1-2024

Machine learning to identify structural motifs in asphaltenes

Arun K. Sharma

Selsela Arsala

James Brady

Madison Franke

Shelby Franke

See next page for additional authors

Authors

Arun K. Sharma, Selsela Arsala, James Brady, Madison Franke, Shelby Franke, Supreet Gandhok, Simon-Olivier Gingras, Ana Gomez, Katelyn Huie, Kayla Katz, Samantha Kozlo, Mateo Longoria, Levi Molnar, Nathaly Peña, and Sarina Regis

Results in Chemistry 7 (2024) 101551

Contents lists available at ScienceDirect

Results in Chemistry

journal homepage: www.sciencedirect.com/journal/results-in-chemistry

Machine learning to identify structural motifs in asphaltenes

Arun K. Sharma *, Selsela Arsala, James Brady, Madison Franke, Shelby Franke, Supreet Gandhok, Simon-Olivier Gingras, Ana Gomez, Katelyn Huie, Kayla Katz, Samantha Kozlo, Mateo Longoria, Levi Molnar, Nathaly Peña, Sarina Regis

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

ARTICLE INFO

Keywords:
Asphaltenes
Machine learning
Image recognition
Molecular topology
Deep learning

ABSTRACT

Asphaltenes are organic compounds that aggregate in crude oil with two dominant molecular architectures: archipelago and continental. Continental architectures possess a single uniform island structure composed of aromatic rings in contrast to archipelago architectures with aromatic cores interconnected through aliphatic chains. The structural composition of asphaltenes varies globally due to geographical differences, posing challenges in their classification due to a lack of uniformity. This study is the first known exploration of using image-based supervised machine learning, particularly the ResNet-50 neural network, for the binary classification of asphaltenes into continental and archipelago motifs. 255 continental and archipelago models underwent structural augmentations to create a sample size of 1,530 asphaltene structures that is robust enough for accurate results in both the training and testing portions of the machine learning. These augmentations included the repeated addition of carbons until a complete pentane chain was added to a specified carbon on each asphaltene structure. Using Mathematica, supervised ResNet-50 image-based classification was used on both original and augmented structure datasets to classify as either archipelago or continental. The classification was also implemented using topological similarity searching for association between atoms and the distance between them for further molecule identification. This study demonstrates the surprising effectiveness of image-based classification compared to traditional topological feature-based methods. Our results reveal that deep learning techniques, especially image-based approaches, provide novel and insightful ways to differentiate complex molecular structures like asphaltenes, challenging the traditional reliance on topological features alone. This research opens new avenues in chemical analysis and molecular characterization, highlighting the potential of machine learning in complex molecular systems.

Introduction

Asphaltenes, a complex component of crude oil, present significant challenges in the petroleum industry due to their structural diversity and propensity to affect processing and refining operations adversely. The structural characterization of asphaltenes remains a pivotal yet challenging aspect, primarily due to their inherent molecular heterogeneity and variation across different geographical sources. This complexity necessitates advanced analytical approaches for accurate classification and understanding. The rapidly evolving field of machine learning (ML) offers promising tools in this regard [1–3].

Machine learning's integration into chemical research has transformed the approach to molecular analysis and characterization. Its application in structural characterization has opened new opportunities for understanding molecular patterns. A recent example demonstrates the utility of ML models to predict shape persistence and cavity size in porous organic cages. The models achieved up to 93 % accuracy, with the [4 + 6] imine condensation of trialdehydes and diamines identified as the most effective method for creating shape persistent structures [4]. This approach represents a significant advancement in understanding and predicting the structural properties of hypothetical organic cages, essential for a range of applications. In this context, supervised learning, known for its effectiveness in pattern recognition, offers a robust framework for classifying complex molecular structures [4,5].

Deep learning, a more advanced form of ML, has further revolutionized data analysis, especially in image-based applications. The Residual Network (ResNet) architecture, particularly the ResNet-50 [6,7] model, is notable for its ability to handle image data efficiently. ResNet-

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

https://doi.org/10.1016/j.rechem.2024.101551
Received 31 January 2024; Accepted 18 May 2024
Available online 19 May 2024

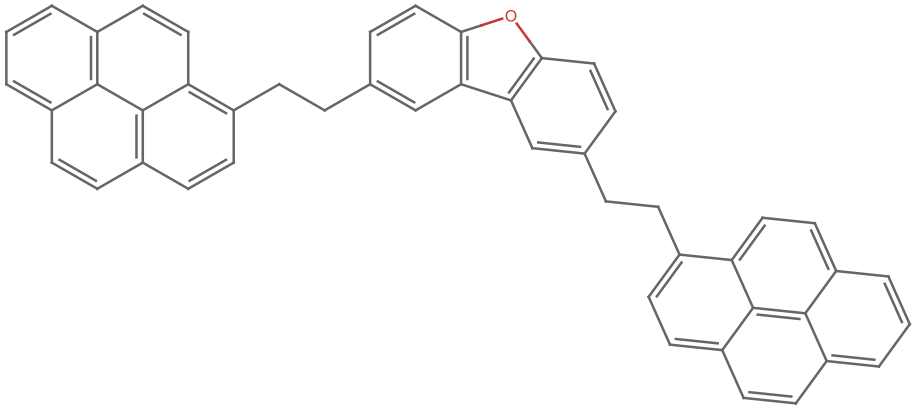

Fig. 1. Example of an archipelago structural motif.

50, a variant of this architecture, stands out for its innovative approach to deep learning, particularly in image classification tasks. A defining feature of ResNet-50 is its use of residual blocks, which incorporate skip connections that allow the network to bypass one or more layers. These skip connections effectively address the vanishing gradient problem, a common challenge in deep neural networks, by facilitating the flow of gradients through the network, thereby enabling the training of much deeper models without a loss in performance. The architecture of ResNet-50 comprises 50 layers, including convolutional layers, pooling layers, and fully connected layers, designed to extract and process a wide range of features from input images. This depth, combined with the efficiency of residual learning, allows ResNet-50 to achieve remarkable accuracy in image recognition tasks, outperforming many other deep learning models, especially in complex datasets like ImageNet [8,9]. ResNet-50 has been widely utilized in diverse fields and some notable examples include chemistry laboratory glassware identification [10], identification of weeds [11], and classification of fungi and bacteria [12].

This study utilizes the ResNet-50 model, trained on a dataset of two-dimensional asphaltene images, to classify them into continental and archipelago motifs. The primary objective is to evaluate the feasibility and effectiveness of using supervised learning and deep learning techniques in differentiating asphaltene structures into continental and archipelago motifs. This study aims to contribute new insights to the applications of convolutional neural networks to chemical structures and cheminformatics.

Molecular representation is a critical factor in ML applications that significantly influences the performance and accuracy of predictive models. These representations are generally classified into four distinct categories: string representations, connection tables, feature-based representations, and computer learning-based representations. Among these, string representations such as the Simplified Molecular-Input Line-Entry System (SMILES), International Chemical Identifier [13,14] (InChI), and MDL Molfile have emerged as significant tools for encoding molecular structures in a compact and interpretable format [15].

