



# Genome-inspired molecular identification in organic matter via Raman spectroscopy



Yun Liu <sup>a,\*</sup>, Nicola Ferralis <sup>a,\*\*</sup>, L. Taras Bryndzia <sup>b</sup>, Jeffrey C. Grossman <sup>a,\*\*\*</sup>

<sup>a</sup> Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

<sup>b</sup> Shell International Exploration and Production Inc., Houston, TX, United States

## ARTICLE INFO

### Article history:

Received 29 October 2015

Received in revised form

2 February 2016

Accepted 4 February 2016

Available online 6 February 2016

## ABSTRACT

Rapid, non-destructive characterization of molecular level chemistry for organic matter (OM) is experimentally challenging. Raman spectroscopy is one of the most widely used techniques for non-destructive chemical characterization, although it currently does not provide detailed identification of molecular components in OM, due to the combination of diffraction-limited spatial resolution and poor applicability of peak-fitting algorithms. Here, we develop a genome-inspired collective molecular structure fingerprinting approach, which utilizes *ab initio* calculations and data mining techniques to extract molecular level chemistry from the Raman spectra of OM. We illustrate the power of such an approach by identifying representative molecular fingerprints in OM, for which the molecular chemistry is to date inaccessible using non-destructive characterization techniques. Chemical properties such as aromatic cluster size distribution and H/C ratio can now be quantified directly using the identified molecular fingerprints. Our approach will enable non-destructive identification of chemical signatures with their correlation to the preservation of biosignatures in OM, accurate detection and quantification of environmental contamination, as well as objective assessment of OM with respect to their chemical contents.

© 2016 Elsevier Ltd. All rights reserved.

## 1. Introduction

Organic matter (OM) is a typical example of a complex, highly chemically diverse geological material whose complexity reflects the inherently chemically diverse biological origin. For such material, current chemical characterization techniques provide either atomic scale identification of individual molecular structures at the cost of losing spatial resolution, or the spatial distribution of average composition at the microscale (Fig. 1). Among all chemical characterization techniques, Raman spectroscopy is unique since it is a non-destructive method capable of identifying the molecular level chemistry of materials through the identification of the vibrational signatures associated with atomic and electronic structures and specific types of chemical bonds [1–3]. However,

when applied to heterogeneous organic matter, the resolution of chemistry from Raman spectroscopy is limited to the average compositional information at the micron scale [4–6], and often based on empirical and indirect correlative analyses with other techniques [4,7–12]. The core of such limitations stems from our inability to process the large amount of information contained within the Raman spectra, which is the collective response of potentially many thousands of possible organic molecular structures [13]. For carbonaceous materials with high aromaticity, this collective response in Raman manifests as two broad bands in the 1100–1700  $\text{cm}^{-1}$  region (Fig. 2a), representative of molecular vibrations among different bonding configurations of carbon and hydrogen atoms. The band in the 1150–1400  $\text{cm}^{-1}$  region is called the D peak, while the one in the 1550–1600  $\text{cm}^{-1}$  region is called the G peak [13–15]. These two bands are empirically associated with the shape and the size of the aromatic structures [5,13]. To date, the averaging across a massively large set of spectral features into single broad bands has prevented the association of any spectral feature to a particular molecular component in these materials.

Here we propose a genome-inspired molecular chemistry

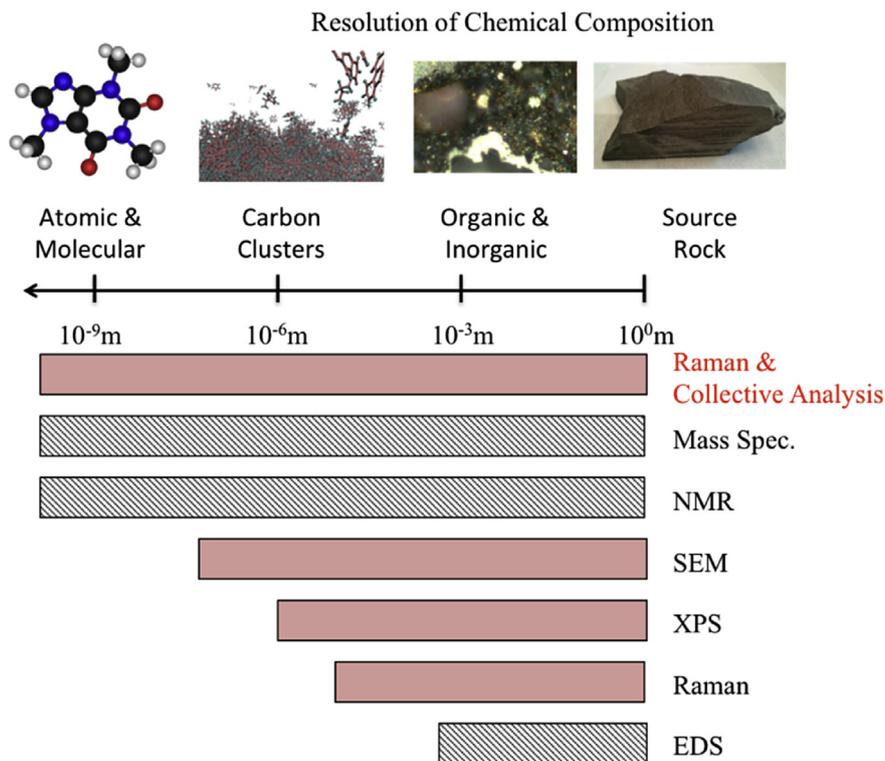
\* Corresponding author.

