Supplemental Information - AWSEM-MD: Protein Structure Prediction Using Coarse-grained Physical Potentials and Bioinformatically Based Local Structure Biasing

Aram Davtyan,[†] Nicholas P. Schafer,[‡] Weihua Zheng,[¶] Cecilia Clementi,[‡] Peter G. Wolynes,[*,‡,¶] and Garegin A. Papoian[*,†]

Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, Department of Chemistry, Rice University, Houston, TX 77251, and Center for Theoretical Biological Physics, University of California in San Diego, La Jolla, CA 92093

E-mail: pwolynes@rice.edu; gpapoian@umd.edu

[*] To whom correspondence should be addressed

[†] University of Maryland

[‡] Rice University

[¶] UCSD

Introduction

Below we present the details of a coarse-grained model for protein simulations dubbed the Associative memory, Water mediated, Structure and Energy Model (AWSEM). This model has been continually developed over approximately two decades and successfully applied to many problems in protein physics.[1–33]

In this text and in our calculations in general we use kcal/mol for units of energy, Angstroms for length and radians for angles.

Description of the coarse-grained protein chain

According to AWSEM, the position and orientation of each amino acid residue is dictated by the positions of its C_α, C_β and O atoms (with the exception of glycine, which lacks a C_β atom). The positions of the other atoms in the backbone are calculated assuming an ideal geometry (Eq. (1)).

$$\begin{aligned}
\mathbf{r}_{N_i} &= 0.48318\mathbf{r}_{C_{\alpha_{i-1}}} + 0.70328\mathbf{r}_{C_{\alpha_i}} - 0.18643\mathbf{r}_{O_{i-1}} \\
\mathbf{r}_{C'_i} &= 0.44365\mathbf{r}_{C_{\alpha_i}} + 0.23520\mathbf{r}_{C_{\alpha_{i+1}}} + 0.32115\mathbf{r}_{O_i} \\
\mathbf{r}_{H_i} &= 0.84100\mathbf{r}_{C_{\alpha_{i-1}}} + 0.89296\mathbf{r}_{C_{\alpha_i}} - 0.73389\mathbf{r}_{O_{i-1}}
\end{aligned} \tag{1}$$

The third line in Eq. (1) gives the position of the hydrogen atom that is attached to the backbone nitrogen. Side chains and solvent are not explicitly present in the model; instead, the effects of side chains and solvent are aliased onto various interactions described in the next section.

The AWSEM Hamiltonian

The solvent averaged free energy function of the protein chain is given in Eq. (2).

$$V_{total} = V_{backbone} + V_{contact} + V_{burial} + V_{HB} + V_{AM} + V_{DSB} \tag{2}$$

The backbone term, $V_{backbone}$, is responsible for restricting the chain to “protein-like" conformations. It consists of several parts, which are shown in Eq. (3).

$$V_{backbone} = V_{con} + V_{chain} + V_{\chi} + V_{rama} + V_{excl} \tag{3}$$

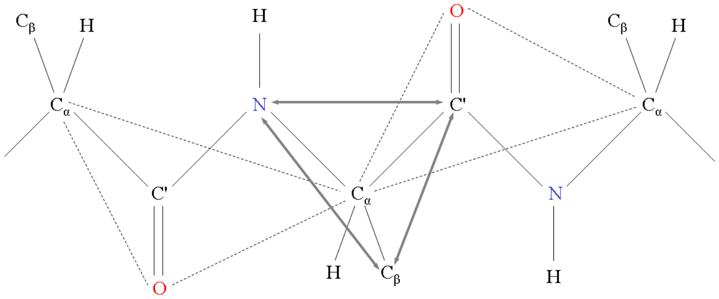


Figure S1: The connectivity of the chain is maintained by a combination of harmonic potentials. The distances contstrained by V_{con} are shown as dashed lines and the distances constrained by V_{chain} are shown as double headed arrows.

The connectivity of the protein chain is maintained by V_{con}, which is a sum of harmonic potentials. Its explicit form is given in Eq. (4), and a schematic of several amino acids is shown in Figure S1.

$$\begin{aligned} V_{con} &= \lambda_{con} \textstyle\sum_{i=1}^{N} [(\mathbf{r}_{C\alpha_i O_i} - \mathbf{r}^0_{C\alpha_i O_i})^2 + (\mathbf{r}_{C\alpha_i C_{\beta_i}} - \mathbf{r}^0_{C\alpha_i C_{\beta_i}})^2] \\ &+ \lambda_{con} \textstyle\sum_{i=1}^{N-1} [(\mathbf{r}_{C\alpha_i C\alpha_{i+1}} - \mathbf{r}^0_{C\alpha_i C\alpha_{i+1}})^2 + (\mathbf{r}_{O_i C\alpha_{i+1}} - \mathbf{r}^0_{O_i C\alpha_{i+1}})^2] \end{aligned} \tag{4}$$

The values of λ_{con} and the equilibrium distances are given in Table S1. In Eq. (4) and elsewhere, unless otherwise noted, i and j are residue indices and N is the total number of residues in the chain.

The correct bond angles around the C_α atom are also achieved using harmonic potentials, as shown in Eq. (5). The values of the parameters in Eq. (5) are given in Table S1.

$$V_{chain} = \lambda_{chain} \left[\sum_{i=2}^{N} (\mathbf{r}_{N_i C_{\beta_i}} - \mathbf{r}^0_{N_i C_{\beta_i}})^2 + \sum_{i=1}^{N-1} (\mathbf{r}_{C'_i C_{\beta_i}} - \mathbf{r}^0_{C'_i C_{\beta_i}})^2 + \sum_{i=2}^{N-1} (\mathbf{r}_{N_i C'_i} - \mathbf{r}^0_{N_i C'_i})^2 \right] \tag{5}$$

Table S1: Protein backbone potential parameters

Parameter	Value	Unites
λ_{con}	10.0	kcal/Å^2 mol
λ_{chain}	5.0	kcal/Å^2 mol
λ_{χ}	20.0	kcal/Å^6 mol
λ_{rama}	2.0	kcal/mol
λ_{excl}	20.0	kcal/Å^2 mol
$\mathbf{r}^0_{C\alpha_i C\alpha_{i+1}}$	3.80	Å
$\mathbf{r}^0_{C\alpha_i CO_i}$	2.43	Å
$\mathbf{r}^0_{CO_i C\alpha_{i+1}}$	2.82	Å
$\mathbf{r}^0_{C\alpha_i C\beta_i}$	1.54	Å
$\mathbf{r}^0_{N_i C_{\beta_i}}$	2.46	Å
$\mathbf{r}^0_{C'_i C_{\beta_i}}$	2.70	Å
$\mathbf{r}^0_{N_i C'_i}$	2.46	Å
χ_0	-0.83	Å^3

The chirality term, V_χ, given in Eq. (6), ensures the correct orientation of the C_β atom relative to the plane formed by the C', C_α and N atoms. A value of $\chi_0 = -0.83\text{Å}^3$, corresponding to an L-amino acid, is used for all residues except Glycine, which is excluded from this potential because it lacks a C_β atom. The values of the parameters in Eq. (6) are given in Table S1.

$$\begin{aligned} V_\chi &= \lambda_\chi \sum_{i=2}^{N-1} (\chi_i - \chi_0)^2 \\ \chi_i &= \left(\mathbf{r}_{C'_i C_{\alpha_i}} \times \mathbf{r}_{C_{\alpha_i} N_i}\right) \cdot \mathbf{r}_{C_{\alpha_i} C_{\beta_i}} \end{aligned} \tag{6}$$

To reproduce the experimental distribution of backbone dihedral angles, we use a Ramachandran potential, V_{rama}, shown in Eq. (7). The resulting potential is plotted in Figure S2. The value of λ_{rama} used is given in Table S1. All other parameters are given in Table S2, where ϕ_0 and ψ_0 are given in radians and W, σ, ω_ϕ and ω_ψ are unitless weights.

