California State University, Monterey Bay

Digital Commons @ CSUMB

Biology and Chemistry Faculty Publications and Presentations

Department of Biology and Chemistry

2-2024

Curated dataset of asphaltene structures

Madison Franke

Selsela Arsala

Frozan Tahiry

Simon-Olivier Gingras

Arun K. Sharma

Follow this and additional works at: https://digitalcommons.csumb.edu/biochem_fac

This Article is brought to you for free and open access by the Department of Biology and Chemistry at Digital Commons @ CSUMB. It has been accepted for inclusion in Biology and Chemistry Faculty Publications and Presentations by an authorized administrator of Digital Commons @ CSUMB. For more information, please contact digitalcommons@csumb.edu.

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Curated dataset of asphaltene structures

Madison Franke, Selsela Arsala, Frozan Tahiry, Simon-Olivier Gingras, Arun K. Sharma*

Department of Biology and Chemistry, California State University, Monterey Bay, Seaside, CA 93955, USA

ARTICLE INFO

Article history:
Received 7 November 2023
Revised 25 November 2023
Accepted 29 November 2023
Available online 6 December 2023

Dataset link: Dataset of asphaltene structures (Original data)

Keywords:
Petroleum
Molecular structures
Computational chemistry
Crude oil

ABSTRACT

Asphaltenes, a distinct class of molecules found in crude oil, exhibit insolubility in nonpolar solvents like n-heptane but are soluble in aromatic solvents such as toluene and benzene. Understanding asphaltenes is crucial in the petroleum industry due to their detrimental effects on oil processing, resulting in significant economic losses and production disruptions. While no singular structure defines asphaltenes, two major molecular architectures, namely archipelago and continental models, have gained wide acceptance for their consistency with various experimental investigations and subsequent use in computational studies.

The archipelago model comprises two or more polyaromatic hydrocarbon entities interconnected via aliphatic side chains. In contrast, the island or continental model features a unified polyaromatic hydrocarbon moiety with 4 to 10 fused aromatic rings, averaging around 7 rings. To establish a comprehensive collection, we meticulously curated over 250 asphaltene structures derived from previous experimental and computational studies in this field. Our curation process involved an extensive literature survey, conversion of figures from publications into molecular structure files, careful verification of conversion accuracy, and structure editing to ensure alignment with molecular formulas. Our database provides digital structure files and optimized geometries for both predominant structural motifs. The optimization procedure commenced with the PM6 semi-empirical method, followed by further optimization utilizing density functional theory employing the B3LYP functional and the 6-31+G(d,p)

* Corresponding author.
E-mail address: arsharma@csumb.edu (A.K. Sharma).

https://doi.org/10.1016/j.dib.2023.109907

2352-3409/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

basis set. Furthermore, we compiled a range of structural and electronic features for these molecules, serving as a valuable foundation for employing machine learning algorithms to investigate asphaltenes. This work provides a ready to use structural database of asphaltenes and sets the stage for future research endeavours in this domain.

© 2023 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Specifications Table

Subject	Chemistry
Specific subject area	Computationally optimized structures of asphaltenes derived from crude oil
Data format	Raw: Gaussian input and output files .chk, .out, .gjf Optimized structure files in .pdb, .mol, .sdf
Type of data	Text files and Binary files
Data collection	Images of asphaltene structures were converted to computable molecular coordinate files through manual editing and partially automated structure conversion tools. Errors in structures were resolved through molecular editing features in Wolfram Mathematica v. 13.3.1. The resulting corrected structures were then optimized using semi-empirical PM6 technique, DFT B3LYP/6-31G(p) and finally at the DFT B3LYP/6-31+G(d,p) level of theory and basis set. All optimized structures were verified to be at a potential energy minimum through evaluation of vibrational frequencies and the absence of negative or imaginary values for those frequencies.
Data source location	Institution: California State University, Monterey Bay City: Seaside, CA Country: United States of America
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.10067907 Direct URL to data: https://zenodo.org/records/10067908
Related research article	None

