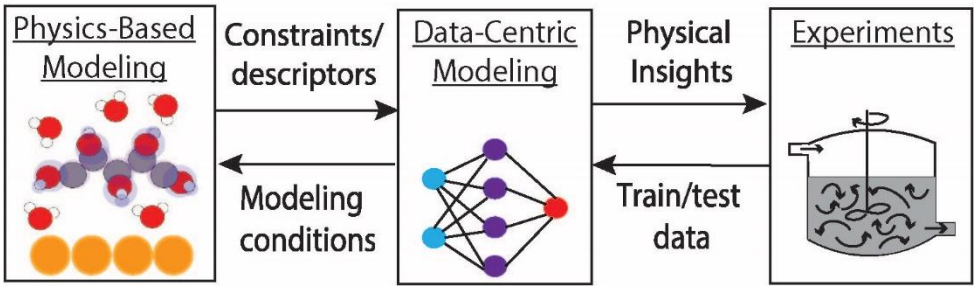
On the Integration of Molecular Dynamics, Data Science, and Experiments for Studying Solvent Effects on Catalysis

Lisa Je[1], George W. Huber[1], Reid C. Van Lehn[1,2], Victor M. Zavala[1,*]

Department of Chemical and Biological Engineering[1] and DOE Great Lakes Bioenergy Research Center[2], University of Wisconsin – Madison, Madison, WI, 53706, USA

*send correspondence to: victor.zavala@wisc.edu

Abstract



Computational workflows that combine molecular dynamics (MD) simulations and emerging data-centric (DC) methods can accelerate the screening and analysis of solvent systems experimentally and computationally. MD simulations provide atomic positions and velocities of reactant, solvent, and catalyst materials that can be manipulated into data representations that in turn can be used by DC techniques to conduct predictive modeling, feature extraction, and experimental design. For liquid-phase catalytic applications, emerging DC techniques such as Convolutional and Graph Neural Networks (CNN/GNN), Topological Data Analysis (TDA), and Active Learning (AL) can leverage MD and experimental data to quickly predict solvent effects on reaction outcomes. For instance, in recent studies, 3D solvent environments obtained with MD have been exploited by CNNs to predict experimental reaction rates for homogeneous acid-catalyzed lignocellulosic processes. In this perspective, we discuss basic principles of DC methods

and how these can be combined with MD to enable high-throughput screening of solvent selection for diverse catalysis applications.

1. Introduction

Driven by environmental concerns, substantial effort has been devoted to developing liquid-phase catalytic processes that produce bio-based fuels and chemicals. Advanced plastic recycling technologies are also being developed that involve liquid-phase catalytic chemistry [1]. In these processes, solvent selection and composition strongly influence reaction rates and product selectivities [2]–[4]. For example, chemical conversion of biomass derivatives such as 5-(hydroxymethyl) furfural is driven by the solvent [5]–[8]. Similarly, the catalytic dehydration of biomass-derived feedstocks using Brønsted solid acid catalysts depends on solvent-mediated pathways [9]–[11]. Unfortunately, elucidating solvent effects experimentally is time-consuming and screening solvents and compositions that achieve desired reaction outcomes is challenging due to the high dimensionality and complexity of solvent design spaces [12]. There is thus a need for computational workflows that can guide and interpret experiments by modeling reactant-solvent-catalyst systems [13].

State-of-the-art methods for computing reaction free-energy landscapes and modeling catalytic materials include density functional theory (DFT) and *ab initio* molecular dynamics (AIMD) simulations [10], [14]. These methods utilize quantum mechanical calculations to predict electronic structure but differ in their treatment of the solvent. DFT calculations typically utilize implicit solvent representations (*e.g.*, polarizable continuum models) and have been used, for instance, to study solvent effects on heterogeneous catalysis [15]–[17]. AIMD simulations calculate forces based on electronic structure calculations performed during a simulation and can directly model the dynamic organization of solvent molecules. Consequently, AIMD simulations

can elucidate the importance of specific molecules within the solvent environment and are applied to model liquid-phase reactions [3], [18] and methanol electrooxidation [19]. Unfortunately, the computational cost associated with DFT and AIMD simulations is substantial.

Screening solvent systems is experimentally and computationally challenging given the number of solvent types and compositions possible for most reactions. Moreover, characterizing 3D solvent environments at the molecular scale is challenging. To address this challenge, data-centric (DC) techniques have the potential to help analyze complex environments and to accelerate the prediction of experimental outcomes by leveraging physics-based simulation data, as demonstrated by the application of DC techniques to screen catalytic design spaces [20]. For example, Neural Networks (NNs) can be combined with DFT descriptors to predict catalytic activity [21]. Machine learning (ML) models have been used for efficiently identifying ternary platinum alloys [22] and ML combined with DFT calculations has been used to predict bimetallic compositions and adsorption energies [23]. These examples focus on systems in which solvent effects are not considered; however, DC tools also have the potential to uncover relationships between molecular-scale inputs and solvent effects to facilitate the computational screening of solvent systems for desired catalysis.

In this perspective paper, we discuss how *classical* molecular dynamics (MD) simulations can be combined with DC approaches - ML and topological data analysis (TDA) - to investigate solvent effects on reaction rates and selectivity [6]–[8]. MD simulations compute forces during a simulation using computationally inexpensive force field methods, enabling access to larger length scales (~10 nm), longer time scales (μs), and more complex solvent environments than AIMD simulations. Classical MD is not typically used for predicting solvent effects because reaction mechanisms, transition states, and electronic structure are not explicitly modeled. However, for

certain systems, DC techniques can be used to characterize solvent environments obtained from MD (e.g., via manual or automated feature extraction) and to learn mappings that correlate environment features to reaction outcomes. This approach has been used, for example, to show that the enrichment of water molecules near a reactant can predict acid-catalyzed reaction rates [6], [7].

