An effective solvent theory connecting the underlying mechanisms of osmolytes and denaturants for protein stability (Supplementary material)

Apichart Linhananta, *,† Shirin Hadizadeh,* Steven S. Plotkin,*

*Department of Physics and Astronomy, The University of British Columbia, 6224 Agricultural Road, Vancouver, B.C. V6T 1Z1, Canada;
†Permanent Address: Physics Department, Lakehead University, 955 Oliver Road, Thunder Bay, O.N., P7B 5E1, Canada

Address reprint requests and inquiries to Steven S. Plotkin steve@phas.ubc.ca, or Apichart Linhananta apichart.linhananta@lakeheadu.ca

Supplementary material for:

An effective solvent theory connecting the underlying mechanisms of osmolytes and denaturants for protein stability

Apichart Linhananta[1,2*], *Shirin Hadizadeh*[1], *Steven Samuel Plotkin*[1*]

[1]Department of Physics and Astronomy, The University of British Columbia, 6224 Agricultural Road, Vancouver, B.C. V6T 1Z1, Canada

[2]Permanent Address: Physics Department, Lakehead University, 955 Oliver Road, Thunder Bay, O.N., P7B 5E1, Canada

E-mail: steve@physics.ubc.ca

E-mail: apichart.linhananta@lakeheadu.ca

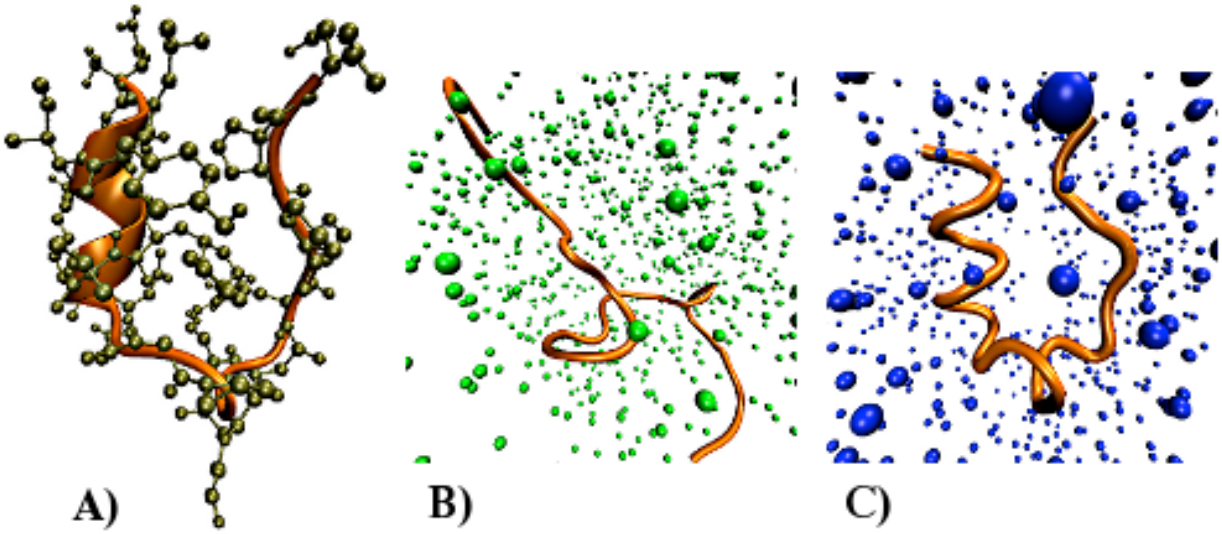

Figure 1: A) Ground-state native structure of the all-atom Gō model of the Trp-cage protein; B) Denatured (unfolded) Trp-cage in denaturing solution; C) Folded Trp-cage in osmolyte solution.

1. Simulation model and method

A brief description of the discrete molecular dynamics (DMD) simulation method is presented here; more details are contained in references (1-7) below. The initial heavy-atom positions were obtained from the NMR structure (structure 1 of PDB 1L2Y(8)) and the missing polar hydrogen molecules are constructed as in reference(2). The resulting structure is comprised of 189 heavy atoms and polar hydrogen atoms (non-polar hydrogen atoms are not represented), which in the discontinuous model are

represented as hard spheres. Two bonded atoms *i* and *j*, as well any 1,3 angle-constrained pair and 1,4 aromatic pair, are constrained to be within $\pm 10\%$ of their distance in the NMR structure by an infinite square-well potential:

$$u_{ij}^{bond} = \begin{cases} \infty, & r < 0.9\sigma_{ij} \\ 0, & 0.9\sigma_{ij} < r < 1.1\sigma_{ij} \\ \infty, & r > 1.1\sigma_{ij}, \end{cases} \quad (1.1)$$

where σ_{ij} is the separation distance of the bonded *i,j* pair in the NMR structure. The model also includes a discontinuous improper dihedral potential, to maintain chirality about tetrahedral heavy atoms (e.g. $C\alpha$, $C\beta$, C, N), and planar moieties such as tryptophan rings.

$$u_{ij}^{improp} = \begin{cases} \infty, & \omega < \omega_0 - 20^\circ \\ 0, & \omega_0 - 20^\circ < \omega < \omega_0 + 20^\circ \\ \infty, & \omega > \omega_0 + 20^\circ. \end{cases} \quad (1.2)$$

Here ω represents the dihedral angles of the constrained atoms, which are restricted to values near $\omega_0 = 35.26439^\circ$ for tetrahedral heavy atoms, and $\omega_0 = 0^\circ$ for planar atoms.

As well, two non-bonded atoms *i,j* may interact by square-well potential with a hard-core radius:

$$u_{ij}^{non-bond} = \begin{cases} \infty, & r < 0.8\sigma_{ij}^{vdW} \\ \varepsilon_{pp} = B_{ij}\varepsilon_{Go}, & 0.8\sigma_{ij}^{vdW} < r < 1.3\sigma_{ij}^{vdW} \\ 0, & r > 1.2\sigma_{ij}^{vdW}, \end{cases} \quad (1.3)$$

where σ_{ij}^{vdW} is the sum of the van der Waals (vdW) radii $r_i + r_j$ for each atom, as given by the CHARMM potential set 19 (9) and B_{ij} and ε_{Go} are Gō interaction strength parameters giving the depth of the square well potential, which may depend on the identities of atoms i and j. Using these parameters we performed a short DMD simulation with fixed native contacts to remove interactions violating equations (1.1)-(1.3) from the initial experimental structure, to produce an equilibrium structure of the model Trp-cage consistent with the above potentials. The ground state native structure is shown in Figure 1A. The Gō model potential (10) is implemented by setting the non-bonded square-well depth ε_{pp} to $-\varepsilon_{Go}$ ($B_{ij} = -1$) for all non-bonded *ij* pairs in the equilibrated structure with van der Waals overlap $r < 1.2\sigma_{ij}^{vdW}$. This gives 1267 non-bonded atomic contacts in the native state. For all other non-bonded *ij* pairs, the square-well depth is set to 0 ($B_{ij} = 0$), so that these atom pairs are purely repulsive. As in previous DMD studies(2, 11), the energy scale is set by the Gō contact energy; thus simulations are

performed with the Gō contact energy ε_{pp}^{*} set to $-\varepsilon_{Go}^{*}=-1$, and all energies and temperatures are scaled in units of ε_{Go} ($E^{*}=E/\varepsilon_{Go}$ and $T^{*}=k_{B}T/\varepsilon_{Go}$). A reduced time unit $t^{*}=t\sqrt{\varepsilon_{Go}/m\sigma_{L}^{2}}$ is also used (m can be taken to be the average atomic mass of the atoms comprising the protein and σ_L = 1 Å).

