Petroleum Science 23 (2026) 2587–2598

Contents lists available at ScienceDirect

Petroleum Science

journal homepage: www.keaipublishing.com/en/journals/petroleum-science

Original Paper

A machine learning-driven interpretative framework for reconstructing hydrocarbon evolution in hybrid petroleum systems

Ke-Yu Tao [a,b,*], Jian Cao [b,**], Yu-Ce Wang [b,c], Wan-Yun Ma [d]

[a] Key Laboratory of Marine Ecosystem Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, 310027, Zhejiang, China
[b] School of Earth Sciences and Engineering, Nanjing University, Nanjing, 210023, Jiangsu, China
[c] PetroChina Hangzhou Research Institute of Geology, Hangzhou, 310027, Zhejiang, China
[d] Research Institute of Experiment and Testing, PetroChina Xinjiang Oilfield Company, Karamay, 834000, Xinjiang, China

ARTICLE INFO

Article history:
Received 12 May 2025
Received in revised form
26 October 2025
Accepted 17 December 2025
Available online 22 December 2025

Edited by Min Li

Keywords:
Machine learning
UMAP
Geochemistry
Hydrocarbons
Hybrid petroleum system
Junggar Basin

ABSTRACT

The genetic identification of hydrocarbons in complex hybrid petroleum systems remains challenging due to overlapping geochemical signatures caused by multi-source inputs and superimposed geological processes. Traditional biomarker-based methodologies often struggle to decouple these nonlinear interactions, leading to interpretive uncertainties in source correlation, thermal maturity assessment, and secondary alteration characterization. This study introduces an unsupervised machine learning framework leveraging manifold learning to resolve these challenges within the hybrid petroleum system of the eastern Junggar Basin. We employed Uniform Manifold Approximation and Projection (UMAP) to analyze high-dimensional molecular fingerprints of hydrocarbons. This approach allowed us to systematically disentangle the genetic signals influenced by multiple factors, including source material, thermal evolution, mixing, biodegradation, and migration-induced phase fractionation. Results identify two primary oil families: Permian-derived and Jurassic-sourced oils, each exhibiting unique evolutionary pathways shaped by differential thermal maturation and post-generation alterations. Spatial mapping of these genetic types reveals systematic trends in hydrocarbon accumulation, highlighting preferential migration pathways and high-potential exploration targets. This workflow not only advances the interpretation of hybrid petroleum systems but also establishes a transferable framework for optimizing exploration strategies in geochemically complex basins. The integration of machine learning with petroleum geochemistry provides a promising pathway to reconcile multi-proxy datasets, reduce interpretive subjectivity, and enhance predictive accuracy in hydrocarbon genetic studies.

1. Introduction

Genetic analysis of hydrocarbons constitutes a core research focus in petroleum geochemistry, providing critical geological constraints for exploration target optimization, resource potential evaluation, and play element characterization (Peters and Fowler, 2002; Magoon, 2004; Curiale, 2008). Molecular fingerprints and isotopic signatures serve as key diagnostic tools for deciphering the genetic information of sedimentary organic matter, with biomarker assemblages offering particularly valuable insights (Peters et al., 2005; Luo et al., 2019). These geochemical proxies provide fundamental evidence for determining hydrocarbon source, constraining thermal evolution stages, and identifying secondary alteration processes (e.g., Wei et al., 2006; Ding et al., 2020). However, with the substantial increase in exploration datasets within petroliferous basins, traditional organic geochemical indicators have exhibited growing complexity in their interrelationships (Curry, 2019). Furthermore, interpretive discrepancies frequently arise when integrating multi-proxy datasets. This is particularly evident in maturity calibration and oil-source correlation. Such inconsistencies introduce substantial uncertainties into hydrocarbon genetic identification (Curiale, 2008; Murray and Peters, 2021).

Extensive sample datasets provide statistically robust insights into the complex geological evolution of petroleum systems. Nevertheless, conventional analytical methodologies necessitate

* Corresponding author. Key Laboratory of Marine Ecosystem Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, 310027, Zhejiang, China.
** Corresponding author.
E-mail addresses: taoky@sio.org.cn (K.-Y. Tao), jcao@nju.edu.cn (J. Cao).

https://doi.org/10.1016/j.petsci.2025.12.029

substantial reliance on discretionary selection and interpretation of geochemical indicators (Curry, 2019; Snodgrass and Milkov, 2020; Su et al., 2025). Such subjectivity introduces potential biases in theoretical models, thereby compromising their applicability in practical exploration scenarios. The analytical challenges primarily arise from the multifaceted evolution of sedimentary organic matter under diverse geological constraints. Organic molecular compositions, initially determined by source material variations, are further modified through post-depositional processes such as thermal maturation, migration-mixing dynamics, and biodegradation (Peters et al., 2005; Naafs et al., 2019). Crucially, individual molecular markers exhibit differential responses to these factors, leading to nonlinear alterations in petroleum geochemical signatures (Tao et al., 2025). Consequently, the diagnostic reliability of single-parameter indices diminishes for tasks such as oil-source correlation, thermal maturity evaluation, and secondary alteration characterization. This complexity is amplified in hybrid petroleum systems, where multiple source inputs and superimposed alteration processes coexist, resulting in significant interpretive ambiguities.

Contemporary geoscience research is witnessing an exponential growth in geochemical datasets. This trend presents unprecedented opportunities to extract advanced information from large-scale sample repositories. This paradigm shift necessitates methodological innovations to overcome inherent constraints in conventional hydrocarbon genetic studies. Machine learning (ML)-driven big data analytics have emerged as powerful tools for investigating complex origins of hydrocarbons within hybrid petroleum systems. These tools are of superior capabilities in decoupling high-dimensional nonlinear relationships. Such advancements have laid the foundation for an innovative analytical paradigm. Notable applications demonstrate this methodological evolution. Alexandrino et al. (2016) effectively differentiated marine versus lacustrine crude oils through support vector machine discriminant analysis (SVM-DA) modeling on two-dimensional gas chromatography-quadrupole mass spectrometry (GC × GC-QMS) datasets. Snodgrass and Milkov (2020) achieved remarkable generalization performance (97% prediction accuracy) in natural gas genetic identification by developing a random forest classifier based on global molecular and isotopic characteristics from >10,000 natural gas samples. Tao et al. (2020, 2025) developed an integrated analytical framework combining: 1) manifold learning techniques including t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP); 2) random forest classifiers/regressors; and 3) Shapley additive explanation (SHAP) interpretative framework. This multi-algorithm framework successfully disentangles and interprets the nonlinear variations in high-dimensional molecular fingerprints and isotopic signals in source rocks and crude oils. Lu et al. (2025) employed a random forest classifier to decipher paleoclimate-controlled organic matter enrichment mechanisms, and effectively discriminate between freshwater and saline lacustrine oil shales based on integrated geochemical and biomarker data.

To advance this emerging paradigm in petroleum geochemistry, this study seeks to demonstrate its promising potential spanning from pattern recognition to exploration strategy development. Focusing on the eastern Junggar Basin, a well-documented hybrid petroleum system with intricate hydrocarbon geochemical characteristics (Chen et al., 2003a, 2003b; Wang et al., 2013), this study addresses two principal objectives. Firstly, to screen the optimal ML algorithm capable of deciphering nonlinear genetic information embedded within hydrocarbon molecular fingerprints. Algorithm selection criteria emphasize the capacity to disentangle overlapping geological influences and to distinctly decouple the associated geochemical signature trends. Secondly, to identify and calibrate four critical features for extensive hydrocarbon samples: source, thermal evolution, secondary alteration, and migration processes. Through this workflow, a comprehensive interpretative framework will be established to elucidate the evolution and spatial occurrence of hydrocarbons in this hybrid petroleum system. This framework will provide actionable insights and a robust theoretical foundation for optimizing exploration strategies.

2. Geological setting

The Junggar Basin is a large petroliferous basin formed under a compressional tectonic setting, exhibiting complex superposition characteristics typical of multi-phase structural evolution (Carroll et al., 1990; Cai et al., 2000). Tectonostratigraphic studies demonstrate four principal orogenic episodes from Late Paleozoic to Quaternary: Hercynian (Late Paleozoic), Indosinian (Triassic-Jurassic), Yanshanian (Cretaceous), and Himalayan (Cenozoic) movements (Feng et al., 1989; Chen et al., 2005; Han and Zhao, 2018). This polycyclic tectonic evolution has generated superimposed structural assemblages and multistage sedimentary systems.

The study area, centered on the central-eastern Junggar Basin, encompasses multiple second-order tectonic units as delineated in Fig. 1. Stratigraphic successions spanning Carboniferous to Cretaceous periods are extensively developed, attaining thicknesses up to 7000 m. Previous investigations have revealed significant heterogeneity in the geochemical compositions of hydrocarbons. Three key factors contribute to the observed hydrocarbon complexity: i) intricate spatial distribution patterns of variably sourced and matured crude oils, ii) widespread mixing events between genetically distinct petroleum systems, and iii) localized secondary alteration processes (Chen et al., 2003a, 2003b; Wang et al., 2013). These superimposed effects collectively govern the heterogeneous physical properties and geochemical signatures of crude oils in the region.

3. Samples and methods

This study analyzed 539 hydrocarbon samples comprising 199 crude oils and 340 reservoir extracts collected from 161 wells across the study area (Fig. 1). These samples provide comprehensive spatial coverage of explored geological extent while effectively representing the full spectrum of geochemical signatures characteristic of this hybrid petroleum system. All experimental work was performed in a single accredited facility following rigorous quality control procedures. This standardized analytical workflow ensures data integrity and analytical reproducibility across the dataset.

