T H E U N I V E R S I T Y O F T U L S A

THE GRADUATE SCHOOL

EXPLAINABLE MACHINE LEARNING FOR OCEAN WORLDS BIOSIGNATURES AND SEAWATER CHEMISTRY

by
Lily A. Clough

A thesis submitted in partial fulfillment of

the requirements for the degree of Doctor of Philosophy

in the Discipline of Computer Science

July 2025

T H E U N I V E R S I T Y O F T U L S A

THE GRADUATE SCHOOL

EXPLAINABLE MACHINE LEARNING FOR OCEAN WORLDS BIOSIGNATURES AND SEAWATER CHEMISTRY

by
Lily A. Clough

A thesis submitted in partial fulfillment of

the requirements for the degree of Doctor of Philosophy

in the Discipline of Computer Science

By Thesis Committee

Brett McKinney, Chair
Bethany Theiling, co-Chair
John Hale, member
Mauricio Papa, member
Ian Riley, member

ACKNOWLEDGEMENTS

I would like to thank my mentors and friends for their support during my studies, and for providing experiences, opportunities, and research that have proven to be out of this world.

Lily A. Clough (Doctor of Philosophy in Computer Science)

Explainable Machine Learning for Ocean Worlds Biosignatures and Seawater Chemistry

Directed by Brett McKinney and Bethany Theiling

182 pp., Chapter 6: Conclusions

(473 words)

Explainable machine learning (ML) methods are needed to support future missions to ocean worlds (OWs) such as Europa and Enceladus for biosignature detection and chemical characterization. Explainable ML methods allow black box model predictions to be interpreted or explained. Such methods are essential for resource-constrained geochemical and astrobiological missions to OWs in which the environment is so remote. Additionally, the stakes of false predictions are high in the detection of an extraterrestrial biosignature, and so ML predictions must include an explanation and an assessment of confidence and trustworthiness to reduce risk and protect mission resources.

For planetary exploration, mass spectrometry (MS) is a ubiquitous tool due to its accuracy and rich spectral data products that provide valuable information about planetary surfaces and subsurfaces, including quantification of elemental and isotopic composition. Elemental and isotopic information allows scientists to deduce crucial geochemical facts about planetary bodies, such as the history, origins, fate, and potential biological content of extraterrestrial rocks, volatiles, and liquid water. Fractionation of light carbon and oxygen isotopes is known to be indicative of microbial and photosynthetic life on Earth, making it a promising tool for biosignature detection. However, biotic signals can be masked by abiotic effects in an environment and abiotic chemistry can mimic biotic environments. To address

these potential false positives and negatives and combat the erosion of trust in ML, this research focuses on developing data processing tools and ML algorithms for isotope ratio mass spectrometry (IRMS) measurements of volatile CO_2 that include explainability and diagnostics for false predictions for future astrobiological investigations of OWs.

This dissertation introduces novel ML methods for the prediction and explanation of biosignatures and seawater chemistry from laboratory derived IRMS data with biotic and abiotic signatures as well as simulated data. The ML methods introduced in this dissertation include novel time-series feature construction, a novel distance metric approach with feature selection that results in a human-interpretable variable (feature) space. In addition, a novel local variable importance method is developed that provides a level of explanation for the prediction of a single sample. This local Nearest-neighbors Projected Distance Regression method (local-NPDR) can detect statistical interactions and can be used to diagnose potential false predictions. We use these methods as well as an interpretable biosignature network visualization to explain predictions by biosignature and seawater chemistry models for OW analogue brine salt components, volatile CO_2 concentration, pH, and ionic strength. A quality analysis/quality control (QA/QC) data processing tool is demonstrated in a simulated Enceladus mission concept to illustrate real time experimental data processing for use in biosignature classification and seawater chemistry models. The local feature importance method is demonstrated on simulated and real geochemical isotopic data and will be demonstrated in the field. While primarily focused on ML applications for icy OWs, the explainable ML methods presented here may be applied to other scientific datasets for any number of planetary environments or analogues. The explainable ML methods presented here are expected to be practical tools that will increase trust for future autonomous planetary exploration.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

"It is the desire for explanations which are at once systematic and controllable by factual evidence that generates science; and it is the organization and classification of knowledge on the basis of explanatory principles that is the distinctive goal of the sciences."
– **Ernest Nagel,** ***The Structure of Science***

Future missions to icy ocean worlds (OW) such as Europa and Enceladus will evaluate the geochemistry, habitability, and potential for biosignatures on these worlds. These missions will benefit from autonomous science and machine learning (ML) methods to process high volumes of data and prioritize signals of interest. Mass spectrometers (MS) are suitable instruments for implementing science autonomy due to their high precision and potential for biosignature detection (evidence of past or present biological activity). In this work, we consider ML methods for the remote prediction of biosignatures from MS measurements of volatile carbon dioxide. This remote detection, as opposed to the direct detection of a biomolecule or organism, makes trust and explainability of the ML approach an important consideration. Trust and explainability are important for scientific applications of ML and even more so for a prediction of extraterrestrial life. In this work, we develop new explainable ML methods for astrobiology applications.

Machine learning (ML) has become a widespread tool for scientific data analysis and is increasingly used in hybrid modeling to predict physical processes (Noordijk et al., 2024). In a utilitarian sense, the goals of ML and science parallel each other: both seek to make accurate and practical predictions. Science and ML achieve this by finding generalizable regularities in data that can be used for prediction. In science, these regularities may become elevated to the status of

a law, which is a distillation of complex data in a form that services another, deeper goal: explanation or understanding. Indeed, one of the most important goals of science is to make nature intelligible to humans by providing insights into the mechanisms by which natural phenomena occur; *i.e.*, to provide scientific explanations (Nagel, 1979).

For ML, regularities found in data are encapsulated in a statistical model or algorithm; however, ML model predictions usually cannot be easily explained like a scientific law and are often likened to "black boxes". Increasingly in critical and high-risk domains, the need for transparency, explainability, and interpretability in ML model predictions has been recognized (Linardatos et al., 2020; Roscher et al., 2020). Transparency means that the algorithmic mechanisms and parameter space through which ML predictions are made are fully understandable and reproducible, while interpretability refers to the ability to draw connections between model predictions and the scientific domain to be understood (Montavon et al., 2018; Roscher et al., 2020). Explainability is less easily defined but perhaps the most crucial aspect of a scientifically-relevant ML model. In ML, explainability can be defined as a highly relevant feature (variable) space (Roscher et al., 2020) incorporated into an algorithm (with some level of transparency) where interpretations about model predictions are able to be made. Explainable ML tools therefore attempt to add some degree of transparency to models, frequently by leveraging a variable space that is mathematically (and if possible, physically) understandable and connecting these variable mathematical abstractions to the ML predictions (*i.e.*, interpretation). These tools provide a series of explanatory principles upon which the ML model prediction can be understood. In this way, trust and explainability in ML are inextricable, as they are in science.

Astrobiology offers an enticing problem for ML – how can we accurately detect the presence of life in an unknown environment of unknown history? The ability to trust an ML model prediction is crucial for such a high-risk scientific question. In the remote locations of proposed astrobiological targets, it is not possible to directly verify whether a biosignature prediction is true or false. Therefore, explanatory false detection tools will be necessary for astrobiology missions. False positive (FP) and false negative (FN) biosignature detection using remote sensing is a well-documented challenge (National Academies of Sciences, Engineering, and Medicine, 2019). Abiotic environments with complex geochemistry can mimic a biosignature, leading to a FP, or the environment can mask a biosignature prediction, leading to a FN (Clough et al., 2025). Autonomous decision making based on ML and artificial intelligence (AI) can make space missions more efficient, but the risk of false predictions must be mitigated, both to protect mission resources and to instill trust in real-time ML analysis of collected data (Theiling et al., 2022a; Da Poian et al., 2025). These examples underscore the importance of interpreting ML predictions in the context of the geochemical environment, using training data that accurately reflects the target deployment environment, and diagnosing false predictions.

However, ML models suffer from an accuracy-explainability tradeoff. As data dimensionality has increased across research fields, ML models have improved in accuracy but grown in complexity, often resulting in "black box" systems with limited transparency of their decision-making process. This high-dimensionality and increased opacity in algorithmic mechanisms results in decreased explainability. For scientific models, explanation is often built into the model in terms of the mathematical symbols that describe physical laws. In this way scientific and ML models have different levels of inherent transparency and explainability. The most transparent model is one whose exact mechanism for prediction is comprehensible to a

human. For example, a decision tree model has a high level of transparency (*i.e.*, a "transparent box"). Its decision-making process can be followed for each variable split in the tree for a given sample, and the structure of the tree gives some explainability: nodes (variables) at the top have the highest variable importance and branches connecting variables may suggest conditional relationships. Unfortunately, its prediction accuracy is not high enough in most applications, which led to resampling methods like Random Forest (RF) (Breiman, 2001). The many trees (forest) used by RF to vote on sample classes is responsible for its improved accuracy but also reduces its explainability.

ML tools can provide global and/or local explainability; global explainability results from generalizations made across all training samples, while local explainability focuses on one sample or a neighborhood of samples (Roscher et al., 2020). While RF is on the opaque end of the transparency spectrum, it does provide tools for global and local explainability such as permutation variable importance. For an important variable, the permutation importance score increases if the out of bag (oob) accuracy of the model decreases after permuting the variable. Permutation importance thus provides a degree of explanation by ranking which variables the RF model finds most necessary for prediction. This importance method is global in that it aggregates information across all training samples and the scores are not specific to explaining an individual sample's prediction. To address this, RF has a local version of permutation importance that gives variable importance scores specific to the prediction for each sample in the training data.

Light stable isotopes like carbon and oxygen in volatile CO_2 molecules are strong candidates for biosignatures due to the large fractionations promoted by biological activity (Park & Epstein, 1960; Vogel, 1980; Krzycki et al., 1987; Schidlowski, 2001). One limitation for methods seeking to detect isotopic biosignatures is that complex abiotic geochemistry may

obscure or mimic biogenic isotope fractionations. While ML may accurately disentangle biosignatures from abiotic mimicry in MS data, ML model predictions can be inscrutable to human interpretation, compromising trust in scientifically significant detections, such as an OW biosignature.

We present a suite of explainable ML models for biosignature detection and seawater chemistry prediction using an isotope ratio mass spectrometry (IRMS) dataset consisting of volatile CO_2 measurements from OW analogue brines (Theiling, 2021; Clough et al., 2025). These experiments are designed to be analogues for Europa and Enceladus seawaters and have a wide range of pH values and ionic strengths to ensure models are robust to multiple possible OW seawater chemistries. Some brines are inoculated with either sulfate-reducing or an uncharacterized heterogeneous mixture of microbes, and others are prepared abiotically. To ensure biosignature models are trained on geochemically complex abiotic data, these data include biotic mimicry achieved by the addition of non-biogenic organics to abiotic brines. Biotic mimicry in an astrobiological context refers to abiotic processes producing patterns or chemical signatures that resemble those produced by life. Such processes on icy OWs could include irradiation, rapid depressurization, high ionic strengths, and phase changes (Craig & Gordon, 1965; Thiemens & Heidenreich, 1983; Cooper et al., 2001; Gaisser et al., 2016).

Our ML approach includes feature, or variable, construction, which provides mathematical and geochemical context for biosignatures (see Sec. 2.1), and our ML feature selection method called Nearest-neighbors Projected Distance Regression (NPDR) that identifies important predictors through the detection of variable main effects as well as statistical interactions (Le et al., 2020) (see Sec. 2.2.1). NPDR feature selection can employ regularization via LASSO (Least Absolute Shrinkage and Selection Operator) or Ridge penalties or can report

feature importance scores using P-values. Using a LASSO penalty reduces the variable space while ensuring selected features are independent (Breiman, 1995; Tibshirani, 1996). Since NPDR feature selection is a nearest-neighbors based algorithm, it requires a distance matrix and can either compute a traditional distance metric such as Euclidean, Mahalanobis, or Manhattan, or accept a user-defined distance. We use a novel distance matrix in NPDR generated by the unsupervised Random Forest proximity (URFP) that is output by unsupervised RF in the original Fortran implementation of RF and in the ranger R package (Breiman, 2001). This distance metric can account for a non-isotropic variable space, unlike traditional distance metrics.

NPDR feature selection often results in feature spaces that are small enough to be visualized in an interpretable network which illustrates the variable main effects and statistical interactions that inform model predictions (Davis et al., 2010; Lareau et al., 2015). Understanding how variables work together globally to inform predictions such as biosignatures increases our ability to understand individual predictions. Local importance scores provide feature importance scores for a single sample. If an individual prediction has local variable importance scores that contradict the global pattern, we hypothesize that there is an increased likelihood that the prediction is false. We extend NPDR feature importance to create a novel local feature importance method called local-NPDR. We compare the ability of local-NPDR to diagnose false predictions with local-RF feature importance for three simulated datasets and the real biosignature data. RF models and local-NPDR false prediction analysis are provided for seawater chemistry predictions, and applications of our interpretable ML models and methods are discussed.

1.1 Mass Spectrometry for Ocean Worlds Exploration

OWs such as Europa and Enceladus have been observed or modeled to have all the necessary ingredients for life, including liquid H_2O, essential elements (C, H, N, O, P, and S), sources of free energy and nutrients required to support microbial metabolisms, and nutrient cycling (McCollom, 1999; Chyba, 2000; Chyba & Phillips, 2001). Because these targets are of high scientific interest, future planned and proposed missions to OWs such as Europa Clipper (*e.g.*, (Vance et al., 2023)) or the Enceladus Orbilander (MacKenzie et al., 2021) plan to assess their habitability and the potential for extant or extinct life by characterizing the surface and subsurface of these worlds using remote and *in situ* methods.

Europa Clipper and future proposed icy OW missions include MSs due to their promise for detecting potential biosignatures on OWs (Chou et al., 2021; Neveu et al., 2018). For example, the Europa Clipper mission is equipped with a time-of-flight mass spectrometer (TOF-MS), the MASPEX (MAss Spectrometer for Planetary EXploration), capable of measuring isotope ratios of ejected plume and exosphere volatiles (Brockwell et al., 2016; Howell & Pappalardo, 2020; Waite et al., 2024).

Isotopes of elements such as oxygen and carbon are of particular interest for biosignatures due to their role in metabolic processes, such as photosynthesis and methanogenesis, during which organisms are shown to uptake lighter isotopes (*e.g.,* Park & Epstein, 1960; Vogel, 1980; Krzycki et al., 1987; Schidlowski, 2001). Analyzing isotope ratios of C-H-O-N-S-bearing compounds can therefore serve as informative indicators of planetary processes and habitability (Miller et al., 2021). However, the nature of planetary exploration frequently requires that compositions and biosignature potential of the surface/subsurface be determined from orbit. It is therefore essential to understand whether the composition of OW

subsurfaces will be reflected in sampled volatiles like CO_2 (Theiling, 2021). Indeed, CO_2 is a major constituent in Enceladus's plumes (Waite et al., 2006; Waite Jr et al., 2009), and recently confirmed on Europa by measurements of the James Webb Space Telescope (Trumbo & Brown, 2023; Villanueva et al., 2023). However, biogenic isotope fractionations (typically single enzymatic steps) may be mimicked by abiotic geological processes in complex geochemical environments (Schidlowski, 2001; Barge et al., 2022). An important goal of biosignature study data is to create abiotic controls that include biotic mimicry by creating laboratory geochemical conditions that produce fractionations resembling those produced by biological mechanisms.

1.2 Machine Learning of Mass Spectrometry Data

Modern MS along with data science and ML methods are potentially capable of conducting autonomous science (Theiling et al., 2022b) to identify isotopic biosignatures since metabolic processes are known to produce large isotope fractionations (Craig, 1953; Park & Epstein, 1960; Vogel, 1980; Schidlowski, 2001). Therefore, to trust in a positive ML prediction as extraordinary as an extraterrestrial biosignature, researchers must have confidence that the experimental signal is not explainable by abiotic geochemical processes. Assumptions used to generate training data can lead to false predictions because the ML model can only learn patterns that exist in the data. If a model has never seen convincing biotic mimicry in training data, the model is susceptible to this type of false prediction. Hence, model interpretability is crucial for ML biosignatures to understand both the isotope chemistry underlying a potential biosignature and the factors that increase the risk of a false prediction. If an ML model is fully interpretable, then a person can follow its logic and make judgements as to the veracity of its prediction. However, the complexity of a model necessary to achieve high accuracy often reduces model

interpretability. For example, the decision tree, the basis of the widely used ML classifier Random Forest (RF) (Breiman, 2001; Shi & Horvath, 2006), has very high interpretability; however, the higher accuracy forest (an ensemble of trees) comes at the cost of interpretability of the individual decision tree models. Thus, additional tools are needed to improve interpretability while maintaining accuracy.

In a recent study, Cleaves et al. used gas chromatography (GC)MS of abiotic and biotic rock samples to train an RF model and visualize sample patterns in the two-dimensional principal component space (Cleaves et al., 2023). Principal component analysis (PCA) allows one to visualize patterns of covariation among samples in a lower dimensional feature space (Maćkiewicz & Ratajczak, 1993). For example, PCA is commonly used in human genetics data to characterize clusters of ancestry, although restricting which PCs are used may affect reproducibility (Elhaik, 2022). PCA does not provide a direct interpretation of RF model predictions, however, since PCA is unsupervised and does not provide information about supervised model predictions. In addition, PCA is a linear transformation of the multivariate space and does not capture interactions between variables, the identification of which can help with model interpretation.

A statistical interaction is defined as the variation between two variables that is conditioned on the outcome variable; for example, the linear correlation between two variables may change based on whether samples are biotic or abiotic (McKinney et al., 2009; Lareau et al., 2015). A main effect, on the other hand, is defined as an association of a variable with the outcome that does not depend on other variables. Statistical interactions and main effects can act as variable importance scores that provide interpretability of a model by weighting variables by the degree to which they contribute to the classification of samples (Guyon & Elisseeff, 2003).

However, most variable importance scores, including those computed by RF, have limited ability to detect interactions (McKinney et al., 2009; Wright et al., 2016). In the presence of biotic mimicry, the main effect of a fractionation variable may have limited ability to distinguish between biotic and abiotic samples, and interactions are expected to play an important role in helping an ML model to discriminate biotic from abiotic samples. This potential for interaction effects motivates our use of Nearest Neighbor Projected Distance Regression (NPDR) feature selection and variable importance (Le et al., 2020).

Variable importance scores are typically computed with respect to all samples. However, RF is also able to compute variable importance scores for a single sample, which we use to explain model predictions of an individual sample and assess the probability that the prediction is a false prediction. A recent approach to single-sample explainability, which can be used for any ML model, is to locally linearize a complex nonlinear ML model in the neighborhood of a single sample. While linearized models are powerful, they cannot fully explain the predictions of a full nonlinear RF model in the same way that RF local variable importance scores might. A limitation of the RF local importance method is that the sample for which predictions are being made must be in the training data, meaning valuable computational resources are needed to provide insight into the nonlinear model prediction.

1.3 Isotope Ratio Mass Spectrometry (IRMS) Data and Processing

Laboratory simulations of OW seawater-CO_2 equilibrations are described in Theiling (2021). In these laboratory experiments, gaseous CO_2 is injected into the headspace of vials containing analogue OW seawater solutions with and without microbes, allowed to equilibrate for seven days, and the equilibrated volatile CO_2 is analyzed using a Thermo Gasbench II trace gas analyzer coupled to a Thermo Delta V Advantage isotope ratio (IR)MS. We create the

analogue OW seawaters using a range of known and hypothesized compositions for the subsurface oceans of Europa and Enceladus (Anderson et al., 1997; Kargel et al., 2000; Fanale et al., 2001; Zolotov & Shock, 2001; Brown & Hand, 2013; Trumbo et al., 2019; Theiling, 2021). Salts include various combinations and concentrations of NaCl, Na_2SO_4, $NaHCO_3$, $MgCl_2$, KCl, and $MgSO_4$ in ultrapure 18 MΩ water (for more details see Sec. A1). In these experiments, a mixture of 0.3% CO_2 in helium is equilibrated with 0.5 mL of analogue OW seawaters (with or without microbes) in chlorobutyl septa-capped 12 mL vials. We prepare all samples in triplicate, which are analyzed in batches of ≤ 96 samples and standards.

Brine experiments span a range of pH values (3.5 – 9.5) and ionic strengths (0.00 – 15.46 M) (see Table A1) to reflect different CO_2 speciations. Brine experiments were designed to reflect hypothesized OW chemistries. For Europa, hypothesized pH values range from acidic (Kargel et al., 2000; Carlson et al., 2009; Zolotov & Kargel, 2009; Pasek & Greenberg, 2012; Muñoz-Iglesias et al., 2013) to slightly alkaline (M.Y. Zolotov, 2008; Zolotov & Kargel, 2009; Vance et al., 2016; Bouquet et al., 2017; Russell et al., 2017). $NaHCO_3$ grains detected from Enceladus's plumes (Postberg et al., 2009) suggest a neutral to alkaline ocean, buffered by detected CO_2 and H_2 (Waite et al., 2006; Zolotov, 2007; Waite Jr et al., 2009; Glein et al., 2015; Glein & Waite, 2020). Similar experiments in Theiling (2021) demonstrate that alkaline experiments have more variable $\delta^{13}C_{CO_2}$, dependent on carbonate species concentrations, while more acidic experiments indicate $\delta^{13}C_{CO_2}$ is more directly related to the fractionation of CO_2 prior to interaction with the brine solutions.

Biotic experiments are inoculated with either collected or commercially-provided microbes with different metabolic activities so the ML model is exposed to different metabolic processes. Anaerobic microbial cultures are cultivated following standardized procedures

(Tanner, 2007) (see Sec. A1 for more information on microbial cultivation). We use a commercially-provided strain of *Desulfotomaculum thermocisternum*, a thermophilic species of sulfate-reducing bacteria, to observe isotopic fractionation due to sulfate reduction that may be promoted on Europa. We also inoculate a subset of experiments with an uncharacterized and heterogeneous mixture of anaerobic microbes from a reducing ephemeral pond environment to simulate data what might be collected from an uncharacterized OW biosphere. Some abiotic experiments include an abiotic organic growth medium to act as geochemical decoys.

The time-of-flight (TOF)-IRMS analyses yield chromatograms, or signal (intensity) versus time plots, with amplitude (mV) on the *y*-axis and time (s) on the *x*-axis. Industry-standard techniques for the trace gas analyzer method of water-equilibration analysis are typified by multiple subsamples of the sample gas (to ensure data precision) and a comparison to multiple injections of a reference gas with known isotopic composition to calculate isotope ratios. Our quality analysis/quality control (QA/QC) pipeline characterizes experiments as pass or fail (for a full description of the QA/QC checks and pass/fail chromatograms see Sec. A1). QA/QC thresholds were informed by experimental and domain expert knowledge.

1.4 Outline of Dissertation

In Chapter 2, we describe a novel ML approach and a Benchmark Ocean Worlds-δCO_2 (BOW-δCO_2) data set of IRMS measurements and derived variables using volatile CO_2 from OW analogue brines. The OW analogue BOW-δCO_2 data set is provided as a resource to the astrobiology research community. The benchmark data is designed to be geochemically complex (i.e., contain biotic mimicry) and includes a range of pH values, ionic strengths, salt mixtures, and salt concentrations. We describe a new ML approach, which includes feature

construction/engineering, nearest-neighbors-based feature selection with a novel distance metric, RF biosignature classification, and multiple modes of interpretability. The ML approach, yields a biosignature model with an average accuracy of 87.3%.

In Chapter 3, we describe the new local-NPDR method for single-sample variable importance, and we describe the simulated and real biosignature data. We compare local-NPDR with other local feature importance methods for the simulated data and real biosignature laboratory data based on the ability to explain and detect false biosignatures. We use local scores to explain which features a classifier might find most important for classifying a specific sample, and we use the discordance between global and local scores combined with single-sample prediction probabilities to flag potential false predictions. In Chapter 4, we apply the above methods to explain and predict seawater chemistry by predicting the presence of NaCl, Na_2SO_4, $NaHCO_3$, $MgCl_2$, KCl, $MgSO_4$, and CO_2 levels, and by predicting continuous outcomes such as pH and ionic strength. Chapter 5 discusses applications of these methods in autonomous astrobiology scenarios. Chapter 6 discusses implications of this work and directions for future research.

CHAPTER 2

MACHINE LEARNING BIOSIGNATURE MODEL USING OCEAN WORLDS ANALOGUE DATA

2.1 Novel Time Series Feature Extraction for IRMS Data

The IRMS analysis methodology uses a mixture of five reference peaks (rectangular peaks, Fig. 1a) and 11 sample peaks (pointed peaks, Fig. 1a). Two sample peaks are discarded for quality purposes: the first is a procedural "injection" at 1 second (measured ~160 seconds later) to enable any possible gas from the previous analysis to be removed before new experiment subsampling, and the second is a true sample peak; however, many users note the potential for a small amount of the previous sample to affect the first sample injection (nanomolar). Therefore, we remove the first sample peak for all samples to treat all data uniformly. In summary, each analysis consists of nine sub-sampled peaks (blue box, Fig. 1) from which to construct IRMS features for ML.

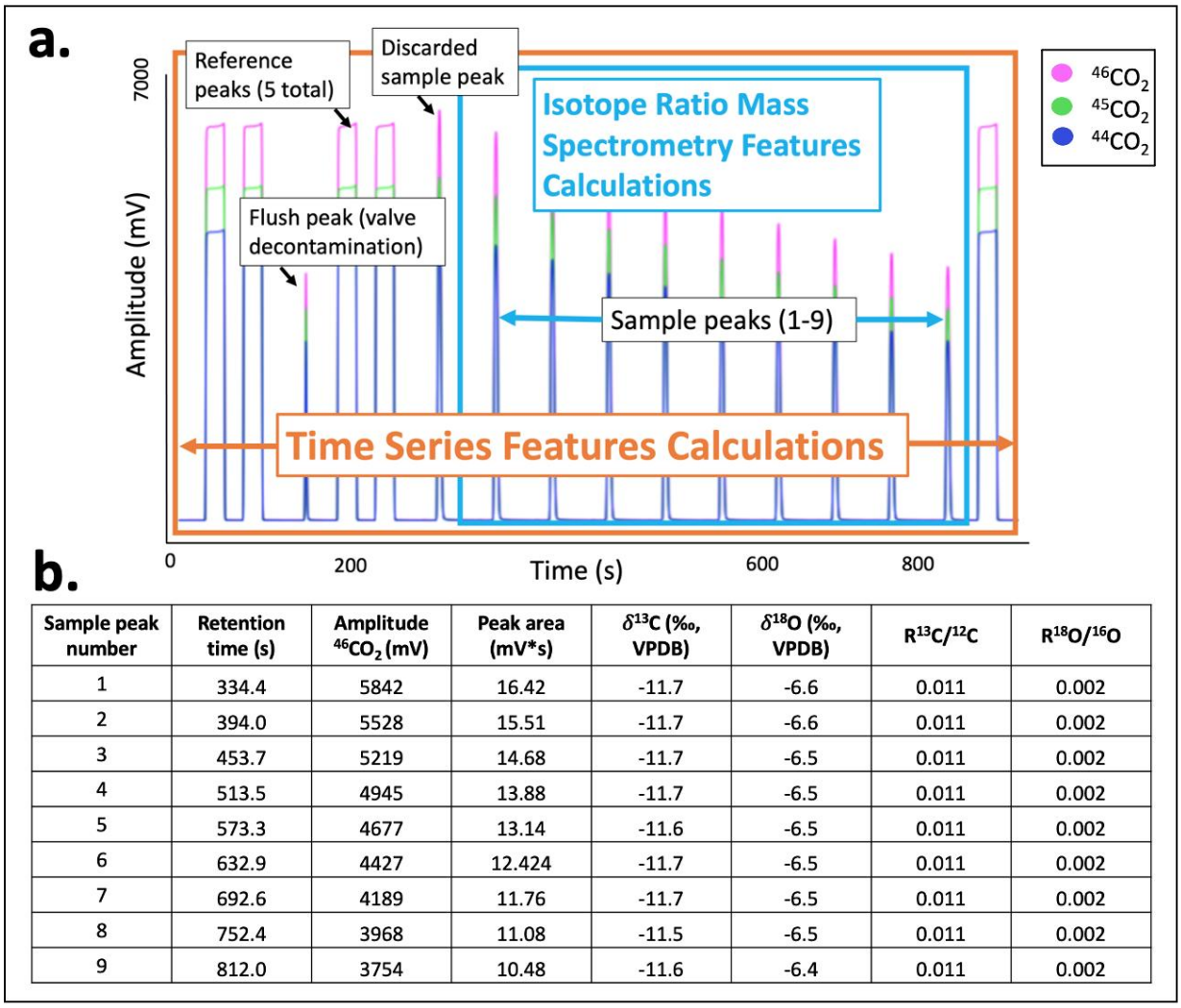
We use 33 Isodat©-output values (*e.g.*, isotope ratios, deltas, peak areas) to calculate IRMS-derived features for ML. Isotopic data are reported in delta (δ) notation, measured in per mil (‰) relative to VPDB (Vienna Peedee Belemnite) for both carbon and oxygen isotopes. For example, $\delta^{13}C_{CO_2(g)}$ is given by:

$$\delta^{13}C_{CO_2(g)} = \left(\frac{R_{CO_2(g)}}{R_{std}} - 1\right) \cdot 10^3, \quad (1)$$

where the factor of 1000 converts the value to per mil, $R_{CO_2(g)}$ is $[^{13}CO_2]/[^{12}CO_2]$ in the sample, and R_{std} is the same ratio for a well-defined standard (*e.g.*, VPDB). The delta formula for oxygen-16 and oxygen-18 isotopes is similar. Internal isotopic standards are analyzed with each batch of samples, which are calibrated to national and international standards VPDB and NBS-

18. The sample vial is measured multiple times, resulting in a type of longitudinal tabular data (Fig. 1b shows a few of the 33 variables). For each of the 33 IRMS variables, we compute two statistics across the nine repeated measurements: the mean and standard deviation. This results in 66 statistically derived IRMS variables per experiment. Examples of typical IRMS features are *avg_$\delta^{13}C$* and *sd_$\delta^{13}C$*, which are the average and standard deviation of the calculated carbon isotope value and are described in more detail in Sec. 2.3.

As TOF-MS chromatograms represent an amplitude in the time domain, we calculate additional features by treating the chromatogram as a time-series (TS). We refer to the combined set of extracted TS and IRMS features as the time-series mass spectrometry (TSMS) feature space. Some examples of TS features include *entropy* (*i.e.*, information entropy) and measures of autocorrelation (*i.e.*, self-similarity). The TS feature extraction is performed using the R library tsfeatures (Fulcher et al., 2013; Hyndman et al., 2015; Fulcher & Jones, 2017; Kang et al., 2017; Henderson & Fulcher, 2022), and the isoreader R library is used to read and process the stable isotope data (Kopf et al., 2021). Effective feature construction can increase the pool of important predictors for improved ML training. ML feature selection, discussed in Sec. 2.2.1, is used to identify which TSMS features are important.



Sample peak number	Retention time (s)	Amplitude $^{46}CO_2$ (mV)	Peak area (mV*s)	$\delta^{13}C$ (‰, VPDB)	$\delta^{18}O$ (‰, VPDB)	$R^{13}C/^{12}C$	$R^{18}O/^{16}O$
1	334.4	5842	16.42	-11.7	-6.6	0.011	0.002
2	394.0	5528	15.51	-11.7	-6.6	0.011	0.002
3	453.7	5219	14.68	-11.7	-6.5	0.011	0.002
4	513.5	4945	13.88	-11.7	-6.5	0.011	0.002
5	573.3	4677	13.14	-11.6	-6.5	0.011	0.002
6	632.9	4427	12.424	-11.7	-6.5	0.011	0.002
7	692.6	4189	11.76	-11.7	-6.5	0.011	0.002
8	752.4	3968	11.08	-11.5	-6.5	0.011	0.002
9	812.0	3754	10.48	-11.6	-6.4	0.011	0.002

Figure 1. (a) Typical isotope ratio mass spectrometry (IRMS) chromatogram of volatile CO_2, a plot of amplitude (mV) versus time in seconds (s). Rectangular peaks represent measurements of CO_2 reference gas of known isotope composition, called reference peaks. Pointed peaks represent repeated measurements of the experimental samples of CO_2 in the vial headspace, or sample peaks. Two sample peaks are discarded during data processing for quality purposes. (b) IRMS data from measurements representing the highest-quality sample peaks (blue box and arrows) are used to calculate IRMS features for ML, and a few of these values are shown to illustrate. For a single experiment, IRMS-output values must be treated to reduce the dimension due to the repeated measurements, resulting in variables such as $avg\_\delta^{13}C$ and $sd\_\delta^{13}C$. Another set of features is constructed by treating the chromatogram as a time-series (TS) (orange box and arrows, a) and calculating features that describe the entire spectrum. The combination of IRMS and TS features results in a feature space with more information to identify biosignatures.

2.2 Interpretable Machine Learning Approach

Here we provide an overview (Fig. 2) of the approach and contributing methods for training, validating, and interpreting the ML biosignature model. After experimental analyses and QA/QC, the BOW-δCO_2 dataset consists of 174 samples (63 biotic/111 abiotic) with 104 TSMS features (Fig. 2 steps I and II). Below we describe how NPDR feature selection is used with a Least Absolute Shrinkage and Selection Operator (LASSO or L_1) penalty (Tibshirani, 1997; Zou & Hastie, 2005; Hesterberg et al., 2008) and Unsupervised RF proximity distance (NPDR-LURF) on this dataset to identify important predictors that participate in statistical interactions for biosignatures while reducing correlation in the selected feature space (Fig. 2, step III). We describe our interaction network approach for visualizing how biosignature model features work together through statistical interactions and individually through main effects to influence model predictions, and we describe false prediction diagnostics based on single sample variable importance scores (Fig. 2, step IV and Sec. 2.2.1). Finally, we discuss the RF classifier used to accurately predict biosignatures, validated on test data (Fig. 2, step IV and Sec. 2.2.1).

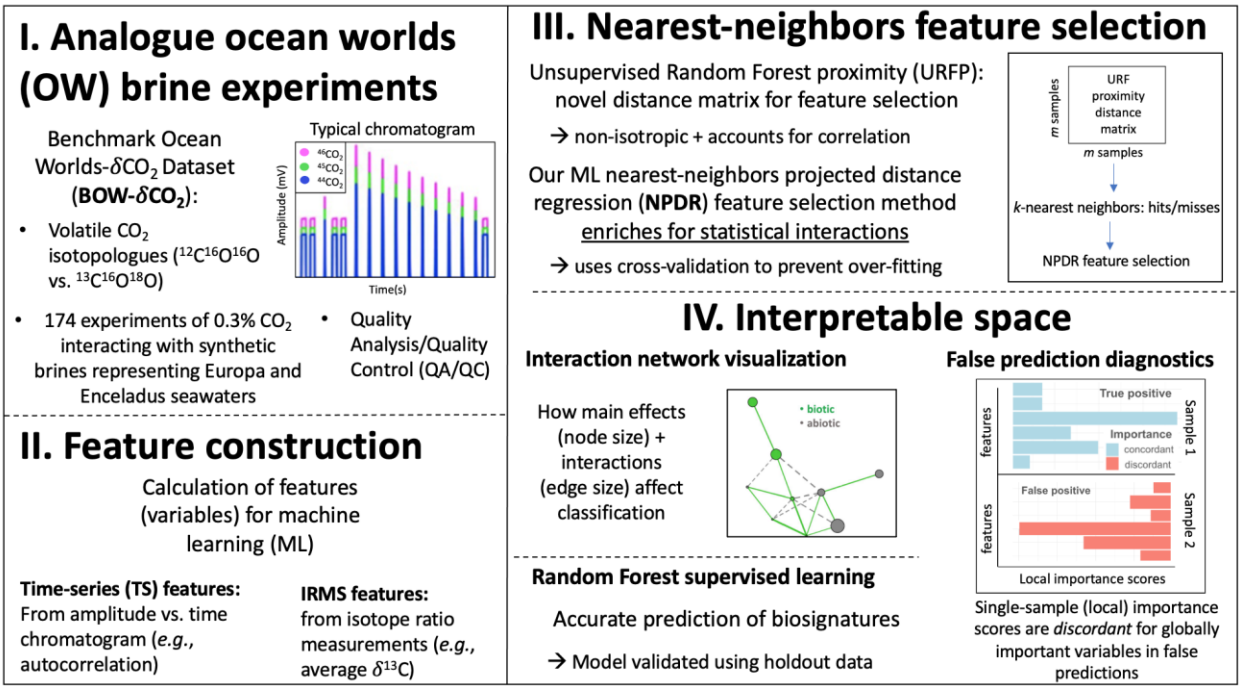


Figure 2. Overview of steps of our interpretable machine learning (ML) methods for biosignature prediction using our BOW-δCO_2 dataset. Step (box) I: Input data. We use 174 chromatograms of IRMS analyses of volatile CO_2 isotopologues interacting with synthetic brines representing OW compositions and perform Quality Analysis/Quality Control (QA/QC) to ensure only high-quality samples are used for ML applications. Step II: Feature construction. We then construct IRMS and TS features from the experimental data for use in predictive models, resulting in 104 variables. We perform feature selection to reduce the number of variables in our model and increase interpretability. Step III: Nearest-neighbors projected distance regression (NPDR); this identifies statistical interactions in high-dimensional data. Nearest-neighbors methods use a distance matrix to find hits/misses in a neighborhood of samples defined by the number of neighbors, *k*. NPDR uses cross-validation to prevent over-fitting and an unsupervised Random Forest proximity (URFP) distance. NPDR may be performed using a penalty that further reduces the feature space by removing highly correlated features, negating the need for this pre-processing step. Step IV: Interpretable space. We use interaction network visualizations to illustrate how features work together to affect model predictions and single-sample importance scores, which help with false prediction diagnostics. We use a Random Forest (RF) classifier, which is validated on holdout data to determine the accuracy of ML predictions of biosignatures.

2.2.1 Nearest-Neighbors Feature Selection: Predictors for Ocean World Biosignatures

ML feature selection adds interpretability to models by reducing the feature space (dimensionality) and providing context for model prediction mechanisms if the features are understandable (Guyon & Elisseeff, 2003). In the current study, we construct a feature space consisting of IRMS-derived variables such as $\delta^{13}C$ or peak area and add extracted TS features such as the relative entropy (time_kl_shift) (Fig. 2, step II). We use NPDR feature selection because of its ability to detect interactions between model variables (e.g., Fig 1b), which is due to NPDR's use of regression on nearest neighbors in the space of all predictor variables (Le et al., 2020). We use the LASSO penalty with NPDR (Fig. 2, step III) because it automatically eliminates irrelevant variables, reduces correlation, and enriches the remaining variable space for variation for classification by RF. The final biosignature we posit as a network of interacting TSMS features (see Fig. 6 in Sec. 3.2.2) with detectable patterns and interrelationships that the RF classifier uses to make predictions (Fig. 2, step IV); our interaction network biosignatures enable us to understand which characteristics of the mass spectral data (time, intensity, ratios, etc.) are the most crucial for accurately identifying biosignatures.

The choice of distance metric in NPDR is crucial for the model because it influences nearest neighbor samples (Dawkins et al., 2021). NPDR typically uses standard distance metrics for nearest-neighbor calculations that treat the multivariate space as isotropic, an assumption that may be violated in IRMS data. Specifically, the TSMS features may have different characteristic scales and non-Gaussian distributions as well as complicated correlation structure. Thus, we introduce Unsupervised Random Forest Proximity (URFP) distance in NPDR to account for correlation structure between variables when computing nearest-neighbors. In addition, URFP is robust to skewed variables and mixed data types (Breiman, 2001; Shi & Horvath, 2006).

To compute the URFP between samples, the unsupervised Random Forest (URF) algorithm generates synthetic data with the same distributional properties as the original data, but with randomized values. In this URF, the class variable is ignored, and the RF is trained to discriminate between randomized synthetic and original data (with real correlation structure) (Fig. 3a, step I). Given an RF model, a proximity between two samples can be estimated for each tree by counting the leaves shared in the path leading from the terminal node of the samples back to the root node. The similarities are averaged over all trees in the forest and inversely weighted by the height of each tree to give the proximity between pairs of instances (Fig. 3a, step II). While both real and synthetic data are used to construct the trees, distances are only computed between real samples since synthetic distances are not relevant for NPDR. The proximity between real samples for a given sample pair *ij* is transformed into a distance, $D_{ij}^{(rf)}$, by

$$D_{ij}^{(rf)} = \sqrt{1 - P_{ij}^{(rf)}}, \tag{2}$$

where $P_{ij}^{(rf)}$ is the URFP (Fig. 3a, step III).

Based on the URFP distance, we compute a matrix of *k*-nearest neighbors (knn) in the feature space (Fig. 3b, step I) for NPDR. We use a theoretical fixed-*k* neighborhood based on the expected number of neighbors in an α radius of the mean (Le et al., 2020):

$$k(m, \alpha) = \left\lfloor \frac{m-1}{2} \left(1 - \text{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right) \right\rfloor, \tag{3}$$

where *m* is the number of samples, α is the number of standard deviations away from the mean neighborhood radius, and erf is the error function. We set $\alpha = 0.5$, which has been shown to balance main effects and interactions during feature selection (Le et al., 2020), and set *m* equal to two times the minority class size to adapt for class imbalance. In our training data, the minority class size (biotic) is 51 samples, so we use $m = 2 \cdot 51$ in Eq. (3), resulting in *k* = 30 for the

training data. From the set of neighboring sample pairs *i* and *j* in neighborhood set $\mathcal{N}$, we compute a projected design matrix $d_{ij,a}$ for the set of predictors ($a = 1, \ldots, p$), and we compute a contrastive difference vector δ_{ij} for the biotic/abiotic class outcome (Fig. 3b, step II). The design matrix $d_{ij,a}$ has $k \cdot m$ rows (all neighbor pairs) and *p* columns representing the variables in the dataset. The contrastive difference vector δ_{ij} for a neighbor pair has a contrastive value of (0/1) if both samples i and j have the (same/different) class.

Similar to regular regression, the columns of the NPDR design matrix represent predictor variables. However, the NPDR rows are *projected distances* between pairs of samples, as opposed to *sample values* for regular regression. A projected distance is the distance between pairs of samples in neighborhood set $\mathcal{N}$ projected onto each variable. Effectively, the projected distance is the difference of the predictor variable values between two samples. As with regular regression, the NPDR formalism allows regularization like LASSO (Tibshirani, 1996; Zou & Hastie, 2005; Tibshirani et al., 2012) to reduce NPDR coefficients to zero. We find the NPDR coefficients $\vec{\beta} = (\beta_1, \beta_2, \ldots, \beta_\mathrm{p})$ that minimize the contrastive logistic classification loss function $\mathcal{L}$ with a LASSO (least absolute shrinkage and selection operator) ($\|\vec{\beta}\|_1$) penalty and hyperparameter, λ (Fig. 3b, step II):

$$\min_{\beta_o, \vec{\beta}} \sum_{ij \in N} \mathcal{L}\left(\delta_{ij}(y); \beta_0 + \vec{\beta} \cdot \vec{d}_{ij}(A)\right) + \lambda \|\vec{\beta}\|_1. \quad (4)$$

The $\beta's$ in the equation are model parameters used to fit the training data, and λ is a hyperparameter tuned prior to fitting the $\beta's$. Hyperparameters are parameters that control properties of the ML algorithm but are determined before final model training typically by cross validation (CV), a method for splitting the training data into folds to reduce error and prevent

overfitting (Parvandeh et al., 2020). Hyperparameters are fixed during final training of the classifier.

For example, given a random split of train/test data in BOW-δCO_2, NPDR feature selection with five-fold CV is performed to identify important features in each fold (ensuring these features generalize). The CV process also yields a value for hyperparameter λ that minimizes the classification error (while making sure the model is not overfitting). The hyperparameter value minimizes the NPDR penalized contrastive CV error, and the magnitude of the non-zero $\beta's$ in the resulting final model represent NPDR feature importance scores. Larger magnitude $\beta's$ indicate a feature is important for classification. The LASSO penalty shrinks the $\beta's$ to zero for features that are not important for classification. If a cluster of highly correlated features represents an important predictor, one variable in the group is retained and the $\beta's$ of the rest of the features are shrunk to zero.

In summary, the optimization of $\beta's$ is performed over sample pairs in the neighborhood set, determined by the URFP distance. We refer to this NPDR approach with the LASSO penalty and URFP distance as NPDR-LURF (Fig. 2, step III and Fig. 3). We combine NPDR-LURF feature selection (using five-fold CV) with supervised RF (using 5000 trees, an RF hyperparameter) over six random 80:20 train:test splits of the BOW-δCO_2 data to estimate an average train and test accuracy. The training data (89 abiotic samples and 51 biotic samples) and testing data (12 biotic and 22 abiotic test samples) are balanced according to the biotic/abiotic ratio in the full dataset.

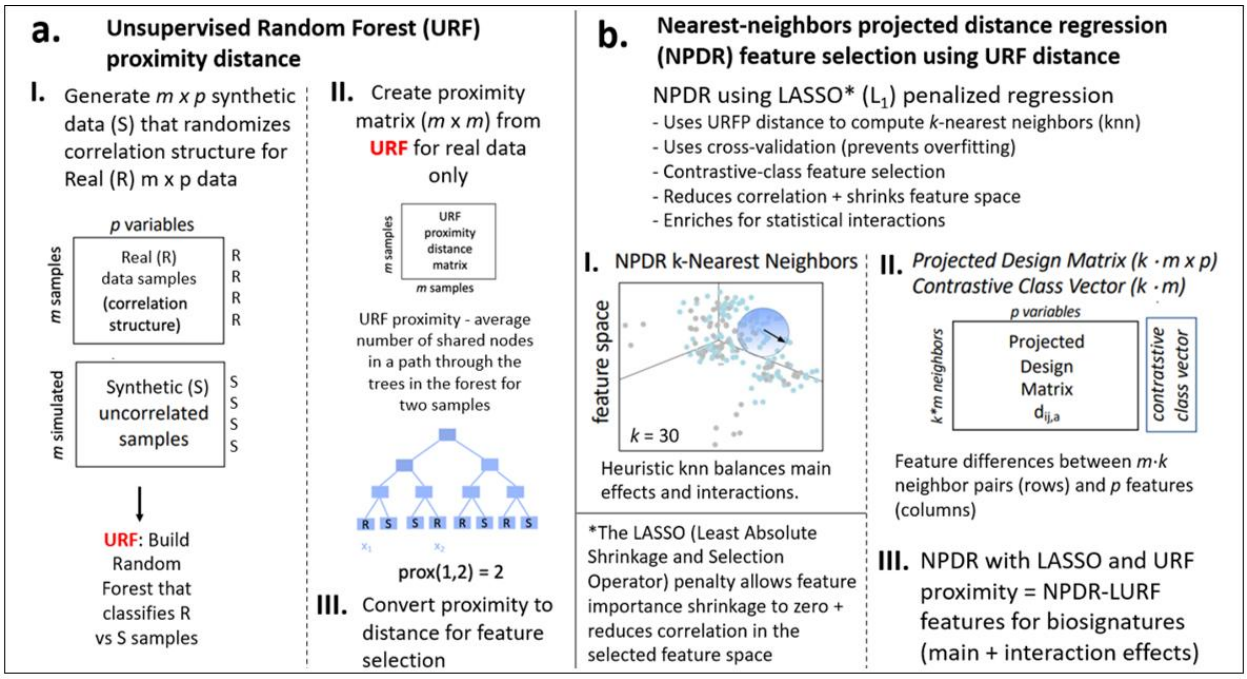


Figure 3. Overview of individual steps to perform NPDR-LURF feature selection: NPDR using a LASSO-penalty and URFP distance, described in **Fig. 2**, step III. (**a**) This study introduces the novel URFP distance for NPDR feature selection. Features constructed from experimental IRMS data ($m\ x\ p$) are used as input (training) data. URFP creates synthetic data of the same size as the original data ($m\ x\ p$). Unsupervised Random Forest (URF) classifies real versus synthetic samples (a, step I). The URFP matrix is returned for real samples, the proximity representing the number of nodes shared in a path through the trees in the forest for two given samples (a, step II). The URFP is converted to a distance for feature selection (a, step III) via Eq. (1). (**b**) NPDR-LURF feature selection allows a maximum shrinkage of the feature space and a maximum reduction in correlation while enriching for statistical interactions. We use the URFP distance to calculate *k*-nearest neighbors (KNN) (b, step I). NPDR creates a statistical projected design matrix of feature differences between $m \cdot k$ neighbor pairs (rows) and p features (columns). NPDR also creates a contrastive class vector containing zeros and ones depending on whether the neighbor pairs are in the same biotic class (b, step II). See methods for selection of k. NPDR-LURF returns input features for biosignatures based on contrastive classes (b, step III).

2.2.2 Network Interpretability

For additional model interpretation, we apply a Regression-based Association-Interaction Network (RAIN) to biosignature variables. The algorithm for RAIN was first described in the context of genetic and gene expression networks (McKinney et al., 2009; Lareau et al., 2015). Nodes in the RAIN are model features that are signed (positive/negative) and weighted by their

main effect on the class (*abiotic*/*biotic*). Edges are signed and weighted by the statistical interaction between the pair of nodes (See Sec. 2.3.3 for the biosignature RAIN).

2.3 Interpretable Machine Learning Biosignature Model Results

In this section we present the results of our interpretable ML approach for biosignature classification. We first discuss how ML is needed for the biosignature dataset because of the prevalence of biotic mimicry; that is, abiotic samples are so similar to biotic samples that no variable main effects are significant enough to separate the two classes. Then we present the results of NPDR-LURF feature selection for biosignature model predictors. Finally we discuss the results of our biosignature model trained on NPDR-LURF selected features and compare the performance of this model to RF classifiers trained using features selected by NPDR-Manhattan (NPDR feature selection using a Manhattan distance metric), the full variable space, and a variable space with highly correlated features removed.

2.3.1 Biotic Mimicry

In the BOW-δCO_2 dataset, when all salt compositions are pooled (Table 1, row 1), the means of $\delta^{13}C$ and $\delta^{18}O$ show almost no difference between biotic and abiotic samples (Fig. 4a). The overall mean of $\delta^{13}C$ for samples of all compositions is -12.8‰ (VPDB) and -5.2‰ (VPDB) for $\delta^{18}O$. Further, there is no statistically significant difference between biotic and abiotic even when samples are stratified by salt composition (Table 1, other rows). This indicates the need for features other than oxygen and carbon isotope fractionations and/or more complex models to discriminate between biotic and abiotic samples. Although one composition showed a significant P-value before adjustment, $\delta^{13}C$ for $NaHCO_3$_NaCl, it was insignificant after adjustment for multiple testing (Table 13, $NaHCO_3$_NaCl).

False discovery rate (FDR) adjusted P-values for t-tests of CO_2 isotope fractionations do not show a significant difference between *biotic* and *abiotic* samples (Table 1). FDR is a procedure that adjusts P-values to limit false positives due to multiple hypothesis testing (Shaffer, 1995). This lack of main effects illustrates the complexity of the abiotic samples, many of which contain non-biogenic organics and multiple carbon sources.

