

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

4      Ranjan V. Mannige<sup>1,\*</sup>

5      <sup>1</sup> Multiscale Institute, Berkeley Lake, GA 30092, U.S.A.

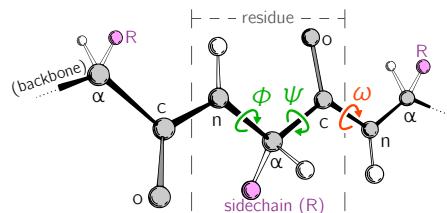
6      \* ranjanmannige@gmail.com

## 7      ABSTRACT

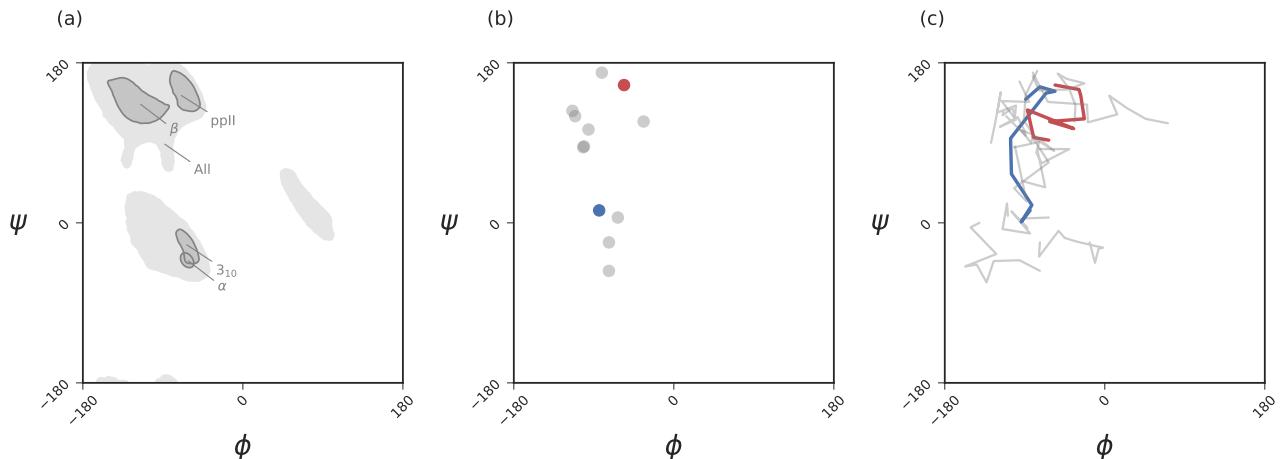
8      Protein backbones occupy diverse conformations, but compact metrics to describe such conformations  
9      and transitions between them have been missing. This report re-introduces the Ramachandran number  
10     ( $\mathcal{R}$ ) as a residue-level structural metric that could simply the life 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 20  
24      naturally occurring amino acids. Proteins are important partially because of the structures that they access:  
25      the conformations (conformational ensemble) that a protein assumes determines the functions available  
26      to that protein. However, all proteins are dynamic: even stable proteins undergo long-range motions  
27      in its equilibrium state; i.e., they have substantial diversity in their conformational ensemble (Mannige,  
28      2014). Additionally, a number of proteins undergo conformational transitions, without which they may  
29      not properly function. Finally, some proteins – intrinsically disordered proteins – display massive disorder  
30      whose conformations dramatically change over time (Uversky, 2003; Fink, 2005; Midic *et al.*, 2009;  
31      Espinoza-Fonseca, 2009; Uversky and Dunker, 2010; Tompa, 2011; Sibille and Bernado, 2012; Kosol  
32      *et al.*, 2013; Dunker *et al.*, 2013; Geist *et al.*, 2013; Baruah *et al.*, 2015), and whose characteristic  
33      structures are still not well-understood (Beck *et al.*, 2008).



**Figure 1. Backbone conformational degrees of freedom** dominantly depend on the dihedral angles  $\phi$  and  $\psi$  (green), and to a smaller degree depend on the third dihedral angle ( $\omega$ ; red) as well as bond lengths and angles (unmarked).



**Figure 2.** While the Ramachandran plot is useful for getting a *qualitative* sense of peptide backbone structure (a, c), it is not a convenient representation for exploring peptide backbone dynamics (c). Secondary structure keys used here and throughout the document: ‘ $\alpha$ ’ –  $\alpha$ -helix, ‘ $3_{10}$ ’ –  $3_{10}$ -helix, ‘ $\beta$ ’ –  $\beta$ -sheet/extension, ‘ppII’ – polyproline II helix.

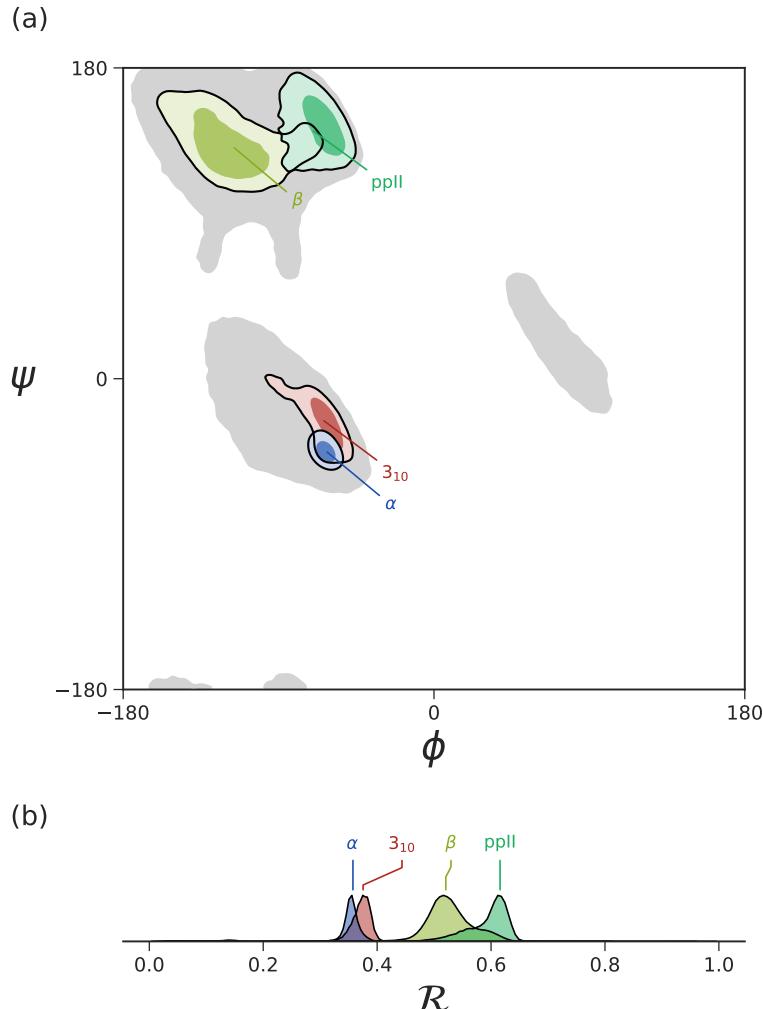
34 Large-scale changes in a protein occur due to changes in protein backbone conformations. Fig. 1 is a  
 35 cartoon representation of a peptide/protein backbone, with the backbone bonds themselves represented  
 36 by darkly shaded bonds. Ramachandran *et al.* (1963) had recognized that the backbone conformational  
 37 degrees of freedom available to an amino acid (residue)  $i$  is almost completely described by only two  
 38 dihedral angles:  $\phi_i$  and  $\psi_i$  (Fig. 1, green arrows). Today, protein structures described in context of the  
 39 two-dimensional  $(\phi, \psi)$ -space are called Ramachandran plots.

