Biomedical Data Science: Mining and Modeling

Globular Protein Structure I

Prof. Corey O'Hern
Department of Mechanical Engineering & Materials Science
Department of Physics
Department of Applied Physics
Program in Computational Biology & Bioinformatics
Integrated Graduate Program in Physical & Engineering Biology
Yale University

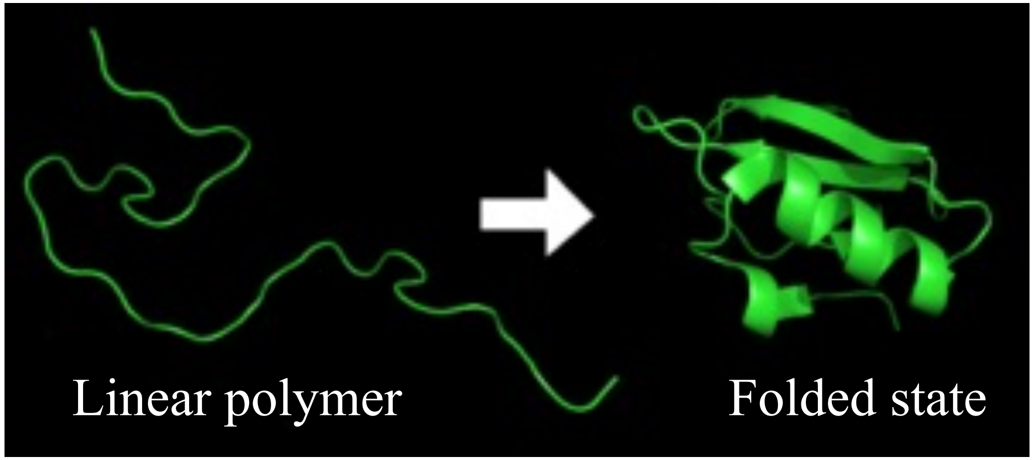
Schedule

Mon./Wed. April 19 and 21: Globular Protein Structure

Thurs./Mon. April 22 and 26: Intrinsically Disordered Proteins

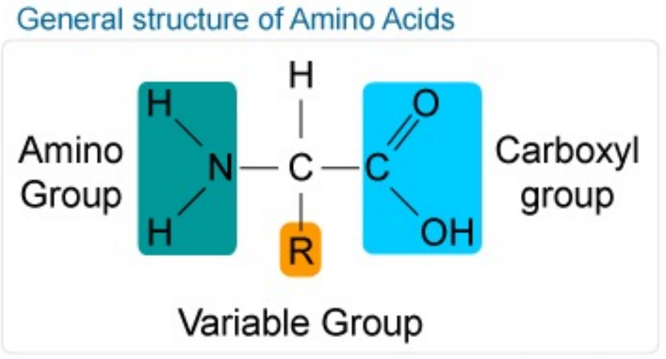
Mon. April 28: Molecular Dynamics Simulations

What are proteins?



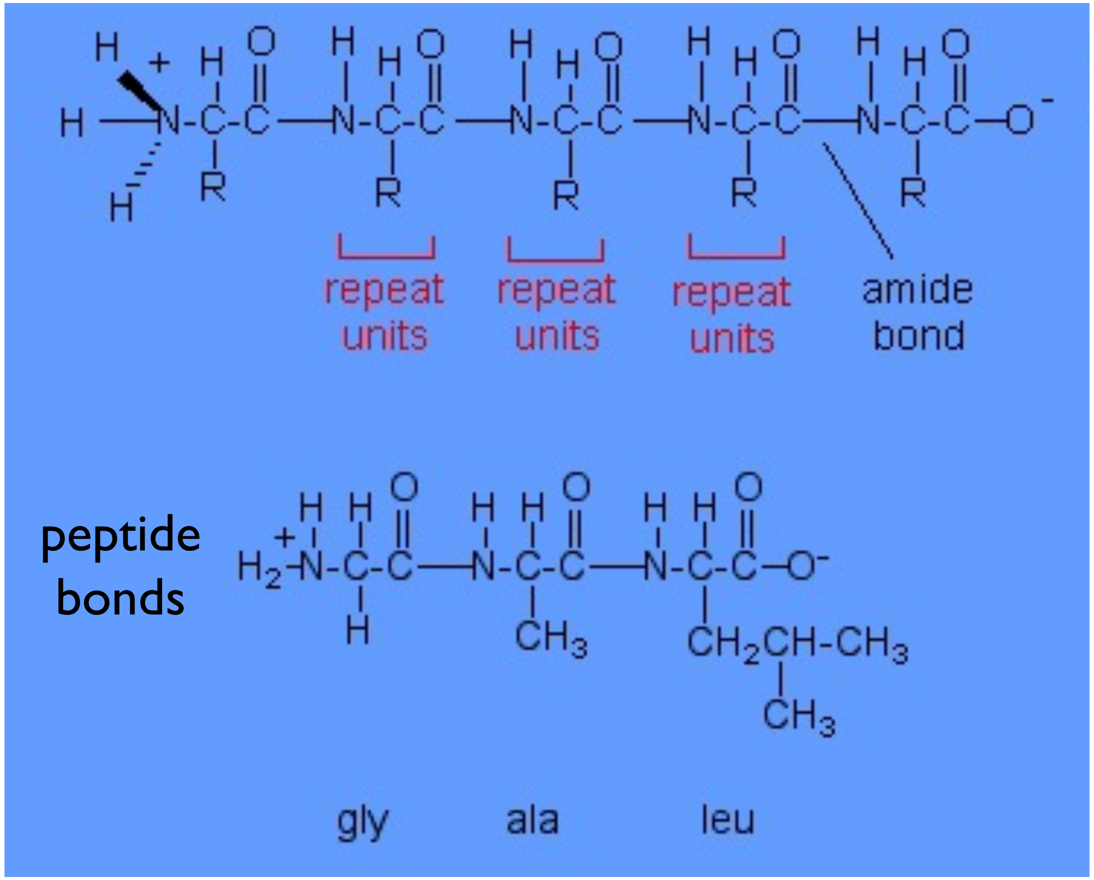
- Proteins are important; e.g. for catalyzing and regulating biochemical reactions, transporting molecules, …
- Linear polymer chain composed of tens (peptides) to thousands (proteins) of monomers
- Monomers are 20 naturally occurring amino acids
- Different proteins have different amino acid sequences
- *Structureless*, extended unfolded state
- Compact, 'unique' native folded state (with secondary and tertiary structure) required for biological function
- Sequence determines protein structure (or lack thereof)
- Proteins unfold or denature with increasing temperature or chemical denaturants

Amino Acids I



N-terminal C_α C-terminal

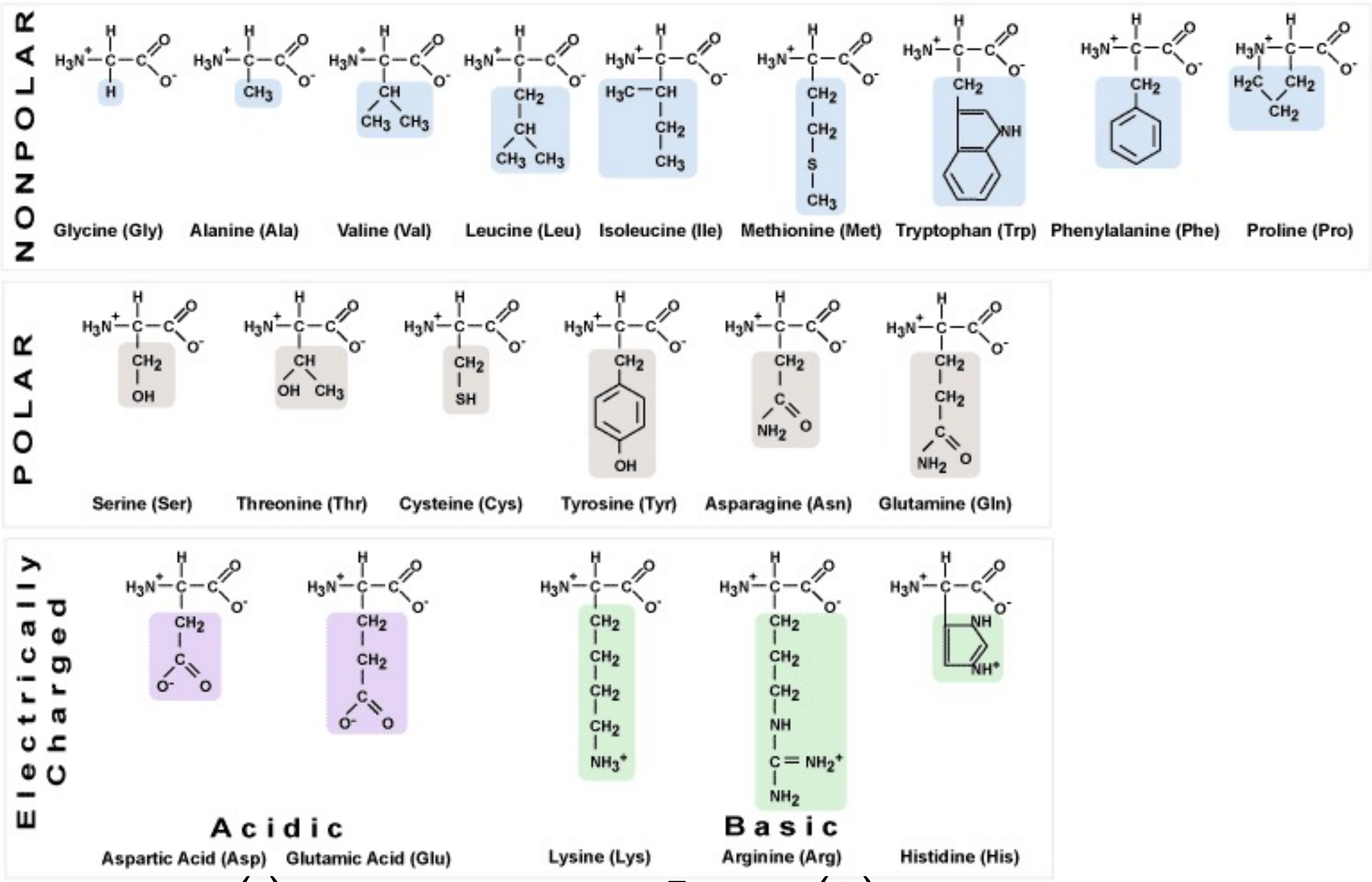
R
variable
side chain



peptide
bonds

- Side chains differentiate amino acid repeat units
- Peptide bonds link residues into polypeptides

Amino Acids II

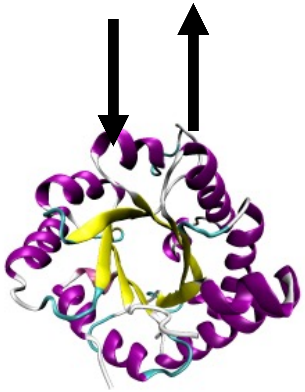


(-) (+)

The Protein Folding Problem:

What is 'unique' folded 3D structure of a protein based on its amino acid sequence? Sequence → Structure

Lys–Asn–Val–Arg–Ser–Lys–Val–Gly–Ser–Thr–Glu–Asn–Ile–Lys– His–Gln–Pro– Gly–Gly–Gly–...



Why do proteins fold (correctly & rapidly)??

