Molecular Dynamics and Machine Learning for Reaction and Nanomaterial Design

by

Alex K. Chew

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemical & Biological Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 05/05/2021

The dissertation is approved by the following members of the Final Oral Committee:

Reid C. Van Lehn, Assistant Professor, Chemical & Biological Engineering
Victor M. Zavala, Associate Professor, Chemical & Biological Engineering
David M. Lynn, Professor, Chemical & Biological Engineering, Chemistry
Joel A. Pedersen, Professor, Soil Science, Chemistry

To my highschool sweetheart, Grace Nguyen, my family, and my late undergraduate adviser, Abhijit Mitra.

ACKNOWLEDGMENTS

This PhD would not be possible without the support of my mentors, group members, family, department, and loved ones. I am beyond grateful for all the help that I received throughout these 5 years, and I hope to express my gratitude here.

First and foremost, I am grateful for my adviser, Reid Van Lehn, who always had the best research ideas that helped shape my projects; while at the same time, he gave me enough room to grow as an independent researcher. I feel very fortunate to be one of the first students to work with Reid, who turned out to be an amazing mentor, teacher, and researcher. Reid was always ready with a new idea whenever I stumbled onto a roadblock in my models, so I am thankful for his suggestions to improve them. He also has an incredible knack of developing understandable scientific images, and I hope to foster that skill throughout my career. Reid has also patiently dealt with my many mistakes and encouraged high-quality research through free food, which I will always remember and hope to carry on these characteristics whenever I am a mentor.

I am also thankful for my other mentors: Professors George Huber, Jim Dumesic, Victor Zavala, and Joel Pedersen. I enjoyed our collaborative projects that both elevated my research and expanded my knowledge in science. I am profoundly grateful to have collaborated with Theodore Walker (Ted) from the Huber and Dumesic group, which led to more than 6 peer-reviewed publications and several chapters within this disseration. From our work together, Ted and I developed a strong collaborative relationships and, even, a life-long friendship. My collaboration with Ted taught me the skills to develop productive collaborations with other groups. As a result, I am thankful to have collaborated with Shengli Jiang (Bruce) from the Zavala group and Christian Lochbaum from the Pedersen group; both collaborations have led to joint publications, where I learned

a great deal of machine learning and experimental methods.

I would also like to express my gratitute to my fellow group members (Brad Dallin, Samartha Patel, Jonathan Sheavly, Atharva Kelkar, Zhizhang Shen, Jianping Li, Zeynep Sumer, Amy Qin, and Panzheng Zhou) for their daily support and for making my work enjoyable. I enjoyed joking around in the office and giving each other feedback to improve ourselves. Together, we have created a Wikipedia page to share code, presentations, and tutorials, which sets the foundation for the next set of Van Lehn group members. We have also built the lab together from ground-up, which was an amazing learning experience for me and has taught me transferable skills that I would forever be grateful for.

I am deeply grateful to my PhD defense committee members: Reid Van Lehn, Victor Zavala, David Lynn, and Joel Pedersen. Thank you (committee members) for your invaluable time to review my dissertation and provide feedback on my defense!

I am thankful for my fiancée, Grace Nguyen, for her never-ending support before and throughout my graduate career. She has been with me since the beginning and supported me daily. Grace is also the reason for any baked goods in the Van Lehn office, which I often say I contributed to, but I honestly just washed the dishes. Grace has brought me joy everyday and provided motivation for me to finish my PhD, so I am humbly grateful for her constant support. I would argue that behind every great scientist is a supportive significant other that made everything possible, and Grace is that person for me.

I am also grateful for my previous mentors and friends from my undergraduate institution, NYU Tandon School of Engineering. I was originally motivated to pursue a PhD during my undergraduate research with my late adviser, Professor Abhijit Mitra. During my research as an undergraduate student, I learned the discipline and focus necessary to finish a research project, and I applied these tools throughout my graduate career.

Professor Mitra's passing due to cancer is what led me to pursue research with Reid at UW-Madison, geared towards using computational tools to guide materials design for cancer cell detection. I will be forever grateful for the opportunity to work with Professor Mitra, and I hope to carry on his legacy and continue "pushing the edge of science," as he would say. I am thankful for my undergraduate mentors (Professors Jin Kim, Jovan Mijovic, Bruce A. Garetz, Rastislav Levicky, and Iwao Teraoka) and friends (ShawnTina Harrod, Carmen Villafane, Volodymyr Krynytskyy, Tatyana Averbukh, Lyubov Chigirinskaya, and Magda Guadalupa), who helped guide me through my undergraduate career that led me to graduate school at UW-Madison.

Finally, many thanks to my family and friends who have supported me throughout these 5 years!

I am grateful for the support by: (1) Department of Chemical and Biological Engineering at the University of Wisconsin-Madison and the Wisconsin Alumni Research Fund; (2) UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science; (3) Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562; (4) Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-SC0018409 and DE-FC02-07ER64494.

Go Badgers!

CONTENTS

Contents **v**

List of Tables **xiv**

List of Figures **xvi**

Abstract **xli**

1 Introduction **1**

1.1 Motivation 1

1.2 Solvent screening for biomass conversion reactions 3

1.2.1 Motivation 3

1.2.2 Background 5

1.2.3 Computational studies on solvent effects in acid-catalyzed reactions 8

1.2.4 Hypothesis and research questions 9

1.3 Screening monolayer-protected gold nanoparticle properties for biomedical applications 10

1.3.1 Motivation 10

1.3.2 Background 13

1.3.3 Design considerations to improve GNP blood circulation lifetimes 14

1.3.4 GNPs in the biological environment: challenges of the corona 15

1.3.5 Hypothesis and research questions 17

1.4 Scope of this disseration 17

1.5 References 22

2 Computational methods **28**

2.1 Classical MD simulations 28
2.1.1 Benefits and limitations 31
2.1.2 Solvation free energy calculations 33
2.1.3 Umbrella sampling simulations 35
2.2 Molecular descriptors and quantitative structure-property relationships 37
2.3 Machine learning models 39
2.3.1 Deep learning models improve prediction accuracy 43
2.4 References 44

3 Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates 49
3.1 Introduction 50
3.2 Methods 53
3.2.1 Reaction kinetics studies 53
3.2.2 Molecular dynamics simulations 54
3.2.3 Hydrogen bonding lifetime 56
3.3 Results and Discussion 57
3.3.1 Universal effects of reactant and cosolvent properties on reactivity 57
3.3.2 Proposed mechanism: reaction rates correlate with formation of water-enriched local solvent domain . 63
3.3.3 MD simulations: formation of water-enriched local domains in solvent mixtures 66
3.3.4 Quantifying water enrichment within the local domain of the reactant 68
3.3.5 Quantifying reactant-water hydrogen bonding strength 71
3.3.6 Multidescriptor correlation between experimental and simulation results 73
3.4 Summary 81
3.5 References 82

4 Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations 91
4.1 Introduction . . . 92
4.2 Methods . . . 99
4.3 Results . . . 102
4.3.1 Comparison between experimental reaction rates and preferential exclusion coefficient . . . 102
4.3.2 Solvation free energy of the hydronium ion in pure solvent systems . . . 107
4.3.3 Solvation free energy of the chloride ion in pure solvent systems . . . 110
4.3.4 Relationship between solvation free energies and solvent parameters . . . 112
4.3.5 Solvation free energies of the hydronium and chloride ion in mixed-solvent systems . . . 115
4.4 Discussion . . . 119
4.4.1 Screening solvent properties using a classical hydronium ion model . . . 119
4.4.2 Incorporation of the stability of the hydronium ion to the predictive model of reaction rates . . . 121
4.5 Summary . . . 125
4.6 References . . . 126

5 Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks 134
5.1 Introduction . . . 135
5.2 Methods . . . 143
5.2.1 Classical molecular dynamics simulation methods . 143
5.2.2 Summary of model training, validation, and testing 144
5.2.3 3D CNN architectures . . . 147

5.2.4 Generation of saliency maps 148
5.3 Results and Discussion 149
5.3.1 Experimental reaction data used for model training and testing 149
5.3.2 Baseline prediction models for reaction rates using human-selected descriptors from classical MD . . . 150
5.3.3 Generation of input data set for interpretation by 3D CNNs using classical MD data 155
5.3.4 3D CNNs improve reaction rate predictions with less required simulation time 160
5.3.5 Generalizability of SolventNet predictions to new solvents and reactants 164
5.3.6 Physical interpretation of SolventNet features 171
5.4 Summary 173
5.5 References 175

6 Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions 182
6.1 Introduction 184
6.2 Methods 188
6.2.1 Reaction kinetics studies 188
6.2.2 Molecular dynamics simulations 190
6.2.3 *Ab initio* calculations 192
6.3 Results and Discussion 192
6.3.1 Proposed reaction mechanism for the acid-catalyzed dehydration of 1,2-propanediol 192
6.3.2 Product selectivities for the acid-catalyzed dehydration of 1,2-propanediol 194
6.3.3 Relating selectivity trends to tabulated cosolvent properties 198

6.3.4 Equilibrium solvation: quantifying reactant and product solvation free energies . . . 202
6.3.5 Equilibrium solvation effects extended to diol dehydration . . . 208
6.3.6 Competition of DMSO for hydroxyl groups on reactant 210
6.3.7 Hypothesized participation of DMSO in dehydration reaction . . . 216
6.4 Summary . . . 221
6.5 References . . . 223

7 Rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using a combined experimental, molecular dynamics and machine learning approach 233
7.1 Introduction . . . 234
7.2 Materials, methods, and definitions . . . 238
7.2.1 Materials . . . 238
7.2.2 Methods . . . 238
7.2.3 Classical molecular dynamics simulations . . . 241
7.3 Computational design tools and a general procedure for screening mixed solvent systems for biomass conversion processes . . . 244
7.3.1 Step 1: Establish reaction network and pre-select candidate solvent systems . . . 247
7.3.2 Step 2: Compute σ to screen mixed solvent systems for improved reactivity . . . 248
7.3.3 Step 3: Compute ΔΔG to estimate selectivities within reaction networks . . . 250
7.3.4 Step 4. Probe selected solvent systems using experiments . . . 254
7.4 Case Study 1: cyclohexanol dehydration to cyclohexene . . . 255
7.5 Case study 2: fructose dehydration to 5-hydroxymethylfurfural 262

7.6 Summary . . . 269
7.7 References . . . 270

8 Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles 278
8.1 Introduction . . . 280
8.2 Methods . . . 283
8.2.1 Gold core . . . 284
8.2.2 Self-assembly process . . . 286
8.2.3 Ligand-exchange and production simulations. . . . 288
8.2.4 Simulation parameters . . . 289
8.3 Results . . . 290
8.3.1 Effect of core morphology on the number of grafted ligands . . . 290
8.3.2 Effect of core morphology on SAM structure . . . 292
8.3.3 Identification of ligand bundles . . . 297
8.3.4 Relationship between gold core faceting and ligand bundling . . . 300
8.3.5 Effect of bundling on ligand properties . . . 303
8.4 Discussion . . . 306
8.5 Summary . . . 309
8.6 References . . . 311

9 Lipophilicity of cationic ligands promotes irreversible adsorption of nanoparticles to lipid bilayers 319
9.1 Introduction . . . 321
9.2 Materials and Methods . . . 324
9.2.1 Ligand Synthesis . . . 324
9.2.2 Gold Nanoparticle Synthesis . . . 325
9.2.3 Ligand Exchange Reaction . . . 325

9.2.4 Characterization of AuNP Hydrodynamic and Electrokinetic Properties . . . 326
9.2.5 Calculation of Ligand R group Lipophilicity 326
9.2.6 Vesicle Preparation and SLB Formation 327
9.2.7 Nanoparticle Adsorption and Desorption Experiments 329
9.2.8 System Setup for Classical MD Simulations 330
9.2.9 Increasing- and Decreasing-z US Simulations and Subsequent Unbiased Simulations 332
9.2.10 Quantifying the Number of Hydrophobic Contacts 333
9.2.11 Hydrophobic Contact US Simulations 334
9.2.12 Simulation Parameters 335
9.3 Results and Discussion 336
9.3.1 Ligand R Group Lipophilicity Governs AuNP Adsorption to Lipid Bilayers 336
9.3.2 Contact between Lipophilic Ligand Groups and Lipid Tails Promotes Adsorption 342
9.3.3 Unbiased Simulations Reveal Two Mechanisms Leading to Prolonged AuNP Adsorption to Bilayers . . . 345
9.3.4 Ligand Intercalation within Bilayer Reduces Barrier for Forming Hydrophobic Contacts 348
9.3.5 Two-state Adsorption Kinetics Describes AuNP Reversibly and Quasi-Irreversibly Adsorbed States . . 352
9.4 Summary 356
9.5 References 357

10 The interplay of ligand properties and core size dictates the hydrophobicity of monolayer-protected gold nanoparticles 364
10.1 Introduction 365
10.2 Methods 370
10.2.1 Planar SAM simulations 370
10.2.2 GNP simulations 373

10.2.3 Propane-water simulations 374
10.2.4 Simulation parameters 375
10.3 Results and Discussion 375
10.3.1 Hydration free energies for planar SAMs predict experimental measurements 375
10.3.2 Local hydration free energies capture hydrophobicity trends in planar SAMs 380
10.3.3 Ligand order in polar-terminated ligands results in broad hydration free energy distributions 384
10.3.4 Inducing surface curvature by tuning GNP size affects hydration free energies 387
10.3.5 Adding unsaturated or branched groups results in spatially distributed clusters 390
10.3.6 Relationship between hydration free energy maps and competitive binding of solvents 394
10.4 Summary 397
10.5 References 398

11 Molecular dynamics-derived descriptors for predicting ligand-coated gold nanoparticle behavior 408
11.1 Introduction 409
11.2 Methods 414
11.2.1 Development of nanoparticle systems 414
11.2.2 Simulation parameters 415
11.2.3 Computing MD-derived descriptors 416
11.2.4 Computational models 416
11.3 Results and discussion 418
11.3.1 Experimental datasets used to develop QNAR relationships 418
11.3.2 Computational workflow to model GNP systems with MD and compute MD-derived descriptors . . . 421

11.3.3 QNAR models using MD-derived descriptors accurately predicts GNP behavior 427
11.3.4 Consensus feature selection informs on GNP design 430
11.4 Model transferability to new datasets 436
11.5 Summary 440
11.6 References 442

12 Conclusion and future research **451**
12.1 Solvent screening for biomass conversion reactions 451
12.1.1 Future work 1: Incorporating product states into convolutional neural networks 453
12.1.2 Future work 2: Integrating process models to fine-tune solvent selection 455
12.2 Screening monolayer-protected gold nanoparticle properties for biomedical applications 456
12.2.1 Future work 1: Develop model to predict protein corona formation around GNPs 458
12.2.2 Future work 2: Incorporate deep learning models to rapidly predict GNP properties 464
12.2.3 Future work 3: Employ active learning to suggest new GNPs for experimentalists 465
12.3 References 466

LIST OF TABLES

2.1 Examples of molecular descriptors at different dimensions. . . 38

3.1 Kinetic solvent parameters for the Brønsted-acid-catalyzed dehydration of XYL in mixtures of water with three organic cosolvents. Reaction rate constant of XYL in water is: $k_{H_2O}^{XYL} = 1.04 \times 10^{-4} \pm 9.3 \times 10^{-6} L\ mol^{-1} s^{-1}$ *Reaction conditions: 403 K; 0.015-1.3 M triflic acid. Confidence intervals were calculated at the 95% confidence level.* N/A means not available due to phase separation between water and organic cosolvent. 61

3.2 Pearson correlation coefficient between the kinetic solvent parameters (σ) and the preferential exclusion coefficient (Γ) for various reactants and cosolvents. 71

3.3 Best-fit slope and root-mean-square error (RMSE) between predicted kinetic solvent parameters (σ_{pred}) and experimental kinetic solvent parameters (σ_{exp}). N/A are omitted values since only three experimental values are available, resulting in an exact solution to Equation 3.9. THF cosolvent mixture results are omitted for the same reason. 75

3.4 Coefficients of the multidescriptor correlation model describing the rates of all seven reactions in each of the three cosolvent mixtures with the best-fit slope and root-mean-square error (RMSE) between predicted (σ_{pred}) and experimentally determined (σ_{exp}) kinetic solvent parameters. The descriptors $\widetilde{\Gamma}$, $\widetilde{\tau}$, and $\widetilde{\delta}$ are equal to Γ, τ, and δ but re-scaled to values between 0 and 1 so the coefficients are comparable. 80

4.1 Dielectric constants and Kamlet-Taft parameters (α, β, π^*) for pure solvents, tabulated according to decreasing transfer free energy of a hydronium ion $\Delta G_{H_3O^+}^{k \to H_2O}$ in these solvents. $\sum \Delta G$ was computed with Equation 4.2. All solvation free energies are in units of kJ/mol. 114

6.1 Apparent rate constants (k_{app}), product kinetic solvent parameters (σ^{PRO}), and selectivities of Brønsted acid-catalyzed 1,2-propanediol (PDO) dehydration in mixed-solvent environments. The cosolvent mass fraction was 90 wt% in all experiments (excluding pure water). Rate constants are derived from Equations 6.1 and 6.2. Selectivities are derived from Equation 6.3. The standard error in selectivities is $\pm$ 5 mol%. *Reaction conditions: ~20 mol% conversion; 433 K; 0.4 M triflic acid; 0.1 M PDO; 150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.* 201

6.2 Dehydration of diols over triflic acid in 90 wt% GVL-water and DMSO-water mixtures. *Reaction conditions: 0.1 M reactant, 433 K, 0.05-0.6 M triflic acid, 500 rpm stirring, reaction time 0-120 min.* 210

LIST OF FIGURES

1.1 Major platform chemicals produced from lignocellulosic biomass. This image was adapted from Ref. 10 with permission from John Wiley and Sons. 5
1.2 Reaction pathway for the decomposition of cellulose in presence/absence of low water concentrations. Cellulose is comprised of ~50% lignocellulosic biomass. 5-hydroxymethylfurfural is highlighted in red, since it is a platform chemical for transportation fuels and a desirable product from biomass conversion processes. This image was adapted from Ref. 18 with permission from The Royal Society of Chemistry. 7
1.3 (a) Applications of GNPs in cancer therapy. This image was adapted from Ref. 36 with permission from Elsevier. (b) Structure of self-assembled monolayers (SAM) ligands, consisting of a sulfur head group, nonpolar backbone, and terminal end group. These ligands spontaneously form SAMs in the presence of a gold core or a flat gold. (c) Examples of the tunable parameters of GNPs, such as ligand backbone, ligand end groups, and gold core size. 12
1.4 Active (solid) and passive (dashed) targeting using GNPs. This image was adapted from Ref. 36 with permission from Elsevier. 14
1.5 Computational overview for applications in solvent screening for biomass conversion reactions. 19
1.6 Computational overview for applications in screening monolayer-protected gold nanoparticle properties. 21

2.1 (a) Schematic overview of MD simulations using four water molecules as an example. (b) Basic algorithm for MD. The abbreviations are: x_i is the atomic position of atom i, E_{pot} is the potential energy, F_i is the force on atom i, m_i is the mass of atom i, t is the current time, Δt is the time step deviation, v_i is the velocity of atom i. This image was adapted from Ref. 1 with permission from Dove Medical Press Limited. 31

2.2 Solvation free energy schematic of transferring a reactant from vacuum to pure water by slowly introducing VDW and Coulombic interactions between reactant and solvent. The solvation free energy is denoted as $G_r^{H_2O}$. 34
2.3 Schematic for computing the free energy of moving a GNP to lipid bilayer using US simulations. 36
2.4 Overview of QSPR modeling with examples of linear and nonlinear models for a binary classification of two properties. . . . 39
2.5 Different nonlinear approaches for developing QSPR models: (a) *k*th nearest neighbor, (b) support vector machines, (c) random forest, (d) neural networks, and (e) convolutional neural networks. 42
2.6 Historical error rate of the best-performing image classification model in the annual ImageNet competition. Deep learning significantly improved the error rate to ~15% in 2012 and reaching close to human-level accuracy ~5% by 2015. This image was adapted from Ref. 31 with permission from John Wiley and Sons. 44

3.1 Schematic depiction of mixed-solvent system preparation. "R" denotes the reactant. Note that the second production trajectory, used to calculate hydrogen bonding lifetimes, was not always 4 ns; some reactants required a longer simulation time to obtain accurate hydrogen-bonding lifetime data. 56
3.2 Brønsted acid-catalyzed reactions of seven model compounds. Rate constants associated with reactions 3, 4, and 6 were taken from prior work.[11,37] Hydroxyl groups are highlighted in red for emphasis. Reactants are arranged according to decreasing hydrophilicity, as estimated by the δ parameter. 58
3.3 Apparent rate constant for XYL dehydration normalized by the rate constant in pure water versus the mass fraction of the organic cosolvent in DIOX-water mixtures. *Reaction conditions: 75 - 200 mM XYL; 0.03 - 1.3 M trifluoromethane sulfonic (triflic) acid; 403 K.* 60
3.4 Kinetic solvent parameters (σ^i_{org}) as a function of: (a) solvent composition in aqueous mixtures of DIOX, GVL, and THF (open symbols = TBA, closed symbols = XYL), and; (b) the accessible hydroxyl fraction (δ) in DIOX-water mixtures. 62

3.5 (a) Role of cosolvent molecules on the distribution of solvent molecules. Favorable interactions with hydrophilic reactants in mixed-solvent environments drive the formation of water-rich local domains around the reactant. While there are fewer water molecules in the local domain relative to pure water, the local water density is enriched relative to the bulk density in the solvent mixture. (b) Proposed effect of cosolvent molecules on a reaction free energy landscape. Stabilization of the proton and transition state in the water-rich local domain, relative to the bulk domain, lowers the apparent free energy barrier for the reaction in a mixed-solvent environment. (c, d) MD simulation snapshots of XYL in (c) pure water and (d) 90 wt% DIOX, which is drawn as a single representative bead to match the schematics in (a). 65

3.6 (a) Schematic depiction of the radial distribution function for XYL. The distance, r, is calculated between the center of mass of the reactant and the oxygen atom of each water molecule. (b) Radial distribution function for XYL in 90 wt% DIOX and pure water ($m_{DIOX} = 0$). The cutoff between local and bulk domain is defined as the distance when the RDF between the reactant and water reaches unity (*i.e.* a random mixture). (c,d) Radial distribution function for TBA (c) and XYL (d) for various wt% of organic solvent. 67

3.7 Relationship between experimentally determined kinetic solvent parameters (σ) and simulated preferential exclusion coefficient (Γ) for (a) TBA and (b) XYL for various wt% of organic cosolvent in GVL-water and DIOX-water mixtures. The gray dotted line denotes when σ and Γ are zero. Kinetic solvent parameters are also plotted against Γ for (c) TBA and (d) XYL. The dashed lines in (c) and (d) represent the best-fit line. Data points are labeled with the wt% of the organic cosolvent. Pure water systems have been omitted from parts (c) and (d) because σ and Γ will always be zero. 70

3.8 Average reactant-water hydrogen bond lifetimes (in picoseconds) for TBA and XYL as a function of mass fraction of organic cosolvent in DIOX-water, GVL-water, and THF-water mixtures. 73

3.9 (a) Comparison of kinetic solvent parameters calculated using the multidescriptor correlation model (σ_{pred}) to experimentally determined values (σ_{exp}) for all seven reactants in DIOX-water mixtures. Each reactant has four data points for 0.25, 0.50, 0.75, and 0.90 mass fractions of DIOX, with the exception of PDO (see Table S1).[2] The slope of the best-fit line for all the data points and the average root-mean-squared error (RMSE) between the values of σ_{pred} and σ_{exp} are shown at the bottom right. The solid black line indicates a perfect correlation ($\sigma_{pred} = \sigma_{exp}$) and dotted lines are drawn at $\sigma_{exp} = 0$ and σ_{pred} to help visualize false positive/negative predicted values. Lines above and below the $\sigma_{pred} = \sigma_{exp}$ line are shifted by ± 0.10, denoting the approximate experimental error. (b) Prediction of kinetic solvent parameters using FRU as a test set with all other reactants taken as a training set. The slope of the best-fit line and RMSE between the values of σ_{pred} and σ_{exp} for the training and test sets are shown at the bottom right. 79

4.1 (a) Brønsted acid-catalyzed reaction of 1,2-propanediol (PDO) to propanal. (b) Schematic of acid-catalyzed reactions in mixed-solvent environments. The reaction proceeds through a charged transition state (TS) formed from the protonation of the reactant by a hydronium ion catalyst. The corresponding free energy diagram schematically depicts the influence of mixed-solvent environments (red and green lines) on acid-catalyzed reactions relative to pure water (black line). Note that this is a generalized representation of a free energy landscape based on prior computational findings[4] and is not specific to 1,2-propanediol dehydration reaction. 98

4.2 (a) Schematic representation of simulation workflow for molecular dynamics and free energy simulations. M denotes either 1,2-propanediol, a hydronium ion, or a chloride ion. (b) Simulation snapshots of hydronium ion in pure water, 90 wt% DIOX, and 90 wt% DMSO. The hydronium ion is located at the center and only solvent molecules within a 5 Å radius is shown. . . . 102

4.3 Radial distribution function between the center of mass of PDO and water in 90 wt% DIOX, 90 wt% DMSO, and pure water. Local and bulk domain cutoffs were determined as the value of r for which the RDF reaches unity. Bin widths for the RDFs were set to 0.02 nm. 104
4.4 (a) Relationship between simulated preferential exclusion coefficient (Γ) and experimentally determined kinetic solvent parameters (σ) for aqueous mixtures of DIOX and DMSO. Experimental values were taken from Ref. 4. (b) Correlation between Γ and σ for aqueous mixtures of DIOX and DMSO. Data points are labeled with the wt% of the organic solvent. 25, 50, 75, and 90 wt% organic solvent was used to correlate Γ and σ as indicated for each point. The best-fit line is drawn and labeled with the corresponding equation and Pearson's r. 107
4.5 (a) Chemical structures of the organic solvents used in this study. (b) Thermodynamic cycle used to compute the free energy for transferring a hydronium or chloride ion from pure water to pure organic solvent. ΔG_j^k and $\Delta G_j^{H_2O}$ are solvation free energies while $\Delta G_i^{H_2O \to k}$ is the transfer free energy computed from Equation 4.1. (c) Transfer free energies for six pure organic solvents. Cyan and purple bars indicate hydronium (H_3O^+) and chloride (Cl^-) ion transfer free energies, respectively. Dashed lines indicate the sum of the transfer energies. Error bars were computed from the standard deviation of two trials; the error is less than 1 kJ/mol and is not visible in the plot. The error is tabulated in Supplementary Table 4.[2] 110
4.6 (a) Solvation free energies for transferring hydronium (H_3O^+, filled lines) and chloride (Cl^-, dashed lines) ions from pure water to aqueous mixtures of dioxane (DIOX, blue lines) and dimethyl sulfoxide (DMSO, red lines). The sums of the transfer free energies ($\sum \Delta G$) are shown as green lines. Error bars are not shown; they range from 0-2.5 kJ/mol when averaging two trials and tabulated in Supplementary Table 5. (b) Radial distribution function (RDF) between the center of mass of the hydronium ion to water (top) and the organic solvent (bottom) in 90 wt% DIOX, 90 wt% DMSO, and pure water. Bin widths for the RDFs were set to 0.02 nm. 118

4.7 (a) Correlation between simulated preferential exclusion coefficient with solvation free energy correction term (Γ'), as expressed in Equation 4.4, and experimentally determined kinetic solvent parameters (σ) for aqueous mixtures of DIOX and DMSO. Experimental values were taken from Ref. 4. Transparent red points and lines show how Γ relates to Γ'. (b) Parity plot between predicted kinetic solvent parameter (σ_{pred}) and experimental kinetic solvent parameter (σ_{exp}) using results from aqueous mixtures of DIOX and DMSO. The predictive model is based on Equation 4.6 as shown above the plot. Data points are labeled with the wt% of the organic solvent. 125

5.1 Overview of solvent effects on acid-catalyzed reactions and model systems. (a) Two example acid-catalyzed reactions: xylitol (XYL) dehydration and levoglucosan (LGA) hydrolysis. (b) Hypothesized effect of mixed-solvent environments on the free energy landscape of acid-catalyzed reactions. The schematic illustrates the formation of a local solvent domain (within the circular dashed line) around the reactant in a mixed-solvent environment that modifies the reaction free energy landscape, thus affecting reaction kinetics.[6,7] (c) Organic, polar aprotic cosolvents modeled in this study, including dioxane (DIOX), γ-valerolactone (GVL), tetrahydrofuran (THF), dimethyl sulfoxide (DMSO), acetonitrile (MeCN), and acetone (ACE). Molecules drawn in black were included in the training set. Molecules drawn in gray were included in the test set. (d) Biomass-derived model reactants modeled in this study, including ethyl tert-butyl ether (ETBE), tert-butanol (TBA), cellobiose (CEL), glucose (GLU), LGA, 1,2-propanediol (PDO), fructose (FRU), and XYL. The color scheme follows part c, except TBA, PDO, and FRU were included as part of some of the reactant-solvent combinations in both training and test sets. 142

5.2 Evaluation of human-selected multidescriptor models. (a) General approach for correlating features from molecular dynamics (MD) simulations to experimental kinetic solvent parameters (σ_{exp}). The simulation configuration shows xylitol (XYL) in 90 wt% dioxane (DIOX) as an example. (b) Schematic illustrating 5-fold cross validation procedure used to train and validate models. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the multidescriptor linear model. The best-fit slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within the plot. The solid black lines indicate perfect correlation ($\sigma_{pred} = \sigma_{exp}$), the dashed black lines indicate approximate experimental error, and the dashed gray lines are drawn at $\sigma_{exp} = 0$ and $\sigma_{pred} = 0$ as a guide to the eye. (d) Parity plot for the nonlinear fully connected neural network model. 153

5.3 Input data representation for 3D CNNs. Approach for converting the atomic positions obtained from a MD simulation to a voxel representation using xylitol in 90 wt% dioxane as an example. (a) For each MD configuration (example at left), a (4 nm)^3 cubic box was centered on the reactant and a $20 \times 20 \times 20$ grid of (0.2 nm)^3 volume elements was used to discretize space. The normalized occurrences of water, oxygens of the reactant, and cosolvent atomic positions within each volume element were stored in different channels to yield a $20 \times 20 \times 20 \times 3$ grid of voxels. Voxels are visualized by showing the water channel in red, the reactant channel in green, and the cosolvent channel in blue. Half of the voxels are transparent to illustrate the solvent distribution around the reactant. (b) Grids of voxels were averaged over 2 ns of MD data (200 MD configurations) to yield a $20 \times 20 \times 20 \times 3$ voxel representation. (c) For each reactant-solvent system, 20 ns of simulation data were used to generate 10 independent voxel representations. 159

5.4 Architecture, training, and performance of 3D CNNs. (a) Architecture of SolventNet, a 3D CNN that inputs a $20 \times 20 \times 20 \times 3$ voxel representation (described in Fig 3) and outputs the predicted kinetic solvent parameter (σ). 3D CNNs were evaluated using the same 5-fold cross validation procedure described in Figure 5.2b. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters using SolventNet. σ_{pred} is the average prediction of 10 voxel representations per label. Error bars show the standard deviation of these predictions. The best-fit slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within the plot. Solid and dashed lines follow the conventions of Figure 5.2. (d) Comparison of the RMSEs between σ_{pred} and σ_{exp} for the multidescriptor linear and nonlinear neural network (NN) models and the 3D CNNs (ORION, VoxNet, and SolventNet) when performing 5-fold cross validation. 161

5.5 Generalizability of SolventNet to new reactants and cosolvents. Parity plots between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the test set. Predictions were made using SolventNet after training with all training set data. σ_{pred} is the average prediction of 2 voxel representations per label. Error bars show the standard deviation of these predictions. The slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within each plot. Solid and dashed lines follow the convention of Figure 5.2. 167

5.6 Leave-one-out cross validation of SolventNet. (a) Schematic illustrating the leave-one-out cross validation procedure in which all data for a single cosolvent or all data for a single reactant were used as the test set and the remaining data were used as the training set. (b) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the leave-one-out cross validation of SolventNet across cosolvents. The RMSE values between σ_{pred} and σ_{exp} labeled within each plot report the values obtained when data for the listed cosolvent-water system were used to as the test set. σ_{pred} is the average prediction of 10 voxel representations per label. Error bars show the standard deviation of these predictions. Solid and dashed lines follow the conventions of Figure 5.2. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the leave-one-out cross validation of SolventNet across reactants. The RMSE values in the table report the values obtained when data for the listed reactant were used as the test set. 170

5.7 Saliency map using SolventNet. Example saliency map generated for a voxel representation of XYL in 90 wt% DIOX (shown in Figure 5.4a) using SolventNet after training with all training set data. The saliency map is visualized on a 3D grid with the same dimensions as the input voxel representation. Each voxel is assigned a saliency value normalized from 0 to 1 that indicates the sensitivity of SolventNet predictions to the normalized occurrences of water, reactant, and cosolvent atoms in that voxel. Larger saliency values indicate greater sensitivity. The saliency map is visualized by separate grids showing the water value in red, the reactant value in green, and the cosolvent value in blue. Half of the voxels are transparent to illustrate the saliency values around the reactant and only the voxels with values greater than 0.10 for each system component are shown. 2D contours are plotted by averaging along the z-axis for normalized values of 0.10, 0.25, 0.50, 0.75, and 0.90.172

6.1 **A** Proposed mechanism for Brønsted acid-catalyzed dehydration of 1,2-propanediol (PDO) to afford either propanal (PRO) or acetone (ACE) in the gas phase over a solid acid catalyst.[62] Red arrows denote the transfer of electrons. **B** Polar aprotic cosolvents used for this study. 194

6.2 Apparent rate constants (k_{app}, dashed lines) and selectivities to propanal (PRO, blue columns) and acetone (ACE, red columns) for Brønsted-acid-catalyzed 1,2-propanediol (PDO) dehydration in mixtures of water with (**A**) 1,4-dioxane (DIOX) and (**B**) dimethyl sulfoxide (DMSO) as a function of the mass fraction of the organic component (m_{DIOX} or m_{DMSO}). Rate constants are derived from Equation 6.1 and selectivities are derived from Equation 6.3. The standard error in selectivities is $\pm$ 5 mol%. *Reaction conditions: ~20 mol% PDO conversion; 433 K; 0.4-0.005 M triflic acid; 0.01 M PDO; 90-150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.* 197

6.3 Apparent rate constants for reactant conversion and product formation for 1,2-propanediol (PDO) dehydration in mixtures of water with dimethyl sulfoxide (DMSO) as a function of the mass fraction of DMSO (m_{DMSO}). Dashed lines are visual aids. *Reaction conditions: ~20 mol% conversion; 433 K; 0.4-0.005 M triflic acid; 0.1 M PDO; 150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.* 198

6.4 **A** Hypothesized effect of mixed-solvent environment on the free energies of reactant, transition, and product states. The change in the relative free energy between the reactant and product states ($\Delta\Delta G$) is proportional to the change in the activation energy ($\Delta\Delta G_{act}$) for a reaction in a mixed-solvent environment compared to the same reaction in pure water. Free energies are schematically drawn relative to the free energy of the reactant state in each solvent environment. **B** Thermodynamic cycle to calculate the free energy difference between a reactant and product in a mixed-solvent environment relative to pure water. Purple arrows indicate solvation free energies computed from MD simulations which are used to calculate the transfer free energies indicated by filled black arrows. The dashed black arrow indicates $\Delta\Delta G$. **C** Simulated $\Delta\Delta G$ for PRO (red bar) and ACE (blue bar) and experimental kinetic solvent parameter for PRO formation (σ^{PRO}, dashed black lines) in 90 wt% organic cosolvents. **D** Correlation between σ^{PRO} and $\Delta\Delta G$ for 90 wt% mass fraction of organic solvent (black) and various wt% mass fractions of DIOX (orange). 207

6.5 **A** Four acid-catalyzed dehydration reactions of representative diols. **B** Experimental kinetic solvent parameter for dehydration product formation (σ^{p}) and $\Delta\Delta G$ for each of the four reactions in 90 wt% GVL- and DMSO-water mixtures. 209

6.6 **A** Spatial distribution maps of the *trans* vicinal diol conformation of PDO in pure water, 90 wt% DIOX, and 90 wt% DMSO. PDO is positioned so that the view angle is along the $C_1 - C_2$ bond as illustrated in the projection diagram at the left. **B** Spatial distribution maps of a *gauche* vicinal diol conformation of PDO in pure water, 90 wt% DIOX, and 90 wt% DMSO. PDO is positioned to the view angle is above the $C_1 - C_2$ bond as shown in the diagram at the left. In both A and B, normalized density isovalues between 1.5-3.0 are shown for water in red and isovalues between 1.3-1.5 are shown for cosolvent in blue. Dashed lines emphasize regions that remain enriched in water in DMSO-water mixtures, whereas other water-enriched regions are missing. Additional angles of these spatial distribution maps are available in Supplementary Movie S1.[2] 215

6.7 **A** Proposed mechanism of PDO dehydration to afford acetone in DMSO-water mixtures. Red arrows denote the transfer of electrons. **B** Relaxed structures of protonated primary or secondary hydroxyl groups of the *trans* vicinal diol conformation of PDO with two DMSO molecules and 1 water. **C** Relaxed structures of protonated primary or secondary hydroxyl groups of the *gauche* vicinal diol conformation of PDO with two DMSO molecules and 1 water molecule. In B and C, black arrows are drawn as a visual guide to identify PDO, the green atom is a proton, and cyan dashed lines indicate hydrogen bonds. 220

7.1 Overview of acid-catalyzed reactions in mixed solvent systems for biomass conversion. (a) Acid-catalyzed reaction example of xylitol (XYL) dehydration to afford 1-4-anhydroxylitol. (b) Example of three organic, polar aprotic cosolvents. (c) Schematic of acid-catalyzed reactions in a mixed solvent systems that proceeds through a charged transition state (TS), formed after the protonation of the reactant by a hydronium ion catalyst.[24] The schematic illustrates the formation of a local solvent domain around the reactant in a mixed solvent system that modifies the reaction free energy landscape, thus affecting reaction kinetics. These images were adapted with permission of 25,26. 237

7.2 Process to screen candidate mixed solvent systems for biomass conversion processes. An arbitrary reaction network as determined using experiments; as an example, product B is desired from reactant A. Polar aprotic cosolvents are selected to mix with water and test the effects of solvent composition on reaction performance. Kinetic solvent parameters (σ) are predicted using molecular dynamics simulations in conjunction with SolventNet, a machine-learning algorithm parameterized to predict kinetic solvent parameters as described in Figure 7.3. The top performing solvents are then tested to see if the reaction selectively forms product B by calculating relative solvation free energies ($\Delta\Delta G$). Negative values of $\Delta\Delta G$ indicate that product B is more stabilized in the mixed solvent system. The top mixed solvent system for the conversion of A and production of B is then selected for experimental testing. 246

7.3 Computational tools used to predict reaction rates and selectivities. (a) Conversion of atomic positions obtained from molecular dynamics simulation trajectories into a voxel representation using XYL in 90 wt% DIOX as an example. For each MD configuration, a $20 \times 20 \times 20$ grid of $(0.2\ \text{nm})^3$ volume elements was centered on the reactant. Voxel representations are visualized by showing the water channel in red, the reactant channel in green, and the cosolvent channel in blue. Half of the voxels are transparent to illustrate the solvent distribution around the reactant. (b) Architecture of SolventNet, a 3D CNN that inputs voxel representations and outputs the predicted kinetic solvent parameter (σ). (c) Hypothesized effect of mixed solvent systems on the free energy landscape of reactant, transition, and product states. The change in the relative free energy between the reactant and product states ($\Delta\Delta G$) is proportional to the change in the activation energy ($\Delta\Delta G_{act}$) for a reaction in a mixed solvent system compared to the same reaction in pure water. The free energies are drawn relative to the reactant state in pure water. (d) Thermodynamic cycle to calculate the free energy difference between a reactant and product in a mixed solvent system relative to pure water. Purple arrows indicate solvation free energies computed from MD simulations which are used to calculate the transfer free energies indicated by filled black arrows. The dashed black arrow indicates $\Delta\Delta G$. These images were adapted with permission of 26,28. 253

7.4 Case study of cyclohexanol conversion to cyclohexene. (a) Acid-catalyzed conversion of cyclohexanol to cyclohexene. (b) Ten organic, polar aprotic cosolvents considered for the initial library of mixed solvent systems. (c) Kinetic solvent parameters predicted by SolventNet for 75 wt% organic cosolvents. Black asterisks indicate solvent systems that are representative of good cosolvents (THF, GVL, and ACE), and poor cosolvents (DIOX and DMSO). (d) Relative solvation free energies ($\Delta\Delta G$) between product (cyclohexene) and reactant state (cyclohexanol) in 75 wt% organic cosolvents relative to pure water. (e) Comparison between kinetic solvent parameters as predicted by SolventNet (red) and determined by experiments (blue). (f) Comparison between $\Delta\Delta G$ (green bars) and experimental percent selectivity (black line) towards cyclohexene. Gray regions denote the error in selectivity measurements. 257

7.5 Case study of fructose conversion to HMF. (a) Acid-catalyzed dehydration of fructose (FRU) to afford 5-hydroxymethylfurfural (HMF), followed by an addition reaction to afford levulinic acid (LA). (b) Kinetic solvent parameters predicted by SolventNet for 90 wt% cosolvent-water mixtures. The difference between kinetic solvent parameters predicted for HMF and FRU is shown in the top axis ($\sigma_{HMF} - \sigma_{FRU}$), where more negative $\sigma_{HMF} - \sigma_{FRU}$ values indicate that the mixed solvent system is predicted to selectively form HMF from FRU. Black asterisks indicate mixed solvent systems that are representative of good cosolvents (THF, GVL, and ACE), and poor cosolvents (DIOX and DMSO). (c) Relative solvation free energies ($\Delta\Delta G_{FRU\rightarrow LA}$) between product (LA) and reactant state (FRU) in 90 wt% organic cosolvents relative to pure water. (d) Comparison between $\sigma_{HMF} - \sigma_{FRU}$ predicted from SolventNet (red) and experimentally determined (blue). (e) Comparison between $\Delta\Delta G_{FRU\rightarrow LA}$ and percent selectivity towards HMF. Gray regions denote the error in $\sigma_{HMF} - \sigma_{FRU}$ in (b) and selectivity measurements in (e). 268

8.1 Schematic of workflow for automatic GNP assembly. 284

8.2 **A** Alkanethiol ligand structures used in this study, where n denotes the number of methylene groups. **B** Snapshots of hollow gold core (HGC), spherical gold core (SGC), and faceted gold core (FGC) for various diameters. Each atom is colored by its coordination number, defined as the number of gold atoms within a cutoff of 0.41, 0.37, and 0.34 nm for HGC, SGC, and FGC, respectively. Colors correspond to coordination numbers of 1-6 (red), 7-8 (orange), 9 (green), 10 (blue) and 11 (violet). . 286
8.3 Number of adsorbed ligands for different gold core morphologies and sizes. Teal triangles are model data verified by thermogravimetric analysis from Ref. 65. Purple squares indicate data from MALDI-MS experiments from Ref. 66. The black line denotes the number of adsorbed ligands on a sphere with a planar gold (111) average surface density of 4.62 ligand/nm^2. Snapshots show gold cores colored with the same scheme as Figure 8.2 and the sulfur atoms of adsorbed ligands in yellow. The color of each arrow corresponds to the gold core shape (*i.e.* HGC, SGC, and FGC models are red, blue, and green, respectively). 292
8.4 **A** Simulation snapshots of 2 nm and 6 nm SGC and FGC GNPs with butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) ligands. Sulfur atoms are represented by yellow beads and carbon chains are represented by gray lines. Hydrogen atoms and surrounding water molecules are not drawn. Gold cores are colored according to coordination number as described in Figure 8.2. **B** Fraction of *trans* dihedral angles as a function of ligand chain length (n) for the SGC (solid lines) and FGC (dashed lines) models. **C** Eccentricity as a function of ligand chain length for the SGC and FGC models. The legend is the same for both plots. Simulation snapshots are shown for the 2 nm FGC model with octanethiol ($n = 7$) and hexadecanethiol ($n = 15$). 296

8.5 **A** Simulation snapshots of 2 nm and 6 nm SGC and FGC GNPs with butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) ligands. Ligands with the same color are in the same bundle, whereas ligands in gray are not in bundles. **B** Number of bundles as a function of ligand chain length (n) for the SGC (solid lines) and FGC (dashed lines) models. **C** Fraction of ligands in bundles as a function of ligand chain length for the SGC and FGC models. The legend is the same for both plots. 299
8.6 **A** Simulation snapshots of 6 nm butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) GNPs for the FGC model. Only the ligand sulfur atoms are shown. At top, each sulfur atom is colored according to if the ligand is in a bundle (black) or not (gray). At bottom, each sulfur atom is colored according to if the nearest gold atom is at an edge (magenta) or facet (cyan). Gold atoms are colored according to coordination number as in Figure 8.2B. **B** The fraction of ligands on facets for bundled (filled lines) and non-bundled (dashed lines) ligands as a function of ligand chain length. 302
8.7 Fraction of *trans* dihedrals for bundled (solid lines) and non-bundled (dashed lines) ligands as a function of ligand chain length (n) for the FGC model. 303
8.8 **A** SASA for bundled (solid lines) and non-bundled (dashed lines) ligands as a function of ligand chain length (n) for the FGC model. **B** Probability density function of the SASA for a 2 nm FGC GNP with butanethiol ligands (top) and a 6 nm FGC GNP with hexadecanethiol ligands (bottom). The probability density function is shown only for one of the three trials. Simulation snapshots are colored by (*i*) bundled (black) and non-bundled (grey) ligands, and (*ii*) ligand SASA. 305

9.1 Experimental and computational systems used to study gold nanoparticle adsorption onto phospholipid bilayers. (a) Ligands are comprised of an alkane group (gray), an oligo(ethylene glycol) spacer group (green), and a cationic quaternary ammonium group (red) substituted with the indicated R group and two methyl groups. The five R groups used are displayed in red and labeled with their calculated log $K_{lip\text{-}w}$ values in parentheses. (b) Schematic of the system used in quartz crystal microbalance experiments to measure nanoparticle adsorption to supported DOPC lipid bilayers. (c) Snapshot of 2-nm gold nanoparticle with C_{10} ligands placed above a DOPC lipid bilayer. The color scheme is illustrated for each of the components at right. The DOPC lipids are comprised of a zwitterionic phosphatidylcholine head group and nonpolar acyl tails consisting primarily of aliphatic carbon atoms. Water is shown in grey. . 338

9.2 Influence of ligand lipophilicity on AuNP attachment to supported DOPC bilayers as determined by QCM-D. (a) Acoustic surface mass density (Γ) maximum and after rinse. (b) Dissipation factor (ΔD) maximum and after rinse. Error bars represent one standard deviation from four replicate QCM-D experiments.341

9.3 Free energy as a function of the z distance between the AuNP and lipid bilayer. Potential mean force (PMF) versus z for C_1- and C_{10}-AuNPs when the gold core is (a) pulled towards (*i.e.* decreasing-z) and (b) away from (*i.e.* increasing-z) the DOPC lipid bilayer. Simulation snapshots show the last configuration from umbrella sampling simulations of C_{10}-AuNPs for different values of z. Water and chlorine atoms are omitted for clarity. Legends are the same for (a) and (b). (c) Number of hydrophobic contacts (c_h) versus z for both decreasing- and increasing-z simulations. Hydrophobic contacts are defined as the number of contacts between nonpolar groups in the ligands and in the DOPC tail groups. Error bars are reported as the standard deviation between two 20 ns blocks for each umbrella sampling window. 344

9.4 Unbiased simulations initiated from umbrella sampling trajectories. (a) Number of hydrophobic contacts (c_h) versus z for unbiased simulations initiated from increasing- and decreasing-z US configurations for C_1- and C_{10}-AuNPs. AuNPs are considered adsorbed if $z < 6$ nm for the last 10 ns of the unbiased simulation (filled markers) and desorbed if $z > 6$ nm (hollow markers). Points in the dashed blue box were for unbiased simulations initiated from increasing-z US configuration, all other points were for simulations initiated from decreasing-z US configurations. (b) Simulation snapshots after 50 ns of unbiased simulation for C_1- and C_{10}-AuNPs. The initial z values and final z values after 50 ns are labeled above the snapshots. Atoms in the DOPC head groups are omitted for clarity. 347

9.5 Free energy as a function of AuNP-bilayer hydrophobic contacts. (a) Potential of mean force (PMF) versus the number of hydrophobic contacts (c_h) for C_1-, Bn-, and C_{10}-AuNPs. Error bars are reported as the standard deviation between two 30 ns trajectories for C_1- and C_{10}-AuNPs and two 20 ns trajectories for Bn-AuNP in each umbrella sampling window. (b) Simulation snapshots with c_h = 5, 40, and 150 for C_1-, Bn-, and C_{10}-AuNPs. DOPC lipids that are within 0.35 nm of the ligand atoms are highlighted in cyan. 351

9.6 The role of ligand end group lipophilicity on adsorption to and desorption from phospholipid bilayers. (a) Adsorption rate constant (k_a) calculated from AuNP adsorption with calculated desorption rate constant (k_d) values. (b) Rate constants for conversion to quasi-irreversibly adsorbed state (k_β) calculated from mass at maximum and after rinse. (c) Desorption rate constants (k_d) calculated from AuNP desorption. Error bars represent one standard deviation of four replicate QCM-D measurements. (d) Schematic showing hypothesized mechanism for preferential adsorption of C_{10}-AuNPs compared to C_1-AuNPs. C_{10}-AuNPs have a longer R group, denoted by the red lines. Alkane is shown as black lines and PEG is shown as green lines. The symbol and color for the gold core and lipid bilayer is the same as Figure 9.1b. 355

10.1 Overview of monolayer-coated gold nanoparticles and planar gold surfaces. (a) Alkanethiol ligand chemical structures studied in this work, including ligands with saturated, unsaturated, and branched nonpolar backbones and six different end groups (indicated by R) of varying polarity. (b) Snapshot of the double planar self-assembled monolayer (SAM) simulations for saturated R = CH_3 ligands (i.e. dodecanethiol). (c) Snapshot of a 2 nm gold core coated with the same ligands as in (b) and snapshot of the equilibrated nanoparticle in water. 372

10.2 Quantifying the hydrophobicity of planar SAMs. (a) Snapshot showing the SAM-water interface (indicated as a purple surface) computed for a planar SAM with saturated R = CH_3 ligands. Two cavity geometries were used to calculate hydration free energies: a $2 \times 2 \times 0.3$ nm^3 cuboidal cavity was used to compute μ_A with INDUS simulations (illustrated at top) and a series of small spherical cavities with radii of 0.33 nm were used to compute μ_L from unbiased MD simulations (illustrated at bottom; only one cavity is shown for clarity). (b) Correlation between μ_A and experimental water contact angles ($\cos\theta$, top) and median values of μ_L (bottom) for ligands with different end group chemistries. Contact angle measurements were taken from Ref. 75,76 and tabulated in the Supporting Information, Table S2.[2] Both μ_L and μ_A were computed for top and bottom planar SAMs and treated as independent samples. Points show the average value of μ_L or μ_A and error bars report the standard deviation of μ_L or μ_A between the two samples. Best-fit lines are drawn with corresponding slopes and Pearson's r values labeled at bottom right. (c) Probability density function, $P(\mu_L)$, for R = CH_3 and R = OH ligands computed from unbiased MD simulations. μ_L for bulk water is shown as a black dashed line as a reference. Dotted lines show the median value of μ_L for each SAM. (d) Hydration free energy maps for the same planar SAMs in (b) with μ_L values ranging between 6 (red) to 16 kT (blue). Each point indicates the value of μ_L calculated using a spherical cavity centered on that point. 379

10.3 Hydrophobicity of 2 nm diameter SAM-protected GNPs with different end group chemistries. (a) Snapshots showing a representative 2 nm diameter GNP coated with R = CH_3 ligands. The SAM-water interface (purple) was computed using the same procedure as the planar SAMs (Figure 10.2a). The hydration free energy map shows μ_L values between 9 (red) to 11 kT (blue) as in Figure 10.2. (b) Simulation snapshots and corresponding hydration free energy maps for 2 nm GNPs with six end group chemistries, following the same color scheme as in (a). (c) Probability density functions, $P(\mu_L)$, showing the distribution of μ_L for GNPs with the six end group chemistries in (b). $P(\mu_L)$ is also shown for planar SAMs in the bottom panel for comparison. Dotted lines show the median value of each μ_L distribution and are color-coded to match the different ligands. The value of μ_L for bulk water is shown as a black dashed line for reference. 383

10.4 Effect of ligand length on the hydrophobicity of planar SAMs. (a) Snapshots of OH-terminated ligands on planar gold with $n = 3$ and $n = 11$ methylene groups. Hydration free energy maps are visualized from a top-down view with μ_L values ranging from 9 (red) to 16 kT (blue). (b) Probability density functions, $P(\mu_L)$, showing the distribution of μ_L for OH-terminated ligands on a 2 nm diameter gold core ($n = 11$, black) and planar gold ($n = 3$, blue and $n = 11$, red). Dotted lines show the median value of μ_L for each distribution and are color-coded to match the different ligand chain lengths and gold core types. The value of μ_L for bulk water is shown as a black dashed line for reference. 386

10.5 Effect of core size on gold nanoparticle hydrophobicity. (a) Example image of a 6 nm diameter gold core. Representative 6 nm GNPs and hydration free energy maps for R = CH_3 ligands and R = OH ligands. Hydration free energy maps show μ_L values between 9 (red) and 11 kT (blue). (b) Probability density functions, $P(\mu_L)$, for the 2 and 6 nm diameter gold cores and planar SAMs for R = CH_3 ligands (top) and R = OH ligands (bottom). Dotted lines denote the median μ_L value for each distribution and are color-coded to match the different gold core sizes. The value of μ_L for bulk water is shown as a black dashed line for reference. 389

10.6 Effect of unsaturated and branched ligands on gold nanoparticle hydrophobicity. (a) Representative 2 nm diameter GNP configurations coated with either saturated, unsaturated, or branched OH-terminated ligands. Hydration free energy maps show μ_L values between 9 (red) and 11 kT (blue) with the same color bar shown in Figure 10.3a. Hydrophilic clusters corresponding to regions where μ_L values are greater than 11.25 kT (the value of μ_L for bulk water) are indicated in unique colors. (b) Probability density function, $P(\mu_L)$, for 2 nm diameter GNPs with either saturated, unsaturated, or branched OH-terminated ligands. Dotted lines show the median value of μ_L for each distribution and are color-coded to match the different ligand structures. The value of μ_L for bulk water is shown as a black dashed line for reference. (c) Median μ_L values for saturated, unsaturated, or branched ligands with different R groups. (d) Number of hydrophilic clusters for saturated, unsaturated, or branched ligands with different R groups. The inset snapshot shows the hydrophilic clusters for unsaturated COOH-terminated ligands. The reported values and error bars in (c) and (d) are the average and standard deviation of values computed for the most and least representative GNPs (Supporting Information, Figure S2).[2] 393

10.7 Relationship between hydration free energy maps and competitive binding of hydrophobic and hydrophilic probe molecules. (a) Simulation snapshot of representative 2 nm diameter GNP with OH-terminated ligands in the presence of 1 mol% propane (red spheres) and 99 mol% water (blue spheres). Hydrogen atoms are omitted for clarity. (b) Occupancy maps showing the fraction of simulation time in which spherical cavities (the same cavities used to compute μ_L) are either occupied by propane (f_P) or water (f_{H_2O}) for the GNP in (a). Hydration free energy maps are shown for μ_L values between 9 (red) to 11 kT (blue) as a comparison to occupancy map. (c) Relationship between f_P or f_{H_2O} to μ_L computed using Gaussian kernel-density estimation to calculate the probability density function; values between 0 to 0.25 were omitted for clarity. 396

11.1 Experimental data used to develop QNAR models. (a) Example of a self-assembled monolayer structure consisting of a sulfur head group, nonpolar backbone, and end group, which are bound to the gold core with a strong gold-sulfur interaction. (b) Schematic representation of experimental observables for GNP behavior, such as logP, cell uptake in A549 cells, and zeta potential in water. (c) Number distribution of GNPs with experimental labels of logP, cell uptake in A549 cells, and zeta potential in water. The total number of GNPs for each experimental observable is shown on the upper right. All experimental data were taken from Ref. 37. 420

11.2 Computational workflow for modeling GNPs using MD simulations. The GNP database consists of diameters of spherical gold cores, number of ligands adsorbed on the surface, and ligand structure in the form of a SMILES string. The diameter and ligand substructure, either butanethiol (SMILES: SCCCC) or 1,2-dithiolane (SMILES: C1CCSS1), were then used to initiate self-assembly simulations. Substructures that are not adsorbed are shown in green. After the substructures are adsorbed onto the gold core surface, excess substructures are removed and adsorbed substructures are replaced with ligands from the GNP database and simulated in pure water (shown as cyan). Sodium or chlorine counterions are shown as purple and included to ensure charge-neutral systems. This schematic uses GNP1 and GNP288 as representative examples with distinct core sizes and ligand substructures. 423

11.3 MD-derived descriptors used to develop QNAR models. (a) Three examples of GNPs with varying core diameters (5.7 *versus* 7.3 nm) and ligand structures. Simulation snapshots are shown without water molecules. (b) Solvent-accessible-surface area (SASA) and number of ligand-water hydrogen bonds (HBonds) *versus* simulation time for the three GNPs in (a). (c) Concise list of of MD-derived descriptors used for developing QNAR models. The full 25 descriptor set is described in Table S1.[2] (d) Pearson's r between the 25 descriptor space, which shows red for highly correlated descriptors and blue for uncorrelated descriptors. Descriptor indexes correspond to descriptors in Table S1.[2] Highly correlated descriptors with | Pearson's r | > 0.9 were removed to output 15 uncorrelated descriptors. 426

11.4 Prediction accuracy of QNAR models. Parity plots of predicted *versus* experimental logP (left), cell uptake in A549 cells (middle), and zeta potential in water (right) values are shown for LASSO (top) and random forest (bottom) models. 5-fold cross validation (5-CV) predictions are shown as black dots and test set predictions are shown as red dots. Pearson's r between predicted and experiments for 5-CV and test sets are shown in the upper left with the same color scheme. Dashed lines indicates when the predicted values equate the experimental values. . . 429

11.5 Feature importance from QNAR models. (a) Top three most important descriptors are shown for logP (left), cell uptake in A549 cells (middle), and zeta potential in water (right) are shown for random forest models. Feature importance was quantified by Shapley values, where the average magnitude of Shapley values (*i.e.* Mean |Shap|) are reported and the sign is determined by the Pearson's r correlation between Shapley and descriptor values. The sign of the feature importance indicates whether the feature positively or negatively impact the experimental observables. Feature importance was computed by training the random forest model using 90% of the training data (randomly sampling without replacement) and iterated for 10 trials; the average of trials is reported and the error is reported as the standard deviation of the trials. (b) Simulation snapshots of GNPs with the highest and lowest top descriptor outputted from the feature importance in (a). Core diameters are shown in parenthesis; number and structure of ligands are shown to the right of the snapshots; and, descriptor and experimental measurements are shown below the snapshots. GNP169 has a mixed-monolayer, but only the structure of the majority ligand is shown for brevity. 435

11.6 Transferability of the cell uptake model to new datasets. (a) Thioalkyl tetra(ethyleneglycol)ated structures with four end group chemistries. (b) Cell uptake of 12 GNPs with core diameters of 2, 4, and 6 nm, and ligand structures from (a). Cell uptake measurements were performed after GNPs were incubated with HeLa cells for 3 hours and quantified using inductively coupled plasma mass spectrometry; all experimental data were taken from Ref. 19. The cell uptake values were converted to have the same cell uptake units as Ref. 37. (c) Predicted *versus* experimental cell uptake values using MD-derived descriptors of the 12 GNPs and the data from (b) as the experimental data. Predicted values were computed using a RF model trained with 15 uncorrelated descriptors as inputs and 65 cell uptake labels from Figure 11.1c as output. Pearon's r between predicted and experimental values using all 12 GNPs are shown in the lower right. Black dashed lines show the best fit line as a guide. The simulation snapshot show a 6 nm GNP with TTMA ligands. . 439

12.1 Future work on incorporating product states into a deep learning framework to predicting selectivities. MD simulations are performed on both reactant and product states. Then, simulations are transformed into voxel representations, as described previously in Chapter 5. The voxel representations are then inputted into 3D convolutional neural networks and merged into fully-connected layers with the regression task of predicting $\Delta\Delta G$. 455
12.2 Future work on integrating US simulations and QCM measurements to measure protein binding affinities on planar SAMs and GNPs. (a) Reaction coordinate of US simulations between GNP and protein (marked as "P"), where the protein is sampled at a distance away from either a curved GNP surface (2-10 nm in core diameter) or planar SAM (estimated for >10 nm core diameters). (b) Expected US simulation and QCM results for a non-binding, weak, and strong protein. 460
12.3 Future work on leveraging cosolvent simulations to predict protein binding sites. (a) Five cosolvents that could be used for cosolvent simulations, which encompass polar, nonpolar, and charged species: water, propane, acetic acid, methylammonium, and formate. Each species (except water) has an amino acid analogue so that these simulations could suggest binding propensities of amino acids located on the protein surface. (b) Simulation workflow of taking the most likely configuration of a GNP (described in Chapter 10), then performing a high temperature simulation at T = 600 K, followed by an *NPT* production simulation at T = 300 K. Cosolvent molecules are color-coded as the same color shown in (a). (c) Occupancy maps of the 5 solvents from (a) for a 2 nm GNP with CH_3- and COOH-terminated ligand end groups. White colors mean no occupancy of the cosolvent, whereas colored surfaces mean high occupancy of the cosolvent. 463
12.4 Future work on integrating deep learning models to predict GNP properties from surface maps. 465

ABSTRACT

Identifying a good set of chemical structures for designing new products or materials is experimentally challenging because of the large range of structures that are possible and the high cost associated with trial-and-error exploration. One solution is to leverage computational tools that could help gain physical insight into how these structures relate to properties and screen for improved properties through structure-property relationships. This dissertation focuses on combining classical molecular dynamics (MD) simulations and machine learning techniques to systematically tune properties for two relevant application areas. In both applications, we develop predictive models that correlate simulation-derived descriptors to experimental data, enabling the screening of properties using computationally efficient methods.

The first application focuses on converting biomass (renewable, organic material from plants) into transportation fuels or commodity chemicals. Biomass conversion is performed through acid-catalyzed reactions in aqueous solution that are hindered by poor reactivity. One way to improve reactivity is to modify the solvent composition by mixing water with organic cosolvents. Thus, we developed MD simulations to understand and predict the effects of adding an organic solvent on biomass conversion reactivity. We used MD and machine learning models to predict the influence of solvents on reaction rates and selectivities in biomass conversion

reactions, and we combined these tools into a workflow for screening solvents in biomass-related reactions.

The second application is motivated by the ability of ligand-coated gold nanoparticles (GNP) to target biomolecules, such as receptors on cancer cells. Due to the vast number of tuning parameters of the GNP (*e.g.* core size, shape, and ligand selection), it is unfeasible to experimentally screen GNPs. Therefore, we developed a generalized workflow for building GNP systems using MD for any arbitrary gold core shape and size, and ligand selection. We used this workflow to study the effect of core and ligand selection on GNP surface properties and GNP-bilayer interactions. We then developed a library of molecular descriptors that characterize GNP properties and leveraged machine learning tools to develop accurate structure-property relationships, which is useful for screening new GNPs for selective behavior.

1 INTRODUCTION

1.1 Motivation

Designing new catalytic reactions, materials, or drugs is often hindered by our lack of understanding in how the underlying chemical structures and their cooperative interactions with other molecules propagate to macroscopic behavior (*e.g.* reaction performance, material durability, or drug specificity). As a result, designing new chemicals is often performed by an empirical trial-and-error approach, which is expensive and does not lend physical understanding into how these chemical structures relate to large-scale properties. These challenges motivate the use of computational approaches to elucidate molecular-level interactions, yield mechanistic insights, and predict macroscale properties prior to synthesizing chemicals in the lab. In the past decade, computer-aided molecular design have been used in a wide range of applications and have shown great success in chemical structure screening through the collaboration between synthetic, process, and computational chemists.[1–3]

In particular, classical molecular dynamics (MD) simulations allow us to explicitly study structures and dynamics of molecules over small length scales (~nm) and short timescales (~μs), which are challenging to assess experimentally or through *ab initio* quantum mechanical calculations. MD simulations are frequently used to study complex biomolecules (*e.g.* proteins or lipid bilayers), where structural and chemical properties are

heavily dictated by their dynamical structure and their interactions with the surrounding environment.[4,5] Significant efforts have been spent on developing accurate force-fields, a set of complex equations and parameters dictating interactions between particles that can capture accurate interactions in complex systems.[6,7] As a result, MD provides a powerful toolset to study a wide range of molecules in solution, where the molecular-level behavior is heavily influenced by cooperative interactions between the chemical of interest and its surrounding environment.

In parallel with the development of MD, improved data storage capabilities have enabled the development of large databases and motivated the development of machine learning algorithms, which are used to develop sophisticated models and perform data mining procedures.[8] The ImageNet Large-Scale Visual Recognition Challenge, which is an annual contest to develop the best model for image detection across 1.2 million images, is a good example of how machine learning could be leveraged to tackle challenging problems, such as object detection or facial recognition.[9] Given that MD simulations provide accurate physics-based representation of chemicals and that machine learning provides models for predicting observables, this dissertation seeks to merge these tools to screen chemical properties for two relevant application areas: (1) acid-catalyzed reactions for biomass conversion and (2) monolayer-coated gold nanoparticles (GNP) for biomedical applications. Although these applications are in different fields, where the former is in catalysis and the latter is in nanomaterials,

they are connected in that their macroscopic behavior depend heavily on interactions between a minor component (*i.e.* solute, such as a reactant or GNP) and a major component (*i.e.* solvent). In addition, both applications have large design spaces, making it difficult to explore experimentally. Therefore, we will use MD simulations to gain physical understanding into how structural parameters affect properties, which we will use to develop structure-property relationships models for screening capabilities. We next motivate each application area and provide a literature review in Sections 1.2 and 1.3, followed by an overview of this disseration in Section 1.4.

1.2 Solvent screening for biomass conversion reactions

1.2.1 Motivation

Biomass is an organic material derived from living organisms that can be used as a source of renewable energy and commodity chemicals. One of the most abundant forms of biomass is lignocellulosic biomass, which can be processed to make fuel or high-value commodity chemicals shown in Figure 1.1.[10,11] However, biomass conversion to more readily usable chemicals still faces economic challenges, such as expensive catalysts, low product selectivity, and cost-inefficient processes. One possible solution is

to modify the solvent composition, which affects reaction rates, product selectivity and stability, and economics of downstream separations. For example, the addition of an organic solvent, such as γ-valerolactone (GVL), has been found to increase the activity of acidic protons and can break down >80% lignin compared to <20% lignin in pure water.[10,12] However, finding an optimal solvent composition empirically by trial-and-error is cost-prohibitive and gives little insight to how the solvent environment will perform in new processes. Therefore, we focused on using MD simulations to study the solvent environment that can give molecular insight into how addition of organic solvents can improve reactivity, which has broader impacts in understanding how to engineer the solvent environment for biomass conversion processes.

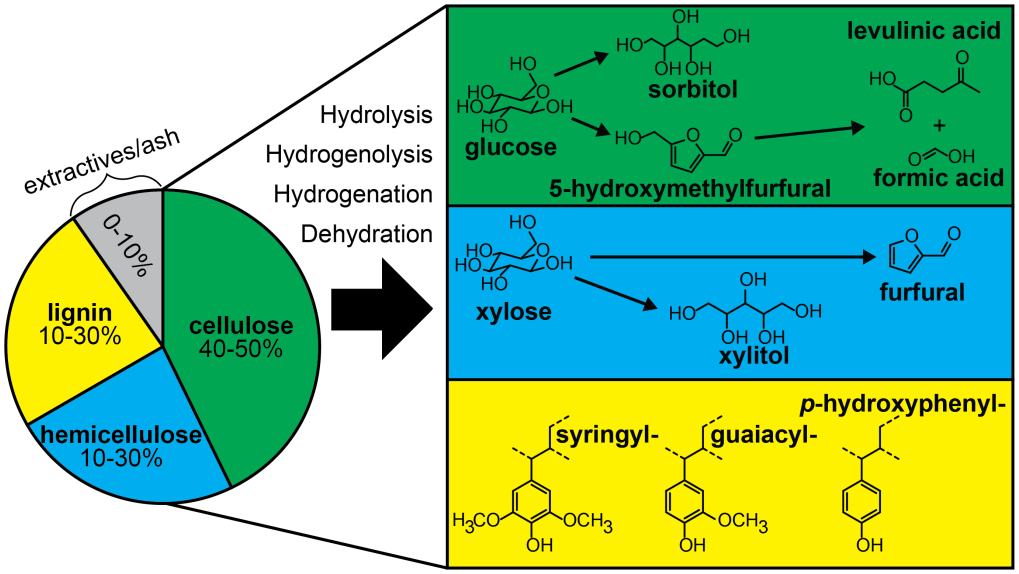


Figure 1.1: Major platform chemicals produced from lignocellulosic biomass. This image was adapted from Ref. 10 with permission from John Wiley and Sons.

1.2.2 Background

The catalytic upgrading of biomass (*e.g.*, wood, crops, *etc.*) is a promising strategy to obtain valuable chemicals from renewable resources while limiting waste products.[10,13–17] For example, cellulose, one of the primary components of lignocellulosic biomass, can be converted through a series of dehydration and hydrolysis reactions to form 5-hydroxylmethylfurfural, a platform chemical for fuels and other commodity chemicals,[18–23] and levoglucosenone (LGO), as depicted in Figure 1.2. These reactions are typically performed in aqueous solution in the presence of an acid-catalyst

(*i.e.* acid-catalyzed reactions), where extensive control over reaction kinetics and selectivity is available by tuning the temperature, catalyst, and solvent composition.[14,18–21,24–27] Recent studies have shown that inclusion of organic solvents can significantly improve the selective conversion of biomass into sugars.[24,28,29] We focus on polar aprotic solvents (*e.g.* dimethyl sulfoxide (DMSO), ketones, tetrahydrofuran (THF), GVL, *etc.*), one type of organic solvent, which are particularly interesting because: (1) they have lone electron pair(s) that can accept hydrogen bonds with water or reactant that can improve miscibility; and (2) they do not have acidic protons, which means that they do not directly contribute to the reaction mechanism.[10]

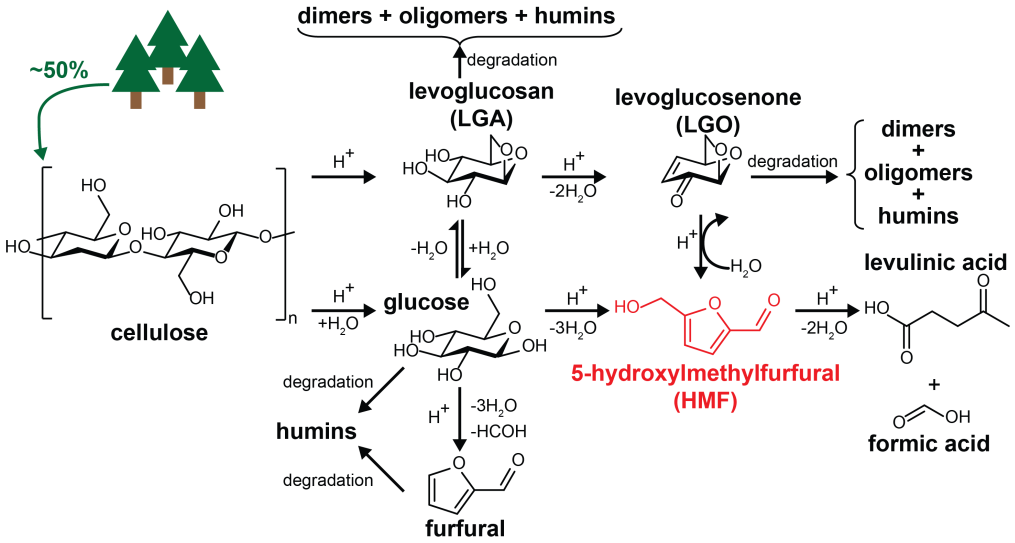


Figure 1.2: Reaction pathway for the decomposition of cellulose in presence/absence of low water concentrations. Cellulose is comprised of ~50% lignocellulosic biomass. 5-hydroxymethylfurfural is highlighted in red, since it is a platform chemical for transportation fuels and a desirable product from biomass conversion processes. This image was adapted from Ref. 18 with permission from The Royal Society of Chemistry.

Solvent effects refer to the changes induced by the solvent environment that can affect chemical reactivity including reaction rate, reaction pathways, product yield and distribution. In addition, solvent effects can have a greater influence on the reaction rate and product selectivities than changing the reaction temperature and/or the catalyst. For example, LGO is the major product of cellulose dehydration in pure THF, whereas HMF is the major product upon increase of water content up to 5 wt%.[18,30] Understanding the effects of changing solvent compositions on acid-catalyzed

reactions is difficult because (1) solute-solvent interactions are difficult to characterize experimentally, and (2) the changes in these interactions and how they translate to changes in reaction barriers are not well-understood. Thus, computational approaches can help us understand molecular-scale effects and explore new solvent compositions, leading to rational design of the solvent environment for biomass conversion.

1.2.3 Computational studies on solvent effects in acid-catalyzed reactions

Computational efforts in the past decade focused on using *ab initio* quantum chemical methods to quantify the effects of solvent on barriers to elementary reaction steps, leading to many important case studies.[31,32] However, *ab initio* techniques are computationally expensive and as a result, their calculations have smaller length scales (Å) and time scales (fs) than MD simulations (nm/ps). Therefore, it is difficult for *ab initio* techniques to capture solvent re-organization that constitute solute-solvent interactions, which can be captured in MD simulations.

Previous work used MD simulations to understand solvent effects in acid-catalyzed reactions.[33] Mushrif *et al.* studied selective conversion of fructose to HMF in aqueous mixtures of DMSO using MD simulations.[34,35] They used volumetric spatial maps, which maps solvent densities around a solute in three-dimensions, and found that there was a competition

between DMSO and water for hydroxyl groups on the solute. Therefore, addition of DMSO increases the rate of fructose conversion to HMF and decreases the rate of undesirable parallel reactions. While volumetric spatial maps are useful to find competing solvent effects around a reactant, it requires a case-by-case study that is difficult for large numbers of solvent compositions, which motivates the necessity of molecular descriptors that can quickly characterize the solvent environment.

1.2.4 Hypothesis and research questions

We hypothesize that MD-derived molecular descriptors encoding information about the solvation environment around the reactant, catalyst, or product can predict solvent-mediated reaction performance in biomass conversion reactions, without having to model the reaction mechanism directly. For this application, we seek to answer the following research questions:

1. How do we use classical MD simulations to gain physical insight into the effects of solvent selection on biomass reactivity? (Chapters 3 and 4)

2. How do we leverage machine learning models to improve accuracy of predicted rates from MD? (Chapter 5)

3. How do we use classical MD to estimate selectivities of biomass conversion reactions? (Chapter 6)

4. How do we combine the computational tools to guide experimental design with minimal experiments? (Chapter 7)

1.3 Screening monolayer-protected gold nanoparticle properties for biomedical applications

1.3.1 Motivation

Gold nanoparticles (GNPs) are promising nanomaterials for cell therapy applications because they could be used to deliver genes or drugs, heated up to kill nearby cells, and surface-modified by ligands for target-specificity toward abnormal cells (see Figure 1.3(a)).[36] Two primary challenges in the applicability of GNPs are: (1) GNPs must avoid elimination by the host immune system, and (2) GNPs must specifically bind to cancer cells, not healthy cells. Challenge (1) can be resolved by using "stealth technologies," which involve coating the GNP with a self-assembled monolayers (SAM), which consists of a thiol head group, nonpolar backbone, and terminal end group shown in Figure 1.3(b). SAM-coated GNPs consisting of thiolated poly(ethylene glycol) (PEG) have been shown to prolong GNP lifetime in the body because the coating is hydrophilic, making the GNP look like the aqueous bulk phase to the immune system.[37] Challenge (2) can be resolved by attaching an antibody or ligands onto the SAM that have a

high affinity to receptors on cancer cells, some of which are overexpressed when compared to normal, healthy cells (*e.g.* epidermal growth factor receptor, folate receptor, *etc.*).[38,39] However, if the GNP is densely coated with PEG, it faces challenge (2) where it loses the ability to bind to receptors on cancer cells.[40] Therefore, there is a competition between being invisible to the immune system by attaching a PEG layer and having a high affinity to cancer cells by attaching ligands that target receptors. Tuning GNP parameters for selective behavior remains a challenge in the field; examples of the ligand backbone, ligand end group, and gold core size is depicted in Figure 1.3(c). We will use MD simulations to understand how the selection of GNP parameters affect its behavior with other molecules, which has broad impacts in the engineering of high affinity GNPs for cancer cell detection.

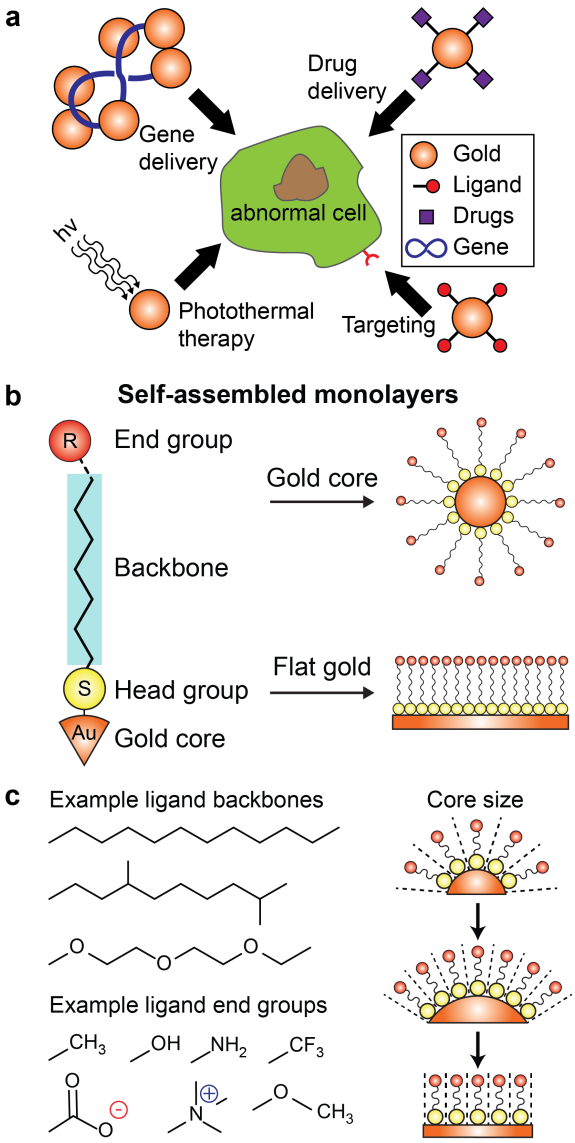


Figure 1.3: (a) Applications of GNPs in cancer therapy. This image was adapted from Ref. 36 with permission from Elsevier. (b) Structure of self-assembled monolayers (SAM) ligands, consisting of a sulfur head group, nonpolar backbone, and terminal end group. These ligands spontaneously form SAMs in the presence of a gold core or a flat gold. (c) Examples of the tunable parameters of GNPs, such as ligand backbone, ligand end groups, and gold core size.

1.3.2 Background

GNPs for drug delivery are ideal because they: (1) have low cytotoxicity; (2) are synthesizable with well-defined sizes; (3) can be readily detected; and more importantly, (4) can be surface-modified by reaction with thiol-containing molecules (*e.g.* -SH). Surface modifications allow changes to physiochemical properties, such as surface charge, morphology, and so on.[41] GNPs can be used for either active targeting or passive targeting, shown in Figure 1.4. Active targeting involves attachment of ligands that recognizes a specific cell receptor displayed on a delivery vehicle, whereas passive targeting involves accumulation of drug carriers in target tissues due to leaky gaps in the endothelium of approximately 600 nm, which are larger compared to healthy tissues. Therefore, for passive targeting, the GNP must be sufficiently small, typically 5-10 nm in diameter, to take advantage of the so-called "enhanced permeation and retention" effect (EPR) where biocompatible macromolecules accumulate at higher concentration near tumor tissues.[42] An ideal GNP would be able to take advantage of both passive and active transport, where the conjugated ligands can specifically bind to cancer cells and the GNP can overcome transport barriers.

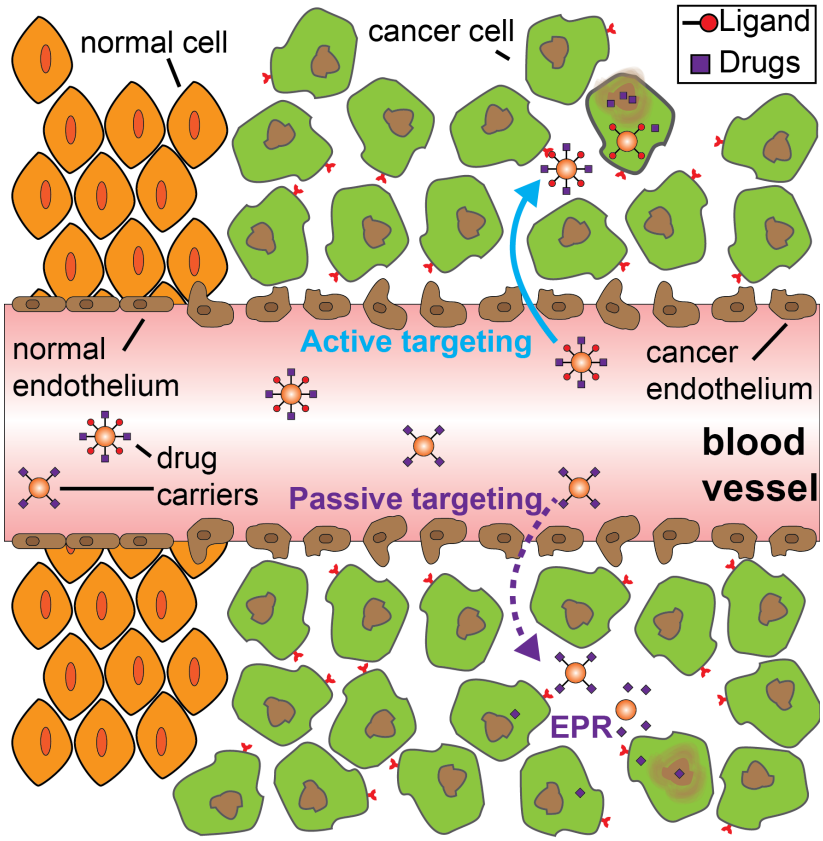


Figure 1.4: Active (solid) and passive (dashed) targeting using GNPs. This image was adapted from Ref. 36 with permission from Elsevier.

1.3.3 Design considerations to improve GNP blood circulation lifetimes

To improve the GNPs blood circulation lifetime, PEG chains can be grafted onto GNPs.[37,41,43] The PEG chains act like a polymer brush that camouflages the GNP from the immune system, preventing protein adsorption

while hydrogen bonding to water. Furthermore, GNP diameter can influence its blood circulation lifetime. Particles smaller than 5 nm in diameter are filtered by the kidneys, a process called renal clearance,[44,45] and hydrophobic GNPs of this size are found to passively penetrate cell membranes and can be trapped within the lipid bilayer.[46] Larger hydrophobic GNPs (5-10 nm) are found to spontaneously translocate across lipid bilayers, and GNPs >10 nm use endocytosis, an active transport mechanism requiring collective action of membrane proteins, to translocate into the cell which is generally slow with characteristic times of seconds to minutes.[47] GNPs > 100 nm are filtered by the spleen and sequestered by the liver.[44,45] Therefore, from a design standpoint, GNPs should include a PEG chains for biocompatibility and should have diameters in the range between 5-100 nm to avoid filtration processes. This dissertation primarily focus on GNPs within the sub-10 nm core size because they are sufficiently small enough to be modeled atomistically, and they have the same length scales to biomolecules.[48]

1.3.4 GNPs in the biological environment: challenges of the corona

Once placed into a biological environment, GNPs face the challenge of proteins adsorbing onto the surface, which can still occur even with PEG chains grafted on the GNP.[49] Adsorbed proteins on the GNP surface form

a protein corona (or simply, corona), typically consisting of soft and hard coronas.[49–51] Soft coronas are loosely bound proteins that can be more readily removed. Conversely, hard coronas are tightly adsorbed onto the surface of the GNP that is more stable and difficult to remove. The formation of coronas on the surface of the GNP can significantly modify surface properties and increase its effective diameter. Therefore, even if we engineered a GNP to target cancer cells, its function may be completely changed due to the formation of a corona, which gives the GNP a biological identity.[51] In addition, proteins in the corona could be displaced by new molecules as the GNP moves to a new environment. Thus, the corona is dependent on not only the current environment, but also all the previous environments that the GNP has passed through.[52] With nearly 3,700 types of proteins in human blood,[53,54] only a few proteins are abundant in the hard corona and they are often difficult to characterize experimentally.[51] Overall, there are many possible compositions of corona formation on GNPs and currently, a lack of tools that can predict, or even understand, the GNP biological identity. This challenge may be addressed with the computational tools developed in this dissertation, which can enable high-throughput screening of protein adsorption on the GNP surface as a way of understanding corona formation.

1.3.5 Hypothesis and research questions

We hypothesize that modeling GNPs with MD simulations can help us understand how gold core and ligand selection affects monolayer properties, and these physical insights could be encoded within molecular descriptors that could be used to predict GNP behavior more broadly. For this application, we seek to answer the following research questions:

1. How do we develop a generalized workflow to account for tunable GNP parameters? (Chapter 8)

2. How do we use classical MD simulations to gain physical insight into the effects of GNP core/ligand selection on surface properties? (Chapters 9 and 10)

3. How do we use MD and machine learning tools to accurately predict GNP behavior and guide experimental design? (Chapter 11)

1.4 Scope of this disseration

The structure of this disseration is outlined as follows:

Chapter 2. This chapter introduces the computational methods, such as classical molecular dynamic simulations, quantitative structure-property relationships, and machine learning models. We will use these tools to generate simulation systems and relate simulation-derived descriptors to experimental observables.

Solvent screening for biomass conversion reactions. The computational tools developed for this application are summarized in Figure 1.5 and described below:

Chapter 3. This chapter introduces the simulation model for screening solvent environments in biomass conversion reactions. We will show that simulation-derived descriptors could predict experimental reaction rates, despite not modeling the reaction mechanism or the catalyst.

Chapter 4. One question from Chapter 3 was whether catalyst information might be informative of how solvents effect biomass conversion reactions. This chapter attempts to incorporate catalyst information by modeling the acid-catalyst (*i.e.* hydronium ion) in various solvent environments.

Chapter 5. This chapter uses the simulation and experimental data from Chapter 3 as inputs and outputs, respectively, for a deep learning framework. We integrated molecular simulations of biomass-relevant compounds and convolutional neural networks to rapidly predict experimental reaction rates without having to pre-define descriptors *a priori* as done in Chapter 3.

Chapter 6. This chapter introduces the use of solvation free energy calculations to estimate selectivities of biomass conversion reactions. These tools are useful for predicting solvent effects on selectivities of parallel reactions, which are generally challenging to anticipate *a priori*.

Chapter 7. This chapter integrates the tools from Chapters 5 and 6 into

a workflow to downselect solvent environments using a combination of molecular simulations and machine learning tools. We then test two case study examples, where these computational tools decrease the number of experiments necessary to select a good mixed-solvent environment for biomass conversion reactions.

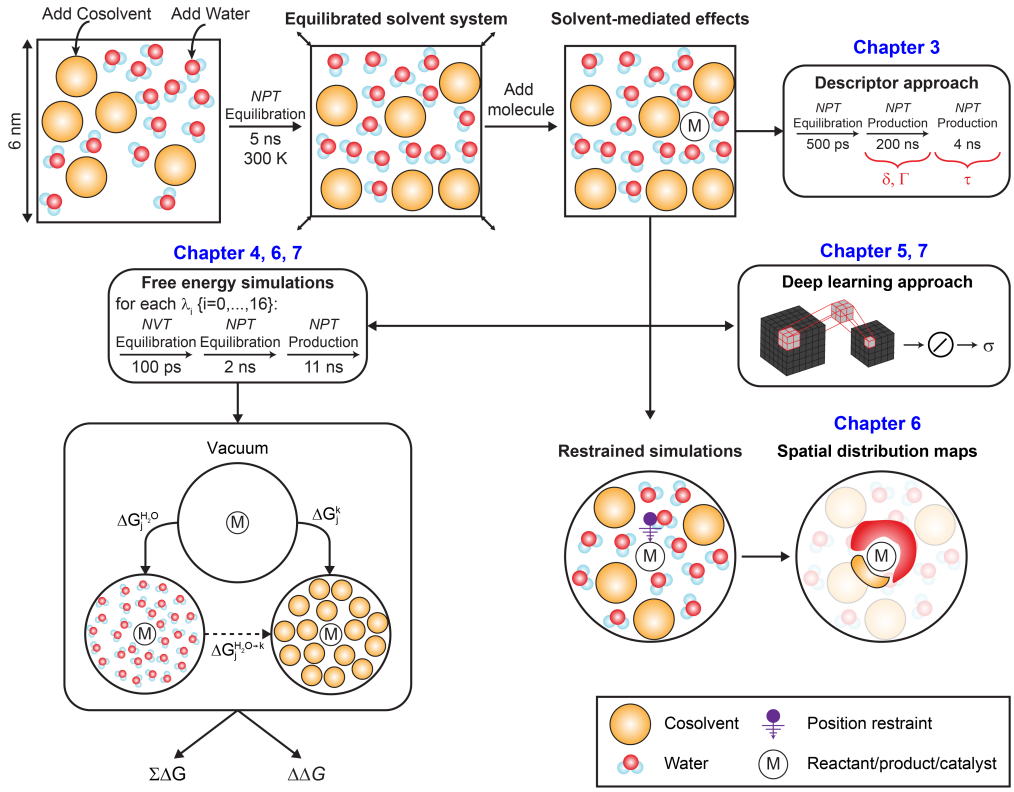


Figure 1.5: Computational overview for applications in solvent screening for biomass conversion reactions.

Screening monolayer-protected gold nanoparticles for biomedical applications. The computational tools developed for this application are

summarized in Figure 1.6 and described below:

Chapter 8. This chapter introduces the generalized workflow for generating GNP systems that account for gold core shape and size, and ligand selection. We used clustering algorithms to quantify the extent of "bundles" or the alignment of ligands observed in small GNPs with long, nonpolar ligands. This chapter highlights the importance of using MD simulations to model GNP systems, which could capture the cooperativity arising from ligand-ligand and ligand-solvent interactions.

Chapter 9. This chapter extends the generalized workflow to studying GNP-lipid-bilayer systems. We investigated the effects of ligand end group chemistry on the adsorption of GNPs onto DOPC lipid bilayer systems. One finding from this chapter is that ligand hydrophobicity (*i.e.* lipophilicity) - the thermodynamic affinity of a molecule to water - correlates with GNP uptake on bilayers, suggesting that hydrophobicity is an important parameter for GNP design.

Chapter 10. Given that hydrophobicity is an important property from Chapter 9, this chapter describes computational methods to quantify the hydrophobicity on the surface of GNPs by analyzing the fluctuations of water at the GNP-water interface.

Chapter 11. This chapter distills the physical insights from Chapters 8-10 in the form of molecular descriptors and describes structure-property relationships used to predict experimental observables (*i.e.* logP, cell uptake, and zeta potential). These models are useful for designing and

screening new GNPs with physically motivated descriptors.

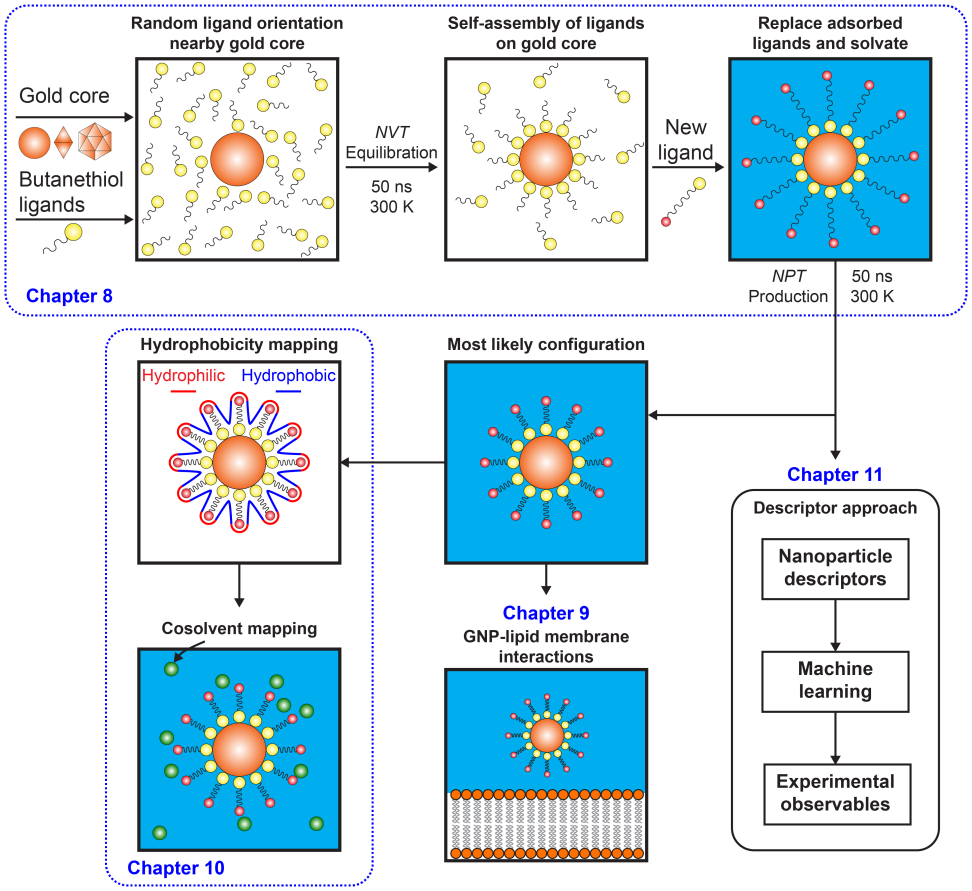


Figure 1.6: Computational overview for applications in screening monolayer-protected gold nanoparticle properties.

Chapter 12. This chapter provides a discussion on the key findings from this dissertation and potential avenues of future research.

1.5 References

[1] Yu, W.; MacKerell, A. D. In *Antibiotics*; Springer, 2017; pp 85–106.

[2] Jhamb, S.; Enekvist, M.; Liang, X.; Zhang, X.; Dam-Johansen, K.; Kontogeorgis, G. M. A review of computer-aided design of paints and coatings. *Current Opinion in Chemical Engineering* **2020**, *27*, 107–120.

[3] Austin, N. D.; Sahinidis, N. V.; Trahan, D. W. Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chemical Engineering Research and Design* **2016**, *116*, 2–26.

[4] Durrant, J. D.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC biology* **2011**, *9*, 1–9.

[5] Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature structural biology* **2002**, *9*, 646–652.

