

1 The BackMAP Python Module: How a 2 Simpler Ramachandran Number Can 3 Simplify the Life of a Protein Simulator

4 Ranjan V. Mannige^{1,*}

5 ¹ Multiscale Institute, Berkeley Lake, GA 30092, U.S.A.

6 * ranjanmannige@gmail.com

7 ABSTRACT

8 Protein backbones occupy diverse conformations, but compact metrics to describe such conformations
9 and transitions between them have been missing. This report re-introduces the Ramachandran number
10 (\mathcal{R}) as a residue-level structural metric that could simply the life of anyone contending with large numbers
11 of protein backbone conformations (e.g., ensembles from NMR and trajectories from simulations).
12 Previously, the Ramachandran number (\mathcal{R}) was introduced using a complicated **closed form**, which made
13 the Ramachandran number difficult to implement. This report discusses a much simpler closed form of
14 \mathcal{R} that makes it much easier to calculate, thereby making it easy to implement. Additionally, this report
15 discusses how \mathcal{R} dramatically reduces the dimensionality of the protein backbone, thereby making it
16 ideal for simultaneously interrogating large number of protein structures. For example, two hundred
17 distinct conformations can easily be described in one graphic using \mathcal{R} (rather than two hundred distinct
18 Ramachandran plots). Finally, a new Python-based backbone analysis tool – BACKMAP – is introduced
19 that reiterates how \mathcal{R} can be used as a simple and succinct descriptor of protein backbones and their
20 dynamics.

21 INTRODUCTION

22 Proteins are a class of biomolecules unparalleled in their functionality (Berg *et al.*, 2010). A natural
23 protein may be thought of as a linear chain of amino acids, each normally sourced from a repertoire of
24 20 naturally occurring amino acids. Proteins are important partially because of the structures that they
25 access: the conformations (conformational ensemble) that a protein assumes determines the functions
26 available to that protein. However, all proteins are dynamic: even stable proteins undergo long-range
27 motions in its equilibrium state; i.e., they have substantial diversity in their conformational ensemble
28 (James and Tawfik, 2003b,a; Oldfield *et al.*, 2005; Tokuriki and Tawfik, 2009; Schad *et al.*, 2011; Vértesy
29 and Orosz, 2011; Mannige, 2014). Additionally, a number of proteins undergo conformational transitions,
30 without which they may not properly function. Finally, some proteins – intrinsically disordered proteins
31 – display massive disorder whose conformations dramatically change over time (Uversky, 2003; Fink,
32 2005; Midic *et al.*, 2009; Espinoza-Fonseca, 2009; Uversky and Dunker, 2010; Tompa, 2011; Sibille and
33 Bernado, 2012; Kosol *et al.*, 2013; Dunker *et al.*, 2013; Geist *et al.*, 2013; Baruah *et al.*, 2015), and whose

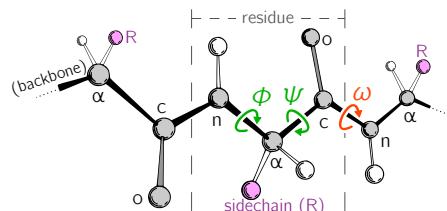


Figure 1. Backbone conformational degrees of freedom dominantly depend on the dihedral angles ϕ and ψ (green), and to a smaller degree depend on the third dihedral angle (ω ; red) as well as bond lengths and angles (unmarked).

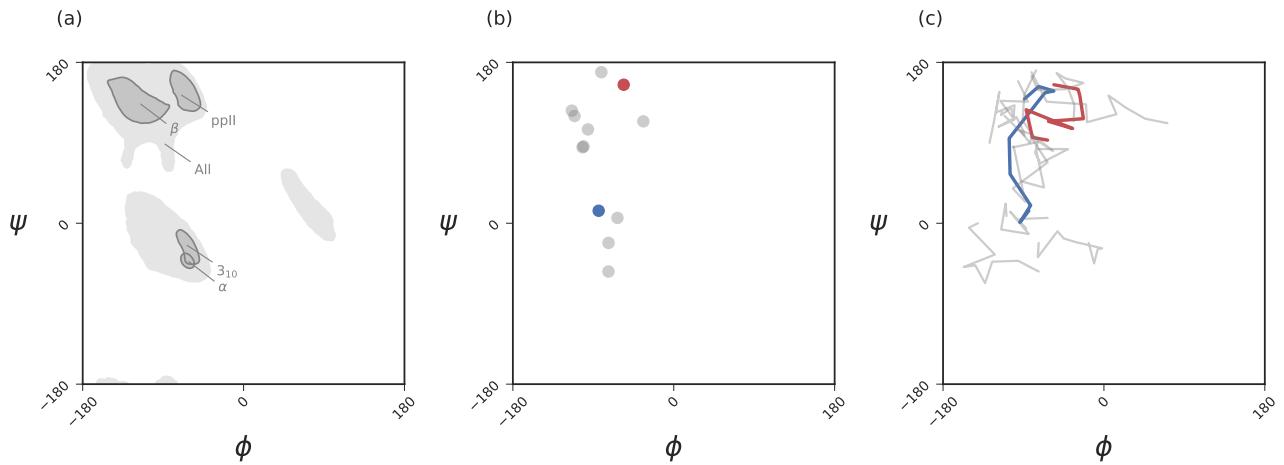


Figure 2. Ramachandran plots allow for the per-residue representation of backbone conformation. Panel (a) represents regions in the plot that are occupied by backbones describing particular regular secondary structures. Panel (b) represents the positions of a 11 residue peptide that describe, with one dot per residue. Panel (c) represents a seven-frame trajectory, where each residue's backbone traces a line, with While the Ramachandran plot is useful for getting a *qualitative* sense of peptide backbone structure (a, b), it is not a convenient representation for exploring peptide backbone dynamics (c).

Secondary structure keys used here and throughout the document: ‘ α ’ – α -helix, ‘ 3_{10} ’ – 3_{10} -helix, ‘ β ’ – β -sheet/extension, ‘ppII’ – polyproline II helix.

34 characteristic structures are still not well-understood (Beck *et al.*, 2008).

35 Large-scale changes in a protein occur due to changes in protein backbone conformations. Fig. 1 is a
 36 cartoon representation of a peptide/protein backbone, with the backbone bonds themselves represented
 37 by darkly shaded bonds. Ramachandran *et al.* (1963) had recognized that the backbone conformational
 38 degrees of freedom available to an amino acid (residue) i is almost completely described by only two
 39 dihedral angles: ϕ_i and ψ_i (Fig. 1, green arrows). Today, Ramachandran plots are used to qualitatively
 40 describe protein backbone conformations.

41 The Ramachandran plot is recognized as a powerful tool for two reasons: 1) it serves as a map
 42 for structural ‘correctness’ (Laskowski *et al.*, 1993; Hooft *et al.*, 1997; Laskowski, 2003), since many
 43 regions within the Ramachandran plot space are energetically not permitted (Momen *et al.*, 2017); and
 44 2) it provides a qualitative snapshot of the structure of a protein (Berg *et al.*, 2010; Alberts *et al.*, 2002;
 45 Subramanian, 2001; Lovell *et al.*, 2003). For example, particular regions within the Ramachandran plot
 46 indicate the presence of particular secondary locally-ordered structures such as the α -helix and β -sheet
 47 (see Fig. 2a).

48 While the Ramachandran plot has been useful as a measure of protein backbone conformation, it is
 49 not popularly used to assess structural dynamism and transitions (unless specific knowledge exists about
 50 whether a particular residue is believed to undergo a particular structural transition). This is because
 51 of the two-dimensionality of the plot: describing the behavior of every residue involves tracking its
 52 position in two-dimensional (ϕ, ψ) space. For example, a naive description of positions of a peptide in a
 53 Ramachandran plot (Fig. 2b) needs more annotations for a per-residue analysis of the peptide backbone’s
 54 structure. Given enough residues, it would be impractical to track the position of each residue within a
 55 plot. This is compounded with time, as each point in (b) becomes a curve (c), further confounding the
 56 situation. The possibility of picking out previously unseen conformational transitions and dynamism
 57 becomes a logistical impracticality. As indicated above, this impracticality arises primarily from the fact
 58 that the Ramachandran plot is a two-dimensional map.