Molecular fingerprints, a pivotal aspect of molecular representation in ML, effectively encapsulate molecular substructures as sparse vectors. These fingerprints are derived in two primary ways: either through the matching of substructures based on expert-defined sets or through algorithmic enumeration and hashing of substructures. A notable example of such fingerprinting techniques is the Extended Connectivity Fingerprints (ECFP) [15], widely recognized for its utility and implemented in the RDKit Python package [16].

The precision of ML models in predicting molecular properties is

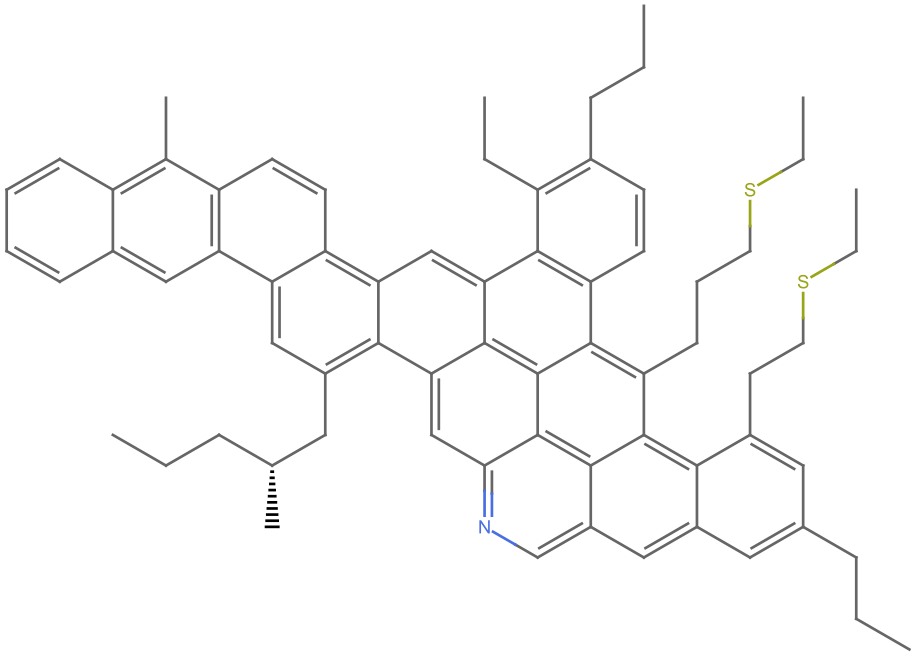

Fig. 2. Example of a continental structural motif.

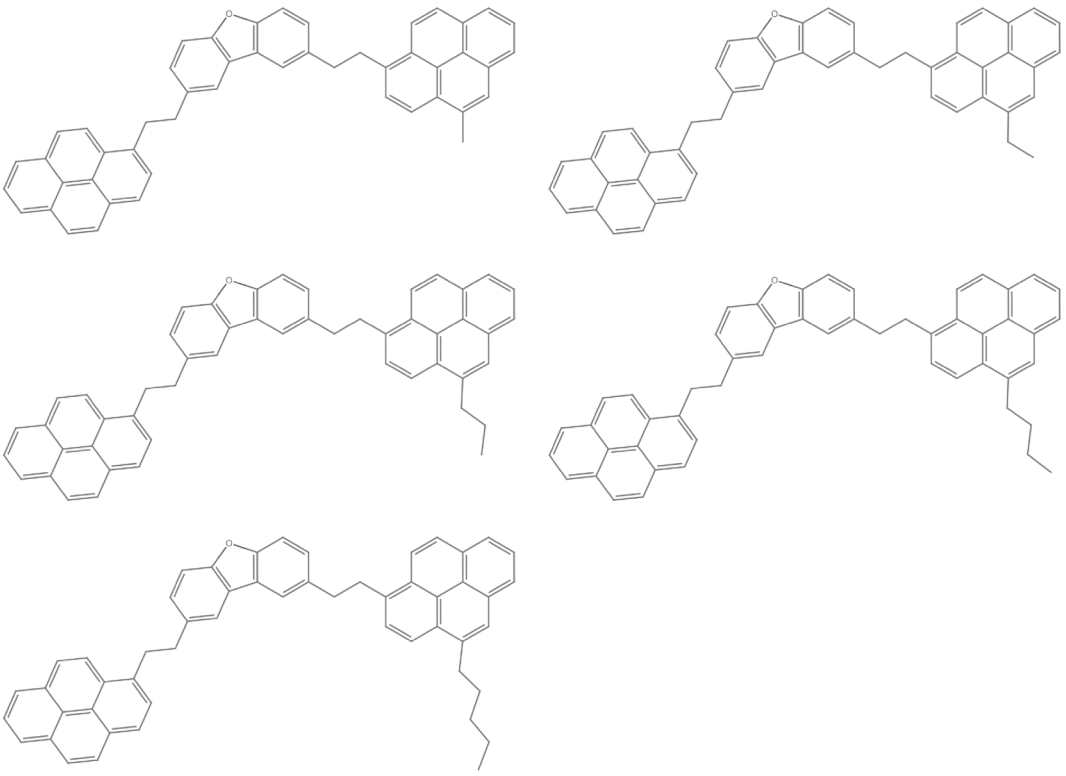

Fig. 3. Example of archipelago model 48 with augmentations. Each structure was edited using the MoleculeDraw function. From left to right, each structure was edited by adding one additional carbon to the same location as before. Modifications began with the addition of a methyl group with the successive addition of carbons up to a pentane chain.

heavily dependent on the choice of molecular representation. A nuanced approach, drawing inspiration from quantum mechanics principles, has led to the development of a hierarchy of representations that adhere to criteria of uniqueness and target similarity. This approach ensures that each molecule is represented in a unique manner while maintaining consistency with the underlying chemical properties. Incorporating higher-order contributions into these representations consistently improves similarity to the true potential energy surface. Consequently, this enhances the predictive accuracy of the resultant ML models, as it allows for a more comprehensive and nuanced understanding of molecular interactions and properties [17,18].

The improvement in the accuracy of machine learning (ML) models for predicting molecular properties has significant implications for addressing complex chemical phenomena, such as the behavior of asphaltenes in crude oil processing. The development of more sophisticated molecular representations, informed by principles of quantum mechanics, allows for a nuanced understanding of molecular interactions and properties. This advancement is critical for industries that deal with complex substances whose behavior is difficult to predict with traditional models. Asphaltenes present a notable challenge in the oil industry, characterized by their unconventional solubility properties, which complicate crude oil processing and refinement. Asphaltenes defy conventional solubility norms by favoring aromatic solvents and precipitating in normal alkanes. This specific behavior is characterized by solubility in aromatic hydrocarbons and insolubility in alkanes and renders them a challenge in crude oil processing. Asphaltenes, consisting primarily of carbon, nitrogen, and sulfur, present a puzzle that spans both chemical and physical dimensions [19]. Archipelago asphaltenes are characterized by their distinct molecular architecture, where smaller and more dispersed aromatic structures are linked by aliphatic chains, resembling a cluster of islands. This configuration leads to relatively lower molecular weights and contributes to their unique solubility and reactivity properties compared to other asphaltene types. Fig. 1 shows an archipelago asphaltene from the dataset used in our investigation. Continental asphaltenes, in contrast, feature larger, more condensed aromatic systems, like a single contiguous landmass. This structure results in higher molecular weights and influences their aggregation behavior, impacting the processing and handling characteristics of crude oil where they are prevalent [20–23]. Fig. 2 illustrates a continental asphaltene structure that displays these characteristics.

