Quantitative Molecular Representation of asphaltenes and Molecular Dynamics simulation of their aggregation

Edo S. Boek, Dmitry S. Yakovlev,*[&] and Thomas F. Headen*[+]*

*Schlumberger Cambridge Research, High Cross, Madingley Road, Cambridge CB3 0EL, UK

[&] DataArt, St.Petersburg, Russia

[+]Dept. Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

AUTHOR EMAIL ADDRESS: e.boek@imperial.ac.uk

RECEIVED DATE

TITLE RUNNING HEAD: QMR representation of asphaltenes and MD simulation of aggregation.

CORRESPONDING AUTHOR FOOTNOTE: EBoek@slb.com; Tel +44 1223 325222

ABSTRACT

We have developed a computer algorithm to generate Quantitative Molecular Representations (QMR) of asphaltenes based on experimental data. First, we generate molecular representations using a Monte Carlo method. For this purpose, we use an extensive set of aromatic and aliphatic building blocks, which are sampled randomly from the corresponding distribution and then linked together using a connection algorithm. The building blocks can be taken from a pre-defined inventory or generated during run-time. Manually pre-fabricated blocks ensure model flexibility while automatically generated blocks allow us to build large aromatic sheets. We allow for both archipelago and peri-condensed structures to be generated. Then, we use a non-linear optimisation procedure to select a small subset of molecules that gives the best match with experimental data. These experimental data consist of Molecular Weight (MW), elemental analysis and NMR spectroscopy, including both ^{1}H and ^{13}C data. First, we validate the method by testing a number of single model compounds. Then we use a real asphaltene data set available in the literature. Different values of the MW were used as input parameter. We tested two specific values of the MW in detail, representing the peri-condensed and archipelago structure respectively: MW= 750 and MW = 4190. For both MWs, we generated 10 sets of 5000 samples each.The samples were then optimized with respect to the experimental objective function. Then we calculate the value of the objective function as an average over all the simulation runs. It turns out that the value of the objective function is significantly smaller for MW=750 than for MW=4190. This means that the lower Molecular Weight of 750 provides the best match with the experimental data. As an example, one of the optimised QMR asphaltene structures generated was then used as input in Molecular Dynamics (MD) simulations to study the formation of nano-aggregates.

KEYWORDS Asphaltene, Quantitative Molecular Representation, QMR, Molecular Dynamics simulation.

1. Introduction

The chemical complexity and diversity of asphaltenes make the analysis and modeling of their chemical structures very challenging. Quantitative Molecular Representations (QMRs) based on analytical data can be used to provide a visual representation of the asphaltenes. They can also be used to generate highly versatile kinetic and reactivity models[1] and to calculate properties such as density, boiling point[1] and solubility[2]. Molecular Dynamics (MD) can be used along with these representations to study how asphaltenes associate and interact with solvents[3] and potential inhibitors. The "average molecule" approach to interpret petroleum analytical data began in the early 1960s. This approach uses correlations that rely on ^{1}H NMR spectroscopy data to calculate average properties, including aromaticity and degree of substitution.[4] When many different molecules represent a petroleum fraction, their process chemistry and product properties can be modeled. More detailed representations have been created using Monte Carlo methods based on data from ^{1}H NMR spectroscopy, elemental analysis, and Vapour Pressure Osmometry (VPO) for various asphaltenes and residues.[1,5] The number of molecules contained in each of these representations ranges from 12 to 10 000 molecules. These methods have been extended to use structural data from quantitative ^{13}C NMR spectroscopy by Sheremata *et al.*[6] The goal of this work was to generate a quantitative molecular representation of asphaltenes that is consistent with data from elemental analysis, ^{1}H and ^{13}C NMR spectroscopy, and VPO. The Sheremata[6] method of construction followed the Monte Carlo approach of Klein and co-workers[1,5] but incorporated ^{13}C NMR data, a distribution of aromatic cluster size following the archipelago framework and both alkyl and thioether bridges. Note that the Sheremata model has been designed to generate archipelago models, due to the limited possibilities to form bridges. Here, we have generalised the Sheremata model and created a more generic model by allowing not only alkyl and thioether bridges, but also arbitrary bridges, within chemical reason. In our model, we can put what we need into the aliphatic building blocks. Our method includes a further extension of the Sheremata method, in the sense that we allow for 3-

dimensional in addition to 2-dimensional asphaltene structures. Particularly in the case of larger structures, it is important to allow for the third dimension because of conformational entropy reasons.

Next, we note the ongoing controversy regarding the molecular weight (MW) of asphaltenes. We refer to a comprehensive recent paper[7] and references therein, framing the discussion between supporters of the archipelago (large MW) and pericondensed (small MW) model. *E.g.* high resolution mass spectroscopy[8] and Fluorescence Depolarisation measurements[9] support the pericondensed model (MW=500-1000), whereas Vapour Pressure Osmometry (VPO) measurements seem to support the archipelago model (MW=4000).[6] In our model, both archipelago and pericondensed model are described by the same chemical logic diagram, so we are in a position to sample both with appropriate selection of sampling parameters and building blocks. We will compare both MW assumptions in our model and we will see that the pericondensed model gives significantly better comparison with the experimental information available. In Appendix A, we present the aromatic building blocks used in our model. In Appendix B, a list of the different atom types, as used in reference 6, is given.

2. Methods

2.1 Model description – Overview

The model is based on the approach described [1,5] and further developed by Sheremata *et al.* [6] The first two papers describe a stochastic approach to converting a set of analytical data into molecular representations of petroleum heavy ends. Molecular structures are generated using a Monte Carlo technique according to a chemical logic diagram which is a way to describe the chemical structure of asphaltenes. At present, there are two widely adopted models for asphaltene structures: pericondensed and archipelago. The pericondensed model is based on a heavy polyaromatic core with attached aliphatic chains. The archipelago model assumes that asphaltenes are formed from small aromatic groups linked by aliphatic chains. The model presented here allows to generate both archipelago and pericondensed structures within the same approach using different sets of input parameters.

2.2 Chemical logic diagram

According to the archipelago model, asphaltene structures are composed of relatively small unit sheets, each built from 8-10 aromatic rings and linked via alkyl chains. In accordance with this approach each asphaltene molecule is described in terms of the following attributes:

1. number of unit sheets
2. number of aromatic rings per unit sheet
3. number of naphthenic rings per unit sheet
4. number of alkyl chains attached to each unit sheet
5. length of each alkyl chain

The scheme in Figure 1 outlines the building blocks for constructing asphaltene molecules. Each attribute is associated with a probability distribution function (hereafter PDF). The PDF can be defined either as a set of tabulated values or in analytical form. It has been shown[1] that the gamma distribution can be successfully used to approximate many kinds of distributions, so we will use the analytical form of the gamma distribution. The gamma distribution function is defined over the $(0, +\infty)$ range, while the attributes are defined over a finite range, and are integers. For this reason, one has to elaborate a technique to project the gamma distribution to a finite region and derive the corresponding Cumulative Distribution Function (hereafter CDF). Because the CDF for the gamma distribution can not be expressed in analytical form, a spline approximation technique is employed. The corresponding algorithms are described in detail below.

Each attribute requires five parameters to define the CDF: minimum (**min**) and maximum **max**) values, average value (**average**), standard deviation (**stddev**) and spline step size **step**. The last parameter defines how accurately the CDF is represented. The smaller the step, the more accurate the representation is. For most of the attributes, a step size of 0.1 is recommended.

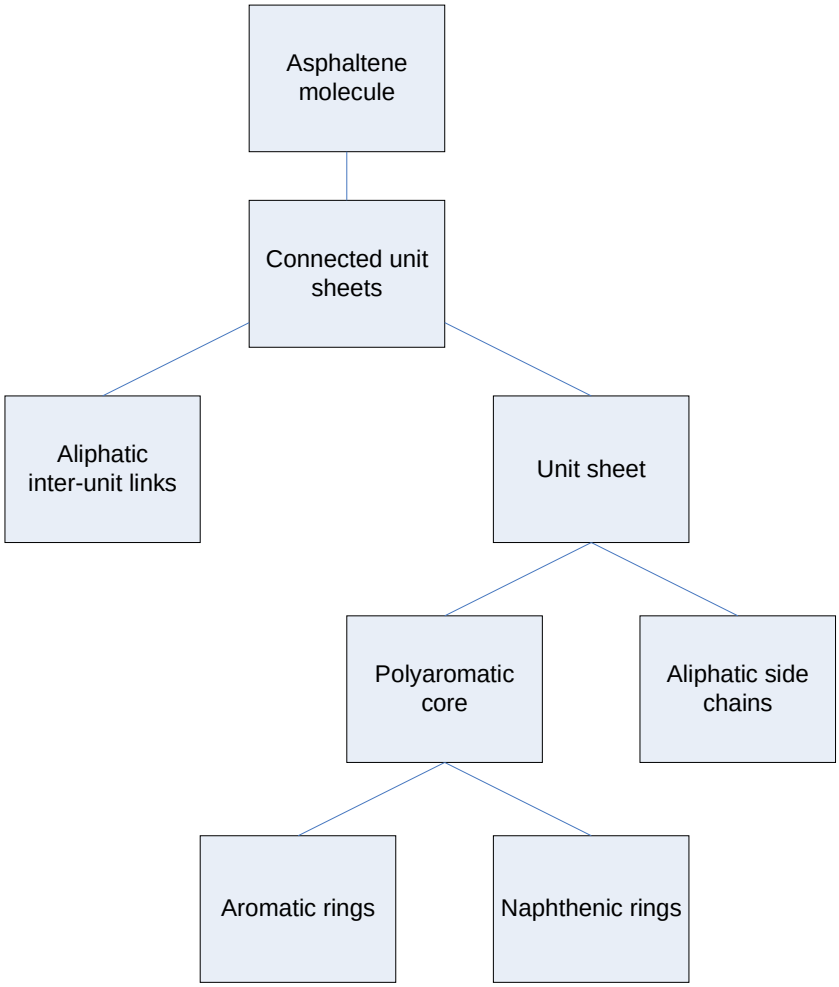


Figure 1. Archipelago chemical logic diagram

2.3 Asphaltene molecule generation

Asphaltene molecule prototypes are generated in accordance with the chemical logic diagram. Prior to a detailed description of the entire process of sampling molecular structures, we are going to explain the concept of **building blocks**.

