Electronic Supporting Information for

Structural Features of Interfacial Water Predict the Hydrophobicity of Chemically Heterogeneous Surfaces

Bradley C. Dallin, Atharva S. Kelkar, and Reid C. Van Lehn*

Department of Chemical and Biological Engineering, University of Wisconsin – Madison,
Madison, WI, 53706, United States.

Table of Contents

S1: Molecular Dynamics Simulation Details

We performed two different types of simulations for this study: unbiased molecular dynamics (MD) simulations and indirect umbrella sampling (INDUS) simulations. We describe the general setup for these simulations and the specific details of each below.

S1.1 General Simulation Details

In both the unbiased and INDUS simulations, a single self-assembled monolayer (SAM) composed of alkanethiol ligands with varying chemistries was created in a 3D periodic simulation box. Binary mixed SAMs with nonpolar methyl end groups and either hydroxyl, amine, or amide polar end groups with mole factions of f_P = 0.0, 0.25, 0.40, 0.50, 0.75, and 1.0 were modeled. All mixed SAMs were modeled in two limiting patterns: checkered and separated (Figure S1). The checkered pattern had ligands arranged such that polar and nonpolar end groups were uniformly distributed across the SAM. The separated pattern had ligands arranged such that nonpolar and polar end groups were separated into two regions. We also modeled single-component "charge-scaled" SAMs for each polar end group. These SAMs contained ligands where the partial charges of the end group were scaled by a multiplicative factor, *k*, to systematically vary the hydrophobicity of the surface (1, 2). The partial charges of end group atoms (defined as the set of atoms in the end group necessary to define a net neutral moiety) were multiplied by values of *k* between 0.0-0.9 in increments of 0.1. We note that Lennard-Jones parameters were not modified so that a SAM with polar end groups and *k* = 0.0 is distinct from a SAM with methyl end groups. Figure S1 illustrates how each ligand was scaled. Together, the checkered, separated, and charge-scaled SAMs defined a data set of 58 unique SAMs that were used for model training.

We performed three independent replicas of the unbiased and INDUS simulations for each of the 58 SAMs to compute error bars (174 simulations in total). In all simulations, we fixed the position of the ligand sulfur atoms by applying a harmonic restraint with a spring constant of 500,000 kJ/mol/nm^2 to mimic the strong Au-S bond that forms in experiments (3, 4). Ligands were positioned to be consistent with an Au(111) lattice with a grafting density of 21.6 Å/ligand consistent with experimental alkanethiol SAMs (5, 6). Ligands were initially directed so the S-to-end group vectors were pointed in the positive *z*-direction. A total of 144 ligands were added to each system spanning an area of 5.2 x 6.0 nm^2. A 5 nm thick water layer was added above the SAM such that the ligand end groups were in contact with the water molecules. The *z*-dimension of the simulation box was expanded by 3.5 nm to include a buffering vacuum layer between the sulfur groups of the SAM and the top of the water layer (7-9). The gold substrate was not explicitly included. In prior work, we found that excluding gold atoms led to better agreement with experimental trends (10). The CHARMM36 force field was used to model all ligand atoms (11-13). Water molecules were modeled using the TIP4P/2005 water model that is recommended with CHARMM36 and CGenFF (14). Electrostatic interactions were calculated using the smooth Particle Mesh Ewald algorithm (15) with short-range Coulomb, van der Waals, and neighbor list cutoffs set to 1.2 nm. All bonds were constrained using the LINCS algorithm (16). MD simulations using the leapfrog integrator with a 2 fs timestep were performed using GROMACS (version 2016.6) (17, 18). Simulations were performed in the *NVT* ensemble. The temperature was maintained at 300 K using a velocity-rescaling thermostat with a time constant of 0.1 ps (19). While pressure coupling was not explicitly performed, the pressure of the aqueous phase was maintained by the buffering vapor phase following previous work (7-9).

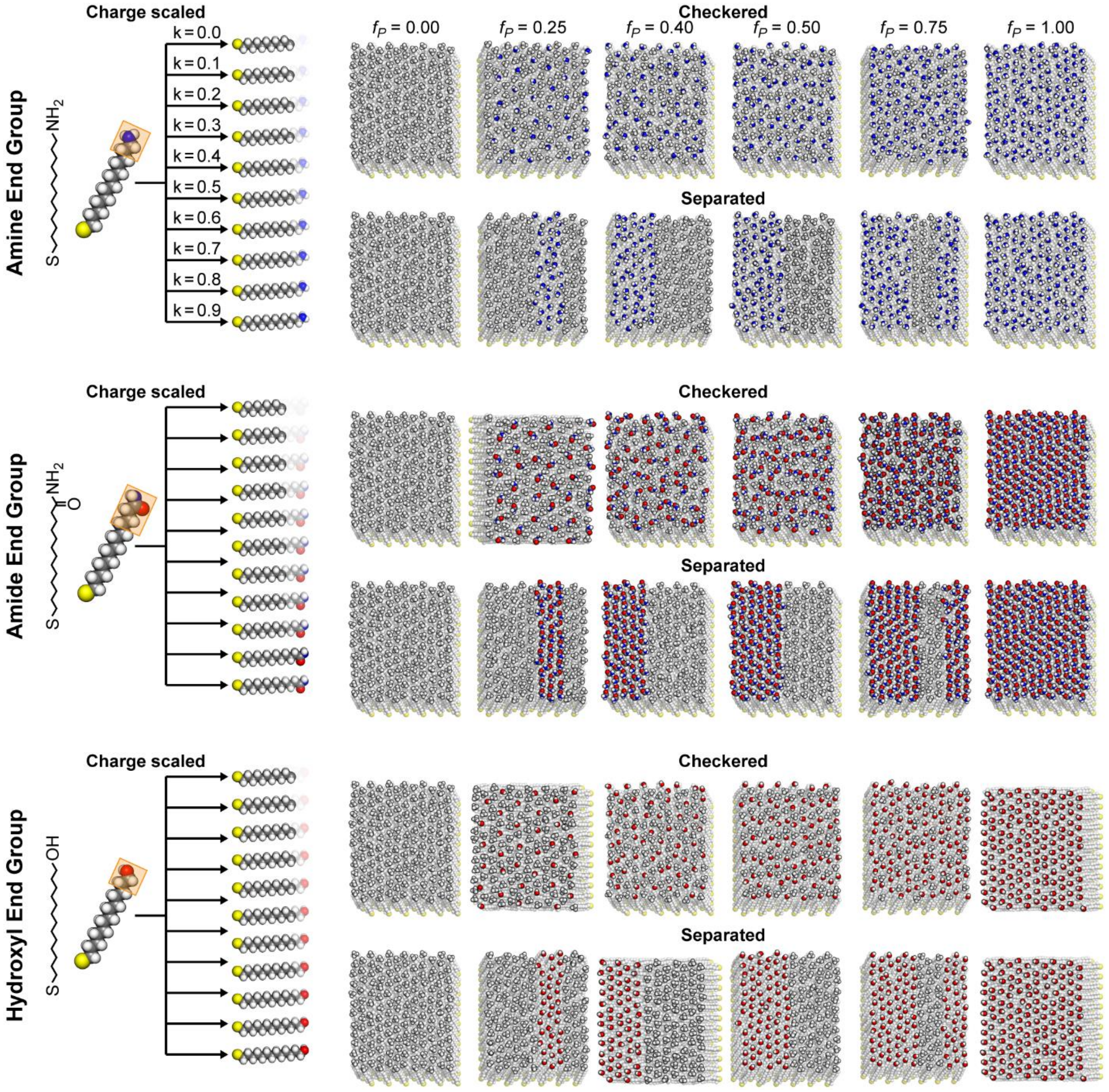

Figure S1. Graphical summary of SAM simulation systems. Images at left depict the charge-scaled ligands. The orange rectangles highlight atoms with scaled partial charges. Images at right show top-down simulation snapshots of checkered and separated SAMs with water molecules excluded to highlight chemical patterns. The fraction of polar end groups increases from left to right with f_P = 0.00 and f_P = 1.00 equivalent to homogeneous SAMs with methyl or polar end groups, respectively.

S1.2 Unbiased Simulations

Unbiased MD simulations were performed using four steps as illustrated in Figure S2. (1) *SAM preparation*: ligand structures were generated using Avogadro (version 1.2.0) and parameterized using the CGenFF/CHARMM force field (14). Next, the ligand structure and force field files were used to create SAM structure and force field files by using in-house Python scripts. Lastly, a 5 nm water layer was added above the SAM by solvating an empty 5.2 x 6.0 x 5.0 nm^3 simulation box using the GROMACS solvate function, then joining the SAM and water files together to create the initial configuration of the SAM-water system. (2) *Energy minimization*: Energy minimization was performed using the steepest descent algorithm until the maximum force between atoms reached the criterion of <200 kJ/mol/nm. (3) *Equilibration*: a 5 ns equilibration simulation was performed to relax the ligand and water atoms. (4) *Production*: a 10 ns production simulation was performed with configurations saved every 1 ps for analysis (generating 10,000 configurations per unbiased simulations).

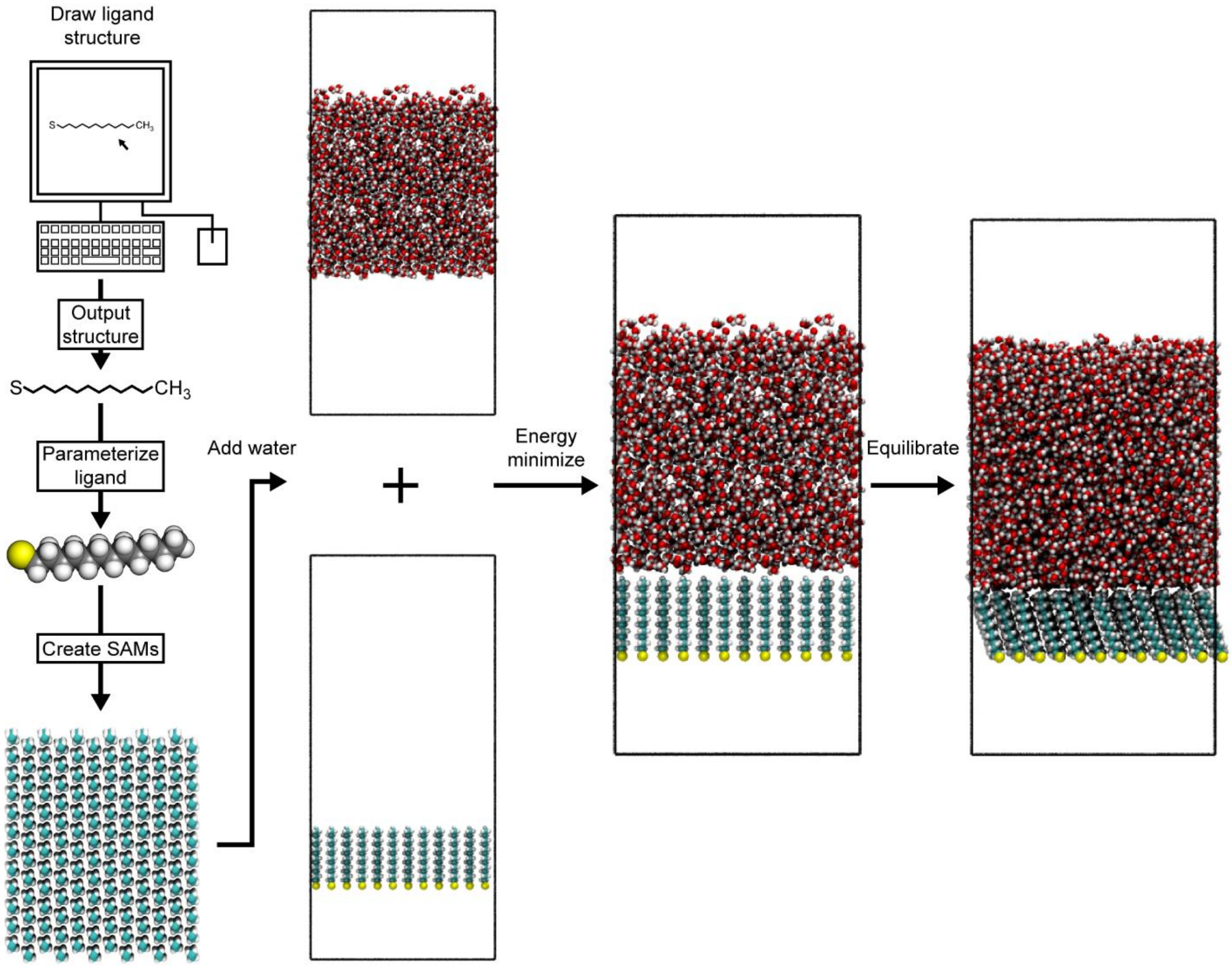

Figure S2. Schematic of SAM preparation and unbiased simulation workflow illustrated using a homogeneous SAM with methyl end groups.

S1.3 INDUS Simulations

The SAMs created for INDUS simulations followed steps 1-3 in Section S1.2. The INDUS method is explained in detail in Section S1.4; here, we describe only aspects of the simulation workflow. INDUS simulations were performed using the final configuration after equilibration as the starting configuration. Two additional steps were followed: (1) *Steered MD*: Steered MD was performed to generate initial configurations for INDUS. We defined a $2.0 \times 2.0 \times 0.3$ nm^3 INDUS cavity with the bottom of the cavity positioned at the SAM-water interface (as defined in Section S1.5). A harmonic potential with a spring constant of 8.5 kJ/mol/nm^3 was then applied to bias the number of water molecules within the cavity. Steered MD was performed by reducing the number of water molecules within the cavity at a rate of 2 ps/molecule for a total simulation time of 120 ps with configurations output every 2 ps. Initial configurations for INDUS were taken from these configurations. 16 total INDUS windows were used to sample the number of water molecules within the cavity: 5 windows (in an increment of 2 water molecules/window) for 0-8 water molecules within the cavity and 11 windows (in an increment of 5 water molecules/window) for 10-60 water molecules within the cavity. Spring constants for INDUS (detailed below) and the number of windows were determined by inspecting histograms used as input to the Weighted Histogram Analysis Method (WHAM). Figure S3 shows histograms for each of the single component SAMs; these histograms are representative of the entire data set.