Levinthal's paradox:

For a protein with N amino acids, number of backbone conformations/minima

$$N_c \sim \mu^{2N}$$

μ = # allowed dihedral angles

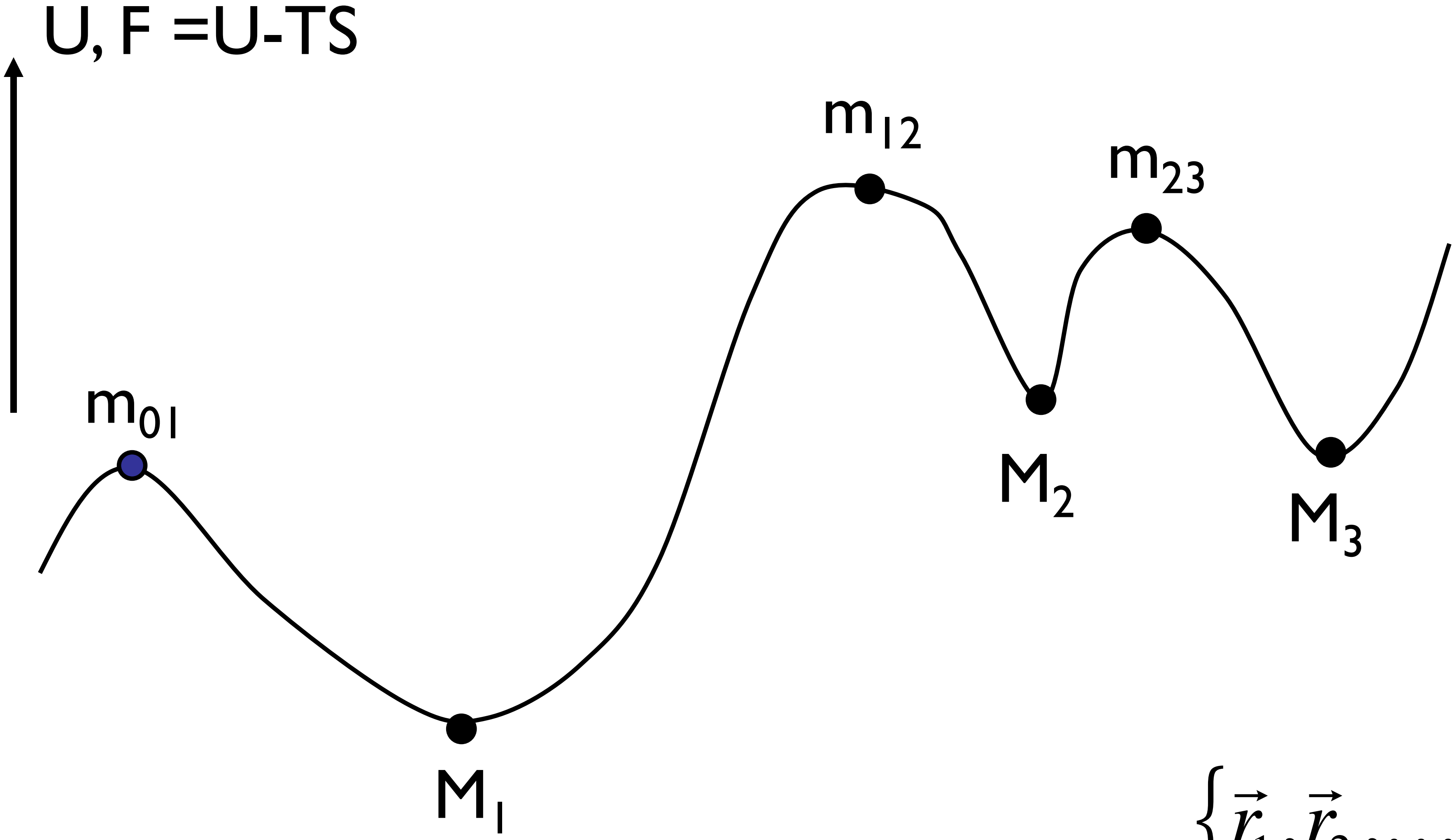
How does a protein find the global optimum w/o global search? Proteins fold much faster.

$$N_c \sim 3^{200} \sim 10^{95}$$

$\tau_{\text{fold}} \sim N_c\, \tau_{\text{sample}} \sim 10^{83}$ s vs $\tau_{\text{fold}} \sim 10^{-6}\text{-}10^{-3}$ s

$\tau_{\text{universe}} \sim 10^{17}$ s

Energy Landscape

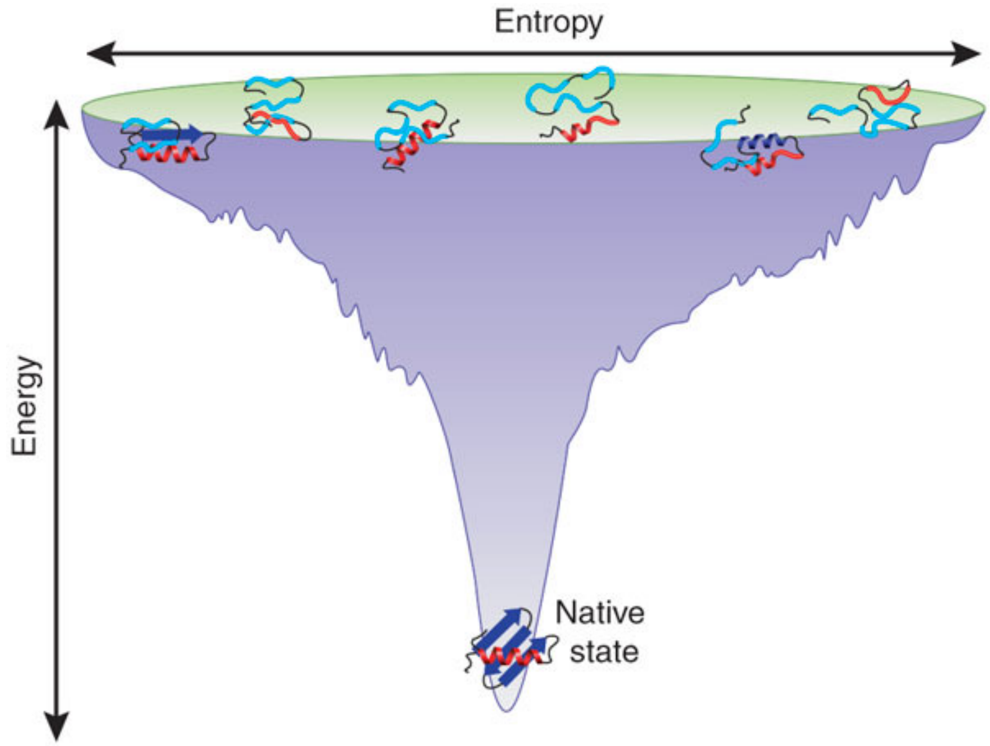


$$\{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\}$$

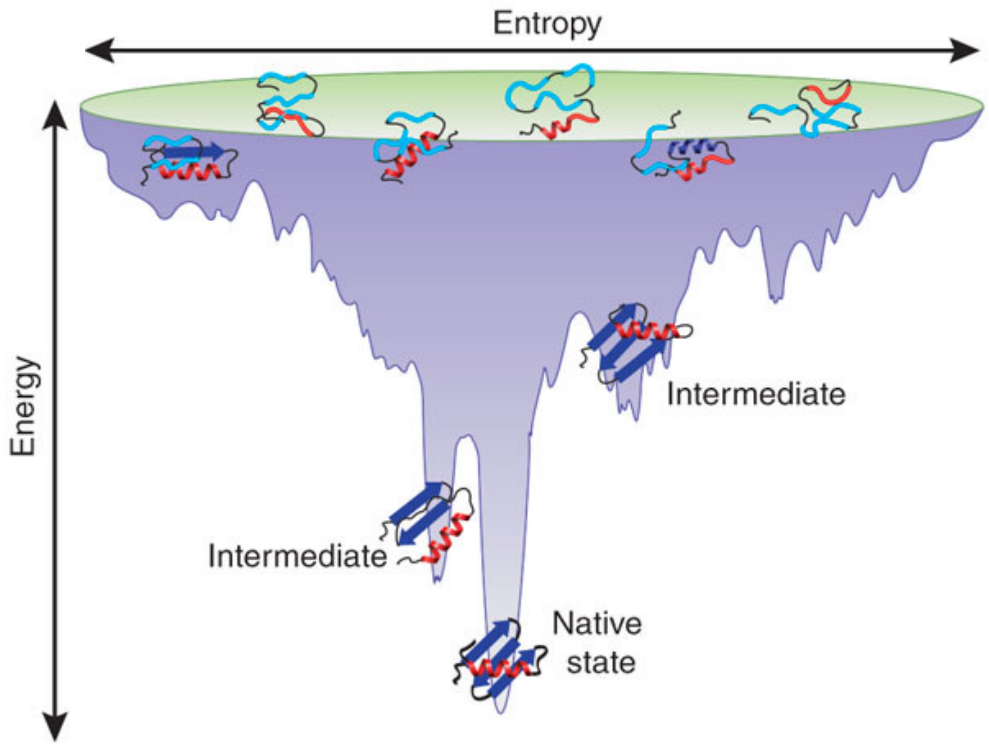
all atomic coordinates; dihedral angles

$$\vec{\nabla} U = 0 \quad \begin{cases} \nabla^2 U > 0 & \text{Minimum (M)} \\ \nabla^2 U = 0 & \text{saddle point} \\ \nabla^2 U < 0 & \text{Maximum (m)} \end{cases}$$