Building blocks are relatively simple molecular structures which are then assembled into asphaltenes. There are two types of building blocks: **aromatic sheets** and **aliphatic chains**. Aromatic sheets have been extended with naphthenic rings, serving as aromatic core of a single unit. Aliphatic chains are attached to the aromatic core, thus forming a single unit. Both types of building blocks (aromatic sheets and aliphatic chains) have a **Size** attribute. In case of an aromatic sheet, the Size attribute stands for the number of benzene rings in the sheet while in case of aliphatic chains it denotes the length of the chain (note that this is not always the same as the number of carbon atoms).

The aromatic building blocks currently available are listed in Appendix A. The aliphatic building blocks include the *n*-alkanes from 1 to 50 carbons, a number of branched alkanes and chains containing heteroatoms. The additional aliphatics are listed in Appendix A.

In order to successfully simulate archipelago molecules, the model was extended with an algorithm generating large polyaromatic sheets with inclusion of nitrogen atoms. First, the algorithm builds a set of connected aromatic rings. Then some of the carbon atoms are converted to nitrogen. The number of nitrogen atoms in the sheet is randomly sampled according to a distribution function.These structures generated are used along with predefined aromatic building blocks to build asphaltene molecules. The demand for such an algorithm originates from the fact that with an increasing number of aromatic rings, the number of possible isomeric structures grows dramatically, thus making it difficult to pre-define each of them.

It should be noted that neither the heteroatoms (N,S,O,V,..) nor C and H, are generated / sampled according to the gamma distribution. Instead the building blocks are sampled. Note that there are two possible approaches to construct asphaltene molecules:

1. select a subset of building blocks and then connect them
2. generate a set of atoms according to their fractions and then connect the atoms.

The most significant drawback of the second approach is that the variety of possible structures is limited (actually hardcoded into the source code). If one wanted to include a new chemical structure, it is necessary to modify the source code (ensuring that the changes are consistent with the existing algorithms) and recompile the application. In this case the end user will not be able to add new kinds of compounds. Sheremata[6] uses the second approach, while we use the first one for flexibility. However, this means more work for the subsequent optimization procedure. We attempted to improve the building block selection algorithm to make the selection depending on the amount of heteroatoms in the building block instead of random selection. However, this resulted in sampling structures formed from the same subset of the building block. The reason is that the elemental composition of, e.g., aliphatic building blocks, significantly differs from the average over the entire molecule supplied with objective parameters. Thus, some building blocks become too favorable for the selection algorithm. Due to the small amounts of heteroatoms, the distribution function has sharp and narrow peaks at particular building blocks and stays nearly zero for the other ones. Currently, the improved selection algorithm is tuned to assign little weight to building blocks due to hereroatoms.

2.4 Optimisation

The goal of this work is to represent an asphaltene fraction with a set containing a minimum number of molecules that is consistent with the experimental data. With the procedure, a relatively large number of molecules are generated. Subsequently, nonlinear optimization is used to select the best set of molecules from this population. Sheremata [6] employed an *L1*, absolute value, objective function for the optimization:

$$F = \sum_{n=1}^{7} \left| \frac{\mu'_n - \mu_n}{\sigma_n} \right| + \frac{1}{11} \sum_{n=8}^{18} \left| \frac{\mu'_n - \mu_n}{\sigma_n} \right| + J \qquad (0)$$

The parameters used in this equation will be described below. We found that the properties of the *L1* function are not favorable for non-linear optimization. First, the gradient of the function does not depend on the proximity to the minimum, thus causing the algorithm to pass the point and leading to problems with convergence. Second, the gradient at the point where the objective function has its minimal value is not a continuous function anymore.

For the reasons mentioned above we chose an *L2* form of the penalty (or objective) function *F*, defined as follows:

$$F = \sum_{n=1}^{7} \left(\frac{\mu'_n - \mu_n}{\sigma_n} \right)^2 + \frac{1}{11} \sum_{n=8}^{18} \left(\frac{\mu'_n - \mu_n}{\sigma_n} \right)^2 + J \qquad (1)$$

where

$$\mu'_n = \sum_{i=1}^{M} x_i \mu'_{n,i} \qquad (2)$$

Here x_i is the mole fraction of each molecule, $\mu'_{n,i}$ is the value of the property *n* from the molecule *i* referring to the mean calculated value of property *n* from the molecular representation. μ_n is the experimentally determined value of property *n*, σ_n is the calculated input error of property *n*. *J* is a

scaling factor, and a function of $\mu'_{n,i}$, required to keep the aromaticity within experimental error boundaries.[6] The properties *n* are listed in Table 2. Numbers 1-7 are related to elemental properties, whereas 8-18 describe the NMR parameters. Note that the contribution of the error in the NMR parameters to the objective function is weighed by a factor of 1/11 (see equation 1), whereas the elemental properties are weighed by a factor of 1. Lowering the importance of the NMR parameters with respect to MW and elemental analysis is reasonable: the elemental compositions have a much smaller experimental error and are more reliable, while the NMR parameters strongly depend on the interpretation of the peaks in the NMR spectra.

A sequential optimization scheme was used to select molecules from the Monte Carlo set. First, the objective function was evaluated for each of the molecules in the population. The molecule that gave the lowest value for the objective function was then selected. In the next step, all combinations between the selected molecule and all of the other molecules in the population were examined. For each pair of molecules, nonlinear optimisation was used to optimize the mole fractions for each molecule against the experimental data (see Equation 1). The pair with the lowest associated value of the objective function was then selected. The algorithm continued adding molecules into the representation until the objective function reached a constant value with a tolerance of 1% or maximum allowed size of the set is reached. The sequential addition of "well fitting" molecules to the representative set is at the heart of the method, and important to making CPU times required for the optimization process manageable. For a discussion on this topic, we refer to the extensive literature on Bayesian statistics.[10]

3. Results

In order to test the methodology, we will first give results for single model compounds. Subsequently, we consider mixtures. We briefly discuss the problems associated with mixtures of 2 or more model compounds. Finally, we will use the data set used by Sheremata *et al.*[6] to show that our model works well and is an improvement compared with the existing model. [6]

3.1 Single compounds

For the cases of benzene, 2-ethyl-anthracene, tetradecyl-benzene and dibenzo-thiophene, we used the calculated properties of the single model compounds. For all four cases, the correct molecular structures emerged from the simulation. The calculated properties and results are presented in Table 1.

Table 1 Calculated properties of single model compounds

	benzene	2-ethyl-anthracene	tetradecyl-benzene	dibenzo-thiophene
MolWt (g/mol)	78	206	274	184
%C (wt %)	92.3	93.2	87.59	78.26
%H (wt %)	7.7	6.8	12.41	4.35
%N (wt %)	0.0	0.0	0.0	0.0
%S (wt %)	0.0	0.0	0.0	17.39
%O (wt %)	0.0	0.0	0.0	0.0
%V (wt %)	0.0	0.0	0.0	0.0
Aromatic C				
Q1	0.0	6.25	5.0	0.0
Q2	0.0	25.00	0.0	33.33
C1	0.0	37.50	0.0	33.33
C2	100.0	18.75	25.0	33.33
Aliphatic C				
(chain) CH_2	0.0	0.0	45.0	0.0
n-CH2	0.0	0.0	0.0	0.0
o-CH2	0.0	6.25	20.0	0.0
CH	0.0	0.0	0.0	0.0
a-CH3	0.0	0.0	0.0	0.0
b-CH3	0.0	0.0	0.0	0.0
g-CH3	0.0	0.0	5.0	0.0

3.2 Mixtures of model compounds

Note that the current QMR model has *not* been designed to predict the unique structures and relative concentrations of mixtures of small numbers of model compounds. Here we consider numbers of model compounds which are small in comparison with the number of compounds in a realistic asphaltene mixture. This can be explained as follows: The "error" parameters associated with the objective function parameters should reflect the experimental standard deviation. Now imagine that one sets the error for the molecular weight (MW) to be small. As an example we take a binary mixture of 2-ethyl-antracene and tetradecyl-benzene. Obviously, the compounds we would like to obtain, have different MW (in this case 206 and 274). If the error one sets in the MW is too low, then a small deviation in weight will result in a high deviation in the objective function, according to the definition of the penalty or objective function *F*, given in Equation (1) so these compounds are effectively excluded from the list. The "error" parameter is of particular importance when sampling mixtures. For this reason, it is not possible to reproduce an exact mixture of e.g. 2-ethyl-antracene and tetradecyl-benzene. For example, a mixture of 2-tetradecyl-antracene and ethyl-benzene has nearly the same average properties (most important difference is in the weight). There is a chance to produce a single compound which is closer to the objective parameters than the mixture of these two compounds.