1. Value of the Data

- The Density Functional Theory (DFT)-optimized molecular database of asphaltenes primes the field for the integration of machine learning and artificial intelligence. AI algorithms can utilize this high-quality data to uncover patterns and relationships that are not immediately apparent, facilitating the discovery of new materials with optimal properties for energy production, storage, and even waste reduction. This intersection of computational chemistry and data science could significantly shorten the research and development cycles in energy industries.
- The extensive database provides a detailed molecular-level understanding of asphaltenes, which are complex and inherently heterogeneous in nature. Through DFT optimization, the data offers precise geometries, electronic structures, and energetic properties. Such granularity is critical for predicting reactivity, understanding aggregation behavior in hydrocarbon mixtures, and tailoring asphaltene management strategies in oil recovery and processing.
- Optimized molecular structures allow for the creation of robust predictive models for asphaltene behavior under various environmental conditions. This is essential for the development of more efficient crude oil processing methods, predicting the likelihood of pipeline clogging, and informing the design of novel dispersants and inhibitors that can effectively mitigate asphaltene deposition.

- The database serves as a foundational resource for computational and experimental chemists in the energy sector, enabling rapid hypothesis testing and simulation. By leveraging these optimized structures, researchers can expedite the development of new materials and processes, including the synthesis of novel asphaltene derivatives with desired properties for industrial applications.
- While the direct implications for petroleum chemistry are evident, the molecular data also has potential applications in related fields. It could influence the development of organic photovoltaic materials, provide insights into carbon-rich molecule behavior in environmental systems, and contribute to the broader understanding of complex molecular systems within organic chemistry.

2. Background

The dataset was compiled with the intention of shedding light on the elusive nature of asphaltene molecules, which play a critical role in the oil industry, particularly in areas such as oil recovery, refining, and pipeline transport. Asphaltenes represent a significant challenge due to their complex, variable structures, and propensities to aggregate, leading to operational issues like deposition and emulsion stabilization. The theoretical backbone of this compilation is based on Density Functional Theory (DFT), a quantum mechanical modeling method used to investigate the electronic structure principally the ground state of complex systems. The methodological approach involved literature review to identify unique structures, conversion of those structures to coordinate files and finally systematic DFT optimization to ascertain stable conformations and electronic characteristics.

The value of this dataset lies in its foundational character, serving both as an extension and a deep-dive elaboration of existing research. It complements published articles by offering a ready-to-use, comprehensive molecular structure library of asphaltenes for theoretical and experimental researchers. This dataset constitutes a fundamental informational repository, with its intrinsic merit anchored in its prospective utilization for computational simulations, predictive modeling, and ensuing empirical corroboration.

3. Data Description

The dataset is presented in two compressed files. The Optimized-Structure-Files-Asphaltenes.zip file has sub-directories: Archipelago and Continental. Each of these directories contains two sub-directories, Experimental and Model, corresponding to the source of the structures, from experimental investigations or modeling studies. Each structure is presented in 3 formats, PDB (Protein Data Bank), MDL Mol (Molecular Design Mol), and SDF (Structure Data File). The Gaussian calculation results are presented in the Asphaltenes-Dataset-Zenodo-Repository.tbz file. This is a bzip2 compressed file and has a similar organizational philosophy. The main folders are called Experimental-Structures and Model-Structures. Each of these directories contains two sub-directories called Archipelago and Continental respectively to separate those structures. For each Gaussian calculation, the input files are .gjf, the checkpoint files are .chk, and the output files are .gjf.out extensions.

4. Experimental Design, Materials and Methods

The data have been classified into two groups of archipelago and continental structures and further subdivided to clearly identify structures from experimental or theoretical sources. In total 255 asphaltene structures have been collected. 70 of these structures correspond to the

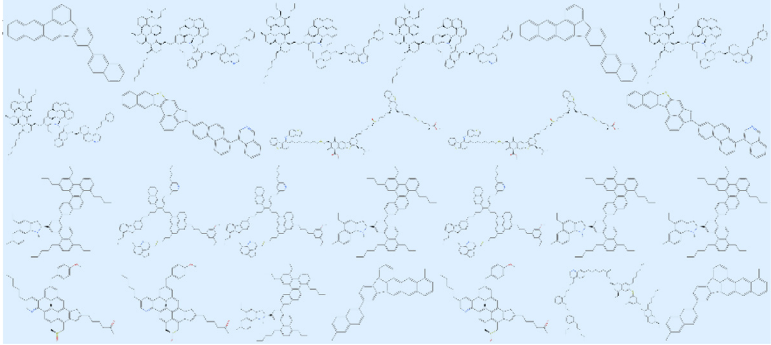

Fig. 1. An image collage of 25 archipelago structures collected from the literature. All structures are not represented here for clarity.

archipelago motif and the rest 185 structures correspond to the continental motif. The number and type of structures from the literature are presented in Table S1.

