necessary to obtain an accurate, residue specific picture of the statistical coil state of proteins, which can be used as a baseline for identifying the occurrence transient secondary structures. In the field of NMR spectroscopy, the Blackledege group developed the so-called flexible meccano program which they used to extract a set of possible conformational ensembles for a given polypeptide/protein sequence by exploiting conformational distributions of amino acid residues in a restricted coil library.(113,118–120) Such ensembles are then used as a starting point for the structure analysis of IDPs by means of the ASTEROIDs algorithm (a selection tool for ensemble representations of intrinsically disordered states) which adjusts the conformational ensemble to constraints imposed by secondary chemical shifts, J-coupling constants (mostly $^3J(H^NH^{C\alpha})$), residual dipole coupling and experimentally obtained radii of gyration. From the description of flexible meccano in the literature (120) it is obvious that nearest neighbour interactions are solely considered by averaging. As already argued in an earlier paper, this might lead to misidentification of transient secondary structures in IDPs such as the tau protein.(71) Second, as already shown in an earlier study on denatured proteins by Jha et al.,(61) correlations between conformational dynamics of residues reduce the configurational entropy of unfolded proteins to significant extent. Being able to quantify the latter is essential for an understanding of the thermodynamics of protein folding and disorder to order transitions. The effect of correlation dynamics on the entropy of GXYG is illustrated in Table S4. It compares the sum of individual configurational entropies of a selected set of X and Y with the total configurational entropy of XY calculated with an interaction model described in refs.(66,70) Most of the obtained values (GAVG and GALG being an exception) are negative, indicating entropy reduction. The majority of the respective room temperature Gibbs energy values are in 10^2 J/mol range. Such values could become much larger for longer peptides and even proteins. GFDG stands out with an entropy reduction of 1.54 kJ/mol. Generally, the values are lower than the ones obtained from MD simulations of Baxa et al., who reported NNI reported entropy changes in the range of -1.5 and 2.1 kJ/residue for the unfolded state of ubiquitin.(60,63)

Author contributions

RSS is the sole author of this article. He performed all the data analyses described therein and wrote the paper.

Conflicts of interest

The author does not declare any conflict of interest.

Data availability

All data analysed in this article have been published and are therefore available. Reported secondary data are either listed in the Tables in the main manuscript and the Supporting Information or will be available from the author upon request.

Acknowledgement

I would like to thank Dr. Roland L. Dunbrack from the Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, for providing me access to the files containing the data points of the Ramachandran plots reported in Ting et al. Dr. Siobhan E. Toal has checked the writing that concerned the content of ref.(72)

Footnotes

[†]In Ting et al., H was multiplied by 100 to express it in units of %.

[‡] A similar discrepancy between the Ting et al. coil library and GXYG was reported earlier by Schweitzer-Stenner and Toal.(71) However, their calculation underestimated the Hellinger distance by a factor of $\sqrt{2}$.

References

1. Uversky VN. Why Are We Interested in the Unfolded Peptides and Proteins . Schweitzer-Stenner R, editor. Hoboken: John Wiley & Sons; 2012. 3–54 p.

2. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: Introducing the D 2 concept. Vol. 37, Annual Review of Biophysics. 2008. p. 215–46.

3. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing Protein Intrinsic Disorder. Chem Rev. 2014 Jul 9;114(13):6561–88.

4. Jakob U, Kriwacki R, Uversky VN. Conditionally and transiently disordered proteins: Awakening cryptic disorder to regulate protein function. Vol. 114, Chemical Reviews. American Chemical Society; 2014. p. 6779–805.

5. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. Journal of Molecular Recognition. 2005;18(5):343–84.

Please do not adjust margins

Please do not adjust margins

6. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets: The roles of intrinsic disorder in protein interaction networks. Vol. 272, FEBS Journal. 2005. p. 5129–48.

7. Deiana A, Forcelloni S, Porrello A, Giansanti A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. PLoS One. 2019 Aug 19;14(8):e0217889.

8. Srivastava A, Yesudhas D, Ahmad S, Gromiha MM. Understanding disorder-to-order transitions in protein–RNA complexes using molecular dynamics simulations. J Biomol Struct Dyn. 2022 Nov 22;40(17):7915–25.

9. Soranno A. Physical basis of the disorder-order transition. Vol. 685, Archives of Biochemistry and Biophysics. Academic Press Inc.; 2020. p. 108305.