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

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

57 Consequently, there has been no single compact descriptor of protein structure. This impedes the  
 58 naïve or hypothesis-free exploration of new trajectories/ensembles. For example, tracking changes in  
 59 protein trajectory is either overly detailed or overly holistic: an example of an overly detailed study is the  
 60 tracking on exactly one or a few atoms over time (this already poses a problem, since we would need  
 61 to know exactly which atoms are expected to partake in a transition); an example of a holistic metric is  
 62 the radius of gyration (this also poses a problem, since we will never know which residues contribute  
 63 to a change in radius of gyration without additional interrogation). With our understanding of protein  
 64 dynamics undergoing a new renaissance – especially due to intrinsically disordered proteins and allosteric  
 65 – having hypothesis-agnostic yet detailed (residue-level) metrics of protein structure has become even  
 66 more relevant.

67 It has recently been shown that the two Ramachandran backbone parameters  $(\phi, \psi)$  may be conve-  
 68 niently combined into a single number – the Ramachandran *number*  $[\mathcal{R}(\phi, \psi)]$  or simply  $\mathcal{R}$  – with little



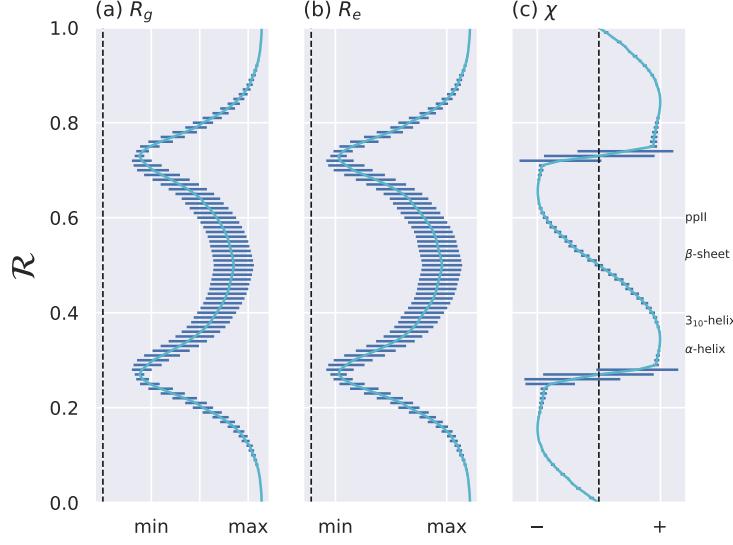
**Figure 3.** The distribution of dominant regular secondary structures are shown in  $[\phi, \psi]$ -space (a) and in  $\mathcal{R}$ -space (b). Ramachandran plots (a) and Ramachandran ‘lines’ (b) equally resolve the secondary structure space, thereby making  $\mathcal{R}$  a compact yet faithful representation of backbone structure (Mannige *et al.*, 2016).

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 MAPs that describe the behavior of a protein backbone within user-inputted conformations, structural ensembles and trajectories. This package 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  $\phi$  and  $\psi$  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 most accurate version of the Ramachandran number is introduced. Section 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



**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 ( $\chi$ ; 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  $R_g$  and  $R_e$  were obtained by computationally generating poly-glycine peptides of length 10 for all possible  $\phi$  and  $\psi \in [-180, -175, \dots, 175, 180]$ . This was done using the Python library PeptideBuilder (Tien *et al.*, 2013). Values for  $R_g$  and  $R_e$  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.

values differ by  $360^\circ$ , 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 number is both compact enough and informative enough to generate immediately useful graphs (multi-angle pictures or MAPs) of a dynamic protein backbone.

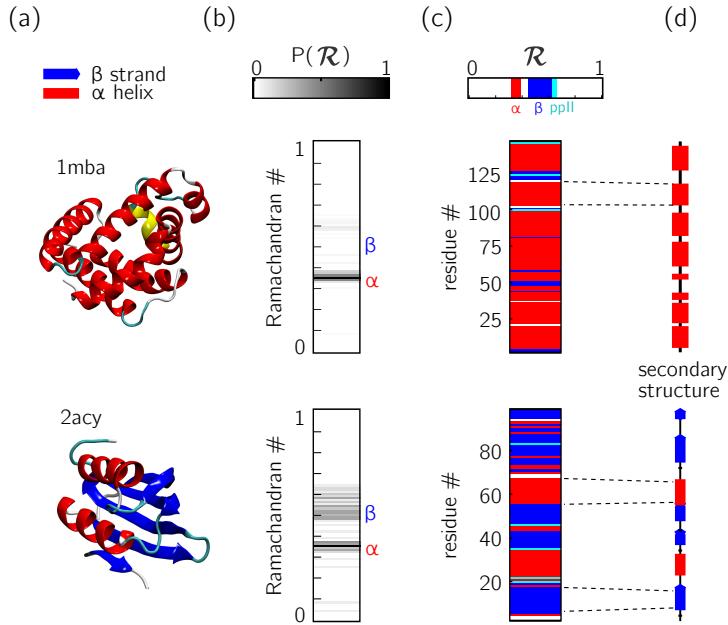
## REASON TO USE THE RAMACHANDRAN NUMBER

### Ramachandran numbers are structurally meaningful

In addition to resolving positions of secondary structures (Fig. 3),  $\mathcal{R}$  relate well to structural measures such as radius of gyration ( $R_g$ ), end-to-end distance ( $R_e$ ) and chirality ( $\chi$ ). These relationships are shown in Fig. 4.