The Gō model protein in explicit solvent is implemented by placing the Trp-cage protein in a 40Å X 40Å X 40Å box, along with a variable number of spherical solvent molecules randomly inserted without hardcore overlaps. A typical simulation consisted of 1000 spherical solvent particles of radius 1.5Å (the approximate radius of water). This is about half the number of water molecules in a 55M solution for a $(40\text{Å})^3$ box. We employ such a dilute concentration for computational convenience; physical concentrations have collision times sufficiently short as to make such simulations prohibitively slow. Diluting the concentration weakens the effects that would be observed by varying solvent qualities from those at 55M, i.e. this simplification effectively places lower bounds on any trends that we predict would be observed. For this reason we find the approximation acceptable as it only strengthens the conclusions of this study. Standard periodic boundary conditions are implemented.

Solvent molecules interact with both protein moieties and with each other by a square-well potential plus hard core radii, having the form:

$$\frac{u_{ij}^{x-s}}{\varepsilon_{Go}}=\begin{cases}\infty, & r<0.8\sigma_{ij}^{xs}\\ \varepsilon_{xs}/\varepsilon_{Go}=\varepsilon_{xs}^{*}, & 0.8\sigma_{ij}^{xs}<r<1.2\sigma_{ij}^{xs}\\ 0, & r>1.2\sigma_{ij}^{xs}.\end{cases} \tag{1.4}$$

In equation (1.4) x may be either a protein atom p or another solvent residue s. For solvent-solvent interactions, σ_{ij}^{ss} is the vdW diameter of the solvent, which we generally set to $\sigma_{ij}^{ss}=3.0$ Å: roughly the size of a water molecule. ε_{ss}^{*} is the solvent-solvent square-well depth in units of the Gō contact energy ε_{Go}. For solvent-protein interactions, $\sigma_{ij}^{ps}=\left(\sigma_{i}^{vdW}+\sigma_{j}^{water}\right)/2$ is the average vdW diameter of the protein-solvent (*i,j*) pair, where σ_{i}^{vdW} is the CHARMM potential set 19 vdW diameter of the i^{th} atom of the protein, and $\sigma_{j}^{water}=3.0$ Å is the vdW diameter of the j^{th} water molecule. ε_{ps}^{*} is the protein-solvent square-well depth in units of the Gō contact energy ε_{Go}. A plot of this potential for several values of protein-solvent interaction energy is shown in Figure 2 below.

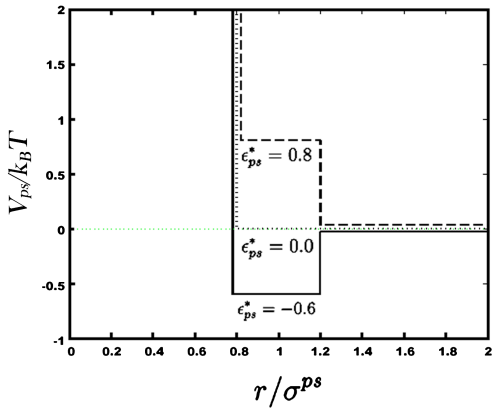

Figure 2: Illustration of the functional form of the potential between solvent-residues and protein atoms given by equation (1.4), for several values of interaction energy ε_{ps}. The protein Gō potential in equation (1.3) has the same functional form with $\varepsilon_{pp} = -1$ for native interactions, $\varepsilon_{pp} = 0$ otherwise.

The DMD methodology benefits from substantial computational speedup over the traditional molecular dynamics that would be required for other solvent models such as TIP3P; the simple functional form of the potential eases the task of evaluating the energy for each time step, and also allows for the fact that configurational updates need only be performed after each collision. On the other hand, features such as the specific steric geometry of various solvents and osmolytes, long-range electrostatics, and angle-dependent hydrogen bonding, are not captured by the present model. More specific stereochemistry as well as angle-dependent hydrogen bonding could be included in more refined DMD solvent models.

1B. Free Energy, Energy, and Entropy Functionals

Free energy surfaces have long been used as a tool to analyze protein folding thermodynamics and kinetics(12, 13). The free energy calculations used here are based on standard multiple-histogram method (14, 15), which calculates thermodynamics quantities by approximating the density of states $g(\varepsilon)$ (i.e. the number of states with energy ε) from simulation data. This gives the partition function $Z(x) = \sum^{*} g(\varepsilon, x, \Delta x)\exp(-\beta\varepsilon)$, where * indicates a summation over all energy states with the order parameter constrained to values ranging from x to $x + \Delta x$, and $g(\varepsilon, x, \Delta x)$ is the density of state of energy ε with the order parameter range x to $x + \Delta x$. Thermodynamic quantities only depend on the bin size Δx by an additive constant, so long as Δx is sufficiently small in the traditional coarse-graining sense. The free energy $F(x) = -k_B T \ln Z(x) = -k_B T \ln P(x) + F$, where $P(x)$ is the probability the system is within the order parameter range x to $x + \Delta x$, and $F = -k_B T \ln Z$ is the total free energy. The internal energy is $U(x) = \langle E(x) \rangle = \sum \varepsilon g(\varepsilon, x, \Delta x)\exp(-\beta\varepsilon)/Z(x)$, and the entropy $S(x) = (U(x) - F(x))/T$.

In this work we use the non-local native fraction of contacts Q as an order parameter, defined by first counting all atom pairs in the native structure that are within 1.2 times the sum of their hard-core radii, and between residues i, j such that $|i-j|>3$. This gives 276 contacts in the native state. The value of Q in a given (partly folded) configuration is the fraction of these contacts present, and varies from 0 (completely unfolded) to 1 (completely folded). A bin size $\Delta Q = 0.02$ was used. Unless stated otherwise, the energy of the system includes intra-protein, protein-solvent, and solvent-solvent energy. The computed energy $u^*(Q)$, free energy $f^*(Q)$, and entropy $s^*(Q)$ functions are in units of the Gō contact energy ε_{GO}.

2. Phase diagram in the temperature, protein-solvent interaction energy plane

The phase diagram in Figure 3 below illustrates that our protein in explicit solvent model reproduces protein stabilization/destabilization by osmolyte/denaturant. The native state becomes more stable as the quality index ε_{ps}^* increases (i.e. become more osmolytic).