The Wolfram language (WL) [1] was utilized for creating the molecules. The structures depicted in the referenced articles were digitally captured via screenshots. These images were subsequently processed through WL's "MoleculeRecognize" [2] function, enabling a digital rendition of each molecular structure. These rendered structures were then cross-referenced with their counterparts in the referenced literature. When discrepancies arose, the "MoleculeDraw" [3] function was invoked, and the structures were edited to correct any errors in the automated conversion process. Around 15% of the structures needed manual adjustments to correct deficiencies.

While the primary references for these depictions were the displayed formulas in the respective papers, the accuracy of each rendered molecule was verified through additional checks. If the molecular weight or formula of the structures was included in the referenced literature, these data points were incorporated into the verification process. Once a structure satisfactorily passed all accuracy checks, it was subjected to a multi-stage molecular optimization process.

To ensure the avoidance of duplicate asphaltene molecules, the WL's "MoleculeMatchQ" [4] function was implemented. This function cross-examines each molecule within the dataset, confirming a lack of identical matches in bonding and connection. Through this scrutiny, redundant asphaltene molecules were identified within our dataset and eliminated from the final collection. The final collection consists of 255 distinct asphaltene structures.

Below we present some descriptive information about the dataset of asphaltenes. The structures of literature collected asphaltenes are displayed in Figs. 1 and 2. Fig. 1 highlights archipelago structures and Fig. 2 displays continental structures.

Asphaltenes derived from different sources share the same solubility characteristics (e.g., toluene soluble, n-heptane insoluble), but the composition of asphaltenes derived from different sources can vary greatly. For example, a key chemical parameter that can differ is the aromaticity which for carbonaceous materials correlates strongly with the ratio of hydrogen to carbon. Petroleum asphaltenes that have not been subjected to any processing such as refining, have an aromaticity near 50% and an H/C ratio of ∼1.1 [5] .The aromaticity of asphaltene molecules plays a crucial role in the aggregation process; heightened aromaticity fosters asphaltene aggregation. Specifically, asphaltenes possessing an elevated aromaticity are prone to aggregate under certain thermodynamic conditions. Conversely, asphaltenes exhibiting notably reduced aromaticity are less likely to aggregate across a broad spectrum of thermodynamic condition envelopes [6]. The H/C ratios of the collected asphaltene structures are shown in Fig. 3. The archipelago structures have a slightly higher unsaturation than the continental structures. The median unsaturation for the archipelago structures is 1.10 and for the continental structures it is about 0.95. However, there is less variance among the archipelago structure H/C ratios and significantly larger variation in the continental structures' dataset.

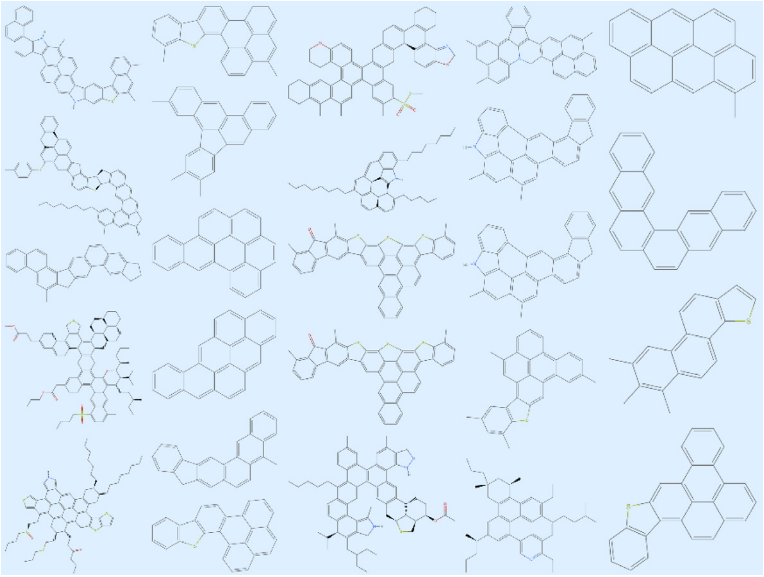

Fig. 2. An image collage of 25 continental structures collected from the literature. All structures are not represented here for clarity.