Methods

Asphaltene structures from experimental and computational sources collected by Franke et al. [24] were utilized in this investigation. That dataset contains 255 asphaltene structures, of which 185 structures correspond to the continental motif and the remaining 70 structures are archipelago type of structures. To compile a sample size that is robust enough for accurate results in both the training and testing portions of machine learning, each structure in the archipelago and continental data sets underwent five structural augmentations. Using Wolfram Language [25], each structure was edited with the MoleculeDraw [26] function. For each modification, we added one additional carbon to the same location as the previous edit, ultimately creating a one, two, three, four, and five-carbon chain addition to the molecule. Fig. 3 shows the modified structures created from the archipelago motif shown in Fig. 1. Similarly, Fig. 4 shows the modified structures derived from the continental motif illustrated in Fig. 2. All carbon additions were made on the exterior of the molecule, distant from aromatic ring islands for continental molecules. Correspondingly, in archipelago structures, the modifications were located distant from the connecting branches between aromatic rings to maintain structural integrity. Aromatic rings were not added to any augmented structures since the arrangement of aromatic rings is the dominant feature of asphaltene molecules. Each structure was then energy minimized and stored. We then compiled the original structures combined with the augmented structures to create a larger

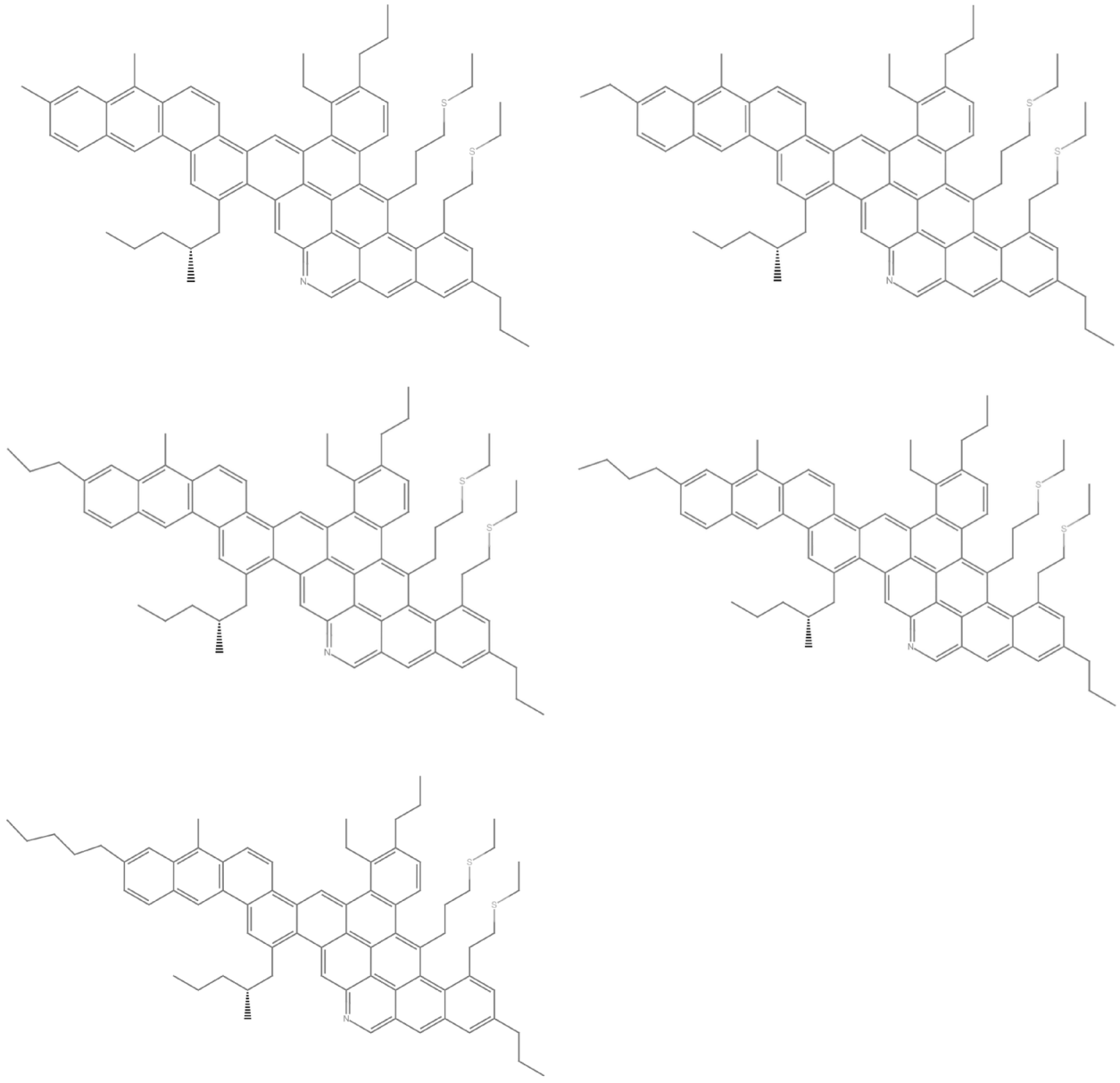

Fig. 4. Example of continental model 2 with augmentations. Each structure was edited using the MoleculeDraw function. From left to right, each structure was edited by adding one additional carbon to the same location as before. Modifications began with the addition of a methyl group with the successive addition of carbons up to a pentane chain.

dataset for our investigation.

Machine learning investigations were implemented using the Wolfram Language [25]and Mathematica notebooks. The results describe classification performance on the original dataset of 255 original asphaltene structures and an augmented dataset composed of 1,530 structures. The augmented dataset contains each structure from the original dataset and the 5 derived structures from each motif. The image-based classification was conducted using the MoleculePlot [27] function to convert molecular structures into visual representations. The line structure images were converted to thumbnail size to limit memory requirements during the process. The asphaltene structures were initially imported into each notebook and labels were applied to the images based on information provided in the asphaltene dataset [24].