2. Physics-based Methods, Data Representations, and Experimental Inputs for Data-Centric Methods

As illustrated in Figure 1, aspects to consider when applying DC techniques to study solvent effects include: (1) the nature of the output data obtained from physics-based models, (2) how output data can be represented mathematically, and (3) how to determine the appropriate DC techniques that are compatible with such data representations. The outputs of physics-based models (DFT, AIMD, and MD) are raw data of different types. DFT outputs are electronic densities resulting in total energies, AIMD outputs are electronic energies, equilibrated positions, and velocities of atoms, and MD outputs are equilibrated positions and velocities of atoms. These raw data outputs can be manipulated into *different data representations* such as scalar values, matrices/tensors, graphs/networks, and fields. For example, total energies from DFT and AIMD are used to calculate binding energy descriptors (scalar values) for predicting catalytic activity. Positions from MD can also be used to calculate descriptors of hydrogen bonding (scalar values) or hydrogen bonding networks. Scalar descriptors are traditional representations (e.g., obtained from physical insight) that are used for predicting catalytic trends (activity and selectivity). Unfortunately, such descriptors are not always easy to identify and may not capture complex spatio-temporal relationships in the raw data. Alternative data representations such as 2D/3D fields can be used to capture spatial densities of molecules (solvent, catalyst, and co-solvent) obtained from DFT, AIMD, and MD simulations. A limitation of 2D/3D fields is that they only partially

encode information on atomistic positions and thus can omit important chemical information such as bonding between atoms and atomic charges. Point clouds are data representations that can encode bonding information (implicitly) because they capture atomistic positions directly. Graphs are also powerful data representations that can directly encode bond information and can be generated from the atomic positions of the DFT, AIMD, and MD simulations to discover topological structures.

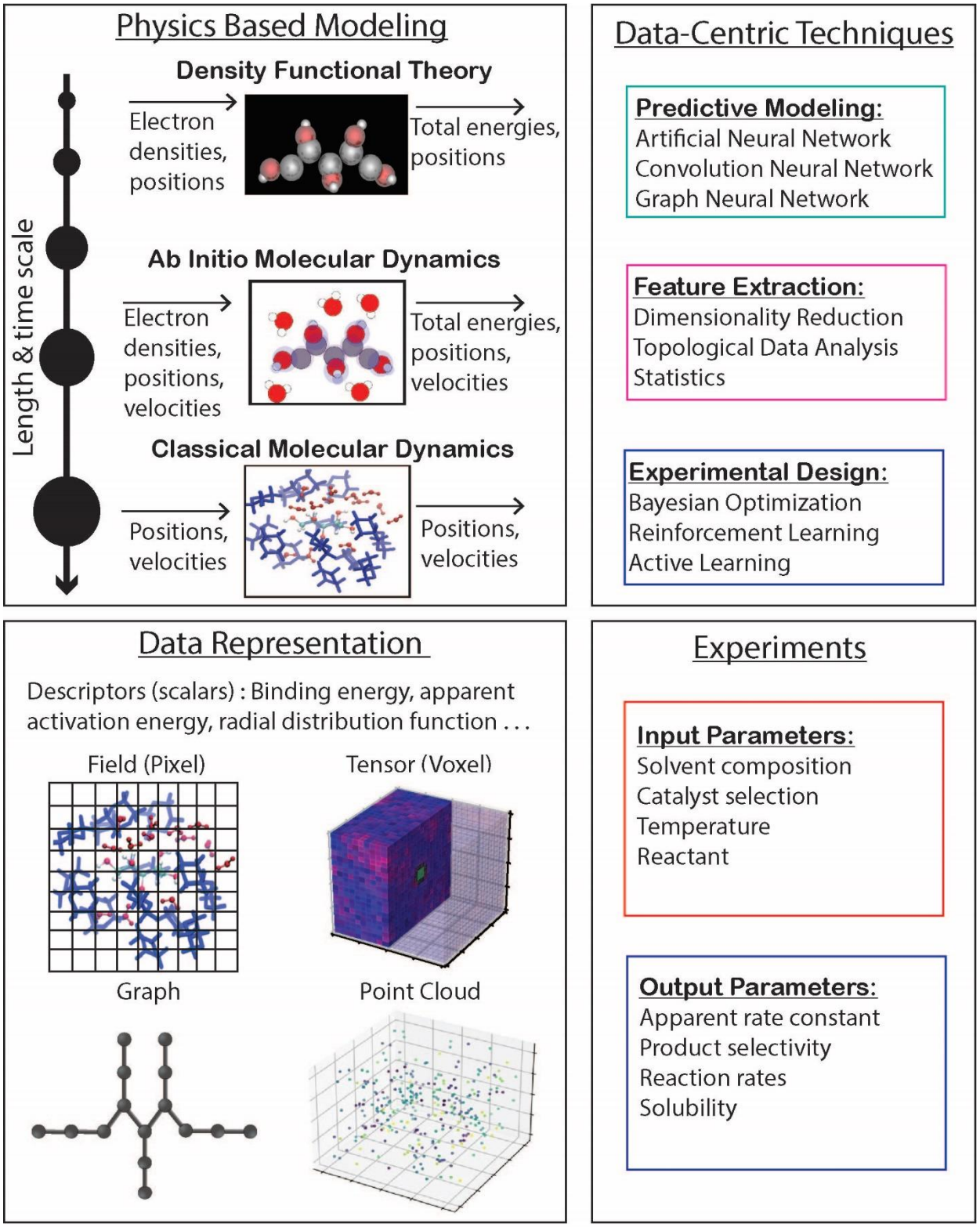


Figure 1. Main components of computational workflows used for analyzing solvent effects: Physics-based Modeling, Data Representations, Data-Centric Techniques, and Experiments. Outputs from physics-based modeling (total energies, positions, and velocities) can be manipulated into different data representations such as scalar descriptors or

into topological representations (fields, tensors, graphs, and point clouds). The data representations can be fed into data-centric techniques to conduct predictive modeling, feature extraction, or experimental design tasks. Data-centric techniques deliver outputs that can be compared to experimental values and can produce insights on physically realizable/promising experiments (which can be captured as constraints). Experimental outputs can also be mapped into data representations to perform data-centric tasks and provide outcomes that help design new experiments.

