PNAS RESEARCH ARTICLE | EARTH, ATMOSPHERIC, AND PLANETARY SCIENCES
BIOCHEMISTRY
OPEN ACCESS

Organic geochemical evidence for life in Archean rocks identified by pyrolysis–GC–MS and supervised machine learning

Michael L. Wong[a,b,1], Anirudh Prabhu[a,1], Conel O'D. Alexander[a], H. James Cleaves II[a,c], George D. Cody[a], Grethe Hystad[d], Marko Bermanec[e], Wouter Bleeker[f], C. Kevin Boyce[g], Andrea Corpolongo[h], Andrew D. Czaja[h], Souvik Das[i], Robert R. Gaines[j], Daniel D. Gregory[k], John A. Jaszczak[l], Emmanuelle J. Javaux[m], Jaganmoy Jodder[n,o], Andrew H. Knoll[p], Martin Van Kranendonk[q], Katie M. Maloney[r], Nora Noffke[s], Robert Rainbird[f], Emersyn Slaughter[t], Eva E. Stüeken[u], Roger E. Summons[v], Frances Westall[w], Jasmina Wiemann[x,y], Shuhai Xiao[z], and Robert M. Hazen[a,2]

Affiliations are included on p. 10.

Edited by John Valley, University of Wisconsin, Madison, WI; received June 4, 2025; accepted September 13, 2025

Throughout Earth's history, organic molecules from both abiogenic and biogenic sources have been buried in sedimentary rocks. Most of these organic molecules have been significantly altered by geologic processes through deep time. Nonetheless, the nature and distribution of those ancient fragmentary organic remains have the potential to reveal diagnostic biomolecular information after billions of years of burial. Here, we analyzed 406 fossil, modern biological, meteoritic, and synthetic samples using pyrolysis gas chromatography and mass spectrometry. We explored these analytical data via supervised machine-learning methods to discriminate samples of biogenic vs. abiogenic origin, plant vs. animal phylogenetic affinity, and photosynthetic vs. nonphotosynthetic physiology. Dividing 272 samples with known phylogenetic affinity and physiology into 9 categories, each further divided into 75% training and 25% testing sets, our random forest models accurately predict pairwise assignments of modern vs. fossil or meteoritic organics (100% correct assignments), fossil plant tissues vs. meteoritic organics (97%), modern vs. fossil plant tissues (98%), and modern plants vs. animal tissues (95%). Pairwise comparisons between fossil biogenic samples vs. abiogenic samples resulted in 93% correct classifications, while analysis of modern and ancient photosynthetic vs. nonphotosynthetic samples also resulted in 93% correct assignments. Our analyses demonstrate that molecular biosignatures can survive in ancient fossils and allow for the identification of organismal origins and traits. Consistent with previous morphological and isotopic inferences, we present evidence for biogenic molecular assemblages in Paleoarchean rocks (3.33 Ga) and for photoautotrophy in Neoarchean rocks (2.52 Ga).

biosignatures | organic chemistry | machine learning | photosynthesis | meteorites

Ancient microfossils, isotopic biosignatures, and microbialites such as stromatolites and microbially induced sedimentary structures support the view that microbial life on Earth originated billions of years before the appearance of complex multicellular organisms (1–5). Molecular fossils have also played a key role in drawing inferences about phylogenetic and metabolic antiquity. However, unambiguous records of complex molecules such as lipids and porphyrins extend back only to about 1.6 billion years (6–10), less than half the age inferred for life from other convincing lines of evidence. Here, we employ an approach in which machine learning facilitates differentiation between highly degraded biogenic and abiogenic samples, significantly extending the record of biomolecule fossilization products. This approach, which has enabled inferences about phylogeny and physiology of Phanerozoic biota (<541 Ma) (11, 12), is here extended to Paleoarchean (>3.5 Ga) organics.

Recent applications of machine learning to the detection of biosignatures in ancient fossils and other molecular suites demonstrate that distributions of organic molecules in living systems are different from organic molecular suites produced by abiogenic processes (12–17). However, diagenesis and metamorphism over billions of years results in the degradation of diagnostic biomolecules (18–24), as well as all but the most robust morphological features (4, 25–28). While it has been experimentally and observationally demonstrated that both phylogenetic and physiological signatures in biomacromolecules can survive over tens to hundreds of millions of years (11, 12, 29–35), the biogenicity of organic molecules older than 2 billion years has not been demonstrated. We postulate that because biomolecules are characterized by their selection for function (36), even highly degraded biomolecular assemblages may retain useful paleobiological information distinct from what emerges from abiogenic processes, even when no individually diagnostic biomolecules are preserved.

Significance

Teasing out biochemical information from ancient organic-rich sediments, notably the timing of the emergence of photosynthesis relative to the inferred oxygenation of Earth's atmosphere, remains a challenging opportunity. To tackle this problem, we analyzed 406 diverse ancient and modern samples and used supervised machine learning to discriminate samples of biogenic vs. abiogenic origin, as well as photosynthetic vs. nonphotosynthetic physiology. Comparing organic-rich samples of uncertain affinity to our training data, ca. 3.33-billion-year-old sedimentary rocks group among microbial samples, and rocks as old as 2.52 billion years ally with more recent photosynthetic life. The application of supervised machine learning thus approximately doubles the interval within which fossil organic matter can be shown to retain molecular information of evolutionary relationships and physiology.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] M.L.W. and A.P. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: rhazen@carnegiescience.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2514534122/-/DCSupplemental.

Published November 17, 2025.

A previous study applied machine learning to pyrolysis–gas chromatography–mass spectrometry (py–GC–MS) analytical data to discriminate molecular suites of biogenic vs. abiogenic remains with ~90% accuracy (15). Here, we extend that approach with a larger sample set and additional specimen attributes, including taxonomic groupings, such as plants vs. animals, and metabolic strategies, notably photosynthesis, to extract biochemical information from ancient samples with highly degraded organic molecules. As in previous work (15, 16, 35), our method entailed four steps: (1) Collecting 406 diverse carbon-bearing samples (*SI Appendix*, Table S1) from varied modern and ancient biogenic and abiogenic sources; (2) concentrating carbonaceous macromolecular material through extraction from meteorites and ancient sedimentary rocks (37, 38); (3) analyzing each sample by pyrolysis gas chromatography coupled to electron impact ionization mass spectrometry (py–GC–MS; see *Analytical Methods*); and (4) training supervised random forest machine-learning models (39) using data from subsets of our sample analyses (*Machine-Learning Methods*). Each sample is represented by a two-dimensional matrix based on chromatographic retention time and mass-to-charge ratio with intensities for each of 489,240 matrix elements (Fig. 1).

The Sample Suite

We analyzed in total 406 natural and synthetic samples containing suites of organic molecules (*SI Appendix*, Table S1). The largest subset includes 141 organic-rich sedimentary rocks, predominantly

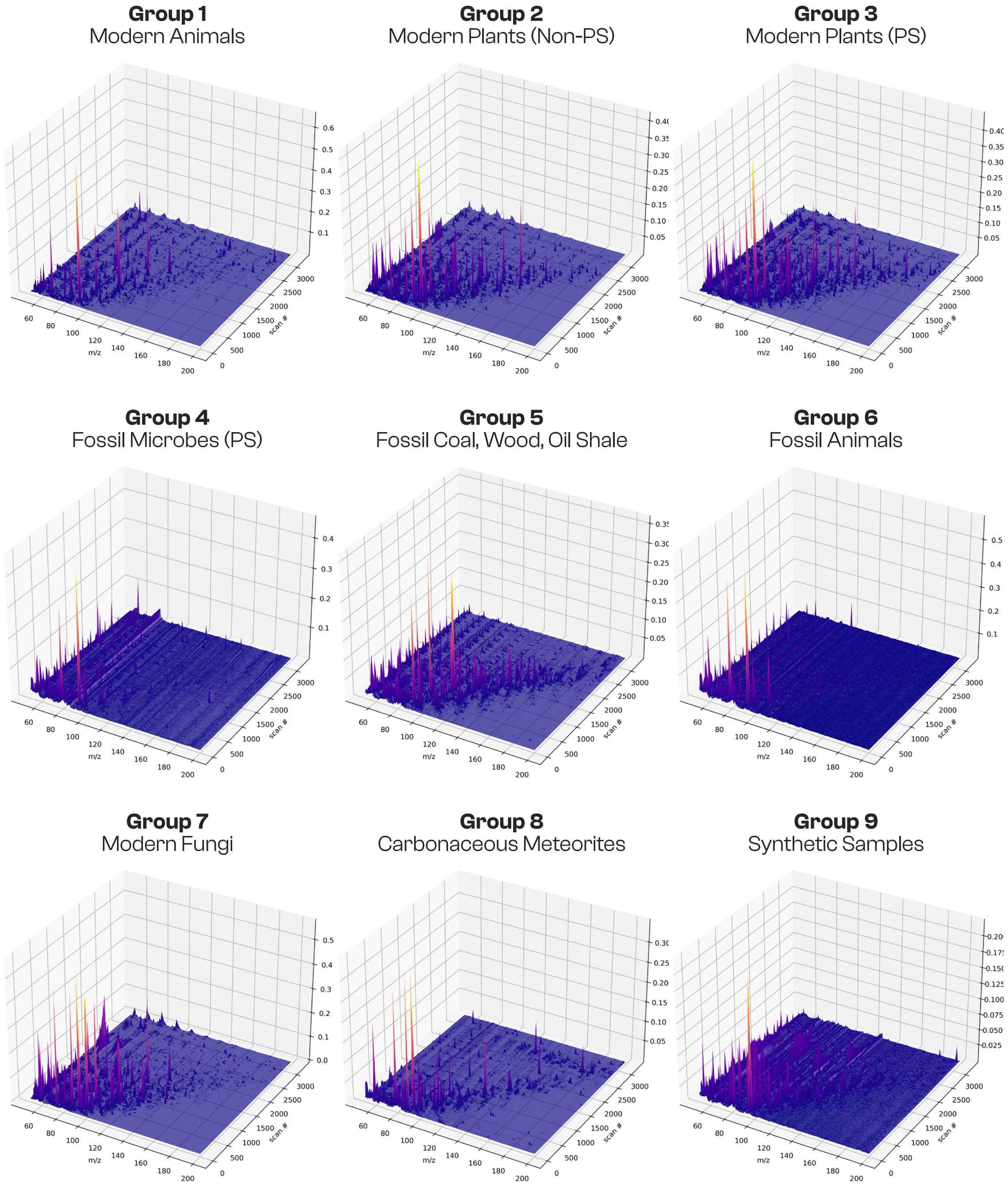

Fig. 1. Aggregated three-dimensional py–GC–MS data for samples in each of our nine categories: modern animals; modern plants (non-photosynthetic); modern plants (photosynthetic); fossil microbes (photosynthetic); fossil coal, wood, oil shale; fossil animals; modern fungi; carbonaceous meteorites; and synthetic samples. These graphs display peak intensities (vertical scale, normalized to the highest peak intensity in each category) for 3,240 elution time bins or "scans" (right-hand scale) and their mass spectra over 150 m/z bins (left-hand scale).