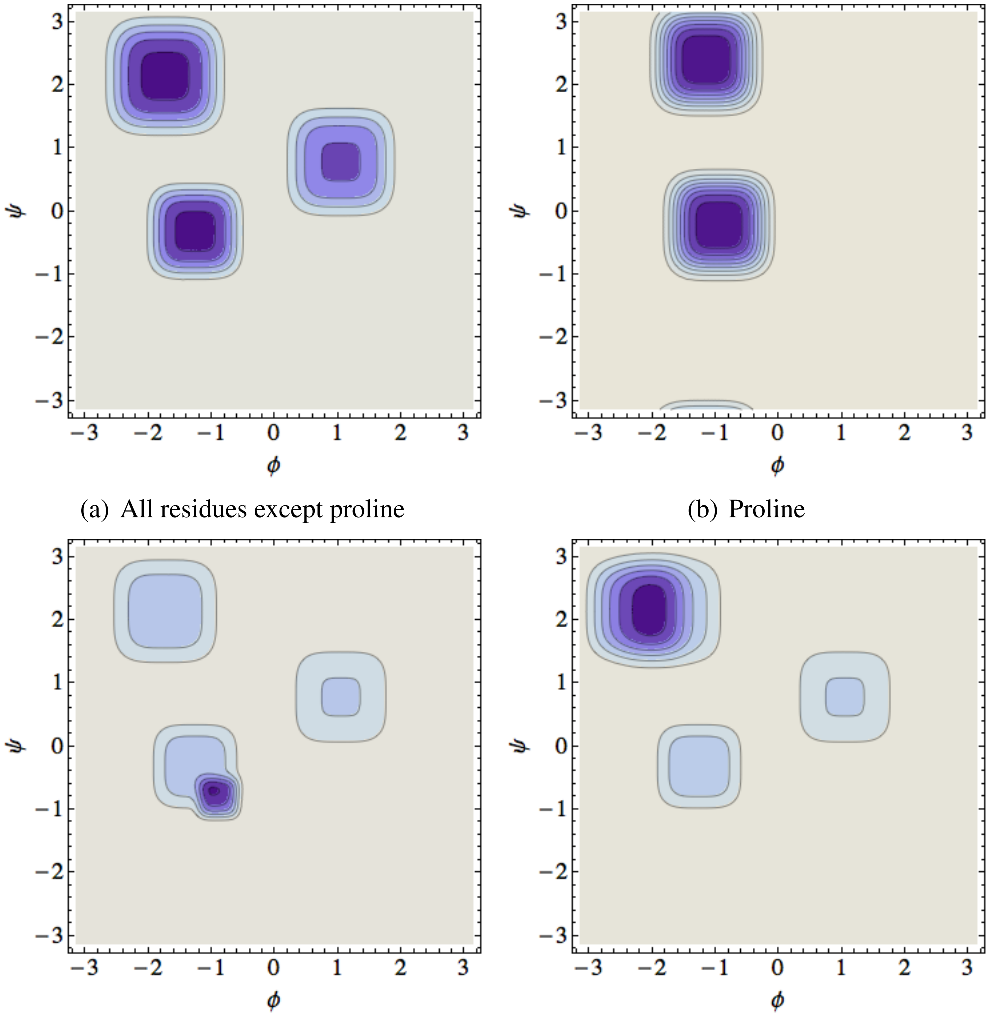
$$V_{rama} = -\lambda_{rama} \sum_{i=2}^{N-1} \sum_j W_j e^{-\sigma_j (\omega_{\phi_j} (\cos(\phi_i - \phi_{0j}) - 1)^2 + \omega_{\psi_j} (\cos(\psi_i - \psi_{0j}) - 1)^2)} \tag{7}$$

The first, last and glycine residues are not included in this potential. ϕ_i is the dihedral angle between the C'_{i-1}, N_i, C_{α_i} and C'_i atoms, and ψ_i is the dihedral angle between the N_i, C_{α_i}, C'_i and N_{i+1} atoms. The first three columns of Table S2 represent the set of the parameters for the general case of non-proline residues. These three columns correspond to right handed helix, left handed helix and β regions of the Ramachandran plot (see Figure S2). Parameters from the next two columns can be used to bias the secondary structure towards right handed alpha helix or beta sheet based on a secondary structure prediction server (*e.g.,* JPRED[34]). The final two columns of Table S2 refer to proline residues, which are known to have different allowed regions for the dihedral angles. The index j in Eq. (6) in this case is not a residue index; instead, it runs over each column of parameters that is appropriate for residue i.

Table S2: Ramachandran potential parameters

	General Case			Alpha Helix	Beta Sheet	Proline	
W	1.3149	1.32016	1.0264	2.0	2.0	2.17	2.15
σ	15.398	49.0521	49.0954	419.0	15.398	105.52	109.09
ω_ϕ	0.15	0.25	0.65	1.0	1.0	1.0	1.0
ϕ_0	-1.74	-1.265	1.041	-0.895	-2.25	-1.153	-0.95
ω_ψ	0.65	0.45	0.25	1.0	1.0	0.15	0.15
ψ_0	2.138	-0.318	0.78	-0.82	2.16	2.4	-0.218

V_{excl} is the excluded volume interaction that provides a repulsion between atoms at short distances, preventing them from overlapping. It has the form given in Eq. (8) where $r^C_{ex} = 3.5\text{Å}$ for sequence separation less then 5 and 4.5Å otherwise, whereas $r^O_{ex} = 3.5\text{Å}$ for any sequence separation. The subscript C refers to both C_α and C_β atoms. In Eq. (8), i and j are atom indices, which run over all pairs of C or O atoms that are not directly connected by V_{con}. $\Theta(x)$ is the Heaviside



(a) All residues except proline

(b) Proline

(c) Biased toward alpha helix

(d) Biased toward beta sheet

Figure S2: Ramachandran potential, V_{rama}. Secondary structure biasing is achieved by adding additional wells to the Ramachandran potential. The colors in (a), (b), (c) and (d) are not normalized to the same scale.