While CO_2 isotope fractionation is not significantly different for biotic classes for any salt composition, carbon and oxygen isotopes fractionate differently relative to analogue seawater composition (Theiling, 2021). For example, compare $\delta^{13}C$ for $MgCl_2$ and $MgSO_4$ brines, and $\delta^{18}O$ for $MgSO_4$_NaCl and NaCl brines (Table 1). Furthermore, CO_2 isotope fractionation for some salt compositions follow established models for the relationship between $\delta^{18}O$-ionic strength, while others do not, prompting the need for further experimental and modeling work (Theiling, 2021); KCl, and $MgCl_2$ brines fractionate oxygen isotopes following established models, while brines with $MgSO_4$, Na_2SO_4, and NaCl compositions deviate from established models.

Table 1

P-values and Adjusted P-values for t-Tests of Mean Difference of $\delta^{13}C$ and $\delta^{18}O$ Between Biotic and Abiotic Class Stratified by Salt Composition in BOW-δCO_2.

$\delta^{18}O$ (‰, VPDB) $\delta^{13}C$ (‰, VPDB)	Sample size (biotic/abiotic)	Overall mean (sd) $\delta^{13}C$	Overall mean (sd) $\delta^{18}O$	Biotic mean (sd) / abiotic mean (sd) $\delta^{13}C$	Biotic mean (sd) / abiotic mean (sd) $\delta^{18}O$	P-values $\delta^{13}C$	P-values $\delta^{18}O$	Adjusted P-values $\delta^{13}C$	Adjusted P-values $\delta^{18}O$
All compositions	174 (63/111)	-12.8 (1.1)	-5.2 (1.7)	-12.8 (1.2) / -12.7 (1.0)	-5.2 (1.6) / -5.3 (1.8)	0.60	0.67	0.66	0.79
NaCl	26 (12/14)	-12.4 (1.4)	-4.0 (0.7)	-12.4 (1.3) / -12.3 (1.3)	-4.0 (0.4) / -4.1 (1.0)	0.85	0.79	0.85	0.79
Na_2SO_4	17 (6/11)	-12.7 (1.1)	-4.5 (0.8)	-13.2 (0.6) / -12.4 (1.2)	-4.4 (0.4) / -4.5 (0.9)	0.08	0.73	0.30	0.79
$MgSO_4$	45 (18/27)	-11.9 (0.9)	-6.2 (0.7)	-11.7 (1.0) / -12.0 (0.8)	-6.1 (0.3) / -6.3 (0.8)	0.28	0.39	0.51	0.79
$NaHCO_3$	55 (18/37)	-13.0 (1.1)	-4.5 (1.0)	-12.8 (1.6) / -13.0 (0.9)	-4.6 (0.9) / -4.4 (1.1)	0.53	0.56	0.65	0.79
$MgSO_4$_$NaHCO_3$	33 (23/10)	-11.1 (0.7)	-6.0 (0.7)	-10.7 (0.5) / -11.5 (0.8)	-5.0 (0.1) / -6.3 (0.9)	0.11	0.29	0.30	0.79
KCl	13 (8/5)	-13.3 (0.5)	-4.1 (0.7)	-13.4 (0.5) / -13.2 (0.5)	-3.9 (0.3) / -4.3 (0.9)	0.42	0.42	0.58	0.79
$MgCl_2$	10 (6/4)	-13.7 (0.7)	-5.1 (1.2)	-14.1 (0.5) / -13.5 (0.7)	-4.6 (0.5) / -5.3 (1.4)	0.09	0.24	0.30	0.79
$MgSO_4$_NaCl	40 (32/8)	-12.0 (0.5)	-6.2 (0.7)	-11.9 (0.05) / -12.1 (0.7)	-6.3 (0.3) / -6.1 (0.8)	0.40	0.64	0.58	0.79
$NaHCO_3$_NaCl	72 (65/7)	-13.4 (0.5)	-3.9 (0.6)	-13.6 (0.2) / -13.3 (0.5)	-3.8 (0.07) / -3.9 (0.7)	0.04	0.43	0.30	0.79
no_salt	17 (11/6)	-13.0 (0.6)	-9.3 (0.4)	-13.3 (0.6) / -12.9 (0.6)	-9.2 (0.1) / -9.4 (0.4)	0.23	0.24	0.51	0.79

The overlap of distributions for $\delta^{13}C$ and $\delta^{18}O$ for biotic and abiotic samples (Fig. 4a) illustrates that our benchmark data include biotic mimicry and that methods for biosignature detection with IRMS-derived features cannot rely on main effects (univariate associations) alone. Rather, biosignature detection should account for statistical interactions (Fig. 4b), which occur when the correlation between features is conditioned on the biotic/abiotic class. A statistical interaction occurs in this dataset between *avg_rR*$^{45}CO_2/^{44}CO_2$, an IRMS feature, and *diff2_acf1*, a time-series (TS) feature (Fig. 4b). The dependence or correlation between these two variables is conditioned on the biotic/abiotic status; the correlation changes from a low 0.8% for abiotic samples to 21.5% for biotic samples. The presence of *diff2_acf1* in the interaction, which represents a measure of self-similarity between peaks in the TS, also illustrates the value of augmenting IRMS-derived features with mathematically constructed TS features that may contain additional variation and function as important predictors and interaction partners for biosignatures.

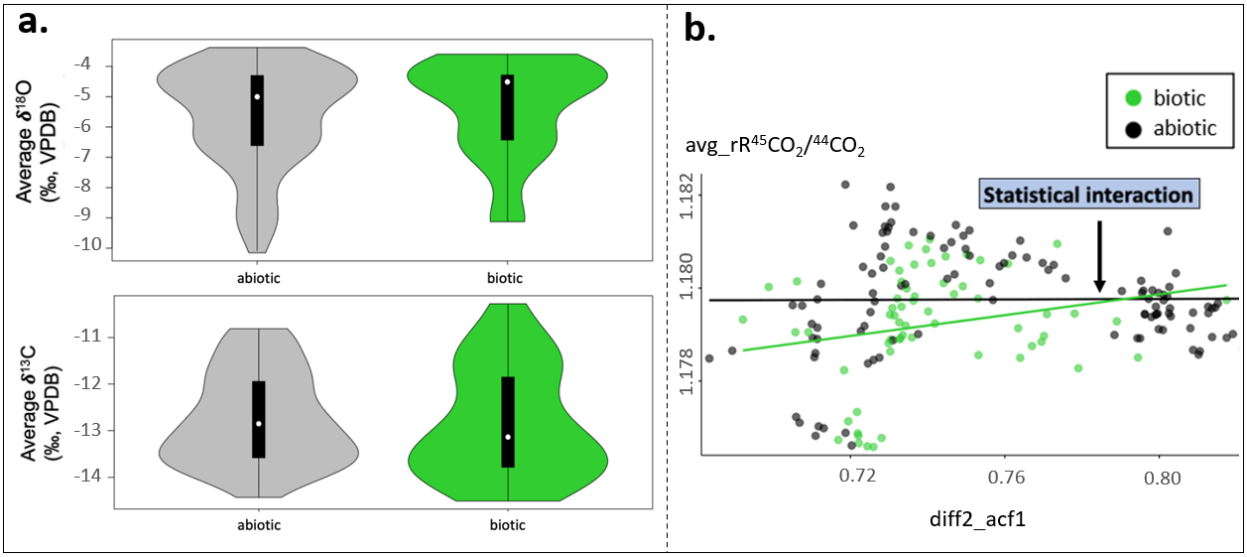


Figure 4. Illustration of the importance of statistical interactions for biosignature detection in IRMS data. (**a**) Violin plots of average $\delta^{18}O$ and $\delta^{18}O$ relative to biotic class show significant overlap, indicating that methods for isotopic biosignature detection must be able to go beyond main effects and detect interactions. (**b**) Scatter plot showing a statistical interaction between an IRMS feature representing $^{45}CO_2/^{44}CO_2$ detected by the instrument (*y*-axis) and a constructed TS feature, *diff2_acf1* (*x*-axis), a measure of self-similarity between peaks in the TS. The black regression line shows a very low positive correlation in abiotic samples, and the green line shows a higher positive correlation in biotic samples.

2.3.2 NPDR-LURF Features for Biosignatures

In the following, we describe the results of our interpretable ML approach to accurately classify biosignatures that is based on the detection of statistical interactions using NPDR-LURF. Our overall feature construction method yields 104 TSMS features (66 IRMS features and 48 TS) (Sec. 2.3.2). This feature space is reduced to on average six features (see Tables. A4-A10), and it is enriched for statistical interactions, as illustrated by our Regression-based Association-Interaction Network (RAIN, Sec. 2.3.3). We then show the results of the RF classifier using NPDR-LURF features to predict biosignatures. We use an 80:20 train:test split on the BOW-δCO_2 data, preserving class imbalance, repeated for six random splits. We present the results of

NPDR feature selection and RF classification for one typical run here. Results of the other random splits can be found in Appendix A4.

We use NPDR feature selection with LASSO penalty and URFP distance metric (NPDR-LURF) to construct an interpretable, parsimonious ML model for biosignature detection. We use hyperparameter penalty $\boldsymbol{\lambda = 2.16 \cdot 10^{-4}}$, chosen by CV. NPDR-LURF yields six predictors, enriched for statistical interactions and with reduced correlation compared with the top six ranked RF importance features (Fig. 5). The meaning of the NPDR-LURF features is discussed later in this section.

We kept highly correlated features in the data to test the ability of different feature selection methods to remove redundancy. The highest correlation between NPDR-LURF features is 53.5% and occurs between *avg_R*$^{45}CO_2/^{45}CO_2$ and *avg_rR*$^{45}CO_2/^{45}CO_2$ (Fig. 5a). The heatmap for the top six RF importance features, however, contains three blocks of high correlation (areas 1, 2, and 3, Fig. 5b) with correlation up to 100%. That is, RF importance selects several features that are so highly correlated they contain virtually no unique information. It would therefore be detrimental for the final model to include the top RF features because they contain so much redundant information the model may not generalize for different splits of the BOW-δCO_2 dataset. Given a correlation cluster, LASSO arbitrarily chooses a representative variable. Therefore, a variable that a human would consider more interpretable or functional might be eliminated by LASSO in favor of another correlated variable. However, the model will have roughly the same accuracy regardless of which variable is retained in a correlation cluster. Based

on these results, we recommend that correlation heat map clusters could be used to help identify more interpretable variables in a cluster.

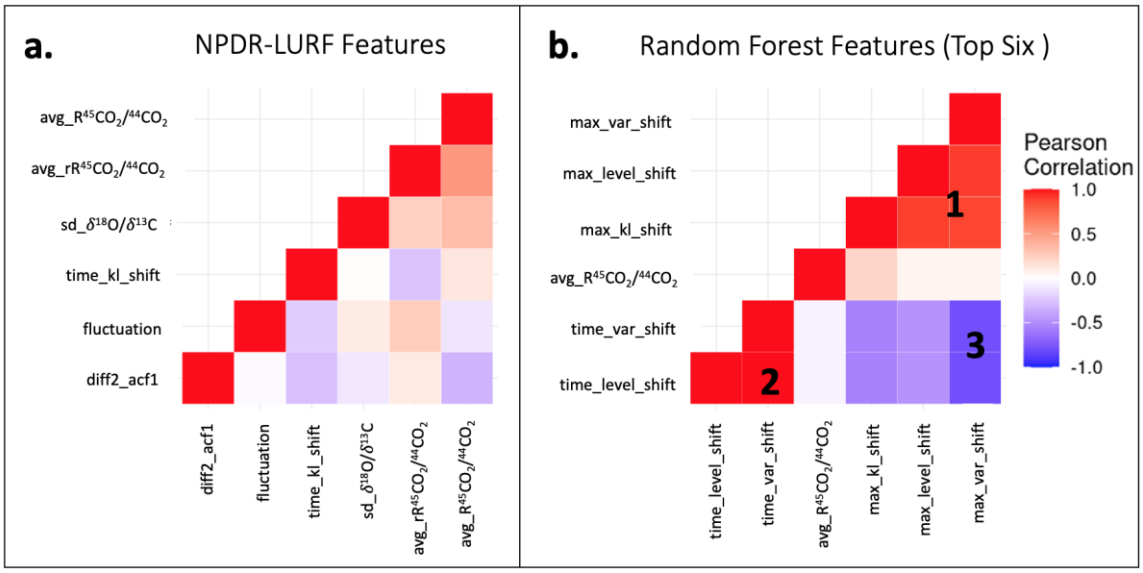


Figure 5. (**a**) Penalized feature selection reduces correlation in the selected feature space. NPDR-LURF yields six uncorrelated features that are an even mixture of TS and IRMS features. The highest amount of correlation is 53.5%, between *avg_$R^{45}CO_2/^{44}CO_2$ and avg_r$R^{45}CO_2/^{44}CO_2$*. (**b**) RF importance, in contrast, selects some highly correlated features for its top six important features. Areas of high correlation correspond to maximum correlations of 90.7% (area 1), 100.0% (area 2), and -79.7% (area 3).

The optimized NPDR-LURF feature space also provides computational benefits for training the model, an important consideration for future science autonomy software for spaceflight. We also find evidence that the URFP distance metric also reduces the effects of correlation between features, even without regression penalties, compared to standard distance metrics (Fig. A6). Additionally, NPDR-LURF features also outperformed feature spaces consisting of delta values for carbon and oxygen isotopes, illustrating the need to go beyond

typical instrument output and for methods that can detect statistical interactions (Figs. A9 and A10).

Here we summarize the six NPDR-LURF selected features for biosignatures on a typical train/test split in our dataset (Table 2). The top NPDR-LURF feature is *avg_*$R^{45}CO_2/^{44}CO_2$. The calculation of $R^{45}CO_2/^{44}CO_2$ employs both the known and measured reference gas isotope ratio to report an adjusted value for the sample peaks. This is different than what is measured as a raw detection, as in *avg_r*$R^{45}CO_2/^{44}CO_2$, the second-ranked NPDR-LURF feature. Although the first and second NPDR-LURF features both represent the ratio of $R^{45}CO_2/^{44}CO_2$, they are calculated differently and capture independent components of the variation about the biotic class. The third-most important NPDR-LURF feature for biosignatures is *sd_*$\delta^{18}O/\delta^{13}C$, the standard deviation of $\delta^{18}O/\delta^{13}C$. See Eq. 1 for calculations of deltas and "R" for CO_2 (Hoefs, 1973; Zeebe & Wolf-Gladrow, 2001). The TS feature *diff2_acf1*, the first autocorrelation coefficient of the twice-differenced time series, is the fourth ranked predictor for biosignatures (Fulcher et al., 2013; Fulcher & Jones, 2017; Kang et al., 2017). The next ranked feature *fluctuation*, calculated through the identification of TS fluctuations using the power spectral density and autocovariance (Talkner & Weber, 2000; Fulcher et al., 2013; Fulcher & Jones, 2017; Kang et al., 2017). The last NPDR-LURF predictor for biosignatures is the TS feature *time_kl_shift*, the time index of

the maximum shift in the Kullback-Leibler (KL) divergence, also called the relative entropy (Talkner & Weber, 2000; Fulcher et al., 2013; Fulcher & Jones, 2017; Kang et al., 2017).

Table 2

Nearest-neighbors Projected Distance Regression (NPDR) Selected Features for Biosignatures

Biosignature feature (NPDR-LURF)	Brief description	References
1. avg_R$^{45}CO_2$/$^{44}CO_2$	The average ratio of $^{45}CO_2$ to $^{44}CO_2$ for sample peaks in an IRMS experiment calculated from the recorded intensity of $^{45}CO_2$ on the instrument detector (mV*s), known reference gas ratios, and linear extrapolation between the reference peaks.	Hoefs 1973; Zeebe & Wolf-Gladrow 2001; Kopf et al. 2021
2. avg_r$R^{45}CO_2$/$^{44}CO_2$	The average ratio of $^{45}CO_2$ to $^{44}CO_2$ for sample peaks in an IRMS experiments as measured from the (r)ecorded intensities of $^{45}CO_2$ and $^{44}CO_2$ on the instrument detector in milliVolt*seconds (mVs).	Kopf et al. 2021
3. sd_$\delta^{18}O/\delta^{13}C$	Standard deviation of the ratio of avg_$\delta^{18}O$ to avg_$\delta^{13}C$ for sample peaks in an IRMS experiment.	Hoefs 1973; Zeebe and Wolf-Gladrow 2001
4. diff2_acf1	Time-series feature that is the first autocorrelation coefficient of the twice-differenced time series.	Fulcher et al. 2013; Fulcher & Jones, 2017; Kang et al. 2017
5. fluctuation	Time-series feature produced from a physics-informed fluctuation analysis where a first order polynomial is fit to the time-series and the range of root mean square fluctuations from the fit are returned (second order fluctuations).	Talkner & Weber 2000; Fulcher et al. 2013; Fulcher & Jones, 2017; Kang et al. 2017
6. time_kl_shift	Time index of the largest shift in the Kullback-Leibler (KL) divergence for a time-series, also called the relative entropy. KL divergence describes the difference between the probability distribution of the time-series and a reference probability distribution.	Fulcher et al. 2013; Fulcher & Jones, 2017; Kang et al. 2017

Note. Nearest-neighbors projected distance regression (NPDR) using LASSO-penalty with unsupervised Random Forest proximity distance (NPDR-LURF) yields six features for biosignatures, ranked in order of highest importance score.

2.3.3 Interpreting Feature Importance and Interaction

The NPDR-LURF selected feature space is small enough to be fully visualized by Regression-based Association-Interaction Network (RAIN), which is a statistical network method to visualize feature main effects and interactions (Lareau et al., 2015; McKinney et al., 2009). For interpretation, a manually-tuned threshold of 1.2 is used for interaction regression coefficients to determine the number of edges in the network (interactions with magnitudes below the threshold are not visualized). Nodes in the RAIN representation (Fig. 6a) are sized by their main effect magnitude on class and colored by the sign (positive or negative) of the

association direction (green for biotic and gray for abiotic). Edges are weighted by the magnitude of the statistical interaction between the nodes and colored by the direction of effect on class. A positive interaction indicates that the joint variation of the two variables increases the probability of a sample being biotic (edge colored green and solid), and a negative interaction indicates that the variation increases the probability of the sample being abiotic (edge colored gray and dashed).

Nodes 2, 3, 4, and 5 have the largest main effects, as indicated by node size (Fig. 6a). Moreover, the node numbers are ordered by NPDR-LURF importance. For nodes 2 and 4, the main effect increases the probability of an abiotic class prediction (gray nodes 2 and 4, Fig. 6a), while for nodes 3 and 5, the main effect increases the probability of a biotic prediction (green nodes 3 and 5, Fig. 6a). Nodes 2 and 4 (*avg_rR*$^{45}CO_2$/$^{44}CO_2$ and *diff2_acf1*) participate in a large interaction with each other that increases the chances of a biotic prediction (green edge between nodes 4 and 2, Fig. 6a). Additionally, node 4 (*diff2_acf1*) is a hub in the interaction network, participating in four interactions (see edges with nodes 2, 3, 5 and 6, Fig. 6a).

Fluctuation (green node 5**,** Fig. 6a) has the largest main effect and participates in a large interaction with *diff2_acf1* that informs abiotic prediction. In contrast, the highest-ranked node by NPDR-LURF (node 1) has the smallest main effect in the RAIN (Fig. 6a), while participating in two statistical interactions above the threshold. Recall that nodes 1 and 2 are related to the same isotope ratio, but they do not have extremely high correlation (0.535, heatmap Fig. 5a), and their combined variation facilitates classification. Node 1 weakly increases the abiotic probability by itself, but its interaction with nodes 2 and 6 reinforce the abiotic effect (Fig. 6a). The NPDR-LURF selected features are therefore enriched for statistical interactions, and the RAIN enables interpretation of the ML biosignature model through visualization of the interplay

of main effects and interactions. Consideration of such interactions between variables can provide crucial evidence and confidence for biosignatures in returned planetary data. In addition, single sample feature importance can help identify false biosignature predictions, which we explore in Chapter 3.

We chose a typical RF model for illustration (Fig. 6b and 6c). Models are trained using NPDR-LURF feature selection and RF. This typical model has a training accuracy of 87.9% (Fig. 6b), with eight false positives (abiotic samples predicted to be biotic) and nine false negatives (biotic samples predicted to be abiotic), corresponding to class errors of 8.0% for abiotic samples and 17.6% for biotic samples. The higher error for the biotic samples is explained by the class imbalance; more biotic samples in future work is expected to improve biotic class accuracy. The test accuracy is 88.2%, with two false positives and two false negatives for biosignatures (Fig. 6c).

To compare the effect of feature selection, we trained an RF classifier using the same training data after removing features that are >99% correlated, resulting in 52 predictors. This model performed closely to the NPDR-LURF model, with 86.4% training accuracy (nine false positives and ten false negatives). On the test data, this larger feature model performed with 91.2% test accuracy (three false positives). The NPDR-LURF features capture important information for biosignatures and perform with similar accuracy using 46 fewer predictors, resulting in a lightweight model with high accuracy.

To ensure the training and testing results were not biased, we repeated the above procedure for five random training and testing splits (still using an 80 train:20 test split) (see Sec. A.5 for features and classification tables). For the other five models, the NPDR-LURF train/test accuracies are consistently similar (mean test accuracy = 87.3%) to the RF classifiers in the

reduced correlation variable space (mean test accuracy = 89.2%), as well as consistently higher than RF classifiers trained using features selected using NPDR-LASSO with Manhattan distance (mean test accuracy = 82.4%), and higher than RF models trained using only delta values for oxygen and carbon isotopes (mean test accuracies for calibrated and uncalibrated values were 78.0% and 72.1%, respectively) (see Figs A.9-10). Our RF classifier for biosignatures is therefore accurate, interpretable, and parsimonious despite being trained on a relatively small sample size compared to features, which is addressed through machine learning feature selection with cross-validation (Hastie et al., 2009; Le et al., 2017; Parvandeh et al., 2020). Furthermore, this model is agnostic to seawater chemistry, meaning that the model can identify biosignatures in analogue OW seawaters of varying chemistries (see Table A.2 for brine compositions).

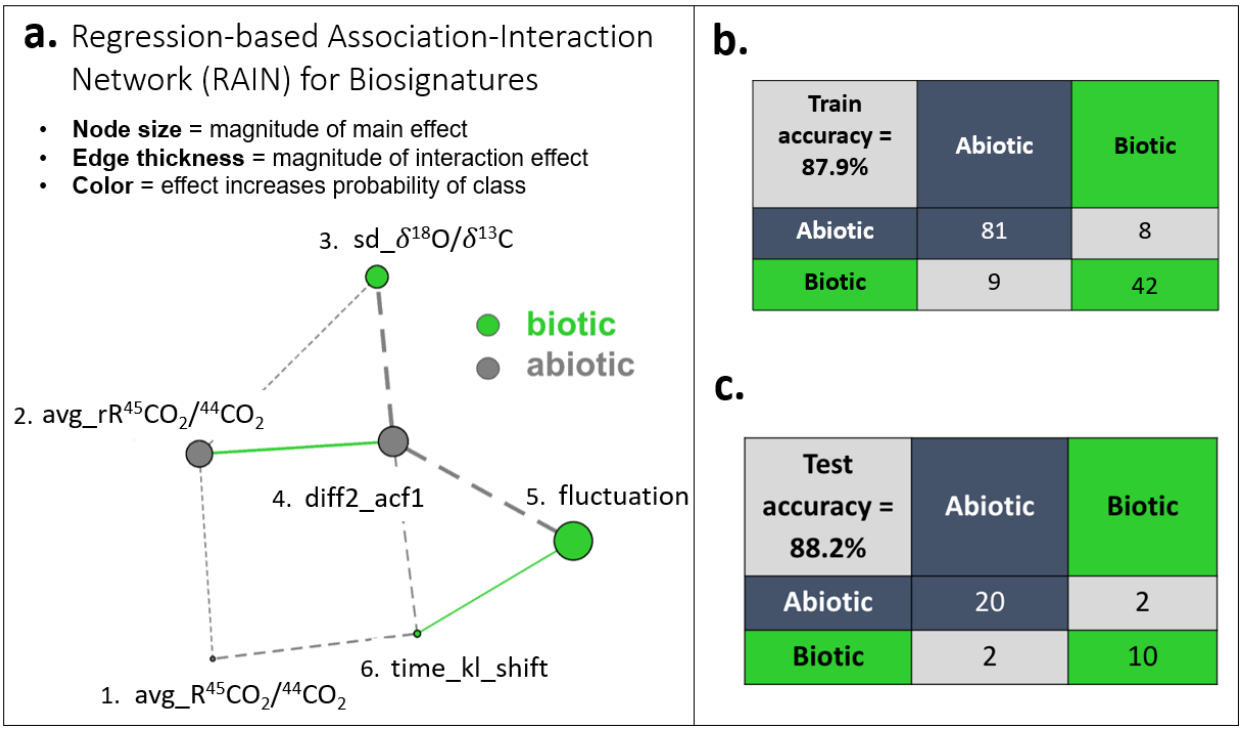


Figure 6. (**a**) Visualization of statistical effects of the highly predictive biosignature features using a Regression-based Association Interaction Network (RAIN) with node numbers in order of NPDR-LURF importance. RAIN encodes feature main effects and statistical interactions in a network. Nodes 2, 3, 4, and 5 have the largest main effects. For nodes 2 and 4, the main effect increases the probability of an abiotic class prediction. The TS feature *diff2_acff1* is also a hub in the interaction network, participating in three large interactions with nodes 2, 3, and 5 that influence the probability for both biotic and abiotic predictions. (**b**) The RF classifier trained on NPDR-LURF features for biosignatures yields a training accuracy of 87.9%, trained on a sample size $m = 140$. Dark grey and green diagonals represent the number of correct predictions: 81/89 abiotic samples are correctly classified, and 42/51 biotic samples are correctly classified. (**c**) The same classifier yields a high test accuracy of 88.2%: 20/22 abiotic samples are correctly classified, and 10/12 biotic samples are correctly classified.

2.4 Conclusions

Isotope fractionation is a powerful indicator of potential biosignatures on future missions to OWs in our solar system. However, abiotic environments with complex geochemistry can mimic biotic signals and increase the false positive rate of biotic prediction of samples, necessitating geochemically complex datasets that include biotic mimicry. In this study, we developed an interpretable ML approach for biosignature detection that can detect statistical interactions, disentangle biotic mimicry, and evaluate the potential for true/false positive and negative biosignature interpretations. To demonstrate this ML approach, we created a benchmark laboratory generated IRMS dataset of volatile CO_2 measurements for seawaters of OW analogue Europa and Enceladus compositions under biotic and abiotic conditions that could be used for future ML training. We developed NPDR-LURF feature selection methods using a novel distance metric, URFP, to create a variable space with physical and mathematical meaning for biosignatures based on IRMS and TS features. NPDR-LURF detected statistical interactions and reduced correlation/redundancy. We used statistical interaction networks to interpret the biosignature model.

In this work, the construction of a higher dimensional feature space was essential for ML models to have the necessary complexity for biosignature detection. The augmentation of IRMS-derived features with TS features provides RF with a rich variable space. However, RF importance has no statistical threshold to reduce the selected number of features, making this feature space too large for human interpretation (p = 104). Here we use NPDR-LURF to reduce this feature space to an average of six predictors, while removing correlated features and enriching for statistical interactions. The NPDR-LURF selected feature space is small enough to be analyzed and understood by researchers, adding interpretability for model prediction mechanisms. Furthermore, these features outperform features selected by both traditional

distance metrics and typically reported IRMS-output variables. Variable spaces in biosignature models for future astrobiology missions therefore need to be both highly reliable (generalizable) and informative for prediction interpretation.

The development of interpretable ML models is paramount to add confidence to detections with as far-reaching implications as extraterrestrial biosignatures on OWs. Even with a reduced variable space, the relationship between features can be complex. Thus, to better understand the statistical mechanisms by which the RF model uses main effects and interactions in TSMS selected features to make biosignature predictions, we used a RAIN (Fig. 6a**)**, which showed that the TS features add predictive information from IRMS spectra to the variable space. For example, *diff2_acf1* captures information about stationarity and self-similarity (see Table 1 for NPDR-LURF feature definitions) and is involved in interactions with many other features. In the context of the IRMS spectra as a time series, this TS variable is likely describing the stability of the peaks, since they are pre-programmed to be taken at specific times— that is, we expect the IRMS peaks to be highly stable, predictable, and self-similar, especially in the abiotic case. If they vary from this expected self-similarity, it could be due to biotic influence that causes variation between successive peaks on the timescale of the IRMS experiments. The main effect of this feature is therefore helping to identify abiotic samples, likely in brines with simpler geochemical conditions (since organics and complex geochemistry may be expected to decrease self-similarity depending on concentrations of species).

This type of interpretation also holds for *fluctuation* (whose main effect increases the probability of a biotic prediction). In the abiotic case, we expect few fluctuations in the peaks that represent random and unpredictable variation in the signal; the biotic samples may deviate from this expectation. We may also expect geochemically complex abiotic samples (ones with

multiple carbon sources or complex salt interactions) to display more random fluctuations detected by fluctuation (and the other TS features) due to biotic mimicry.

Two of the top NPDR-LURF features represent $R^{13}C/^{12}C$ (related to the abundance of CO_2 isotopologues), *avg_rR*$^{45}CO_2/^{44}CO_2$ and *avg_R*$^{45}CO_2/^{44}CO_2$. As noted previously, these features are not highly correlated despite representing the same isotope ratio (see Fig. 5a). These differences provide independent variation that the ML model can use for biosignature detection. This has implications for data interpretation and development of flight MS for astrobiology missions in terms of handling experimental measurements and constructing new variables to improve the ML detection of biosignatures by identifying time-dependent or mass and spectral-dependencies for biosignatures.

The effect *of sd_*$\delta^{18}O/\delta^{13}C$ can be understood by noting that different salts fractionate carbon and oxygen isotopes differently and noting that variables representing $R^{13}C/^{12}C$ are important for biotic prediction based on NPDR-LURF. This standard deviation of mixed deltas (*sd_*$\delta^{18}O/\delta^{13}C$) may therefore be capturing information about variation in isotope fractionation related to salt composition and biotic class. It has a large main effect in addition to participating in interactions with the first and second-most important features (see Fig. 6a).

Our BOW-δCO_2 dataset begins to address the concerns of deducing baseline OW subsurface chemistry from volatile isotope measurements of exospheres and plumes. Analogue OW brines display complex isotope fractionation for CO_2, indicating that understanding what to expect as an abiotic or prebiotic baseline under different OW conditions is a minimum requirement for the success of ML for extraterrestrial biosignature detection (Chou et al., 2021; Barge et al., 2022). Furthermore, an understanding of the isotope effects of geologic processes relevant for OWs should also be investigated (Theiling, 2021). Our plans for future laboratory

work will therefore investigate different and more complex geochemical environments. Future computational work will integrate ML with biogeochemical modeling to establish expected isotopic signals due to metabolic differences, incorporating bioinformatics, thermodynamics and kinetics methods. Such realistic modeling and simulation can also be used to better understand, interpret, and improve ML modeling for biosignature detection.

Despite using only six features, our NPDR-LURF RF classifier achieved high test accuracy (88.2%), comparable to that of an RF model trained on all 104 features (91.2%). The NPDR-LURF model may be a more desirable model for astrobiology missions because it has fewer noise features and redundancy with comparable accuracy to the full-feature model. The NPDR-LURF RF model importantly lends itself more to interpretation, having just six predictors.

Our NPDR-LURF RF classifier for biosignatures is therefore highly accurate, interpretable, and parsimonious. Furthermore, this model is agnostic to seawater chemistry and is lightweight in terms of memory and computational resources. Our proof-of-principle NPDR feature selection and high accuracy RF classifier, validated on holdout data, are a first step towards creating ML models with interpretable mechanisms and false prediction diagnostics for biosignature detection on future astrobiology missions to OWs with results immediately applicable to the conception of such missions.

In the next chapter, local feature importance methods are used to help diagnose false predictions in the biosignature data. Local-RF feature importance is compared with local-NPDR for simulated and the biosignature data.

CHAPTER 3

LOCAL-NPDR: A NEW INTERPRETABLE ML METHOD FOR SINGLE SAMPLES

Explainable ML is important for biosignature prediction on future astrobiology missions to OWs to minimize the risk of false positives due to geochemical biotic mimicry and false negatives due to environmental factors that mask biosignatures. ML models frequently use feature importance scores to provide insights into model prediction mechanisms by quantifying each variable's contribution to the prediction. Variable importance methods typically aggregate information globally across all training samples and therefore do not provide interpretation for the classification of a single sample. In contrast, local variable importance scores quantify the contribution of variables to the classification of a single sample and can therefore help explain why the sample was predicted to be in a certain class and diagnose whether it is a false prediction. We present a new local variable importance method that handles nonlinearity, statistical interactions, and includes penalized feature selection. Our approach represents a local version of Nearest-neighbor Projected Distance Regression (NPDR) feature selection. We evaluate local-NPDR on complex simulated data and real data from a study of carbon and oxygen isotopic biosignatures using laboratory-generated ocean world analogue brines. We use the concordance between global- and local-NPDR scores to diagnose classifier predictions, allowing the mechanisms of a true or false prediction to be explained. We illustrate the capacity of local-NPDR to integrate scientific explanations of single-sample ML predictions to support a more comprehensive framework for biosignature detection.

3.1 Local-RF Feature Importance

Local variable importance scores quantify variable contributions to the prediction of a single sample and can help explain why a sample is predicted to be in a certain class. Additionally, local feature importance methods can be used to diagnose whether sample predictions are likely true or false predictions, which is of obvious interest for ML applications for astrobiology.

The RF algorithm as implemented in the ranger R package (and the original Fortran) can return local variable importance scores for each sample in the training data (Breiman, 2001). RF uses a variable permutation change in accuracy while the sample is out-of-bag (oob) to determine both global and local feature importance. In global-RF variable permutation importance, the oob samples are fed into each tree of the forest to compute classification accuracy. By definition, the oob samples (about one-third of the training samples) are not seen by a particular tree during training (and varies depending on the tree in the forest). This accuracy calculation is repeated, but the order of the values for each variable is permuted in separate iterations. The change in average classification accuracy before and after permuting the variable is a measure of the variable's importance. Because permutation of an important variable is expected to decrease classification accuracy, the greater the decrease in accuracy after permutation, the more important the variable is considered (globally) for prediction.

The local-RF variable importance procedure also computes changes in accuracy before and after variable permutation but instead of permuting the variable for all oob samples, the value of each variable is permuted for a single sample and the sample is run through all trees in the forest *for which that sample is oob* to yield an average accuracy before and after variable permutation, the difference of which is the local RF variable importance for that sample.

3.2 Local-NPDR Feature Importance and False Prediction Diagnostics

In this section, we describe the local-NPDR algorithm and formalism in the context of global-NPDR feature selection along with an illustration of its use for diagnosing true and false predictions. Then, we describe the procedure for designating a prediction as likely true or false based on the total local scores (TLSs): a TLS may be concordant with the globally important features (positive) or discordant (negative). We hypothesize that samples with TLSs near zero or negative are more likely to be false predictions, since their feature importance scores do not match the global pattern for feature importance.

3.2.1 *Local-NPDR: Feature Importance for a Single Sample*

Consider a pair of samples or neighbors *i* and *j* that are distinct rows an *m* x *p* data matrix **X** with *m* samples and *p* variables. The class vector *y* has length *m*. To describe the NPDR contrastive loss, we use the contribution to the binary cross-entropy for a pair of neighbors given a set of regression coefficients represented by β,

$$\mathcal{L}_{ij}\left(\beta_o, \vec{\beta}\right) = -\delta_{ij}(y)\ln\left(\hat{d}_{ij}(X)\right) - \left(1-\delta_{ij}(y)\right)\ln\left(1-\hat{d}_{ij}(X)\right), \qquad (5)$$

where δ_{ij} is the hit/miss indicator variable and $\hat{d}_{ij}(X)$ is the predicted probability that the two samples are in different classes (*e.g.*, for the probability of a miss, $\delta_{ij} = 1$). The indicator variable can have two values: $\delta_{ij} = 1$ if the pair of samples are in a different class ($y_i \neq y_j$) and $\delta_{ij} = 0$ if they are in the same class ($y_i == y_j$). The predicted probability is computed using the following logit transformation

$$\hat{d}_{ij}(X) = \frac{1}{1+e^{-\left(\beta_0+\vec{\beta}\cdot\vec{d}_{ij}(X)\right)}} \qquad (6)$$

of the multivariate model of projected distances, $\vec{d}_{ij}(X)$, of all independent variables in X. In other words, for a fixed pair of *ij* neighbors, each element of the vector, $\vec{d}_{ij}(X)$, is an absolute difference between their values for each independent variable in X. We refer to these differences as projected distances onto a variable axis in the *p*-dimensional space. For example, if X were a numeric data matrix, the vector of projected distances would be

$$\vec{d}_{ij}(X) = \left(\left|X_{i1} - X_{j1}\right|, \left|X_{i2} - X_{j2}\right|, \cdots, \left|X_{ip} - X_{jp}\right|\right). \qquad (7)$$

The goal of local-NPDR is to find the variable importance scores ($\vec{\beta}$) that minimize the penalized negative log-likelihood (or cross entropy) over the neighborhood $N_k(i)$ of sample *i*

$$\beta_i^{local} = \min_{\beta_o, \vec{\beta}} \left(\sum_{j \in N_k(i)} \mathcal{L}_{ij}\left(\beta_o, \vec{\beta}\right) + \lambda \left(\alpha \left\| \vec{\beta} \right\|_1 + (1-\alpha) \left\| \vec{\beta} \right\|_2 \right) \right). \quad (8)$$

The penalty is implemented via the R library glmnet, which L_1 and L_2 regularization methods, or Ridge (L_2, $\alpha = 0$) and LASSO (L1, $\alpha = 1$) regression penalties in a method called the "elastic net" that effectively provides a feature importance ranking (Tibshirani, 1996). We typically use LASSO (Least Absolute Shrinkage and Selection Operator) for global-NPDR feature selection and tune λ via cross-validation (CV). This reduces the selected feature space and ensures variable independence. For local-NPDR, we typically employ a Ridge penalty because we have already reduced the feature space using global-NPDR and want the rankings of the selected features. The quantity $N_k(i)$ is the set of neighbors of sample *i*, and the resulting NPDR attribute scores, $\vec{\beta}^{local}$, are local to each sample *i*. The neighborhood is computed independently of the class status of samples and is defined using a distance matrix, discussed more below. These local variable importance scores indicate the importance of features that allow the single sample to discriminate whether neighbor samples are in the same or a different class as the target sample. If a variable were involved in an interaction, NPDR would reflect this in the importance score

because it uses nearest neighbors that are computed in the higher dimensional space of all other variables. This makes NPDR multivariate even when scoring a single variable for a single sample. The Ridge or LASSO version of NPDR includes additional multivariate effects in its model.

We illustrate how local-NPDR feature selection can add support to true predictions of an ML model (Fig. 7) and can help identify false positive predictions (Fig. 8). For both cases, we assume that the purple variable *A* on the vertical axis is important for classification. The importance can be determined, with varying confidence, using global-NPDR. For completeness, the global NPDR scores are computed by minimizing the following penalized cross entropy

$$\vec{\beta}^{global} = \min_{\beta_o,\vec{\beta}} \left(\sum_{i=1}^{m} \sum_{j \in N_k(i)} \mathcal{L}_{ij}(\beta_o, \vec{\beta}) + \lambda \left(\alpha \left\| \vec{\beta} \right\|_1 + (1-\alpha) \left\| \vec{\beta} \right\|_2 \right) \right), \quad (9)$$

which, in contrast to Eq. (8), includes the sum over all samples *i* from 1 to *m*.

The global importance for discrimination or classification of purple variable *A* (Fig. 7) can be seen by noticing that the mean for the 'x' class is larger than the mean for the 'o' class. Next, we consider how the globally important variable is affected locally in the local-NPDR contrastive loss (Eq. 5) for sample 1, when it is in the correct 'x' class (x_1 in Fig. 7). For illustration purposes, we estimate the contributions of the contrastive loss (Eq. 5) for variable *A* for Sample-1 using $k = 2$ neighbors, which, in this case, are Sample-2 (same class 'x') and Sample-3 (opposite class 'o'). The neighbor-pair loss L_{12} is low because the projected distance d_{12} is small (leading to a small $\hat{d}_{12}$ miss-probability) and their actual miss state is $\delta_{12} = 0$. That is, the quantity $-\ln\left(1 - \hat{d}_{12}(A)\right)$ will be a small positive loss (high accuracy), and the contribution to the local score from A would be relatively large. The neighbor-pair loss L_{13} is also low (high accuracy) because, while d_{13} is large, their actual miss state is $\delta_{13} = 1$. The

relevant part of the loss $-\delta_{13}\ln\left(\hat{d}_{13}(A)\right)$ will be a small positive quantity, and the contribution of neighbors Sample-1 and Sample-3 to the local score for variable *A* will be relatively large. This high importance score local to sample x_1 for the globally important variable A (concordant local and global scores) is supporting evidence that sample x_1 is a true positive. In contrast, if we incorrectly label Sample-1 ('o' instead of 'x' in Fig. 8), the neighbor losses will be high (low accuracy) and the importance of variable *A* local Sample-1 will be low, discordant with the global importance of *A*, and suggesting that Sample-1 might be a false prediction.

The quantity *k* in $N_k(i)$ is the number of nearest neighbors used for sample *i* in NPDR, sometimes referred to as *knn* (*k*-nearest neighbors). This number can vary from sample to sample or be uniform (same for all samples). For global-NPDR *knn*, we use k–D$\sigma_{1/2}$, which is the expected number of neighbors that are within ½ standard deviation of the mean distance (D$\sigma_{1/2}$) between all sample pairs. For local-NPDR, which focuses on only one sample, we use *knn-max* = *m* – 1 because it maximizes the statistical power by using all possible samples in the neighborhood of the single sample. The tradeoff is a decreased ability to detect statistical interactions: using *kin-max* causes NPDR and Relief-based methods to become myopic; that is, focused on the importance of single variables (McKinney et al., 2013; Dawkins & McKinney, 2025). Once an appropriate neighborhood is determined, the imbalance between hit and miss groups in the neighborhood of the sample is accounted for by regression model weights using the ratio 1 - *num_in_class*/*num_samples*.

The nearest neighbors are determined from a chosen distance metric in the full space of variables. For the current study, we employ a novel distance metric called the Unsupervised Random Forest Proximity (URFP), chosen due to its ability to account for a non-isotropic

variable space and its performance in the biosignature dataset compared with a traditional Manhattan distance metric (Clough et al., 2025).

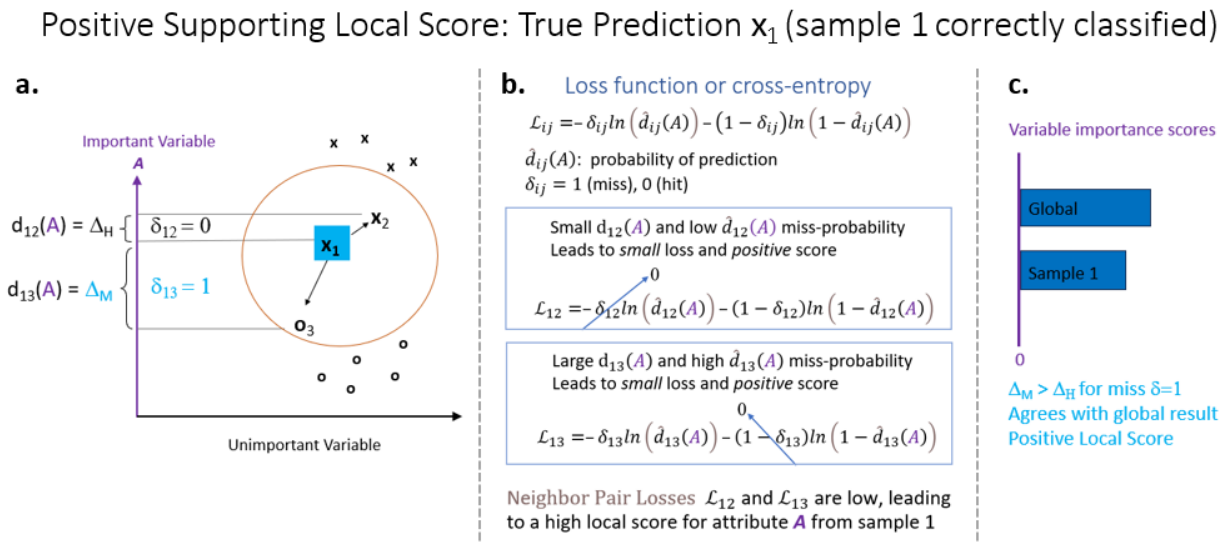


Figure 7. Local-NPDR mathematics for a correct classification with a positive (supporting) local feature importance score. (**a**) Hypothetical Sample 1 of Class *X* (blue highlighted $\mathbf{x}_1$) is plotted with two features, one simulated with a main effect for classification, *i.e.*, for discriminating between class *X* and *O* samples (*Variable A*, purple) and one unimportant variable with no effect. The nearest neighbors for Sample 1, indicated inside the orange neighborhood circle, are Sample 2 ($\mathbf{x}_2$, same class as Sample 1) and Sample 3 ($\mathbf{o}_3$, different class than Sample 1). The projected distances between Samples 1 and 2 for *Variable A*, $d_{12}(A)$, and Samples 1 and 3, $d_{13}(A)$, are indicated by Δ_H (hit) and Δ_M (miss), resulting in $\delta_{12} = 0$ (hit) and $\delta_{13} = 1$ (miss). The projected distances for these same samples onto the horizontal axis (unimportant variable) are negligible because this variable cannot discriminate between samples in Class *X* or Class *O*. (**b**) Local-NPDR loss function for the two nearest neighbors of Sample 1 (see Eq. 1). The total loss for *Variable A* (for Sample 1) is the sum of all pairwise loss functions for all local neighbors. The loss functions for two pairs of neighbors ($\mathcal{L}_{12}$ and $\mathcal{L}_{13}$) are small when Sample 1 is correctly classified and the δ's are correctly assigned as $\delta_{12} = 0$ (hit) and $\delta_{13} = 1$ (miss). (**c**) These low losses for the true classification of Sample 1 as Class *X* lead to positive local importance scores for important *Variable A*.

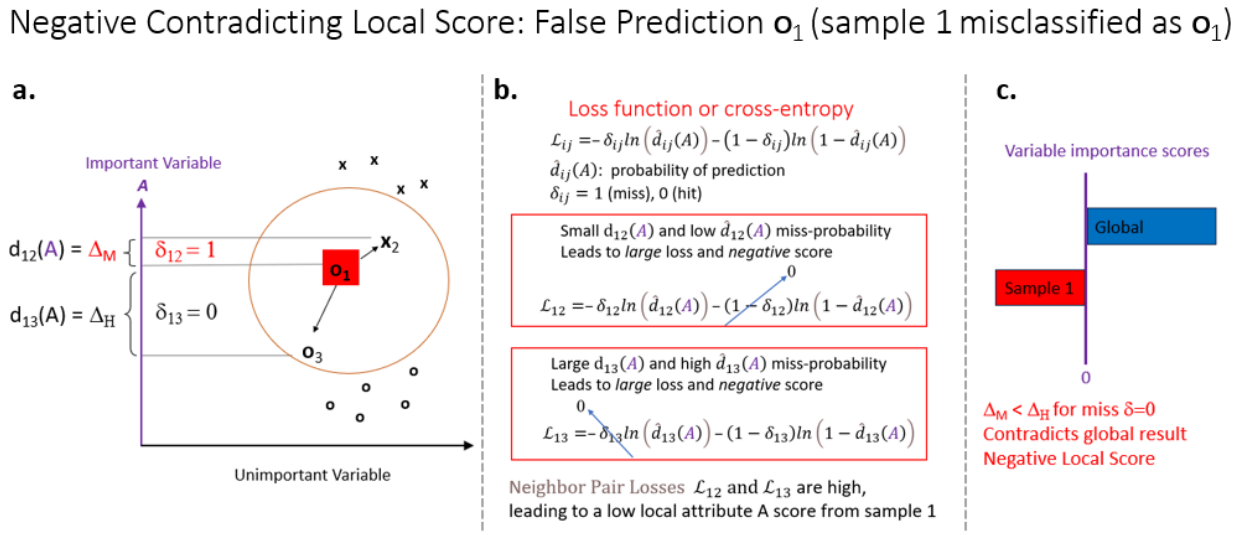


Figure 8. Local-NPDR mathematics for an incorrect classification with a negative (contradicting) local feature importance score. (**a**) Consider hypothetical Sample 1 of Class *X* that is incorrectly assigned Class *O* (red highlighted $\mathbf{o}_1$) and two features, one simulated with a main effect for classification (*Variable A*, purple) and one unimportant variable with no effect. The two nearest neighbors for Sample 1, indicated by the neighborhood circle, are Sample 2 ($\mathbf{x}_2$, different class as Sample 1 prediction) and Sample 3 ($\mathbf{o}_3$, same class as Sample 1 prediction). The projected distances between Samples 1 and 2 for *Variable A*, $d_{12}(A)$, and Samples 1 and 3, $d_{13}(A)$, are indicated by Δ_M (miss) and Δ_H (hit), resulting in $\delta_{12} = 1$ (miss) and $\delta_{13} = 0$ (hit). (**b**) Local-NPDR loss function for the two nearest neighbors of Sample 1 (see Eq. 1). Note the total loss for *Variable A* (for Sample 1) is the sum of all pairwise loss functions for all local neighbors. The loss functions for two pairs of neighbors ($\mathcal{L}_{12}$ *and* $\mathcal{L}_{13}$) are large when Sample 1 is incorrectly classified and the δ's are incorrectly assigned as $\delta_{12} = 1$ (miss) and $\delta_{13} = 0$ (hit). (**c**) These large losses for the false classification of Sample 1 as Class *O* lead to negative local importance scores for *Variable A*.

3.2.2 *Procedure for Reporting False Predictions and Validation*

We use NPDR to determine local and global feature importance. Because NPDR is a contrastive method, it predicts the class difference of neighbors, not the class of a given sample. To predict the class of individual samples, we use RF classification because of its robustness to skewed variables and mixed data types and its resistance to over-fitting. For each sample, we compute the local-NPDR importance scores for the features that were selected globally by

LASSO-NPDR using the URFP distance metric; these are the variables on which the RF classifier is trained, ensuring that the feature selection method is independent from the classification method. In this step, a new URFP distance metric using only the global-NPDR features is used. The local-NPDR variable importance scores can be concordant with the global-NPDR scores (manifested as positive variable importance scores) or discordant (negative importance scores). If the sum of the local scores is negative (overall discordant), the sample was likely classified based on variables that were not part of the general (global) pattern of the classifier. We hypothesize that such samples are more likely to be false predictions because they do not follow the general pattern learned by the classifier from the global dataset. We combine local-NPDR feature importance scores with RF prediction probabilities to further constrain which samples are identified as potential false predictions, hypothesizing that samples classified with lower prediction probabilities are more likely to be incorrectly classified.

We further compare false prediction diagnosis of individual samples in holdout data using local-NPDR variable importance with local-RF. We compute the overall local variable importance scores for correctly and incorrectly classified samples (based on the RF classifier) to see whether discordance is associated with false predictions. An initial question is which globally important variables to include in the concordance calculation. NPDR can use a LASSO penalty that results in a statistical threshold for importance. However, RF does not have a threshold for feature selection. Thus, we use the global-NPDR features as the RF model variables to determine local-RF feature importance. NPDR feature selection thresholds can be defined either through P-values or via regularization.