shales and cherts, ranging in age from ~3.8 Ga to 10 Ma. The sources of organic molecules in many of the Precambrian samples are uncertain. In addition, 65 samples represent fossil wood, coal, oil shale, and other primarily Phanerozoic organic-rich samples. Modern living land plants, animals, and fungi are represented by 123 specimens. Unambiguously abiogenic samples include organic molecular suites concentrated by chemical digestion from 42 meteorites (39 of which are carbonaceous chondrites), as well as 35 suites of organic molecules produced in laboratory synthesis experiments. Of these 406 samples, 272 can be confidently assigned to one of nine categories that we employed in supervised machine learning, as follows (Fig. 1):

1. *Modern Animals*: 21 samples from diverse recently deceased invertebrate and vertebrate animals.
2. *Modern Plants (non-photosynthetic tissues)*: 40 samples include roots, seeds, flowers, fruits, and tree sap—i.e., non-photosynthesizing plant tissues and secretions.
3. *Modern Plants (leaves)*: 36 samples of green leaves and other photosynthesizing tissues.
4. *Sedimentary Rocks with Confirmed Fossils of Photosynthetic Cyanobacteria or Algae*: 24 samples of HCl and HF acid-concentrated organic residues from shale or chert with reliable morphological evidence for cyanobacteria or algal fossils.
5. *Fossil Wood, Coal, and Oil shale*: 49 primarily Phanerozoic (<541 Ma) samples, as well as less well understood hydrocarbon-rich deposits in Proterozoic rocks such as "shungite" (40, 41) and "anthraxolite" (42–46).
6. *Animal Fossils*: 9 Phanerozoic samples, including carbonized remains of fossil fish and trilobites, as well as shell-binding protein dissolved from Miocene gastropods (47).
7. *Modern Fungi*: 16 samples, including various wood fungi and yeast.
8. *Meteorites*: 42 meteorites, mostly carbonaceous chondrites (e.g., ref. 48).
9. *Synthetic Samples*: 35 organic suites produced in Maillard reactions (49), formose reactions (50), and other laboratory synthesis processes.

We employed the 272 samples of Groups 1 to 9 in training and testing sets to explore the extent to which methods of supervised machine learning can discriminate between different suites of organic molecules. Note that we also designate two samples of modern cyanobacteria as Group 10 and one sample of the modern halophile *Halobacter* as Group 11, which are used in machine-learning Model #4 (*Photosynthetic vs. Non-photosynthetic Organisms*). Most of the remaining 131 samples are acid-concentrated residues from organic-bearing Archean or Proterozoic sedimentary rocks. Most, if not all, of these ancient samples are thought to have biogenic origins based on microfossil populations, isotopic abundances, and/or petrological evidence of microbial mats, but especially in our oldest samples, taxonomic and metabolic inferences have, to date, been limited (1, 3, 5–15, 51–61). Thus, these samples provide a useful testing ground for the application of our machine-learning analyses.

Pairwise Discriminations Among Nine Categories

As an initial test of our approach, we applied the random forest method, an established robust supervised approach to machine-learning. We built a binary classifier for each of the 36 pairwise sets of data among the nine categories of labeled samples. To evaluate the applicability of the model, we split the data into a randomly selected training set (75% of the samples) and test set (25% of the samples). To illustrate this approach, in the case of Group 3 (36 samples of modern plant leaves) vs. Group 8 (42 carbonaceous meteorites), the method distinguished plants from meteorites with 100% accuracy. We calculated a class probability from 0.00 to 1.00 for each of the 78 samples. (Note that class probability is calculated from the number of times out of 1,000 random forest cycles that a class is assigned. For example, in the above case "a meteorite class probability" of 0.763 means that a sample was classified as a meteorite in 763 of 1,000 trees, vs. a plant in 237 of 1,000 trees.) The extent of discrimination between these two groups is illustrated by a histogram with a well-defined bimodal separation (Fig. 2*A*). In this example, all samples have values of meteorite class probability either >0.60 or <0.40—a result that underscores the significant differences between the organic suites of modern plants vs. carbon-rich meteorites, as found in a previous study (15). Although random forest class predictions are based on a 0.50 class probability threshold, a value of class probability >0.60 or <0.40 is more reliable evidence for group membership owing to the reduced chances of class switch for lower probability predictions. To improve the reliability of our predictions, future studies will explore various probability thresholds to choose the optimal threshold based on a combination of factors including F1 score, ROC curve, and Youden's J index (62).

In contrast, when comparing the mass spectra of Group 2 (41 samples of non-photosynthetic plant tissues, such as roots, seeds, and tree sap) vs. the 36 green leaves of Group 3, the success of this method is only 79%, and the probability distribution of class assignments for this case is unimodal (Fig. 2*B*). This result reflects the molecular similarities of varied plant tissues and corresponding difficulty of discriminating between these two groups based on py–GC–MS alone.

Table 1 records the percentage of correct labels for pairwise tests of all 36 possible combinations of nine groups. In 25 of 36 tests, we achieved at least 90% correct assignments of both training and test sets, with 18 tests at least 95% correct. However, a problem often arises with the random forest approach when comparing two sample sets of very different sizes, as the algorithm will attempt to place most of the samples of the smaller group into the larger group. In nine cases with significantly different numbers of samples in the two groups, we note a failed discrimination with "F." All nine failed instances in Table 1 relate to pairings with the smallest two Groups 6 or 7, containing only 9 and 16 samples, respectively. For example, in the analysis of 42 meteorites vs. 9 fossil animals, 8 of the fossil animals were incorrectly assigned as meteorites. An objective for future studies is to increase the numbers of samples in these underrepresented groups. Other planned approaches to address class imbalance include downsampling by selecting the most informative samples in the given imbalanced dataset through an active learning strategy to mitigate the effect of imbalanced class labels (63) and adjusting the class probability threshold using a method such as GHOST (Generalized tHresh-Old ShifTing) (64).

The pairwise analyses represented in Table 1 suggest opportunities for discriminating among a wide variety of organic molecular suites. An important conclusion of this investigation of pairwise discrimination is that, given relatively balanced numbers of samples in two groups, machine-learning methods can successfully discriminate between two categories in most instances. Setting aside the underrepresented Group 6 (9 fossil animals) and Group 7 (16 fungi), this machine-learning method finds greater than 90% differences between plants and animals (Group 1 vs. Groups 2, 3, and 5), greater than 95% differences between most

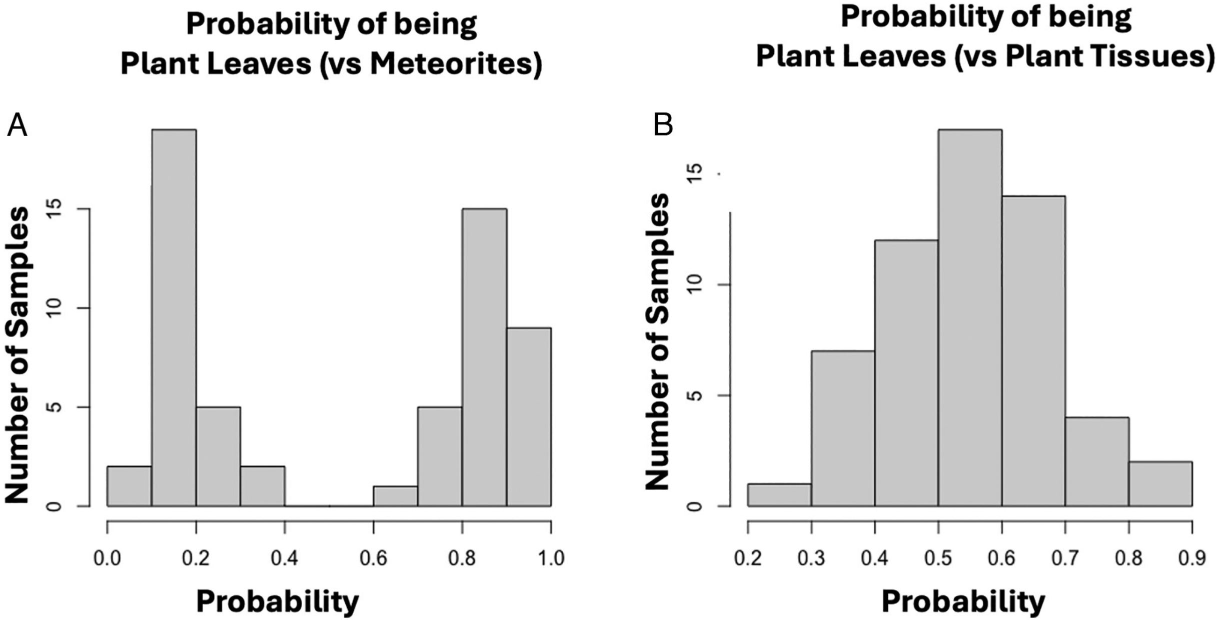

Fig. 2. Histograms reveal the probabilities that individual samples in a training set (a randomly selected 75% of all samples) fall in one of two categories. The *x* axis indicates the class probability that a sample lies within one of two groups in bins of width 0.1, while the *y* axis records the number of samples in each bin. (*A*) The training set includes 28 (of 37) specimens of green plant leaves vs. 31 (of 42) carbonaceous meteorites. In this example, with two contrasting suites of organic molecules, all (100%) of the samples are correctly assigned as plants or meteorites with reliable class probabilities >0.60 or <0.40. Furthermore, all but 3 samples have class probabilities >0.70 or <0.30. (*B*) The training set includes 27 (of 36) specimens of plant leaves vs. 30 (of 40) specimens of non-photosynthetic plant tissues—similar molecular suites that are predicted correctly in only 79% of samples. Furthermore, only 14 of 57 training set samples are assigned to one of the two groups with a reliable class probability >0.60 or <0.40.

groups of biogenic and abiogenic samples (e.g., Groups 1, 2, and 3 vs. Groups 8 and 9), and greater than 95% differences between most groups of photosynthetic vs. non-photosynthetic specimens (e.g., Groups 2, 3, 4, and 5 vs. Groups 1, 8, and 9).

Note that one advantage of the machine-learning approaches employed in this study is that the risk of false group classification resulting from widespread molecular contamination (for example by plasticizers or fingerprints) is minimal, provided that such contamination is not restricted to one group of samples. Similarly, our methods might be affected by systematic processing procedures applied to only one group, such as sample digestion protocols used on only one carbon-rich lithology. Therefore, we have applied the same procedures for all digested samples. Additionally, to address the effect of systematic contamination on our machine-learning approach, we are currently conducting a follow-up study in which controlled amounts of contaminants are introduced across multiple training samples. This study will allow us to assess the robustness of our models to such interference. Future work will also explore alternative data processing strategies and model architectures aimed at mitigating or eliminating the effects of systematic contamination.

In the following two sections, we combine groups in Table 1 to examine the discrimination between biogenic vs. abiogenic and photosynthetic vs. non-photosynthetic samples.