Three broad categories of DC techniques are discussed here: (1) predictive modeling, (2) feature extraction, and (3) experimental design. Modeling methods such as NNs and CNNs can process data objects (experimental and computational) in the form of scalar descriptors or 2D/3D fields as inputs for predicting solvent and catalytic trends. Feature extraction methods such as principal component analysis (PCA), convolutions, autoencoders, and TDA can reduce the data objects into low-dimensional forms that can be used to visualize solvent environments and to make predictions. Some of these feature extraction techniques are embedded within predictive models (e.g., convolutions are performed inside CNNs). The features extracted from these techniques are expected to be informative descriptors of the original raw data objects but are not always easy to interpret. For instance, autoencoders and PCA obtain descriptors in the form of latent variables that do not have a direct physical meaning. Techniques such as TDA obtain descriptors that encode topological/geometric information such as Betti numbers (number of voids and connected components), Euler characteristic curves [24], or persistence diagrams [25], [26]. Experimental design methods such as Bayesian optimization and active learning (AL) aim to propose new experimental inputs that can help refine data objects and associated features and predictive models.

Here, we do not provide an extensive review of available techniques but instead describe a few emerging DC techniques - Graph Neural Networks (GNNs), TDA, and AL – and highlight how they can be combined with MD simulations to understand and predict solvent effects on catalysis. GNNs and TDA can process graph-structured data and thus naturally capture topological features of molecular structures and intramolecular bonding. Graphs are flexible data

representations and can encode significant information; for instance, nodes in a molecular graph can represent atoms and encode information such as chemical identity, graph degree (number of connected edges to the node), and charge, while edges can represent chemical bonding. [27] However, graphs can also represent other data objects such as domains in a solvent structure or a hydrogen bonding network. A graph representation can be input into a GNN and information is extracted by using nested convolution and activation operations; the information is reduced using pooling operations to obtain a feature vector that can be used to make predictions (regression or classification) using a NN or statistical models [28]. TDA can extract information from graphs, point clouds, and fields. The features obtained from TDA are the number of voids and connected components, which are captured in topological descriptors such as the Euler characteristic and Betti numbers [25]. These descriptors can be used for conducting regression, classification, and clustering tasks; this approach has been used for identifying and predicting desired properties relevant to solvent-catalyst interactions, porous zeolites [29], and liquid crystals [30]. AL accelerates simulations and experimental studies by exploring the information space in less time and is used in molecular property prediction [31], generating interatomic potentials [32], [33], and identifying catalytic candidate surfaces with high activity [34]. AL creates a surrogate model (e.g., support vector machine, NN, or Gaussian process) from the input dataset which selects the next dataset. These input descriptors are iteratively added to the original dataset and updated as a new surrogate model to identify labels of interest as outputs. AL aims to suggest experiments to balance the often-conflicting goals of exploring the design space (to gain confidence in predictions) and of exploiting the design space (to optimize an objective of interest). This approach can enable accelerated, flexible, and automated materials screening for catalysis and reaction engineering.

3. Current Applications of MD and Data-Centric Techniques for Investigating Solvent Effects

To illustrate concepts discussed, here we present a case study from our recent work that illustrates how MD simulations and DC can analyze solvent effects and their impact on reaction outcomes. This work was motivated by experimental findings from Dumesic *et. al.*, who demonstrated that rates for the Brønsted-acid-catalyzed hydrolysis and dehydration of biomass-derived oxygenates increase in mixtures of water and polar aprotic cosolvents [7]. Here, reactivity was quantified by the so-called *kinetic solvent parameter*, which measures the relative rate of a reaction in a solvent mixture normalized by the rate in pure water. In our first DC approach, we sought to predict this output by training predictive models (regression and feedforward NNs) that used MD-derived descriptors of the solvent environment to predict experimental reactivity for seven biomass-derived oxygenates in mixtures of water and three different polar aprotic cosolvents in varying ratios; in total, the data set comprised 76 unique reactant-solvent combinations and associated kinetic solvent parameters.

To obtain the solvent descriptors, we performed MD simulations for each of the reactant-solvent combinations to compute three *human-selected* descriptors that quantified: (1) water enrichment near the reactant, (2) the hydrogen bonding lifetime between water molecules and the reactant, and (3) the fraction of the reactant surface area associated with hydroxyl groups. These descriptors (scalar values) were selected based on physical insight from prior liquid-phase AIMD simulations that identified a relationship between properties of the local solvent domain near a reactant and corresponding activation energies [3]. The descriptors were then used as input to linear and non-linear regression models to predict kinetic solvent parameters (Figure 2). Model accuracy was quantified by computing the root-mean-squared error (RMSE) and the slope of a parity line between predicted and experimentally measured values. When trained on mixtures of water and dioxane, a linear regression model was accurate, obtaining an RMSE of 0.23 and parity

line slope of 0.89; however, this model was not as accurate when including mixtures with other cosolvents (gamma-valerolactone or tetrahydrofuran), obtaining an RMSE of 0.58 and decreasing the slope to 0.49. Utilizing a NN to train a nonlinear regression model did not improve model accuracy (RMSE of 0.62 and parity line slope of 0.46); this indicated that the descriptors selected provided valuable information to predict reactivity but failed to capture some features of the solvent environment. Nonetheless, the encouraging results for dioxane-water mixtures indicated that MD-derived solvent environments contain relevant information that can be exploited to predict reactivity trends. This relationship between MD and solvent environment structure is possible because MD simulations generate high-dimensional datasets that encode both spatial and temporal correlations between molecules.

