Molecular dynamics simulations of asphaltene aggregation: machine learning identification of representative molecules, polydispersity and inhibitor performance

Rémi Pétuya,[*,†] Abhishek Punase,[‡] Emanuele Bosoni,[†] Antonio Pedro de Oliveira Filho,[‡] Juan Sarria,[¶] Nirupam Purkayastha,[¶] Jonathan Wylde,[‡] and Stephan Mohr[†]

†*Nextmol (Bytelab Solutions SL), Barcelona, Spain*

‡*Clariant Oil Services, Clariant Corporation, Houston, Texas, USA*

¶*Clariant Produkte (Deutschland) GmbH, Frankfurt, Germany*

§*Heriot-Watt University, Edinburgh, Scotland, UK*

E-mail: remi.petuya@nextmol.com

Abstract

Molecular Dynamics simulations have been employed to investigate the effect of polydispersity on the aggregation of asphaltene. To make the large combinatorial space of possible asphaltene blends accessible to a systematic study via simulation, an upfront unsupervised machine learning approach (clustering) was employed to identify a reduced set of model molecules representative of the diversity of asphaltene. For these molecules, monodisperse asphaltene simulations have shown a broad range of aggregation behavior, driven by their structural features: size of the aromatic core,

length of the aliphatic chains and presence of heteroatoms. Then, the combination of these model molecules in a series of polydisperse mixtures have highlighted the complex and diverse effects of polydispersity on the aggregation process of asphaltene, which yielded both antagonistic, synergistic and seed effects. These findings illustrate the necessity of accounting for polydispersity when studying the asphaltene aggregation process and have permitted to establish a robust protocol for the *in-silico* evaluation of the performance of asphaltene inhibitors, as illustrated for the case of a nonylphenol resin.

1 Introduction

As a consequence of global governmental policies, the increasing popularity of electric vehicles and the momentum for hydrogen as a clean source of energy, the consumption of gasoline and other fuels is set to steadily decline. On the other hand, the proportion of the average oil barrel dedicated to petrochemicals will grow up to an estimated 20% by 2040.[1] Therefore, ensuring a sustainable production of fossil resources will continue to be an objective of paramount importance. A predominant challenge for the Oil & Gas industry is the deposition of asphaltene,[2] a class of compounds defined as the fraction of crude oil that is soluble in toluene but not in *n*-heptane.[3] Asphaltenes, considered to be among the heaviest and more polar components of crude oils, are generally described as a very polydisperse class of organic solids made of a variety of polyaromatic structures with aliphatic chains or heteroatoms, either organic or metallic.[2] Their interaction with water, clay and between themselves can result in critical issues in oil fields. Overall, they can precipitate in the reservoir and plug production and transportation flowlines, risking economic loss due to flow interruption and environmental damage.[4] An efficient and economical mitigation strategy consist of injecting chemical additives, referred to as asphaltene inhibitors, to stabilize asphaltene in crude oil. Hence, the development of these additives is of high industrial relevance. They are generally surfactants or polymers,[5–8] but the potential of more exotic

chemistries, such as amphiphilic macromolecules[9] or deep eutectic solvents[10] (a class of products formed by the hydrogen bonding between cheap and safe components, e.g. an amine and a carboxylic acid, which represents an alternative to the expensive traditional ionic liquids) has also been evaluated. The development of efficient asphaltene inhibitors is often hindered because of the oilfield dependence of asphaltene stability and aggregation behavior. Therefore, it seems critical to understand and rationalize the underlying mechanisms of the asphaltene aggregation process to better address its inhibition.

After decades of research, two main interpretations of the asphaltene aggregation process are usually presented. On the one hand, the "Yen-Mullins" model argues a hierarchical description in which asphaltenes are predicted to form dense nanoaggregates of less than 10 molecules, driven by interactions between aromatic centers, which then aggregate less strongly into larger clusters.[11,12] Such macroaggregates end up being too heavy and produce solid deposits on the pipeline wall. On the other hand, as the "Yen-Mullins" model fails to predict or reconcile a series of experimental observations, such as the complexity of asphaltene molecular structure or the heterogeneous distribution of nanoaggregates sizes, Gray *et al.*[13] proposed an alternate supramolecular model to better capture the high complexity of the aggregation process. In this paradigm, aromatic $\pi - \pi$ stacking is not considered to be the dominant aggregation driving force but a contributing factor alongside other interactions relevant to petroleum such as acid-base interactions, hydrogen bonding, metal coordination complexes and interactions between cycloalkyl and alkyl groups.[14] In this model, strongly bound nanoaggregates can continue to grow beyond 10 asphaltenes. As encouraged in the research article from Gray *et al.*,[13] this paradigm has been put to the test both experimentally and computationally over the last decade.

Historically, asphaltene research has been mostly experimentally led,[15] but, in the context of the global digital transition, there is undoubtedly a momentum for *in-silico* approaches, thanks to their contributions to the understanding and the rationalization of complex chemical and physical processes, as well as their relative affordability in comparison to laboratory

experiments. Seminal works such as the ones by Headen et al.[16,17] and Seghdi et al.[18] have established robust molecular dynamics (MD) simulations protocols capable of providing valuable insights into the first stage of the asphaltene aggregation process. More recently, via a series of MD studies[19–24] Santos Silva et al. have worked on decoding the complex relationships between asphaltene molecular structures and their aggregation, studying the role of heteroatoms positioned either on the core or on the lateral chains,[19,20] the role of metalloporphyrins and demulsifier molecules,[21,24] and the effect of variations in the size of the aromatic core and lateral chains length.[22] Overall, they have shown that the formation of nanoaggregates depends on the size of the conjugated core and on the possible presence of an H-bonds forming polar group, whereas macroaggregation is determined by the length of the lateral chains and their possible terminal polar group.[23] Given that their observations lay outside the domain of the "Yen-Mullins" model, they consequently argued that this colloidal model, even though it is capable of describing the aggregation process of standard asphaltenes, might be a particular case of the more general supramolecular model, as proposed by Gray *et al.*,[13] which is better suited to address the complete chemical diversity of asphaltenes. Furthermore, in a few studies, MD simulations have also been deployed to investigate the inhibitor action of chemical additives such as dodecylbenzenesulfonic acid,[25] limonene,[17] *n*-octylphenol,[26–28] and a series of polymers (two succinimide-based structures and one maleic anhydride),[29] on asphaltene aggregation. Finally, Headen et al. have demonstrated that MD aggregation simulations for monodisperse asphaltene systems qualitatively reproduce neutron total scattering data.[30]

However, two types of limitations have been identified for MD simulations of asphaltene:[31,32] i) the size and time scales limitations of MD simulations restrain this approach to the first stage of aggregation, and coarse-grained simulations would be necessary to simulate the second and later stages of assembly (e.g., flocculation), and ii) polydispersity in the asphaltene molecular structures is of great importance for the prediction of aggregation structures and should be included in any simulation. Indeed, in order to thoroughly investigate

the effect of functional groups on the aggregation process, many previous works—with the exception of Javanbakht et al.[32]—had to be limited to series of asphaltene model molecules with similar aromatic cores[18–20,22,33] to prevent the complexity of the systems from hindering the decoding of interaction mechanisms that underlie such process. To provide a different perspective on the asphaltene aggregation process, one objective of this work is to account for the global asphaltene polydispersity.

The molecular structure of asphaltene has been a long standing debate and important focus during decades of investigation, thus over time their molecular weight has been estimated at values spanning six orders of magnitude.[34] The development of asphaltene model molecules, which is not the focus of this work, designed for performing MD simulations, has been addressed and reviewed in a series of studies.[31,35–38] In recent years, two experimental techniques have provided insights of unprecedented quality on the molecular structures of asphaltenes. On the one hand, atomic force microscopy studies have permitted to visualize more than 100 asphaltene motifs,[39] confirming the presence of structures made of polynuclear aromatic hydrocarbons with alkyl side-chains, usually referred to as the "island" or "continental" model. On the other hand, extrographic fractionation and ultrahigh-resolution mass spectrometry studies advocate that in petroleum island-type asphaltenes coexist with less generally accepted "archipelago" motifs, in which multiple aromatic cores are bridged together and include multiple functionalities.[40,41] Such works argue that while the island motifs are readily accessible, asphaltene purification is required to detect and characterize archipelago asphaltenes.

In this work, we have performed a series of MD simulations to investigate the aggregation of different systems of asphaltene. After identifying, via unsupervised machine learning (ML) approaches, a set of asphaltenes that are representative of the diversity of the catalogue of Law et al.,[38] we have have focused first on the aggregation of monodisperse asphaltene systems. Then, the selected model molecules have been combined to study the aggregation of polydisperse asphaltene mixtures. Finally, the most aggregating polydisperse mixture

of asphaltene has been selected as test case to study the action of a nonylphenol resin asphaltene inhibitor at a reasonably low concentration of 1 wt%. In what follows, the results are reported and discussed while the details of the computational methods employed are presented in the final section.

2 Results and discussions

2.1 Identification of representative asphaltene models

The asphaltene molecular models used in this work have been selected from a catalogue of 100 plausible molecular models designed for MD simulations of asphaltenes (89 models) and resins (11 models) that comprises both island and archipelago motifs.[38] To generate these models, the quantitative molecular representation approach implemented by Boek et al.[42] had been applied to elemental analysis and ^{1}H–^{13}C nuclear magnetic resonance spectroscopy experimental data.[38] Furthermore, as shown by Law et al.,[38] their catalogue of molecules reasonably covers the same chemical space as previous asphaltene models collected from an in-depth literature review. Simulating all 89 asphaltene molecules and multiple combinations of these is currently beyond the reach of affordable MD simulations. Therefore, identifying representative asphaltene models out of this catalogue is a way of accounting for polydispersity at a reasonable cost. After digitalization of the catalogue of molecules,[38] an unsupervised machine learning (ML) strategy relying on unsupervised clustering, detailed in section 4.1, has been implemented to identify the 6 model molecules displayed in Figure 1. These molecules are representative of the large diversity of asphaltene structures, with 4 island and 2 archipelago asphaltenes. In particular the island models have very different structural features: A3 has long side-chains, A54 has many heterocycles and A29 is very bulky. The detailed composition of each cluster is given in the Supporting Information.

