Predicting Hydrophobicity by Learning Spatiotemporal Features of Interfacial Water Structure: Combining Molecular Dynamics Simulations with Convolutional Neural Networks

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

Department of Chemical and Biological Engineering, University of Wisconsin – Madison, 1415 Engineering Drive, Madison, WI, 53706 USA

*send correspondence to: vanlehn@wisc.edu

Abstract

The hydrophobicity of functionalized interfaces can be quantified by the structure and dynamics of water molecules using molecular dynamics (MD) simulations, but existing methods to quantify interfacial hydrophobicity are computationally expensive. In this work, we develop a new machine learning approach that leverages convolutional neural networks (CNNs) to predict the hydration free energy (HFE) as a measure of interfacial hydrophobicity based on water positions sampled from MD simulations. We construct a set of idealized self-assembled monolayers (SAMs) with varying surface polarities and calculate their HFEs using indirect umbrella sampling calculations (INDUS). Using the INDUS-calculated HFEs as labels and physically informed representations of interfacial water density from MD simulations as input, we train and evaluate a series of neural networks to predict SAM HFEs. By systematically varying model hyperparameters, we demonstrate that a 3D CNN trained to analyze both spatial and temporal correlations between interfacial water molecule positions leads to HFE predictions that require an order of magnitude less MD simulation time than INDUS. We showcase the power of this model to explore a large design space by predicting HFEs for a set of 71 chemically heterogeneous SAMs with varying patterns and mole fractions.

Introduction

The hydrophobicity of an interface reflects its thermodynamic tendency to minimize contact with surrounding water molecules[1]. Interfacial hydrophobicity determines the magnitude of attractive, water-mediated hydrophobic interactions that drive numerous processes, such as protein folding[2-4], molecular recognition[4-6], colloidal aggregation[6, 7], and nanoparticle adsorption on surfaces[8-10]. However, the hydrophobicity of interfaces that are chemically heterogeneous at the nanoscale – *i.e.*, surfaces with polar and nonpolar groups in close spatial proximity – is poorly understood and difficult to predict[11-14]. Typical methods quantify interfacial hydrophobicity at the nanoscale based on the nonpolar solvent-accessible surface area[15-19] or by group-specific parameters (*e.g.*, hydrophobicity scale values[20] or octanol-water partition coefficients[21-24]). These methods neglect non-additive contributions to hydrophobicity that emerge from the complex interplay of water-water and water-solute interactions that determines interfacial water structure. Accordingly, methods that can account for these nanoscale effects are needed to predict the hydrophobicity of both synthetic and biological materials.

The hydrophobicity of idealized nonpolar solutes (*i.e.*, rigid solutes unable to form hydrogen bonds with water) manifests from the restructuring of water at two limiting length scales. Near nonpolar solutes at sub-nanometer sizes, interfacial water molecules restructure while preserving a hydrogen bond network similar to that of bulk water, leading to a negative entropy of hydration[1]. Conversely, near nonpolar solutes larger than ~1 nm, hydrogen bonds between interfacial water molecules are broken, leading to a decrease in the enthalpy of hydration[1]. These length-scale dependent perturbations to water structure and hydration thermodynamics are well-described by Lum-Chandler-Weeks (LCW) theory, which emphasizes the importance of water density fluctuations that induce the dewetting of hydrophobic interfaces[25]. Simulations have further confirmed the LCW prediction that interfacial water molecules near extended hydrophobic interfaces exhibit structural and dynamical correlations consistent with a soft, fluctuating interface[26, 27]. However, the hydrophobicity of more complex interfaces is less readily characterized. For example, experiments and simulations have shown that hydrophobicity is strongly modulated by the number and spatial positions of polar and nonpolar amino acid residues in proteins[28-30] and peptides[31-33] and the composition and patterning of polar end groups in mixed self-assembled monolayers (SAMs)[34, 35]. Interfacial physical properties, such as the length and saturation of alkanethiol ligands, have also been shown to influence the hydrophobicity of

uniformly nonpolar, planar SAMs by modulating the surface order[36-38]. These examples illustrate that realistic interfaces cannot be described as idealized nonpolar solutes, requiring alternative techniques to predict their hydrophobicity.

Several recent studies have used molecular dynamics (MD) simulations to quantify interfacial hydrophobicity. Based on the insight that water structure is a key determinant of hydrophobicity, Shin and Willard[39] defined the intrinsic hydropathy as a measure of the deviation of interfacial water structure at non-ideal hydrophobic surfaces from water structure at an ideal hydrophobic surface, while Remsing and Weeks[40] calculated the solvent electrostatic potential to quantify the local reorientation of water molecules at heterogeneous interfaces. However, neither of these techniques have been applied to differentiate the hydrophobicity of different surfaces. Recent work from our group measured the hydrophobic force between two uniformly nonpolar SAMs, and demonstrated correlation with hydrophobic forces between nonpolar SAMs measured experimentally[36]. Unfortunately, calculating the SAM-SAM hydrophobic force is too computationally expensive to be widely applied to a large range of SAM compositions and chemistries. Recognizing that water density fluctuations are enhanced near hydrophobic interfaces, Patel *et al* developed indirect umbrella sampling (INDUS) to calculate the excess chemical potential, or hydration free energy (HFE), of a cavity near an interface by measuring the free energy cost for expelling water molecules from the cavity[41, 42]. Because interfacial water molecules near hydrophobic interfaces are susceptible to dewetting, the HFE decreases near more hydrophobic interfaces. HFEs calculated using INDUS have been shown to correlate with experimental measurements[36] and the method has been applied to study the hydrophobicity of SAM surfaces[36, 43], nanostructured solutes[44], and proteins[44]. Nonetheless, INDUS is still computationally expensive due to the need for multiple simulation windows for convergence. There remains a need for a computationally efficient, quantifiable, generalizable method to compute the hydrophobicity of interfaces with varying chemical features.

As an alternative to human-guided analysis, the large data sets generated by atomistic MD simulations (comprising of thousands of atomic positions for millions of timesteps) provide the opportunity to apply machine learning (ML) techniques to sample and estimate thermodynamic properties from MD simulation output. ML has been used to accelerate MD simulations, improve force field parameterization[45, 46], and increase the efficiency of equilibrium sampling[47], but there are relatively few cases of ML being used to replace statistical mechanics formalisms through

feature-learning from labeled data[48, 49]. As one recent example, DeFever *et al.* demonstrated the use of a neural network called *PointNet* to identify local structural environments in molecular simulations[48]. In another study, Oh *et al.*[50] used a deep learning model to predict surface charge densities based on water orientation angle distributions computed from MD simulations. These studies demonstrate the power of using ML in conjunction with MD to learn features of water structure at interfaces and predict surface properties. Among different neural network architectures, Convolutional Neural Networks (CNNs) are well-suited for analyzing spatially correlated data (*e.g.*, 2D images) which allows them to outperform other ML methods at image classification tasks[51]. CNNs learn spatial features over both short and long length scales and efficiently learn generalizable features of data[52-54]. CNNs have been applied to physically relevant systems by studying local protein microenvironments around amino acids[55], evaluating protein-ligand binding sites[56], and studying the configurational chemical space for stability of molecules[57]. Importantly, CNNs have also been exploited for the analysis of sequences of images (*i.e.*, videos) by recognizing temporal correlations between distinct images[58]. A CNN is thus a potentially powerful tool for predicting thermodynamic observables from MD data based on both structural and temporal information, rather than structural information alone as in prior methods[48-50].

In this work, we investigate the hypothesis that a CNN can efficiently predict interfacial hydrophobicity based on analysis of the spatial correlations (*e.g.*, due to water structure) and temporal correlations (*e.g.*, due to water density fluctuations) encoded within the positions of interfacial water molecules sampled using MD. We utilize MD simulations and INDUS to calculate the HFEs of idealized SAMs of varying hydrophobicity. We then train a series of ML models to predict HFEs based on the positions of interfacial water molecules obtained from short MD trajectories and systematically vary the input data representation to determine the impact of spatial and temporal information on HFE predictions. We find that the prediction accuracy of a CNN exceeds that of an artificial neural network, which we attribute to the recognition of spatial correlations between interfacial water molecules. We further find that providing time-series data to either a 2D or 3D CNN increases prediction accuracy, which we attribute to recognition of short-timescale fluctuations. CNN predictions generalize across interfaces with distinct chemical properties while requiring only 200 ps of MD data per HFE prediction. We leverage this computational efficiency to compute HFEs for 71 mixed SAMs composed of methyl- and hydroxy-terminated alkanethiols and identify trends associated with the spatial heterogeneity of end groups.

These results demonstrate the ability of MD simulations with ML analysis to rapidly screen the hydrophobicity of chemically heterogeneous surfaces.