Roughness of Energy Landscape



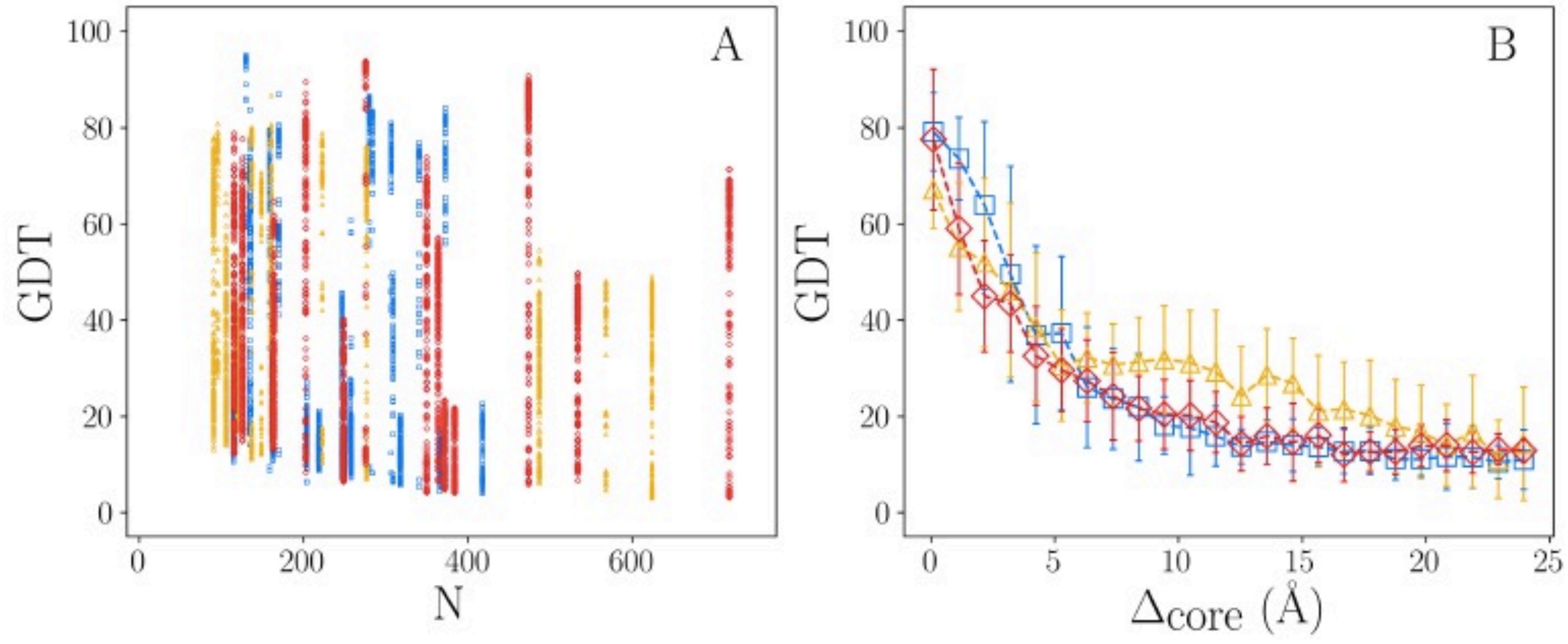
smooth, funneled



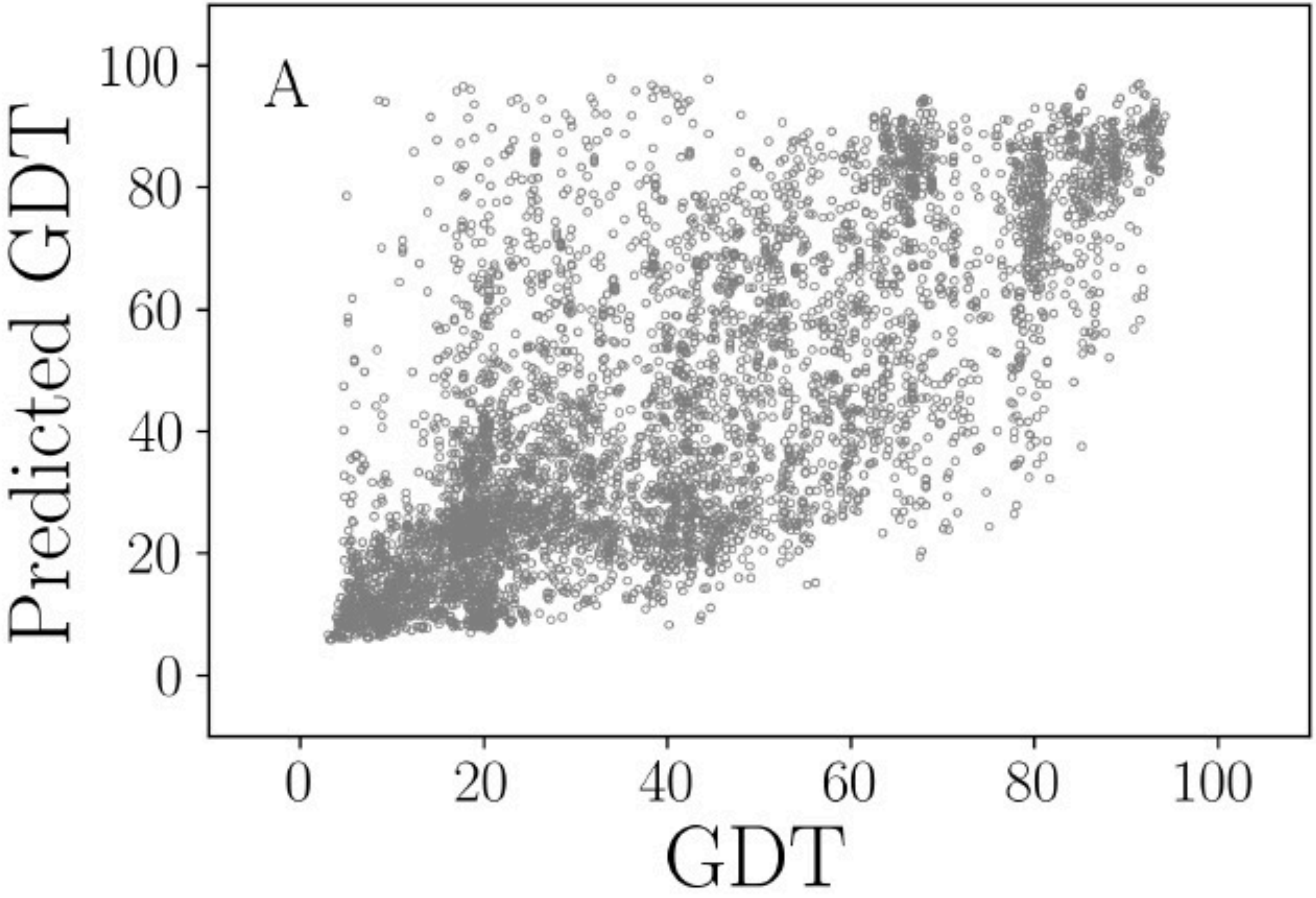
rough

(Wolynes et. al. 1997)

Critical Assessment of Structure Prediction (CASP)



Hoval et. al., *Protein Science* (2018)
Moult et. al., *Protein Science* (2018)


A
Predicted GDT
100
80
60
40
20
0
0
20
40
60
80
100
GDT

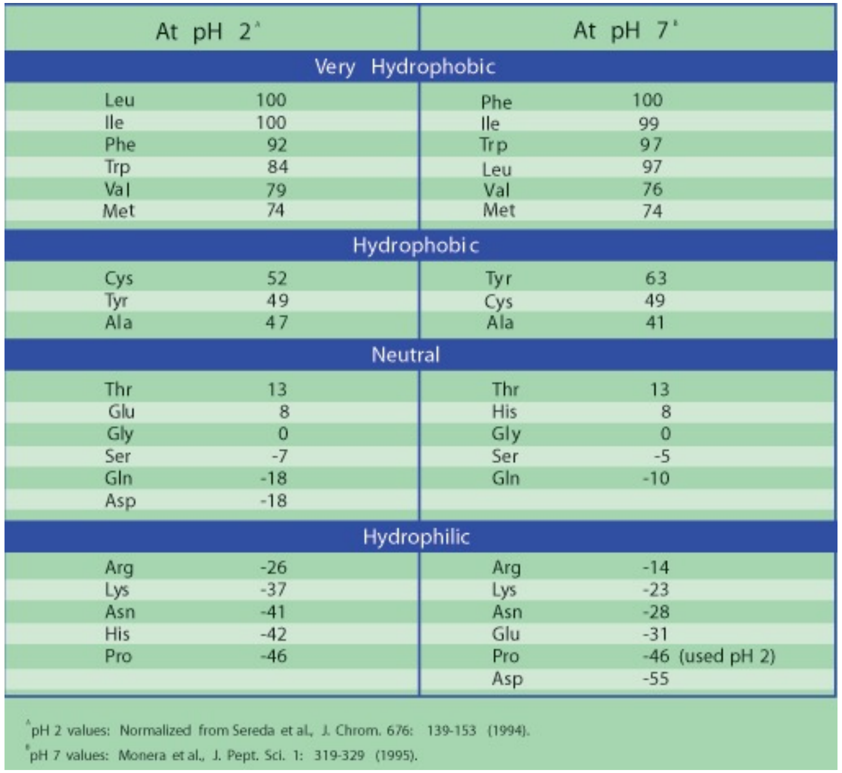
Driving Forces

- Folding: hydrophobicity, hydrogen bonding, van der Waals interactions, …
- Unfolding: increase in conformational entropy, electric charge…

inside H (hydrophobic)

outside P (polar)

Hydrophobicity index



At pH 2[A]		At pH 7[B]	
Very Hydrophobic			
Leu	100	Phe	100
Ile	100	Ile	99
Phe	92	Trp	97
Trp	84	Leu	97
Val	79	Val	76
Met	74	Met	74
Hydrophobic			
Cys	52	Tyr	63
Tyr	49	Cys	49
Ala	47	Ala	41
Neutral			
Thr	13	Thr	13
Glu	8	His	8
Gly	0	Gly	0
Ser	-7	Ser	-5
Gln	-18	Gln	-10
Asp	-18		
Hydrophilic			
Arg	-26	Arg	-14
Lys	-37	Lys	-23
Asn	-41	Asn	-28
His	-42	Glu	-31
Pro	-46	Pro	-46 (used pH 2)
		Asp	-55

[A] pH 2 values: Normalized from Sereda et al., J. Chrom. 676: 139-153 (1994).

[B] pH 7 values: Monera et al., J. Pept. Sci. 1: 319-329 (1995).

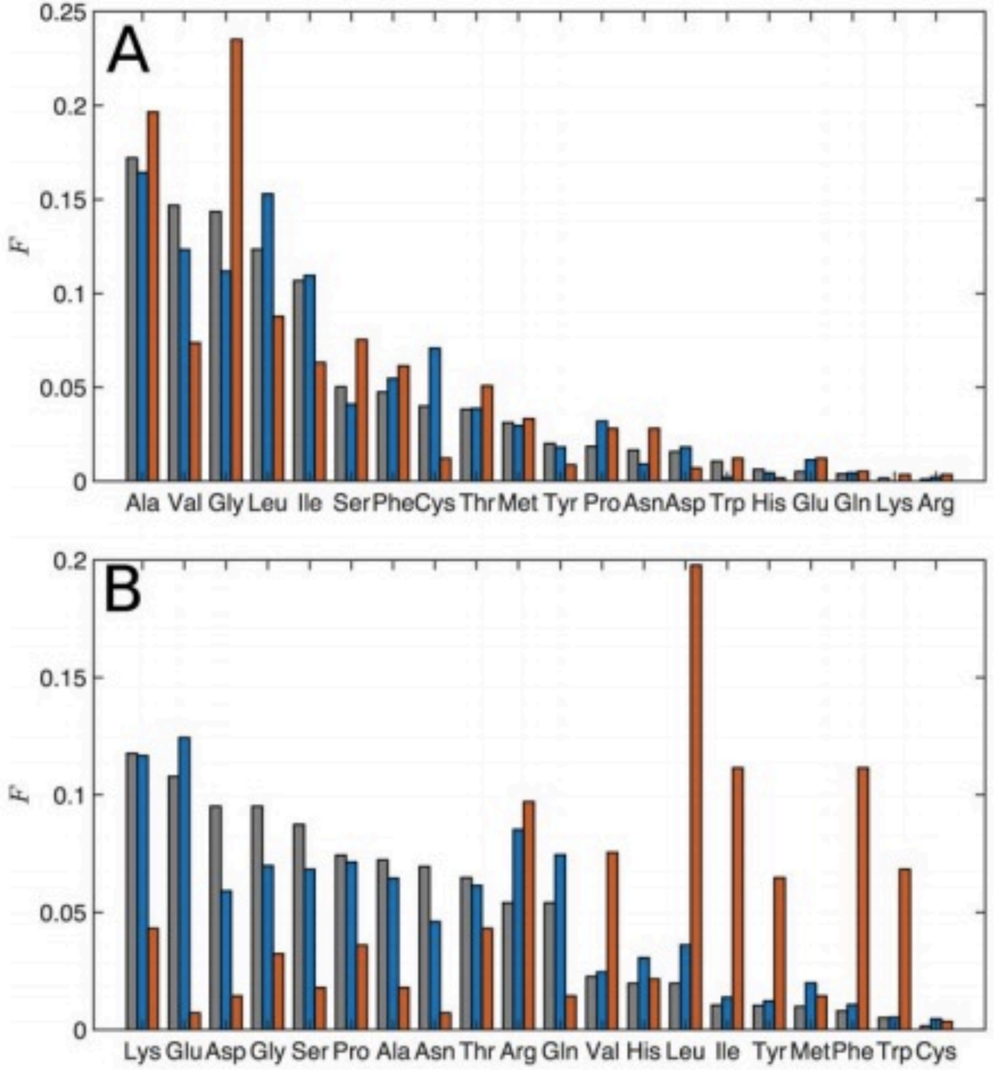
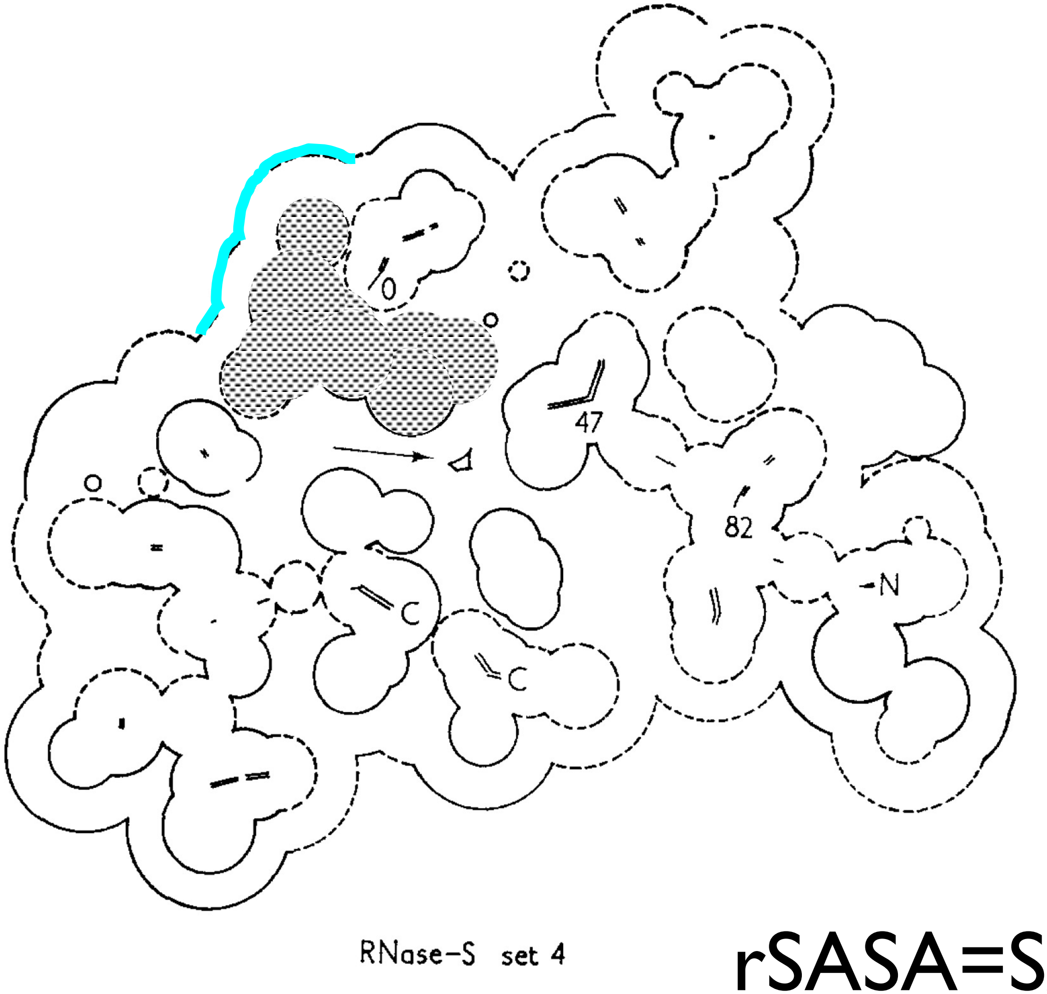
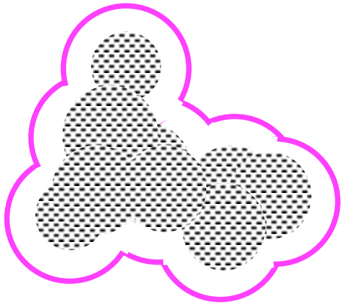


