Revitalizing Simulations of Colloidal and Interfacial Systems: From Atomistic Trajectories to Data-Centric Insights

Hasan Imani Parashkooh

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MECHANICAL ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

April 2025

Abstract

This dissertation advances methodological frameworks for investigating complex systems characterized by competing interactions. Structured around three interrelated themes, the research develops methods to tackles challenges in water-in-oil emulsions, chemical inhibition in asphaltene aggregation, and adsorption energy prediction. While each theme addresses unique challenges, they all contribute to quantitative understanding of interactions across scales, integrating structural, thermodynamic, and energetic perspectives to establish governing principles for system behavior.

The first theme elucidates the interplay between asphaltene aggregation and water-in-oil droplet coalescence. Molecular Dynamics (MD) simulations conducted in pentane solvents reveal the mutual influences of asphaltene and water droplets, highlighting a nonmonotonic trend in polyaromatic stacking with increasing droplet sizes. An innovative in-house tool is introduced to analyze droplet coalescence modes, demonstrating that droplet growth predominantly occurs around a nucleation site, which is the largest droplet.

In the second theme, a novel approach for calculating partial molar volumes (PMVs) directly from MD simulation trajectories is presented. This approach has been validated against experimental data, yielding an average error of 4.41%. When applied to systems containing model asphaltenes, organic solvents, and chemical inhibitors, the PMV analysis elucidates molecular-level inhibition mechanisms. Specifically, it shows that inhibitors enhance solubility by altering nanoaggregate size and number, with their polar and nonpolar segments interacting with different regions of the asphaltene molecules and the solvent.

The third and final theme highlights the critical need for efficient and accurate adsorption energy predictions, a fundamental aspect of catalysis, materials design, and a key factor in mitigating asphaltene deposition. Traditional methods of calculating adsorption energy such as Density Functional Theory (DFT) are computationally demanding, particularly for large, complex adsorbates like aromatics. To overcome these limitations, this work applies pretrained graph neural networks (GNNs) with a directional message passing architecture. The model is trained to capture geometric relationships between small adsorbates and metallic or metal oxide surfaces, linking

them to energy and atomic forces. Here, it is fine-tuned to adapt these learned interactions for larger molecules, including aromatics. Two curated datasets, a diverse adsorbate-substrate collection and a specialized aromatic subset, are employed to balance generalizability with specificity. Results indicate that the sheer volume of fine-tuning data has a greater impact on model adaptation than using smaller but more domain-specific datasets. Moreover, the results show that selectively fine-tuning the early layers, which focus on geometric features, achieves performance comparable to full model retraining. This highlights the crucial role of geometric feature adaptation.

Collectively, these themes contribute to the refinement of data-driven methodologies for complex interfacial and amorphous systems, providing actionable insights into structural, thermodynamic, and energetic properties that were previously inaccessible. Nano-scale simulations with extensive temporal and spatial sampling are essential, as competing interactions make it impossible to predict dominant factors without comprehensive analysis. By reutilizing existing simulation data, the innovations presented here enhance sustainability and efficiency, with broad implications for physical chemistry, materials science, and industrial applications.

Acknowledgment

As I'm reaching the end of my PhD journey, I am filled not only with pride in the academic milestones achieved but with profound gratitude for those who shaped this experience into one of growth, discovery, and resilience.

First, I want to thank **my advisor, Dr. Cuiying Jian**, whose mentorship has been a guiding force throughout this academic journey. From the outset of the program, she encouraged me to think boldly and refine not only my voice as a researcher but also the clarity and impact with which I present my work. Her emphasize on crafting compelling narratives and engaging diverse audiences transformed my approach to sharing ideas, a skill as vital as the research itself. Moreover, her faith in my potential turned uncertainty into confidence and taught me the power of communicating science with purpose.

Gratitude is extended to **my committee members, Dr. Alidad Amirfazli and Dr. Song Wang**, as they have played a crucial role in shaping this work. Their discerning feedback has sharpened the quality of my work and broadened my perspective. Their willingness to engage with my research helped me navigate complexities with clarity and purpose.

The unwavering support of **my family** has been an anchor through every challenge and triumph. In moments of self-doubt, their unwavering belief in me rekindled my determination. I carry their lessons of courage and kindness far beyond this chapter of my life.

Last but not least, I want to thank **my partner, Farnoosh,** whose unwavering support and patience proved indispensable throughout the challenges of graduate study. Through countless nights of coding, writing, and troubleshooting, she stood steadfastly by my side, offering quiet resilience and encouragement that transformed solitary challenges into shared endeavors. I will forever carry the memory of her contributions as a testament to the power of committed partnership.

This journey has taught me that a PhD is more than a degree. It is a testament to the communities that sustain us. As I step forward, I do so not just with new knowledge, but with a deepened appreciation for collaboration, perseverance, and the extraordinary power of human support. To all who shared this path: you have my deepest, most enduring appreciation.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Janus Systems

Physical systems often display complex, non-linear behaviors due to being impacted by different interactions from their components. These interactions, which may simultaneously act in opposing directions, create significant challenges in predicting system outcomes without direct experimentation. They can be categorized as conflicting "actors", defined as factors or entities that exert opposing effects on a given outcome. The interplay between these actors often results in emergent behaviors that are highly sensitive to control variables, such as temperature, concentration, or composition. An example of such systems is polymer-water systems, specifically polymer hydrogels, which have widespread biomedical applications in drug delivery [1,2], biosensing [3,4], and bone regeneration [5,6]. In polymer hydrogels, the gel strength or stiffness (outcomes of interest) are influenced by both hydrophilic interactions between polymer and water and crosslinking interactions between polymer chains. These two interactions exert opposite effects on the gel strength and are sensitive to polymer concentration (control variables). A tipping point or threshold of polymer concentrations exists where the dominance of these two interactions changes, leading to opposite trends in the gel strength with increasing polymer concentrations.

We propose the term Janus systems to describe such systems. In Roman mythology, Janus is the deity of transitions and dualities, symbolizing beginnings and endings, or dual perspectives. By analogy, Janus systems encapsulate the concept of two (or more) principal interactions (actors) pulling the system in opposite directions. This metaphor also aligns with the idea of tipping points or thresholds, reflecting the sudden, sometimes dramatic, shifts in system behavior when one actor overtakes the other. To systemically study and represent Janus systems, this dissertation defines four key quantities: **A**, **B**, **P**, and **R**. Here, **A** and **B** are opposing actors influencing a measurable outcome **P**. They are called opposing because actor **A** seeks to increase **P**, while **B** works to decrease it. The equilibrium value of **P** is modulated by a set of control parameters **R**, which govern how **A** and **B** influences **P**. In the polymer-water systems mentioned above, **P** could represent the gel strength or stiffness, actor **A** could denote hydrophilic interactions, and **B** could correspond to

crosslinking interactions between polymer chains. At low polymer concentrations (a component of **R**), hydrophilic interactions dominate, suppressing gel formation. As the concentration increases, crosslinking interactions take over, drastically altering the gel's properties.

Another example of a Janus system is the formation and stability of emulsions, such as oil-water mixtures stabilized by surfactants. These emulsions have diverse applications, including their use in enhanced oil recovery [7]. Here, **P** is the stability of emulsion droplets (e.g., their resistance to coalescence). Actor **A** is surfactant adsorption at the oil-water interface, which reduces surface tension and stabilizes droplets (increasing **P**), while **B** is van der Waals attractions between oil droplets, which promote coalescence and destabilize the emulsion (decreasing **P**). Furthermore, **R** is the surfactant concentration and temperature, which modulate the strength of **A** and **B**. At low surfactant concentrations (low **R**), **B** dominates, leading to rapid coalescence and poor emulsion stability. As surfactant concentration increases (high **R**), **A** becomes dominant, stabilizing the emulsion by forming a protective layer around the droplets. However, at very high surfactant concentrations, micelle formation in the bulk phase can compete with interfacial adsorption, reducing the effectiveness of **A** and potentially destabilizing the emulsion again.

As demonstrated in the above example, non-monotonic trends are a common feature of Janus systems, arising from the alternating dominance of opposing interactions (actors) as control variables are varied. Another observation is that **P** can be very sensitive to even minor changes in **R**, especially in regions where small adjustments in **R** lead to major shifts in the interplay between **A** and **B**, resulting in significant changes in **P**. These intricate interdependencies and divergent behaviors characteristic of such systems impose significant challenges on unravelling their complexity.

The classical approach to studying Janus systems begins with designing experiments to systematically explore how control parameters (e.g., **R**, the polymer concentration in the first example) influence the system's behavior (e.g., **P**, gel strength or stiffness). These experiments aim to observe how variations in **R**, shift the balance of influence between competing actors **A** (hydrophilic interactions) and **B** (crosslinking interactions). By identifying regions of **R** where each actor dominates, empirical data can be gathered to understand the system's dynamics. Subsequently, this data is used in conjunction with theoretical frameworks to explain why and how

A and **B** affect **P**. These insights can further guide the development of mathematical models to predict **P** in unexplored regions of **R**.

The classical experimental approach remains essential for investigating complex systems such as asphaltene aggregation [8] and hydrogel swelling [9], especially when real-world conditions must be accurately replicated. Experiments enable direct measurement of macroscopic phenomena, including aggregation and coalescence rates [8,10], viscosity variations [11], and swelling behavior [9], under diverse environmental conditions such as temperature, pressure, and chemical heterogeneity. However, experimental studies are inherently resource-intensive, demanding substantial time, financial investment, and specialized equipment to systematically probe even a limited portion of the related **R** space.

A major limitation of experimental methods lies in the difficulty to reveal the nano-scale interactions responsible for observed macroscopic phenomena. For instance, while experiments can measure asphaltene precipitation rates in the presence of chemical inhibitors [12], they cannot directly trace the structural interactions, such as how inhibitor molecules disrupt π–π stacking among asphaltene molecules. The disconnect between macroscopic outcomes and molecular mechanisms complicates the interpretation of results and often necessitates complementary approaches. Additionally, experiments are limited to specific parameter ranges, so researchers must rely on extrapolation methods that may not be accurate given the sensitivity of Janus systems to changes in **R**. This makes it challenging to apply the findings to untested regions without further, time-consuming testing.

Computational simulations offer a powerful alternative, providing molecular-level insights and enabling efficient exploration of broader parameter spaces. For example, in the case of chemicals used to inhibit asphaltene aggregation, simulations can model inhibitor's interactions with asphaltenes in unprecedented detail, revealing specific molecular alignments, hydrogen bonding patterns, or steric effects that hinder aggregation [13]. The high-dimensional outputs of simulations, including atomic-level positional data, can then be analyzed using statistical tools to identify dominant interaction mechanisms and extract high-level conclusions. For instance, simulations may show how a small number of inhibitor molecules align with asphaltene surfaces to disrupt aggregation [14], offering insights that are difficult to obtain experimentally.

Additionally, simulations can predict behavior under extreme conditions, such as ultra-high temperatures, which pose significant challenges for experimental studies due to safety and equipment limitations.

Simulations are not always preferable to experiments. They can be computationally demanding, especially for large systems or long timescales. To make them tractable, researchers often use approximations, such as coarse-grained models in MD [15] or RANS models in CFD, to simplify complex systems. In some cases, simulations are even more limited than experiments in exploring certain regions of the control space. For instance, simulating asphaltene aggregation at low inhibitor concentrations necessitates a large number of solute and asphaltene molecules to maintain statistical reliability, significantly increasing system size and computational cost. In contrast, experimental techniques can directly measure inhibition efficiency at low concentrations, providing valuable insights into system behavior within crude oil reservoirs [12]. Additionally, simulations struggle to fully capture impurities, multi-scale interactions, and the complexity of real-world environments, while these factors are naturally incorporated in experimental setups. Despite these limitations, the ability of simulations to resolve molecular interactions with high precision and predict behaviors beyond experimental reach makes them an indispensable complement to experimental research.

1.2. Molecular Dynamics and the Role of Data Analysis

Simulations can replicate experimental conditions and provide a wealth of data on the relationships between A, B, P, and R. For instance, molecular dynamics (MD) simulations have been widely used to study polymer-water and emulsion systems [16–20]. One major advantage of MD simulations is their ability to generate extensive, atomistic datasets, which embed all dynamical details of the simulated system. Examining these details could help reveal the time-dependent effects of **A** and **B** on **P** (system behaviors) at the atomistic levels that are inaccessible using experimental techniques. They will allow us to capture the dependence of A and B on molecular structures and the associated effects on P (the system outcomes).

In practice, extracting specific insights from MD data can be challenging because the information is embedded in high-dimensional datasets. For example, the systems described in Chapter 2 of this thesis are simulated over 100 nanoseconds to ensure that equilibrium is reached.

During these simulations, the three-dimensional coordinates of every atom are recorded every 2 picoseconds, resulting in 50,000 frames. In the case of united-atom modeling for a medium-sized system—comprising in excess of 25,000 sites—the simulation produces on the order of 4×10^9 coordinate data points, corresponding to a storage requirement slightly greater than 15 GB when using single-precision (float) representations.

Recently, Machine Learning (ML) methods have been developed to help manage these challenges of analyzing MD data. For instance, there have been major advancements in learning slow modes associated with the Variational Approach to Conformation Dynamics (VAC) [21], Variational Approach for Markov Processes (VAMP), and Time-Lagged Autoencoders (TAE) [22]. These methods have enabled the analysis of MD trajectories to identify long-lived states and transition pathways. These methods, often implemented using Artificial Neural Networks (ANNs), are used to project high-dimensional data from MD simulations into a lower-dimensional space, revealing long-lived states and transition pathway. As one example, VAMP principle was used in conjunction with ANNs to create VAMPnets [23], a framework for automated construction of Markov State Models (MSMs). This framework was specifically applied to study the folding dynamics of the NTL9 protein.

Similarly, enhanced sampling methods like using autoencoders [24,25] help overcome time-scale limitations by generating statistically accurate data for slow biological processes. These methods are essential for studying events that occur beyond the reach of typical molecular dynamics (MD) simulations. Such processes include ligand dissociation [26], where a ligand (a molecule that binds to a protein) detaches from its binding site, and slow conformational exchanges, where proteins transition between different stable states. Investigating these mechanisms is essential for a deeper understanding of both drug interactions and protein functionality.

While the methods discussed above have advanced our ability to extract slow modes from MD simulations, they primarily focus on generating abstract, low-dimensional embeddings through techniques such as autoencoders. In contrast, the scope of this work is not limited to such descriptive embeddings; rather, it aims to address existing gaps in ML advancements by incorporating innovative approaches that combine both abstract and heuristic embeddings. Here,

"heuristic embeddings" refer to features that are directly constructed based on domain-specific knowledge (such as coalescence modes, partial molar volumes, or adsorption energies) that carry clear physical interpretations.

Furthermore, an aspect that remains relatively less unexploited is how to fully exert the atomistic details of simulations beyond their equilibrium properties. For instance, standard MD simulations, where compositions remain unchanged throughout the simulations, are often employed to study atom/molecule distributions. They are often treated as unsuitable for studying thermodynamic properties such as partial molar volume. The underlying reason is partially due to that the high-dimensional datasets require automated tools to disentangle interplays among different dimensions. As a result, thermodynamic properties require perturbations to the system to allow the calculation of system responses and further the investigation of thermodynamic properties. However, simulating perturbations using MD is often time-consuming and requires extensive resources. As one example, free energy perturbation molecular dynamics (FEP/MD) simulations can be used to obtain binding free energy of proteins. This is done by calculating free energy differences when slowly changing a system from one state to another, which requires multiple simulations across a series of intermediate states. Each of these simulations is computationally intensive and adds to the overall cost [27]. In this regard, automated tools to exploit standard MD simulations could save tremendous amount of simulation time and resources.

Moreover, recent advancements in adsorption energy predictions have been propelled by datasets and models developed under the Open Catalyst challenges [28–30]. These initiatives have delivered state-of-the-art predictive frameworks that excel in evaluating adsorption energies for small adsorbates—typically molecules containing up to around 20 atoms [31,32]. However, a significant gap persists: the current models are not readily applicable to larger molecular systems. As the complexity and size of adsorbates increase beyond this threshold, predictive accuracy and transferability diminish.

The above limitations have driven the focus of this dissertation toward developing methods that address specific challenges in understanding Janus systems and predicting diverse **P** outcomes, including structural, thermodynamic, and energetic properties. These methods were designed to extract new insights from legacy or existing simulation data that are inaccessible by heuristic

approaches. They enabled the testing of novel theories without requiring sophisticated experiments or simulations. By applying these approaches, the utility of existing data was maximized with reduced resource consumption. These approaches were validated using colloidal and interfacial systems containing amorphous materials, e.g. surface-active compounds in oil productions. These systems were chosen for addressing real-world challenges, such as mitigating pipe clogging and improving the efficiency of oil-water separation in the oil industry as described in Section 1.3. Three distinct aspects of these materials, structural, thermodynamic, and energy properties, were studied to provide a comprehensive understanding of Janus systems.

1.3. Asphaltenes in the Case Studies

Asphaltenes, the heaviest and most surface-reactive fraction of non-volatile petroleum, pose significant challenges in oil pipelines and processing equipment due to their strong tendency to aggregate and adhere to surfaces. Their strong propensity to aggregate and adhere to surfaces can lead to deposits that clog pipelines and disrupt production [33]. Moreover, their surface activity promotes the formation of layers at the oil-water interface, which stabilizes emulsions and complicates separation processes.

In addition to the importance of asphaltene studies for industrial processing, the inherent properties of asphaltenes render them as ideal candidates for investigating size-dependent behaviors in Janus systems. Asphaltenes, by themselves, characterized by multiple structural features that often introduce competing interaction modes, such as π-π stacking, polar headgroup interactions, and tail-tail interactions. These interactions can lead to non-monotonic or multi-stage bulk and interfacial behaviors, making their interpretation more complex than that of simpler molecules with a single dominant interaction mechanism. For example, asphaltenes can adsorb onto water droplet surfaces to facilitate or prevent droplet coalescence depending on control variables (conditions). By contrast, smaller molecules with a single dominating feature may manifest non-significant effects on the droplet coalescence process.

Given the complexity of asphaltenes, the selection of asphaltenes for study underscores the importance of method compatibility with diverse structural characteristics. This compatibility simplifies the adaptation of the methods to similar complex systems that often present analytical challenges. Thus, it helps extend the application of the developed methods beyond oil industries.

1.4. Objectives of the Dissertation

The overarching goal of this dissertation is to develop and refine data-driven methods for studying Janus systems in the context of amorphous and interfacial physics. A key emphasis is placed on leveraging the results of existing, standard Molecular Dynamics (MD) simulation and Density Functional Theory (DFT) calculation. By doing so, it eliminates the need for additional computationally demanding simulations while retaining insights into nanoscale interactions and dynamics. It is worth emphasizing that *Janus systems* is used here as an umbrella term to describe systems governed by two or more competing interactions. However, the focus of this dissertation is specifically on Janus systems within interfacial and amorphous physical systems. The dissertation is structured around three main themes, each addressing distinct challenges in physical chemistry and materials science. The contributions made in these themes are as follows:

1. Investigating competing interactions in water-in-oil systems and characterizing droplet coalescence dynamics

2. Developing a novel method for calculating partial molar volume (PMV) to examine the impact of chemical inhibitors on asphaltene aggregation

3. Exploring fine-tuning methods in pretrained graph neural networks (GNNs) for predicting adsorption energy of aromatic compounds on metallic surfaces

Through these three themes, the dissertation aims to bridge experimental observations, molecular-level understanding, and predictive modeling in systems with competing interactions. By taking structural, thermodynamic, and energetic perspectives, it establishes principles that dictate emergent phenomena in Janus systems. Below is an overview of each theme and its alignment with Janus system framework:

Theme 1: Investigations on the Coalescence of Water-in-Oil Droplets

Amorphous systems, such as water-in-oil emulsions, are central to many industrial processes, particularly in the oil and energy sectors [34,35]. These systems exhibit complex dynamics driven by competing interactions. For instance, in water-in-oil systems stabilized by asphaltenes, the degree of asphaltene aggregation and the coalescence behavior of water droplets are strongly

influenced by the delicate balance between adsorption at interfaces and aggregation among asphaltene molecules. In this theme, the control parameter (**R**) is the concentration of water. It explores:

- The aggregation of asphaltenes, quantified as the average size of clusters (**P**).
    - Actors **A** and **B** are the interactions between asphaltene and water surface (adsorption), and asphaltene's polyaromatic cores with each other (aggregation), respectively.
- The coalescence rate (**P'**) and coalescence mechanisms of water droplets
    - Here, actor **A'** is again adsorption, which causes **P'** (coalescence rate) to go down, and **B'** is the intermolecular attractions between water molecules causing the **P'** to go up.

Theme 2: A novel method for calculating partial molar volume from a standard molecular trajectory

The aggregation of asphaltenes in oil is a significant challenge in the oil industry, often leading to precipitation and blockages. Chemical inhibitors are widely used to address this issue [16,36], but their molecular-level mechanisms remain inadequately understood. A novel approach is proposed to calculate the Partial Molar Volume (PMV) of system components directly from MD simulation data, using techniques such as linear regression and small-volume sampling. Recalling the Janus systems framework, this theme focuses on:

- Assessing the effects of inhibitor concentration (**R**) on asphaltene aggregation, initially measured as the average size of clusters (**P**).
- Presenting PMV as a more thermodynamically meaningful parameter (**P**) for assessing the solubility of asphaltene-inhibitor mixtures, compared to the initial use of average cluster size.

- Investigating the interplay between 3 actors: **A**, inhibitor-asphaltene interactions (which can either disrupt or stabilize clusters), **B**, inhibitor-inhibitor interactions (which vary with concentration), and **C,** asphaltene-asphaltene interactions (which promote aggregation).

Theme 3: Fine-Tuning Equivariant Graph Neural Networks for Adsorption Energy Prediction

Accurate prediction of adsorption energies for compounds on surfaces is critical for applications ranging from catalysis to CO_2 capture and materials design to asphaltene deposition. Adsorption energies are traditionally calculated using first-principles methods like Density Functional Theory (DFT), which provides high accuracy but is computationally intensive. This limits its practicality for large systems, complex adsorbates, or extensive high-throughput screenings. While significant progress has been made using machine learning models, particularly 3D-equivariant graph neural networks (GNNs), their application to large, complex adsorbates remains underexplored.

Here, the emphasis is on extending the application of 3D-equivariant graph neural networks (GNNs) to aromatic molecules, which have historically been underrepresented in existing datasets and models. The competing interactions in this framework are not physical but captured by two properties of the training dataset: actor **A**, representing fine-tuning on a specialized aromatic dataset that accentuates the nuanced aromatic interactions essential for accurate adsorption modeling, and actor **B**, representing training on a broader, more diverse dataset that promotes overall generalizability with more training data. The control parameter (**R**) in this scenario includes factors such as dataset composition and fine-tuning strategy, both of which influence the predictive performance (**P**) of the model as measured by specific evaluation metrics.

By systematically exploring various fine-tuning methods, this theme seeks to better capture the energetic and geometric nuances governing adsorbate–substrate interactions, thereby advancing our ability to predict adsorption energies for industrially relevant systems.

1.5. Outline of the Dissertation

The rest of the dissertation is structured as follows: **Chapter 2** covers the first theme, which presents a series of MD simulations examining the simultaneous aggregation/adsorption of a

model asphaltene and the coalescence of water-in-oil droplets, revealing the simultaneity of asphaltene aggregation/adsorption and water droplet growth. This theme emphasizes the development of analysis methods that uncover new insights into droplet coalescence modes and their dependence on molecular interactions. Such insights are crucial for improving the design of water-in-oil emulsions and mitigating issues like pipeline clogging.

In **Chapter 3**, a novel method is introduced for calculating partial molar volumes (PMV) from molecular simulation trajectories, validated against experimental data. Its application was demonstrated using asphaltene-containing with the presence of inhibitors to probe molecular-level mechanisms of inhibition. This would conclude the second objective of the dissertation. The PMV method offers a quantitative tool for evaluating solubility changes. This approach enhances our understanding of how molecular positioning within aggregates influences the extent of aggregation and offers a predictive framework for optimizing inhibitor formulations.

Finally, **Chapter 4** explores the prediction of adsorption energies for large aromatic adsorbates using Equivariant Graph Neural Networks, detailing challenges such as data availability and model scalability, and showcasing how advanced machine learning techniques can enhance our understanding of adsorption energy.

2. Data Mining Guided Molecular Investigations on the Coalescence of Water-in-Oil Droplets

Published as **Hasan Imani Parashkooh**, Cuiying Jian, *Data Mining Guided Molecular Investigations on the Coalescence of Water-in-Oil Droplets*, Energy & Fuels 36 (4) (2022) 1811-1824

Name	Contributions
Hasan Imani Parashkooh	Conceptualization Formal analysis Investigation Software Visualization Writing - original draft
Cuiying Jian	Funding acquisition Methodology Project administration Resources Writing - original draft, Review and editing Supervision

2.1. Introduction

Asphaltenes have been known for a long time to play a major role in stabilizing emulsions [34,35,37]. For instance, Mclean et al. [38] identified asphaltene as the primary agent in stabilizing water-in-crude oil emulsions. Correspondingly, asphaltene behaviors at the oil-water interface, as well as the stabilizing mechanisms, have drawn considerable attention in the literature.

In the work of Yu et al. [39], it was found that asphaltenes are more active at seawater/oil interfaces compared to pure water/oil interfaces, and these interfacial activities are inversely correlated with the molecular weight of asphaltenes. Alvarez et al. [40] correlated the turning point of interfacial tension with the breakage of asphaltene layers formed on the oil-water interface. Poteau et al. [41] showed that high or low water pH can enhance the interfacial activity of asphaltenes by inducing electric charges to functional groups in their molecular structures. Using microfluidic experimental setups, Lin et al. [42] and Zhang et al. [10] revealed that increasing asphaltene concentrations can decrease the coalescence rate of water-in-oil emulsions. Yarranton et al. [43] also showed that interfacial behaviors of asphaltenes depend on concentrations. Specifically, at low concentrations, asphaltenes reversibly absorb as molecules and form a film with low elasticity that cannot stabilize emulsions, and in contrast, at high concentrations, asphaltenes irreversibly adsorb as nanoaggregates that can resist compression and stabilize emulsions. To overcome these phenomena, microwave heating [36] and chemical inhibitors [16] have been used to diminish the emulsification caused by asphaltenes.

Generally, asphaltenes consist of diverse components [44] and not all components are interfacially active [45,46]. Instead, it was proposed that asphaltenes can be separated into interfacially active asphaltenes (IAA) and remaining asphaltenes (RA) [45,47]. Rahham et al. [48] compared the structures of IAA and RA, and showed that they have a similar molecular weight, but IAA have smaller aromatic cores with more functional groups (such as sulfuric or carboxylic groups). Similar results have also been reported in the work of Ballard et al. [49] These differences lead to the different interfacial structures formed by asphaltenes. Yang et al. [45] showed that IAA forms rigid layers and irreversibly adsorbs to the oil−water interface that become thicker and more rigid over time. On the other hand, interfacial films formed by RA are softer and exhibit reversible adsorption [45].

To fundamentally understand effects of asphaltene molecular structures on their behaviors and properties, computational studies have also been carried out at the atomistic levels. Due to the complexity of asphaltenes, various model compounds have been proposed as proxies [50–53], such as Violanthrones [54], perylene bisimides [55], Hexabenzocoronenes [56], "average" model structures representing specific experimental data [57], etc. These models aimed to reveal the effect of diverse molecular moieties, and were first employed to investigate asphaltene aggregations that simultaneously accompany asphaltene interfacial activities. In this regard, $\pi - \pi$ interactions between the cores [58,59], hydrogen bonds between heteroatoms [17,51], hydrophobic interactions [60], etc. have all been observed to favor asphaltene aggregation, and the aggregated structures range from parallel stacking to T-shaped stacking depending on the driving forces [61]. Following these works, significant amounts of simulation works have been further performed on the interfacial properties of asphaltenes.

Using molecular dynamics (MD), Teklebrhan et al. [62] studied the effect of the terminal moiety structure and aromaticity of the organic solvent on the partition of asphaltene molecules at the interface. Their results showed hydrophobic aromatic moieties can hinder the adsorption process, while carboxylic functional groups on the side chains will promote asphaltene adsorption. Kuznicki et al. [19] employed MD to investigate the configurations of asphaltene molecules at water/toluene interfaces and showed that asphaltenes that absorbed onto interface prefer to be parallel to each other but perpendicular to the planar interface. This configuration was also observed on curved interfaces, for instance, in the works of Liu et al. [63] and Jian et al. [20]. In those two works, asphaltenes were reported to form a steady protective film wrapping the water droplets, which is responsible for the stabilization of water-in-oil emulsions. In addition, Jian et al. [20] found that unabsorbed asphaltene aggregates that are floating around can also introduce barriers to impede droplet coalescences and thus stabilize emulsions.

From the above, a fair amount of research has been done on the interfacial behaviors of asphaltenes in conjunction with their aggregation activities. However, the mutual effects of droplet coalescence and asphaltene activities are relatively less discussed in the literature. In fact, for all the MD works reviewed above [19,20,61–63], either planar interfaces or pre-formed water droplets were used to represent emulsions. Planar interfaces [15,19,61,62,64–66] directly prevent the investigation of droplet coalescences. In the works where pre-formed droplets were adopted

[20,63], the maximum number of droplets is no more than 2, and asphaltene molecules were directly placed near the interface, diminishing the simultaneity of droplet coalescences and asphaltene aggregation/adsorption. From computational perspectives, this is due to the difficulties in extracting and quantifying collective behaviors of interests from the multi-dimensional space in a simulation system. As one example, for a typical MD system containing ~50,000 atoms, a total of 150,000 coordinates will be generated at each time frame. This is, to identify collective behaviors among 50,000 atoms (e.g., the aggregation pathways of these 50,000 atoms), 150,000 points are being considered at each time frame. Common analyses of MD trajectories typically need to be performed over multiple frames, which would involve significantly more data points. One way to tackle this challenge is to use machine learning tools such as diffusion map to reduce the dimensions of the system [67]. However, these tools abstract features that may not have physical meanings [23,68].

To bridge the above gap, in this study, both water and asphaltene molecules are randomly distributed in the simulation box (see section 2.2), leading to simultaneous aggregation/adsorption and droplet coalescences. We then reported detailed analysis on the behaviors of both asphaltene and water molecules, with a major focus on water molecules to reveal coalescence pathways. A data map was developed to investigate growth modes from atomistic levels. The remainder of this chapter is organized as following: simulation methods are introduced in section 2.2, results are presented in section 2.3, and final conclusions are given in section 2.4.

2.2. Method

2.2.1. Simulated Systems

The asphaltene model simulated here is based on Violanthrone-79 (VO-79, $C_{50}H_{48}O_4$), which has been widely used in the literature as a surrogate for asphaltenes [19,69–71]. The molecular structure of VO-79 (see Figure 2-1) consists of one large polyaromatic region and two long aliphatic side chains; with 9.0 wt% of oxygen contents, it can represent asphaltene fractions that are responsible for stabilizing water-in-oil emulsions [72]. Pentane was chosen as the oil phase to promote asphaltene aggregation and adsorption [73], and thus allow us to probe the simultaneity of aggregation and the affected coalescence. It also needs to be pointed out here that at nanometer

scales, the precipitation effect of pentane is insignificant due to the negligible presence of gravitational forces.

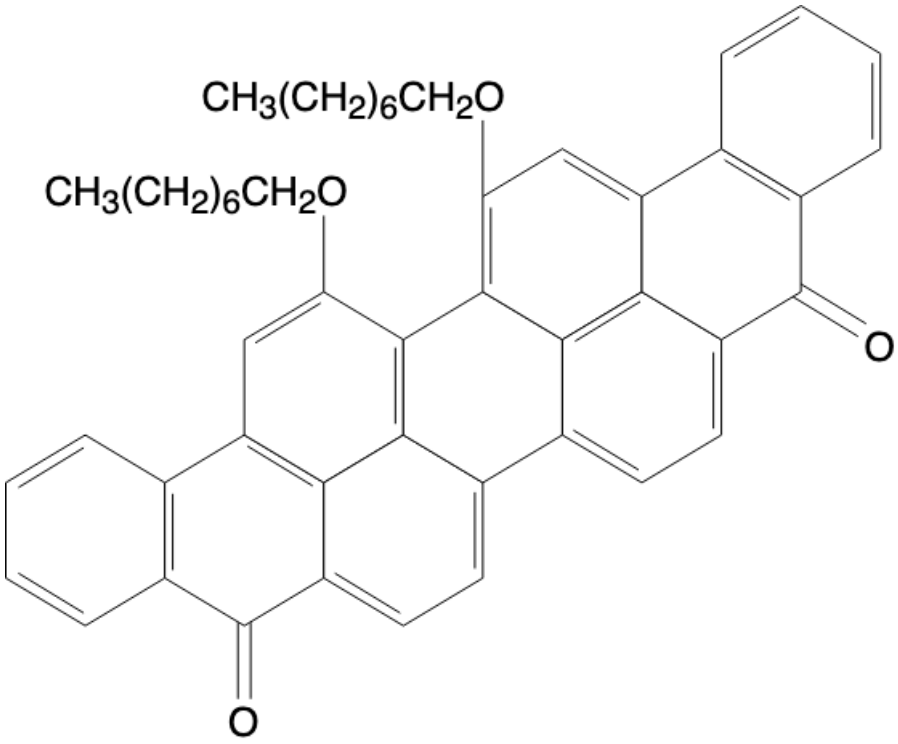

Figure 2-1 Chemical structure of asphaltene model compound (VO-79)

Adopting VO-79 as a model compound for asphaltenes, a total of 6 systems were built to investigate asphaltene aggregation and water-in-oil coalescence. The details of these systems are listed in Table 2-1. The first system (A16-W0) in Table 2-1 serves as the control system, which doesn't have water molecules in it. The following 3 systems were designed to study the effect of number of water molecules. Each of these systems has 16 VO-79 molecules, and a certain number of water molecules. For instance, system A16_W300 consists of 16 VO-79 molecules and 300 water molecules. During the construction of initial configurations, the 16 asphaltene molecules were first inserted at random positions inside the simulation box (Figure 2-2), followed by randomly adding water molecules. Then, the box was filled with pentane solvents. Following these systems, to study the effect of asphaltene concentrations, two additional systems (A80_W600 and A160_W600 in Table 2-1) were then built using the same procedure. Both systems are of 600 water molecules but a larger number of VO-79 molecules (80 and 160, respectively, for system A80_W160 and A160_W160).

Table 2-1 Properties of the simulated systems

System	Asphaltene Molecules	Pentane Molecules	Water Molecules	Initial Dimension (nm)	Final Dimension (nm)
A16_W0	16	4800	0	10^3	9.808^3
A16_W300	16	4800	300	10^3	9.845^3
A16_W600	16	4800	600	10^3	9.88^3
A16_W900	16	4800	900	10^3	9.913^3
A80_W600	80	8677	600	10^3	12.15^3
A160_W600	160	7964	600	10^3	12.05^3

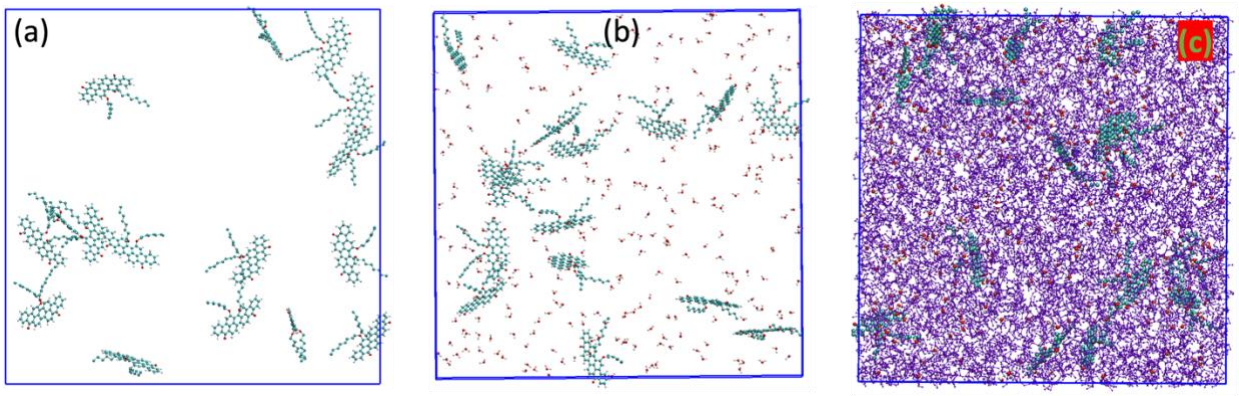

Figure 2-2 Construction of initial configurations: (a) asphaltene molecules were randomly distributed in the simulation box, (b) water molecules were added, and (c) pentane molecules were added.

2.2.2. Simulation Details

The topology of VO-79 molecule was first obtained using PRODRG [74]. Then, it was modified by manually adjusting the partial charges and charge groups in order to be compatible with the force field parameter set 53A6 of GROMOS96 [75]. This approach has been showed to be suitable for probing the molecular dynamics of polyaromatic cores [76]. The simple-point-charge model was chosen here for water molecules, which has been shown to be suitable for emulsion simulations [20,77]. For pentane molecules, the topology was generated from dipalmitoylphosphatidylcholine in the GROMOS96 force field parameter set 53A6 [75], through the *gmx pdb2gmx* routine available in GROMACS[78].

All simulations were performed using the MD package GROMACS (version 2018.6) [78]. For each system, static energy minimization was first performed to ensure that the maximum force is less than 1000.0 kJ/(mol·nm). Then, while restraining VO-79 molecules, water and pentane molecules were allowed to relax for 1 ns at 300 K and 1 bar. This relaxation helps to remove close contacts that may exist in the systems when building initial configurations, and thus solvate VO-79 molecules and stabilize the simulation systems. In addition, it also helps to obtain a distribution of droplet sizes (see section 3 for detailed results). Finally, all restraints were removed, and an NPT ensemble simulation was performed for 100 ns for each system. To compute electrostatic interactions, the particle-mesh Ewald method [79] was adopted, and a cut-off distance of 1.4 nm was imposed for van der Waals interactions. Periodic boundary conditions, the SETTLE algorithm [80] to constrain all bonds for water molecules, the LINCS algorithm [81] to constrain all bonds for VO-79 and pentane molecules, and a timestep of 2 fs were used for all dynamics simulations. Trajectories are saved per 10 ps, leading to consecutive frames being separated by 10 ps.

2.2.3. Droplet Analysis

To probe coalescence, water molecules in each system were first mapped to vertices of a graph called *G*. Then, an edge was created between any two vertices (water molecules), if the distance between the corresponding oxygen atoms is smaller than 0.32 nm. This distance criterion was chosen based on the plot of radial distribution function (RDF) for water oxygens (see section 7.1.1). Thus, the interconnected vertices of graph *G* represent water molecules in the same droplet. Examining graph *G* at different frames can thus provide an effective way to understand the coalescence of water-in-oil droplets (details in section 2.3). However, while we are focusing on water molecules here, the number of atoms involved during building graph *G* still exceeds that can be handled by heuristic search and direct visualization. Taking system A16_W300 as one example, which has the smallest number of water molecules, there are $300 \times 3 = 900$ coordinates for oxygen atoms. For such a system, to analyze 100 frames, 90,000 coordinates should be handled. We also note that for systems of 600 and 900 water molecules, the coordinates that need to be analyzed are doubled and tripled, respectively.

To perform the above analysis in an automated way, a customized method for data mining was developed. Briefly, this method inputs the trajectory of water molecules (more precisely the oxygen atom in each molecule) and simulation dimensions over time, and constructs graph *G* based

on the pairwise distances at each frame along the simulation trajectories. To address periodic boundary conditions, the convention of minimum image distance is applied. By assigning an ID number to each water molecule, changes in each droplet (i.e., interconnected components of G) are monitored by tracking ID variations (see section 7.1.2 for implementation details). To validate this method, we compared the number of droplets obtained with that calculated using the GROMACS module gmx *clustsize*. Our results almost overlap with those from the standard routine (see Appendix 7.1.3). Beyond tracking numbers of droplets, the superiority of our method is to pinpoint molecules in the same droplet (emulsions) and thus allow us to track the coalescence pathways. Given the large number of molecules involved, our method provides a simple, yet efficient way for visualization and quantification (details in section 2.3).

2.3. Results and Discussion

In this section, we will first present our results on asphaltene aggregation and adsorption, followed by coalescence studies using in-house developed tools.

2.3.1. Asphaltene Aggregation and Adsorption.

To probe the effect of water on asphaltene behaviors, Figure 2-4 shows the RDF (g(r)) plots for the separation of center of masses (COM) between polyaromatic cores during the last 5 ns of the 100 ns NPT production. We note that this last 5 ns (95-100 ns) is where our simulations have achieved equilibrium, given that the number of water droplets already approaches 1 within 30 ns of the simulation time (see Figure A1-3 in the Appendix 7.1). Furthermore, as long as equilibrium is attained, the trends presented below will remain the same irrespective of the time periods that the analyses were performed over. As can be seen in Figure 2-4, clearly, RDFs have the most significant peak around 0.45 nm in all systems. This is, adjacent polyaromatic cores have formed $\pi - \pi$ stackings, consistent with the results from literature [59,82,83]. However, the peak heights are evidently different among these systems. In fact, the following order is observed for the first RDF peak: A16_W0 > A16_W900 > A16_W300 > A16_W600. Additionally, at a larger COM separation distance (> 0.6 nm), system A16_W0 has two more evident peaks at ~0.7 nm and ~1.1 nm, respectively; among systems with water molecules, only system A16_W900 presents evident peaks at ~0.7 nm and ~1.1 nm. Thus, the presence of water indeed can affect the

aggregation patterns of VO-79 molecules. To further probe this, we further quantified VO-79 aggregates as presented below.

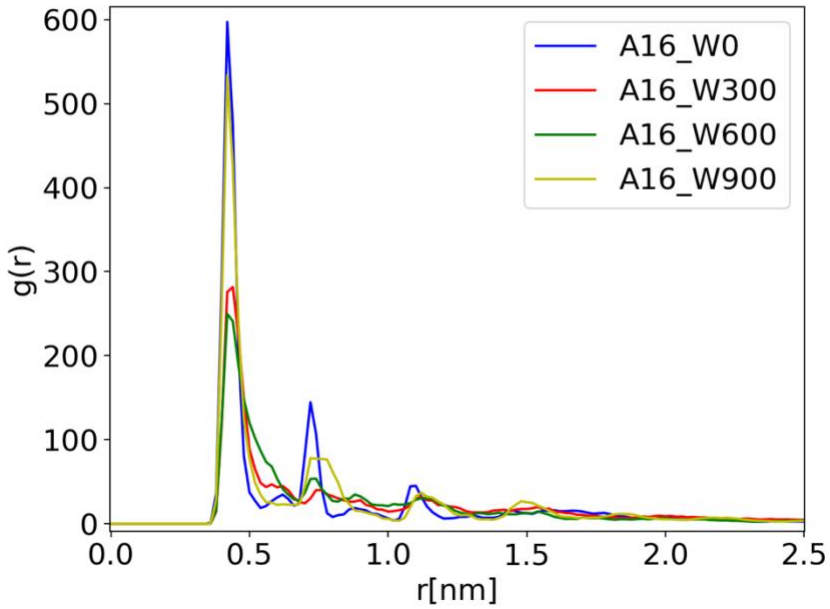

Figure 2-3 RDFs for systems containing 16 asphaltene molecules during the last 5 ns of the NPT production

To identify parallel stacking pairs, we calculated the COM distance (d_{ij}), as well as the angle σ_{ij}, between any two VO-79 cores (see Figure 2-5 (a) for a schematic). Following our previous works [59], a parallel stacking (PS) pair, including both direct parallel stacking and shifted parallel stacking, is formed between two VO-79 molecules, if COM distance is $\leq$ 0.75 nm and $\cos\sigma \geq 0.9$. Based on this, we further define a stacking block (see Figure 2-5 (b) for a schematic) as the structure consisting of consecutively stacked VO-79 molecules with their polyaromatic cores parallel with one another. The size of a stacking block was then quantified by the number of VO-79 molecules involved. As can be seen, this definition essentially describes the persistency of parallel stacking. By averaging over the last 5 ns of the production runs, the numbers of PS pairs and sizes of stacking blocks are plotted in Figure 2-5 (c) for systems A16_W0, A16_W300, A16_W600, and A16_W900.

