Topological Analysis of Molecular Dynamics Simulations using the Euler Characteristic

Alexander Smith,[†] Spencer Runde,[†] Alex K. Chew,[†] Atharva S. Kelkar,[†] Utkarsh Maheshwari,[‡] Reid C. Van Lehn,[†] and Victor M. Zavala[*,†]

[†]*Department of Chemical and Biological Engineering, University of Wisconsin, Madison*

[‡]*Department of Electrical and Computer Engineering, University of Wisconsin, Madison*

E-mail: victor.zavala@wisc.edu

Abstract

Molecular dynamics (MD) simulations are used in diverse scientific and engineering fields such as drug discovery, materials design, separations, biological systems, and reaction engineering. These simulations generate highly complex datasets that capture the 3D spatial positions, dynamics, and interactions of thousands of molecules. Analyzing MD datasets is key for understanding and predicting emergent phenomena and in identifying key drivers and tuning design knobs of such phenomena. In this work, we show that the Euler characteristic (EC) provides an effective topological descriptor that facilitates MD analysis. The EC is a versatile, low-dimensional, and easy-to-interpret descriptor that can be used to reduce, analyze, and quantify complex data objects that are represented as graphs/networks, manifolds/functions, and point clouds. Specifically, we show that the EC is an informative descriptor that can be used for machine learning and data analysis tasks such as classification, visualization, and regression. We demonstrate the benefits of the proposed approach through case studies that aim to understand and predict the hydrophobicity of self-assembled monolayers and the reactivity of complex solvent environments.

Introduction

The development of advanced molecular dynamics (MD) simulation methods has provided researchers the ability to rapidly screen for new chemistry, biological interactions, and materials.[1–3] For example, large-scale molecular simulations have been used in the screening of molecular organic frameworks (MOFs) for hydrogen storage.[4] These techniques are also employed in the study of soft materials such as proteins and polymers,[5–7] and in the design of self-assembled colloidal systems.[8] However, the analysis of MD datasets is challenging due to both their size and complexity; specifically, MD simulations can produce terabytes of data that require computationally efficient and scalable analysis methods, while the complexity of the data requires methods that are generalizable to a broad range of systems and that are robust to data heterogeneity and noise.[9]

Quantification and reduction of molecular simulation data has been traditionally conducted via order parameters and summarizing statistics such as radial distribution functions and correlation fields, particularly for condensed-phase systems.[10] These descriptors are usually computationally efficient, physically interpretable, and are derived from principles of physics and statistical mechanics. Moreover, such descriptors usually correlate to emergent properties of interest and thus can be used to construct predictive models. However, order parameters are typically designed for particular applications that meet specific assumptions (e.g., spatial isotropy or crystallinity) and are thus limited in scope.[11–14]

Another approach for quantifying molecular simulation data consists of using machine learning (ML) tools such as convolutional neural networks (CNNs) and autoencoders to extract informative descriptors from data.[15–19] ML tools are versatile in that they require few assumptions on the application and can be used for processing diverse data formats (e.g., images, tensors, graphs). However, descriptors extracted using ML

tools can be difficult to interpret and training predictive models based on such descriptors may require large numbers of parameters and large amounts of labeled MD data, which can be computationally costly to produce.

In this work, we investigate the application of tools from topology for the analysis of MD simulation data. Topology focuses on characterizing the global structure (e.g., connectivity, continuity) of shapes and objects and has been gaining attention in diverse scientific and engineering fields.[20,21] In the context of molecular simulations, connectivity and continuity of molecules such as polymers, proteins, and other molecules have been studied via topology. These topological methods have mostly relied on the use of applied knot theory, which can quantify the entangled structures of molecules and use this information to predict thermodynamic bulk properties. These techniques have also been used for connecting the knotting of molecules with their reactivity and function (e.g., DNA recombinase enzymes).[22–24] In these applications, however, the connection between the molecular data object and its topological representation (i.e., knots) is limited to specific settings. Recent work has also focused on the application of persistence homology to molecular simulation data.[20,25] Persistent homology has been shown to provide powerful characterizations of the topology and geometry of data.[25–30] However, the outputs of these methods (e.g., persistence diagrams) can be difficult to directly integrate into data analysis and ML tasks without further transformation (e.g., vectorization and smoothing) and hyperparameter optimization.[21]

The main goal of this work is to demonstrate that topology can be applied to a broad range of molecular simulation settings. At the core of this approach is the observation that one can represent data as graphs and manifolds, which are versatile topological objects that can be efficiently quantified using a topological descriptor known as the Euler Characteristic (EC).[21,31] Graph representations for molecular simulation data have been widely

applied and are easy to justify from a physical standpoint.[32,33] An example arises in the analysis of non-covalent bonding networks (e.g., hydrogen bonding networks in water). Here, the individual water molecules can be considered vertices of a graph with non-covalent bonds (e.g., hydrogen bonds) representing edges between the vertices. Manifold representations for molecular simulations arise when there is a continuous function describing behavior over a space (or surface) of a simulation. An example of a manifold representation arises in the analysis of time-averaged spatial density, where density is computed at each spatial location resulting in a continuous function over the entire simulation space.[20,21] Manifolds can be extended to more complex spaces, such as the surface of a molecule, nanoparticle, polymer, or protein.[34] In these approaches, the surface is treated as a manifold and physical characteristics (e.g., hydrophobicity, charge, forces, curvature) represent functions on the manifold. However, graph and manifold representations of molecular simulation data can be high-dimensional and not directly amenable to common analysis tasks (e.g. classification, regression, visualization). Thus, there is need for simple and computationally efficient methods for reducing and quantifying these topological data representations. This characterization can be accomplished by performing a decomposition of the space into a set of independent *topological bases* that capture basic topological features such as holes, connected components, and voids. The EC is a scalar integer quantity that is defined as the alternating sum of the rank of these topological bases of an object. The EC is often combined with a data processing technique known as *filtration*, which enables the characterization of more complex topological objects such as matrices, images, fields/functions, and weighted graphs.[21,35] The filtration process gives rise to the so-called *EC curve*, which is a function that summarizes how topological features emerge and disappear through the filtration process. Compared to descriptors obtained from persistent homology (e.g., persistence diagrams), the EC curve provides a quantifiable and easy-to-interpret descriptor of complex data objects.

To illustrate the efficiency and effectiveness of the EC, we provide studies arising in a couple of complex molecular simulation systems. The first simulation system is the measurement of hydrophobicity on the surface of 2D self-assembled monolayers (SAMs), and the second system is the analysis of 3D solvation effects on acid-catalyzed reactions systems for biomass processing. Previous work suggests that the topology and geometry of water plays a critical role in understanding and predicting emergent physical and chemical properties for both of these systems.[36,37] Thus, in both examples, we are focused on quantifying the topological structures and patterns emerging from solvent behavior at the surface of a SAM and around biomass-relevant reactants to quantify hydrophobicity and reactivity, respectively. We also show that the EC can be used to quantify different forms of data representations typically used in these types of MD simulations (hydrogen bond networks and density fields). Specifically, we show that the EC is a descriptor that correlates strongly with emergent properties and this enables the construction of low-dimensional and effective predictive models that are significantly more computationally tractable than recently-developed ML models such as CNNs. We show that simple regression models that take the EC as input are able to accurately predict the hydration free energy of simulated 2D SAMs and the change in reactivity due to solvation effects in acid-catalyzed reaction systems. These studies also illustrate the stability of the EC in quantifying noisy MD data and the physical intuition that can be gained through topological analysis. Moreover, we show that the EC can be used to monitor the dynamic evolution of topology in these MD systems which can be used, for instance, to determine when a system has achieved a topological steady-state. We also note that solvent-rich processes (like those studied here) are common in a diverse range of systems analyzed through MD, further supporting the general relevance of these methods. All code and data needed for reproducing our results are provided.

Topology of Graphs and Manifolds

To analyze MD simulation data with the EC, we begin by defining a couple of fundamental topological data representations: graphs and manifolds.

A graph is a 2D topological object that consists of an ordered pair $G(V, E)$, where V represents a set of *vertices* and E represents a set of paired vertices known as *edges*. Edges represent relationships (connectivity) between vertices. A graph representation for an atomistic simulation of water is shown in Figure 1. Here, the water molecules are represented as vertices and hydrogen bonds between molecules are represented as edges. This graph (network) representation can be used to understand how water is interacting both locally and globally, by quantifying specific features, such as the number of *cycles* and *connected components* of the graph. A cycle represents a path that traverses edges on a graph starting at a particular vertex v_j and ending at that same vertex v_j. Physical examples of graph cycles are found in tetramer, pentamer, and hexamer water structures.[38,39] A connected component is a subset of a graph $C(V_C, E_C) \subseteq G(V, E)$ in which any vertex $v_i \in V_C$ of the subgraph can reach any other vertex $v_j \in V_C$ by traversing edges of the subgraph $\{v_i, v_j\} \in E_C$, and is disconnected from all other subsets of the graph. In other words, the number of connected components is the number of connected partitions of a graph. In a hydrogen bonding network for water, connected components help us understand the physical state of the system; for example, in a condensed (or crystalline) state, there will be many hydrogen bonds present, reducing the total number of connected components but increasing the total number of molecules in each connected component. The opposite would hold true for a system acting as an ideal gas; here, no bonds are formed and thus each molecule exists as its own connected component. Data can also be encoded in a graph object (in nodes and edges) using functions $f : V \to \mathbb{R}$ and $f : E \to \mathbb{R}$. Values attached to nodes or edges are typically called weights or features; as such, graphs that encode data are also known as weighted graphs.

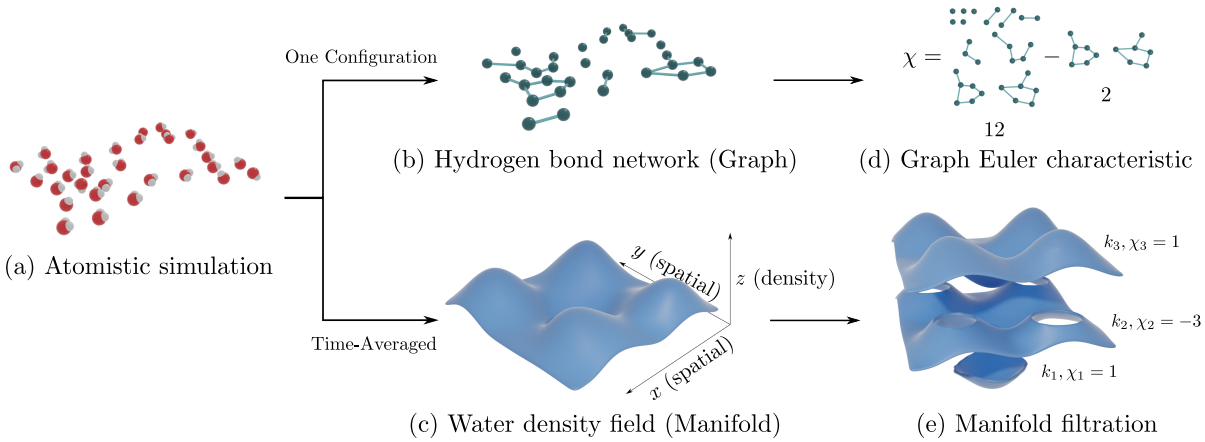


Figure 1: Graph and manifold representations of a molecular simulation of water. (a) Snapshot of an atomistic simulation of water (only some molecules are shown). (b) A graphical representation of the hydrogen bonding network formed between water molecules within the simulation, and (c) a density field derived from time-averaging water molecule positions during the simulation. The density field is represented as a manifold $\mathcal{M}$, with a continuous function $f : \mathcal{M} \to \mathbb{R}$ that maps each point of the manifold to a corresponding water density value; visualized here by changes in the height of the surface. (d,e) Represents the EC (χ) quantification of the graph and manifold data representations. (d) The graph is quantified by subtracting the total number of cycles from the total number of connected components in the graph ($\chi = 12 - 2 = 10$). (e) The manifold is quantified through a filtration. At multiple increasing density thresholds $k_i \in \mathbb{R}$, the EC χ_i is computed by subtracting the total number of holes from the total number of connected components in the filtered manifold $x \in \mathcal{M} : f(x) \leq k_i$. We note that filtered manifolds all originate from the same data object and that the vertical layout is meant to illustrate the topological changes as the filtration is performed. The paired values $\{k_i, \chi_i\}$ are used to construct an Euler characteristic curve.