10. Hadži S, Lah J. The free energy folding penalty accompanying binding of intrinsically disordered α-helical motifs. Protein Science. 2022 Jul 16;31(7).

11. Oldfield CJ, Cue B, Dunker AK, Uversky VN. Binding Promiscurity of Unfolded Peptides. In: Schweitzer-Stenner R, editor. Protein and Peptide Folding, Misfolding and Non-Folding. Hoboken: Wiley & Sons, Inc; 2012. p. 241–78.

12. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic´ Z. Biochemistry 2002;41(21):6543–82.

13. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of Molecular Recognition Features (MoRFs). J Mol Biol. 2006;362(5):1043–59.

14. Morris OM, Torpey JH, Isaacson RL. Intrinsically disordered proteins: modes of binding with emphasis on disordered domains. Open Biol. 2021 Oct 6;11(10).

15. Dumarieh R, Lagasca D, Krishna S, Kragelj J, Xiao Y, Ansari S, et al. Structural Context Modulates the Conformational Ensemble of the Intrinsically Disordered Amino Terminus of α-Synuclein. J Am Chem Soc. 2025 Apr 9;147(14):11800–10.

16. Dunker AK, Obradovic Z. The protein trinity - Linking function and disorder. Nat Biotechnol. 2001;19(9):805–6.

17. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the association of p53 and 14-3-3 with their partners. BMC Genomics. 2008;9:1–20.

18. Dobson CM. Protein misfolding, evolution and disease. Trends Biochem Sci [Internet]. 1999/09/02. 1999;24(9):329–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10470028

19. Balasco N, Diaferia C, Morelli G, Vitagliano L, Accardo A. Amyloid-Like Aggregation in Diseases and Biomaterials: Osmosis of Structural Information. Front Bioeng Biotechnol. 2021;9:641372.

20. Avni A, Swasthi HM, Majumdar A, Mukhopadhyay S. Intrinsically disordered proteins in the formation of functional amyloids from bacteria to humans. Prog Mol Biol Transl Sci. 2019;166:109–43.

21. Jensen MR, Salmon L, Nodet G, Blackledge M. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J Am Chem Soc 2010;132(4):1270–2.

22. Jensen MR, Zweckstetter M, Huang JR, Blackledge M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. Vol. 114, Chem. Rev.. American Chemical Society; 2014. p. 6632–60.

23. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RWH, Blackledge M. 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. 2008 Jun 25;130(25):8055–61.

24. Sziegat F, Silvers R, Hähnke M, Jensen MR, Blackledge M, Wirmer-Bartoschek J, et al. Disentangling the coil: Modulation of conformational and dynamic properties by site-directed mutation in the non-native state of hen egg white lysozyme. Biochemistry. 2012;51(16):3361–72.

25. Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, et al. NMR characterization of long-range order in intrinsically disordered proteins. J Am Chem Soc. 2010;132(24):8407–18.

26. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen J, et al. The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics. 2008;9(Suppl 2):S1.

27. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins: Structure, Function and Genetics. 2001;42(1):38–48.

28. Vitalis A, Lyle N, Pappu R V. Thermodynamics of β-sheet formation in polyglutamine. Biophys J. 2009;97(1):303–11.

29. Das RK, Pappu R V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of appositely charged residues. Proc. Nat. Acad. of Sci. USA. 2013;110:13392–7.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

30. Lyle N, Das RK, Pappu R V. A quantitative measure for protein conformational heterogneity. J Chem Phys. 2013;139.

31. Holehouse AS, Garai K, Lyle N, Vitalis A, Pappu R V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. J Am Chem Soc. 2015;137(8):2984–95.

32. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu R V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2010;107(18):8183–8.

33. Bianchi G, Mangiagalli M, Barbiroli A, Longhi S, Grandori R, Santambrogio C, et al. Distribution of Charged Residues Affects the Average Size and Shape of Intrinsically Disordered Proteins. Biomolecules. 2022 Apr 9;12(4):561.

34. Hofmann H, Soranno A, Borgia A, Gast K, Nettels D, Schuler B. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. Proc Natl Acad Sci U S A. 2012;109(40):16155–60.

35. Wuttke R, Hofmann H, Nettels D, Borgia MB, Mittal J, Best RB, et al. Temperature-dependent solvation modulates the dimensions of disordered proteins. Proc Natl Acad Sci U S A. 2014;111(14):5213–8.

