Electronic Supplementary Material (ESI) for Chemical Science.

Supporting Information for: A generalized deep learning approach for local structure identification in molecular simulations

Ryan S. DeFever,[†] Colin Targonski,[‡] Steven W. Hall,[†] Melissa Smith,[‡] and Sapna Sarupria[*,†]

[†]*Department of Chemical & Biomolecular Engineering, Clemson University, Clemson, SC 29634*

[‡]*Department of Electrical & Computer Engineering, Clemson University, Clemson, SC 29634*

E-mail: ssarupr@g.clemson.edu

Simulations for crystal structure identification

Simulations of pure phases were performed to generate training data for the PointNet. The pure phases were simulated in a range of temperature and pressure conditions to expose the network to conditions with varied density and magnitude of thermal fluctuations. Though temperature and pressure conditions sometimes exceeded the thermodynamic stability of the simulated phases, we confirmed that all phases remained mechanically stable for the duration of the simulations. Simulation details specific to the different systems are provided in the following sections.

Lennard–Jonesium

Simulations of bulk liquid, face-centered cubic (fcc), hexagonal close-packed (hcp), and body-centered cubic (bcc) phases were performed in a range of conditions both above and below the melting point. Initial configurations for the solid phases were generated by replicating the unit cell and resulted in 16384, 14976, and 17496 atoms for the fcc, hcp, and bcc phases, respectively. The initial configuration for the liquid phase consisted of 16384 atoms randomly placed in a cubic simulation box of length 25σ, where σ is the size parameter in the LJ potential. All values for the LJ system are reported in reduced units.

Simulations of the liquid, fcc, and hcp phases were performed in the NpT ensemble at a range of temperatures between 0.5 and $1.7\epsilon/k_\mathrm{B}$ and pressures between 0 and $15\epsilon/\sigma^3$. Simulations of the bcc phase were performed in the NVT ensemble with a range of temperatures between 0.6 and 1.6 ϵ/k_B and densities between 0.95 and $1.2\sigma^{-3}$. Each NpT system was first equilibrated for 500τ to the target conditions, followed by a 2000τ simulation with the Bussi thermostat[1] and Parrinello-Rahman barostat,[2] each with coupling constant 0.5τ. Since the bcc phase is unstable with respect to transformation to the close-packed phases, a slab of frozen particles in bcc arrangement was used to stabilize the crystal. Analysis was only performed on particles several layers from the frozen slab. Simulations were performed in GROMACS 2018.[3] A time step of 0.001τ was used. Group cutoff scheme was used with neighbor list updates every 10 steps and cutoff radius of 4.0σ. The LJ potential was force-switched from a distance of 3.0σ to the cutoff at 3.5σ.

Water

All-atom simulations of water were performed for the liquid phase, five ice phases, and two guest-free hydrate phases. Liquid and ice phases were simulated at temperatures spaced between 200 K and 300 K and at pressures between 1 bar and 12000 bar. Hydrate phases were sampled at temperatures 230–270 K, with pressures -2000–1000 bar. Initial configurations for the solid phases were generated by replicating the unit cell, and resulted in the following numbers of water molecules in each system: ice Ih: 768, ice Ic: 512, ice III: 768, ice V: 1792, ice VI: 640, hydrate sI: 1242, hydrate sII: 1088. The initial liquid configuration consisted of 909 water molecules. Following an energy minimization step, systems were simulated for 25 ns in the NpT ensemble at the target temperature and pressure. Temperature was maintained with the thermostat of Bussi *et al.*[1] with a coupling constant of 0.5 ps. Anisotropic (isotropic) pressure coupling was applied for the solid(liquid) phase(s) with the Berendsen barostat[4] with a coupling time constant of 5 ps. The first 5 ns of the simulation was treated as equilibration and not used for data collection.

