Contents lists available at ScienceDirect

Computers and Geosciences

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

Chroma: A MATLAB package and open-source platform for biomarker data processing and automatic index calculations

Julian Traphagan [*], Guangsheng Zhuang

Louisiana State University, USA

ARTICLE INFO

Keywords:
Biomarkers
Hydrocarbon index
MATLAB
Chromatography
Biogeochemistry

ABSTRACT

The molecular ratio indices of biological markers (biomarkers), such as the Carbon Preference Index (CPI) or P_{aq}, are frequently used as proxies for paleoclimatic and palaeoecological conditions. These indices are regularly extracted from the relative abundances of target molecules detected by a Gas Chromatography analyzer with a Flame Ionization Detector (GC-FID). Despite their use in biogeochemical studies for over a half-century, it remains common procedure to quantify the abundance of individual compounds by manual integration of chromatogram peaks (i.e., interpret baselines visually and characterize peaks by hand), which is time consuming and can lead to inconsistent results. Here, we introduce a new MATLAB package (Chroma) for the automatic detection and integration of standard-referenced biomarker abundances and the calculation of a variety of established hydrocarbon indices commonly reported in the published literature. The algorithm identifies the detector response timing of specific target peaks in a sample chromatogram by cross-referencing to a standard (e.g., Mix-A6, Schimmelmann, Indiana University Bloomington), then calculates the peak areas for an approximation of molecular abundance. This new toolkit for automatic and rapid integration of GC-acquired data provides a consistent and reproducible approach for the calculation of hydrocarbon indices and offers a standardized inter-laboratory platform for data comparisons and exchange. We validate the utility of the Chroma package with the chromatograms of plant wax *n*-alkanes, a widely used proxy for ecology and hydrology, from six stratigraphic sections in the Tibetan Plateau. Chroma is an effective tool for efficient data processing and will continuously evolve to accommodate extended uses in related areas of biomarker research beyond *n*-alkanes.

1. Introduction

Indices determined from the molecular distribution of biological markers (biomarkers), such as *n*-alkanes (Bray and Evans, 1961; Eglinton and Hamilton, 1967; El Nemr et al., 2016; Ficken et al., 2000; Lee et al., 2019; Poynter and Eglinton, 1990), are an expanding class of proxies in organic geochemistry frequently utilized for assessing paleoenvironmental and palaeoecological interpretations of organic compounds preserved in sediments (Aichner et al., 2018; El Nemr et al., 2016; Ortiz et al., 2021). These molecular ratios provide constraints on biological source (Bush and McInerney, 2013; Li et al., 2020; Zhang et al., 2017), extent of diagenesis (Kang et al., 2020; Ofiti et al., 2021; Thomas et al., 2021), and the ecological responses to changes in environmental conditions (Leider et al., 2013; Seki et al., 2009, 2012; Wang et al., 2016), and are powerful tools for contextualization of the sedimentary record. Many studies have used hydrocarbon indices, such as carbon preference index (CPI), to supplement the hydroclimatic and ecological interpretations of the stable isotopic compositions of *n*-alkanes (Duan and Jinxian, 2011; Niedermeyer et al., 2016; Tibbett et al., 2021; Yan et al., 2021), polycyclic aromatic hydrocarbons (Égüez et al., 2022; Kang et al., 2020), and saccharides (Stolpnikova et al., 2020) in sediments and soils, and to characterize the state of preservation of organic matter in the stratigraphic record (Elson et al., 2022; Feakins et al., 2016).

Standard strategies for quantifying the distribution of lipid compounds in a sample analyzed by gas-chromatography (GC) and detected by a flame ionization detector (GC-FID) generally involve manual integration of sample chromatogram peaks by visually comparing them to peaks in the chromatogram of an external standard. While interpretations made from the quantification of molecular abundance rely heavily on consistent and quality peak characterization, manual approaches in determining their properties can be highly variable between research groups and individuals, and typically demand a substantial amount of processing time. In some cases, peak integrations are only

[*] Corresponding author.
E-mail address: jtraph1@lsu.edu (J. Traphagan).

https://doi.org/10.1016/j.cageo.2024.105675
Received 9 January 2024; Received in revised form 20 June 2024; Accepted 26 June 2024
Available online 29 June 2024
0098-3004/Published by Elsevier Ltd.

roughly estimated by manually identifying the detector response of the target compounds, an approach that is inherently subjective, tedious, time-consuming, inconsistent, and is prone to large and difficult to quantify uncertainties. Additionally, current platforms for post-processing of GC-acquired chromatograms are commonly limited by low customizability or prohibitive costs. These software applications also lack the capability to automatically identify specific biomarker components (e.g., *n*-alkane homologs of different carbon chain lengths) by cross-referencing them with a known standard and smoothly transfer this information to the calculation of molecular abundance ratios from the estimated integrations. This has led research groups within the biomarker geochemistry community to develop independent and differing internal procedures which can vary widely from practice to practice.

To streamline the conversion of GC-FID data (e.g., *n*-alkane intensity) to homolog-specific (e.g., carbon chain lengths C_{25}, C_{27}, C_{29}, etc.) determinations of abundance and improve procedures for post-integration index calculations, we developed a new package of MATLAB functions (Chroma) for reproducible and efficient processing of GC-generated chromatograms. The main objective of Chroma is to offer a free and universally accessible platform for the inter-laboratory standardization of protocols for peak integration, data comparison, and quality control (Fig. 1). The output data generated by this subroutine can be applied to multiple samples simultaneously and visualized using built-in plotting options for data comparisons and interpretation (e.g., age-correlated profiles of hydrocarbon indices).

We demonstrate the utility of Chroma using chromatograms from previously reported Oligocene-Pleistocene age plant wax *n*-alkanes from fluvial and lacustrine sediments of six stratigraphic sections (Fig. 2A) in the Qaidam Basin and Hexi Corridor (Hou et al., 2021; Wu et al., 2019, 2021). The consistency of the tool and its capacity for handling large datasets is demonstrated by comparing the reported CPI values from these studies to those calculated in Chroma. We also report the results of additional hydrocarbon indices which were not included in the original analyses. Our results demonstrate Chroma's broad applicability for the automated processing of biomarker data to produce high quality and reproducible estimations of frequently reported proxy indices. This approach promotes the inter-laboratory standardization of data production in biomarker analysis and provides a new systematic method for the treatment of GC data from initial acquisition to full processing and implementation. The peak identification and integration computations in Chroma can be used for the GC data of any biomarker for which there is a corresponding standard; currently, calculations of molecular ratio indices are limited to those of *n*-alkanes, as they are the most prevalently reported biomarker in the literature, but the package will be continuously expanded to include index and post-integration processing functionalities for other biomarkers. Chroma is a community-based platform for which users are encouraged to expand on the base function library for future downloadable versions of the toolkit.

2. Methods and computational structure

Sediments sampled from six sections of the western Qaidam Basin and Hexi Corridor (Fig. 2A) spanning Oligocene to Pleistocene in age were used in this analysis: Honggouzi (HGZ), Qigequan (QGQ), Caogou (CG), Laojunmiao (LJM), Xichagou (XCG), and Wenshushan (WSS). These sections were previously studied for compound-specific isotope analyses of carbon and hydrogen in plant wax *n*-alkanes and stable oxygen and carbon in sedimentary carbonates for reconstructions of paleohydrology and paleoecology in the Tibetan Plateau (Hou et al., 2021; Kent-Corson et al., 2009; Wu et al., 2019, 2021). CPI values for all sections were manually characterized using the Thermo Scientific Xcalibur software for data acquisition and processing (Hou et al., 2021). We compare the originally reported CPI values of these formations to the values calculated in the automatic detection and integration algorithm of Chroma. Calculations of other hydrocarbon indices were not performed in the original analyses; their values determined by Chroma are also reported here for additional insight on regional paleoecology and