36. Müller-Späth S, Soranno A, Hirschfeld V, Hofmann H, Rüegger S, Reymond L, et al. Charge interactions can dominate the dimensions of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2010;107(33):14609–14.

37. Flory PJ. Statistical Mechanics of Chain Molecules. Ithaca: Cornell University Press; 1953.

38. Brant DA, Flory PJ. The Configuration of Random Polypeptide Chains. II. Theory. J Am Chem Soc. 1965;87(13):2791–2780.

39. Brant DA, Flory PJ. The Configuration of Random Polypeptide Chains. I. Experimental Results. J Am Chem Soc. 1965;87(13):2788–91.

40. Ramakrishnan C, Ramachandran GN. Stereochemical Criteria for Polypeptide and Protein Chain Conformation. II. Allowed Conformations for a Pair of Peptide Units. Biophys J. 1965;5:909–33.

41. Ramachandran GN, Sasisekharan V. Conformation of Polypeptides and Proteins. Adv Protein Chem. 1968;23(C):283–437.

42. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7(1):95–9.

43. Ramachandran GN, Venkatachalam CM, Krimm S. Stereochemical Criteria for Polypeptide and Protein Chain Conformations: III. Helical and Hydrogen-Bonded Polypeptide Chains. Biophys J. 1966;6(6):849–72.

44. Schweitzer-Stenner R. The Physics of Protein Structure and Dynamics. When and Why Proteins Fold and Don't Fold. San Diego: Elsevier; 2024.

45. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, et al. Random-coil behavior and the dimensions of chemically unfolded proteins. Proceedings of the National Academy of Sciences. 2004 Aug 24;101(34):12491–6.

46. Ding F, Jha RK, Dokholyan N V. Scaling Behavior and Structure of Denatured Proteins. Structure. 2005 Jul;13(7):1047–54.

47. Dill KA, Bromberg S. Molecular Driving Forces. Statistical Thermodynamics in Chemistry and Biology. New York: Garland Science; 2002.

48. Flory PJ. Statistical Mechanics of Chain Molecules. Vol. 23, Physics Today. New York: Wiley & Sons; 1953. 71 p.

49. Schweitzer-Stenner R. The relevance of short peptides for an understanding of unfolded and intrinsically disordered proteins. Vol. 25, Physical Chemistry Chemical Physics. Royal Society of Chemistry; 2023. p. 11908–33.

50. Schweitzer-Stenner R. Local and Global Behavior of Unfolded and Intrinsically Disordered Peptides and Proteins. ChemBioChem. 2025 Aug 22;26(15).

51. Garcia AE. Characterization of non-alpha helical conformations in Ala peptides. Polymer (Guildf). 2004;45(2):669–76.

52. Gnanakaran S, Garcia AE. Validation of an All-Atom Protein Force Field: From Dipeptides to Larger Peptides. J Phys Chem B. 2003;107(46):12555–7.

53. Best RB, Hummer G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J Phys Chem B [Internet]. 2009/06/12. 2009;113(26):9004–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19514729

54. Best RB, Zheng W, Mittal J. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. J Chem Theory Comput. 2014 Nov 11;10(11):5113–24.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

55. Best RB, Mittal J. Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse. J Phys Chem B. 2010 Nov 25;114(46):14916–23.

56. Toal S, Meral D, Verbaro D, Urbanc B, Schweitzer-Stenner R. PH-independence of trialanine and the effects of termini blocking in short peptides: A combined vibrational, NMR, UVCD, and molecular dynamics study. J Phys Chem B 2013;117(14):3689–706.

57. Zhang S, Schweitzer-Stenner R, Urbanc B. Do Molecular Dynamics Force Fields Capture Conformational Dynamics of Alanine in Water? J Chem Theory Comput. 2020;16(1):510–27.

58. Suresh A, Schweitzer-Stenner R, Urbanc B. Amber ff24EXP-GA, Based on Empirical Ramachandran Distributions of Glycine and Alanine Residues in Water. J Chem Theory Comput. 2025 Feb 20;

59. Schweitzer-Stenner R, Toal SE. Entropy reduction in unfolded peptides (and proteins) due to conformational preferences of amino acid residues. PhysChemChemPhys. 2014;16(41):22527–36.