[6] Lyubartsev, A. P.; Rabinovich, A. L. Force field development for lipid membrane simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2016**, *1858*, 2483–2497.

[7] Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC* **2015**, *8*, 37.

[8] Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **2021**, *8*, 1–74.

[9] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al.. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.

[10] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[11] Zhao, X.; Zhang, L.; Liu, D. Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. *Biofuels, Bioproducts and Biorefining* **2012**, *6*, 465–482.

[12] Luterbacher, J. S.; Rand, J. M.; Alonso, D. M.; Han, J.; Youngquist, J. T.; Maravelias, C. T.; Pfleger, B. F.; Dumesic, J. A. Nonenzymatic sugar production from biomass using biomass-derived γ-valerolactone. *Science* **2014**, *343*, 277–280.

[13] Walker, T. W.; Motagamwala, A. H.; Dumesic, J. A.; Huber, G. W. Fundamental catalytic challenges to design improved biomass conversion technologies. *Journal of Catalysis* **2018**, *369*.

[14] Huber, G. W.; Iborra, S.; Corma, A. Synthesis of transportation fuels from biomass: chemistry, catalysts, and engineering. *Chemical reviews* **2006**, *106*, 4044–4098.

[15] Stöcker, M. Biofuels and biomass-to-liquid fuels in the biorefinery: Catalytic conversion of lignocellulosic biomass using porous materials. *Angewandte Chemie International Edition* **2008**, *47*, 9200–9211.

[16] Tock, L.; Gassner, M.; Maréchal, F. Thermochemical production of liquid fuels from biomass: Thermo-economic modeling, process design and process integration analysis. *Biomass and bioenergy* **2010**, *34*, 1838–1854.

[17] Nguyen, T. Y.; Cai, C. M.; Kumar, R.; Wyman, C. E. Overcoming factors limiting high-solids fermentation of lignocellulosic biomass to ethanol. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11673–11678.

[18] He, J.; Liu, M.; Huang, K.; Walker, T. W.; Maravelias, C. T.; Dumesic, J. A.; Huber, G. W. Production of levoglucosenone and 5-hydroxymethylfurfural from cellulose in polar aprotic solvent–water mixtures. *Green Chemistry* **2017**, *19*, 3642–3653.

[19] Chheda, J. N.; Huber, G. W.; Dumesic, J. A. Liquid-phase catalytic processing of biomass-derived oxygenated hydrocarbons to fuels and chemicals. *Angewandte Chemie International Edition* **2007**, *46*, 7164–7183.

[20] Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chemical reviews* **2007**, *107*, 2411–2502.

[21] Román-Leshkov, Y.; Barrett, C. J.; Liu, Z. Y.; Dumesic, J. A. Production of dimethylfuran for liquid fuels from biomass-derived carbohydrates. *Nature* **2007**, *447*, 982–985.

[22] Mellmer, M. A.; Gallo, J. M. R.; Martin Alonso, D.; Dumesic, J. A. Selective production of levulinic acid from furfuryl alcohol in THF solvent systems over H-ZSM-5. *ACS Catalysis* **2015**, *5*, 3354–3359.

[23] Pagan-Torres, Y. J.; Wang, T.; Gallo, J. M. R.; Shanks, B. H.; Dumesic, J. A. Production of 5-hydroxymethylfurfural from glucose using a combination of Lewis and Brønsted acid catalysts in water in a biphasic reactor with an alkylphenol solvent. *Acs Catalysis* **2012**, *2*, 930–934.

[24] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[25] Motagamwala, A. H.; Won, W.; Maravelias, C. T.; Dumesic, J. A. An engineered solvent system for sugar production from lignocellulosic biomass using biomass derived γ-valerolactone. *Green Chemistry* **2016**, *18*, 5756–5763.

[26] Won, W.; Motagamwala, A. H.; Dumesic, J. A.; Maravelias, C. T. A co-solvent hydrolysis strategy for the production of biofuels: process synthesis and technoeconomic analysis. *Reaction Chemistry & Engineering* **2017**, *2*, 397–405.

[27] Sener, C.; Motagamwala, A. H.; Alonso, D. M.; Dumesic, J. A. Enhanced furfural yields from xylose dehydration in the γ-Valerolactone/water solvent system at elevated temperatures. *ChemSusChem* **2018**, *11*, 2321–2331.

[28] Mellmer, M. A.; Alonso, D. M.; Luterbacher, J. S.; Gallo, J. M. R.; Dumesic, J. A. Effects of γ-valerolactone in hydrolysis of lignocellulosic biomass to monosaccharides. *Green Chemistry* **2014**, *16*, 4659–4662.

[29] Wang, Y.; Wang, H.; Lin, H.; Zheng, Y.; Zhao, J.; Pelletier, A.; Li, K. Effects of solvents and catalysts in liquefaction of pinewood sawdust for the production of bio-oils. *Biomass and bioenergy* **2013**, *59*, 158–167.

[30] Cao, F.; Schwartz, T. J.; McClelland, D. J.; Krishna, S. H.; Dumesic, J. A.; Huber, G. W. Dehydration of cellulose to levoglucosenone using polar aprotic solvents. *Energy & Environmental Science* **2015**, *8*, 1808–1815.

[31] Assary, R. S.; Kim, T.; Low, J. J.; Greeley, J.; Curtiss, L. A. Glucose and fructose to platform chemicals: understanding the thermodynamic landscapes of acid-catalysed reactions using high-level ab initio methods. *Physical Chemistry Chemical Physics* **2012**, *14*, 16603–16611.

[32] Zhang, J.; Das, A.; Assary, R. S.; Curtiss, L. A.; Weitz, E. A combined experimental and computational study of the mechanism of fructose dehydration to 5-hydroxymethylfurfural in dimethylsulfoxide using Amberlyst 70, PO43-/niobic acid, or sulfuric acid catalysts. *Applied Catalysis B: Environmental* **2016**, *181*, 874–887.

[33] Varghese, J. J.; Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering* **2019**, *4*, 165–206.

[34] Mushrif, S. H.; Caratzoulas, S.; Vlachos, D. G. Understanding solvent effects in the selective conversion of fructose to 5-hydroxymethylfurfural: a molecular dynamics investigation. *Physical Chemistry Chemical Physics* **2012**, *14*, 2637–2644.

[35] Vasudevan, V.; Mushrif, S. H. Insights into the solvation of glucose in water, dimethyl sulfoxide (DMSO), tetrahydrofuran (THF) and N, N-dimethylformamide (DMF) and its possible implications on the conversion of glucose to platform chemicals. *Rsc Advances* **2015**, *5*, 20756–20763.

[36] Ghosh, P.; Han, G.; De, M.; Kim, C. K.; Rotello, V. M. Gold nanoparticles in delivery applications. *Advanced drug delivery reviews* **2008**, *60*, 1307–1315.

[37] Alexis, F.; Pridgen, E.; Molnar, L. K.; Farokhzad, O. C. Factors affecting the clearance and biodistribution of polymeric nanoparticles. *Molecular pharmaceutics* **2008**, *5*, 505–515.

[38] Peer, D.; Karp, J. M.; Hong, S.; Farokhzad, O. C.; Margalit, R.; Langer, R. Nanocarriers as an emerging platform for cancer therapy. *Nature nanotechnology* **2007**, *2*, 751–760.

[39] Steichen, S. D.; Caldorera-Moore, M.; Peppas, N. A. A review of current nanoparticle and targeting moieties for the delivery of cancer therapeutics. *European journal of pharmaceutical sciences* **2013**, *48*, 416–427.

[40] Pissuwan, D.; Valenzuela, S. M.; Cortie, M. B. Therapeutic possibilities of plasmonically heated gold nanoparticles. *TRENDS in Biotechnology* **2006**, *24*, 62–67.

[41] Bergen, J. M.; Von Recum, H. A.; Goodman, T. T.; Massey, A. P.; Pun, S. H. Gold nanoparticles as a versatile platform for optimizing physicochemical parameters for targeted drug delivery. *Macromolecular Bioscience* **2006**, *6*, 506–516.

[42] Maeda, H. The enhanced permeability and retention (EPR) effect in tumor vasculature: the key role of tumor-selective macromolecular drug targeting. *Advances in enzyme regulation* **2001**, *41*, 189–207.

[43] Hamad, I.; Al-Hanbali, O.; Hunter, A. C.; Rutt, K. J.; Andresen, T. L.; Moghimi, S. M. Distinct polymer architecture mediates switching of complement activation pathways at the nanosphere- serum interface: implications for stealth nanoparticle engineering. *ACS nano* **2010**, *4*, 6629–6638.

[44] Gatmaitan, Z.; Varticovski, L.; Ling, L.; Mikkelsen, R.; Steffan, A.-M.; Arias, I. M. Studies on fenestral contraction in rat liver endothelial cells in culture. *The American journal of pathology* **1996**, *148*, 2027.

[45] Longmire, M.; Choyke, P. L.; Kobayashi, H. Clearance properties of nano-sized particles and molecules as imaging agents: considerations and caveats. **2008**.

[46] Guo, Y.; Terazzi, E.; Seemann, R.; Fleury, J. B.; Baulin, V. A. Direct proof of spontaneous translocation of lipid-covered hydrophobic nanoparticles through a phospholipid bilayer. *Science advances* **2016**, *2*, e1600261.

[47] Liu, J.; Sun, Y.; Drubin, D. G.; Oster, G. F. The mechanochemistry of endocytosis. *PLoS Biol* **2009**, *7*, e1000204.

[48] You, C.-C.; De, M.; Rotello, V. M. Monolayer-protected nanoparticle–protein interactions. *Current opinion in chemical biology* **2005**, *9*, 639–646.

[49] 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*, 6748–6755.

[50] Melby, E. S.; Lohse, S. E.; Park, J. E.; Vartanian, A. M.; Putans, R. A.; Abbott, H. B.; Hamers, R. J.; Murphy, C. J.; Pedersen, J. A. Cascading effects of nanoparticle coatings: Surface functionalization dictates the assemblage of complexed proteins and subsequent interaction with model cell membranes. *ACS nano* **2017**, *11*, 5489–5499.

[51] Monopoli, M. P.; Åberg, C.; Salvati, A.; Dawson, K. A. Biomolecular coronas provide the biological identity of nanosized materials. *Nature nanotechnology* **2012**, *7*, 779–786.

[52] Lundqvist, M.; Stigler, J.; Cedervall, T.; Berggard, T.; Flanagan, M. B.; Lynch, I.; Elia, G.; Dawson, K. The evolution of the protein corona around nanoparticles: a test study. *ACS nano* **2011**, *5*, 7503–7509.

[53] Anderson, N. L.; Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Molecular & cellular proteomics* **2002**, *1*, 845–867.

[54] Pieper, R.; Gatlin, C. L.; Makusky, A. J.; Russo, P. S.; Schatz, C. R.; Miller, S. S.; Su, Q.; McGrath, A. M.; Estock, M. A.; Parmar, P. P.; et al.. The human serum proteome: Display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *PROTEOMICS: International Edition* **2003**, *3*, 1345–1364.

2 COMPUTATIONAL METHODS

This chapter reviews the computational methods used throughout this dissertation.

2.1 Classical MD simulations

This section gives an overview of classical MD simulations, which have been well-summarized in several review articles[1–3] and described briefly here. For more than 40 years, classical MD simulations have been used to understand how proteins or other molecular systems will move over time, based the physics governing the interactions between atoms. The first MD simulations were performed on simple gases in the late 1950s,[4] followed by the first MD simulations of proteins in the late 1970s.[5] Due to the advancement of graphical processing units (GPUs) in the past decade, the performance of MD simulations have dramatically improved, which led to a boom in the utility of MD in applications like drug discovery or materials design.[1,2]

The algorithm of MD simulations is fairly straightforward, as depicted in Figure 2.1(a) and summarized in 2.1(b). Given the atomic position of all the atoms in the system (*e.g.* four water molecules as an example in Figure 2.1(a)), the force (F_i) exerted on atom i by all other atoms is calculated using the potential energy of the system (E_{pot}). The potential energy is deduced from the atomic positions and molecular structure that are calibrated to

reproduce energies from quantum mechanical calculations or experimental observables (*e.g.* bond stretching); the equations for computing E_{pot} are known as forcefields. Equation 2.1 shows an example an equation to approximate E_{pot} using forcefield equations.[3] The bonded energies (E_{bonded}) consists of simple ball-spring models for bonding and angle interactions, and sinusodal function for dihedral angle interactions, which distinguishes between eclipsed and staggered conformations. The nonbonded energies ($E_{nonbonded}$, marked in red) are modeled using Lennard-Jones potential and electrostatic interactions, modeled by Coulomb's law.

$$\begin{aligned} E_{pot} &= E_{bonded} + E_{nonbonded} \\ &= \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq}) + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \\ &\quad \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right] \end{aligned} \tag{2.1}$$

These forcefields are calibrated to capture quantum mechanical/experimental results by the tuning the constants (*e.g.* K_r, K_θ, V_n, A_{ij}, B_{ij}) or the equilibrium values between atoms (*e.g.* equilibrium bond length, r_{eq}). Different forcefields may vary in the functional form of Equation 2.1; the most popular force fields for MD simulations are CHARMM,[6] AMBER,[7] GROMOS,[8] and OPLS-AA.[9] This dissertation primarily uses CHARMM

forcefields because it is calibrated for modeling biomolecular systems and small molecules.[6,10]

After the forces acting on each atom of the system is computed, the atomic positions are moved according Newton's law of motion (*i.e.* $F = ma$) by computing velocities v_i, then the new position $x_i(t + \Delta t)$. The MD algorithm in Figure 2.1(b) is iterated until the simulation reaches the timescales of the desired behavior. The MD trajectory outputs position, velocities, and forces over time, which could be used analyze interactions that lead to molecular behavior, as described in Section 2.2. To ensure numerical accuracy and stability, the time step of the MD simulation, Δt, must be small relative to the fastest coordinate change in the system, such as bond vibrations.[11] Since bond vibrations of hydrogen atoms are a few tens of femtoseconds, Δt is typically 1-2 femtoseconds for MD simulations and is a major bottleneck in the simulation procedure.[1] As a result, microsecond-long processes that are barely capturing biological processes is challenging to reach without massively parallel computers that exceed 10^9 steps (*i.e.* 1 μs). Fortunately, the present generation of computers with GPUs allow us to achieve microsecond timescales using popular MD simulation packages, such as GROMACS,[12] NAMD,[13] AMBER,[14] CHARMM,[15] or DESMOND.[16] This dissertation primarily uses the GROMACS Version 2016[12] software package to perform MD simulations.

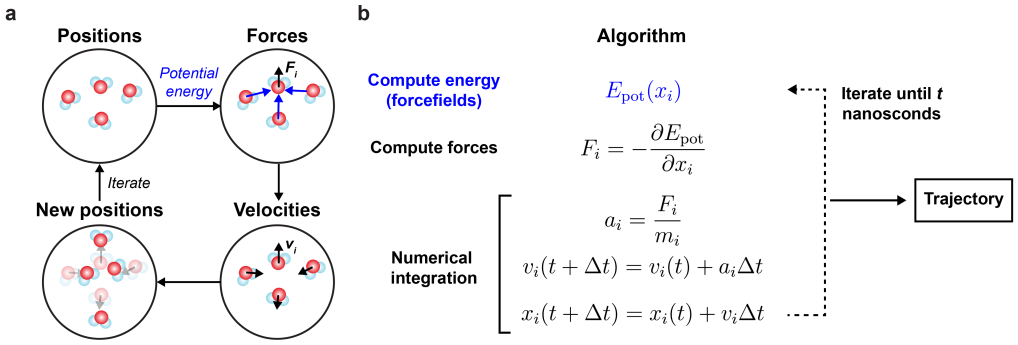


Figure 2.1: (a) Schematic overview of MD simulations using four water molecules as an example. (b) Basic algorithm for MD. The abbreviations are: x_i is the atomic position of atom i, E_{pot} is the potential energy, F_i is the force on atom i, m_i is the mass of atom i, t is the current time, Δt is the time step deviation, v_i is the velocity of atom i. This image was adapted from Ref. 1 with permission from Dove Medical Press Limited.

2.1.1 Benefits and limitations

Classical MD simulations are beneficial because:

- MD can explicitly simulate solvent molecules, which can help interrogate solvent-mediated interactions that may be missed when using implicit solvation models.
- MD can capture cooperative interactions arising from multiple species (*e.g.* protein flexibility that may affect how strongly a small molecule might bind onto active site).[17]
- Forcefields are parameterized for studying biomolecular systems and small molecules, which are highly relevant to the systems in this

dissertation.

Despite the success of MD, the utility of this method is limited by the following challenges:

- Forcefields may sometimes require further refinement to capture experimental behavior.[3]

- MD does not model electron degrees of freedom; hence, bond breaking/forming or reactions cannot be observed in a simulation. In addition, polarization or charge transfer is missed in a typical MD simulation. As a result, if reactions are of interest, *ab initio* MD is frequently used to enable bond breaking/forming events at a higher computational cost compared to classical MD.[18,19]

- Since MD is limited by 1-2 femtosecond timesteps, it is computationally challenging to reach the microsecond timescale necessary to observe biological behavior. To resolve this limitation, coarse-grained models[20] or enhanced sampling techniques[21] are often employed to improve the sampling capabilities of MD. Alternatively, improvements in hardware could help facilitate faster atomistic MD simulations, such as Anton machines.[22]

2.1.2 Solvation free energy calculations

Classical MD simulations could inform the stability of a molecule in a solvent environment by computing solvation free energies. Figure 2.2 depicts the introduction of a reactant molecule (labelled "R") that is introduced in water by slow introduction of van der Waals (VDW) and electrostatic interactions. Solvation free energies quantify the free energy for introducing a solute in solvent environment by accounting for solute-solvent interactions (*e.g.*, hydrogen bonding, ion-dipole interactions, and van der Waals forces) and the solvent reorganization necessary to accomodate the solute. We use solvation free energy calculations in Chapters 4 and 6, and their simulation parameters are briefly summarized below.

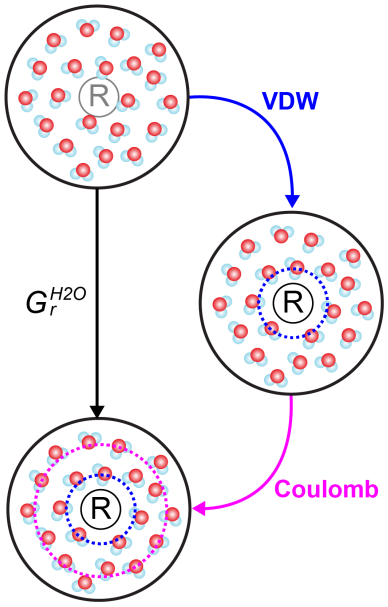


Figure 2.2: Solvation free energy schematic of transferring a reactant from vacuum to pure water by slowly introducing VDW and Coulombic interactions between reactant and solvent. The solvation free energy is denoted as $G_r^{H_2O}$.

The total potential of the system is defined as a function of two parameters, λ_{LJ} and λ_{elec}, which scale the LJ and electrostatic potentials between the solute and solvent, as shown in Equation 2.2.

$$\begin{aligned} U(\lambda_{LJ},\lambda_{elec}) = & U^{LJ}_{M,solv}(\lambda_{LJ}) + U^{elec}_{M,solv}(\lambda_{elec}) + U^{bonded}_{M} + \\ & U^{nonbonded}_{M} + U^{bonded}_{solv} + U^{nonbonded}_{solv} \end{aligned} \tag{2.2}$$

$U^{LJ}_{M,solv}$ and $U^{elec}_{M,solv}$ are the LJ and electrostatic potentials between so-

lute and solvent, U_M^{bonded} and $U_M^{nonbonded}$ are intramolecular bonded and non-bonded potentials of the solute, and U_{solv}^{bonded} and $U_{solv}^{nonbonded}$ are the bonded and non-bonded potentials between all solvent molecules.[23] We performed seventeen independent simulations for each solvation free energy: fourteen in which $\lambda_{elec} = 0.00$ and $\lambda_{LJ} = 0.00$, 0.00922, 0.04794, 0.11505, 0.260634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078, or 1.00, and three in which $\lambda_{LJ} = 1.00$ and $\lambda_{elec} = 0.25$, 0.75, or 1.00. The LJ coupling parameters represent a 12-point Gaussian sequence, used previously to verify ion model parameters.[24,25] In summary, we use solvation free energy calculations to quantify the free energy of introducing a reactant, product, or catalyst in a solvent environment, which is useful to investigate relative stabilities of each species as we vary the solvent environment.

2.1.3 Umbrella sampling simulations

Umbrella sampling (US) is an enhanced sampling technique to estimate free energy barriers along a reaction coordinate. For example, if we were interested in the adsorption of GNP to lipid bilayer, we may set the reaction coordinate as the distance between the GNP to the bilayer, as depicted in Figure 2.3. US simulations are performed by restraining the GNP to a series of reaction coordinate values using a weak harmonic spring, which results in a histogram of positions and forces along the reaction coordinate. Given the forces or positions over time and extent of bias for each reaction

coordinate value, we can use the Weighted Histogram Analysis Method (WHAM)[26] to retrieve the free energy along the reaction coordinate. We use US simulations to investigate GNP adsorption onto bilayers in Chapter 9.

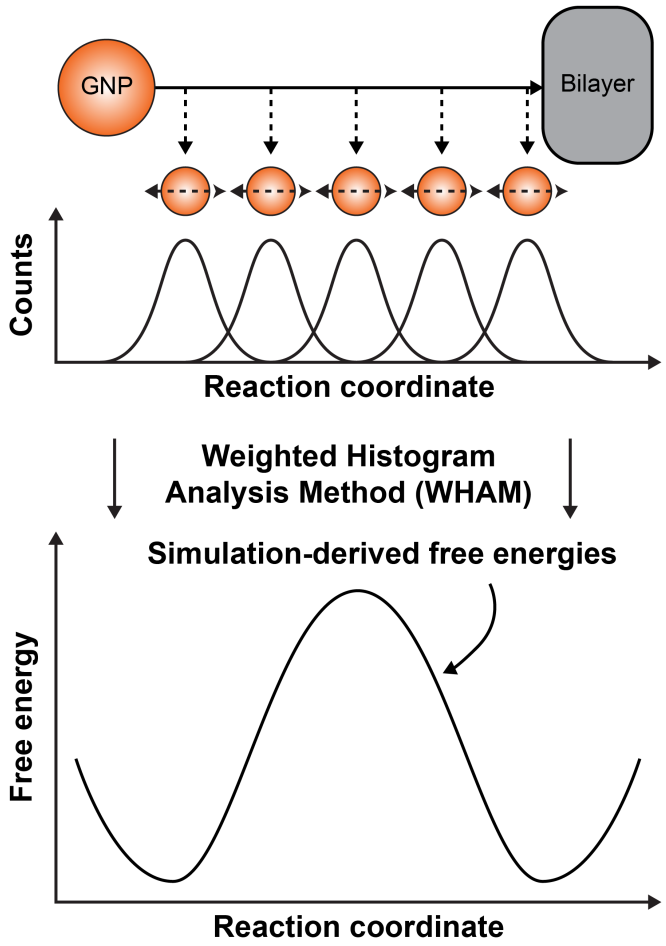


Figure 2.3: Schematic for computing the free energy of moving a GNP to lipid bilayer using US simulations.

2.2 Molecular descriptors and quantitative structure-property relationships

One approach to distilling the molecular information from MD simulations is by developing molecular descriptors, a mathematical tool that transforms chemical information into numbers. For example, accessible surface area (ASA) is a molecular descriptor that quantifies the surface area of a molecule exposed to a solvent. The information encoded within molecular descriptors is highly dependent on the chemical resolution that it is derived from. Table 2.1 lists molecular descriptors obtained from different dimensions from 1D to 4D; generally, the higher the dimension, the higher the molecular resolution that could be characterized by a molecular descriptor. For instance, a 1D descriptor (*e.g.* the molecular formula) cannot distinguish between propanal and acetone molecules that have the same composition of atoms (*e.g.* C_3H_6O), whereas a 2D descriptor (*e.g.* chemical structure) could easily distinguish between propanal and acetone structures. 4D descriptors from MD simulations would give the highest molecular resolution, further highlighting the advantages of using MD.

Once computed, molecular descriptors can also be used to predict chemical properties using quantitative structure-property relationships (QSPR), summarized in Figure 2.4.[27,28] QSPR modeling relates properties of interest (*e.g.* solubility) to molecular descriptors (*e.g.* ASA) and

have been frequently employed in medicinal chemistry, with applications such as drug delivery[29] or nanoparticle cytotoxicity.[28] Furthermore, QSPR modeling is expected to capture complex relationships between molecular structures (microscopic) and properties (macroscopic) of the chemicals without requiring detailed knowledge of the mechanisms of interactions. More importantly, QSPR modeling enables high-throughput screening approaches to tune system properties. The success of QSPR modeling is highly dependent on the quality of the molecular descriptors and the selected model. However, selecting relevant molecular descriptors and choosing the best QSPR model to predict a property is generally challenging, often requiring domain expertise and trial-and-error approaches. These challenges could be addressed with the use of machine learning algorithms that use a data-centric approach to identifying good features and best models for a property of interest.

Table 2.1: Examples of molecular descriptors at different dimensions.

Dimension	Type of information	Example
1D	molecular formula	Propanal/acetone (C_3H_6O)
2D	two-dimensional structure	ChemDraw structure
3D	conformation-dependent structure	PDB structure
4D	time-dependent dynamics of a molecule	MD trajectory

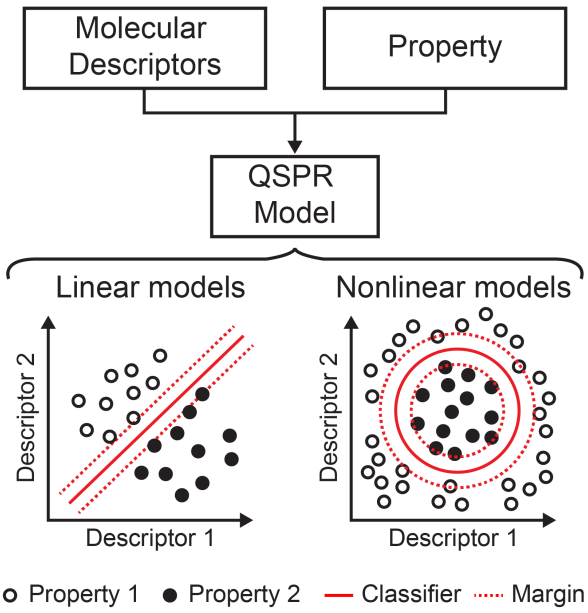


Figure 2.4: Overview of QSPR modeling with examples of linear and nonlinear models for a binary classification of two properties.

2.3 Machine learning models

One of the challenges tackled in this dissertation is finding a good model between the molecular descriptors and the property of the system. If the model is linear (see Figure 2.4), it is easy to interpret how the molecular descriptors relate to the property by analyzing the weights of the regression model, which could provide understanding into the underlying mechanism of action and ways to engineer the system for improved properties.[27,30] However, linearity is not always possible, and nonlinear

approaches using more complex machine learning algorithms may be necessary, such as *k*th nearest neighbor (*k*NN), support vector machines (SVM), random forest (RF), neural networks (NN), and convolutional neural networks (CNN). Each of the models are schematically depicted in Figure 2.5 and succinctly described below:

- ***k*NN**: classification of property 1 (open circles) *versus* 2 (filled circles) is based on how many *k*th neighbors have either property 1 or 2. The number of *k* neighbors is a tunable hyperparameter that varies based on the dataset and often selected using a cross-validation procedure.

- **SVM**: classification of property 1 *versus* 2 is performed by defining a hyperplane that maximizes the margins (denoted in dotted lines) between the two properties. SVMs can also leverage kernel functions (known as the "kernel trick" or "kernel method") that could map the input data into a nonlinear space, which might be easier to define the hyperplane. While *k*NN and SVM models are described here as a classification task, they could also be similarly used for a regression task to predict a property value.

- **RF**: consists of a collection of decision trees that differ by limiting the extent of data shown to each tree; upon prediction of a new input data, the decision trees collectively "vote" to predict the property value. The number of decision trees for RF models is a tunable hy-

perparameter, but generally, more trees equate to a better predictive model.

- **NN**: consists of interconnected layers that have weights learned from an algorithm called "backpropagation," where the weights of each layer are modified based on the error of the output. The backpropagation procedure is iterated across the training data multiple times, where a complete pass over the data is called an "epoch."[31] The number of hidden layers and activation functions of the nodes are tunable hyperparameters; higher numbers of hidden layers equate to a larger number of total weights and a more nonlinear model.

- **CNN**: consists of a collection of convolutional layers, max-pooling layers, and fully connected layers that are designed to extract spatial relationships within images and predict an output class (*e.g.* car in Figure 2.5(e)). CNNs can extract spatial relationships using a "local receptive field," which is a fixed array (*e.g.* 2×2) that iteratively loops through the image and outputs a high or low value based on whether a specific edge/feature is detected. CNNs have been successfully applied to image detection applications[32] and reviewed thoroughly in Ref. 33.

The models presented in Figure 2.5 are not an exhaustive list; there are many other models outside the scope of this dissertation, such as graph-convolutional neural networks,[34] recurrent neural networks,[35] and so on.

Selecting the best machine learning model is based on the application and dataset available, so domain expertise remains crucial to developing good molecular descriptors and hypothesis-based QSPR models.

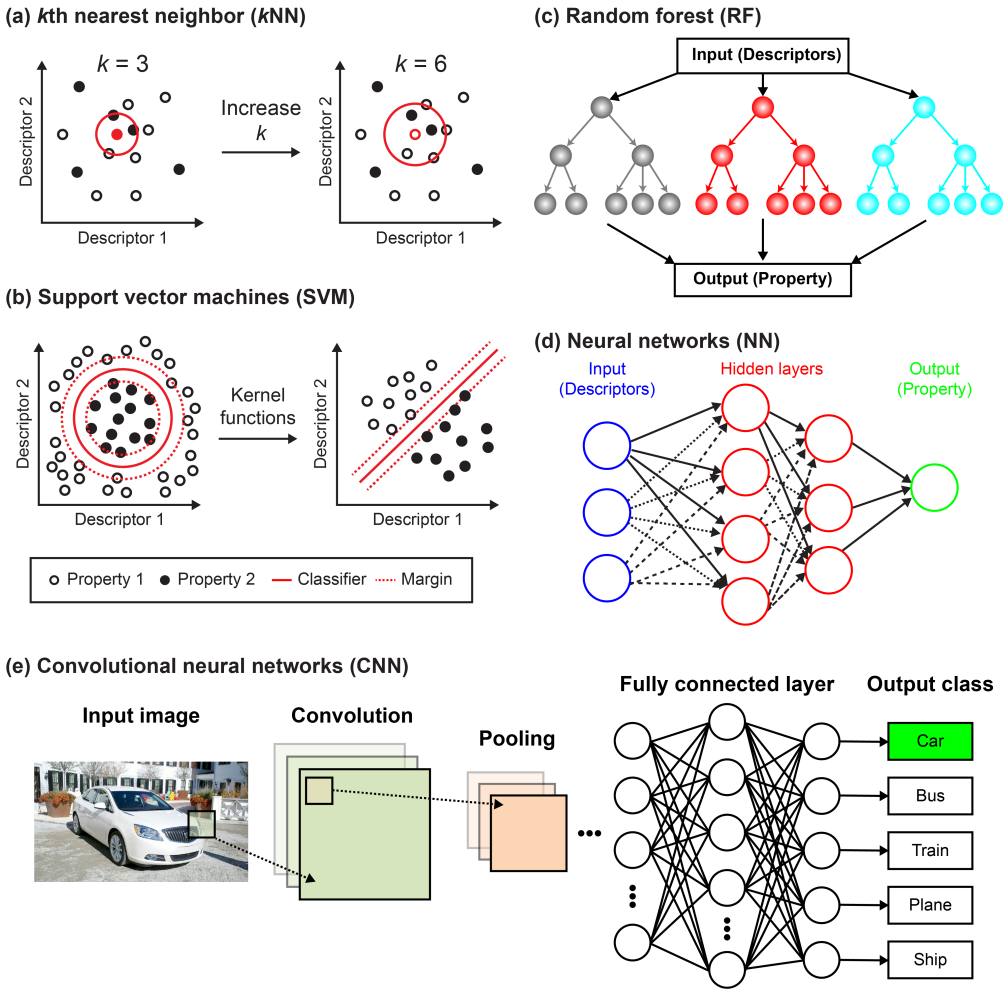


Figure 2.5: Different nonlinear approaches for developing QSPR models: (a) *k*th nearest neighbor, (b) support vector machines, (c) random forest, (d) neural networks, and (e) convolutional neural networks.

2.3.1 Deep learning models improve prediction accuracy

The prediction performance boost when using deep learning models,[36] such as NN or CNN, are best highlighted in the ImageNet competition,[37] a contest requiring a model to classify more than 1.2 million images. Figure 2.6 shows the historical error rate of the best-performing model for the annual ImageNet competition. Prior to deep learning models, the error rate of the best models stagnated around ~25%. Deep learning models were introduced in 2012 and significantly improved the prediction accuracy close to human error of ~5% by 2015. These findings suggest that if deep learning models could accurately classify images, they could be used to drastically improve QSPR models for chemical design.[31]

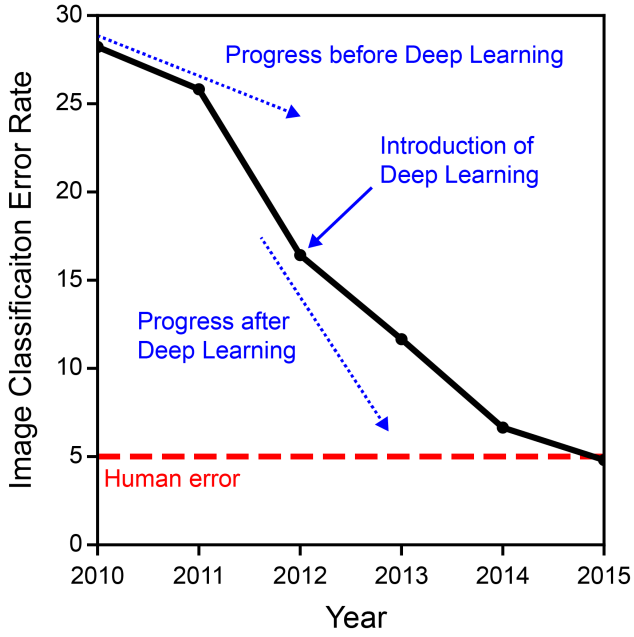


Figure 2.6: Historical error rate of the best-performing image classification model in the annual ImageNet competition. Deep learning significantly improved the error rate to ~15% in 2012 and reaching close to human-level accuracy ~5% by 2015. This image was adapted from Ref. 31 with permission from John Wiley and Sons.

2.4 References

[1] Hospital, A.; Goñi, J. R.; Orozco, M.; Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC* **2015**, *8*, 37.

[2] Hollingsworth, S. A.; Dror, R. O. Molecular dynamics simulation for all. *Neuron* **2018**, *99*, 1129–1143.

[3] Durrant, J. D.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC biology* **2011**, *9*, 1–9.

[4] Alder, B. J.; Wainwright, T. E. Phase transition for a hard sphere system. *The Journal of chemical physics* **1957**, *27*, 1208–1209.

[5] McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590.

[6] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods* **2017**, *14*, 71–73.

[7] Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* **2018**, *115*, E4758–E4766.

[8] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; et al.. The GROMOS software for biomolecular simulation: GROMOS05. *Journal of computational chemistry* **2005**, *26*, 1719–1751.

[9] Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al.. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of chemical theory and computation* **2016**, *12*, 281–296.

[10] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **2008**, *29*, 1859–1865.

[11] Elber, R. Perspective: Computer simulations of long time dynamics. *The Journal of chemical physics* **2016**, *144*, 060901.

[12] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[13] Nelson, M. T.; Humphrey, W.; Gursoy, A.; Dalke, A.; Kalé, L. V.; Skeel, R. D.; Schulten, K. NAMD: a parallel, object-oriented molecular dynamics program. *The International Journal of Supercomputer Applications and High Performance Computing* **1996**, *10*, 251–268.

[14] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3*, 198–210.

[15] Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al.. CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **2009**, *30*, 1545–1614.

[16] Bowers, K. J.; Chow, D. E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; et al. In *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*; IEEE; pp 43–43.

[17] Lexa, K. W.; Carlson, H. A. Full protein flexibility is essential for proper hot-spot mapping. *Journal of the American Chemical Society* **2011**, *133*, 200–202.

[18] Paquet, E.; Viktor, H. L. Computational methods for Ab initio molecular dynamics. *Advances in Chemistry* **2018**, *2018*, 9839641.

[19] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[20] Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chemical Society Reviews* **2013**, *42*, 6801–6822.

[21] Fujisaki, H.; Moritsugu, K.; Matsunaga, Y.; Morishita, T.; Maragliano, L. Extended phase-space methods for enhanced sampling in molecular simulations: a review. *Frontiers in bioengineering and biotechnology* **2015**, *3*, 125.

[22] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; et al..

Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM* **2008**, *51*, 91–97.

[23] Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. Prediction of absolute solvation free energies using molecular dynamics free energy perturbation and the OPLS force field. *Journal of chemical theory and computation* **2010**, *6*, 1509–1519.

[24] Bonthuis, D. J.; Mamatkulov, S. I.; Netz, R. R. Optimization of classical nonpolarizable force fields for OH- and H3O+. *The Journal of chemical physics* **2016**, *144*, 104503.

[25] Horinek, D.; Mamatkulov, S. I.; Netz, R. R. Rational design of ion force fields based on thermodynamic solvation properties. *The Journal of chemical physics* **2009**, *130*, 124507.

[26] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **1995**, *16*, 1339–1350.

[27] Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al.. QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry* **2014**, *57*, 4977–5010.

[28] Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. Nano-QSAR modeling for predicting the cytotoxicity of metal oxide nanoparticles using novel descriptors. *RSC advances* **2016**, *6*, 25766–25775.

[29] Zakharov, A. V.; Varlamova, E. V.; Lagunin, A. A.; Dmitriev, A. V.; Muratov, E. N.; Fourches, D.; KuzâŁ™min, V. E.; Poroikov, V. V.; Tropsha, A.; Nicklaus, M. C. QSAR modeling and prediction of drug–drug interactions. *Molecular pharmaceutics* **2016**, *13*, 545–556.

[30] Guha, R.; Stanton, D. T.; Jurs, P. C. Interpreting computational neural network quantitative structure- activity relationship models: A detailed interpretation of the weights and biases. *Journal of chemical information and modeling* **2005**, *45*, 1109–1121.

[31] Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *Journal of computational chemistry* **2017**, *38*, 1291–1307.

[32] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.

[33] Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **2017**, *29*, 2352–2449.

[34] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **2019**, *10*, 370–377.

[35] Müller, A. T.; Hiss, J. A.; Schneider, G. Recurrent neural network model for constructive peptide design. *Journal of chemical information and modeling* **2018**, *58*, 472–479.

[36] Alzubaidi, L.; Zhang, J.; Humaidi, A. J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M. A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* **2021**, *8*, 1–74.

[37] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al.. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.

3 UNIVERSAL KINETIC SOLVENT EFFECTS IN ACID-CATALYZED REACTIONS OF BIOMASS-DERIVED OXYGENATES

This chapter focuses on developing classical molecular dynamics simulations to investigate solvent-mediated effects on biomass conversion reactions. For each chapter henceforth, we will begin by asking the research questions that we are tackling, followed by a short summary of the findings:

- How do we develop a workflow to model biomass-derived species in aqueous mixtures with cosolvents?
- What physically motivated molecular descriptors could we compute from molecular simulations that may relate to experimental reaction rates?
- How do we merge these descriptors into a predictive model for reaction rates and can the information from molecular dynamics alone be predictive of reactivity (despite not modeling the catalyst or reaction mechanism)?

In this chapter, the rates of Brønsted-acid-catalyzed reactions of ethyl tert-butyl ether, tert-butanol, levoglucosan, 1,2-propanediol, fructose, cel-

This chapter was reproduced from Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628 with permission from the Royal Society of Chemistry.[1] The electronic supporting information is cited as Ref. 2. T. W. Walker and A. K. Chew contributed equally to this work. T. W. Walker, H. Li, B. Demir, Z. C. Zhang, G. W. Huber, and J. A. Dumesic designed and performed the reaction kinetic studies.

lobiose, and xylitol were measured in solvent mixtures of water with three polar aprotic cosolvents: γ-valerolactone; 1,4-dioxane; and tetrahydrofuran. As the water content of the solvent environment decreases, reactants with more hydroxyl groups have higher catalytic turnover rates for both hydrolysis and dehydration reactions. We present classical molecular dynamics simulations to explain these solvent effects in terms of three simulation-derived observables: (1) the extent of water enrichment in the local solvent domain of the reactant (Γ); (2) the average hydrogen bonding lifetime between water molecules and the reactant (τ); and (3) the fraction of the reactant accessible surface area occupied by hydroxyl groups (δ), all as a function of solvent composition. We develop a model, constituted by linear combinations of these three observables, that predicts experimentally determined rate constants as a function of solvent composition for the entire set of acid-catalyzed reactions.

3.1 Introduction

Acid-catalyzed reactions in the liquid phase are ubiquitous in the production and upgrading of renewable biomass-derived oxygenates,[3–5] which have garnered interest as sources of organic carbon for the production of renewable liquid fuels.[5–10] An important variable to control the reactivity and selectivity for these catalytic processes is the solvent composition, which affects reaction rates,[11,12] product selectivities,[13,14] the stability of

desired products,[3,15–17] and the economics of downstream separations. Prior studies have shown that aqueous mixtures containing polar aprotic cosolvents (*i.e.*, mixed-solvent environments) are of particular interest in this context; a minimum amount of water is often required to facilitate the solvation of biomass-derived materials, while the cosolvent can enhance reaction performance.[11,18,19] For example, the rate of xylose dehydration to furfural in aqueous mixtures of γ-valerolactone increases 30-fold compared to the same reaction in pure water, while the formation of undesired humins via degradation of the reactant and product is suppressed,[12] improving the furfural selectivity by over 50%.

While extensive knowledge has been accumulated regarding optimal solvent compositions for key applications,[12,13,15,20,21] it is not generally possible to anticipate how mixed-solvent environments will perform in new processes a priori, because the mechanistic details underlying kinetic solvent effects in multicomponent systems are not well understood.[22–24] Computational efforts in the past decade have relied on *ab initio* quantum chemical methods to quantify the effects of solvent on the barriers to elementary reaction steps.[25] These studies have provided detailed insights in important case studies,[26–29] but broadly applicable design rules have not been developed due (in part) to the limitation of *ab initio* techniques to capture the slow intermolecular re-organizations that constitute solvent-solute interactions.[29–35]

In contrast to *ab initio* methods, classical molecular dynamics (MD)

simulations can access longer time scales (~μs) and larger length scales (~nm), and at a lower computational cost. MD simulations therefore permit characterization of the solvent environment in the immediate vicinity of a reactant molecule (*i.e.*, the local solvent domain of the reactant), which can be compared to the solvent environment far from the reactant (*i.e.*, the bulk solvent domain).[18,21,36] For example, MD simulations have examined the preferred configurations of solvent molecules at biomass-relevant reactant surfaces.[18,21] Based on key reaction rate measurements in this study and insights from our recent work,[37] we hypothesize that trends in acid-catalyzed reaction rates as a function of solvent environment can be related to the formation and properties of a water-rich or -deficient local domain near the reactant. These properties can be quantified via classical MD simulations, and moreover can be determined for a large series of possible reactants and solvent compositions given the computational efficiency of classical MD simulations.

In this chapter, we report the effects of mixed-solvent environments, consisting of water mixed with a polar aprotic cosolvent, on experimentally determined rates of Brønsted-acid-catalyzed reactions for seven biomass-derived model compounds. Irrespective of the selected cosolvent, consisting of 1,4-dioxane, γ-valerolactone, and tetrahydrofuran, we find that reactants with more available hydroxyl groups become more reactive as the water content of the solvent environment is decreased, and this behavior is true for both hydrolysis and dehydration reactions considered in this

study. By contrasting properties between local and bulk solvent domains using MD simulations, we develop a computational approach to predict reaction rates as a function of solvent composition. We predict experimentally determined reaction rates using three computationally determined observables: (1) the extent of water enrichment in the local solvent domain of the reactant (Γ); (2) the average hydrogen bonding lifetime between water molecules and the reactant (τ); and (3) the fraction of the reactant accessible surface area occupied by hydroxyl groups (δ).

To our knowledge, the approach developed in this study provides the first tool of its kind in the context of biomass conversion in multicomponent solvent environments. As such, this study represents a step toward the model-predictive design of liquid-phase biomass conversion technologies. Moreover, this approach demonstrates that contrasting properties between local and bulk solvent domains using MD simulations can provide insight into the kinetics of acid-catalyzed reactions in mixed-solvent environments.

3.2 Methods

3.2.1 Reaction kinetics studies

Reactions were carried out in closed, 10 mL thick-walled glass reactors. In a typical experiment, an appropriate amount of reactant (*e.g.*, xylitol (XYL)), acid catalyst, water and organic cosolvent (*e.g.*, 1,4-dioxane (DIOX))

were charged to the reactors, which were then sealed and placed in an oil bath at the appropriate temperature. Reactors were removed at times corresponding to the desired reaction time, and quenched in an ice bath at 273 K. Reaction products were analyzed using high-performance liquid chromatographs equipped with differential refractometers and photo diode array detectors, or gas chromatographs equipped with flame ionization detectors. All products were quantified using calibration curves with external standards. Conditions for each reaction (temperature, fractional conversion, *etc.*) were chosen so that each reaction was selective (>90%) to a single product. This procedure allowed for reliable measurements of rate constants based on the rate reactant consumption in a MATLAB-based optimization routine as detailed in the ESI.[2] Trifluoromethane sulfonic (triflic) acid (pK_a, H_2O = 14.7, pK_a, DMSO = -14.3, pK_a, MeCN = 0.7) was used as catalyst in all experiments, which has been shown to behave as a strong acid even in pure polar aprotic solvents.[37–39] We thus assume complete dissociation of the acidic proton in all mixed solvent environments, allowing for normalization of the apparent rate constants on a per-proton basis.

3.2.2 Molecular dynamics simulations

Classical molecular dynamics simulations were performed using Gromacs version 2016.[40] Reactants and cosolvents were parameterized using the CGenFF/CHARMM36 force fields[41–43] while water was modeled using

the SPC/E model.[44] Solvent mixtures were initially equilibrated in an *NPT* ensemble using a Berendsen barostat at 1 bar and velocity-rescale thermostat at 300 K. The initial simulation box size containing the cosolvent and water was set to $(6\ \text{nm})^3$ in all simulations as schematically illustrated in Figure 3.1. A single reactant molecule was then added to the system, equilibrated with the same barostat and thermostat at the temperature of the reaction for 500 ps, followed by a *NPT* MD simulation with Parrinello-Rahman and Nóse-Hoover as the barostat and thermostat, respectively, for 200 ns. The accessible surface area, radial distribution functions, and preferential exclusion coefficients were calculated using the final 190 ns of simulation data. For preferential exclusion coefficients, the simulation trajectory was partitioned into two separate trajectories of 95 ns to obtain the error in the calculations as the standard deviation of the calculated values. Simulation analysis was performed using the MDTraj tool box[45] and analysis tools developed in-house. A total of 91 MD simulations (~18.2 μs simulation time) was performed. Details of the simulation parameters are available in the ESI, as well as methods for computing simulation observables.[2]

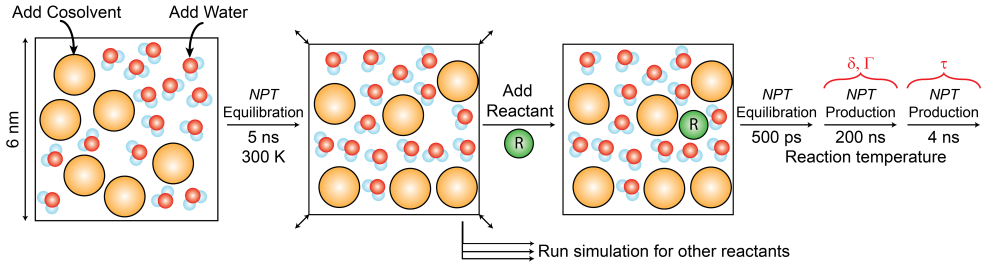


Figure 3.1: Schematic depiction of mixed-solvent system preparation. "R" denotes the reactant. Note that the second production trajectory, used to calculate hydrogen bonding lifetimes, was not always 4 ns; some reactants required a longer simulation time to obtain accurate hydrogen-bonding lifetime data.

3.2.3 Hydrogen bonding lifetime

Hydrogen bonding lifetimes were calculated based on the Luzar and Chandler approach.[46] The kinetics of hydrogen bond formation and breaking is described in Equation 3.1.

$$\frac{dc(t)}{dt} = -kc(t) + k'n(t) \tag{3.1}$$

k and k' are rate constants for breaking and making hydrogen bonds, respectively. $c(t)$ is the hydrogen bond correlation function that outputs the probability a hydrogen bond is intact at time t given that it was intact at $t = 0$. $n(t)$ is the probability that a hydrogen bond is broken but the hydrogen bonding groups are still within hydrogen bonding distance. The average bonding lifetime (τ_{HB}) could be found by the reciprocal of the forward rate constant ($1/k$) and calculated using Gromacs 5.0.1.[47–51]

Additional details of hydrogen bonding lifetime calculations are available in the ESI.[2]

3.3 Results and Discussion

3.3.1 Universal effects of reactant and cosolvent properties on reactivity

Figure 3.2 shows the seven Brønsted-acid catalyzed reactions considered in this study, including those reported in our prior work.[37] For brevity, the reactant abbreviations in Figure 3.2 are used throughout this text. Reactions were carried out in pure water, and in aqueous mixtures of the three polar aprotic cosolvents: DIOX; γ-valerolactone (GVL); and tetrahydrofuran (THF). Reactants in Figure 3.2 are organized according to decreasing hydrophilicity, as estimated by the accessible hydroxyl fraction (δ), which we have defined as the accessible surface area (ASA) occupied by the (N) hydroxyl groups in a reactant molecule ($ASA_{OH,k}$) normalized by the ASA occupied by the (M) total atoms in the molecule, shown in Equation 3.2.

$$\delta = \frac{\sum_{k=1}^{N} ASA_{OH,k}}{\sum_{l=1}^{M} ASA_{l}} \tag{3.2}$$

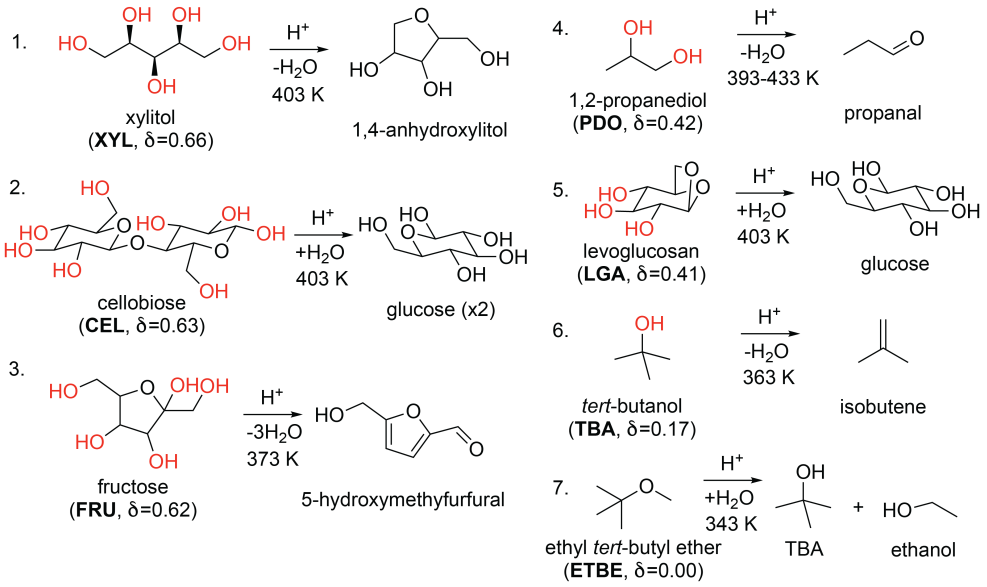


Figure 3.2: Brønsted acid-catalyzed reactions of seven model compounds. Rate constants associated with reactions 3, 4, and 6 were taken from prior work.[11,37] Hydroxyl groups are highlighted in red for emphasis. Reactants are arranged according to decreasing hydrophilicity, as estimated by the δ parameter.

The forward rates of all reactions (r_i) in Figure 3.2 are described by apparent first-order kinetics with respect to the concentrations of the reactants (C_i) and acidic protons in solution (C_{H^+}).[37,52–54] Accordingly, we measured apparent rate constants in each solvent environment by fitting the experimental reactions kinetics data (obtained in batch reactors) to expressions in the form of Equation 3.3, where ($k^i_{org,j}$) is the apparent rate constant for the ith reaction, and the subscript denotes the identity and composition (in jth mass fraction) of the organic cosolvent shown in

Equation 3.3.

$$r_i = -\frac{dC_i}{dt} = k^i_{org,j} C_i C_{H^+} \tag{3.3}$$

Figure 3.3 shows the measured apparent rate constants for XYL dehydration in DIOX-water mixtures ($k^{XYL}_{DIOX,j}$), normalized by the rate constant in pure water ($k^{XYL}_{H_2O}$), as a function of the mass fraction of DIOX in the solvent environment (m_{DIOX}). As the mass fraction of DIOX increases, the value of the rate constant for XYL dehydration increases by nearly two orders of magnitude. In general, the values of the rate constants measured in this study are a strong function of solvent composition, and this phenomenon has previously been noted elsewhere.[11,12,37]

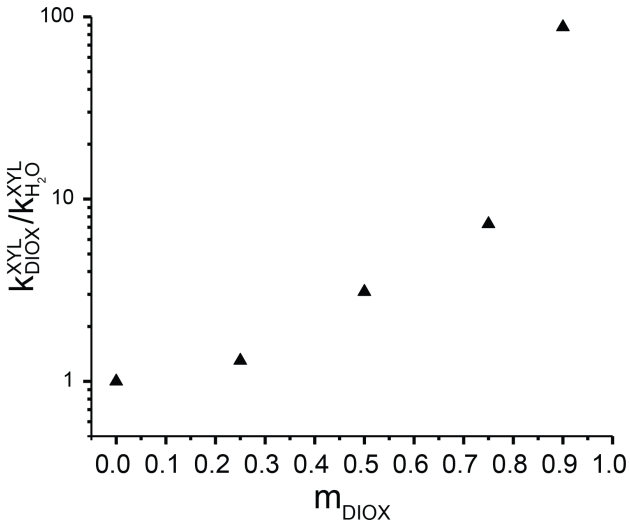


Figure 3.3: Apparent rate constant for XYL dehydration normalized by the rate constant in pure water versus the mass fraction of the organic cosolvent in DIOX-water mixtures. *Reaction conditions: 75 - 200 mM XYL; 0.03 - 1.3 M trifluoromethane sulfonic (triflic) acid; 403 K.*

To compare reactivity trends across different reactions and solvent environments, the rate constants associated with each of the reactions in Figure 3.2 were described in terms of a kinetic solvent parameter ($\sigma^{i}_{org,j}$) as shown in Equation 3.4.

$$\sigma^{i}_{org,j} = \log_{10}\left(\frac{k^{i}_{org,j}}{k^{i}_{H_2O}}\right) \tag{3.4}$$

Positive kinetic solvent parameters indicate an increase in the rate of reaction in a particular solvent environment compared to the same reaction in

Table 3.1: Kinetic solvent parameters for the Brønsted-acid-catalyzed dehydration of XYL in mixtures of water with three organic cosolvents. Reaction rate constant of XYL in water is: $k_{H_2O}^{XYL} = 1.04 \times 10^{-4} \pm 9.3 \times 10^{-6}$ L mol^{-1}s^{-1} *Reaction conditions: 403 K; 0.015-1.3 M triflic acid. Confidence intervals were calculated at the 95% confidence level.* N/A means not available due to phase separation between water and organic cosolvent.

m_{org}	$\sigma_{GVL,j}^{XYL}$	$\sigma_{DIOX,j}^{XYL}$	$\sigma_{THF,j}^{XYL}$
0.90	2.05 ± 0.07	1.80 ± 0.08	1.85 ± 0.10
0.75	1.02 ± 0.08	1.02 ± 0.08	0.74 ± 0.13
0.50	2.05 ± 0.07	0.50 ± 0.08	N/A
0.25	0.11 ± 0.10	0.18 ± 0.08	0.23 ± 0.07

pure water, while negative values have a converse implication. Table 3.1 presents the kinetic solvent parameters for the rate of XYL dehydration in aqueous mixtures of up to 90 wt% DIOX, GVL, and THF. These results demonstrate a general trend of increasing reactivity with decreasing water content for each of the three cosolvents. A complete list of the kinetic solvent parameters collected in this study is presented in the ESI.[2]

Figure 3.4(a) presents the kinetic solvent parameters for XYL and tert-butanol (TBA) dehydration in aqueous mixtures of DIOX, GVL, and THF as a function of the mass fraction of the organic cosolvent (m_{org}). These two reactions represent upper and lower limits in our dataset with respect to reactant hydrophilicity and kinetic behavior. XYL has an accessible hydroxyl fraction (δ) of 0.66 and becomes monotonically more reactive with decreasing water content of the solvent environment. In contrast, TBA has an accessible hydroxyl fraction of 0.17, and its reactivity is a

non-monotonic function of solvent composition.[37] In general, the extent to which the rate of TBA dehydration is affected by the increasing addition of organic cosolvent is smaller than that of XYL dehydration.

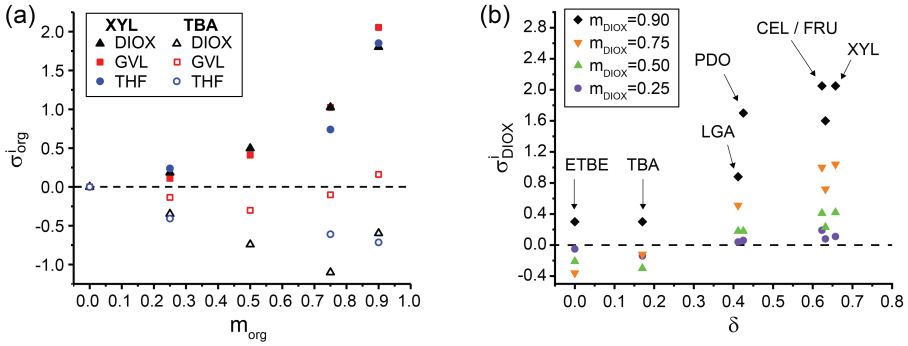


Figure 3.4: Kinetic solvent parameters (σ^i_{org}) as a function of: (a) solvent composition in aqueous mixtures of DIOX, GVL, and THF (open symbols = TBA, closed symbols = XYL), and; (b) the accessible hydroxyl fraction (δ) in DIOX-water mixtures.

The comparison of XYL and TBA dehydration suggests that reactivity trends with respect to the water content of the solvent environment may depend on the hydrophilicity of the reactant, as we have postulated elsewhere.[37] Figure 3.4(b) presents the kinetic solvent parameters as a function of the accessible hydroxyl fraction for all seven reactions in aqueous DIOX mixtures. At a fixed solvent composition, the kinetic solvent parameters generally become more positive, indicating increasing rates of reaction compared to pure water, with an increasing accessible hydroxyl fraction. Similar behavior was observed in GVL and THF mixtures (see ESI).[2] Note that this behavior is observed for both the dehydrations and hydrolysis

reactions considered in this study.

3.3.2 Proposed mechanism: reaction rates correlate with formation of water-enriched local solvent domain

To understand the aforementioned trends in reactivity, we note that rates of chemical reactions are controlled by the thermodynamic properties of the reactants, catalyst, and transition states in the elementary steps.[24] A quantitative understanding of solvent effects therefore requires knowledge of the reaction mechanism, and a rigorous characterization of the transition-state-solvent interactions in each of the kinetically relevant steps.[29,34,37]

In many cases, however, catalytic reactions occur via a sequence of elementary steps where a single step is rate-limiting, and the rate is then controlled by the thermodynamic properties of the transition state for this step.[55–57] In these cases,[58,59] the reaction mechanism may be analyzed via a sequence of quasi-equilibrated steps, such that the reactant and proton are treated as being in equilibrium with the transition state. Furthermore, transition states in acid-catalyzed reactions typically display strong carbocation-like character.[57,60] Accordingly, as shown in Figure 3.5(b), we consider the acid-catalyzed reactions in this study as being composed of two generalized steps, both of which can be impacted by the solvent:

(*i*) the transfer of a proton from the bulk domain to the reactant, and;

(*ii*) the formation of a carbocation-like transition state.

For step (*i*), we have shown that water-enriched local solvent domains form around hydroxyl groups in the presence of polar aprotic cosolvents.[60] In mixed-solvent environments, hydrophilic reactants thus drive the formation of local solvent domains in which the local density of water molecules near the reactant is greater than the density of water molecules in the bulk solvent, as shown in Figure 3.5(a). A proton is therefore destabilized in the bulk solvent relative to the local domain, because of its higher affinity for water than for the organic phase, leading to a thermodynamic driving force for transferring the proton to the reactant.[12,61–63] Vlachos and co-workers have postulated a similar mechanism for fructose dehydration in DMSO-water mixtures.[27]

For step (*ii*), we hypothesize that the stability of the reactant, proton, and transition state are correlated in these local solvent domains, because water molecules that bind strongly to the reactant are preorganized into configurations that stabilize charged transition states.[34,64] Similar solvent preorganization is thought to contribute to enzyme catalytic efficiency,[65] has been speculated as a key effect in the acid-catalyzed glucose to fructose isomerization reaction,[34] and may contribute to the enhanced reactivity of hydronium ions in confined environments (zeolites).[66,67]

We now explore whether reactant-solvent-cosolvent interactions can be tuned to deliberately drive the formation of water-enriched local domains near hydrophilic reactants. Water and the hydrophilic reactant are then characterized as being confined to a local domain, with confinement

enhancing reactivity by increasing reactant-proton association, and stabilizing the carbocation-like transition states common to acid-catalyzed reaction mechanisms. With MD simulations, we probe our hypotheses by analyzing the local solvent domain near the reactant and deriving simulation measurables that can be used to predict experimental reaction rates. Simulation snapshots for XYL in pure water and 90 wt% DIOX are shown in Figure 3.5(c) and Figure 3.5(d), respectively.

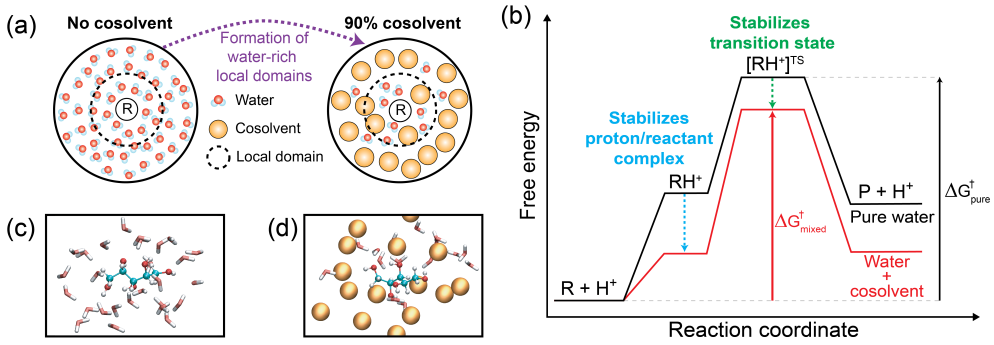


Figure 3.5: (a) Role of cosolvent molecules on the distribution of solvent molecules. Favorable interactions with hydrophilic reactants in mixed-solvent environments drive the formation of water-rich local domains around the reactant. While there are fewer water molecules in the local domain relative to pure water, the local water density is enriched relative to the bulk density in the solvent mixture. (b) Proposed effect of cosolvent molecules on a reaction free energy landscape. Stabilization of the proton and transition state in the water-rich local domain, relative to the bulk domain, lowers the apparent free energy barrier for the reaction in a mixed-solvent environment. (c, d) MD simulation snapshots of XYL in (c) pure water and (d) 90 wt% DIOX, which is drawn as a single representative bead to match the schematics in (a).

3.3.3 MD simulations: formation of water-enriched local domains in solvent mixtures

Figure 3.6 shows the radial distribution function (RDF), which quantifies the density of water molecules at a radius r away from a central molecule normalized by the bulk water density. The RDF is schematically depicted in Figure 3.6(a) for a mixed-solvent environment in which XYL is the central molecule. The RDF for XYL in a 90 wt% DIOX-water mixture and in pure water (0 wt% DIOX-water) is shown in Figure 3.6(b). We define the cutoff, r_{cutoff}, between the local and bulk solvent domains as the distance at which the RDF reaches unity.

RDFs between TBA-water and XYL-water in various DIOX-water mixtures are shown in Figure 3.6(c) and Figure 3.6(d), respectively. From the RDFs, we find that the water content in the local solvent domain of each reactant increases compared to the bulk solvent domain when a high concentration of cosolvent is present in the solvent environment. This behavior is apparent from the increase in the magnitude of the first solvation peak for systems containing large concentrations of DIOX relative to its magnitude for a pure water system, indicating that water preferentially partitions to the local solvent domain around the reactant. The RDF for XYL has a broader first solvation peak than TBA, which indicates greater water enrichment in the local solvent domain when the reactant has more hydroxyl groups.

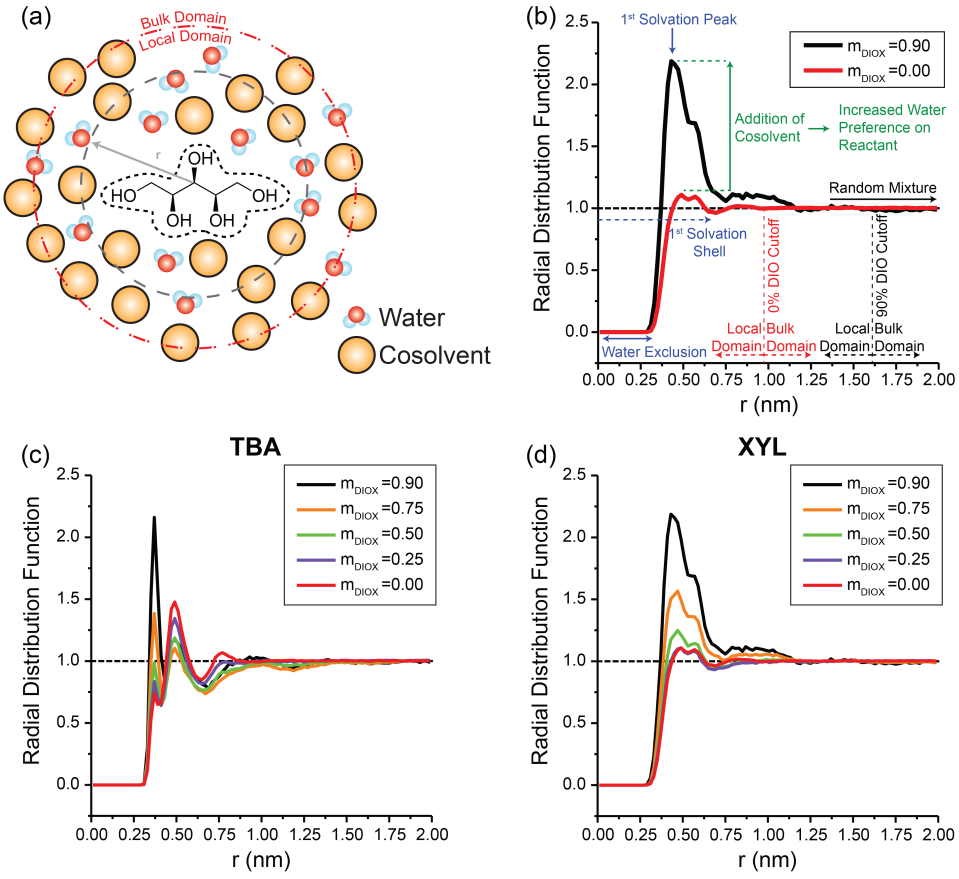


Figure 3.6: (a) Schematic depiction of the radial distribution function for XYL. The distance, r, is calculated between the center of mass of the reactant and the oxygen atom of each water molecule. (b) Radial distribution function for XYL in 90 wt% DIOX and pure water ($m_{DIOX} = 0$). The cutoff between local and bulk domain is defined as the distance when the RDF between the reactant and water reaches unity (*i.e.* a random mixture). (c,d) Radial distribution function for TBA (c) and XYL (d) for various wt% of organic solvent.

3.3.4 Quantifying water enrichment within the local domain of the reactant

Because RDFs are difficult to compare across reactants with different concentrations of cosolvents, we calculated the preferential exclusion coefficient (Γ) as a molecular descriptor to quantify the local domain composition around the reactant. We define Γ as the excess number of cosolvent molecules within the local solvent domain of the reactant relative to the bulk solvent domain. Preferential exclusion coefficients are calculated from the MD simulations according to Equation 3.5, where n_C and n_W are the total number of cosolvent and water molecules, and the superscripts L and B indicate molecules within the local and bulk domains, respectively.[68–71]

$$\Gamma = -\left\langle n_C^L - n_W^L \left(\frac{n_C^B}{n_W^B} \right) \right\rangle \tag{3.5}$$

Positive values of Γ indicate that the concentration of cosolvent is lower in the local solvent domain of the reactant than in the bulk solvent domain. Positive Γ is also referred to as preferential hydration,[71] because the exclusion of cosolvent indicates that the reactant has a higher affinity for water. Negative values of Γ indicate that the concentration of cosolvent is higher in the local solvent domain of the reactant than in the bulk solvent domain and that the reactant has a higher affinity for the cosolvent.

Preferential exclusion coefficients (Γ) calculated for TBA and XYL in DIOX-water and GVL-water mixtures at various cosolvent concentrations

are shown as filled lines in Figures 3.7(a) and (b). Experimentally determined kinetic solvent parameters (σ) are depicted as dashed lines for comparison. For TBA, Γ is negative with the exception of 90% GVL-water. Conversely, Γ is positive for XYL across the range of cosolvent compositions, which means that XYL preferentially excludes cosolvent and has a higher affinity for water. We find that Γ correlates well with σ, even capturing the non-monotonic behavior in TBA.

Figures 3.7(c) and 3.7(d) show a strong linear correlation between simulation-derived Γ and the experimentally determined σ, as indicated by Pearson correlation coefficients (Pearson's r). Pearson correlation coefficients close to 1 indicate a total positive linear correlation, whereas values near -1 indicate a linear negative correlation, and zero indicates no linear correlation. Pearson correlation coefficients for all reactants in each of the three cosolvent mixtures are summarized in Table 3.2. We find that Γ and σ are highly correlated (Pearson's $r \geqslant 0.80$) for the majority of the systems, with the exception of THF-water systems.

The agreement between simulations and experiments suggests that higher reaction rates correspond to the enrichment of water near the reactant despite differences in reactant hydrophilicity and reaction mechanisms. However, the poor correlation between Γ and σ in some reactant/cosolvent environments suggests that additional parameters that characterize the local solvent domain will improve our understanding of the solvent effects that contribute to experimentally determined reaction

rates.

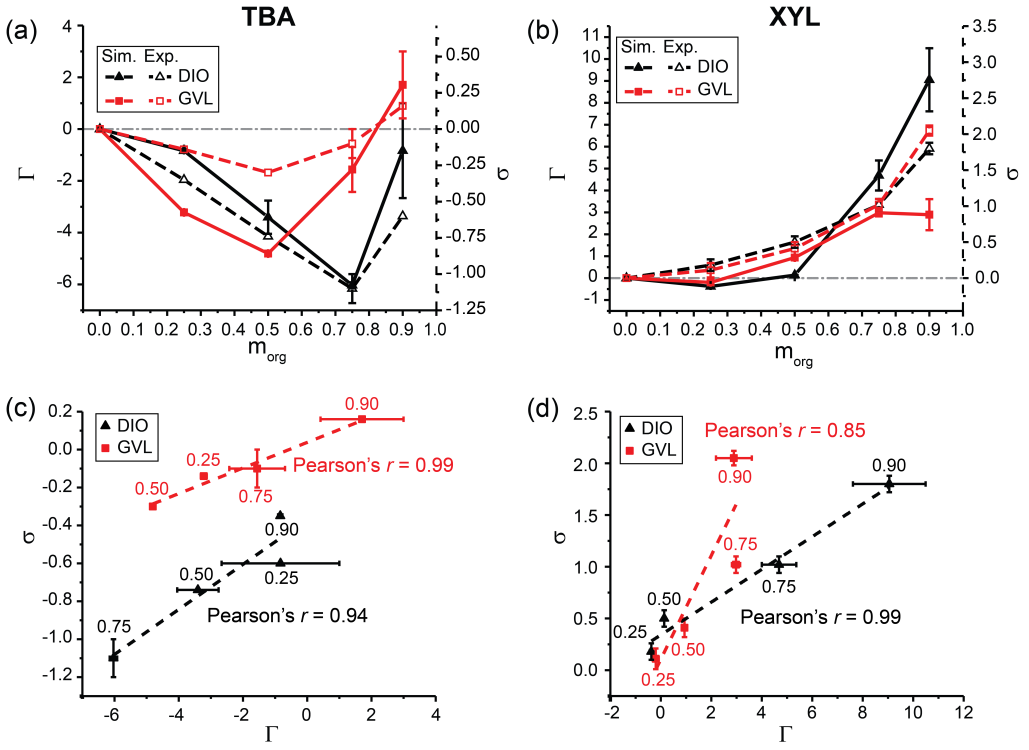


Figure 3.7: Relationship between experimentally determined kinetic solvent parameters (σ) and simulated preferential exclusion coefficient (Γ) for (a) TBA and (b) XYL for various wt% of organic cosolvent in GVL-water and DIOX-water mixtures. The gray dotted line denotes when σ and Γ are zero. Kinetic solvent parameters are also plotted against Γ for (c) TBA and (d) XYL. The dashed lines in (c) and (d) represent the best-fit line. Data points are labeled with the wt% of the organic cosolvent. Pure water systems have been omitted from parts (c) and (d) because σ and Γ will always be zero.

Table 3.2: Pearson correlation coefficient between the kinetic solvent parameters (σ) and the preferential exclusion coefficient (Γ) for various reactants and cosolvents.

	Reactant							
Cosolvent	ETBE	TBA	LGA	PDO	FRU	CEL	XYL	All*
DIOX	-0.08	0.94	0.84	0.98	0.96	0.98	0.99	0.84
GVL	0.80	0.99	0.91	1.00	0.93	0.80	0.85	0.72
THF	0.49	-0.68	0.76	0.55	0.45	0.55	0.26	0.60

* Uses data from all reactants and all four cosolvent wt%

3.3.5 Quantifying reactant-water hydrogen bonding strength

Following the hypothesis that charged transition states may be stabilized by water molecules in the local solvent domain that are preorganized into favorable binding configurations, we next calculated the average reactant-water hydrogen bonding lifetime (τ_{HB}) as a molecular descriptor to quantify the strength of water binding to the reactant in the mixed-solvent systems.[72] We expect that hydrogen bonds between the reactant and water are stronger (*i.e.*, longer-lasting) in mixed-solvent environments that have large cosolvent concentrations, because water-water hydrogen bonds are unable to form, increasing the preference of reactant-water hydrogen bonds. Stronger interactions between water and reactant may translate to a lower transition state free energy and thus, an increase in reaction rate.[73,74]

Hydrogen bonding lifetimes for TBA and XYL are shown in Figure 3.8. When increasing the concentration of the organic cosolvent, hydrogen bonding lifetimes between the reactant and the surrounding water molecules increase monotonically across different cosolvent environments, indicating stronger reactant-water hydrogen bonds. XYL has a lower hydrogen bonding lifetime compared to TBA, possibly due to a higher reaction temperature for XYL (403 K vs. 363 K); higher system temperatures often result in lower hydrogen bonding lifetimes. To remove temperature effects, we define the hydrogen bonding lifetime ratio, τ, by normalizing the hydrogen bonding lifetime in the organic cosolvent mixtures ($\tau_{HB,\,org}$) by the hydrogen bonding lifetime of the same reactant in pure water (τ_{HB,H_2O}), as shown in Equation 3.6.

$$\tau = \frac{\tau_{HB,\,org}}{\tau_{HB,H_2O}} \tag{3.6}$$

A monotonic increase in hydrogen bonding lifetime ratio with respect to cosolvent fraction was observed for reactants across all cosolvent environments (Table S5 in the ESI),[2] confirming that reactant-water hydrogen bonding strength generally increases as the amount of available water in the mixture decreases.

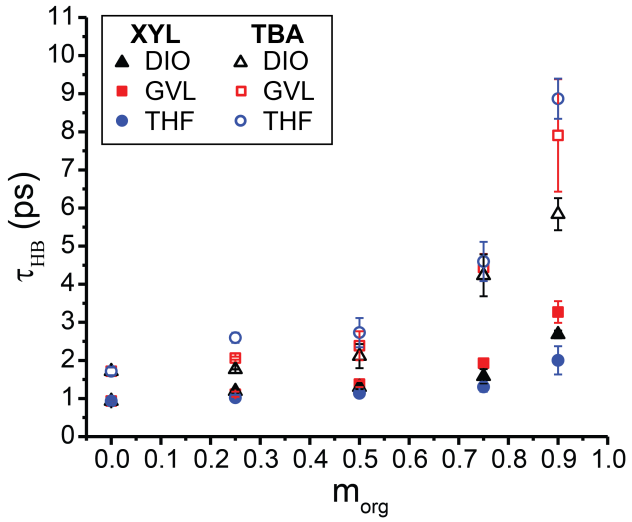


Figure 3.8: Average reactant-water hydrogen bond lifetimes (in picoseconds) for TBA and XYL as a function of mass fraction of organic cosolvent in DIOX-water, GVL-water, and THF-water mixtures.

3.3.6 Multidescriptor correlation between experimental and simulation results

Since both Γ and τ measure contributions to experimentally determined reaction rates, we explored the use of these descriptors in combination to improve the correlations in Table 3.2, as shown in Equation 3.7.

$$\sigma_{pred} = A + Bb + Cc + \ldots \tag{3.7}$$

A, B, and C are coefficients that quantify the relationship between the simulated molecular descriptors b and c and experimental data (σ). Coefficients for the multidescriptor correlation are obtained by linear regression between the simulated descriptors and experimental kinetic solvent parameters. The model accuracy is then assessed by using the coefficients to calculate a predicted kinetic solvent parameter, σ_{pred}, and plotting it against the experimentally determined kinetic solvent parameter σ_{exp}.

To determine if hydrogen bonding strength improves the correlation between simulation results and experimental models, we define two models: Equation 3.8 uses a single descriptor, Γ, and Equation 3.9 uses two descriptors, Γ and τ.

$$\sigma_{pred} = A + B(\Gamma) \tag{3.8}$$

$$\sigma_{pred} = A + B(\Gamma) + C(\tau) \tag{3.9}$$

As shown in Table 3.3, the slope between σ_{pred} and σ_{exp} for most reactant/-cosolvent systems is close to unity and the root-mean-square error (RMSE) between the predicted and experimental values is lowered when using the two-descriptor model in Equation 3.9 compared to only fitting a single descriptor in Equation 3.8. Therefore, the addition of the hydrogen bonding lifetime ratio, which characterizes the binding strength between the reactant and water within the local domain, improves the overall correlation between σ_{pred} and σ_{exp}.

Table 3.3: Best-fit slope and root-mean-square error (RMSE) between predicted kinetic solvent parameters (σ_{pred}) and experimental kinetic solvent parameters (σ_{exp}). N/A are omitted values since only three experimental values are available, resulting in an exact solution to Equation 3.9. THF cosolvent mixture results are omitted for the same reason.

		1 Descriptor Fit		2 Descriptor Fit	
		Equation 3.8		Equation 3.9	
Cosolvent	Reactant	Slope[a]	RMSE[b]	Slope[a]	RMSE[b]
DIOX	ETBE	0.01	0.11	0.17	0.10
	TBA	0.89	0.09	0.99	0.03
	LGA	0.70	0.11	0.98	0.03
	PDO	0.96	0.05	N/A	N/A
	FRU	0.92	0.16	1.00	0.03
	CEL	0.97	0.08	0.97	0.07
	XYL	0.98	0.09	0.98	0.08
GVL	ETBE	0.63	0.14	0.63	0.14
	TBA	0.97	0.03	0.98	0.02
	LGA	0.82	0.14	0.95	0.07
	PDO	0.99	0.07	N/A	N/A
	FRU	0.86	0.27	1.00	0.00
	CEL	0.64	0.35	0.97	0.11
	XYL	0.73	0.39	1.00	0.00

The results in Table 3.3 suggest that the two-descriptor correlation model can accurately reproduce reaction rates as a function of cosolvent concentration for a single reactant in a single solvent mixture. To probe if Γ and τ can be used to predict reaction performance for a series of various reactants, we calculated best-fit parameters for Equation 3.9 using the combined data for all seven reactants in DIOX-water mixtures and used

the resulting two-descriptor correlation model to calculate values of σ. Comparing the calculated and experimental values of σ resulted in a slope of 0.73 and RMSE of 0.36 (Figure S7 in the ESI).[2] We further found that incorporating a reactant-specific descriptor, the accessible hydroxyl fraction (δ), led to the improved multidescriptor correlation model described in Equation 3.10.

$$\sigma_{\text{pred}} = A + B(\Gamma) + C(\tau) + D(\delta) \tag{3.10}$$

The three molecular descriptors in Equation 3.10 are statistically uncorrelated (Figure S9 in the ESI)[2] with one another, and moreover are physically motivated vis-à-vis the generalized reaction mechanism proposed above. The preferential exclusion coefficient captures the extent to which the reactant accumulates excess water in its immediate vicinity, which creates a thermodynamic driving force for the transfer of a proton from the bulk phase to the reactant, thereby initiating the acid-catalyzed reaction mechanism. The hydrogen bonding lifetime ratio captures information regarding the binding strength of water to the reactant, which we interpret as a measure of the ability of the encapsulated water cluster to stabilize the carbocation-like transition states common to acid-catalyzed reactions, as shown in Figure 3.5. The accessible hydroxyl fraction captures information relating to the percentage of the surface area of the reactant molecule occupied by hydroxyl groups, which we interpret as a qualitative descriptor of the reactant's hydrogen bonding capacity, normalized by its molecular

size. To further validate the addition of each parameter, we compared several simpler models (see Table S8 in the ESI)[2] using the Akaike Information Criterion (AIC).[75] Comparison of the AIC's associated with each of the competing models indicated that Equation 3.10 affords a sufficient improvement in the model's ability to predict experimental rate constants so that its greater complexity is statistically justified.

Figure 3.9(a) compares kinetic solvent parameters calculated using the multidescriptor correlation model in Equation 3.10 to experimentally determined values for DIOX-water mixtures, affording a slope of 0.89 and RMSE of 0.23 with few false positive/negatives. As the correlation is maintained across several different reactants, we suggest that acid-catalyzed reactions behave similarly in each of these cosolvent mixtures, where knowledge of simulation-derived Γ, τ, and δ can predict experimental σ_{exp} with accuracy. To demonstrate the predictive power of the model in Equation (9), we selected one reactant (the test set), fit the parameters in Equation 3.10 using the remaining six reactants (the training set), and then calculated kinetic solvent parameters for the test set reactant using the parameters derived from the training set. This procedure assesses the ability of the multidescriptor correlation model to predict kinetic solvent parameters for reactants that are not used to determine model parameters. Figure 3.9(b) shows the results of this procedure for DIOX-water mixtures using FRU as the test set reactant. We find that the best-fit slope and RMSE between experimental and predicted kinetic solvent parameters are nearly the same

when using the six-reactant training set compared to using all seven reactants. The test set RMSE is 0.12, which is lower than the average RMSE of the training set. These results confirm the robustness and predictability of the multidescriptor correlation model. Table S9 further extends these results to all reactants in all solvent systems.[2]

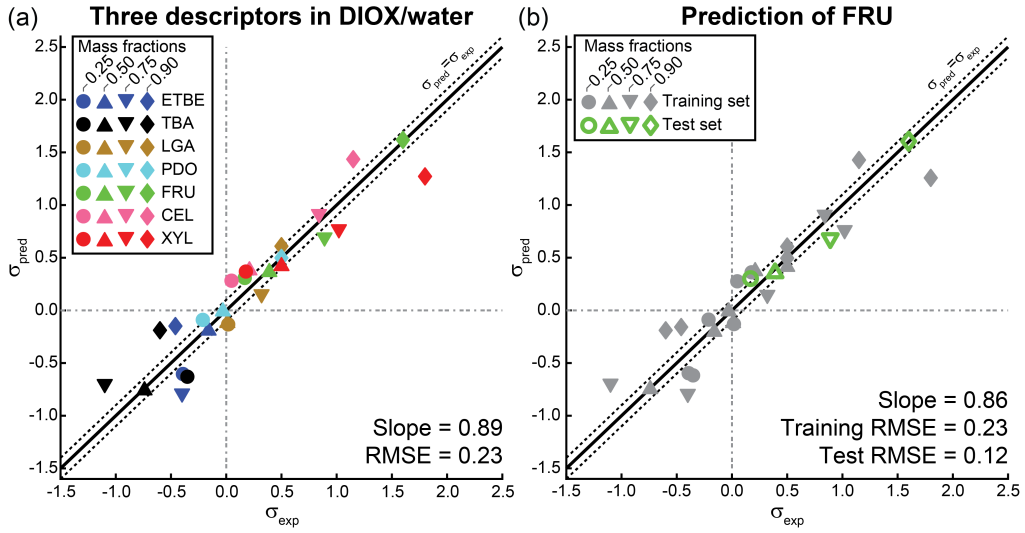


Figure 3.9: (a) Comparison of kinetic solvent parameters calculated using the multidescriptor correlation model (σ_{pred}) to experimentally determined values (σ_{exp}) for all seven reactants in DIOX-water mixtures. Each reactant has four data points for 0.25, 0.50, 0.75, and 0.90 mass fractions of DIOX, with the exception of PDO (see Table S1).[2] The slope of the best-fit line for all the data points and the average root-mean-squared error (RMSE) between the values of σ_{pred} and σ_{exp} are shown at the bottom right. The solid black line indicates a perfect correlation ($\sigma_{pred} = \sigma_{exp}$) and dotted lines are drawn at $\sigma_{exp} = 0$ and σ_{pred} to help visualize false positive/negative predicted values. Lines above and below the $\sigma_{pred} = \sigma_{exp}$ line are shifted by ± 0.10, denoting the approximate experimental error. (b) Prediction of kinetic solvent parameters using FRU as a test set with all other reactants taken as a training set. The slope of the best-fit line and RMSE between the values of σ_{pred} and σ_{exp} for the training and test sets are shown at the bottom right.

Table 3.4 summarizes the model coefficients and error for the different cosolvent environments. Simulated parameters were re-scaled to values between 0 and 1 so that the values of the coefficients in Equation 3.10 can be compared (see ESI for re-scaling details).[2] DIOX-water mixtures

were found to have the best correlation when fitting. GVL-water and THF-water cosolvent environments have the highest error. The coefficients differ across cosolvents, suggesting that the simulated parameters contribute to a different degree when changing cosolvent environments. In general, Γ has the largest weight in the multidescriptor correlation model for DIOX-water and GVL-water systems, indicating that the formation of a water-rich local domain is the dominant contributor to reactivity, in agreement with the correlations in Table 3.2. Γ has a smaller weight in THF-water system further supported by low correlations between Γ and σ in Table 3.2. The low correlation implies that there may be other descriptors that could be included to improve the model, such as additional cosolvent-specific descriptors (*e.g.*, dielectric constants, size, dipole moment, *etc.*).[76]

Table 3.4: Coefficients of the multidescriptor correlation model describing the rates of all seven reactions in each of the three cosolvent mixtures with the best-fit slope and root-mean-square error (RMSE) between predicted (σ_{pred}) and experimentally determined (σ_{exp}) kinetic solvent parameters. The descriptors $\widetilde{\Gamma}$, $\widetilde{\tau}$, and $\widetilde{\delta}$ are equal to Γ, τ, and δ but re-scaled to values between 0 and 1 so the coefficients are comparable.

	$\sigma_{pred} = A + B(\widetilde{\Gamma}) + C(\widetilde{\tau}) + D(\widetilde{\delta})$					
Cosolvent	A	B	C	D	Slope	RMSE
DIOX	-1.484	1.536	1.244	0.999	0.89	0.23
GVL	-1.416	1.696	0.760	1.018	0.71	0.36
THF	-0.826	0.349	1.592	0.410	0.51	0.59

3.4 Summary

We have analyzed the effects of three polar aprotic cosolvents, in aqueous mixtures of varying composition, on the acid-catalyzed reactions of seven biomass-derived model compounds. General trends in reactivity, as expressed by changes in apparent rate constants as a function of solvent composition, were correlated to simulation-derived observables obtained from classical molecular dynamics simulations. We find that the presence of organic cosolvents in the solvent mixture leads to the formation of water-enriched local solvent domains near hydrophilic reactants and increases the strength of hydrogen bonding between reactants and local water molecules. These effects are quantified by molecular descriptors describing: (1) the local density of water near the reactant (Γ); (2) the average hydrogen-bond lifetime between the reactant and neighboring water molecules (τ); and (3) the accessible surface area occupied by hydroxyl groups on the reactant (δ). By combining these three observables in a multiple linear regression scheme, we have developed a multidescriptor correlation model that predicts rate constants as a function of solvent composition. This development represents an important step toward the computational design of new liquid-phase biomass conversion processes, informed by a first principles approach.

3.5 References

[1] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628.

[2] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates [Supporting Information]. *Energy & Environmental Science* **2018**, *11*, 617–628.

[3] Chheda, J. N.; Huber, G. W.; Dumesic, J. A. Liquid-phase catalytic processing of biomass-derived oxygenated hydrocarbons to fuels and chemicals. *Angewandte Chemie International Edition* **2007**, *46*, 7164–7183.

[4] Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chemical reviews* **2007**, *107*, 2411–2502.

[5] Huber, G. W.; Iborra, S.; Corma, A. Synthesis of transportation fuels from biomass: chemistry, catalysts, and engineering. *Chemical reviews* **2006**, *106*, 4044–4098.

[6] Román-Leshkov, Y.; Barrett, C. J.; Liu, Z. Y.; Dumesic, J. A. Production of dimethylfuran for liquid fuels from biomass-derived carbohydrates. *Nature* **2007**, *447*, 982–985.

[7] Stöcker, M. Biofuels and biomass-to-liquid fuels in the biorefinery: Catalytic conversion of lignocellulosic biomass using porous materials. *Angewandte Chemie International Edition* **2008**, *47*, 9200–9211.

[8] Tock, L.; Gassner, M.; Maréchal, F. Thermochemical production of liquid fuels from biomass: Thermo-economic modeling, process design and process integration analysis. *Biomass and bioenergy* **2010**, *34*, 1838–1854.

[9] Simonetti, D. A.; Dumesic, J. A. Catalytic production of liquid fuels from biomass-derived oxygenated hydrocarbons: catalytic coupling at multiple length scales. *Catalysis Reviews* **2009**, *51*, 441–484.

[10] Nguyen, T. Y.; Cai, C. M.; Kumar, R.; Wyman, C. E. Overcoming factors limiting high-solids fermentation of lignocellulosic biomass to ethanol. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11673–11678.

[11] Mellmer, M. A.; Alonso, D. M.; Luterbacher, J. S.; Gallo, J. M. R.; Dumesic, J. A. Effects of γ-valerolactone in hydrolysis of lignocellulosic biomass to monosaccharides. *Green Chemistry* **2014**, *16*, 4659–4662.

[12] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[13] Román-Leshkov, Y.; Chheda, J. N.; Dumesic, J. A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose. *Science* **2006**, *312*, 1933–1937.

[14] Wei, Z.; Li, Y.; Thushara, D.; Liu, Y.; Ren, Q. Novel dehydration of carbohydrates to 5-hydroxymethylfurfural catalyzed by Ir and Au chlorides in ionic liquids. *Journal of the Taiwan Institute of Chemical Engineers* **2011**, *42*, 363–370.

[15] Mellmer, M. A.; Gallo, J. M. R.; Martin Alonso, D.; Dumesic, J. A. Selective production of levulinic acid from furfuryl alcohol in THF solvent systems over H-ZSM-5. *ACS Catalysis* **2015**, *5*, 3354–3359.

[16] Pagan-Torres, Y. J.; Wang, T.; Gallo, J. M. R.; Shanks, B. H.; Dumesic, J. A. Production of 5-hydroxymethylfurfural from glucose using a combination of Lewis and Brønsted acid catalysts in water in a biphasic reactor with an alkylphenol solvent. *Acs Catalysis* **2012**, *2*, 930–934.

[17] Zhang, T.; Kumar, R.; Wyman, C. E. Enhanced yields of furfural and other products by simultaneous solvent extraction during thermochemical treatment of cellulosic biomass. *Rsc Advances* **2013**, *3*, 9809–9819.

[18] Mostofian, B.; Cai, C. M.; Smith, M. D.; Petridis, L.; Cheng, X.; Wyman, C. E.; Smith, J. C. Local phase separation of co-solvents enhances

pretreatment of biomass for bioenergy applications. *Journal of the American Chemical Society* **2016**, *138*, 10869–10878.

[19] Nguyen, T. Y.; Cai, C. M.; Kumar, R.; Wyman, C. E. Co-solvent pretreatment reduces costly enzyme requirements for high sugar and ethanol yields from lignocellulosic biomass. *ChemSusChem* **2015**, *8*, 1716–1725.

[20] He, J.; Liu, M.; Huang, K.; Walker, T. W.; Maravelias, C. T.; Dumesic, J. A.; Huber, G. W. Production of levoglucosenone and 5-hydroxymethylfurfural from cellulose in polar aprotic solvent–water mixtures. *Green Chemistry* **2017**, *19*, 3642–3653.

[21] Vasudevan, V.; Mushrif, S. H. Insights into the solvation of glucose in water, dimethyl sulfoxide (DMSO), tetrahydrofuran (THF) and N, N-dimethylformamide (DMF) and its possible implications on the conversion of glucose to platform chemicals. *Rsc Advances* **2015**, *5*, 20756–20763.

[22] Cesarotti, E.; Ugo, R.; Kaplan, L. A discussion of the different kinds of solute-solute and solute-solvent interactions acting in homogeneous catalysis by transition metal complexes. *Coordination Chemistry Reviews* **1982**, *43*, 275–298.

[23] Dyson, P. J.; Jessop, P. G. Solvent effects in catalysis: rational improvements of catalysts via manipulation of solvent interactions. *Catalysis Science & Technology* **2016**, *6*, 3302–3316.

[24] Madon, R. J.; Iglesia, E. Catalytic reaction rates in thermodynamically non-ideal systems. *Journal of Molecular Catalysis A: Chemical* **2000**, *163*, 189–204.

[25] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[26] Behtash, S.; Lu, J.; Walker, E.; Mamun, O.; Heyden, A. Solvent effects in the liquid phase hydrodeoxygenation of methyl propionate over a Pd (1 1 1) catalyst model. *Journal of Catalysis* **2016**, *333*, 171–183.

