frontiers | Frontiers in Astronomy and Space Sciences

TYPE Original Research
PUBLISHED 24 September 2025
DOI 10.3389/fspas.2025.1651953

OPEN ACCESS

EDITED BY
Josep M. Trigo-Rodríguez,
Spanish National Research Council (CSIC), Spain

REVIEWED BY
Pierfrancesco Novielli,
University of Bari Aldo Moro, Italy
Floyd Nichols,
Virginia Tech, United States

*CORRESPONDENCE
Brett A. McKinney,
brett.mckinney@gmail.com

RECEIVED 22 June 2025
ACCEPTED 13 August 2025
PUBLISHED 24 September 2025

CITATION
Clough LA, Major JD, Seyler LM, Da Poian V, Theiling BP and McKinney BA (2025) Local-NPDR: a novel variable importance method for explainable machine learning and false discovery diagnosis for ocean worlds biosignatures.
Front. Astron. Space Sci. 12:1651953.
doi: 10.3389/fspas.2025.1651953

Local-NPDR: a novel variable importance method for explainable machine learning and false discovery diagnosis for ocean worlds biosignatures

Lily A. Clough[1,2,3], Jonathan D. Major[4], Lauren M. Seyler[5], Victoria Da Poian[3,6,7], Bethany P. Theiling[3] and Brett A. McKinney[1]*

[1]Tandy School of Computer Science, The University of Tulsa, Tulsa, OK, United States, [2]Aurora Engineering, Reston, VA, United States, [3]Planetary Environments Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, United States, [4]School of Geosciences, University of South Florida, Tampa, FL, United States, [5]School of Natural Sciences and Mathematics, Stockton University, Galloway, NJ, United States, [6]Earth and Planetary Science, Johns Hopkins University, Baltimore, MD, United States, [7]Tyto Athene LLC, Reston, VA, United States

Explainable machine learning (ML) is important for biosignature prediction on future astrobiology missions to minimize the risk of false positives due to geochemical biotic mimicry and false negatives due to environmental factors that mask biosignatures. ML models often use feature importance scores to provide insights into model prediction mechanisms by quantifying each variable's contribution to the prediction. Global variable importance methods aggregate information 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. The ability of local-NPDR to differentiate between true and false predictions is compared with other common local importance methods. Local-NPDR is able to diagnose individual false predictions using the concordance between global and local scores, and it can explain mechanisms of true and false predictions. These features allow local-NPDR to integrate scientific explanations of single-sample ML predictions to support a more comprehensive framework for biosignature detection.

KEYWORDS

explainable machine learning, biosignature detection, local importance scores, ocean worlds, biotic mimicry

1 Introduction

> "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

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 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 can be defined as a highly relevant feature (variable) space (Roscher et al., 2020) through which interpretations about model predictions can be made. To add explainability to ML model predictions, a variable space that is mathematically (and if possible, physically) understandable can be leveraged to connect the variables (and their physical/mathematical meanings) to the particular ML predictions being made and to the scientific problem at hand (*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., 2022; 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.

Although not universally the case, ML tends to suffer from an accuracy-explainability tradeoff (Ali et al., 2023). As data dimensionality (number of features) 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 and relevant predictors (variables). 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 as well: 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 (Breiman, 2001). 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.

Recently, we showed that local (single-sample) RF variable importance has the potential to add to the explainability of ML biosignature model predictions and can help diagnose false predictions (Clough et al., 2025). We used our Nearest-neighbors

Projected Distance Regression (NPDR) global feature selection with an RF classifier for biosignature prediction (Le et al., 2020), and we computed the discordance between global and local scores for single samples to diagnose false predictions. This global-local discordance provided important insights; however, existing local variable (or feature) importance methods face limitations such as needing samples to be in the training data, the lack of a statistical threshold for feature selection, and limited ability to account for statistical interactions. A statistical interaction occurs when the effect of a feature on the outcome variable depends on one or more other features (McKinney et al., 2006). An example of an interaction is conditional correlation between pairs of variables that depends on the outcome variable and may occur without either variable having a main effect. These interactions are likely to be important for uncovering biotic mimicry (Clough et al., 2025), and therefore techniques for evaluating the probability of true/false positive or negative predictions of biosignatures are needed for future astrobiology missions.

In high-dimensional variable spaces expected for data of astrobiological relevance (e.g., mass spectrometry and spectroscopy), RF has low power to detect statistical interactions between variables that may be important for classification because variables are selected in trees preferentially based on main effects (McKinney et al., 2009; Wright et al., 2016). However, our global feature selection method, NPDR, has shown high power to detect both main effects and statistical interactions in the biosignature model, and it uses a regression penalty to yield a reduced dimensionality space of independent features (Clough et al., 2025). In the current study, we present an additional mechanism for evaluating the reliability of biosignature predictions. We extend NPDR to compute local or single-sample importance scores to take advantage of NPDR's ability to detect complex relationships between variables. Our global variable importance ML tool helps satisfy the deeper goal of science to provide explanations by allowing an ML model to be trained in a selected features space that is as relevant as possible to the outcome being analyzed. Our local feature importance ML tool introduced in the current study provides explanatory analysis for a single sample in the context of globally important variables and a given ML prediction. Crucially, this analysis allows for a determination of whether a single sample prediction is likely to be true or false without knowing the actual sample label.

The critical need for explainable ML methods across multiple disciplines has given rise to additional local methods since the advent of local-RF, such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME creates local importance scores for models by fitting a surrogate linear model to synthetic samples in the local neighborhood of a query sample (Ribeiro et al., 2016). SHAP uses game theory concepts to provide model agnostic local scores that contribute to the prediction of a sample, while TreeSHAP provides local feature explanations specifically for tree based models like RF (Lundberg et al., 2020). These core tools have been used in a variety of scientific and medical domains including those related to geosciences such as physical oceanography (Navarra et al., 2025), flood susceptibility (Choubin et al., 2025), and CO_2 changes in the soil (Novielli et al., 2025). We compare local-NPDR and local-RF with these prominent methods.