3.3 Sheremata results

In Table 2, we present the optimised asphaltene representations consisting of molecules from starting populations of 3000 and 4000 molecules. The calculated aromaticity and elemental compositions of the molecular representation matched the experimental values almost to within the range of experimental error. The calculated molecular weight was slightly smaller than the MW found from the VPO experiments. We note that an error of at least 15% is normally associated with these VPO measurements [3], so that our calculated results are within the experimental error margin. We also note that VPO measurements are carried out under conditions when non-covalent association is likely.[13]

Table 2 Comparison between experimental and calculated properties of the molecular representations

Property	Objective Function Index	EXPERIMENT			Calculated Properties			
					5 of 3000 molecules		6 of 4000 molecules	
		Property	Error	%Error	Property	%Error	Property	%Error
Mol Weight (g/mol)	1	4190	630	15.04	3802.11	9.26	3541.79	15.47
% C (wt %)	2	81.4	1.6	1.97	81.09	0.38	81.21	0.23
% H (wt %)	3	8.45	0.17	2.01	9.04	6.98	8.97	6.15
% N (wt %)	4	1.17	0.02	1.71	1.16	0.85	1.16	0.85
% S (wt %)	5	7.95	0.16	2.01	7.68	3.40	7.65	3.77
% O (wt %)	6	1.03	0.02	1.94	1.01	1.94	1.01	1.94
V (PPM)	7	877	18	2.05	0.01	14.03	0.01	14.03
Total Aromaticity (% mol C)		50.1	3	5.99	45.59	9.00	46.54	7.11
Q1 (% mol C)	8	10.4	3.6	34.62	12.26	17.88	12.03	15.67
Q2 (% mol C)	9	17.6	3.3	18.75	14.52	17.50	14.98	14.89
C1 (% mol C)	10	10.8	3.3	30.56	12.02	11.30	12.53	16.02
C2 (% mol C)	11	11.3	2.3	20.35	6.78	40.00	7	38.05
Total Aliphatic Content (% mol C)		49.9	3	6.01	54.41	9.04	53.46	7.13
Other aliphatic (% molC)	12	17.7	4.6	25.99	37.43	111.47	36.49	106.16
Total a-CH3 (% mol C)	13	6.6	0.4	6.06	0	100.00	0	100.00
b-CH3 (% mol C)	14	1.3	0.2	15.38	0.2	84.62	0.16	87.69
Chain CH2 (% mol C)	15	7.7	0.3	3.90	8.8	14.29	8.94	16.10
Aliphatic CH (% mol C)	16	9.8	2.1	21.43	1.25	87.24	1.2	87.76
Naphthenic CH2 (% mol C)	17	4.5	0.3	6.67	0.64	85.78	0.66	85.33
g-CH3 (% mol C)	18	2.3	0.2	8.70	6.1	165.22	6.01	161.30
Objective function total					97.03		94.94	

Table 3 lists the molecular weights and the elemental compositions of the individual optimized molecules (5 out of 3000). The same data for 6 optimized molecules out of 4000 generated is presented in Table 4. Note that the overall error (objective function total) decreases with the total number of molecules increasing from 3000 to 4000. We note that most of the experimental parameters unrelated to the NMR experiments were modeled to within experimental error. This is due to the fact that the contribution of the error in the NMR parameters to the objective function is weighed by a factor of 1/11 (see equation 1). Lowering the importance of the NMR parameters is reasonable: the elemental compositions have a much smaller experimental error and are more reliable, while the NMR parameters strongly depend on the interpretation of the peaks in the NMR spectra.

Table 3: MW and elemental compostion of optimised molecules (5 out of 3000)

Molecule	mol%	Mol weight, g/mol	Elemental composition, wt%					
			C	H	N	S	O	V
1	20.2	911	84.3	8.89	1.54	3.51	1.76	0
2	9.554	4060	80.1	8.62	1.03	10.25	0	0
3	4.045	997	79.44	9.93	4.21	6.42	0	0
4	65.63	4821	80.4	9.11	0.87	8.63	1	0
5	0.572	4525	75.85	8.04	1.55	12.73	0.71	1.13

Table 4: MW and elemental composition of optimised molecules (6 out of 4000).

Molecule	mol%	Mol weight, g/mol	Elemental composition, wt%					
			C	H	N	S	O	V
1	17.949	911	84.3	8.89	1.54	3.51	1.76	0
2	7.554	997	79.44	9.93	4.21	6.42	0	0
3	53.152	4821	80.4	9.11	0.87	8.63	1	0
4	0.605	4525	75.85	8.04	1.55	12.73	0.71	1.13
5	13.144	4135	81.26	7.96	0.34	10.06	0.39	0
6	7.596	2233	81.68	9.09	0.63	7.17	1.43	0

The optimized molecules are presented in Table 3 and Table 4. Note that there is significant overlap between the molecular representations between both cases. This lends credibility to the optimisation process.

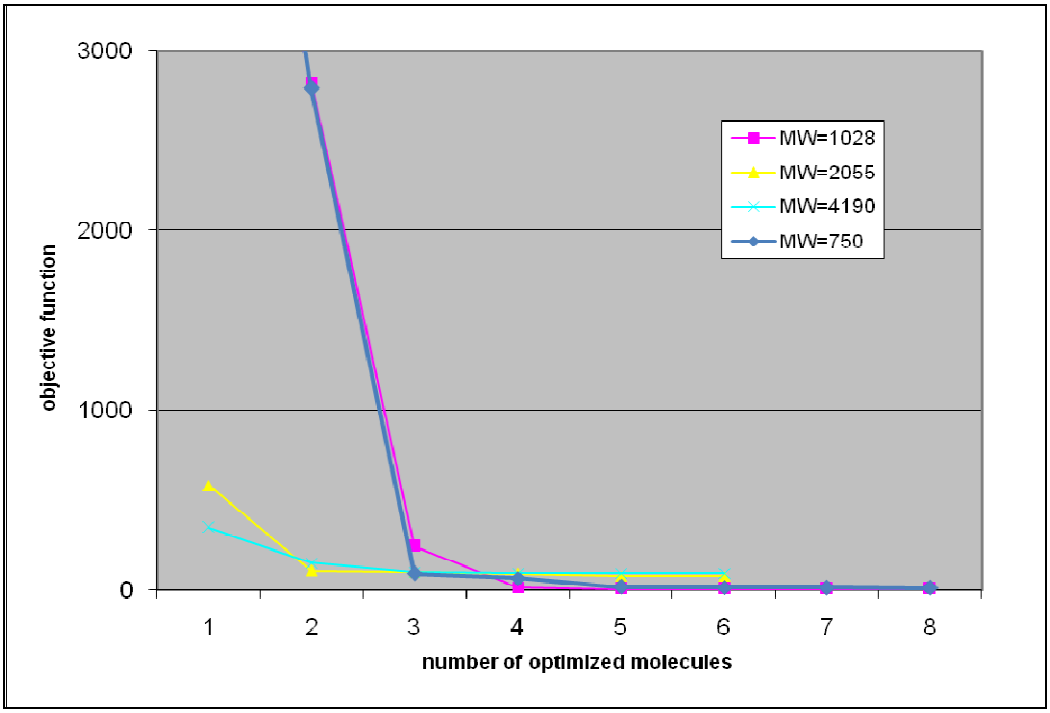
To examine the ability of molecular representations to fit the experimental data, independent populations of 3000 and 4000 molecules were created. Each population was optimized using the

sequential optimization approach. For each independent population, the objective function decreased as the number of molecules in the representation increased from one to five. This is shown in Table 5 and Figure 2. This means that a good quality representation requires four to six asphaltene molecules depending on the complexity of the structure. Please note that the difference between the values of the objective function for MW=750 and MW=1028 is not statistically significant.

Table 5: Value of objective function for different number of optimised molecules

Number of optimized molecules		Objective function		
	MW=750	MW=1028	MW=2055	MW=4190
1	6275.7	5349	582.48	347.24
2	2789.9	2826.95	109.24	150.19
3	90.88	245.35	101.5	100.8
4	66.84	14.81	90.71	95.63
5	17.14	6.75	77.72	95.18
6	14.73	5.59	77.7	94.94
7	13.77	5.59		
8	12.42	5.59		

Figure 2: Value of the objective function as a function of increasing number of optimised molecules for different Molecular Weights.



Molecular Weight

As mentioned above, there is a serious controversy in the literature regarding the Molecular Weight (MW) of aspahltenes. The following experimental observations have been made in the recent literature:

- Mass Spectrometric and Molecular diffusion methods show that the MW of asphaltenes is approximately 750 $\pm$ 200 amu. [11]

- Molecular Orbital Calculations combined with Optical Spectroscopy shows that Petroleum asphaltenes consist of Polycyclic Aromatic Hydrocarbons (PAHs) mostly with 6-8 Fused Aromatic Rings (FAR). [12]

In order to investigate this controversy, we examined the sensitivity of the molecular representations to the molecular weight. Two additional different target molecular weights were selected: half the molecular weight as determined by VPO (2055 g/mol) and one-quarter (1028 g/mol). The data in **Table 6** shows a comparison between the calculated properties at different target molecular weights and the experimental data. From **Table 5**, it can be observed that our method gives significantly smaller values for the objective function for MWs of 1028 and 750, when the number of optimised molecules is 4 or larger. This means that we obtain a more reliable representation for lower values of the MW. Note that this result is different from Sheremata[6], and is due to the fact that our linkage algorithm allows for both archipelago and pericondensed structures to be generated. In fact, the optimal values for the Mol. Weight of 1028 and 750 g/mol are supportive of the values suggested by Mullins[13] et al., associated with pericondensed structures. The details of the optimized molecules from **Table 6** for MW=1028 are shown in **Table 7**. The total number of sampled compounds is 1000; only 5 molecules were left after optimization. Note that we did not consider further reductions in the target MW, such as MW = 514. It is not clear whether the Hirsch model[14] would still work for such low molecular weights. For such substances one has to set the sampling parameters explicitly. Also, we need to consider how to take care

of vanadium for such low MW. Currently, the only building block which contains vanadium is porphyrine. Therefore porphyrine is required to reduce the objective function deviation due to the presence of V. However, porphyrine is a relatively heavy compound, especially when long aliphatic chains are attached. This leads to a larger deviation in the molecular weight. To solve the problem, we will have to either introduce smaller building blocks containing vanadium or set the V fraction to zero.

Table 6: Calculated properties of Asphaltene representations of different target molecular weights

Property	Experimental	Target molecular weight, g/mol		
		4190	2055	1028
Calculated MW (g/mol)	4190	3541.79	1733.66	913.67
C (wt %)	81.4	81.21	80.79	81.25
H (wt %)	8.45	8.97	9.12	8.5
N (wt %)	1.17	1.16	1.17	1.17
S (wt %)	7.95	7.65	7.9	8.05
O (wt %)	1.03	1.01	1.02	1.03
Aromacity (% mol C)	50.1	46.54	43.86	50.23

Table 7: Calculated Elemental composition of the molecules in the asphaltene representation: MW=1028

Molecule	mol%	Mol weight, g/mol	Elemental composition, wt%					
			C	H	N	S	O	V
1	30.54	1082	79.85	8.50	1.29	8.87	1.48	0
2	22.95	1033	82.48	8.42	1.36	6.20	1.55	0
3	20.48	1252	86.26	8.95	2.24	2.56	0	0
4	13.39	240	80.00	6.67	0	13.33	0	0
5	12.64	1206	80.60	8.79	0	10.61	0	0

Table 8: Calculated Elemental composition of the molecules in the asphaltene representation: MW=750. The average MW for this set is 549

Molecule	mol%	Mol weight, g/mol	Elemental composition, wt%					
			C	H	N	S	O	V
1	45.29	597	84.42	7.87	2.35	5.36	0	0
2	26.32	520	80.77	10	0	6.15	3.08	0
3	22.78	436	77.06	8.26	0	14.68	0	0
4	5.61	749	83.31	6.28	1.87	8.54	0	0

In order to estimate the ability of the model to reproduce results, we did a set of 10 simulations with the same parameters, assuming MW=750 and MW=4190. First, we generated 5000 molecules for each simulation, then removed the duplicates, and optimized the results using sequential optimization, to obtain a maximum of 6 optimised structures for each run. We calculated the value of the objective function and molecular weight for each run. Finally the average values for both parameters were calculated as an average over the 10 simulations. The results are presented in Tables 9 and 10. It turns out that the value of the objective function is 29 (+/-8) for MW=750 and 59 (+/- 11) for MW=4190. This means that the lower MW of 750 provides the better match with the experimental data. Please note that for MW=4190, the average QMR MW is significantly lower than the experimental value. This is due to the fact that the NMR parameters force the molecules generated to be small. Please note that the simulations in Tables 6 and 9 were done for different sampling parameters and MW errors. Therefore the resulting values are also different.