FIGURE 5 Fractions of amino acids with A, rSASA $\leq 10^{-3}$ and B, rSASA>0.5 for residues in the Dun1.0 (grey), PPI (blue), and TM (red) datasets. The fractions are defined relative to the total number of residues in each rSASA category. C, The fractions of core residues (light bars) and non-core residues (rSASA>0.5, dark bars) among the 11 non-charged residues (Ala, Gly, Ile, Leu, Met, Phe, Ser, Thr, Trp, Tyr, and Val) [Color figure can be viewed at wileyonlinelibrary.com]

Solvent Accessible Surface Area and rSASA





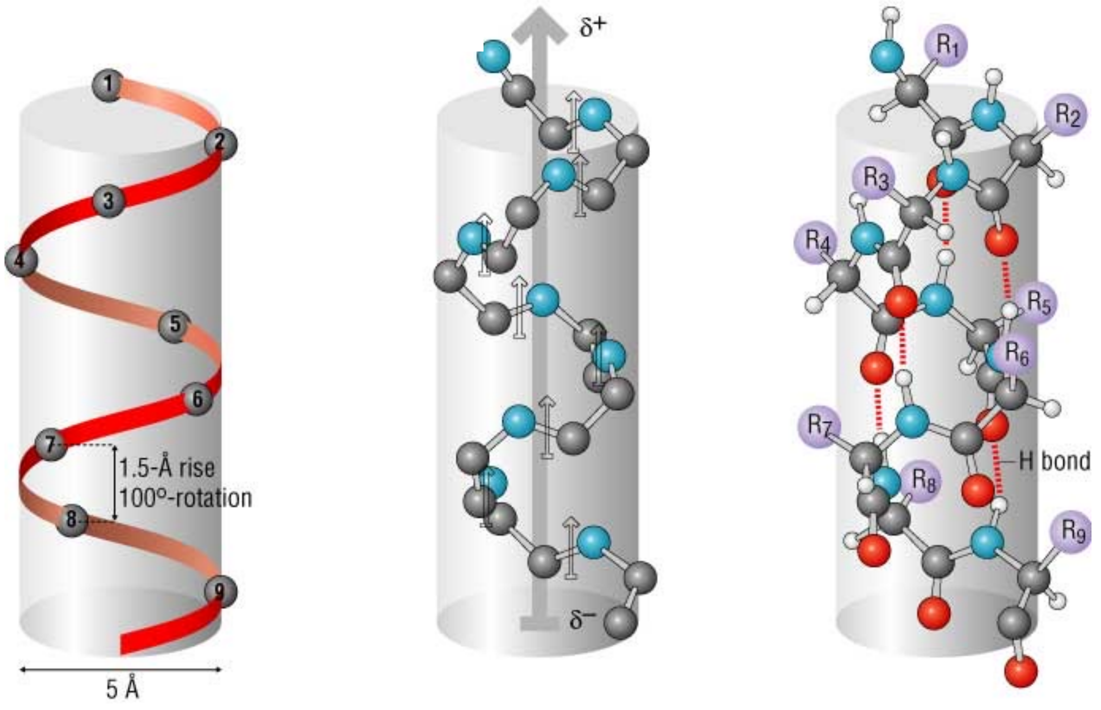
— $SASA_{res}$

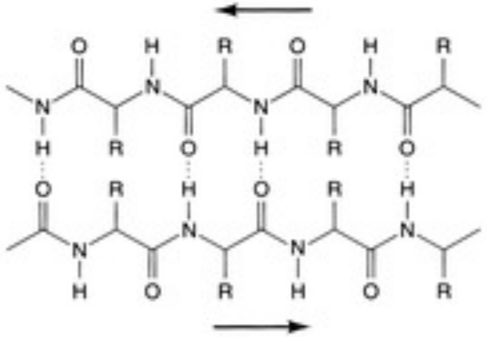
— $SASA_{dip}$

$$rSASA=SASA_{res}/SASA_{dip} = [0,1]$$

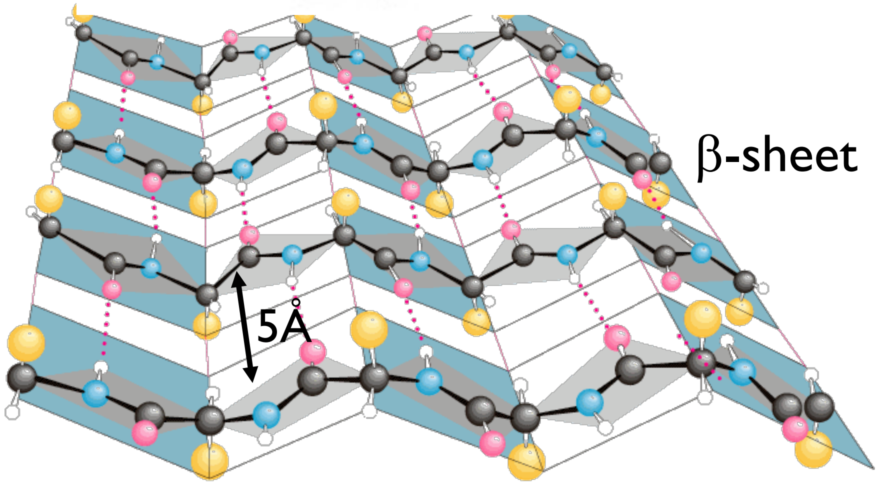
Secondary Structure: Loops, α-helices, β-strands/sheets