\*\* Corresponding author.

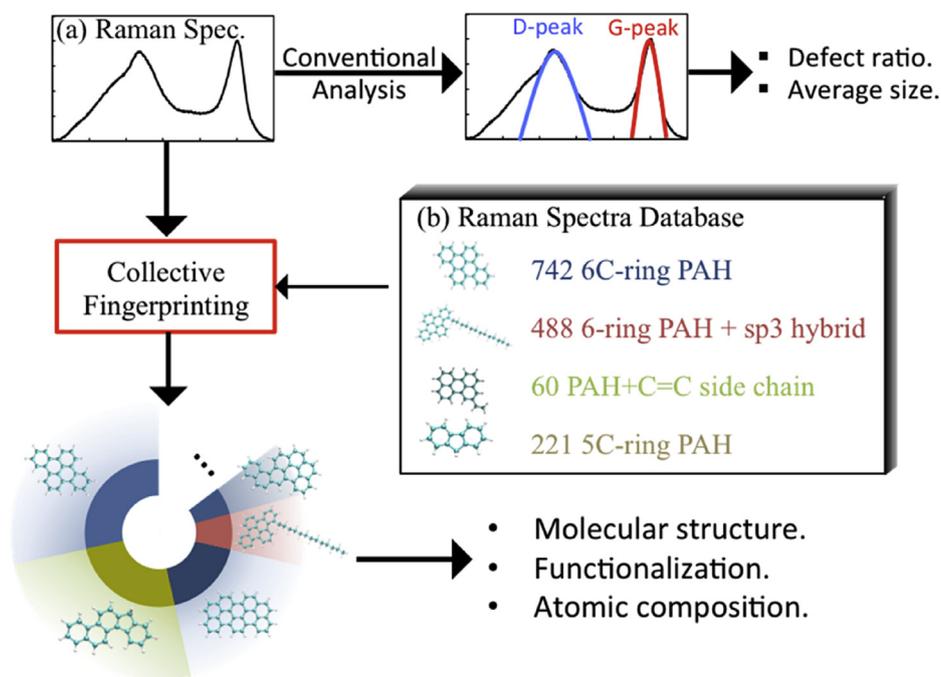
\*\*\* Corresponding author.

E-mail addresses: [yunl@mit.edu](mailto:yunl@mit.edu) (Y. Liu), [ferralis@mit.edu](mailto:ferralis@mit.edu) (N. Ferralis), [jcg@mit.edu](mailto:jcg@mit.edu) (J.C. Grossman).

<sup>1</sup> These authors contributed equally to this work.



**Fig. 1.** Comparison of current experimental characterization techniques for organic matter with their resolution of chemistry. Techniques that can provide spatial information (original location of the content) are marked with solid red bars (and shaded bars for those which cannot, respectively). Current techniques sacrifice spatial information for resolving molecular structures. Our collective analysis approach provides not only identification of molecule structures, but also their original spatial locations. Molecular structures are visualized using the VMD software package [16]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



**Fig. 2.** Comparison between the conventional analysis and the collective fingerprinting approach we are taking to resolve the molecular level chemistry information from a regular Raman spectrum. (a) A typical experimental Raman spectrum of chemically heterogeneous carbonaceous materials in the 1100–1700  $\text{cm}^{-1}$  region. (b) List of molecular structures currently available in the “kerogen genome database” [17]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analysis approach that combines *ab initio* calculations with data-mining techniques to explicitly capture the full molecular

complexity from the Raman spectrum. Unlike the top-down methods of conventional analysis algorithms that decompose broad spectral features into peaks with specific functional shapes (Fig. 2), our approach analyzes the information from a bottom-up approach, by searching for ensembles of molecular structures of which the collective Raman spectra reproduce the broad spectral features observed. Analogous to identifying genes to understand the functionality of living species, we identify “molecular fingerprints” to understand the chemistry of materials, where each “molecular fingerprint” corresponds to a single molecular structure within the ensemble. The identified ensemble of molecular fingerprints essentially serves as the “genome” of molecular chemistry for the target material, which can then be used for further analysis of materials properties and functionalities.