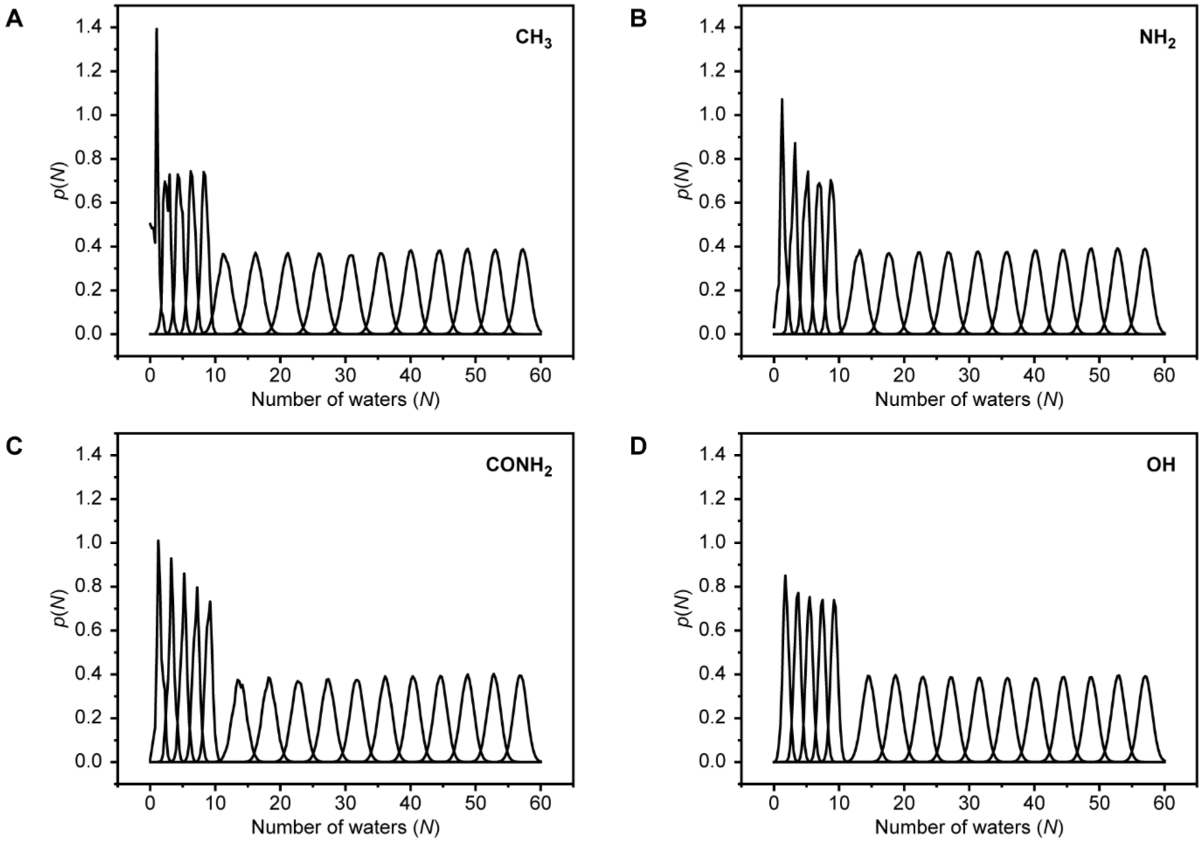

Figure S3. WHAM histograms for single-component SAMs with (A) methyl, (B) amine, (C) amide, and (D) hydroxyl end groups. Similar histograms were obtained for all 58 SAMs.

(2) *Equilibration/Production:* A 5 ns *NVT* simulation was performed for each INDUS window (16 per SAM) with harmonic potentials applied to restrain the number of water molecules within the cavity. INDUS windows with fewer than 10 water molecules required a spring constraint of 8.5 kJ/mol/nm^3. Windows with more than 10 water molecules required a spring constant of 2.0 kJ/mol/nm^3. The discrete and coarse-grained numbers of water molecules (described in Section S1.4) within the cavity were sampled every 0.2 ps. The sampling time necessary for convergence of the INDUS calculations was determined by varying both the sampling time (*i.e.*, the amount of simulation time used to generate input for WHAM) and the equilibration time (*i.e.*, the amount of simulation time used to initially equilibrate the system). Figure S4 shows representative curves for hydration free energy versus sampling time. By inspection, it was determined that 3 ns is sufficient sampling time for the INDUS calculations to converge. Figure S5 shows hydration free energy versus equilibration time curves for the same SAMs. By inspection, 2 ns of simulation time is sufficient for INDUS simulation to equilibrate.

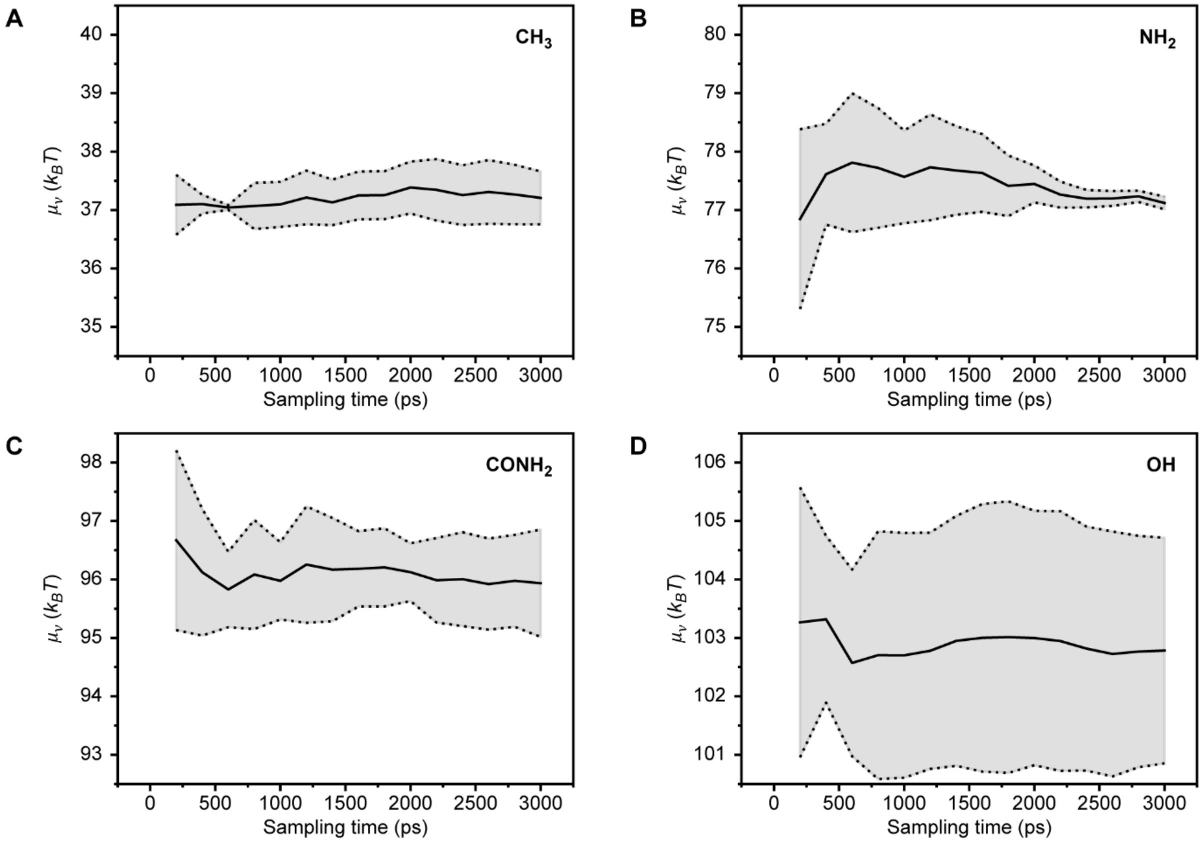

Figure S4. INDUS convergence time for the (A) methyl-, (B) amine-, (C) amide-, (D) hydroxyl-terminated SAMs. 2 ns of equilibration time was used to test convergence. Error bars were computed as the standard deviation of three independent INDUS calculations.

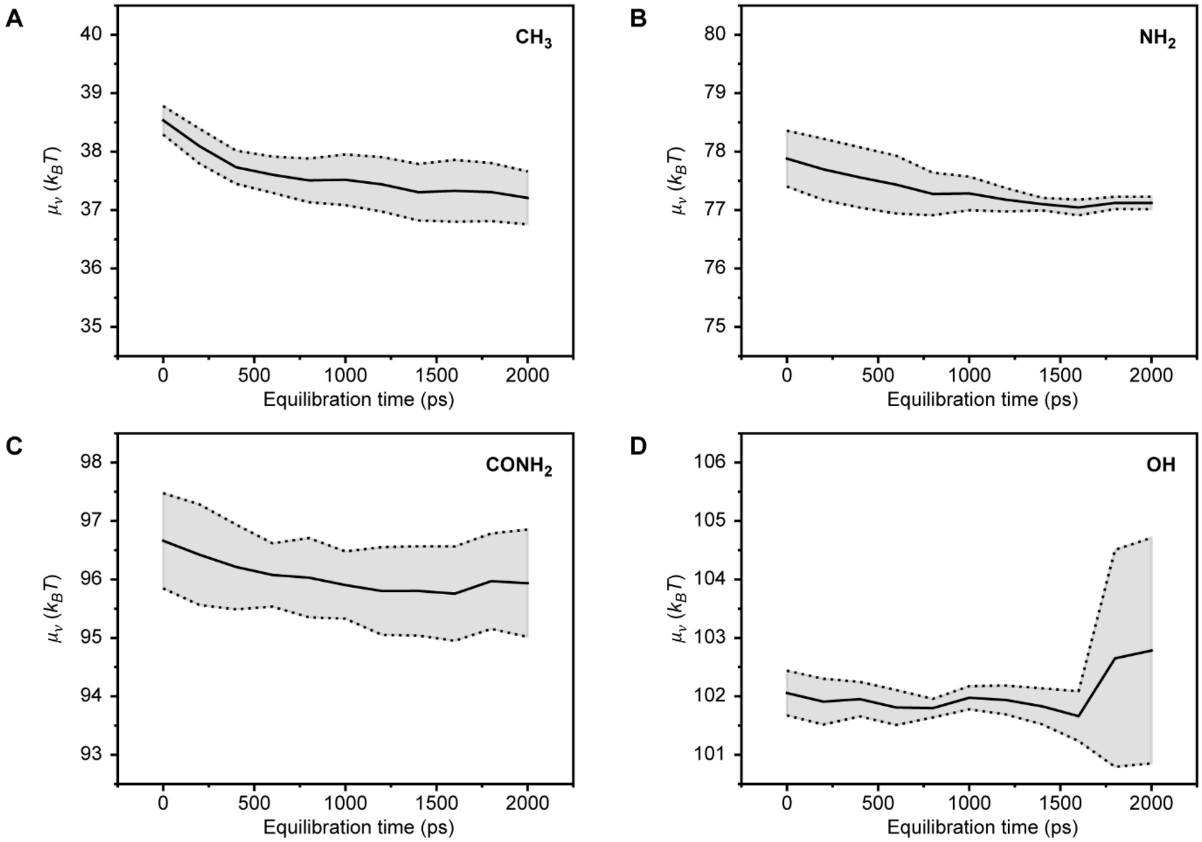

Figure S5. INDUS equilibration time for the (A) methyl-, (B) amine-, (C) amide-, (D) hydroxyl-terminated SAMs. Error bars were computed as the standard deviation of three independent INDUS calculations.