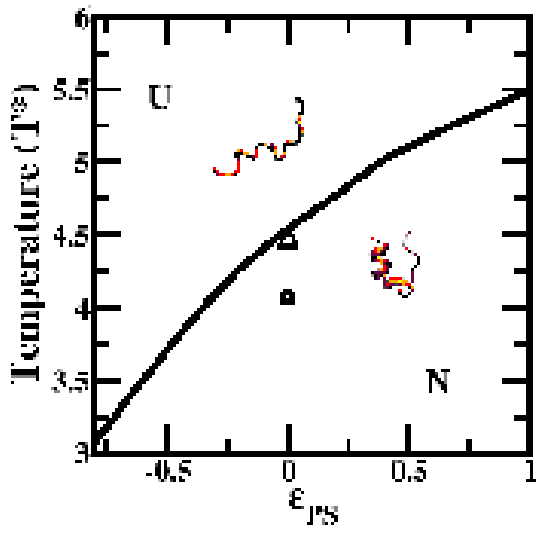

Figure 3. Solvent quality temperature (T^*) vs. solvent quality (ε_{ps}^*) phase diagram. N denotes region where the native state is stable, and U, region where the unfolded state is stable. The phase boundary (solid line) is determined by folding temperature T_f^* obtained from histogram analysis of the simulation data (figure 2A main text). The two symbols in the figure indicate 2 systems with different solvent conditions: O, location in the phase diagram of the Gō model in implicit solvent; and Δ, location in the phase diagram of the Go model in hard sphere solvent ($\varepsilon_{ps}^* = \varepsilon_{ss}^* = 0$).

3. Comments on the specific heat $C^*(T^*)$ and folding cooperativity

The specific heat $C^*\left(T^*\right)$ (in reduced units described in the text) is obtained from $\left\langle \left(E^* - \left\langle E^* \right\rangle\right)^2 \right\rangle \Big/ T^{*2}$. The energy here is generally the full system energy: protein-protein interaction energy plus protein-solvent interaction energy plus solvent-solvent interaction energy. For implicit solvent and hard-sphere solvent, this is equivalent to the protein-protein interaction energy only. The protein-protein interaction energy corresponds to all 1267 atom-atom native non-bonded contacts: *276* of these are non-local (between residues *i* and *j* with $|i-j| \geq 4$) and show significantly more change upon folding than the *991* local contacts. Scanning the temperature results in a single peak of $C^*\left(T^*\right)$, corresponding to a first-order-like folding transition. The high temperature baseline is generally observed to be lower than the low temperature baseline, a feature that is specific to discontinuous potentials and has been observed previously by many authors (1, 6, 7, 11, 16-18). This phenomenon is likely due to in part to the reduced energetic fluctuations in the unfolded state as compared to continuum models. For example the heat capacities of the angle and dihedral potentials, having no energy scale, are temperature-independent; moreover the repulsive potentials between self-avoiding particles no longer have temperature-dependent energetic fluctuations on the $\sim r^{-12}$ part of the LJ potential, but have no energetic fluctuations on the hard-wall potential. This phenomenon is generally seen in the temperature-dependence of energetic fluctuations in models involving square-well interactions. For example, square-well fluids, with pair potentials having the form of equation (1.3), have a heat capacity which is a decreasing function of temperature (19-22), and as well show a heat capacity drop across the liquid-gas transition (19), consistent with the liquid gas transition observed in clusters and homopolymers with square-well interactions (23).

We compute the cooperativity of the folding transition in our model from the ratio of the van't Hoff to calorimetric enthalpy, employing the method of Kaya and Chan (24) which gives this ratio as $2T_{\max}\sqrt{k_B C_p\left(T_{\max}\right)}\Big/\Delta H_{cal}$. The protein's internal energy is a nearly linear function of Q in the Gō model (see figure 5B in the text). We use internal energy as a way of accounting for baseline subtraction by removing bath heat capacity. Moreover a scan of the heat capacity, as shown in Figure 4 below, and calculated from fluctuations in Q as $C^*\left(T^*\right) = A\left\langle \left(Q - \left\langle Q \right\rangle\right)^2 \right\rangle \Big/ T^{*2}$ with A a proportionality constant, gives approximately the same heat capacity as the usual enthalpic definition (Figure 4). The quantity Q accounts only for protein-protein interactions, while the internal energy in Figure 4 contains both protein-protein and protein-solvent interactions. This accounts in part for the discrepancy between the two curves.

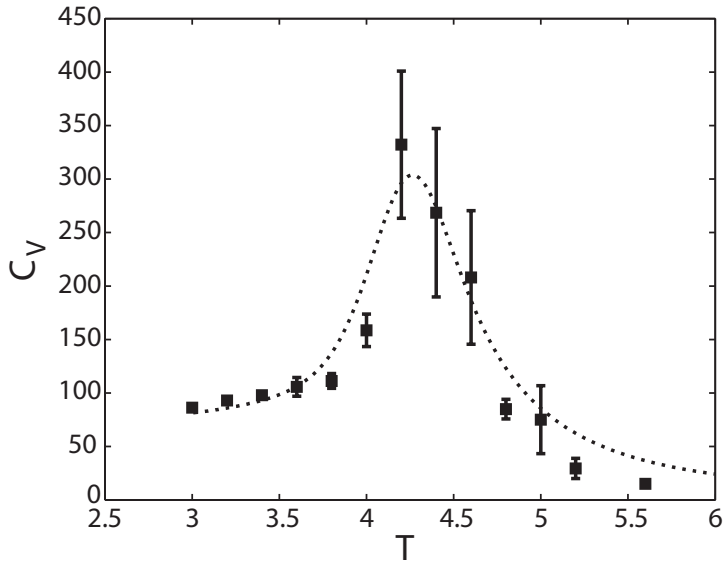

Figure 4: Internal heat capacity scan using either the fluctuations in Q, or fluctuations in internal energy of protein-protein plus protein-solvent interactions, for the system with $\varepsilon^* = -0.2$. One can see generally good agreement between the two methods. The poorer statistics for Q is simply the result of less structural data being recorded compared to the energetic data.

With Q as an energetic proxy, the above van't Hoff ratio $2T_{\max}\sqrt{k_B C_p(T_{\max})}/\Delta H_{cal}$ becomes $2\delta Q/(Q_F - Q_U)$, i.e. twice the standard deviation of Q at the peak of the heat capacity, divided by the difference in Q between the folded and unfolded states, below and above the transition. Thermodynamic perturbation theory was used when needed to obtain accurate values of the temperature at the heat capacity peak. As mentioned in the text, the calorimetric Q values were obtained from thermal averages well above and below the transition midpoint for each value of solvent-protein interaction energy ε^*_{ps}, and these values did not vary significantly as ε^*_{ps} was varied. Thermodynamic perturbation theory was again employed as needed to obtain low and high temperature ensembles (though temperature differences were always less than 15%). We did not apply base-line subtraction to the heat capacity curves (24), which would result in an overall increase in cooperativity, but would not change the trend in figure 4. Statistical error bars were obtained from the standard deviations of the respective quantities (R_{GY} or $\Delta H_{vH}/\Delta H_{cal}$) between non-overlapping data sets.

4. Energy, entropy, and free energy surfaces

The energy and entropy surfaces can be plotted for the protein-protein and protein-solvent interactions energy alone, yielding what the effective energies and entropies would be in an implicit solvent model. These results, shown in Figure 5 below, corroborate the conclusions drawn from the total system energy and entropy in Figure 6 in the main text.