[27] Caratzoulas, S.; Vlachos, D. G. Converting fructose to 5-hydroxymethylfurfural: a quantum mechanics/molecular mechanics

study of the mechanism and energetics. *Carbohydrate research* **2011**, *346*, 664–672.

[28] Murzin, D. Y. Solvent effects in catalysis: implementation for modelling of kinetics. *Catalysis Science & Technology* **2016**, *6*, 5700–5713.

[29] Zhang, J.; Das, A.; Assary, R. S.; Curtiss, L. A.; Weitz, E. A combined experimental and computational study of the mechanism of fructose dehydration to 5-hydroxymethylfurfural in dimethylsulfoxide using Amberlyst 70, PO43-/niobic acid, or sulfuric acid catalysts. *Applied Catalysis B: Environmental* **2016**, *181*, 874–887.

[30] Assary, R. S.; Kim, T.; Low, J. J.; Greeley, J.; Curtiss, L. A. Glucose and fructose to platform chemicals: understanding the thermodynamic landscapes of acid-catalysed reactions using high-level ab initio methods. *Physical Chemistry Chemical Physics* **2012**, *14*, 16603–16611.

[31] Choudhary, V.; Mushrif, S. H.; Ho, C.; Anderko, A.; Nikolakis, V.; Marinkovic, N. S.; Frenkel, A. I.; Sandler, S. I.; Vlachos, D. G. Insights into the interplay of Lewis and Brønsted acid catalysts in glucose and fructose conversion to 5-(hydroxymethyl) furfural and levulinic acid in aqueous media. *Journal of the American Chemical Society* **2013**, *135*, 3997–4006.

[32] Curioni, A.; Sprik, M.; Andreoni, W.; Schiffer, H.; Hutter, J.; Parrinello, M. Density functional theory-based molecular dynamics simulation of acid-catalyzed chemical reactions in liquid trioxane. *Journal of the American Chemical Society* **1997**, *119*, 7218–7229.

[33] Meijer, E. J.; Sprik, M. A density functional study of the addition of water to SO3 in the gas phase and in aqueous solution. *The Journal of Physical Chemistry A* **1998**, *102*, 2893–2898.

[34] Mushrif, S. H.; Varghese, J. J.; Krishnamurthy, C. B. Solvation dynamics and energetics of intramolecular hydride transfer reactions in biomass conversion. *Physical Chemistry Chemical Physics* **2015**, *17*, 4961–4969.

[35] Qian, X.; Wei, X. Glucose isomerization to fructose from ab initio molecular dynamics simulations. *The Journal of Physical Chemistry B* **2012**, *116*, 10898–10904.

[36] Mushrif, S. H.; Caratzoulas, S.; Vlachos, D. G. Understanding solvent effects in the selective conversion of fructose to 5-hydroxymethylfurfural: a molecular dynamics investigation. *Physical Chemistry Chemical Physics* **2012**, *14*, 2637–2644.

[37] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[38] Trummal, A.; Lipping, L.; Kaljurand, I.; Koppel, I. A.; Leito, I. Acidity of strong acids in water and dimethyl sulfoxide. *The Journal of Physical Chemistry A* **2016**, *120*, 3663–3669.

[39] Raamat, E.; Kaupmees, K.; Ovsjannikov, G.; Trummal, A.; Kütt, A.; Saame, J.; Koppel, I.; Kaljurand, I.; Lipping, L.; Rodima, T.; et al.. Acidities of strong neutral Brønsted acids in different media. *Journal of Physical Organic Chemistry* **2013**, *26*, 162–170.

[40] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[41] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[42] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[43] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[44] Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[45] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.

[46] Luzar, A.; Chandler, D. Hydrogen-bond kinetics in liquid water. *Nature* **1996**, *379*, 55–57.

[47] Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Computer physics communications* **1995**, *91*, 43–56.

[48] Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation* **2008**, *4*, 435–447.

[49] Lindahl, E.; Hess, B.; Van Der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual* **2001**, *7*, 306–317.

[50] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. GROMACS: fast, flexible, and free. *Journal of computational chemistry* **2005**, *26*, 1701–1718.

[51] van der Spoel, D.; van Maaren, P. J.; Larsson, P.; Tîmneanu, N. Thermodynamics of hydrogen bonding in hydrophilic and hydrophobic media. *The Journal of Physical Chemistry B* **2006**, *110*, 4393–4398.

[52] Sasaki, M.; Furukawa, M.; Minami, K.; Adschiri, T.; Arai, K. Kinetics and mechanism of cellobiose hydrolysis and retro-aldol condensation in subcritical and supercritical water. *Industrial & engineering chemistry research* **2002**, *41*, 6642–6649.

[53] Oltmanns, J. U.; Palkovits, S.; Palkovits, R. Kinetic investigation of sorbitol and xylitol dehydration catalyzed by silicotungstic acid in water. *Applied Catalysis A: General* **2013**, *456*, 168–173.

[54] O'Reilly, K. T.; Moir, M. E.; Taylor, C. D.; Smith, C. A.; Hyman, M. R. Hydrolysis of tert-butyl methyl ether (MTBE) in dilute aqueous acid. *Environmental science & technology* **2001**, *35*, 3954–3961.

[55] Swift, T. D.; Bagia, C.; Choudhary, V.; Peklaris, G.; Nikolakis, V.; Vlachos, D. G. Kinetics of homogeneous Brønsted acid catalyzed fructose dehydration and 5-hydroxymethyl furfural rehydration: a combined experimental and computational study. *Acs Catalysis* **2014**, *4*, 259–267.

[56] Akien, G. R.; Qi, L.; Horváth, I. T. Molecular mapping of the acid catalysed dehydration of fructose. *Chemical Communications* **2012**, *48*, 5850–5852.

[57] Bennet, A. J.; Sinnott, M. L. Complete kinetic isotope effect description of transition states for acid-catalyzed hydrolyses of methyl. alpha.- and. beta.-glucopyranosides. *Journal of the American Chemical Society* **1986**, *108*, 7287–7294.

[58] Dumesic, J. Analyses of reaction schemes using De Donder relations. *Journal of Catalysis* **1999**, *185*, 496–505.

[59] Dumesic, J. Reply to finding the rate-determining step in a mechanism: Comparing DeDonder relations with the" degree of rate control". *Journal of Catalysis* **2001**, *204*, 525–529.

[60] Lee, J. K.; Bain, A. D.; Berti, P. J. Probing the transition states of four glucoside hydrolyses with 13C kinetic isotope effects measured at natural abundance by NMR spectroscopy. *Journal of the American Chemical Society* **2004**, *126*, 3769–3776.

[61] Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Single-ion solvation free energies and the normal hydrogen electrode potential in methanol, acetonitrile, and dimethyl sulfoxide. *The Journal of Physical Chemistry B* **2007**, *111*, 408–422.

[62] Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. The proton's absolute aqueous enthalpy and Gibbs free energy of solvation from cluster-ion solvation data. *The Journal of Physical Chemistry A* **1998**, *102*, 7787–7794.

[63] Kalidas, C.; Hefter, G.; Marcus, Y. Gibbs energies of transfer of cations from water to mixed aqueous organic solvents. *Chemical reviews* **2000**, *100*, 819–852.

[64] Warshel, A. Energetics of enzyme catalysis. *Proceedings of the National Academy of Sciences* **1978**, *75*, 5250–5254.

[65] Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. Electrostatic basis for enzyme catalysis. *Chemical reviews* **2006**, *106*, 3210–3235.

[66] Bruice, T. C.; Lightstone, F. C. Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions. *Accounts of chemical research* **1999**, *32*, 127–136.

[67] Liu, Y.; Vjunov, A.; Shi, H.; Eckstein, S.; Camaioni, D. M.; Mei, D.; Baráth, E.; Lercher, J. A. Enhancing the catalytic activity of hydronium ions through constrained environments. *Nature communications* **2017**, *8*, 1–8.

[68] Kang, M.; Smith, P. E. Preferential interaction parameters in biological systems by Kirkwood–Buff theory and computer simulation. *Fluid Phase Equilibria* **2007**, *256*, 14–19.

[69] Schneider, C. P.; Trout, B. L. Investigation of cosolute- protein preferential interaction coefficients: New insight into the mechanism by which arginine inhibits aggregation. *The Journal of Physical Chemistry B* **2009**, *113*, 2050–2058.

[70] Shukla, D.; Trout, B. L. Preferential interaction coefficients of proteins in aqueous arginine solutions and their molecular origins. *The Journal of Physical Chemistry B* **2011**, *115*, 1243–1253.

[71] Shulgin, I. L.; Ruckenstein, E. Local composition in the vicinity of a protein molecule in an aqueous mixed solvent. *The Journal of Physical Chemistry B* **2007**, *111*, 3990–3998.

[72] Smolin, N.; Winter, R. Effect of temperature, pressure, and cosolvents on structural and dynamic properties of the hydration shell of SNase: A molecular dynamics computer simulation study. *The Journal of Physical Chemistry B* **2008**, *112*, 997–1006.

[73] Sutton, J. E.; Vlachos, D. G. A theoretical and computational analysis of linear free energy relations for the estimation of activation energies. *ACS Catalysis* **2012**, *2*, 1624–1634.

[74] Wang, S.; Petzold, V.; Tripkovic, V.; Kleis, J.; Howalt, J. G.; Skulason, E.; Fernández, E.; Hvolbæk, B.; Jones, G.; Toftelund, A.; et al.. Universal transition state scaling relations for (de) hydrogenation over transition metals. *Physical Chemistry Chemical Physics* **2011**, *13*, 20760–20765.

[75] Akaike, H. A new look at the statistical model identifications. *IEEE transactions on automatic control* **1974**, *19*, 716–723.

[76] Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters,. pi.*,. alpha., and. beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry* **1983**, *48*, 2877–2887.

4 QUANTIFYING THE STABILITY OF THE HYDRONIUM ION IN ORGANIC SOLVENTS WITH MOLECULAR DYNAMICS SIMULATIONS

The previous chapter provides a molecular simulation framework to predict experimental reaction rates using molecular descriptors derived from simulations of a single reactant in mixed-solvent environments. However, this model does not incorporate any information about the acid catalyst or the product state. Hence, this chapter focuses on incorporating acid catalyst information by modeling a hydronium ion in different solvent environments. This chapter aims to tackle the following questions:

- How does the stability of an acid catalyst change when we vary the solvent environment?
- What physical insights can we gain about the acid catalyst that might inform on solvent selection?
- How do we incorporate catalyst information within the descriptor framework developed in Chapter 3?

In this chapter, classical molecular dynamics simulations were performed to quantify the stability of hydronium and chloride ions by measuring their solvation free energies in water, 1,4-dioxane (DIOX), tetrahy-

This chapter was reproduced from Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations. *Frontiers in chemistry* **2019**, *7*, 439, with permission from Frontiers under the Creative Commons Attribution ("CC BY") license.[1] The supplementary material is cited as Ref. 2.

drofuran (THF), γ-valerolactone (GVL), N-methyl-2-pyrrolidone (NMP), acetone (ACE), and dimethyl sulfoxide (DMSO). By measuring the free energy for transferring a hydronium ion from pure water to pure organic solvent, we found that the hydronium ion is destabilized in DIOX, THF, and GVL and stabilized in NMP, ACE, and DMSO relative to water. The distinction between these organic solvents can be used to predict which phase the hydronium ion prefers in aqueous mixtures of organic solvents. We then incorporated the stability of the hydronium ion into a predictive model for the acid-catalyzed conversion of 1,2-propanediol to propanal. The revised model is able to predict experimental reaction rates across solvent systems with different organic solvents. These results demonstrate the ability of classical molecular dynamics simulations to screen solvent systems for improved acid-catalyzed reaction performance.

4.1 Introduction

Solution-phase biomass conversion reactions ubiquitously require an acidic proton catalyst (H^+), which exists in solution as a hydronium ion (H_3O^+). In homogenous reactions, the catalyst is obtained from the addition of a Brønsted acid[3] and recognized to follow a specific catalysis mechanism since a protonated solvent is the catalyst. Figure 4.1A shows an example reaction for the acid-catalyzed dehydration of 1,2-propanediol to propanal, which is representative of acid-catalyzed reactions for biomass-

derived model compounds.[4,5] In these reactions, the hydronium ion catalyst (H_3O^+) protonates the reactant (R) to form a reactant/proton complex (RH^+). The reaction proceeds to a charged transition state ($[RH^+]^{TS}$) and subsequently forms the product (P) with the hydronium ion reformed (Figure 4.1(b)). The relative stabilities of the reactant, transition state, and catalyst in solution are thus critical for determining reaction kinetics.[6] Understanding how these solvent effects influence reaction kinetics is necessary to guide the optimization of solvent compositions and reactor conditions and maximize the productivity of biomass conversion reactions.

Previous studies have found that mixtures of water and organic, polar aprotic cosolvents (*i.e.* mixed-solvent environments) can increase or decrease the rates of Brønsted acid catalyzed reactions depending on the stability of the acid catalyst.[4,6–8] One mechanism by which the solvent composition affects catalyst stability is by shifting the acid dissociation equilibrium, which is quantified by the acid disassociation constant (K_a). For example, weak acids with small K_a values, such as formic acid or acetic acid, were found to decrease acid-catalyzed reaction rates in mixed-solvent environments compared to pure water due to the reduced availability of catalytic hydronium ions.[7] Conversely, strong acids with large K_a values, such as triflic acid, dissociate in a small fraction of water and were found to improve xylose conversion reaction rates by 40 fold in 90 wt% γ-valerolactone,[7] suggesting an alternative role for the solvent. Based on

combined classical and ab initio molecular dynamics (MD) simulations, Mellmer *et al.* found that the hydronium ion catalyst in mixed-solvent systems is destabilized in the bulk solvent relative to the local solvent domain of the reactant due to unfavorable interactions between the hydronium ion and the cosolvent.[4] As a result, the acid catalyst is thermodynamically driven to water-enriched local solvent domains formed by hydrophilic reactants when a high mass fraction of the organic phase is present, effectively lowering activation energy barriers and increasing reaction rates relative to pure water (Figure 4.1(b)). Together, these results indicate that the solvent composition can modulate reaction kinetics by both modulating catalyst availability and the interactions of the catalyst with the reactant.

Building upon these studies, we hypothesized that acid-catalyzed reaction rates correlate with the formation of water-enriched local solvent domains because the catalyst is assumed to be stabilized by interactions with water and thus the formation of water-enriched local solvent domains would drive the partitioning of the catalyst to the reactant. We derived a correlative model that used descriptors derived from classical MD to predict experimental reaction kinetics for seven biomass-derived model compounds in aqueous mixtures of dioxane and γ-valerolactone.[4] While reaction free energies are typically determined from ab initio level studies,[4,9] this hypothesis allowed us to use classical MD simulations to more rapidly screen through multiple solvent compositions and reactions.

However, the assumption that the hydronium ion preferentially interacts with water is not always true. For instance, more basic organic solvents, such as dimethyl sulfoxide (DMSO), have been shown to participate in the reaction mechanism by stabilizing the proton.[4] The addition of DMSO has also been shown to diminish acid-catalyzed conversion of tert-butanol,[4] indicating that addition of organic solvents can also destabilize the reactant/proton complex, raise energy barriers, and consequently slower reaction kinetics relative to pure water (Figure 4.1(b)). Therefore, understanding and quantifying the thermodynamic stability of the hydronium ion in mixed-solvent environments is essential to accurate predictions of acid-catalyzed reaction kinetics.

It is experimentally difficult to directly measure the free energy of an isolated hydronium ion in solution since electroneutrality must be maintained.[10] To obtain single-ion thermodynamics, nonclassical techniques such as atomic and molecular spectroscopy combined with statistical mechanics are utilized.[11] To broaden the range of possible systems, computational tools have been developed to model the hydronium ion and isolate the role of the solvent on the acid catalyst.[12] For solution-phase reactions, we quantify the stability of the acid catalyst in terms of its solvation free energy, or the free energy for introducing the catalyst in solution. The solvation free energy accounts for interactions between the catalyst and solvent (*e.g.* hydrogen bonding, ion-dipole interactions, and van der Waals forces) and the solvent reorganization necessary to accommodate

the catalyst. Solvation free energies are also important in determining the partitioning of ions between different phases.[13] Typically, ab initio level simulations are performed to accurately compute solvation free energies of the acid catalyst.[14–18] However, these simulations are computationally expensive and thus challenging to perform for multiple solvent compositions. Recently, Bonthuis *et al.* developed a classical hydronium ion model that accurately reproduces experimental solvation free energies in pure water.[19] We thus hypothesize that this classical hydronium ion model can be used to compute solvation free energies of the acid catalyst and leverage the computational efficiency of MD simulations to screen stability in different solvent compositions, assuming that the hydronium ion maintains its structure in these solvents. These calculations can then be used to predict the relationship between solvent composition and reaction kinetics for acid-catalyzed reactions. Since our previous work found favorable agreement between MD simulation-derived descriptors with experimental reaction rates without mechanistic details of the reaction,[5] we are focused on studying how water-enrichment (or cosolvent-enrichment) can improve reaction performance by favorably facilitating a hydronium ion.

Herein, we use classical MD simulations to study the stability of a hydronium ion in six organic polar aprotic cosolvents: dioxane (DIOX), tetrahydrofuran (THF), γ-valerolactone (GVL), N-methyl pyrrolidine (NMP), acetone (ACE), and dimethyl sulfoxide (DMSO). We also study the stability

of a chloride ion in the same solvents to calculate the effect of the conjugate base. We use previous literature values for the reaction rates of the acid-catalyzed conversion of 1,2-propandiol (PDO) as a model reaction to study the influence of the different cosolvents. We then quantify the stabilities of the hydronium and chloride ions in pure and mixed-solvent environments by computing the solvation free energies. We find that the free energy for transferring a hydronium ion from pure water to organic solvent can distinguish between solvents that favorably (NMP, ACE, DMSO) and unfavorably (DIOX, THF, GVL) solvate the acid catalyst. With this information, we improve our previously developed predictive model for the conversion of PDO[5] by including a cosolvent-specific descriptor that incorporates information about the stability of the hydronium ion in the solvent system.

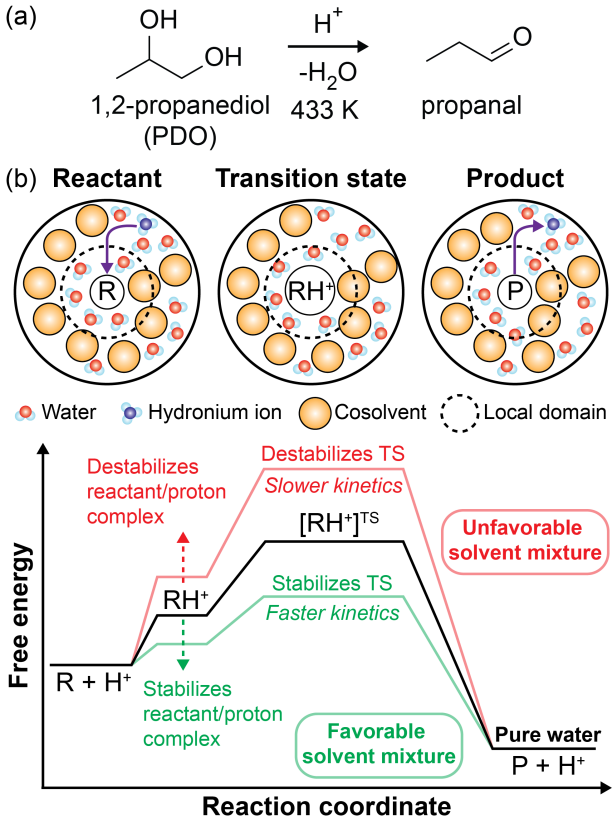


Figure 4.1: (a) Brønsted acid-catalyzed reaction of 1,2-propanediol (PDO) to propanal. (b) Schematic of acid-catalyzed reactions in mixed-solvent environments. The reaction proceeds through a charged transition state (TS) formed from the protonation of the reactant by a hydronium ion catalyst. The corresponding free energy diagram schematically depicts the influence of mixed-solvent environments (red and green lines) on acid-catalyzed reactions relative to pure water (black line). Note that this is a generalized representation of a free energy landscape based on prior computational findings[4] and is not specific to 1,2-propanediol dehydration reaction.

4.2 Methods

Classical MD simulations were performed using GROMACS 2016.[20] We used the classical hydronium and chloride ion models parameterized by Bonthius *et al.*,[19] which have been found to reproduce experimental solvation free energies in pure water systems modeled using the Single Point Charge/Extended (SPC/E) water model.[21] Bond constraints for the hydronium ion were modified to improve simulation performance by using the more efficient LINCS constraint algorithm[22] instead of the SHAKE constraint algorithm[23] (SI, Section 1).[2] PDO and all cosolvents were parameterized using the CGenFF/CHARMM36 forcefields,[24–26] while water was modeled using the SPC/E model.[21] For all simulations, Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones (LJ) potential with a 1.2 nm cutoff that was smoothly shifted to zero between 1.0 nm and 1.2 nm. Electrostatic interactions were calculated using the smooth Particle Mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and 4th order interpolation. Bonds were constrained using the LINCS algorithm.[22] All thermostats used a 1.0 ps time constant and all barostats used a 5.0 ps time constant with an isothermal compressibility of 5.0×10^{-5} bar^{-1}.

We initialized simulation configurations using the protocol schematically depicted in Figure 4.2(a). The initial simulation box containing water and cosolvent (if applicable) had dimensions of (6 nm)^3 in all simulations

and was equilibrated in a *NPT* simulation for 5 ns at $T = 300$ K and $P = 1$ bar with a velocity-rescale thermostat and Berendsen barostat. A single reactant or ion molecule (designated as "M" in Figure 4.2(a)) was then added to the system and equilibrated with the same barostat and thermostat for 500 ps. *NPT* production simulations were performed for all systems for 200 ns with a Parrinello-Rahman barostat and Nóse-Hoover thermostat; simulations of the reactant, PDO, were performed at $T = 433.15$ K to match the experimental reaction temperature[4] while simulations of the hydronium/chloride ions were performed at $T = 300$ K. Simulation configurations were output every 10 ps and the final 190 ns of each production trajectory were used for analysis. Simulation analysis was performed using the MDTraj library[27] and analysis tools developed in house. MD simulations were performed using a leapfrog integrator with a 2-fs time step. Figure 4.2(b) shows simulation snapshots of the nearby solvent environment around a hydronium ion in 90 wt% DIOX, 90 wt% DMSO, and pure water.

Each solvation free energy was computed from a series of stochastic dynamics simulations (Figure 4.2(a)). Simulations were initialized using an equilibrated solvent system (as described above) with a hydronium or chloride ion added to the system. The total potential energy of the system for solvation free energy calculations are described in Section 2.1.2. All free energy simulations used a soft-core LJ potential as described in the SI, Section 2.1.[2,28] For each simulation, the system was energy minimized with the steepest descent algorithm and equilibrated with a 100 ps *NVT*

simulation followed by a 2 ns *NPT* simulation with the Berendsen barostat. An 11 ns *NPT* production simulation was then performed with the Parrinello-Rahman barostat. All simulations were performed at $T = 300$ K and $P = 1$ bar. Energy differences computed between all pairs of windows were collected every 0.2 ps and solvation free energies were computed with the Multistate Bennett Acceptance Ratio[29] method using the python alchemical-analysis tool.[30] The 11 ns of each *NPT* production simulation was split into two 5.5 ns trajectories and treated as two independent trials. All solvation free energy results and error bars are reported as the average and standard deviation of the two trials, respectively. We further calculated three analytical correction terms to account for:

1. finite-size effects due to system interactions with periodic images,

2. the compression free energy for transferring an ion from a 1 atm ideal gas phase to 1 mol/L ideal solution, and

3. the electrostatic energy required to pass through an interfacial potential when the ion transfers from vacuum to bulk solution. These correction terms are included to account for differences between simulation and experiments as described in the SI, Section 2.2.[2]

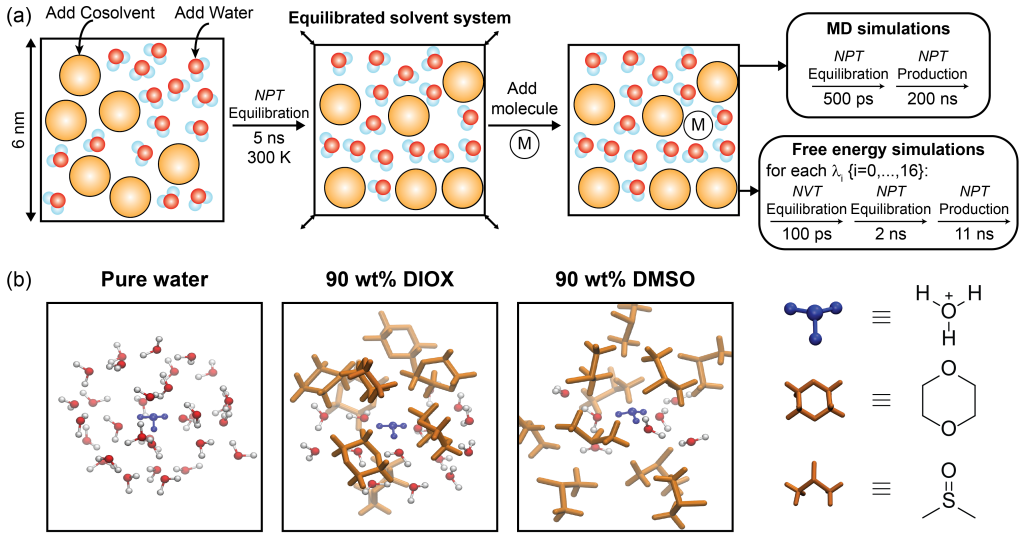


Figure 4.2: (a) Schematic representation of simulation workflow for molecular dynamics and free energy simulations. M denotes either 1,2-propanediol, a hydronium ion, or a chloride ion. (b) Simulation snapshots of hydronium ion in pure water, 90 wt% DIOX, and 90 wt% DMSO. The hydronium ion is located at the center and only solvent molecules within a 5 Å radius is shown.

4.3 Results

4.3.1 Comparison between experimental reaction rates and preferential exclusion coefficient

In our previous study of solution-phase acid-catalyzed reactions,[5] we hypothesized that the transition state is lower in free energy relative to the initial reactant state in mixed-solvent environments due to two reasons: (1) the catalyst is destabilized in bulk solvent relative to a water-enriched local

domain near the reactant, leading to a thermodynamic driving force for the transfer of catalytic protons to the local domain, and (2) the transition state is stabilized by water confined within this domain. We then developed a predictive model for experimental reaction rates by quantifying water enrichment in the vicinity of the reactant, supporting the hypothesis for aqueous mixtures of DIOX and GVL. This hypothesis assumes that the hydronium ion catalyst has a higher affinity for water than the cosolvent. However, this assumption may not be accurate for more basic cosolvents, such as DMSO, which can favorably stabilize the acid catalyst in bulk solution.[4] We thus test the validity of this assumption by determining if experimental reaction rates correlate with water enrichment for a model reaction, the Brønsted acid-catalyzed conversion of 1,2 propanediol (PDO) to propanal (Figure 4.1(a)), in DIOX and DMSO mixed-solvent environments. These cosolvents represent extremes in polarity: the dielectric constant of DIOX is 2.20, whereas the dielectric constant of DMSO is 48.90.[31]

We first analyze the solvent environment around PDO by calculating the radial distribution function (RDF). The RDF quantifies the solvent density, normalized by the bulk solvent density, at a distance r away from a central point. Figure 4.3 shows the RDF between the center of mass of PDO and water for 90 wt% DIOX, 90 wt% DMSO, and pure water. In 90 wt% DIOX, the peak of the RDF is significantly higher than in pure water, indicating that water preferentially partitions to the local solvent domain around PDO in high concentrations of DIOX. Conversely, in 90 wt%

DMSO, the first peak of the RDF is almost the same as in pure water and the RDF then drops below unity at 0.60 nm, indicating the local depletion of water. The diminished water content near PDO in aqueous mixtures of DMSO is due to the cosolvent's high affinity for oxygen groups, resulting in a competition between water and DMSO for the hydroxyl groups of PDO.[32] These findings confirm that DMSO and DIOX significantly influence the extent to which the reactant preferentially recruits water to the local domain.

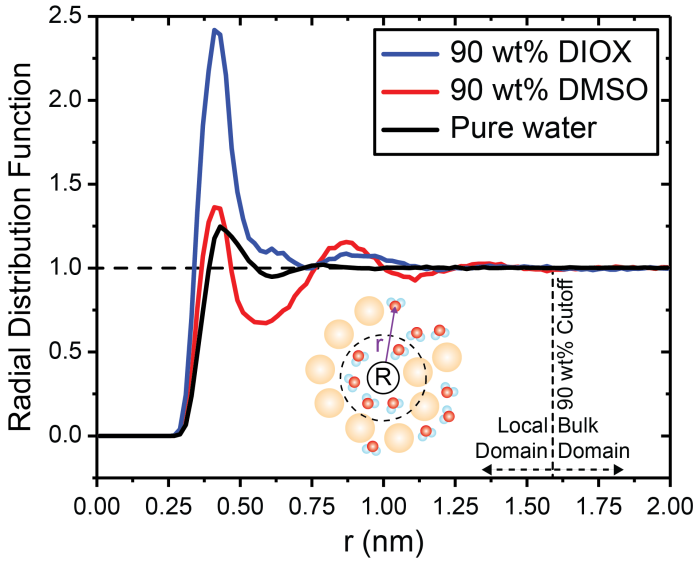


Figure 4.3: Radial distribution function between the center of mass of PDO and water in 90 wt% DIOX, 90 wt% DMSO, and pure water. Local and bulk domain cutoffs were determined as the value of r for which the RDF reaches unity. Bin widths for the RDFs were set to 0.02 nm.

Since RDFs are difficult to compare across different cosolvent concentra-

tions, we previously computed the preferential exclusion coefficient (Γ), a molecular descriptor that quantifies the local domain composition around the reactant.[5] Γ is defined as the excess number of cosolvent molecules within the local solvent domain of the reactant relative to the bulk solvent domain, computed using Equation 3.5.[33–36]

We define the boundary between local and bulk solvent domains as the value of r at which the RDF reaches unity (Figure 4.3), which occurs at $r = 1.59$ nm for both solvent systems. Positive Γ values indicate lower concentrations of cosolvent in the local solvent domain of the reactant compared to the bulk solvent domain. Therefore, positive Γ indicates the reactant has a higher affinity for water. Conversely, negative values of Γ indicate that the reactant has a higher affinity for the cosolvent.

We previously found that simulation-derived Γ and correlates with experimental reaction rates quantified by the kinetic solvent parameter (σ) shown in Equation 3.4.[5] For simplicity, we denote $\sigma^{i}_{org,j}$ as σ. Positive σ values indicate that the reaction occurs more favorably in aqueous mixtures with organic solvents compared to pure water. Negative σ values indicate the converse. We take experimental reaction rates from Ref. 4, which are tabulated in the SI, Section 3.[2]

Figure 4.4(a) compares values of simulation-derived Γ (filled lines) and experimentally measured[4] σ (dashed lines) for aqueous mixtures of DIOX and DMSO for the PDO dehydration reaction. In each separate mixed-solvent environment, Γ and σ are correlated across the solvent

composition range as shown in Figure 4.4(b). We report the Pearson correlation coefficient (Pearson's r) as an indicator of linear correlation: values close to 1 indicate total positive linear correlation, values close to -1 indicate total negative linear correlation, and values close to 0 indicate no linear correlation. We find $r = 0.97$ for aqueous mixtures of DIOX, indicating strong positive linear correlation. However, we find $r = 0.97$ for aqueous mixtures of DMSO, indicating strong negative correlation and that the depletion of water around PDO in DMSO mixtures still leads to enhanced reaction kinetics. These results indicate that in either solvent system the preferential exclusion coefficient can predict reaction kinetics; however, the negative correlation between Γ and σ in DMSO suggests that increased reaction rates are not due to water enrichment. This finding suggests that the assumption that the acid catalyst preferentially partitions to water-enriched regions of the system is not valid for all cosolvents and must be revised to derive a correlative model for reaction rates that can be broadly applied to any cosolvent of interest.

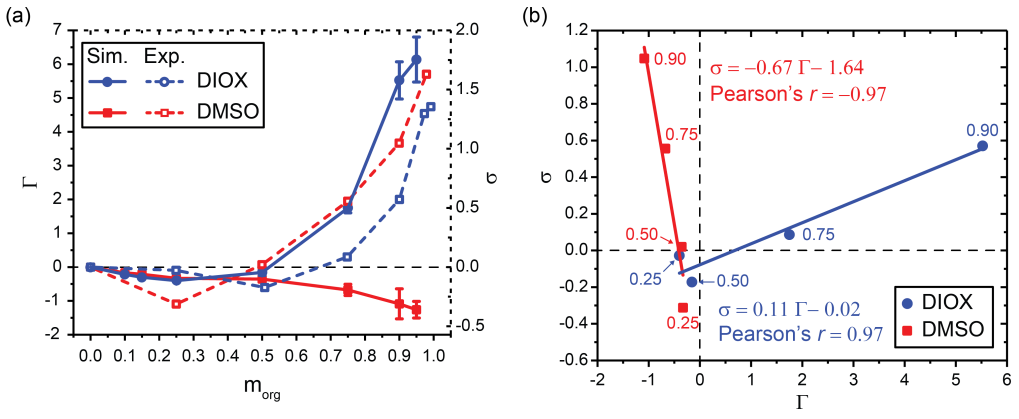


Figure 4.4: (a) Relationship between simulated preferential exclusion coefficient (Γ) and experimentally determined kinetic solvent parameters (σ) for aqueous mixtures of DIOX and DMSO. Experimental values were taken from Ref. 4. (b) Correlation between Γ and σ for aqueous mixtures of DIOX and DMSO. Data points are labeled with the wt% of the organic solvent. 25, 50, 75, and 90 wt% organic solvent was used to correlate Γ and σ as indicated for each point. The best-fit line is drawn and labeled with the corresponding equation and Pearson's r.

4.3.2 Solvation free energy of the hydronium ion in pure solvent systems

Previous studies have found that the hydronium ion is more stable in DMSO than water based on lower solvation free energies.[18] Since our simulations show that PDO preferentially interacts with DMSO rather than water and PDO reaction rates are increased in high concentrations of DMSO, the reactant/proton complex may be stabilized in the organic phase compared to the water phase, leading to increased reaction rates.

Therefore, we hypothesize that the solvation free energy of the hydronium ion catalyst in the organic solvent can be used to classify the preference for the catalyst for either water or organic phase and develop an updated correlative model between Γ and σ for a range of cosolvents.

We calculated the solvation free energy of the hydronium ion in six organic, polar aprotic solvents (Figure 4.5(a)) and performed the same calculations for a chloride ion to determine the solvation free energy for a conjugate base. We selected polar aprotic solvents due to their relevance to acid-catalyzed biomass conversion processes, where inclusion of these solvents has been found to enhance reaction performance.[4,5,7] To test the simulation approach, we first calculated solvation free energies in pure water as -465.1 kJ/mol (experimentally measured as -453.2 kJ/mol)[37] for the hydronium ion and 286.4 kJ/mol (experimentally measured as -304.6 kJ/mol)[38] for the chloride ion; their sum of -751.5 kJ/mol is comparable to the estimated experimental value of -757.8 kJ/mol.[37,38] The experimental values reproduced from Ref. 37 and 38 are modified to include the 7.9 kJ/mol correction term associated with transferring an ion from 1 atm ideal gas phase to 1 mol/L ideal solution to compare with our results (SI, Section 2.2).[2] The relative differences between solvation free energies are more important than absolute values[39] for inferring the behavior of the ions in different solvents; therefore, we focus on relative transfer free energies between pure water and organic solvents.

We computed the free energy of transferring the hydronium or chloride

ion from water to solvent systems with organic solvents using Equation 4.1 (schematically illustrated in Figure 4.5(b)):

$$\Delta G_j^{H_2O \to k} = \Delta G_j^k - \Delta G_j^{H_2O} \tag{4.1}$$

where k is the solvent system of interest and j is either hydronium or chloride ion. A negative value of $\Delta G_j^{H_2O \to k}$ indicates that the ion is thermodynamically stabilized in the kth solvent system compared to pure water. Figure 4.5(c) shows $\Delta G_j^{H_2O \to k}$ of the hydronium and chloride ions in each pure solvent. For the hydronium ion (cyan bars), $\Delta G_{H_3O^+}^{H_2O \to k}$ is positive for DIOX, GVL, and THF, indicating that the hydronium ion is unfavorable in these solvents. These results support our prior assumption that the hydronium ion prefers water rather than the organic phase in these solvents, allowing us to correlate the formation of water-enriched local domains to reaction kinetics.[5] Conversely, $\Delta G_{H_3O^+}^{H_2O \to k}$ is negative for NMP, ACE and DMSO, indicating that the hydronium ion is favorable in these solvents. Notably, these solvents are more basic than water[40] based on several solvent scales (*e.g.* B parameter of Koppel and Palm[41] or Kamlet-Taft β scale)[40,42] (discussed below and in Table 4.1). The negative free energy for transferring a hydronium ion from water to DMSO agrees with prior results[4,18] and supports the hypothesis that the sign of this free energy change determines the relationship between Γ and σ for DIOX and DMSO mixtures (Figure 4.4). In a similar fashion, we suspect that ACE

and DMSO would exhibit similar solvent effects.

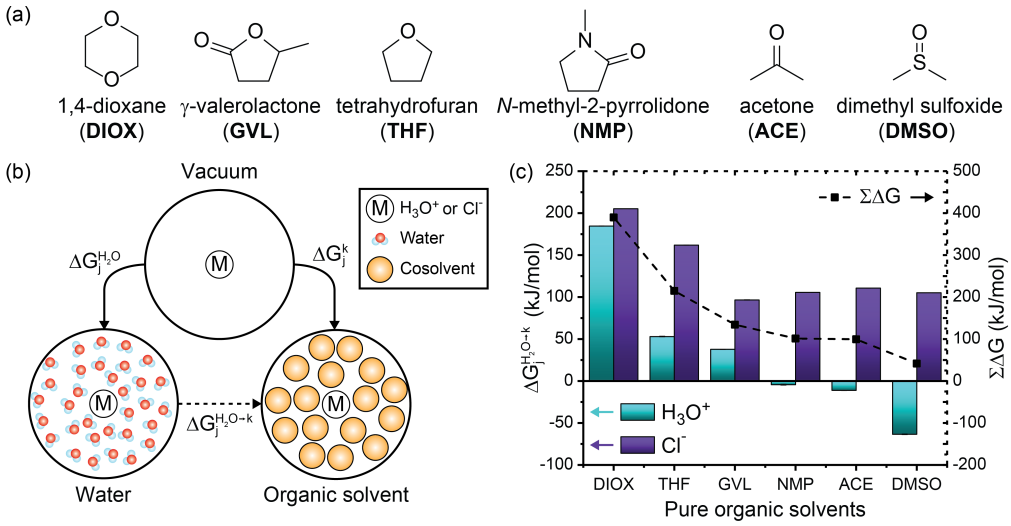


Figure 4.5: (a) Chemical structures of the organic solvents used in this study. (b) Thermodynamic cycle used to compute the free energy for transferring a hydronium or chloride ion from pure water to pure organic solvent. ΔG_j^k and $\Delta G_j^{H_2O}$ are solvation free energies while $\Delta G_i^{H_2O \to k}$ is the transfer free energy computed from Equation 4.1. (c) Transfer free energies for six pure organic solvents. Cyan and purple bars indicate hydronium (H_3O^+) and chloride (Cl^-) ion transfer free energies, respectively. Dashed lines indicate the sum of the transfer energies. Error bars were computed from the standard deviation of two trials; the error is less than 1 kJ/mol and is not visible in the plot. The error is tabulated in Supplementary Table 4.[2]

4.3.3 Solvation free energy of the chloride ion in pure solvent systems

While the solvation free energy of the hydronium ion alone quantifies catalyst stability, the effect of the solvent on acid dissociation equilibrium

depends on the solvation free energy of the hydronium ion and its conjugate base (*i.e.*, the chloride ion). Figure 4.5(c) shows that $\Delta G_{Cl^-}^{H_2O\rightarrow k}$ is positive for all pure organic solvent systems (purple bars), indicating that the chloride ion thermodynamically prefers water over each of these solvents. Furthermore, the solvation free energies for the chloride ion do not vary significantly for the different organic solvents, with the exception of DIOX and THF. The difference in the solvation of the hydronium and chloride ions is likely due to differences in hydrogen bonding capabilities: the hydronium ion can donate and accept hydrogen bonds while the chloride ion can only accept hydrogen bonds. Since water is the only solvent in this study that can donate hydrogen bonds, it is expected that the chloride ion is most stable in water.

The effect of the solvent on the solvation free energies of the hydronium and chloride ions relative to their solvation free energies in water is quantified via the term $\sum \Delta G$, which we define in Equation 4.2.

$$\sum \Delta G = \sum \Delta G_{H_3O^+Cl^-}^{k\rightarrow H_2O} = \Delta G_{H_3O^+}^{H_2O\rightarrow k} + \Delta G_{Cl^-}^{H_2O\rightarrow k} \tag{4.2}$$

We expect that positive values of $\sum \Delta G$ would reduce acid dissociation relative to pure water due to the decreased stability of the dissociated ions in the pure organic solvent. Figure 4.5(c) shows that $\sum \Delta G$ is positive for each organic solvent (black dashed lines). This result suggests that all of the polar aprotic solvents would decrease acid dissociation, leading

to the reduced catalyst availability associated with weak acids based on experiments.[7] The sign of $\sum \Delta G$ is largely dictated by the solvation free energies of the chloride ion, indicating that the selection of the conjugate base is important for acid disassociation,[7] although the choice of conjugate base would not affect the solvation free energies of the hydronium ion itself.

4.3.4 Relationship between solvation free energies and solvent parameters

Given the computational expense of free energy calculations, we next sought to relate the transfer free energy results ($\Delta G_{H_3O^+}^{k\rightarrow H_2O}$ and $\sum \Delta G$) to tabulated solvent properties to determine if these properties could accelerate solvent screening. Table 4.1 compares transfer free energy values to solvent dielectric constants and Kamlet-Taft parameters (α, β, π^*). We use the dielectric constant to quantify the polarizability of the solvents and the Kamlet-Taft parameters to quantify hydrogen-bond donating ability (acidity, α), hydrogen-bond accepting ability (basicity, β), and polarity/polarizability (π^*).[42–45] Each of the Kamlet-Taft parameters are scaled from 0 to 1 based on two reference solvents. For instance, π^* uses cyclohexane and DMSO as a reference for 0 and 1, respectively.[43–46] We expect that the stability of a hydronium ion can be influenced by the polarity of the solvent; however, neither dielectric constant nor π^* quantitatively correlate with

$\Delta G^{k\rightarrow H_2O}_{H_3O^+}$ or $\sum \Delta G$. Furthermore, basicity is expected to be an important metric of whether a hydronium ion is favored in a solvent environment, where larger β values indicate a more basic solvent that would favorably solvate the acidic hydronium ion, but there is no clear correlation between β and the free energy results. We also do not find a correlation between α and the free energy results, which is expected since acidity does not directly relate to the stability of a hydronium ion in a solvent system.

These data suggest that typical solvent-specific parameters cannot easily describe the interplay of solute-solvent interactions and solvent reorganization that dictate the measured transfer free energies. We further computed the RDF between the hydronium ion and pure solvents (Supplementary Figure 4.7)[2] to determine if solvent structure correlated with the transfer free energies, but we do not find a clear trend to explain the results found in Figure 4.5(c). This data thus suggests that the free energy calculations are providing new information that can be used to predict the preference of the hydronium ion for either water or an organic solvent and quantify the effect of solvent composition on acid dissociation. It is also possible that the MD workflow is insufficiently accurate to predict these values, particularly given the classical model of the hydronium ion. However, the good agreement between the calculated solvation free energies of the hydronium and chloride ions in water with experimental data suggests that the model is reasonable. We also emphasize that DIOX, THF, and GVL have positive values of $\Delta G^{k\rightarrow H_2O}_{H_3O^+}$ and lead to increased reaction

rates in mixed-solvent systems when water is enriched near the reactant, while DMSO has a negative value of $\Delta G_{H_3O^+}^{k\rightarrow H_2O}$ and leads to increased reaction rates in mixed-solvent systems when the cosolvent is enriched near the reactant. The distinct behavior of these cosolvents mirrors the difference in the sign of the calculated transfer free energies, suggesting that the transfer free energies are correctly capturing differences in the preference of the hydronium ion for bulk organic solvent.

Table 4.1: Dielectric constants and Kamlet-Taft parameters (α, β, π^*) for pure solvents, tabulated according to decreasing transfer free energy of a hydronium ion $\Delta G_{H_3O^+}^{k\rightarrow H_2O}$ in these solvents. $\sum \Delta G$ was computed with Equation 4.2. All solvation free energies are in units of kJ/mol.

		Kamlet-Taft Parameters				
Solvent	Dielectric constant[a]	α[b]	β[b]	π^*[c]	$\Delta G_{H_3O^+}^{k\rightarrow H_2O}$	$\sum \Delta G$
DIOX	2.20	0.00	0.37	0.49	184.6	389.85
THF	7.40	0.00	0.55	0.55	53.0	214.91
GVL	36.47	0.00	0.60	0.83	37.8	134.36
Water	78.50	1.17	0.47	1.14	0.0	0.00
NMP	32.16	0.00	0.77	0.92	-4.3	101.25
ACE	20.70	0.08	0.43	0.62	-11.2	99.55
DMSO	48.90	0.00	0.76	1.00	-63.5	41.69

[a] Values are from Ref. 31, except for GVL[47] and NMP.[48]

[b] Values from Ref. 49, except for GVL.[50]

[c] Values from Ref. 46, except for GVL[50] and water.[51]

4.3.5 Solvation free energies of the hydronium and chloride ion in mixed-solvent systems

Figure 4.6(a) shows the free energies of transferring either hydronium or chloride ion to aqueous mixtures of DMSO and DIOX from pure water; these solvents represent extrema of low and high affinity cosolvents for the hydronium ion. In aqueous mixtures of DMSO, Figure 4.6(a) shows a monotonic decrease in the hydronium ion transfer free energy (*i.e.*, an increase in hydronium ion stability relative to pure water) as the mass fraction of the organic phase increases. Since the free energy calculations in pure organic solvents (Figure 4.5(c)) show that pure DMSO stabilizes the hydronium ion more than water, these results agree with the expectation that increasing concentrations of DMSO results in improved stability of the hydronium ion. Figure 4.6(b) shows RDFs between the hydronium ion and both water and DMSO in 90 wt% DMSO. The peak of the ion-water RDF in 90 wt% DMSO is higher than in pure water, showing a local enrichment of water around the ion; however, there is also a cosolvent peak at ~0.38 nm, showing an enrichment in DMSO. Therefore, water and DMSO both favorable solvate the hydronium ion (visually shown in Figure 4.2(c)), leading to its increased stability relative to pure water. These results are consistent with experimental trends that find increasing concentration of DMSO monotonically increases basicity.[52] The results further suggest that there should be a driving force to partition the hydronium ion to regions

of the solvent system that have the highest concentration of DMSO to reduce its free energy to the greatest extent, agreeing with the hypothesis that local enrichment of DMSO around a reactant leads to an increase in acid-catalyzed reaction rates.

In aqueous mixtures of DIOX, Figure 4.6(a) shows a non-monotonic trend in the hydronium ion transfer free energy as the mass fraction of the organic phase increases. The transfer free energy is negative for all mixed compositions indicating that the hydronium ion is more stable than in either pure solvent. In the RDFs presented in Figure 4.6(b), the peak of the ion-water RDF in 90 wt% DIOX is almost ten-fold larger than the peak in pure water, indicating a significant enrichment of water around the hydronium ion (visually shown in Figure 4.2(b)). In addition, the cosolvent RDF (Figure 4.6(b), bottom) shows that DIOX is depleted near the hydronium ion up to distances of about 1 nm. These results together indicate that the hydronium ion nucleates a local domain of water molecules confined within the vicinity of the ion by the surrounding cosolvent. We attribute the decreased free energy of the hydronium ion in the mixed-solvent environment to the formation of this domain, which effectively sequesters water molecules to eliminate unfavorable water-cosolvent interactions that are not present in either pure solvent. Surprisingly, this data suggests that there should not be a driving force for hydronium ions to partition from bulk mixed-solvent environments to water-enriched domains near hydrophilic reactants as previously hypothesized because the solvation

free energy of the ion in pure water is higher than in the mixed-solvent environment. This finding suggests that the stabilization of the charged transition state by confined water molecules in the water-enriched local domain might be the dominant factor leading to increased reaction rates. However, these calculations omit explicit modeling of the reactant, which could affect partitioning thermodynamics.

Finally, Figure 4.6(a) shows that the chloride ion is not favored in any mixed-solvent composition, resulting in positive $\Delta G_j^{H_2O \to k}$ and $\sum \Delta G$ values for all DIOX and DMSO mass fractions. These data again indicate that acid dissociation is preferred in pure water rather than any mixed-solvent environment and thus weak acids are less likely to dissociate, diminishing reaction performance.

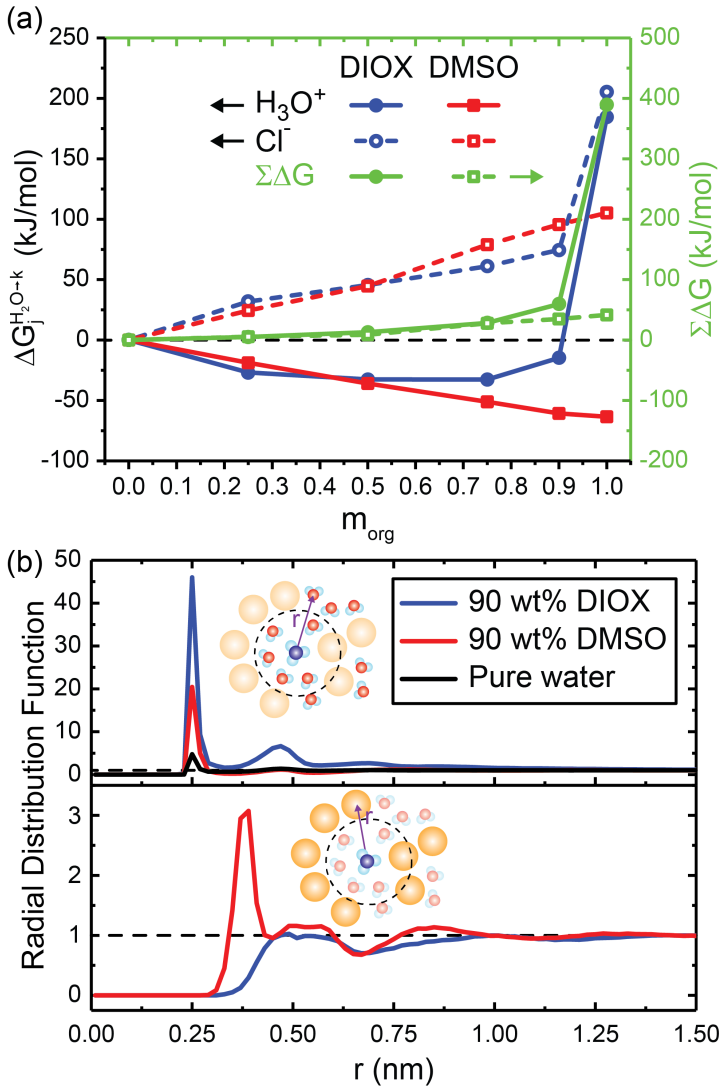


Figure 4.6: (a) Solvation free energies for transferring hydronium (H_3O^+, filled lines) and chloride (Cl^-, dashed lines) ions from pure water to aqueous mixtures of dioxane (DIOX, blue lines) and dimethyl sulfoxide (DMSO, red lines). The sums of the transfer free energies ($\sum \Delta G$) are shown as green lines. Error bars are not shown; they range from 0-2.5 kJ/mol when averaging two trials and tabulated in Supplementary Table 5. (b) Radial distribution function (RDF) between the center of mass of the hydronium ion to water (top) and the organic solvent (bottom) in 90 wt% DIOX, 90 wt% DMSO, and pure water. Bin widths for the RDFs were set to 0.02 nm.

4.4 Discussion

4.4.1 Screening solvent properties using a classical hydronium ion model

Recent studies of acid-catalyzed biomass conversion reactions have illustrated that the stability of the hydronium ion catalyst in various mixed-solvent environments can dramatically affect reaction rates.[3,4,6,7,12,53,54] Computational tools have been developed to study the hydronium ion in different solvent systems,[12,15,19] with ab initio molecular dynamics emerging as a powerful method to study interactions between the hydronium ion and solvent molecules due to the method's accuracy and ability to capture quantum mechanical effects.[55–63] However, ab initio simulations are computationally expensive and difficult to expand across multiple solvent systems. Therefore, we used a classical hydronium ion model[19] to compute the stability of the hydronium ion by measuring its solvation free energy in solvent systems with organic, polar aprotic solvents. Our findings suggest that the hydronium ion is unfavorable in DIOX, THF, and GVL solvents but favorable in NMP, ACE, and DMSO solvents (Figure 4.5(c)). These results can classify whether a solvent favorably facilitates a hydronium ion to help determine which phase the acid catalyst prefers in mixed-solvent systems. Furthermore, we could not identify a tabulated cosolvent-specific descriptor (*e.g.* dielectric constant, Kamlet-Taft parameters) that correlates with the hydronium ion solvation free ener-

gies, although the solvation free energies qualitatively capture features of solvent scales, such as the large distinction between DIOX and DMSO solvents. The lack of correlation suggests that the solvation free energy calculated from a MD simulation may provide unique information on proton-solvent interactions and can act as a cosolvent-specific descriptor for the stability of the acid catalyst.

In mixed-solvent environments, the solvation structure around the hydronium ion show that water-enriched local domains are formed, analogous to water-enrichment around hydrophilic reactants,[4,5] but the magnitude of enrichment is dependent on the choice of organic solvent. DMSO molecules compete with water for binding sites around the hydronium ion, whereas DIOX molecules are depleted around the hydronium ion. The hydronium ion solvation free energies in aqueous mixtures of DIOX suggest that small amounts of water can stabilize the hydronium ion to a greater degree than pure water or DIOX. This stabilization originates from the hydronium ion being confined by water, a solvent environment also found in water-enriched local domains formed by hydrophilic reactants. This finding suggests that stabilization of charged transition states by confined water in mixed-solvent environments may contribute to the increased reaction rates observed experimentally.

In all solvent environments studied, the sum of the transfer free energies of the hydronium and chloride ions from water was positive. This result indicates that non-aqueous environments tend to suppress acid

dissociation, leading to lower catalyst availability for weak acids that translates to lower reaction rates.[7] However, in this work we only studied the solvation free energy of a chloride ion conjugate base, and thus investigating the effect of alternative conjugate bases on acid dissociation could yield different effects on acid dissociation. For example, triflic acid is known to readily disassociate even in high concentrations of DMSO.[4] Future work will thus extend the framework developed here to further screen conjugate bases to determine the effect on acid dissociation, enabling the incorporation of these values into predictive models for reaction optimization.

4.4.2 Incorporation of the stability of the hydronium ion to the predictive model of reaction rates

Figure 4.4 showed that the acid-catalyzed dehydration of PDO depends on the choice of cosolvent, with the experimentally measured kinetic solvent parameter (σ) correlating with the simulation-derived preferential exclusion parameter (Γ) in aqueous mixtures of DIOX and DMSO. This correlation is based on the physical understanding that catalytic hydronium ions preferentially partition to the water-enriched local domain around the reactant, increasing reaction performance and leading to a positive correlation between σ and Γ (Figure 4.4(b)). However, the correlation between σ and Γ is negative in DMSO, a solvent for which water depletion around the

reactant is observed while reaction rates still increase. We hypothesized that the negative correlation may be because the hydronium ion preferentially interacts with DMSO rather than water and thus partitions to the water-depleted, DMSO-enriched local domain. This hypothesis is supported by the negative free energy for transferring a hydronium ion from water to DMSO as shown in Figure 4.5(b). Thus, the correlation between Γ and σ must be adjusted to account for the stability of the hydronium ion in the local domain.

We include the sign of the hydronium ion transfer free energy between pure organic solvent to water, $\Delta G_{H_3O^+}^{H_2O \rightarrow \text{pure org.}}$, as a correction term in the preferential exclusion coefficient (Γ') by using Equation 4.3.

$$\Gamma' = \Gamma \times sign(\Delta G_{H_3O^+}^{H_2O \rightarrow \text{pure org.}}) \tag{4.3}$$

Equation 4.3. ensures that Γ' and σ are positively correlated for aqueous mixtures of DMSO, shown in Figure 4.7(a). Since Γ' and σ are positively correlated for both aqueous mixtures of DIOX and DMSO, we can then write a correlative model for σ_{pred} that bridges these distinct solvents using Equation 4.4.

$$\sigma_{pred} = A(\Gamma') \tag{4.4}$$

where A is a constant. Supplementary Figure 8 shows the correlation between σ_{pred} and σ_{exp} when combining results from both DIOX-water and DMSO-water mixtures, resulting in a best-fit slope of 0.25 (ideally this

value should be unity), and a root-mean-square error (RMSE) between predicted and experimental value of 0.39.[2] Similar to our previous work,[5] we explored the use of multiple descriptors in combination to improve the correlations. In particular, we define $\Delta G_{HYD}^{k/H_2O}$ as the ratio of the transfer free energy of the hydronium ion in kth solvent system (ΔG_{HYD}^{k}) to the transfer free energy of hydronium ion in pure water ($\Delta G_{HYD}^{H_2O}$).

$$\Delta G_{HYD}^{k/H_2O} = \frac{\Delta G_{HYD}^{k}}{\Delta G_{HYD}^{H_2O}} \tag{4.5}$$

We interpret $\Delta G_{HYD}^{k/H_2O}$ as a unitless metric that quantifies the hydronium ion stability in mixed-solvent system relative to pure water. We expect that improved stability of the hydronium ion in mixed-solvent systems results in improved reaction kinetics. In addition, since acid-catalyzed reactions generally form a charged transition state after protonation of the reactant (Figure 4.1(b)), we hypothesize that the stability of the hydronium ion is an estimate to the stability of forming the transition state in mixed-solvent systems compared to pure water. Supplementary Figure 9 shows $\Delta G_{HYD}^{k/H_2O}$ as a function of mass content aqueous mixtures of DIOX and DMSO.[2] For these cosolvents, $\Delta G_{HYD}^{k/H_2O}$ is greater than unity, indicating that the hydronium ion is further stabilized in mixed-solvent environments than in pure water.

We combined Γ' and $\Delta G_{HYD}^{k/H_2O}$ into a multilinear regression frame work shown in Equation 4.6, where A, B, and C are coefficient constants. To en-

able comparison between the constants, we standardized Γ' and $\Delta G_{HYD}^{k/H_2O}$ by subtracting their mean and dividing by their standard deviations, described in Supplementary Section 5.2.[2] All standardized variables are denoted by a hat accent.

$$\sigma_{pred} = A(\widehat{\Gamma'}) + B(\widehat{\Delta G_{HYD}^{k/H_2O}}) + c \tag{4.6}$$

Figure 4.7(b) shows the correlation between σ_{pred} and σ_{exp} when using Equation 4.6 and combining results from both DIOX-water and DMSO-water mixtures. The best-fit slope is 0.91, close to the ideal value of unity, and the RMSE between predicted and experimental values is 0.13. These values are both similar to the correlations obtained in our previous work for seven hydrophilic compounds in single solvent systems.[5] Furthermore, the coefficients between $\widehat{\Gamma'}$ and $\widehat{\Delta G_{HYD}^{k/H_2O}}$ are comparable (0.33 vs. 0.38), showing that solvent enrichment around the reactant that favors the hydronium ion catalyst and the stability of the hydronium ion (or analogously, the transition state) are important variables for the prediction of acid-catalyzed reaction kinetics. We thus find that including information on the hydronium ion solvation free energy in an organic solvent can improve the correlation between Γ and σ when considering aqueous mixtures with different polar aprotic cosolvents. We note that additional solvent-specific descriptors (*e.g.* hydrogen bonding between water and the organic phase, *etc.*) may improve the correlation between different solvent systems and is

a subject of future research.

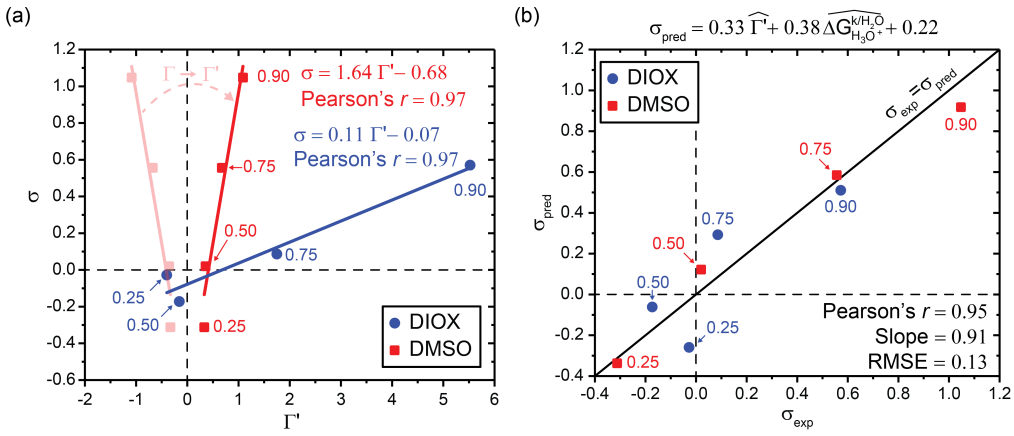


Figure 4.7: (a) Correlation between simulated preferential exclusion coefficient with solvation free energy correction term (Γ′), as expressed in Equation 4.4, and experimentally determined kinetic solvent parameters (σ) for aqueous mixtures of DIOX and DMSO. Experimental values were taken from Ref. 4. Transparent red points and lines show how Γ relates to Γ′. (b) Parity plot between predicted kinetic solvent parameter (σ_{pred}) and experimental kinetic solvent parameter (σ_{exp}) using results from aqueous mixtures of DIOX and DMSO. The predictive model is based on Equation 4.6 as shown above the plot. Data points are labeled with the wt% of the organic solvent.

4.5 Summary

We performed classical molecular dynamics simulations and solvation free energy calculations to quantify the stability of hydronium and chloride ions in six organic, polar aprotic solvents. We found that the hydronium ion is favorably solvated in pure NMP, ACE, and DMSO solvents, but

unfavorably solvated in pure DIOX, THF, and GVL solvents. In mixed-solvent environments, the inclusion of water with DIOX stabilizes the hydronium ion more than their pure solvent counterparts. We attribute this increased stabilization to the formation of water-enriched local solvent domains around the hydronium ion. In aqueous mixtures of DMSO, the hydronium ion is further stabilized with increasing concentration of the organic phase. Conversely, the chloride ion is destabilized in all pure organic solvents and mixed solvent systems, inhibiting acid dissociation. By quantifying the stability of the hydronium ion in organic solvents, we obtained a new cosolvent-specific descriptor that quantifies acid catalyst stability. We incorporated this descriptor into a predictive model for 1,2-propanediol dehydration reaction rates to demonstrate that the solvation free energy results can be used to bridge reaction rate predictions across different cosolvent systems. Incorporating information about the acid catalyst stability in different solvent mixtures represents an important step toward the rational design of mixed-solvent environments for acid-catalyzed reaction schemes and has the potential to alleviate time-intensive experimentation that accompanies the optimization of biomass conversion reactions for maximum productivity.

4.6 References

[1] Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations.

Frontiers in chemistry **2019**, *7*, 439.

[2] Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations [Supplementary Material]. *Frontiers in chemistry* **2019**, *7*, 439.

[3] He, J.; Liu, M.; Huang, K.; Walker, T. W.; Maravelias, C. T.; Dumesic, J. A.; Huber, G. W. Production of levoglucosenone and 5-hydroxymethylfurfural from cellulose in polar aprotic solvent–water mixtures. *Green Chemistry* **2017**, *19*, 3642–3653.

[4] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[5] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628.

[6] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[7] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[8] Sener, C.; Motagamwala, A. H.; Alonso, D. M.; Dumesic, J. A. Enhanced furfural yields from xylose dehydration in the γ-Valerolactone/water solvent system at elevated temperatures. *ChemSusChem* **2018**, *11*, 2321–2331.

[9] Caratzoulas, S.; Vlachos, D. G. Converting fructose to 5-hydroxymethylfurfural: a quantum mechanics/molecular mechanics study of the mechanism and energetics. *Carbohydrate research* **2011**, *346*, 664–672.

[10] Reif, M. M.; Hünenberger, P. H. Computation of methodology-independent single-ion solvation properties from molecular simulations. IV. Optimized Lennard-Jones interaction parameter sets for the alkali and halide ions in water. *The Journal of chemical physics* **2011**, *134*, 144104.

[11] Hünenberger, P.; Reif, M. *Single-ion solvation: experimental and theoretical approaches to elusive thermodynamic quantities*; Royal Society of Chemistry, 2011; Vol. 3.

[12] Varghese, J. J.; Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering* **2019**, *4*, 165–206.

[13] Duignan, T. T.; Baer, M. D.; Schenter, G. K.; Mundy, C. J. Real single ion solvation free energies with quantum mechanical simulation. *Chemical science* **2017**, *8*, 6131–6140.

[14] Mejıas, J.; Lago, S. Calculation of the absolute hydration enthalpy and free energy of H+ and OH-. *The Journal of Chemical Physics* **2000**, *113*, 7306–7316.

[15] Tawa, G.; Topol, I.; Burt, S.; Caldwell, R.; Rashin, A. Calculation of the aqueous solvation free energy of the proton. *The Journal of chemical physics* **1998**, *109*, 4852–4863.

[16] Pliego, J. R.; Riveros, J. M. The cluster- continuum model for the calculation of the solvation free energy of ionic species. *The Journal of Physical Chemistry A* **2001**, *105*, 7241–7247.

[17] Tunon, I.; Silla, E.; Bertran, J. Proton solvation in liquid water: an ab initio study using the continuum model. *The Journal of Physical Chemistry* **1993**, *97*, 5547–5552.

[18] Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Single-ion solvation free energies and the normal hydrogen electrode potential in methanol, acetonitrile, and dimethyl sulfoxide. *The Journal of Physical Chemistry B* **2007**, *111*, 408–422.

[19] Bonthuis, D. J.; Mamatkulov, S. I.; Netz, R. R. Optimization of classical nonpolarizable force fields for OH- and H3O+. *The Journal of chemical physics* **2016**, *144*, 104503.

[20] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[21] Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[22] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[23] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics* **1977**, *23*, 327–341.

[24] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[25] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[26] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[27] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.

[28] Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; Van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free

energy calculations based on molecular simulations. *Chemical physics letters* **1994**, *222*, 529–539.

[29] Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* **2008**, *129*, 124105.

[30] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of computer-aided molecular design* **2015**, *29*, 397–411.

[31] Fowler, F.; Katritzky, A.; Rutherford, R. The correlation of solvent effects on physical and chemical properties. *Journal of the Chemical Society B: Physical Organic* **1971**, 460–469.

[32] Vishnyakov, A.; Lyubartsev, A. P.; Laaksonen, A. Molecular dynamics simulations of dimethyl sulfoxide and dimethyl sulfoxide- water mixture. *The Journal of Physical Chemistry A* **2001**, *105*, 1702–1710.

[33] Kang, M.; Smith, P. E. Preferential interaction parameters in biological systems by Kirkwood–Buff theory and computer simulation. *Fluid Phase Equilibria* **2007**, *256*, 14–19.

[34] Schneider, C. P.; Trout, B. L. Investigation of cosolute- protein preferential interaction coefficients: New insight into the mechanism by which arginine inhibits aggregation. *The Journal of Physical Chemistry B* **2009**, *113*, 2050–2058.

[35] Shukla, D.; Trout, B. L. Preferential interaction coefficients of proteins in aqueous arginine solutions and their molecular origins. *The Journal of Physical Chemistry B* **2011**, *115*, 1243–1253.

[36] Shulgin, I. L.; Ruckenstein, E. Local composition in the vicinity of a protein molecule in an aqueous mixed solvent. *The Journal of Physical Chemistry B* **2007**, *111*, 3990–3998.

[37] Pliego Jr, J. R. Thermodynamic cycles and the calculation of pKa. *Chemical physics letters* **2003**, *367*, 145–149.

[38] Pliego Jr, J. R.; Riveros, J. M. New values for the absolute solvation free energy of univalent ions in aqueous solution. *Chemical Physics Letters* **2000**, *332*, 597–602.

[39] Horinek, D.; Mamatkulov, S. I.; Netz, R. R. Rational design of ion force fields based on thermodynamic solvation properties. *The Journal of chemical physics* **2009**, *130*, 124507.

[40] Fawcett, W. R. Acidity and basicity scales for polar solvents. *The Journal of Physical Chemistry* **1993**, *97*, 9540–9546.

[41] Koppel, I.; Palm, V. In *Advances in linear free energy relationships*; Springer, 1972; pp 203–280.

[42] Kamlet, M. J.; Taft, R. The solvatochromic comparison method. I. The. beta.-scale of solvent hydrogen-bond acceptor (HBA) basicities. *Journal of the American chemical Society* **1976**, *98*, 377–383.

[43] Kamlet, M. J.; Abboud, J. L.; Taft, R. The solvatochromic comparison method. 6. The. pi.* scale of solvent polarities. *Journal of the American Chemical Society* **1977**, *99*, 6027–6038.

[44] Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters,. pi.*,. alpha., and. beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry* **1983**, *48*, 2877–2887.

[45] Taft, R.; Kamlet, M. J. The solvatochromic comparison method. 2. The. alpha.-scale of solvent hydrogen-bond donor (HBD) acidities. *Journal of the American Chemical Society* **1976**, *98*, 2886–2894.

[46] Laurence, C.; Nicolet, P.; Dalati, M. T.; Abboud, J.-L. M.; Notario, R. The empirical treatment of solvent-solute interactions: 15 Years of. pi. *The Journal of Physical Chemistry* **1994**, *98*, 5807–5816.

[47] Wohlfarth, C. In *Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures*; Springer, 2015; pp 91–91.

[48] Uosaki, Y.; Kawamura, K.; Moriyoshi, T. Static Relative Permittivities of Water + 1-Methyl-2-pyrrolidinone and Water + 1,3-Dimethyl-2-imidazolidinone Mixtures under Pressures up to 300 MPa at 298.15 K. *Journal of Chemical & Engineering Data* **1996**, *41*, 1525–1528.

[49] Marcus, Y. The properties of organic liquids that are relevant to their use as solvating solvents. *Chemical Society Reviews* **1993**, *22*, 409–416.

[50] Jessop, P. G.; Jessop, D. A.; Fu, D.; Phan, L. Solvatochromic parameters for solvents of interest in green chemistry. *Green Chemistry* **2012**, *14*, 1245–1259.

[51] Buhvestov, U.; Rived, F.; Ràfols, C.; Bosch, E.; Rosés, M. Solute–solvent and solvent–solvent interactions in binary solvent mixtures. Part 7. Comparison of the enhancement of the water structure in alcohol–water mixtures measured by solvatochromic indicators. *Journal of physical organic chemistry* **1998**, *11*, 185–192.

[52] Catalán, J.; Díaz, C.; García-Blanco, F. Characterization of binary solvent mixtures of DMSO with water and other cosolvents. *The Journal of organic chemistry* **2001**, *66*, 5846–5852.

[53] Walker, T. W.; Motagamwala, A. H.; Dumesic, J. A.; Huber, G. W. Fundamental catalytic challenges to design improved biomass conversion technologies. *Journal of Catalysis* **2018**, *369*.

[54] Mellmer, M. A.; Alonso, D. M.; Luterbacher, J. S.; Gallo, J. M. R.; Dumesic, J. A. Effects of γ-valerolactone in hydrolysis of lignocellulosic biomass to monosaccharides. *Green Chemistry* **2014**, *16*, 4659–4662.

[55] Izvekov, S.; Voth, G. A. Ab initio molecular-dynamics simulation of aqueous proton solvation and transport revisited. *The Journal of chemical physics* **2005**, *123*, 044505.

[56] Marx, D. Proton transfer 200 years after von Grotthuss: Insights from ab initio simulations. *ChemPhysChem* **2006**, *7*, 1848–1870.

[57] Marx, D.; Tuckerman, M. E.; Hutter, J.; Parrinello, M. The nature of the hydrated excess proton in water. *Nature* **1999**, *397*, 601–604.

[58] Morrone, J. A.; Tuckerman, M. E. Ab initio molecular dynamics study of proton mobility in liquid methanol. *The Journal of chemical physics* **2002**, *117*, 4403–4413.

[59] Sagnella, D. E.; Laasonen, K.; Klein, M. L. Ab initio molecular dynamics study of proton transfer in a polyglycine analog of the ion channel gramicidin A. *Biophysical journal* **1996**, *71*, 1172–1178.

[60] Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. Ab initio molecular dynamics simulation of the solvation and transport of hydronium and hydroxyl ions in water. *The Journal of chemical physics* **1995**, *103*, 150–161.

[61] Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. Ab initio molecular dynamics simulation of the solvation and transport of H3O+ and OH-ions in water. *The Journal of Physical Chemistry* **1995**, *99*, 5749–5752.

[62] Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. Ab initio simulations of water and water ions. *Journal of Physics: Condensed Matter* **1994**, *6*, A93.

[63] Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. On the quantum nature of the shared proton in hydrogen bonds. *Science* **1997**, *275*, 817–820.

5 FAST PREDICTIONS OF LIQUID-PHASE ACID-CATALYZED REACTION RATES USING MOLECULAR DYNAMICS SIMULATIONS AND CONVOLUTIONAL NEURAL NETWORKS

Chapters 3 and 4 leveraged molecular dynamics-derived molecular descriptors to predict reaction rates. While these molecular descriptors provide us with physical understanding to how solvents could affect reactivity, these descriptors do not generalize across different cosolvent-water mixtures (*e.g.* GVL-water and THF-water mixtures in Figure S8 of ESI).[3] This chapter uses the simulation data available from Chapter 3 as input data to a deep learning framework to predict experimental reaction rates (σ_{exp}). In this chapter, we seek to answer the following questions:

- How could we transform molecular simulation information into a readable form for a deep learning model, described in Section 2.3?
- How could we use deep learning models to predict experimental reaction rates across different cosolvent-water systems?
- How do we use deep learning models to give us physical intuition into the important features from molecular dynamics simulations?

This chapter was reproduced from Chew, A. K.; Jiang, S.; Zhang, W.; Zavala, V. M.; Van Lehn, R. C. Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks. *Chemical Science* **2020**, *11*, 12464–12476, with permission from the Royal Society of Chemistry.[1] The supplementary information is cited as Ref. 2. S. Jiang, W. Zhang, and V. M. Zavala developed SolventNet and other 3D CNN architectures used to analyze voxel representations.

In this chapter, we show that the complex atomistic configurations of reactant-solvent environments generated by classical molecular dynamics simulations can be exploited by 3D convolutional neural networks to enable accurate predictions of Brønsted acid-catalyzed reaction rates for model biomass compounds. We develop a 3D convolutional neural network, which we call SolventNet, and train it to predict acid-catalyzed reaction rates using experimental reaction data and corresponding molecular dynamics simulation data for seven biomass-derived oxygenates in water-cosolvent mixtures. We show that SolventNet can predict reaction rates for additional reactants and solvent systems an order of magnitude faster than prior simulation methods. This combination of machine learning with molecular dynamics enables the rapid, high-throughput screening of solvent systems and identification of improved biomass conversion conditions.

5.1 Introduction

The catalytic conversion of lignocellulosic biomass is a promising strategy to obtain transportation fuels and high-value chemicals from renewable feedstocks.[4] The conversion of biomass-derived molecules is typically facilitated by liquid-phase, acid-catalyzed reactions (examples shown in Figure 5.1a) that are hindered by low reactivity in aqueous solution. One method to increase acid-catalyzed reaction rates is to modify the solvent

composition by mixing organic, polar aprotic cosolvents with water to create mixed-solvent environments; such environments have been shown to improve reaction rates up to 100-fold compared to rates for the same reactions in pure water.[5–7] Unfortunately, identifying solvent environments that improve reaction rates by trial-and-error experimentation is time-consuming, costly, and provides limited physical insight into the physical basis of solvent effects. Instead, computational tools have been applied to understand solvent effects on chemical reactivity and guide the design of solvent mixtures for efficient biomass conversion processes.[8]

In the past decade, *ab initio* quantum chemical methods have been used to quantify solvent effects on barriers to elementary steps for biomass conversion reactions.[4,8,9] Using *ab initio* molecular dynamics (MD) simulations, Mellmer *et al.* found that the inclusion of organic cosolvents increases biomass conversion reaction kinetics by lowering the activation energy due to changes to the solvation environment around the acidic proton catalyst, the reactant, and charged transition states (Figure 5.1b).[6] The simulations revealed that hydrophilic reactants drive the formation of water-enriched local domains in mixed-solvent environments that preferentially solvate the acid catalyst and stabilize subsequent carbocation-like transition states.[5–7] These findings suggest that the extent of water enrichment around the reactant is directly correlated with acid-catalyzed reaction performance. Similar studies have used *ab initio* techniques to understand how the solvent environment alters reaction kinetics for key

acid-catalyzed reactions,[8,10] such as the conversion of fructose to afford 5-hydroxymethylfurfural[11,12] (a platform chemical for polymer precursors and transportation fuels).[13] Unfortunately, while *ab initio* methods can directly probe mechanistic details of reactions, their high computational expense renders the screening of multiple solvent compositions infeasible.

Relative to *ab initio* MD, classical MD simulations can access longer timescales (μs) and larger length scales (nm) with significantly reduced computational budgets, allowing a more rapid characterization of complex solvent environments around even large reactants.[14–16] Classical MD simulations are suitable for modeling the spatial organization of mixed-solvent environments that emerges from the interplay of reactant-solvent-cosolvent interactions and may impact reaction kinetics (*e.g.*, due to preferential solvation of the reactant as described above) but may not be captured by bulk solvent descriptors (*e.g.*, the dielectric constant).[8,17] On the other hand, a key limitation of classical MD simulations is their inability to directly model chemical reactions. Nonetheless, we recently utilized classical MD simulations to understand and predict solvent effects on experimental reaction rates for the conversion of biomass-derived model compounds in aqueous mixtures of 1,4-dioxane (DIOX), γ-valerolactone (GVL), and tetrahydrofuran (THF).[7] Based on the hypothesis that classical MD simulations could quantify the reactant-water-cosolvent interactions that dictated reactivity in prior *ab initio* simulations,[6] we developed an MD model consisting of only reactant, water, and cosolvent molecules

and calculated three simulation-derived descriptors that quantified the extent of water-enrichment around the reactant, reactant-water hydrogen bonding, and reactant hydrophilicity. We then derived a linear regression model that used these three descriptors to predict experimental reaction rates and found good agreement in DIOX-water mixtures.[7] These findings showed that classical MD simulations can be used to predict solvent effects on reaction rates without explicitly modeling the acid catalyst or the reaction mechanism. The regression model was less accurate for GVL- and THF-water mixtures, indicating that either descriptors computed with classical MD cannot quantify reaction rates in these systems or that more complex descriptors must be defined to capture reactivity trends. However, designing new descriptors of reaction kinetics based on human intuition is challenging, often requiring complex and time-consuming data analysis tools (*e.g.* solvation free energies[17,18] or three-dimensional solvent mapping)[17] that cannot be readily generalized across a range of solvent compositions.

As an alternative to designing descriptors via human intuition, machine learning methods have been increasingly used to infer molecular properties by automatically extracting features from complex sources of data.[19–25] For example, convolutional neural networks (CNNs) can be used to identify and quantify patterns within two-dimensional (2D) spatial datasets such as images.[26] By training on a suitable set of labeled image data, CNNs extract spatial features without requiring human supervision

and can then utilize these features to classify image contents. CNNs have been shown to outperform other machine learning methods (*e.g.* fully connected neural networks and support vector machines)[27] in the ImageNet Large-Scale Visual Recognition Challenge,[28] a contest requiring a model to classify more than 1.2 million images. CNNs can be further generalized to extract features from three-dimensional (3D) volumetric data,[29] which can facilitate the analysis of 3D molecular structures. For example, 3D CNNs have recently been used to detect protein functional sites,[30] evaluate protein-ligand binding sites,[31] and quantify protein-ligand binding affinities[32] by training on protein database structures. Based on these examples and our prior success using classical MD simulations to predict acid-catalyzed reaction outcomes,[7,17] we hypothesize that 3D CNNs can exploit the output of classical MD simulations to more accurately predict solvent effects on acid-catalyzed reaction rates.

In this chapter, we developed 3D CNNs that utilize information on atomic positions obtained from classical MD simulation trajectories to predict the rates of liquid-phase, acid-catalyzed biomass conversion reactions in mixed-solvent environments. To develop our training procedure, we use 76 experimentally determined reaction rates for 7 biomass-derived model reactants in DIOX-, GVL-, and THF-water solvent mixtures as labels. For each experimental reaction rate and associated solvent mixture, we record configurations from a corresponding MD simulation trajectory (each configuration contains the 3D positions of atoms in reactant, solvent,

and water molecules). Configurations collected from the simulation and their spatial rotations are used to obtain multiple voxel representations that map to the same experimental reaction rate and are used as input data for the 3D CNN. This procedure allows us to construct a rich training dataset that consists of 18,240 voxel representations (over 240 distinct voxel representations mapping to each of the 76 experimental reaction rates). This approach seeks to demonstrate that MD trajectory data embeds rich information that explains reaction rates and develops early in the MD simulation to drastically reduce computational time. We find that all 3D CNNs considered - including a new 3D CNN that we developed, which we call SolventNet, and two previously developed 3D CNNs (ORION[33] and VoxNet)[34] - predict experimental reaction rates more accurately than models based on human-selected, MD-derived descriptors4. SolventNet predictions generalize to a test set consisting of 32 experimentally determined reaction rates obtained from the literature, including rates for reactants in three additional polar aprotic cosolvents not included in model training - dimethyl sulfoxide, acetonitrile, and acetone - with distinct properties (*e.g.*, functional groups, basicity, and polarizability) from the cosolvents used to train the model. We further find that accurate reaction rate predictions with SolventNet require as little as 4 ns of classical MD trajectory data, a 50-fold improvement from the original 205 ns of MD data used in models based on human-selected descriptors.[7] We thus conclude that 3D atomistic configurations contain significant information

that explains reaction rates and that such configurations develop early in the MD simulation. We envision that the computational efficiency associated with the combination of 3D CNNs and classical MD simulations will enable the integration of these tools with process models to screen solvents and optimize reactor conditions for biomass conversion processes.[35]

This chapter is organized as follows. We first describe the set of 108 experimentally determined reaction rates, each associated with a unique combination of reactant and mixed-solvent environment, that are used as labels for the training set (76 reaction rates) and test set (32 reaction rates). We then describe the training and validation of baseline linear and nonlinear models using data consisting of human-selected descriptors computed from MD trajectories, with one trajectory and corresponding set of descriptors per label. We next describe an alternative input data set generated by splitting each MD trajectory into 10 independent voxel representations for interpretation by 3D CNNs. After data augmentation, this procedure yields 240 voxel representations per label that are used to train and validate 3D CNNs for comparison to the baseline models. We then assess the test set and leave-one-out cross-validation accuracy of SolventNet to establish model generalizability. Finally, we visualize spatial features that influence SolventNet accuracy using saliency maps.

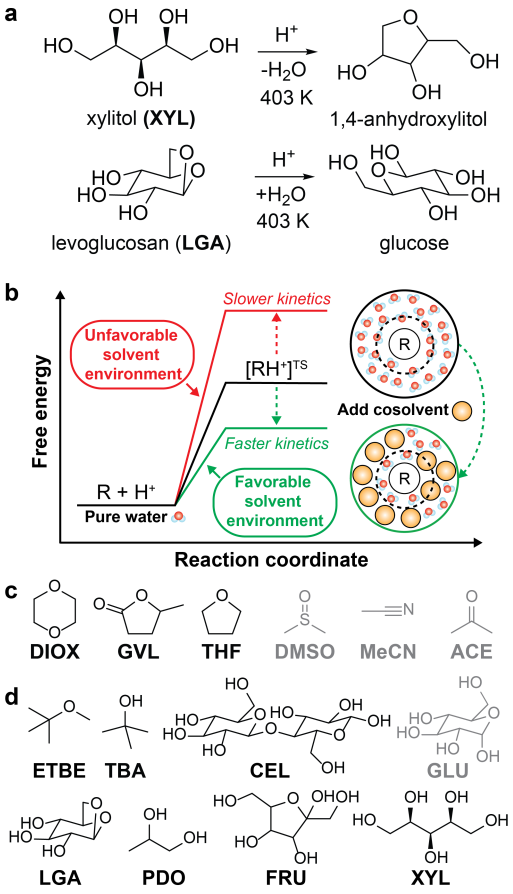


Figure 5.1: Overview of solvent effects on acid-catalyzed reactions and model systems. (a) Two example acid-catalyzed reactions: xylitol (XYL) dehydration and levoglucosan (LGA) hydrolysis. (b) Hypothesized effect of mixed-solvent environments on the free energy landscape of acid-catalyzed reactions. The schematic illustrates the formation of a local solvent domain (within the circular dashed line) around the reactant in a mixed-solvent environment that modifies the reaction free energy landscape, thus affecting reaction kinetics.[6,7] (c) Organic, polar aprotic cosolvents modeled in this study, including dioxane (DIOX), γ-valerolactone (GVL), tetrahydrofuran (THF), dimethyl sulfoxide (DMSO), acetonitrile (MeCN), and acetone (ACE). Molecules drawn in black were included in the training set. Molecules drawn in gray were included in the test set. (d) Biomass-derived model reactants modeled in this study, including ethyl tert-butyl ether (ETBE), tert-butanol (TBA), cellobiose (CEL), glucose (GLU), LGA, 1,2-propanediol (PDO), fructose (FRU), and XYL. The color scheme follows part c, except TBA, PDO, and FRU were included as part of some of the reactant-solvent combinations in both training and test sets.

5.2 Methods

5.2.1 Classical molecular dynamics simulation methods

Classical MD simulations were performed using a leapfrog integrator with a 2-fs timestep. The initial simulation box dimensions were set to $(6\ \text{nm})^3$ in all simulations, and water and cosolvent molecules were added in the desired proportions. All reactants and cosolvents were parameterized using the CGenFF/CHARMM36 forcefields.[36–38] Water was modeled using the Single Point Charge/Extended (SPC/E) model.[39] Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential that was smoothly shifted to zero between 1.0 nm and 1.2 nm. Electrostatic interactions were calculated using the smooth Particle Mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and 4th order interpolation. Bonds were constrained using the LINCS algorithm.[40] The solvent system was equilibrated in a *NPT* simulation for 5 ns at T = 300 K and P = 1 bar with a velocity-rescale thermostat and Berendsen barostat. A single reactant molecule was added to the system and equilibrated with the same barostat and thermostat for 500 ps. *NPT* production simulations were then performed at the reaction temperature and P = 1 bar using a Nose-Hoover thermostat and Parrinello-Rahman barostat. All thermostats used a 1.0 ps time constant and all barostats used a 5.0 ps time constant with an isothermal compressibility of $5.0 \times 10^{-5}\ \text{bar}^{-1}$. Simulation configurations were

output every 10 ps. All simulations were performed using GROMACS 2016[41] and visualized using VMD.[42]

Simulation data obtained using this protocol for reactant-solvent combinations included in the training set (including molecules drawn in black in Figure 5.1c and Figure 5.1d) were taken from Ref. 7. The production simulations for these molecules were performed for 200 ns. This simulation time was necessary due to the slow convergence of Γ as described in Ref. 7. The last 190 ns were used to compute Γ and δ for the multidescriptor models shown in Figure 5.2. An additional 5 ns production simulation was performed to compute τ. Descriptors were calculated as described in Ref. 7 (values listed in Table S1).[2] The first 20 ns of the 200 ns were used to generate voxel representations for training and validating the 3D CNNs. New simulations were performed following the above simulation protocol for the reactant-solvent combinations included in the test set (including molecules drawn in gray in Figure 5.1c and Figure 5.1d). The production simulations for these molecules were performed for 4 ns at the reaction temperatures (listed in Table S2)[2] and used to generate voxel representations for testing the 3D CNNs.

5.2.2 Summary of model training, validation, and testing

108 experimentally determined reaction rates from literature sources were converted to kinetic solvent parameters using Equation 3.4 and used as labels. Each reaction rate (label) was determined for a unique reactant-

solvent combination, which is defined as a single reactant in a binary mixture of a single cosolvent and water (in varying wt%). The labels were split into a training set with 72 labels (70% of the labels), all from Ref. 7, and a test set with 32 labels (30% of the labels) from Refs. 6, 43. A single MD trajectory was associated with each label. All models were evaluated using a 5-fold cross-validation procedure in which 80% of the labels and associated descriptors/voxel representations (60-61 labels) were used as training data and the remaining 20% of the labels (15-16 labels) and associated descriptors/voxel representations were used as validation data. This procedure was iterated 5 times such that each label was used for validation once. Model performance was evaluated based on the RMSE of the predicted values of the kinetic solvent parameter for the validation set.

Training data for the multidescriptor models were generated by using each MD trajectory to compute 3 descriptors from 205 ns of MD data as described above. The multidescriptor linear model was trained by regressing a line to the training data. The multidescriptor fully connected neural network was trained for 500 epochs (for each epoch, the neural network trained one cycle of the entire dataset) using the Keras deep learning library on top of Tensorflow.[44] Training was performed using the Adam optimizer with a learning rate of 0.001, mean-squared loss function, and training batches of size 18 (one epoch equates to four backpropagation steps).

Training data for the 3D CNNs were generated by splitting each MD

trajectory associated with a training set label into 10 independent partitions containing 2 ns (200 MD configurations) of consecutive MD data. For each partition, all MD configurations were converted to 3D grids of voxels that were averaged to obtain voxel representations as described in the Results and Discussion (Figure 5.3). The 10 voxel representations per label used for training were augmented by including all 24 unique cube rotations as training data, leading to a total of 240 voxel representations per label or 14,400-14,640 training voxel representations. The voxel representations per label used for validation were not augmented, leading to 150-160 validation voxel representations. All 3D CNNs were trained for 500 epochs using the Keras deep learning library on top of Tensorflow.[44] Models were trained using the Adam optimizer with a learning rate of 0.00001, mean-squared loss function, and training batches of size 18 (one epoch equates to 814 backpropagation steps). Learning curves for all 3D CNNs are shown in Fig. S5 of the Supplementary Information.[2]

The generalizability of SolventNet was assessed using the test set and leave-one-out cross-validation of the training set. For each test set label, 4 ns of consecutive MD data was converted to two independent voxel representations that were not augmented. SolventNet was re-trained using augmented voxel representations for all training set data (18,240 voxel representations) and used to predict kinetic solvent parameters for the 2 voxel representations per test set label (64 voxel representations). In the leave-one-out cross-validation procedure, all labels associated with

a single reactant or cosolvent in the training set were held out as the test set, then SolventNet was trained using all data for the remaining labels and used to predict kinetic solvent parameters for 10 voxel representations per held out label. In both procedures, model performance was assessed based on the RMSE of the predicted values of the kinetic solvent parameter for the test set.

5.2.3 3D CNN architectures

Three 3D CNN architectures that vary in complexity and number of parameters were considered in this work: VoxNet[34] (5 layers, 150,689 parameters), ORION[33] (8 layers, 908,833 parameters), and SolventNet (9 layers, 172,417 parameters), which we developed in-house. The architectures of VoxNet and ORION are described in the Supporting Information (Fig. S4).[2] SolventNet has three stages (Figure 5.4a). The first stage has two convolutional layers with 8 and 16 $3 \times 3 \times 3$ filters, respectively, and a $2 \times 2 \times 2$ max-pooling layer. The second stage has the same structure as the first, except the two convolutional layers have 32 and 64 filters, respectively. The results from the last max-pooling layer are passed to a batch normalization layer and flattened. The flattened data is then input to a neural network with three fully connected layers with 128 nodes in each layer. The ReLU activation function is used for the fully connected layers.

5.2.4 Generation of saliency maps

Saliency maps were computed using a fully trained SolventNet with the integrated gradient approach.[45] We define the voxel representation input as a tensor $x \in \mathbb{R}^{20\times20\times20\times20\times3}$. SolventNet has a loss function $L : \mathbb{R}^{20\times20\times20\times20\times3} \to \mathbb{R}$, where the output is the RMSE loss value. Equation 5.1 defines the saliency map function ($E : \mathbb{R}^{20\times20\times20\times20\times3} \to \mathbb{R}^{20\times20\times20\times20\times3}$):

$$E(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial L(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha \tag{5.1}$$

$E(x)$ is the saliency value of x. $\bar{x}$ is the baseline input, which we select as $\bar{x} = 0$. Equation 5.1 accounts for the change of the loss function caused by the change in the normalized occurrences of atoms in a voxel in the original input voxel representation. If the loss function does not significantly change, then that voxel is considered to be unimportant for the prediction. Values of $E(x)$, normalized to lie within the range 0-1, are shown on the saliency maps in Figure 5.7.

5.3 Results and Discussion

5.3.1 Experimental reaction data used for model training and testing

108 previously reported, experimentally determined reaction rates for acid-catalyzed dehydration or hydrolysis reactions in mixed-solvent environments were used as labels for the training and testing of predictive models for acid-catalyzed reaction rates. Each reaction rate was measured for a unique reactant-solvent combination (*i.e.*, a reactant in a mixed-solvent environment). 76 experimental reaction rates (70% of the labels) were taken from Ref. 7 and used as labels for the training set. The training set includes rates for 7 biomass-relevant model reactants: ethyl tert-butyl ether (ETBE), tert-butanol (TBA), 1,2-propanediol (PDO), levoglucosan (LGA), fructose (FRU), cellobiose (CEL), and xylitol (XYL). Solvent systems include aqueous mixtures with 25 wt%, 50 wt%, 75 wt%, or 90 wt% of one of three polar aprotic cosolvents: 1,4-dioxane (DIOX), γ-valerolactone (GVL), and tetrahydrofuran (THF). 32 experimental reaction rates (30% of the labels) were taken from Refs. 6 and 43 and used as labels for the test set. The test set includes rates for 4 model reactants: TBA, FRU, PDO, and glucose (GLU). Solvent systems include aqueous mixtures with dimethyl sulfoxide (DMSO), acetonitrile (MeCN), and acetone (ACE). Training set experiments were performed at temperatures between 343 K - 433 K and test set experiments were performed at temperatures between 363 K - 433

K. All experiments used triflic acid as a catalyst, except for test set experiments from Ref. 43 (for FRU and GLU in aqueous mixtures of ACE) which used hydrochloric acid. The test set labels thus consist of independent reaction rates from different literature sources for reactants and cosolvents not included in the training set but at comparable experimental conditions. Chemical structures of all reactants and cosolvents are shown in Figure 5.1c and Figure 5.1d, respectively. Table S1 and Table S2 list the reaction conditions for each label.[2]

To compare the effect of the mixed-solvent environment on reaction rates in different systems, Equation 3.4 defines the kinetic solvent parameter (σ) as the log-ratio between the apparent rate constant for the dehydration or hydrolysis of reactant r in a given mixed-solvent environment and the apparent rate constant for the same reaction in pure water.[7] $\sigma > 1$ indicates that the reaction rate is faster in the mixed-solvent environment than in pure water and $\sigma < 1$ indicates the converse. All labels for the training and test sets were defined as σ values (σ_{exp} to facilitate the training of regression models.