Manifolds are also versatile topological data representations that can capture continuous forms of information (e.g., 3D density fields) in high-dimensional spaces. This contrasts with graph representations, which capture discrete characteristics of a 2D data object (e.g., number of bonds, molecules, clusters). A manifold $\mathcal{M}$ is a topological space that *locally* resembles a Euclidean space; this means that the neighborhood of a point $x \in \mathcal{U}$ in an n-dimensional manifold (with $\mathcal{U} \subseteq \mathcal{M}$) can be mapped to $n-$dimensional Euclidean space through a continuous, bijective function. These neighborhoods and associated mappings are also known as *charts*. For example, the surface of the Earth is a 2D manifold and we can map the curved surface of the Earth to a flat Euclidean plane (i.e., a 2D Euclidean space) using a chart in order to measure properties such as distances or areas. The general nature of manifolds allows them to represent a broad range of structures, shapes, and complex geometric objects in molecular simulations (e.g., surface of a protein or a nanoparticle). Manifolds can also have encoded data on them (e.g., Earth surface temperature), which is captured using a continuous function $f : \mathcal{M} \to \mathbb{R}$. In Figure 1, we present a manifold representation for a 2D simulation of water. Here, the simulation domain (e.g., a 2D plane) is a 2D manifold and we define a continuous function that captures the time-averaged density of water at each location in the domain.

The Euler Characteristic

Graph and manifold representations are able to capture both discrete and continuous information within a given simulation and their topology can be directly quantified/summarized using a descriptor known as the Euler characteristic (EC).[21] The EC is denoted as $\chi \in \mathbb{Z}$ and is mathematically defined as the alternating sum of the rank of topological bases for a given space known as Betti numbers $\beta_i \in \mathbb{Z}_+$, where $i \in \mathbb{Z}_+$ represents the dimensionality of the topological basis:

$$\chi := \sum_{i=0}^{n} (-1)^i \beta_i \tag{1}$$

Importantly, the topological bases of a space (e.g., connected components, holes, voids) are preserved under deformations such as stretching, twisting, and bending (are topological invariants). For any topological space of n-dimensions, there can only exist topological bases up to that given dimension. For example, a 3D space can only contain β_0 (representing connected components), β_1 (representing holes and cycles), and β_2 (representing voids and cavities) in the space. Figure 1 represents a hydrogen bonding network for an atomistic MD simulation of water as a graph. A graph is a 2D topological space, and thus has a couple of Betti numbers β_0 and β_1. Figure 1 illustrates that, for a simulation snapshot, the number of connected components is 12, the number of cycles is 2, and thus the EC is $\chi = 12 - 2 = 10$.

Manifold Filtrations

Analysis of data represented as manifolds (or weighted graphs) requires an added processing step known as a *filtration*. A filtration quantifies the topology of *sublevel sets* of the manifold. Given an n-dimensional manifold $\mathcal{M}$ and a continuous function $f : \mathcal{M} \to \mathbb{R}$, a *sublevel set* of the manifold is defined as $\mathcal{M}_{k_i}$ that contains points $\{x \in \mathcal{M} : f(x) \leq k_i\}$, where $k_i \in \mathbb{R}$ represents our *filtration threshold*. Hence, we can construct nested sublevel sets at increasing filtration thresholds for the manifold:

$$\mathcal{M}_{k_1} \subseteq \mathcal{M}_{k_2} \subseteq ... \subseteq \mathcal{M}_{k_n} \subseteq \mathcal{M} \tag{2}$$

where $k_1 < k_2 < ... < k_n$ represent our filtration thresholds, and $\mathcal{M}$ represents the original

manifold. We can measure/quantify the topology of these nested sublevel sets with the EC at each filtration threshold $\{\chi_1, \chi_2, ..., \chi_n\}$. We ultimately obtain an ordered pair of values $\{k_i, \chi_i\}$, which characterize the topology of the manifold and its associated function. An illustration of the filtration process is found in Figure 1 for a 2D atomistic simulation, where time-averaged water density is analyzed over the space of the simulation. We have selected three different filtration values $k_1 \leq k_2 \leq k_3$ corresponding to the three sublevel sets. Similar to the graph example, we compute the EC by counting the total number of n-dimensional topological bases (β_0, β_1 for a 2D manifold). The bottom most sublevel set at filtration value k_1 represents a single connected component, capturing a local minima in the function f, and resulting in $\chi_1 = 1 - 0 = 1$. As the filtration threshold increases to k_2, a single connected component remains but four holes (i.e., cycles) are formed indicating the presence of local maxima of the function f, which results in an EC value of $\chi_2 = 1 - 4 = -3$. The final filtration threshold k_3 returns the original manifold, which is a single connected component: $\chi_3 = 1 - 0 = 1$ (further filtration of the space will not change the manifold topology). The filtration of a weighted graph is conducted in an analogous manner (by eliminating nodes or edges in which the data is below a certain threshold value). We note that filtration operations are easy to conduct and are thus scalable.

Applications of Topology in MD Simulations

The EC of graphs and manifolds provides a topological descriptor that quantifies complex structures and patterns that arise in MD simulations. The EC can be used to conduct a wide variety of ML and data analysis tasks such as visualization, clustering, regression, and classification. Here, we demonstrate that the EC of a molecular system correlates strongly to emergent physical and chemical characteristics. As such, we show that the EC can be used as an informative descriptor to predict emergent behavior. Moreover, we show that such predictions can be conducted using simple linear regression models,

which contrasts with existing approaches based on CNNs.

The first set of simulations studied involve self-assembled monolayers (SAMs); here, we use the EC to predict the hydration free energy of the 2D SAM surface. The second set of simulations aims to predict the reactivity of a molecule based on the topology of a solvent environment composed of water and a cosolvent. These examples were specifically chosen because they were previously studied using advanced CNNs and thus have a frame of reference.[36,40] We also highlight that these molecular systems are solvent-dominated; as such, their emergent properties are known to be influenced by the spatial structure and correlations of the solvent environment.[41,42] This information will be quantified directly using the EC of graph and manifold representations of such environments.

Implementation details for both case studies can be found in the Supplementary Information. All code and data needed to reproduce the results can be found in https://github.com/zavalab/ML/tree/master/MD_Euler.

Hydrophobicity on the Surface of Self-Assembled Monolayers

We study the surfaces of SAMs using an MD simulation dataset obtained from recent work of Kelkar and co-workers.[36] The SAM structures are built from a planar array of alkanethiol ligands with hydroxyl, amine, or amide end groups. Each simulation consists of a single SAM solvated by bulk water. A simulation snapshot can be found in Figure 2a. A total of 50 different SAMs were created (22 having hydroxyl groups, 14 with amine end groups, and 14 with amide end groups). The partial charges of the end groups are modulated using a scaling factor that simulates changes in the polarity of the SAM surface. Additional details on the MD simulation methodology and parameters are available in the work of work of Kelkar and co-workers.[36]

Our goal is to study the topology of water in a thin interfacial layer located at the SAM surface. To do so, we leverage the topological structures formed by water at the SAM-water interface to directly predict the hydration free energy (HFE) of the SAM through linear regression. The HFE is a property that captures surface hydrophobicity behavior and is key in understanding protein adsorption.[43,44] A common method for computing the HFE in molecular simulations is indirect umbrella sampling (INDUS). This method is highly accurate but it is computationally expensive, as it requires sampling of a low-probability event.[45] The dataset developed in previous work leveraged INDUS to create a set of SAM simulations with computed HFE values. Such simulations were used to train and test a CNN that directly predicts HFE from the SAM interfacial water structure (represented as a 2D water density field). Here, we instead develop a linear regression model using the EC of the SAM interfacial water structure.

To quantify the topology of the SAM structure, we subsample each SAM simulation using non-overlapping sets of 800 simulation time steps (800 picoseconds). We then compute a time-averaged EC value for the hydrogen bonding network G_{ww} within the interfacial layer denoted as $\langle \chi(G_{ww}) \rangle$ for each subsample. We will represent the EC value for graphs G as $\chi(G)$ to distinguish it from the EC values computed for manifolds (denoted as χ). The presence (or absence) of hydrogen bonds between water molecules was computed using the Luzar-Chandler criterion.[46,47] For an interval of simulation time points $t \in [a, b)$ where $b, a \in \mathbb{Z}_+$ and $b - a = 800$, we compute the EC value of the hydrogen bonding network $\chi(G_{ww})_t$ at each simulation time point $t \in 0, 1, ..., 800$. The time-averaged hydrogen bonding EC value is computed as:

$$\langle \chi(G_{ww}) \rangle := \frac{1}{(b-a)} \sum_{t=a}^{b} \chi(G_{ww})_t \tag{3}$$

In other words, $\langle \chi(G_{ww}) \rangle$ captures the time-average topology of the hydrogen bonding network G_{ww}. The details outlining the practical computation of the EC for these simulations can be found in the supporting information and code shared in this manuscript.

The selection of hydrogen bonding criterion can impact the computed EC value. However, time averaged water density information is unaffected by this choice. Thus, incorporating both hydrogen bonding and density information will provide a data representation that is robust to particular choices of criterion. Thus, we represent the SAM surface as a 2D manifold $\mathcal{M}$ and treat the time-averaged water density at the interfacial layer as a continuous function on the manifold $f : \mathcal{M} \to \mathbb{R}$ as shown in Figure 2. The manifolds are constructed by binning water molecule positions at the interfacial layer in a 20×20 grid, where each grid point accounts for 0.1 nm^2 area, with depth of 0.3 nm, on the surface of a SAM over the period of 800 simulation time steps. The accumulated bin data are then averaged to obtain a continuous water density function for our surface manifold $\mathcal{M}$. This representation matches the representation of Kelkar and co-workers used as input to a CNN to allow for direct model comparison.[36] We performed a manifold filtration to quantify spatial density fluctuations in water at the SAM surface with an EC curve. We recall that the EC curve is the set of paired filtration thresholds k_i and EC values χ_i of the filtered sublevel sets $\mathcal{M}_{k_i}$. The filtration thresholds k_i represent water densities given in units of molecules/nm^2.

