energy&fuels

Open Access

pubs.acs.org/EF

Article

Machine Learning for FT-ICR MS Analysis: Predicting Origin, Thermal Evolution, and Biodegradation of Crude Oils

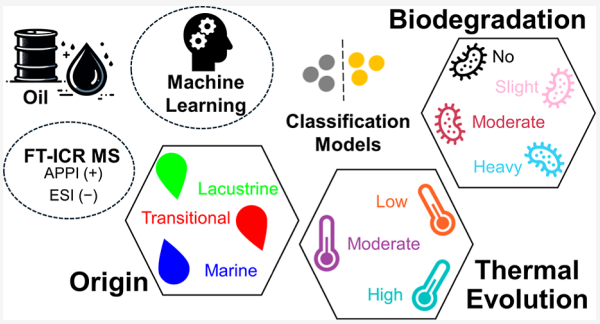
Jussara V. Roque,* Lidya C. da Silva, Danielle M. M. Franco, Gesiane da S. Lima, Mario D. Rangel, Joelma P. Lopes, Ygor dos S. Rocha, Igor V. A. F. de Souza, and Boniek Gontijo*

Cite This: *Energy Fuels* 2026, 40, 13761−13776

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This study provides a detailed assessment of the integration of machine learning with Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) for the characterization of crude oil properties. Utilizing data from both positive-ion atmospheric pressure photoionization (APPI) and negative-ion electrospray ionization (ESI) techniques, this study systematically modeled geochemical characteristics (origin, thermal evolution, and biodegradation) using partial least-squares for discriminant analysis (PLS-DA) and ordered predictor selection for discriminant analysis (OPSDA). The models were initially trained using subsets in which the variable characteristic under investigation was isolated. Under these conditions, very high classification performance was obtained for several heteroatom classes, with some models showing complete class separation within the evaluated data set. This performance illustrates the ability of the models to discern subtle geochemical signatures within the oil samples. The subsequent application of these models to the entire data set confirmed their ability to generalize and maintain high predictive accuracy across a broader spectrum of samples. In the final stages of analysis, interpretative insights were drawn from the most significant variables identified by the models. This discussion delves into the molecular constituents that play pivotal roles in the geochemical differentiation of oils, shedding light on potential markers for oil classification. These findings underscore the effectiveness of APPI (+) and ESI (−) in identifying molecular features and demonstrate the synergy between advanced mass spectrometry techniques and machine learning. Such integrative approaches promise to enhance oil classification processes, offer valuable insights for the petroleum industry, and aid environmental management by providing a robust framework for identifying molecular markers that distinguish different types of crude oils. Overall, this study confirms the robust capabilities of machine learning for high-resolution petroleum molecular characterization and paves the way for future innovations in analytical and data-driven approaches for complex crude oil analysis.



1. INTRODUCTION

Petroleum comprises a complex mixture of hydrocarbons. Its chemical composition varies according to the different associations of organic matter preserved in the source rocks. These source rocks were formed millions of years ago in various geological settings, including lakes (lacustrine) and marine environments.[1]

Understanding the paleodepositional conditions of petroleum is crucial for unraveling its geological history, which in turn aids in identifying potential reservoirs and optimizing extraction.[2,3] In addition, different environments have distinct organic matter compositions, which can affect the quality and quantity of petroleum; therefore, they have implications for exploration and production strategies.[4,5]

Thermal evolution is another factor responsible for the changes in the composition and characteristics of oils. It encompasses hydrocarbon fluids' generation, transformation, and maturation within the sedimentary basins.[6] Therefore, understanding the thermal evolution of reservoir oils can help predict the types of hydrocarbons and their physical properties. This knowledge guides the selection of the extraction and refining methods.[7]

Additionally, petroleum forms and accumulates in reservoir rocks and can undergo biodegradation due to microorganisms' activity. This biodegradation process can significantly alter the composition of petroleum, resulting in the loss of lighter, more valuable hydrocarbons and enrichment of heavier fractions.[8] Therefore, understanding the biodegradation of oils is paramount in petroleum exploration, as it directly impacts hydrocarbon preservation in reservoirs within reservoirs.[9]

Received: January 8, 2026
Revised: May 20, 2026
Accepted: May 29, 2026
Published: June 3, 2026

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

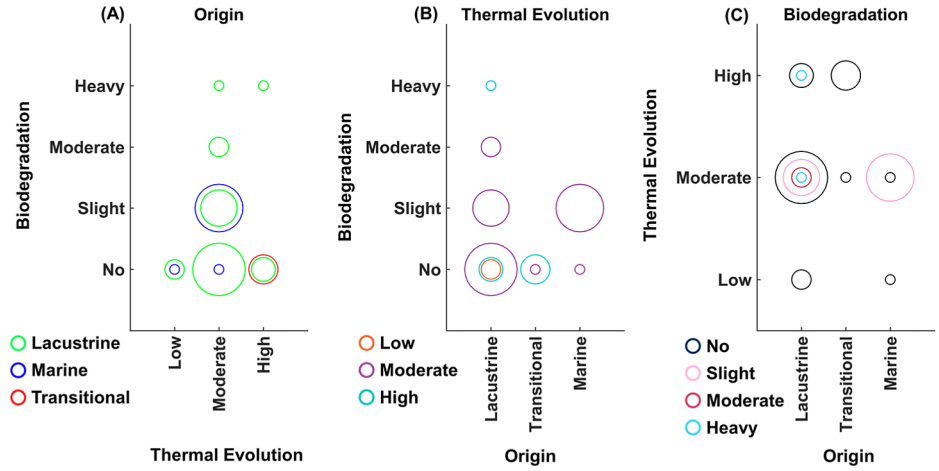


Figure 1. Bubble plots illustrating the origin (A), thermal evolution (B), and biodegradation (C) sample distribution against other petroleum characteristics. Each bubble size represents a relative number of samples with the same properties.

Traditionally, gas chromatography−mass spectrometry (GC−MS) analysis of classical biomarkers has been the industry standard for petroleum geochemical characterization. Biomarker families such as steranes, terpanes, and aromatic steroids are widely used to evaluate depositional environment, source facies, thermal maturity, and biodegradation processes in crude oils. These compounds provide highly diagnostic molecular indicators and have been extensively validated in petroleum geochemistry over several decades.[10,11] However, GC−MS analyses are typically limited to a relatively small subset of target molecules that can be chromatographically separated and detected, and the approach focuses primarily on hydrocarbon biomarkers that often occur at relatively low abundance.

In contrast, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) enables the direct detection of thousands of molecular species in complex petroleum mixtures without prior chromatographic separation. Owing to its ultrahigh resolving power and subppm mass accuracy, FT-ICR MS allows the assignment of molecular formulas for a vast range of heteroatom-containing compounds (e.g., N-, S-, and O-containing species) that are largely inaccessible using conventional GC−MS methods.[12,13] As a result, FT-ICR MS provides a comprehensive molecular fingerprint of petroleum, capturing compositional variations across entire classes of compounds rather than focusing solely on specific biomarker families.

Using complementary ionization techniques such as atmospheric pressure photoionization in positive mode (APPI (+)) and electrospray ionization in negative mode (ESI (−)), FT-ICR MS has been successfully applied to investigate key geochemical processes in crude oils.[14] Researchers have utilized FT-ICR MS to distinguish the paleodepositional origin of oils by analyzing NSO compounds and carbon number distributions.[15,16] Similarly, FT-ICR MS has been used to investigate the effects of thermal maturation and microbial degradation in oils. Studies have demonstrated differences in class distribution, double-bond equivalents (DBE), and carbon numbers concerning maturity levels.[17−19] Similarly, FT-ICR MS analyses have revealed alterations in NSO compounds due to biodegradation effects.[20,21] Thus, FT-ICR MS does not replace traditional biomarker analysis but expands the accessible molecular information, enabling multivariate approaches that exploit the full compositional fingerprint of crude oils.

However, interpreting FT-ICR MS data remains a significant challenge due to the extreme chemical complexity and high dimensionality of petroleomic data sets. In recent years, machine learning approaches have been increasingly applied to mass spectrometry data to extract meaningful patterns, enabling improved characterization, classification, and prediction of petroleum properties.[22−25] As a result of these advancements, machine learning has been employed with mass spectrometry to provide insights into oil composition and address challenges within the oil industry.[26−28]

Recent reviews investigated the current state of the use of machine learning aplications for high-resolution petroleum molecular characterization using high-resolution instruments such as FT-ICR MS, compiling numerous studies across the field.[29,30] These reviews highlights the wide range of approaches employed, spanning from exploratory data analysis[31−33] and classical multivariate techniques[34] to more advanced machine learning strategies,[35] including neural networks[36] and other predictive models.[37,38] In this context, the integration of data science with high-resolution petroleum molecular characterization has also been referred to as petroinformatics,[39] reflecting the growing role of computational approaches in the interpretation of complex petroleum data sets. Despite these advances, the direct application of machine learning models to predict key geochemical characteristics such as origin, thermal evolution, and biodegradation using FT-ICR MS data remains limited.

Based on the successful application of machine learning in petroleum research, this study evaluates the use of supervised classification models combined with APPI (+) and ESI (−) FT-ICR MS data to predict crude oil origin, thermal evolution, and biodegradation. The data set includes oils from different depositional contexts and varying degrees of maturity and biodegradation. This approach aims to relate high-resolution molecular profiles to established geochemical classifications and identify molecular features associated with each property.

2. EXPERIMENTAL SECTION

2.1. Samples

A comprehensive analysis was conducted on a data set of ninety-five oil samples collected from three distinct Brazilian basins. The oil

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

samples were supplied by the Centre of Research, Development, and Innovation Leopoldo Americo Miguez de Mello (CENPES, Petrobras, RJ, Brazil). These samples were chosen to represent a broad spectrum of petroleum characteristics, shedding light on the diverse origins (lacustrine, marine, and transitional), thermal evolution (low, moderate, and high), and biodegradation (no biodegradation, slight, moderate, and heavy) profiles of the Brazilian oil industry. A detailed table containing the geochemical characteristics and biomarker parameters of all crude oil samples is provided in Table S1 (Supporting Information). Representative normalized C8−C40 hydrocarbon distributions for selected oils are additionally shown in Figure S1 (Supporting Information).

Among the sampled basins, one predominantly featured lacustrine-sourced oils, whereas the other mainly yielded transitional oils. One basin stands out for its dual nature of producing both lacustrine and marine oils. This diverse data set encompasses oils with varying degrees of thermal evolution and reflects the different geological histories of the basins.

Lacustrine-origin oils are characterized by moderate thermal evolution, which indicates specific geological settings that influence their formation. These oils typically display slight or no biodegradation, suggesting a preservation environment that minimizes biological transformations.

Marine-origin oils exhibit a more comprehensive range of thermal evolution, from moderate to high, reflecting the varied geological histories encountered in marine depositional environments. Biodegradation levels in these samples varied, with some showing slight to moderate signs, potentially due to greater exposure to biologically active environments.

Transitional-origin oils, derived from environments influenced by both lacustrine and marine depositional processes, show predominantly moderate thermal evolution. This mirrors the intermediate nature of source rock environments. Biodegradation in these samples was generally absent, which may indicate rapid burial and preservation conditions that limit biological activity.

Finally, the subset of samples from the basin with mixed lacustrine and marine oils demonstrates the intricate interplay between these two origins. Here, oils with moderate and high thermal evolutions can be found, suggesting a complex geological history with varying degrees of thermal stress. The biodegradation profiles of these oils are similarly diverse, ranging from nonexistent to moderate, outlining the heterogeneity of reservoir conditions.

A bubble plot representation (Figure 1) was employed to elucidate the relationships between the Brazilian oil samples' origin, thermal evolution, and biodegradation. Figure 1 provides a visual summary of the interrelated characteristics across oil samples. The size of each bubble corresponded to the number of samples exhibiting traits within the data set.

The bubble plots in Figure 1 offer insightful visualizations of the diversity of Brazilian petroleum resources. Figure 1A highlights the distribution of oil origins and their associated biodegradation levels, suggesting that lacustrine environments tend to yield oils with lower biodegradation. Figure 1B shows the varying thermal maturity of oils across different origins, with a notable concentration of moderate thermal evolution in transitional oils. Finally, Figure 1C shows the biodegradation levels across the spectrum of thermal evolution and origin, revealing a predominance of nonbiodegraded oils irrespective of their geological history.