The manuscript is organized as follows. We first discuss methods for labeling idealized SAMs with HFEs, converting MD data to density matrices, and using these density matrices to train 2D and 3D CNNs to predict HFEs. We then present results for HFE calculations, evaluation of different model architectures that analyze spatial and temporal features of the input MD data, and results demonstrating the utility of a 3D CNN to learn spatial and temporal correlations that most accurately predict HFEs. Finally, we use the 3D CNN trained on ideal SAMs to predict HFEs of experimentally representative mixed SAMs of different patterns.

Methods

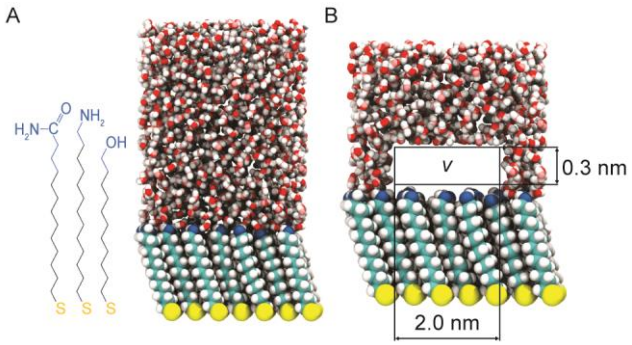


Figure 1: Idealized SAMs. (A) Chemical structures of ligands with amide, amine, and hydroxyl end groups that were used to construct idealized SAMs. The partial charges of the atoms drawn in blue are multiplied by the polarity scaling factor, k, to generate a range of SAMs of varying surface polarity. A representative MD simulation snapshot of a solvated SAM is shown at right. (B) Representative simulation snapshot from an INDUS window in which the cavity, indicated by v, is dewet ($N = 0$).

Generating a set of idealized SAMs

We generated a set of single-component, idealized SAMs similar to the SAMs studied by Kanduc *et al*[59]. Each idealized SAM consisted of 64 alkanethiol ligands with either a hydroxyl, amine, or amide end group and with a backbone containing either 11, 12, or 13 methylene groups for the hydroxyl, amide, and amine end groups respectively (Figure 1A). We multiplied the partial charges of all ligand end group atoms (defined as the atoms in the end group and atoms in the backbone as necessary to ensure a net-neutral group of atoms) by a scaling factor, k, to modulate surface polarity. SAMs were generated using values of k between 0 (most hydrophobic) to 1 (most

hydrophilic) for each of the end groups (values of k are listed in Table S1). 50 idealized SAMs were generated: 22 with hydroxyl end groups, 14 with amine end groups, and 14 with amide end groups.

Each SAM was constructed by positioning the 64 ligands in the *x*-*y* plane to mimic self-assembly onto a gold (111) lattice with a grafting density of 21.6 Å^2/ligand[60]. Gold atoms were not modeled because our prior work showed that removing the gold substrate leads to SAM properties in better agreement with experiments[36]. Ligands were oriented with the end-groups pointing in the positive *z*-direction. A 5-nm thick water layer was placed above the SAM such that the ligand end groups were in contact with the water. A 3-nm thick buffering vacuum layer was then added above the top of the water layer. Periodic boundary conditions were applied in all 3 dimensions. Ligands were modeled using the CHARMM36 force field[61] with the TIP4P/2005 water model[62]. Electrostatic interactions were calculated using the smooth Particle Mesh Ewald algorithm[63] with short-range Coulomb, van der Waals, and neighbor list cutoffs set to 1.2 nm. Because gold-sulfur bonds do not exist in the simulations, harmonic restraints with a spring constant of 50,000 kJ/mol/nm^2 were applied to the sulfur atoms to hold the SAM in place.

Indirect Umbrella Sampling (INDUS) to generate labels for idealized SAMs

We performed INDUS simulations[41, 42] to calculate HFEs for the idealized SAMs to use as labels for the machine learning model. Equation 1 defines the HFE in terms of the probability, $P_\nu(0)$, that a cavity (indicated by subscript ν) placed near the SAM is occupied by 0 water molecules:

$$\mathrm{HFE} = -k_B T \ln\left(P_\nu(0)\right) \tag{1}$$

For cavities larger than typical molecular sizes, HFEs cannot be sampled from unbiased simulations because $P_\nu(0)$ is extremely small. Instead, INDUS samples the probability of observing 0 molecules in a cavity ν by biasing the number of water molecules in the cavity in a series of independent windows. The HFE is sensitive to the placement and size of the cavity. To ensure consistency across SAMs, all INDUS calculations used a 2.0×2.0×0.3 nm^3 cavity (Figure 1B) positioned with its base on a constant water (number) density isosurface computed using the method developed by Willard and Chandler[64]. SI Section S1 further discusses the cavity placement and INDUS simulation details.

INDUS simulations were performed using the simulation package GROMACS (version 2016)[65] patched with the PLUMED plugin (version 2.5.1). Simulations used the leap-frog integrator with a 2-fs timestep. The temperature was maintained at 300 K using a velocity rescaling thermostat with a temperature-coupling time of 0.1 ps[66]. Each system was equilibrated for 5 ns, then INDUS simulations were performed using ~13 windows per SAM. Values of η_j and κ_j for all windows are listed in SI Section S1. Each window was simulated for 5 ns, with the first 2 ns discarded as equilibration. This number of windows and simulation time is sufficient for convergence (SI Figure S1), leading to a total of 65 ns of INDUS simulation time to calculate the HFE for each SAM.

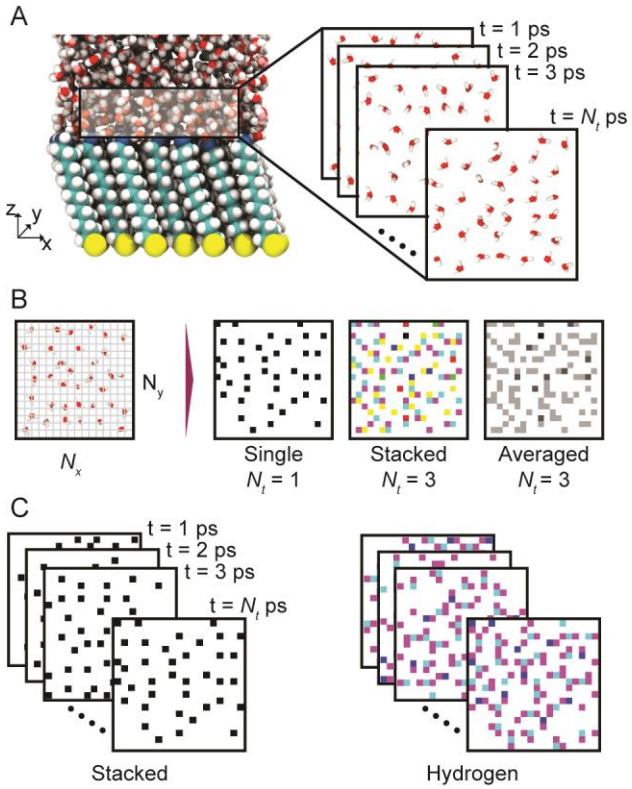


Figure 2: Conversion of MD data to CNN input. (A) Illustration of interfacial water molecules, defined as those within the INDUS cavity (defined in Figure 1B). The enlarged panel shows views of interfacial water molecules projected into the *x*-*y* plane for multiple consecutive simulation configurations from one MD trajectory. (B) Input representations for 2D CNNs. All input representations are generated by projecting the oxygen atom positions of interfacial water molecules onto a grid with N_x and N_y segments in the *x*- and *y*-dimensions, respectively, to generate a water density matrix. Either a single matrix is input (the "single" representation) or matrices are generated for N_t consecutive configurations and stored as separate

channels (for the "stacked" representation) or averaged together (for the "averaged" representation). (C) Input representations for 3D CNNs. Water density matrices for N_t consecutive configurations are stored in the 3rd (temporal) dimension for the "stacked" representation. Density matrices containing both oxygen and hydrogen atom positions are stored for the "hydrogen" representation.

Generating MD data for training and testing CNNs

Two different types of CNN architectures were used in this study – a 2D CNN and a 3D CNN. A 2D CNN, which is typically used for image analysis, interprets input data consisting of $N_x \times N_y \times N_c$ matrices, where N_x and N_y refer to the number of pixels in the x- and y-dimensions of the input image and N_c is the number of color channels (*e.g.*, $N_c = 3$ for red-green-blue images). A 3D CNN, which can be used for video analysis, interprets input data consisting of $N_x \times N_y \times N_t \times N_c$ matrices, where N_x, N_y, and N_c have the same interpretation as for a 2D CNN and N_t is the number of images (which would typically be from sequential timepoints for video analysis).

Unbiased MD simulations of each of the idealized SAMs were performed to generate atomic positions that were converted to input matrices that could be interpreted by CNNs, as summarized in Figure 2. Each simulation was performed for 40 ns using GROMACS (version 2016)[67] with the same force field and runtime parameters described above. Configurations were output every 1 ps. To convert these configurations to CNN input representations, we divided the projected area of the INDUS cavity into a $N_x \times N_y$ grid of bins, with each bin occupying a $2.0/N_x \times 2.0/N_y$ nm^2 area. For each MD configuration, we defined the set of interfacial water molecules as those water molecules with oxygen atoms within the cavity, then recorded the number of interfacial water molecules with oxygen atoms that had x- and y-positions within each bin. We defined this value as the water density within that bin, yielding a $N_x \times N_y \times 1$ water density matrix for each MD configuration.