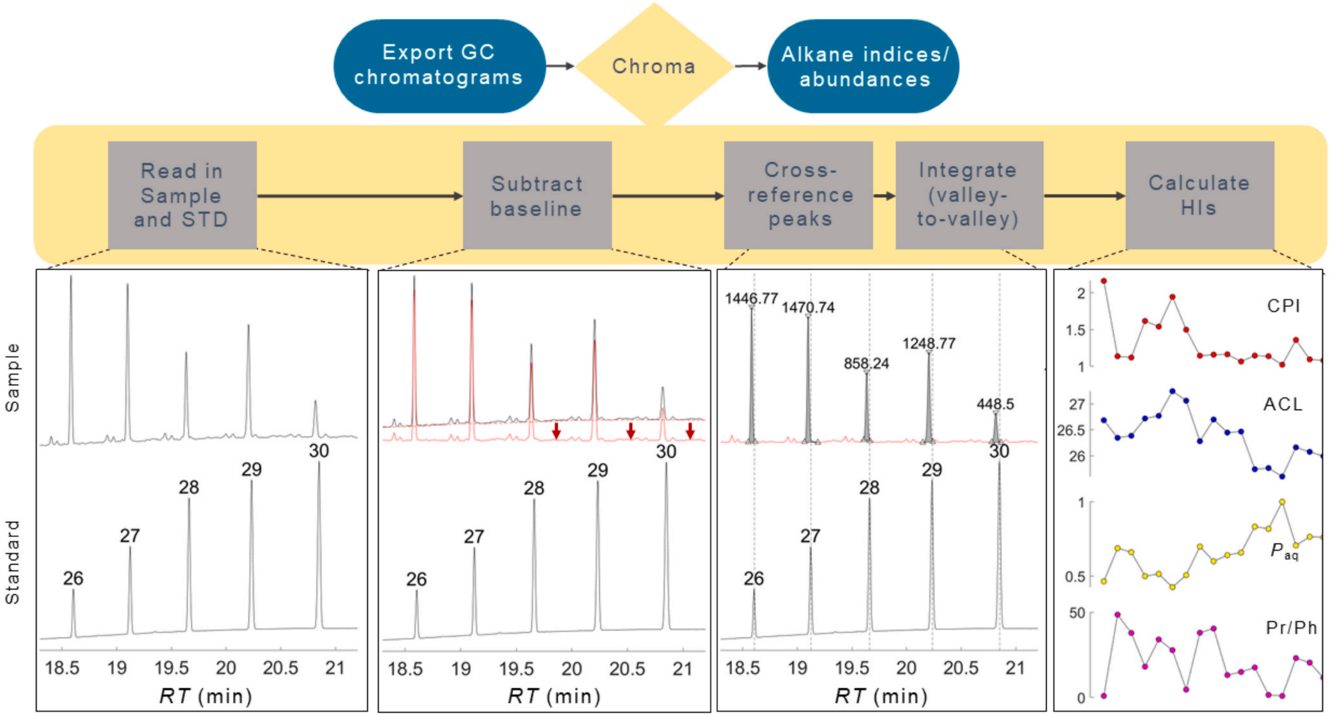

Fig. 1. Schematic flowchart of basic calculations performed in Chroma. Operations in the Chroma package are outlined by yellow regions. Example sample and standard chromatograms are shown below each step description. Peak areas are reported in units of intensity and time (fA x min). Homologs of the *n*-alkanes are labeled by number. STD = reference standard; *RT* = retention time (min); HI = hydrocarbon index; CPI = carbon preference index (C_{23-33}); ACL = average chain length (C_{16-33}).

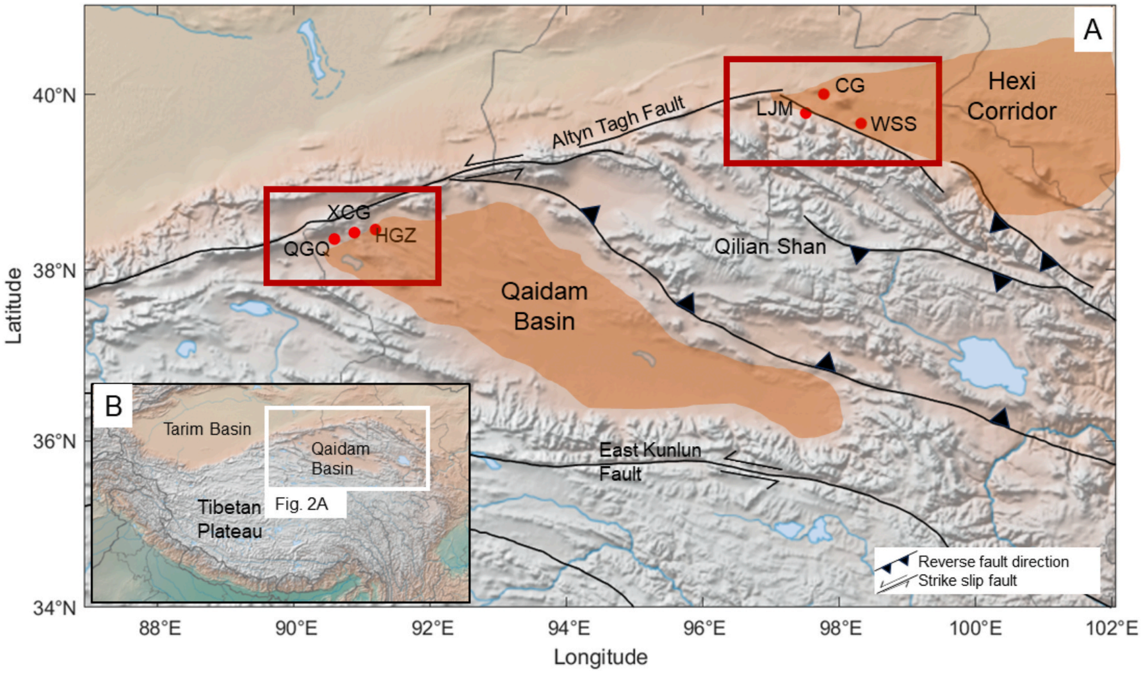

Fig. 2. (A) Map of the Northern Tibetan Plateau and stratigraphic sections in the Qaidam Basin and Hexi Corridor (orange shaded regions). The western Qaidam Basin contains the Qigequan (QGQ), Honggouzi (HGZ), and Xichagou (XCG) sections. The Hexi Corridor contains the Caogou (CG), Laojunmiao (LJM), and Wenshushan (WSS) sections. Solid black lines delineate major faults. (B) Inset map of the Tibetan Plateau.

hydroclimate. All *n*-alkane detector responses were generated using a Thermo Scientific Trace 1310 Gas-Chromatography analyzer with a Flame Ionization Detector (GC-FID), fitted with a programmable-temperature vaporization (PTV) injector and TG-1MS column (60 m length, 0.25 mm inner diameter, 0.25 μm film thickness), and digitized in Xcalibur (Fig. 3). A general sample processing flow from GC-FID analysis to Chroma is schematically depicted in Fig. 3. Specific GC conditions and operational settings are detailed in the reporting literature of the original chromatograms and outlined in the supplementary documentation (Hou et al., 2021; Wu et al., 2019, 2021).

2.1. Chroma algorithm

Peak identification of biomarkers of interest (e.g., *n*-alkanes) and integration calculations of biomarker peak areas are performed in the title function *chroma* included in the Chroma package. The fundamental flow structure of the Chroma algorithm follows these automated processing steps in MATLAB: (1) read in raw chromatogram data generated

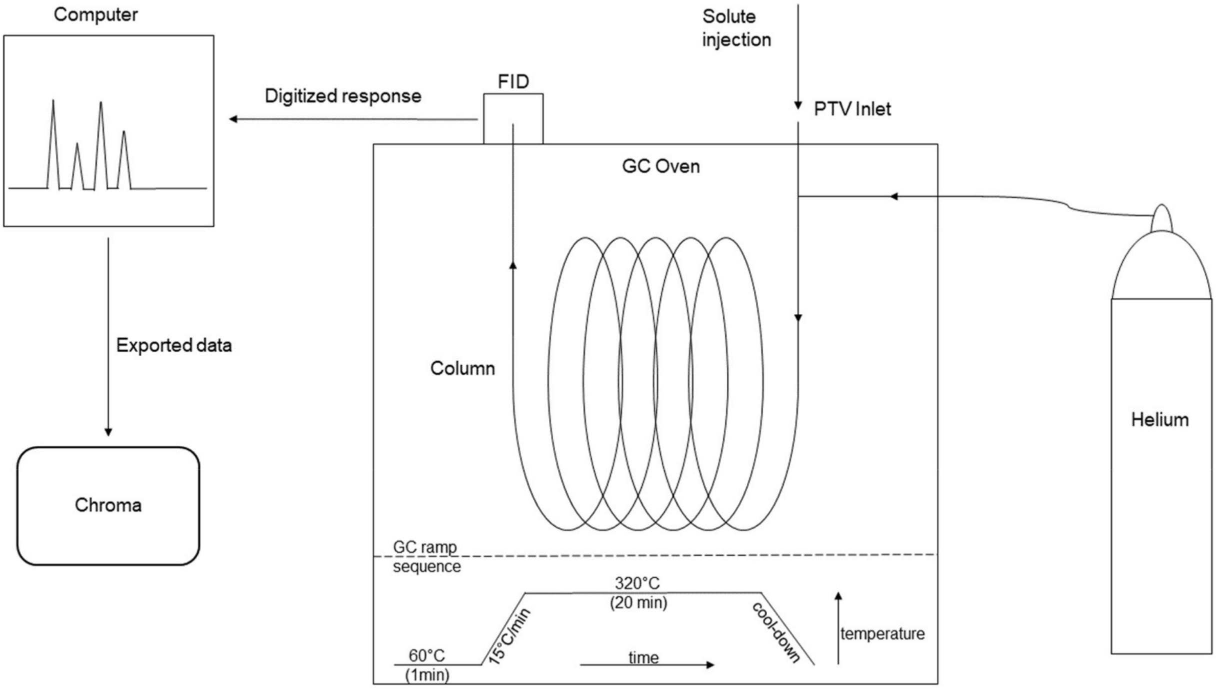

Fig. 3. Schematic diagram of a GC-FID system. A summary of the GC oven ramp sequence is illustrated at the bottom.