### Ramachandran numbers are more compact than one might realize

An important aspect of the Ramachandran number ( $\mathcal{R}$ ) lies in its compactness compared to the traditional Ramachandran pair  $(\phi, \psi)$ . Say we have an  $N$ -residue peptide. Then, switching from  $(\phi, \psi)$  to  $\mathcal{R}$  appears to only reduce the number of variables from  $2N$  to  $N$ , and hence by half. However,  $(\phi, \psi)$  values are coupled, i.e., for any  $N$ -length peptide, any ordering of  $[\phi_1, \phi_2, \dots, \phi_N, \psi_1, \psi_2, \dots, \psi_N]$  can not describe



**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  $\alpha$ -helical, while [2acy](#) is likely a mix of  $\alpha$ -helices and  $\beta$ -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.

the structure, it is only *pairs* –  $[(\phi_1, \psi_1), (\phi_2, \psi_2), \dots, (\phi_N, \psi_N)]$  – that can. Therefore, we must think of switching from  $(\phi, \psi)$ -space to  $\mathcal{R}$ -space as a switch in structure space per residue from  $N$  two-tuples  $(\phi_i, \psi_i)$  that reside in  $\phi \times \psi$  space to  $N$  single-dimensional numbers ( $\mathcal{R}_i$ ).

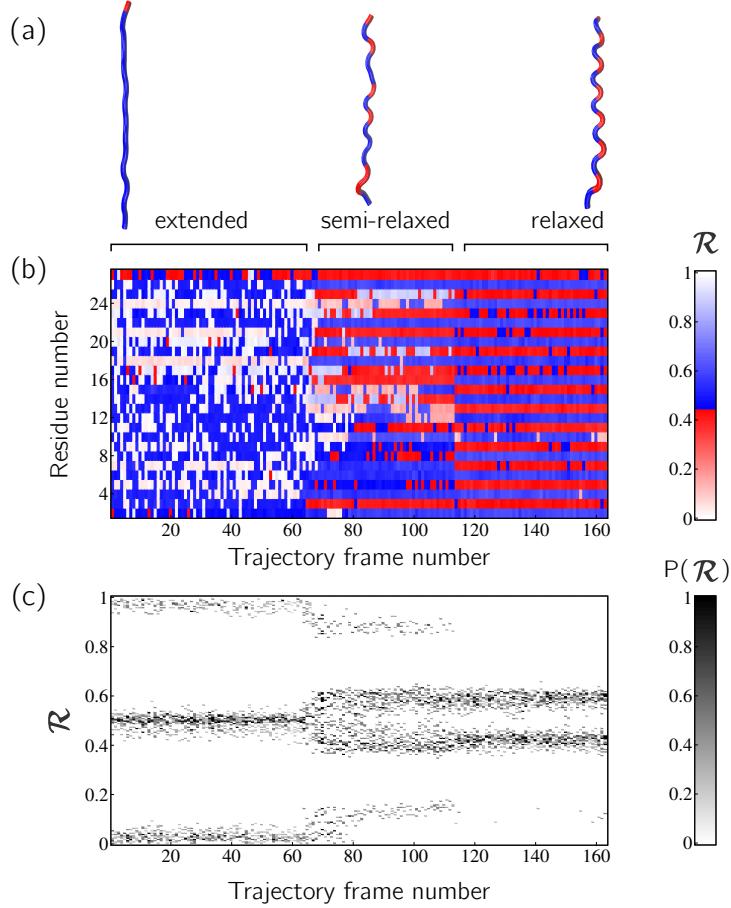
The value of this conversion is that the structure of a protein can be described in various one-dimensional arrays (per-structure “Ramachandran codes” or “ $\mathcal{R}$ -codes”), which, when arranged vertically/columnnally, describe easy to digest/interpret structural patterns. See, e.g., Fig. 5.

### Ramachandran codes are stackable

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. 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



**Figure 6. Stacked  $\mathcal{R}$ -codes provide useful information at a glance.**

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.

By utilizing  $\mathcal{R}$ , maps such as those in Fig. 6 provide information about every  $\phi$  and  $\psi$  within the backbone. As such, these maps are dubbed MAPs, for Multi Angle Pictures. A Python package called 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 structure, or multiple structures separated by the code ‘MODEL’.

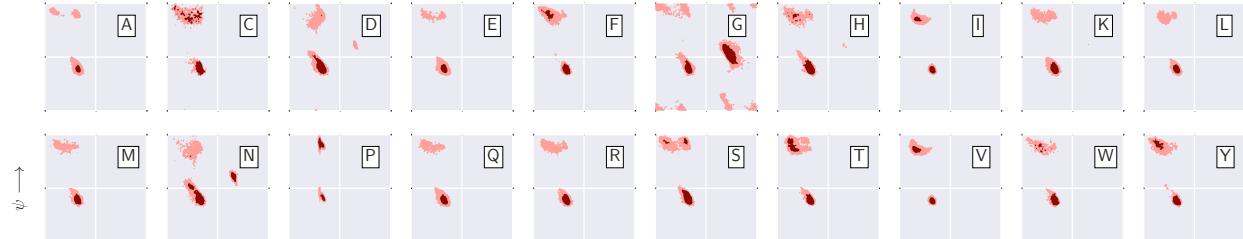
#### Case study: picking out subtle differences from high volume of data

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

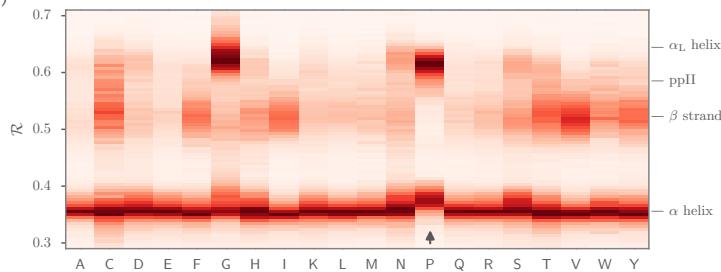
It is also known that amino acids preceding prolines display unusual shift in backbone twist/chirality. For example, Fig. 8 shows that amino acids appearing before prolines and glycines behave differently than they would otherwise (discussed further in the figure caption). While these results have been discussed previously (Gunasekaran *et al.*, 1998; Ho and Brasseur, 2005), they were reported more than 30 years after the first structures were published; they would have been relatively easy to find if  $\mathcal{R}$ -codes were to be used regularly.