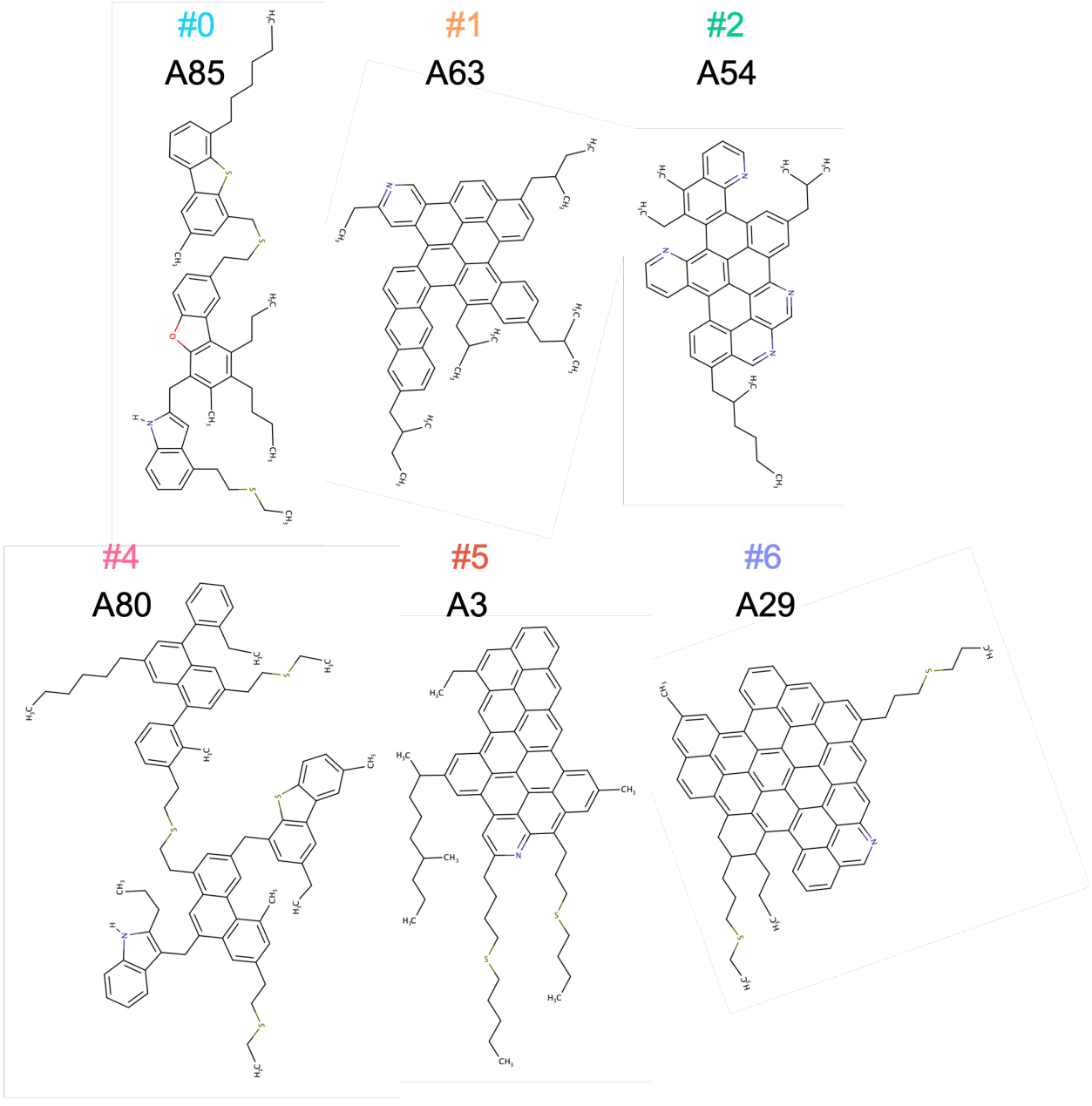


Figure 1: 2D representations of the representative asphaltene model molecules selected after the cluster analysis. The molecule name is consistent with the original work from Law et al.,[38] and cluster labels and colors are consistent with Figure 8. No molecule from cluster #3 has been selected as they are resins.

2.2 Monodisperse asphaltene systems

The aggregation of representative asphaltene model molecules from Figure 1 has been studied by a series of MD simulations in toluene and heptane. The detail of the set-up is given in section 4.2. To investigate the trade-off between convergence and simulation cost, monodisperse asphaltene simulations with 40, 100, and 200 asphaltene molecules at a concentration of 7 wt % were performed. The aggregation state of asphaltenic systems is monitored consistently with previous works: the series of observables defined by Headen et al.[31] have been implemented in a homemade python script using the package MDAnalysis version 2.0[43] and are defined in section 4.3. The aggregation number, g_n, which corresponds to the number of molecules per aggregate, also referred to as clusters, permits to monitor the equilibrium state of the system while estimating the size of a cluster of asphaltenes. In Figure 2, the evolution of g_n during the 240ns simulation of 100 molecules of model A29 in heptane is represented along with 5ns and 20ns moving averages to guide the eye.

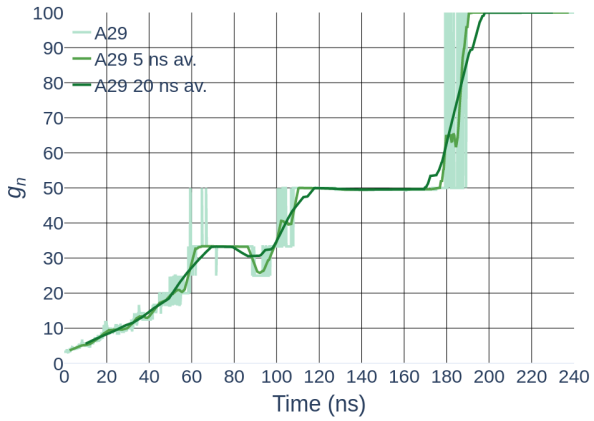


Figure 2: Aggregation number, i.e., average asphaltene cluster size from the simulation of 100 molecules of asphaltene A29 in heptane. 5ns and 20ns moving averages are used to filter short term fluctuations of the aggregation dynamics.

For all other monodisperse simulations performed in this study, the evolution of g_n is displayed in the Supporting Information Figure S5 to Figure S9, and Table 1 reports the final values of the 20ns average window of the aggregation number, thus filtered from short

time oscillations. Additionally, intermediate values for the same observable after 120ns are provided to evidence the necessity of extending these simulations up to 240ns. As always in molecular modelling, identifying the reasonable system size and time scale to simulate is crucial to balance the trade-off between simulation cost and accuracy. In recent studies, i) Headen et al.[31] have performed simulations of 27 asphaltene molecules during 80ns (even tough they occasionally extended up to 160ns and 500ns, ii) with the same model molecules Ghamartale et al.[27] simulated 50 molecules of asphaltene during 120ns, and iii) Villegas et al.,[33] who focused on the aggregation of a subfraction of asphaltene in toluene, showed by comparison with simulations of system sizes up to 160 asphaltene molecules that in their case simulating 20 asphaltene molecules during 120ns already yielded converged results.

For the model molecules studied in this work, it appears that simulating systems of 100 asphaltenes during 240ns, a system size and simulation time that is well in the top tier of the current state-of-the-art, is the best compromise between cost and accuracy. Indeed, as shown in Table 1 and in the Supporting Information (Figure S7 and Figure S8), with 100 model molecules finite size effects are only observed for A29 in heptane, which arrives at full segregation within 240ns (aggregate size of 100 molecules in Figure 2), whereas aggregate size does not reach 200 when 200 asphaltenes are in the system. However, even at the largest system size studied (200 asphaltenes), aggregation is clearly much stronger with A29 than any other model asphaltene, which is already well captured with 100 asphaltene molecules. When using only 40 asphaltene molecules, finite size effects are also observed with molecules A3, A54 and A85 in heptane. Although very insightful, the simulations performed with 200 asphaltene molecules are currently still too expensive (for instance up to 621,082 atoms for A3 in heptane) to be performed on a regular basis and must be reserved to case-by-case studies. Additionally, when the aggregation behavior of a system is uncertain, as for example with 40 molecules of A85 in heptane (see the Supporting Information Figure S6), extending the simulation time further than 240ns can provide more trustworthy insights. In what follows, we focus our analysis and comments on simulations of 240ns performed with

100 asphaltene molecules. While these system sizes and simulation time scales are accessible on High Performance Computer facilities, though costly over an entire study, it is worth mentioning some initiatives that have been pushing the system size and time scale limits. On the side of the simulation time scale, Glova et al.[44] performed simulations of 50 asphaltene molecules during 5μs to identify the best partial charge method, namely AM1-BCC, to use in combination with the general AMBER force field, whereas on the system size aspect, Javanbakht et al.[32] studied the aggregation of polydisperse asphaltene mixtures of size up to 1005 model molecules during 200ns and concluded that, under their simulation conditions, 375 asphaltene molecules were enough to capture all possible nanoaggregate shapes.

Table 1: Values of the 20ns average window of the aggregation number of all monodisperse simulations performed in this study after half of the MD production, g_n^{120} at 120ns, and at the end of the simulations, g_n^{240} at 240ns.

		40 asph.		100 asph.		200 asph.	
asph.	solvent	g_n^{120}	g_n^{240}	g_n^{120}	g_n^{240}	g_n^{120}	g_n^{240}
A3	toluene	10.2 ± 0.9	7.6 ± 0.5	7.6 ± 0.2	8.3 ± 0.2	7.1 ± 0.1	6.7 ± 0.3
A3	heptane	20.6 ± 0.0	40.0 ± 0.0	12.6 ± 0.6	28.1 ± 1.7	12.0 ± 0.1	28.2 ± 1.5
A29	toluene	40.0 ± 0.3	29.2 ± 3.7	16.9 ± 0.7	60.2 ± 2.8	14.3 ± 0.2	40.8 ± 0.7
A29	heptane	22.1 ± 4.6	40.0 ± 0.0	45.6 ± 4.9	100.0 ± 0.0	37.9 ± 4.5	66.9 ± 0.3
A54	toluene	4.6 ± 0.1	4.0 ± 0.1	4.6 ± 0.1	5.3 ± 0.1	/	/
A54	heptane	15.2 ± 0.9	19.8 ± 1.3	16.5 ± 0.6	23.8 ± 1.4	/	/
A63	toluene	3.9 ± 0.1	3.7 ± 0.1	3.8 ± 0.1	3.9 ± 0.1	3.8 ± 0.0	3.5 ± 0.0
A63	heptane	5.4 ± 0.2	4.9 ± 0.2	6.4 ± 0.2	6.8 ± 02	5.4 ± 0.1	6.4 ± 0.1
A80	toluene	3.3 ± 0.2	3.1 ± 0.1	3.1 ± 0.1	3.4 ± 0.1	/	/
A80	heptane	6.1 ± 0.2	8.7 ± 0.8	10.2 ± 0.4	14.0 ± 0.4	/	/
A85	toluene	3.4 ± 0.1	3.7 ± 0.2	3.4 ± 0.2	3.5 ± 0.1	/	/
A85	heptane	7.8 ± 1.0	23.3 ± 4.2	10.0 ± 0.2	13.7 ± 1.0	/	/

Further than the average number of asphaltene molecule per cluster, the size of these clusters can also be characterized via their radius of gyration, R_g, while an estimate of their density and their relative shape anisotropy κ^2, which takes values between 0 for a spherical cluster and 1 for a linear chain, provide information relative to their shape. Details of the implementation of these metrics are presented in section 4.3. Besides, as these observables have been defined to charaterize the equilibrated state of the asphaltenic systems that were simulated, we have focused on the last 40ns of each simulation in order to compare what we

consider equilibrated asphaltenic systems.

