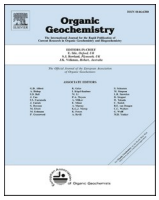
Organic Geochemistry 160 (2021) 104298

Contents lists available at ScienceDirect

Organic Geochemistry

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



Diamondoids in petroleum: Their potential as source and maturity indicators

Wenmin Jiang [a,b], Yun Li [a,b,*], Chenchen Fang [c], Zhiqiang Yu [a,b], Yongqiang Xiong [a,b,*]

[a] State Key Laboratory of Organic Geochemistry (SKLOG), Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, PR China
[b] CAS Center for Excellence in Deep Earth Science, Guangzhou 510640, PR China
[c] PetroChina Research Institute of Petroleum Exploration and Development, Beijing 100083, PR China

ARTICLE INFO

Associate Editor—**Clifford Walters**

Keywords:
Petroleum
Diamondoids
Source type
Thermal maturity
Multivariate statistical analysis

ABSTRACT

Assessing the origins and thermal maturity of oils is an important issue in applied petroleum geochemistry. In this study, we extracted information regarding the sources and thermal maturity of oils from diamondoid data using multivariate statistical analysis based on four series of thermal simulation experiments conducted on Type I, II, and III kerogens and a crude oil. Using principal component analysis (PCA) and discriminant analysis (DA), we established the relationship between diamondoid indices and kerogen type, which was then used to identify the source of oils in some oil fields in China. PCA and regression analysis (RA) were used to determine the quantitative relationship between diamondoid indices and thermal maturity, which was then used to estimate the thermal maturity of the studied oils. Our results indicate that this approach is effective for determining the source and maturity of oils, particularly for condensates.

1. Introduction

Diamondoids are a group of rigid fused-ring alkanes with diamond-like structures that have been widely detected in petroleum and sediment extracts (Landa and Machacek, 1933; Williams et al., 1986; Wingert, 1992; Chen et al., 1996; Dahl et al., 1999, 2003; Schulz et al., 2001; Wei et al., 2006; Gentzis and Carvajal-Ortiz, 2018; Scarlett et al., 2019; Spaak et al., 2020; Atwah et al., 2021b; Botterell et al., 2021; Forkner et al., 2021). Due to their strong resistance to thermal degradation and biodegradation, diamondoids have great potential for determining the thermal maturity of source rocks and oils (Chen et al., 1996; Zhang et al., 2005; Wei et al., 2007b; Gentzis and Carvajal-Ortiz, 2018), establishing oil–oil and oil–source rock correlations (Sassen and Post, 2008; Moldowan et al., 2015; Spaak et al., 2020; Atwah et al., 2021a, b; Botterell et al., 2021; Forkner et al., 2021), estimating the extent of oil cracking (Dahl et al., 1999), and evaluating the degree of biodegradation of oils (Grice et al., 2000; Wei et al., 2007a). Although various diamondoid indices, including absolute concentrations, concentration ratios, and $\delta^{13}C$ values, have been used in previous studies, their application is not as universal and effective as might be expected. This may be due to: (1) the lack of a universal calibrated relationship between diamondoid indices and thermal maturity; (2) a limited understanding of the effects of source on the composition and distribution of diamondoids; and (3) the superposition of multiple factors that complicates the interpretation of diamondoid data.

Concentrations of individual diamondoids and the distribution of homologues and isomers can vary widely in natural samples. Diamondoids in oils include adamantanes, diamantanes, triamantanes, tetramantanes, pentamantanes, hexamantanes and so forth (Dahl et al., 2003; Moldowan et al., 2015) and 32 or more individual compounds can be detected in a typical analysis (Fang et al., 2012, 2013; Liang et al., 2012; Zhu et al., 2013). However, most of this information is not used in many geochemical studies, which typically report only few parameters, such as the concentration of 4- + 3-methyldiamantane (Dahl et al., 1999) and ratios of methyladamantanes/adamantane (MAs/A), methyldiamantanes/diamantane (MDs/D), the methyl adamantane index (MAI; 1-MA/[1-MA + 2-MA]) and methyl diamantane index (MDI; 4-MD/[4-MD + 1-MD + 3-MD]). This can be attributed to the fact that several factors can affect the composition and distribution of diamondoids in natural petroleum or source rock samples that hinder easy interpretation of diamondoid data. For example, some diamondoid indices, which are usually used as maturity indicators, can also be substantially affected by source rock facies (Schulz et al., 2001), cracking (Wei et al., 2006), mixing (Jiang et al., 2020), and

* Corresponding authors at: State Key Laboratory of Organic Geochemistry (SKLOG), Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, PR China.
E-mail addresses: liyun@gig.ac.cn (Y. Li), xiongyq@gig.ac.cn (Y. Xiong).

https://doi.org/10.1016/j.orggeochem.2021.104298
Received 1 June 2021; Received in revised form 31 July 2021; Accepted 2 August 2021
Available online 8 August 2021

biodegradation (Cheng et al., 2018). Most studies only take the influence of a single factor into consideration and to a large extent, the choice of diamondoid indices depends on the research objectives, which may be unsuitable for natural samples that have been affected by multiple factors. Therefore, making full use of diamondoid data and evaluating the effects of different factors on these data remains challenging.

Multivariate statistical analysis is a powerful data processing approach that has been used to extract useful information from large datasets in many fields (e.g., finance, biology, geoscience, forensic science, and chemistry; Komura et al., 2005; Kujawinski et al., 2009; Banas et al., 2010; Hur et al., 2010; Härdle and Simar, 2015). For example, principal component analysis (PCA) is one of the most common types of multivariate statistical analysis and can simplify a dataset and retain most of the original information by reducing the dimensions of the observations through linear combinations of original variables with the largest variances. This is usually used to extract the correlations of variables and group the observations. Discriminant analysis (DA) is used to establish an allocation rule from known groups of objects and allocate a new observation involving the same variables into one of the groups. Regression analysis (RA) determines correlations between dependent and independent variables in a dataset and establishes regression equations with good correlation, which can then be used to quantitatively predict the dependent variable based on new independent variables. These statistical methods are used in this study to analyze diamondoid data.