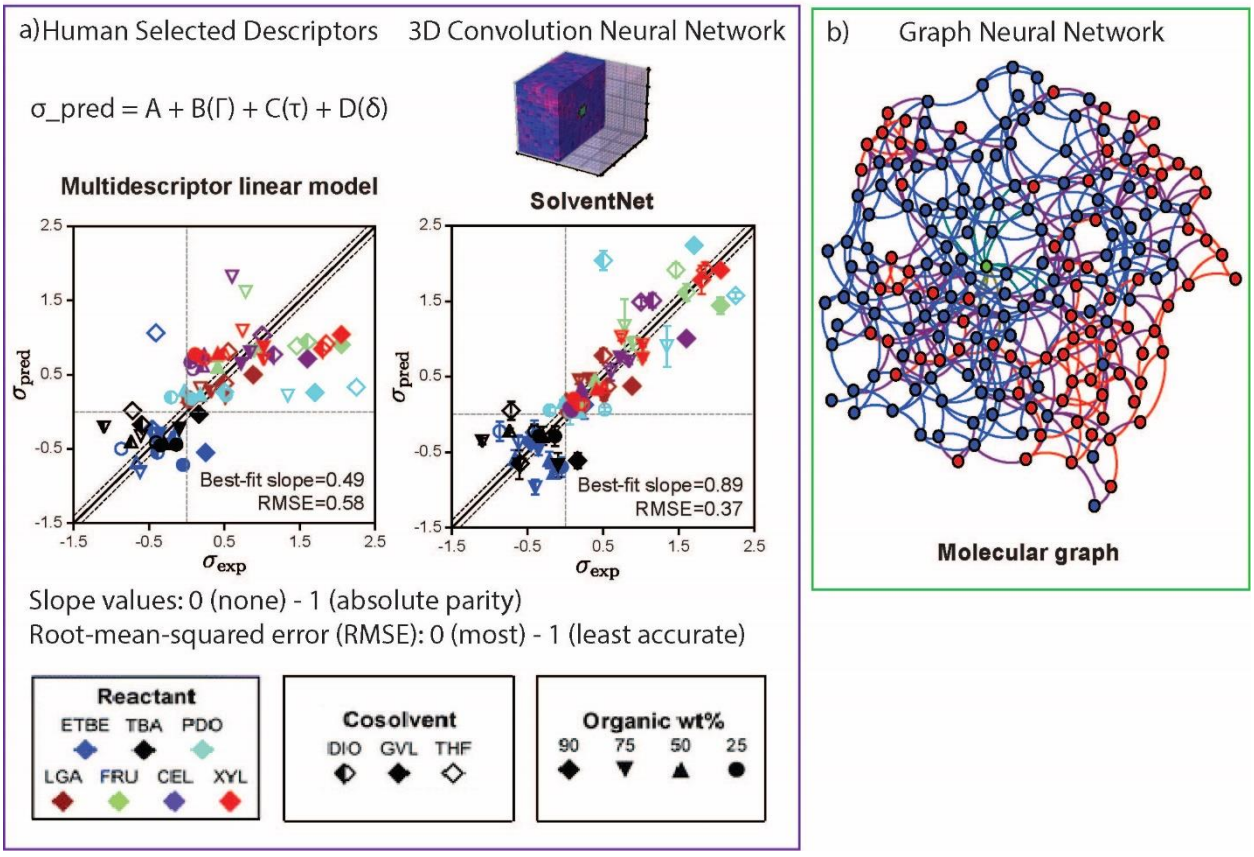


Figure 2. a) Evaluating human-selected descriptors vs 3D CNN techniques from MD simulations to predict experimental kinetic solvent parameters. Human-selected descriptors were obtained from each MD trajectory

reflecting physical reactant hydrophilicity and solvent interactions: the preferential exclusion coefficient (Γ); the hydrogen bonding lifetime between the reactant and water (τ); and the accessible hydroxyl fraction (δ). These human-selected descriptors were applied to a linear regression model to predict values of the kinetic solvent parameter and a parity plot between the predicted and experimental kinetic solvent parameters (σ_{pred} vs. σ_{exp}) resulted in the best-fit slope and root-mean-squared error (RMSE) shown within the plot below. The 3D CNN inputs a 20×20×3 voxel tensor and outputs predicted kinetic solvent parameters. This 3D CNN called SolventNet predicted kinetic solvent parameters which were applied to a parity plot with the best-fit slope and RMSE shown within the plot below. b) A graph neural network is proposed as an alternative DC architecture for predicting values of the kinetic solvent parameter. The data representation proposed is a graph network as illustrated.

In a second approach, we sought to automatically extract features of the MD-derived solvent environments (without designing human-selected descriptors). To do so, we gathered raw data from MD outputs – that is, time trajectories for atomistic positions of the reactant, co-solvent, and water in a 3D domain. From the atomistic positions, we determined time trajectories for the densities of water, co-solvent, and water in the 3D domain, which resulted in time trajectories of three different 3D fields that were discretized into voxels. We computed time-averages over short time periods of these fields; here each voxel encoded the density of either water, co-solvent, and water in a particular region of the domain and over a particular time period [6]. The three fields were fed into a 3D CNN; this architecture automatically learned spatial features (by learning filters) from the 3D fields that best explained reactivity. The architecture, which we called SolventNet, exhibited substantially improved accuracy (RMSE of 0.37 and parity line slope of 0.89) compared to the regression models [6]. Moreover, SolventNet generalized to predict kinetic solvent parameters for an unseen test set of 32 reactant-solvent combinations (obtained independently from the literature), which included three additional cosolvents not included in the training set. SolventNet also required substantially less MD simulation time than the descriptor-based model; specifically, we found that the early dynamics of the 3D solvent environment encoded sufficient information to predict reactivity. This computational efficiency enabled solvent screening for the acid-catalyzed dehydration of cyclohexanol to cyclohexene and the partial dehydration of fructose to 5-hydroxymethylfurfural [8]. Unfortunately, a downside of SolventNet

is the number of model parameters (172,417 for SolventNet vs. 3 for linear regression), which hinders physical interpretability and complicates training.