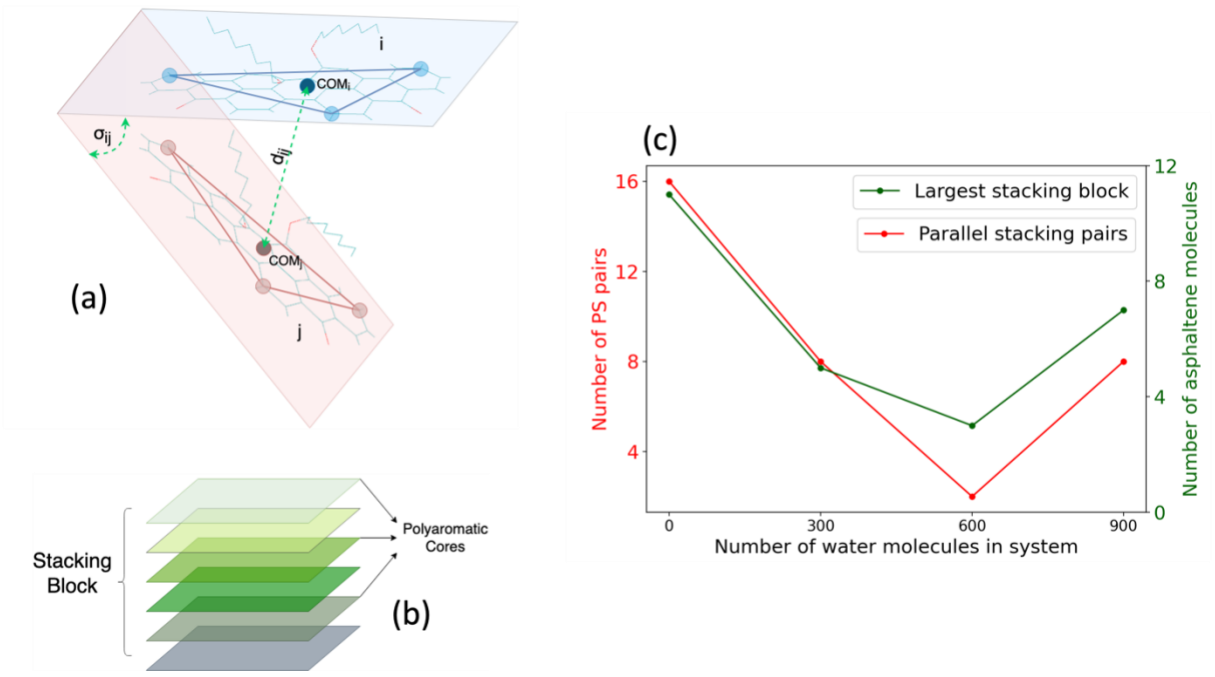

Figure 2-4 (a) The distance and angle measurements between 2 VO-79 molecules. Here, the distance was measured between the COMs of the cores, and 3 atoms were selected to represent the polyaromatic plane in each VO-79 molecule for measuring the angle. (b) A schematic to represent a stacking block of asphaltene molecules. (c) The effect of water molecules on aggregation in 16 VO-79 systems.

From Figure 2-5 (c), water molecules show non-monotonic effects on the number of PS pairs and the size of the largest stacking blocks. Specifically, with increasing the number of water molecules, the number of PS pairs, as well as the size of stacking blocks first decreases, and then increases. This observation is consistent with the peaks of RDFs shown in Figure 2-4, suggesting that the presence of water molecules will disrupt stackings at low concentrations, and help with stackings at high concentrations (more information on the stacking blocks is available in the section 7.1.4) To visualize the overall aggregation patterns under the aforementioned influence, Figure 2-6 shows the corresponding snapshots for aggregated structures formed in each system. Here, the aggregated structures are determined by distance only. In other words, two VO-79 molecules were defined to be in the same aggregate if the distance between COMs of their polyaromatic cores is less than 0.75 nm. As can be seen from Figure 2-6, while stacking is formed in all systems, distinct differences exist in the aggregated shapes.

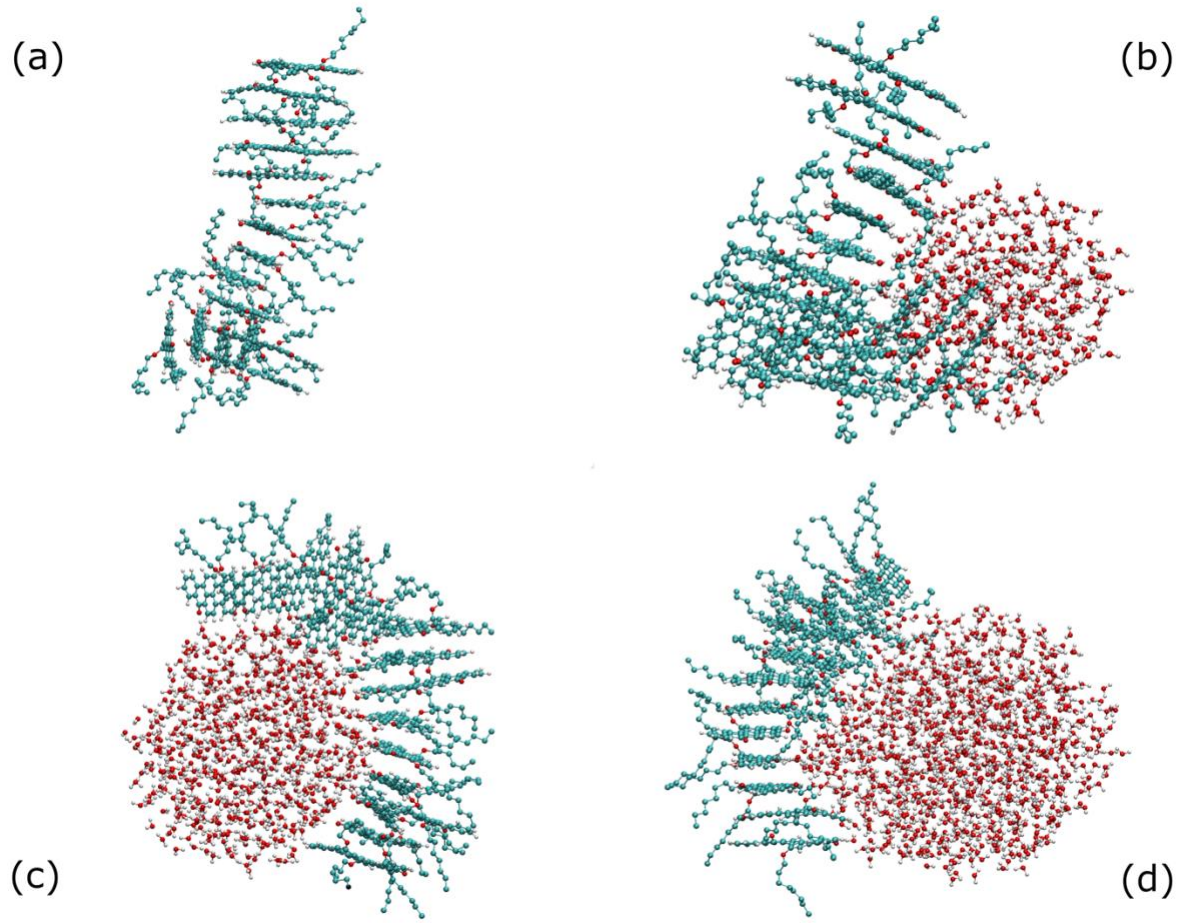

Figure 2-5 Bending effect of the water droplet on VO-79 aggregates visualized by snapshots taken near the end of the simulation: (a) system A16_W0, (b) system A16_W300, (c) system A16_W600, and (d) system A16_W900.

Without water molecules (system A16_W0, Figure 2-6 (a)), the aggregate renders a straight, rod-like geometry. With the addition of water molecules, model asphaltene molecules migrate to the water/pentane interface due to the interfacial activities of VO-79 molecules. As can be seen in Figure 2-6 (b)-(d), VO-79 molecules absorbed onto the droplet with their aromatic cores perpendicular to the interface, consistent with literature results [82]. This particular adsorption in conjunction with $\pi - \pi$ stacking has created different shapes of stacking blocks, depending on the size of the water droplet in each of the simulated systems. Specifically, in system A16_W0 during the time 95-100ns of the production phase, 15 out of 16 asphaltene molecules have stacked with at least one of the other asphaltenes, which is the highest fraction in all systems and corresponds to the largest stacking block shown in Figure 2-5 (c). In system A16_W300 (Figure 2-6 (b)), with the presence of water molecules, the tendency of polyaromatic cores to be perpendicular to the

water/oil interface causes the aggregate to bend, and further deviates the adjacent polyaromatic cores from the parallel configuration. Furthermore, it seems that limited by the size of the droplet, there is no sufficient surface area for the entire aggregate to be completely adsorbed. With increasing the number of water molecules (system A16_W600, Figure 2-6 (c)), all the VO-79 molecules adsorbed onto the water droplet, leading to evident bending of the aggregate. With further increasing the number of water molecules (system A16_W900, Figure 2-6 (d)), the increase in the droplet diameter results in less bending in the aggregate and meanwhile provides sufficient surface areas for VO-79 adsorption. This is, the aggregation patterns of VO-79 cores are resultant from two competing factors, parallel aggregation driven by interactions among polyaromatic cores, and bended shapes driven by adsorption to be perpendicular to the interface. The synergistic effects of these two factors lead to the non-linear behaviors in the number of PS pairs as well as the size of the largest stacking block (Figure 2-5 (c)). Considering our simulation length (100 ns), it may also be possible that the combined water-asphaltene interactions slow down the diffusion of VO-79 molecules, leading to that parallel stacking in systems A16_W300 and A16_W600 would need much longer time to be formed compared to that in system A16_W0. Nevertheless, with 900 water molecules, the diffusion barrier could be overcome by sufficiently large water-asphaltene interactions.

For systems (A80_W600 and A160_W600) with an increased number of VO-79 molecules, the final configurations are shown in section 7.1.4 of the Appendix. As can be seen, unlike Figure 2-6 where each system only has one droplet, these two systems are having more than 1 droplet by the end of the 100 ns simulation. To understand the effect of VO-79 molecules on droplet coalescences, we will present detailed analysis on the droplet coalescence using our in-house developed tool.

2.3.2. Water Droplet Growth.

To probe the effect of VO-79 molecules on the coalescence of water-in-oil droplets, the number of water droplets at each simulation frame was first determined using the method described in section 2.3. The corresponding results were plotted in Figure 2-7 as a function of simulation time for systems A16_W300, A16_W600, and A16_W900. Figure 2-7 (a) shows the equilibrations stage, where the positions of VO-79 molecules were restrained. Therefore, the time interval for the equilibration phase was selected to cover the first half of this phase (for water droplet sizes, see

section 7.1.5). Figure 2-7 (b) represents the production stage of the simulation, where all molecules are allowed to freely move. Here, t = 6.14 ns was chosen, since it is the first moment at which a single droplet can be observed (see section 7.1.4 in the Appendix).

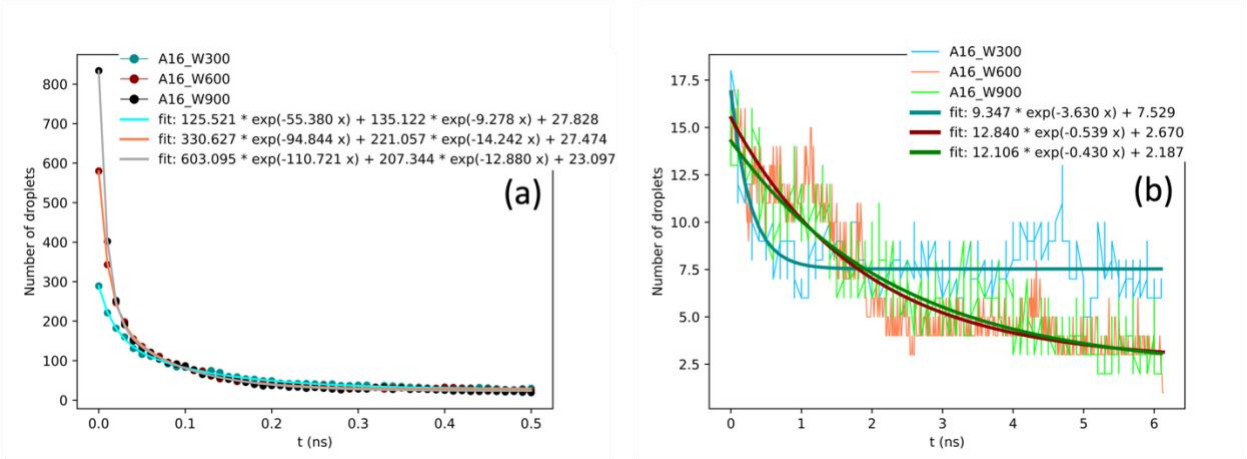

Figure 2-6 Number of water droplets during: (a) the first 0.5 ns of the equilibration phase and (b) the first 6.14 ns of the production phase. The thicker lines represent fitted curves.

Following above, a double exponential function, equation (2-1) is then employed to fit these curves in the equilibration phase. For the production phase, since the number of water droplets is smaller, a single exponential function (equation (2-2)) was used for fitting. The average values of the derivatives (defined in equation (2-3)) of the fitted function are then calculated to represent the average coalescence rates, and equations (2-4) and (2-5) are the specific forms for equilibration and production phases, respectively.

$$f_1(t) = a_1\, e^{-b_1 t} + a_2\, e^{-b_2 t} + \mathrm{c} \tag{2-1}$$

$$f_2(t) = a\, e^{-bt} + \mathrm{c} \tag{2-2}$$

$$\overline{\left(\frac{df}{dt}\right)} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{df}{dt}\, dt = \frac{1}{t_2 - t_1}\left[f(t)\right]_{t_1}^{t_2} \tag{2-3}$$

$$\overline{\left(\frac{df_1}{dt}\right)} = \frac{1}{t_2 - t_1}[a_1(\, e^{-b_1 t_2} - \, e^{-b_1 t_1}) + a_2(\, e^{-b_2 t_2} - \, e^{-b_2 t_1})] \tag{2-4}$$

$$\overline{\left(\frac{df_2}{dt}\right)} = \frac{1}{t_2 - t_1}[a(\, e^{-b t_2} - \, e^{-b t_1})] \tag{2-5}$$

In the equations above, $f_1(t)$ is the double exponential function with coefficients a_1, a_2, b_1, b_2 and c to be fitted to the data from the equilibration phase, while $f_2(t)$ represents the single exponential function that was fitted to the production data. It needs to be pointed out here that while experimentally, the coalescence rate is measured using the decrement in the surface area of the droplet covered by the medium [10], the decrement in the total number of water droplets is used here to calculate the coalescence rate. This is due to the fact that the droplets studied here are in the initial stage of formation, meaning that each water droplet only consists of a few molecules. Nevertheless, the rates obtained here (see Table 2-2) show an increasing trend with increasing the number of water molecules, consistent with experimental works reported in the literature [84]. Following the same method, we also calculated the coalescence rate in the production stage of the simulation, where all molecules are allowed to freely move. Again, an overall increasing trend is observed in the coalescence rate.

Table 2-2 Coalescence rates in systems with fixed number of asphaltene molecules

System	Coalescence Rates ($\frac{1}{ns}$)	
	Equilibration Phase	**Production Phase**
A16_W300	519	1.57
A16_W600	1103	2.02
A16_W900	1620	1.83

To show the effects of asphaltene concentrations on coalescence rates, we performed similar analysis on systems A80_W600 and A160_W600. In conjunction with system A16_W600, the plots obtained are shown in Figure 2-8, and the average coalescence rates are shown in Table 2-3. First, for the equilibration stage, with increasing the number of VO-79 molecules, the

coalescence rate shows a decreasing trend, consistent with literature results that increasing asphaltene concentration can decrease coalescence rates [10,85]. However, in the production stage, the trend is reversed, in that with increasing the number of VO-79 molecules, an increasing trend is observed in the coalescence rate. This is due to the reason that at the start of the production phase, because of faster coalescences in the equilibration phase, system A16_W600 has already reached to a state that not too many water droplets were left around, while systems A80_W600 and A160_W600 still have a fair amount of droplets and thus can maintain a higher rate (referring to the later stage of Figure 2-8 (a) and the earlier stage of Figure 2-8 (b)).

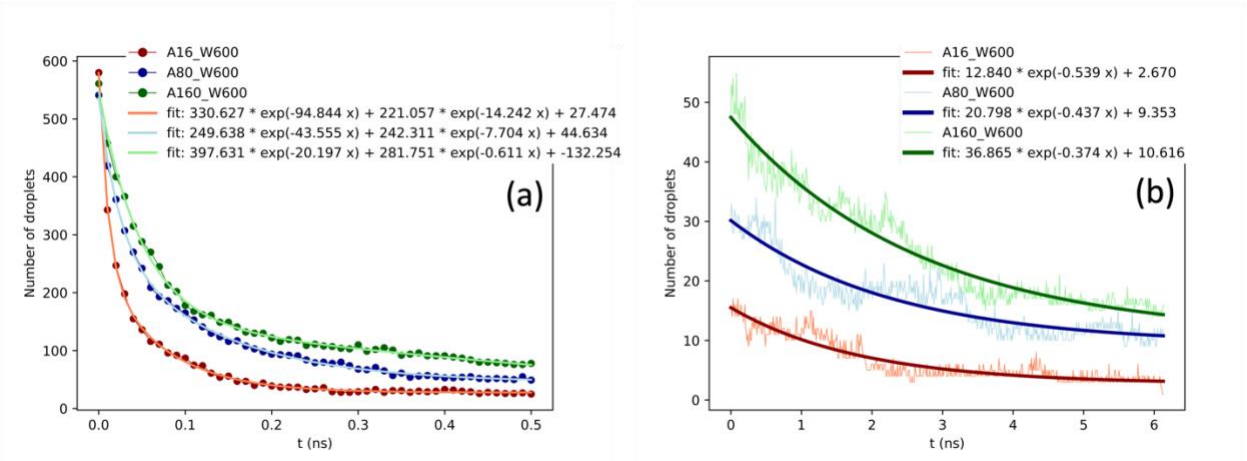

Figure 2-7 Number of water droplets during: (a) the first 0.5 ns of the equilibration phase, and (b) the first 6.14 ns of the production phase, for systems containing 600 water molecules. The thicker lines represent fitted curves.

Table 2-3 Coalescence rates in systems with fixed number of water molecules

System	Coalescence Rates ($\frac{1}{ns}$)	
	Equilibration Phase	Production Phase
A16_W600	1103	2.02
A80_W600	973	3.73
A160_W600	943	5.41

To probe whether water and asphaltene concentrations can affect the coalescence modes, we categorized the growth of water droplets into different modes by comparing the largest droplets

at two consecutive frames (time interval: 10 ps). The following scenario was named as the growing mode, where the largest droplet absorbs distinctly smaller ones and merges with them to create an even larger droplet at the next frame. The criteria used to determine the growing mode are: 1) the largest droplet at the current frame must belong to the largest droplet at the next frame, and 2) the largest droplet at the current frame must account for more than 50% of molecules in the largest droplet at the next frame. The two criteria were implemented by comparing the IDs of water molecules contained in the largest droplets at two consecutive frames being analyzed (see section 2.3 about how water molecules are grouped based on the droplet analysis). This is, criterion 1) is set to be met, if molecular IDs found in the largest droplet at the current frame are also present in the largest droplet at the next frame; criterion 2) is set to be met, if molecular IDs from the largest droplet at the current frame represent more than 50% of the total IDs in the largest droplet at the next frame. The rationale behind these definitions is to identify the prevalence of the case where the largest droplet serves as nucleus sites to promote coalescence. In other words, unlike in the calculation of coalescence rates, all water droplets were taken into consideration, for the analysis on coalescence modes, we are focusing on the largest droplet in each system. The scenario, where the largest droplet remained the same at consecutive frames, was named as the stalling mode. This is, the largest droplets at the current frame and the next frame contain the same water molecules. This is implemented, again, by comparing molecular IDs. Figure 2-9 shows the growing mode observed in systems A16_W600, and the stalling mode in A16_W900 as representatives. Note that in the stalling mode, while smaller droplets can collide in this scenario, this collision doesn't surpass the size of the largest droplet that is already in existence. Since molecular IDs are recorded during implementing our criteria, the stalling mode also confirms that the composition of the largest droplet remains unchanged.

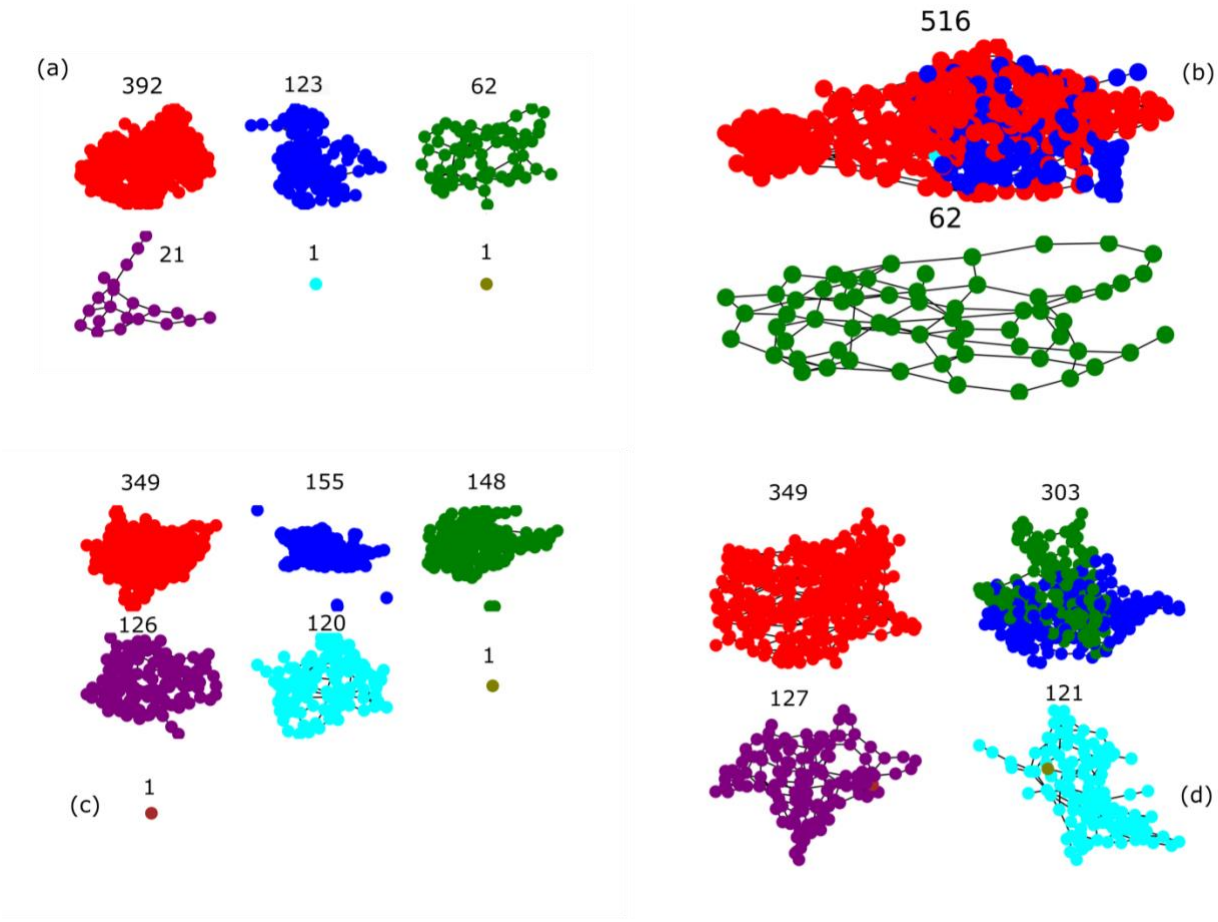

Figure 2-8 (a) and (b): growing mode observed in system A16_W600. At $t = 2.19$ ns shown in (a), the largest droplet contains 392 molecules, and growing from $t = 2.19$ ns, this largest droplet merged with the second largest droplet (123 molecules) and absorbed a single water molecule. This leads to that the largest droplet at $t = 2.2$ ns shown in (b) contains 516 molecules. (c) and (d): stalling mode observed in system A16_W900. At both timesteps (t = 2.59 ns shown in (c) and t = 2.6 ns shown in (d)), the largest droplet remained the same size, while smaller droplets merged from t = 2.59 ns to t = 2.6 ns. Noe that in the stalling mode, while smaller droplets can collide in this scenario, this collision doesn't surpass the size of the largest droplet that is already in existence.

Using the above definitions, Figure 2-10 summarizes these modes observed in the initial stages of equilibration and production phases for systems A16_W300, A16_W600, and A16_W900. Figure 2-10 (a) and (c) plots the stalling and growing modes as a function of simulation time. In these two figures, the coalescence mode at each frame is determined, classified, and colored according to the legend; blank means coalescence modes other than stalling and growing are in presence. Meanwhile, Figure 2-10 (b) and (d) summarize the total percentage of these modes, which defined as the total time, where stalling or growing modes are observed, over

the total time periods, 500 ps and 6140 ps for equilibration and production phases, respectively. Here, the total time of stalling or growing modes being observed is determined by the total length of the corresponding streaks (details on these calculations and streak lengths are available in Appendix, section 7.1.6).

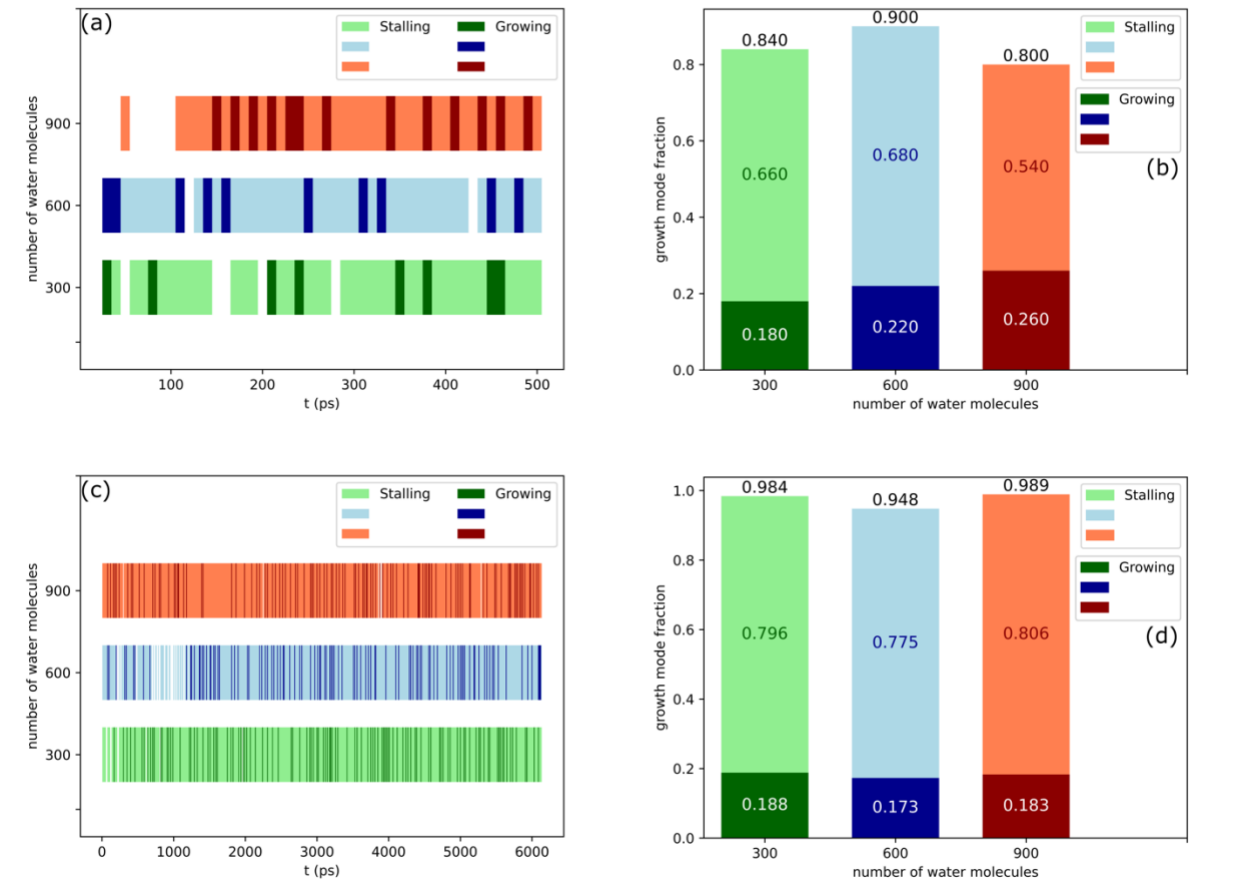

Figure 2-9 Coalescence mode analysis for systems A16_W300, A16_W600, and A16_W900. (a) and (c) show the distribution of coalescence modes in equilibration and production phases, respectively; (b) and (d) plots the percentage of different modes.

As can be seen from Figure 2-10, stalling modes dominate in both phases. This is due to the presence of VO-79 molecules, which would act like walls to impose barriers on the diffusion of water droplets as well as reduce the related collision probability. At the equilibration stage, with increasing the number of water molecules, while the percentage of the stalling mode shows a non-monotonic trend, the percentage of the growing mode exhibit an increasing trend. The latter is due to the fact that increasing the number of water molecules would increase the probability of collision

and merging that can lead to growing modes. Entering the production phase, while the percentage of growing mode in system A16_W300 system remain approximately the same in both equilibration and production, for systems A16_W600 and A16_W900, the percentage of growing modes is evidently reduced compared to that in the equilibration stage. Contrarily, for the same system, the stalling mode is more prevalent in the production phase compared to that in the equilibration phase. More interestingly, the consecutive nature of the growing mode is more diversified in the production phase (see Appendix, section 7.1.6 for quantitative comparisons). In the production stage, all systems have reduced number of droplets with, on average, larger droplet sizes compared to the equilibration stages (see Figure 2-7 and section 7.1.5 in Appendix). This is, the collision probability is intrinsically limited by the smaller number of water droplets, regardless of the number of water molecules. Furthermore, VO-79 molecules start to aggregate, which would further prevent the collision of water droplets (note that coalescence rates in the production stage is much smaller than those in the equilibration stage, see Table 2-2). Correspondingly, the growing mode is reduced, and the stalling mode is increased.

We also performed mode analysis for systems A80_W600 and A160_W600, and plotted the corresponding results in Figure 2-11 together with system A16_W600 to probe the effect of asphaltene concentrations. Similar to the observations in Figure 2-10, stalling modes dominate in both equilibration and production phases, and for all systems, the percentage of stalling mode is increased in the production phase compared to that in the equilibration phase, with an opposite trend observed in the population of growing modes. Furthermore, in the equilibration stage, increasing asphaltene concentrations leads to a decrement in the percentage of the "nucleus" mode (stalling plus growing). More specifically, the percentage of stalling modes is marginally reduced, while the percentage of growing modes is slightly increased. This seems to contradict with the analysis on the decreasing trend in the coalescence rate by increasing the number of VO-79 molecules, where it was proposed that more VO-79 molecules would create more barriers to impede coalescence. However, as mentioned earlier, when calculating coalescence rates, all water droplets were taken into consideration, while for the analysis on coalescence modes, we are focusing on the largest droplet at each frame in each system. This is, the increment observed in the percentage of the growing mode might be resultant from the fact that increasing number of VO-79 molecules would create much more droplets with much smaller sizes (see Tables A7.1.2 and A7.1.3 in the Appendix). Overall, these abundant, small droplets have lower coalescence rates due

to the presence of more VO-79 molecules, but the presence of extremely small water droplets (e.g., the ones with less than 10 molecules) in system A80_W600 and A160_W600 may actually collide with others more easily (note that in the equilibrations stage, VO-79 molecules are restrained). In the production phase, as the number of asphaltene molecules increases, the growing mode exhibit a decreasing trend while the stalling mode has an increasing trend. This trend could be caused by the aggregation of floating VO-79 molecules, which brings in dynamic barriers to impede collisions among the water droplets, and thus promote the stalling mode. In addition, the protective layers formed by VO-79 on the droplet surface can also prevent droplet coalescence (see section 7.1.5 in the Appendix). We also note that systems A80_W600 and A160_W600 have more than one water droplet by the end of the 100 ns production simulation, confirming the impeding role of VO-79 molecules.

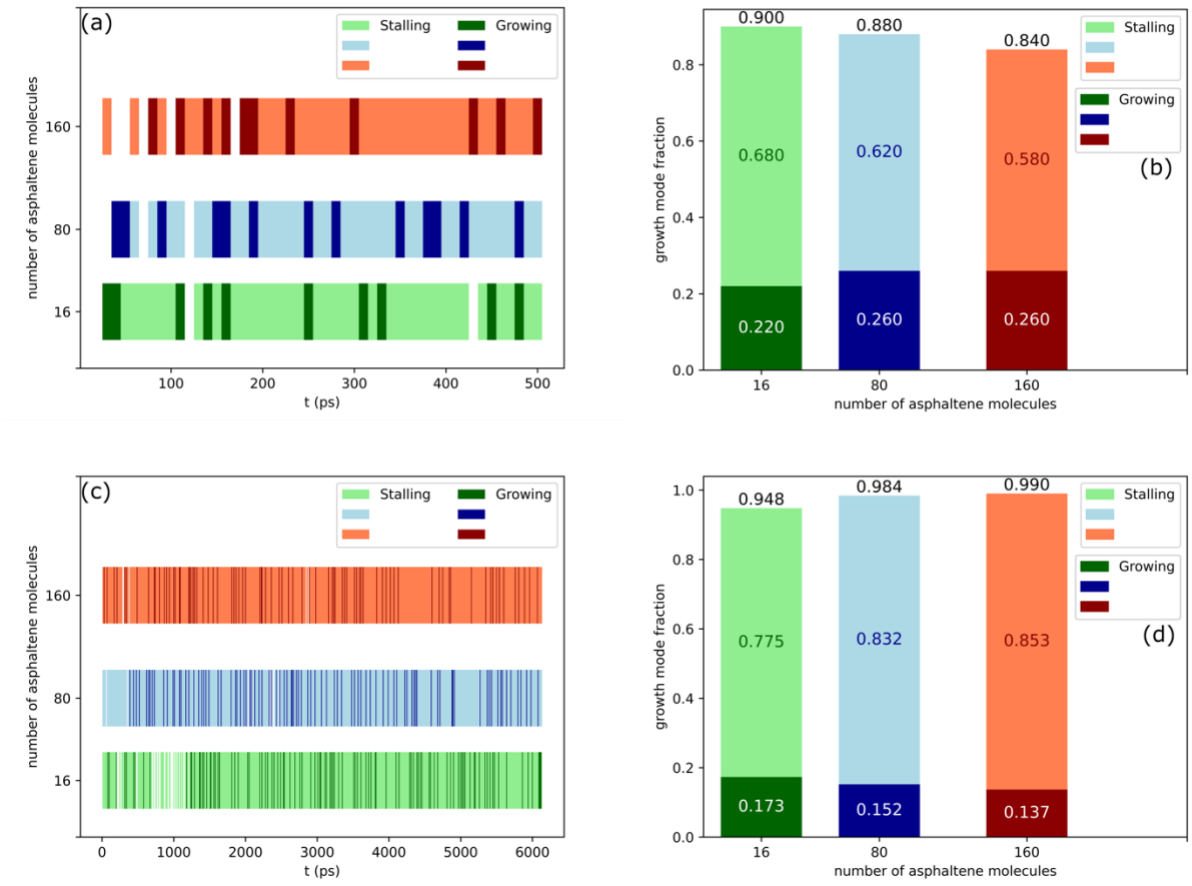

Figure 2-10 Coalescence mode analysis for systems A16_W600, A80_W600, and A160_W600. (a) and (c) show the distribution of coalescence modes in equilibration and production phases, respectively; (b) and (d) plots the percentage of different modes.

To further validate the above observation, we adjusted the criterion in the definition of the growing mode presented earlier. Specifically, the criterion was further increased in that the largest droplet at the current frame must account for more than 70% of molecules in the largest droplet at the next frame. Figure A1.7 in the Appendix (section 7.1.7) shows the effect of this adjustment on the abundances of growing modes in all systems for both equilibration and production phases. As it can be seen, almost all trends observed before still hold, confirming the dominant role of the largest droplet as the nucleus site.

To probe other modes existing in the system, we took a detailed look at the "blank" parts in the streak plots. The following mode is identified: small droplets from the current frame collide and merge into a larger one at the next frame, and this newly formed droplet surpass the size of the largest droplet at the current frame. This scenario is plotted in Figure 2-12 using system A80_W600 as a representative. Figure 2-12 (a) shows the largest droplets at t=0.05 ns, and (b) shows the largest droplets at t=0.06 ns. From (a) to (b), small droplets (colored in yellow) at (a) merged into a larger one at (b), which exceeds the size of the largest droplet observed at the frame shown in (a). In other words, the largest droplet is not the nucleus site for this particular frame. Similarly, (c) shows the largest droplets at t=2.46 ns, (d) shows the largest droplets at t=2.47 ns. From (c) to (d), small droplets (colored in yellow and cyan) at (c) merged into a larger one at (d), which exceeds the size of the largest droplet observed at the frame shown in (c). Again, this means the largest droplet is not the nucleus site. It needs to be pointed out that the mode shown in Figure 2-12 could be easily found in other A16 systems (see section 7.1.8 in the Appendix).

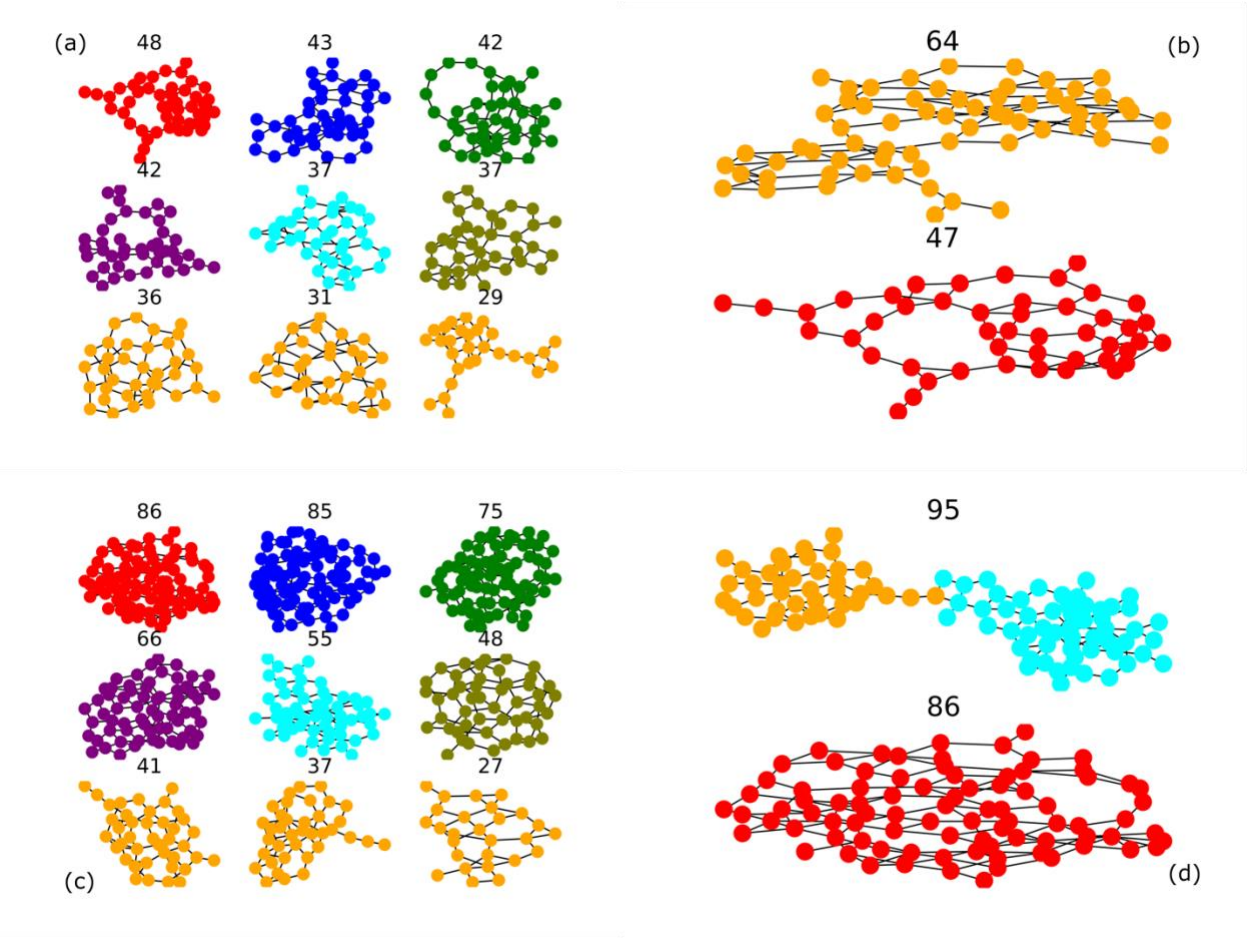

Figure 2-11 Non-nucleus mode observed in the production phase of system A80_W600: (a) shows the largest droplets at $t = 0.05$ ns, (b) shows the largest droplets at $t = 0.06$ ns. (c) shows the largest droplets at $t = 2.46$ ns, (d) shows the largest droplets at $t = 2.47$ ns. See the main text for detailed analysis.

2.3.3. Implications

From aggregation perspectives, asphaltenes belong to the family of polyaromatic compounds that are of wide applications in electronics. For instance, aggregates consisting of polyaromatic compounds have been used in organic thin-film (or field-effect) solar cells [86,87], conductive organosilica films [88], and photovoltaics liquid crystals [89]. These applications require polyaromatic aggregates to be of specific geometries, i.e., one-dimensional, rod-like, parallel stacked, etc.[90]. Our results reported here suggest that water droplets can serve as an effective medium to tune the aggregate shapes. From emulsion perspectives, compared to similar computational works in literature where the droplets are preformed [20,63,91,92], our work here started with dispersed water molecules. While number of molecules studied are still incomparable

with the length scale used in experiments, the coalescence mode analysis provides direct observations of the coalescence process that may be inaccessible from experiments. Indeed, the growth of water droplets is dominated by using the larger ones as the nucleus site, confirming experimental postulations reported in the work of Nowbahar et al. [93].

From method perspectives, our work can provide a new approach to quantitatively study coalescence phenomena in colloidal systems. It needs to be pointed out here our method is automated, in that the only inputs needed are atomistic trajectories with predefined criteria. Thus, the method here can provide a platform for analyzing cluster/coalescence of nanoscale emulsions for use in food industries [94], fuel systems[95], and pharmaceutical applications [96]. For instance, in the work of Moghaddasi et al. [96], the oil-in-water nanoemulsions were studied in the absence and presence of curcumin, and our methods can directly help to enable an automated review of simulation trajectories, and more importantly reveal the binding modes among oil, surfactant, and curcumin. In addition, the cluster analysis on water molecules can also help to address aggregation pathways in broad areas, such as the binding of drugs and carriers in drug delivery [97]. We also emphasize that for analyzing coalescence modes, our method is focusing on the largest droplet in each system to emphasize its prevalence and effects. While this method omits coalescence happening among smaller droplets, it can still identify them (see blank parts in Figures 2-10 and 2-11). It is also extendable to simultaneously track the stalling mode of coalescence and the collision among small droplets.

The classification of droplet dynamics into 'growing' and 'stalling' modes serves as a powerful quantitative tool to connect microscopic coalescence events with macroscopic system properties like emulsion stability. By defining a proxy for nucleation-driven growth, this method allows for a direct, comparative analysis of how different system components, such as asphaltene and water concentrations, impact the underlying mechanisms of phase separation. The utility of this framework is demonstrated by its ability to produce trends consistent with physical intuition; for example, correctly identifying that asphaltenes promote 'stalling' modes, thereby acting as emulsifying agents.

However, it is important to acknowledge the limitations of this approach. The classification is inherently dependent on the chosen time interval for analysis (Δt = 10 ps). While this window

was selected to balance the observation of meaningful dynamic events against thermal noise, the absolute frequencies of the identified modes would change with a different Δt. Nonetheless, the *relative trends* observed when comparing different systems are expected to be robust. Furthermore, this analysis simplifies the complex dynamics into a binary classification and does not explicitly capture other possible events, such as the simultaneous merger of multiple large droplets or droplet fragmentation. Despite these limitations, the method successfully fulfills its intended purpose as a robust indicator for tracking the prevalence of the dominant coalescence mechanism within the simulated systems.

Additionally, we note that using a single model compound such as VO-79 to represent complex asphaltenes may intrinsically have some limitations. For instance, our results may overemphasize the importance of parallel stacking, considering that polydisperse compounds may have a smaller extent due to the presence of dissimilar molecular structures [44]. Furthermore, the adsorbed configuration reported here may only represent one of the possible structures formed by asphaltenes at the oil/water interface [61]. This is, the usage of a single model may omit other information that can only be accessible by polydisperse models. However, adopting a simplified model is a standard approach in MD to unentangle complexity and gradually understand different factors [98]. In our current work, VO-79 is a proxy that can mimic the interfacial activities of asphaltenes, and thus allow us to probe the simultaneity of emulsion formation and asphaltene aggregation under the influence of asphaltene adsorption. Inclusion of other models is entirely feasible, but may obscure the synergistic effects of emulsion, adsorption, and aggregation because of the entanglement among asphaltene proxies. Similarly, adopting model compounds from a different source (e.g., the ones proposed in [99]) may lead to different stacking extents as well as adsorbed configurations. However, since parallel stacking is commonly observed among all asphaltene molecules [53], provided that the model asphaltene can mimic interfacial activities of asphaltenes, we expect similar modes to be observed for coalescence, and the synergistic effects of aggregation, emulsion, and adsorption should stay the same.