Water density matrices were combined using different procedures to test multiple input representations for CNN analysis. In the "averaged" representation, density matrices from N_t consecutive MD configurations were averaged to create a single $N_x \times N_y \times 1$ water density matrix. In the "stacked" representation, water density matrices from N_t consecutive MD configurations were stored in separate channels to create a single $N_x \times N_y \times N_t$ water density matrix. The averaged representation was only used as an input to the 2D CNN. The stacked representation was used as an input to both the 2D and 3D CNN (as a $N_x \times N_y \times N_t \times 1$ water density matrix). Finally, we

defined a "hydrogen" representation, in which separate sets of bins were used to record both the number of interfacial oxygen atoms and interfacial hydrogen atoms, leading to a $N_x \times N_y \times N_t \times 2$ density matrix that was only input to the 3D CNN. For all results in the main text, we used $N_x = N_y = 20$; the effect of alternative values of N_x and N_y on model predictions is discussed in the SI (Figure S2-S4). N_t was varied as described in the Results. Alternative approaches, such as one in which water molecule densities were obtained using a Gaussian kernel function, were also tested without significant improvement in resulting model accuracy (discussed in SI Section S7).

Training, cross-validating, and testing the CNNs

Each CNN was trained for the regression task of predicting the HFE using density matrices generated from unbiased MD simulations as input. HFEs (labels) and corresponding 40-ns MD trajectories (each with 40,000 MD configurations) for 40 idealized SAMs were used for model training and validation. HFEs and MD trajectories for 10 idealized SAMs (4 with hydroxyl end groups, 3 with amine end groups, and 3 with amide end groups) were held out for model testing. Table S1 lists values of k for all 50 SAMs and their division into training/validation and test sets.

A 5-fold cross-validation procedure was employed to evaluate each model. In this approach, the 40 SAMs and associated MD data designated for model training and validation were randomly split into 5 equal folds. 4 folds were used as a training set and 1 fold was used as the validation set. For each SAM in the training set, the final 36 ns of the trajectory were used to generate between 1,200 and 36,000 independent density matrices depending on the value of N_t for the specific input representation, as further discussed in the Results. For each SAM in the validation set, up to the first 3 ns of the MD trajectory were used to generate between 100 and 3,000 independent density matrices depending on the value of N_t for the specific input representation. The first part of the trajectory was chosen for validation set predictions since this helps determine the ability of the CNN to screen new SAMs with minimal simulation data. Thus, in total each model was trained using between 38,400 and 1.1 million density matrices, then validated by predicting HFEs for between 800 and 24,000 density matrices. The validation set HFE for each SAM was calculated by averaging predicted HFEs for multiple density matrices, with the total amount of simulation time per HFE prediction (*i.e.*, the amount of simulation time per density matrix multiplied by the number of density matrices per HFE) treated as a

hyperparameter. This procedure was repeated five times such that each SAM was included in the validation set exactly once.

CNNs were developed using the Keras wrapper for Tensorflow, implemented in Python 3.4[68]. For training the 2D CNNs, we used an *Adamax* optimizer with a learning rate of 0.001, batch size of 128, and training over 50 epochs. For training the 3D CNNs, we used an *Adamax* optimizer with a learning rate of 0.00002, batch size of 256, and training over 500 epochs. Model architecture and training loss curves are discussed in more detail in SI Section S4.

Results and Discussion

Hydration free energies of idealized SAMs as labels for machine learning

We performed INDUS simulations to calculate HFEs for all idealized SAMs that were then used as labels for the ML model. We chose the HFE as a label because it is a thermodynamically well-defined, continuous quantity that can be used to rank-order interfacial hydrophobicity and can be related to experimental measurements. A lower HFE implies that the SAM has a lower affinity for water (leading to enhanced water density fluctuations) and is accordingly more hydrophobic than a SAM with a larger HFE. Previous studies have shown that HFEs are correlated with experimental hydrophobic force measurements between SAMs[36, 43] and contact angles[69]. We calculated the HFE using a 2.0×2.0×0.3 nm^3 cavity (Figure 1B) because this size falls within the large length scale regime appropriate for quantifying the hydrophobicity of extended planar interfaces[41]. We chose to label three sets of idealized SAMs with different end groups because these systems enable a wide range of HFEs to be sampled while testing generalizability across end group chemistries. We scaled the polarity parameter k from 0 (nonpolar) to 1 (polar) for each end group with values of k selected to distribute HFE values within +/- 5 k_BT of each other. This approach causes the HFEs (labels) to sample the entire range of relevant values for each end group to facilitate accurate regression predictions. Incorporating diverse end group chemistries ensures that features that could be associated with distinct water-end group interactions are included in the training data.

Figure 3 compares HFEs for the 40 idealized SAMs in the training set. HFEs increase monotonically with k from $k = 0$ (most hydrophobic) to $k = 1$ (most hydrophilic). For $k = 0$, the amine SAM is more hydrophobic than the hydroxyl and amide SAMs due to the distinct Lennard-Jones interactions of the various end groups with interfacial water molecules. As k increases, there

is a marginal increase in HFEs for all SAMs until k reaches approximately 0.3-0.4, at which point the partial charges in the SAM end groups lead to increased SAM-water hydrogen bonding and a consequential greater increase in the HFE. The amine SAMs remain more hydrophobic than the hydroxyl and amide SAMs for all values of k because the hydrogen atoms in the amine end group have a lower partial positive charge, which reduces the strength of hydrogen bonding with amine end group ligands. HFEs for the hydroxyl and amine SAMs increase more rapidly than the HFEs for the amide SAM, leading to a cross-over between the HFEs of the amide and hydroxyl SAMs at $k = 0.7$. We hypothesize that this behavior is due to the onset of hydrogen bonding between end groups, thus reducing hydrogen bonding with water, that is unique to the amide end group and decreases the HFE compared to the hydroxyl SAM[59]. Finally, the HFE for the hydroxyl SAM plateaus at large values of k, which is consistent with the plateauing of contact angles in previous studies[70]. These results indicate that the idealized SAMs span a range of HFEs as desired and each end group exhibits distinct behavior contributing to a diverse training set for CNN analysis.

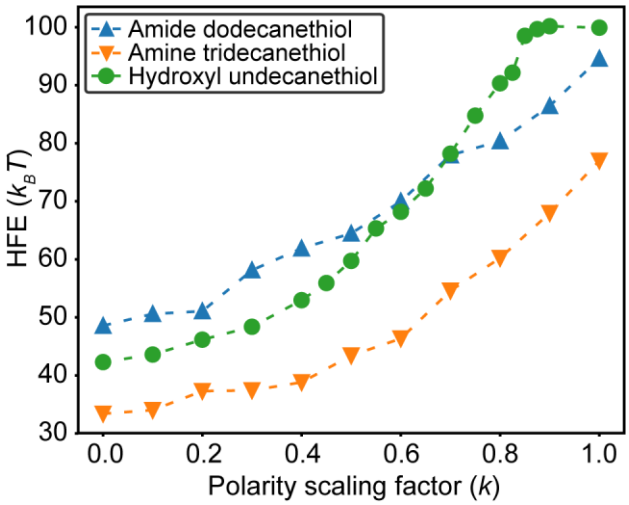


Figure 3: Labels for idealized SAMs. Hydration free energies (HFEs) of the 40 idealized SAMs included in the training set as a function of the polarity scaling factor (k) of the end group. Three different end groups (amide, amine, and hydroxyl) are modeled with structures shown in Figure 1A.

Approach for evaluating neural network architectures and input data representations

Using the idealized SAM HFEs as labels, we developed CNNs for the regression task of predicting the HFE based on input density matrices obtained from unbiased MD simulations. Depending on the input representation, each density matrix quantifies water positions in either a single MD

configuration or a series of N_t consecutive configurations (illustrated in Figure 2). Once trained, the CNN can predict an HFE for a single density matrix or can average together predictions for multiple density matrices generated with sequential MD simulation configurations. The amount of simulation time per prediction is then the product of the number of density matrices used to compute the average HFE and the simulation time per density matrix. There are thus many hyperparameters that can impact CNN accuracy, which poses a challenge due to the complexity of testing all possible hyperparameter combinations. We focused on systematically varying the CNN architecture, input representation, and the amount of simulation time per prediction to test hypotheses related to the importance of spatial and temporal correlations and to determine the efficiency of HFE predictions. We found that other hyperparameters (learning rate, batch size, *etc.*) had minimal effect on CNN accuracy.