Table 2: Summary of the monodispere simulations performed with 100 asphaltene molecules. For the aggregation number, g_n^{240}, we report the 20ns average window at the end of the simulation. For the radius of gyration, R_g (in Å), the estimated density and the relative shape anisotrpy, κ^2, we report the average value over the last 40ns of the simulations.

asph.	solvent	g_n^{240}	av. R_g	av. density	av. κ^2
A3	toluene	8.3 ± 0.2	14.5 ± 1.3	0.4 ± 0.0	0.2 ± 0.0
A3	heptane	28.1 ± 1.7	19.6 ± 2.3	0.4 ± 0.0	0.1 ± 0.0
A29	toluene	60.2 ± 2.8	30.0 ± 4.0	0.3 ± 0.0	0.2 ± 0.1
A29	heptane	100.0 ± 0.0	35.0 ± 0.3	0.4 ± 0.0	0.2 ± 0.0
A54	toluene	5.3 ± 0.1	10.3 ± 0.6	0.5 ± 0.0	0.2 ± 0.0
A54	heptane	23.8 ± 1.4	18.2 ± 1.6	0.4 ± 0.0	0.3 ± 0.0
A63	toluene	3.9 ± 0.1	11.6 ± 0.8	0.3 ± 0.0	0.3 ± 0.0
A63	heptane	6.8 ± 0.2	12.6 ± 0.8	0.4 ± 0.0	0.2 ± 0.0
A80	toluene	3.4 ± 0.1	15.6 ± 1.1	0.2 ± 0.0	0.3 ± 0.0
A80	heptane	14.0 ± 0.4	17.4 ± 1.6	0.3 ± 0.0	0.2 ± 0.0
A85	toluene	3.5 ± 0.1	13.1 ± 0.8	0.3 ± 0.0	0.3 ± 0.0
A85	heptane	13.7 ± 1.0	15.8 ± 1.6	0.3 ± 0.0	0.2 ± 0.0

Along with the aggregation number, the different observables displayed in Figure 3, accumulated during the last 40ns of a simulation with 100 asphaltenes A29, permit to analyze the aggregation of the different monodisperse systems. Figures for all other simulations are available in the Supporting Information Figure S10 to Figure S18 and results (average values of radius of gyration, estimated density and shape anisotropy) with 100 asphaltene model molecules are summarized in Table 2 (see Supporting Information Table S3 and Table S4 for equivalent summaries for simulations with 40 and 200 asphaltene molecules, respectively). As could be expected, in toluene the dispersion of the model molecules studied in this work is generally very stable. For simulations with 100 asphaltenes, aggregation numbers larger than 10 are only obtained for one exception: molecule A29 (see Table 1 and Figure S7). A29 is also very clearly the most strongly aggregating molecule in heptane. In heptane, Figure 2 shows different stages of aggregation of A29, with 3 plateaus around $g_n = 33$, $g_n = 50$ and $g_n = 100$, the later corresponding to a complete segregation of the asphaltenes from the solvent. As observed in other simulations,[18,23] the aggregation process of this system can be described hierarchically. Initially, in the stage that Sedghi et al.[18] named

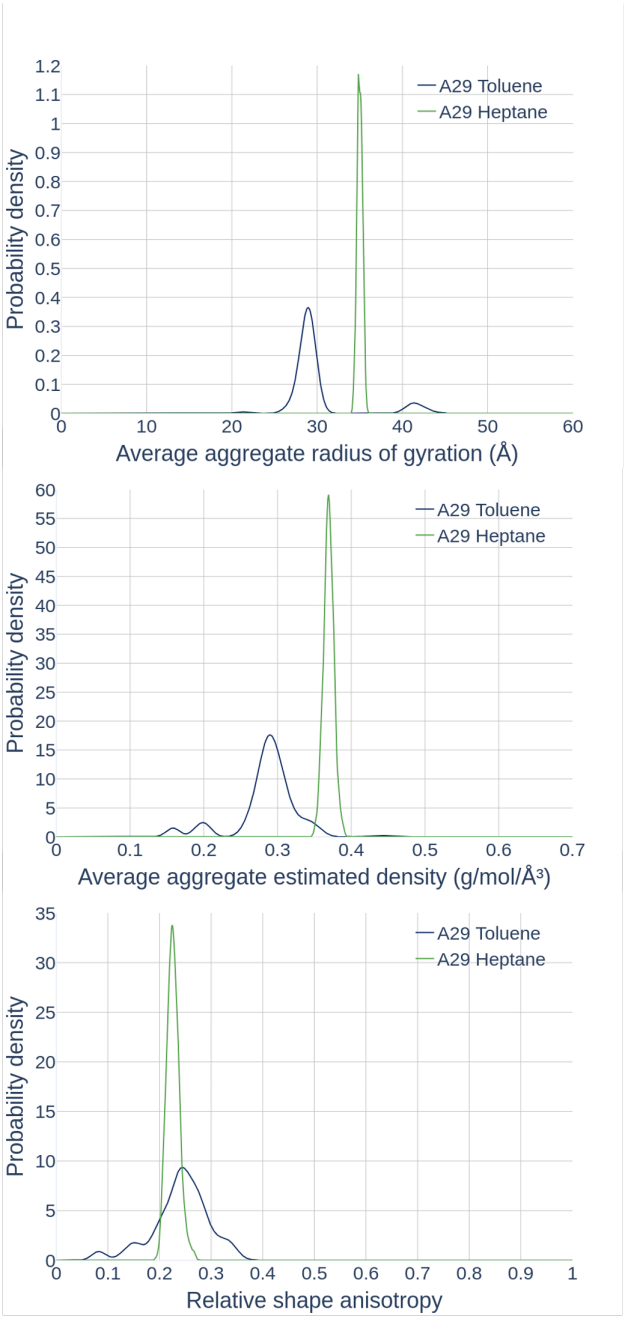


Figure 3: Properties of the asphaltene clusters from simulations of 100 molecules of A29 in toluene and heptane. Top panel: distributions of the cluster radius of gyration. Middle panel: distribution of the estimated cluster density. Bottom panel: distributions of the cluster relative shape anisotropy.

nanoaggregation, the size of the aggregates smoothly increases up to g_n values around 15-17 at 40ns, which is larger than the definition of the Yen-Mullins model.[11,12] Then, between 40ns and 120ns, g_n starts exhibiting a step-wise increase characteristic of the beginning of a clustering stage, in which aggregation occurs between asphaltene clusters. As the simulation time increases, g_n steps increase, illustrating the merger into asphaltene clusters of increasing size until the event at 180ns that results in the combination of the two last remaining clusters. In heptane, the asphaltene model molecules can be classified in two groups: i) the stable A63, A80 and A85 that do not aggregate and ii) the unstable A3, A29 and A54 that aggregate. The relative aggregation ranking of these systems is consistent with previous studies decoding the relationship between structure of the asphaltenes and aggregation.[18,23,31] Indeed, as nanoaggregation has been shown to primarily depend on the size of the aromatic core of asphaltenes, it is to be expected that A29 reaches the highest level of aggregation in this study. Additionally, A29 contains polar aliphatic chains with sulfur heteroatoms that, in spite of their length, contribute favorably to macroaggregation. A29 is followed by A3, which also possess long polar aliphatic chains, and A54, which is made of many heterocycles and short apolar aliphatic chains.

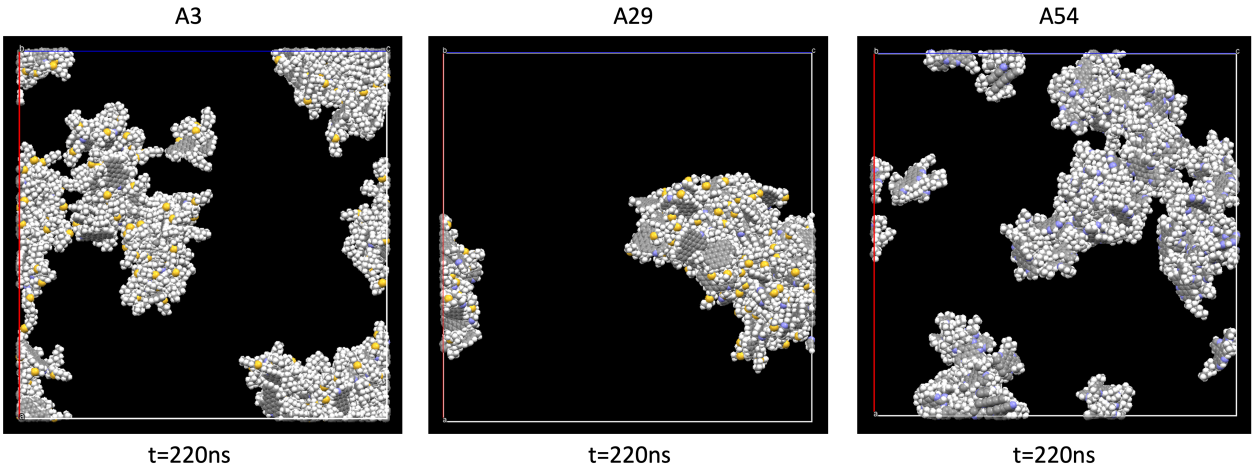


Figure 4: Snapshots from monodisperse simulations of 100 model molecules in heptane for A3 (left), A29 (center), A54 (right) after 220ns. Carbon atoms are represented in gray, nitrogen atoms in blue, sulfur atoms in yellow and hydrogen atoms in white. Full page images for each system are available in Supporting Information Figure S21 to Figure S23.

Over the 240ns simulations, asphaltene aggregates of $g_n \geq 20$ are formed in these three unstable systems. To illustrate their equilibrium states, snapshots of the simulations after 220ns are displayed in Figure 4 and some full page zooms are provided in the Supporting Information Figure S21 to Figure S23. Even tough A3 long polar aliphatic chains do not seem to limit aggregation in comparison with A54, they govern the packing of the aggregates. Indeed, the aromatic cores of A3 and A54 are of similar size, but A3 exhibits ordered parallel stacks of 5 to 10 molecules, sometimes referred as "pancakes stacking", whereas A54 aggregates in a more disordered manner with numerous T-shape interactions between parallel stacks of smaller size, typically 2 to 5 molecules. It seems that even though T-shape $\pi - \pi$ interactions occur for A3, they are not as frequent as with A54. This can be attributed to a combination of factors: the heterocycles of A54 strengthen both parallel and T-shape $\pi - \pi$ interactions between aromatic cores, while the long aliphatic chains of A3 hinder T-shape $\pi - \pi$ interactions. In the case of A29, in spite of the limitations of the static 2-dimensional view, we clearly see a single aggregate of 100 molecules, whereas many clusters are present in the two other systems. Consistently with its larger aggregation number, A29 in heptane also yields aggregates with a much larger radius of gyration: average $R_g = 35.0 \pm 0.3$ Å in comparison with $R_g = 19.6 \pm 2.3$ Å, $R_g = 18.2 \pm 1.6$ Å and $R_g = 12.6 \pm 0.8$ Å, respectively, for A3, A54 and A63 in heptane (Table 2). Nevertheless, despite these differences, Figure S14 shows similar distributions for the estimated density of the aggregates produced by these island systems. The more peaked R_g and density distributions of A29 in comparison with A3, A54 and A63 confirms that A29 has reached a completely equilibrated segregation from the heptane solvent. With respect to the relative shape of the aggregates, as evidenced by the average values in Table 2, the clusters are more elongated than spherical (values close to 0). Among the three asphaltenes stable in heptane, A80 and A85 are of archipelago type, a type of molecules known for its stability, and A63, with its aromatic core of moderate size and absence of heteroatoms, do not present any structural feature favouring aggregation. As mentioned previously, the finite size effects of the simulations affect quantitatively the values