by GC acquisition, (2) subtract the baseline for improved peak identification, (3) cross-reference peaks in the sample to peaks in a known standard based on user-defined identification criteria, (4) integrate referenced sample peaks to approximate relative abundance, and (5) calculate hydrocarbon index values from the ratios of the integrated peak areas (Fig. 1). The function scans the GC traces from one sample and one reference standard (ideally acquired during the same analysis sequence), such as Mix-A6 (Schimmelmann, Indiana University Bloomington), then identifies the individual *n*-alkane homologs by referencing the sample chromatogram peaks to the known standard peaks (Fig. 1). Peaks in the sample chromatogram with retention times inside a user-specified time window (*ds*) of the available standard peaks are integrated to determine the molecular abundance of each carbon chain length (Fig. 4). The parameter *ds* (the number of data points away from the reference peak center in either direction) should be large enough to capture the target peaks in the sample but narrow enough to exclude the distal peaks which are not well-aligned with the standard – this value will vary depending on instrument sampling rates and the type of biomarker. The GC-FID system used to produce the chromatograms analyzed in this study acquired combustion response data at a rate of 0.1 s per data point (e.g., $ds = 40$ indicates a detection window of 8 s). Sample peaks are also filtered by specifying a minimum peak height, below which all peaks (e.g., noise, contaminants, etc.) are removed. The combined use of sample-to-standard peak referencing and a minimum peak threshold provides a strict and consistent set of criteria for target peak identification (Figs. 1 and 4). Hydrocarbon indices are then determined by calculating the ratios of particular chain lengths of *n*-alkanes integrated in the algorithm.

By default, the Chroma algorithm is initialized by subtracting the baseline of the sample chromatogram using a not-a-knot cubic spline interpolation of the local minima – if a blank trace from a solvent analysis is available, *chroma* is able to accept this chromatogram as the baseline instead. All peaks in the sample detected above a specified minimum intensity threshold are identified, and homolog compounds not contained in the standard are removed to retain only the targeted biomarker peaks (Figs. 1, 4 and 5). The target peaks are integrated using the native MATLAB trapezoidal integration function *trapz* (a basic valley-to-valley integration method is shown in Fig. 4, but *chroma* contains several built-in integration method options) and correlated with the standard to determine the homolog-specific abundances (Fig. 4). The *chroma* function analyzes each peak by direct sample-to-standard peak retention time correlation. Peaks are then individually assessed for changes in gradient to reduce false interpretations of noisy or coeluted peaks (Fig. 5). In cases where multiple peaks satisfy the detection criteria, the peak nearest in retention time to the external standard is selected. Compounds will not be automatically identified if the standard material is not available. For known peaks observed in the sample chromatogram but not in the standard (i.e., chain-lengths greater than 30 in Mix-A6), *chroma* can be instructed to include these homologs during the analysis by manually entering their expected retention times and compound number (Fig. 5). A simple valley-to-valley integration method is shown in Fig. 5 to highlight peak-base identifications in Chroma, however the integration method for identified peaks is user-definable and includes additional integration approaches such as drop-to-baseline and tangential-drop (descriptions of integration method options are provided in the supplementary documentation).

Default values for most input parameters, such as the detection window or minimum intensity threshold, are built-in but can be adjusted on a case-by-case basis to accommodate differences in instrumental sampling rates (i.e., frequency of data acquisition from the GC-FID), available standard compounds, variability in compound abundances, and condition settings for the GC and FID. Required inputs include the time and intensity information of the sample and standard chromatograms and the available homolog numbers (i.e., C_{29}, C_{30}, C_{31}, etc.) in the

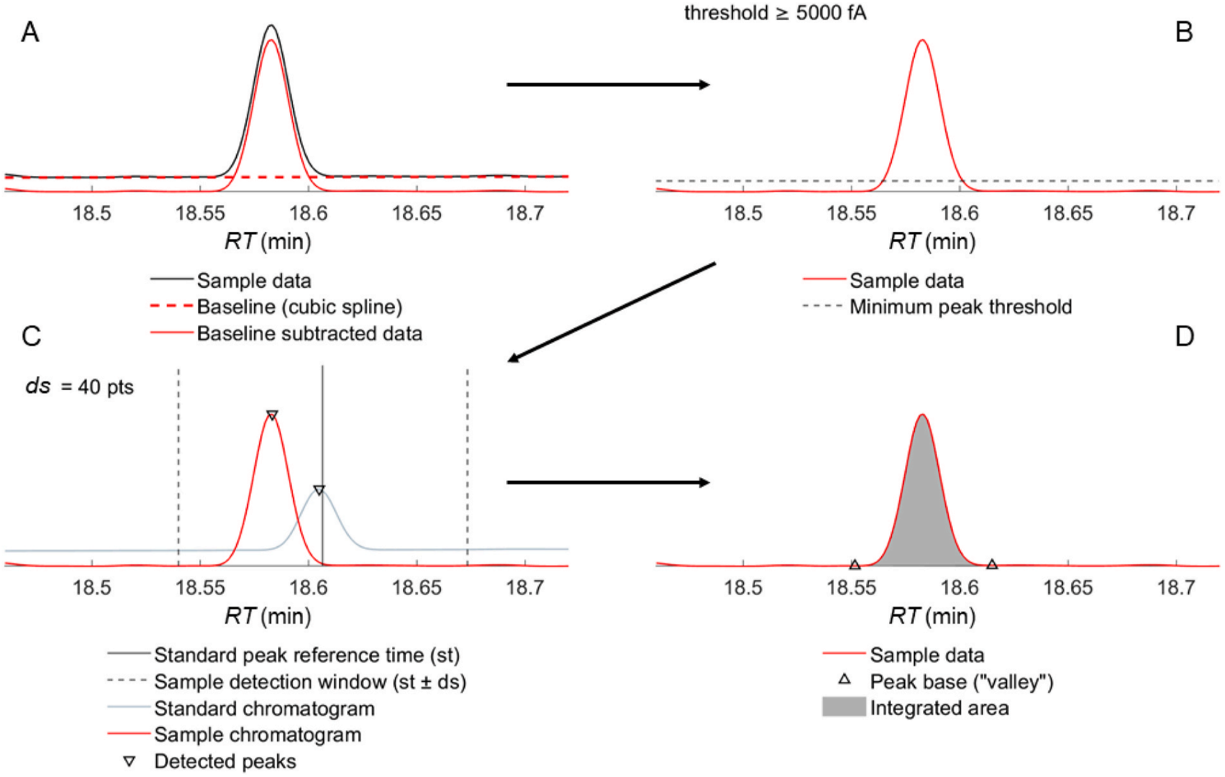

Fig. 4. Example of sample processing flow in chroma. (A) Baseline subtraction using a not-a-knot cubic spline interpolation of the local minima. (B) Minimum peak height threshold, below which all peaks are filtered out. (C) Detection window for standard to sample peak cross-referencing. The parameter *ds* is defined as the number of data points (i.e., distance from the standard peak center) within which a sample peak detection is permitted. Sample peaks outside of this window are filtered out of the analysis. (D) Valley-to-valley base identification and peak integration.