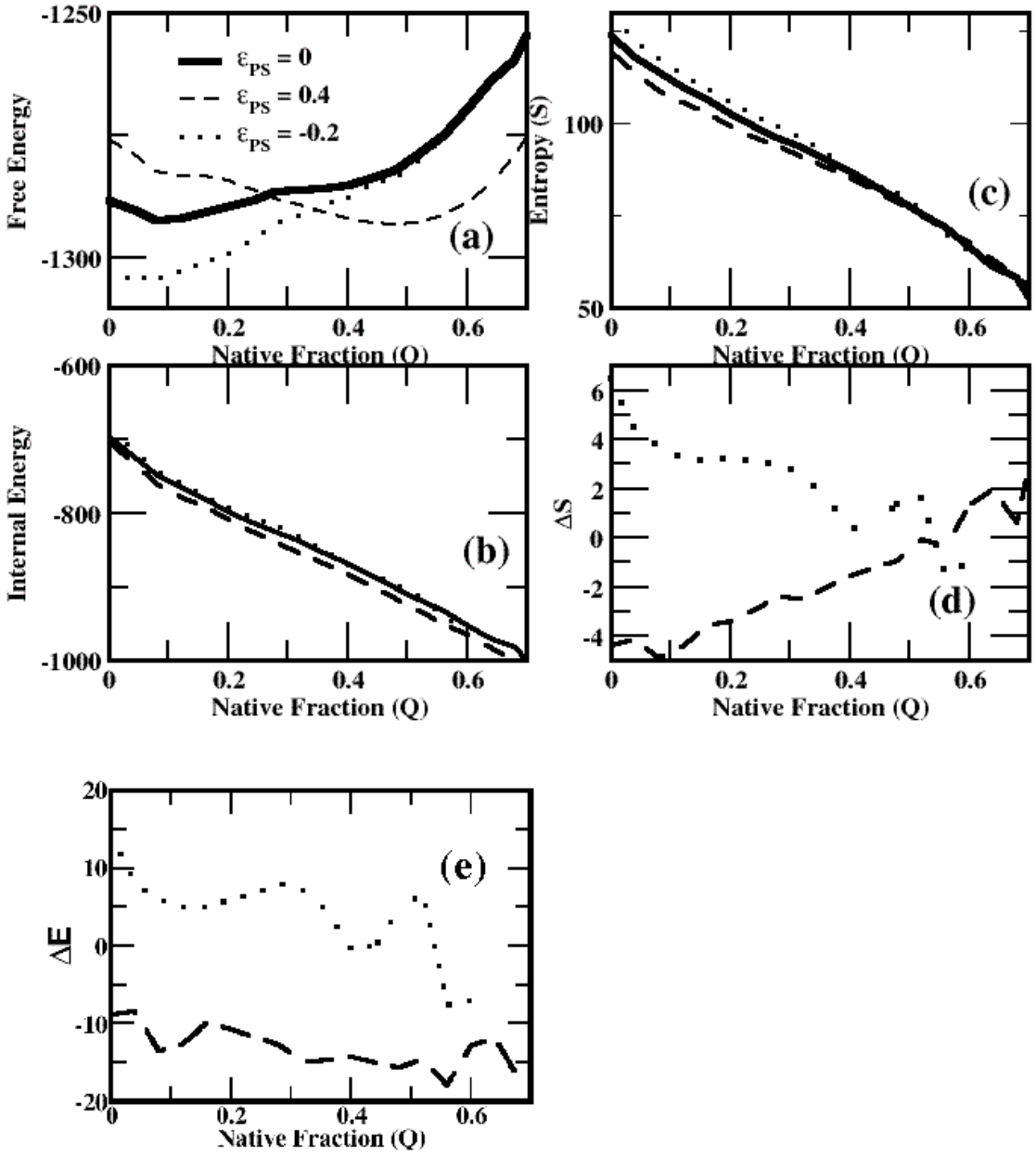

Figure 5: Energy and entropy surfaces for protein-protein + protein-solvent interactions, as a measure of the internal energy and entropy of the effective system. (a) Free energy vs. protein native fraction (Q) at $T^* = 4.8$ for $\varepsilon_{ps}^* = 0$ (solid line), $\varepsilon_{ps}^* = 0.4$ (dashed line), and $\varepsilon_{ps}^* = -0.2$ (dotted line), with solvent-solvent contact energy fixed at $\varepsilon_{ss}^* = -1$. (b) Protein internal energy $\langle E \rangle$ vs. Q. (c) Entropy(S) vs. Q. (d) Change in entropy (ΔS) vs. Q, for osmolyte solvent ($\varepsilon_{ps}^* = 0.4$) compared to neutral solvent ($\varepsilon_{ps}^* = 0$), i.e. $\Delta S = S\left(\varepsilon_{ps}^* = 0.4\right) - S\left(\varepsilon_{ps}^* = 0\right)$ (dashed line), and for denaturant ($\varepsilon_{ps}^* = -0.2$) solvent compare to neutral ($\varepsilon_{ps}^* = 0$) solvent $\Delta S = S\left(\varepsilon_{ps}^* = -0.2\right) - S\left(\varepsilon_{ps}^* = 0\right)$. One can see almost no change in the internal entropy for the folded state, but significant changes in the unfolded ensemble. (e).

Changes in enthalpy between osmolyte and neutral solvent $\Delta E = E\left(\varepsilon_{ps}^{*} = 0.4\right) - E\left(\varepsilon_{ps}^{*} = 0\right)$ **(dashed line), and between denaturant solvent and neutral solvent** $\Delta E = E\left(\varepsilon_{ps}^{*} = -0.2\right) - E\left(\varepsilon_{ps}^{*} = 0\right)$ **(dotted line).**

5. The correspondence between explicit and implicit osmolyte models

As summarized in section 2C in the main text, a particular effective solvent system, characterized by three contact energies $\varepsilon_{pp}^{*}, \varepsilon_{ps}^{*}, \varepsilon_{ss}^{*}$, can be shown to be equivalent to a given solution with explicit solute, characterized by six parameters $\varepsilon_{pp}, \varepsilon_{ps}, \varepsilon_{op}, \varepsilon_{oo}, \varepsilon_{os}, \varepsilon_{ss}$. The above 6 interaction enthalpies, corresponding to pair interactions between *p*, *o*, and *s*, along with the temperature T, are the relevant energy scales in the problem. There are also 3 length scales in the problem corresponding to the sizes of the various species: r_p, r_s, r_o corresponding to the monomer, solvent, and osmolyte radius. The effects of the corresponding radii of the various constituents in the model, r_p, r_s, r_o, are manifested simply as different coordination numbers q_p, q_s, q_o for each of the species.