3.1. Hydrocarbon geochemical characterization

The molecular composition of hydrocarbon samples was characterized to investigate their molecular signatures. Reservoir rock samples were crushed to 80-mesh particles and subjected to Soxhlet extraction using a 93:7 (*v*/*v*) dichloromethane-methanol solution. Asphaltene fractions in both crude oils and extracts were precipitated through excess *n*-hexane addition. Subsequently, deasphalted hydrocarbons underwent sequential chromatographic separation on silica-alumina columns via an eluotropic series of *n*-hexane, *n*-hexane-dichloromethane (7:3 *v*/*v*), and methanol, yielding distinct saturated hydrocarbon, aromatic hydrocarbon, and resin fractions. For molecular fingerprinting, saturated hydrocarbon fractions were analyzed using an Agilent 7890B gas chromatography equipped with a flame ionization detector (GC-FID). Compound resolution was achieved using a DB-1MS column (60 m × 0.32 mm i.d.; 0.25 μm film

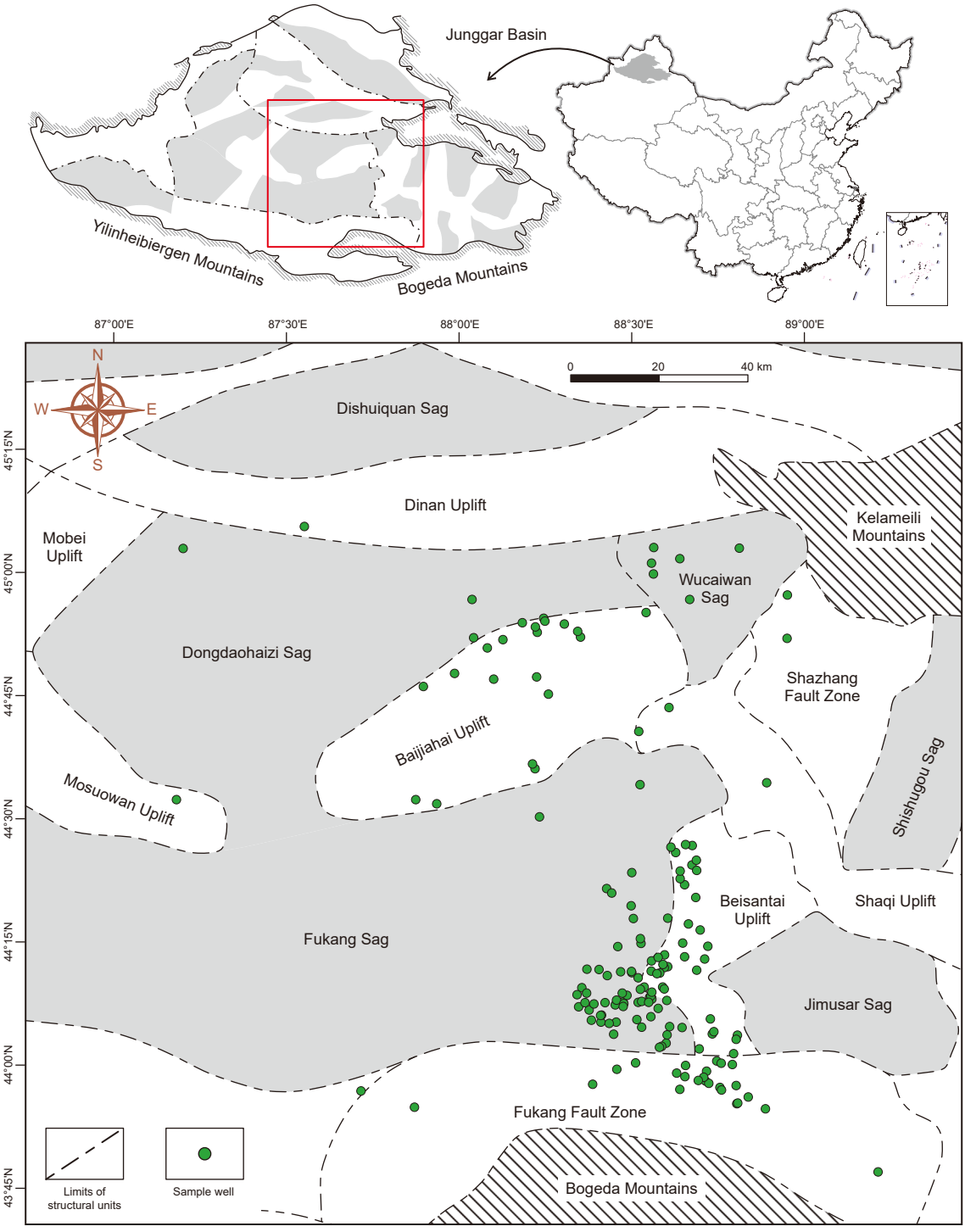

Fig. 1. Location and structural setting of the study area within the Junggar Basin. The green dot indicates the well sites of the analytic samples.

thickness) under nitrogen carrier gas flow. The GC oven temperature was initially set to 30 °C for 15 min, increased to 310 °C at 3 °C/min, and finally held at 310 °C for 30 min. Subsequent structural identification was performed using an Agilent 7890B-5977A gas chromatography-mass spectrometry (GC-MS) system with identical column specifications. The GC oven temperature program initiated at 80 °C (2 min isothermal), followed by sequential ramping to 220 °C at 3 °C/min and 295 °C at 2 °C/min, with a final 30-min hold to ensure complete elution of high-molecular-weight components.

3.2. Machine learning analysis

To elucidate latent genetic information encoded in petroleum molecular compositions, three unsupervised machine learning approaches were systematically implemented: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP). The analytical molecular dataset contains relative abundance profiles of 92 distinct compounds, segmented into three categories: 1) relative peak intensity percentages of major

constituents per sample as determined by GC-FID results, including *n*-alkanes and branched alkanes from C_8 to C_{38}, and carotanes; 2) relative peak intensity percentages of terpanes as identified in the *m/z* 191 GC-MS spectra; and 3) relative peak intensity percentages of steranes in the *m/z* 217 GC-MS spectra. These three compositional subsets were integrated to form the final input dataset for PCA, t-SNE and UMAP analyses. Given that these molecular data is fundamental in the field of petroleum molecular geochemistry, the proposed method is highly applicable and suitable for broader adoption. Computational implementations leveraged dedicated R packages: the "vegan" package for PCA, "Rtsne" for t-SNE, and "umap" for UMAP. Through empirical optimization, the UMAP configuration adopted critical hyperparameters of metric = "euclidean", n_neighbors = 30 and min_dist = 0.1 to balance global structure preservation with local pattern resolution.

4. Results and discussion

4.1. Dimensionality reduction of molecular compositions in hydrocarbons

To elucidate potential genetic correlations and compositional disparities among hydrocarbon samples (including crude oils and extracts), we systematically evaluated three dimensionality-reduction algorithms for visualizing their molecular complexity. Comparative analysis revealed significant limitations in PCA, a linear dimensionality-reduction method constrained by its cumulative explanatory variance of merely 45% for the first two principal components (Fig. 2(a)). This inadequacy suggests the presence of intricate molecular compositions and potential polynomial relationships within the dataset, necessitating the application of nonlinear approaches. Among nonlinear algorithms, t-SNE demonstrated partial success by identifying discrete endmember clusters in oil samples, though extract samples exhibited less coherent clustering patterns (Fig. 2(b)). Notably, UMAP demonstrated optimal performance, resolving three dominant genetic endmembers for the oils, with transitional samples displaying intermediate characteristics (Fig. 2(c)). Extracts exhibited divergent geochemical signatures, with approximately 62% aligning with established oil clusters while the remainder formed unique compositional groupings (Fig. 2(c)).

Comparative analysis of three-dimensional t-SNE and UMAP outputs revealed substantial structural congruence between the dimensionality-reduction patterns (Table S1). Mechanistically, UMAP outperformed t-SNE in preserving global structural features during 2D projection of high-dimensional nonlinear relationships (Liu et al., 2024). The t-SNE's stochastic nature potentially compromises topological fidelity through excessive local cluster emphasis. The validity of UMAP has been extensively verified across multidisciplinary applications in petroleum geology (Liu et al., 2024; Zhang et al., 2024). These findings collectively establish UMAP as a superior analytical framework for resolving nonlinear molecular heterogeneity in complex petroleum systems where geochemical signatures exhibit multidimensional variance patterns (Tao et al., 2025).

4.2. Genetic classification and identification of petroleum

The validity and geochemical implication of the UMAP output can be verified by the expression of biomarker proxies. We emphasize that conventional biomarker interpretation faces inherent limitations in complex petroleum systems due to multidimensional interference effects—where single parameters frequently record overlapping source, maturity, and alteration signals (Curiale, 2008). Notably, the UMAP analysis successfully established a multivariate differentiation framework capable of resolving these entangled biomarker evolution pathways through nonlinear pattern recognition (Tao et al., 2025). Comprehensive integration of diverse proxies within this framework revealed five geochemically distinct petroleum groups: two protogenetic families (I and II) and three alteration-dominated types exhibiting (i) mixing-controlled hybridization, (ii) biodegradation-induced compositional changes, and (iii) phase fractionation-driven component segregation (Fig. 3).