The remainder of the manuscript is organized as follows. 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 predictions. The local-NPDR method is not specific to a given (ML) classifier, and it is able to model nonlinear decision boundaries and detect statistical interactions between features. 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. We then discuss the necessity of explainability and false prediction assessment for high-stakes predictions such as astrobiology biosignatures.

2 Methods

In this section, we first 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. Next, we describe the procedure for designating a prediction as likely true or false based on the total local scores for concordance (positive) and discordance (negative) for globally important features. Finally, we describe the simulated and real biosignature datasets for validation of the local-NPDR algorithm.

2.1 Local-NPDR: feature importance for a single sample

Consider a pair of samples or neighbors i and j that are distinct rows of an m x p data matrix $\mathbf{X}$ 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,

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

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 Equation 2 is computed using the following logit transformation

$$\hat{d}_{ij}(X) = \frac{1}{1+e^{-\left(\beta_0+\vec{\beta}\cdot\vec{d}_{ij}(X)\right)}} \quad (2)$$

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 (Equation 3) would be

$$\vec{d}_{ij}(X) = \left(\left|X_{i1}-X_{j1}\right|,\left|X_{i2}-X_{j2}\right|,\cdots,\left|X_{ip}-X_{jp}\right|\right). \quad (3)$$

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

$$\beta_i^{local} = \min_{\beta_o,\vec{\beta}} \left(\sum_{j \in N_k(i)} \mathcal{L}_{ij}\left(\beta_o,\vec{\beta}\right) + \lambda\left(\alpha\left\|\vec{\beta}\right\|_1 + (1-\alpha)\left\|\vec{\beta}\right\|_2\right)\right). \quad (4)$$

The penalty is implemented via the R library *glmnet*, which allows for a blend of L_1 and L_2 regularization in a method called the "elastic net" (Tibshirani, 1996), and it can be used for Ridge (L_2, $\alpha = 0$) or LASSO (L_1, $\alpha = 1$) penalized regression. We typically use LASSO (Least Absolute Shrinkage and Selection Operator) for global-NPDR feature selection and tune λ via cross-validation. This reduces the selected feature space and increases variable independence. The nature of the derivative of the absolute value function in LASSO prevents regression coefficients from shrinking further once reaching zero as λ increases, but rather they stay zero. For local-NPDR, we typically employ a Ridge penalty because we have already reduced the feature space using global-NPDR and want to keep the rankings of all 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 for true ML predictions of single samples (blue box sample 1, Figure 1a) and can help identify false positive predictions (red box sample 1, Figure 2a) by comparing the local score for a globally important variable (purple variable A on the vertical axis, Figure 1). The global importance can be determined using global-NPDR. For completeness, the global NPDR scores are computed by minimizing the following penalized cross entropy

$$\vec{\beta}^{global} = \min_{\beta_o,\vec{\beta}} \left(\sum_{i=1}^{m} \sum_{j \in N_k(i)} \mathcal{L}_{ij}\left(\beta_o,\vec{\beta}\right) + \lambda\left(\alpha\left\|\vec{\beta}\right\|_1 + (1-\alpha)\left\|\vec{\beta}\right\|_2\right)\right), \quad (5)$$

which, in contrast to Equation 4, includes the sum over all samples i from 1 to m.

The fact that the purple variable A is globally important for classification can be seen by noticing that its mean for the 'x' class is larger than its mean for the 'o' class (Figure 1). Note that in contrastive feature selection methods such as NPDR, a sample contributes positively to a variable's importance score if the projected distance along that variable axis to its opposite-class nearest neighbor (Δ_M, delta miss) is greater than the projected distance to its same-class nearest neighbor (Δ_H, delta hit). This differential $\Delta_M - \Delta_H$ quantifies how well the variable keeps hits close together and misses farther apart in a neighborhood (McKinney et al., 2013).

First we consider how the globally important variable is affected locally in the local-NPDR contrastive loss (Equation 1) for Sample-1 when the sample is in the correct 'x' class (x_1 in blue box, Figure 1a). Specifically, we estimate the contributions to the contrastive loss (Equation 1) for variable A and Sample-1 using $k = 2$ neighbors. In this case, the two neighbors are Sample-2 (same class as Sample 1 (hit): 'x') and Sample-3 (opposite class of Sample 1 (miss): 'o'). The neighbor-pair loss for the miss L_{12} is low (good fit) because the projected distance d_{12} is small (Figure 1b) leading to a small $\hat{d}_{12}$ miss-probability, and their actual miss state is $\delta_{12} = 0$, which causes the first term to be zero. That is, the non-zero quantity $-\ln\left(1-\hat{d}_{12}(A)\right)$ will be a small positive loss (good fit), and the contribution to the local score from A for Sample-1would be relatively large. The neighbor-pair loss for the hit L_{13} is also low (good fit) because, while d_{13} is large, their actual miss state is $\delta_{13} = 1$, causing the second term to be zero. The remaining non-zero part of the loss $-\delta_{13}\ln\left(\hat{d}_{13}(A)\right)$ will be a small positive quantity, and the contribution of Sample-1 and Sample-3 neighbors to the local score for variable A will be relatively large. This high importance score in the local neighborhood of 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 (Figure 1c). In contrast, if we incorrectly label Sample-1 ('o' instead of 'x' in Figure 2A), the neighbor losses will be high (bad fit, Figure 2b) and the importance of variable A in the local neighborhood of Sample-1 will be low (Figure 2c), 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 *knn_max* causes NPDR and Relief-based methods to become myopic; that is, focused on the importance of single variables (McKinney et al., 2013; Dawkins and 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).