A visualization of the EC curve for a time-averaged water density field derived from a SAM simulation (HFE = $33k_BT$) is shown in Figure 2. We illustrate the topological changes in the manifold as we perform our filtration: $\mathcal{M}_{0.05} \subseteq \mathcal{M}_{0.07} \subseteq \mathcal{M}_{0.12} \subseteq \mathcal{M}_{0.20}$. We see in the first sublevel set $\mathcal{M}_{0.05}$ that connected components start to form. These are a direct result of local minima (e.g., areas of low water density) near the SAM surface, and result in a positive EC value. As the filtration threshold increases, there is an increase

in the number of connected components representing areas of low water density and an increased EC value $\mathcal{M}_{0.07}$. As the filtration threshold increases, we begin to pass saddle points in the density function where individual components merge, resulting in a single connected component with many holes (representing local maxima of the water density) and a corresponding negative EC value $\mathcal{M}_{0.12}$. The filtration then reaches a threshold in which the topology of the sublevel set is equal to the topology of the original manifold $\mathcal{M}_{0.20} = \mathcal{M}$. In this case, the original manifold $\mathcal{M}$ is a single connected component with an EC value $\chi = 1$. Details on the practical computation of the EC for sublevel sets can be found in the supporting information and code.

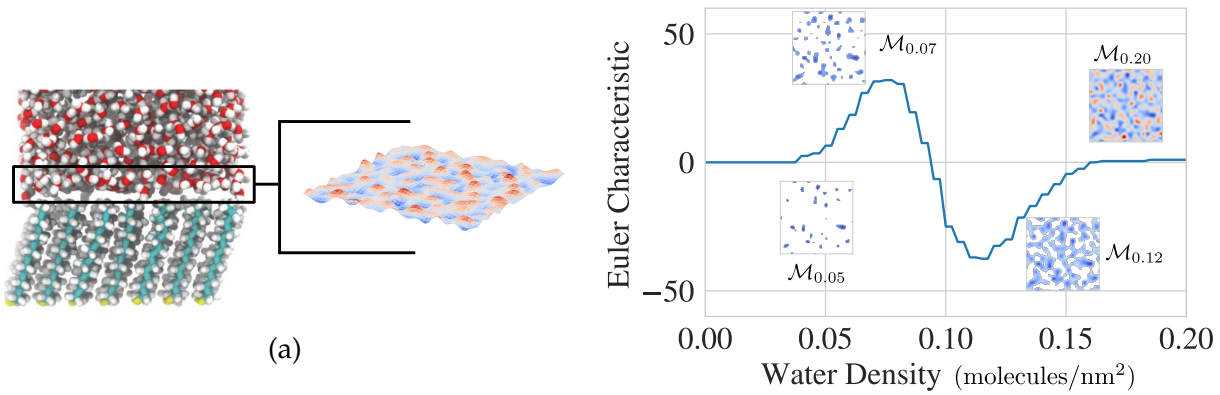


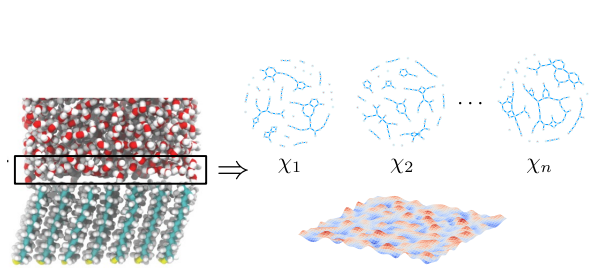
Figure 2: (a) Interfacial water density field derived from a SAM simulation. The 2D density field is represented as a manifold $\mathcal{M}$ with a continuous function $f : \mathcal{M} \to \mathbb{R}$ that maps each manifold location to its corresponding water density value visualized here by color (blue = low, red = high) and surface height. (b) The EC curve obtained from level set filtration of the density field. The EC curve is created by thresholding the density field function/manifold, creating multiple nested submanifolds $\mathcal{M}_{k_i}$, and then computing the EC of each submanifold. The EC curve is constructed from the paired values $\{k_i, \chi_i\}$. We visualize the corresponding submanifolds as the density threshold increases from $k_0 = 0$ to $k_n = 0.2$.

Figure 3 illustrates the input for a regression model derived from subsets of a SAM simulation (HFE = $33k_BT$ as labeled by INDUS). Figure 3 also reveals that topological representations are invariant to different types of deformation of the data.[48] Specifically,

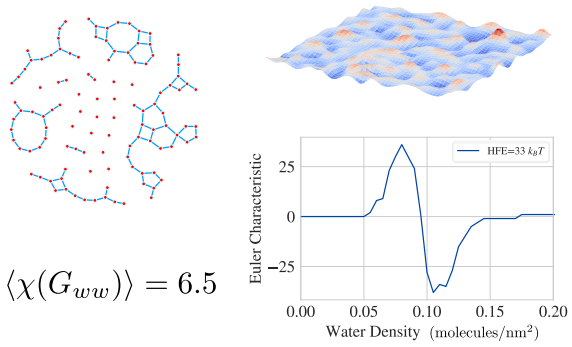
we see that both the hydrogen bond network and water density manifold topology varies significantly over time, but their corresponding topologies (captured by $\langle \chi(G_{ww}) \rangle$ and EC curves) are similar. This result is of practical relevance, because it shows that the EC can be used to monitor the dynamics of topology (e.g., to determine when the system is undergoing a topological transition or has reached steady-state). For example, Figure 4 illustrates the topological convergence of $\chi(G_{ww})$ for a simulation of a SAM with hydroxy-undecanethiol endgroups. The topology of the hydrogen bonding networks converges after approximately 1500 picoseconds of simulation time. This portion of the simulation is not included in the training/testing data.

Before building a linear regression model that predicts HFE from the SAM structure, we first determined if there was indeed a relationship between the topology of the SAM interfacial water structure and the HFE. In Figure 5, we find that there is a strong correlation between the topology of the time-averaged water density field at the SAM surface and its emergent HFE. Specifically, the local minima and maxima (i.e., critical points) of the density change in both shape and magnitude as the HFE for the SAM is changed. These topological changes are captured effectively using the EC curve. At low HFE values (e.g., HFE = $33k_BT$) we see that there are many critical points with relatively low magnitude; as HFE is increased (e.g., HFE = $100k_BT$) we see fewer critical points but the corresponding magnitude of these critical points is increased, which we see is reflected in the EC curves.

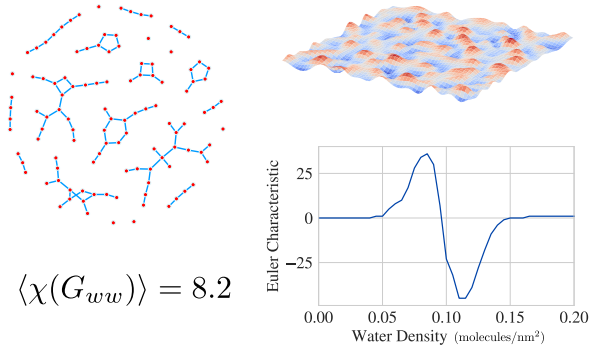
We further highlight the relationship between the SAM topology and the HFE by performing principal component analysis (PCA) on all 400 sampled EC curves derived from 40 simulations with precomputed HFE values. The EC curve of each sample is represented as a vector $x_j \in \mathbb{R}^m$ where $m = 20$ is our number of filtration thresholds and the entries of x_j are the EC values of the sublevel sets. Each vector is stacked into a matrix



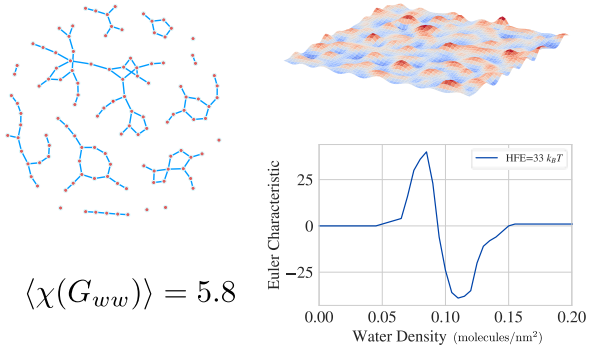
(a) Graph and Manifold Representations



(b) $t \in [0, 800)$



(c) $t \in [800, 1600)$



(d) $t \in [1600, 2400)$

Figure 3: (a) Illustration of the graph and manifold data representations derived from the SAM simulation shown in Figure 2 at HFE = $33k_BT$,. (b),(c) and (d) Representative hydrogen bond networks (graphs) and time-averaged density fields (manifolds) derived from the interfacial water in SAM molecular simulations over different time frames of the same simulation. They also contain the time-averaged graph EC $\langle\chi(G_{ww})\rangle$ and the EC curve created from a filtration of the associated density field. Each SAM simulation is split into multiple subsets of a single simulation $t \in [a, b)$, for which a corresponding density field EC curve and time-averaged graph EC is computed. We note the stability of both the time-averaged graph EC and the EC curve. The density fields and graphs are visually very different, but the topological measures of the graphs and density fields are almost identical throughout the simulation. This demonstrates the robustness of these topological descriptors in capturing the underlying characteristics of molecular simulations.

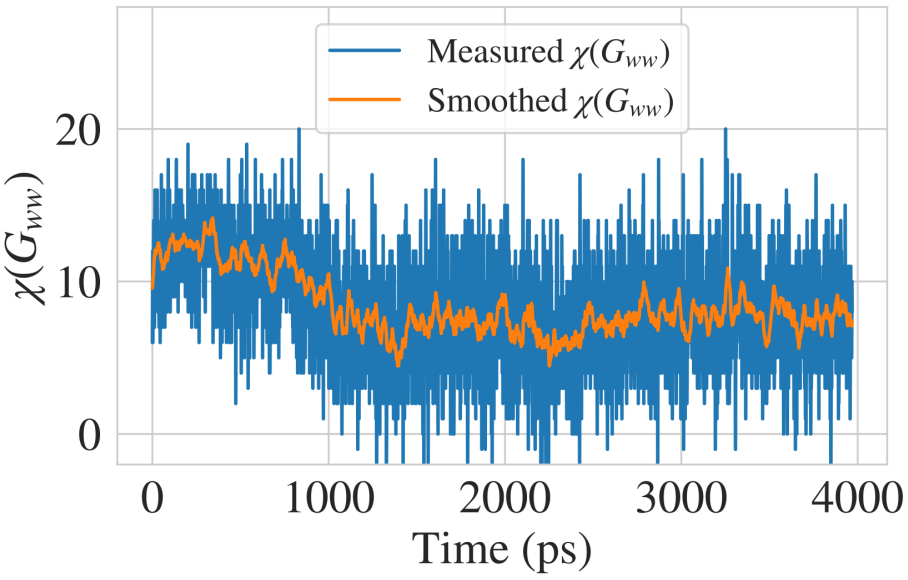


Figure 4: Demonstration of the topological convergence of the hydrogen bonding network in a simulation of a SAM with hydroxy-undecanethiol endgroups. We illustrate both the directly measured $\chi(G_{ww})$ and a smoothed value (rolling window average) to better emphasize the trending of $\chi(G_{ww})$. There is a clear topological convergence of the hydrogen bonding network found in the initial 1500 picoseconds of the simulation. This initial portion of the simulation is not included in the training/testing datasets.