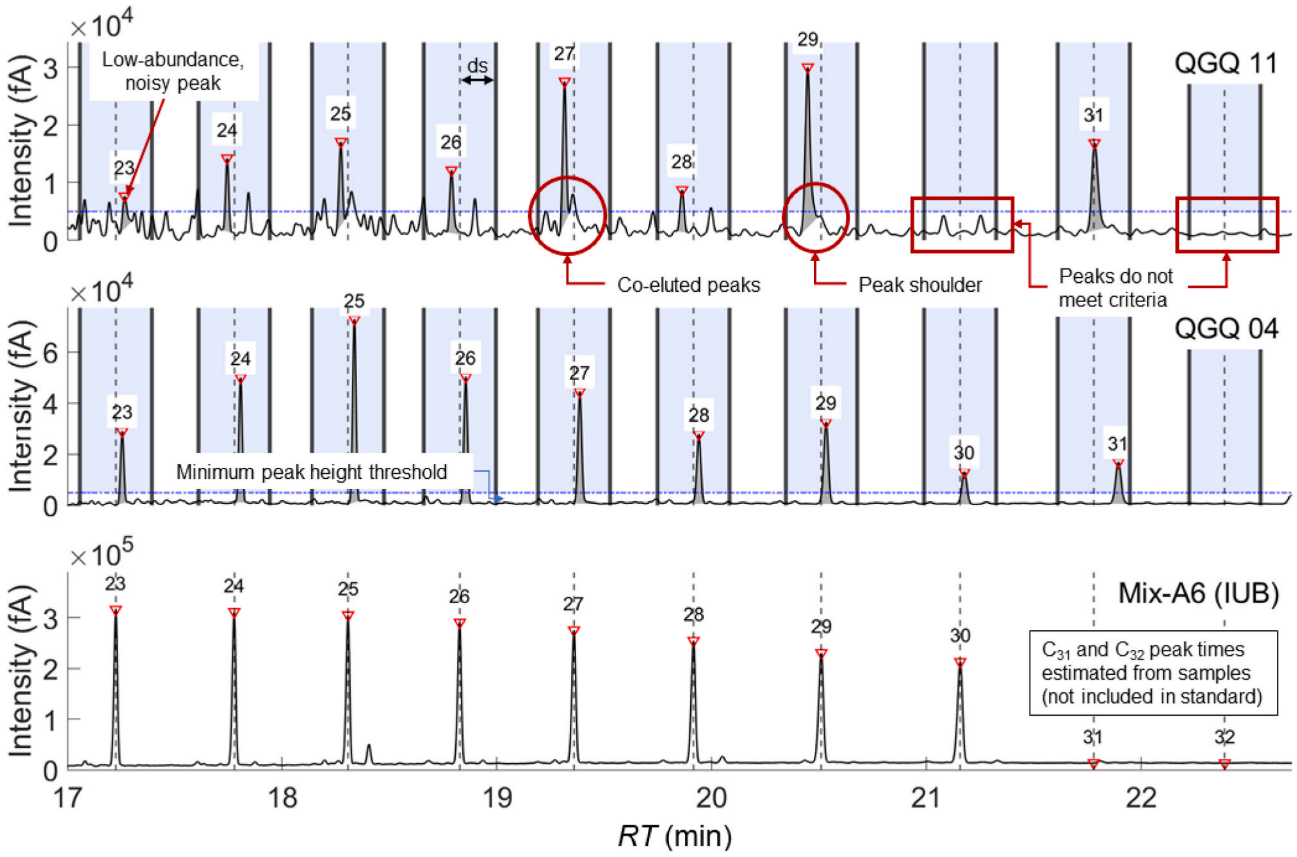

Fig. 5. Examples of automatic target peak detection criteria in chroma for a noisy sample (QGQ 11) and non-noisy sample (QGQ 04) from section QGQ. Sample peaks are filtered by minimum peak height threshold (blue dashed line) and cross-referencing to the standard peaks (vertical dashed lines). Solid vertical lines and blue shaded areas indicate the boundaries of the sample detection window (*ds* = 40). Homologs in the standard are correlated with the corresponding sample peak to assign the homolog number (e.g., C_{25}, C_{27}, etc.) labeled above each peak. Mix-A6 was used as the standard for analysis (homologs $C_{16\text{-}30}$). Additional homologs C_{31} and C_{32} were added based on the approximate timing of peaks observed in the samples. If not included, these peaks will not be detected.

standard. All other input parameters are optional and may be adjusted from the default values; the descriptions for these parameters can be found in the supplementary vignette. Performance times for running the algorithm will vary depending on the vector length of the input spectra. The Chroma package includes a function for performing this operation over multiple samples at once (*chromall*). The *chromall* function applies the *chroma* algorithm to each sample and automatically calculates the hydrocarbon indices. This functionality can also output a suite of figures for assessing the performance of the run, diagnosing potential adjustments in input parameters, and summary statistics of the analysis (see Fig. S3 in supplementary document). Full descriptions of each function in the package and their capabilities are provided in supplementary documentation.

2.2. Built-in index calculations

Chroma accommodates a suite of built-in hydrocarbon indices which can be calculated automatically during the peak integrations of a series of sample chromatogram traces. A list of all available output indices generated by Chroma can be found in the supplementary documentation. Here, we discuss the calculations of CPI, average chain length (ACL), P_{aq}, and odd-over-even predominance (OEP). Currently, Chroma can calculate 19 hydrocarbon indices as a part of its basic package but can be expanded to include additional automatic calculations.

CPI is a measure of the relative predominance of odd-numbered and even-numbered chains of carbon atoms of *n*-alkanes in a sample (Bray and Evans, 1961; Marzi et al., 1993):

$$\mathrm{CPI} = \left(\Sigma C_{23\text{-}31\,|\,\mathrm{odd}} + \Sigma C_{25\text{-}33\,|\,\mathrm{odd}}\right) \Big/ \left(2 \text{ x } \Sigma C_{24\text{-}32\,|\,\mathrm{even}}\right)$$

where the abundance (or concentration) of each *n*-alkane homolog is denoted as C_n hereafter, with *n* being the homolog number. The ratio is commonly used in petroleum geochemical analysis as an indicator for organic maturity and source, where higher values (CPI >1) of the index reflect a preference toward odd-numbered carbon chains and therefore indicate terrestrially sourced organic matter and thermal immaturity (Eglinton and Hamilton, 1967). Terrigenous vascular (higher) plants typically have CPI values over 3, while lower CPI values (CPI <1) may indicate a higher contribution of organic matter from petroleum and aquatic bacteria. Sediment and organic matter input from marine microorganisms and recycled material are represented in the CPI ≈1 range (Herrera-Herrera et al., 2020). We report CPI values using a homolog range of $C_{23\text{-}33}$.

OEP is an alternative ratio for the relative proportion of odd-numbered and even-numbered chains (Scalan and Smith, 1970); the index is typically given in the form $\mathrm{OEP} = (C_{27} + C_{29} + C_{31} + C_{33})/(C_{26} + C_{28} + C_{30} + C_{32})$. Like CPI, OEP can be used as an indicator for the presence of kerogens. OEP values of 4–8 typically indicate terrestrial plant wax *n*-alkanes, while lower values may indicate crude oil or aquatic lipids (Hoefs et al., 2002; Struck et al., 2020; Zech et al., 2009).

ACL is the weighted average of available chain lengths of *n*-alkane molecules (Jeng, 2006; Poynter and Eglinton, 1990):

$$\mathrm{ACL} = (\Sigma C_n \text{ x } n)/(\Sigma C_n)$$

Lipids of vascular plants typically produce a greater number of *n*-alkanes with chain lengths in the $C_{23\text{-}33}$ range and a strong predominance of odd-numbered homologs; *n*-alkanes produced by marine organisms like bacteria are typically shorter in chain length. Thus, ACL is commonly used as a proxy for vegetation type and associated climates. For example, grassland-derived plant lipids produced in more arid climates may have higher ACL values than lipids produced by plants in relatively humid forests (Cranwell, 1973). The production of longer chain *n*-alkanes is climatically associated with warmer or more arid climates (Leider et al., 2013; Simoneit et al., 1991).

P_{aq} is a proxy for the relative input of submerged floating aquatic macrophytes and emergent or terrestrial plants in lacustrine environments (Ficken et al., 2000):

$$P_{aq} = (C_{23} + C_{25})/(C_{23} + C_{25} + C_{29} + C_{31})$$

Typical values of P_{aq} in modern plants generally follow these ranges: $P_{aq} = 0$–0.1 for terrestrial plants, $P_{aq} = 0.1$–0.4 for emergent macrophytes or a mixture of terrestrial and aquatic plants, and $P_{aq} = 0.4$–1 for submerged and floating species (Ficken et al., 2000).

3. Results

Hydrocarbon index values and peak areas of *n*-alkanes were determined using the *chroma* algorithm. Calculations were performed on a total of 152 samples for six sections: 14 samples for QGQ, 23 samples for HGZ, 48 samples for XCG, 23 samples for CG, 23 samples for LJM, and 21 samples for WSS (see Fig. 2 for study locations).

3.1. Chroma performance

The hydrocarbon index calculations, including peak identification and peak area integration, were performed at a rate of 0.076 s per sample and a total of ~11.5 s for the complete analysis. Performance times will vary depending on the operating system, computer specifications and input parameters. Peaks outside those available in the reference standard (i.e., noise or contaminants) were successfully filtered out of the analysis by standard-referenced target peak identification, and *n*-alkanes were well-captured (i.e., not omitted or missed) in all automatic detections and integrations.