4.2.1. Protogenetic petroleum families and thermal evolution patterns

The systematic evaluation of source-related biomarker assemblages reveals distinct differentiation in organic facies between petroleum Families I and II. Although these parameters may be influenced by secondary geochemical processes such as thermal maturation, the UMAP-derived dimensionality reduction successfully resolves their multivariate variability. Notably, β-carotane abundance emerges as a diagnostic discriminator, with Family I exhibiting β-carotane as a dominant component in its saturated hydrocarbon fraction (mean β-carotane/n-C_{max} ratio = 0.28), whereas it is nearly undetectable in Family II (mean

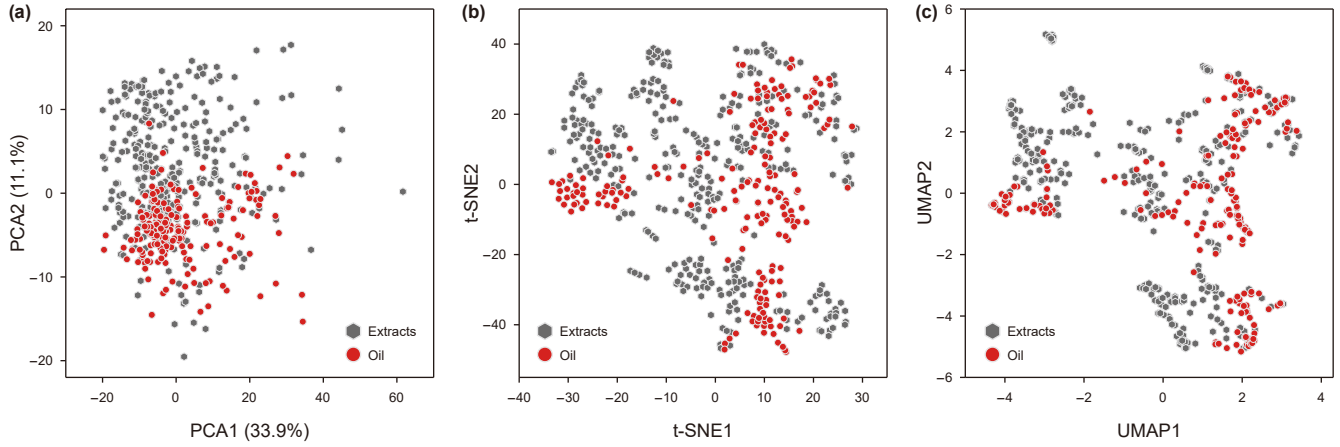

Fig. 2. Comparative evaluation of dimensionality-reduction techniques in visualizing multivariate structural relationships of crude oil and extract molecular fingerprints. Dimensionless coordinates represent low-dimensional embedding of high-dimensional feature space. **(a)** Principal component analysis (PCA); **(b)** t-distributed stochastic neighbor embedding (t-SNE); **(c)** Uniform manifold approximation and projection (UMAP).

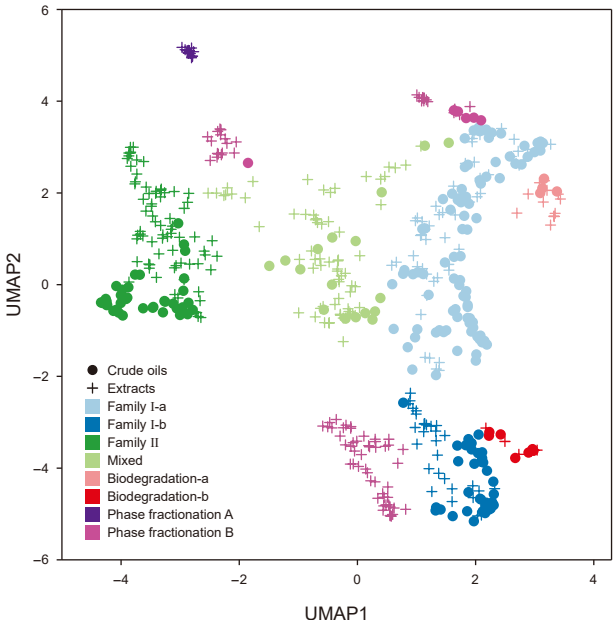

Fig. 3. UMAP framework visually illustrating the genetic disparities and relationships among the crude oils and extracts.

ratio = 0.02; Fig. S1(a)). This contrast is further substantiated by comparative analysis of followed parameters: Family I exhibits lower pristane/phytane (Pr/Ph) ratios (mean = 1.54) relative to Family II (mean = 2.51; Fig. S1(b)). The gammacerane index (gammacerane/C_{30} hopane) shows elevated values in Family I (mean = 0.21) compared to Family II (mean = 0.08; Fig. S1(c)). The C_{24}TeT/C_{26}TT (C_{24} tetracyclic terpane/C_{26} tricyclic terpane) ratio is lower in Family I (mean = 0.75) compared to Family II (mean = 3.08; Fig. S1(d)). Sterane distributions reveal significant inter-family variations, with Family I showing notably higher relative abundances of $\alpha\alpha\alpha C_{28}$ 20R and reduced proportions of $\alpha\alpha\alpha C_{29}$ 20R compared to Family II (Fig. S1(e) and (f)).

The convergent evidence from multiple biomarker types permits robust paleoenvironmental reconstructions. Family I geochemical signatures indicate deposition under relatively high salinity and anoxic conditions, as evidenced by enhanced gammacerane indices and β-carotane preservation (Moldowan et al., 1985; Fu et al., 1986; Grice et al., 1998). The associated organic matter (OM) primarily originated from aquatic algae and bacterial sources, as evidenced by hopane and sterane distributions (Peters et al., 2005). Conversely, Family II appears to originate from more oxygenated, lower salinity depositional systems with substantial terrestrial OM input, as evidenced by high Pr/Ph ratio and abundant C_{29} steranes and C_{24}TeT (Moldowan et al., 1985; Noble et al., 1985a; Hughes et al., 1995).

Furthermore, the UMAP framework effectively captures petroleum molecular heterogeneity through differential inter-sample dispersion patterns. It is observed that Family I exhibits broad dispersion within the UMAP space (Fig. 3), a pattern strongly correlated with thermal maturation gradients. This interpretation is supported by progressive variations in multiple maturity-sensitive parameters across the Family I cluster, including: $\beta\beta/(\alpha\alpha+\beta\beta)$ stereoisomerization ratios (Fig. S1(g)), TT/C_{30}H (peak of tricyclic terpane to C_{30} hopane ratio; Fig. S1(h)), pregnane index (pregnane/$\alpha\alpha\alpha C_{29}$ 20R sterane ratio; Fig. S1(i)), and Ts/Tm (18α-trisnorneohopane/17α-trisnorhopane; Fig. S1(j)). These systematic variations facilitate subdivision of Family I into two subgroups: Family I-a (low-maturity to peak mature) and Family I-b (late mature to post-mature) as detailed in Section 4.3. Of note, thermal maturation exerts certain effects on source diagnostic biomarkers, typically including β-carotane/n-C_{max} (Fig. S1(a)) and C_{24}TeT/C_{26}TT (Fig. S1(d)). These ratios exhibit progressive depletion with increasing thermal stress, attributable to differential thermal stability between compound classes (Peters and Moldowan, 2017).

In contrast, Family II displays a narrow thermal maturation spectrum, as evidenced by its compact clustering in UMAP space (Fig. 3). Nevertheless, subtle maturity gradients are discernible along the UMAP2-axis (Fig. S1(g)–S1(j)), mirroring the pattern observed in Family I. Notably, a compositional dichotomy emerges between sample types: extracts are predominantly characterized by early-maturity signatures, whereas crude oils represent more mature phases (see Section 4.3 for detailed characterization).

To link the hydrocarbon families in our UMAP results with specific sources, an important background consideration is required. In petroliferous systems, hydrocarbon fluid samples that can be collected are generally more abundant and diverse than source rock samples. Thus, compositions of hydrocarbon fluid can capture a broader spectrum of generative processes and maturity stages, whereas source rock data often represent a limited stage of generative processes (e.g., low-maturity end members). In our framework, characteristics of known source rock can be viewed as points along a continuous hydrocarbon evolution strand. A reliable correlation is established only when the entire strand (hydrocarbon fluids) can be rationally linked to a specific source point, and when variations along that strand are consistently explained by recognized secondary processes.

Geochemical evidence from previous source rock studies in the eastern Junggar Basin suggests distinct origins for the two petroleum families: Family I derived from the Early-Middle Permian Lucaogou Fm, whereas Family II predominantly originates from Jurassic coal-bearing strata. Specifically, the Lucaogou Fm is well documented as containing highly organic-rich, oil-prone source rocks (Chen et al., 2003a; Wang et al., 2020, 2023; Wu et al., 2021). This unit was deposited in a salinity-stratified lacustrine environment characterized by suboxic-anoxic conditions and hydrothermal influences (Liu et al., 2022; Wu et al., 2022; Xia et al., 2023). The Lucaogou source rock is the only unit in the study area shares a range of highly matched characteristics in terms of source-related biomarkers with Family I, typically including high concentration of β-carotane, a relatively high gammacerane index, and a relatively high proportion of C_{28} within C_{27-29} regular steranes (Chen et al., 2003a; Wu et al., 2021; Wang et al., 2023). In contrast, Jurassic source rocks comprise coal measures and interbedded dark mudstones deposited in oxic, freshwater lacustrine-deltaic environments (Li et al., 2014). These rocks exhibit numerous highly consistent biomarker characteristics with Family II, typically including extremely high Pr/Ph and C_{24}TeT/C_{26}TT ratios, as well as the absolute predominance of C_{29} among C_{27-29} regular steranes (Chen et al., 2003a). Although coal seams exhibit limited oil generation potential, the associated mudstones may contribute to localized oil accumulations in the study area (Chen et al., 2003a).

The observed maturity disparity between the two families is most likely attributed to contrasting hydrocarbon generation capacities of their source rocks. Lucaogou source rocks, dominated by Type I-II kerogens, demonstrate sustained liquid hydrocarbon generation across the full oil window (Wu et al., 2021). Conversely, Jurassic mudstones contain Type II_2-III kerogens, which generate liquid hydrocarbons predominantly during peak maturation stages (Qian et al., 2018). As a result, limited early-formed oils typically remain as dispersed phases within reservoirs.

4.2.2. Mixing effect

The observed hydrocarbons mixing process in the UMAP arises from migration interactions between these two petroleum families. The UMAP framework effectively delineates geochemical variations within these mixed systems, where spatial proximity to a specific petroleum family correlates with compositional similarity. Analytical results demonstrate that mixed oils/extracts in the study area exhibit stronger geochemical affinity with Family I (Fig. 3), indicating predominant contributions from Lucaogou-derived hydrocarbons. This predominance is particularly evident in parameters such as β-carotane/n-C_{max} (Fig. S1(a)) and steranes distributions (Fig. S1(e) and (f)). Notably, the mixing continuum primarily occurs between Family I-a (lower-maturity) and Family II endmembers, forming transitional signatures along the UMAP1-axis (Fig. 3). In contrast, Family I-b (higher-maturity) samples maintain distinct clustering patterns, suggesting limited interaction with other petroleum populations during migration.