step function. The form of V_{excl} for single pair of oxygens is plotted in Figure 3(a).

$$\begin{aligned} V_{excl} = \; & \lambda_{excl} \sum_{ij} \Theta(r_{C_i,C_j} - r_{ex}^{C})(r_{C_i,C_j} - r_{ex}^{C})^2 \\ & + \lambda_{excl} \sum_{ij} \Theta(r_{O_i,O_j} - r_{ex}^{O})(r_{O_i,O_j} - r_{ex}^{O})^2 \end{aligned} \quad (8)$$

When describing $V_{contact}$, it is useful to define two C_β-C_β distance ranges (replaced by C_α atom

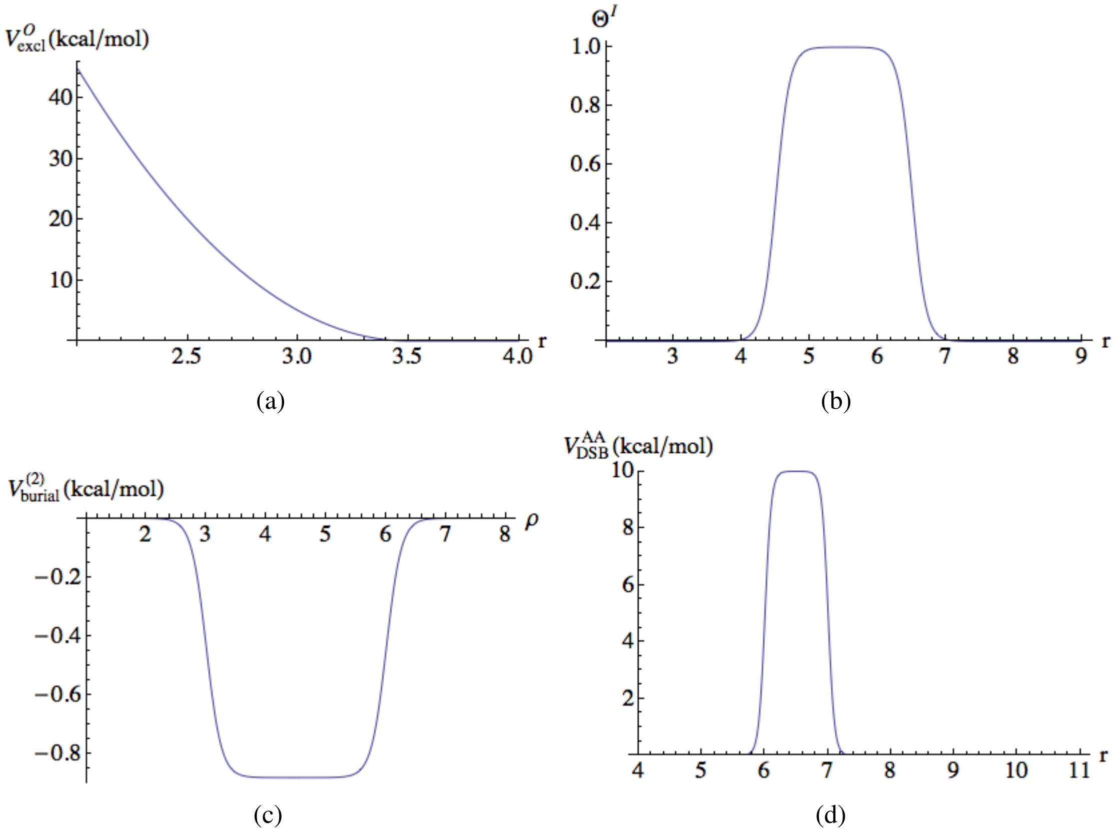


Figure S3: (a) Plot of excluded volume potential vs. distance between two oxygens. (b) Plot of Θ function defined in Eq. (9) for direct contact well vs. distance between two residues. (c) Plot of burial potential function for $\mu = 2$ vs. local density value. (d) Plot of desolvation barrier between two alanines vs. distance between them.

in the case of glycine), hereafter identified by the superscripts I and II. The first distance range, the “direct contact well", goes from $r^{I}_{min} = 4.5$Å to $r^{I}_{max} = 6.5$Å. The second distance range, the “water or protein mediated well", goes from $r^{II}_{min} = 6.5$Å to $r^{II}_{max} = 9.5$Å. If the C_{β} atoms of two residues i and j are separated by a distance between r^{μ}_{min} and r^{μ}_{max}, then the function Θ^{μ}_{ij}, given in Eq. (9), will be equal to 1; otherwise, it will be 0. It switches smoothly from 1 to 0 near the extremes of the distance ranges (see Figure 3(b)).

$$\Theta^{\mu}_{ij} = \frac{1}{4}\left(1+\tanh\left[\eta\left(r_{ij}-r^{\mu}_{min}\right)\right]\right)\left(1+\tanh\left[\eta\left(r^{\mu}_{max}-r_{ij}\right)\right]\right) \tag{9}$$

By summing Θ_{ij}^{μ} over j, you can obtain the number of residues in the μ-well of residue i. The local density, ρ_i, of residue i is defined as $\rho_i = \sum_{j=1}^{N} \Theta_{ij}^{I}$, which is equal to the number of residues in its “direct contact well".

$V_{contact}$ is a contact interaction term between residues far apart in sequence.[6] It consists of V_{direct} and V_{water}. V_{direct} is a pairwise additive potential with the form given in Eq. (10)

$$V_{direct} = -\lambda_{direct} \sum_{j-i>9}^{N} \gamma_{ij}(a_i, a_j)\Theta_{ij}^{I} \tag{10}$$

where r_{ij} is the C_β-C_β distance between residues i and j, and $\gamma(a_i, a_j)$ is a residue type specific constant. The γ parameters were optimized to maximize the ratio of the folding temperature to the glass transition temperature of the model, $\frac{T_f}{T_g}$.[6] In Eq. (10) and elsewhere, a_i refers to the residue type of residue i.

V_{water} is a many-body interaction term that switches between water-mediated and protein-mediated interaction weights depending on the local density around the interacting residues. The explicit form is given in Eq. (11)

$$V_{water} = -\lambda_{water} \sum_{j-i>9}^{N} \Theta_{ij}^{II} \left(\sigma_{ij}^{wat} \gamma_{ij}^{wat}(a_i, a_j) + \sigma_{ij}^{prot} \gamma_{ij}^{prot}(a_i, a_j) \right) \tag{11}$$

where σ_{ij}^{wat} and σ_{ij}^{prot} are the switching functions defined in Eq. (12).

$$\begin{aligned} \sigma_{ij}^{wat} &= \tfrac{1}{4}\left(1 - \tanh\left[\eta_\sigma\left(\rho_i - \rho_0\right)\right]\right)\left(1 - \tanh\left[\eta_\sigma\left(\rho_j - \rho_0\right)\right]\right) \\ \sigma_{ij}^{prot} &= 1 - \sigma_{ij}^{wat} \end{aligned} \tag{12}$$

σ_{ij}^{prot} and σ_{ij}^{wat} switch smoothly from 0 to 1 and 1 to 0, respectively, as either of the local densities, ρ_i or ρ_j, exceeds a threshold $\rho_0 = 2.6$. A plot of σ_{ij}^{wat} is given in Figure S4.