5.3.2 Baseline prediction models for reaction rates using human-selected descriptors from classical MD

In prior work, we performed classical MD simulations of one reactant molecule in a mixed-solvent environment to generate a single 205 ns simu-

lation trajectory for each of the 76 reactant-solvent combinations included in the training set.[7] Figure 5.2a illustrates the general approach for extracting descriptors from these MD simulations to predict experimental kinetic solvent parameters. From each MD trajectory, we computed three physically motivated, human-selected descriptors that capture reactant hydrophilicity and solvent interactions: the preferential exclusion coefficient (Γ), which quantifies the local enrichment of water in the spatial region near the reactant; the hydrogen bonding lifetime between the reactant and water (τ), which quantifies the stabilization of a putative charged transition state by nearby water molecules; and the accessible hydroxyl fraction (δ), which quantifies reactant hydrophilicity by dividing the accessible surface area (ASA) of the reactant's hydroxyl groups by the ASA of the overall molecule. Descriptor calculations are described in Chapter 3[7] and descriptor values are listed in Table S1.[2]

Descriptor values were used as input data to train a linear regression model (Equation 5.2) to predict values of the kinetic solvent parameter (σ_{pred}):

$$\sigma_{pred} = A = B(\widetilde{\Gamma}) + C(\widetilde{\tau}) + D(\widetilde{\delta}) \tag{5.2}$$

A, B, C, and D are regression coefficients and descriptors with a tilde are re-scaled between 0 and 1 by min-max scaling. Figure 5.2b illustrates the 5-fold cross-validation procedure used to evaluate if the linear model and all following models described in this work could generalize to new

reactant-solvent combinations not included in the training data.[46] In this procedure, the 76 labels in the training set were randomly split into five folds, each containing approximately 20% of the labels. All input data (*i.e.*, descriptor values) associated with the labels in one of the five folds were used as the validation set, the input data associated with the labels in the remaining four folds were used to train the model (*i.e.*, regression coefficients were fit), and values of σ_{pred} were calculated for the validation set. This procedure was iterated five times such that each training set label was included in the validation set exactly once.

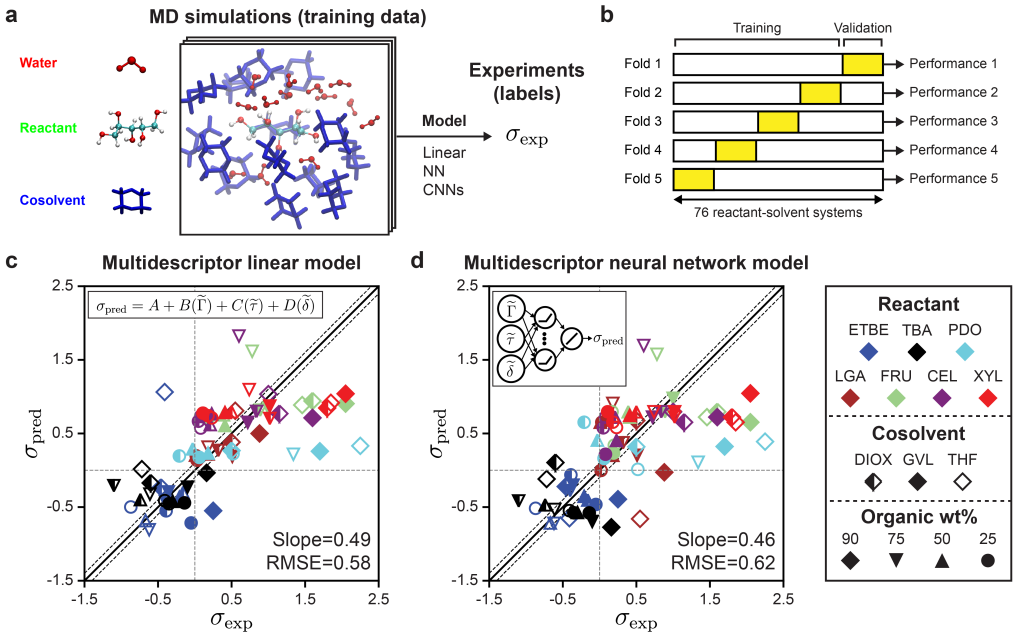


Figure 5.2: Evaluation of human-selected multidescriptor models. (a) General approach for correlating features from molecular dynamics (MD) simulations to experimental kinetic solvent parameters (σ_{exp}). The simulation configuration shows xylitol (XYL) in 90 wt% dioxane (DIOX) as an example. (b) Schematic illustrating 5-fold cross validation procedure used to train and validate models. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the multidescriptor linear model. The best-fit slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within the plot. The solid black lines indicate perfect correlation ($\sigma_{pred} = \sigma_{exp}$), the dashed black lines indicate approximate experimental error, and the dashed gray lines are drawn at $\sigma_{exp} = 0$ and $\sigma_{pred} = 0$ as a guide to the eye. (d) Parity plot for the nonlinear fully connected neural network model.

Figure 5.2c shows the parity plot between σ_{pred} and σ_{exp} for the linear regression model, with each value of σ_{pred} corresponding to a validation set prediction from the 5-fold cross validation procedure. We use the

best-fit slope and root-mean-square error (RMSE) between σ_{pred} and σ_{exp} to evaluate model accuracy; a perfect model would have a slope of one and a RMSE of zero. Since the RMSE of the experimental data is 0.10,[7] any predictive model with RMSE near 0.10 is sufficiently accurate. The linear model has a slope of 0.49 and RMSE of 0.58; most outliers are reactions occurring in THF-water (hollow symbols) or in 90 wt% cosolvent systems (diamond symbols). Fitting the linear model using only data for DIOX-water mixtures without 5-fold cross validation leads to an RMSE of 0.23,[7] but this approach is less accurate for GVL-water (RMSE of 0.36) and THF-water (RMSE of 0.59) mixtures (SI, Table S3),[2] indicating that the performance of the linear model depends strongly on the cosolvent. This result suggests that the linear model has limited generalizability across cosolvents. We also tested if a nonlinear model could improve upon the predictions of the linear model using the same input data. We performed 5-fold cross-validation to evaluate a fully connected neural network with three hidden layers, each with ten rectified linear units (ReLU), followed by a linear unit for the regression task of predicting σ using the three human-selected descriptors as input. Figure 5.2d shows the parity plot between σ_{pred} and σ_{exp} using the fully connected neural network model. The behavior of the fully connected neural network model was comparable to that of the linear model with a slope of 0.46 and RMSE of 0.62. We use these multidescriptor models as a baseline for comparison to alternative models.

These findings show that reaction rates predicted with the multidescriptor models lie in the correct quadrants with few false-positive or false-negative σ values and are accurate for some cosolvents (*i.e.*, DIOX). The descriptors underscore the importance of spatial information: Γ quantifies the solvent composition in a region near the reactant, τ describes the relative locations of reactant hydroxyl groups and water molecules, and δ is related to the relative surface areas of hydrophilic and hydrophobic regions of the reactant. However, both models have significant outliers for systems corresponding to larger values of σ_{exp}, suggesting that the descriptors fail to capture important information that may be encoded within the complex geometrical (3D) features of the reactant-solvent environment. In addition, the identification of these descriptors requires domain expertise and is time-consuming.

5.3.3 Generation of input data set for interpretation by 3D CNNs using classical MD data

To improve upon the human-selected multidescriptor models, we hypothesized that 3D CNNs can be used to establish mappings between atomic positions sampled from classical MD (which emerge from the combination of reactant-solvent, solvent-cosolvent, and reactant-cosolvent interactions) to experimental reaction rates. We expect that 3D CNNs are appropriate for analyzing these systems because:

(*i*) CNNs can extract non-intuitive features of the reactant-solvent environment by identifying spatial correlations in the input data

(*ii*) liquid-phase systems exhibit pronounced spatial correlations due to intermolecular interactions between solvent molecules

(*iii*) the importance of spatial information encoded within the human-selected descriptors suggests that spatial correlations are relevant to solvent effects

(*iv*) 3D CNNs can analyze atomic positions without transforming the domain to a 2D space (flattening), thereby capturing the detailed geometry of the reactant and local solvation environment

3D CNNs thus provide a natural framework for identifying complex features of reactant-solvent environments that affect reaction rates but that may not be easily identified using human intuition.

We first developed a protocol for converting trajectory data of atomic positions of reactant, solvent, and cosolvent molecules obtained from classical MD simulations into a data representation that is suitable for 3D CNN analysis. 3D CNNs interpret data consisting of a series of voxels arranged in a 3D grid, with each voxel containing normalized intensities in several independent channels. The channels can convey different types of field information. The relative positioning of the voxels in the grid confers spatial information. We thus converted the spatially continuous atomic

positions output by MD to voxels that record the normalized occurrences of water, cosolvent, and reactant oxygen atoms within $(0.2\ \text{nm})^3$ volume elements. This data representation is motivated by the physical intuition obtained from the success of the human-selected multidescriptor models: the importance of the descriptor Γ suggests that the positions of water and cosolvent atoms should be recorded to quantify preferential enrichment of solvent molecules near the reactant while the descriptors τ and δ suggest that the positions of reactant oxygen atoms should be recorded to quantify potential hydrogen bonding and reactant hydrophilicity. The volume associated with each voxel was selected to be comparable to typical atomic radii to ensure that molecular geometry could be resolved.

Figure 5.3 illustrates the approach used to convert MD positions to a grid of voxels. For each set of atomic positions corresponding to a single time sampled during a MD trajectory (*i.e.*, a MD configuration), we centered a 3D histogram on the center-of-mass of the reactant. The histogram covered a cubic $(4\ \text{nm})^3$ volume (a volume smaller than the total simulation box size to avoid crossing the simulation box boundaries) that was divided into a $20 \times 20 \times 20$ grid of bins corresponding to $(0.2\ \text{nm})^3$ volume elements. For each bin, we calculated the normalized occurrence of water atoms by counting the number of water atoms within the bin and normalizing by the maximum number of water atoms within any bin. The same procedure was separately performed to calculate the normalized occurrence of cosolvent and reactant oxygen atoms in each bin. The

normalized water, reactant, and cosolvent occurrences were then stored in the "red, green, and blue" color channels, respectively (Figure 5.3a) to obtain a $20 \times 20 \times 20 \times 3$ grid of voxels for a single MD configuration.

To capture the preferential locations of solvent molecules relative to the reactant as they diffuse within the cubic volume, and to prevent voxels from being unoccupied, we averaged grid values obtained from multiple consecutive MD configurations to generate a single averaged grid of voxels that we define as a voxel representation. Specifically, each voxel representation was generated by averaging grid values from 2 ns of MD data (corresponding to 200 consecutive MD configurations) as illustrated in Figure 5.3b. The 2 ns simulation time was selected as a balance between maximizing model accuracy and minimizing computational expense. This simulation time is substantially shorter than the 205 ns trajectories associated with each training label. Thus, we split the first 20 ns of each trajectory into 10 independent 2-ns partitions and generated a voxel representation for each partition, yielding 10 voxel representations per training label. These choices were based on extensive robustness tests to determine the best-performing input data representations as described in the Supplementary Information, Section S5 (Table S4).[2] Because 3D CNNs are not rotationally invariant, we further augmented the training data by rotating each voxel representation to generate 24 unique cube rotations per voxel representation (Fig. S1),[2] leading to 240 (augmented) voxel representations per training set label.

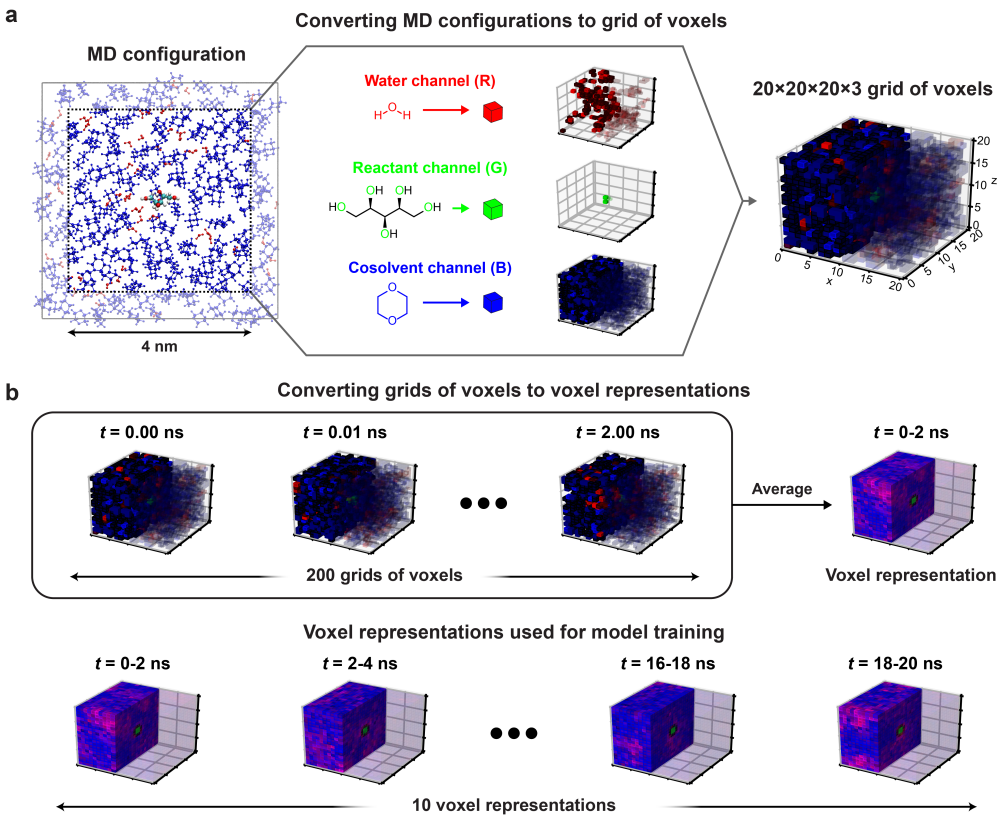


Figure 5.3: Input data representation for 3D CNNs. Approach for converting the atomic positions obtained from a MD simulation to a voxel representation using xylitol in 90 wt% dioxane as an example. (a) For each MD configuration (example at left), a $(4\ \text{nm})^3$ cubic box was centered on the reactant and a $20 \times 20 \times 20$ grid of $(0.2\ \text{nm})^3$ volume elements was used to discretize space. The normalized occurrences of water, oxygens of the reactant, and cosolvent atomic positions within each volume element were stored in different channels to yield a $20 \times 20 \times 20 \times 3$ grid of voxels. Voxels are visualized by showing the water channel in red, the reactant channel in green, and the cosolvent channel in blue. Half of the voxels are transparent to illustrate the solvent distribution around the reactant. (b) Grids of voxels were averaged over 2 ns of MD data (200 MD configurations) to yield a $20 \times 20 \times 20 \times 3$ voxel representation. (c) For each reactant-solvent system, 20 ns of simulation data were used to generate 10 independent voxel representations.

5.3.4 3D CNNs improve reaction rate predictions with less required simulation time

We next developed a 3D CNN model, which we call SolventNet, that inputs voxel representations and outputs predicted kinetic solvent parameters. The SolventNet architecture consists of four convolutional layers, two max-pooling layers, and three fully connected layers as shown schematically in Figure 5.4a. This architecture is based on the previously developed VoxNet 3D CNN but replaces the final fully connected layer with two convolutional layers, one max-pooling layer, and three fully connected layers.[34] We also compared SolventNet to the VoxNet[34] and ORION[33] 3D CNNs (SI, Fig. S4)[2] to investigate how the 3D CNN architecture influences prediction accuracy. All three models include a final layer with a linear activation unit for the regression task of predicting σ.

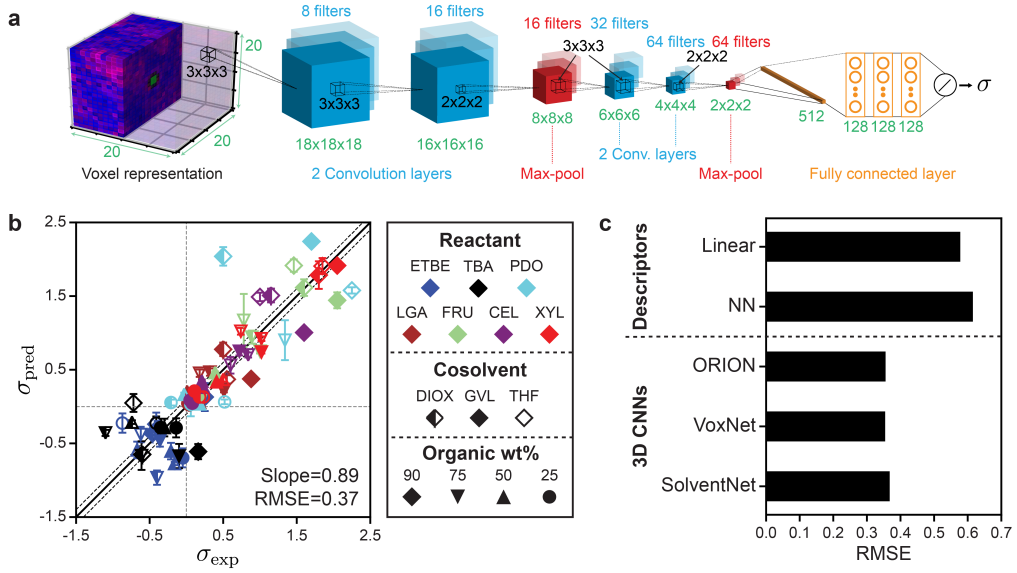


Figure 5.4: Architecture, training, and performance of 3D CNNs. (a) Architecture of SolventNet, a 3D CNN that inputs a $20 \times 20 \times 20 \times 3$ voxel representation (described in Fig 3) and outputs the predicted kinetic solvent parameter (σ). 3D CNNs were evaluated using the same 5-fold cross validation procedure described in Figure 5.2b. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters using SolventNet. σ_{pred} is the average prediction of 10 voxel representations per label. Error bars show the standard deviation of these predictions. The best-fit slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within the plot. Solid and dashed lines follow the conventions of Figure 5.2. (d) Comparison of the RMSEs between σ_{pred} and σ_{exp} for the multidescriptor linear and nonlinear neural network (NN) models and the 3D CNNs (ORION, VoxNet, and SolventNet) when performing 5-fold cross validation.

The 3D CNNs were evaluated using 5-fold cross-validation, similar to the procedure used for the human-selected descriptor models. For the 3D CNNs, the training data included all augmented voxel representations

for 80% of the labels (14,400 or 14,640 voxel representations) and the validation data included only non-augmented voxel representations for the remaining 20% of the labels (150 or 160 voxel representations). Figure 5.4b shows the parity plot between σ_{pred} and σ_{exp} using SolventNet. Values and error bars of σ_{pred} report the average and standard deviation of the values predicted for each of the 10 non-augmented validation set voxel representations per label. Compared to the baseline human-selected multidescriptor models (Figure 5.2), SolventNet significantly improves the prediction of kinetic solvent parameters with a best-fit slope of 0.89 and RMSE of 0.37. Figure 5.4c further indicates that the RMSEs obtained using all three 3D CNNs (SolventNet, VoxNet, and ORION) are comparable and outperform the multidescriptor models. Moreover, the 3D CNNs were trained using only 20 ns of MD data per reactant-solvent combination compared to the 205 ns of MD data per reactant-solvent combination used to compute the three descriptors in Equation 5.2. The 3D CNNs required 1.6-2.4 hours to train using one GPU and one CPU core (Table S3), whereas the MD simulations required 216 hours for all 76 reactant-solvent combinations using one GPU and 28 CPU cores, a substantially longer time. Therefore, the 10-fold reduction in MD data translates to a comparable decrease in real time required for model training.

A potential issue when training CNNs is the large number of learned parameters, which may lead to overfitting. We note, however, that the 3D CNNs used in this work are relatively compact; SolventNet has 172,417

parameters compared to 33,601,345 parameters for VGG16, a common 2D CNN[47] (Methods and Table S3).[2] The ratio of parameters to the number of augmented training voxel representations for SolventNet is 11.8-11.9; for comparison, the ratio of parameters to the number of training descriptor sets for the fully connected neural network model is 4.4. The small ratio of parameters to training examples for SolventNet is comparable to alternative 3D CNN architectures and suggests that we should not expect significant overfitting during training. For example, 3D CNNs have been used to assess the quality of protein tertiary structures with a ratio of parameters to training examples of 34 (~5.4 million parameters, ~160,000 grids of voxels)[48] and 85 (~170 million parameters, ~2 million grids of voxels).[49] We also observe that increasing the amount of training voxel representations by a factor of 10 (by splitting 200 ns of MD data into 100 voxel representations) does not impact 3D CNN performance (Fig. S3).[2] Moreover, the differences in RMSE between ORION, VoxNet, and SolventNet show that CNN architecture minimally affects model performance, even though ORION has nearly five times as many parameters as SolventNet. Therefore, the specific 3D CNN architecture is not critical; rather, it is the use of a 3D CNN to analyze MD simulations that yield improved model performance. We also tested alternative neural network architectures (SI),[2] including networks with both descriptors and voxel representations as input, 2D CNNs, and 3D CNNs with alternative voxel representations, but did not observe increases in accuracy. Together, these results show that

3D CNNs are more accurate than prior models based on human-selected descriptors (Figure 5.2) while requiring less MD data to train, that the amount of training data used is sufficient, and that prediction accuracy does not significantly vary with model architecture. For the remainder of this chapter, we focus on SolventNet because it has the median number of parameters and performs well when predicting the test set as discussed in the next section (see Table S3 for model comparisons).[2]

5.3.5 Generalizability of SolventNet predictions to new solvents and reactants

We next assessed the accuracy of SolventNet for the 32 reactant-solvent combinations included in the test set. For each test set reactant-solvent combination, we performed new MD simulations to obtain the 4 ns of simulation data necessary to generate two voxel representations (following the same procedure illustrated in Figure 5.3). Only two voxel representations were used to assess the ability of SolventNet to rapidly generalize to new solvent combinations with minimal MD simulation data. The test set voxel representations were not augmented. We re-trained SolventNet using all training set data (76 labels and 18,240 augmented voxel representations) and then predicted kinetic solvent parameters for the test set voxel representations.

Figure 5.5 shows the parity plot between σ_{pred} and σ_{exp} for the test

set using SolventNet. Values and error bars of σ_{pred} report the average and standard deviation of the values predicted for each of the 2 test set voxel representations per label. The best-fit slope is 0.72 and RMSE is 0.48, indicating that while prediction accuracy slightly degrades compared to predictions for the validation set, SolventNet still generalizes well. Notably, the test set accuracy exceeds the validation set accuracy of the baseline multidescriptor models. The parity plot also shows high linear correlation with Pearson's $r = 0.80$ for the test set data. Thus, even for systems for which SolventNet predictions are less accurate, the model still captures qualitative trends regarding solvent compositions that improve reaction rates, despite including reactants (GLU), cosolvents (MeCN, DMSO, and ACE), and organic weight fractions (44, 65, and 88 wt%) not present for any of the reactant-solvent combinations in the training set. These findings suggest that SolventNet extracted features of the reactant-solvent environment that are important to reaction performance and that are generalizable across different reactant-solvent combinations. The reduced accuracy may be due, in part, to the different literature sources for the test set data (which may introduce experimental error) and the use of a hydrochloric acid rather than triflic acid catalyst in Ref. 43, since chloride can impact reaction kinetics.[9]

The test set results show that SolventNet performs well for DMSO-water mixtures with a RMSE of 0.43. This result is somewhat surprising because DMSO is more basic than the cosolvents used for training[6] and is

known to compete against water for binding sites around hydroxyl groups on hydrophilic reactants.[15,17] In addition, we have previously found that reactant-DMSO interactions can be favored over reactant-water interactions in DMSO-water mixtures, whereas reactant-water interactions are favored in the other cosolvent-water systems.[18] Despite these unique behaviors, the features learned by SolventNet can translate to reaction rate predictions for DMSO-water mixtures with accuracy comparable to predictions for all other systems tested. Of the reactants in the test set, the worst prediction accuracy was obtained for GLU with a RMSE of 0.88. Since GLU was not part of the training set, we expect that the prediction accuracy for this reactant would be lower than FRU; moreover, this system used a hydrochloric acid catalyst as noted above. Nonetheless, the qualitative trend is again captured for GLU conversion with a Pearson's $r = 0.93$ (computed only for systems with GLU).

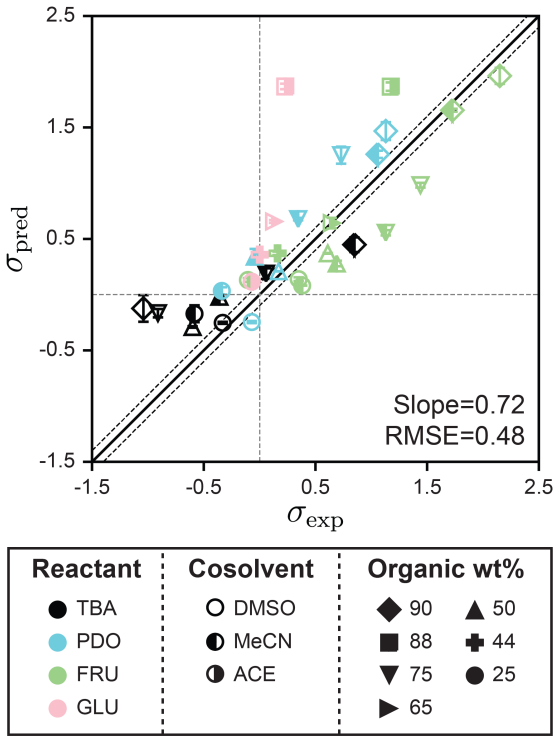
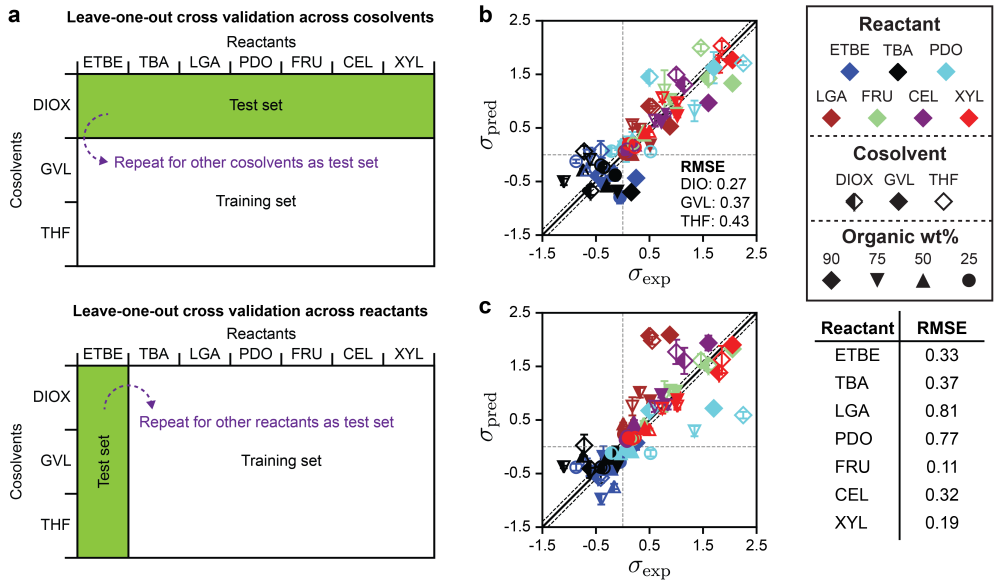


Figure 5.5: Generalizability of SolventNet to new reactants and cosolvents. Parity plots between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the test set. Predictions were made using SolventNet after training with all training set data. σ_{pred} is the average prediction of 2 voxel representations per label. Error bars show the standard deviation of these predictions. The slope and root-mean-squared error (RMSE) between σ_{pred} and σ_{exp} values are shown within each plot. Solid and dashed lines follow the convention of Figure 5.2.

Based on the DMSO and GLU analysis, we also used leave-one-out cross-validation to determine if SolventNet predictions were sensitive to particular reactants or cosolvents included in the training set, further motivated by the weak performance of descriptor-based models for THF-

and GVL-water mixtures. In this procedure (illustrated in Figure 5.6a), we held out all labels and associated voxel representations for reactant-solvent combinations in the original training set that either contained a given cosolvent (*e.g.*, all DIOX-water mixtures) or a given reactant (*e.g.*, XYL in all solvent systems) and used these data as the test set. SolventNet was trained using the remaining data and used to predict kinetic solvent parameters for 10 voxel representations per test set reactant-solvent combination. This procedure was repeated by iteratively using data for each cosolvent or reactant as the test set. Figure 5.6b shows the parity plot between σ_{pred} and σ_{exp} for leave-one-out cross validation across cosolvents. The RMSE varies between 0.27-0.43, which is comparable to the predictions for DMSO-water mixtures. These results indicate that SolventNet predictions are comparable for a wide range of mixed-solvent environments, including the THF- and GVL-water mixtures that were poorly predicted by the linear multidescriptor model. Figure 5.6c shows the parity plot between σ_{pred} and σ_{exp} for leave-one-out cross validation across reactants. The RMSE varies between 0.11-0.81 depending on the specific reactant, which is comparable to the results for leave-one-out cross validation across cosolvents. The largest RMSE is obtained for LGA which is comparable to the test set results for GLU. As with the GLU results, σ_{pred} and σ_{exp} for LGA exhibit strong linear correlation with a r of 0.90, indicating that quantitative trends of reactivity are captured. LGA may be an outlier due to the overestimation of its hydrophilicity since the voxel representations account for all oxygens

of LGA, including oxygens in ether bonds. Taken together, the results from the independent test set and from leave-one-out cross-validation suggest that SolventNet predictions generalize well across all cosolvents tested and all reactants other than LGA.



Reactant	RMSE
ETBE	0.33
TBA	0.37
LGA	0.81
PDO	0.77
FRU	0.11
CEL	0.32
XYL	0.19

Figure 5.6: Leave-one-out cross validation of SolventNet. (a) Schematic illustrating the leave-one-out cross validation procedure in which all data for a single cosolvent or all data for a single reactant were used as the test set and the remaining data were used as the training set. (b) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the leave-one-out cross validation of SolventNet across cosolvents. The RMSE values between σ_{pred} and σ_{exp} labeled within each plot report the values obtained when data for the listed cosolvent-water system were used to as the test set. σ_{pred} is the average prediction of 10 voxel representations per label. Error bars show the standard deviation of these predictions. Solid and dashed lines follow the conventions of Figure 5.2. (c) Parity plot between predicted (σ_{pred}) and experimental (σ_{exp}) kinetic solvent parameters for the leave-one-out cross validation of SolventNet across reactants. The RMSE values in the table report the values obtained when data for the listed reactant were used as the test set.

5.3.6 Physical interpretation of SolventNet features

While SolventNet offers improved prediction accuracy and computational efficiency compared to models based on human-selected descriptors, it is difficult to physically interpret features extracted by the model.[50] For example, representative voxel representations for different reactant-solvent combinations do not have visually distinctive features (Fig. S2).[2] As an initial step toward interpreting features recognized by SolventNet, we generated saliency maps to visualize the sensitivity of SolventNet predictions to different voxels and thus to specific spatial regions around the reactant. A saliency map consists of saliency values (normalized between 0-1) for each voxel that indicate how sensitive the SolventNet prediction is to the normalized occurrences of water, reactant, and cosolvent atoms in that voxel. Larger saliency values indicate increased sensitivity. Figure 5.7 shows a saliency map that was generated using the integrated gradient approach[45] (described in the Methods) by inputting a voxel representation of XYL in 90 wt% DIOX to a fully trained SolventNet model. The saliency map is split into separate 3D grids of voxels for reactant, cosolvent, and water channels, with each voxel colored according to its normalized saliency value using the same color scheme as the input voxel representation (Figure 5.3). Transparent voxels are unimportant to model predictions (normalized saliency values less than 0.10).

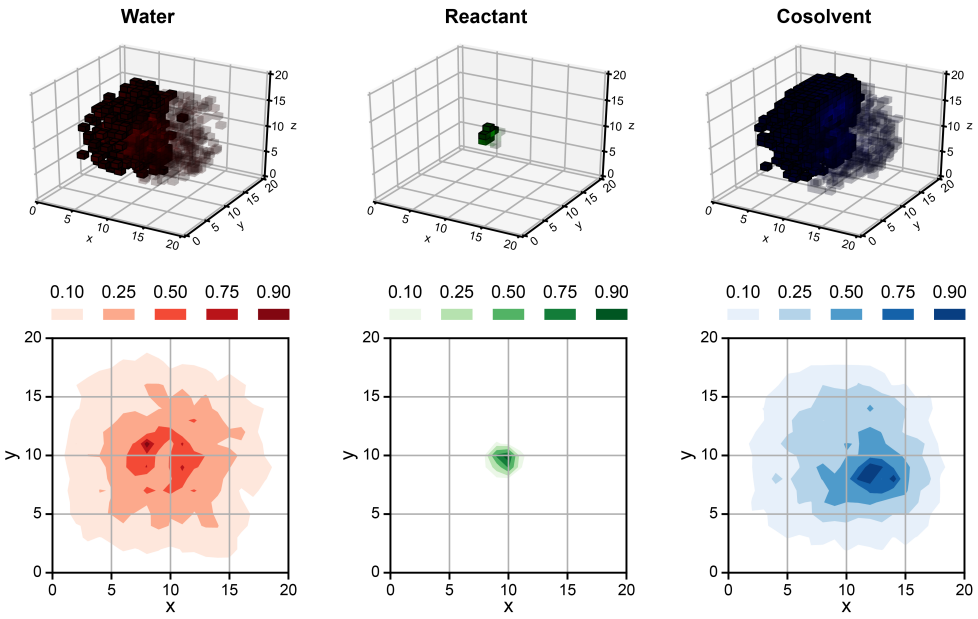


Figure 5.7: Saliency map using SolventNet. Example saliency map generated for a voxel representation of XYL in 90 wt% DIOX (shown in Figure 5.4a) using SolventNet after training with all training set data. The saliency map is visualized on a 3D grid with the same dimensions as the input voxel representation. Each voxel is assigned a saliency value normalized from 0 to 1 that indicates the sensitivity of SolventNet predictions to the normalized occurrences of water, reactant, and cosolvent atoms in that voxel. Larger saliency values indicate greater sensitivity. The saliency map is visualized by separate grids showing the water value in red, the reactant value in green, and the cosolvent value in blue. Half of the voxels are transparent to illustrate the saliency values around the reactant and only the voxels with values greater than 0.10 for each system component are shown. 2D contours are plotted by averaging along the z-axis for normalized values of 0.10, 0.25, 0.50, 0.75, and 0.90.

Inspection of the saliency map in Figure 5.7 confirms that SolventNet primarily recognizes features of the solvent environment within a local

domain near the reactant, in agreement with the assumption underlying the definition of human-selected descriptors. Saliency maps for each channel are further projected from 3D to 2D by averaging saliency values along the z-axis (selected arbitrarily) to generate the contour plots shown in Figure 5.7. These plots clearly show that regions near the reactant are most important for predictions, that the size of the simulated volume is sufficiently large that regions far from the reactant are unimportant, and that SolventNet extracts non-intuitive geometrical features that are not captured by the assumptions of spherical symmetry. Similar saliency maps may be useful for guiding future *ab initio* calculations by identifying important solvent regions to be studied, thus minimizing the number of molecules needed for quantum chemical calculations.

5.4 Summary

In this chapter, we combined machine learning tools with classical MD simulations to develop SolventNet, a 3D CNN which interprets MD simulation data that has been transformed into a voxel representation to predict acid-catalyzed reaction rates in aqueous mixtures with polar aprotic cosolvents. SolventNet does not require the human selection of descriptors a priori and can capture spatial correlations without assuming spherical symmetry. SolventNet (and other 3D CNNs) more accurately predicts reaction rates than models based on human-selected descriptors and generalizes to new

polar aprotic cosolvents, cosolvent mass fractions, and reactants not used during model training. These findings indicate that SolventNet extracts features from the spatial configurations of reactant, solvent, and cosolvent molecules determined by classical MD simulations that encode sufficient information to predict solvent effects on acid-catalyzed reactions, even though the reaction mechanisms and possible transition states are not explicitly modeled.

A significant advantage of this newly developed approach is computational efficiency. Once trained, SolventNet requires as little as 4 ns of MD simulation data to predict a reaction rate for a single reactant-solvent combination. For the system sizes considered in this work, these trajectories can be simulated in less than an hour on single node of a typical high-performance computing cluster, greatly diminishing the time necessary to predict reaction rates compared to *ab initio* methods. This reduced computational expense suggests that SolventNet could be leveraged for solvent screening, potentially in combination with process models,10 to design more efficient biomass conversion processes without costly experimental trial-and-error. However, all systems studied so far involve mixtures of water and polar aprotic cosolvents; extending SolventNet to predict reaction rates in substantially different solvent systems (*e.g.*, ionic liquids) will likely require additional training.

The voxel representations input to SolventNet also only contain atomic positions, thus omitting important chemical information such as the pres-

ence of covalent bonds between atoms and atomic charges. These data could possibly be interpreted by alternative network architectures. For example, a graph neural network could represent atoms as nodes and atomic interactions as edges.[16,19,20,51] Thus, we anticipate that further model development can continue to increase prediction accuracy. Furthermore, the voxel representations input to SolventNet are only generated from MD simulations of reactant states; future work will explore whether including voxel representations from simulations of product states can improve the accuracy of reaction rate predictions or enable predictions of reaction selectivities.[17] Classical MD can also model molecules much larger than the small-molecule reactants studied here, such as biomass-relevant polymers (*e.g.*, cellulose, hemicellulose, or lignin).[14,15] Future work will explore the use of SolventNet with semantic segregation techniques, which enable individual regions of data input to a CNN to be separately classified,[52] to predict the reactivity of multiple reactive sites simultaneously based on MD simulations of biomass-relevant polymers. Screening solvents for these larger systems is difficult with *ab initio* methods.

5.5 References

[1] Chew, A. K.; Jiang, S.; Zhang, W.; Zavala, V. M.; Van Lehn, R. C. Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks. *Chemical Science* **2020**, *11*, 12464–12476.

[2] Chew, A. K.; Jiang, S.; Zhang, W.; Zavala, V. M.; Van Lehn, R. C. Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks [Supporting Information]. *Chemical Science* **2020**, *11*, 12464–12476.

[3] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates [Supporting Information]. *Energy & Environmental Science* **2018**, *11*, 617–628.

[4] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[5] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[6] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[7] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628.

[8] Varghese, J. J.; Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering* **2019**, *4*, 165–206.

[9] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Ma, K.; Elliott, W. A.; Bai, P.; Johnson, R. L.; Walker, T. W.; Shanks, B. H.; Rioux, R. M.; et al.. Effects of chloride ions in acid-catalyzed biomass dehydration reactions in polar aprotic solvents. *Nature communications* **2019**, *10*, 1–10.

[10] Mushrif, S. H.; Varghese, J. J.; Krishnamurthy, C. B. Solvation dynamics and energetics of intramolecular hydride transfer reactions

in biomass conversion. *Physical Chemistry Chemical Physics* **2015**, *17*, 4961–4969.

[11] Caratzoulas, S.; Vlachos, D. G. Converting fructose to 5-hydroxymethylfurfural: a quantum mechanics/molecular mechanics study of the mechanism and energetics. *Carbohydrate research* **2011**, *346*, 664–672.

[12] Tsilomelekis, G.; Josephson, T. R.; Nikolakis, V.; Caratzoulas, S. Origin of 5-hydroxymethylfurfural stability in water/dimethyl sulfoxide mixtures. *ChemSusChem* **2014**, *7*, 117–126.

[13] He, J.; Liu, M.; Huang, K.; Walker, T. W.; Maravelias, C. T.; Dumesic, J. A.; Huber, G. W. Production of levoglucosenone and 5-hydroxymethylfurfural from cellulose in polar aprotic solvent–water mixtures. *Green Chemistry* **2017**, *19*, 3642–3653.

[14] Patri, A. S.; Mostofian, B.; Pu, Y.; Ciaffone, N.; Soliman, M.; Smith, M. D.; Kumar, R.; Cheng, X.; Wyman, C. E.; Tetard, L.; et al.. A multifunctional cosolvent pair reveals molecular principles of biomass deconstruction. *Journal of the American Chemical Society* **2019**, *141*, 12545–12557.

[15] Mushrif, S. H.; Caratzoulas, S.; Vlachos, D. G. Understanding solvent effects in the selective conversion of fructose to 5-hydroxymethylfurfural: a molecular dynamics investigation. *Physical Chemistry Chemical Physics* **2012**, *14*, 2637–2644.

[16] Vermaas, J. V.; Petridis, L.; Ralph, J.; Crowley, M. F.; Beckham, G. T. Systematic parameterization of lignin for the CHARMM force field. *Green Chemistry* **2019**, *21*, 109–122.

[17] Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions. *ACS Catalysis* **2019**, *10*, 1679–1691.

[18] Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations. *Frontiers in chemistry* **2019**, *7*, 439.

[19] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **2019**, *10*, 370–377.

[20] Duvenaud, D.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. Advances in Neural Information Processing Systems 28. *Cortes C., Lawrence ND, Lee DD, Sugiyama M., Garnett R., Eds* **2015**, 2224–2232.

[21] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268–276.

[22] Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. Electronic structure at coarse-grained resolutions from supervised machine learning. *Science advances* **2019**, *5*, eaav1190.

[23] Lee, E. Y.; Fulan, B. M.; Wong, G. C.; Ferguson, A. L. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences* **2016**, *113*, 13588–13593.

[24] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

[25] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances* **2017**, *3*, e1603015.

[26] Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **2017**, *29*, 2352–2449.

[27] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*, 1097–1105.

[28] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al.. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.

[29] Singh, R. D.; Mittal, A.; Bhatia, R. K. 3D convolutional neural network for object recognition: a review. *Multimedia Tools and Applications* **2019**, *78*, 15951–15995.

[30] Torng, W.; Altman, R. B. High precision protein functional site detection using 3D convolutional neural networks. *Bioinformatics* **2019**, *35*, 1503–1512.

[31] Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042.

[32] Jiménez, J.; Skalic, M.; Martinez-Rosell, G.; De Fabritiis, G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling* **2018**, *58*, 287–296.

[33] Sedaghat, N.; Zolfaghari, M.; Amiri, E.; Brox, T. Orientation-boosted voxel nets for 3d object recognition. *arXiv preprint arXiv:1604.03351* **2016**.

[34] Maturana, D.; Scherer, S. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE; pp 922–928.

[35] Alonso, D. M.; Hakim, S. H.; Zhou, S.; Won, W.; Hosseinaei, O.; Tao, J.; Garcia-Negron, V.; Motagamwala, A. H.; Mellmer, M. A.; Huang, K.; et al.. Increasing the revenue from lignocellulosic biomass: Maximizing feedstock utilization. *Science advances* **2017**, *3*, e1603301.

[36] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[37] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[38] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[39] Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[40] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[41] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[42] Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, *14*, 33–38.

[43] Motagamwala, A. H.; Huang, K.; Maravelias, C. T.; Dumesic, J. A. Solvent system for effective near-term production of hydroxymethylfurfural (HMF) with potential for long-term process improvement. *Energy & Environmental Science* **2019**, *12*, 2212–2222.

[44] Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al.. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* **2016**.

[45] Sundararajan, M.; Taly, A.; Yan, Q. In *International Conference on Machine Learning*; PMLR; pp 3319–3328.

[46] Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808* **2018**.

[47] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.

[48] Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics* **2018**, *34*, 4046–4053.

[49] Sato, R.; Ishida, T. Protein model accuracy estimation based on local structure quality assessment using 3D convolutional neural network. *PloS one* **2019**, *14*, e0221347.

[50] Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; Lipson, H. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* **2015**.

[51] Ying, R.; You, J.; Morris, C.; Ren, X.; Hamilton, W. L.; Leskovec, J. Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804* **2018**.

[52] Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M. S. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval* **2018**, *7*, 87–93.

6 EFFECT OF MIXED-SOLVENT ENVIRONMENTS ON THE SELECTIVITY OF ACID-CATALYZED DEHYDRATION REACTIONS

The preceding chapters focused on using molecular dynamics simulations to predict the reactivity of biomass conversion reactions, which is useful for screening solvent systems for a series reaction, but these tools are not informative for reactions performing in parallel that result in multiple products. As a result, this chapter focuses on developing computational tools to estimate selectivity by analyzing solvation environments of both reactant and product states. This chapter seeks to answer the following questions:

- How can solvents affect the selectivity of parallel reactions?
- How do we estimate the selectivity *a priori* using molecular dynamics simulations?
- What are the limitations of using molecular dynamics simulations towards predicting reaction selectivity?

This chapter was reproduced with permission from Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions. *ACS Catalysis* **2019**, *10*, 1679–1691.[1] Copyright 2021 American Chemical Society. The electronic supporting information is cited as Ref. 2. A. K. Chew and T. W. Walker contributed equally to this work. T. W. Walker, B. Demir, L. Witteman, G. W. Huber, and J. A. Dumesic designed and performed the experimental work of this chapter. J. Euclide assisted with the ALAMO calculations. Z. Shen performed the *ab initio* calculations.

In this chapter, we report the kinetics and selectivity of Brønsted acid-catalyzed 1,2-propanediol dehydration in pure water and in aqueous mixtures of the polar aprotic cosolvents γ-valerolactone, 1,4-dioxane, tetrahydrofuran, N-methyl-2-pyrrolidone, tetramethylene sulfoxide, and dimethyl sulfoxide at 433 K. We find that the major product of 1,2-propanediol dehydration is propanal in most mixed-solvent environments with selectivities between 1-68 mol%. In contrast, 1,2-propanediol dehydration in aqueous mixtures of dimethyl sulfoxide affords acetone as the major product with up to 48% selectivity with minimal propanal formation. We use classical molecular dynamics simulations to probe these solvent effects by computing the difference between the solvation free energies of 1,2-propanediol and propanal in aqueous mixtures of polar aprotic cosolvents and in pure water. We find that the difference in the solvation free energies is correlated with the rates of propanal formation in all mixed-solvent environments, indicating that the solvent-mediated stabilization of the product state relative to the reactant state translates to increased selectivity toward the same product. Similar agreement between simulated solvation free energies and experimental reaction rates/selectivities is established for the acid-catalyzed dehydration of cis- and trans-1,2-cyclohexanediol and 1,3-cyclohexanediol. Finally, analysis of the solvation environment around 1,2-propanediol shows that dimethyl sulfoxide uniquely competes against water to solvate reactive hydroxyl groups, which causes a change in reaction mechanism in this solvent system that leads to the formation

of acetone rather than propanal. These results represent a step toward the computationally efficient screening of solvent systems for acid-catalyzed, liquid-phase processes.

6.1 Introduction

The liquid-phase catalytic conversion of lignocellulosic biomass to transportation fuels and valuable chemicals is a promising strategy for reducing global dependence on fossil resources.[3–8] However, achieving high selectivity toward desired products is a fundamental challenge inhibiting efficient biomass conversion. One strategy to improve the performance of liquid-phase catalytic processes is to modify the solvent composition, which can alter the rates and selectivities of the underlying chemical reactions.[9–11] In particular, mixtures of water with organic cosolvents (*i.e.*, mixed-solvent environments) have garnered interest for their ability to improve the yields of acid-catalyzed biomass conversion processes,[3,12] and a corresponding body of empirical knowledge has been accumulated regarding advantageous solvent compositions for specific applications.[13,14] However, it is not typically possible to anticipate how mixed-solvent environments will affect new processes *a priori*.[15,16] Therefore, solvent selection often requires trial-and-error experimentation or computationally expensive simulation methods.

We recently utilized classical molecular dynamics (MD) simulations

in combination with experimental reaction kinetics measurements to estimate the rates of acid-catalyzed reactions of biomass-analogous model oxygenates in mixtures of water and polar aprotic cosolvents,[17] based on insights obtained from prior experiments and *ab initio* MD simulations.[18] These simulation-derived estimates of reactivity translate into predictions of selectivity for series reactions in which an intermediate rather than a terminal product is desired. For example, the acid-catalyzed dehydration of fructose to afford 5-hydroxymethylfurfural (HMF), a platform chemical for fuel and other commodity chemicals production,[13,19–23] can be difficult to control due to the subsequent hydrolysis of HMF to form levulinic and formic acids.[24] Inclusion of dimethyl sulfoxide (DMSO) has been shown to increase HMF selectivity, in part by preventing the subsequent, undesirable hydrolysis step.[25] However, the reaction networks underlying liquid-phase processes are often comprised of reactions occurring both in series and in parallel. Therefore, it is desirable to understand how solvent composition affects not just the rates of acid-catalyzed reactions, but also the selectivity of reactions occurring in parallel.

The mechanisms by which solvent molecules influence acid-catalyzed reaction selectivity can be broadly understood in terms of solvent effects that we divide into three categories:[26–28] equilibrium solvation,[18,29–33] cosolvent participation,[25,34–36] and solvent dynamics.[18,33,34,37] Equilibrium solvation refers to the effect of the solvent environment on the stabilities (*i.e.*, relative free energies) of the reactant, transition, and product states.[3,30]

For example, polar aprotic cosolvents can lower the activation energy barriers for certain biomass conversion processes by stabilizing key transition states, resulting in enhanced reaction rates compared to the same reactions in pure water.[27,31,32] Cosolvent participation refers to the influence of cosolvent molecules on reaction selectivity by participating in reaction steps or by sterically "shielding" reactive sites around the reactant.[25,34] For example, classical MD simulations suggest that the competition between DMSO and water molecules for the hydroxyl groups of fructose molecules promotes conversion to HMF by protecting fructose from undesirable parallel reactions and shielding HMF from subsequent rehydration to levulinic acid.[25] Solvent dynamics are relevant when charged transition states polarize nearby solvent molecules and the rates at which these polarized solvent molecules relax towards more stable spatial and/or electronic configurations control the stability of the charged transition state. For example, the slow orientational relaxation of methanol molecules compared to water molecules increases the activation energy barrier for the isomerization of glucose to fructose by leading to poor solvation of the transition state.[34] These examples illustrate some of the processes by which the solvent environment can influence reaction selectivity. However, this mechanistic analysis has only been performed for a limited set of reactant-solvent combinations. Developing a quantitative understanding of solvent effects that can apply across a range of mixed-solvent environments and acid-catalyzed reaction mechanisms remains a challenge.

In this chapter, we investigate how the solvent composition can be systematically varied to influence the selectivity of model acid-catalyzed dehydration reactions to afford different products, and we explore how the stability and local solvent environments of reactants and products, as probed using classical MD simulations, can be used to anticipate these effects. We demonstrate that the selectivity of Brønsted acid-catalyzed 1,2-propanediol (PDO) dehydration varies in six different cosolvent-water mixtures with different mass fractions of cosolvent, affording low-to-moderate yields of propanal (PRO) or acetone (ACE) as the major product. We find that ACE is formed as the major product in aqueous solutions of DMSO, whereas all other solvent environments produce PRO over ACE with the rate of PRO formation depending on the specific cosolvent composition. Using experimental reaction kinetics measurements, we demonstrate that these solvent-induced changes in reaction selectivity can be understood in terms of the rate of PRO formation, relative to ACE, as a function of solvent composition. Furthermore, we find that MD-derived reactant and product solvation free energies, which quantify the effect of equilibrium solvation on selectivity, can explain these solvent-mediated changes to the rate of PRO formation. This finding generalizes to reaction rate trends for other representative diols in 90 wt% GVL- and DMSO-water mixtures. However, equilibrium solvation calculations do not explain the formation of ACE in DMSO-water mixtures. Analysis of the unique spatial distribution of solvent molecules around PDO in DMSO-water mixtures

instead suggests that DMSO competes with water for reactive sites which may lead to changes in the reaction mechanism to favor the production of ACE via a cosolvent participation effect. Supporting this hypothesis, density functional theory calculations indicate that the local solvent environment in DMSO-water mixtures preferentially stabilizes the protonated primary hydroxyl group of PDO, which is consistent with a mechanism that affords ACE. These results demonstrate how the combination of experimental reaction kinetic measurements, classical MD simulations, and *ab initio* calculations can generate insight into the effect of mixed-solvent environments on the selectivity of acid-catalyzed reactions. Furthermore, these results represent a step toward anticipating how solvent composition affects the selectivity of reactions occurring in parallel using computationally efficient simulation methods capable of analyzing a range of solvent systems.

6.2 Methods

6.2.1 Reaction kinetics studies

Reactions were carried out in closed, 10 mL thick-walled glass reactors. In a typical experiment, an appropriate amount of reactant (*i.e.* PDO), acid catalyst, water and organic cosolvent (*e.g.* DMSO) were charged to the reactors, which were then sealed and placed in an oil bath at 433 K. Reactors were removed at the desired reaction time and quenched in

an ice bath at 273 K. The conversion of the reactant was monitored as a function of reaction time using a Shimadzu high-performance liquid chromatograph equipped with a differential refractometer. Products were quantified using a Shimadzu gas chromatograph equipped with a flame ionization detector. All products were quantified using calibration curves with external standards. Rate constants for the conversion of reactants and the formation products were derived from Equation 6.1 using the method of initial rates. Trifluoromethane sulfonic (triflic) acid (pK_a, H_2O = 14.7, pK_a, DMSO = -14.3) was used as catalyst in all experiments. Triflic acid behaves as a strong acid even in pure polar aprotic solvents.[18,38,39] We thus assume complete dissociation of the acidic proton in all mixed-solvent environments, allowing for normalization of the apparent rate constants on a per-proton basis.

The sum of reaction selectivities is less than 100% in all solvent systems displayed in Figure 6.2. This is due to the formation of a range of condensation products, such as 2-ethyl-4-methyl-1,3-dioxolane, 2-methoxy-1,3-dioxolane, 4-methyl-1,3-dioxolane and 1,3-dioxolane. In our prior work, we have also detected the presence of multiple degradation products such as 3-hydroxyl-2-methyl-pentanal, 2-ethyl-2-butenal, heptanal and 2-(1-methylethoxy)-1-propanol using gas chromatography-mass spectrometry analysis.[18] However, due to difficulties in isolating and quantifying these species, their formation rates were not investigated in this study.

6.2.2 Molecular dynamics simulations

All classical molecular dynamics simulations were performed using Gromacs version 2016.[40] Reactant, products, and cosolvents were parameterized using the CGenFF/CHARMM36 force fields.[41–43] Water was modeled using the Single Point Charged/Extended (SPC/E) model.[44] We initialized simulations of mixed-solvent environments following the protocol discussed in our previous work.[17,45] The initial simulation box dimensions were set to (6 nm)^3 in all simulations and water and cosolvent molecules were added at the desired composition. The system was equilibrated in a *NPT* simulation for 5 ns at 300 K using a velocity-rescale thermostat and 1 bar using a Berendsen barostat.

Solvation free energies were computed from a series of stochastic dynamics simulations, as described in Section 2.1.2. For each simulation, the system was energy minimized with the steepest descent algorithm and equilibrated for 100 ps at constant temperature followed by 2 ns at constant temperature and constant pressure using the Berendsen barostat. An 11-ns production simulation at constant temperature and pressure was then performed with the Parrinello-Rahman barostat. All simulations were performed at 433.15 K and 1 bar. Conformations of the reactants and products were not restrained during the solvation free energy calculations. Energy differences computed between all pairs of windows were collected every 0.2 ps and solvation free energies were computed with the Multistate Bennett Acceptance Ratio[46] method using the python alchemical-analysis

tool.[47] The last 10 ns of each production simulation were split into two 5 ns trajectories and treated as two independent trials. All solvation free energy results and error bars are reported as the average and standard deviation of the two trials, respectively.

Spatial distribution maps were generated using simulations with the reactant restrained and the solvent free to explore reactive sites. A single reactant was added to an equilibrated solvent system and then the system was equilibrated again for 500 ps at 433.15 K using a velocity-rescale thermostat and 1 bar using a Berendsen barostat. The system was simulated for 300 ns for PDO and 50 ns for *cis*- and *trans*-isomers of 1,2-cyclohexanediol at the same temperature and pressure, controlled by the Parrinello-Rahman barostat and Nóse-Hoover thermostat, respectively. We extracted the most likely configuration of the reactant from unbiased simulations and restrained the reactant's atomic positions with a force constant of 10,000 kJ/(mol-nm) to eliminate rotational degrees of freedom (ESI, Section S7).[2] We then performed an additional 200 ns simulation with the restrained reactant in solution, where the final 190 ns of the trajectory was used to calculate spatial distribution maps. Simulation analysis was performed using the MDTraj package[48] and analysis tools developed in-house.

6.2.3 *Ab initio* calculations

All *ab initio* calculations were carried out using Gaussian16.[49] For calculations with a single protonated PDO molecule (HPDO), geometry optimizations were performed at the unrestricted B3LYP,[50–52] PBE[53,54] and MP2[55–59] levels of theory in vacuum with the 6-311++G(d) basis set to compare results obtained using multiple levels of theory. These structures were subsequently optimized using the conductor-like polarizable continuum model (CPCM) with either water or DMSO as an implicit solvent.[60,61] For calculations with HPDO and explicit DMSO and/or water molecules, all geometry optimizations were performed with the PBE functional and 6-311++G(d) basis set, followed by subsequent optimization with either DMSO or water as an implicit solvent model using CPCM. Frequency analyses were performed to ensure true energy minima.

6.3 Results and Discussion

6.3.1 Proposed reaction mechanism for the acid-catalyzed dehydration of 1,2-propanediol

Figure 6.1A depicts a possible mechanism for the Brønsted acid-catalyzed dehydration of PDO in the gas phase over an acid catalyst.[62] Depending upon which hydroxyl moiety is removed in the form of water, this reaction affords either PRO or ACE as the major product, passing through

a secondary- or primary-carbocation-like intermediate, respectively. If subsequent conversion of the products is neglected, then the selectivity of this reaction depends upon the relative rates at which corresponding products are formed; or, equivalently, upon the relative energy differences between the reactant, transition, and product states. Accordingly, PRO is the major observable product in the gas phase because a secondary carbocation is more stable than a primary carbocation.[63] However, liquid solvents can alter the reaction mechanism by stabilizing structures that may be unstable in the gas phase.[27–29] Liquid-phase reactions may also proceed through a different mechanism, such as a reaction mechanism that is more concerted rather than stepwise, or a reaction mechanism with intermediates that are chemically distinct from those shown in Figure 6.1A. Following these insights, we now explore whether the composition of mixed-solvent environments can be tuned to modulate the relative rates of PDO dehydration to selectively afford either PRO or ACE. We examine the effects of mixed-solvent environments composed of water and one of six polar aprotic cosolvents; their chemical structures are shown in Figure 6.1B.

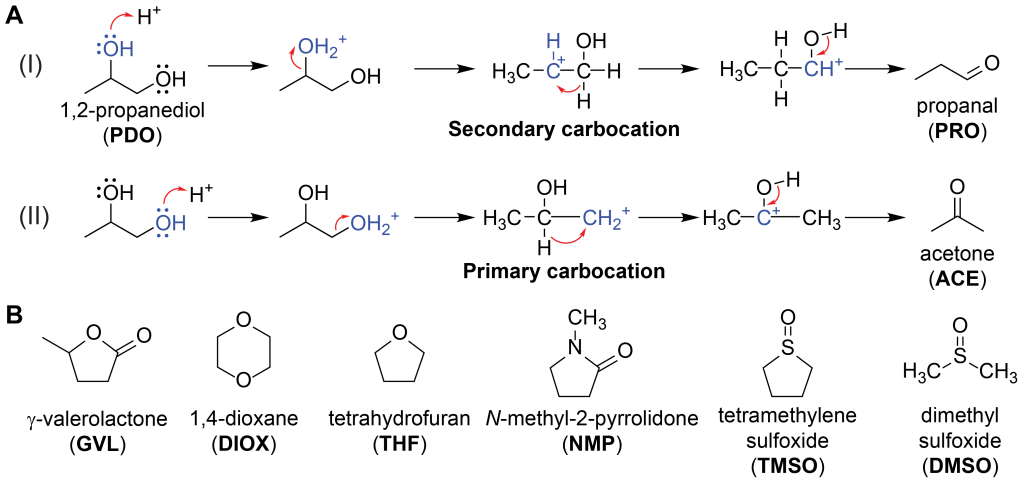


Figure 6.1: **A** Proposed mechanism for Brønsted acid-catalyzed dehydration of 1,2-propanediol (PDO) to afford either propanal (PRO) or acetone (ACE) in the gas phase over a solid acid catalyst.[62] Red arrows denote the transfer of electrons. **B** Polar aprotic cosolvents used for this study.

6.3.2 Product selectivities for the acid-catalyzed dehydration of 1,2-propanediol

We carried out PDO dehydration in batch reactors over a triflic acid catalyst and monitored reactant conversion as a function of time. The kinetics of PDO dehydration have been shown to be first order with respect to the concentration of the reactant and acidic protons in solution.[18] Accordingly, Equation 6.1 relates the initial rate of reactant conversion (r_i) to the molar concentration of reactant i (C_i), the molar concentration of acidic protons (C_{H^+}) in solution, and the apparent first-order rate constant (k_{app}) at a

fixed solvent composition and temperature:

$$r_i = -\frac{dC_i}{dt} = k_{app} C_i C_{H^+} \tag{6.1}$$

Simultaneously, the formation of product p (*i.e.*, PRO or ACE) was monitored as a function of time, and the selectivity to each product (S_p) was quantified as the ratio of the initial rate of product formation (r_p) to the initial rate of reactant conversion (r_i) or, equivalently, as a ratio of rate constants (Equations 6.2-6.3):

$$r_p = \frac{dC_p}{dt} = k_p C_i C_{H^+} \tag{6.2}$$

$$S_p = \frac{r_p}{r_i} = \frac{k_p}{k_{app}} \tag{6.3}$$

The rate constant for the formation of each product is therefore obtained by multiplying k_{app} by the selectivity to either ACE or PRO.

Figure 6.2 shows the selectivities and apparent rate constants for PDO dehydration in water and in aqueous mixtures with DIOX or DMSO. The reaction selectivities do not sum to one due to the formation of condensation products that are difficult to quantify (see Methods); thus, we focus on the effect of the solvent composition on the formation of PRO and ACE. In pure water, PRO is formed with 41% selectivity on a molar basis and ACE is formed with 9% selectivity. In both solvent systems, the value of the apparent rate constant for PDO conversion increases monotonically

with the mass fraction of the organic phase, an effect that was explored in prior work.[17,18] In 25 wt% DIOX, the selectivity to ACE decreases to zero and the selectivity to PRO decreases to 15% (Figure 6.2A). Upon further addition of DIOX, the selectivity to PRO partially recovers, passing through a maximum of 30% at 90 wt% DIOX. The selectivity to PRO trends towards zero for mass fractions of DIOX above 90 wt%, indicating that water might play a role in the reaction mechanism that affords this product. This observation is consistent with prior studies indicating that protic solvents such as water can facilitate acid-catalyzed reaction mechanisms by stabilizing carbocation-like intermediates.[18,34] In contrast, almost no PRO is formed in 25 wt% DMSO, whereas the selectivity to ACE increases to 49% (Figure 6.2B). The further addition of DMSO results in a monotonic decrease in the selectivity to ACE, again indicating that this reaction is at least partially facilitated by water.

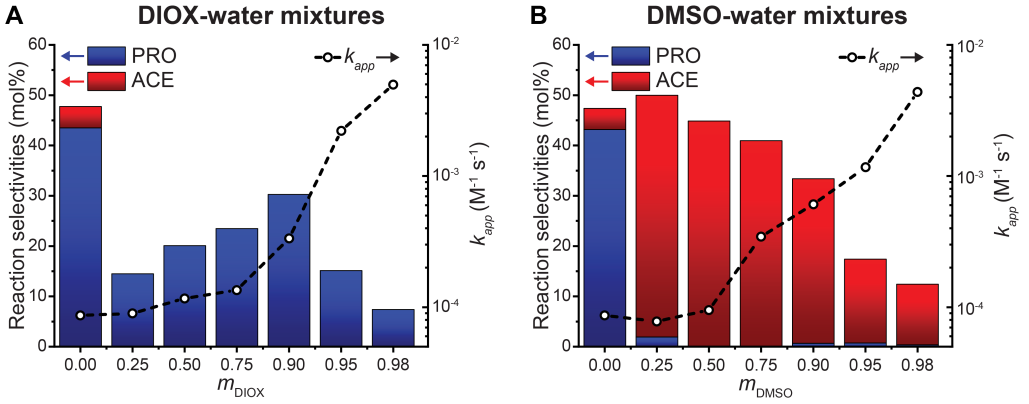


Figure 6.2: Apparent rate constants (k_{app}, dashed lines) and selectivities to propanal (PRO, blue columns) and acetone (ACE, red columns) for Brønsted-acid-catalyzed 1,2-propanediol (PDO) dehydration in mixtures of water with (**A**) 1,4-dioxane (DIOX) and (**B**) dimethyl sulfoxide (DMSO) as a function of the mass fraction of the organic component (m_{DIOX} or m_{DMSO}). Rate constants are derived from Equation 6.1 and selectivities are derived from Equation 6.3. The standard error in selectivities is $\pm$ 5 mol%. *Reaction conditions: ~20 mol% PDO conversion; 433 K; 0.4-0.005 M triflic acid; 0.01 M PDO; 90-150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.*

To understand the trends in selectivity, we probe the role of DMSO water mixtures in altering the relative rates of PRO and ACE formation. Figure 6.3 shows the apparent rate constants for PDO consumption, PRO formation, and ACE formation as a function of solvent composition in DMSO-water mixtures. As the mass fraction of DMSO is increased from 7 to 25 wt%, the overall rate of PDO conversion remains roughly constant, while the rate of ACE formation increases by a factor of four. In contrast, the rate of PRO formation decreased by a factor of more than 30 in this

same mass fraction range. As the mass fraction of DMSO is increased from 25 to 90 wt%, the rate constants for PDO consumption and the formation of both products generally increase together.

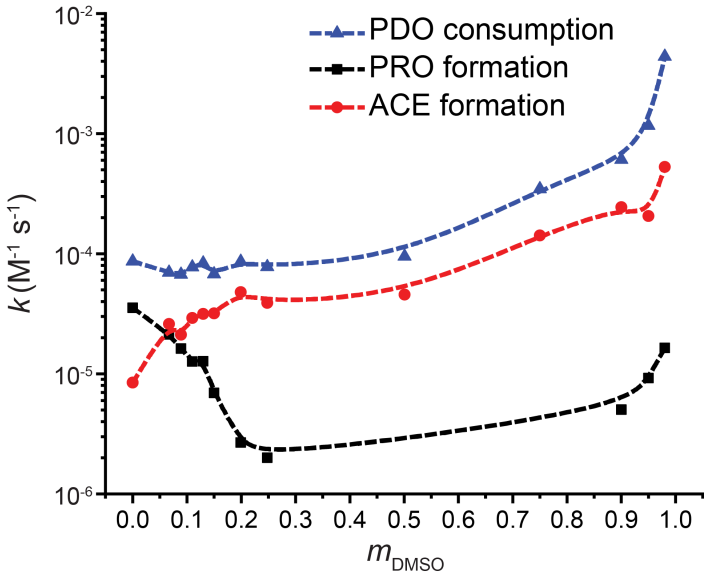


Figure 6.3: Apparent rate constants for reactant conversion and product formation for 1,2-propanediol (PDO) dehydration in mixtures of water with dimethyl sulfoxide (DMSO) as a function of the mass fraction of DMSO (m_{DMSO}). Dashed lines are visual aids. *Reaction conditions: ~20 mol% conversion; 433 K; 0.4-0.005 M triflic acid; 0.1 M PDO; 150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.*

6.3.3 Relating selectivity trends to tabulated cosolvent properties

The results shown in Figure 6.3 suggest that DMSO suppresses the rate of PRO formation, or equivalently, increases the activation energy bar-

rier for PRO formation.[64] We sought to determine if this behavior could generalize to other cosolvents based on the hypothesis that the polarizability or dielectric constant of a solvent could be related to activation energy barriers that lead to a specific product.[28] Previous work has also shown that more nucleophilic (basic) or polar solvents, such as DMSO, can facilitate distinct reaction mechanisms compared to pure water.[65,66] We thus carried out the acid-catalyzed dehydration of PDO in four other cosolvent-water mixtures at 90 wt% cosolvent to determine if there is a relationship between selectivity and cosolvent polarity or basicity. Selected cosolvents include γ-valerolactone (GVL), tetrahydrofuran (THF), *N*-methyl-2-pyrrolidone (NMP), and tetramethylene sulfoxide (TMSO) because these cosolvents have dielectric constants that lie between those of DIOX (2.20) and DMSO (46.46) (Table 6.1). To facilitate comparison of relative reaction rates between mixed-solvent environments, Equation 6.4 defines the product kinetic solvent parameter (σ^p, with p indicating a specific product) based on the rate of product formation in the mixed-solvent environment (k^p_{org}) and the rate of product formation in pure water ($k^p_{H_2O}$):[17]

$$\sigma^p = \log_{10}\left(\frac{k^p_{org}}{k^p_{H_2O}}\right) \tag{6.4}$$

$\sigma^p > 0$ indicates increased product formation in a particular mixed-solvent environment compared to the same reaction in pure water, while $\sigma^p < 0$ has the converse implication.

Table 6.1 presents apparent rate constants for acid-catalyzed PDO dehydration, selectivities to PRO and ACE, and σ^{PRO} for each of the cosolvent-water mixtures. Entries in Table 6.1 are organized in descending order based on σ^{PRO}. Table 6.1 also displays the dielectric constant[67–70] and basicity of the cosolvents, as expressed by the pK_a's of their protonated conjugate acids (pK_a^*).[71,72] Larger dielectric constants indicate a more polar solvent and pK_a^* values below -1.71 indicate solvents that are more basic than pure water. With the exception of DMSO-water mixtures, PRO is the preferred product in all solvent systems, as expected based on analysis of the reaction mechanism presented in Figure 6.1.[63,73] Furthermore, neither the product selectivities nor σ^{PRO} correlate with the dielectric constant or the basicity of the organic cosolvent. Finally, we found no correlation between σ^{PRO} and Kamlet-Taft parameters (π^*, β, α),[74] which are commonly used to characterize solvent properties, for the cosolvents listed in Table 6.1 (ESI, Figure S3).[2] These observations illustrate that properties of mixed-solvent environments cannot generally be understood in terms tabulated descriptors of the pure individual components.

Table 6.1: Apparent rate constants (k_{app}), product kinetic solvent parameters (σ^{PRO}), and selectivities of Brønsted acid-catalyzed 1,2-propanediol (PDO) dehydration in mixed-solvent environments. The cosolvent mass fraction was 90 wt% in all experiments (excluding pure water). Rate constants are derived from Equations 6.1 and 6.2. Selectivities are derived from Equation 6.3. The standard error in selectivities is $\pm$ 5 mol%. *Reaction conditions: ~20 mol% conversion; 433 K; 0.4 M triflic acid; 0.1 M PDO; 150 min reaction time; 500 rpm stirring rate, 2 mL total solvent volume.*

Cosolvent	pK_a^{*a}	Dielectric constant[b]	k_{app} $(M^{-1}ks^{-1})$	Conversion (mol%)	S_{PRO}	S_{ACE}	σ^{PRO}
GVL	-7.00	36.47	0.880	55	68	0	1.24
DIOX	-3.00	2.20	0.336	12	34	0	0.52
THF	-2.05	7.40	0.434	23	8	0	0.48
NMP	-0.71	32.16	0.084	24	52	1	0.09
TMSO	N/A	42.84	0.281	14	3	0	-0.44
DMSO	-1.80	46.46	0.670	37	1	33	-0.57
Pure water	-1.71	80.10	0.084	9	41	9	N/A

*basicity is quantified as the pK_a of the solvent's protonated conjugate acid. For example, -2.05 is the pK_a of the protonated ether group on THF.

[a] pK_a values are from Ref. 71 and 72.

[b] Dielectric constants are from Ref. 67, except for GVL,[69] NMP,[70] and TMSO.[68]

6.3.4 Equilibrium solvation: quantifying reactant and product solvation free energies

Because the selectivity of PDO dehydration in mixed-solvent environments cannot be rationalized by the tabulated bulk properties of the pure cosolvents, we use classical MD to investigate mechanisms by which mixed-solvent environments might influence selectivity to ACE and PRO. We first hypothesize that solvent relaxation is fast relative to the timescales associated with reaction elementary steps such that the activation energy for the formation of a specific product is correlated with the difference between the equilibrium free energies of the reactant and product. Similar hypotheses, all of which follow from the Hammond postulate, lead to the typical linear free energy relationships that relate changes in reaction rates to changes in the energy of reactant and product states in catalytic processes.[75–77] This hypothesis is equivalent to assuming that equilibrium solvation effects dictate selectivity and suggests that changes in selectivity can be related to the free energies of the reactants and products without knowledge of the reaction mechanism or explicitly modeling the transition state itself. Figure 6.4A schematically illustrates the hypothesized effect of a mixed-solvent environment on a reaction free energy landscape that would influence selectivity to a particular product. Furthermore, *ab initio* MD simulations have shown that the PDO dehydration mechanism that forms PRO is characterized by a late transition state[18] that is analogous to

the late transition state formed in the acid-catalyzed conversion of fructose to HMF.[29] This finding emphasizes the importance of quantifying product stability because in many instances the structure of a late transition state more closely resembles the product than the reactant.

To investigate equilibrium solvation effects, we leverage the computational efficiency of classical MD simulations to quantify the solvation free energy of each product and reactant in each mixed-solvent environment studied experimentally. The solvation free energy of a species is defined as the change in free energy associated with transferring the species from vacuum to solvent; the difference in the solvation free energy of a species in two solvent environments is the change in the free energy associated with transferring the species from one solvent environment to the other. Figure 6.4B illustrates the thermodynamic cycle used to determine the influence of the solvent environment on the reactant and product free energies. The free energy of transferring species s from water to a mixed-solvent environment ($\Delta G_s^{H_2O \rightarrow org}$) is related to the solvation free energy of s in pure water ($G_s^{H_2O}$) and the solvation free energy of s in a mixed-solvent environment (G_s^{org}) by Equation 6.5.

$$\Delta G_s^{H_2O \rightarrow org} = G_s^{org} - G_s^{H_2O} \tag{6.5}$$

A negative value of $\Delta G_s^{H_2O \rightarrow org}$ indicates that s is more favorably solvated in the mixed-solvent environment than in pure water. We then quantify

the free energy difference between the reactant (r) and product (p) in a mixed-solvent environment ($\Delta\Delta G$) by performing four solvation free energy calculations that are related by Equation 6.6.

$$\begin{aligned} \Delta\Delta G &= \Delta G_p^{H_2O \rightarrow org} - \Delta G_r^{H_2O \rightarrow org} \\ &= (G_p^{org} - G_r^{org}) - (G_p^{H_2O} - G_r^{H_2O}) \end{aligned} \tag{6.6}$$

$\Delta\Delta G$ quantifies the Gibbs free energy difference between the reactant and product in the mixed-solvent environment ($G_p^{org} - G_r^{org}$) relative to the same free energy difference in a pure water reference state ($G_p^{H_2O} - G_r^{H_2O}$). If $\Delta\Delta G < 0$, the difference between the free energy of the product and the reactant is more negative in the mixed-solvent environment than in pure water, indicating that the product is stabilized to a greater degree than the reactant (Figure 6.4A). Therefore, we hypothesize that a negative value of $\Delta\Delta G$ for a product in a given mixed-solvent environment should increase its rate of formation compared to pure water due to equilibrium solvation effects, and conversely a positive value of $\Delta\Delta G$ for a product should decrease its rate of formation. This hypothesis is qualitatively consistent with the free energy landscapes previously computed in mixed-solvent environments for *tert*-butanol and PDO dehydration using *ab initio* MD simulations.[18] We emphasize that $\Delta\Delta G$ quantifies changes to the relative stabilities of the reactant and products in mixed-solvent environments and does not directly report the relative stabilities of the reactant and products

in pure water. Because the experiments (Figure 6.2) indicate that PRO is preferentially formed in pure water, we assume that this product is more favorable according to the equilibrium solvation hypothesis.

Figure 6.4C plots $\Delta\Delta G$ for PRO and ACE against experimentally determined values of σ^{PRO} in mixed-solvent environments. $\Delta\Delta G$ for PRO formation is negative for aqueous mixtures containing 90 wt% GVL, DIOX, THF, or NMP, indicating that PRO formation should be enhanced in these mixed-solvent environments compared to pure water. For these same mixtures, σ^{PRO} is positive, indicating that experimental reaction rates for PRO formation are enhanced relative to pure water and agreeing with the simulation predictions. Conversely, $\Delta\Delta G > 0$ and $\sigma^{PRO} < 0$ for 90 wt% TMSO and DMSO mixtures, indicating that in these systems equilibrium solvation effects predict the suppression of PRO formation. Figure 6.4D plots the correlation between $\Delta\Delta G$ and σ^{PRO} for each of the solvent mixtures from Figure 6.4C and for aqueous mixtures containing 25, 50, and 75 wt% DIOX. As hypothesized, $\Delta\Delta G$ and σ^{PRO} exhibit a negative linear correlation (Pearson's $r = -0.81$) with a root-mean-squared error (RMSE) of 0.33 between values of σ^{PRO} predicted using the linear correlation and experimentally determined values. These results are consistent with the hypothesis that equilibrium solvation can account for the effect of mixed-solvent environments on the rate of PRO formation. However, $\Delta\Delta G$ does not capture the non-monotonic behavior found in σ^{PRO} across different mass fractions of DIOX and DMSO (ESI, Section S6.1)[2] and $\Delta\Delta G$

for PRO is more negative than ACE for all solvents. Because PRO is the preferred product in pure water (Figure 6.2), the more negative values of $\Delta\Delta G$ for PRO than ACE in all other solvent systems studied indicate that PRO should always be the preferred product compared to ACE. These data thus do not explain the experimental finding that ACE is preferentially formed over PRO in DMSO-water mixtures, suggesting that the formation of ACE in the presence of DMSO is due to alternative solvent effects as will be further discussed below.

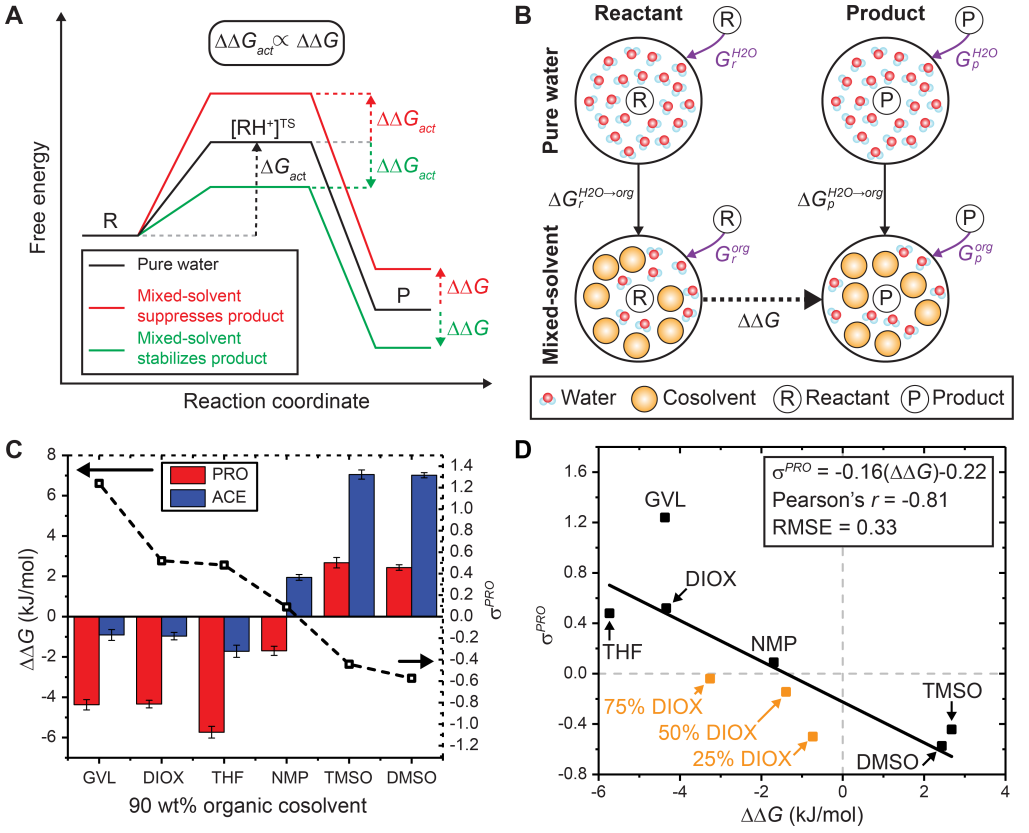


Figure 6.4: **A** Hypothesized effect of mixed-solvent environment on the free energies of reactant, transition, and product states. The change in the relative free energy between the reactant and product states (ΔΔG) is proportional to the change in the activation energy ($\Delta\Delta G_{act}$) for a reaction in a mixed-solvent environment compared to the same reaction in pure water. Free energies are schematically drawn relative to the free energy of the reactant state in each solvent environment. **B** Thermodynamic cycle to calculate the free energy difference between a reactant and product in a mixed-solvent environment relative to pure water. Purple arrows indicate solvation free energies computed from MD simulations which are used to calculate the transfer free energies indicated by filled black arrows. The dashed black arrow indicates ΔΔG. **C** Simulated ΔΔG for PRO (red bar) and ACE (blue bar) and experimental kinetic solvent parameter for PRO formation (σ^{PRO}, dashed black lines) in 90 wt% organic cosolvents. **D** Correlation between σ^{PRO} and ΔΔG for 90 wt% mass fraction of organic solvent (black) and various wt% mass fractions of DIOX (orange).

6.3.5 Equilibrium solvation effects extended to diol dehydration

To determine if ΔΔG can quantify product formation for different dehydration reactions, we performed the acid-catalyzed dehydration of three other representative diols in 90 wt% GVL- and DMSO-water mixtures: *cis*-1,2-cyclohexanediol, *trans*-1,2-cyclohexanediol, and *trans*-1,3-cyclohexenediol; dehydration reaction schemes are shown in Figure 6.5A. GVL and DMSO were selected as cosolvents because they correspond to the highest and lowest rates of PRO formation from PDO dehydration (Table 6.1). Table 6.2 shows the rates and the selectivities of these reactions towards the dehydration products in 90 wt% GVL- and DMSO-water mixtures over a triflic acid catalyst at 433 K. For all reactions, we find that the selectivity to the dehydration products in Table 6.2 is higher in 90 wt% GVL than 90 wt% DMSO and ΔΔG is more negative in 90 wt% GVL than in 90 wt% DMSO. Figure 6.5B shows σ^p (where p refers generically to the dehydration product of each reaction) and ΔΔG for each different reaction in 90 wt% GVL- and DMSO-water mixtures. σ^p and ΔΔG are again inversely related for each separate reaction, agreeing with the results for PDO dehydration (Figure 6.4D). This result indicates that reactant and product solvation free energies in mixed-solvent environments can provide insight into the selectivity for a dehydration product.

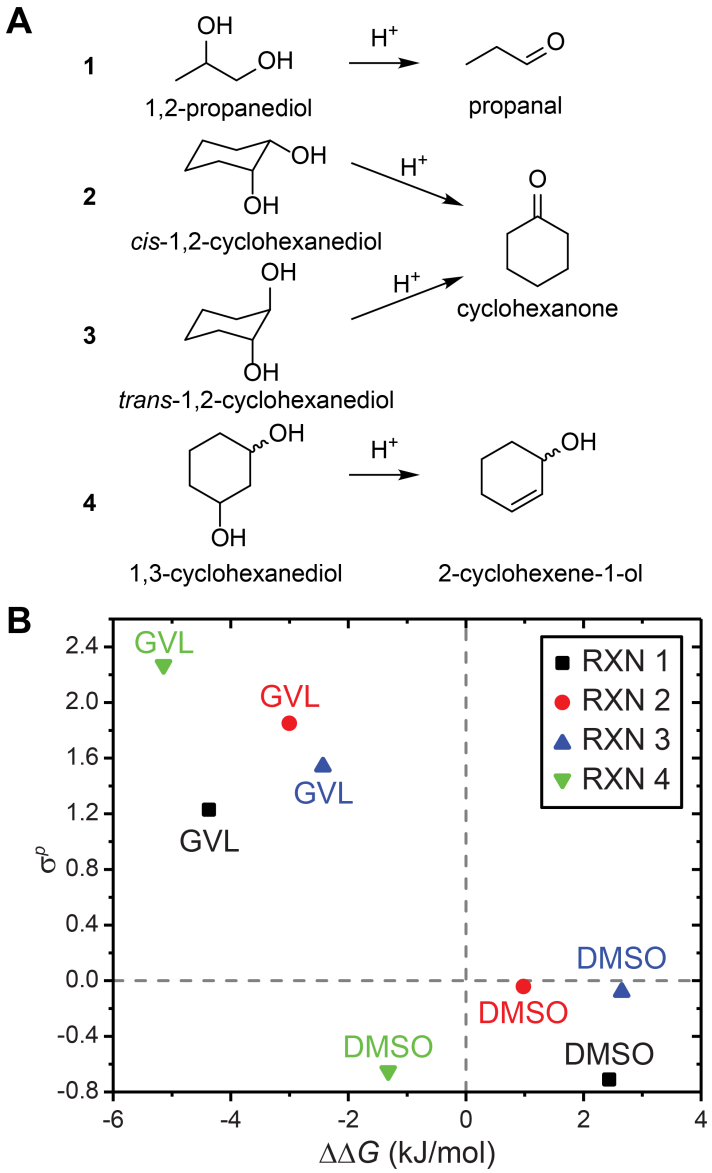


Figure 6.5: **A** Four acid-catalyzed dehydration reactions of representative diols. **B** Experimental kinetic solvent parameter for dehydration product formation (σ^p) and $\Delta\Delta G$ for each of the four reactions in 90 wt% GVL- and DMSO-water mixtures.

Table 6.2: Dehydration of diols over triflic acid in 90 wt% GVL-water and DMSO-water mixtures. *Reaction conditions: 0.1 M reactant, 433 K, 0.05-0.6 M triflic acid, 500 rpm stirring, reaction time 0-120 min.*

Reaction	Solvent	k_{app} ($M^{-1}ks^{-1}$)	Conversion (mol%)	Selectivity[a] (mol%)	σ^p	$\Delta\Delta G$
1	GVL	0.88	55	68	1.23	-4.37
	DMSO	0.67	37	1	-0.71	2.44
2	GVL	45.74	56	70	1.85	-3.01
	DMSO	1.52	31	27	-0.04	0.98
3	GVL	3.13	18	3	1.54	-2.43
	DMSO	0.34	11	1	-0.08	2.65
4[b]	GVL	68.19	33	99	2.27	-5.14
	DMSO	0.32	44	25	-0.65	-1.32

[a] dehydration products shown in Figure 6.5A.

[b] consists of equal *cis* and *trans* isomers of 1,3-cyclohexanediol. The $\Delta\Delta G$ value is the average result from both isomers.

6.3.6 Competition of DMSO for hydroxyl groups on reactant

Figure 6.4C shows that $\Delta\Delta G$ is more negative for PRO than ACE in all solvent mixtures, which indicates that equilibrium solvation effects can explain trends underlying PRO formation in multiple solvent systems but cannot explain the formation of ACE in DMSO-water mixtures. We thus hypothesize that DMSO molecules can participate directly in the reaction mechanism to promote ACE formation. The unique ability of DMSO to

mediate distinct reaction mechanisms compared to other polar aprotic solvents has been previously reported. For instance, *ab initio* molecular dynamics simulations have shown that DMSO can act as a proton acceptor to mediate the proton transfer steps in glucose dehydration reactions.[78] As noted previously, DMSO is also thought to promote fructose conversion to HMF by shielding HMF from rehydration.[25] These processes require that DMSO molecules localize near the reactant in mixed-solvent environments.

Figure 6.6 shows spatial distribution maps that quantify the density of water and either DIOX or DMSO molecules in the three-dimensional volume near PDO. Densities are normalized by the density of the bulk solvent. Spatial distribution maps are computed for PDO in pure water, 90 wt% DIOX, and 90 wt% DMSO and show normalized density values between 1.5-3.0 for water (shown in red) and 1.3-1.5 for cosolvent (shown as blue). These values were selected to show spatial regions enriched in either water or cosolvent molecules relative to the composition of the bulk solvent. Spatial distribution maps are shown for both the *trans* conformation (Figure 6.6A) and *gauche* conformation (Figure 6.6B) of PDO vicinal diols. The *trans* conformation is the most probable conformation obtained in unbiased MD simulations in all solvent systems, while the *gauche* conformation is obtained in ~20-40% of all simulation configurations (ESI, Section S7.1)[2] and thus still frequently occurs. Supplementary Movie S1 further shows several alternative rotations of the spatial distribution maps.[2]

Focusing first on the *trans* conformation, Figure 6.6A shows that in pure water there is slight enrichment of water around the hydroxyl groups of PDO, indicating a near-uniform distribution of water molecules. In 90 wt% DIOX, there are two water-enriched regions associated with each hydroxyl moiety, in agreement with our previous work in which water-enriched local domains were shown to form around hydrophilic reactants in the presence of DIOX.[17] These regions correspond to positions occupied by water molecules that act as both hydrogen bond donors and acceptors for the PDO hydroxyl groups. Conversely, in 90 wt% DMSO, there is enrichment of both water and DMSO molecules around the hydroxyl groups. Importantly, the water-enriched regions corresponding to water molecules that donate hydrogen bonds to the hydroxyl groups are missing. Figure S9 and Supplementary Movie S2 show spatial distribution maps for 90 wt% mixtures of the other cosolvents listed in Table 6.1;[2] DMSO is the only cosolvent that strongly competes with water for reactive sites around PDO. Figure 6.6B shows that similar behavior is observed for the *gauche* conformation in pure water and in 90 wt% DIOX, with water-enriched regions observed near the hydroxyl groups. Three, rather than four, water-enriched regions are present because one region contains water molecules that act as both hydrogen bond donors and acceptors to the two hydroxyl groups; minimal intramolecular hydrogen bonding between the vicinal diols occurs (ESI, Table S5).[2] DMSO again competes with water molecules in regions near the hydroxyl groups. Notably, a water-enriched region

corresponding to water molecules that accept hydrogen bonds from the primary hydroxyl group is present, but a similar region near the secondary hydroxyl group is diminished.

The spatial distribution maps indicate that DMSO exhibits unique solvation behavior among all studied cosolvents by occluding regions containing water molecules that act as hydrogen bond donors, which would correspond to the protonation site if an acidic proton were transferred from a hydronium ion to the reactant.[25] DMSO itself can act as a base[79] and the stability of charged intermediates in base-catalyzed reaction mechanisms follow opposite trends compared to their acid-catalyzed counterparts (*i.e.*, a primary carbanion is relatively stable compared to a primary carbocation).[66] We thus hypothesized that DMSO may suppress PRO formation from PDO by occluding access of the hydronium ion to the hydroxyl groups of the reactant, while simultaneously acting as a base to remove the primary OH group in the form of water. To probe this hypothesis, we experimentally studied PDO dehydration in 90 wt% DMSO-water mixtures in the absence of triflic acid catalyst and found that PDO underwent no reaction. Thus, we reject the hypothesis that DMSO is acting as a base catalyst to afford ACE from PDO.

We next probed the generality of the influence of DMSO beyond PDO by experimentally characterizing the products of 1-butanol (with one primary hydroxyl group) and glycerol (with three hydroxyl groups) dehydration at 433 K over a triflic acid catalyst in 90 wt% DMSO-water mixtures.

No dehydration product (butene) was formed from 1-butanol. Conversely, glycerol dehydration afforded hydroxyacetone, a primary dehydration product analogous to ACE in PDO dehydration, with 97% selectivity (ESI, Section S3).[2] We further found that glycerol dehydration afforded hydroxyacetone with only 3% selectivity in pure water and no hydroxyacetone was formed in 90 wt% GVL-water mixtures. These results indicate that vicinal diols may promote dehydration for the linear alcohols considered.

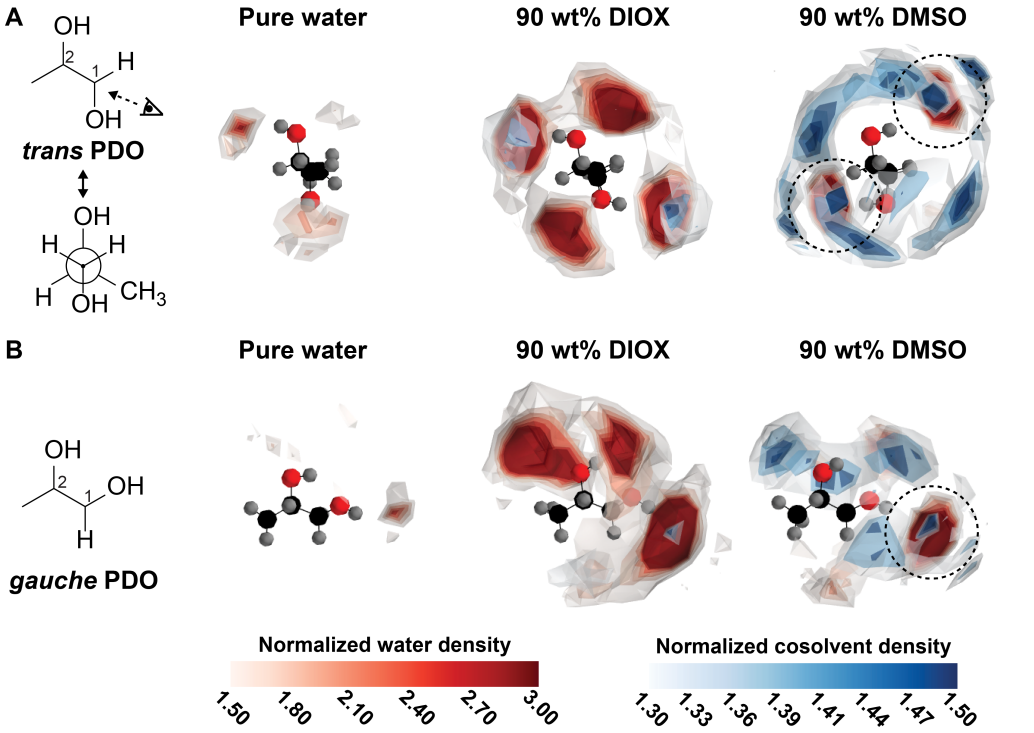


Figure 6.6: **A** Spatial distribution maps of the *trans* vicinal diol conformation of PDO in pure water, 90 wt% DIOX, and 90 wt% DMSO. PDO is positioned so that the view angle is along the $C_1 - C_2$ bond as illustrated in the projection diagram at the left. **B** Spatial distribution maps of a *gauche* vicinal diol conformation of PDO in pure water, 90 wt% DIOX, and 90 wt% DMSO. PDO is positioned to the view angle is above the $C_1 - C_2$ bond as shown in the diagram at the left. In both A and B, normalized density isovalues between 1.5-3.0 are shown for water in red and isovalues between 1.3-1.5 are shown for cosolvent in blue. Dashed lines emphasize regions that remain enriched in water in DMSO-water mixtures, whereas other water-enriched regions are missing. Additional angles of these spatial distribution maps are available in Supplementary Movie S1.[2]

6.3.7 Hypothesized participation of DMSO in dehydration reaction

Based on the simulation and experimental findings in this work and prior literature results, Figure 6.7A shows a possible mechanism for the DMSO-mediated formation of ACE in DMSO-water mixtures. An acidic proton, stabilized in solution either by DMSO-water clusters or by DMSO molecules alone,[18] first adds to the primary hydroxyl group of PDO. The protonated hydroxyl group is stabilized by a nearby DMSO molecule, which acts as a proton acceptor, and a nearby water molecule, which donates a hydrogen bond. Previous *ab initio* studies of acid-catalyzed dehydration reactions have shown that DMSO stabilizes protonated hydroxyl groups16,[78] and DMSO-water clusters stabilize protons more strongly than DMSO alone,[18] suggesting that the proposed intermediate is most favorably stabilized by both solvating water and DMSO molecules. The acidic proton then catalyzes a semipinacol rearrangement[80] to yield the final products, consistent with the observation that a vicinal hydroxyl group promotes dehydration.