## 2. Computational and experimental details

### 2.1. Overview

Our approach is realized in two steps. Using *ab initio* calculations, we first build a library of computationally derived Raman spectra for a large number of molecules, the variety of which in principle covers the potential compositional phase space of the target material. Specifically for the analysis of the molecular chemistry of OM, we built a “kerogen genome database” [17] containing 1511 hydrocarbon molecules (Fig. 2b). Construction of these hydrocarbon compounds is described in Section 2.2. Next, we identify molecular fingerprints using data mining techniques to determine the correlation between the experimental Raman spectrum and the Raman spectra of ensembles of molecules from the library. The ensemble with the highest correlation is most closely identified as the molecular chemistry of the target material. Specifically, our “collective fingerprinting analysis” is performed by searching for the ensemble of molecules that minimizes the normalized deviation  $D$  between the experimental Raman spectrum and the collective Raman spectrum of molecules:

$$D = \frac{\sum (I_{\text{exp}} - \sum \beta_i I_i)^2}{\sum I_{\text{exp}}^2} \quad (1)$$

where  $I_{\text{exp}}$  is the experimental Raman intensity at a given wavenumber,  $I_i$  is the Raman intensity of the  $i$ -th molecule in the ensemble at the same wavenumber,  $\beta_i$  is the linear coefficient for the  $i$ -th molecule, and the summation runs through all wavenumber points of the experimental Raman spectrum. The value of  $D$  ranges between 0 and 1 [18]. From our results, Raman spectra comprised of ensembles of molecules with  $D$  values below 0.1 typically represent well the main features of the experimental spectrum. However, practically, the exponential number of possible combinations with respect to the ensemble size rapidly overwhelms any attempt at a sequential search through the Raman library. To overcome this hurdle, a genetic algorithm [19] is employed to accelerate the search (see Section 2.4).

### 2.2. Construction of hydrocarbon molecular structures with random walk algorithm

To ensure a systematic generation of hydrocarbon molecules with various types of aromatic and aliphatic components, we designed an algorithm that takes into account independently the aromatic size, aliphatic ratio, and the position where the aliphatic chains are attached to the aromatic flake. The aromatic flakes are generated from a random walk on a two dimensional graphene lattice. The aliphatic chains are generated at a random length

between  $[1, N]$  until the total chain length reaches  $N$ , with  $N$  being the designated total number of aliphatic carbon atoms. Redundant generations are detected and removed via a “rotate and compare” algorithm, which rotates the molecule by  $\pi/3$  each iteration and compares it with all existing molecules with the same aromatic size and aliphatic ratio.

These 1511 hydrocarbon molecules currently available in the “kerogen genome database” can be categorized into four groups based on their molecular structures: 742 molecules are poly aromatic hydrocarbon (PAH) molecules purely composed of six-member rings, ranging from 1 (benzene) to 10 rings either from the NIST online library of PAH molecules [20,21] or generated computationally from our own random generation algorithm. 488 molecules are composed of six-ring PAH with fully saturated  $sp^3$  carbon side chains randomly attached to edges of the structures. 211 molecules are 5-carbon ring structures, also retrieved from the NIST online library. 60 PAH structures with unsaturated side chains containing  $C=C$  bonds are included to test the effects of extended conjugation with conjugated side chains.