The burial term, V_{burial}, given in Eq. (13), is a many body interaction which is based on a particular residue type's propensity to be in a low ($\mu = 1$, $\rho_{min}^{1} = 0.0$, $\rho_{max}^{1} = 3.0$), medium ($\mu = 2$,

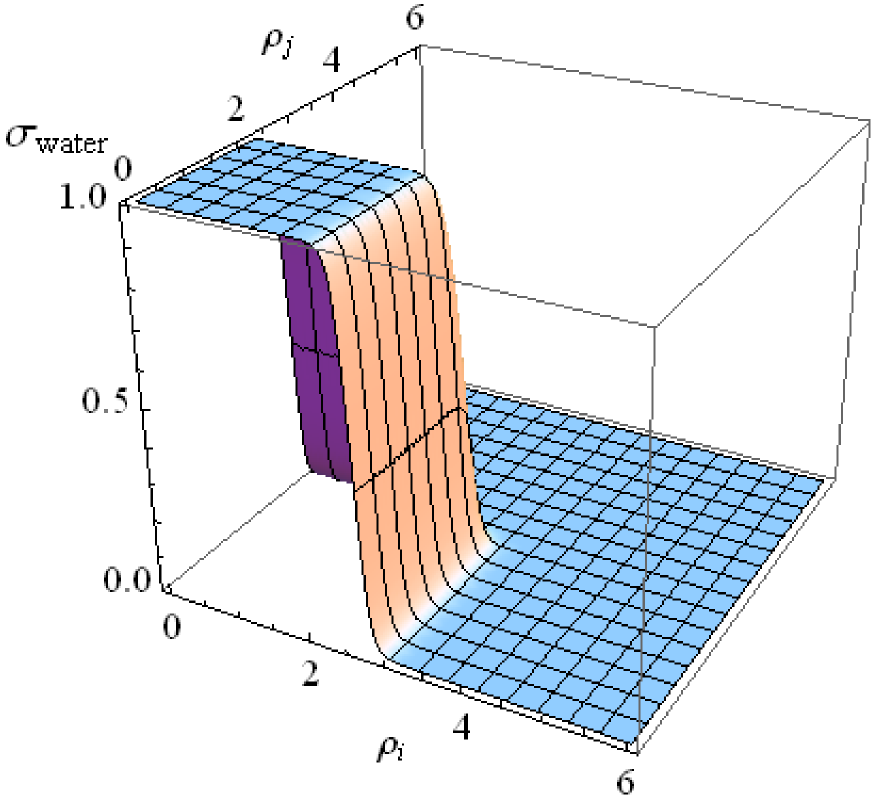


Figure S4: Plot of σ_{ij}^{wat} in Eq. (12); adapted with permission from.[6]

$\rho_{min}^2 = 3.0$, $\rho_{max}^2 = 6.0$), or high ($\mu = 3$, $\rho_{min}^3 = 6.0$, $\rho_{max}^3 = 9.0$) density environment.[6] These propensities are given by the $\gamma_{burial}(a_i, \rho_i)$ coefficients in Table S4.

$$V_{burial} = -\frac{1}{2}\lambda_{burial}\sum_{i=1}^{N}\sum_{\mu=1}^{3}\gamma_{burial}(a_i,\rho_i)\left(\tanh\left[\eta\left(\rho_i - \rho_{min}^{\mu}\right)\right] + \tanh\left[\eta\left(\rho_{max}^{\mu} - \rho_i\right)\right]\right) \tag{13}$$

The hydrogen bonding potential, V_{HB}, given in Eq. (14), is a sum of three terms.

$$V_{HB} = V_{\beta} + V_{P-AP} + V_{helical} \tag{14}$$

The first two terms of Eq. (14) are β hydrogen bonding terms. The V_{β} potential has the form given in Eqs. (15) to (18) where r_{ij}^{ON} is the distance from the carbonyl oxygen of residue i to the nitrogen of residue j, and r_{ij}^{OH} is the distance from carbonyl oxygen of residue i to the backbone amide

Table S3: Potential parameters

Parameter	Value	Unites	Parameter	Value	Unites
V_{direct}					
λ_{direct}	1.0	kcal/mol	η	5.0	Å^{-1}
V_{water}					
λ_{water}	1.0	kcal/mol	η	5.0	Å^{-1}
ρ_0	2.6		η_σ	7.0	
$V_{helical}$					
$\lambda_{helical}$	1.5	kcal/mol	ρ_0	3.0	
γ_{prot}	2.0		$\left\langle r^{ON} \right\rangle$	2.98	Å
γ_{wat}	-1.0		$\left\langle r^{OH} \right\rangle$	2.06	Å
η	7.0	Å^{-1}	σ_{ON}	0.68	Å
η_σ	7.0		σ_{OH}	0.76	Å
V_β					
$\left\langle r^{ON} \right\rangle$	2.98	Å	η^I	1.0	Å^{-1}
$\left\langle r^{OH} \right\rangle$	2.06	Å	η^{II}	0.5	Å^{-1}
σ_{ON}	0.68	Å	r_c^{HB}	12.0	Å
σ_{OH}	0.76	Å			
V_{P-AP}					
γ_{APH}	1.0	kcal/mol	η	7.0	Å^{-1}
γ_{AP}	0.4	kcal/mol	r_0	8.0	Å
γ_P	0.4	kcal/mol			
V_{burial}					
λ_{burial}	1.0	kcal/mol	η	4.0	
V_{AM}					
λ_{AM}	1.0	kcal/mol			
V_{DSB}					
λ_{DSB}	10.0	kcal/mol	r^0_{min}	6.0	Å
κ_{DSB}	10.0	Å^{-1}	r^0_{max}	7.0	Å

hydrogen of residue j. $\left\langle r^{ON} \right\rangle$ and $\left\langle r^{OH} \right\rangle$ are the corresponding equilibrium bond lengths, and σ_{NO} and σ_{HO} are their variances.

$$V_\beta^{ij} = -[\Lambda_1(|j-i|)\theta_{i,j} + \Lambda_2(a_i, a_j, |j-i|)\theta_{i,j}\theta_{j,i} + \Lambda_3(a_i, a_j, |j-i|)\theta_{i,j}\theta_{j,i+2}]\nu_i^I \nu_j^{II} \tag{15}$$