The relationships in Figs. 7 and 8 show how subtle changes in structure can be easily picked out when structures are stacked side-by-side in the form of  $\mathcal{R}$ -codes. Such subtle changes are often witnessed when

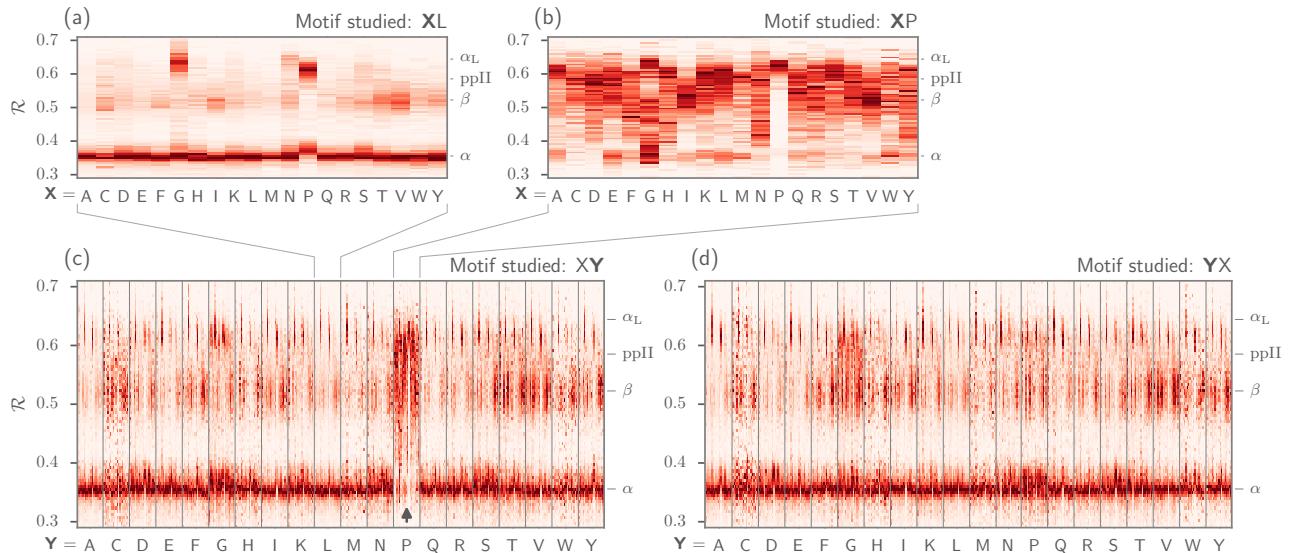
(a)



(b)



**Figure 7. Ramachandran lines are stackable – Part I.** 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  $\alpha$ -helical region of the proline plot (P). However, when we convert Ramachandran plots to Ramachandran 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  $\alpha$ -helix region, which is not evident to an untrained eye in (a).



**Figure 8. Ramachandran lines are stackable – Part II.** 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  $\mathcal{R} > 0.5$ ). Panel (c) shows the behavior of individual amino acids when situated *before* each of the 20 amino acids. This graph shows a major benefit of side-by-side Ramachandran line “stacking”: general trends become much more obvious. For example, it is evident that glycines and prolines dramatically modify the structure of an amino acid preceding it (compared to average behavior of amino acids in Fig. 7b). This trend is not as strong when considering amino acids that *follow* glycines or prolines (c). Such trends, while previously discovered [e.g., Gunasekaran *et al.* (1998); Ho and Brasseur (2005)], would not be accessible when naively considering Ramachandran plots because one would require 400 ( $20 \times 20$ ) distinct Ramachandran plots to compare.

147 protein backbones transition from one state to another.

## 148 USING THE BACKMAP PYTHON MODULE

### 149 Installation

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

### 153 Usage

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

#### 156 In-script usage I: first simple test

157 The simplest test would be to generate Ramachandran numbers from  $(\phi, \psi)$  pairs:

```
158 # Import module
159 import backmap
160 # Convert (phi, psi) to R
161 print backmap.R(phi=0,psi=0) # Expected output: 0.5
162 print backmap.R(-180,-180) # Expected output: 0.0
163 print backmap.R(180,180) # Expected output: 1.0 (equivalent in meaning to 0)
```

1  
2  
3  
4  
5  
6

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

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

##### 168 1. Select and read a protein PDB structure

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

```
171 import backmap
172 pdbfn = './pdbs/nanosheet_birth_U7.pdb' # Set pdb name
173 data = backmap.read_pdb(pdbfn) # READ PDB in the form of a matrix with columns
```

1  
2  
3

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

##### 178 2. Select color scheme (color map)

179 In addition to custom colormaps listed in the next section, one can also use standardly available  
180 colormaps at [matplotlib.org](#) (e.g., ‘Reds’ or ‘Reds\_r’).

```
181 # setting the name of the colormap
182 cmap = "SecondaryStructure"
```

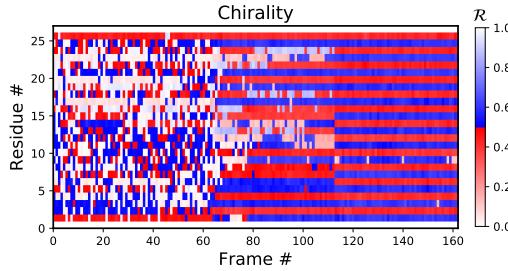
4  
5

##### 185 3. Draw per-chain MAPs

```
186 # Grouping by chain
187 grouped_data = backmap.group_by(data, group_by='chain',
188                                 columns_to_return=['model','resid','R'])
189 for chain in grouped_data.keys(): # Going through each chain
190     # Getting the X,Y,Z values for each entry
191     models, residues, Rs = grouped_data[chain]
192     # Finally, creating (but not showing) the graph
193     backmap.draw_xyz(X = models, Y = residues, Z = Rs
194                      , xlabel = 'Frame #', ylabel = "Residue #", zlabel = '$\mathcal{R}$'
195                      ,cmap = cmap, title = "Chain: "+chain+""
196                      ,vmin=0,vmax=1)
197     # Now, we display the graph:
198     plt.show() # ... one can also use plt.savefig() to save to file
```

6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18

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



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

### 207 In-script usage III: Creating custom graphs

208 Other types of grpahs can be easily created by modifying part three of the code above. For example, the  
 209 following code creates histograms of R, one for each model (starting from line 9 above).