For the proposed mechanism to selectively produce ACE, the primary hydroxyl group should be preferentially protonated over the secondary hydroxyl group. We thus performed *ab initio* calculations to determine if DMSO-water mixtures more favorably stabilize the protonated primary hydroxyl group by computing the energy difference (ΔE) between pro-

tonated primary and secondary hydroxyl groups of PDO, as defined in Equation 6.7.

$$\Delta E = E_{HPDO}^{secondary} - E_{HDPO}^{primary} \tag{6.7}$$

$\Delta E > 0$ indicates that the protonation of the primary hydroxyl group is more favorable than the protonation of the secondary hydroxyl group of PDO, which would favor ACE formation. In both pure water and pure DMSO (to represent 90 wt% DMSO-water mixtures), $\Delta E < 0$ when using implicit solvent models at multiple levels of theory (ESI, Table S6).[2] These findings may indicate that implicit solvent models are insufficient to capture the influence of the local solvent environment encoded within the spatial distribution maps computed using MD. We next performed density functional theory (DFT) calculations using a cluster-continuum approach in which a small number of solvent molecules near the reactant were explicitly modeled while the rest of the solvent was modeled implicitly. Similar techniques have been previously used to significantly improve DFT predictions of pK_a's in aqueous solution,[81–83] and moreover prior *ab initio* MD simulations have shown only a small number of DMSO molecules contribute to dehydration reactions.[78] System configurations were extracted from the MD trajectories, the solvent molecules in close proximity to the two hydroxyl groups of PDO were retained, and either the primary or secondary hydroxyl group was protonated. The system was then relaxed via DFT to obtain energy-minimizing structures and

calculate ΔE.

We first calculated ΔE for PDO in a *trans* vicinal diol conformation with a solvation shell containing DMSO molecules and a single water molecule to represent 90 wt% DMSO configurations. The number of DMSO molecules was varied between one and four to consider multiple possible solvent configurations. Focusing on configurations with 2 DMSO molecules and 1 water molecule to avoid effects associated with solvent molecules far from the protonated hydroxyl group, we find that the protonated hydroxyl group is stabilized by a nearby DMSO molecule and a water molecule that acts as a hydrogen bond acceptor, in agreement with the proposed mechanism (Figure 6.7A). If a similar solvent structure is maintained around each hydroxyl group, $\Delta E = 2.59$ kJ/mol, favoring the protonated primary hydroxyl group (Figure 6.7B). When the same solvent structure is maintained for the *gauche* conformation (Figure 6.7C), the energy difference increases to $\Delta E = 7.64$ kJ/mol, indicating that protonation of the primary hydroxyl group is preferred to a greater extent for the *gauche* conformation compared to the *trans* conformation. Figure 6.6B further indicates that in the *gauche* conformation, the primary hydroxyl group is preferentially solvated by both a DMSO molecule and a water molecule, which is the preferred solvent structure in the DFT calculations. This finding suggests that the *gauche* conformation of PDO may both energetically prefer protonation of the primary hydroxyl group and drive the formation of solvent structures that stabilize this group. In support

of this hypothesis, Table 6.2 also indicates that *cis*-1,2-cyclohexanediol is significantly more reactive and selective to its dehydration product in DMSO-water mixtures than *trans*-1,2-cyclohexanediol. Spatial distribution maps for these two isomers (ESI, Figure S11)[2] exhibit similar trends as for the *trans* and *gauche* conformations of PDO, with only a single water domain observed near *cis*-1,2-cyclohexanediol and DMSO molecules enriched in regions corresponding to hydrogen bonding donors.

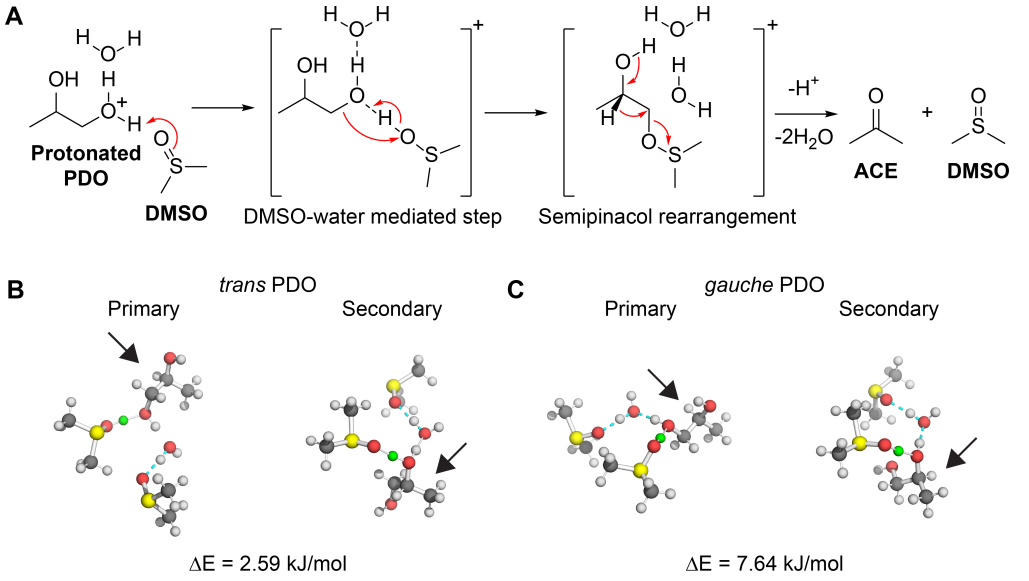


Figure 6.7: **A** Proposed mechanism of PDO dehydration to afford acetone in DMSO-water mixtures. Red arrows denote the transfer of electrons. **B** Relaxed structures of protonated primary or secondary hydroxyl groups of the *trans* vicinal diol conformation of PDO with two DMSO molecules and 1 water. **C** Relaxed structures of protonated primary or secondary hydroxyl groups of the *gauche* vicinal diol conformation of PDO with two DMSO molecules and 1 water molecule. In B and C, black arrows are drawn as a visual guide to identify PDO, the green atom is a proton, and cyan dashed lines indicate hydrogen bonds.

Overall, the simulations and experiments are consistent with the mechanism in Figure 6.7A, in which the unique localization of DMSO and water molecules near reactive sites of PDO contribute to the preferential formation of ACE over PRO in DMSO-water simulations. The *gauche* conformation in particular favors protonation of the primary hydroxyl group while simultaneously driving the formation of preferential solvation

environments. These results are also consistent with the "shielding" mechanism because the *gauche* conformation inhibits access of acidic protons to secondary hydroxyl groups. An alternative possibility is that solvent dynamic effects may affect reaction selectivity. DMSO-water mixtures have been shown to exhibit slow solvent relaxation for ~19-33 and ~52-70 wt% DMSO,[37] which may lead to non-equilibrium solvation of charged transition states during PRO dehydration that could increase the activation energy barrier for forming PRO. Caratzoulas *et al.* found that solvent reorganization is the primary contribution to hydride transfer activation energy barriers in the acid-catalyzed dehydration of fructose to HMF,[33] which may similarly affect PDO dehydration. As these effects are neglected by the static relaxation calculations performed here, more detailed *ab initio* MD simulations coupled with enhanced sampling techniques may be necessary to further validate the mechanism for the of PDO to ACE in DMSO-water mixtures.

6.4 Summary

We have studied the Brønsted acid-catalyzed dehydration of PDO as a model compound for biomass conversion. Reactions were carried out in pure water and in aqueous mixtures of γ-valerolactone, 1,4-dioxane, tetrahydrofuran, *N*-methyl-2-pyrrolidone, tetramethylene sulfoxide, and dimethyl sulfoxide at 433 K over a triflic acid catalyst to demonstrate that

the selectivity of PDO dehydration can be modulated by solvent composition. PRO is the major product in all mixed-solvent environments, except in mixtures of water with DMSO, in which ACE is the major product. We find that the presence of DMSO suppresses the rate of PRO formation. We attribute this suppression behavior to increases in activation energy barriers for transition states corresponding to PRO formation in mixed-solvent environments. We probe changes to these activation barriers by assuming that the reactant and product states of PDO and PRO, respectively, are correlated in energy with their corresponding transition states. Using classical molecular dynamics simulations, we use solvation free energy calculations to quantify the free energy difference between the reactant and product in a mixed-solvent environment relative to water ($\Delta\Delta$G) and find that this free energy difference correlates with the experimental rate of propanal formation in five cosolvent systems. $\Delta\Delta$G also captures trends in the suppression of the main dehydration product in three other acid-catalyzed dehydration reactions: *cis*-1,2-cyclohexanediol, *trans*-1,2-cyclohexanediol, and 1,3-cyclohexanediol. However, $\Delta\Delta$G does not predict the preference for ACE formation in DMSO-water mixtures. Analysis of the spatial distribution of solvent molecules around PDO instead shows that DMSO uniquely competes with water for reactive sites, which can lead to changes in mechanisms that favor steps towards forming acetone. Static relaxation calculation using density functional theory show that a protonated hydroxyl group is stabilized by DMSO and water molecules, pointing to the

importance of the local solvent environment. Calculations of the energy difference between protonated primary and secondary hydroxyl groups on PDO further demonstrate that protonation of the primary hydroxyl group is favorable in DMSO-water mixtures, particularly for PDO in the *gauche* vicinal diol conformation, which is consistent with a mechanism producing acetone.

These findings show that classical MD simulations can efficiently screen multiple solvent systems to determine trends underlying the suppression of the main dehydration product, without requiring the explicit modeling of reaction mechanisms. This methodology represents a step toward the rational design of mixed-solvent environments for liquid-phase reaction schemes and has the potential to alleviate the time-intensive exercise of experimentation that typically accompanies the development of new processes. Moreover, from the perspective of solvent screening, classical MD appears to be sufficient to capture trends and can be a useful guide for *ab initio* studies, such as the density functional theory calculations performed here, or experimental investigations.

6.5 References

[1] Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions. *ACS Catalysis* **2019**, *10*, 1679–1691.

[2] Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions [Supporting Information]. *ACS Catalysis* **2019**, *10*, 1679–1691.

[3] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[4] Walker, T. W.; Motagamwala, A. H.; Dumesic, J. A.; Huber, G. W. Fundamental catalytic challenges to design improved biomass conversion technologies. *Journal of Catalysis* **2018**, *369*.

[5] Huber, G. W.; Iborra, S.; Corma, A. Synthesis of transportation fuels from biomass: chemistry, catalysts, and engineering. *Chemical reviews* **2006**, *106*, 4044–4098.

[6] Stöcker, M. Biofuels and biomass-to-liquid fuels in the biorefinery: Catalytic conversion of lignocellulosic biomass using porous materials. *Angewandte Chemie International Edition* **2008**, *47*, 9200–9211.

[7] Tock, L.; Gassner, M.; Maréchal, F. Thermochemical production of liquid fuels from biomass: Thermo-economic modeling, process design and process integration analysis. *Biomass and bioenergy* **2010**, *34*, 1838–1854.

[8] Nguyen, T. Y.; Cai, C. M.; Kumar, R.; Wyman, C. E. Overcoming factors limiting high-solids fermentation of lignocellulosic biomass to ethanol. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11673–11678.

[9] Mukherjee, S.; Vannice, M. A. Solvent effects in liquid-phase reactions: I. Activity and selectivity during citral hydrogenation on Pt/SiO2 and evaluation of mass transfer effects. *Journal of Catalysis* **2006**, *243*, 108–130.

[10] Foresman, J. B.; Keith, T. A.; Wiberg, K. B.; Snoonian, J.; Frisch, M. J. Solvent effects. 5. Influence of cavity shape, truncation of electrostatics, and electron correlation on ab initio reaction field calculations. *The Journal of Physical Chemistry* **1996**, *100*, 16098–16104.

[11] Reichardt, C.; Welton, T. *Solvents and solvent effects in organic chemistry*; John Wiley & Sons, 2011.

[12] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[13] He, J.; Liu, M.; Huang, K.; Walker, T. W.; Maravelias, C. T.; Dumesic, J. A.; Huber, G. W. Production of levoglucosenone and 5-hydroxymethylfurfural from cellulose in polar aprotic solvent–water mixtures. *Green Chemistry* **2017**, *19*, 3642–3653.

[14] Wei, Z.; Li, Y.; Thushara, D.; Liu, Y.; Ren, Q. Novel dehydration of carbohydrates to 5-hydroxymethylfurfural catalyzed by Ir and Au chlorides in ionic liquids. *Journal of the Taiwan Institute of Chemical Engineers* **2011**, *42*, 363–370.

[15] Madon, R. J.; Iglesia, E. Catalytic reaction rates in thermodynamically non-ideal systems. *Journal of Molecular Catalysis A: Chemical* **2000**, *163*, 189–204.

[16] Cesarotti, E.; Ugo, R.; Kaplan, L. A discussion of the different kinds of solute-solute and solute-solvent interactions acting in homogeneous catalysis by transition metal complexes. *Coordination Chemistry Reviews* **1982**, *43*, 275–298.

[17] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628.

[18] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[19] Chheda, J. N.; Huber, G. W.; Dumesic, J. A. Liquid-phase catalytic processing of biomass-derived oxygenated hydrocarbons to fuels and chemicals. *Angewandte Chemie International Edition* **2007**, *46*, 7164–7183.

[20] Corma, A.; Iborra, S.; Velty, A. Chemical routes for the transformation of biomass into chemicals. *Chemical reviews* **2007**, *107*, 2411–2502.

[21] Román-Leshkov, Y.; Barrett, C. J.; Liu, Z. Y.; Dumesic, J. A. Production of dimethylfuran for liquid fuels from biomass-derived carbohydrates. *Nature* **2007**, *447*, 982–985.

[22] Mellmer, M. A.; Gallo, J. M. R.; Martin Alonso, D.; Dumesic, J. A. Selective production of levulinic acid from furfuryl alcohol in THF solvent systems over H-ZSM-5. *ACS Catalysis* **2015**, *5*, 3354–3359.

[23] Pagan-Torres, Y. J.; Wang, T.; Gallo, J. M. R.; Shanks, B. H.; Dumesic, J. A. Production of 5-hydroxymethylfurfural from glucose using a combination of Lewis and Brønsted acid catalysts in water in a biphasic reactor with an alkylphenol solvent. *Acs Catalysis* **2012**, *2*, 930–934.

[24] Román-Leshkov, Y.; Chheda, J. N.; Dumesic, J. A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose. *Science* **2006**, *312*, 1933–1937.

[25] Mushrif, S. H.; Caratzoulas, S.; Vlachos, D. G. Understanding solvent effects in the selective conversion of fructose to 5-hydroxymethylfurfural: a molecular dynamics investigation. *Physical Chemistry Chemical Physics* **2012**, *14*, 2637–2644.

[26] Varghese, J. J.; Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering* **2019**, *4*, 165–206.

[27] Shrivastav, G.; Khan, T. S.; Agarwal, M.; Haider, M. A. A Car–Parrinello Molecular Dynamics Simulation Study of the Retro Diels–Alder Reaction for Partially Saturated 2-Pyrones in Water. *The Journal of Physical Chemistry C* **2018**, *122*, 11599–11607.

[28] Guo, N.; Caratzoulas, S.; Doren, D. J.; Sandler, S. I.; Vlachos, D. G. A perspective on the modeling of biomass processing. *Energy & Environmental Science* **2012**, *5*, 6703–6716.

[29] Caratzoulas, S.; Courtney, T.; Vlachos, D. G. Hybrid quantum mechanics/molecular mechanics-based molecular dynamics simulation of acid-catalyzed dehydration of polyols in liquid water. *The Journal of Physical Chemistry A* **2011**, *115*, 8816–8821.

[30] Giorgianni, G.; Abate, S.; Centi, G.; Perathoner, S.; van Beuzekom, S.; Soo-Tang, S.-H.; Van der Waal, J. C. Effect of the solvent in enhancing the selectivity to furan derivatives in the catalytic hydrogenation of furfural. *Acs Sustainable Chemistry & Engineering* **2018**, *6*, 16235–16247.

[31] Gupta, S.; Alam, M. I.; Khan, T. S.; Sinha, N.; Haider, M. A. On the mechanism of retro-Diels–Alder reaction of partially saturated 2-pyrones to produce biorenewable chemicals. *RSC advances* **2016**, *6*, 60433–60445.

[32] Khan, T. S.; Gupta, S.; Alam, M. I.; Haider, M. A. Reactivity descriptor for the retro Diels–Alder reaction of partially saturated 2-pyrones: DFT study on substituents and solvent effects. *RSC advances* **2016**, *6*, 101697–101706.

[33] Caratzoulas, S.; Vlachos, D. G. Converting fructose to 5-hydroxymethylfurfural: a quantum mechanics/molecular mechanics study of the mechanism and energetics. *Carbohydrate research* **2011**, *346*, 664–672.

[34] Mushrif, S. H.; Varghese, J. J.; Krishnamurthy, C. B. Solvation dynamics and energetics of intramolecular hydride transfer reactions in biomass conversion. *Physical Chemistry Chemical Physics* **2015**, *17*, 4961–4969.

[35] Nikolakis, V.; Mushrif, S. H.; Herbert, B.; Booksh, K. S.; Vlachos, D. G. Fructose–water–dimethylsulfoxide interactions by vibrational spectroscopy and molecular dynamics simulations. *The Journal of Physical Chemistry B* **2012**, *116*, 11274–11283.

[36] Vasudevan, V.; Mushrif, S. H. Insights into the solvation of glucose in water, dimethyl sulfoxide (DMSO), tetrahydrofuran (THF) and N, N-dimethylformamide (DMF) and its possible implications on the conversion of glucose to platform chemicals. *Rsc Advances* **2015**, *5*, 20756–20763.

[37] Hazra, M. K.; Bagchi, B. Non-equilibrium solvation dynamics in water-DMSO binary mixture: Composition dependence of non-linear relaxation. *The Journal of chemical physics* **2018**, *149*, 084501.

[38] Trummal, A.; Lipping, L.; Kaljurand, I.; Koppel, I. A.; Leito, I. Acidity of strong acids in water and dimethyl sulfoxide. *The Journal of Physical Chemistry A* **2016**, *120*, 3663–3669.

[39] Raamat, E.; Kaupmees, K.; Ovsjannikov, G.; Trummal, A.; Kütt, A.; Saame, J.; Koppel, I.; Kaljurand, I.; Lipping, L.; Rodima, T.; et al.. Acidities of strong neutral Brønsted acids in different media. *Journal of Physical Organic Chemistry* **2013**, *26*, 162–170.

[40] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[41] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[42] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[43] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[44] Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[45] Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations. *Frontiers in chemistry* **2019**, *7*, 439.

[46] Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* **2008**, *129*, 124105.

[47] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of computer-aided molecular design* **2015**, *29*, 397–411.

[48] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.

[49] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.; *GaussianËœ16 Revision C.01*; 2016; Gaussian Inc. Wallingford CT.

[50] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A* **1988**, *38*, 3098.

[51] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **1988**, *37*, 785.

[52] Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation energy density functionals of Becke and Lee, Yang and Parr. *Chemical Physics Letters* **1989**, *157*, 200–206.

[53] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77*, 3865.

[54] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.* **1997**, *78*, 1396–1396.

[55] Frisch, M. J.; Head-Gordon, M.; Pople, J. A. A direct MP2 gradient method. *Chemical Physics Letters* **1990**, *166*, 275–280.

[56] Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Semi-direct algorithms for the MP2 energy and gradient. *Chemical physics letters* **1990**, *166*, 281–289.

[57] Head-Gordon, M.; Head-Gordon, T. Analytic MP2 frequencies without fifth-order storage. Theory and application to bifurcated hydrogen bonds in the water hexamer. *Chemical Physics Letters* **1994**, *220*, 122–128.

[58] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical physics letters* **1988**, *153*, 503–506.

[59] Sæbø, S.; Almlöf, J. Avoiding the integral storage bottleneck in LCAO calculations of electron correlation. *Chemical Physics Letters* **1989**, *154*, 83–89.

[60] Barone, V.; Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A* **1998**, *102*, 1995–2001.

[61] Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of computational chemistry* **2003**, *24*, 669–681.

[62] Zhang, D.; Barri, S. A.; Chadwick, D. Dehydration of 1, 2-propanediol to propionaldehyde over zeolite catalysts. *Applied Catalysis A: General* **2011**, *400*, 148–155.

[63] Akiyama, M.; Sato, S.; Takahashi, R.; Inui, K.; Yokota, M. Dehydration–hydrogenation of glycerol into 1, 2-propanediol at ambient hydrogen pressure. *Applied Catalysis A: General* **2009**, *371*, 60–66.

[64] Lowry, T. H.; Richardson, K. S. *Mechanism and theory in organic chemistry*; Harper & Row New York, 1987; Vol. 294.

[65] Jencks, W. Ingold Lecture. How does a reaction choose its mechanism? *Chemical society reviews* **1981**, *10*, 345–375.

[66] Richard, J. P.; Toteva, M. M.; Amyes, T. L. What is the stabilizing interaction with nucleophilic solvents in the transition state for solvolysis of tertiary derivatives: nucleophilic solvent participation or nucleophilic solvation? *Organic letters* **2001**, *3*, 2225–2228.

[67] Fowler, F.; Katritzky, A.; Rutherford, R. The correlation of solvent effects on physical and chemical properties. *Journal of the Chemical Society B: Physical Organic* **1971**, 460–469.

[68] Wohlfarth, C. *Static dielectric constants of pure liquids and binary liquid mixtures: supplement to IV/6*; Springer Science & Business Media, 2008; Vol. 17.

[69] Wohlfarth, C. In *Static Dielectric Constants of Pure Liquids and Binary Liquid Mixtures*; Springer, 2015; pp 91–91.

[70] Uosaki, Y.; Kawamura, K.; Moriyoshi, T. Static Relative Permittivities of Water + 1-Methyl-2-pyrrolidinone and Water + 1,3-Dimethyl-2-imidazolidinone Mixtures under Pressures up to 300 MPa at 298.15 K. *Journal of Chemical & Engineering Data* **1996**, *41*, 1525–1528.

[71] Klein, D. R. *Organic Chemistry*, 3rd ed.; Wiley: USA, 2017.

[72] Bagno, A.; Scorrano, G. Acid-base properties of organic solvents. *Journal of the American Chemical Society* **1988**, *110*, 4577–4582.

[73] Nakamura, K.; Osamura, Y. Theoretical study of the reaction mechanism and migratory aptitude of the pinacol rearrangement. *Journal of The American Chemical Society* **1993**, *115*, 9112–9120.

[74] Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters,. pi.*,. alpha., and. beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry* **1983**, *48*, 2877–2887.

[75] Sutton, J. E.; Vlachos, D. G. A theoretical and computational analysis of linear free energy relations for the estimation of activation energies. *ACS Catalysis* **2012**, *2*, 1624–1634.

[76] Van Santen, R. A.; Neurock, M.; Shetty, S. G. Reactivity theory of transition-metal surfaces: a Brønsted- Evans- Polanyi linear activation energy- free-energy analysis. *Chemical reviews* **2009**, *110*, 2005–2048.

[77] Wang, S.; Petzold, V.; Tripkovic, V.; Kleis, J.; Howalt, J. G.; Skulason, E.; Fernández, E.; Hvolbæk, B.; Jones, G.; Toftelund, A.; et al.. Universal transition state scaling relations for (de) hydrogenation over transition metals. *Physical Chemistry Chemical Physics* **2011**, *13*, 20760–20765.

[78] Qian, X.; Liu, D. Free energy landscape for glucose condensation and dehydration reactions in dimethyl sulfoxide and the effects of solvent. *Carbohydrate research* **2014**, *388*, 50–60.

[79] Martin, D.; Weise, A.; Niclas, H.-J. The solvent dimethyl sulfoxide. *Angewandte Chemie International Edition in English* **1967**, *6*, 318–334.

[80] Song, Z.-L.; Fan, C.-A.; Tu, Y.-Q. Semipinacol rearrangement in natural product synthesis. *Chemical reviews* **2011**, *111*, 7523–7556.

[81] Bryantsev, V. S.; Diallo, M. S.; Goddard Iii, W. A. Calculation of solvation free energies of charged solutes using mixed cluster/continuum models. *The Journal of Physical Chemistry B* **2008**, *112*, 9709–9719.

[82] Thapa, B.; Schlegel, H. B. Density functional theory calculation of p K aâĽ™s of thiols in aqueous solution using explicit water molecules and the polarizable continuum model. *The Journal of Physical Chemistry A* **2016**, *120*, 5726–5735.

[83] Thapa, B.; Schlegel, H. B. Improved p K a Prediction of Substituted Alcohols, Phenols, and Hydroperoxides in Aqueous Medium Using Density Functional Theory and a Cluster-Continuum Solvation Model. *The Journal of Physical Chemistry A* **2017**, *121*, 4698–4706.

7 RATIONAL DESIGN OF MIXED SOLVENT SYSTEMS FOR ACID-CATALYZED BIOMASS CONVERSION PROCESSES USING A COMBINED EXPERIMENTAL, MOLECULAR DYNAMICS AND MACHINE LEARNING APPROACH

Chapter 5 provides a computational workflow for using molecular dynamcis simulations and convolutional neural networks (*i.e.* SolventNet) to predict solvent-mediated reaction rates, and Chapter 6 provides a method for estimating selectivities of products using solvation free energy calculations. This chapter combines the tools from the preceding chapters into a workflow for screening solvents in biomass conversion reactions. This chapter aims to address the following questions:

- How do we merge the computational tools from Chapters 5 and 6 into a systematic workflow for screening solvents for biomass-related reactions?
- How effective is this predictive framework for select case studies?
- What are the limitations of the computational workflow?

This chapter was reproduced from Walker, T. W.; Chew, A. K.; Van Lehn, R. C.; Dumesic, J. A.; Huber, G. W. Rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using a combined experimental, molecular dynamics and machine learning approach. *Topics in Catalysis* **2020**, *63*, 649–663 with permission from Springer Nature.[1] The supporting information is cited as Ref. 2. T.W. Walker and A.K. Chew contributed equally to this work. T.W. Walker performed all experimental reaction kinetic studies.

In this chapter, we summarize our efforts to estimate solvent effects on the rates and selectivities of liquid-phase, acid-catalyzed biomass conversions reactions using experiments, classical molecular dynamics simulations, and machine learning tools. We then synthesize these insights into a workflow that allows for the rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using computationally efficient methods and minimal experiments. We demonstrate this design framework by analyzing two case studies: the acid-catalyzed dehydration of cyclohexanol to cyclohexene, and the partial dehydration of fructose to 5-hydroxymethylfurfural.

7.1 Introduction

Owing to the development of advanced spectroscopic and computational methods, it is now possible to resolve molecular-level details regarding how elementary reaction sequences proceed on the surface of a catalyst,[3,4] and to produce accurate, a priori estimates of the corresponding reaction barriers.[5,6] These molecular details are typically quantified and combined into quantitative structure-property relationships that enable the rational design of new catalytic materials tailored for specific applications.[7–10] Accordingly, catalyst research and design has assumed a key role in addressing challenges across the full spectrum of societal needs.[10]

Recent interest in liquid-phase, acid-catalyzed biomass conversion reac-

tions[11] has produced a corresponding interest in the role of solvents in mediating these processes[12,13] (*e.g.*, conversion of xylitol to 1,4-anhydroxylitol, Figure 7.1a). In particular, aqueous mixtures with organic, polar aprotic cosolvents (*e.g.*, tetrahydrofuran (THF), Figure 7.1b) have been shown to enable the complete dissolution of biomass-derived materials,[14] while also enhancing reaction selectivity toward desired products[15] compared to the same processes in pure water. However, our understanding of the role of mixed solvent systems in mediating acid-catalyzed biomass conversion reactions can, in some ways, be compared to the state of catalyst science and technology forty years ago:[16] experimental efforts have identified optimal solvent compositions for key applications[17–20] while, in parallel, state-of-the-art quantum chemical methods have produced important, molecular-level insights regarding the role of the solvent system in representative case studies.[21–23] However, a general framework to anticipate solvent effects in biomass conversion processes is currently lacking. Designing mixed solvent systems for biomass conversion processes therefore typically involves laborious, empirical screening of the continuous space of possible water-cosolvent compositions.

Toward developing a framework to effect the rational design of solvent compositions for biomass conversion reactions, we recently investigated the mechanistic role of mixed solvent systems on liquid-phase, acid-catalyzed hydrolysis and dehydration reactions for several biomass-derived model compounds.[24] We found that the addition of polar aprotic

cosolvents to water leads to the formation of *water-enriched local domains* around hydrophilic reactants.[24,25] The acid catalyst is drawn into these regions due to preferential catalyst-water interactions,[24,26] essentially co-locating reactants and catalysts in solution, and resulting in enhanced reaction rates (Figure 7.1c). We then generalized these insights into a predictive framework that allows for the rates of acid-catalyzed reactions of biomass-derived oxygenates to be estimated as a function of the water content of the solvent systems using computationally efficient classical molecular dynamics (MD) simulations and machine learning tools.[25–27] We also demonstrated that MD simulations can be used to estimate the selectivity of reactions occurring in parallel by computing difference in the solvation free energies of the reactant and possible products as a function of solvent composition.[28]

In this chapter, we build upon these past fundamental studies to develop a new workflow for selecting mixed solvent systems for acid-catalyzed biomass conversion processes. We combine the prior MD- and machine-learning based tools to efficiently screen possible solvent systems using minimal experimentation and computationally efficient methods. We demonstrate the usage of this workflow by analyzing two case studies: the acid-catalyzed dehydration of cyclohexanol to cyclohexene, and the partial dehydration of fructose to 5-hydroxymethylfurfural.

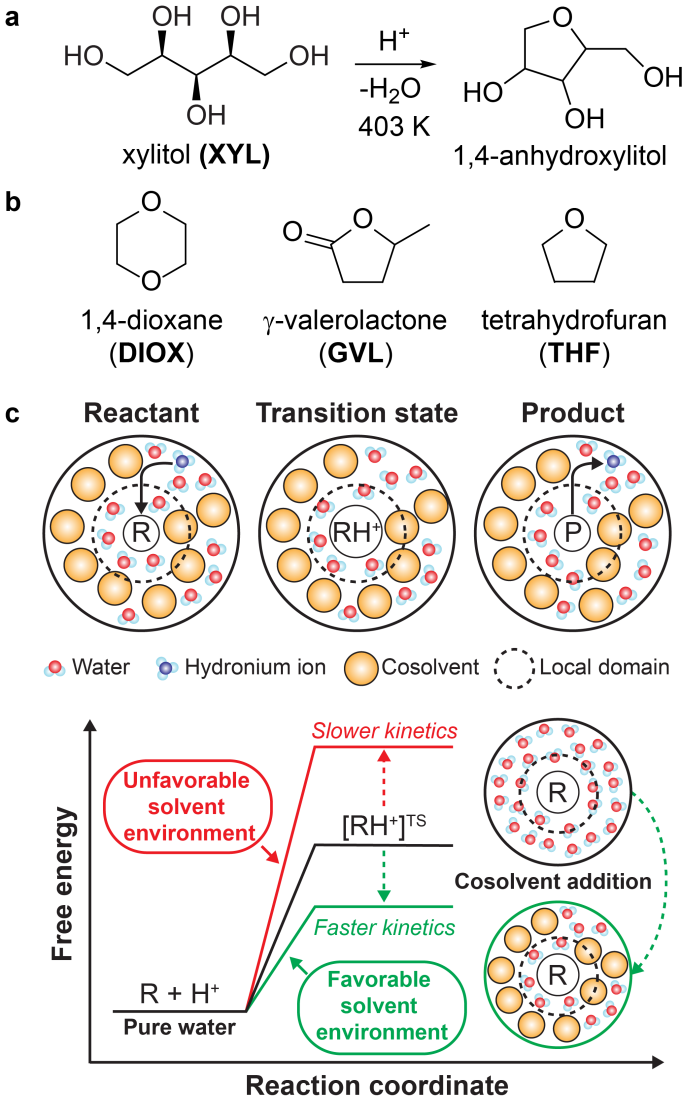


Figure 7.1: Overview of acid-catalyzed reactions in mixed solvent systems for biomass conversion. (a) Acid-catalyzed reaction example of xylitol (XYL) dehydration to afford 1-4-anhydroxylitol. (b) Example of three organic, polar aprotic cosolvents. (c) Schematic of acid-catalyzed reactions in a mixed solvent systems that proceeds through a charged transition state (TS), formed after the protonation of the reactant by a hydronium ion catalyst.[24] The schematic illustrates the formation of a local solvent domain around the reactant in a mixed solvent system that modifies the reaction free energy landscape, thus affecting reaction kinetics. These images were adapted with permission of 25,26.

7.2 Materials, methods, and definitions

7.2.1 Materials

Solvents (tetrahydrofuran, 99% with 200 ppm BHT; gamma-valerolactone, 98%+; acetone, 99% anhydrous; dioxane, 99%+ with 200 ppm BHT; and dimethyl sulfoxide, 98%+) were obtained from vendors and used as received. Fructose, 5-hydroxymethylfurfural, levulinic acid, cyclohexanol and cyclohexene (all reagent grade) were obtained from Sigma and used as received. Triflic acid (99%) was obtained from Acros Organics and used as received.

7.2.2 Methods

Reaction kinetics studies of fructose dehydration, 5-hydroxymethylfurfural hydrolysis, and cyclohexanol dehydration were carried out in glass reactors. Briefly, an appropriate amount of reactant material (*e.g.*, 1 wt% fructose) was dissolved in the desired solvent system (*e.g.*, 90% tetrahydrofuran with 10 wt% water) containing an appropriate amount of dissolved acid catalyst (*e.g.*, 0.1 M). This reaction mixture was then charged into closed, 10 mL thick-walled glass reactors equipped with magnetic stir bars, which were then sealed with Teflon-lined screw-top caps and submerged in an oil bath at the desired reaction temperature (*e.g.*, 130°C). Reactors were agitated using a magnetic stir plate at 500 rpm. Reactors were then removed at intervals corresponding to the desired reaction time

points, and quenched in an ice bath at 0°C.

The contents of the reactors were filtered with a 0.2 micron PTFE syringe filter, and analyzed using high-performance liquid chromatographs (HPLC, for fructose, 5-hydroxymethylfurfual, and levulinic acid) or gas-chromatographs (GC, for cyclohexanol and cyclohexene). Analytes were quantified using calibration curves using known external standards. The gas chromatograph was a Shimadzu GC-2010 equipped with a flame ionization detector and a RTX-VMS column. The HPLC was a Shimadzu 1020 series equipped with a refractive index detector and a photodiode array. The column was a proton-based Aminex HP87-X ion exclusion column with a mobile phase of 0.005 M sulfuric acid in water at a flow rate of 0.6 mL/min.

Triflic acid was used as an acid catalyst in all experiments, as it has been shown to fully dissociate even in organic solvents, which allows kinetic parameters (rate constants) to be estimated by normalizing the experimentally determined reaction rate on an accurate per-proton basis using Equation 7.1. r_i is the rate at which the reactant i is consumed in units of mol L^{-1} s^{-1}, C_i is the molar concentration of the ith species, C_{H^+} is the molar concentrations of excess protons in solution (assumed to be equal to the molar concentration of dissolved triflic acid in these experiments), and k_i is the rate constant associated with this reaction step at a fixed solvent composition and temperature.

$$r_i = -\frac{dC_i}{dt} = k_i C_i C_{H^+} \tag{7.1}$$

The rate constant associated with formation of the corresponding product (j) from reactant (i) is then shown in Equation 7.2.

$$r_j = -\frac{dC_j}{dt} = k_j C_i C_{H^+} \tag{7.2}$$

Equation 7.3 defines the selectivity to product j as the ratio of the yield of product j to the conversion of the reactant i, where the conversion (X) of species i is defined in Equation 7.4.

$$S_j = \frac{C_j^{final}}{C_i^{initial}} X_i^{-1} \tag{7.3}$$

$$X_i = 1 - \frac{C_j^{final}}{C_i^{initial}} \tag{7.4}$$

Herein, for every case but fructose conversion in pure water, we report rate constants measured in the kinetic regime, meaning that subsequent conversion of the product j can be neglected. Under these conditions, the selectivity (S_j) of the reaction going from the reactant i to the product j is independent of conversion, so that selectivity is equivalent to the ratio of initial rates or, equivalently, the ratio of rate constants as shown in Equation 7.5. Throughout this investigation, reported reaction selectivities are therefore consistent with and equal to the definitions provided in both

Equations 7.3 and 7.5. See Table S1 and the associated discussion in the ESI for details.[2]

$$S_j = \frac{r_j}{r_i} = \frac{k_j}{k_i} \tag{7.5}$$

Finally, to facilitate comparison of reactivity across multiple mixed solvent systems, we quantify solvent effects on acid-catalyzed reaction rates in terms of a kinetic solvent parameter (σ), which we have defined in previously as Equation 3.4. Positive σ values indicate an increase the rate of the corresponding reaction step in the mixed solvent system compared to pure water; negative values have the converse implication.

7.2.3 Classical molecular dynamics simulations

All classical MD simulations were performed using Gromacs 2016.[29] Reactant, product, and cosolvent molecules were parameterized using the CGenFF/CHARMM36 force fields.[30–32] Water was modeled using the Single Point Charged/Extended (SPC/E) model.[33] A $(6\ \text{nm})^3$ simulation box was initialized with water and cosolvent molecules at the desired composition. The system was equilibrated in a *NPT* simulation for 5 ns at T = 300 K and P = 1 bar with a velocity-rescale thermostat and Berendsen barostat. Then, a single reactant molecule was added to the system and equilibrated with the same barostat and thermostat for 500 ps. 4 ns *NPT* productions were performed at the reaction temperature (T = 433.15 K

for cyclohexanol dehydration or T = 403.15 K for fructose dehydration) and P = 1 bar using a Nóse-Hoover thermostat and Parrinello-Rahman barostat. The 4 ns trajectory was partitioned into two 2 ns trajectories. Each partition was used to generate voxel representations as described in Chapter 5[27] and summarized here. For each simulation configuration, a 3D histogram was generated with the system centered on the center-of-mass of the reactant. The histogram covered a $(4\ \text{nm})^3$ volume that was divided into a $20 \times 20 \times 20$ grid of bins corresponding to $(0.2\ \text{nm})^3$ volume elements. For each bin, normalized occurrences of reactant, cosolvent, and water molecules were stored in three separate channels to obtain a $20 \times 20 \times 20 \times 3$ grid of voxels for a single MD configuration. These grid values were averaged using 2 ns of simulation data, equivalent to 200 MD configurations, to generate a single voxel representation that captures the spatial distribution of atoms within the system. Voxel representations were then inputted into a pre-trained 3D convolutional neural network called SolventNet, as described in Ref. 27, which outputs the predicted kinetic solvent parameters. Each 2 ns partition was treated as an independent trial. We report the average predicted kinetic solvent parameter for these two trials and report the standard deviation of the predictions as the error.

Solvation free energies were computed from a series of stochastic dynamics simulations following the procedure described in Section 2.1.2.[26,28] Starting from an equilibrated solvent system, a reactant or product was

added to the system. The total potential of the system was defined in Equation 2.2 as a function of Lennard-Jones (λ_{LJ}) and electrostatic (λ_{elec}) coupling parameters. For each simulation, the system was energy minimized with the steepest descent algorithm and equilibrated for 100 ps at constant temperature followed by 2 ns at constant temperature and constant pressure using the Berendsen barostat. A 11 ns production simulation at constant reaction temperature and pressure (P = 1 bar) was then performed with the Parrinello-Rahman barostat. The last 10 ns of each *NPT* production simulation were split into two 5 ns trajectories and treated as two independent samples; the average and standard deviation of these samples are reported as the free energy and error. Energy differences computed between all pairs of windows were collected every 0.2 ps and solvation free energies were computed with the Multistate Bennett Acceptance Ratio[34] method using the python alchemical-analysis tool.[35]

All simulations were performed using a leapfrog integrator with a 2-fs timestep. Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential that was smoothly shifted to zero between 1.0 nm and 1.2 nm. Electrostatic interactions were calculated using the smooth Particle Mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and 4th order interpolation. Bonds were constrained using the LINCS algorithm.[36] All thermostats used a 1.0 ps time constant and all barostats used a 5.0 ps time constant with an isothermal compressibility of 5.0×10^{-5}

bar^{-1}.

7.3 Computational design tools and a general procedure for screening mixed solvent systems for biomass conversion processes

Liquid-phase, acid-catalyzed transformations of biomass-derived materials are often characterized by a complex network of reaction steps occurring both in series and in parallel.[37] In these cases, the general challenge in effecting the selective transformation of the raw material into a single, desired product is to selectively enhance the rates of the desirable reaction steps over the undesirable ones. In principle, this challenge can be addressed by modulating the properties of the liquid solvent,[15,38] which is connected to the free-energy landscape that controls the kinetics of each reaction step through a series of non-covalent solvent-solute interactions (solvation energies). However, few predictive frameworks exist to anticipate these effects, and this is particularly true for mixtures of water with organic cosolvents, wherein the relevant mechanistic details are myriad, complex, and not well understood.

Herein, we propose a general process to screen mixed solvent systems for acid-catalyzed biomass conversion reactions using a combination of the experimental and computational approaches described in our prior

work, which can reduce the overall experimental burden associated with designing new solvent systems using empirical screening methods alone. This process is outlined in Figure 7.2 and is composed of four key steps:

1. Establish the reaction network in a reference solvent system and pre-select an initial library of possible mixed solvent systems;

2. Use MD and machine-learning-based tools to screen the initial library of candidate mixed solvent systems, differentiating those that promote the rates of the desired reaction steps;

3. Perform solvation free energy calculations to estimate the thermodynamic selectivity preference for a desired product in candidate mixed solvent systems, and;

4. Experimentally validate the model-predicted, best-performing mixed solvent systems.

We discuss these four generalized steps in more detail, and then demonstrate the procedure outlined in Figure 7.2 by analyzing two case studies: (1) the conversion of cyclohexanol to cyclohexene and (2) the conversion of fructose (FRU) to 5-hydroxymethylfurfural (HMF).

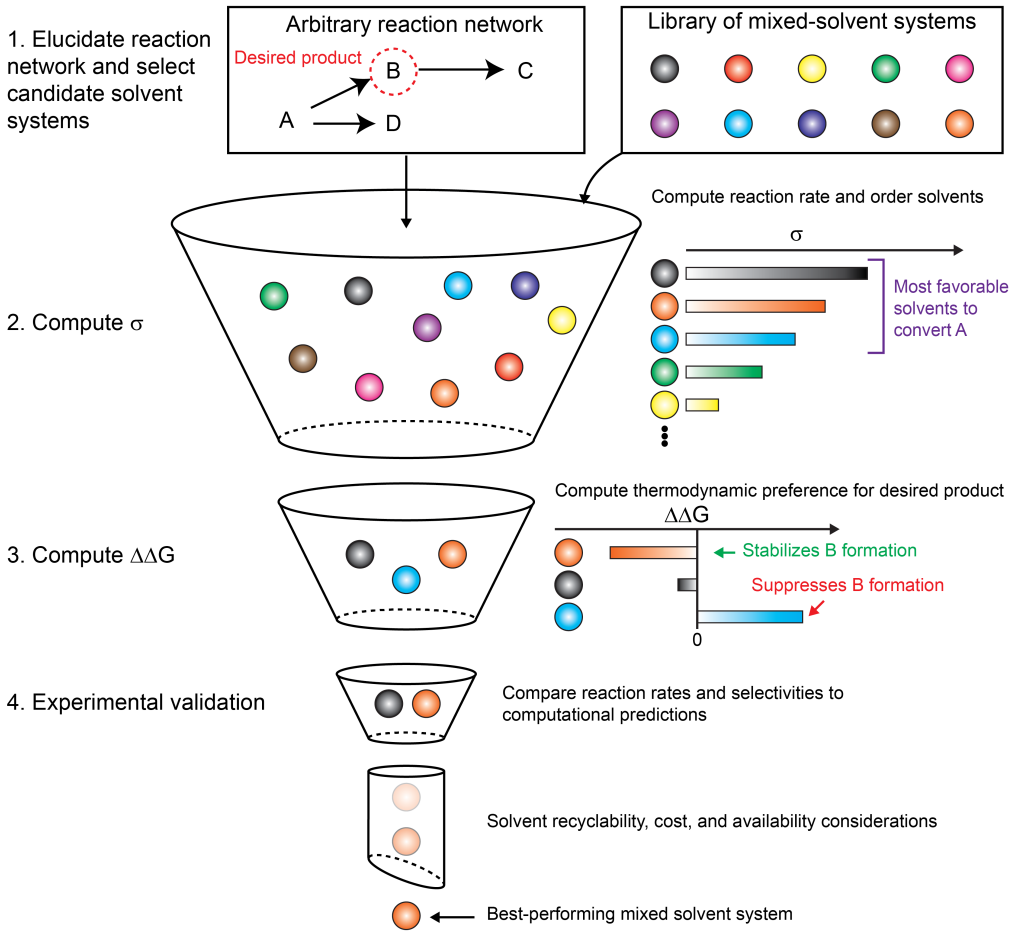


Figure 7.2: Process to screen candidate mixed solvent systems for biomass conversion processes. An arbitrary reaction network as determined using experiments; as an example, product B is desired from reactant A. Polar aprotic cosolvents are selected to mix with water and test the effects of solvent composition on reaction performance. Kinetic solvent parameters (σ) are predicted using molecular dynamics simulations in conjunction with SolventNet, a machine-learning algorithm parameterized to predict kinetic solvent parameters as described in Figure 7.3. The top performing solvents are then tested to see if the reaction selectively forms product B by calculating relative solvation free energies ($\Delta\Delta G$). Negative values of $\Delta\Delta G$ indicate that product B is more stabilized in the mixed solvent system. The top mixed solvent system for the conversion of A and production of B is then selected for experimental testing.

7.3.1 Step 1: Establish reaction network and pre-select candidate solvent systems

The first step in the solvent design process is to elucidate the reaction network underlying the desired chemical transformations in a reference solvent system. An obligatory reference solvent would be pure water, owing to its affordability and ease of handling. However, the reference solvent may contain an organic cosolvent; the minimum criterion is that the reactants and products are soluble at the desired concentrations. For example, cyclohexanol is soluble in pure water, but a minimum amount of organic cosolvent (*e.g.*, ~75 wt% THF) is required to solubilize its dehydration product (cyclohexene). In such cases, this minimum, threshold organic cosolvent content could dictate the reference state.

The reaction network underlying the desired transformation should include all discernable reaction steps between the reactant and products. Accordingly, kinetic parameters at a fixed reaction temperature (rate constants as described in Equations 7.1 and 7.2) should be obtained for all known steps in the reaction network. Note that here we define reaction steps as the transformation between a quantifiable reactant and corresponding product, not elementary steps. This reaction kinetics model is constituted by rate laws that describe the rates of each reaction step as a function of temperature and reactant concentrations,[39,40] and forms the basis for the MD-enabled screening process described below.

Finally, once the reaction kinetics model is formulated in a reference solvent system, a library of candidate mixed solvent systems is proposed. Candidate mixed solvent systems at this stage are selected based on simple design criteria, such as: compatibility with the anticipated reaction conditions (*e.g.*, thermal stability at the anticipated reaction temperature), miscibility with water, toxicity limits, affordability or availability, and so on. Multiple cosolvent-water compositions are considered for each organic cosolvent, or a fixed cosolvent composition can be specified to expedite the MD-screening process outlined in the next step. If a minimum amount of organic cosolvent is required in the reference solvent system, then this cosolvent-water composition can be used. If pure water is used as a reference solvent system then, as above, the base-case cosolvent-water composition can be fixed based on the solubility limits of the reactants and products. For example, fructose is highly soluble in pure water, but is only minimally soluble in solvent systems containing more than 90 wt% of common organic solvents (*e.g.* THF).

7.3.2 Step 2: Compute σ to screen mixed solvent systems for improved reactivity

Once the reaction network and library of candidate mixed solvent systems have been selected, MD and machine learning based tools are used to sort mixed solvent systems based on predictions of reaction rate enhance-

ments and selectivity. In recent work, we have established methods to analyze solvent effects on the rates of acid-catalyzed biomass conversion reactions in mixed solvent systems using classical MD simulations.[25,26] For each mixed solvent system of interest, one reactant molecule is added, and simulated at the desired reaction temperature and pressure in the *NPT* ensemble. We then developed a machine learning environment to correlate the atomic positions of reactant, water, and cosolvent molecules to experimentally determined kinetic solvent parameters (σ), outlined in Figure 7.3a,b.[27] Of the different machine learning tools (*e.g.* support vector machines, deep neural networks), convolutional neural networks (CNNs) have been found to be the best performer in identifying spatial patterns within 2D images.[41] Thus, we first converted atomic positions from MD simulations into an input representation that could be used for a CNN, called voxel representations. These representations are normalized 3D histograms of water, cosolvent, and reactant atomic positions that are stored in red, blue, and green channels (Figure 7.3a). We then input the voxel representations into a 3D CNN, called SolventNet (Figure 7.3b),[27] which is trained on experimental kinetic solvent parameters for 7 biomass-relevant reactants (ethyl tert-butyl ether, tert-butanol, levoglucosan, 1,2-propandiol, cellobiose, FRU, and XYL), 3 cosolvents (Figure 7.1b), and 4 cosolvent mass fractions (25, 50, 75, 90 wt%). We found that SolventNet outperformed other machine learning techniques and human-engineered MD observables in predicting reaction rates.[27] Given a new

reactant-solvent combination, σ values are extracted from MD simulations using SolventNet by: (1) performing a MD simulation of the reactant in the desired mixed solvent composition; (2) generating voxel representations based on the atomic positions of the MD trajectory; and (3) inputting the voxel representations into the trained SolventNet to predict σ. SolventNet only requires 2 ns production simulations to make a prediction, requiring less than an hour in a supercomputing environment for a single reactant-solvent combination. Therefore, we use SolventNet to rapidly screen through mixed solvent systems and predict solvent effects on reaction rates, as expressed by σ, for biomass-derived reactants as a function of solvent system.

7.3.3 Step 3: Compute $\Delta\Delta G$ to estimate selectivities within reaction networks

While SolventNet can rapidly predict forward reaction rates as a function of solvent compositions, it does not directly quantify the selectivity toward a specific product. Based on *ab initio* molecular dynamics studies,[24] we hypothesized that the modulation of solvent composition could affect the reaction energy barriers towards specific products, and we estimate changes in these energy barriers by measuring the relative stability of the reactant and product states in mixed solvent systems (Figure 7.3c). Therefore, we have developed a separate framework using solvation free energy

calculations to estimate the selectivity of parallel reactions as a function of solvent composition, as shown in Figure 7.3.[28] We define the stability of reactant and product states with solvation free energy calculations, which measures the free energy associated with transferring a reactant or product molecule from vacuum to a solvent system. We quantify the free energy difference between reactant (r) and product (p) in mixed solvent systems ($\Delta\Delta$G) by performing four solvation free energy calculations that are related by Equation 6.6 and shown in Figure 7.3d. If $\Delta\Delta G < 0$, the free energy difference between product and reactant state is more negative in mixed solvent systems compared to pure water, indicating that the product state is stabilized to a greater degree than the reactant. We have shown that $\Delta\Delta$G values capture the suppression or enhancement of a main dehydration product without having to model the reaction mechanism or the catalyst explicitly.[28]

Solvation free energy calculations are computationally more demanding than the aforementioned machine-learning tools, which limits their applicability as a high-throughput screening tool. Since $\Delta\Delta$G values take a significantly longer time to compute, requiring ~85 ns of production simulation time (~12 hours on a supercomputer) for a single reactant-solvent combination, initial screening with SolventNet is necessary to lower the necessary number of $\Delta\Delta$G calculations. However, once the larger library of candidate mixed solvent systems has been prescreened using SolventNet to differentiate systems that best promote the reactivity of a reactant

molecule, solvation free energy calculations can be used to quantify a subset of mixed solvent systems that might best enhanced selectivity towards a desired product.

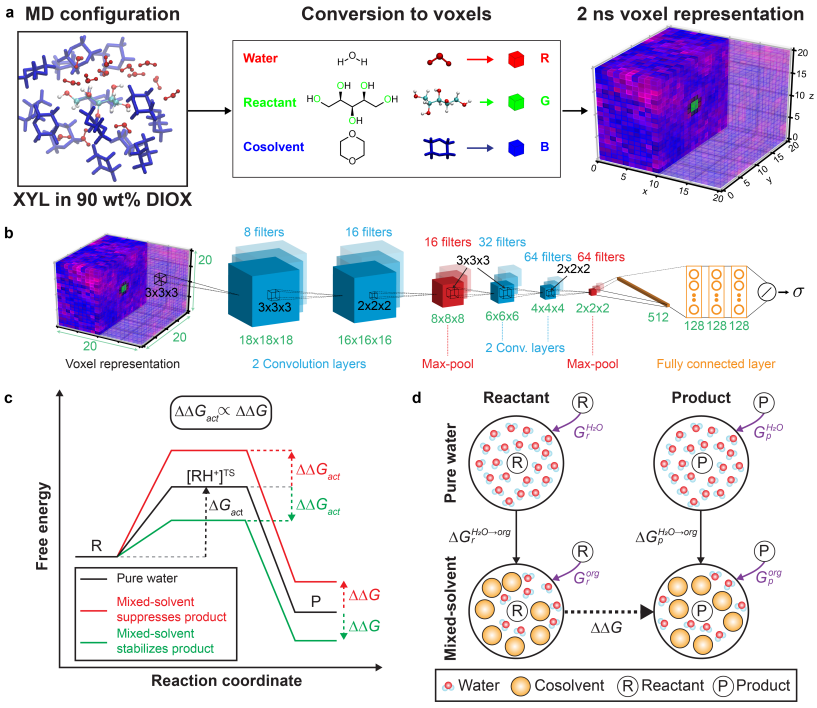


Figure 7.3: Computational tools used to predict reaction rates and selectivities. (a) Conversion of atomic positions obtained from molecular dynamics simulation trajectories into a voxel representation using XYL in 90 wt% DIOX as an example. For each MD configuration, a $20 \times 20 \times 20$ grid of $(0.2\ \text{nm})^3$ volume elements was centered on the reactant. Voxel representations are visualized by showing the water channel in red, the reactant channel in green, and the cosolvent channel in blue. Half of the voxels are transparent to illustrate the solvent distribution around the reactant. (b) Architecture of SolventNet, a 3D CNN that inputs voxel representations and outputs the predicted kinetic solvent parameter (σ). (c) Hypothesized effect of mixed solvent systems on the free energy landscape of reactant, transition, and product states. The change in the relative free energy between the reactant and product states ($\Delta\Delta G$) is proportional to the change in the activation energy ($\Delta\Delta G_{act}$) for a reaction in a mixed solvent system compared to the same reaction in pure water. The free energies are drawn relative to the reactant state in pure water. (d) Thermodynamic cycle to calculate the free energy difference between a reactant and product in a mixed solvent system relative to pure water. Purple arrows indicate solvation free energies computed from MD simulations which are used to calculate the transfer free energies indicated by filled black arrows. The dashed black arrow indicates $\Delta\Delta G$. These images were adapted with permission of 26,28.

7.3.4 Step 4. Probe selected solvent systems using experiments

The results of the MD-enabled solvent screening processes outlined in Steps 2 and 3 above are:

1. a list of mixed solvent systems ranked by SolventNet-generated predictions of reaction rates as a function of solvent composition for each step in the reaction network and;

2. a separate list of mixed solvent systems ranked by solvation-free-energy-enabled predictions as to which systems may best enhance the selectivity of each reaction step toward a desired product.

These ordered lists of candidate mixed solvent systems is then used to guide experiments toward confirming best performers, as determined by solvent systems that maximize selectivity toward a desired product. Once the most selective solvent system has been confirmed by experimentation, other reaction conditions, such as temperature or reactant concentration, can be varied to check for the effect of these parameters on reaction performance. If desired, this optimized set of reaction conditions can be used as a new reference solvent system, and the solvent selection process can be repeated from step one, forming a feedback loop to further improve reaction performance, or generate reaction performance data for techno-

economic analyses.

7.4 Case Study 1: cyclohexanol dehydration to cyclohexene

The Brønsted-acid-catalyzed dehydration of cyclohexanol affords cyclohexene, as displayed in Figure 7.4a. This reaction proceeds via a known sequence of elementary steps, so that some of the factors which control the reactivity of cyclohexanol are understood from first principles.[42,43] However, the reactive, carbocation-like intermediates underlying this reaction mechanism readily participate in a series of side reactions, so that the selectivity to cyclohexanol is often less than 100%.[44] Together, these factors make cyclohexanol dehydration of general interest as a probe reaction to understand the factors which control liquid-phase, acid catalyzed biomass conversion reactions.[45] Accordingly, we now demonstrate how the tools described above can be used to anticipate the effects of mixed solvent systems in modulating the selectivity of cyclohexanol dehydration to afford cyclohexene.

Cyclohexanol and cyclohexene are both converted to unaccountable degradation products (humins) in the presence of an acid catalyst, but for simplicity only the formation of cyclohexene is considered here. The rate constant associated with this reaction was obtained in pure water at

160°C and is in agreement with the values reported elsewhere.[42,45] Water was chosen as a reference state solvent system for this example, though it should be noted that the product (cyclohexene), is insoluble in water. In THF-, DIOX-, dimethyl sulfoxide (DMSO)-, GVL-, and acetone (ACE)-water mixtures, a minumum amount of ~75 wt% organic cosolvent was added to water to achieve complete mixing of a 1 wt% cyclohexene solution. To facilitate the computational solvent screening process, the mass fraction of the organic phase for each candidate cosolvent-water mixture is fixed at 75 wt% of the organic phase, as a lower limit. Accordingly, we consider 75 wt% mixtures of ten common organic cosolvents, which are shown in Figure 7.4b.

Following the computation screening steps outlined in Figure 7.2, we first sort this initial library of ten cosolvent-water systems using SolventNet-predicted σ values for cyclohexanol dehydration as shown in Figure 7.4c. SolventNet predicts negative σ values for all mixed solvent systems that indicate the reactivity of cyclohexanol is suppressed compared to the same reaction in pure water. Despite the suppression of reactivity, a minimum amount of organic cosolvent is necessary to facilitate the solubilization of cyclohexene. In this context, the solvent systems which might best promote the reactivity of cyclohexanol are therefore those with the least negative σ values. Following this criterion, THF, GVL and ACE are anticipated to best promote the reactivity of cyclohexanol, whereas DMSO and DIOX are anticipated to suppress the reactivity of cyclohexanol.

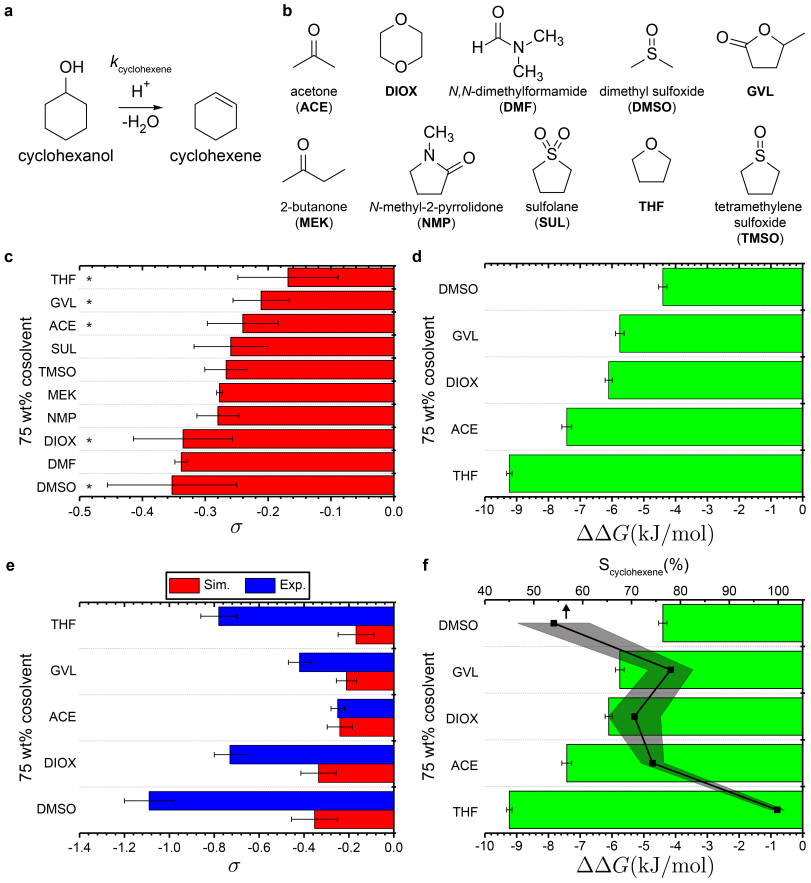


Figure 7.4: Case study of cyclohexanol conversion to cyclohexene. (a) Acid-catalyzed conversion of cyclohexanol to cyclohexene. (b) Ten organic, polar aprotic cosolvents considered for the initial library of mixed solvent systems. (c) Kinetic solvent parameters predicted by SolventNet for 75 wt% organic cosolvents. Black asterisks indicate solvent systems that are representative of good cosolvents (THF, GVL, and ACE), and poor cosolvents (DIOX and DMSO). (d) Relative solvation free energies (ΔΔG) between product (cyclohexene) and reactant state (cyclohexanol) in 75 wt% organic cosolvents relative to pure water. (e) Comparison between kinetic solvent parameters as predicted by SolventNet (red) and determined by experiments (blue). (f) Comparison between ΔΔG (green bars) and experimental percent selectivity (black line) towards cyclohexene. Gray regions denote the error in selectivity measurements.

As discussed above, SolventNet-predicted changes in reactivity (as expressed by σ values) apply only to the overall reactivity of the cyclohexanol, and cannot explicitly distinguish between changes in selectivity towards a desired product (*e.g.*, cyclohexene). To anticipate which mixed solvent systems promote reactivity toward a desired product, we therefore interrogate a subset of these ordered cosolvent-water systems using the more computationally demanding solvation free energy calculations, as described above. Note that in practice we would prescribe analyzing solvation free energies for only those solvent systems that are predicted to most enhance the overall reactivity of cyclohexanol. In this example, a subset of candidate solvent systems that meets this criterion would be THF, ACE and GVL. Herein, however, we select DMSO, DIOX, GVL, ACE and THF, which span the range of SolventNet-predicted σ values, to demonstate whether trends in overall reactivity can predict trends in selectivity for a given reaction step.

Figure 7.4d shows solvation free energy differences (ΔΔG) between cyclohexanol and cyclohexene in 75 wt% DMSO, DIOX, GVL, ACE, and THF relative to pure water. Negative ΔΔG values are obtained for all five mixed solvent systems, indicating that the solvent-induced reaction free energy change is negative in all mixed solvent systems relative to pure water. 75 wt% THF has the most negative ΔΔG, implying that this solvent mixture is predicted to best promote the selectivity towards cyclohexene. In contrast, 75 wt% DMSO is predicted to least promote selectivity toward

cyclohexene. Note that the ordering of the solvents based on σ and ΔΔG values (Figures 7.4b and 7.4c, respectively) are not the exactly the same. This result is expected, since σ only captures solvent-induced changes in the rate of reactant conversion, without any consideration of the thermodynamic preference towards a specific product as a function of solvent composition. In contrast, ΔΔG considers the stability of a specific product state relative to the reactant state as function of solvent composition. Despite this difference, σ and ΔΔG both predict that 75 wt% THF will best facilitate the selective conversion of cyclohexanol into cyclohexene, whereas 75 wt% DMSO would be the worst-performing solvent mixture in this regard.

Figures 7.4e and 7.4f compare the predicted and experimentally determined values for the rate and selectivity of cyclohexanol conversion to cyclohexene, respectively, for 75 wt% DMSO, DIOX, GVL, ACE, and THF. As shown in Figure 7.4e, the experimental kinetic solvent parameters of cyclohexanol agree with the sign predicted by SolventNet, where σ is negative for all mixed solvent mixtures in this study. However, SolventNet does not quantitatively capture the experimental trends that have σ values less than -0.5. This behavior is likely attributable to the fact that SolventNet was trained on datasets that did not include rate values below σ.[27] The errors in predicted σ values during SolventNet validation were all less than or equal to ~0.15,[27] but we suspect the error is larger for the case studies presented here due to the limitations of the training data.

This limitation could be addressed by an integrated experimental and computational feedback approach to improve predictions by retraining SolventNet with reactants or solvent mixtures that have distinct properties from the original training set. THF also appears to be a strong outlier in the predictions. This behavior may be due to the unusual phase behavior of THF-water mixtures, which exhibit a closed-loop miscibility gap.[46] Our previous work also found that 90 wt% THF-water systems show significant water-enrichment around 1,2-propanediol that is distinct from from other cosolvents, such as GVL and DIOX.[28] These unique properties of THF-water mixtures may explain the large error in values of σ predicted by SolventNet relative to the experimental measurements. Despite this quantitative disagreement, SolventNet captures the general reactivity behaviors, where 75 wt% DMSO supresses the reactivity of cyclohexanol to the greatest extent, while GVL and DIOX suppress the reactivity of cyclohexanol to a lesser extent.

Comparing the ΔΔG values in the selected solvent systems to the experimentally determined reaction selectivities, we see that the solvation free energy calculations are able to capture the key behaviors that 75 wt% THF produces cyclohexene with the greatest selectivity, while 75 wt% DMSO produces cyclohexene with the lowest selectivity. In practice and without experimental data to guide the solvent design process, the computational tools have correctly identified THF-water mixtures as "good" candidate solvent systems to facilitate the selective dehydration of cyclohexanol into

cyclohexene, and this prediction would then be confirmed with a single experiment. Together, these results demonstrate how the MD- and machine-learning based tools can be used to guide the solvent selection process for acid-catalyzed dehydration reactions, reducing the experimental burden of empirically screening a large library possible cosolvent-water combinations.

It should be noted that fully characterizing the solvent's role in controlling selectivity towards a specfic product would be best enabled by calculating the solvent-induced reaction free energy change for every possible product generated in parallel. Modeling all possible products is often not possible in practice because polymeric degredation products (humins) afforded by acid-catalyzed reactions of oxygenated compounds can be complex and difficult to characterize, which is especially true of acid-catalyzed reaction schemes underlying the decomposition of real biomass. For example, the acid-catalyzed degradation of cychexanol likely affords a range of condensation products such as such as dicyclohexyl ether or higher oligmoers, or etherified aggregates of cyclohexene and organic cosolvent molecules.[44,45] However, these degradation products were not observed by GC-FID, likely owing to their high boiling points. Despite this limitation, the agreement between simulations and experiments indicate that MD-derived solvation free energy calculations that describe the thermodynamic relationship between the reactant state and a single, desired product state can predict a mixed solvent system's ability

to promote selectivity toward the same product. In the context of solvent design for biomass converion processes, this predictive power is desirable even if all possible products cannot be considered, given that the full set of possible products being formed is difficult to acertain and the full space of possible cosolvent-water systems cannot realistically be interrogated using experiments alone.

7.5 Case study 2: fructose dehydration to 5-hydroxymethylfurfural

HMF is an important platform molecule derived from biomass, which is produced by the partial dehydration of FRU over Brønsted-acid catalysts.[47] However, in a pure aqueous solvent, HMF yields are limited by: (1) the formation of humins,[48] which are produced in parallel with HMF via the acid-catalyzed polymerization of FRU, and (2) the hydrolysis of HMF to form stoichiometric amounts of formic and levulinic acids (LA).[49] As a result, the highest yields of HMF obtained from FRU in pure water are about 40% at 100% fructose conversion.[50] Exhaustive experimental efforts over the past decade have identified a number of solvent systems that will facilitate this reaction in higher yields than pure water.[17,51,52] Important among these alternative solvent systems are mixtures of water with DMSO.[21,53] Following the generalized solvent design procedure outlined above, we now demonstrate how DMSO-water mixtures could be readily

identified from among a subset of other common industrial solvents as a promising candidate to facilitate the production of HMF in high yield.

Figure 7.5a shows a simplified reaction scheme describing the acid-catalyzed conversion of FRU to HMF, and subsequent conversion of HMF into LA. FRU and HMF are both converted to humins, but for simplicity only the formations of HMF and LA are considered here. The rate constants associated with each step in the reaction scheme were obtained in pure water at 130°C, and are in agreement with the values reported elsewhere.[54] Water was chosen as a reference state solvent system for this example, because the reactant and products are all water-soluble. However, FRU is only minimally soluble in most cosolvent-water mixtures when the organic phase is present in mass fractions above 90 wt%, resulting in an upper limit to the cosolvent concentration. Therefore, to facilitate the computational solvent screening process, we focus on 90 wt% cosolvent for each candidate mixed solvent system.

Following the procedure outlined in Figure 7.2, we first analyze MD simuations of FRU and HMF in the 90 wt% cosolvent-water systems using SolventNet. In this example, since the intermediate dehydration product (HMF) is desired, we are interested in cosolvent-water systems that will selectively enhance the reactivity of FRU over HMF. Therefore, we sort the candidate cosolvent-water systems based on the differences in the predicted kinetic solvent parameters between HMF and FRU as a function of solvent system ($\sigma_{HMF} - \sigma_{FRU}$). More negative values of $\sigma_{HMF} - \sigma_{FRU}$

indicate that the rate of HMF production is enhanced to a greater extent than the rate of LA formation, and selectivity to HMF is therefore promoted. Figure 7.5b shows these SolventNet-predicted $\sigma_{HMF} - \sigma_{FRU}$ values across the same ten cosolvent-water systems used in the previous case study (Figure 7.4b) at 90 wt% cosolvent composition. SolventNet-predicted $\sigma_{HMF} - \sigma_{FRU}$ values indicate that 90 wt% GVL, TMSO and DMSO might best facilitate the selective conversion of FRU into HMF while suppressing the formation of LA, whereas 90 wt% ACE and THF are predicted to be the worst-performing mixed solvent systems. 90 wt% DIOX is predicted to perform in a capacity intermediate between these extrema. Therefore, as in the previous example, we select GVL, DMSO, DIOX, ACE and THF for further analysis using the solvation free energy calculations. We note again that, in practice, only those mixed solvent systems that SolventNet predicts to most enhance the reactivity of the desired reaction step would be interrogated further using solvation free energy calculations, whereas in this work, we select solvent systems spanning the full range of predictied reactivities for validation purposes.

Since this is a series reaction, we focus on suppression of the LA product. Accordingly, Figure 7.5d shows the $\Delta\Delta G$ values between FRU and LA ($\Delta\Delta G_{FRU \rightarrow LA}$) for 90 wt% GVL, DMSO, DIOX, ACE and THF. Note that positive $\Delta\Delta G_{FRU \rightarrow LA}$ values correspond to the suppression of the undesired product (LA). Of the five cosolvent-water systems, we find that 90 wt% DMSO is the only mixed solvent systems that is anticipated to suppress

the formation of LA, whereas all other mixed solvent systemsstabilize the formation of LA from FRU. Therefore, between the SolventNet-predicted $\sigma_{HMF} - \sigma_{FRU}$ values and the MD-estimated $\Delta\Delta G_{FRU \rightarrow LA}$ values, 90 wt% THF is predicted to effect the lowest selectivity towards HMF, whereas 90 wt% DMSO is predicted to effect the highest selectivity towards HMF.

Figure 7.5d compares the predicted and experimentally determined $\sigma_{HMF} - \sigma_{FRU}$ values. While SolventNet does not capture quantitative agreement with experiments, the predicted values capture the general trend that 90 wt% GVL and DMSO are the best-performing mixed solvent systems and 90 wt% DIOX, THF, and ACE are the worst-performing mixed solvent systems. Figure 7.5e compares the $\Delta\Delta G_{FRU \rightarrow LA}$ values to the percent selectivity towards HMF (S_{HMF}) in 90 wt% DMSO, GVL, ACE, DIOX, and THF. The suppression behavior found from positive $\Delta\Delta G_{FRU \rightarrow LA}$ values in 90 wt% DMSO correctly corresponds with the highest selectivity towards HMF. $\Delta\Delta G_{FRU \rightarrow LA}$ also correctly predicts that 90 wt% THF would be the worst-performing solvent with low HMF selectivities of ~20%.

Together, these results demonstrate how the generalized workflow described in Figure 7.2 could be used to identify 90 wt% DMSO as a promising solvent system to improve the yield HMF from FRU as compared to pure water, in agreement with experimental findings from multiple reports. However, these results also indicate that 90 wt% ACE and THF solvent systems perform poorly in selectively converting FRU to HMF. This finding is in constrast to recent reports that ACE-water mixtures

are able to produce HMF from FRU in high yield.[51] Other studies found that THF-water mixtures facilitate the conversion of whole biomass into fermentable sugars and furanic fuels precersuors in high yield.[55,56] We rationalize these discrepancies by noting that these reports utilized different water contents, temperatures, and acid catalysts than those investigated in this work. Furthermore, the reactant material in the reports by Wyman and coworkers was whole biomass, a complex mixture of carbohydrate- and lignin-derived molecules,[55,56] so that the chemical intermediates between this starting material and the final products are likely distinct from the well-defined FRU/HMF system studied herein. Together, these results illuminate some imporant limitations of our approach: (1) reaction kinetics can only be estimated at a fixed temperature as a function of solvent systems; (2) differences in the acidic proton's conjugate base are not incorporated into our predictions, and; (3) model reactions do not capture the complex reaction schemes underlying the transformation of real biomass-derived materials. As a result, future work will focus on incorporating the effects of temperature and catalyst selection to the design of efficient reaction systems for biomass conversion processes.

Previous *ab initio* MD simulations have shown that hydride shifts in FRU dehydration are key steps to forming HMF, which may occur intramolecularly or through solvent participation.[57–59] A limitation of the current framework described in Figure 7.2 is that it does not probe changes in reaction pathway due to the addition of cosolvents. However, the com-

putational approach can identify top candidate mixed solvent systems that are predicted to have enhanced reactivity or selectivity compared to pure water that could then be further interrogated with *ab initio* calculations to capture potential mechanistic changes due to solvent participation.[28] Analysis of the top candidate mixed solvent systems could also be used to understand the behavior that underlies improved reactivity and selectivity. For example, the preferential solvation of hydrophilic reactants by water and long reactant-water hydrogen bonding lifetimes are generally found in the presence of high cosolvent concentration,[25,60] leading to improved reaction rates due to preferred acid catalyst-water interactions.[24–26] However, significant differences in solvation environment between the reactant and product could lead to decreased reactivity because of the extent of solvent rearrangement required to stabilize the product state, especially in the presence of highly polarizable cosolvents such as DMSO.[61] Future work will thus focus on characterizing the solute-solvent interactions for top candidate mixed solvent systems identified via this computational approach, which could then give insight into solvent effects on improved reaction performance.

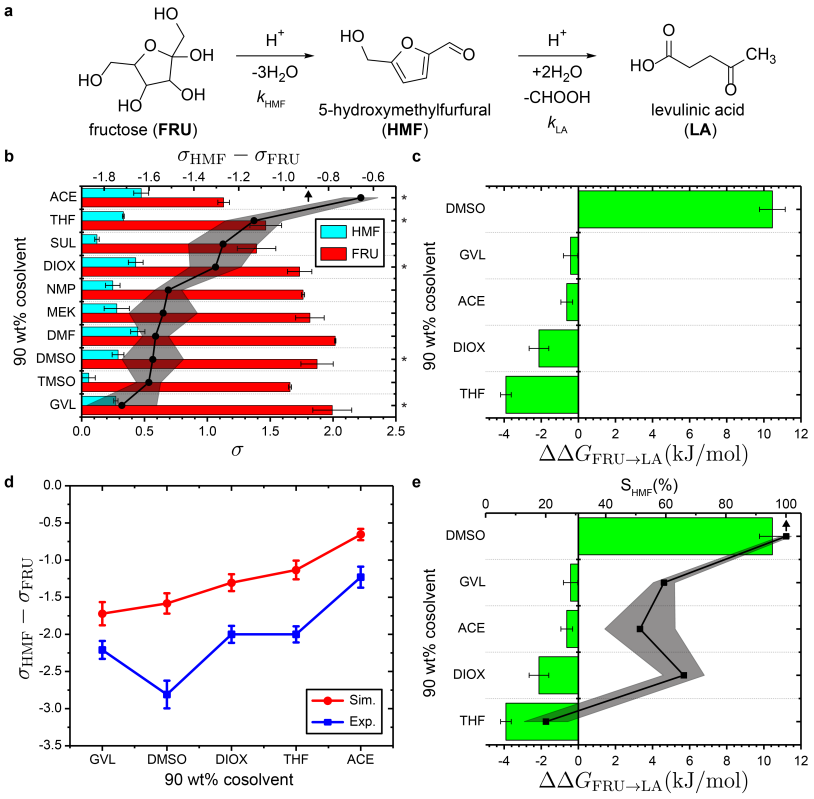


Figure 7.5: Case study of fructose conversion to HMF. (a) Acid-catalyzed dehydration of fructose (FRU) to afford 5-hydroxymethylfurfural (HMF), followed by an addition reaction to afford levulinic acid (LA). (b) Kinetic solvent parameters predicted by SolventNet for 90 wt% cosolvent-water mixtures. The difference between kinetic solvent parameters predicted for HMF and FRU is shown in the top axis ($\sigma_{HMF} - \sigma_{FRU}$), where more negative $\sigma_{HMF} - \sigma_{FRU}$ values indicate that the mixed solvent system is predicted to selectively form HMF from FRU. Black asterisks indicate mixed solvent systems that are representative of good cosolvents (THF, GVL, and ACE), and poor cosolvents (DIOX and DMSO). (c) Relative solvation free energies ($\Delta\Delta G_{FRU\rightarrow LA}$) between product (LA) and reactant state (FRU) in 90 wt% organic cosolvents relative to pure water. (d) Comparison between $\sigma_{HMF} - \sigma_{FRU}$ predicted from SolventNet (red) and experimentally determined (blue). (e) Comparison between $\Delta\Delta G_{FRU\rightarrow LA}$ and percent selectivity towards HMF. Gray regions denote the error in $\sigma_{HMF} - \sigma_{FRU}$ in (b) and selectivity measurements in (e).

7.6 Summary

We have developed a general workflow to select mixed solvent systems for enhanced reaction rates and selectivities of acid-catalyzed reactions for biomass conversion processes using computationally efficient methods and minimal experiments. Classical MD simulations and machine learning tools were used to estimate the influence of solvent composition on the reactivity of biomass-derived molecules by analyzing the solvent environment around the reactant. Furthermore, MD simulations can estimate solvent-induces changes in the selectivity of biomass conversion reactions to specific products by quantifying solvation free energies for reactant and product states as a function of solvent composition. Together, these MD- and machine-learning-based tools are combined into a workflow for selecting mixed solvent systems for biomass conversion applications with minimal experiments.

We have demonstrated the utility of this workflow by analyzing two case studies: the dehydration of cyclohexanol into cyclohexene and the acid-catalyzed partial dehydration of fructose to 5-hydroxymethylfurfural. In both case studies, the MD- and machine-learning based tools predict trends that enable the selection of best-performing solvent systems (as expressed by their ability to faciliate the desired reactions in high yield) to be down-selected from a large library of candidate solvent systems that would be laborous and cost-inefficient to investigate using experimental

screening methods alone. These results demonstrate that computationally efficient screening methods and predictive design tools allow for the selection of top candidate mixed solvent systems, which could be then further interrogated to determine how solute-solvent interactions drive reactivity in liquid phase catalytic processes. These efforts therefore represent a step toward a molecular-level understanding the role of mixed solvent systems in controlling reactivity of liquid phase reactions, and toward the rational design of mixed solvent systems in a framework that reduces the experimental burden that accompanies the development of new chemical processes.

7.7 References

[1] Walker, T. W.; Chew, A. K.; Van Lehn, R. C.; Dumesic, J. A.; Huber, G. W. Rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using a combined experimental, molecular dynamics and machine learning approach. *Topics in Catalysis* **2020**, *63*, 649–663.

[2] Walker, T. W.; Chew, A. K.; Van Lehn, R. C.; Dumesic, J. A.; Huber, G. W. Rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using a combined experimental, molecular dynamics and machine learning approach [Supporting Information]. *Topics in Catalysis* **2020**, *63*, 649–663.

[3] Dumesic, J.; Topsøe, H.; Khammouma, S.; Boudart, M. Surface, catalytic and magnetic properties of small iron particles: II. Structure sensitivity of ammonia synthesis. *Journal of Catalysis* **1975**, *37*, 503–512.

[4] Ledesma, C.; Yang, J.; Chen, D.; Holmen, A. Recent approaches in mechanistic and kinetic studies of catalytic reactions using SSITKA technique. *ACS Catalysis* **2014**, *4*, 4527–4547.

[5] Fan, L.; Ziegler, T. Nonlocal density functional theory as a practical tool in calculations on transition states and activation energies. Applications to elementary reaction steps in organic chemistry. *Journal of the American Chemical Society* **1992**, *114*, 10890–10897.

[6] Gokhale, A. A.; Kandoi, S.; Greeley, J. P.; Mavrikakis, M.; Dumesic, J. A. Molecular-level descriptions of surface chemistry in kinetic models using density functional theory. *Chemical Engineering Science* **2004**, *59*, 4679–4691.

[7] Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nature chemistry* **2009**, *1*, 37–46.

[8] Zhang, L.; Roling, L. T.; Wang, X.; Vara, M.; Chi, M.; Liu, J.; Choi, S.-I.; Park, J.; Herron, J. A.; Xie, Z.; et al.. Platinum-based nanocages with subnanometer-thick walls and well-defined, controllable facets. *Science* **2015**, *349*, 412–416.

[9] Nørskov, J. K.; Bligaard, T.; Logadottir, A.; Bahn, S.; Hansen, L. B.; Bollinger, M.; Bengaard, H.; Hammer, B.; Sljivancanin, Z.; Mavrikakis, M.; et al.. Universality in heterogeneous catalysis. *Journal of catalysis* **2002**, *209*, 275–278.

[10] Centi, G.; Perathoner, S. Catalysis: role and challenges for a sustainable energy. *Topics in Catalysis* **2009**, *52*, 948–961.

[11] Chheda, J. N.; Huber, G. W.; Dumesic, J. A. Liquid-phase catalytic processing of biomass-derived oxygenated hydrocarbons to fuels and chemicals. *Angewandte Chemie International Edition* **2007**, *46*, 7164–7183.

[12] Shuai, L.; Luterbacher, J. Organic solvent effects in biomass conversion reactions. *ChemSusChem* **2016**, *9*, 133–155.

[13] Walker, T. W.; Motagamwala, A. H.; Dumesic, J. A.; Huber, G. W. Fundamental catalytic challenges to design improved biomass conversion technologies. *Journal of Catalysis* **2018**, *369*.

[14] Serrano-Ruiz, J. C.; Dumesic, J. A. Catalytic routes for the conversion of biomass into liquid hydrocarbon transportation fuels. *Energy & Environmental Science* **2011**, *4*, 83–99.

[15] Mellmer, M. A.; Sener, C.; Gallo, J. M. R.; Luterbacher, J. S.; Alonso, D. M.; Dumesic, J. A. Solvent effects in acid-catalyzed biomass conversion reactions. *Angewandte chemie international edition* **2014**, *53*, 11872–11875.

[16] Maugh, I.; et al.. Catalysis: no longer a black art. *Science (Washington, DC);(United States)* **1983**, *219*.

[17] Román-Leshkov, Y.; Chheda, J. N.; Dumesic, J. A. Phase modifiers promote efficient production of hydroxymethylfurfural from fructose. *Science* **2006**, *312*, 1933–1937.

[18] Luterbacher, J. S.; Rand, J. M.; Alonso, D. M.; Han, J.; Youngquist, J. T.; Maravelias, C. T.; Pfleger, B. F.; Dumesic, J. A. Nonenzymatic sugar production from biomass using biomass-derived γ-valerolactone. *Science* **2014**, *343*, 277–280.

[19] Shuai, L.; Questell-Santiago, Y. M.; Luterbacher, J. S. A mild biomass pretreatment using γ-valerolactone for concentrated sugar production. *Green Chemistry* **2016**, *18*, 937–943.

[20] Motagamwala, A. H.; Won, W.; Maravelias, C. T.; Dumesic, J. A. An engineered solvent system for sugar production from lignocellulosic biomass using biomass derived γ-valerolactone. *Green Chemistry* **2016**, *18*, 5756–5763.

[21] Mushrif, S. H.; Caratzoulas, S.; Vlachos, D. G. Understanding solvent effects in the selective conversion of fructose to 5-hydroxymethylfurfural: a molecular dynamics investigation. *Physical Chemistry Chemical Physics* **2012**, *14*, 2637–2644.

[22] Christianson, J. R.; Caratzoulas, S.; Vlachos, D. G. Computational insight into the effect of Sn-beta Na exchange and solvent on glucose isomerization and epimerization. *ACS Catalysis* **2015**, *5*, 5256–5263.

[23] Assary, R. S.; Redfern, P. C.; Hammond, J. R.; Greeley, J.; Curtiss, L. A. Computational studies of the thermochemistry for conversion

of glucose to levulinic acid. *The Journal of Physical Chemistry B* **2010**, *114*, 9002–9009.

[24] Mellmer, M. A.; Sanpitakseree, C.; Demir, B.; Bai, P.; Ma, K.; Neurock, M.; Dumesic, J. A. Solvent-enabled control of reactivity for liquid-phase reactions of biomass-derived compounds. *Nature Catalysis* **2018**, *1*, 199–207.

[25] Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Van Lehn, R. C.; Dumesic, J. A. Universal kinetic solvent effects in acid-catalyzed reactions of biomass-derived oxygenates. *Energy & Environmental Science* **2018**, *11*, 617–628.

[26] Chew, A. K.; Van Lehn, R. C. Quantifying the stability of the hydronium ion in organic solvents with molecular dynamics simulations. *Frontiers in chemistry* **2019**, *7*, 439.

[27] Chew, A. K.; Jiang, S.; Zhang, W.; Zavala, V. M.; Van Lehn, R. C. Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks. *Chemical Science* **2020**, *11*, 12464–12476.

[28] Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Van Lehn, R. C. Effect of mixed-solvent environments on the selectivity of acid-catalyzed dehydration reactions. *ACS Catalysis* **2019**, *10*, 1679–1691.

[29] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[30] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[31] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules

compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[32] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[33] Berendsen, H.; Grigera, J.; Straatsma, T. The missing term in effective pair potentials. *Journal of Physical Chemistry* **1987**, *91*, 6269–6271.

[34] Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics* **2008**, *129*, 124105.

[35] Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of computer-aided molecular design* **2015**, *29*, 397–411.

[36] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[37] Daoutidis, P.; Marvin, W. A.; Rangarajan, S.; Torres, A. I. Engineering biomass conversion processes: a systems perspective. *AIChE Journal* **2013**, *59*, 3–18.

[38] Madon, R. J.; Iglesia, E. Catalytic reaction rates in thermodynamically non-ideal systems. *Journal of Molecular Catalysis A: Chemical* **2000**, *163*, 189–204.

[39] Krishna, S. H.; Walker, T. W.; Dumesic, J. A.; Huber, G. W. Kinetics of levoglucosenone isomerization. *ChemSusChem* **2016**, *10*.

[40] Lindsay, M. J.; Walker, T. W.; Dumesic, J. A.; Rankin, S. A.; Huber, G. W. Production of monosaccharides and whey protein from acid whey waste streams in the dairy industry. *Green Chemistry* **2018**, *20*, 1824–1834.

[41] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al.. Imagenet

large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.

[42] Liu, Y.; Vjunov, A.; Shi, H.; Eckstein, S.; Camaioni, D. M.; Mei, D.; Baráth, E.; Lercher, J. A. Enhancing the catalytic activity of hydronium ions through constrained environments. *Nature communications* **2017**, *8*, 1–8.

[43] Mei, D.; Lercher, J. A. Effects of Local Water Concentrations on Cyclohexanol Dehydration in H-BEA Zeolites. *The Journal of Physical Chemistry C* **2019**, *123*, 25255–25266.

[44] He, J.; Zhao, C.; Lercher, J. A. Impact of solvent for individual steps of phenol hydrodeoxygenation with Pd/C and HZSM-5 as catalysts. *Journal of catalysis* **2014**, *309*, 362–375.

[45] Wang, X.; Rinaldi, R. A route for lignin and bio-oil conversion: dehydroxylation of phenols into arenes by catalytic tandem reactions. *Angewandte Chemie* **2013**, *125*, 11713–11717.

[46] Smith, M. D.; Mostofian, B.; Petridis, L.; Cheng, X.; Smith, J. C. Molecular driving forces behind the tetrahydrofuran–Water miscibility gap. *The Journal of Physical Chemistry B* **2016**, *120*, 740–747.

[47] van Putten, R.-J.; Van Der Waal, J. C.; De Jong, E.; Rasrendra, C. B.; Heeres, H. J.; de Vries, J. G. Hydroxymethylfurfural, a versatile platform chemical made from renewable resources. *Chemical reviews* **2013**, *113*, 1499–1597.

[48] Patil, S. K.; Lund, C. R. Formation and growth of humins via aldol addition and condensation during acid-catalyzed conversion of 5-hydroxymethylfurfural. *Energy & Fuels* **2011**, *25*, 4745–4755.

[49] Asghari, F. S.; Yoshida, H. Kinetics of the decomposition of fructose catalyzed by hydrochloric acid in subcritical water: formation of 5-hydroxymethylfurfural, levulinic, and formic acids. *Industrial & Engineering Chemistry Research* **2007**, *46*, 7703–7710.

[50] Teong, S. P.; Yi, G.; Zhang, Y. Hydroxymethylfurfural production from bioresources: past, present and future. *Green Chemistry* **2014**, *16*, 2015–2026.

[51] Motagamwala, A. H.; Huang, K.; Maravelias, C. T.; Dumesic, J. A. Solvent system for effective near-term production of hydroxymethylfurfural (HMF) with potential for long-term process improvement. *Energy & Environmental Science* **2019**, *12*, 2212–2222.

[52] Qi, X.; Watanabe, M.; Aida, T. M.; Smith Jr, R. L. Selective conversion of D-fructose to 5-hydroxymethylfurfural by ion-exchange resin in acetone/dimethyl sulfoxide solvent mixtures. *Industrial & engineering chemistry research* **2008**, *47*, 9234–9239.

[53] Tsilomelekis, G.; Josephson, T. R.; Nikolakis, V.; Caratzoulas, S. Origin of 5-hydroxymethylfurfural stability in water/dimethyl sulfoxide mixtures. *ChemSusChem* **2014**, *7*, 117–126.

[54] Weingarten, R.; Cho, J.; Xing, R.; Conner Jr, W. C.; Huber, G. W. Kinetics and reaction engineering of levulinic acid production from aqueous glucose solutions. *ChemSusChem* **2012**, *5*, 1280–1290.

[55] Cai, C. M.; Zhang, T.; Kumar, R.; Wyman, C. E. THF co-solvent enhances hydrocarbon fuel precursor yields from lignocellulosic biomass. *Green Chemistry* **2013**, *15*, 3140–3145.

[56] Smith, M. D.; Mostofian, B.; Cheng, X.; Petridis, L.; Cai, C. M.; Wyman, C. E.; Smith, J. C. Cosolvent pretreatment in cellulosic biofuel production: effect of tetrahydrofuran-water on lignin structure and dynamics. *Green Chemistry* **2016**, *18*, 1268–1277.

[57] Mushrif, S. H.; Varghese, J. J.; Krishnamurthy, C. B. Solvation dynamics and energetics of intramolecular hydride transfer reactions in biomass conversion. *Physical Chemistry Chemical Physics* **2015**, *17*, 4961–4969.

[58] Nikbin, N.; Caratzoulas, S.; Vlachos, D. G. A First Principles-Based Microkinetic Model for the Conversion of Fructose to 5-Hydroxymethylfurfural. *ChemCatChem* **2012**, *4*, 504–511.

[59] Varghese, J. J.; Mushrif, S. H. Origins of complex solvent effects on chemical reactivity and computational tools to investigate them: a review. *Reaction Chemistry & Engineering* **2019**, *4*, 165–206.

[60] Vasudevan, V.; Mushrif, S. H. Insights into the solvation of glucose in water, dimethyl sulfoxide (DMSO), tetrahydrofuran (THF) and N, N-dimethylformamide (DMF) and its possible implications on the conversion of glucose to platform chemicals. *Rsc Advances* **2015**, *5*, 20756–20763.

[61] Hazra, M. K.; Bagchi, B. Non-equilibrium solvation dynamics in water-DMSO binary mixture: Composition dependence of non-linear relaxation. *The Journal of chemical physics* **2018**, *149*, 084501.

8 EFFECT OF CORE MORPHOLOGY ON THE STRUCTURAL ASYMMETRY OF ALKANETHIOL MONOLAYER-PROTECTED GOLD NANOPARTICLES

Chapter 7 shows the success of integrating molecular dynamics simulations and machine learning to guide solvent selection in biomass conversion reactions. One could envision that those computational tools could be applied to more complex systems. This chapter and subsequent chapters (9-11) focuses on integrating molecular dynamics simulation and machine learning to guide the design of monolayer-protected gold nanoparticles for biomedical applications. This chapter seeks to address the following questions:

- How do we develop a generalized workflow to model gold nanoparticle systems, which could account for variations in gold core shape and size, and ligand selection?
- How does the shape and size of the gold core affect monolayer characteristics?
- How do we encode complex, cooperative behavior (*e.g.* formation of bundles) into a molecular descriptor that could provide physical description of the monolayer?

This chapter was reproduced with permission from Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.[1] Copyright 2021 American Chemical Society. The supporting information is cited as Ref. 2.

In this chapter, atomistic molecular dynamics simulations were used to study the influence of gold core morphology, size, and ligand length on the structure of uniformly nonpolar alkanethiol monolayer-protected GNPs in water. Using a generalized system preparation workflow, three gold core models were selected for this study: (1) a uniformly spherical hollow gold core, (2) a spherical gold core cut from a bulk gold lattice, and (3) a faceted gold core obtained from variance-constrained semigrand-canonical simulations. Independent of the gold core morphology, we found that long alkanethiol ligands exhibit increased ligand order and form quasi-crystalline domains, or bundles, in which ligands orient in the same direction, leading to asymmetric monolayer structures. Faceted gold cores promote the formation of ligand bundles for short ligand lengths, but the influence of the gold core is diminished for long ligand lengths. We used a clustering algorithm to identify the subpopulation of bundled ligands and found that bundling leads to heterogeneous surface properties, whereby bundled ligands have a higher ligand order and lower surface area accessible for solvent interactions. These findings demonstrate the importance of GNP compositional features on monolayer structure, which could impact GNPs interactions with other molecules.