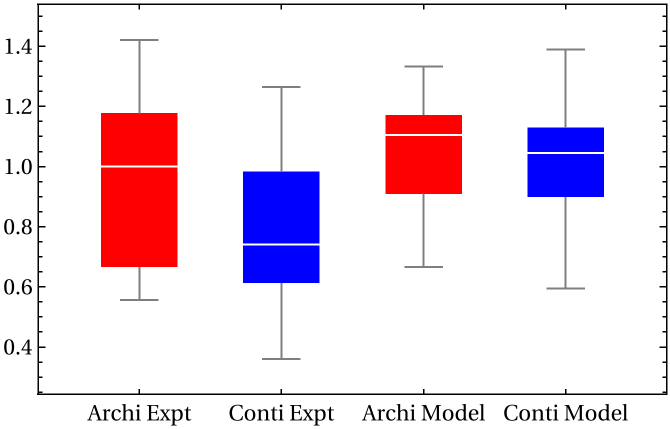

Fig. 3. H/C ratios of archipelago and continental structures.

There has been considerable debate on the aggregation tendency of asphaltenes and their constitution as monomeric or polymeric species. Presently, the consensus is that asphaltenes are monomeric [7], and techniques such as time-resolved fluorescence depolarization and mass spectrometry have converged on molecular weights between ∼250 and 1200 g/mol, with an average of approximately 750 g/mol [8].

The molar masses of archipelago and continental structures shown in Fig. 4 in our dataset follow this trend. The minimum, median and the maximum values for the mean molar mass of the archipelago structures are 270, 700 and 1500 g/mol, respectively. Furthermore, the minimum, median and the maximum values for the molar mass of the continental structures are 230, 750 and 1800 g/mol, respectively. Overall, the range in the molar mass of the continen-

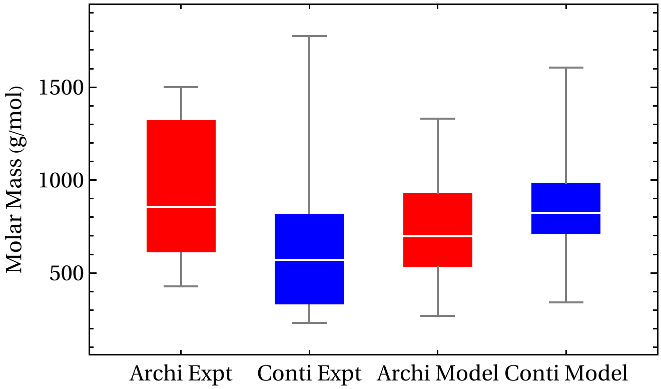

Fig. 4. Molar mass of archipelago and continental structures.

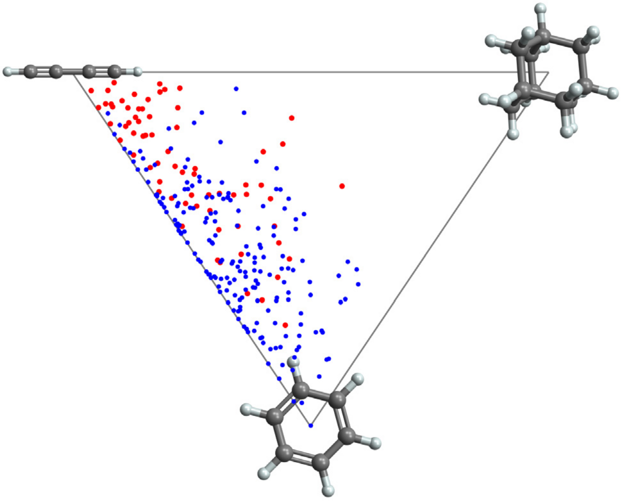

Fig. 5. Molecular shape analysis of archipelago and continental structures. Red dots represent archipelago structures whereas continental structures are represented by blue dots.

tal structure is larger, but overall continental structures have a higher mean molar mass than archipelago structures.