We validate the local-NPDR variable importance method on both real and simulated datasets. RF classifiers and global-NPDR feature selection for all datasets are trained using an

80:20 train:test split that preserves the class imbalance. RF hyperparameters are tuned using 5-fold CV. Simulated datasets allow us to compare effects of variable correlation, main and interaction effects, and class imbalance on ML models. Furthermore, since it is known whether variables in the simulated datasets are functional (*i.e.*, whether they have a main and/or an interaction effect), we can quantify the performance of our methods. Real datasets ensure that our methods work in real applications on imperfect or complex data.

The real dataset of interest for this study is the biosignature Benchmark Ocean Worlds-δCO_2 dataset (BOW-δCO_2), consisting of isotope ratio mass spectrometry (IRMS) measurements of volatile CO_2 evolved from laboratory-generated ocean world (OW) analogue brines of *biotic* and *abiotic* samples (Clough et al., 2025). This dataset contains 174 samples of IRMS experiments (111 *abiotic* and 63 *biotic*), generated with 0.3% CO_2 by volume and containing different salt compositions relevant for both Europa and Enceladus. The imbalance in this dataset is 0.64, with *biotic* samples making up the minority class.

We generate three simulated datasets using the `createSimulation2` function from our NPDR R library. These simulated data have the advantage of having known (ground truth) functional features (i.e., features associated with the outcome variable) while incorporating realistic effects found in real data. The first two simulated datasets are designed to have similar properties to the real BOW-δCO_2 dataset. Like our real BOW-δCO_2 dataset, these two simulated datasets have similar dimensions and class imbalance (m = 300 samples, p = 100 variables, class imbalance = 0.6), a realistic correlation structure between variables, and includes both interaction and main effects. We simulated 20% of the features to be functional, with 10 main effects ("mainvars") and 10 interaction variables ("intvars"). These two simulations have effect sizes of 1.5 for both main effects and interaction effects. The remaining features are noise variables that

have no effect on classification outcome. Since the main and interaction effects are known for particular variables, this allows us to assess whether feature selection methods are selecting relevant variables for classification. We simulate a third dataset with larger sample size and balanced classes ($m = 500$ samples, $p = 100$ variables). This dataset has the same number of main effects and interactions as the other two but has smaller main effect sizes (main effects = 0.8, interaction effects = 1.5).

3.3 Global-NPDR and RF Classifier Training

Before performing local-NPDR on individual samples, we first perform global-NPDR using all training samples and train RF classification models for the real biosignature data and the three simulated datasets. For global-NPDR feature selection, we use a LASSO penalty (see Eqn. 9 in Sec. 3.2.1) and URFP distance (global-NPDR-LURF) for training data splits. We train RF classifiers with tuned hyperparameters in the selected-feature spaces. For all datasets, we use weights to compensate for class imbalance in RF classifier training where *class_weights* = *1/num_samples*.

For the biosignature training data (m = 140 samples, 89 *abiotic* and 51 *biotic*), global-NPDR with hyperparameter $\lambda = 0.01$ results in five selected features (Table 3) out of 104 total predictors. Using these five features, the RF classifier with tuned hyperparameters *mtry* = 5, *splitrule* = "extratrees", *min.node.size* = 7, and *ntrees* = 5000 yields a training accuracy of 90.7% and a test accuracy of 91.2% (Fig. 9a). The accuracy breakdown by class (*biotic*/*abiotic*) shows high prediction accuracies for both classes in the train and test data alike for the biosignature dataset, despite the class imbalance. The *abiotic* class accuracy in the training data is 91.0% and

in the test data it is 95.5%. For the *biotic* class, in the training data the RF prediction accuracy using NPDR-LURF features is 90.2% and in the test data it is 83.3%, slightly lower.

In the three simulated datasets, global-NPDR selected nine, eight, and ten features out of 100 (Table 3) with hyperparameter λ = {0.02, 0.013, and 0.01}. The simulated datasets include functional features, which begin with "main" and "int" for main effects and interactions, respectively (Table 1). Simulated features that begin with "var" are noise features and are not functional. The two simulated datasets that contain two noise variables each are the imbalanced datasets, and the class-balanced dataset with more samples yielded no noise variables from global-NPDR.

The resulting RF training (test) accuracies for the simulated data are 77.9% (71.7%), 82.9% (78.3%), and 84.3% (80.0%), respectively (Fig. 9 b-d). The dataset with the highest accuracy (Fig. 9d) is balanced between classes and has a higher sample size. In addition, main effects play a more prominent role in feature selection (Table 3, last column). For the respective simulated data, the tuned RF hyperparameters for the simulated datasets were: *mtry* = {5, 8, 2}, *splitrule* = {"gini", "extratrees", "extratrees"}, *min.node.size* = {12, 3, 7}, and *ntrees* = {5000, 6000, 6000}. The two imbalanced simulated datasets show a discrepancy in class accuracy that is most notable in the test data (Fig. 9b and c), while the balanced simulated dataset shows a more balanced RF class prediction accuracy in both the train and test data (Fig. 9d).

Table 3.

Global-NPDR-LURF selected features for real biosignature data and three simulated datasets. Biosignature features include IRMS and time-series derived features. Simulated data include functional features, which begin with "main" and "int" for main effects and interactions, respectively. Simulated features that begin with "var" are not functional. Model accuracy details provided in Fig. 9.

Global-NPDR-LURF importance ranking	**Biosignature data: 91.0% Train Accuracy**	**Simulated data: 77.9% RF Train Accuracy**	**Simulated data: 82.9% RF Train Accuracy**	**Simulated data: 84.3% RF Train Accuracy**
1.	avg_r$R^{45}CO_2/^{44}CO_2$	mainvar9	mainvar8	mainvar5
2.	sd_$\boldsymbol{\delta}^{18}O/\boldsymbol{\delta}^{13}C$	mainvar4	intvar8	mainvar9
3.	diff2_acf1	mainvar1	mainvar9	mainvar1
4.	fluctuation	intvar8	mainvar7	mainvar4
5.	time_kl_shift	intvar3	var14	mainvar7
6.	-	mainvar8	var64	mainvar10
7.	-	mainvar2	intvar7	mainvar3
8.	-	var64	var6	mainvar8
9.	-	var35	-	intvar4
10.	-	-	-	mainvar6

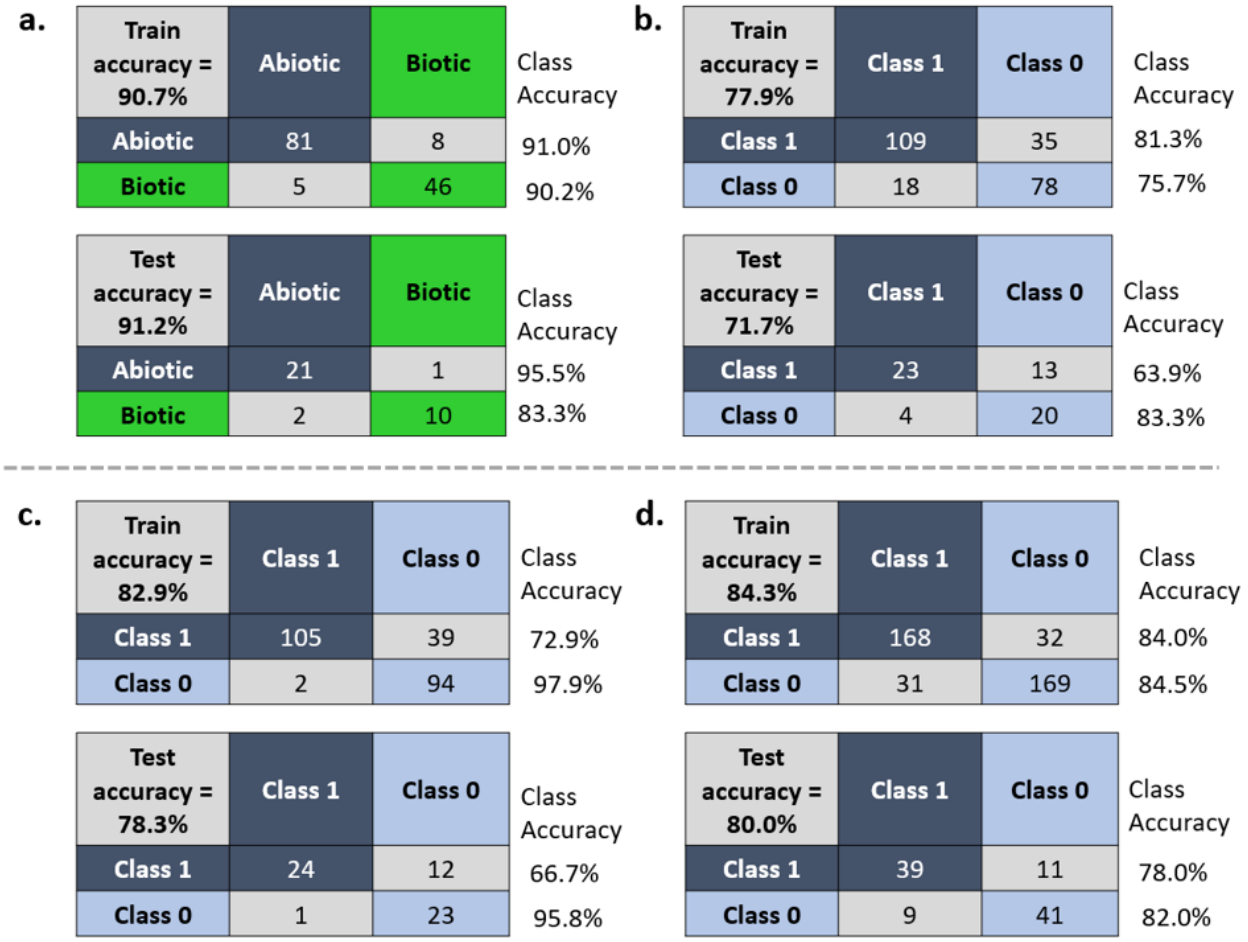


Figure 9. Random Forest (RF) train and test accuracies for real biosignature data and simulated datasets. (**a**) The RF classifier for biosignatures yields a 90.7% training accuracy. There are 51 *biotic* samples and 89 *abiotic* samples in the training data; five *biotic* samples are misclassified as *abiotic* (false negatives) and eight *abiotic* sample are predicted to be *biotic* (false positives). The biosignature train and test data show a similar high-accuracy performance despite class imbalance (class imbalance = 0.64), where the overall test accuracy is 91.2%. (**b**) This imbalanced simulated dataset (class imbalance = 0.6) contains 144 class *1* training samples and 96 class *0* training samples, yielding an overall training accuracy of 77.9%. This dataset shows a more balanced class accuracy in the training data than the testing data. (**c**) This imbalanced simulated dataset (class imbalance = 0.6) contains the same class breakdown as the dataset in (b) and shows similar behavior in terms of class accuracy imbalance in the test data but is even more pronounced. (**d**) This balanced simulated dataset contains more samples (the training data contains 200 class *0* and class *1* samples each) and balanced class accuracies in both train and test data.

3.4 Local-NPDR Analysis

In the following sections we present the results of our local-NPDR feature importance method to discriminate between true and false ML predictions in the three simulated datasets as well as the real biosignature BOW-δCO_2 dataset.

3.4.1 *Local-NPDR Feature Importance for Simulated Data*

For each sample in the train and test data, we calculate local-NPDR feature importance scores using a Ridge penalty, "lambda.1se" hyperparameter, and URFP distance based on the set of global-NPDR-LURF features. Results from the test data are discussed here; see Appendix B.3 for training data results. For each sample, the total local-NPDR variable importance scores are computed (for the globally important features), and we determined whether the sample is correctly or incorrectly classified by the RF model.

We then perform a t-test for the total local-NPDR variable importance scores between true and false predictions. For the three simulated datasets, the total local-NPDR scores are higher in the true versus false prediction groups for both training (Appendix Fig. B.3) and test samples (Fig. 10). The elevated total local score (TLS) in true versus false groups in the test data is statistically significant (with P-values: $4.6 \cdot 10^{-6}$ to $7.9 \cdot 10^{-12}$). However, there is the potential for score overlap between individual true and false predictions, especially in imbalanced datasets. For this reason, it can be beneficial to incorporate other information for diagnosing sample predictions, such as the RF classifier probability.

True and false predictions can be further broken down into true positive/true negatives and false positives/false negatives. For the simulated datasets, class *0* is taken to be the positive class. For the imbalanced datasets, this corresponds to the minority class, chosen to mimic the

study design for the biosignature data. This means the true positive (TP) and false negative (FN) predictions involve the minority class *0* for the two imbalanced simulated datasets, while true negative (TN) and false positive (FP) predictions involve samples of majority class *1*.

Mean total local-NPDR variable importance scores for the imbalanced simulated datasets show different values for false negative versus false positive predictions in both the train (Appendix Fig. B.4) and test data (Fig.10a and b). In both cases, the FP group, composed of class *1* samples incorrectly predicted to be class *0,* has higher mean TLS than the FN group, made of class *0* samples incorrectly predicted to be class *1*. This effect is likely due to class imbalance, and is absent in the balanced simulated dataset (compare Fig. 10a and b with Fig. 10c), suggesting that local importance methods may be less reliable in the presence of imbalance, which is also a perennial challenge for classification methods. During RF classifier training of imbalanced datasets, the majority class may be penalized, and the algorithm attempts to maximize the classification accuracy of the minority class. This results in a higher classification error for the majority class.

However, the RF prediction probabilities can help differentiate the true and false predictions in these cases (Fig. 10d and e), where the average probabilities for false predictions are lower than those for true predictions. For the imbalanced datasets, the mean RF prediction probability for true positives (~70%), representing samples of the minority class, is lower than for true negatives (>90%), samples of the majority class.

The balanced simulated dataset is less prone to the discrepancies in false prediction mean total local-NPDR variable importance scores (Fig. 10c). In this case, both the FP and FN predictions have similarly low mean total local-NPDR importance scores and both the TP and TN predictions have much higher scores. For this dataset, the average RF prediction probabilities

are also more balanced among prediction types, with false prediction probabilities both appearing at ~65% or below and mean true prediction probabilities both being ~75% (Fig. 10e). This average probability being lower than the probability for the majority class in the imbalanced simulated datasets could be related to the lower sample size.

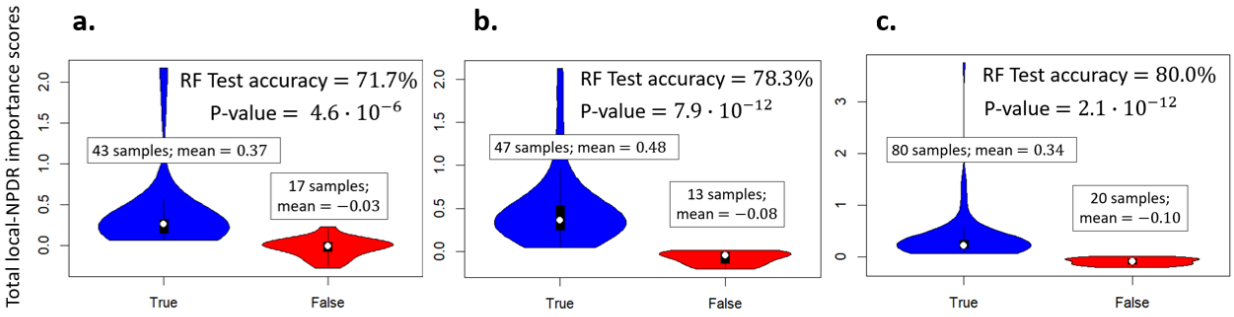


Figure 10. Total local-NPDR variable importance scores for true and false predictions in three simulated test (holdout) datasets. In each dataset, the local-NPDR scores are higher in the true (blue) versus false (red) prediction samples (all t-tests statistically significant). (**a**) This class-imbalanced simulated dataset has 60 test samples; 43 were correctly classified by the RF classifier trained in the global-NPDR selected feature space, yielding a test accuracy of 71.7%. The mean total local-NPDR feature importance scores are higher in the true prediction group ($P=4.6 \cdot 10^{-6}$). (**b**) The mean local-NPDR variable importance scores for this imbalanced dataset is higher in the true prediction group ($P = 7.9 \cdot 10^{-12}$). The RF classification test accuracy is 78.3% on 60 test samples. (**c**) This simulated dataset has a larger sample size, and the classes are balanced. Local-NPDR importance scores are also higher for true predictions ($P = 2.1 \cdot 10^{-12}$).

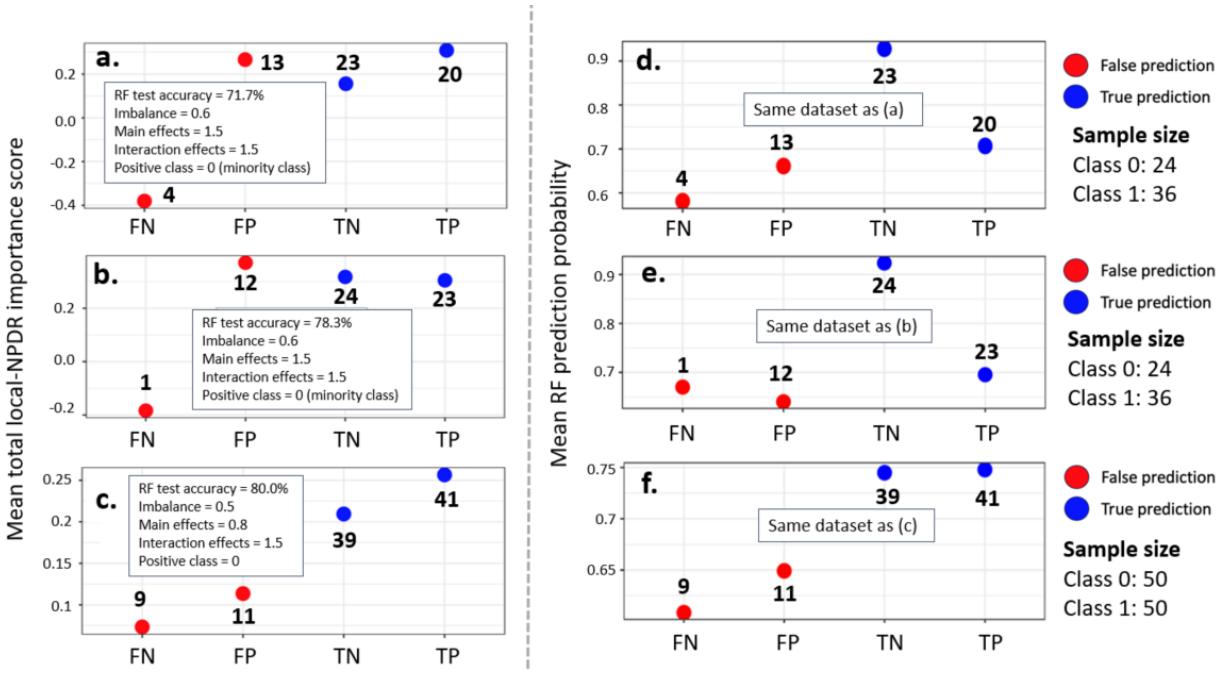


Figure 11. Mean total local-NPDR variable importance scores for three simulated datasets (left panel, a-c) and mean RF prediction probability for the same three simulated datasets (right panel, d-f). Values are broken down by prediction type on the x-axes (FN = false negative, FP = false positive, TN = true negative, TP = true positive). False predictions are indicated by red circles, true predictions by blue, and the number of samples of each prediction type are given next to the points. Class sample sizes are indicated in the legend for each dataset. Each row of figures is one of the three simulated datasets (accuracies summarized in Figs. 4b, 4c, and 4d). (**a**) The distribution of mean total local-NPDR variable importance scores by prediction type is affected by the class imbalance. The mean total local-NPDR variable importance score is high for false positives (class *1* samples incorrectly predicted to be class *0*), true negative samples (class *1* samples correctly classified), and false positives (class *1* samples incorrectly predicted to be class *0*). (**b**) The mean total local-NPDR feature score distribution by prediction type is similarly distributed to those in (a) (compare simulation parameters). (**c**) The mean total local-NPDR variable importance scores show a different distribution. In this case, the classes are balanced, there are more samples, and the main effect size is decreased to 0.8, while the interaction effects are kept at 1.5. For this dataset, the false negative and false positive scores are both lower than the true negative and true positive scores. (**d**) The mean RF prediction probability for the same simulated test samples as in (a) shows lowest prediction probability for false negatives, followed by false positives. (**e**) False positive and false negative mean RF prediction probabilities are lower than for true predictions for the same dataset as in (b). (**f**) For the simulated samples in (c), mean RF prediction probabilities are again lower for false predictions than for true predictions.

3.4.2 *Local-NPDR Feature Importance for Biosignature Data*

For the biosignature data, the *biotic* class corresponds to the positive class. This means that TP and FN predictions involve the minority *biotic* class, while TN and FP predictions involve samples of the majority *abiotic* class. Since the sample sizes are small for the biosignature test data (for example, one prediction type, FP, has only one sample), we will discuss the mean total local-NPDR variable importance scores for the training data. In Sec. 3.3, we present results using local-NPDR variable importance in combination with RF prediction probabilities to diagnose false predictions on holdout simulated and biosignature data.

For the biosignature training data, both the FP and FN mean total local-NPDR importance score is higher than the TN, representing the *abiotic* samples (Fig. 12a). This could be due to the much smaller sample size. If more samples were added, we might expect a distribution that more resembles that of the local-NPDR mean TLS in the imbalanced simulated datasets. Like the imbalanced simulated data, the mean RF prediction probabilities for the false predictions in the biosignature data are much lower than those for the true predictions (Fig. 12b)

The combination of indicators provided by both the mean total local-NPDR variable importance scores and the RF prediction probabilities provide a complimentary approach for identifying possible false or problematic predictions (in which the model is unsure of classification) in samples whose actual class is unknown.

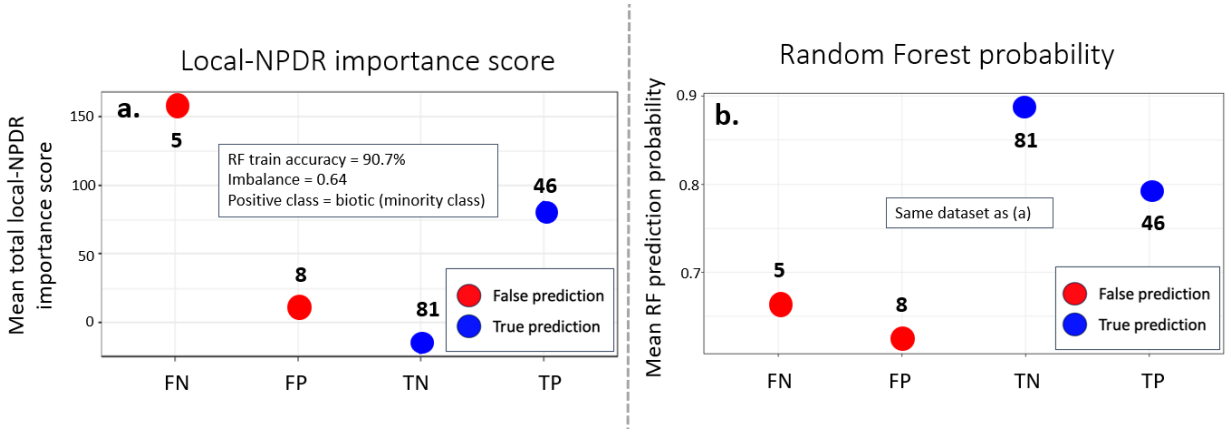


Figure 12. Mean total local-NPDR variable importance and mean RF prediction probability for the four different prediction types in the biosignature training data. The number of samples representing each type of prediction are indicated next to the points. (**a**) Local-NPDR mean total local importance score for each prediction type for training samples in the RF biosignature classification model. The dataset has a class imbalance of 0.64, where *biotic* is the minority class (and is also designated the positive class). This dataset shows a mean total local-NPDR importance score for FP samples that is larger than the score for TN samples, which mirrors behavior seen in the imbalanced simulated data (compare a and Fig. 11a and b). (**b**) The mean RF prediction probability is below 70% for both FN and FP samples and higher for TN (>85%) and TP (>77%) samples.

3.4.3 *Comparison of Local Feature Importance Methods*

Using our RF models trained in the global-NPDR-LURF feature space as the base classifier for each dataset, we compare local-NPDR with local-RF feature importance. Details for each algorithm can be found in the Methods section. As with local-NPDR, we perform a t-test for the local-RF TLS between true and false predictions. For the three simulated datasets, the total local-RF scores are higher in the true versus false prediction groups for both training (Fig. B.3) and test samples (Fig. 13). The elevated TLS in true versus false groups in the test data is statistically significant with P-values ranging from $1.6 \cdot 10^{-6}$ to $<2.2 \cdot 10^{-16}$. Again, there is the potential for score overlap between individual true and false predictions, meaning that

information provided by the RF probability model could be useful in identifying false predictions.

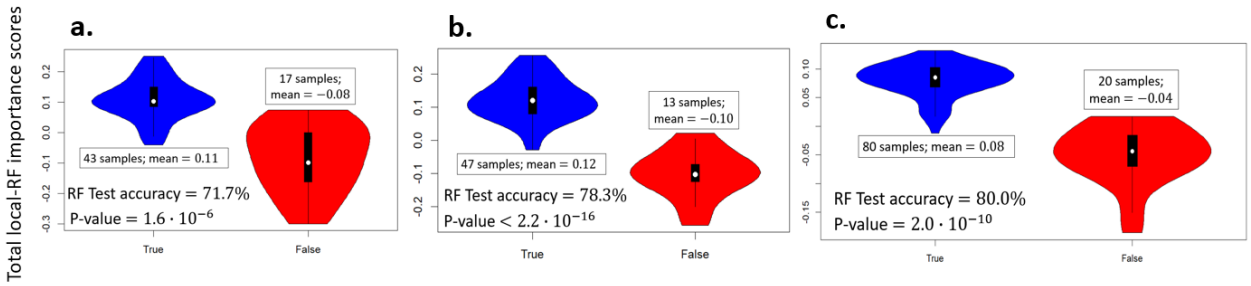


Figure 13. Total local-RF variable importance scores for true and false predictions in three simulated test (holdout) datasets. In each dataset, the local-RF scores are higher in the true (blue) versus false (red) prediction samples (all t-tests statistically significant). (**a**) This class-imbalanced simulated dataset has mean total local-RF importance scores that are higher in the true prediction group (P=$1.6 \cdot 10^{-6}$). (**b**) The mean local-RF variable importance scores for this imbalanced dataset are higher in the true prediction group ($P < 2.2 \cdot 10^{-16}$). (**c**) Local-RF importance scores are in this balanced simulated dataset are higher for true predictions ($P = 2.0 \cdot 10^{-10}$).

An analysis of the mean TLS for local-RF for train and test samples in the three simulated and biosignature datasets versus prediction type (FP, FN, TP, and TN) shows separation between both classes of false predictions and true predictions (Fig. 14 and Fig. B.4). While mean TLS for local-NPDR in some false predictions are higher than mean TLS for some true predictions, local-RF mean-TLS for false predictions are always lower than mean-TLS for true predictions (compare Figs. 11 and 12). Local-RF variable importance is expected to have a good performance at identifying false predictions, since this method is native to the classifier, and these results show that local-RF importance is less affected by class imbalance than local-NPDR. Limitations to local-RF variable importance were mentioned in Sec. 3.1 are discussed more in Sec. 3.5.

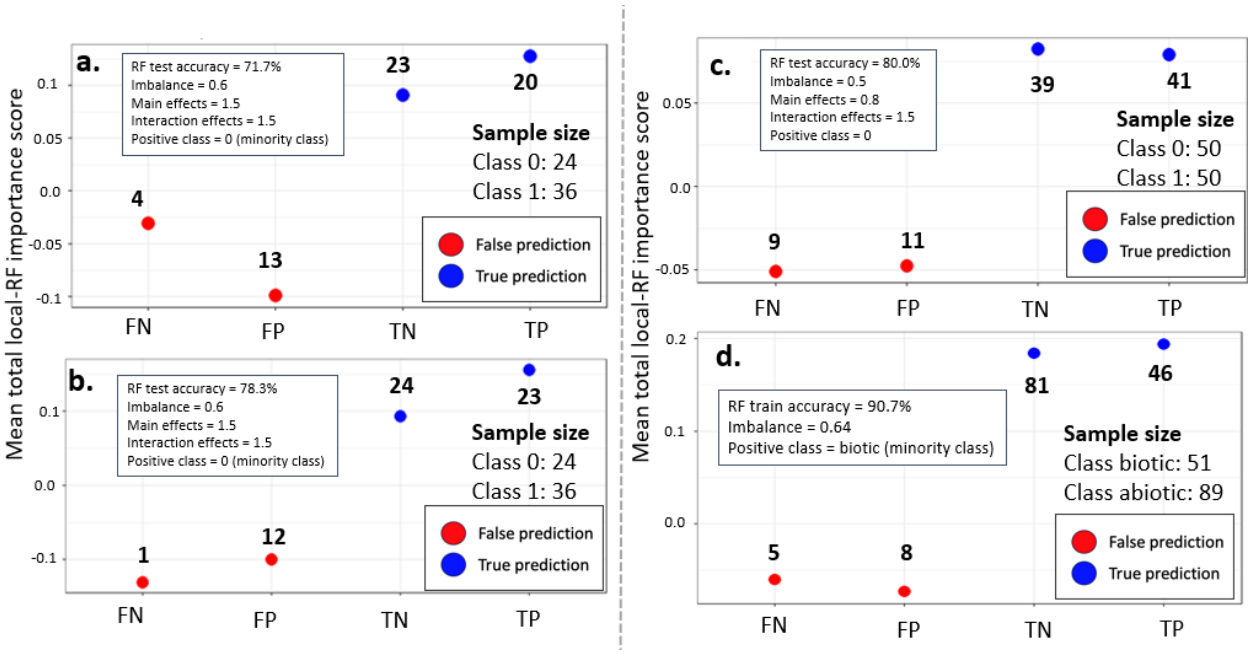


Figure 14. Mean total local score (TLS) by prediction type for local-RF variable importance for four datasets. The number of samples representing each type of prediction are indicated next to the points. (**a**) Local-RF mean TLS for each prediction type for test samples in the simulated dataset with 71.7% RF test accuracy. Both FP and FN predictions have lower scores than TN and TP predictions. **(b)** Local-RF mean TLS for each prediction type for test samples in the simulated dataset with 78.3% RF test accuracy. Again, both false prediction groups show lower mean TLS than true predictions. **(c)** Mean TLS using local-RF for the simulated dataset with 80.0% RF test accuracy shows the most similar distribution between the two false prediction groups and the two true prediction groups separately, likely driven by class balance. **(d)** The mean local-RF TLS for the biosignature data training samples (depicted for increased sample sizes) shows lower FN and FP scores than TN and TP scores. Although imbalanced, this dataset shows the most similar TLS distribution to the balanced simulated dataset in (c).

Both local-NPDR and local-RF result in statistically significant differences in mean-TLS between true and false predictions in the simulated datasets (Fig. 10 and Fig. 13). For the biosignature data, local-NPDR variable importance indicates less clear separation between true and false prediction scores (compare Fig. 7a and Fig. 9d), indicating the sensitivity of the multivariate regression to both imbalance and small sample sizes, discussed in more detail in Sec. 3.5. In the next section, we compare the ability of local-RF and local-NPDR to diagnose false predictions in "unknown" samples across the four datasets. While the results in this section

may lead one to conclude that local-RF will always outperform local-NPDR, two additional sets of analyses on the test samples reveal a comparable performance between the two methods.

3.4.4 *Diagnosing False Predictions in "Unknown" Samples*

Four samples, each representing the four prediction types, are chosen for further analysis from test samples in each of the simulated datasets and the biosignature data. Local-NPDR and local-RF total local importance scores are calculated for each of the four samples, as well as the prediction probabilities; for each method, we attempt to characterize the results as either indicative of a true prediction or a false one. Results for the simulated datasets can be found in Appendix B.4, and we present the results of this analysis for the biosignature dataset here.

Consider the case where the actual class of the four test samples is unknown. Local importance methods and classification probabilities can help us identify potentially false predictions in this scenario by analyzing the variable importance TLS for the samples, the individual local feature importance scores, and the RF classifier prediction probability (Fig. 15). For example, given the fact that an "unknown" sample has a positive total local-NPDR score of 19.4, that only one of the features has a negative local importance score, and that the classifier reports an 82.1% probability that the sample is *biotic*, we accept this prediction as a likely true positive (Fig. 15a). Likewise, for an *abiotic* classification with a high-magnitude positive local-NPDR score of 100.6, small-magnitude negative local scores for individual variables, and a RF prediction probability of 81.3%, we accept this classification as a likely true negative (Fig. 15b). If the TLS for local-NPDR is close to zero or negative, if there are large-magnitude negative local variable importance scores, and if the RF prediction probability is low, these samples are subject to being flagged as potential false predictions (Fig. 10c and d). For a sample with a local-

NPDR score of -33.35, a large-magnitude negative score for the top global-NPDR ranked feature and an RF prediction probability of 55.5%, this *biotic* prediction is flagged as a potential false positive (Fig. 15c). For an *abiotic* prediction with a similar large-magnitude negative score for the top global-NPDR feature, a TLS of -67.6, and an RF classification probability of 63.3%, this sample is flagged as a likely false negative (Fig. 15d).

An analogous analysis using local-RF as the importance method for the same four test samples uses similar reasoning (Fig. 16). The magnitude of local-RF and local-NPDR variable importance scores differs because the nature of the methods is fundamentally different; local-RF importance scores are changes in accuracy after and before variable permutation, while local-NPDR scores represent regression coefficients for the pairwise sample regression. This means the magnitudes for the local-NPDR scores will vary by dataset, while the local-RF importance scores will always represent a change in percent accuracy. While the particular variables that are negative differ between local-RF and local-RF, the TLS for the true predictions are positive, in agreement with local-NPDR for the true positive (Fig. 16a) and true negative sample predictions (Fig. 16b). The TLS is negative for the false positive (Fig. 16c) and false negative (Fig. 16d) predictions, again in agreement with local-NPDR for these samples. In general, the concordance/discordance in true/false predictions is more pronounced in the local-RF scores for these samples than for the local-NPDR scores (compare the amount of blue in true and red in false predictions in Figs. 15 and 16).

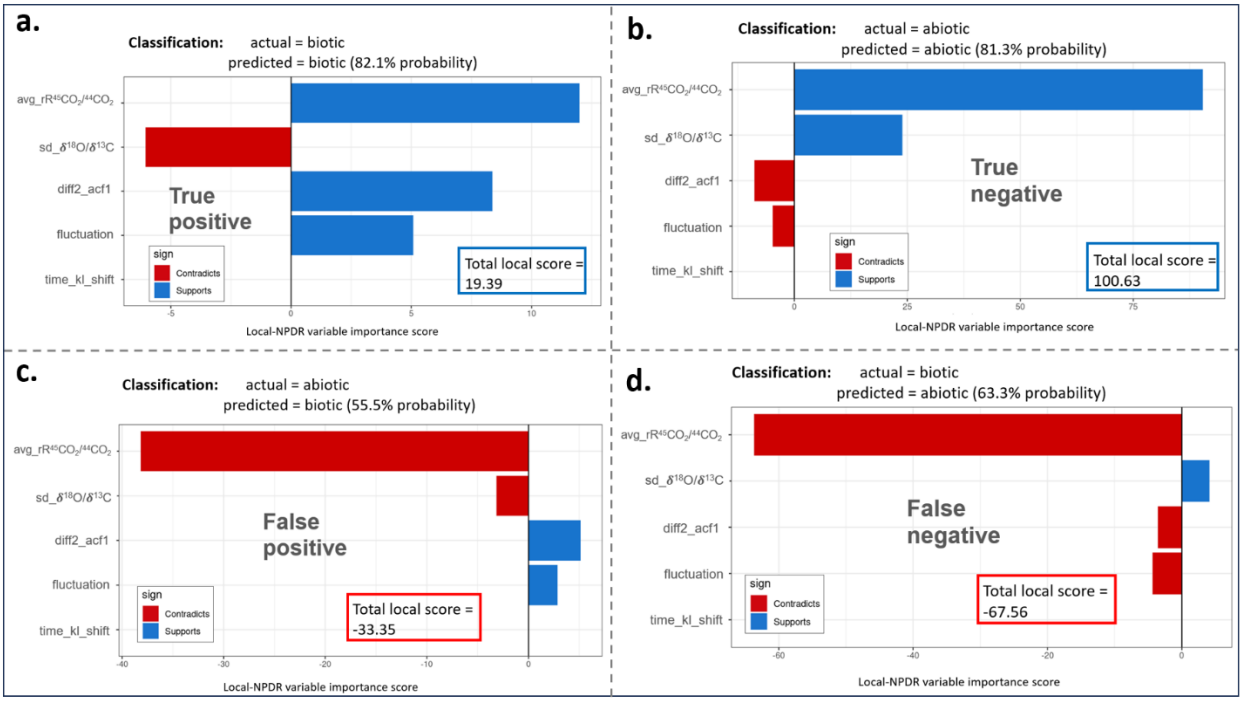


Figure 15. Typical distributions of local-NPDR variable importance scores for true and false predictions. Variables shown are listed according to global-NPDR importance ranking, with the most important feature for classification at the top. "Contradicts" (red) means the local-NPDR score of a variable is negative, in disagreement with the global-NPDR importance. "Supports" (blue) means the local-NPDR score for a variable is positive, in agreement with the global-NPDR importance. (**a**) For a *biotic* sample correctly classified as *biotic* by the RF biosignature model, a true positive, the overall local variable importance score is 19.4. This positive score indicates the local variable importance scores are mostly concordant with the assigned class for this sample (*biotic*). Additionally, the RF probability model reports a high prediction probability of 98.8%, increasing the likelihood that this is a correct prediction. (**b**) For an *abiotic* sample correctly predicted to be *abiotic*, a true negative, the total local score (TLS) is 100.6 and the RF prediction probability is 81.3%, suggesting this is a correct prediction. (**c**) For an *abiotic* sample incorrectly predicted to be *biotic*, a false positive, the TLS is -33.4. This large negative score flags the sample as a potential false prediction in which the assigned classed, *biotic*, is not concordant with the local variable importance scores of the two most important global-NPDR features. In this case the RF probability model yields a low prediction probability of 55.5%, further increasing doubt in the validity of this classification. (**d**) For a *biotic* sample incorrectly predicted to be *abiotic*, a false negative, the local-NPDR TLS is -67.6, with a large-magnitude negative score for the top global-NPDR feature. The RF probability is reported as 63.3%, and this along with the large negative TLS, indicates this sample is likely incorrectly classified.

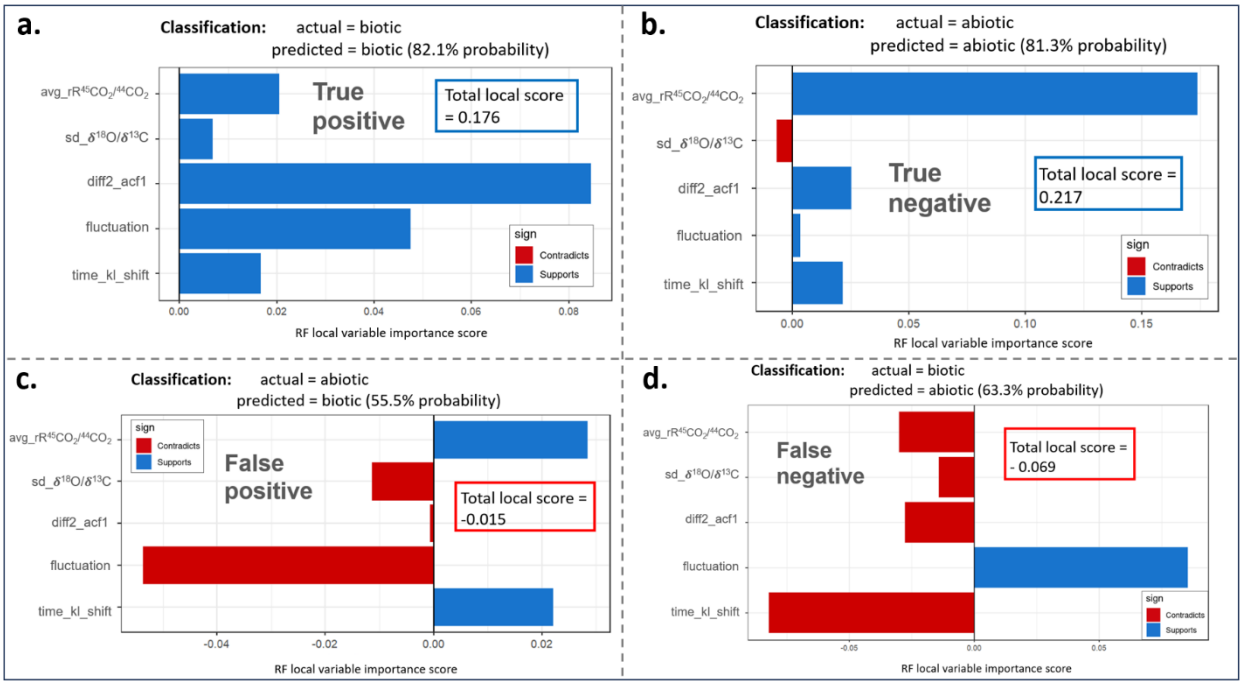


Figure 16. Local-RF variable importance scores for the cases analyzed by local-NPDR. The variables are given according to global-NPDR importance scores. Note the magnitudes of local-RF and local-NPDR scores differ due to the fundamentally different methods of local variable importance calculation. The prediction probabilities are the same as reported for local-NPDR because the same RF probability model is used; likewise, the same RF classifier is used to make the class assignment. "Contradicts" (red bars) means the local-RF score is negative (unimportant for classification) for a variable that is globally important, while "Supports" (blue bars) means the local variable importance score is positive, agreeing with the global importance of the variable. (**a**) A sample classified as *biotic* shows all positive local-RF importance scores in agreement with (supporting) the global variable importance scores. The total local score (TLS) is 0.18 and the RF prediction probability is 82.1%, indicating that this is likely a true classification of a biosignature. (**b**) With only one small negative variable importance score, this *abiotic* prediction has a TLS of 0.22 and a RF probability of 81.3%. We accept this abiotic classification as a likely true prediction. (**c**) A sample classified as *biotic* with a local-RF TLS of -0.015 and a RF prediction probability of 55.5% is flagged as a potential false positive prediction. (**d**) An *abiotic* prediction with a TLS of -0.07 and several negative variable importance scores has a 63.3% RF probability. This *abiotic* prediction is flagged as a potential false negative.

To illustrate an applied comparative analysis, consider the case where we have a set of test samples with unknown actual classes and a trained RF classifier and probability model with global-NPDR selected features that labels each test sample with a prediction and probability. We would like to "quarantine" the test samples that could be falsely predicted from the samples that we are confident are correctly classified. To do this, we apply both the RF probability model and the TLS discussed above (Fig. 15) and define an acceptable probability and a TLS for which we accept samples. An immediate complication is defining an appropriate RF probability and TLS threshold to flag samples. A detailed discussion of these considerations is deferred until Sec. 3.5 ; for now, consider that the TLS threshold will be local importance method dependent (compare the differences in local-RF and local-NPDR importance scores) and subject to user preference, as will the probability threshold. One way to decide a reasonable TLS threshold is to analyze the distribution of scores in the training datasets for each local importance method and decide a threshold that will result in not "too many" samples being flagged (Figs. B.6 and B.7), since it is expected that some true predictions with low probabilities or low/negative TLSs will also be quarantined. For this analysis, we use an RF prediction probability threshold of 75% for all datasets and both local importance methods.

For the three simulated datasets, we use local-NPDR importance score thresholds of {0.25, 0.35, 0.35} for the datasets with RF test accuracies of {71.7%, 78.3%, and 80.0%} respectively. These thresholds differ because of the different distributions of local-NPDR scores in the three datasets (see Fig. B.6). For the biosignature data, we use a local-NPDR score threshold of 25. Analysis of the distribution of local-RF importance scores for the training datasets show a similar distribution among the datasets due to the nature of the importance score calculation. We therefore define a threshold of 0 for local-RF importance for all datasets. False

samples in the test data for each dataset can be detected using both methods (Table 4). For the arbitrary thresholds defined, local-RF and local-NPDR perform comparably well, showing similar or comparable overall false prediction diagnostic rates (compare third and fifth columns, Table 4). For the biosignature data, the two methods flag the same falsely predicted samples, one false positive and one false negative (compare biosignature data in columns 4 and 6), and the miss the same false negative.

In this example analysis, local-RF flags fewer true predictions than local-NPDR in all datasets. For example, in the simulated dataset with 71.7% RF test accuracy, local-NPDR flags 26 total samples while local-RF only flags 12. Out of the 26 flagged by local-NPDR, 14 are true predictions with low TLS; out of the 12 flagged by local-RF, only two are true predictions. It is worth mentioning that global-NPDR feature selection has significantly enabled local-RF in this analysis; this will be discussed more in Sec. 3.5. To summarize this applied analysis, local-RF and local-NPDR do a comparable job flagging false predictions in the three simulated datasets and the biosignature dataset. However, local-RF, when supplied with a model trained on global-NPDR selected features, flags significantly fewer true predictions than local-NPDR for all datasets.

Table 4. False prediction diagnosis rates for local-RF and local-NPDR variable importance methods for four datasets using either a RF prediction probability <75% or an TLS less than an arbitrary threshold.

Dataset	RF test accuracy	Local-RF false prediction diagnostic rate	FP/FN diagnosis rate: local-RF	Local-NPDR false prediction diagnostic rate	FP/FN diagnosis rate: local-NPDR
Simulated	71.7%	52.9%	61.5% / 25.0%	70.6%	61.5% / 100.0%
Simulated	78.3%	76.9%	75.0% / 100.0%	61.5%	58.3% / 100.0%
Simulated	80.%	75.0%	72.7% / 77.8%	90.0%	81.8% / 100.0%
Biosignature	91.2%	66.7%	100.0% / 50.0%	66.7%	100.0% / 50.0%

3.5 Conclusions

To explain individual sample predictions, we extended (global) NPDR variable importance to compute importance scores in the context of the neighborhood of a single sample. Local-NPDR uses a generalized linear model to contrastively determine whether neighbors of a sample of interest are in the same or different class (hits or misses). Variable importance scores are coefficients in the contrastive loss optimization, which can include LASSO or Ridge penalties. NPDR is sensitive to detecting interactions, a significant advantage in feature selection and importance ranking for high-dimensional datasets. We used an URFP distance metric to define the neighborhood because it effectively handles non-isotropic variable spaces and reduces correlation between variables; however, NPDR can accept any number of different distance metrics. In addition to the distance metric, the choice of *k* affects NPDR's ability to detect interactions (Dawkins & McKinney, 2025). For global-NPDR, we used a default sample-size-dependent *k* that reliably balances interaction and main effects and accounts for class imbalance. For local-NPDR, we used the maximum number of neighbors to increase power to detect important variables.

We used the local-NPDR scores to explain RF predictions of individual samples, and we used the total local score (TLS) of the globally important variables to help diagnose false predictions. We showed that high positive TLS for samples is associated with true predictions, and low or negative TLS is associated with false predictions. The RF sample prediction probability provided additional true/false diagnostic evidence.

For comparison with local-NPDR, we used local-RF variable importance, which uses the change in classification accuracy after variable permutation to score each variable for trees in which the sample is oob; if the classification accuracy increases or stays the same upon random permutation, that variable is not considered important for classification; however, if the classification accuracy decreases upon random permutation of the variable, that variable is considered important for classification. For RF permutation importance, the sample must be part of the training data, meaning that a new RF model must be trained to generate local-RF variable importance scores for a single new test sample. Local-RF does better with class imbalance and the small biosignature dataset in terms of separating true and false scores, but because it is a method used during classifier training, it could change the nature of the model during the re-training process itself. For the analysis of a single new sample, the change to the base classifier is expected to be negligible; however, if many new samples are introduced in order to generate local-RF importance scores, this could significantly alter the model from the one that originally classified the sample, altering the explainability and in a worst-case scenario, the truth of the outcome variable.

RF importance has limited ability to detect interaction effects or quantify importance due to main effects versus interactions, meaning that a more in-depth explanation for variable effects on an individual sample prediction as illustrated in the previous paragraphs would be difficult

with local RF. This limited ability can be seen by comparing the results of local-NPDR with local-RF in the balanced simulated dataset (compare Fig. 10c, where this dataset has the lowest P-value for local-NPDR between true and false predictions, with Fig. 13c, where local-RF variable importance results in the second-lowest P-value for this dataset). This balanced simulated dataset was therefore not the best performing dataset for the local-RF method in separating true and false predictions, in contrast with local-NPDR, despite the better class balance and increased sample size. This is likely because we decreased the magnitude of variable main effects in this dataset compared with interactions, making it a more difficult classification problem for RF. In this way, this simulated dataset is more like the real biosignature data. The ability to compare the differential effects of class imbalance, sample size, and the relative magnitudes of main/interaction effects for variables is enabled by the simulated data, and we can see through this analysis that it is a combination of interaction effects and class imbalance that affect both classification and feature importance methods in the biosignature dataset. This has immediate implications for the deployment of ML methods for astrobiology missions: for real complex datasets, as geochemical isotopic data for biosignatures is, the ability to detect statistical interactions is a significant advantage.

Local-NPDR variable importance scores also have the advantage of being calculated independently from the classifier. This could be a potential limitation, since it will not be a perfect explainer for the classifier. But if a classifier is trained on global-NPDR selected features, local-NPDR will be informative as to whether the outcome label generated by the classifier matches what is expected by the algorithm in the context of a neighborhood of samples in the global-NPDR feature space. Since local-NPDR requires a distance matrix, it still requires access to the training data, like local-RF. However, no model re-training or hyperparameter tuning is

needed. Local-NPDR showed evidence of being sensitive to imbalance and small sample sizes where local-RF was less susceptible. The NPDR multivariate regression on sample pairs would benefit from increased sample sizes and an improved way to handle data imbalance, which can be an intractable ML problem. Despite these challenges, local-NPDR performs similarly well to local-RF in diagnosing false test predictions in the biosignature dataset and in the simulated datasets (see Table 4).

In practice, diagnosing false samples using either local-RF or local-NPDR requires the user to define an appropriate RF probability and TLS to use to flag samples. Future work will incorporate suggested statistical thresholds that may be reasonable; however, the acceptable risk for false predictions will be user and application dependent. For example, in terms of biosignature detection in a future astrobiology mission to an OW, there may be very little tolerance for a false ML prediction. Researchers may use a much stricter RF probability threshold than 75%, and a lower TLS to diagnose potential false predictions. The tradeoff of using a stricter threshold is an increase in false negatives (true predictions quarantined). Additionally, knowledge of statistical interactions and the relative importance of each variable (*e.g.*, as quantified in NPDR's Epistasis Rank) for each prediction type may be incorporated into an analysis of a ML prediction, ensuring a more robust explanation of the likelihood of a true or false prediction. This type of nuanced analysis can be leveraged to preserve mission resources for future astrobiology missions by increasing the fidelity of the local-NPDR false prediction diagnosis method.

In the current study we show that local-RF variable importance benefits from global-NPDR feature selection. In our application example, local-RF flagged fewer true predictions than local-NPDR, which was enabled by global-NPDR feature selection. It has been previously

shown that the RF biosignature classifier benefits from global-NPDR feature selection using an URFP distance metric (Clough et al., 2025). Since RF has no statistical threshold for limiting the number of variables, if feature selection were not performed, the user would have to decide which of the 100 features to use in a local importance analysis. Using all 100 features, some of which are noise or highly correlated, is likely to result in overfitting, potentially compromising the local-RF false prediction diagnosis method. The LASSO version of NPDR provides a feature selection threshold and helps local RF detect false predictions.