8.1 Introduction

Gold nanoparticles (GNPs) have attracted significant interest for biomedical applications because of their low cytotoxicity, stability, and detectability *in vivo*.[3–5] Particularly prominent are GNPs protected by self-assembled monolayers (SAMs) consisting of alkanethiol ligands that are bound to the gold surface via a gold-sulfur interaction.[6,7] SAM-protected GNPs are advantageous because of their ease of fabrication and the ability to synthetically control a large number of parameters, including the gold core size, shape, ligand grafting density, and ligand chemistry, to tune GNP properties and behavior.[8,9] However, identifying how specific components of the GNP design space affect GNP properties in biological environments, or even in aqueous solution, remains an ongoing challenge, particularly because the experimental characterization of GNPs is currently limited in the level of molecular detail that can be obtained[10] and often requires a combination of multiple techniques.[11,12] New techniques are needed to develop structure-function relationships that can be used to guide GNP design.

Recently, molecular simulations have emerged as a valuable tool for informing materials design, as GNPs are small enough that their behavior can be characterized *in silico* at atomistic resolution.[13–25] Molecular simulations are particularly valuable for quantifying GNP properties that emerge from the collective behavior of SAM ligands and cannot be inferred from

single-ligand properties. An important example of how collective ligand interactions affect behavior is the observation that alkanethiol ligands grafted to GNPs spontaneously form highly asymmetric structures, even if the SAM composition is chemically homogeneous.[26,27] Simulations have shown that this asymmetry arises due to preferred interactions between nonpolar ligand backbones that drive the formation of quasi-crystalline ligand domains in vacuum,[27–29] water,[30] and nonpolar solvents.[26,31,32] This structural asymmetry plays an important role in various nanoscale processes; for example, SAM asymmetry has been shown to affect GNP-induced DNA bending,[19] GNP aggregation and colloidal stability,[30,31] GNP-protein binding,[33] and GNP interactions with lipid membranes.[34–37] The relevance of SAM asymmetry to these disparate processes has further inspired studies of factors affecting SAM asymmetry, including the temperature, GNP size, and alkanethiol tail length.[27–29]

While these prior findings demonstrate the importance of SAM structural features, most GNP simulation studies rely on assumptions regarding the morphology of the GNP core and the surface grafting density of alkanethiol ligands that could affect simulation observations. Previous authors have represented the gold core implicitly,[26,30] as a perfect, hollow sphere,[27,38] as a sphere cut from a bulk gold lattice,[39] or as a facetted geometry.[28,29] As a result, the effect of the gold core morphology on SAM asymmetry is still unclear. Specifically, it is possible that gold core facets, which are predicted computationally[40] and observed experimentally[41–44]

promote the formation of quasi-crystalline ligand domains. Previous authors have also used different approaches to approximate SAM coverage, typically either assuming uniform ligand grafting[26,31,32] or self-assembling ligands onto the surface via molecular simulations.[28,29,45,46] These different assumptions regarding the treatment of the gold core and SAM coverage make comparisons across the literature challenging, especially since ligand grafting densities depend on GNP size.[39,47] Finally, parameters used to quantify asymmetry have included the free volume per ligand,[26] center of mass differences between the ligand and GNP core,[48] and eigenvalues of the moment of inertia tensor.[49] These ensemble-average quantities do not directly identify ligand domains and cannot easily classify differences in properties between various regions on the GNP surface. There is thus a need for new approaches to prepare GNP simulations that account for variations in GNP morphology and ligand grafting density, enabling improved characterization of SAM structure.

In this chapter, we developed a workflow to streamline the preparation of GNPs with arbitrary core sizes, morphologies, and ligand coatings for atomistic molecular dynamics (MD) simulations. We then performed MD simulations to compare the effects of the core and ligand properties on the structural characteristics of alkanethiol SAMs. These simulations confirm that ligands organize into quasi-crystalline domains, or bundles, that lead to spatially heterogeneous, asymmetric surface properties even for uniform coatings. We used the Hierarchical Density-based Spatial Clus-

tering of Applications with Noise (HDBSCAN) clustering algorithm[50,51] to quantify ligand bundling as a function of ligand length for three gold core morphologies. Our results show that bundle formation is promoted by gold core facets for sufficiently short ligand lengths. Finally, we characterized distinguishing structural features of bundled and non-bundled ligand subpopulations, including ligand order and the available surface area for ligand water interactions. These findings provide new insight into the role of gold core physical characteristics, and specifically morphology, on SAM structure and provide a new workflow to standardize future GNP studies.

8.2 Methods

The workflow for preparing GNP simulations is summarized in Figure 8.1. A gold core of a desired size and morphology is first selected and added to a simulation box containing an excess of butanethiol ligands. These ligands then self-assemble onto the gold core during atomistic MD simulations based on protocols explored previously,[28,29,45] which allows the ligand grafting density to adjust to the gold core size and morphology rather than being predefined. The excess, unattached butanethiol ligands are removed and the remaining adsorbed ligands are replaced with desired ligands. The GNP is then solvated and equilibrated in the *NPT* ensemble. In this chapter, we focus on GNPs with nonpolar alkanethiol ligands

(see Figure 8.2A) in water, although this workflow can be adapted to develop GNPs with more complex ligands. All MD simulations were performed using GROMACS version 2016.[52] Three trials of this workflow were performed for each core morphology, size, and ligand selection. All simulation measurables are reported as the average of three trials. Error bars report the standard deviation of each measurement between the three trials.

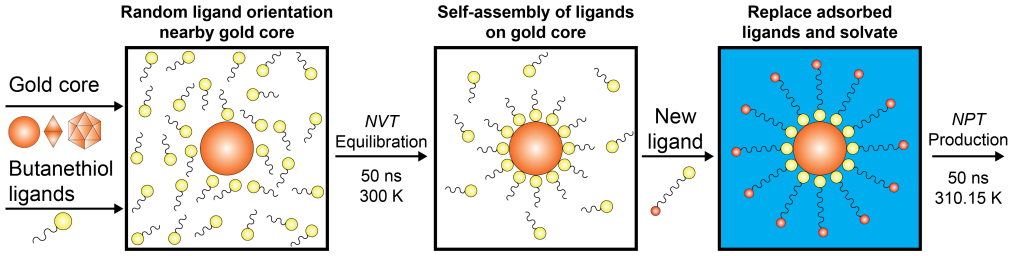


Figure 8.1: Schematic of workflow for automatic GNP assembly.

8.2.1 Gold core

Three gold core morphologies of increasing complexity were selected for this study:

1. a hollow gold core (HGC), in which gold atoms were uniformly distributed on a spherical surface;

2. a spherical gold core (SGC), in which an approximately spherical core was truncated from the bulk gold face-centered cubic lattice; and

3. a faceted gold core (FGC), in which gold atom positions were obtained from the lowest-energy structure for a given core diameter obtained using variance-constrained semigrand-canonical simulations[53] with an embedded atom potential.[54]

Additional details of how the gold core morphologies were developed are included in the SI.[2] The three different morphologies differ in the coordination of surface atoms; snapshots of each morphology for varying core diameters are shown in Figure 8.2B. Notably, the FGC models are dominated by planar (111) facets and form structures similar to the icosahedral and truncated octahedral morphologies predicted by density functional theory calculations[40] and observed experimentally;[43] the HGC and SGC models thus represent extremes of small facets, while the FGC model represents the extreme of large facets. We approximate the diameter of each gold core as the maximum gold-gold pairwise distance. Diameters between 2-7 nm were modeled for each morphology. The FGC models have average diameters of 2.0, 2.9, 3.8, 4.8, 5.7, and 6.7 nm, which we denote for simplicity as 2, 3, 4, 5, 6, and 7 nm, respectively.

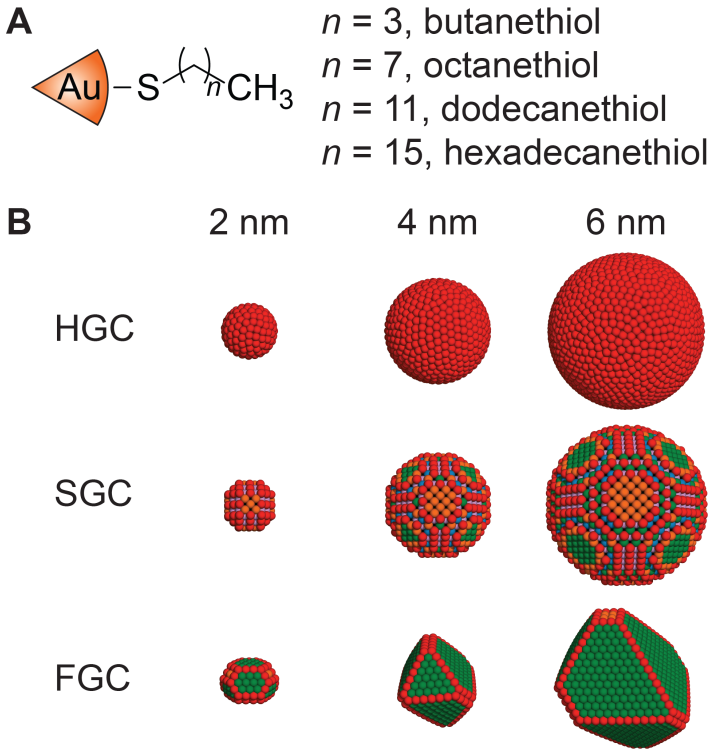


Figure 8.2: **A** Alkanethiol ligand structures used in this study, where n denotes the number of methylene groups. **B** Snapshots of hollow gold core (HGC), spherical gold core (SGC), and faceted gold core (FGC) for various diameters. Each atom is colored by its coordination number, defined as the number of gold atoms within a cutoff of 0.41, 0.37, and 0.34 nm for HGC, SGC, and FGC, respectively. Colors correspond to coordination numbers of 1-6 (red), 7-8 (orange), 9 (green), 10 (blue) and 11 (violet).

8.2.2 Self-assembly process

We selected butanethiol as the model ligand to self-assemble onto the gold core because prior studies have found that varying the alkanethiol chain length (*i.e.*, the number of methylene groups in the alkanethiol backbone)

does not significantly affect the number of adsorbed ligands during this process.[45] For this procedure, butanethiol ligands were modeled using a united atom model, where the SH, CH_2, and CH_3 groups were each represented as a single interaction site. Gold atoms and ligands were parameterized based on the model by Djebaili and coworkers,[45] which yields good agreement with experimental results.[55] Au-S interactions were modeled using a Lennard-Jones potential, which ignores features of the Au-S bonds (*i.e.*, adatom coordination,[56] top-site grafting,[7] *etc.*), but we justify our model a posteriori based on favorable comparison to experiments as described below. This workflow also correctly leads to a grafting density of 4.62 ligands/nm^2 for a planar gold surface (SI, Figure S8).[2] Force field parameters used for the ligands and the gold for the self-assembly process are available in the SI.[2]

The gold core was placed at the center of a cubic, periodic simulation box and frozen for all self-assembly simulations. An excess number of ligands, corresponding to 20 ligands/nm^2 of GNP surface area, were randomly positioned in the simulation box. The simulation box was then expanded, leaving a 2 nm distance between the ligands and the periodic boundaries (SI, Figure S3).[2] This expansion allowed the butanethiol ligands to bind to or leave the gold surface. The system was energy minimized with the steepest descent algorithm and a force tolerance of 1000.0 kJ/(mol-nm), then MD was performed in the *NVT* ensemble using a leapfrog integrator with a 1 femtosecond time step for a total of 30 ns. The temperature was

maintained at 300 K with the Nóse-Hoover thermostat with a 0.4 ps time constant.

8.2.3 Ligand-exchange and production simulations.

Upon completion of the self-assembly process, excess butanethiol ligands were removed, leaving only butanethiol ligands that had adsorbed to the gold core (SI, Figure S9).[2] A ligand was considered adsorbed if its sulfur atom was within 0.327 nm of any gold atom.[45] Each adsorbed butanethiol ligand was then replaced with a desired alkanethiol ligand by aligning the position of the new ligand's sulfur atom with the sulfur atom of the adsorbed ligand and orienting the new ligand radially away from the center of mass of the gold core. The GNP was then solvated with water using the TIP3P model.[57,58] Adsorbed sulfur atoms were bonded to all gold atoms within 0.327 nm and gold atoms were bonded to other gold atoms within 0.290 nm. All Au-S and Au-Au bonds were modeled using a harmonic potential with a spring constant of 50,000 kJ mol^{-1} nm^{-2}. The system was energy minimized with the steepest descent algorithm and a force tolerance of 1000.0 kJ/(mol nm), then MD was performed in the *NPT* ensemble using a leapfrog integrator with a 2 femtosecond time step. The simulation was split into 5 ns of equilibration followed by 50 ns of production. Simulation configurations were output every 100 ps and the final 40 ns of each production trajectory were used for analysis. The pressure was maintained at 1 bar using the Berendsen barostat for equilibration

and the Parrinello-Rahman barostat for production. A physiological temperature of 310.15 K was maintained using a velocity rescale thermostat. All thermostats used a 0.1 ps time constant and all barostats used a 1.0 ps time constant with an isothermal compressibility of 4.5×10^{-5} bar^{-1}.

8.2.4 Simulation parameters

Gold atoms were parameterized using the INTERFACE force field[59] and ligands were parameterized using the CGenFF/CHARMM36 force fields.[60–62] We selected the INTERFACE force field because of its compatibility with the CHARMM force fields, fast performance by using simple 12-6 LJ potentials for gold parameters, and its validity at atmospheric pressures and temperatures within 298 ± 200 K[59] that are within the range of this study. The CGenFF/CHARMM36 force fields are comparable to other force fields, such as GAFF and OPLS-AA, in benchmark studies of a range of small-molecule liquids[63] and are compatible with a range of biomolecules for future studies. In all simulations, Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential using a 1.2 nm cutoff that was smoothly shifted to zero between 1.0 nm and 1.2 nm. Electrostatic interactions were calculated using the Smooth Particle Mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and 4th order interpolation. Bonds were constrained using the LINCS algorithm.[64] Periodic boundary conditions were enabled in all directions.

8.3 Results

8.3.1 Effect of core morphology on the number of grafted ligands

We first determined how the gold core morphology affects ligand surface coverage after self-assembly. Figure 8.3 compares the number of adsorbed ligands as a function of GNP core diameter (estimated based on the maximum gold-gold distance) between simulations and published experiments. Increasing the core diameter results in a monotonic increase in the number of adsorbed ligands for all three morphologies. The number of adsorbed ligands for the spherical SGC and HGC models is greater than the number of adsorbed ligands for the FGC model for all diameters and converges for diameters greater than ~5 nm. The results for the SGC and HGC models agree well with data by Hostetler *et al.*,[65] who approximated the number of adsorbed ligands by dividing the surface areas of ideal truncated octahedra by the surface area footprint of an alkanethiol ligand. The authors used this data to predict experimental thermogravimetric analysis measurements and found reasonable agreement, although the calculations overestimated experimental measurements. The similarity between the results for the SGC and HGC models and the Hostetler data suggests that these spherical models represent an upper bound on the number of adsorbed ligands for a given GNP core diameter. Since the SGC and HGC models have a very similar number of adsorbed ligands, we

found that their surface properties are nearly indistinguishable. Therefore, we focus the subsequent sections solely on differences between the SGC and FGC models. Results for the HGC model can be found in the SI.[2]

Jiang *et al.*[66] used matrix-assisted laser desorption/ionization mass spectroscopy (MALDI-MS) to directly determine the number of adsorbed ligands for 2, 4, and 6 nm GNPs (shown as purple squares in Figure 8.3). The MALDI-MS results agree well with the number of ligands adsorbed for the FGC model, which is significantly lower than the values for the spherical HGC and SGC models. Treating the core as uniformly spherical with a grafting density of 4.62 ligand/nm^2 - the grafting density of a planar gold (111) surface[67,68] - also yields a similar number of adsorbed ligands (black line in Figure 8.3). However, this approach ignores spatial variations in ligand grafting density; ligands adsorb with a higher density at the facets, where there are more favorable gold-sulfur interactions (SI, Figure S7).[2] Since the experimental MALDI-MS results report directly on the number of adsorbed ligands without assumptions regarding core geometry, the agreement between the MALDI-MS results and the FGC model simulations suggests that this morphology is representative of the highly faceted shapes observed experimentally.[41–44]

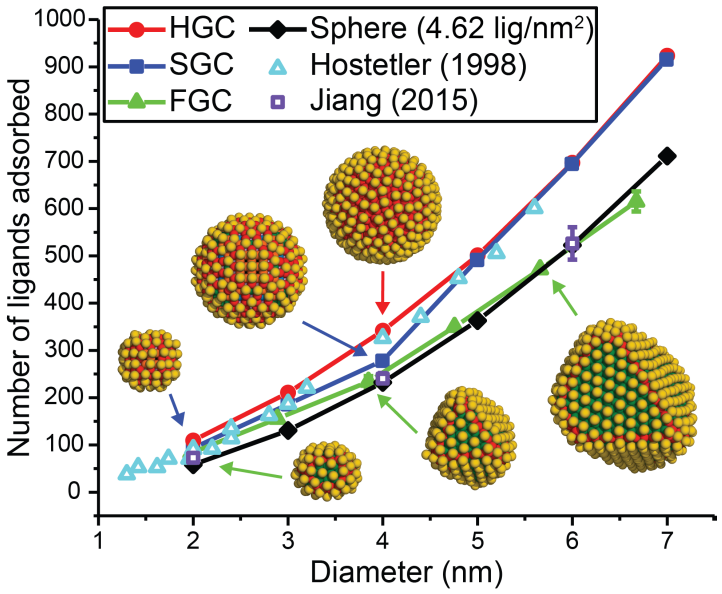


Figure 8.3: Number of adsorbed ligands for different gold core morphologies and sizes. Teal triangles are model data verified by thermogravimetric analysis from Ref. 65. Purple squares indicate data from MALDI-MS experiments from Ref. 66. The black line denotes the number of adsorbed ligands on a sphere with a planar gold (111) average surface density of 4.62 ligand/nm^2. Snapshots show gold cores colored with the same scheme as Figure 8.2 and the sulfur atoms of adsorbed ligands in yellow. The color of each arrow corresponds to the gold core shape (*i.e.* HGC, SGC, and FGC models are red, blue, and green, respectively).

8.3.2 Effect of core morphology on SAM structure

We next investigated the effect of core morphology on the structure of nonpolar alkanethiol SAMs in aqueous solution as the core size and ligand chain length (*i.e.*, the number of methylene groups, n, in the ligand backbone) were varied. Increasing either the gold core size or the ligand chain

length leads to a gradual transition from a SAM containing an approximately uniform, spherical distribution of disordered ligands to a SAM containing anisotropic, aspherical "bundles" of quasi-crystalline ligands that are oriented in the same direction, as illustrated in Figure 8.4A. This transition was quantified by measuring ligand order and SAM eccentricity.

As a measure of ligand order, we calculated the fraction of *trans* dihedral angles, which is defined as the number of *trans* dihedral angles ($120° < \phi < 240°$) normalized by the total number of dihedral angles in the ligand chain. We only calculated the dihedral angles of the heavy atoms, ignoring all hydrogen atoms. A fraction of *trans* dihedral angles value near one indicates a highly ordered ligand structure and a value near zero indicates a highly disordered ligand structure. Figure 8.4B shows the fraction of *trans* dihedral angles for the SGC and FGC models. For a given core diameter, increasing n increases ligand order, as is also observed for planar SAMs (SI, Figure S14).[2] For a given n, increasing the core diameter also increases ligand order due to the decreased free volume accessible to each ligand. The combination of these two effects leads to ordered quasi-crystalline configurations for large GNPs with long ligands. For $n < 11$, ligands are more ordered for the FGC model than the SGC model due to the reduced free volume for neighboring ligands on the faceted surfaces, although this effect is diminished for $n \geqslant 11$.

To quantify SAM sphericity, we calculated the eccentricity, e, shown in Equation 8.1, where I_{min} and I_{avg} are the smallest moment of inertia and

average moment of inertia along the principal axes of the ligands (excluding the gold core and computed with the Gromacs tool gmx principal), respectively.[69,70]

$$e = 1 - \left\langle \frac{I_{min}}{I_{max}} \right\rangle \tag{8.1}$$

Eccentricity values range from zero, indicating a spherical SAM, to one, indicating a highly aspherical (*e.g.* ellipsoidal) SAM. Figure 8.4C shows the eccentricity for the SGC and FGC models. The dominant feature is the increase in the eccentricity with increasing n due to the formation of oppositely oriented ligand bundles (snapshots shown in Figure 8.4C). Eccentricity also increases as core diameter decreases for $n \geqslant 11$. On average, the SGC model leads to a smaller eccentricity compared to the FGC model for $n \leqslant 11$. For $7 < n < 11$, there is a transition where the SAM begins to form highly asymmetric strutures that is highly dependent on the gold core diameter. At the extremum ligand length ($n = 16$), the SGC and FGC models produce similar eccentricity values, indicating that long ligands form asymmetric SAMs regardless of the core morphology.

This analysis of SAM structure indicates that increasing n increases ligand order due to the formation of quasi-crystalline ligand bundles which then increases SAM eccentricity. This behavior is expected given the formation of quasi-crystalline ligand arrangements on planar SAMs for long ligand lengths. The large amount of free volume accessible to ligands

grafted to small cores further permits bundling into highly asymmetric configurations that further increases eccentricity. The core morphology primarily affects GNPs with shorter ligand lengths; the FGC model promotes increased ligand order and eccentricity compared to the SGC model for $n \leqslant 11$, suggesting that ligand bundling is promoted as a consequence of the facets present in the FGC model.

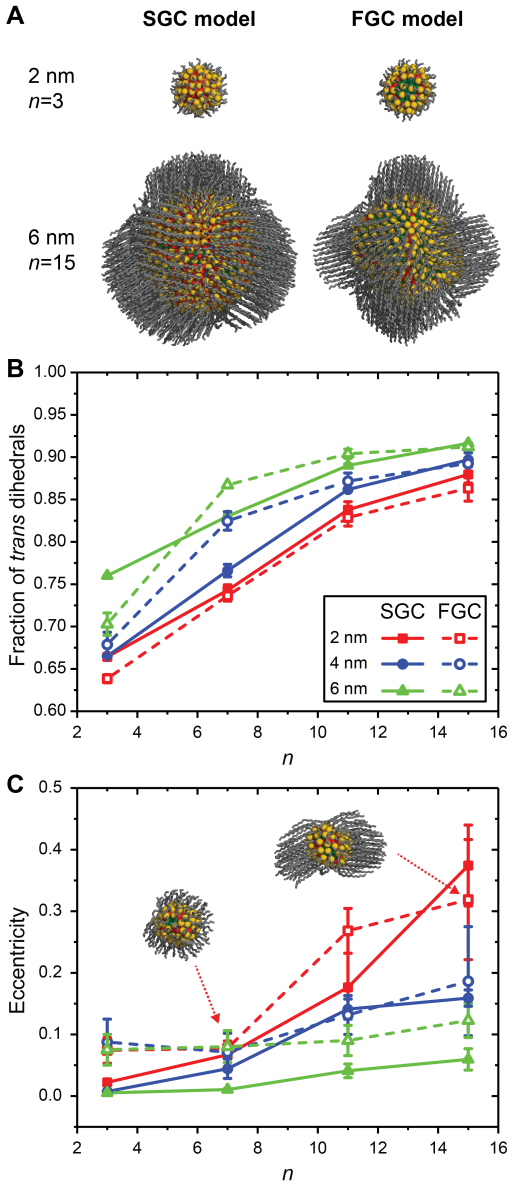


Figure 8.4: **A** Simulation snapshots of 2 nm and 6 nm SGC and FGC GNPs with butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) ligands. Sulfur atoms are represented by yellow beads and carbon chains are represented by gray lines. Hydrogen atoms and surrounding water molecules are not drawn. Gold cores are colored according to coordination number as described in Figure 8.2. **B** Fraction of *trans* dihedral angles as a function of ligand chain length (n) for the SGC (solid lines) and FGC (dashed lines) models. **C** Eccentricity as a function of ligand chain length for the SGC and FGC models. The legend is the same for both plots. Simulation snapshots are shown for the 2 nm FGC model with octanethiol ($n = 7$) and hexadecanethiol ($n = 15$).

8.3.3 Identification of ligand bundles

Eccentricity is an ensemble-average parameter that does not directly identify ligand bundles, which are the most prominent structural features observed for the alkanethiol SAMs. To complement the prior structural observables, we used the HBDSCAN clustering algorithm[50,51] to identify specific sets of ligands that form bundles, enabling the analysis of ligand subpopulations to quantify SAM structural properties. Ligands are assigned to the same bundle based on their relative orientations and end group distances; ligands that have small deviations in angle (*i.e.*, point in the same direction) and have small end group distances are considered in the same bundle. Some ligands are not assigned to bundles and are labeled as "non-bundled." Complete details on the algorithm and its implementation are provided in the SI.[2] To illustrate bundles identified using this approach, Figure 8.5A shows simulation snapshots of 2 nm and 6 nm SGC and FGC GNPs with butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) ligands. Ligands with the same color are in the same bundle, and each bundle is assigned a unique color. Non-bundled ligands are colored as gray and are found at the perimeter of bundles where they can access the surrounding free volume.

Figure 8.5B shows the number of bundles for the SGC and FGC models as a function of n and core diameter. Increasing n results in a decrease in the number of bundles and a corresponding increase in the fraction of ligands that are in bundles (Figure 8.5C). For long ligand chains, a small

number of bundles encompass a large fraction of all ligands. A lower number of bundles correlates with a higher eccentricity due to the formation of highly asymmetric, oppositely oriented bundles as shown in Figure 8.4C. Conversely, for short ligand chains, a much smaller fraction of ligands is in bundles and the number of bundles is much larger, both results consistent with disordered, uniformly distributed ligand end groups. Increasing the core diameter increases the number of bundles without changing the fraction of ligands that are in bundles. The FGC model in general has fewer bundles compared to the SGC model, and at smaller ligand lengths ($n \leqslant 7$), the fraction of ligands in bundles is higher for the FGC model than the SGC models, which may be attributed to the promotion of bundling by the facets. Together, these results suggest that the tendency for ligands to form bundles is determined primarily by ligand length, but the distribution and number of bundles depends on the core size and morphology for $n \leqslant 7$ while the influence of the gold core is diminished for $n \geqslant 11$.

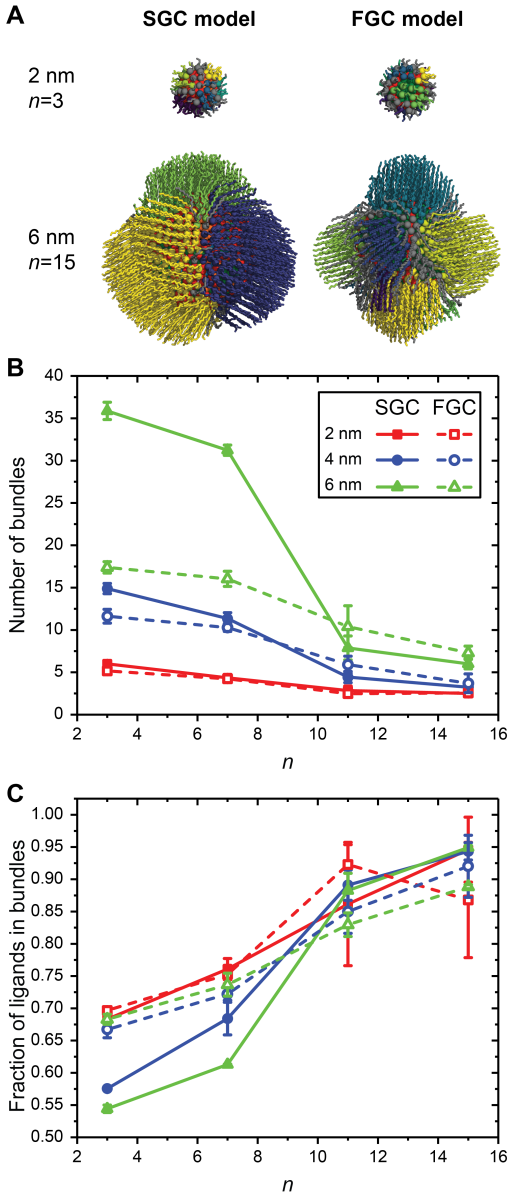


Figure 8.5: **A** Simulation snapshots of 2 nm and 6 nm SGC and FGC GNPs with butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) ligands. Ligands with the same color are in the same bundle, whereas ligands in gray are not in bundles. **B** Number of bundles as a function of ligand chain length (n) for the SGC (solid lines) and FGC (dashed lines) models. **C** Fraction of ligands in bundles as a function of ligand chain length for the SGC and FGC models. The legend is the same for both plots.

8.3.4 Relationship between gold core faceting and ligand bundling

Figure 8.5 indicates that the FGC model favors the formation of a small number of ligand bundles containing a larger fraction of all ligands relative to the SGC model, particularly for shorter ligand lengths. We next investigate the relationship between faceting and ligand bundling to determine if facets locally promote bundle formation. We define gold atoms with coordination numbers greater than seven as facet atoms (*e.g.* green/orange atoms in Figure 8.2B) and the remaining gold atoms as edge atoms (red atoms in Figure 8.2B). Figure 8.6A shows simulation snapshots of 6 nm FGC GNPs with butanethiol and hexadecanethiol ligands. The top snapshots show the sulfur head groups of bundled and non-bundled ligands, colored in black and grey respectively. The bottom snapshots show the sulfur head groups colored in magenta if the nearest gold atom is an edge atom and in cyan if the nearest gold atom is a facet atom. Comparing these two representations illustrates that non-bundled ligands are clustered near edges for the butanethiol ligands, while there is no clear correlation between non-bundled ligands and either edges or facets for the longer hexadecanethiol ligands. This comparison suggests that facets promote ligand bundling for shorter ligand chain lengths.

Figure 8.6B shows the fraction of ligands for which the nearest gold atom is a facet atom, f_{facet}, for the FGC model, distinguishing bundled

(solid lines) and non-bundled (dashed lines) ligands. For short ligands ($n \leqslant 7$) and large core diameters (4 and 6 nm), a much larger fraction of bundled ligands are coordinated by facet gold atoms than non-bunded ligands, indicating that bundles preferentially form on facets while non-bundled ligands reside at edges. For longer ligands ($n \geqslant 11$), both bundled and non-bundled ligands exhibit a similar preference for facets, indicating that the gold morphology no longer influences which ligands are bundled together. There is also no clear preference of bundled or non-bundled ligands for facets for the 2 nm GNP; note that the total fraction of facet gold atoms increases with diameter (SI, Figure S10).[2] Overall, this data supports the argument that facets promote bundle formation for short ligands on large GNPs.

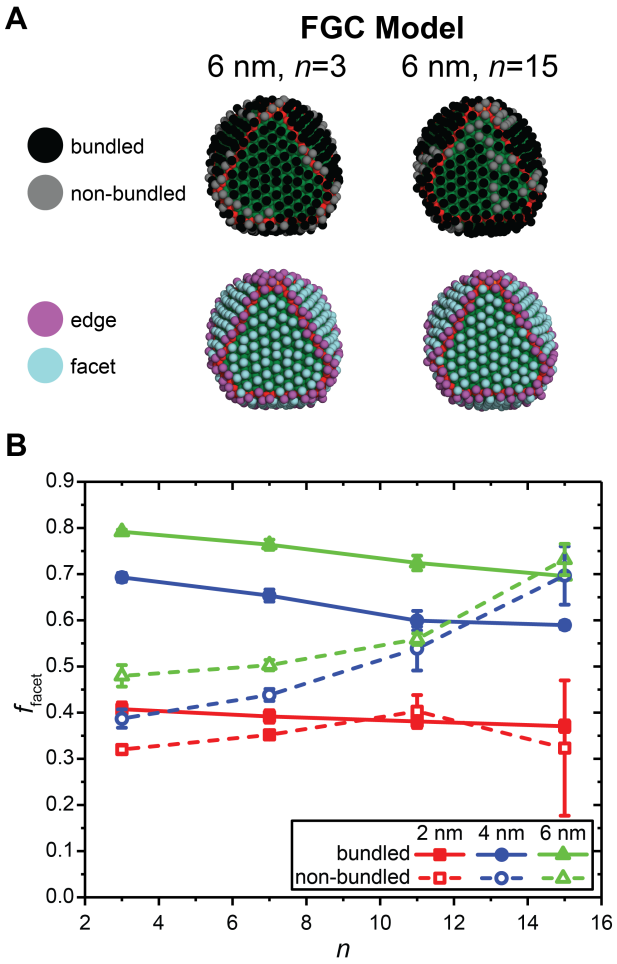


Figure 8.6: **A** Simulation snapshots of 6 nm butanethiol ($n = 3$) and hexadecanethiol ($n = 15$) GNPs for the FGC model. Only the ligand sulfur atoms are shown. At top, each sulfur atom is colored according to if the ligand is in a bundle (black) or not (gray). At bottom, each sulfur atom is colored according to if the nearest gold atom is at an edge (magenta) or facet (cyan). Gold atoms are colored according to coordination number as in Figure 8.2B. **B** The fraction of ligands on facets for bundled (filled lines) and non-bundled (dashed lines) ligands as a function of ligand chain length.

8.3.5 Effect of bundling on ligand properties

The analysis of the preceding sections indicates that the core morphology affects ligand bundling, leading to a heterogeneous population of bundled and non-bundled ligands that are expected to have distinct properties. Using the clustering algorithm, we can distinguish the average properties of these two populations. Figure 8.7 shows the fraction of *trans* dihedral angles for bundled and non-bundled ligands for the FGC model. The observed trends are similar to those in Figure 8.4A, but bundled ligands are more ordered on average than non-bundled ligands for all core diameters. This effect is amplified for long ligand lengths ($n \geqslant 11$), where the ligands form fewer, but more dense bundles (Figure 8.5). These results indicate that non-bundled ligands are disordered due to their alignment with high free volume edges (Figure 8.6), which disrupts ligand order.

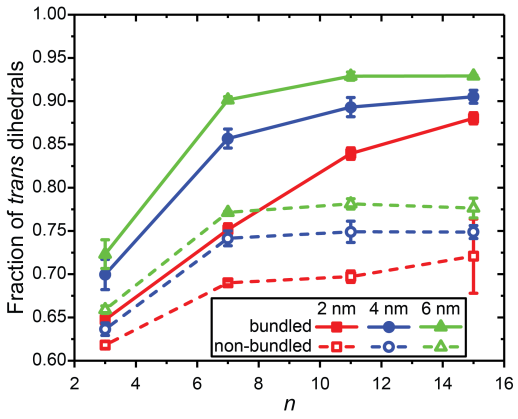


Figure 8.7: Fraction of *trans* dihedrals for bundled (solid lines) and non-bundled (dashed lines) ligands as a function of ligand chain length (n) for the FGC model.

We further quantify the differences between bundled and non-bundled ligands by calculating the solvent-accessible surface area (SASA), which describes the surface area available for solvent molecules to contact the ligands. Figure 8.8A shows the SASA per ligand for bundled and non-bundled ligands for the FGC model. Increasing the ligand chain length increases the SASA while increasing the core diameter decreases the SASA due to the reduced free volume accessible to each ligand. Bundled ligands on average have a lower SASA than non-bundled ligands for all ligand lengths and core sizes. To further elucidate these differences, Figure 8.8B shows a probability density function of the SASA for a 2 nm butanethiol GNP (top) and a 6 nm hexadecanethiol GNP (bottom). Short alkanethiol chains ($n = 3$) have small, transient bundles due to low degrees of ligand order, leading to a SASA distribution that is almost indistinguishable between bundled and non-bundled ligands. Conversely, there is a clear bimodal distribution in the SASA between bundled and non-bundled ligands for long ligand chains, with non-bundled ligands exhibiting a broader distribution of SASA values indicating highly variable solvent accessibility. These findings are illustrated by the simulation snapshots that are (i) colored by bundled (black) versus non-bundled (grey) ligands and (ii) colored by the SASA of each ligand. This analysis thus indicates that GNPs with ligand bundles dynamically display spatially heterogeneous surface properties.

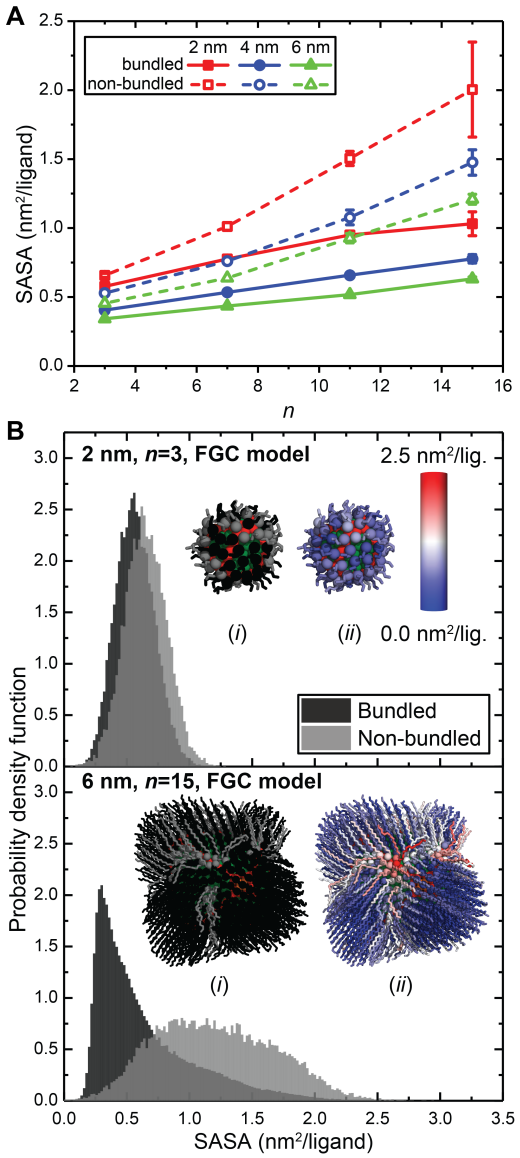


Figure 8.8: **A** SASA for bundled (solid lines) and non-bundled (dashed lines) ligands as a function of ligand chain length (n) for the FGC model. **B** Probability density function of the SASA for a 2 nm FGC GNP with butanethiol ligands (top) and a 6 nm FGC GNP with hexadecanethiol ligands (bottom). The probability density function is shown only for one of the three trials. Simulation snapshots are colored by (*i*) bundled (black) and non-bundled (grey) ligands, and (*ii*) ligand SASA.

8.4 Discussion

Using a generalized workflow for GNP system preparation, we modeled a series of single-component alkanethiol SAM-protected GNPs with varying core diameters and core morphologies. By allowing ligands to self-assemble onto the GNP surface, rather than specifying a ligand surface density a priori, we found that the number of ligands adsorbed to the highly faceted FGC models agrees well with experimental estimates. Moreover, the choice of gold core morphology affects SAM structure; facets promote the formation of asymmetric structures even for single-component SAMs. SAM asymmetry is amplified for small gold core diameters (~2 nm) and long alkanethiol chain lengths ($n \geqslant 11$). In these cases, ligands bundle together and orient in the same direction to form quasi-crystalline domains, a phenomenon reported extensively in the literature.[14,26–32,48,49,71] The segregation of SAM ligands into bundles and non-bundles introduces spatially heterogeneous surface properties due to the excess free volume accessible to non-bundled ligands.

Furthermore, our findings suggest that gold core morphology is important to the formation of ligand bundles for gold core diameters > 2 nm and short ligand lengths ($n < 11$). We used a clustering algorithm to identify bundled ligands and found that for long ligands ($n \geqslant 11$), the SAM consists of a small number of bundles that contain nearly all ligands, independent of the core morphology. However, core facets promote bundle formation

for short ligands ($n < 11$) and for 4 and 6 nm core diameters; the influence of facets on bundling for a 2 nm core diameter is minimized due to the small size of each facet. This behavior can be interpreted in terms of the order-disorder transition between disordered and quasi-crystalline ligand conformations for planar SAMs, which occurs as n increases from 3 to 4 using the force field parameters in this work (SI, Figure S14).[2] On planar surfaces, sufficiently long ligands form quasi-crystalline arrangements to maximize favorable interchain interactions which overcome the entropic penalty associated with increased ligand order.[72] Ligand bundling reflects a similar order-disorder transition, but ligand packing is inhibited by the curvature of the core. As a result, planar facets facilitate chain packing and promote bundle formation for shorter ligand lengths, while edges act as disclinations that inhibit chain packing due to the excess free volume. Bundles form regardless of the core morphology and size if the ligand chain length is sufficiently long.

Our results suggest that the formation of ligand bundles can be controlled by selecting the gold core size and alkanethiol ligand length, or by controlling the faceting of the gold core. Promoting bundle formation can introduce spatially heterogeneous surface properties and influence overall GNP behavior. For example, hexadecanethiol bundles have been found to affect the colloidal stability of small GNPs (~4 nm) by allowing SAMs to interdigitate and pack into energetically stable configurations.[31] In another example, the formation of ligand bundles was shown to bend and

compact DNA.[19] Spatial heterogeneities in ligand arrangements, similar to the bundles identified here, can also lead to the formation of protein like binding pockets in the SAM.[33] We have limited our studies to homogenous alkanethiols on GNPs, but future studies will focus on the inclusion of complex ligands or mixed SAMs which can further tune surface properties to tailor interactions in biological systems.

Finally, from the standpoint of developing accurate GNP models, we summarize some recommendations below for future GNP simulations:

- The structural properties of long alkanethiol ligands ($n \geqslant 11$) are not significantly influenced by the gold core morphology. For these systems, the SGC model is recommended for its ease of generation and reproducibility. Despite the FGC model being a more accurate representation of experimental systems, generating these morphologies requires a separate simulation.

- Short ligand lengths ($n < 11$) are heavily influenced by the gold core morphology. Our results suggest that the FGC model, or potentially other faceted gold core morphologies, would more accurately represent the structural properties of these GNPs.

- The SGC and HGC models produce similar results. However, since the HGC model has a vacuum core, it imposes technical issues related to the need for strong gold-gold bonds, weaker van der Waals interactions with surrounding molecules due to missing core atoms, and

the unphysical introduction of molecules within the core molecule (*e.g.*, during system preparation). Previous authors have overcome these issues by fixing the positions of the gold and sulfur atoms and adding alternative potentials to account for interactions between the gold core, ligands, and solvent.[31] Nonetheless, we recommend the SGC model over the HGC model to avoid having to implement correction for these issues given the similar results produced by both models.

- Ligand self-assembly leads to higher surface densities for all gold core morphologies than the surface density of a planar (111) surface. Since experimental results also suggest that higher grafting densities are expected on curved surfaces, we recommend permitting ligand self-assembly rather than pre-defining surface densities based on planar estimates.

8.5 Summary

In this chapter, we developed a workflow to generate GNP models for any arbitrary size and ligand selection, enabling the systematic evaluation of GNP properties using MD simulations. Using this workflow, we first investigated the effect of three different gold core morphologies - a hollow gold core (HGC) model, a spherical gold core (SGC) model, and a faceted gold core (FGC) model - on alkanethiol SAM grafting densities. We found

that the SGC and HGC models produce similar results that overestimate the experimental number of grafted ligands, while the FGC model leads to grafting ligands in good agreement with recent experimental measurements. We next investigated the effect of gold core size, ligand chain length, and morphology on SAM structural properties and determined that the dominant parameters are the alkanethiol ligand chain length and gold core diameter; in general, longer chain lengths increase ligand order, while larger gold cores lead to more spherical SAM structures. SAM asymmetry was promoted for long ligand chains on smaller gold cores and attributed to the formation of discrete, quasi-crystalline ligand bundles.

To further investigate SAM asymmetry and domain properties, we identified the specific subset of ligands that are in bundles using the HDBSCAN clustering algorithm, which classified ligands that were oriented in the same direction with nearby end groups. This approach demonstrated that longer alkanethiols tend to have fewer, densely packed bundles, while bundle formation for short ligand lengths (fewer than 11 methylene groups in the ligand chain) is highly influenced by the gold morphology. Specifically, we found that bundles are promoted on planar facets due to favorable packing that promotes ligand order. Finally, we found that ligands in bundles have increased order and a smaller surface area accessible to solvent compared to non-bundled ligands. Therefore, the formation of bundles can introduce heterogeneous surface properties, even though the monolayer is in principal homogenous. Future work will explore the use of

more complex ligands with the workflow developed here to begin tuning GNPs for biomedical applications.

8.6 References

[1] Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.

[2] Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles [Supporting Information]. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.

[3] Bergen, J. M.; Von Recum, H. A.; Goodman, T. T.; Massey, A. P.; Pun, S. H. Gold nanoparticles as a versatile platform for optimizing physicochemical parameters for targeted drug delivery. *Macromolecular Bioscience* **2006**, *6*, 506–516.

[4] Kinnear, C.; Moore, T. L.; Rodriguez-Lorenzo, L.; Rothen-Rutishauser, B.; Petri-Fink, A. Form follows function: nanoparticle shape and its implications for nanomedicine. *Chemical reviews* **2017**, *117*, 11476–11521.

[5] Toy, R.; Bauer, L.; Hoimes, C.; Ghaghada, K. B.; Karathanasis, E. Targeted nanotechnology for cancer imaging. *Advanced drug delivery reviews* **2014**, *76*, 79–97.

[6] Pensa, E.; Cortés, E.; Corthey, G.; Carro, P.; Vericat, C.; Fonticelli, M. H.; Benítez, G.; Rubert, A. A.; Salvarezza, R. C. The chemistry of the sulfur–gold interface: in search of a unified model. *Accounts of chemical research* **2012**, *45*, 1183–1192.

[7] Häkkinen, H. The gold–sulfur interface at the nanoscale. *Nature chemistry* **2012**, *4*, 443.

[8] Saha, K.; Agasti, S. S.; Kim, C.; Li, X.; Rotello, V. M. Gold nanoparticles in chemical and biological sensing. *Chemical reviews* **2012**, *112*, 2739–2779.

[9] Alkilany, A. M.; Lohse, S. E.; Murphy, C. J. The gold standard: gold nanoparticle libraries to understand the nano–bio interface. *Accounts of chemical research* **2013**, *46*, 650–661.

[10] Pengo, P.; Şologan, M.; Pasquato, L.; Guida, F.; Pacor, S.; Tossi, A.; Stellacci, F.; Marson, D.; Boccardo, S.; Pricl, S.; et al.. Gold nanoparticles with patterned surface monolayers for nanomedicine: current perspectives. *European Biophysics Journal* **2017**, *46*, 749–771.

[11] Nicolardi, S.; Van Der Burgt, Y. E.; Codée, J. D.; Wuhrer, M.; Hokke, C. H.; Chiodo, F. Structural characterization of biofunctionalized gold nanoparticles by ultrahigh-resolution mass spectrometry. *ACS nano* **2017**, *11*, 8257–8264.

[12] Zhou, H.; Li, X.; Lemoff, A.; Zhang, B.; Yan, B. Structural confirmation and quantification of individual ligands from the surface of multifunctionalized gold nanoparticles. *Analyst* **2010**, *135*, 1210–1213.

[13] Charchar, P.; Christofferson, A. J.; Todorova, N.; Yarovsky, I. Understanding and designing the gold–bio interface: Insights from simulations. *Small* **2016**, *12*, 2395–2418.

[14] Bolintineanu, D. S.; Lane, J. M. D.; Grest, G. S. Effects of functional groups and ionization on the structure of alkanethiol-coated gold nanoparticles. *Langmuir* **2014**, *30*, 11075–11085.

[15] Cui, Q.; Hernandez, R.; Mason, S. E.; Frauenheim, T.; Pedersen, J. A.; Geiger, F. Sustainable nanotechnology: opportunities and challenges for theoretical/computational studies. *The Journal of Physical Chemistry B* **2016**, *120*, 7297–7306.

[16] Ghorai, P. K.; Glotzer, S. C. Atomistic simulation study of striped phase separation in mixed-ligand self-assembled monolayer coated nanoparticles. *The Journal of Physical Chemistry C* **2010**, *114*, 19182–19187.

[17] Heikkilä, E.; Martinez-Seara, H.; Gurtovenko, A. A.; Vattulainen, I.; Akola, J. Atomistic simulations of anionic Au144 (SR) 60 nanoparticles interacting with asymmetric model lipid membranes. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2014**, *1838*, 2852–2860.

[18] Hong, J.; Hamers, R. J.; Pedersen, J. A.; Cui, Q. A hybrid molecular dynamics/multiconformer continuum electrostatics (MD/MCCE) approach for the determination of surface charge of nanomaterials. *The Journal of Physical Chemistry C* **2017**, *121*, 3584–3596.

[19] Nash, J. A.; Singh, A.; Li, N. K.; Yingling, Y. G. Characterization of nucleic acid compaction with histone-mimic nanoparticles through all-atom molecular dynamics. *ACS nano* **2015**, *9*, 12374–12382.

[20] Olmos-Asar, J. A.; Ludueña, M.; Mariscal, M. Monolayer protected gold nanoparticles: The effect of the headgroup–au interaction. *Physical Chemistry Chemical Physics* **2014**, *16*, 15979–15987.

[21] Peters, B. L.; Lane, J. M. D.; Ismail, A. E.; Grest, G. S. Fully atomistic simulations of the response of silica nanoparticle coatings to alkane solvents. *Langmuir* **2012**, *28*, 17443–17449.

[22] Salassi, S.; Simonelli, F.; Bochicchio, D.; Ferrando, R.; Rossi, G. Au nanoparticles in lipid bilayers: a comparison between atomistic and coarse-grained models. *The Journal of Physical Chemistry C* **2017**, *121*, 10927–10935.

[23] Sen, S.; Han, Y.; Rehak, P.; Vuković, L.; Král, P. Computational studies of micellar and nanoparticle nanomedicines. *Chemical Society Reviews* **2018**, *47*, 3849–3860.

[24] Van Lehn, R. C.; Alexander-Katz, A. Structure of mixed-monolayer-protected nanoparticles in aqueous salt solution from atomistic molecular dynamics simulations. *The Journal of Physical Chemistry C* **2013**, *117*, 20104–20115.

[25] 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*, 3165–3175.

[26] Lane, J. M. D.; Grest, G. S. Spontaneous asymmetry of coated spherical nanoparticles in solution and at liquid-vapor interfaces. *Physical review letters* **2010**, *104*, 235501.

[27] Ghorai, P. K.; Glotzer, S. C. Molecular dynamics simulation study of self-assembled monolayers of alkanethiol surfactants on spherical gold nanoparticles. *The Journal of Physical Chemistry C* **2007**, *111*, 15857–15862.

[28] Luedtke, W.; Landman, U. Structure, dynamics, and thermodynamics of passivated gold nanocrystallites and their assemblies. *The Journal of Physical Chemistry* **1996**, *100*, 13323–13329.

[29] Luedtke, W.; Landman, U. Structure and thermodynamics of self-assembled monolayers on gold nanocrystallites. *The Journal of Physical Chemistry B* **1998**, *102*, 6566–6572.

[30] Matthew D áLane, J.; et al.. Assembly of responsive-shape coated nanoparticles at water surfaces. *Nanoscale* **2014**, *6*, 5132–5137.

[31] Kister, T.; Monego, D.; Mulvaney, P.; Widmer-Cooper, A.; Kraus, T. Colloidal stability of apolar nanoparticles: the role of particle size and ligand shell structure. *ACS nano* **2018**, *12*, 5969–5977.

[32] Monego, D.; Kister, T.; Kirkwood, N.; Mulvaney, P.; Widmer-Cooper, A.; Kraus, T. Colloidal stability of apolar nanoparticles: Role of ligand length. *Langmuir* **2018**, *34*, 12982–12989.

[33] Riccardi, L.; Gabrielli, L.; Sun, X.; De Biasi, F.; Rastrelli, F.; Mancin, F.; De Vivo, M. Nanoparticle-based receptors mimic protein-ligand recognition. *Chem* **2017**, *3*, 92–109.

[34] Ding, H.-m.; Tian, W.-d.; Ma, Y.-q. Designing nanoparticle translocation through membranes by computer simulations. *ACS nano* **2012**, *6*, 1230–1238.

[35] Van Lehn, R. C.; Alexander-Katz, A. Membrane-embedded nanoparticles induce lipid rearrangements similar to those exhibited by biological membrane proteins. *The Journal of Physical Chemistry B* **2014**, *118*, 12586–12598.

[36] Gkeka, P.; Angelikopoulos, P.; Sarkisov, L.; Cournia, Z. Membrane partitioning of anionic, ligand-coated nanoparticles is accompanied by ligand snorkeling, local disordering, and cholesterol depletion. *PLoS Comput Biol* **2014**, *10*, e1003917.

[37] Van Lehn, R. C.; Alexander-Katz, A. Fusion of ligand-coated nanoparticles with lipid bilayers: Effect of ligand flexibility. *The Journal of Physical Chemistry A* **2014**, *118*, 5848–5856.

[38] Jiménez, A.; Sarsa, A.; Blázquez, M.; Pineda, T. A molecular dynamics study of the surfactant surface density of alkanethiol self-assembled monolayers on gold nanoparticles as a function of the radius. *The Journal of Physical Chemistry C* **2010**, *114*, 21309–21314.

[39] Copie, G.; Biaye, M.; Diesinger, H.; Melin, T.; Krzeminski, C.; Cleri, F. Deformation Localization in Molecular Layers Constrained between Self-Assembled Au Nanoparticles. *Langmuir* **2017**, *33*, 2677–2687.

[40] Barnard, A. S. Direct comparison of kinetic and thermodynamic influences on gold nanomorphology. *Accounts of chemical research* **2012**, *45*, 1688–1697.

[41] Buffat, P.-A.; Flüeli, M.; Spycher, R.; Stadelmann, P.; Borel, J.-P. Crystallographic structure of small gold particles studied by high-resolution electron microscopy. *Faraday Discussions* **1991**, *92*, 173–187.

[42] Elechiguerra, J. L.; Reyes-Gasga, J.; Yacaman, M. J. The role of twinning in shape evolution of anisotropic noble metal nanostructures. *Journal of Materials Chemistry* **2006**, *16*, 3906–3919.

[43] Marks, L. Experimental studies of small particle structures. *Reports on Progress in Physics* **1994**, *57*, 603.

[44] Whetten, R. L.; Khoury, J. T.; Alvarez, M. M.; Murthy, S.; Vezmar, I.; Wang, Z.; Stephens, P. W.; Cleveland, C. L.; Luedtke, W.; Landman, U. Nanocrystal gold molecules. *Advanced materials* **1996**, *8*, 428–433.

[45] Djebaili, T.; Richardi, J.; Abel, S.; Marchi, M. Atomistic simulations of the surface coverage of large gold nanocrystals. *The Journal of Physical Chemistry C* **2013**, *117*, 17791–17800.

[46] Pool, R.; Schapotschnikow, P.; Vlugt, T. J. Solvent effects in the adsorption of alkyl thiols on gold structures: A molecular simulation study. *The Journal of Physical Chemistry C* **2007**, *111*, 10201–10212.

[47] Hill, H. D.; Millstone, J. E.; Banholzer, M. J.; Mirkin, C. A. The role radius of curvature plays in thiolated oligonucleotide loading on gold nanoparticles. *ACS nano* **2009**, *3*, 418–424.

[48] Bozorgui, B.; Meng, D.; Kumar, S. K.; Chakravarty, C.; Cacciuto, A. Fluctuation-driven anisotropic assembly in nanoscale systems. *Nano letters* **2013**, *13*, 2732–2737.

[49] Salerno, K. M.; Ismail, A. E.; Lane, J. M. D.; Grest, G. S. Coating thickness and coverage effects on the forces between silica nanoparticles in water. *The Journal of chemical physics* **2014**, *140*, 194904.

[50] Campello, R. J.; Moulavi, D.; Sander, J. In *Pacific-Asia conference on knowledge discovery and data mining*; Springer; pp 160–172.

[51] McInnes, L.; Healy, J. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*; IEEE; pp 33–42.

[52] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[53] Rahm, J. M.; Erhart, P. Beyond magic numbers: atomic scale equilibrium nanoparticle shapes for any size. *Nano letters* **2017**, *17*, 5775–5781.

[54] Grochola, G.; Russo, S. P.; Snook, I. K. On fitting a gold embedded atom method potential using the force matching method. *The Journal of chemical physics* **2005**, *123*, 204719.

[55] Djebaili, T.; Abel, S.; Marchi, M.; Richardi, J. Influence of Force-Field Parameters on the Atomistic Simulations of Metallic Surfaces and Nanoparticles. *The Journal of Physical Chemistry C* **2017**, *121*, 27758–27765.

[56] Voznyy, O.; Dubowski, J. J.; Yates Jr, J. T.; Maksymovych, P. The role of gold adatoms and stereochemistry in self-assembly of methylthiolate on Au (111). *Journal of the American Chemical Society* **2009**, *131*, 12989–12993.

[57] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

[58] Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *The Journal of chemical physics* **1996**, *105*, 1902–1921.

[59] Heinz, H.; Vaia, R.; Farmer, B.; Naik, R. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12- 6 and 9- 6 Lennard-Jones potentials. *The Journal of Physical Chemistry C* **2008**, *112*, 17281–17290.

[60] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[61] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[62] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[63] Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force field benchmark of organic liquids: density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *Journal of chemical theory and computation* **2012**, *8*, 61–74.

[64] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[65] Hostetler, M. J.; Wingate, J. E.; Zhong, C.-J.; Harris, J. E.; Vachet, R. W.; Clark, M. R.; Londono, J. D.; Green, S. J.; Stokes, J. J.; Wignall, G. D.; et al.. Alkanethiolate gold cluster molecules with core diameters from 1.5 to 5.2 nm: core and monolayer properties as a function of core size. *Langmuir* **1998**, *14*, 17–30.

[66] Jiang, Y.; Huo, S.; Mizuhara, T.; Das, R.; Lee, Y.-W.; Hou, S.; Moyano, D. F.; Duncan, B.; Liang, X.-J.; Rotello, V. M. The interplay of size and surface functionality on the cellular uptake of sub-10 nm gold nanoparticles. *ACS nano* **2015**, *9*, 9986–9993.

[67] Ulman, A. Formation and structure of self-assembled monolayers. *Chemical reviews* **1996**, *96*, 1533–1554.

[68] Nuzzo, R. G.; Zegarski, B. R.; Dubois, L. H. Fundamental studies of the chemisorption of organosulfur compounds on gold (111). Implications for molecular self-assembly on gold surfaces. *Journal of the American Chemical Society* **1987**, *109*, 733–740.

[69] Lebecque, S.; Crowet, J.-M.; Nasir, M. N.; Deleu, M.; Lins, L. Molecular dynamics study of micelles properties according to their size. *Journal of Molecular Graphics and Modelling* **2017**, *72*, 6–15.

[70] Marchi, M.; Abel, S. Modeling the self-aggregation of small AOT reverse micelles from first-principles. *The journal of physical chemistry letters* **2015**, *6*, 170–174.

[71] Koch, A. H.; Lévêque, G.; Harms, S.; Jaskiewicz, K.; Bernhardt, M.; Henkel, A.; Sönnichsen, C.; Landfester, K.; Fytas, G. Surface asymmetry of coated spherical nanoparticles. *Nano letters* **2014**, *14*, 4138–4144.

[72] Chen, J.; Chang, B.; Oyola-Reynoso, S.; Wang, Z.; Thuo, M. Quantifying gauche defects and phase evolution in self-assembled monolayers through sessile drops. *ACS omega* **2017**, *2*, 2072–2084.

9 LIPOPHILICITY OF CATIONIC LIGANDS PROMOTES IRREVERSIBLE ADSORPTION OF NANOPARTICLES TO LIPID BILAYERS

Chapter 8 described a generalized workflow for generating GNP systems. We now extend this workflow to interrogate the effects of ligand end group chemistry on the adsorption of GNPs onto lipid bilayers. This chapter seeks to answer the following questions:

- How does the hydrophobicity (or similarly, lipophilicity) of the ligand end group affect GNP adsorption onto bilayers?
- How do we use enhanced sampling simulations (*e.g.* umbrella sampling from Section 2.1.3) to measure GNP-bilayer adsorption free energies?
- How can molecular simulations yield physical insight into GNP-bilayer adsorption behavior?

In this chapter, we use a library of cationic ligands coated on 2-nm gold nanoparticles to probe the impact of ligand end group lipophilicity on interactions with supported phosphatidylcholine lipid bilayers as a model for cytoplasmic membranes. Nanoparticle adsorption to and desorption

This chapter was reproduced from with permission from Lochbaum, C. A.; Chew, A. K.; Zhang, X.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A. Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers. *ACS Nano* **2021**, *15*, 6562–6572.[1] The supporting information is cited as Ref. 2. C. A. Lochbaum and A. K. Chew contributed equally to this work. C. A. Lochbaum, X. Zhang, V. Rotello, and J. A. Pedersen design, synthesized, and performed the experiments for this chapter.

from the model membranes were investigated by quartz crystal microbalance with dissipation monitoring. We find that nanoparticle adsorption to model membranes increases with ligand lipophilicity. The effects of ligand structure on gold nanoparticle attachment were further analyzed using atomistic molecular dynamics simulations, which showed that increasing ligand lipophilicity promotes ligand intercalation into the lipid bilayer. Together, the experimental and simulation results could be described by a two-state model that accounts for initial attachment and subsequent conversion to a quasi-irreversibly bound state. We find that only nanoparticles coated with the most lipophilic ligands in our nanoparticle library undergo conversion to the quasi-irreversible state. We propose that initial attachment is governed by interaction between the ligands and phospholipid tail groups, while conversion into the quasi-irreversibly bound state reflects ligand intercalation between phospholipid tail groups and eventual lipid extraction from the bilayer. Systematic variation of ligand lipophilicity enabled us to demonstrate that the lipophilicity of cationic ligands correlates with nanoparticle-bilayer adsorption and suggests that changing the nonpolar ligand R group promotes a mechanism of ligand intercalation into the bilayer associated with irreversible adsorption.

9.1 Introduction

Ligand-coated engineered nanomaterials (ENMs) have been used in bioimaging, drug delivery, and consumer goods, inspiring investigations into understanding how ENMs interact with biological interfaces.[3,4] In particular, establishing relationships between the properties of ENMs and their interactions with cellular membranes is essential for designing safe ENMs.[5,6] For example, interactions between ENMs and cellular membranes can result in lipid extraction[7,8] and membrane disruption,[9,10] events that can lead to cytotoxicity.[9,11,12] However, predicting such behaviors from ligand properties remains challenging, inhibiting ENM design. Toward this end, ligand lipophilicity - the property quantifying the partitioning of a ligand between aqueous media and lipid, which correlates strongly with hydrophobicity[13] - has been found to dictate ENM interactions with biological interfaces.[14–17] However, the specific interactions between lipophilic ENM ligands and cellular membranes-and the degree to which ligand lipophilicity drives these interactions-remain unclear.

Ligand-coated gold nanoparticles (AuNPs) have been used as model ENMs because of their ease of fabrication and tunable surface chemistry.[18] AuNPs can be synthesized at sizes commensurate with biomolecules (< 10 nm), enabling the study of interactions with biological interfaces at the same length scale.[19] To probe interactions between ENMs and cell membranes, supported lipid bilayers (SLBs) have been used as model

membrane systems.[20,21] For example, SLBs have been used to link nanoparticle size, core composition, and surface chemistry to increased cellular interaction, internalization, and cytotoxicity.[9,11] Quartz crystal microbalance with dissipation monitoring (QCM-D) has been used to monitor ENM behavior at the supported lipid bilayer-solution interface through measuring changes in mass and energy dissipation of material coupled to the motion of the sensor.[22] Changes in energy dissipation reflect the viscoelasticity of laterally homogeneous adlayers such as SLBs or to the stiffness of particle-surface contacts.[22,23] QCM-D has been used to deduce mechanisms of peptide incorporation into SLBs,[24] as well as the kinetics of ENM interactions with surfaces.[25,26]

While the kinetics of AuNP adsorption onto SLBs can be quantified via QCM-D, other approaches are needed to reveal the molecular-scale AuNP-bilayer interactions that lead to adsorption. As a result, computational models, such as classical molecular dynamics (MD) simulations, have been used to simulate AuNP-bilayer interactions at multiple length scales.[16,27–30] Previous work used atomistic MD simulations to study the free energy of AuNP insertion into the bilayer for AuNPs coated with varying ratios of neutral octanethiol and negatively charged 11-mercapto-1-undecanesulphonate (MUS).[31] Inclusion of more lipophilic ligands (*viz.* octanethiol) lowered the free energy of AuNP insertion into the lipid bilayer, suggesting that AuNP lipophilicity is critical to fusion with the lipid bilayer.[31] Subsequent studies found that AuNP insertion was driven pri-

marily by the lipid membrane core shielding lipophilic ligands from the aqueous environment.[31,32] These studies, in conjunction with related computational results,[16] suggest that the proportion and spatial distribution of lipophilic ligands on AuNP surfaces represent important parameters that modulate potential AuNP insertion into the lipid bilayer. However, the influence of ligand structure, specifically the degree of lipophilicity, on the interaction of AuNPs with lipid membranes remains to be determined.

In this chapter, we combine QCM-D experiments and atomistic MD simulations to systematically investigate the influence of ligand lipophilicity on AuNP interactions with single-component lipid bilayers. The adsorption of AuNPs to lipid bilayers and their long-term attachment stability increased with ligand lipophilicity. Simulations revealed a mechanism for the long-term stability observed for AuNPs with increased ligand lipophilicity: hydrophobic contacts between ligand lipophilic groups and the lipid tail groups drive the intercalation of more lipophilic ligands into the bilayer or lipid extraction from the bilayer. Kinetic analysis of the QCM-D data accounting for reversible and quasi-irreversible AuNP adsorption showed agreement with free energy calculations of AuNP-SLB systems. Our findings demonstrate that AuNPs with increased ligand lipophilicity have increased rates of conversion to a quasi-irreversibly bound state, providing design rules for developing ENMs with tailored membrane interactions.

9.2 Materials and Methods

9.2.1 Ligand Synthesis

Figure S1 depicts the synthesis pathway for the ligands used in AuNP functionalization.[2] Compound (1) was synthesized as described by Miranda *et al.*[33] Compound (1) (1.0 g, 2 mmol, 1 eq) was dissolved in 5 mL of ethyl acetate in a 20 mL vial. To each vial, we added 30 eq of the corresponding amine, and the entire solution was sealed properly and heated gently to 50°C for 3-5 days. Afterward, the solvent was evaporated and the residue was washed three times by hexanes, heptanes, or a 1:1 mixture of hexanes and heptanes to obtain (2) as a yellowish oil-like liquid.

We dissolved 200 mg of (2) in 3 mL of dichloromethane under a nitrogen atmosphere. To the solution we added 20 eq of trifluoroacetic acid followed by addition of 1.2 eq of triisopropylsilane. The solution was stirred at room temperature overnight. Afterward, the solvent was removed, and the residue was washed with hexane or heptane three times and ether for three more times to obtain ligands as colorless or yellowish liquid. To validate the ligand structure, ^{1}H NMR spectroscopy of ligands was performed. Spectra are shown in the Supporting Information, Figures S2-S6.[2]

9.2.2 Gold Nanoparticle Synthesis

Gold nanoparticles were synthesized by the Brust-Schiffrin two-phase method as described in previous literature.[34,35] In brief, 1 g of $HAuCl_4$ was dissolved in 300 mL 1:1 water toluene. We added 2.1 g of tetraoctylammonium bromide directly with maximum stirring speed. We added 0.7 mL of pentanethiol dropwise until the whole solution turned into white. Then 2.0 g of sodium borohydride was dissolved in around 8 mL of water and immediately added into the white solution. After stirring overnight, the organic layer was separated and dried under reduced pressure in room temperature. The residue was precipitated in cold ethanol and re-dissolved in hexanes. The solution washed with acetonitrile 120 times until all TOAB was fully removed to obtain gold core.

9.2.3 Ligand Exchange Reaction

Gold cores (40 mg) were dissolved in 4 mL dichloromethane under nitrogen atmosphere. To this solution we added 120 mg of the corresponding ligand in a mixed dichloromethane/methanol (2 mL/2 mL) solution in a dropwise manner under nitrogen and stirred the mixture for 72 h at room temperature. Solvents were removed under reduced pressure, and precipitations were washed with hexanes three times, and hexanes/dichloromethane mixture (1:1 v/v), or pure dichloromethane three times. The solid was suspended in ultrapure water, dialyzed for 3 days, and

concentrated by ultracentrifugation. The concentration of gold nanoparticles was determined based on the absorption at 506 nm as previously reported by Haiss *et al.*[36] Adsorption spectra and compiled absorbance at 506 for all AuNPs are shown in Supporting Information, Figure S7 and Table S1 respectively.[2]

9.2.4 Characterization of AuNP Hydrodynamic and Electrokinetic Properties

The hydrodynamic diameters and apparent zeta potentials of the AuNPs were determined by dynamic light scattering and laser Doppler electrophoresis, respectively (Malvern Zetasizer Nano ZS). Gold NPs (10 nM) were suspended in water or 10 mM HEPES 10 mM NaCl buffered at pH 7.4 for 30 minutes prior to measurement (Supporting Information, Figure S8).[2] In preliminary experiments, we determined that the AuNPs remained colloidally stable in 10 mM NaCl for the duration of the QCM-D experiments. We therefore selected this salt concentration to study AuNP interaction with bilayers. Higher salt concentrations led to much more pronounced aggregation and destabilized the colloidal suspension.

9.2.5 Calculation of Ligand R group Lipophilicity

We calculated the lipid-water partition coefficient ($K_{lip\text{-}w}$) of ligand R groups as a measure of ligand lipophilicity using the poly-parameter

linear free energy relationship:[13]

$$\log K_{\text{lip-w}} = c + eE + sS + aA + bB + vV \tag{9.1}$$

where E, S, A, B, and V describe the ligand head group excess molar refraction, dipolarity/polarizability, H-bond acidity, H-bond basicity, and molar volume, respectively, and the corresponding lowercase letters are specific for the water-lipid partitioning system.[13] The constant c is obtained from the multiple linear regression used to establish the system descriptors. The chemical descriptors for the ligand R groups (E, S, A, B, and V) were obtained from the Helmholtz Centre for Environmental Research Linear Solvation Energy Relationship.[37] Values for the systems parameters c, *e*, s, a, b, and *v* were taken from equation 3 of Endo *et al.*[13] The values for the chemical descriptors and system parameters are compiled in Supporting Information, Table S2.[2]

9.2.6 Vesicle Preparation and SLB Formation

Solutions for all experiments employing lipids were buffered to pH 7.4 with 10 mM HEPES. Vesicles were formed from 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC; Avanti Polar Lipids, 850375) via vesicle extrusion.[38,39] In short, chloroform was removed from DOPC by evaporation in a vacuum chamber for 1 h. We resuspended DOPC (2.5 mg·mL^{-1}) in 1 mM NaCl. To form vesicles, lipid suspensions were sonicated for 30 min

followed by three freeze-thaw cycles (incubation in liquid nitrogen for 5 min followed by sonication at room temperature for 5 min). Vesicles were extruded 11 times through 50 nm polycarbonate filters. The DOPC vesicles had hydrodynamic diameters between 90 and 110 nm as determined by DLS and ζ-potentials of 0 to -6 mV as determined by laser Doppler electrophoresis. Vesicles were stored at 4 °C and used within 10 days.

Supported lipid bilayers were formed on SiO_2-coated QCM-D sensors (QSX203) in a Q-Sense E4 instrument (Biolin Scientific) by vesicle fusion.[22,39] In short, vesicles (0.125 mg·mL^{-1}) in 100 mM NaCl were flowed (0.100 mL·min^{-1}) over sensors freshly cleaned in an UV/ozone chamber for 20 min. The DOPC vesicles attained a critical surface concentration on the sensors after ~5 min, at which point the vesicles fused and ruptured and a stable bilayer was formed. After signal stabilization, the stable bilayer was rinsed for 10 min with 100 mM NaCl to remove any loosely adhered vesicles. Example frequency and energy dissipation traces for bilayer formation are shown in Supporting Information, Figure S9.[2] The final frequency change for the bilayers was 25 ± 0.5 Hz and the dissipation factor was $0.2(\pm 0.1) \times 10^{-6}$, consistent with values previosuly reported for supported DOPC bilayers.[38]

9.2.7 Nanoparticle Adsorption and Desorption Experiments

After formation of stable DOPC bilayers on SiO_2-coated QCM-D sensors, we flowed 10 mM NaCl over bilayer until a stable baseline was achieved. We then introduced AuNPs (10 nM) at the same flow rate and in solution of the same composition until a stable plateau in frequency was attained (within 20 min). At this point, AuNP-free solution was introduced into the flow cell and the bilayer was rinsed until a stable baseline was observed. In preliminary experiments with C_{10}-AuNPs, we determined that a AuNP concentration of 10 nM and an exposure time of 20 min was sufficient for a plateau in frequency to be attained. Figure S10 shows changes in acoustic surface mass density and energy dissipation upon exposure of supported lipid bilayers to AuNPs decorated with each of the indicated ligands studied.[2]

We calculated surface mass densities from frequency shifts (Δf_n) using the Sauerbrey equation shown in Equation 9.2.[40]

$$\Gamma_{QCMD} = -C\frac{\Delta f_n}{n} \tag{9.2}$$

where C is the mass sensitivity constant and n is the harmonic number.[40] Data for AuNP-bilayer interaction were determined to fall within the Sauerbrey regime, $4x10^{-7}Hz^{-1} >> \frac{-\Delta D_n}{\frac{\Delta f_n}{n}}$ for a 5 MHz crystal.[22] All data taken between a bilayer baseline and final baseline fit the Sauerbrey regime. For

all analysis the 5th harmonic was used. We calculated the maximum surface mass density, Γ_{max}, at the end of the adsorption phase. We determined $d\Gamma/dt$ by taking the derivative of Γ with respect to time: a linear regression algorithm using 33 points (~30 s) centered around one point was used to calculate the derivative. To model the rate coefficients, k_a and k_d, a single-parameter optimized least squares model was used with equations 9.5 and 9.6 respectively. Adsorption data were fitted starting from the time the AuNP nanoparticle suspension had displaced AuNP-free solution in the flow chamber (indicated by a positive peak in $d\Gamma/dt$) and ending when the AuNP suspension had been displaced from the flow chamber by AuNP-free solution (indicated by a negative peak in $d\Gamma/dt$). The time period over which adsorption data were fitted (14 min) corresponds to the duration of AuNP flow minus the time of AuNP addition and removal from the flow cell. Desorption data were fitted for 8 min after observation of the negative peak in $d\Gamma/dt$ associated with displacement of the AuNP suspension from the flow chamber. The results of these fits are shown in Supporting Information, Figure S11 and Figure S12 for adsorption and desorption curves respectively.[2] Goodness of fit was calculated as a non-linear R^2 value and reported in Supporting Information, Table S3.[2]

9.2.8 System Setup for Classical MD Simulations

Interactions of AuNPs with DOPC bilayers were modeled with classical MD simulations using Gromacs 2016.[41] The simulation workflow for

developing the AuNP-DOPC systems is summarized in Supporting Information, Figure S13.[2] AuNPs were modeled using a self-assembly protocol described previously,[42] which outputs 2-nm diameter AuNPs (226 gold atoms) with 83 ligands. AuNPs were modeled by using the INTERFACE force field[43] for gold atoms and the CHARM36/CGenFF force fields[44–46] for the ligands. The AuNPs were solvated with the TIP3P water model[47] with sufficient chlorine counterions to ensure the system is charge-neutral then equilibrated for 5 ns at a temperature T = 300 K (controlled by a velocity-rescale thermostat) and pressure P = 1 bar (controlled by a Berendsen barostat). A 50 ns *NPT* simulation was subsequently performed at the same temperature and pressure, controlled by the same thermostat and Parrinello-Rahman barostat. The last configuration of the 50 ns *NPT* simulation was used to initiate AuNP-DOPC lipid membrane simulations.

The DOPC bilayer was generated using CHARMM36-GUI web-interface[48] with 196 lipids in each of the top and bottom leaflets. The dimensions of the DOPC bilayer were selected to avoid interactions between AuNPs due to the periodic boundary conditions (see Supporting Information, Figure S14).[2] The bilayer was equilibrated with water at T = 300 K with semi-isotropic pressure coupling in the x-y dimensions at P = 1 bar, controlled by Nose-Hoover thermostat and Parinello-Rahman barostat. One AuNP was then inserted into the DOPC system such that the gold core center-of-mass was 6.9 nm away from DOPC center-of-mass and solvated with water (Figure 9.1c). The AuNP-DOPC lipid membrane system was

then equilibrated for 10 ns in the *NPT* ensemble with the AuNP restrained at a 6.9 nm distance from the surface of DOPC with a harmonic potential of 2000 kJ·mol^{-1}·nm^{-2} controlled by the Berendsen thermostat and barostat.

9.2.9 Increasing- and Decreasing-z US Simulations and Subsequent Unbiased Simulations

We defined z as the center-of-mass distance between the gold core and DOPC bilayer. Initial configurations for the decreasing-z US simulations were generated by pulling the gold core towards the bilayer at 0.0005 nm·ps^{-1} using a harmonic potential with a spring constant of 2000 kJ·mol^{-1}·nm^{-2}. Pulling was performed in the *NPT* ensemble at T = 300 K and P = 1 bar, controlled by the velocity-rescale thermostat and semi-isotropic Berendsen barostat. Configurations from this trajectory were used to perform the decreasing-z US simulations with z varying from 1.3 nm to 6.5 nm. Initial configurations for the increasing-z US simulations were generated by pulling the gold core away from the bilayer starting from the final configuration of the z = 1.3 nm window from the decreasing-z US simulations. This pulling simulation used the same harmonic potential and pull rate as the decreasing-z pulling simulations. Configurations from this trajectory were used to perform increasing-z US simulations with z varying from 1.5 to 6.5 nm.

All US simulation windows were equilibrated for 500 ps in the *NPT*

ensemble using the velocity-rescale thermostat and Berendsen barostat then simulated for 50 ns using the same thermostat and the Parinello-Rahman barostat. Some windows were extended to an additional 100 ns as discussed in the Supporting Information.[2] The last 40 ns of the production simulation for each window were used to compute the PMF with the Weighted Histogram Analysis Method (WHAM).[49] Additional details on the US protocol and number of simulation windows are provided in the Supporting Information.[2]

Unbiased MD simulations were performed using the 50 ns configuration from either a decreasing- or increasing-z US simulation window. The configuration was equilibrated for 500 ps with the z value restrained using the same spring constant as US simulations. Then, a 50 ns unbiased *NPT* simulation was performed at P = 1 bar controlled by the Parinello-Rahman barostat and T = 300 K controlled by the velocity-rescale thermostat.

9.2.10 Quantifying the Number of Hydrophobic Contacts

The total number of hydrophobic contacts (c_h) was defined in Equation 9.3 by summing all possible contacts between alkane and R group atoms of the ligands (i) and tail group atoms of DOPC (j).

$$c_h = \sum_i \sum_j s_{ij} \tag{9.3}$$

Equation 9.4 defines s_{ij} as a continuous function that smoothly decays between 1 (corresponding to a hydrophobic contact) and 0 as a function of distance between atoms i and j (r_{ij}).

$$s_{ij} = \frac{1 - \left(\frac{r_{ij}}{r_o}\right)^6}{1 - \left(\frac{r_{ij}}{r_o}\right)^{12}} \tag{9.4}$$

r_o is defined as the cutoff when s_{ij} approaches zero and was set to 0.35 nm. Hydrogen atoms were not considered when quantifying the total number of hydrophobic contacts. Hydrophobic contacts were computed using PLUMED Version 2.5.1.[50]

9.2.11 Hydrophobic Contact US Simulations

We performed US simulations using hydrophobic contacts (c_h) as the collective variable using PLUMED Version 2.5.1 in conjunction with Gromacs 2016.6. Initial AuNP-DOPC configurations used the last configuration from the decreasing-z US simulations at z = 5.1 nm, which has $c_h = 0$ for C_1- and C_{10}-AuNPs (Figure 9.3c). Initial configurations for US simulations were generated using a *NVT* pulling trajectory with a spring constant of 50 kJ·mol^{-1}·contacts^{-2} (see Supporting Information, Table S7).[2] US simulations were initiated using configurations from the pulling simulations with c_h values between 0 - 150 contacts in increments of 2.5 contacts, totaling up to 61 windows. A spring constant of 10 kJ·mol^{-1}·contacts^{-2} was

used for all US simulations. Each simulation window was equilibrated for 500 ps using the velocity-rescale thermostat then simulated for 80 ns using the same thermostat and the Parinello-Rahman barostat. The last 60 ns of the production simulation for each window were used to compute the PMF with WHAM.[49] For Bn-AuNPs, the last 40 ns of 50 ns production simulations for each window was used to compute the PMF, which was sufficient for convergence (see Supporting Information, Figure S19).[2]

9.2.12 Simulation Parameters

For all simulations, Verlet lists were generated using a 1.2-nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential using a 1.2-nm cutoff that was smoothly shifted to zero between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the smooth particle mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and fourth order interpolation. Bonds were constrained using the LINCS algorithm.[51] Periodic boundary conditions were enabled in all directions.

9.3 Results and Discussion

9.3.1 Ligand R Group Lipophilicity Governs AuNP Adsorption to Lipid Bilayers

Figure 9.1a shows the library of cationic ligands used to functionalize the 2-nm diameter AuNPs employed in this study. As described in the Methods, ligand lipophilicity is expressed as the equilibrium partition coefficient between phosphatidylcholine liposomes and water ($K_{lip\text{-}w}$) calculated for each ligand R group: methyl (C_1), ethyl (C_2), butyl (C_4), benzyl (Bn), and decyl (C_{10}).[13] Larger log $K_{lip\text{-}w}$ values correspond to a higher propensity for the ligand R group to partition into phosphatidylcholine liposomes from water. By modulating only the ligand R group, we systematically study the effects of R group lipophilicity while keeping the gold core size, charge, and ligand backbone constant. We determined AuNP hydrodynamic diameter and apparent zeta potential by dynamic light scattering and laser Doppler electrophoresis, respectively. Figure S8 shows that in the aqueous solution used for our experiments (10 mM NaCl buffered to pH 7.4 with 10 mM HEPES), the hydrodynamic diameters, and therefore degrees of aggregation, of the AuNPs in our library were comparable. The only statistically significant difference in hydrodynamic diameter was between the Bn- and C_{10}-AuNPs ($p = 0.314$), which we attribute to differences in the polarizability and flexibility of the ligand R groups (*i.e.*, higher polarizability of the benzyl group leading to stronger van der Waals

interactions, higher flexibility of the C_{10} R group allowing it to minimize solvent exposure by folding back on itself). The apparent zeta potentials of the AuNPs in our library were statistically indistinguishable and did not differ from zero ($p > 0.05$, Figure S8d;[2] zeta potential mostly between -10 and $+10$ mV).[52] Given the near neutral apparent zeta potentials of the AuNPs, we do not expect electrostatics to dominate their interaction with bilayers. Similar libraries of AuNPs have been used to correlate cytotoxicity and hydrophobicity.[17]

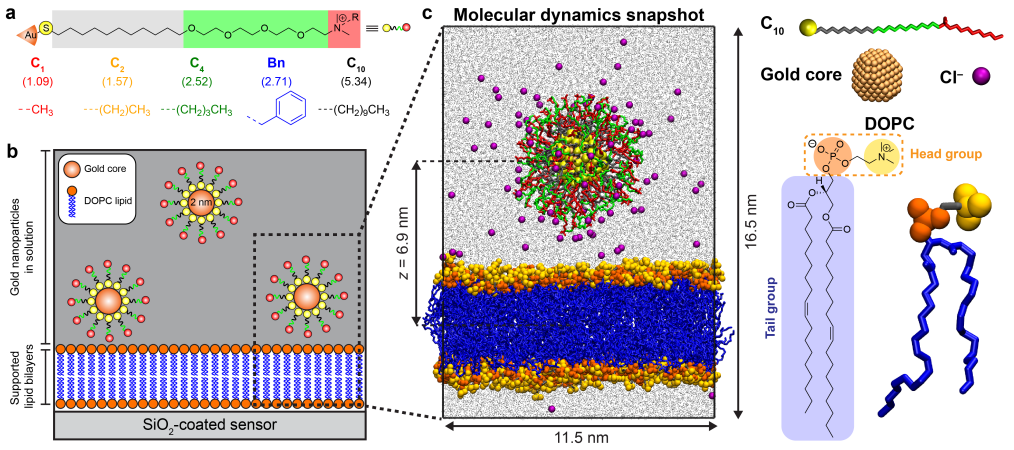


Figure 9.1: Experimental and computational systems used to study gold nanoparticle adsorption onto phospholipid bilayers. (a) Ligands are comprised of an alkane group (gray), an oligo(ethylene glycol) spacer group (green), and a cationic quaternary ammonium group (red) substituted with the indicated R group and two methyl groups. The five R groups used are displayed in red and labeled with their calculated log $K_{lip\text{-}w}$ values in parentheses. (b) Schematic of the system used in quartz crystal microbalance experiments to measure nanoparticle adsorption to supported DOPC lipid bilayers. (c) Snapshot of 2-nm gold nanoparticle with C_{10} ligands placed above a DOPC lipid bilayer. The color scheme is illustrated for each of the components at right. The DOPC lipids are comprised of a zwitterionic phosphatidylcholine head group and nonpolar acyl tails consisting primarily of aliphatic carbon atoms. Water is shown in grey.

As described in the Methods, SLBs were prepared from 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) using the vesicle fusion method,[52] after which AuNPs (10 nM) were flowed over DOPC lipid bilayers for 20 minutes (schematically represented in Figure 9.1b) followed by rinsing with AuNP-free solution for extended periods of time. We characterize AuNP-SLB interactions at two distinct time points: after 20 minutes

flow-experimentally determined as sufficient time to attain an adsorption plateau-and after extended rinsing. We report QCM-D results as acoustic surface mass density (Γ) and the change in energy dissipation (ΔD). We obtained Γ from frequency changes using the Sauerbrey equation.[22]

Figure 9.2a shows the acoustic surface mass density for the AuNPs after 20 minutes flow (Γ_{max}) and after rinsing (Γ_{rinse}). We find that C_2-, C_4-, Bn-, and C_{10}-AuNPs adsorb to the DOPC bilayer, whereas any adsorption of C_1-AuNPs was not detectable. For the series of AuNP ligands used, Γ_{max} correlates positively with $\log K_{lip\text{-}w}$. For C_2-AuNPs, $\Gamma_{rinse} = 0$, indicating that these AuNPs reversibly adsorb to the lipid bilayer and rinse away using buffer. In contrast, $\Gamma_{rinse} > 0$ for C_4-, Bn, and C_{10}-AuNPs, suggesting that a population of these AuNPs remain quasi-irreversibly bound to the bilayer. Given the similarities in particle core size, apparent zeta potential, ligand structure (with the exception of the R group), and hydrodynamic diameter, we expect that the nature of the ligand R group was the dominant contributor to AuNP-bilayer interactions; however, we cannot exclude that aggregation state had a minor influence in the case of the Bn-AuNPs. Increases in AuNP-bilayer interaction with increasing lipophilicity is consistent with established ideas of lipophilic ligand-mediated NP toxicity.[17]

Figure 9.2b shows change in energy dissipation after 20 minutes flow (ΔD_{max}) and after rinse (ΔD_{rinse}). The C_1-AuNPs do not effect a detectable dissipation change, consistent with the lack of observed mass attachment

(Figure 9.2a). For C_2-AuNPs, zero Γ_{rinse} corresponds with nonzero ΔD_{rinse}. Without a quantifiable population of AuNPs adsorbed to the SLB, either small undetectable populations of AuNPs induce detectable viscoelastic changes to the bilayer, or AuNP adsorption induces a permanent change in the energy dissipation of the bilayer that persists after AuNPs have rinsed away. For C_4-, Bn-, and C_{10}-AuNPs, ΔD_{max} is statistically indistinguishable from ΔD_{rinse} ($p < 0.05$). A population of reversibly adsorbed AuNPs leave the bilayer upon rinsing without producing a detectable change in energy dissipation. From the QCM-D results, we hypothesize that the lipophilicity of C_2-, C_4-, Bn-, and C_{10}-AuNPs leads to spontaneous adsorption onto DOPC lipid bilayers and subsequent formation of a quasi-irreversibly bound state in a subset of the adsorbed AuNP population. Ligand lipophilicity appears to determine the degree of quasi-irreversible interaction, with the least lipophilic ligands leading to negligible quasi-irreversible adsorption. We employ classical molecular dynamics simulations to measure the free energy barriers to forming these states and to gain insight into the mechanism of quasi-irreversibly binding.

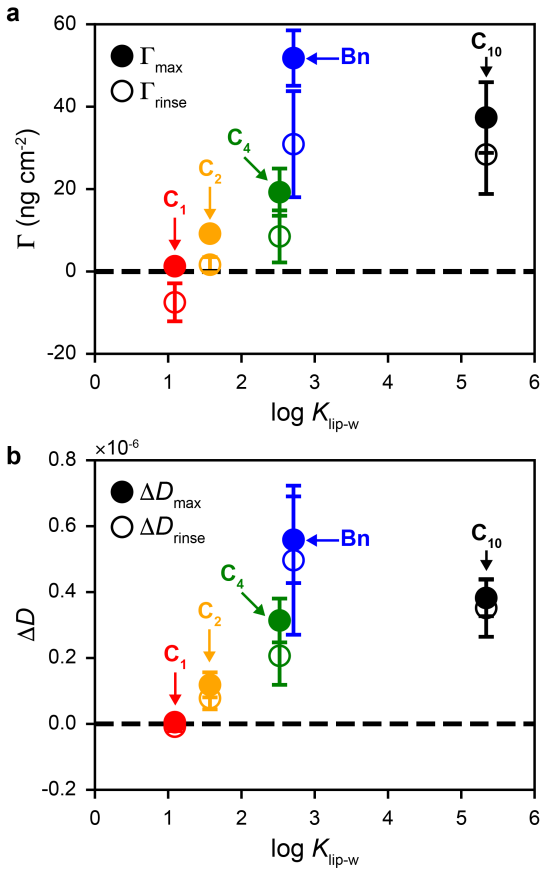


Figure 9.2: Influence of ligand lipophilicity on AuNP attachment to supported DOPC bilayers as determined by QCM-D. (a) Acoustic surface mass density (Γ) maximum and after rinse. (b) Dissipation factor (ΔD) maximum and after rinse. Error bars represent one standard deviation from four replicate QCM-D experiments.

9.3.2 Contact between Lipophilic Ligand Groups and Lipid Tails Promotes Adsorption

We first modeled interactions between C_1- and C_{10}-AuNPs and DOPC lipid bilayers to understand the molecular interactions that drive AuNP adsorption. These AuNPs were selected because the ligands represent the extremes of $\log K_{lip\text{-}w}$ in this study and the AuNPs exhibit distinct adsorption behavior: C_1-AuNPs do not adsorb to DOPC bilayers, whereas C_{10}-AuNPs adsorb quasi-irreversibly (Figure 9.2a). We first quantified potentials of mean force (PMFs) for C_1- and C_{10}-AuNP adsorption using umbrella sampling (US) simulations. The PMF measures the free energy as a function of a collective variable, z, which we define as the distance in the direction normal to the bilayer between the gold core and DOPC center-of-mass (Figure 9.1c) following past literature.[27–30] We calculated PMFs using initial simulation configurations generated by pulling the AuNP either from an initial position in water toward the bilayer (decreasing-z simulations) or from an initial position in the bilayer toward water (increasing-z simulations). These two methods were used to interrogate potential hysteresis associated with long timescale bilayer rearrangements.[53,54]

Figure 9.3 compares decreasing-z and increasing-z PMFs for C_1- and C_{10}-AuNPs. Both sets of PMFs are comparable: the decreasing-z PMFs monotonically increase as z decreases, whereas the increasing-z PMFs exhibit free energy minima at positions near the water-bilayer interface

($z \approx 5$ nm). Simulation snapshots indicate that the free energy minima correspond to configurations in which nonpolar lipid tail groups are in contact with lipophilic groups on the C_1- and C_{10}-AuNPs, suggesting that these favorable contacts lead to thermodynamically preferred adsorbed states for both AuNPs. These minima are consistent with the experimentally observed quasi-irreversibly bound states for C_{10}-AuNPs but do not explain the inability of the C_1-AuNPs to adsorb. However, the pronounced hysteresis between the increasing-z and decreasing-z PMFs suggests that using z as a collective variable does not capture potential free energy barriers that could inhibit adsorption, as previously observed in simulations of lipid insertion into a bilayer.[55]

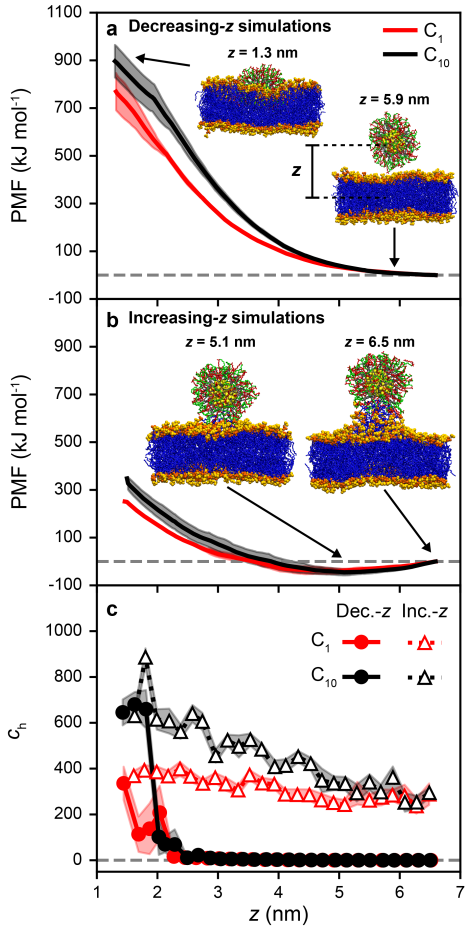


Figure 9.3: Free energy as a function of the z distance between the AuNP and lipid bilayer. Potential mean force (PMF) versus z for C_1- and C_{10}-AuNPs when the gold core is (a) pulled towards (*i.e.* decreasing-z) and (b) away from (*i.e.* increasing-z) the DOPC lipid bilayer. Simulation snapshots show the last configuration from umbrella sampling simulations of C_{10}-AuNPs for different values of z. Water and chlorine atoms are omitted for clarity. Legends are the same for (a) and (b). (c) Number of hydrophobic contacts (c_h) versus z for both decreasing- and increasing-z simulations. Hydrophobic contacts are defined as the number of contacts between nonpolar groups in the ligands and in the DOPC tail groups. Error bars are reported as the standard deviation between two 20 ns blocks for each umbrella sampling window.

9.3.3 Unbiased Simulations Reveal Two Mechanisms Leading to Prolonged AuNP Adsorption to Bilayers

We hypothesized that barriers to the formation of hydrophobic contacts between lipids and the AuNP could explain the differences between the increasing-z and decreasing-z PMFs; previous studies have also identified hydrophobic contacts as important for favorable AuNP-bilayer interactions.[32,56,57] Therefore, we calculated the total number of hydrophobic contacts (c_h) between the alkane and R groups of the ligands and the tail groups of DOPC. Figure 9.3c plots c_h as a function of z for both increasing-z and decreasing-z simulations. For both AuNPs, many hydrophobic contacts persist during the increasing-z simulations, indicating that hydrophobic contacts are highly favorable. Conversely, hydrophobic contacts are observed for only small values of z during the decreasing-z simulations, indicating the presence of hidden barriers that prevent contacts from forming. To confirm that hydrophobic contacts are important for adsorption, we performed 50-ns unbiased simulations initiated from configurations with different z and c_h values to determine if the C_1- and C_{10}-AuNPs desorb from the bilayer. Figure 9.4a shows that both C_1- and C_{10}-AuNPs remain adsorbed to the lipid bilayer if the initial value of c_h exceeds ~40 contacts, even for large values of z. These unbiased simulations are consistent with the hypothesis that hydrophobic contacts between the AuNP ligands and lipid membrane are important for adsorption but not captured by z alone.