Table S4: Burial potential, V_{burial}, coefficients $\gamma_{burial}(a_i, \rho_i)$

a_i		ρ_i	
	0.0-3.0	3.0-6.0	6.0-9.0
Ala	0.84	0.88	0.57
Arg	0.94	0.83	0.13
Asn	0.96	0.79	0.25
Asp	0.98	0.75	0.20
Cys	0.67	0.94	0.66
Gln	0.96	0.79	0.24
Glu	0.97	0.78	0.16
Gly	0.94	0.81	0.34
His	0.92	0.85	0.13
Ile	0.78	0.92	0.55
Leu	0.78	0.94	0.46
Lys	0.98	0.75	0.00
Met	0.82	0.92	0.46
Phe	0.81	0.94	0.33
Pro	0.97	0.76	0.25
Ser	0.94	0.79	0.38
Thr	0.92	0.82	0.40
Trp	0.85	0.91	0.34
Tyr	0.83	0.92	0.34
Val	0.77	0.93	0.55

$$\theta_{i,j} = \exp\left[-\frac{(r_{ij}^{ON} - \langle r^{ON}\rangle)^2}{2\sigma_{NO}^2} - \frac{(r_{ij}^{OH} - \langle r^{OH}\rangle)^2}{2\sigma_{HO}^2}\right] \tag{16}$$

$$\nu_i^{\mu} = \frac{1}{2}\left(1 + \tanh\left[\eta^{\mu}\left(r_{i-2,i+2}^{C\alpha} - r_c^{HB}\right)\right]\right) \tag{17}$$

$$\begin{aligned}
\Lambda_1(|j-i|) &= \lambda_1(|j-i|) \\
\Lambda_2(a_i, a_j, |j-i|) &= \lambda_2(|j-i|) - 0.5\alpha_1(|j-i|)lnP_{HB}(a_i, a_j) \\
&\quad -0.25\alpha_2(|j-i|)[lnP_{NHB}(a_{i+1}, a_{j-1}) + lnP_{NHB}(a_{i-1}, a_{j+1})] \\
&\quad -\alpha_3(|j-i|)[lnP_{anti}(a_i) + lnP_{anti}(a_j)] \\
\Lambda_3(a_i, a_j, |j-i|) &= \lambda_3(|j-i|) - \alpha_4(|j-i|)lnP_{parHB}(a_{i+1}, a_j) \\
&\quad -\alpha_5(|j-i|)lnP_{par}(a_{i+1}) + \alpha_4(|j-i|)lnP_{par}(a_j)
\end{aligned} \tag{18}$$

The first term in the Eq. (15) describes simple pairwise additive hydrogen bonding interactions. The second term gives additional cooperative stabilization to anti-parallel β conformations and the third term gives additional cooperative stabilization to parallel β conformations. All of the Λ_k coefficients depend on the sequence separation of residues i and j, and the coefficients Λ_2 and Λ_3 are also amino acid type (a_i and a_j) dependent. The constants $\langle r^{ON} \rangle$, $\langle r^{OH} \rangle$, σ_{NO}, σ_{HO} (see Table S3) and probabilities, P, for amino acids to be hydrogen bonded (HB) or not hydrogen bonded (NHB) were extracted from a database of well-resolved protein structures.[28] The parameters λ and α of Eq. (18) were optimized to maximize the T_f/T_g ratio.[6] Their values for different sequence separation classes are given in Table S5. For $|j-i| < 18$, $\lambda_3 = 0$ because parallel hydrogen bonds rarely form between residues which are less than 18 amino acids apart. The ν_i and ν_j terms ensure that β hydrogen bonding does not occur between residues that are in the middle of a five residue segment that is shorter than $r_c^{HB} = 12.0$Å, as β hydrogen bonding networks tend not to form between chain segments that are not at least somewhat extended.

Table S5: Hydrogen bonding potential λ and α coefficients, in *$kcal/mol$*

sequence separation	λ_1	λ_2	λ_3	α_1	α_2	α_3	α_4	α_5
$4 \leq \|j-i\| < 18$	1.37	3.89	0.0	1.30	1.32	1.22	0.0	0.0
$18 \leq \|j-i\| < 45$	1.36	3.50	3.47	1.30	1.32	1.22	0.33	1.01
$\|j-i\| \geq 45$	1.17	3.52	3.62	1.30	1.32	1.22	0.33	1.01

V_β will stabilize an already formed β hydrogen bonding network, but small deviations from an ideal β-sheet geometry will be significantly higher in energy. However, during secondary structure formation it is necessary to search through many possible conformations. The “liquid-crystal potential", V_{P-AP}, enables a protein chain to adopt approximate parallel or antiparallel β-sheet conformations before the hydrogen bonds are fully formed. The strength of this potential is chosen so that structures can easily fall apart and reassemble. The general form of this potential is given in Eq. (19).

$$\begin{aligned} V_{P-AP} = & -\gamma_{APH} \sum_{i=1}^{N-13} \sum_{j=i+13}^{min(i+16,N)} \nu_{i,j}\nu_{i+4,j-4} \\ & -\gamma_{AP} \sum_{i=1}^{N-17} \sum_{j=i+17}^{N} \nu_{i,j}\nu_{i+4,j-4} - \gamma_{P} \sum_{i=1}^{N-13} \sum_{j=i+9}^{N-4} \nu_{i,j}\nu_{i+4,j+4} \end{aligned} \tag{19}$$

V_{P-AP} favors contacts between residues i and j if residues $i+4$ and $j+4$ (parallel, P) or $i+4$ and $j-4$ (antiparallel, AP) are already in contact. Formation of β-hairpins (APH) is separate from the general antiparallel case to allow for the possibility of assigning it a different weight. Two residues are considered to be in contact with each other if the distance between their C_α atoms is less than r_0. Thus, $\nu_{i,j}$ is defined as the smooth switching function $\nu_{i,j} = \frac{1}{2}\left(1+\tanh\left[\eta\left(r_0 - r_{C\alpha_i,C\alpha_j}\right)\right]\right)$, where $\eta = 7.0\text{Å}^{-1}$ and $r_0 = 8.0\text{Å}$. γ_{AP} and γ_P usually take the value of 0.4 kcal/mol. Only in the case when secondary structure prediction information is available and both residues i and j are predicted to be in a β-strand do we use a value of $\gamma_{AP} = \gamma_P = 0.6$ kcal/mol instead.

The $V_{helical}$ term, given in Eq. (20), is responsible for the formation of alpha helices.[35]

$$\begin{aligned} V_{helical} = & -\lambda_{helical} \sum_{i=1}^{N-4} (f(a_i) + f(a_{i+4}))(\gamma_{prot}\sigma_{i,i+4}^{prot} + \gamma_{wat}\sigma_{i,i+4}^{wat}) \times \\ & \exp\left[-\frac{\left(r_{i,i+4}^{ON} - \langle r^{ON}\rangle\right)^2}{2\sigma_{ON}^2} - \frac{\left(r_{i,i+4}^{OH} - \langle r^{OH}\rangle\right)^2}{2\sigma_{OH}^2}\right] \end{aligned} \tag{20}$$

In Eq. (20), $f(a_i)$ (see Table S6) is the probability of finding residue i in a helix. All residue types have positive values between 0 and 1 except for proline, as it lacks a backbone amide hydrogen and therefore can only be a hydrogen bond acceptor, but never a donor. To reflect this we use $f(a_{i+4}) = -3.0$ if the $i+4$ residue is a proline. σ_{ij}^{prot} and σ_{ij}^{wat} are the same as in Eq. (12). γ_{prot} is the strength of the interaction when both residues are buried. When residues are exposed to water, they are allowed to form hydrogen bonds with surrounding water molecules and forming hydrogen bonds with each other is not as favorable. Thus γ_{wat} is negative, as shown in Table S3.