α-helix





β-strand

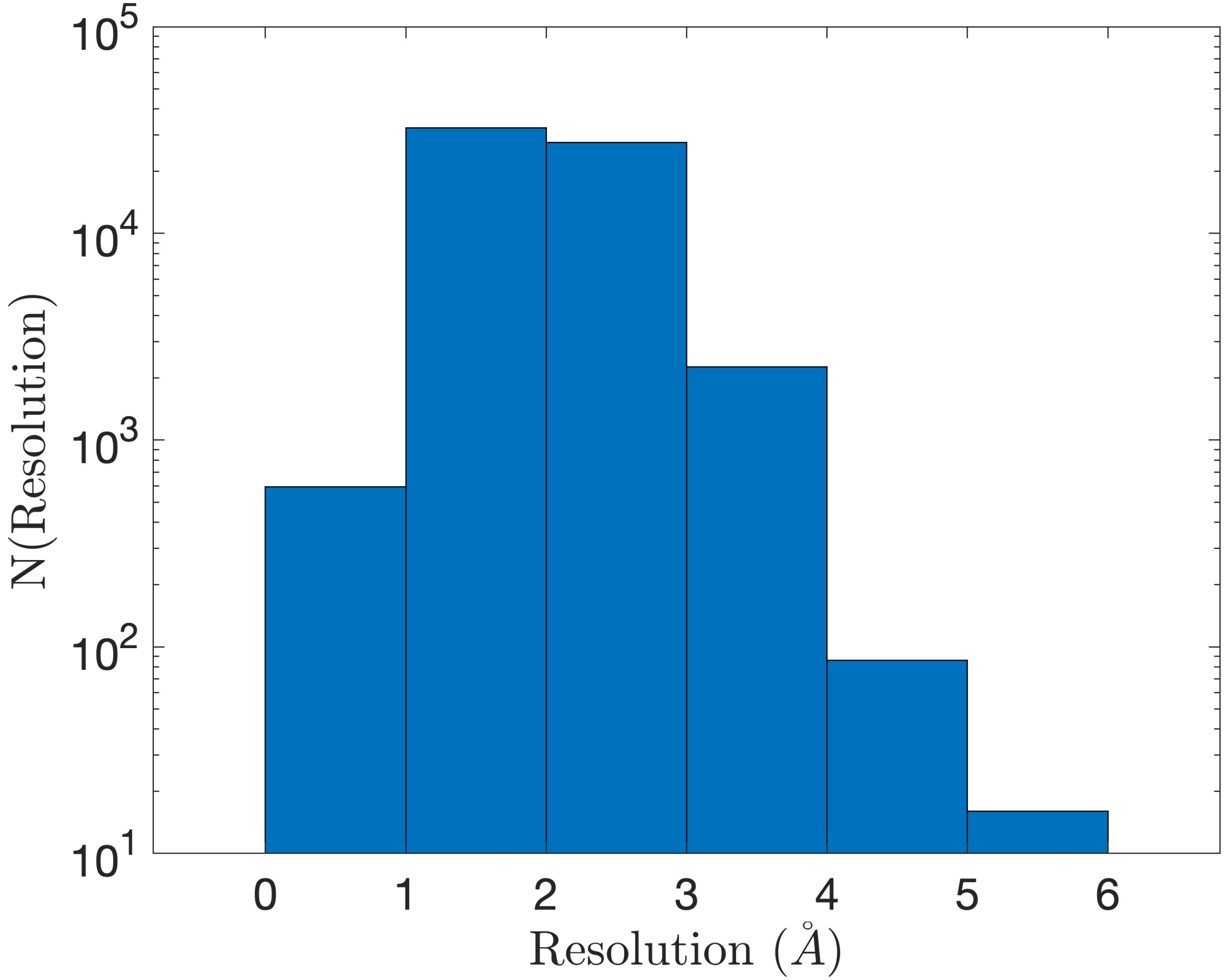


β-sheet

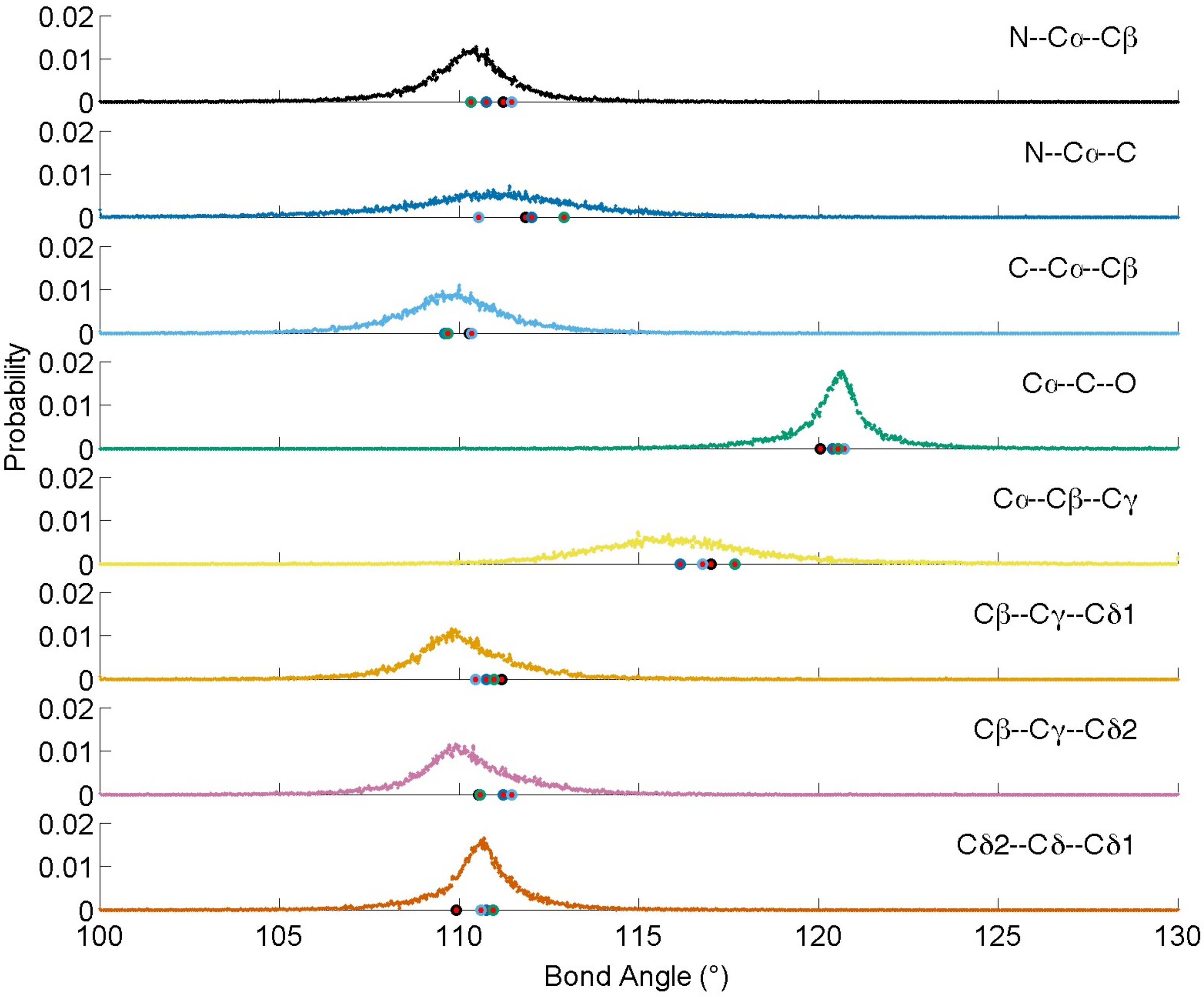
- Right-handed; three turns
- Vertical hydrogen bonds between NH$_2$ (teal/white) backbone group and C=O (grey/red) backbone group four residues earlier in sequence
- Side chains (R) on outside; point upwards toward NH$_2$
- Each amino acid corresponds to 100°, 1.5Å, 3.6 amino acids per turn
- (ϕ,ψ)=(-60°,-45°)
- α-helix propensities: Met, Ala, Leu, Glu

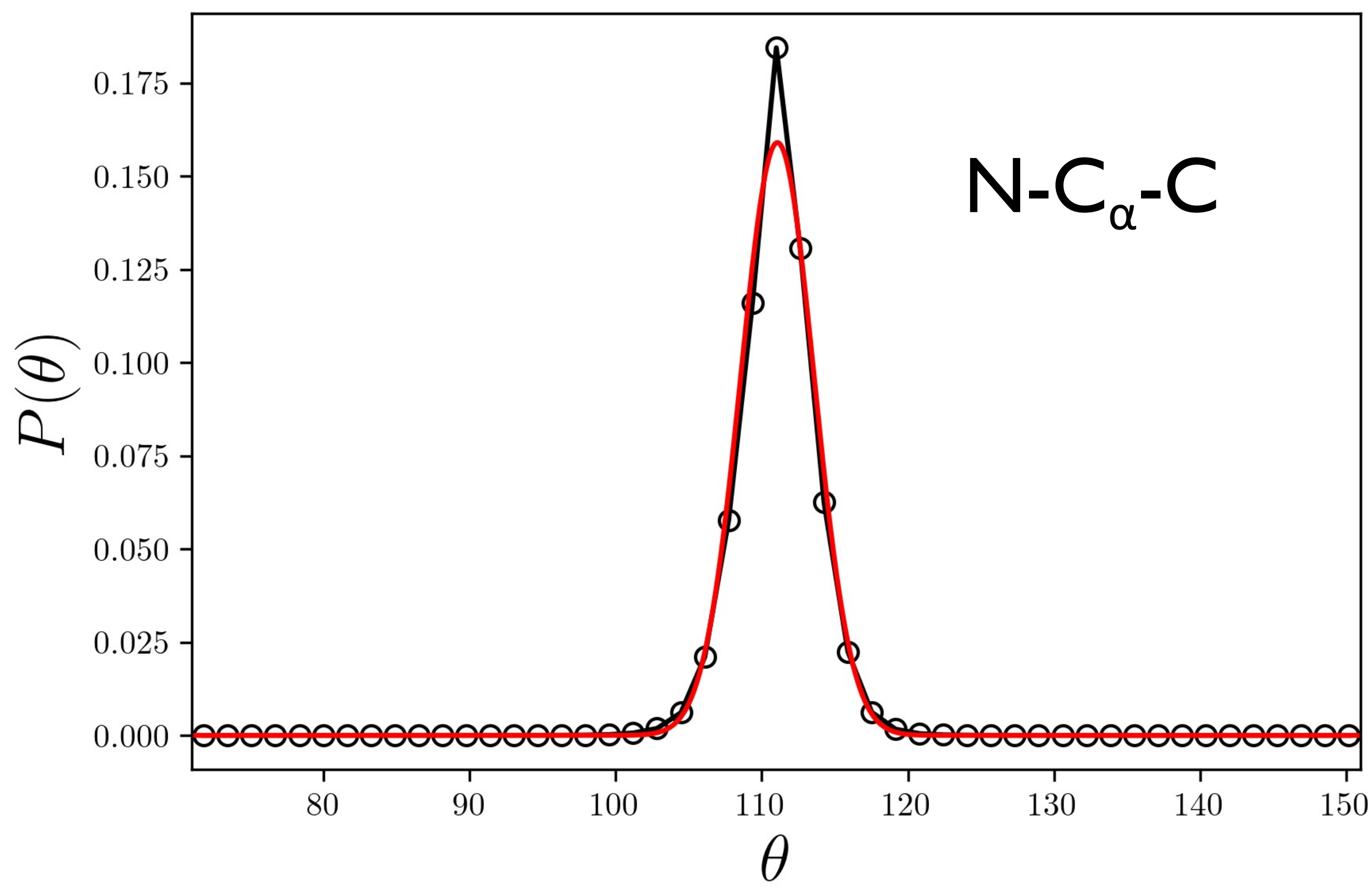
- 5-10 residues; peptide backbones fully extended
- NH (blue/white) of one strand hydrogen-bonded to C=O (black/red) of another strand
- C_α ,side chains (yellow) on adjacent strands aligned; side chains along single strand alternate up and down
- (ϕ,ψ)=(-135°,135°)
- β-strand propensities: Val, Thr, Tyr, Trp, Phe, Ile