Global-NPDR feature selection also allows the RF classifiers to be considerably more lightweight than if the classifier were required to use the full variable space, so that if training a new model is required, as is the case for the local-RF variable importance method, that process is much less computationally intensive. This has implications for applications such as online learning in automated space exploration. NDPR variable importance methods, both global and local, can contribute to ensuring ML models and data products are an appropriate size for use on flight computers while maintaining accuracy and increasing interpretability.

It is therefore the best practice to use some form of feature selection and global-NPDR with LASSO penalty, a method sensitive to statistical interactions in high-dimensional datasets, is a natural choice in complex real datasets, like IRMS measurements for astrobiology or gene expression data for disease prediction. The URFP distance metric adds an additional ability to construct a neighborhood in a non-isotropic variable space, an advantage over traditional distance metrics. NPDR provides additional information about each variable's contributions in terms of main effects and interactions, which enables a more in-depth analysis of each prediction based on the local-NPDR importance scores. This allows us to gain much more than a prediction label for each experimental sample we wish to classify. We can start to articulate how the black

box RF is using individual variables to inform classification, and exactly how particular variables may be fooling the model in the case of false predictions. The implications for this increased understanding in the search for OW biosignatures are that we can encode our quantitative understanding of variable effects and each variable's ability to classify samples of a particular class into our science autonomy framework for exploration. Additionally, the use of a novel distance matrix enables even more analysis of individual samples, including outlier and anomaly detection, while preserving non-isotropic variable relationships, which are expected in complex real datasets.

CHAPTER 4

MACHINE LEARNING CHEMISTRY PREDICTION FROM OCEAN WORLD ANALOGUE DATA

As discussed previously, future missions to remote OWs in our solar system will benefit from trustworthy methods for science autonomy. The high-accuracy RF biosignature classifier presented in Chapter 3 is trained on an interpretable variable space using features selected by global-NPDR-LURF that are related to carbon and oxygen isotope ratios and time series features of the chromatograms resulting from measurements of volatile CO_2 generated from the OW analogue brines. We added a local feature importance tool, local-NPDR, which reports variable importance scores for single samples to help explain ML predictions and identify false predictions. We showed that this method is comparable to local-RF variable importance and is more dependable in datasets with lower variable main effects and more statistical interactions but is affected by class imbalance. We now apply our suite of explainable ML tools to develop additional RF models that can accurately predict chemical properties of the OW analogue data. Outcome properties of interest include volatile CO_2 concentration and the presence of important anions in the brines, such as sulfate, bicarbonate, and chloride, as well as analogue seawater pH and ionic strength. We present each of these chemical RF models in an interpretable global-NPDR selected features space, which is small enough in many cases to be visualized in our interpretable statistical network, RAIN. Finally, our local-NPDR feature importance tool is used to diagnose false chemical predictions in holdout data. These tools build trust in ML predictions by enabling explanation of individual sample predictions and will be important methods for future astrobiology and geochemical exploration of OWs.

4.1 Ocean Worlds Seawater Chemistry

The prediction of OW seawater chemistry from orbiting instrumentation is important for understanding the geochemistry of these worlds without the ability to deploy a lander. It is therefore essential to understand whether subsurface OW chemistry can be deduced from sampling volatiles evolved from the subsurface. Although it is expected that the transportation mechanisms for subsurface liquids into exosphere ices and volatiles will alter the original chemistry, it is hoped that we can still make predictions about the original chemistry. Future experimental work will focus on understanding these transportation effects on *e.g.,* carbon and oxygen isotope ratios; for this initial effort, we characterize the brine chemistry using volatile CO_2 evolved from the brine at Earth temperatures and pressures using our explainable RF models trained using NPDR feature selection.

The following analysis constitutes a proof-of-principle ML chemical characterization of OW analogue brines using explainable ML methods. First, unsupervised learning techniques such as KNN network clustering are applied to the IRMS and time series data to understand the inherent similarities and differences between samples in the OW analogue brine dataset and to visualize how these patterns map to their chemical characteristics. Techniques like this can be used to understand whether pre-trained ML models are likely to be successful in new environments by allowing us to understand if a new environment is inherently similar to or significantly different from any samples in the training data. If a new environment has a similar chemistry to samples in the training data, a pre-trained model may be appropriate and successful. If, however, a new environment is very different from samples in the training data, an online learning process may be appropriate and pre-trained models should be used with caution and

should incorporate explainability tools like local-NPDR and local-RF employed. This unsupervised approach can also help identify systematic biases in the experimental design that might affect supervised learning. We then present the results of supervised learning for the chemical characterization of the OW analogue brines. The supervised ML procedure is similar to the biosignature approach, which uses LURF-NDPR feature selection, RF classification and regression, RAIN interaction network visualization, and local-NPDR single sample explainability and false prediction diagnosis. We successfully detect major anions such as bicarbonate and sulfate and predict volatile CO_2 concentration. The detection of chloride and prediction of continuous outcomes such as pH and ionic strength prove more challenging. We discuss reasons for these challenges and ways to address them in the future. Finally, we discuss implications of this analysis for geochemical investigations on OWs.

4.2 Machine Learning Methods and Dataset

Unsupervised learning methods include KNN network representation of samples with Louvain (network-based) and Greedy (Newman) clustering. In our KNN cluster analysis, network nodes are samples plotted on a 2-D plane, with relational edges between nodes defined based on proximity in the space of selected IRMS and time-series features. Network-based algorithms can then be used to identify the number of clusters in the network and the number of samples (and which samples) are in each cluster. This ensures that the number of clusters is determined in a fully automated manner and with limited manual intervention or bias.

In Chapters 2 and 3, supervised learning was performed using RF classification for two categorical outcomes and regression for continuous outcomes (Breiman, 2001). Instead of two classes, *e.g.*, *biotic* and *abiotic*, some RF chemistry classifiers in the current chapter may have

three or more outcome classes. These classifiers are expected to present the RF algorithm with a more difficult classification task because of potential sparseness of classes, especially for small datasets such as the ones discussed here.

Instead of using samples where the injected amount of volatile CO_2 is constant, as in the biosignature model presented in Chapter 2, here we use the entire range of experimental $[CO_2](g)$, which spans from 0 to 2% $CO_2(g)$ by volume (12mL vials were used in experiments, see Sections 1.3 and A.1). This increases the sample size compared to the biotic/abiotic analysis, which used a fix value for CO_2 concentration. Maximizing the sample size is important for training/testing models with more than two classes, as well as for regression models, which require larger sample sizes to represent all classes and achieve high accuracy. Furthermore, the inclusion of samples with variable injected CO_2 concentrations allows the models to be more agnostic to the atmospheric concentration of carbon dioxide. We use an extension of the dataset presented in Chapters 2 and 3. There are 290 samples in each training dataset and 71 samples in each test dataset for the development of the RF models to predict analogue seawater chemical parameters.

4.3 Unsupervised Learning Results

In this section we present the results of unsupervised learning to find inherent chemical patterns in the OW analogue brine data. Previously we compared UMAP, t-SNE and KNN network clustering; here we show the KNN network clustering results, as this network had the most similarities with the chemical information about the samples in each cluster and is therefore the most informative for subsequent ML analyses. In the KNN network method, an unweighted network is made using a distance matrix with the NPDR function `nearestNeighbors`. We

compute a distance matrix between samples and use it to create an edge list by specifying a threshold of k nearest neighbors for each sample/node. We use R *igraph* library function `graph_from_edgelist` to convert the edgelist to an adjacency matrix.

Clustering or community detection is a main goal of network analysis. If samples belong to a shared community, that means they are nodes that have a higher likelihood of being connected to each other than to nodes in other clusters. Furthermore, a community can be considered a locally dense subgraph (Barabási, 2016), meaning there are inherent similarities within the community compared with samples in other clusters. There are several different methods for clustering, including hierarchical clustering, recursive partitioning, and a statistical mechanical approach based on information theory and entropy (Jaynes, 1957). Here we use a network-based type of hierarchical clustering called Louvain (Bhowmick et al., 2020). This method is known to be able to detect substructure in larger clusters by using a network embedding to detect communities. In this case the hierarchical clustering uses an aggregation approach (*i.e.*, combines embeddings of nodes) over different levels in the hierarchy to find final embedding vectors. We compare Louvain clustering with another community detection approach based on recursive partitioning, called Greedy Modularity (Newman, 2006). Modularity, sometimes called Newman's modularity, in network analysis is a matrix used to discover network community structure and is defined as the difference between an adjacency matrix element, which represents the observed linkage between two nodes, and the expected number of connections by random chance. The larger the difference between the observed connections and those expected by chance, the more inherent structure a cluster is considered to have. It should be noted that Louvain clustering, a more modern form of community detection, also employs Newman's modularity, but rather than a divisive hierarchical clustering, where the algorithm

selects nodes in different communities to partition the network, as implemented in the original Girvan-Newman algorithm, Louvain uses an aggregative hierarchical approach.

4.3.1 *KNN-Network Clustering Using URFP Distance in the Full Variable Space*

A KNN network was created in the full variable space using the URFP distance matrix. Greedy and Louvain clustering are then performed on the resulting KNN network, yielding 11 Greedy clusters and 18 Louvain clusters (Fig. 17). That Louvain clustering results in more communities than Greedy is consistent with the claim that it is better able to detect substructure; however, verifying whether this substructure is meaningful or a manifestation of real similarities or differences in the sample nodes is more difficult to determine. Many of the clusters resulting from Greedy modularity make sense in terms of the network layout, except for the grey/light blue clusters in the top middle of the graph (Fig. 17a). The mixed clusters in this area could indicate real substructure here and are consistent with the Louvain clusters in that area (Fig. 17b, red and green clusters at the top of the graph). In the area immediately below this, Louvain and Greedy disagree; where Greedy finds substructure in the middle area (dark blue and green clusters, Fig. 17a) while Louvain does not (light blue middle cluster, Fig. 17b). Overall, however, it is clear Louvain finds more substructure within visibly clustered areas than Greedy modularity.

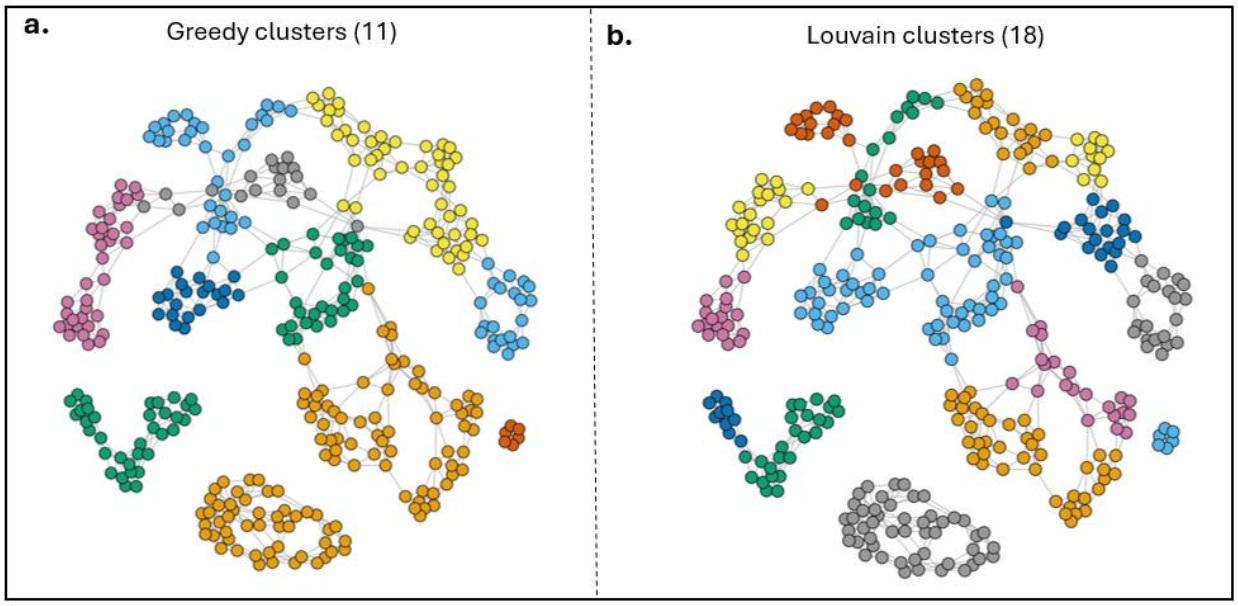


Figure 17. Greedy and Louvain clusters in the KNN network created in the full variable space using the URFP distance. (**a**) Clustering using Greedy Modularity results in 11 clusters for this network. Most of these clusters make sense in terms of the network layout, and the algorithm finds some substructure in the top part of the graph (grey and light blue clusters). (**b**) Community detection using the Louvain method and modularity yields 18 clusters. The Louvain clusters overall show more substructure than the Greedy clusters.

To determine whether the substructures detected by Louvain clustering are meaningful according to known parameters about the nodes, we compare the Louvain-detected clusters on the KNN network with known salt content of the sample brine, [CO2](g), and environmental dataset label (Fig. 18). From the KNN network colored by sample brine salt content, it can be seen that there is real substructure correlated with major salts that is detected by Louvain clustering (Fig. 18b). For example, clusters *XIV* and *XV* have real compositional differences (sulfate versus bicarbonate). However, not all Louvain-identified substructures make sense in terms of salt content (compare clusters *XVI* and *XVII*, Fig. 18b). In some cases, the substructure not explained by salt content can be explained by differences in sample volatile carbon dioxide concentration, such as clusters *XVI* and *XVII*, which contain 2% and 0.3% CO_2(g), respectively (Fig. 18c). The substructure at the top of the KNN graph detected by both Louvain and Greedy

clustering can be explained by differences in biotic class (compare *biotic* green cluster *I* with *XI*, *XII*, and *XIII abiotic* clusters, Fig. 18d).

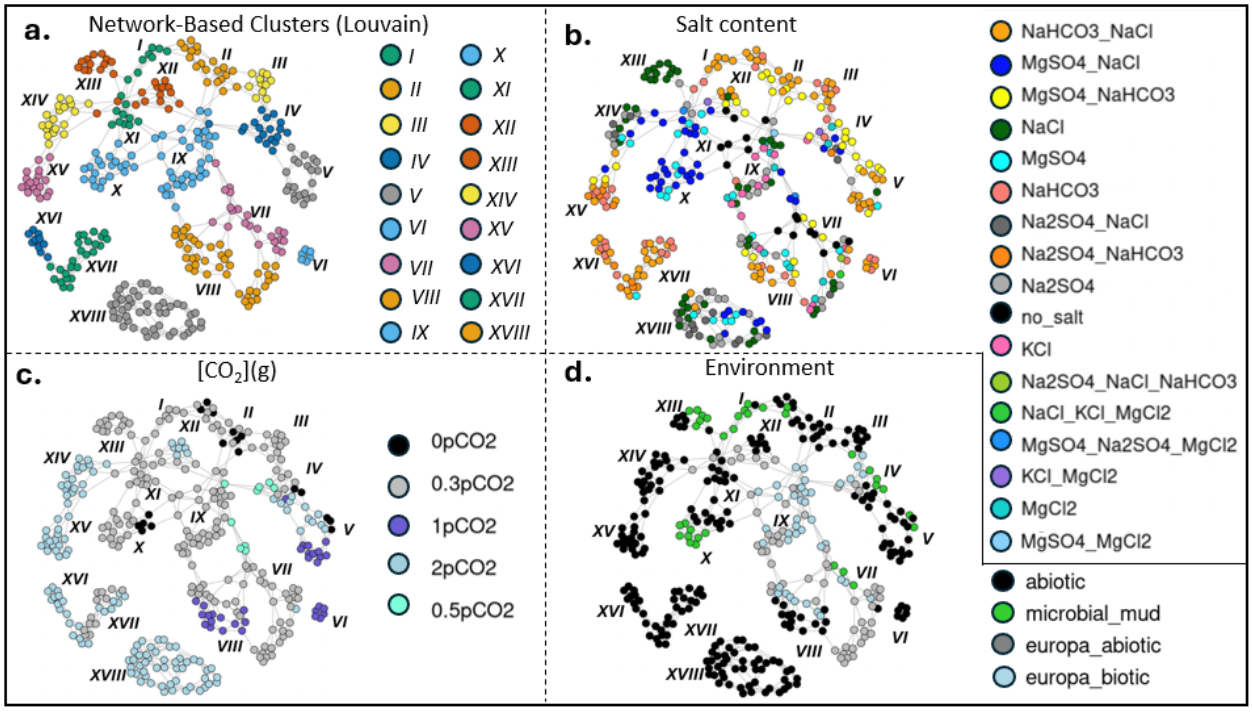


Figure 18. Comparison of Louvain cluster membership in the KNN network created using the full variable space and URFP distance by salt content, carbon dioxide concentration, and environmental dataset. (**a**) The 18 Louvain clusters from Fig. 17b are numbered for comparison with other plots. (**b**) The KNN network colored by salt content shows that some real substructure correlating with salt content has been found. For instance, clusters *XIV* and *XV* have real differences in composition (sulfate versus bicarbonate compositions), as do *I*, *II*, *II*, (bicarbonate and sulfate dominated) and *IV* (sulfate and chloride dominated). However, not all Louvain-identified substructures make sense in terms of salt content (compare clusters *XVI* and *XVII*). (**c**) Some Louvain clusters can be explained by differences in samples volatile carbon dioxide concentration, such as clusters *XVI* and *XVII*, which contain 2% and 0.3% CO_2(g) while the brines contain predominately bicarbonate ions. (**d**) The substructure at the top of the graph detected by both Louvain and Greedy clustering can be explained by differences in biotic class (compare *biotic* green cluster *I* with *XI*, *XII*, and *XIII abiotic* clusters.

One advantage of Louvain and Greedy clustering is that individual samples from each cluster can be analyzed without the bias introduced by arbitrarily asserting the number of clusters or neighbors, as is required in some methods (see Sec. 2.2.1 for discussion on our algorithmic method of determining *k*). Analyzing cluster membership by salt content reveals more clearly which clusters are driven by salt content rather than carbon dioxide or environmental dataset label (Fig. 19). For example, samples in cluster 2 all contain $MgSO_4$, making this one of the purest clusters in the graph, along with cluster 10, whose samples all contain $NaHCO_3$. Interestingly, samples that contain no salt are not clustered together, meaning there are other factors contributing to the structure, like CO_2 (g) concentration, biosignature presence, and likely other unknown factors.

Louvain	Clust 1	Clust 2	Clust 3	Clust 4	Clust 5	Clust 6	Clust 7	Clust 8	Clust 9	Clust 10	Clust 11	Clust 12	Clust 13	Clust 14	Clust 15	Clust 16	Clust 17	Clust 18
KCl	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	2	8
KCl_MgCl2	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
MgCl2	0	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	2	4
MgSO4	0	5	3	0	0	0	1	1	3	0	1	2	0	1	0	5	0	1
MgSO4_MgCl2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MgSO4_Na2SO4 _MgCl2	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
MgSO4_NaCl	0	16	5	0	3	0	3	0	0	0	0	3	0	2	1	7	0	0
MgSO4_ NaHCO3	7	0	0	1	6	0	3	6	4	0	0	0	0	3	3	0	0	0
Na2SO4	0	0	2	0	0	1	1	1	4	0	0	5	0	0	0	11	3	6
Na2SO4_NaCl	0	0	0	0	2	0	0	0	0	0	0	3	0	0	0	12	1	0
Na2SO4_NaCl_ NaHCO3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Na2SO4_ NaHCO3	0	0	0	0	1	0	0	2	1	0	4	0	1	1	4	0	1	0
NaCl	0	0	0	0	0	12	0	1	2	0	0	5	0	1	0	11	4	8
NaCl_KCl_ MgCl2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
NaHCO3	2	0	2	4	2	0	1	0	2	3	7	0	4	0	6	0	1	1
NaHCO3_NaCl	11	0	6	6	0	0	1	10	7	3	13	1	5	3	4	2	0	0
no_salt	0	0	0	0	0	0	9	0	0	0	0	0	0	2	0	0	0	6

Figure 19. Salt content of Louvain clusters. Cluster 2, consisting of all samples that contain $MgSO_4$, is one of the purest clusters in terms of salt content. Likewise, cluster 10 samples all contain $NaHCO_3$. Samples that contain no salt are not clustered together, meaning there are other factors contributing to the structure, such as volatile carbon dioxide concentration.

4.3.2 *KNN-Network Clustering Using URFP Distance in the NPDR-URF Selected Variable Space*

Because of the small sample size and large relative number of classes in the bulk salt dataset, NPDR-URF feature selection is performed using P-values rather than a LASSO or Ridge penalty. Using a P-value threshold <0.05, 30 features are selected, detailed in the next section (see Fig. 24). To understand what kind of inherent information is provided by the NPDR-URF selected features for bulk salt content, we perform the same KNN network clustering as described in the previous section. In this case, Greedy modularity results in five clusters while Louvain clustering results in seven, fewer in each case than detected in the network created using the full variable space (Fig. 20). Again comparing the Louvain clusters with salt content, carbon dioxide concentration and environmental labels, we find that there are still substructures highly correlated with salt content and other chemical labels (Fig. 21).

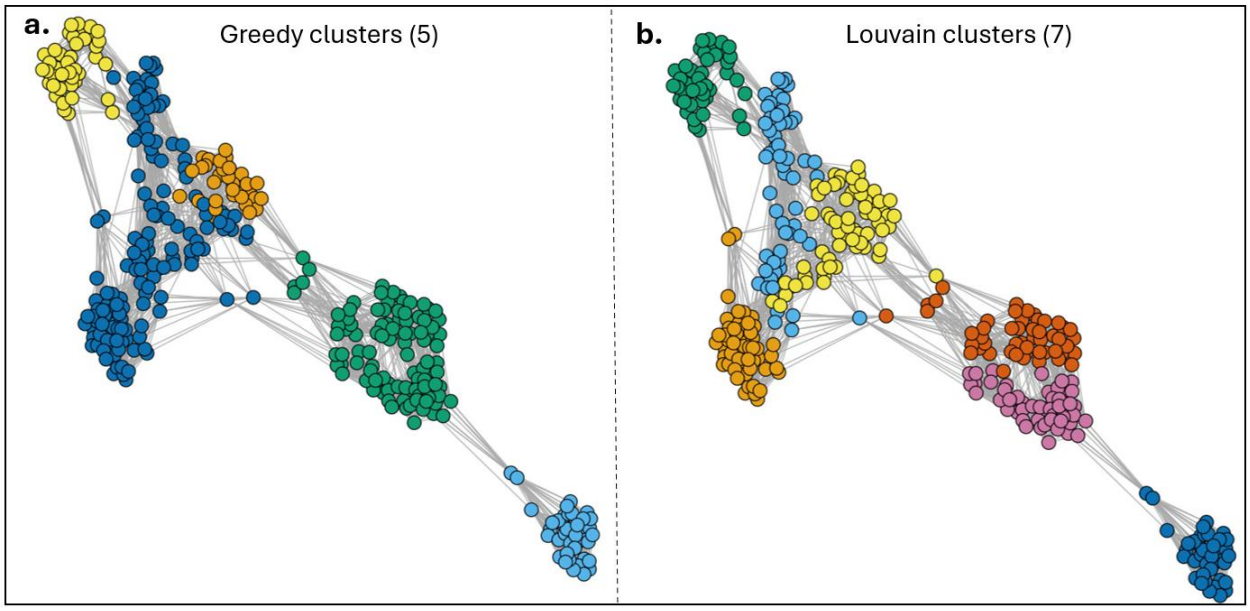


Figure 20. Greedy and Louvain clustering of the KNN network created in the NPDR-URF feature space with URFP distance. (**a**) Greedy modularity produces five clusters, most of which are spatially separated in the network. (**b**) Louvain clustering results in seven clusters in the selected-features space for salt content.

Even in this reduced features space, Louvain clustering is finding substructure related to salt content, carbon dioxide concentration, and environmental dataset labels. For instance, the substructure detected in clusters *VI* and *VII* seem to be due to a difference in chemical composition, namely bicarbonates versus sulfates and chlorides, respectively (Fig. 21b). The larger cluster to which these samples belong, treated as one cluster by Greedy modularity (compare Fig. 20a green cluster and Fig. 20b red and pink clusters), is composed of 2% CO_2 (g) samples. Furthermore, cluster *IV* is composed of samples mostly in the *Europa biotic* or *Europa abiotic* datasets (Fig. 21d), meaning they have been prepared with sulfate and chloride-rich brines and similar non-biogenic organics.

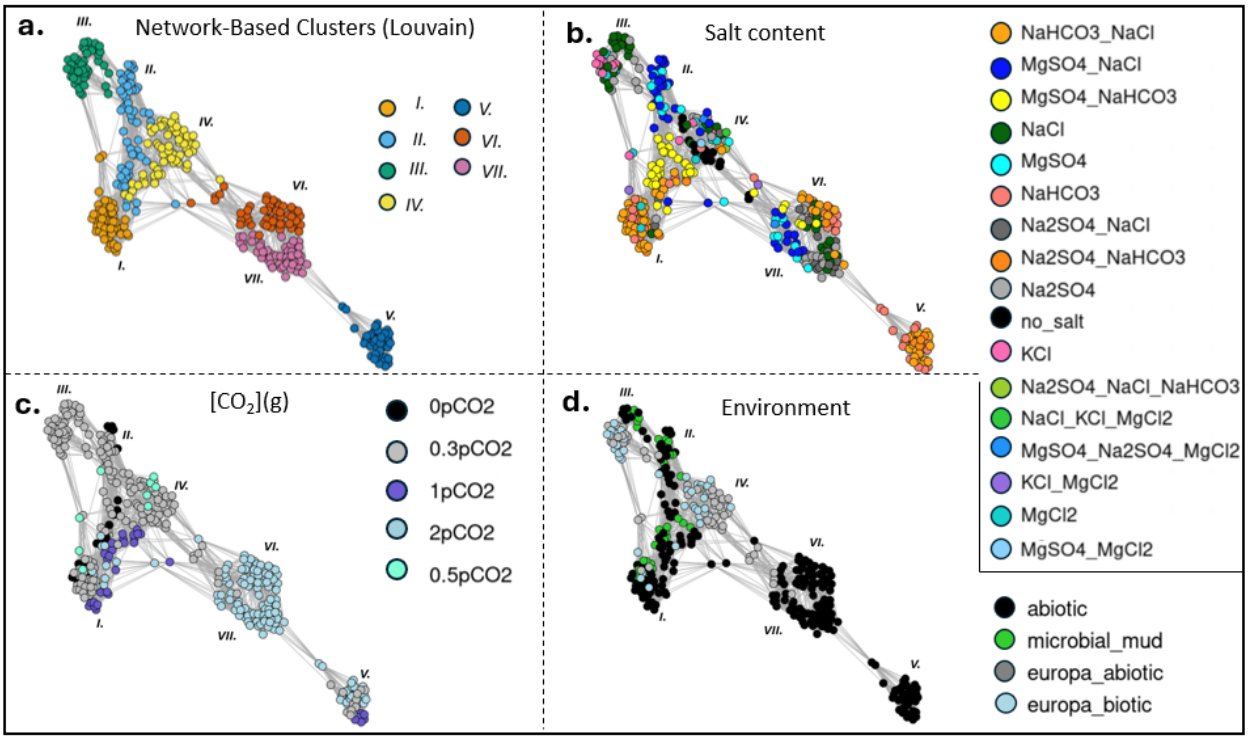


Figure 21. Comparison of salt content, [CO_2], and environmental dataset in the NPDR-LURF feature space Louvain clusters of the KNN network created using the URFP distance. (**a**) Louvain-detected lusters presented in Fig. 20 are numbered. (**b**) Salt content differences are responsible for clusters *VI* (bicarbonate dominated) and *VII* (sulfate and chloride dominated). (**c**) The larger cluster that clusters *VI* and *VII* belong to has the same carbon dioxide concentration of 2%. (**d**) The substructure detected in clusters *II* and *IV* is due to differences in environmental dataset.

4.4 Supervised Learning Results for Bulk Salt Composition

In this section we discuss the results of global-NPDR-URF feature selection and RF model prediction for bulk salt composition. There are 16 different bulk salt compositions in our dataset, although several have so few samples that those classes are not present in the test dataset (Fig. 22). These classes are included to present a further challenge to the classifier. Using an 80:20 train:test split results in three classes having fewer than 5 samples to train on (Fig. 22). Noting that there is considerable overlap between some classes, since for example, the *MgSO4* class (royal blue samples, Fig. 22) and the *MgSO4_NaHCO3* (yellow samples, Fig. 22) class

both share MgSO4 salt, we expect similarities between samples in these classes and potential for model confounding. Indeed, these samples can cluster together (see Fig. 21b, cluster *II*).

RF classification of bulk salt content using the NPDR-URF selected features yields a training accuracy of 66.6% and a test accuracy of 63.4%, considerably lower than the models created using two classes with more samples in each class (Fig. 23). However, as we discuss below, the model is actually better than the accuracy would indicate because it exactly detects the presence of chemical components. In this figure, samples that have been correctly predicted appear on the diagonal. Sample predictions appear as colored elements in the rows. For example, a correct prediction of *KCl* brines appear on the diagonal (bolded pink boxes, Fig. 23a and b), while incorrect *KCl* classifications appear as color mismatches on the first row. In the training data, two *KCl* brines are incorrectly classified as *NaCl* (Fig. 23a, first row pink box under the dark green NaCl column label). Misclassifications can be seen to contain common salt components or ionic content. Despite the modest reported test accuracy, all misclassifications contain significant common ionic content as the actual class. We define a "hit" for this model as a prediction that contains at least half of the ions as the actual class, even if the predicted label is not an exact match. By this definition, each prediction for the test samples is at least partly true, resulting in a test "hit accuracy" of 100% (Fig. 23c), illustrating the capacity of the model to predict ionic content.

	KCl	KCl_ MgCl2	MgCl2	MgSO4	MgSO4_ Na2SO4_ MgCl2	MgSO4_ NaCl	MgSO4_ NaHCO3	Na2SO4	Na2SO4_ NaCl	Na2SO4_ NaCl_ NaHCO3	Na2SO4_ NaHCO3	NaCl	NaCl_ KCl_ MgCl2	NaHCO3	NaHCO3 _NaCl	no_salt
training	9	1	9	19	2	32	28	26	15	1	11	39	1	27	56	14
testing	4	1	1	4	0	8	5	8	3	0	4	5	0	8	16	3

Figure 22. Training and test data samples for the bulk salt content RF classifier. There are 16 different brine salt compositions in the dataset. *NaHCO3_NaCl* is the class with the most training samples, while *KCl_MgCl2*, *MgCl2*, and the three-component brines have the fewest training and test samples.

Even though there is a 25.0% test classification accuracy for *MgSO4* prediction (Fig. 23b, light blue samples), we can accept that the model has learned something about the ionic content of these brines because they are all predicted to be *MgSO4_NaCl* (Fig. 23b, blue samples), a class which is expected to have geochemical similarities with *MgSO4*. The model correctly classifies the single *MgCl2* sample in the test data (Fig. 23b, diagonal bold turquoise box). These samples likely have a unique isotopic signature in the measured volatile CO_2 that has been learned by the model, even though only 9 samples were available for training and all but one of those was misclassified during model training validation (Fig. 23a, turquoise boxes off the diagonal in the *MgCl2* row). Notably, no other samples are incorrectly predicted to be *MgCl2* in the test data.

a. ↓ actual

Train_acc = 66.6% Predicted →	KCl	KCl_ MgCl2	MgCl2	MgSO4	MgSO4_ Na2SO4_ MgCl2	MgSO4 _NaCl	MgSO4_ NaHCO3	Na2SO4	Na2SO4_ NaCl	Na2SO4_ NaCl_ NaHCO3	Na2SO4_ NaHCO3	NaCl	NaCl_ KCl_ MgCl2	NaHCO3	NaHCO3 _NaCl	no_salt
KCl	4	0	1	0	0	0	0	0	0	0	0	2	0	1	1	0
KCl_MgCl2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
MgCl2	1	0	1	0	0	0	0	0	0	0	0	5	0	2	0	0
MgSO4	0	0	0	11	0	8	0	0	0	0	0	0	0	0	0	0
MgSO4_Na2SO4 _MgCl2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
MgSO4_NaCl	0	0	0	5	3	24	0	0	0	0	0	0	0	0	0	0
MgSO4_ NaHCO3	0	0	0	0	0	3	25	0	0	0	0	0	0	0	0	0
Na2SO4	1	0	2	0	0	0	0	12	7	0	0	4	0	0	0	0
Na2SO4_NaCl	0	0	0	0	0	1	0	5	7	0	0	2	0	0	0	0
Na2SO4_NaCl_ NaHCO3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Na2SO4_ NaHCO3	0	0	0	0	0	0	0	0	0	0	3	0	0	8	0	0
NaCl	1	0	2	0	0	0	0	3	1	0	1	31	0	0	0	0
NaCl_ KCl_MgCl2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
NaHCO3	0	0	1	0	0	0	0	1	0	0	1	0	0	12	12	0
NaHCO3_ NaCl	0	0	0	0	0	0	0	0	0	1	3	0	0	3	49	0
no_salt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14
class_acc	0.44	0.00	0.11	0.58	0.00	0.75	0.89	0.46	0.47	0.00	0.27	0.80	0.00	0.44	0.88	1.00

b. ↓ actual

Test_acc = 63.4% Predicted →	KCl	KCl_ MgCl2	MgCl2	MgSO4	MgSO4_ Na2SO4_ MgCl2	MgSO4 _NaCl	MgSO4_ NaHCO3	Na2SO4	Na2SO4_ NaCl	Na2SO4_ NaCl_ NaHCO3	Na2SO4_ NaHCO3	NaCl	NaCl_ KCl_ MgCl2	NaHCO3	NaHCO3 _NaCl	no_salt
KCl	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
KCl_MgCl2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
MgCl2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MgSO4	0	0	0	1	0	2	1	0	0	0	0	0	0	0	0	0
MgSO4_Na2SO4 _MgCl2	0	0	0	0	-	0	1	0	0	0	0	0	0	0	0	0
MgSO4_NaCl	0	0	0	0	0	6	2	0	0	0	0	0	0	0	0	0
MgSO4_ NaHCO3	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
Na2SO4	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0	0
Na2SO4_NaCl	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0
Na2SO4_NaCl_ NaHCO3	0	0	0	0	0	0	0	0	0	-	0	1	0	0	0	0
Na2SO4_ NaHCO3	0	0	0	0	0	0	0	0	0	0	1	0	0	1	2	0
NaCl	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
NaCl_ KCl_MgCl2	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0
NaHCO3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	0
NaHCO3_ NaCl	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	0
no_salt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
class acc	0.25	0.00	1.00	0.25	-	0.75	1.00	0.46	0.67	-	0.25	1.00	-	0.44	0.25	1.00

c.

Test hit acc*	KCl	KCl_ MgCl2	MgCl2	MgSO4	MgSO4_ Na2SO4_ MgCl2	MgSO4_ NaCl	MgSO4_ NaHCO3	Na2SO4	Na2SO4_ NaCl	Na2SO4_ NaCl_ NaHCO3	Na2SO4_ NaHCO3	NaCl	NaCl_ KCl_ MgCl2	NaHCO3	NaHCO3 _NaCl	no_salt
	1.00	0.00	1.00	1.00	-	1.00	1.00	1.00	1.00	-	1.00	1.00	-	1.00	1.00	1.00

*Hit: half of the ions in the actual class are present in the predicted class

Figure 23. RF classification of bulk salt content using NPDR-URF (**a**) Training data classification table for RF bulk salt content prediction trained in the NPDR-URF selected features space. The RF classifier for 16 different bulk brine compositions yields a train accuracy of 66.6%. Correct predictions are shown in bold boxes on the diagonal. Model predictions are shown as colored boxes in the rows; color mismatches indicate misclassifications. The class prediction accuracy for each brine composition is indicated beneath the table. The high class accuracies for some salts indicate strong predictive signals for brines that contain MgSO4 and NaHCO3, or that have salinity = 0 (absence of salt signals). (**b**) The RF brine composition model yields a test accuracy of 63.4%. there are many samples correctly classified. (**c**) All of the misclassified samples make sense because they have a similar ionic content with the incorrectly predicted class (compare *MgSO4* and *MgSO4_NaCl*).

To verify that the NPDR-URF selected features for brine salt content capture important predictive information, we train compare the RF classifier trained in this feature space with a full variable model. There are 28 NPDR-selected features used in both the Louvain clustering discussed in Sec. 4.3.2 and the RF classifier for salt content (Fig. 24). The top ten NPDR features are expected to be more independent than the top ten RF features reported by the full-variable model (Fig. 24, compare first and second columns) (Clough et al., 2025). For comparison, the NPDR-URF features ranked by RF are shown in the third column (Fig. 24). The RF classifier created in the NPDR-URF selected features space for bulk salt content had a slightly higher test accuracy (63.4%) than the model trained in the full variable space (61.9%) (Fig. 23). Although this is a modest increase, it indicates that NPDR-URF successfully identifies important predictors for salt content that perform similarly well to the model trained in the full variable space.

NPDR-URF features	RF importance (full variable space)	RF importance (NPDR-URF features)
1. avg_r$R^{46}CO_2/^{44}CO_2$	avg_calib_$\delta^{18}O$	avg_calib_$\delta^{18}O$
2. avg_$R^{46}CO_2/^{44}CO_2$	avg_$\delta^{18}O/\delta^{13}C$	avg_$\delta^{18}O/\delta^{13}C$
3. avg_r$\delta^{46}CO_2$	sd_$\delta^{13}C$	sd_$\delta^{13}C$
4. avg_$\delta^{46}CO_2$	sd_$R^{45}CO_2/^{44}CO_2$	sd_$R^{45}CO_2/^{44}CO_2$
5. avg_$R^{18}O/^{16}O$	avg_$R^{18}O/^{16}O$	avg_$R^{18}O/^{16}O$
6. avg_$\delta^{18}O$	avg_$\delta^{17}O$	avg_$\delta^{17}O$
7. avg_$R^{17}O/^{16}O$	avg_$R^{17}O/^{16}O$	avg_$R^{17}O/^{16}O$
8. avg_$\delta^{17}O$	sd_r$R^{45}CO_2/^{44}CO_2$	sd_$R^{13}C/^{12}C$
9. avg_calib_$\delta^{18}O$	avg_$\delta^{18}O$	sd_r$\delta^{45}CO_2$
10. avg_$\delta^{18}O/\delta^{13}C$	sd_r$R^{45}CO_2/^{44}CO_2$	avg_$\delta^{18}O$
... (28 features)	... (104 features)	... (28 features)

Figure 24. Comparison of the top ten features for salt content reported by NPDR-URF, RF importance, and RF importance when trained using NPDR-URF features. NPDR-URF selects 28 total features. When given all 104 variables, RF variable importance may rank highly correlated features in the top ten when correlation is not accounted for before model training. The top ten RF variable importance NPDR-URF features are different than the NPDR-URF rankings. In each case the top ten features are all related to isotope ratios (mass spectrometry features).

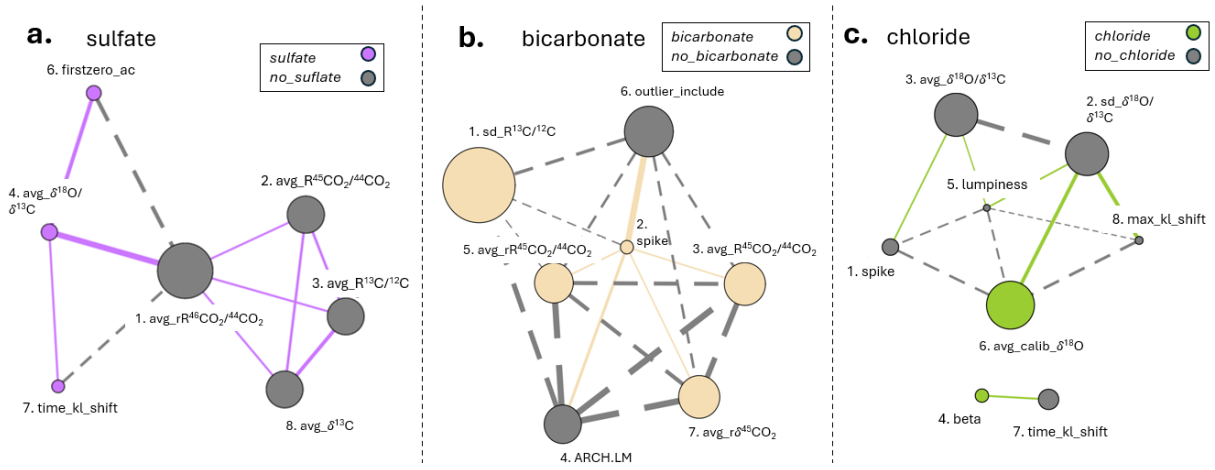


Figure 25. RAINs for sulfate, bicarbonate, and chloride NPDR-LURF selected features. Node size indicates the magnitude of the variable main effect and edge thickness indicates the interaction magnitude between variables. Node and edge color indicates the direction of the effect. Global-NPDR feature ranks are indicated by numbers next to each variable name. (**a**) The RAIN for sulfate prediction shows main effects that largely favor the negative prediction, with two large magnitude interactions. (**b**) The bicarbonate RAIN shows a large main effect for the top NPDR-LURF feature, *sd_R^{13}C/^{12}C* (Node 1) that favors *bicarbonate* prediction. There are three large-magnitude interactions that favor the negative prediction and one large interaction that favors bicarbonate prediction. (**c**) The RAIN for chloride prediction shows modest main effects and interactions, with mixed effects.

4.5 *Supervised Learning Results*

In this section we present the results of supervised learning feature selection and classification for major anions and volatile [CO_2]. Interpretable network visualization of two-outcome models are presented to illustrate how selected features work independently through main effects and in pairs to affect model predictions. These networks, called RAINs (Regression-based Association-Interaction Networks, see Sec. 2.3.3 for description), add to the interpretability of ML model predictions by visualizing variable effects (Fig 25). For example, NPDR-LURF selected predictors for sulfate have modest main effects, with the largest being the top NPDR-LURF feature (Fig. 25a, grey Node 1). The direction of this effect is toward the *no_sulfate* class, meaning it increases the probability of a negative prediction. However,

avg_rR^{46}*CO*$_2$/^{44}*CO*$_2$ also participates in a large magnitude interaction with *avg_d18O/d13C* (Fig. 25a, purple Node 4), which has a direction that favors sulfate prediction.

In contrast, the RAIN for bicarbonate prediction shows a very large main effect for the top NPDR-LURF selected feature, *sd_R*^{13}*C/*^{12}*C* (Fig. 25b, wheat-colored Node 1). This network has many large-magnitude interactions that mostly favor the negative prediction. All features except for *spike* have at least moderately large main effects, the majority of which increase the likelihood of *bicarbonate* prediction. The presence of large main effects and many interactions in the bicarbonate RAIN is expected to have implications for the ease of training and subsequent performance of classification and explanation methods. The RAIN for chloride has three variables with modest main effects with mixed prediction directions (Fig. 25c, Nodes 2, 3, and 6). There are several interactions in this network with moderate to small magnitudes. The lack of very large variable main effects and interactions for chloride is similarly expected to have implications for the ease of training classifiers for this outcome.

RF classifiers were trained in the NPDR-LURF selected features space and full variables spaces to predict chloride, bicarbonate, sulfate, and volatile CO_2 concentration using an 80:20 train:test split that preserves class imbalance (Fig. 26). For the anion models (sulfate, bicarbonate, and chloride prediction), the classes are reasonably balanced (Fig. 26 a, b, and c). For the CO_2 concentration model, however, there are few samples in the *0.5%_CO2*, *1%_CO2* and *0%_CO2* classes compared with *0.3%_CO2* and *2%_CO2* classes, potentially posing a challenge for ML methods for feature selection, classification, and explanation (Fig. 26d).

NPDR-LURF feature selection for sulfate yields eight predictors that are a mixture of features related to ^{18}O, ^{13}C, and TS features (Fig. 27a). While the sulfate classifier produces a test accuracy of 87.3%, the accuracy of the individual classes is different, with the negative

prediction (*no_sulfate*) having a much higher test class accuracy of 94.7% compared with 78.8% for the positive prediction (*sulfate*) (Fig. 27b). There are seven false negative predictions, which are *sulfate* brines incorrectly predicted to be *no_sulfate*, and two false positives, or *no_sulfate* brines predicted to be *sulfate*.

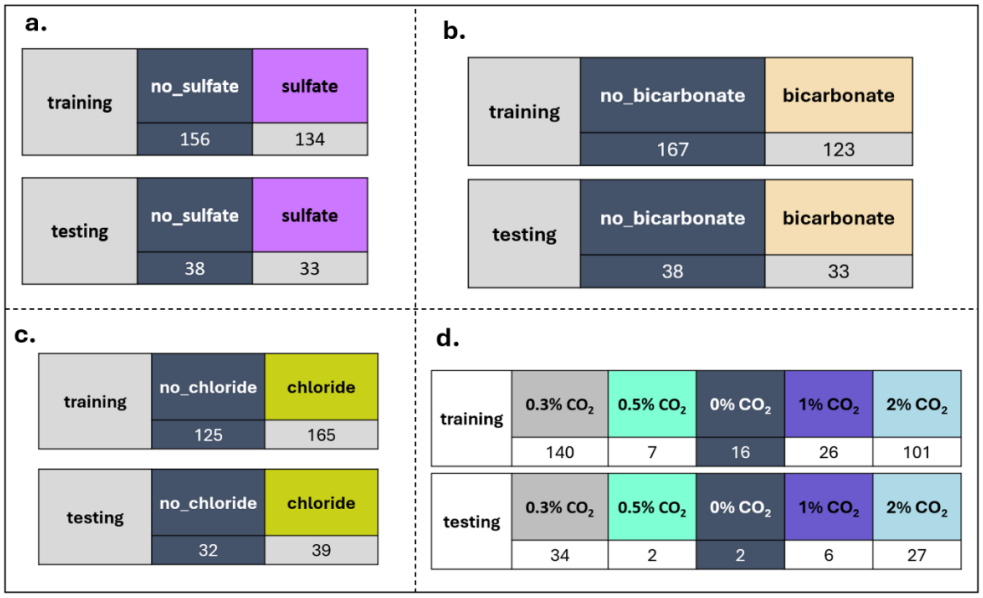


Figure 26. Training and testing samples for sulfate, bicarbonate, chloride, and volatile [CO_2] RF classifier development.

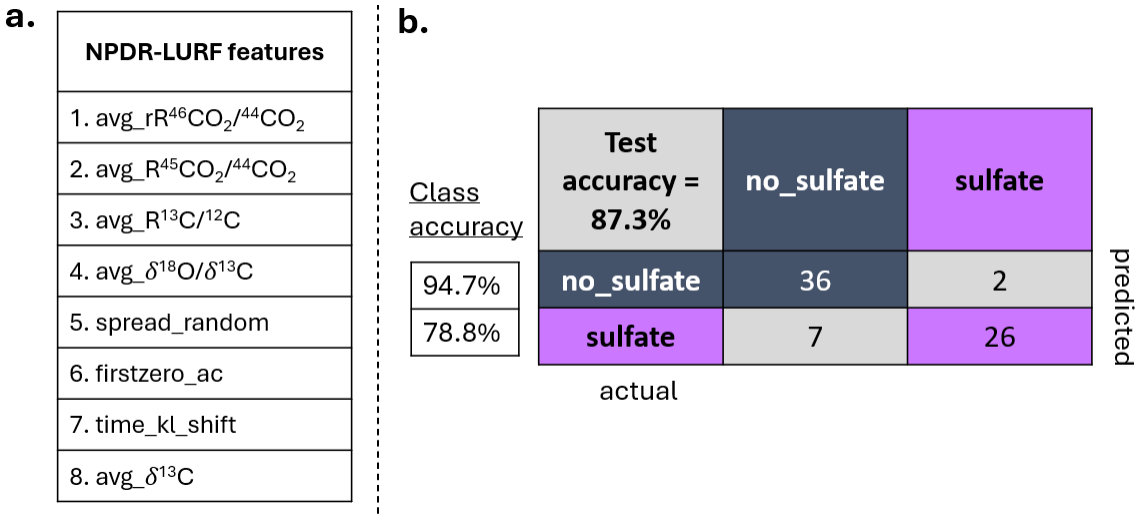


Figure 27. NPDR-LURF selected features and RF classification results for sulfate prediction. (**a**) Eight features are selected using NPDR-LURF and are a mixture of features related to O^{18}, C^{13} and TS features. (**b**) The RF classifier trained in the selected features space has a test accuracy of 87.3%, and the negative class has a much higher accuracy than the positive class. There are seven false negative predictions for the test samples (*sulfate* brines incorrectly predicted to be *no_sulfate*) and two false positives (*no_sulfate* brines predicted to be *sulfate*).

NPDR-LURF feature selection for bicarbonate classification produces seven features (Fig. 28a). The mass spectrometry features are predominately related to ^{13}C, and there are two TS features, *spike* and *outlier_include*. The RF model to predict bicarbonate ion presence in the OW analogue brines yields a high test accuracy of 95.8% (Fig. 28b). There are no false negatives in the test sample predictions, or *bicarbonate* samples predicted to be *no_bicarbonate*, and there are three false positives, or *no_bicarbonate* sample brines predicted to be *bicarbonate*.

Feature selection for the chloride model yields eight predictors, which are related to both ^{18}O and ^{13}C, and that include five TS features, a large number compared with the other anion predictions (Fig. 29a). The RF classifier for chloride anion presence in the brines produces a moderate test accuracy of 71.8%, with the positive class, *chloride*, showing a much higher test accuracy (84.6%) than the negative class, *no_chloride* (56.3%) (Fig. 29b). There are 14 false positives (*no_chloride* brines predicted to be *chloride*), while there are only six false negatives (*chloride* samples predicted to be *no_chloride*).

NPDR-LURF feature selection for volatile CO_2 concentration, which has five categories (see Fig. 26d), yields 18 predictors (Fig. 29a). Both TS and IRMS features are well-represented in this feature space. The RF model to classify CO_2 concentration yields a high test accuracy of 95.8%, with only test three samples being misclassified (Fig. 29b). All three misclassifications belong to one of two classes that have a higher CO_2 concentration relative to all the other test samples, the *1%_CO2* and *2%_CO2* classes.

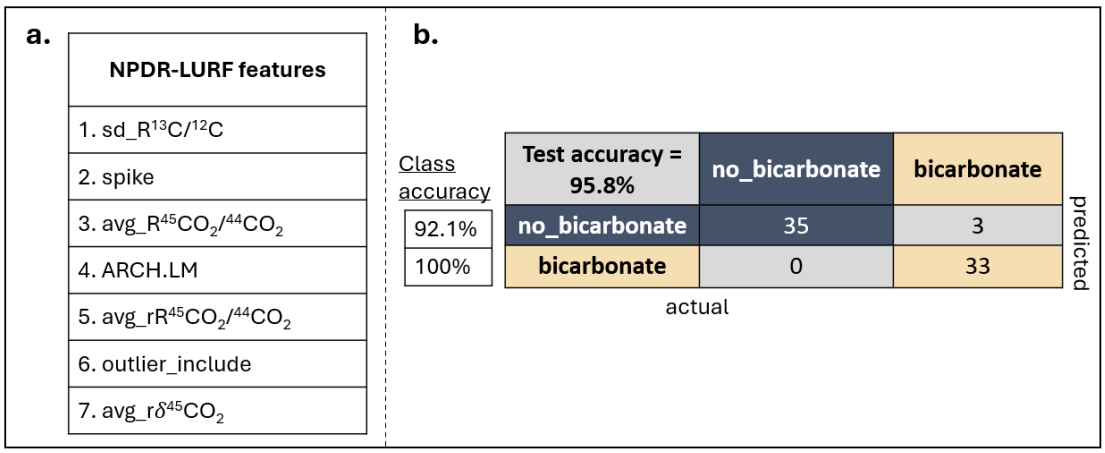


Figure 28. NPDR-LURF selected features and RF classification results for bicarbonate prediction. (**a**) Feature selection yields seven predictors for bicarbonate ion presence that are related to ^{13}C and include two time series features. (**b**) The RF classifier for bicarbonate trained on NPDR-LURF features yields a high test accuracy of 95.8%. There are no false negative predictions for the test samples (*bicarbonate* samples incorrectly predicted to be *no_bicarbonate*) and there are three false positives (*no_bicarbonate* brines predicted to be *bicarbonate*).