$[x_1\ x_2\ \cdots\ x_n]^T \in \mathbb{R}^{n\times m}$. We apply a singular value decomposition to this matrix and visualize the data projected onto the two leading principal components; these components capture the low-dimensional structure of the EC and show that this is highly correlated with the HFE. These results confirm that the topology of the SAM affects the HFE and that such topology can be captured using the EC.

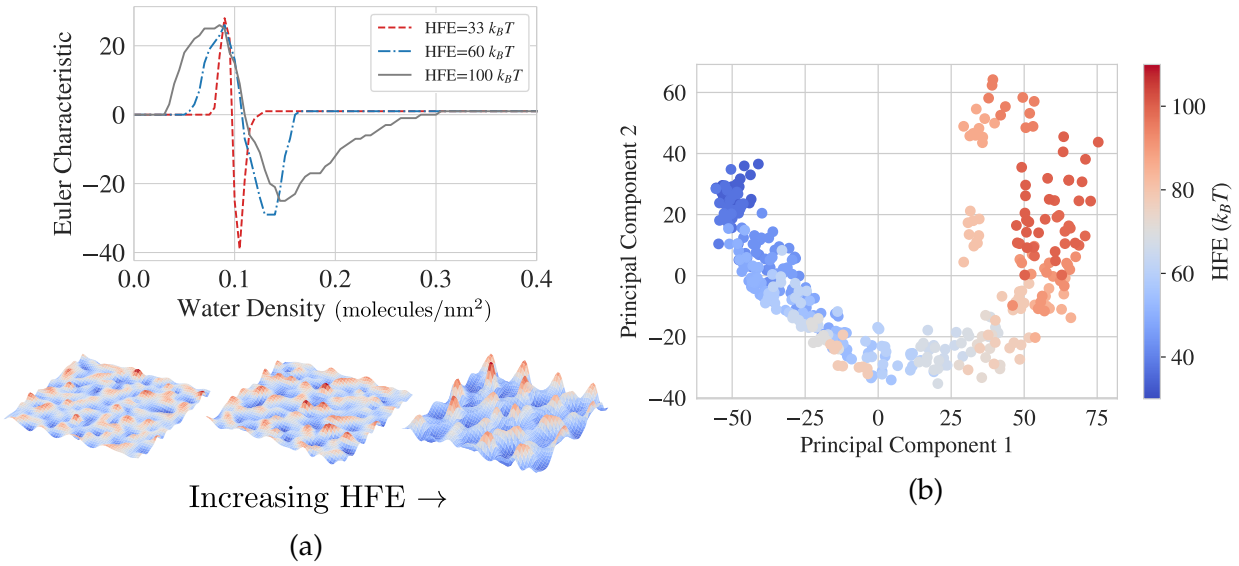


Figure 5: (a) EC curves for multiple density fields taken from simulations of SAMs with increasing HFE values. The impact of increased HFE on the SAM interfacial water layer is directly correlated with changes in the resulting EC curve. (b) Principal Component Analysis (PCA) is conducted on the EC curves for each density field from the set of SAM simulations. From (a) we see a continuous change in the EC curve that correlates with the HFE of the given SAM, in (b) we capture this continuous change and visualize a data structure that correlates directly with the HFE of the simulations through the first two principal components.

We next develop a linear regression model by using the water density EC curves and hydrogen bonding network EC values $\langle\chi(G_{ww})\rangle$ as model inputs. The model chosen is a linear support vector machine (SVM) model taken from the `LIBSVM` library.[49] We train the linear model using a set of 40 simulations with precomputed HFE values via the INDUS method (400 samples). Once the model has been trained, we test its prediction accuracy on a completely separate set of 10 SAM simulations (100 subsampled points). Our goal

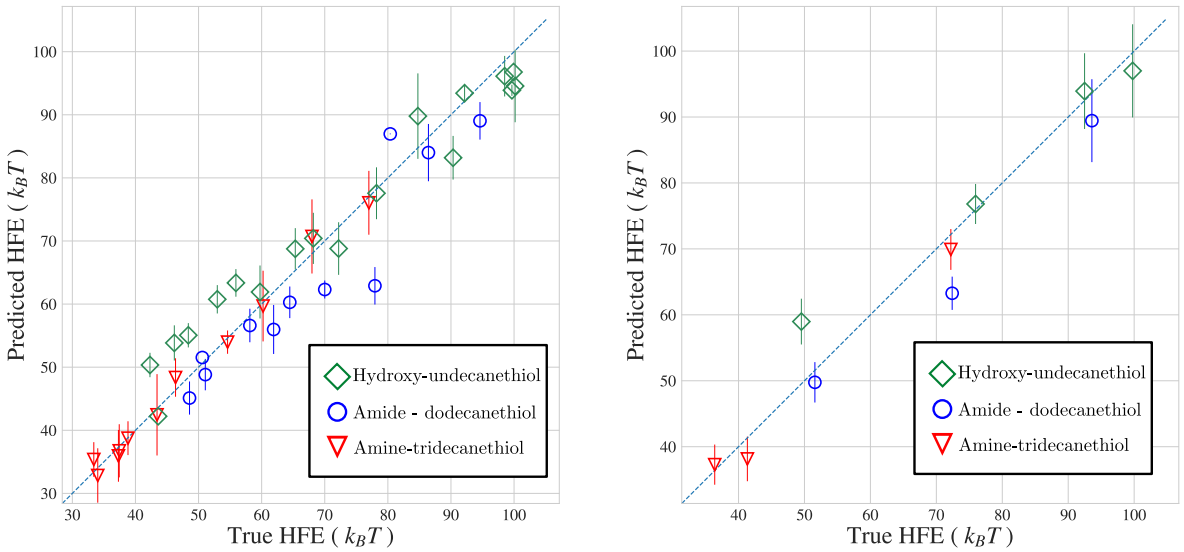
is to accurately predict the true HFE (computed via INDUS) for this separate set of simulations. The regression results for both model training and model testing are shown in Figure 6. We note that each point represents the mean HFE value predicted for the entire simulation, with the error bars representing a single standard deviation in the HFE computed from each of the simulation subsamples. We see that the linear model is able to predict the HFE for the testing set of simulations with little error (RMSE = 6.4 k_BT). Moreover, we have found that this model improves the results of a previously-developed 2D-CNN (RMSE = 8.0 k_BT), and is similar in performance to a 3D-CNN model (RMSE = 5.9 k_BT).[36] The computations required to both train and predict with the linear model are minimal compared to INDUS and to both machine learning models. Thus, this topological approach can be used in the analysis of high-throughput simulations or in screening for surfaces with optimal chemical or physical properties. We have also compared the results obtained using the EC with models trained on water density profiles that capture fluctuations in water density as a function of distance from the SAM surface. All details of this analysis are found in the supplementary information and code. The water density profile models have a large error (RMSE = 20 k_BT). This confirms previous results that water density profiles are unable to characterize surface hydrophobicity.[50] This finding is also supported by results from the areas of stochastic geometry and statistical thermodynamics which demonstrate that the EC can be used to distinguish molecular configurations and phase transitions that the density profiles and radial distribution functions cannot distinguish.[51,52]

The linear prediction model uses topological information obtained from both graph and manifold representations of the SAM. We have found that when either representation (density field or hydrogen bond network) is used independently for developing a model, the resulting prediction RMSE values increase. Removing the graph EC $\langle\chi(G_{ww})\rangle$ increases the RMSE of the model to 7.7 k_BT. Removing the manifold EC curve χ increases

the RMSE to 18 k_BT. These results indicate that spatial fluctuations in water density play a much larger role in characterizing hydration free energy, which is consistent with past studies of water density fluctuations.[36,44,45] Moreover, we have found that such models lead to severe underprediction of HFE for amide-dodecanethiol simulations. This may be explained by the tendency for amide end groups to form substantial hydrogen bonds with other amide end groups rather than water, which is unique for the surfaces studied.[53–55] This added complexity is captured effectively when combining topological information from both graph and manifold representations and highlights how such representations can provide complementary information.

Euler Characteristic Interpretability

An added benefit of linear models and our topological characterization of MD simulation data is interpretability, which in the physical sciences is often as important as prediction accuracy.[56] Linear models provide a level of interpretability that is not accessible when using more complex machine learning models, such as the previously trained CNN. For example, we can train a non-linear SVM using a radial basis function (RBF) kernel on this same dataset. The RBF SVM model performs slightly worse on the testing set than the linear SVM model (RMSE 6.9 k_BT - details found in supplementary code). Moreover, the non-linear SVM is no longer interpretable because the model coefficients do not correspond directly to the data features (e.g., the EC curve).[57] This issue is not present in the application of linear models because the magnitude and sign of the model coefficients correspond directly to the significance of a particular feature of the data. We can interpret the trained linear model's output $F \in \mathbb{R}$ (HFE) as the sum of the EC values weighted by the model coefficients $F = \sum_i c_i \chi_i + b\langle \chi(G_{ww}) \rangle + I$. Here, $c_i \in \mathbb{R}$ represent the model coefficients for the input EC curve, $b \in \mathbb{R}$ represents the model coefficient for the averaged graph EC, and $I \in \mathbb{R}$ is a constant value. Thus, a coefficient c_i, b with high magnitude sug-



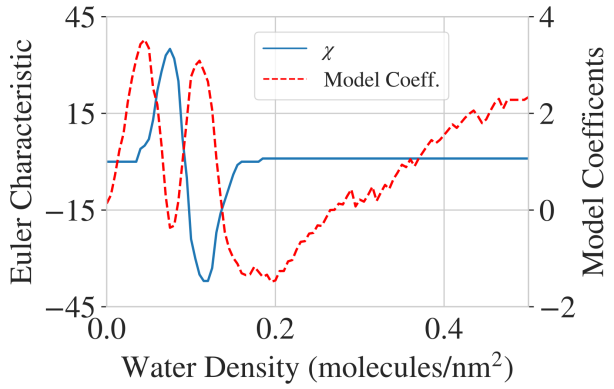
(a) Training.

(b) Testing.