Water was described by the TIP4P/Ice[5] model. Simulations were performed in GROMACS 2018.[3] Dynamics were propagated by the leap-frog integrator with an integration time step of 2 fs. Linear center-of-mass motion was removed every 10 integration steps. Cutoffs for LJ and Coloumbic interactions were set to 1.0 nm. The Verlet cutoff scheme was employed with the Verlet buffer tolerance set to 0.005.[6] Long-range electrostatics were treated with particle mesh Ewald.[7] Geometry of water molecules was maintained with SETTLE.[8]

Mesophases

Simulations of six mesophases were performed: liquid, lamellar, lxs, hexagonal, gyroid, and body-centered cubic. The systems were described by the model presented in Ref. 9, which is comprised of pairwise interactions using the two-body term of the Stillinger–Weber potential.[10] The systems comprise of two particle types, denoted A and B. Different mesophases form from tuning the A–B interactions and the fraction of type A, χ_{A}. All simulations were performed with $\varepsilon_{\mathrm{AA}} = \varepsilon_{\mathrm{BB}} = 1.0$ kcal mol^{-1}, $\sigma_{\mathrm{AA}} = \sigma_{\mathrm{BB}} = 1.0$ and $\sigma_{\mathrm{AB}} = 1.15$. Values other than temperature and energy are reported as dimensionless quantities. All simulations are performed at $T = 300$ K and $p = 0$. Simulations are performed in the $\chi_{\mathrm{A}} > 0.5$ portion of the phase diagram so type B is the minor component.

Guided by the phase diagram presented in Fig. 6 of Ref. 9, we select the following conditions for each phase. Liquid: $\chi_{\mathrm{A}} = 0.5$, $\varepsilon_{\mathrm{AB}} = 0.85$ kcal mol^{-1}, lamellar: $\chi_{\mathrm{A}} =$

0.5, $\varepsilon_{\mathrm{AB}} = 1.4$ kcal mol^{-1}, shifted layered crystal (lxs): $\chi_{\mathrm{A}} = 0.5$, $\varepsilon_{\mathrm{AB}} = 1.9$ kcal mol^{-1}, hexagonal: $\chi_{\mathrm{A}} = 0.77$, $\varepsilon_{\mathrm{AB}} = 1.8$ kcal mol^{-1}, gyroid: $\chi_{\mathrm{A}} = 0.67$, $\varepsilon_{\mathrm{AB}} = 1.8$ kcal mol^{-1}, body-centered cubic (bcc): $\chi_{\mathrm{A}} = 0.86$, $\varepsilon_{\mathrm{AB}} = 3.8$ kcal mol^{-1}. Except for the body-centered cubic phase, all phases were generated through nucleation from the isotropic liquid. All systems except bcc contained 16384 atoms. The bcc systems contained 14000 atoms.

Simulations were performed in GROMACS 2018[3] using tabulated potentials. The cut-off was set to the theoretical maximum for the Stillinger–Weber potential. Equations of motion were integrated with the leap-frog integrator with a time step of 0.005. Systems were equilibrated for 500,000 steps in the NpT ensemble with temperature and pressure coupling maintained by the Bussi thermostat[1] ($\tau_T = 2.0$) and Berendsen barostat[4] ($\tau_p = 4.1$), respectively. For the production simulations, temperature and pressure were maintained with the Bussi thermostat[1] and Parrinello–Rahman barostat[2] with damping constants of $\tau_T = 2.0$ and $\tau_p = 10.2$, respectively. Production simulations were performed for 2.5×10^8 steps. Only the portion of the simulations after the crystal phase had grown to occupy the entire simulation box were used for analysis.

Simulations for hydrophobicity identification

Self-assembled monolayer systems

Self-assembled monolayer (SAM) surfaces are flexible organic surfaces composed of alkane chains attached to a metal surface. All SAM surfaces were constructed to be approximately 6×7 nm with 192 alkane chains total. Each chain contains a sulfur atom attached to one end of a 10-carbon alkane chain and a terminal group at the other end, in this case CH3 and OH. Sulfur atoms were restrained to positions corresponding to their hypothetical spacing when adsorbed to a Au (111) surface. The in-plane structure of the sulfur atoms was $\sqrt{3} \times \sqrt{3}$R30 with a 0.497 nm distance between neighboring sulfur atoms.[11] The surfaces are periodic in x and y directions. Partial charges were taken from the OPLS-AA force field.[12] All other bonded and nonbonded parameters were taken from the General Amber force field.[13] The surface with vacuum space on either side in the z direction was equilibrated in the NVT ensemble for 5 ns at 300 K. A slab of 6000 TIP3P water molecules was placed in contact with the surface terminal groups. The vacuum space above the water acts as a natural barostat, maintaining the pressure at 0 bar. The surface–water system was equilibrated in the NVT ensemble (300 K) for 5 ns. Training and testing samples were collected from a subsequent production run of 25 ns in the NVT ensemble (300 K). Simulations were performed in GROMACS[3] with a time step of 0.002 ps. The Bussi thermostat[1] maintained temperature with time constant $\tau_T = 0.5$ ps. Hydrogen bonds were constrained with LINCS.[14] LJ and Coulombic cutoffs were set to 1.0 nm. Particle mesh Ewald was used to calculate long-range electrostatics.[15]