---

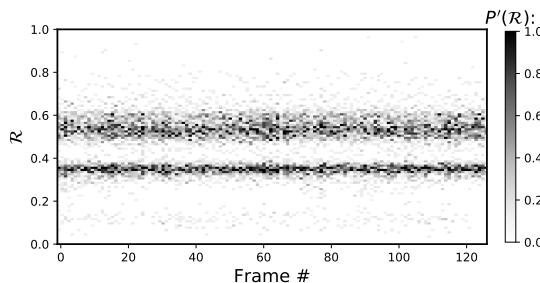
```

210 for chain in grouped_data.keys():
211     models, residues, Rs = grouped_data[chain]
212
213     'Begin custom code'
214     X = []; Y = []; Z =[]; # Will set X=model, Y=R, Z=P(R)
215     # Bundling the three lists into one 2d array
216     new_data = np.array(zip(models, residues, Rs))
217     # Getting all R values, model by model
218     for m in sorted(set(new_data[:,0])): # column 0 is the model column
219         # Getting all Rs for that model #
220         current_rs = new_data[np.where(new_data[:,0]==m)][:,2] # column 2 contains R
221         # Getting the histogram
222         a,b = np.histogram(current_rs, bins=np.arange(0,1.01,0.01))
223         max_count = float(np.max(a))
224         for i in range(len(a)):
225             X.append(m); Y.append((b[i]+b[i+1])/2.0); Z.append(a[i]/float(np.sum(a)));
226     'End custom code'
227
228     # Finally, creating (but not showing) the graph
229     draw_xyz(X = X, Y = Y, Z = Z
230             , xlabel = 'Frame #', ylabel = "$\mathcal{R}$", zlabel = "$P(\mathcal{R})$"
231             ,cmap = 'Greys', ylim=[0,1])
232     plt.yticks(np.arange(0,1.00001,0.2))
233     # Now, we display the graph:
234     plt.show() # ... one can also use plt.savefig() to save to file

```

---

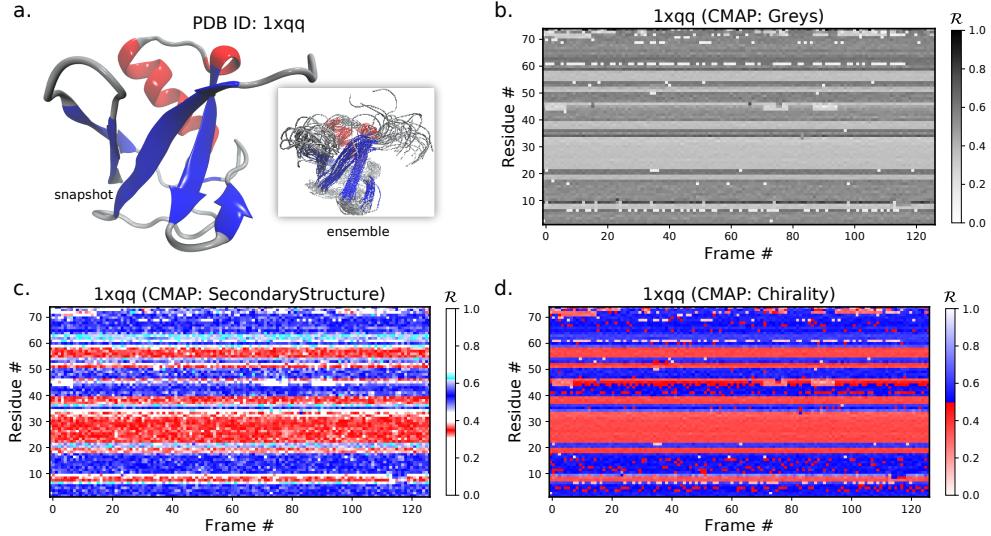
237 The code above results in the following graph:



### 239 In-script usage IV: Available color schemes (CMAPs)

240 Aside from the general color maps (cmaps) that exist in matplotlib (e.g., ‘Greys’, ‘Reds’, or, god forbid,  
 241 ‘jet’), BACKMAP provides two new colormaps: ‘Chirality’ (key: +twists – red; -ve twists: blue),  
 242 and ‘SecondaryStructure’ (key: potential helices – red; sheets – blue; ppII helices – cyan). right  
 243 twisting backbones are shown in red; left twisting backbones are shown in blue). Fig. 9 shows how

244 a single protein ensemble may be described using these schematics. As illustrated in Fig. 9b, cmaps  
 245 available within the standard matplotlib package do not distinguish between major secondary structures  
 246 well, while those provided by BACKMAP do. In case it is known that the protein backbone accesses  
 247 non-traditional regions of the Ramachandran plot, a four-color schematic will be needed (see below for  
 248 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),  $\beta$ -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.

### 249 Stand Alone Usage

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

### 252 Stand Alone Example I: A Stable Protein

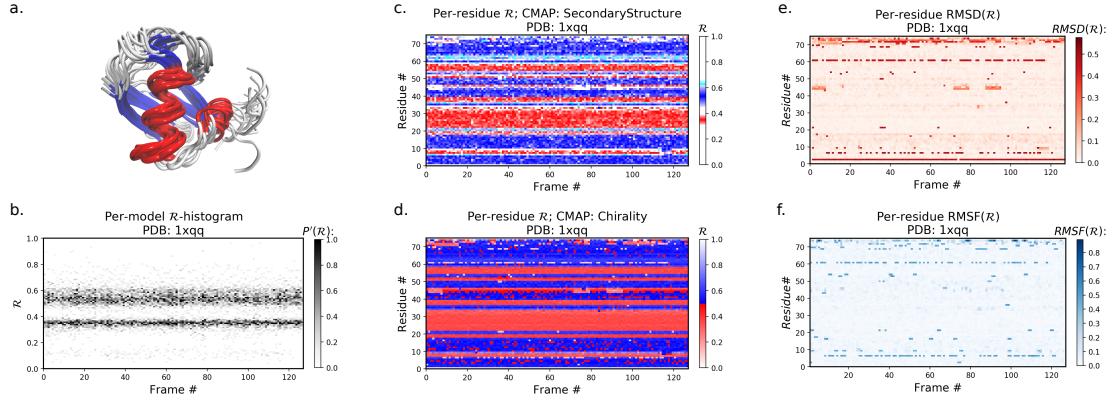
253 Panels (b) through (f) of Fig. 10 below were created by running '`> python -m backmap ./tests/pdfs/1xqq.pdb`'  
 254 (Panel (b) was created using VMD). These graphs indicate that protein 1xqq describes a conformationally  
 255 stable protein, since each residue fluctuates little in color (structure) over 'time' (c,d; here and below, it is  
 256 assumed that discrete models represent distinct states of the protein over 'time'), show little change in the  
 257  $R$  histogram over time (b) and show few enduring fluctuations in RMSD (e) and RMSF (f).

258 In particular, each column in Panel (b) describes the histogram in Ramachandran number ( $R$ ) space  
 259 for a single model/timeframe. These histograms show the presence of both  $\alpha$ -helices (at  $R \approx 0.34$ )  
 260 and  $\beta$ -sheets (at  $R \approx 0.52$ ). Additionally, Panels (c) and (d) describe per-residue conformational plots  
 261 (colored by two different metrics or CMAPS), which show that most of the protein backbone remains  
 262 relatively stable over time (e.g., few fluctuations in state or 'color' are evident over frame #). Finally,  
 263 Panel (e) describes the extent towards which a single residue's state has deviated from the first frame,  
 264 and Panel (f) describes the extent towards which a single residue's state has deviated from its state in the  
 265 previous frame. All these graphs, show that this protein is relatively conformationally stable.