The molecular shapes of asphaltenes are classified using the principal moments of inertia (PMI) and the shapes are visualized on a PMI plot [9–11]. The three corners of the triangle represent rod, disk, and sphere shapes respectively. These shapes are represented by buta-1,3-diyne, benzene, and adamantane molecules respectively. The principal moments of inertia are calculated from the 3D coordinates and are normalized by dividing the two smaller moments by the largest value. Using these ratios as coordinates, the shapes form the corners of a triangle as shown in Fig. 5.

The red dots in Fig. 5 represent the archipelago structures and the blue dots represent continental structures. It is clearly seen that the archipelago structures have a tendency towards the rod shape, whereas the continental structures have a tendency towards the flat-disk shape. And overall, neither the archipelago structures nor the continental structures have a tendency towards the spherical shape. This trend is in excellent agreement with the definitions of archipelago and continental structures. Archipelago structures possess multiple aliphatic chains,

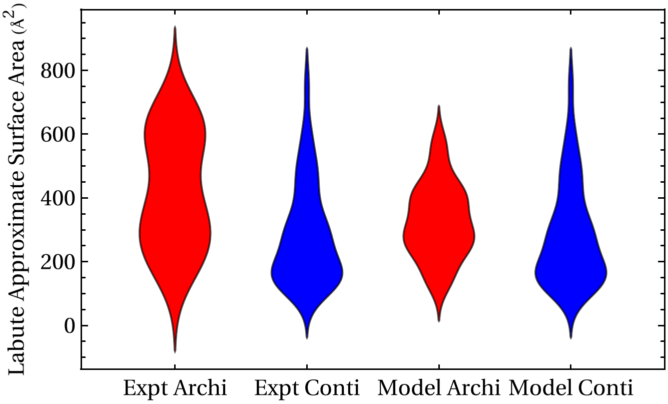

Fig. 6. Violin plots representing the distribution of Labute approximate surface area computed for all molecules in the dataset.

whereas continental structures have a large aromatic core and only a few aliphatic periphery chains.

After these preliminary data analysis from the collected structures, the geometries were optimized using Gaussian 16 [12]. In the presented study, a two-step optimization process was employed for the asphaltene structures, starting with the PM6 semi-empirical method and followed by optimization using Density Functional Theory (DFT). The PM6 method was selected for its computational efficiency, which is crucial when dealing with the large size of asphaltene molecules. This method strikes a balance between accuracy and computational demand, enabling the rapid generation of reasonable molecular geometries. The role of PM6 was to provide a reliable starting geometry for the subsequent optimization steps, which is particularly important for complex molecules in the dataset. DFT was then utilized for refining these geometries. Owing to its effectiveness in handling large molecules, DFT offers an elevated level of accuracy, especially in scenarios where electron correlation plays a significant role. Despite its increased computational intensity, DFT provides an optimal balance between computational load and precision, a critical aspect for the study of asphaltenes. The combination of the PM6 method and DFT ensures both scientific rigor and practical feasibility in the curation of the dataset. This approach allows for the efficient processing of numerous molecules while maintaining the accuracy of the final optimized structures. The initial optimization was carried out using the semi-empirical PM6 approach [13]. The geometries were carefully analyzed to ensure that there were no imaginary frequencies returned during the calculation. Following this successful optimization, these geometries were used as initialization geometries for a hybrid Density Functional Theory (DFT) calculation with the B3LYP functional and the 6-31+G(d,p) basis set [14]. The geometry optimization and frequency calculation protocols were implemented again, and the final optimized geometries have been provided as part of the dataset along with Gaussian output files and chk files. The geometries were verified to be minimums on the potential energy surface through the calculation of vibrational frequencies. There were no negative or imaginary frequencies for these structures. These optimized structures were then analyzed to calculate the approximate surface areas and molecular volumes. The isotropic polarizability of the molecules was calculated with the “polar” keyword. All calculations were carried out on the Expanse [15] supercomputer at SanDiego Supercomputer Center provided through an ACCESS [16] allocation.