2.2 Random forest variable importance

We compare local-NPDR with other local importance scores, including local-RF, a method native to the RF classifier (Breiman, 2001). 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

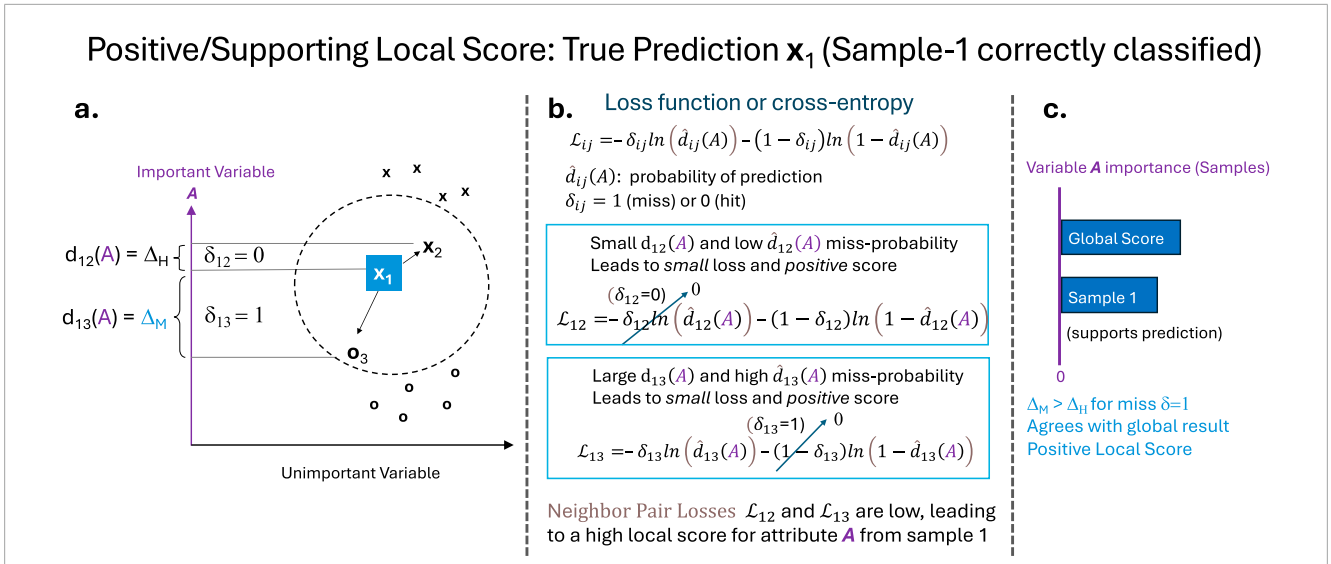

FIGURE 1
Local-NPDR mathematics for a correct classification with a positive (supporting) local feature importance score. **(a)** Consider hypothetical Sample-1 of class *'x'* (blue highlighted $\mathbf{x}_1$) and two features, one simulated with a main effect for classification, *i.e.*, for discriminating 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 dashed 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 also indicated by Δ_H (hit) and Δ_M (miss) because their actual hit/miss statuses are δ_{12} = 0 (hit) and δ_{13} = 1 (miss). Note that 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 Equation 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 small because Sample-1 is correctly classified and the δ's are correctly assigned as δ_{12} = 0 (hit) and δ_{13} = 1 (miss). **(c)** These low losses for the true classification of Sample-1 as class *'x'* lead to positive local scores for important variable *A* in agreement with the global score.

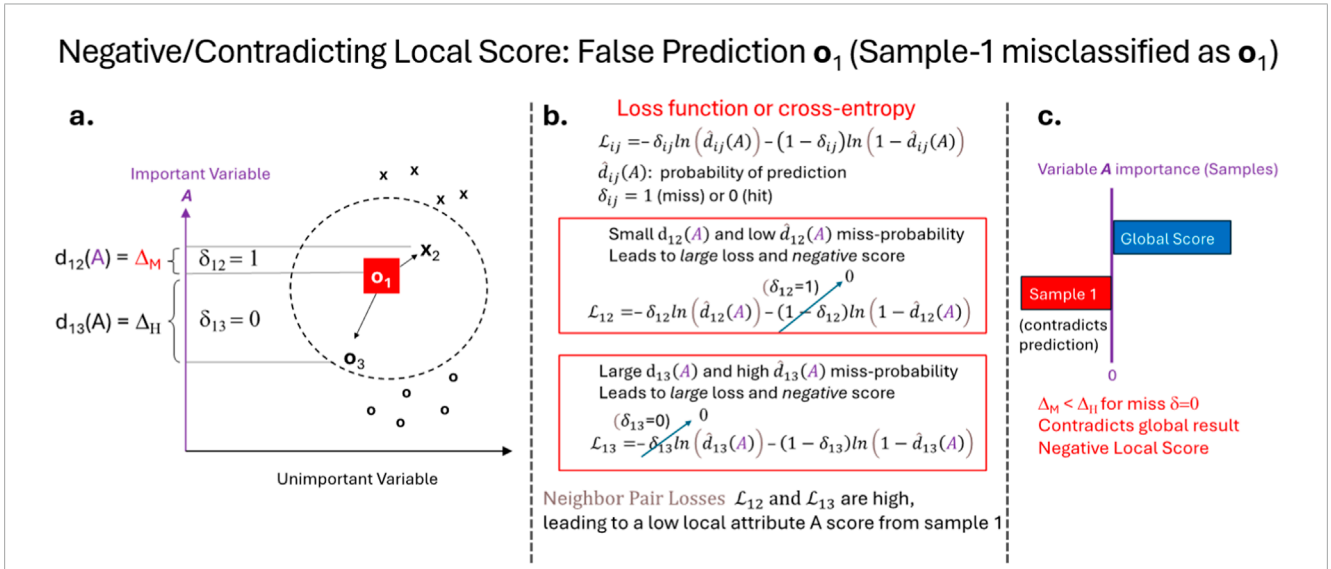

FIGURE 2
Local-NPDR mathematics for an incorrect classification with a negative (contradicting) local feature importance score. **(a)** Same as Figure 1 except hypothetical Sample-1 is now incorrectly assigned class *'o'* (red highlighted $\mathbf{o}_1$). Two features are simulated, one with a main effect 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) and Sample-3 ($\mathbf{o}_3$, same class as 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_M (miss) and Δ_H (hit)because δ_{12} = 1 (miss) and δ_{13} = 0 (hit). **(b)** Local-NPDR loss function for the two nearest neighbors of Sample 1 (see Equation 1). The loss functions for two pairs of neighbors ($\mathcal{L}_{12}$ *and* $\mathcal{L}_{13}$) are large because Sample 1 is incorrectly classified and the δ's are incorrectly assigned as δ_{12} = 1 (miss) and δ_{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*.