The findings obtained with SolventNet highlight the ability of DC techniques to extract relevant information (e.g., geometrical and statistical) from complex datasets to enable predictions. Notably, classical MD simulations only model interactions between solvent molecules and the neutral reactants and cannot directly model interactions between solvent molecules and charged transition states, indicating that DC techniques overcome this limitation despite the difference in reactant/transition state properties. More recently, we have been using TDA to more directly extract topological features from 3D solvent configurations and with this overcome the challenges encountered with the 3D CNN. We have found that the topology of solvent environments can be characterized using Betti numbers and the Euler characteristic and that such descriptors can be used to predict reactivity using simple linear regression models [30]. An alternative is to replace 3D field representations with graph-based representations of intermolecular interactions between solvent molecules (e.g., hydrogen bond networks); these representations would enable the use of GNNs or regression models (Figure 2). The latter approach could also permit the integration of graph-based representations of intermolecular interactions with graph-based representations of molecular structure to identify solvent effects on selectivities [35]. Another potential approach is to apply an AL framework to supplement the 3D CNN for solvent screening. In the work of Van Lehn *et. al.*, an AL framework used MD simulation inputs to predict hydration free energies, a hydrophobicity descriptor, for a dataset of 200,000 self-assembled monolayers [36]. Similarly, solvent compositions can also be identified using this AL framework to guide efficient experimental data collection; potential future research could also investigate integrating AL with high-throughput experiments. Finally, an implicit assumption in our prior work is that the rate-

determining step and reaction mechanism remain constant as the solvent composition changes, and future work will need to consider if DC techniques can remain accurate if solvent selection changes the reaction mechanism. However, we anticipate that the inability of DC techniques to generalize to new solvent compositions could also be informative by spurring investigations of how specific solvent compositions influence reaction mechanisms.

4. Future applications of MD and Data-Centric Techniques for Investigating Solvent Effects

The successful application of MD and DC techniques for predicting solvent effects motivates the pursuit of multiple avenues; here, we identify three potential application areas where solvent effects are critical and hence the combination of MD and DC could be beneficial. In analogy to biomass deconstruction, MD and DC tools could potentially predict solvent effects relevant to the liquid-phase upcycling of plastic waste. For example, molecular simulations have been used to screen solvents for the selective dissolution of polymers from multilayer plastic films to enable their recovery in a chemically pure form [37]–[39]. This approach requires calculation of polymer solubility and could potentially be accelerated by DC techniques that utilize graph-based representations of polymers and their chemical features as input to permit analysis by GNNs or TDA. Similarly, the base-catalyzed depolymerization of polyethylene terephthalate in organic solvents has been developed industrially (*i.e.*, the VolCat process, the Eastman Chemical methanolysis process [40], and the Loop industry polyethylene terephthalate process), but solvent effects remain unexplored. AL could be used to guide the selection of solvent systems and dissolution temperatures to optimize process parameters.

MD and DC tools can also explore large information spaces to investigate how solvent effects influence catalytic adsorption in zeolite or other microporous materials. For example,

Flaherty *et. al.*, demonstrated that the structure of water within zeolite pores substantially impacts epoxidation turnover rates [41], [42]. The hydrogen bond structures formed under these confined conditions could be modeled using MD and interpreted using GNNs or TDA to relate such structures to a wider range of zeolite geometries or experimental reaction conditions identifying molecular clustering and solvent kinetic effects [18]. TDA has also been used to screen zeolite structures for methane storage and carbon capture [29]; there thus exists the potential to combine TDA-based analysis of zeolite structures with either TDA analysis of MD data or MD-derived descriptors of solvent structure to identify zeolites with high catalytic activity.

The choice of liquid electrolyte also impacts the activity and selectivity of electrocatalytic reactions, such as carbon dioxide electroreduction (CO_2RR) [43]–[45]. Solvent effects on electrocatalysis are currently studied by implicit solvent first-principles techniques, but these studies could be supplemented using explicit solvent MD and DC techniques to study solvent structure near electrode interfaces [46], [47]. For example, the structure of the electrical double layer in aqueous solutions with different supporting electrolytes could be modeled with MD and represented as voxel representations suitable for interpretation by TDA or a 3D CNN to predict CO_2RR activity. Exploration of broader classes of solvent systems, including non-aqueous solvents, could be guided using AL to identify systems with improved chemical reactivity or selectivity to products. Such systems would thus complement studies focused on tuning electrocatalyst properties to additionally optimize the surrounding electrolyte.

5. Concluding Remarks

Molecular dynamic (MD) simulations combined with emerging data-centric (DC) techniques can provide valuable insights on the effect of solvents in reaction engineering and catalysis. MD provides atomistic positions of solvents and catalytic materials as inputs for DC techniques.

Techniques such as GNNs, TDA, and AL have the potential to predict solvent effects more accurately by supplementing current MD and other physics-based tools like DFT. Together, MD and DC tools can guide the high-throughput screening of solvent selection for catalytic reactions and minimize experimental effort. Previously, MD with DC tools predicted catalytic rates and selectivities on homogeneous, acid-catalyzed lignocellulosic conversion processes. Future directions MD and DC are in the field of heterogenous biomass catalytic processes, zeolite catalysis, and electrocatalysis.

6. Acknowledgements

The work is supported in part by the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018409. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1747503. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. Disclosure

The authors declare no conflict of interest.

8. References

(1) Ellis, L. D.; Rorrer, N. A.; Sullivan, K. P.; Otto, M.; McGeehan, J. E.; Román-Leshkov, Y.; Wierckx, N.; Beckham, G. T. Chemical and Biological Catalysis for Plastics Recycling and Upcycling. *Nat. Catal. 2021 47* **2021**, *4* (7), 539–556. https://doi.org/10.1038/s41929-021-00648-4.