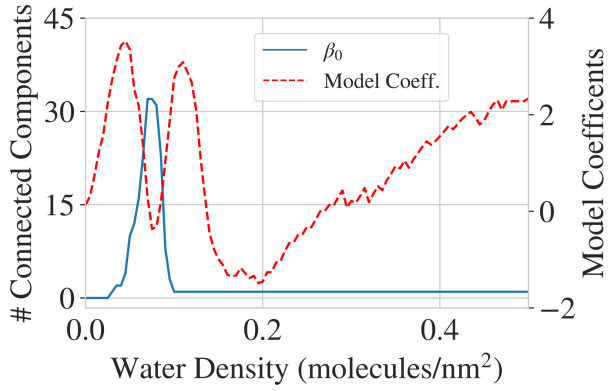
Figure 6: (a) Training data parity plot of predicted versus INDUS derived HFE. Linear regression is conducted using the corresponding EC curve and averaged graph EC $\langle \chi(G_{ww}) \rangle$ as inputs. The training data set is split into two portions, one for model training and the other for model validation. Predictions on the validation dataset are very accurate and suggest the linear model can obtain high accuracy in HFE prediction (RMSE = 5.1 k_BT). (b) Testing data parity plot of predicted versus INDUS derived HFE. The testing data consists of a completely separate set of SAM simulations not used in model training. For each simulation EC curves and $\langle \chi(G_{ww}) \rangle$ are measured. The trained linear model is then used to predict the HFE for the separate set of SAM simulations. The results demonstrate a high level of accuracy and low prediction error (RMSE = 6.4 k_BT), which is comparable to validation set accuracy as expected. Error bars in both plots represent a single standard deviation from the mean.

gests an area of the filtration that has significance in the prediction of HFE. We explore this in an analysis of water density EC curves at both low HFE (HFE = $33k_BT$) and high HFE (HFE = $100k_BT$). In Figure 7a and 8a, the solid (blue) line represents the EC values during a filtration χ_i, and the dashed (red) line represents the model coefficients $c_i \in \mathbb{R}$ associated with the EC at each level of the filtration i.

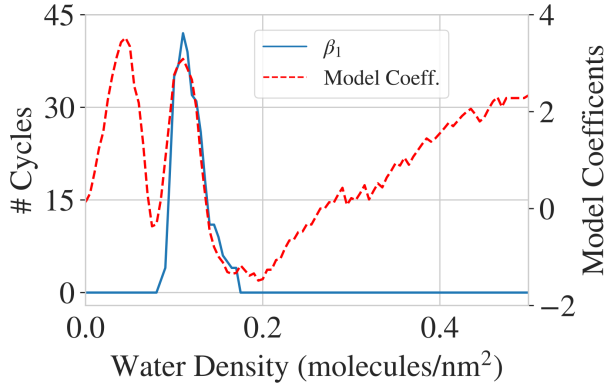
To interpret the results of our linear regression model, we recall that each value of the EC curve χ_i represents the alternating sum of the Betti numbers β_{0i} and β_{1i}. We recall that β_{0i} represents the number of connected components at filtration level i and β_{1i} represents the number of cycles (holes) at filtration level i. Thus, we can decompose the EC curve (Figures 7a & 8a) into two curves that reflect the individual Betti numbers during the filtration. These curves are known as Betti curves. The Betti curve for β_0 is found in Figure 7b for the low HFE simulation and in Figure 8b for the high HFE simulation. Viewing these individual curves provides a better understanding of how the linear model captures and interprets the topological characteristics of the data. For the low HFE simulation we see there is minimal weighting of β_0 during the filtration, which will result in a low HFE value prediction. For the high HFE sample, we see that the β_0 values during the filtration almost perfectly overlap with positive model coefficients. This will contribute to a high HFE value prediction. We see a similar trend in the analysis of β_1 (Figure 7c & 8c). For low HFE we see a large *positive* weight of β_1, this will result in a *negative* value in the model output F (low HFE prediction) because the EC is an alternating sum. For the high HFE sample we see a large *negative* weighting of β_1, which will result in a *positive* value in the model output (and a larger HFE prediction). Physically, these changes in topology represent differences in the spatial distribution of water on the SAM surface. For a high HFE sample we observe alternating areas of high and low density, which represent cycles β_1 and connected components β_0 that exist during a large portion of the filtration. For the simulation with low HFE, we observe water as being distributed homogeneously across



(a) $\chi = \beta_0 - \beta_1$ analysis

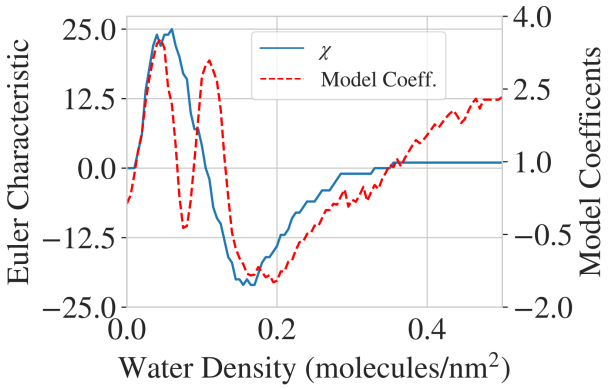


(b) β_0 analysis

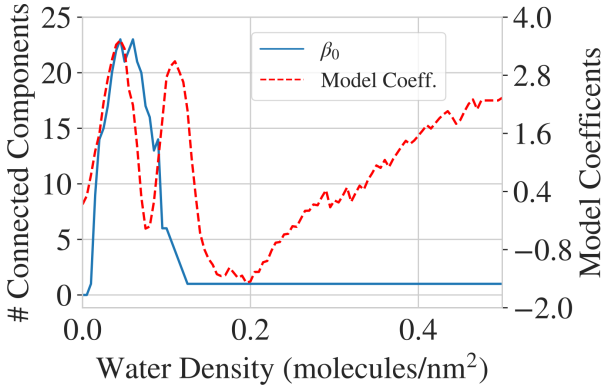


(c) β_1 analysis

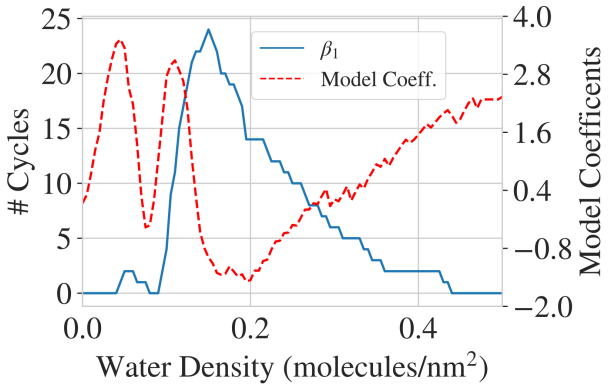
Figure 7: An analysis of the trained linear SVM model coefficients for a SAM surface with low HFE ($33k_BT$). (a) Overlay of model coefficients with the EC curve. (b) Overlay of model coefficients with the Betti curve representing β_0 (connected components) during the filtration. The model coefficients place almost no weight on the β_0 peak which corresponds with a low HFE prediction. (c) Overlay of model coefficients with the Betti curve representing β_1 (cycles) during the filtration. We observe a positive weighting of the β_1 peak which will result in a large *negative* value (low HFE) in the model output because the EC is an alternating sum of the Betti numbers. The shape of these curves is a direct result of the homogeneity of the water density distribution on a low HFE surface.



(a) $\chi = \beta_0 - \beta_1$ analysis



(b) β_0 analysis



(c) β_1 analysis

Figure 8: An analysis of the trained linear SVM model coefficients for a SAM surface with high HFE ($100k_BT$). (a) Overlay of model coefficients with the EC curve. (b) Overlay of model coefficients with the Betti curve representing β_0 (connected components) during the filtration. The model coefficients place a large positive weight on the β_0 peak which corresponds with a high HFE prediction. (c) Overlay of model coefficients with the Betti curve representing β_1 (cycles) during the filtration. We observe a negative weighting of the β_1 peak which will result in a large *positive* value (high HFE) in the model output because the EC is an alternating sum of the Betti numbers. The shape of these curves is a direct result of the heterogeneity of the water density distribution on a high HFE SAM surface. We see large fluctuations in the density on the SAM surface, which result in connected components and holes that exist during a large portion of the filtration.

the SAM surface. This results in small spatial fluctuations in water density on the surface which induce cycles and connected components that appear during a small portion of the filtration. This further illustrates how the EC curve can allow a simple linear model to identify complex changes in molecular density distribution in a simulation.

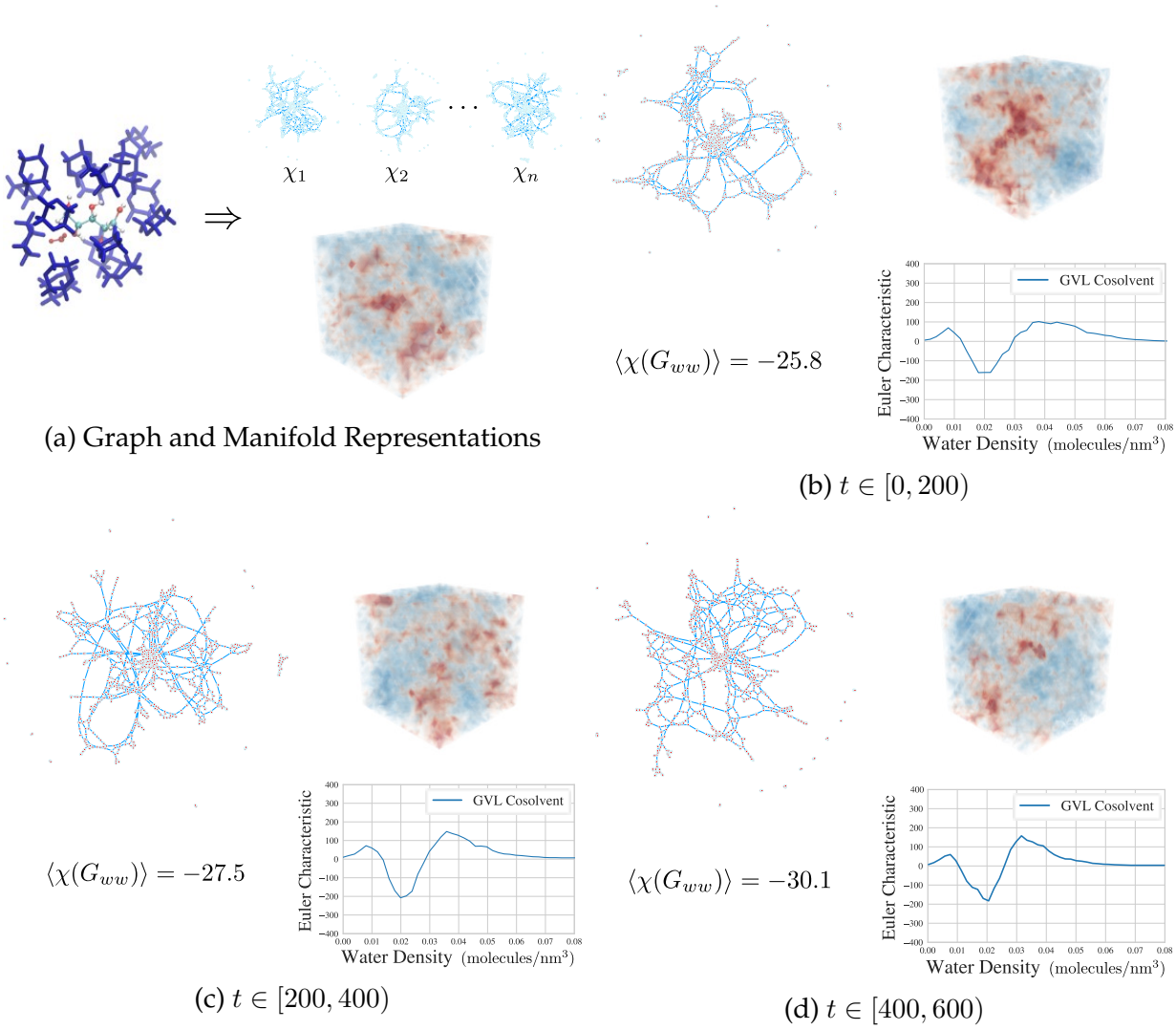
Solvent-Mediated Reactivity in Acid-Catalyzed Reactions