S1.4 INDUS Calculation Details

INDUS simulations are performed by applying a harmonic potential to bias the number of water molecules, N, within a predefined volume (*i.e.*, cavity) (8). To avoid impulsive forces associated with discrete changes in N, a new coarse-grained number of water molecules, $\widehat{N}$, is defined by Equation S1:

$$\widehat{N} = \sum_{i=1}^{M} \int_{v} \phi(\boldsymbol{r}; \boldsymbol{r}_i)\, d\boldsymbol{r} \tag{S1}$$

M is the total number of water molecules, $\boldsymbol{r}$ is the coordinate of a position within the simulation box, $\boldsymbol{r}_i$ is the position of the oxygen atom of the i^{th} water molecule, and the integral is evaluated over the volume of the cavity (denoted by v). The function $\phi(\boldsymbol{r}; \boldsymbol{r}_i)$ is a Gaussian kernel function of the form shown in Equation S2 which converts the discrete water molecule positions to continuous densities:

$$\phi(\boldsymbol{r}; \boldsymbol{r}_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{\frac{3}{2}} \exp\left(-\frac{(\boldsymbol{r} - \boldsymbol{r}_i)^2}{2\sigma^2}\right) \tag{S2}$$

$\boldsymbol{r}$ and $\boldsymbol{r}_i$ are the same as in Equation S1 and σ sets the width of the Gaussian function. We set σ to 0.1 nm to approximate the radius of a water molecule (~0.15 nm). Values of $\widehat{N}$ were then biased using a harmonic potential of the form:

$$U_j(t) = \frac{1}{2}\kappa_j\left(\widehat{N} - \eta_j\right)^2 \tag{S3}$$

η_j is the control variable in the j^{th} window and κ_j is the spring constant for the corresponding biasing potential. The PLUMED (v2.5.1) plug-in for GROMACS was used to apply the biasing potential to water molecules inside the cavity (20). The unbiased joint probability function $p_v(N, \widehat{N})$ was obtained using the Weighted Histogram Analysis Method (WHAM) (21), then integrated over $\widehat{N}$ to obtain $p_v(N)$, or the probability associated with observing N water molecules within the cavity. Figure S6 illustrates example distributions of $\ln p_v(N)$ for five SAMs of increasing hydrophilicity along with a corresponding quadratic fit line for the most hydrophilic SAM (100 mol% OH) to demonstrate that the $p_v(N)$ distribution near this surface is Gaussian as expected. While all SAMs have the same most probable value of N, the other SAMs exhibit broader, non-Gaussian $p_v(N)$ distributions with fat low-N tails that are indicative of increasing hydrophobicity and an increased probability that the cavity dewets. The hydration free energy, μ_v, is then obtained from the value of $p_v(N = 0)$ by Equation S4 and used to quantify the relative hydrophobicity between SAMs in the main text (7, 22, 23).

$$\mu_v = -k_B T \ln p_v(0) \tag{S4}$$

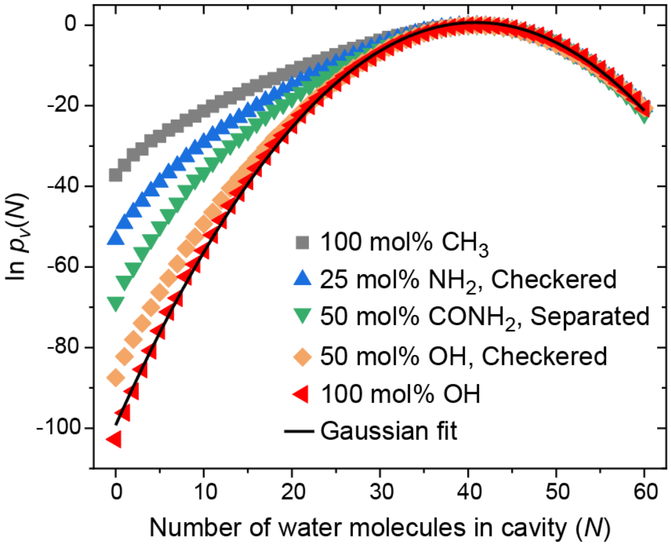

Figure S6. Probability distributions after WHAM reweighting for the number of water molecules within the INDUS cavity (N) for several representative SAMs of varying hydrophobicity (including mixed SAMs in both checkered and separated patterns, as shown in Figure 1A of the main text). The black line indicates a Gaussian (quadratic) fit, which accurately describes the most hydrophilic (100 mol% OH) surface. While all SAMs share the most probable value of N, the more hydrophobic surfaces exhibit non-Gaussian distributions with fat tails at low values of N.

S1.5 SAM-Water Interface Definition

We define the SAM-water interface as the constant water density isosurface adjacent to the SAM (24). We calculated the isosurface by generating a coarse-grained density field of the water molecules by applying a Gaussian kernel function (Equation S2) to the positions of the water molecules. The width (σ) was set to 0.24 nm. This value was selected because it is approximately the bulk correlation length of liquid water such that this value reduces the number of voids in the coarse-grained density profile. The water density was histogrammed by discretizing the simulation box into a 3D grid with grid points separated by 0.1 nm in each Cartesian direction. The isosurface was calculated from the averaged density profile from the final 1000 configurations of the initial equilibration simulation by interpolating between points on this grid to obtain the contour of which the water density was equal to half the density of bulk water (16 water molecules/nm^3).

S2: Data-Centric Analysis Description

We calculated 10 water structural parameters (152 water structural features) to train the data-centric regression methods used in this study. This section provides details about the water structural parameter calculations and validation of the regression models. Water structure parameters are summarized in Table S1.

Table S1. Summary of water structure parameters calculated in unbiased MD simulations.

Water structure parameter	**Symbol**	**Number of Features**
Triplet angle distribution	$p(\theta)$	90
OH bond orientational angle distribution	$p(\phi)$	18
Number of all hydrogen bonds	N_{All}^{hb}	1
Number of SAM-SAM hydrogen bonds	$N_{\mathrm{SAM-SAM}}^{hb}$	1
Number of SAM-water hydrogen bonds	$N_{\mathrm{SAM-water}}^{hb}$	1
Number of water-water hydrogen bonds	$N_{\mathrm{water-water}}^{hb}$	1
All hydrogen bonds distributions	$p\left(N_{\mathrm{All}}^{hb}\right)$	10
SAM-SAM hydrogen bonds distribution	$p\left(N_{\mathrm{SAM-SAM}}^{hb}\right)$	10
SAM-water hydrogen bonds distribution	$p\left(N_{\mathrm{SAM-water}}^{hb}\right)$	10
Water-water hydrogen bonds distribution	$p\left(N_{\mathrm{Water-water}}^{hb}\right)$	10

S2.1 Calculation of Water Structural Parameters

Water structural parameters were computed from interfacial water molecule positions obtained from the unbiased MD simulations. For all calculations, interfacial water molecules were defined as water molecules with centers of mass within 0.3 nm of the SAM-interface. This distance is equal to the z-dimension of the INDUS cavity. The SAM-water interface is defined in Section S1.5. Post-analysis of the MD trajectories was performed using Python scripts to calculate the interfacial water triplet angle distribution (Figure S7), OH bond orientational angle distribution (Figure S8), and the distribution of the number of hydrogen bonds between SAM-SAM (Figure S9), SAM-water (Figure S10), water-water (Figure S11), and all (Figure S12). Details on each of these parameters are provided in the following sections.

S2.1.1 Triplet angle distribution

The triplet angle distribution, $p(\theta)$, for interfacial water molecules was calculated according to Equation S5 (25, 26)

$$p(\theta) = \frac{1}{N_F} \int_t \left[\frac{1}{N_W(t)} \int_{|\boldsymbol{r}_j| < r_c} \int_{|\boldsymbol{r}_k| < r_c} \delta \left(\frac{\boldsymbol{r}_j}{|\boldsymbol{r}_j|} \cdot \frac{\boldsymbol{r}_k}{|\boldsymbol{r}_k|} - \cos\theta \right) d\boldsymbol{r}_j d\boldsymbol{r}_k \right] dt \qquad (S5)$$

$|\boldsymbol{r}_j|$ and $|\boldsymbol{r}_k|$ are the distances between a central water oxygen atom (denoted by index i) and two neighboring water oxygen atoms (denoted by the indices j, k) with positions $\boldsymbol{r}_j$ and $\boldsymbol{r}_k$ relative to the central water oxygen atom. The two inner integrals are evaluated with respect to all possible triplets of water molecules such that both distances are within the neighbor cutoff distance r_c = 0.33 nm. The delta function selects angles consistent with the given value of the triplet angle, θ, which is thus defined as the angle between atoms j, i, and k. The number of interfacial water molecules, $N_W(t)$, is time dependent because it varies with each simulation configuration. The outer integral is thus over simulation time and N_F is equal to the number of simulation frames. $p(\theta)$ values were histogrammed with θ ranging from 0°-180° in increments of 2°, leading to 90 water structural features that correspond to the probabilities of different values of θ.

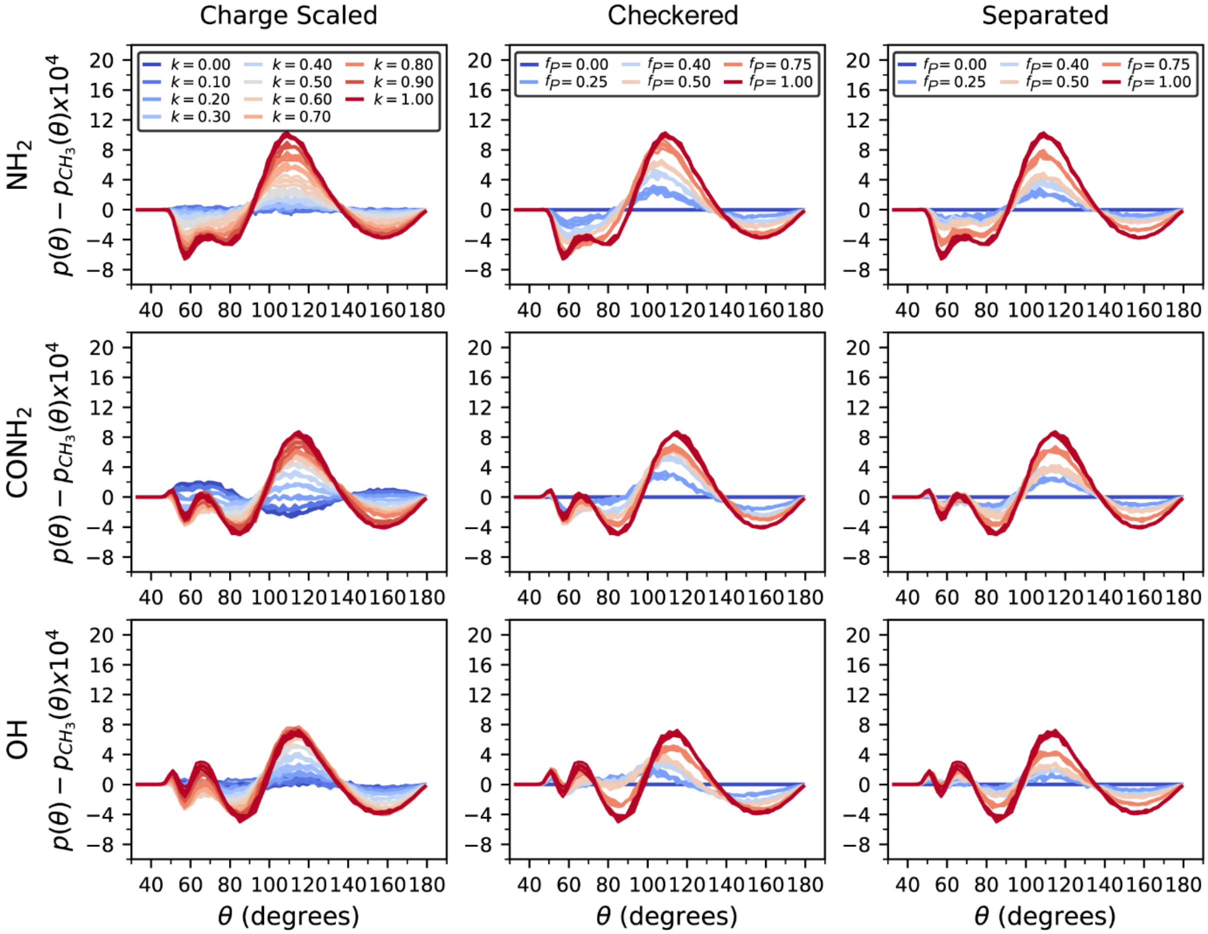

Figure S7. Triplet angle distributions for each end group chemistry and for charge-scaled, checkered, and separated SAMs. The distribution for SAMs with methyl end groups is treated as a baseline and subtracted from each curve to highlight deviations in $p(\theta)$ for different end group chemistries.

S2.1.2 OH bond orientational angle distribution

The water OH bond orientational angle distribution, $p(\phi)$, was calculated according to Equation S6. This metric quantifies the distribution of the angle of the OH bond vector with respect to the SAM surface normal.

$$p(\phi) = \frac{1}{N_F}\frac{1}{\sin\phi}\int_t \left[\frac{1}{N_W(t)}\int_i \delta\left(\frac{\boldsymbol{r}_i^{\mathrm{OH}}}{\left|\boldsymbol{r}_i^{\mathrm{OH}}\right|}\cdot\boldsymbol{n} - \cos\phi\right) d\boldsymbol{r}_W\right] dt \qquad (S6)$$

The outer integral is the same as in the triplet angle distribution (Equation S5). The integral is also normalized by $1/\sin\phi$ to correct for the contribution of the solid angle variation. The inner integral is evaluated for all interfacial water molecules for a given simulation configuration. $\boldsymbol{r}_i^{\mathrm{OH}}$ is an OH bond vector for the *i*th water molecule and $\boldsymbol{n}$ is the surface normal vector, which is always in the positive *z*-direction. The delta function selects angles consistent with the given value of the OH bond orientation angle, ϕ, which is thus defined as the angle between an OH bond vector and the surface normal. $p(\phi)$ values were histogrammed with ϕ ranging from 0°-180° in increments of 10°, leading to 18 water structural features that correspond to the probabilities of different values of ϕ.