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

2.3 Procedure for reporting false predictions

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 (Tibshirani, 1996; Zou and Hastie, 2005).

2.4 Validating local-NPDR: real and simulated datasets

We validate the local-NPDR variable importance method on both real and simulated datasets. 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 and simulated datasets used in this study are summarized in Table 1.

We perform RF classification and global-NPDR feature selection for all datasets using an 80:20 train:test split that preserves the class imbalance. Previously, we found 80:20 splits have very stable test accuracies across repeated 5 folds (Clough et al., 2025). In the current study, we use a single 80:20 split of the data to simplify the interpretation of the results while comparing the local score methods. We choose a split of the real data with a typical (median) test accuracy. RF hyperparameters are tuned using 5-fold cross-validation in the training set. The real astrobiology dataset consists 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 biosignature dataset, referred to as Benchmark Ocean Worlds-δCO_2 dataset (BOW-δCO_2), 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 (summarized at the top of Table 1) using the `createSimulation2` function from our *npdr* R library based on the approach in Ref. (Lareau et al., 2015). 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. Simulations 1 and 2 datasets are designed to have similar properties to the real BOW-δCO_2 dataset. Like our real dataset, these two simulated datasets have a similar number of features (p = 100), sample size (m = 240 train and 60 test samples) and class imbalance (0.6). In addition, the simulated data have 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. The Simulation 3 dataset has the same number of features (p = 100) but a larger sample size (m = 400 training samples and 100 test samples) with balanced classes instead of imbalance. This dataset has the same number of main effects and interactions as Simulations 1 and 2 but has smaller main effect sizes (main effect strength = 0.8, interaction effect strength = 1.5).

10.3389/fspas.2025.1651953

TABLE 1 Summary of the real biosignature data and three simulated datasets (top). For the simulated data, the number and type of functional features are given (main effects, interaction effects or noise features). Random Forest is used for training and testing accuracies for all data (additional accuracy information in Figure 3) using global-NPDR-LURF feature selection (features listed at bottom). 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 noise variables not involved in classification except by chance. Dashes are used for data that have fewer important features selected.

	Biosignature data	Simulation 1	Simulation 2	Simulation 3
class1: class0 (test) samples	89:51 (22:12) abio:bio	144:96 (36:24)	144:96 (36:24)	200:200 (50:50)
main: interact: noise	104 total features	10:10:80 features	10:10:80 features	10:10:80 features
Strength: (main/interact)	—	1.5/1.5	1.5/1.5	0.8/1.5
Train (Test) Accuracy	90.7% (91.2%)	77.9% (71.1%)	82.9% (78.3%)	84.3% (80.0%)
Global NPDR Features				
1	avg_rR$^{45}CO_2$/$^{44}CO_2$	mainvar9	mainvar8	mainvar5
2	sd_$\delta^{18}O/\delta^{13}C$	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

3 Results

3.1 Global-NPDR feature selection and RF classifiers

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 Equation 5 in Section 2.1) and URFP distance (global-NPDR-LURF) for training data splits. The dataset was split into train (89 *abiotic* and 22 *biotic*) and test (51 *abiotic* and 12 *biotic*) sets that reflect the global imbalance (see Table 1 for further details). 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_class*. For the biosignature training data (m = 140 samples, 89 *abiotic* and 51 *biotic*), global-NPDR with hyperparameter λ = 0.01 results in five selected features (bottom left of Table 1) 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* = 5,000 yields a training accuracy of 90.7% and a test accuracy of 91.2% (Figure 3a). 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.

For datasets Simulations 1-3, global-NPDR selected nine, eight, and ten features out of 100 (bottom of Table 1) with hyperparameter λ = {0.02, 0.013, and 0.01} respectively. The simulated datasets include functional features, which begin with "main" and "int" for main effects and interactions, respectively. Simulated features that begin with "var" are noise features and are not functional. The two simulated datasets (Simulations 1 and 2) that contain noise variables are also imbalanced datasets. The Simulation 3 dataset has balanced classes, and global-NPDR found no noise variables. This lower false positive rate for functional variables may be due to class balance or larger sample size. The resulting RF training (test) accuracies for data Simulations 1-3 are 77.9% (71.7%), 82.9% (78.3%), and 84.3% (80.0%), respectively (Figures 3b–d). The dataset with the highest accuracy (Figure 3d) is balanced between classes and has a higher sample size. In addition, main effects play a more prominent role in feature selection (Table 1, bottom last column). For the respective simulated data, the tuned RF hyperparameters were: *mtry* = {5, 8, 2},