The resulting lists of labeled molecules were collated and shuffled to provide a dataset of labeled asphaltene structures. Subsequently, the dataset containing labeled molecules underwent a division into training and testing sets ensuring an 80–20 split, with 80 % of the molecules allocated to the training set and the remaining 20% to the testing set. Since the dataset is unbalanced, it was decided to split the archipelago and continental structures into testing and training sets individually. The training sets were combined to form the aggregate training set and similarly the testing sets were combined to form the aggregate testing set. This ensures that the final testing set contains 20 % of structures from each category and the training set contains 80 % of the structures from each category, archipelago, and continental. The classification was carried out using two different techniques. The first method involved utilizing the topological features of the molecules codified in molecular fingerprints. The fingerprints were computed using the RDKit [16] functionality implemented in Mathematica.

The fingerprint-based classification was performed using the Classify [28] function in Mathematica. This function applies a variety of methods to the training data and returns the best performing model with hyperparameter values using cross-validation on the dataset. The resulting classifier was tasked with classifying the testing set, comprising molecules that were not part of the training dataset. This procedural sequence was iteratively executed for each classification, with feature

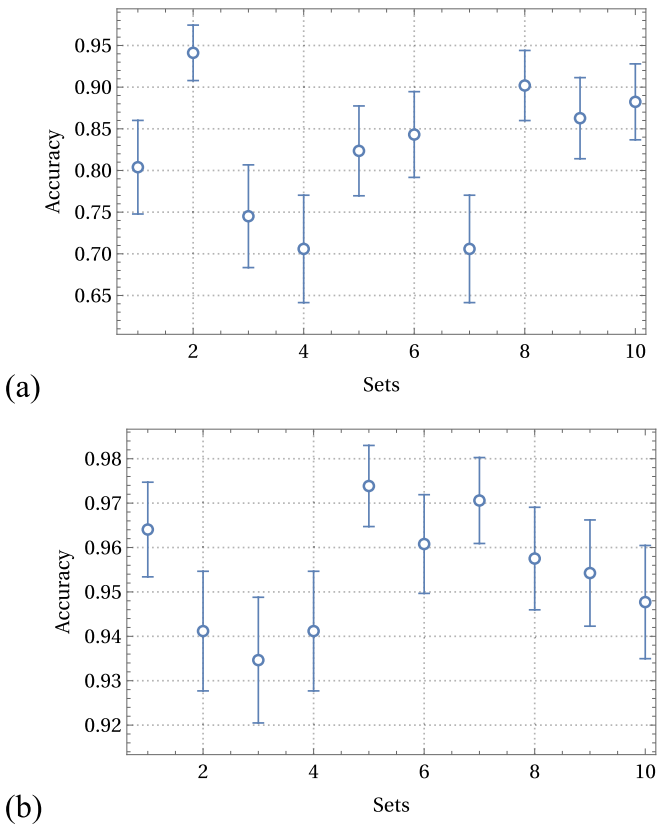

Fig. 5. Accuracy scores for the original figures (a) and composite figures (b) using ResNet-50. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

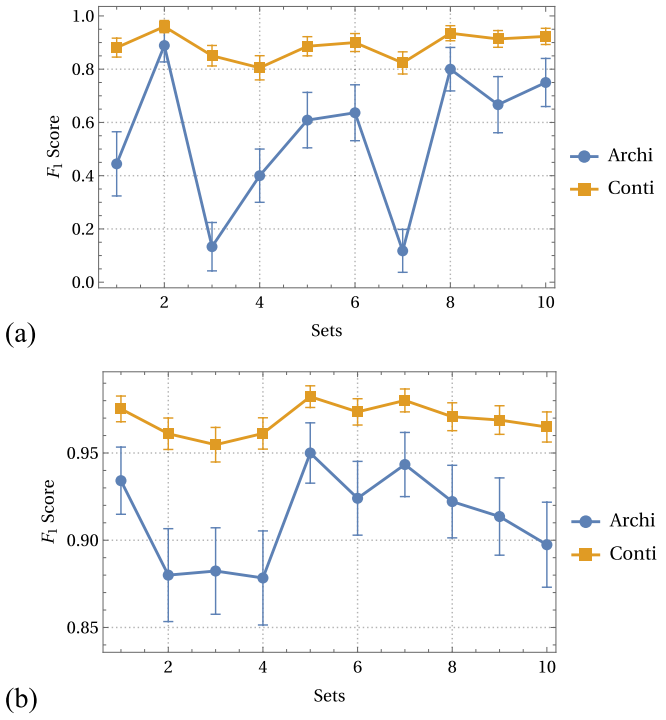

Fig. 6. F_1 scores for the original figures (a) and composite figures (b) using ResNet-50. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

specification incorporated within the Classify function. ClassifierMeasurements [29] function was employed to obtain comprehensive measurements and assessments of the classification results. This function provided an output detailing the efficacy and accuracy of the classifier, offering insights into the performance of the machine learning model across various classifications. The entire classification and testing methodology was carried out 10 times with random sampling from the entire dataset of structures. This classification was carried out for both datasets, the original structure collection, and the composite collection with modified molecules. The Logistic Regression technique with regularization performed the best in our investigations to classify asphaltenes using fingerprint information as the encoding property.

The second method implemented supervised learning to exploit the visual differences in the shapes of the archipelago and continental motifs. Supervised learning, a subset of machine learning methodologies, trains neural networks using labeled data, enabling the network to perform specific tasks on new, unlabeled data with learned patterns. This approach, particularly potent in deep learning architectures, has significantly advanced fields such as computer vision and natural language processing [30]. Image classification, a primary application of deep learning, aims to categorize images into predefined classes and has seen remarkable success. Transfer learning, especially relevant in computer vision, leverages pre-existing models trained on extensive datasets like ImageNet, facilitating rapid and efficient model development without the need to construct neural networks from scratch. These pre-trained models, often sourced from existing literature, are adapted to new problems, demonstrating considerable efficacy in various computer vision tasks, as detailed in comprehensive analyses of ImageNet-based models [9,31]. ResNet-50, the pre-trained neural network used for this study was sourced from the Wolfram Neural Net Repository [32]. The demonstrated versatility and performance of ResNet-50 over larger neural networks and across various domains, including its successful application in laboratory glassware recognition [10], weed identification [11], and the classification of fungi and bacteria [12] motivated its selection for this investigation. This choice was underpinned by ResNet-50′s proven capability in handling diverse image-based classification tasks, making it an apt model for exploring asphaltene categorization.

The training process involved modifying the network by removing the final classification layer and incorporating a new classifier designed to handle the two molecular classes, along with a SoftMax layer for probability computation. These modifications were executed using the NetDrop [33] function, and the subsequent training was conducted through the NetTrain [34] function, aligning the network with the specific requirements of our classification task. The final models, Mathematica coding notebooks, and all data used to carry out the investigations are provided in the Zenodo data repository [35].

Results and discussion

The training of each network resulted in the classifier being able to distinguish between continental and archipelago asphaltenes successfully. The following subsections describe classification metrics to quantify the performance of these models. The original dataset is comprised of 255 structures and the composite dataset includes the original structures and those created by modifying each structure 5 times to create an augmented set of 1,530 structures. Between these extremes of original dataset and the composite dataset, we have also examined the classification efficiency of both techniques at each stage using incremental edits. The overall composition of the dataset and the F1 scores at each stage are shown in Table 1.