59 For example, tracking changes in protein trajectory is either overly detailed or overly holistic: an
 60 example of an overly detailed study is the tracking on exactly one or a few atoms over time (this
 61 already poses a problem, since we would need to know exactly which atoms are expected to partake
 62 in a transition); an example of a holistic metric is the radius of gyration (this also poses a problem,
 63 since we will never know which residues contribute to a change in radius of gyration without additional
 64 interrogation). With our understanding of protein dynamics undergoing a new **renaissance** – especially due
 65 to intrinsically disordered proteins and allosteric – having hypothesis-agnostic yet detailed (residue-level)

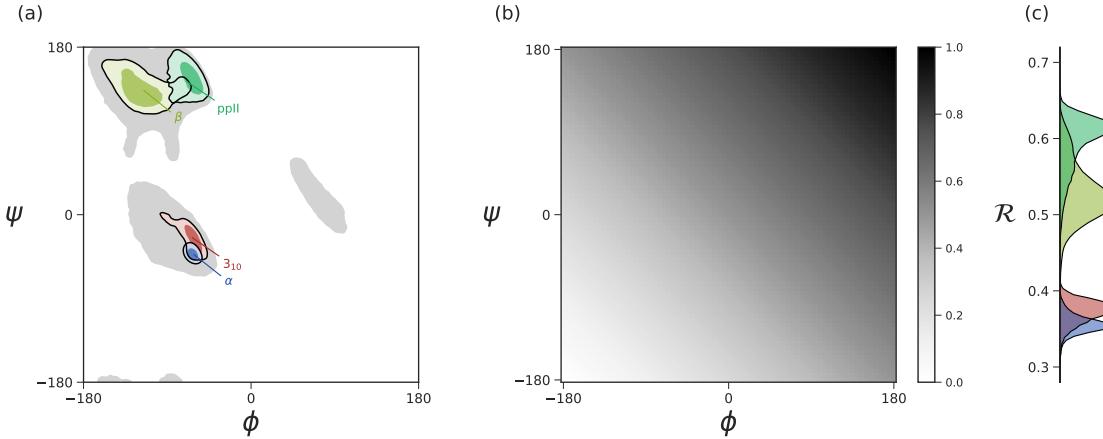


Figure 3. Panel (a) describes the distribution of dominant regular secondary structures. Panel (b) shows the mapping between the (ϕ, ψ) and \mathcal{R} . In Particular, \mathcal{R} increases in negative-sloping sweeps from the bottom left to the top right of the Ramachandran plot. Panel (c) describes the distribution of secondary structures in \mathcal{R} space. Both Ramachandran plots (a) and Ramachandran ‘lines’ (c) equally resolve the secondary structure space, thereby making \mathcal{R} a compact yet faithful representation of backbone structure (Mannige *et al.*, 2016).

metrics of protein structure has become even more relevant. Consequently, there has been no single compact descriptor of protein structure. This impedes the naïve or hypothesis-free exploration of new trajectories/ensembles.

It has recently been shown that the two Ramachandran backbone parameters (ϕ, ψ) may be conveniently combined into a single number – the Ramachandran *number* [$\mathcal{R}(\phi, \psi)$ or simply \mathcal{R}] – with little loss of information (Fig. 3; Mannige *et al.* (2016)). In a previous report, detailed discussions were provided regarding the reasons behind and derivation of \mathcal{R} (Mannige *et al.*, 2016). This report provides a simpler version of the equation previously published (Mannige *et al.*, 2016), and further discusses how \mathcal{R} may be used to provide information about protein ensembles and trajectories. Finally, this report introduces a software package – BACKMAP – that can be used by to produce *pictograms* that describe the behavior of a protein backbone within user-inputted conformations, structural ensembles and trajectories. Given that each pictogram provides a picture of the whole protein backbone (i.e., all ϕ and ψ angles), these pictograms are named *multi-angle pictures* (or MAPs). BACKMAP is presently available on GitHub (<https://github.com/ranjanmannige/BackMAP>).

INTRODUCING THE SIMPLIFIED RAMACHANDRAN NUMBER (\mathcal{R})

The Ramachandran number is both an idea and an equation. Conceptually, the Ramachandran number (\mathcal{R}) is any closed form that collapses the dihedral angles ϕ and ψ into one structurally meaningful number (Mannige *et al.*, 2016). Mannige *et al.* (2016) presented a version of the Ramachandran number (shown in the appendix as Eqn. 7) that was complicated in closed form, thereby reducing its utility. Here, a simpler and more accurate version of the Ramachandran number is introduced. The appendix shows how this simplified form was derived from the original closed form (Eqns. 7).

Given arbitrary limits of $\phi \in [\phi_{\min}, \phi_{\max}]$ and $\psi \in [\psi_{\min}, \psi_{\max}]$, where the minimum and maximum values differ by 360° , the most general and accurate equation for the Ramachandran number is

$$\mathcal{R}(\phi, \psi) \equiv \frac{\phi + \psi - (\phi_{\min} + \psi_{\min})}{(\phi_{\max} + \psi_{\max}) - (\phi_{\min} + \psi_{\min})}. \quad (1)$$

For consistency, we maintain throughout this paper that $\phi_{\min} = \psi_{\min} = -180^\circ$ or $-\pi$ radians, which makes

$$\mathcal{R}(\phi, \psi) = \frac{\phi + \psi + 2\pi}{4\pi}. \quad (2)$$

As evident in Fig. 3, the distributions within the Ramachandran plot are faithfully reflected in corresponding distributions within Ramachandran number space. This paper shows how the Ramachandran

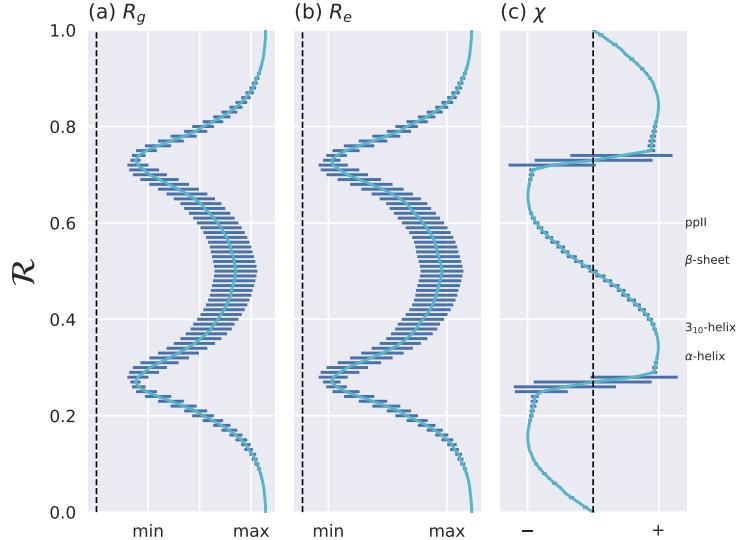


Figure 4. The Ramachandran number \mathcal{R} displays smooth relationships with respect to radius of gyration (R_g ; a), end-to-end distance (R_e ; b) and chirality (χ ; c), as calculated within Mannige (2017). Light blue lines are average trends, dark blue horizontal lines are error bars. Average positions of dominant secondary structures are shown to the right. These trends explain why \mathcal{R} is a useful and compact structural measure. Structural measures R_g , R_e , and χ were obtained by computationally generating polyglycine peptides of length 10 for all possible ϕ and $\psi \in [-180, -175, \dots, 175, 180]$. This was done using the Python library PeptideBuilder (Tien *et al.*, 2013). Values for R_g , R_e , and χ were obtained for each peptide and binned with respect to its $\mathcal{R}(\phi, \psi)$ (each bin represents a region in \mathcal{R} space that is 0.01 \mathcal{R} in width). Given that actual values for R_g and R_e mean little (since one rarely deals with polyglycines of length 10), actual values are omitted. χ ranges from -1 to +1.

89 number is both compact enough and informative enough to generate immediately useful graphs (multi-angle pictures or MAPs) of a dynamic protein backbone.
90

91 REASON TO USE THE RAMACHANDRAN NUMBER

92 Ramachandran numbers are structurally meaningful

93 In addition to resolving positions of secondary structures (Fig. 3), \mathcal{R} relates well to structural measures
94 such as radius of gyration (R_g), end-to-end distance (R_e) and chirality (χ). These relationships are shown
95 in Fig. 4. Note that chirality comes in many forms, e.g., one could be talking about different stereo-isomers,
96 such as L vs D amino acids, or one could be concerned with left-twisting versus right-twisting backbones,
97 i.e., handedness Mannige (2017). This report will primarily be focused on chirality in context of backbone
98 twist/handedness.

99 The trends in Fig. 4 show that as one progresses from low to high \mathcal{R} , various structural properties also
100 progress smoothly. Additionally, backbones that display similar \mathcal{R} also show little variation in structural
101 properties, as evidenced by the small standard deviation bars. It is also important to note that the standard
102 deviations shown in Fig. 4 were calculated by first populating every possible region of (ϕ, ψ) -space.
103 However, in reality, most regions of (ϕ, ψ) -space are unoccupied due to steric/electrostatic constraints,
104 which means that these error bars are likely to be even smaller than depicted. Finally, the \mathcal{R} number
105 is calculated by taking ‘sweeps’ of the (ϕ, ψ) -space in lines that are parallel to the negatively-sloping
106 diagonal. Interestingly, such ‘sweeps’ encounter only one major (dense) region within (ϕ, ψ) -space (e.g.,
107 \mathcal{R} ’s in the general vicinity of 0.34 represent structures that resemble α helices. This means that \mathcal{R} can
108 also be used to assess the types of secondary structure present in a protein conformation.