Given the six explicit-system (ES) parameters $\varepsilon_{pp}, \varepsilon_{ps}, \varepsilon_{op}, \varepsilon_{oo}, \varepsilon_{os}, \varepsilon_{ss}$, we seek the three implicit-system (IS) parameters $\varepsilon_{pp}^{*}, \varepsilon_{ps}^{*}, \varepsilon_{ss}^{*}$ that would give the same average interaction probabilities for the system. The mean number of interactions N_{ij} between species i and j for the explicit system can be found from the sum rules for the total number of nearest neighbours:

$$\begin{aligned} q_p N_p &= 2N_{pp} + N_{ps} + N_{op} \\ q_s N_s &= 2N_{ss} + N_{ps} + N_{os} \\ q_o N_o &= 2N_{oo} + N_{op} + N_{os} \end{aligned} \tag{1.5}$$

and the 3 quasichemical equations of mass balance for the reactions $po + so \rightleftarrows ps + oo$, $so + so \rightleftarrows ss + oo$, $po + po \rightleftarrows pp + oo$, which give equations of the form:

$$\frac{N_{ps} N_{oo}}{N_{po} N_{so}} = e^{-\beta\left(\varepsilon_{ps} + \varepsilon_{oo} - \varepsilon_{po} - \varepsilon_{so}\right)} \tag{1.6}$$

There are 3 independent rate equations, giving along with (1.5) a total of 6 equations for the 6 N_{ij}. Several approximations are made in the above heuristic approach. The different sizes of the particles are only manifested in the different coordination numbers in (1.5). More accurate models would modify the sum rules to more accurately account for the geometry of nearest neighbours, which would also result in modified stoichiometric coefficients in the mass balance equations. The parameter ε_{pp} should be interpreted as a chemical potential, because the conformational entropy for protein monomers is smaller than that for free osmolyte or solvent particles.

To map to the implicit system, first let the IS particle numbers be given by: $N_p^* = N_p$, and $N_s^* = N_o + N_s$. Adding the solvent and solute equations in (1.5) gives a relation between the coordination numbers and solvent particles which can be mapped to the implicit system:

$$\begin{gathered} q_s N_s + q_o N_o = 2\left(N_{ss} + N_{oo} + N_{os}\right) + N_{ps} + N_{op} \\ \text{and } \; q_s^* N_s^* = 2N_{ss}^* + N_{sp}^* \end{gathered} \tag{1.7}$$

Equating the numbers on the left hand side of (1.7) gives the coordination number q_s^* for effective solvent particles as $q_s x_s + q_o x_o$, i.e. the solvent and osmolyte coordination numbers weighted by their respective mol fractions x_i. A pair of equations analogous to (1.7) holds for protein monomers. Equating protein-protein, protein-bath, and bath-bath contacts in both models gives the correspondence between N_{ij} and N_{ij}^*:

$$\begin{aligned} N_{pp}^* &= N_{pp} \\ N_{ss}^* &= N_{ss} + N_{oo} + N_{os} \\ N_{ps}^* &= N_{ps} + N_{op} \end{aligned} \tag{1.8}$$

The N_{ij}^* then determine the transfer energy in the IS model through:

$$\frac{N_{pp}^* N_{ss}^*}{\left(N_{ps}^*\right)^2} = e^{-\beta\left(\varepsilon_{pp}^* + \varepsilon_{ss}^* - 2\varepsilon_{ps}^*\right)} \tag{1.9}$$

The other 2 equations determining the 3 energy scales in the IS model may be found from the conservation of total energy between the two models:

$$N_{pp}^* \varepsilon_{pp}^* + N_{ss}^* \varepsilon_{ss}^* + N_{ps}^* \varepsilon_{ps}^* = \sum_{i \le j} N_{ij} \varepsilon_{ij} \tag{1.10}$$

And by the ansatz that the total bath interaction energy (or equivalently protein-bath interaction energy) be the same in both models:

$$N_{ss}^* \varepsilon_{ss}^* = N_{ss} \varepsilon_{ss} + N_{oo} \varepsilon_{oo} + N_{os} \varepsilon_{os} \tag{1.11}$$

Equations (1.9)-(1.11) then determine $\varepsilon_{pp}^*, \varepsilon_{ps}^*, \varepsilon_{ss}^*$. For example, a system with $\varepsilon_{pp}, \varepsilon_{ps}, \varepsilon_{op}, \varepsilon_{ss}, \varepsilon_{os}, \varepsilon_{oo} = -1.1, +0.25, +1.1, -1.25, -0.8, -0.4$ in units of $k_B T$, with $q_p, q_o, q_s = 8, 6, 4$ and mol fractions $x_p, x_o, x_s = 0.05, 0.2, 0.75$ has effective energy scales $\varepsilon_{pp}^*, \varepsilon_{ps}^*, \varepsilon_{ss}^* = -1, +0.4, -1$; a set of parameters that we often use in our simulations. We have thus

shown that for any explicit osmolyte-solvent system, there exists an effective solvent model that captures the thermodynamics of protein contacts in the original system.

6. Solvation barriers

Figure 6 and Figure 7 show histograms of the atomic distances for representative pairs of atoms participating in native contacts, as a function of distance in Å. The sampling is obtained for simulations of the protein in a neutral solvent that is self-attractive ($\varepsilon_{ss}^{*} = -1$) but has hard-sphere interactions with the protein ($\varepsilon_{sp}^{*} = 0$). The temperature is set to the transition temperature T_f . Histograms are obtained by binning the distances and dividing by the phase space factor of $4\pi r^2$. The plots generally show a maximum near the van der Waals distance between atoms σ_{ij}^{vdW} , indicated above each subplot panel. Also apparent for some of the contacts (e.g. 1, 2, 3, 7, 8, 51, 52) is a shoulder indicating a preferential occupation at the solvent separated distance $\sigma_{ij}^{vdW} + \sigma_{j}^{water}$. For some contacts this preferential occupation appears as a secondary peak (contacts 48 and 49), for others such as contacts 4-6, 9, 46,47,50,53,54 the effect of a solvent separated minimum in the potential of mean force is manifested as a long tail in the distribution, or a shift in the peak of the distribution.