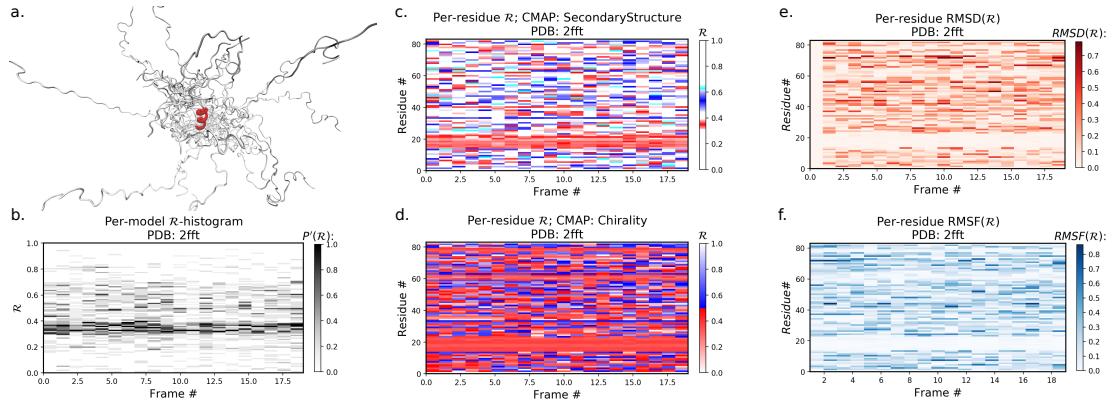
### 266 Stand Alone Example II: An Intrinsically Disordered Protein

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

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



**Figure 10.** Protein 1xqq describes a stable protein.



**Figure 11.** Protein 2fft describes an intrinsically disordered protein, with one stable helix in red.

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

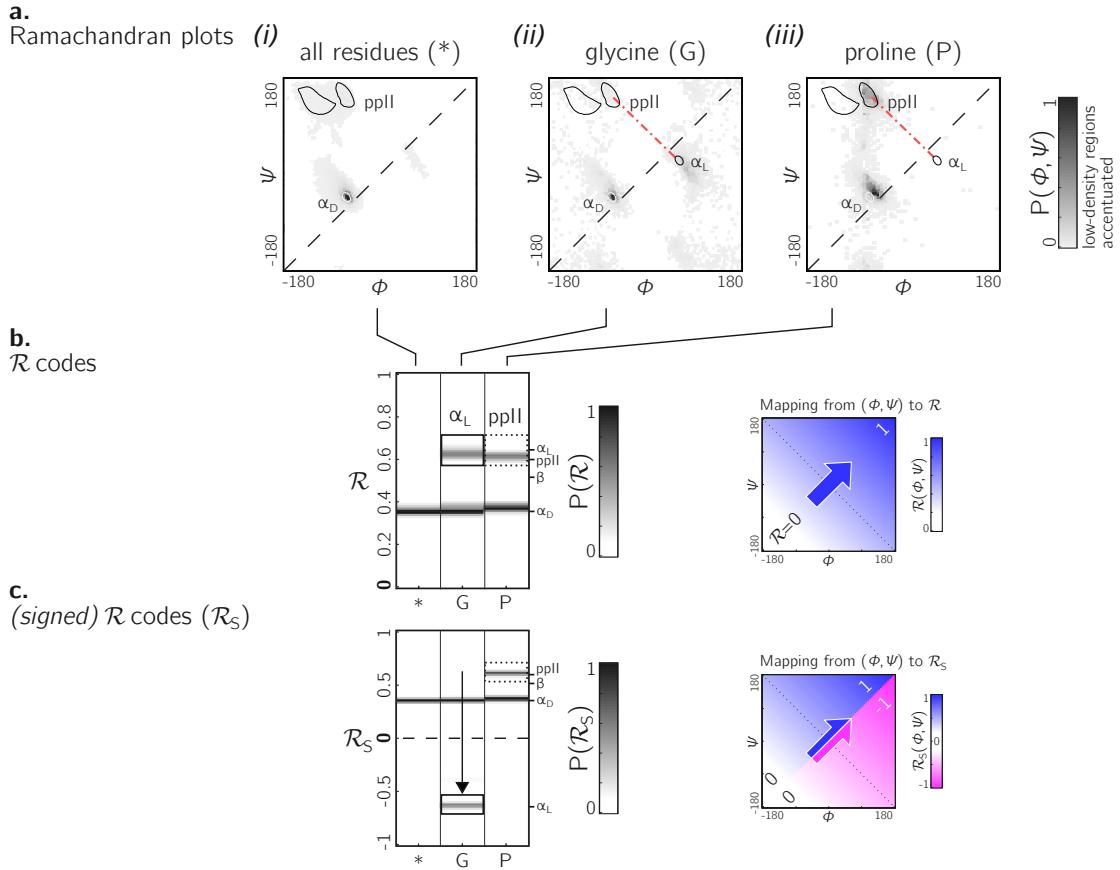
275     Yet, interestingly, Panels (c) through (f) also show an unusually stable region – residues 15 through  
276     25 – which consistently display the same conformational ( $\alpha$ -helical) state at  $R \approx 0.34$  (interpreted as the  
277     color red in Panel (c)). This trend would be hard to recognize by simply looking at the structural ensemble  
278     (Panel (a)).

#### 279     A signed Ramachandran number for ‘misbehaving’ backbones

280     The Ramachandran number increases in value from the bottom left of the Ramachandran plot to the  
281     top right in sweeps that are parallel to the negative sloping diagonal. As discussed in Mannige *et al.*  
282     (2016), this method of mapping a two-dimensional space into one number is still structurally meaningful  
283     and descriptive because 1) most structural features of the protein backbone – e.g. radius of gyration  
284     (Mannige *et al.*, 2016), end-to-end distance (Mannige *et al.*, 2016), and chirality (Mannige, 2017) – vary  
285     little along lines parallel to the negatively-sloping diagonal (this is indicated by relatively small standard  
286     deviations in structural metrics for similar  $R$ s; Fig. 4), and 2) most protein backbones display chiral  
287     centers and therefore predominantly appear on the top left region of the Ramachandran plot (above the  
288     dashed diagonal in Fig. 12a-(i)).

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

295     To accomodate the situation where achiral backbones are expected (eg., if peptoids or polygycines



**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  $\alpha_D$  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 available by the separation of polyproline-II and  $\alpha_D$  helices (c). The mapping of  $(\phi, \psi)$  to  $\mathcal{R}$  and  $\mathcal{R}_S$  are shown to the right of each respective  $\mathcal{R}$ -plot (b,c).

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.,

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

As an example of the utility of  $\mathcal{R}_S$ , Fig. 12b shows that  $\mathcal{R}_S$  easily distinguishes  $\alpha_D$  from ppII.