Accuracy

Accuracy defines the fraction of the images that were correctly identified from the testing set. Fig. 5 shows a graphical representation of the mean accuracy of the data sets after they were divided into small subsets of data, using ResNet-50. There was an increase in the accuracy

Table 1
Composition of the dataset at each stage of structure editing and the resulting mean F_1 score for archipelago and continental structures. The data in the last 4 columns shows the minimum, maximum, and the mean ± standard deviation values for the classification. The data clearly shows that image-based classification reaches high levels of reliability with much larger datasets compared to the fingerprint-based classification technique.

Dataset	Archipelago structures	Continental structures	F_1 score Archipelago structures image based	F_1 score Continental structures image based	F_1 score Archipelago structures fingerprint based	F_1 score Continental structures fingerprint based
Original	69	186	0.13, 0.92, 0.51 ± 0.330	0.81, 0.97, 0.89 ± 0.06	0.53, 0.80, 0.70 ± 0.09	0.88, 0.94, 0.92 ± 0.02
Original + 1 edit	138	372	0.40, 0.945, 0.80 ± 0.17	0.81, 0.98, 0.93 ± 0.05	0.82, 1.0, 0.96 ± 0.05	0.94, 1.0, 0.98 ± 0.02
Original + 2 edits	207	558	0.39, 0.921, 0.82 ± 0.16	0.86, 0.97, 0.95 ± 0.03	0.92, 1.0, 0.96 ± 0.02	0.97, 1.0, 0.99 ± 0.01
Original + 3 edits	276	744	0.48, 0.923, 0.84 ± 0.135	0.80, 0.97, 0.94 ± 0.05	0.93, 1.0, 0.97 ± 0.03	0.97, 1.0, 0.99 ± 0.01
Original + 4 edits	345	930	0.79, 0.96, 0.87 ± 0.05	0.93, 0.98, 0.96 ± 0.01	0.98, 1.0, 0.99 ± 0.01	0.99, 1.0, 0.99 ± 0.01
Original + 5 edits (Composite dataset)	414	1116	0.88, 0.95, 0.91 ± 0.03	0.96, 0.98, 0.97 ± 0.01	1.0, 1.0, 1.0 ± 0.0	1.0, 1.0, 1.0 ± 0.0

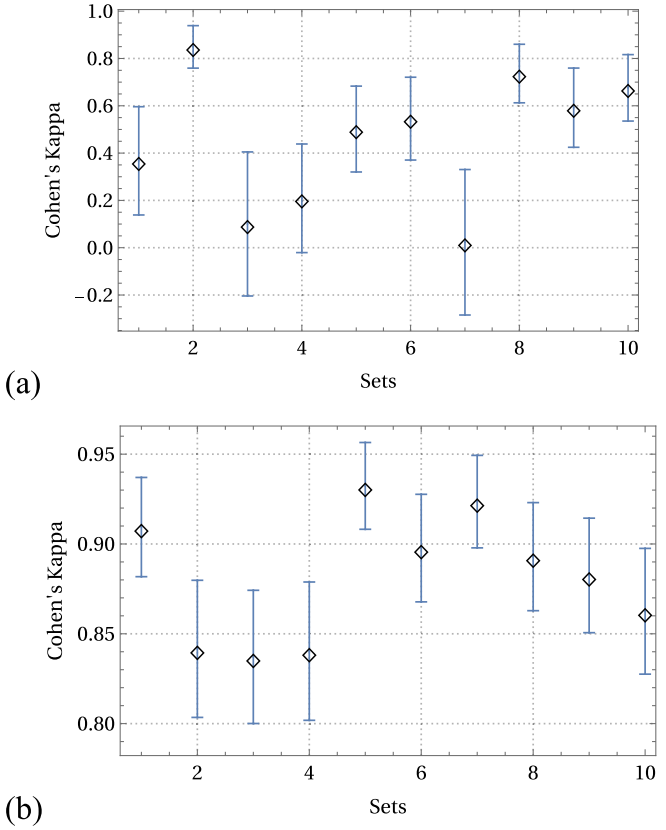

Fig. 7. Cohen's Kappa scores for the original figures (a) and composite figures (b) using ResNet-50. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

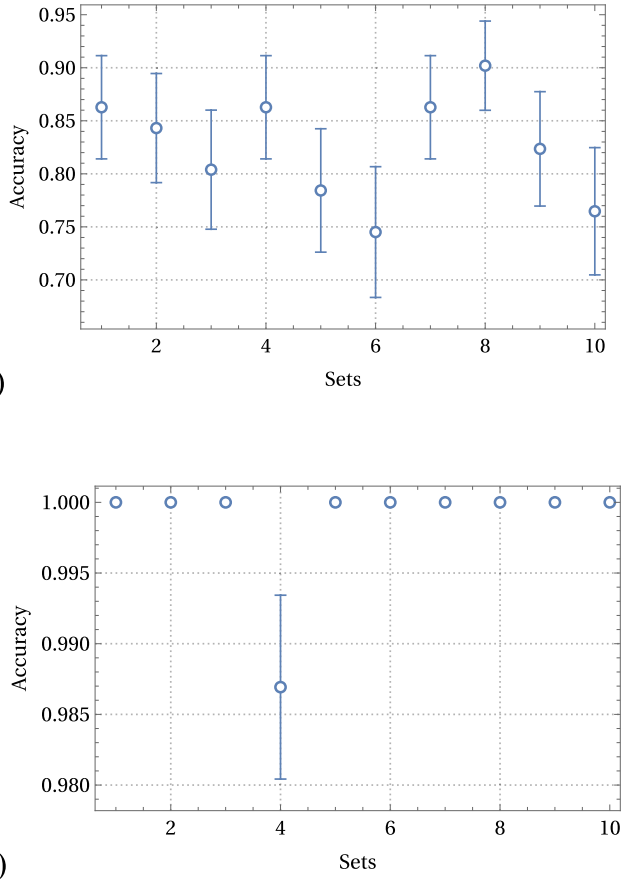

Fig. 8. Accuracy scores for the original figures (a) and composite figures (b) using Topology. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

of our results in the composite figures dataset, primarily due to an increase in sample size being tested. The lowest accuracy for the original figures was 65 % with the highest accuracy being 98 %. (Fig. 5a). The highest accuracy for the composite figures was 92 % and the highest accuracy was 99 %. (Fig. 5b).

We also analyzed the structural classification of the original dataset and the composite dataset through topological information. The accuracy results using the fingerprints are seen in Fig. 8. The lowest accuracy recorded for our original dataset was 68 % and the highest recorded accuracy for this dataset was 94 %. Compared to the composite figures' dataset, our lowest accuracy was 98 % and the highest was 100 % accuracy.