60. Baxa MC, Haddadian EJ, Jha AK, Freed KF, Sosnick TR. Context and force field dependence of the loss of protein backbone entropy upon folding using realistic denatured and native state ensembles. J Am Chem Soc. 2012;134(38):15929–36.

61. Jha AK, Colubri A, Freed KF, Sosnick TR. Statistical coil model of the unfolded state: Resolving the reconciliation problem. Proc Natl Acad Sci U S A 2005;102(37):13099–104.

62. Fitzkee NC, Rose GD. Reassessing random-coil statistics in unfolded proteins. Proc Natl Acad Sci U S A 2004;101(34):12497–502.

63. Baxa MC, Haddadian EJ, Jumper JM, Freed KF, Sosnick TR. Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. Proc Natl Acad Sci U S A. 2014;111(43):15396–401.

64. Schweitzer-Stenner R, Kurbaj R, O'Neill N, Andrews B, Shah R, Urbanc B. Conformational Manifold Sampled by Two Short Linear Motif Segments Probed by Circular Dichroism, Vibrational, and Nuclear Magnetic Resonance Spectroscopy. Biochemistry. 2023 Sep 5;62(17):2571–86.

65. Schweitzer-Stenner R, Milorey B, Schwalbe H. Randomizing of Oligopeptide Conformations by Nearest Neighbor Interactions between Amino Acid Residues. Biomolecules. 2022;12:684.

66. Milorey B, Schwalbe H, O'Neill N, Schweitzer-Stenner R. Repeating Aspartic Acid Residues Prefer Turn-like Conformations in the Unfolded State: Implications for Early Protein Folding. J Phys Chem B. 2021;125(41):11392–407.

67. Milorey B, Malyshka D, Schweitzer-Stenner R. pH Dependence of Ferricytochrome c Conformational Transitions during Binding to Cardiolipin Membranes: Evidence for Histidine as the Distal Ligand at Neutral pH. J Phys Chem Lett. 2017;8(9):1993–8.

68. Tran HT, Wang X, Pappu R V. Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins. Biochemistry. 2005;44(34):11369–80.

69. Schweitzer-Stenner R, Toal SE. Anticooperative Nearest-Neighbor Interactions between Residues in Unfolded Peptides and Proteins. Biophys J. 2018;114(5):1046–57.

70. Schweitzer-Stenner R, Toal SE. Construction and comparison of the statistical coil states of unfolded and intrinsically disordered proteins from nearest-neighbor corrected conformational propensities of short peptides. Mol Biosyst. 2016;12(11):3294–306.

71. Toal SE, Kubatova N, Richter C, Linhard V, Schwalbe H, Schweitzer-Stenner R. Randomizing the Unfolded State of Peptides (and Proteins) by Nearest Neighbor Interactions between Unlike Residues. Chemistry - A European Journal. 2015;21:5173–92.

72. Milorey B, Schweitzer-Stenner R, Andrews B, Schwalbe H, Urbanc B. Short peptides as predictors for the structure of polyarginine sequences in disordered proteins. Biophys J. 2021;120(4):662–76.

73. Serrano L. Comparison between the ϕ distribution of the amino acids in the protein database and NMR data indicates that amino acids have various ϕ propensities in the random coil conformation. J Mol Biol. 1995;254(2):322–33.

74. Scholtz JM, Baldwin RL. The Mechanism of alpha-Helix Formation by Peptides. Annu Rev Biophys Biomol Struct. 1992 Jun;21(1):95–118.

75. Sosnick TR. Sampling library [Internet]. Available from: http://godzilla.uchicago.edu/cgi-bin/rama.cgi

76. Shen Y, Bax A. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

homology. The Journal of Biomolecular NMR. 2007 Aug;38(4):289–302.

77. Shen Y, Roche J, Grishaev A, Bax A. Prediction of nearest neighbor effects on backbone torsion angles and NMR scalar coupling constants in disordered proteins. Prot. Sci.. 2018 Jan 25;27(1):146–58.

78. Jiang F, Han W, Wu YD. The intrinsic conformational features of amino acids from a protein coil library and their applications in force field development. PhysChemChem Phys. 2013;15(10):3413–28.

79. Zhou CY, Jiang F, Wu YD. Residue-specific force field based on protein coil library. RSFF2: Modification of AMBER ff99SB. J Phys Chem B. 2015;119(3):1035–47.

80. Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. PLoS Comput Biol. 2010;6(4):e1000763_1-21.