Abiogenic/Biogenic Discrimination

Recognizing biogenic vs. abiogenic samples is a key objective in both paleobiology and astrobiology research (12, 15, 65–70). In the case of organic matter preserved in ancient sedimentary rocks such as shales and cherts, controversies may arise regarding biogenicity when no definitive morphological evidence of microfossils (e.g., cellular fossils or mat textures) is preserved (56, 71–73), particularly in Archean (>2.5 Ga) samples. While claims of

Table 1. Pairwise comparisons of discriminations among nine groups of samples

Group number[*]	2	3	4	5	6	7	8	9	Number[†]
1—Modern animals	91[‡]	95	100	94	F[§]	F	98	98	21
2—Modern plants (non-PS)	.	79	96	94	92	F	97	95	40
3—Modern plants (PS)	.	.	100	97	F	F	100	98	36
4—Fossil microbes (PS)	.	.	.	96	F	97	92	89	24
5—Fossil wood, coal, and oil shale	.	.	.	.	F	98	97	97	49
6—Fossil animals	.	.	.	.	.	94	F	F	9
7—Modern fungi	.	.	.	.	.	.	98	97	16
8—Carbonaceous meteorites	.	.	.	.	.	.	.	98	42
9—Synthetic samples	.	.	.	.	.	.	.	.	35

[*] See text for full description of groups.
[†] Number of samples in that group.
[‡] Percentage of correctly classified samples.
[§] "F" indicates a failed classification because >50% of the samples in the smaller group are incorrectly assigned to the larger group.

abiogenic origins for most Phanerozoic hydrocarbons (74, 75) have been largely dismissed (76, 77), experimental and geochemical evidence unequivocally points to the importance of natural abiogenic organic synthesis in some deep hydrothermal environments (77–84). Distinguishing biogenic from abiogenic suites of organics therefore remains an important challenge in understanding Earth's oldest sediment-bound organic matter.

Combining two or more of the sample groups allows us to examine the ability to discriminate major classes of samples. Accordingly, we compared random forest models for three contrasting groupings to test our ability to identify unknown biogenic vs. abiogenic samples.

Model #1. In Model #1, we tested modern living animals and plants in Groups 1, 2, and 3 (97 samples) vs. abiogenic Groups 8 and 9 (77 samples). Of the 174 samples in the combined training (130 samples) and testing sets (44 samples), 171 (98%) were correctly categorized (*SI Appendix*, Table S2), with a strongly bimodal distribution of probabilities (Fig. 3*A*), reflecting the distinct character of organic suites from modern organisms vs. those from meteorites or synthetic mixtures. The three misclassified samples include biogenic *Mimosa hostilis* root and Lawson cypress (*Chamaecyparis lawsoniana*) sap (biogenic class probabilities 0.35 and 0.40, respectively) and one set of abiogenic Maillard reaction products with an indeterminate biogenic class probability of 0.56. Therefore, among the 174 training and testing set samples employed in Model #1, we find two biogenic false negatives and no biogenic false positives.

In order to assess the performance of our models, we used the receiver operating characteristic (ROC) curves and also calculated the associated area under curve (AUC) (85, 86). ROC curves connect coordinate points with [(1 – specificity) = false positive rate] as the *x*-axis and sensitivity as the *y*-axis at all cut-off values measured from the test results (87). The area under the ROC curve (AUC) is used as a measure of accuracy in binary classifiers. The closer the ROC curve is to the upper left corner of the graph, the higher the accuracy of the test because in the upper left corner, the sensitivity = 1 and the false positive rate = 0 (specificity = 1). The ideal ROC curve thus has an AUC = 1.0 (87). We plotted the ROC curves (*SI Appendix*, Fig. S1*A*) and the calculated associated AUC for the training and testing sets (AUC = 0.977 and 1.000, respectively).

To assess the generalizability of our model and to showcase a more accurate representation of our model's errors, we additionally ran a 10-fold repeated cross validation (CV) on all 174 samples in this dataset. This step involved repeating K-fold CV multiple times and reporting the mean results from all the folds and all runs. After repeated CV, Model #1 shows an accuracy of 98.3%, which aligns well with our training-test split runs. Detailed results of our repeated CV can be found in *SI Appendix*, Table S6.

Despite this accuracy, Model #1 was unsuccessful at correctly predicting the biogenicity of many ancient organic remains with confirmed biogenic origins (i.e., samples from Groups 4 and 5), because the random forest methodology is not able to correctly characterize types of samples that are not included in the training set. For example, 34 of 49 Group 5 coal and fossil wood specimens in our inventory were incorrectly classified as abiogenic. Consequently, Model #1 is not suitable for determining the biogenicity of fossil organics.

Model #2. In a second model, we sought to discriminate between ancient biogenic samples and organic-rich abiotic samples. Model #2 employed a biogenic training/test set with 87 organic-rich Phanerozoic, Proterozoic, and Archean sedimentary rocks, including 25 coal and fossil wood samples in Group 4 plus 62 Group 5 shales and cherts identified as biogenic based on their geological settings, significant organic content, and diagnostic carbon isotopic values. These 87 ancient biogenic samples were compared to 77 abiogenic samples from Groups 8 (42 meteorites) and 9 (35 synthetics; *SI Appendix*, Table S4). Of the 87 biogenic paleo-organic samples, 83 (95%) were correctly classified. Furthermore, 70 of these samples (80%) had high-confidence biogenic class probabilities

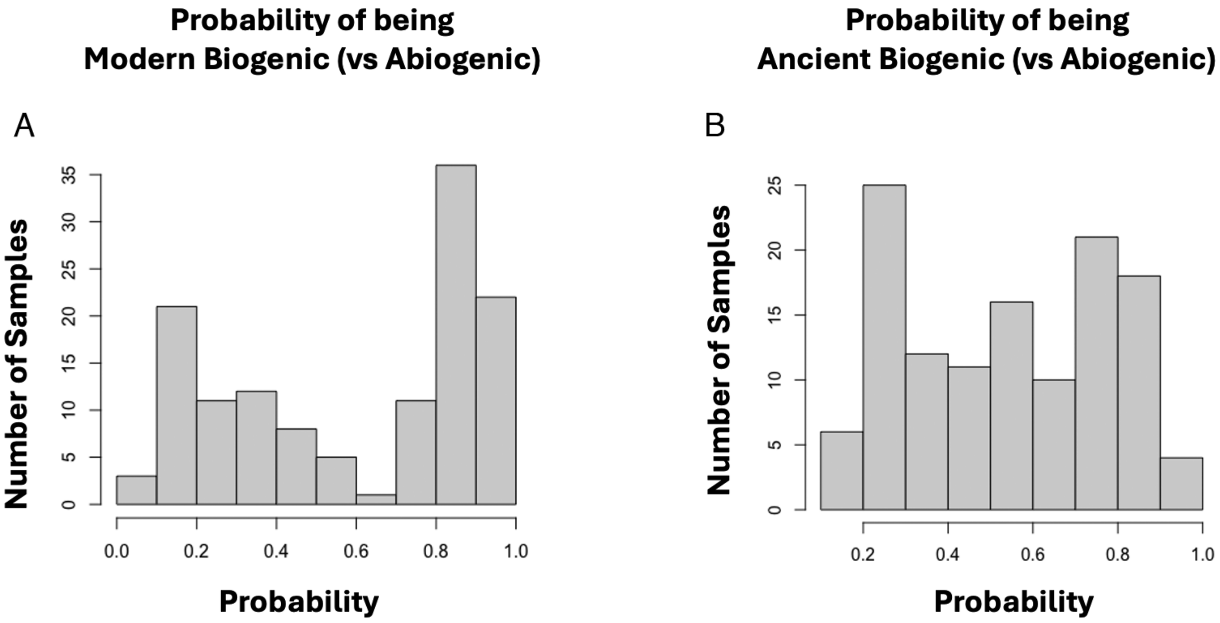

Fig. 3. Histograms of the probabilities that individual samples in a training set (a randomly selected 75% of all samples) is biogenic vs. abiogenic. The *x* axis indicates the class probability that a sample lies within one of two groups in bins of width 0.1, while the *y* axis records the number of samples in each bin. (*A*) Model #1. The training set includes 130 samples divided between modern animals and plants vs. abiogenic meteorites and laboratory synthetic mixtures. All but 7 samples have reliable biogenic class probabilities >0.60 or <0.40, as reflected in the bimodal distribution of class probabilities. (*B*) Model #2. The training set includes 123 samples divided between ancient molecular suites of known biogenic origins vs. abiogenic meteorites and laboratory synthetic mixtures. While 25 samples have indeterminate biogenic class probabilities between 0.6 and 0.4, 98 samples have reliable class probabilities >0.6 or <0.4.

>0.60. ROC curves (*SI Appendix*, Fig. S1*B*) yield AUC = 0.924 and 0.926 for the training and testing sets, respectively. After repeated CV, Model #2 shows an accuracy of 92.7%, which aligns well with our training-test split runs. Detailed results of our repeated CV can be found in the *SI Appendix*, Table S6.

Of 77 abiogenic samples, 69 (90%) in the training and testing sets were correctly classified, including 39 of 42 meteorites and 30 of 35 synthetic mixtures. Only 3 of the 12 misclassified samples had an incorrect class probability >0.57, including 1 sample (of 6) of formose reaction products, which was incorrectly classified as biogenic (biogenic class probability 0.67) and is the only biogenic false positive in Model #2. In addition, the 1.5 Ga Yusmastakh Formation of Northern Siberia and the 0.64 Ga Nantuo Formation of Hubei Province, China, both of presumed biogenic origins, were incorrectly classified as abiogenic (abiogenic class probabilities 0.68 and 0.67, respectively). A histogram of class probabilities (Fig. 3*B*) displays a modest bimodal distribution, albeit with significantly more samples in the indeterminate 0.40 to 0.60 range than in Fig. 2*A*.

Applying Model #2 to 109 ancient organic-rich sedimentary rocks of uncertain biogenicity (*SI Appendix*, Table S3), we find 68 samples (61%) with biogenic class probabilities >0.50, of which 32 samples have biogenic class probabilities >0.60. The latter examples include 2 samples from the Paleoproterozoic (2.3 Ga) Gowganda Formation, Cobalt Basin, Ontario, Canada (biogenic class probabilities 0.65 and 0.73); 4 samples (of 5) from the Neoarchean (2.52 Ga) Gamohaan Formation, Kaapvaal Craton, South Africa (0.65 to 0.69); 1 sample (of 2) from the Neoarchean (2.66 Ga) Jerrinah Formation, Western Australia (0.62); 2 samples (of 3) from the Paleoarchean (3.33 Ga) Josefsdal Chert (0.65 and 0.68); and 1 sample (of 2) from the Paleoarchean (3.51 Ga) Singhbhum Craton, India (0.61).

A subset of 186 ancient organic-rich sedimentary rocks (*SI Appendix*, Table S3)—those for which reliable ages are available—display systematic trends vs. age distribution (Fig. 4). Of 82 Phanerozoic specimens, 76 (93%) fall in the biogenic range. By contrast, 43 of 59 (73%) of Proterozoic samples and 21 of 45 (47%) of Archean samples are labeled as biogenic with a similar degree of confidence. The significant decrease in percentage of labeled biogenic samples with increasing age may reflect the progressive degradation of biomolecules in older samples that have been subjected to extensive diagenesis (i.e., physical and chemical changes during lithification), metamorphism (alteration by temperature and pressure), and/or metasomatism (chemical alteration by hydrothermal fluids). At the same time, we cannot rule out the possibility that the oldest samples in our survey incorporated a higher percentage of abiogenic organic molecules from meteorites or other sources (88).