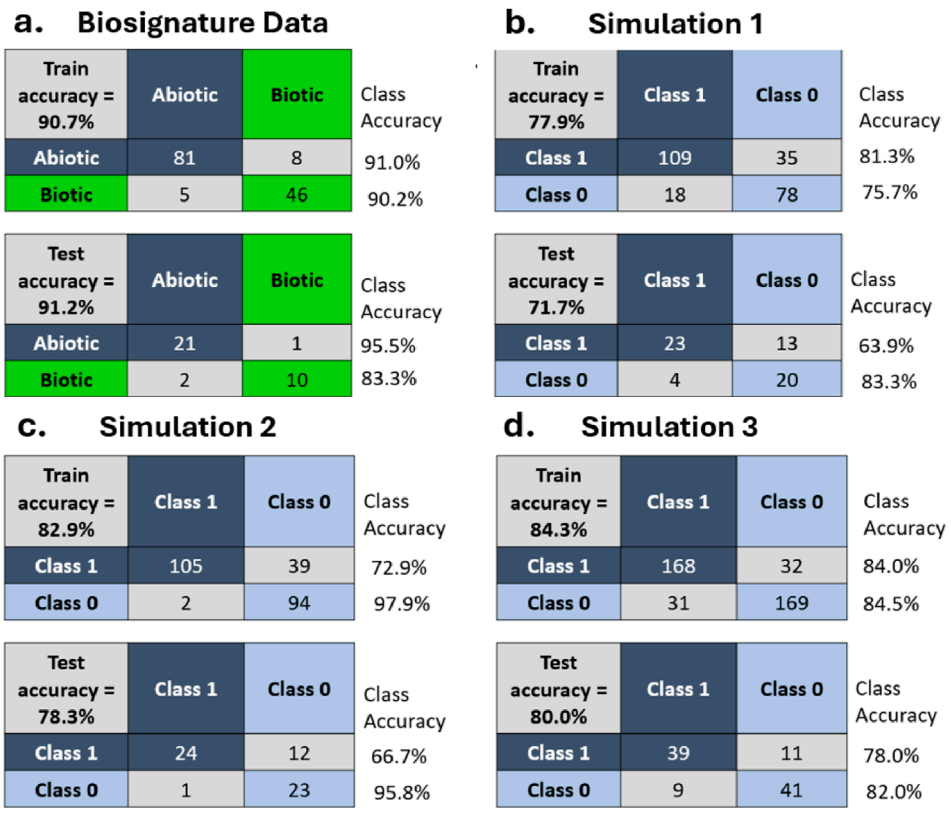

FIGURE 3
Random Forest (RF) train and test accuracies for real biosignature data and simulated datasets. **(a)** The RF classifier for biosignatures yields a 90.7% training accuracy. There are 51 *biotic* samples and 89 *abiotic* samples in the training data; five *biotic* samples are misclassified as *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)** Simulation 1 is an imbalanced simulated dataset (class imbalance = 0.6) that 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)** Simulation 2 data is also imbalanced (class imbalance = 0.6) with the same The two imbalanced simulated datasets show a discrepancyclass breakdown as Simulation 1 and shows similar behavior in terms of class accuracy imbalance in the test data but is even more pronounced. **(d)** Simulation 3 data is balanced and has a higher sample size (training data contains 200 class-*0* and class-*1* samples each). The class accuracies are balanced in both train and test data.

splitrule = {"gini", "extratrees", "extratrees"}, *min. node.size* = {12, 3, 7}, and *ntrees* = {5,000, 6,000, 6,000}. The two imbalanced simulated datasets show a discrepancy in class accuracy that is most notable in the test data (Figures 3b,c), while the balanced simulated dataset shows a more balanced RF class prediction accuracy in both the train and test data (Figure 3d).

3.2 Local-NPDR feature importance for true and false ML predictions

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.

3.2.1 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 for the set of global-NPDR-LURF features. Results from the test data are discussed here; see Supplementary Section S3 for training data results. For each sample, the total local-NPDR variable importance scores are computed (for the globally important features), and we use a t-test to compare the total local scores (TLS) between samples with a true and false prediction by the RF model. For the three simulated datasets Table 1, the total local-NPDR scores are higher in the true versus false prediction groups for both training (Supplementary Figure S3) and test samples (Figure 4). The elevated TLS in true versus false groups in the test data is statistically significant (with P-values: $4.6 \cdot 10^{-6}$ to $7.9 \cdot 10^{-12}$).

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,

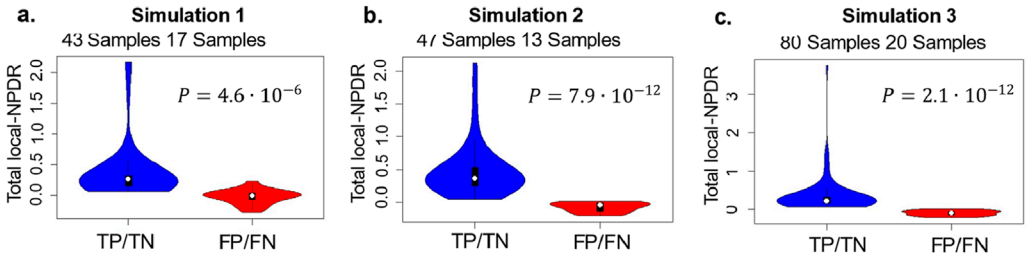

FIGURE 4
Total local-NPDR variable importance scores for true and false predictions in three simulated test (holdout) datasets. Total Local Scores (TLS) are computed for the globally important features based on LASSO NDPR. In each dataset, the local-NPDR scores are higher in the true (blue) versus false (red) prediction samples with very low overlap (all t-tests statistically significant). Detailed properties of Simulation 1-3 are given in Table 1. **(a)** Simulation 1 is class-imbalanced and the mean total local-NPDR feature importance scores are higher in the true prediction group ($P = 4.6 \cdot 10^{-6}$). **(b)** Simulation 2 data is also imbalance and the mean local-NPDR variable importance scores is higher in the true prediction group ($P = 7.9 \cdot 10^{-12}$). **(c)** Simulation 3 has a larger sample size, and the classes are balanced. Local-NPDR importance scores are also higher for true predictions ($P = 2.1 \cdot 10^{-12}$).