(2) Adams, J. S.; Chemburkar, A.; Priyadarshini, P.; Ricciardulli, T.; Lu, Y.; Maliekkal, V.; Sampath, A.; Winikoff, S.; Karim, A. M.; Neurock, M.; Flaherty, D. W. Solvent Molecules Form Surface Redox Mediators in Situ and Cocatalyze O2reduction on Pd. *Science (80-.).* **2021**, *371* (6529), 626–632. https://doi.org/10.1126/SCIENCE.ABC1339.

(3) 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. *Nat. Catal. 2018 13* **2018**, *1* (3), 199–207. https://doi.org/10.1038/s41929-018-0027-3.

(4) John Varghese, J.; H. Mushrif, S. Origins of Complex Solvent Effects on Chemical Reactivity and Computational Tools to Investigate Them: A Review. *React. Chem. Eng.* **2019**, *4* (2), 165–206. https://doi.org/10.1039/C8RE00226F.

(5) Chang, H.; Motagamwala, A. H.; Huber, G. W.; Dumesic, J. A. Synthesis of Biomass-Derived Feedstocks for the Polymers and Fuels Industries from 5-(Hydroxymethyl)Furfural (HMF) and Acetone. *Green Chem.* **2019**, *21* (20), 5532–5540. https://doi.org/10.1039/C9GC01859J.

(6) K. Chew, A.; Shengli Jiang; Weiqi Zhang; M. Zavala, V.; Lehn, R. C. V. Fast Predictions of Liquid-Phase Acid-Catalyzed Reaction Rates Using Molecular Dynamics Simulations and Convolutional Neural Networks. *Chem. Sci.* **2020, *11* (46), 12464–12476. https://doi.org/10.1039/D0SC03261A.

The authors demonstrate that classical molecular dynamic simulations can quantify the reactant–water–cosolvent interactions that dictate reactivity in prior ab initio simulations. Molecular dynamics can then be combined with 3-dimensional convolution neural networks to enable the rapid high throughput screening of solvent systems for biomass conversion processes.

*(7) Walker, T. W.; Chew, A. K.; Li, H.; Demir, B.; Zhang, Z. C.; Huber, G. W.; Lehn, R. C. Van; Dumesic, J. A. Universal Kinetic Solvent Effects in Acid-Catalyzed Reactions of Biomass-Derived Oxygenates. *Energy Environ. Sci.* **2018**, *11* (3), 617–628. https://doi.org/10.1039/C7EE03432F.

The authors demonstrate a multidescriptor model to predict rate constants as a function of solvent composition for acid-catalyzed reactions via physics-based methods. Three aprotic solvents and seven biomass derivatives were investigated using molecular dynamics to identify three descriptors (1) water enrichment in the local solvent domain of the reactant, (2) average hydrogen bonding lifetime in the water molecules and (3) portion of the reactant accessible surface area occupied by hydroxyls

*(8) 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. *Top. Catal. 2020 637* **2020**, *63* (7), 649–663. https://doi.org/10.1007/S11244-020-01260-9.

The authors demonstrate a method to estimate solvent effects on liquid-phase, acid-catalyzed biomass conversion reaction rates/selectivity by combining experiments, classical molecular dynamic simulations, and data-centric methods. These solvent effects were investigated in two case studies (1) acid-catalyzed dehydration of cyclohexanol to cyclohexene and (2) partial dehydration of fructose to 5-hydroxymethylfurfurnal.

(9) Zhao, Z.; Bababrik, R.; Xue, W.; Li, Y.; Briggs, N. M.; Nguyen, D.-T.; Nguyen, U.;

Crossley, S. P.; Wang, S.; Wang, B.; Resasco, D. E. Solvent-Mediated Charge Separation Drives Alternative Hydrogenation Path of Furanics in Liquid Water. *Nat. Catal. 2019 25* **2019**, *2* (5), 431–436. https://doi.org/10.1038/s41929-019-0257-z.

(10) Oyola-Rivera, O.; He, J.; Huber, G. W.; Dumesic, J. A.; Cardona-Martínez, N. Catalytic Dehydration of Levoglucosan to Levoglucosenone Using Brønsted Solid Acid Catalysts in Tetrahydrofuran. *Green Chem.* **2019**, *21* (18), 4988–4999. https://doi.org/10.1039/C9GC01526D.

(11) Rojas-Buzo, S.; Corma, A.; Boronat, M.; Moliner, M. Unraveling the Reaction Mechanism and Active Sites of Metal–Organic Frameworks for Glucose Transformations in Water: Experimental and Theoretical Studies. *ACS Sustain. Chem. Eng.* **2020**, *8* (43), 16143–16155. https://doi.org/10.1021/ACSSUSCHEMENG.0C04398.

*(12) Li, G.; Wang, B.; Resasco, D. E. Water-Mediated Heterogeneously Catalyzed Reactions. *ACS Catal.* **2019**, *10* (2), 1294–1309. https://doi.org/10.1021/ACSCATAL.9B04637.

The authors emphasize the importance of water as a solvent which can influence the solubility and diffusivity of reactants and products, selectivity, and overall rates of reactants and products. This review demonstrates several catalytic mechanisms in the presence of water which have the following effects on: (1) modification of reactant solubility and rate of mass transport (2) reactant, product, and transisition state stability and (3) cocatalyst abilities.

(13) Chen, B. W. J.; Xu, L.; Mavrikakis, M. Computational Methods in Heterogeneous Catalysis. *Chem. Rev.* **2020**, *121* (2), 1007–1048. https://doi.org/10.1021/ACS.CHEMREV.0C01060.

(14) Varghese, J. J. Computational Design of Catalysts for Bio-Waste Upgrading. *Curr. Opin. Chem. Eng.* **2019**, *26*, 20–27. https://doi.org/10.1016/J.COCHE.2019.08.002.