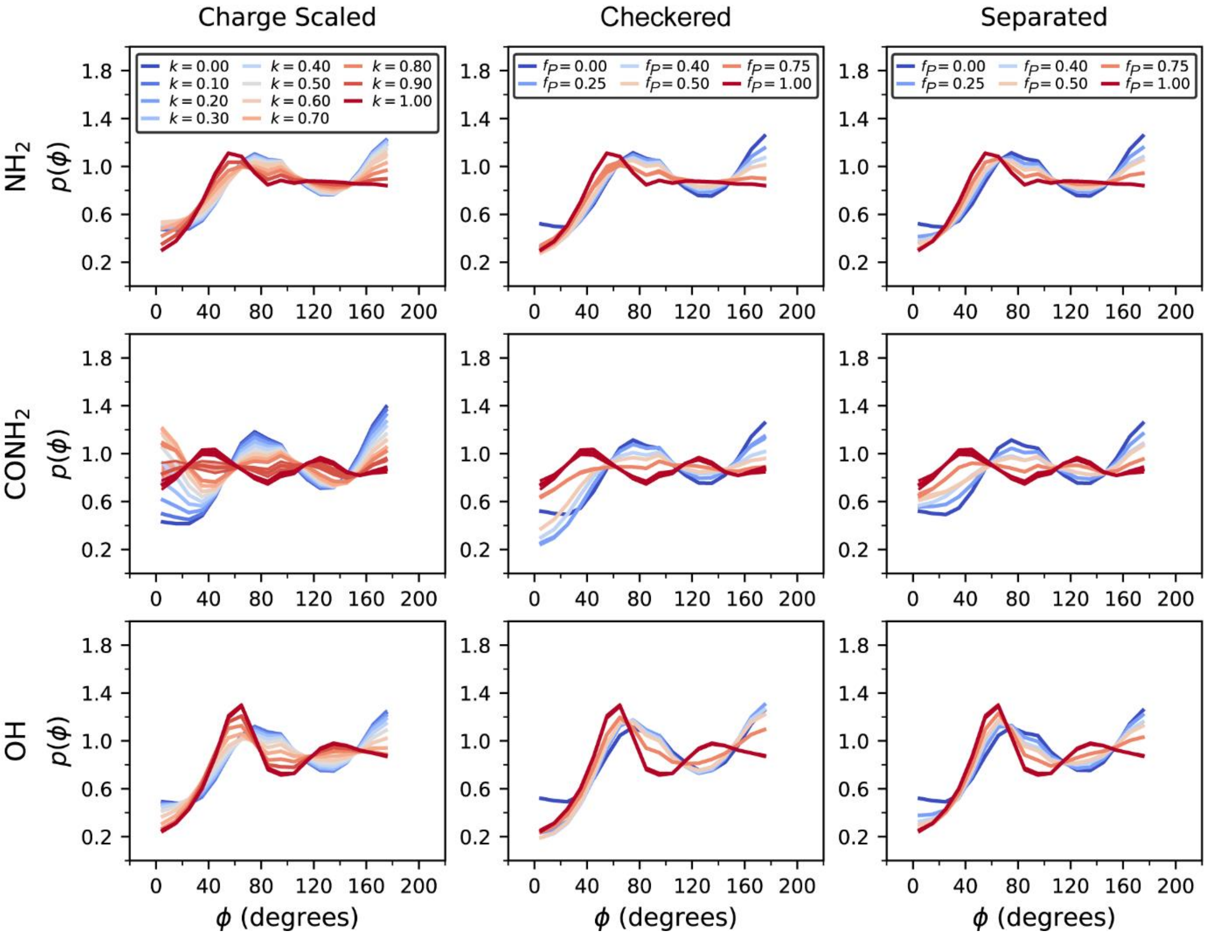

Figure S8. OH bond orientational angle distribution for each end group chemistry and for charge-scaled, checkered, and separated SAMs.

S2.1.3 Number of hydrogen bonds

The number of hydrogen bonds was calculated using the Luzar-Chandler geometric criteria (27). A hydrogen bond was counted when the distance between the donor and acceptor atoms was less than 0.35 nm and the hydrogen-donor-acceptor angle was less than 30°. For example, a hydrogen bond between two water molecules would be counted if the distance between the oxygen atoms is less than 0.35 nm and the angle between the hydrogen and two oxygen atoms is less than 30°. The number of hydrogen bonds formed between interfacial water molecules with each other (water-water), interfacial water molecules and SAM ligand atoms (SAM-water), SAM ligand atoms with each other (SAM-SAM), and all hydrogen bonds formed by SAM ligands and interfacial water molecules were separately calculated, contributing 4 water structural features. All values were time-averaged over the entire simulation trajectory.

S2.1.4 Hydrogen bond distributions

The distribution of the number of hydrogen bonds, N^{hb}, was calculated according to Equation S7

$$p\left(N^{\mathrm{hb}}\right) = \frac{1}{N_F} \int\limits_t \left[\frac{1}{N_W(t)} \int\limits_i \delta\left(N_i^{\mathrm{hb}} - N^{\mathrm{hb}}\right) d\boldsymbol{r}_W\right] dt \qquad (S7)$$

The outer integral is the same as in Equations S5-S6. The inner integral is evaluated for all hydrogen bonding groups (polar end group or interfacial water molecule) for a given simulation configuration. N_i^{hb} is the number of hydrogen bonds formed by the *i*th hydrogen bonding group. The delta function selects the number of hydrogen bonds consistent with the value of interest (with hydrogen bonds defined following the approach in Section 2.1.2). $p\left(N^{\mathrm{hb}}\right)$ values were histogrammed with N^{hb} ranging from 0-9 in increments of 1. Separate histograms were computed for SAM-SAM, SAM-water, water-water, and all hydrogen bonds (as defined in Section 2.1.2), leading to 10 water structural features for each different category of hydrogen bonds and 40 water structural features in total. Figures S9-S12 show the distributions for SAM-SAM, SAM-water, water-water, and all hydrogen bonds, respectively.

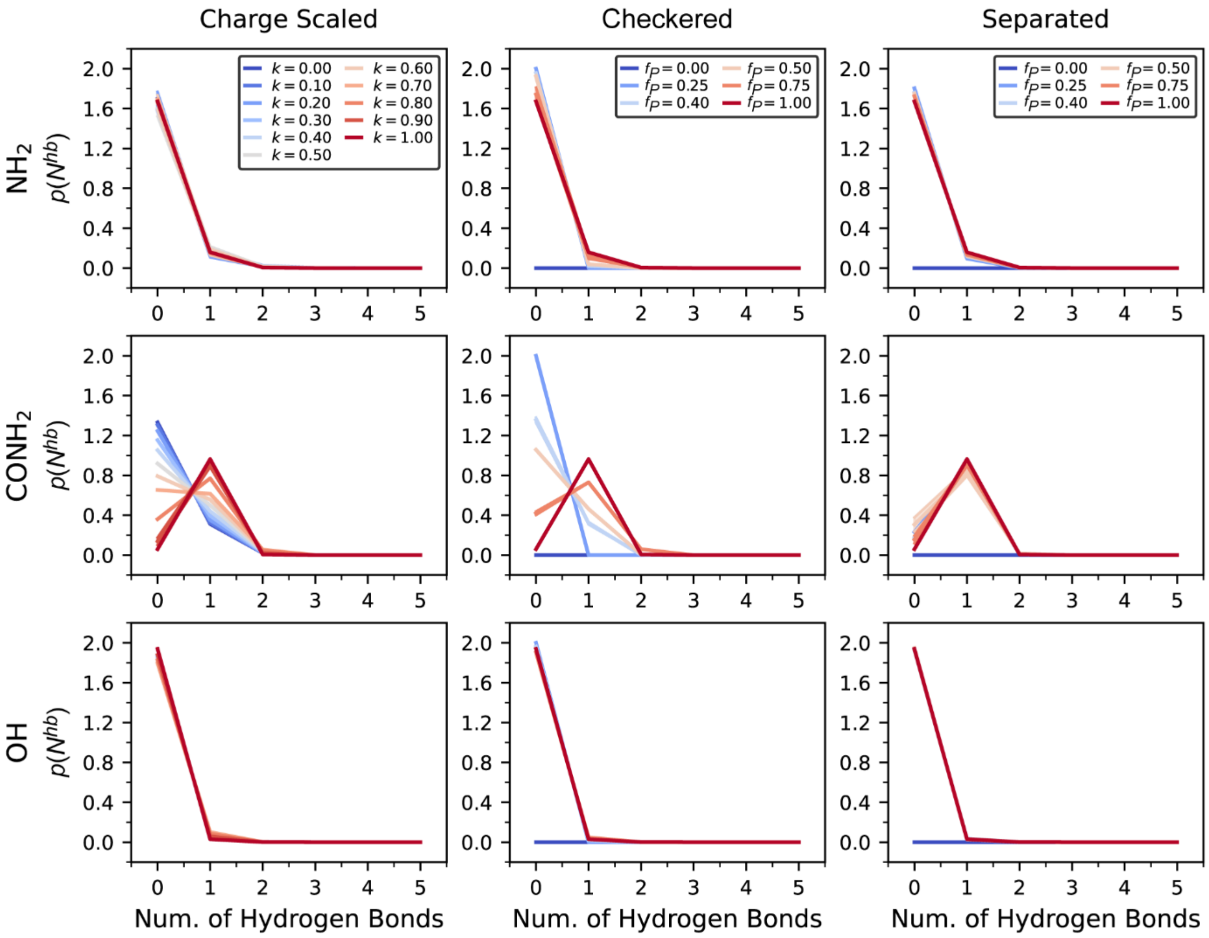

Figure S9. SAM-SAM hydrogen bond distributions for each end group chemistry and charge-scaled, checkered, and separated SAMs.

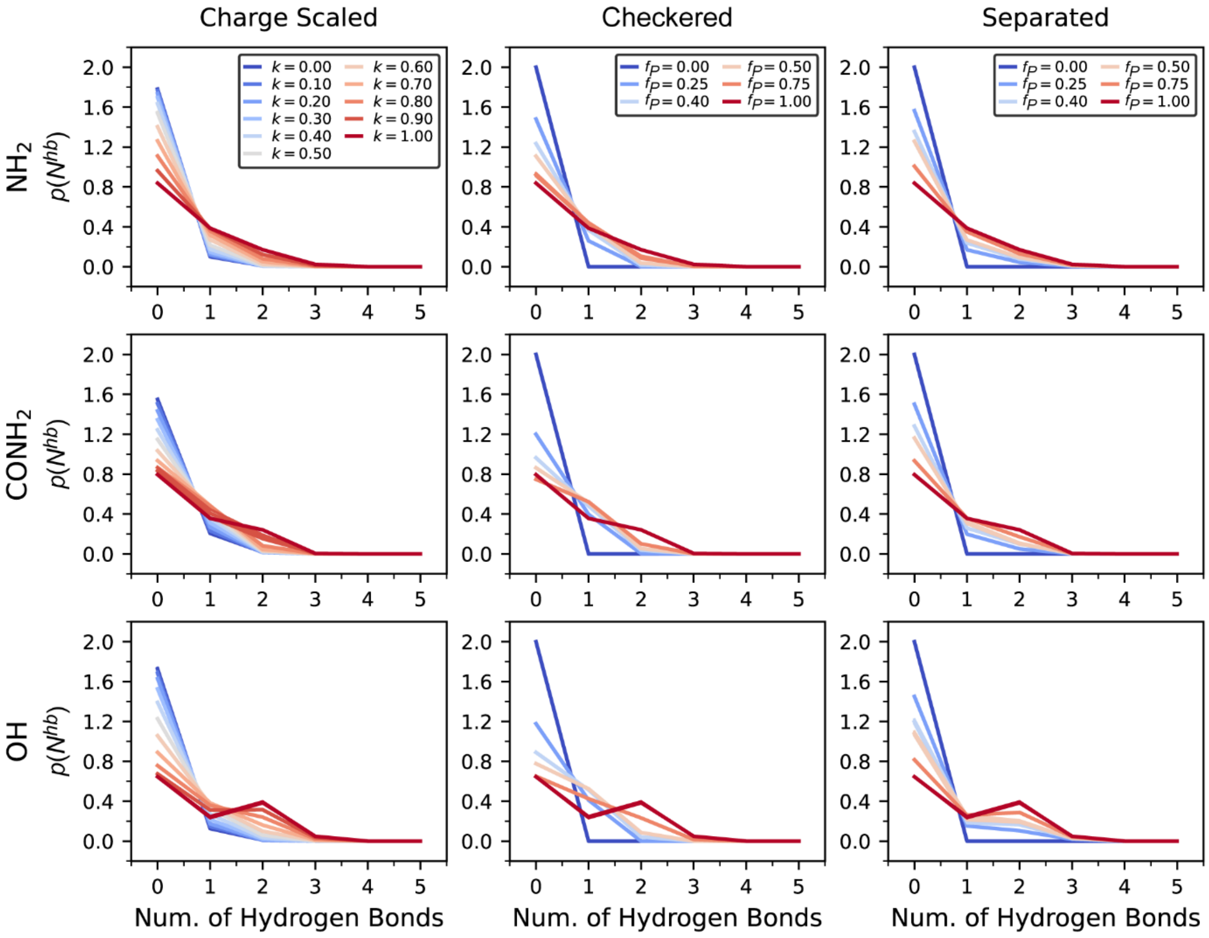

Figure S10. SAM-water hydrogen bond distributions for each end group chemistry and charge-scaled, checkered, and separated SAMs.