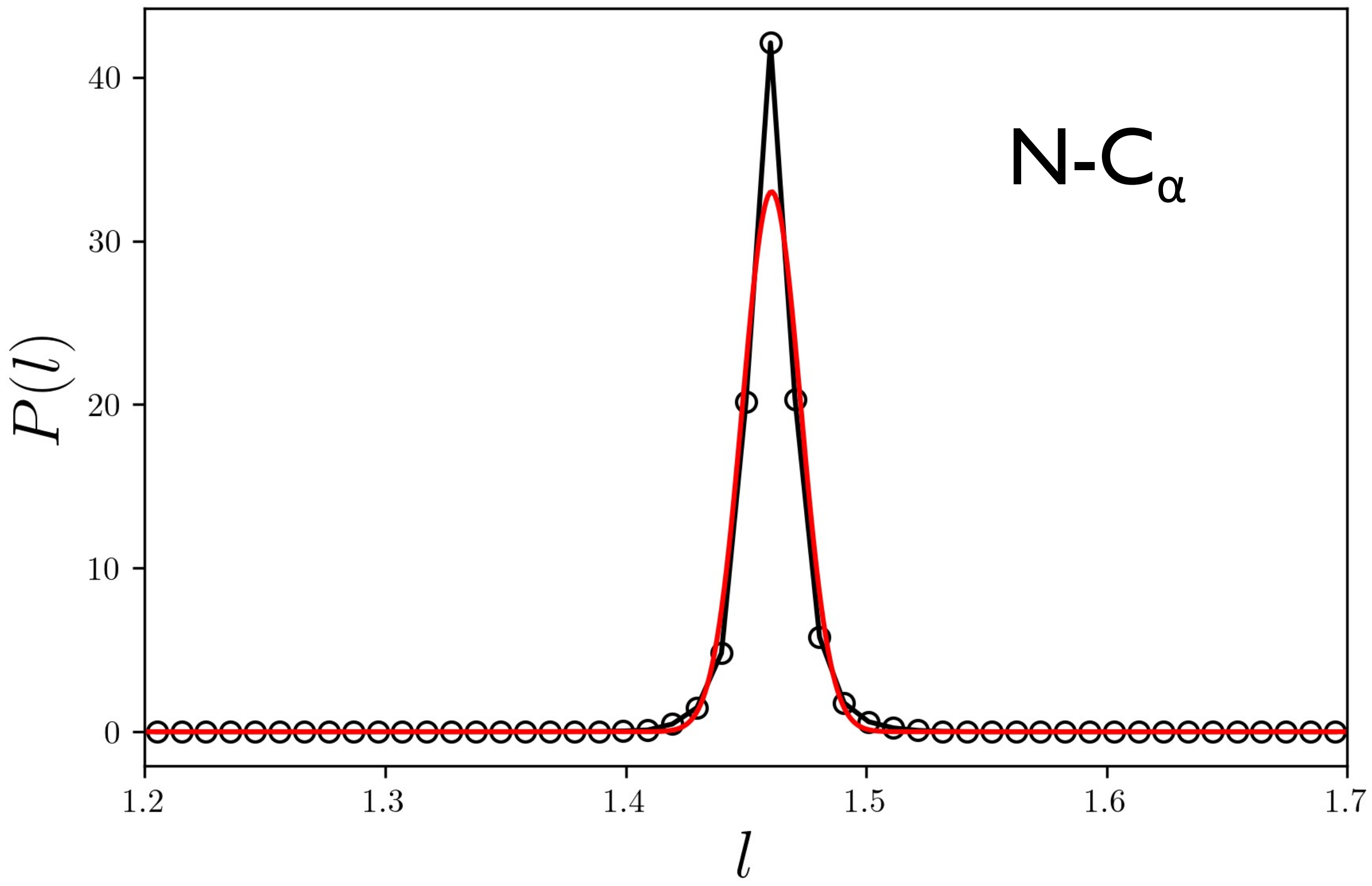
N_s=62,938 monomeric xtal structures



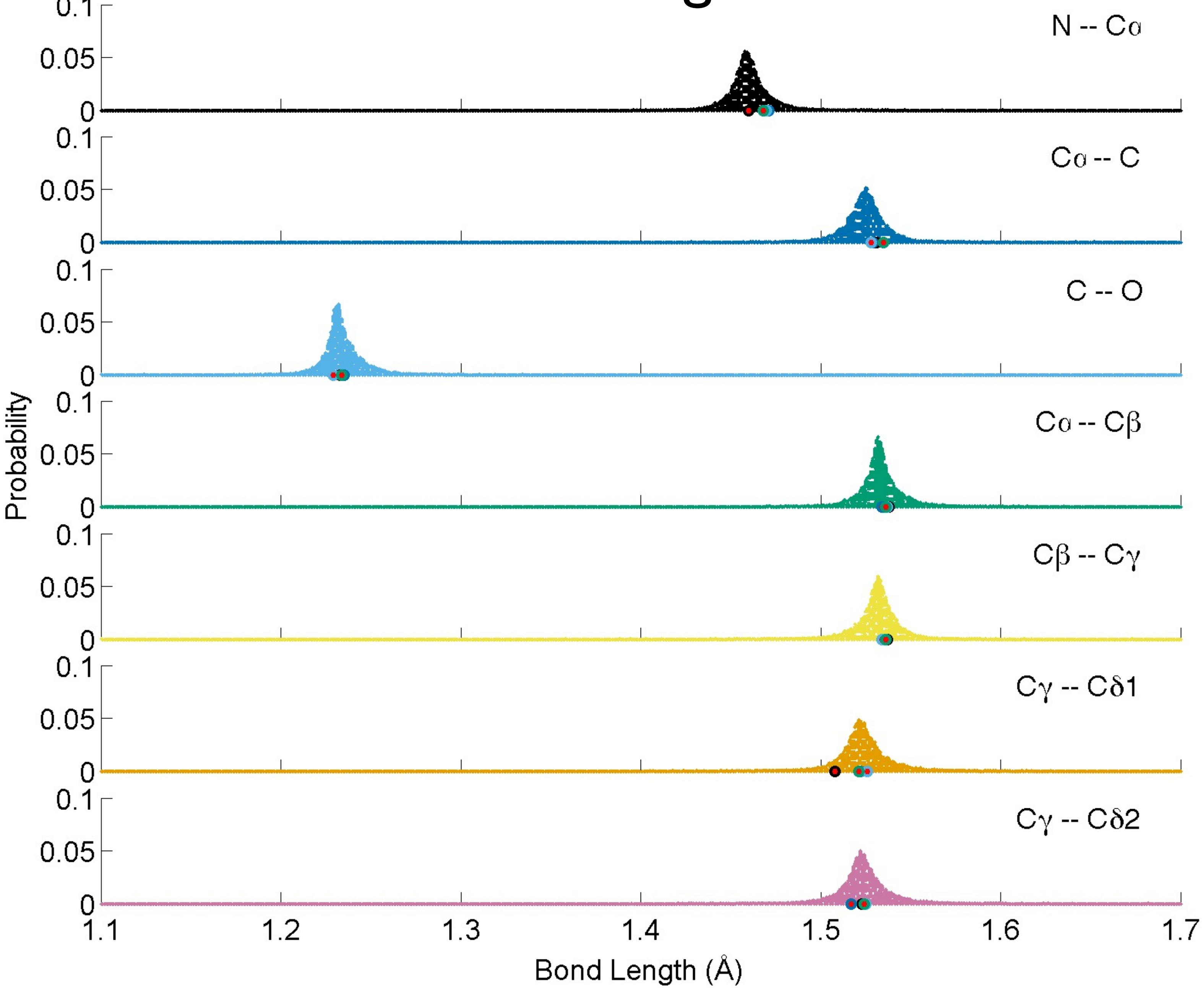
Bond Angles


Probability
0.02
0.01
0
N--Cα--Cβ
N--Cα--C
C--Cα--Cβ
Cα--C--O
Cα--Cβ--Cγ
Cβ--Cγ--Cδ1
Cβ--Cγ--Cδ2
Cδ2--Cδ--Cδ1
100
105
110
115
120
125
130
Bond Angle (°)

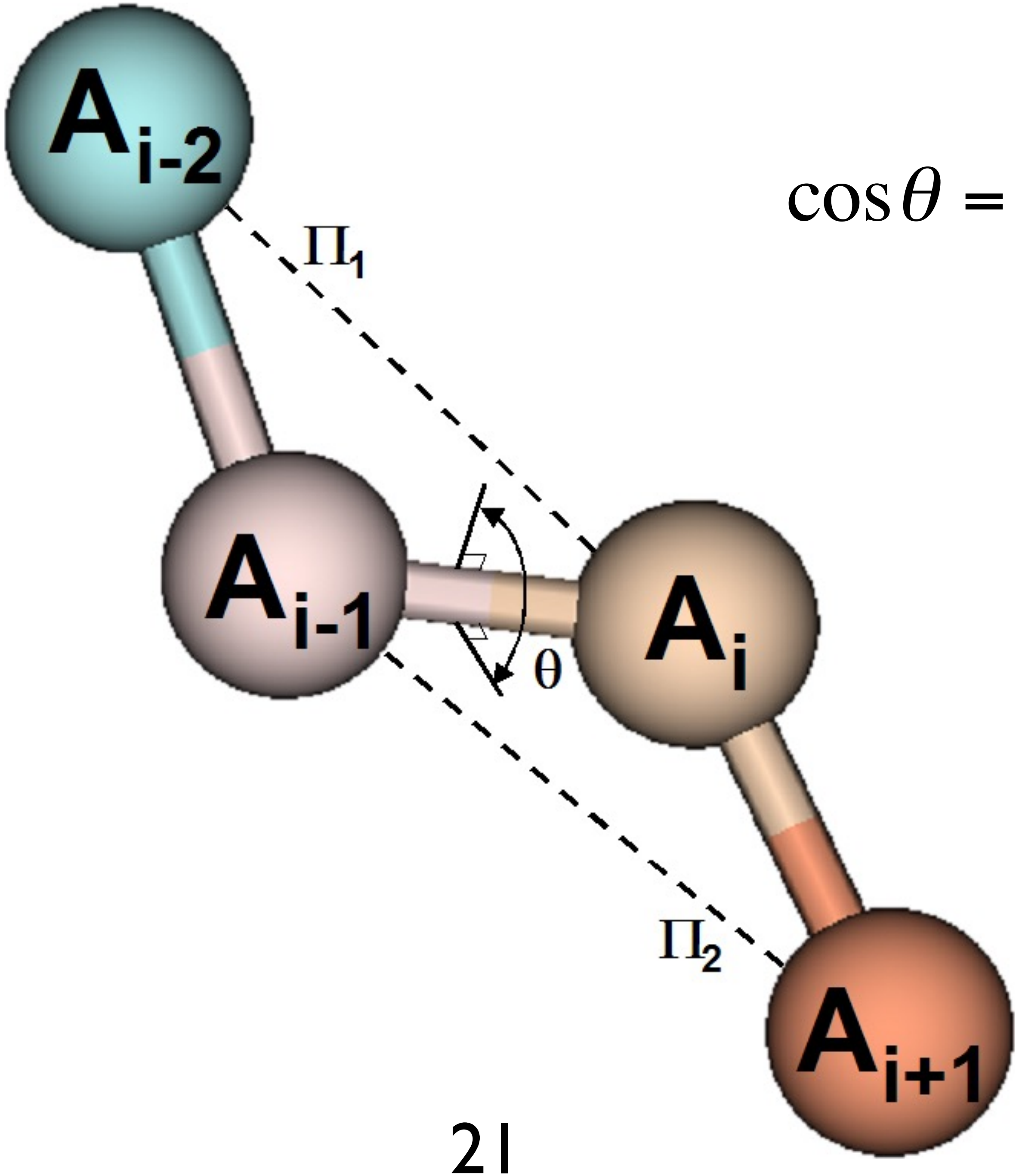

N-C_α-C
$P(\theta)$
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
θ
80
90
100
110
120
130
140
150