We now use the EC for understanding and predicting solvent-mediated reactivity of acid-catalyzed reactions based on the topology of water and cosolvent mixtures. Previous work has demonstrated that varying the cosolvent type and concentration in a cosolvent/water mixture impacts the relative reactivity of acid-catalyzed reactions for biomass conversion.[37,40] Walker and co-workers analyzed the influence of solvation towards reactivity by studying the structure of water in molecular simulations of a single reactant molecule in different cosolvent/water mixtures (snapshot shown in Figure 9). We demonstrate that the EC can be used to predict the solvent-mediated changes in reactivity as the concentration and type of cosolvent are varied, a task that previously used 3D CNNs.[40] The organic, polar aprotic cosolvents modeled in this study are dioxane (DIO), γ-valerolactone (GVL), tetrahydrofuran (THF), dimethyl sulfoxide (DMSO), acetonitrile (ACN), and acetone (ACE). Biomass-derived reactants modeled in this study are ethyl tert-butyl ether (ETBE), tert-butanol (TBA), cellobiose (CEL), glucose (GLU), levoglucosan (LGA), 1,2-propanediol (PDO), fructose (FRU), and xylitol (XYL); further details about the dataset can be found in the supporting information and work of Chew and co-workers.[40]

We again represent the MD simulation data as both graphs and manifolds and assess whether topological descriptors alone can predict solvent-mediated reaction rates (Figure 9). We propose this method of analysis because there is an established history of water enriched structures playing an important role in understanding and predicting reactivity.[37,58–60] These structures can be quantified through the EC and EC filtrations and used

directly in prediction. The simulations contain a single reactant molecule centered in a 4 nm^3 cube surrounded by water and cosolvent in specified weight percentages. We subsample the 2 ns simulations in sets of 200 picoseconds (i.e. 20 frames) and produce both graph (hydrogen bonding) and manifold (water density) representations. To understand the topology of the hydrogen bonding networks in the simulations, we consider the EC for the water-water hydrogen bonding network $\chi(G_{ww})$, the cosolvent-reactant hydrogen bonding network $\chi(G_{cr})$, and the water-reactant hydrogen bonding network $\chi(G_{wr})$. Hydrogen bonds in each case were computed using the Luzar-Chandler criterion. For each of these networks, we construct a time-averaged EC value for the subsampled simulations $\langle\chi(G_{ww})\rangle, \langle\chi(G_{cr})\rangle, \langle\chi(G_{wr})\rangle$. The manifold $\mathcal{M}$ for this system is now the entire simulation space (versus the surface of the SAMs described previously), which consists of a 3D cube with a continuous function $f : \mathcal{M} \to \mathbb{R}$ representing the time-averaged density of water in each simulation subsample. We ignore the structure of the corresponding cosolvent topology because it is directly related to the water topology (high density water implies low-density cosolvent). We construct our manifold and function representation by placing a $20 \times 20 \times 20$ grid centered on the reactant molecule. We then accumulate and bin water positions within each grid point (each representing a 0.2 nm^3 volume) over the 200 picoseconds of simulation time. The accumulated data is then averaged over the time frame and represents a continuous density function over the simulation space. The EC curves in Figure 9 look slightly different to those for our previous 2D system. This system, now in 3D, has a 3rd Betti number (β_2), which quantifies the number of voids/cavities that appear during filtration. These voids are associated with pockets of high water density (e.g., local maxima) within our 3D manifold and result in a second peak in the EC curve $\chi = \beta_0 - \beta_1 + \beta_2$. Details for the practical computation of the EC for these 3D manifolds can be found in the supporting information and code.

Figure 9 provides another demonstration of the robustness of these topological meth-



(a) Graph and Manifold Representations

(b) $t \in [0, 200)$

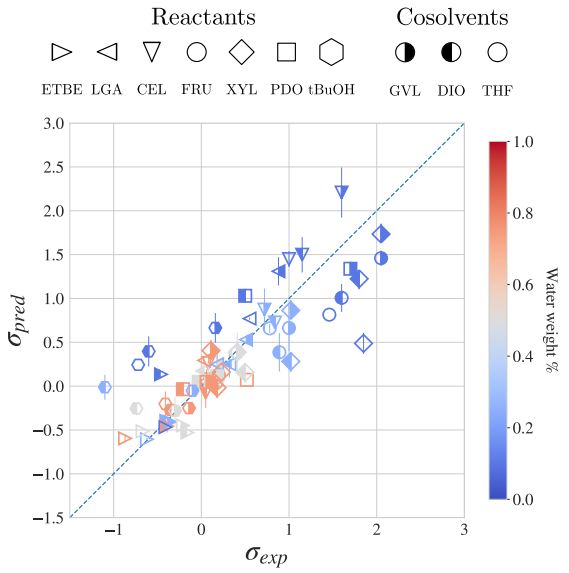
(c) $t \in [200, 400)$

(d) $t \in [400, 600)$

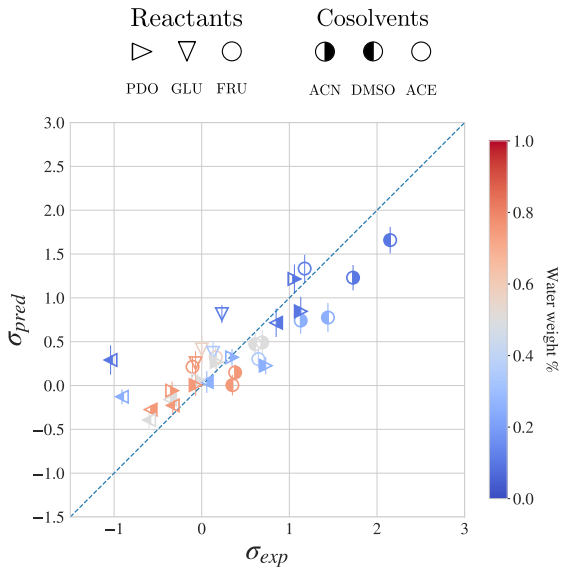
Figure 9: (a) Illustration of the graph and manifold data representations derived from the acid-catalyzed reaction simulations. (b),(c), and (d) Representative water-water hydrogen bond networks (graphs) and time-averaged density fields (manifolds) derived from the water in the molecular simulations over different time frames. They also contain the time-averaged graph EC $\langle \chi(G_{ww}) \rangle$ and the EC curve created from a filtration of the water density field. The density fields and graphs are visually different, but the EC values are similar throughout the simulation. These results demonstrate the robustness of topological descriptors.

ods. We illustrate the changes in the water density and hydrogen bonding network $\langle\chi(G_{ww})\rangle$ during the course of a single MD simulation. Visually, these graphs and manifold functions appear to be distinct, but when quantified and compared through the EC they are almost identical (indicating that they are topologically close). From the ECs of Figure 9, for instance, it appears that the system quickly reaches a topological steady-state and thus the MD simulation can be terminated early to reduce computational time.

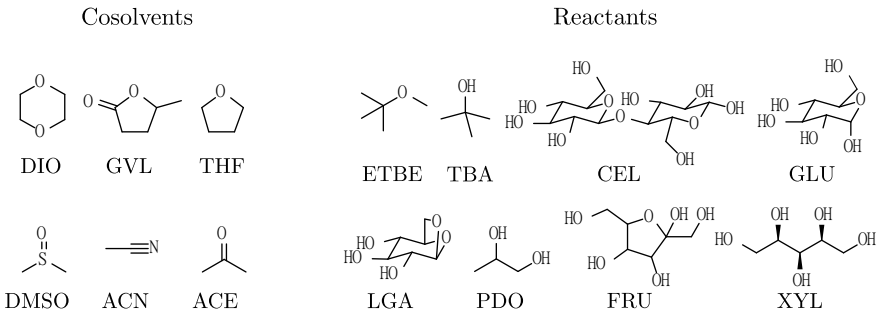
We developed a linear regression model that takes as input the corresponding EC curves and hydrogen bonding EC values $\langle\chi(G_{ww})\rangle, \langle\chi(G_{cr})\rangle, \langle\chi(G_{wr})\rangle$ and outputs the experimentally determined change in reaction rate $\sigma = \log_{10}(k_{org}/k_{H_2O})$ where k_{org} represents the reaction rate in a cosolvent/water mixture and k_{H_2O} represents the reaction rate in pure water.[37] We train the linear regression model on a set of 76 cosolvent and reactant combinations (760 subsampled points) and test our model on a set of 32 different reactant and solvent combinations (320 subsampled points), which is the same data training/testing split used to evaluate the CNN developed in the work of Chew and coworkers.[40] Figure 10 lists the different potential combinations of reactant, cosolvent, and cosolvent/water ratios that are used in both training and testing. Figure 10 also shows the accuracy of the linear model in both training (RMSE = 0.39) and testing (RMSE = 0.42), from which we can conclude that the simple linear model is able to accurately predict the change in reactivity for these chemical systems. We can compare these results directly to the work of Chew and co-workers, where the authors used 3D CNNs that contain up to $\sim$172,417 parameters, compared to the 23 parameters used in our linear model. The topological approach achieves accuracy superior to the trained 3D CNN on the same testing set (RMSE = 0.48). We also note significant improvements in the computational resources needed for training the models. Our linear model takes approximately 2 minutes to train (this time includes computation of ECs), while the 3D CNN can take up to 2 hours. Furthermore, our linear model does not require a search for optimal hyper-parameters or 3D



(a) Training Data Regression.



(b) Testing Data Regression.



(c) Structures of reactants and cosolvents.

Figure 10: (a) Training data parity plot of predicted versus experimental σ (change in reaction rate). The EC curve and averaged graph EC values $\langle\chi(G_{ww})\rangle$, $\langle\chi(G_{cr})\rangle$, $\langle\chi(G_{wr})\rangle$ are used as inputs to a linear model. Predictions on the training dataset are accurate (RMSE = 0.39) and suggest a linear model could be used to obtain high accuracy in the prediction of reactivity trends (σ). (b) Testing data parity plot of predicted versus experimental σ. An unseen test set of acid-catalyzed reaction simulations are created for different cosolvents and solutes. From this data, the corresponding EC curve and graph EC values are computed. The trained linear model is used to predict the experimentally verified reactivity increase for the separate set of acid-catalyzed reaction simulations. The results demonstrate a high level of test set accuracy with low prediction error (RMSE = 0.42). Error bars in both plots represent a single standard deviation from the mean.