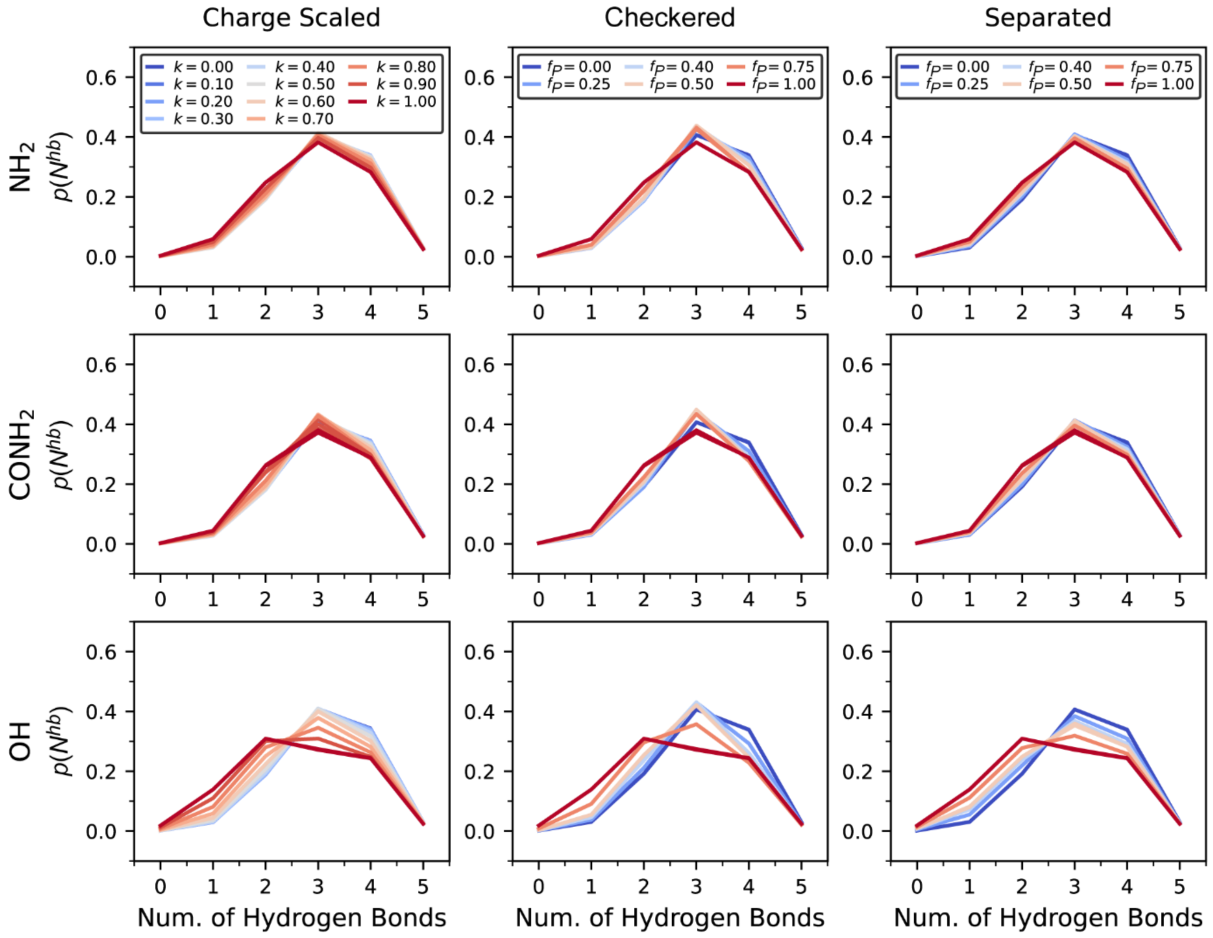

Figure S11. Water-water hydrogen bond distributions for each end group chemistry and charge-scaled, checkered, and separated SAMs.

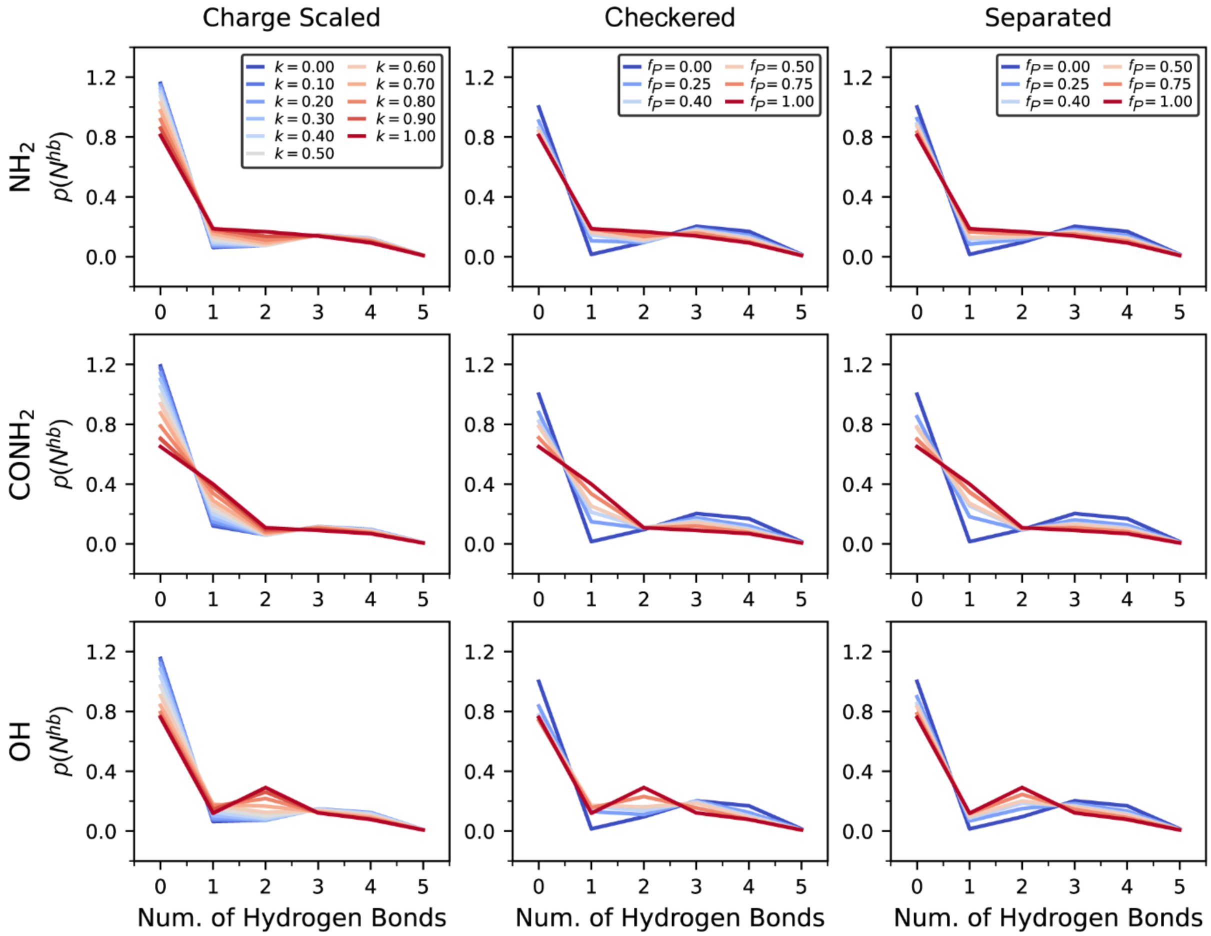

Figure S12. Total hydrogen bond distributions for each end group chemistry and charge-scaled, checkered, and separated SAMs.

S2.2 Data-Centric Feature Selection Models

The most important features for predicting hydration free energy values were determined using Lasso and Random Forest regression models. Details of each model and hyperparameter tuning are provided below. Three independent data sets were generated from the three independent INDUS and unbiased simulations performed (Data Set 1, 2, and 3). Each data set consisted of hydration free energy labels and water structure feature vectors for each of the SAMs. The hydration free energy labels assigned a relative hydrophobicity value to each SAM. The water structural feature vector was created by appending each water structural feature (*i.e.*, the numerical values described in Section S2) into a single vector. For each feature, values were standardized by subtracting the mean and dividing by the standard deviation of that feature's values across the training data. Our goal was to train regression models to predict the set of labels using the set of feature vectors as input as described in the following sections.

S2.2.1 Removal of Correlated Features

Before training the regression models, the water structural feature vectors were pre-processed to remove correlated features. We calculated the pairwise Pearson's correlation coefficient between all features for each data set. The second feature in the pair was removed if the coefficient absolute value was larger than 0.9. The removed feature is essentially arbitrary since the high correlation implies that information is preserved. The absolute value was used to remove both positively and negatively correlated features. If a feature appeared in one data set but not the others, that feature was retained for analysis of all data sets. This procedure reduced the total number of features from 152 to 45 features. Table S2 lists each of the features used to train the regression models.

Table S2. Set of features retained after removing highly correlated features (each associated with different structural parameters) used to train the regression models. Some features were only present in the specific data set indicated by the subscripts.

Retained Features Used for Model Training	
Triplet Angle Probabilities, $p(\theta)$	
42, 44, 46[1], 48, 60, 90, 110, 136, 138, 140[1], 178	
OH Bond Orientation Angle Probabilities, $p(\phi)$	
25[2,3], 35, 45[3], 65, 105, 115[2,3], 135, 145, 155	
Number of Hydrogen Bonds, N^{hb}	
All, SAM-SAM, SAM-Water	
Hydrogen Bond Probabilities, $p(N^{\mathrm{hb}})$	
All:	0, 1, 2[2,3], 3, 6
SAM-SAM:	0, 2, 3, 4, 5
SAM-Water:	0, 1, 2[3], 3[1], 4, 5
Water-Water:	0[2], 2[1], 3, 5, 6, 7

[1] Feature is only present for Data Set 1
[2] Feature is only present for Data Set 2
[3] Feature is only present for Data Set 3

S2.2.2 Lasso Regression

We wish to predict the label (μ_v) using the minimum number of structural features. To do this, we formulated the problem using a Lasso regression model. Lasso regression, or L1 regression, is based on linear regression (Equation S8) with a regularization term added to the cost function (Equation S9)

$$\mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} \tag{S8}$$

$$\text{cost} = \sum_{i=1}^{N} \left(\mu_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{p} \left| \beta_j \right| \tag{S9}$$

μ_i is the hydration free energy of the *i*th SAM, x_{ij} is the value of the *j*th water structural feature computed for the *i*th SAM, β_j is the *j*th coefficient, or weight, of the linear model, α is the regularization parameter, N is the number of SAMs used in the model training (N = 58), and p is the total number of water structural features (*i.e.*, the size of the feature vector) input into the model (p = 45). The first term in Equation S9 is the sum of squared errors and the second term is the L1 penalty on the coefficient values. The regularization term, because of the L1 penalty, forces coefficients with values below a certain threshold to zero. This enables prediction of labels using a small number of features with coefficients for all other features reduced to 0; such features are unimportant for model predictions. The Lasso model has 46 model parameters: each linear coefficient ($\beta_0, \beta_1, \ldots, \beta_{45}$) and 1 hyperparameter, the regularization parameter (α). The linear coefficients were optimized by minimizing the cost function (Equation S9) using the gradient descent algorithm built into the Scikit-Learn Lasso regression function (28). The regularization term, α, was tuned by iteratively testing values for α between 0.1 and 10 with a step size of 0.1. All SAMs were used as input to the Lasso model and the linear coefficients were determined for each tested value of α. The root-mean-squared error (RMSE) between the predicted and INDUS calculated hydration free energies was computed as a function of α (Figure S13A). Increasing α also influences the number of features not eliminated by Lasso (*i.e.*, the number of non-zero coefficients). Figure S13B shows the number of features with non-zero coefficients as a function of α.

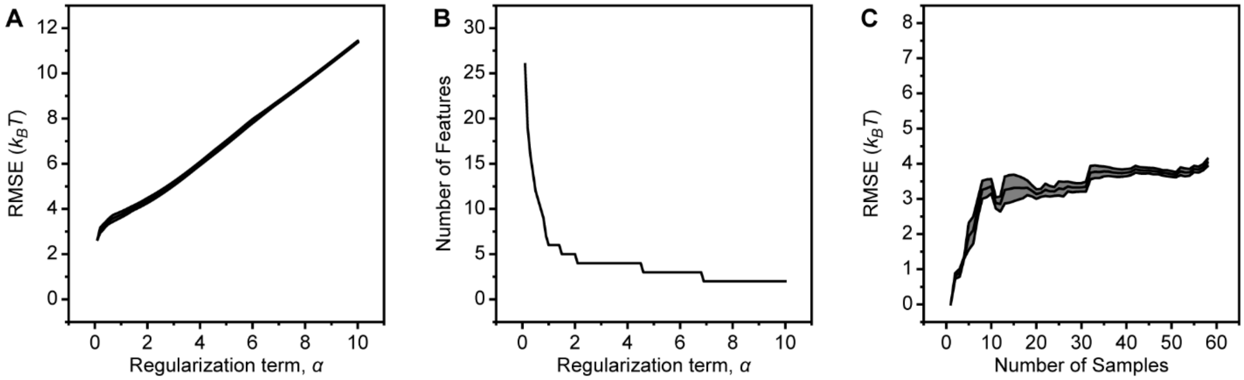

Figure S13. Lasso hyperparameter optimization. (A) Root-mean-square error (RMSE) as a function of regularization term (α). (B) Number of features with weights greater than zero as a function of α. (C) RMSE as a function of the number of SAMs in the training set (samples, N). Error bars in each plot are the standard deviation calculated from the three independent data sets.

Based on these plots, we chose a value of α = 1.45 which eliminates all but 5 features with an RMSE of approximately 4 k_BT. We also determined the minimum value of N needed to accurately train the Lasso model by iteratively training the model with 1 to 58 SAMs. Figure S13C shows how RMSE varies with increasing N. The lower RMSE values observed for N < 10 indicate a highly overfit model, whereas model convergence around N = 35 indicates a better fit. Knowing the model is fitted well for N > 35 is important because we performed 5-fold cross validation using $N \approx 45$ in each fold. We note that increasing the value of α to 3.0 eliminates a feature, but also increases the RMSE by about 10% as described in more detail below. We elected to retain 5 features to reduce the RMSE and because all 5 features are physically interpretable; moreover, this number of features is still small enough to avoid overfitting. Figure S14A shows the parity plot of labels predicted by the Lasso model versus INDUS-calculated hydration free energies determined by training on the full data set (N = 58) with α = 1.45, then also predicting on the same full data set. The weights determined by Lasso by training on the full data set are shown in Figure S14B. Full descriptions for each feature are detailed in the main text.