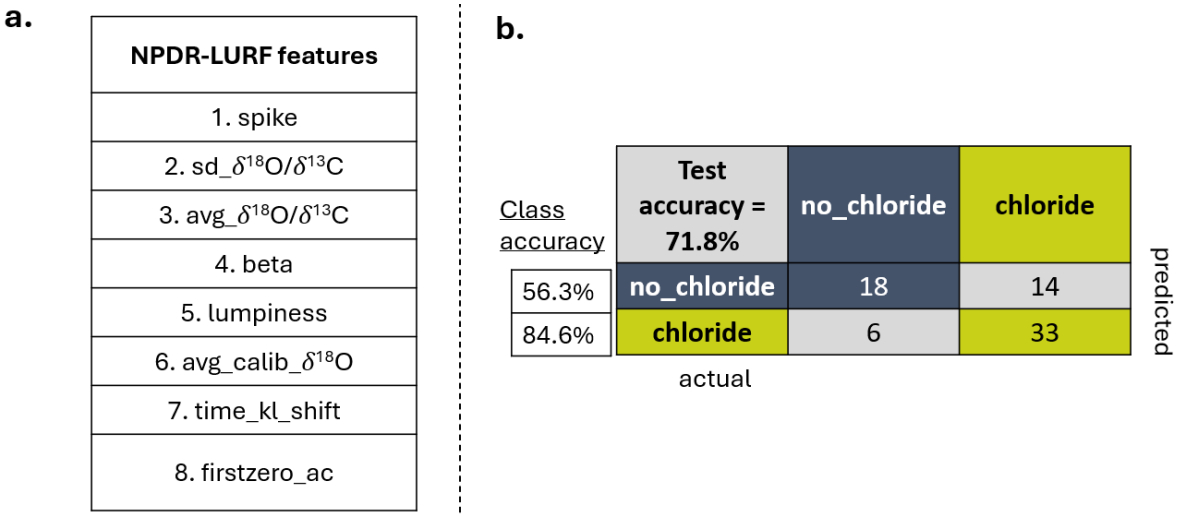


Figure 29. NPDR-LURF selected features and RF classification results for chloride prediction. (**a**) Feature selection for chloride prediction produces eight predictors that are enriched for TS features relative to IRMS features. (**b**) The RF model for chloride prediction yields a moderate test accuracy of 71.8%, with the positive prediction (*chloride*) displaying a much higher test class accuracy of 84.6% compared with the negative prediction (*no_chloride*); there are 14 false positives (*no_chloride* brines predicted to be *chloride*) and six false negatives (*chloride* samples predicted to be *no_chloride*).

a.

NPDR-LURF Selected Features for $[CO_2](g)$	
1. sd_$R^{46}CO_2/^{44}CO_2$	10. fluctuation
2. sd_$R^{45}CO_2/^{44}CO_2$	11. motif_entropy
3. sd_$rR^{45}CO_2/^{44}CO_2$	12. avg_$\delta^{18}O/\delta^{13}C$
4. sampenc	13. diff1_acf10
5. avg_$rR^{45}CO_2/^{44}CO_2$	14. spread_random
6. avg_$rR^{45}CO_2/^{44}CO_2$	15. crossing_points
7. walker_cross	16. sd_$r\delta^{45}CO_2/\delta^{44}CO_2$
8. garch_r2	17. max_kl_shift
9. garch_acf	18. sampen_first

b.

Class accuracy	Acc = 95.8 %	0.3% CO_2	0.5% CO_2	0% CO_2	1% CO_2	2% CO_2
100%	0.3% CO_2	34	0	0	0	0
100%	0.5% CO_2	0	2	0	0	0
100%	0% CO_2	0	0	2	0	0
83.3%	1% CO_2	0	0	0	5	1
92.6%	2% CO_2	0	0	0	2	25

actual

predicted

Figure 30. NPDR-LURF selected features and RF classification results for volatile CO_2 concentration prediction. (**a**) There are 18 features selected for CO_2 concentration classification. Both TS and IRMS features are well-represented in this feature space. (**b**) The RF model to classify CO_2 concentration yields a high test accuracy of 95.8%, with only test three samples being misclassified.

RF regression models were trained using NPDR-LURF selected features for pH (Fig. 31a) and ionic strength (Fig. 31b). These are the only predictions in which IRMS features related to ^{17}O are enriched in the selected features space. These feature spaces both contain many more IRMS features than TS features. The RF model to predict pH yields a mean absolute error (MAE) of 0.4 pH units, and there is sparse data in the middle of the plot (Fig. 32a). The RF model to predict brine ionic strength yields an MAE of 1.1M (Fig. 32b). This data is even more sparse than the pH data (see empty area around ionic strength = 4 M, Fig. 32b). The deviation from the 45-degree perfect fit line (red line, Fig. 32) is likely due to the small sample sizes, which can affect regression model performance. Even with a lack of data points, these models perform reasonably well and are expected to improve with improved measurement techniques and increased sample sizes, discussed more in Sec. 4.7.

a.

NPDR-LURF Selected Features for pH	
1. sd_$R^{13}C/^{12}C$	12. sd_$\delta^{45}CO_2/\delta^{44}CO_2$
2. sd_$R^{45}CO_2/^{44}CO_2$	13. sd_r$\delta^{45}CO_2/\delta^{44}CO_2$
3. spike	14. outlier_include
4. avg_$R^{17}O/^{16}O$	15. avg_$\delta^{45}CO_2/\delta^{44}CO_2$
5. avg_$R^{46}CO_2/^{44}CO_2$	16. avg_r$\delta^{45}CO_2/\delta^{44}CO_2$
6. avg_$R^{45}CO_2/^{44}CO_2$	17. avg_$\delta^{13}C$
7. avg_$R^{18}O/^{16}O$	18. avg_r$\delta^{46}CO_2/\delta^{44}CO_2$
8. garch_acf	19. avg_$\delta^{46}CO_2/\delta^{44}CO_2$
9. avg_$\delta^{18}O/\delta^{13}C$	20. avg_$\delta^{17}O$
10. ARCH.LM	21. avg_$\delta^{18}O$
11. sd_$\delta^{13}C$	22. lumpiness

b.

NPDR-LURF Selected Features for Ionic Strength (M)	
1. avg_$R^{17}O/^{16}O$	12. sd_$\delta^{13}C$
2. avg_$R^{18}O/^{16}O$	13. avg_$\delta^{17}O$
3. avg_$R^{46}CO_2/^{44}CO_2$	14. avg_r$\delta^{46}CO_2/\delta^{44}CO_2$
4. sd_$R^{13}C/^{12}C$	15. avg_$\delta^{46}CO_2/\delta^{44}CO_2$
5. avg_$R^{13}C/^{12}C$	16. avg_$\delta^{18}O$
6. avg_$R^{45}CO_2/^{44}CO_2$	17. avg_$\delta^{45}CO_2/\delta^{44}CO_2$
7. avg_r$R^{46}CO_2/^{44}CO_2$	18. avg_r$\delta^{45}CO_2/\delta^{44}CO_2$
8. garch_acf	19. avg_$\delta^{13}C$
9. avg_$\delta^{18}O/\delta^{13}C$	20. spread_random
10. walker_cross	21. time_kl_shift
11. outlier_include	-

Figure 31. NPDR-LURF selected features for pH and ionic strength. (**a**) Twenty-two features are selected as predictors for pH, dominated by IRMS variables related to different isotopes, such as ^{17}O, ^{18}O, and ^{13}C. (**b**) Twenty-one features are selected as predictors for ionic strength and are similarly dominated by IRMS variables related to ^{17}O, ^{18}O, and ^{13}C.

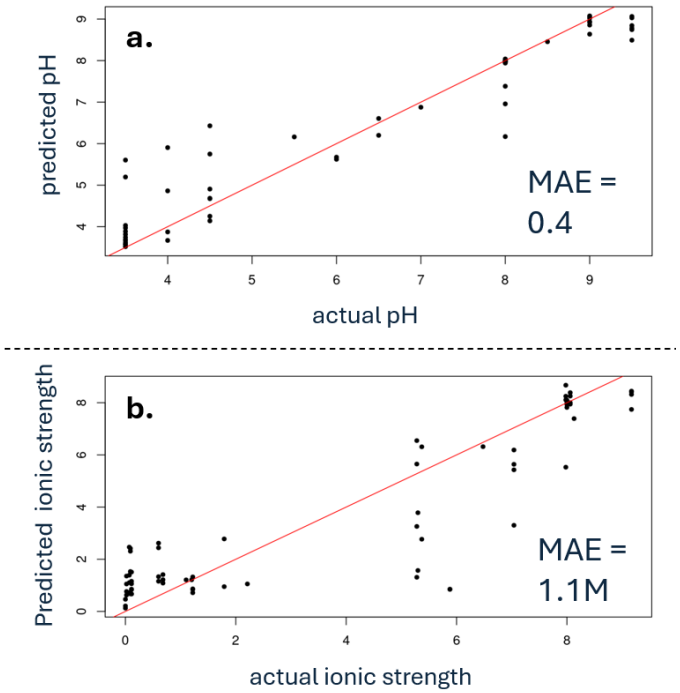


Figure 32. RF regression results for pH and ionic strength prediction. The 45-degree red line indicates a perfect match between predicted values and actual values. Model performance is given by the mean absolute error (MAE) in pH units or ionic strength (M). (**a**) The RF regression model performs moderately well for pH prediction, even with sparse data, yielding an MAE of 0.4 pH units. (**b**) The ionic strength RF model performs less well with an MAE of 1.1 M. This data is even more sparse than the pH data (see empty area around ionic strength = 4 M).

4.6 Local-NPDR False Prediction Diagnostics

Local-NPDR for individual sample prediction explanation and false prediction diagnostics was applied to models for bicarbonate and CO_2 concentration to illustrate the feasibility of the method on real seawater chemistry datasets. This method was previously described in Sec. 3.2.1. First, global-NPDR-LURF feature importance scores for this dataset were created using scaled versions of these features. This ensures that the feature importance scores are all on the same order of magnitude, and no single variable can become essentially the only contributor to the total local score (TLS), a metric by which false predictions can be identified.

T-tests for statistical significance between TLS for true and false predictions in the bicarbonate and $[CO_2](g)$ training datasets yielded significant P-values for bicarbonate ($P = 1.1 \cdot 10^{-5}$), but not for $[CO_2](g)$ prediction (P = 0.08) (Fig. 33a). When all $[CO_2](g)$ samples (train and test) were included in the analysis, the TLS for true and false predictions were statistically different yielding a P-value of 0.005 (Fig. 33b).

For the bicarbonate dataset, the behavior of true and false predictions can be further analyzed by considering the types of predictions that can occur. For two-outcome classifiers like the bicarbonate model, we consider four prediction types: true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs). Combining information like total local-NPDR variable importance scores and RF prediction probabilities allows for a more sophisticated consideration of individual sample predictions in terms of both explainability and interpretability (Figs. 34 and 35). For well-trained models with global-NPDR selected feature spaces, we find that total local-NPDR scores (Figs. 34a and 35a) and RF prediction probabilities (Figs. 34b and 35b) are higher for true predictions than for false predictions.

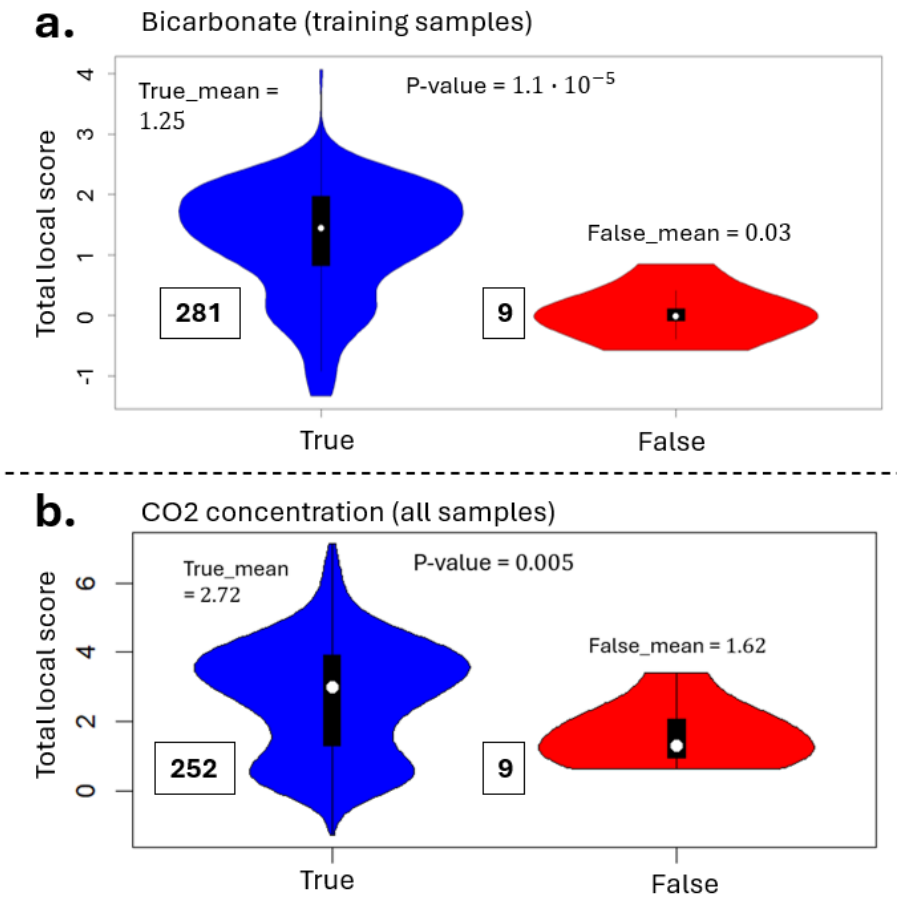


Figure 33. Total local-NPDR variable importance scores for true and false predictions using the bicarbonate and carbon dioxide RF models. True prediction scores are shown in blue and false prediction scores are red. The number of samples in each prediction category are boxed next to the plots. (**a**) TLS for bicarbonate training samples are higher in true predictions than false predictions ($P = 1.1 \cdot 10^{-5}$). (**b**) TLS for all samples in the CO2 concentration prediction model are higher for true predictions than false predictions (P = 0.005).

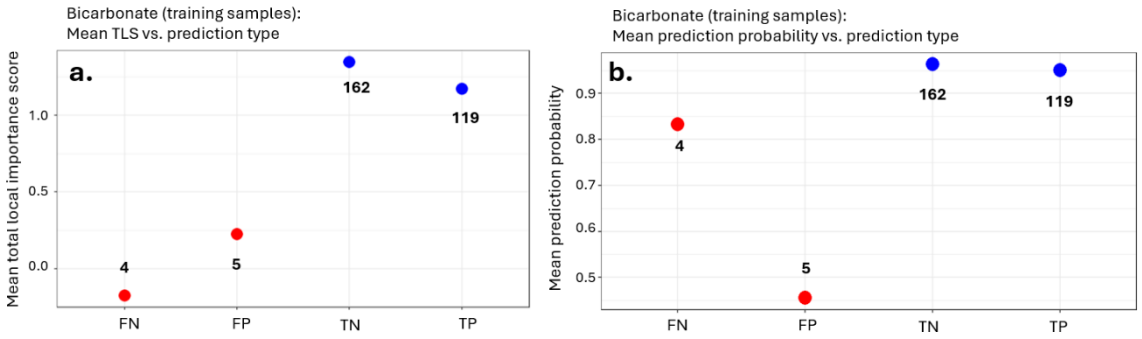


Figure 34. Bicarbonate training sample mean local-NPDR total scores and RF prediction probabilities for four prediction types. False prediction values are shown in red and true prediction values are blue. The number of samples in each prediction category are indicated in bold in the plots. (**a**) Mean TLSs for FN and FP predictions are over a point lower than for true predictions. (**b**) Mean RF prediction probability for TN and TP predictions are higher (over 90%) than for FP and FN predictions, with FP predictions having much lower probabilities than FNs.

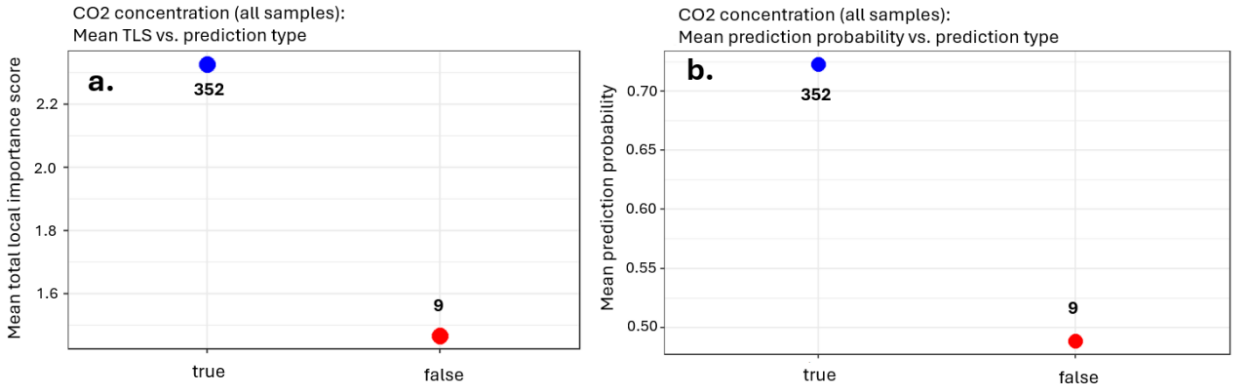


Figure 35. Mean local-NPDR total scores and RF prediction probabilities for four prediction types in carbon dioxide concentration dataset samples. False prediction values are shown in red and true prediction values are blue. The number of samples in each prediction category are indicated in bold in the plots. (**a**) Mean TLS for true predictions are higher than for false predictions. (**b**) Mean RF prediction probabilities are much lower for false predictions than for

Because the bicarbonate selected features space is smaller than that of the CO_2 concentration model, containing seven features compared with the 18 variables, we use the bicarbonate data to analyze local-NPDR scores for individual features for the four different prediction types (Fig. 36). For TP predictions, the correct prediction of bicarbonate anions, all feature importance scores are on average positive except for *spike*, which is slightly negative (Fig. 36a). For TN predictions, there are more variables with small negative scores, but the larger magnitude positive variables result in an overall positive score (Fig. 36b). The results are more complicated for the FP and FN predictions, where the variable *ARCH.LM* has a large magnitude positive score for FNs on average and a negative score for FP predictions (compare Fig. 36c and d, purple bar). This type of information can be used to assess prediction trustworthiness, discussed more in Sec. 4.7.

To illustrate what local-NPDR variable importance scores look like for individual samples in the test data, we graph local importance scores for true and false predictions in the bicarbonate and carbon dioxide concentration datasets (Figs. 37 and 38). For a true *bicarbonate* prediction, a TP, several local-NPDR feature importance scores are large and positive (blue supporting features, Fig. 37a), and the TLS is 1.5. Additionally, the RF model prediction probability is 99.9%, providing this positive *bicarbonate* prediction with high model confidence. This prediction is therefore accepted as a likely true positive prediction. For an FP (incorrect *bicarbonate* prediction), the TLS is 0.1 and three out of seven globally important features for bicarbonate detection have negative local-NPDR importance scores (red contradicting features, Fig. 37b). Furthermore, the RF prediction probability is below 50%, meaning that the classifier and probability forest disagree, an indication of low model confidence in the predicted label. This test sample is therefore quarantined as a likely false prediction.

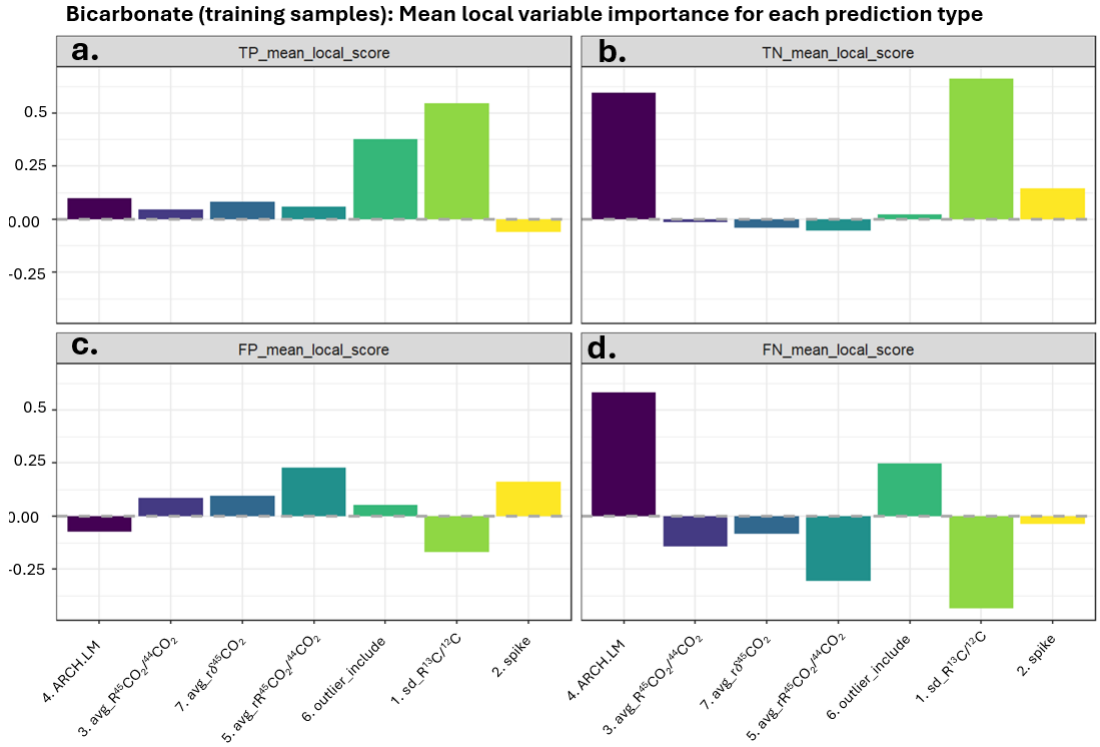


Figure 36. Average local-NPDR feature importance for four prediction types in the bicarbonate dataset. Global-NPDR feature importance ranks are indicated by numbers next to the feature name. (**a**) TP predictions have largely positive average local-NPDR scores. (**b)** TN predictions have three variables that have on average negative local-NPDR scores, while *ARCH.LM* and *sd_*$R^{13}C/^{12}C$ are large and positive on average. (**c**) There are several variables with average positive local-NPDR scores for FP predictions, all related to ^{13}C. (**d**) For FN predictions, *ARCH.LM* has a large positive score on average, as does *outlier_include*, while the IRMS features have negative scores.

For the volatile carbon dioxide concentration test samples, there are five possible outcomes and 18 globally important features for predicting sample [CO_2](g). The TLS for a true classification of a test sample as 0.3% volatile carbon dioxide by volume is 3.5 and the RF probability is 89.1% (Fig. 37a). For a false classification of a *2%_CO2* sample as *1%_CO2*, the TLS is less than one and the RF probability is only 37.2%, with the *1%_CO2* class having a similar but lower probability of 26.6%. This sample is quarantined for having a low-confidence

prediction, though it seems likely the actual concentration of CO2 could be somewhere in-between the two values.

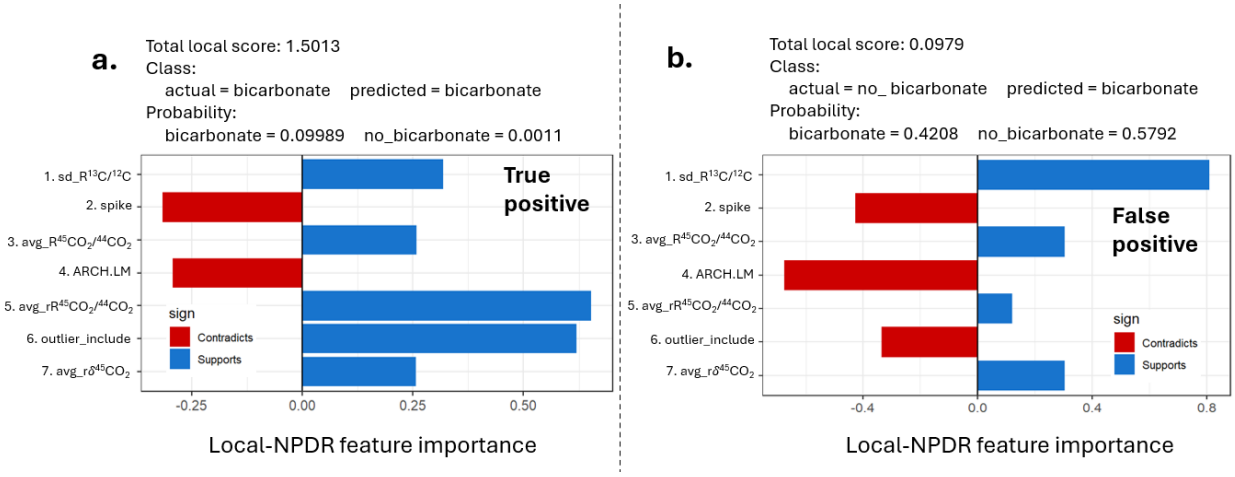


Figure 37. Local-NPDR feature importance scores for a true and a false bicarbonate detection. (**a**) The TLS for a TP (correct prediction of bicarbonate) is 1.5 and the RF probability is 99.9%. This sample is accepted as likely true prediction. (**b**) For a FP (incorrect prediction of bicarbonate), the TLS is much lower, 0.1, and the RF probability is below 50%. The disagreement between the RF classifier and probability forest indicates a low model confidence in this prediction, and this sample prediction is quarantined and due to the low TLS, labeled a likely false positive.

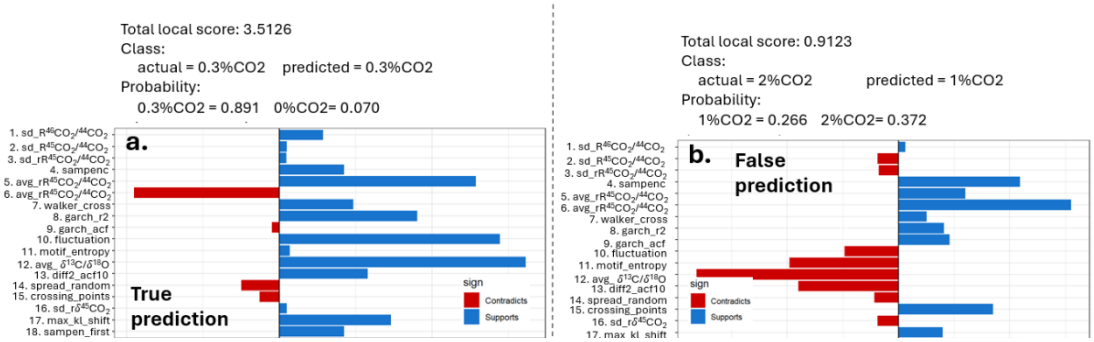


Figure 38. Local-NPDR feature importance scores for a true and a false classification of volatile carbon dioxide concentration. (**a**) The TLS for a true classification of a sample as 0.3% volatile carbon dioxide 3.5 and the RF probability is 89.1%. (**b**) For a false classification of a 2% CO_2 sample as 1% CO_2, the TLS is less than one and the RF probability is only 26.6%, with the 2% class having a nearly equal probability. This sample is quarantined for having a low-confidence prediction, though it seems likely the actual concentration of CO_2 could be somewhere in-between the two values.

Because the distribution of TLSs is dataset-dependent, in practice it is best to define thresholds for diagnosing possible false predictions and predictions with low model confidence. We examine the distribution of TLS in bicarbonate and carbon dioxide concentration datasets and define thresholds that will not result in an unacceptable number of quarantined test samples (Fig. 39). For bicarbonate, using a TLS threshold of 0.6 and a RF probability threshold of 80%, two out of three false predictions in the test samples can be diagnosed. Four true predictions with low TLS and low RF probabilities are also quarantined. For the carbon dioxide dataset, all three of the false predictions in the test dataset can be identified using a lower threshold of 70% for RF probability (due to the five classes) and a TLS threshold of 1.5. These thresholds also quarantine 21 true predictions (out of 71 total test samples). The inclusion of more samples is likely due to the lower prediction probabilities in the five-class model.

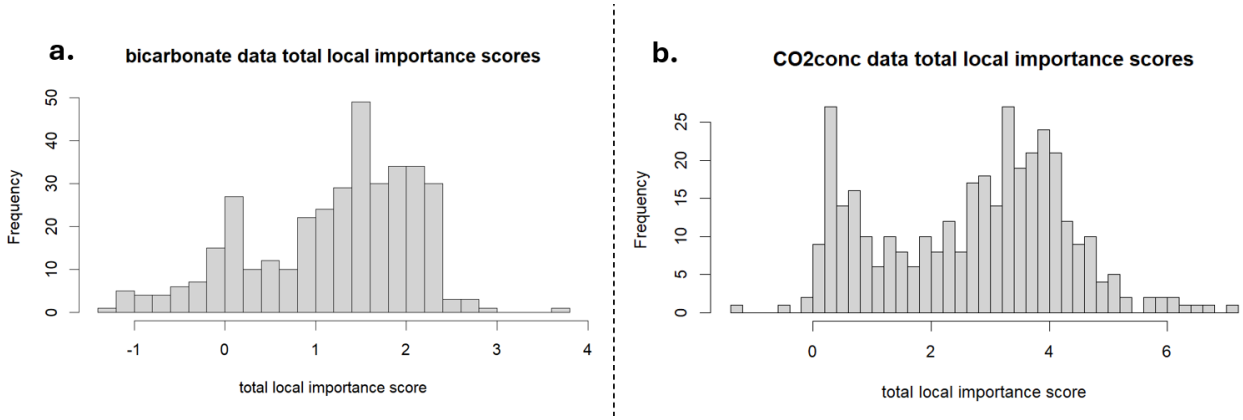


Figure 39. Histograms showing total local-NPDR feature importance score distributions in the bicarbonate and carbon dioxide concentration datasets. (**a**) Total local-NPDR scores in the bicarbonate data range from around -1.5 to greater than 3.5. Most samples have TLS between 1 and 2.5. (**b**) The distribution of TLS for carbon dioxide prediction ranges from less than -1 to greater than 6.

4.7 Conclusions

Geochemical investigations on icy OWs will seek to predict the subsurface seawater chemistry from orbital measurements of exosphere volatiles and ices. The results of our interpretable and explainable ML analysis for OW seawater chemistry prediction from the measurement of volatile CO_2 show that it is possible to predict many chemical parameters of interest from measuring the carbon and oxygen isotope ratios of volatiles evolved from the brines. Furthermore, our ML feature selection reduces high dimensional feature spaces to provide a small number of predictors whose interactions and main effects can be visualized in a RAIN (see Fig. 25). This informs model prediction explanation, especially in terms of whether predictions are likely to be true or false; we can see which effects dominate our feature space and therefore affect model prediction mechanisms and ultimately model prediction accuracy. Along with local-NPDR for false prediction diagnostics, these tools provide powerful insights into our data and models that allows us to develop an informed trust in our model predictions. Furthermore, and most importantly, it allows us to understand what our data and model limitations are; *i.e.*, what types of class predictions the model is most likely to fail on (see Figs. 27-30, class prediction accuracies, and Fig. 36, TLS for different prediction types), what the model probabilities look like for true versus false predictions, and what the local-NPDR total scores present for those predictions (see Figs. 34 and 35). Information from unsupervised learning methods like KNN-network Louvain clustering revealed complex inherent structure in the data in terms of analogue brine biogeochemistry (see Figs. 18 and 21) that allows further interpretability in understanding why certain features cause false predictions and provides insights into the reasons for model confounding.

Understanding the details of classifier performance in terms of predictor effects and class accuracy is important when considering the implementation of ML models to predict seawater chemistry for future missions to OWs. For example, understanding the reasons for the imbalance in our sulfate model class accuracies, where the *sulfate* test class accuracy is only 78.8% compared with 94.7% for *no_sulfate* (see Fig. 27b) comes from analyzing the RAIN for sulfate (Fig. 25a), where all large main effects increase the likelihood of a *no_sulfate* prediction. The RAIN for bicarbonate implies the RF model and local importance methods are likely to be successful (Fig. 25b): there are several variables with moderate to large main effects for both *bicarbonate* and *no_bicarbonate* classes and there are multiple large-magnitude interactions that favor both class predictions.

The chloride RAIN provides a few strong but mostly modest interactions as well as a few moderate-sized main effects that inform prediction, but it should be no surprise that this model has the lowest model prediction accuracy of 71.8%. The lack of global-NPDR-LURF selected features related to isotope ratios and relative enrichment in TS features indicates that this outcome (chloride ion presence) does not have many IRMS features that are good predictors. This makes sense, since the chloride ion contains neither oxygen nor carbon, so the direct measurement of carbon and oxygen isotopes from volatile CO_2 is unlikely to produce strong predictors for the detection of chloride in the sample brine. Additionally, the positive class, *chloride*, shows a much higher test accuracy (84.6%) than the negative class, *no_chloride* (56.3%) (Fig. 29b). This could be driven by the two large-magnitude interactions in the chloride RAIN that increase the likelihood of *chloride* prediction (see Fig. 25c, edges between Nodes 2 and 6, and Nodes 2 and 8). The addition of the TS features to the variable space therefore allows an increased ability to detect chloride.

The ability to select globally important features and analyze them in terms of individual samples (local feature importance) provides more than data insights; it can inform subsequent lines of experimental investigations. If certain features cause model confounding, perhaps there is a geochemical reason. For example, three features related to C^{13} in the bicarbonate selected features space (*avg_*$R^{45}CO_2/^{44}CO_2$, *avg_r*$R^{45}CO_2/^{44}CO_2$, and *avg_r*$\delta^{45}CO2$) have mean positive local-NPDR scores for true and false positive predictions and are negative on average for true negative and false negative predictions, indicating that this variable confuses the model for samples that do not contain bicarbonate. These features all have main effects that favor the bicarbonate-prediction direction (see Fig. 25b, Nodes 3, 5, and 7), potentially explaining their direction of bias and the reason for them being on average negative for the *no_bicarbonate* predictions (see Fig. 34). However, features that are discovered to be on average positive for true predictions and negative for false predictions can also open lines of experimental investigation, such as *sd_*$R^{13}C/^{12}C$ for the bicarbonate model (Fig. 34). For example, why should this feature be highly predictive while the others related to C^{13} are problematic for some samples? And for the sulfate prediction, features related to C^{13} are selected, indicating unique sulfate effects on carbon fractionation in the experimental brines. Therefore, if certain features are highly predictive, as global-NPDR suggests of O^{17}-related features and pH/ionic strength predictions, the reasons for this could and should be explained through geochemical reasoning and further investigations.

The addition of our novel local variable importance method, local-NPDR, adds false prediction diagnostics in addition to explainability for individual sample predictions (see Figs. 37 and 38). In combination with RF prediction probabilities, local-NPDR allows us to diagnose predictions that have either low model prediction confidence or low TLS, which indicate the prediction has an increased likelihood of being a false prediction (see Figs. 34 and 35). As

discussed above, local importance methods can provide information about what types of predictions the model is best at, and which variables can cause model confusion. This has implications for mission resource conservation in terms of not spending resources reporting predictions or further analyzing data that are unlikely to be true predictions or that have low prediction confidence. Although several or many true predictions can be quarantined along with false predictions, these either have low TLS or low prediction probabilities. Thresholds that incorporate an acceptable risk assessment are recommended; for example, in terms of high-risk predictions like potential OWs biosignatures, very strict probability and TLS thresholds may be implemented to keep the risk of a false prediction at a minimum. This will have the effect of quarantining potentially true predictions but acknowledges the inherent risk with such an ML prediction. In contrast, an ML prediction of bicarbonate, sulfate, or chloride ions is much lower risk and could trigger follow-on science measurements from other onboard instrumentation to confirm the chemical prediction as part of a possible implementation of science autonomy for future missions to OWs.

Future work will extend local-NPDR to continuous outcomes (regression models) and explore statistical thresholds for TLS and prediction probabilities that can be implemented in both simulated mission concepts and real field demonstrations. Limitations of local-NPDR and our classifiers include challenges presented by imbalanced, small datasets. Future experimental work will seek to add more samples to our ML investigations to further validate our existing models and extend them to more complicated and realistic geochemical scenarios appropriate for OWs.

CHAPTER 5
TOWARDS AUTONOMOUS SCIENCE IN ASTROBIOLOGY: SOFTWARE SOLUTIONS

This chapter discusses the applications of the ML methodology developed in the previous chapters. The first section covers the need for science autonomy in the field of astrobiology, risks and concerns that such applications of science autonomy raise, and how these may be mitigated through the deployment of explainable and interpretable ML methods. The second section discusses the use of the OW biosignature and seawater chemistry models presented in previous chapters in a simulated Enceladus mission concept which involves the implementation of automated decision making based on the ML prediction results. The last section presents real use cases of ML models trained with environmental sensor data using our explainable and interpretable methods and discusses how local-NPDR false prediction diagnostics will be used in a field demonstration.

5.1 Science Autonomy for Astrobiology

Recent advancements in data science and ML have made new analytical methods available to astrobiology research. However, as datasets become more complex and high-dimensional, more complicated models are needed to achieve high accuracy predictions. This often results in the use of "black box" ML models with limited bespoke methods for prediction explanation. For high-risk future applications of ML for science autonomy, such as future astrobiology and geochemical investigations of OWs, methods for assessing individual sample prediction trustworthiness are essential. There is much inherent risk in deploying a black box ML model to detect extraterrestrial biosignatures trained in a high dimensional feature space with no feature selection or induvial sample prediction analysis employed. To base mission-level

decisions upon such detections adds to this risk and may prove to be untenable. The ability to provide not just a prediction label and a quantitative assessment of prediction probability, but to explain whether an individual prediction is likely to be true or false and what features are contributing to that likelihood in terms of global and local NPDR feature importance scores provides powerful information for crucial mission decisions that must be made to support both mission and science interests. The ability to assess model prediction bias is essential; the NPDR global and local feature importance method provides a way of quantifying each feature's effects on the model and their effect direction. Because this method is independent of the model, it provides an unbiased assessment of an induvial sample. This can be combined with methods native to the model (if such methods exist, as they do in RF) to make more informed decisions in an automated manner. For example, local-RF and local-NPDR total scores for a sample can be analyzed and compared; if both agree, this bolsters confidence in the ultimate decision about this sample. This type of granular understanding of model performance in terms of selected features and how they contribute to model outcomes (and whether that contribution is likely to lead to a true or false prediction) leads to increased trust in ML for science autonomy and a more responsible deployment of ML on future astrobiology and geochemical missions to OWs.

5.2 MLMS: Simulated Enceladus Mission Concept

To address the need for accurate and explainable ML methods for astrobiology OWs isotopic data, we developed the machine learning for mass spectrometry (MLMS) suite of interpretable models and tools, and incorporated the QA/QC IRMS data processing tool into it. RF models for biosignatures and seawater chemistry trained on global-NPDR-LURF or URF selected features described in Chapters 2 and 4 were included in a simulated Enceladus mission concept (Yu & Hecht, 2024) demonstration in September 2024 that ran through the OnAIR

(Onboard Artificial Intelligence Research) software (Gizzi et al., 2022, 2025). The demonstration utilized simulated Enceladus orbital telemetry for eight orbiters (Fig. 40a) and a mothership with a simulated communication network between orbiting agents utilizing different possible communication bands for different data types and priorities. In the simulation, a pre-planned IRMS measurement of volatile CO_2 by one science agent results in the detection of a novel seawater chemistry and a possible biosignature. The detection is communicated to the mothership, whose onboard science autonomy software provides another orbiter with the location and maneuver parameters to make a follow-on measurement to confirm the measurement by the previous orbiter (Fig. 40b). The IRMS data used in the simulated follow-on measurement is run through the data processing pipeline in real time, and is depicted as a chromatogram in the image (Fig. 40b). The ML model output includes a variety of predictions and probabilities, including biosignature presence, bicarbonate and sulfate predictions, volatile CO_2 concentration, and an outlier score (Fig 40b).

This demonstration illustrates the feasibility of using these models in real time on a computer with limited resources, and how the implementation of science autonomy methods can enhance science return and preserve mission resources by making onboard decisions about collected data immediately after an experiment and data processing (*e.g.,* QA/QC), such as whether a biosignature or novel seawater chemistry is detected.

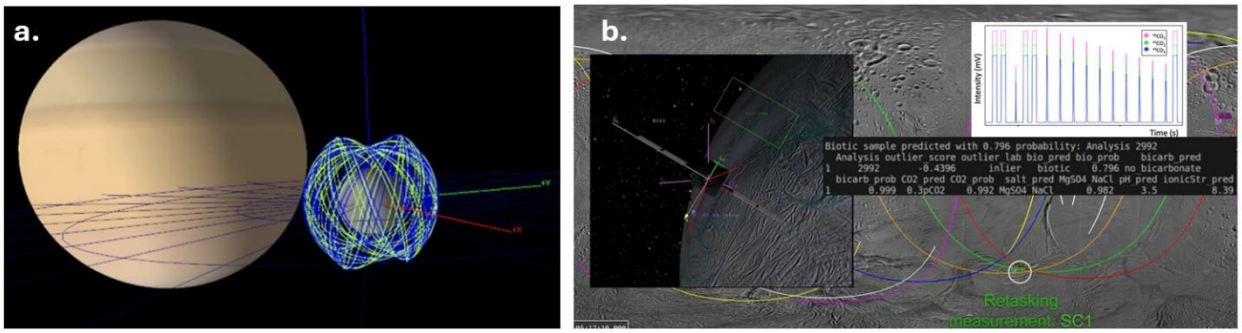


Figure 40. Illustration of simulated Enceladus mission concept with MLMS output. (**a**) Image of simulated orbital trajectories for the IRMS-enabled science agents in the simulation. (**b**) Illustration of an orbiter re-tasking to take a follow-on IRMS measurement based on another agent's assessment of a novel seawater with a possible biosignature. The IRMS data used in the simulated follow-on measurement is depicted as a chromatogram and the ML model output is shown below the graph.

5.3 Field Demonstrations of Machine Learning Algorithms

A supervised ML algorithm to detect evaporites was validated in a field demonstration on July 2, 2024 at the Great Salt Plains near Jet, Oklahoma (Fig. 41). The Great Salt Plains is a wildlife refuge and state park in Northwest Oklahoma consisting of kilometers of salt-encrusted flats, marked by both permanent and transient brine streams and ponds (K. S. Johnson, 1988; K. Johnson, 2013). Although not traditionally recognized as such, the site is an underutilized OW and Mars analogue site due to its hypersaline environment that includes active subsurface to surface brine transport and evaporation. Evaporation of brines and subsurface processes result in surface salt crusts and unique subsurface selenite crystals (Fig. 42a). The Great Salt Plains hosts novel extremophile life and geochemistry relevant to hypersaline planetary environments, providing an environment suited to habitability and geochemical investigations for both Mars and OWs (Postberg et al., 2009; Martínez & Renno, 2013; Toner et al., 2015; D. T. Vaniman et al., 2018; Glein & Waite, 2020; D. Vaniman et al., 2024).

ML models trained to detect evaporites using our explainable methodology successfully detected evaporite minerals at the Great Salt Plains in real time after taking a reflected-light spectroscopic measurement and processing the data. In addition to a spectrometer, other environmental sensors for atmospheric CO_2 concentration, pressure, temperature, visible and UV light collected data for further ML method development working on a software platform called YAAS (Yet Another Array of Sensors).

Similar RF models to predict sulfur rocks were demonstrated onboard Turtlebot rovers. In this demonstration, a positive sulfur detection was communicated in real time to another rover, which was called on to relocate and confirm the detection (Fig. 42b). The successful demonstration of RF classifiers in the field using real sensor data on resource-limited computers illustrates how our methods can run on flight-like computers and be used for a variety of scientific data products and predictions (Fig. 42c and d).

Supervised and unsupervised ML methods will be tested at Green Lakes State Park in New York (PI: Bethany Theiling, NASA-GSFC). ML methods designed for the environmental sensor array will be integrated with OnAIR and other intelligent software onboard an aerial flight platform to predict habitable conditions and detect important minerals. Local-NPDR false prediction diagnostics will be added to assess model performance and to inform subsequent data collection day-to-day in the field.

Figure 41. Photograph of July 2, 2024 Great Salt Plains ML demonstration. The sensor array is taking environmental measurements near a brine stream (blue box, foreground). The salt flats can be seen in the background.

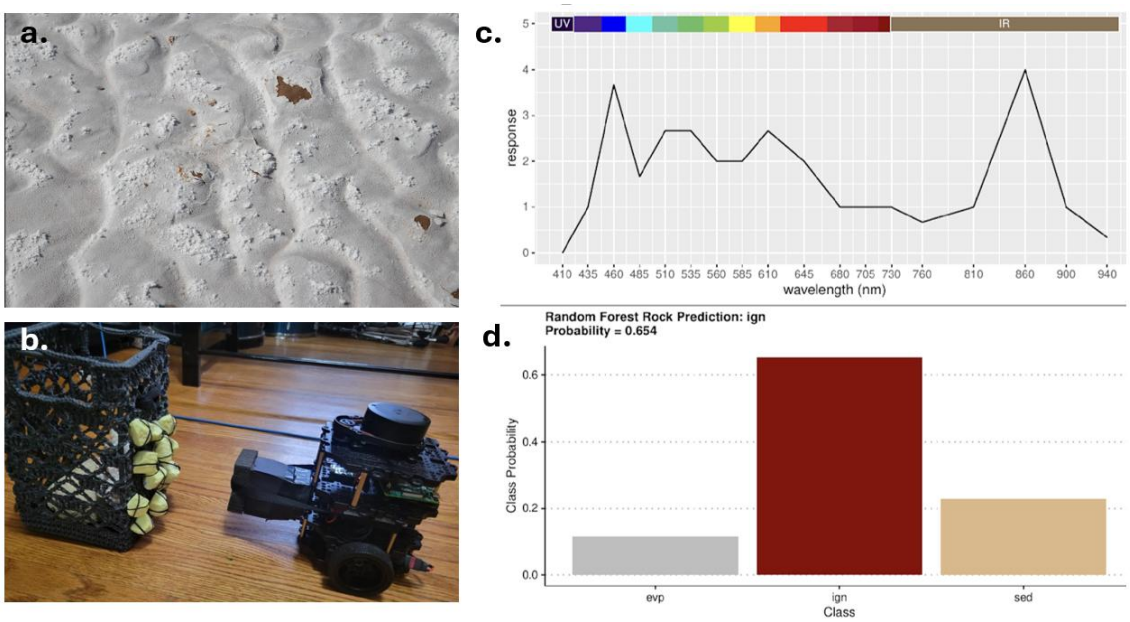


Figure 42. Example of ML output from YAAS and different mineral detections. (**a**) Evaporite-encrusted soil at the Great Salt Plains was confirmed and communicated in real time using RF classification. (**b**) Sulfur rocks were detected with a reflected light spectrometer (seen mounted to the Turtlebot) and RF classification. (**c**) An example of an 18-wavelegnth spectrum of a rock. (**d**) An example of the YAAS ML output for rock type classification, showing an igneous rock prediction.

CHAPTER 6

DISCUSSION AND CONCLUSIONS

6.1 Explainable ML Classifiers for OW Biosignatures and Seawater Chemistry Prediction

Explainable ML methods for OW biosignatures and seawater chemistry have been developed and validated for both simulated and real data domains. A high accuracy biosignature classifier was trained in a global-NPDR-LURF selected features space that achieved higher accuracy than the full-variable model and higher accuracy than models created using features selected with a traditional Manhattan distance metric. Results from ML chemical characterization of the OW analogue brines show that it is possible to create high-accuracy seawater chemistry models using the same feature selection and model training methodology.

Additional explainability of sample prediction can be provided by putting the local-NPDR feature importance scores in the context of main and interaction effects in a statistical interaction network. This interaction network (computed by the `regain` function in the NPDR R library) is a *pxp* matrix that contains each variable's main effect on the diagonal and interaction effects with other features on the off-diagonal entries (Appendix Fig. A.1). The centrality scores of the variables in this interaction network give the cumulative effects of the interactions and main effects of each variable (Appendix Table A.1) (Davis et al., 2010; Lareau et al., 2015). The ability to go beyond an importance score and see the individual effects of each variable and interaction provides additional explainability.

6.2 Local-NPDR Variable Importance for False Prediction Diagnosis

One of the goals of science is explanation, but black box ML models in scientific applications are antithetical to this goal. We developed a new ML interpretation tool, local-

NPDR, and tested it on three simulated datasets, two of which were imbalanced and one with decreased main effects, and one real biosignature dataset that had a small sample size and was similarly imbalanced and is known to have lower variable main effects. This tool can be used to explain why a single sample was predicted to be in a given class based on the variable importance weights. These weights are computed based on their ability to model the contrastive probability (hits and misses) for samples in the target sample's neighborhood. The sign and magnitude of the NPDR TLS of globally important variables can be used for diagnosing the likelihood of the sample being false. In conjunction with single sample RF prediction probability, local-NPDR can be a useful tool for future astrobiology missions. If resources are available, the consensus of local-NPDR and local-RF importance methods could improve the ability to detect false samples and discriminate them from flagged true predictions.

A potential limitation of local-NPDR is its independence of the classifier that one is trying to interpret. In other words, it may be preferrable to have a local explainer that explains the mechanisms of a particular classifier. A way to link local-NDPR to a specific classifier could be to use nested-CV feature selection (Parvandeh et al., 2020). Another limitation is the need to have a neighborhood of labeled samples, rather than a statistical model, to compute local-NPDR score of a single sample. As mentioned, local-NPDR could be combined with classifier-specific local importance methods, if available, to further improve the detection of false predictions.

Another practical challenge in ML training is class imbalance. Class imbalance can affect both global and local feature importance methods. Notably, the simulated dataset with class balance and an increased sample size contained zero noise features in the global-NPDR-LURF selected features space and was the best performing simulated dataset for local-NPDR analyses. An important area of future research will be to improve performance for imbalanced data.

Additional future work will extend NPDR and local-NPDR to non-tabular data such as images or time series using representation learning, both of which would enable a lightweight version of traditionally computationally intensive methods with added interpretability. In addition to improvements in ML algorithms and explainability, improvements in biosignature detection require close attention to data collection. Both scientific and ML models inherently include some degree of bias as they rely on initial assumptions or hypotheses before data collection. The resulting data with their biases are then used to identify generalizable patterns. Such assumptions, for example, make it challenging to develop fully agnostic biosignature models from laboratory data. However, both science and ML can guard against biases and discover more general models by making predictions on new data that test the limitations of the existing theory or model. For instance, precise measurements of blackbody radiation exposed shortcomings of classical mechanics (the ultraviolet catastrophe), ultimately leading to a more general atomic theory. Similarly, in ML, using data outside the training domain can identify limits of a model's validity for prediction. For example, depending on the initial training data, an OW biosignature model may not be valid for certain ranges and combinations of temperature, volatiles, pressure, and salinity.

