SI Text

Details of the System. We consider a hydrophobic chain in explicit solvent. The unbranched chain is composed of $N_{\rm s} = 12$ spherical atoms of diameter 7.2 Å and mass 70.73 amu. The interactions between the chain atoms were treated with the purely repulsive WCA potential ($\sigma = 6.4145$ Å$, \epsilon = 15$ kJ/mol) (1-4). This potential, which is obtained by truncating and shifting the Lennard-Jones 12-6 potential at its minima, provides a slightly softened hard-sphere interaction that is convenient for molecular dynamics (MD) simulations. Neighboring atoms are linked by harmonic bonds, using

$$V_{\rm bond} = \sum_{k=1}^{N_{\rm s}-1} \frac{1}{2} k_{\rm b} (l_{\rm b} - |\mathbf{r}_{k+1} - \mathbf{r}_k|)^2, \tag{12}$$

where $|\mathbf{r}_{k+1} - \mathbf{r}_k|$ is the distance between chain atoms at positions $\mathbf{r}_k$ and $\mathbf{r}_{k+1}$, $l_{\rm b} = 7.2$ Å, and $k_{\rm b} = 84.9102$ kJ·mol^{-1}·Å^{-2}. To add some rigidity to the chain, a harmonic potential is included to penalize its curvature,

$$V_{\rm angle} = \sum_{k=2}^{N_{\rm s}-1} \frac{1}{2} k_{\phi} \phi_k^2, \tag{13}$$

where ϕ_k is the angle between neighboring bond vectors $(\mathbf{r}_{k+1} - \mathbf{r}_k)$ and $(\mathbf{r}_k - \mathbf{r}_{k-1})$, and $k_\phi = 11.1407$ kJ·mol^{-1}·radian^{-2}.

The chain was solvated with $33,912$ water molecules in an orthorhombic simulation cell with periodic boundary conditions and dimensions of 99.5 Å × 99.5 Å × 116.1 Å. Interactions among water molecules are described with the SPC/E rigid water potential (5), and the chain atoms interact with the oxygen atoms in the water molecules via WCA repulsions ($\sigma = 4.617$ Å$, \epsilon = 10$ kJ/mol) (1-4). Electrostatic interactions are included using the smoothed particle mesh Ewald method (6). Given the generality of the lengthscale-dependent hydrophobic effect for liquids near liquid-vapor phase coexistence and far from the critical point (7), we expect the SPC/E model of water to give a qualitatively reliable picture of hydrophobic collapse. Huang *et al.* (8) have explored the thermodynamics of hydrophobic solvation for the SPC/E water model in detail.

All MD simulations were performed at 300 K using the Nosé-Hoover thermostat and a timestep of 2 fs (9). Constant pressure conditions were maintained in the simulations using the liquid-vapor coexistence technique described in the text. To ensure that the liquid-vapor interface remains flat, as opposed to collapsing under the surface tension of the liquid, the solvent density restraint potential in main text was applied to the top two layers of lattice cells in the simulation box with parameters $P_k^* = 0$ and $\kappa = 100$ kJ·mol^{-1}·molecule^{-2}. All MD calculations were performed using a modified version of the DL_POLY_3 molecular simulation package (10).

String Method in Collective Variables. The string method (11-13) is a technique for calculating the committor function for a dynamical process. The constant-value contours of the committor function are approximated by a collection of hyperplanes that are perpendicular to a given path (the string). This approximation can be justified within the framework of transition path theory (TPT) (14), which yields a variational criterion for the string such that the perpendicular hyperplanes optimally approximate the isocommittor surfaces.

The string method in collective variables characterizes the committor function projected onto the space of collective variables. Provided that the collective variables adequately describe the reaction and that the reactive trajectories projected onto the space of collective variables remain confined to a relatively narrow tube, the approximation is not only optimal but also accurate. In this case, it can be shown that the string coincides with the minimum free energy path (MFEP) in the space of collective variables, which is also the path of maximum likelihood for the reaction monitored in these variables.

Unlike other free-energy mapping techniques, the string method focuses only on a linear subset of the space of collective variables. It thus scales independently of the dimensionality of the full free energy surface. The method does not require performance of long, reactive MD trajectories. Instead, a double-ended approach is employed in which the path is updated using short, restrained MD simulations.