81. Schweitzer-Stenner R. Distribution of conformations sampled by the central amino acid residue in tripeptides inferred from amide i band profiles and NMR scalar coupling constants. J Phys Chem B. 2009;113(9):2922–32.

82. Schweitzer-Stenner R. Secondary structure analysis of polypeptides based on an excitonic coupling model to describe the band profile of amide I' of IR, raman, and vibrational circular dichroism spectra. J Phys Chem B. 2004;108(43):16965–75.

83. Hu JS, Bax A. Determination of ϕ and χ 1 angles in proteins from 13C- 13C three- bond J couplings measured by three-dimensional heteronuclear NMR. How planar is the peptide bond? J Am Chem Soc. 1997;119(27):6360–8.

84. Wirmer J, Schwalbe H. Angular dependence of 1J(Ni,Cαi) and 2J(Ni,Cα(i-1)} couplings constants measured in J-modulated HSQCs. J Biomol NMR. 2002;23(1):47–55.

85. Debartolo J, Jha A, Freed KF, Sosnick TR. Local Backbone Preferences and Nearest-Neighbor Effects in the Unfolded and Native States. In: Schweitzer-Stenner R, editor. Protein and Peptide Folding, Misfolding, and Non-Folding. Chichester: Wiley & Sons; 2012. p. 79–98.

86. Hagarman A, Measey TJ, Mathieu D, Schwalbe H, Schweitzer-Stenner R. Intrinsic propensities of amino acid residues in GxG peptides inferred from amide I' band profiles and NMR scalar coupling constants. J Am Chem Soc. 2010;132(2):540–51.

87. Schweitzer-Stenner R, Hagarman A, Toal S, Mathieu D, Schwalbe H. Disorder and order in unfolded and disordered peptides and proteins: A view derived from tripeptide conformational analysis. I. Tripeptides with long and predominantly hydrophobic side chains. Proteins: Structure, Function and Bioinformatics. 2013;81(6):955–67.

88. Hagarman A, Mathieu D, Toal S, Measey TJ, Schwalbe H, Schweitzer-Stenner R. Amino acids with hydrogen-bonding side chains have an intrinsic tendency to sample various turn conformations in aqueous solution. Chemistry - A European Journal [Internet]. 2011/05/07. 2011;17(24):6789–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21547966

89. Penkett CJ, Redfield C, Dodd I, Hubbard J, McBay DL, Mossakowska DE, et al. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. J Mol Biol [Internet]. 1997;274(2):152–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9398523

90. Swindells MB, Macarthur MW, Thornton JM. Intrinsic ϕ,ψ propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol. 1995;2(7):569–603.

91. Zaman MH, Shen MY, Berry RS, Freed KF, Sosnick TR. Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. J Mol Biol. 2003;331(3):693–711.

92. Avbelj F, Baldwin RL. Origin of the neighboring residue effect on peptide backbone conformation. Proc Natl Acad Sci U S A [Internet]. 2004/07/16. 2004;101(30):10967–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15254296

93. Pappu R V., Srinivasan R, Rose GD. The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding. Proc Natl Acad Sci U S A. 2000;97(23):12565–70.

94. Bundi A, Wüthrich K. ^{1}H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. Biopolymers. 1979;18(2):285–97.

95. Braun D, Wider G, Wüthrich K. Sequence-Corrected ^{15}N "Random Coil" Chemical Shifts. J. Am. Chem. Soc. 1994;116:8466–9.

96. Schwarzinger S, Kroon GJA, Foss TR, Chung J, Wright PE, Dyson HJ. Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc. 2001;123(13):2970–8.

Please do not adjust margins

Please do not adjust margins

97. Kjaergaard M, Poulsen FM. Sequence correction of random coil chemical shifts: Correlation between neighbor correction factors and changes in the Ramachandran distribution. J Biomol NMR. 2011;50(2):157–65.

98. Jung YS, Oh KI, Hwang GS, Cho M. Neighboring residue effects in terminally blocked dipeptides: Implications for residual secondary structures in intrinsically unfolded/disordered proteins. Chirality. 2014;26(9):443–52.

99. Gómez S, Egidi F, Puglisi A, Giovannini T, Rossi B, Cappelli C. Unlocking the power of resonance Raman spectroscopy: The case of amides in aqueous solution. J Mol Liq. 2022 Jan;346:117841.