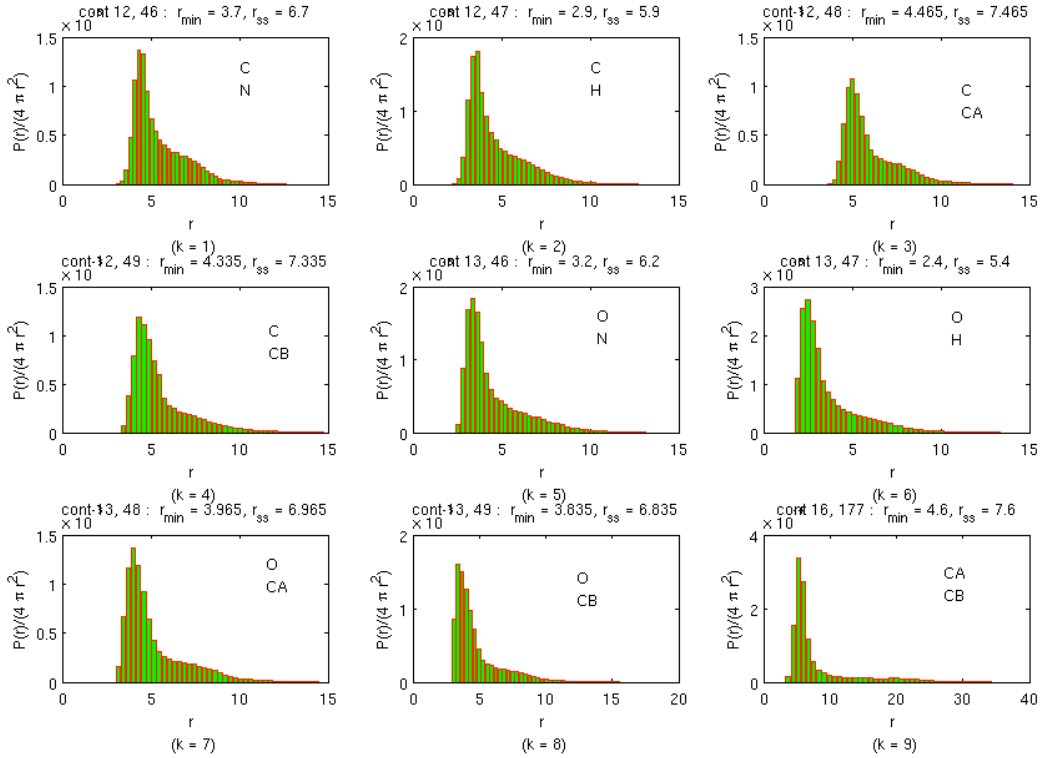

Figure 6: Histograms of atomic distances, for atoms participating in the first 9 native contacts (enumerated starting from the N terminus). Under each subplot is the contact label k. Above each subplot are the atoms indices i, j participating in the contact, and the corresponding atomic identities are listed inside each panel. Also above each subplot is the van der Waals distance between atoms $r_{\min} \equiv \sigma_{ij}^{vdW}$, and the solvent separated distance $r_{ss} \equiv \sigma_{ij}^{vdW} + \sigma_{j}^{water}$.

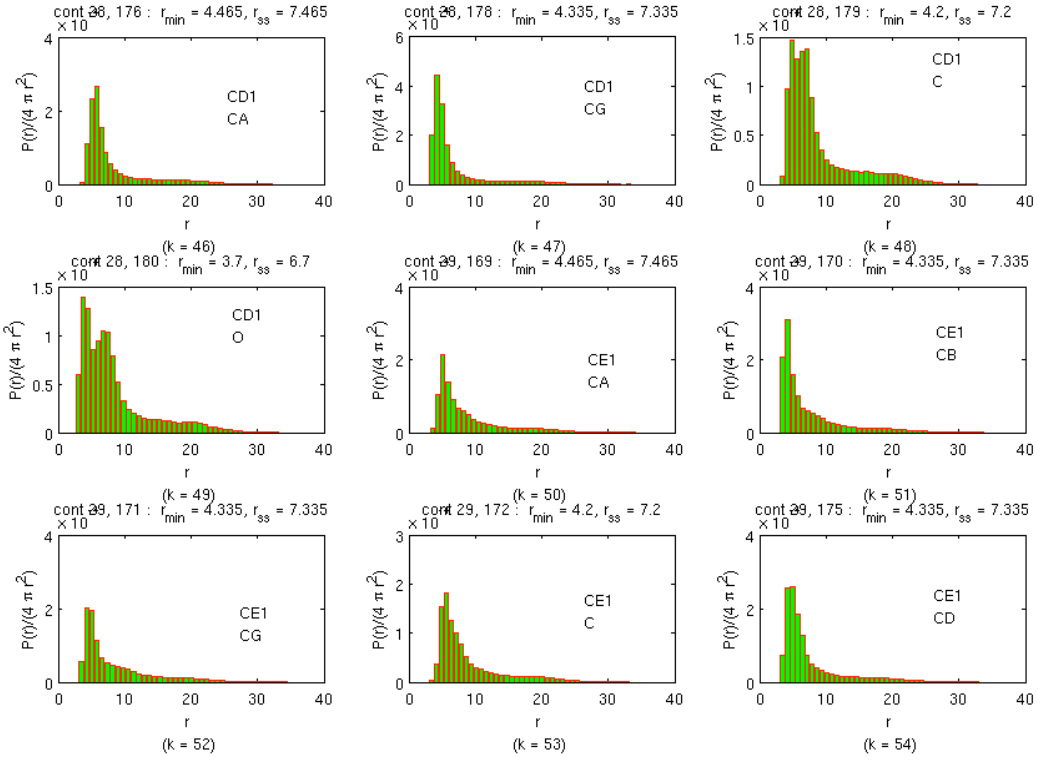

Figure 7: Histograms of atomic distances, for atoms participating in native contacts 46-54. Further description is given in the text and in the caption to Figure 6. Note that contacts 48 and 49 show secondary maxima near the solvent-separated distance, which implies strong solvation effects for these contacts.

The effects of solvent on the potentials of mean force (PMF) can be seen most clearly by investigating the transfer of the system from implicit solvent, to explicit solvent. The difference in the PMF is given by

$$\Delta W_{vac\to HS} = W_{HS} - W_{vac} = -\log\left(\frac{g_{HS}(r)}{g_{vac}(r)}\right) = \log\left(\frac{P_{vac}(r)}{P_{HS}(r)}\right). \qquad (1.12)$$

Here $g_i(r)$ is the pair distribution function in each respective solvent and $P_i(r)$ is the probability density a pair of atoms resides at position r. The last equality arises because factors of V and $4\pi r^2$ cancel for the ratio inside the log. Here $P_i(r)$ is obtained at the respective transition temperature T_f for each model solvent (at the same temperature the two systems explore completely different ensembles). A similar procedure to the above was employed by Sobolewski et al to find solvent contribution to the PMFs for nonpolar dimers in water (25).

Figure 8 and Figure 9 shows $\Delta W_{vac\to HS}$ for several representative native contacts (28-45). A ubiquitous feature of the transfer profiles for these and the other native contacts (data not shown) is the existence of a desolvation barrier, which appears typically between the van der Waals distance σ_{ij}^{vdW} and the

solvent separated distance $\sigma_{ij}^{vdW} + \sigma_{j}^{water}$. Occasionally an additional bias appears indicating a stronger native contact in the presence of explicit solvent (e.g. contacts 30, 34, 40, 41). At larger distances, the probability is small for both HS and implicit solvent, so the data becomes statistically unreliable. Nevertheless, there is also the ubiquitous feature of a bias favouring intermediate values of distance separation in the presence of explicit solvent. This bias represents the increased sampling of intermediate distances in HS solvent, which is itself a consequence of the induced collapse of the unfolded state in the presence of HS solvent.