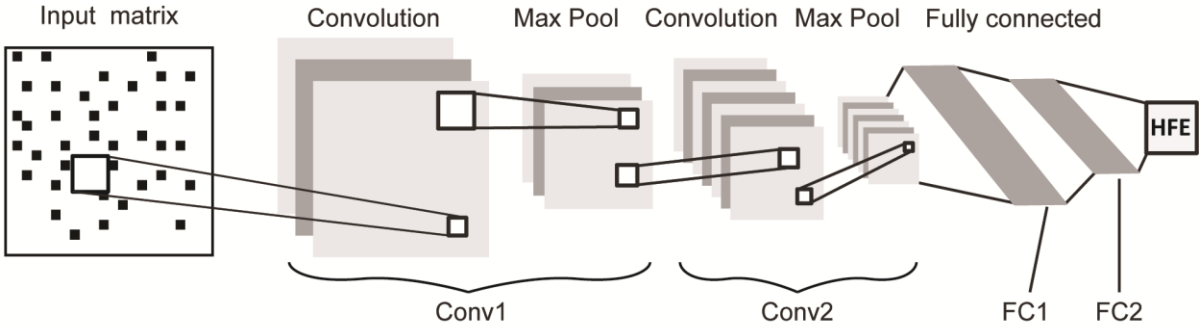
We varied one of the three selected hyperparameters at a time to identify values that improve CNN accuracy and then checked if previously selected hyperparameters were robust after selection of a new hyperparameter. For each set of hyperparameters, we trained and evaluated a corresponding CNN using the 5-fold cross-validation procedure described in the Methods. We assessed CNN accuracy by computing the root-mean-squared error (RMSE) and Pearson's R (as a measure of linear correlation) between the CNN-predicted and INDUS-calculated HFEs for SAMs in the validation set and repeated this procedure five times such that each SAM was in the validation set once. This procedure ensures the generalizability of CNN predictions and permits the comparison of CNN accuracy for different hyperparameters as explored in the following sections.

Analysis of spatial correlations improves HFE prediction accuracy

We first varied the CNN architecture by sequentially adding layers and neurons to a single-layer artificial neural network (ANN) to match the architecture of LeNet5, a 2D CNN first introduced by LeCun *et al.*[71]. Figure 4A schematically illustrates the types of layers added to the ANN: convolutional layers, which detect spatial correlations in input data; max pooling layers, which down sample input data to reduce dimensionality and the number of model parameters; and fully-connected layers, which calculate the final output given flattened input data. We hypothesized that adding convolutional layers would improve model accuracy by recognizing short length-scale spatial correlations between interfacial water molecules that have been identified at hydrophobic

interfaces[26, 27]. All CNNs were trained using the "single" representation (Figure 2) with 800 ps of simulation time per prediction (corresponding to 800 density matrices).

A



B

Conv1	Conv2	FC1	FC2	# parameters	RMSE (k_BT)	Pearson's R
0	0	0	10	40,021	15.6	0.81
0	0	0	50	200,101	14.3	0.83
0	0	180	50	81,281	14.1	0.85
1	1	180	50	10,278	19.8	0.33
2	5	180	50	13,638	15.4	0.83
25	12	180	50	31,708	13.8	0.85

Figure 4: Variations to network architecture. (A) Schematic of a 2D CNN architecture. The types of layers and numbers of neurons/filters were varied to find the best-performing architecture. (B) Table summarizing variations in network architecture and corresponding changes to validation set accuracy. The highlighted row is the chosen architecture.

Figure 4B lists the number of learned model parameters, cross-validation RMSE, and Pearson's R for different architectures. A large number of learned parameters is undesirable as it increases the likelihood of overfitting to the training data, so we evaluate model selections based on both the number of learned parameters and RMSE. The baseline ANN with a single fully connected layer has a RMSE of 15.6 k_BT. Adding neurons to the ANN increases the number of learned parameters linearly but the RMSE decreases only marginally to 14.3 k_BT. The addition of convolutional layers drastically reduces the number of parameters (~200,000 to ~10,000) but increases the RMSE. However, adding more filters to the convolutional layers leads to a RMSE substantially lower than that of the baseline ANN (13.8 k_BT) with fewer learned parameters, supporting our hypothesis that a CNN can extract features from spatial correlations in water density to improve HFE predictions. Based on these comparisons, we chose the CNN architecture highlighted in green in Figure 4B (discussed in more detail in the SI in Section S4).

Analysis of temporal correlations significantly improves HFE prediction accuracy

The previous analysis was based on input water density matrices generated from single MD configurations (N_t = 1) lacking temporal information. We next varied the input representation to test the hypothesis that including data from multiple consecutive MD configurations would capture water density fluctuations and improve CNN accuracy. We thus compared the accuracy of 2D and 3D CNNs for input representations as a function of N_t using the CNN architecture (number of layers, filters, and neurons) established in the previous section with 800 ps of simulation time per prediction; the same architecture was found to perform well for 3D CNNs (Table S2). Since density matrices are generated from unique MD configurations, increasing the value of N_t with a fixed 800 ps of simulation time decreases the number of density matrices per prediction for the stacked representation. For the averaged representation, we use overlapping configurations to generate consecutive input representations such that increasing the value of N_t decreases the number of density matrices per prediction only marginally.

Figure 5 shows the validation set RMSE for 2D CNNs with the "averaged" and "stacked" input representations as a function of N_t. For both representations, the RMSE decreases with increasing N_t and Pearson's R increases from ~0.85 for N_t = 1 (Figure 4B) to a maximum of 0.92-0.93 for larger values of N_t (Figure S7 of the SI). These results illustrate the importance of incorporating information from multiple MD configurations into the input representation. Surprisingly, the averaged and stacked representations have comparable RMSEs for all values of N_t, even though the averaged representation captures time-averaged water densities instead of density fluctuations. The comparable performance of the two representations does not invalidate the hypothesis that density fluctuations play an important role in predicting the HFE because the convolutional filters in a 2D CNN analyze all N_t configurations simultaneously (since data are stored in color channels) and thus cannot capture correlations at timescales shorter than the time associated with N_t. Alternatively, convolutional filters in a 3D CNN analyze data stored in a 3rd temporal dimension and can learn correlations between a smaller number of consecutive configurations (*e.g.*, 3) rather than all N_t configurations. A 3D CNN also has the added advantage of introducing data in color channels with only a marginal increase in the number of parameters (SI Section S4.1). We thus stored hydrogen atom density in the "hydrogen" representation that we

hypothesized would provide information on the orientation of water molecules and hydrogen bonding in the system which is absent from the oxygen density alone.

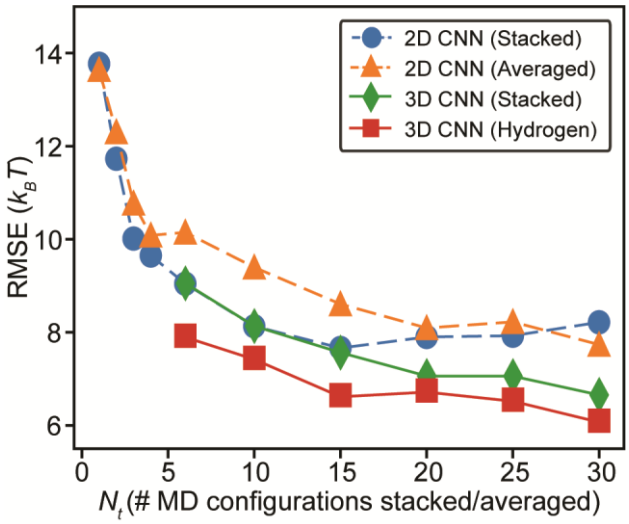


Figure 5: Variations to input representations. Validation set RMSE as a function of the number of MD configurations (N_t) included in the input representations used for the 2D and 3D CNNs with 800 ps of simulation time per prediction.

Figure 5 shows the validation set RMSE for 3D CNNs with the "stacked" and "hydrogen" input representations as a function of N_t for $N_t >= 6$ (which is required for analysis by the second convolutional layer). Both 3D CNN input representations outperform the 2D CNN input representations for $N_t > 15$. The improved accuracy of the 3D CNN compared to the 2D CNN for the "stacked" representation is notable because the input data are identical for these two CNNs. We attribute the difference in accuracy to the ability of the 3D convolutional filters to identify short time correlations for larger values of N_t that permits recognition of water density fluctuations near hydrophobic surfaces[14, 72]. The addition of hydrogen density information further reduces the RMSE to ~6 k_BT. Because density fluctuations are enhanced at more hydrophobic surfaces, we expect that the improvement in 3D CNN accuracy is due to improved predictions for the most hydrophobic surfaces. Comparing the RMSE of the best performing 2D CNN representation and best performing 3D CNN representation for the 8 most hydrophobic surfaces confirms this expectation: the RMSE for the 2D CNN is ~9.1 k_BT whereas the RMSE for the 3D CNN is ~7.2 k_BT. Based on this analysis, we conclude that a 3D CNN with N_t = 30 and the "hydrogen" representation yields the highest accuracy HFE predictions by analyzing both spatial and temporal

correlations in the input water density matrices. Further discussion of the differences between these two models is included in SI Section S6.