109 Ramachandran codes are stackable

110 An important aspect of the Ramachandran number (\mathcal{R}) lies in its compactness compared to the traditional
111 Ramachandran pair (ϕ, ψ) . The value of the conversion from (ϕ, ψ) -space to \mathcal{R} -space is that the structure

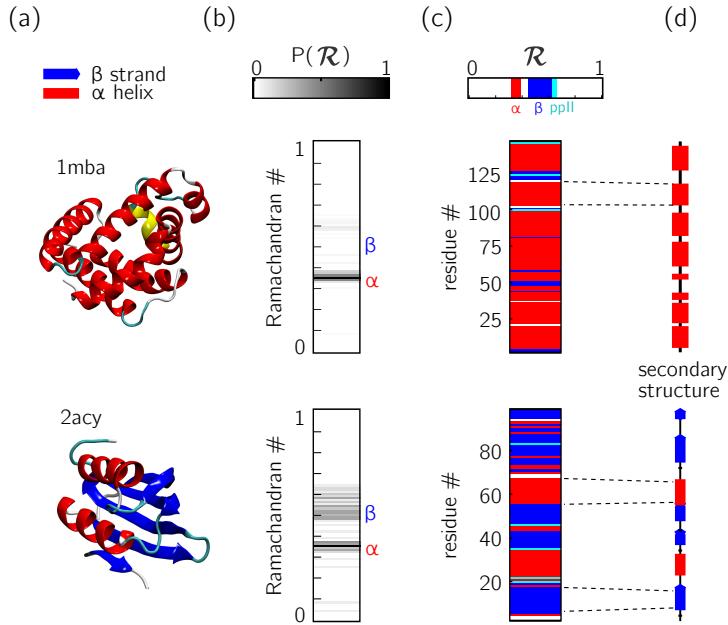


Figure 5. Two types of \mathcal{R} -codes. Digesting protein structures (a) using \mathcal{R} numbers either as histograms (b) or per-residue codes (c) allow for compact representations of salient structural features. For example, a single glance at the histograms indicate that protein 1mba is likely all α -helical, while 2acy is likely a mix of α -helices and β -sheets. Additionally, residue-specific codes (c) not only indicate secondary structure content, but also exact **secondary** structure stretches (compare to d), which gives a more complete picture of how the protein is linearly arranged.

of a protein can be described in various one-dimensional arrays (per-structure “Ramachandran codes” or “ \mathcal{R} -codes” or multi-angle maps); see, e.g., Fig. 5.

In addition to assuming a small form factor, \mathcal{R} -codes may then be *stacked* side-by-side for visual and computational analysis. There lies its true power.

For example, the one- \mathcal{R} -to-one-residue mapping means that the entire residue-by-residue structure of a protein can be shown using a string of \mathcal{R}_i s (which would show regions of secondary structure and disorder, for starters). Additionally, an entire protein’s backbone makeup can be shown as a histogram in \mathcal{R} -space (which may reveal a protein’s topology). The power of this format lies not only in the capacity to distill complex structure into compact spaces, but in its capacity to display *many* complex structures in this format, side-by-side (stacking).

Peptoid nanosheets (Mannige *et al.*, 2015) will be used here as an example of how multiple structures, in the form of \mathcal{R} -codes, may be stacked to provide immediately useful pictograms. Peptoids are stereoisomers of peptides, where the sidechain is attached to the backbone nitrogen rather than the α carbon atom. Since both peptoids and peptides share identical backbone connectivity, the analysis described below could be applied to both peptides and peptoids.

Peptoid nanosheets are a recently discovered peptide-mimic that, in one molecular dynamics simulation (Mannige *et al.*, 2015), were shown to display a novel secondary structure. In the reported model (Mannige *et al.*, 2015), each peptoid within the nanosheet displays backbone conformations that alternate in chirality, causing the backbone to look like a meandering snake that nonetheless maintains an overall linear direction. This secondary structure was discovered by first setting up a nanosheet where all peptoid backbones were restrained to be fully extended (Fig. 6a, left), after which the restraints were energetically softened (a, middle) and completely **released** (a, right). As evident in Fig. 6b and Fig. 6c, the two types of \mathcal{R} -code stacks display salient information at first glance: 1) Fig. 6b shows that the extended backbone first undergoes some rearrangement with softer restraints, and then becomes much more binary in arrangement as we look down the backbone (excepting the low-order region in the middle, unshown in Fig. 6a); and 2) Fig. 6c shows that lifting restraints on the backbone causes a dramatic change in backbone topology, namely a birth of a bimodal distribution evident in the two parallel horizontal bands.

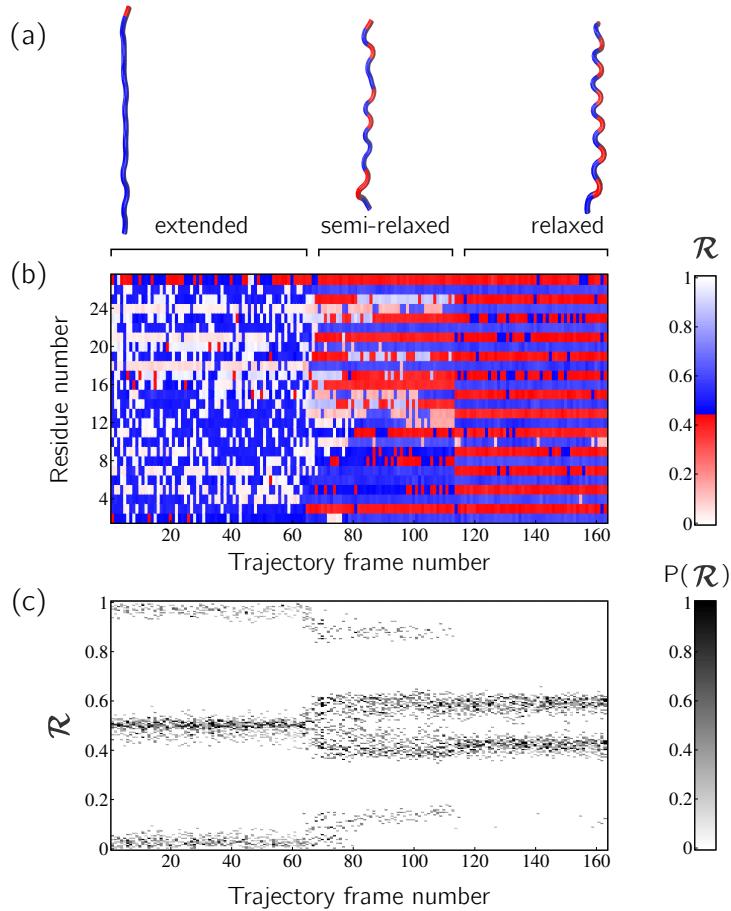


Figure 6. Stacked \mathcal{R} -codes provide useful information at a glance. Each panel represents a molecular dynamics simulation of a peptoid nanosheet (Mannige *et al.*, 2016), where each peptoid backbone was held (energetically restrained) in extended state in the beginning, upon which each backbone was allowed to relax by lifting the restraints. Panel (a) displays representative structures from each stage of the simulation. Panel (b) represents how the per-residue structure of the peptide evolved over ‘time’ (the progression of time is represented as increasing frame number). Panel (c) represents how the general distribution of backbone conformations in the peptoid (as evident by the \mathcal{R} histogram) evolves over time.

139 By utilizing \mathcal{R} , maps such as those in Fig. 6 provide information about every ϕ and ψ within the
 140 backbone. As such, these maps are dubbed MAPs, for Multi Angle Pictures. A Python package called
 141 BACKMAP created Fig. 6a and b, which is provided as a GitHub repository at <https://github.com/ranjanmannige/BackMAP>. BACKMAP takes in a PDB structure file containing a single
 142 structure, or multiple structures separated by the code ‘MODEL’.
 143

144 Case study: picking out subtle differences from high volume of data

145 This section expands on the notion that \mathcal{R} -numbers – due to their compactness/stackability – can be used
 146 to pick out backbone structural trends that would be hard to decipher using any other metric. For example,
 147 it is well known that prolines (P) display unusual backbone behavior: in particular, proline backbones
 148 occupy structures that are close to but distinct from α -helical regions. Due to the two-dimensionality
 149 of Ramachandran plots (Fig. 7a), such distinctions are hard to visually pick out from Ramachandran
 150 plots. However, stacking per-amino-acid \mathcal{R} -codes side by side make such differences patent (Fig. 7b; see
 151 arrow).

152 It is also known that amino acids preceding prolines display unusual shift in backbone twist/chirality.
 153 For example, Fig. 8c shows that amino acids appearing before prolines behave differently than they would
 154 otherwise (see the upward-facing arrow). Additionally, amino acids *following* glycines also appear to