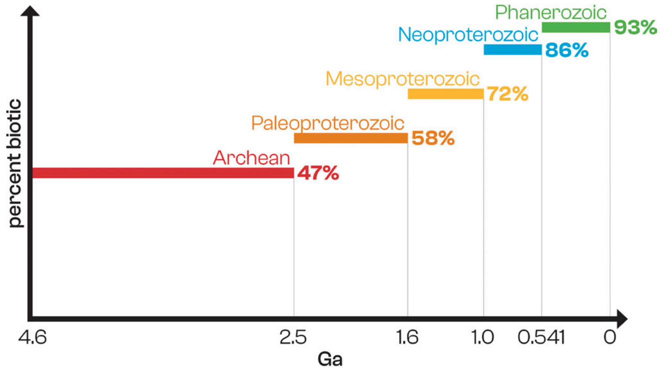

Fig. 4. The percentage of biogenic (vs. abiogenic) samples classified by random forest Model #2 increases significantly through time for rocks from Archean to Proterozoic to Phanerozoic eons.

Model #3. A third test of biogenicity employed a biogenic training/test set with 89 organic-rich Phanerozoic, Proterozoic, and Archean sedimentary rocks, including 25 samples in Group 4 plus 64 shale and chert samples identified as biogenic based on their geological settings, significant organic content, and diagnostic carbon isotopic values. These 89 ancient biogenic samples were compared to 77 abiogenic samples from Groups 8 (42 meteorites) and 9 (35 synthetics; *SI Appendix*, Table S4). Model #3 differs from Model #2 in that it does not include Phanerozoic fossil wood, coal, and other hydrocarbon-rich materials. All 89 biogenic fossil samples were correctly classified. Furthermore, 70 of these samples (80%) had high-confidence biogenic class probabilities >0.60.

However, 18 of 77 abiogenic samples (23%) in the training and testing sets, including 8 of 42 meteorites and 10 of 35 synthetic mixtures, were incorrectly classified as biogenic. Five of these misclassified samples have an incorrect biogenic class probability >0.60, including synthetic Kraton (0.61) and synthetic polystyrene (0.61), both of which are employed as GC–MS standards. Also misidentified are three shock-heated meteorites: the Sutters Mill CM2 carbonaceous chondrite (0.63), the Y 86720 C2 carbonaceous chondrite (0.61), and the Tissint Martian shergottite (0.62) (89–91). These five false positive results point to the need for additional samples, especially a more diverse suite of abiogenic samples in future studies. Note that ROC curves (*SI Appendix*, Fig. S1*C*) yield AUC = 0.873 and 0.863 for the training and testing sets, respectively. In addition to assessing the generalizability of our model we ran a 10-fold repeated CV on all 166 relevant samples to showcase a more accurate representation of our model's errors. After repeated CV, Model #3 shows an accuracy of 91.6%, which shows a slightly increased accuracy when compared with our training-test split runs (*SI Appendix*, Table S6).

We applied Model #3 to ancient organic molecular suites from 42 sedimentary rocks, including 9 Phanerozoic, 9 Proterozoic, and 24 Archean samples of unknown biogenicity (*SI Appendix*, Table S4). Model #3 classified all 42 unknown samples as biogenic, with 30 samples (71%) having ≥0.60 biogenic class probability and 23 with ≥0.65 biogenic class probability (55%). Of note, several Eoarchean and Paleoarchean samples have biogenic class probabilities >0.65, including 2 (of 3) samples of 3.8 Ga metasediments from the Isua Greenstone Belt, Greenland [(92, 93); biogenic class probabilities 0.67 and 0.68]; 3 samples from the 3.5 Ga Apex Chert and Dresser Formation, Pilbara Craton, Western Australia [(23, 27, 71, 94–97); 0.65, 0.69, and 0.75]; and 1 sample from the 3.5 Ga Theespruit Formation, Barberton Greenstone Belt, South Africa [(98); 0.67]. Thus, consistent with morphological and/or isotopic indicators of biogenicity, our analyses suggest that the organic matter in these samples is biogenic in origin.

Models #2 and #3 Combined. Together, Models #2 and #3 described above suggest a conservative strategy for assessing biomolecular signatures among significantly degraded suites of organic molecules in Proterozoic and Archean chert and shale. At this stage of our investigations, the biogenicity of many samples remains indeterminate by one or more of the approaches described above. However, we suggest that any unknown ancient organic-rich sample that scores ≥0.60 biogenic class probability in both Model #2 and Model #3 above is likely to be biogenic. Note that with these criteria none of the abiogenic samples in Groups 8 or 9 yield false positives. Applying this ≥0.60 criterion to the unknown ancient samples from our inventory recorded in *SI Appendix*, Tables S3 and S4, we assign a biogenic origin to the organic molecular suites in 11 samples of ancient sedimentary rocks (Table 2).

Table 2. **Predicted biogenic samples based on ≥60% biogenic class probabilities from both Models #2 and #3 above**

Sample #	Sample Description	Age (Ma)	Probability Model #2	Probability Model #3
RMH189	Ramsay Crossing	44	0.809	0.638
RMH221	South Oman Salt Basin, U-shale	542	0.731	0.759
RMH233	South Oman Salt basin, Sharum Fm.	560	0.757	0.702
RMH273	HCs on native Cu*	~1,000	0.700	0.668
RMH274	HCs on native Cu*	~1,000	0.731	0.642
RMH239	Velkerri Fm., Roper Gp, Australia	1,400	0.723	0.801
RMH282	Gowganda Fm.	2,300	0.726	0.830
RMH372	Gamohaan Fm.	2,521	0.711	0.804
RMH243	Jerrinah Fm., Fortescue Group	2,660	0.617	0.755
RMH178	Josefsdal Chert	3,330	0.682	0.767
RMH179	Josefsdal Chert	3,330	0.648	0.735

*Hydrocarbon deposits on native copper (99).

Our results thus support the biogenic origins of organic molecular suites from the Paleoproterozoic (2.30 Ga) Gowganda Formation of Ontario, Canada (100–102); the Neoarchean (2.52 Ga) Gamohaan Formation, Kaapvaal Craton, South Africa (103, 104); the Neoarchean (2.66 Ga) Jerrinah Formation of the Fortescue Group, Western Australia (105, 106); and the Paleoarchean (3.33 Ga) Josefsdal Chert of the Barberton Greenstone Belt, South Africa (53). The latter example is of special interest because extraterrestrial organics have been described as occurring in a 3.33 Ga sediment layer from the Josefsdal Chert (88). Based on the use of meteorite organics in our training and testing sets, we are confident that the organics in this sample are not predominantly of meteoritic origins. This ability to discriminate between biogenic and meteoritic organics has special relevance to the analysis of astrobiological samples, for example sediments analyzed on or returned from Mars.

Photosynthetic vs. Non-Photosynthetic Organisms

Photosynthesis has been a significant metabolic strategy for most of Earth's history, as revealed by geological and geochemical evidence for the Neoarchean and Paleoproterozoic rise of atmospheric and oceanic oxygen (107, 108), the preservation of cyanobacteria and algal fossils in many Proterozoic cherts and shales (28, 109–111), and bacterial genomic trees (112). A major effort in molecular paleontology has been to detect evidence of cellular metabolic strategies, such as photosynthesis, via the survival of diagnostic molecules and molecular structures, including porphyrins (113–115), thylakoids (116), steranes (117), and constituents of leaf cuticles (118, 119). However, no molecular study has provided unambiguous biomolecular evidence for phototrophy earlier than 1.75 Ga, although a photosynthetic origin for more ancient organic matter is routinely accepted based on morphological and isotopic evidence, as well as molecular clock analyses that suggest an Archean origin for oxygenic photosynthesis (120). Furthermore, some experts argue that photosynthesis is the only plausible origin for organic molecules in any sedimentary rock with a total organic carbon content greater than a few weight percent (121).

Despite the complexities of the evolution of photosynthesis and the occurrence of multiple extant phototrophic biochemical pathways (122–125), the detection of photosynthesis and other metabolic strategies is a promising target for machine-learning applications. We echo previous studies on Phanerozoic fauna (12, 32, 34), albeit with much more ancient samples, that patterns of molecular fragmentation may preserve signals of metabolic processes long after specific diagnostic molecules have degraded. Accordingly, in random forest Model #4 we considered the comparison of Groups 1, 6, 8, 9, and 11 (107 samples of modern and fossil non-photosynthetic organisms, meteorites, and synthetic mixtures) vs. Groups 2, 3, 4, 5, and 10 (151 samples of modern and fossil photosynthetic organisms). Of the 259 training and test set samples, 242 were correctly classified (93% correct; *SI Appendix*, Table S5). ROC curves (*SI Appendix*, Fig. S1*D*) yield AUC = 0.922 and 0.924 for the training and testing sets, respectively. When we ran a 10-fold repeated CV on all 259 relevant samples, Model #4 showed an accuracy of 92.3%, which aligns well with our training-test split runs (*SI Appendix*, Table S6).

The confidence with which these samples were assignable as photosynthetic vs. non-photosynthetic differs by group. All 42 meteorite samples (Group 8) were correctly assigned, with an average photosynthetic class probability of 0.28, while all but 1 of 35 synthetic samples (Group 9) were correctly assigned with an average photosynthetic class probability of 0.41. All but 2 of 48 Group 5 fossil coal, wood, and other Phanerozoic hydrocarbon samples in the training set were correctly assigned as photosynthetic, with a high average photosynthetic class probability of 0.72. The only misclassified samples were coal slag (photosynthetic class probability 0.46) that had been heated to temperatures >500 °C and charcoal (0.46) heated to >400 °C; therefore, any molecular biosignatures in these 2 samples were highly degraded.

All 76 modern plant specimens (Groups 2 and 3) were correctly classified as photosynthetic with average class probability 0.84. Of 21 modern animals, 19 were correctly assigned as non-photosynthetic with an average non-photosynthetic class probability of 0.69. By contrast, 6 of 8 fossil animal samples (Group 6) were incorrectly assigned, though with an indeterminate average photosynthetic class probability of 0.51.

With only two exceptions, no incorrectly assigned sample in the training set had a class probability >0.60 or <0.40. The first exception is Mesoproterozoic (1.5 Ga) chert from the Yusmastakh Formation, Billyakh Group, northern Siberia, which is thought to incorporate organics from photosynthetic microorganisms (126), yet which has a photosynthetic class probability of only 0.34. Interestingly, this sample also had a biogenic class probability of 0.32 using Model #3, suggesting that any photosynthetic signature has been erased by diagenesis. Otherwise, this sample represents the only false negative in Model #4.