2.4. Conclusions

In this work, we probed the mutual effects of asphaltene aggregation/adsorption and water-in-pentane droplet coalescence. To simultaneously capture those phenomena, water and model asphaltene (VO-79) molecules were both randomly distributed in the simulation box. By the end

of the production run, it was confirmed that VO-79 molecules would absorb to water/oil interface with their polyaromatic cores perpendicular to water surface. This specific orientation competes with the parallel stacking among polyaromatic cores, leading to the non-monotonic trends observed in the preference of parallel stacking with increasing the concentration of water molecules. Specifically, to achieve the aforementioned perpendicular orientation at the interface, droplets of small radii would bend the aggregates of VO-79, and thus disrupt the parallel stacking among VO-79 molecules. With increasing the radii of water droplets, the bending extents needed to achieve perpendicular orientation are decreased, which helps to restore the parallel stacking among polyaromatic cores. We then employed in-house developed tools to investigate the coalescence of water-in-oil droplets. Coalescence rates were first obtained by fitting exponential functions to simulation data, and it was found that increasing the number of water molecules would increase the coalescence rate, while VO-79 molecules could impede droplet coalescence rates, consistent with literature results. More importantly, in all systems simulated, the droplet growth was dominant by using the largest droplet as the nucleus site, through stalling (where the largest droplet retains its sizes) and growing (where the largest droplet adsorbs smaller ones) modes. We emphasize again that for stalling modes, while smaller droplets may merge, the largest droplet remains unchanged. Thus, it is still the nucleus site from the perspective of tracking the largest droplet. The dynamic barriers introduced by floating VO-79 molecules promote the stalling mode, compared to the scenario where VO-79 molecules were restrained. Subtle differences also exist among those systems, in that both large (induced by more water molecules, see systems A16_W600 and A16_W900) and abundant small (induced by more asphaltene molecules, see systems A80_W600 and A160_W600) water droplets would help with growing mode in the equilibration stage of the simulation, where VO-79 molecules are restrained and only impose static barriers. The results reported here provide quantitative insights into the simultaneity of VO-79 (asphaltene) aggregation/adsorption and the coalescence of water droplets, and the automated method developed can also help to reveal nucleation mechanisms for nanoemulsion systems in broad areas. Altogether, this work is of both practical and fundamental importance. It can not only help to understand and further control emulsions in, for example, food, cosmetic, and petroleum industries, but also shed lights on how synergistic effects can be revealed and extracted in a quantitative way.

3. A method for calculating partial molar volume from a standard molecular trajectory

Published as **Hasan Imani Parashkooh**, Cuiying Jian, *An innovative method for calculating partial molar volume from a standard molecular trajectory*, Journal of Molecular Liquids 359 (2024) 123879.

Name	Contributions
Hasan Imani Parashkooh	Conceptualization Formal analysis Investigation Software Visualization Writing - original draft
Cuiying Jian	Funding acquisition Methodology Project administration Resources Writing - original draft, Review and editing Supervision

3.1. Introduction

The definition of partial molar properties pertains to the components of a mixture, where changes in a particular property occur as the quantity of that component is altered while keeping pressure and temperature constant [100]. These partial molar properties can be defined for a range of thermodynamic properties, including volume. Out of these properties, the partial molar volume (PMV) of mixture components has significant implications in various fields as reviewed below.

In biophysics, Le Chatelier's law has significance in explaining the pressure-induced denaturation of proteins, where the PMV plays a crucial role [101]. Upon applying pressure, proteins undergo structural transformation to a lower PMV state [102]. The PMV of proteins has also been utilized as an order parameter in protein folding, as demonstrated by Kitahara et al. [103] Furthermore, Kawama et al. [104] have investigated the PMV of proteins during the hydration process of 27 different proteins, which controls biophysical phenomena such as folding and ligand binding. In the food industry, the PMV of sugars is related to their sweetness response [105]. Additionally, Ivanov et al. [106] have demonstrated that investigating volume characteristics such as PMV in multi-component solutions can aid in studying the quality of interactions between solutes. Specifically, their work focuses on the impact of urea and tetramethylurea on the structure of water and formamide solvents. It was found that the effects of urea and tetramethylurea on the solvent structure are more pronounced at lower temperatures.

Examining molecular interactions through Partial Molar Volume (PMV) reveals valuable insights, particularly when isolating a solute molecule from interactions with other solute molecules. At infinite dilution, solute-solute interactions become negligible. The determination of partial molar volume (at infinite dilatation) becomes instrumental in providing clear information about various interactions in solutions, including ion-solvent, solute-solvent, and ion-ion interactions. For example, Kaur et al. [107] conducted experiments with propylene and hexylene glycols in methanol solutions of chlorhexidine. They obtained density value and calculated apparent molar volume (AMV) and then utilized linear regression to approximate the PMV value at infinite dilatation. Their findings reported positive PMV values at all temperatures and concentrations, indicating robust solute-solvent interactions. Similarly, Yan et al. [108] observed positive PMV (at infinite dilatation) values, when experimenting with a mixture of some amino acids and a pharmaceutically active ionic liquid, noting an increase with rising temperature. They

attributed this trend to the thickness of solvation layers enveloping the terminal groups of zwitterions at different temperatures.

As Liu et al describe in their work [109] the deviation between the volume of the solution from the ideal behavior was resultant from two mechanisms acting simultaneously. First, it is the interaction of solute and solvent molecules. Second, smaller molecules filling the void of larger molecules tend to shrink the volume. These two mechanisms can act in the same direction, or they can compete, and the result of competition can be investigated from the excess molar volume.

While PMV provides information about the volume change associated with adding one mole of a component to a solution, excess molar volume (EMV) V_m^E gives insights into the overall volume behavior of a mixture compared to an ideal solution. The link between them lies in understanding how the interactions at the molecular level, as reflected by PMV, contribute to the overall deviation from ideal behavior observed in V_m^E. Therefore, while PMV is a property of individual components in a solution, EMV is a property of the entire mixture, and it describes the variation in volume when two components are mixed to form a solution from the volume that would be expected if the mixture followed ideal behavior.

The EMV values are mainly affected by physical, chemical and structural factors. The physical factors consist of dispersion forces and other non–specific physical interactions that increase the value of EMV. On the other hand, the chemical and structural contributions like disruption of H–bonding, charge–transfer (donor–acceptor) complexes, and strong dipole–dipole interactions among the unlike molecules, mostly decrease EMV [110–112]. For example, Verma et al. [111] obtained volumetric properties of mesitylene and alkanol mixtures. They observed positive values for EMV across the whole composition range for mesitylene (1) + alkanol (2) mixtures, which they explained by the rupture of H-bonding in self-associated alkanol as well as dipole–dipole interactions in monomers and dimers of alkanol and due to a change in favorable orientation in mesitylene. On the other hand, in their other work [110], where the intermolecular interactions between the mixture of p–chlorotoluene and alkanol was studied, they reported that the structural contribution such as formation of H–bonding, charge–transfer (donor–acceptor) complexes, and strong dipole-dipole interactions among the unlike molecules, led to negative EMV. An interesting observation arises when connecting EMV with PMV. In these studies,

positive EMVs correlated with a decrease in the PMV of the non-alkanol component as its molar fraction increased. Conversely, negative EMVs were associated with an increase in the component's PMV with molar fraction. Notably, the PMV of alkanol showed a slight increase in both cases when its molar fraction decreased.

These findings highlight the capability of PMV to dissect interactions related to each component independently. In contrast, EMV is shown to reflect the overall collapse or expansion of the entire system. This observation underscores the distinct roles of PMV and EMV, emphasizing their complementary nature in elucidating the intricate interactions within complex mixtures.

In experiment, PMV is measured by numerical differentiation. That is, numerically differentiating the total volume while changing the amount of substance of interest and maintaining the temperature and the pressure of the system. For example, Cadena et al. conducted an experiment to measure the partial molar volume of CO_2 in two imidazolium-based ionic liquids [113]. The experimental setup employed was previously described in their published works [114]. To maintain a constant temperature throughout the experiment, they utilized a constant temperature bath, while the pressure was monitored with a capacitance manometer [115]. Several methods have been developed to predict PMV in numerical simulations. Hypothetically, the method of numerical differentiation, as described in experimental settings, can also be applied to simulations. However, it demands multiple independent systems to be simulated, and then the total volume to be measured and numerically differentiated with respect to the number of molecules of the specific component [116]. This prerequisite renders the method inefficient for numerical simulations. Moreover, the imprecision in measuring the total property has a substantial influence on the value of PMV [117]. The Kirkwood-Buff (KB) integrals can also be used to measure PMV and offer an advantage over other techniques in that they can be employed for closed systems with constant composition [118]. This is accomplished by implementing the KB integrals for a specific portion of the entire system [119]. An alternative set of techniques relies on a mathematical approach that centers on the partition function of a binary mixture in the isobaric-isothermal ensemble [120]. The partition function is a statistical function that characterizes the thermodynamic state of a system and can be represented as a sum of contributions from each

component in the mixture [121]. The methods used in the work of Widom [122] and Sinzingre et al. [120] are examples from this category.

Here we developed a novel method to calculate PMV of the components from the existing trajectories, outputted by molecular dynamics and as a demonstration, this novel method was tested on asphaltenes to illustrate its applications in understanding aggregation. Asphaltene has been chosen as the system to test due to its propensity for aggregation, which poses significant challenges in clean oil production [33]. Precipitation resulting from asphaltene aggregates can obstruct transportation pipelines [46]. To mitigate asphaltene aggregation, microwave heating [36] and chemical inhibitors [16] have been explored. Inhibitors are specifically designed to interact with asphaltenes and inhibit their aggregation [123,124]. Among these inhibitors, dodecylbenzene sulfonic acid (DBSA) has garnered significant attention due to its excellent performance as an asphaltene inhibitor. It has been demonstrated that even a low dosage of 5%wt of DBSA can completely reverse the aggregation effect [125] and improve the rheology of bitumen [126]. The complexity of asphaltene-inhibitor systems provides a platform to test the applicability of our novel method in providing physical insights. As presented in sections 2 and 3, the PMV serves as a valuable thermodynamic descriptor, providing insights into the molecular interactions and volumetric changes occurring within a system [106].

Current methodologies for determining or computing Partial Molar Volume (PMV) exhibit limitations, with some being constrained to specific molecular interactions, as exemplified by Widom's method [122]. Alternatively, computationally demanding techniques, such as numerical differentiation, pose challenges, especially in the context of post-processing legacy simulation datasets. Recognizing this gap in the literature, Josephson et al. [116] proposed a method that leverages molecular trajectories to derive PMV values for simulated systems. Building upon this foundation, our work introduces an efficient method for PMV calculation, which can also be applied to legacy simulation trajectories.

The remainder of this chapter is organized as following: In the following section, we introduce and discuss the methodology used in this work, along with its implementation. Additionally, we provide an overview of the properties of the systems that were analyzed in this study in section 3.3. These systems can be categorized into two groups: one group comprising a

solvent and asphaltene, and the other group including DBSAs as well. In section 3.4, we conclude the chapter, summarizing the findings and implications drawn from the study.

3.2. Method

3.2.1. Formulation

For a thermodynamical property like volume V, the parameter $\bar{V}_i$ can be defined as PMV of component i in a mixture. Equation (3-1) is the formal definition of the PMV [127], where N_i is the amount of substance i in the mixture.

$$\bar{V}_i = \left(\frac{\delta V}{\delta N_i}\right)_{T,P,N_{j\neq i}} \tag{3-1}$$

By integrating the total volume over the PMV of all components, one can write the integral form of the definition in equation (3-2).

$$V = \sum_i N_i \bar{V}_i \tag{3-2}$$

To determine the PMVs of an N-component system, one can create a system of equations using the total volume and the quantities of all components for n independent instances. Although this approach is mathematically sound, it is difficult to implement numerically due to the requirement that the instances be independent, or else the system becomes ill-conditioned. This approach shares similar challenges with the numerical differentiation method. In addition, simulating multiple independent systems incurs a significant computational cost. Furthermore, since PMVs depend on the system's composition, the obtained values would represent the mean of the sampled state points. This creates a trade-off, as noted in [116]: if the sampled instances are too distant, the average value becomes highly uncertain, while if they are too close, the system is prone to ill-conditioning.

To avoid the need of simulating multiple systems, we choose to generate n instances of equation (3-2) by utilizing the fluctuation of state parameters within a single simulation. If the ensemble allows for significant fluctuations in N_i and X, it is possible to identify independent state points for the system to build equation (3-2). The primary source of inaccuracy in this setup arises

from measuring the total property (i.e., volume), which is susceptible to numerical noise in the simulation.

One approach to handling this noise is to acquire more than n instances and use linear regression, as outlined in the work of Josephson et al. [116] In solving equation (3-2), if the values of X and N_is for more than n instances are available, multivariate linear regression can be applied to obtain $\bar{X}_i$ is, i.e., solving for the left-hand side vector in equation (3-3). Here, the subscript i indicates the relationship to component i of the mixture, and the superscript j shows the relationship to instance j. It is important to note that this method can only be employed under the assumption that the partial molar properties (PMV in this work) remain invariant when the composition changes.

$$\begin{bmatrix} N_1^1 & N_2^1 & \dots & N_n^1 \\ N_1^2 & N_2^2 & \dots & N_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ N_1^m & N_2^m & \dots & N_n^m \end{bmatrix} \begin{bmatrix} \bar{V}_1 \\ \vdots \\ \bar{V}_n \end{bmatrix} = \begin{bmatrix} V^1 \\ \vdots \\ V^m \end{bmatrix} \qquad (3\text{-}3)$$

3.2.2. Implementation

The molecular trajectories outputted by molecular dynamic simulations provide an opportunity to setup equation (3-3). That is, for simulation of a system containing n compounds, $m > n$ instances with the number of all components and the value of total property are needed. Here, the focus is on PMV and for this value the total volume should be measured at all instances. To alter the number of molecules one could simulate the system in ensembles where the number of molecules is not constant, although the control over the number of molecules is still limited. The work of Josephson et al. [116] introduced the method of linear regression for partial molar properties uses systems in which the composition changes by chemical reaction or phase transform. Our systems are not susceptible to chemical reactions; hence, there will be no variations in the number of molecules for the components. As a result, it is not feasible to sample instances throughout the simulation period that are linearly independent concerning N_is. This implies that the method utilized in the original paper cannot be directly implemented in this case.

Here, that method is modified so it can be applied on systems simulated in ensembles where system compositions are not changed. The sampling that produces the rows of the matrix and the left-hand side vector of equation (3-3) is done by inserting cubic control volumes inside the simulation domain. As shown in Figure 3-1, instead of using the whole simulation box volume as the total property, the volume of the sub-system (control volume) is used. The trajectories are compiled to count the number of molecules of each substance that is located inside the control volume. To introduce variance to the total property the dimension of the control volumes randomly changes between the values in the interval of [0.6,0.8] and [0.1,0.9] times the simulation box dimensions for the test systems and the validation systems, respectively. Since there are molecules on the edges of the control volumes that are partially inside them, non-integer is allowed for the values of N_i^j s, which for every molecule is determined by the fraction of its atoms that are inside the control volume to the total number of atoms of that molecule. In each snapshot, a certain number of identical control volumes is generated. The dimension of these control volumes is such that they overlap and cover the whole simulation box. Then, the number of molecules of all components in the system is counted in all control volumes. Based on those numbers, a certain number of control volumes are selected to provide their samples (more details in Figure A7.2.1 and section 3.4.1).

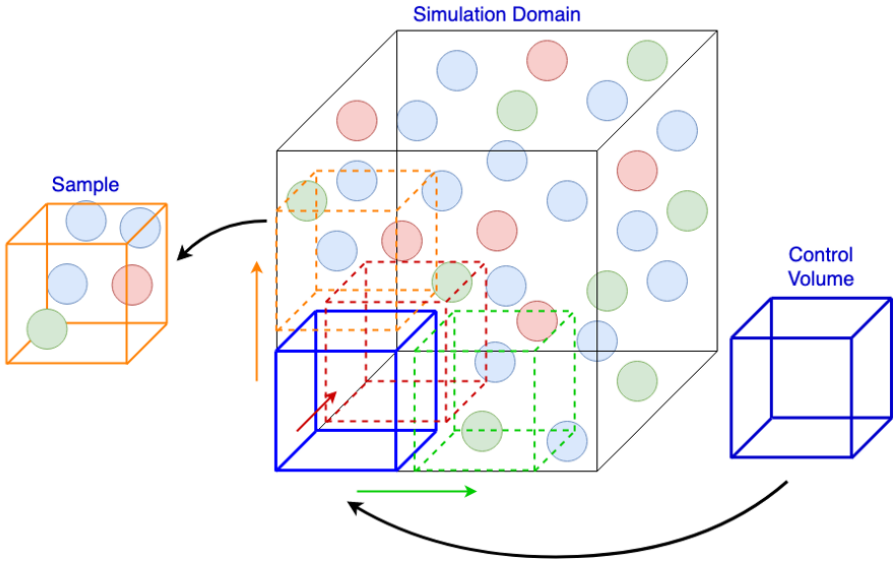

Figure 3-1 Sampling the data by control volumes for calculating partial molar volumes of all components. The control volumes used in the actual setup overlap with each other.

The final step is averaging the numbers provided by the samples for each of the components forming one row of matrix of equation (3-3) for the related snapshot. The volume of the identical control volumes is then written on the element of total property vector (right hand side of equation (3)) corresponding to that snapshot. This process is done for as many snapshots as needed to provide the m rows of the matrix. As said before the control volumes are identical in each instance but randomly change at different instances.

3.2.3. Validation Systems

To validate the method, it was applied on the 2 systems discussed in the work of Verma et al[111]. The properties of the systems are brought in Table 3-1. The systems are binary mixtures of mesitylene and isopropanol with 2 different compositions.

Table 3-1 The properties of simulated validation systems.

System	Mesitylene (1)	Isopropanol (2)	x_1	Simulation box length [nm]
Mesit_20	750	2868	0.21	7.892
Mesit_80	2000	500	0.8	8.066

In a study by Verma et al [111], the temperature-dependent density of both pure components and the mixture was measured across various compositions. Subsequently, excess molar volume (EMV) values were computed using equation (3-4). Following this, a Redlich-Kister polynomial was employed to fit the EMV values, and the resulting coefficients were utilized to derive partial molar volumes (PMV) through equations (3-5). Further details on these calculations are elucidated in Section 7.2.2 of the Appendix.

$$V_m^E = \frac{M_1 x_1 + M_2 x_2}{\rho} - \frac{M_1 x_1}{\rho_1} - \frac{M_2 x_2}{\rho_2} \tag{3-4}$$

$$\begin{cases} \bar{V}_1 = V_m^E + V_1 + x_2 (\frac{\partial V_m^E}{\partial x_1})_{P,T} \\ \bar{V}_2 = V_m^E + V_2 - x_1 (\frac{\partial V_m^E}{\partial x_1})_{P,T} \end{cases} \tag{3-5}$$

In equations (3-4) and (3-5), V_m^E is the excess and $\bar{V}_i$s are the partial molar volumes, while M_is represent molar masses, x_is denote mole fractions, V_1 and V_2 represent the molar volumes of the pure components.

3.2.4. Test Systems

As mentioned in the introduction, asphaltene plays a pivotal role in various challenges within the oil industry, such as pipeline blockages and the stabilization of water-in-oil emulsions. To gain insights into their properties, researchers employ model molecules as proxies for asphaltenes in numerical simulations, and since asphaltenes possess significant aromatic components, these model molecules are designed to reflect this characteristic [51–53]. Based on the positioning and distribution of aromatic parts, these model molecules can be categorized into two main types: the island or continental type, where a single polyaromatic core (PAC) is surrounded by various functional groups, and the archipelago type, where smaller polyaromatic groups are connected with aliphatic chains or similar groups [128,129]. The presence of polar aromatic parts [58], polar functional groups, and heteroatoms is identified as a crucial factor contributing to the self-aggregation of asphaltenes [17,51], mediated by interactions such as hydrogen bonding, acid-base interactions, and π-π stacking. Particularly, in the continental-type model molecules, π-π interactions play a significant role in attracting adjacent PACs of asphaltenes, facilitating their aggregation [59]. The stacking process through π-π interactions initiates when a few molecules come into contact with each other, aligning their polyaromatic cores in parallel configuration, forming a group of molecules known as parallelly stacked (PS) clusters. One of the common model molecules that is frequently used as a continental type asphaltene proxy is Violathrone-79 (VO-79) [70,71].

The systems studied here contain a mixture of VO-79 and solvents. To study the effect of inhibitors, 4-Dodecylbenzenesulfonic Acid (DBSA) was adopted here, since it has demonstrated exceptional performance and has drawn considerable attention from researchers [13,14,130]. Asphaltenes are represented using Violanthrone-79 (VO-79) and their chemical structures are shown in Figure 3-2.

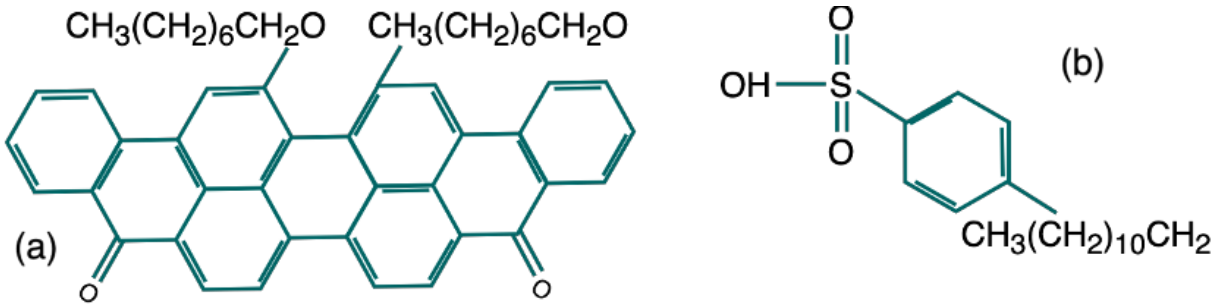

Figure 3-2 The chemical structures of violanthrone 79 (a) and 4-dodecylbenzenesulfonic acid (b).

Table 3-3 shows the composition of the systems studied. One of them is named A16_Wt and contains 50301 water and 16 V0-79 molecules. The inclusion of water and asphaltene in one of the systems is due to the fact that water is a polar solvent, while asphaltene is hydrophobic [131]. This difference in polarity can significantly affect the arrangement of water molecules around asphaltene molecules compared to organic solvents. Therefore, the next system A16_Wt contains 16 Vo-79 molecules solvated in n-Pentane. The components of A16_Pn, are set to make its output comparable to A16_Wt. The study can investigate how the presence of water affects the aggregation behavior of asphaltene molecules and more importantly how PMV changes with altering solvents.

The systems Hn0, Hn60, Hn120 and Hn180 contain 24 Vo-79 molecules and n-heptane as the solvent, and 0, 60, 120 and 180 DBSA molecules respectively. The composition of the systems that are simulated in this study has been selected with the aim of investigating the effect of varying inhibitor concentrations on the aggregation of asphaltene molecules in organic solvents. Specifically, the systems have the same amount of VO-79 but different amounts of DBSAs, allowing for a comparison of the effect of inhibitor concentration on asphaltene aggregation as well as PMV values.

Similarly, Pn0, Pn60, Pn120 and Pn180 were composed with 24 VO-79, n-pentane and 0, 60, 120 and 180 DBSA molecules. Incorporating different solvents in addition to varying inhibitor concentrations in the simulations can help to increase the confidence of the analysis by providing a more comprehensive understanding of how asphaltene aggregation is influenced by different solvent environments. By simulating the mix of VO-79 and DBSA in both n-heptane and n-

pentane, the study can investigate how variations in solvent properties and inhibitors affect the stability of asphaltene molecules through the PMV analysis.

Table 3-2 The properties of simulated test systems.

System	Violanthrone 79	n-Pentane	n-Heptane	Water	4-Dodecylbenzenesulfonic Acid	Simulation box length [nm]
A16_Wt	16	0	0	50301	0	11.607
A16_Pn	16	4800	0	0	0	9.811
Hn0	24	0	6637	0	0	11.807
Hn60	24	0	6360	0	60	11.722
Hn120	24	0	6107	0	120	11.645
Hn180	24	0	5883	0	180	11.582
Pn0	24	8420	0	0	0	11.813
Pn60	24	8146	0	0	60	11.750
Pn120	24	7903	0	0	120	11.709
Pn180	24	7638	0	0	180	11.655

Similarly, Pn0, Pn60, Pn120 and Pn180 were composed with 24 VO-79, n-pentane and 0, 60, 120 and 180 DBSA molecules. Incorporating different solvents in addition to varying inhibitor concentrations in the simulations can help to increase the confidence of the analysis by providing a more comprehensive understanding of how asphaltene aggregation is influenced by different solvent environments. By simulating the mix of VO-79 and DBSA in both n-heptane and n-pentane, the study can investigate how variations in solvent properties and inhibitors affect the stability of asphaltene molecules through the PMV analysis.

The VO-79 molecule's initial topology was created with PRODRG [74] and refined by manually adjusting partial charges and charge groups to align with GROMOS96 forcefield parameter set 53A6 [75]. This method, known for exploring polyaromatic core molecular dynamics [76], was chosen for compatibility. Topologies for n-pentane and n-heptane molecules were derived from dipalmitoylphosphatidylcholine in the GROMOS96 forcefield parameter set

53A6 using the gmx pdb2gmx routine in GROMACS [78], This method has been validated in our previous works [58,59,132].

For validation system components, topology files were obtained from the 54a7 force field [133] using the Automated Topology Builder [134]. This tool automatically refines topologies for small molecules. All simulations were conducted using the MD package GROMACS (version 2018.6).

For each system, an initial static energy minimization ensured a force below 1000.0 kJ/(mol·nm). The systems underwent a 1 ns relaxation period at 300 K (for test systems) and 298 K (for validation systems) under a pressure of 1 bar. This process effectively eliminated close contacts, ensuring thorough solvation, and stabilizing the system. All restraints were then removed, and NPT ensemble simulations ran for 30 ns for validation systems and 100 ns for the test system. Electrostatic interactions used the particle-mesh Ewald method [79], with a 1.4 nm cutoff for van der Waals interactions. Periodic boundary conditions, the SETTLE algorithm [80] for water molecule bond constraints, the LINCS algorithm [81] for general bond constraints, and a 2 fs time step were employed in all dynamic simulations.

3.2.5. Calculation Methods for Apparent Molar Volumes

To illustrate the unique implications of PMV, we also calculated apparent molar volume (AMV), which is another thermodynamic property used to describe the change in the volume of a solution when a certain amount of solute is added. It is defined as the change in the volume of the solution per unit change in the amount of solute added, while keeping the pressure and temperature constant. The AMV method considers the volume alteration caused by both the solute and the interactions between the solvent and solute. However, it attributes the entire volume change to the solute and assumes a constant molar volume for the solvent pre- and post-solvation [100]. Nonetheless, the AMV value serves as a valuable tool for investigating solution behaviors in comparison with PMV.

$$V_i^{\theta} = \frac{M_i}{\rho} - \frac{\rho - \rho_0}{m_i \rho \rho_0} \tag{3-6}$$

In equation (3-6), M_i and m_i are molar mass and molality of component i respectively. Also, ρ is the density of the solution and ρ_0 is the density of the pure solvent. The use of AMV has been used to evaluate solute-solvent interactions and aggregate formation in oil systems [135] as well as the critical aggregation concentration of liquid surfactant [C(12)mim]Br in different organic solvents [136] In this work, we compared AMV and PMV (details in section 3.4).

3.3. Results and Discussion

3.3.1. Validation of the Method

As mentioned in the method section, two systems were simulated to assess the accuracy of our approach. Due to the challenge of replicating identical systems during composition construction, the concentrations of components differ slightly from the corresponding experimental systems by Verma et al. [111]. We utilized the experimental results to interpolate the Partial Molar Volumes (PMVs) at the compositions of our validation systems. Figure 3-3 illustrates the fitted curve and interpolated points.

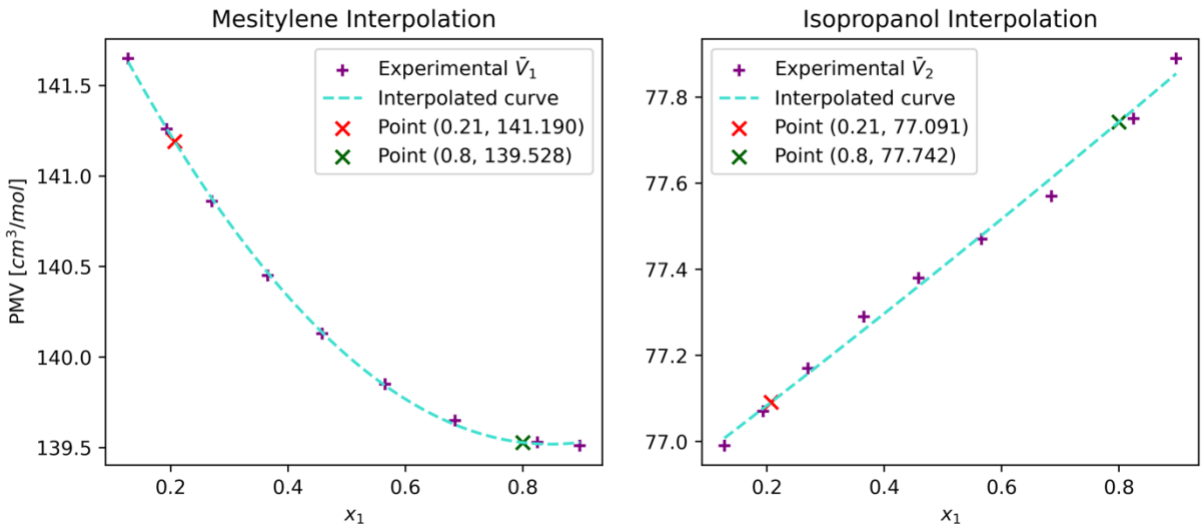

Figure 3-3 The interpolation of the experimental partial molar volumes at the compositions corresponding to the simulated validation systems. Purple markers represent experimental results, while red and green points denote concentrations in low and high mesitylene, respectively (Mesit_20, Mesit_80). The values for mesitylene are depicted in the left plot, while the right plot pertains to isopropanol.

In Figure 3-4, we present a comparative analysis of our calculation results for the validation systems listed in Table 3-1, juxtaposed with values interpolated from the experimental findings of

Verma et al. [111]. Our calculations involved the insertion of 64 control volumes at equidistant snapshots extracted from the simulation of binary mixtures. The dimensions of these control volumes were randomly selected between 0.1 and 0.9 of the simulation box dimensions at each sampled snapshot.

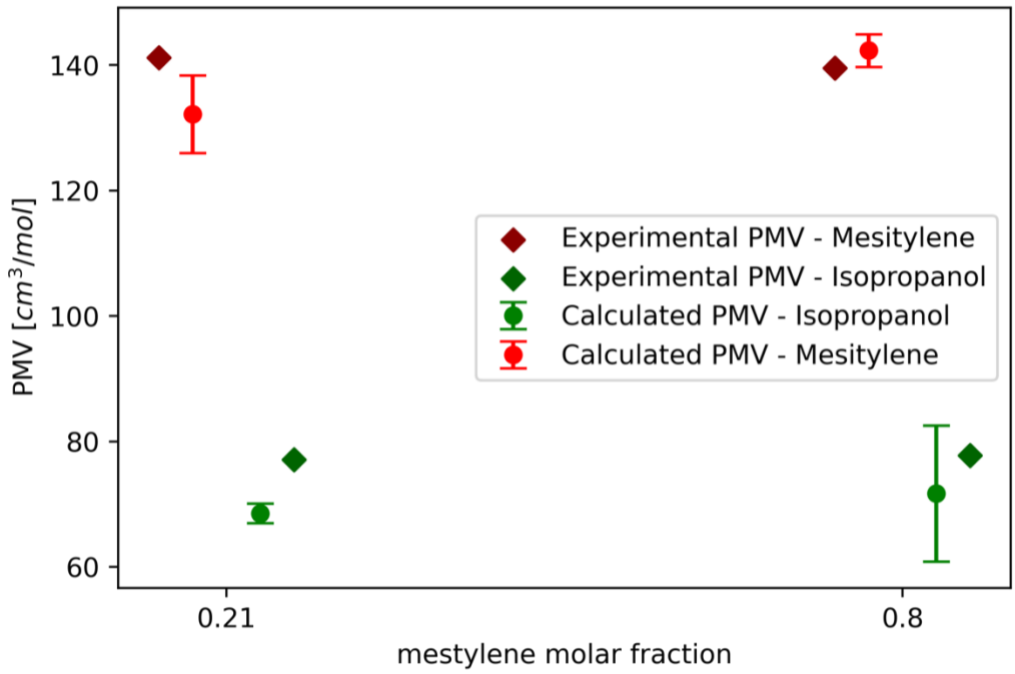

Figure 3-4 The validation of the method by comparing the calculated results with the experimental partial molar volumes of two binary mixtures. The diamonds are the interpolated values from the experimental results and the error bars show the mean and standard deviation of calculated values using different time samplings.

For each time step, 16 out of the 64 control volumes were strategically chosen and averaged to yield the quantities of molecules corresponding to that specific time step. The selection criteria were based on ensuring that the 16 control volumes closely matched the composition of the entire system. Time sampling was performed at intervals of 20, 50, 100, 200, and 250 ps. The resultant averages and standard deviations from these methods are depicted as circles and error bars in Figure 3-4. A notable observation is that, in each system, the Partial Molar Volume (PMV) of the component with a larger quantity is calculated with higher precision, as evidenced by the noticeably smaller error bars associated with it.

Table 3-4 presents a comparison between the values obtained through each time sampling method and those derived from the interpolation of experimental results. A notable observation

from the table is that the errors exhibit no correlation with the time sampling steps. This lack of correlation arises from the fact that, while decreasing the time sampling step provides more samples, it also results in snapshots that are closer in states to each other. Consequently, the rows in equation (3) become more linearly dependent. Another noteworthy observation is a significant error of 27.14% within the calculated values, specifically associated with the PMV of isopropanol in the Mesit_80 system using 200 ps time sampling steps. Its magnitude, nearly twice as large as the next highest error in all calculations, identifies it as an outlier. This underscores the importance of employing multiple time sampling methods and averaging their outcomes rather than relying on the results of a single time sampling. A further refinement post-averaging could involve eliminating results that deviate beyond a certain threshold from the average value.

In summarizing the data from Table 3-4, across all time samplings, we obtained a 6.33% error for mesitylene PMV in Mesit_20 and a 2.49% deviation in Mesit_80. For isopropanol, the average deviation was 11.09% in Mesit_20 and 13.96% in Mesit_80. Once again, it is worth emphasizing that these errors are more prominent for components with smaller quantities in each system. However, by excluding values that diverge by more than 20% from the average, the deviation for the PMV of isopropanol in Mesit_80 can be reduced to 9.06%.

Table 3-3 The value and relative error compared with the experimental results in different time sampling methods, for both mesitylene (1) and isopropanol (2)

Time Sampling Steps	**20 ps**		**40 ps**		**100 ps**		**200 ps**		**250 ps**	
	Mesit_20	Mesit_80	Mesit_20	Mesit_80	Mesit_20	Mesit_80	Mesit_20	Mesit_80	Mesit_20	Mesit_80
$\bar{V}_1$ $[cm^3/mol]$	139.486	143.561	122.038	137.883	137.794	141.634	130.110	145.823	131.435	142.794
Experimental Error of $\bar{V}_1$[%]	1.21	2.89	13.56	1.18	2.41	1.51	7.85	4.51	6.91	2.34
$\bar{V}_2$ $[cm^3/mol]$	66.673	70.001	71.093	89.756	67.068	75.165	69.048	56.639	68.808	66.927
Experimental Error of $\bar{V}_2$[%]	13.51	9.96	7.78	15.45	13.00	3.32	10.43	27.14	10.74	13.91

3.3.2. Optimizations Using the Test Systems

We calculated PMVs for solvents (results shown in Figure A7.2.2, A7.2.3). The results show that for all the systems the PMV value remains almost constant for the solvent. That is, the PMV of n-heptane is just above 0.145 $^{L}/_{mol}$ in Hn0, Hn60, Hn120 and Hn180. The PMV is between 0.115 and 0.12 $^{L}/_{mol}$ in systems where n-pentane is the solvent, and a little over 0.0185 $^{L}/_{mol}$ for water in system A16_Wt. These values align with those found in existing literature. For example, Bazile et al. [137] conducted measurements on the PMV of n-heptane when mixed with methane in a binary mixture. They reported comparable results, particularly when the methane mol percentage is below 60%. The molar volume measurements for n-pentane [138] and water [139] are also consistent with these values, indicating minimal alterations in the solvent configuration when compared to their pure states. For VO-79, extra care should be taken to select the control volume so that it properly represents the molecular composition of the system on that snapshot.

The criteria for selecting the control volumes are discussed. Also, as mentioned previously, the whole simulation domain is divided by a certain number of overlapping control volumes and a certain number of these control volumes is selected to provide sampling for equation (3-3). However, the selection is not arbitrary. All the systems of Table 3-1 contain asphaltene molecules, they self-aggregate and form clusters during the simulation time. In each time step, the trajectories are scanned to identify parallel stacking pairs based on distance and angles criteria proposed in the work of Jian et al. [59] For asphaltene molecules, an adjacency matrix is created in each time step based on the criteria, then the number of connected components (d_{cc}) of that adjacency matrix is computed using depth first search [140]. Finally in an instant containing d_{cc} connected components, the first $8d_{cc}$ control volumes with the largest number of asphaltene molecules are selected to provide the samples. Figure A7.2.1 shows the comparison between this selection method and randomly selected control volumes on the standard deviation of computed PMVs.

The systems contain molecules with different sizes and the number of small molecules like water and pentane in systems containing them, are larger than the big molecules in those systems. For this reason, number of molecules of each component would have different scales in the setup of equation (3-3). It is known that normalizing the features would help to get more accurate results in linear regression [141]. The normalizing is done by scaling every feature by its maximum value

of all instances. Therefore, every element of the matrix will be in [0,1]. After the calculations the PMV values are multiplied by the maximum values to satisfy equation (3-3).

The results are shown in Figure 3-5. As can be seen, they have large error bars. This stems from the inherent limitation of accommodating more than a few of these sizable molecules within control volumes, which can be as small as 0.216 of the simulation box volume. Despite accounting for non-integer values and factoring in the fractional contribution of atoms from these molecules within the control volume, the constraints on variability associated with their presence endure.

To further improve the precision of big molecules PMV, the PMV of different parts were disentangled. That is, dividing the larger molecules into different parts and approaching them as separate molecules. The insets in Figure 3-5 show the way that disentanglement of the PMV of different part of the large molecules is done. The parts were disentangled based on their polarity. In the setup of the Equation (3), one column is added for each of these molecules to account for the number of each part in the control volumes and the vector of partial molar properties would have a new element for each part. After computing the values of PMVs, the PMV of different parts are summed to give the PMV of the whole molecule. Figure 3-5 shows the comparison of the variance in asphaltene PMV computed with and without disentanglement. As can be seen, the variance is reduced. Four different calculations are done using 40, 50, 100 and 200 snapshots uniformly distributed in the last 10 ns interval. This reduction indicates a significant enhancement in the performance of our method.

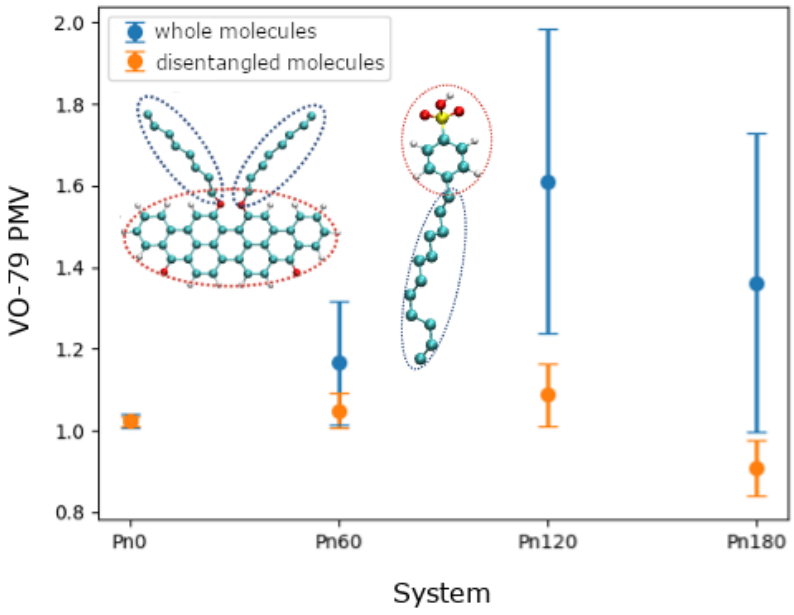

Figure 3-5 Comparison of the partial molar volume values obtained by disentangling molecule parts versus using the whole molecular structure. Partial molar volume calculations were performed treating the polar (red ovals) and non-polar (blue ovals) parts of each molecule as separate entities.

3.3.3. Application in Water and Pentane Systems

The PMV and AMV values of asphaltene are shown in Figure 3-6 for systems A16_Pn and A16_Wt. The snapshots are taken to calculate the PMV are from the time interval of 90-100 ns. Asphaltene PMV values calculated using different time samplings are shown in dark red circles. The number displayed within each circle represents the count of 50 ps time intervals that were skipped between successive snapshots during the time sampling. Figure 3-6 also displays the average and standard deviation of PMVs, represented by error bars, for all the time samplings. AMV of VO-79 as a solute in water and pentane was calculated for two different systems, A16_Pn and A16_Wt, respectively.

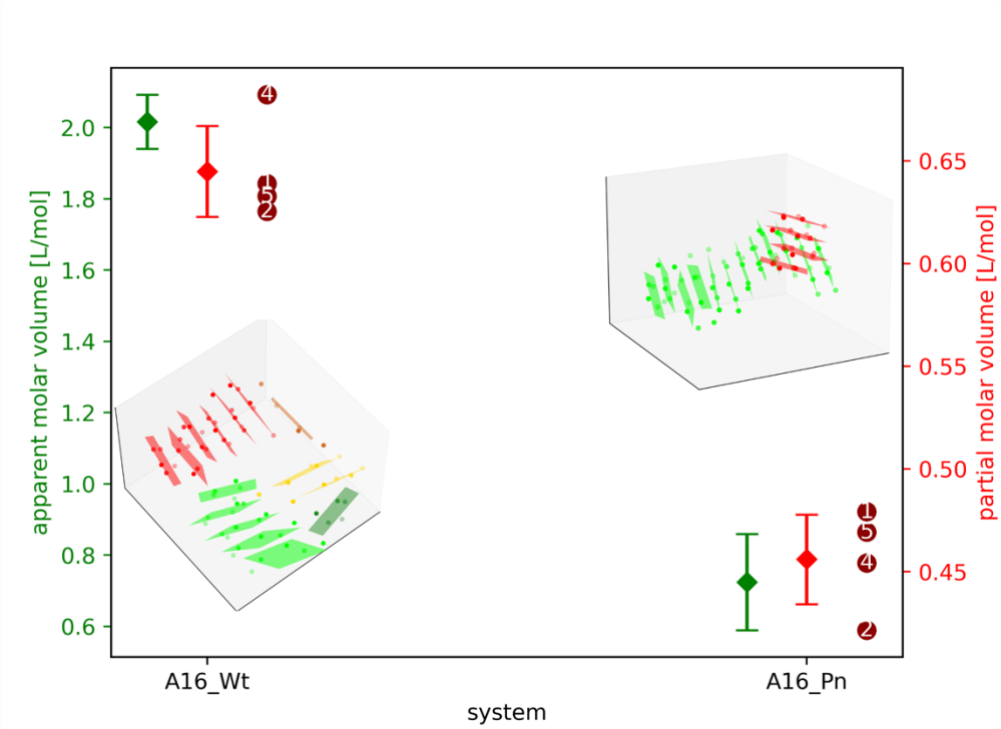

Figure 3-6 The values of partial molar volume and apparent molar volume for A16_Wt and A16_Pn calculated for the last 10 ns of the simulation. The dots show the average value of the different time-sampling method in partial molar volumes and average of the time interval in apparent molar volumes. The error bars show the standard deviation.

First, it is evident that the AMV values exhibit a greater magnitude compared to the PMV values. This outcome is anticipated since, in computing the AMV, the additional space generated by the mutual repulsion of solute and solvent molecules in both solvents is attributed to the solute. On the other hand, within the PMV approach, a portion of this surplus space is represented through an elevation in the solvent's PMV. It is important to note that due to the substantially higher number of solvent molecules, any increase in its PMV might not be noticeable, and it could potentially be overshadowed by the inherent numerical fluctuations stemming from the measurements of the overall volume.