Table 9: Reproducibility results for MW=750 from 10 simulations: # mol = number of molecules generated, #mol no dups= number of molecules after duplicates have been removed, # mol opt = **number of molecules after optimization, Obj dev = value of objective function. At the bottom, the average over the 10 runs and standard deviation is given for both the deviation from objective function and the Molecular Weight.**

	#mol	#mol no dups	#mol opt	Obj dev	Average MW
1	5000	4996	4	39	915
2	5000	4998	5	27	745
3	5000	4998	4	32	549
4	5000	4997	4	26	1003
5	5000	5000	5	29	687
6	5000	4997	6	36	875
7	5000	4998	5	37	650
8	5000	4999	5	31	859
9	5000	4994	4	18	595
10	5000	5000	5	14	675
	AVG			**29 (+/- 8)**	**755 (+/- 150)**

Table 10: Reproducibility results for MW=4190 from 10 simulations. See Table 9 for details.

	#mol	#mol no dups	#mol opt	Obj dev	Average MW
1	5000	5000	5	44	1664
2	5000	5000	4	57	1766
3	5000	5000	5	69	1644
4	5000	5000	4	67	2048
5	5000	4999	4	56	2188
6	5000	5000	5	63	1549
7	5000	5000	4	63	1936
8	5000	5000	4	36	1925
9	5000	5000	4	70	1874
10	5000	5000	5	63	1571
	AVG.			**59 (+/- 11)**	**1817 (+/- 213)**

We collected the final optimal structures with the heighest weight fraction. These are shown in Figure 3 - Figure 10. Overall, the model demonstrated a good reproducibility with respect to deviation function, average molecular weight and similarity of the resulting asphaltene structures.

Figure 3. Optimized structure run 1, MW=727, 45 wt %.

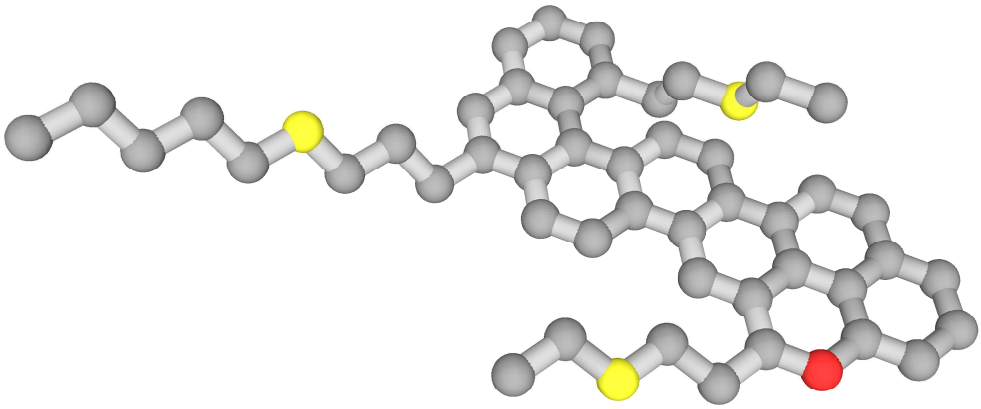


Figure 4. Optimized structure run 3 MW=597, 45 %. This molecule will be used in the subsequent MD simulations.

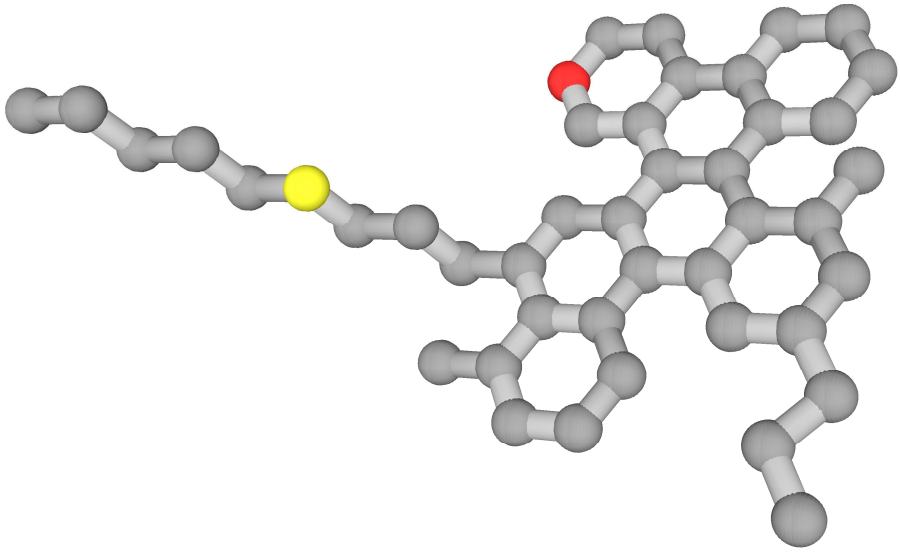


Figure 5. Optimized structure run 4, MW=1443, 41 wt%.

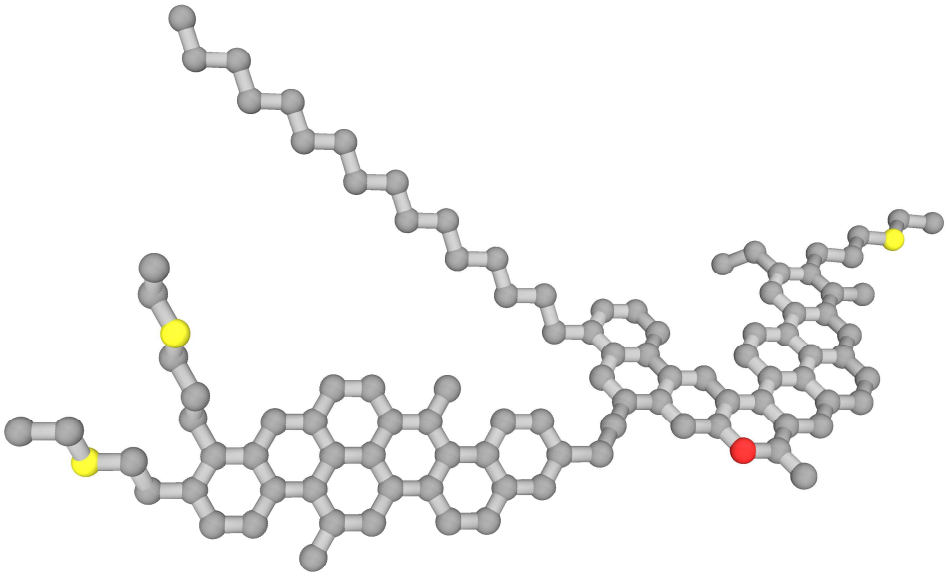


Figure 6. Optimized structure run 5 MW=755, 56 wt %.

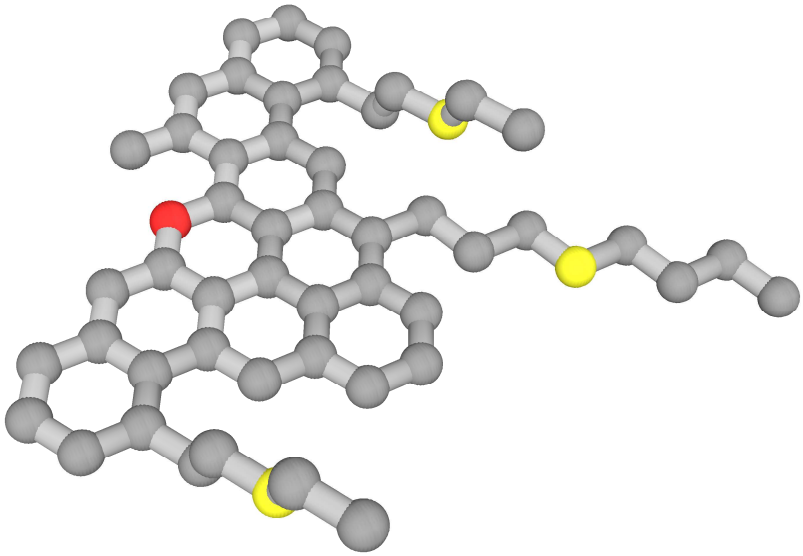


Figure 7. Optimized structure run 7 MW=747, 61 wt %.