Both global- and local-NPDR feature importance methods add to the interpretability tools currently available for ML methods, which is especially important for high-risk prediction domains. For such ML applications, it is expected that researchers (humans in the loop) will work with the ML algorithm output to ultimately reach the most informed conclusion possible about the prediction in question and mitigate risks associated with false predictions. It is essential to ensure the training data is representative of the deployment environment and its limitations understood, that models are responsibly trained and validated, that models limit noise features

and highly correlated features, and tools are included that can help understand individual predictions and diagnose false predictions. Producing a prediction is only a first step in ML; as in science, it is more important to understand the nature of the prediction, how it was made, and whether it should be trusted.

Additional validation of our interpretable ML methods was achieved through application to OW seawater chemistry models and in field demonstrations, where models trained in a similar way using environmental sensor data were deployed to predict minerals of interest for OW exploration, such as evaporites. Together, the rigorous validation of methods through both simulated and real datasets illustrates that our explainable ML approach is both practical and essential for future OWs astrobiology missions.

REFERENCES

Anderson, J. D., Lau, E. L., Sjogren, W. L., Schubert, G., & Moore, W. B. (1997). Europa's Differentiated Internal Structure: Inferences from Two Galileo Encounters. *Science*, *276*(5316), 1236–1239. https://doi.org/10.1126/science.276.5316.1236

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press. http://networksciencebook.com/

Barge, L. M., Rodriguez, L. E., Weber, J. M., & Theiling, B. P. (2022). Determining the "Biosignature Threshold" for Life Detection on Biotic, Abiotic, or Prebiotic Worlds. *Astrobiology*, *22*(4), 481–493. https://doi.org/10.1089/ast.2021.0079

Bhowmick, A. K., Meneni, K., Danisch, M., Guillaume, J.-L., & Mitra, B. (2020). LouvainNE: Hierarchical Louvain Method for High Quality and Scalable Network Embedding *. *WSDM '20: Proceedings of the 13th International Conference on Web Search and Data Mining*, 43–51. https://doi.org/10.1145/3336191.3371800

Bouquet, A., Glein, C. R., Wyrick, D., & Waite, J. H. (2017). Alternative Energy: Production of H2 by Radiolysis of Water in the Rocky Cores of Icy Bodies. *The Astrophysical Journal Letters*, *840*(1), L8. https://doi.org/10.3847/2041-8213/aa6d56

Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, *37*(4), 373–384. https://doi.org/10.1080/00401706.1995.10484371

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brockwell, T. G., Meech, K. J., Pickens, K., Waite, J. H., Miller, G., Roberts, J., Lunine, J. I., & Wilson, P. (2016). The mass spectrometer for planetary exploration (MASPEX). *2016 IEEE Aerospace Conference*, 1–17. https://doi.org/10.1109/AERO.2016.7500777

Brown, M. E., & Hand, K. P. (2013). Salts and radiation products on the surface of Europa. *The Astronomical Journal*, *145*(4), 110. https://doi.org/10.1088/0004-6256/145/4/110

Carlson, R. W., Calvin, W. M., Dalton, J. B., Hansen, G. B., Hudson, R. L., Johnson, R. E., & McCord, T. B. (2009). Europa's Surface Composition. In *Europa* (pp. 283–328). University of Arizona Press. https://www.jstor.org/stable/j.ctt1xp3wdw

Chou, L., Mahaffy, P., Trainer, M., Eigenbrode, J., Arevalo, R., Brinckerhoff, W., Getty, S., Grefenstette, N., Da Poian, V., Fricke, G. M., Kempes, C. P., Marlow, J., Sherwood Lollar, B., Graham, H., & Johnson, S. S. (2021). Planetary Mass Spectrometry for Agnostic Life Detection in the Solar System. *Frontiers in Astronomy and Space Sciences*, *8*. https://www.frontiersin.org/articles/10.3389/fspas.2021.755100

Chyba, C. F. (2000). Energy for microbial life on Europa. *Nature*, *403*(6768), 381. https://doi.org/10.1038/35000281

Chyba, C. F., & Phillips, C. B. (2001). Possible ecosystems and the search for life on Europa. *Proceedings of the National Academy of Sciences*, *98*(3), 801–804. https://doi.org/10.1073/pnas.98.3.801

Cleaves, H. J., Hystad, G., Prabhu, A., Wong, M. L., Cody, G. D., Economon, S., & Hazen, R. M. (2023). A robust, agnostic molecular biosignature based on machine learning. *Proceedings of the National Academy of Sciences*, *120*(41), e2307149120. https://doi.org/10.1073/pnas.2307149120

Clough, L. A., Da Poian, V., Major, J. D., Seyler, L. M., McKinney, B. A., & Theiling, B. P. (2025). Interpretable Machine Learning Biosignature Detection From Ocean Worlds Analogue CO2 Isotopologue Data. *Earth and Space Science*, *12*(3), e2024EA003966. https://doi.org/10.1029/2024EA003966

Cooper, J. F., Johnson, R. E., Mauk, B. H., Garrett, H. B., & Gehrels, N. (2001). Energetic Ion and Electron Irradiation of the Icy Galilean Satellites. *Icarus*, *149*(1), 133–159. https://doi.org/10.1006/icar.2000.6498

Craig, H. (1953). The geochemistry of the stable carbon isotopes. *Geochimica et Cosmochimica Acta*, *3*(2), 53–92. https://doi.org/10.1016/0016-7037(53)90001-5

Craig, H., & Gordon, L. I. (1965). *Deuterium and oxygen 18 variations in the ocean and the marine atmosphere*. Consiglio nazionale delle richerche, Laboratorio de geologia nucleare. http://catalog.hathitrust.org/api/volumes/oclc/8019537.html

Da Poian, V., Theiling, B., Lyness, E., Burtt, D., Azari, A. R., Pasterski, J., Chou, L., Trainer, M., Danell, R., Kaplan, D., Li, X., Clough, L., McKinney, B., Mandrake, L., Diamond, B., & Freissinet, C. (2025). *Science Autonomy using Machine Learning for Astrobiology* (arXiv:2504.00709). arXiv. https://doi.org/10.48550/arXiv.2504.00709

Davis, N. A., Crowe, J. E., Pajewski, N. M., & McKinney, B. A. (2010). Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. *Genes & Immunity*, *11*(8), 630–636. https://doi.org/10.1038/gene.2010.37

Dawkins, B. A., Le, T. T., & McKinney, B. A. (2021). Theoretical properties of distance distributions and novel metrics for nearest-neighbor feature selection. *PLOS ONE*, *16*(2), e0246761. https://doi.org/10.1371/journal.pone.0246761

Dawkins, B. A., & McKinney, B. A. (2025). Multivariate Optimization of k for k-Nearest-Neighbor Feature Selection With Dichotomous Outcomes: Complex Associations, Class Imbalance, and Application to RNA-Seq in Major Depressive Disorder. *IEEE Transactions on Computational Biology and Bioinformatics*, *22*(01), 39–51. https://doi.org/10.1109/TCBBIO.2024.3494599

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*, *12*(1), 14683. https://doi.org/10.1038/s41598-022-14395-4

Fanale, F. P., Li, Y.-H., De Carlo, E., Farley, C., Sharma, S. K., Horton, K., & Granahan, J. C. (2001). An experimental estimate of Europa's "ocean" composition independent of Galileo orbital remote sensing. *Journal of Geophysical Research: Planets*, *106*(E7), 14595–14600. https://doi.org/10.1029/2000JE001385

Fulcher, B. D., & Jones, N. S. (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Systems*, *5*(5), 527-531.e3. https://doi.org/10.1016/j.cels.2017.10.001

Fulcher, B. D., Little, M. A., & Jones, N. S. (2013). Highly comparative time-series analysis: The empirical structure of time series and their methods. *Journal of The Royal Society Interface*, *10*(83), 20130048. https://doi.org/10.1098/rsif.2013.0048

Gaisser, T. K., Engel, R., & Resconi, E. (2016). *Cosmic Rays and Particle Physics* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139192194

Gizzi, E., Jr, T. C., Cassamajor-Paul, C., Chertok, R., Clough, L., Firth, C., Gibson, A., Haroon, I., Marshall, J., Maynard, P., Monaghan, M., Owens, H., Rogers, D., Sultana, M., Sinapov, J., & Theiling, B. (2025). OnAIR: Applications of the NASA On-Board Artificial Intelligence Research Platform. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(28), Article 28. https://doi.org/10.1609/aaai.v39i28.35156

Gizzi, E., Nair, L., Chernova, S., & Sinapov, J. (2022). Creative Problem Solving in Artificially Intelligent Agents: A Survey and Framework. *Journal of Artificial Intelligence Research*, *75*, 857–911. https://doi.org/10.1613/jair.1.13864

Glein, C. R., Baross, J. A., & Waite, J. H. (2015). The pH of Enceladus' ocean. *Geochimica et Cosmochimica Acta*, *162*, 202–219. https://doi.org/10.1016/j.gca.2015.04.017

Glein, C. R., & Waite, J. H. (2020). The Carbonate Geochemistry of Enceladus' Ocean. *Geophysical Research Letters*, *47*(3), e2019GL085885. https://doi.org/10.1029/2019GL085885

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*(null), 1157–1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Henderson, T., & Fulcher, B. D. (2022). *Feature-Based Time-Series Analysis in R using the theft Package* (arXiv:2208.06146). arXiv. https://doi.org/10.48550/arXiv.2208.06146

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys*, *2*(none), 61–93. https://doi.org/10.1214/08-SS035

Hoefs, J. (1973). *Stable Isotope Geochemistry* (Vol. 9). Springer. https://doi.org/10.1007/978-3-540-70708-0

Howell, S. M., & Pappalardo, R. T. (2020). NASA's Europa Clipper—A mission to a potentially habitable ocean world. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15160-9

Hyndman, R. J., Wang, E., & Laptev, N. (2015). Large-Scale Unusual Time Series Detection. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 1616–1619. https://doi.org/10.1109/ICDMW.2015.104

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

Johnson, K. (2013). Great Salt Plains, Alfalfa County, Oklahoma—Geology, brines, and hourglass-selenite crystals. *Oklahoma City Geological Survey, Shale Shaker*, *64*, 86–93.

Johnson, K. S. (1988). Great Salt Plains and hourglass selenite crystals, Salt Fork of the Arkansas River, northwest Oklahoma. In O. T. Hayward (Ed.), *South-Central Section of the Geological Society of America* (Vol. 4, p. 0). Geological Society of America. https://doi.org/10.1130/0-8137-5404-6.135

Johnson, P. V., Hodyss, R., Vu, T. H., & Choukroun, M. (2019). Insights into Europa's ocean composition derived from its surface expression. *Icarus*, *321*, 857–865. https://doi.org/10.1016/j.icarus.2018.12.009

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358. https://doi.org/10.1016/j.ijforecast.2016.09.004

Kargel, J. S., Kaye, J. Z., Head, J. W., Marion, G. M., Sassen, R., Crowley, J. K., Ballesteros, O. P., Grant, S. A., & Hogenboom, D. L. (2000). Europa's Crust and Ocean: Origin, Composition, and the Prospects for Life. *Icarus*, *148*(1), 226–265. https://doi.org/10.1006/icar.2000.6471

Kopf, S., Davidheiser-Kroll, B., & Kocken, I. (2021). Isoreader: An R package to read stable isotope data files for reproducible research. *Journal of Open Source Software*, *6*(61), 2878. https://doi.org/10.21105/joss.02878

Krzycki, J. A., Kenealy, W. R., DeNiro, M. J., & Zeikus, J. G. (1987). Stable Carbon Isotope Fractionation by Methanosarcina barkeri during Methanogenesis from Acetate, Methanol,

or Carbon Dioxide-Hydrogen. *Applied and Environmental Microbiology*, *53*(10), 2597–2599.

Lareau, C. A., White, B. C., Oberg, A. L., & McKinney, B. A. (2015). Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure. *BioData Mining*, *8*(1), 5. https://doi.org/10.1186/s13040-015-0040-x

Le, T. T., Dawkins, B. A., & McKinney, B. A. (2020). Nearest-neighbor Projected-Distance Regression (NPDR) for detecting network interactions with adjustments for multiple tests and confounding. *Bioinformatics*, *36*(9), 2770–2777. https://doi.org/10.1093/bioinformatics/btaa024

Le, T. T., Simmons, W. K., Misaki, M., Bodurka, J., White, B. C., Savitz, J., & McKinney, B. A. (2017). Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. *Bioinformatics*, *33*(18), 2906–2913. https://doi.org/10.1093/bioinformatics/btx298

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

MacKenzie, S. M., Neveu, M., Davila, A. F., Lunine, J. I., Craft, K. L., Cable, M. L., Phillips-Lander, C. M., Hofgartner, J. D., Eigenbrode, J. L., Waite, J. H., Glein, C. R., Gold, R., Greenauer, P. J., Kirby, K., Bradburne, C., Kounaves, S. P., Malaska, M. J., Postberg, F., Patterson, G. W., … Spilker, L. J. (2021). The Enceladus Orbilander Mission Concept: Balancing Return and Resources in the Search for Life. *The Planetary Science Journal*, *2*(2), 77. https://doi.org/10.3847/PSJ/abe4da

Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303–342. https://doi.org/10.1016/0098-3004(93)90090-R

Martínez, G. M., & Renno, N. O. (2013). Water and Brines on Mars: Current Evidence and Implications for MSL. *Space Science Reviews*, *175*, 29–51. https://doi.org/10.1007/s11214-012-9956-3

McCollom, T. M. (1999). Methanogenesis as a potential source of chemical energy for primary biomass production by autotrophic organisms in hydrothermal systems on Europa. *Journal of Geophysical Research: Planets*, *104*(E12), 30729–30742. https://doi.org/10.1029/1999JE001126

McKinney, B. A., Jr, J. E. C., Guo, J., & Tian, D. (2009). Capturing the Spectrum of Interaction Effects in Genetic Association Studies by Simulated Evaporative Cooling Network Analysis. *PLOS Genetics*, *5*(3), e1000432. https://doi.org/10.1371/journal.pgen.1000432

McKinney, B. A., White, B. C., Grill, D. E., Li, P. W., Kennedy, R. B., Poland, G. A., & Oberg, A. L. (2013). ReliefSeq: A Gene-Wise Adaptive-K Nearest-Neighbor Feature Selection Tool for Finding Gene-Gene Interactions and Main Effects in mRNA-Seq Gene Expression Data. *PLOS ONE*, *8*(12), e81527. https://doi.org/10.1371/journal.pone.0081527

Miller, K. E., Theiling, B., Hofmann, A. E., Castillo-Rogez, J., Neveu, M., Hosseini, S., Barnes, J., Kleer, K. de, Barrett, T. J., Franz, H. B., Glein, C. R., House, C. H., Blase, R. C., Libardoni, M. J., Spilker, L. J., Choukroun, M., & Drouin, B. J. (2021). Vol. 53, Issue 4 (Planetary/Astrobiology Decadal Survey Whitepapers). *Bulletin of the AAS*, *53*(4). https://doi.org/10.3847/25c2cfeb.6c7da905

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Muñoz-Iglesias, V., Bonales, L. J., & Prieto-Ballesteros, O. (2013). pH and Salinity Evolution of Europa's Brines: Raman Spectroscopy Study of Fractional Precipitation at 1 and 300 Bar. *Astrobiology*, *13*(8), 693–702. https://doi.org/10.1089/ast.2012.0900

M.Y. Zolotov. (2008). Oceanic Composition on Europa: Contraints from Mineral Solubilities. *Abstract*. Lunar and Planetary Science XXXIX, Lunar and Planetary Institute, Houston.

Nagel, E. (1979). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Hackett Publishing Co. https://hackettpublishing.com/philosophy/philosophy-science/the-structure-of-science

National Academies of Sciences, Engineering, and Medicine. (2019). Biosignature Identification and Interpretation. In *An Astrobiology Strategy for the Search for Life in the Universe*. National Academies Press (US). https://doi.org/10.17226/25252.

Neveu, M., Hays, L. E., Voytek, M. A., New, M. H., & Schulte, M. D. (2018). The Ladder of Life Detection. *Astrobiology*, *18*(11), 1375–1402. https://doi.org/10.1089/ast.2017.1773

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Noordijk, B., Garcia Gomez, M. L., ten Tusscher, K. H. W. J., de Ridder, D., van Dijk, A. D. J., & Smith, R. W. (2024). The rise of scientific machine learning: A perspective on combining mechanistic modelling with machine learning for systems biology. *Frontiers in Systems Biology*, *4*. https://doi.org/10.3389/fsysb.2024.1407994

Park, R., & Epstein, S. (1960). Carbon isotope fractionation during photosynthesis. *Geochimica et Cosmochimica Acta*, *21*(1), 110–126. https://doi.org/10.1016/S0016-7037(60)80006-3

Parvandeh, S., Yeh, H.-W., Paulus, M. P., & McKinney, B. A. (2020). Consensus features nested cross-validation. *Bioinformatics*, *36*(10), 3093–3098. https://doi.org/10.1093/bioinformatics/btaa046

Pasek, M. A., & Greenberg, R. (2012). Acidification of Europa's Subsurface Ocean as a Consequence of Oxidant Delivery. *Astrobiology*, *12*(2), 151–159. https://doi.org/10.1089/ast.2011.0666

Postberg, F., Kempf, S., Schmidt, J., Brilliantov, N., Beinsen, A., Abel, B., Buck, U., & Srama, R. (2009). Sodium salts in E-ring ice grains from an ocean below the surface of Enceladus. *Nature*, *459*(7250), 1098–1101. https://doi.org/10.1038/nature08046

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Russell, M. J., Murray, A. E., & Hand, K. P. (2017). The Possible Emergence of Life and Differentiation of a Shallow Biosphere on Irradiated Icy Worlds: The Example of Europa. *Astrobiology*, *17*(12), 1265–1273. https://doi.org/10.1089/ast.2016.1600

Schidlowski, M. (2001). Carbon isotopes as biogeochemical recorders of life over 3.8 Ga of Earth history: Evolution of a concept. *Precambrian Research*, *106*(1), 117–134. https://doi.org/10.1016/S0301-9268(00)00128-5

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, *46*(1), 561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021

Shi, T., & Horvath, S. (2006). Unsupervised Learning With Random Forest Predictors. *Journal of Computational and Graphical Statistics*, *15*(1), 118–138. https://doi.org/10.1198/106186006X94072

Talkner, P., & Weber, R. O. (2000). Power spectrum and detrended fluctuation analysis: Application to daily temperatures. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *62*(1 Pt A), 150–160. https://doi.org/10.1103/physreve.62.150

Tanner, R. S. (2007). Cultivation of Bacteria and Fungi. In *Manual of Environmental Microbiology* (pp. 69–78). John Wiley & Sons, Ltd. https://doi.org/10.1128/9781555815882.ch6

Theiling, B. P. (2021). The effect of Europa and Enceladus analog seawater composition on isotopic measurements of volatile CO2. *Icarus*, *358*, 114216. https://doi.org/10.1016/j.icarus.2020.114216

Theiling, B. P., Chou, L., Da Poian, V., Battler, M., Raimalwala, K., Arevalo, R., Neveu, M., Ni, Z., Graham, H., Elsila, J., & Thompson, B. (2022a). Science Autonomy for Ocean Worlds Astrobiology: A Perspective. *Astrobiology*, *22*(8), 901–913. https://doi.org/10.1089/ast.2021.0062

Theiling, B. P., Chou, L., Da Poian, V., Battler, M., Raimalwala, K., Arevalo, R., Neveu, M., Ni, Z., Graham, H., Elsila, J., & Thompson, B. (2022b). Science Autonomy for Ocean Worlds Astrobiology: A Perspective. *Astrobiology*, *22*(8), 901–913. https://doi.org/10.1089/ast.2021.0062

Thiemens, M. A., & Heidenreich, J. E. (1983). The Mass-Independent Fractionation of Oxygen: A Novel Isotope Effect and its Possible Cosmochemical Implications. *Science*, *219*(4588), 1073–1075.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), 385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong Rules for Discarding Predictors in Lasso-Type Problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *74*(2), 245–266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

Toner, J. D., Catling, D. C., & Light, B. (2015). Modeling salt precipitation from brines on Mars: Evaporation versus freezing origin for soil salts. *Icarus*, *250*, 451–461. https://doi.org/10.1016/j.icarus.2014.12.013

Trumbo, S. K., & Brown, M. E. (2023). The distribution of CO2 on Europa indicates an internal source of carbon. *Science*, *381*(6664), 1308–1311. https://doi.org/10.1126/science.adg4155

Trumbo, S. K., Brown, M. E., & Hand, K. P. (2019). Sodium chloride on the surface of Europa. *Science Advances*, *5*(6), eaaw7123. https://doi.org/10.1126/sciadv.aaw7123

Vance, S. D., Craft, K. L., Shock, E., Schmidt, B. E., Lunine, J., Hand, K. P., McKinnon, W. B., Spiers, E. M., Chivers, C., Lawrence, J. D., Wolfenbarger, N., Leonard, E. J., Robinson, K. J., Styczinski, M. J., Persaud, D. M., Steinbrügge, G., Zolotov, M. Y., Quick, L. C.,

Scully, J. E. C., … Elder, C. M. (2023). Investigating Europa's Habitability with the Europa Clipper. *Space Science Reviews*, *219*(8), 81. https://doi.org/10.1007/s11214-023-01025-2

Vance, S. D., Hand, K. P., & Pappalardo, R. T. (2016). Geophysical controls of chemical disequilibria in Europa. *Geophysical Research Letters*, *43*(10), 4871–4879. https://doi.org/10.1002/2016GL068547

Vaniman, D., Chipera, S., Rampe, E., Bristow, T., Blake, D., Meusburger, J., Peretyazhko, T., Rapin, W., Berger, J., Ming, D., Craig, P., Castle, N., Downs, R. T., Morrison, S., Hazen, R., Morris, R., Pandey, A., Treiman, A. H., Yen, A., … Fraeman, A. (2024). Gypsum on Mars: A Detailed View at Gale Crater. *Minerals*, *14*(8), Article 8. https://doi.org/10.3390/min14080815

Vaniman, D. T., Martínez, G. M., Rampe, E. B., Bristow, T. F., Blake, D. F., Yen, A. S., Ming, D. W., Rapin, W., Meslin, P.-Y., Morookian, J. M., Downs, R. T., Chipera, S. J., Morris, R. V., Morrison, S. M., Treiman, A. H., Achilles, C. N., Robertson, K., Grotzinger, J. P., Hazen, R. M., … Sumner, D. Y. (2018). Gypsum, bassanite, and anhydrite at Gale crater, Mars. *American Mineralogist*, *103*(7), 1011–1020. https://doi.org/10.2138/am-2018-6346

Villanueva, G. L., Hammel, H. B., Milam, S. N., Faggi, S., Kofman, V., Roth, L., Hand, K. P., Paganini, L., Stansberry, J., Spencer, J., Protopapa, S., Strazzulla, G., Cruz-Mermy, G., Glein, C. R., Cartwright, R., & Liuzzi, G. (2023). Endogenous CO2 ice mixture on the surface of Europa and no detection of plume activity. *Science*, *381*(6664), 1305–1308. https://doi.org/10.1126/science.adg4270

Vogel, J. C. (1980). *Fractionation of the Carbon Isotopes During Photosynthesis*. Springer. https://doi.org/10.1007/978-3-642-46428-7

Waite, J. H., Burch, J. L., Brockwell, T. G., Young, D. T., Miller, G. P., Persyn, S. C., Stone, J. M., Wilson, P., Miller, K. E., Glein, C. R., Perryman, R. S., McGrath, M. A., Bolton, S. J., McKinnon, W. B., Mousis, O., Sephton, M. A., Shock, E. L., Choukroun, M., Teolis, B. D., … Siegmund, O. H. W. (2024). MASPEX-Europa: The Europa Clipper Neutral Gas Mass Spectrometer Investigation. *Space Science Reviews*, *220*(3), 30. https://doi.org/10.1007/s11214-024-01061-6

Waite, J. H., Combi, M. R., Ip, W.-H., Cravens, T. E., McNutt, R. L., Kasprzak, W., Yelle, R., Luhmann, J., Niemann, H., Gell, D., Magee, B., Fletcher, G., Lunine, J., & Tseng, W.-L. (2006). Cassini Ion and Neutral Mass Spectrometer: Enceladus Plume Composition and Structure. *Science*, *311*(5766), 1419–1422. https://doi.org/10.1126/science.1121290

Waite Jr, J. H., Lewis, W. S., Magee, B. A., Lunine, J. I., McKinnon, W. B., Glein, C. R., Mousis, O., Young, D. T., Brockwell, T., Westlake, J., Nguyen, M.-J., Teolis, B. D., Niemann, H. B., McNutt Jr, R. L., Perry, M., & Ip, W.-H. (2009). Liquid water on Enceladus from observations of ammonia and 40Ar in the plume. *Nature*, *460*(7254), Article 7254. https://doi.org/10.1038/nature08153

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*, 1–17. https://doi.org/10.18637/jss.v077.i01

Wright, M. N., Ziegler, A., & König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, *17*(1), 145. https://doi.org/10.1186/s12859-016-0995-8

Yu, W., & Hecht, G. R. (2024). *Onboard GNC Concept Design Study for Cooperative Event Driven Distributed Systems* (p. 25) [Earth Resources and Remote Sensing]. https://hs-niederrhein.digibib.net/search/eds/record/edsnas:edsnas.20240001747?be-eds-

expander=fulltext&be-eds-facetfilter=1%2CContentProvider%3ANASA+Technical+Reports&be-eds-sort=relevance&q-au=%22Grant%2C+R.%22&start=1&count=20&hitcount=34&pos=1

Zeebe, R., & Wolf-Gladrow, D. (2001, October 29). *CO2 in Seawater: Equilibrium, Kinetics, Isotopes*. https://www.semanticscholar.org/paper/CO2-in-Seawater%3A-Equilibrium%2C-Kinetics%2C-Isotopes-Zeebe-Wolf-Gladrow/171c4a87efd2630eae567d92d2d9ece4541d980d

Zolotov, M. Y. (2007). An oceanic composition on early and today's Enceladus. *Geophysical Research Letters*, *34*(23). https://doi.org/10.1029/2007GL031234

Zolotov, M. Y., & Kargel, J. S. (2009). On the Chemical Composition of Europa's Icy Shell, Ocean, and Underlying Rocks. In *Europa* (pp. 431–458). University of Arizona Press. https://www.jstor.org/stable/j.ctt1xp3wdw.24

Zolotov, M. Y., & Shock, E. L. (2001). Composition and stability of salts on the surface of Europa and their oceanic origin. *Journal of Geophysical Research: Planets*, *106*(E12), 32815–32827. https://doi.org/10.1029/2000JE001413

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

APPENDIX A

BIOSIGNATURE DATA PROCESSING, ANALYSIS AND PREDICTION

This Supplemental Information Appendix contains details about the collection, processing, and full machine learning (ML) biosignature prediction results of our Benchmark Ocean Worlds volatile CO_2 isotope ratio mass spectrometry (IRMS) dataset, referred to as BOW-δCO_2 (Sec. A1). This data is provided in csv-file format for accessibility after processing from the instrument-output dxf-format for Isodat©, and includes the full extracted time-series mass spectrometry (TSMS) feature space, as well as associated metadata such as pH, ionic strength, and salt components. For some samples, pH values are not available (NA). We discuss the automated quality analysis/quality control pipeline implemented in R that is used to select quality samples for ML and further data analysis from high-throughput CO_2 IRMS experimental data. The QA/QC pipeline runs checks for the expected number of peaks, their expected retention times, relative intensities, and standard deviation of isotope measurements for both reference and sample peaks (Sec. A2). Samples that pass QA/QC are handed to a calibration method for automated oxygen isotope calibration using internal laboratory standards (Sec A2).

Our ML biosignature model is agnostic to this variable CO_2 isotope fractionation and accurately classifies biosignatures in different brines, prepared with variable salt components (and thus ionic strengths), with varying pH, and non-biogenic organic content (Sec. A3).

Then we describe the known correlation clusters in our dataset and provide correlation heatmaps showing how the inclusion of the LASSO penalty during feature selection reduces correlation in the feature space, and the use of the novel unsupervised Random Forest proximity (URFP) distance results in the selection of fewer features, reducing the feature space further than standard distance metrics without the use of the LASSO penalty (Sec. A4).

Finally we present the full results of five additional replicates of random train/test splits to confirm that our results are unbiased and to compare the performance of our nearest-neighbors projected distance regression (NPDR) feature selection (Le et al., 2020) using LASSO penalty and URFP distance (NPDR-LURF) with different feature spaces (Sec. A5). We compare the performance of NPDR-LURF with NPDR-LMan (NPDR with LASSO penalty and Manhattan distance), RR (reduced redundancy, a feature space with features >99.0% correlated removed), deltas (*avg_*$\delta^{18}O$ and *avg_*$\delta^{13}C$), and calib-deltas (*avg_calib_*$\delta^{18}O$ and *avg_*$\delta^{13}C$) features. We report that NPDR-LURF selected features outperform the deltas and calib-deltas features significantly and additionally outperforms NPDR-LMan features with comparable performance to the RR features, a feature space that captures all of the information in the full TSMS feature space with redundancy removed. The NPDR-LURF models generally have more than forty fewer features than the RR features (NPDR-LURF and NPDR-LMan features are selected on every train/test split to prevent bias), making these models more lightweight in terms of computational resources in addition to being more human-interpretable. Finally, we present more local-RF importance scores for true and biosignature predictions (Sec. A6).

A.1 Generation of the BOW-δCO_2 Dataset

Ocean world (OW) analogue seawater solutions were prepared and analyzed as described in (Theiling, 2021), and a brief overview of the solution preparation is given here.

All experimental data were collected using a Thermo Finnigan Gasbench II trace gas analyzer coupled to a Thermo Delta V Advantage Isotope Ratio Mass Spectrometer (IRMS). Seawater analogue samples were prepared as 250 mL stock solutions, and 0.5 mL of brine solution was pipetted from these into 12 mL chlorobutyl septa-capped vials. CO_2 (g) was added by volume (0.3% for the samples in the current study) after flushing vials with ultra-high purity

He to remove atmospheric gases. Both abiotic and biotic (inoculated with microbes) samples were prepared in triplicate to measure within-sample variability and natural heterogeneity. Sample seawater and gas were allowed to equilibrate for seven days in a temperature-controlled room before analysis. The solutions contained a distribution of salt compositions and non-biogenic organics (microbial substrates, included in both biotic and abiotic samples as controls and to assess the potential for abiotic biosignature decoys).

Quality experiments from three batches of samples were combined to create BOW-δCO_2: "Abiotic", "Europa Bacteria", and "Microbial Mud". The "Europa Bacteria" batch represented samples closest to hypothesized compositions of Europa's subsurface ocean (P. V. Johnson et al., 2019; Pasek & Greenberg, 2012) and contained both non-biogenic laboratory organics and substrates. While the abiotic and microbial mud batches only contained abiotic and collected biotic samples, respectively, the Europa bacteria batch contained both biotic and abiotic samples, sub-datasets referred to as "Europa Abiotic" and "Europa biotic" (Table A.1). The "Europa Biotic" samples additionally contained laboratory cultivated thermophilic sulfate reducers, *Desulfotomaculum thermocisternum* (type strain DSM 10259, obtained from the Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures GmbH) in various media. Microbial cultures were cultivated under anaerobic conditions following standardized procedures (Tanner, 2007) in septa-capped serum bottles and recommended media for *D. thermocisternum* at 25°C (modified Desulfonauticus medium 383b): in 1 L distilled water, 3.0 g Na_2SO_4, 0.2 g KH_2PO_4, 0.3 g NH_4Cl, 21.0 g NaCl, 3.00 $MgCl_2$ x $6H_2O$, 0.5 g KCl, 0.15 g $CaCl_2$ x $2H_20$, 2.5 g $NaHCO_3$, 0.4 g Na_2S x $9H_2O$, 1 mL trace element solution (ATCC), and 10 ml vitamin solution (ATCC), pH 7.0-7.2, under 80% H_2 and 20% CO_2. Initial growth medium is a salinity modified version of DSM medium 113: in 1 L distilled water, 2.0 g KH_2PO_4, 4.0 g KNO_3, 1.0 g NH_4Cl,

0.8 g $MgSO_4$ x $7H_2O$, 5.0 g $Na_2S_2O_3$ x $5H_2O$, 1.0 g $NaHCO_3$, 2.0 mg $FeSO_4$ x $7H_2O$, 25.0 g NaCl, and 2 mL trace element solution (ATCC), pH 7.0, under 100% N_2. Septa-capped bottles are slightly overpressurized with headspace gas to prevent infiltration of atmospheric O_2. Na-resazurin (0.00005% wt./volume final concentration) was added to cultures as an oxidation-reduction indicator. Microbial growth in initial stock was tracked by direct cell counting using DAPI stain and fluorescent microscopy.

Once exponential growth was achieved, a ~500 mL culture volume was transferred to the University of Tulsa for the establishment of Europa microcosms. Seawater microcosms were inoculated with culture (or cell-free blanks "media" "control") using a gastight syringe. *D. thermocisternum* samples were stored in a 40°F refrigerator to slow growth, and used for IRMS analysis within one month of cultivation.

Table A.1.

Number of samples in four sub-datasets of BOW-δCO_2.

Abiotic	Europa Abiotic	Europa Biotic	Microbial Mud
57	59	36	39

Note. Samples consisted of both abiotic and biotic samples. Salts used for all datasets included various combinations and concentrations of KCl, NaCl, Na_2SO_4, $NaHCO_3$, $MgCl_2$, and $MgSO_4$ following a range of hypothesized and measured Europa and Enceladus seawater salt compositions (detailed in Theiling 2021). The 'Europa Biotic' samples were prepared with *Delsulfotomaculum thermocisternum* microbes in solutions with and without different types of substrates. The 'Microbial Mud' samples were prepared with an uncharacterized heterogenous sample of microbes collected from a reducing lake environment. Abiotic samples that may act as biotic decoys were samples prepared with the *D. thermocisternum* growth medium organic substrate without microbes.

The addition of microbes did not significantly change the pH of the prepared analogue OW brines, which we report on a total scale (Table A.2). Microbe experiments were performed under the same experimental conditions as all abiotic experiments (Theiling, 2021). The distribution of salt components among the biotic and abiotic classes and among the four batches of data were roughly balanced by the proportion of abiotic experiments to biotic. The ionic strength of our solutions ranges from 0.0 – 15.46 M (Table A.2), and pH values from 3.5 – 9.5. We note that pH values were measured with strips because the vials could not accommodate probes, and we used acid- (<7) and base- (>7) specific strips to enable a precision to +/- 0.5.

Table A.2.

Salt Components, pH Values, and Ionic Strengths in BOW-δCO_2.

Salt composition	pH (total)	Ionic strength (M)
KCl	4.5	0.07 - 0.94
KCl + $MgCl_2$	4.5	0.22 - 3.14
$MgCl_2$	5.0	0.16 - 2.21
$MgSO_4$	3.5	7.98
$MgSO_4$ + Na_2SO_4 + $MgCl_2$	3.5	13.41 - 15.46
$MgSO_4$ + NaCl	3.5	8.06 - 9.17
$MgSO_4$ + $NaHCO_3$	8.0 - 9.0	8.0 - 8.97
Na_2SO_4	3.5 - 4.0	5.28 - 7.04
Na_2SO_4 + NaCl + $NaHCO_3$	8.0	5.39
Na_2SO_4 + $NaHCO_3$	8.0 - 9.0	5.30 - 5.88
Na_2SO_4 + NaCl	3.5	5.37 - 6.48
NaCl	4.5	0.09
NaCl + KCl + $MgCl_2$	4.5	0.31 - 2.36
$NaHCO_3$	8.0 - 9.5	0.02 - 0.6
$NaHCO_3$ + NaCl	8.0 - 9.0	0.11 - 1.79

The “Microbial mud” experiments were performed specifically to understand the potential heterogeneity of biosignatures that could be generated by an uncharacterized microbial ecosystem, such as one that could exist in an unknown planetary environment. The “Microbial Mud” batch therefore contained microbial mud collected from a natural source – a small, restricted, ephemeral pond outside of Tulsa, Oklahoma (approximate coordinates 36.234143, -95.923816) in April 2017. Collection of anaerobic mud was performed by slowly submerging 6 collection vials in anoxic-suboxic mud and water and capping the collection vial while still submerged. Aliquots for experiments were extracted from the vials through a septa cap using a gastight syringe and transferred to prepared, septa-capped analogue seawater vials that had been flushed with He and CO_2. Six mud collection vials were capped tightly during collection to prevent leaking of atmosphere into the vials. However, we noted a minor visual oxidation at the top ~1 cm of two vials within a few days of collection, as indicated by alteration of dark brown-black mud to red-orange. The four remaining mud collection vials did not demonstrate any evidence of alteration over several weeks, even after subsampling for experiments, demonstrating that anaerobic conditions were maintained. Microbes (microbial mud and *D. thermocisternum* samples) were stored in a 40°F refrigerator to slow growth. IRMS experiments using ‘microbial mud’ were performed within 1-2 weeks of collection to minimize the potential for oxidation of samples during storage. The *D. thermocisternum* samples were used for IRMS analysis within one month of cultivation.

A2: Building a Quality Dataset for Machine Learning: QA/QC and Calibration

A fully automated QA/QC pipeline was performed on raw IRMS experimental output to build high-quality datasets for ML using the R statistical programming language. The R library

isoreader (Kopf et al., 2021) was used to read in specially formatted output files (.dxf files) from the instrument software, Isodat©. Data from a given experiment consisted of repeated measurements of volatiles over time (900 seconds) in the headspace of the 12 mL vial containing 0.5 mL of an OW analogue seawater. The spectra, or chromatograms, produced for these experiments are created from signal versus time measurements (inset spectrum, Fig. A.1). During the experiment, a reference gas with known CO_2 isotope ratios was also repeatedly sampled for the calculation of sample (unknown) isotope ratios. We took five reference peak measurements (rectangular peaks, inset spectrum Fig. A.1) and a total of eleven sample peaks (sharp peaks, inset spectrum Fig. A.1). The first occurrence of a sample peak represents valve decontamination efforts (between groups of two reference peaks, inset spectrum Fig. A.1). The first occurrence of a sharp "sample" peak is due to planned valve decontamination efforts, whereby leftover gas in the system's sampling loop is briefly switched to a transfer mode to rid the system of any previous sample (between groups of two reference peaks, inset spectrum Fig. A.1). It is well-understood that higher concentration samples can require an additional decontamination transfer, and therefore we also removed the first sample peak after this group of reference peaks to remove the potential for the first true sample peak to have any trace of the prior analysis. This reduces the ten planned subsamples to obtain nine sample peaks for data analysis. This peak filtering was performed during the automated QA/QC (numbered checks, Fig. A.1), as was oxygen isotope calibration (Fig. A.3).

Typical failures demonstrate extra peaks due to air contamination (Fig. A.1a, left panel: Analysis 3080) or a lack of sample peaks due to no CO_2 being present because of preparation error (Fig. A.1a, left panel: spectra with no peaks). In some cases, the relative difference in intensities between the reference and sample peaks will cause an experiment to fail QC if it is

severe enough, and in other cases the reason for an experiment being flagged cannot be seen from the chromatogram. The most frequent reason for a 'normal' looking experiment (compare Analyses 2097, Fig. A.1a right panel, and 3046, Fig. A.1b, left panel) to fail QC is that either the standard deviation of $\delta^{18}O$ or $\delta^{13}C$ (or both) are above the accepted threshold (Fig. A.1, checks 3 and 6).

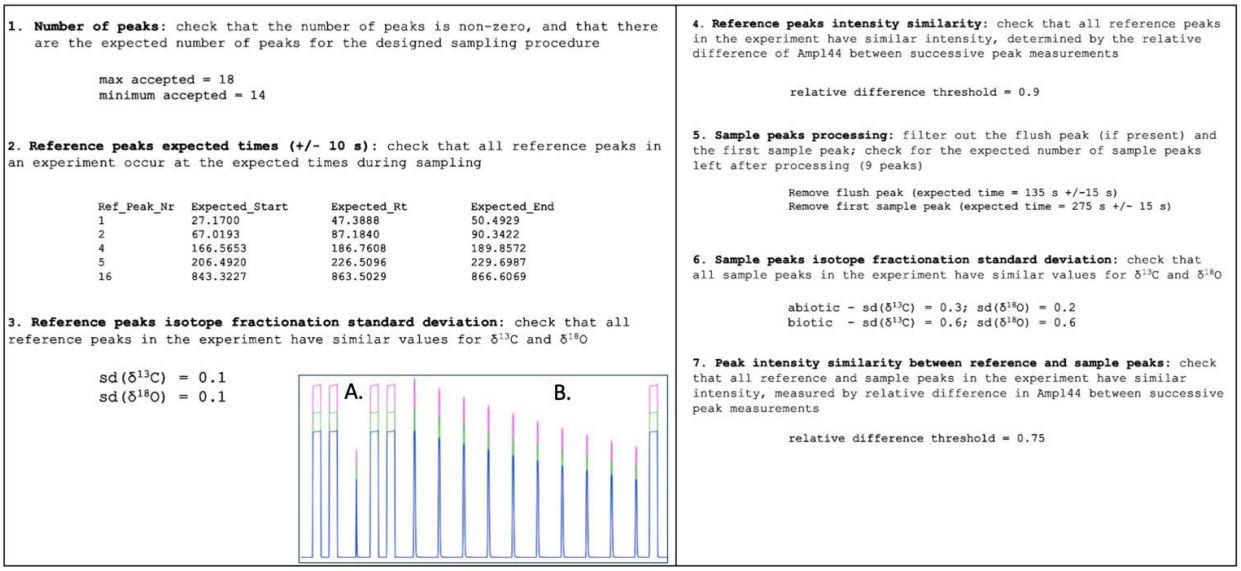
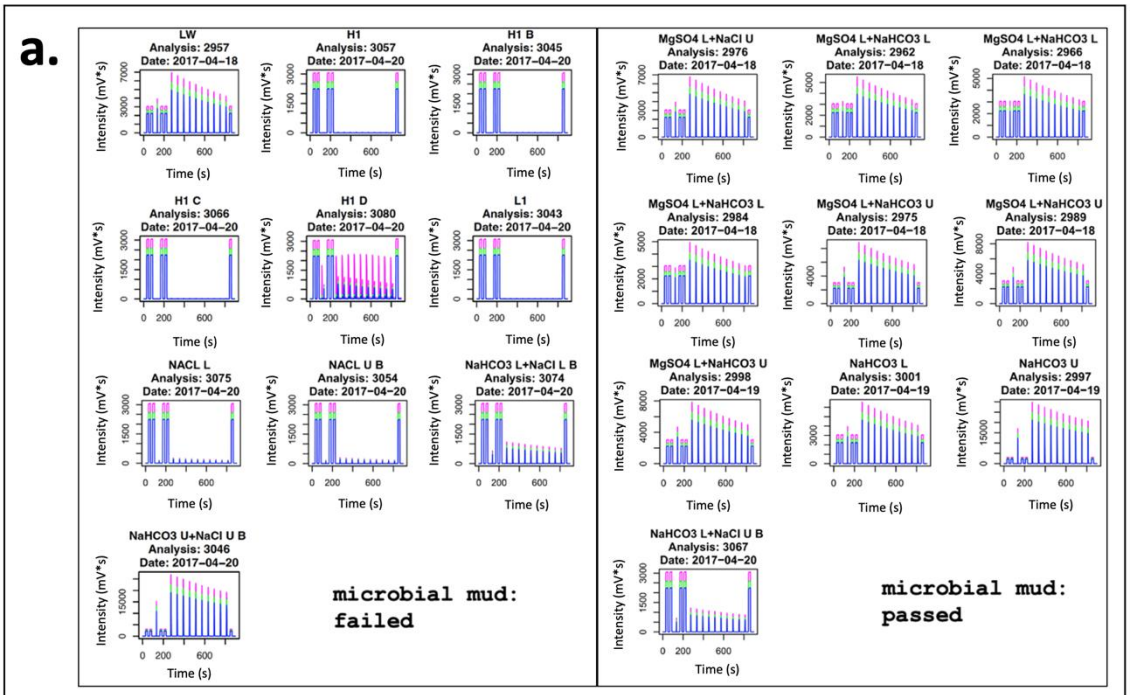


Figure A.1. Checks implemented in the QA/QC pipeline that determine whether an IRMS experiment is of acceptable quality. A typical quality experiment in our benchmark dataset (inset spectrum of rectangular and sharp peaks of CO_2 isotopologues (stacked peaks, masses = 46 (pink), 45 (green) and 44 (blue)) shows five reference peaks (A, four rectangular peaks and last peak) and 11 sharp peaks (B). The reference peaks correspond to reference CO_2 with known isotope fractionation. The sharp peaks, referred to as sample peaks, correspond to repeated sampling (auto-sampling) of the headspace (i.e., of volatile CO_2) in the experimental vial containing 0.5 mL of OW analogue seawater. A typical experimental batch will contain around 96 vials of analogue seawaters. A variety of problems can occur during vial preparation and during the IRMS analysis itself that can lead to a poor-quality spectrum, including vial leaks (atmospheric contamination), water contamination (condensed water on the bottom of vial lids being unintentionally sampled which restricts the sample transfer system), and instrument errors (sampling failure leading to no peaks). Our automated QA/QC pipeline implements seven checks (checks 1-7) informed by expert knowledge of the experimental design and of expectations for quality IRMS measurements to autonomously inspect instrument output from hundreds of experiments to report experimental quality parameters and determine whether experiments pass or fail QA/QC.

The thresholds for the checks implemented in the automated QC were informed by experimental and expert knowledge. Differences in thresholds for expected quality in biotic versus abiotic experiments (Fig. A.1, check 6) ensured that normal isotope variations in the biotic samples that could indicate poor quality in abiotic experiments were preserved. Analyses showed that our biotic experiments may indeed have different baselines for quality due to the heterogeneity with which microbes metabolize and/or differences in total cell counts per experiment, but when prior knowledge of the sample is not possible, we have found that the majority of quality biotic experiments were retained using the abiotic thresholds (the tradeoff being that several quality microbial experiments were filtered out). We therefore used a strict threshold for isotope fractionation standard deviation between repeated measures in check 6 for "Abiotic" experiments and a more relaxed value for "Microbial Mud" experiments, and abiotic thresholds for the "Europa Bacteria" samples, which contained both biotic and abiotic samples. Results from this approach showed that high quality data were obtained for both biotic and abiotic samples for ML (Fig. A.2a) and the relative proportions of experiments from each class being filtered out were similar (Fig. A.2b).



b.

Batch	Abiotic	Europa bacteria	Microbial mud
Before QC	784	216	86
Passed QC	435	110	42

Figure A.2. Results from automated QA/QC on three batches of laboratory generated OW IRMS data. Images of chromatograms that passed (a, right) and failed (a, left) the automated checks from the "Microbial Mud" samples show visible differences in quality. A failure may typically demonstrate extra peaks from air contamination (as seen in Analysis 3080, left panel a) or a lack of sample peaks due to no CO_2 being present. In some cases, the relative difference in intensities between the reference and sample peaks will cause an experiment to fail QC if it is severe enough, and in other cases the reason for an experiment being flagged cannot be seen from the chromatogram. The most frequent reason for a 'normal' looking experiment (compare Analyses 2097, right panel and 3046 left panel) to fail QA/QC is that either the standard deviation of $\delta^{18}O$ or $\delta^{13}C$ (or both) are above the accepted threshold (Fig. A.1, checks 3 and 6). (b) The results after running the automated QA/QC on the "Abiotic", "Europa Bacteria" (consisting of both "Europa Abiotic" and "Europa Biotic" samples), and "Microbial Mud" datasets for all concentrations of CO_2 are shown below the chromatograms and indicate the large number of experiments that are of poor quality in real laboratory work compared with experiments of high quality. This illustrates the high value of quality ML data generated from experimentation.

Experiments that pass QA/QC are handed to a method for oxygen isotope calibration (Fig. A.3). Calibration uses internal laboratory standards, referred to as H1, L1, and LW, with known oxygen isotope ratios relative to national and international standards. Linear regression of measured oxygen isotope fractionation ($\delta^{18}O$) with known values is used to produce the calibration model. This also produces two more features for ML upon feature extraction, *avg_calib_* $\delta^{18}O$ and *sd_calib_* $\delta^{18}O$. Automated methods like our QA/QC pipeline for CO_2 IRMS data could be important for data prioritization and compression for future resource-limited missions to OWs to enhance science return.

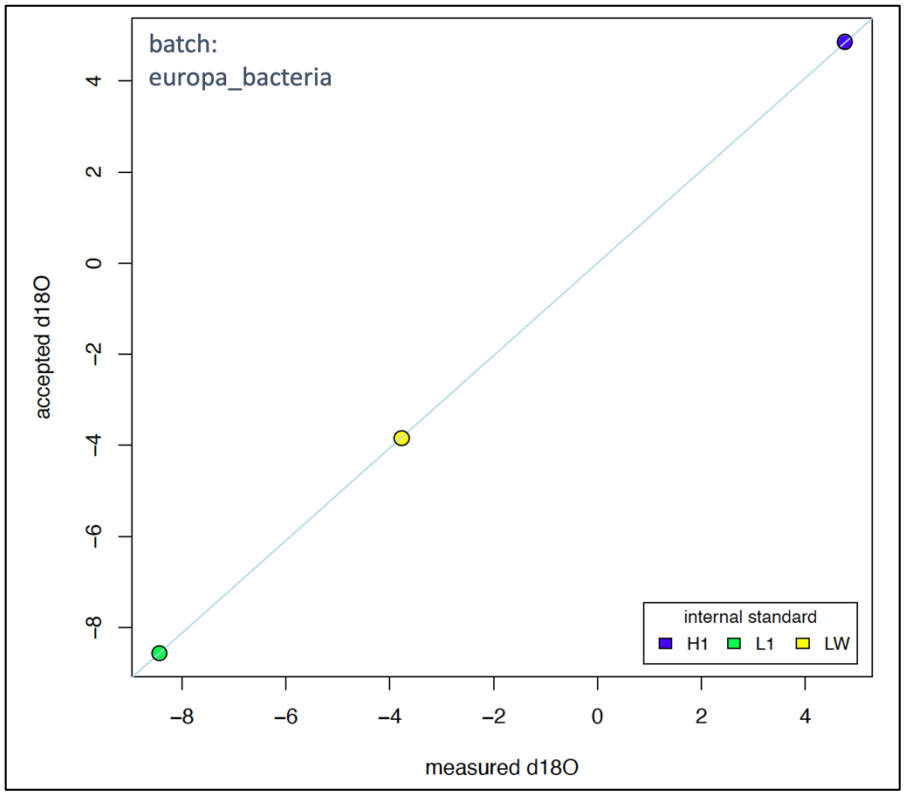


Figure A.3. Linear calibration of oxygen isotope fractionation using internal laboratory standard solutions. Several vials containing laboratory standards (referred to as H1, L1, and LW) with known $\delta^{18}O$ are regularly interspersed with prepared OW analogue brines during high-throughput IRMS analyses. Linear regression is used for measured $\delta^{18}O$ of the standards to produce a calibration model and two more IRMS-derived features for ML, *avg_calib_*$\delta^{18}O$ and *sd_calib_*$\delta^{18}O$.