3.2. Index calculations

Hydrocarbon indices determined by Chroma are reported in Fig. 6 by section and basin – red for Hexi Corridor samples and blue for Qaidam samples. CPI values in the Qaidam Basin are relatively more variable from 30 Ma to 20 Ma (XCG section), stabilize to lower values averaging ~1.2 by 17 Ma, then increase between 17 and 13 Ma and after ~5 Ma. CPI values in the Hexi Corridor remain mostly stable at ~1.24 and generally overlap with the CPI values of Qaidam, with sporadic higher values at 9-7 Ma and 3-2 Ma (Fig. 6). ACL values were automatically calculated using all available homologs in the standard ($C_{16\text{-}33}$) – however, Chroma allows for a user-defined homolog range if preferred (i.e., average higher plant chain length; AHPCL). They range from 22.57 to 29.16 with a mean of 26.59 in Qaidam and 22.98 to 28.57 with a mean of 26.18 in the Hexi Corridor. P_{aq} values in the Qaidam Basin increase gradually until ~11 Ma, stabilize at ~0.8 before decreasing to ~0.5 by 2.5Ma (Fig. 6). Hexi Corridor P_{aq} values remain relatively stable at ~0.7–0.8, with increased variability after ~12 Ma. OEP values for both regions are largely covaried with CPI (Fig. 6). Other index values calculated during this analysis are available in the supplementary materials.

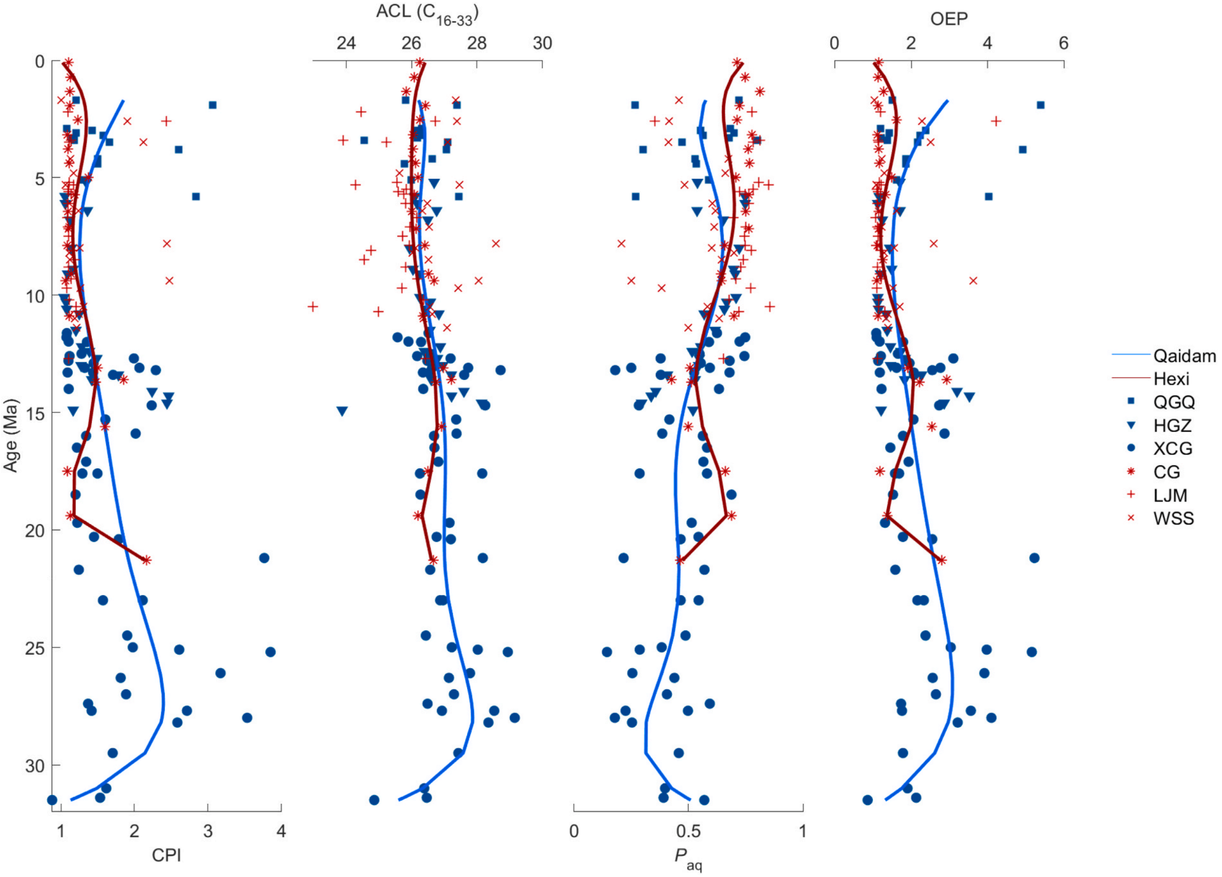

Fig. 6. Selected hydrocarbon indices of the Qaidam Basin (blue) and Hexi Corridor (red) determined by Chroma. The total analysis time including peak identifications, integrations, and index calculations was ~11.5 s for 152 samples across 6 sections using *chromall*. Trend lines were calculated by polynomial regression.

4. Discussion

4.1. Comparison of CPI determinations

We use the manually calculated CPI values from the original analysis for all sections by Hou et al. (2021) and Wu et al. (2019) to validate the peak area measurements and index calculations performed by Chroma (Figs. 7 and 8). While the reprocessed CPI profiles are generally in good agreement with their originally reported values, notable differences in the analyses include individual samples at 13.1 Ma (CG), 9.4 Ma (WSS), and 1.9 Ma (QGQ) (Fig. 8). The discrepant CG and WSS samples exhibit a dominantly terrestrial higher plant distribution, with chain lengths appearing only in the C_{23} to C_{33} range, while the QGQ sample exhibits relatively less odd-over-even predominance and slightly larger abundances in the lower alkane chain lengths (i.e., below C_{23}). Chroma's diagnostic features and visual inspection of the automatic integrations reveal well-captured baselines and peak areas and consistent retention time residuals between sample and standard homolog matches. We anticipate that these deviations may occur in part due to a greater tendency of the *chroma* algorithm to identify homologs with low abundance or target peak ambiguity, in contrast to a manual approach in which the individual may exercise a higher degree of restraint when visually evaluating peaks with considerable noise or coelution. Discrepancies between methods may also reflect potential sources of uncertainty inherent to manual approaches for determining peak areas; while GC-FID is a powerful tool for quantifying relative lipid biomarker abundance, it is indeed limited by the subjectivity of comparing retention times of an unknown peak to that of an external standard. While a more robust peak identification approach would entail access to their mass spectra, Chroma aims to reduce these sources of subjectivity especially when mass spectrometer measurements are not available, as is the case for the examples in this study. Certainly, such discrepancies could arise due to algorithmic inconsistencies in baseline or peak identifications, but in such instances, users of Chroma are recommended to investigate the output diagnosis figures and data generated by the code and adjust the input peak identification parameters accordingly (i.e., changing the peak detection thresholds or integration method).

Overall, the reprocessed CPI values using Chroma sufficiently reproduce the manually integrated values. The robust multiple linear regression of the manual and Chroma-derived CPI values returns a slope of 0.975 (r^2 = 0.868, p = 0.0366), and the majority of values are centered near the intersection of the regression and 1:1 reference line (Fig. 7). Empirical cumulative distribution functions (ECDF) for each profile are shown in Fig. 8, which show that the calculated CPI values mostly fall between 1 and 2. We also examined the retention time residuals for paired sample and external standard homolog peaks to assess the consistency of peak identifications (see Fig. S4 in the supplementary documentation); this built-in diagnostic evaluation of Chroma's output shows that the retention time offset of equivalent homolog peaks in the standard and sample are generally consistent, ranging between ~0.2 and 5 s depending on the sample.