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 (Supplementary Figure S4) and test data (Figures 5a,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 Figures 5a,b with Figure 5c), 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 (Figures 5d,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 (Figure 5c). 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% (Figure 5e). 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. However, because of the potential for TLS overlap between individual true and false predictions, especially in imbalanced datasets, it can be beneficial to incorporate other information for diagnosing sample predictions, such as the RF classifier probability.

3.2.2 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 Section 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 (Figure 6a). 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 (Figure 6b). 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.

3.2.3 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 (Supplementary Figure S3) and test samples (Figure 7). 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

10.3389/fspas.2025.1651953

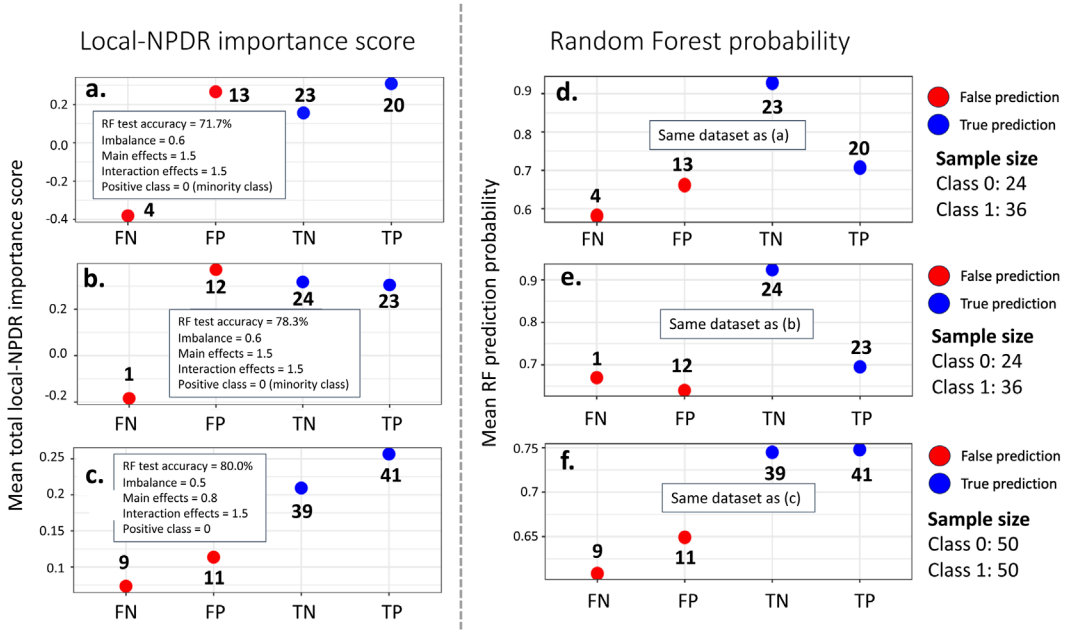

FIGURE 5
Mean total local-NPDR variable importance scores for three simulated datasets [left panel, **(a–c)**] and mean RF prediction probability for the same three simulated datasets [right panel, **(d–f)**]. Values are broken down by prediction type on the x-axes (FN = false negative, FP = false positive, 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 Figures 3b–d). **(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 true positives (class-*0* samples correctly 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.

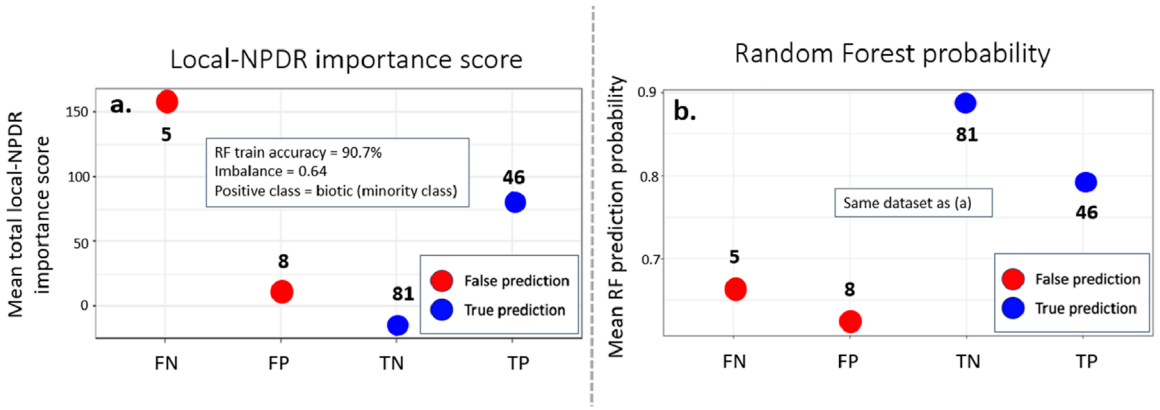

FIGURE 6
Mean total local-NPDR variable importance and mean RF prediction probability for the four different prediction types in the biosignature training data. The number of samples representing each type of prediction are indicated next to the points. **(a)** Local-NPDR mean total local importance 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 Figures 4a,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.