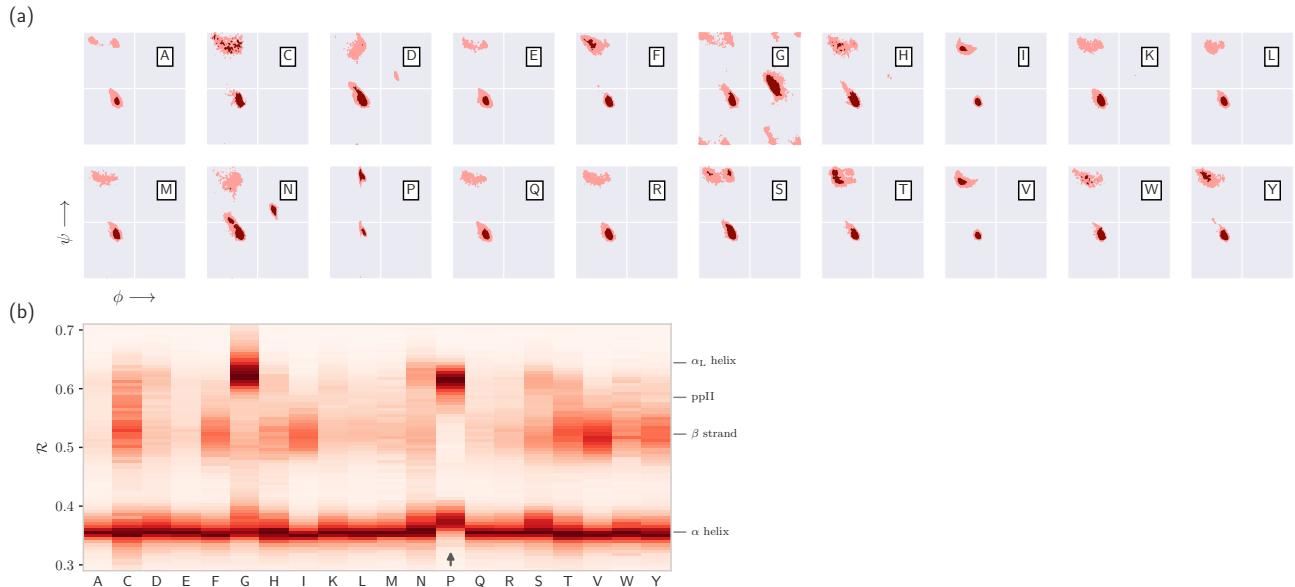


Figure 7. Combining Ramachandran plots for all amino acids into one graph. Panel (a) shows the per-amino acid backbone behavior of an average protein found in the protein databank (PDB). While these plots are useful, it is difficult to compare such plots. For example, it is hard to pick out the change in the α -helical region of the proline plot (P). However, when we convert Ramachandran plots to Rama^{chandran} lines [by converting $(\phi_i, \psi_i) \rightarrow \mathcal{R}_i$], we are able to conveniently “stack” Ramachandran lines calculated for each residue. Then, even visually, it is obvious that proline does not occupy the canonical α -helix region, which is not evident to an untrained eye in (a).

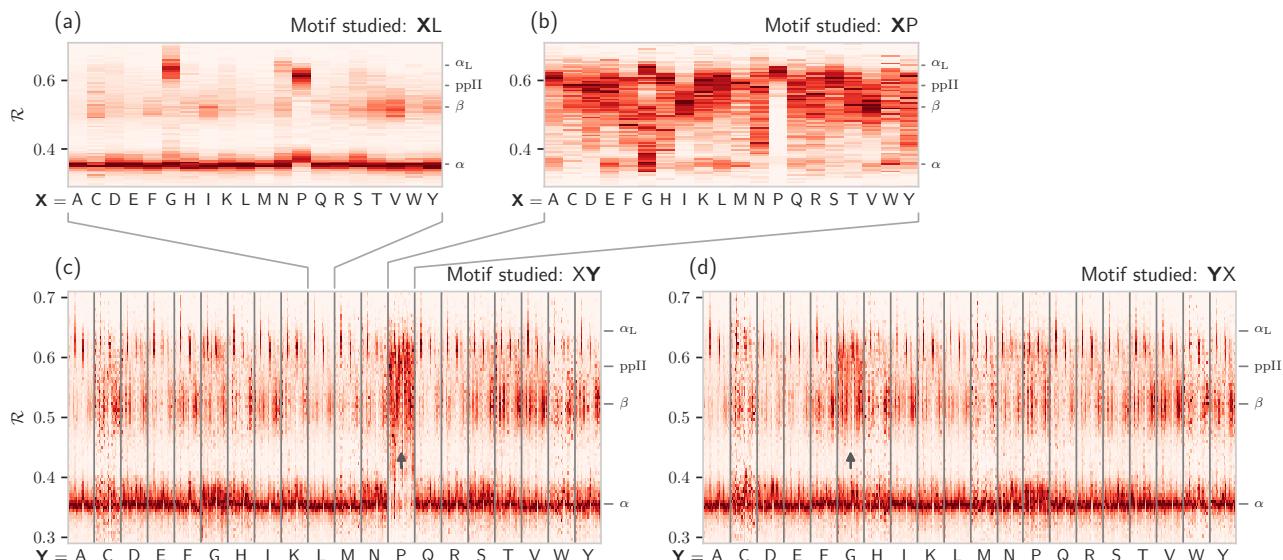


Figure 8. How residue neighbors modifies structure. Similar to Fig. 7b, Panel (a) represents the behavior of an amino acid ‘X’ situated *before* a leucine (XL; assuming that we are reading a sequence from the N terminal to the C terminal). Panel (b) similarly represents the behavior of specific amino acids situated before a proline (XP). While residues preceding a leucine behave similarly to their average behavior (Fig. 7a), most residues preceding prolines appear to be enriched in structures that change ‘direction’ or backbone chirality (this is evident by many amino acids switching from $\mathcal{R} < 0.5$ to $\mathcal{R} > 0.5$). Panels (c) and (d) show the behavior of individual amino acids when situated before and after each of the 20 amino acids, respectively. Panels (c) and (d) show a major benefit of side-by-side Ramachandran line “stacking”: general trends become much more obvious. For example, it is evident that prolines dramatically modify the structure of an amino acid preceding it (compared to average behavior of amino acids in Fig. 7b), while residues following glycines also have a higher prevalence of $\mathcal{R} > 0.5$ conformations (both trends are indicated by small arrows). Such trends, while previously discovered (see text), would not be accessible when naïvely considering Ramachandran plots because one would require 400 (20×20) distinct Ramachandran plots to compare. Note that the statistics for each \mathcal{R} -line in (c) and (d) are dependent on the joint prevalence of the residues being considered. For this reason, some \mathcal{R} -lines (e.g., those associated with cysteines) look ⁷⁷¹⁶ rough or ‘dotty’ than others..

155 have their structures modified (Fig. 8d; upward arrow). Note that these results are not new, and it has
 156 already been confirmed that, e.g., nearest neighbors affect the conformational behavior of an amino
 157 acid as witnessed within Ramachandran plots (Ting *et al.*, 2010), and proline changes the backbone
 158 conformation of the preceding residue (Gunasekaran *et al.*, 1998; Ho and Brasseur, 2005). However,
 159 Figs. 7 and 8 indicate that such information can be more concisely shown/identified when structures
 160 are stacked side-by-side in the form of \mathcal{R} -codes. Such subtle changes are often witnessed when protein
 161 backbones transition from one state to another.

162 USING THE BACKMAP PYTHON MODULE

163 Installation

164 BACKMAP may either be installed locally by downloading the [GitHub repository](#), or installed directly
 165 by running the following line in the command prompt (assuming that pip exists): > pip install
 166 backmap

167 Usage

168 The module can either be imported and used within existing scripts, or used as a standalone package using
 169 the command ‘python -m backmap’. First the in-script usage will be discussed.

170 In-script usage I: first simple test

171 The simplest test would be to generate Ramachandran numbers from (ϕ, ψ) pairs:

172	# Import module	1
173	import backmap	2
174	# Convert (phi, psi) to R	3
175	print backmap.R(phi=0,phi=0) # Expected output: 0.5	4
176	print backmap.R(-180,-180) # Expected output: 0.0	5
177	print backmap.R(-180, 180) # Expected output: 1.0 (equivalent in meaning to 0)	6

180 In-script usage II: basic usage for creating Multi-Angle Pictures (MAPs)

181 The following code shows how Multi-Angle Pictures (MAPs) of protein backbones can be generated:

182 1. Select and read a protein PDB structure

183 Each trajectory frame must be a set of legitimate protein databank "ATOM" records separated by
 184 "MODEL" keywords (distinct models show up as distinct frames on the x-axis or abscissa).

185	import backmap	1
186	pdbfn = './pdbs/nanosheet_birth_U7.pdb' # Set pdb name	2
187	data = backmap.read_pdb(pdbfn) # READ PDB in the form of a matrix with columns	3

190 Here, ‘data’ is a 2d array with four columns [‘model’, ‘chain’, ‘resid’, ‘R’]. The first row of
 191 ‘data’ is the header (i.e., the name of the column, e.g., ‘model’), with values that follow.

192 2. Select color scheme (color map)

193 In addition to custom colormaps listed in the next section, one can also use traditionally available
 194 colormaps at [matplotlib.org](#) (e.g., ‘Reds’ or ‘Reds_r’).

195	# setting the name of the colormap	4
196	cmap = "SecondaryStructure"	5

199 3. Draw per-chain MAPs

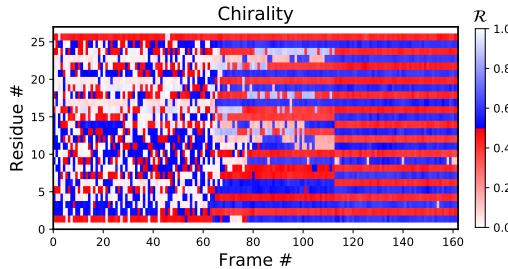
200	# Grouping by chain	6
201	grouped_data = backmap.group_by(data, group_by='chain',	7
202	columns_to_return=['model', 'resid ', 'R'])	8
203	for chain in grouped_data.keys(): # Going through each chain	9
204	# Getting the X,Y,Z values for each entry	10
205	models , residues , Rs = grouped_data[chain]	11
206	# Finally , creating (but not showing) the graph	12
207	backmap.draw_xyz(X = models , Y = residues , Z = Rs	13
208	, xlabel ='Frame #' , ylabel ="Residue #" , zlabel ='\$\mathcal{R}\$')	14

```

210 ,cmap = cmap      , title = "Chain: '"+chain+"'"
211 ,vmin=0,vmax=1)
212 # Now, we display the graph:
213 plt.show() # ... one can also use plt.savefig() to save to file

```

215 Running the module as a standalone script would produce all these graphs automatically. ‘plt.show()’
216 would result in the following image being rendered:



217

218 Additionally, by changing how one assigns values to ‘X’ and ‘Y’, one can easily construct and draw
219 other types of graphs such as time-resolved histograms, root mean squared fluctuations, root mean
220 squared deviation, etc.