4.2.3. Biodegradation

The UMAP framework identified two distinct clusters characterized by exceptionally elevated β-carotane abundances (Fig. S1(a)). Chromatographic analysis confirmed biodegradation signatures in these samples, manifested through notable depletion of n-alkanes and concomitant enrichment of β-carotane in saturated hydrocarbon fractions (Fig. S2). This fractionation pattern reflects the superior biodegradation resistance of β-carotane relative to straight-chain alkanes, resulting in progressive β-carotane/n-C_{max} ratio amplification with increasing biodegradation intensity (Wenger et al., 2002). Similarly, branched/isoprenoid alkanes demonstrate enhanced microbial recalcitrance compared to n-alkanes (Peters and Fowler, 2002), as evidenced by elevated Ph/n-C_{18} ratios (mean of 4.8 vs. mean of 0.4 in non-degraded samples; Fig. S1(k)).

Based on the UMAP framework, it is evident that the two clusters (Biodegradation-a and Biodegradation-b) are respectively associated with Family I-a and Family I-b (Fig. 3). This suggests that they represent biodegraded products derived from lower-maturity and higher-maturity oils originating from the Lucaogou Fm, respectively. The biodegraded oils predominantly exhibit moderate alteration levels, showing alterations in alkanes with preserved terpane/sterane distributions (Fig. S2). Notably, some Biodegradation-a samples display advanced degradation, characterized by 17α-hopane depletion and 25-norhopane generation (Noble et al., 1985b; Fig. S2(a)). This biodegradation gradient drives systematic density variations, with Biodegradation-a oils averaging 0.94 g/cm^3 (API 19°) vs. 0.88 g/cm^3 (API 30°) for Biodegradation-b.

4.2.4. Phase fractionation

Within the UMAP framework, Phase fractionation A emerges as a distinct outlier cluster exclusively composed of hydrocarbon extracts. Three transitional clusters (Phase fractionation B) form connectivity bridges between this end-member and Family I-a, Family I-b, and Mixed Oil groups, respectively (Fig. 3). Notably, the identification of Phase fractionation B is through 3D UMAP visualization, as their transitional relationships with Phase fractionation A involve multidimensional variations exceeding the representational capacity of 2D UMAP embedding.

Phase fractionation A exhibits diagnostic geochemical signatures, most prominently characterized by exceptionally high Terrigenous/Aquatic Ratio (TAR) values (Fig. S1(l)). Chromatograms reveal a complete absence of $<C_{16}$ n-alkanes in these extracts, contrasting with well-preserved high-molecular-weight homologues (C_{27}–C_{29} dominance; Fig. S3(a)). This pronounced waxy hydrocarbon composition strongly suggests intensive phase fractionation during petroleum migration. Such fractionation likely occurred in fracture systems where rapid pressure drops induced sequential precipitation of heavy components (Volk et al., 2000).

Phase fractionation B encompasses residual petroleum phases demonstrating progressive elimination of high-molecular-weight alkanes (C_{25+}; Fig. S3(b)–(d)). These phases retain continuous alkane distributions spanning light to intermediate fractions, a characteristic fingerprint of migration-induced phase partitioning (Suchý et al., 2010; Han et al., 2015). Subdivision of Phase fractionation B within the UMAP framework effectively discriminates source-related variations among original oil compositions (Fig. 3). This classification is corroborated by the conserved distribution patterns of hopanes and steranes—biomarkers exhibiting relative stability during phase fractionation processes (van Grass et al., 2000; Fig. S3(b)–(d)). Collectively, the Phase fractionation A-B continuum documents a comprehensive sequence of hydrocarbon fractionation events during secondary migration.

4.3. Decoupling multivariate controls on geochemical proxies

The UMAP analytical framework enables effective discrimination of distinct response patterns exhibited by geochemical proxies under multifactorial geological controls. This methodology demonstrates particular efficacy in establishing robust correlation trends among proxies. This capability enhances predictions of petroleum evolutionary stages, even when non-specific geochemical parameters are employed. Such advancement addresses a critical challenge. Given that most molecular indicators exhibit multivariate responses to diverse geological processes, substantially limiting their diagnostic utility in complex petroleum systems.

The pregnane index, while proposed as a potential thermal maturity indicator (Tao et al., 2021), demonstrates significant covariation with organic facies and biodegradation (Requejo et al., 1997; Peters et al., 2005; Zhang et al., 2021). Similarly, the diasteranes/regular steranes ratio (DiaS/S) responds sensitively to organic facies, thermal stress, and biodegradation (Seifert and Moldowan, 1979). Cross-plotting these two parameters reveals source-dependent evolutionary trajectories: Family I and Family II exhibit distinct nonlinear correlation trends (Fig. 4(a)). These trends correspond to differential molecular evolution pathways of OM from discrete source facies during thermal maturation. Notably within Family I, the dataset splits into two maturity-defined subgroups along the correlation trend: Family I-a (lower maturity) and Family I-b (higher maturity) (Fig. 4(a)). Intriguingly, biodegraded samples form a third linear trend that diverges from the maturation trajectories. This trend specifically indicates that progressive biodegradation elevates both pregnane index and DiaS/S values. These biodegraded oils, designated Biodegradation-a and Biodegradation-b, originate respectively from Family I-a and I-b endmembers of the Lucaogou Fm. The observed trend separation provides insights for differential biodegradation intensity between lower-maturity and higher-maturity petroleum phases. Of them, Biodegradation-a commonly underwent severer biodegradation, resulting in broader variations in pregnane index and DiaS/S (Fig. 4(a)).

The interference of source-facies heterogeneity on maturity parameters can be particularly evidently discerned by examining the correlation between pregnane index and TT/C_{30}H ratio. Logarithmic cross-plots of these parameters reveal that Family I and Family II oils cluster along distinct evolutionary trajectories (Fig. 4(b)). This provides independent validation of the source-facies differentiation achieved through UMAP analysis. Notably, mixed hydrocarbons

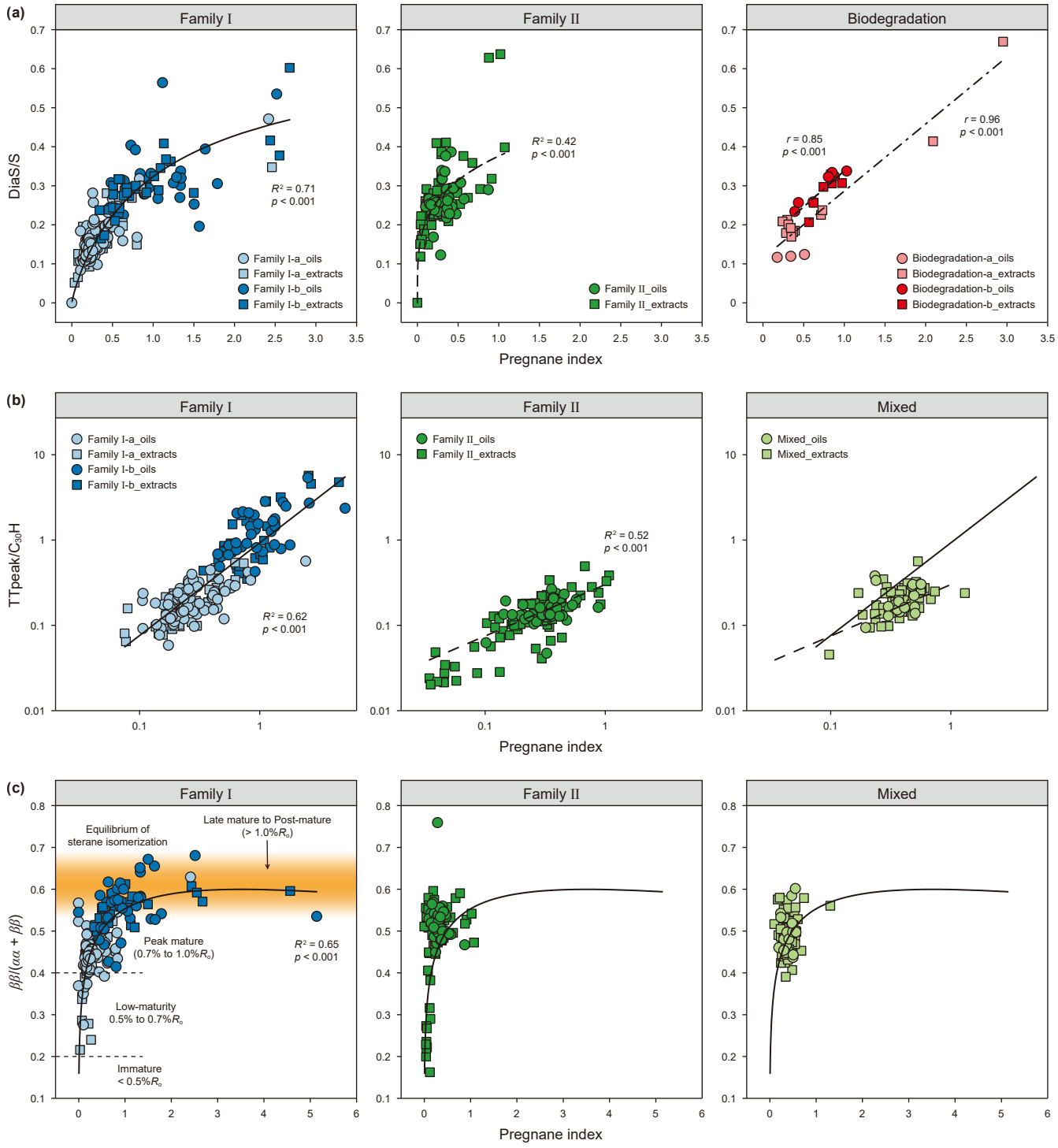

Fig. 4. Cross plots of geochemical proxies based on the genetic classification of hydrocarbons derived from molecular UMAP analysis. **(a)** Pregnane index versus DiaS/S; **(b)** Pregnane index versus TT/C_{30}H; **(c)** Pregnane index versus $\beta\beta/(\alpha\alpha+\beta\beta)$ stereoisomerization ratio.

exhibit intermediate compositional signatures, systematically distributed between the two dominant family trends (Fig. 4(b)).