It is difficult to extract the influence of a specific factor on diamondoid data for natural samples. However, laboratory simulation experiments can solve this problem by constraining the reaction conditions. Given that source and thermal maturity are the two main factors that could influence the composition and distribution of diamondoids in petroleum, the simulation experiments were designed with a focus on these two factors. Four series of thermal simulation experiments were conducted to model the formation and evolution of diamondoids from three different types of kerogen and a crude oil in our previous studies (Fang et al., 2012; Jiang et al., 2018). The diamondoid data obtained from these simulation experiments are used as a calibration dataset to establish the source and thermal maturity determination models using multivariate statistical methods. Thus, the main objective of this study is to develop diamondoid-based models to discriminate oil sources by calibrating the relationship between diamondoid indices and oil source types, and to determine thermal maturities of oils by establishing the quantitative relationship between diamondoid indices and oil maturities. Finally, the developed models are applied to natural oil samples collected from different basins to assess their effectiveness. This is the first study to undertake both oil source determination and thermal maturity evaluation using the same diamondoid dataset.

2. Materials and methods

2.1. Simulation samples

The diamondoid data of thermal simulation samples were taken from our previous simulation experiments undertaken on different types of organic matter (i.e., Type I kerogen, Type II kerogen, and Type III kerogen, and a normal crude oil; Fang et al., 2012; Jiang et al., 2018). The thermal maturity of the samples was calculated based on the temperature and time conditions of the experiments using the EasyRo% method proposed by Sweeney and Burnham (1990). Thus, these samples have a known source type and thermal maturity.

2.2. Oil sample pretreatment and diamondoid determination

Diamondoids in 67 natural oils collected from three typical basins in China (i.e., the Tarim, Junggar, and Pear River Mouth basins) were analyzed in the present study. All the studied oil samples were carefully selected and were considered to be from a single source and free of secondary alteration. For example, the oils from the Tazhong Uplift were typical Cambrian–Lower Ordovician sourced marine oils within the Tarim Basin (Li et al., 2018 and References therein); the oils from the central Junggar Basin were generated mainly from Permian source rocks in the Pen 1 Jingxi Depression (Jiang et al., 2018 and References therein). Preparation of the oil samples involved: (1) ~50 mg of oil was first dissolved in isooctane in a 4 mL vial, and then 50 μL of an internal standard solution of *n*-dodecane-d_{26} and *n*-hexadecane-d_{34} in isooctane was injected into the sample vial; (2) after 10 min of ultrasonic treatment to dissolve any material, the vial was placed in a centrifuge for 10 min to separate asphaltenes; and (3) the resulting supernatant was transferred to a 2 mL vial for diamondoid detection. The sample information is provided in Supplementary Table S1.

The diamondoids were identified and analyzed by gas chromatography triple quadrupole mass spectrometry (GC–MS-MS; Thermo Fisher TSQ Quantum XLS) following the method of Liang et al. (2012). In brief, a 1 μL aliquot of each sample was injected into the GC system with an AS3000 autosampler. The GC instrument was equipped with a PTV injector, and the PTV splitless mode was used for the analyses, during which the inlet temperature was 300 °C and the split flow was 15 mL/min after 1 min of splitless flow. Helium (99.999% purity) was used as the carrier gas at a constant flow rate of 1.5 mL/min. The diamondoids were separated on a DB-1 fused silica capillary column (50 m × 0.32 mm i.d. × 0.52 μm film thickness). The GC oven temperature was held at 50 °C for 2 min, then increased to 80 °C at 15 °C/min, 250 °C at 2.5 °C/min, and to 300 °C at 15 °C/min, and was finally held at 300 °C for 10 min. Quantification of diamondoid compounds was undertaken by comparing the peak areas for unknowns with two internal standards (i.e., *n*-dodecane-d_{26} for adamantanes and *n*-hexadecane-d_{34} for diamantanes) in selected reaction monitoring (SRM) mode. The response factors of different types of diamondoids relative to the internal standards were obtained by the external standardization method. A total of 32 diamondoid compounds, including 22 adamantane and 10 diamantane compounds, were detected (Table 1, A1–A12 and D1–D12).

2.3. Statistical analyses

The multivariate statistical analyses used in this study included PCA, DA, and RA (Härdle and Simar, 2015). This was performed using IBM SPSS Statistics software. A total of 50 diamondoid indices, including 32 relative diamondoid compound abundances and 18 concentration ratios calculated according to previous studies (Fang et al., 2012; Jiang et al., 2018), were used as variables for the statistical analyses (Table 1). Diamondoid data from the simulation samples were considered to be the calibration dataset, and the data for the natural oil samples were regarded as the test dataset.

2.3.1. Principal component analysis

PCA was used to reduce the dimensionality of the diamondoid data. The raw diamondoid data were standardized using the Z-score method prior to PCA. Firstly, principal components (PCs) were extracted and a PCA model was established from the calibration data. The PC score values for the calibration data were calculated by linear combination of normalized original variables with loadings. The test data were then projected into the established PC model to calculate the score values of the new data using the same parameters (i.e., mean, standard deviation, and loadings; Table 2) that were used for the calibration data. Finally, the test data (i.e., the oil samples) could be directly compared with the calibration data (i.e., the simulation samples).

2.3.2. Fisher discriminant analysis

The PC score values obtained from the PCA model were further used to discriminate the source types of the simulation samples using a Fisher DA model. Simulation samples were divided into four groups according to source type, which were labeled as Group 1, 2, 3, and 4, corresponding to oil, Type I kerogen, Type II kerogen, and Type III kerogen,

Table 1
Definitions and abbreviations of diamondoid indices used in this study.