The second exception was the hard exterior tissues of an assumed non-photosynthetic sea squirt (paraphyletic class *Ascidiacea*), with a photosynthetic class probability 0.71. By contrast, the soft interior tissues had photosynthetic class probability 0.33—a typical

Table 3. Predicted photosynthetic/non-photosynthetic samples based on photosynthetic class probabilities

Sample #	Sample Description	Age (Ma)	Photosynthetic Class Probability
Predicted photosynthetic			
RMH233	South Oman Salt Basin	560	0.688
RMH079	Rysso Fm., Svalbard	750	0.626
RMH091	Svanbergfjellet Fm., Spitsbergen	810	0.607
RMH242	Velkerri Fm., Australia	1,400	0.723
RMH093	Irregully Fm., Australia	1,500	0.619
CW045	Shungite, Karelia, Russia	2,000	0.654
RMH282	Gowganda Group, Ontario, Canada	2,300	0.644
RMH372	Gamohaan Fm., South Africa	2,520	0.579
Predicted non-photosynthetic			
RMH193	Kockatea Shale, Australia	254	0.305
RMH244	Duck Creek Dolomite, Australia	1,800	0.290
RMH095	Michigamme Fm., Michigan	1,850	0.381
RMH200	Mount McRae Shale, Australia	2,500	0.294
RMH208	Nauga Fm., Kaapvaal Craton, South Africa	2,501	0.298
RMH211	Kamden Mbr, Kaapvaal Craton, South Africa	2,504	0.304
RMH180	Dresser Fm., Pilbara Craton, Australia	3,480	0.373
RMH248	Theespruit Fm., South Africa	3,500	0.389

value for an animal in Model #4. However, we have subsequently learned that sessile sea squirts are often coated with photosynthetic algae (127). Therefore, invoking the criteria that a confident prediction of a sample's photosynthetic origins requires a photosynthetic class probability >0.60, we conclude that there are no false positives in this model's training set.

When we apply Model #4 to 131 Precambrian organic-rich sedimentary rocks of unknown affinities, we find that most samples have intermediate photosynthetic class probabilities <0.60 and >0.40 (*SI Appendix*, Table S5)—values that do not allow a confident assessment of phototrophy. Nevertheless, several samples display reliable evidence for photosynthesis with class probabilities >0.60 (Table 3). Among these specimens is the Paleoproterozoic (2.30 Ga) Gowganda Group of Ontario, Canada (photosynthetic class probability 0.644).

In addition, while they do not meet the 0.6 threshold, all five specimens from different horizons of the Neoarchean (2.52 Ga) Gamohaan Formation, Campbellrand Group, South Africa have photosynthetic class probabilities between 0.54 and 0.58, which collectively are suggestive of a photosynthetic classification. This formation is of special interest because its unique, three-dimensionally preserved carbonate microbialite fossils present compelling morphological evidence for Neoarchean photosynthesis (128, 129). Our machine-learning analyses thus complement these morphological observations by demonstrating that molecular evidence for photosynthesis has been preserved in the distribution of this sample's diagenetically altered molecular fragments, even though no diagnostic biomolecules remain intact.

On the other hand, several Proterozoic and Archean samples, including the 1.8 Ga Duck Creek Dolomite (130, 131), the 2.5 Ga Mount McRae Shale (132), several other units from the 2.5 Ga Kaapvaal Craton (133), the 3.48 Ga Dresser Formation of the Pilbara Craton (23, 95), and the 3.5 Ga Theespruit Formation (98), lie in the likely non-photosynthetic range, all with photosynthetic class probabilities <0.40 (Table 3). As noted above, molecular analyses of such samples in some instances may primarily reflect significant diagenesis of biomolecules rather than a non-photosynthetic origin (130, 131).

Discussion

In this study, we used py–GC–MS analyses of 406 varied organic-rich samples combined with supervised machine learning to determine phylogenetic and physiological attributes of ancient life based on the distribution of diagenetically altered molecular fragments. Rather than search for specific biomolecules, we explored the prospects of supervised machine learning to recognize patterns among molecular distributions—data that reveal both the intensities of numerous peaks and regions of no intensity. Key findings include the earliest biomolecular evidence for:

- The photosynthetic origins of organic molecules in the 2.52 Ga Gamohaan Formation, Campbellrand Group, South Africa, and the 2.30 Ga Gowganda Group, Ontario, Canada;
- The biogenicity of organic molecules preserved in the ~3.51 Ga Singhbhum Craton, India; the 3.33 Ga Josefsdal Chert of the Barberton Greenstone Belt, South Africa; and the 2.66 Ga Jerrinah Formation, Fortescue Group, Pilbara Craton, Australia;
- The apparently non-photosynthetic origin of organic species in the 3.5 Ga Theespruit Formation, Barberton Greenstone Belt, South Africa, and the 3.48 Ga Dresser Formation, Pilbara Craton, Australia (with the caveat that diagenesis and/or metamorphism may have degraded any molecular evidence of photosynthesis).

The most significant outcome of this research is demonstration of the potential for supervised machine-learning methods to discover biochemical information from highly degraded molecular suites as old as the Paleoarchean Eon—samples in which no intact biomolecules have been preserved. Two opportunities hold the promise for significant advances in this effort.

First, substantially more samples need to be evaluated, especially more proportionately balanced numbers of samples in varied groups for training and testing. Furthermore, with the exceptions of carbonaceous meteorites, our sample inventory is notably lacking in ancient abiogenic samples. While it is difficult to be confident of the abiogenicity of organic matter in any Proterozoic or

Archean sample, we can synthesize a range of Fischer–Tropsch products that might simulate ancient abiogenic organosynthesis (134, 135).

Additionally, our photosynthetic vs. non-photosynthetic discrimination was performed using a machine-learning classifier trained dominantly on oxygenic photosynthetic organisms (e.g., plants, cyanobacteria, and algae). The inclusion of anoxygenic photosynthesizers in our training set might allow for discrimination among different kinds of photosynthesis as well as the identification of anoxygenic photosynthesis in paleobiological samples. Critically, the binning of organisms based on Linnean ranks and other arbitrary features may dilute the true contribution of a phylogenetic signal to our analytical data. An important opportunity is thus the determination of phylogenetic information from fossil biomolecules (8, 12). Advances in the understanding of Precambrian eukaryotes (136–140), terrestrial flora (141–144), and vertebrates (11, 29, 32, 145), demonstrate the potential of machine-learning applications.

Furthermore, we have yet to fully exploit the opportunities provided by machine learning. In future work, our py–GC–MS analyses will be combined with additional attributes of organic molecular assemblages, including C-H-N compositions; stable isotope ratios of C, H, N, O, and S; morphological data on well-preserved fossils; Raman spectra; Fourier Transform IR spectra; and other information. Looking ahead, the use of isotopologues (146), especially position-specific isotopologues (147, 148), would provide an added dimension to such studies of paleotaxonomy.

We conclude that information-rich attributes of ancient organic matter, even though highly degraded and with few if any surviving biomolecules, have much to reveal about the nature and evolution of life.

Methods

Sample Characterization. We collected and curated 406 carbon-bearing samples from varied sources. A list of these samples and their attributes (e.g., biogenic vs. abiogenic, photosynthetic vs. non-photosynthetic, age, etc.) is provided in *SI Appendix*, Table S1.

Analytical Methods. py–GC–MS analyses were performed with a CDS 6150 pyroprobe (CDS Analytical, Inc., Oxford, PA) interfaced with an Agilent 8860 series gas chromatograph interfaced with an Agilent 5999 quadrupole mass spectrometer. An Agilent 30 M 5 % phenyl PDMS column was used for chromatographic separation. The GC oven temperature was programmed to hold at 50 °C for 1 min, then increase from 50 °C to 300 °C at a rate of 5 °C min^{-1}, and then remain at 300 °C for 15 min. Helium (UHP 5.5 grade) was used as the carrier gas, operating in constant flow mode. Samples (each 10 to 100 μg) were loaded into preashed (combusted under air at 550 °C for 3 h) quartz tubes, which were then inserted into the coil of the pyroprobe and flash pyrolyzed (ramp rate 500 °C s^{-1}) to 610 °C and held for 10 s. The pyrolysates were immediately swept onto the GC column by the He gas and analyzed. The source was operated in electron ionization (EI) mode with 70 eV ionization energy at 250 °C. The mass selective detector scan rate was 0.80 s/decade over a range of m/z 45 to 700, with an interscan delay of 0.20 s.

MS data were not collected for the first 2 min after injection to avoid overloading the detector with small volatiles such as CO_2 and H_2O. In addition, since many of the samples were curated independently and displayed signals from C16 (palmitic) and C18 (stearic) fatty acids–common components of fingerprints and "slip agents" added to plastic sample bags to keep them from sticking together–we excluded the regions of the chromatograms after the region where these common contaminants elute. Note, however, that there is still a contribution to the examined chromatographic complexity from derivatives of such compounds (for example straight and branched long-chain alkanes, alkenes, and rearranged aldehyde and ketone derivatives of these species) that contribute to the precutoff region's molecular complexity. We found there was little signal beyond *m/z* 200, thus the region considered in the computational methods was limited from 2 to 35 min and *m/z* 45 to 200. Each sample was reduced to a two-dimensional matrix with 489,240 elements representing signal intensities as a function of mass and retention time, though many of the intensity values were zero (Fig. 1).

As in ref. 15, the data output are not immediately compared to existing libraries to identify specific molecular compounds. Instead, the full three-dimensional output files are preprocessed and used in our machine-learning analysis. In other words, in our methodology, precise compound identification is useful but not necessary: rather, relational aspects of chromatographic and mass peaks are of interest. This approach underscores this method's utility in identifying high-dimensional and potentially unexpected patterns that are most indicative of a certain kind of sample, rather than relying on precise molecular fragments that may not be the most advantageous for making such a discrimination. In addition to identifying the provenance of ancient samples on Earth, data-driven discovery represents a promising pathway for detecting alien biology (15).

Machine-Learning Analysis. We trained machine-learning models using three-dimensional data that included chromatographic retention time, mass-to-charge ratio, and intensity values for each sample. The data generated and analyzed in this manuscript can be found on the Open Science Framework repository titled "Organic geochemical evidence for life in Archean rocks identified by py–GC–MS and supervised machine learning" (https://doi.org/10.17605/OSF.IO/G93CS). The code for the paper can be found at https://github.com/PrabhuLab/PyGCMS-Biosign-ML. All data, code, and materials used in the analysis are available to any researcher for purposes of reproducing or extending the analysis. Licenses for the data and code usage and relevant attribution information will be updated on the respective repositories.

Peak intensity was normalized on a scale from 0 to 1, while time and mass/charge were analyzed as collected in 3,240 scan time steps (with each scan representing 3 s) and one m/z increments over the range from m/z 50 to 200, respectively. Our machine-learning approach allows us to identify diagnostic sets of feature sets of retention time and m/z values that distinguish between our assigned classes in a model. The K-fold cross-validation method has applied overcome the concern of overfitting to the training data.

Preprocessing. Each of the 406 analyzed samples was represented as a two-dimensional matrix, where the rows and columns represent the scan numbers and m/z ratios, respectively, and the entries are the corresponding intensities. For each sample and m/z value, we performed the preprocessing steps in the chromatographic direction, stabilized the variance of the intensity values by taking the square root, smoothed the values by taking the moving average of its current and its immediate five nearest observations on each side, and subtracted the baseline, where the baseline estimation was based on the Statistics-sensitive Non-linear Iterative Peak-clipping algorithm (SNIP) (149), using the R-library, MALDIquant (150).

Intensity values were normalized via min–max normalization (151), followed by peak detection in the chromatographic direction for each *m/z* ratio. Peaks were detected as local maxima above 4 × signal-to-noise ratio, where the noise was estimated by calculating the median absolute deviation using the R-library, MALDIquant (150). After eliminating near-zero variance and strongly correlated features using the R-library caret (152), the data were reduced to 8,149 features, which are the detected combination of scan number and *m/z* values. These 406 py–GC–MS data files are available at (https://doi.org/10.17605/OSF.IO/G93CS).