of the observables for A29 in heptane. All details for simulations with 200 asphaltenes in heptane are provided in the Supporting Information (Figure S9, Figure S12, Figure S15, Figure S18 and Table S4) and most of the qualitative conclusions on the differences between A29 and other model molecules still apply. However, it is worth mentioning that the differences in R_g and in density between A3 and A29 do not seems as important with simulations performed with 200 asphaltene molecules (Figure S12 and Figure S15 and Table S4), even though the difference in final g_n is already striking with $g_n = 66.8 \pm 4.5$ for A29 and $g_n = 28.2 \pm 0.1$. This shows that with larger system sizes, more representative of the reality, the simulation time required for the system to settle into the clustering stage of the aggregation process and adopt its characteristics, beyond the step-wise increases visible with 200 asphaltenes A29 in heptane (Figure S9), can be longer than 240ns. To obtain confirmation we have extended up to 500ns the simulations in heptane with 200 molecules of A3 and with 200 molecules of A29, which yields aggregation numbers $g_n = 28.9 \pm 0.6$ and $g_n = 200.0 \pm 0.0$, respectively. The observables for these simulations, plotted in Figure S19 and Figure S20, confirm an important quantitative difference between these two asphaltene model molecules already observed in simulations with 100 asphaltene molecules in spite of the finite size effects. Consequently, the multi-stage aggregation processes observed in early simulation works may be artefacts from simulation finite size effects, but could still be confirmed by longer or larger simulations. However, with the computational power currently available, gaining insights into the late stages of the aggregation process would probably require to resort to coarse-grained simulations.[45] Overall, the representative asphaltene model molecules identified by unsupervised ML show a diversity of aggregation behavior that confirms their suitability for describing asphaltene polydispersity once mixed together.

2.3 Polydisperse mixtures of asphaltene

To study the effect of asphaltene polydispersity on the aggregation process, a series of polydisperse mixtures, made of 3 to 4 types of model molecules for a total concentration of

7 wt % of asphaltene, have been designed based on the results from monodisperse simulations (details in Table 3). While simulations of the first four quaternary mixtures (Q1, Q2, Q3 and Q4) contain 25 molecules of each model, in ternary mixtures (T1, T2, T3 and T4) 34 molecules of the first listed model and 33 of the other two are combined to reach 100 asphaltene molecules per simulation. Q5 and Q6 have been designed slightly differently than the other quaternary mixtures, and contain, respectively, 23 molecules of A23, 33 molecules of A29, and 22 molecules of A54 and A85 for Q5, and 33 molecules of A29, 23 molecules of A63 and 22 molecules of A80 and A85 for Q6. The reasoning behind the specific selection of model molecules in the ternary mixtures is the following: T1 and T4 were designed to mix together the asphaltene model molecules that are stable (A63, A80 and A85 in T1) and unstable (A3, A29 and A54 in T4) in heptane. Then, T2 mixes the two most unstable molecules (A3 and A29) with the less stable of the stable ones, namely A85. Finally, T3 mixes two aggregating asphaltenes (A3 and A54) with A85, but does not include the very strongly aggregating molecule A29. Table 3 summarizes the results of the simulations, and all plots, in analogy to those of the monodisperse simulations, can be found in the Supporting Information Figure S24 to Figure S33.

As could be anticipated from the monodisperse simulations, aggregation in toluene is always quite low in polydisperse simulations (g_n always ≤ 11.0 in Table 3). The aggregation level observed for T1 in heptane is quite low with $g_n = 8.8 \pm 0.1$, which is even lower than the average aggregation from monodisperse simulations ($g_n^{avT1} = 0.34 \times 6.8 + 0.33 \times 14.0 + 0.33 \times 13.7 = 11.5$). Thus, already in that case, we observe that polydispersity does not necessarily yield an aggregation level corresponding to the weighted average of the contributions from monodisperse simulations, but instead an antagonistic effect to the aggregation process can be observed. This antagonistic effect is even more pronounced in T4, for which aggregation only reaches $g_n = 18.8 \pm 0.6$, whereas individually, all monodisperse simulations of the constituents of this mixture exceed $g_n = 23.8$ (see Table 2). Conversely, in T3, aggregation reaches $g_n = 20.0 \pm 2.9$, which is the same range than the weighted average of contributions

Table 3: Summary of the polydisperse simulations performed with 100 asphaltene molecules. In ternary mixtures the are 34 molecules of the first listed model and 33 of the other two. Quaternary mixtures Q1 to Q4 contain 25 molecules of each model. Q5 is a mixture of 23 molecules of A23, 33 molecules of A29, and 22 molecules of A54 and A85. Q6 is a mixture of 33 molecules of A29, 23 molecules of A63 and 22 molecules of A80 and A85. For the aggregation number, g_n^{240}, we report the 20ns average window at the end of the simulation. For the radius of gyration, R_g (in Å), the estimated density and the relative shape anisotrpy, κ^2, we report the average value over the last 40ns of the simulations.

mixture	asph. models	solvent	g_n^{240}	av. R_g	av. density	av. κ^2
T1	A63 A80 A85	toluene	3.3 ± 0.0	13.2 ± 0.8	0.3 ± 0.0	0.3 ± 0.0
T1	A63 A80 A85	heptane	8.8 ± 0.1	14.8 ± 1.3	0.4 ± 0.0	0.2 ± 0.0
T2	A3 A29 A85	toluene	9.5 ± 0.2	15.0 ± 1.5	0.4 ± 0.0	0.3 ± 0.0
T2	A3 A29 A85	heptane	52.6 ± 1.3	27.2 ± 3.7	0.3 ± 0.0	0.3 ± 0.1
T3	A3 A54 A85	toluene	6.2 ± 0.1	12.8 ± 1.0	0.4 ± 0.0	0.2 ± 0.0
T3	A3 A54 A85	heptane	20.0 ± 2.9	15.9 ± 2.8	0.4 ± 0.0	0.2 ± 0.0
T4	A3 A29 A54	toluene	8.3 ± 1.6	15.2 ± 2.7	0.4 ± 0.0	0.2 ± 0.1
T4	A3 A29 A54	heptane	18.8 ± 0.6	17.0 ± 1.1	0.4 ± 0.0	0.1 ± 0.0
Q1	A3 A29 A54 A85	toluene	7.0 ± 0.9	13.6 ± 1.3	0.4 ± 0.0	0.2 ± 0.0
Q1	A3 A29 A54 A85	heptane	18.2 ± 1.1	15.5 ± 1.4	0.4 ± 0.0	0.2 ± 0.0
Q2	A54 A63 A80 A85	toluene	3.9 ± 0.1	12.6 ± 0.8	0.3 ± 0.0	0.3 ± 0.0
Q2	A54 A63 A80 A85	heptane	12.1 ± 0.9	16.3 ± 1.9	0.4 ± 0.0	0.2 ± 0.1
Q3	A29 A63 A80 A85	toluene	4.5 ± 0.1	13.2 ± 1.0	0.3 ± 0.0	0.3 ± 0.0
Q3	A29 A63 A80 A85	heptane	18.2 ± 1.8	18.0 ± 3.3	0.4 ± 0.0	0.3 ± 0.1
Q4	A3 A63 A80 A85	toluene	4.1 ± 0.1	13.1 ± 0.8	0.3 ± 0.0	0.2 ± 0.0
Q4	A3 A63 A80 A85	heptane	10.4 ± 0.3	15.4 ± 1.2	0.4 ± 0.0	0.2 ± 0.0
Q5	A3 A29 A54 A85	toluene	9.0 ± 0.7	13.7 ± 1.3	0.4 ± 0.0	0.2 ± 0.0
Q5	A3 A29 A54 A85	heptane	35.3 ± 1.9	20.3 ± 2.8	0.5 ± 0.0	0.2 ± 0.0
Q6	A29 A63 A80 A85	toluene	4.6 ± 0.0	12.7 ± 0.8	0.4 ± 0.0	0.3 ± 0.0
Q6	A29 A63 A80 A85	heptane	20.0 ± 0.2	16.8 ± 1.1	0.4 ± 0.0	0.2 ± 0.0

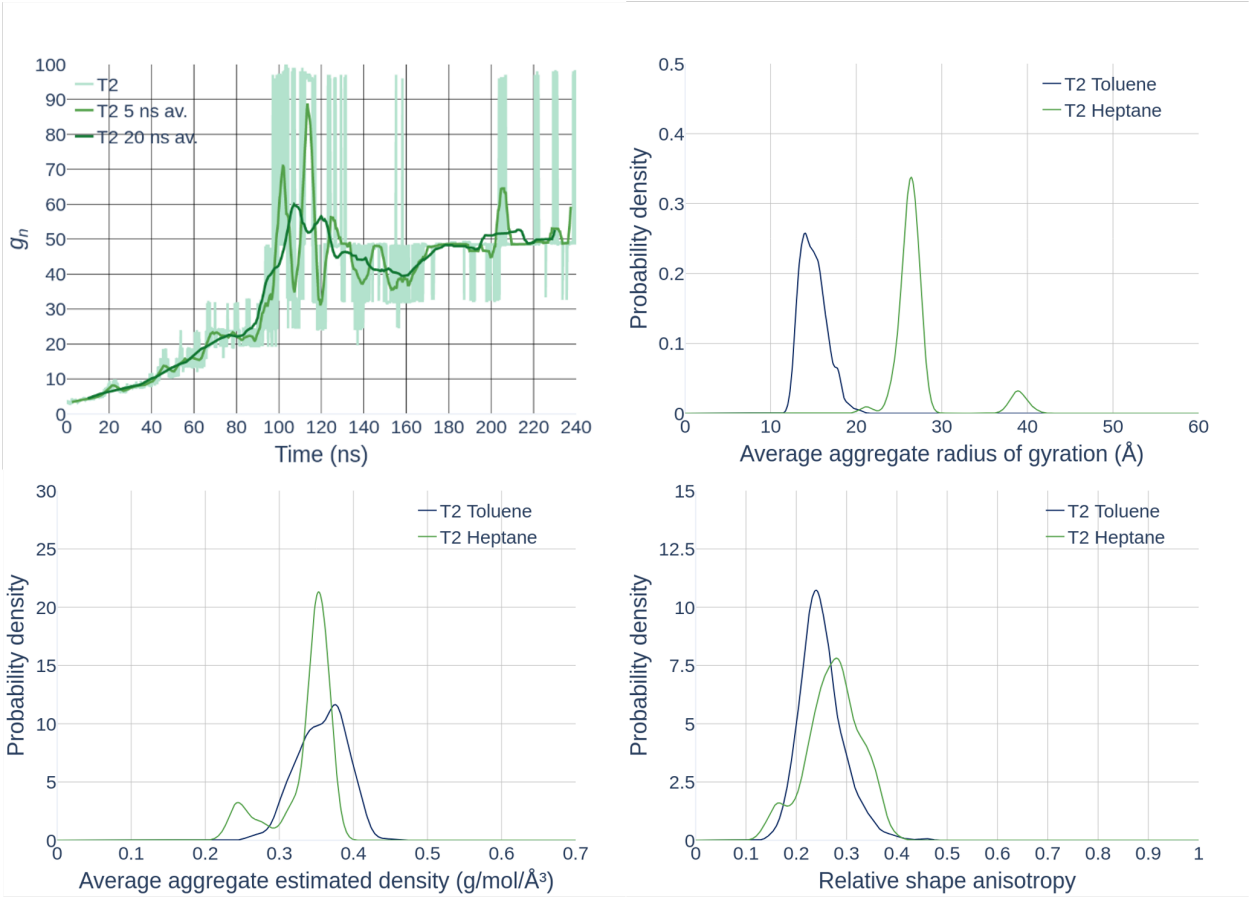


Figure 5: Aggregation number g_n (top left) from simulations of polydisperse mixture T2 in heptane. 5ns and 20ns moving averages are used to filter short term fluctuations of the aggregation dynamics. The distributions of the properties of the asphaltene clusters from simulations of T2 in both heptane and toluene are also represented, namely the radius of gyration (top right), the estimated density (bottom left) and the relative shape anisotropy (bottom right), accumulated over the last 40ns of the simulations.