221 In-script usage III: Creating custom graphs

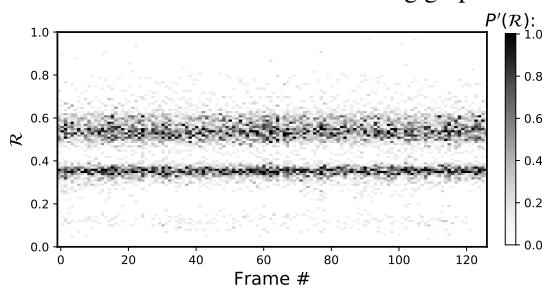
222 Other types of [graphs](#) can be easily created by modifying part three of the code above. For example, the
223 following code creates histograms of R, one for each model (starting from line 9 above).

```

224 for chain in grouped_data.keys():
225     models, residues, Rs = grouped_data[chain]
226
227     'Begin custom code'
228     X = []; Y=[]; Z=[]; # Will set X=model, Y=R, Z=P(R)
229     # Bundling the three lists into one 2d array
230     new_data = np.array(list(zip(models, residues, Rs)))
231     # Getting all R values, model by model
232     for m in sorted(set(new_data[:,0])): # column 0 is the model column
233         # Getting all Rs for that model #
234         current_rs = new_data[np.where(new_data[:,0]==m)][:,2] # column 2 contains R
235         # Getting the histogram
236         a,b = np.histogram(current_rs, bins=np.arange(0,1.01,0.01))
237         max_count = float(np.max(a))
238         for i in range(len(a)):
239             X.append(m); Y.append((b[i]+b[i+1])/2.0); Z.append(a[i]/float(np.sum(a)));
240     'End custom code'
241
242     # Finally, creating (but not showing) the graph
243     draw_xyz(X = X      , Y = Y      , Z = Z
244             , xlabel = 'Frame #' , ylabel = "$\\mathcal{R}$" , zlabel = "$P'(\\mathcal{R})$"
245             ,cmap = 'Greys' , ylim=[0,1])
246     plt.yticks(np.arange(0,1.00001,0.2))
247     # Now, we display the graph:
248     plt.show() # ... one can also use plt.savefig() to save to file

```

251 The code above results in the following graph:



252

253 **In-script usage IV: Available color schemes (CMAPs)**

254 Aside from the general color maps (cmaps) that exist in matplotlib (e.g., ‘Greys’, ‘Reds’, or, god forbid,
 255 ‘jet’), BACKMAP provides two new colormaps: ‘Chirality’ (key: +twists – red; -ve twists: blue),
 256 and ‘SecondaryStructure’ (key: *potential* helices – red; sheets – blue; ppII helices – cyan). right
 257 twisting backbones are shown in red; left twisting backbones are shown in blue). Fig. 9 shows how
 258 a single protein ensemble may be described using these schematics. As illustrated in Fig. 9b, cmaps
 259 available within the standard matplotlib package do not distinguish between major secondary structures
 260 well, while those provided by BACKMAP do. In case it is known that the protein backbone accesses
 261 non-traditional regions of the Ramachandran plot, a four-color schematic will be needed (see below for
 262 more discussions).

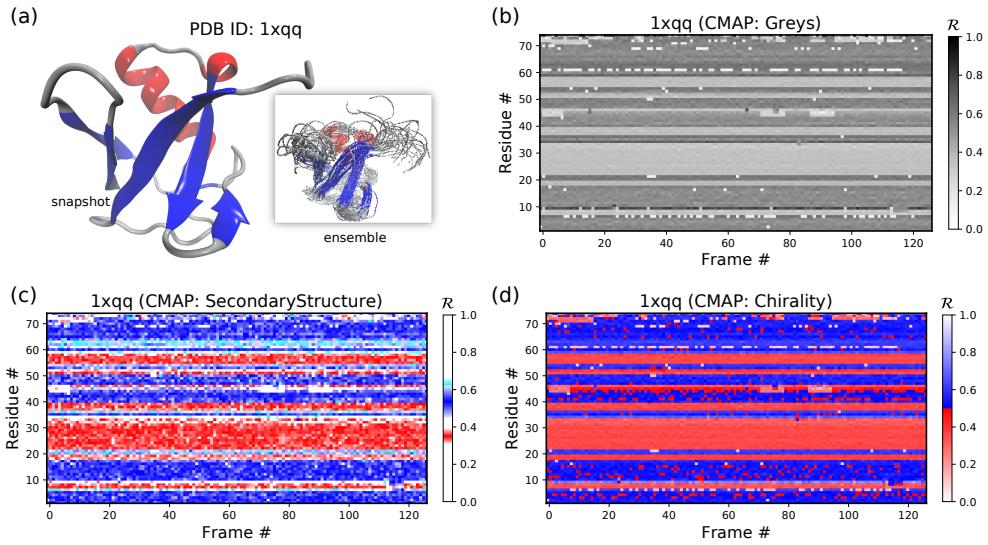


Figure 9. A protein ensemble (a) along with some MAPs colored with different themes (b-d). Panels (c) and (d) are provided by the BACKMAP module. In Panel (c), β -sheets are shown in blue and all helices are shown in red. In Panel (d), right-handed and left-handed backbone twists are shown as red and blue respectively.

263 **Stand Alone Usage**

264 BACKMAP can be used as a stand alone package by running ‘> python -m backmap -pdb <pdb_dir_or_file>’.
 265 The sections below describes the expected outputs and how they may be interpreted.

266 **Stand Alone Example I: A Stable Protein**

267 Panels (b) through (f) of Fig. 10 below were created by running ‘> python -m backmap ./tests/pdfs/1xqq.pdb’
 268 (Panel (a) was created using VMD). These graphs indicate that protein 1xqq describes a conformationally
 269 stable protein, since each residue fluctuates little in color (structure) over ‘time’ (c,d; here and below, it is
 270 assumed that discrete models represent distinct states of the protein over ‘time’), show little change in the
 271 \mathcal{R} histogram over time (b) and show few enduring fluctuations in RMSD (e) and RMSF (f).

272 In particular, each column in Panel (b) describes the histogram in Ramachandran number (\mathcal{R}) space
 273 for a single model/timeframe. These histograms show the presence of both α -helices (at $\mathcal{R} \approx 0.34$)
 274 and β -sheets (at $\mathcal{R} \approx 0.52$). Additionally, Panels (c) and (d) describe per-residue conformational plots
 275 (colored by two different metrics or CMAPs), which show that most of the protein backbone remains
 276 relatively stable over time (e.g., few fluctuations in state or ‘color’ are evident over frame #). Finally,
 277 Panel (e) describes the extent towards which a single residue’s state has deviated from the first frame,
 278 and Panel (f) describes the extent towards which a single residue’s state has deviated from its state in the
 279 previous frame. All these graphs, show that this protein is relatively conformationally stable.

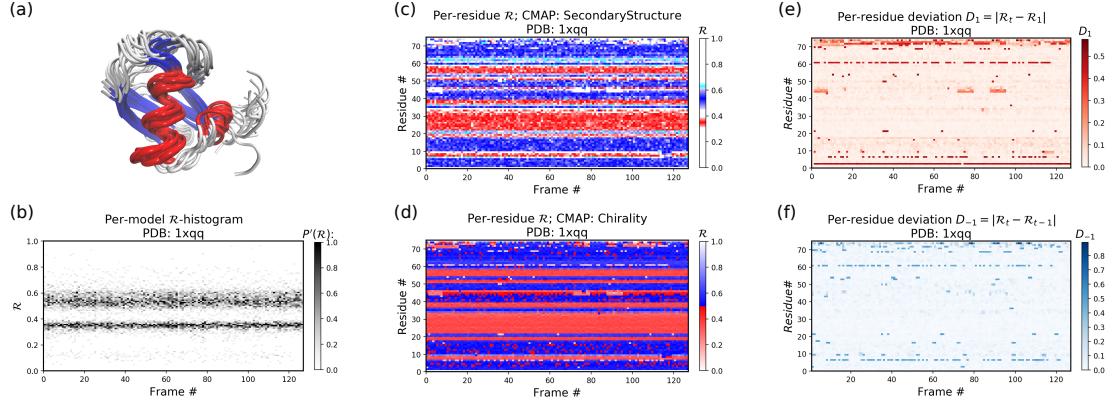


Figure 10. Protein 1xqq describes a stable protein. Panel (a) represents the entire ensemble, Panel(b) represents a histogram distribution of \mathcal{R} , Panels (c) and (d) represent two ways color per-residue \mathcal{R} plots, and Panels (e) and (f) are two ways to describe backbone fluctuation over time.