When evaluating biomarker-based maturity indicators, C_{29} sterane stereoisomerization ratios [20S/(20S + 20R) and $\beta\beta/(\alpha\alpha+\beta\beta)$] demonstrate superior thermal specificity within specific windows of hydrocarbon generation (Seifert and Moldowan, 1986). However, their diagnostic utility becomes constrained beyond peak oil generation thresholds (ca. 1.0%R_o; Peters and Moldowan, 2017). Comparatively, cracking-derived parameters such as pregnane index and TT/C_{30}H maintain responsiveness across broader maturation ranges, albeit with increased susceptibility to non-thermal alteration processes (van Graas, 1990; Price, 1993). Importantly, the molecular UMAP framework overcomes traditional limitations by establishing high-fidelity correlations even with low-specificity biomarkers. This methodological advancement enables precise reconstruction of differential

petroleum evolution pathways, effectively decoupling superimposed geochemical effects through multidimensional pattern recognition.

For Family I, a statistically robust correlation between pregnane index and $\beta\beta/(\alpha\alpha+\beta\beta)$ emerges after eliminating source facies interference and secondary alteration effects (Fig. 4(c)). The pregnane index demonstrates particular efficacy in assessing higher maturity stages (R_o > 1.0%), where $\beta\beta/(\alpha\alpha+\beta\beta)$ stereoisomerization attains equilibrium (Fig. 4(c)). This diagnostic capability stems from the exceptional thermodynamical stability of pregnane. This stability allows it to become sustainedly enriched during thermal maturation, as alkyl side chains are cleaved from long-chain steranes. (Eglinton et al., 1988; Requejo et al., 1997; Tao et al., 2021). Supporting evidence from source rock pyrolysis experiments reveals pregnanes as the predominant residual steroids in post-mature stage pyrolysates (Wingert and Pomerantz, 1986; Wei et al., 2007). The dual-maturity indicator system reveals distinct applications: $\beta\beta/(\alpha\alpha+\beta\beta)$ ratio effectively constrains lower maturity ranges, while pregnane index becomes diagnostic at R_o > 1.0% (Fig. 4(c)). Visual inspection reveals clear differentiation between Family I-a (low-maturity to peak mature) and Family I-b (late mature to post-mature) within this parameter space, reflecting systematic thermal evolution trends (Fig. 4(c)).

Family II samples display a restricted maturity spectrum predominantly within the low-maturity to peak oil window, of which all oil samples are peak mature (Fig. 4(c)). The observed mixing patterns in pregnane index vs. $\beta\beta/(\alpha\alpha+\beta\beta)$ plot suggest petroleum mixing primarily occurred between peak mature members of two genetic families (Fig. 4(c)). This interpretation aligns with previous UMAP clustering results discussed in Section 4.2.2, reinforcing the consistency of our analytical framework.

4.4. Spatial distribution heterogeneity of petroleum categories

The five petroleum categories exhibit significant spatial partitioning characteristics. Family I demonstrates predominant occurrence within the eastern slope of Fukang Sag, Beisantai Uplift, and Fukang Fault Zone, while showing limited presence in the northern study area (Fig. 5(a)). Notably, the depositional patterns of Lucaogou source rocks display a distinct southward-thickening and northward-thinning configuration, spatially correlating with Family I distribution (Fig. 5(a)). This spatial correspondence suggests that Fukang Sag and Jimusar Sag constitute the principal hydrocarbon generation centers for the Lucaogou Fm. In recent years, significant exploration advances in the tight oils within the Lucaogou Fm. in the Jimusar Sag (Wang et al., 2020, 2023; Wu et al., 2021), as well as discoveries in the Permian petroleum system in the Fukang Sag/Fault Zone (Tao et al., 2012; Liu et al., 2023), have provided support for this perspective.

Furthermore, substantial variations are observed between the two subcategories of Family I. Geographically, the higher-maturity Family I-b primarily concentrates along the Fukang Sag slope, with limited occurrence in structural highs such as Beisantai Uplift and Fukang Fault Zone (Fig. 5(a)). Vertical distribution patterns reveal that the lower-maturity Family I-a is restricted to relatively shallow stratigraphic intervals, whereas Family I-b exhibits extensive accumulation across both deep and shallow reservoir units (Fig. 6). These distribution characteristics imply that vertical migration primarily controls the accumulation of Family I, while the burial-thermal evolution stage of Lucaogou strata governs the oil maturity within overlying reservoirs. Such an accumulation model is widely documented in the entire basin, such as the Mahu Sag area of northwestern Junggar Basin (Tao et al., 2021). From the tectonic uplift zones toward the center of sag, variations in thermal evolution degree of the source rocks control the zonal distribution of the derivative oils with differing maturities across the geographic extent.

Family II displays a distinct distribution pattern compared to Family I, with widespread occurrence across the northern region of the study area (Fig. 5(b)). Previous studies also suggest that the crude oils in these northern regions (e.g., the Baijiahai Uplift) were primarily sourced from the Jurassic strata (Chen et al., 2003b). However, the distribution of Family II becomes progressively restricted in southern regions, primarily concentrated along the Fukang Sag slope with negligible accumulation observed in eastern sectors such as Beisantai Uplift (Fig. 5(b)). Geochemical evidence identifies the organic-rich mudstone succession within the Lower Jurassic Badaowan Fm as the principal oil source for Jurassic coal-bearing sequences (Chen et al., 2003b). This source unit exhibits maximum thickness in the western study area, gradually thinning eastward (Fig. 5(b)). The observed Family II distribution pattern consequently reflects a pronounced east-west differentiation in Jurassic oil-generation potential. This geospatial relationship further implies potential undiscovered hydrocarbon accumulations derived from Jurassic source rocks in the underexplored central Fukang Sag and Dongdaohaizi Sag regions. Vertical distribution analysis indicates that Family II primarily accumulated at shallow to moderate depths, facilitating mixing processes with Family I hydrocarbons within equivalent stratigraphic intervals (Fig. 6).

Biodegradation processes are preferentially observed within tectonic uplift zones, particularly in the Beisantai Uplift and Fukang Fault Zone (Fig. 5(c), Fig. S4). Notably, the biodegradation of crude oils and the subsequent formation of secondary microbial gases in the Beisantai Uplift have been extensively documented in previous studies (Lu et al., 2015; Hou et al., 2021). These structural domains experienced intensive fault network development during Yanshanian-Himalayan tectonic movements, characterized by multiphase uplift-thrust activities (Fig. 5(d)). Vertical distribution analysis reveals that biodegraded oils among the five petroleum categories are basically exclusively confined to shallow reservoir intervals (Fig. 6). The well-developed fault systems serve dual functions: 1) enabling efficient vertical hydrocarbon migration through conductive pathways, and 2) establishing favorable hydrodynamic circulation regimes within clastic reservoirs. This geological configuration creates persistent low-temperature conditions combined with active groundwater systems in shallow stratigraphic units, collectively promoting extensive biodegradation of hydrocarbon accumulations in these areas.

Hydrocarbons exhibiting migration-induced phase fractionation demonstrate genetic linkage with fault system evolution in the study area. The regional fault architecture predominantly originated through three major tectonic episodes: Late Hercynian (Late Permian), Yanshanian (Late Jurassic-Early Cretaceous), and Himalayan movements. Crucially, phase fractionation phenomena are strictly confined to Yanshanian-age fault networks (Fig. 5(d)), indicating this tectonic phase as a pivotal event in oil charging and migration dynamics. The Yanshanian faults typically reactivate deeper Hercynian faults, forming complex fault architectures characterized by vertically superimposed Y-shaped configurations. This structural inheritance enabled the Yanshanian fault system to exert predominant control on vertical hydrocarbons migration. Moreover, the vertical distribution of phase-fractionated hydrocarbons (Fig. 6) provides insights for assessing both scale and extent of oil migration-charging processes within the study area.

4.5. Practical implications for petroleum exploration

Through systematic genetic identification of hydrocarbon samples in this study, we have established statistically robust