Concentration	Abbreviation	Diamondoid compound	Concentration Ratio	Abbreviation	Formula
A1	A	Adamantane	IR1	MAI	1-MA/(1-MA + 2-MA)
A2	1-MA	1-Methyladamantane	IR2	EAI	1-EA/(1-EA + 2-EA)
A3	1,3-DMA	1,3-Dimethyladamantane	IR3	DMAI-1	1,3-DMA/(1,3-DMA + 1,2-DMA
A4	1,3,5-TMA	1,3,5-Trimethyladamantane	IR4	DMAI-2	1,3-DMA/(1,3-DMA + 1,4-DMA)
A5	1,3,5,7-TeMA	1,3,5,7-Tetramethyladamantane	IR5	TMAI-1	1,3,5-TMA/(1,3,5-TMA + 1,3,4-TMA)
A6	2-MA	2-Methyladamantane	IR6	TMAI-2	1,3,5-TMA/(1,3,5-TMA + 1,3,6-TMA)
A7	1,4-DMA(*cis*)	1,4-Dimethyladamantane(cis)	IR7	MDI	4-MD/(4-MD + 1-MD + 3-MD)
A8	1,4-DMA(*trans*)	1,4-Dimethyladamantane(trans)	IR8	DMDI-1	4,9-DMD/(4,9-DMD + 3,4-DMD)
A9	1,3,6-TMA	1,3,6-Trimethyladamantane	IR9	DMDI-2	4,9-DMD/(4,9-DMD + 4,8-DMD)
A10	1,2-DMA	1,2-Dimethyladamantane	CR1	A/MAs	Adamantane/Methyladamantanes
A11	1,3,4-TMA(*cis*)	1,3,4-Trimethyladamantane(cis)	CR2	MAs/DMAs	Methyladamantanes/Dimethyladamantanes
A12	1,3,4-TMA(*trans*)	1,3,4-Trimethyladamantane(trans)	CR3	DMAs/TMAs	Dimethyladamantanes/Trimethyladamantanes
A13	1,2,5,7-TeMA	1,2,5,7-Tetramethyladamantane	CR4	A/D	Adamantane/Diamantane
A14	1-EA	1-Ethyladamantane	CR5	MAs/MDs	Methyladamantanes/Methyldiamantanes
A15	2,6 + 2,4-DMA	2,6- + 2,4-Dimethyladamantane	CR6	DMAs/MDs	Dimethyladamantanes/Methyldiamantanes
A16	1-E,3-MA	1-Ethyl-3-methyladamantane	CR7	DMAs/DMDs	Dimethyladamantanes/Dimethyldiamantanes
A17	1,2,3-TMA	1,2,3-Trimethyladamantane	CR8	MDs/DMDs	Methyldiamantanes/Dimethyldiamantanes
A18	1-E,3,5-DMA	1-Ethyl-3,5-dimethyladamantane	CR9	As/Ds	Total Adamantanes/Total Diamantanes
A19	2-EA	2-Ethyladamantane			
A20	1,3,5,6-TeMA	1,3,5,6-Tetramethyladamantane			
A21	1,2,3,5-TeMA	1,2,3,5-Tetramethyladamantane			
A22	1-E,3,5,7-TMA	1-Ethyl-3,5,7-trimethyladamantane			
D1	D	Diamantane			
D2	4-MD	4-Methyldiamantane			
D3	4,9-DMD	4,9-Dimethyldiamantane			
D4	1-MD	1-Methyldiamantane			
D5	1,4 + 2,4-DMD	1,4- + 2,4-Dimethyldiamantane			
D6	4,8-DMD	4,8-Dimethyldiamantane			
D7	1,4,9-TMD	1,4,9-Trimethyldiamantane			
D8	3-MD	3-Methyldiamantane			
D9	3,4-DMD	3,4-Dimethyldiamantane			
D10	3,4,9-TMD	3,4,9-Trimethyldiamantane			

respectively. Fisher DA is a linear DA technique that searches for a projection vector to maximize the difference between groups. The projection vector (i.e., the canonical discriminant functions) was calculated based on the calibration data and used to create a new set of data (i.e., scores of functions) from the original data (i.e., in this case PC score values from the PCA model). A territorial map showing the range of different groups was then constructed according to the distance between groups. Finally, the test data were projected onto the territorial map to calculate scores according to the canonical discriminant functions and were then classified.

2.3.3. Regression analysis

Linear and polynomial regression analyses were conducted on the calibration data, in which the extracted PCs and their corresponding EasyRo% values were regarded as independent and dependent variables, respectively. This allowed construction of the maturity model, which was then applied to the test data.

3. Results and discussion

3.1. Principal component analysis

To examine the potential relationships between diamondoid variables and the main factors influencing these variables, PCA was conducted on the diamondoid data from thermal simulation samples of known source type and thermal maturity. Three PCs were extracted based on the eigenvalue plot (Supplementary Fig. S1). The first three PCs explain 81.0% of the total variance of the dataset (i.e., 57.2% for PC1, 13.6% for PC2, and 10.2% for PC3). The loading plot in Fig. 1 shows that most of the variables have small angles (<45°) to PC1, meaning they are highly positively or negatively correlated with PC1, but weakly correlated with PC2 (Aitchison and Greenacre, 2002). PC1 explains 57.2% of the total data variance and separates the adamantanes of A3–A5 and diamantanes of D1–D8 with a positive loading from the adamantanes of A6–A22 with a negative loading. PC2 explains 13.6% of the total data variance and separates adamantanes of A1 and A2 with a positive loading from the diamantanes of D9 and D10 with a negative loading. The projection of the simulation samples onto PCs (Fig. 2) shows that the first PC is significantly associated with the maturity of the simulation samples (i.e., the maturity increases with PC1), whereas the second PC contains information regarding source type (i.e., the source type correlation becomes worse with increasing PC2). PC2 can effectively separate different groups of the simulation samples. Positive loadings for PC2 are representative of Type III kerogen-derived diamondoids, whereas negative loadings are representative of oil-derived diamondoids. The Type I–II kerogen-derived diamondoids are distributed along the PC2 axis. Thus, PC1 and PC2 represent mainly maturity and source type, respectively. The meaning of PC3 is ambiguous and is not discussed further here. Thus, the PC model can explain the factors that affect the diamondoid indices for the simulation samples.

The test data were also projected onto the PCA model constructed from the calibration data to test the ability of the model to group unknown samples (Fig. 2). Based on the PCA model, we can make some inferences about these oil samples. For example, the oil samples from the central Junggar Basin were sourced from Type I–II kerogens, which is consistent with previous studies that have shown these oils were mainly derived from high-quality Permian source rocks (Pan et al., 1999; Jiang et al., 2019). Oils from the Tazhong Uplift in the Tarim Basin were generated from Type I kerogen, which is also consistent with a previous study that showed these oils were mainly derived from marine source rocks (Li et al., 2018). Oil samples from the Baiyun Sag in the Pearl River Mouth Basin can be inferred to have been derived from Type II kerogen, which is consistent with their derivation from shallow lacustrine source rocks in the Baiyun Sag (Long et al., 2020; Jiang et al., 2021). The source of oil in the Kuqa Depression in the Tarim Basin is still unclear, due to their low biomarker abundances (Yang et al., 2016; Ji et al., 2017). Based on the PCA model, these oils are mainly derived from Type II–III kerogens. Hence, the oil samples can be broadly classified using the PCA

Table 2
Parameters used in the PCA model.