2.2. MS Analyses and Processing

In this study, ninety-five crude oil samples were subjected to APPI (+) and ESI (−) FT-ICR MS analyses. MS analyses were carried out using a 7T Bruker SolariX 2xR FT-ICR MS instrument. The detailed acquisition parameters used for both ionization sources, including instrumental settings and operating conditions, are summarized in Table S2 (Supporting Information). To ensure analytical reproducibility and instrument stability, the performance of the FT-ICR MS system is continuously monitored through a multivariate control chart developed in our laboratory. This control chart is based on repeated acquisitions of a reference crude oil sample under the same analytical conditions used in routine measurements, including the ionization modes applied in this study (APPI (+) and ESI (−)). Prior to each analytical sequence, the reference sample is acquired and evaluated against the statistical limits of the control chart, and the instrument is released for analysis only when the acquisition falls within the acceptable control limits.

Briefly, the control chart is constructed using principal component analysis (PCA) of the reference sample spectra, which is continuously updated as new acquisitions are included. The Hotelling's T^2 and Q-residual statistics[40,41] are then calculated to evaluate the multivariate consistency of each acquisition relative to the established reference distribution. These parameters allow the detection of deviations in instrumental performance and ensure the reproducibility of the petroleomic profiles obtained in this study. The multivariate control charts showing the Hotelling's T^2 and Q-residual limits used to monitor instrument stability are provided in Figure S2 in the Supporting Information.

For ESI (−) analyses, the samples were prepared at 0.5 mg mL^{-1} in toluene/methanol (1:1, *v*/*v*). 50 μL of NH_4OH was added to each sample to improve ionization. Nitrogen was used as drying gas at a flow rate of 1 L min^{-1} and a temperature of 200 °C and as nebulizing gas with 1 bar. The samples were injected at a flow rate of 120 $\mu L.h^{-1}$. The capillary voltage was set to 3.8 kV, and the spray shield voltage was set to −500 V. Ions were accumulated in the collision cell for 0.05 s and transferred to the ICR cell within 0.6−0.8 ms. This time-of-flight corresponds to the ion transfer time from the collision cell to the ICR cell. Spectra were recorded in broadband mode using 8 mega word data sets. 300 scans were accumulated in a mass range from m/z 128 to 2000 for each mass spectrum.

For APPI (+), the samples were prepared in toluene at the same concentrations (0.5 mg mL^{-1}), and the injection flow rate was maintained at 500 $\mu L\ h^{-1}$. Nitrogen was used as a drying gas at 200 °C and a nebulizing gas at 2 bar. The capillary voltage was set to 4 kV. Ions were accumulated in the collision cell for 0.05 s and transferred to the ICR cell within 0.8−1 ms. This time-of-flight corresponds to the ion transfer time from the collision cell to the ICR cell. Spectra were recorded in broadband mode using 8 mega word data sets. 300 scans were accumulated in a mass range from m/z 128 to 2000 for each mass spectrum.

Raw FT-ICR MS spectra in magnitude mode were internally recalibrated using DataAnalysis 5.0 SR1 software (Bruker Daltonik GmbH). During this step, centroiding and peak picking were performed to extract monoisotopic mass signals from the broadband spectra and generate a peak list for each sample. Following recalibration, the spectra were processed using Composer64 software (Sierra Analytics, CA, United States, version 1.5.3) for molecular formula assignment. Composer applies elemental constraints and mass accuracy criteria to assign molecular formulas to the detected peaks. The parameters used for molecular formula assignment are summarized in Table S3 of the Supporting Information.

Instrumental noise and low-intensity signals were controlled during the molecular assignment step by applying intensity thresholds within the Composer software. For APPI (+) spectra, a threshold of 0.6% relative abundance was applied, while for ESI (−) spectra thresholds between 0.1 and 0.3% were used depending on the data set. These thresholds were used to remove low-intensity background signals and instrumental noise while preserving chemically meaningful peaks.

After molecular assignment and filtering, the resulting molecular formulas and their corresponding monoisotopic abundances were exported as composition tables for each sample. These tables constituted the initial data set used for both classical petroleomic characterization and subsequent multivariate statistical analyses.

2.3. Data Organization

2.3.1. Variables. After molecular formula assignment and filtering, the APPI (+) and ESI (−) FT-ICR MS data were exported from Composer64 as composition tables containing the molecular formulas and their corresponding monoisotopic abundances for each sample. These composition tables constituted the initial data set used for subsequent data analysis. The resulting data sets from both ionization

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

sources were then imported into MATLAB 2024a (MathWorks, Natick, USA) for further processing and multivariate statistical analyses.

The imported data included information such as the monoisotopic abundance, molecular formulas, m/z values, and classes of heteroatoms. A data matrix was then built using the monoisotopic abundance values, referred to as the **X** matrix, representing the independent variables. In the **X** matrix, the rows correspond to the individual samples, whereas the columns represent unique molecular formulas.

The APPI (+) and ESI (−) data sets were treated independently throughout the study, as these ionization techniques provide complementary molecular information and selectively ionize different compound families.

Classical petroleomic approaches were applied to characterize the samples' origin, thermal evolution, and biodegradation characteristics.[42,43] During this characterization, the abundances of various classes of heteroatoms were evaluated. To reduce sparsity and improve the robustness of the multivariate models, a filtering step was applied to the heteroatom classes. Classes detected in fewer than 25% of the samples were excluded, except in cases where their occurrence was clearly associated with a specific geochemical condition that could provide discriminant information. In addition, classes presenting relative abundances lower than 1% in more than 90% of the samples were also removed, as these variables likely correspond to extremely low-intensity signals with limited statistical relevance. This criterion was adopted to remove rare or sporadic classes likely associated with analytical noise or extremely low-abundance signals, while preserving classes relevant for geochemical interpretation. After this filtering step, specific data matrices were constructed for each ionization source. These matrices included only the molecular formulas belonging to the selected heteroatom classes and were subsequently used as input for the multivariate analyses. The tailored data matrices generated are summarized in Table S2 (Supporting Information), which delineates the variable subsets used across the different ionization techniques, APPI (+) and ESI (−), highlighting the specificity of the molecular classes considered in each case. In total, the variable organization process yielded five distinct subsets for the APPI (+) ionization technique and four subsets for the ESI (−) technique, each designed to analyze the variables crucial for understanding the complex molecular composition of the oil samples.

2.3.2. Samples Geochemical Classification. The analysis of the 95 oil samples illustrates the diversity of the three selected sedimentary basins, with each contributing a distinct geochemical signature to the data set. Supporting geochemical evidence based on conventional GC−MS biomarker parameters is provided in Table S1, including commonly used origin, thermal evolution and biodegradation indicators that help contextualize the classifications adopted for the samples. Complementary representative C8−C40 hydrocarbon distributions, including pristane and phytane profiles, are shown in Figure S1.

The samples were grouped according to their geological origin: 59 lacustrine, 10 marine, and 26 transitional. Regarding thermal evolution, five samples were classified as low, indicating less mature oil. Most of the 74 samples exhibited moderate thermal evolution, suggesting a more advanced maturation stage. Meanwhile, 16 samples were marked by high thermal evolution, indicating significant geological aging and transformation.

In the context of biodegradation, these data revealed a wide spectrum of alterations. Fifty-one samples exhibited no signs of biodegradation, indicating they were likely unaffected by microbial activity. Thirty-eight samples showed slight biodegradation, indicating a minimal biological impact. Four samples demonstrated moderate biodegradation, and two showed heavy biodegradation, indicating extensive microbial degradation that can markedly alter the properties of the oil.

To evaluate each geochemical characteristic independently, sample subsets were created by varying one property at a time while keeping the others as constant as possible. This strategy reduced confounding effects among origin, thermal evolution, and biodegradation during model training. The subsets used for modeling are summarized in Table 1.

Table 1. Sample Subsets for Independent Geochemical Characteristic Modeling[a]

geochemical characteristic	origin	thermal evolution	biodegradation
origin	lacustrine (14) transitional (24)	moderate	slight
	lacustrine (6) marine (9)	high	no
thermal Evolution	lacustrine	low (4) moderate (29) high (6)	no
biodegradation	lacustrine	moderate	no (29) slight (14) moderate (4) high (1)

[a]The values in parentheses are the number of oil samples in each class.

In the pursuit of independently modeling each geochemical characteristic, we encountered certain constraints that necessitated the development of two separate models for the characteristics of origin. The inherent variability within our data set did not permit the fixation of other characteristics while simultaneously representing all three classes of origin (lacustrine, marine, and transitional) in a single model. This limitation led to the development of two distinct models of origin. The first model encompasses 14 lacustrine and 24 transitional samples, exhibiting moderate thermal evolution and slight biodegradation. This model allowed us to investigate the nuances between lacustrine and transitional origins under controlled thermal maturity and biodegradation conditions. The second model contrasts six lacustrine with nine marine samples, characterized by high thermal evolution and no biodegradation. This comparison is instrumental in distinguishing the geochemical signatures of lacustrine versus marine origins when subjected to a higher degree of thermal stress without the confounding influence of biodegradation.

The samples were successfully segregated for the thermal evolution and biodegradation characteristics to reflect the full spectrum of low, moderate, and high thermal evolution and from no biodegradation to moderate biodegradation, respectively. However, it is noteworthy that the high biodegradation category could not be modeled because of insufficient samples; only one sample fell into this category, rendering it statistically insignificant for independent modeling. This approach underscores the complexities and limitations inherent in geochemical modeling, mainly when dealing with data sets that exhibit extensive natural variability.

By scrutinizing the data in this manner, we aimed to delineate each geochemical parameter's contribution to the oils' complex composition. The following sections detail the statistical methodologies employed to dissect these influences and the resultant findings, which provide fresh insights into the geochemical differentiation of petroleum sources.

2.4. Multivariate Modeling

Combining each subset of samples with a corresponding subset of variables resulted in a comprehensive suite of 48 unique data sets, each poised for a detailed analysis. From the subsets of data formed based on origin, thermal evolution, and biodegradation, for both the APPI (+) and ESI (−) ionization sources, classification models were built using partial least-squares for discriminant analysis (PLS-DA)[44] combined with a variable selection method called ordered predictors selection for discriminant analysis (OPSDA).[45]

PLS-DA is a supervised pattern recognition method that employs the independent variable matrix **X** and dependent variable vector **Y** (representing sample classes) to create classification models.

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

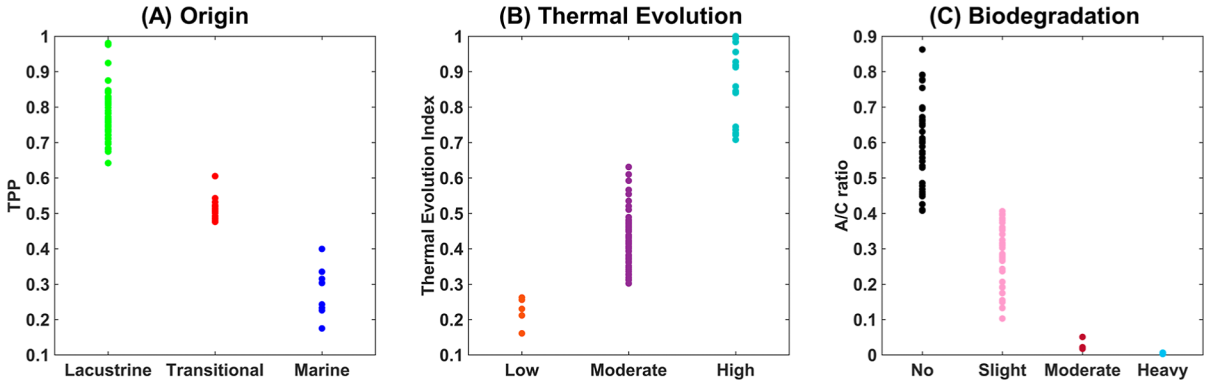


Figure 2. Geochemical classification of oil samples. (A) Distribution of samples by origin using tetracyclic polyprenoid (TPP) biomarker ratios. (B) Thermal evolution of the samples using an index adapted from Noah et al., calculated from relative abundance ratios of hydrocarbon classes detected by APPI (+) FT-ICR MS. (C) Biodegradation levels evaluated using the acyclic/cyclic (A/C) ratio, calculated from the relative abundance of acyclic and cyclic O2 class carboxylic acids detected by ESI (−) FT-ICR MS.