Based on Figure 3-6, whether the changes in the volume of the mixture is completely attributed to the solute (AMV approach) or the contribution of the interactions on solvents volume is also considered (PMV approach), the asphaltenes are occupying more space in water compared to n-Pentane. This might be surprising, since asphaltenes are shown in the inset of Figure 3-6 to

form dense aggregate in water compared to in n-pentane. To understand why, we will take a look at aggregated structures.

As shown in the left inset of Figure 3-6, they curl into a spherical like structure. On the other hand, the asphaltenes in A16_Pn are mostly stacked in a parallel form, creating a couple of rod-like structures (right inset). This is, the PS clusters are more curled up in the A16_Wt system. To demonstrate this point using quantitative measures, the cosine between angles and the distance between pairs of asphaltene molecules in each cluster is plotted in Figure 3-7. The angle between molecules is the angle between the planes that passes through their polyaromatic core. The Figure 3-7 is plotted by obtaining the average distance and the cosine of the pairs that passed the criteria for parallel stacking during the 90-100 ns of the simulation. Specifically, pairs were considered if their average distance was below $0.75\ nm$ and the absolute value of their cosine value exceeded 0.9. As shown, in both systems as the distance between the molecules in the PS cluster increases, their planes deviate more from the parallel structure. However, in system A16_Wt, we have a PS cluster (shown in yellow) that has angled planes even in close distances. Another point that can be noted from Figure 3-7 is the large length of the biggest cluster in A16_Pn (furthest pairs are almost 4 nm apart) compared to A16_Wt. Combining Figures 3-4 and 3-5, it is evident that in A16_Wt, the smaller PS clusters are coming together, whereas in A16_Pn, the formation of the big rod-like cluster, has prevented the asphaltene molecules to gather in a non-parallel configuration. These remarks regarding the structure of aggregation are provided to exemplify the contrasting aggregation patterns in the two solvents. This distinction serves to emphasize the resilience of describing solute-solvent interactions using PMV values derived from the method, irrespective of alterations in the aggregation configuration within the control volumes. To be more precise, upon casual observation of the aggregation structures, one might anticipate the PMV in pentane to exceed that in water.

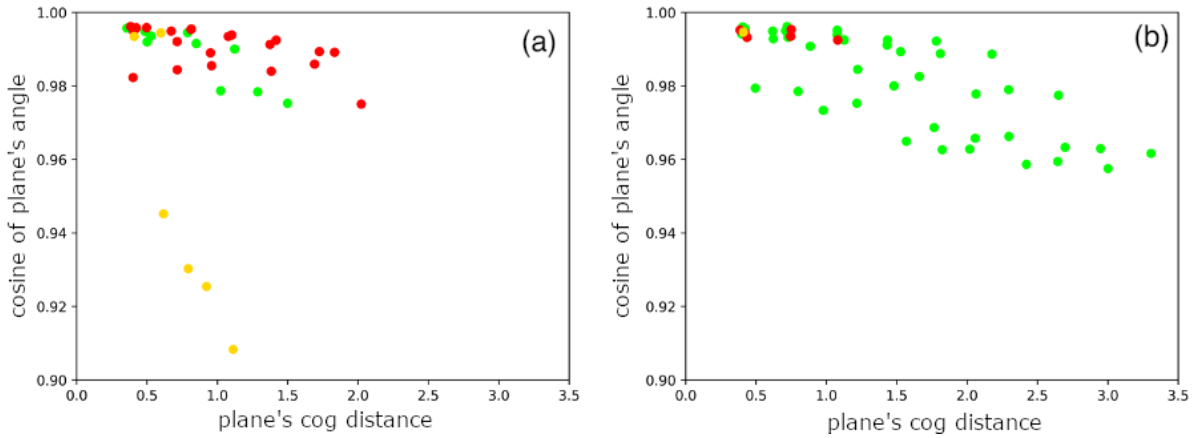

Figure 3-7 The cosine of the angle between pairs of asphaltene polyaromatic planes in the same parallelly stacked cluster for A16_Wt (a) and A16_Pn (b). The clusters are recognized by the criteria on the average distance and angle in the last 10 ns of the simulation. The pairs of each cluster are shown with the same color.

Due to hydrophobicity of asphaltene molecules, asphaltenes arrange to form a structure that minimizes the surface area of the interface with water. Their aliphatic side chains would be less exposed to water due to their hydrophobicity. However, what was seen in the snapshots (Figure A 7.2.4) was that some of the side chains were pointed to the outside of the spherical cluster. These side chains would repulse the water molecules and would cause more volume to be assigned to asphaltene molecules. However, this repulsion is not pushing the water molecules together. As it can be seen in Figure 7.2.2, the calculated PMV for water is larger than pure water's molar volume. Meaning more volume is assigned to water in A16_Wt compared to water's pure state. Also, in A16_Pn the PMV values calculated for n-Pentane are slightly higher than its molar volume. The rises observed in the solvent PMVs illustrate how solvents are absorbing a portion of the additional space generated by mixing, elucidating the reason for the higher magnitude of AMV values, which exclusively associate the surplus space with the solutes, as depicted in Figure 3-6.

Based on the preceding discussion, it becomes evident that PMVs offer additional insights into solute-solvent interactions. Although asphaltene aggregates exhibit greater severity in n-pentane, this does not necessarily suggest that these asphaltene aggregates would occupy less volume in water. In order to accommodate diverse interactions, the total volume expands when asphaltene is placed in water, which is demonstrated by the notably greater increase in asphaltene PMV in the aqueous medium.

3.3.4. Applications in Different DBSA systems

The main driving force behind island-type asphaltene aggregation in alkanes has been proposed to be the insolubility of the polyaromatic cores (PACs) of asphaltenes[142].

Figure 3-8 illustrates the normalized PMV values for systems containing 0 and 180 DBSAs, calculated during the 90-100 ns simulation period. The equation (3-7) is used for normalizing the PMV values, in which N_j and m_j are the number of molecules and the number of atoms per molecule for component j in the system:

$$\bar{V}_p = \frac{N_a m_a \bar{V}_a + N_d m_d \bar{V}_d}{N_a m_a + N_d m_d} \qquad (3\text{-}7)$$

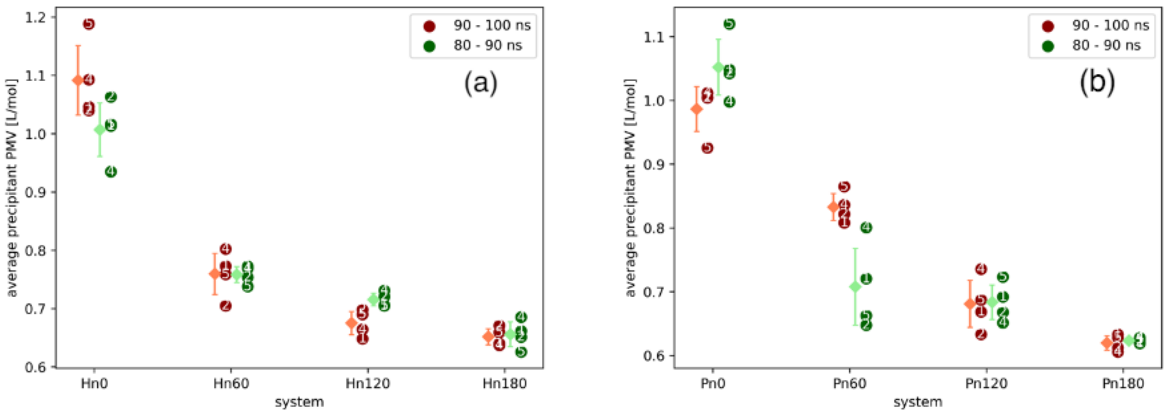

Figure 3-8 Partial molar volume of precipitant phase during the two time intervals for all heptane (a) and pentane (b) systems containing 4-dodecylbenzenesulfonic acid.

As evident from the results of Figure 3-8, PMVs of precipitants decrease as the number of DBSAs increases for systems employing either heptane or pentane as solvents, across both time intervals (80-90 ns and 90-100 ns). Creek et al. [143] demonstrated a linear correlation between the Hildebrand solubility parameter and the square root of PMV of the precipitant in alkanes mixtures. This is, a decrease in PMV would indicate a decrease in the solubility parameter of the precipitant. Painter et al. [144] found that asphaltene is miscible in solvents with a solubility parameter greater than 20 $(MPa)^{0.5}$, while alkane solvents such as propane, pentane, heptane, and octane fall below this threshold. In other words, reducing the solubility parameter of precipitant

(asphaltene) phase in these solvents would increase their solubility in alkanes. Therefore, our results of decreased PMVs indicate that increasing DBSA concentrations can reduce the Hildebrand solubility parameter of DBSA and asphaltene mixtures. Consequently, the solubility parameter approaches closer to that of the solvent, resulting in enhanced solubility of the precipitants and reduced susceptibility to fluctuations, as anticipated.

In order to quantify DBSA inhibiting effects, the VO-79 nanoclusters were identified in each case. Specifically, if two adjacent molecules had atoms within a distance of $0.5\,nm$ from each other, they were considered members of the same cluster. It should be noted that these nanoclusters are different from PS clusters where in addition to the distance criteria, there was also an angle criteria for the relative orientation of their polyaromatic cores. Figure 3-9 displays the number of VO-79 clusters during the simulation time interval of $20 - 100\,ns$. To determine the average size of the clusters in each case, we followed the methodology employed by Sedghi et al. [145] and calculated the $z-$average aggregation number, g_z, using equation (3-8). In this equation, n_i is the number of aggregates i that contain g_i asphaltenes.

$$g_z = \frac{\sum n_i g_i^3}{\sum n_i g_i^2} \tag{3-8}$$

Figure 3-9 demonstrates that the number of clusters in 0 DBSA systems is significantly lower, resulting in larger average cluster sizes compared to the systems containing DBSAs. Additionally, we observe that both Pn0 and Hn0 systems have reached equilibrium in terms of the number of VO-79 clusters before the simulation time of $t = 50\,ns$, whereas the 120 and 180 DBSA systems continue to exhibit considerable fluctuations in the number of clusters even at $t = 100\,ns$. On average it can be deducted that with increase in the number of DBSAs the number of clusters increases and their average size decrease. This can be interpreted as the increase of solubility of asphaltene nano aggregates as the number of DBSA in the system is increasing, consistent with the PMV implications discussed above. We note that using Hildebrand solubility parameter for solutions has limitations, especially for substances with polar interactions and hydrogen bonds [146]. Nevertheless, our observations are consistent with the work of Jiang et al. [13], which reported the inhibiting effects of DBSA (Dodecylbenzenesulfonic acid).

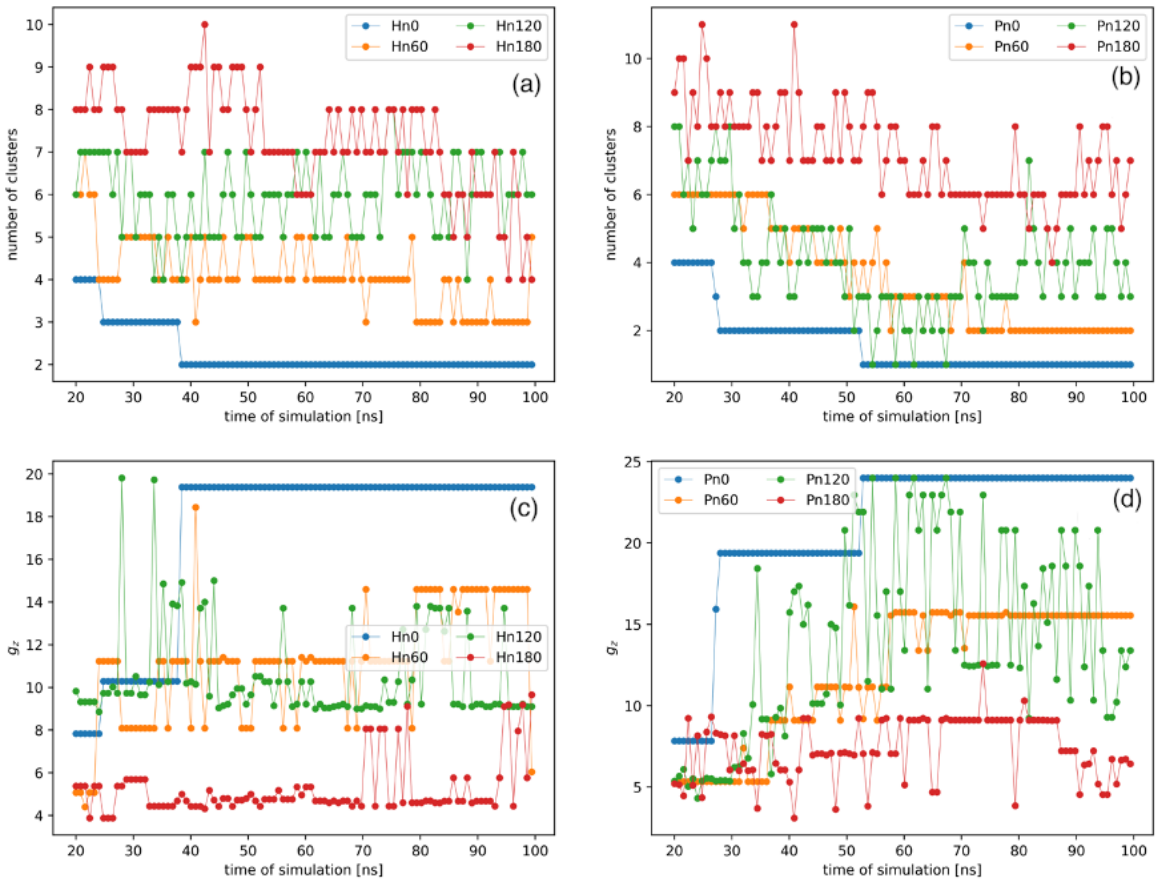

Figure 3-9 The number of the clusters ((a) for heptane and (b) for pentane) and their z-average size g_z ((c) for heptane and (d) for pentane) in time for systems containing 4-dodecylbenzenesulfonic acid.

To further understand the interaction mechanisms of DBSA, asphaltene, and solvents, Figure 3-10 presents the configurations of 0 and 180 DBSA systems for a specific instance. In this figure, distinct segments of the molecules are color-coded to facilitate visualization. The first notable point in Figure 3-10 is the contrasting structure of VO-79 clusters in 0 and 180 DBSA systems. In the case of 0 DBSA systems, the clusters are remarkably large. In Figure 3-10(c), all VO-79 molecules are observed to merge together, forming a single extensive cluster in the system Pn0. Conversely, Figures 3-10(b) and (d) demonstrate that VO-79 clusters in Hn180 and Pn180 systems are distinctly separated by the presence of DBSA molecules. These observations are consistent with previous PMV implications. Further examinations show that all configurations exhibit a consistent pattern where the non-polar regions of both VO-79 and DBSA are facing outward, away from the clusters and towards the solvents.

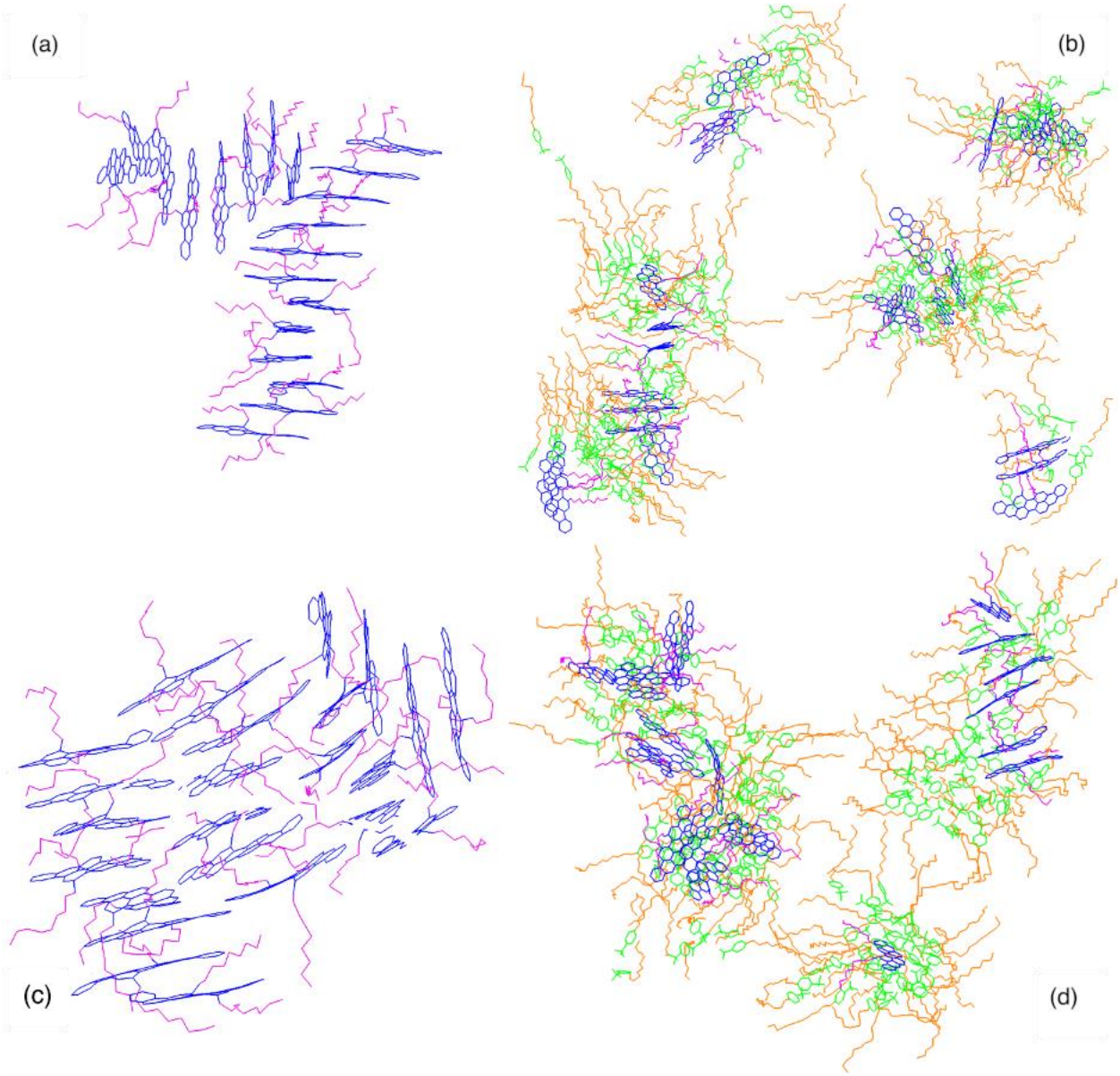

Figure 3-10 The snapshots taken from (a) Hn0 at $t = 96\ ns$, (b) Hn180 at $t = 89\ ns$, (c) Pn0 at $t = 96\ ns$ and (d) Pn180 at $t = 80\ ns$. The polar cores of the violanthrone 79s are shown in blue, their side chains are colored magenta and the 4-dodecylbenzenesulfonic acid' polar head, and non-polar tail are colored green and orange, respectively.

After comparing the averaged PMV values for precipitants between systems and reviewing the structure of asphaltene nanoclusters in them, the underlying interactions can be revealed. First, it should be noted that according to equation (3), when the number of the other component ($j \neq i$) molecules in the close proximity of the component i increases, $\bar{V}_i$ is decreased. In the systems examined in this section, the solvent is organic and non-polar. As a result, the solvent molecules

tend to approach the solutes (VO-79 and DBSA) from their non-polar sides. Consequently, when the non-polar sides of the solutes are exposed to the solvent, an increase in the number of solvent molecules in their vicinity leads to a decrease in their PMV, as mentioned earlier. Figure 3-5 illustrates that both VO-79 and DBSA possess a non-polar aliphatic chain, referred to as the "tail" for DBSA in this study. Due to the larger size ratio of the non-polar part to the polar part in DBSA compared to VO-79, an increase in the amount of DBSA while maintaining a constant number of VO-79 molecules would lead the solvent molecules to approach the solute flocculates more closely.

Figure 3-11 provides evidence supporting the observation that organic solvents tend to approach DBSAs from their non-polar side. The radial distribution function is plotted for solvent molecules around different segments of DBSA molecules in systems containing heptane and DBSA during two time intervals. The first interval represents the initial phase of the simulation when the DBSAs are not yet fully attached to VO-79 molecules. The other interval corresponds to the final stages of the simulation. As depicted in the figure, regardless of the simulation stage, the solvent molecules exhibit a higher density in the first layer surrounding the non-polar tail of the DBSA molecules. This can be attributed to both intermolecular interactions and the fact that the length of the tail allows it to come into contact with a greater number of solvent molecules compared to the polar head.

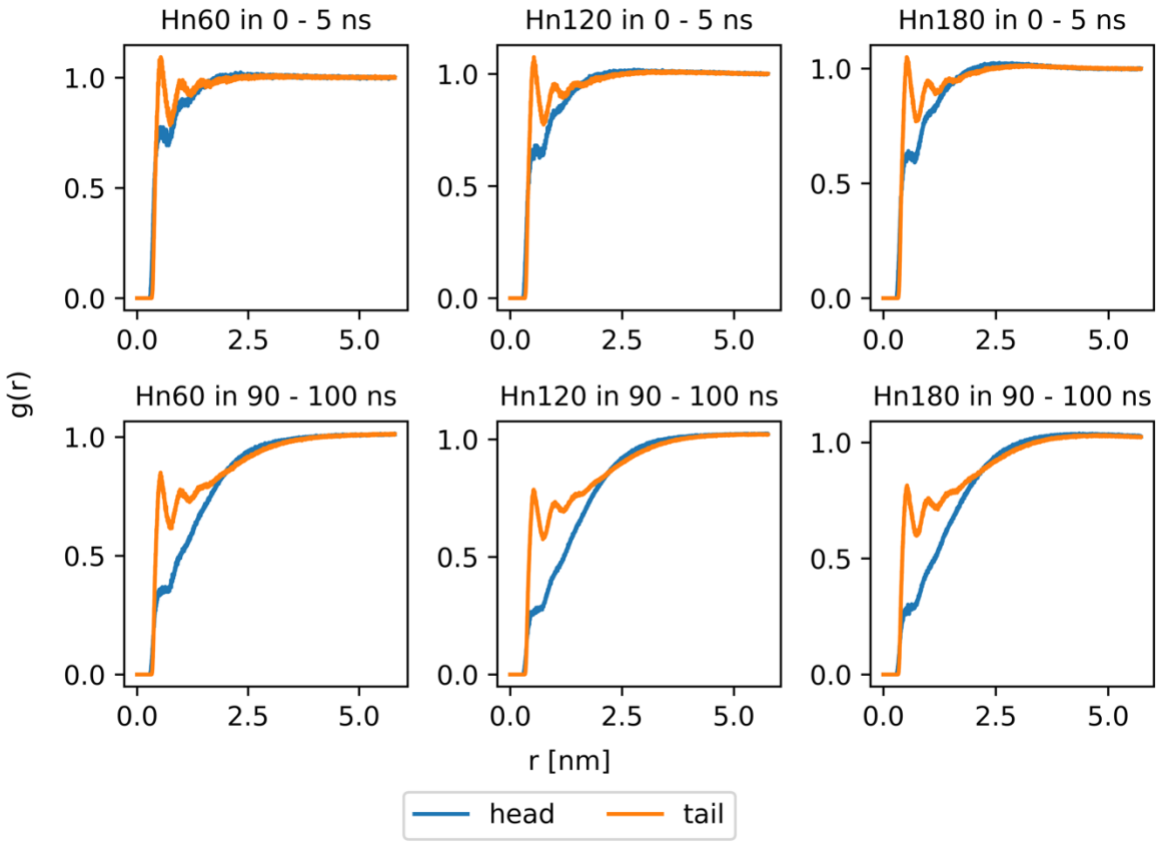

Figure 3-11 The radial distribution function of solvent molecules with reference to different parts of the 4-dodecylbenzenesulfonic acid molecules in Hn60, Hn120 and Hn180 systems, drawn in 2 different time intervals.

Regarding the interactions VO-79 and DBSA, in Section 7.2.5 of the Appendix, the RDFs of the head and tail segments of DBSA molecules are plotted with respect to the polar cores and non-polar side chains of VO-79 molecules. The RDFs are presented for the same time intervals examined in Figure 3-11. It is evident that, in all systems and for both time intervals, the RDFs associated with the polar parts of the solutes exhibit the highest peaks. This can be attributed to the stronger interactions between the polar head of DBSA and the polar core of VO-79 compared to the interactions involving the non-polar segments or the polar-non-polar interactions between these molecules. Consequently, it can be concluded that the polar heads of DBSAs tend to attach themselves to the polyaromatic cores (PACs) of the VO-79s, causing their tails to face away from the VO-79s and towards the solvent. This alignment further supports the higher presence of solvent

molecules surrounding the non-polar parts of the DBSAs in Figure 3-10. The RDF plots for pentane systems are also included in Figure A 7.2.5 of Section 7.2.5 of the Appendix.

From the above, the addition of DBSA to the systems containing VO-79 and solvent has led to a decrease in the PMV value, suggesting a decreased Hildebrand solubility parameter with increase in DBSA number. This is further confirmed by the decreased cluster size and increased number of clusters in the simulated systems with increasing the number of DBSA molecules. The presence of DBSA and its interaction with the VO-79 molecules alter the structural organization and behavior of the system. Consequently, the polyaromatic cores that were previously covered by each other in parallelly stacked clusters, as observed in 0 DBSA systems, have become exposed to DBSA molecules. These DBSA molecules are inclined to attach themselves to the polyaromatic cores using their polar heads. While the amount of DBSA used in the studied systems is increased, the VO-79 clusters become smaller in size and provide more surface area on their aromatic cores for most of the DBSA polar heads to attach in all systems. This arrangement attracts a high number of solvent molecules to their proximity, and a decrease is observed in the PMV value.

3.4.5. Limitations and Implications

First, it is important to acknowledge the theoretical underpinnings and limitations of using a system of summability relations, equation (3-3) to approximate partial molar volumes. As this method provides a numerical approximation and does not directly enforce the Gibbs-Duhem equation, which is a mathematical consequence of the formal definition of partial molar properties (equation (3-1)), the results should be interpreted as robust approximations rather than exact thermodynamic quantities derived from first principles.

Specifically, the accuracy of this regression-based approximation is vulnerable to several factors. For instance, the method is susceptible to numerical instability if the system of equations is ill-conditioned, which can occur if the compositions of the different instances used are not sufficiently distinct from one another. Also, any statistical or systematic errors in the input total volumes V, which are derived from molecular simulations, will inevitably propagate into the calculated partial molar volumes. Therefore, the precision of the results is fundamentally limited by the precision of the input simulation data. While this approach has proven effective for the systems studied herein, these potential vulnerabilities should be considered when applying the

method to other contexts, particularly those involving highly non-ideal mixtures or limited input data.

Secondly, it is noted that we used the assumption of invariance of PMVs when sampling a specific system. The underlying rationale is that examination was performed on equilibrated systems where solute molecules achieved a uniform PMV, and as such altering the position and dimension of control volumes in the system will not affect the value of asphaltene PMV. Contrarily, changing the composition of the whole system, for example by changing the number of DBSA molecules, would cause the asphaltene molecules to aggregate in a different structure and that would vary the PMV value.

Thirdly, it is important to note that the systems simulated in this study were relatively small and consisted of a limited number of large molecules, such as asphaltene. Consequently, the accuracy of the solvent PMV was observed to be higher in these cases, as the control volumes were capable of accommodating a larger quantity of these molecules. This suggests that in larger systems with a significant abundance of molecules like VO-79, the PMV results would likely be more accurate and exhibit reduced variability. It is crucial to acknowledge that the accuracy of the method is constrained by the size of the simulation systems.

Finally, it should be emphasized that the method developed in this study is that it only requires simulating a single system to calculate the PMVs. As elucidated in the discussion, this approach effectively captures a multitude of processes occurring within the systems, encompassing interactions between solvent-solute and solute-solute entities. It can capture the effect of both number of clusters and cluster sizes. Therefore, despite the limitation of system size, the method was demonstrated to provide a comprehensive insight into various processes and their interplay. Moreover, it is worth highlighting that the method employed to calculate the PMV and analyze the results holds potential for wider applications beyond the specific systems studied in this paper. For instance, it can be effectively employed to investigate other systems such as ionic liquids to describe the expansion of the liquid caused by the dissolved gas [147,148]. Similarly, this method can be applied to processes such as protein folding, where PMV values can be utilized to locate various folded structures on the free energy landscape [149]. This versatility further underscores the utility and relevance of the method beyond the examples presented in this current study.

3.4. Conclusions

This work demonstrated the utilization of linear regression for determining the PMV (partial molar volume) of multicomponent mixtures in dynamic simulations. The method presented here builds upon a previous approach used for systems that allowed for composition changes via phase transfer or reactive moves. However, in this study, we generalized the method to be applicable to ensembles where compositions are not changed along the simulation trajectory by incorporating control volumes within the simulation domain.

We validated our method by simulating two of the systems studied by Verma et al. [111] and applying our method on the outcome of the simulation. Through a comparison with experimental data, we achieved average errors as low as 2.49%, demonstrating the effectiveness of our method in generating reliable results. This method was then applied to examine the influence of solute-solvent interactions on PMV values by comparing the values of VO-79 PMV when mixed with a polar solvent (water) versus a non-polar one (n-pentane). Moreover, we successfully calculated the weighted average PMV of asphaltene and DBSA under various environmental conditions. It was shown that PMV has direct correlations with solute solubilities. A decreased PMV corresponds to a decreased solubility parameter, and increased solubilities. This is further confirmed by examining cluster size and solvent interactions. This exploration involved comparing PMV trends with quantitative measures characterizing the interaction between different segments of asphaltene and DBSA molecules and the size of their nanoaggregates. Increasing the concentration of DBSA and the interaction between their polar heads and the polyaromatic cores of model asphaltenes result in the formation of smaller asphaltene nanoaggregates. These nanoaggregates are enveloped by DBSA molecules, and due to the interaction between the non-polar tails of DBSA and the solvents, they become more readily solvated.

Overall, this study presents a simple yet efficient method to investigate volumetric thermodynamic parameters of solutions. Its applications demonstrated by asphaltenes suggest potential significance in the study of other colloidal systems across diverse fields, including molecular biology.

4. Predictions of Adsorption Energy of Aromatic Adsorbates Using Equivariant Networks

4.1. Introduction

The depletion of lighter conventional crude oils has led to the increased use of heavy, extra-heavy, and other unconventional crudes as the primary feedstock for refining [150]. This shift has continuously brought asphaltene—the heaviest and most surface-reactive non-volatile petroleum fraction—into the research spotlight. Asphaltenes present significant challenges in oil pipelines and processing equipment due to their propensity to aggregate and adhere to surfaces, leading to fouling and flow disruptions.

Traditional methods for removing asphaltene deposits, such as solvents, surfactants, and mechanical treatments, are costly and often provide only temporary relief, as asphaltenes tend to redeposit [151]. Nanoparticles (NPs) have emerged as a promising alternative in the oil industry, offering unique properties such as high surface area-to-volume ratios, functionalizable surfaces, and effective mobility in porous media. These features enable NPs to adsorb and disperse asphaltenes efficiently, potentially providing a long-term solution to deposition challenges. Additionally, their high adsorption capacity and catalytic potential make NPs valuable for enhancing heavy oil recovery [152]. Developing effective nanoparticle inhibitors for asphaltene aggregation relies on precise modeling of adsorption interactions, where accurate predictions of adsorption energies for heavy aromatic compounds are critical.

Beyond oil industries, scalable and accurate methods for predicting adsorption energies have also become a growing focus for hydrogen energy storage (HES) processes [28,153–155]. HES addresses a key challenge in the transition to carbon-neutral energy: storing energy generated at times that do not align with fluctuating demand. Despite its potential for scalability, long-term storage, and portability, HES adoption has been limited by low efficiency and high costs [154]. Central to the HES process are electrocatalysts, which enhance the efficiency of electrochemical reactions such as water splitting and hydrogen oxidation. Current state-of-the-art electrocatalysts, typically noble metals like iridium and platinum, are highly effective but prohibitively expensive [156]. In the search for improved electrocatalysts, understanding the relationship between a material's structure and its catalytic activity is vital. While traditional descriptors, such as M-OH

bond strength or the heat of metal oxide formation, provide useful insights, they primarily reflect bulk properties rather than surface phenomena, where catalytic reactions occur [156]. To reveal surface characteristics, adsorption energy has emerged as a more precise descriptor, linking structural properties to catalytic performance [157].

Calculating adsorption energies traditionally relies on Density Functional Theory (DFT), which provides high accuracy but is computationally expensive, scaling poorly with system size. To overcome these challenges, machine learning (ML) models have been developed and trained on first-principles computation results, enabling accurate adsorption energy predictions at significantly reduced computational costs. Graph Neural Networks (GNNs) have been particularly successful due to their ability to capture both local and global chemical environments effectively.

Despite significant advances, particularly with equivariant neural networks that incorporate physical symmetries into their architectures, current ML models are limited in their application to larger, more complex adsorbates like asphaltenes (heavy aromatic molecules of 60 to 200 atoms). Most existing models are trained on datasets such as the Open Catalyst dataset [29] and Open Direct Air Capture dataset [158], which primarily include small adsorbates of 2 to 50 atoms. These datasets lack sufficient representation of larger functional groups, resulting in diminished predictive performance when models are extended to larger molecules prevalent in practical scenarios. While recent studies have sought to address the modeling of adsorption energies for larger molecules, they often fail to leverage the full potential of advanced equivariant neural networks. These architectures excel in capturing intricate interatomic interactions and adhering to conservation laws, offering significant improvements in prediction accuracy and efficiency.

To advance adsorption energy predictions of large adsorbates such as asphaltenes, this theme focuses on the fine-tuning advanced equivariant models using datasets with larger molecules. Fine-tuning refers to the process of taking a pre-trained machine learning model and adjusting its weights slightly by training it further on a new, often smaller dataset. The pre-trained model has already learned general features from a large dataset, so fine-tuning adapts it for a specific, related task rather than training from scratch so while prior knowledge is leveraged, the training is accelerated. Pre-trained models, utilizing datasets with diverse functional groups, will be adapted to predict adsorption energies for complex adsorbates.

The remaining of this chapter is organized as follows: Section 4.2 provides the necessary background, including a detailed explanation of adsorption energy (Section 4.2.1), an overview of equivariant GNNs for adsorption energy prediction (Section 4.2.2), an introduction to the DFT datasets involved in pretraining and fine-tuning (Section 4.2.3), and a discussion of the challenges associated with applying GNNs to large molecules (Section 4.2.4). Section 4.3 presents the methods, results, and discussions, beginning with the dataset processing strategies (Section 4.3.1) and followed by an in-depth description of the fine-tuning approach, including relevant training details and performance insights (Section 4.3.2). Conclusions are given in Section 4.4.

4.2. Background

4.2.1. Adsorption Energy

Adsorption energy (E_{ads}) quantifies the strength of the interaction between an adsorbate molecule and a substrate surface. This interaction typically results from attractive forces such as van der Waals forces, electrostatic interactions, or chemical bonding. One way to obtain E_{ads} is using:

$$E_{ads} = E_{tot} - E_S - E_A \tag{4-1}$$

Here, E_{tot} is the energy of the adsorption system and E_S and E_A are the energy of the substrate by itself and the gas-phase molecule energy of the adsorbate, respectively [159]. When an adsorbate interacts with a solid surface, it forms a specific configuration based on the interactions with the substrate. Accurately determining adsorption energy requires identifying the global minimum energy configuration—the arrangement that minimizes the system's total energy across all possible adsorbate placements and orientations.

Traditionally, adsorption energies are calculated using first-principles methods such as DFT [160]. Identifying the global minimum often involves sampling various adsorbate-surface configurations. Thus, selecting optimal configurations has relied on expert intuition or heuristic methods leveraging surface symmetry. However, “brute-intuition” approaches depend on manually selecting plausible starting geometries and are inherently biased by the user's assumptions [161]. Meanwhile, "brute-force" methods aim to exhaustively screen all

configurations [162] but are computationally prohibitive and still require initial assumptions about binding sites and molecular conformations, introducing further bias [163]. As such, while these strategies have been effective in descriptor-based studies [28], they lack scalability for complex systems with numerous local energy minima. Furthermore, large adsorbates often present additional challenges, including flexible internal structures, multidentate binding geometries, and diverse interaction sites, particularly on defected or amorphous surfaces. To address these challenges, this theme adopts a graph-based ML strategy to predict adsorption energies.

4.2.2. Equivariant GNNS for Adsorption Energy Prediction

Equivariant GNNs are designed to predict adsorption energies by capturing the intricate interactions between adsorbates and surfaces while adhering to the physical symmetries of these systems [164]. In adsorption studies, the adsorbate-surface system can be represented as a graph using GNN, where nodes correspond to atoms, and edges represent bonds or interactions between them. To accurately represent each atom, node embeddings can be enriched with various chemical descriptors. While basic options include atomic numbers or one-hot encodings to differentiate chemical elements [159], embeddings can also incorporate additional atomic properties like electronegativity, atomic radius, and ionization energy. Structural characteristics, such as hybridization state, formal charge, and valence electrons, could potentially add further detail to node and edge representations [165]. Furthermore, self-supervised [166] and unsupervised [167] learning approaches aim for learnable feature representations, instead of relying on manually constructed material descriptors.

Recently, a significant advancement in GNNs is the incorporation of physical symmetries through equivariant neural networks, where equivariance refers to the property of a function that transforms predictably under certain input transformations [164,168]. For physical systems, essential symmetries include translational, rotational, and reflection invariance, which ensure that properties like energy remain unchanged irrespective of a system's orientation or position. Equivariant networks integrate these symmetries into the model's architecture, ensuring that predictions remain physically consistent and adhere to fundamental laws [168]. Equivariant neural networks for 3D systems, such as SE(3)/E(3)-equivariant networks, leverage transformations represented by tensor fields and spherical harmonics, which project spatial information in a way

that respects these symmetries (for details, see Appendix 7.3.1). Despite the advantages of equivariant GNNs, certain challenges exist when applying them to predicting the adsorption energies of heavy aromatic molecules.

In ML-driven energy and force predictions, different tasks are designed to approximate computational chemistry simulations efficiently. Among them, Structure to Energy and Forces (S2EF) [29] focuses on predicting the total energy of an atomic configuration along with the per-atom forces acting on each atom. This task is widely used to accelerate molecular dynamics simulations and structure relaxations. This work specifically focuses on S2EF (for other available tasks, see Appendix 7.3.1), given its foundational role in enabling faster and more accurate force evaluations, which are critical for large-scale molecular simulations.

4.2.3. DFT Datasets

In this subsection, we introduce the key DFT datasets utilized in this work, covering their methodologies for calculating adsorption energies and forces, as well as the structures, adsorbates, and substrates involved. Figure 4-1 shows an overview of the datasets and materials in each of them. Next, a brief introduction for each of the datasets is presented.

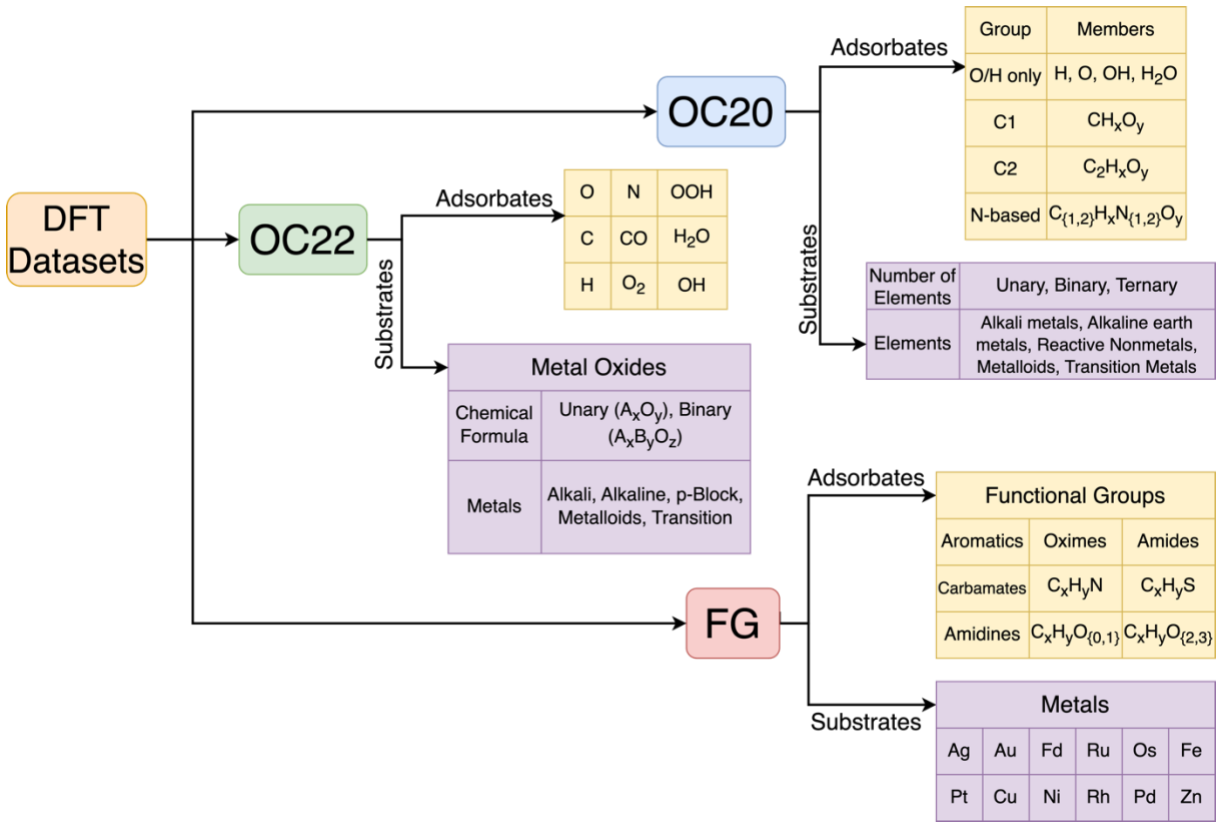

Figure 4-1 Overview of the DFT datasets used in this work and their corresponding materials. The models are pre-trained on OC20 and OC22, while FG serves as the fine-tuning dataset.

a) Open Catalyst Datasets (OC20 and OC22)

The OC20 dataset [29] comprises over 1.2 million DFT relaxations, totaling approximately 250 million single-point calculations, across a diverse range of materials, surfaces, and adsorbates. This dataset focuses on small adsorbates, including C1 and C2 compounds (containing 1 and 2 carbon atoms, respectively), as well as nitrogen and oxygen-containing intermediates, adsorbed onto various catalyst surfaces. Recognizing the limited representation of oxides in OC20, the same authors introduced the Open Catalyst 2022 (OC22) dataset[30]. OC22 addresses this gap by incorporating 62,331 DFT relaxations (~9.85 million single-point calculations) spanning a wide range of oxide materials, surface coverages, and adsorbates.

The dataset generation process applied three critical filters during DFT relaxation: excluding desorption events (non-binding adsorbates), dissociated adsorbates (the breaking of an adsorbate into different atoms or fragments), and systems with significant adsorbate-induced surface

distortions[28]. Crucially, *adsorbates* in this context encompass not only intact molecules but also molecular fragments (e.g., functional groups, radicals) that adsorb via distinct binding sites. However, dissociative adsorption (where fragments form *during* relaxation) was explicitly excluded to preserve the integrity of single-molecule adsorption energy calculations. By retaining systems where pre-defined fragments or molecules adsorb intact (without further dissociation), the dataset captures realistic adsorption mechanisms while avoiding misleading energy artifacts. This approach ensures that models trained on the data learn transferable relationships between adsorbate structure (including fragment geometries and binding-site variability) and adsorption energy. For aromatic compounds, which often adsorb as intact ring systems or functionalized derivatives, the constraints align with their typical non-dissociative binding behavior, enabling robust generalization of the GNN model to aromatic adsorption phenomena.

b) FG Dataset

The Functional Groups (FG) dataset [159] provides DFT-calculated adsorption energies and forces for a selection of functional groups adsorbed onto various substrates. It includes 207 organic molecules adsorbed onto 14 transition metals, featuring diverse functional groups and aromatic structures with heteroatoms. This dataset includes detailed information on the structural configurations of both the adsorbates and substrates.