Variable	Basic Statistics		Loadings		
	Mean	Standard Deviation	PC1 (57.2%)	PC2 (13.6%)	PC3 (10.2%)
IR1	0.6449	0.1687	0.1717	0.0981	0.1181
IR2	0.6179	0.2213	0.1208	−0.0225	0.2549
IR3	0.6622	0.2003	0.1677	0.1244	0.1143
IR4	0.4761	0.2229	0.1774	0.0939	0.0275
IR5	0.3297	0.1858	0.1748	0.1082	0.0293
IR6	0.4698	0.2004	0.1642	0.1613	0.0303
IR7	0.2831	0.1072	0.1196	−0.1578	−0.0928
IR8	0.2415	0.0988	0.1673	0.0677	−0.0662
IR9	0.6122	0.1585	−0.0101	−0.0858	−0.1230
CR1	0.5390	0.9348	0.0681	0.1681	−0.2603
CR2	0.7149	0.2674	−0.1238	0.2039	0.0520
CR3	1.5571	0.4646	0.1035	0.0765	−0.3134
CR4	6.5277	5.7947	−0.0880	0.2626	−0.0832
CR5	2.1690	1.5889	−0.1455	0.1566	0.0056
CR6	2.7253	1.5557	−0.1508	0.1001	−0.0237
CR7	7.1643	5.1259	−0.1288	0.1827	0.0393
CR8	2.5440	0.7113	−0.0373	0.2814	0.1344
CR9	5.8668	3.9033	−0.1468	0.1569	0.0090
A1	7.8199	6.1569	0.0078	0.3046	−0.2249
A2	10.7984	3.8541	−0.0558	0.2519	0.1344
A3	9.5440	6.0312	0.1719	0.0864	0.0347
A4	3.3813	2.3228	0.1511	0.0660	0.1716
A5	0.3724	0.3611	0.1454	0.1234	−0.0537
A6	6.8372	3.8444	−0.1779	0.0278	−0.0690
A7	4.5391	1.7338	−0.1708	−0.0712	0.0367
A8	4.8042	1.8179	−0.1741	−0.0649	0.0522
A9	3.3707	1.2833	−0.1188	−0.0737	0.2408
A10	3.7772	1.9481	−0.1698	−0.1192	−0.0772
A11	3.2212	0.9547	−0.1484	−0.0094	0.2361
A12	3.2812	0.9816	−0.1479	−0.0335	0.2308
A13	1.8055	0.5238	0.0670	−0.0633	0.3248
A14	1.6430	0.8736	−0.1630	−0.1229	−0.0173
A15	2.1824	1.8883	−0.1317	−0.1818	−0.1895
A16	2.4326	1.0197	−0.1346	−0.0987	0.1747
A17	3.6333	2.1486	−0.1704	−0.1034	−0.1043
A18	0.6724	0.2959	−0.0314	−0.0666	0.3196
A19	1.5540	1.3925	−0.1440	−0.1205	−0.1509
A20	0.3587	0.1264	−0.1206	−0.0315	0.2290
A21	1.3518	0.7972	−0.1650	−0.1277	−0.0873
A22	2.1154	1.2020	−0.1790	−0.0410	−0.0332
D1	1.7601	0.9960	0.1706	−0.1211	0.0020
D2	4.1229	3.5344	0.1700	−0.1341	−0.0182
D3	1.1140	1.1191	0.1687	−0.1157	−0.0582
D4	1.7683	1.2351	0.1618	−0.0120	0.0857
D5	0.8447	0.8919	0.1687	−0.0901	0.0189
D6	0.7201	0.6915	0.1648	−0.1301	−0.0013
D7	0.3603	0.3327	0.1516	−0.1266	0.0352
D8	6.8220	3.3567	0.1611	−0.1254	0.0802
D9	2.9628	1.9554	0.1066	−0.2664	−0.0255
D10	0.0286	0.0693	0.0155	−0.2756	−0.0927

model constructed from the calibration data.

3.2. Fisher discriminant analysis and identification of oil source types

Assessing the origins and maturity of condensates, particularly for highly mature condensates, is challenging due to a lack of conventional biomarkers. Given that the PCA model could only broadly classify the oil samples, DA was applied to these samples. A Fisher DA model was constructed based on the calibration data with known groups and three canonical discriminant functions (Table 3). The first canonical discriminant function can explain 98.4% of the total variance, meaning that it can almost entirely discriminate the groups of simulation samples. In addition, the coefficient of PC2 in the first canonical discriminant function is much larger than that of PC1 or PC3, which also supports the view that PC2 mainly reflects the source type. The scores of the canonical discriminant functions for all data were calculated by combining the functions and original PC scores. Finally, a territorial map was constructed according to the distance between known groups in the calibration data.

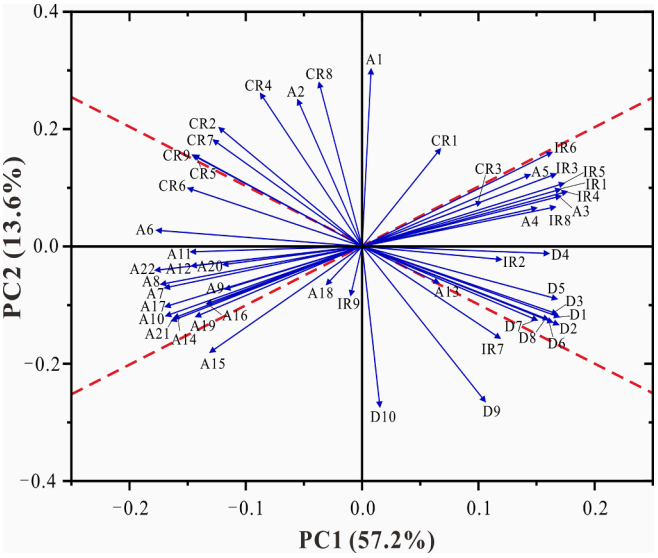


Fig. 1. Loading plot for the PCA model calculated from the simulation samples. The abbreviations are defined in Table 1.

The test data were projected onto the territorial map (Fig. 3), which contains four regions representing the range of four groups. The calibration data were classified into four groups and the classification results show that 95.3% of the original grouped cases and 86.0% of the cross-validated grouped cases were correctly classified, indicating the high accuracy of the Fisher DA model. Thus, the grouping of the test data can be predicted based on its projection onto this plot. Oils from the central Junggar Basin, Tazhong Uplift, Kuqa Depression, and Baiyun Sag are mainly classified into Groups 2–3, 2, 3–4, and 3 (Fig. 3 and Supplementary Table S1), which correspond to Type I–II kerogens, Type I kerogen, Type II–III kerogens, and Type II kerogen, respectively. Thus, the PCA–Fisher DA model can objectively and accurately group the oil samples and identify their source.