V_{AM} is the associative memory potential. When combined with known protein structures and

Table S6: $f(a_i)$ **values**

a_i	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
$f(a_i)$	0.77	0.68	0.07	0.15	0.23	0.33	0.27	0.0	0.06	0.23
a_i	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
$f(a_i)$	0.62	0.65	0.50	0.41	0.4/-3.0	0.35	0.11	0.45	0.17	0.14

an algorithm for aligning a target sequence to those structures, it can be used to limit the local (secondary structure) conformational search. Each portion of a known structure which is aligned to a particular set of residues in the target sequence is known as a “memory". In this paper, we have used a “fragment memory" approach wherein the memories are short (9 residues or less) and the fragments are chosen using BLAST.[36] The maximum sequence separation of interacting residues is determined either by the length of the memory or a maximum cutoff, whichever is shorter. The form of the V_{AM} potential is given in Eq. (21) where i and j go over all C_α and C_β atoms up to a maximum sequence separation, which in this case includes the entire fragment. In Eq. (21), i and j are not residue indices, but atom indices. ω_m is the weight of the memory, $\gamma_{AM}(a_i,a_j)$ is a residue type dependent interaction strength, and r_{ij}^m is the distance between the i and j atoms in the memory structure. In the simplest case, as was used in this paper, both ω_m and $\gamma_{AM}(a_i,a_j)$ are 1.0 for all memories and all residue types. λ_{AM} is an overall scaling factor for the associative memory term, which can be used to adjust the weight of the term relative to others in the Hamiltonian, and $\sigma_{IJ} = |I-J|^{0.15}$ is a sequence separation dependent width.

$$V_{AM} = -\lambda_{AM} \sum_m \omega_m \sum_{ij} \gamma_{ij} \exp\left[-\frac{(r_{ij}-r_{ij}^m)^2}{2\sigma_{IJ}^2} \right] \tag{21}$$

V_{DSB} is a desolvation barrier potential. When pairs of residues are separated by a distance that is less than the width of a water, but they are not in direct contact, there is an energetic barrier that comes from the formation of a vacuum.[37] The form of the potential is given in Eq. (22) where r_{ij}

is the $C_\beta - C_\beta$ distance, except when a glycine is involved, in which case the C_α coordinates for the glycine are used.

$$\begin{aligned} V_{DSB} &= \lambda_{DSB} \sum_{j-i>9}^{N} \frac{1}{2} \Big(\tanh \left[\kappa_{DSB} (r_{ij} - r_{min}^{DSB}(a_i, a_j)) \right] + \\ &\qquad \tanh \left[\kappa_{DSB} (r_{max}^{DSB}(a_i, a_j) - r_{ij}) \right] \Big) \\ r_{min}^{DSB}(a_i, a_j) &= r_{min}^{0} + r_{shift}(a_i) + r_{shift}(a_j) \\ r_{max}^{DSB}(a_i, a_j) &= r_{max}^{0} + r_{shift}(a_i) + r_{shift}(a_j) \end{aligned} \tag{22}$$

Typical values for the parameters in Eq. (22) are given in Table S3. A sample plot of V_{DSB} interaction potential between two alanines is shown in Figure 3(d). As indicated, the minimum and maximum distances at which the desolvation barrier is activated, r_{min}^{DSB} and r_{max}^{DSB}, are residue type dependent. The details of r_{shift} are given in Table S7.

Table S7: $r_{shift}(a_i)$ values, in Å

a_i	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
$r_{shift}(a_i)$	0.00	2.04	0.57	0.57	0.36	1.11	1.17	−1.52	0.87	0.67
a_i	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
$r_{shift}(a_i)$	0.79	1.47	1.03	1.00	−0.10	0.26	0.37	1.21	1.15	0.39

Simulation protocol

We performed all molecular dynamics simulations using the Nose-Hoover thermostat as implemented in the open source simulation package LAMMPS. We recently extended LAMMPS by implementing all of the AWSEM potentials described in this supplement and adding a special atom style (called peptide), which is suitable for heteropolymeric systems such as proteins. All of our extensions to LAMMPS, as well as all analysis tools used for the current study, are available under the GNU General Public License at http://code.google.com/p/awsemmd/.

We started all structure prediction simulations from an extended conformation at a temperature well above the folding temperature. The simulations ran for 4×10^6 steps under non-periodic boundary conditions to a temperature well below the folding temperature. We used a timestep of 3 femtoseconds and saved the coordinates of the system every 1000 steps. For each saved snapshot, we calculated Q and RMSD values relative to an experimentally determined structure.

Table S8: Direct contact potential, V_{direct}, and Water potential, V_{water}, coefficients $\gamma^{dir}(a_i,a_j)$, $\gamma^{prot}(a_i,a_j)$, $\gamma^{wat}(a_i,a_j)$