CNN architecture, further reducing the needed computational resources. These results highlight the desirable scalability of topological characterizations based on the EC.

As noted previously, the linear regression model trained here has an added benefit of interpretability. Figure 11 contains an analysis of simulations of fructose in varying cosolvent/water mixtures, all at the same cosolvent/water weight ratio (90%/10%). Figure 11 illustrates the differences in water topology that occurs when the chemistry of the cosolvent is altered. We focus on two particular cosolvents: THF and DMSO. For fructose, the change in reaction rate is highest when in a DMSO/water mixture ($\sigma = 1.7$) and lowest when in a THF/water mixture ($\sigma = 0.8$). We note that this corresponds with large changes in the topology of the water density. In THF, water is agglomerated in large clusters near the reactant, which reduces the total number of topological features in the water density function (e.g., connected component, holes, voids) and dampens the magnitude of the peaks and valleys in the EC curve, which is consistent with the findings of Chew and co-workers.[61] The sublevel sets at points in the filtration are also illustrated (Figure 11), further confirming this result. The opposite holds true for water in DMSO, here we see a larger number of high-density and low-density areas on our manifold increasing the peak and valley magnitude of the EC curve. This indicates that DMSO is interrupting the interactions between water molecules and reducing the total amount of water molecules near fructose. This behavior can increase selectivity of the acid-catalyzed reaction of fructose, where the shielding of subsequent products (e.g., 5-hydroxymethylfurfural) inhibits the formation of undesired products (e.g., levulinic acid) as shown in the findings by Mushrif and co-workers.[58]

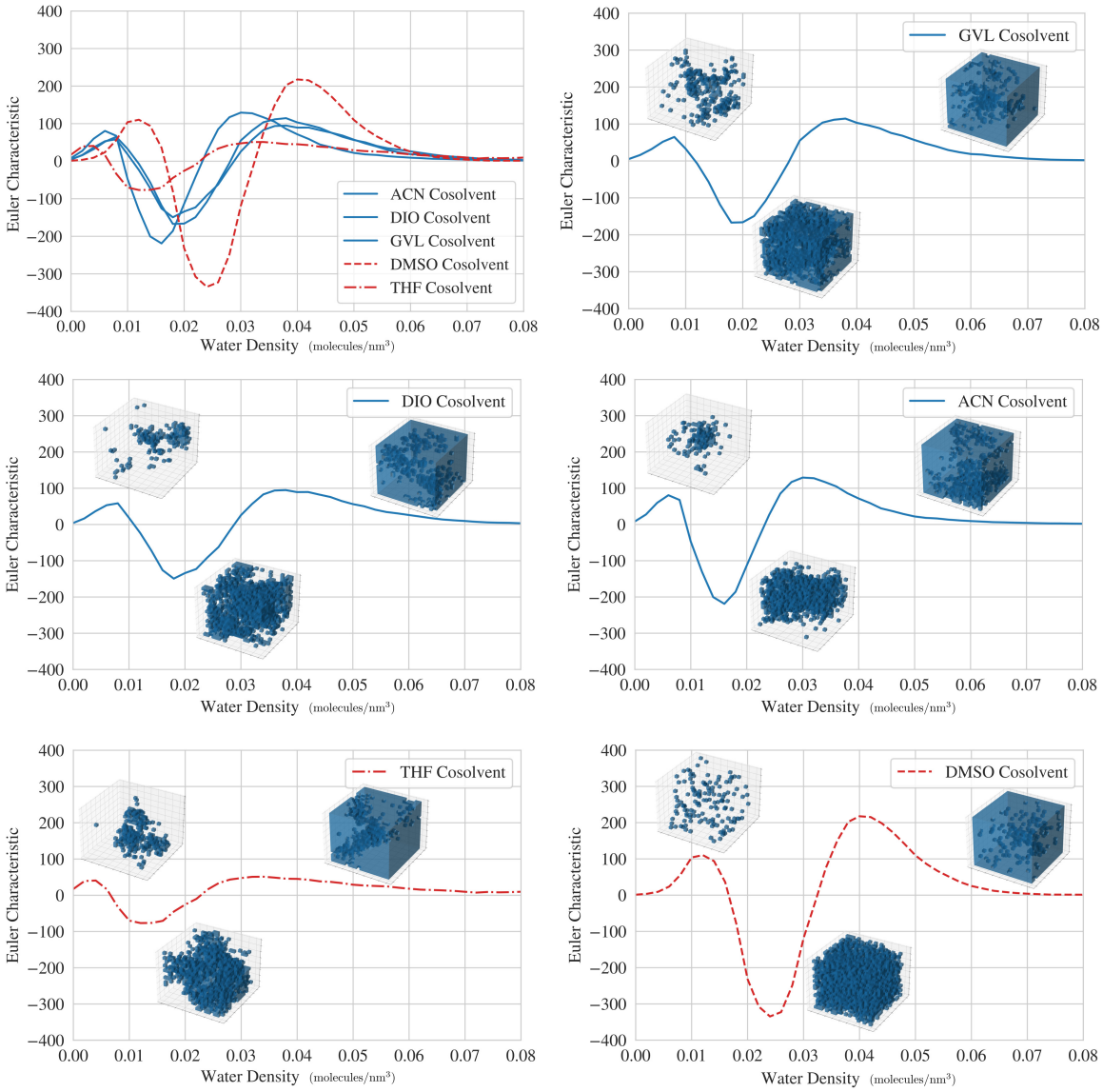


Figure 11: EC curves for fructose solvated in different cosolvent/water mixtures along with the representative submanifolds during various points in the filtration. Each of the EC curves are computed with a 90 wt% cosolvent, 10 wt% water solution. Many of the simulations (ACN, DIO, GVL) behave similarly from a topological perspective, and each have a similar impact on the reactivity of fructose. The EC curves for THF and DMSO differ from the EC curves of the previously mentioned solvents. DMSO interrupts water-water interactions resulting in a larger number of high and low density areas, which manifests in an increased number of connected components, holes, and voids in the density field captured by the EC curve. The opposite occurs with THF where we see fewer, but larger, clusters of high and low density water; this suggests that THF increases the interactions between water molecules and causes larger clusters of water to form.

Discussion and Future Work

Manifold and graph representations of molecular simulation data provide a flexible avenue for capturing both discrete and continuous information sources. In this work, we analyze the topology associated with simulations of self-assembled monolayers and acid-catalyzed reaction systems. We characterize these topological representations with the Euler characteristic curve, which is a simple and robust data descriptor that can be directly integrated into data analysis tasks such as dimensionality reduction or regression. We show that this method results in improvements in computational efficiency, generalizability, and simplicity which can be leveraged to reduce the complexity of models needed for analysis of molecular simulation data (e.g., 3D convolutional neural networks are reduced to linear regression models). These improvements also reduce the need for large labeled datasets needed for ML model training and provide physical intuition. This provides opportunities to improve the information gained from the analysis of high-throughput or large-scale simulation data, which can be used in screening for new materials and chemistry or in optimizing physical and chemical characteristics of existing systems.

The methods developed in this work can extend to other scientific domains because manifolds and graphs can represent many different materials and chemistry simulation datasets, permitting analysis via topology and the Euler characteristic. For example, Density functional theory (DFT) calculations can provide direct graphical representations of molecular bonds.[62] The topology of these molecular graphs provides insight into physical and chemical characteristics.[62–64] Topology and the EC allow us to summarize these complex molecular graphs for tasks such as high-throughput screening.[28,65] For example, topological analysis can improve the high-throughput screening of nanoporous materials used for gas storage and separation, like zeolites[28,66] and metal organic frameworks (MOFs).[67] DFT and molecular dynamics simulations also provide powerful manifold rep-

resentations of data through molecular surfaces.[68] Molecular surfaces are one way to capture complex interactions between molecules or molecules and their environment.[68–70] These molecular surfaces are manifolds with associated functions. Thus, topological methods can quantify and summarize molecular surface data. Topology also provides complementary information that is not captured in statistical summaries (e.g., sigma profiles).[71] We envision that such topological analysis can thus a wide range of systems and processes involving analysis of molecular surfaces, including in studies of protein folding and docking, nanoparticles, membranes, pharmaceuticals, and catalysts.[34,72–75]

In future work, we will address some of the existing challenges faced by the Euler characteristic in its application to the study of physical and chemical systems. For example, we aim to form deeper relationships between the Euler characteristic and the thermodynamics of chemical systems through connections between topology and geometry that are found in stochastic geometry.[52,76,77] We will also further explore statistical frameworks such as hypothesis testing in the analysis of the Euler characteristic, similar to those found in the study of random fields.[31,78] This information will aid in providing comparisons between the Euler characteristic and other topological measures such as those found in persistence homology, and when either method can be efficiently applied. These advancements will continue to establish the Euler characteristic as a physically meaningful measure for the study of other chemical systems such as colloidal polymers, gels, surfactants, multi-phase systems, and proteins that demonstrate complex topology and geometry.[79–83]

Acknowledgement

The authors acknowledge funding from the U.S. National Science Foundation (NSF) under BIGDATA grant IIS-1837812. A. K. C, A. S. K., and R. C. V. further acknowledge

support from NSF under grant number DMR-2044997.

Supporting Information Available

Implementation details for both case studies can be found in the Supplementary Information. All code and data needed to reproduce the results can be found in https://github.com/zavalab/ML/tree/master/MD_Euler.

References

(1) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; Van Der Spoel, D., et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(2) Doerr, S.; Harvey, M.; Noé, F.; De Fabritiis, G. HTMD: high-throughput molecular dynamics for molecular discovery. *Journal of chemical theory and computation* **2016**, *12*, 1845–1852.

(3) Harvey, M. J.; De Fabritiis, G. High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug discovery today* **2012**, *17*, 1059–1062.

(4) Bobbitt, N. S.; Snurr, R. Q. Molecular modelling and machine learning for high-throughput screening of metal-organic frameworks for hydrogen storage. *Molecular Simulation* **2019**, *45*, 1069–1081.

(5) Peter, C.; Kremer, K. Multiscale simulation of soft matter systems. *Faraday discussions* **2010**, *144*, 9–24.

(6) Attig, N.; Binder, K.; Grubmüller, H.; Kremer, K. Computational soft matter: from

synthetic polymers to proteins. *John von Neumann Institute for Computing (NIC), Juelich* **2004**,

(7) Scheraga, H. A.; Khalili, M.; Liwo, A. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.

(8) Orsi, M. *Self-assembling Biomaterials*; Elsevier, 2018; pp 305–318.

(9) Je, L.; Huber, G. W.; Van Lehn, R. C.; Zavala, V. M. On the integration of molecular dynamics, data science, and experiments for studying solvent effects on catalysis. *Current Opinion in Chemical Engineering* **2022**, *36*, 100796.

(10) 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.

(11) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *The Journal of chemical physics* **2008**, *129*, 114707.

(12) van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. Molecular simulation as an aid to experimentalists. *Current opinion in structural biology* **2008**, *18*, 149–153.

(13) Anwar, J.; Zahn, D. Uncovering molecular processes in crystal nucleation and growth by using molecular simulation. *Angewandte Chemie International Edition* **2011**, *50*, 1996–2013.