4.2. Advantages of chroma for improved biomarker analysis

The principal advantage of the Chroma package is the ability to generate index calculations efficiently and consistently from automatically integrated biomarker peaks of GC-FID chromatogram data. The package is equipped for the synthesis of large chromatogram datasets across multiple samples and aims to reduce sources of uncertainty and irreproducibility associated with manual integration methods. Tools for data visualization and figure-making are included in the package, which are optimized for post-analysis diagnostics and pre-processing evaluations. Chroma is open-source, flexible, and improves inter-laboratory communications and data sharing, requiring only the chromatography

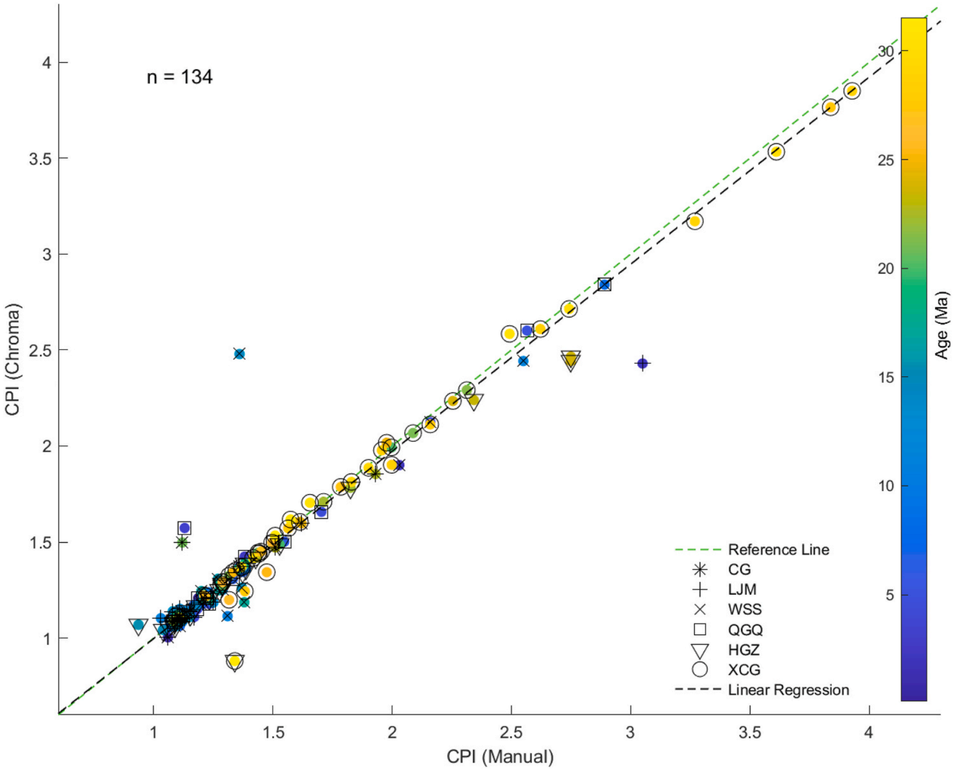

Fig. 7. Comparison of CPI values calculated manually (horizontal axis) and by Chroma (vertical axis) for all sections (134 samples). Robust multiple linear regression (black dashed line) and 1:1 (green dashed line) lines are plotted for reference (r^2 = 0.868, p = 0.0366). Ages of the samples are shown by the color gradient. All manual integrations are reported from Hou et al. (2021) and Wu et al. (2019), with the exception of the chromatograms of section XCG which were manually integrated for this study.

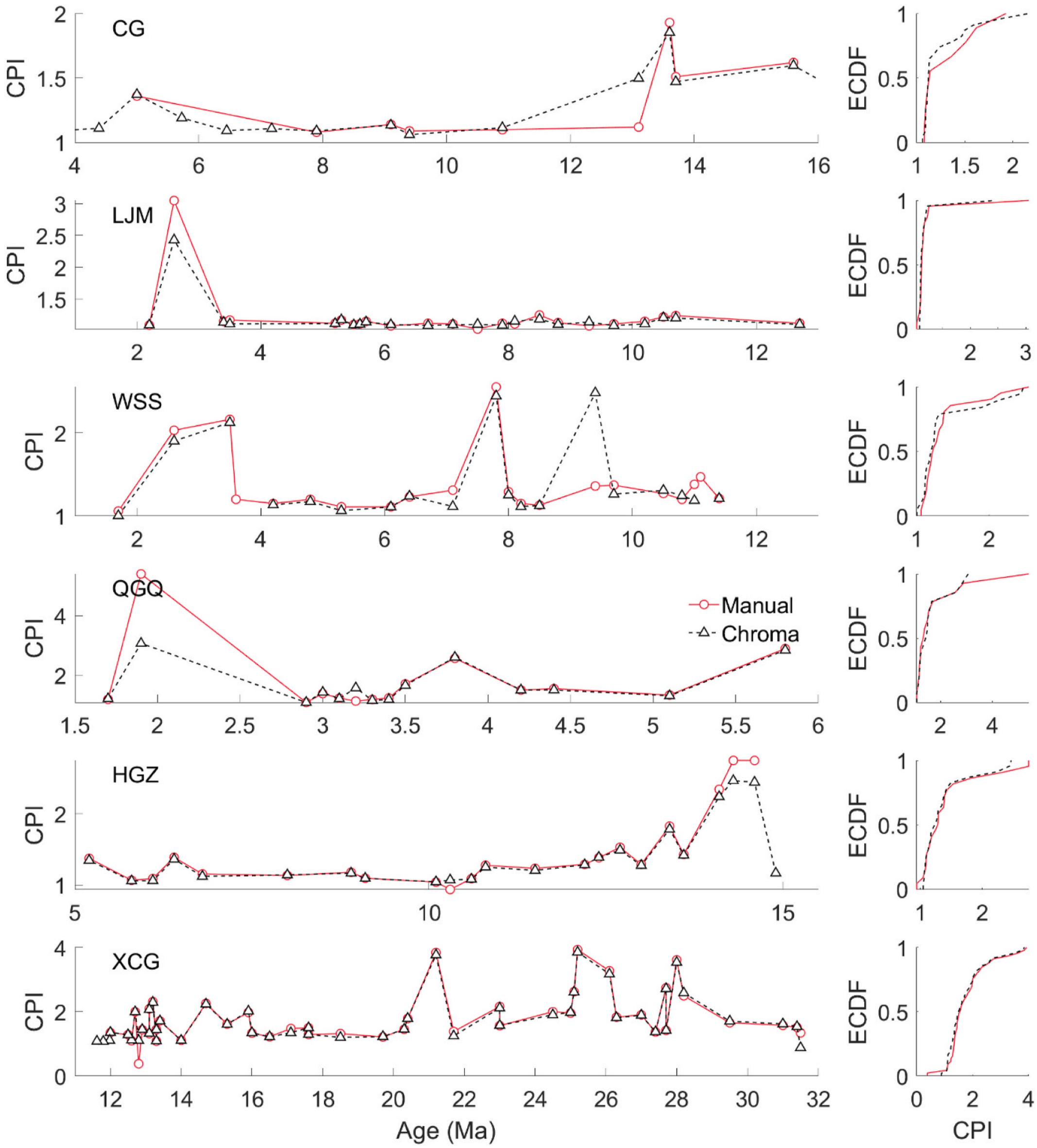

Fig. 8. Comparison of originally reported CPI values determined manually (red) from Hou et al. (2021) and Wu et al. (2019) and CPI values determined automatically by Chroma (black) for all sections. The average calculation time per sample in Chroma was ~0.08 s, including automatic peak identification and integration. Empirical cumulative distribution functions (ECDF) are shown to the right for each section, plotted against probability. The automated CPI calculations performed by Chroma are generally consistent with those calculated manually. Manual peak integrations for section XCG were performed in this study.

data system software for digitizing and exporting detector response signals into readable (e.g., .txt, .csv, .xlsx) file formats (Figs. 1 and 3). The full function suite for the Chroma package and additional data and tutorials are found in the supplementary materials.

Current software or third-party packages often cannot automatically produce hydrocarbon index calculations from raw chromatogram data, and typically do not carry flexible tools for output customization. Software applications commonly used for GC-FID analysis also do not typically carry options for automatic referencing of standard to sample peaks. The Chroma package also includes useful diagnostic tools for analysis-to-analysis comparison (e.g., before and after urea adduction) and the automatic identification of pristane and phytane. The target peak detection algorithm is capable of identifying peaks that are difficult to distinguish manually due to high background noise levels or relatively low abundance (Fig. 5). Chroma is designed to increase data production reliability and establish consistent processing criteria for all chromatogram analyses.

4.3. *Future development and current Limitations of chroma*

Chroma is designed to be adaptable and to continuously evolve as a comprehensive analysis application in all areas of biomarker research. Currently, the calculation of indices are limited to those of *n*-alkanes, but the package will continue to expand to include useful index calculations for other biomarkers – for example, additional functionalities for the calculation of the U^{k}_{37} index from the GC-FID analysis of alkenones (Brassell et al., 1986; Chen et al., 2021; Prahl et al., 1988). Some other lipid, pigment, and biomembrane biomarkers for potential addition to the base Chroma functions include the index calculations of *n*-alkanoic acids and *n*-alkanols, which have detectable abundances by GC-FID analysis (Freimuth et al., 2019; Simoneit, 2004). Options for automatic concentration calculations of individual homologs with the use of an internal standard may also be incorporated. Some disciplines in the study of *n*-alkanes may also benefit from expanded functionality for the analysis of crude oil and petrogenic contamination of ecosystems. While these extended uses will bring more capabilities for post-peak integration data production to the Chroma package, the base peak correlation and integration algorithm (Fig. 4) of the sample and standard is already applicable for all GC-generated chromatogram datasets. Updated versions with new additions to the base function suite will be added to the downloadable directory in the supplementary link.