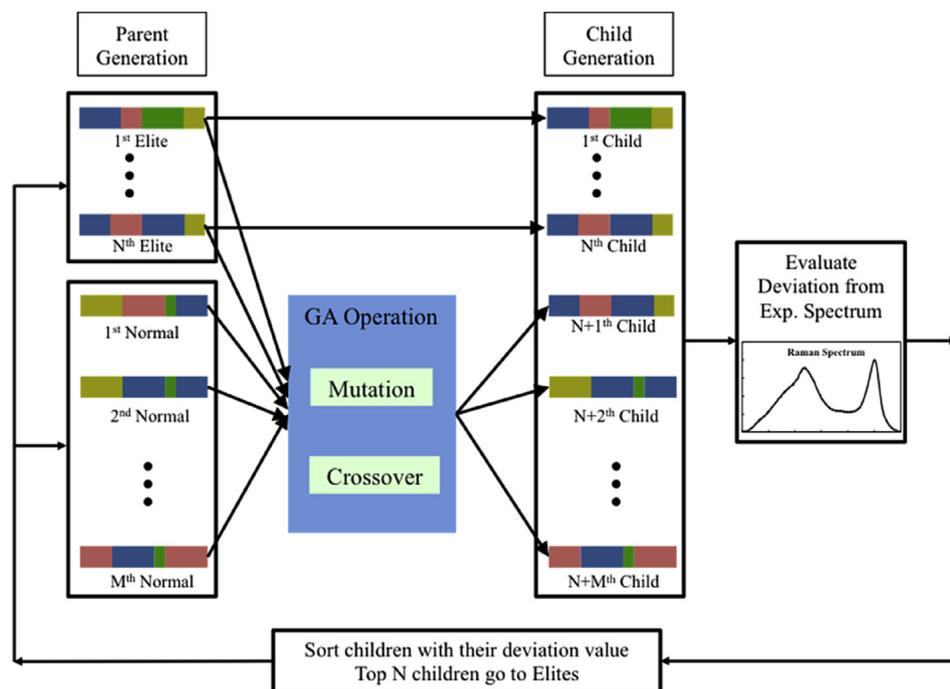
We note that the database employed here contains molecular compounds with only hydrogen and carbon. While this limits the ability to represent the complete hetero-chemistry of natural OM that contain other heteroatoms such as oxygen, nitrogen and sulfur, the database realistically captures the molecular bonding configurations represented in the relevant portion of the Raman spectra of OM ( $1100\text{--}1700\text{ cm}^{-1}$ ). We emphasize that the collective analysis approach itself is not limited by the size of the database and can be readily applied to expanded libraries containing additional molecular structures.

### 2.3. Calculation of Raman spectra

All the Raman spectra in this database are calculated with density functional theory as implemented in the Gaussian'09 package. We used the Perdew–Burke–Ernzerhof functional [22] to describe the exchange–correlation interaction of the electrons. Atomic orbitals are expanded with the cc-pVTZ basis set [23]. Raman activities are calculated from the differentiation of dipole derivatives with respect to the external electric field. All the molecular structures are fully relaxed using the Bery algorithm before calculating their Raman spectra. We compute Raman intensity from Raman activity following the same method used by Murphy et al. [24]. For visualization and comparison with experimental spectrum, discrete lines are converted to continuous Raman spectrum using Gaussian smearing of 5 wavenumbers.

### 2.4. Genetic algorithm

An overview of the genetic algorithm (GA) we implemented in our approach is shown in Fig. 3. At the beginning of each GA iteration, the best  $N$  parent ensembles ( $N = 30$ ) are carried over directly to the child generation, increasing the chance for its “genes” (or molecular structures) to “survive” (to participate in the next generation). The rest of the child generation is filled by random mutations and crossover from the full parent generation. Here, the two operations of “mutation” and “crossover” are defined as randomly swapping one molecular fingerprint in the ensemble with another one in the database, and exchanging half of the molecular fingerprints with another ensemble, respectively. In this process the best candidate of the child generation will be no worse than the best candidate of the parent generation, and results can improve as the genetic algorithm evolves. However, the use of only iterative crossover and mutation risks locking the ensemble into a local minimum far away from better solutions. To solve this problem, we introduced the “Genesis Flood” mechanism, wherein once every



**Fig. 3.** Flow chart of the genetic algorithm iterations as implemented in our collective analysis code. Each colored bar represents one ensemble of molecular fingerprints in which each color block represents one molecular fingerprint. Mutation and Crossover operations are explained in the Methods section. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

thousand generations, every ensemble in the parent generation is destroyed except for the most fit, and the rest of the population is filled with random ensembles. A typical GA run consists of 50,000 iterations in order to ensure convergence. More Details and the performance of our implementation can be found in the Supplementary Information.

### 2.5. Raman spectroscopy

Experimental Micro-Raman spectra were acquired using a Horiba LabRAM 800 HR spectrometer equipped with a He–Ne (632.817 nm) laser as the excitation source and a Peltier-cooled CCD detector. The laser was focused on the sample with a 400 nm confocal hole using the 100 $\times$  objective under reflected illumination. The laser spot on the sample was  $\sim$ 800 nm in diameter and had a power of  $\sim$ 4 mW at the sample surface. A calibrated edge high band filter (lowest wavenumber:  $\sim$ 70  $\text{cm}^{-1}$ ) was used to minimize the elastic backscattered signal. A minimum of 10 independent spots was analyzed on each sample and data were collected from 5 to 60 s per spot depending upon the Raman/Fluorescence intensity. The full spectral window for each acquisition is from  $-50$  to  $4000 \text{ cm}^{-1}$ . To reduce the amount of artifacts introduced by the background subtraction, due to the highly non-linear background over the full spectra, the spectra was broken down into several regions of interest. The first-order spectral window for the organic region was usually taken from  $1000$  to  $1800 \text{ cm}^{-1}$ . The background subtraction in this spectral window is performed using 2nd order polynomial functions.