Constructing an effective PLS-DA model involves deriving a binary matrix **Y** with columns matching the model's class count. Each column assigns a value of 1 to rows representing a specific class and 0 to rows not belonging to that class.[44] On the other hand, OPSDA is a variable selection method based on the construction of an informative vector that ranks variables according to their relevance for classification.[45,46] In this study, the variable importance on projection (VIP) scores were used as the informative vector, allowing the variables to be ordered from most to least relevant. This approach provides insight into the relative importance of molecular features and facilitates the identification of key molecular formulas responsible for class separation.

For the construction of PLS-DA models, random 10-fold cross-validation was employed. In this approach, the data set was randomly divided into ten subsets, where approximately 90% of the samples were used for model calibration and the remaining 10% were used for validation in each iteration. This procedure was repeated until all samples were used once for validation. Class assignment in the PLS-DA models was performed using a threshold defined from a normal probability density function.[47] The predicted values obtained for each class were compared to this threshold, where samples with predicted values above the threshold were assigned to the class, and those below the threshold were considered not to belong to that class.

Variable selection was performed using OPSDA, applying a window of 10 variables with increments of 5 and a convergence criterion of 2%. The variables were ranked according to their VIP scores, and classification models were iteratively built by progressively including variables following this ranking. Model performance was evaluated at each step, with the selection of variables guided by the minimization of misclassification during cross-validation. Importantly, the OPSDA procedure was performed within each cross-validation iteration, ensuring that variable selection was conducted exclusively using the training data of each fold and thereby preventing information leakage during model evaluation. This strategy avoids arbitrary variable selection thresholds and ensures that the selected variables are directly linked to predictive performance.

Different normalization strategies were tested on matrix **X** to evaluate their influence on the classification models. Model performance was assessed using accuracy, class-wise sensitivity, and balanced accuracy. The balanced accuracy metric was used to account for potential class imbalance, as it represents the average sensitivity across all classes. In addition, misclassification rates were evaluated to provide a direct measure of prediction errors.

After model optimization, the selected models were applied to the complete data set to evaluate their classification behavior beyond the controlled training subsets.

Chemometric modeling and variable selection procedures were implemented in MATLAB. The algorithms used for model optimization and variable selection follow previously published methodologies, including the OPS strategy and its extension for discriminant analysis. Detailed descriptions of these algorithms and their MATLAB implementations are available in the literature,[45,48] which provides the computational framework adopted in this study.

3. RESULTS AND DISCUSSION

3.1. Samples Classification

The classified oil data set enabled the investigation of how depositional origin, thermal evolution, and biodegradation influence the molecular composition of crude oils based on FT-ICR MS data and machine learning analysis.

The classifications provided by Petrobras were supported by conventional biomarker parameters summarized in Table S1. Depositional origin was validated through tetracyclic polyprenoid (TPP) distributions,[10] while representative C8−C40 hydrocarbon distributions for selected oils are presented in Figure S1, highlighting differences in *n*-alkane profiles and relative pristane and phytane abundances among the oil groups.

Thermal evolution was additionally evaluated using an index adapted from Noah et al. (2020) based on APPI (+) data.[49] The index was calculated from hydrocarbon compositional trends associated with increasing aromaticity and molecular condensation during thermal maturation and interpreted alongside established maturity indicators, such as the Ts/(Ts + Tm) ratio, defined as the ratio between 18α(H)-22,29,30-trisnorhopane (Ts) and 17α(H)-22,29,30-trisnorhopane (Tm), which is widely used as a thermal maturity proxy. Biodegradation trends were assessed using the acyclic/cyclic ratio (A/C ratio) of O2 class carboxylic acids measured by ESI (−), following correlations previously reported in the literature and consistent with the Peter & Moldowan biodegradation scale.[50]

Figure 2 presents a comprehensive overview of these classifications and validations, showing the robust methodology employed to affirm the sample classifications provided by Petrobras and further elucidate the complex characteristics of these oil samples.

As shown in Figure 2A, the biomarker analysis revealed a pronounced distinction in TPP ratios, with lacustrine crude oils exhibiting higher TPP values than their marine counterparts. This observation aligns with the existing literature, which asserts that lacustrine environments are conducive to high TPP ratios.[10]

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

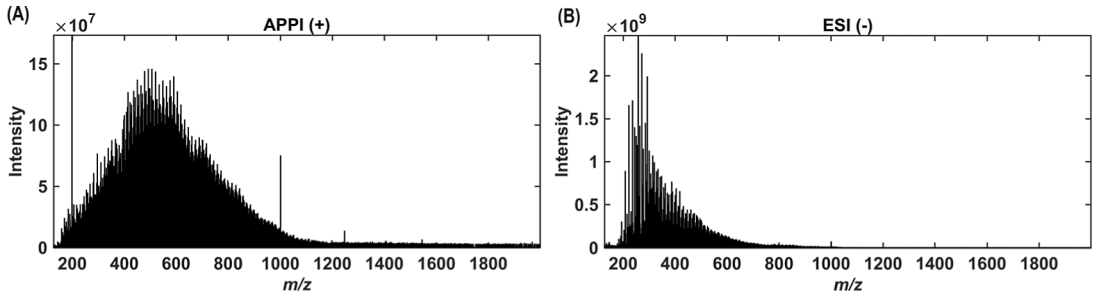


Figure 3. Mass spectra of a representative oil sample obtained by APPI (+) FT-ICR MS (A) and ESI (−) FT-ICR MS (B).

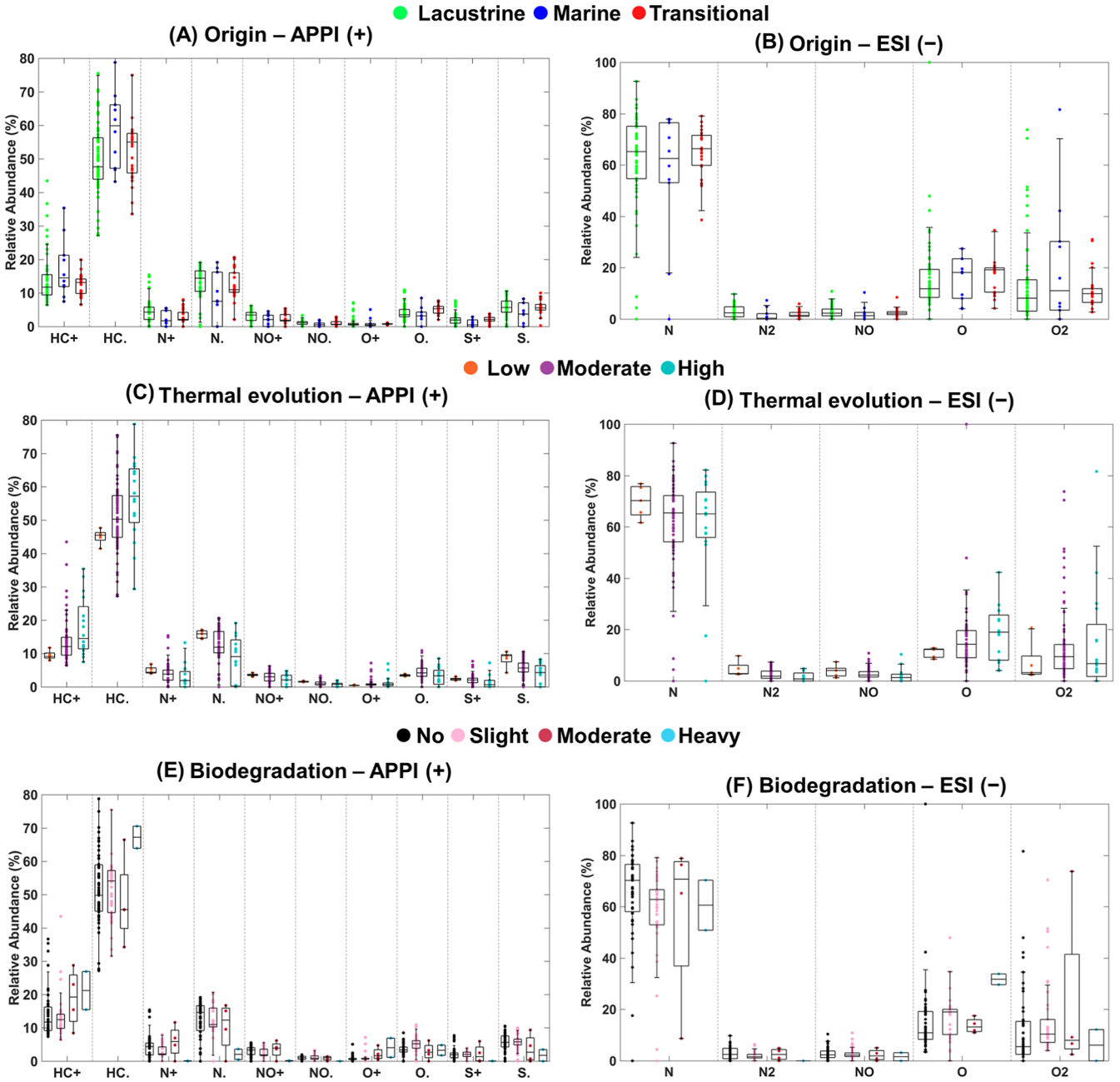


Figure 4. Class distribution box plots for geochemical characteristics. (A, C and E) Relative abundance of compound classes in samples analyzed by APPI (+), categorized by origin, thermal evolution, and biodegradation. (B, D and F) Relative abundance of ESI (−) ionization, classified by origin, thermal evolution, and biodegradation.

As shown in Figure 2B, the thermal evolution index showed a distinct correlation with the thermal evolution of the samples. This index, adapted from Noah's et al.[49] was calculated from the relative abundance ratios of hydrocarbon classes detected

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

Table 2. Accuracy of Classification Models for Origin, Thermal Evolution, and Biodegradation Obtained by PLS-DA and OPSDA for Each Dataset Obtained by APPI (+) and ESI (−) FT-ICR MS Analyses[a]

	variable subsets					
samples subsets	$X_{APPI\text{-}class}$	$X_{APPI\text{-}HC}$	$X_{APPI\text{-}N}$	$X_{APPI\text{-}NO}$	$X_{APPI\text{-}O}$	$X_{APPI\text{-}S}$
origin 1[c]	100%[b]	100%	87%	67%*	67%*	73%*
origin 2[d]	100%*	100%	100%	100%	100%	100%
thermal Evolution	97%*	93%	88%*	100%*	100%*	100%*
biodegradation	100%*	97%	100%	99%	100%	100%
samples subsets	$X_{ESI\text{-}class}$	$X_{ESI\text{-}N}$	$X_{ESI\text{-}N2}$	$X_{ESI\text{-}NO}$	$X_{ESI\text{-}O}$	$X_{ESI\text{-}O2}$
origin 1[c]	100%*	100%	NM	67%	100%	87%
origin 2[d]	100%*	100%	100%*	100%	100%	100%
thermal evolution	100%*	97%*	97%*	85%*	98%*	85%*
biodegradation	100%*	100%	100%	100%	100%	100%*

[a]NM: Not modeled. [b]Results achieved only with PLS-DA. [c]Subset about lacustrine and marine separation. [d]Subset about lacustrine and transitional separation.

by APPI (+) FT-ICR MS, using compositional trends associated with increasing aromaticity and molecular condensation during thermal maturation. Higher index values indicate higher levels of thermal evolution.

Figure 2C illustrates a discernible downward trend in the acyclic/cyclic (A/C) ratio as the biodegradation intensity increased. The A/C ratio was calculated as the relative abundance of acyclic O2 compounds divided by the abundance of cyclic O2 compounds detected by ESI (−) FT-ICR MS. This trend is underpinned by findings from a key study that investigated the effects of petroleum biodegradation on polar acidic compounds, particularly those containing two oxygen atoms.[50] The outcomes of this study led to the proposition of a biodegradation scale that utilizes acidic compounds as reliable indicators. Our analysis showed that as biodegradation progressed, the proportion of acyclic compounds diminished relative to cyclic compounds, which is consistent with the established scale and supports its applicability.

Based on the observations presented in Figure 2, it is important to emphasize that the geochemical classification adopted in this study does not rely solely on a single diagnostic parameter but is supported by multiple geochemical indicators derived from conventional biomarker analysis. The classifications of origin, thermal evolution, and biodegradation were initially provided by Petrobras and were supported by the evaluated biomarker ratios. As commonly discussed in petroleum geochemistry,[5,10] such classifications do not always present strict boundaries, since geochemical processes such as mixing, biodegradation, and varying maturity levels can produce transitional molecular signatures. Consequently, a certain degree of classification uncertainty is inherent to these systems.