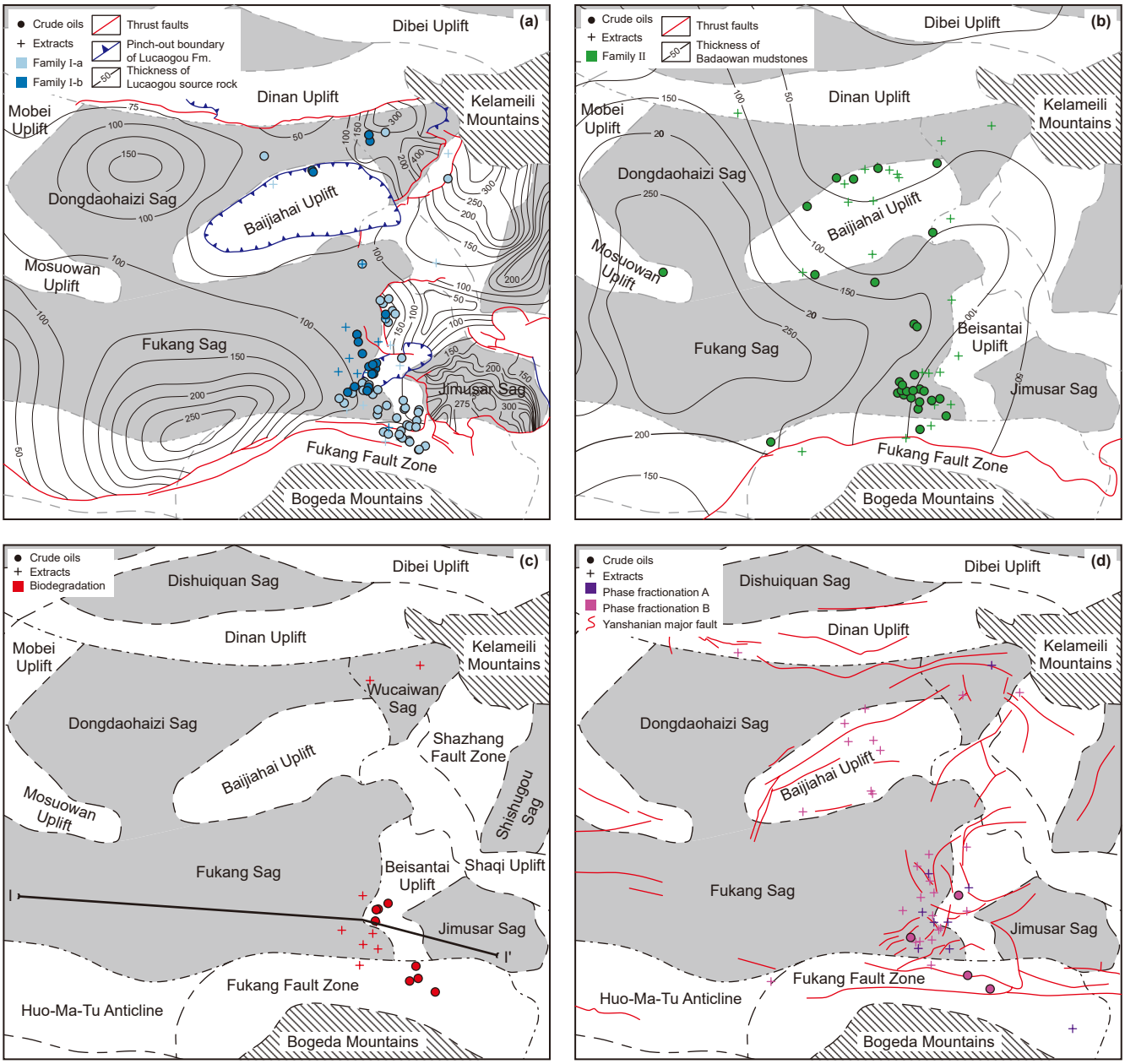

Fig. 5. Spatial occurrences of different petroleum categories identified through molecular UMAP analysis. **(a)** Family I, including subcategories Family I-a and Family I-b; **(b)** Family II; **(c)** Biodegraded oils and extracts; **(d)** Phase fractionation phenomena vs. the distributions of Yanshanian-age fault networks.

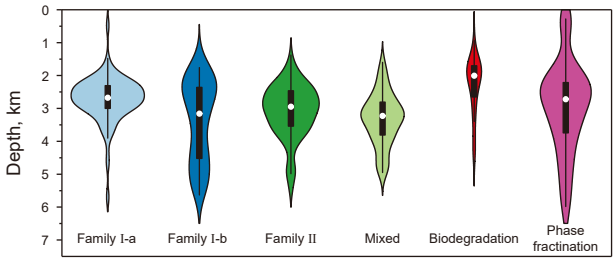

Fig. 6. Comparison of the vertical distribution extent for different petroleum categories.

geochemical features with exploration implications. The two oil families exhibit marked disparities in physical properties attributable to divergent source biomaterial assemblages. For Lucaogou-derived crude oils (Family I), low-maturity members display elevated densities reaching 1.0 g/cm^3 (Fig. 7(a)). Progressive thermal maturation drives density reduction in this family, with late mature to post-mature phases showing significant lightning (0.72–0.87 g/cm^3; Fig. 7(a)). In contrast, Jurassic-sourced oils (Family II) predominantly represent peak maturity stages, exhibiting systematically lower densities (0.75–0.88 g/cm^3) compared to Lucaogou counterparts at equivalent maturation levels. Notably, this density range overlaps with that of late mature to post-mature Lucaogou oils (Fig. 7(b)). Physical property of mixed oils reveals that oil mixing predominantly occurs between Family II and Family I-a, as evidenced by the characteristic density distribution pattern (Fig. 7(b)). This phenomenon is likely attributed to the vertical juxtaposition of these two members in the depth profile, facilitating extensive mixing.

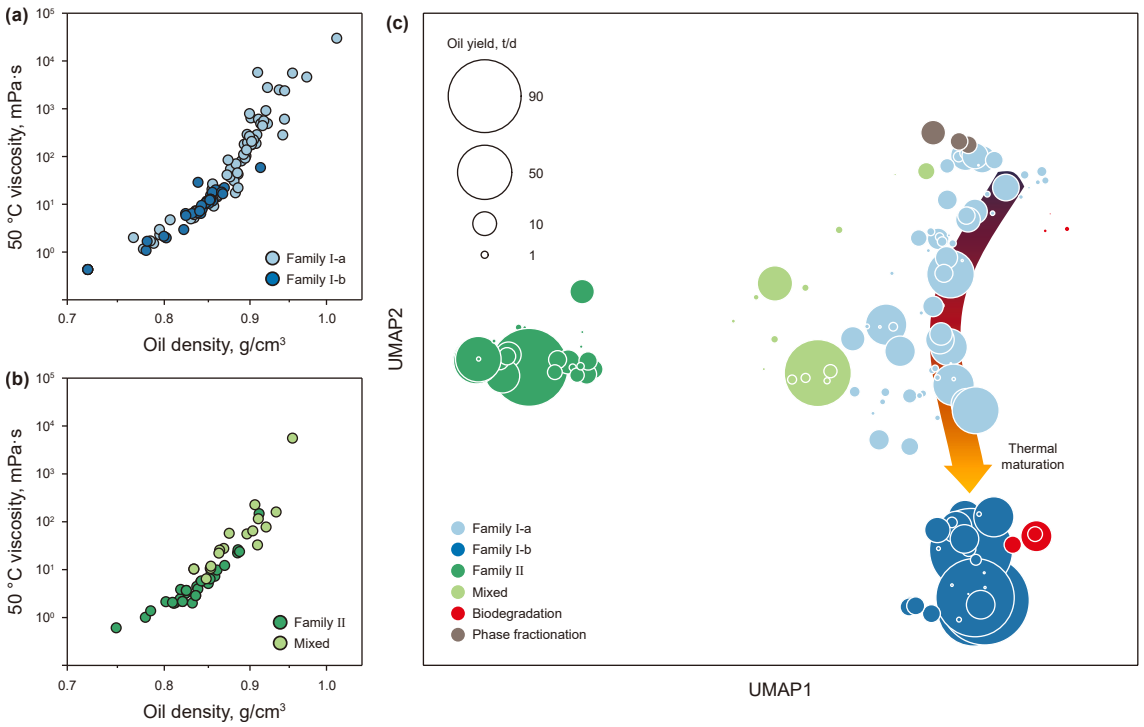

Fig. 7. **(a, b)** Evolution trajectories of physical properties with thermal maturation regarding two genetically discrete oil families and the mixed member; **(c)** Expression of oil production data on the UMAP framework, showing the controls of source biomaterials, thermal maturity and secondary alteration.

Our sample-constrained analysis of oil yield distributions reveals distinct generation patterns across oil families. For Family I-a, the Lucaogou Fm exhibits relatively limited petroleum generation during low-maturity phases. While oil productivity undergoes marked escalation with advancing thermal evolution, particularly during peak maturation (Fig. 7(c)). For Family II, the Jurassic Fm demonstrate broad comparability with Family I-a in oil generation characteristics, although notable productivity anomalies emerge in specific structural domains (Fig. 7(c)). Notably, late to post-mature phases of the Lucaogou Fm (Family I-b) demonstrate substantially enhanced oil generation capacity that far more exceeds those of both Family I-a and Family II (Fig. 7(d)).

Considering the spatial distribution patterns of these oil families discussed earlier, there remains substantial underexploited potential for high-maturity Lucaogou -derived oils in central Fukang Sag's mid-deep zones. The continuous stratigraphic-lithologic traps for this resource type warrant prioritized investigation. The lower viscosity characteristics of these high-maturity oils theoretically enhance reservoir connectivity, favoring large-scale accumulation. Concomitantly, Jurassic oils present prospective targets in shallower sequences across both Fukang and Dongdaohaizi sags, particularly where structural-lithologic configurations optimize migration pathways.

4.6. Broader applicability and transferability of the framework

The powerful capability of UMAP in disentangling complex geological influences, as demonstrated in this study for the hybrid petroleum system of eastern Junggar Basin, is not limited to such scenarios. It is broadly applicable to various complex systems characterized by distinct geological characteristics. A compelling case in point is the extensive areas involving central and western Junggar Basin (Tao et al., 2025). In this complex system, UMAP was successfully employed to decipher the complex effects of source disparities, mixing, thermal maturation, biodegradation, and evaporative fractionation on crude oil compositions. These applications across diverse geological settings underscore the method's robustness in addressing a wide spectrum of geological complexities.

A common concern regarding ML approaches is their reliance on large datasets. However, as an unsupervised dimensionality reduction tool, the validity of UMAP's output depends more on the representativeness of the samples than on their absolute number. Its primary value lies in uncovering the intrinsic topological structure of the data. For a geologically well-constrained sub-sag or localized region, a smaller but representative dataset (e.g., 30–50 samples encompassing key geological end-members) could be sufficient for UMAP to reveal meaningful patterns and relationships. While a larger sample size yields a smoother manifold and sharper cluster boundaries, it does not fundamentally alter the underlying data structure that UMAP captures.

When applying this framework to studies with limited sample availability, parameter tuning becomes particularly important. The key parameter n_neighbors, which controls the balance between capturing local and global structure, can be optimized for smaller datasets. We recommend using a smaller n_neighbors value (e.g., 5–15) for smaller datasets. This adjustment forces the algorithm to focus on finer-grained, local structures, effectively preventing over-smoothing and helping to reveal robust patterns even when data points are sparse. Therefore, through appropriate parameter tuning, the UMAP framework maintains its analytical power and is fully applicable to smaller-scale studies.