Note that the signed  $\mathcal{R}_S$ , while useful, would be important in very limited scenarios, as more than 96% of the amino acids in the Protein Databank (PDB) occupy the upper-left region of the Ramachandran plot (with the 3% of ‘rule breakers’ contributed mostly by glycines).

## CONCLUSION

A simpler Ramachandran number is reported –  $\mathcal{R} = (\phi + \psi + 2\pi)/(4\pi)$  – which, while being a single number, provides much information. For example, as discussed in Mannige *et al.* (2016),  $\mathcal{R}$  values above 0.5 are left-handed, while those below 0.5 are right handed,  $\mathcal{R}$  values close to 0, 0.5 and 1 are extended,  $\beta$ -sheets occupy  $\mathcal{R}$  values at around 0.52, right-handed  $\alpha$ -helices hover around 0.34.

304 Given the Ramachandran number's 'stackability', single graphs can hold detailed information of the  
 305 progression/evolution of molecular trajectories. Indeed, Fig. 8 shows how 400 distinct Ramachandran  
 306 plots can easily be fit into one graph when using  $\mathcal{R}$ . Finally, a python script/module (BACKMAP) has  
 307 been provided in an online [GitHub repository](#) to promote the utility of  $\mathcal{R}$  as a universal metric.

## 308 MATERIALS

309 The following protein structures were obtained from the Protein DataBank (PDB): [1mba](#), [2acy](#), [1xqq](#), and  
 310 [2fft](#). The first two in the list ([1mba](#), [2acy](#)) describe single conformations and the last two ([1xqq](#), [2fft](#))  
 311 describe ensembles.

312 Statistics about single amino acid conformations and secondary structures (excepting polyproline II  
 313 helices) were derived from the Structural Classification of Proteins or SCOPe website [Release 2.06; [Fox](#)  
 314 *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  
 315 conformations (one domain per conformation) with lower than 40% sequence identity. Secondary structure  
 316 annotations were assigned using the DSSP algorithm (Kabsch and Sander, 1983), although the STRIDE  
 317 algorithm (Frishman and Argos, 1995) provides qualitatively identical distributions.

318 Given the absence of polyproline II helix (ppII) annotation in the present version of DSSP, statistics  
 319 for polyproline II helices (used to generate the green distributions in Figs X) were obtained from segments  
 320 within 16,535 proteins annotated by PolyprOnline ([Chebrek et al.](#), 2014) to contain three or more residues  
 321 of the secondary structure.

322 Fig X represents a trajectory of a portion of a single peptoid backbone within a 'relaxing' peptoid  
 323 nanosheet bilayer. The conformation of this backbone – derived from work by [Mannige et al.](#) (2015) and  
 324 [Mannige et al.](#) (2016) – is also available as '[/tests/pdbs/nanosheet\\_birth\\_U7.pdb](#)' within the companion  
 325 [GitHub repository](#).

326 Root mean squared deviation (RMSD) and fluctuation (RMSF) are measures of change in structure  
 over 'time' when respectively compared to the initial conformation or the preceding conformation. Their  
 equations are as follows:

$$\text{RMSD}_{r,t} = \sqrt{(\mathcal{R}_{r,t} - \mathcal{R}_{r,1})^2}, \quad \text{RMSF}_{r,t} = \sqrt{(\mathcal{R}_{r,t} - \mathcal{R}_{r,t-1})^2}. \quad (4)$$

327 Here,  $\mathcal{R}_{r,t}$  is the Ramachandra nnumber associated with residue number  $r$  at 'time'  $t$ . Since we are  
 328 only considering deviation and fluctuations within individual resides, these numbers are normalized by  
 329 dividing by 1.

## 330 ACKNOWLEDGMENTS

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

## 337 APPENDIX

### 338 Simplifying the Ramachandran number ( $\mathcal{R}$ )

339 This section will derive the simplified Ramachandran number presented in this paper from the more  
 340 complicated looking Ramachandran number introduced previously ([Mannige et al.](#), 2016).

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

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

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

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

341 Here,  $\lfloor x \rfloor$  rounds  $x$  to the closest integer value,  $\sigma$  is a scaling factor, discussed below, and  $\lambda$  is the  
 342 range of an angle in degrees (i.e.,  $\lambda = \phi_{\max} - \phi_{\min}$ ). Effectively, this equation does the following. 1) It  
 343 divides up the Ramachandran plot into  $(360^\circ \sigma^{1/\sigma})^2$  squares, where  $\sigma$  is a user-selected scaling factor  
 344 that is measured in reciprocal degrees [see Fig. 8b in [Mannige et al. \(2016\)](#)]. 2) It then assigns integer  
 345 values to each square by setting the lowest integer value to the bottom left of the Ramachandran plot  
 346 ( $\phi = -180^\circ, \psi = -180^\circ$ ) and proceeding from the bottom left to the top right by iteratively slicing down  
 347 -1/2 sloped lines and assigning increasing integer values to each square that one encounters. 3) Finally,  
 348 the equation assigns any  $(\phi, \psi)$  pair within  $\phi, \psi \in [-\phi_{\min}, \phi_{\max}]$  to the integer value ( $R_Z$ ) assigned to the  
 349 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:

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

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  $\sigma$  tends towards  $\infty$ . In particular, when  $\sigma \rightarrow \infty$ , Eqn. 7 becomes

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

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

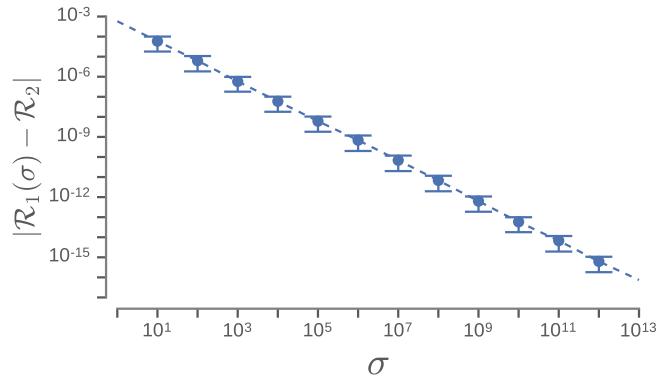
$$\mathcal{R}(\phi, \psi) = \frac{\phi + \psi + 2\pi}{4\pi}. \quad (9)$$

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

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

However, in changing the ranges, the meaning of the Ramachandran number will change. This manuscript assumes that all angles  $(\phi, \psi, \omega)$  range between  $-\pi$  ( $-180^\circ$ ) and  $\pi$  ( $180^\circ$ )