from monodisperse simulations ($g_n^{avT3} = 0.34 \times 28.1 + 0.33 \times 23.8 + 0.33 \times 13.7 = 21.9$) and in T2, aggregation reaches $g_n = 52.6 \pm 1.3$, which even exceeds the respective average ($g_n^{avT2} = 0.34 \times 28.1 + 0.33 \times 100 + 0.33 \times 13.7 = 47.1$). Therefore, the effect of polydispersity is complex and the aggregation between molecules of different types is driven by a sum of correlated contributions from aromatic cores, aliphatic chains or heteroatoms. Comparing T2 and T4, it seems that there might be a lower compatibility in the aggregation of molecules A29 with molecules A54 than in the aggregation between A29 and A85 that results to be more favorable. Simulations of binary mixtures in heptane, not reported in this paper for

the sake of conciseness, mixing the three model molecules A29, A54 and A85 have confirmed this observation with $g_n[0.50 \times \text{A29} + 0.50 \times \text{A54}] \leq g_n[0.50 \times \text{A29} + 0.50 \times \text{A85}]$, whereas in monodisperse simulations the aggregation of A54 is clearly superior to the aggregation of A85 ($g_n = 23.8 \pm 1.4$ for A54 *vs* $g_n = 13.7 \pm 1.0$ for A85). A similar conclusion can be drawn looking at the compatibility of A54 and A85 with molecule A3. Thus, the archipelago nature of A85, even though usually associated with stable asphaltenic systems, seems to facilitate the aggregation – when combined with bulky asphaltene such A3 and A29 as in T2 – more than the island nature of A54 with very short aliphatic chain, as in T4. Figure 5 displays all aggregation observables for T2, which exhibits the largest aggregation among the polydisperse mixtures studied, with $g_n = 52.6 \pm 1.3$. It is the only mixture to reach the clustering stage of the aggregation as confirmed by the step-wise increases of its aggregation number (see Figure 25).

To further test the effects of polydispersity, the aggregation of quaternary mixtures has been studied. In particular, Q5 and Q6 mixtures have been designed to test both the driving potential of A29 and the synergistic action of A29 and A85 with respect to aggregation. Conversely to T2, no synergistic effect is observed at the specific compositions in Q5 nor Q6 (aggregation does not exceed the weighted average contributions from the monodisperse simulations) even tough a seed effect from the molecules A29 can be detected comparing T1 with Q3 and Q6 and T3 with Q1 and Q5. However, this seed effect is not linearly correlated with the composition of the systems. For example in Q3, there is the same number of each model molecule (25 molecules) A63, A80, A85, and A29, and an important increase in aggregation ($g_n = 18.2 \pm 1.1$) is observed in comparison with T1 ($g_n = 8.8 \pm 0.1$). However, in Q6 there are more molecules of the very strongly aggregating A29 (33 molecules) in comparison with Q3, but yielding a proportionally smaller increase in aggregation ($g_n = 20.0 \pm 0.2$). Thus it seems that in this case the seed effect of molecules A29 reaches a saturation point. Conversely, in Q3, which contains the same number of each model molecule (25 molecules) A3, A54, A85, and A29, the presence of 25 of the strongly aggregating A29

molecules does not lead to a larger aggregation ($g_n = 18.2 \pm 1.8$) than in T3 ($g_n = 20.0 \pm 2.9$), which only contains molecules of A3, A54, and A85. Meanwhile, in Q5, where the number of molecules of A29 is set to 33, a clear increase in aggregation is observed ($g_n = 35.3 \pm 1.9$), as if there were a concentration threshold to overcome within this polydisperse mixture for the seed effect to be activated. Besides, comparing Q2 and Q4 with Q3 shows that A29 is clearly a stronger aggregation seed than A3 and A54, but no synergistic effect is observed among the quaternary mixtures investigated in this work. Finally, a word of caution must be mentioned here: even though the conclusions presented just above seem reasonable, it is important to keep in mind that they are only based on small differences obtained from single run simulations. Ideally, one would like to perform many simulations per system in order to draw stronger conclusions. However, the computational cost of such MD simulations makes a systematic n-repetition process prohibitive.

Overall, the investigation of the asphaltene aggregation process in these polydisperse mixtures has shown a variety of complex and correlated effects: antagonistic and synergistic effects within mixtures and seed effects of specific model molecules, A29 in particular, have been detected. These findings highlight once again the necessity of accounting for asphaltene polydispersity even though the level of asphaltene aggregation in polydisperse mixtures never reaches the largest aggregation level of monodisperse simulations of A29. Moreover, as the ultimate goal is to contribute to the design of asphaltene inhibitors, it is important to ensure that the *in-silico* evaluation of their performance is not biased by a specific asphaltene model molecule with which the inhibitor could interact a lot and limit its aggregation whereas it could interact more moderately with other molecules, depending on their chemistry. Accounting for polydispersity in this characterization permits to limit this risk. Besides, in order to be able to capture the action of an inhibitor on the polydisperse mixture of asphaltenes, it is necessary to use a mixture which reaches significant level of aggregation. Therefore, T2 is the most suited polydisperse mixture of asphaltenes for this task and has been employed in the simulations presented in the next section.

2.4 Showcase of inhibition simulation

Beyond the investigation of asphaltene aggregation and the effect of polydispersity on this process, another objective of this work is to set up a robust protocol for the *in-silico* characterization of the action of asphaltene inhibitors. To illustrate that the workflow implemented in this study has permitted to reach this objective, a simulation of the aggregation in heptane of the polydisperse mixture T2 in presence of a nonylphenol resin asphaltene inhibitor, at a concentration of 1 wt %, has been performed. Due to Intellectual Property restrictions, the exact form of the inhibitor cannot be published, but we show in Figure 6 its general structure, which is sufficient for the purpose of this showcase.

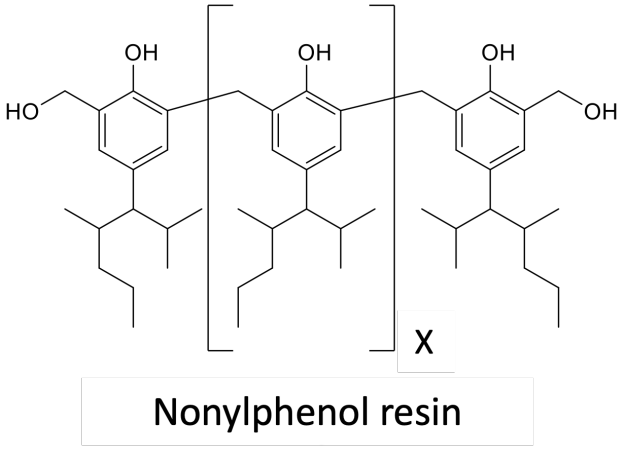


Figure 6: 2D representation of the molecular structure of the inhibitor.

The aggregation behavior of the mixture T2 in presence of the inhibitor (in black) is compared in Figure 7 to the case without (in green) already discussed in the previous section. To facilitate the comparison and avoid short term fluctuations, the 20ns moving averages are presented for the aggregation number. Moreover, while in the aggregation number only the asphaltene molecules are considered, the inhibitor molecules are included in the calculation of the radius of gyration of the cluster they belong to, in order to avoid drawing erroneous conclusions from artificially low aggregation numbers resulting from the limits of this metrics. An example of this type of artefact would be the following case: if two clusters of asphaltene were connected by one or two inhibitors in between while not interacting directly with each

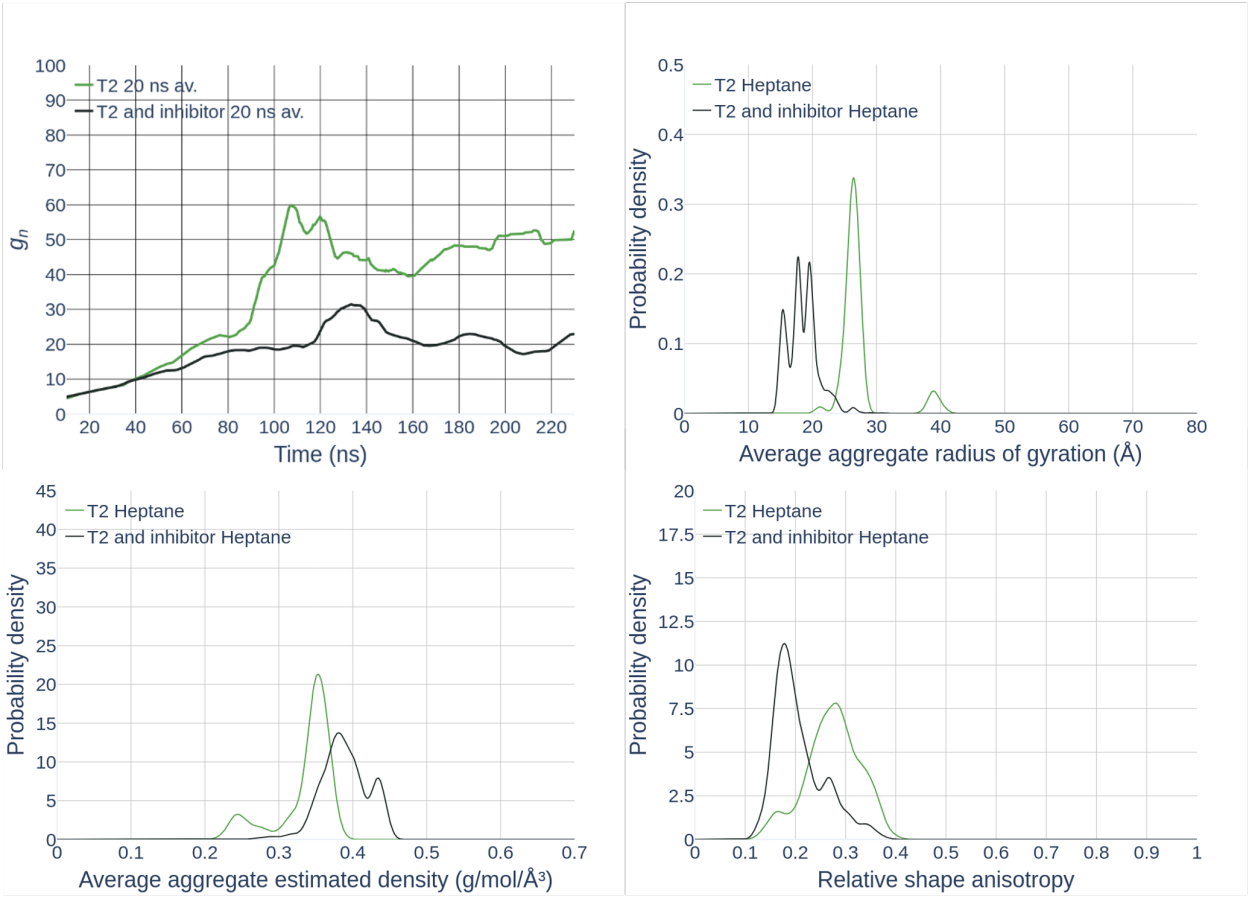


Figure 7: Aggregation number g_n (top left) from simulations of polydisperse mixture T2 in heptane with (in black) and without (in green) inhibitor molecules. 20ns moving averages are used to filter short term fluctuations of the aggregation dynamics. The distributions of the properties of the asphaltene clusters from these simulations are also represented (using the same color code), namely the radius of gyration (top right), the estimated density (bottom left) and the relative shape anisotropy (bottom right), accumulated over the last 40ns of the simulations.