Chroma performs optimally when external standard chromatograms contain well-defined peaks of only the targeted lipid compounds. Identification of contaminant peaks in the standard can lead to mismatched peak identifications or failure of the algorithm. Provided that the standard chromatogram is uncontaminated, the Chroma algorithm will return a detection of any standard-matched sample peak within the user-defined parameters by minimized retention time distance from the targeted compound. This can lead to misidentification in high-noise samples where the input detection parameters may not be sufficient to isolate target peaks from the background. We have found this to be a rare occurrence in our analyses, but it is generally advisable to exercise additional scrutiny of the diagnostic features in such cases.

4.4. *Other automatic peak integration approaches*

Previous automation approaches for GC trace analysis have effectively highlighted the advantages of programmable quantification and evaluation of the relative abundances of lipid compounds; namely, toolboxes for MATLAB such as TEXPRESS and ORIGAmI, developed for determination of the TEX_{86} and U^{k}_{37} sea-surface temperature indices using liquid or gas chromatographic techniques, demonstrate the capacity of standardizing peak integrations to generate quality-controlled outputs efficiently and reliably (Dillon and Huang, 2015; Fleming and Tierney, 2016). The peak integration methods in these packages rely on Gaussian models to describe the peak structure. While ideal detector responses for a given compound are well-captured by such models, we anticipate that varying GC conditions and sample quality may necessitate alternate modes of integration; Chroma contains multiple options for integration, including a Gaussian fit approach, for added flexibility. The code for TEXPRESS and ORIGAmI were specifically designed for the analysis of glycerol dialkyl glycerol tetraethers (GDGTs) or alkenone lipids, and to date there exists no equivalent software for the calculations of *n*-alkane indices.

5. Conclusions

We demonstrate the capabilities of the Chroma package in MATLAB for automatic detections, integrations, and processing of biomarker chromatogram data. This package is a powerful tool for generating reproducible and reliable index data more efficiently and precisely than many conventional methods. The Chroma package was able to complete the integration analysis and all index calculations of six stratigraphic sections in the Qaidam Basin and Hexi Corridor in under 0.08 s per sample. This approach offers a platform for better inter-laboratory index comparisons, provides flexible and customizable functionalities for performing multiple hydrocarbon index calculations over large datasets, and eliminates uncertainties introduced by manual peak characterization. The Chroma package was initially developed for the analysis of *n*-alkanes; however, additional tools may be added in the future for processing of other lipid fractions. The base MATLAB code and peripheral functions will be continuously updated and supplemented to incorporate new techniques for the characterization of biomarker distributions.

Code availability section

All reported data and functions are available in the Chroma GitHub repository. The directory includes downloadable files for the Chroma functions, all data used in this study, and documentation including tutorial vignettes for function descriptions and basic usage. Additional information or data may be made available upon request.

Contact: jtraph1@lsu.edu.

Program language: MATLAB.

The source codes are available for downloading at the link: https://github.com/jwt218/Chroma.

CRediT authorship contribution statement

Julian Traphagan: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Guangsheng Zhuang:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Data curation, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

Zhuang acknowledges financial support from U.S. National Science Foundation (NSF) grants EAR 2022282 and OCE 2217529 and LSU Robey H Clark Distinguished Professorship funds. Traphagan thanks GSA Graduate Student Research Grant 00012457. The authors would also like to sincerely thank the reviewers for their time and constructive feedback on the development of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2024.105675.

1. Fig. 1: Schematic flowchart of basic calculations performed in Chroma. Operations in the Chroma package are outlined by yellow regions. Example sample and standard chromatograms are shown below each step description. Peak areas are reported in units of intensity and time (fA x min). Homologs of the *n*-alkanes are labeled by number. STD = reference standard; *RT* = retention time (min); HI = hydrocarbon index; CPI = carbon preference index ($C_{23\text{-}33}$); ACL = average chain length ($C_{16\text{-}33}$).
2. Fig. 2: (A) Map of the Northern Tibetan Plateau and stratigraphic sections in the Qaidam Basin and Hexi Corridor (orange shaded regions). The western Qaidam Basin contains the Qigequan (QGQ), Honggouzi (HGZ), and Xichagou (XCG) sections. The Hexi Corridor contains the Caogou (CG), Laojumiao (LJM), and Wenshushan (WSS) sections. Solid black lines delineate major faults. (B) Inset map of the Tibetan Plateau.
3. Fig. 3: Schematic diagram of a GC-FID system. A summary of the GC oven ramp sequence is illustrated at the bottom.
4. Fig. 4: Example of sample processing flow in chroma. (A) Baseline subtraction using a not-a-knot cubic spline interpolation of the local minima. (B) Minimum peak height threshold, below which all peaks are filtered out. (C) Detection window for standard to sample peak cross-referencing. The parameter *ds* is defined as the number of data points (i.e., distance from the standard peak center) within which a sample peak detection is permitted. Sample peaks outside of this window are filtered out of the analysis. (D) Valley-to-valley base identification and peak integration.
5. Fig. 5: Examples of automatic target peak detection criteria in chroma for a noisy sample (QGQ 11) and non-noisy sample (QGQ 04) from section QGQ. Sample peaks are filtered by minimum peak

height threshold (blue dashed line) and cross-referencing to the standard peaks (vertical dashed lines). Solid vertical lines and blue shaded areas indicate the boundaries of the sample detection window (*ds*). Homologs in the standard are correlated with the corresponding sample peak to assign the homolog number (e.g., C_{25}, C_{27}, etc.) labeled above each peak. Mix-A6 was used as the standard for analysis (homologs $C_{16\text{-}30}$). Additional compounds C_{31} and C_{32} were added based on the approximate timing of peaks observed in the samples. If not included, these peaks will not be detected.

6. Fig. 6: Selected hydrocarbon indices of the Qaidam Basin (blue) and Hexi Corridor (red) determined by Chroma. The total analysis time including peak identifications, integrations, and index calculations was ~11.5 s for 152 samples across 6 sections using *chromall.* Trend lines were calculated by polynomial regression.
7. Fig. 7: Comparison of CPI values calculated manually and by Chroma for all sections (134 samples). Robust multiple linear regression (black dashed line) and 1:1 (green dashed line) lines are plotted for reference. Ages of the samples are shown by the color gradient. All manual integrations are reported from Hou et al. (2021) and Wu et al. (2019), with the exception of the chromatograms of section XCG which were manually integrated for this study.
8. Fig. 8: Comparison of originally reported CPI values determined manually (red) from Hou et al. (2021) and Wu et al. (2019) and CPI values determined automatically by Chroma (black) for all sections. The average calculation time per sample in Chroma was ~0.08 s, including automatic peak identification and integration. Empirical cumulative distribution functions (ECDF) are shown to the right for each section, plotted against probability. The automated CPI calculations performed by Chroma are generally consistent with those calculated manually. Manual peak integrations for section XCG were performed in this study.

References

Aichner, B., et al., 2018. Leaf wax n -alkane distributions record ecological changes during the Younger Dryas at Trzechowskie paleolake (Northern Poland) without temporal delay. Clim. Past Discuss 1–29.

Brassell, S.C., Eglinton, G., Marlowe, I.T., Pflaumann, U., Sarnthein, M., 1986. Molecular stratigraphy: a new tool for climatic assessment. Nature 320 (6058), 129–133.

Bray, E.E., Evans, E.D., 1961. Distribution of n-paraffins as a clue to recognition of source beds. Geochem. Cosmochim. Acta 22 (1), 2–15.

Bush, R.T., McInerney, F.A., 2013. Leaf wax n-alkane distributions in and across modern plants: implications for paleoecology and chemotaxonomy. Geochem. Cosmochim. Acta 117, 161–179.

Chen, X., et al., 2021. A potential suite of climate markers of long-chain n-alkanes and alkenones preserved in the top sediments from the Pacific sector of the Southern Ocean. Prog. Earth Planet. Sci. 8 (1), 23.

Cranwell, P.A., 1973. Chain-length distribution of n-alkanes from lake sediments in relation to post-glacial environmental change. Freshw. Biol. 3 (3), 259–265.

Dillon, J.T., Huang, Y., 2015. TEXPRESS v1.0: a MATLAB toolbox for efficient processing of GDGT LC–MS data. Org. Geochem. 79, 44–48.