For unbiased simulations in which the AuNPs remain adsorbed, z increases until $z \approx 5$ nm, thus reaching states consistent with the free energy minima obtained from the increasing-z PMFs. Figure 9.4b shows final snapshots from the unbiased simulations of C_1- and C_{10}-AuNPs for $z \approx 2$ nm. The snapshots show two mechanisms that promote continued AuNP adsorption: (1) the C_1-AuNP remains adsorbed due to lipid extraction, where lipids are pulled away from the bilayer, and (2) the C_{10}-AuNP remains adsorbed due to both lipid extraction and ligand intercalation, where some C_{10} ligands extend into the hydrophobic core of the bilayer. In both mechanisms, hydrophobic contacts drive the rearrangement of lipids to facilitate AuNP adsorption. C_1 ligands are less capable of facilitating ligand intercalation relative to C_{10} ligands, because the C_1 hydrophobic chain length is shorter than C_{10}.

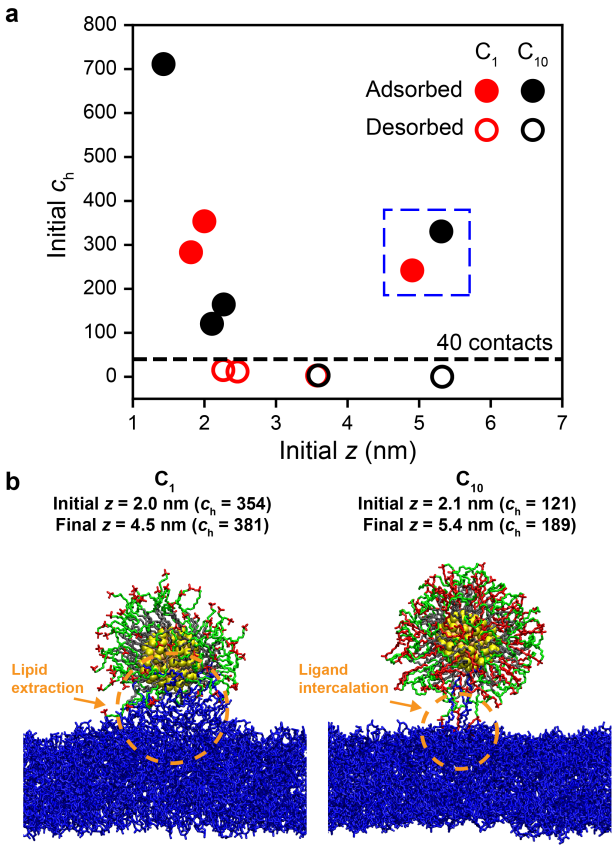


Figure 9.4: Unbiased simulations initiated from umbrella sampling trajectories. (a) Number of hydrophobic contacts (c_h) versus z for unbiased simulations initiated from increasing- and decreasing-z US configurations for C_1- and C_{10}-AuNPs. AuNPs are considered adsorbed if $z < 6$ nm for the last 10 ns of the unbiased simulation (filled markers) and desorbed if $z > 6$ nm (hollow markers). Points in the dashed blue box were for unbiased simulations initiated from increasing-z US configuration, all other points were for simulations initiated from decreasing-z US configurations. (b) Simulation snapshots after 50 ns of unbiased simulation for C_1- and C_{10}-AuNPs. The initial z values and final z values after 50 ns are labeled above the snapshots. Atoms in the DOPC head groups are omitted for clarity.

9.3.4 Ligand Intercalation within Bilayer Reduces Barrier for Forming Hydrophobic Contacts

The results from the US and unbiased simulations suggest that hydrophobic contacts are important for AuNP adsorption but are not sampled when using z as a collective variable. We therefore performed US using c_h as a collective variable to calculate a corresponding PMF. We included Bn-AuNP in this analysis, which has intermediate ligand lipophilicity compared to the other ligands and adsorbs quasi-irreversibly (Figure 9.2a). Figure 9.5a shows PMFs for C_1-, Bn-, and C_{10}-AuNPs as a function of c_h. The PMF for the C_1-AuNP monotonically increases with c_h, indicating that the initial formation of hydrophobic contacts is unfavorable. The PMF for Bn-AuNP also monotonically increases with larger c_h values, but with a substantially lower slope than C_1-AuNP, which indicates that the more lipophilic Bn ligands reduce the free energy barrier to forming hydrophobic contacts. The PMF for C_{10}-AuNPs has an initial free energy barrier at $c_h = 5$ contacts and a local minimum at $c_h \approx 40$ contacts, then increases monotonically at larger c_h values. The local minimum indicates a metastable state due to the formation of favorable hydrophobic contacts. This metastable state for C_{10}-AuNPs may lead to subsequent ligand intercalation or lipid extraction steps that result in the quasi-irreversibly bound states observed from experiments and the global minimum indicated by the increasing-z PMFs (Figure 9.3b).

Figure 9.5b shows simulation snapshots of C_1-, Bn-, and C_{10}-AuNPs for various c_h values. For C_1-AuNPs, ligands contact the bilayer for $c_h \leqslant 30$ contacts, then a single lipid molecule is extracted for $c_h \geqslant 40$ contacts. At $c_h = 150$ contacts, the C_1-AuNP desorbs from the DOPC lipid membrane even after extracting two lipid molecules, suggesting that lipid extraction is not sufficiently favorable to promote adsorption without more substantial bilayer deformations (like those observed in Figure 9.4b). Conversely, the snapshots of the Bn- and C_{10}-AuNPs show that ligands intercalate within the bilayer. For the Bn-AuNP, multiple ligands intercalate within the bilayer with increasing c_h. We attribute the smaller slope of the Bn-AuNP PMF compared to the C_1-AuNP PMF to favorable intercalation. For the C_{10}-AuNP, a single ligand intercalates within the bilayer at $c_h \approx 40$. This value of c_h corresponds to the local minimum in the PMF and is comparable to the threshold for stable adsorption identified from unbiased simulations (Figure 9.4a). Unlike the Bn-AuNP, multiple C_{10} ligands only intercalate within the bilayer for large values of c_h ($\approx$ 150 contacts). We further tested if ligand intercalation is sufficient to keep the C_{10}-AuNP adsorbed to the DOPC lipid bilayer. We performed four unbiased simulations of each of the three AuNPs initiated with c_h = 40 and found that the C_{10}-AuNPs remain adsorbed, whereas most C_1-AuNPs rapidly desorb and Bn-AuNPs desorb but at a slower rate than C_1-AuNPs (Supporting Information, Figure S22).[2]

Taken together, the simulation results suggest that C_1-, Bn- and C_{10}-

AuNPs can favorably adsorb to the bilayer if sufficient hydrophobic contacts are formed between the AuNP and the bilayer (Figure 9.3b). However, C_1-AuNPs can form hydrophobic contacts with the bilayer only via a lipid extraction mechanism that is associated with a large free energy barrier (Figure 9.5). This barrier thus inhibits adsorption in agreement with experimental measurements. Conversely, the more lipophilic R groups present in the Bn and C_{10} ligands intercalate within the bilayer to promote the formation of hydrophobic contacts (Figure 9.5). Additional snapshots illustrating these two mechanisms are shown in the Supporting Information, Figure S20.[2] Intercalation could facilitate the initial favorable adsorption of Bn- and C_{10}-AuNPs followed by longer timescale interconversion to a quasi-irreversibly adsorbed state associated with many hydrophobic contacts (Figure 9.2a). Furthermore, ligand intercalation for C_{10}-AuNPs could explain the experimentally observed bilayer viscoelastic change (Figure 9.2b).

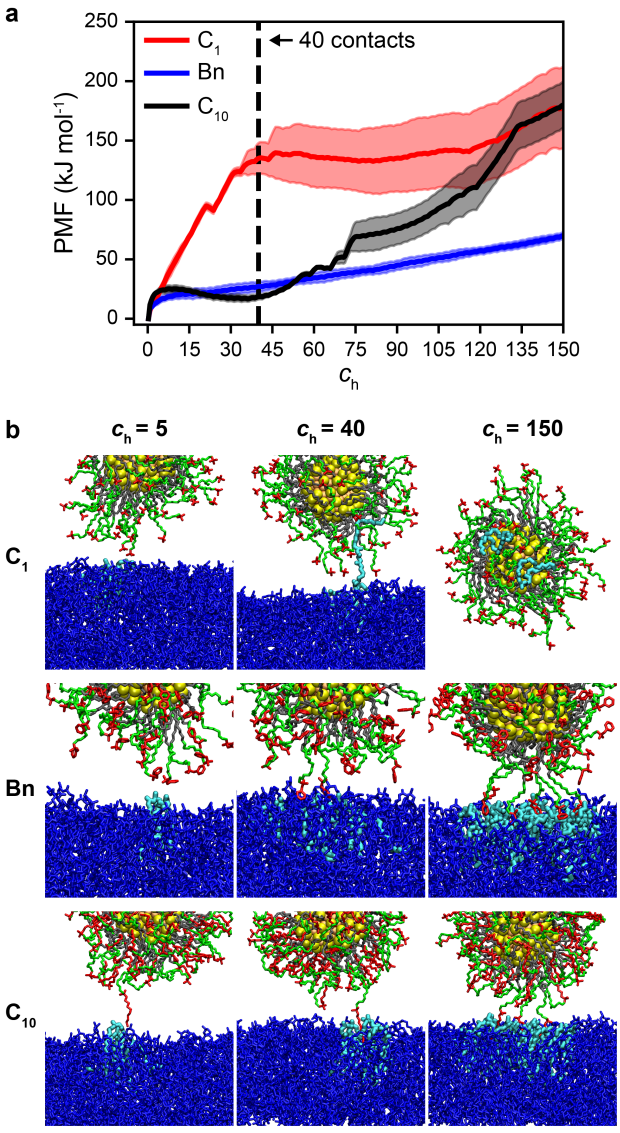


Figure 9.5: Free energy as a function of AuNP-bilayer hydrophobic contacts. (a) Potential of mean force (PMF) versus the number of hydrophobic contacts (c_h) for C_1-, Bn-, and C_{10}-AuNPs. Error bars are reported as the standard deviation between two 30 ns trajectories for C_1- and C_{10}-AuNPs and two 20 ns trajectories for Bn-AuNP in each umbrella sampling window. (b) Simulation snapshots with c_h = 5, 40, and 150 for C_1-, Bn-, and C_{10}-AuNPs. DOPC lipids that are within 0.35 nm of the ligand atoms are highlighted in cyan.

9.3.5 Two-state Adsorption Kinetics Describes AuNP Reversibly and Quasi-Irreversibly Adsorbed States

We constructed an analytical kinetic model to investigate the connection between the findings from MD simulations and the experimentally observed quasi-irreversible binding of AuNPs to lipid bilayers. QCM-D cannot directly measure the extent of ligand intercalation, but it does allow two distinct adsorbed states to be distinguished.[25,26,58] We define a three-state model to describe the kinetics of AuNP-SLB interactions which includes a (1) metastable, reversibly adsorbed state, which can undergo either further ligand intercalation, lipid extraction, or desorption; (2) quasi-irreversibly adsorbed state, which corresponds to a high number of hydrophobic contacts consistent with a global free energy minimum (Figure 9.3b); and (3) desorbed state, which could reflect barriers to forming initial hydrophobic contacts necessary for adsorption.

The two adsorbed states represent a reversibly adsorbed (Γ_α) and a quasi-irreversibly adsorbed (Γ_β) population of nanoparticles. At any given time, the total mass of adsorbed particles equals the sum of the masses of the reversibly and quasi-irreversibly adsorbed particles: $\Gamma_{\text{total}}(t) = \Gamma_\alpha(t) + \Gamma_\beta(t)$. Assuming first-order adsorption kinetics, the rate of mass adsorption (the first derivative of mass adsorption with respect to time) can be expressed as a function of the first-order rates of adsorption (k_a), conversion to quasi-irreversibly adsorbed state (k_β), and desorption (k_d).

Equation 9.5 describes the adsorption kinetics, as modified from Zhang *et al.*[25]

$$\frac{d\Gamma_{\alpha}}{dt} = k_a m_s \frac{\Gamma_{max} - \Gamma_{total}(t)}{\Gamma_{max}} - k_d \theta \Gamma_{total}(t) - k_\beta \theta \Gamma_{total}(t) \tag{9.5}$$

where m_s is the mass density of AuNPs in solution, and θ is the ratio of reversibly to total adsorbed particles, approximated as $\theta = \frac{\Gamma_{rinse}}{\Gamma_{max}}$ The adsorption capacity of the bilayer was taken as Γ_{max}. Particles that rinse from the bilayer were considered to have been in the reversibly adsorbed α state. Equation 9.6 describes the desorption of AuNPs from a SLB after AuNPs have been removed from solution.

$$\frac{d\Gamma}{dt} = -k_d \Gamma_{total}(t) \tag{9.6}$$

To determine the rate coefficient for conversion to the quasi-irreversible adsorption state (k_β), we consider the rate of change of Γ_β independently described by Equation 9.7.

$$\frac{d\Gamma_{\beta}}{dt} = k_\beta \theta \Gamma_{total}(t) \tag{9.7}$$

We have two distinct time points for Γ_β ($t = 0$ min and $t = 20$ min). By integrating over the total time of NP-bilayer exposure ($t_{max} = 20$ min), we determine k_β assuming first-order kinetics shown in Equation 9.8.

$$k_\beta = \frac{\Gamma_\beta}{\Gamma_{total} \theta t} \tag{9.8}$$

Figure 9.6 shows the rate coefficients k_a, k_β, and k_d for the different AuNPs. The rate coefficient for adsorption, k_a, increases monotonically with increasing ligand R group lipophilicity as expressed by $\log K_{lip\text{-}w}$ (Figure 9.6a). k_a is dictated by the free energy of binding, which is positively correlated to ligand lipophilicity. The rate coefficient for conversion from the reversible to the quasi-irreversible adsorbed state, k_β, appears to increase with lipophilicity for C_2-, C_4-, Bn- and C_{10}-AuNPs (Figure 9.6b). After initial adsorption, subsequent ligand intercalation into the bilayer and lipid extraction from the bilayer facilitates the formation of a quasi-irreversibly adsorbed state (Figure 9.4); therefore, we expect that k_β is dictated by the free energy of hydrophobic contact between ligand R groups and DOPC tail groups. As explored in the MD simulations, increasing ligand R group lipophilicity leads to a decreased free energy barrier for forming hydrophobic contacts (Figure 9.5). The rate coefficient for desorption of AuNPs in the reversibly adsorbed state, k_d, appears to be independent of ligand R group lipophilicity (Figure 9.6c), suggesting the mechanism for desorption remains constant between AuNPs of varying ligand lipophilicity. These findings indicate that selection of ligand R group lipophilicity could drive reversible or quasi-irreversible AuNP adsorption onto DOPC bilayers.

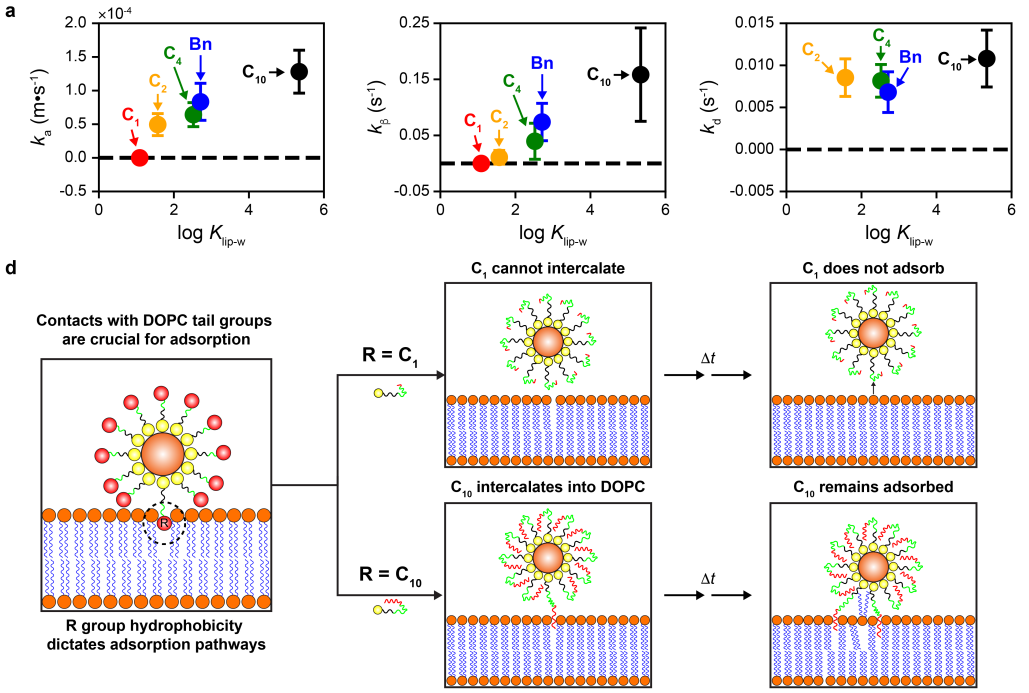


Figure 9.6: The role of ligand end group lipophilicity on adsorption to and desorption from phospholipid bilayers. (a) Adsorption rate constant (k_a) calculated from AuNP adsorption with calculated desorption rate constant (k_d) values. (b) Rate constants for conversion to quasi-irreversibly adsorbed state (k_β) calculated from mass at maximum and after rinse. (c) Desorption rate constants (k_d) calculated from AuNP desorption. Error bars represent one standard deviation of four replicate QCM-D measurements. (d) Schematic showing hypothesized mechanism for preferential adsorption of C_{10}-AuNPs compared to C_1-AuNPs. C_{10}-AuNPs have a longer R group, denoted by the red lines. Alkane is shown as black lines and PEG is shown as green lines. The symbol and color for the gold core and lipid bilayer is the same as Figure 9.1b.

9.4 Summary

We explored the role of ligand lipophilicity on the adsorption of 2 nm diameter AuNPs to zwitterionic phospholipid bilayers using QCM-D measurements and atomistic MD simulations. The experiments indicated that AuNPs coated with lipophilic ligands can adsorb quasi-irreversibly to DOPC bilayers. The extent of conversion to the quasi-irreversibly adsorbed state scales with ligand functional group lipophilicity. The simulations revealed that AuNP adsorption depends on the number of hydrophobic contacts between the AuNP ligands and the phospholipid acyl chains, which can emerge either due to the intercalation of ligands within the bilayer or the extraction of lipids from the bilayer. Increasing ligand end group lipophilicity promotes intercalation within the bilayer rather than lipid extraction, which reduces the free energy barrier for forming hydrophobic contacts. Analytical modeling of experimental AuNP-bilayer interaction kinetics relates the MD findings to the formation of both reversibly and quasi-irreversibly adsorbed states.

Together, the experiments and simulations suggest a mechanism for AuNP adsorption that depends critically on ligand lipophilicity (Figure 9.6d). The first step for adsorption is hydrophobic contact between the ligand R group and DOPC tail groups. Since Bn and C_{10} ligands have lipophilic R groups, they can intercalate within the bilayer to promote hydrophobic contacts after overcoming a ~30 kJ·mol^{-1} free energy bar-

rier. At long time scales, Bn- and C_{10}-AuNPs form a quasi-irreversible adsorbed state corresponding to a large number of hydrophobic contacts, consistent with the global free energy minimum in Figure 9.3b. Conversely, C_1 ligands are unable to make sufficient hydrophobic contacts to maintain adsorption due to the large barrier required for lipid extraction (Figure 9.5), resulting in desorption. Our results show that functionalized nanomaterials coated with lipophilic end group ligands could have enhanced adsorption and long-term stability on cell membranes. Tuning nanomaterial ligand lipophilicity may provide a means to enhance targeted nanodrug delivery through selective intercalation of ligands into cell membranes; alternatively, decreasing ligand lipophilicity may allow for creation of environmentally benign nanomaterials.

9.5 References

[1] Lochbaum, C. A.; Chew, A. K.; Zhang, X.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A. Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers. *ACS Nano* **2021**, *15*, 6562–6572.

[2] Lochbaum, C. A.; Chew, A. K.; Zhang, X.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A. Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers [Supporting Information]. *ACS Nano* **2021**, *15*, 6562–6572.

[3] Maurer-Jones, M. A.; Gunsolus, I. L.; Murphy, C. J.; Haynes, C. L. Toxicity of engineered nanoparticles in the environment. *Analytical chemistry* **2013**, *85*, 3036–3049.

[4] Nel, A.; Xia, T.; Mädler, L.; Li, N. Toxic potential of materials at the nanolevel. *science* **2006**, *311*, 622–627.

[5] Hamers, R. J. Nanomaterials and global sustainability. *Accounts of chemical research* **2017**, *50*, 633–637.

[6] Johnston, L. J.; Gonzalez-Rojano, N.; Wilkinson, K. J.; Xing, B. Key challenges for evaluation of the safety of engineered nanomaterials. *NanoImpact* **2020**, *18*, 100219.

[7] Zhu, W.; von dem Bussche, A.; Yi, X.; Qiu, Y.; Wang, Z.; Weston, P.; Hurt, R. H.; Kane, A. B.; Gao, H. Nanomechanical mechanism for lipid bilayer damage induced by carbon nanotubes confined in intracellular vesicles. *Proceedings of the National Academy of Sciences* **2016**, *113*, 12374–12379.

[8] Chong, G.; Foreman-Ortiz, I. U.; Wu, M.; Bautista, A.; Murphy, C. J.; Pedersen, J. A.; Hernandez, R. Defects in Self-Assembled Monolayers on Nanoparticles Prompt Phospholipid Extraction and Bilayer-Curvature-Dependent Deformations. *The Journal of Physical Chemistry C* **2019**, *123*, 27951–27958.

[9] Bailey, C. M.; Kamaloo, E.; Waterman, K. L.; Wang, K. F.; Nagarajan, R.; Camesano, T. A. Size dependence of gold nanoparticle interactions with a supported lipid bilayer: A QCM-D study. *Biophysical chemistry* **2015**, *203*, 51–61.

[10] Liu, X.; Chen, K. L. Interactions of graphene oxide with model cell membranes: Probing nanoparticle attachment and lipid bilayer disruption. *Langmuir* **2015**, *31*, 12076–12086.

[11] Mensch, A. C.; Hernandez, R. T.; Kuether, J. E.; Torelli, M. D.; Feng, Z. V.; Hamers, R. J.; Pedersen, J. A. Natural organic matter concentration impacts the interaction of functionalized diamond nanoparticles with model and actual bacterial membranes. *Environmental science & technology* **2017**, *51*, 11075–11084.

[12] Olenick, L. L.; Troiano, J. M.; Vartanian, A.; Melby, E. S.; Mensch, A. C.; Zhang, L.; Hong, J.; Mesele, O.; Qiu, T.; Bozich, J.; et al.. Lipid corona formation from nanoparticle interactions with bilayers. *Chem* **2018**, *4*, 2709–2723.

[13] Endo, S.; Escher, B. I.; Goss, K.-U. Capacities of membrane lipids to accumulate neutral organic chemicals. *Environmental science & technology* **2011**, *45*, 5912–5921.

[14] Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. Understanding biophysicochemical interactions at the nano–bio interface. *Nature materials* **2009**, *8*, 543–557.

[15] 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*, 3965–3967.

[16] Rossi, G.; Monticelli, L. Gold nanoparticles in model biological membranes: A computational perspective. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2016**, *1858*, 2380–2389.

[17] Kim, S. T.; Saha, K.; Kim, C.; Rotello, V. M. The role of surface functionality in determining nanoparticle cytotoxicity. *Accounts of chemical research* **2013**, *46*, 681–691.

[18] Sengani, M.; Grumezescu, A. M.; Rajeswari, V. D. Recent trends and methodologies in gold nanoparticle synthesis–A prospective review on drug delivery aspect. *OpenNano* **2017**, *2*, 37–46.

[19] You, C.-C.; De, M.; Rotello, V. M. Monolayer-protected nanoparticle–protein interactions. *Current opinion in chemical biology* **2005**, *9*, 639–646.

[20] Troiano, J. M.; Olenick, L. L.; Kuech, T. R.; Melby, E. S.; Hu, D.; Lohse, S. E.; Mensch, A. C.; Dogangun, M.; Vartanian, A. M.; Torelli, M. D.; et al.. Direct probes of 4 nm diameter gold nanoparticles interacting with supported lipid bilayers. *The Journal of Physical Chemistry C* **2015**, *119*, 534–546.

[21] Melby, E. S.; Lohse, S. E.; Park, J. E.; Vartanian, A. M.; Putans, R. A.; Abbott, H. B.; Hamers, R. J.; Murphy, C. J.; Pedersen, J. A. Cascading effects of nanoparticle coatings: Surface functionalization dictates the assemblage of complexed proteins and subsequent interaction with model cell membranes. *ACS nano* **2017**, *11*, 5489–5499.

[22] Reviakine, I.; Johannsmann, D.; Richter, R. P.; *Hearing what you cannot see and visualizing what you hear: interpreting quartz crystal microbalance data from solvated interfaces*; 2011.

[23] Yousefi, N.; Tufenkji, N. Probing the interaction between nanoparticles and lipid membranes by quartz crystal microbalance with dissipation monitoring. *Frontiers in chemistry* **2016**, *4*, 46.

[24] McCubbin, G. A.; Praporski, S.; Piantavigna, S.; Knappe, D.; Hoffmann, R.; Bowie, J. H.; Separovic, F.; Martin, L. L. QCM-D fingerprinting of membrane-active peptides. *European Biophysics Journal* **2011**, *40*, 437–446.

[25] Zhang, M.; Akbulut, M. Adsorption, desorption, and removal of polymeric nanomedicine on and from cellulose surfaces: effect of size. *Langmuir* **2011**, *27*, 12550–12559.

[26] Zhang, M.; Soto-Rodríguez, J.; Chen, I.-C.; Akbulut, M. Adsorption and removal dynamics of polymeric micellar nanocarriers loaded with a therapeutic agent on silica surfaces. *Soft Matter* **2013**, *9*, 10155–10164.

[27] Heikkila, E.; Martinez-Seara, H.; Gurtovenko, A. A.; Javanainen, M.; Häkkinen, H.; Vattulainen, I.; Akola, J. Cationic Au nanoparticle binding with plasma membrane-like lipid bilayers: potential mechanism for spontaneous permeation to cells revealed by atomistic simulations. *The Journal of Physical Chemistry C* **2014**, *118*, 11131–11141.

[28] Lolicato, F.; Joly, L.; Martinez-Seara, H.; Fragneto, G.; Scoppola, E.; Baldelli Bombelli, F.; Vattulainen, I.; Akola, J.; Maccarini, M. The role of temperature and lipid charge on intake/uptake of cationic gold nanoparticles into lipid bilayers. *Small* **2019**, *15*, 1805046.

[29] Nakamura, H.; Sezawa, K.; Hata, M.; Ohsaki, S.; Watano, S. Direct translocation of nanoparticles across a model cell membrane by nanoparticle-induced local enhancement of membrane potential. *Physical Chemistry Chemical Physics* **2019**, *21*, 18830–18838.

[30] Lin, J.; Zhang, H.; Chen, Z.; Zheng, Y. Penetration of lipid membranes by gold nanoparticles: insights into cellular uptake, cytotoxicity, and their relationship. *ACS nano* **2010**, *4*, 5421–5429.

[31] 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 letters* **2013**, *13*, 4060–4067.

[32] Simonelli, F.; Bochicchio, D.; Ferrando, R.; Rossi, G. Monolayer-protected anionic Au nanoparticles walk into lipid membranes step by step. *The Journal of Physical Chemistry Letters* **2015**, *6*, 3175–3179.

[33] Miranda, O. R.; Chen, H.-T.; You, C.-C.; Mortenson, D. E.; Yang, X.-C.; Bunz, U. H.; Rotello, V. M. Enzyme-amplified array sensing of proteins in solution and in biofluids. *Journal of the American Chemical Society* **2010**, *132*, 5285–5289.

[34] Kanaras, A. G.; Kamounah, F. S.; Schaumburg, K.; Kiely, C. J.; Brust, M. Thioalkylated tetraethylene glycol: a new ligand for water soluble monolayer protected gold clusters. *Chemical communications* **2002**, 2294–2295.

[35] Brust, M.; Walker, M.; Bethell, D.; Schiffrin, D. J.; Whyman, R. Synthesis of thiol-derivatised gold nanoparticles in a two-phase liquid–liquid system. *Journal of the Chemical Society, Chemical Communications* **1994**, 801–802.

[36] Haiss, W.; Thanh, N. T.; Aveyard, J.; Fernig, D. G. Determination of size and concentration of gold nanoparticles from UV- Vis spectra. *Analytical chemistry* **2007**, *79*, 4215–4221.

[37] Ulrich, S.; Brown, T.; Watanabe, N.; Bronner, G.; Abraham, M.; Goss, K. NE UFZ-LSER database v 3.2. Leipzig, Deutschland, Helmholtz Zentrum für Umweltforschung-UFZ. **2017**.

[38] Cho, N.-J.; Frank, C. W.; Kasemo, B.; Höök, F. Quartz crystal microbalance with dissipation monitoring of supported lipid bilayers on various substrates. *Nature protocols* **2010**, *5*, 1096–1106.

[39] Richter, R. P.; Bérat, R.; Brisson, A. R. Formation of solid-supported lipid bilayers: an integrated view. *Langmuir* **2006**, *22*, 3497–3505.

[40] Kankare, J. Sauerbrey equation of quartz crystal microbalance in liquid medium. *Langmuir* **2002**, *18*, 7092–7094.

[41] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[42] Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.

[43] Heinz, H.; Vaia, R.; Farmer, B.; Naik, R. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12- 6 and 9-6 Lennard-Jones potentials. *The Journal of Physical Chemistry C* **2008**, *112*, 17281–17290.

[44] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[45] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[46] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[47] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

[48] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of computational chemistry* **2008**, *29*, 1859–1865.

[49] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional free-energy calculations using the weighted

histogram analysis method. *Journal of Computational Chemistry* **1995**, *16*, 1339–1350.

[50] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185*, 604–613.

[51] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[52] Melby, E. S.; Mensch, A. C.; Lohse, S. E.; Hu, D.; Orr, G.; Murphy, C. J.; Hamers, R. J.; Pedersen, J. A. Formation of supported lipid bilayers containing phase-segregated domains and their interaction with gold nanoparticles. *Environmental Science: Nano* **2016**, *3*, 45–55.

[53] Neale, C.; Bennett, W. D.; Tieleman, D. P.; Pomes, R. Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers. *Journal of Chemical Theory and Computation* **2011**, *7*, 4175–4188.

[54] Filipe, H. A.; Moreno, M. J.; Róg, T.; Vattulainen, I.; Loura, L. M. How to tackle the issues in free energy simulations of long amphiphiles interacting with lipid membranes: convergence and local membrane deformations. *The Journal of Physical Chemistry B* **2014**, *118*, 3572–3581.

[55] Rogers, J. R.; Geissler, P. L. Breakage of Hydrophobic Contacts Limits the Rate of Passive Lipid Exchange Between Membranes. *The Journal of Physical Chemistry B* **2020**, *124*, 5884–5898.

[56] Van Lehn, R. C.; Ricci, M.; Silva, P. H.; Andreozzi, P.; Reguera, J.; Voïtchovsky, K.; Stellacci, F.; Alexander-Katz, A. Lipid tail protrusions mediate the insertion of nanoparticles into model cell membranes. *Nature communications* **2014**, *5*, 1–11.

[57] 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*, 3165–3175.

[58] Van Tassel, P. R.; Viot, P.; Tarjus, G. A kinetic model of partially reversible protein adsorption. *The Journal of chemical physics* **1997**, *106*, 761–770.

10 THE INTERPLAY OF LIGAND PROPERTIES AND CORE SIZE DICTATES THE HYDROPHOBICITY OF MONOLAYER-PROTECTED GOLD NANOPARTICLES

The previous chapter found that ligand lipophilicity/hydrophobicity correlates with gold nanoparticle uptake onto lipid bilayers. This suggests that quantifying the extent of hydrophobicity of gold nanoparticles is informative towards predicting their behavior with other biomolecules. This chapter tackles the following questions:

- How do we quantify hydrophobicity at the nanoscale and how do we apply it to nanoparticle systems?
- How does modulation of the gold core and ligand selection affect interfacial hydrophobicity?
- What information could we gain from quantifying the hydrophobicity of gold nanoparticle systems and how does that relate to interactions with other molecules?

In this chapter, we quantify nanoparticle hydrophobicity by using atomistic molecular dynamics simulations to calculate local hydration free energies at the nanoparticle-water interface. The simulations reveal that the hydrophobicity of large gold nanoparticles is determined primarily

This chapter was reproduced from Chew, A. K.; Dallin, B. C.; Van Lehn, R. C. The Interplay of Ligand Properties and Core Size Dictates the Hydrophobicity of Monolayer-Protected Gold Nanoparticles. *ACS Nano* **2021**, *15*, 4534–4545.[1] Copyright 2021 American Chemical Society. The supporting information is cited as Ref. 2. B. C. Dallin performed the indirect umbrella sampling simulations for this chapter.

by ligand end group chemistry, as expected. However, for small gold nanoparticles, long alkanethiol ligands interact to form anisotropic bundles that lead to substantial spatial variations in hydrophobicity even for homogeneous monolayer compositions. We further show that nanoparticle hydrophobicity is modulated by changing the ligand structure, ligand chemistry, and gold core size, emphasizing that single-ligand properties alone are insufficient to characterize hydrophobicity. Finally, we illustrate that hydration free energy measurements correlate with the preferential binding of propane as a representative hydrophobic probe molecule. Together, these results show that both physical and chemical properties influence the hydrophobicity of small nanoparticles and must be considered together when predicting gold nanoparticle interactions with biomolecules.

10.1 Introduction

Gold nanoparticles (GNPs) protected by self-assembled monolayers (SAMs) have attracted significant interest for biomedical,[3] pollutant-detection,[4] and antimicrobial[5] applications because of their ease of fabrication, detectability, and tunable properties.[6,7] A SAM comprises multiple organic ligands, each typically consisting of a sulfur head group, long carbon chain backbone, and terminal end group, that exothermically adsorb onto GNPs via a favorable gold-sulfur interaction, resulting in a stable, robust mono-

layer.[8,9] The physical and chemical properties of SAMs can be synthetically tuned by changing the ligand chain length, structure (*e.g.* branched versus linear), and end group chemistry (*e.g.* methyl versus alcohol group), providing a powerful and versatile toolset to influence GNP surface properties. The GNP core size is another critical parameter that can be synthetically tuned. For biological applications, small GNPs (< 10 nm in diameter) are of interest because they have been shown to avoid removal from the body through renal clearance[10] and are approximately the same size as proteins (*e.g.*, hemoglobin is about 5 nm in diameter).[11] For small GNPs, however, the free volume available to ligands in the SAM can lead to variations to GNP surface properties that can be difficult to anticipate.[12–15] There is thus a need to predict how the interplay of ligand properties and core size impact GNP surface properties to avoid time-consuming experimental GNP synthesis and characterization.

One important GNP property is hydrophobicity, which characterizes the thermodynamic affinity of a GNP for water. Hydrophobicity plays an important role in GNP behavior, such as the self-assembly of GNPs to form aggregates,[16–18] the adsorption of hydrophobic proteins onto the GNP surface,[19] and the insertion of GNPs into the hydrophobic core of cell membranes.[14,20] Unfortunately, quantifying hydrophobicity is challenging because it emerges from collective water-water and ligand-water interactions at the GNP-water interface. At the macroscale, hydrophobicity is typically quantified by the water droplet contact angle on a flat surface,[21–23]

but contact angles cannot be computed for single GNPs. Alternatively, hydrophobic forces between two planar SAMs can be directly measured using atomic force microscopy,[24–27] but extending these measurements to GNPs is again challenging due their non-planar geometries. Experimental surface free energy characterization techniques have been developed to capture the hydrophobicity of nanoparticles by measuring their colloidal stability within a probe liquid, but this technique requires careful selection of the probe liquid.[28] Dye adsorption methods that introduce a hydrophobic or hydrophilic probe and measure the changes to adsorption in the presence of nanoparticles are also limited by probe selection and cannot distinguish between hydrophobic or electrostatic effects since these probes are charged.[29] These limitations suggest that alternative characterization techniques are necessary to quantify GNP hydrophobicity.

As an alternative to experimental measurements, previous authors have quantified GNP hydrophobicity by computing the octanol-water partition coefficient (logP) for ligand end groups.[17] Values of logP for ligand end groups have been found to correlate with the immune response activity,[30] cell uptake,[17] selective binding with protein isoforms,[31] and antimicrobial efficacy of small GNPs.[5] While these examples show the value of rationally tuning GNP hydrophobicity, single-ligand descriptors cannot account for nonadditive contributions to GNP surface properties that may emerge from the organization of ligands on the surface,[32] cooperative interactions between ligands,[32–36] or the curvature of the GNP

surface.[13,37] For example, cooperative ligand interactions drive the formation of bundles[36,38–40] in which ligands are oriented in the same direction as a result of ligand-ligand interactions. These bundles could affect GNP behavior, such as the aggregation of GNPs that align to minimize exposure of hydrophobic regions.[34,35] Furthermore, the specific spatial organization of ligands has been shown to lead to nonadditive contributions to GNP interfacial energies which cannot be accounted for using single-ligand parameters.[32,41] Recent computational work has instead utilized "virtual" GNPs, in which ligands are oriented radially and placed randomly around a gold core, to calculate the solvent accessible surface area and predict logP measurements.[42] However, virtual GNPs are static models that do not account for the interplay of ligand-water and ligand-ligand interactions that influence hydrophobicity.

To analyze the nanoscale fluctuations of ligands and water molecules that contribute to interfacial hydrophobicity, atomistic molecular dynamics (MD) simulations can be used to model the properties of functionalized GNPs.[38–40,42–49] To quantify hydrophobicity, recent MD simulations have calculated interfacial water density fluctuations.[21,41,50–61] Increased water density fluctuations lead to a higher probability of dewetting for cavities placed near an interface and indicate that the surface is more hydrophobic. The probability of dewetting can be thermodynamically related to the hydration free energy, which has been shown to correlate with macroscopic contact angle[21] and atomic force microscopy measurements.[24,60]

Calculating interfacial hydration free energies thus permits the characterization of hydrophobicity for interfaces of arbitrary geometry and has been successfully applied to study idealized surfaces,[57] planar SAMs, carbon nanotubes, and proteins,[41] but has yet to be applied to understand factors influencing the hydrophobicity of SAM-protected GNPs.

In this chapter, we systematically investigate the effect of GNP curvature and ligand properties on the hydrophobicity of small 2 nm and 6 nm GNPs as well as planar gold SAMs that represent larger GNPs. We perform unbiased MD simulations to quantify local hydration free energies near planar SAMs and find that these calculations are correlated with results from indirect umbrella sampling (INDUS) simulations that capture experimental trends.[24] By quantifying the hydration free energies of GNPs, we show that the surface curvature, ligand disorder, GNP size, and ligand conformation all influence GNP hydrophobicity, resulting from an interplay of ligand-ligand and ligand-water interactions that cannot be captured by single-ligand descriptors (*e.g.* logP). Lastly, we relate the local hydration free energies of GNPs to the competitive binding between water and propane molecules to illustrate how hydration free energies predict the preferential binding of hydrophobic moieties. These results suggest that quantifying GNP hydrophobicity at the nanoscale can be used to engineer GNP surface properties towards selective interactions with biomolecules, such as proteins or lipid bilayers.

10.2 Methods

10.2.1 Planar SAM simulations

Figure 10.1b shows the simulation system for planar SAMs with R = CH_3 ligands. A planar gold surface was constructed from a face-centered cubic lattice and two SAMs were grafted on opposite sides of the gold layer. The gold and sulfur atoms were separated by a 0.327 nm vertical distance. 200 total ligands were initiated in an all-trans configuration along the vector perpendicular to the gold surface. A 5 nm water layer was added above each SAM. All gold and sulfur atoms were restrained with a spring constant of 500,000 kJ/(mol nm^2). Unless otherwise stated, all simulations were performed at a temperature T = 300 K, controlled by the velocity rescale thermostat, and P = 1 bar (if *NPT* ensemble), controlled by either Berendsen or Parrinello-Rahman barostat. The system was then energy minimized, equilibrated for 6 ns in the *NPT* ensemble with the pressure controlled by the Berendsen barostat, then equilibrated for 5 ns in the same ensemble with the pressure controlled by an anisotropic Parrinello-Rahman barostat that only allowed changes to the z-dimension of the simulation box (see Supporting Information, Figure S1).[2] Two 5 ns restrained *NPT* simulations were then performed with the pressure controlled by the Berendsen barostat for the first 5 ns and Parrinello-Rahman barostat for the next 5 ns. The last configuration from this workflow was extracted, a 5 nm buffering vacuum layer[62] was added above each of the

SAMs (illustrated in Figure 10.1b), and the resulting configuration was used to initiate a 50 ns restrained *NVT* production simulation. For restrained simulations, all ligand heavy atoms (except sulfur atoms) were restrained with a weaker spring constant of 50 kJ/(mol nm^2). The addition of ligand heavy atom restraints did not significantly impact local hydration free energies (see Supporting Information, Figure S4).[2] Configurations from all restrained 50 ns *NVT* production simulations were output every 1 ps.

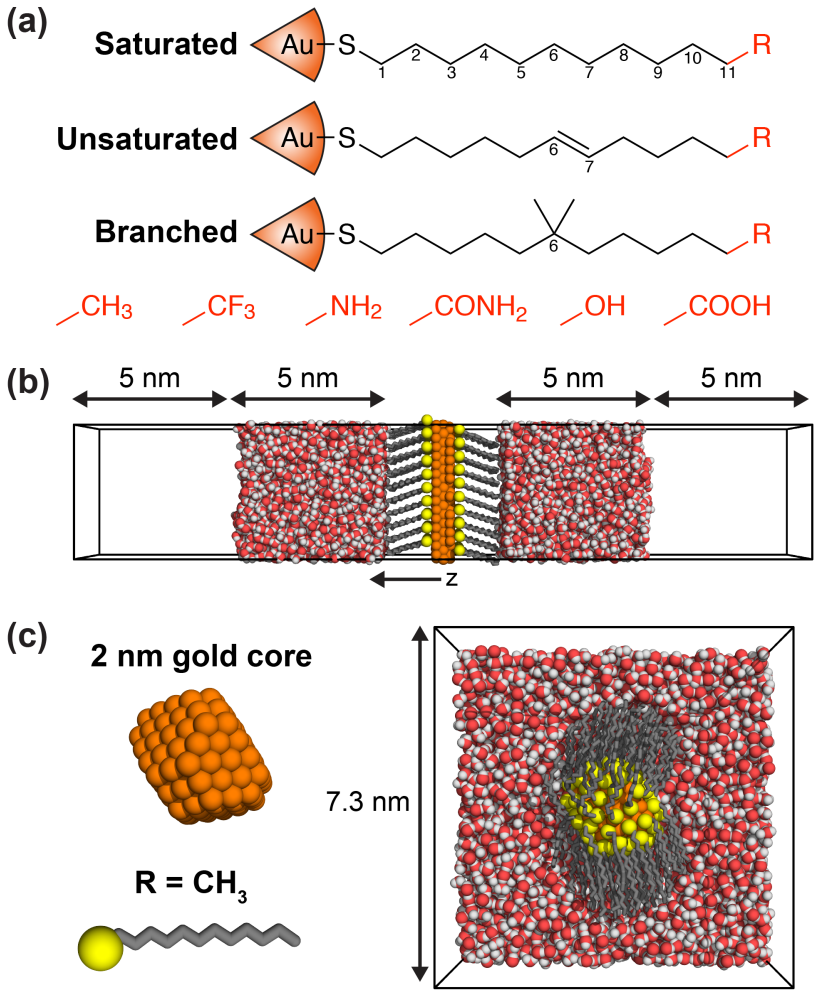


Figure 10.1: Overview of monolayer-coated gold nanoparticles and planar gold surfaces. (a) Alkanethiol ligand chemical structures studied in this work, including ligands with saturated, unsaturated, and branched non-polar backbones and six different end groups (indicated by R) of varying polarity. (b) Snapshot of the double planar self-assembled monolayer (SAM) simulations for saturated R = CH_3 ligands (i.e. dodecanethiol). (c) Snapshot of a 2 nm gold core coated with the same ligands as in (b) and snapshot of the equilibrated nanoparticle in water.

10.2.2 GNP simulations

Faceted GNPs with 2 and 6 nm gold core diameters were constructed as described in Ref. 36. A 5 ns *NPT* equilibration simulation was performed with the pressure controlled by the Berendsen barostat, then a 50 ns *NPT* production simulation was performed using the Parrinello-Rahman barostat. The last 40 ns of the production simulation were used to select the most representative configurations of the GNP as described in the Supporting Information.[2] Two 2 ns restrained *NPT* simulations were then performed using the most representative GNP configurations with the pressure controlled by the Berendsen barostat for the first 2 ns and the Parrinello-Rahman barostat for the next 2 ns. The last configuration was used to initiate restrained *NVT* production simulations for 50 ns (see Figure 10.1c). Gold atoms and sulfur atoms were restrained with a spring constant of 500,000 kJ/(mol nm^2), whereas ligand heavy atoms (except sulfur atoms) were restrained with a weaker spring constant of 50 kJ/(mol nm^2). Restraints on the ligand heavy atoms were selected to maintain ligand fluctuations and a well-defined surface (see Supporting Information, Figure S3).[2] The weaker spring constant is commensurate with previous simulations that map ligand binding sites in proteins.[63] In addition, previous simulations found that spurious solvation preferences on the surface of fully restrained proteins,[64] necessitating the use of weak spring constants to minimize these effects.

10.2.3 Propane-water simulations

Hydrophobic sites around a GNP were mapped by simulating the GNP in an aqueous mixture with propane. The cubic simulation box was resized with 1 nm between the edge of the box and the GNP and solvated using Packmol[65] with 1 mol% propane (96 molecules) and 99 mol% water (9,602 molecules). To maintain a constant SAM-water interface, gold and ligand atoms were restrained using the same restraints as described for the restrained *NVT* simulations. The simulation protocol was as follows: (1) 2 ns *NPT* equilibration at P = 1 bar (controlled by a Berendsen barostat) and T = 300 K, (2) 4 ns *NVT* temperature annealing, and (3) 12 ns *NPT* production at P = 1 bar (controlled by the Parrinello-Rahman barostat) and T = 300 K. The *NVT* temperature annealing step was included to allow the solvent to explore the surface at a higher temperature and avoid any bias from the initial configurations. For the *NVT* temperature annealing step, all gold core and SAM ligand atoms were frozen, the temperature was ramped from 300 K to 600 K for 1 ns, held constant at 600 K for 1 ns, ramped down from 600 K to 300 K for 1 ns, and held constant at 300 K for 1 ns. The last 10 ns of the 12 ns *NPT* production simulation were used for analysis with configurations output every 10 ps. The simulation procedure was iterated five times with different initial solvent configurations (see Supporting Information, Figure S10 for convergence details).[2]

10.2.4 Simulation parameters

Ligands were parameterized with the CGenFF/CHARM36 forcefield[66–68] and gold atoms were parameterized with the INTERFACE force field.[69] Water was modeled using the TIP3P model.[70] In all simulations, Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential using a 1.2 nm cutoff that was smoothly shifted to zero between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the smooth particle mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and fourth order interpolation. Bonds were constrained using the LINCS algorithm.[71] Periodic boundary conditions were enabled in all directions. All classical MD simulations were performed using Gromacs 2016[72] using the leapfrog integrator with a 2-fs timestep.

10.3 Results and Discussion

10.3.1 Hydration free energies for planar SAMs predict experimental measurements

We first quantified the hydrophobicity of single-component planar SAMs containing saturated undecanethiol backbones and six end groups of varying polarity (listed in Figure 10.1a) because the simple geometry of planar SAMs facilitates comparisons between simulation-derived hydrophobicity

calculations and experimental measurements. Planar SAMs also represent the limit of large GNPs (>60 nm in diameter) with low surface curvatures,[73] enabling comparisons between the hydrophobicity of large GNPs and small (<10 nm diameter) GNPs with identical SAM compositions. R = COOH ligands were included to capture effects associated with highly polar end groups and were modeled as neutral (corresponding to acidic conditions) in this study to avoid the complexity associated with counterions. Figure 10.1b shows the planar SAM simulation systems consisting of two SAMs above and below a planar gold (111) surface with R = CH_3 ligands. A 5 nm water box was simulated above each SAM to capture SAM-water interactions and a 5 nm vacuum layer was included to provide a buffering boundary condition.[62] Figure 10.2a shows the SAM-water interface for R = CH_3 ligands as an example. For all systems, the SAM-water interface was defined as a constant water density isosurface using the Willard-Chandler method.[62] We selected a constant density of 26 molecules/nm^3, which is sufficiently large to avoid overlap between the grid points and the ligand heavy atoms for R = CH_3 ligands (see Supporting Information, Figure S5).[2]

We quantified SAM hydrophobicity by calculating the hydration free energy (μ_v) of a cavity placed at the SAM-water interface. μ_v is related to the probability of vacating all heavy atoms within the cavity, $p_v(N = 0)$, by Equation 10.1:[24,57]

$$\mu_v = -kT \ln p_v(N = 0) \tag{10.1}$$

N is the number of heavy atoms in the cavity, k is the Boltzmann constant, and T is the temperature. Lower μ_ν values indicate that the cavity is more likely to dewet and thus more hydrophobic. The subscript ν indicates a specific cavity because in general μ_ν depends on the cavity geometry, volume, and placement.

As a first approach, we computed hydration free energies for $2 \times 2 \times 0.3$ nm^3 cuboidal cavities (Figure 10.2a) placed at the SAM-water interface. We denote these free energies as μ_A to indicate the hydration free energy of a cavity spanning a large area. Values of μ_A have been shown to correlate with experimentally determined hydrophobic forces between planar SAMs, indicating that this approach captures macroscopic measurements of hydrophobicity.[24] The challenge in computing μ_A is accurately calculating $p_A(0)$ for large cavities, which have a low likelihood of spontaneously dewetting in unbiased MD simulations. We thus performed indirect umbrella sampling (INDUS) simulations in which a biasing potential is used to vacate water molecules from the cavity and the corresponding value of $p_A(0)$ (and μ_A) is obtained using the Weighted Histogram Analysis Method (see Supporting Information, Figure S6).[74]

As validation, Figure 10.2b compares values of μ_A to experimentally determined water contact angles[75,76] for single-component SAMs prepared on gold for the same set of ligand end groups as those listed in Figure 10.1a. Figure 10.2b shows that the selected end groups in Figure 10.1b vary substantially in their hydrophobicity as expected. CH_3 and CF_3 end

groups are more hydrophobic (smaller values of μ_A) and the other end groups are less hydrophobic. Larger water contact angles indicate more hydrophobic surfaces;[21] accordingly, the positive correlation between μ_A and the contact angle indicates that the simulation results predict interfacial hydrophobicity in good agreement with macroscopic experiments, resulting in a high Pearson's r correlation coefficient of 0.96. μ_A also can distinguish between SAMs that are indistinguishable using the water contact angle; specifically the R = COOH surface has a larger value of μ_A than the R = OH and R = $CONH_2$ SAMs, reflecting a stronger affinity for water, even though these SAMs are all macroscopically wet and have similar contact angles. These results indicate that the simulations capture SAM hydrophobicity in good agreement with experiments, permitting comparison of planar SAMs and small GNPs to determine how changes in GNP geometry and ligand structure, together with end group chemistry, influence interfacial hydrophobicity.

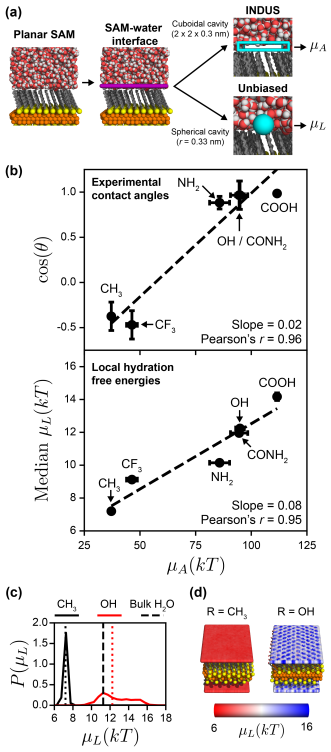


Figure 10.2: Quantifying the hydrophobicity of planar SAMs. (a) Snapshot showing the SAM-water interface (indicated as a purple surface) computed for a planar SAM with saturated R = CH_3 ligands. Two cavity geometries were used to calculate hydration free energies: a $2 \times 2 \times 0.3$ nm^3 cuboidal cavity was used to compute μ_A with INDUS simulations (illustrated at top) and a series of small spherical cavities with radii of 0.33 nm were used to compute μ_L from unbiased MD simulations (illustrated at bottom; only one cavity is shown for clarity). (b) Correlation between μ_A and experimental water contact angles ($\cos\theta$, top) and median values of μ_L (bottom) for ligands with different end group chemistries. Contact angle measurements were taken from Ref. 75,76 and tabulated in the Supporting Information, Table S2.[2] Both μ_L and μ_A were computed for top and bottom planar SAMs and treated as independent samples. Points show the average value of μ_L or μ_A and error bars report the standard deviation of μ_L or μ_A between the two samples. Best-fit lines are drawn with corresponding slopes and Pearson's r values labeled at bottom right. (c) Probability density function, $P(\mu_L)$, for R = CH_3 and R = OH ligands computed from unbiased MD simulations. μ_L for bulk water is shown as a black dashed line as a reference. Dotted lines show the median value of μ_L for each SAM. (d) Hydration free energy maps for the same planar SAMs in (b) with μ_L values ranging between 6 (red) to 16 kT (blue). Each point indicates the value of μ_L calculated using a spherical cavity centered on that point.

10.3.2 Local hydration free energies capture hydrophobicity trends in planar SAMs

We next extended the previous calculations from planar SAMs to small GNPs, which exhibit complex, non-planar geometries due to their faceted gold cores and the tendency of ligands to form bundles (Figure 10.1c). Performing INDUS simulations using cavities with irregular geometries is challenging; previous work has instead used INDUS to quantify the hydrophobicity of proteins by defining a series of cuboid cavities placed around the protein surface.[59] We developed a similar approach to map the hydrophobicity of GNPs by placing a series of small spherical cavities at the SAM-water interface and estimating hydration free energies in each cavity from a single unbiased simulation. We denote these free energies as μ_L to indicate the local hydration free energy at a particular region of the SAM-water interface. All spherical cavities had radii of 0.33 nm, which is approximately the van der Waals radius of a methane molecule and has been previously used to quantify on the interfacial hydrophobicity of planar SAMs.[21] An example spherical cavity is shown in Figure 10.2a. For cavities this size, $p_L(0)$ can be sampled in an unbiased simulation but can fail to converge near hydrophilic regions where the likelihood of dewetting is low. Therefore, for each cavity, we calculated $p_L(N)$ during a 50-ns unbiased simulation, fit $p_L(N)$ to a normal distribution because water density fluctuations are Gaussian,[21] then extrapolated to obtain

$p_L(0)$ and μ_L (see Supporting Information, Figure S8).[2] This method enables calculations of μ_L for cavities across the entire SAM-water interface from a single unbiased simulation, enabling the rapid characterization of interfacial hydrophobicities for a series of GNPs.

Figure 10.3b shows representative GNPs and hydration free energy maps for the different end group chemistries (Figure 10.1a). Since a long methylene chain is present in these ligands, even ligands terminated with polar end groups form bundles.[35] Hydration free energy maps for non-polar R = CH_3 ligands show that the poles, where the majority of methyl groups reside, are uniformly hydrophobic. However, the annular shell in between the poles is less hydrophobic because of the exposure of sulfur atoms to interfacial water molecules and because of the surface curvature, which has been shown introduce differences in hydrophobicity.[41,77] Hydration free energy maps show that highly polar-terminated ligands (*e.g.* $CONH_2$, OH, and COOH) have distinct regions of low hydrophobicity at the poles, where the terminal ends of the ligands are less hydrophobic and the exposed methylene groups are more hydrophobic. One striking feature is that the R = $CONH_2$, OH and COOH hydration free energy maps reveal a heterogeneous landscape with regions of high (blue regions) and low (red regions) hydrophobicity. These results show that homogenously patterned GNPs form bundles that are independent of the end group and result in spatially heterogenous domains of hydrophobicity, even for CH_3-terminated GNPs.

Figure 10.3c shows the probability density functions of μ_L for the 2 nm GNPs (top panel) and planar SAMs (bottom panel) for the different ligand end group chemistries listed in Figure 10.1a. Planar SAMs show broad μ_L distributions for polar ligands, which is due to the formation of checkerboard hydration free energy maps like those in Figure 10.2d (shown in the Supporting Information, Figure S9),[2] but with median values of μ_L that differ significantly in the order CH_3 < CF_3 < NH2 < $CONH_2$ < OH < COOH. For GNPs, the median values of μ_L follow the same order as for the planar SAMs with the exception of $CONH_2$, indicating that the end group does influence hydrophobicity as expected. In contrast to the planar SAMs, however, GNPs exhibit narrower μ_L distributions with a majority of μ_L values below that of bulk water (black dashed lines, Figure 10.3c) and with median values that are more similar for different polar end groups. The narrower μ_L distribution of GNPs compared to planar SAMs is due to the excess free volume in GNPs that allows solvent molecules to contact the nonpolar methylene backbones of the ligands. As a result, large hydrophobic regions are found for all ligand end group chemistries on GNPs, even for the polar-terminated ligands. These hydrophobic regions around the exposed methylene groups are consistent with the orientation of multiple aggregated GNPs that preferentially adsorb to each other to minimize the extent of hydrophobic regions exposed to the solvent.[35] Previous studies have used the logP of the end groups as a metric of GNP hydrophobicity;[5,30,31] these results show that while end group chemistries

do influence GNP hydrophobicity, the hydrophobicity of small GNPs differs substantially from the hydrophobicity of equivalent planar SAMs because the cooperative interactions between ligands in the monolayer result in spatially heterogenous interfacial properties.

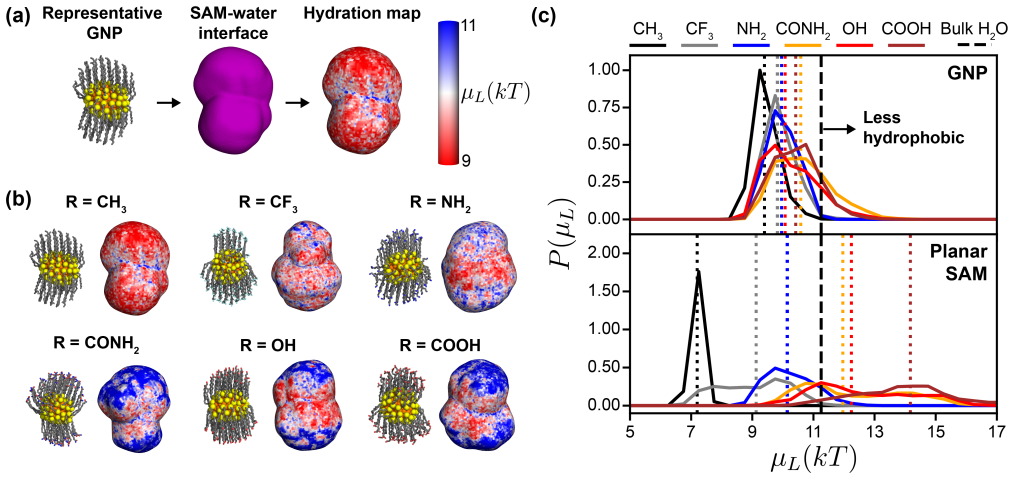


Figure 10.3: Hydrophobicity of 2 nm diameter SAM-protected GNPs with different end group chemistries. (a) Snapshots showing a representative 2 nm diameter GNP coated with R = CH_3 ligands. The SAM-water interface (purple) was computed using the same procedure as the planar SAMs (Figure 10.2a). The hydration free energy map shows μ_L values between 9 (red) to 11 kT (blue) as in Figure 10.2. (b) Simulation snapshots and corresponding hydration free energy maps for 2 nm GNPs with six end group chemistries, following the same color scheme as in (a). (c) Probability density functions, $P(\mu_L)$, showing the distribution of μ_L for GNPs with the six end group chemistries in (b). $P(\mu_L)$ is also shown for planar SAMs in the bottom panel for comparison. Dotted lines show the median value of each μ_L distribution and are color-coded to match the different ligands. The value of μ_L for bulk water is shown as a black dashed line for reference.

10.3.3 Ligand order in polar-terminated ligands results in broad hydration free energy distributions

We further investigated what factors influenced the different μ_L distributions found for the planar SAMs and 2 nm GNPs. We hypothesized that the broad distributions found for the homogeneous planar SAMs could be due to the ordering of the long ligand SAMs, as suggested by the hydration free energy map in Figure 10.1c. Previous experiments and simulations have shown that decreasing the number of methylene groups in the ligand backbone (n) decreases monolayer order, which introduces a rougher surface that affects interfacial hydrophobicity.[24,27] We thus modeled alkanethiol ligands with only 3 methylene groups to determine the effect of order on the μ_L distributions. Figure 10.4a displays the side-view of the monolayer structure for $n = 3$ and $n = 11$ OH-terminated ligands and shows that longer chain lengths are more ordered than the shorter chain lengths. Figure 10.4a shows hydration free energy maps for $n = 3$ and $n = 11$ OH-terminated ligands and reveals that decreasing n results in more hydrophobic regions and a less heterogenous hydration free energy map. This change in hydrophobicity is apparent in Figure 10.4b, which shows the probability density function of μ_L for OH-terminated ligands on a 2 nm diameter gold core ($n = 11$) and on planar gold ($n = 3$ and $n = 11$). Decreasing the ligand length for planar SAMs results in a narrower μ_L distribution that is similar to those found for the GNPs.

These findings show that highly ordered SAMs like $n = 11$ give rise to surface heterogeneity that causes the broadening of the hydration free energy distributions, whereas more disordered SAMs like $n = 3$ and the SAM-coated GNPs appear more homogenous with narrower hydration free energy distributions.

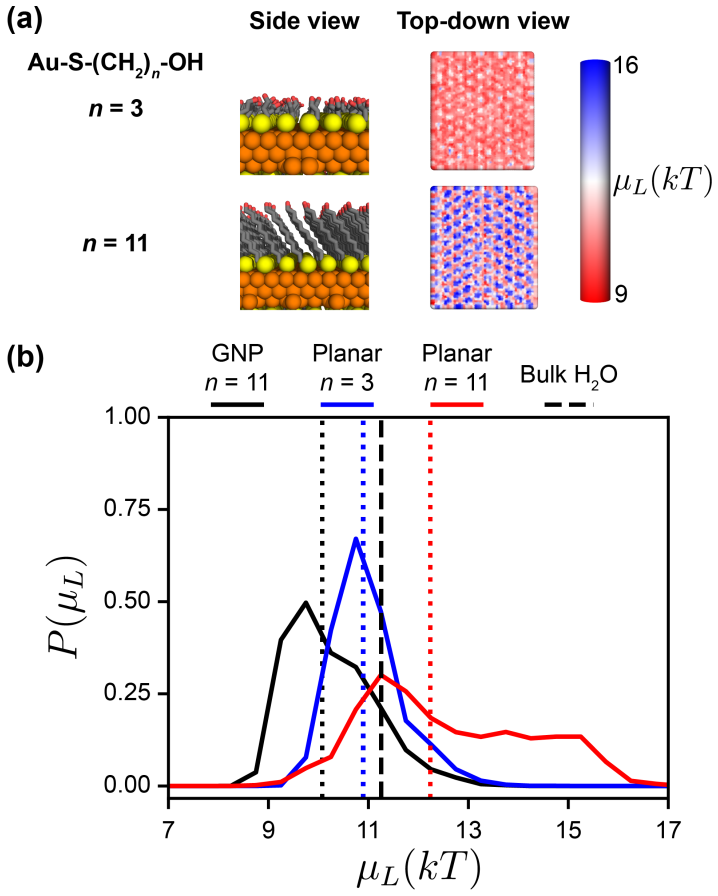


Figure 10.4: Effect of ligand length on the hydrophobicity of planar SAMs. (a) Snapshots of OH-terminated ligands on planar gold with $n = 3$ and $n = 11$ methylene groups. Hydration free energy maps are visualized from a top-down view with μ_L values ranging from 9 (red) to 16 kT (blue). (b) Probability density functions, $P(\mu_L)$, showing the distribution of μ_L for OH-terminated ligands on a 2 nm diameter gold core ($n = 11$, black) and planar gold ($n = 3$, blue and $n = 11$, red). Dotted lines show the median value of μ_L for each distribution and are color-coded to match the different ligand chain lengths and gold core types. The value of μ_L for bulk water is shown as a black dashed line for reference.

10.3.4 Inducing surface curvature by tuning GNP size affects hydration free energies

Given that GNP hydrophobicity depends on ligand order due to variations in SAM ligand lengths (Figure 10.4b), we also expected the GNP core diameter to affect hydrophobicity because smaller GNP core diameters result in a higher excess free volume per ligand and decreased ligand order.[36] The 2.2 kT shift in median μ_L for the CH_3-terminated planar SAM and GNP shown in Figure 10.3c also suggests surface curvature plays a significant role in determining GNP hydrophobicity. To investigate the effect of GNP size, we modeled a 6 nm diameter gold core GNP (shown in Figure 10.5a), which is intermediate in size between the 2 nm diameter gold core GNP and the planar gold surface. Figure 10.5a shows representative 6 nm GNPs with CH_3- and OH-terminated ligands and corresponding hydration free energy maps. The hydration free energy maps for CH_3-terminated ligands show primarily hydrophobic regions with less hydrophobic regions closer to the gold core and near the methylene groups of the ligand backbones, similar to the findings for 2 nm diameter GNPs (Figure 10.3b). Conversely, OH-terminated ligands on 6 nm diameter GNP shows less hydrophobic regions where the end groups are bundled and more hydrophobic regions around the methylene groups where the ligand backbones are exposed.

Figure 10.5b shows the probability density functions of μ_L for CH_3- and OH-terminated ligands on a 2 nm diameter gold core, 6 nm diameter gold

core, and a planar gold surface. For CH_3-terminated ligands, decreasing the gold core diameter results in slightly less hydrophobic surfaces, with minor differences in the μ_L distribution between 2 and 6 nm diameter gold cores. Conversely, for OH-terminated ligands, decreasing the gold core diameter results in more hydrophobic surfaces that arises from increased exposure of methylene groups around the GNP and reduced facet size where the ligands tend to form bundles. Together, these findings show that decreasing the gold core size influences hydrophobicity due to the exposure of methylene backbones and surface curvature, which can be attributed to the increased free volume present in the monolayer.[24,41]

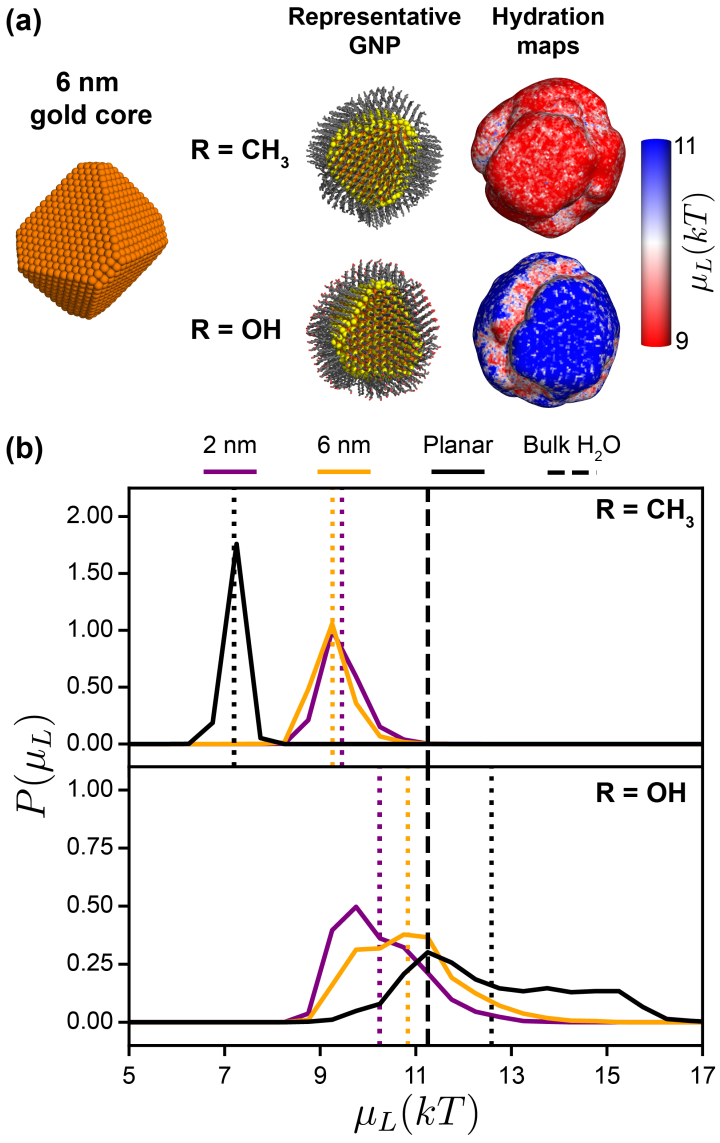


Figure 10.5: Effect of core size on gold nanoparticle hydrophobicity. (a) Example image of a 6 nm diameter gold core. Representative 6 nm GNPs and hydration free energy maps for R = CH_3 ligands and R = OH ligands. Hydration free energy maps show μ_L values between 9 (red) and 11 kT (blue). (b) Probability density functions, $P(\mu_L)$, for the 2 and 6 nm diameter gold cores and planar SAMs for R = CH_3 ligands (top) and R = OH ligands (bottom). Dotted lines denote the median μ_L value for each distribution and are color-coded to match the different gold core sizes. The value of μ_L for bulk water is shown as a black dashed line for reference.

10.3.5 Adding unsaturated or branched groups results in spatially distributed clusters

The preceding sections showed that long alkanethiol ligands form bundles that result in spatial variations in hydrophobicity (Figure 10.3b). We hypothesized that introducing surface disorder by adding an unsaturated bond or branched methylene groups to a ligand backbone would disrupt the formation of bundles to influence hydrophobicity without varying the end group chemistry (Figure 10.1a). The addition of unsaturated bonds to the ligand backbone has been shown to decrease surface order and reduce hydrophobicity for planar SAM systems.[24,27] Figure 10.6a shows representative 2 nm diameter GNP with saturated, unsaturated, or branched OH-terminated ligands. The representative GNP images show bundles for saturated ligands, whereas no distinct bundles are found for GNPs with unsaturated and branched ligands. The same behavior was observed for all six ligand end groups (Figure S7).[2] As a result, the hydration free energy maps for unsaturated or branched ligands are more spherical and the less hydrophobic regions are more dispersed compared to those of the saturated ligands.

Figure 10.6b shows the probability density functions of μ_L for 2 nm diameter GNPs with saturated, unsaturated, or branched OH-terminated ligands. Inclusion of unsaturated or branched ligands shifts the μ_L distributions to larger values slightly compared to the saturated ligands,

indicating that disordered, nonbundled surfaces are less hydrophobic. Figure 10.6c shows the median μ_L values for different end groups with either saturated, unsaturated, or branched ligands. Overall, the median μ_L values do not significantly change when introducing ligand disorder, indicating that the overall surface hydrophobicity is not affected by ligand conformation. However, the hydration free energy maps in Figure 10.6a indicate that the spatial heterogeneity of less hydrophobic regions is lost when introducing either unsaturated or branched ligands; the less hydrophobic regions are spatially distributed across the surface for unsaturated or branched ligands. To quantify this effect, we used the DBSCAN clustering algorithm[78] to calculate the number of regions with μ_L greater than bulk water ($\mu_L \geqslant 11.25$ kT), which we defined as hydrophilic regions (see Supporting Information for clustering parameters). Figure 10.6a shows the different hydrophilic regions identified for the three GNPs described above with each region colored uniquely. Unsaturated and branched ligands have hydrophilic regions that are distributed uniformly around the GNP, whereas saturated ligands have two distinct hydrophilic regions nearby the OH groups. Figure 10.6d shows the number of hydrophilic clusters for the different end groups and ligand conformations. Ligands terminated with R = $CONH_2$, OH, or COOH show increased number of clusters for unsaturated and branched ligands compared to the saturated ligands. These results indicate that adding an unsaturated bond or branching methylene groups disrupt the bundles formed by long

alkanethiol ligands and result in hydrophilic regions that are distributed around the GNP, minimizing the spatial heterogeneity found for GNPs with saturated alkanethiol ligands.

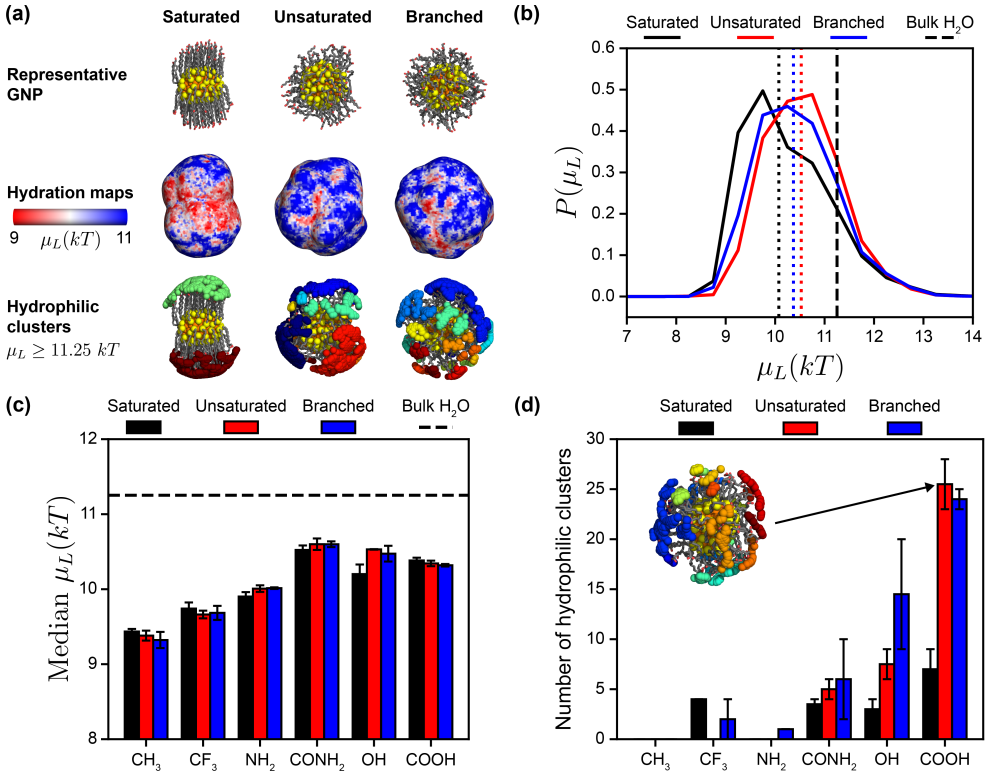


Figure 10.6: Effect of unsaturated and branched ligands on gold nanoparticle hydrophobicity. (a) Representative 2 nm diameter GNP configurations coated with either saturated, unsaturated, or branched OH-terminated ligands. Hydration free energy maps show μ_L values between 9 (red) and 11 kT (blue) with the same color bar shown in Figure 10.3a. Hydrophilic clusters corresponding to regions where μ_L values are greater than 11.25 kT (the value of μ_L for bulk water) are indicated in unique colors. (b) Probability density function, $P(\mu_L)$, for 2 nm diameter GNPs with either saturated, unsaturated, or branched OH-terminated ligands. Dotted lines show the median value of μ_L for each distribution and are color-coded to match the different ligand structures. The value of μ_L for bulk water is shown as a black dashed line for reference. (c) Median μ_L values for saturated, unsaturated, or branched ligands with different R groups. (d) Number of hydrophilic clusters for saturated, unsaturated, or branched ligands with different R groups. The inset snapshot shows the hydrophilic clusters for unsaturated COOH-terminated ligands. The reported values and error bars in (c) and (d) are the average and standard deviation of values computed for the most and least representative GNPs (Supporting Information, Figure S2).[2]

10.3.6 Relationship between hydration free energy maps and competitive binding of solvents

The hydration free energy maps quantify the differences in the thermodynamic affinity of water for different regions of the SAM-water interface. To illustrate how hydration free energies relate to physical binding processes, we quantified the competitive binding between hydrophilic (water) and hydrophobic (propane) small-molecule probes at the SAM-water interface. This approach was inspired by similar computational methods used to identify potential ligand-binding sites on proteins.[79] Figure 10.7a shows a snapshot of a representative 2 nm diameter GNP with saturated OH-terminated ligands in the presence of 1 mol% propane and 99 mol% water. Dilute amounts of propane were selected to prevent aggregation of hydrophobic probes in water.68 The extent of preferential binding on the SAM-water interface was measured by the fraction of simulation time that each spherical cavity used to compute μ_L was either occupied by at least one molecule of propane (f_P) or water (f_{H_2O}). f_P and f_{H_2O} range from 0 to 1, with a value of 1 indicating that at least one solvent molecule occupies the cavity for all MD configurations.

Figure 10.7b shows occupancy maps of propane and water for a representative 2 nm GNP with saturated OH-terminated ligands. The hydration free energy map for the same GNP is shown for comparison. The occupancy maps show that propane preferentially binds to the exposed

methylene groups and outcompetes water for these regions. The occupancy maps for propane are consistent with the hydrophobic areas found in the hydration free energy maps for $\mu_L \leqslant 9$ kT. The occupancy maps also show that water molecules prefer the OH end groups at the poles, consistent with the less hydrophobic regions on the hydration free energy map for $\mu_L \geqslant 11$ kT. Figure 10.7c shows the relationship between values of f_P and f_{H_2O} and values of μ_L by using Gaussian kernel density estimation to calculate the probability density function. f_P and f_{H_2O} are non-zero for all values of μ_L, indicating that propane and water compete for binding sites on the GNP. At approximately μ_L = 10 kT, f_P reaches 0 and f_{H_2O} reaches 1 with a maximum at μ_L = 11 kT, close to the value for bulk water (μ_L = 11.25 kT). These findings indicate that the subtle variations in μ_L encoded within the hydration free energy maps drive large differences in the preferential binding of model hydrophilic and hydrophobic probes. Occupancy maps for a 2 nm GNP with branched OH-terminated ligands show that the surface is primarily occupied by water molecules for all μ_L values (see Supporting Information, Figure S11),[2] indicating that introducing ligand disorder could prevent the adsorption of small hydrophobic molecules in agreement with the uniform distribution of hydrophilic regions observed in Figure 10.6a. Together, these results indicate that the competitive binding between solvents of varying hydrophobicity follows trends established by the hydration free energy maps, suggesting that variations in μ_L can provide insight into the binding of GNPs to other

biomolecules (*e.g.* proteins).

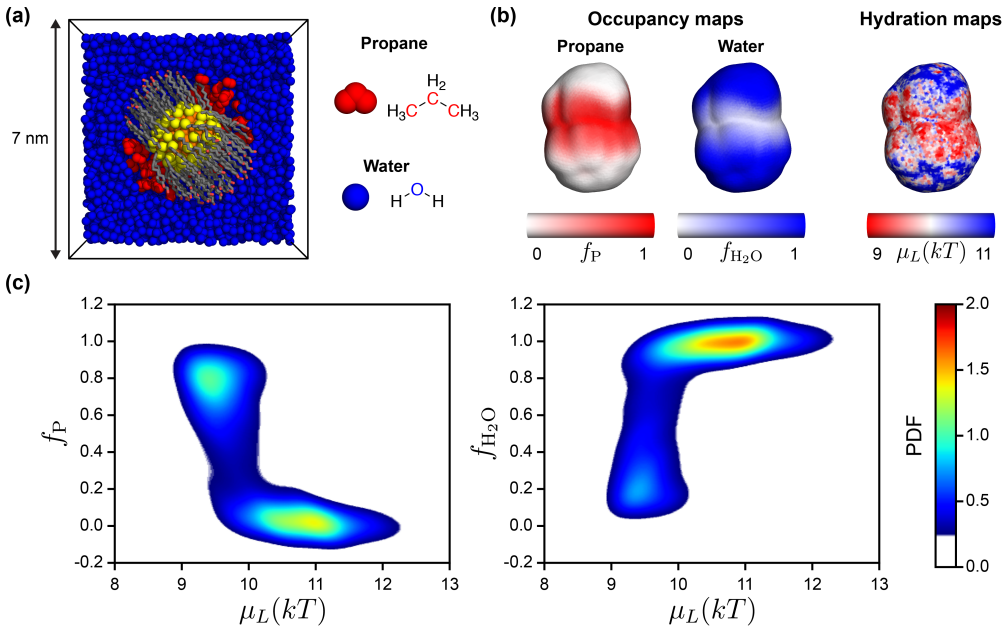


Figure 10.7: Relationship between hydration free energy maps and competitive binding of hydrophobic and hydrophilic probe molecules. (a) Simulation snapshot of representative 2 nm diameter GNP with OH-terminated ligands in the presence of 1 mol% propane (red spheres) and 99 mol% water (blue spheres). Hydrogen atoms are omitted for clarity. (b) Occupancy maps showing the fraction of simulation time in which spherical cavities (the same cavities used to compute μ_L) are either occupied by propane (f_P) or water (f_{H_2O}) for the GNP in (a). Hydration free energy maps are shown for μ_L values between 9 (red) to 11 kT (blue) as a comparison to occupancy map. (c) Relationship between f_P or f_{H_2O} to μ_L computed using Gaussian kernel-density estimation to calculate the probability density function; values between 0 to 0.25 were omitted for clarity.

10.4 Summary

In this chapter, we used atomistic molecular dynamics simulations to quantify the hydrophobicity of small (<10 nm) and large (planar, representing >60 nm diameters) SAM-protected GNPs by computing hydration free energies in cavities placed at the SAM-water interface. Ligand end group properties (*e.g.* logP) have been conventionally thought to dictate hydrophobicity; however, the simulations revealed that varying the ligand end group chemistry, gold core size, and ligand backbone structure collectively lead to spatially varying hydration free energies. We found broad distributions of these hydration free energies in polar-terminated planar SAMs due to the quasi-crystalline ordering of ligands, suggesting that flat, chemically homogeneous SAMs could exhibit spatial variations in hydrophobicity. For 2 nm gold nanoparticles, the excess free volume accessible to ligands decreases their order compared to ligands in planar SAMs and results in a narrower hydration free energy distribution arising from the exposure of methylene groups to water. Spatial variations in hydrophobicity are further amplified for these GNPs because ligands tend to form "bundles," giving rise to anisotropic structures despite homogeneous surface coatings. Similar spatial variations in hydrophobicity were also found for larger, 6 nm GNPs, indicating that small GNPs exhibit similar behavior for a range of core diameters. We further showed that introducing ligand disorder by adding either an unsaturated bond

or branched methylene group disrupts ligands bundles and results in a spatially homogenous surface. Hydration free energy calculations were further validated by observing the preferential binding of propane on the surface to illustrate how analysis of hydration free energies translate into interactions with other molecules.

Altogether, these results reveal that end group chemistry dictates hydrophobicity for large length scale systems like planar SAMs, but at small length scales, the emergence of non-uniform structures (*e.g.* bundles) dictates spatial heterogeneities in hydrophobicity for all ligands regardless of the end group. These spatially heterogenous properties cannot be captured by single-ligand properties and spontaneously occur without explicitly patterning the SAMs using mixed-SAM systems.[32] Similar spatially heterogeneous properties have been shown in simulations[24,60] and experiments[25,26] to drive large differences in the hydrophobic interactions between apposed SAMs and between SAMs and synthetic peptides, further highlighting the importance of quantifying the spatial variations in GNP hydrophobicity. These results thus provide guidelines for understanding how ligand chemical properties and GNP core size can cooperatively dictate hydrophobic interactions between GNPs and other biomolecules.

10.5 References

[1] Chew, A. K.; Dallin, B. C.; Van Lehn, R. C. The Interplay of Ligand Properties and Core Size Dictates the Hydrophobicity of Monolayer-

Protected Gold Nanoparticles. *ACS Nano* **2021**, *15*, 4534–4545.

[2] Chew, A. K.; Dallin, B. C.; Van Lehn, R. C. The Interplay of Ligand Properties and Core Size Dictates the Hydrophobicity of Monolayer-Protected Gold Nanoparticles [Supporting Information]. *ACS Nano* **2021**, *15*, 4534–4545.

[3] Duncan, B.; Kim, C.; Rotello, V. M. Gold nanoparticle platforms as drug and biomacromolecule delivery systems. *Journal of controlled release* **2010**, *148*, 122–127.

[4] Wang, C.; Yu, C. Detection of chemical pollutants in water using gold nanoparticles as sensors: a review. *Reviews in Analytical Chemistry* **2013**, *32*, 1–14.

[5] Li, X.; Robinson, S. M.; Gupta, A.; Saha, K.; Jiang, Z.; 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*, 10682–10686.

[6] Sengani, M.; Grumezescu, A. M.; Rajeswari, V. D. Recent trends and methodologies in gold nanoparticle synthesis–A prospective review on drug delivery aspect. *OpenNano* **2017**, *2*, 37–46.

[7] Jans, H.; Huo, Q. Gold nanoparticle-enabled biological and chemical detection and analysis. *Chemical Society Reviews* **2012**, *41*, 2849–2866.

[8] Häkkinen, H. The gold–sulfur interface at the nanoscale. *Nature chemistry* **2012**, *4*, 443.

[9] Saha, K.; Agasti, S. S.; Kim, C.; Li, X.; Rotello, V. M. Gold nanoparticles in chemical and biological sensing. *Chemical reviews* **2012**, *112*, 2739–2779.

[10] Longmire, M.; Choyke, P. L.; Kobayashi, H. Clearance properties of nano-sized particles and molecules as imaging agents: considerations and caveats. **2008**.

[11] Erickson, H. P. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biological procedures online* **2009**, *11*, 32–51.

[12] Jiang, Y.; Huo, S.; Mizuhara, T.; Das, R.; Lee, Y.-W.; Hou, S.; Moyano, D. F.; Duncan, B.; Liang, X.-J.; Rotello, V. M. The interplay of size and surface functionality on the cellular uptake of sub-10 nm gold nanoparticles. *ACS nano* **2015**, *9*, 9986–9993.

[13] Pengo, P.; Şologan, M.; Pasquato, L.; Guida, F.; Pacor, S.; Tossi, A.; Stellacci, F.; Marson, D.; Boccardo, S.; Pricl, S.; et al.. Gold nanoparticles with patterned surface monolayers for nanomedicine: current perspectives. *European Biophysics Journal* **2017**, *46*, 749–771.

[14] 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 letters* **2013**, *13*, 4060–4067.

[15] Zhang, Y.; Hudson-Smith, N. V.; Frand, S. D.; Cahill, M. S.; Davis, L. S.; Feng, Z. V.; Haynes, C. L.; Hamers, R. J. Influence of the spatial distribution of cationic functional groups at nanoparticle surfaces on bacterial viability and membrane interactions. *Journal of the American Chemical Society* **2020**, *142*, 10814–10823.

[16] Grzelczak, M.; Liz-MarzÃ¡n, L. M. Exploiting hydrophobic interactions at the nanoscale. *The journal of physical chemistry letters* **2014**, *5*, 2455–2463.

[17] Li, S.; Zhai, S.; Liu, Y.; Zhou, H.; Wu, J.; Jiao, Q.; Zhang, B.; Zhu, H.; Yan, B. Experimental modulation and computational model of nano-hydrophobicity. *Biomaterials* **2015**, *52*, 312–317.

[18] Sánchez-Iglesias, A.; Grzelczak, M.; Altantzis, T.; Goris, B.; Perez-Juste, J.; Bals, S.; Van Tendeloo, G.; Donaldson Jr, S. H.; Chmelka, B. F.; Israelachvili, J. N.; et al.. Hydrophobic interactions modulate self-assembly of nanoparticles. *ACS nano* **2012**, *6*, 11059–11065.

[19] 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*, 6748–6755.

[20] Sun, S.; Huang, Y.; Zhou, C.; Chen, S.; Yu, M.; Liu, J.; Zheng, J. Effect of hydrophobicity on nano-bio interactions of zwitterionic luminescent

gold nanoparticles at the cellular level. *Bioconjugate chemistry* **2018**, *29*, 1841–1846.

[21] Godawat, R.; Jamadagni, S. N.; Garde, S. Characterizing hydrophobicity of interfaces by using cavity formation, solute binding, and water correlations. *Proceedings of the National Academy of Sciences* **2009**, *106*, 15119–15124.

[22] Law, K.-Y.; *Definitions for hydrophilicity, hydrophobicity, and superhydrophobicity: getting the basics right*; 2014.

[23] Shenogina, N.; Godawat, R.; Keblinski, P.; Garde, S. How wetting and adhesion affect thermal conductance of a range of hydrophobic to hydrophilic aqueous interfaces. *Physical review letters* **2009**, *102*, 156101.

[24] 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*, 2078–2088.

[25] Ma, C. D.; Wang, C.; Acevedo-Vélez, C.; Gellman, S. H.; Abbott, N. L. Modulation of hydrophobic interactions by proximally immobilized ions. *Nature* **2015**, *517*, 347–350.

[26] Wang, C.; Ma, C.-K. 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. *Journal of the American Chemical Society* **2017**, *139*, 18536–18544.

[27] Yeon, H.; Wang, C.; Van Lehn, R. C.; Abbott, N. L. Influence of order within nonpolar monolayers on hydrophobic interactions. *Langmuir* **2017**, *33*, 4628–4637.

[28] Cao, Z.; Tsai, S. N.; Zuo, Y. Y. An Optical Method for Quantitatively Determining the Surface Free Energy of Micro-and Nanoparticles. *Analytical chemistry* **2019**, *91*, 12819–12826.

[29] Crandon, L. E.; Boenisch, K. M.; Harper, B. J.; Harper, S. L. Adaptive methodology to determine hydrophobicity of nanomaterials in situ. *PloS one* **2020**, *15*, e0233844.

[30] 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*, 3965–3967.

[31] Chen, K.; Rana, S.; Moyano, D. F.; Xu, Y.; Guo, X.; Rotello, V. M. Optimizing the selective recognition of protein isoforms through tuning of nanoparticle hydrophobicity. *Nanoscale* **2014**, *6*, 6492–6495.

[32] Luo, Z.; Murello, A.; Wilkins, D. M.; Kovacik, F.; Kohlbrecher, J.; Radulescu, A.; Okur, H. I.; Ong, Q. K.; Roke, S.; Ceriotti, M.; et al.. Determination and evaluation of the nonadditivity in wetting of molecularly heterogeneous surfaces. *Proceedings of the National Academy of Sciences* **2019**, *116*, 25516–25523.

[33] Batista, C. A. S.; Larson, R. G.; Kotov, N. A. Nonadditivity of nanoparticle interactions. *Science* **2015**, *350*.

[34] Kister, T.; Monego, D.; Mulvaney, P.; Widmer-Cooper, A.; Kraus, T. Colloidal stability of apolar nanoparticles: the role of particle size and ligand shell structure. *ACS nano* **2018**, *12*, 5969–5977.

[35] Matthew D áLane, J.; et al.. Assembly of responsive-shape coated nanoparticles at water surfaces. *Nanoscale* **2014**, *6*, 5132–5137.

[36] Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.

[37] Walker, D. A.; Leitsch, E. K.; Nap, R. J.; Szleifer, I.; Grzybowski, B. A. Geometric curvature controls the chemical patchiness and self-assembly of nanoparticles. *Nature nanotechnology* **2013**, *8*, 676–681.

[38] Ghorai, P. K.; Glotzer, S. C. Molecular dynamics simulation study of self-assembled monolayers of alkanethiol surfactants on spherical gold nanoparticles. *The Journal of Physical Chemistry C* **2007**, *111*, 15857–15862.

[39] Lane, J. M. D.; Grest, G. S. Spontaneous asymmetry of coated spherical nanoparticles in solution and at liquid-vapor interfaces. *Physical review letters* **2010**, *104*, 235501.

[40] Luedtke, W.; Landman, U. Structure, dynamics, and thermodynamics of passivated gold nanocrystallites and their assemblies. *The Journal of Physical Chemistry* **1996**, *100*, 13323–13329.

[41] 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. *Proceedings of the National Academy of Sciences* **2017**, *114*, 13345–13350.

[42] Wang, W.; Yan, X.; Zhao, L.; Russo, D. P.; Wang, S.; Liu, Y.; Sedykh, A.; Zhao, X.; Yan, B.; Zhu, H. Universal nanohydrophobicity predictions using virtual nanoparticle library. *Journal of cheminformatics* **2019**, *11*, 1–5.

[43] Das, M.; Dahal, U.; Mesele, O.; Liang, D.; Cui, Q. Molecular dynamics simulation of interaction between functionalized nanoparticles with lipid membranes: Analysis of coarse-grained models. *The Journal of Physical Chemistry B* **2019**, *123*, 10547–10561.

[44] Heikkila, E.; Martinez-Seara, H.; Gurtovenko, A. A.; Javanainen, M.; Häkkinen, H.; Vattulainen, I.; Akola, J. Cationic Au nanoparticle binding with plasma membrane-like lipid bilayers: potential mechanism for spontaneous permeation to cells revealed by atomistic simulations. *The Journal of Physical Chemistry C* **2014**, *118*, 11131–11141.

[45] Lolicato, F.; Joly, L.; Martinez-Seara, H.; Fragneto, G.; Scoppola, E.; Baldelli Bombelli, F.; Vattulainen, I.; Akola, J.; Maccarini, M. The role of temperature and lipid charge on intake/uptake of cationic gold nanoparticles into lipid bilayers. *Small* **2019**, *15*, 1805046.

[46] Pan, S.; Li, T.; Olvera de la Cruz, M. Molecular dynamics simulation of DNA-directed assembly of nanoparticle superlattices using patterned templates. *Journal of Polymer Science Part B: Polymer Physics* **2016**, *54*, 1687–1692.

[47] Tollefson, E. J.; Allen, C. R.; Chong, G.; Zhang, X.; Rozanov, N. D.; Bautista, A.; Cerda, J. J.; Pedersen, J. A.; Murphy, C. J.; Carlson, E. E.; et al.. Preferential binding of cytochrome c to anionic ligand-coated gold nanoparticles: A complementary computational and experimental approach. *ACS nano* **2019**, *13*, 6856–6866.

[48] Heikkilä, E.; Gurtovenko, A. A.; Martinez-Seara, H.; Häkkinen, H.; Vattulainen, I.; Akola, J. Atomistic simulations of functional Au144 (SR) 60 gold nanoparticles in aqueous environment. *The Journal of Physical Chemistry C* **2012**, *116*, 9805–9815.