(14) Allen, M. P. Molecular simulation and theory of the isotropic–nematic interface. *The Journal of Chemical Physics* **2000**, *112*, 5447–5453.

(15) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annual review of physical chemistry* **2020**, *71*, 361–390.

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

(17) 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.

(18) 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.

(19) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems* **2017**, *30*.

(20) Smith, A. D.; Dłotko, P.; Zavala, V. M. Topological data analysis: concepts, computation, and applications in chemical engineering. *Computers & Chemical Engineering* **2021**, *146*, 107202.

(21) Smith, A.; Zavala, V. M. The Euler characteristic: A general topological descriptor for complex data. *Computers & Chemical Engineering* **2021**, *154*, 107463.

(22) Horner, K. E.; Miller, M. A.; Steed, J. W.; Sutcliffe, P. M. Knot theory in modern chemistry. *Chemical Society Reviews* **2016**, *45*, 6432–6448.

(23) Sumners, D. The knot theory of molecules. *Journal of mathematical chemistry* **1987**, *1*, 1–14.

(24) Liang, C.; Mislow, K. Knots in proteins. *Journal of the American Chemical Society* **1994**, *116*, 11189–11190.

(25) Sørensen, S. S.; Biscio, C. A.; Bauchy, M.; Fajstrup, L.; Smedskjaer, M. M. Revealing

hidden medium-range order in amorphous materials using topological data analysis. *Science Advances* **2020**, *6*, eabc2320.

(26) Steinberg, L.; Russo, J.; Frey, J. A new topological descriptor for water network structure. *Journal of cheminformatics* **2019**, *11*, 1–11.

(27) Tang, W. S.; da Silva, G. M.; Kirveslahti, H.; Skeens, E.; Feng, B.; Sudijono, T.; Yang, K. K.; Mukherjee, S.; Rubenstein, B.; Crawford, L. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS computational biology* **2022**, *18*, e1010045.

(28) 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. *Journal of chemical theory and computation* **2018**, *14*, 4427–4437.

(29) Pirashvili, M.; Steinberg, L.; Belchi Guillamon, F.; Niranjan, M.; Frey, J. G.; Brodzki, J. Improved understanding of aqueous solubility modeling through topological data analysis. *Journal of cheminformatics* **2018**, *10*, 1–14.

(30) Topaz, C. M.; Ziegelmeier, L.; Halverson, T. Topological data analysis of biological aggregation models. *PloS one* **2015**, *10*, e0126383.

(31) Adler, R. J. Some new random field tools for spatial analysis. *Stochastic Environmental Research and Risk Assessment* **2008**, *22*, 809–822.

(32) Bernstein, J.; Davis, R. E.; Shimoni, L.; Chang, N.-L. Patterns in hydrogen bonding: functionality and graph set analysis in crystals. *Angewandte Chemie International Edition in English* **1995**, *34*, 1555–1573.

(33) Jeffrey, G. A. Hydrogen-bonding: an update. *Crystallography Reviews* **2003**, *9*, 135–176.

(34) 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.

(35) Chung, M. K.; Smith, A.; Shiu, G. Reviews: Topological Distances and Losses for Brain Networks. *arXiv preprint arXiv:2102.08623* **2021**,

(36) 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.

(37) 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.

(38) McDonald, S.; Ojamäe, L.; Singer, S. J. Graph theoretical generation and analysis of hydrogen-bonded structures with applications to the neutral and protonated water cube and dodecahedral clusters. *The Journal of Physical Chemistry A* **1998**, *102*, 2824–2832.

(39) Radhakrishnan, T.; Herndon, W. C. Graph theoretical analysis of water clusters. *The Journal of Physical Chemistry* **1991**, *95*, 10609–10617.

(40) 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.

(41) Gun'ko, V.; Turov, V.; Bogatyrev, V.; Zarko, V.; Leboda, R.; Goncharuk, E.; Novza, A.; Turov, A.; Chuiko, A. Unusual properties of water at hydrophilic/hydrophobic interfaces. *Advances in Colloid and Interface Science* **2005**, *118*, 125–172.

(42) Kusalik, P. G.; Svishchev, I. M. The spatial structure in liquid water. *Science* **1994**, *265*, 1219–1221.

(43) Rasaiah, J. C.; Garde, S.; Hummer, G. Water in nonpolar confinement: From nanotubes to proteins and beyond. *Annu. Rev. Phys. Chem.* **2008**, *59*, 713–740.

(44) Rego, N. B.; Patel, A. J. Understanding hydrophobic effects: Insights from water density fluctuations. *Annual Review of Condensed Matter Physics* **2022**, *13*, 303–324.

(45) 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.

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

(47) Shenogina, N.; Keblinski, P.; Garde, S. Strong frequency dependence of dynamical coupling between protein and water. *The Journal of chemical physics* **2008**, *129*, 10B614.

(48) Zomorodian, A. Topological data analysis. *Advances in applied and computational topology* **2012**, *70*, 1–39.

(49) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* **2011**, *2*, 1–27.

(50) 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.

(51) Mecke, K. R.; Sofonea, V. Morphology of spinodal decomposition. *Physical Review E* **1997**, *56*, R3761.

(52) Mecke, K. R. *Statistical Physics and Spatial Statistics*; Springer, 2000; pp 111–184.

(53) Valiokas, R.; Östblom, M.; Svedhem, S.; Svensson, S. C.; Liedberg, B. Thermal stability of self-assembled monolayers: Influence of lateral hydrogen bonding. *The Journal of Physical Chemistry B* **2002**, *106*, 10401–10409.

(54) Lewis, P. A.; Smith, R. K.; Kelly, K. F.; Bumm, L. A.; Reed, S. M.; Clegg, R. S.; Gunderson, J. D.; Hutchison, J. E.; Weiss, P. S. The role of buried hydrogen bonds in self-assembled mixed composition thiols on Au {111}. *The Journal of Physical Chemistry B* **2001**, *105*, 10630–10636.

(55) Clegg, R. S.; Reed, S. M.; Hutchison, J. E. Self-assembled monolayers stabilized by three-dimensional networks of hydrogen bonds. *Journal of the American Chemical Society* **1998**, *120*, 2486–2487.

(56) Box, G. E. Science and statistics. *Journal of the American Statistical Association* **1976**, *71*, 791–799.

(57) Noble, W. S. What is a support vector machine? *Nature biotechnology* **2006**, *24*, 1565–1567.

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

(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) Li, G.; Wang, B.; Resasco, D. E. Water-mediated heterogeneously catalyzed reactions. *Acs Catalysis* **2019**, *10*, 1294–1309.

(61) 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.

(62) Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Physical review letters* **1985**, *55*, 2471.

(63) Corminboeuf, C.; Tran, F.; Weber, J. The role of density functional theory in chemistry: Some historical landmarks and applications to zeolites. *Journal of Molecular Structure: THEOCHEM* **2006**, *762*, 1–7.

(64) Nicholas, J. B. Density functional theory studies of zeolite structure, acidity, and reactivity. *Topics in Catalysis* **1997**, *4*, 157–171.

(65) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-data science in porous materials: materials genomics and machine learning. *Chemical reviews* **2020**, *120*, 8066–8129.

(66) Krishnapriyan, A. S.; Haranczyk, M.; Morozov, D. Topological descriptors help predict guest adsorption in nanoporous materials. *The Journal of Physical Chemistry C* **2020**, *124*, 9360–9368.

(67) Krishnapriyan, A. S.; Montoya, J.; Haranczyk, M.; Hummelshøj, J.; Morozov, D. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Scientific reports* **2021**, *11*, 1–11.

(68) Connolly, M. L. Analytical molecular surface calculation. *Journal of applied crystallography* **1983**, *16*, 548–558.

(69) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface

charge to calculate reaction field energies: Applications to the molecular systems and geometric objects. *Journal of computational chemistry* **2002**, *23*, 128–137.

(70) Jiang, F.; Kim, S.-H. "Soft docking": matching of molecular surface cubes. *Journal of molecular biology* **1991**, *219*, 79–102.

(71) Mullins, E.; Oldland, R.; Liu, Y.; Wang, S.; Sandler, S. I.; Chen, C.-C.; Zwolak, M.; Seavey, K. C. Sigma-profile database for using COSMO-based thermodynamic methods. *Industrial & engineering chemistry research* **2006**, *45*, 4389–4415.

(72) Palm, K.; Luthman, K.; Unge, A.-L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *Journal of pharmaceutical sciences* **1996**, *85*, 32–39.

(73) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences* **1992**, *89*, 2195–2199.

(74) 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.

(75) Stansfeld, P. J.; Sansom, M. S. Molecular simulation approaches to membrane proteins. *Structure* **2011**, *19*, 1562–1572.

(76) Mecke, K. R. Integral geometry in statistical physics. *International Journal of Modern Physics B* **1998**, *12*, 861–899.

(77) Mecke, K. R. A morphological model for complex fluids. *Journal of Physics: Condensed Matter* **1996**, *8*, 9663.

(78) Adler, R. J.; Taylor, J. E., et al. *Random fields and geometry*; Springer, 2007; Vol. 80.

(79) Dhakal, S.; Sureshkumar, R. Topology, length scales, and energetics of surfactant micelles. *The Journal of Chemical Physics* **2015**, *143*, 024905.

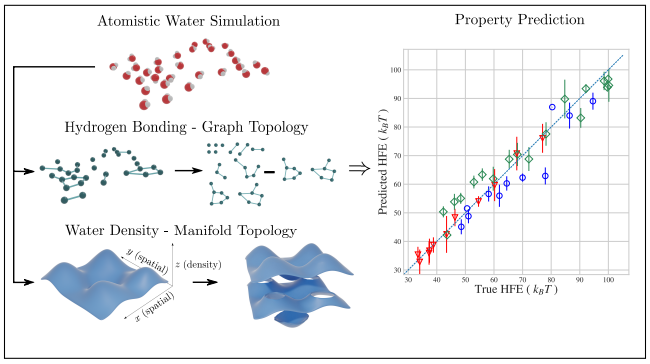
(80) Qin, S.; Jin, T.; Van Lehn, R. C.; Zavala, V. M. Predicting critical micelle concentrations for surfactants using graph convolutional neural networks. *The Journal of Physical Chemistry B* **2021**, *125*, 10610–10620.

(81) Colombo, J.; Del Gado, E. Stress localization, stiffening, and yielding in a model colloidal gel. *Journal of rheology* **2014**, *58*, 1089–1116.

(82) Statt, A.; Kleeblatt, D. C.; Reinhart, W. F. Unsupervised learning of sequence-specific aggregation behavior for a model copolymer. *Soft matter* **2021**, *17*, 7697–7707.

(83) Amézquita, E. J.; Quigley, M. Y.; Ophelders, T.; Munch, E.; Chitwood, D. H. The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics* **2020**, *249*, 816–833.

TOC Graphic


Atomistic Water Simulation
Property Prediction
Hydrogen Bonding - Graph Topology
Water Density - Manifold Topology
y (spatial)
z (density)
x (spatial)
Predicted HFE (k_BT)
True HFE (k_BT)