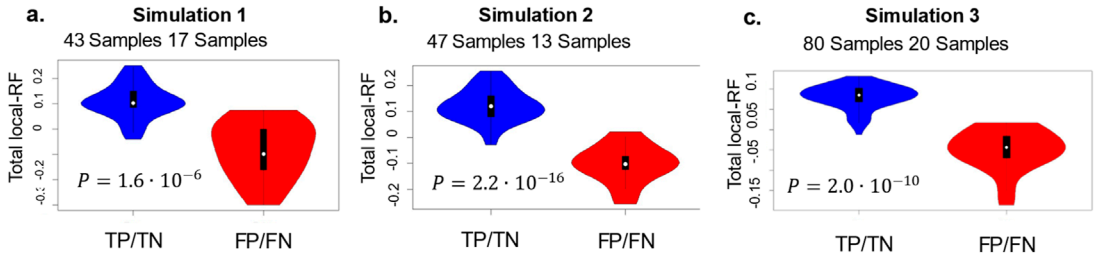

FIGURE 7
Total local-RF variable importance scores for true and false predictions in three simulated test (holdout) datasets. Total Local Scores (TLS) are computed for the globally important features based on LASSO NDPR. In each dataset, the local-RF scores are higher in the true (blue) versus false (red) prediction samples (all t-tests statistically significant). Detailed properties of Simulation 1-3 are given in Table 1. **(a)** Simulation 1 is class-imbalanced has and the mean total local-RF importance scores that are higher in the true prediction group ($P = 1.6 \cdot 10^{-6}$). **(b)** Simulation 3 is also class imbalanced and the mean local-RF variable importance scores are higher in the true prediction group ($P < 2.2 \cdot 10^{-16}$). **(c)** Local-RF importance scores are in the balanced dataset Simulation 3 are higher for true predictions ($P = 2.0 \cdot 10^{-10}$).

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

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 (Figure 8; Supplementary Figure S4). 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 Figures 5, 8). 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 Section 1 are discussed more in Section 4.

Both local-NPDR and local-RF result in statistically significant differences in mean-TLS between true and false predictions in the simulated datasets (Figures 4, 7). For the biosignature data, local-NPDR variable importance indicates less clear separation between true and false prediction scores (compare Figures 6a, 8d), indicating the sensitivity of the multivariate regression to both imbalance and small sample sizes, discussed in more detail in Section 4. 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.

We perform LIME and treeSHAP explanation on the three simulations (Table 1) using the same tuned RF model restricted to LURF features that was used for local-NPDR and local-RF. We train the LIME and treeSHAP explainers on this data and test TLS scores on the holdout data for differences between true and false RF predictions (Figure 9). The total LIME and tree SHAP total scores are higher in the true prediction aggregated group (TP and TN combined) than the false prediction groups (FP and FN combined). However, the differences are less significant than local-NPDR (Figure 4) and local-RF (Figure 7), with some of the LIME (Figure 9b) and treeSHAP (Figure 9f) true/false differences not being statistically significant. The treeSHAP TLS distributions in the true prediction groups are consistently bimodal. This is due to the TN scores being positive and TP scores being negative. The TP scores are consistently in the opposite of the desired direction.

3.3 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 Supplementary Section S4, 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 (Figure 10). 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 (Figure 10a). 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 (Figure 10b). 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 (Figures 10c,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 (Figure 10c). For an *abiotic* prediction with

10.3389/fspas.2025.1651953

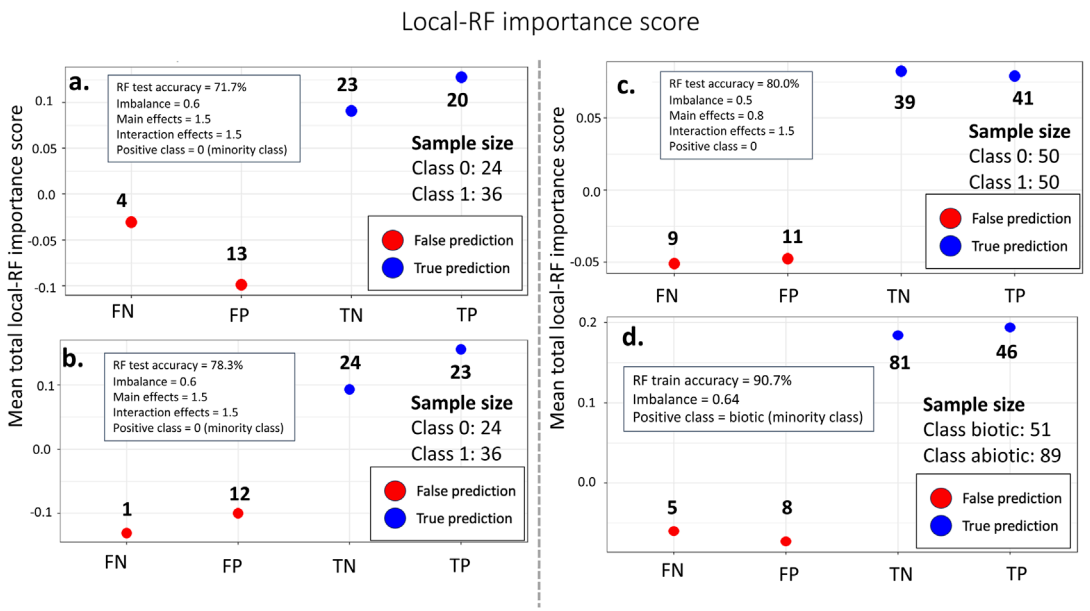

FIGURE 8
Mean total local score (TLS) for local-RF variable importance for four datasets. The number of samples representing each type of prediction are indicated next to the points. **(a)** Local-RF mean TLS for each prediction type for test samples in the simulated dataset with 71.7% RF test accuracy. 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)**.