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
ALA	ALA	0.72	0.09	0.02	ALA	ARG	-0.27	0.04	-0.00
ALA	ASN	-0.26	0.01	-0.00	ALA	ASP	-0.40	0.00	-0.07
ALA	CYS	0.62	0.27	0.29	ALA	GLN	-0.24	-0.02	-0.12
ALA	GLU	-0.35	0.02	-0.09	ALA	GLY	-0.11	0.05	-0.04
ALA	HIS	-0.13	0.03	-0.16	ALA	ILE	1.00	0.12	0.21
ALA	LEU	1.00	0.10	0.26	ALA	LYS	-0.45	0.02	0.08
ALA	MET	0.51	0.16	0.06	ALA	PHE	0.57	0.31	0.31
ALA	PRO	-0.53	-0.00	-0.00	ALA	SER	-0.21	-0.00	0.04
ALA	THR	0.08	0.05	0.03	ALA	TRP	0.40	0.09	-0.08
ALA	TYR	0.11	0.19	0.14	ALA	VAL	0.92	0.33	0.25
ARG	ARG	-0.64	-0.05	0.62	ARG	ASN	-0.28	-0.05	0.64
ARG	ASP	0.41	0.02	1.00	ARG	CYS	-0.40	0.43	0.46
ARG	GLN	-0.21	-0.04	0.43	ARG	GLU	-0.03	-0.03	0.97
ARG	GLY	-0.33	-0.01	0.32	ARG	HIS	-0.53	-0.06	0.32
ARG	ILE	-0.14	-0.04	0.07	ARG	LEU	-0.25	-0.07	-0.04
ARG	LYS	-0.96	-0.08	0.47	ARG	MET	-0.02	-0.16	0.14
ARG	PHE	-0.18	-0.13	-0.11	ARG	PRO	-0.82	0.01	0.43
ARG	SER	-0.33	0.01	0.32	ARG	THR	-0.23	-0.01	0.35
ARG	TRP	-0.30	-0.20	-0.05	ARG	TYR	0.14	0.15	-0.47
ARG	VAL	-0.17	0.01	0.11	ASN	ASN	0.16	-0.03	0.58
ASN	ASP	0.02	-0.01	0.28	ASN	CYS	-0.09	0.16	0.17
ASN	GLN	-0.19	-0.02	0.39	ASN	GLU	-0.56	-0.03	0.27
ASN	GLY	-0.14	0.01	0.10	ASN	HIS	-0.07	0.00	0.13
ASN	ILE	-0.72	-0.22	0.24	ASN	LEU	-0.58	-0.13	0.19
ASN	LYS	-0.45	-0.05	0.44	ASN	MET	-0.60	-0.10	-0.10
ASN	PHE	-0.52	-0.11	0.10	ASN	PRO	-0.69	-0.01	0.57
ASN	SER	-0.02	0.00	0.31	ASN	THR	-0.31	-0.02	0.30
ASN	TRP	-0.37	0.08	-0.30	ASN	TYR	-0.27	0.14	-0.45
ASN	VAL	-0.59	-0.11	-0.00	ASP	ASP	-0.57	0.00	0.23
ASP	CYS	-0.37	-0.24	0.52	ASP	GLN	-0.39	-0.03	0.31
ASP	GLU	-0.85	-0.04	0.20	ASP	GLY	-0.30	-0.02	0.25
ASP	HIS	-0.08	0.01	0.61	ASP	ILE	-0.72	-0.18	0.27
ASP	LEU	-0.78	-0.20	0.24	ASP	LYS	0.11	-0.03	0.84
ASP	MET	-0.58	-0.18	-0.02	ASP	PHE	-0.76	-0.19	0.00
ASP	PRO	-0.82	-0.02	0.48	ASP	SER	-0.03	-0.00	0.09

Table S8 – continue

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
ASP	THR	-0.22	-0.01	0.18	ASP	TRP	-0.74	-0.13	-0.14
ASP	TYR	-0.78	0.05	-0.43	ASP	VAL	-0.74	-0.15	0.18
CYS	CYS	0.98	0.39	0.64	CYS	GLN	-0.43	0.16	0.66
CYS	GLU	-0.36	0.15	-0.15	CYS	GLY	0.43	0.39	-0.08
CYS	HIS	0.69	0.03	-0.04	CYS	ILE	0.70	0.33	0.91
CYS	LEU	0.98	0.31	0.25	CYS	LYS	-0.58	-0.01	0.30
CYS	MET	0.30	0.73	-0.52	CYS	PHE	0.85	0.88	0.77
CYS	PRO	0.09	0.39	0.02	CYS	SER	0.47	0.52	0.15
CYS	THR	-0.18	0.34	-0.11	CYS	TRP	0.10	0.58	1.00
CYS	TYR	0.87	0.52	0.42	CYS	VAL	0.95	0.62	0.00
GLN	GLN	-0.29	0.03	0.32	GLN	GLU	-0.49	-0.04	0.59
GLN	GLY	-0.37	0.01	0.11	GLN	HIS	-0.72	0.04	0.57
GLN	ILE	-0.43	-0.09	0.11	GLN	LEU	-0.29	-0.13	0.02
GLN	LYS	-0.49	-0.07	0.44	GLN	MET	-0.33	-0.13	-0.07
GLN	PHE	-0.35	0.04	-0.08	GLN	PRO	-0.60	0.01	0.46
GLN	SER	-0.34	-0.02	0.33	GLN	THR	-0.03	-0.03	0.37
GLN	TRP	-0.56	-0.06	-0.27	GLN	TYR	-0.21	-0.10	-0.69
GLN	VAL	-0.28	0.09	-0.02	GLU	GLU	-0.86	-0.04	0.38
GLU	GLY	-0.55	-0.01	0.09	GLU	HIS	-0.50	-0.05	0.40
GLU	ILE	-0.49	-0.11	0.22	GLU	LEU	-0.56	-0.26	0.13
GLU	LYS	0.13	-0.03	1.00	GLU	MET	-0.77	-0.23	0.22
GLU	PHE	-0.75	-0.16	-0.07	GLU	PRO	-0.78	-0.02	0.48
GLU	SER	-0.31	-0.01	0.18	GLU	THR	0.05	-0.01	0.14
GLU	TRP	-0.46	0.00	-0.29	GLU	TYR	-0.32	-0.04	-0.47
GLU	VAL	-0.38	-0.12	0.14	GLY	GLY	0.37	0.09	-0.08
GLY	HIS	-0.42	-0.03	0.29	GLY	ILE	0.04	-0.05	0.17
GLY	LEU	-0.22	-0.05	0.17	GLY	LYS	-0.48	-0.03	0.27
GLY	MET	0.13	0.21	0.05	GLY	PHE	-0.05	-0.08	0.32
GLY	PRO	-0.42	0.06	0.37	GLY	SER	0.02	0.03	0.14
GLY	THR	-0.14	0.02	0.18	GLY	TRP	0.04	-0.03	0.13
GLY	TYR	0.15	0.08	0.00	GLY	VAL	-0.11	0.05	0.20
HIS	HIS	-0.16	0.11	0.76	HIS	ILE	-0.30	-0.00	0.37
HIS	LEU	0.08	-0.00	-0.00	HIS	LYS	-0.55	-0.10	0.63
HIS	MET	0.20	0.09	-0.12	HIS	PHE	0.37	0.39	-0.11
HIS	PRO	-0.60	0.03	0.53	HIS	SER	-0.03	0.05	0.13
HIS	THR	-0.09	0.02	0.41	HIS	TRP	-0.01	0.48	-0.29
HIS	TYR	0.26	0.35	-0.28	HIS	VAL	0.16	0.03	0.03