In this context, the role of machine learning is not to redefine the geochemical classes but to evaluate whether the complex molecular fingerprints obtained by FT-ICR MS can reproduce these established geochemical classifications. By integrating thousands of molecular features simultaneously, machine learning models can capture subtle compositional patterns that are difficult to evaluate using a limited number of traditional biomarkers. Therefore, the machine learning metrics obtained in this work should be interpreted as a measure of how consistently the ultrahigh resolution petroleomic data reflects the geochemical classifications established through conventional petroleum geochemistry approaches.

3.2. Petroleomic Characterization

The FT-ICR MS analysis generated complex spectra containing thousands of detectable peaks for each crude oil sample. Following internal recalibration and molecular formula assignment, an average of 7105 molecular formulas per sample was obtained for the APPI (+) data sets, whereas the ESI (−) data sets yielded an average of 2986 assigned formulas per sample. The mass accuracy obtained for the assigned formulas remained within the expected range for ultrahigh-resolution FT-ICR MS measurements, with root-mean-square (RMS) mass errors of 0.151 ppm for APPI (+) and 0.110 ppm for ESI (−). A detailed summary of the number of assigned peaks and the corresponding mass error statistics for each sample is provided in Table S4 (Supporting Information).

In our petroleomic approach exploration, spectral analysis serves as a pivotal window in the molecular composition of oil samples. Initially, we presented representative spectra obtained through APPI and ESI (Figure 3), which are key tools for unraveling the chemical complexity of these samples. Despite the diversity in classes on origin, thermal evolution, and biodegradation, the ESI and APPI spectra exhibited remarkable uniformity, reflecting the inherently complex and multifaceted nature of petroleum.

For each ionization source, unique molecular formulas were obtained for the 95 oil samples, resulting in 25,027 formulas for APPI (+) and 12,830 formulas for ESI (−). Regarding a more detailed analysis of compound class abundances, Figure S3 (Supporting Information) offers a comprehensive view of these classes, where we have meticulously selected only those that are consistent across all samples (appears in a minimum of 25% of samples) for further investigation (HC, N, NO, O, and S, both protonated and radical for APPI (+), and N, N2, NO, O, and O2, for ESI (−)). Building on this refined data set, we used box plots (Figure 4) to discern the variable classes associated with each geochemical characteristic. This method, departing from traditional visualization techniques, not only accentuates subtle distinctions between classes but also facilitates a straightforward and intuitive comparative analysis. The resulting box plots reveal patterns and variances that extend our understanding of how geochemical attributes influence the molecular landscape of petroleum.

Through this analysis, we began to discern significant variations among the samples despite their initial spectral similarity. These differences in variable classes enable us to delve deeper into understanding how the origin, thermal evolution, and biodegradation processes shape the molecular

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

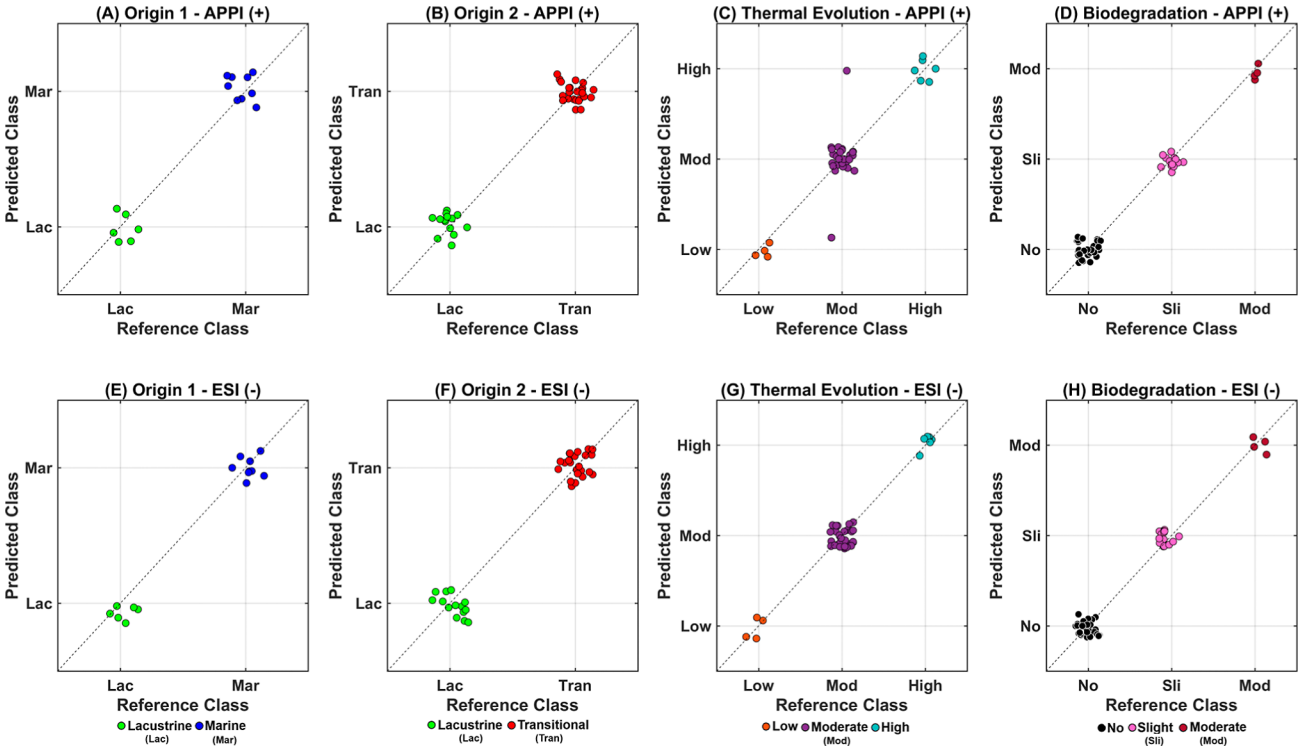


Figure 5. Classification performance of PLS-DA and OPSDA models for geochemical characterization of oil samples using APPI (+) and ESI (−) FT-ICR MS data sets containing all heteroatom classes. (A–D) correspond to models obtained from APPI (+) data, (E–H) correspond to ESI (−) data. Points located along the diagonal indicate correct classifications, while off-diagonal points correspond to misclassified samples.

signature of crude oil samples, expanding the geochemical interpretation of FT-ICR MS data and providing new insights into petroleum molecular characterization.

Moreover, it is crucial to highlight that the selected classes were subsequently utilized to build machine learning models (Table S5, Supporting Information) to identify informative variables for each geochemical characteristic. For the APPI (+) data, radical and protonated species were treated as independent molecular formulas in the data set. However, they were not further separated into distinct radical and protonated classes during the modeling step.

3.3. Modeling

In our analytical exploration, the machine learning models constructed using PLS-DA and OPSDA methods stand at the forefront of our geochemical characterization efforts. These models were meticulously calibrated to unravel complex petroleomic data, providing a robust classification of the samples based on their geochemical properties.

The results of our modeling efforts are summarized in Table 2, which presents the classification performance of the PLS-DA and OPSDA models for the different geochemical properties evaluated in this study, including origin, thermal evolution, and biodegradation. The accuracy values marked with an asterisk (*) correspond to models constructed using PLS-DA without the variable selection step, while the remaining results correspond to models built using the OPSDA variable selection procedure. Additionally, the notation NM (not modeled) indicates cases where model construction was not feasible due to the sparsity of the data set, which resulted in highly sparse matrices. To provide a more comprehensive assessment of model performance, additional metrics including class-wise sensitivity, balanced accuracy, and balanced error rate for both calibration and cross-validation are reported in the Supporting Information (Tables S6–S8), corresponding to the models developed for origin, thermal evolution, and biodegradation, respectively. These metrics allow the evaluation of predictive performance while accounting for potential class imbalance in the data sets.

Table 2 summarizes the classification accuracies obtained for origin, thermal evolution, and biodegradation using different APPI (+) and ESI (−) variable subsets. Figure 5 shows the corresponding predictions for the data sets containing all selected heteroatom classes ($\mathbf{X}_{\text{APPI-class}}$ and $\mathbf{X}_{\text{ESI-class}}$), while the results for the remaining variable subsets are provided in Figures S3–S6 (Supporting Information).

Ideally, for a model exhibiting perfect classification accuracy, all samples should align precisely on the corresponding dashed lines designated for each class on the *y*-axis, indicating a flawless match between the predicted and actual class labels. The figure highlights instances of misclassification where the samples do not align with the expected dashed lines.

The graphical results presented in Figure 5, together with the accuracy values summarized in Table 2, provide a complementary view of the classification performance of the PLS-DA and OPSDA models. While Table 2 quantitatively reports the overall accuracy obtained for each data set, Figure 5 visually illustrates the agreement between predicted and reference class memberships for the data sets containing all heteroatom classes. Additional performance metrics, reported in the Supporting Information (Tables S6–S8) complement the results presented in the main manuscript and allow a more detailed evaluation of model performance. Overall, the consistency observed between calibration and cross-validation results indicates stable predictive behavior and suggests a low risk of model overfitting.

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

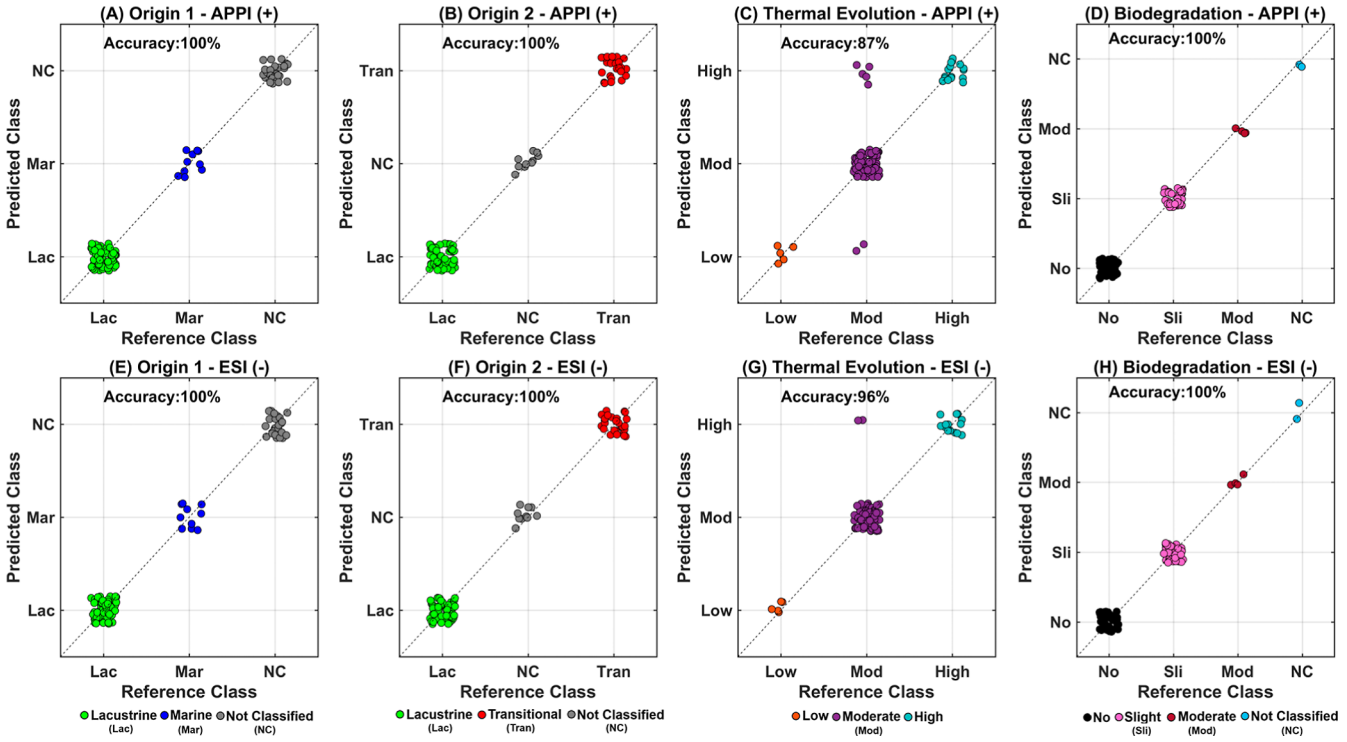


Figure 6. Classification performance of PLS-DA and OPSDA models for the remaining samples within the complete data set of 95 crude oil samples using APPI (+) and ESI (−) FT-ICR MS data with all variable classes. (A−D) correspond to models obtained from APPI (+) data, (E−H) correspond to ESI (−) data. Points located along the diagonal indicate correct classifications, while off-diagonal points correspond to misclassified samples.