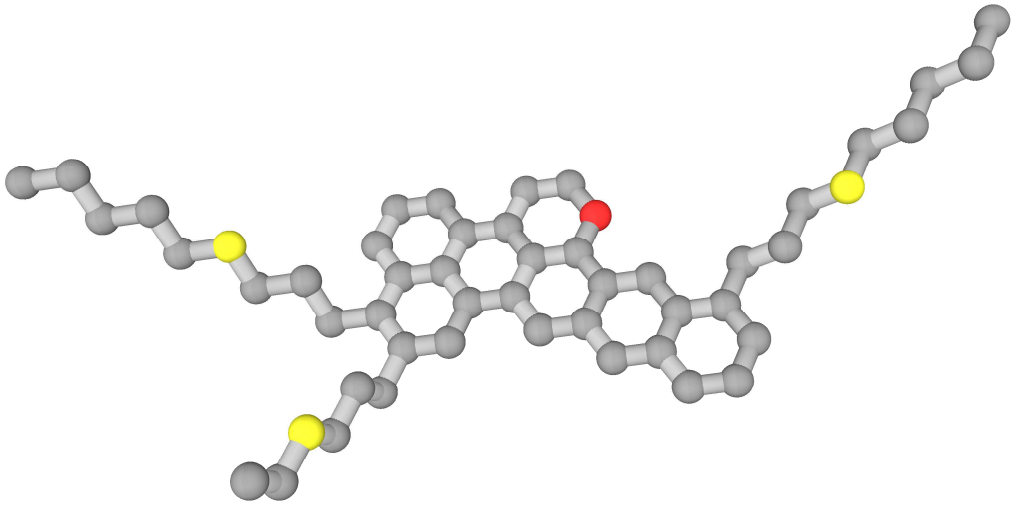


Figure 8. Optimized structure run 8 –MW=755, 34 wt %

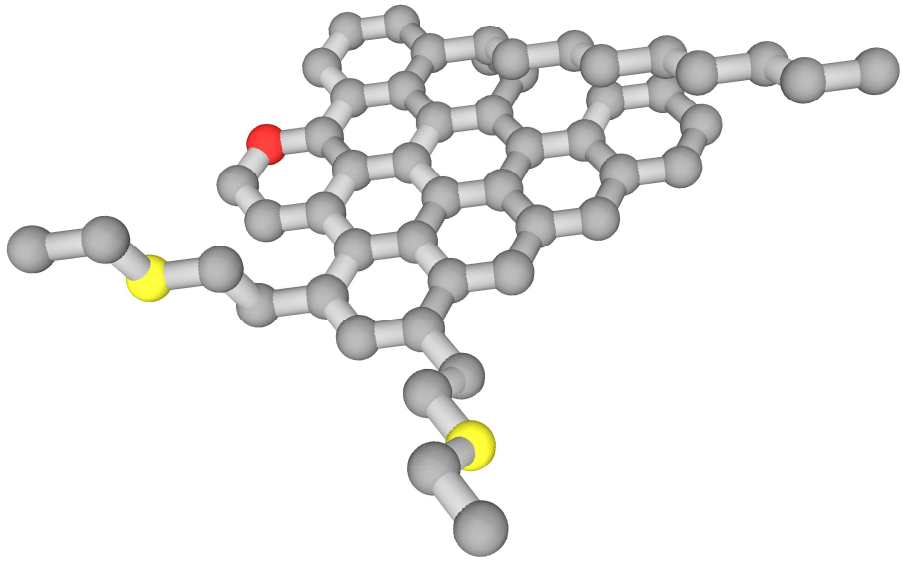


Figure 9. Optimized structure run 9, MW=567, 33 wt%.

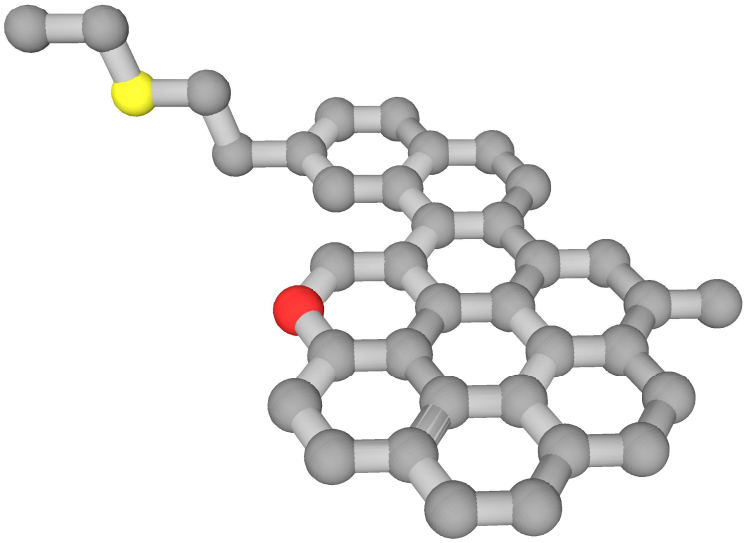
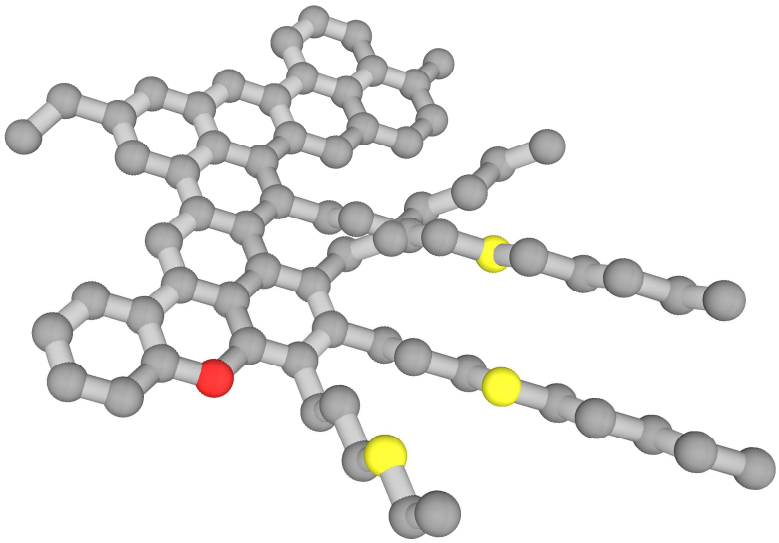


Figure 10. Optimized structure run 10, MW=1027, 35 wt %



Correlation between objective parameters

One of the problems which can be investigated with the model application is the relations between the objective parameters (correlation) and feasibility of experimental data. The idea of the exercise is to first generate a broad set of asphaltene structures, providing a good representation of the most generic set. Then we calculate the average values, correlation coefficients and limits of the objective parameters and compare them with the experiment. This will help to evaluate the reliability and reveal systematic errors and inconsistencies of the experimental data. A representative set of asphaltenes is generated by setting a very flat distribution (high STD value) of the sampling parameters. To ensure the distributions are uniform, the standard deviation was set to twice the average value. A summary of the sampling parameters for the representation is given in Table 11.

Table 11. Sampling parameters for feasibility check

	Min	Max	Avg	Std
Unit sheets per asphaltene molecule	1	20	10	20
Aromatic rings per unit sheet	1	15	8	16
Number of naphthenic rings per unit sheet	0	10	5	10
Alkyl chain length	1	25	12	24
Substitution of aromatic atoms with alkyl chains (%)	0	40	20	40
Substitution of naphthenic atoms with alkyl chains (%)	0	50	25	50

The total number of asphaltene molecules generated was 100000. The results of the data processing are given in Table 12 and Table 13. We found a strong correlation between:

- Molecular weight and number of units (0.96)
- Q2 (see Appx. C) and number of aromatic rings per unit (0.73)
- gamma-CH3 and the number of chains per unit (0.78)
- naphthenic CH2 and number of naphthenic rings per unit (0.78)

This list allows us to state the following:

1. The asphaltene MW is mostly controlled by the number of units the molecule is composed of. At the same time, the size of the units is about the same.

2. The number of aromatic rings per unit is driven by the amount of bridgehead aromatic quaternary carbon (Q2), which is directly available from NMR experiments

Table 13 contains averages, standard deviation, min/max values and expected ranges for all objective and sampling parameters. With some caution (one has to keep in mind that adding essential fractions of heteroatoms will shift/scale most of the values) this table may be used as a reference to validate experimental data.

Having obtained representative structures for the asphaltene experimental data, it is now possible to use these structures in Molecular Dynamics and Monte Carlo simulations to measure physical properties from ensemble averages.

In the following section, we will carry out Molecular Dynamics (MD) simulations for the optimized QMR structures generated.

Table 12. Correlation coefficients

	Units	Links	Aromatic rings	Naphthenic rings	Chains	Naphthenic rings/Unit	Naphthenic rings / Aromatic rings	Aromatic rings / Unit	Chains / Units
Mol Weight	0.96	0.96	0.97	0.86	0.88	0.09	-0.08	0.13	0.14
w%C	-0.09	-0.09	0.01	-0.09	-0.20	-0.05	-0.21	0.42	-0.28
w%H	0.19	0.19	0.08	0.25	0.39	0.26	0.39	-0.40	0.53
w%N	-0.08	-0.08	-0.12	-0.09	-0.11	-0.10	0.01	-0.25	-0.14
w%S	0.05	0.05	0.02	0.03	0.10	-0.04	-0.01	-0.11	0.14
w%O	-0.04	-0.04	-0.06	-0.05	-0.06	-0.05	-0.01	-0.12	-0.08
w%V	0.00	0.00	-0.02	-0.01	-0.02	-0.02	-0.01	-0.05	-0.04
Ar Tot%	-0.24	-0.24	-0.15	-0.36	-0.41	-0.44	-0.51	0.29	-0.51
Ar Q1%	0.07	0.07	-0.02	0.28	0.02	0.67	0.77	-0.33	-0.03
Ar Q2%	-0.08	-0.08	0.09	-0.16	-0.18	-0.24	-0.54	0.73	-0.25
Ar C1%	-0.20	-0.20	-0.11	-0.36	-0.32	-0.54	-0.59	0.31	-0.38
Ar C2%	-0.22	-0.22	-0.24	-0.32	-0.29	-0.43	-0.28	-0.28	-0.32
Al Tot%	0.24	0.24	0.15	0.36	0.41	0.44	0.51	-0.29	0.51
Al aCH3%	0.05	0.05	0.03	0.03	0.20	-0.01	0.02	-0.06	0.37
Al bCH3%	-0.01	-0.01	-0.02	-0.01	0.14	0.01	0.03	-0.03	0.33
Al gCH3%	0.03	0.03	0.03	0.05	0.40	0.04	0.03	-0.01	0.78
Al CH2%	0.25	0.25	0.19	0.19	0.37	-0.01	-0.02	-0.10	0.41
Naph CH2%	-0.09	-0.09	-0.12	0.18	-0.18	0.67	0.78	-0.20	-0.24
Al CH2o%	0.19	0.19	0.09	0.33	0.31	0.51	0.60	-0.32	0.40
Al CH%	0.19	0.19	0.10	0.39	0.29	0.47	0.49	-0.22	0.33