Model Choice. We picked the random forest method because it is highly accurate, computational lightweight, and interpretable. The random forest method is an ensemble classification method that constructs a collection of decorrelated decision trees (38). The exploration of better classifiers and their benefits with respect to tradeoffs between accuracy, applicability, and interpretability will be one of the next steps in our paleobiological study. We used the random forest model from the R-library RandomForest (153).

Model Validation. We used two validation strategies for the trained machine-learning model. First, we used a 75/25 training-test split using stratified random sampling, where the model was trained on 75% of the samples and the accuracy assessed on the remaining 25% of the samples. Additionally, we also tested the accuracy of our four models using three repeats on a 10-fold cross-validation. Repeated k-fold cross-validation provides a way to improve the estimated performance of a machine-learning model. This step involved repeating K-fold CV multiple times and reporting the mean results from all the folds and all runs. Such a result is expected to be a more accurate representation of a model's error. The results from the 10-fold repeated CV are presented in *SI Appendix*, Table S6.

Data, Materials, and Software Availability. The data generated and analyzed in this manuscript can be found on the Open Science Framework repository titled "Organic geochemical evidence for life in Archean rocks identified by py–GC–MS and supervised machine learning" (https://doi.org/10.17605/OSF.IO/G93CS). The code for the paper can be found at https://github.com/PrabhuLab/PyGCMS-Biosign-ML. All data, code, and materials used in the analysis are available to any researcher for purposes of reproducing or extending the analysis. Licenses for the data and code usage and relevant attribution information will be updated on the respective repositories.

ACKNOWLEDGMENTS. We thank our three anonymous peer reviewers for providing insightful comments and suggestions that strengthened our manuscript. We also thank Terry Collins, Woody Fisher, Steve Shirey, and Michael Walter for valuable discussions. We are grateful to John Saul and Jake Skabelund, for the donation of samples employed in this study. Dr. Michelle Jordan provided invaluable help during the acid dissolution of rock samples. Studies of molecular biosignatures have been supported by the John Templeton Foundation Grant 61783. Additional support has been provided by the NASA Astrobiology Institute ENIGMA Team Grant 80NSSC18M0093, a private foundation, and the Carnegie Institution for Science. A.C. and A.D.C. are supported by NSF-Earth Sciences grant 2029521. M.L.W.'s research is funded by NASA through the NASA Hubble Fellowship Program Grant HST-HF2-51521.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. Any opinions, findings, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the NASA.

Author affiliations: [a]Earth and Planets Laboratory, Carnegie Institution for Science, Washington, DC 20015; [b]Sagan Fellow, NASA Hubble Fellowship Program, Space Telescope Science Institute, Baltimore, MD 21218; [c]Department of Chemistry, Howard University, Washington, DC 20059; [d]Department of Mathematics and Statistics, Purdue University Northwest, Hammond, IN 46323; [e]Department of Earth Sciences, University of Graz, Graz 8010, Austria; [f]Geological Survey of Canada, Ottawa, ON K1A 0E8, Canada; [g]Department of Geological Sciences, Stanford University, Stanford, CA 94305; [h]Department of Geosciences, University of Cincinnati, Cincinnati, OH 45221; [i]State Key Laboratory of Critical Earth Material Cycling and Mineral Deposits, Nanjing University, Nanjing 210023, China; [j]Geology Department, Pomona College, Claremont, CA 91711; [k]Department of Earth Sciences, University of Toronto, Toronto, ON M5S 3B1, Canada; [l]A. E. Seaman Mineral Museum and Department of Physics, Michigan Technological University, Houghton, MI 49931; [m]Department of Geology, University of Liège, Liège 4000, Belgique; [n]Department of Geosciences, University of Oslo, Oslo 0316, Norway; [o]Evolutionary Studies Institute, University of the Witwatersrand, Wits 2050, South Africa; [p]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; [q]School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia; [r]Department of Earth and Environmental Sciences, Michigan State University, East Lansing, MI 48824; [s]Department of Ocean and Earth Sciences, Old Dominion University, Norfolk, VA 23529; [t]Department of Geological Sciences, University of Florida, Gainesville, FL 32611; [u]School of Earth and Environmental Sciences, University of St. Andrews, St. Andrews KY16 9TS, United Kingdom; [v]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [w]Centre de Biophysique Moléculaire, CNRS-UPR4301, Orléans 45071, France; [x]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD 21218; [y]Center for Functional Anatomy and Evolution, Johns Hopkins School of Medicine, Baltimore, MD 21218; and [z]Department of Geosciences, Virginia Tech, Blacksburg, VA 24060

Author contributions: M.L.W., A.P., H.J.C., and R.M.H. designed research; M.L.W., A.P., H.J.C., G.D.C., W.B., C.K.B., A.C., S.D., D.D.G., J.A.J., A.H.K., K.M.M., R.R., E.S., F.W., J.W., S.X., and R.M.H. performed research; M.L.W., A.P., C.O'D.A., H.J.C., G.D.C., M.B., W.B., A.C., A.D.C., R.R.G., E.J.J., J.J., M.V.K., N.N., E.E.S., R.E.S., and R.M.H. contributed new reagents/analytic tools; M.L.W., A.P., G.D.C., G.H., A.C., E.S., and R.M.H. analyzed data; M.L.W. data visualization; A.P. data visualization programming; C.O'D.A. contributed samples and edited manuscript; H.J.C. and G.D.C. collected samples and edited manuscript; G.H. and E.E.S. edited manuscript; M.B., W.B., C.K.B., A.C., A.D.C., S.D., R.R.G., D.D.G., J.A.J., E.J.J., J.J., A.H.K., M.V.K., K.M.M., N.N., R.R., E.E.S., R.E.S., F.W., J.W., and S.X. sample acquisition; R.M.H. collected samples and processed samples; and M.L.W., A.P., M.B., W.B., C.K.B., A.C., A.D.C., S.D., R.R.G., D.D.G., J.A.J., E.J.J., J.J., A.H.K., M.V.K., K.M.M., N.N., R.R., E.E.S., R.E.S., F.W., J.W., S.X., and R.M.H. wrote the paper.

1. M. T. Rosing, 13C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
2. A. C. Allwood, J. P. Grotzinger, A. H. Knoll, I. Kanik, Controls on development and diversity of Early Archean stromatolites. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9548–9555 (2009).
3. J. W. Schopf *et al.*, SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 53–58 (2018).
4. N. Noffke *et al.*, "Microbially induced sedimentary structures (MISS)" in *Treatise of Invertebrate Paleontology, vol. B: Prokaryota*, N. Noffke, Ed. (University of Kansas, 2022), 178p.
5. F. Westall, S. Xiao, Precambrian earth: Co-evolution of life and geodynamics. *Precambrian Res.* **414**, 107589 (2024).
6. J. J. Brocks *et al.*, Biomarker evidence for green and purple sulphur bacteria in a stratified Paleoproterozoic sea. *Nature* **437**, 866–870 (2005).
7. B. Rasmussen, I. R. Fletcher, J. J. Brocks, M. R. Kilburn, Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**, 1101–1104 (2008).
8. D. E. G. Briggs, R. E. Summons, Ancient biomolecules: Their origins, fossilization, and role in revealing the history of life. *BioEssays* **36**, 482–490 (2014).
9. K. L. French *et al.*, Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5915–5920 (2015).
10. J. J. Brocks *et al.*, Lost world of complex life and the late rise of the eukaryotic crown. *Nature* **618**, 767–773 (2023).
11. J. Wiemann *et al.*, Fossilization transforms vertebrate hard tissue proteins into N-heterocyclic polymers. *Nat. Commun.* **9**, 4741 (2018).
12. J. Wiemann, J. M. Crawford, D. E. G. Briggs, Phylogenetic and physiological signals in metazoan fossil biomolecules. *Sci. Adv.* **6**, eaba6883 (2020).
13. E. D. Dorn, G. D. McDonald, M. C. Storrie-Lombardi, K. H. Nealson, Principal component analysis and neural networks for detection of amino acid biosignatures. *Icarus* **166**, 403–409 (2003).
14. S. M. Marshall *et al.*, Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat. Commun.* **12**, 3033 (2021).
15. H. J. Cleaves II *et al.*, A robust, agnostic molecular biosignature based on machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2307149120 (2023).
16. G. Hystad *et al.*, Detecting biosignatures in complex molecular mixtures from pyrolysis-gas-chromatography-mass spectrometry data using machine learning. *J. Geophys. Res.: Mach. Learn. Comput.* **2**, e2024JH000441 (2025).
17. A. Kahana *et al.*, Constructing the molecular tree of life using assembly theory and mass spectrometry. arXiv [Preprint] (2025). https://arxiv.org/abs/2408.09305 (Accessed 27 October 2025).
18. D. R. Lawrence, Taphonomy and information losses in fossil communities. *Geol. Soc. Am. Bull.* **79**, 1315–1330 (1968).
19. A. K. Behrensmeyer, S. M. Kidwell, R. A. Gastaldo, "Taphonomy and paleobiology" in *Deep Time: Paleobiology's Perspective*, D. H. Erwin, S. L. Wing, Eds. (The Paleontological Society, 2000), pp. 103–147.
20. C. P. Marshall *et al.*, Structural characterization of kerogen in 3.4 Ga archaean cherts from the Pilbara Craton, Western Australia. *Precambrian Res.* **155**, 1–23 (2007).
21. M. H. Schweitzer, Soft tissue preservation in terrestrial Mesozoic vertebrates. *Annu. Rev. Earth Planet. Sci.* **39**, 187–216 (2011).
22. M. Bourbin, S. Derenne, F. Robert, Limits in pyrolysis–GC–MS analysis of kerogen isolated from Archean cherts. *Org. Geochem.* **52**, 32–34 (2012).
23. N. Noffke, D. Christian, D. Wacey, R. M. Hazen, Microbially induced sedimentary structures recording an ancient ecosystem in the ca. 3.48 billion-year-old Dresser Formation, Pilbara, Western Australia. *Astrobiol. J.* **13**, 1103–1124 (2013).
24. S. L. Potter-McIntyre *et al.*, Taphonomy of microbial biosignatures in spring deposits: A comparison of modern. Quaternary and Jurassic examples. *Astrobiology* **17**, 216–230 (2017).
25. J. P. Grotzinger, A. H. Knoll, Stromatolites in Precambrian carbonates: Evolutionary mileposts or environmental dipsticks? *Annu. Rev. Earth Planet. Sci.* **27**, 313–358 (1999).
26. R. S. Sansom, S. E. Gabbott, M. A. Purnell, Non-random decay of chordate characters causes bias in fossil interpretation. *Nature* **463**, 797–800 (2010).
27. T. Djokic *et al.*, Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits. *Nat. Commun.* **8**, 15263 (2017).
28. Z. Guo *et al.*, Cellular taphonomy of well-preserved Gaoyuzhuang microfossils: A window into the preservation of ancient cyanobacteria. *Precambrian Res.* **304**, 88–98 (2018).
29. E. M. Boatman *et al.*, Mechanisms of soft tissue and protein preservation in *Tyrannosaurus rex*. *Sci. Rep.* **9**, 15678 (2019).
30. V. E. McCoy *et al.*, Chemical signatures of soft tissues distinguish between vertebrates and invertebrates from the Carboniferous Mazon Creek Lagerstätte of Illinois. *Geobiology* **18**, 560–565 (2020).
31. C. Colleary *et al.*, Molecular preservation in mammoth bone and variation based on burial environment. *Sci. Rep.* **11**, 2662 (2021).
32. J. Wiemann *et al.*, Fossil biomolecules reveal an avian metabolism in the ancestral dinosaur. *Nature* **606**, 522–526 (2022).
33. A. Mojarro *et al.*, Comparative soft-tissue preservation in Holocene-age capelin concretions. *Geobiology* **20**, 377–398 (2022).
34. M. Tripp *et al.*, Fossil biomarkers and biosignatures preserved in coprolites reveal carnivorous diets in the carboniferous Mazon Creek ecosystem. *Biology* **11**, 1289 (2022).
35. J. Wiemann, P. R. Heck, Quantifying the impact of sample, instrument, and data processing on biological signatures in modern and fossil tissues detected with Raman spectroscopy. *J. Raman Spectrosc.* **55**, 761–773 (2024).
36. M. L. Wong *et al.*, On the roles of selection and function in evolving systems. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2310223120 (2023).
37. C. M. O'D. Alexander, M. L. Fogel, H. Yabuta, G. D. Cody, The origin and evolution of chondrites recorded in the elemental and isotopic compositions of their macromolecular organic matter. *Geochim. Cosmochim. Acta* **71**, 4380–4403 (2007).
38. C. M. O'D. Alexander *et al.*, The provenances of asteroids, and their contributions to the volatile inventories of the terrestrial planets. *Science* **337**, 721–723 (2012).
39. T. K. Ho, "Random decision forest" in *Proceedings of the International Conference on Document Analytical Recognition* (1995), vol. 3, https://doi.org/10.1109/ICDAR.1995.598994.
40. G. Khavari-Khorassani, D. G. Murchison, The nature of Siberian shungite. *Chem. Geol.* **26**, 165–182 (1979).
41. V. A. Melezhik, M. M. Filippov, A. E. Romashkin, A giant paleoproterozoic deposit of shungite in NW Russia: Genesis and practical applications. *Ore Geol. Rev.* **24**, 135–154 (2004).
42. G. H. Eldridge, "The asphalt and bituminous rock deposits of the United States: Part 1-Director's report and a paper on asphalt and bituminous rock deposits" in *Twenty-Second Annual Report of the United States Geological Survey to the Secretary of the Interior, 1900-01: no. 22; pt. 1-01-14* (1901), pp. 209–364.