### 2.6. Organic matter sample selection

OM samples are obtained from the Argonne Premium Coal Bank [25] and Penn State Coal Bank [26]. All samples were delivered in 1  $\mu\text{m}$  size powder sealed under Ar. To prevent oxidation, the coals were preserved in an inert environment when not actively used for

the experiments. Detailed information of all samples is listed in Table 1. While chemically heterogeneous at the molecular scale, the choice of type III kerogens, whose Raman spectra is homogenous and reproducible at the macroscale, is made to allow a meaningful validation of our approach, with available bulk scale comprehensive analytical information. As a result, the Raman spectra acquired at the microscale do not differ when acquired in different locations within the same specimen.

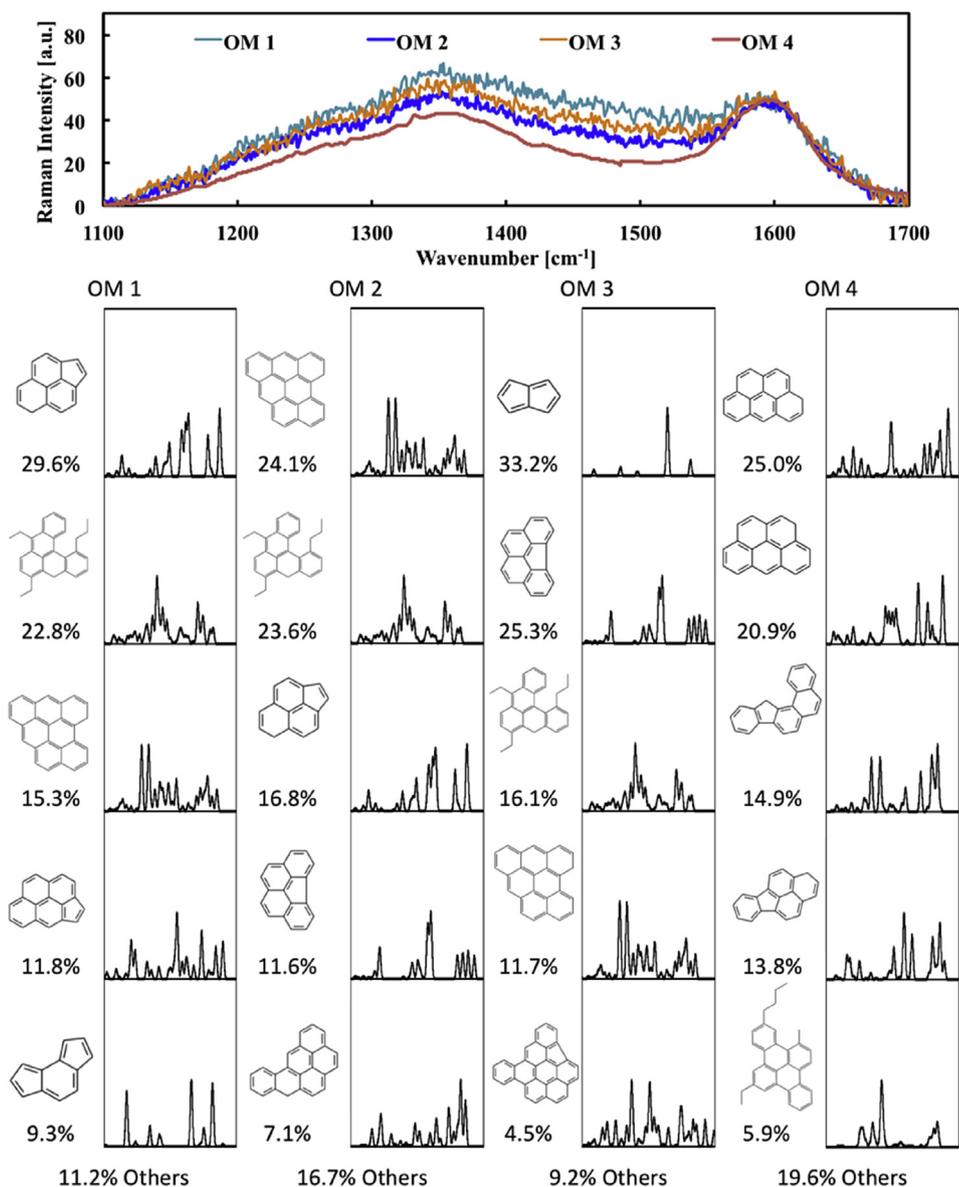
## 3. Results and discussion

To evaluate this approach, we perform collective fingerprinting analysis on the Raman spectra of four well-characterized plant-based OM (type III kerogen) with ascending maturity. An ensemble size of 12 is used in our fingerprinting analysis as it allows for the main features of the experimental Raman spectrum to be well represented (see Supplementary Fig. 1). For each sample, we carry out five separate GA searches and record the identified fingerprints. Fig. 4 shows one set of molecular fingerprints for each of the samples identified at the end of one collective analysis, together with the Raman responses of each molecular composition. Notably, a few key chemical quantities related to the molecular level chemistry can now be measured directly from the fingerprints

**Table 1**

Additional information for all four samples used in this study. HVB, MVB, and LVB are abbreviations for High-, Medium-, and Low-volatile bituminous, respectively. Vitrinite reflectance value (VRo) is a commonly used thermal maturity indicator, with higher VRo being geochemically more mature and vice versa.

	Source	VRo	Type
OM 1	Pittsburgh No. 8	0.72%	HVB
OM 2	Stockton seam	0.77%	HVB
OM 3	Upper Freeport	0.99%	MVB
OM 4	Pocahontas No. 3	1.71%	LVB