3D CNN HFE predictions require minimal simulation data

The final hyperparameter that we tuned is the number of density matrices used to compute average validation set HFEs during 5-fold cross validation. In principle, only 1 density matrix is required to predict the HFE for a surface, corresponding to 30 ps of MD simulation data for the best-performing 3D CNN. However, averaging HFEs calculated for a larger number of density matrices yields a more accurate HFE prediction at the cost of additional simulation time per prediction. Figure 6 shows the validation set RMSE as a function of the amount of MD simulation time per prediction for the 3D CNN with $N_t = 30$. The RMSE plateaus once the amount of simulation time per prediction exceeds 180 ps (corresponding to 6 input density matrices), which is less than the 800 ps of simulation time per prediction used in the previous hyperparameter tests (Figures 4 and 5). We also tested the effect of reducing the simulation time per MD configuration from 1 ps to 0.2 ps, but this change decreased CNN accuracy (SI Section S8). For comparison, INDUS requires 65 ns of MD simulation time to calculate the HFE (for 13 windows with 5 ns per window as described in the Methods). The 3D CNN prediction thus requires 361 times less MD simulation time than INDUS. If the 5 ns of initial SAM equilibration (largely required for the convergence of the water density isosurface used to place the INDUS cavity as discussed in the Methods) is included in both calculations, the 3D CNN still requires 13.5 times less MD simulation time than INDUS. Techniques to decrease this equilibration time (*e.g.,* by initializing the SAM ligands in tilted configurations) could be explored in future work to reduce the computational expense needed for the 3D CNN.

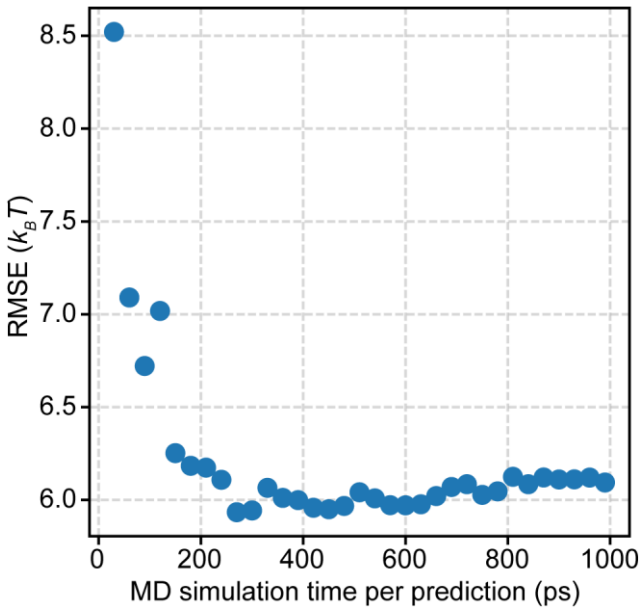


Figure 6: Variation in simulation time per prediction. Validation set RMSE as a function of the total amount of MD simulation time per prediction for a 3D CNN with N_t = 30 and the "hydrogen" input representation. Increasing the amount of simulation time leads to multiple density matrices per prediction; the reported HFE is the average of the HFEs predicted for each density matrix.

Test set predictions confirm the generalizability of 3D CNN predictions

Having identified a best-performing 3D CNN (with the "hydrogen" input representation, N_t = 30, and 180 ps of simulation time per prediction) based on 5-fold cross-validation, we compared the accuracy of this model on the unseen test set (10 SAMs) after re-training using all data in the training set (40 SAMs). Figure 7 shows a parity plot between CNN-predicted HFEs and INDUS-calculated HFEs. Black points indicate test set SAMs whereas gray points indicate predictions from cross validation of the training set SAMs. Confidence bars for each prediction of the 3D CNN shown in Figure 7 are generated by calculating the standard deviation over 4 independent 180 ps intervals. It must be noted that the confidence bars are not very wide, and only 180 ps of data suffices to make a robust HFE prediction. The test set RMSE is 5.85 k_BT, which is comparable to the RMSE of the cross-validation set (6.1 k_BT). There is an uncertainty of 1-3 k_BT associated with INDUS HFEs; thus, the 3D CNN RMSE is approximately twice the error of INDUS. The HFEs of the test set are also predicted with near-perfect rank-ordering and a Pearson's R of 0.97. These results imply generalizability of the 3D CNN beyond the training data. Finally, we note that the number of learned parameters in the chosen 3D CNN is 524,208 and the training data include

48,000 density matrices. The ratio of learned parameters to training examples is thus 10.9, implying that the 3D CNN is overparameterized as is frequently done with neural networks. This ratio is comparable to prior studies using CNNs, for which ratios of parameters to training examples range from ~1[71] to ~50[52, 54], and lies in the "interpolating" regime (corresponding to a ratio > 10) in which most model neural networks operate and which permits accurate generalization to test set data[73]. The cross-validation and regularization performed during model training avoids overfitting the data and leads to accurate test set predictions.

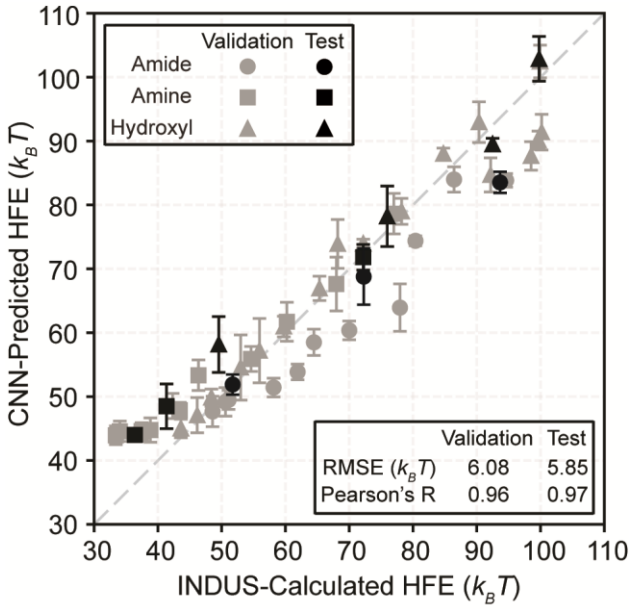


Figure 7: Validation and test set predictions. Parity plot for best-performing 3D CNN (N_t = 30 with the hydrogen representation and 180 ps of data per prediction). Prediction accuracy is comparable for 5-fold cross-validation of the training set and for the independent test set, indicating model generalizability. Confidence bars indicate the standard deviation computed from four independent samples.

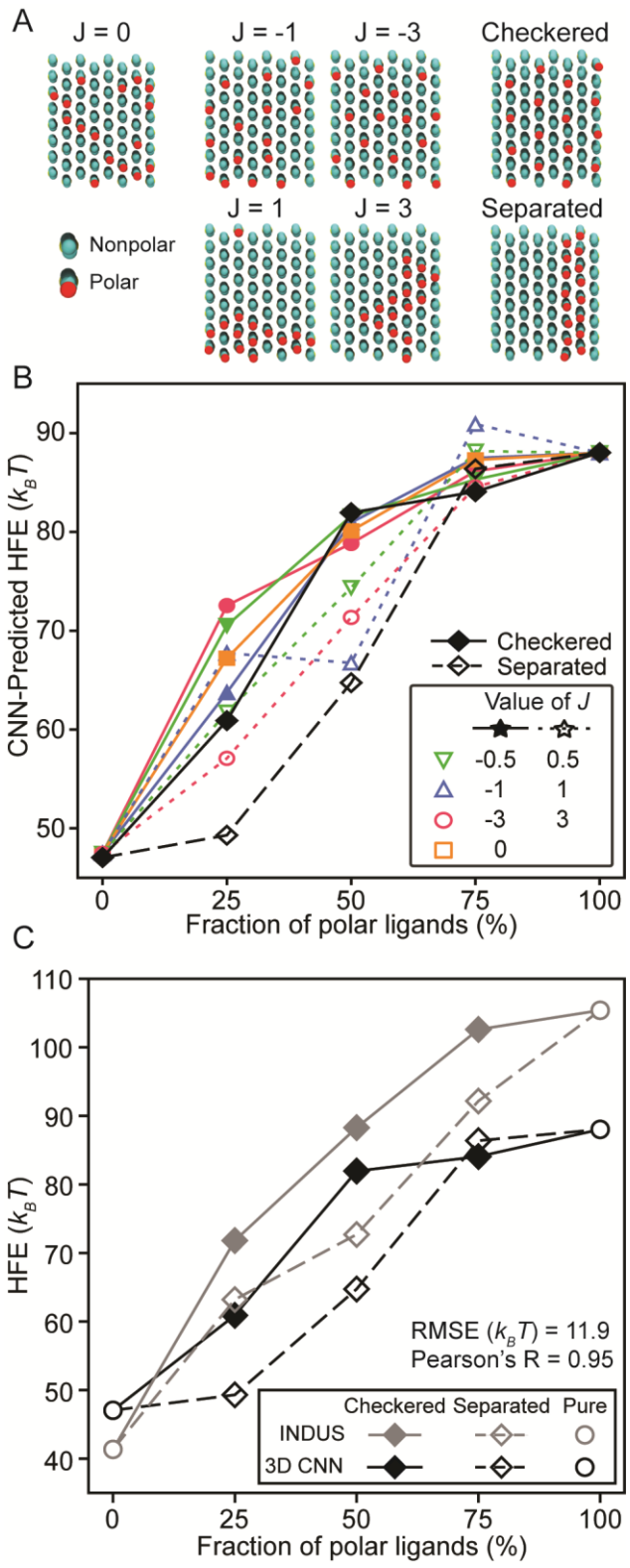


Figure 6: Hydration free energies for chemically heterogeneous SAMs. (A) Sample patterns for a SAM with a mole fraction of 25% polar end groups. The "checkered" and "separated" patterns are shown along