F_1 score

The F_1 score is a robust statistical measure used in binary classification to assess the balance between precision and recall, especially in datasets with uneven class distributions. It is calculated as the harmonic mean of precision and recall, thereby providing a single metric that encapsulates both the false positive and false negative aspects of the classification model. We measured the F_1 score using ResNet-50 and topology encoding on both the original and composite datasets. Fig. 6 (a) shows the F_1 Score for the original dataset using ResNet-50 for both archipelago and continental asphaltene structures. The values for the original dataset regarding the F_1 Score were lower indicating a suboptimal balance between precision and recall. The F_1 Score for the

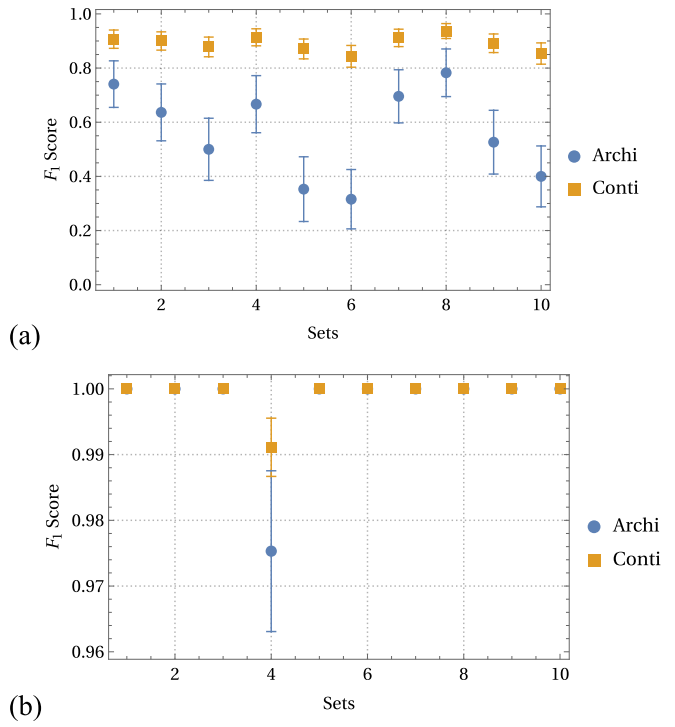

Fig. 9. F_1 scores for the original figures (a) and composite figures (b) using Topology. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

composite dataset shown in Fig. 6(b) using ResNet-50 for both forms of asphaltene structures exhibited higher values indicating robust classification.

The structural classification of archipelago and continental asphaltenes was also accomplished using molecular fingerprints. The corresponding F_1 score results are displayed in Fig. 9, with the original dataset as Fig. 9(a) and the composite dataset as Fig. 9(b). The lowest F_1 Score for the original dataset was 0.0, indicating poor classification efficacy. The lowest F_1 Score for the composite dataset was 0.961, indicating excellent classification efficacy.

Cohen's Kappa

In the context of binary classification of molecules, Cohen's Kappa offers a nuanced measure of agreement, crucial for comparing the efficacy of using molecular fingerprints versus image-based representations. This statistical tool accounts for chance agreement in categorical classification, thereby providing a reliable assessment of the classification methods' accuracy. A value above 0.80 is often associated with a strong agreement between the classification model's predictions and actual values. Cohen's Kappa was measured on the original and composite datasets using ResNet-50 for image classification and topology for structural classification. Fig. 7a exhibits Cohen's Kappa for the original dataset with the lowest value being −0.2, signifying extremely poor classification performance. Fig. 7b exhibits Cohen's Kappa for the composite dataset with the lowest value being 0.80. This indicates that the model's output had a high level of agreement compared to the actual values, when using ResNet-50 for image classification purposes. Fig. 10 (a) shows Cohen's Kappa for the original dataset, with the lowest value being −0.01, indicating a very poor performance in that instance. Fig. 10 (b) shows Cohen's Kappa for the composite dataset, with the lowest value being 0.95, indicating a very strong and consistent level of classification performance.

These indicators already present a clear picture of the success of computer vision-based classification of molecular motifs in the case of

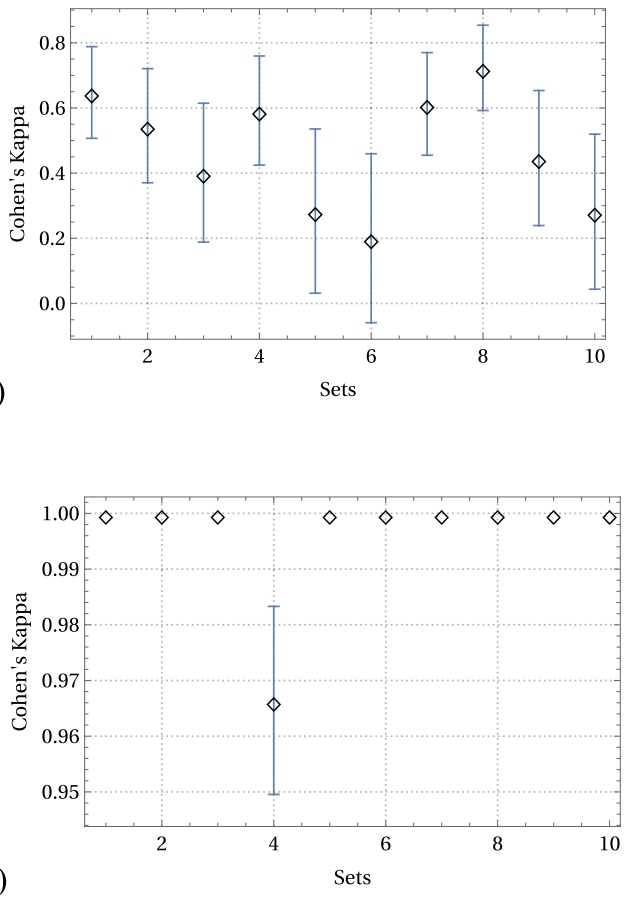

Fig. 10. Cohen's Kappa scores for the original figures (a) and composite figures (b) using Topology. Please note the Y-axis in both figures. In (b) the Y-axis range is much smaller compared to that in (a).

asphaltenes. Even though there is a clear indication of the success of these methods the relative performance of image recognition and fingerprint-based techniques are shown in Fig. 11. An important consideration in all machine learning applications is the computational cost of training the models and here the image-recognition based technique is at a clear disadvantage. Fig. 12 illustrates the total training time required on the same computing hardware for each technique and the much larger computational cost of training the ResNet-50 neural network is evident. However, to the best of our knowledge this is the first documented application of ResNet-50 neural network architecture to identify molecular motifs. Furthermore, we have also analyzed the validation error rate, precision, recall, and geometric mean probability for each classifier. These metrics and their corresponding values for each classifier model are provided in the Supporting Information.