The molecules included in the FG dataset cover key functional groups in organic chemistry, featuring nitrogen, oxygen, and sulfur heteroatoms. These functional groups are divided into several categories to reflect the diversity of chemical interactions relevant to surface adsorption. The categories include non-cyclic hydrocarbons, O-functionalized compounds (such as alcohols, ketones, aldehydes, ethers, carboxylic acids, and carbonates), and N-functionalized compounds (amines, imines, and amidines). Additionally, S-functionalized compounds, such as thiols, thioaldehydes, and thioketones, are included, as well as N- and O-functionalized combinations like amides, oximes, and carbamate esters. The dataset also contains aromatic molecules with up to two rings, which may also include heteroatoms. This dataset serves as a comprehensive resource for fine-tuning and testing ML models, particularly in predicting adsorption energies for large and diverse adsorbates.

4.2.4. Challenges in Applying GNNs to Predict Adsorption Energies for Large Molecules

Most existing GNNs for adsorption energy prediction are pretrained on datasets containing small adsorbates, such as the OC20. These models are optimized for relatively simple systems with fewer atoms and less structural complexity. Applying them directly to large molecules, such as those in the FG-dataset, introduces significant challenges.

One major issue is the mismatch in scale and complexity. Large adsorbates often contain flexible bonds, diverse functional groups, and intricate interaction sites, significantly increasing the number of possible adsorption configurations. This added complexity can overwhelm existing models, resulting in poor predictions. Additionally, the architectural capacity of pretrained GNNs may be insufficient to process such extensive input sizes effectively. Without adaptations, these models are unable to capture the detailed interactions present in larger molecules, leading to diminished performance. Computationally, processing the full structures of large adsorbates requires significantly more memory and processing time, making direct application infeasible for high-throughput studies or large-scale datasets. In other words, the challenges associated with applying GNNs to predict adsorption energies for large molecules stem from the increased scale and complexity of the systems, as well as the limitations of models pretrained on small adsorbates. Preprocessing the FG-dataset to focus on critical interactions and segregating it into molecular families for specialized fine-tuning offer practical solutions. These strategies reduce computational demands and enhance model accuracy, enabling more effective predictions for complex adsorption systems. In Section 4.3, we will present in detail on the development of our approaches for predicting adsorption energies of large molecules.

4.3. Methods, Results, and Discussions

4.3.1. Processing the Dataset

The primary dataset in focus is the FG-dataset. This section outlines the methods used to process the structures (adsorbate + substrate pairs) within the FG-dataset.

a) Ensemble Extraction

The FG-dataset entries each consist of a relaxed network of substrate atoms paired with a single adsorbate molecule, with each substrate built from 48 atoms. Although the aim is to eventually train models that can handle larger structures, including every adsorbate atom as input is both unnecessary and computationally expensive during fine-tuning and inference. Following an approach similar to Pablo-Garcia et al. [159], the dataset entries were streamlined to retain only the adsorbate atoms and the substrate atoms deemed essential.

As shown in Figure 4-2, the process starts by generating graph representation for the given structure. To generate the graph representation for each entry, Pablo-Garcia et al. [159] applied the Voronoi tessellation method. This technique partitions the three-dimensional space by assigning every atom a region that comprises all points closer to it than to any other atom. In the resulting graph, atoms are represented as nodes, and a connection is drawn between two nodes if their corresponding atoms share a Voronoi facet and if the distance between them is less than the sum of their covalent radii plus an additional tolerance. In this work, the covalent radii values from the original FG-dataset publication were used, supplemented with a tolerance of 0.5 Å to better detect metal–adsorbate connections. Next, rather than constructing a full graph with nodes and edges for all atoms, metal atoms that were not directly connected to the adsorbate were removed from the graph. These omitted connections are represented as dotted edges in Figure 4-2. Finally, the refined structure is obtained by retaining only the atoms corresponding to the nodes of the reduced graph representation, while discarding the remaining atoms.

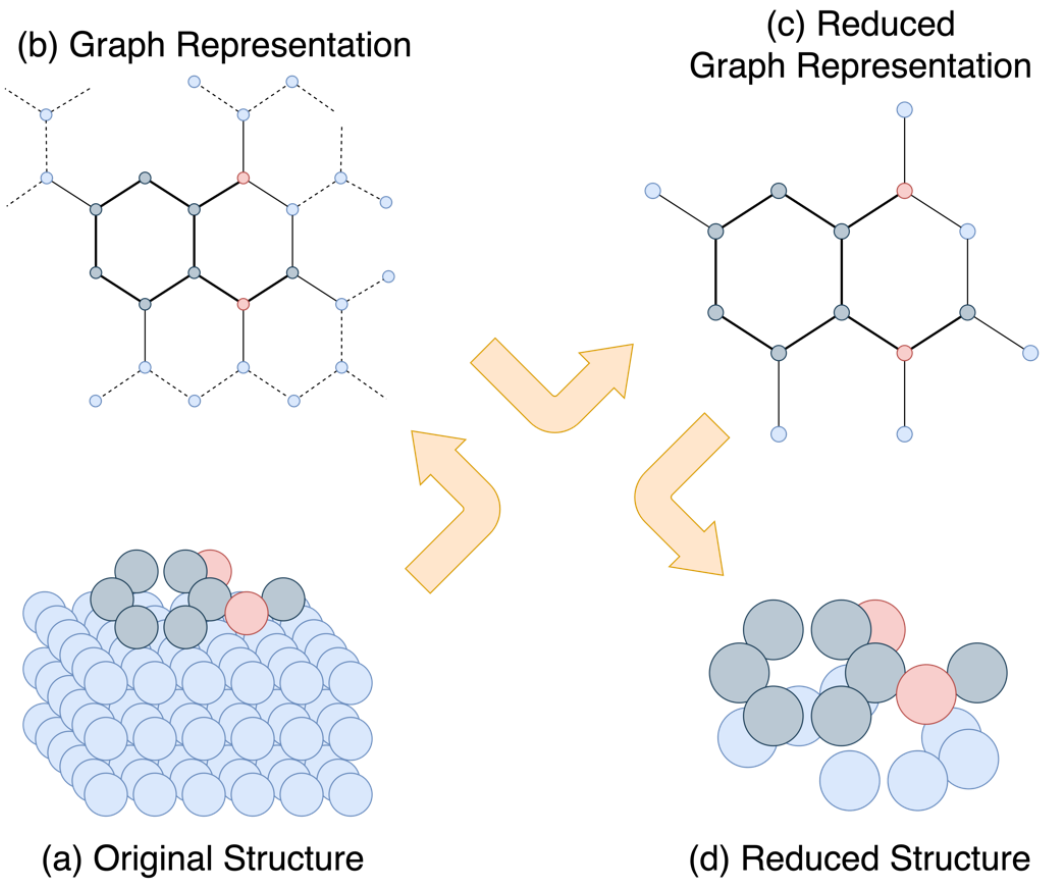

Figure 4-2 Ensemble extraction on the entries of FG dataset

b) Segregating the Functional Groups

The FG-dataset comprises a diverse collection of adsorbate–substrate pairs, with each pair represented in one or two structural configurations, culminating in a total of 6,866 entries. These adsorbates originate from nine distinct families of organic molecules, encompassing a diverse range of chemical functionalities (see the summary in Figure 4-1). Unfortunately, in the publicly available version of the FG-dataset [159], these nine categories are not explicitly segregated. From a ML standpoint, there are advantages to fine-tuning a model exclusively on structurally and chemically similar adsorbates, since such a dataset distribution can align well with transfer learning principles, improving performance on molecules akin to those in the target category. For instance, one might hypothesize that training a model specifically on aromatic compounds could yield better predictions for a new aromatic molecule compared to training on a broader mix of functional groups.

However, there is a significant trade-off. While filtering for specific families (e.g., segregating aromatic compounds from the rest) can concentrate the training data on structurally relevant examples, it inherently reduces the overall training set size. In many ML contexts, especially in deep learning, a larger dataset, even if somewhat heterogeneous, can often outperform a smaller but more homogeneous subset. This occurs partly because the model can still learn to generalize across a range of similar chemical interactions present in all FG-dataset entries. Consequently, there is an inherent balancing act: Refining the dataset to closely match the desired chemical family versus retaining the breadth and volume of training samples to avoid overfitting or insufficient coverage of relevant chemical space.

In our work, we explored this balance by applying certain composition-based criteria, like the number of carbon atoms and the ratio of carbon to hydrogen atoms, to distinguish aromatic compounds from the broader FG-dataset, calling it FG-aromatics with 1,140 entries. The aim is to assess whether restricting the training set to those compounds would yield better model performance, compared to fine-tuning on larger and more diverse dataset. After applying the ensemble extraction on FG-dataset and FG-aromatics, we obtain **Extracted FG** and **Segregated aromatics,** respectively. These two datasets are the ones used for fine-tuning the model in this work.

4.3.2. Fine-Tuning Details

a) The Pretrained Model

The GemNet-OC-S2EFS-OC20+OC22 configuration builds upon the GemNet-OC [31] model by training on a combined dataset consisting of 133,934,018 frames from OC20 and 9,854,504 single-point calculations from OC22. Leveraging the Geometric Message Passing Neural Network (GemNet) architecture [169], GemNet-OC represents atomic systems as graphs, where atoms are depicted as nodes and connections between atoms within a specified distance serve as edges. This enhanced model is one of the most high-performing approaches on the OC20 dataset, delivering a 16% improvement in performance over the original GemNet model [169] while reducing training time by a factor of ten. The model is designed to accurately predict both the energy of the system and the forces acting on each individual atom, making it a powerful tool

for fine-tuning on custom datasets. Below, we will describe the specifics of GemNet-OC related to our work here (for details about this model, see Appendix 7.3.2.).

Figure 4-3 provides a streamlined overview of the GemNet-OC model architecture. At its core, GemNet-OC begins by embedding both atoms and the edges between them into high-dimensional vectors. This initial embedding step encodes geometric relationships via three types of Bessel functions: radial (RBF), circular (CBF), and spherical (SBF). The purpose of these functions is to capture pairwise distances, three-body angles, and four-body dihedral-like configurations. Polynomial envelopes are applied to these Bessel functions for smooth differentiability, following the approach of the DimeNet family of models[168,170].

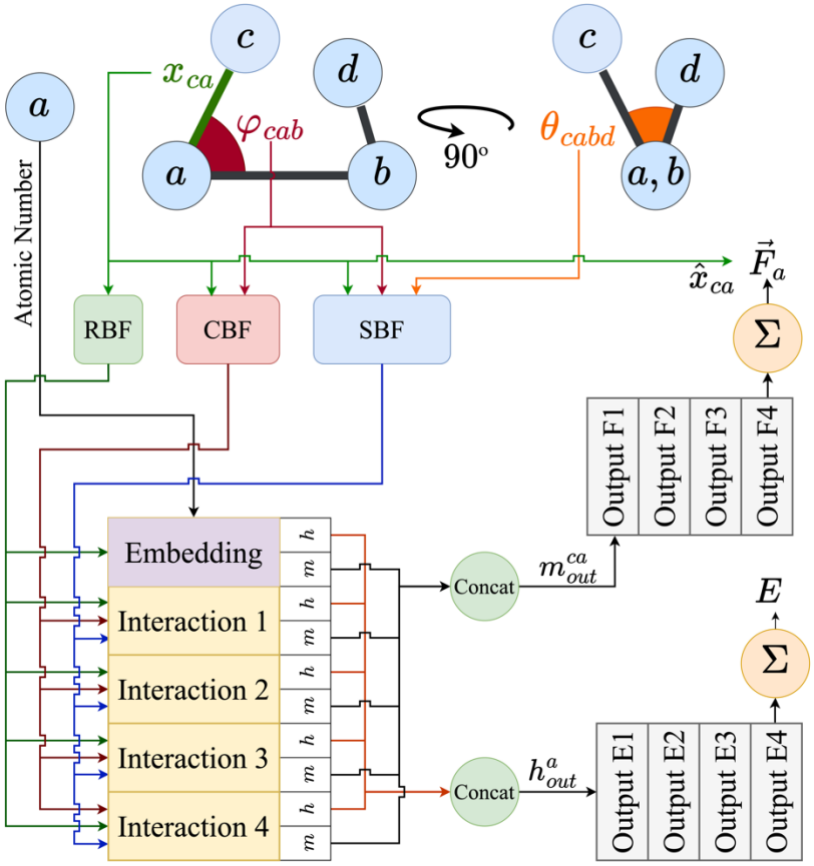

Figure 4-3 Simplified architecture of GemNet-OC. The model starts by embedding atoms and edges with Basis functions. These embeddings pass through an Embedding layer and multiple Interaction layers that refine representations for energy and force prediction. The Interaction stage outputs are then fed into the Output blocks, which convert the concatenated features into per-atom energy and force contributions that are then aggregated across all atoms.

The embeddings are then passed through an **Embedding** layer and multiple **Interaction** layers. Each of these layers produces two representations, h and m, which in GemNet-OC are used to compute energy and force information, respectively. Stacking several Interaction layers refines these representations, enabling the model to learn progressively richer chemical and geometric features. Finally, each Interaction stage feeds into **Output** blocks, consisting of dense and residual layers (labeled "Output E1-4" and "Output F1-4" in Figure 4-3). These blocks convert the concatenated representation from the interaction blocks into energy and force contributions for each node (atom). The total energy and forces on each atom are obtained by aggregating the energy and force outputs across all atoms.

The GemNet-OC model demonstrated strong performance on the pretraining data, achieving a 50.05% success rate and an energy mean absolute error (MAE) of 0.1694 eV when trained on the OC20 test dataset as reported in the AdsorbML paper [28], and an energy MAE of 0.483 eV when trained on both OC20 and OC22 datasets. However, its performance significantly declines when evaluated directly on the extracted FG dataset. Specifically, the success rate plummets to just 1.59%, and the energy MAE rises sharply to 5.562 eV. This stark contrast highlights the essential needs of fine-tuning in order to transfer the model's capabilities to different or more constrained datasets.

b) Fine-Tuning Strategies

In adapting the GemNet-OC model to the extracted FG dataset, three distinct fine-tuning strategies were explored, each differing in how much of the pretrained network is updated (shown in Table 4-1). These strategies are developed based on the fact that deep learning models typically learn features in a hierarchical manner, where the representations in early layers capture low-level details and progressively build up to more abstract concepts in deeper layers. This hierarchical progression of learned features has been widely observed and discussed in the literature [171,172] Similarly, in the GemNet-OC model, the design of its blocks follows a comparable principle. The interaction blocks (see Figure 4-2), which are positioned earlier in the network, primarily focus on encoding the geometrical relationships and local interactions of the input. In contrast, the output blocks are located deeper in the network and are tasked with synthesizing the processed information into representations that are directly related to predicting forces and energies. For

instance, Yang et al. [173] demonstrate the application of intermediate embeddings of a modified GemNet-dT model. By treating the intermediate representations as fingerprints, they enable efficient similarity searches in large databases, showcasing the versatility of GemNet-OC's architecture. Inspired by these works, the fine-tune strategies developed here are designed to probe the model's ability to capture nuanced geometric relationships as well as to uncover its potential for applications in materials discovery and structure-property mapping.

As shown in Table 4-1, the first strategy involved freezing the weights in the Interaction Blocks (FIB), which means these layers responsible for multi-level message passing remained fixed. Only the Output Blocks and any additional layers outside of the Interaction Blocks were trained. This approach preserves the geometric and chemical insights learned during pretraining, assuming that the fundamental understanding of interatomic relationships remains applicable to the new dataset. At the same time, the Output Blocks and the other parts can adjust their parameters to better align the existing embeddings with the new target task.

Table 4-1 The summary of the fine-tuning strategies

Strategy	**Frozen Blocks**	**Trainable Blocks** (Number of Trainable Parameters)	**Total Number of Trainable Parameters** (Fraction of the Model's Total Parameters)
Freezing the Interaction Blocks (FIB)	Interaction	Basis Functions (12,288) Atom and Edge Embeddings (21,248 + 327,680) Outputs (5×3,031,040) Energy and Force MLPs (2,949,888)	18,466,304 (~0.45)
Freezing the Output Blocks (FOB)	Output	Basis Functions (12,288) Atom and Edge Embeddings (21,248 + 327,680) Interactions (4×5,688,856) Energy and Force MLPs (2,949,888)	26,066,528 (~0.63)
Full Fine-Tuning (FFT)	-	Basis Functions (12,288) Atom and Edge Embeddings (21,248 + 327,680) Interactions (4×5,688,856) Outputs (5×3,031,040) Energy and Force MLPs (2,949,888)	41,221,728 (1.0)

The second strategy froze the weights in the Output Blocks (FOB), allowing only the Interaction Blocks and remaining modules to update. Since the Output Blocks produce the final transformations to energy and force predictions, freezing them retains the final layers' pretrained behavior. In contrast, the Interaction Blocks must adapt their embeddings to match these fixed output transformations while incorporating new data from the Extracted FG dataset. Lastly, the third strategy was full fine-tuning (FFT), where every learnable parameter in the GemNet-OC

model was allowed to update. This complete approach lets the network recalibrate both the geometry-learning and prediction layers, though it risks overfitting if the reduced dataset lacks sufficient coverage relative to the model's complexity. Each of these strategies balances preserving valuable pretraining knowledge against allowing enough flexibility for the new domain.

c) Evaluation Metrics

During the training and validation process, a range of metrics is used to monitor the accuracy of both energy and force predictions. Specifically, the mean absolute error (MAE) is measured for the predicted energies, while separate MAEs are calculated for the x, y, and z components of the forces. An overall forces MAE is also tracked, along with the cosine similarity between predicted and true force vectors, the magnitude error, and a combined metric that evaluates whether both energy and forces remain within certain thresholds. This assortment of metrics offers a detailed view of the model's performance across various dimensions of the prediction task, with the primary metric during training being the forces MAE.

For final evaluations, the key indicators are the success rate and the energy MAE. The success rate is defined as the percentage of test frames whose predicted energy lies within 0.1 eV of the DFT reference, following the metric proposed in the AdsorbML paper [28] and other seminal references [29,174,175]. Meanwhile, the energy MAE provides a direct measurement of how closely the model's energy predictions match the ground-truth DFT values. By considering both the success rate and the energy MAE, it is possible to balance the need for overall accuracy with the stricter requirement of matching DFT-level precision on a high proportion of test examples.

4.3.3. Results

a) Comparing Training Data Purity

Table 2 shows that the two fine-tuned models, each fully trained for 16 epochs, outperform the pretrained baseline by a large margin in both the Segregated aromatics and the Extracted FG tests. The pretrained network exhibits very poor performance, with success rates under 2% and MAE values above 5 eV in both test sets. This result highlights how relying on a model trained solely on the original OC20 data is not sufficient for either the FG or the aromatics distribution. In

Table 2, extracted FG refers to ensembles extracted from the full FG dataset (6,866 entries), while segregated aromatics consists exclusively of ensembles containing an aromatic molecule adsorbate from the FG dataset (1,140 entries).

To establish a baseline trained on the FG dataset, GAME-Net is utilized. GAME-Net, a GNN introduced in the same study[159] as the FG dataset for adsorption energy prediction, incorporates three core components: (1) fully connected layers for node-level feature transformation, (2) convolutional layers to aggregate neighbor node information, and (3) a pooling layer to generate graph-level energy predictions. With only 285,761 parameters, the architecture prioritizes efficiency while capturing local and global interactions. To evaluate its performance, a testing protocol analogous to ensemble extraction was employed, ensuring robustness in predictions.

GAME-Net's results were compared against fine-tuned versions of the larger GemNet-OC model (41 million parameters) on the same dataset. The comparison is limited to MAE, as the authors of GAME-Net did not include success rate. GAME-Net achieved a MAE of **0.34 eV** for adsorption energy prediction of aromatics, where GemNet-OC, when fine-tuned on segregated aromatic compounds, reduced the MAE to **0.125 eV**, and further to **0.084 eV** when trained on extracted FG data. This substantial performance gap highlights the efficacy of fine-tuning large pre-trained models. However, GAME-Net demonstrates competitive utility as a lightweight alternative for scenarios prioritizing computational efficiency over accuracy gains. The comparison underscores the trade-offs between model scale, generalizability, and precision in adsorption energy prediction tasks.

Comparing the two fully fine-tuned models, the one trained on the full FG dataset gives the highest success rate and the lowest MAE in both segregated-aromatics and extracted-FG tests. Even in the segregated-aromatics test, where one might expect the network fine-tuned solely on aromatics to excel, the FG-trained model still performs better. This suggests that using a larger and more varied dataset (FG) helps the network learn geometry and energy relationships that transfer reasonably well to aromatics. Even when the additional data is not aromatics, they're more similar to aromatics compared to the pretraining data (OC20 and OC22). In contrast, the model trained strictly on aromatics becomes too specialized for aromatic systems and struggles with FG test cases

that deviate from aromatic structures, leading to a substantial drop in performance on the reduced-FG test. Nonetheless, its accuracy on aromatics is not far behind that of the FG-trained network, confirming that it is indeed better adapted to those particular molecular configurations but lacks the broader coverage provided by the more diverse FG data.

Table 4-2 The evaluation metrics for the pretrained checkpoint and fine-tuned model using Extracted FG and FG-aromatics

	Segregated aromatics		Extracted FG	
Model	**Success Rate [%]**	**MAE [eV]**	**Success Rate [%]**	**MAE [eV]**
Pretrained GemNet-OC	0.88	10.975	1.59	5.562
GemNet-OC fine-tuned on **Extracted FG**	70.80	0.084	75.83	0.074
GemNet-OC fine-tuned on **Segregated aromatics**	60.18	0.125	23.15	0.456
GAME-Net trained on **Extracted FG**	-	0.34	-	0.18

b) Comparing Fine-Tuning Strategies

In the plots of Figure 4-4, each point shows the model's predicted energy versus the true DFT value after only one epoch of fine-tuning. The diagonal line indicates perfect agreement: points closer to that line are more accurate. The pretrained model (blue squares) remains widely scattered, reflecting a high MAE in both extracted FG and segregated aromatics. This is because it has not yet adapted to the new domain. In contrast, the fully fine-tuned model (orange squares) lies tightly around the diagonal after just one epoch, achieving the lowest MAE in both training sets.

As stated before and highlighted in the work of Yang et al. [173], the early layers of GemNet models, specifically Embedding and Interaction blocks, can be used to generate the atomic fingerprints to describe the local environment of an atom within a chemical system. This

suggests a functional distinction between the layers: the earlier blocks primarily focus on capturing geometric relationships (e.g., bond angles, distances, and directional interactions), while the Output blocks are more involved in transforming these geometric representations into predictions of forces and energies. However, this distinction is not absolute, as both sets of layers inherently contain information related to geometry and forces. Thus, when analyzing fine-tuning strategies, it is reasonable to associate earlier layers with geometric encoding and deeper layers with property prediction, while acknowledging the interconnected nature of these tasks.

In FIB strategy, when the Interaction Blocks are frozen (green squares), there is some improvement over the baseline, but the geometry layers remain stuck in their original configuration, so predictions are still more spread out than the fully fine-tuned case. Freezing the Output Blocks (FOB, red squares) helps the geometry layers adapt, though the static output layer imposes a temporary mismatch between newly learned embeddings and the final predictions. As a result, red points are closer than blue but less accurate than green for now.

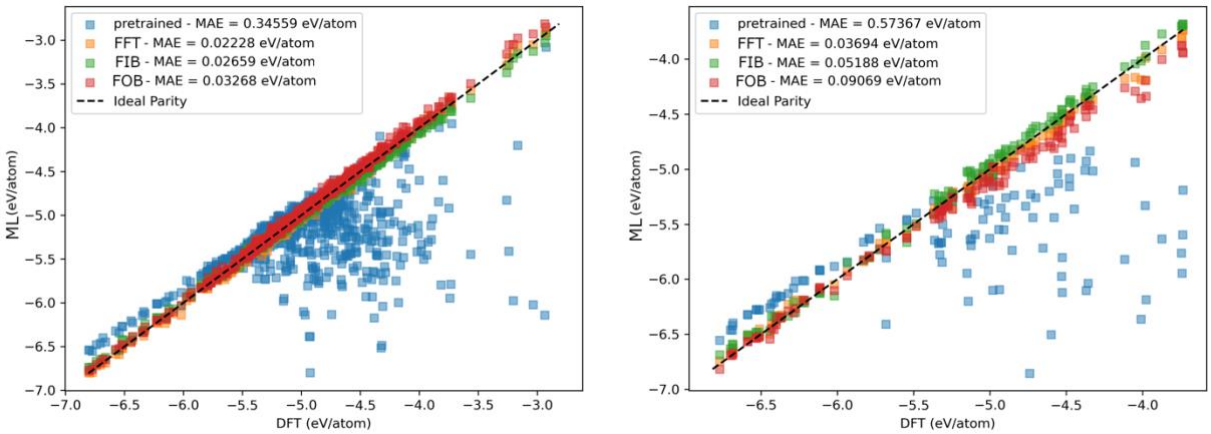

Figure 4-4 Parity plots showing each model's energy predictions versus the true DFT values after a single epoch of training on the **Extracted FG** (left) and **Segregated aromatics** (right) subsets.

For more detail study of different fine-tuning strategies, the evaluation metrics are thoroughly investigated over different training epochs. Figure 4-5 shows that fully fine-tuning the entire GemNet-OC network (FFT) yields the best improvement in accuracy on the Extracted FG dataset, as evidenced by the highest success rate (over 70% by epoch 16) and the lowest MAE. When the Interaction Blocks are frozen (FIB), the model's success rate actually declines over time,

and the MAE steadily increases. This suggests that the geometry-focused portion of the model (i.e., the message-passing layers) must be allowed to adapt to the new domain for good performance. Simply retraining the Output Blocks cannot compensate if the underlying geometric embeddings remain stuck in their pretrained state. By contrast, freezing the Output Blocks while allowing the Interaction Blocks to train (FOB) leads to moderate improvement. Here, the model can re-learn or refine the geometric relationships in the new dataset, although the final layers are still inherited from pretraining. This strategy yields better results than leaving the geometry layers frozen, but it still lags behind fully fine-tuning the entire network. Allowing both the geometry (Interaction Blocks) and the final prediction stages (Output Blocks) to adapt appears necessary to maximize accuracy on a domain that differs significantly from the data on which the model was originally trained.

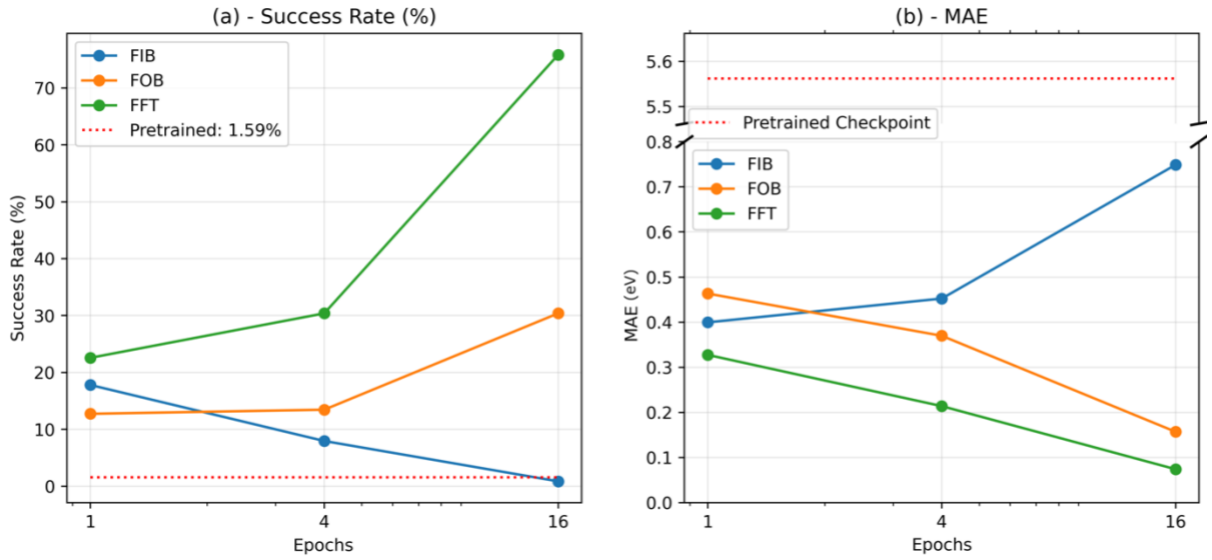

Figure 4-5 Evaluation metrics for fine-tuned models on Extracted FG in different training epochs. (a): Success rate, and (b): Energy's mean absolute error (MAE)

When the Interaction Blocks remain frozen, the model is locked into a geometric representation tuned to the original pretraining data. If the extracted FG dataset differs significantly, either in atomic configurations or chemical contexts, the fixed geometry layers cannot adapt to these new patterns. Although the Output Blocks may adjust initially, they do not have enough flexibility by themselves to reconcile the mismatch between an old geometric representation and the new domain. As training proceeds, this mismatch can actually grow worse.

The Output Blocks, trying to compensate for an outdated geometric embedding, might overfit or make inconsistent corrections. Consequently, the overall accuracy begins to decline instead of improving with additional epochs. By contrast, FOB and FFT allow either the Interaction Blocks or the entire network to update and learn geometry more suited to the Extracted FG dataset, which explains why their performance improves over time while the FIB approach degrades.

Figure 4-6 shows the results for the reduced set of only aromatic molecules (FG-aromatics). It can be seen that the trends differ slightly from the full dataset. In the case of full fine-tuning, the situation is similar to the FG-dataset: there is constant improvement with more epochs, though the absolute values of the evaluation metrics are worse compared to fine-tuning on all FG data. As explained in the previous section, this stems from the model underfitting due to less data, despite that data being more homogeneous.

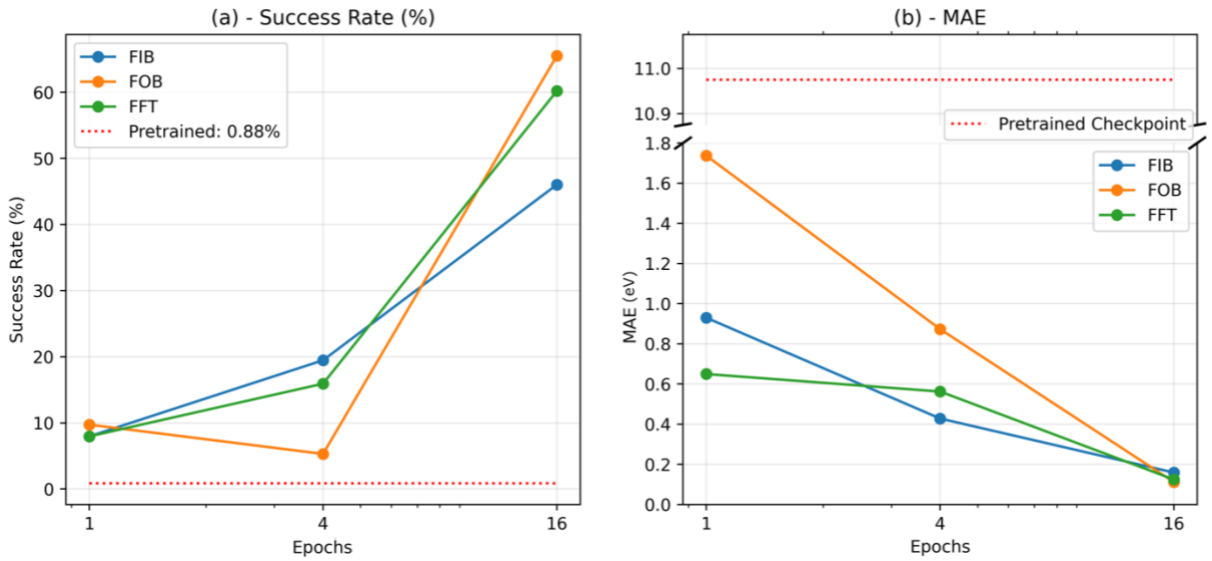

Figure 4-6 Evaluation metrics for fine-tuned models on a section of Extracted FG-aromatics in different training epochs. (a): Success rate, and (b): energy's mean absolute error

When the interaction blocks are frozen (FIB), although the geometric side of the model cannot be adjusted, its output (the input to the output blocks) is more consistent because of the more homogeneous geometry of aromatics, which is reflected in the output of the interaction blocks. Therefore, the output blocks can align themselves with this consistent signal, and this alignment leads to the continuous improvement of the FIB approach in both evaluation metrics.

The case of FOB is more complex. Because the aromatics are more homogeneous and closer in type to one another, the geometry representations may converge smoother toward a suitable embedding. The suitability of the intermediate embedding (the output of the interaction blocks and the input to the output blocks) must account for two factors: First, learning a good geometric representation of the aromatics; secondly, ensuring that representation matches the frozen output layer's transformation into force and energy values. Achieving both of these goals with a smaller dataset can still benefit from more epochs, which is evident in the steady and steep decrease in the MAE as shown in Figure 4-6.

However, there are two observations that need explanation in Figure 4-6: a) Unlike Figure 4-5, the FOB method trained for 16 epochs performs better than FFT on FG-aromatics, and its success rate is on par with full fine-tuning on the entire FG dataset, b) There is a dip in the success rate of the FOB method from epoch 1 to epoch 4, followed by a steep rise at epoch 16. For explaining the observation (a), one likely reason is that for a smaller, more homogeneous domain such as FG-aromatics, the pretrained output blocks still provide a robust mapping from geometry to force/energy values. Meanwhile, the trainable interaction blocks can converge relatively good toward representations that match these frozen output blocks. In contrast, fully fine-tuning on a small dataset may lead to more significant parameter shifts in both geometry and output layers, sometimes causing slower convergence or partial underfitting. Because the FOB approach only refines the geometry to align with a stable, pretrained output block, it can ultimately exceed the performance of fully fine-tuning under these conditions.

About the dip in performance at epoch 4, note that the success rate metric is extremely sensitive. As shown in the Appendix 7.3.3 (where "Energy, Forces within threshold" metric was used, similar to success rate but incorporating both energy and forces), performance can fluctuate throughout training. These fluctuations arise from transient misalignments between the evolving geometry representation and the frozen output stage, which do not necessarily show up as strongly in a relatively smoother error metric like MAE.

4.4. Conclusion

In this chapter, the capabilities of equivariant graph neural networks for predicting the adsorption energies of aromatic molecules on metal substrates were investigated. A GemNet-OC

model [31], which had been pretrained on the OC20/OC22 datasets [29,30] of smaller molecules, was fine-tuned on the FG dataset [159] that comprises molecules with up to 12 carbon atoms, including aromatics. Preprocessing was applied to the FG dataset to extract adsorbate atoms and nearby substrate atoms, thereby reducing structural complexity. Furthermore, the aromatic entries were segregated to form a dedicated FG-aromatics dataset.

Full fine-tuning of the GemNet-OC model on both the extracted FG and FG-aromatics datasets was performed. Superior performance was observed when the model was fine-tuned on the FG dataset, as demonstrated by evaluations on the test splits of both FG and FG-aromatics. Although improvements were noted for the FG-aromatics dataset, more significant gains were recorded on FG, likely due to the greater structural diversity. The improvements on Extracted FG can be associated with the larger volume of training data (non-aromatic entries of FG dataset) that still more closely resembled aromatic configurations compared to the original pretraining data (OC20/OC22). Ultimately, the decision to segregate was shown to depend on the specific objectives, which requires balancing the importance of detecting nuanced aromatic interactions against the potential loss of predictive power that might arise from not utilizing a more comprehensive dataset.

The architecture of the GemNet-OC model was comprised of distinct modules. Interaction blocks, which learned geometrical attributes through directional message passing between embedded coordinates, bond angles, and dihedral angles, were complemented by output blocks that transformed these embeddings into atom-level representations for energy and force prediction. To elucidate the role of these modules in domain adaptation, three fine-tuning strategies were adopted: freezing the interaction blocks (FIB), freezing the output blocks (FOB), and fully fine-tuning all modules (FFT). It was observed that full fine-tuning yielded the best overall performance, while the decline in performance under the FIB approach underscored the critical contribution of the interaction blocks. The success of full fine-tuning highlights the necessity of updating both geometric (interaction blocks) and task-specific (output blocks) components when transferring models to new molecular systems.

For the FG-aromatics subset, further observations were recorded regarding the fine-tuning strategies. When the interaction blocks were frozen (FIB), the geometric outputs were rendered

more consistent because of the homogeneous nature of aromatic molecules. This consistency enabled the output blocks to be effectively aligned with the stable signal, resulting in continuous improvements in both evaluations. On the other hand, with the FOB strategy, the convergence of the geometric representations toward a suitable embedding had to be achieved while ensuring that the fixed output layer correctly transformed this embedding into force and energy values. This caused larger errors in earlier steps of the training, but extended training epochs contributed to the alignment, as reflected by a steady decrease in mean absolute error.

In summary, it was demonstrated that enhanced predictive performance could be achieved by fine-tuning an equivariant graph neural network on a diverse dataset and that the interaction blocks played a pivotal role in adapting the model to new molecular domains. For accurate and transferable predictions, balancing data diversity with domain specificity and architecting models is key to probing the adsorptions of complex molecules onto metal substrates. These findings establish a roadmap for optimizing graph neural networks in computational surface science.

5. Conclusions and Future Perspectives

5.1. Conclusions

This dissertation advances the understanding of complex physical systems by integrating data-driven methods that range from analytical techniques to cutting-edge deep learning frameworks. Central to this work is the concept of *Janus systems,* governed by competing interactions that lead to emergent, threshold-driven behaviors. To demonstrate the cohesiveness and broader implications of the developed methodologies, this dissertation is structured around three interconnected objectives. Each objective employs innovative approaches to extract insights from existing nanoscale simulations, avoiding computationally intensive calculations while repurposing data originally generated for more complex theoretical inquiries. The work presented reveals how molecular-level competitions, such as adsorption versus aggregation, or polar versus non-polar interactions, create tipping points that dictate macroscopic outcomes. A unifying theme across these methods is their ability to revitalize legacy simulation data, transforming underutilized outputs into tools for probing new investigations. By bridging molecular interactions with system-scale phenomena, this dissertation provides a framework to decode ambiguity in amorphous systems and predict threshold-driven transitions.

The first objective explored the Janus-like interplay between asphaltene aggregation/adsorption and water-in-oil droplet coalescence through analyzing the outputs of standard molecular dynamics (MD) simulations. Tracking the behavior of water and asphaltene molecules revealed a competition between two forces: (1) the adsorption of asphaltenes at the water/oil interface, where their polyaromatic cores orient **perpendicularly to the water surface**, and (2) their intrinsic tendency to stack **parallel to one another**. This duality creates a nonmonotonic relationship between water concentration and asphaltene aggregation. While water disrupts aggregation by interacting with asphaltene polar groups, larger water droplets provide expanded interfacial areas that paradoxically promote simultaneous adsorption and aggregation.

To quantify coalescence dynamics, automated tools combining a customized clustering algorithm and exponential curve fitting were developed. The analysis revealed that water molecules accelerate coalescence via polar interactions, while asphaltenes impede it by forming dynamic barriers at the interface. By tracking individual molecules, two distinct growth modes

emerged: *stalling* (the largest droplet remains stable) and *growing* (the largest droplet absorbs smaller ones), governed by the dominance of interfacial forces over aggregation resistance. These results bridge molecular-scale competitions to macroscopic emulsion behavior, demonstrating how opposing forces create threshold-driven transitions. This mechanistic understanding offers a blueprint for industries like petroleum processing and cosmetics to predict and control emulsion stability by manipulating molecular interactions.

The second objective addressed the Janus dynamics of solute–solvent interactions by developing a novel method to calculate partial molar volumes (PMV) in multicomponent mixtures from molecular dynamics (MD) trajectories. Building on prior approaches, this method introduced control volumes to enable PMV determination in systems with fixed compositions, overcoming a key limitation of existing techniques. Validation against experimental data achieved an average error of 2.49%, confirming its reliability. When applied to systems containing oil, asphaltene models, and aggregation inhibitors, the PMV method revealed a competition between two opposing forces: solute–solute aggregation (promoting insolubility) and solute–solvent interactions (enhancing dissolution). A decrease in PMV correlated with improved solubility, reflecting a tipping point where solvent interactions overpower aggregation tendencies. Radial distribution function (RDF) analyses and cluster size tracking demonstrated that inhibitors disrupt asphaltene stacking by binding to their polyaromatic cores, favoring smaller, solvated aggregates. Higher inhibitor concentrations amplified this effect, as non-polar inhibitor tails preferentially interacted with asphaltenes, destabilizing larger aggregates. This PMV method is simple, efficient, validated, and provides a universal metric to quantify competing interactions in multicomponent systems. Its ability to identify molecular drivers of solubility transitions offers practical value for industries seeking to optimize formulations, from stabilizing colloids to enhancing dissolution in complex mixtures.

The third objective focused on applying advanced deep learning to predict adsorption energies of aromatic molecules, such as small asphaltene models, on metal substrates. To achieve this, the study utilized GemNet-OC [31], a graph neural network (GNN) that incorporates equivariance (to rotational/translational transformations) and directional message passing to model geometric relationships. In GemNet-OC, *interaction blocks* process atomic configurations by iteratively updating node features through directional messages between atoms, capturing nuanced

geometric details of adsorbate-substrate pairs. The remaining modules then transform these learned representations into energy and force predictions. The core contribution of the work presented lies in adapting this pretrained model to aromatic systems through strategic fine-tuning. Two datasets were curated: the **FG dataset** (diverse adsorbate-substrate pairs) and the **FG-aromatics subset** (specialized for aromatic interactions). A key Janus dynamic emerged: the FG dataset's size and diversity improved generalizability, while the FG-aromatics subset enhanced specificity for aromatic interactions critical to adsorption.

An ablation-like study revealed that updating *only the interaction blocks* during fine-tuning while preserving pretrained weights in other modules, would result in comparable performance with full fine-tuning. This highlights that geometric feature adaptation is pivotal for transfer learning, whereas energy/force decoders rely on conserved physical principles. Full fine-tuning of all modules on the FG dataset achieved broad applicability, while the FG-aromatics subset prioritized precision in aromatic systems. These insights provide a roadmap for balancing data specificity and scale when repurposing large GNNs for niche domains, with implications for oil transportations and catalyst design where interfacial engineering is the key.

At its core, this dissertation deciphers how competitions among actors in the systems (adsorption versus aggregation, dissolution versus precipitation, and model generalizability versus specificity) create tipping points that define system behavior. The three chapters presented demonstrate how integrating classical data analysis with modern machine learning can unravel the behavior of complex Janus systems. By repurposing existing MD simulations and adapting pretrained models, this work bridges molecular-scale interactions to macroscopic properties, offering quantitative insights into phenomena such as emulsion stability, solubility transitions, and adsorption.

The methodologies developed here offer scalable tools to advance the theoretical understanding of interfacial and colloidal dynamics. In summary, key methodological advancements in this work include: 1) a microscale approach to monitor coalescence, 2) a simplified method for calculation of partial molar properties, and 3) a transfer learning strategy that leverages sparse, domain-specific first-principles calculations for energy prediction. Beyond enhancing nanoscale insights into complex interfacial and amorphous systems **without requiring**

additional simulations, these methods provide practical tools for industries seeking to control such systems. In petroleum processing, they enable predictive screening of emulsion stabilizers and asphaltene inhibitors, helping to reduce costly pipeline fouling and prevent asphaltene deposition. Additionally, they contribute to catalyst design by accelerating the discovery of surface-active materials with precise binding affinities.

5.2. Future Perspectives

While this dissertation has successfully streamlined the analysis of various Janus systems, several promising avenues remain for further investigation to enhance and broaden the applicability of these methodologies.

One key area for development is the refinement of droplet coalescence and asphaltene aggregation analysis. Moving beyond a simple dichotomy of stalling versus growth, future research should aim to develop a more nuanced classification of droplet growth modes. For instance, employing higher-resolution tracking of coalescence events, integrating detailed morphological metrics, and probing transient states during droplet merging can uncover subtle nucleation and growth mechanisms. The techniques used for monitoring graph dynamics, like node/edge matching and graph differencing can be used in this front. This approach will help analyze the merging dynamics of smaller ones, in addition to the current focus on the growing of the largest droplets.

Another significant direction involves extending the methodology to encompass additional partial molar properties, such as energies, enthalpies, and entropies. Capturing a broader range of solute–solvent interactions would not only enrich the thermodynamic profile of complex mixtures but also enhance predictive capabilities. For instance, partial molar enthalpy can be computed directly for a given simulation frame, from instantaneous measures of internal energy and volume [116]

Finally, for adsorption energy prediction, exploring alternative deep learning architectures represents a critical step forward. Investigating models like Equiformers [32,164], alongside systematic evaluation of different fine-tuning strategies, could offer valuable comparisons to the

current model and highlight the key components necessary for robust domain adaptation. For instance, this study used only two datasets and showed that full fine-tuning maximizes accuracy for a 40M-parameter GemNet model on larger datasets (e.g., Extracted FG), while partial parameter freezing achieves similar results with 40% less compute on smaller datasets (e.g., Segregated aromatics). Expanding to hierarchical DFT datasets (progressively smaller and more specialized) paired with advanced models like Equiformer v2, could systematically reveal how model size, data granularity, and tuning strategies interact. By mapping these relationships across architectures, empirical scaling laws could emerge, defining thresholds where specific model-data pairings optimize accuracy or efficiency. Identifying precise thresholds for varying model sizes and architectures is particularly relevant as state-of-the-art models now exceed 150M parameters [176].