Table 13. Objective parameters feasibility regions

	MIN	MAX	STD	AVG-STD	AVG	AVG+STD
Units	1	20	5.17	**1.55**	**6.71**	**11.88**
Links	0	19	5.17	**0.55**	**5.71**	**10.88**
Aromatic rings	1	166	29.67	**6.66**	**36.32**	**65.99**
Aromatic rings/Unit	1	15	2.13	**3.28**	**5.41**	**7.55**
Naphthenic rings	0	60	8.50	**0.62**	**9.12**	**17.62**
Naphthenic rings/Unit	0	10	1.03	**0.33**	**1.36**	**2.38**
Naphthenic rings/Aromatic rings	0	3	0.29	**0.00**	**0.29**	**0.58**
Chains	0	57	7.01	**0.08**	**7.09**	**14.10**
Chains/Units	0	17	0.94	**0.12**	**1.06**	**2.00**
Atoms	6	1255	232.05	**49.34**	**281.39**	**513.43**
Mol Weight	78	16984	3118.20	**656.22**	**3774.42**	**6892.63**
w%C	64	96.89	2.62	**87.05**	**89.68**	**92.30**
w%H	3.11	13.47	1.39	**7.04**	**8.43**	**9.82**
w%N	0	14.93	1.17	**0**	**0.45**	**1.62**
w%S	0	21.62	1.54	**0**	**1.20**	**2.75**
w%O	0	9.52	0.48	**0**	**0.13**	**0.60**
w%V	0	13.6	0.61	**0**	**0.11**	**0.73**
Ar Tot%	7.5	100	15.03	**40.76**	**55.79**	**70.83**
Ar Q1%	0	33.33	4.30	**6.56**	**10.86**	**15.16**
Ar Q2%	0	61.54	7.89	**13.69**	**21.57**	**29.46**
Ar C1%	0	60	7.55	**7.26**	**14.81**	**22.35**
Ar C2%	0	100	8.27	**0.29**	**8.56**	**16.82**
Al Tot%	0	92.5	15.03	**29.17**	**44.21**	**59.24**
Al aCH3%	0	25	0.91	**0**	**0.84**	**1.75**
Al bCH3%	0	20	0.64	**0**	**0.36**	**0.99**
Al gCH3%	0	16.67	1.34	**0.44**	**1.78**	**3.12**
Al CH2%	0	64.52	9.70	**5.79**	**15.49**	**25.19**
Naph CH2%	0	33.33	4.08	**0.34**	**4.42**	**8.50**
Al CH2o%	0	54.55	6.38	**13.01**	**19.40**	**25.78**
Al CH%	0	16.67	1.57	**0.35**	**1.92**	**3.49**

4. Molecular Dynamics Simulation of QMR Generated Asphaltene Structures.

In recent years there has been an increasing effort to understand asphaltene aggregation on a molecular level using molecular simulation[15,16,17,18,19,20]. The Quantitative Molecular Representation (QMR) method is a vital tool for generating molecular structures for asphaltene simulation.[6] It is only possible to conduct simulations on a small number of molecular structures due to the large amount of processor time required for each simulation. Therefore it is essential that we represent the full spectrum of asphaltenes by the smallest possible number of representative molecules. As shown in the previous sections, the QMR method provides a good method of achieving this. A full study for a set of 3 molecules generated by QMR has been conducted by Headen et al[22]. Here we provide an example simulation of one molecule. The molecule chosen is asphaltene 2 of simulation run 3, its chemical structure is given in Figure 4. We carry out Molecular Dynamics simulations of 6 asphaltene molecules in toluene and heptanes. From these simulations, we obtain the distance-time relationship over 60 ns between asphaltene pairs in order to reveal the timescale of aggregation. We also obtain the asphaltene-asphaltene g(r) and average angle between the aromatic planes as a function of distance to directly detect the structure of asphaltene aggregation in dimers and trimers – this can indicate the structure of larger asphaltene nanoaggregates. Lastly we use the asphaltene-asphaltene g(r) to calculate the asphaltene-asphaltene potential of mean force (PMF).

4.1 Simulation Methods.

Classical molecular dynamics uses well defined classical intra- and inter-molecular potentials to calculate interatomic forces. The system is allowed to evolve over time by stepwise integration of the equations of motion. It is important that the step-size be smaller than the time period of the fastest motion in the simulation. Using this method it is possible to simulate systems containing ~10000 atoms over a few nanoseconds on ~4 processor nodes in reasonable times. For this study the GROMACS MD code[23] with the OPLS- AA force field parameters[24,25] was used. The OPLS force field has been shown to work well for aromatic liquids in reproducing experimental data[25]. Rigid bonds were used to remove the fastest moving molecular motions, a time step of 2fs was used for all simulations. Periodic boundary conditions with the minimum image convention were used so that a small box of ~10000 atoms can represent the bulk. Long range coulomb intermolecular forces are treated using the Particle-Mesh Ewald (PME)[26] technique which allows the use of Fast-Fourier Transforms (FFT). Before the simulation cell was built, explicit hydrogen atoms where added to the asphaltene structure generated by the QMR algorithm followed by a geometry optimization using ArgusLab[27]. The 3-dimensional structure of the asphaltene is shown in Figure 11. Please note that the structure of the asphaltene molecule is sterically constrained; the aromatic core is curved out of shape due to a methyl group on the aromatic core pointing back towards an aromatic hydrogen. This emphasizes the need for allowing 3-d configurations.

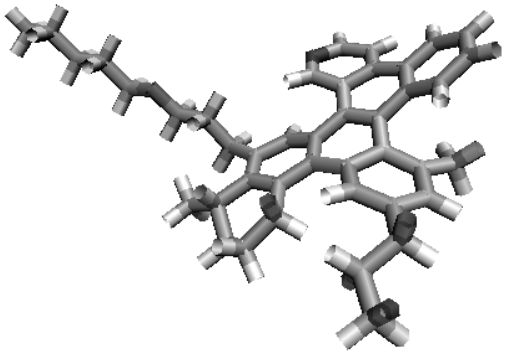


Figure 11: 3-dimensional structure of asphaltene 2 from simulation 3 used from molecular dynamics simulations. Aromatic plane is curved out of shape due to methyl group pointing back towards adjacent aromatic ring.

We carried out two set of simulations of 6 asphaltene molecules in either toluene or heptane at 7wt% over 20ns. Analysis of the trajectory of the full NVT simulation allows the calculation of the asphaltene-asphaltene radial distribution function, g(r). The g(r) is defined as ratio of the local density of atoms/molecules ρ(*r*) at distance *r* from an atom at the origin to the average density of atoms in the bulk, ρ:

$$g(r) = \frac{\rho(r)}{\rho} \tag{3}$$

In this case *r* is the distance between a single atom defined on each asphaltene molecule to represent its centre. The cumulative coordination number can be calculated from the g(r) by integration over spherical shells of thickness *dr*:

$$N(r) = \int_0^r \rho g(r).4\pi r^2 dr \tag{4}$$

In essence this is the number of asphaltene molecules within the sphere radius *r* centered on the original molecules. Another useful quantity that may be calculated from the g(r) is the Potential of Mean Force (PMF). In the low density limit, the pair potential *u(r)* is given by:

$$u(r) = -kT \ln(g(r)) \quad (5)$$

This concept can be extended to dense fluids by defining the potential of mean force *W* as[28]:

$$W = -kT \ln(g(r)) \quad (6)$$

The PMF is equivalent to the Helmholtz free energy (plus a constant). We can therefore calculate the free energy of dimer formation by taking the difference of the potential of mean force at maximum separation and at equilibrium separation (where the PMF is a minimum).

For the simulations, the system was created by periodically arranging the asphaltene molecules in a box of the appropriate size (dependent on how many solvent molecules needed to be added to make 7wt% asphaltene). A periodic arrangement of solvent molecules was then added with any molecules overlapping the asphaltene molecules being removed. The system was initially equilibrated by a 100ps NVT simulation, followed by a 500ps (*NpT*) simulation for the system to reach equilibrium density. Sampling was conducted during the main 60ns (*NVT*)simulation. Here (*NpT*) and (*NVT*) indicate the isothermal-isobaric and constant volume ensembles respectively. The temperature was maintained at 300K using the Nose-Hoover thermostat, the Parrinello-Rahman barostat was used for the NPT equilibration.

4.2 Simulation Results and Discussion

Over the course of the simulation, the atomic positions are recorded every 4 ps. From this the distance between the centre of mass of the asphaltene molecules over the course of the simulation can be obtained. Figure 12 and Figure 13 show the distance between pairs formed between 1 asphaltene and the 5 others in the simulation box over a 60ns simulation, in toluene and heptane respectively. For all distance/time plots a running average every 500 ps is taken to reduce the scatter in the data.

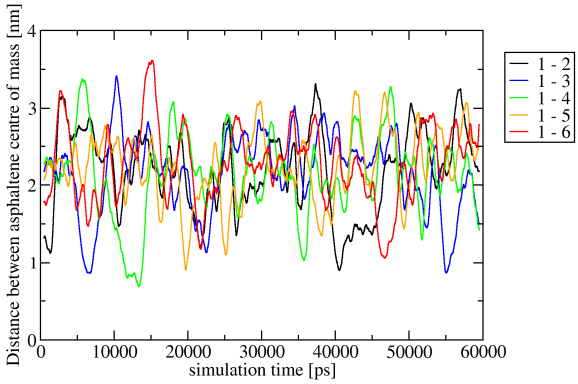


Figure 12: Distance between centre of mass of one asphaltene with 5 others in the simulation box during 60ns NVT simulation in toluene.

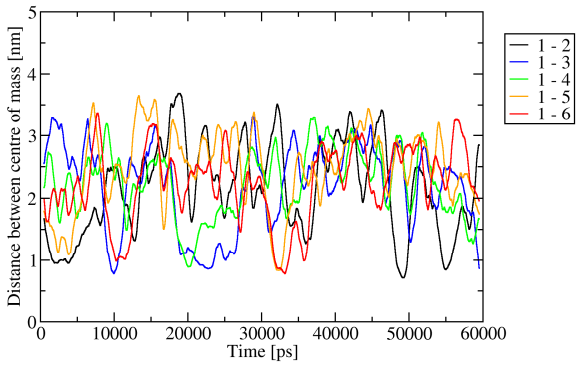


Figure 13: Distance between centre of mass of one asphaltene with 5 others in the simulation box during 60ns NVT simulation in toluene.