(15) Hansen, M. H.; Nørskov, J. K.; Bligaard, T. First Principles Micro-Kinetic Model of Catalytic Non-Oxidative Dehydrogenation of Ethane over Close-Packed Metallic Facets. *J. Catal.* **2019**, *374*, 161–170. https://doi.org/10.1016/J.JCAT.2019.03.034.

(16) Ludwig, T.; Gauthier, J. A.; Brown, K. S.; Ringe, S.; Nørskov, J. K.; Chan, K. Solvent–Adsorbate Interactions and Adsorbate-Specific Solvent Structure in Carbon Dioxide Reduction on a Stepped Cu Surface. *J. Phys. Chem. C* **2019**, *123* (10), 5999–6009. https://doi.org/10.1021/ACS.JPCC.8B11571.

(17) Mathew, K.; Sundararaman, R.; Letchworth-Weaver, K.; Arias, T. A.; Hennig, R. G. Implicit Solvation Model for Density-Functional Study of Nanocrystal Surfaces and Reaction Pathways. *J. Chem. Phys.* **2014**, *140* (8), 084106. https://doi.org/10.1063/1.4865107.

*(18) Bates, J. S.; Bukowski, B. C.; Greeley, J.; Gounder, R. Structure and Solvation of Confined Water and Water–Ethanol Clusters within Microporous Brønsted Acids and Their Effects on Ethanol Dehydration Catalysis. *Chem. Sci.* **2020**, *11* (27), 7102–7122. https://doi.org/10.1039/D0SC02589E.

The authors distinguish the unique effects of water in a confined and clustered structure

which is reflected in a case study of ethanol dehydration to form diethyl ether. Water clusters in microporous Brønsted acids solvate hydrogen bonded water networks and stabilize intermediates and transisition states compared to non-extended water networks.

(19) Herron, J. A.; Morikawa, Y.; Mavrikakis, M. Ab Initio Molecular Dynamics of Solvation Effects on Reactivity at Electrified Interfaces. *Proc. Natl. Acad. Sci.* **2016**, *113* (34), E4937–E4945. https://doi.org/10.1073/PNAS.1604590113.

(20) Kitchin, J. R. Machine Learning in Catalysis. *Nat. Catal. 2018 14* **2018, *1* (4), 230–232. https://doi.org/10.1038/s41929-018-0056-y.

The author discusses current physics-based model development as thermodynamic and quantum methcanical derivations to fit parameters with experimental data. The author proposes machine learning model development as flexible and non linear methods trained from reference data to predict reactivity descriptors for high throughput catalytic screening.

*(21) Smith, A.; Keane, A.; Dumesic, J. A.; Huber, G. W.; Zavala, V. M. A Machine Learning Framework for the Analysis and Prediction of Catalytic Activity from Experimental Data. *Appl. Catal. B Environ.* **2020**, *263*, 118257. https://doi.org/10.1016/J.APCATB.2019.118257.

The authors demonstrate a machine learning framework for analyzing and predicting catalytic activity for the water-gas shift reaction which expands the information space of heterogenous catalysts. Using neural networks and dimension reduction in the form of principle component analysis, the authors pave a path to bridge the information space from physics-based theory to experiments and guide catalyst screening.

(22) Lee, J.; Jinnouchi, R. Machine Learning-Based Screening of Highly Stable and Active Ternary Pt Alloys for Oxygen Reduction Reaction. *J. Phys. Chem. C* **2021**, *125*, 16963–16974. https://doi.org/10.1021/ACS.JPCC.1C02890.

(23) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; Chan, K.; Nørskov, J. K. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO2 Reduction. *ACS Catal.* **2017**, *7* (10), 6600–6608. https://doi.org/10.1021/ACSCATAL.7B01648.

(24) Smith, A.; Zavala, V. The Euler Characteristic: A General Topological Descriptor for Complex Data. *Comput. Chem. Eng.* **2021**, *154*.

(25) Townsend, J.; Micucci, C. P.; Hymel, J. H.; Maroulas, V.; Vogiatzis, K. D. Representation of Molecular Structures with Persistent Homology for Machine Learning Applications in Chemistry. *Nat. Commun. 2020 111* **2020**, *11* (1), 1–9. https://doi.org/10.1038/s41467-020-17035-5.

(26) Meng, Z.; Anand, D. V.; Lu, Y.; Wu, J.; Xia, K. Weighted Persistent Homology for Biomolecular Data Analysis. *Sci. Reports 2020 101* **2020**, *10* (1), 1–15. https://doi.org/10.1038/s41598-019-55660-3.

(27) Qin, S.; Jin, T.; Van Lehn, R. C.; Zavala, V. M. Predicting Critical Micelle Concentrations for Surfactants Using Graph Convolutional Neural Networks.

(28) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking Graph Neural Networks for Materials Chemistry. *npj Comput. Mater. 2021 71* **2021**, *7* (1), 1–8. https://doi.org/10.1038/s41524-021-00554-0.

(29) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites. *J. Chem. Theory Comput.* **2018**, *14* (8), 4427–4437. https://doi.org/10.1021/ACS.JCTC.8B00253.

(30) Smith, A. D.; Dłotko, P.; Zavala, V. M. Topological Data Analysis: Concepts, Computation, and Applications in Chemical Engineering. *Comput. Chem. Eng.* **2021**, *146*, 107202. https://doi.org/10.1016/J.COMPCHEMENG.2020.107202.

(31) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine Learning of Molecular Properties: Locality and Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241727. https://doi.org/10.1063/1.5005095.

(32) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-Guzik, A. Machine-Learned Potentials for next-Generation Matter Simulations. *Nat. Mater. 2021 206* **2021**, *20* (6), 750–761. https://doi.org/10.1038/s41563-020-0777-6.