Among the geochemical properties evaluated, the classification of oil origin provides an important starting point for assessing the ability of the petroleomic data sets to capture depositional signatures. For the Origin 1 subset (lacustrine and marine), both APPI (+) and ESI (−) models showed some limitations in classification performance, as evidenced accuracy values reported in Table 2. The misclassification patterns for this subset can also be observed in Figure S4 (Supporting Information). These results suggest that distinguishing between these two origins may be more challenging due to overlapping molecular signatures. In contrast, for the Origin 2 subset (lacustrine and transitional), the models exhibited very high classification performance across both ionization methods, with clear separation between classes observed in Figure 5B,F. The alignment of samples along the diagonal indicates consistent classification, highlighting the ability of the models to capture distinct molecular features associated with these origins. A similar behavior is observed in Figure S5 (Supporting Information), where all samples are correctly classified for this subset.

The thermal evolution of the samples was predicted with high accuracy using both ionization methods, with APPI slightly outperforming ESI, as seen in the thermal evolution models and depicted in Figure 5C,G. The detailed classification results for the thermal evolution models across all variable subsets are presented in Figure S6 (Supporting Information). APPI's high accuracy of APPI across variable classes underscores its sensitivity to molecular changes associated with thermal evolution, which is further confirmed by the near-perfect clustering along the dashed lines in the figure.

In the context of biodegradation, the ESI models stood out, achieving a uniform 100% accuracy across all variable classes. This suggests that ESI ionization is particularly attuned to oxygen-rich compounds that are indicative of biodegradation processes, a conclusion supported by the tight groupings observed in Figure 5H. The detailed classification results for the biodegradation models across all variable subsets are presented in Figure S7 (Supporting Information). The APPI models also displayed strong performance, indicating their utility in detecting biodegradation effects, although slightly less consistent than ESI, as shown in Figure 5D.

Overall, the models showed high classification accuracy for the evaluated geochemical characteristics. Complementary behavior was observed between the ionization sources, with APPI (+) contributing more significantly to thermal evolution models and ESI (−) showing stronger performance for biodegradation-related classifications. The selected models were subsequently applied to the complete set of 95 samples, and the resulting classification behavior is presented in Figure 6. In this expanded analysis, the performance of each model was gauged against the entire data set to ascertain how well the models could predict the geochemical characteristics of samples that were not part of the original modeling process.

In the previous section, samples selected for training the models varied only in the characteristics of interest, such as origin, thermal evolution, or biodegradation, which was intended to create a focused and controlled training environment. By isolating each variable, the models could learn to identify unique geochemical signatures associated with each characteristic without the confounding influence of others.

This methodological choice is strategic for model training because it ensures that the model's predictive capabilities

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

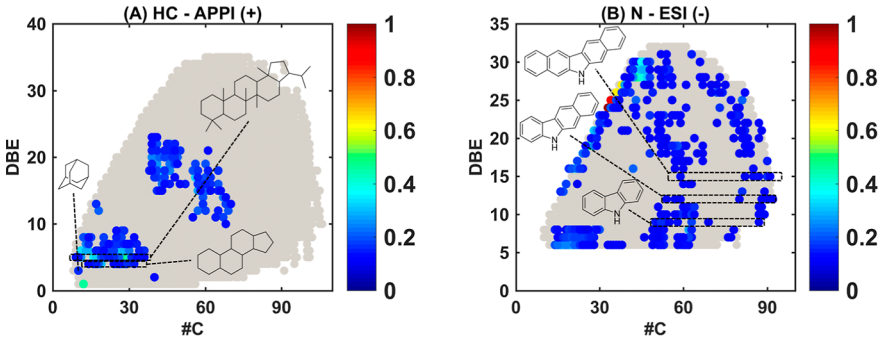


Figure 7. Carbon number vs double-bond equivalent (DBE) for the HC class in APPI (+) and N class in ESI (−). All variables (●) are shown, along with the most important variables for origin classification, and are indicated by importance, ranging from red to blue (● → ●). The chemical structures illustrate the respective classes and represent the DBE values.

independently recognize each variable's nuances. For instance, samples with consistent thermal evolution and biodegradation levels were used when training the model to predict origin. This targeted approach was designed to minimize noise in the data and allow the model to capture the most relevant patterns for classification.

When these models are applied to the entire data set, where all characteristics vary simultaneously, the situation is much more reflective of real-world scenarios, where samples are not controlled for one variable at a time. The nearly 100% accuracy for origin and biodegradation across the APPI and ESI techniques suggests that the models effectively learned to generalize from the controlled conditions they were trained under to the more complex, variable conditions in the full data set. This indicates a well-trained model that is robust against overfitting the training data and can handle the inherent variability of real samples.

The slight decrease in the accuracy of thermal evolution, while still high, may reflect the increased complexity of predicting this characteristic in a data set with multiple varying factors. Thermal evolution is likely a more nuanced geochemical feature to capture, possibly because it is a continuous process rather than a discrete classification, such as origin or biodegradation. The high but imperfect accuracy indicates that there might be more overlapping geochemical signatures between different stages of thermal evolution, making it a more challenging task for the model.

These results underscore the value of the modeling approach. By first ensuring that models can successfully predict characteristics when other variables are controlled, strong classifiers that retain their predictive power are created, even when faced with the full complexity of the data. The success of the models in accurately predicting the geochemical characteristics of the entire data set demonstrates their potential utility in practical petroleomic applications such as oil exploration and analysis. It also opens the door for further refinement, especially in the thermal evolution model, which could potentially be improved with additional training data or more sophisticated modeling techniques that can better handle the complexities associated with this characteristic.

3.4. Interpretative Insights

The selected variables from the comprehensive machine learning models, trained in a full suite of variable classes, offer interpretative insights into the molecular factors that characterize the geochemical properties of oil samples. This part of the analysis examines the significant predictors identified by the models, which are crucial to understanding the intricate geochemical signatures of petroleum.

Variables that stand out in their predictive capacity can illuminate the underlying processes of oil formation, thermal evolution, and biodegradation. Their examination contributes not only to predictive modeling based on FT-ICR MS molecular data but also to broader applications in petroleum geology and reservoir management, offering a nuanced perspective on the molecular intricacies of oil samples. To further explore the distribution of these relevant molecular features, DBE versus carbon number plots were used to visualize their location within the chemical space of each heteroatom class. In these representations, the full set of detected molecular formulas is shown as a background distribution, while the variables selected by the OPSDA procedure are highlighted. The color scale reflects the relative importance of the selected variables, as determined by their ranking in the informative vector, allowing the identification of the most relevant molecular features associated with class discrimination.

Figure 7 provides a focused view of the key variables for the hydrocarbon class assessed by APPI (+) and nitrogen class by ESI (−). These classes provided the most robust models for distinguishing between the origins of the oil samples. The variables were determined using the OPSDA method, which assigns an importance score to each variable in the classification models. Variables with higher importance scores were considered more significant for class separation.

The important variables identified by the OPSDA method for all heteroatom classes, which are paramount in classifying oil samples by geochemical origin, are displayed in Figure S8 (Supporting Information) for both the APPI (+) and ESI (−) ionization techniques. These variables and their molecular formulas are detailed in the Supporting Information Tables S9 and S10 for APPI (+) and ESI (−), respectively. These tables clarify the molecular entities that play a significant role in the geochemical differentiation of the samples.

For interpretative insights, the selected variables from both models for origin (Origins 1 and 2) were concatenated to form a combined data set. From this amalgamated pool, variables with the highest importance scores were chosen. This selection process ensures that the most influential variables across both models are brought to the forefront of a more profound interpretative analysis. The most impactful of these, in the context of geochemical origin, is used to enhance our understanding of the molecular characteristics that distinguish different types of oil.

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

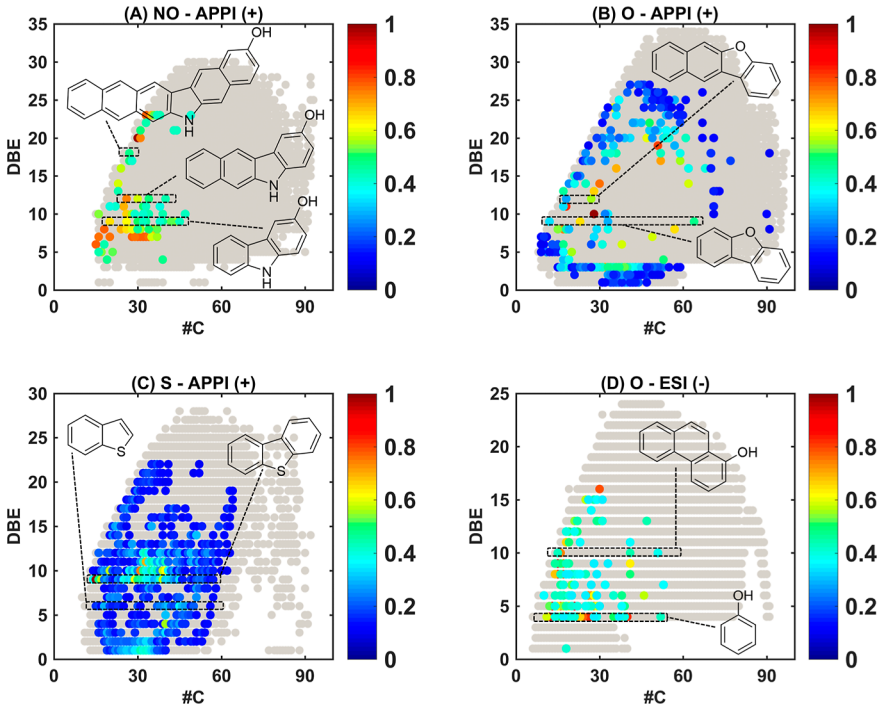


Figure 8. Carbon number vs Double Bond Equivalent (DBE) for the NO, O, and S classes in APPI (+) and O class in ESI (−). All variables (●) are shown, and the most important variables for thermal evolution classification are indicated by their importance, ranging from red to blue (● → ●). The chemical structures illustrate the respective classes and represent the DBE values.

In Figure 7, each point represents a variable, with the color gradient from red to blue indicating the variable's importance in the origin classification (red being the most important). The gray dots represent all variables, providing a backdrop for the selected important variables. The chemical structures depicted are representative molecules from each class and were not explicitly annotated in the data set.

Regarding the HC class assessed by APPI (+), the results underscore that the paramount molecular formulas utilized in modeling are linked to hopane-type pentacyclic triterpenoids. These compounds exhibit variations corresponding to the origins of organic matter and depositional processes in the paleoenvironment.[51−53] This evidence indicates that the OPSDA method accurately identifies variables that are intertwined with biomarkers, which is essential for differentiating samples based on their origins. Additionally, molecular formulas associated with adamantane compounds were selected. Notably, these compounds exhibit thermodynamic stability, suggesting their potential application in oil-source correlation.[54,55]

Class N, selected by OPSDA in ESI (−), is consistently linked to carbazole and benzo homologue compounds. Carbazoles are nitrogen-containing heterocyclic aromatic compounds that are commonly found in crude oils and sedimentary rocks. Previous studies investigated their geochemical significance in petroleum exploration.[56,57] These compounds are suitable for distinguishing crude oils based on their origins and variations in facies and depositional environments while remaining unaffected by maturity.[58] Several studies indicate that lacustrine freshwater−oils tend to be enriched in N-containing compounds, potentially attributed to substantial algae input. Essentially, the precursors of these compounds may stem from amino acids within bacterial and planktonic proteins. Furthermore, nitrogen also originates from chlorophyll and is commonly detected in bitumen and source rock oils preserved within porphyrins.[22,53,59]

Figure 8 offers insights into the thermal evolution of the oil samples by showing the distribution of key variables for the NO, O, and S classes detected by APPI (+) and the O class detected by ESI (−). The OPSDA method's selection of these variables based on their importance scores reflects their relevance in modeling the thermal evolution stages. For a comprehensive understanding of how thermal evolution influences the molecular composition of oil samples, Figure S9 in the Supporting Information extends this analysis across all heteroatom classes. The molecular formulas of all variables, selected for their significant roles in indicating varying degrees of thermal evolution, are listed in Supporting Information Tables S11 and S12 for both the APPI (+) and ESI (−) ionization techniques.

These selections underscored the variables that were most correlated with the thermal history of the samples. By integrating OPSDA's variable importance rankings with the model's predictions, we can discern the molecular transformations that occur because of thermal processes. This knowledge is instrumental in advancing our understanding of the molecular signatures associated with different thermal evolution stages and their implications for oil genesis and maturation.