5. Conclusions

The genetic origins and spatial distribution of hydrocarbon fluids in sedimentary basins remain inadequately constrained due to limited reliable prior knowledge in existing geological frameworks. Therefore, this study employs unsupervised learning methodologies to extract the critical genetic information encoded

within hydrocarbon molecular fingerprints. UMAP emerges as an effective manifold learning technique. It excels at preserving both global data structures and local topological relationships within high-dimensional geochemical datasets. This approach can systematically disentangle the nonlinear evolutionary patterns in hydrocarbon molecular composition that arise from complex, superimposed geological processes.

Through comprehensive UMAP analysis, we successfully delineate the hydrocarbon genetic types and evolutionary pathways within the hybrid petroleum system of the eastern Junggar Basin. Two primary oil families, namely Permian-derived and Jurassic-derived oils, were identified. Additionally, their post-depositional processes were effectively discriminated, including thermal maturation, mixing, biodegradation, and phase fractionation induced by migration. This methodological advancement facilitates the decoupling of interdependent geochemical signals. As a result, it establishes robust correlations between molecular fingerprints and specific geological drivers. By spatially mapping hydrocarbon genetic types and evolutionary features across the study area, our findings provide critical insights into the accumulation patterns of various resource types. This work also establishes a theoretical basis for identifying high-potential exploration targets.

CRediT authorship contribution statement

Ke-Yu Tao: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Jian Cao:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Conceptualization. **Yu-Ce Wang:** Validation, Resources, Investigation, Funding acquisition. **Wan-Yun Ma:** Resources, Investigation.

Declaration of competing interest

We have no known competing financial interests that could have appeared to influence the work reported in this manuscript.

Acknowledgements

We thank the technical staff from the Research Institute of Petroleum Exploration and Development of PetroChina Xinjiang Oilfield Company for their cooperation in completing the work, and Research Institute of Experimental and Testing for cooperation in conducting geochemical analyses. This work was jointly funded by the National Natural Science Foundation of China (Grant Nos. 42203033 and 42302137).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petsci.2025.12.029.

References

Alexandrino, G.L., Prata, P.S., Augusto, F., 2016. Discriminating lacustrine and marine organic matter depositional paleoenvironments of Brazilian crude oils using comprehensive two-dimensional gas chromatography–quadrupole mass spectrometry and supervised classification chemometric approaches. Energy Fuels. 31, 170–178. https://doi.org/10.1021/acs.energyfuels.6b01925.

Cai, Z.X., Chen, F.J., Jia, Z.Y., 2000. Types and tectonic evolution of Junggar Basin. Earth Sci. Front. 7, 431–440 (in Chinese).

Carroll, A.R., Yunhai, L., Graham, S.A., Xuchang, X., Hendrix, M.S., Jinchi, C., McKnight, C.L., 1990. Junggar basin, northwest China: trapped late Paleozoic Ocean. Tectonophysics 181, 1–14. https://doi.org/10.1016/0040-1951(90)90004-R.

Chen, F.J., Wang, X.W., Wang, X.W., 2005. Prototype and tectonic evolution of the Junggar Basin, northwestern China. Earth Sci. Front. 12, 77–89 (in Chinese).

Chen, J., Liang, D., Wang, X., Zhong, N., Song, F., Deng, C., Shi, X., Jin, T., Xiang, S., 2003a. Mixed oils derived from multiple source rocks in the Cainan oilfield, Junggar Basin, Northwest China. Part I: genetic potential of source rocks, features of biomarkers and oil sources of typical crude oils. Org. Geochem. 34, 889–909. https://doi.org/10.1016/S0146-6380(03)00030-5.

Chen, J., Deng, C., Liang, D., Wang, X., Zhong, N., Song, F., Shi, X., Jin, T., Xiang, S., 2003b. Mixed oils derived from multiple source rocks in the Cainan oilfield, Junggar Basin, Northwest China. Part II: artificial mixing experiments on typical crude oils and quantitative oil-source correlation. Org. Geochem. 34, 911–930. https://doi.org/10.1016/S0146-6380(03)00031-7.

Curiale, J.A., 2008. Oil-source rock correlations – Limitations and recommendations. Org. Geochem. 39, 1150–1161. https://doi.org/10.1016/j.orggeochem.2008.02.001.

Curry, D.J., 2019. Future directions in basin and petroleum systems modeling: A survey of the community. AAPG Bull. 103, 2285–2293. https://doi.org/10.1306/1208171615217152.

Ding, W., Hou, D., Jiang, L., Jiang, Y., Wu, P., 2020. High abundance of carotanes in the brackish-saline lacustrine sediments: a possible Cyanobacteria source? Int. J. Coal Geol. 219, 103373. https://doi.org/10.1016/j.coal.2019.103373.

Eglinton, T.I., Douglas, A.G., Rowland, S.J., 1988. Release of aliphatic, aromatic and sulphur compounds from Kimmeridge kerogen by hydrous pyrolysis: A quantitative study. Org. Geochem. 13, 655–663. https://doi.org/10.1016/0146-6380(88)90086-1.

Feng, Y., Coleman, R.G., Tilton, G., Xiao, X., 1989. Tectonic evolution of the west Junggar region, Xinjiang, China. Tectonics 8, 729–752. https://doi.org/10.1029/TC008i004p00729.

Fu, J.M., Sheng, G.Y., Xu, J.Y., Eglinton, G., Gowar, A.P., Jia, R.F., Fan, S.F., Peng, P.A., 1986. Peculiarities of salt lake sediments as potential source rocks in China. Org. Geochem. 10, 119–126. https://doi.org/10.1016/0146-6380(86)90015-X.

Grice, K., Schouten, S., Peters, K.E., Sinninghe Damsté, J.S., 1998. Molecular isotopic characterisation of hydrocarbon biomarkers in Palaeocene-Eocene evaporitic, lacustrine source rocks from the Jianghan Basin, China. Org. Geochem. 29, 1745–1764. https://doi.org/10.1016/S0146-6380(98)00075-8.

Han, Y., Mahlstedt, N., Horsfield, B., 2015. The Barnett Shale: compositional fractionation associated with intraformational petroleum migration, retention, and expulsion. AAPG Bull. 99, 2173–2202. https://doi.org/10.1306/06231514113.

Han, Y.G., Zhao, G.C., 2018. Final amalgamation of the Tianshan and Junggar orogenic collage in the southwestern Central Asian Orogenic belt: constraints on the closure of the Paleo-Asian Ocean. Earth Sci. Rev. 186, 129–152. https://doi.org/10.1016/j.earscirev.2017.09.012.

Hou, M.G., Zha, M., Ding, X.J., Yin, H., Bian, B.L., Liu, H.L., Jiang, Z.F., 2021. Source and accumulation process of Jurassic biodegraded oil in the Eastern Junggar Basin, NW China. Pet. Sci. 18, 1033–1046. https://doi.org/10.1016/j.petsci.2021.07.010.

Hughes, W.B., Holba, A.G., Dzou, L.I.P., 1995. The ratios of dibenzothiophene and pristane to phytane as indicators of depositional environment and lithology of petroleum source rocks. Geochem. Cosmochim. Acta 59, 3581–3598. https://doi.org/10.1016/0016-7037(95)00225-O.

Li, S.L., Yu, X.H., Tan, C.P., Steel, R., 2014. Jurassic sedimentary evolution of southern Junggar Basin: Implication for palaeoclimate changes in northern Xinjiang Uygur Autonomous Region, China. J. Palaeogeogr. 3, 145–161. https://doi.org/10.3724/SP.J.1261.2014.00049.

Liu, D., Fan, Q., Zhang, C., Gao, Y., Du, W., Song, Y., Zhang, Z., Luo, Q., Jiang, Z., Huang, Z., 2022. Paleoenvironment evolution of the Permian Lucaogou Formation in the southern Junggar Basin, NW China. Palaeogeogr. Palaeoclimatol. Palaeoecol. 603, 111198. https://doi.org/10.1016/j.palaeo.2022.111198.

Liu, D., Wang, Y., Yang, H., Li, S., Liu, C., Han, Y., Chen, M., 2023. Genesis types and distribution of crude oil in Fukang Sag and its peripheral bulges, Junggar Basin. China Pet. Explor. 28, 94–107. https://doi.org/10.3969/j.issn.1672-7703.2023.01.009.

Liu, N., Zhang, Z., Zhang, H., Wang, Z., Gao, J., Liu, R., Zhang, N., 2024. Multiple-frequency attribute blending via adaptive uniform manifold approximation and projection and its application on hydrocarbon reservoir delineation. Geophysics 89, WA195–WA206. https://doi.org/10.1190/geo2023-0111.1.

Lu, J., Wang, L., Chen, S., Han, H., Zhang, H., Huang, Y., He, X., Zhan, P., Zhou, S., Zhang, A., Li, X., 2015. Features and origin of oil degraded gas of Santai field in Junggar Basin, NW China. Petrol. Explor. Dev. 42, 466–474. https://doi.org/10.1016/S1876-3804(15)30039-2.

Lu, M., Duan, G., Zhang, T., Liu, N., Song, Y., Zhang, Z., Qiao, J., Wang, Z., Fang, Z., Luo, Q., 2025. Influences of paleoclimatic changes on organic matter enrichment mechanisms in freshwater and saline lacustrine oil shales in China: a machine learning approach. Earth Sci. Rev. 262, 105061. https://doi.org/10.1016/j.earscirev.2025.105061.

Luo, G., Yang, H., Algeo, T.J., Hallmann, C., Xie, S., 2019. Lipid biomarkers for the reconstruction of deep-time environmental conditions. Earth Sci. Rev. 189, 99–124. https://doi.org/10.1016/j.earscirev.2018.03.005.

Magoon, L.B., 2004. In: Cleveland, C., Ayres, R.U. (Eds.), Petroleum System—nature's Distribution System of Oil and Gas. Encyclopedia of Energy. Elsevier Academic Press, Amsterdam, pp. 823–836.

Moldowan, J.M., Seifert, W.K., Gallegos, E.J., 1985. Relationship between petroleum composition and depositional environment of petroleum source rocks. AAPG Bull. 69, 1255–1268.