The approximate Van der Waals surface area also known as the Labute approximate surface area [17,18] for all asphaltene molecules is displayed in Fig. 6 as a violin plot for each set of molecules. A violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data. The continental asphaltenes dataset shows much broader deviation and larger average surface area compared to the archipelago structures. These values were computed using the 3D coordinates obtained from the optimized geometry.

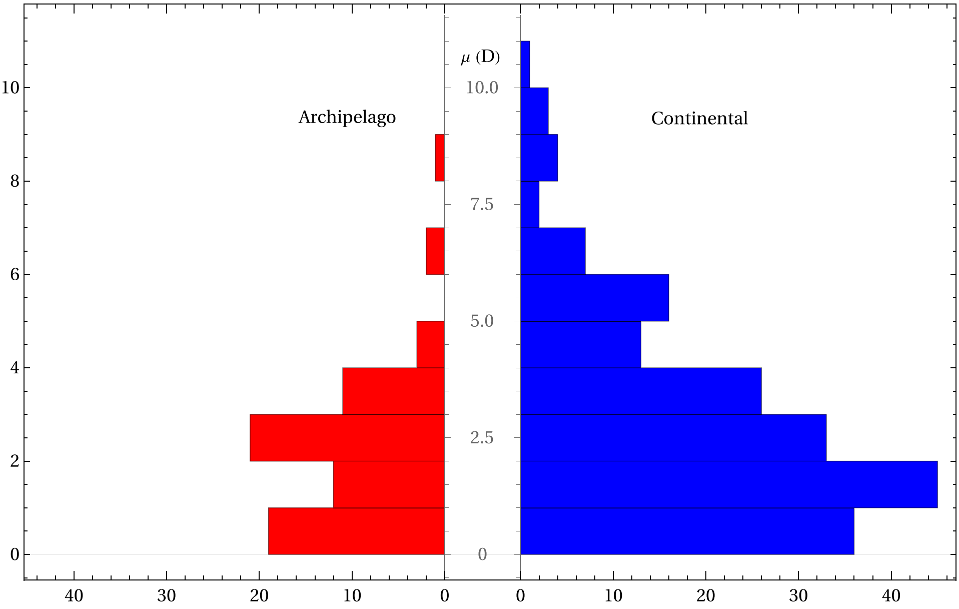

Fig. 7. Paired histogram of dipole moment values for asphaltenes. The larger number of continental asphaltenes naturally results in wider spread of dipole moment values.

The dipole moments of the molecules were also extracted from the Gaussian output files. Gaussian reports the dipole moment vector for the molecule and the value was calculated as:

$$\mu = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2}$$

A summary of the dipole moment values of the asphaltenes in the form of a paired histogram is represented in Fig. 7. The dataset contains roughly double the number of continental molecules as archipelago molecules and consequently the variation in dipole moments and other properties is far richer for continental molecules. The larger surface area on average and the higher spread of dipole moments for continental asphaltenes alludes to the polarizability trends that will be shown later.

A bubble chart of the molecular volume connected with the molar mass and the surface area of asphaltenes is quite informative and shown in Fig. 8. The molecular volume was calculated using the equation:

$$V_{vdw} = \sum \text{all atom contributions} - 5.92N_B - 14.7R_A - 3.8R_{NA}$$

where N_B is the number of bonds, R_A is the number of aromatic rings, and R_{NA} is the number of nonaromatic rings. This formulation has been used successfully in calculation of molecular volumes in literature [19]. This analysis was carried out to ensure internal consistency in the dataset. The size of the bubbles is proportional to the molecular volume. The volumes of both continental and archipelago asphaltenes appear to increase in direct proportionality to their respective surface area and molar masses. The distribution of molecular volumes is provided in supporting information Figure S1.

The energy difference between the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO), is called the HOMO-LUMO gap. The size of the HOMO-LUMO gap can be used to predict the strength and stability of structures. HOMO-LUMO gap values were extracted from the Gaussian calculation results and are displayed as a paired histogram for all molecules in the dataset in Supporting Information Figure S2. All structure files and optimized geometries have been provided through the Zenodo [20] data repository.