A.3 Biotic Class and Salt Composition in BOW-δCO_2

The biotic and abiotic samples represented in each salt composition roughly represent the class imbalance (about 64% of all samples are abiotic), except for some small-sample size compositions, which only have abiotic samples (Fig. A.4a). For purposes of interpretation, we also refer to the "Abiotic", "Europa Abiotic", "Europa Biotic", and "Microbial Mud" sub-datasets as "environments", since they reflect different biogeochemical environments and conditions. Inspection of biotic classes by environment shows that the "Microbial Mud" and "Europa Biotic" samples are roughly balanced and reflect various posited OW seawater compositions (Fig. A.4b). Although this data is not perfectly balanced, it represents a high-quality benchmark dataset of OW analogue brine IRMS experiments of volatile CO_2 for astrobiology ML research.

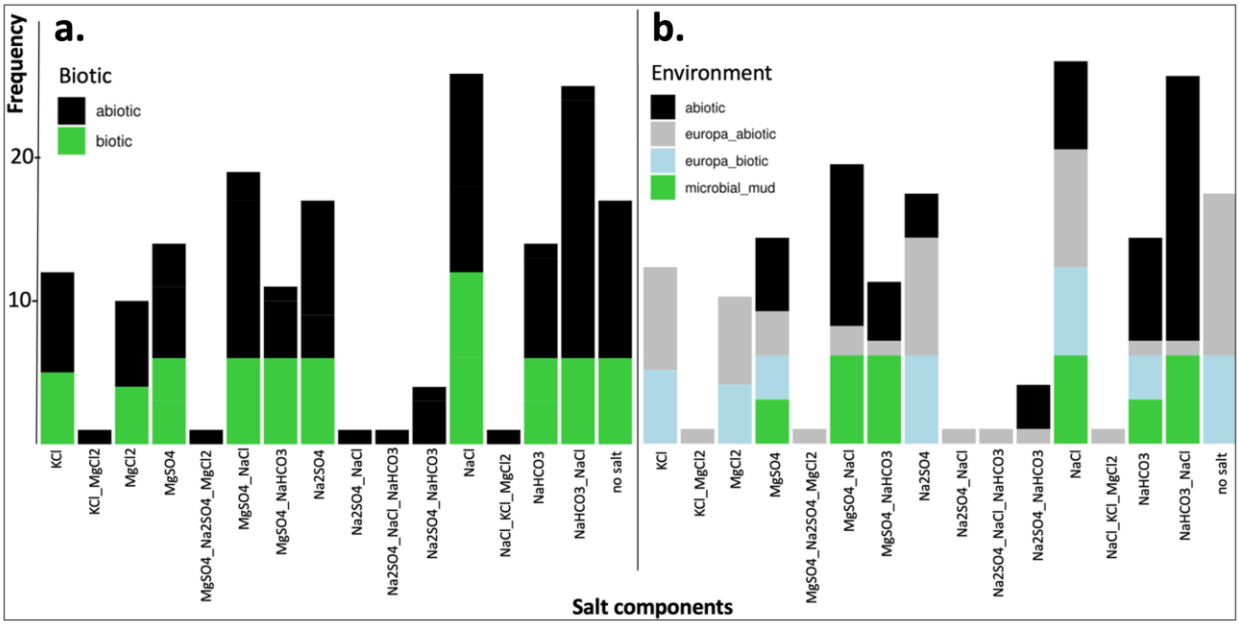


Figure A.4. Frequency of biotic vs. abiotic samples (a) and environment (b) for all salt compositions prepared using 0.3% CO_2. Environment is a more granular label for biotic and abiotic samples. There are about 64% abiotic samples in our dataset, and the class imbalance is roughly equal across different compositions. This dataset represents a high-quality benchmark dataset of OW analogue brine IRMS experiments of volatile CO_2 for astrobiology ML research.

A.4 Machine Learning Feature Spaces and Correlation Clusters

In the main manuscript we illustrate the limitations of feature selection methods that do not account for high correlation between variables (see Fig. 5). While RF returns feature importance scores and could be used for feature selection, it cannot find statistical interactions in high-dimensional feature spaces (McKinney et al., 2009; Wright et al., 2016) and it returns scores for all features used to train the model, so there is no way to reduce the feature space except through an arbitrary choice of *e.g.*, taking the top twenty features. If RF is used for feature selection, one must therefore do the non-trivial work of manually reducing high correlation by selecting one variable in a correlation cluster of features. For example, the TSMS features have 11 clusters of >99.0% correlation, mostly in the IRMS-derived features output by the instrument (Fig. A.5). If RF importance is used for feature selection, then one should select one variable out of each correlation cluster (Fig. A.5 a-k) to keep in the feature space and remove the rest. There is still some level of arbitrariness to this process, as the 99.0% correlation threshold is simply a choice.

We generate a feature space we refer to as reduced redundancy (RR) in which we make that choice based on a subjective notion of which feature may provide the most interpretability in a correlation cluster, indicated by "*****" (Fig. A.5). This process is also time-consuming, whereas feature selection methods like NPDR with a LASSO penalty automatically reduce correlation and redundancy while enriching the selected feature space for statistical interactions, without the

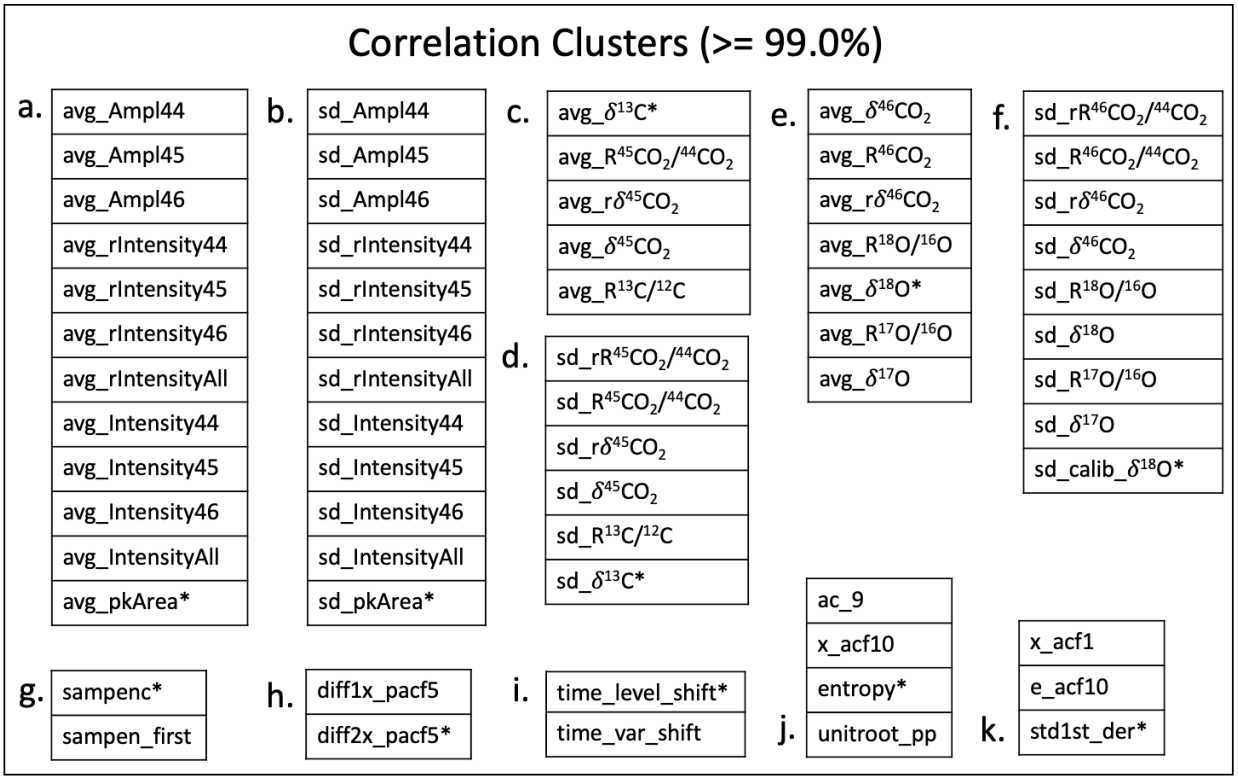


Figure A.5. There are 11 clusters of >99.0% correlation in the BOW-δCO_2 dataset, mostly in the IRMS-derived features. There are two clusters of 12 features (a and b) representing average peak intensities and the standard deviation of peak intensities in chromatograms. Two clusters represent average $\delta^{13}C$ and the standard deviation of $\delta^{13}C$ between sample peaks (c and d), and two clusters likewise represent average $\delta^{18}O$ (and $\delta^{17}O$) and the standard deviation of $\delta^{18}O$ (and $\delta^{17}O$) between sample peaks (e and f). There are five smaller correlation clusters of time-series (TS) features in this dataset (g – k). While features in correlation clusters are highly correlated in BOW-δCO_2, this may or may not be true for other datasets, so correlation analysis must be repeated for different datasets. We curate one feature from each correlation cluster a – k, indicated by "*" to create the reduced redundancy (RR) feature space.

need for additional RF models or the manual reduction of correlation clusters. Instead, features are selected using an adjusted P-value of < 0.05.

For this dataset, NPDR-URF selected features show a reduced feature space compared with NPDR feature selection using standard distance metrics even without the use of the LASSO penalty (Fig. A.6). For example, NPDR-URF feature selection yields ten features for biosignatures, only four more than selected using NPDR-LURF. NPDR feature selection using a standard Manhattan distance metric, however, resulted in 22 features for biosignatures. Five of

the NPDR-URF features are in a highly correlated cluster (Fig. A.6a), and selecting one yields a feature space identical to NPDR-LURF. The NPDR-Manhattan features show several areas of high correlation (areas 1-5, Fig. A.6b) and a feature space that is over twice as large as NPDR-URF. Future work is needed to determine if this reduction in selected features and thereby redundancy is a characteristic trait of the URFP distance metric in NPDR.

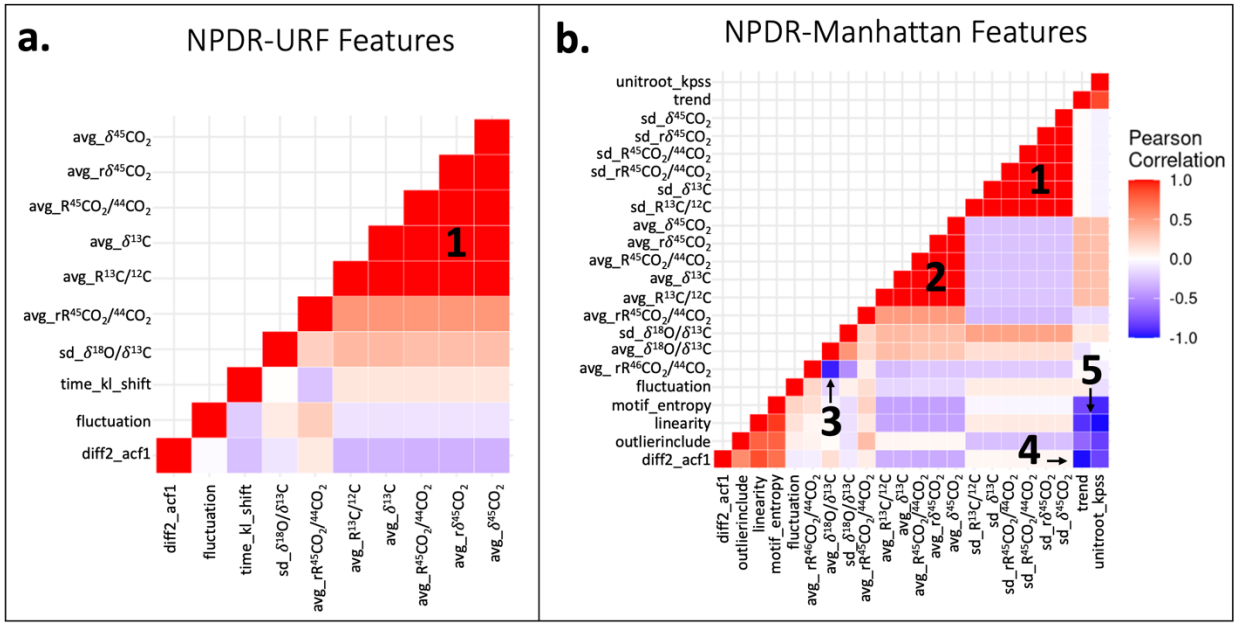


Figure A.6. Correlation heatmaps for NPDR-URF and NPDR-Manhattan selected features (without LASSO penalty). NPDR with unsupervised Random Forest (NPDR-URF) distance metric **(a)** results in fewer features and less correlation than a standard Manhattan metric (b). Red numbered clusters have high positive correlation. Blue labeled clusters have very high negative correlation.

We therefore compare the RF performance of biosignature models created using NPDR-LURF, NPDR-LMan, and the reduced redundancy (RR) feature spaces, as well as feature spaces composed of only delta features to illustrate the difficulties of a ML prediction of biosignatures using only IRMS-derived features (Fig. A.7). The RR feature space was determined as described above (Fig. A.7a). The "deltas" feature space is composed of *avg_δ13C* and *avg_δ18O* (Fig. A.7b), and the calibrated deltas feature space, "calib-deltas", is composed of avg_δ^{13}C and

avg_calib_$\delta^{18}O$ (Fig. A.7c). NPDR-selected feature spaces are determined on each new train and test split to prevent model bias (Fig. A.7d, see also Tables A.3 – A.8 for NPDR feature selection results).

a. Reduced Redundancy (RR) Feature Space

diff1_acf1	trev_num	unitroot_kpss
diff2_acf1	curvature	flat_spots
diff1_acf10	firstzero_ac	spike
diff2_acf10	localsimple_taures	lumpiness
alpha	histogram_mode	nonlinearity
beta	spreadrandomlocal	linearity
fluctuation	stability	avg_r$R^{45}CO_2/^{44}CO_2$
sd1st_der	outlierinclude	avg_r$R^{46}CO_2/^{44}CO_2$
trend	walker_propcross	avg_$\delta^{13}C$
entropy	garch_acf	sd_$\delta^{13}C$
motif_entropy	arch_r2	avg_$\delta^{18}O$
sampenc	garch_r2	avg_$\delta^{18}O/\delta^{13}C$
max_kl_shift	arch_acf	sd_$\delta^{18}O/\delta^{13}C$
time_level_shift	ARCH.LM	avg_calib_$\delta^{18}O$
max_var_shift	x_pacf5	sd_calib_$\delta^{18}O$
time_kl_shift	diff2x_pacf5	avg_pkArea
max_level_shift	e_acf1	sd_pkArea

b. Deltas Feature Space

avg_$\delta^{18}O$
avg_$\delta^{13}C$

c. Calibrated Deltas (calib-deltas) Feature Space

avg_$\delta^{13}C$
avg_calib_$\delta^{18}O$

d. NPDR-LURF and NPDR-LMan Feature Spaces: Selected on Each Train/Test Random Split (See Tables S4 – S9)

Figure A.7. Feature spaces used to confirm ML results. (a) Reduced redundancy (RR) features were pre-selected to limit the amount of correlation among features. We perform comparison analyses with average deltas (b), calibrated deltas (c), and NPDR selected features (d).

A.5 Machine Learning Model Replicates: Prediction of Biosignatures in Different Feature Spaces

Here we compare the RF biosignature model performance of the five feature spaces discussed in the previous section. NPDR-LURF and NPDR-LMan feature spaces are defined on each split using the training data to prevent bias, and RR feature space was defined as discussed in Sec. A.4. As mentioned before, NPDR-LASSO-Manhattan (NPDR-LMan) uses a standard distance metric for NPDR, while NPDR-LURF uses the non-isotropic URF proximity distance. The feature spaces composed of only average delta features are used to assess the difficulties of a ML prediction of biosignatures using only the average CO_2 isotope fractionation values and the improvement in accuracy of including additional IRMS-derived features like the standard deviations as well as the extracted TS features.

First, we present feature ranking results from NPDR and RF for the random 80%:20% train/test split used for analysis in the main manuscript using the five different feature spaces, referred to as Run 0 (Table A.3). NPDR-LURF and NPDR-LMan features are ranked by NPDR importance, while the Deltas, Calib-Deltas, and RR features are ranked by RF variable importance in the biosignature model, since NPDR was not used to select these feature spaces. Note that RF importance returns importance scores for all features used to train the model, in contrast to feature selection methods that use LASSO, in which the penalty shrinks regression coefficients for unimportant or highly correlated features to zero, removing them from the feature space. We therefore report the top ten RF importance features for the RR feature space. For NPDR-LURF feature selection in 80/20 Split Run 0, the hyperparameter penalty $\lambda = 2.16 \cdot 10^{-4}$ and for NPDR-LMan, $\lambda = 1.92 \cdot 10^{-4}$. To assess the stability of top-ranked NPDR and RR features, we include a random-split consistency fraction (RSCF) to express the percentage of splits for which the feature was selected as an important predictor for biosignatures. For

example, for NPDR-LURF, diff2_acf1 is 5/5, yielding RSCF=1 (Tables A.3-A.8, see also Table A.9).

For Run 0, NPDR-LMan selects several of the same features as NPDR-LURF, but also includes additional features not selected by NPDR-LURF, such as *motif_entropy*, *walker_propcross*, and *avg_rR$^{46}CO_2$/$^{44}CO_2$*. For both Deltas and Calib-Deltas feature spaces, RF variable importance ranks *avg_$\delta^{13}C$* as more important for biosignatures than *avg_$\delta^{18}O$* or *avg_calib_$\delta^{18}O$*, respectively. This is always true for the Calib-Deltas feature space across our total of six replicates; however, for the Deltas feature space, sometimes *avg_$\delta^{18}O$* is ranked higher than *avg_$\delta^{13}C$* (see Tables A.3 and A.5). This could indicate that the calibration of oxygen isotopes in the Calib-Deltas feature space is imparting some increased stability across sample folds for the biosignature model, compared with the Deltas feature space. The top ten RF features in the RR feature space are enriched in TS features, with only two IRMS-derived features selected, *avg_rR$^{45}CO_2$/$^{44}CO_2$* and *avg_$\delta^{13}C$*.

Table A.3.

Comparison of NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: Run 0

Rank	NPDR-LURF	RR (top ten)	NPDR-LMan	Deltas	Calib-Deltas
1.	avg_$R^{45}CO_2/^{44}CO_2$	max_kl_shift	avg_$R^{45}CO_2/^{44}CO_2$	avg_$\delta^{13}C$	avg_$\delta^{13}C$
2.	avg_$rR^{45}CO_2/^{44}CO_2$	localsimple_taures	avg_$rR^{45}CO_2/^{44}CO_2$	avg_$\delta^{18}O$	avg_calib_$\delta^{18}O$
3.	sd_$\delta^{18}O/\delta^{13}C$	fluctuation	sd_$\delta^{18}O/\delta^{13}C$	-	-
4.	diff2_acf1	avg_$rR^{45}CO_2/^{44}CO_2$	motif_entropy	-	-
5.	fluctuation	time_level_shift	avg_$rR^{46}CO_2/^{44}CO_2$	-	-
6.	time_kl_shift	avg_$\delta^{13}C$	fluctuation	-	-
7.	-	diff2x_pacf5	walker_propcross	-	-
8.	-	diff2_acf10	time_kl_shift	-	-
9.	-	diff2_acf1	-	-	-
10.	-	max_level_shift	-	-	-

Note. Run 0 corresponds to the random 80%:20% train/test split used for analysis in the main manuscript. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR), Deltas, and Calib-Deltas features are ranked by RF importance.

In 80/20 Split Run 1, for NPDR-LURF, the hyperparameter penalty $\lambda = 3.09 \cdot 10^{-5}$ and for NPDR-LMan, $\lambda = 2.07 \cdot 10^{-2}$. For this split, NPDR-LURF selects seven features as important predictors for biosignatures, adding *avg_rR^{46}CO$_2$/^{44}CO$_2$* and *walker_propcross* compared with Run 0 (Table A.4). NPDR-LMan again selects *avg_rR^{46}CO$_2$/^{44}CO$_2$*, *avg_rR^{45}CO$_2$/^{44}CO$_2$*, *avg_R^{45}CO$_2$/^{44}CO$_2$*, *fluctuation*, *walker_propcross*, and *motif_entropy*. In Run 1, RF using the Deltas feature space ranks *avg_δ^{18}O* as more important for biosignature prediction than *avg_δ^{13}C*. Again, the top RF features in the RR feature space are enriched for TS features, with the same two IRMS-derived features being selected.

Table A.4.

Comparison of NPDR-LURF, RR, *NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: Run 1*

Rank	NPDR-LURF	RR (top ten)	NPDR-LMan	Deltas	Calib-Deltas
1.	avg_rR^{45}CO$_2$/^{44}CO$_2$	max_kl_shift	sd_R^{13}C/^{12}C	avg_δ^{18}O	avg_δ^{13}C
2.	sd_δ^{18}O/δ^{13}C	fluctuation	avg_R^{45}CO$_2$/^{44}CO$_2$	avg_δ^{13}C	avg_calib_δ^{18}O
3.	walker_propcross	avg_rR^{45}CO$_2$/^{44}CO$_2$	avg_rR^{45}CO$_2$/^{44}CO$_2$	-	-
4.	diff2_acf1	localsimple_taures	motif_entropy	-	-
5.	avg_rR^{46}CO$_2$/^{44}CO$_2$	time_level_shift	avg_rR^{46}CO$_2$/^{44}CO$_2$	-	-
6.	fluctuation	diff2_acf10	fluctuation	-	-
7.	time_kl_shift	diff2x_pacf5	walker_propcross	-	-
8.	-	diff2_acf1	-	-	-
9.	-	avg_δ^{13C}	-	-	-
10.	-	max_var_shift	-	-	-

Note. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR, Deltas, and Calib-Deltas features are ranked by RF importance.

In 80/20 Split Run 2, for NPDR-LURF, $\lambda = 8.92 \cdot 10^{-5}$ and for NPDR-LMan, $\lambda = 4.42 \cdot 10^{-2}$. On this split, NPDR-LURF selects eleven features, adding features like *sd_δ^{13}C*, *sd_R^{13}C/^{12}C*, *avg_rδ^{45}CO$_2$*, *avg_δ^{45}CO$_2$*, and *motif_entropy* (Table A.5). When the entire dataset

is analyzed, *sd_$\delta^{13}C$* and *sd_$R^{13}C/^{12}C$*, and *avg_r$\delta^{45}CO_2$* and *avg_$\delta^{45}CO_2$* are in highly correlated clusters (see Fig. A.4), but for this train/test split, NPDR with the LASSO penalty selects both. A similar occurrence happens for NPDR-LMan, in which the algorithm selects both *avg_$R^{13}C/^{12}C$* and *avg_$\delta^{45}CO_2$*. On this split, it could be that there is enough variation in globally correlated features that the LASSO penalty preserves both. This is an atypical result in our set of six total runs for NPDR-LURF but occurs twice again with NPDR-LMan (see Tables A.6 and A.7). More research should be done on the nature of the distance metrics, and why NPDR-LURF seems to be more successful at reducing redundancy and is more selective overall in returning important predictors than NPDR-Manhattan. The top RF importance features in the RR space are again enriched for TS features, and two IRMS-derived features are selected, in this case *avg_$R^{45}CO_2/^{44}CO_2$* and *sd_pkArea*, the latter of which has never been selected as an important predictor for biosignatures.

In 80/20 Split Run 3, for NPDR-LURF, $\lambda = 1.31 \cdot 10^{-4}$ and for NPDR-LMan, $\lambda = 1.04 \cdot 10^{-3}$. For this split, NPDR-LURF selects five features while NPDR-LMan selects eleven (Table A.5). NPDR-LMan selects *sampenc* and *sampen_first*, which occur in a highly correlated cluster in the global data (see Fig. A.4). For the Deltas feature space, RF ranks *avg_$\delta^{18}O$* as more important for biosignatures than *avg_$\delta^{13}C$*, while the ranking for the Calib-Deltas remains consistent. The top RF features in the RR space again contain *avg_r$R^{45}CO_2/^{44}CO_2$* and *avg_$\delta^{13}C$*, and the rest are TS features.

Table A.5.

Comparison of NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: *Run 2*

Rank	NPDR-LURF	RR (top 11)	NPDR-LMan	Deltas	Calib-Deltas
1.	sd_$R^{13}C/^{12}C$	fluctuation	avg_$R^{45}CO_2/^{44}CO_2$	avg_$\delta^{13}C$	avg_$\delta^{13}C$
2.	avg_$R^{45}CO_2/^{44}CO_2$	max_kl_shift	avg_$R^{13}C/^{12}C$	avg_$\delta^{18}O$	avg_calib_$\delta^{18}O$
3.	avg_$rR^{45}CO_2/^{44}CO_2$	localsimple_taures	motif_entropy	-	-
4.	sd_$\delta^{18}O/\delta^{13}C$	time_level_shift	fluctuation	-	-
5.	motif_entropy	avg_$rR^{45}CO_2/^{44}CO_2$	avg_$\delta^{13}C$	-	-
6.	fluctuation	diff2_acf10	-	-	-
7.	diff2_acf1	diff2_acf1	-	-	-
8.	sd_$\delta^{13}C$	nonlinearity	-	-	-
9.	time_kl_shift	sd_pkArea	-	-	-
10.	avg_$r\delta^{45}CO_2$	max_var_shift	-	-	-
11.	avg_$\delta^{45}CO_2$	max_level_shift	-	-	-

Note. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR, Deltas, and Calib-Deltas features are ranked by RF importance.

Table A.6.

Comparison of NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: Run 3

Rank	NPDR-LURF	RR (top 11)	NPDR-LMan	Deltas	Calib-Deltas
1.	avg_$rR^{45}CO_2/^{44}CO_2$	fluctuation	avg_$R^{45}CO_2/^{44}CO_2$	avg_$\delta^{18}O$	avg_$\delta^{13}C$
2.	sd_$\delta^{18}O/\delta^{13}C$	max_kl_shift	sd_$R^{13}C/^{12}C$	avg_$\delta^{13}C$	avg_calib_$\delta^{18}O$
3.	diff2_acf1	avg_$rR^{45}CO_2/^{44}CO_2$	motif_entropy	-	-
4.	fluctuation	localsimple_taures	sd_$\delta^{18}O/\delta^{13}C$	-	-
5.	time_kl_shift	diff2x_pacf5	avg_$rR^{45}CO_2/^{44}CO_2$	-	-
6.	-	time_level_shift	sampenc	-	-
7.	-	avg_$\delta^{13}C$	fluctuation	-	-
8.	-	diff2_acf10	avg_$\delta^{18}O/\delta^{13}C$	-	-
9.	-	diff2_acf1	outlierinclude	-	-
10.	-	max_level_shift	time_kl_shift	-	-
11.	-	x_pacf5	sampen_first	-	-

Note. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR, Deltas, and Calib-Deltas features are ranked by RF importance.

In Run 4, for NPDR-LURF, $\lambda = 2.22 \cdot 10^{-2}$ and for NPDR-LMan, $\lambda = 2.90 \cdot 10^{-2}$. For this split, NPDR-LURF selects just three predictors for biosignatures, while NPDR-LMan selects eight (Table A.7). Again NPDR-LMan selects predictors that occur in highly correlated clusters together in the global data (*avg_r*$\delta^{45}CO_2$, *avg_*$R^{45}CO_2/^{44}CO_2$, and *avg_*$\delta^{45}CO_2$). Once again, the top RF features in the RR space contain *avg_r*$R^{45}CO_2/^{44}CO_2$ and *avg_*$\delta^{13}C$, and the rest are TS features.

Table A.7.

Comparison of NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: Run 4

Rank	NPDR-LURF	RR (top ten)	NPDR-LMan	Deltas	Calib-Deltas
1.	avg_$R^{45}CO_2/^{44}CO_2$	max_kl_shift	avg_$R^{45}CO_2/^{44}CO_2$	avg_$\delta^{13}C$	avg_$\delta^{13}C$
2.	avg_r$R^{45}CO_2/^{44}CO_2$	fluctuation	motif_entropy	avg_$\delta^{18}O$	avg_calib_$\delta^{18}O$
3.	diff2_acf1	avg_r$R^{45}CO_2/^{44}CO_2$	avg_r$R^{46}CO_2/^{44}CO_2$	-	-
4.	-	localsimple_taures	avg_r$R^{45}CO_2/^{44}CO_2$	-	-
5.	-	time_level_shift	fluctuation	-	-
6.	-	diff2x_pacf5	sd_$\delta^{18}O/\delta^{13}C$	-	-
7.	-	diff2_acf10	avg_r$\delta^{45}CO_2/\delta^{44}CO_2$	-	-
8.	-	avg_$\delta^{13}C$	avg_$\delta^{45}CO_2/\delta^{44}CO_2$	-	-
9.	-	diff2_acf1	-	-	-
10.	-	x_pacf5	-	-	-

Note. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR, Deltas, and Calib-Deltas features are ranked by RF importance.

In 80/20 Split Run 5, for NPDR-LURF, $\lambda = 1.63 \cdot 10^{-3}$ and for NPDR-LMan, $\lambda = 2.52 \cdot 10^{-3}$. NPDR-LURF and NPDR-LMan both select six features in this split, with some differences, with NPDR-LMan selecting *motif_entropy* instead of *fluctuation* and *diff2_acf1* like NPDR-LURF (Table A.8). RF importance again ranks *avg_r*$R^{45}CO_2/^{44}CO_2$ and *avg_*$\delta^{13}C$ as the only top IRMS-derived features.

Table A.8.

Comparison of NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas Feature Spaces: Run 5

Rank	NPDR-LURF	RR (top ten)	NPDR-LMan	Deltas	Calib-Deltas
1.	avg_rR$^{45}CO_2$/$^{44}CO_2$	max_kl_shift	avg_R$^{45}CO_2$/$^{44}CO_2$	avg_δ^{13}C	avg_δ^{13}C
2.	sd_δ^{18}O/δ^{13}C	fluctuation	sd_δ^{18}O/δ^{13}C	avg_δ^{18}O	avg_calib_δ^{18}O
3.	diff2_acf1	avg_rR$^{45}CO_2$/$^{44}CO_2$	avg_rR$^{45}CO_2$/$^{44}CO_2$	-	-
4.	fluctuation	time_level_shift	motif_entropy	-	-
5.	avg_rR$^{46}CO_2$/$^{44}CO_2$	max_var_shift	avg_rR$^{46}CO_2$/$^{44}CO_2$	-	-
6.	time_kl_shift	diff2x_pacf5	time_kl_shift	-	-
7.	-	max_level_shift		-	-
8.	-	avg_δ^{13}C	-	-	-
9.	-	e_acf1	-	-	-
10.	-	arch_acf	-	-	-

Note. NPDR-LURF and NPDR-LMan features are ranked by NPDR importance scores, while RR, Deltas, and Calib-Deltas features are ranked by RF importance.

Across the total of six runs, NPDR-LURF selects *diff2_acf1* and *avg_rR$^{45}CO_2$/$^{44}CO_2$* every time, yielding a random-split consistency fraction (RSCF) =1 (Table A.9). While *diff2_acf1* is not selected by NPDR-LMan throughout the repeated splits, *motif_entropy* and *avg_R$^{45}CO_2$/$^{44}CO_2$* have RSCF = 1 using this feature selection method (Table A.9). In the RR feature space, RF permutation importance yields RSCF = 1 for *max_kl_shift*, *time_level_shift*, and *avg_rR$^{45}CO_2$/$^{44}CO_2$* (Table A.9). NPDR-LURF additionally selects the fewer features across the six replicates compared with NPDR-LMan, indicating that it has a tendency to create a more reduced selected feature space. In the RR space, RF importance has no statistical threshold to reduce the selected number of features (ten is an arbitrary threshold), again illustrating the usefulness of the LASSO penalty for model interpretation.

Table A.9.

Random Split Consistency Fraction (RSCF) for Three Feature Spaces

NPDR-LURF	RR	NPDR-LMan
diff2_acf1 (1.0)	max_kl_shift (1.0)	avg_$R^{45}CO_2/^{44}CO_2$ (1.0)
avg_r$R^{45}CO_2/^{44}CO_2$ (1.0)	avg_r$R^{45}CO_2/^{44}CO_2$ (1.0)	motif_entropy (1.0)
sd_$\delta^{18}O/\delta^{13}C$ (0.8)	time_level_shift (1.0)	avg_r$R^{45}CO_2/^{44}CO_2$ (0.8)
fluctuation (0.8)	fluctuation (1.0)	fluctuation (0.8)
time_kl_shift (0.8)	avg_$\delta^{13}C$ (0.8)	avg_r$R^{46}CO_2/^{44}CO_2$ (0.7)
avg_$R^{45}CO_2/^{44}CO_2$ (0.5)	diff2_acf10 (0.8)	sd_$\delta^{18}O/\delta^{13}C$ (0.7)
avg_r$R^{46}CO_2/^{44}CO_2$ (0.3)	diff2_acf1 (0.8)	sd_$R^{13}C/^{12}C$ (0.5)
avg_$\delta^{45}CO_2$ (0.2)	localsimple_taures (0.7)	time_kl_shift (0.3)
walker_propross (0.2)	diff2x_pacf5 (0.7)	walker_propcross (0.3)
sd_$R^{13}C/^{12}C$ (0.2)	max_level_shift (0.5)	avg_$\delta^{13}C$ (0.2)
sd_$\delta^{13}C$ (0.2)	max_var_shift (0.5)	sampenc (0.2)
avg_r$\delta^{45}CO_2$ (0.2)	nonlinearity (0.2)	avg_$\delta^{18}O/\delta^{13}C$ (0.2)
motif_entropy (0.2)	sd_pkArea (0.2)	outlierinclude (0.2)
	x_pacf5 (0.2)	time_kl_shift (0.2)
	e_acf1 (0.2)	sampenfirst (0.2)
	arch_acf (0.2)	avg_r$\delta^{45}CO_2/\delta^{44}CO_2$ (0.2)
		avg_$\delta^{45}CO_2/\delta^{44}CO_2$ (0.2)

Note. The RSCF is the percentage of times the feature is selected out of the six random splits. NPDR-LURF and NPDR-LMan features are selected using NPDR with the LASSO penalty and URFP and Manhattan distances, respectively. The RR feature importance was determined using the top ten RF permutation importance features for the six random splits.

To construct RF classifiers for biosignatures, we use the R library ranger (Wright & Ziegler, 2017), a fast implementation of RF based on the original Fortran implementation (Breiman, 2001). We use class imbalance weighting (class weight = 1/class size) and cross-validation (CV) with five folds to tune standard RF hyperparameters *mtry* (number of variables to randomly sample at each split), *splitrule* (method to split nodes), *min.node.size* (minimum number of samples in a terminal node), and *ntrees* (number of trees). We tuned the hyperparameters for each RF model on every train/test random split. We use a minimum of 5000

trees for every RF model and more if tuning results show increased model performance with more trees.

a. Run 0

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	8	extratrees	1	5000
Deltas	2	extratrees	4	13000
Calib-Deltas	2	extratrees	3	5000
NPDR-LURF	2	hellinger	18	5000
NPDR-LMan	3	hellinger	5	5000

b. Run 1

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	6	extratrees	1	5000
Deltas	2	extratrees	1	5000
Calib-Deltas	2	extratrees	1	5000
NPDR-LURF	2	extratrees	1	5000
NPDR-LMan	2	extratrees	5	5000

c. Run 2

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	6	extratrees	1	5000
Deltas	2	extratrees	1	5000
Calib-Deltas	2	extratrees	1	5000
NPDR-LURF	2	extratrees	1	5000
NPDR-LMan	2	extratrees	5	5000

d. Run 3

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	14	extratrees	2	5000
Deltas	2	extratrees	1	5000
Calib-Deltas	2	extratrees	1	5000
NPDR-LURF	2	extratrees	1	5000
NPDR-LMan	2	extratrees	5	5000

e. Run 4

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	6	extratrees	13	5000
Deltas	2	extratrees	1	5000
Calib-Deltas	2	extratrees	1	5000
NPDR-LURF	2	extratrees	1	5000
NPDR-LMan	2	extratrees	5	5000

f. Run 5

Variable Space	mtry	splitrule	min.node.size	ntrees
RR	11	gini	1	5000
Deltas	2	extratrees	1	5000
Calib-Deltas	2	extratrees	1	5000
NPDR-LURF	2	extratrees	1	5000
NPDR-LMan	2	extratrees	5	5000

Figure A.8. Results of hyperparameter tuning for 80:20 train/test splits Runs 0-5 (**a-f**) in the five feature spaces used for comparison of RF biosignature model performance.

Final RF models are fit using tuned hyperparameters (Fig. A.8 a-f) for the five feature spaces for 80/20 train/test splits Runs 0-5 (Fig. A.9). For Run 0, as discussed in the main manuscript, the NPDR-LURF, RR, and NPDR-LMan RF biosignature models all achieve high training accuracy, outperforming the Delts and Calib-Deltas features models by more than 10%. Overall, NPDR-LURF selected features outperform NPDR-LMan selected features, achieving a mean training accuracy of 90.0%, compared with 86.4%. The mean and median NPDR-LURF training accuracies are comparable to the RR mean test accuracy of 89.2% (Fig. A.9). The NPDR-LURF selected features are therefore capturing predictive information for biosignatures

using significantly fewer features than the RR model, and furthermore, this feature space is enriched for statistical interactions compared with the RR feature space.

The NPDR-LURF selected features yield similarly high test accuracies, achieving a mean test accuracy of 87.3%, comparable to the RR feature space mean test accuracy of 89.7%. This confirms that NPDR-LURF selects important predictors for biosignatures on holdout data in a reproducible way, performing similarly to a much more high-dimensional feature space and significantly outperforming RF models for biosignatures created using only delta features (Fig. A.10). Classification tables, or confusion matrices, for Runs 0-5 illustrate that NPDR-LURF biosignature models have a more proportional class error for predictions than the other models, in addition to being more parsimonious and performing with comparably high accuracy as the RR features models (Figs. A.11 – A.16).

RF Biosignature Training Accuracies for Five Feature Spaces

	NPDR-LURF	RR	NPDR-LMan	Deltas	Calib-Deltas
**Run 0	**87.9%**	**86.4%**	**87.1%**	**72.9%**	**74.3%**
Run 1	90.0%	88.6%	86.4%	70.7%	70.0%
Run 2	92.9%	88.2%	77.9%	72.1%	74.3%
Run 3	92.9%	90.0%	88.6%	70.7%	72.9%
Run 4	85.7%	87.1%	84.3%	72.9%	70.7%
Run 5	90.7%	87.9%	86.4%	71.4%	69.3%
median	90.4%	88.1%	86.4%	71.8%	71.8%
mean	90.0%	88.0%	85.1%	71.8%	71.9%

** **Run 0** uses the train/test data analyzed in the main manuscript.
- **Runs 1 – 5** use additional random train/test splits (80 train:20 test)
- mean and median are reported for all 6 runs

Figure A.9. Training accuracies for NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas feature spaces for random 80/20 train/test splits Runs 0-5.

RF Biosignature Test Accuracies for Five Feature Spaces

	NPDR-LURF	RR	NPDR-LMan	Deltas	Calib-Deltas
**Run 0	**88.2%**	**91.2%**	**78.8%**	**61.8%**	**73.5%**
Run 1	88.2%	91.2%	82.4%	67.6%	85.3%
Run 2	85.3%	88.2%	79.4%	79.4%	85.3%
Run 3	82.4%	85.3%	82.4%	73.5%	79.4%
Run 4	88.2%	88.2%	91.2%	70.6%	76.5%
Run 5	91.2%	91.2%	82.4%	79.4%	76.5%
median	88.2%	89.7%	82.4%	72.1%	78.0%
mean	87.3%	89.2%	82.8%	72.1%	79.4%

** **Run 0** uses the train/test data analyzed in the main manuscript.
- **Runs 1 – 5** use additional random train/test splits (80 train:20 test)
- mean and median are reported for all 6 runs

Figure A.10. Test accuracies for NPDR-LURF, RR, NPDR-LMan, Deltas, and Calib-Deltas feature spaces for random 80/20 train/test splits Runs 0-5.

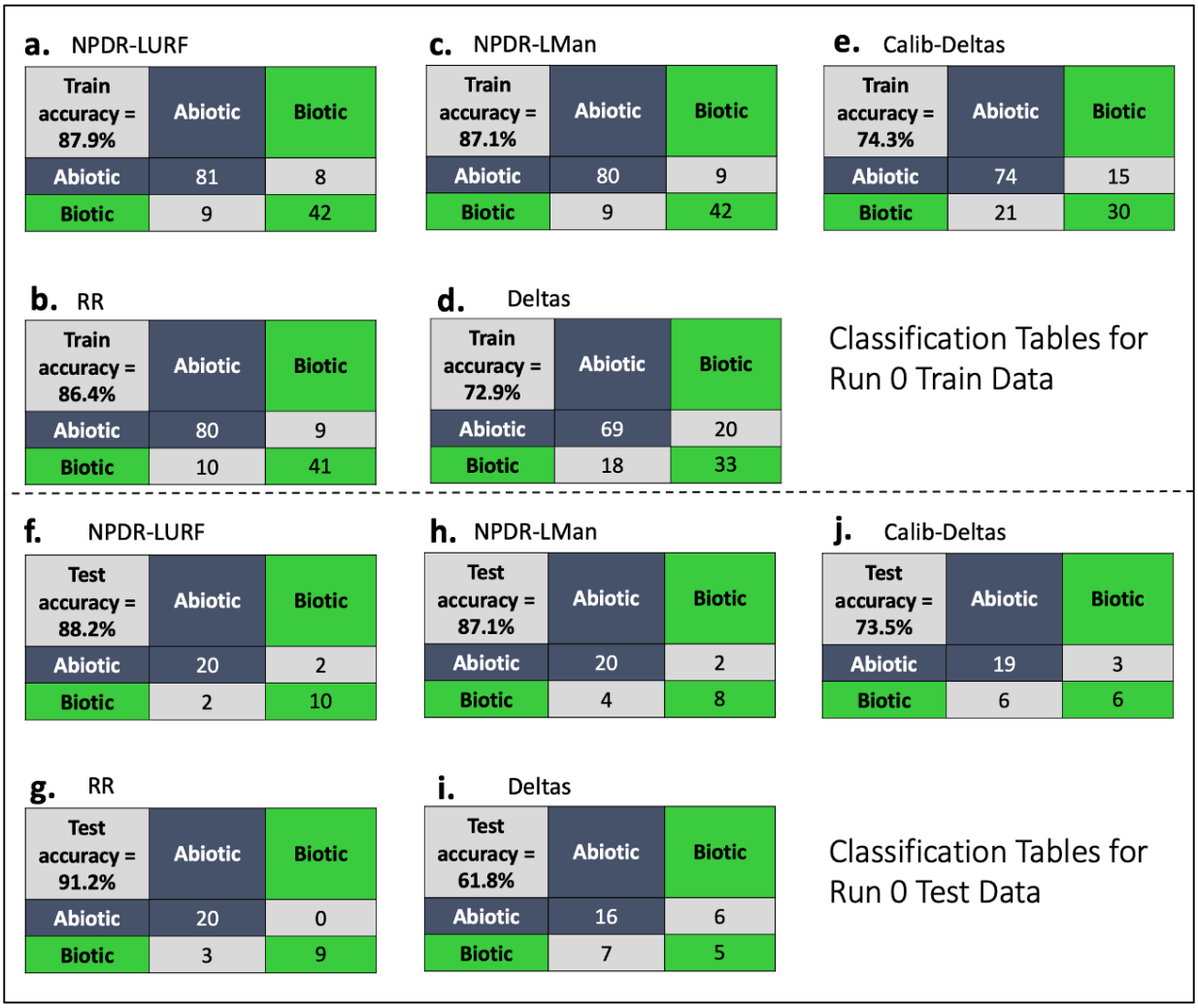


Figure A.11. Classification tables for RF models trained and tested on the random 80:20 train/test split used for analysis in the main manuscript, Run 0. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

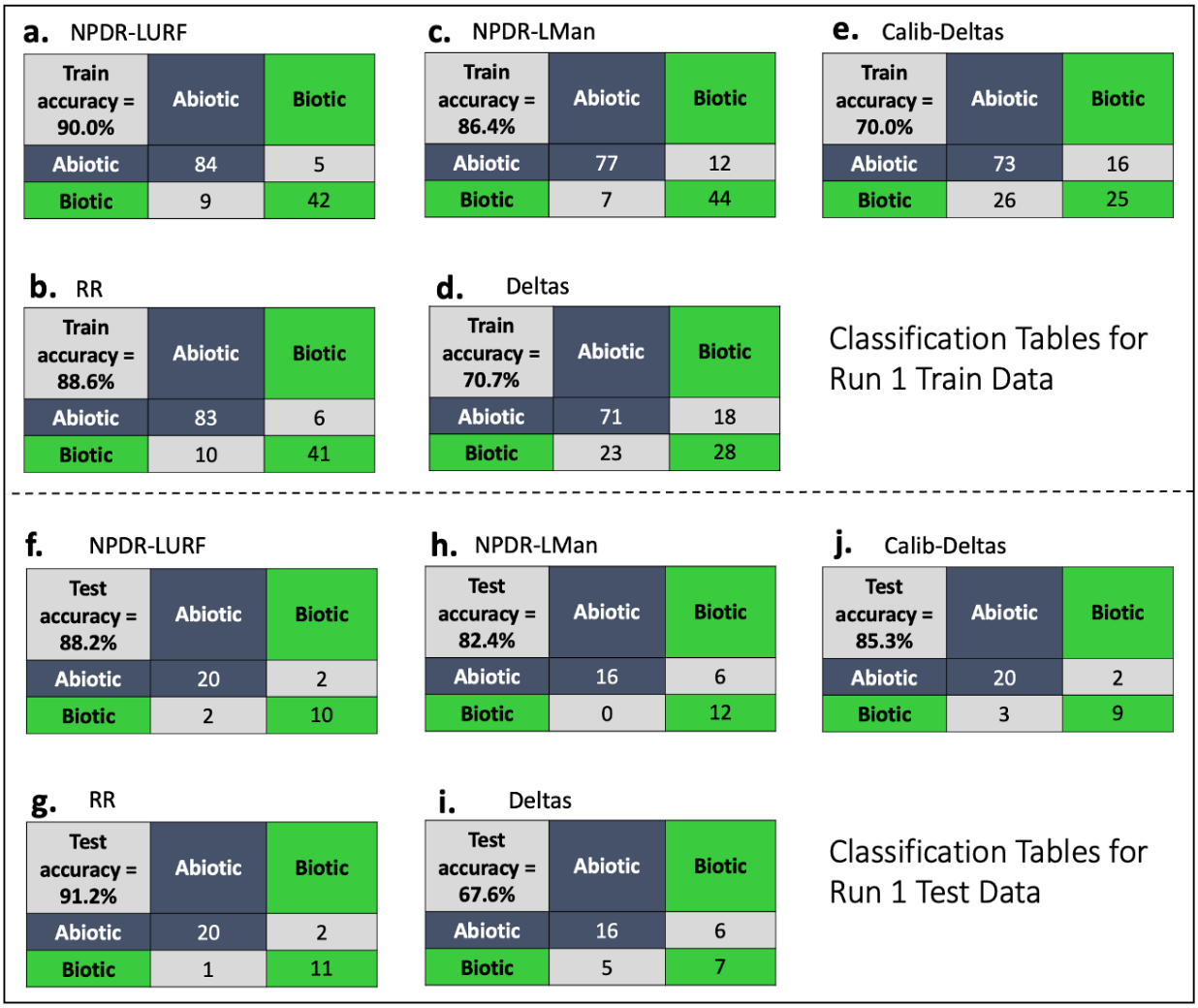


Figure A.12. Classification tables for RF models trained and tested on the random 80:20 train/test split Run 1. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

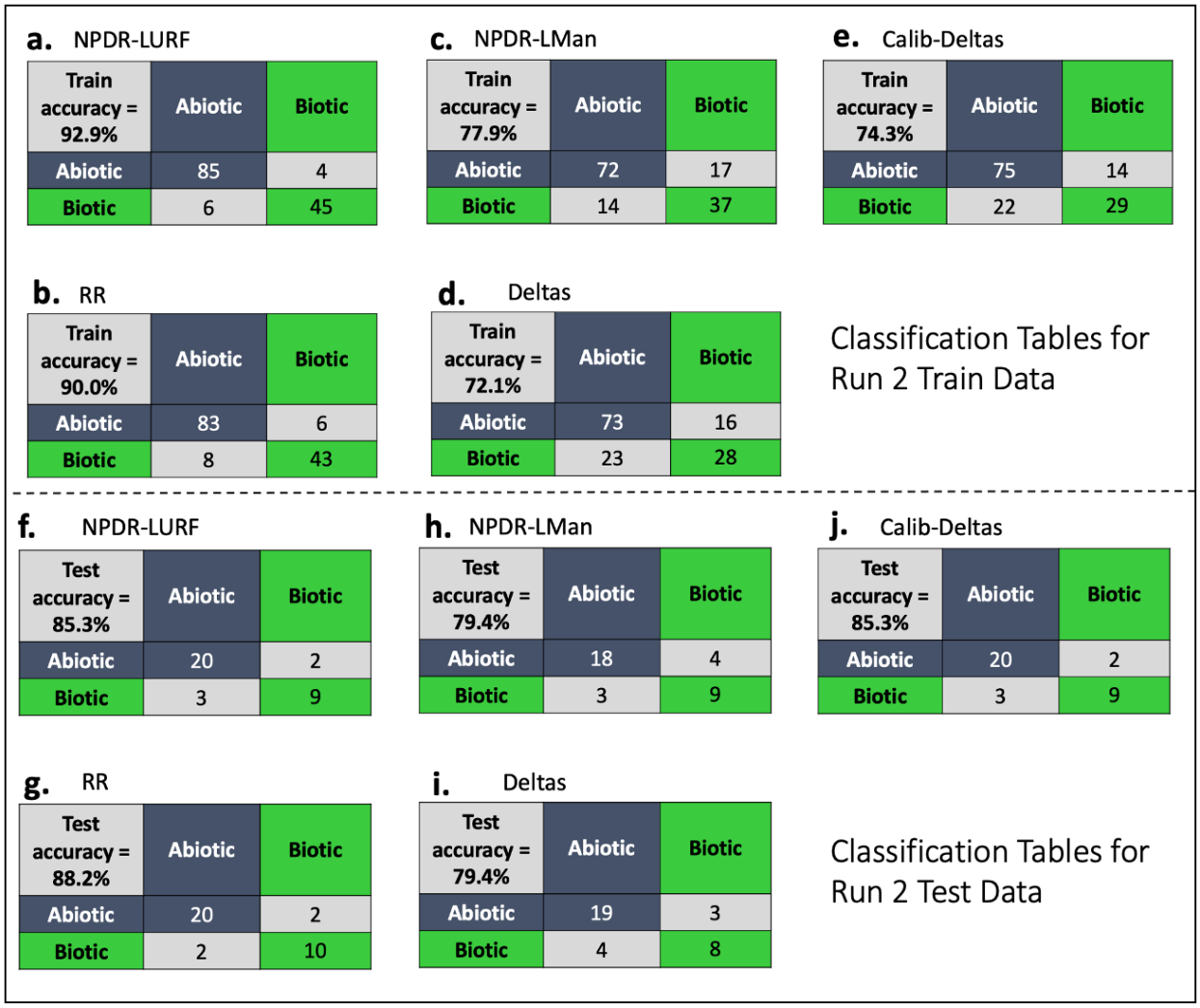


Figure A.13. Classification tables for RF models trained and tested on the random 80:20 train/test split Run 2. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