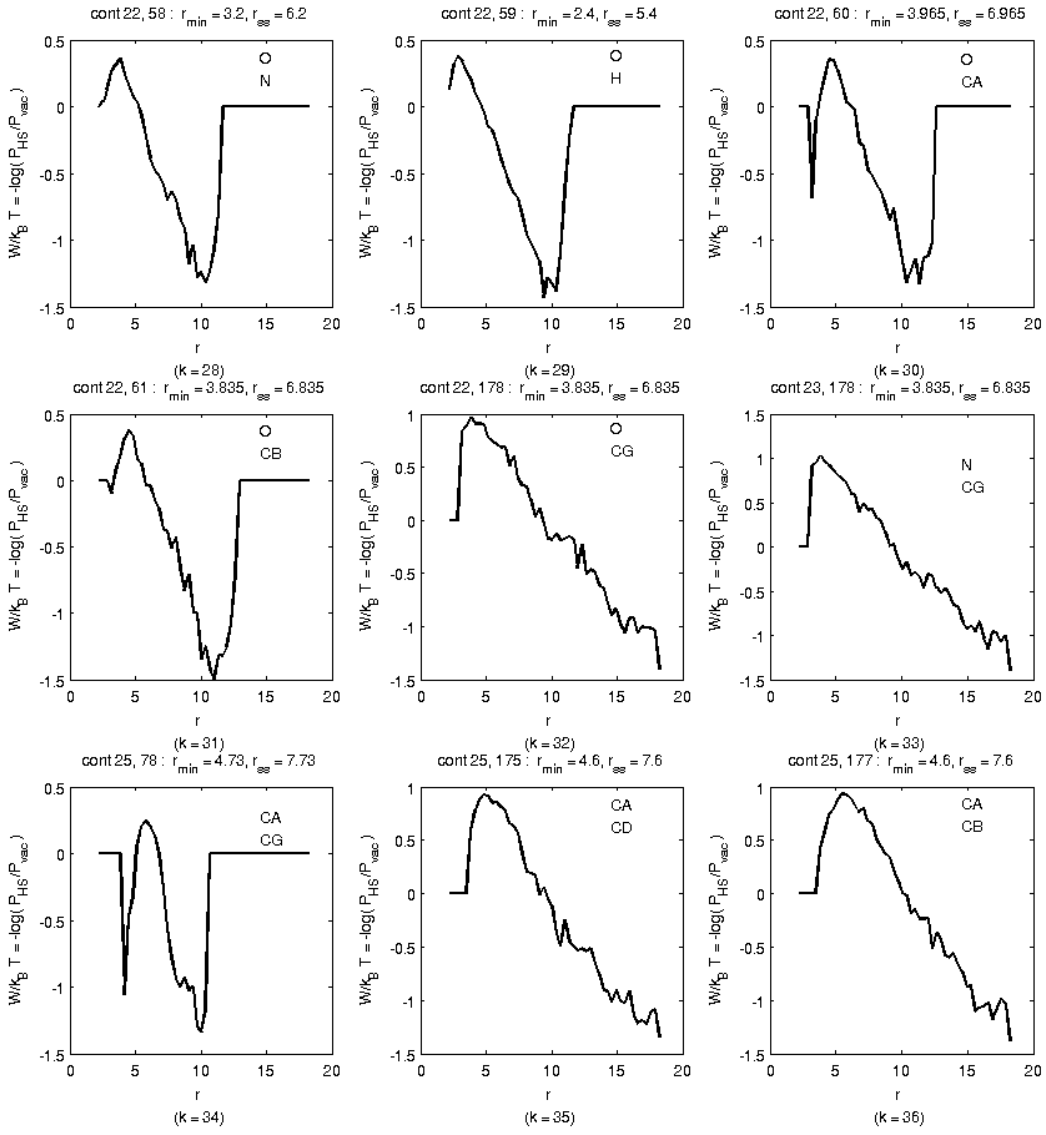

Figure 8: Difference in PMFs in the transfer from implicit solvent to neutral solvent, for native contacts 28-36. Notation is the same as in Figure 6.

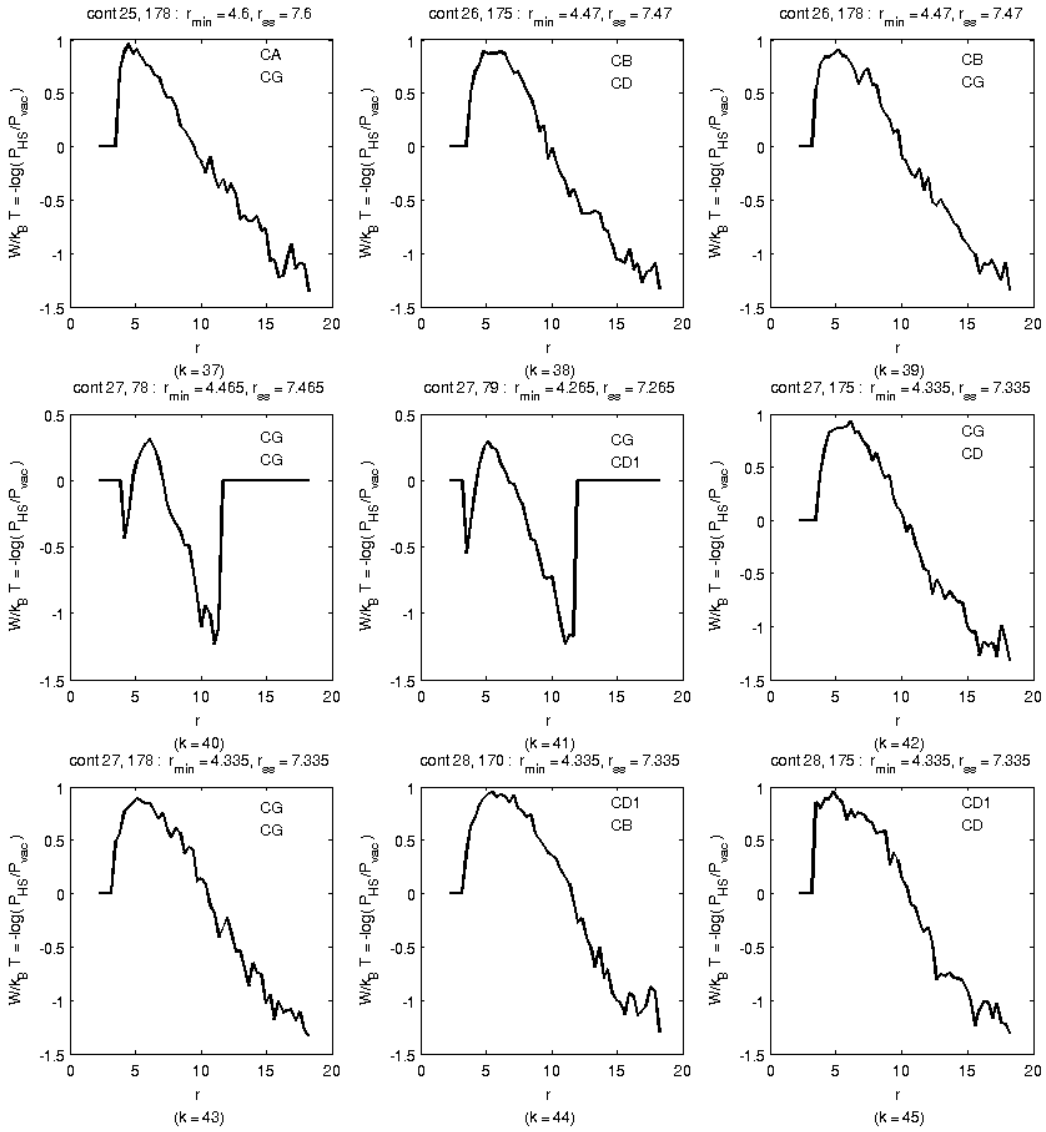

Figure 9: Difference in PMFs in the transfer from implicit solvent to neutral solvent, for native contacts 37-45.