The simulations in both toluene and heptane show that asphaltene dimers/trimers form, split apart and re-aggregate with other asphaltenes. There are no dimers living longer than roughly 10ns. In an attempt to quantify the *aggregation time* of asphaltene dimers, we have defined a critical aggregation distance of 1nm and a critical aggregation time of 0.5 ns. When the distance between an asphaltene pair is below this critical aggregation distance for longer than the critical aggregation time, this is identified as an aggregation event. The length of time of each aggregation event is counted and then averaged over all the aggregation events yielding the average aggregation time. **Table 14** gives the average aggregation times and number of aggregation events for simulations in toluene and heptane. Similar average aggregation times are seen in both toluene and heptane. There is a slight increase in the number of aggregation events in heptane indicating slightly enhanced aggregation.

Table 14: Average aggregation time and number of aggregation events for asphaltene pairs during 60ns simulations in toluene and heptane.

	Heptane	*Toluene*
Average aggregation time (ps)	1129	1082 ± 128
Number of aggregation events over 60 ns	25	22

Figure 14 is a snapshot of the simulation in heptane showing the formation an asphaltene cluster at ~32ns. Here *loose* rather than *rigid* stacking of the aromatic cores is observed, *i.e.* over the course of the trajectory, a dimer pair will change conformations around each other continually. Although parallel stacking of the asphaltene aromatic cores is observed, this is not the only mode of aggregation: alkyl-alkyl, alkyl-aromatic and T-shaped aggregates are also observed.

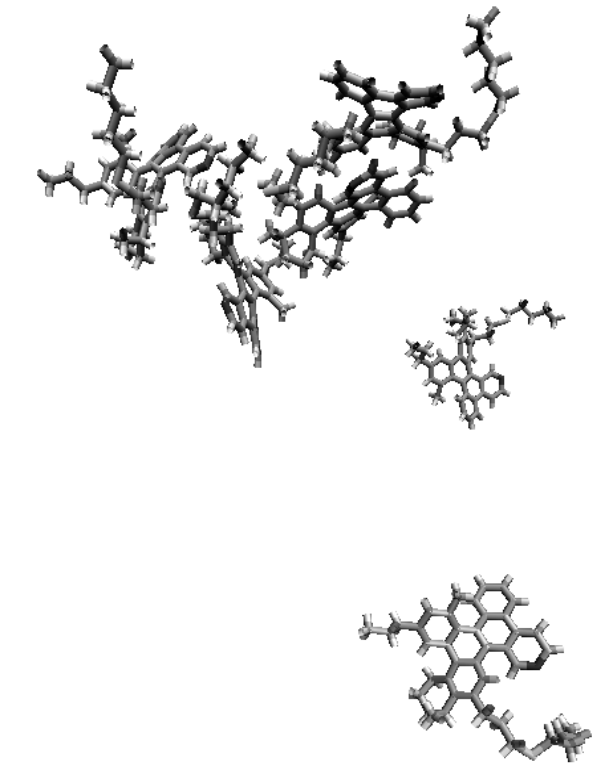


Figure 14: Snapshot of MD simulation of 6 asphaltene molecules in heptane, showing formation of an asphaltene cluster at approximately 32ns. Heptane molecules are not shown for clarity.

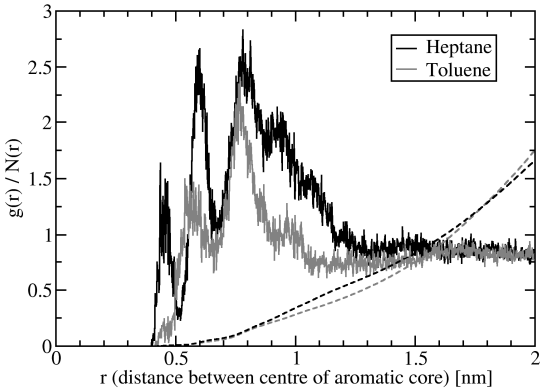


Figure 15: Asphaltene-asphaltene g(r) (solid line) and N(r) (dashed line) for simulations in toluene (grey) and heptane (black)

The asphaltene-asphaltene g(r) for simulations in toluene and heptane is shown in Figure 15. Interestingly, the g(r) shows 2 distinct peaks in heptane at ~0.60 and 0.78nm. The small peak at 0.6nm is most probably from a conformation of parallel stacked aromatic cores; this is not seen in toluene. This conclusion is supported by the average angle between the aromatic planes at the smallest distances between asphaltene centre of mass given in Figure 16. At ~0.6nm the average cosine of the angle between the aromatic planes is close to 1 in heptane, indicating mostly parallel arrangement of the aromatic planes. In toluene, the average cosine of the angle between the aromatic planes is ~0.65. Because the average value is larger than 0.5, we can argue that there is some preference for a parallel arrangement, but considerably less so than in heptane. The broader second peak centred at ~0.78nm in the g(r) is common to both simulations in toluene and heptane. At this distance there is little orientational preference for the aromatic planes. Therefore T-shaped and offset stacked dimer conformations are likely. The cumulative aggregation number, N(r), demonstrates that although there are distinct peaks in the g(r) at low *r*, these only correspond to a small number of molecules. Indeed it is only at 1.5nm separation that the average number of asphaltene molecules reaches one.

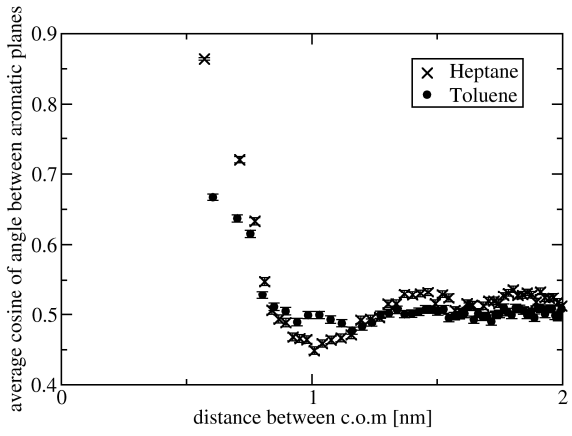


Figure 16: Average cosine of the angle between aromatic planes as dunction of distance between asphaltene centre of mass for simulations in toluene (filled circle) and heptane (cross)

The potential of mean force between asphaltene molecules may be calculated from the asphaltene-asphaltene g(r) using equation 6. This is shown in Figure 17. It is evident from the potential of mean force that the free energy of dimer formation between the asphaltene molecules is small, at most ~1kT. This is smaller than has been seen for other asphaltene molecules of this size generated by the QMR method[22]. A probable reason for this is that this molecule has a sterically constrained aromatic core, forcing it into a curved conformation and thus making it more difficult for asphaltene molecules to stack effectively.

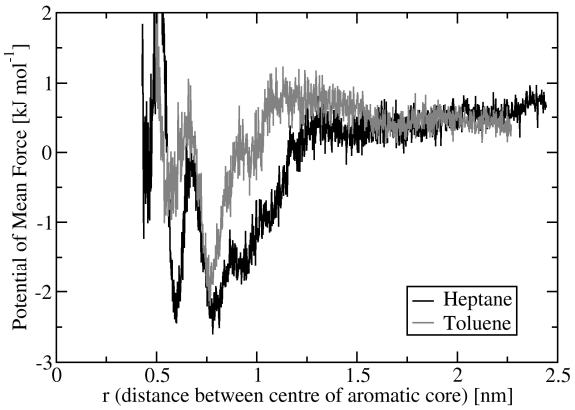


Figure 17: Asphaltene-asphaltene Potential of Mean force for simulations in toluene (grey) and heptane (black)

5. Conclusions and Discussion

We have created an improved computer algorithm to generate molecular representations of asphaltenes. Three-dimensional molecular representations were generated using a Monte Carlo method, which constructs asphaltene molecules from aromatic and aliphatic building blocks. The building blocks were sampled randomly and then linked together using a connection algorithm. We use a non-linear optimisation procedure to select a subset of molecules that gives the best match with experimental data. These experimental data consist of molecular weight, elemental analysis and NMR spectroscopy, including both ^{1}H and ^{13}C data. First, we validate the method by testing a number of single model compounds, for which we have accurate input data available. For these small compounds, we are able to predict the structures correctly. Subsequently, we used a data set available in the literature for a real Athabasca asphaltene sample, containing a mixture of a very large number of asphaltene compounds. We generated sets of 3000 and 4000 samples, which were then optimized with respect to the experimental penalty or objective function. The optimization procedure shows that a set of 5 – 6 asphaltene structures is sufficient to represent the experimental data. The method gives a significantly better match with the experimental data for a MW of 750 than for MW=4190, as measured by the value of the objective function averaged over 10 simulation runs. Note that this result is different from Sheremata,[6] and due to the fact that our linkage algorithm allows for both archipelago and pericondensed structures to be generated. The optimal value for the MW of 750 g/mol is supportive of the values recently suggested by Mullins *et al.*[9] associated with pericondensed asphaltene structures. Finally, the QMR structures generated from bulk asphaltene samples or deposits can be directly used to calculate the interaction between asphaltene molecules using MD simulations. As an example, we have done this for one of the optimized QMR structures in the final section of this paper. For a more extensive MD simulation study using QMR structures, we refer to ref. 22.

Acknowledgments

We are grateful to DBR Edmonton (Canada) for suggesting this line of research and the development of a practical software implementation which may form the basis of a future product. We are also grateful to John Crawshaw (SCR) for valuable comments. T. Headen acknowledges financial support of NERC UK.

[1] M.Neurock, A.Nigam, D.Trauth and M.T.Klein, *Chem. Eng. Sci.*, **1994**, 49 (24A), 4153-4177

[2] E. Rogel, *Energy & Fuels*, **1997**, 11, 920-925.

[3] E. Rogel, *Energy & Fuels*, **2000**, 14, 566-574.

[4] Brown, J. K.; Ladner, W. R.; Sheppard, N. *Fuel* **1960**, *39*, 79-86; ibid 87-96.