3.3. Establishment and application of the maturity model

PC1 of the PCA model represents mainly the thermal maturity and is independent of source (PC2). Thus, a maturity regression model was constructed based on the known maturity of the simulation samples and their PC1 scores. The calibration curve was fitted with a cubic regression analysis as it yielded a better fit than linear and quadratic regression analysis. In the maturity model (Fig. 4a), the r^2 value of the fitted curve is 0.89, indicating a good correlation between the known EasyRo% value and PC1 score. Taking into account the 95% confidence interval and maturity range of the original data, the optimum application range of the model is EasyRo = 1.0–2.7%, which is in the range of the maturity of light oil and condensate oil. The thermal maturation stages of kerogen are shown in Supplementary Fig. S2 based on our previous study (Jiang et al., 2018). A good linear relationship was also observed between the model-predicted and known EasyRo values (Fig. 4b), which has a slope of approximately one and an intercept of approximately zero. Thus, the maturity model constructed from the simulation samples is robust.

The maturity model was used to predict the maturity of the oil samples. The predicted results are shown in Fig. 4c and Supplementary Table S1, and reveal that oils from the central Junggar Basin have the lowest maturity with EasyRo = 0.8–1.2%, corresponding to the mature stage of the oil window; oils from the Tazhong Uplift in the Tarim Basin have EasyRo = 1.8–2.1%, corresponding to the late-mature stage of the oil window; oils from the Baiyun Sag in the Pearl River Mouth Basin (EasyRo = 1.4–1.9%) have maturities between the above two groups;

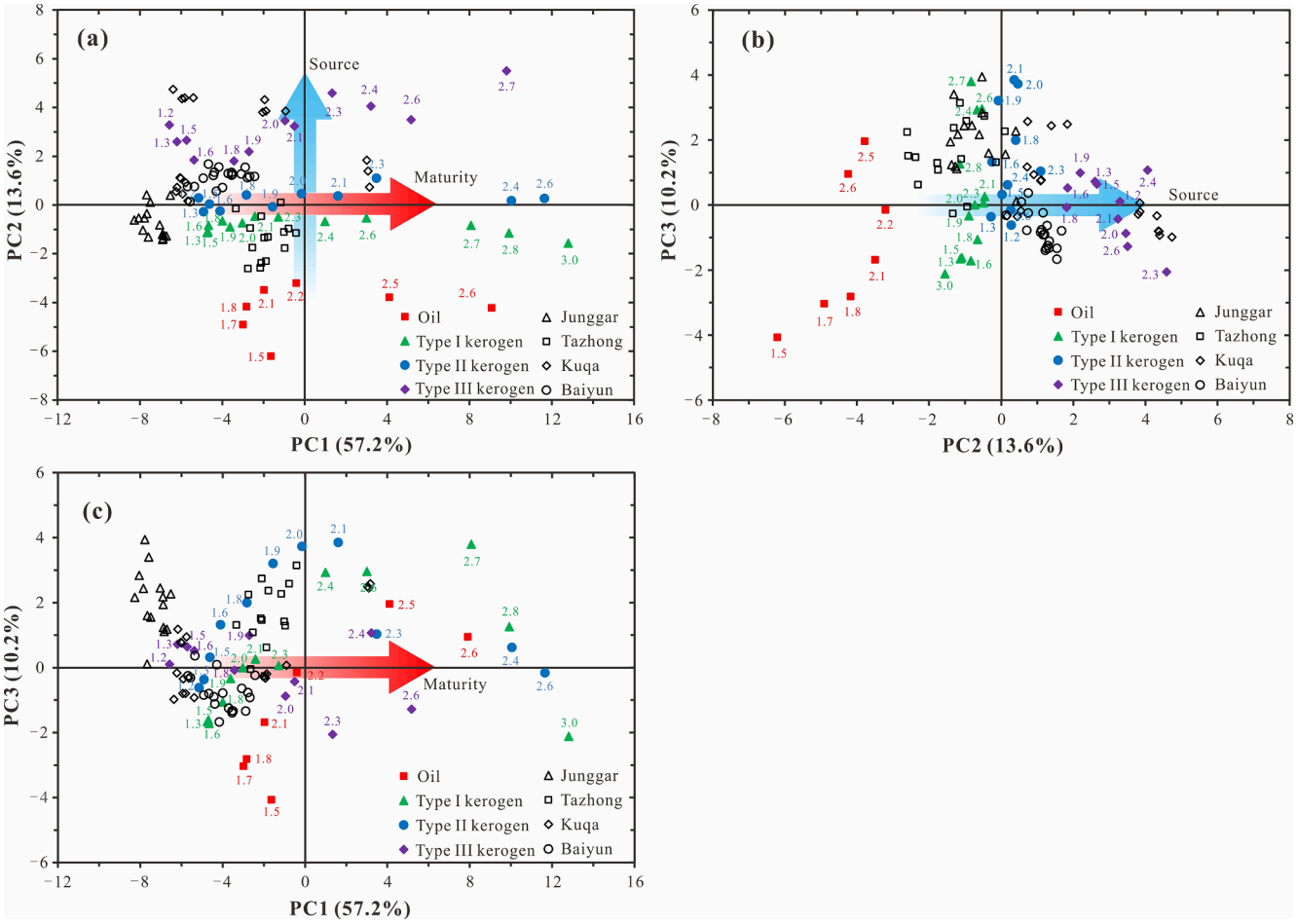


Fig. 2. Plots of (a) principal component (PC) 1 vs PC2, (b) PC2 vs PC3, and (c) PC1 vs PC3 for the principal component analysis model calculated from the simulation samples.

Table 3
Canonical discriminant function coefficients (unstandardized) of the Fisher discriminant analysis.