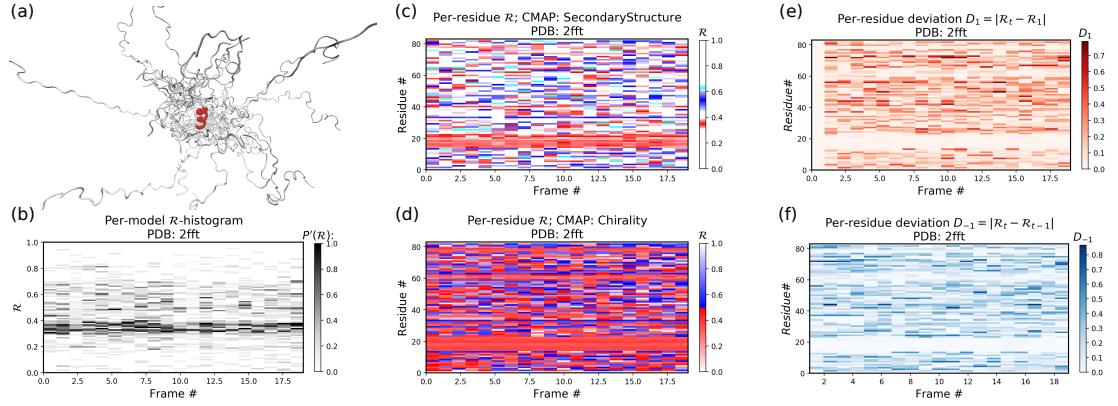


Figure 11. Protein 2fft describes an intrinsically disordered protein, with one stable helix in red. Descriptions of each panel are identical to that of Fig. 10.

280 Stand Alone Example II: An Intrinsically Disordered Protein

281 Fig. 11 is identical to Fig. 10, except that the panels pertain to an intrinsically disordered protein 2fft
282 whose structural ensemble describes dramatically distinct conformations.

283 As compared to the conformationally stable protein above, protein 2fft is much more flexible. Panel
284 (b) shows that the states accessed per model are diverse and dramatically fluctuate over the entire range of
285 \mathcal{R} (this is especially true when compared to a stable protein, see Fig. 10b).

286 The diverse states occupied by each residue (Panels (c) and (d)) confirm the conformational variation
287 displayed by most of the backbone (Panels (e) and (f)) similarly show how most of the residues fluctuate
288 dramatically.

289 Yet, interestingly, Panels (c) through (f) also show an **unusually** stable region – residues 15 through
290 25 – which consistently display the same conformational (α -helical) state at $\mathcal{R} \approx 0.34$ (interpreted as the
291 color red in Panel (c)). This trend would be hard to recognize by simply looking at the structural ensemble
292 (Panel (a)).

293 A signed Ramachandran number for ‘misbehaving’ backbones

294 The Ramachandran number increases in value from the bottom left of the Ramachandran plot to the
295 top right in sweeps that are parallel to the negative sloping diagonal. As discussed in Mannige *et al.*
296 (2016), this method of mapping a two-dimensional space into one number is still structurally meaningful
297 and descriptive because 1) most structural features of the protein backbone – e.g. radius of gyration
298 (Mannige *et al.*, 2016), end-to-end distance (Mannige *et al.*, 2016), and chirality (Mannige, 2017) – vary
299 little along lines parallel to the negatively-sloping diagonal (this is indicated by relatively small standard
300 deviations in structural metrics for similar \mathcal{R} s; Fig. 4), and 2) most protein backbones display chiral

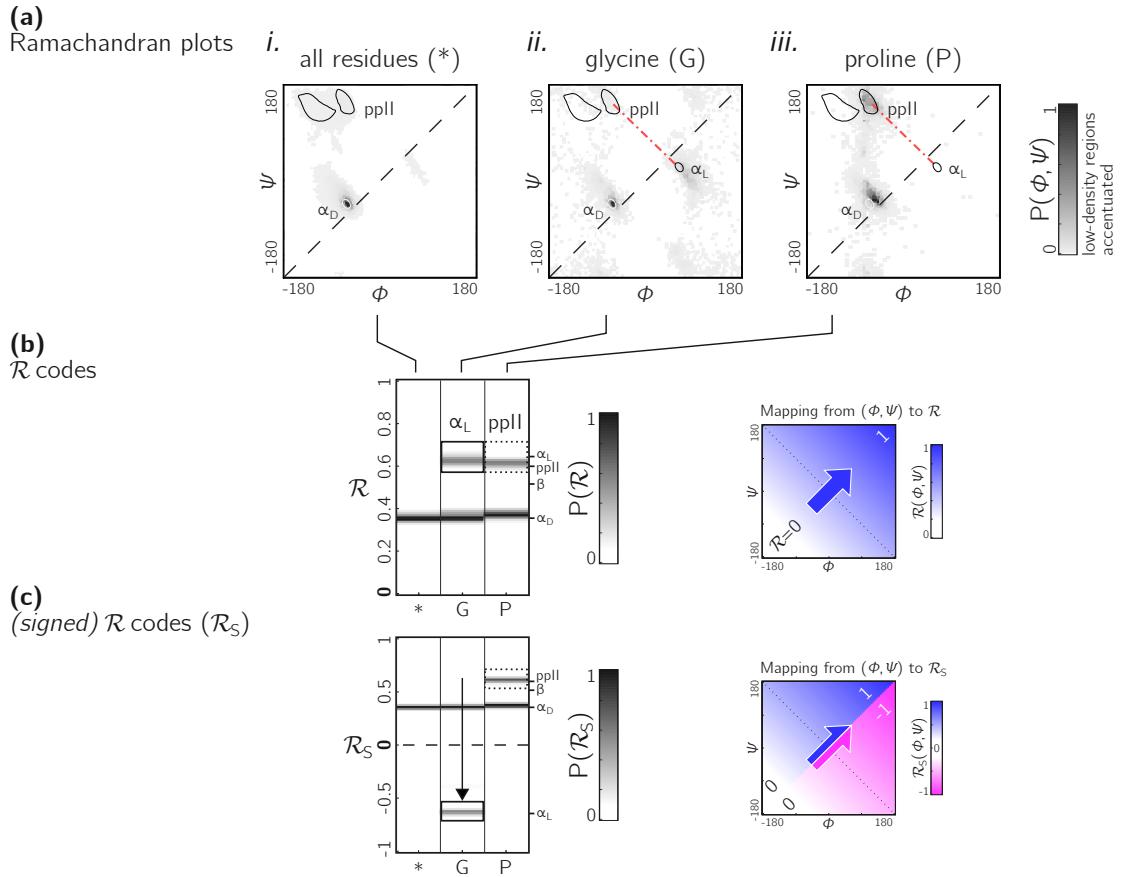


Figure 12. Signed \mathcal{R}_S are required for non-chiral backbones. While the backbones of most amino acids occupy the top of the positively sloped diagonal (dashed in b), non chiral amino acids such as Glycines (or their N-substituted variants – peptoids) display no such preference, which causes distinct secondary structures that lie on the same ‘sweep’ to be localized at similar regions in \mathcal{R} (e.g., in b, polyproline-II and α_D helices both localize at $\mathcal{R} \approx 0.6$). However, a signed Ramachandran number (\mathcal{R}_S) solves this overlap by multiplying those \mathcal{R} ’s derived from backbones with $\phi > \psi$ by -1 . The resolving power of \mathcal{R}_S is evident by the separation of polyproline-II and α_D helices (c). The mapping of (ϕ, ψ) to \mathcal{R} and \mathcal{R}_S are shown to the right of each respective \mathcal{R} -plot (b,c).

301 centers and therefore predominantly appear on the top left region of the Ramachandran plot (above the
302 dashed diagonal in Fig. 12a-(i)).

303 However, not all backbones localize in only one half of the Ramachandran plot. Particularly, among
304 biologically relevant amino acids, glycine occupies both regions of the Ramachandran plot (Fig. 12a-(ii);
305 of note, the α_L helix region becomes relatively prominent). On the other hand, prolines are known to form
306 polyproline-II helices (ppII in Fig. 12a-(iii)), which falls on almost the same ‘sweep’ as glycine rich pep-
307 tides (red dot-dashed line). In situations where both prolines and glycines are abundant, the Ramachandran
308 number (\mathcal{R}) would fail to distinguish α_L from ppII (Fig. 12b; regions outlined by rectangles).

To accomodate the situation where achiral backbones are expected (eg., if peptoids or polyglycines are being studied), an additional Ramachandran number – the *signed* Ramachandran number \mathcal{R}_S – is introduced here. \mathcal{R}_S is identical to the original number in magnitude, but which changes sign from + to – as you approach \mathcal{R} numbers that are to the right (or below) the positively sloped diagonal. I.e.,

$$\mathcal{R}_S = \begin{cases} \mathcal{R} & , \text{if } \psi \geq \phi \\ \mathcal{R} \times -1 & , \text{if } \psi < \phi \end{cases} \quad (3)$$

309 As an example of the utility of \mathcal{R}_S , Fig. 12c shows that \mathcal{R}_S easily distinguishes α_D from ppII.

310 Note that the signed \mathcal{R}_S , while useful, would be important in very limited scenarios, as more than

311 96% of the amino acids in the Protein Databank (PDB) occupy the upper-left region of the Ramachandran
312 plot (with the 3% of ‘rule breakers’ contributed mostly by glycines).

313 CONCLUSION

314 A simpler Ramachandran number is reported – $\mathcal{R} = (\phi + \psi + 2\pi)/(4\pi)$ – which, while being a single
315 number, provides much information. For example, as discussed in Mannige *et al.* (2016), \mathcal{R} values
316 above 0.5 are left-handed **in twist**, while those below 0.5 are right handed, \mathcal{R} values close to 0, 0.5 and
317 1 are extended, β -sheets occupy \mathcal{R} values at around 0.52, right-handed α -helices hover around 0.34.
318 Given the Ramachandran number’s ‘stackability’, single graphs can hold detailed information of the
319 progression/evolution of molecular trajectories. Indeed, Fig. 8 shows how 400 distinct Ramachandran
320 plots can easily be fit into one graph when using \mathcal{R} . Finally, a python script/module (BACKMAP) has
321 been provided in an online [GitHub repository](#) to promote the utility of \mathcal{R} as a universal metric.