[5] D.M. Campbell, M.T. Klein, *Applied Catalysis A*, **1997**, 160, 41-54

[6] J.M.Sheremata, M.R.Gray, H,D.Dettman and W.C.McCaffrey, *Energy & Fuels* **2004,** 18, 1377 (2004).

[7] O.C. Mullins, *Fuel* **2007,** *86*, 309-312.

[8] Qian K, Rodgers RP, Hendrickson CL, Emmett MR, Marshall AG. *Energ Fuel* **2001** *15*:492

[9] H.Groenzin and O.C. Mullins, *J. Phys. Chem. A* **1999,** *103*, 11237-11245; *Energy & Fuels* **2000,** *14*, 677-684;

[10] "Bayesian Statistics 7", Proceedings of the 7th Valencia Int'l Meeting Edited by J. M. Bernardo, A.P. Dawid, J. O. Berger, M. West, D. Heckerman, M.J. Bayarri and A. F.M. Smith (2003)

[11] O.C.Mullins, B. Martínez-Haya and A. G. Marshall, *Energy & Fuels* **2008,** *22*, 1765–1773;

[12] Y. Ruiz-Morales and O.C. Mullins, *Energy & Fuels* **2007,** *21*, 256-265;

[13] H.Groenzin and O.C. Mullins, *J. Phys. Chem. A* **1999,** *103,* 11237-11245; *Energy & Fuels* **2000,** *14,* 677-684;

[14] E. Hirsch and K. H. Altgelt, *Anal. Chemistry*, **42**, 1970, 1330-1339.

[15] Pacheco-Sánchez, J. H.; Zaragoza, I. P.; Martínez-Magadán, J. *Energy and Fuels*, **2003**, *17*, 1346.

[16] Pacheco-Sánchez, J. H.; Álvarez-Ramírez, F.; Martínez-Magadán, J. M.; *Energy & Fuels* **2004,** *18*, 1676-1686

[17] Carauta, A. N. M.; Seidl, P. R.; Chrisman, E. C. A. N.; Correia, J. C. G.; Menechini, P. O.; Silva, D. M.; Leal, K. Z.; de Menezes, S..; de Souza, W. F.; Teixeira, M. A. G. *Energy & Fuels*, **2005,** *19*, 1245

[18] Carauta, A. N. M.; Correia, J. C.G.; Seidl, P. R.; Silva, D. M. *THEOCHEM*, **2005**, *755*, 1–8

[19] Zhang, L.; Greenfield, M. L. *Energy & Fuels*, **2007,** *21*, 1102-1111

[20] Zhang, L.; Greenfield, M. L. *J. Chem. Phys*, **2007**, *127*, 194502

[22] Headen, T. F.; Boek, E. S.; Skipper, N. T. Submitted to *Energy and Fuels*

[23] van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comp. Chem*, **2005**, *26*, 1701-1718

[24] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc*. **1996**, *118*, 11225-11236

[25] Jorgensen, W. L.; Laird, E. R.; Nguyen, T. B.; Tirado-Rive, J. *J. Comp. Chem*, **1993**, *14*, 206 - 215

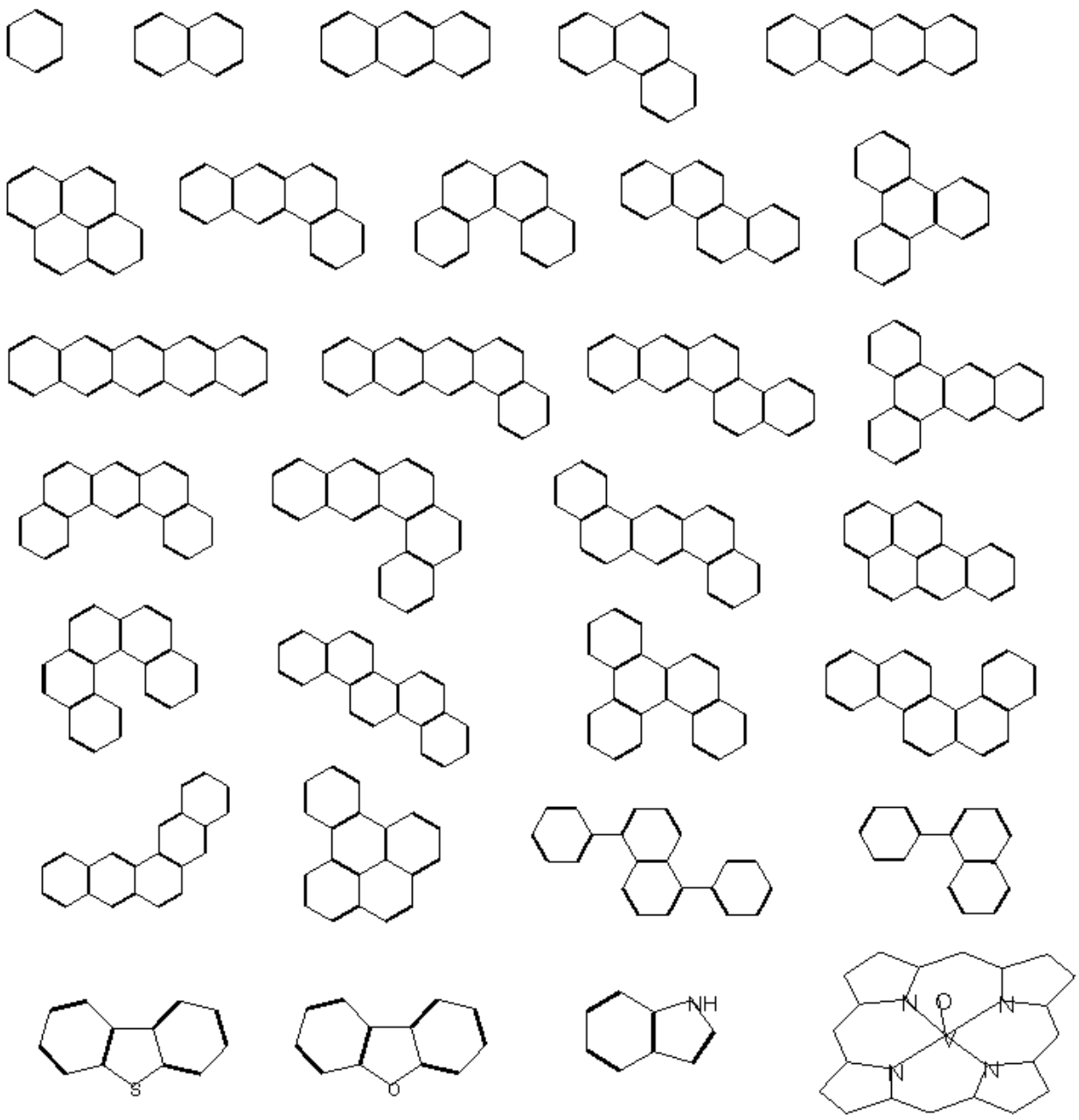
[26] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.*, **1993**, *98*, 10089 - 10092

[27] ArgusLab 4.0.1, Mark A. Thompson, Planaria Software LLC, Seattle, WA, http://www.arguslab.com

[28] Lee, L. L.; Molecular Thermodynamics of Nonideal Fluids. Butterworth-Heinemann, **1988.**

Appendix A: Building Blocks

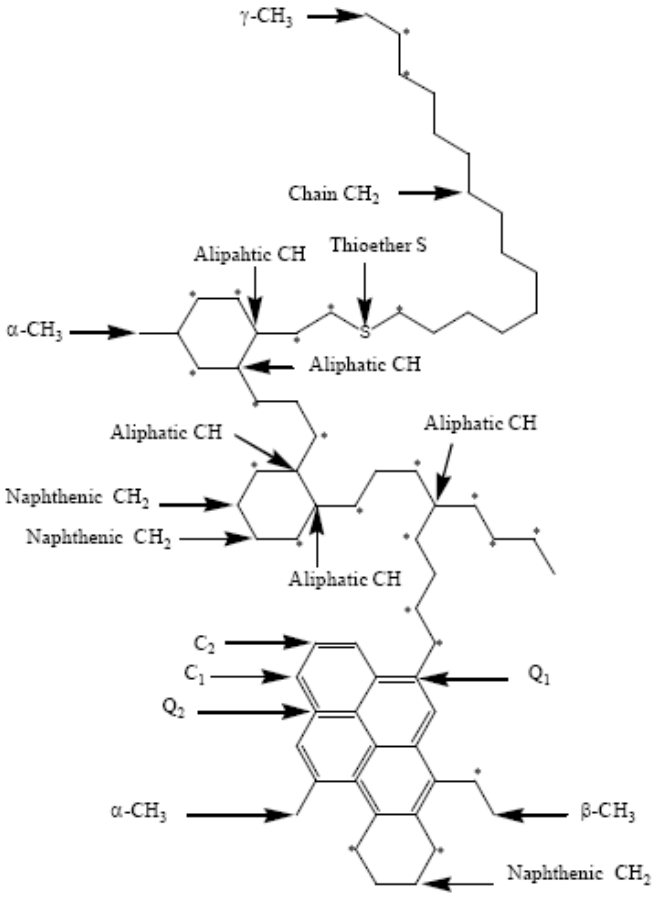
Here we present the predefined aromatic building blocks used in our model.



Appendix B: Atom types

The atom types are defined according to Sheremata[6] :

Atom type	Description
Aromatic	
Q1	Alkyl-substituted aromatic quaternary carbon
Q2	Bridgehead aromatic quaternary carbon
C1	Aromatic CH beside a Q2 carbon
C2	All aromatic CH that are not a C1 carbon
Aliphatic	
CH2	Chain CH_2
n-CH2	Naphthenic CH_2
o-CH2	Other aliphatic CH_2
CH	Aliphatic CH
a-CH3	α-CH_3
b-CH3	β-CH_3
g-CH3	γ-CH_3


γ-CH3
Chain CH2
Alipahtic CH
Thioether S
α-CH3
S
Aliphatic CH
Aliphatic CH
Aliphatic CH
Naphthenic CH2
Naphthenic CH2
Aliphatic CH
C2
C1
Q2
Q1
α-CH3
β-CH3
Naphthenic CH2