N-C_α
$P(l)$
40
30
20
10
0
1.2
1.3
1.4
1.5
1.6
1.7
l

Bond Lengths

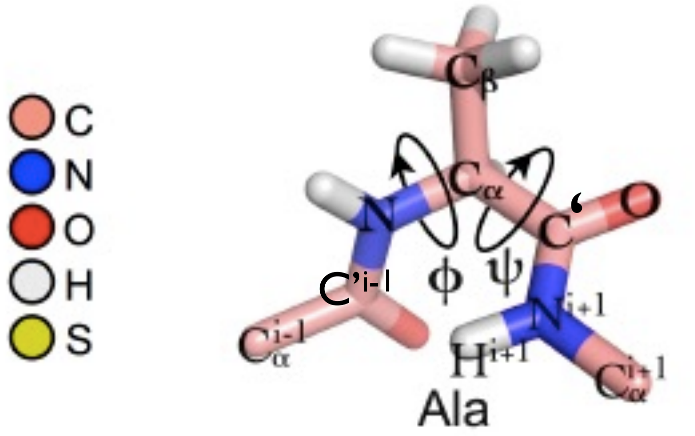


Backbonde Dihedral Angles



$$\cos\theta = \hat{\pi}_1 \bullet \hat{\pi}_2$$

$3N-6$ DoF	
$-(N-1)$	Bond lengths
$-(N-2)$	Bond angles
$=N-3$	Dihedral angles



ϕ: $C'^{i-1}NC_{\alpha}C'$

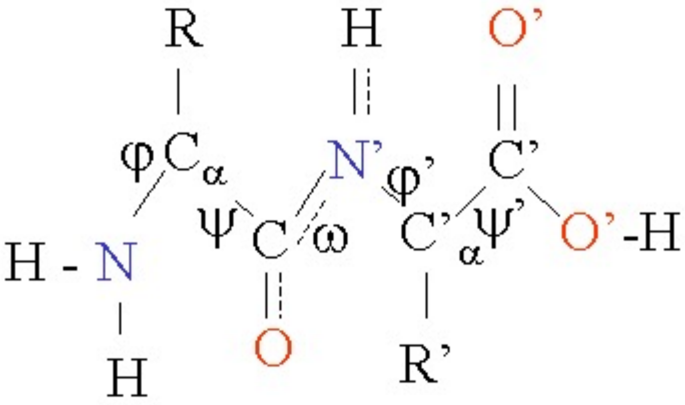
ψ: $NC_{\alpha}C'N^{i+1}$

ω_{I}:$C^{i-1}{}_{\alpha}C'^{i-1}NC_{\alpha}$

ω_{2}:$C_{\alpha}C'N^{i+1}C^{i+1}{}_{\alpha}$

Ramachandran Plot: Determining Steric Clashes

Backbone dihedral angles

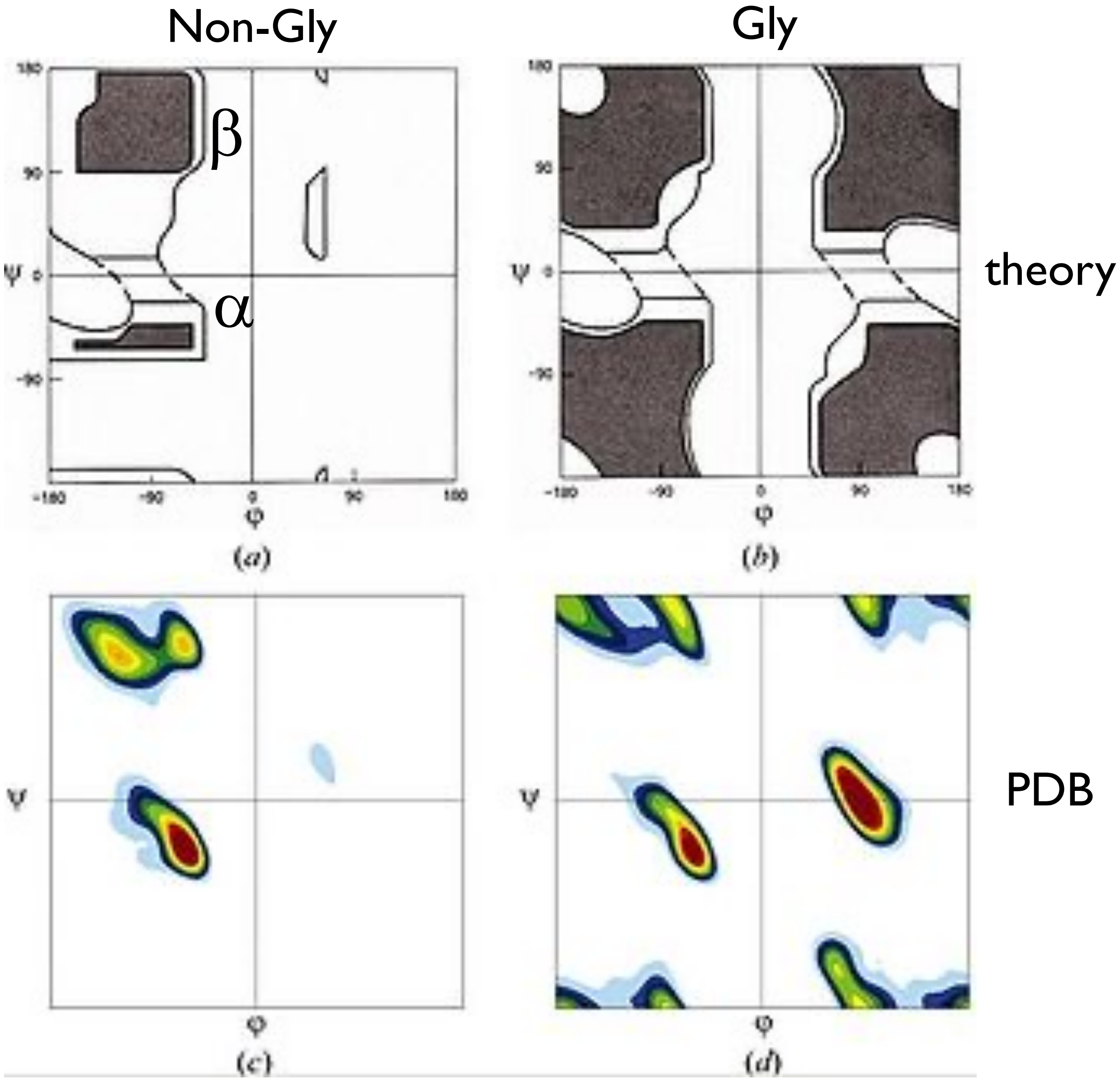


4 atoms define dihedral angle:

$C_{-1}NC_{\alpha}C$ ϕ

$C_{\alpha,-1}C_{-1}NC_{\alpha}$ $\omega=0,180°$

$NC_{\alpha}CN_{+1}$ ψ

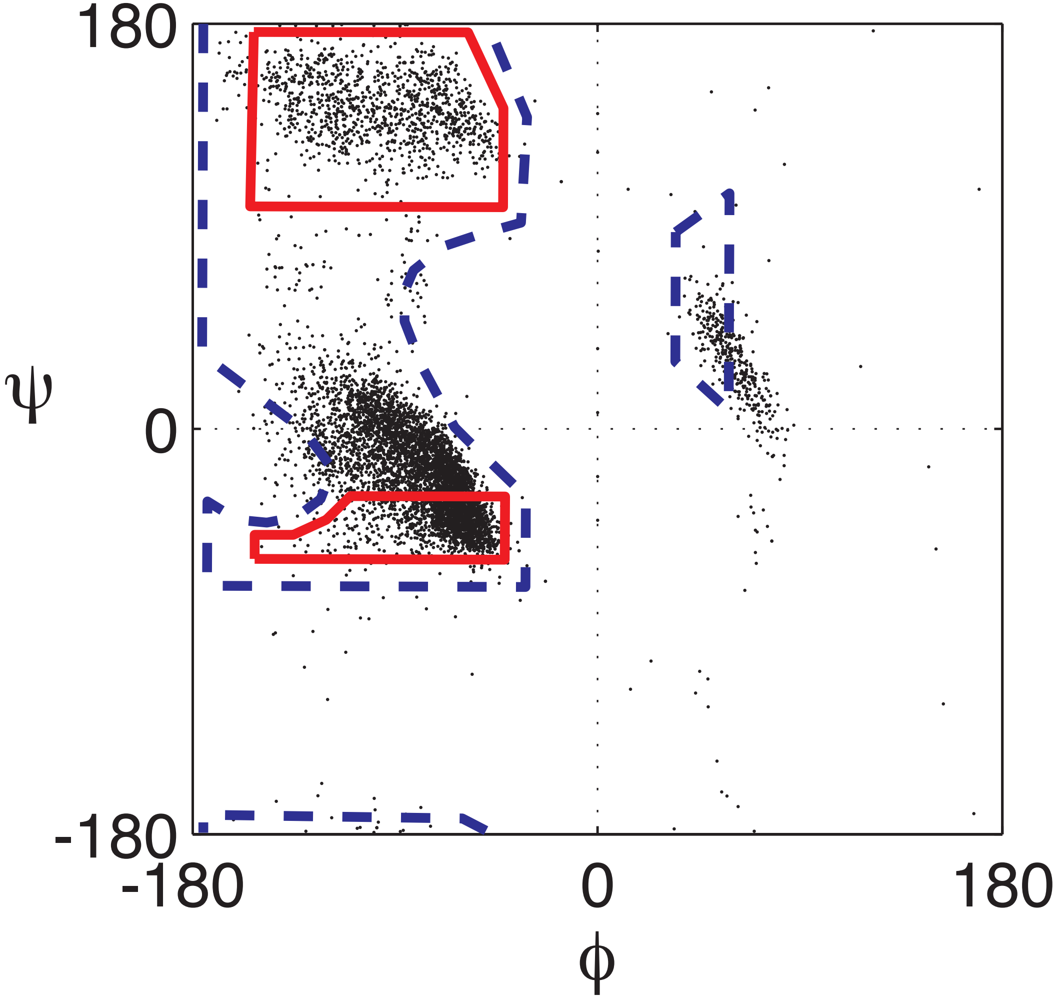


▨ vdW radii

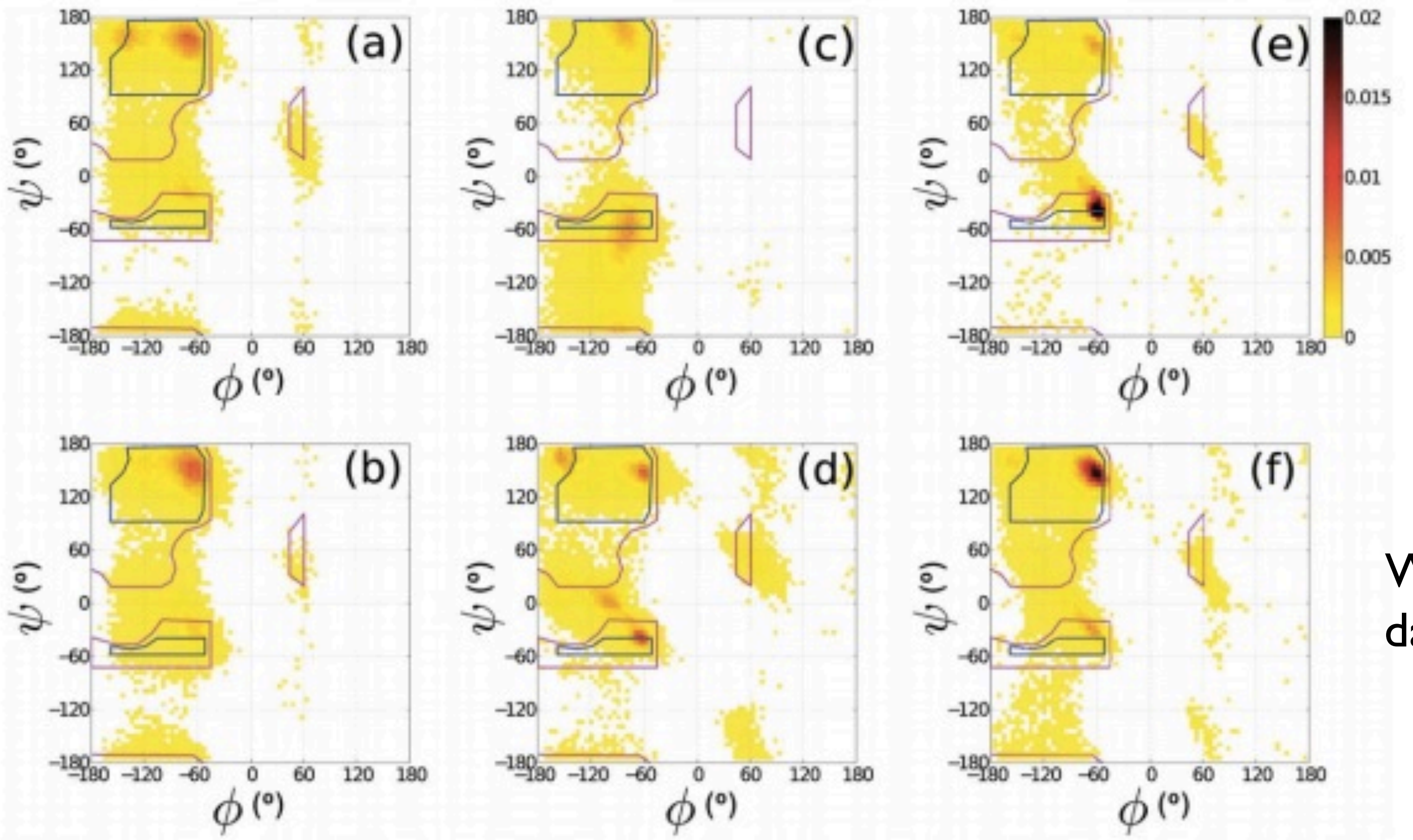
—— < vdW radii

- - - backbone flexibility

Backbone dihedral angles from PDB



Dunbrack 1.0



Wu coil database

Figure 5. Probability distributions $P(\phi, \psi)$ for the backbone dihedral angles ϕ and ψ obtained from MD simulations of an Ala dipeptide mimetic using recent versions of the CHARMM and Amber force fields, their associated optimized water models, and with and without the "ILDN-NMR" and "CMAP" dihedral angle potential corrections: (a) Amber99sb + TIP4P-Ew, (b) Amber99sb-ILDN-NMR + TIP4P-Ew, (c) CHARMM27 + TIP3SP, and (d) CHARMM27-CMAP+TIP3SP. Subpanels (e) and (f) correspond to the Ala ϕ-ψ distributions from the Dunbrack Database[38] and the Wu "Coil-3" library,[10] respectively. The Ramachandran hard-sphere[3] normal and outer limits (pink and blue lines, respectively) for $\tau = 110°$ are overlaid on each panel. The Amber and CHARMM MD simulations were thermally equilibrated at 303 K and sampled for 500 ns.