A complete discussion of the implementation of the string method in collective variables can be found in ref. 13. We represent the string in the variables of the chain atom positions and the solvent cell densities, as is explained in the primary text. Data needed for the calculation of the minimum free energy path were obtained from restrained MD simulations,

$$\frac{\partial F(\mathbf{z})}{\partial z_i} = \lim_{\kappa\to\infty} \kappa \int_{\Re^n} (z_i - z_i(\mathbf{x}))\rho_{\kappa,z}(\mathbf{x})d\mathbf{x} \tag{14}$$

and

$$M_{ij}(\mathbf{z}) = \lim_{\kappa\to\infty} \int_{\Re^n} \sum_{k=1}^{n} m_k^{-1} \frac{\partial z_i(\mathbf{x})}{\partial x_k} \frac{\partial z_j(\mathbf{x})}{\partial x_k} \rho_{\kappa,z}(\mathbf{x})d\mathbf{x}, \tag{15}$$

where

$$\begin{aligned} \rho_{\kappa,z}(\mathbf{x}) &= \exp(-\beta U_{\kappa,z}(\mathbf{x}))/Z_{\kappa,z}, & (16)\\ Z_{\kappa,z} &= \int_{\Re^n} \exp(-\beta U_{\kappa,z}(\mathbf{x}))d\mathbf{x}, & (17)\\ \text{and}\quad U_{\kappa,z}(\mathbf{x}) &= V(\mathbf{x}) + \frac{\kappa}{2}\sum_{i=1}^{\mathcal{N}}(z_i - z_i(\mathbf{x}))^2. & (18) \end{aligned}$$

The string is converged to the MFEP using Eqs. **5** and **6**. Between timesteps of these dynamics, the string is smoothed and reparameterized. The smoothing procedure is performed using

$$\mathbf{z}^{m,*} = (1-\eta)\mathbf{z}^{m} + \frac{\eta}{2}(\mathbf{z}^{m-1} + \mathbf{z}^{m+1}), \tag{19}$$

where $\mathbf{z}^{m,*}$ and $\mathbf{z}^{m}$ indicate the m^{th} configuration of the smoothed and unsmoothed string, respectively. The parameter $\eta = 0.05$ was chosen to be sufficiently small ($O(N_{\mathrm{d}}^{-1})$, as is discussed in ref. 13) that the smoothing procedure does not effect the accuracy of the calculated MFEP. Reparameterization of the string was performed using the linear interpolation scheme described in ref. 13.

The string was initialized from configurations of an MD trajectory in which the hydrated chain was artificially extended from a collapsed configuration with the aid of a greatly magnified force constant in the chain potential energy term V_{angle}. In the first stage of the string calculation, the string was discretized using $N_d = 20$ configurations and evolved using large timesteps and weak collective variable restraints. Specifically, we employed chain atom restraint force constants of 2 kJ·mol^{-1}·Å^{-2}, solvent restraints of 10 kJ·mol^{-1}·molecule^{-2}, and a steepest descent timestep of 300 fs. Restrained MD trajectories of 10 ps were performed during this stage, including 0.5 ps of equilibration. This first stage of the string calculation included nine steepest descent timesteps.

In a second stage of the string calculation, the number of configurations used to represent the path was increased to $N_d = 40$, the chain atom restraints were increased to 5 kJ·mol^{-1}·Å^{-2}, the solvent restraints were increased to 40 kJ·mol^{-1}·molecule^{-2}, the steepest descent timestep was decreased to 40 fs, and the restrained MD trajectory time was increased to 20 ps, including 0.5 ps of equilibration. The second stage of the string calculation was terminated after six optimization steps, at which point reasonable convergence, as determined by monitoring changes in the path and its corresponding free energy profile, was obtained. Throughout both stages of the string calculation, the path endpoint corresponding to the extended chain was (after local relaxation) held fixed, and the other endpoint corresponding to the collapsed chain was allowed to relax on the free energy surface according to Eq. **6**. The final, converged string bears little resemblance to the initial guess.

Statistical error in the evaluation $\partial F(\mathbf{z})/\partial z_i$ and $M_{ij}(\mathbf{z})$ give rise to noise in the evolution of the string according to Eq. **5**, and the former also introduces statistical error in the calculation of free energy profiles via Eq. **7**. By averaging these quantities over trajectories of length 10-20 ps, we have presumed that the atomistic coordinates equilibrate on considerably faster time scales. SPC/E water diffuses at over 1 Å/ps and and its center of mass velocity autocorrelation function decays in approximately 1 ps, so it is clear that in a trajectory of 20 ps, the molecular distribution of water will be well equilibrated on the 0.3 Å length scale of the solvent density field. That the motion of the chain atoms also equilibrates on this time scale is illustrated in SI Fig. 6, which shows the cumulative average of the x, y, and

z components of the mean force on a particular chain atom (atom number 7, found at the middle of the chain) at a particular point (configuration number 5) along the optimized string. Also plotted is the variance of the mean calculated over picosecond blocks.

We now describe our estimation of the statistical error in the free energy profile. In the final optimization step of the string, the variance in the mean force for each solvent and solute collective variable was calculated, sampling at every molecular dynamics time step. But these data points are highly correlated, giving rise to an underestimation in the statistical error that must be corrected. Using the time-series data for the mean forces (such as that shown in SI Fig. 6), we calculate the statistical inefficiency, or the length of time between entirely independent samples of the force, to be approximately 0.5-1 ps (15). The value of 1 ps for the statistical inefficiency was then used to correct the oversampling bias in the calculated variance of the mean forces on the collective variables (15). With this unbiased estimate for the statistical error in the mean forces, we use Eq. **7** to obtain the accumulated statistical error in the free energy profile, shown in SI Fig. 7. Although the absolute value of the statistical error in the free energy profile becomes sizable over the length of the entire string, the uncertainty in the relative free energy of consecutive configurations is only about the size of the circles used in the plots. Therefore, local features of the path, such as the barrier at configuration 22, are resolved within statistical error.