43. C. Richardson, Grahamite, a solid native bitumen. *J. Am. Chem. Soc.* **32**, 1032–1049 (1910).
44. J. R. Dunn, D. W. Fisher, Occurrence, properties, and paragenesis of anthraxolite in the Mohawk Valley [New York]. *Am. J. Sci.* **252**, 489–501 (1954).
45. A. C. Hutton, Petrographic classification of oil shales. *Int. J. Coal Geol.* **8**, 203–231 (1987).
46. T. Boden, B. T. Tripp, "Gilsonite veins of the Uinta Basin, Utah" in *Utah Geological Survey Special Study* (2012), vol. **141**, 50p.
47. J. R. Nance *et al.*, Preserved shell-binding protein and associated pigment in the Middle Miocene (8 to 18 Ma) gastropod *Ecphora*. *Geochem. Perspect. Lett.* **1**, 1–8 (2015).
48. P. Schmitt-Kopplin *et al.*, High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 2763–2768 (2010).
49. G. P. Ellis, The maillard reaction. *Adv. Carbohydr. Chem.* **14**, 63–134 (1959).
50. Y. Kebukawa, A. D. Kilcoyne, G. D. Cody, Exploring the potential formation of organic solids in chondrites and comets through polymerization of interstellar formaldehyde. *Astrophys. J.* **771**, 19 (2013).
51. M. M. Tice, D. R. Lowe, Photosynthetic microbial mats in the 3, 416-Myr-old ocean. *Nature* **431**, 549–552 (2004).
52. E. E. Stüeken, R. Buick, B. M. Guy, M. C. Koehler, Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature* **520**, 666–669 (2015).
53. F. Westall *et al.*, Archean (3.33 Ga) microbe-sediment systems were diverse and flourished in a hydrothermal context. *Geology* **43**, 615–618 (2015).
54. J. Alleon, R. E. Summons, Organic geochemical approaches to understanding early life. *Free Radic. Biol. Med.* **140**, 103–112 (2019).
55. J. Alleon *et al.*, Organo-mineral associations in chert of the 3.5 Ga Mount Ada Basalt raise questions about the origin of organic matter in paleoarchean hydrothermally influenced sediments. *Sci. Rep.* **9**, 16712 (2019).
56. E. J. Javaux, Challenges in evidencing the earliest traces of life. *Nature* **572**, 451–460 (2019).
57. K. Lepot, Signatures of early microbial life from the Archean (4 to 2.5 Ga) eon. *Earth-Sci. Rev.* **209**, 103296 (2020).
58. B. Cavalazzi *et al.*, Cellular remains in a ~3.42-billion-year-old subseafloor hydrothermal environment. *Sci. Adv.* **7**, eabf3963 (2021).
59. A. H. Knoll, *A Brief History of Earth: Four Billion Years in Eight Chapters* (HarperCollins, New York, NY, 2021).
60. T. W. Lyons *et al.*, Co-evolution of early Earth environments and microbial life. *Nat. Rev. Microbiol.* **22**, 572–586 (2024).
61. J. Jodder, A. Hofmann, P. Durand, "The Archaean record of the Singhbhum Craton: A new window into early life on Earth" in *The Archean Earth, 2nd Edition of The Precambrian Earth. Earth and Environment–Geology*, M. Homann *et al.*, Eds. (Elsevier, 2025).
62. E. E. Schisterman, D. Faraggi, B. Reiser, J. Hu, Youden index and the optimal threshold for markers with mass at zero. *Stat. Med.* **27**, 297–315 (2008).
63. W. Lee, K. Seo, Downsampling for binary classification with a highly imbalanced dataset using active learning. *Big Data Res.* **28**, 100314 (2022).
64. C. Esposito *et al.*, GHOST: Adjusting the decision threshold to handle imbalanced data in machine learning. *J. Chem. Inf. Model.* **61**, 2623–2640 (2021).
65. S. S. Johnson *et al.*, Fingerprinting non-terran biosignatures. *Astrobiology* **18**, 915–922 (2018).
66. M. A. Chan *et al.*, Deciphering biosignatures in planetary contexts. *Astrobiology* **19**, 1075–1102 (2019).
67. N. Guttenberg, H. Chen, T. Mochizuki, H. J. Cleaves II, Classification of the biogenicity of complex organic mixtures for the detection of extraterrestrial life. *Life* **11**, 461 (2021).
68. N. Noffke, Microbially induced sedimentary structures in clastic deposits: Implication for the prospection for fossil life on Mars. *Astrobiology* **21**, 866–892 (2021).
69. T. L. Salter, B. A. Magee, J. H. Waite, M. A. Sephton, Mass spectrometric fingerprints of bacteria and archaea for life detection on icy moons. *Astrobiology* **22**, 143–157 (2022).
70. M. C. Figueroa *et al.*, A machine-learning approach to biosignature exploration on early Earth and Mars using sulfur isotope and trace element data in pyrite. *Astrobiology* **24**, 1110–1127 (2024).
71. J. W. Schopf, Microfossils of the early Archean Apex chert: New evidence of the antiquity of life. *Science* **260**, 640–646 (1993).
72. M. D. Brasier *et al.*, Questioning the evidence for Earth's oldest fossils. *Nature* **416**, 76–81 (2002).
73. S. Derenne *et al.*, Molecular evidence for life in the 3.5 billion year old Warrawoona chert. *Earth Planet. Sci. Lett.* **272**, 476–480 (2008).
74. T. Gold, *The Deep Hot Biosphere* (Copernicus, New York, NY, 1999).
75. J. F. Kenney *et al.*, Dismissal of claims of a biological connection for natural petroleum. *Energia* **22**, 26–34 (2001).
76. A. Brown, Upwelling of hot gas. *Am. Sci.* **87**, 372 (1999).
77. M. Sephton, R. M. Hazen, On the origins of deep hydrocarbons. *Rev. Mineral. Geochem.* **75**, 449–465 (2013).
78. A. I. Rushdi, B. R. T. Simoneit, Lipid formation by aqueous Fisher–Tropsh type synthesis over a temperature range of 100–400 °C. *Origins Life Evol. Biosphere* **31**, 103–118 (2001).
79. T. M. McCollom, J. S. Seewald, Carbon isotope composition of organic molecules produced by abiotic synthesis under hydrothermal conditions. *Earth Planet. Sci. Lett.* **243**, 74–84 (2006).
80. Q. Fu *et al.*, Abiotic formation of hydrocarbons under hydrothermal conditions: Constraints from chemical and isotopic data. *Geochim. Cosmochim. Acta* **71**, 1982–1998 (2007).
81. C. E. Manning, E. L. Shock, D. A. Sverjensky, The chemistry of carbon in aqueous fluids at crustal and upper mantle conditions: Experimental and theoretical constraints. *Rev. Mineral. Geochem.* **75**, 109–148 (2013).
82. T. M. McCollum, Laboratory simulations of abiotic hydrocarbon formation in Earth's deep subsurface. *Rev. Mineral. Geochem.* **75**, 467–494 (2013).
83. Ménez *et al.*, Abiotic synthesis of amino acids in the recesses of the oceanic lithosphere. *Nature* **564**, 59–63 (2018).
84. B. Rasmussen, J. R. Muhling, Organic carbon generation in 3.5-billion-year-old basalt-hosted seafloor hydrothermal vent systems. *Science. Sci. Adv.* **9**, eadd7925 (2023).
85. J. Fan, S. A. Upadhye, A. Worster, Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* **8**, 19–20 (2006).
86. S. Narkhede, Understanding AUC–ROC curve. *Towards Data Sci.* **26**, 220–227 (2018).
87. F. S. Nahm, Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean J. Anesthesiol.* **75**, 25–36 (2022).
88. D. Gourier *et al.*, Extraterrestrial organic matter preserved in 3.33 Ga sediments from Barberton, South Africa. *Geochim. Cosmochim. Acta* **258**, 207–225 (2019).
89. D. P. Glavin *et al.*, "Chapter 3. The origin and evolution of organic matter in carbonaceous chondrites and their links to parent bodies" in *Primitive Meteorites and Asteroids*, N. M. Abreu, Ed. (Elsevier, Amsterdam, The Netherlands, 2018), pp. 205–271.
90. T. Schulz *et al.*, The history of the Tissint meteorite, from its crystallization on Mars to its exposure in space: New geochemical, isotopic, and cosmogenic nuclide data. *Meteorit. Planet. Sci.* **55**, 294–311 (2020).
91. G. D. Cody *et al.*, The nature of insoluble organic matter in Sutter's Mill and Murchison carbonaceous chondrites: Testing the effect of X-ray computed tomography and exploring parent body organic molecular evolution. *Meteorit. Planet. Sci.* **59**, 3–22 (2024).
92. M. T. Rosing, N. M. Rose, D. Bridgwater, H. S. Thompsen, Earliest part of Earth's stratigraphic record: A reappraisal of the >3.7 Ga Isua (Greenland) supracrustal sequence. *Geology* **24**, 43–46 (1996).
93. A. P. Nutman, V. C. Bennett, C. R. L. Friend, M. van Kranendonk, The Eoarchean legacy of Isua (Greenland) worth preserving for future generations. *Earth-Sci. Rev.* **198**, 102923 (2019).
94. F. Westall, R. L. Folk, Exogenous carbonaceous microstructures in early Archaean cherts and BIFs from the Isua greenstone belt: Implications for the search for life in ancient rocks. *Precambrian Res.* **126**, 313–330 (2003).
95. M. van Kranendonk *et al.*, Geological setting of Earth's oldest fossils in the Ca. 3.5 Ga Dresser Formation, Pilbara Craton, Western Australia. *Precambrian Res.* **167**, 93–124 (2006).
96. M. D. Brasier *et al.*, "Geology and putative microfossil assemblage of the c. 3460 Ma 'Apex chert', Chinaman Creek, Western Australia" in *Geological Survey of Western Australia, Record 2011/7* (2011), 60p.
97. A. O. Marshall, J. R. Emry, C. P. Marshall, Multiple generations of carbon in the Apex Chert and implications for preservation of microfossils. *Astrobiology* **12**, 160–166 (2012).
98. M. J. van Kranendonk, A. Kroner, E. Hegner, J. Connelly, Age, lithology and structural evolution of the 3.53 Ga Theespruit Formation in the Tjakastad area, southwestern Barberton Greenstone Belt, South Africa, with implications for Archaean tectonics. *Chem. Geol.* **261**, 115–139 (2009).
99. S. M. Jones, Fluid flow, alteration, and timing of Cu-Ag mineralization at the White Pine sediment-hosted copper deposit, Michigan, USA. *Econ. Geol.* **118**, 1431–1465 (2023).
100. R. H. Rainbird, J. A. Donaldson, Nonglaciogenic deltaic deposits in the early Proterozoic Gowganda Formation, Cobalt Basin. *Can. J. Earth Sci.* **25**, 710–724 (1988).
101. G. M. Young, H. W. Nesbitt, The gowganda formation in the southern part of the Huronian outcrop belt, Ontario, Canada: Stratigraphy, depositional environments and regional tectonic significance. *Precambrian Res.* **29**, 265–301 (1985).
102. K. Kennedy, N. Eyles, The Paleoproterozoic (c.2.3 Ga) Gowganda Formation: Deep water, glacially-influenced debrites and related mass flow along a passive margin. *Earth-Sci. Rev.* **261**, 105033 (2025).
103. A. Corpolongo, A. Czaja, Extra-large and morphologically unique microfossils of the 2.52 Ga Gamohaan Formation, South Africa. *Authorea*, June 21, 2019.
104. N. McLoughlin *et al.*, Microbial sulphur-cycling and atmospheric signatures in the 2.52 Ga Gamohaan Formation, South Africa. *Earth Planet. Sci. Lett.* **602**, 117941 (2023).
105. A. M. Thorne, A. F. Trendall, Geology of the Fortescue Group Pilbara Craton Western Australia. *Geol. Surv. Western Australia Bull.* 144, 1–266 (2001).
106. J. Kasbohm *et al.*, Paleogeography and high-precision geochronology of the Neoarchean Fortescue Group, Pilbara, Western Australia. *Precambrian Res.* **394**, 107114 (2023).
107. A. D. Anbar *et al.*, A whiff of oxygen before the great oxidation event? *Science* **317**, 1903–1906 (2007).
108. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
109. S. M. Awramik, The oldest records of photosynthesis. *Photosynth. Res.* **33**, 75–89 (1992).
110. J. W. Schopf, The paleobiological record of photosynthesis. *Photosynth. Res.* **107**, 87–101 (2011).
111. N. J. Butterfield, Proterozoic photosynthesis–A critical review. *Paleontology* **58**, 953–972 (2015).
112. A. A. Davin *et al.*, A geological timescale for bacterial evolution and oxygen adaptation. *Science* **388**, 6742 (2025).
113. M. C. Sforna *et al.*, Intracellular bound chlorophyll residues identify 1 Gyr-old fossils as eukaryotic algae. *Nat. Commun.* **134**, 146 (2022).
114. J. D. Ayala, E. R. Schroeter, M. H. Schweitzer, Porphyrin-based molecules in the fossil record shed light on the evolution of life. *Minerals* **14**, 201 (2024).
115. C. Demoulin *et al.*, *Polysphaeroides filiformis*, a proterozoic cyanobacterial microfossil and implications for cyanobacteria evolution. *iScience* **27**, 2 (2024).
116. C. F. Demoulin, Y. J. Lara, A. Lamblon, E. J. Javaux, Oldest thylakoids in fossil cells directly evidence oxygenic photosynthesis. *Nature* **625**, 529–534 (2024).
117. J. J. Brocks *et al.*, The rise of algae in Cryogenian oceans and the emergence of animals. *Nature* **548**, 578–581 (2017).
118. D. E. G. Briggs, Molecular taphonomy of animal and plant cuticles: Selective preservation and diagenesis. *Philos. Trans. R. Soc. Lond.* **B354**, 2–17 (1999).
119. V. Vajda *et al.*, Molecular signatures of fossil leaves provide unexpected new evidence for extinct plant relationships. *Nat. Ecol. Evol.* **1**, 1093–1099 (2017).
120. G. P. Fournier *et al.*, The Archean origin of oxygenic photosynthesis and extant cyanobacterial lineages. *Proc. R. Soc. B* **288**, 20210675 (2021).
121. R. E. Summons *et al.*, Preservation of Martian organic and environmental records: Final report of the Mars biosignature working group. *Astrobiology* **11**, 157–181 (2011).
122. B. J. Tipple, M. Pagani, The early origins of terrestrial C_4 photosynthesis. *Annu. Rev. Earth Planet. Sci.* **35**, 435–461 (2007).
123. M. F. Hohmann-Marriott, R. E. Blankenship, Evolution of photosynthesis. *Annu. Rev. Plant Biol.* **62**, 515–548 (2011).
124. R. S. Gupta, Chapter two—Molecular markers for photosynthetic bacteria and insights into the origin and spread of photosynthesis. *Adv. Bot. Res.* **66**, 37–66 (2013).
125. A. Nishihara, Y. Tsukatani, C. Azai, M. K. Nobu, Illuminating the coevolution of photosynthesis and Bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2322120121 (2024).
126. V. N. Sergeev, A. H. Knoll, J. P. Grotzinger, Paleobiology of the Mesoproterozoic Billyakh Group, Anabar Uplift, northern Siberia. *J. Paleontol.* **69**, 1–37 (1995).
127. E. H. Newcomb, T. D. Pugh, Blue-green algae associated with ascidians of the Great Barrier Reef. *Nature* **253**, 533–534 (1975).
128. D. W. Sumner, Late archean calcite-microbe interactions: Two morphologically distinct microbial communities that affected calcite nucleation differently. *PALAIOS* **12**, 302 (1997).
129. A. Corpolongo, *Neoarchean Microfossils and Microbialites Inform the Search for Extraterrestrial Life in the Solar System* (University of Cincinnati, 2024).