with examples for five different values of the Ising parameter, *J*. (B) CNN-predicted HFE as a function of the fraction of polar ligands for a set of mixed hydroxyl-tridecanethiol and tridecanethiol SAMs. Each point is the average of the HFEs computed for 3 separate SAM patterns generated using the same value of *J*. (C) Comparison of CNN-predicted and INDUS-calculated HFEs for the checkered and separated patterns.

Application of 3D CNN to predict HFEs for chemically heterogenous SAMs

Having trained and tested 3D CNN generalizability using idealized SAMs of varying hydrophobicity, we next demonstrate the use of the new model to predict HFEs for a set of chemically heterogeneous "real" SAMs. These SAMs consist of binary mixtures of two ligands: tridecanethiol, which is nonpolar, and hydroxyl-tridecanethiol, which is polar and identical to the hydroxyl-terminated ligand used for the idealized SAM with $k = 1$. The hydrophobicity of mixed SAMs constructed from these two ligands is sensitive to the mole fractions of the two components (*i.e.,* the SAM composition) and their relative spatial positioning (*i.e.*, the SAM patterning) influencing hydrophobicity. Prior studies have shown a complex dependence of interfacial hydrophobicity on SAM composition and patterning that is difficult to predict *a priori* due to nonadditive contributions to hydrophobicity from the polar and nonpolar ligands[34, 35, 44]. We thus leveraged the computational efficiency of the 3D CNN to screen a range of SAM compositions and patterns to reveal the consequent impact on hydrophobicity.

We constructed mixed SAMs containing 64 total ligands with 0, 0.25, 0.5, 0.75, and 1.0 mole fractions of the polar component following the approach used to construct idealized SAMs (see Methods). Unlike the single-component idealized SAMs, the pattern must be determined for each mixed SAM. We first generated two patterns representing opposite extremes: a "checkered" pattern, in which polar and nonpolar ligands were distributed in a checkerboard-like arrangement to achieve as uniform an arrangement as possible, and a "separated" pattern, in which polar and nonpolar ligands were distributed in two distinct domains (Figure 8A). To obtain patterns between these extremes that are not manually generated, we determined ligand positions using a Monte Carlo algorithm with an Ising model energy function in which the Ising parameter, *J*, controls the positions of nonpolar and polar ligands. A large positive value of *J* leads to patterns similar to the manually generated "separated" pattern whereas a large negative value of *J* leads to patterns similar to the "checkered" pattern. We do not expect the hydrophobicities of SAMs generated with large positive or negative values of *J* to correspond exactly to the hydrophobicities of SAMs with

manually generated patterns because an Ising model does not generate "perfect" patterns (for example, a perfectly smooth boundary between polar and nonpolar ligands, or a perfectly checkered SAM), thus enabling comparisons to the manually generated patterns. We generated patterns using the Ising model with 6 values of J between -3 and 3 for each mole fraction (excluding 0 and 1) in addition to the uniform and mixed extremes. Three independent patterns were generated for each mole fraction and value of J, leading to a total of 71 "real" SAMs. Figure 8A shows example mixed SAM patterns generated with this methodology. More details about the Ising model for SAM preparation are detailed in Section S9 of the SI. Each SAM was equilibrated for 5 ns and then simulated for another 800 ps to generate data for CNN analysis.

Figure 8B shows CNN-predicted HFEs as a function of the mole fraction of polar ligands and J for the mixed SAMs. Each point is the average of the HFE predictions for three independently generated patterns, excluding the uniform and separated patterns that were constructed without the Ising model. Increasing the mole fraction of polar ligands increases the HFE, reflecting the intuitive trend that more polar surfaces are less hydrophobic. The HFE roughly plateaus for the "checkered" pattern at a mole fraction of 0.75, in general agreement with the plateau of the HFE for the idealized hydroxyl SAMs at large values of k (Figure 3). As expected, the SAM pattern substantially impacts the HFE. SAMs in which the polar and nonpolar groups are separated (the separated pattern and large positive values of J) have a lower HFE than SAMs in which the polar and nonpolar groups are uniformly distributed (the checkered pattern and large negative values of J). These differences lead to an envelope of HFEs for a given mole fraction and highlight the ability of the CNN to determine variations in hydrophobicity associated with SAM patterns that would not be predicted by a simple linear combination model (discussed in SI Section S10.4). The trends are in agreement with previous studies which found that surfaces with clusters of polar groups appear more hydrophobic than surfaces with uniformly distributed polar groups[34, 35, 44]. However, HFEs are not monotonic in J, and in particular we observe non-monotonic changes in the HFEs with respect to mole fraction for $J = 1$. We hypothesize that these deviations arise due to defects in the underlying SAM pattern because the Ising model for small positive values of J produces a disordered boundary between the polar and nonpolar ligands (an example shown in Figure S11). Previous studies have shown that slight variations in the spatial positioning of nonpolar and polar groups can significantly impact HFEs, suggesting that defects may have a large

impact on hydrophobicity[35, 44]. The effect of defects on the HFE are an avenue for exploration in future work.

For comparison to the CNN-predicted results, we calculated HFEs using INDUS for the 6 mixed SAMs with the "checkered" and "separated" compositions. Figure 8C compares the CNN-predicted HFEs for these surfaces to the INDUS-calculated HFEs. Both techniques produce envelopes that exhibit similar trends: like the CNN-predicted HFEs, INDUS-calculated HFEs are larger for the checkered pattern compared to the separated pattern and the HFE of the checkered pattern roughly plateaus for large mole fractions. The CNN-predicted envelope is compressed relative to the INDUS-calculated envelope, with HFEs for large mole fractions underpredicted and HFEs for small mole fractions overpredicted (as discussed further in SI Section S10). This compression is expected as there is less training data for the extremes of the HFEs (Figure 3) which can bias predictions towards the mean of the data[74]. Predictions of these values might be improved by adding additional surfaces at these extremes or by including unphysical systems designed to be more hydrophobic (*e.g.*, SAMs lacking van der Waals interactions) or more hydrophilic (*e.g.*, SAMs with $k > 1$) than those studied here to expand the range of HFEs studied. Nonetheless, the overall correlation between the values is high (Pearson's R of 0.95; see SI Figure S14). Since the HFE is a relative metric used to rank-order surfaces of varying hydrophobicity, and its relationship to experimental quantities is also via linear correlation[43], the ability of the CNN to predict the same HFE trends as INDUS suggests that CNN-predicted HFEs can be used to screen the hydrophobicity of real surfaces (particularly those that lie within the extremes of hydrophobicity) with minimal computational cost.

Conclusions

In this work, we have presented a machine learning approach that uses a CNN to predict the HFE of a functionalized surface based on interfacial water densities obtained from short unbiased MD simulations. We computed HFEs for a set of idealized SAMs with varying end groups and surface polarities to use as labels for CNN training. We converted interfacial water configurations from MD simulations of the same SAMs into input density matrices for CNN analysis and then tuned hyperparameters including the network architecture, input data representation, and amount of simulation time per HFE prediction. Based on prior studies of interfacial hydrophobicity, we hypothesized that a CNN designed to learn spatial correlations (*e.g.*, due to water structure) and

temporal correlations (*e.g.*, due to water density fluctuations) between interfacial water molecules would improve prediction accuracy. This hypothesis was validated by showing that the most accurate HFE predictions were obtained using a 3D CNN which analyzed input matrices with two spatial dimensions and one temporal dimension. The 3D CNN further permitted analysis of both oxygen and hydrogen atomic densities to gain insight into factors such as water orientation. Generalizability of the 3D CNN was demonstrated through predictions on an independent test set. Critically, the 3D CNN reduces the simulation time required to predict the HFE to as little as 200 ps after SAM equilibration, reducing the computational expense by 1-2 orders of magnitude compared to indirect umbrella sampling and enabling the screening of a larger design space. We thus applied the 3D CNN to screen a set of "real" SAMs of mixed polar and nonpolar ligands with varying patterns to demonstrate that SAMs with uniformly distributed polar and nonpolar ligands are less hydrophobic than SAMs with separated ligands. The computational efficiency associated with the newly developed 3D CNN will enable the study of hydrophobicity of a wider array of patterns in future work to further interrogate the origin of these non-monotonic effects.