**Fig. 4.** List of top 5 out of 12 molecular fingerprints from one fingerprint set identified for each kerogen sample. Original Raman spectra of the four kerogen samples are shown on top. The molar ratio of each fingerprint in the 12-molecule fingerprint set is shown below the structure picture of each molecule. Raman spectrum of each molecular fingerprint in the 1100–1700 [cm<sup>-1</sup>] wavenumber regions is shown to the right of each fingerprint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

obtained. Consequently, by comparing the chemical quantities measured from the molecular fingerprints against the bulk values of the corresponding sample obtained using other experimental approaches, we can directly evaluate the accuracy of our fingerprints. For instance, by counting the number of aromatic carbon atoms in each molecule, we are able to calculate directly the average aromatic cluster size. Taking OM 1 as an example, its average aromatic cluster size calculated from our analysis is  $17 \pm 2$  C-atom/PAH, in good agreement with the value of  $15 \pm 3$  C-atom/PAH estimated from carbon-13 nuclear magnetic resonance (C NMR) spectroscopy [27]. Further, by counting the bridgehead aromatic carbon atoms in each structure, we calculate the average mole fraction of bridgehead aromatic carbons to be  $0.36 \pm 0.03$ , in close agreement with the value of  $0.31 \pm 0.06$  measured from the C NMR experiment [27]. The atomic hydrogen/carbon ratio, which is one of the most important parameters for determining the type and

origin of kerogen from the Van Krevelen diagram [28], is  $0.76 \pm 0.05$  measured from the collective search results, in good agreement with the estimated H/C ratio of 0.77 from elemental analysis [25]. The direct comparison and excellent agreement between C NMR and the results obtained through our genome-inspired Raman approach illustrates the effectiveness and accuracy of the fingerprinting method.

A full comparison between the chemical quantities measured from our collective analysis approach and experimental results for the samples examined in this work, are shown in Table 2. Note that the aromatic cluster sizes and H/C ratios calculated from our analysis are significantly different between different OM samples, and most of our predicted values agree well with the C NMR experiment. OM sample 2 shows a higher aromatic cluster size, which agrees with the lower D-peak height in its Raman spectrum but differs from C NMR experiments in the literature [27], possibly

**Table 2**  
Molecular chemistry quantities calculated from our collective fingerprinting analysis. Molar H/C ratio (“H/C”), aromatic cluster size (“AC”), and mole fraction of bridgehead aromatic carbon (“ $\chi_b$ ”) are in good agreement with the experimental values from the literature [25–27].