100. Avbelj F, Baldwin RL. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: Distributions of phi. Proc Natl Acad Sci U S A. 2003;100(10):5742–7.

101. Mezei M, Fleming PJ, Srinivasan R, Rose GD. Polyproline II Helix Is the Preferred Conformation for Unfolded Polyalanine in Water. Proteins: Structure, Function and Genetics. 2004;55(3):502–7.

102. Eker F, Cao X, Nafie L, Huang Q, Schweitzer-Stenner R. The structure of alanine based tripeptides in water and dimethyl sulfoxide probed by vibrational spectroscopy. J Phys Chem B. 2003;107(1):358–65.

103. Meral D, Toal S, Schweitzer-Stenner R, Urbanc B. Water-Centered Interpretation of Intrinsic pPII Propensities of Amino Acid Residues: In Vitro-Driven Molecular Dynamics Study. J Phys Chem B 2015;119(42):13237–51.

104. Milorey B, Farrell S, Toal SE, Schweitzer-Stenner R. Demixing of water and ethanol causes conformational redistribution and gelation of the cationic GAG tripeptide. Chemical Communications. 2015;51(92):16498–501.

105. Diguiseppi D, Milorey B, Lewis G, Kubatova N, Farrell S, Schwalbe H, et al. Probing the Conformation-Dependent Preferential Binding of Ethanol to Cationic Glycylalanylglycine in Water/Ethanol by Vibrational and NMR Spectroscopy. J Phys Chem B. 2017;121(23):5744–58.

106. Shi Z, Chen K, Liu Z, Ng A, Bracken WC, Kallenbach NR. Polyproline II propensities from GGXGG peptides reveal an anticorrelation with β-sheet scales. Proc Natl Acad Sci U S A. 2005;102(50):17964–8.

107. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet Processes. J Am Stat Assoc. 2006 Dec 1;101(476):1566–81.

108. Gnanakaran S, García AE. Helix-coil transition of alanine peptides in water: Force field dependence on the folded and unfolded structures. Proteins: Structure, Function and Genetics 2005;59(4):773–82.

109. Verbaro D, Ghosh I, Nau WM, Schweitzer-Stenner R. Discrepancies between conformational distributions of a polyalanine peptide in solution obtained from molecular dynamics force fields and amide I' band profiles. J Phys Chem B. 2010;114(51):17201–8.

110. Graf J, Nguyen PH, Stock G, Schwalbe H. Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study. J Am Chem Soc 2007;129(5):1179–89.

111. Jiang F, Han W, Wu YD. Influence of side chain conformations on local conformational features of amino acids and implication for force field development. J Phys Chem B 2010;114(17):5840–50.

112. Mukrasch MD, Markwick P, Biernat J, Von Bergen M, Bernadó P, Griesinger C, et al. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. J Am Chem Soc. 2007;129(16):5235–43.

113. Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, et al. Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering. Structure. 2014;22(2):238–49.

114. Mantsyzov AB, Maltsev AS, Ying J, Shen Y, Hummer G, Bax A. A maximum entropy approach to the study of residue-specific backbone angle distributions in α-synuclein, an intrinsically disordered protein. Protein Science. 2014 Sep 22;23(9):1275–90.

115. Roche J, Shen Y, Lee JH, Ying J, Bax A. Monomeric Aβ 1–40 and Aβ 1–42 Peptides in Solution Adopt Very Similar Ramachandran Map Distributions That Closely Resemble Random Coil. Biochemistry. 2016 Feb 9;55(5):762–75.

116. Andrews B, Guerra J, Schweitzer-Stenner R, Urbanc B. Do molecular dynamics force fields accurately model Ramachandran distributions of amino acid residues in water? PhysChemChemPhys. 2022;24(5):3259–79.

117. Bernadó P, Bertoncini CW, Griesinger C, Zweckstetter M, Blackledge M. Defining long-range order and local disorder in native α-synuclein using residual dipolar couplings. J Am Chem Soc. 2005;127(51):17968–9.

118. Ozenne V, Schneider R, Yao M, Huang JR, Salmon L, Zweckstetter M, et al. Mapping the potential energy

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins

Please do not adjust margins

landscape of intrinsically disordered proteins at amino acid resolution. J Am Chem Soc. 2012 Sep 12;134(36):15138–48.

119. Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, et al. Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics. 2012 Jun;28(11):1463–70.

This journal is © The Royal Society of Chemistry 20xx

Please do not adjust margins