Variable	Function		
	1	2	3
PC1 (57.2%)	−0.0664	0.0318	0.1771
PC2 (13.6%)	1.4125	−0.0146	0.0379
PC3 (10.2%)	0.0848	0.4663	−0.0662
Constant	0.0000	0.0000	0.0000

oils from the Kuqa Depression in the Tarim Basin define three groups with EasyRo = 1.2–1.4%, 1.9–2.1%, and 2.4%. The predicted maturity of these oil samples from the maturity model is consistent with the conclusions of previous studies (Zhang et al., 2010; Li et al., 2018; Wang et al., 2018; Jiang et al., 2019). Different maturities predicted for the oils from the Kuqa Depression suggests that there might be multiple hydrocarbon injections in the petroleum reservoir. This is consistent with the study of Zhang et al. (2010), which indicated that there were two expulsion events in the Kuqa Sag and the gas and condensate charged later than the normal oils.

3.4. General methodology for oil source and maturity prediction

For any oil sample without obvious secondary alteration, its source and maturity can be determined based on the following steps: (1) diamondoids in the sample are analyzed and related indices (Table 1) are calculated; (2) these indices (i.e., variables) should be normalized and

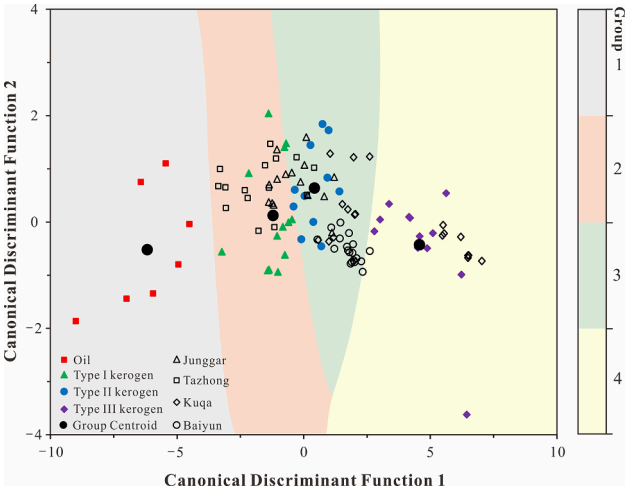


Fig. 3. Plot of canonical discriminant function 1 vs 2 for the Fisher discriminant analysis model calculated from the simulation samples.

are used to calculate PC scores based on the PCA model (Table 2); (3) after conversion by the canonical discriminant functions (Table 3), the sample can be projected onto the territorial map of the Fisher DA model (Fig. 3) and then the source can be determined; and (4) the maturity of the oil sample can be predicted from the maturity model (Fig. 4a).

The PCA–Fisher DA and PCA–RA models constructed from the

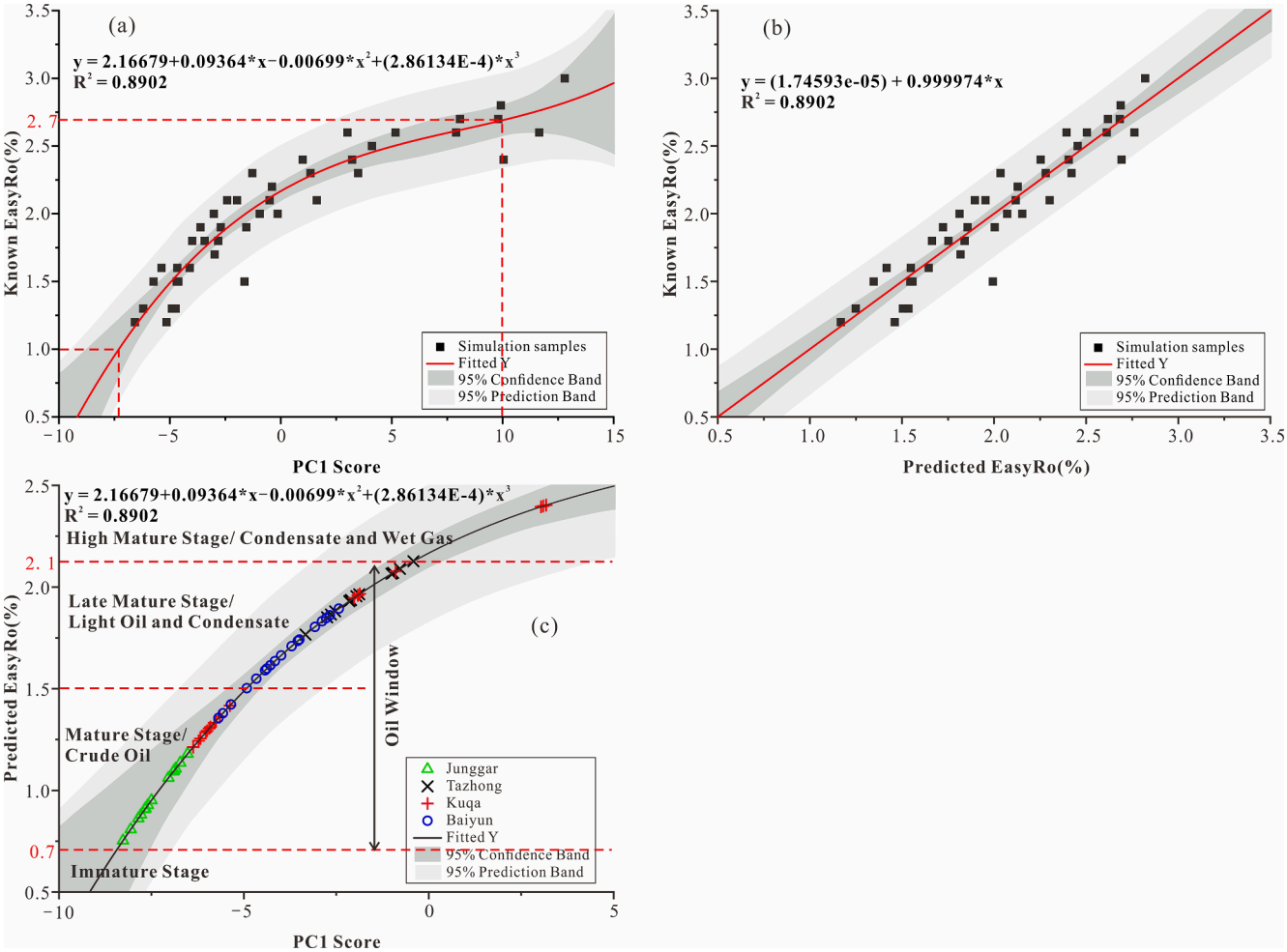


Fig. 4. (a) Maturity regression model calculated from the simulation samples; (b) Plot of the model predicted EasyRo vs the known EasyRo for simulation samples; (c) Model predicted EasyRo results for oil samples from different basins in China.

diamondoid data for the simulation samples can be used to identify the oil source and maturity, particularly for highly mature oils (i.e., light oils or condensates), which have low abundances of conventional biomarkers, but are rich in diamondoids. Given that the effects of secondary alteration (i.e., biodegradation, evaporation) or other factors (i.e., mixing, contamination from oil-based mud) on the diamondoid data were not considered during construction of the models, it is unclear whether these models are suitable for oil samples that have been subjected to any of these factors. Theoretically, multivariate statistical analysis on diamondoid indices would eliminate or diminish the impacts of some abnormal indices caused by secondary alteration or other factors. Certainly, the effects of secondary alteration or other factors on these models requires further study. In addition, more studies are needed to test the validity of this approach and extend it to other petroleum basins.