322 MATERIALS

323 Statistics about single amino acid conformations and secondary structures (excepting polyproline II
324 helices) were derived from the Structural Classification of Proteins or SCOPe website [Release 2.06; Fox
325 *et al.* (2014)]. This database, currently available at <http://scop.berkeley.edu/downloads/pdbstyle/pdbstyle-sel-gs-bib-40-2.06.tgz>, contains 13,760 three-dimensional protein
326 conformations (one domain per conformation) with lower than 40% sequence identity. Secondary structure
327 annotations were assigned using the DSSP algorithm (Kabsch and Sander, 1983), although the STRIDE
328 algorithm (Frishman and Argos, 1995) provides qualitatively identical distributions. **These statistics were**
329 **used to produce distributions within Fig. 2a and Fig. 3a,c.**

330 Given the absence of polyproline II helix (ppII) annotation in the present version of DSSP, statistics
331 for polyproline II helices (**used to generate the ppII distributions in Fig. 2a and Fig. 3a,c**) were obtained
332 from segments within 16,535 proteins annotated by PolyprOnline (Chebrek *et al.*, 2014) to contain three
333 or more residues of the secondary structure.

334 Fig. 6 represents a trajectory of a portion of a single peptoid backbone within a ‘relaxing’ peptoid
335 nanosheet bilayer. The conformation of this backbone – derived from work by Mannige *et al.* (2015) and
336 Mannige *et al.* (2016) – is also available as ‘/tests/pdbs/nanosheet_birth_U7.pdb’ within the companion
337 [GitHub repository](#).

338 The following protein structures were obtained from the Protein DataBank (PDB): 1mba, 2acy,
339 1xqq, and 2fft. The first two in the list (1mba, 2acy) describe single conformations and the last two
340 (1xqq, 2fft) describe ensembles. \mathcal{R} -based multi-angle pictures (MAPs) were created for each structure
341 $X \in [\text{nanosheet_birth_U7.pdb}, 2fft, 2acy, 1xqq, 1mba]$ using the following command line code:

342 > python -m backmap -pdb tests/pdbs/X.pdb

343 The output of this command line implementation were used in panels (b) onwards of Figs. 5, 6, 9, 10
344 and 11.

345 In order to describe change in structural, this report uses two metrics for structural deviation: deviation
346 in structure when compared to the first conformation in the trajectory (D_1), and the previous conformation
347 in the trajectory (D_{-1}). For any residue r at time t , these equations can be described as follows:

$$D_1 = |\mathcal{R}_t - \mathcal{R}_1|, \quad D_{-1} = |\mathcal{R}_t - \mathcal{R}_{t-1}|. \quad (4)$$

348 All three-dimensional representation of proteins (Panel (a) in Figs. 5, 6, 9, 10 and 11) were created
349 using VMD (Humphrey *et al.*, 1996). Finally, all other figures – excepting Fig. 1 that is derived from Man-
350 nige *et al.* (2016) – were created using helper Python scripts available in [manuscript/python_generators/](#)
351 within the companion [GitHub repository](#).

352 ACKNOWLEDGMENTS

353 During the development of this paper, RVM was partially supported by the Defense Threat Reduction
354 Agency under contract no. IACRO-B0845281. RVM thanks Alana Canfield Mannige for her critique. The
355 notion of the signed Ramachandran number emerged from discussions with Joyjit Kundu and Stephen
356 Whitelam while at the Molecular Foundry at Lawrence Berkeley National Laboratory (LBNL). This work
357 was partially done at the Molecular Foundry at LBNL, supported by the Office of Science, Office of Basic
358 Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

357 **APPENDIX**

358 **Simplifying the Ramachandran number (\mathcal{R})**

359 This section will derive the simplified Ramachandran number presented in this paper from the more
360 complicated looking Ramachandran number introduced previously (Mannige *et al.*, 2016).

Assuming the bounds $\phi \in [\phi_{\min}, \phi_{\max}]$ and $\psi \in [\psi_{\min}, \psi_{\max}]$, the previously described Ramachandran number takes the form

$$\mathcal{R}(\phi, \psi) \equiv \frac{R_{\mathbb{Z}}(\phi, \psi) - R_{\mathbb{Z}}(\phi_{\min}, \phi_{\min})}{R_{\mathbb{Z}}(\phi_{\max}, \phi_{\max}) - R_{\mathbb{Z}}(\phi_{\min}, \phi_{\min})}, \quad (5)$$

where, $\mathcal{R}(\phi, \psi)$ is the Ramachandran number with range $[0, 1]$, and $R_{\mathbb{Z}}(\phi, \psi)$ is the *unnormalized* integer-spaced Ramachandran number whose closed form is

$$R_{\mathbb{Z}}(\phi, \psi) = \left\lfloor (\phi - \psi + \lambda)\sigma/\sqrt{2} \right\rfloor + \left\lfloor \sqrt{2}\lambda\sigma \right\rfloor \left\lfloor (\phi + \psi + \lambda)\sigma/\sqrt{2} \right\rfloor. \quad (6)$$

361 Here, $\lfloor x \rfloor$ rounds x to the closest integer value, σ is a scaling factor, discussed below, and λ is the
362 range of an angle in degrees (i.e., $\lambda = \phi_{\max} - \phi_{\min}$). Effectively, this equation does the following. 1) It
363 divides up the Ramachandran plot into $(360^\circ \sigma^{1/2})^2$ squares, where σ is a user-selected scaling factor
364 that is measured in reciprocal degrees [see Fig. 8b in Mannige *et al.* (2016)]. 2) It then assigns integer
365 values to each square by setting the lowest integer value to the bottom left of the Ramachandran plot
366 ($\phi = -180^\circ, \psi = -180^\circ$) and proceeding from the bottom left to the top right by iteratively slicing down
367 -1/2 sloped lines and assigning increasing integer values to each square that one encounters. 3) Finally,
368 the equation assigns any (ϕ, ψ) pair within $\phi, \psi \in [-\phi_{\min}, \phi_{\max}]$ to the integer value ($R_{\mathbb{Z}}$) assigned to the
369 divvied-up square that they it exists in.

Combining the two equations (Eqns. 5 and 6) results in the following, rather imposing, equation for the Ramachandran number:

$$\mathcal{R}(\phi, \psi) = \frac{\begin{pmatrix} \left\lfloor (\phi - \psi + \lambda)\sigma/\sqrt{2} \right\rfloor & + \left\lfloor \sqrt{2}\lambda\sigma \right\rfloor \left\lfloor (\phi + \psi + \lambda)\sigma/\sqrt{2} \right\rfloor \\ - \left\lfloor (\phi_{\min} - \psi_{\min} + \lambda)\sigma/\sqrt{2} \right\rfloor & - \left\lfloor \sqrt{2}\lambda\sigma \right\rfloor \left\lfloor (\phi_{\min} + \psi_{\min} + \lambda)\sigma/\sqrt{2} \right\rfloor \end{pmatrix}}{\begin{pmatrix} \left\lfloor (\phi_{\max} - \psi_{\max} + \lambda)\sigma/\sqrt{2} \right\rfloor & + \left\lfloor \sqrt{2}\lambda\sigma \right\rfloor \left\lfloor (\phi_{\max} + \psi_{\max} + \lambda)\sigma/\sqrt{2} \right\rfloor \\ - \left\lfloor (\phi_{\min} - \psi_{\min} + \lambda)\sigma/\sqrt{2} \right\rfloor & - \left\lfloor \sqrt{2}\lambda\sigma \right\rfloor \left\lfloor (\phi_{\min} + \psi_{\min} + \lambda)\sigma/\sqrt{2} \right\rfloor \end{pmatrix}} \quad (7)$$

However useful Eqn. 7 is, the complexity of the equation may be a deterrent towards utilizing it. This paper reports a simpler equation that is derived by taking the limit of Eqn. 7 as σ tends towards ∞ . In particular, when $\sigma \rightarrow \infty$, Eqn. 7 becomes

$$\mathcal{R}(\phi, \psi) = \lim_{\sigma \rightarrow \infty} \bar{\mathcal{R}}(\phi, \psi) = \frac{\phi + \psi - (\psi_{\min} + \psi_{\max})}{(\phi_{\max} + \psi_{\max}) - (\phi_{\min} + \psi_{\min})}. \quad (8)$$

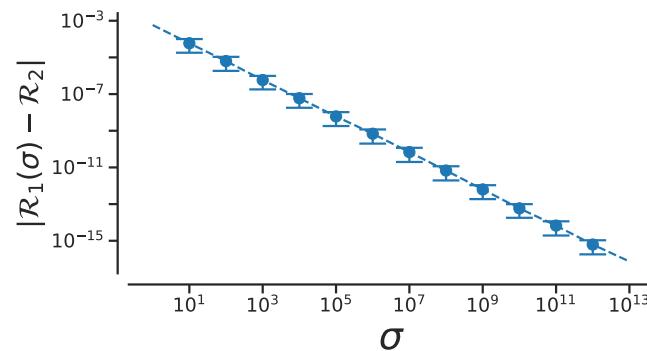


Figure 13. The increase in the accuracy measure (σ) for the original Ramachandran number (Eqn. 6) results in values that tend towards the new Ramachandran number in this paper (Eqn. 2).