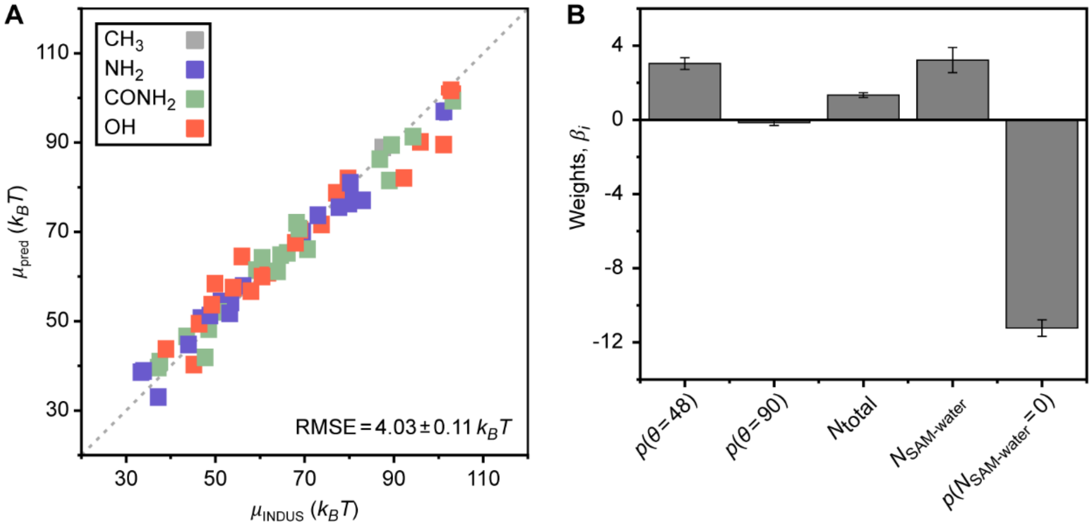

Figure S14. (A) Parity plot of the Lasso-predicted hydration free energies compared with the INDUS calculations. (B) Bar plot comparing the weights of the linear coefficients determined by Lasso regression. Error bars are the standard deviation from the three independent data sets.

S2.2.3 5-Fold Cross Validation of Lasso Regression Model

We used 5-fold cross validation to ensure that the features identified by Lasso were robust. Cross validation was performed by splitting the data set into five folds, each containing a training set of roughly 46 SAMs and a validation set of roughly 12 SAMs not present in the training set. The five folds were selected such that each SAM was held out of the training set and included in the validation set exactly once. For each fold, a Lasso regression model with a set value of α was independently fit on the training set and used to predict hydration free energies for the SAMs in the validation set. As was done for the linear regression model, the same five folds were used with ~46 SAMs for training and ~12 SAMs for validation. The points in the parity plot in Figure S15A are the model predictions for the validation sets for α = 1.45. The weights of the features for all five folds are shown in Figure S15B. The same five important features identified when the Lasso regression model is trained on all SAMs are also identified during 5-fold cross validation, confirming that these five features are universal and not specific to a single training set. The probability of 5 water-water hydrogen bonds also appears as a very small feature in some of the folds. This feature is eliminated when considering all SAMs (Figure S14). Error bars were calculated as the standard deviation from applying this procedure to all three data sets. Errors bars not visible are small than the points. We also performed 5-fold cross-validation using α = 3.0, which eliminates a feature. We find that the cross-validation RMSE for α = 3.0 increases by about 10% (from 4.86 k_BT to 5.33 k_BT) while eliminating the probability of observing a triplet angle of 90° as a feature in all five folds (Figure S15C), which is consistent with this feature having the lowest weight of the five features identified using α = 1.45. Because this probability has a clear physical interpretation (as described in the main text and in section S3.3, below), we opted to retain five features for all models in the main text in order to reduce the RMSE.

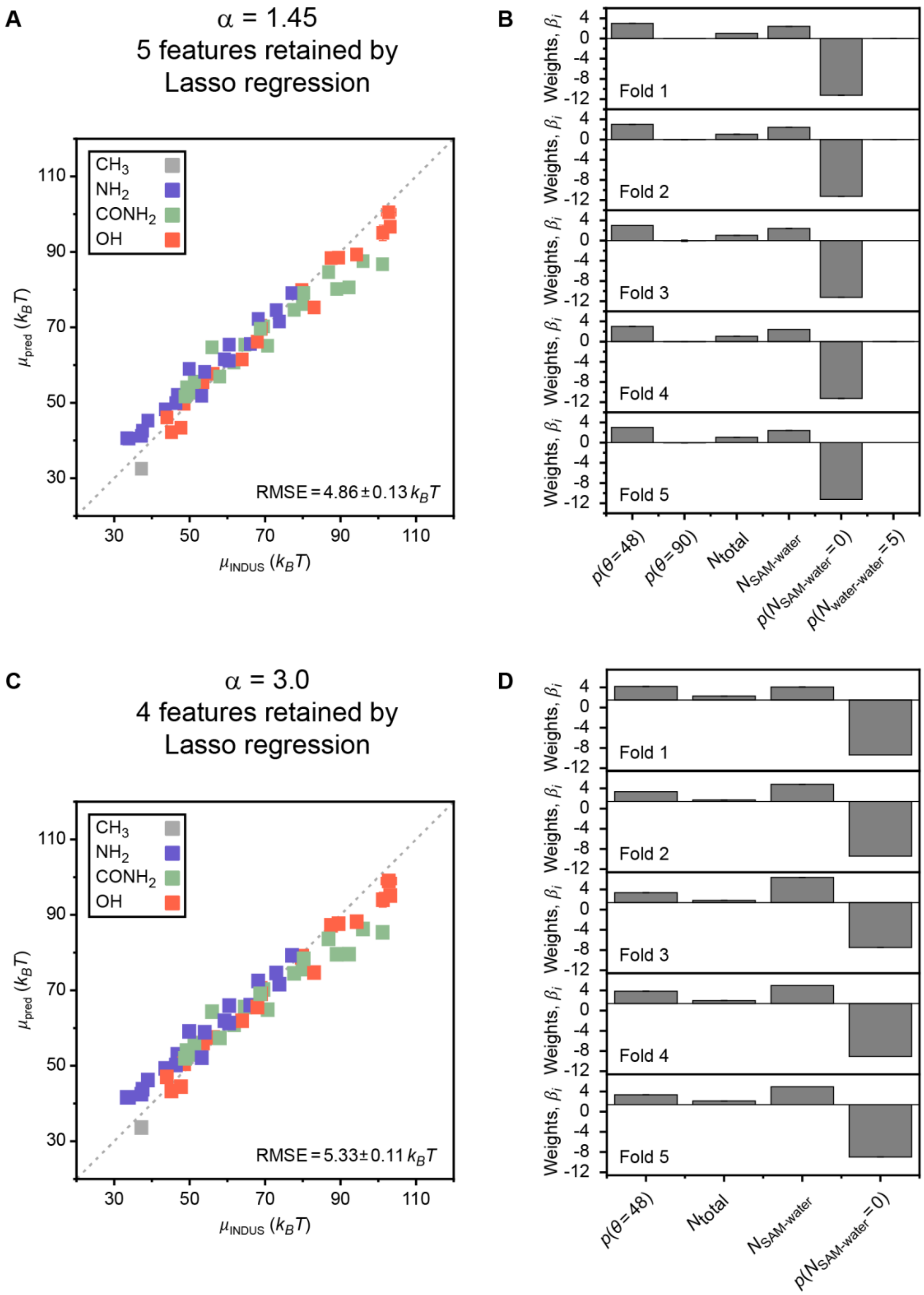

Figure S15. Parity plot comparing the predicted hydration free energies determined by 5-fold Lasso regression and INDUS estimated values for (A) α = 1.45 and (C) α = 3.0. Bar plots for each of the folds comparing the nonzero coefficient weights for (B) α = 1.45 and (D) α = 3.0.

S2.2.4 5-Fold Cross Validation of the Linear Regression Model

In the main text, Figure 2 shows the parity plot comparing INDUS-calculated and linear regression-predicted hydration free energies. Each data point is that plot corresponds to the validation set prediction obtained during 5-fold cross validation of the linear regression model using the minimum set of features identified by Lasso. We performed 5-fold cross validation to address model accuracy on data not seen by the model during training (to assess model generalizability). Cross validation was performed using the same procedure described in the preceding section for Lasso regression. For each fold, a linear regression model (using the five Lasso-identified features as input) was independently fit on the training set and used to predict hydration free energies for the SAMs in the validation set. Each point in Figure 2 of the main text corresponds to the hydration free energy predicted for the SAM for the fold in which it was included in the validation set. That is, Figure 2 only shows predictions for SAMs not included in model training, highlighting the potential of the model to generalize to unseen data.

S2.2.5 Correlation between Identified Features

Although highly correlated features were removed prior to model training, the threshold value of the Pearson's correlation coefficient used to remove highly correlated features was itself high (0.9). We thus further computed correlation coefficients between the five features identified by Lasso regression. Figure S16 shows a matrix of pairwise Pearson's correlation coefficient for these five features. Correlation coefficients are all low or modest (absolute values < 0.6-0.7) with the exception of the coefficient between the probability that an interfacial water molecule forms zero SAM-water hydrogen bonds, $p(N_{\mathrm{SAM-water}} = 0)$, and the average number of SAM-water hydrogen bonds, $N_{\mathrm{SAM-water}}$, which exhibit a higher correlation coefficient of -0.84. However, both features have a high weight in the Lasso model, suggesting that their correlation is sufficiently low that each feature contributes useful information to model predictions.

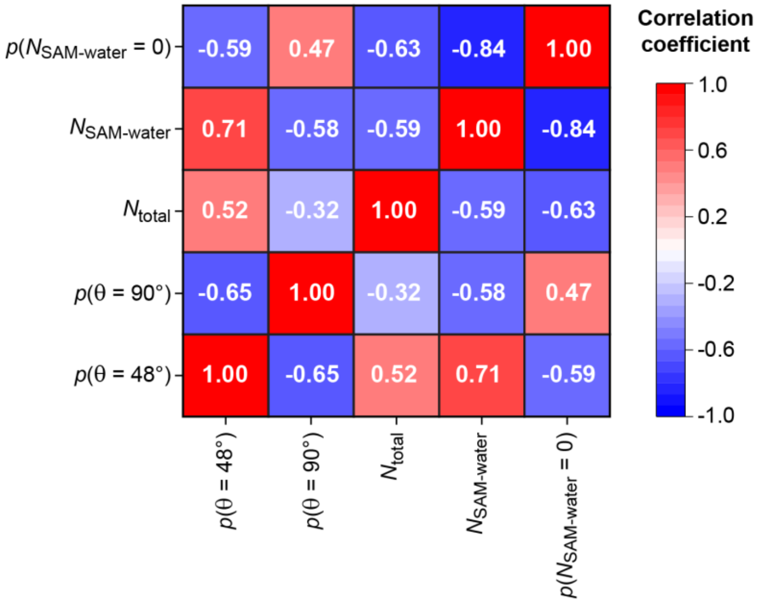

Figure S16. Matrix of Pearson's correlation coefficients between 5 features identified by Lasso regression.

S2.2.6 Comparison of Identified Features to Mole Fraction of Polar Groups

To demonstrate that the water structural features identified by Lasso regression are uncorrelated with mole fraction of polar end groups on the surface (f_P), we calculated the Pearson correlation coefficient between (f_P) and the 5 features identified by Lasso regression. Table S3 lists the correlation coefficients. The magnitudes of all coefficients are less than 0.65 and three features have a magnitude less than 0.1; from this analysis, we conclude that only the total number of hydrogen bonds (N_{total}) is weakly correlated with f_P, as might be expected, whereas the other structural features are all uncorrelated.

Table S3. Correlation of water structural features with mole fraction of polar groups (f_P).

Feature	**Pearson's correlation coefficient with f_P**
N_{total}	-0.620
$p(\theta = 90°)$	-0.370
$N_{\text{SAM-water}}$	0.097
$p(\theta = 48°)$	-0.079
$p(N_{\text{SAM-water}} = 0)$	-0.062

S2.2.7 Random Forest Regression

Random Forest regression was used to further validate the features identified by Lasso regression. Random Forest is a nonlinear model that applies an ensemble of decision trees to predict hydration free energies. The Random Forest algorithm builds M decision trees then makes a prediction by averaging the predictions of the individual trees. The variance of the decision trees is reduced by using bootstrap aggregation (bagging). The bagging process generates an ensemble of bootstrapped sample data sets. A bootstrapped sample is a random sample of the data taken with replacement (*i.e.*, after a data point is selected for a sample, it remains available for further selection). Once the M trees are built, their predictions are averaged to obtain the final prediction (Figure S17A). Figure S17B shows prediction RMSE as a function of the number of trees. From inspection, it was determined that 100 trees were sufficient for a converged prediction. During the bagging process, the decision trees were built using split-variable randomization where each time a split is performed, the search for the split variable is limited to a random subset of m of the original p features. Figure S17C shows the RMSE as a function of the fraction of features for each split. It was determined that limiting a split to 60% gave satisfactory performance.

Random Forest alone does not eliminate nonimportant features. Consequently, we added a recursive feature elimination (RFE) wrapper to the Random Forest workflow. RFE eliminates features by recursively determining the importance of each feature from the initial set of features, then iteratively removing the least important features until the stopping criterion is met (*e.g.*, minimum number of features). Figure S17D shows the RMSE as a function of the number of features retained by the RFE algorithm. The RMSE in Figure S17D has a minimum at approximately five features, indicating that five features are enough to accurately predict hydration free energies. Figure S17E shows the prediction accuracy as a function of the maximum tree depth.

Tree depth is the number of times a split (decision) with the data can be done (*i.e.*, more splits equal larger tree depth). Figure S17F shows prediction accuracy versus number of SAM samples input as training data. Like the Lasso model, the Random Forest model is highly overfit when the number of samples is small (Figure S13C). The model converges when more than 20 SAM samples are used. This indicates that >20 SAM samples are needed to accurately predict hydration free energies with this model.