4. Conclusions

Multivariate statistical analyses were used to investigate diamondoid indices in oil samples. PCA revealed that PC1 reflects oil maturity and PC2 reflects oil source, based on simulation samples. Based on the PCA results, Fisher DA was used to further classify oil samples into groups, and RA was used to construct a maturity model. The models presented in this study can be used to identify the oil source type and quantitatively determine its thermal maturity, which significantly advances the use of diamondoids as source and maturity indicators.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 41902129 and 41773034), the Natural Science Foundation of Guangdong Province (Grant No. 2018B030306006), and the Youth Innovation Promotion Association CAS (Grant No. 2018386). This is contribution No. IS-3057 from GIGCAS. We really appreciate Drs John K. Volkman (Co-Editor-in-Chief), Clifford Walters (Associate Editor), Rob Forkner (Reviewer) and an anonymous reviewer for their helpful comments and suggestions which substantially improved our manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.orggeochem.2021.104298.

References

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 51 (4), 375–392.

Atwah, I., Mohammadi, S., Moldowan, J.M., Dahl, J., 2021a. Episodic hydrocarbon charge in tight Mississippian reservoirs of Central Oklahoma, USA: Insights from oil inclusion geochemistry. Marine and Petroleum Geology 123, 104742. https://doi.org/10.1016/j.marpetgeo.2020.104742.

Atwah, I., Moldowan, J.M., Koskella, D., Dahl, J., 2021b. Application of higher diamondoids in hydrocarbon mudrock systems. Fuel 284, 118994. https://doi.org/10.1016/j.fuel.2020.118994.

Banas, K., Banas, A., Moser, H.O., Bahou, M., Li, W., Yang, P., Cholewa, M., Lim, S.K., 2010. Multivariate analysis techniques in the forensics investigation of the postblast residues by means of Fourier Transform-Infrared Spectroscopy. Analytical Chemistry 82 (7), 3038–3044.

Botterell, P.J., Houseknecht, D.W., Lillis, P.G., Barbanti, S.M., Dahl, J.E., Moldowan, J. M., 2021. Geochemical advances in Arctic Alaska oil typing – North Slope oil correlation and charge history. Marine and Petroleum Geology 127, 104878. https://doi.org/10.1016/j.marpetgeo.2020.104878.

Chen, J., Fu, J., Sheng, G., Liu, D., Zhang, J., 1996. Diamondoid hydrocarbon ratios: novel maturity indices for highly mature crude oils. Organic Geochemistry 25 (3-4), 179–190.

Cheng, X., Hou, D., Xu, C., 2018. The effect of biodegradation on adamantanes in reservoired crude oils from the Bohai Bay Basin, China. Organic Geochemistry 123, 38–43.

Dahl, J.E., Liu, S.G., Carlson, R.M.K., 2003. Isolation and structure of higher diamondoids, nanometer-sized diamond molecules. Science 299, 96–99.

Dahl, J.E., Moldowan, J.M., Peters, K., Claypool, G., Rooney, M., Michael, G., Mello, M., Kohnen, M., 1999. Diamondoid hydrocarbons as indicators of oil cracking. Nature 399, 54–56.

Fang, C., Xiong, Y., Li, Y., Chen, Y., Liu, J., Zhang, H., Adedosu, T.A., Peng, P., 2013. The origin and evolution of adamantanes and diamantanes in petroleum. Geochimica et Cosmochimica Acta 120, 109–120.

Fang, C., Xiong, Y., Liang, Q., Li, Y., 2012. Variation in abundance and distribution of diamondoids during oil cracking. Organic Geochemistry 47, 1–8.

Forkner, R., Fildani, A., Ochoa, J., Moldowan, J.M., 2021. Linking source rock to expelled hydrocarbons using diamondoids: An integrated approach from the Northern Gulf of Mexico. Journal of Petroleum Science and Engineering 196, 108015. https://doi.org/10.1016/j.petrol.2020.108015.

Gentzis, T., Carvajal-Ortiz, H., 2018. Comparative study of conventional maturity proxies with the methyldiamondoid ratio: Examples from West Texas, the Middle East, and northern South America. International Journal of Coal Geology 197, 115–125.

Grice, K., Alexander, R., Kagi, R.I., 2000. Diamondoid hydrocarbon ratios as indicators of biodegradation in Australian crude oils. Organic Geochemistry 31 (1), 67–73.

Härdle, W., Simar, L., 2015. Applied Multivariate Statistical Analysis, 4th ed. Springer-Verlag, Berlin.

Hur, M., Yeo, I., Park, E., Kim, Y.H., Yoo, J., Kim, E., No, M.-H., Koh, J., Kim, S., 2010. Combination of statistical methods and Fourier transform ion cyclotron resonance mass spectrometry for more comprehensive, molecular-level interpretations of petroleum samples. Analytical Chemistry 82 (1), 211–218.

Ji, H., Huang, G., Cheng, D., Xu, S., 2017. Geochemical application of light hydrocarbons in Kuqa Depression of Tarim Basin: case study of Dawanqi-Dabei Areas. Natural Gas Geoscience 28, 965–974 (in Chinese with English abstract).

Jiang, W., Li, Y., Xiong, Y., 2018. The effect of organic matter type on formation and evolution of diamondoids. Marine and Petroleum Geology 89, 714–720.

Jiang, W., Li, Y., Xiong, Y., 2019. Source and thermal maturity of crude oils in the Junggar Basin in Northwest China determined from the concentration and distribution of diamondoids. Organic Geochemistry 128, 148–160.