Murray, A.P., Peters, K.E., 2021. Quantifying multiple source rock contributions to petroleum fluids: Bias in using compound ratios and neglecting the gas fraction. AAPG Bull. 105, 1661–1678. https://doi.org/10.1306/03122120056.

Naafs, B.D.A., Inglis, G.N., Blewett, J., McClymont, E.L., Lauretano, V., Xie, S., Evershed, R.P., Pancost, R.D., 2019. The potential of biomarker proxies to trace climate, vegetation, and biogeochemical processes in peat: A review. Global Planet. Change 179, 57–79. https://doi.org/10.1016/j.gloplacha.2019.05.006.

Noble, R.A., Alexander, R., Kagi, R.I., Knox, J., 1985a. Tetracyclic diterpenoid hydrocarbons in some Australian coals, sediments and crude oils. Geochem. Cosmochim. Acta 49, 2141–2147. https://doi.org/10.1016/0016-7037(85)90072-9.

Noble, R., Alexander, R., Kagi, R.I., 1985b. The occurrence of bisnorhopane, trisnorhopane and 25-norhopanes as free hydrocarbons in some Australian shales. Org. Geochem. 8, 171–176. https://doi.org/10.1016/0146-6380(85)90035-X.

Peters, K.E., Fowler, M.G., 2002. Applications of petroleum geochemistry to exploration and reservoir management. Org. Geochem. 33, 5–36. https://doi.org/10.1016/S0146-6380(01)00125-5.

Peters, K.E., Moldowan, J.M., 2017. Biomarkers: Assessment of source rock thermal maturity. In: Sorkhabi, R. (Ed.), Encyclopedia of Petroleum Geoscience. Springer, Cham. https://doi.org/10.1007/978-3-319-02330-4_10-1.

Peters, K.E., Walters, C.C., Moldowan, J.M., 2005. In: The Biomarker Guide: Interpreting Molecular Fossils in Petroleum and Ancient Sediments, second ed. Cambridge University Press.

Price, L.C., 1993. Thermal stability of hydrocarbons in nature: limits, evidence, characteristics, and possible controls. Geochem. Cosmochim. Acta 57, 3261–3280. https://doi.org/10.1016/0016-7037(93)90539-9.

Qian, Y., Zhang, T., Wang, Z., Tuo, J., Zhang, M., Wu, C., Tian, C., 2018. Organic geochemical characteristics and generating potential of source rocks from the Lower-middle Jurassic coal-bearing strata in the East Junggar Basin, NW China. Mar. Petrol. Geol. 93, 113–126. https://doi.org/10.1016/j.marpetgeo.2018.02.036.

Requejo, A.G., Allan, J., Creaney, S., Gray, N.R., Cole, K.S., 1997. Short-chain (C_{21} and C_{22}) diasteranes in petroleum and source rocks as indicators of maturity and depositional environment. Geochem. Cosmochim. Acta 61, 2653–2667. https://doi.org/10.1016/S0016-7037(97)00106-3.

Seifert, W.K., Moldowan, J.M., 1979. The effect of biodegradation on steranes and terpanes in crude oils. Geochem. Cosmochim. Acta 43, 111–126. https://doi.org/10.1016/0016-7037(79)90051-6.

Seifert, W.K., Moldowan, J.M., 1986. Use of biological markers in petroleum exploration. Methods Geochem. Geophys. 24, 261–290.

Snodgrass, J.E., Milkov, A.V., 2020. Web-based machine learning tool that determines the origin of natural gases. Comput. Geosci. 145, 104595. https://doi.org/10.1016/j.cageo.2020.104595.

Su, K., Xu, Y., Luo, Q., Liu, Y., Li, Y., Yan, G., 2025. Mini-review on petroleum molecular geochemistry: Opportunities with digitalization, machine learning, and artificial intelligence. Energy Fuels. 39, 5034–5050. https://doi.org/10.1021/acs.energyfuels.4c05402.

Suchý, V., Dobeš, P., Sýkorová, I., Machovič, V., Stejskal, M., Kroufek, J., Matysová, P., 2010. Oil-bearing inclusions in vein quartz and calcite and bitumens in veins: testament to multiple phases of hydrocarbon migration in the Barrandian Basin (Lower Palaeozoic), Czech Republic. Mar. Petrol. Geol. 27, 285–297. https://doi.org/10.1016/j.marpetgeo.2009.08.017.

Tao, K.Y., Cao, J., Wang, Y.C., Ma, W., 2025. Disentangling and interpreting nonlinear molecular and isotopic variations in petroleum using machine learning. Mar. Petrol. Geol. 171, 107175. https://doi.org/10.1016/j.marpetgeo.2024.107175.

Tao, K.Y., Cao, J., Wang, Y.C., Xiang, B., Bian, C., Hu, W., 2021. Petroleum system for the continuous oil play in the lacustrine Lower Triassic, Junggar Basin, China. AAPG Bull. 105, 2349–2380. https://doi.org/10.1306/07022119211.

Tao, K.Y., Cao, J., Wang, Y.C., Mi, J.L., Ma, W.Y., Shi, C.H., 2020. Chemometric classification of crude oils in complex petroleum systems using t-distributed stochastic neighbor embedding machine learning algorithm. Energy Fuels. 34, 5884–5889. https://doi.org/10.1021/acs.energyfuels.0c01333.

Tao, S., Wang, Y., Tang, D., Wu, D., Xu, H., He, W., 2012. Organic petrology of Fukang Permian Lucaogou formation oil shales at the northern foot of Bogda Mountain, Junggar Basin, China. Int. J. Coal Geol. 99, 27–34. https://doi.org/10.1016/j.coal.2012.05.001.

Van Graas, G.W., 1990. Biomarker maturity parameters for high maturities: Calibration of the working range up to the oil/condensate threshold. Org. Geochem. 16, 1025–1032. https://doi.org/10.1016/0146-6380(90)90139-Q.

van Graas, G.W., Gilje, A.E., Isom, T.P., Tau, L.A., 2000. The effects of phase fractionation on the composition of oils, condensates and gases. Org. Geochem. 31, 1419–1439. https://doi.org/10.1016/S0146-6380(00)00128-5.

Volk, H., Mann, U., Burde, O., Horsfield, B., Suchý, V., 2000. Petroleum inclusions and residual oils: Constraints for deciphering petroleum migration. J. Geochem. Explor. 69, 595–599. https://doi.org/10.1016/S0375-6742(00)00156-4.

Wang, X., Zhi, D., Wang, Y., Chen, J., Qin, Z., Liu, D., Xiang, Y., Lan, W., Li, N., 2013. Geochemistry of Source Rock and Petroleum in the Junggar Basin. Petroleum Industry Press, Beijing, pp. 1–565.

Wang, Y., Cao, J., Tao, K., Li, E., Ma, C., Shi, C., 2020. Reevaluating the source and accumulation of tight oil in the middle Permian Lucaogou Formation of the Junggar Basin, China. Mar. Petrol. Geol. 117, 104384. https://doi.org/10.1016/j.marpetgeo.2020.104384.

Wang, Y., Cao, J., Tao, K., Zhang, C., Xiang, B., Li, E., Pan, C., 2023. Origin of heavy shale oil in saline lacustrine basins: insights from the Permian Lucaogou formation, Junggar Basin. AAPG Bull. 107, 1553–1579. https://doi.org/10.1306/10242222027.

Wei, Z.B., Moldowan, J.M., Jarvie, D.M., Hill, R., 2006. The fate of diamondoids in coals and sedimentary rocks. Geology 34, 1013–1016. https://doi.org/10.1130/G22840A.1.

Wei, Z.B., Moldowan, J.M., Zhang, S.C., Hill, R., Jarvie, D.M., Wang, H.T., Song, F.Q., Fago, F., 2007. Diamondoid hydrocarbons as a molecular proxy for thermal maturity and oil cracking: Geochemical models from hydrous pyrolysis. Org. Geochem. 38, 227–249. https://doi.org/10.1016/j.orggeochem.2006.09.011.

Wenger, L.M., Davis, C.L., Isaksen, G.H., 2002. Multiple controls on petroleum biodegradation and impact on oil quality. SPE Reservoir Eval. Eng. 5, 375–383. https://doi.org/10.2118/80168-PA.

Wingert, W.S., Pomerantz, M., 1986. Structure and significance of some twenty-one and twenty-two carbon petroleum steranes. Geochem. Cosmochim. Acta 50, 2763–2769. https://doi.org/10.1016/0016-7037(86)90225-5.

Wu, A., Cao, J., Zhang, J., Wu, T., Wang, Y., 2022. Origin of microbial–hydrothermal bedded dolomites in the Permian Lucaogou formation lacustrine shales, Junggar Basin, NW China. Sediment. Geol. 440, 106260. https://doi.org/10.1016/j.sedgeo.2022.106260.

Wu, H., Hu, W., Wang, Y., Tao, K., Tang, Y., Cao, J., Kang, X., 2021. Depositional conditions and accumulation models of tight oils in the middle Permian Lucaogou Formation in Junggar Basin, northwestern China: New insights from geochemical analysis. AAPG Bull. 105, 2477–2518. https://doi.org/10.1306/06222118094.

Xia, L., Cao, J., Hu, W., Tang, Y., Zhang, C., He, W., 2023. Paleo-environmental conditions and organic carbon accumulation during glacial events: new insights from saline lacustrine basins. Global Planet. Change 227, 104162. https://doi.org/10.1016/j.gloplacha.2023.104162.

Zhang, Y.D., Sun, Y.G., Liu, Q., 2021. Distribution and carbon isotope composition of pregnane in carbonate-evaporitic rocks from the Bonan Sag, Bohai Bay Basin, Eastern China: Insights into sources and associated lake environments. Org. Geochem. 151, 104127. https://doi.org/10.1016/j.orggeochem.2020.104127.

Zhang, Z., Liu, N., Liu, R., Lu, M., Wei, T., Gao, J., 2024. Adaptive multifrequency attribute analysis and its application on reservoir characterization. IEEE Trans. Geosci. Rem. Sens. 62, 1–10. https://doi.org/10.1109/TGRS.2024.3373393.