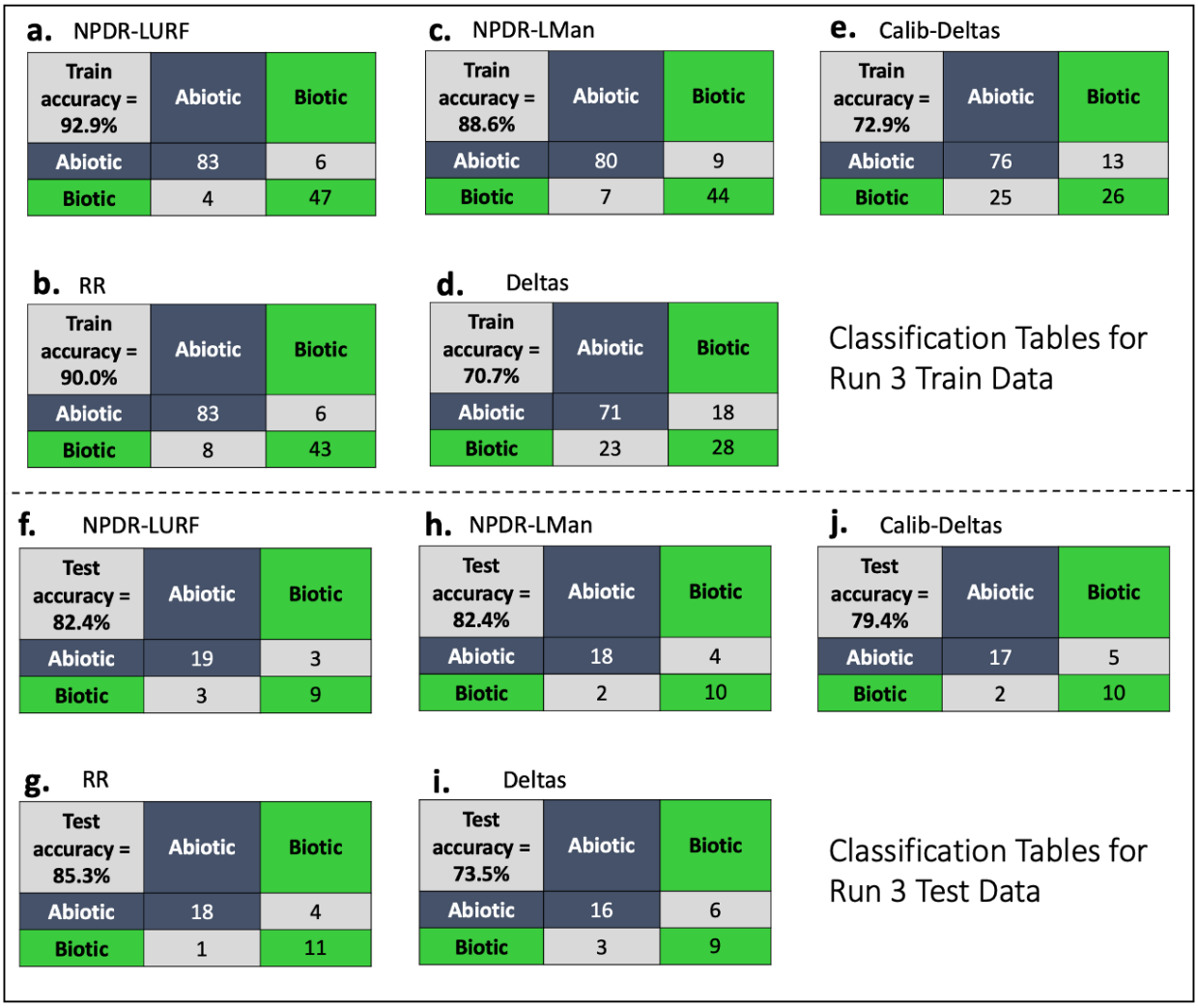


Figure A.14. Classification tables for RF models trained and tested on the random 80:20 train/test split Run 3. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

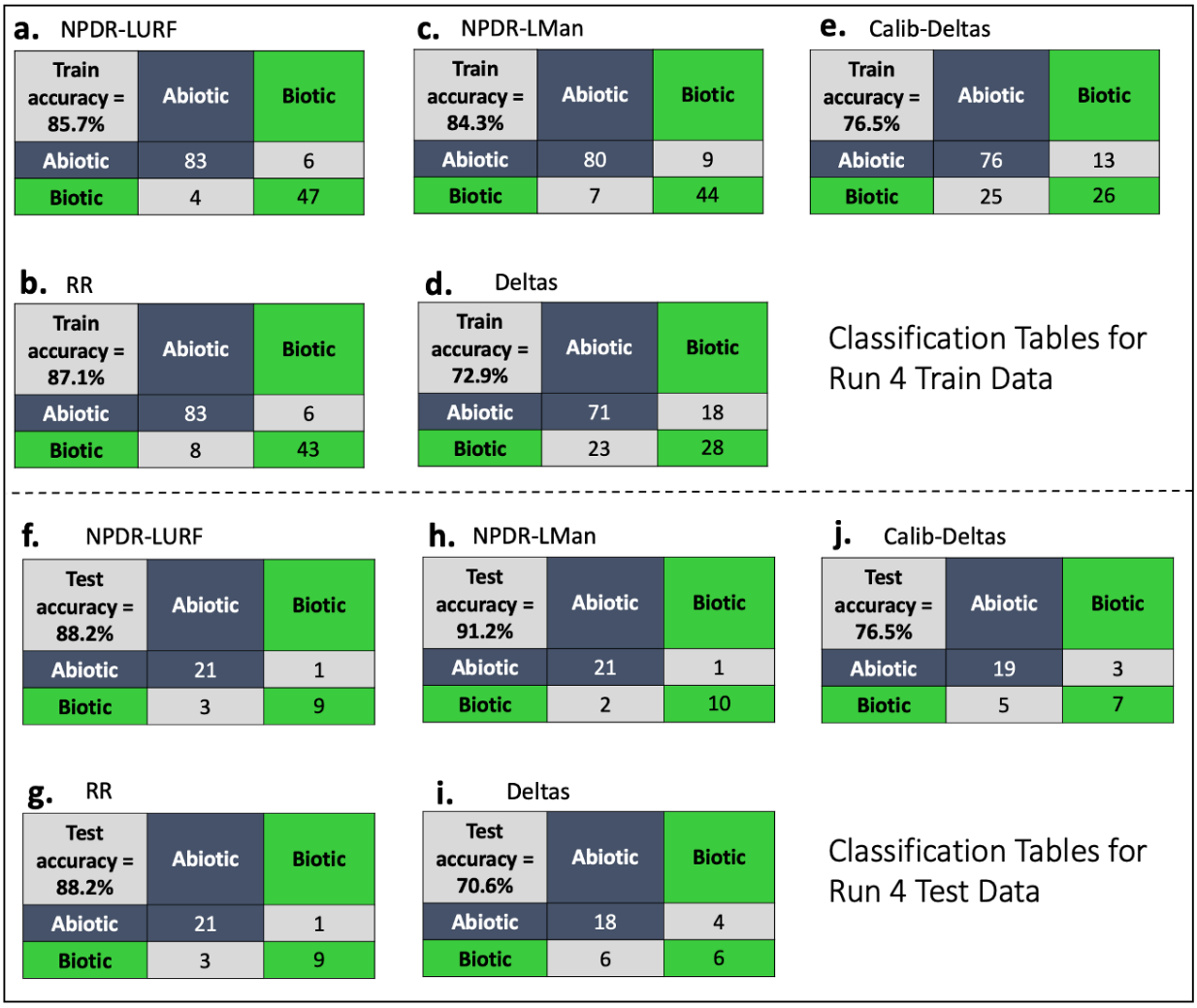


Figure A.15. Classification tables for RF models trained and tested on the random 80:20 train/test split Run 4. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

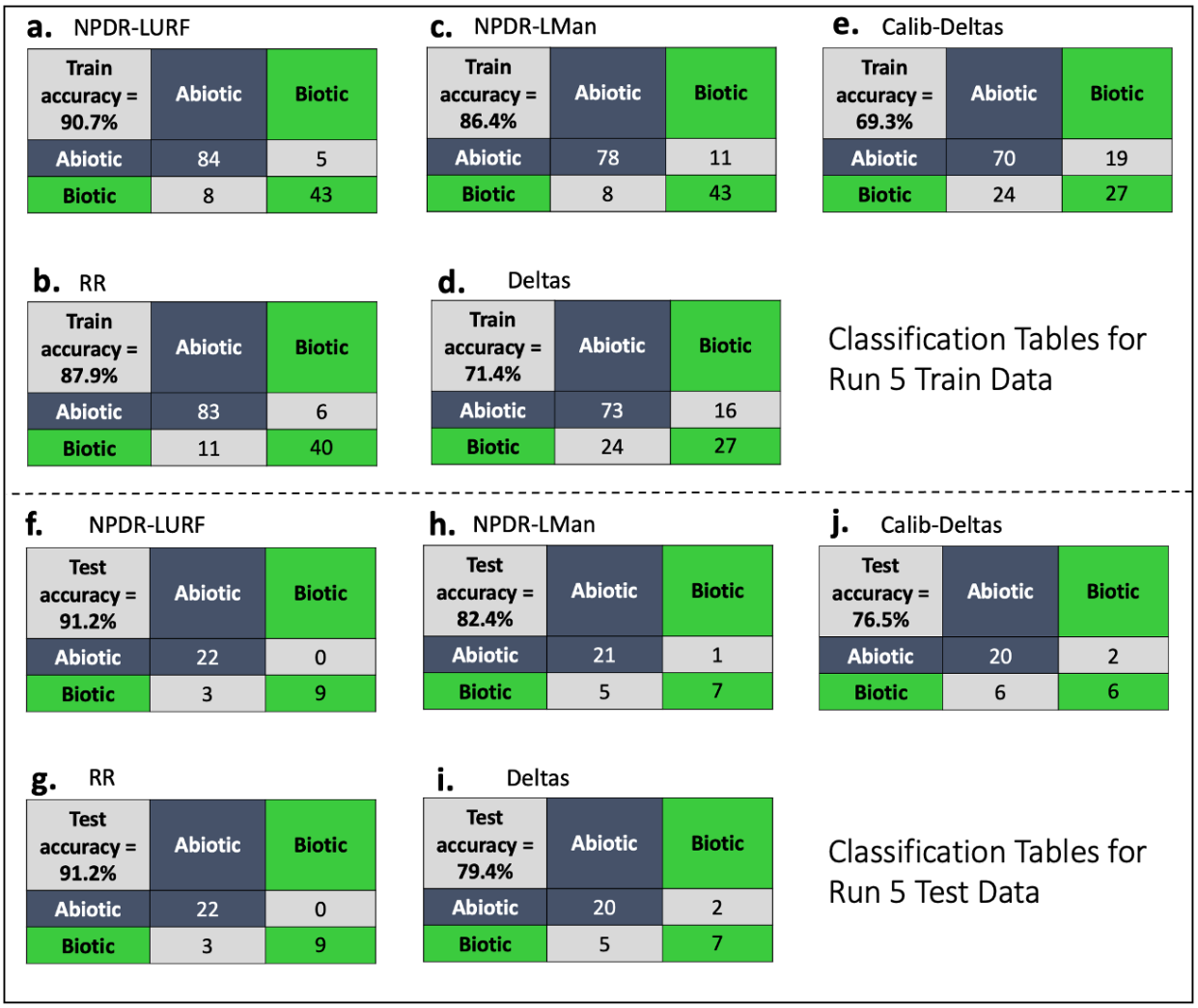


Figure A.16. Classification tables for RF models trained and tested on the random 80:20 train/test split Run 5. Dark grey and green diagonals represent the number of correct predictions, and light grey diagonals represent incorrect predictions (actual class is represented by the rows; predicted class by the columns).

A6: True and False Predictions using Local Random Forest Variable Importance Scores

To further illustrate the general behavior of false predictions in the BOW-δCO_2 dataset and the utility of single-sample variable importance for identifying false predictions, we include additional sample-level variable importance cases of true and false predictions for NPDR-LURF features selected in a typical run, Run 0 (Figs. A.17 and A.18). Recall from Appendix A.1 that our dataset contains four subsets of samples that are related to each other through brine chemistry and/or microbial content, which we label Microbial Mud, Abiotic, Europa Abiotic, and Europa Biotic. These subsets themselves contain a variety of salt compositions but under similar biogeochemical conditions. If the features for all samples were identically distributed, we would expect that the misclassification rate would be the same for each group. However, we find that the RF biosignature classifier has more difficulty with some subsets of our experimental dataset than others; indicating some abiotic or biotic samples may be geochemically straightforward to classify while others are good mimickers of biotic fractionations.

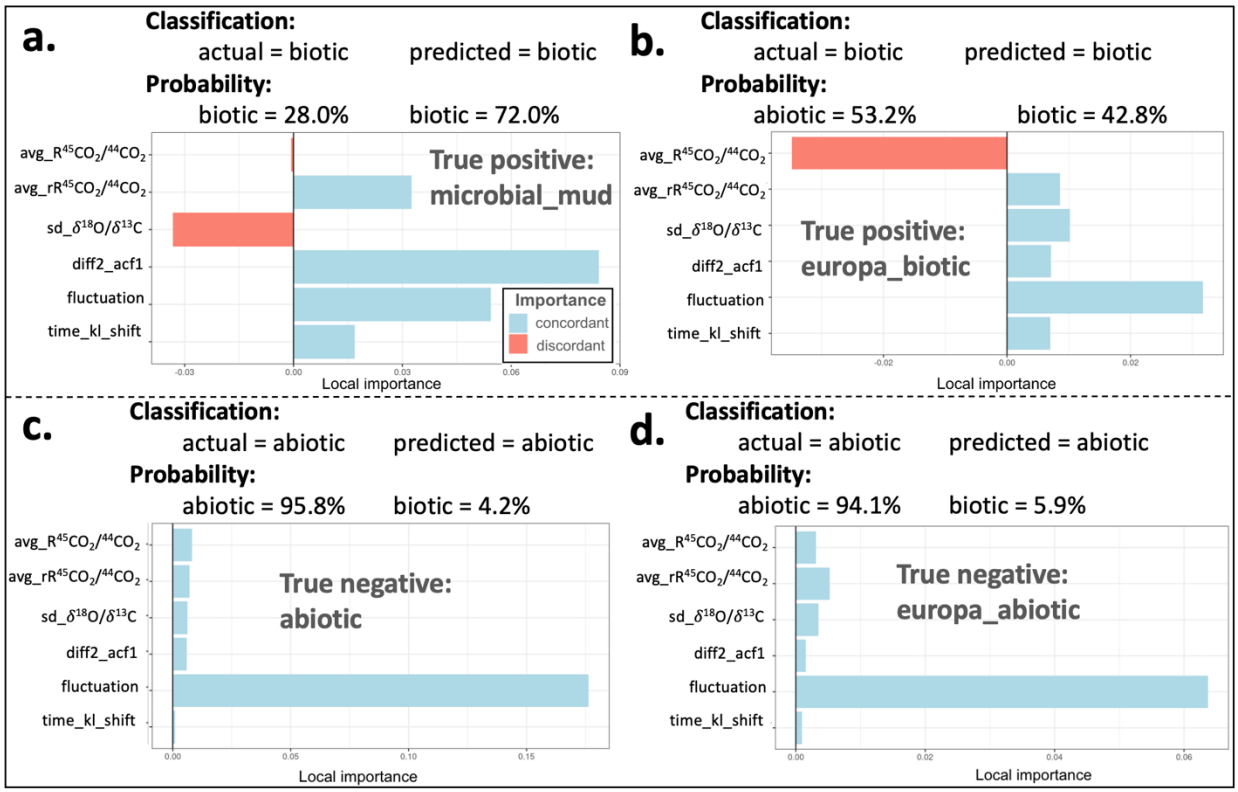


Figure A.17. True predictions in the biosignature data. (a) A true positive prediction of a Microbial Mud samples shows large-magnitude concordant variable importance scores and a high prediction probability. (b) A true positive prediction of a Europa Biotic sample shows a large magnitude discordant importance score, with a probability forest prediction probability that contradicts the classifier, indicating this is a difficult sample to classify. True negatives for Abiotic and Europa Abiotic samples (c) and (d) show all concordant local variable importance scores and high prediction probabilities.

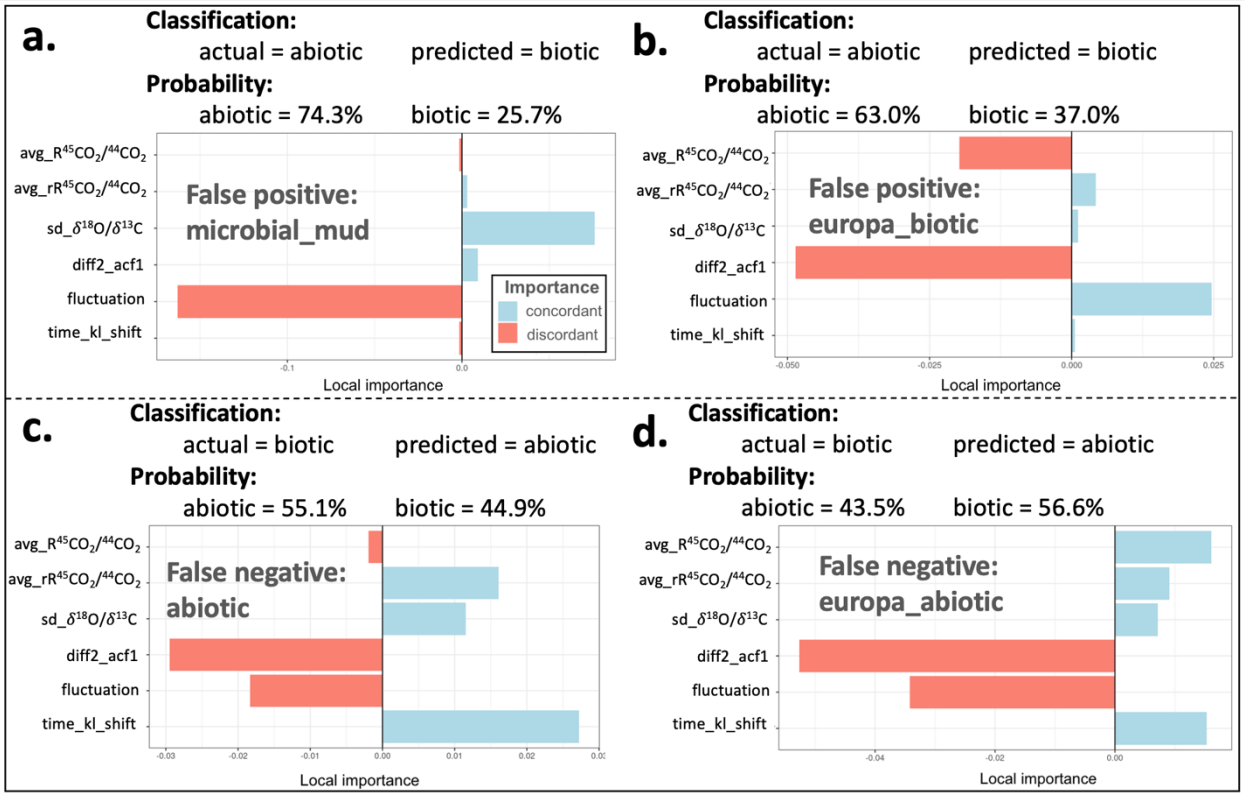


Figure A.18. False predictions in the biosignature data. (a) A false positive prediction of a Microbial Mud samples shows a large-magnitude discordant variable importance score and a high prediction probability. (b) A false negative prediction of a Europa Biotic sample shows two larger magnitude discordant importance scores. (c) A false negative for an Abiotic sample shows large magnitude discordant local variable importance scores and a low probability forest prediction probability. (d) A false negative for a Europa Abiotic sample likewise shows discordant scores and a low prediction probability.

As we might expect, we find that samples into which we introduced non-biogenic organics such as microbial substrate both with and without microbes (*i.e.*, the *Europa biotic* and *Europa abiotic* subsets) are more difficult for the model to classify than the simpler samples found in the *abiotic* dataset or those in the *microbial mud* samples, which likely stand out to the model due to the biological heterogeneity in the samples. For our dataset, the expected number of misclassifications is equal to the number of samples in each subset times the overall misclassification rate (MR) of the model, MR = 12.1% (Table A.10). For samples in the *Europa abiotic* group, the number of misclassified samples (7) is slightly higher but consistent with the

expected number (6.5). However, for samples in the *Europa biotic* group, the number of misclassified samples (10) is significantly higher than the expected number (4). In contrast, for samples in the *abiotic* and *microbial mud* groups, the actual misclassification rates are significantly lower than the expected rates.

Table A.10.

Analysis of misclassified samples by subset.

Data subsets (174 total samples)	Number of samples	Percent of total samples	Actual number of misclassified	Percent misclassified	Expected number of misclassified
Microbial_mud	30	17.2 (30/174)	1	4.8 (1/21)	3.6
Europa_biotic	33	19.0 (33/174)	10	47.6 (10/21)	4.0
Europa_abiotic	54	31.0 (54/174)	7	33.3 (7/21)	6.5
Abiotic	57	32.8 (57/174)	3	14.3 (3/21)	6.88

Note. Values are for misclassified samples from Run 0. The total number of misclassified samples is M = 21 and the total number of samples is N = 174. Therefore, the overall misclassification rate is MR = M/N = 21/174 = 0.121. For each group (row), the "Expected number of misclassified" = "Number of samples" · MR. The actual number of misclassified samples is higher than expected for Europa biotic samples, while the actual and expected misclassified samples for Europa abiotic are similar. Abiotic and Microbial mud samples are misclassified at lower rate than expected.

APPENDIX B

ADDITIONAL LOCAL-NPDR FEATURE IMPORTANCE RESULTS

This Supplemental Information Appendix contains additional tools for data and machine learning (ML) model prediction interpretation, as well as more details about the local-NPDR (Nearest-neighbors Projected Distance Regression) feature importance scores and how the method can be used to understand ML datasets better. In the next section (B.1), we present an example of a Regression-based Association-Interaction Network (RAIN) for the biosignature data, generated by a tool in the NPDR R library. This network visualization illustrates how variable main effects and interactions between features may work together to inform model predictions. In Sec. B.2, we present additional local-NPDR analysis that can be performed to better understand which variables are driving false positive (FP), false negative (FN), true positive (TP), and true negative (TN) predictions. Understanding which variables drive true and false predictions can be helpful in interpretation of falsely classified samples which have high local variable importance scores. In Sec. B.3, the distribution of local-RF (Random Forest) and local-NPDR for different prediction types and overall is discussed. The distribution of local importance scores is used to manually provide a threshold for a total local score (TLS) by which to quarantine samples that may be falsely predicted.

B.1 RAIN for Biosignature Data using NPDR-LURF Selected Features

The NPDR R library contains functions to quantify the main and interaction effects for variables and variable pairs, and additionally provides a network representation of these effects in a graph called a Regression-based Association-Interaction Network (RAIN), first used to understand how interactions in genetic and gene expression data affect disease-state outcomes (McKinney et al., 2009; Lareau et al., 2015). An interaction-magnitude threshold may be applied

to determine which interactions to visualize, if the dataset has many. For the biosignature data, we find that we can represent all interactions in the global-NPDR selected feature space (Fig. B.1). In an RAIN, the node size represents the magnitude of the variable's main effect, and the color of the node indicates the direction of the effect. For example, the top-ranked global-NPDR feature, *avg_rR*$^{45}CO_2/^{44}CO_2$ (gray node 1. Fig. B.1) has a main effect that increases the chances of an *abiotic* classification and participates in four interactions with other variables. The largest two interactions of *avg_rR*$^{45}CO_2/^{44}CO_2$ occur with *time_kl_shift* and *diff2_acf1*, and have a *biotic* and *abiotic* direction, respectively. While *diff2_acf1* has a similarly large *abiotic* main effect as *avg_rR*$^{45}CO_2/^{44}CO_2$, *time_kl_shift* has a small, *biotic* main effect. The interpretation of the global-NPDR selected features is discussed in Ref. (Clough et al., 2025); in summary, *time_kl_shift* and *diff2_acf1* are measures of information entropy and self-similarity, respectively.

Main and interaction effects for variables can be quantified using the NPDR function, `regain`, which computes information about how each variable may be contributing, individually and jointly, to model predictions (Table B.1). The output of the `regain` function is a p x p matrix (where p = number of variables) whose diagonal elements represent variable main effects and non-diagonal elements represent the interaction effect between the row and column variables. A network centrality called Epistasis Rank (ER) can be calculated from the RAIN matrix for each variable. ER was originally called SNPrank (Davis et al., 2010) because it was developed for genetic data. ER is similar to Google's PageRank for web searches, but rather than binary connections between nodes, ER uses signed, weighted interaction strengths, and rather than disallowing self-connections, ER uses the RAIN diagonal of main effects.

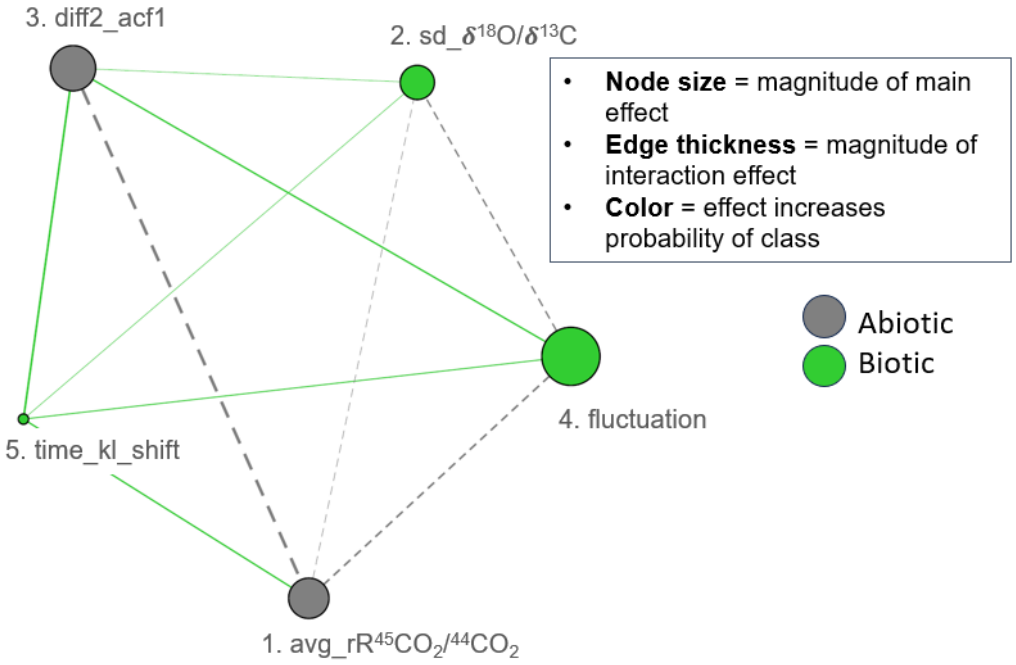


Figure B.1. Regression-based Association-Interaction Network (RAIN) for the biosignature dataset. RAINs show the variable main effects (magnitude indicated by node size) and interaction effects between pairs of variables (magnitude indicated by edge width) as well as a classification direction, that is, whether the effect increases the probability of *biotic* or *abiotic* prediction (indicated by color). Variables are ordered according their global-NPDR rank. This network shows that each feature participates in four interactions, and there is no "hub" in the network.

Table B.1. Main and interaction effects of global-NPDR-LURF selected features for the biosignature dataset. Epistasis rank quantifies the main effect of a variable and the cumulative network effect of interactions with other variables.

Global-NDPR feature	Epistasis rank	Mean interaction effect (magnitude)	Main effect
1. avg_rR$^{45}CO_2$/$^{44}CO_2$	5.11	0.96	-2.83
2. sd_$\boldsymbol{\delta}^{18}$O/$\boldsymbol{\delta}^{13}$C	1.98	0.39	2.40
3. diff2_acf1	-1.56	0.98	-3.18
4. fluctuation	0.92	0.83	4.08
5. time_kl_shift	-5.45	0.64	0.72

B.2 Local-NPDR Mean Scores by Prediction Type for Each Feature

In this section we present another way to analyze local feature importance scores grouped by prediction type (FP/FN/TP/TN). Using the biosignature data as an example, we calculate the mean variable importance score for each prediction type (Fig. B.2). Recall that the RF model trained on the global-NPDR-LURF selected features yields a train/test accuracy of 90.7% / 91.2%. This dataset has a class imbalance = 0.64 where the minority class is *biotic*. Recall also that local-NPDR scores may be affected by class imbalance. This effect can be seen in the large positive mean local-NPDR variable importance score for *avg_rR$^{45}CO_2$/$^{44}CO_2$* in both true positive samples (Fig. B.2a) and false negative samples (Fig. B.2d). Additionally, this variable has a negative score for true negative predictions (Fig. B.2c). This helps the interpretation of biosignature model predictions. For example, a large positive score for *avg_rR$^{45}CO_2$/$^{44}CO_2$* could mean either a likely true positive or false negative prediction, both of which involve the biotic class. To decide between which is most likely, other variable local scores can be examined, such as *diff2_acf1*, which has on average a positive local-NPDR importance score for true positive predictions but is on average negative for false negative predictions.

Analyses such as the one above can help researchers build trust in their ML model predictions. Furthermore, understanding which variables tend to be informative for which types of predictions can help researchers decide if quarantined samples should be "released" or not, and used for future model training. Additionally, these plots show that no individual variable is positive in all "true" predictions nor negative in all "false" predictions. This could be due to class imbalance and/or small sample size in this case.

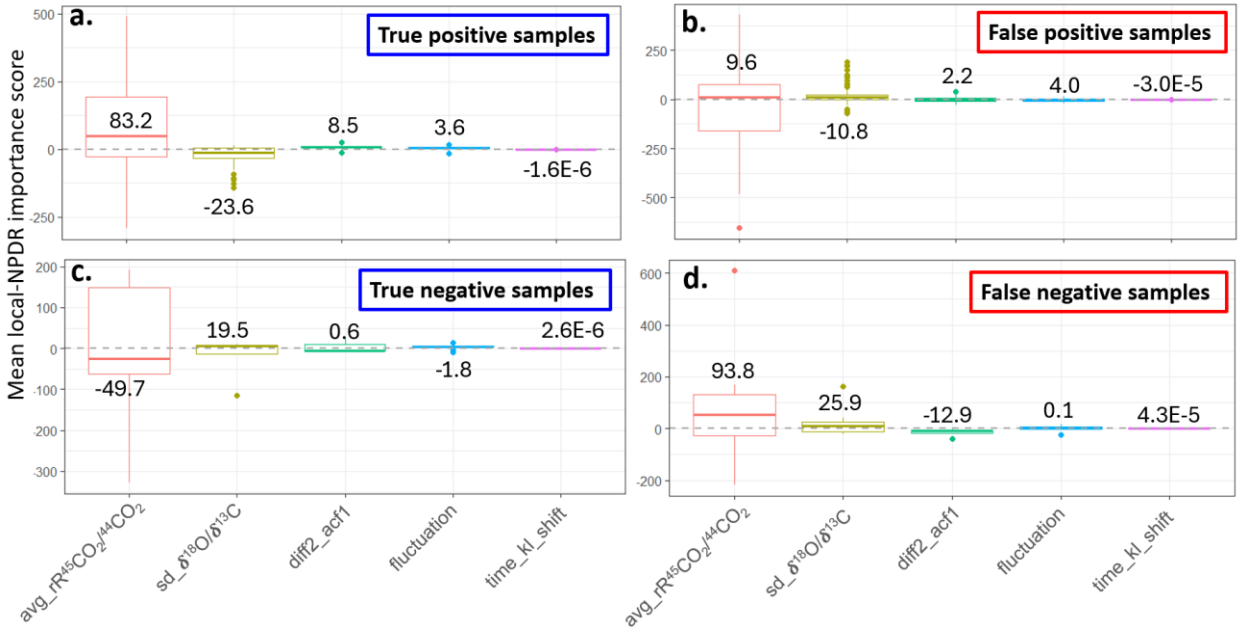


Figure B.2. Mean total local-NPDR importance scores for each variable by prediction type (TP, TN, FP, and FN). **(a)** The mean local-NPDR score for variables in true positive predictions show mostly positive scores, with one large negative mean score, for *sd_$\delta^{18}O$/ $\delta^{13}C$*. **(b)** For false positive predictions, the only variable with a large negative score is *sd_$\delta^{18}O$/ $\delta^{13}C$*. This indicates that while *sd_$\delta^{18}O$/ $\delta^{13}C$* is not informative for diagnosing true positive predictions, it is important for diagnosing false positives. **(c)** The variable *avg_r$R^{45}CO_2/^{44}CO_2$* is on average negative for true negative predictions, while *sd_$\delta^{18}O$/ $\delta^{13}C$* is positive. **(d)** For false negative predictions, neither *avg_r$R^{45}CO_2/^{44}CO_2$* nor *sd_$\delta^{18}O$/ $\delta^{13}C$* have negative average local-NPDR scores, while the variable *diff2_acf1* is negative.

B.3 Total Local Scores for Training Samples

Total local-NPDR and total local-RF importance scores for training samples are discussed in this section. The mean values of total local-NPDR and local-RF variable importance scores with respect to true and false predictions are statistically significant for both methods in all training datasets (Fig. B.3).

Local-NPDR scores for each prediction type (FP/FN/TP/TN) for training samples in the three simulated datasets show similar trends to the test data discussed in the main manuscript (Fig. B.4, compare with Fig. 12 in Sec. 3.4.2). For the two simulated datasets with class imbalance, local-NPDR mean TLS for FP predictions are higher than either FN or TN scores, because of the effects of class imbalance (Fig. B.4a and b). In contrast, the simulated dataset with class balance shows that FP and FN scores are both lower than TP and TN scores (Fig. B.4c). The average RF prediction probabilities for each prediction type in the three simulated training datasets show that, as with the test data, the RF probabilities along with the local-NPDR scores will help diagnose false predictions (Fig. B.4d, e, and f).

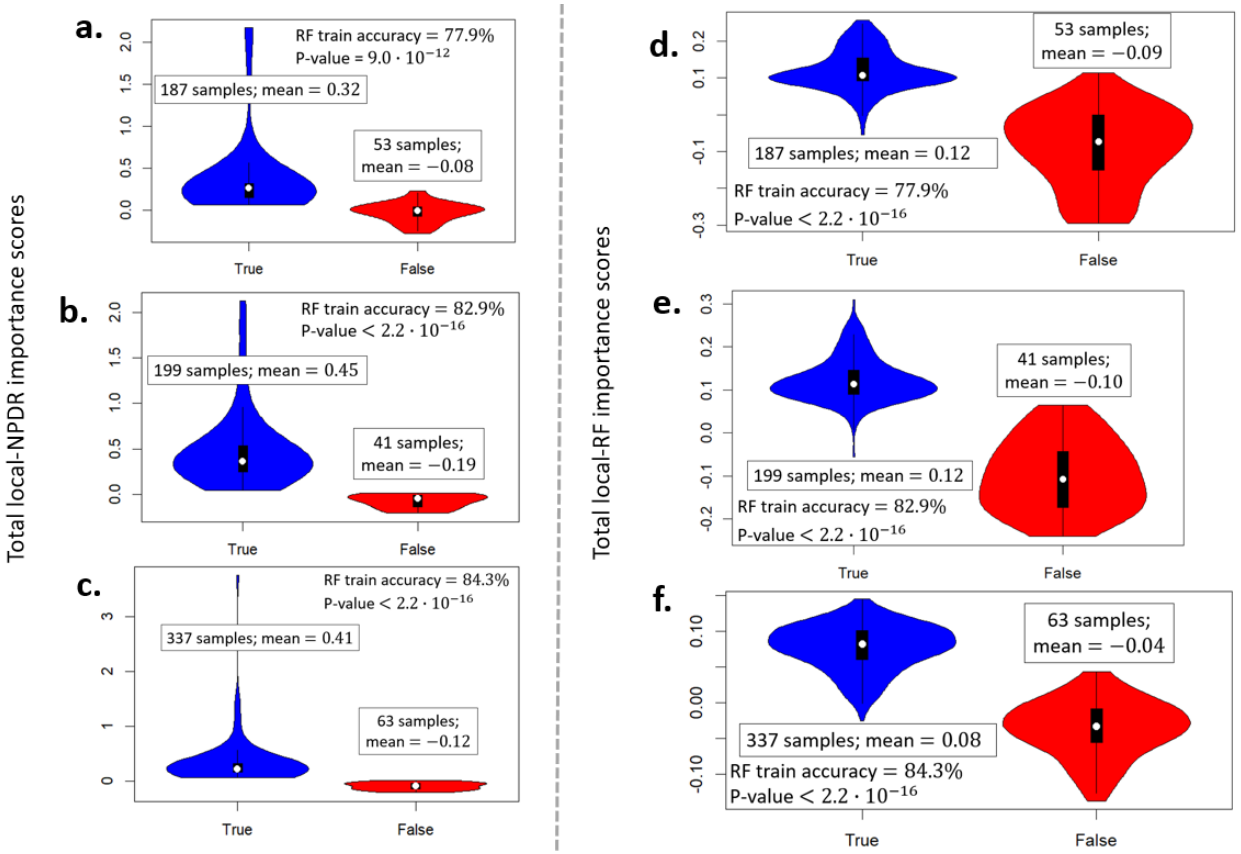


Figure B.3. Total local-NPDR and total local-RF variable importance scores for true and false predictions for training samples in the three simulated datasets. In each dataset and for both methods, local-NPDR and local-RF, the scores are higher in the true (blue) versus false (red) prediction samples (all t-tests are statistically significant). **(a)** This class-imbalanced simulated dataset with RF training accuracy of 77.9% produces 187 samples with true classifications and 53 samples with false classifications in the training data. The mean local-NPDR variable importance scores are higher in the true prediction group ($P=1.6 \cdot 10^{-6}$). **(b)** This simulated dataset with class imbalance yields RF training accuracy of 82.9% with 199 true predictions and 41 false predictions. The mean total local-NPDR scores for true predictions are higher than false prediction scores ($P<2.2 \cdot 10^{-16}$). **(c)** The simulated dataset with balanced classes and increased sample size yields an RF training accuracy of 84.3% with 337 true predictions and 63 false predictions in the training data. Local-NPDR importance scores are higher for true predictions in this balanced dataset ($P<2.2 \cdot 10^{-16}$). **(d)** For the same imbalanced simulated training dataset depicted in (a), total local-RF variable importance scores for true predictions are higher than for false predictions ($P<2.2 \cdot 10^{-16}$). **(e)** Local-RF variable importance scores for true predictions are higher than false predictions in the simulated dataset shown in (b) ($P<2.2 \cdot 10^{-16}$). **(f)** The class-balanced simulated dataset in (c) has higher local-RF scores for true classifications ($P<2.2 \cdot 10^{-16}$).

Local-RF scores for the simulated data training samples are also similar to the test samples (Fig. B.4, compare with Fig. 14 in Sec. 3.4.3). However, instead of the scores for FP predictions being higher, local-RF produces scores for FN predictions that are higher than FP, the opposite of the trend in local-NPDR for training samples in the two imbalanced simulated datasets (Fig. B.4a and b). Like local-NPDR scores for the training samples in the class-balanced simulated data, local-RF scores are more similarly distributed between FN/FP predictions (Fig. B.5c).

Analyzing the distribution of local-NPDR and local-RF scores for all training samples in the three simulated datasets can help determine reasonable thresholds for procedures like false prediction diagnosis (see Sec. 3.4.4) in addition to being generally informative for understanding particular datasets (Fig. B.5). For example, the magnitude and distribution of local-NDPR variable importance scores can vary, since the scores are contrastive-loss optimized regression coefficients for pairs of samples (compare scores in Fig. B.5 a, b, and c with Fig. B6a). In contrast, local-RF importance scores are a difference in classification accuracy after and before

variable permutation, so the expected magnitudes for these scores may vary less than the local-NPDR scores (compare Fig. B.5d, e, and f with Fig. B.6b).

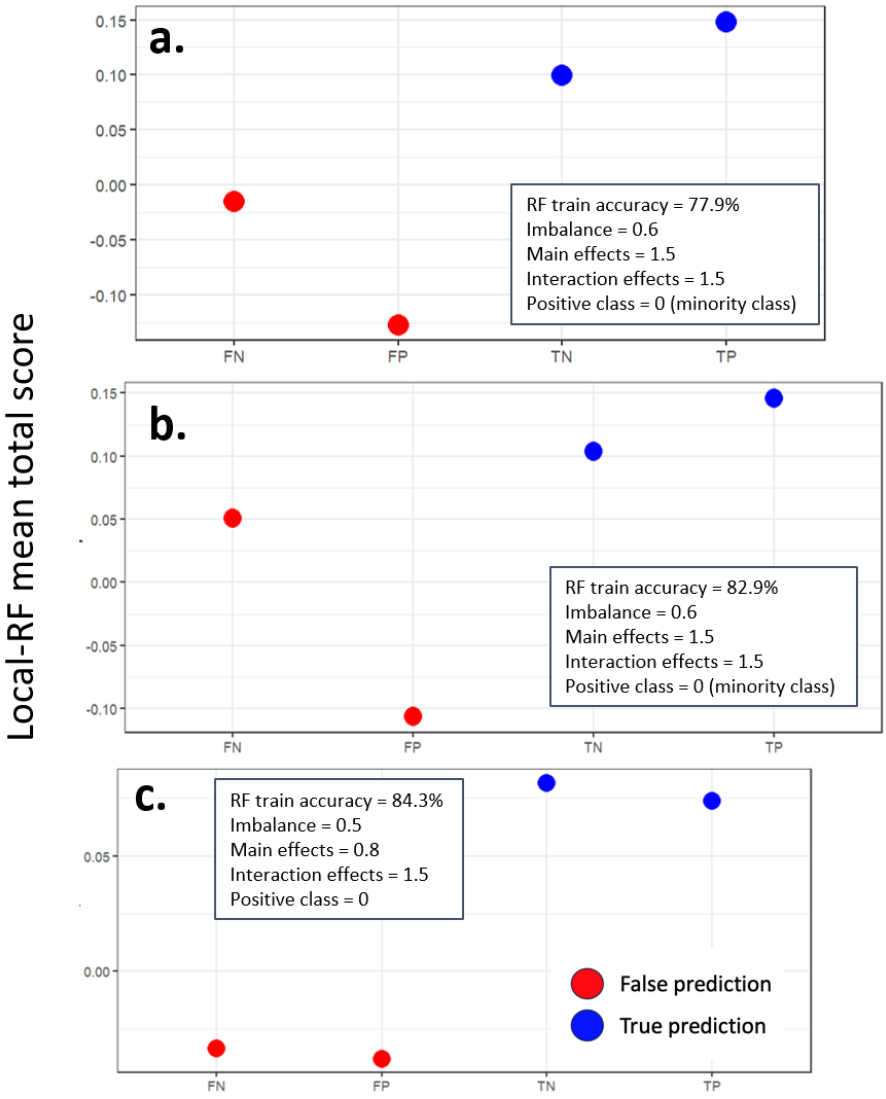


Figure B.4. Local-RF mean total importance scores for each prediction type for training samples in the three simulated datasets. **(a)** For this class-imbalanced dataset, training samples have FN local-RF scores that are higher than the FP scores. The mean scores for the true predictions are higher than either of the false. **(b)** This simulated dataset shows a similar distribution of local-RF mean TLS to (a). **(c)** The class-balanced simulated training samples show local-RF scores for FN and FP predictions that are similar to each other and lower than the TN and TP scores.

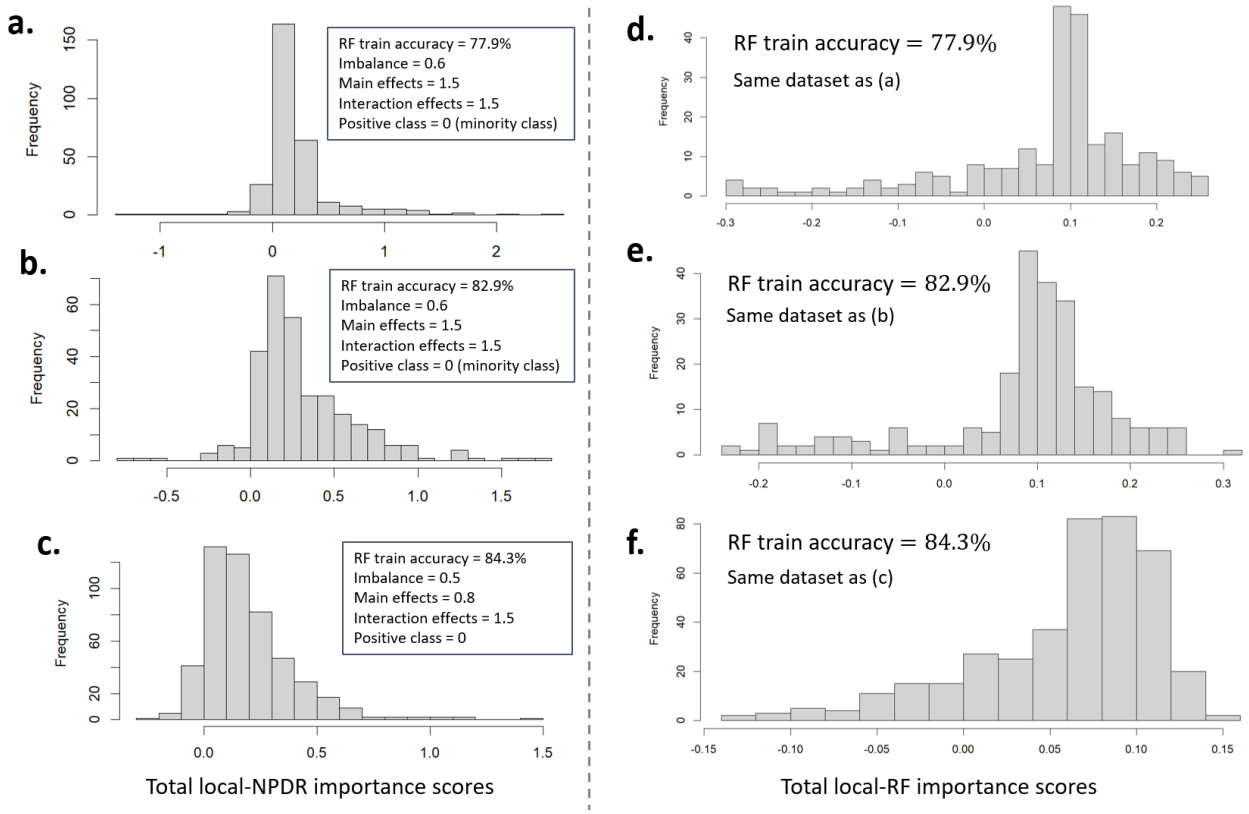


Figure B.5. Distributions of local importance scores for the three simulated datasets. **(a)** Local-NPDR TLS for training samples in this class-imbalanced dataset shows most samples have scores slightly greater than zero, and the positive distribution extends to around 3. **(b)** The local-NPDR TLS for training samples in other class-imbalanced simulated data are more spread-out and less clustered near zero than in (a). The scores in this dataset have smaller maximum and minimum values than (a). **(c)** The local-NPDR scores in the class-balanced simulated dataset have similar maximum and minimum values as (b), but there are more negatively-scored samples and fewer samples with scores > 1. **(d)** The same training samples shown in (a) have local-RF TLS centered around 0.1. **(e)** The simulated training samples depicted in (b) also have local-RF scores centered near 0.1. **(f)** The samples in the balanced simulated dataset shown in (c) also have local-RF TLS around 0.1.

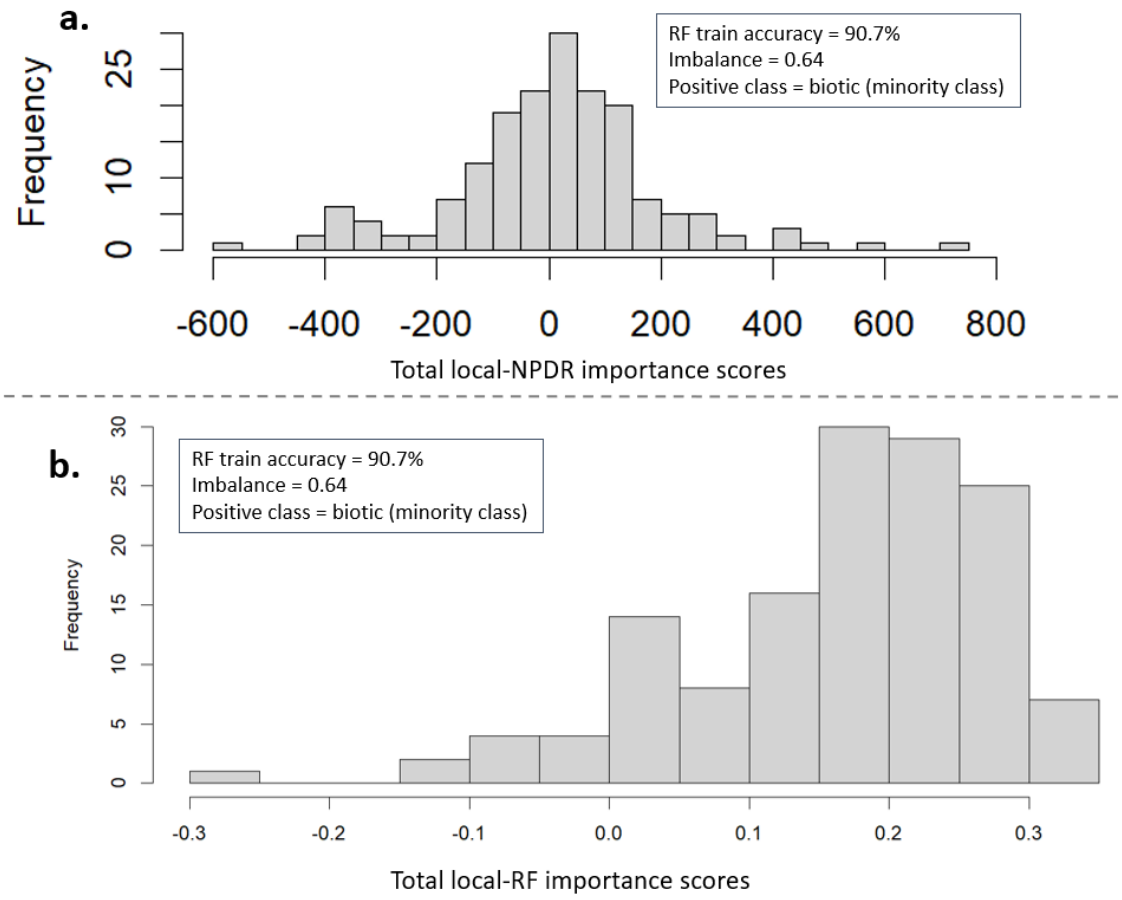


Figure B.6. Distribution of local-NPDR and local-RF importance scores for the biosignature training data. **(a)** Total local-NPDR range from -600 to over 700 scores for the training samples in the biosignature dataset, with the most frequent scores occurring between 0 and 100. **(b)** The total local-RF scores for the same biosignature data training samples are centered around 0.2.