	OM no. 1		OM no. 2		OM no. 3		OM no. 4	
	Cal.	Exp.	Cal.	Exp.	Cal.	Exp.	Cal.	Exp.
H/C	0.76 ± 0.05	0.77	0.73 ± 0.04	0.76	0.72 ± 0.09	0.66	0.65 ± 0.02	0.65
AC	17 ± 2	15 ± 3	20 ± 1	14 ± 3	18 ± 2	18 ± 3	20 ± 2	N/A
$\chi_b$	0.36 ± 0.03	0.31 ± 0.06	0.41 ± 0.01	0.29 ± 0.06	0.39 ± 0.02	0.36 ± 0.03	0.38 ± 0.05	N/A

due to inhomogeneity of the sample. Our predicted H/C ratio is significantly lower for OM 4 than OM 1 or 2, in good agreement with elemental analysis experiments [25,26]. The aromatic cluster sizes are larger for OM 4 than other samples, which is not surprising as OM 4 is more mature than other samples used in this work (see section 2.6).

We also notice that among the molecular fingerprints identified for OM 4, less than 10% has an aliphatic component. In contrast, aliphatic compounds appear with a significant predicted percentage (>20%) in the fingerprints identified for OM 1 or OM 2. This comparison resonates with the fact that strong reduction of aliphatic components is a consequence of catagenesis, which leaves a heavily aromatic backbone as the major component in OM [29]. Furthermore, we have found aromatic carbon clusters with rings of five carbon atoms, which are major components in molecular representations of type III kerogens with similar maturity [30,31]. Five-carbon rings have a higher mole ratio of bridgehead carbon atoms than six-carbon ring aromatic clusters with the same number of carbon atoms. Molecules containing five-carbon ring structures were not explicitly considered in previous C NMR experiments [27]. One example of such molecules can be found in the fourth group in Fig. 2b. We note that although the comparison is focused on C NMR measurements, the method developed here can be complementary to other experimental techniques that probe C–H functionality in OM (such as FT-IR) or elemental composition (mass spectrometry).

#### 4. Conclusion

In this work, we present a genome-inspired method to identify from the Raman spectra the most representative molecular chemistry of complex heterogeneous carbonaceous materials. Successful application of our approach critically depends on two major components: an extensible Raman spectra database of individual molecular structures that covers the chemistry represented in the Raman spectra of the target material, and an efficient collective analysis algorithm that allows for identification of an ensemble of molecules. We show that using a “kerogen genome database” to represent the OM chemistry and using a genetic algorithm for collective analysis, it is possible to extract accurate molecular information from the OM Raman spectra: The micron-scale chemical fingerprinting can be directly compared with bulk chemical data. The genomic nature of our approach and the statistical convergence of the molecular fingerprints suggest that similarly to genetics for bio species, genetics of molecular structures can be established for OM to systematically quantify its molecular composition. The molecular structures identified through this method allows for the construction of representative molecular mixtures that can serve as the basis for large-scale atomistic simulations of OM [30–32]. More broadly, our work demonstrates the power of combining *ab initio* calculations and data-mining methods with information-rich experiments to elevate our understanding of the underlying fundamental science. While it is currently being applied to evaluate Raman spectra, the inclusion of

additional computed spectral data (infrared absorption, for instance) would allow for potential application of the same collective fingerprinting approach in other characterization techniques such as FT-IR.

#### Acknowledgments

This work is supported by Shell Oil Company and Extramural Research (EMR) funding provided by Shell Innovation Research and Development under the MIT Energy Initiative (grant number: PT32401).

#### Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.carbon.2016.02.017>.