other, the aggregation number would be low whereas there would actually be a very large cluster of asphaltene and inhibitors. On the other hand, the inclusion of the inhibitors in the calculation of the radius of gyration would yield large values, thus revealing the limitation of the inhibitor performance. The nonylphenol resin inhibitor manages to limit the aggregation of the T2 mixture to $g_n = 23.0 \pm 1.9$, which is less than half the aggregation number of the case without inhibitor, as summarized in Table 4. The average radius of gyration is also reduced, which confirms the very good performance of this inhibitor. Considering that the concentration of inhibitor is only 1 wt %, we can conclude that the nonylphenol resin inhibitor is qualitatively (a different definition of the aggregation number was used) a better performing asphaltene inhibitor than n-Octylphenol,[27] which needed 7 wt % to limit the aggregation of less aggregating asphaltenic systems. Nevertheless, it is worth noting that such concentration, namely 1 wt %, is still two orders of magnitude larger than usual operating conditions. However, further decreasing the concentration of the inhibitor would require to perform the simulations on much larger systems, yielding a computational cost far beyond the reach of any systematic study.

Table 4: Summary of the polydisperse simulations of T2 with and without inhibitor. For the aggregation number, g_n^{240}, we report the 20ns average window at the end of the simulation. For the radius of gyration, R_g (in Å), the estimated density and the relative shape anisotrpy, κ^2, we report the average value over the last 40ns of the simulations.

mixture	asph. models	inhibitor	solvent	g_n^{240}	av. R_g	av. density	av. κ^2
T2	A3 A29 A85	No	heptane	52.6 ± 1.3	27.2 ± 3.7	0.3 ± 0.0	0.3 ± 0.1
T2	A3 A29 A85	1 wt %	heptane	23.0 ± 1.9	18.5 ± 2.4	0.4 ± 0.0	0.2 ± 0.1

3 Conclusion

In this study, the combination of unsupervised machine learning and molecular dynamics simulation has permitted to thoroughly investigate the role of asphaltene polydispersity on the aggregation process. Indeed, we first performed an upfront selection, via unsupervised machine learning, of a series of asphaltene model molecules representative of a broad and

diverse catalogue specifically designed for the purpose of molecular dynamics simulations. Then, we have studied the aggregation of these molecules via monodisperse simulations in toluene and heptane solvents. At the exception of the most strongly aggregating molecule, namely A29, aggregation in toluene has been weak. In heptane, monodisperse simulations—in agreement with recent simulation works—have shown that even though $\pi - \pi$ interactions can be a very strong driver for asphaltene aggregation other molecular features such as polar aliphatic chains and heteroatoms significantly contribute to this process. Furthermore, the different aggregation behavior of the representative asphaltene model molecules in heptane have confirmed their ability to capture the diversity of asphaltene. Afterwards, the effect of asphaltene polydispersity on the aggregation process has been investigated by simulations of ternary and quaternary polydisperse mixtures of these molecules. Overall the effect of asphaltene polydispersity is complex, difficult to disentangle and diverse: depending on the composition of the mixture both antagonistic, synergistic and seed effects have been observed. These findings illustrate again the necessity to account for polydispersity when studying the asphaltene aggregation process and its inhibition. Finally, this work has also permitted to deploy a robust simulation protocol for the *in-silico* evaluation of the performance of asphaltene inhibitors, as demonstrated by the case study presented with the nonyphenol resin inhibitor. In future works we intend to build on the developments presented here to investigate and compare the behavior of a series of asphaltene inhibitors.

4 Computational methods

4.1 Unsupervised machine learning

The first step to select a series of representative asphaltene model molecules from the catalogue of Law et al.[38] was to generate digital molecular structures of the model molecules released in the form of 2D and 3D molecular representations. To accelerate the digitalization, we have used an open-source optical chemical structure recognition (OCSR) Java-based

tool called MolVec.[46] Despite such tool, this is still a cumbersome process, but it is worth pointing out that OCSR tools have been reviewed recently,[47] and that many developments are currently under process in this field. Additional comments on the information of the original catalogue of 100 molecules are presented in the Supporting Information. SMILES codes (Simplified Molecular Input Line Entry System)[48] of the model molecules have been obtained via MolVec and their 3D molecular structures have been generated using RDKit.[49] The 3D molecular structures have been used as input for the calculation of 3D molecular descriptors by the free software Mordred,[50] and later on to perform MD simulations when deemed relevant.

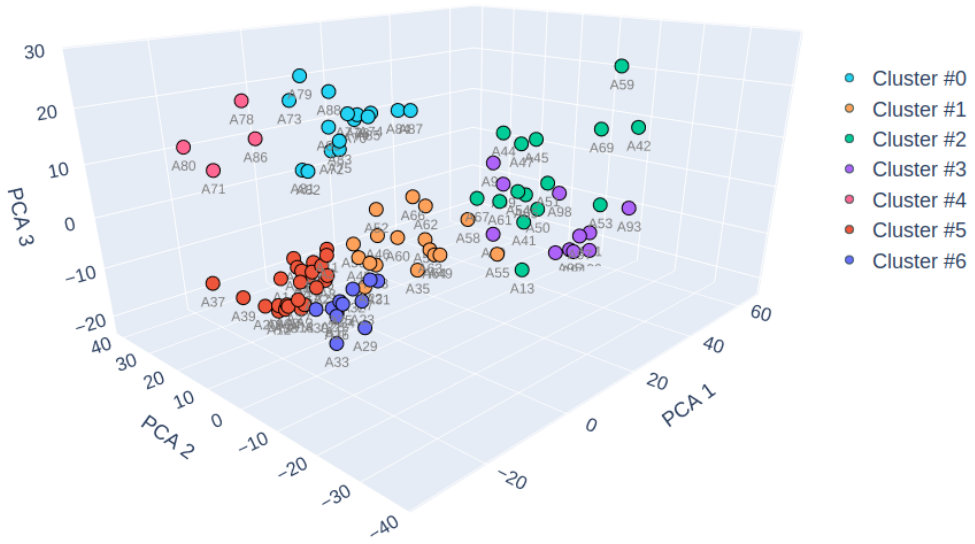


Figure 8: Representation of the catalogue of 100 molecules from Law et al.[38] after 3D principal component analysis and Kmeans clustering analysis. Each point represents a molecule and colors reflect cluster assignation. The labels of the molecules are consistent with the original work.[38]

At this point, each of the 100 molecules from the catalogue of Law et al.[38] is described by 1826 3D molecular descriptors. Then, this highly multidimensional representation of the chemical space of asphaltene is reduced to three dimensions via principal component analysis (PCA),[51] a linear dimensionality reduction method, using the implementation of the scikit-learn library.[52,53] In order to identify groups of similar molecules, an unsupervised cluster analysis has been performed, using the standard Kmeans algorithm of scikit-learn. Figure 8

displays the 3D PCA representation of the catalogue of 100 molecules using one color per cluster. It is worth mentioning that the first 3 PCA components account for 37.5%, 21.9% and 8.8% of the explained variance ratio, respectively, hence a total of 68.2%. In this case the Kmeans algorithm had been set up to identify 7 clusters, and it is interesting to point out that the 11 resin molecules have correctly been assigned to a cluster of their own, namely cluster #3 in Figure 1, leaving 6 clusters of asphaltene. More details, such as the effect of the dimensionality reduction method (either using PCA or the uniform manifold approximation and projection method, UMAP, which is an non-linear method[54]), the difference between performing the dimensionality reduction before or after the cluster analysis, and the effect of the parameters of the clustering analysis (choice of number of clusters and choice of clustering algorithm) are reported in the Supporting Information (Figure S1 to Figure S4 and Table S1 and Table S2). However, it is important to mention that we have verified that such details only affected the cluster assignation of a few molecules at the frontier between clusters. When large numbers of molecules are considered in cluster analysis, the closest molecule to each cluster centroid is often chosen as representative for the cluster. In our specific case, the number of molecules is moderate, and therefore we have looked at all the assignations and have chosen the molecules represented in Figure 1 as representative of their clusters, ensuring that none of these was at the frontier between clusters and affected by the setup of the clustering analysis. These molecules have subsequently been used in the MD simulations. The approach reported here is general and can be extended to include future developments of asphaltene models. As shown in Figure 1, the 6 molecules selected as representative are very diverse, with 4 island and 2 archipelago asphaltenes. In particular the island models have very different structural features: A3 has long side-chains, A54 has a many heterocycles and A29 is very bulky. The detailed composition of each cluster is given in Supporting Information.

4.2 Simulation details

While a variety of force fields (FF), both all atoms, united-atoms, and coarse-grained, have been used for the simulation of asphaltene aggregation,[27] overall any modern and well-validated FF can be considered a reasonable choice, as pointed out by Headen et al.[31] Indeed, by way of comparison, there are still uncertainties about larger issues, such as asphaltene structures and the exact composition of their systems. In this work, we have used the GAFF force field[55] in combination with AM1-BCC atomic partial charges, as validated by Glova et al.[44] Besides, the GAFF force field had already been used in asphaltene simulations.[56–58] The topologies of the simulated molecules (available in Supporting Information additional files) were generated using the ACPYPE tool,[59] which builds on Antechamber.[60] GPU-accelerated MD simulations were performed with the GROMACS simulation code (version 2020.4),[61–64] which also served to construct the simulation boxes via random insertion of molecules (both for position and orientation). Monodisperse simulations with 40, 100 and 200 asphaltene molecules served as benchmark, before settling on 100 molecules as the best compromise between system size and simulation cost. The number of solvent molecules (toluene and heptane) in each cubic simulation box was defined such as to ensure an asphaltene concentration of 7 wt % for each system. Therefore, the largest simulated system, namely 200 molecules of asphaltene model A3 in heptane, contained 621,082 atoms. When the aggregation inhibitor, a nonylphenol resin, was included in the simulations, its number of molecules was obtained from the 7:1 targeted ratio between asphaltenes and inhibitor. Then the number of solvent molecules was tuned to adjust the concentration of the inhibitor at 1 wt %.