For the NO class, molecular formulas indicative of nitrogen compounds adorned with hydroxyl groups (benzacaidinol, hydroxydibenzocarbazole and hydroxybenzonapthocarbazole) stand out among the variables significant for discerning the thermal evolution of samples.[60] The distribution of the NO class within oils is intricately linked to the depositional environment and the maturity level of organic matter, highlighting its significance in reflecting the geochemical

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

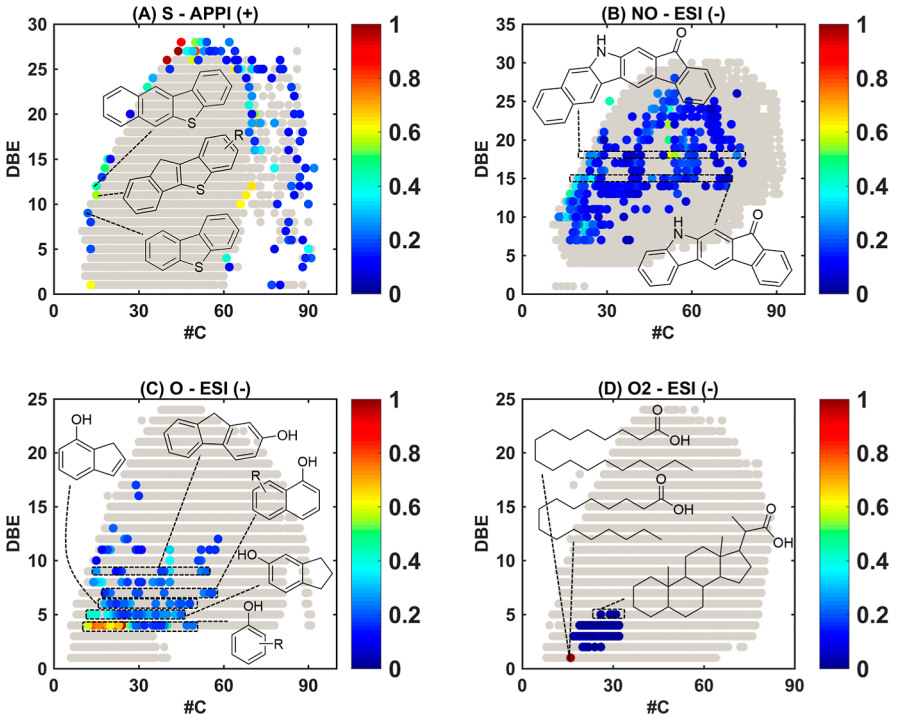


Figure 9. Carbon number vs double-bond equivalent (DBE) for the S class in APPI (+) and NO, O, and O2 classes in ESI (−). All variables (●) are shown, along with the most important variables for biodegradation classification, and are indicated by importance, ranging from red to blue (● → ●). The chemical structures illustrate the respective classes and represent the DBE values.

history of the oil. This connection underscores the relevance of the NO class not only in classifying thermal maturity and providing insights into the geological context and transformation processes of crude oils. Generally, there is an observable trend of increasing NO compound presence as maturity levels rise. This trend can be ascribed to the gradual buildup of nitrogen-rich compounds with lengthy carbon chains and minimal unsaturation within the organic matter as it ages. Such accumulation occurs without reaching a phase of instability, leading to a pronounced presence of these unsaturated nitrogen oxides within the source rocks.[61]

The key variables for the O class in APPI (+) are believed to be primarily composed of aliphatic and aromatic aldehydes, alcohols, ketones, and furans.[57,62] Furthermore, O compounds exhibiting a DBE range of 1 to 4 are likely to be acyclic or nonaromatic cyclic compounds characterized by at least one hydroxyl or carbonyl group.[53,61,63]

In ESI (−), O compounds predominantly comprise phenols, sterols, or alcohols.[16,20,64] Phenols and their alkylated derivatives are recognized as significant contributors to the thermal evolution of crude oils.[53,64] The species with DBE 1 to 3 are most likely components with a hydroxyl functional group able to be deprotonated.[16,53] The core structures for compounds within DBE 4 are likely phenols and their alkylated forms, whereas those in DBE 5 and 6 may correspond to indanols and indenols, respectively. Furthermore, the DBE of 7, 9, and 11 are typically associated with naphthols, fluorenols, and phenanthrenols. Compounds with DBE 4, 5, and 7 could also encompass nonaromatic, monoaromatic, or triaromatic sterols.[64]

For class S, the most significant variables detected through APPI (+) mainly comprised dibenzothiophenes, with some featuring smaller-sized molecules, such as $C_{15}H_{14}S$, and others having elongated alkyl chains, such as $C_{40}H_{62}S$. These compounds serve as petroleum biomarkers for evaluating the thermal evolution of oils.[63] The literature suggests that the relative distributions of methylated dibenzothiophenes can offer insights into the thermal maturity of oils and condensates originating from various source rock types.[65,66]

For an analogous exploration of the effects of biodegradation on oil samples, Figure 9 provides a detailed representation of the significant variables for the S class using APPI (+) and the O2 class using ESI (−). These variables were meticulously chosen through the OPSDA method based on their importance scores, which are particularly indicative of the biodegradation processes the samples have undergone.

In a broader context, Figure S10 in the Supporting Information provides a more extensive picture by illustrating the key variables for all heteroatom classes influenced by biodegradation across APPI (+) and ESI (−) techniques. The molecular formulas of these pivotal variables are listed in Supporting Information Tables S13 and S14. These tables shed light on the molecular fragments most affected by biodegradation, thereby contributing significantly to the differentiation of oil samples based on the extent of their biodegradation.

By closely examining these variables and their patterns, we can begin to unravel the complex interplay between the biodegradation processes and the molecular constitution of petroleum. This understanding is crucial, as it enhances the accuracy of biodegradation assessments and informs recovery strategies and the management of petroleum reservoirs. The figure below highlights the molecular indicators of biodegradation, providing a nuanced view of how these processes alter the chemical landscape of oil samples.

For APPI (+), the most significant variables with all heteroatom classes primarily consisted of species with aromatic nuclei linked to the extended carbon chains. These compounds

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

are of particular significance as potential biodegradation markers, considering that demethylation is a plausible reaction process during the biodegradation of aromatic hydrocarbons.[67] In the case of class S compounds, molecular formulas related to dibenzothiophenes were also identified as highly significant by the OPSDA method. According to a recent study, these compounds are also potential indicators of biodegradation as their concentrations tend to increase with the progression of biodegradation.[68]

For the ESI (−) technique, the OPSDA method prominently highlighted molecular formulas tied to carbazoles with phenol/alcohol groups within the NO class.[20] This selection aligns with prior research that attests to the biodegradability of the carbazole series. Due to their susceptibility to biodegradation, such variables are highly relevant for differentiating between biodegradation levels in oil samples.[57,69] Additionally, fluorenonocarbazoles have been indicated as potential structures within the NO class, further enriching the pool of molecular markers significant for assessing biodegradation stages.[70]

For the O class, both phenols or aromatic alcohols emerge as crucial indicators for microbial degradation processes in crude oils.[70,71] These compounds are pivotal in shedding light on the biodegradation activities within an oil sample. Additionally, the DBE values among O compounds offer detailed insights into their specific chemical structures concerning biodegradation stages. DBE values of 4 are predominantly associated with alkylphenols, while DBE values of 5 and 6 are generally linked to alkylphenols featuring indanol rings, and a DBE of 8 suggests the presence of alkylnaphthol compounds.[63,70] The concentration of O compounds diminishes significantly, becoming trace or undetectable in oils that have undergone severe to extreme biodegradation, yet these compounds are abundant in samples that have not experienced heavy biodegradation.[70] This pattern underscores the role of alcohol groups as intermediary products in oil biodegradation. Alkylphenols, in particular, are prone to complete biodegradation and consumption in environments susceptible to intense biodegradation processes.[70,72]

Finally, for the O2 class, the most important variables are related to acyclic fatty acids and naphthenic acids. These findings are noteworthy, as the ratio of acyclic fatty acids to naphthenic acids has been used in several studies to investigate the biodegradation of crude oils.[20,73−75] Studies have highlighted the impact of fatty acids on the biodegradation process, noting that certain short-chain fatty acids may exhibit toxicity toward petroleum-degrading bacteria.[76] The augmentation of biodegradation in oil-contaminated environments by adding specific fatty acids such as myristic, oleic, linoleic, stearic, and palmitic acids has been demonstrated to enhance the degradation of hydrocarbons.[77,78] This enhancement is supported by evidence showing that some microbial strains, capable of degrading petroleum hydrocarbons, produce fatty alcohols, fatty aldehydes, and fatty acids as intermediate products.[79] Furthermore, the fatty acid composition in various substances has been linked to differences in biodegradation rates, with a higher active group content correlating with increased degradation.[80]

Similarly, naphthenic acids present in petroleum have been identified as potential indicators of crude oil's biodegradation status.[81,82] Aerobic degradation of naphthenic acids, particularly those of lower molecular mass, by rhizosphere microorganisms suggests a preferential breakdown pattern that can be significantly influenced by environmental factors and microbial activity.[83] The presence of naphthenic acids has been pinpointed as the key factor contributing to the acidity of oil, where biodegradation stands out as the primary mechanism resulting in elevated levels of these acids.[84] These findings underscore the importance of both fatty and naphthenic acids in the microbial degradation pathways of petroleum and their utility as biochemical markers for evaluating the extent of biodegradation in contaminated sites.

The discussion in this section thoroughly explores the molecular classes pertinent to the origin, thermal evolution, and biodegradation of the oil samples. This analysis sheds light on the intricate molecular dynamics underlying each geochemical feature. By leveraging the power of machine learning, we have enhanced our understanding of petroleomic complexities. The findings here serve as a foundation for ongoing research seeking to decipher the nuanced molecular narratives that trace the geological journey of petroleum.

4. CONCLUSION

The findings of this study highlight the significant potential of integrating machine learning algorithms with FT-ICR MS to discern and predict the origin, thermal evolution, and biodegradation of oil samples. This integrative approach enhances the accuracy of oil characterization and offers actionable insights that can be leveraged within the petroleum industry and environmental stewardship.

Comparative analysis of the ESI (−) and APPI (+) ionization techniques revealed a nuanced understanding of their respective strengths in capturing the geochemical properties of the oil. The APPI (+) technique demonstrated exemplary accuracy, which is consistently reliable for modeling the thermal evolution of oil, whereas ESI (−) is unmatched in its ability to represent biodegradation processes.

Moreover, this study successfully illuminated the pivotal role of specific molecular formulas within heteroatom classes. These formulas have emerged as indicators in machine learning models, proving to be instrumental in accurately predicting geochemical characteristics. Such molecular markers are invaluable for the petroleum industry as they are key to refining the classification and assessment of oil samples, which can inform extraction, processing, and environmental impact assessments.

In conclusion, this study not only validates the synergistic application of machine learning and FT-ICR MS as a powerful tool for petroleomic analysis but also serves as a stepping stone for identifying definitive molecular markers. These markers can potentially revolutionize the classification and understanding of oils, offering a refined lens through which we can view the intricate tapestry of geochemical processes. As the search for distinct molecular signatures continues, insights from this study will undoubtedly inform future explorations and enhance our ability to navigate the complex world of petroleum geochemistry.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.energyfuels.6c00156.

Summarizing the classification of oil samples by geological origin, thermal evolution, and biodegradation levels, along with the corresponding GC−MS biomarker

https://doi.org/10.1021/acs.energyfuels.6c00156
Energy Fuels 2026, 40, 13761−13776

ratios used to support these classifications. Additional tables provide the acquisition parameters for FT-ICR MS analyses using APPI (+) and ESI (−) ionization, as well as the criteria and performance of molecular formula assignment, including the number of assigned peaks and mass error statistics (ppm). The material also reports extended model performance metrics for both calibration and cross-validation. The figures include the distribution of compound classes, additional model prediction results, carbon number versus DBE diagrams highlighting selected variables, and a multivariate control chart used to assess data consistency (PDF)

AUTHOR INFORMATION

Corresponding Authors

Jussara V. Roque − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil;* orcid.org/0000-0002-5220-959X; Email: jussararoque@gmail.com

Boniek Gontijo − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil;* orcid.org/0000-0003-1197-4284; Email: boniek@ufg.br

Authors

Lidya C. da Silva − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil;* orcid.org/0000-0002-0058-1744

Danielle M. M. Franco − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil*

Gesiane da S. Lima − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil*

Mario D. Rangel − *Center of Excellence in Molecular Studies, Energy and Petroleum (CEMEP), Universidade Federal de Goiás, Goiânia 74690-900 Goiás, Brazil*

Joelma P. Lopes − *Geochemistry Division, PETROBRAS, Rio de Janeiro 21941-915, Rio de Janeiro, Brazil*

Ygor dos S. Rocha − *Geochemistry Division, PETROBRAS, Rio de Janeiro 21941-915, Rio de Janeiro, Brazil*

Igor V. A. F. de Souza − *Geochemistry Division, PETROBRAS, Rio de Janeiro 21941-915, Rio de Janeiro, Brazil*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.energyfuels.6c00156

Funding

The Article Processing Charge for the publication of this research was funded by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), Brazil (ROR identifier: 00x0ma614).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge financial support from Petróleo Brasileiro SA-Petrobras.