370 Assuming that $\phi, \psi \in [-180^\circ, 180^\circ]$ or $[-\pi, \pi]$,

$$\mathcal{R}(\phi, \psi) = \frac{\phi + \psi + 2\pi}{4\pi}. \quad (9)$$

371 Conformation of this limit is shown numerically in Fig. 13. Since larger σ s indicate higher accuracy,
372 $\lim_{\sigma \rightarrow \infty} \mathcal{R}(\phi, \psi)$ represents an exact representation of the Ramachandran number. Using this closed form,
373 this report shows how both static structural features and complex structural transitions may be identified
374 with the help of Ramachandran number-derived plots.

Assuming, a different range (say, $\phi, \psi \in [0, 2\pi]$), the Ramachandran number in that frame of reference
will be

$$\mathcal{R}(\phi, \psi)_{\phi, \psi \in [0, 2\pi]} = \frac{\phi + \psi}{4\pi}. \quad (10)$$

375 However, in changing the ranges, the meaning of the Ramachandran number will change. This manuscript
376 assumes that all angles (ϕ, ψ, ω) range between $-\pi$ (-180°) and π (180°)

377 REFERENCES

- 378 **Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2002.** Molecular biology of the cell.
379 new york: Garland science; 2002. *Classic textbook now in its 5th Edition*.
- 380 **Baruah A, Rani P, Biswas P. 2015.** Conformational entropy of intrinsically disordered proteins from
381 amino acid triads. *Scientific reports* **5**.
- 382 **Beck DA, Alonso DO, Inoyama D, Daggett V. 2008.** The intrinsic conformational propensities of the
383 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proceedings of the
384 National Academy of Sciences* **105**(34):12259–12264.
- 385 **Berg JM, Tymoczko JL, Stryer L. 2010.** *Biochemistry, International Edition*. WH Freeman & Co.,
386 New York, 7 edition.
- 387 **Chebrek R, Leonard S, de Brevern AG, Gelly JC. 2014.** Polyproline: polyproline helix ii and
388 secondary structure assignment database. *Database* **2014**:bau102.
- 389 **Dunker A, Babu M, Barbar E, Blackledge M, Bondos S, Dosztányi Z, Dyson H, Forman-Kay J,
390 Fuxreiter M, Gsponer J, Han KH, Jones D, Longhi S, Metallo S, Nishikawa K, Nussinov R,
391 Obradovic Z, Pappu R, Rost B, Selenko P, Subramaniam V, Sussman J, Tompa P, Uversky V.
392 2013. What's in a name? why these proteins are intrinsically disordered? *Intrinsically Disordered
393 Proteins* **1**:e24157.**
- 394 **Espinosa-Fonseca LM. 2009.** Reconciling binding mechanisms of intrinsically disordered proteins.
395 *Biochemical and biophysical research communications* **382**(3):479–482.
- 396 **Fink AL. 2005.** Natively unfolded proteins. *Curr Opin Struct Biol* **15**(1):35–41.
- 397 **Fox NK, Brenner SE, Chandonia JM. 2014.** Scope: Structural classification of proteins—extended,
398 integrating scop and astral data and classification of new structures. *Nucleic Acids Res* **42**(Database
399 issue):D304–D309. doi:10.1093/nar/gkt1240.
- 400 **Frishman D, Argos P. 1995.** Knowledge-based protein secondary structure assignment. *Proteins:
401 Structure, Function, and Bioinformatics* **23**(4):566–579.
- 402 **Geist L, Henen MA, Haiderer S, Schwarz TC, Kurzbach D, Zawadzka-Kazimierczuk A, Saxena
403 S, Żerko S, Koźmiński W, Hinderberger D, et al. 2013.** Protonation-dependent conformational
404 variability of intrinsically disordered proteins. *Protein Science* **22**(9):1196–1205.
- 405 **Gunasekaran K, Nagarajaram H, Ramakrishnan C, Balaram P. 1998.** Stereochemical punctuation
406 marks in protein structures: glycine and proline containing helix stop signals. *Journal of molecular
407 biology* **275**(5):917–932.
- 408 **Ho BK, Brasseur R. 2005.** The ramachandran plots of glycine and pre-proline. *BMC structural biology*
409 **5**(1):1.
- 410 **Hooft RW, Sander C, Vriend G. 1997.** Objectively judging the quality of a protein structure from a
411 ramachandran plot. *Computer applications in the biosciences: CABIOS* **13**(4):425–430.
- 412 **Humphrey W, Dalke A, Schulten K. 1996.** VMD – Visual Molecular Dynamics. *Journal of Molecular
413 Graphics* **14**:33–38.

- 414 **James LC, Tawfik DS.** 2003a. Conformational diversity and protein evolution—a 60-year-old hypothesis
415 revisited. *Trends Biochem Sci* **28**(7):361–368.
- 416 **James LC, Tawfik DS.** 2003b. The specificity of cross-reactivity: promiscuous antibody binding involves
417 specific hydrogen bonds rather than nonspecific hydrophobic stickiness. *Protein Sci* **12**(10):2183–2193.
- 418 **Kabsch W, Sander C.** 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-
419 bonded and geometrical features. *Biopolymers* **22**(12):2577–2637. doi:10.1002/bip.360221211.
- 420 **Kosol S, Contreras-Martos S, Cedeño C, Tompa P.** 2013. Structural characterization of intrinsically
421 disordered proteins by nmr spectroscopy. *Molecules* **18**(9):10802–10828.
- 422 **Laskowski RA.** 2003. Structural quality assurance. *Structural Bioinformatics, Volume 44* pages 273–303.
- 423 **Laskowski RA, MacArthur MW, Moss DS, Thornton JM.** 1993. Procheck: a program to check the
424 stereochemical quality of protein structures. *Journal of applied crystallography* **26**(2):283–291.
- 425 **Lovell SC, Davis IW, Arendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS,
426 Richardson DC.** 2003. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*
427 **50**(3):437–450. ISSN 1097-0134. doi:10.1002/prot.10286.
- 428 **Mannige RV.** 2014. Dynamic new world: Refining our view of protein structure, function and evolution.
429 *Proteomes* **2**(1):128–153.
- 430 **Mannige RV.** 2017. An exhaustive survey of regular peptide conformations using a new metric for
431 backbone handedness (*h*). *PeerJ* **5**:e3327. ISSN 2167-8359. doi:10.7717/peerj.3327.
- 432 **Mannige RV, Haxton TK, Proulx C, Robertson EJ, Battigelli A, Butterfoss GL, Zuckermann RN,
433 Whitelam S.** 2015. Peptoid nanosheets exhibit a new secondary structure motif. *Nature* **526**:415–420.
- 434 **Mannige RV, Kundu J, Whitelam S.** 2016. The Ramachandran number: an order parameter for protein
435 geometry. *PLoS One* **11**(8):e0160023.
- 436 **Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN.** 2009. Protein disorder in the human
437 diseasome: unfoldomics of human genetic diseases. *BMC Genomics* **10 Suppl 1**:S12. doi:10.1186/
438 1471-2164-10-S1-S12.
- 439 **Momen R, Azizi A, Wang L, Yang P, Xu T, Kirk SR, Li W, Manzhos S, Jenkins S.** 2017. The role
440 of weak interactions in characterizing peptide folding preferences using a qtaim interpretation of the
441 ramachandran plot (ϕ - ψ). *International Journal of Quantum Chemistry* .
- 442 **Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK.** 2005. Comparing and
443 combining predictors of mostly disordered proteins. *Biochemistry* **44**(6):1989–2000.
- 444 **Ramachandran G, Ramakrishnan C, Sasisekharan V.** 1963. Stereochemistry of polypeptide chain
445 configurations. *Journal of molecular biology* **7**(1):95–99.
- 446 **Schad E, Tompa P, Hegyi H.** 2011. The relationship between proteome size, structural disorder and
447 organism complexity. *Genome Biol* **12**(12):R120.
- 448 **Sibille N, Bernado P.** 2012. Structural characterization of intrinsically disordered proteins by the
449 combined use of nmr and saxs. *Biochemical society transactions* **40**(5):955–962.
- 450 **Subramanian E.** 2001. On ramachandran. *Nature Structural & Molecular Biology* **8**(6):489–491.
- 451 **Tien MZ, Sydykova DK, Meyer AG, Wilke CO.** 2013. Peptidebuilder: A simple python library to
452 generate model peptides. *PeerJ* **1**:e80.
- 453 **Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL Jr.** 2010. Neighbor-dependent
454 ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process
455 model. *PLOS Computational Biology* **6**(4):1–21. doi:10.1371/journal.pcbi.1000763.
- 456 **Tokuriki N, Tawfik DS.** 2009. Protein dynamism and evolvability. *Science* **324**(5924):203–207.
- 457 **Tompa P.** 2011. Unstructural biology coming of age. *Curr Opin Struct Biol* **21**(3):419–425. doi:
458 10.1016/j.sbi.2011.03.012.
- 459 **Uversky VN.** 2003. Protein folding revisited. a polypeptide chain at the folding-misfolding-nonfolding
460 cross-roads: which way to go? *Cell Mol Life Sci* **60**(9):1852–1871.
- 461 **Uversky VN, Dunker AK.** 2010. Understanding protein non-folding. *Biochim Biophys Acta*
462 **1804**(6):1231–1264.
- 463 **Vértessy BG, Orosz F.** 2011. From “fluctuation fit” to “conformational selection”: evolution, rediscovery,
464 and integration of a concept. *Bioessays* **33**(1):30–34.