The inset to figure 2A in the main text is reproduced below. It is obtained by taking an average over a representative set of native contacts that had short-range desolvation barriers (less than $\sim 8\ \AA$). These contacts were 1-8, 38-31, 34, 40, 41, 57-65, 70-72, 82-95, 226-239. This was then added to the square well potential to obtain the new effective potential in the presence of HS solvent. The main features are a slight elevation of the contact potential well, a desolvation barrier at around 5 Å, and a broad minimum centered around 9 Å corresponding to the induced collapse of the unfolded state and concurrent loss of cooperativity.

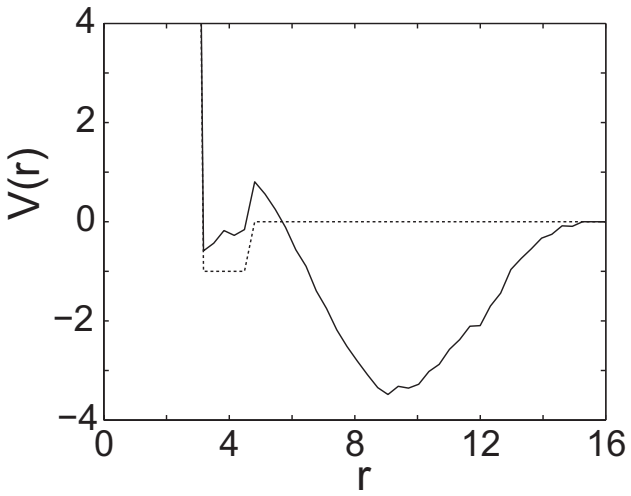

Figure 10: Dashed line: Square well pair potential (taken for contact 1 between atoms C(12) and N(46)). Solid line: Modified effective pair potential in the presence of explicit neutral solvent, obtained by adding the averaged transfer PMF described above to the square well potential.

1. Zhou, Y. Q., and M. Karplus. 1999. Folding of a model three-helix bundle protein: A thermodynamic and kinetic analysis. Journal of Molecular Biology 293:917-951.
2. Linhananta, A., J. Boer, and I. MacKay. 2005. The equilibrium properties and folding kinetics of an all-atom G(o)over-bar model of the Trp-cage. Journal of Chemical Physics 122.
3. Zhou, Y. Q., C. Zhang, G. Stell, and J. Wang. 2003. Temperature dependence of the distribution of the first passage time: Results from discontinuous molecular dynamics simulations of an all-atom model of the second beta-hairpin fragment of protein G. Journal of the American Chemical Society 125:6300-6305.
4. Borreguero, J. M., N. V. Dokholyan, S. V. Buldyrev, E. I. Shakhnovich, and H. E. Stanley. 2002. Thermodynamics and folding kinetics analysis of the SH3 domain from discrete molecular dynamics. Journal of Molecular Biology 318:863-876.
5. Jang, H. B., C. K. Hall, and Y. Q. Zhou. 2004. Assembly and kinetic folding pathways of a tetrameric beta-sheet complex: Molecular dynamics simulations on simplified off-lattice protein models. Biophysical Journal 86:31-49.
6. Ding, F., D. Tsao, H. F. Nie, and N. V. Dokholyan. 2008. Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure 16:1010-1018.
7. Ding, F., J. J. LaRocque, and N. V. Dokholyan. 2005. Direct observation of protein folding, aggregation, and a prion-like conformational conversion. Journal of Biological Chemistry 280:40235-40240.
8. Neidigh, J. W., R. M. Fesinmeyer, and N. H. Andersen. 2002. Designing a 20-residue protein. Nature Structural Biology 9:425-430.
9. Neria, E., S. Fischer, and M. Karplus. 1996. Simulation of activation free energies in molecular systems. Journal of Chemical Physics 105:1902-1921.
10. Go, N. 1983. THEORETICAL-STUDIES OF PROTEIN FOLDING. Annual Review of Biophysics and Bioengineering 12:183-210.
11. Zhou, Y. Q., and A. Linhananta. 2002. Role of hydrophilic and hydrophobic contacts in folding of the second beta-hairpin fragment of protein G: Molecular dynamics simulation studies of an all-atom model. Proteins-Structure Function and Genetics 47:154-162.

12. Plotkin, S. S., J. Wang, and P. G. Wolynes. 1997. Statistical mechanics of a correlated energy landscape model for protein folding funnels. Journal of Chemical Physics 106:2932-2948.
13. Plotkin, S. S., and J. N. Onuchic. 2002. Understanding protein folding with energy landscape theory - Part II: Quantitative aspects. Quarterly Reviews of Biophysics 35:205-286.
14. Kumar, S., D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. 1992. THE WEIGHTED HISTOGRAM ANALYSIS METHOD FOR FREE-ENERGY CALCULATIONS ON BIOMOLECULES .1. THE METHOD. Journal of Computational Chemistry 13:1011-1021.
15. Socci, N. D., and J. N. Onuchic. 1995. KINETIC AND THERMODYNAMIC ANALYSIS OF PROTEINLIKE HETEROPOLYMERS - MONTE-CARLO HISTOGRAM TECHNIQUE. Journal of Chemical Physics 103:4732-4744.
16. Dokholyan, N. V., S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. 1998. Discrete molecular dynamics studies of the folding of a protein-like model. Folding & Design 3:577-587.
17. Zhou, Y. Q., and A. Linhananta. 2002. Thermodynamics of an all-atom off-lattice model of the fragment B of Staphylococcal protein A: Implication for the origin of the cooperativity of protein folding. Journal of Physical Chemistry B 106:1481-1485.
18. Shimada, J., E. L. Kussell, and E. I. Shakhnovich. 2001. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. Journal of Molecular Biology 308:79-95.
19. Alder, B. J., D. A. Young, and M. A. Marx. 1972. STUDIES IN MOLECULAR DYNAMICS .10. CORRECTIONS TO AUGMENTED VAN-DER-WAALS THEORY FOR SQUARE-WELL FLUID. Journal of Chemical Physics 56:3013-&.
20. Mishra, B. M., and S. K. Sinha. 1984. THERMODYNAMIC PROPERTIES OF A TWO-DIMENSIONAL SQUARE-WELL FLUID. Pramana 23:79-90.
21. Kiselev, S. B., J. F. Ely, L. Lue, and J. R. Elliott. 2002. Computer simulations and crossover equation of state of square-well fluids. Fluid Phase Equilibria 200:121-145.
22. Khanna, K. N., A. Tewari, and D. K. Dwivedee. 2006. A new coordination number model and heat capacity of square-well fluids of variable width. Fluid Phase Equilibria 243:101-106.
23. Zhou, Y. Q., M. Karplus, J. M. Wichert, and C. K. Hall. 1997. Equilibrium thermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. Journal of Chemical Physics 107:10691-10708.
24. Kaya, H., and H. S. Chan. 2000. Polymer principles of protein calorimetric two-state cooperativity. Proteins-Structure Function and Genetics 40:637-661.
25. Sobolewski, E., M. Makowski, C. Czaplewski, A. Liwo, S. Oldziej, and H. A. Scheraga. 2007. Potential of mean force of hydrophobic association: Dependence on solute size. Journal of Physical Chemistry B 111:10765-10774.