Future work will extend this approach to other types of chemistries and surfaces, including but not restricted to planar and more heterogeneous systems. This machine learning model can also be augmented with chemical information (*e.g.*, hydrogen bonds) not explicitly captured in the input representations that we have considered. Such chemical information could potentially be captured by using physically intuitive and spatially non-restrictive data representations like hydrogen bond network graphs, rather than the water density matrices, in conjunction with alternative machine learning methods like Graph Neural Networks. Comparison of alternative networks and input representations to those studied in this work can further help understand the manifestation of hydrophobicity. For example, features that differentiate highly hydrophobic and hydrophilic surfaces based on invariants in water structure can be extracted from the trained CNN and compared to features recognized by alternative network architectures. Finally, we anticipate that this machine learning approach could be extended to predict interfacial behavior (*e.g.*, protein binding) by providing input like the electrostatic potential in addition to water structure.

Supplementary Information

Additional details on INDUS simulations, values of HFEs of idealized SAMs, parameter tests for size of input representation and INDUS cavity size, CNN architecture and loss curves, Pearson's

R for cross-validation (2D and 3D CNN), use of a Gaussian kernel to approximate water molecules, details about the Ising model used to generate SAM morphology, results of Ising SAMs on varying training data, other observations of HFE predictions made using CNNs

Acknowledgments

The authors gratefully acknowledge partial support of this research through the University of Wisconsin Materials Research Science and Engineering Center (DMR-1720415). Support for this research was provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562, and the Center for High Throughput Computing at the University of Wisconsin-Madison. The authors also thank Alexander Smith and Alex Chew from the University of Wisconsin-Madison for their insightful discussion.

References

1. Chandler, D., Interfaces and the driving force of hydrophobic assembly. *Nature* **2005,** *437* (7059), 640-7.
2. Dill, K. A., Dominant forces in protein folding. *Biochemistry* **1990,** *29* (31), 7133-55.
3. Dobson, C. M., Protein folding and misfolding. *Nature* **2003,** *426* (6968), 884-90.
4. Nicholls, A.; Sharp, K. A.; Honig, B., Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **1991,** *11* (4), 281-96.
5. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **2006,** *49* (21), 6177-96.
6. Snyder, P. W.; Mecinovic, J.; Moustakas, D. T.; Thomas, S. W., 3rd; Harder, M.; Mack, E. T.; Lockett, M. R.; Heroux, A.; Sherman, W.; Whitesides, G. M., Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. *Proc Natl Acad Sci U S A* **2011,** *108* (44), 17889-94.
7. Van Lehn, R. C.; Alexander-Katz, A., Ligand-mediated short-range attraction drives aggregation of charged monolayer-protected gold nanoparticles. *Langmuir* **2013,** *29* (28), 8788-98.
8. Nakanishi, K.; Sakiyama, T.; Imamura, K., On the adsorption of proteins on solid surfaces, a common but very complicated phenomenon. *J Biosci Bioeng* **2001,** *91* (3), 233-44.
9. Van Lehn, R. C.; Atukorale, P. U.; Carney, R. P.; Yang, Y. S.; Stellacci, F.; Irvine, D. J.; Alexander-Katz, A., Effect of particle diameter and surface composition on the spontaneous fusion of monolayer-protected gold nanoparticles with lipid bilayers. *Nano Lett* **2013,** *13* (9), 4060-7.

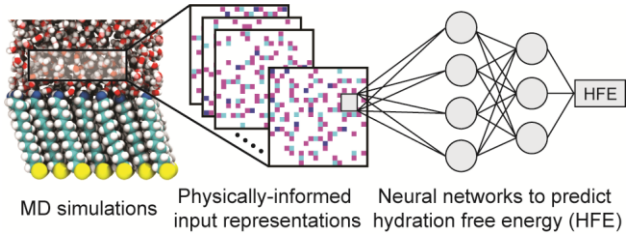
10. Van Lehn, R. C.; Alexander-Katz, A., Pathway for insertion of amphiphilic nanoparticles into defect-free lipid bilayers from atomistic molecular dynamics simulations. *Soft Matter* **2015,** *11* (16), 3165-3175.
11. Wang, J. H.; Bratko, D.; Luzar, A., Probing surface tension additivity on chemically heterogeneous surfaces by a molecular approach. *P Natl Acad Sci USA* **2011,** *108* (16), 6374-6379.
12. Giovambattista, N.; Lopez, C. F.; Rossky, P. J.; Debenedetti, P. G., Hydrophobicity of protein surfaces: Separating geometry from chemistry. *P Natl Acad Sci USA* **2008,** *105* (7), 2274-2279.
13. Chong, S. H.; Ham, S., Impact of chemical heterogeneity on protein self-assembly in water. *P Natl Acad Sci USA* **2012,** *109* (20), 7636-7641.
14. Jamadagni, S. N.; Godawat, R.; Garde, S., Hydrophobicity of Proteins and Interfaces: Insights from Density Fluctuations. *Annual Review of Chemical and Biomolecular Engineering* **2011,** *2*, 147-171.
15. Reynolds, J. A.; Gilbert, D. B.; Tanford, C., Empirical Correlation between Hydrophobic Free-Energy and Aqueous Cavity Surface-Area. *P Natl Acad Sci USA* **1974,** *71* (8), 2925-2927.
16. Abraham, M. H.; Whiting, G. S.; Fuchs, R.; Chambers, E. J., Thermodynamics of Solute Transfer from Water to Hexadecane. *J Chem Soc Perk T 2* **1990,** (2), 291-300.
17. Stephenson, B. C.; Goldsipe, A.; Beers, K. J.; Blankschtein, D., Quantifying the hydrophobic effect. 1. A computer simulation-molecular-thermodynamic model for the self-assembly of hydrophobic and amphiphilic solutes in aqueous solution. *J Phys Chem B* **2007,** *111* (5), 1025-1044.
18. Jiang, L. G.; Cao, S. Q.; Cheung, P. P. H.; Zheng, X. Y.; Leung, C. W. T.; Peng, Q.; Shuai, Z. G.; Tang, B. Z.; Yao, S. H.; Huang, X. H., Real-time monitoring of hydrophobic aggregation reveals a critical role of cooperativity in hydrophobic effect. *Nat Commun* **2017,** *8*.
19. Chen, J.; Brooks, C. L., Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys Chem Chem Phys* **2008,** *10* (4), 471-481.
20. MacCallum, J. L.; Tieleman, D. P., Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions. *Trends Biochem Sci* **2011,** *36* (12), 653-662.
21. Moyano, D. F.; Saha, K.; Prakash, G.; Yan, B.; Kong, H.; Yazdani, M.; Rotello, V. M., Fabrication of Corona-Free Nanoparticles with Tunable Hydrophobicity. *ACS Nano* **2014,** *8* (7), 6748-6755.
22. Li, X. N.; Robinson, S. M.; Gupta, A.; Saha, K.; Jiang, Z. W.; Moyano, D. F.; Sahar, A.; Riley, M. A.; Rotello, V. M., Functional Gold Nanoparticles as Potent Antimicrobial Agents against Multi-Drug-Resistant Bacteria. *ACS Nano* **2014,** *8* (10), 10682-10686.
23. Moyano, D. F.; Goldsmith, M.; Solfiell, D. J.; Landesman-Milo, D.; Miranda, O. R.; Peer, D.; Rotello, V. M., Nanoparticle Hydrophobicity Dictates Immune Response. *Journal of the American Chemical Society* **2012,** *134* (9), 3965-3967.
24. Chen, K. M.; Rana, S.; Moyano, D. F.; Xu, Y. S.; Guo, X. H.; Rotello, V. M., Optimizing the selective recognition of protein isoforms through tuning of nanoparticle hydrophobicity. *Nanoscale* **2014,** *6* (12), 6492-6495.
25. Lum, K.; Chandler, D.; Weeks, J. D., Hydrophobicity at small and large length scales. *J Phys Chem B* **1999**.
26. Mittal, J.; Hummer, G., Static and dynamic correlations in water at hydrophobic interfaces. *Proc Natl Acad Sci U S A* **2008,** *105* (51), 20130-5.
27. Meyer, E. E.; Rosenberg, K. J.; Israelachvili, J., Recent progress in understanding hydrophobic interactions. *Proc Natl Acad Sci U S A* **2006,** *103* (43), 15739-46.