REFERENCES

(1) Klein, C. P. A. Formation, Transport, and Lithification of Sediment. In *Earth Materials*; Cambridge University Press, 2012; pp 308−337.

(2) Cui, J.; Zhu, R.; Hu, J. Identification and Geochemical Significance of Polarized Macromolecular Compounds in Lacustrine and Marine Oils. *Chin. J. Geochemistry* **2014**, *33* (4), 431−438.

(3) De La Parra, F.; Pinzon, D.; Mantilla-Duran, F.; Rodriguez, G.; Caballero, V. Marine-Lacustrine Systems during the Eocene in Northern South America − Palynological Evidence from Colombia. *J. South Am. Earth Sci.* **2021**, *108*, 103188.

(4) Tissot, B. P.; Welte, D. H. *Petroleum Formation and Occurrence; Earth and Environmental Science*; Springer Berlin Heidelberg, 2013.

(5) Peters, K. E.; Moldowan, J. M. *The Biomarker Guide: Interpreting Molecular Fossils in Petroleum and Ancient Sediments*; Prentice Hall, 1993.

(6) Dembicki, H., Jr. Interpreting Crude Oil and Natural Gas Data. In *Practical Petroleum Geochemistry for Exploration and Production*; Elsevier, 2017; pp 135−188.

(7) Peters, K. E.; Fowler, M. G. Applications of Petroleum Geochemistry to Exploration and Reservoir Management. *Org. Geochem.* **2002**, *33* (1), 5−36.

(8) Head, I. M.; Jones, D. M.; Larter, S. R. Biological Activity in the Deep Subsurface and the Origin of Heavy Oil. *Nature* **2003**, *426* (6964), 344−352.

(9) Hartono, B. M.; Subroto, E. A.; Kesumajana, A. H. P.; Andrianto, R.; Wahyudiono, J.; Malvinas, G.; Priyanto, B. The Significant Meaning of Biodegradation to the Petroleum System of the Seram Basin, Eastern Indonesia: Evaluation on the Basis of Physio-Chemical Composition Variation of Seram Oils. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1047* (1), 012032.

(10) Peters, K. E.; Walters, C. C.; Moldowan, J. M. *The Biomarker Guide. II Biomarkers and Isotopes in Petroleum Systems and Earth History*, 2nd. ed.; Cambridge University Press: Cambridge, 2004.

(11) Philp, R. P.; Mansuy, L. Petroleum Geochemistry: Concepts, Applications, and Results. *Energy Fuels* **1997**, *11* (4), 749−760.

(12) Rodgers, R. P.; Marshall, A. G. Asphaltenes, Heavy Oils, and Petroleomics. In *Petroleomics: Advanced Characterization of Petroleum-Derived Materials by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS) BT*; Mullins, O. C., Sheu, E. Y., Hammami, A., Marshall, A. G., Eds.; Springer New York: New York, NY, 2007; pp 63−93.

(13) Marshall, A. G.; Rodgers, R. P. Petroleomics: Chemistry of the Underworld. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (47), 18090−18095.

(14) Lima, N. M.; Machado, H. G.; da Silva Lima, G.; da Silva Junior, J. B.; Sousa Cardoso, J. A.; dos Santos, G. F.; Chaves, A. R.; da Silva, R. C.; de Andrade Ferreira, A.; Gontijo, B. Liquid Chromatographic Separation and FT-ICR MS of Crude Oils: Insights from Polar Fractions. *Energy Fuels* **2025**, *39* (42), 20199−20212.

(15) Orrego-Ruiz, J. A.; Marquez, R. E.; Rojas-Ruiz, F. A. New Insights on Organic Geochemistry Characterization of the Putumayo Basin Using Negative Ion Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2020**, *34* (5), 5281−5292.

(16) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Qian, K.; Robbins, W. K. Identification of Acidic NSO Compounds in Crude Oils of Different Geochemical Origins by Negative Ion Electrospray Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Org. Geochem.* **2002**, *33* (7), 743−759.

(17) Vaz, B. G.; Covas, T. R.; Freitas, C. S.; Tose, L. V.; Rangel, M. D.; Silva, R. C.; Pereira, R. C. L.; Rocha, Y. S. Evaluating Orbitrap Mass Spectrometry for Geochemical Characterization of Crude Oils. In *29th International Meeting on Organic Geochemistry*; European Association of Geoscientists & Engineers, 2019; pp 1−2.

(18) Poetz, S.; Horsfield, B.; Wilkes, H. Maturity-Driven Generation and Transformation of Acidic Compounds in the Organic-Rich Posidonia Shale as Revealed by Electrospray Ionization Fourier

Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2014**, *28* (8), 4877−4888.
(19) Hughey, C. A.; Rodgers, R. P.; Marshall, A. G.; Walters, C. C.; Qian, K.; Mankiewicz, P. Acidic and Neutral Polar NSO Compounds in Smackover Oils of Different Thermal Maturity Revealed by Electrospray High Field Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Org. Geochem.* **2004**, *35* (7), 863−880.
(20) Kim, S.; Stanford, L. A.; Rodgers, R. P.; Marshall, A. G.; Walters, C. C.; Qian, K.; Wenger, L. M.; Mankiewicz, P. Microbial Alteration of the Acidic and Neutral Polar NSO Compounds Revealed by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Org. Geochem.* **2005**, *36* (8), 1117−1134.
(21) Zuber, J.; Rathsack, P.; Otto, M. Structural Characterization of Acidic Compounds in Pyrolysis Liquids Using Collision-Induced Dissociation and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2018**, *90* (21), 12655−12662.
(22) Rocha, Y.; Pereira, R. C. L.; Mendonça Filho, J. G.; Mendonça Filho, J. G. Geochemical Characterization of Lacustrine and Marine Oils from Off-Shore Brazilian Sedimentary Basins Using Negative-Ion Electrospray Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (ESI FTICR-MS). *Org. Geochem.* **2018**, *124*, 29−45.
(23) Oliveira, J. V. A.; Roque, J. V.; Franco, D. M. M.; Matos, L. F.; Rangel, M. D.; Lopes, J. P.; Rocha, Y. d. S.; Gontijo, B. Crude Oil Mixture Resolution by FT-ICR MS and Multivariate Modeling: Quantitative Source Attribution Considering Different Thermal Maturities. *Energy Fuels* **2025**, *39* (33), 15648−15658.
(24) Lacroix-Andrivet, O.; Maillard, J.; Mendes Siqueira, A. L.; Hubert-Roux, M.; Loutelier-Bourhis, C.; Afonso, C. Molecular Characterization of Aged Bitumen with Selective and Nonselective Ionization Methods by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. 2. Statistical Approach on Multiple-Origin Samples. *Energy Fuels* **2021**, *35* (20), 16442−16451.
(25) Guillemant, J.; Lacoue-Nègre, M.; Berlioz-Barbier, A.; de Oliveira, L. P.; Albrieux, F.; Joly, J.-F.; Duponchel, L. Evaluating the Benefits of Data Fusion and PARAFAC for the Chemometric Analysis of FT-ICR MS Data Sets from Gas Oil Samples. *Energy Fuels* **2020**, *34* (7), 8195−8205.
(26) Gjelsvik, E. L.; Fossen, M.; Brunsvik, A.; Tøndel, K. Using Machine Learning-Based Variable Selection to Identify Hydrate Related Components from FT-ICR MS Spectra. *PLoS One* **2022**, *17* (8), No. e0273084.
(27) Roque, J. V.; Cardoso, W. J.; de Aguiar, D. V. A.; dos Santos, G. F.; Gomes, A. de O.; Medeiros Júnior, I.; Lima, G. da S.; Gontijo, B. Integrating FT-ICR MS and Machine Learning to Forecast Acid Content Across Boiling Cuts. *Anal. Chem.* **2025**, *97* (11), 5965−5974.
(28) Vaz, B. G.; Abdelnur, V.; Rocha, W. F. C.; Gomes, A. O.; Pereira, R. C. L. Predictive Petroleomics: Measurement of the Total Acid Number by Electrospray Fourier Transform Mass Spectrometry and Chemometric Analysis. *Energy Fuels* **2013**, *27* (4), 1873−1880.
(29) Folli, G. S.; Rafalscky, L.; Feu, A. E.; Chinelatto Junior, L. S.; Filgueiras, P. R.; Romão, W. Petroleomics by High-Resolution Mass Spectrometry: Review and Outlook. *Energy Fuels* **2026**, *40* (4), 1884−1901.
(30) Gjelsvik, E. L.; Fossen, M.; Tøndel, K. Current Overview and Way Forward for the Use of Machine Learning in the Field of Petroleum Gas Hydrates. *Fuel* **2023**, *334*, 126696.
(31) de Aguiar, D. V. A.; Lima, G. d. S.; Roque, J. V.; Oliveira, J. V. A.; Medeiros Júnior, I.; Gomes, A. d. O.; Mendes, L. A. N.; Vaz, B. G. Quantitative Evaluation of Non-Basic Nitrogen-Containing Compounds in Petroleum-Derived Samples by Direct Injection ESI (−) Orbitrap MS. *Anal. Chem.* **2023**, *95* (16), 6507−6513.
(32) Witt, M.; Timm, W. Determination of Simulated Crude Oil Mixtures from the North Sea Using Atmospheric Pressure Photoionization Coupled to Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2016**, *30* (5), 3707−3713.
(33) Corilo, Y. E.; Rowland, S. M.; Rodgers, R. P. Calculation of the Total Sulfur Content in Crude Oils by Positive-Ion Atmospheric Pressure Photoionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2016**, *30*, 3962−3966.
(34) Valente Roque, J.; Rodrigues, M. F.; Dufrayer, G. H. M.; Medeiros Júnior, I.; de Carvalho, R. M.; da Silva Lima, G.; dos Santos, G. F.; Vaz, B. G. Characterization and Quantification of Naphthenic Acids in Produced Water by Orbitrap MS and a Multivariate Approach. *J. Am. Soc. Mass Spectrom.* **2024**, *35* (9), 2128−2135.
(35) Guillemant, J.; Berlioz-Barbier, A.; Albrieux, F.; De Oliveira, L. P.; Lacoue-Nègre, M.; Joly, J. F.; Duponchel, L. Low-Level Fusion of Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Data Sets for the Characterization of Nitrogen and Sulfur Compounds in Vacuum Gas Oils. *Anal. Chem.* **2020**, *92* (3), 2815−2823.
(36) Solihat, N. N.; Son, S.; Williams, E. K.; Ricker, M. C.; Plante, A. F.; Kim, S. Assessment of Artificial Neural Network to Identify Compositional Differences in Ultrahigh-Resolution Mass Spectra Acquired from Coal Mine Affected Soils. *Talanta* **2022**, *248*, 123623.
(37) Li, H.; Ma, J.; Wang, Y.; Zhang, C.; Wu, S.; Su, H. A Combinatorial Machine Learning Models -Based Method for Predicting the Viscosity-Temperature Relationship of Crude Oil. *2025 IEEE 7th Int. Conf. Commun. Inf. Syst. Comput. Eng.*; IEEE 2025, 881−885.
(38) Le, K.; Radović, J. R.; MacCallum, J. L.; Larter, S. R.; Van Humbeck, J. F. Machine Learning in Complex Organic Mixtures: Applying Domain Knowledge Allows for Meaningful Performance with Small Data Sets. *J. Am. Chem. Soc.* **2024**, *146* (32), 22563−22569.
(39) Hur, M.; Kim, S.; Hsu, C. S. Springer Handbook of Petroleum Technology. In *Petroinformatics BT*; Hsu, C. S., Robinson, P. R., Eds.; Springer International Publishing: Cham, 2017; pp 173−198.
(40) Morrison, D. F. Encyclopedia of Biostatistics. In *Hotelling's T 2*; Wiley, 2005.
(41) Riffenburgh, R. H. Tests on the Distribution Shape of Continuous Data. In *Statistics in Medicine*; Elsevier, 2006; pp 369−386.
(42) Palacio Lozano, D. C.; Thomas, M. J.; Jones, H. E.; Barrow, M. P. Petroleomics: Tools, Challenges, and Developments. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **2020**, *13* (1), 405−430.
(43) Marshall, A. G.; Rodgers, R. P. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.* **2004**, *37* (1), 53−59.
(44) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17* (3), 166−173.
(45) Roque, J. V.; Cardoso, W.; Peternelli, L. A.; Teófilo, R. F. Comprehensive New Approaches for Variable Selection Using Ordered Predictors Selection. *Anal. Chim. Acta* **2019**, *1075*, 57.
(46) Wartha, C. A.; de Aguiar Porto, N.; Tomaz, A. C.; Roque, J. V.; Teixeira Diniz, M. B.; Lopes Ribeiro de Queiroz, M. E.; Teófilo, R. F.; Pereira Barbosa, M. H. Classification of Sugarcane Genotypes Susceptible and Resistant to the Initial Attack of Sugarcane Borer Diatraea Saccharalis Using Epicuticular Wax Composition. *Phytochemistry* **2022**, *199*, 113175.
(47) Pérez, N. F.; Ferré, J.; Boqué, R. Calculation of the Reliability of Classification in Discriminant Partial Least-Squares Binary Classification. *Chemom. Intell. Lab. Syst.* **2009**, *95* (2), 122−128.
(48) Cardoso, W.; Roque, J. V.; Jansen, J. J.; Teng, S. Y.; Teófilo, R. F. Combinatorial Order Pre-Processing Search (COPS): A New Pre-Processing Strategy for Large-Scale Interpretable Data Analysis in Process Analytical Technologies. *Comput. Chem. Eng.* **2025**, *192*, 108892.
(49) Noah, M.; Horsfield, B.; Han, S.; Wang, C. Precise Maturity Assessment over a Broad Dynamic Range Using Polycyclic and Heterocyclic Aromatic Compounds. *Org. Geochem.* **2020**, *148*, 104099.
(50) Angolini, C. F. F.; Capilla, R.; Marsaioli, A. J. Petroleum Biodegradation Effects on Polar Acidic Compounds and Correlation with Their Corresponded Hydrocarbon Fractions. *Energy Fuels* **2015**, *29* (8), 4886−4892.
(51) Mello, M. R.; Telnaes, N.; Gaglianone, P. C.; Chicarelli, M. I.; Brassell, S. C.; Maxwell, J. R. Organic Geochemical Characterisation of Depositional Palaeoenvironments of Source Rocks and Oils in Brazilian Marginal Basins. *Org. Geochem.* **1988**, *13* (1−3), 31−45.