Table S8 – continue

a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}	a_i	a_j	γ^{dir}	γ^{prot}	γ^{wat}
ILE	ILE	0.98	1.00	1.00	ILE	LEU	0.98	1.00	0.38
ILE	LYS	-0.71	-0.09	0.20	ILE	MET	0.74	0.72	0.74
ILE	PHE	0.88	0.93	0.35	ILE	PRO	-0.43	-0.19	0.27
ILE	SER	-0.43	0.02	0.31	ILE	THR	-0.02	0.11	0.24
ILE	TRP	0.82	0.34	0.37	ILE	TYR	0.90	0.23	0.37
ILE	VAL	0.98	0.69	0.77	LEU	LEU	0.98	1.00	0.37
LEU	LYS	-0.66	-0.07	0.07	LEU	MET	0.85	0.74	0.27
LEU	PHE	0.79	0.70	0.25	LEU	PRO	-0.54	-0.15	0.11
LEU	SER	-0.34	-0.13	0.29	LEU	THR	0.01	0.13	0.26
LEU	TRP	0.98	0.38	0.76	LEU	TYR	0.69	0.35	0.32
LEU	VAL	0.98	0.64	0.43	LYS	LYS	-0.97	-0.06	0.42
LYS	MET	-0.70	-0.14	0.06	LYS	PHE	-0.66	-0.17	-0.26
LYS	PRO	-1.00	-0.03	0.55	LYS	SER	-0.62	-0.03	0.33
LYS	THR	-0.55	-0.03	0.47	LYS	TRP	-0.40	-0.21	-0.62
LYS	TYR	-0.18	-0.29	-0.58	LYS	VAL	-0.62	-0.13	0.03
MET	MET	0.52	0.32	-1.00	MET	PHE	0.69	0.72	0.30
MET	PRO	-0.50	0.01	0.13	MET	SER	-0.33	0.07	-0.03
MET	THR	-0.09	0.06	0.22	MET	TRP	0.12	0.50	-0.85
MET	TYR	0.64	0.27	-0.14	MET	VAL	0.63	0.40	0.62
PHE	PHE	0.98	1.00	0.52	PHE	PRO	-0.22	-0.21	0.26
PHE	SER	-0.27	0.01	0.13	PHE	THR	-0.16	0.12	0.16
PHE	TRP	0.67	0.66	0.54	PHE	TYR	0.62	0.27	-0.11
PHE	VAL	0.78	0.83	0.20	PRO	PRO	-0.51	-0.01	0.33
PRO	SER	-0.56	-0.00	0.52	PRO	THR	-0.47	-0.01	0.07
PRO	TRP	0.01	0.47	-0.56	PRO	TYR	0.06	-0.07	-0.34
PRO	VAL	-0.33	-0.10	0.21	SER	SER	-0.10	0.02	0.23
SER	THR	-0.10	-0.01	0.19	SER	TRP	-0.32	0.11	0.05
SER	TYR	-0.30	-0.03	-0.09	SER	VAL	-0.25	-0.00	0.10
THR	THR	0.16	-0.01	0.37	THR	TRP	-0.44	0.13	-0.13
THR	TYR	-0.22	-0.06	-0.37	THR	VAL	0.18	-0.10	0.19
TRP	TRP	0.07	0.43	-1.00	TRP	TYR	0.21	0.15	-0.95
TRP	VAL	0.52	0.44	1.00	TYR	TYR	0.55	0.21	-0.45
TYR	VAL	0.62	0.59	0.38	VAL	VAL	0.98	0.73	0.87

References

[1] Sasai, M.; Wolynes, P. *Phys. Rev. A* **1992**, *46*, 7979–7997.

[2] Ferreiro, D.; Hegler, J.; Komives, E.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19819–19824.

[3] Sasai, M.; Wolynes, P. *Phys. Rev. Lett.* **1990**, *65*, 2740–2743.

[4] Eastwood, M.; Wolynes, P. *J. Chem. Phys.* **2001**, *114*, 4702–4716.

[5] Papoian, G.; Ulander, J.; Wolynes, P. *J. Am. Chem. Soc.* **2003**, *125*, 9170–9178.

[6] Papoian, G.; Ulander, J.; Eastwood, M.; Luthey-Schulten, Z.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3352–3357.

[7] Zong, C.; Papoian, G.; Ulander, J.; Wolynes, P. *J. Am. Chem. Soc.* **2006**, *128*, 5168–5176.

[8] Latzer, J.; Shen, T.; Wolynes, P. *Biochemistry* **2008**, *47*, 2110–2122.

[9] Weinkam, P.; Pletneva, E.; Gray, H.; Winkler, J.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 1796–1801.

[10] Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. *Proteins: Structure, Function, and Bioinformatics* **1999**, *34*, 281–294.

[11] Latzer, J.; Eastwood, M.; Wolynes, P. *J. Chem. Phys.* **2006**, *125*, 214905–1–214905–12.

[12] Hegler, J.; Weinkam, P.; Wolynes, P. *HFSP J.* **2008**, *2*, 307–313.

[13] Ferreiro, D.; Hegler, J.; Komives, E.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3499–3503.

[14] Papoian, G.; Wolynes, P. *Biopolymers* **2003**, *68*, 333–349.

[15] Weinkam, P.; Zong, C.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12401–12406.

[16] Sutto, L.; Latzer, J.; Hegler, J.; Ferreiro, D.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19825–19830.

[17] Weinkam, P.; Romesberg, F.; Wolynes, P. *Biochemistry* **2009**, *48*, 2394–2402.

[18] Weinkam, P.; Zimmermann, J.; Romesberg, F.; Wolynes, P. *Acc. Chem. Res.* **2010**, *43*, 652–660.

[19] Friedrichs, M.; Wolynes, P. *Science* **1989**, *246*, 371–373.

[20] Friedrichs, M.; Wolynes, P. *Tetrahedron Comput. Methodol.* **1990**, *3*, 175–190.

[21] Friedrichs, M.; Goldstein, R.; Wolynes, P. *J. Mol. Biol.* **1991**, *222*, 1013–1034.

[22] Goldstein, R.; Luthey-Schulten, Z.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 9029–9033.

[23] Koretke, K.; Luthey-Schulten, Z.; Wolynes, P. *Protein Sci.* **1996**, *5*, 1043–1059.

[24] Koretke, K.; Luthey-Schulten, Z.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 2932–2937.

[25] Eastwood, M.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. *IBM J. Res. Dev.* **2001**, *45*, 475–497.

[26] Hardin, C.; Eastwood, M.; Luthey-Schulten, Z.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 14235–14240.

[27] Eastwood, M.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. *J. Chem. Phys.* **2002**, *117*, 4602–4615.

[28] Hardin, C.; Eastwood, M.; Prentiss, M.; Luthey-Schulten, Z.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1679–1684.

[29] Eastwood, M.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. *J. Chem. Phys.* **2003**, *118*, 8500–8512.

[30] Prentiss, M.; Hardin, C.; Eastwood, M.; Zong, C.; Wolynes, P. *J. Chem. Theory Comput.* **2006**, *2*, 705–716.

[31] Kwac, K.; Wolynes, P. *Bull. Korean Chem. Soc.* **2008**, *29*, 2172–2182.

[32] Hegler, J.; Lätzer, J.; Shehu, A.; Clementi, C.; Wolynes, P. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15302–15307.

[33] Oklejas, V.; Zong, C.; Papoian, G.; Wolynes, P. *Methods* **2010**, *52*, 84–90.

[34] Cuff, A. J.; Clamp, E. M.; Siddiqui, S. A.; Finlay, M.; Barton, G. J. *Bioinformatics* **1998**, *14*, 892–893.

[35] Oklejas, V.; Zong, C.; Papoian, G. A.; Wolynes, P. G. *Methods* **2010**, *52*, 84–90.

[36] Altschul, S.; Madden, T.; Schäffer, A.; Zhang, J.; Zhang, A.; Miller, W.; Lipman, D. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

[37] Hummer, G.; Garde, S.; García, A.; Paulaitis, M.; Pratt, L. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 1552–1555.