(33) Jinnouchi, R.; Miwa, K.; Karsai, F.; Kresse, G.; Asahi, R. On-the-Fly Active Learning of Interatomic Potentials for Large-Scale Atomistic Simulations. *J. Phys. Chem. Lett.* **2020**, *11* (17), 6946–6955. https://doi.org/10.1021/ACS.JPCLETT.0C01061.

*(34) Tran, K.; Ulissi, Z. W. Active Learning across Intermetallics to Guide Discovery of Electrocatalysts for CO2 Reduction and H2 Evolution. *Nat. Catal. 2018 19* **2018**, *1* (9), 696–703. https://doi.org/10.1038/s41929-018-0142-1.

The authors emphasize a need to explore the electrocatalyst information space systematically due to the large number of possible electrocatalyst candidates that are active, stable, and selective for CO_2 reduction and H_2 evolution. Machine learning through an active learning framework is introduced to explore adsorption sites, surfaces, and material species which reduces these physical features into numerical representations or fingerprints for reduced dimension correlations and more managable search spaces.

(35) Chew, A. K.; Walker, T. W.; Shen, Z.; Demir, B.; Witteman, L.; Euclide, J.; Huber, G. W.; Dumesic, J. A.; Lehn, R. C. Van. Effect of Mixed-Solvent Environments on the Selectivity of Acid-Catalyzed Dehydration Reactions. *ACS Catal.* **2019**, *10* (3), 1679–1691. https://doi.org/10.1021/ACSCATAL.9B03460.

The authors utilize solvation free energy calculations and experiments to reveal the influence of solvent composition on an individual reaction step's rate resulting in a change of selectivity for Brønsted acid-catalyzed 1,2-propanediol dehydration in pure water and aqueous mixtures of polar aprotic cosolvents. This is the first step towards a systematic screening of aqueous and non-aqueous solvent systems for acid-catalyzed processes.

(36) Kelkar, A.; Dallin, B.; Lehn, R. Van. Identifying Nonadditive Contributions to the Hydrophobicity of Chemically Heterogeneous Surfaces via Dual-Loop Active Learning. **2021**. https://doi.org/10.33774/CHEMRXIV-2021-668S4.

(37) Walker, T. W.; Frelka, N.; Shen, Z.; Chew, A. K.; Banick, J.; Grey, S.; Kim, M. S.; Dumesic, J. A.; Lehn, R. C. Van; Huber, G. W. Recycling of Multilayer Plastic Packaging Materials by Solvent-Targeted Recovery and Precipitation. *Sci. Adv.* **2020**, *6* (47). https://doi.org/10.1126/SCIADV.ABA7599.

(38) Sánchez-Rivera, K. L.; Zhou, P.; Kim, M. S.; Chávez, L. D. G.; Grey, S.; Nelson, K.; Wang, S.-C.; Hermans, I.; Zavala, V. M.; Lehn, R. C. Van; Huber, G. W. Reducing Antisolvent Use in the STRAP Process by Enabling a Temperature-Controlled Polymer Dissolution and Precipitation for the Recycling of Multilayer Plastic Films. *ChemSusChem* **2021**, *14* (19), 4317–4329. https://doi.org/10.1002/CSSC.202101128.

(39) Zhou, P.; Sánchez-Rivera, K. L.; Huber, G. W.; Van Lehn, R. C. Computational Approach for Rapidly Predicting Temperature-Dependent Polymer Solubilities Using Molecular-Scale Models. *ChemSusChem* **2021**, *14* (19), 4307–4316. https://doi.org/10.1002/CSSC.202101137.

(40) Tullo, A. H. Eastman Will Build a $250 Million Plastics Recycling Plant. *Chemical & Engineering News*. 2021, pp 1–2.

(41) Bregante, D. T.; Tan, J. Z.; Schultz, R. L.; Ayla, E. Z.; Potts, D. S.; Torres, C.; Flaherty, D. W. Catalytic Consequences of Oxidant, Alkene, and Pore Structures on Alkene Epoxidations within Titanium Silicates. *ACS Catal.* **2020**, *10* (17), 10169–10184. https://doi.org/10.1021/ACSCATAL.0C02183.

(42) Bregante, D.; Chan, M.; Tan, J. Z.; Ayla, E. Z.; Nicholas, C. P.; Shukla, D.; Flaherty, D. The Shape of Water in Zeolites and Its Impact on Epoxidation Catalysis. **2020**. https://doi.org/10.26434/CHEMRXIV.13288805.V1.

(43) König, M.; Vaes, J.; Klemm, E.; Pant, D. Solvents and Supporting Electrolytes in the Electrocatalytic Reduction of CO2. *iScience* **2019**, *19*, 135–160. https://doi.org/10.1016/J.ISCI.2019.07.014.

(44) Vasilyev, D. V.; Dyson, P. J. The Role of Organic Promoters in the Electroreduction of Carbon Dioxide. *ACS Catal.* **2021**, *11* (3), 1392–1405. https://doi.org/10.1021/ACSCATAL.0C04283.

(45) Monteiro, M. C. O.; Dattila, F.; Hagedoorn, B.; García-Muelas, R.; López, N.; Koper, M. T. M. Absence of CO2 Electroreduction on Copper, Gold and Silver Electrodes without Metal Cations in Solution. *Nat. Catal. 2021 48* **2021**, *4* (8), 654–662. https://doi.org/10.1038/s41929-021-00655-5.

(46) Goodpaster, J. D.; Bell, A. T.; Head-Gordon, M. Identification of Possible Pathways for C–C Bond Formation during Electrochemical Reduction of CO2: New Theoretical Insights from an Improved Electrochemical Model. *J. Phys. Chem. Lett.* **2016**, *7* (8), 1471–1477. https://doi.org/10.1021/ACS.JPCLETT.6B00358.

(47) Sundararaman, R.; Schwarz, K. Evaluating Continuum Solvation Models for the Electrode-Electrolyte Interface: Challenges and Strategies for Improvement. *J. Chem. Phys.* **2017**, *146* (8), 084111. https://doi.org/10.1063/1.4976971.