#### References

- [1] M.B. Denton, R.P. Sperline, J.H. Giles, D.A. Gilmore, C.J.S. Pommier, R.T. Downs, Advances in the application of array detectors for improved chemical analysis, part I. Comparison of qualitative analyses using large, computer-based Raman spectral libraries, *Aust. J. Chem.* 56 (2–3) (2003) 117–131.
- [2] J.D. Rodriguez, B.J. Westenberger, L.F. Buhse, J.F. Kauffman, Quantitative evaluation of the sensitivity of library-based Raman spectral correlation methods, *Anal. Chem.* 83 (11) (2011) 4061–4067.
- [3] B.K. Lavine, J. Workman, *Chemom. Anal. Chem.* 85 (2) (2013) 705–714.
- [4] R.M. Bustin, J.V. Ross, J.N. Rouzaud, Mechanisms of graphite formation from kerogen: experimental evidence, *Int. J. Coal Geol.* 28 (1) (1995) 1–36.
- [5] S.R. Kelemen, H.L. Fang, Maturity trends in Raman spectra from kerogen and coal, *Energy Fuels* 15 (3) (2001) 653–658.
- [6] J. Jehlička, O. Urban, J. Pokorný, Raman spectroscopy of carbon and solid bitumens in sedimentary and metamorphic rocks, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 59 (10) (2003) 2341–2352.
- [7] G.D. Smith, R.J.H. Clark, Raman microscopy in archaeological science, *J. Archaeol. Sci.* 31 (8) (2004) 1137–1160.
- [8] E. Quirico, J.-N. Rouzaud, L. Bonal, G. Montagnac, Maturation grade of coals as revealed by Raman spectroscopy: Progress and problems, *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 61 (10) (2005) 2368–2377.
- [9] P. Vandenabeele, J. Jehlička, P. Vitek, H.G.M. Edwards, On the definition of Raman spectroscopic detection limits for the analysis of biomarkers in solid matrices, *Planet. Space Sci.* 62 (1) (2012) 48–54.
- [10] S. Potgieter-Vermaak, N. Maledi, N. Wagner, J.H.P. Van Heerden, R. Van Grieken, J.H. Potgieter, Raman spectroscopy for the analysis of coal: a review, *J. Raman Spectrosc.* 42 (2) (2011) 123–129.
- [11] A. Cuesta, P. Dhamelincourt, J. Laureyns, A. Martínez-Alonso, J.M.D. Tascón, Raman microprobe studies on carbon materials, *Carbon* 32 (8) (1994) 1523–1532.
- [12] X. Li, Y. Ju, Q. Hou, Z. Li, J. Fan, FTIR and Raman spectral research on metamorphism and deformation of coal, *J. Geol. Res.* 2012 (2012) 8.
- [13] A.C. Ferrari, J. Robertson, Interpretation of Raman spectra of disordered and amorphous carbon, *Phys. Rev. B* 61 (20) (2000) 14095–14107.
- [14] F. Tuinstra, J.L. Koenig, Raman spectrum of graphite, *J. Chem. Phys.* 53 (3) (1970) 1126–1130.
- [15] R.O. Dillon, J.A. Woollam, V. Katkanant, Use of Raman scattering to investigate disorder and crystallite formation in as-deposited and annealed carbon films, *Phys. Rev. B* 29 (6) (1984) 3482–3489.
- [16] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, *J. Mol. Graph.* 14 (1) (1996) 33–38.
- [17] Kerogen Genome Database. Available from: <https://sites.google.com/site/yunliuscj/databases>.
- [18] A. Kwiatkowski, M. Gnyba, J. Smulko, P. Wierzbza, Algorithms of chemicals detection using Raman spectra, *Metrol. Meas. Syst.* (2010) 549.
- [19] Genetic Algorithm for Collective Fingerprinting using Raman Spectroscopy. Available from: <https://sites.google.com/site/yunliuscj/code>.
- [20] L.C. Sander, S.A. Wise, Polycyclic aromatic hydrocarbon structure index, NIST

- Spec. Publ. (1997) 922.
- [21] T.C. Allison, NIST Polycyclic Aromatic Hydrocarbon Structure Index, Available from: <http://pah.nist.gov/>.
- [22] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* 77 (18) (1996) 3865–3868.
- [23] R.A. Kendall, T.H. Dunning, R.J. Harrison, Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions, *J. Chem. Phys.* 96 (9) (1992) 6796–6806.
- [24] W.F. Murphy, W. Holzer, H.J. Bernstein, Gas phase Raman intensities: a review of “Pre-Laser” data, *Appl. Spectrosc.* 23 (3) (1969) 211–218.
- [25] K.S. Vorres, The argonne premium coal sample program, *Energy Fuels* 4 (5) (1990) 420–426.
- [26] Penn State Coal Bank, Energy Institute, Penn State University, Available from: [http://www.energy.psu.edu/sites/default/files/files/CSB\\_FactSheet.pdf](http://www.energy.psu.edu/sites/default/files/files/CSB_FactSheet.pdf).
- [27] M.S. Solum, R.J. Pugmire, D.M. Grant, Carbon-13 solid-state NMR of Argonne-premium coals, *Energy Fuels* 3 (2) (1989) 187–193.
- [28] D.W. van Krevelen, Graphical-statistical method for the study of structure and reaction processes of coal, *Fuel* 29 (1950) 269–283.
- [29] B.P. Tissot, D.H. Welte, *Petroleum Formation and Occurrence*, second ed., Springer, 1984.
- [30] J.P. Mathews, A.C.T. van Duin, A.L. Chaffee, The utility of coal molecular models, *Fuel Process Technol.* 92 (4) (2011) 718–728.
- [31] J.P. Mathews, A.L. Chaffee, The molecular representations of coal – A review, *Fuel* 96 (0) (2012) 1–14.
- [32] N. Ferralis, Y. Liu, K.D. Bake, A.E. Pomerantz, J.C. Grossman, Direct correlation between aromatization of carbon-rich organic matter and its visible electronic absorption edge, *Carbon* 88 (0) (2015) 139–147.