130. A. H. Knoll, P. K. Strother, S. Rossi, Distribution and diagenesis of microfossils from the lower Proterozoic Duck Creek Dolomite. *Precambrian Res.* **38**, 257–279 (1988).
131. J. P. Wilson *et al.*, Geobiology of the late Paleoproterozoic Duck Creek Formation, Western Australia. *Precambrian Res.* **179**, 135–149 (2010).
132. T. Kakegawa, H. Kawai, H. Ohmoto, Origins of pyrites in the ~2.5 Ga Mt. McRae Shale, the Hamersley District, Western Australia. *Geochim. Cosmochim. Acta* **62**, 3205–3220 (1998).
133. D. Y. Sumner, N. J. Beukes, Sequence stratigraphic development of the Neoarchean Transvaal carbonate platform, Kaapvaal Craton, South Africa. *S. Afr. J. Geol.* **109**, 11–22 (2006).
134. H. Schulz, Short history and present trends of Fischer–Tropsch synthesis. *Appl. Catal. A: Gen.* **186**, 3–12 (1999).
135. M. E. Dry, The Fischer–Tropsch process: 1950–2000. *Catal. Today* **71**, 227–241 (2002).
136. A. H. Knoll, The early evolution of eukaryotes: A geological perspective. *Science* **256**, 622–627 (1992).
137. J. J. Brocks, G. A. Logan, R. Buick, R. E. Summons, Archean fossils and the early rise of eukaryotes. *Science* **285**, 1033–1036 (1999).
138. E. J. Javaux, The early eukaryotic fossil record. *Adv. Med. Biol.* **607**, 1–19 (2007).
139. E. J. Javaux, A. H. Knoll, Micropaleontology of the lower Mesoproterozoic Roper Group, Australia, and implications for early eukaryotic evolution. *J. Paleontol.* **91**, 199–229 (2017).
140. E. J. Javaux, A diverse Paleoproterozoic microbial ecosystem implies early eukaryogenesis. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **380**, 20240092 (2025).
141. C. K. Boyce *et al.*, Chemical evidence for cell wall lignifications and the evolution of tracheids in early Devonian plants. *Int. J. Plant Sci.* **164**, 691–702 (2003).
142. C. K. Boyce *et al.*, Devonian landscape heterogeneity recorded by a giant fungus. *Geology* **35**, 399–402 (2007).
143. J. Panczak, P. Kosakowski, A. Zakrzewski, Biomarkers in fossil resins and their palaeoecological significance. *Earth Sci. Rev.* **242**, 104455 (2023).
144. T. O. Akinsanpe, S. A. Bowden, J. Parnell, Molecular and mineral biomarker record of terrestrialization in the Rhynie chert. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **640**, 112101 (2024).
145. J. Lindgren *et al.*, Molecular composition and ultrastructure of Jurassic paravian feathers. *Sci. Rep.* **5**, 13520 (2015).
146. L. Y. Yeung, J. L. Ash, E. D. Young, Biological signatures in clumped isotopes of O_2. *Science* **348**, 431–434 (2015).
147. J. M. Eiler *et al.*, Analysis of molecular isotopic structures at high precision and accuracy by orbitrap mass spectrometry. *Int. J. Mass Spectrom.* **422**, 126–142 (2017).
148. S. S. Zeichner *et al.*, Position-specific carbon isotopes of Murchison amino acids elucidate extraterrestrial abiotic organic synthesis networks. *Geochim. Cosmochim. Acta* **355**, 210–221 (2023).
149. C. G. Ryan *et al.*, SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrum. Methods Phys. Res. B* **34**, 396–402 (1988).
150. S. Gibb, K. Strimmer, MALDIquant: A versatile R package for the analysis of mass spectrometry data. *Bioinformatics* **28**, 2270–2271 (2012).
151. S. G. Patro, K. K. Sahu, Normalization: A preprocessing stage. arXiv [Preprint] (2015). https://doi.org/10.48550/arXiv.1503.06462 (Accessed 27 October 2025).
152. M. Kuhn, Classification and regression training, package, caret (2022). https://github.com/topepo/caret/. Accessed 27 October 2025.
153. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).