Duan, Y., Jinxian, H., 2011. Distribution and isotopic composition of n-alkanes from grass, reed and tree leaves along a latitudinal gradient in China. Geochem. J. 45.

Eglinton, G., Hamilton, R.J., 1967. Leaf Epicuticular waxes. Science 156 (3780), 1322–1335.

Égüez, N., Mallol, C., Makarewicz, C.A., 2022. n-Alkanes and their carbon isotopes (δ13C) reveal seasonal foddering and long-term corralling of pastoralist livestock in eastern Mongolia. J. Archaeol. Sci. 147, 105666.

El Nemr, A., Moneer, A.A., Ragab, S., El Sikaily, A., 2016. Distribution and sources of n-alkanes and polycyclic aromatic hydrocarbons in shellfish of the Egyptian Red Sea coast. The Egyptian Journal of Aquatic Research 42 (2), 121–131.

Elson, A.L., Rohrssen, M., Marshall, J., Inglis, G.N., Whiteside, J.H., 2022. Hydroclimate variability in the United States continental interior during the early Eocene Climatic Optimum. Palaeogeogr. Palaeoclimatol. Palaeoecol. 595, 110959.

Feakins, S.J., et al., 2016. Production of leaf wax n-alkanes across a tropical forest elevation transect. Org. Geochem. 100, 89–100.

Ficken, K.J., Li, B., Swain, D.L., Eglinton, G., 2000. An n-alkane proxy for the sedimentary input of submerged/floating freshwater aquatic macrophytes. Org. Geochem. 31 (7), 745–749.

Fleming, L., Tierney, J., 2016. An automated method for the determination of the TEX86 and U37K` paleotemperature indices. Org. Geochem. 92, 84–91.

Freimuth, E.J., Diefendorf, A.F., Lowell, T.V., Wiles, G.C., 2019. Sedimentary n-alkanes and n-alkanoic acids in a temperate bog are biased toward woody plants. Org. Geochem. 128, 94–107.

Herrera-Herrera, A.V., Leierer, L., Jambrina-Enríquez, M., Connolly, R., Mallol, C., 2020. Evaluating different methods for calculating the Carbon Preference Index (CPI): implications for palaeoecological and archaeological research. Org. Geochem. 146, 104056.

Hoefs, M.J.L., Rijpstra, W.I.C., Sinninghe Damsté, J.S., 2002. The influence of oxic degradation on the sedimentary biomarker record I: evidence from Madeira Abyssal Plain turbidites. Geochem. Cosmochim. Acta 66 (15), 2719–2735.

Hou, M., Zhuang, G., Wu, M., 2021. Isotopic fingerprints of mountain uplift and global cooling in paleoclimatic and paleoecological records from the northern Tibetan Plateau. Palaeogeogr. Palaeoclimatol. Palaeoecol. 578, 110578.

Jeng, W.-L., 2006. Higher plant n-alkane average chain length as an indicator of petrogenic hydrocarbon contamination in marine sediments. Mar. Chem. 102 (3), 242–251.

Kang, M., Kim, K., Choi, N., Kim, Y., Lee, J., 2020. Recent occurrence of PAHs and n-alkanes in PM2.5 in Seoul, Korea and Characteristics of their sources and Toxicity. Int. J. Environ. Res. Publ. Health 17, 1397.

Kent-Corson, M.L., et al., 2009. Stable isotopic constraints on the tectonic, topographic, and climatic evolution of the northern margin of the Tibetan Plateau. Earth Planet Sci. Lett. 282 (1), 158–166.

Lee, D.-H., et al., 2019. Evaluation of alkane indexes for quantifying organic source from end member mixing experiments based on soil and algae. Ecol. Indicat. 107, 105574.

Leider, A., Hinrichs, K.-U., Schefuß, E., Versteegh, G.J.M., 2013. Distribution and stable isotopes of plant wax derived n-alkanes in lacustrine, fluvial and marine surface sediments along an Eastern Italian transect and their potential to reconstruct the hydrological cycle. Geochem. Cosmochim. Acta 117, 16–32.

Li, C., et al., 2020. Assessment of the relationship between ACL/CPI values of long chain n-alkanes and climate for the application of paleoclimate over the Tibetan Plateau. Quat. Int. 544, 76–87.

Marzi, R., Torkelson, B.E., Olson, R.K., 1993. A revised carbon preference index. Org. Geochem. 20 (8), 1303–1306.

Niedermeyer, E.M., et al., 2016. The stable hydrogen isotopic composition of sedimentary plant waxes as quantitative proxy for rainfall in the West African Sahel. Geochem. Cosmochim. Acta 184, 55–70.

Ofiti, N., et al., 2021. Warming promotes loss of subsoil carbon through accelerated degradation of plant-derived organic matter. Soil Biol. Biochem. 156, 108185.

Ortiz, J.E., et al., 2021. Bulk and compound-specific δ13C and n-alkane indices in a palustrine intermontane record for assessing environmental changes over the past 320 ka: the Padul Basin (Southwestern Mediterranean realm). J. Iber. Geol. 47 (4), 625–639.

Poynter, J., Eglinton, G., 1990. 14. Molecular composition of three sediments from hole 717c: the Bengal fan. Proceedings of the Ocean Drilling Program. Scientific results, pp. 155–161.

Prahl, F.G., Muehlhausen, L.A., Zahnle, D.L., 1988. Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. Geochem. Cosmochim. Acta 52 (9), 2303–2310.

Scalan, E.S., Smith, J.E., 1970. An improved measure of the odd-even predominance in the normal alkanes of sediment extracts and petroleum. Geochem. Cosmochim. Acta 34 (5), 611–620.

Seki, O., et al., 2012. Assessment for paleoclimatic utility of terrestrial biomarker records in the Okhotsk Sea sediments. Deep-sea Research Part Ii-topical Studies in Oceanography - DEEP-SEA RES PT II-TOP ST OCE 61–64.

Seki, O., Meyers, P.A., Kawamura, K., Zheng, Y., Zhou, W., 2009. Hydrogen isotopic ratios of plant wax n-alkanes in a peat bog deposited in northeast China during the last 16kyr. Org. Geochem. 40 (6), 671–677.

Simoneit, B.R.T., 2004. Biomarkers (molecular fossils) as geochemical indicators of life. Adv. Space Res. 33 (8), 1255–1261.

Simoneit, B.R.T., et al., 1991. Molecular marker study of extractable organic matter in aerosols from urban areas of China. Atmospheric Environment. Part A. General Topics 25 (10), 2111–2129.

Stolpnikova, E., Kovaleva, N., Kovalev, I., 2020. n-Alkane distribution—a Paleovegetation change indicator during the Period from late glacial to late Holocene on Russian plain (Bryansk region). Geosciences 10 (3), 86.

Struck, J., et al., 2020. Leaf wax n-alkane patterns and compound-specific δ13C of plants and topsoils from semi-arid and arid Mongolia. Biogeosciences 17, 567–580.

Thomas, C.L., Jansen, B., van Loon, E.E., Wiesenberg, G.L.B., 2021. Transformation of n-alkanes from plant to soil: a review. SOIL 7 (2), 785–809.

Tibbett, E.J., et al., 2021. Late Eocene record of hydrology and temperature from Prydz Bay, East Antarctica. Paleoceanogr. Paleoclimatol. 36 (4), e2020PA004204.

Wang, X., Huang, X., Sachse, D., Ding, W., Xue, J., 2016. Molecular paleoclimate reconstructions over the last 9 ka from a peat sequence in South China. PLoS One 11 (8), e0160934.

Wu, M., Zhuang, G., Hou, M., Liu, Z., 2021. Expanded lacustrine sedimentation in the Qaidam Basin on the northern Tibetan Plateau: Manifestation of climatic wetting during the Oligocene icehouse. Earth Planet Sci. Lett. 565, 116935.

Wu, M., Zhuang, G., Hou, M., Miao, Y., 2019. Ecologic shift and aridification in the northern Tibetan Plateau revealed by leaf wax n-alkane δ2H and δ13C records. Palaeogeogr. Palaeoclimatol. Palaeoecol. 514, 464–473.

Yan, Y., Zhao, B., Xie, L., Zhu, Z., 2021. Trend reversal of soil n-alkane Carbon Preference Index (CPI) along the precipitation gradient and its paleoclimatic implication. Chem. Geol. 581, 120402.

Zech, M., et al., 2009. Reconstructing Quaternary vegetation history in the Carpathian Basin, SE Europe, using n-alkane biomarkers as molecular fossils. Quaternary Science Journal 58, 148–155.

Zhang, H., Wang, R., Xiao, W., 2017. Paleoenvironmental implications of Holocene long-chain n-alkanes on the northern Bering sea slope. Acta Oceanol. Sin. 36 (8), 137–145.