Protein systems

Structures of hydrophobin II (PDB: 2B97) and CheY (PDB: 3CHY) were taken from the Protein Data Bank (PDB). Hydrophobin and CheY were solvated with TIP3P water in $5 \times 5 \times 5$ nm^3 and $8 \times 8 \times 8$ nm^3 simulation boxes, respectively. Four sodium counter ions were added to the CheY system. The proteins were described by the AMBER99SB-ILDN force field.[16] Heavy atoms of the proteins were position restrained and the systems were energy minimized. Following the energy minimization, the systems were equilibrated for 5 ns in the NpT ensemble (300 K, 1 bar) with the protein heavy atoms position restrained. Temperature coupling was only applied to the solvent (Bussi thermostat,[1] $\tau_T = 0.5$ ps). The Berendsen barostat[4] was used during equilibration with $\tau_p = 1.0$ ps. Systems were simulated in production for 25 ns in the NpT ensemble (300 K, 1 bar) with no position restraints. Temperature coupling was applied with the Bussi thermostat[1] ($\tau_T = 1.0$ ps) and the Parrinello-Rahman barostat[2] ($\tau_p = 5.0$ ps). Temperature coupling was only applied to the solvent. All simulations were performed in GROMACS 2018.[3] Equations of motion were integrated with the leap frog algorithm with a time step of 0.002 ps. LJ and Coulombic cutoffs were set to 1.0 nm. Particle mesh Ewald was used to calculate long-range electrostatics.[15] Hydrogen bonds were constrained with the LINCS algorithm.[14]

Details of Geiger–Dellago network implementation

The network described in Geiger and Dellago[17] was constructed in Keras.[18] We used a cutoff of 2.6 and 0.6 nm for the LJ and water systems, respectively. The symmetry functions and their parameters were taken directly from Ref. 17. The network consisted of two hidden layers with 35 neurons each, followed by a softmax layer to determine the final classification. The two hidden layers had ReLU activation functions with batch normalization. The model was trained for fifty epochs with the Adam optimizer[19] with a learning rate 0.0005 and default parameters. Roughly ∼500,000 and ∼3,000,000 training examples were used for the LJ and water systems, respectively.

References

(1) Bussi, G.; Donadio, D.; Parrinello, M. J. Chem. Phys. **2007**, 126, 014101.

(2) Parrinello, M.; Rahman, A. J. Appl. Phys. **1981**, 52, 7182–7190.

(3) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. SoftwareX **2015**, 1, 19–25.

(4) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J. Chem. Phys. **1984**, 81, 3684–3690.

(5) Abascal, J. L. F.; Sanz, E.; García Fernández, R.; Vega, C. J. Chem. Phys. **2005**, 122, 234511.

(6) Verlet, L. Phys. Rev. **1967**, 159, 98.

(7) Darden, T.; York, D.; Pedersen, L. J. Chem. Phys. **1993**, 98, 10089–10092.

(8) Miyamoto, S.; Kollman, P. A. J. Comput. Chem. **1992**, 13, 952–962.

(9) Kumar, A.; Molinero, V. J. Chem. Phys. Lett. **2017**, 8, 5053–5058.

(10) Stillinger, F. H.; Weber, T. A. Phys. Rev. B **1985**, 31, 5262.

(11) Love, J. C.; Estroff, L. A.; Kriebel, J. K.; Nuzzo, R. G.; Whitesides, G. M. Chem. Rev. **2005**, 105, 1103–1170.

(12) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. J. Am. Chem. Soc. **1996**, 118, 11225–11236.

(13) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. J. Comput. Chem. **2004**, 25, 1157–1174.

(14) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. J. Comput. Chem. **1997**, 18, 1463–1472.

(15) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. J. Chem. Phys. **1995**, 103, 8577.

(16) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Proteins: Structure, Function, and Bioinformatics **2010**, 78, 1950–1958.

(17) Geiger, P.; Dellago, C. J. Chem. Phys. **2013**, 139, 164105.

(18) others,, et al. Keras. `https://keras.io`, 2015.

(19) Kingma, D. P.; Ba, J. arXiv preprint arXiv:1412.6980 **2014**,