These future directions based on the current dissertation will extend the impacts of data-driven strategies across multiple fields, deepen our understandings of Janus systems, and facilitate more precise control in practical applications.

Beyond enhancing the current applications, the foundational data-driven approaches of this dissertation can be generalized to a wider range of scientific challenges. The graph-based analysis of droplet dynamics, for example, is not limited to fluid systems; it could be adapted to study other complex time-series processes like crystallization nucleation [177,178], protein folding pathways, or polymer self-assembly [179,180]. Likewise, the fine-tuning methodologies for equivariant GNNs are directly transferable to predicting other critical material properties, such as thermal conductivity [181]or mechanical strength [182], extending the impact of this work far beyond its original scope.

Perhaps the most transformative future direction involves shifting the paradigm from predictive analysis to active, generative discovery using state-of-the-art AI. One major avenue is the development of physics-conditioned diffusion models [183] to act as surrogate molecular dynamics simulators. By incorporating physical laws such as energy conservation, directly into the training and sampling procedures, as demonstrated in autonomous driving through manual bridges [184] or Oracle-style discriminators for enforcing physical validity [185], diffusion models can be conditioned to generate realistic molecular trajectories. This approach offers the potential

to capture rare events and long-timescale phenomena at a fraction of the traditional computational cost.

A complementary approach involves using models like variational autoencoders (VAEs) to learn a probabilistic "map" of chemical possibilities, which is especially powerful for inverse design [186]. By learning a latent space of stable molecules, such models could generate entirely new chemical structures tailored to exhibit specific target properties, such as high catalytic activity or optimal adsorption energy. Crucially, this probabilistic framework provides built-in uncertainty quantification, guiding researchers to focus experimental efforts on the most promising and high-confidence candidates. Together, these generative strategies represent a paradigm shift from analyzing existing systems to designing novel ones from the ground up, leveraging the foundation of this dissertation to pioneer the future of autonomous materials discovery.

6. References

[1] Q. Chai, Y. Jiao, X. Yu, Hydrogels for biomedical applications: Their characteristics and the mechanisms behind them, Gels 3 (2017). https://doi.org/10.3390/gels3010006.

[2] J. Li, D.J. Mooney, Designing hydrogels for controlled drug delivery, Nat Rev Mater 1 (2016) 16071. https://doi.org/10.1038/natrevmats.2016.71.

[3] R.R. Nair, S. Debnath, S. Das, P. Wakchaure, B. Ganguly, P.B. Chatterjee, A Highly Selective Turn-On Biosensor for Measuring Spermine/Spermidine in Human Urine and Blood, ACS Appl Bio Mater 2 (2019) 2374–2387. https://doi.org/10.1021/acsabm.9b00084.

[4] H. Wang, H. Wang, Y. Li, C. Jiang, D. Chen, Y. Wen, Z. Li, Capillarity self-driven DNA hydrogel sensor for visual quantification of microRNA, Sens Actuators B Chem 313 (2020). https://doi.org/10.1016/j.snb.2020.128036.

[5] X. Xue, Y. Hu, S. Wang, X. Chen, Y. Jiang, J. Su, Fabrication of physical and chemical crosslinked hydrogels for bone tissue engineering, Bioact Mater 12 (2022) 327–339. https://doi.org/10.1016/j.bioactmat.2021.10.029.

[6] H. Pan, H. Gao, Q. Li, Z. Lin, Q. Feng, C. Yu, X. Zhang, H. Dong, D. Chen, X. Cao, Engineered macroporous hydrogel scaffolds: Via pickering emulsions stabilized by MgO nanoparticles promote bone regeneration, J Mater Chem B 8 (2020) 6100–6114. https://doi.org/10.1039/d0tb00901f.

[7] K.C.B. Maia, A. Densy dos Santos Francisco, M.P. Moreira, R.S.V. Nascimento, D. Grasseschi, Advancements in Surfactant Carriers for Enhanced Oil Recovery: Mechanisms, Challenges, and Opportunities, ACS Omega 9 (2024) 36874–36903. https://doi.org/10.1021/acsomega.4c04058.

[8] X. Li, Y. Guo, E.S. Boek, X. Guo, Experimental Study on Kinetics of Asphaltene Aggregation in a Microcapillary, Energy and Fuels 31 (2017) 9006–9015. https://doi.org/10.1021/acs.energyfuels.7b01170.

[9] G. Graeber, C.D. Díaz-Marín, L.C. Gaugler, Y. Zhong, B. El Fil, X. Liu, E.N. Wang, Extreme Water Uptake of Hygroscopic Hydrogels through Maximized Swelling-Induced Salt Loading, Advanced Materials 36 (2024). https://doi.org/10.1002/adma.202211783.

[10] Z. Zhang, J. Song, Y.J. Lin, X. Wang, S.L. Biswal, Comparing the Coalescence Rate of Water-in-Oil Emulsions Stabilized with Asphaltenes and Asphaltene-like Molecules, Langmuir 36 (2020) 7894–7900. https://doi.org/10.1021/acs.langmuir.0c00966.

[11] N. Asy-Syifa, Kusjuriansah, W.X. Waresindo, D. Edikresnha, T. Suciati, K. Khairurrijal, The Study of the Swelling Degree of the PVA Hydrogel with varying concentrations of PVA, in: J Phys Conf Ser, Institute of Physics, 2022. https://doi.org/10.1088/1742-6596/2243/1/012053.

[12] Z. Zhao, X. Xia, Y. Li, D. Liu, W. Cai, C. Li, G. Sun, B. Yao, F. Yang, Effect of Dodecylbenzenesulfonic Acid as an Asphaltene Dispersant on the W/O Emulsion Stabilized by Asphaltenes and Paraffin Wax, Energy and Fuels 37 (2023) 4244–4255. https://doi.org/10.1021/acs.energyfuels.2c03687.

[13] B. Jiang, R. Zhang, N. Yang, L. Zhang, Y. Sun, C. Jian, L. Liu, Z. Xu, Molecular mechanisms of suppressing asphaltene aggregation and flocculation by dodecylbenzenesulfonic acid probed by molecular dynamics simulations, Energy and Fuels 33 (2019) 5067–5080. https://doi.org/10.1021/acs.energyfuels.9b00821.

[14] R. Skartlien, S. Simon, J. Sjöblom, A DPD study of asphaltene aggregation: The role of inhibitor and asphaltene structure in diffusion-limited aggregation, J Dispers Sci Technol 38 (2017) 440–450. https://doi.org/10.1080/01932691.2016.1172972.
[15] Y. Ruiz-Morales, O.C. Mullins, Coarse-Grained Molecular Simulations to Investigate Asphaltenes at the Oil−Water Interface, Energy & Fuels 29 (2015) 1597–1609.
[16] J. Djuve, X. Yang, I.J. Fjellanger, J. Sjöblom, E. Pelizzetti, Chemical destabilization of crude oil based emulsions and asphaltene stabilized emulsions, Colloid Polym Sci 279 (2001) 232–239. https://doi.org/10.1007/s003960000413.
[17] S. Yaseen, G.A. Mansoori, Molecular dynamics studies of interaction between asphaltenes and solvents, J Pet Sci Eng 156 (2017) 118–124. https://doi.org/10.1016/j.petrol.2017.05.018.
[18] N. Patel, D.N. Dubins, R. Pomès, T. V. Chalikian, Parsing Partial Molar Volumes of Small Molecules: A Molecular Dynamics Study, J Phys Chem B 115 (2011) 4856–4862. https://doi.org/10.1021/jp2012792.
[19] T. Kuznicki, J.H. Masliyah, S. Bhattacharjee, Aggregation and Partitioning of Model Asphaltenes at Toluene−Water Interfaces: Molecular Dynamics Simulations, Energy & Fuels 23 (2009) 5027–5035. https://doi.org/10.1021/ef9004576.
[20] C. Jian, H. Zeng, Q. Liu, T. Tang, Probing the Adsorption of Polycyclic Aromatic Compounds onto Water Droplets Using Molecular Dynamics Simulations, The Journal of Physical Chemistry C 120 (2016) 14170–14179. https://doi.org/10.1021/acs.jpcc.6b03850.
[21] Y. Wang, J.M. Lamim Ribeiro, P. Tiwary, Machine learning approaches for analyzing and enhancing molecular dynamics simulations, Curr Opin Struct Biol 61 (2020) 139–145. https://doi.org/10.1016/j.sbi.2019.12.016.
[22] W. Chen, H. Sidky, A.L. Ferguson, Capabilities and Limitations of Time-lagged Autoencoders for Slow Mode Discovery in Dynamical Systems, (2019). https://doi.org/10.1063/1.5112048.
[23] A. Mardt, L. Pasquali, H. Wu, F. Noé, VAMPnets for Deep Learning of Molecular Kinetics, Nat Commun 9 (2018) 5. https://doi.org/10.1038/s41467-017-02388-1.
[24] W. Chen, A.L. Ferguson, Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration, J Comput Chem 39 (2018) 2079–2102. https://doi.org/10.1002/jcc.25520.
[25] J.M.L. Ribeiro, P. Bravo, Y. Wang, P. Tiwary, Reweighted autoencoded variational Bayes for enhanced sampling (RAVE), Journal of Chemical Physics 149 (2018). https://doi.org/10.1063/1.5025487.
[26] Y. Wang, J.M.L. Ribeiro, P. Tiwary, Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics, Nat Commun 10 (2019). https://doi.org/10.1038/s41467-019-11405-4.
[27] J. Wang, Y. Deng, B. Roux, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials, Biophys J 91 (2006) 2798–2814. https://doi.org/10.1529/biophysj.106.084301.
[28] J. Lan, A. Palizhati, M. Shuaibi, B.M. Wood, B. Wander, A. Das, M. Uyttendaele, C.L. Zitnick, Z.W. Ulissi, AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials, NPJ Comput Mater 9 (2023). https://doi.org/10.1038/s41524-023-01121-5.
[29] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C.L.

Zitnick, Z. Ulissi, Open Catalyst 2020 (OC20) Dataset and Community Challenges, ACS Catal 11 (2021) 6059–6072. https://doi.org/10.1021/acscatal.0c04525.
[30] R. Tran, J. Lan, M. Shuaibi, B.M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, A. Sriram, F. Therrien, J. Abed, O. Voznyy, E.H. Sargent, Z. Ulissi, C.L. Zitnick, The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts, in: AIChE Annual Meeting, Conference Proceedings, American Institute of Chemical Engineers, 2023: pp. 3066–3084. https://doi.org/10.1021/acscatal.2c05426.
[31] J. Gasteiger, M. Shuaibi, A. Sriram, S. Günnemann, Z. Ulissi, C.L. Zitnick, A. Das, GemNet-OC: Developing Graph Neural Networks for Large and Diverse Molecular Simulation Datasets, (2022). http://arxiv.org/abs/2204.02782.
[32] Y.-L. Liao, B. Wood, A. Das, T. Smidt, EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations, (2023). http://arxiv.org/abs/2306.12059.
[33] O.C. Mullins, The Asphaltenes, Annual Review of Analytical Chemistry 4 (2011) 393–418. https://doi.org/10.1146/annurev-anchem-061010-113849.
[34] D.P. Ortiz, E.N. Baydak, H.W. Yarranton, Effect of Surfactants on Interfacial Films and Stability of Water-in-Oil Emulsions Stabilized by Asphaltenes, J. Colloid Interface Sci. 351 (2010) 542–555.
[35] M. Fingas, B. Fieldhouse, Studies of the Formation Process of Water-in-Oil Emulsions, Mar. Pollut. Bull. 47 (2003) 369–396.
[36] L. Xia, S. Lu, G. Cao, Stability and demulsification of emulsions stabilized by asphaltenes or resins, J Colloid Interface Sci 271 (2004) 504–506. https://doi.org/10.1016/j.jcis.2003.11.027.
[37] O. V Gafonova, H.W. Yarranton, The Stabilization of Water-in-Hydrocarbon Emulsions by Asphaltenes and Resins, J. Colloid Interface Sci. 241 (2001) 469–478.
[38] J.D. McLean, P.M. Spiecker, A.P. Sullivan, P.K. Kilpatrick, The Role of Petroleum Asphaltenes in the Stabilization of Water-in-Oil Emulsions, in: Structures and Dynamics of Asphaltenes, Springer, Boston, MA, 1998.
[39] G. Yu, K. Karinshak, J.H. Harwell, B.P. Grady, A. Woodside, M. Ghosh, Interfacial Behavior and Water Solubility of Various Asphaltenes at High Temperature, Colloids Surf A: Physicochemical and Engineering Aspects 441 (2014) 378–388.
[40] G. Alvarez, S. Poteau, J.-F. Argillier, D. Langevin, J.-L. Salager, Heavy Oil-Water Interfacial Properties and Emulsion Stability: Influence of Dilution, Energy & Fuels 23 (2009) 294–299.
[41] S. Poteau, J.-F. Argillier, D. Langevin, F. Pincet, E. Perez, Influence of pH on Stability and Dynamic Properties of Asphaltenes and Other Amphiphilic Molecules at the Oil−Water Interface, Energy & Fuels 19 (2005) 1337–1341.
[42] Y. Lin, A. Perrard, S.L. Biswal, R.M. Hill, S. Trabelsi, Microfluidic Investigation of Asphaltenes-Stabilized Water-in-Oil Emulsions, Energy & Fuels (2018).
[43] H.W. Yarranton, P. Urrutia, D.M. Sztukowski, Effect of Interfacial Rheology on Model Emulsion Coalescence: II, J. Colloid Interface Sci. 310 (2007) 253–259.
[44] M.R. Gray, H.W. Yarranton, M.L. Chacón-Patiño, R.P. Rodgers, B. Bouyssiere, P. Giusti, Distributed Properties of Asphaltene Nanoaggregates in Crude Oils: A Review, Energy & Fuels 35 (2021) 18078–18103.

[45] F. Yang, P. Tchoukov, E. Pensini, T. Dabros, J. Czarnecki, J. Masliyah, Z. Xu, Asphaltene Subfractions Responsible for Stabilizing Water-in-Crude Oil Emulsions. Part 1: Interfacial Behaviors, Energy & Fuels 28 (2014) 6897–6904.
[46] L. Goual, B. Zhang, Y. Rahham, Nanoscale Characterization of Thin Films at Oil/Water Interfaces and Implications to Emulsion Stability, Energy & Fuels 35 (2021) 444–455. https://doi.org/10.1021/acs.energyfuels.0c03466.
[47] P. Qiao, D. Harbottle, P. Tchoukov, X. Wang, Z. Xu, Asphaltene Subfractions Responsible for Stabilizing Water-in-Crude Oil Emulsions. Part 3. Effect of Solvent Aromaticity, Energy & Fuels 31 (2017) 9179–9187.
[48] Y. Rahham, K. Rane, L. Goual, Characterization of the Interfacial Material in Asphaltenes Responsible for Oil/Water Emulsion Stability, Energy & Fuels 34 (2020) 13871–13882.
[49] D.A. Ballard, J.H. Pickering, I. Rosbottom, S. Tangparitkul, K.J. Roberts, R. Rae, P.J. Dowding, R.B. Hammond, D. Harbottle, Molecular Survey of Strongly and Weakly Interfacially Active Asphaltenes: An Intermolecular Force Field Approach, Energy & Fuels 35 (2021) 17424–17433.
[50] E.Y. Sheu, Petroleum Asphaltene Properties, Characterization, and Issues, Energy & Fuels 16 (2002) 74–82.
[51] X. Tan, H. Fenniri, M.R. Gray, Water Enhances the Aggregation of Model Asphaltenes in Solution via Hydrogen Bonding, Energy & Fuels 23 (2009) 3687–3693. https://doi.org/10.1021/ef900228s.
[52] Y. Ruiz-Morales, X. Wu, O.C. Mullins, Electronic Absorption Edge of Crude Oils and Asphaltenes Analyzed by Molecular Orbital Calculations with Optical Spectroscopy, Energy & Fuels 21 (2007) 944–952. https://doi.org/10.1021/ef0605605.
[53] M.R. Gray, R.R. Tykwinski, J.M. Stryker, X. Tan, Supramolecular Assembly Model for Aggregation of Petroleum Asphaltenes, Energy & Fuels 25 (2011) 3125–3134. https://doi.org/10.1021/ef200654p.
[54] J.D. Cyran, A.T. Krummela, Probing Structural Features of Self-Assembled Violanthrone-79 Using Two Dimensional Infrared Spectroscopy, J. Chem. Phys. 142 (2015) 212435.
[55] I.D. Mackie, G.A. DiLabio, Importance of the Inclusion of Dispersion in the Modeling of Asphaltene Dimers, Energy & Fuels 24 (2010) 6468–6475.
[56] F. Rakotondradany, H. Fenniri, P. Rahimi, K.L. Gawrys, P.K. Kilpatrick, M.R. Gray, Hexabenzocoronene Model Compounds for Asphaltene Fractions: Synthesis & Characterization, Energy & Fuels 20 (2006) 2439–2447.
[57] E. Rogel, Simulation of Interactions in Asphaltene Aggregates, Energy & Fuels 14 (2000) 566–574.
[58] C. Jian, T. Tang, S. Bhattacharjee, Molecular Dynamics Investigation on the Aggregation of Violanthrone78-Based Model Asphaltenes in Toluene, Energy & Fuels 28 (2014) 3604–3613. https://doi.org/10.1021/ef402208f.
[59] C. Jian, T. Tang, S. Bhattacharjee, Probing the Effect of Side-Chain Length on the Aggregation of a Model Asphaltene Using Molecular Dynamics Simulations, Energy & Fuels 27 (2013) 2057–2067. https://doi.org/10.1021/ef400097h.
[60] D. Ji, G. Liu, X. Zhang, C. Zhang, S. Yuan, Molecular Dynamics Study on the Adsorption of Heavy Oil Drops on a Silica Surface with Different Hydrophobicity, Energy & Fuels 34 (2020) 7019–7028.

[61] F. Gao, Z. Xu, G. Liu, S. Yuan, Molecular Dynamics Simulation: The Behavior of Asphaltene in Crude Oil and at the Oil/Water Interface, Energy & Fuels 28 (2014) 7368–7376.
[62] R.B. Teklebrhan, L. Ge, S. Bhattacharjee, Z. Xu, J. Sjöblom, Initial Partition and Aggregation of Uncharged Polyaromatic Molecules at the Oil−Water Interface: A Molecular Dynamics Simulation Study, J. Phys. Chem. B 118 (2014) 1040–1050.
[63] J. Liu, Y. Zhao, S. Ren, Molecular Dynamics Simulation of Self-Aggregation of Asphaltenes at an Oil/Water Interface: Formation and Destruction of the Asphaltene Protective Film, Energy & Fuels 29 (2015) 1233–1242.
[64] C. Jian, Q. Liu, H. Zeng, T. Tang, A Molecular Dynamics Study of the Effect of Asphaltenes on Toluene/Water Interfacial Tension: Surfactant or Solute?, Energy & Fuels 32 (2018) 3225–3231.
[65] Y. Mikami, Y. Liang, T. Matsuoka, E.S. Boek, Molecular Dynamics Simulations of Asphaltenes at the Oil–Water Interface: From Nanoaggregation to Thin-Film Formation, Energy & Fuels 27 (2013) 1838–1845.
[66] M. Kunieda, K. Nakaoka, Y. Liang, C.R. Miranda, A. Ueda, S. Takahashi, H. Okabe, T. Matsuoka, Self-Accumulation of Aromatics at the Oil−Water Interface through Weak Hydrogen Bonding, J. Am. Chem. Soc. 132 (2010) 18281–18286.
[67] J. Wang, M.A. Gayatri, A.L. Ferguson, Mesoscale Simulation and Machine Learning of Asphaltene Aggregation Phase Behavior and Molecular Assembly Landscapes, J. Phys. Chem. B 121 (2017) 4923–4944.
[68] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, J.C. Grossman, Graph Dynamical Networks for Unsupervised Learning of Atomic Scale Dynamics in Materials, Nat. Commun. 10 (2019) 2667.
[69] X. Sun, H. Zeng, T. Tang, Molecular simulation of folding and aggregation of multi-core polycyclic aromatic compounds, J. Mol. Liq. 310 (2020) 113248.
[70] T. Lan, H. Zeng, T. Tang, Understanding Adsorption of Violanthrone-79 as a Model Asphaltene Compound on Quartz Surface Using Molecular Dynamics Simulations, The Journal of Physical Chemistry C 122 (2018) 28787–28796. https://doi.org/10.1021/acs.jpcc.8b09712.
[71] X. Sun, H. Zeng, T. Tang, Effect of non-ionic surfactants on the adsorption of polycyclic aromatic compounds at water/oil interface: A molecular simulation study, J Colloid Interface Sci 586 (2021) 766–777. https://doi.org/10.1016/j.jcis.2020.10.146.
[72] S. Subramanian, S. Simon, B. Gao, J. Sjöblom, Asphaltene Fractionation Based on Adsorption Onto Calcium Carbonate: Part 1. Characterization of Sub-Fractions and QCM-D Measurements, Colloids Surf. 495 (2016) 136–148.
[73] B.B. Nielsen, W.Y. Svrcek, A.K. Mehrotra, Effects of Temperature and Pressure on Asphaltene Particle Size Distributions in Crude Oils Diluted with n-Pentane, Ind. Eng. Chem. Res. 33 (1994) 1324–1330.
[74] A.W. Schüttelkopf, D.M.F. van Aalten, PRODRG: a Tool for High-Throughput Crystallography of Protein-Ligand Complexes, Acta Cryst. 60 (2004) 1355–1363. https://doi.org/10.1107/S0907444904011679.
[75] C. Oostenbrink, A. Villa, A.E. Mark, W.F. Van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6, J Comput Chem 25 (2004) 1656–1676. https://doi.org/10.1002/jcc.20090.

[76] C. Jian, M.R. Poopari, Q. Liu, N. Zerpa, H. Zeng, T. Tang, Mechanistic Understanding of the Effect of Temperature and Salinity on the Water/Toluene Interfacial Tension, Energy & Fuels 30 (2016) 10228–10235. https://doi.org/10.1021/acs.energyfuels.6b01995.
[77] J. Zielkiewicz, Structural Properties of Water: Comparison of the SPC, SPCE, TIP4P, and TIP5P Models of Water, J. Chem. Phys. 123 (2005) 104501.
[78] D.V.D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, GROMACS: Fast, Flexible, and Free, J Comput Chem 26 (2005) 1701–1712. https://doi.org/10.1002/jcc.20291.
[79] U. Essmann, L. Perera, M.L. Berkowitz, A Smooth Particle Mesh Ewald Method, J. Chem. Phys. 103 (1995) 8577–8593. https://doi.org/10.1063/1.470117.
[80] S. Miyamoto, P.A. Kollman, Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models, J. Comput. Chem. 13 (1992) 952–962. https://doi.org/10.1002/jcc.540130805.
[81] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, LINCS: A Linear Constraint Solver for Molecular Simulations, J. Comput. Chem. 18 (1997) 1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
[82] C. Jian, Q. Liu, H. Zeng, T. Tang, Effect of Model Polycyclic Aromatic Compounds on the Coalescence of Water-in-Oil Emulsion Droplets, J. Phys. Chem. C 121 (2017) 10382–10391.
[83] Y.-L. Wang, A. Laaksonen, M.D. Fayer, Hydrogen Bonding versus π−π Stacking Interactions in Imidazolium−Oxalatoborate Ionic Liquid, J. Phys. Chem. B 121 (2017) 7173–7179.
[84] I. Vinckier, P. Moldenaers, A.M. Terracciano, N. Grizzuti, Droplet Size Evolution During Coalescence in Semiconcentrated Model Blends, AIChE Journal 44 (1998) 951–958.
[85] S. Mozaffari, H. Ghasemi, P. Tchoukov, J. Czarnecki, N. Nazemifard, Lab-on-a-Chip Systems in Asphaltene Characterization: A Review of Recent Advances, Energy & Fuels 35 (2021) 9080–9101.
[86] W. Wong, T. Singh, D. Vak, W. Pisula, C. Yan, X. Feng, E. Williams, K. Chan, Q. Mao, D. Jones, C.-Q. Ma, K. Müllen, P. Bäuerle, A. Holmes, Solution Processable Fluorenyl Hexa-peri-hexabenzocoronenes in Organic Field-Effect Transistors and Solar Cells, Adv. Funct. Mater 20 (2010) 927–938.
[87] H. Choi, S. Paek, J. Song, C. Kim, N. Cho, J. Ko, Synthesis of Annulated Thiophene Perylene Bisimide Analogues: Their Applications to Bulk Heterojunction Organic Solar Cells, Chem. Commun. 47 (2011) 5509–5511.
[88] N. Mizoshita, T. Tani, S. Inagaki, Highly Conductive Organosilica Hybrid Films Prepared from a Liquid-Crystal Perylene Bisimide Precursor, Adv. Funct. Mater 21 (2011) 3291–3296.
[89] L. Schmidt-Mende, A. Fechtenkötter, K. Müllen, E. Moons, R.H. Friend, J.D. MacKenzie, Self-Organized Discotic Liquid Crystals for High-Efficiency Organic Photovoltaics, Science (1979) 293 (2001) 1119–1122.
[90] C. Jian, T. Tang, One-Dimensional Self-Assembly of Polyaromatic Compounds Revealed by Molecular Dynamics Simulations, J. Phys. Chem. B 118 (2014) 12772–12780.
[91] F. Song, H. Niu, J. Fan, Q. Chen, G. Wang, L. Liu, Molecular Dynamics Study on the Coalescence and Break-up Behaviors of Ionic Droplets Under DC Electric Field, J. Mol. Liq. 312 (2020) 113195.

[92] L. Zhao, P. Choi, Molecular Dynamics simulation of the coalescence of nanometer-sized water droplets in n-heptane, J. Chem. Phys. 120 (2004).
[93] A. Nowbahar, K.A. Whitaker, A.K. Schmitt, T.-C. Kuo, Mechanistic Study of Water Droplet Coalescence and Flocculation in Diluted Bitumen Emulsions with Additives Using Microfluidics, Energy & Fuels 31 (2017) 10555–10565.
[94] D.J. McClements, J. Rao, Food-Grade Nanoemulsions: Formulation, Fabrication, Properties, Performance, Biological Fate, and Potential Toxicity, Crit. Rev. Food Sci. Nutr. 51 (2011) 285–330.
[95] B.K. Debnath, U.K. Saha, N. Sahoo, A Comprehensive Review on the Application of Emulsions as an Alternative Fuel for Diesel Engines, Renew. Sust. Energ. Rev. 42 (2015) 196–211.
[96] F. Moghaddasi, M.R. Housaindokht, M. Darroudi, M.R. Bozorgmehr, A. Sadeghi, Soybean Oil-Based Nanoemulsion Systems in Absence and Presence of Curcumin: Molecular Dynamics Simulation Approach, J. Mol. Liq. 264 (2018) 242–252.
[97] A. Ainurofiq, S. Choir, Drug Release Mechanism of Slightly Soluble Drug from Nanocomposite Matrix Formulated with Zeolite/ Hydrotalcite as Drug Carrier, Trop J Pharm Res 14 (2015) 1129–1135.
[98] T.F. Headen, E.S. Boek, G. Jackson, T.S. Totton, E.A. Müller, Simulation of Asphaltene Aggregation through Molecular Dynamics: Insights and Limitations, Energy & Fuels 31 (2017) 1108–1125.
[99] E.S. Boek, D.S. Yakovlev, T.F. Headen, Quantitative Molecular Representation of Asphaltenes and Molecular Dynamics Simulation of Their Aggregation, Energy & Fuels 23 (2009) 1209–1219.
[100] J. Wawer, J. Krakowiak, W. Grzybkowski, Apparent molar volumes, expansibilities, and isentropic compressibilities of selected electrolytes in methanol, J Chem Thermodyn 40 (2008) 1193–1199. https://doi.org/10.1016/j.jct.2008.04.008.
[101] T. Imai, A. Kovalenko, F. Hirata, Partial Molar Volume of Proteins Studied by the Three-Dimensional Reference Interaction Site Model Theory, J Phys Chem B 109 (2005) 6658–6665. https://doi.org/10.1021/jp045667c.
[102] C. Balny, P. Masson, K. Heremans, High pressure effects on biological macromolecules: from structural changes to alteration of cellular processes, Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1595 (2002) 3–10. https://doi.org/10.1016/S0167-4838(01)00331-4.
[103] R. Kitahara, H. Yamada, K. Akasaka, P.E. Wright, High Pressure NMR Reveals that Apomyoglobin is an Equilibrium Mixture from the Native to the Unfolded, J Mol Biol 320 (2002) 311–319. https://doi.org/10.1016/S0022-2836(02)00449-7.
[104] K. Kawama, Y. Fukushima, M. Ikeguchi, M. Ohta, T. Yoshidome, gr Predictor: A Deep Learning Model for Predicting the Hydration Structures around Proteins, J Chem Inf Model 62 (2022) 4460–4473. https://doi.org/10.1021/acs.jcim.2c00987.
[105] S. Karimi, H. Shekaari, I. Ahadzadeh, The sweetness response and thermophysical properties of glucose and fructose in the aqueous solution of some deep eutectic solvents at T= (288.15–318.15) K, Carbohydr Res 495 (2020) 108083. https://doi.org/10.1016/j.carres.2020.108083.
[106] E. V. Ivanov, A. V. Kustov, E.Y. Lebedeva, Solutions of Urea and Tetramethylurea in Formamide and Water: A Comparative Analysis of Volume Characteristics and Solute–Solute Interaction Parameters at Temperatures from 288.15 to 328.15 K and Ambient

Pressure, J Chem Eng Data 64 (2019) 5886–5899. https://doi.org/10.1021/acs.jced.9b00794.
[107] P. Kaur, N. Chakraborty, K.C. Juglan, H. Kumar, M. Singla, Temperature dependent physicochemical studies propylene and hexylene glycols in methanol solutions of chlorhexidine, J Mol Liq 339 (2021) 116810. https://doi.org/10.1016/j.molliq.2021.116810.
[108] Z. Yan, X. Chen, L. Liu, X. Cao, Molecular interactions between some amino acids and a pharmaceutically active ionic liquid domiphen DL-mandelic acid in aqueous medium at temperatures from 293.15 K to 313.15 K, J Chem Thermodyn 138 (2019) 1–13. https://doi.org/10.1016/j.jct.2019.06.002.
[109] Y. Liu, M. Li, C. Zhu, T. Fu, X. Gao, Y. Ma, Volumetric and viscometric properties of aqueous choline chloride + methyldiethanolamine deep eutectic solvents, J Chem Thermodyn 188 (2024) 107179. https://doi.org/10.1016/j.jct.2023.107179.
[110] S. Verma, A. Sharma, S. Maken, Volumetric, transport and acoustic properties of binary mixtures containing alkanol at 298.15–318.15 K, J Mol Liq 390 (2023) 123029. https://doi.org/10.1016/j.molliq.2023.123029.
[111] S. Verma, S. Gahlyan, P. Bhagat, M. Rani, M. Bhagat, S. Rana, V.K. Rattan, Y. Lee, S. Maken, Thermodynamic study of mesitylene + alkanol mixtures: Insights into molecular interactions, J Mol Liq 386 (2023) 122498. https://doi.org/10.1016/j.molliq.2023.122498.
[112] Y. Wang, X. Wang, L. Chai, E. Wang, X. Wei, Z. Wu, J. Zhang, Excess properties, spectral analyses, and CO2 capture performance of N-methyldiethanolamine + polyethylene glycol 300 binary system, J Mol Liq 390 (2023) 123165. https://doi.org/10.1016/j.molliq.2023.123165.
[113] C. Cadena, J.L. Anthony, J.K. Shah, T.I. Morrow, J.F. Brennecke, E.J. Maginn, Why Is CO_2 So Soluble in Imidazolium-Based Ionic Liquids?, J Am Chem Soc 126 (2004) 5300–5308. https://doi.org/10.1021/ja039615x.
[114] M.D. Macedonia, D.D. Moore, E.J. Maginn, M.M. Olken, Adsorption Studies of Methane, Ethane, and Argon in the Zeolite Mordenite: Molecular Simulations and Experiments, Langmuir 16 (2000) 3823–3834. https://doi.org/10.1021/la9912500.
[115] J.L. Anthony, E.J. Maginn, J.F. Brennecke, Solution Thermodynamics of Imidazolium-Based Ionic Liquids and Water, J Phys Chem B 105 (2001) 10942–10949. https://doi.org/10.1021/jp0112368.
[116] T.R. Josephson, R. Singh, M.S. Minkara, E.O. Fetisov, J.I. Siepmann, Partial molar properties from molecular simulation using multiple linear regression, Mol Phys 117 (2019) 3589–3602. https://doi.org/10.1080/00268976.2019.1648898.
[117] A. Rahbari, T.R. Josephson, Y. Sun, O.A. Moultos, D. Dubbeldam, J.I. Siepmann, T.J.H. Vlugt, Multiple linear regression and thermodynamic fluctuations are equivalent for computing thermodynamic derivatives from molecular simulation, Fluid Phase Equilib 523 (2020) 112785. https://doi.org/10.1016/j.fluid.2020.112785.
[118] C. Calero-Rubio, C. Strab, G. V. Barnett, C.J. Roberts, Protein Partial Molar Volumes in Multicomponent Solutions from the Perspective of Inverse Kirkwood–Buff Theory, J Phys Chem B 121 (2017) 5897–5907. https://doi.org/10.1021/acs.jpcb.7b02553.
[119] N. Dawass, P. Krüger, S.K. Schnell, J.-M. Simon, T.J.H. Vlugt, Kirkwood-Buff integrals from molecular simulation, Fluid Phase Equilib 486 (2019) 21–36. https://doi.org/10.1016/j.fluid.2018.12.027.

[120] P. Sindzingre, G. Ciccotti, C. Massobrio, D. Frenkel, Partial enthalpies and related quantities in mixtures from computer simulation, Chem Phys Lett 136 (1987) 35–41. https://doi.org/10.1016/0009-2614(87)87294-9.
[121] E.A. Guggenheim, Grand Partition Functions and So-Called ``Thermodynamic Probability'', J Chem Phys 7 (1939) 103–107. https://doi.org/10.1063/1.1750386.
[122] B. Widom, Structure of interfaces from uniformity of the chemical potential, J Stat Phys 19 (1978) 563–574. https://doi.org/10.1007/BF01011768.
[123] T.A. Al-Sahhaf, M.A. Fahim, A.S. Elkilani, Retardation of asphaltene precipitation by addition of toluene, resins, deasphalted oil and surfactants, Fluid Phase Equilib 194–197 (2002) 1045–1057. https://doi.org/10.1016/S0378-3812(01)00702-6.
[124] S.M. Hashmi, K.X. Zhong, A. Firoozabadi, Acid–base chemistry enables reversible colloid-to-solution transition of asphaltenes in non-polar systems, Soft Matter 8 (2012) 8778. https://doi.org/10.1039/c2sm26003d.
[125] D. Subramanian, A. Firoozabadi, Effect of Surfactants and Water on Inhibition of Asphaltene Precipitation and Deposition, in: Day 4 Thu, November 12, 2015, SPE, 2015. https://doi.org/10.2118/177669-MS.
[126] F.J. Ortega, F.J. Navarro, M. García-Morales, Dodecylbenzenesulfonic Acid as a Bitumen Modifier: A Novel Approach To Enhance Rheological Properties of Bitumen, Energy & Fuels 31 (2017) 5003–5010. https://doi.org/10.1021/acs.energyfuels.7b00419.
[127] Arieh Ben-Naim, Relations between thermodynamic quantities and generalized molecular distribution functions, in: Molecular Theory of Solutions, Oxford University Press, 2006: pp. 105–111.
[128] B. Schuler, S. Fatayer, G. Meyer, E. Rogel, M. Moir, Y. Zhang, M.R. Harper, A.E. Pomerantz, K.D. Bake, M. Witt, D. Peña, J.D. Kushnerick, O.C. Mullins, C. Ovalles, F.G.A. van den Berg, L. Gross, Heavy Oil Based Mixtures of Different Origins and Treatments Studied by Atomic Force Microscopy, Energy & Fuels 31 (2017) 6856–6861. https://doi.org/10.1021/acs.energyfuels.7b00805.
[129] M.L. Chacón-Patiño, S.M. Rowland, R.P. Rodgers, Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs, Energy & Fuels 31 (2017) 13509–13518. https://doi.org/10.1021/acs.energyfuels.7b02873.
[130] D. Liu, C. Li, L. li, L. Dong, X. Chen, F. Yang, G. Sun, Effect of the Interactions between Asphaltenes and Amphiphilic Dodecylbenzenesulfonic Acid on the Stability and Interfacial Properties of Model Oil Emulsions, Energy & Fuels 34 (2020) 6951–6961. https://doi.org/10.1021/acs.energyfuels.0c00833.
[131] A.B. Patel, S. Shaikh, K.R. Jain, C. Desai, D. Madamwar, Polycyclic Aromatic Hydrocarbons: Sources, Toxicity, and Remediation Approaches, Front Microbiol 11 (2020). https://doi.org/10.3389/fmicb.2020.562813.
[132] H.I. Parashkooh, C. Jian, Data Mining Guided Molecular Investigations on the Coalescence of Water-in-Oil Droplets, Energy & Fuels 36 (2022) 1811–1824. https://doi.org/10.1021/acs.energyfuels.1c03358.
[133] N. Schmid, A.P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A.E. Mark, W.F. van Gunsteren, Definition and testing of the GROMOS force-field versions 54A7 and 54B7, European Biophysics Journal 40 (2011) 843–856. https://doi.org/10.1007/s00249-011-0700-9.

[134] A.K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P.C. Nair, C. Oostenbrink, A.E. Mark, An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0, J Chem Theory Comput 7 (2011) 4026–4037. https://doi.org/10.1021/ct200196m.
[135] C. Ovalles, E. Rogel, H. Morazan, M.E. Moir, Synthesis, characterization, and mechanism of asphaltene inhibition of phosphopropoxylated asphaltenes, Fuel 180 (2016) 20–26. https://doi.org/10.1016/j.fuel.2016.03.084.
[136] Q. Feng, H. Wang, S. Zhang, J. Wang, Aggregation behavior of 1-dodecyl-3-methylimidazolium bromide ionic liquid in non-aqueous solvents, Colloids Surf A Physicochem Eng Asp 367 (2010) 7–11. https://doi.org/10.1016/j.colsurfa.2010.05.032.
[137] J.-P. Bazile, D. Nasri, H. Hoang, G. Galliero, J.-L. Daridon, Density, Speed of Sound, Compressibility and Related Excess Properties of Methane + n-Heptane at T = 303.15 K and p = 10 to 70 MPa, Int J Thermophys 41 (2020) 115. https://doi.org/10.1007/s10765-020-02694-9.
[138] R.S. Hutchings, W. Alexander Van Hook, Molar volumes in the homologous series of normal alkanes at two temperatures, Fluid Phase Equilib 21 (1985) 165–170. https://doi.org/10.1016/0378-3812(85)90067-6.
[139] G. Guevara-Carrion, R. Fingerhut, J. Vrabec, Density and Partial Molar Volumes of the Liquid Mixture Water + Methanol + Ethanol + 2-Propanol at 298.15 K and 0.1 MPa, J Chem Eng Data 66 (2021) 2425–2435. https://doi.org/10.1021/acs.jced.1c00070.
[140] Depth-First Search, in: Graph Algorithms, Cambridge University Press, 2011: pp. 46–64. https://doi.org/10.1017/CBO9781139015165.006.
[141] G. James, D. Witten, T. Hastie, R. Tibshirani, Linear Regression, in: 2021: pp. 59–128. https://doi.org/10.1007/978-1-0716-1418-1_3.
[142] E. Rogel, Effect of Inhibitors on Asphaltene Aggregation: A Theoretical Framework, Energy & Fuels 25 (2011) 472–481. https://doi.org/10.1021/ef100912b.
[143] J.L. Creek, J. Wang, J.S. Buckley, Verification of Asphaltene-Instability-Trend (ASIST) Predictions for Low-Molecular-Weight Alkanes, SPE Production & Operations 24 (2009) 360–368. https://doi.org/10.2118/125203-PA.
[144] P. Painter, B. Veytsman, J. Youtcheff, Guide to Asphaltene Solubility, Energy & Fuels 29 (2015) 2951–2961. https://doi.org/10.1021/ef502918t.
[145] M. Sedghi, L. Goual, W. Welch, J. Kubelka, Effect of Asphaltene Structure on Association and Aggregation Using Molecular Dynamics, J Phys Chem B 117 (2013) 5765–5776. https://doi.org/10.1021/jp401584u.
[146] A. Sohani, M. Zamani Pedram, S. Hoseinzadeh, Determination of Hildebrand solubility parameter of pure 1-alkanols up to high pressures, J Mol Liq 297 (2020) 111847. https://doi.org/10.1016/j.molliq.2019.111847.
[147] J. Kumełan, D. Tuma, G. Maurer, Partial molar volumes of selected gases in some ionic liquids, Fluid Phase Equilib 275 (2009) 132–144. https://doi.org/10.1016/j.fluid.2008.09.024.
[148] A. Finotello, J.E. Bara, D. Camper, R.D. Noble, Room-Temperature Ionic Liquids: Temperature Dependence of Gas Solubility Selectivity, Ind Eng Chem Res 47 (2008) 3453–3459. https://doi.org/10.1021/ie0704142.
[149] K. Akasaka, R. Kitahara, Y.O. Kamatari, Exploring the folding energy landscape with pressure, Arch Biochem Biophys 531 (2013) 110–115. https://doi.org/10.1016/j.abb.2012.11.016.

[150] J. Liggio, S.M. Li, K. Hayden, Y.M. Taha, C. Stroud, A. Darlington, B.D. Drollette, M. Gordon, P. Lee, P. Liu, A. Leithead, S.G. Moussa, D. Wang, J. O'Brien, R.L. Mittermeier, J.R. Brook, G. Lu, R.M. Staebler, Y. Han, T.W. Tokarek, H.D. Osthoff, P.A. Makar, J. Zhang, D.L. Plata, D.R. Gentner, Oil sands operations as a large source of secondary organic aerosols, Nature 534 (2016) 91–94. https://doi.org/10.1038/nature17646.
[151] C.A. Franco, N.N. Nassar, M.A. Ruiz, P. Pereira-Almao, F.B. Cortés, Nanoparticles for inhibition of asphaltenes damage: Adsorption study and displacement test on porous media, Energy and Fuels 27 (2013) 2899–2907. https://doi.org/10.1021/ef4000825.
[152] M. Sadegh Mazloom, A. Hemmati-Sarapardeh, M.M. Husein, H. Shokrollahzadeh Behbahani, S. Zendehboudi, Application of nanoparticles for asphaltenes adsorption and oxidation: A critical review of challenges and recent progress, Fuel 279 (2020). https://doi.org/10.1016/j.fuel.2020.117763.
[153] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.T. Dinh, P. De Luna, Z. Yu, A.S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.C. Lo, A. Ip, Z. Ulissi, E.H. Sargent, Accelerated discovery of CO2 electrocatalysts using active machine learning, Nature 581 (2020) 178–183. https://doi.org/10.1038/s41586-020-2242-8.
[154] C.L. Zitnick, L. Chanussot, A. Das, S. Goyal, J. Heras-Domingo, C. Ho, W. Hu, T. Lavril, A. Palizhati, M. Riviere, M. Shuaibi, A. Sriram, K. Tran, B. Wood, J. Yoon, D. Parikh, Z. Ulissi, An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage, (2020). http://arxiv.org/abs/2010.09435.
[155] Z.W. Ulissi, A.J. Medford, T. Bligaard, J.K. Nørskov, To address surface reaction network complexity using scaling relations machine learning and DFT calculations, Nat Commun 8 (2017). https://doi.org/10.1038/ncomms14621.
[156] J. Zhang, H. Bin Yang, D. Zhou, B. Liu, Adsorption Energy in Oxygen Electrocatalysis, Chem Rev 122 (2022) 17028–17072. https://doi.org/10.1021/acs.chemrev.1c01003.
[157] J.K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J.R. Kitchin, T. Bligaard, H. Jónsson, Origin of the overpotential for oxygen reduction at a fuel-cell cathode, Journal of Physical Chemistry B 108 (2004) 17886–17892. https://doi.org/10.1021/jp047349j.
[158] A. Sriram, S. Choi, X. Yu, L.M. Brabson, A. Das, Z. Ulissi, M. Uyttendaele, A.J. Medford, D.S. Sholl, The Open DAC 2023 Dataset and Challenges for Sorbent Discovery in Direct Air Capture, ACS Cent Sci 10 (2024) 923–941. https://doi.org/10.1021/acscentsci.3c01629.
[159] S. Pablo-García, S. Morandi, R.A. Vargas-Hernández, K. Jorner, Ž. Ivković, N. López, A. Aspuru-Guzik, Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks, Nat Comput Sci 3 (2023) 433–442. https://doi.org/10.1038/s43588-023-00437-y.
[160] A. Torres, J. Amaya Suárez, E.R. Remesal, A.M. Márquez, J. Fernández Sanz, C. Rincón Cañibano, Adsorption of Prototypical Asphaltenes on Silica: First-Principles DFT Simulations Including Dispersion Corrections, Journal of Physical Chemistry B 122 (2018) 618–624. https://doi.org/10.1021/acs.jpcb.7b05188.
[161] A.A. Peterson, Global optimization of adsorbate-surface structures while preserving molecular identity, Top Catal 57 (2014) 40–53. https://doi.org/10.1007/s11244-013-0161-8.