Conclusion

In this study, we developed an extensive dataset comprising 1,530 modified archipelago and continental asphaltene structures, augmenting the original set of 255 structures sourced from literature. This composite dataset, enriched with modifications on all structures, was instrumental in investigating the feasibility of molecular identification through both image and structural classification. Utilizing ResNet-50, we trained a model for image classification of these asphaltenes, while topology-based methods facilitated the structural classification. Our approach utilized the MoleculeDraw code in Wolfram Mathematica for molecular modifications, ensuring diversity in the dataset.

The classification process, conducted using ResNet-50 and topology-based models, was applied to both the original and the augmented

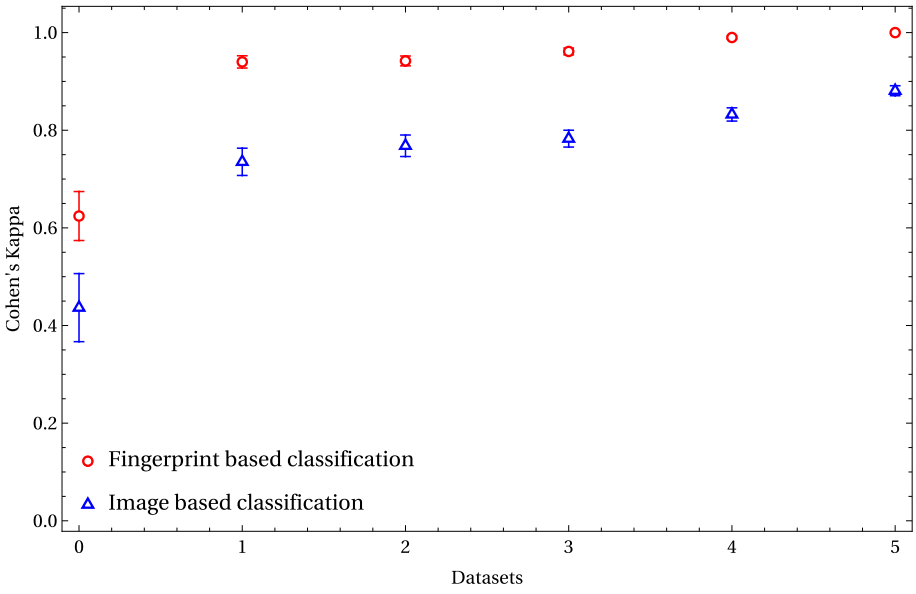

Fig. 11. Cohen's Kappa measurement for each of the 5 edited datasets with each type of classification technique. The fingerprint-based recognition outperforms the image-based recognition of molecules in the datasets examined in this study. However, this figure also makes it clear that the original dataset with 255 structures is inadequate for reliable classification.

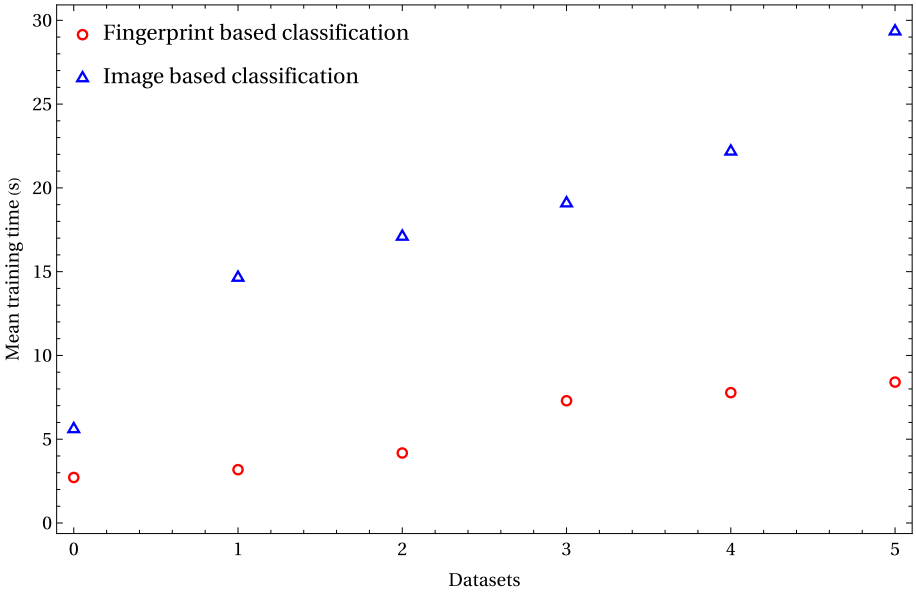

Fig. 12. The mean training time in seconds required for training the classifier at each level of the data augmentation procedure. The mean is computed from ten independent iterations of the classification procedure. The fingerprint-based classification procedure is computationally cheaper relative to the image-based classification technique.

datasets, split into training and testing sets. This methodology allowed for a comprehensive evaluation of the classifiers' performance. Notably, the results were more reliable in the composite dataset, likely due to its larger size and variability. This finding underscores the importance of dataset diversity and size in machine learning applications for chemical analysis, suggesting that the augmented dataset better captures the complexity inherent in asphaltene structures. Our research thus provides valuable insights into the application of machine learning techniques in the classification of complex molecular systems, demonstrating the potential of image-based approaches in enhancing the accuracy and efficiency of such tasks.

This research also provides a novel contribution to the field of asphaltene analysis by demonstrating the unexpected effectiveness of image-based classification using deep learning algorithms. Contrary to the established reliance on topological features, our findings with ResNet-50 on an extensive dataset underscore the feasibility and precision of image-based methods in distinguishing between continental and archipelago asphaltenes. Despite the higher cost of training the image-based classification technique, it presents an interesting opportunity for future applications. This breakthrough suggests a broader applicability and potential for image-based machine learning techniques in complex molecular characterization, opening new avenues for future research in chemical analysis. Future research directions could explore broader applications in chemical informatics, further solidifying the role of image-based machine learning in molecular science.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used GPT4 to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Arun K. Sharma: Conceptualization, Software, Validation, Resources, Project administration, Supervision. **Selsela Arsala:** Formal analysis, Data curation, Methodology, Writing – original draft. **James Brady:** Software, Validation, Data curation, Investigation, Formal analysis. **Madison Franke:** Software, Validation, Data curation, Investigation, Formal analysis. **Shelby Franke:** Writing – original draft, Writing – review & editing, Visualization. **Supreet Gandhok:** Writing – original draft, Writing – review & editing, Visualization. **Simon-Olivier Gingras:** Formal analysis, Data curation, Methodology, Writing – original draft. **Ana Gomez:** Conceptualization, Resources, Writing – original draft. **Katelyn Huie:** Formal analysis, Data curation, Methodology, Writing – original draft. **Kayla Katz:** Conceptualization, Resources, Writing – original draft. **Samantha Kozlo:** Writing – original draft, Writing – review & editing, Visualization. **Mateo Longoria:** Software, Validation, Data curation, Investigation, Formal analysis. **Levi Molnar:** Software, Validation, Data curation, Investigation, Formal analysis. **Nathaly Peña:** Conceptualization, Resources, Writing – original draft. **Sarina Regis:** Formal analysis, Data curation, Methodology, Writing – original draft.