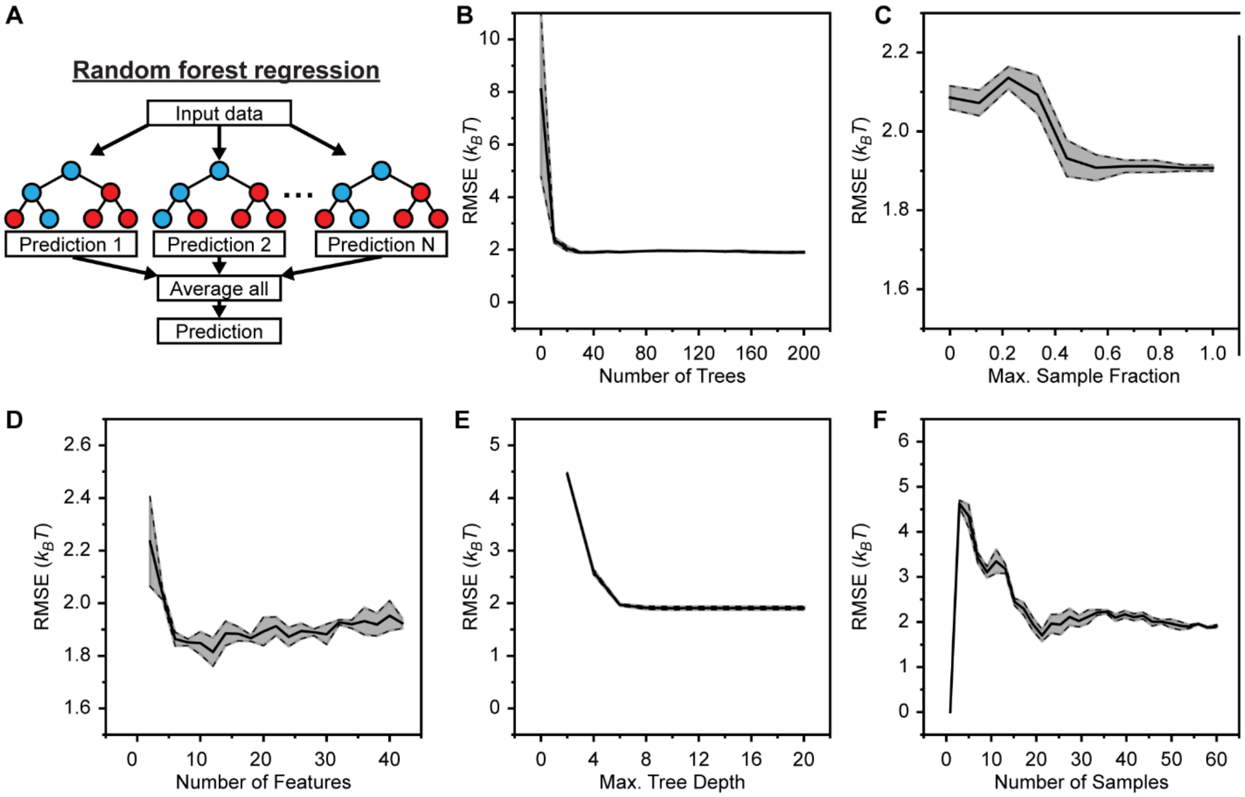

Figure S17. (A) Schematic of Random Forest model. Model accuracy versus (B) number of trees, (C) Maximum sample fraction in split, (D) number of features retained, (E) maximum tree depth, and (F) number of samples in training data.

All SAMs were used to train the Random Forest model. Figure S18A shows the comparison between the model predicted hydration free energies and the INDUS estimates. Figure S18B shows importance metrics for each feature retained by the RFE algorithm. Three of the features found by the Random Forest model are also found by the Lasso model suggesting $p(\theta = 48°)$, $N_{\mathrm{SAM-water}}$, and $p(N_{\mathrm{SAM-water}} = 0)$ are robust features across the two models. The $p(N_{\mathrm{water-water}} = 5)$ feature found as third-most important in the Random Forest model is also found as a nonzero feature in some of the folds when 5-fold cross validation was performed using Lasso, indicating that this feature is slightly robust across models (Figure S18B). The N_{total} feature identified by Lasso and the $p(N_{\mathrm{total}} = 0)$ features identified by Random Forest are highly correlated with an absolute Pearson's correlation coefficient of 0.89. The criteria to eliminate features was 0.9. The strong correlation between the features suggests each of these features encodes similar information in the prediction.

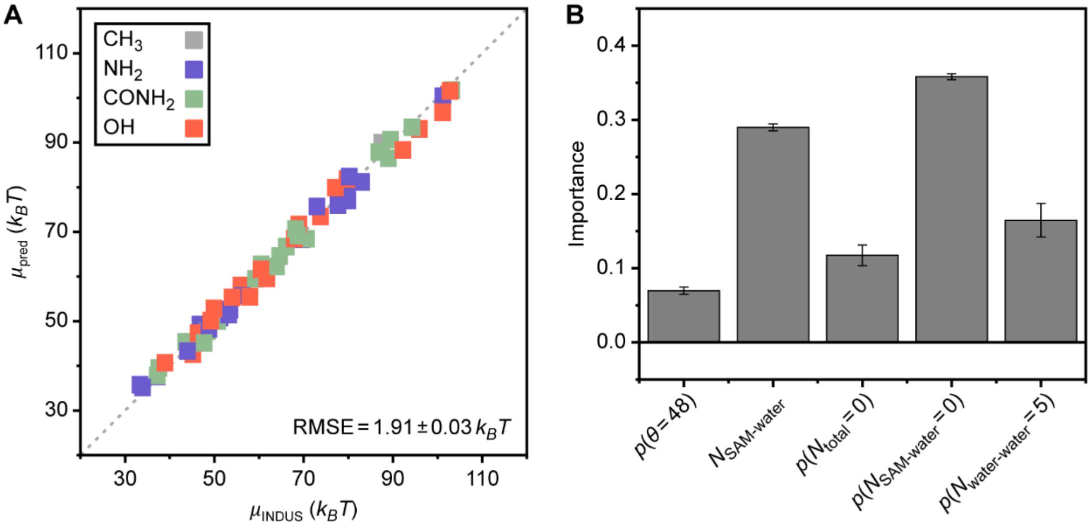

Figure S18. (A) Parity plot of Random Forest model trained and bagged using all SAMs. (B) Bar plot comparing features not eliminated by the RFE algorithm.

Important features were also determined using 5-fold cross validation of the Random Forest model (Section S2.2.3). Figure S19A shows a parity plot comparing the Random Forest-predicted and INDUS-measured hydration free energies. The predicted values were computed from the validation set during 5-fold cross validation. Figure S19B compares the averaged features weights across the 5 folds. The same five important features arise as when the model is trained on all SAMs providing an additional robustness check to these five features being universal across the surfaces. The error bars are the standard deviation from calculations performed with each of the three independent data sets.

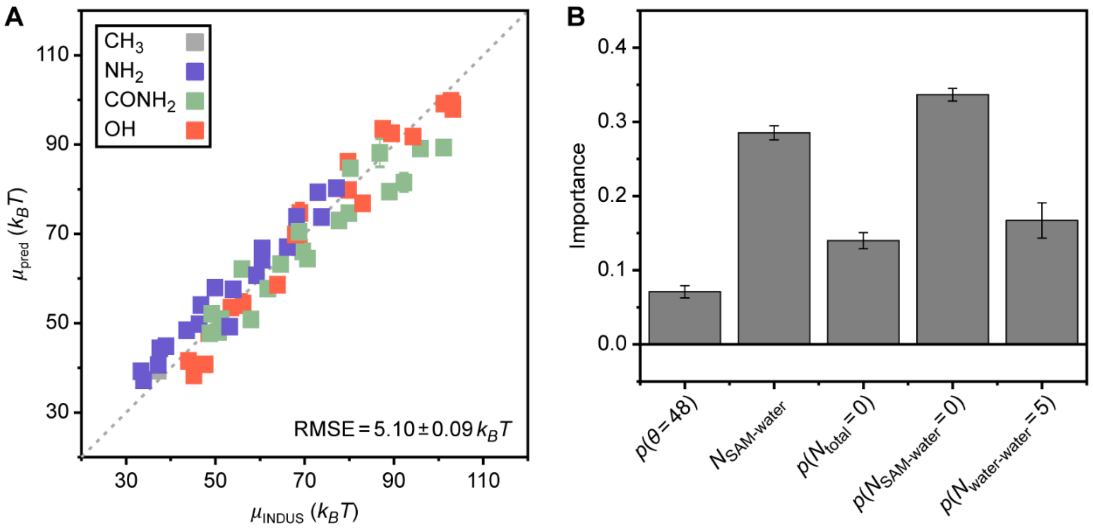

Figure S19. (A) Parity plot comparing hydration free energies predicted by 5-fold Lasso regression (with each point representing a validation set prediction) and hydration free energies calculated with INDUS. (B) Bar plot comparing feature importance in the Random Forest model. Error bars represent standard deviations across the five folds.

S3: Additional Results

S3.1 Additional INDUS Results and Discussion

Figure S20 shows μ_v as a function of the scaling factor, k, for the charge-scaled SAMs. As expected, μ_v increases with increasing k in all cases, thereby sampling a continuous range of values.

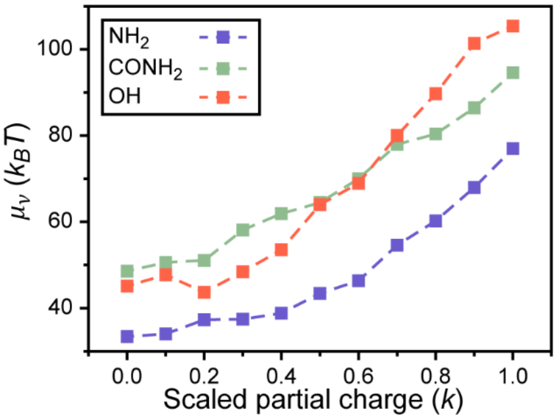

Figure S20. Hydration free energies (μ_v) calculated with INDUS as a function of the multiplicative scaling factor used to modulate the end group partial charges in charge-scaled SAMs.

S3.2 Test of Linear Regression Model on Unseen Chemically Heterogeneous SAMs

The ability of the linear regression model trained in the main text to generalize to unseen chemically heterogeneous SAMs was tested by predicting values of μ_v for 153 SAMs obtained from Ref. (29). These SAMs contain the same set of polar end groups studied in the main text but in a broader range of mole fractions and patterns than studied in the 58 SAM training data set. Specifically, there are three sets of 51 mixed SAMs, each containing a mixture of methyl-functionalized and either amine-, amide-, or hydroxyl-functionalized alkanethiol ligands (akin to the chemically heterogeneous SAMs used for model training). Each SAM includes a central 4×4 patch of polar and nonpolar ligands surrounded by an additional 48 nonpolar ligands; the relative fraction of polar and nonpolar ligands and their spatial positions within the patch were randomly selected to generate 50 distinct SAMs (with 1 additional SAM including a patch containing only polar ligands). For each SAM, a 10 ns unbiased MD simulation was performed following the protocol from Section S1.2 and used to compute the 5 water structural features that were identified from Lasso regression following the same approach as all SAMs in the training data. Each feature was standardized using the same mean and standard deviation obtained from the 58 SAM training data set, then μ_v values were predicted for all SAMs using the linear regression model with the weights obtained from the training data. No parameters were refit specifically to the new set of SAMs. INDUS-calculated μ_v values were taken from Ref. (29), in which μ_v was calculated following the same procedure described in Section S1.3.

Figure S21 shows the parity plot between INDUS-calculated and predicted μ_v values for the test set of unseen chemically heterogeneous SAMs. The RMSE of the linear regression model predictions for the test set is 5.21 k_BT. This RMSE is only slightly larger than the RMSE for cross-validation predictions of the 24 chemically heterogeneous SAMs with separated or checkered patterns (4.57 k_BT) even though there is a much larger number of SAMs in the test set than in the training set, indicating the ability of the regression model to generalize to new patterned SAMs. The slight increase in the RMSE may be because the patterned chemically heterogeneous patch in these SAMs is embedded within a perimeter of purely nonpolar ligands, which may contribute to slight variations in water order parameters for those regions that do not reflect the patterned surface. Nonetheless, the encouraging RMSE suggests that the trained regression model is correctly capturing variations in the interfacial water structure across this wide range of surfaces.

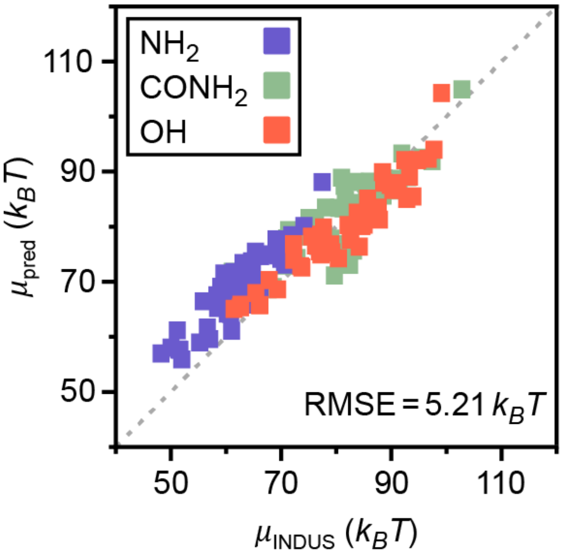

Figure S21. Parity plot comparing hydration free energies (μ_v) predicted from multivariate linear regression to those calculated by INDUS for a test set of 153 chemically heterogeneous SAMs not seen during model training.