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

## 357 REFERENCES

- 358 **Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P.** 2002. Molecular biology of the cell.  
359 new york: Garland science; 2002. *Classic textbook now in its 5th Edition* .
- 360 **Baruah A, Rani P, Biswas P.** 2015. Conformational entropy of intrinsically disordered proteins from  
361 amino acid triads. *Scientific reports* **5**.
- 362 **Beck DA, Alonso DO, Inoyama D, Daggett V.** 2008. The intrinsic conformational propensities of the  
363 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proceedings of the  
364 National Academy of Sciences* **105**(34):12259–12264.
- 365 **Berg JM, Tymoczko JL, Stryer L.** 2010. *Biochemistry, International Edition*. WH Freeman & Co.,  
366 New York, 7 edition.
- 367 **Chebrel R, Leonard S, de Brevern AG, Gelly JC.** 2014. Polypronline: polyproline helix ii and  
368 secondary structure assignment database. *Database* **2014**:bau102.
- 369 **Dunker A, Babu M, Barbar E, Blackledge M, Bondos S, Dosztányi Z, Dyson H, Forman-Kay J,  
370 Fuxreiter M, Gsponer J, Han KH, Jones D, Longhi S, Metallo S, Nishikawa K, Nussinov R,  
371 Obradovic Z, Pappu R, Rost B, Selenko P, Subramaniam V, Sussman J, Tompa P, Uversky V.  
372 2013. What's in a name? why these proteins are intrinsically disordered? *Intrinsically Disordered  
373 Proteins* **1**:e24157.**
- 374 **Espinoza-Fonseca LM.** 2009. Reconciling binding mechanisms of intrinsically disordered proteins.  
375 *Biochemical and biophysical research communications* **382**(3):479–482.
- 376 **Fink AL.** 2005. Natively unfolded proteins. *Curr Opin Struct Biol* **15**(1):35–41.
- 377 **Fox NK, Brenner SE, Chandonia JM.** 2014. Scope: Structural classification of proteins—extended,  
378 integrating scop and astral data and classification of new structures. *Nucleic Acids Res* **42**(Database  
379 issue):D304–D309. doi:10.1093/nar/gkt1240.
- 380 **Frishman D, Argos P.** 1995. Knowledge-based protein secondary structure assignment. *Proteins:  
381 Structure, Function, and Bioinformatics* **23**(4):566–579.
- 382 **Geist L, Henen MA, Haiderer S, Schwarz TC, Kurzbach D, Zawadzka-Kazimierczuk A, Saxena  
383 S, Źerko S, Koźmiński W, Hinderberger D, et al.** 2013. Protonation-dependent conformational  
384 variability of intrinsically disordered proteins. *Protein Science* **22**(9):1196–1205.
- 385 **Gunasekaran K, Nagarajaram H, Ramakrishnan C, Balaram P.** 1998. Stereochemical punctuation  
386 marks in protein structures: glycine and proline containing helix stop signals. *Journal of molecular  
387 biology* **275**(5):917–932.
- 388 **Ho BK, Brasseur R.** 2005. The ramachandran plots of glycine and pre-proline. *BMC structural biology*  
389 **5**(1):1.
- 390 **Hooft RW, Sander C, Vriend G.** 1997. Objectively judging the quality of a protein structure from a  
391 ramachandran plot. *Computer applications in the biosciences: CABIOS* **13**(4):425–430.
- 392 **Kabsch W, Sander C.** 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-  
393 bonded and geometrical features. *Biopolymers* **22**(12):2577–2637. doi:10.1002/bip.360221211.
- 394 **Kosol S, Contreras-Martos S, Cedeño C, Tompa P.** 2013. Structural characterization of intrinsically  
395 disordered proteins by nmr spectroscopy. *Molecules* **18**(9):10802–10828.
- 396 **Laskowski RA.** 2003. Structural quality assurance. *Structural Bioinformatics, Volume 44* pages 273–303.
- 397 **Laskowski RA, MacArthur MW, Moss DS, Thornton JM.** 1993. Procheck: a program to check the  
398 stereochemical quality of protein structures. *Journal of applied crystallography* **26**(2):283–291.
- 399 **Mannige RV.** 2014. Dynamic new world: Refining our view of protein structure, function and evolution.  
400 *Proteomes* **2**(1):128–153.
- 401 **Mannige RV.** 2017. An exhaustive survey of regular peptide conformations using a new metric for  
402 backbone handedness (*h*). *PeerJ* **5**:e3327. ISSN 2167-8359. doi:10.7717/peerj.3327.
- 403 **Mannige RV, Haxton TK, Proulx C, Robertson EJ, Battigelli A, Butterfoss GL, Zuckermann RN,  
404 Whitelam S.** 2015. Peptoid nanosheets exhibit a new secondary structure motif. *Nature* **526**:415–420.
- 405 **Mannige RV, Kundu J, Whitelam S.** 2016. The Ramachandran number: an order parameter for protein  
406 geometry. *PLoS One* **11**(8):e0160023.
- 407 **Midic U, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN.** 2009. Protein disorder in the human  
408 diseasesome: unfoldomics of human genetic diseases. *BMC Genomics* **10 Suppl** **1**:S12. doi:10.1186/  
409 1471-2164-10-S1-S12.
- 410 **Momen R, Azizi A, Wang L, Yang P, Xu T, Kirk SR, Li W, Manzhos S, Jenkins S.** 2017. The role  
411 of weak interactions in characterizing peptide folding preferences using a qtaim interpretation of the

- 412 ramachandran plot ( $\phi$ - $\psi$ ). *International Journal of Quantum Chemistry* .
- 413 **Ramachandran G, Ramakrishnan C, Sasisekharan V.** 1963. Stereochemistry of polypeptide chain  
414 configurations. *Journal of molecular biology* **7**(1):95–99.
- 415 **Sibille N, Bernado P.** 2012. Structural characterization of intrinsically disordered proteins by the  
416 combined use of nmr and saxs. *Biochemical society transactions* **40**(5):955–962.
- 417 **Subramanian E.** 2001. Gn ramachandran. *Nature Structural & Molecular Biology* **8**(6):489–491.
- 418 **Tien MZ, Sydykova DK, Meyer AG, Wilke CO.** 2013. Peptidebuilder: A simple python library to  
419 generate model peptides. *PeerJ* **1**:e80.
- 420 **Tompa P.** 2011. Unstructural biology coming of age. *Curr Opin Struct Biol* **21**(3):419–425. doi:  
421 10.1016/j.sbi.2011.03.012.
- 422 **Uversky VN.** 2003. Protein folding revisited. a polypeptide chain at the folding-misfolding-nonfolding  
423 cross-roads: which way to go? *Cell Mol Life Sci* **60**(9):1852–1871.
- 424 **Uversky VN, Dunker AK.** 2010. Understanding protein non-folding. *Biochim Biophys Acta*  
425 **1804**(6):1231–1264.