(52) Yue, H.; Horsfield, B.; Schulz, H.-M.; Yang, S.; Vieth-Hillebrand, A.; Poetz, S. Preservation of Biotic and Palaeoenvironmental Signatures in Organosulfur Compounds of Immature Fine-Grained Sedimentary Rocks. *Int. J. Coal Geol.* **2023**, *265*, 104168.
(53) Martins, L. L.; Schulz, H.-M.; Noah, M.; Poetz, S.; Severiano Ribeiro, H. J. P.; da Cruz, G. F. New Paleoenvironmental Proxies for the Irati Black Shales (Paraná Basin, Brazil) Based on Acidic NSO Compounds Revealed by Ultra-High Resolution Mass Spectrometry. *Org. Geochem.* **2021**, *151*, 104152.
(54) Li, Y.; Xiong, Y.; Chen, Y.; Tang, Y. The Effect of Evaporation on the Concentration and Distribution of Diamondoids in Oils. *Org. Geochem.* **2014**, *69*, 88−97.
(55) Yang, C.; Wang, Z. D.; Hollebone, B. P.; Peng, X.; Fingas, M.; Landriault, M. GC/MS Quantitation of Diamondoid Compounds in Crude Oils and Petroleum Products. *Environ. Forensics* **2006**, *7* (4), 377−390.
(56) Snyder, L. R. Distribution of Benzcarbazole Isomers in Petroleum as Evidence for Their Biogenic Origin. *Nature* **1965**, *205* (4968), 277.
(57) Han, Q.; Li, M.; Liu, X.; Jiang, W.; Shi, S.; Tang, Y.; He, D. Fractionation of Alkylated Carbazoles in Petroleum during Subsurface Migration: Evidence from Molecular Simulation and Application in Sandstone Reservoirs. *J. Pet. Sci. Eng.* **2020**, *191*, 107308.
(58) Bakr, M. Occurrence and Geochemical Significance of Carbazoles and Xanthones in Crude Oil from the Western Desert, Egypt. *J. King Abdulaziz Univ., Sci.* **2009**, *20* (2), 127−159.
(59) Baxby, M.; Patience, R. L.; Bartle, K. D. The Origin And Diagenesis Of Sedimentary Organic Nitrogen. *J. Pet. Geol.* **1994**, *17* (2), 211−230.
(60) Yuan, M.; Pan, S.; Jing, Z.; Poetz, S.; Shi, Q.; Han, Y.; Zou, C. Geochemical Distortion on Shale Oil Maturity Caused by Oil Migration: Insights from the Non-Hydrocarbons Revealed by FT-ICR MS. *Int. J. Coal Geol.* **2023**, *266*, 104142.
(61) Han, S.; Qiao, Y.; Xiang, C.; Zhang, J.; Wang, Y.; Huo, M.; Mu, X.; Huang, J.; Zhu, J. Geochemical Characteristics of Mesoproterozoic Source Rocks in North China: Insights for Organic Matter Enrichment and Thermal Evolution. *Energies* **2024**, *17* (3), 596.
(62) Huba, A. K.; Huba, K.; Gardinali, P. R. Understanding the Atmospheric Pressure Ionization of Petroleum Components: The Effects of Size, Structure, and Presence of Heteroatoms. *Sci. Total Environ.* **2016**, *568*, 1018−1025.
(63) Oldenburg, T. B. P.; Brown, M.; Bennett, B.; Larter, S. R. The Impact of Thermal Maturity Level on the Composition of Crude Oils, Assessed Using Ultra-High Resolution Mass Spectrometry. *Org. Geochem.* **2014**, *75*, 151−168.
(64) Ziegs, V.; Noah, M.; Poetz, S.; Horsfield, B.; Hartwig, A.; Rinna, J.; Skeie, J. E. Unravelling Maturity- and Migration-Related Carbazole and Phenol Distributions in Central Graben Crude Oils. *Mar. Pet. Geol.* **2018**, *94*, 114−130.
(65) Chakhmakhchev, A.; Suzuki, M.; Takayama, K. Distribution of Alkylated Dibenzothiophenes in Petroleum as a Tool for Maturity Assessments. *Org. Geochem.* **1997**, *26* (7−8), 483−489.
(66) Li, M.; Wang, T.-G.; Shi, S.; Zhu, L.; Fang, R. Oil Maturity Assessment Using Maturity Indicators Based on Methylated Dibenzothiophenes. *Pet. Sci.* **2014**, *11* (2), 234−246.
(67) Huang, H.; Li, Z.; Wen, Z.; Han, D.; Pan, R. Biodegradation Influence on Alkylphenanthrenes in Oils from Bongor Basin, SW Chad. *Sci. Rep.* **2019**, *9* (1), 12960.
(68) Shi, S.; Chen, J.; Zhu, L.; Wang, T. Selective Biodegradation of Dibenzothiophene and Alkyl Dibenzothiophenes in Crude Oils from the Linpan Oilfield, Bohai Bay Basin, Eastern China. *Asian J. Earth Sci.* **2022**, *7*, 100078.
(69) Liao, Y.; Shi, Q.; Hsu, C. S.; Pan, Y.; Zhang, Y. Distribution of Acids and Nitrogen-Containing Compounds in Biodegraded Oils of the Liaohe Basin by Negative Ion ESI FT-ICR MS. *Org. Geochem.* **2012**, *47*, 51−65.
(70) Chang, X.; Liu, T.; Shi, B.; Liu, Z.; Zhang, P.; Xu, Y.; Chen, G. NSO-Bearing Compounds in Severely−Extremely Biodegraded Crude Oil Revealed by ESI(−) FT-ICR MS. *Energy Reports* **2023**, *9*, 1077−1092.
(71) Martins, L. L.; Pudenzi, M. A.; da Cruz, G. F.; Nascimento, H. D. L.; Eberlin, M. N. Assessing Biodegradation of Brazilian Crude Oils via Characteristic Profiles of O 1 and O 2 Compound Classes: Petroleomics by Negative-Ion Mode Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Energy Fuels* **2017**, *31* (7), 6649−6657.
(72) Watson, J. S.; Jones, D. M.; Swannell, R. P. J.; van Duin, A. C.T. Formation of Carboxylic Acids during Aerobic Biodegradation of Crude Oil and Evidence of Microbial Oxidation of Hopanes. *Org. Geochem.* **2002**, *33* (10), 1153−1169.
(73) Cheng, L.; Shi, S.; Yang, L.; Zhang, Y.; Dolfing, J.; Sun, Y.; Liu, L.; Li, Q.; Tu, B.; Dai, L.; Shi, Q.; Zhang, H. Preferential Degradation of Long-Chain Alkyl Substituted Hydrocarbons in Heavy Oil under Methanogenic Conditions. *Org. Geochem.* **2019**, *138*, 103927.
(74) Fafet, A.; Kergall, F.; Da Silva, M.; Behar, F. Characterization of Acidic Compounds in Biodegraded Oils. *Org. Geochem.* **2008**, *39* (8), 1235−1242.
(75) Deng, S.; Li, S.; Li, X.; Jiang, F.; Liu, J. New Geochemical Proxies of Nitrogen- and Oxygen-Containing Compounds Characterized by ESI (−) FT-ICR MS. *Energy Fuels* **2023**, *37* (20), 15505−15522.
(76) Atlas, R. M.; Bartha, R. Inhibition by Fatty Acids of the Biodegradation of Petroleum. *Antonie Van Leeuwenhoek* **1973**, *39* (1), 257−271.
(77) Obbard, J. P.; Ng, K. L.; Xu, R. Bioremediation of Petroleum Contaminated Beach Sediments: Use of Crude Palm Oil and Fatty Acids to Enhance Indigenous Biodegradation. *Water, Air, Soil Pollut.* **2004**, *157* (1−4), 149−161.
(78) Breger, I. A. Diagenesis of Metabolites and a Discussion of the Origin of Petroleum Hydrocarbons. *Geochim. Cosmochim. Acta* **1960**, *19* (4), 297−308.
(79) Yanto, D. H. Y.; Tachibana, S. Biodegradation of Petroleum Hydrocarbons by a Newly Isolated Pestalotiopsis Sp. NG007. *Int. Biodeterior. Biodegrad.* **2013**, *85*, 438−450.
(80) Luo, Z.; Xia, C.; Fan, H.; Chen, X.; Peng, B. The Biodegradabilities of Different Oil-Based Fatliquors. *J. Am. Oil Chem. Soc.* **2011**, *88* (7), 1029−1036.
(81) Hughey, C. A.; Minardi, C. S.; Galasso-Roth, S. A.; Paspalof, G. B.; Mapolelo, M. M.; Rodgers, R. P.; Marshall, A. G.; Ruderman, D. L. Naphthenic Acids as Indicators of Crude Oil Biodegradation in Soil, Based on Semi-quantitative Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2008**, *22* (23), 3968−3976.
(82) Nascimento, L. R.; Rebouças, L. M. C.; Koike, L.; de A.M Reis, F.; Soldan, A. L.; Cerqueira, J. R.; Marsaioli, A. J. Acidic Biomarkers from Albacora Oils, Campos Basin, Brazil. *Org. Geochem.* **1999**, *30* (9), 1175−1191.
(83) Biryukova, O. V.; Fedorak, P. M.; Quideau, S. A. Biodegradation of Naphthenic Acids by Rhizosphere Microorganisms. *Chemosphere* **2007**, *67* (10), 2058−2064.
(84) Meredith, W.; Kelland, S.-J.; Jones, D. Influence of Biodegradation on Crude Oil Acidity and Carboxylic Acid Composition. *Org. Geochem.* **2000**, *31* (11), 1059−1073.