S3.3 Physical Basis for the Importance the θ = 90° Feature

We further examined the significance of the $p(\theta = 90°)$ feature to understand how this feature distinguishes SAM patterns. Our analysis was motivated by two observations from past simulation studies: (1) $p(\theta = 90°)$ is a maximum in the triplet angle distribution of an ideal gas, suggesting that large values of $p(\theta = 90°)$ indicate random, vapor-like structural arrangements

(25, 26); and (2) interfacial water molecules near hydrophobic surfaces exhibit larger density fluctuations compared to more hydrophilic surfaces and form a liquid-vapor-like interface (8, 22, 30). These two observations motivate our hypothesis that the water molecules near the large nonpolar domains of the separated SAMs will have higher values of $p(\theta = 90°)$ because water molecules near large nonpolar domains exhibit behavior more similar to a liquid-vapor interface. To test this hypothesis, we measured $p(\theta = 90°)$ for each interfacial water molecule and histogrammed these values based on the *x*-position of the molecule in the simulation box; note that ligands in separated SAMs were placed such that the polar and nonpolar ligands were on opposite ends of the *x*-axis. This approach thus quantifies spatial variations in $p(\theta = 90°)$ for the nonpolar and polar domains of the separated SAMs. Figure S21 shows histogrammed values of $p(\theta = 90°)$ as a function of x for the separated and checkered SAMs. Each value is normalized by the equivalent value computed for a purely nonpolar SAM ($f_P = 0$) to highlight deviations from a large nonpolar domain. Figures S21 reveals that there is minimal spatial variation in $p(\theta = 90°)$ for all checked SAMs, reflecting the uniform distribution of polar and nonpolar end groups for this pattern. Conversely, the *x*-positions associated with nonpolar domains in the separated SAMs have higher probabilities of $p(\theta = 90°)$, with values close to 1.00 (*i.e.*, similar to values near a nonpolar SAM). These increased values lead to an average increase in $p(\theta = 90°)$ for separated SAMs compared to checkered SAMs for nearly all SAMs in the data set, explaining the ability of this feature to distinguish checkered and separated SAMs.

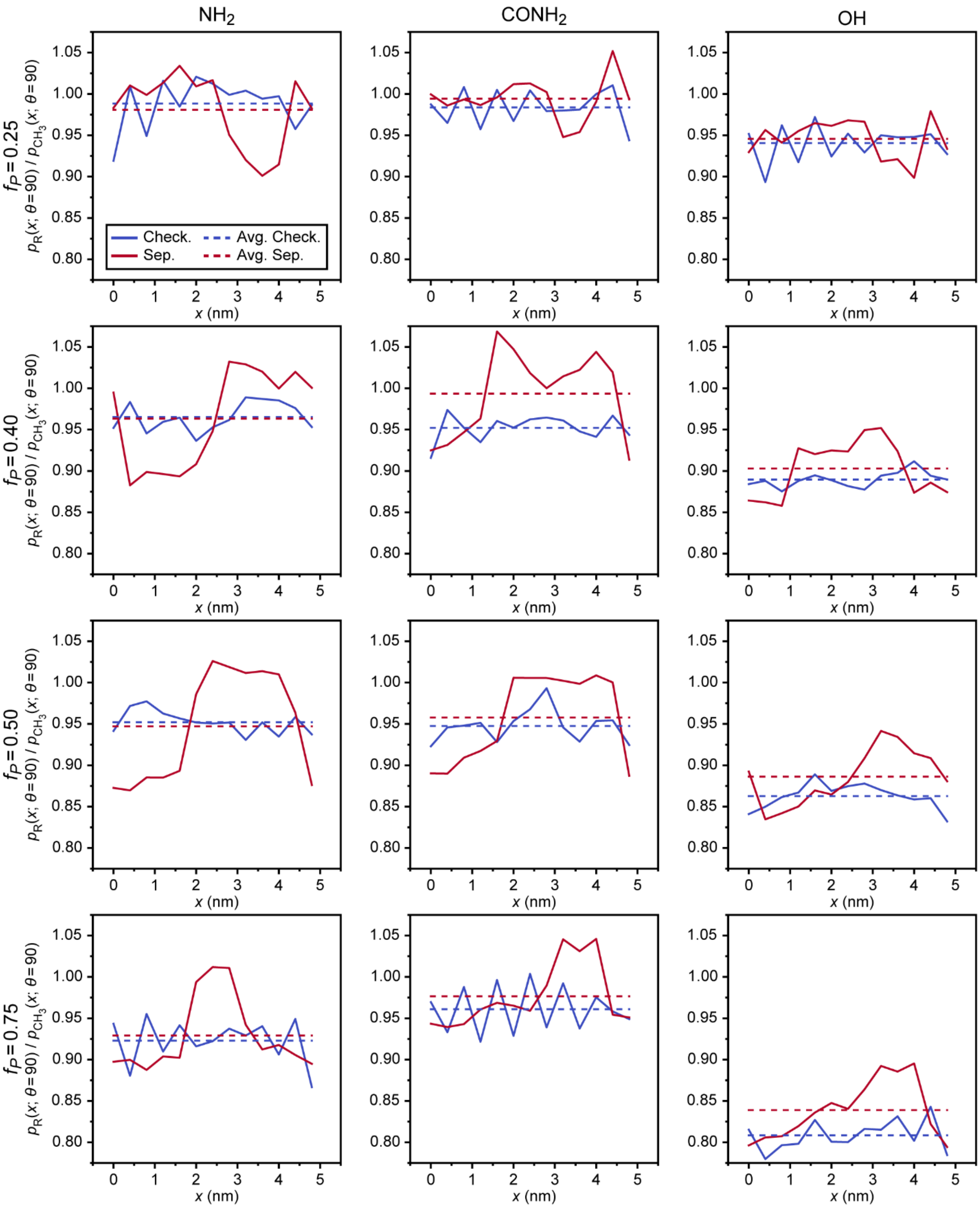

Figure S22. Probability of water molecules with a triplet angle of 90° as a function of x-position. Values are normalized by those of a nonpolar SAM ($f_P = 0$). Average values are shown as horizontal dashed lines.

S3.4 Coordination Number Probability Distributions for Separated SAMs

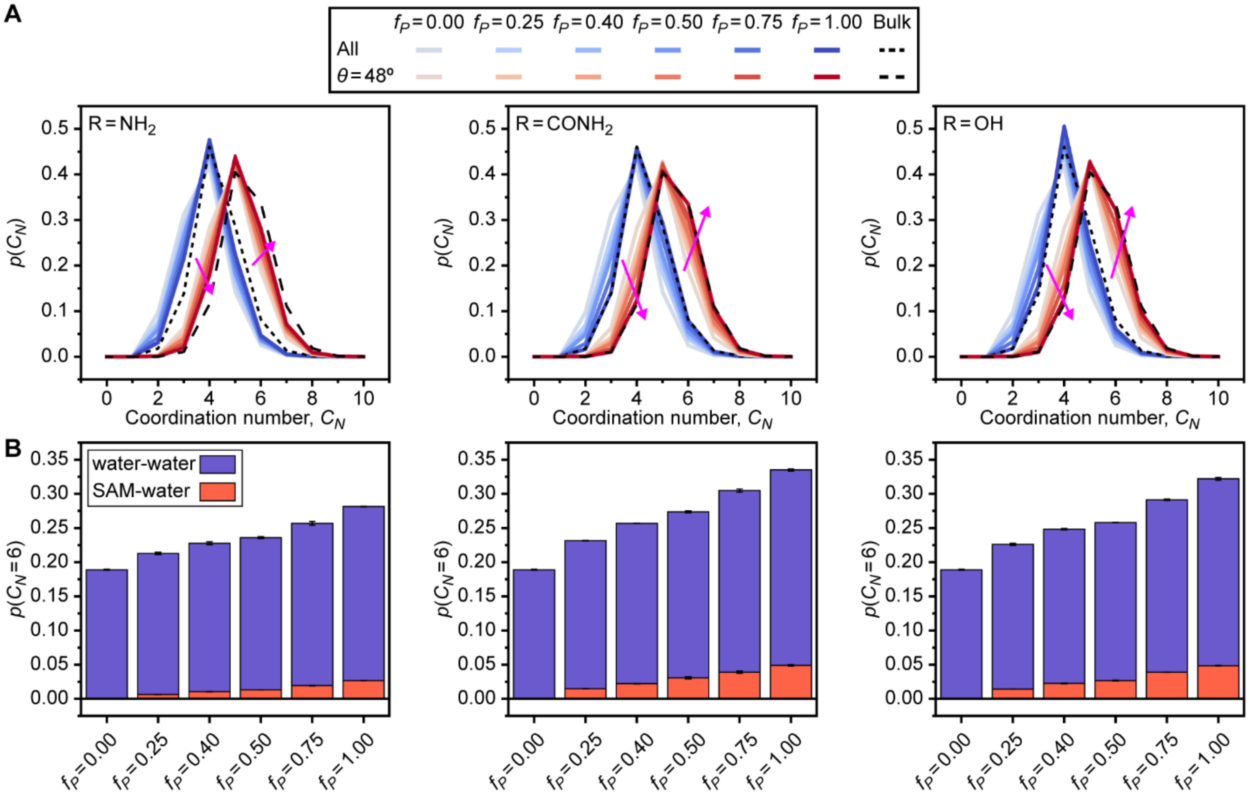

Figure S23. (A) Water coordination number (C_N) probability density functions for all interfacial water molecules (blue lines) and only interfacial water molecules with a triplet angle of 48° (red lines). Bulk water probability density functions for all water molecules (dotted line) and water molecules with a triplet angle of 48° (dashed line) are included for reference. Shifts with increasing f_P are indicated by the purple arrows. (B) Probability density function values for C_N = 6. Stacked columns indicate the contributions from water-water coordination (blue columns) and SAM-water coordination (red columns). A and B both consider only separated SAMs.

References

1. Giovambattista N, Debenedetti PG, & Rossky PJ (2007) Effect of surface polarity on water contact angle and interfacial hydration structure. *J Phys Chem B* 111(32):9581-9587.
2. Kelkar AS, Dallin BC, & Van Lehn RC (2020) Predicting Hydrophobicity by Learning Spatiotemporal Features of Interfacial Water Structure: Combining Molecular Dynamics Simulations with Convolutional Neural Networks. *J Phys Chem B* 124(41):9103-9114.
3. Saha K, Agasti SS, Kim C, Li X, & Rotello VM (2012) Gold nanoparticles in chemical and biological sensing. *Chem Rev* 112(5):2739-2779.
4. Hakkinen H (2012) The gold-sulfur interface at the nanoscale. *Nat Chem* 4(6):443-455.
5. Porter MD, Bright TB, Allara DL, & Chidsey CED (1987) Spontaneously organized molecular assemblies. 4. Structural characterization of n-alkyl thiol monolayers on gold by optical ellipsometry, infrared spectroscopy, and electrochemistry. *Journal of the American Chemical Society* 109(12):3559-3568.
6. Love JC, Estroff LA, Kriebel JK, Nuzzo RG, & Whitesides GM (2005) Self-assembled monolayers of thiolates on metals as a form of nanotechnology. *Chem Rev* 105(4):1103-1169.
7. Patel AJ, Varilly P, & Chandler D (2010) Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *J Phys Chem B* 114(4):1632-1637.
8. Patel AJ, Varilly P, Chandler D, & Garde S (2011) Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling. *J Stat Phys* 145(2):265-275.
9. Dallin BC & Van Lehn RC (2019) Spatially Heterogeneous Water Properties at Disordered Surfaces Decrease the Hydrophobicity of Nonpolar Self-Assembled Monolayers. *J Phys Chem Lett* 10(14):3991-3997.
10. Dallin BC, Yeon H, Ostwalt AR, Abbott NL, & Van Lehn RC (2019) Molecular Order Affects Interfacial Water Structure and Temperature-Dependent Hydrophobic Interactions between Nonpolar Self-Assembled Monolayers. *Langmuir* 35(6):2078-2088.
11. Vanommeslaeghe K, *et al.* (2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 31(4):671-690.
12. Yu W, He X, Vanommeslaeghe K, & MacKerell AD, Jr. (2012) Extension of the CHARMM General Force Field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J Comput Chem* 33(31):2451-2468.
13. Soteras Gutierrez I, *et al.* (2016) Parametrization of halogen bonds in the CHARMM general force field: Improved treatment of ligand-protein interactions. *Bioorg Med Chem* 24(20):4812-4825.
14. Abascal JL & Vega C (2005) A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys* 123(23):234505.
15. Essmann U, *et al.* (1995) A smooth particle mesh Ewald method. *The Journal of Chemical Physics* 103(19):8577-8593.
16. Hess B, Bekker H, Berendsen HJC, & Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* 18(12):1463-1472.
17. Berendsen HJC, van der Spoel D, & van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* 91(1-3):43-56.

18. Abraham MJ*, et al.* (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19-25.
19. Bussi G, Donadio D, & Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126(1):014101.
20. Tribello GA, Bonomi M, Branduardi D, Camilloni C, & Bussi G (2014) PLUMED 2: New feathers for an old bird. *Computer Physics Communications* 185(2):604-613.
21. Roux B (1995) The calculation of the potential of mean force using computer simulations. *Computer Physics Communications* 91(1-3):275-282.
22. Godawat R, Jamadagni SN, & Garde S (2009) Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations. *Proc Natl Acad Sci U S A* 106(36):15119-15124.
23. Patel AJ*, et al.* (2012) Sitting at the edge: how biomolecules use hydrophobicity to tune their interactions and function. *J Phys Chem B* 116(8):2498-2503.
24. Willard AP & Chandler D (2010) Instantaneous liquid interfaces. *J Phys Chem B* 114(5):1954-1958.
25. Head-Gordon T & Stillinger FH (1993) An orientational perturbation theory for pure liquid water. *The Journal of Chemical Physics* 98(4):3313-3327.
26. Stock P*, et al.* (2017) Unraveling Hydrophobic Interactions at the Molecular Scale Using Force Spectroscopy and Molecular Dynamics Simulations. *ACS Nano* 11(3):2586-2597.
27. Luzar A & Chandler D (1996) Hydrogen-bond kinetics in liquid water. *Nature* 379(6560):55-57.
28. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267-288.
29. Kelkar AS, Dallin BC, & Van Lehn RC (2022) Identifying nonadditive contributions to the hydrophobicity of chemically heterogeneous surfaces via dual-loop active learning. *J Chem Phys* 156(2).
30. Willard AP & Chandler D (2014) The molecular structure of the interface between water and a hydrophobic substrate is liquid-vapor like. *J Chem Phys* 141(18):18C519.