[49] Riccardi, L.; Gabrielli, L.; Sun, X.; De Biasi, F.; Rastrelli, F.; Mancin, F.; De Vivo, M. Nanoparticle-based receptors mimic protein-ligand recognition. *Chem* **2017**, *3*, 92–109.

[50] Giovambattista, N.; Debenedetti, P. G.; Rossky, P. J. Hydration behavior under confinement by nanoscale surfaces with patterned hydrophobicity and hydrophilicity. *The Journal of Physical Chemistry C* **2007**, *111*, 1323–1332.

[51] Hua, L.; Zangi, R.; Berne, B. Hydrophobic interactions and dewetting between plates with hydrophobic and hydrophilic domains. *The Journal of Physical Chemistry C* **2009**, *113*, 5244–5253.

[52] Kanduč, M.; Schlaich, A.; Schneck, E.; Netz, R. R. Water-mediated interactions between hydrophilic and hydrophobic surfaces. *Langmuir* **2016**, *32*, 8767–8782.

[53] Monroe, J. I.; Shell, M. S. Computational discovery of chemically patterned surfaces that effect unique hydration water dynamics. *Proceedings of the National Academy of Sciences* **2018**, *115*, 8093–8098.

[54] Patel, A. J.; Garde, S. Efficient method to characterize the context-dependent hydrophobicity of proteins. *The Journal of Physical Chemistry B* **2014**, *118*, 1564–1573.

[55] Patel, A. J.; Varilly, P.; Chandler, D.; Garde, S. Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling. *Journal of statistical physics* **2011**, *145*, 265–275.

[56] Shin, S.; Willard, A. P. Characterizing hydration properties based on the orientational structure of interfacial water molecules. *Journal of chemical theory and computation* **2018**, *14*, 461–465.

[57] Patel, A. J.; Varilly, P.; Chandler, D. Fluctuations of water near extended hydrophobic and hydrophilic surfaces. *The Journal of Physical Chemistry B* **2010**, *114*, 1632–1637.

[58] 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.

[59] 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. *The Journal of Physical Chemistry B* **2012**, *116*, 2498–2503.

[60] Dallin, B. C.; Van Lehn, R. C. Spatially heterogeneous water properties at disordered surfaces decrease the hydrophobicity of nonpolar self-assembled monolayers. *The journal of physical chemistry letters* **2019**, *10*, 3991–3997.

[61] Kelkar, A. S.; Dallin, B. C.; Van Lehn, R. C. Predicting Hydrophobicity by Learning Spatiotemporal Features of Interfacial Water Structure: Combining Molecular Dynamics Simulations with Convolutional Neural Networks. *The Journal of Physical Chemistry B* **2020**, *124*, 9103–9114.

[62] Willard, A. P.; Chandler, D. Instantaneous liquid interfaces. *The Journal of Physical Chemistry B* **2010**, *114*, 1954–1958.

[63] Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell Jr, A. D. Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach. *Journal of chemical information and modeling* **2013**, *53*, 3384–3398.

[64] Lexa, K. W.; Carlson, H. A. Full protein flexibility is essential for proper hot-spot mapping. *Journal of the American Chemical Society* **2011**, *133*, 200–202.

[65] Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry* **2009**, *30*, 2157–2164.

[66] Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation* **2012**, *8*, 3257–3273.

[67] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[68] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[69] Heinz, H.; Vaia, R.; Farmer, B.; Naik, R. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12- 6 and 9-6 Lennard-Jones potentials. *The Journal of Physical Chemistry C* **2008**, *112*, 17281–17290.

[70] Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *The Journal of chemical physics* **1996**, *105*, 1902–1921.

[71] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[72] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *International conference on exascale applications and software*; Springer; pp 3–27.

[73] Hill, H. D.; Millstone, J. E.; Banholzer, M. J.; Mirkin, C. A. The role radius of curvature plays in thiolated oligonucleotide loading on gold nanoparticles. *ACS nano* **2009**, *3*, 418–424.

[74] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry* **1992**, *13*, 1011–1021.

[75] Sigal, G. B.; Mrksich, M.; Whitesides, G. M. Effect of surface wettability on the adsorption of proteins and detergents. *Journal of the American Chemical Society* **1998**, *120*, 3464–3473.

[76] Wang, H.; Chen, S.; Li, L.; Jiang, S. Improved method for the preparation of carboxylic acid and amine terminated self-assembled monolayers of alkanethiolates. *Langmuir* **2005**, *21*, 2633–2636.

[77] Sarupria, S.; Garde, S. Quantifying water density fluctuations and compressibility of hydration shells of hydrophobic solutes and proteins. *Physical Review Letters* **2009**, *103*, 037803.

[78] Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* **2017**, *42*, 1–21.

[79] Ghanakota, P.; Carlson, H. A. Driving structure-based drug discovery through cosolvent molecular dynamics: Miniperspective. *Journal of medicinal chemistry* **2016**, *59*, 10383–10399.

11 MOLECULAR DYNAMICS-DERIVED DESCRIPTORS FOR PREDICTING LIGAND-COATED GOLD NANOPARTICLE BEHAVIOR

Chapters 8-10 provided a framework for developing gold nanoparticle systems and investigated the effects of gold core and ligand selection on their interactions with other molecules. A critical question is whether these computational workflows could guide the design of gold nanoparticles. This chapter seeks to address the following questions:

- What simulation-derived molecular descriptors are important for characterizing monolayer-protected gold nanoparticle systems?
- Given the molecular descriptors, what machine learning model is best to predict experimentally determined gold nanoparticle properties?
- How could we leverage machine learning to gain physical insight into how the selection of gold core and ligand selection translate to gold nanoparticle behavior?

In this chapter, we use atomistic molecular dynamics simulations to model 154 monolayer-protected gold nanoparticles in aqueous solution to understand how gold core and ligand selection can affect properties

This chapter was reproduced with permission from *ACS Nano*, submitted for publication, and cited as Chew, A. K.; Pedersen, J. A.; Van Lehn, R. C. Predicting Nano-bio Activities of Monolayer-Protected Gold Nanoparticles using Molecular Dynamics-Derived Descriptors. *In preparation*. Unpublished work copyright 2021 American Chemical Society. The supporting information is cited as Ref. 2.

relevant to nanoparticle behavior. We developed a library of 25 simulation-derived molecular descriptors that capture structural and chemical properties of the monolayer. We then leveraged interpretable, data-centric models (*i.e.* LASSO and random forest) to develop structure-property relationships and predict experimental GNP behavior from a curated dataset, such as logP, cell uptake, and zeta potential. These models reveal that simulation-derived molecular descriptors could be used to accurately predict experimental trends, and they yield physical insight into how GNP parameters can be tuned for a selective behavior. Finally, we show that these models could predict cell uptake trends for 12 unseen nanoparticles, suggesting that molecular simulations of monolayer-protected gold nanoparticles could be used to screen and guide the design of nanoparticles.

11.1 Introduction

Gold nanoparticles (GNPs) are promising materials for biological applications, such as drug delivery, biosensing, and photothermal therapy, because their interactions with biological materials can be tailored by grafting organic ligands to the GNP surface to form a self-assembled monolayer (SAM).[3–5] Small GNPs with core diameters less than 10 nm are of particular interest because this size limits renal clearance from the body.[6] This size is also commensurate with that of typical biomolecules (*e.g.*, proteins)

and the thickness of the cell membrane, enabling the study of related nano-bio at comparable length scales.[7] Because such interactions can be modulated by tailoring the ligand composition of the protecting SAM, substantial effort has been devoted to understanding how ligand propertes impact interactions at the nano-bio interface. For example, experimentally modulating the ligand hydrophobicity - or the thermodynamic affinity of a ligand for water - by changing the extent of the nonpolar groups on the ligand have been correlated to changes in immune response,[8] binding affinities toward proteins,[9] and cellular uptake kinetics.[10,11] Furthermore, modulating ligand electrostatic interactions, such as introducing positively charged groups within the ligand, have been found to increase the interactions between small GNPs and slightly negatively charged lipid bilayers, resulting in GNP adsorption to or internalization within lipid bilayers.[12–14] In contrast, small GNPs with negatively charged ligands do not favorably penetrate through the lipid bilayer.[13–16] Small GNPs with zwitterionic ligands have been shown to prevent strong protein adsorption onto the surface, which could avoid the masking of surface properties due to the formation of protein coronas.[17] However, despite these advancements, engineering small GNPs for targeted nano-bio interactions[18] (*e.g.*, selective protein binding or favorable cell uptake) remains challenging because subtle variations to the gold core size and ligand properties can manifest as substantial changes to GNP behavior,[8,14,17,19–27] which are challenging to predict *a priori*.[28] Experimentally exploring the vast design space of

possible GNP compositions is also time-consuming and difficult, in part because of the lack of molecular level detail that could be obtained at nanometer length scales, especially for complex, non-planar geometries arising from small GNPs.[29,30]

Alternatively, computational modeling that yield accurate quantitative nanostructure-activity relationships (QNAR) is critical to rationally designing effective GNPs and alleviates extensive trial-and-error experimentation.[31,32] QNAR models use numerical parameters called descriptors that capture characteristics of GNPs and relate them to the activity of GNPs in biological environments through a varity of machine learning algorithms (*e.g.* multilinear regression, support vector machines, and so on).[31] Descriptors could consist of experimental descriptors (*e.g.* size, shape, zeta potentials, *etc.*)[32] or classical molecular descriptors for small organic molecules (*e.g.* constitutional, topological, electrostatic, *etc.*);[33] however, experimental descriptors are challenging to measure for a large range of particles and classical molecular descriptors do not take into account mixtures of organic and inorganic molecules.[31] Hence, atomistically modeling GNPs have been employed to better describe their structure, provide informative descriptors, and improve the accuracy of QNAR models.

Virtual GNP (vGNP) models, which are static models consisting of a gold core with ligands randomly placed on the core surface, were developed to more accurately capture the diversity of GNPs, such as the gold core size, ligand density, and surface chemistry.[34] Novel descriptors were

developed to capture surface properties of vGNPs and used to develop QNAR models that could predict both biophysiochemical properties (*e.g.* logP and zeta potentials) and bioactivity (*e.g.* cell uptake, GNP-enzyme bindings, *etc.*).[34–37] Furthermore, deep learning methods, such as convolutional neural networks, have been used to analyze vGNPs and predict GNP observables, avoiding tedious descriptor calculations.[38] While vGNPs enable rapid screening across a large range of particles, these models do not account for non-obvious, cooperative interactions between the ligands and their surrounding environment that are important to characterizing monolayer properties.

Alternative to vGNPs, atomistic molecular dynamics (MD) simulations have been employed to model the GNP in a variety of explicit solvents that could capture ligand-ligand and ligand-solvent interactions.[11,30,39–45] For example, GNPs coated with long, nonpolar alkanethiol ligands were experimentally[46] and computationally[39–45] found to form bundles, where these ligands align in the same direction due to preferred methylene interactions. These bundles dictate how GNPs bind with one another to minimize the extent of hydrophobic area exposed,[39,40] how GNPs can be used to bend single stranded nucleic acids,[47] and how spatially heterogenous surface properties emerge despite the monolayer being chemically homogenous.[45] Therefore, we hypothesize that MD simulations could better capture monolayer properties arising from these non-obvious, collective interactions as compared to static models, which could improve

QNAR predictions of experimental data. Another added-benefit of using MD is the ability to analyze GNPs in the presence of lipid bilayers or proteins,[25–27] allowing for in-depth mechanistic studies that is not possible with vGNPs. While these past studies, along with many other studies of SAM-coated GNPs,[25–28,30,48] have provided useful insights into the interplay of gold core and ligand selection on SAM properties that influence GNP behavior with other biomolecules, these studies have primarily focused on a limited subset of GNPs, in part because there was a lack of an experimentally curated dataset.[37] Therefore, the utility of MD to predict GNP activity more broadly across a large range of GNPs have yet to be employed.

In this work, we modeled 154 sub-10-nm GNPs in aqueous solution using atomistic MD simulations and developed a library of 25 MD-derived descriptors that characterizes structural and chemical properties of GNPs. We leverage interpretable, data-centric models, such as least absolute shrinkage and selection operator (LASSO) regression and random forest (RF) models, to develop QNAR models using the descriptors as input and experimental labels as output, such as logP, cell uptake in A549 cells, and zeta potential measurements from Ref. 37. We then use feature extraction tools to pinpoint the most important GNP descriptors that relate to these experimental observables, which provide useful information in designing new GNPs for selective behavior. Finally, we test the generalizability of the cell uptake model by predicting GNP behavior in a separate dataset[19]

consisting of 12 GNPs and found that the RF model could capture cell uptake trends for unseen GNPs. These models are useful for predicting GNP behavior *a priori* using computationally efficient tools and enables the rational design of new GNPs for selective nano-bio behavior.

11.2 Methods

11.2.1 Development of nanoparticle systems

SAM-coated GNPs were developed using a self-assembly approach as described previously in Ref. 41. Either butanethiol (SMILES: SCCCC) or 1,2-dithiolane (SMILES: C1CCSS1) ligands were self-assembled onto a spherical gold core as shown in Figure 11.2. Ligands were considered adsorbed if all sulfur atoms have a sulfur-gold interaction within 0.327 nm. These adsorbed ligands were then replaced with ligands used in this study. Ligand atoms were modeled with the CGenFF/CHARMM36 forcefield (July 2020 version),[49–51] gold atoms were modeled with the Interface force field,[52] and water molecules were modeled using the TIP3P model.[53] To eliminate atomic clashes between ligands, a short 4 ps NVT simulation was performed with all interactions turned off and the last heavy atom of the ligand restrained radially from the gold core center to the box edge with a spring constant of 5,000 kJ/mol/nm^2. Then, van der Waals interactions were slowly turned on in a series of energy minimization steps, as described in the Supporting Information.[2] Rhombic dodecahedron peri-

odic boundaries were set with a 1 nm distance between the ligand atoms and the box edge to maximize computational efficiency for GNP systems with approximate spherical symmetry. The system was then solvated with water, and, for charged systems, sodium or chlorine counterions replaced water molecules to ensure charge-neutral systems. All subsequent simulations were performed with temperature T = 300 K and pressure P = 1 bar. A 1 ns *NPT* equilibration was performed with the temperature controlled by the velocity-rescale thermostat and the pressure controlled by the Berendsen barostat. Then, a 50 ns *NPT* production simulation was performed with the same thermostat and the pressure controlled by the Parinello-Rahman thermostat. The last 40 ns of simulation data was used to compute MD-derived descriptors, which was sufficient simulation time for descriptors to converge (see SI, Figure S4).[2]

11.2.2 Simulation parameters

In all simulations, Verlet lists were generated using a 1.2 nm neighbor list cutoff. Van der Waals interactions were modeled with a Lennard-Jones potential using a 1.2 nm cutoff that was smoothly shifted to zero between 1.0 and 1.2 nm. Electrostatic interactions were calculated using the smooth particle mesh Ewald method with a short-range cutoff of 1.2 nm, grid spacing of 0.12 nm, and fourth order interpolation. Bonds were constrained using the LINCS algorithm.[54] Periodic boundary conditions were enabled in all directions. All classical MD simulations were per-

formed using Gromacs 2016[55] using the leapfrog integrator with a 2-fs timestep.

11.2.3 Computing MD-derived descriptors

All MD-derived descriptors were generated with a combination of in-house Python (MDTraj[56] and MDAnalysis)[57,58] and Gromacs analysis tools.[55] The full list of the 25 MD-derived descriptors computed are tabulated in Table S1.[2] 15 uncorrelated descriptors shown in Table S2 were used to develop QNAR models.[2]

11.2.4 Computational models

Each experimental dataset was divided into training (80% of the original dataset) and test sets (20% of the original dataset). The training sets were used to build the models and the test sets were used to evaluate model transferability. To evaluate the generalizability of the models across the training data, 5-fold cross validation was performed where, for each fold, 80% of the training set was used to train the model and the remaining 20% was used to validate the model; the process is iterated until all of the training labels had a chance to be in the validation set. When predicting the test set, the models were trained using all the training set data. Two models were used in this work: (1) LASSO and (2) RF models. These models were selected for their ability to down-select the most important

features associated with their predictions. Feature importance was determined using the SHapley Additive exPlanation (SHAP) method,[59,60] where the average magnitude of the Shapley values is reported. The sign of the feature importance was determined by the Pearson's r value between Shapley and descriptor values, where Pearson's r greater than 0 indicate a positive sign and less than 0 indicate a negative sign on the feature importance. Feature importance and its associated error bars were estimated by a bootstrapping method, where 90% of the training data (randomly selected without replacement) were used to re-train the model and iterated for a series of ten trials. The average feature importance of the trials were reported and the error was estimated by computing the standard deviation of the trials. To test robustness of using MD-derived descriptors to predict experimental observables, a second trial of the GNP systems were produce using the same simulation protocol and starting from self-assembly simulations. The second trial was performed with a shorter GNP-water simulation of 20 ns, with the last 10 ns used for MD-derived descriptor calculations. The second trial performed similarly to the longer 50 ns GNP-water simulation in this text, suggesting robustness in this computational workflow and a lower simulation time necessary for MD-derived descriptors to make accurate predictions (see SI, Figure S9-S11).[2]

11.3 Results and discussion

11.3.1 Experimental datasets used to develop QNAR relationships

To develop QNAR relationships, we used experimental data from a curated database consisting of 414 unique GNPs with physiochemical (*e.g.* logP and zeta potential) and bioactivity (*e.g.* cell uptake) values.[37] These three tabulated GNP observables are schematically summarized in Figure 11.1b. LogP is the partitioning of GNPs between octanol and water phases, where larger logP values indicate that the GNPs prefer octanol phases more than the water phases, suggesting that the surface is hydrophobic. Cell uptake in A549 cells is the propensity of GNPs to be internalized within the cell. Finally, zeta potential in water is the effective charge of the GNP, where higher absolute zeta potentials values are suggestive of stable suspensions.[61] Altogether, these experimental observables provide insight into the activity of GNPs within a biological environment, which are useful for designing GNPs and provide labels for developing QNAR models.

For this work, we focused on spherical GNPs with $<$ 10 nm in core diameter, resulting in 154 unique GNPs (96 single- and 58 multi-component) with 105 unique ligands and core diameters ranging between 2 - 8.5 nm. Figure 11.1c shows the number distribution of GNPs with experimental values of 110 logP, 65 cellular uptake in A549 cells, and 102 zeta poten-

tial values for the 154 unique GNPs. LogP and cell uptake datasets have histograms that are well-distributed, whereas the zeta potential dataset have histograms that are skewed with GNPs primarily having negative zeta potential values. The skewed distribution in the zeta potential dataset towards negative values could be attributed to the negative zeta potential of citrate-reduced gold nanoparticles, where a 13.3 nm diameter core has a zeta potential of approximately -39.7 $\pm$ 0.7 mV at pH 9.3.[62] The experimental datasets from Ref. 37 provide information about the core diameter, ligand structure, and ligand grafting density, which were used to model GNPs with MD simulations.

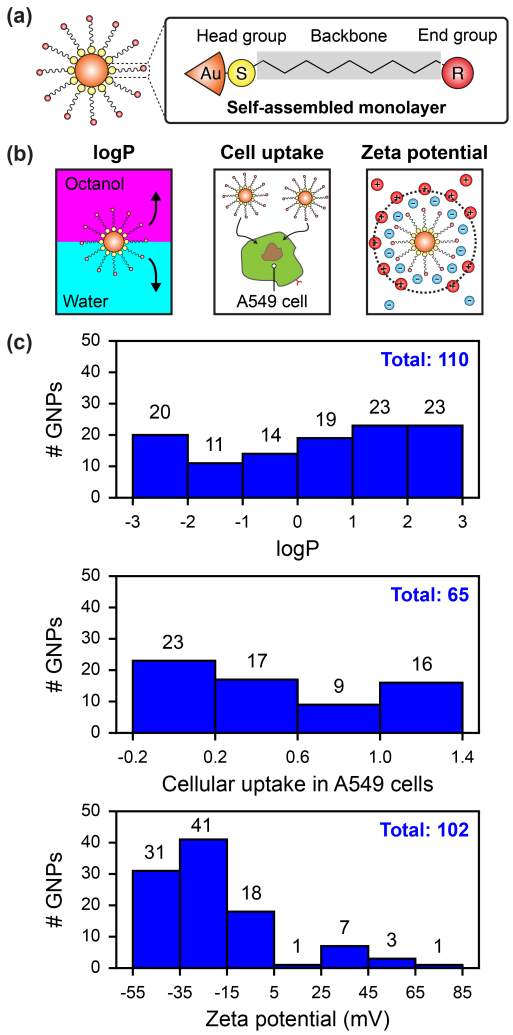


Figure 11.1: Experimental data used to develop QNAR models. (a) Example of a self-assembled monolayer structure consisting of a sulfur head group, nonpolar backbone, and end group, which are bound to the gold core with a strong gold-sulfur interaction. (b) Schematic representation of experimental observables for GNP behavior, such as logP, cell uptake in A549 cells, and zeta potential in water. (c) Number distribution of GNPs with experimental labels of logP, cell uptake in A549 cells, and zeta potential in water. The total number of GNPs for each experimental observable is shown on the upper right. All experimental data were taken from Ref. 37.

11.3.2 Computational workflow to model GNP systems with MD and compute MD-derived descriptors

Given the large space of gold core sizes and ligand chemistries, we used a generalized workflow to model GNP systems for any arbitrary gold core shape and size, and ligand selection as described in Ref. 41, which was modified to model the ligand chemistries observed experimentally.[37] Figure 11.2 summarizes the workflow from generating GNP models using a database consisting of spherical gold core diameter and ligand simplified molecular-input line-entry (SMILES) strings; GNP1 and GNP288 are used as representative examples with distinct gold core sizes and ligand structures (all nomenclature is the same as Ref. 37). The database from Ref. 37 consists of ligand SMILES strings with one of two unique substructure patterns: (1) butanethiol (SMILES: SCCCC) and (2) 1,2-dithiolane (SMILES: C1CCSS1). To position the ligand on the gold surface, we performed self-assembly simulations where an excess of butanethiol or 1,2-dithiolane molecules were placed around a gold core and assembled onto the surface *via* a strong sulfur-gold Lennard-Jones (LJ) interaction.[41,63] Substructures that have sulfur atom(s) within 0.327 nm of the gold atoms were considered bound onto the gold core, whereas all other substructures were considered unbound (depicted as green molecules in Figure 11.2). After self-assembly simulations, unbound substructures are removed and bound substructures were randomly removed to match the total number of the ligands

in the database. If the total number of ligands from the database is larger than the number of bound substructures from self-assembly simulations, then we use all bound substructures, which provided reasonable grafting densities (Figure S1).[2] The substructures are then replaced with ligand structures from the GNP database, and the GNP system is then simulated in the presence of water molecules and counterions, depicted as cyan in Figure 11.2. We used this workflow to systematically model 154 GNPs in aqueous solution and compute physically motivated MD-derived descriptors that characterize GNP properties. GNP-water simulations were performed for 50 ns in the *NPT* ensemble (see Methods) and the last 40 ns of simulation time were used to compute MD-derived descriptors.

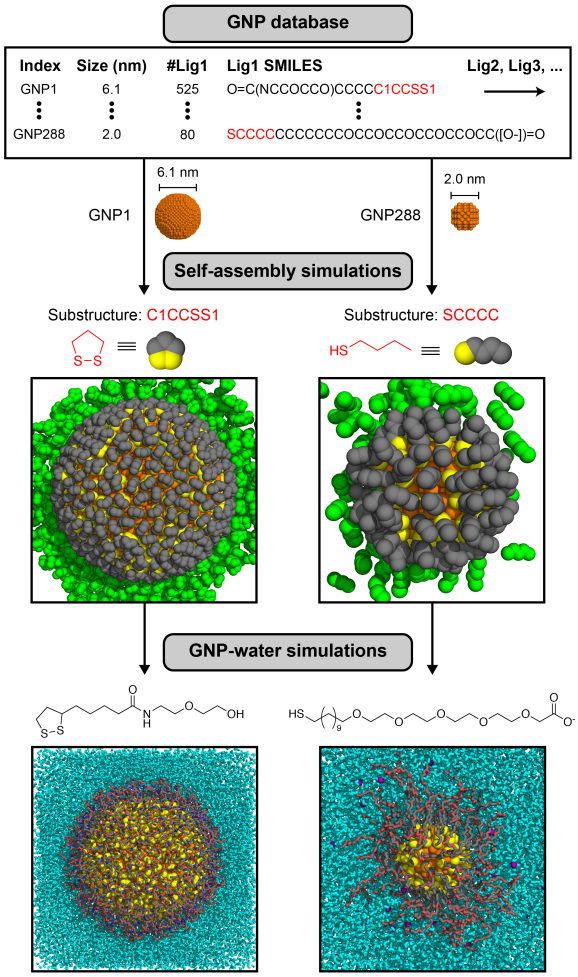


Figure 11.2: Computational workflow for modeling GNPs using MD simulations. The GNP database consists of diameters of spherical gold cores, number of ligands adsorbed on the surface, and ligand structure in the form of a SMILES string. The diameter and ligand substructure, either butanethiol (SMILES: SCCCC) or 1,2-dithiolane (SMILES: C1CCSS1), were then used to initiate self-assembly simulations. Substructures that are not adsorbed are shown in green. After the substructures are adsorbed onto the gold core surface, excess substructures are removed and adsorbed substructures are replaced with ligands from the GNP database and simulated in pure water (shown as cyan). Sodium or chlorine counterions are shown as purple and included to ensure charge-neutral systems. This schematic uses GNP1 and GNP288 as representative examples with distinct core sizes and ligand substructures.

We developed a library of simulation-derived molecular descriptors that capture physiochemical properties of GNPs. Figure 11.3a shows examples of three GNPs (labeled as GNP14, GNP123, and GNP151) with varying core diameters and ligand structures; the simulation snapshots are shown on the right with complex SAM geometries. Figure 11.3b shows the solvent-accessible-surface area (SASA) and number of ligand-water hydrogen bonds *versus* simulation time for the three GNP examples. These descriptors capture non-obvious structural characteristics arising from the interplay of the gold core and ligand selection. For instance, GNP14 has the smallest core diameter, but it has approximately the same SASA as GNP123 due to the longer ligand attached on GNP14. Furthermore, GNP123 has the largest number of ligand-water hydrogen bonds, despite the ligand on GNP123 having fewer oxygen and nitrogen atoms as compared to that of GNP14. These results suggests that MD simulations and corresponding descriptor calculations capture non-obvious characteristics of GNPs, which could be used to capture trends that could predict GNP behavior. All descriptors were standardized by subtracting the mean and dividing by the standard deviation, so they could be compared on the same magnitude.

Figure 11.3c shows a truncated set of 25 total MD-derived descriptors that encompass structural and chemical properties of GNPs. These descriptors were selected based on previous literature;[18,34–36,41,64] a full table of the descriptors is available in Table S1.[2] The descriptor space was then filtered by removing all correlated descriptors, which was performed us-

ing a Pearson's r correlation matrix between the 25 descriptors, as shown in Figure 11.3d. Red regions in Figure 11.3d mean that Pearson's r values are close to 1, indicating that the descriptors are highly correlated, whereas blue regions mean the converse. Descriptors with the magnitude of Pearson's r values greater than 0.90 were removed, resulting in 15 uncorrelated descriptors tabulated in Table S2.[2] These uncorrelated descriptors were used to develop QNAR models to predict GNP experimental observables.

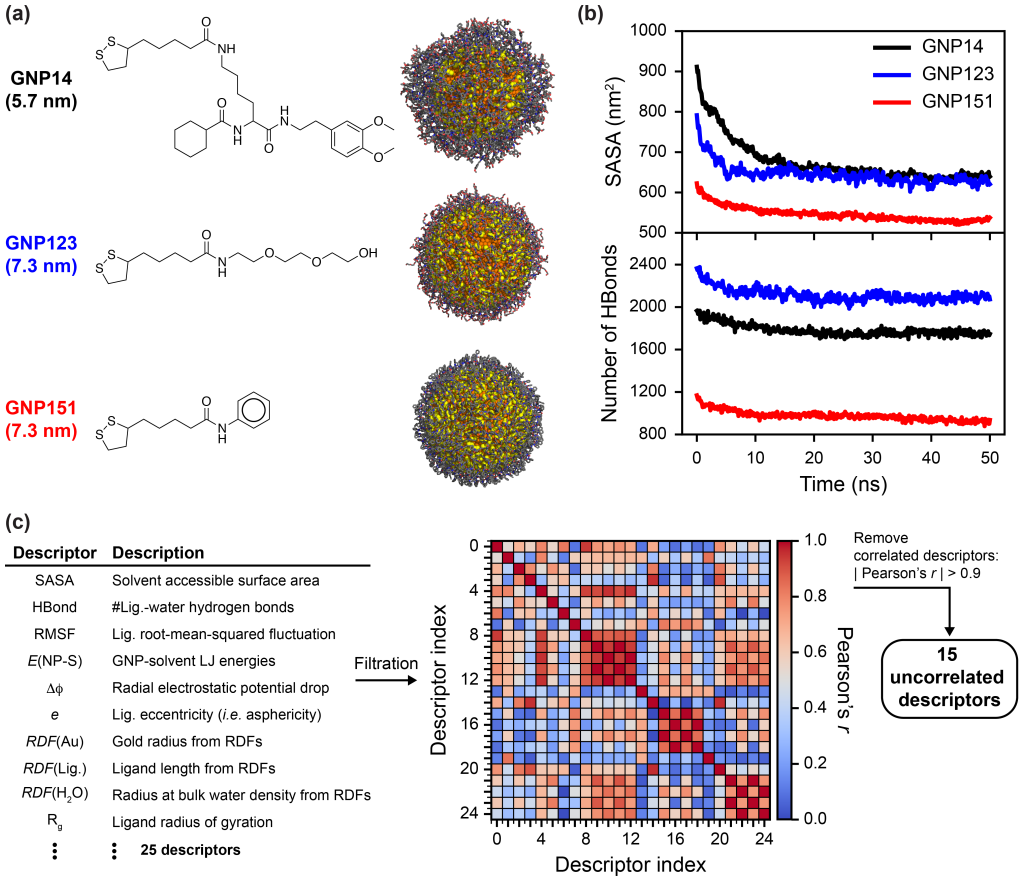


Figure 11.3: MD-derived descriptors used to develop QNAR models. (a) Three examples of GNPs with varying core diameters (5.7 *versus* 7.3 nm) and ligand structures. Simulation snapshots are shown without water molecules. (b) Solvent-accessible-surface area (SASA) and number of ligand-water hydrogen bonds (HBonds) *versus* simulation time for the three GNPs in (a). (c) Concise list of of MD-derived descriptors used for developing QNAR models. The full 25 descriptor set is described in Table S1.[2] (d) Pearson's r between the 25 descriptor space, which shows red for highly correlated descriptors and blue for uncorrelated descriptors. Descriptor indexes correspond to descriptors in Table S1.[2] Highly correlated descriptors with | Pearson's r | > 0.9 were removed to output 15 uncorrelated descriptors.

11.3.3 QNAR models using MD-derived descriptors accurately predicts GNP behavior

We sought to develop QNAR models that use the uncorrelated descriptors from Table S2[2] as the input and experimental labels (*i.e.*, logP, cell uptake, zeta potential) as the output. We compared two algorithms to probe the prediction capabilities of using MD-derived descriptors: (1) LASSO and (2) RF models. LASSO is a linear model that minimizes the residual sum squared and the L_1 norm of the regression weights. An advantage of using the LASSO model, as compared to a typical linear regression model, is its ability to remove descriptors that do not significantly contribute to the prediction of the experimental observable, which is useful for down-selecting the most important descriptors associated with GNP behavior. Conversely, RF is a non-linear model consisting of a collection of decision trees that are each trained using different subsets of the training data; these trees then collectively vote on a predicted output value. We expect that since the RF model is non-linear, it may capture non-obvious correlations between the descriptors that could improve the prediction accuracy as compared to the linear model from LASSO. To test the transferability of the models to unseen data, we split the experimental labels with 80% of the labels as the training set and 20% of the labels as the testing set. Test set predictions were performed after the models were fully trained with the training set. To test generalizability of the models across the training set,

5-fold cross validation (5-CV) was performed on the training set, where the training set was further split into five subsets. For each fold, the model was trained on four out of five subsets and tested on the remaining subset (*i.e.* validation set). The procedure was iterated five times until all of the subsets was used in the validation set once.

Figure 11.4 shows the predicted *versus* experimental logP (left), cell uptake in A549 cells (middle), and zeta potential values in water (right) using LASSO (top) and RF (bottom) models; 5-CV predictions are shown as black dots and test set predictions are shown as red dots with the corresponding color-coded Pearson's r between predicted and experimental values are shown in the upper left of each plot. For logP and cell uptake data sets, LASSO models perform well with high 5-CV Pearson's r greater than 0.80 and slightly diminished test set performance with Pearson's $r \geqslant 0.70$. However, LASSO models performed poorly in predicting zeta potential in water with 5-CV and test set Pearson's r of 0.66 and 0.56, respectively. Conversely, RF models significantly improve predictions for logP, cell uptake, and zeta potential datasets with 5-CV Pearson's $r \geqslant 0.86$ and test set Pearson's $r \geqslant 0.74$, suggesting that non-linear models could improve the prediction accuracy. These results also show that simulation-derived molecular descriptors from GNP in solution could be used to predict GNP experimental observables, despite not explicitly modeling the environments that the GNPs were measured experimentally, such as the octanol-water phases for logP values or the lipid membrane for cell uptake

values. Furthermore, these descriptors are not specific to GNP systems (other than radial distribution functions), which suggests that they could be more broadly used for different types of ligand-coated nanomaterials.

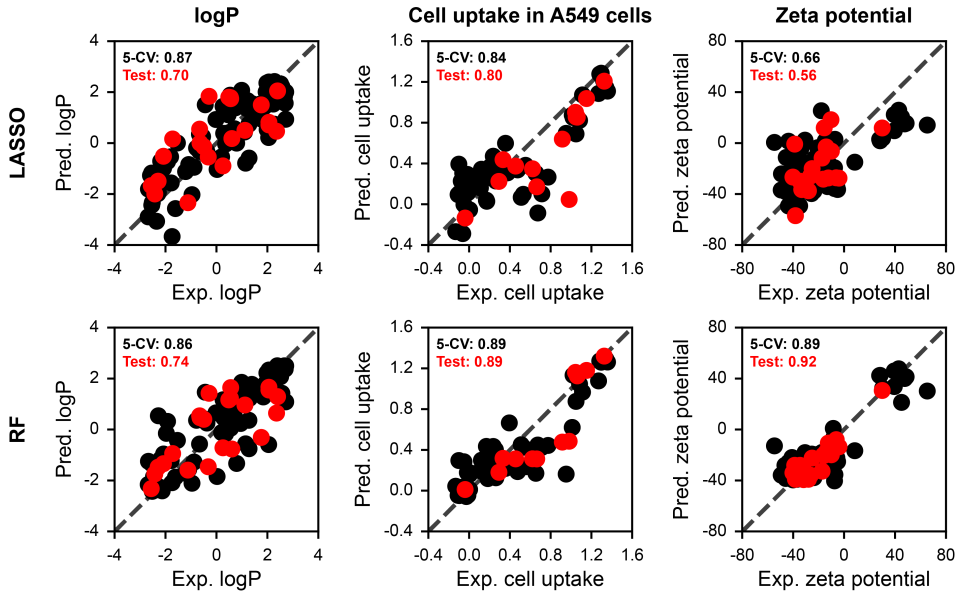


Figure 11.4: Prediction accuracy of QNAR models. Parity plots of predicted *versus* experimental logP (left), cell uptake in A549 cells (middle), and zeta potential in water (right) values are shown for LASSO (top) and random forest (bottom) models. 5-fold cross validation (5-CV) predictions are shown as black dots and test set predictions are shown as red dots. Pearson's r between predicted and experiments for 5-CV and test sets are shown in the upper left with the same color scheme. Dashed lines indicates when the predicted values equate the experimental values.

11.3.4 Consensus feature selection informs on GNP design

Given the success of LASSO and RF models in predicting GNP experimental trends (Figure 11.4), we next sought to identify the top features that are important to capturing the trends in the experiments. The advantage of using LASSO and RF models is their ability to inform on feature importance, which is typically measured as the magnitude of the regression weights for LASSO models, or measured by computing the gini information gain of the nodes within decision trees for RF models.[65] However, these approaches are model-dependent and makes comparing feature importance between models challenging. Alternative to these approaches, the SHapley Additive exPlanation (SHAP) method was recently introduced, which is a model-agnostic approach that have been used measure feature importance even for "black box" models, such as deep neural networks.[66] The SHAP method assigns an importance value for a feature by comparing model predictions with and without the feature across all possible permutations of the feature space, then computing Shapley values by averaging the marginal contribution of the feature.[59,60] Shapley values estimate feature importance by the magnitude of these values, as well as the positive or negative contribution to the prediction output through the sign (*e.g.* a negative Shapley value means the feature contributes negatively to the prediction).[66,67] We use Shapley values to prescribe feature importance on

the descriptor space, whereby for each descriptor, the average magnitude of the Shapley values across all instances (*i.e.* Mean |Shap|) is reported and the sign of the feature importance is determined by the Pearson's r value between Shapley and descriptor values. To estimate the accuracy of the feature importance outputted from LASSO and RF models with the SHAP method, a bootstrapping procedure was implemented,[68] which involves re-training LASSO and RF models with 90% of the training set (randomly sampled without replacement). This procedure was iterated 20 times such that the average of the trials is reported and the error is estimated by computing the standard deviation of the trials.

Given that the usefulness of feature importance is dependent on model accuracy, we primarily focus on the feature importance from RF models, which outperformed LASSO models in predicting experimental trends (Figure 11.4). Figure 11.5a shows the top three important descriptors with the highest feature importance for predicting logP (left), cell uptake in A549 cells (middle), and zeta potential (right) datasets using the RF model; red bars indicates that the descriptor negatively contributes to the experimental values, where as blue bar indicates the converse. For the logP dataset, the RF model identifies the RMSF normalized by SASA ($RMSF_{SASA}$) as the top descriptor, which is similar to the LASSO model that identifies RMSF as the top descriptor (Figure S6)[2] and suggests a consensus on the top descriptor critical for predicting logP. The RF model identifies the second and third top descriptors as LJ energies between SASA-normalized GNP-

solvent ($E(\text{NP-S})_{\text{SASA}}$) and ligand-normalized GNP-GNP ($E(\text{NP-NP})_{\#\text{Lig.}}$) for the logP dataset, whereas the LASSO model identifies ligand-water hydrogen bonding normalized by SASA ($\text{HBond}_{\text{SASA}}$) and LJ energies between GNP-GNP ($E(\text{NP-NP})_{\text{SASA}}$ as second and third top descriptors (Figure S6).[2] There is no clear consensus of the second and third top descriptor between RF and LASSO models for logP predictions.

For the cell uptake dataset, $\text{HBond}_{\#\text{Lig.}}$ and $\text{HBond}_{\text{SASA}}$ appear within the top three descriptors for both RF and LASSO models, suggesting that the extent of ligand-water hydrogen bonding is important the propensity of GNPs to be taken into a cell. The connection between ligand-water hydrogen bonding and cell uptake is non-obvious, and it has been previously observed in cell-penetrating peptides, where surface hydrogen bonding affects adsorption onto lipid bilayers.[69] For the zeta potential dataset, there are no consensus between RF and LASSO models on the top descriptors, possibly due to the poor prediction accuracy of LASSO for this dataset (Figure 11.4). The RF model identifies SASA-normalized GNP-GNP LJ energies ($E(\text{NP-NP})_{\text{SASA}}$) as the most important descriptor, followed by ligand-water hydrogen bonding (HBond), and the electrostatic potential drop between bulk water and gold core ($\Delta\phi$). Since zeta potential measures the effective charge on the nanoparticle surface,[61,70] we expect that information about charges (*e.g.* $\Delta\phi$) would be useful to predicting the zeta potential,[64] which is consistent with the predicted top descriptors from the RF model.

We next focus on select GNPs that exhibit the extremes of the top descriptor for each dataset. Figure 11.5b shows the simulation snapshots of GNPs with the highest (top) and lowest (bottom) top descriptor and the corresponding experimental observable for logP (left), cell uptake in A549 cells (middle), and zeta potential (right) datasets. For logP datasets, GNP132 and GNP146 have the highest and lowest $RMSF_{SASA}$ values, where increasing $RMSF_{SASA}$ values lead to more negative logP values. A clear distinguishing feature between these two particles is that GNP132 has fewer ligands on the surface as compared to GNP146, which results in a larger free volume accessible for the ligands to fluctuate. These results suggest that larger ligand fluctuations in water (*i.e.* good solvation) would lead to a higher preference in the water phase (*i.e.* lower logP values). For cell uptake datasets, GNP169 and GNP152 have the highest and lowest $HBond_{SASA}$ values, where increasing $HBond_{SASA}$ values result in a decrease in the experimental cell uptake values. Since the ligands on GNP169 have more oxygen and nitrogen atoms as compared to the ligands on GNP152, they could form more hydrogen bonds with water. Hence, GNP169 may prefer interacting with water more than with the lipid bilayer, resulting in a lower cell uptake value as compared to GNP152. For zeta potential datasets, GNP132 and GNP129 has the highest and lowest $E(\text{NP-NP})_{SASA}$ values, where higher $E(\text{NP-NP})_{SASA}$ values lead to more positive zeta potentials. More negative $E(\text{NP-NP})_{SASA}$ values suggest favorable ligand-ligand interactions, which is shown as densely packed ligands for GNP129.

Conversely, less negative $E(\text{NP-NP})_{\text{SASA}}$ suggest diminished ligand-ligand interactions, shown as GNP132 with fewer ligands on the surface. These findings suggest varying grafting densities, which could affect the charge on a GNP surface, strongly impact zeta potential measurements.

The feature importance results in Figure 11.5a are useful for designing new GNPs. For example, tuning the ligand RMSF would affect the propensity of GNPs to partition into octanol/water phases (*i.e.* logP), which agrees with previous literature that found ligand fluctuations is closely related to the surface hydrophobicity.[71] Similarly, tuning extent of ligand-water hydrogen bonds, by including or removing oxygen/nitrogen atoms in the ligand structure, would affect the propensity of cell uptake. Furthermore, accurately computing descriptors like HBond or RMSF is only possible using a MD simulation, further supporting the use of MD to interrogate surface characteristics of GNPs rather than static GNP models.[24,34,37] Furthermore, we tested the robustness of these models by repeating the entire workflow for a second trial consisting of a shorter GNP-water simulation of 20 ns and using the last 10 ns to compute MD-derived descriptors. We found that the model prediction (Figure S9) and feature selection (Figure S10) are consistent with the findings presented for the longer 50 ns simulations.[2] These findings show that using MD-derived descriptors could robustly predict experimental observables, and even shorter simulations could be used to screen for enhanced GNP properties.

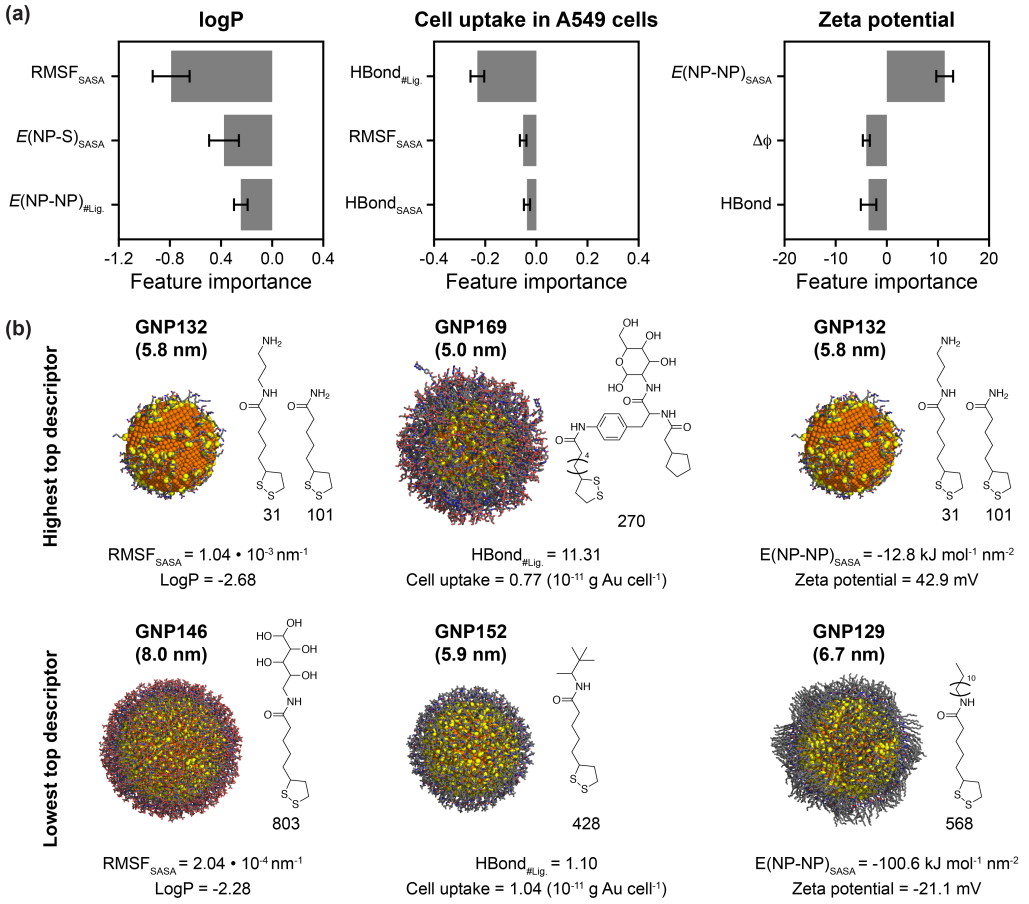


Figure 11.5: Feature importance from QNAR models. (a) Top three most important descriptors are shown for logP (left), cell uptake in A549 cells (middle), and zeta potential in water (right) are shown for random forest models. Feature importance was quantified by Shapley values, where the average magnitude of Shapley values (*i.e.* Mean |Shap|) are reported and the sign is determined by the Pearson's r correlation between Shapley and descriptor values. The sign of the feature importance indicates whether the feature positively or negatively impact the experimental observables. Feature importance was computed by training the random forest model using 90% of the training data (randomly sampling without replacement) and iterated for 10 trials; the average of trials is reported and the error is reported as the standard deviation of the trials. (b) Simulation snapshots of GNPs with the highest and lowest top descriptor outputted from the feature importance in (a). Core diameters are shown in parenthesis; number and structure of ligands are shown to the right of the snapshots; and, descriptor and experimental measurements are shown below the snapshots. GNP169 has a mixed-monolayer, but only the structure of the majority ligand is shown for brevity.

11.4 Model transferability to new datasets

We next sought to test the model prediction capabilities highlighted in Figure 11.4 to new datasets that are outside of the experimentally database of Ref. 37. We focus on testing the generalizability of the cell uptake model because cell uptake is the most biologically relevant observable and most useful in terms of designing safe, effective GNPs. Jiang *et al.* measured the cell uptake of 12 GNPs with core diameters of 2, 4, 6 nm, coated by thioalkyl tetra(ethyleneglycol)ated structures with four distinct end groups shown in Figure 11.6a.[19] Trimethylammonium (TTMA) and carboxylate (COO) ligands were representative cationic and anionic ligands, respectively; NS and SN ligands are zwitterionic ligands. Figure 11.6b shows the cell uptake of these 12 GNPs in human cervical carcinoma (HeLa) cells; all experimental data were taken from Ref. 19 and converted to the cell uptake units in used for cell uptake models in Figure 11.4. Cationic GNPs with TTMA ligands were found to have higher cell uptake values compared to the other GNPs, and increasing the GNP size from 2 nm to 6 nm generally increases the extent of cell uptake, especially for TTMA-coated GNPs.

Given that the RF model performed well in correlating MD-derived descriptors to cell uptake in A549 cells (Figure 11.4), we sought to see if a RF model could predict these trends after training it with 65 cell uptake in A549 labels (Figure 11.1c). We note that there are several discrepancies in the experimental cell uptake measurements, such as differences in cell

line (A549 cells in Ref. 37 *versus* HeLa cells in Ref. 19) or initial GNP concentration (50 μg/mL in Ref. 37 *versus* ≈ 1.2 μg/mL in Ref. 19). Hence, we do not expect the models trained with labels from Figure 11.1c to quantitatively predict cell uptake values from Figure 11.6b; rather, the trained RF model should capture qualitative cell uptake trends.

We trained a RF model using all 65 experimental labels from Figure 11.1c as outputs and the corresponding GNPs with uncorrelated descriptors as inputs. We then performed MD simulations of the 12 GNPs from Ref. 19 using the same workflow and descriptor calculations described in Figures 11.2 and Figure 11.3, we used uncorrelated descriptors to predict cell uptake trends. Figure 11.6c shows the predicted *versus* experimental cell uptake for the 12 GNPs using the trained RF model, which had a high Pearson's r between predicted and experimental values of 0.85. The RF model predicted that the 6 nm GNP with TTMA ligands (simulation snapshot is shown in Figure 11.6c) had the highest cell uptake, consistent with experimental values in Figure 11.6b. One striking feature is that the RF model can capture experimental trends despite differences in experimental environments (*e.g.* cell lines), suggesting that GNP features related to cell uptake (*e.g.* surface hydrogen bonding) may be fundamental characteristics associated with GNP translocation across or adsorption onto lipid membranes.

We have also tested the prediction of LASSO model on these 12 GNPs (Figure S7),[2] which performed poorer than the RF model with a Pearon's

r between predicted and experimental cell uptake of 0.26. These findings suggest that RF models outperform LASSO models in predicting cell uptake values for unseen GNPs, which is consistent with the better prediction performance of RF models in the training and testing data (Figure 11.4). To ensure robustness of the transferrability of cell uptake models, we predicted cell uptake trends for these 12 GNPs using a second trial with a shorter GNP-water simulations of 20 ns as described before, and we found similar prediction trends (Figure S11).[2]

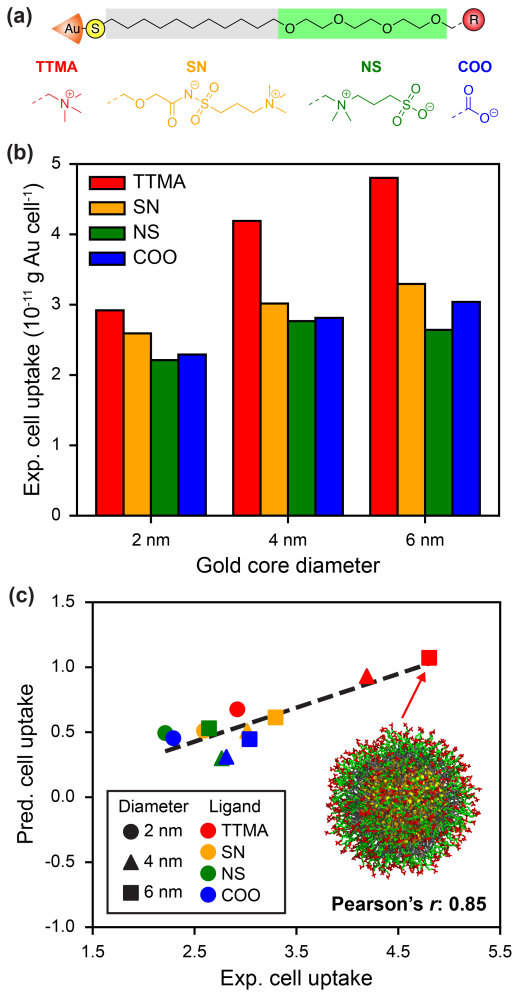


Figure 11.6: Transferability of the cell uptake model to new datasets. (a) Thioalkyl tetra(ethyleneglycol)ated structures with four end group chemistries. (b) Cell uptake of 12 GNPs with core diameters of 2, 4, and 6 nm, and ligand structures from (a). Cell uptake measurements were performed after GNPs were incubated with HeLa cells for 3 hours and quantified using inductively coupled plasma mass spectrometry; all experimental data were taken from Ref. 19. The cell uptake values were converted to have the same cell uptake units as Ref. 37. (c) Predicted *versus* experimental cell uptake values using MD-derived descriptors of the 12 GNPs and the data from (b) as the experimental data. Predicted values were computed using a RF model trained with 15 uncorrelated descriptors as inputs and 65 cell uptake labels from Figure 11.1c as output. Pearon's r between predicted and experimental values using all 12 GNPs are shown in the lower right. Black dashed lines show the best fit line as a guide. The simulation snapshot show a 6 nm GNP with TTMA ligands.

11.5 Summary

In this chapter, we modeled 154 SAM-coated GNPs in aqueous solution using atomistic MD simulations, developed a library of simulation-derived descriptors, and leveraged data-centric models (*e.g.* LASSO and RF) to predict GNP observables (*i.e.* log P, cell uptake, and zeta potential) from an experimentally curated dataset. By modeling GNPs with MD, the interplay between the gold core and ligand selection that affect monolayer characteristics is captured, specifically the non-obvious, cooperative interactions arising from ligand-ligand and ligand-water interactions. These effects are encoded in the form of simulation-derived descriptors that capture structural and chemical properties of the monolayer, which are difficult to estimate *a priori* with the gold core and ligand structure or static models alone. We found that LASSO and RF models performed well in predicting experimental measurements using simulation-derived descriptors as inputs, which suggests that these models could be used to screen GNPs more broadly. Furthermore, by analyzing feature importance of these interpretable models, we found a consensus between the models of the most important descriptors related to experimental measurements, such as the importance of ligand fluctuations for logP measurements and the importance of surface hydrogen bonding for cell uptake measurements. Finally, we tested the transferability of the cell uptake models to 12 unseen GNPs and found that the RF model can accurately predict cell uptake

trends, despite differences in the experimental methods for measuring cell uptake between the trained and unseen GNPs (*e.g.* different cell line, initial GNP concentration, *etc.*).

Altogether, these results reveal that simulation-derived molecular descriptors could be used to rank-order GNPs for selective properties, which would aid in screening GNPs with computationally efficient tools towards guiding experimental design. The added benefit of modeling GNPs in solution with MD is that their behavior with biomolecules could be further interrogated by modeling GNP in presence of lipid bilayers[11,12,16,25,72–74] or proteins,[75] enabling a bottom-up modeling approach towards identifying promising GNPs. Future research will focus on leveraging these workflows to predict the how the selection of gold core and ligand influence the formation of protein coronas, which is the strong adsorption of protein upon introducing GNPs in physiological environments. For instance, zwitterionic ligands were found to prevent strong protein adsorption (*i.e.* hard corona) at physiological serum concentrations,[17] but it is unclear how the monolayer structure prevents hard corona formation. Additionally, the expansive simulation data available from this work would enable deep learning approaches that may improve predictions of experimental observables with MD.

11.6 References

[1] Chew, A. K.; Pedersen, J. A.; Van Lehn, R. C. Predicting Nano-bio Activities of Monolayer-Protected Gold Nanoparticles using Molecular Dynamics-Derived Descriptors. *In preparation*.

[2] Chew, A. K.; Pedersen, J. A.; Van Lehn, R. C. Predicting Nano-bio Activities of Monolayer-Protected Gold Nanoparticles using Molecular Dynamics-Derived Descriptors [Supporting Information]. *In preparation*.

[3] Ghosh, P.; Han, G.; De, M.; Kim, C. K.; Rotello, V. M. Gold nanoparticles in delivery applications. *Advanced drug delivery reviews* **2008**, *60*, 1307–1315.

[4] Jans, H.; Huo, Q. Gold nanoparticle-enabled biological and chemical detection and analysis. *Chemical Society Reviews* **2012**, *41*, 2849–2866.

[5] Riley, R. S.; Day, E. S. Gold nanoparticle-mediated photothermal therapy: applications and opportunities for multimodal cancer treatment. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **2017**, *9*, e1449.

[6] Longmire, M.; Choyke, P. L.; Kobayashi, H. Clearance properties of nano-sized particles and molecules as imaging agents: considerations and caveats. **2008**.

[7] You, C.-C.; De, M.; Rotello, V. M. Monolayer-protected nanoparticle–protein interactions. *Current opinion in chemical biology* **2005**, *9*, 639–646.

[8] 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*, 3965–3967.

[9] Chen, K.; Rana, S.; Moyano, D. F.; Xu, Y.; Guo, X.; Rotello, V. M. Optimizing the selective recognition of protein isoforms through tuning of nanoparticle hydrophobicity. *Nanoscale* **2014**, *6*, 6492–6495.

[10] Sun, S.; Huang, Y.; Zhou, C.; Chen, S.; Yu, M.; Liu, J.; Zheng, J. Effect of hydrophobicity on nano-bio interactions of zwitterionic luminescent gold nanoparticles at the cellular level. *Bioconjugate chemistry* **2018**, *29*, 1841–1846.

[11] Lochbaum, C. A.; Chew, A. K.; Zhang, X.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A. Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers. *ACS nano* **2021**.

[12] Lolicato, F.; Joly, L.; Martinez-Seara, H.; Fragneto, G.; Scoppola, E.; Baldelli Bombelli, F.; Vattulainen, I.; Akola, J.; Maccarini, M. The role of temperature and lipid charge on intake/uptake of cationic gold nanoparticles into lipid bilayers. *Small* **2019**, *15*, 1805046.

[13] Tatur, S.; Maccarini, M.; Barker, R.; Nelson, A.; Fragneto, G. Effect of functionalized gold nanoparticles on floating lipid bilayers. *Langmuir* **2013**, *29*, 6606–6614.

[14] Chong, G.; Foreman-Ortiz, I. U.; Wu, M.; Bautista, A.; Murphy, C. J.; Pedersen, J. A.; Hernandez, R. Defects in Self-Assembled Monolayers on Nanoparticles Prompt Phospholipid Extraction and Bilayer-Curvature-Dependent Deformations. *The Journal of Physical Chemistry C* **2019**, *123*, 27951–27958.

[15] Heikkilä, E.; Martinez-Seara, H.; Gurtovenko, A. A.; Javanainen, M.; Häkkinen, H.; Vattulainen, I.; Akola, J. Cationic Au nanoparticle binding with plasma membrane-like lipid bilayers: potential mechanism for spontaneous permeation to cells revealed by atomistic simulations. *The Journal of Physical Chemistry C* **2014**, *118*, 11131–11141.

[16] Lin, J.; Zhang, H.; Chen, Z.; Zheng, Y. Penetration of lipid membranes by gold nanoparticles: insights into cellular uptake, cytotoxicity, and their relationship. *ACS nano* **2010**, *4*, 5421–5429.

[17] 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*, 6748–6755.

[18] Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. Understanding

biophysicochemical interactions at the nano–bio interface. *Nature materials* **2009**, *8*, 543–557.

[19] Jiang, Y.; Huo, S.; Mizuhara, T.; Das, R.; Lee, Y.-W.; Hou, S.; Moyano, D. F.; Duncan, B.; Liang, X.-J.; Rotello, V. M. The interplay of size and surface functionality on the cellular uptake of sub-10 nm gold nanoparticles. *ACS nano* **2015**, *9*, 9986–9993.

[20] Saha, K.; Rahimi, M.; Yazdani, M.; Kim, S. T.; Moyano, D. F.; Hou, S.; Das, R.; Mout, R.; Rezaee, F.; Mahmoudi, M.; et al.. Regulation of macrophage recognition through the interplay of nanoparticle surface functionality and protein corona. *ACS nano* **2016**, *10*, 4421–4430.

[21] Li, X.; Robinson, S. M.; Gupta, A.; Saha, K.; Jiang, Z.; 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*, 10682–10686.

[22] Melby, E. S.; Lohse, S. E.; Park, J. E.; Vartanian, A. M.; Putans, R. A.; Abbott, H. B.; Hamers, R. J.; Murphy, C. J.; Pedersen, J. A. Cascading effects of nanoparticle coatings: Surface functionalization dictates the assemblage of complexed proteins and subsequent interaction with model cell membranes. *ACS nano* **2017**, *11*, 5489–5499.

[23] Yu, Q.; Zhao, L.; Guo, C.; Yan, B.; Su, G. Regulating protein corona formation and dynamic protein exchange by controlling nanoparticle hydrophobicity. *Frontiers in bioengineering and biotechnology* **2020**, *8*, 210.

[24] Bai, X.; Wang, S.; Yan, X.; Zhou, H.; Zhan, J.; Liu, S.; Sharma, V. K.; Jiang, G.; Zhu, H.; Yan, B. Regulation of cell uptake and cytotoxicity by nanoparticle core under the controlled shape, size, and surface chemistries. *ACS nano* **2019**, *14*, 289–302.

[25] Contini, C.; Schneemilch, M.; Gaisford, S.; Quirke, N. Nanoparticle–membrane interactions. *Journal of Experimental Nanoscience* **2018**, *13*, 62–81.

[26] Rossi, G.; Monticelli, L. Gold nanoparticles in model biological membranes: A computational perspective. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2016**, *1858*, 2380–2389.

[27] Rossi, G.; Monticelli, L. Simulating the interaction of lipid membranes with polymer and ligand-coated nanoparticles. *Advances in Physics: X* **2016**, *1*, 276–296.

[28] Kim, S. T.; Saha, K.; Kim, C.; Rotello, V. M. The role of surface functionality in determining nanoparticle cytotoxicity. *Accounts of chemical research* **2013**, *46*, 681–691.

[29] Bunker, A.; Magarkar, A.; Viitala, T. Rational design of liposomal drug delivery systems, a review: combined experimental and computational studies of lipid membranes, liposomes and their PEGylation. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2016**, *1858*, 2334–2352.

[30] Pengo, P.; Şologan, M.; Pasquato, L.; Guida, F.; Pacor, S.; Tossi, A.; Stellacci, F.; Marson, D.; Boccardo, S.; Pricl, S.; et al.. Gold nanoparticles with patterned surface monolayers for nanomedicine: current perspectives. *European Biophysics Journal* **2017**, *46*, 749–771.

[31] Fourches, D.; Pu, D.; Tropsha, A. Exploring quantitative nanostructure-activity relationships (QNAR) modeling as a tool for predicting biological effects of manufactured nanoparticles. *Combinatorial Chemistry & High Throughput Screening* **2011**, *14*, 217–225.

[32] Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. Quantitative nanostructure- activity relationship modeling. *ACS nano* **2010**, *4*, 5703–5712.

[33] Singh, K. P.; Gupta, S. Nano-QSAR modeling for predicting biological activity of diverse nanomaterials. *RSC Advances* **2014**, *4*, 13215–13230.

[34] Wang, W.; Sedykh, A.; Sun, H.; Zhao, L.; Russo, D. P.; Zhou, H.; Yan, B.; Zhu, H. Predicting nano–bio interactions by integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS nano* **2017**, *11*, 12641–12649.

[35] Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* **2019**, *11*, 8352–8362.

[36] Wang, W.; Yan, X.; Zhao, L.; Russo, D. P.; Wang, S.; Liu, Y.; Sedykh, A.; Zhao, X.; Yan, B.; Zhu, H. Universal nanohydrophobicity predictions using virtual nanoparticle library. *Journal of cheminformatics* **2019**, *11*, 1–5.

[37] Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nature communications* **2020**, *11*, 1–10.

[38] Yan, X.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B. Prediction of Nano–Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images. *ACS Sustainable Chemistry & Engineering* **2020**.

[39] Kister, T.; Monego, D.; Mulvaney, P.; Widmer-Cooper, A.; Kraus, T. Colloidal stability of apolar nanoparticles: the role of particle size and ligand shell structure. *ACS nano* **2018**, *12*, 5969–5977.

[40] Matthew D áLane, J.; et al.. Assembly of responsive-shape coated nanoparticles at water surfaces. *Nanoscale* **2014**, *6*, 5132–5137.

[41] Chew, A. K.; Van Lehn, R. C. Effect of core morphology on the structural asymmetry of alkanethiol monolayer-protected gold nanoparticles. *The Journal of Physical Chemistry C* **2018**, *122*, 26288–26297.

[42] Ghorai, P. K.; Glotzer, S. C. Molecular dynamics simulation study of self-assembled monolayers of alkanethiol surfactants on spherical gold nanoparticles. *The Journal of Physical Chemistry C* **2007**, *111*, 15857–15862.

[43] Lane, J. M. D.; Grest, G. S. Spontaneous asymmetry of coated spherical nanoparticles in solution and at liquid-vapor interfaces. *Physical review letters* **2010**, *104*, 235501.

[44] Luedtke, W.; Landman, U. Structure, dynamics, and thermodynamics of passivated gold nanocrystallites and their assemblies. *The Journal of Physical Chemistry* **1996**, *100*, 13323–13329.

[45] Chew, A. K.; Dallin, B. C.; Van Lehn, R. C. The Interplay of Ligand Properties and Core Size Dictates the Hydrophobicity of Monolayer-Protected Gold Nanoparticles. *ACS nano* **2021**, *15*, 4534–4545.

[46] Koch, A. H.; Lévêque, G.; Harms, S.; Jaskiewicz, K.; Bernhardt, M.; Henkel, A.; Sönnichsen, C.; Landfester, K.; Fytas, G. Surface asymmetry of coated spherical nanoparticles. *Nano letters* **2014**, *14*, 4138–4144.

[47] Nash, J. A.; Tucker, T. L.; Therriault, W.; Yingling, Y. G. Binding of single stranded nucleic acids to cationic ligand functionalized gold nanoparticles. *Biointerphases* **2016**, *11*, 04B305.

[48] Sridhar, D. B.; Gupta, R.; Rai, B. Effect of surface coverage and chemistry on self-assembly of monolayer protected gold nanoparticles: a molecular dynamics simulation study. *Physical Chemistry Chemical Physics* **2018**, *20*, 25883–25891.

[49] Huang, J.; MacKerell Jr, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, *34*, 2135–2145.

[50] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; et al.. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* **2010**, *31*, 671–690.

[51] Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell Jr, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry* **2012**, *33*, 2451–2468.

[52] Heinz, H.; Vaia, R.; Farmer, B.; Naik, R. Accurate simulation of surfaces and interfaces of face-centered cubic metals using 12- 6 and 9- 6 Lennard-Jones potentials. *The Journal of Physical Chemistry C* **2008**, *112*, 17281–17290.

[53] Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *The Journal of chemical physics* **1996**, *105*, 1902–1921.

[54] Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry* **1997**, *18*, 1463–1472.

[55] Páll, S.; Abraham, M. J.; Kutzner, C.; Hess, B.; Lindahl, E. In *2nd International Conference on Exascale Applications and Software (EASC), APR 02-03, 2014, Stockholm, SWEDEN*; Springer Publishing Company; pp 3–27.

[56] McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophysical journal* **2015**, *109*, 1528–1532.

[57] Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D. L.; Buchoux, S.; Kenney, I. M.; et al.; *MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations*; Tech. Rep.; Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2019.

[58] Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry* **2011**, *32*, 2319–2327.

[59] Lundberg, S.; Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* **2017**.

[60] Lundberg, S.; Lee, S.-I. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478* **2016**.

[61] Předota, M.; Machesky, M. L.; Wesolowski, D. J. Molecular origins of the zeta potential. *Langmuir* **2016**, *32*, 10189–10198.

[62] Ivanov, M. R.; Bednar, H. R.; Haes, A. J. Investigations of the mechanism of gold nanoparticle stability and surface functionalization in capillary electrophoresis. *ACS nano* **2009**, *3*, 386–394.

[63] Djebaili, T.; Richardi, J.; Abel, S.; Marchi, M. Atomistic simulations of the surface coverage of large gold nanocrystals. *The Journal of Physical Chemistry C* **2013**, *117*, 17791–17800.

[64] Heikkilä, E.; Gurtovenko, A. A.; Martinez-Seara, H.; Häkkinen, H.; Vattulainen, I.; Akola, J. Atomistic simulations of functional Au144 (SR) 60 gold nanoparticles in aqueous environment. *The Journal of Physical Chemistry C* **2012**, *116*, 9805–9815.

[65] Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Perez-Sanchez, H.; Benediktsson, J. A.; Thapa, A.; Barr, A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications* **2017**, *72*, 151–159.

[66] Rodríguez-Pérez, R.; Bajorath, J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of computer-aided molecular design* **2020**, *34*, 1013–1026.

[67] Boehmke, B.; Greenwell, B. M. *Hands-on machine learning with R*; CRC Press, 2019.

[68] An, Y.; Sherman, W.; Dixon, S. L. Kernel-based partial least squares: application to fingerprint-based QSAR with model visualization. *Journal of chemical information and modeling* **2013**, *53*, 2312–2321.

[69] Jing, X.; Yang, M.; Kasimova, M. R.; Malmsten, M.; Franzyk, H.; Jorgensen, L.; Foged, C.; Nielsen, H. M. Membrane adsorption and binding, cellular uptake and cytotoxicity of cell-penetrating peptidomimetics with α-peptide/β-peptoid backbone: Effects of hydrogen bonding and α-chirality in the β-peptoid residues. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2012**, *1818*, 2660–2668.

[70] Vinothini, K.; Rajan, M. In *Characterization and Biology of Nanomaterials for Drug Delivery*; Mohapatra, S. S.; Ranjan, S.; Dasgupta, N.; Mishra, R. K.; Thomas, S., Eds.; Micro and Nano Technologies; Elsevier, 2019; pp 219–263.

[71] Dallin, B. C.; Van Lehn, R. C. Spatially heterogeneous water properties at disordered surfaces decrease the hydrophobicity of nonpolar self-assembled monolayers. *The journal of physical chemistry letters* **2019**, *10*, 3991–3997.

[72] Nakamura, H.; Sezawa, K.; Hata, M.; Ohsaki, S.; Watano, S. Direct translocation of nanoparticles across a model cell membrane by nanoparticle-induced local enhancement of membrane potential. *Physical Chemistry Chemical Physics* **2019**, *21*, 18830–18838.

[73] 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 letters* **2013**, *13*, 4060–4067.

[74] Van Lehn, R. C.; Ricci, M.; Silva, P. H.; Andreozzi, P.; Reguera, J.; Voïtchovsky, K.; Stellacci, F.; Alexander-Katz, A. Lipid tail protrusions mediate the insertion of nanoparticles into model cell membranes. *Nature communications* **2014**, *5*, 1–11.

[75] Simonelli, F.; Rossi, G.; Monticelli, L. Role of Ligand Conformation on Nanoparticle–Protein Interactions. *The Journal of Physical Chemistry B* **2019**, *123*, 1764–1769.

12 CONCLUSION AND FUTURE RESEARCH

This chapter highlights the main takeaways from this dissertation and suggests potential avenues of future research, which is divided into two sections for each application area.

12.1 Solvent screening for biomass conversion reactions

Chapters 3-7 formulated a computational framework that enables solvent-screening for biomass-relevant reactions using classical MD simulations and machine learning tools. The most notable takeaways are:

- MD-derived descriptors correlated with experimental reaction rates (Chapters 3) and selectivity (Chapter 6), despite not explicitly modeling the catalyst or reaction mechanism.
- Modeling the acid catalyst could yield physical understanding into how the catalyst might behave in different solvent environments, which may improve prediction capabilities of reactivity (Chapter 4).
- Deep learning models, such as convolutional neural networks (*e.g.* SolventNet), could improve and accelerate reaction rate predictions made from MD (Chapter 5). In addition, a fully trained SolventNet could be analyzed to inform us about the important characteristics from MD that relates to reaction rates, which could potentially guide a descriptor-based approach or yield a deeper understanding of solvent effects. The approach of integrating deep learning models

to autonomously extract features from MD and make predictions of experimental observables could accelerate model generation for a large range of applications.

- By merging SolventNet (Chapter 5) and solvation free energy calculations (Chapter 6), we now have a computational workflow that could down-select solvents for biomass conversion processes using a combination of MD and machine learning methods, which could be used to lower the extent of experiments required to identify good solvent compositions (Chapter 7).

While these computational tools provide us with a way of predicting solvent-mediated effects on acid-catalyzed reactions, there are also a number of challenges that could to be addressed in future work:

- The model uses lengthy solvation free energy calculations to predict trends in selectivity. Given the success of 3D CNNs to predict reactivity (Chapter 5), one could envision a similar approach performed to predict product selectivity, which motivates the future direction described in Section 12.1.1.

- The model does not incorporate costs associated with downstream separation, costs of the solvents, or toxicity of solvents. For example, acetone-water mixtures have been found to produce high yields of 5-hydroxylmethylfurfural (HMF), and HMF could be easily separated from the mixture in downstream processes.[1] Thus, this computational workflow could be improved by integrating process models / techno-economic analysis approaches to guide solvent selection, which motivates the future direction described in Section 12.1.2.

- Reaction temperature and catalyst selection was not heavily probed by the models, which are both important parameters for reactor de-

sign. For the acid-catalyzed reactions, we primarily used triflic acid, which is a strong acid that disassociates readily in solution. However, weaker acids with low disassociation constants would influence the reactivity significantly; hence, modeling the acid catalyst and its conjugate base may be useful for predicting effects on reactivity (Chapter 4).

- The models primarily used small biomass-derived compounds. However, biomass consists of large macromolecules, such as lignin or cellulose. Hence, incorporating larger compounds into the model would enable flexibility of the model for degrading real biomass material.
- The computational workflow may miss important solvent-mediated interactions that are not captured using classical MD simulations. Chapter 6 highlights this issue, where DMSO-water mixtures change the product selectivity of 1,2-propanediol dehydration reaction due to the presence of DMSO. Therefore, the workflow could be improved by incorporating *ab initio* calculations that may better capture the influence of solvents in acid-catalyzed reaction mechanisms.

12.1.1 Future work 1: Incorporating product states into convolutional neural networks

Given the success of SolventNet in predicting acid-catalyzed reaction rates from MD simulations of a single reactant in various solvent environments (Chapter 5), incorporating information about the product state may enable rapid prediction of selectivities in parallel reactions. Since selectivity

information may be limited in the literature, we could focus on predicting $\Delta\Delta G$, which is computed from a series of solvation free energies between reactant and product states and was found to correlate with reaction selectivities (Chapter 6). One major bottleneck of computing $\Delta\Delta G$ is that it is computationally expensive to perform (*e.g.* > 12 hours on a supercomputer). We could potentially speed up selectivity predictions by combining workflows from Chapters 5 and 6. Figure 12.1 depicts a deep learning framework towards combining molecular information from reactant (*e.g.* 1,2-propanediol) and product (*e.g.* propane) states to predict $\Delta\Delta G$. If this approach is successful, the ability of SolventNet to predict reactivity and selectivity would provide powerful tools to rapidly screen solvent compositions for biomass conversion reactions. Inclusion of catalyst simulations might also improve the generalizability of this deep learning framework to enable the screening of catalysts. In addition, exploring different machine learning models (*e.g.* graph-convolutional neural networks)[2] may improve prediction accuracy by encoding important chemical information missed in the SolventNet approach, such as hydrogen bonding or orientation of the hydroxyl groups relative to the solvent.

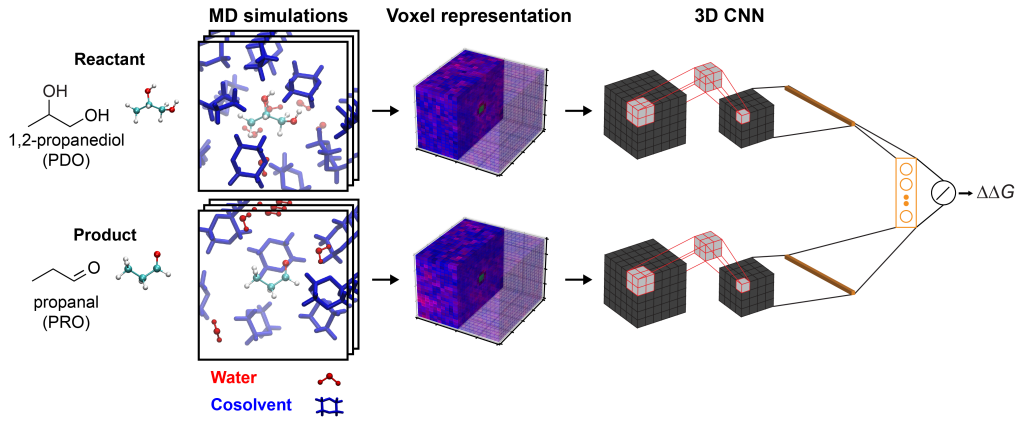


Figure 12.1: Future work on incorporating product states into a deep learning framework to predicting selectivities. MD simulations are performed on both reactant and product states. Then, simulations are transformed into voxel representations, as described previously in Chapter 5. The voxel representations are then inputted into 3D convolutional neural networks and merged into fully-connected layers with the regression task of predicting ΔΔG.

12.1.2 Future work 2: Integrating process models to fine-tune solvent selection

While we could use the computational tools in this dissertation to predict reactivity and selectivity in biomass conversion reaction, these tools alone do not suggest a good reaction workflow for converting biomass to useful chemicals. For instance, while a specific cosolvent-water mixture might have high reactivity, these cosolvents may be expensive, toxic, or difficult to separate. As a result, recent literature have focused on identifying green solvents that have lower toxicity and possibly lower cost.[3] These

solvent factors and downstream separation processes would be useful for designing a series of biomass conversion reactors. Hence, future work could focus on integrating molecular-level modeling and process modeling tools (*e.g.* ASPEN),[4,5] where MD simulations/machine learning could be used to rapidly predict reactivity/selectivity and process models could be used to inform cost expenses and challenges in downstream separation processes.

12.2 Screening monolayer-protected gold nanoparticle properties for biomedical applications

Chapters 8-11 provide computational tools to model monolayer-protected gold nanoparticle (GNP) systems using classical MD simulations and the analysis of these systems enables the prediction of GNP behavior with other biomolecules. The main takeaways of this work are:

- A generalized workflow to build GNP systems that accounts for gold core and ligand selection was developed (Chapter 8).
- Cooperative interactions arising from ligand-ligand or ligand-environment interactions necessitate the use of MD simulations to more accurately describe monolayer characteristics. These cooperative interactions (*e.g.* formation of bundles) could be quantified using clustering algo-

rithms (Chapter 8) or surface characterization techniques (Chapter 10) and distilled in the form of molecular descriptors (Chapter 11).

- MD-derived molecular descriptors and machine learning models could be used to predict GNP behavior with other biomolecules (Chapter 11), which enables a rational design and screening of GNPs for selective behavior.
- Given MD simulations of GNP in solution, subsequent GNP-bilayer (Chapter 9) or GNP-protein interactions could be further interrogated, which could yield physical insight into how gold core or ligand selection might affect GNP interactions with biomolecules.

These computational tools provide a good step towards an *in silico* workflow for designing better GNPs, but these models also face many challenges:

- The model is missing the ability to predict protein corona formation. Given that the first challenge that a GNP faces in a biological environment is the adsorption of proteins on the GNP surface (discussed in Section 1.3.4), we need to understand how gold core and ligand selection affect the distribution of proteins adsorbed onto the surface, motivating future research in Section 12.2.1.
- The model uses physically motivated molecular descriptors to predict GNP behavior (Chapter 11), but human-selected descriptors may miss important characteristics that may better describe the GNP system. As a result, one may envision that a deep learning model (*e.g.* SolventNet in Chapter 5) may capture complex features that could better predict GNP behavior, motivating future research in Section 12.2.2.

- The models could generalize to nanoparticle behavior that has already been observed, but they have yet to be used to inform an experimentalist whether a GNP is worthwhile to synthesize. Hence, we need an active learning framework to suggest new experiments of gold nanoparticles that would exhibit characteristics distinct from the available training data, motivating future research in Section 12.2.3.
- Given that GNP systems are large (*e.g.* > 100 K atoms), these atomistic models may face a computational bottleneck. This issue could be resolved with hardware improvements (*e.g.* Anton machines designed for MD simulations),[6] coarse-grained modeling,[7,8] enhanced sampling techniques (*e.g.* umbrella sampling simulations in Chapter 9), or more sophisticated analysis procedures (*e.g.* deep learning models in Chapter 5).

12.2.1 Future work 1: Develop model to predict protein corona formation around GNPs

Protein adsorption onto the GNP surface, forming a "protein corona," is a highly relevant issue in the design of GNPs for biomedical applications because protein corona formation effectively gives the GNP a new "biological identity" with surface properties that are different from bare GNPs.[9–12] A suggested future direction is the development of computational tools to predict protein adsorption onto GNP. A recent experimental study quantified the extent of protein adsorption on planar SAM surfaces,[12] which is a

good starting point tuning a computational workflow to predict protein adsorption propensity on planar systems. One way to estimate protein binding propensity to the SAM surface *in silico* is by performing umbrella sampling (US) simulations, which measures the free energy of moving a molecule along a reaction coordinate. Figure 12.2(a) shows a schematic of a GNP-protein US simulation with z as the reaction coordinate, defined as the center-of-mass distance between the gold core and protein (denoted as "P"). For small GNPs (<10 nm), we could potentially speed up the simulation by only considering half the GNP, which would still account for curvature effects of the GNP. Figure 12.2(b) shows the expected US simulation and QCM results for non-binding, weak, and strong proteins. US simulations should provide a direct comparison to QCM measurements, which would be a good validation that the simulations is capturing experimental trends.

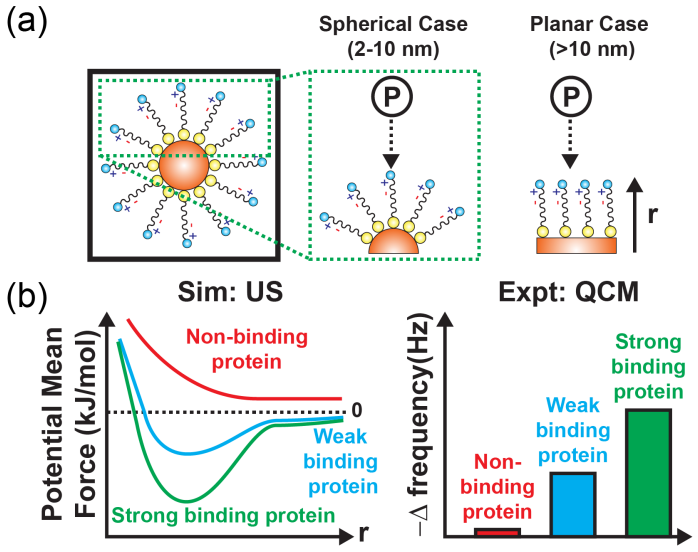


Figure 12.2: Future work on integrating US simulations and QCM measurements to measure protein binding affinities on planar SAMs and GNPs. (a) Reaction coordinate of US simulations between GNP and protein (marked as "P"), where the protein is sampled at a distance away from either a curved GNP surface (2-10 nm in core diameter) or planar SAM (estimated for >10 nm core diameters). (b) Expected US simulation and QCM results for a non-binding, weak, and strong protein.

One challenge in US simulations is that they are computationally expensive (>1 day for a single SAM-protein combination on a high-performance cluster), and their results are largely dependent in the initialization of the simulation. Therefore, a single umbrella sampling calculation might not account for the different rotations of the protein and conformations of the ligand that might affect the binding affinity of a protein to a SAM surface. We could tackle this by using enhanced sampling techniques with a different reaction coordinate (*e.g.* hydrophobic contacts reaction

coordinate in Chapter 9), but identifying the best reaction coordinate is tedious and may not translate for different SAM-protein combinations.

An alternative strategy to measure the free energy for introducing a molecule at an interface is by performing cosolvent simulations, which consists of simulations that include water and cosolvent molecules surrounding a protein or SAM.[13] The cosolvent molecules act as molecular probes that can inform the locations of hydrophilic or charged locations, depending on the chemical property of the probe. A good example of molecular probes relevant to natural amino acids is described in Ref. 14. One major benefit of using cosolvent simulations is the reduced computational expense of measuring surface properties of across the entire surface of a GNP, which may extrapolate to binding propensities to different proteins and accounts for all possible rotations. We have previously used cosolvent simulations to estimate preferred binding sites of hydrophobic probes (*e.g.* propane) around GNPs in Chapter 10.

Figure 12.3 shows a computational workflow for using cosolvent simulations to estimate protein binding propensity. Figure 12.3(a) shows 5 distinct cosolvents that encompass polar, nonpolar, and charged groups. These cosolvents were selected for their amino acid analogies, and they have been used in previous literature for cosolvent mapping simulations.[13] Figure 12.3(b) displays a computational workflow for developing cosolvent simulations by first initiating the cosolvents around the nanoparticle, performing a high temperature *NVT* simulation at T = 600 K to remove

cosolvent aggregation[15] and allow for cosolvents to explore the GNP surface, and finally, a *NPT* simulation at T = 300 K and P = 1 bar to allow for cosolvents to bind to the surface. These simulations output the fraction of simulation time occupied by each cosolvent as a surface map illustrated in 12.3(c). Low occupancies are shown as white color, whereas high occupancies are shown by the color-codes of each cosolvent species. With these occupancy maps, we could begin to correlate GNP binding with proteins observed in experiments for small 2 nm GNP systems.[11] By developing a protein corona prediction model from either US simulations or cosolvent simulations, we could begin screening GNPs for selective protein adsorption, which could then help us predict GNP behavior in the biological environment.

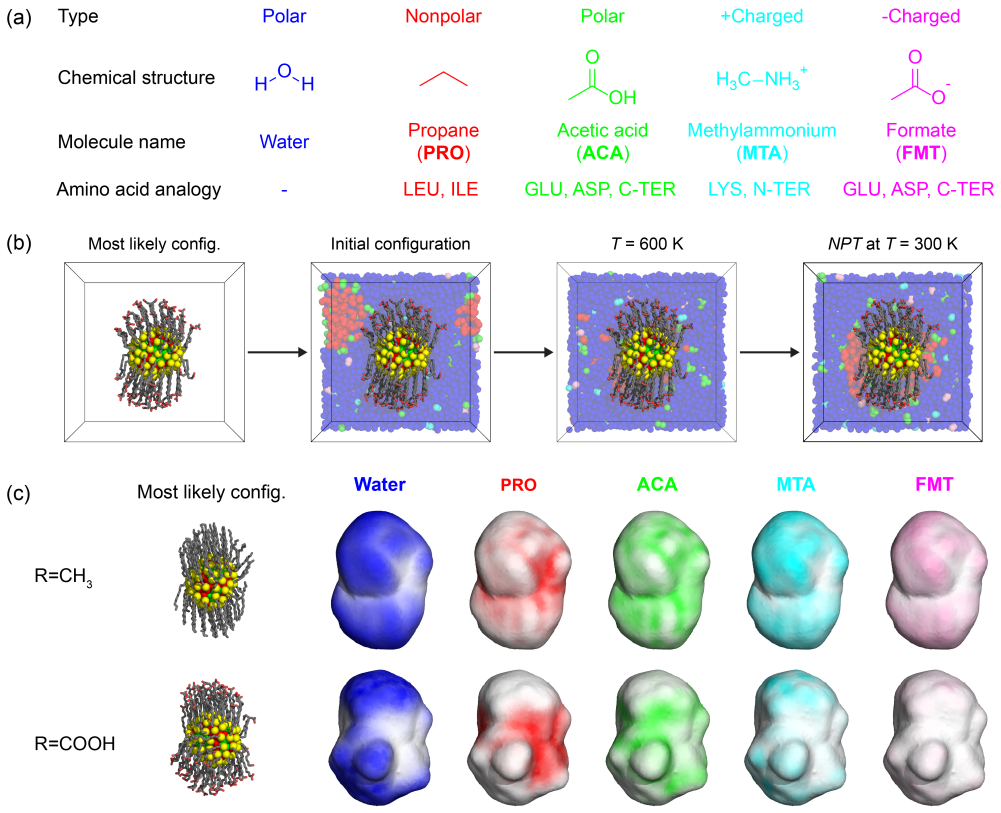


Figure 12.3: Future work on leveraging cosolvent simulations to predict protein binding sites. (a) Five cosolvents that could be used for cosolvent simulations, which encompass polar, nonpolar, and charged species: water, propane, acetic acid, methylammonium, and formate. Each species (except water) has an amino acid analogue so that these simulations could suggest binding propensities of amino acids located on the protein surface. (b) Simulation workflow of taking the most likely configuration of a GNP (described in Chapter 10), then performing a high temperature simulation at T = 600 K, followed by an *NPT* production simulation at T = 300 K. Cosolvent molecules are color-coded as the same color shown in (a). (c) Occupancy maps of the 5 solvents from (a) for a 2 nm GNP with CH_3- and COOH-terminated ligand end groups. White colors mean no occupancy of the cosolvent, whereas colored surfaces mean high occupancy of the cosolvent.

12.2.2 Future work 2: Incorporate deep learning models to rapidly predict GNP properties

Given that quantifying the hydrophobicity of GNPs correlated with preferential binding of hydrophobic moeties like propane (Chapter 10), one could envision that surface maps like the hydration maps in Figure 10.3 encode important information that could be predictive of GNP behavior with other biomolecules. Recent literature has integrated surface maps that encode hydrophobicity, electrostatics, and geometric features with deep learning techniques (*e.g.* convolutional neural networks (CNN)) to obtain difficult-to-predict protein properties, such as protein-protein binding.[16] This approach is a promising method for automatically extracting molecular features from surface information directly and relating them to experimental observables, overcoming extensive time required to develop molecular descriptors by domain experts. Future work could focus on using deep learning frameworks to improve prediction accuracy of MD-derived molecular descriptors from Chapter 11. Figure 12.4 summarizes the use of deep learning models, such as CNNs to predict experimental observables. The input data for the deep learning model would be surface properties, such as the hydration maps from Chapter 10, which are then converted into a 2D image using polar coordinates. The output data would be experimental observables (*e.g.* logP, cell uptake, zeta potential, *etc.*).

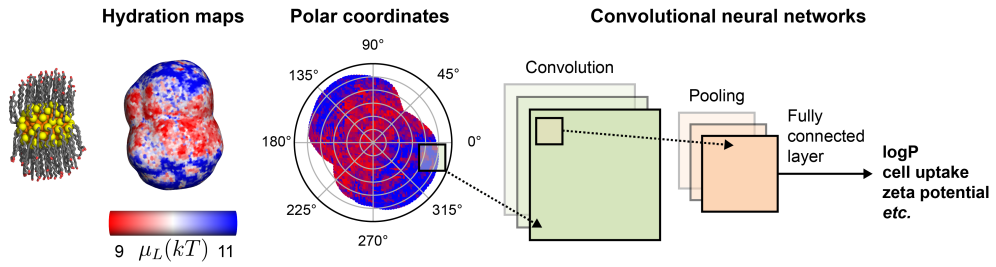


Figure 12.4: Future work on integrating deep learning models to predict GNP properties from surface maps.

12.2.3 Future work 3: Employ active learning to suggest new GNPs for experimentalists

A critical challenge of GNP applications is the selection of GNPs for experimental synthesis and testing. Hence, the computational tools developed in this dissertation may aid in screening GNPs for enhanced properties. However, given the limited experimental data, we need an algorithm to intelligently explore the design space that could "exploit" areas of promising GNPs and "explore" areas of GNPs that are different from the current data set. These algorithms are known as active learning, which are designed to iteratively ask the user for new data point labels based on the current data available.[17–19] With an active learning framework, we could begin tuning GNPs for a large-range of properties, such as cell uptake, and use the computational models (*e.g.* descriptor approach in Chapter 11 or deep learning approach in Section 12.2.2) to help explore the design

space. Then, we could downselect the top 5 GNPs for an experimentalist to validate. This synergistic approach between computation and experiment would be a powerful approach to rapidly identify promising GNPs while lowering the number of experiments required.

12.3 References

[1] Motagamwala, A. H.; Huang, K.; Maravelias, C. T.; Dumesic, J. A. Solvent system for effective near-term production of hydroxymethylfurfural (HMF) with potential for long-term process improvement. *Energy & Environmental Science* **2019**, *12*, 2212–2222.

[2] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **2019**, *10*, 370–377.

[3] Jessop, P. G.; Jessop, D. A.; Fu, D.; Phan, L. Solvatochromic parameters for solvents of interest in green chemistry. *Green Chemistry* **2012**, *14*, 1245–1259.

[4] Walker, T. W.; Frelka, N.; Shen, Z.; Chew, A. K.; Banick, J.; Grey, S.; Kim, M. S.; Dumesic, J. A.; Van Lehn, R. C.; Huber, G. W. Recycling of multilayer plastic packaging materials by solvent-targeted recovery and precipitation. *Science advances* **2020**, *6*, eaba7599.

[5] Begum, S.; Rasul, M. G.; Akbar, D.; Ramzan, N. Performance analysis of an integrated fixed bed gasifier model for different biomass feedstocks. *Energies* **2013**, *6*, 6508–6524.

[6] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; et al..

Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM* **2008**, *51*, 91–97.

[7] Sheavly, J. K.; Pedersen, J. A.; Van Lehn, R. C. Curvature-driven adsorption of cationic nanoparticles to phase boundaries in multicomponent lipid bilayers. *Nanoscale* **2019**, *11*, 2767–2778.

[8] Sheavly, J. K.; Van Lehn, R. C. Bilayer-mediated assembly of cationic nanoparticles adsorbed to lipid bilayers: Insights from molecular simulations. *AIChE Journal* **2020**, *66*, e16993.

[9] Lu, X.; Xu, P.; Ding, H.-M.; Yu, Y.-S.; Huo, D.; Ma, Y.-Q. Tailoring the component of protein corona via simple chemistry. *Nature communications* **2019**, *10*, 1–14.

[10] Lundqvist, M.; Stigler, J.; Cedervall, T.; Berggard, T.; Flanagan, M. B.; Lynch, I.; Elia, G.; Dawson, K. The evolution of the protein corona around nanoparticles: a test study. *ACS nano* **2011**, *5*, 7503–7509.

[11] Saha, K.; Rahimi, M.; Yazdani, M.; Kim, S. T.; Moyano, D. F.; Hou, S.; Das, R.; Mout, R.; Rezaee, F.; Mahmoudi, M.; et al.. Regulation of macrophage recognition through the interplay of nanoparticle surface functionality and protein corona. *ACS nano* **2016**, *10*, 4421–4430.

[12] Attwood, S. J.; Kershaw, R.; Uddin, S.; Bishop, S. M.; Welland, M. E. Understanding how charge and hydrophobicity influence globular protein adsorption to alkanethiol and material surfaces. *Journal of Materials Chemistry B* **2019**, *7*, 2349–2361.

[13] Ghanakota, P.; Carlson, H. A. Driving structure-based drug discovery through cosolvent molecular dynamics: Miniperspective. *Journal of medicinal chemistry* **2016**, *59*, 10383–10399.

[14] Di Felice, R.; Corni, S. Simulation of peptide–surface recognition. *The Journal of Physical Chemistry Letters* **2011**, *2*, 1510–1519.

[15] Raman, E. P.; Yu, W.; Lakkaraju, S. K.; MacKerell Jr, A. D. Inclusion of multiple fragment types in the site identification by ligand competitive saturation (SILCS) approach. *Journal of chemical information and modeling* **2013**, *53*, 3384–3398.

[16] Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; Bronstein, M.; Correia, B. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **2020**, *17*, 184–192.

[17] Brochu, E.; Cora, V. M.; De Freitas, N. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599* **2010**.

[18] Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B.; Jerome, S. Efficient Exploration of Chemical Space with Docking and Deep-Learning. **2021**.

[19] Du, P.; Bai, X.; Tan, K.; Xue, Z.; Samat, A.; Xia, J.; Li, E.; Su, H.; Liu, W. Advances of four machine learning methods for spatial data handling: A review. *Journal of Geovisualization and Spatial Analysis* **2020**, *4*, 1–25.