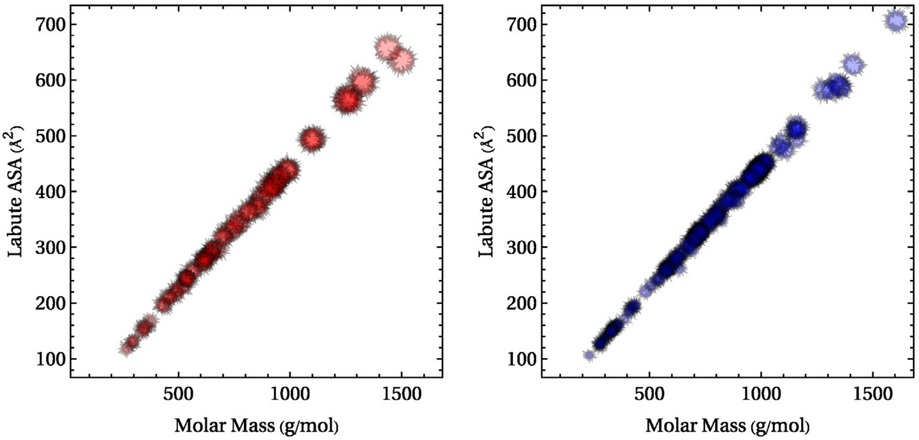

Fig. 8. Bubble chart of the molecular volume with surface area and molar mass. The size of the bubble is indicative of the volume of that molecule. On the left are the archipelago molecules while continental molecules are on the right. The bubbles are rendered with some transparency- darker areas in the figures imply larger numbers of molecules within that range of molar mass, surface area, and molecular volume. The clear trend is within expectation and helps to establish consistency of these internal metrics.

Limitations

A significant limitation is the dataset's reliance on figures from previous publications. In selecting the 255 asphaltene structures for our study, we conducted a thorough literature search, focusing on models and experimental studies from the past two decades. This approach ensured that our dataset reflects contemporary understanding and applications in the field of asphaltene research. However, it's important to acknowledge that some of these structures have also been utilized in studies prior to this period. While our selection criteria aimed for a broad representation of current trends and knowledge, this temporal focus may omit certain historical perspectives and applications of these structures. Recognizing this limitation is essential for a comprehensive understanding of the scope and applicability of our dataset. Despite implementing various checks, there remains a possibility that errors may have crept in during the interpretation of these figures. Another primary source of error identified is the initial conversion of molecular structures from 2D to 3D, where minor inaccuracies could lead to discrepancies. Mitigation efforts included the use of established computational methods and rigorous verification steps. Variability in experimental conditions reported in literature also introduces uncertainty. To counter this, our dataset included only structures with comprehensive experimental details, and cross-referencing was employed for consistency. It is recognized, however, that computational optimization methods like DFT, while robust, are not without limitations.

Ethics statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Dataset of asphaltene structures (Original data) (Zenodo).

CRediT Author Statement

Madison Franke: Data curation, Validation, Visualization, Investigation; **Selsela Arsala:** Data curation, Validation, Visualization, Investigation; **Frozan Tahiry:** Data curation, Writing – original draft, Investigation; **Simon-Olivier Gingras:** Validation, Visualization; **Arun K. Sharma:** Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing, Conceptualization, Methodology, Software.

Acknowledgements

This work was supported by the donors of ACS Petroleum Research Fund under Undergraduate Research Grant 61864-UR4. This work used EXPANSE at the San Diego Supercomputer Center through allocation TG-CHE230074 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2023.109907.

References

[1] S. Wolfram, What We've built Is a computational language (and that's very important!), J. Comput. Sci. 46 (2020), doi:10.1016/j.jocs.2020.101132.

[2] Wolfram Research, MoleculeRecognize—Wolfram Language Documentation, (n.d.). https://reference.wolfram.com/language/ref/MoleculeRecognize.html (Accessed 7 April 2022).

[3] Wolfram Research, MoleculeDraw—Wolfram Language Documentation, (n.d.). https://reference.wolfram.com/language/ref/MoleculeDraw.html (Accessed 22 May 2023).