[162] J.H. Montoya, K.A. Persson, A high-throughput framework for determining adsorption energies on solid surfaces, NPJ Comput Mater 3 (2017). https://doi.org/10.1038/s41524-017-0017-z.
[163] H. Jung, L. Sauerland, S. Stocker, K. Reuter, J.T. Margraf, Machine-learning driven global optimization of surface adsorbate geometries, NPJ Comput Mater 9 (2023). https://doi.org/10.1038/s41524-023-01065-w.
[164] Y.-L. Liao, T. Smidt, Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs, (2022). http://arxiv.org/abs/2206.11990.
[165] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural Message Passing for Quantum Chemistry, (2017). http://arxiv.org/abs/1704.01212.
[166] J.-G. Kong, K.-L. Zhao, J. Li, Q.-X. Li, Y. Liu, R. Zhang, J.-J. Zhu, K. Chang, Self-supervised Representations and Node Embedding Graph Neural Networks for Accurate and Multi-scale Analysis of Materials, (2022). http://arxiv.org/abs/2211.03563.
[167] L. Gutiérrez-Gómez, J.C. Delvenne, Unsupervised network embeddings with node identity awareness, Appl Netw Sci 4 (2019). https://doi.org/10.1007/s41109-019-0197-1.
[168] J. Gasteiger, J. Groß, S. Günnemann, Directional Message Passing for Molecular Graphs, (2020). http://arxiv.org/abs/2003.03123.
[169] J. Gasteiger, F. Becker, S. Günnemann, GemNet: Universal Directional Graph Neural Networks for Molecules, (2021). http://arxiv.org/abs/2106.08903.
[170] J. Gasteiger, S. Giri, J.T. Margraf, S. Günnemann, Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules, (2020). http://arxiv.org/abs/2011.14115.
[171] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, Y. Yu, HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition, n.d.
[172] L. Jian-min, Y. Min-hua, X. Yu, Hierarchical features learning with convolutional neural networks based on aircraft recognition on images from remote sensing image, in: 2016 IEEE International Conference on Consumer Electronics-China (ICCE-China), IEEE, 2016: pp. 1–5. https://doi.org/10.1109/ICCE-China.2016.7849736.
[173] Y. Yang, M. Liu, J.R. Kitchin, Neural network embeddings based similarity search method for atomistic systems, Digital Discovery 1 (2022) 636–644. https://doi.org/10.1039/d2dd00055e.
[174] A. Kolluru, M. Shuaibi, A. Palizhati, N. Shoghi, A. Das, B. Wood, C.L. Zitnick, J.R. Kitchin, Z.W. Ulissi, Open Challenges in Developing Generalizable Large-Scale Machine-Learning Models for Catalyst Discovery, ACS Catal 12 (2022) 8572–8581. https://doi.org/10.1021/acscatal.2c02291.
[175] M. Schaarschmidt, M. Riviere, A.M. Ganose, J.S. Spencer, A.L. Gaunt, J. Kirkpatrick, S. Axelrod, P.W. Battaglia, J. Godwin, Learned Force Fields Are Ready For Ground State Catalyst Discovery, (2022). http://arxiv.org/abs/2209.12466.
[176] L. Barroso-Luque, M. Shuaibi, X. Fu, B.M. Wood, M. Dzamba, M. Gao, A. Rizvi, C.L. Zitnick, Z.W. Ulissi, Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models, (2024). http://arxiv.org/abs/2410.12771.
[177] L. Kollias, R. Rousseau, V.-A. Glezakou, M. Salvalaglio, Understanding MOF nucleation from solution with Evolving Graphs, (2022). https://doi.org/10.26434/chemrxiv-2021-00kkd-v2.

[178] F.M. Dietrich, X.R. Advincula, G. Gobbo, M.A. Bellucci, M. Salvalaglio, Machine Learning Nucleation Collective Variables with Graph Neural Networks, J Chem Theory Comput 20 (2024) 1600–1611. https://doi.org/10.1021/acs.jctc.3c00722.
[179] B. Liu, M. Xue, Y. Qiu, K.A. Konovalov, M.S. O'Connor, X. Huang, GraphVAMPnets for uncovering slow collective variables of self-assembly dynamics, J Chem Phys 159 (2023). https://doi.org/10.1063/5.0158903.
[180] C. Yoshikawa, D.A. Nguyen, T. Nakaji-Hirabayashi, I. Takigawa, H. Mamitsuka, Graph Network-Based Simulation of Multicellular Dynamics Driven by Concentrated Polymer Brush-Modified Cellulose Nanofibers, ACS Biomater Sci Eng 10 (2024) 2165–2176. https://doi.org/10.1021/acsbiomaterials.3c01888.
[181] S.-H. Lee, J. Li, V. Olevano, B. Sklénard, Equivariant graph neural network interatomic potential for Green-Kubo thermal conductivity in phase change materials, Phys Rev Mater 8 (2024) 033802. https://doi.org/10.1103/PhysRevMaterials.8.033802.
[182] G. Hu, M.I. Latypov, AnisoGNN: graph neural networks generalizing to anisotropic properties of polycrystals, (2024). http://arxiv.org/abs/2401.16271.
[183] J. Ho, A. Jain, P. Abbeel, Denoising Diffusion Probabilistic Models, (2020). http://arxiv.org/abs/2006.11239.
[184] S. Naderiparizi, X. Liang, B. Zwartsenberg, F. Wood, Constrained Generative Modeling with Manually Bridged Diffusion Models, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 19607–19615. https://doi.org/10.1609/aaai.v39i18.34159.
[185] S. Naderiparizi, X. Liang, B. Zwartsenberg, F. Wood, Don't be so negative! Score-based Generative Modeling with Oracle-assisted Guidance, (2023). http://arxiv.org/abs/2307.16463.
[186] B. Wang, S. Zheng, J. Wu, J. Li, F. Pan, Inverse design of catalytic active sites via interpretable topology-based deep generative models, NPJ Comput Mater 11 (2025) 147. https://doi.org/10.1038/s41524-025-01649-8.

7. Appendix

7.1. Appendix for Chapter 2

7.1.1. Cutoff Distance for Water Oxygen Atoms

Figure A7.1-1 shows the radial distribution function (RDF, g(r)) for water oxygen atoms in systems A16_W300, A16_W600, and A16_W900, which was used to determine the cutoff distance for the analysis of water droplets (see section 2.3). From Figure A7.1-1, RDF has its first shallow at $r = 0.32$ nm; therefore, two water molecules are defined to be in the same droplet if the distance between their oxygen atoms is less than 0.32 nm.

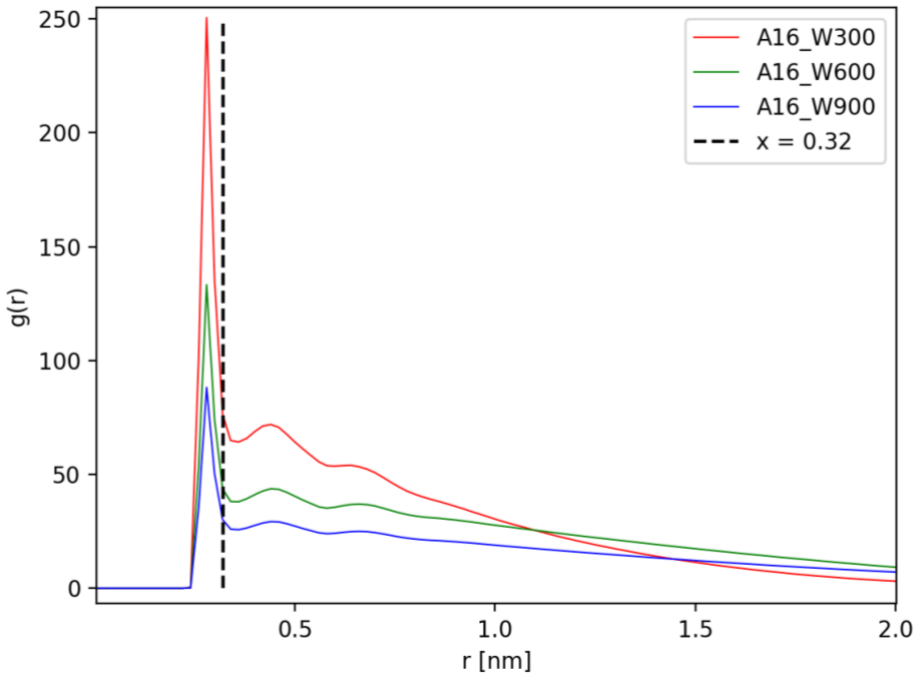

Figure A7.1-1 RDF for distances between water oxygen atoms in A16 systems.

7.1.2. Details on the Implementation of Droplet Analysis

The implementation of our analysis method is based on Python (version 3.9) with library NetworkX (version 2.5.1). The chart shown in Figure A 7.1.2 represents the flow of our method.

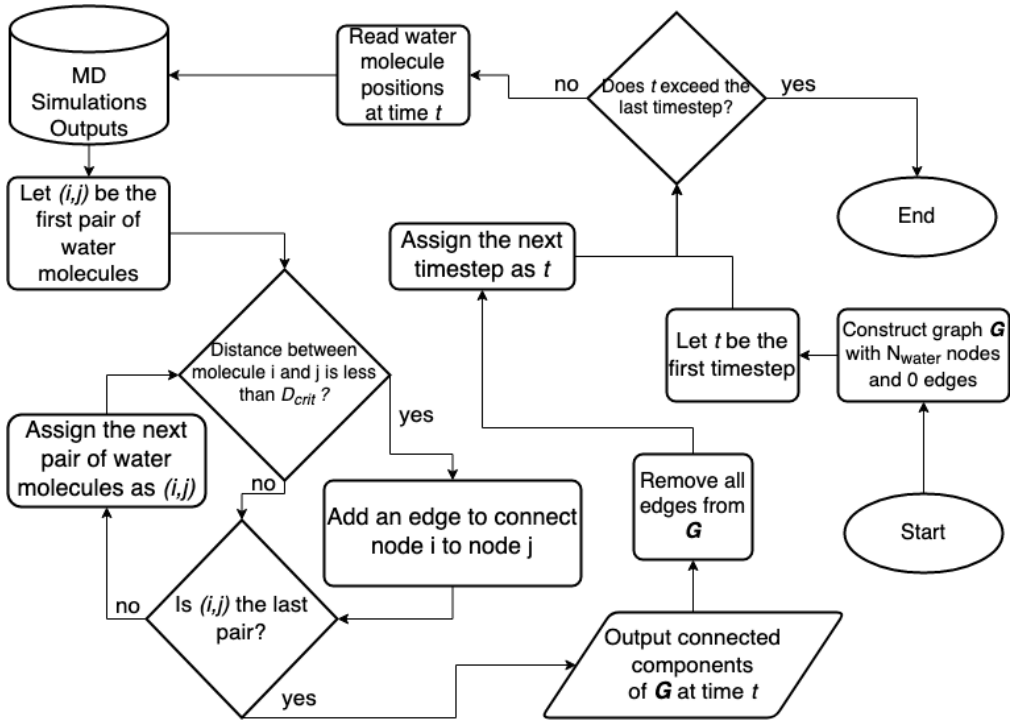

Figure A7.1-2 Flowchart for implementing the droplet analysis method.

7.1.3. Validation of Our Customized Tool for Droplet Analysis

Figure A7.1.3 shows the comparison on the number of droplets obtained using our in-house developed tool with that from the standard routine *gmx clustsize* in GROMACS. As it can be seen, for all cases, the results essentially overlap with one another, confirming the accuracy of our method. The slight difference between *gmx clustsize* and our method lies in the distance calculations. The *gmx clustsize* module in GROMACS performs the cluster analysis based on the distance between molecules. For two water molecules, the distance is calculated between any atom from one of the two molecule and any atom from the other water molecule, leading to 9 pairs of distances between two water molecules. The two water molecules are determined to be in the same droplet as long as one of the 9 pairs of distances is less than a cutoff distance. On the other hand, the in-house developed method calculated the distance between oxygen atoms in water molecules. Using oxygen-oxygen distances significantly reduced the calculation load and meanwhile maintains satisfactory accuracies

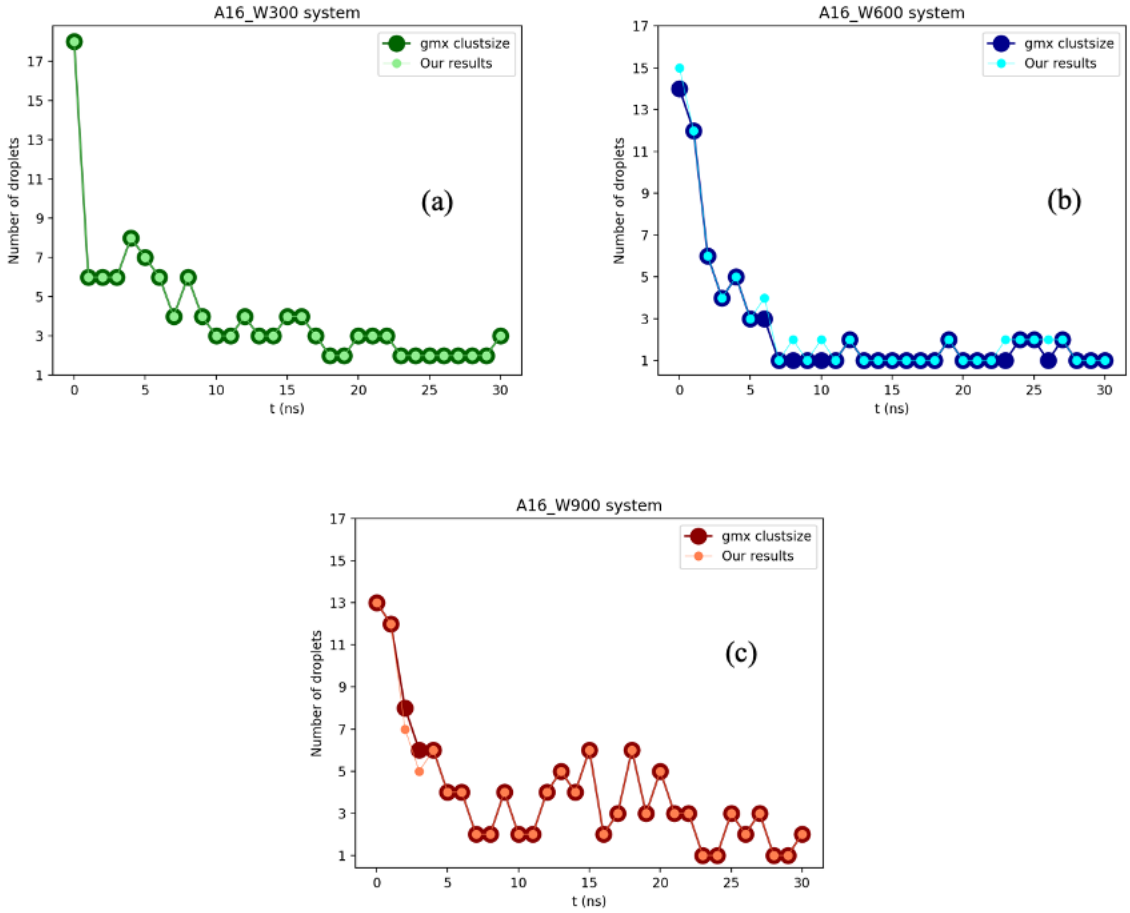

Figure A7.1-3 Comparison on the number of droplets obtained from our customized tool with that of GROMACS module *gmx clustsize* for systems (a) A16_W300, (b) A16_W600, and (c) A16_W900.

7.1.4. Parallel Stackings Identified in Each System and Additional Configurations

Table 7.1.1 summarizes the number of VO-79 molecules in each stacking block (sorted by sizes), for all systems averaged over the last 5 ns of the production phase. The number in the bracket indicates the occurrence of the corresponding block. For instance, in system A16_W0, 11(×1) means this type of stacking bock has 11 VO-79 molecules, and the total count of such a stacking block in 1; meanwhile in system A16_W300, 1(×5) means this stacking block only has 1 VO-79 molecule, and the total count of such a stacking block is 5. As it can be seen, system A16_W0 has the largest stacking block among all systems (including A80_W600 and A160_W600), because of the absence of water droplets that interfere with the $\pi - \pi$ interactions

between polyaromatic cores. With increasing the number of water molecules, the size of stacking block is first decreased, and then increased (see Figure 2.4).

Table A7.1-1 Stacking blocks in all simulated systems

System	A16_W0	A16_W300	A16_W600	A16_W900	A80_W600	A160_W600
1st block	11(×1)	5(×1)	3(×1)	7(×1)	7(×1)	10(×1)
2nd block	4(×1)	4(×1)	1(× 13)	2(×1)	6(×1)	9(×1)
3rd block	1(×1)	2(×1)	-	1(× 7)	5(× 2)	8(×1)
4th block	-	1(×5)	-	-	3(× 3)	7(×1)
5th block	-	-	-	-	2(×1)	5(×1)
6th block	-	-	-	-	1(× 46)	4(×5)
7th block	-	-	-	-	-	3(× 11)
8th block	-	-	-	-	-	2(× 9)
9th block	-	-	-	-	-	1(×50)
Number of stacking block	3	8	14	9	54	80

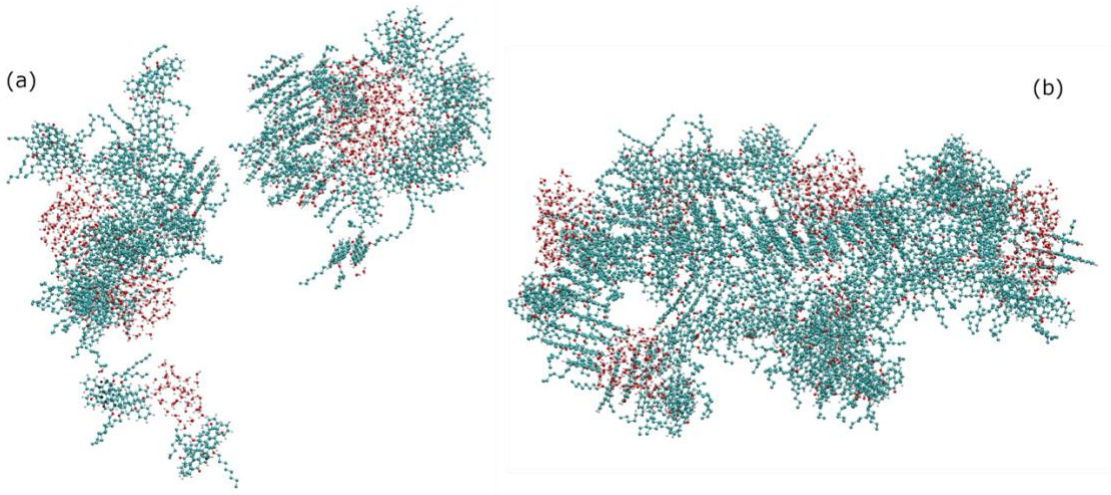

Figure A7.1-4 Final configurations of VO-79 and water molecules at the end of the simulation in (a) system A80_W600, and (b) system A160_W600. In both systems, more than 1 droplet is in presence.

7.1.5. Distribution of Water Droplet Sizes

In the main text, we focused on the first half of the equilibration phase when calculating coalescence rates. Table 7.1.2 shows the distribution of water droplets at at $t = 0.5$ ns of equilibration. As can be seen, with increasing numbers of water molecules, fewer water droplets

were formed with larger sizes. On the other hand, with increasing numbers of asphaltene molecules, more water droplets were formed with smaller sizes. These observations are consistent with coalescence rates discussed in the main text.

Table A7.1-2 Water droplets in all simulated systems at t=0.5 ns of the equilibration phase

System	A16_W300	A16_W600	A16_W900	A80_W600	A160_W600
1st - 5th Droplets	49(×1) 35(×1) 28(×1) 22(×2) 18(×1)	93(×1) 54(×1) 47(×1) 45(×1) 41(×1)	131(×1) 125(×1) 102(×1) 73(×1) 71(×1)	42(×1) 36(×1) 29(×1) 28(×1) 27(×1)	35(×2) 32(×1) 30(×1) 22(×1) 20(×1)
6th - 10th droplets	16(×1) 11(×1) 10(×2) 9(×1) 8(×3)	30(×2) 26(×1) 25(×2) 22(×1) 21(×1)	70(×1) 64(×1) 36(×1) 35(×1) 29(×1)	25(×2) 22(×3) 21(×1) 19(×3) 18(×2)	19(×3) 18(×2) 16(×1) 15(×1) 14(×3)
11th - 15th droplets	7(×2) 6(×2) 5(×1) 4(×1) 2(×1)	19(×1) 18(×4) 15(×1) 12(×1) 11(×1)	28(×1) 27(×1) 22(×2) 21(×1) 20(×1)	17(×3) 16(×1) 15(×1) 14(×1) 13(×1)	13(×3) 12(×2) 11(×1) 10(×2) 9(×1)
16th - 20th droplet	1(×9) - - - -	5(×1) 4(×1) 1(×3) - -	17(×1) 6(×1) 1(×1) - -	12(×1) 10(×1) 8(×2) 7(×2) 5(×2)	8(×4) 7(×1) 6(×4) 5(×5) 4(×6)
21st - 25th droplet	-	-	-	4(×4) 3(×3) 2(×1) 1(×10) -	3(×4) 2(×4) 1(×25) - -
Number of Droplets	30	25	19	49	81

To determine time intervals for the production phase, we tracked the time at which each system only has one droplet left. For systems A16_W300, A16_W600, and A16_W900, the simulation time is 8.51 ns, 6.14 ns, and 11.47 ns, respectively. For systems A80_W600 and A160_W600, given the abundance of VO-79 molecules, they both have more than 1 droplet at the

end of the 100 ns simulation (see Figure A 7.1.3). As a comparison, Table 7.1.3 lists the distribution of droplet sizes at $t = 6.14$ ns for all systems. Again, as can be seen, with increasing numbers of water molecules, fewer water droplets were formed with larger sizes; on the other hand, with increasing numbers of asphaltene molecules, more water droplets were formed with smaller sizes.

Table A7.1-3 Water droplets in all simulated systems at t=6.14 ns of the production phase

System	A16_W300	A16_W600	A16_W900	A80_W600	A160_W600
1st Droplet	121(×1)	600(×1)	773(×1)	160(×1)	159(×1)
2nd droplet	64(×1)	-	126(×1)	124(×1)	106(×1)
3rd droplet	43(×1)	-	1(×1)	96(×1)	104(×1)
4th droplet	12(×1)	-	-	89(×1)	76(×1)
5th droplet	1(×1)	-	-	64(×1)	68(×1)
6th droplet	-	-	-	48(×1)	25(×2)
7th droplet	-	-	-	14(×1)	21(×1)
8th droplet	-	-	-	2(×1)	9(×1)
9th droplet	-	-	-	1(×3)	1(×7)
Number of Droplets	5	1	3	11	15

7.1.6. Mode Quantifications Based on Streak Lengths

Our simulation trajectories are saved per 10 ps, which is the time interval of consecutive frames used in Figures 9 and 10. To quantitatively compare, all individual streak lengths (unit: 10 ps) are determined and summarized in tables S4-8 with the corresponding occurrences. For example, in system A16_W300 (Table 7.1.4), at the equilibration phase, the stalling mode contains: 6×10 ps streak (3 occurrences), 3×10 ps streak (3 occurrences), and 2×10 ps streak (3 occurrence); correspondingly, the growth mode only contains two types of streaks: 2×10 ps streak

(1 occurrence) and 1×10 ps streak (7 occurrences). Thus, the percentage of stalling is: $(3 \times 6 \times 10\ \text{ps} + 3 \times 3 \times 10\ \text{ps} + 3 \times 2 \times 10\ \text{ps})/500\text{ps} = 0.66$, and the growing percentage is: $(2 \times 10\ \text{ps} + 7 \times 1 \times 10\ \text{ps})/500\text{ps} = 0.18$.

Table A7.1-4 Number and length of all stalling and growing modes in system A16_W300

Equilibration Phase				**Production Phase**			
Stalling		Growing		Stalling		Growing	
number	length (10 ps)	number	length (10 ps)	number	length (10 ps)	number	length (10 ps)
3	6	1	2	1	17	2	2
3	3	7	1	1	13	113	1
3	2			3	12		
				1	11		
				2	10		
				7	9		
				2	8		
				6	7		
				8	6		
				10	5		
				15	4		
				12	3		
				25	2		
				25	1		
Max	6	Max	2	Max	17	Max	2
Min	1	Min	1	Min	1	Min	1
Average	3.4	Average	1.1	Average	4.1	Average	1

Table A7.1-5 Number and length of all stalling and growing modes in system A16_W600

Equilibration Phase				**Production Phase**			
Stalling		Growing		Stalling		Growing	
number	length (10 ps)	number	length (10 ps)	number	length (10 ps)	number	length (10 ps)
1	9	2	2	1	21	1	4
1	8	7	1	1	16	7	2
1	6			1	15	90	1
1	5			2	14		
1	2			2	13		
4	1			2	12		
				1	11		
				1	10		
				3	9		
				3	8		
				4	7		
				6	6		
				10	5		
				12	4		
				15	3		
				20	2		
				27	1		
Max	9	Max	2	Max	21	Max	4
Min	1	Min	1	Min	1	Min	1
Average	3.5	Average	1.2	Average	4.3	Average	1.1

Table A7.1-6 Number and length of all stalling and growing modes in system A16_W900

Equilibration Phase				**Production Phase**			
Stalling		Growing		Stalling		Growing	
number	length (10 ps)	number	length (10 ps)	number	length (10 ps)	number	length (10 ps)
1	6	1	2	1	39	1	3
1	4	11	1	1	20	6	2
1	3			1	14	98	1
4	2			2	13		
5	1			1	12		
				2	11		
				2	10		
				4	9		
				3	8		
				2	7		
				10	6		
				12	5		
				10	4		
				17	3		
				19	2		
				19	1		
Max	6	Max	2	Max	39	Max	3
Min	1	Min	1	Min	1	Min	1
Average	2	Average	1.1	Average	4.6	Average	1.1

Table A7.1-7 Number and length of all stalling and growing modes in system A80_W600

Equilibration Phase				**Production Phase**			
Stalling		Growing		Stalling		Growing	
number	length (10 ps)	number	length (10 ps)	number	length (10 ps)	number	length (10 ps)
1	6	3	2	1	35	2	2
2	5	7	1	1	28	90	1
6	2			1	19		
2	1			1	16		
				2	13		
				2	12		
				2	10		
				6	9		
				4	8		
				4	7		
				8	6		
				11	5		
				7	4		
				19	3		
				15	2		
				9	1		
Max	6	Max	2	Max	35	Max	2
Min	1	Min	1	Min	1	Min	1
Average	2.6	Average	1.3	Average	5.4	Average	1

Table A7.1-8 Number and length of all growth modes in system A160_W600

Equilibration Phase				**Production Phase**			
Stalling		Growing		Stalling		Growing	
number	length (10 ps)	number	length (10 ps)	number	length (10 ps)	number	length (10 ps)
1	12	1	4	1	46	5	2
1	6	1	2	1	28	75	1
2	3	6	1	3	19		
2	2			1	18		
3	1			1	17		
				2	15		
				1	12		
				1	11		
				4	10		
				3	9		
				4	8		
				5	7		
				5	6		
				10	5		
				8	4		
				9	3		
				12	2		
				9	1		
Max	12	Max	4	Max	46	Max	2
Min	1	Min	1	Min	1	Min	1
Average	3.2	Average	1.5	Average	6.5	Average	1.1

From these tables, it can be seen that stalling mode dominates in each system, and comparing production with equilibration, while growing modes have similar streak lengths, those of stalling modes are more diversified in the production phase. The minimum, maximum, and average streak lengths are also plotted in Figure A7.1.5 and A7.1.6. From these two figures, again, it can be seen that while streak lengths (unit: 10 ps) of growing modes remain more or less the same (no more than 4 in all systems), streak lengths of growing modes are significantly diversified in the production phase compared to the equilibration phase.

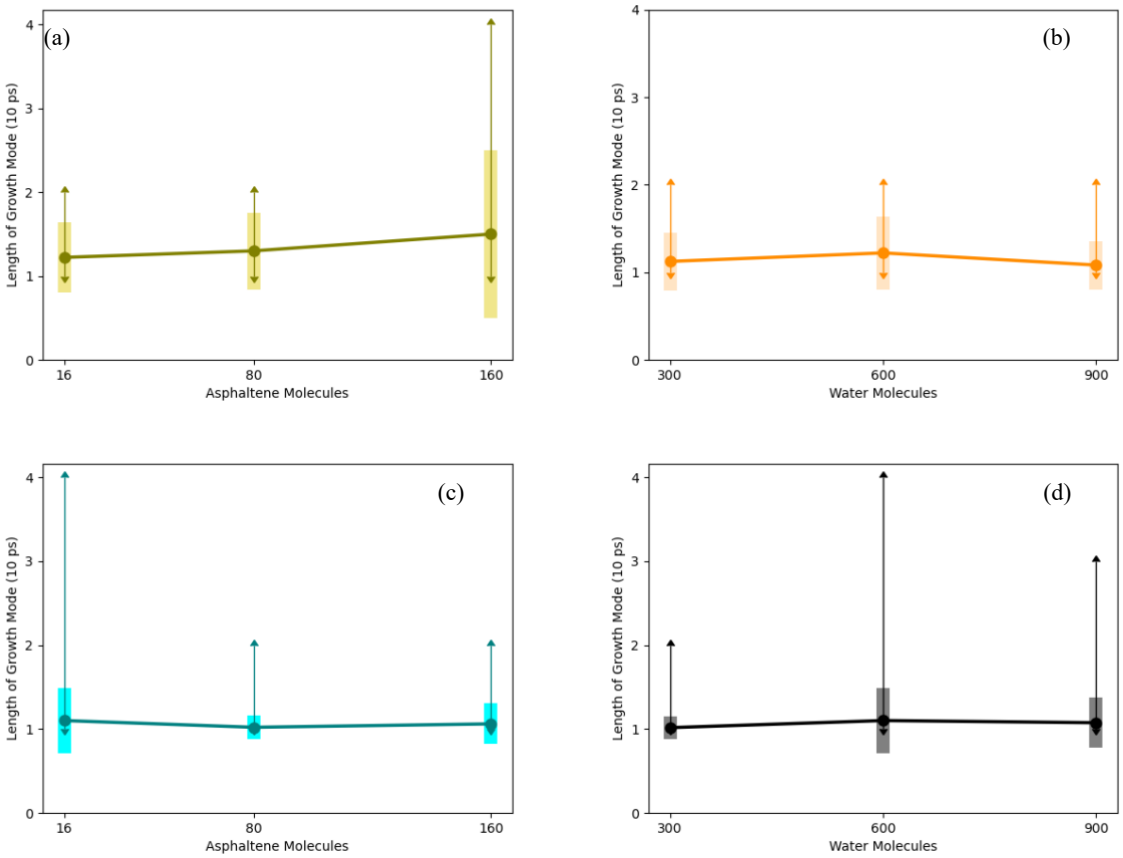

Figure A7.1-5 Minimum (upper arrows), maximum (lower arrows), average (markers) and standard deviation (error bars) of steak lengths for the growing mode in equilibration (a and b) and production phases (c and d). The left column (a and c) is for systems A16_W600, A80_W600, A160_W600, and the right column (b and d) is for systems A16_W300, A16_W600, A16_W900, respectively.

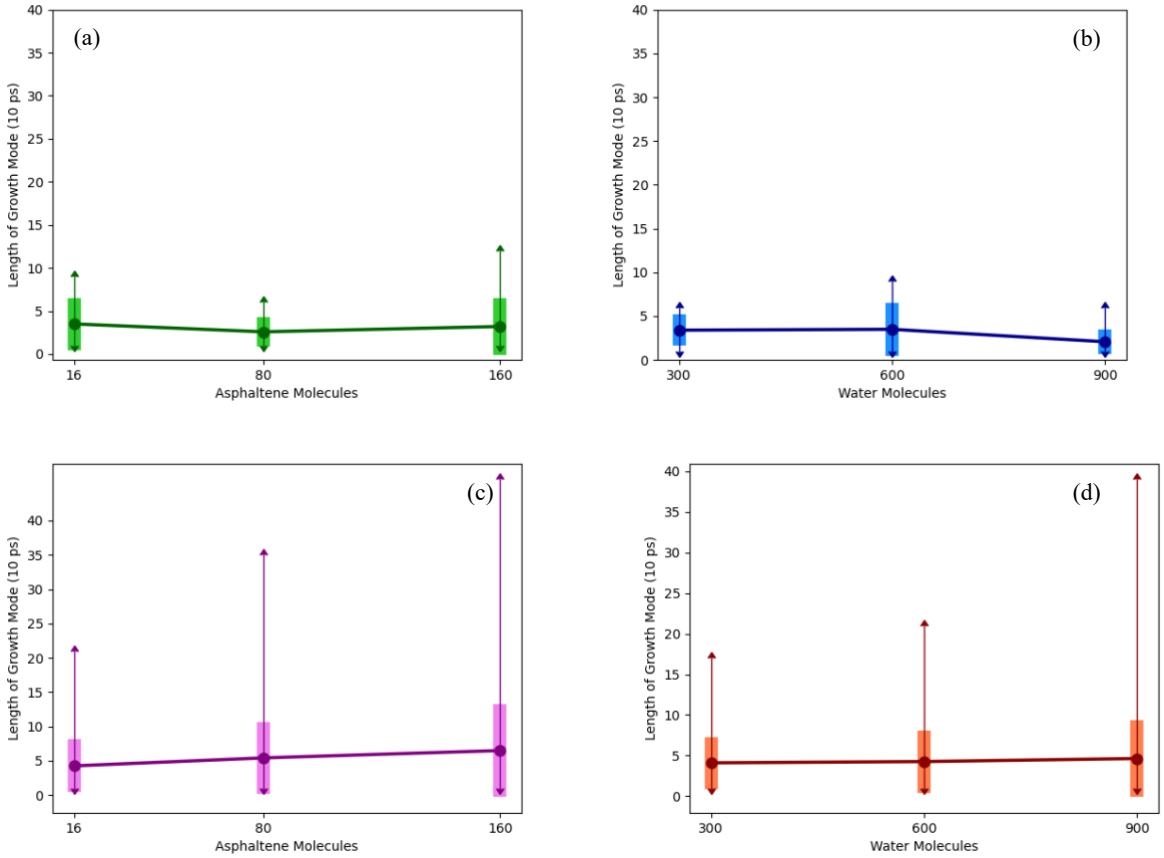

Figure A7.1-6 Minimum (upper arrows), maximum (lower arrows), average (markers) and standard deviation (error bars) of streak lengths for stalling modes in equilibration (a and b) and production phases (c and d). The left column (a and c) is for systems A16_W600, A80_W600, A160_W600, and the right column (b and d) is for systems A16_W300, A16_W600, A16_W900, respectively.

7.1.7. Coalescence Modes Quantified by Using 70% as the Criteria

To confirm the nucleus effect of the largest droplets, we increased the criteria used in the growing mode from 50% to 70 % (see section 2.3.2 for the definition of growing modes), and Figure A7.1.4 compares the results obtained using these two criteria. As can be seen, all trends observed before still hold except Figure A7.1.3(c), confirming the dominant role of the largest droplets as the nucleus sites. The relatively drastic change in system A160_W600 might be caused by sizes of water droplets. As shown in Tables A7.1.2 and A7.1.3, system A160_W600 has much more droplets with much smaller sizes compared to other systems. In return, the role of the largest droplet might be weakened.

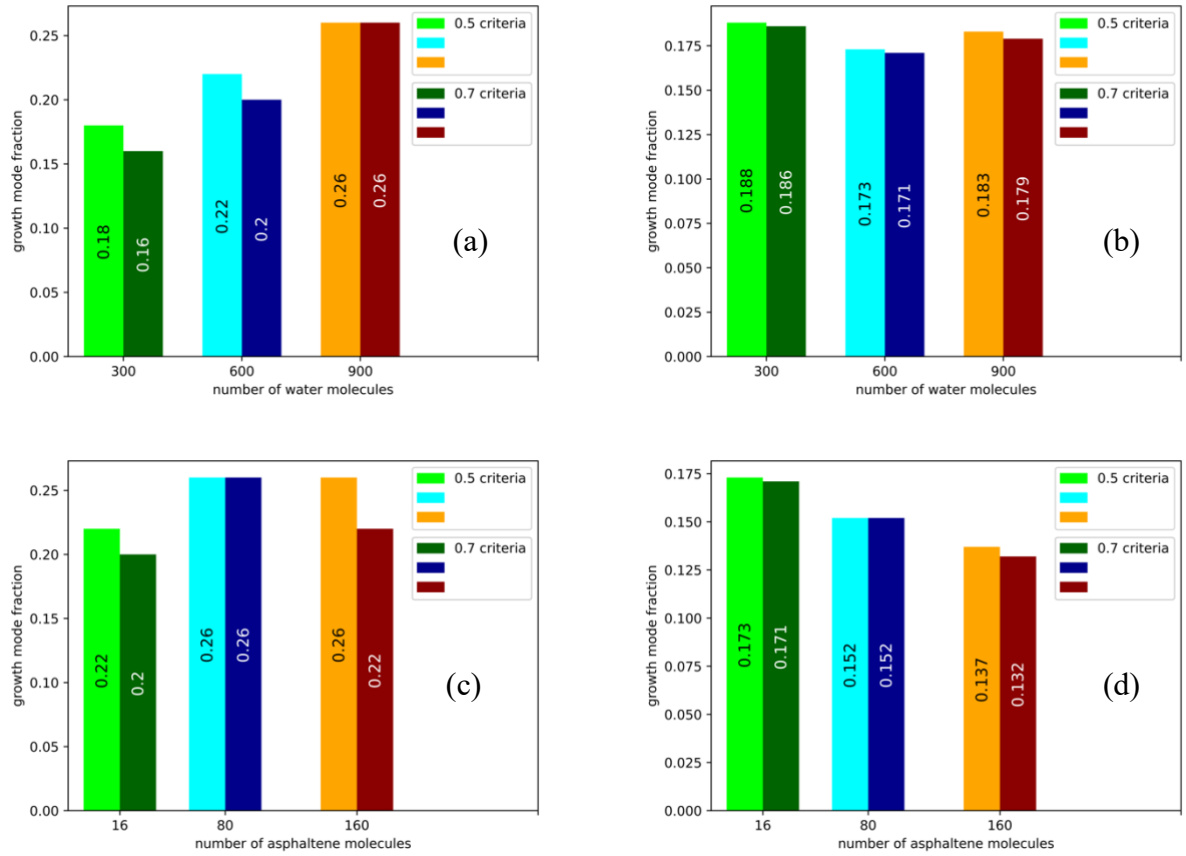

Figure A7.1-7 Comparison on criteria used to define the growing mode: (a) equilibration and (b) production of systems A16_W300, A16_W600, and A16_W900; and (c) equilibration and (d) production of systems A16_W600, A80_W600 and A160_W600.

7.1.8. Non-Nucleus Modes Observed in A16 Systems

Figure A7.1.8 show the non-nucleus mode (i.e., neither growing or stalling modes) observed in systems A16_W300, A16_W600, and A16_W900.

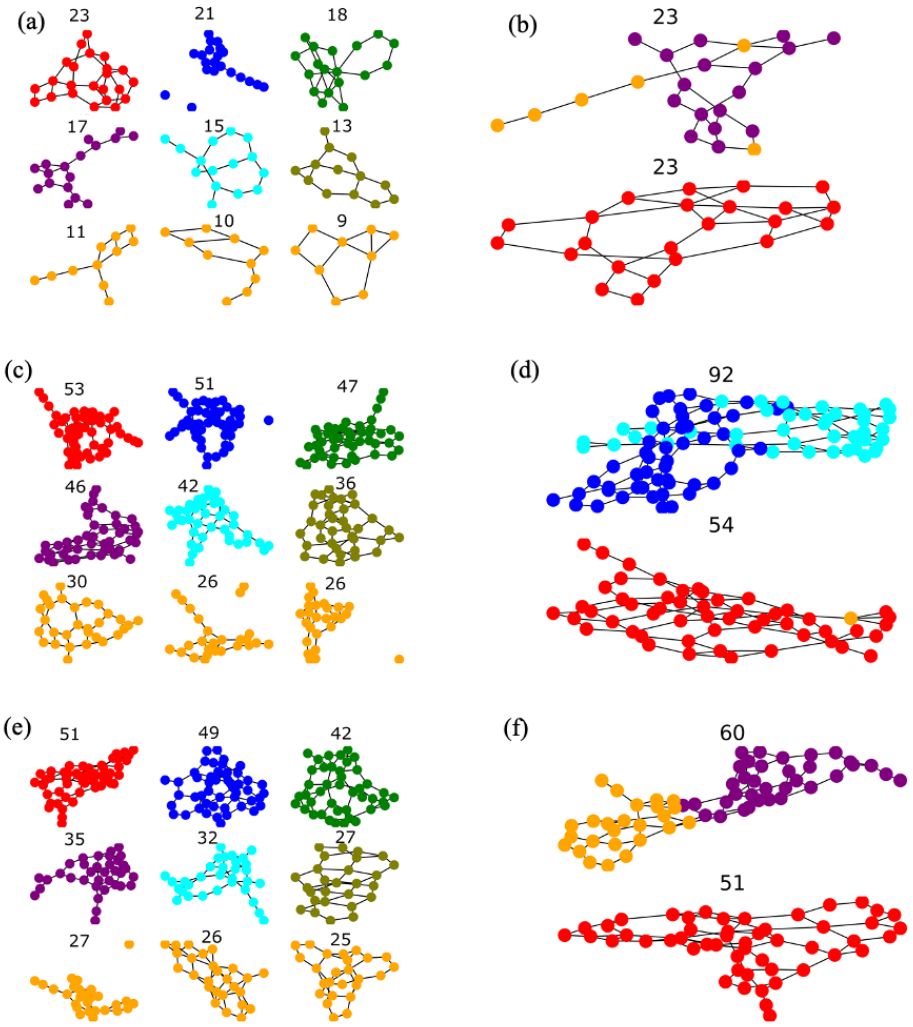

Figure A7.1-8 Snapshots of non-nucleus modes: (a) - (b) in system A16_W300, (c) - (d) in system A16_W600, and (e) - (f) in system A16_W900. Briefly, smaller droplets in (a), (c), and (e) merged into the droplets in (b), (d), and (f), respectively. Furthermore, these newly formed droplets in (b), (d), and (f) surpass the sizes of largest droplets at previous steps, i.e., the ones shown in (a), (c), and (e), and became the largest droplet in the corresponding system.

7.2. Appendix for Chapter 3

7.2.1. Volume Sampling Method

Figure A7.2.1 is included to illustrate the process and rationale behind setting the number of control volumes and their selection. As mentioned in the main text, control volume sizes are randomly chosen within a specific interval, but not all of them are utilized for obtaining molecular composition samples. Instead, the focus is on the structure of the precipitants, and the number of their clusters is determined at a given snapshot instance. Subsequently, a proportional number of control volumes are selected for sampling based on this cluster count. The snapshots chosen for sampling are those with the highest abundance of precipitants. This approach, as depicted on the right side of Figure A7.2.1, ensures that the flocculates are adequately represented within the sampled control volumes.

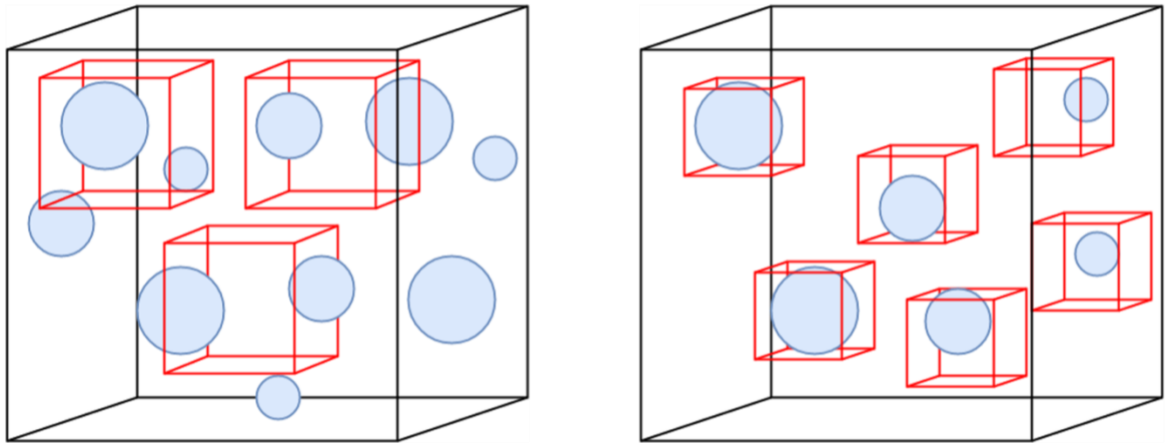

Figure A7.2-1 The visual representation on the left shows a random selection of volume samples, while the one on the right demonstrates a method where control volumes are selected in proportion to the number of flocculates in the system, and then further refined by choosing the control volumes with the highest number of solvents.