Jiang, W., Li, Y., Xiong, Y., 2020. Reservoir alteration of crude oils in the Junggar Basin, northwest China: insights from diamondoid indices. Marine and Petroleum Geology 119, 104451. https://doi.org/10.1016/j.marpetgeo.2020.104451.

Jiang, W., Li, Y., Yang, C., Xiong, Y., 2021. Organic geochemistry of source rocks in the Baiyun Sag of the Pearl River Mouth Basin, South China Sea. Marine and Petroleum Geology 124, 104836. https://doi.org/10.1016/j.marpetgeo.2020.104836.

Komura, D., Nakamura, H., Tsutsumi, S., Aburatani, H., Ihara, S., 2005. Multidimensional support vector machines for visualization of gene expression data. Bioinformatics 21 (4), 439–444.

Kujawinski, E.B., Longnecker, K., Blough, N.V., Vecchio, R.D., Finlay, L., Kitner, J.B., Giovannoni, S.J., 2009. Identification of possible source markers in marine dissolved organic matter using ultrahigh resolution Mass Spectrometry. Geochimica et Cosmochimica Acta 73 (15), 4384–4399.

Landa, S., Machacek, V., 1933. Adamantane, a new hydrocarbon extracted from petroleum. Collection of Czechoslovak Chemical Communications 5, 1–5.

Li, Y., Xiong, Y., Liang, Q., Fang, C., Chen, Y., Wang, X., Liao, Z., Peng, P., 2018. The application of diamondoid indices in the Tarim oils. American Association of Petroleum Geologists Bulletin 102 (02), 267–291.

Liang, Q., Xiong, Y., Fang, C., Li, Y., 2012. Quantitative analysis of diamondoids in crude oils using gas chromatography-triple quadrupole mass spectrometry. Organic Geochemistry 43, 83–91.

Long, Z.L., Chen, C., Ma, N., Zhai, P.Q., Huang, Y.P., Shi, C., 2020. Genesis and accumulation characteristics of hydrocarbons in Baiyun Sag, deep water area of Pearl River Mouth Basin. China Offshore Oil and Gas 32, 36–45 (in Chinese with English abstract).

Moldowan, J.M.(., Dahl, J., Zinniker, D., Barbanti, S.M., 2015. Underutilized advanced geochemical technologies for oil and gas exploration and production-1: the diamondoids. Journal of Petroleum Science and Engineering 126, 87–96.

Pan, C.C., Fu, J.M., Sheng, G.Y., Yang, J.Q., 1999. The determination of oil sources and its significance in the central Junggar Basin. Acta Petrolei Sinica 20, 27–32 (in Chinese with English abstract).

Sassen, R., Post, P., 2008. Enrichment of diamondoids and ^{13}C in condensate from Hudson Canyon, US Atlantic. Organic Geochemistry 39 (1), 147–151.

Scarlett, A.G., Spaak, G., Mohamed, S., Plet, C., Grice, K., 2019. Comparison of tri-, tetra- and pentacyclic caged hydrocarbons in Australian crude oils and condensates. Organic Geochemistry 127, 115–123.

Schulz, L.K., Wilhelms, A., Rein, E., Steen, A.S., 2001. Application of diamondoids to distinguish source rock facies. Organic Geochemistry 32 (3), 365–375.

Spaak, G., Edwards, D.S., Grosjean, E., Scarlett, A.G., Rollet, N., Grice, K., 2020. Identifying multiple sources of petroleum fluids in Browse Basin accumulations using diamondoids and semi-volatile aromatic compounds. Marine and Petroleum Geology 113, 104091. https://doi.org/10.1016/j.marpetgeo.2019.104091.

Sweeney, J.J., Burnham, A.K., 1990. Evaluation of a simple model of vitrinite reflectance based on chemical kinetics. American Association of Petroleum Geologists Bulletin 74, 1559–1570.

Wang, C., Zeng, J., Zhang, Z., Shi, N., Lao, M., Zhao, Q., Dai, J., Wang, F., Liu, X., 2018. Origin and distribution of natural gas and oil in the Baiyun Depression, Pearl River Mouth Basin, South China Sea. Journal of Petroleum Science and Engineering 170, 467–475.

Wei, Z., Moldowan, J.M., Jarvie, D.M., Hill, R., 2006. The fate of diamondoids in coals and sedimentary rocks. Geology 34 (12), 1013–1016. https://doi.org/10.1130/G22840A.1.

Wei, Z., Moldowan, J.M., Peters, K.E., Wang, Y.e., Xiang, W., 2007a. The abundance and distribution of diamondoids in biodegraded oils from the San Joaquin Valley: implications for biodegradation of diamondoids in petroleum reservoirs. Organic Geochemistry 38 (11), 1910–1926.

Wei, Z., Moldowan, J.M., Zhang, S., Hill, R., Jarvie, D.M., Wang, H., Song, F., Fago, F., 2007b. Diamondoid hydrocarbons as a molecular proxy for thermal maturity and oil cracking: geochemical models from hydrous pyrolysis. Organic Geochemistry 38 (2), 227–249.

Williams, J.A., Bjorøy, M., Dolcater, D.L., Winters, J.C., 1986. Biodegradation in south Texas Eocene oils—effects on aromatics and biomarkers. Organic Geochemistry 10 (1-3), 451–461.

Wingert, W.S., 1992. GC-MS analysis of diamondoid hydrocarbons in Smackover petroleums. Fuel 71 (1), 37–43.

Yang, L., Li, M., Zhang, C., 2016. Influence of biodegradation on light hydrocarbon parameters in crude oil of Kuqa Formation from Dawanqi Oilfield. Geological Journal of China Universities 22, 549–554 (in Chinese with English abstract).

Zhang, B., Huang, L., Wu, Y., Wang, H., Cui, J., 2010. Quantitative evaluation of crude oil composition changes caused by strong gas washing: a case study of natural gas pool in Kuqa Depression. Earth Science Frontiers 17, 270–279 (in Chinese with English abstract).

Zhang, S., Huang, H., Xiao, Z., Liang, D., 2005. Geochemistry of Palaeozoic marine petroleum from the Tarim Basin, NW China. Part 2: Maturity assessment. Organic Geochemistry 36 (8), 1215–1225.

Zhu, G., Wang, H., Weng, N., Huang, H., Liang, H., Ma, S., 2013. Use of comprehensive two-dimensional gas chromatography for the characterization of ultra-deep condensate from the Bohai Bay Basin, China. Organic Geochemistry 63, 8–17.