The string method is a local, rather than global, optimization scheme. Different minimized free energy paths might have been obtained from string calculations started with different initial paths. However, the simple topology of the chain, as well as the unrestrained MD simulations presented in the paper, suggest that the calculated MFEP is a reasonable characterization of the reaction mechanism for the hydrophobic collapse of the hydrated chain.

All computations were performed in parallel using 2.2 GHz AMD Operton processors. Each step in the first of the stage of the string calculation required 600 CPU hours. Each step in the second stage required $2,400$ CPU hours. Each evaluation of the committor function described in “The Committor Function and a Proof of Principle for Coarse-Graining” required $15,000$ CPU hours.

Estimation of the Barrier Between Reactive Channels. An estimate for the energetic cost of translating the bend along the chain suggests that trajectories that collapse by forming asymmetric bends in the chain do not belong to a distinct reactive channel from those that form a symmetric bend. We compared the bending energy for the chain geometry from configuration 22 in Fig. 2 with the bending energy for the geometry of the chain obtained by placing a hydrophobe at the midpoint of each hydrophobe-hydrophobe bond in configuration 22 (and with the position of the last hydrophobe at the end of the chain being determined by linear extrapolation). The difference between these bending energies is small in comparison to the uncertainty in energy due to thermal fluctuations.

Alternative Models for the Solvation Free Energy. Even with the aide of a one-parameter fit, the solvation free energy profiles estimated on the basis of only solvent-depleted surface area or only solvent-depleted volume do not match the accuracy obtained using Eq. **9**. In SI Fig. 9A, we see the profile (dotted-dashed line) obtained by fitting $\tilde{\gamma}$ in the expression

$$F_{\mathrm{h}}(\alpha) = \tilde{\gamma} A_{\mathrm{tot}}(\alpha). \tag{20}$$

The height of the shoulder at configurations 14-18 and the barrier peak region do not reproduce the simulated results (solid line) as well as using Eq. **9** (dotted line). As is seen SI Fig. 9B, the solvation free energy profile (dotted-dashed) that is obtained using the expression

$$F_{\mathrm{h}}(\alpha) = \tilde{\lambda} V_{\mathrm{tot}}(\alpha) \tag{21}$$

is less accurate than that obtained using either Eqs. **9** or **20**.

Convergence of the Free Energy Profile with Increasing Solvent Box Sizes. We show that the free energy profile calculated using Eq. **11** converges with respect to the solvent contribution. The curves in SI Fig. 10 correspond to free energy profiles obtained with V equal to the middle $b \times b \times b$ lattice cells of the simulation box. The profile changes dramatically with b until V fully encompasses the region of the bending chain, where the dewetting transition occurs. For $b > 16$, no major changes are seen in the profile, since the added contributions are from solvent cells that are distant from the collapsing chain. The small changes that are found with large b are due to statistical noise. This noise increases with larger b since the number of included solvent cells increases as b^3.

1. Chandler D, Weeks JD (1970) *Phys Rev Lett* 25:149-152.

2. Weeks JD, Chandler D, Andersen HC (1971) *J Chem Phys* 54:5237-5247.

2. Weeks JD, Chandler D, Andersen HC (1971) *J Chem Phys* 55:5422-5423.

4. Andersen HC, Weeks JD, Chandler D (1971) *Phys Rev A* 4:1597-1607.

5. Berendsen HJC, Grigera JR, Straatsma TP (1987) *J Chem Phys* 91:6269-6271.

6. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) *J Chem Phys* 103:8577-8593.

7. Chandler D (2005) *Nature* 437:640-647.

8. Huang DM, Geissler PL, Chandler D (2001) *J Phys Chem B* 105:6704-6709.

9. Hoover WG (1985) *Phys Rev A* 31:1695-1697.

10. Todorov I, Smith W (2004) *Phil Trans R Soc London A* 362:1835-1852.

11. E W, Ren W, Vanden-Eijnden E (2002) *Phys Rev B* 66:52301.

12. E W, Ren W, Vanden-Eijnden E (2005) *J Phys Chem B* 109:6688-6693.

13. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) *J Chem Phys* 125:024106.

14. Vanden-Eijnden E (2006) in *Computer Simulations in Condensed Matter: From Materials to Chemical Biology*, eds Ferrario M, Ciccoti G, Binder K (Springer, Berlin), Vol 2, pp 439-478.

15. Allen MP, Tildesley DF (1987) *Computer Simulation of Liquids* (Oxford Univ Press, Oxford), pp 192-195.