28. Patel, A. J.; Garde, S., Efficient method to characterize the context-dependent hydrophobicity of proteins. *J Phys Chem B* **2014,** *118* (6), 1564-73.
29. Patel, A. J.; Varilly, P.; Jamadagni, S. N.; Hagan, M. F.; Chandler, D.; Garde, S., Sitting at the Edge: How Biomolecules use Hydrophobicity to Tune Their Interactions and Function. *J Phys Chem B* **2012,** *116* (8), 2498-2503.
30. Giovambattista, N.; Lopez, C. F.; Rossky, P. J.; Debenedetti, P. G., Hydrophobicity of protein surfaces: Separating geometry from chemistry. *Proc Natl Acad Sci U S A* **2008,** *105* (7), 2274-9.
31. Stock, P.; Monroe, J. I.; Utzig, T.; Smith, D. J.; Shell, M. S.; Valtiner, M., Unraveling Hydrophobic Interactions at the Molecular Scale Using Force Spectroscopy and Molecular Dynamics Simulations. *ACS Nano* **2017,** *11* (3), 2586-2597.
32. Ma, C. D.; Wang, C.; Acevedo-Velez, C.; Gellman, S. H.; Abbott, N. L., Modulation of hydrophobic interactions by proximally immobilized ions. *Nature* **2015,** *517* (7534), 347-50.
33. Wang, C.; Ma, C. D.; Yeon, H.; Wang, X.; Gellman, S. H.; Abbott, N. L., Nonadditive Interactions Mediated by Water at Chemically Heterogeneous Surfaces: Nonionic Polar Groups and Hydrophobic Interactions. *J Am Chem Soc* **2017,** *139* (51), 18536-18544.
34. Monroe, J. I.; Shell, M. S., Computational discovery of chemically patterned surfaces that effect unique hydration water dynamics. *Proc Natl Acad Sci U S A* **2018,** *115* (32), 8093-8098.
35. Schrader, A. M.; Monroe, J. I.; Sheil, R.; Dobbs, H. A.; Keller, T. J.; Li, Y.; Jain, S.; Shell, M. S.; Israelachvili, J. N.; Han, S., Surface chemical heterogeneity modulates silica surface hydration. *Proc Natl Acad Sci U S A* **2018,** *115* (12), 2890-2895.
36. Dallin, B. C.; Yeon, H.; Ostwalt, A. R.; Abbott, N. L.; Van Lehn, R. C., Molecular Order Affects Interfacial Water Structure and Temperature-Dependent Hydrophobic Interactions between Nonpolar Self-Assembled Monolayers. *Langmuir* **2019,** *35* (6), 2078-2088.
37. Yeon, H.; Wang, C.; Van Lehn, R. C.; Abbott, N. L., Influence of Order within Nonpolar Monolayers on Hydrophobic Interactions. *Langmuir* **2017,** *33* (19), 4628-4637.
38. Cui, X.; Liu, J.; Xie, L.; Huang, J.; Liu, Q.; Israelachvili, J. N.; Zeng, H., Modulation of Hydrophobic Interaction by Mediating Surface Nanoscale Structure and Chemistry, not Monotonically by Hydrophobicity. *Angew Chem Int Ed Engl* **2018,** *57* (37), 11903-11908.
39. Shin, S.; Willard, A. P., Water's Interfacial Hydrogen Bonding Structure Reveals the Effective Strength of Surface–Water Interactions. *J Phys Chem B* **2018,** *122* (26), 6781-6789.
40. Remsing, R. C.; Weeks, J. D., Hydrophobicity Scaling of Aqueous Interfaces by an Electrostatic Mapping. *J Phys Chem B* **2015,** *119* (29), 9268-77.
41. Patel, A. J.; Varilly, P.; Chandler, D.; Garde, S., Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling. *J Stat Phys* **2011,** *145* (2), 265-275.
42. Patel, A. J.; Varilly, P.; Chandler, D., Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *J Phys Chem B* **2010,** *114* (4), 1632-7.
43. Dallin, B. C.; Van Lehn, R. C., Spatially Heterogeneous Water Properties at Disordered Surfaces Decrease the Hydrophobicity of Nonpolar Self-Assembled Monolayers. *J Phys Chem Lett* **2019,** *10* (14), 3991-3997.
44. Xi, E.; Venkateshwaran, V.; Li, L.; Rego, N.; Patel, A. J.; Garde, S., Hydrophobicity of proteins and nanostructured solutes is governed by topographical and chemical context. *Proc Natl Acad Sci U S A* **2017,** *114* (51), 13345-13350.
45. Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R., Machine learning force fields: Construction, validation, and outlook. *J Phys Chem C* **2016,** *121* (1), 511-522.

46. Gastegger, M.; Behler, J.; Marquetand, P., Machine learning molecular dynamics for the simulation of infrared spectra. *Chem Sci* **2017,** *8* (10), 6924-6935.
47. Noe, F.; Olsson, S.; Kohler, J.; Wu, H., Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019,** *365* (6457), 1001-+.
48. DeFever, R. S.; Targonski, C.; Hall, S. W.; Smith, M. C.; Sarupria, S., A generalized deep learning approach for local structure identification in molecular simulations. *Chem Sci* **2019,** *10* (32), 7503-7515.
49. Chew, A.; Jiang, S.; Zhang, W.; Zavala, V.; Van Lehn, R., *Fast Predictions of Liquid-Phase Acid-Catalyzed Reaction Rates Using Molecular Dynamics Simulations and Convolutional Neural Networks*. arXiv preprint, 2019.
50. Oh, M. I.; Oh, C. I.; Gupta, M.; Weaver, D. F., Decoding Interfacial Water Orientation to Predict Surface Charge Density on a Model Sheet Using a Deep Learning Algorithm. *Journal of Physical Chemistry C* **2020,** *124* (4), 2574-2582.
51. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015,** *521* (7553), 436-44.
52. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems, 2012; pp 1097-1105.
53. Simonyan, K.; Zisserman, A., Very deep convolutional networks for large-scale image recognition. *arXiv preprint* **2014**.
54. He, K.; Zhang, X.; Ren, S.; Sun, J. In *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp 770-778.
55. Torng, W.; Altman, R. B., 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics* **2017,** *18* (1), 302.
56. Jimenez, J.; Doerr, S.; Martinez-Rosell, G.; Rose, A. S.; De Fabritiis, G., DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017,** *33* (19), 3036-3042.
57. Schutt, K. T.; Arbabzadah, F.; Chmiela, S.; Muller, K. R.; Tkatchenko, A., Quantum-chemical insights from deep tensor neural networks. *Nat Commun* **2017,** *8*, 13890.
58. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. In *Learning spatiotemporal features with 3d convolutional networks*, Proceedings of the IEEE international conference on computer vision, 2015; pp 4489-4497.
59. Kanduc, M.; Schlaich, A.; Schneck, E.; Netz, R. R., Water-Mediated Interactions between Hydrophilic and Hydrophobic Surfaces. *Langmuir* **2016,** *32* (35), 8767-82.
60. Love, J. C.; Estroff, L. A.; Kriebel, J. K.; Nuzzo, R. G.; Whitesides, G. M., Self-assembled monolayers of thiolates on metals as a form of nanotechnology. *Chem Rev* **2005,** *105* (4), 1103-69.
61. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **2010,** *31* (4), 671-90.
62. Abascal, J. L.; Vega, C., A general purpose model for the condensed phases of water: TIP4P/2005. *J Chem Phys* **2005,** *123* (23), 234505.
63. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A Smooth Particle Mesh Ewald Method. *Journal of Chemical Physics* **1995,** *103* (19), 8577-8593.

64. Willard, A. P.; Chandler, D., Instantaneous liquid interfaces. *J Phys Chem B* **2010,** *114* (5), 1954-8.
65. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. J. S., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **2015,** *1*, 19-25.
66. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *The Journal of chemical physics* **2007,** *126* (1), 014101.
67. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *J Comput Chem* **2005,** *26* (16), 1701-18.
68. Chollet, F. O., Keras: Deep learning library for theano tensorflow. **2015**.
69. Godawat, R.; Jamadagni, S. N.; Garde, S., Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations. *Proc Natl Acad Sci U S A* **2009,** *106* (36), 15119-24.
70. Kanduc, M.; Netz, R. R., Atomistic simulations of wetting properties and water films on hydrophilic surfaces. *J Chem Phys* **2017,** *146* (16), 164705.
71. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P., Gradient-based learning applied to document recognition. *Proceedings of the Ieee* **1998,** *86* (11), 2278-2324.
72. Acharya, H.; Vembanur, S.; Jamadagni, S. N.; Garde, S., Mapping hydrophobicity at the nanoscale: applications to heterogeneous surfaces and proteins. *Faraday Discuss* **2010,** *146*, 353-65; discussion 367-93, 395-401.
73. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S., Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc Natl Acad Sci U S A* **2019,** *116* (32), 15849-15854.
74. Mac Namee, B.; Cunningham, P.; Byrne, S.; Corrigan, O. I., The problem of bias in training data in regression problems in medical decision support. *Artif Intell Med* **2002,** *24* (1), 51-70.

TOC Graphic


HFE
MD simulations
Physically-informed input representations
Neural networks to predict hydration free energy (HFE)