[4] Wolfram Research, MoleculeMatchQ—Wolfram Language Documentation, (n.d.). https://reference.wolfram.com/language/ref/MoleculeMatchQ.html (Accessed 22 May 2023).

[5] H. Wang, H. Xu, W. Jia, J. Liu, S. Ren, Revealing the intermolecular interactions of asphaltene dimers by quantum chemical calculations, Energy Fuels 31 (2017) 2488–2495, doi:10.1021/acs.energyfuels.6b02738.

[6] M. Ahmadi, Z. Chen, Molecular interactions between asphaltene and surfactants in a hydrocarbon solvent: application to asphaltene dispersion, Symmetry 12 (2020) 1–18, doi:10.3390/sym12111767.

[7] A.B. Andrews, R.E. Guerra, O.C. Mullins, P.N. Sen, Diffusivity of asphaltene molecules by fluorescence correlation spectroscopy, J. Phys. Chem. A 110 (2006) 8093–8097, doi:10.1021/jp062099n.

[8] M.R. Gray, M.L. Chacón-Patiño, R.P. Rodgers, Structure-reactivity relationships for petroleum asphaltenes, Energy Fuels 36 (2022) 4370–4380, doi:10.1021/acs.energyfuels.2c00486.

[9] W.H.B. Sauer, M.K. Schwarz, Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity, J. Chem. Inf. Comput. Sci. 43 (2003) 987–1003, doi:10.1021/CI025599W.

[10] I. Motoc, Molecular shape descriptors, in: Steric Effects in Drug Design, 2007, pp. 93–105, doi:10.1007/bfb0111214.9.

[11] MoleculePrincipalMomentPlot | Wolfram Function Repository, (n.d.). https://resources.wolframcloud.com/FunctionRepository/resources/MoleculePrincipalMomentPlot/(Accessed 1 June 2023).

[12] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi,

J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Gaussian16 Revision B.01, 2016.

[13] J.J.P. Stewart, Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements, J. Mol. Model. 13 (2007) 1173–1213, doi:10.1007/S00894-007-0233-4.

[14] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, J. Chem. Phys. 98 (1993) 5648–5652, doi:10.1063/1.464913.

[15] S. Strande, H. Cai, M. Tatineni, W. Pfeiffer, C. Irving, A. Majumdar, D. Mishin, R.S. Sinkovits, M.M. Norman, N. Wolter, T. Cooper, I. Altintas, M. Kandes, I. Perez, M. Shantharam, M. Thomas, S. Sivagnanam, T. Hutton, H. Cai, T. Cooper, C. Irving, T. Hutton, M. Kandes, A. Majumdar, D. Mishin, I. Perez, W. Pfeiffer, M. Shantharam, R.S. Sinkovits, S. Sivagnanam, S. Strande, M. Tatineni, M. Thomas, N. Wolter, M.M. Norman, T. Hut-ton, Expanse: computing without boundaries: architecture, deployment, and early operations experiences of a supercomputer designed for the rapid evolution in science and engineering, Practice and Experience in Advanced Research Computing, Association for Computing Machinery, New York, NY, USA, 2021, doi:10.1145/3437359.

[16] T.J. Boerner, S. Deems, T.R. Furlani, S.L. Knuth, J. Towns, ACCESS: advancing innovation, in: Association for Computing Machinery (ACM), 2023, pp. 173–176, doi:10.1145/3569951.3597559.

[17] P. Labute, Derivation and Applications of Molecular Descriptors Based on Approximate Surface Area, 2004, pp. 261–278, doi:10.1385/1-59259-802-1:261.

[18] P. Labute, A widely applicable set of descriptors, J. Mol. Graph Model. 18 (2000) 464–477, doi:10.1016/S1093-3263(00)00068-1.

[19] Y.H. Zhao, M.H. Abraham, A.M. Zissimos, Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds, J. Org. Chem. 68 (2003) 7368–7373, doi:10.1021/jo034808o.

[20] A.K. Sharma, Dataset of asphaltene structures, (n.d.). https://zenodo.org/records/10067908 (Accessed 2 November 2023).