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.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rechem.2024.101551.

References

[1] Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; de Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digital Discovery*. Royal Society of Chemistry August 8, 2023. https://doi.org/10.1039/d3dd00113j.

[2] Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning, Quantum Chemistry, and Chemical Space. **2017**, *30*, 225–256. https://doi.org/10.1002/9781119356059.ch5.

[3] D. Kuntz, A.K. Wilson, Machine Learning, Artificial Intelligence, and Chemistry: How Smart Algorithms Are Reshaping Simulation and the Laboratory, Pure Appl. Chem. 94 (8) (2022) 1019–1054, https://doi.org/10.1515/PAC-2022-0202/ASSET/GRAPHIC/J_PAC-2022-0202_FIG_003.JPG.

[4] L. Turcani, R.L. Greenaway, K.E. Jelfs, Machine Learning for Organic Cage Property Prediction, Chem. Mater. 31 (3) (2019) 714–727, https://doi.org/10.1021/acs.chemmater.8b03572.

[5] Raghunathan, S.; Priyakumar, U. D. Molecular Representations for Machine Learning Applications in Chemistry. *International Journal of Quantum Chemistry*. John Wiley and Sons Inc April 5, 2022. https://doi.org/10.1002/qua.26870.

[6] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. **2015.**

[7] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, 2016; Vol. 2016-Decem, pp 770–778. https://doi.org/10.1109/CVPR.2016.90.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database, Inst. Electr. Electron. Eng. (IEEE) (2010) 248–255, https://doi.org/10.1109/cvpr.2009.5206848.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[10] A.K. Sharma, Laboratory Glassware Identification: Supervised Machine Learning Example for Science Students, J. Comput. Sci. Educ. 12 (1) (2021) 8–15, https://doi.org/10.22369/issn.2153-4136/12/1/2.

[11] Olsen, A.; Konovalov, D. A.; Philippa, B.; Ridd, P.; Wood, J. C.; Johns, J.; Banks, W.; Girgenti, B.; Kenny, O.; Whinney, J.; Calvert, B.; Azghadi, M. R.; White, R. D. DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Scientific Reports 2019 9:1* **2019**, *9* (1), 1–12. https://doi.org/10.1038/s41598-018-38343-3.

[12] Zawadzki, P. Deep Learning Approach to the Classification of Selected Fungi and Bacteria. *Proceedings of 2020 IEEE 21st International Conference on Computational Problems of Electrical Engineering, CPEE 2020* **2020**, 1–4. https://doi.org/10.1109/CPEE50798.2020.9238764.

[13] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, J. Cheminform. 7 (1) (2015) 1–34, https://doi.org/10.1186/S13321-015-0068-4/FIGURES/11.

[14] S. Heller, InChI – the Worldwide Chemical Structure Standard, J. Cheminform. 6 (S1) (2014) 1–9, https://doi.org/10.1186/1758-2946-6-s1-p4.

[15] Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. John Wiley and Sons Inc September 1, 2022. https://doi.org/10.1002/wcms.1603.

[16] *RDKit*. https://www.rdkit.org/ (accessed 2024-01-27).

[17] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Anatole Von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space.

[18] B. Huang, O.A. Von Lilienfeld, Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity, J. Chem. Phys. 145 (16) (2016), https://doi.org/10.1063/1.4964627.

[19] A.H. Alshareef, Asphaltenes: Definition, Properties, and Reactions of Model Compounds, Energy Fuel 34 (1) (2020), https://doi.org/10.1021/acs.energyfuels.9b03291.

[20] A.A. Yushkin, A.V. Balynin, A.I. Nekhaev, A.V. Volkov, Separation of Archipelago- and Continent-Type Asphaltenes on Ultrafiltration Membranes, Membr. Membr. Technol. 3 (2) (2021), https://doi.org/10.1134/S2517751621020098.

[21] O.C. Mullins, H. Sabbah, J. Eyssautier, A.E. Pomerantz, L. Barré, A.B. Andrews, Y. Ruiz-Morales, F. Mostowfi, R. McFarlane, L. Goual, R. Lepkowicz, T. Cooper, J. Orbulescu, R.M. Leblanc, J. Edwards, R.N. Zare, Advances in Asphaltene Science and the Yen-Mullins Model, Energy Fuels 26 (2012) 3986–4003, https://doi.org/10.1021/ef300185p.

[22] O.C. Mullins, The Modified Yen Model, Energy Fuel 24 (4) (2010) 2179–2207, https://doi.org/10.1021/ef900975e.

[23] O.C. Mullins, The Asphaltenes, Annu. Rev. Anal. Chem. 4 (1) (2011) 393–418, https://doi.org/10.1146/annurev-anchem-061010-113849.

[24] M. Franke, S. Arsala, F. Tahiry, S.-O. Gingras, A.K. Sharma, Curated Dataset of Asphaltene Structures, Data Brief 109907 (2023), https://doi.org/10.1016/j.dib.2023.109907.

[25] Wolfram, S. What We've Built Is a Computational Language (and That's Very Important!). *Journal of Computational Science*. Elsevier B.V. October 1, 2020. https://doi.org/10.1016/j.jocs.2020.101132.

[26] Wolfram Research. *MoleculeDraw—Wolfram Language Documentation*. https://reference.wolfram.com/language/ref/MoleculeDraw.html (accessed 2023-05-22).

[27] *MoleculePlot—Wolfram Language Documentation*. https://reference.wolfram.com/language/ref/MoleculePlot (accessed 2024-01-24).

[28] *Classify—Wolfram Language Documentation*. https://reference.wolfram.com/language/ref/Classify (accessed 2024-01-24).

[29] *ClassifierMeasurements—Wolfram Language Documentation*. https://reference.wolfram.com/language/ref/ClassifierMeasurements.html (accessed 2024-01-24).

[30] W. Rawat, Z. Wang, Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, Neural Comput. 29 (9) (2017) 2352–2449, https://doi.org/10.1162/neco_a_00990.

[31] Canziani, A.; Paszke, A.; Culurciello, E. *An Analysis of Deep Neural Network Models for Practical Applications*. http://arxiv.org/abs/1605.07678 (accessed 2020-07-15).

[32] Wolfram-Research. *ResNet-50 - Wolfram Neural Net Repository*. https://resources.wolframcloud.com/NeuralNetRepository/resources/ResNet-50-Trained-on-ImageNet-Competition-Data/ (accessed 2021-08-03).

[33] *NetDrop—Wolfram Language Documentation*. https://reference.wolfram.com/language/ref/NetDrop.html?q=NetDrop (accessed 2024-01-27).
[34] *NetTrain: Train a given neural net—Wolfram Documentation*. https://reference.wolfram.com/language/ref/NetTrain.html?q=NetTrain (accessed 2024-01-27).
[35] *Code and models for asphaltene structural identification*. https://zenodo.org/records/10602597 (accessed 2024-01-30).