7.2.2. Validation

Here, first we explain the process of calculating partial molar volumes from the excess molar volumes (EMV) according to the method implemented in Journal of Molecular Liquids 386 (2023) 122498. From Redlich-Kister polynomial fitting to EMV values, the coefficients (A_is) are obtained.

$$V_m^E = x_1 x_2 \sum_{i=0}^{n} A_i (x_1 - x_2)^i = x_1(1 - x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i$$

For binary systems, by substituting $x_1 + x_2 = 1$:

$$V_m^E = x_1(1 - x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i$$

Now we substitute V_m^E in equations (5) of the main text, we start by calculating and simplifying $(\frac{\partial V_m^E}{\partial x_1})_{P,T}$:

$$\frac{\partial V_m^E}{\partial x_1} = (1 - x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i - x_1 \sum_{i=0}^{n} A_i (2x_1 - 1)^i + x_1(1 - x_1) \sum_{i=0}^{n} A_i (2i)(2x_1 - 1)^{i-1}$$
$$= (1 - 2x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i + 2x_1(1 - x_1) \sum_{i=0}^{n} A_i (i)(2x_1 - 1)^{i-1}$$

Therefore, the third term of the equation (5) related to $\bar{V}_1$ would be:

$$x_2 (\frac{\partial V^E}{\partial x_1})_{P,T} = (1 - x_1)(\frac{\partial V^E}{\partial x_1})_{P,T}$$
$$= (1 - x_1)(1 - 2x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i + 2x_1(1 - x_1)^2 \sum_{i=0}^{n} A_i (i)(2x_1 - 1)^{i-1}$$

Therefore:

$$\bar{V}_1 = V_m^E + x_2 (\frac{\partial V^E}{\partial x_1})_{P,T}$$
$$= x_1(1 - x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i + (1 - x_1)(1 - 2x_1) \sum_{i=0}^{n} A_i (2x_1 - 1)^i$$
$$+ 2x_1(1 - x_1)^2 \sum_{i=0}^{n} A_i (i)(2x_1 - 1)^{i-1}$$
$$= (1 - x_1)^2 \sum_{i=0}^{n} A_i (2x_1 - 1)^i + 2x_1(1 - x_1)^2 \sum_{i=0}^{n} A_i (i)(2x_1 - 1)^{i-1}$$

Similarly, for the third term of equation (5) related to $\bar{V}_2$, we would have:

$$x_1(\frac{\partial V^E}{\partial x_1})_{P,T} = x_1(1-2x_1)\sum_{i=0}^{n} A_i(2x_1-1)^i + 2x_1^2(1-x_1)\sum_{i=0}^{n} A_i(i)(2x_1-1)^{i-1}$$

And Finally:

$$\bar{V}_2 = V_m^E - x_1(\frac{\partial V^E}{\partial x_1})_{P,T} = x_1^2\sum_{i=0}^{n} A_i(2x_1-1)^i - 2x_1^2(1-x_1)\sum_{i=0}^{n} A_i(i)(2x_1-1)^{i-1}$$

The partial molar volumes calculated from the method above by Verma et al. are listed in Table A7.2-1. These values were used for interpolation of the PMV at composition of the validation systems introduced in table 3-1.

Table A7.2-1 The experimental values from Journal of Molecular Liquids 386 (2023) 122498, for binary mixtures of mesitylene (1) + isopropanol (2) at T=298.15 K.

x_1	$\bar{V}_1$ $[cm^3/mol]$	$\bar{V}_2$ $[cm^3/mol]$
0.1275	141.65	76.99
0.1934	141.26	77.07
0.2702	140.86	77.17
0.3654	140.45	77.29
0.4585	140.13	77.38
0.5657	139.85	77.47
0.685	139.65	77.57
0.8251	139.53	77.75
0.8975	139.51	77.89

7.2.3. The Partial Molar Volumes of the Solvents

To validate our calculations, we present the partial molar volume (PMV) results for the solvents in the systems. Figure A7.2.2 illustrates the PMV values for solvents in both A16_Pn and A16_Wt systems. The circles represent values obtained from each time sampling method, while the diamonds with error bars represent the average and standard deviation, respectively. Additionally, Figure A7.2.2 includes the molar volume of pure solvents (water and pentane) for reference.

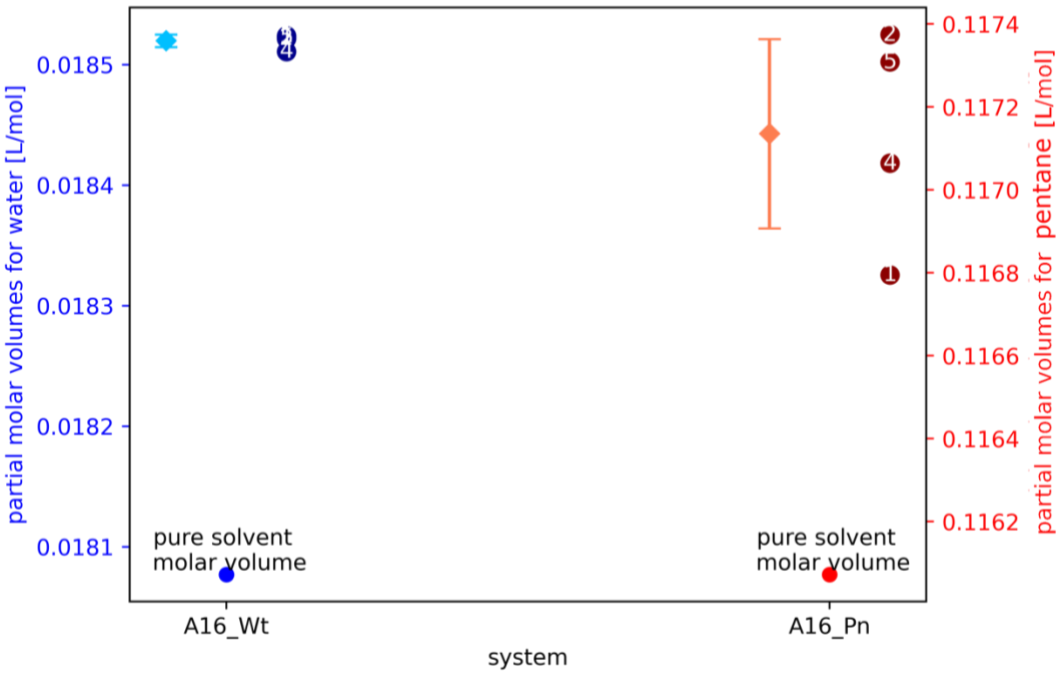

Figure A7.2-2 Comparing the calculated PMV and molar volume of solvents in A16_Wt and A16_Pn

Furthermore, Figure A7.2.3 displays the PMV of solvents in DBSA systems, including Hn0, Hn60, Hn120, Hn180, Pn0, Pn60, Pn120, and Pn180. The plot reveals that the PMV values for each solvent exhibit convergence. As elaborated in the main text, smaller molecules like solvents pose no issues with representation in the volume samples, as they are adequately present and do not encounter the challenges faced by larger molecules in smaller control volumes.

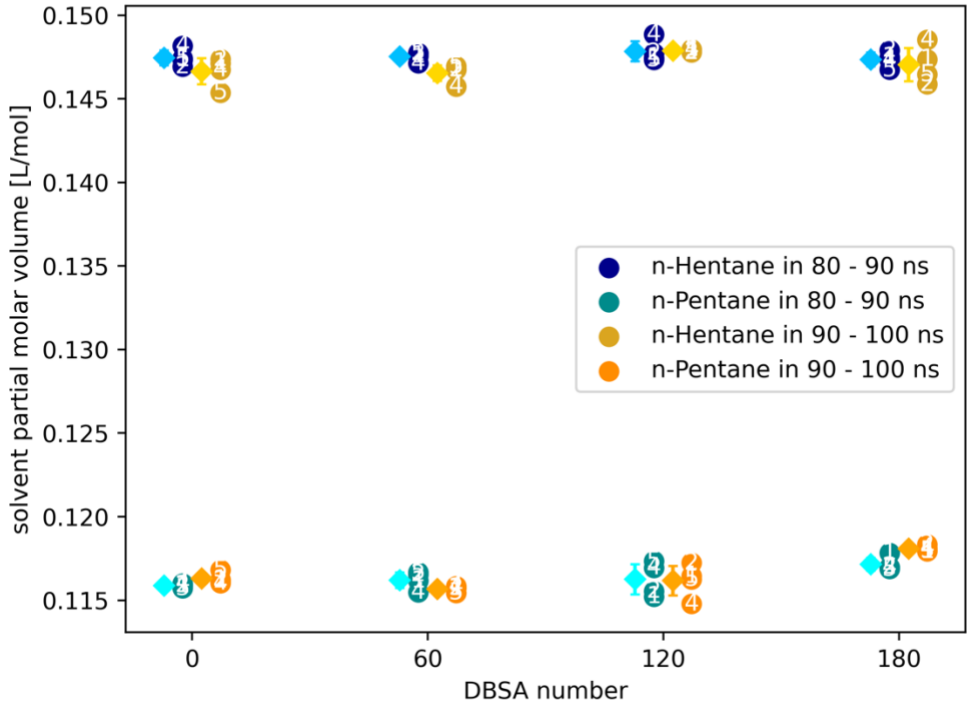

Figure A7.2-3 The PMV of heptane and pentane in different DBSA systems for two separate time intervals.

7.2.4. Snapshots of the Systems at the Final Stage of their Simulation

In Figure A7.2.4, the configuration of the VO-79 molecules at the simulation's final stage is presented for both A16_Wt and A16_Pn systems. As explained in the main text, in pentane, the majority of asphaltenes adopt their typical rod-like structure. Conversely, in water, the side chains tend to position themselves further away from the water molecules, resulting in a central placement within the sphere-like configuration formed by the polyaromatic cores.

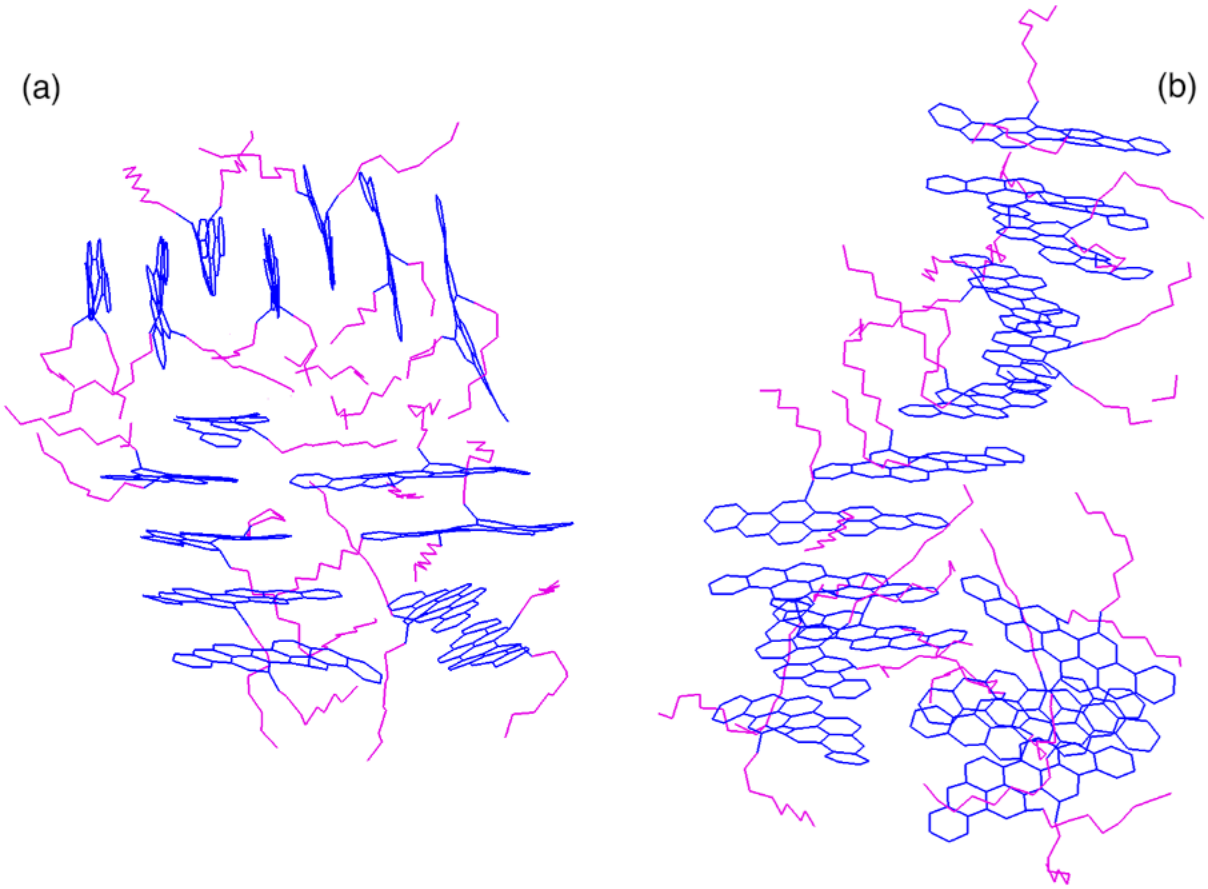

Figure A7.2-4 The configuration of VO-79 molecules in systems A16_Wt (a) and A16_Pn (b)

7.2.5. The Radial Distribution Function Plots for Pentane Systems

As shown in Figure 3-9, the solvent molecules tend to position themselves closer to the non-polar part of the DBSAs, while the DBSA and VO-79 molecules exhibit a preference for interacting with each other via their polar regions. These observations were demonstrated using heptane systems. Figure A 7.2.5 further validates these findings by confirming that the same principles hold true for systems in which pentane serves as the solvent.

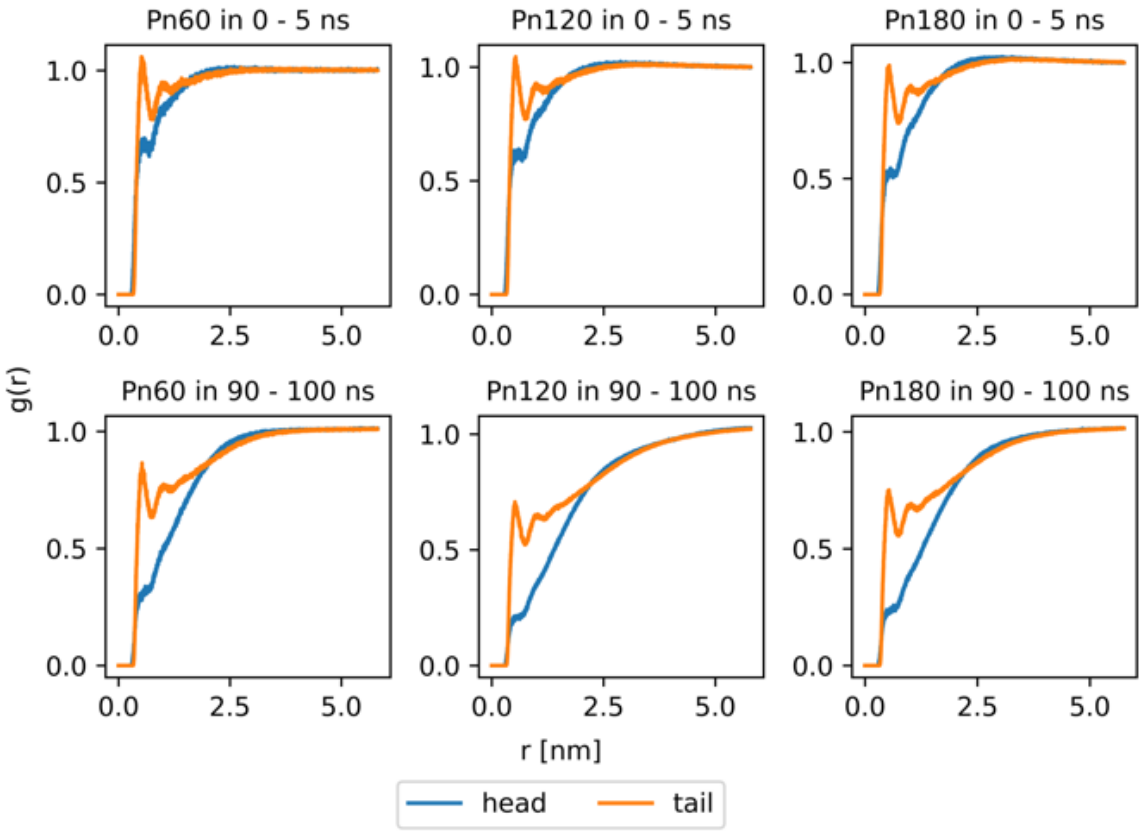

Figure A7.2-5 The RDF of solvent molecules with reference to different parts of the DBSA molecules drawn for Pn60, Pn120 and Pn180 systems and in 2 different time intervals.

Furthermore, Figure A7.2.6 illustrates the inclination of the polar components of VO-79s and DBSAs to mutually influence each other. As mentioned in the main text, the RDFs portraying the interactions of the head and tail segments of DBSA molecules are graphed relative to the core and aliphatic side chains of VO-79 molecules. These RDFs are showcased for both the initial and final phases of the simulation. As shown in Figure A7.2.6, across all systems in the selected temporal intervals, the RDFs linked with the polar constituents of the precipitants manifest the most significant peaks.

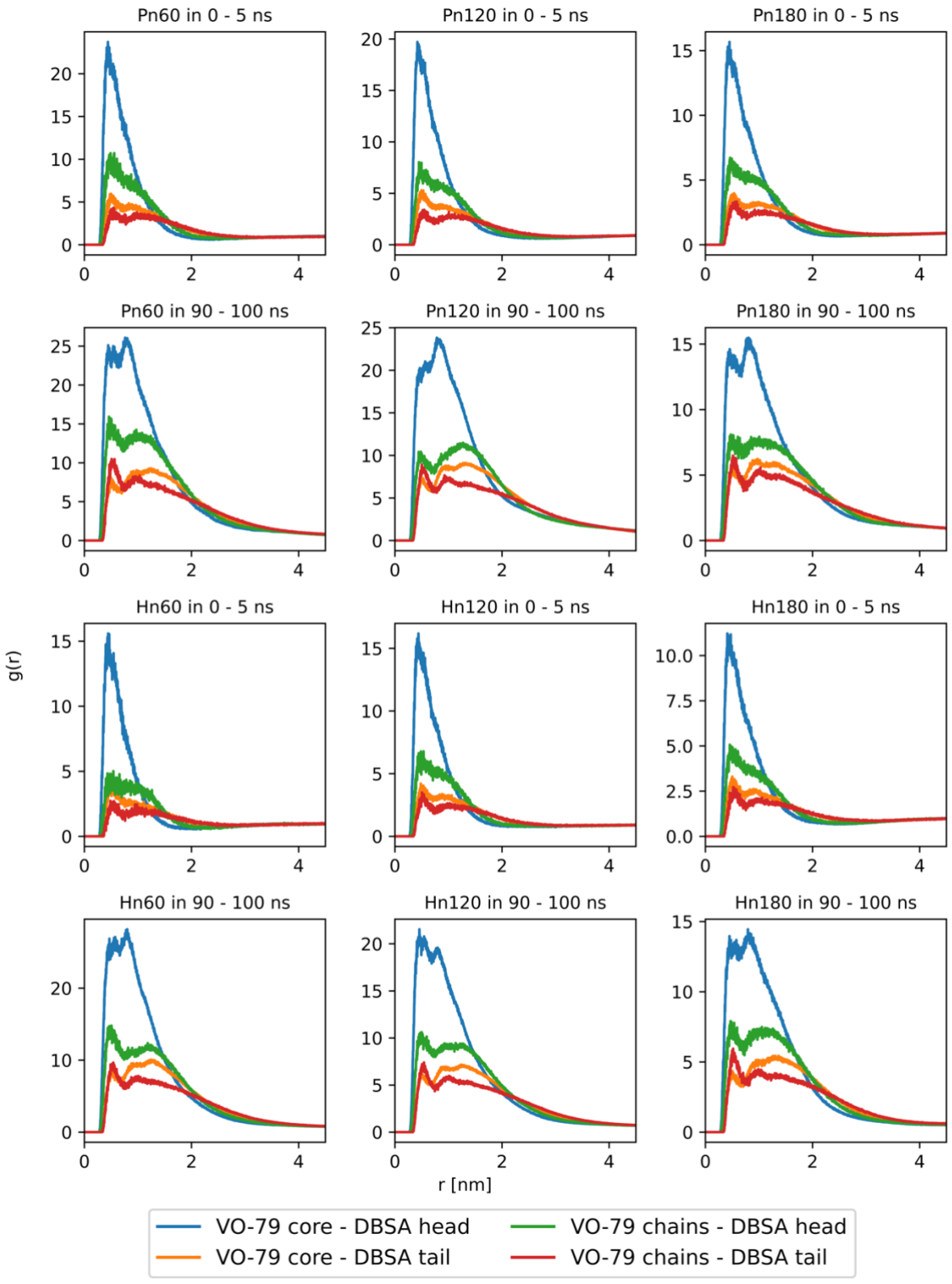

Figure A7.2-6 The RDF of different parts of the DBSA molecules with reference to different parts of the VO-79 molecules.

7.3. Appendix for chapter 4

7.3.1. Equivariant GNNs

Equivariant Graph Neural Networks (GNNs) are designed to predict adsorption energies by capturing the intricate interactions between adsorbates and surfaces while adhering to the physical symmetries of these systems. In adsorption studies, the adsorbate-surface system can be represented as a graph where nodes correspond to atoms, and edges represent bonds or interactions between them. Equivariant GNNs enhance traditional GNNs by incorporating rotational, translational, and reflection symmetries directly into their architectures, ensuring that predictions remain invariant under physical transformations.

The way GNNs work is to convert atomic structure of the adsorbate-surface system is into a graph. In the graph, atoms are treated as nodes, and the bonds or interactions between them are edges, effectively capturing the molecular or crystal structure as a graph. To accurately represent each atom, node embeddings can be enriched with various chemical descriptors. While basic options include atomic numbers or one-hot encodings to differentiate chemical elements, embeddings can also incorporate additional atomic properties like electronegativity, atomic radius, and ionization energy. Structural characteristics, such as hybridization state, formal charge, and valence electrons, add further detail to node representations, and environmental factors, e.g., the number of bonded neighbors or distances to adjacent atoms, provide additional refinement.

Representing atoms as nodes in a graph has become a common practice in the literature. However, determining the edges within this framework involves various approaches. The most straight forward approach is to derive edges from the molecular graph, specifically the chemical bonds. Despite its prevalence, this approach is not MD compatible as it introduces discontinuity and disrupts the underlying physics [A1] MD compatibility requires the predicted force field generated by the model, when applicable, t constitute a conservative vector field. A straightforward approach to ensure this is to have the model predict the potential energy function, from which the force can be derived through gradient computation (utilizing back-propagation with respect to atom coordinates).

A significant advancement in GNNs is the incorporation of physical symmetries through equivariant neural networks, where equivariance refers to the property of a function that transforms

predictably under certain input transformations. For physical systems, essential symmetries include translational, rotational, and reflection invariance, which ensure that properties like energy remain unchanged irrespective of a system's orientation or position. Equivariant networks integrate these symmetries into the model's architecture, ensuring that predictions remain physically consistent and adhere to fundamental laws. For a detailed explanation of equivariance, symmetries in 3D space (SE(3)/E(3)), and the requirements for models to adhere to these symmetries, see Appendix A of the original Equiformer paper [A2].

Equivariant neural networks for 3D systems, such as SE(3)/E(3)-equivariant networks, leverage transformations represented by tensor fields and spherical harmonics, which project spatial information in a way that respects these symmetries. This approach offers several advantages:

Enhanced Modeling of Physical Laws: Equivariant networks respect conservation laws and symmetries, ensuring their predictions align with physical reality. These networks incorporate inductive biases that make their internal representations and predictions equivariant to 3D translations, rotations, and optionally inversions. They build equivariant features for each node using vector spaces of irreducible representations (irreps) and enable interactions through equivariant operations like tensor products. This is crucial for accurately predicting properties such as forces and energies, as improper symmetry handling can lead to unrealistic results. For example, models like Tensor Field Networks [A3] achieve 3D rotational and translational equivariance, using spherical harmonics and irreps to ensure predictions adhere to these symmetries.

Reduction of Model Complexity: By embedding physical constraints directly into the model's architecture, equivariant networks achieve high accuracy with fewer parameters, simplifying the architecture. This constraint-based approach reduces overfitting, improves computational efficiency, and often avoids the need for manual feature engineering. Models like DimeNet [A1] use a basis representation that combines directional and distance information, enabling compact and efficient representations of atomic interactions.

Improved Data Efficiency: Encoding symmetry information into the model structure itself enhances data efficiency, reducing the amount of data required to capture underlying physics. This efficiency is valuable in domains like adsorption modeling, where datasets may be limited in scope

or costly to generate. Since these models understand rotational and translational invariance as part of their structure, they require fewer training examples to learn accurate predictions for complex molecules [A2].

Incorporating these symmetries requires constructing node features through invariant or equivariant representations. Applying directional embeddings and tensor products to ensure that message-passing steps preserve symmetry across layers. For instance, in DimeNet [A1], directional message passing is refined by joint 2D representations of interatomic distances and angles using spherical harmonics and Bessel functions, creating rotationally invariant features. Furthermore, to maintain differentiability—a requirement in physical simulations—equivariant models avoid non-continuous activations (like ReLU) and opt for basis representations that stabilize predictions even under small deformations [A1].

In ML-driven energy and force predictions, different tasks are designed to approximate computational chemistry simulations efficiently. Among them, Structure to Energy and Forces (S2EF), Initial Structure to Relaxed Structure (IS2RS), and Initial Structure to Relaxed Energy (IS2RE) serve distinct yet interconnected purposes across various applications beyond catalysis [A4].

S2EF focuses on predicting the total energy of an atomic configuration along with the per-atom forces acting on each atom. This task is fundamental in approximating computationally expensive quantum chemistry calculations and is widely used to accelerate molecular dynamics simulations and structure relaxations. IS2RS, on the other hand, takes an initial molecular or atomic configuration and predicts the final relaxed atomic positions. This is particularly relevant for modeling stable molecular structures, understanding reaction mechanisms, and simulating physical processes that involve structural rearrangements. IS2RE bypasses the explicit relaxation process by directly predicting the energy of the final relaxed configuration from an initial structure, providing a rapid alternative for estimating system stability and adsorption properties.

Chapter 4 specifically focuses on S2EF, given its foundational role in enabling faster and more accurate force evaluations, which are critical for large-scale molecular simulations.

7.3.2. Details on the Architecture of GemNet-OC

The GemNet-OC model is engineered to predict the energy and forces within atomic systems through a sophisticated graph-based framework. It begins by receiving the atomic numbers and spatial positions of all atoms in the system as input, eliminating the need for additional information such as bond types. In this representation, atoms are modeled as nodes, and edges are established between atoms based on their proximity, forming a graph that encapsulates the system's structural information.

As explained in section 4.3.2, GemNet-OC begins by embedding atoms and the edges between them into high-dimensional vectors that encode geometric relationships using three distinct Bessel functions—radial, circular, and spherical—to capture pairwise distances, three-body angles, and four-body dihedral-like configurations, with polynomial envelopes ensuring smooth differentiability. These initial embeddings are then processed through an embedding layer and further refined by multiple interaction layers, each producing dual representations that are leveraged to compute energy and force information, thereby progressively building richer chemical and geometric features. Finally, each Interaction stage feeds into output blocks that predict individual contributions to energy and forces, with the total energy and forces on each atom being obtained by aggregating these contributions across all atoms.

It should be noted that ML models usually output energy and then predict forces by calculating $F_a = -\partial E/\partial x_a$ via backpropagation. This approach ensures a conservative force field, which is crucial for the stability of molecular dynamics simulations. However, equivariant neural networks, such as GemNet-OC, can also directly predict forces and other vector quantities, leveraging their symmetry-preserving architectures. For further details on this direct force prediction approach, refer to Section 3 of the original GemNet paper.

The interaction blocks of the model employ a multi-level message-passing mechanism that enhances traditional two-level schemes [A1, A5, A6]. This hierarchy includes atom-to-atom interactions, which efficiently exchange information using only the distances between atoms without imposing neighbor restrictions. Additionally, atom-to-edge and edge-to-atom interactions are added to GemNet-Oc and did not exist in the original implementation of GemNet. These interactions facilitate detailed information flow by incorporating directional embeddings,

geometric data, and learnable filters. During this process, the model calculates interatomic distances, angles between neighboring edges, and dihedral angles formed by triplets of edges, enabling it to understand complex geometric relationships within the atomic system.

Unlike many models that use a fixed distance cutoff to define interatomic connections, GemNet-OC selects a predetermined number of nearest neighbors for each atom. While this might initially raise concerns about differentiability if atoms switch neighbor order, the authors report no practical problems. This approach ensures a consistent neighborhood size, enhancing the model's scalability and performance across diverse systems. To further optimize computational efficiency, the model utilizes simplified basis functions by decoupling the radial and angular components. Gaussian or 0-order Bessel functions represent radial dependencies, while the angular dependencies are captured using the outer product of order 0 spherical harmonics, simplifying to Legendre polynomials.

GemNet-OC's architecture includes four Interaction Blocks, each comprising 52 layers (about 22.7 million parameters in total), which facilitate extensive multi-level message passing. These blocks integrate geometric information such as distances, angles, and dihedral angles to progressively refine the embeddings. Following the Interaction Blocks are five Output Blocks, each containing 22 layers (about 15.1 million parameters in total). Instead of making direct predictions, these blocks generate intermediate embeddings that are concatenated and processed through multi-layer perceptrons (MLPs) to produce the final predictions for energy and forces. Separate MLP pathways convert the concatenated embeddings into precise output values, with a learnable MLP enhancing the atom embeddings and ensuring an efficient information flow from embeddings to energy predictions.

In addition to the Interaction and Output Blocks, GemNet-OC incorporates 28 additional layers that would add up to approximately 3.3 million parameters are distributed across various modules. These include initial embedding layers that initialize the atomic and edge embeddings, specialized modules that handle different basis functions for radial and angular components, and components that finalize the prediction outputs. This comprehensive layering ensures a seamless transition from input embeddings through complex interaction layers to the final predictive outputs.

Tables A 7.3-1 to 4 show the details of each layer of the network including the modules, layer types and number of neurons (input and output dimensions) for each layer in different blocks.

Table A 7.3-1 Embedding block

Module	Layer Name	Type	Neurons (in→out)
Embedding Init	`atom_emb.embeddings`	Embedding	(83 elements → 256-dim vector)
	`edge_emb.dense.linear`	Linear + SiLU	640 → 512
Basis Embedding (Quadruplet)	`mlp_rbf_qint.linear`	Linear	128 → 16
	`mlp_cbf_qint`	Basis Embedding	(implicit)
	`mlp_sbf_qint`	Basis Embedding	(implicit)
Basis Embedding (Atom-Edge Interaction)	`mlp_rbf_aeint.linear`	Linear	128 → 16
	`mlp_cbf_aeint`	Basis Embedding	(implicit)
	`mlp_rbf_eaint.linear`	Linear	128 → 16
	`mlp_cbf_eaint`	Basis Embedding	(implicit)
	`mlp_rbf_aint`	Basis Embedding	(implicit)
Basis Embedding (Triplet)	`mlp_rbf_tint.linear`	Linear	128 → 16
	`mlp_cbf_tint`	Basis Embedding	(implicit)
Misc / Basis Embedding	`mlp_rbf_h.linear`	Linear	128 → 16
	`mlp_rbf_out.linear`	Linear	128 → 16

Table A 7.3-2 Interaction block

Module	Layer Name	Type	Neurons (in→out)
Core Update (edge)	dense_ca.linear	Linear + SiLU	512 → 512
Triplet Interaction (trip_interaction)	dense_ba.linear	Linear + SiLU	512 → 512
	mlp_rbf.linear	Linear	16 → 512
	mlp_cbf.bilinear.linear	Efficient Bilinear	1024 → 64
	down_projection.linear	Linear + SiLU	512 → 64
	up_projection_ca.linear	Linear + SiLU	64 → 512
	up_projection_ac.linear	Linear + SiLU	64 → 512
Quadruplet Interaction (quad_interaction)	dense_db.linear	Linear + SiLU	512 → 512
	mlp_rbf.linear	Linear	16 → 512
	mlp_cbf.linear	Linear	16 → 32
	mlp_sbf.bilinear.linear	Efficient Bilinear	1024 → 32
	down_projection.linear	Linear + SiLU	512 → 32
	up_projection_ca.linear	Linear + SiLU	32 → 512
	up_projection_ac.linear	Linear + SiLU	32 → 512
Atom-Edge Interaction (atom_edge_interaction)	dense_ba.linear	Linear + SiLU	256 → 256
	mlp_rbf.linear	Linear	16 → 256
	mlp_cbf.bilinear.linear	Efficient Bilinear	1024 → 64
	down_projection.linear	Linear + SiLU	256 → 64
	up_projection_ca.linear	Linear + SiLU	64 → 512
	up_projection_ac.linear	Linear + SiLU	64 → 512
Edge-Atom Interaction (edge_atom_interaction)	dense_ba.linear	Linear + SiLU	512 → 512
	mlp_rbf.linear	Linear	16 → 512
	mlp_cbf.bilinear.linear	Efficient Bilinear	1024 → 64
	down_projection.linear	Linear + SiLU	512 → 64
	up_projection_ca.linear	Linear + SiLU	64 → 256
Atom-Atom Interaction (atom_interaction)	bilinear.linear	Bilinear	1024 → 64
	down_projection.linear	Linear + SiLU	256 → 64
	up_projection.linear	Linear + SiLU	64 → 256
Edge Residual Stack (before skip ×2)	Residual Layers (Edge)		512 → 512 ×2 per Residual
Atom Residual Stack (atom_emb_layers ×2)	Residual Layers (Atom)		256 → 256 ×2 per Residual
Atom Update (atom_update)	dense_rbf.linear	Linear	16 → 512
	layers[0].linear	Linear + SiLU	512 → 256
	layers[1-3]	Residual (2× Linear each)	256 → 256 ×2 per Residual
Post-Concat. Layer	concat_layer.dense.linear	Linear + SiLU	1024 → 512
Final Residual Layer	residual_m[0]	Residual (2× Linear)	512 → 512 ×2

Table A 7.3-3 Output block

Module	Layer Name	Type	Neurons (in→out)
Output Block	`dense_rbf.linear`	Linear	16 → 512
	`layers[0].linear`	Linear + SiLU	512 → 256
	`layers[1-3]`	Residual (2× Linear each)	256 → 256 ×2 per Residual
Energy Head (per block)	`seq_energy_pre[0].linear`	Linear + SiLU	512 → 256
	`seq_energy_pre[1-3]`	Residual (2× Linear each)	256 → 256 ×2 per Residual
	`seq_energy2[0-2]`	Residual (2× Linear each)	256 → 256 ×2 per Residual
Force Prediction Head	`dense_rbf_F.linear`	Linear	16 → 512
	`seq_forces[0-2]`	Residual (2× Linear each)	512 → 512 ×2 per Residual

Table A 7.3-4 Global output MLP

Module	Layer Name	Type	Neurons (in→out)
Global Output MLP	`out_mlp_E[0].linear`	Linear + SiLU	1280 → 256
	`out_mlp_E[1-2]`	Residual (2× Linear each)	256 → 256 ×2 per Residual
Final Energy Output	`out_energy.linear`	Linear	256 → 1
Global Force MLP	`out_mlp_F[0].linear`	Linear + SiLU	2560 → 512
	`out_mlp_F[1-2]`	Residual (2× Linear each)	512 → 512 ×2 per Residual
Final Force Output	`out_forces.linear`	Linear	512 → 1

7.3.3. Training Curves and Details

All fine-tuning experiments were conducted on the Narval supercomputing cluster, managed by the Digital Research Alliance of Canada. Each training job was allocated significant computational resources, utilizing a single node with 16 CPU cores, 256 GB of RAM, and one NVIDIA A100 GPU. Full fine-tuning of the model on the larger Extracted FG dataset for 16

epochs with a batch size of 64 required approximately 28 minutes of computation time. For the smaller, more specialized Segregated Aromatics dataset, a full 16-epoch training run with a batch size of 16 was completed in approximately 18 minutes.

To illustrate the consistency of the results, Table A 7.3-5 shows the standard deviation (STD) of the errors for the pre-trained model and each of the fine-tuning strategies. A lower STD indicates that the prediction errors are tightly clustered around the average error (MAE), signifying a more reliable and consistent model. The extremely high STDs for the pre-trained model on both datasets highlight its unreliability, with prediction errors that are large and highly erratic. In contrast, all fine-tuning strategies dramatically reduce the STD, demonstrating that the adapted models provide not only more accurate but also significantly more dependable predictions.

When comparing the fine-tuning methods, on the diverse Extracted FG dataset, full fine-tuning achieves the lowest STD (0.0766), making it the most consistent and trustworthy model. For the specialized Segregated Aromatics dataset, an interesting trend emerges: all three fine-tuning strategies yield similar STDs (around 0.18). This suggests that while the average accuracy (MAE) can be optimized by freezing the output blocks, the inherent difficulty in predicting complex aromatic systems leads to a comparable level of prediction variability across all adapted models.

Table A 7.3-5 The standard deviation of the errors for the pretrained baseline and fine-tuning strategy

Fine-Tuning and Test Dataset	Fine-Tuning Strategy	STD [eV]
Extracted FG	pretrained	7.686
	FFT	0.0766
	FOB	0.0956
	FIB	0.2930
Segregated Aromatics	pretrained	12.052
	FFT	0.189
	FOB	0.177
	FIB	0.176

Figures A7.3-1 and A7.3-2 illustrate the training curves corresponding to each fine-tuning strategy applied to the Segregated aromatics and Extracted FG datasets, respectively. In both cases, the training set constitutes 80% of the dataset, while the validation and test sets each account for 10%.

For the Segregated aromatics dataset, which comprises 1,140 entries, the training set includes 912 entries (1,140 × 0.8). Using a batch size of 8, this yields 114 steps per epoch (912/8) and a total of 1,824 steps over the fine-tuning process. Similarly, for the Extracted FG dataset, with 6,866 entries, the training set consists of 5,536 entries. With a batch size of 32, this results in 173 training steps per epoch (5,536/32) and a total of 2,768 steps over 16 epochs (173 × 16).

The metrics displayed in Figures A7.3-1 and A7.3-2 include the training loss, energy MAE (Mean Absolute Error), and the fraction of systems where both force and energy predictions meet a specified accuracy threshold, denoted as Energy, Forces Within Threshold (EFWT). The training loss is defined as the MAE of forces across all atoms in every entry of the training set. Although the models are fine-tuned to predict energies, using force MAE as the primary loss function ensures optimization at the atomic level, which inherently leads to more accurate and physically consistent energy predictions.

The energy MAE and EFWT metrics are evaluated on the validation set after every 50 training steps. EFWT is calculated based on the absolute error of per-atom forces and energy per system. Specifically, it represents the fraction of systems where the maximum force error is below 0.03 eV/Angstrom and the maximum energy error is below 0.02 eV.

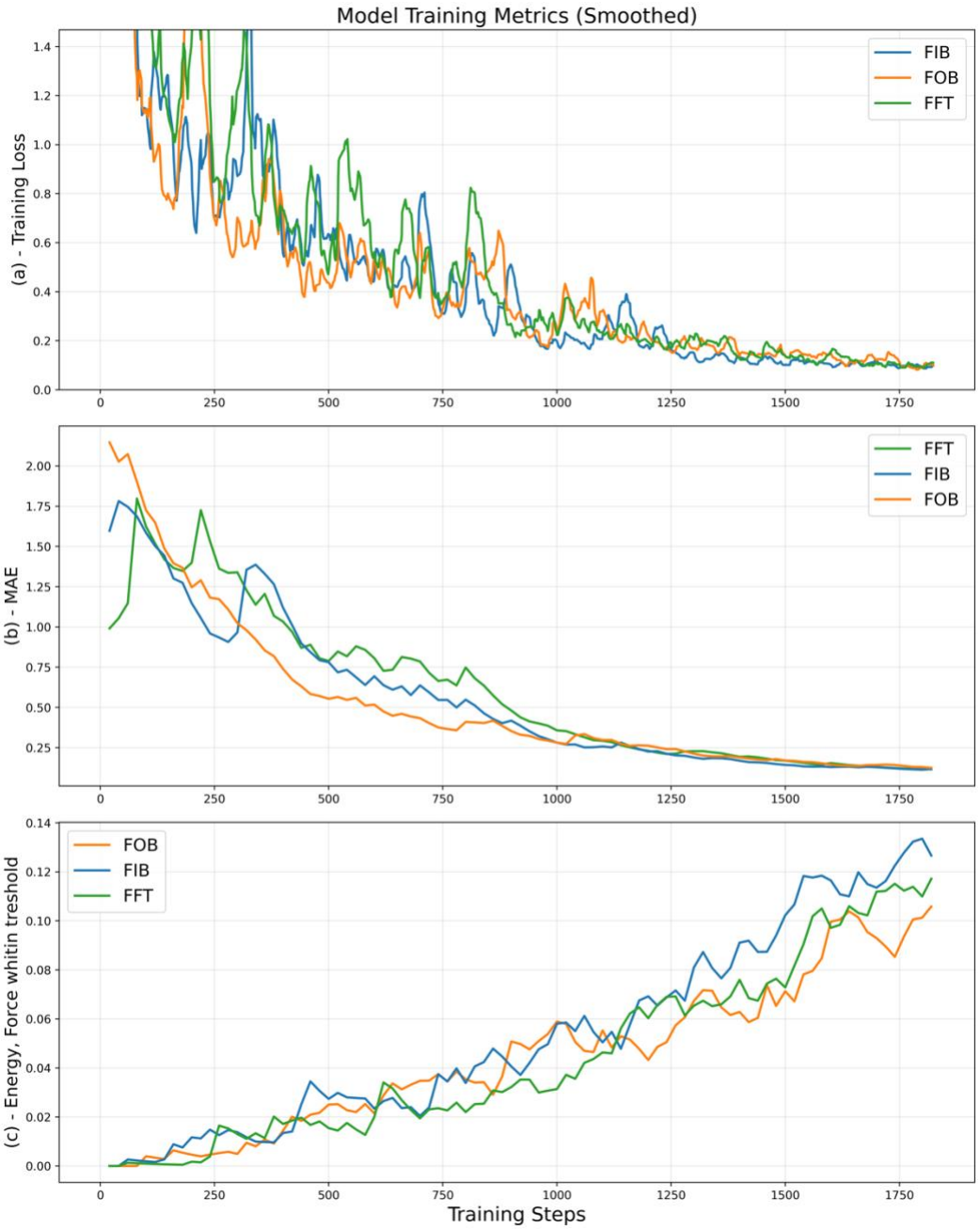

Figure A7.3-1 The training curves related to fine-tuning the model on segregated aromatics dataset. Subfigure (a) shows the primary loss (force MAE) on the training set during training, (b) shows the energy MAE calculated on the validation set every 50 steps, and (c) shows the fraction of validation set entries where the maximum energy and force errors are below 0.02 eV and 0.03 eV/Å, respectively

Figure A7.3-2 The training curves related to fine-tuning the model on extracted FG dataset. (a) shows the primary loss (force MAE) on the training set during training. (b) shows the energy MAE calculated on the validation set. (c) shows the fraction of the validation set entries where energy and force errors are below certain thresholds

References for Appendix

[A1] J. Gasteiger, J. Groß, S. Günnemann, Directional Message Passing for Molecular Graphs, (2020). http://arxiv.org/abs/2003.03123.

[A2] Y.L. Liao, T. Smidt, Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs, (2022). http://arxiv.org/abs/2206.11990.

[A3] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, P. Riley, Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, (2018). http://arxiv.org/abs/1802.08219.

[A4] Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C.L. Zitnick, Z. Ulissi, Open Catalyst 2020 (OC20) Dataset and Community Challenges, ACS Catal 11 (2021) 6059–6072. https://doi.org/10.1021/acscatal.0c04525.

[A5] K.T. Schütt, O.T. Unke, M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, (2021). http://arxiv.org/abs/2102.03150.

[A6] K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – A deep learning architecture for molecules and materials, J Chem Phys 148 (2018). https://doi.org/10.1063/1.5019779.