The adopted simulation protocol can be summarized as follow. After construction of a cubic periodic simulation system, a steepest decent energy minimization is performed until all forces decrease below 100kJ/mol/nm. Then, a 3ns MD equilibration simulation in the isobaric-isothermal (NPT) ensemble is run using GROMACS' velocity rescaling thermostat[65] and a Berendsen barostat.[66] Afterwards, production MD simulation are carried out for 240ns with Nosé-Hoover thermostat[67,68] and Parrinello-Rahman barostat.[69,70] All simu-

lations are performed at a temperature of 300K and a pressure of 1bar. Equations of motion are integrated using the leapfrog algorithm[71] with a timestep of 2ps while keeping hydrogen bonds rigid via the LINCS algorithm.[72] To account for long-range electrostatic interactions the particle-mesh Ewald (PME) algorithm[73] is employed, whereas a plain cut-off (PME cannot be used with GPU yet) with standard correction for energy and pressure is adopted for long-range dispersion interactions, in both cases with a cutoff value of 1.25nm.

4.3 Aggregation observables

The aggregation number captures the aggregation state of asphaltenic systems by counting the number of asphaltene molecules constituting an aggregate (also called cluster). In this study, two molecules are considered to belong to the same aggregate if the shortest distance between atoms of the two molecules is inferior to a threshold value of 3.5 Å. This definition follows the findings of Headen et al.,[31] who showed that, when asphaltene molecules are clustered, their shortest distance clearly decreases below 3.5 Å. Moreover, Ghamartale et al.,[27] who studied the same asphaltene molecules, argued that such threshold is applicable because the range of hydrogen bond length is between 2.70 and 3.30 Å. Furthermore, they provided an interesting discussion about the different possibilities for defining such criterion, with a focus on the effect of using distances calculated between the center of mass between molecules (instead of interatomic distances) which can be less suited to properly account for irregular packing of aggregates. Both number-average aggregation number, g_n and z-average aggregation number, g_z have been used in asphaltene publications. Even though some prefer using g_z[18,27] over g_n,[31] no compelling argument was found in the papers nor in the original reference[74] that actually contains three definitions, namely, apart from the two presented here, the weight-average aggregation number g_w, which can all be obtained from different experimental techniques: g_n via membrane osmometry, g_w via static light scattering and g_z via intrinsic viscosity measurements. As this study relies on MD simulations, we have followed – as for other observables – the definitions from Headen et al.,[31] and used g_n as aggregation

number, which is strictly the average number of asphaltene molecules per aggregate:

$$g_n = \frac{\sum_i n_i g_i}{\sum_i n_i} \tag{1}$$

with n_i the number of aggregates of g_i molecules. It is important to precise that the sums of (1) start from 2, thus the monomers are excluded. To quantify the size of polymers or macromolecules in solution,[75] the radius of gyration (R_g) is defined as:

$$R_g^2 = \frac{1}{N}\sum_i^N (\boldsymbol{r}_i - \boldsymbol{r}_{cm})^2 \tag{2}$$

with $\boldsymbol{r}_i$ the position vector of atom i and $\boldsymbol{r}_{cm}$ the position vector of the center of mass of the aggregate. Furthermore, information relative to the shape of the aggregate can be extracted from the gyration tensor ($\boldsymbol{S}$):[76]

$$\boldsymbol{S} = \frac{1}{N}\begin{pmatrix} \sum_i (x_i - x_{cm})^2 & \sum_i (x_i - x_{cm})(y_i - y_{cm}) & \sum_i (x_i - x_{cm})(z_i - z_{cm}) \\ \sum_i (y_i - y_{cm})(x_i - x_{cm}) & \sum_i (y_i - y_{cm})^2 & \sum_i (y_i - y_{cm})(z_i - z_{cm}) \\ \sum_i (z_i - z_{cm})(x_i - x_{cm}) & \sum_i (z_i - z_{cm})(y_i - y_{cm}) & \sum_i (z_i - z_{cm})^2 \end{pmatrix} \tag{3}$$

in which the sums run over all atoms i of the aggregate and cm again refers to the center of mass. The diagonalization of the gyration tensor, $\boldsymbol{S} = diag(\lambda_1, \lambda_2, \lambda_3)$, permits to obtain the eigenvalues (principal moments) ordered as $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Alternatively to (2), the radius of gyration can directly be obtained from the sum of the eigenvalues: $R_g^2 = \lambda_1 + \lambda_2 + \lambda_3$. From these eigenvalues, an estimation of the dimensionality and the symmetry of the aggregates can be provided by κ^2, the relative shape anisotropy:

$$\kappa^2 = 1 - 3\frac{(\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_3\lambda_1)}{(\lambda_1 + \lambda_2 + \lambda_3)^2} \tag{4}$$

κ^2 values span between 0, for a perfectly spherical cluster, and 1, for a linear chain. Still from the eigenvalues, it is possible to estimate the density of the asphaltene aggregates. The

volume of each aggregate is approximated by the volume of an hypothetical effective ellipsoid having the same principal moments as the gyration tensor. Hence the axes a, b and c of such ellipsoid would be equal to $\sqrt{5^3\lambda_n}(n = 1 - 3)$ and the volume encompassing the aggregate is:

$$V_{cluster} = \frac{4}{3}\pi\sqrt{5^3\lambda_1\lambda_2\lambda_3} \tag{5}$$

The mass of the aggregate is calculated as the sum of the mass of each asphaltene molecule populating the cluster. Therefore, the estimated density writes as:

$$\rho_{cluster} = \frac{\sum_i m_i}{V_{cluster}} \tag{6}$$

where m_i is the mass of the ith molecule of the aggregate. In this work, the atomic positions have been written out each 10ps (5,000 time steps), and thus all observables of the aggregation state of the system, namely the aggregation number, radius of gyration, relative shape anisotropy and density estimate, have been computed for trajectory frames each 10ps. The aggregation number is represented with both 5ns and 20ns moving average to better guide the eye. Additionally, in order to describe the equilibrium state of the asphaltenic systems, the average of the other observables in each recorded trajectory frame and their distributions are calculated during the last 40ns of each run (from 200 to 240ns of MD simulation).

Acknowledgement

The authors thank Clariant for the financial support and to allow for the work to be published. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801342 (Tecniospring INDUSTRY) and the Government of Catalonia's Agency for Business Competitiveness (ACCIÓ). We acknowledge PRACE for awarding us access to JUWELS at GCS@FZJ, Germany

Supporting Information Available

See the Supporting Information for Tables and Figures with additional details and results from the ML procedure and MD simulations.

References

(1) Alexander H. Tullo, The future of oil is in chemicals, not fuels. *C&EN Global Enterprise* **2019**, *97*, 26–29.

(2) Kelland, M. A. *Production Chemicals for the Oil and Gas Industry*, second edition ed.; Taylor & Francis Group, 2014.

(3) Mullins, O. C.; Pomerantz, A. E.; Andrews, A. B.; Dutta Majumdar, R.; Hazendonk, P.; Ruiz-Morales, Y.; Goual, L.; Zare, R. N. In *Springer Handbook of Petroleum Technology*; Hsu, C. S., Robinson, P. R., Eds.; Springer International Publishing: Cham, 2017; pp 221–250.

(4) Wylde, J.; Punase, A. Asphaltenes: A Complex and Challenging Flow Assurance Issue To Measure and Quantify Risk. *Journal of Petroleum Technology* **2020**, *72*, 45–48.

(5) Cheng, R.; Zou, R.; He, L.; Liu, L.; Cao, C.; Li, X.; Guo, X.; Xu, J. Effect of Aromatic Pendants in a Maleic Anhydride- co-Octadecene Polymer on the Precipitation of Asphaltenes Extracted from Heavy Crude Oil. *Energy and Fuels* **2021**, *35*, 10562–10574.

(6) Zhu, Q.; Lin, B.; Yan, Z.; Yao, Z.; Cao, K. Influences of Molecular Structure of Poly(styrene-co-octadecyl maleimide) on Stabilizing Asphaltenes in Crude Oil. *Energy and Fuels* **2020**, *34*, 3057–3064.

(7) Liu, D.; Zhang, H.; Li, C.; Yang, F.; Sun, G.; Yao, B. Experimental Investigation on the Interactions between Asphaltenes and Comb-like Octadecyl Acrylate (OA) Polymeric Flow Improvers at the Model Oil/Water Interface. *Energy and Fuels* **2020**,

(8) Firoozinia, H.; Fouladi Hossein Abad, K.; Varamesh, A. A comprehensive experimental evaluation of asphaltene dispersants for injection under reservoir conditions. *Petroleum Science* **2016**, *13*, 280–291.

(9) Wang, X.; Zhang, H.; Liang, X.; Shi, L.; Chen, M.; Wang, X.; Liu, W.; Ye, Z. New Amphiphilic Macromolecule as Viscosity Reducer with Both Asphaltene Dispersion and Emulsifying Capacity for Offshore Heavy Oil. *Energy and Fuels* **2021**, *35*, 1143–1151.

(10) Kashefi, S.; Shahrabadi, A.; Jahangiri, S.; Lotfollahi, M. N.; Bagherzadeh, H. Investigation of the performance of several chemical additives on inhibition of asphaltene precipitation. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects* **2016**, *38*, 3647–3652.

(11) Mullins, O. C. The modified yen model. *Energy and Fuels* **2010**, *24*, 2179–2207.

(12) Mullins, O. C. *et al.* Advances in asphaltene science and the Yen-Mullins model. *Energy and Fuels* **2012**, *26*, 3986–4003.

(13) Gray, M. R.; Tykwinski, R. R.; Stryker, J. M.; Tan, X. Supramolecular assembly model for aggregation of petroleum asphaltenes. *Energy and Fuels* **2011**, *25*, 3125–3134.

(14) Murgich, J. Intermolecular forces in aggregates of asphaltenes and resins. *Petroleum Science and Technology* **2002**, *20*, 983–997.

(15) Mullins, O. C. The asphaltenes. *Annual Review of Analytical Chemistry* **2011**, *4*, 393–418.

(16) Headen, T. F.; Boek, E. S.; Skipper, N. T. Evidence for asphaltene nanoaggregation in toluene and heptane from molecular dynamics simulations. *Energy and Fuels* **2009**, *23*, 1220–1229.

(17) Headen, T. F.; Boek, E. S. Potential of mean force calculation from molecular dynamics

simulation of asphaltene molecules on a calcite surface. *Energy and Fuels* **2011**, *25*, 499–502.

(18) Sedghi, M.; Goual, L.; Welch, W.; Kubelka, J. Effect of asphaltene structure on association and aggregation using molecular dynamics. *Journal of Physical Chemistry B* **2013**, *117*, 5765–5776.

(19) Santos Silva, H.; Sodero, A. C.; Bouyssiere, B.; Carrier, H.; Korb, J. P.; Alfarra, A.; Vallverdu, G.; Bégué, D.; Baraille, I. Molecular Dynamics Study of Nanoaggregation in Asphaltene Mixtures: Effects of the N, O, and S Heteroatoms. *Energy and Fuels* **2016**, *30*, 5656–5664.