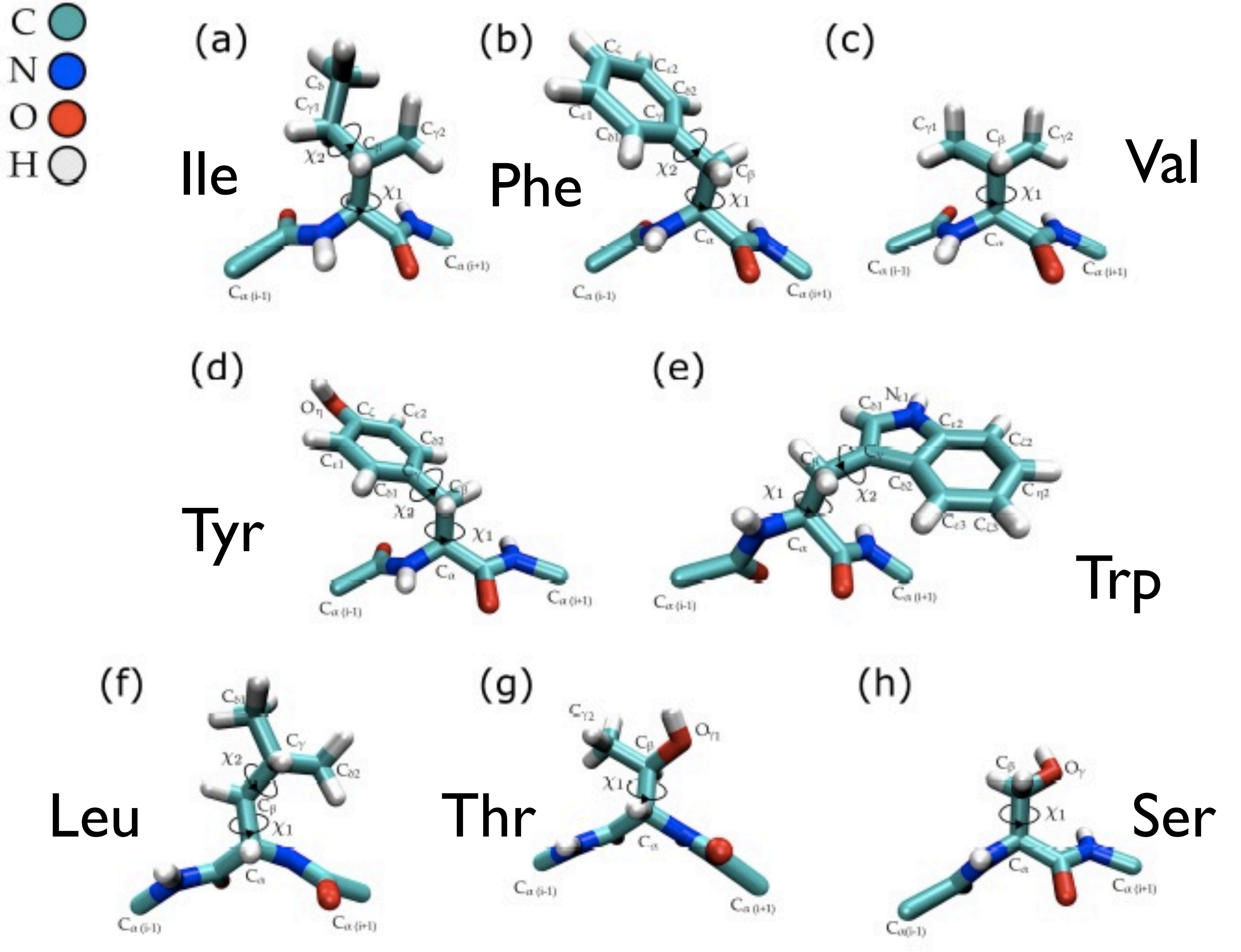
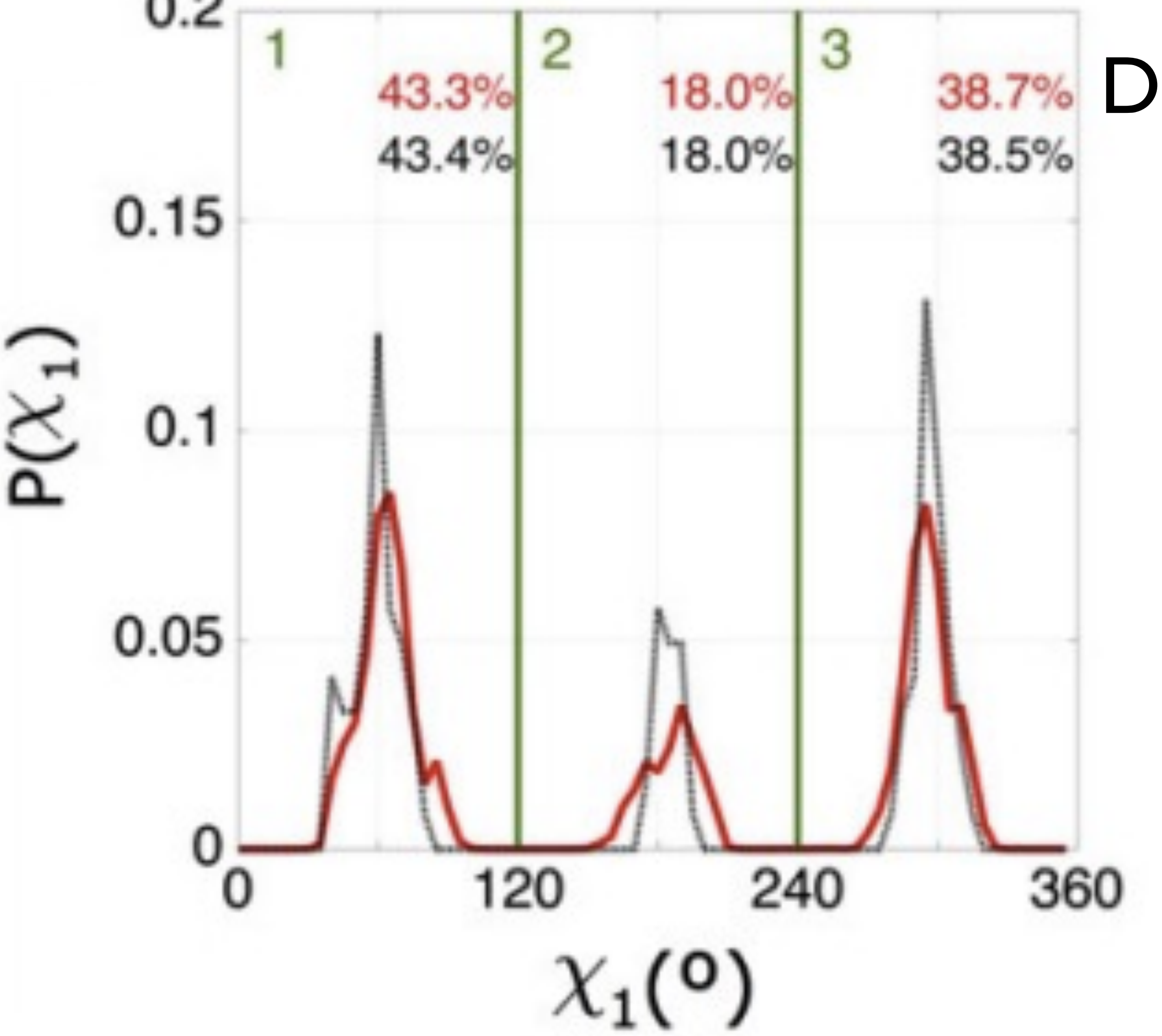


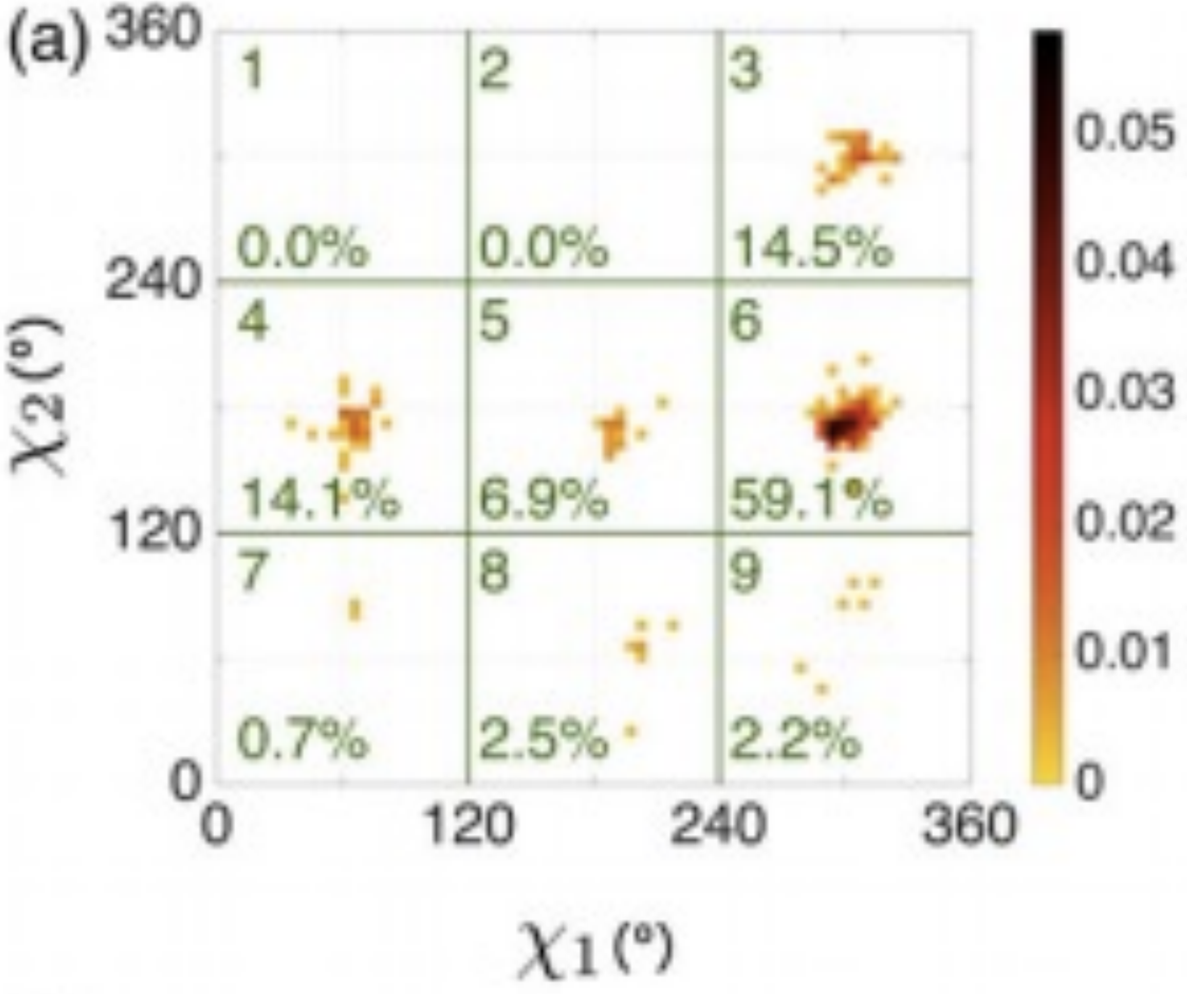
Figure S1: Stick representations of (a) Ile, (b) Phe, (c) Val, (d) Tyr, (e) Trp, (f) Leu, (g) Thr, and (h) Ser dipeptide mimetics. The carbon, nitrogen, oxygen, and hydrogen atoms are shaded green, blue, red, and white, respectively. The side chain dihedral angles χ_1 and χ_2 and several key atoms are labeled. The residues before (i-1) and after (i+1) the ith central residue are labeled at the C_α atom.

Thr



Dunbrack 1.0

Ile


(a)
360
240
120
0
χ2(°)
1
2
3
4
5
6
7
8
9
0.0%
0.0%
14.5%
14.1%
6.9%
59.1%
0.7%
2.5%
2.2%
0
120
240
360
χ1(°)
0.05
0.04
0.03
0.02
0.01
0

1. Can the structural properties of protein cores be quantitatively modeled using hard-spheres?

2. What is the packing fraction in protein cores?

3. Can simple hard-sphere model improve computational design of protein-protein interactions?