(20) Sodero, A. C.; Santos Silva, H.; Guevara Level, P.; Bouyssiere, B.; Korb, J. P.; Carrier, H.; Alfarra, A.; Bégué, D.; Baraille, I. Investigation of the Effect of Sulfur Heteroatom on Asphaltene Aggregation. *Energy and Fuels* **2016**, *30*, 4758–4766.

(21) Santos Silva, H.; Sodero, A. C.; Korb, J. P.; Alfarra, A.; Giusti, P.; Vallverdu, G.; Bégué, D.; Baraille, I.; Bouyssiere, B. The role of metalloporphyrins on the physical-chemical properties of petroleum fluids. *Fuel* **2017**, *188*, 374–381.

(22) Santos Silva, H.; Alfarra, A.; Vallverdu, G.; Bégué, D.; Bouyssiere, B.; Baraille, I. Sensitivity of Asphaltene Aggregation toward the Molecular Architecture under Desalting Thermodynamic Conditions. *Energy and Fuels* **2018**, *32*, 2681–2692.

(23) Santos Silva, H.; Alfarra, A.; Vallverdu, G.; Bégué, D.; Bouyssiere, B.; Baraille, I. Asphaltene aggregation studied by molecular dynamics simulations: role of the molecular architecture and solvents on the supramolecular or colloidal behavior. *Petroleum Science* **2019**, *16*, 669–684.

(24) Santos Silva, H.; Alfarra, A.; Vallverdu, G.; Bégué, D.; Bouyssiere, B.; Baraille, I. Role of the porphyrins and demulsifiers in the aggregation process of asphaltenes at water/oil

interfaces under desalting conditions: a molecular dynamics study. *Petroleum Science* **2020**, *17*, 797–810.

(25) Goual, L.; Sedghi, M. Role of ion-pair interactions on asphaltene stabilization by alkylbenzenesulfonic acids. *Journal of Colloid and Interface Science* **2015**, *440*, 23–31.

(26) Goual, L.; Sedghi, M.; Wang, X.; Zhu, Z. Asphaltene aggregation and impact of alkylphenols. *Langmuir* **2014**, *30*, 5394–5403.

(27) Ghamartale, A.; Zendehboudi, S.; Rezaei, N. New Molecular Insights into Aggregation of Pure and Mixed Asphaltenes in the Presence of n-Octylphenol Inhibitor. *Energy and Fuels* **2020**, *34*, 13186–13207.

(28) Ghamartale, A.; Zendehboudi, S.; Rezaei, N.; Chatzis, I. Effects of inhibitor concentration and thermodynamic conditions on n-octylphenol-asphaltene molecular behaviours. *Journal of Molecular Liquids* **2021**, *340*, 116897.

(29) Lowry, E.; Sedghi, M.; Goual, L. Polymers for asphaltene dispersion: Interaction mechanisms and molecular design considerations. *Journal of Molecular Liquids* **2017**, *230*, 589–599.

(30) Headen, T. F.; Hoepfner, M. P. Predicting Asphaltene Aggregate Structure from Molecular Dynamics Simulation: Comparison to Neutron Total Scattering Data. *Energy and Fuels* **2019**, *33*, 3787–3795.

(31) Headen, T. F.; Boek, E. S.; Jackson, G.; Totton, T. S.; Müller, E. A. Simulation of Asphaltene Aggregation through Molecular Dynamics: Insights and Limitations. *Energy and Fuels* **2017**, *31*, 1108–1125.

(32) Javanbakht, G.; Sedghi, M.; Welch, W. R.; Goual, L.; Hoepfner, M. P. Molecular polydispersity improves prediction of asphaltene aggregation. *Journal of Molecular Liquids* **2018**, *256*, 382–394.

(33) Villegas, O.; Salvato Vallverdu, G.; Bouyssiere, B.; Acevedo, S.; Castillo, J.; Baraille, I. Molecular Cartography of A1 and A2 Asphaltene Subfractions from Classical Molecular Dynamics Simulations. *Energy and Fuels* **2020**, *34*, 13954–13965.

(34) Mullins, O. C.; Martínez-Haya, B.; Marshall, A. G. Contrasting perspective on asphaltene molecular weight. This Comment vs the Overview of A. A. Herod, K. D. Bartle, and R. Kandiyoti. *Energy and Fuels* **2008**, *22*, 1765–1773.

(35) Greenfield, M. L. Molecular modelling and simulation of asphaltenes and bituminous materials. *International Journal of Pavement Engineering* **2011**, *12*, 325–341.

(36) Sjöblom, J.; Simon, S.; Xu, Z. Model molecules mimicking asphaltenes. *Advances in Colloid and Interface Science* **2015**, *218*, 1–16.

(37) Martín-Martínez, F. J.; Fini, E. H.; Buehler, M. J. Molecular asphaltene models based on Clar sextet theory. *RSC Advances* **2015**, *5*, 753–759.

(38) Law, J. C.; Headen, T. F.; Jiménez-Serratos, G.; Boek, E. S.; Murgich, J.; Müller, E. A. Catalogue of Plausible Molecular Models for the Molecular Dynamics of Asphaltenes and Resins Obtained from Quantitative Molecular Representation. *Energy and Fuels* **2019**, *33*, 9779–9795.

(39) Schuler, B.; Meyer, G.; Peña, D.; Mullins, O. C.; Gross, L. Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *Journal of the American Chemical Society* **2015**, *137*, 9870–9876.

(40) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs. *Energy and Fuels* **2017**, *31*, 13509–13518.

(41) Acevedo, N.; Moulian, R.; Chacón-Patiño, M. L.; Mejia, A.; Radji, S.; Daridon, J. L.; Barrère-Mangote, C.; Giusti, P.; Rodgers, R. P.; Piscitelli, V.; Castillo, J.; Carrier, H.;

Bouyssiere, B. Understanding Asphaltene Fraction Behavior through Combined Quartz Crystal Resonator Sensor, FT-ICR MS, GPC ICP HR-MS, and AFM Characterization. Part I: Extrography Fractionations. *Energy and Fuels* **2020**, *34*, 13903–13915.

(42) Boek, E. S.; Yakovlev, D. S.; Headen, T. F. Quantitative molecular representation of asphaltenes and molecular dynamics simulation of their aggregation. *Energy and Fuels* **2009**, *23*, 1209–1219.

(43) Naughton, F. B.; Alibay, I.; Barnoud, J.; Barreto-Ojeda, E.; Beckstein, O.; Bouysset, C.; Cohen, O.; Gowers, R. J.; MacDermott-Opeskin, H.; Matta, M.; Melo, M. N.; Reddy, T.; Wang, L.; Zhuang, Y. MDAnalysis 2.0 and beyond: fast and interoperable, community driven simulation analysis. *Biophysical Journal* **2022**, *121*, 272a–273a.

(44) Glova, A. D.; Larin, S. V.; Nazarychev, V. M.; Kenny, J. M.; Lyulin, A. V.; Lyulin, S. V. Toward Predictive Molecular Dynamics Simulations of Asphaltenes in Toluene and Heptane. *ACS Omega* **2019**, *4*, 20005–20014.

(45) Jiménez-Serratos, G.; Totton, T. S.; Jackson, G.; Müller, E. A. Aggregation Behavior of Model Asphaltenes Revealed from Large-Scale Coarse-Grained Molecular Simulations. *Journal of Physical Chemistry B* **2019**, *123*, 2380–2396.

(46) Peryea, T.; Katzel, D.; Zhao, T.; Southall, N.; Nguyen, D.-T. MOLVEC: Open source library for chemical structure recognition. ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY. 2019.

(47) Rajan, K.; Brinkhaus, H. O.; Zielesny, A.; Steinbeck, C. A review of optical chemical structure recognition tools. *Journal of Cheminformatics* **2020**, *12*, 1–13.

(48) Weininger, D. SMILES, a Chemical Language and Information System: 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(49) Landrum, G. A. RDKit: Open-source cheminformatics. 2006.

(50) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10*, 4.

(51) Jolliffe, I. T. *Principal Component Analysis*, springer-verlag ed.; 1986.

(52) Tipping, M. E.; Bishop, C. M. Mixtures of Probabilistic Principal Component Analysers Michael. *Neural Computation* **1999**, *11*, 443–482.

(53) Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(54) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**,

(55) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(56) Venkataraman, P.; Zygourakis, K.; Chapman, W. G.; Wellington, S. L.; Shammai, M. Molecular Insights into Glass Transition in Condensed Core Asphaltenes. *Energy and Fuels* **2017**, *31*, 1182–1192.

(57) Wang, W. *et al.* Nanoaggregates of Diverse Asphaltenes by Mass Spectrometry and Molecular Dynamics. *Energy and Fuels* **2017**, *31*, 9140–9151.

(58) Lyulin, S. V.; Glova, A. D.; Falkovich, S. G.; Ivanov, V. A.; Nazarychev, V. M.; Lyulin, A. V.; Larin, S. V.; Antonov, S. V.; Ganan, P.; Kenny, J. M. Computer Simulation of Asphaltenes. *Petroleum Chemistry* **2018**, *58*, 983–1004.

(59) Sousa Da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012**, *5*, 1–8.

(60) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling* **2006**, *25*, 247–260.

(61) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

(62) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 435–447.

(63) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(64) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19 – 25.

(65) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **2007**, *126*, 014101.

(66) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **1984**, *81*, 3684–3690.

(67) Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* **1984**, *52*, 255–268.

(68) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(69) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(70) Nosé, S.; Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics* **1983**, *50*, 1055–1076.

(71) Hockney, R. W.; Goel, S. P.; Eastwood, J. W. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics* **1974**, *14*, 148–158.

(72) Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **2008**, *4*, 116–122.

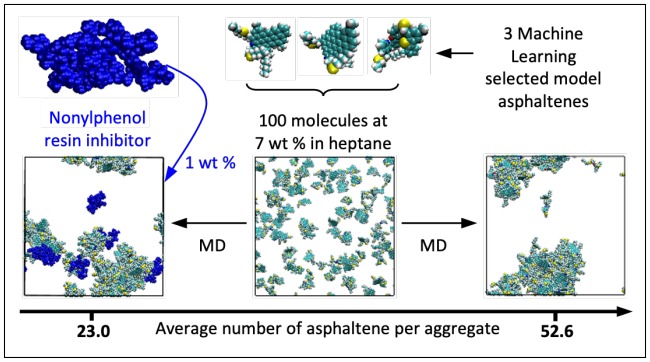
(73) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(74) Nagarajan, R.; Ruckenstein, E. In *Equations of State for Fluids and Fluid Mixtures*; Sengers, J. V., Kayser, R. F., Peters, C. J., White, H. J., Eds.; Experimental Thermodynamics; Elsevier, 2000; Vol. 5; Chapter 15, pp 589–749.

(75) Fixman, M. Radius of gyration of polymer chains. *The Journal of Chemical Physics* **1962**, *36*, 306–310.

(76) Theodorou, D. N.; Suter, U. W. Shape of Unperturbed Linear Polymers: Polypropylene. *Macromolecules* **1985**, *18*, 1206–1214.

TOC Graphic


3 Machine
Learning
selected model
asphaltenes
Nonylphenol
resin inhibitor
1 wt %
100 molecules at
7 wt % in heptane
MD
MD
23.0
Average number of asphaltene per aggregate
52.6