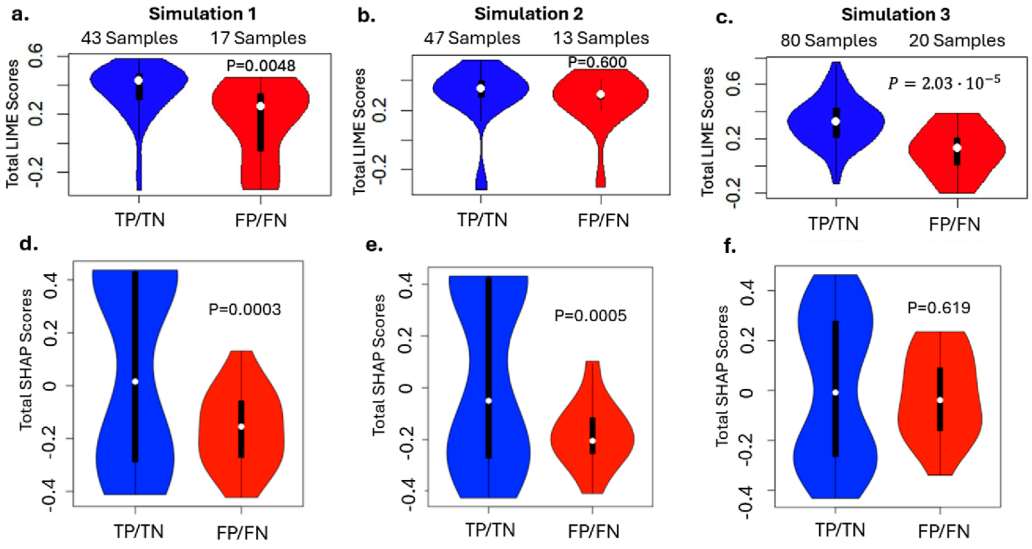

FIGURE 9
Total local variable importance scores for true and false predictions in three simulated test (holdout) datasets for LIME **(a–c)** and treeSHAP **(d–f)**. Total Local Scores (TLS) are computed for the globally important features based on LASSO NDPR. The mean TLS are higher in the true (blue) versus false (red) prediction samples, but the P-values and distribution overlap are higher between true and false groups than local-NPDR and local-RF. Detailed properties of Simulation 1-3 are given in Table 1.

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 (Figure 10d).

An analogous analysis using local-RF as the importance method for the same four test samples uses similar reasoning (Figure 11). The magnitude of local-RF and local-NPDR variable importance scores

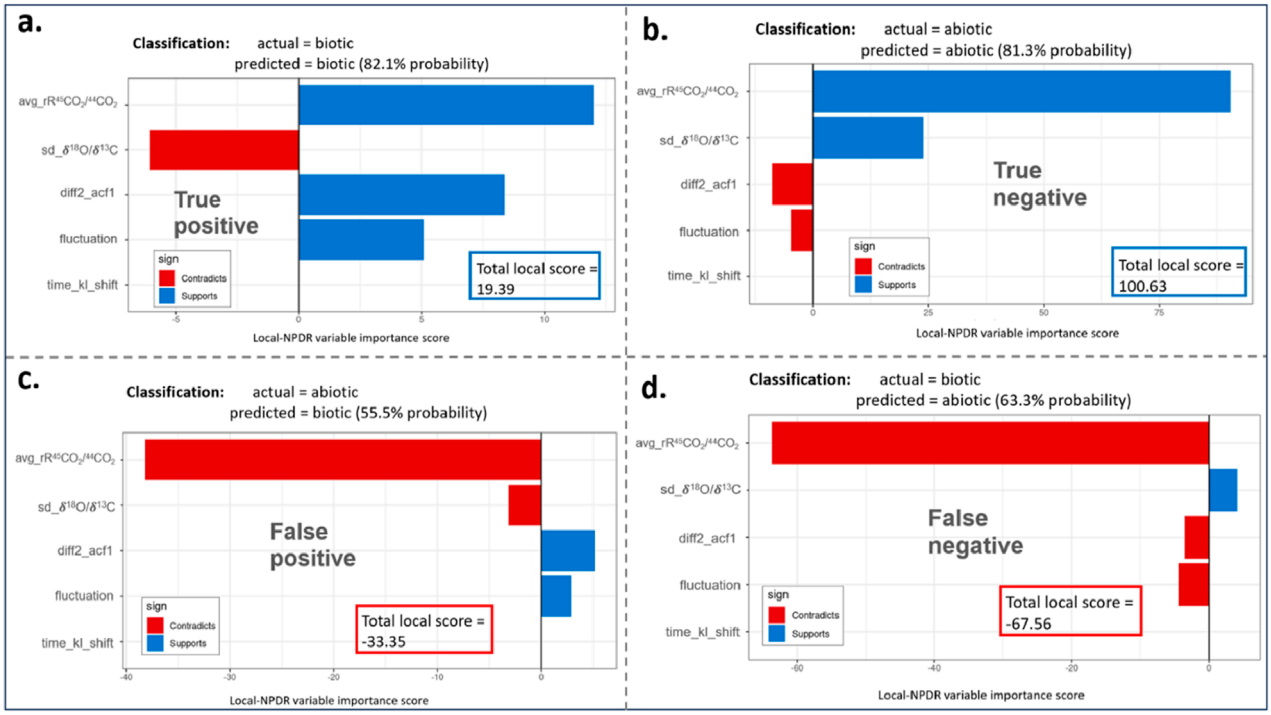

FIGURE 10
Typical distributions of local-NPDR variable importance scores for true and false predictions. Variables shown are listed according to global-NPDR importance ranking, with the most important feature for classification at the top. "Contradicts" (red) means the local-NPDR score of a variable 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 82.1%, 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 the sample is likely incorrectly classified.

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 (Figure 11a) and true negative sample predictions (Figure 11b). The TLS is negative for the false positive (Figure 11c) and false negative (Figure 11d) 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 Figures 10, 11).

To illustrate an applied comparative analysis, consider the case where we have a set of test samples with unknown actual classes and trained RF classifier and probability models. Suppose models are trained with global-NPDR selected features; each test sample is run through the classifier and probability RF models and is labeled with a prediction and probability. The goal of the local variable importance analysis is to "quarantine" the test samples that could be falsely predicted from the samples that we are most confident are correctly classified. To do this, we consider the class and probability output of the RF models in the context of the TLSs discussed above (Figures 10, 11) and define an RF probability and an NPDR/RF TLS for which we accept samples. A challenge is to define an appropriate RF probability and TLS threshold to flag samples. A detailed discussion of potential strategies is deferred until the Discussion; 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. For this analysis, we use an RF prediction probability threshold of 75% for all datasets and both local importance methods.

We use local-NPDR importance score thresholds of {0.25, 0.35, 0.35} for Simulations 1-3 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

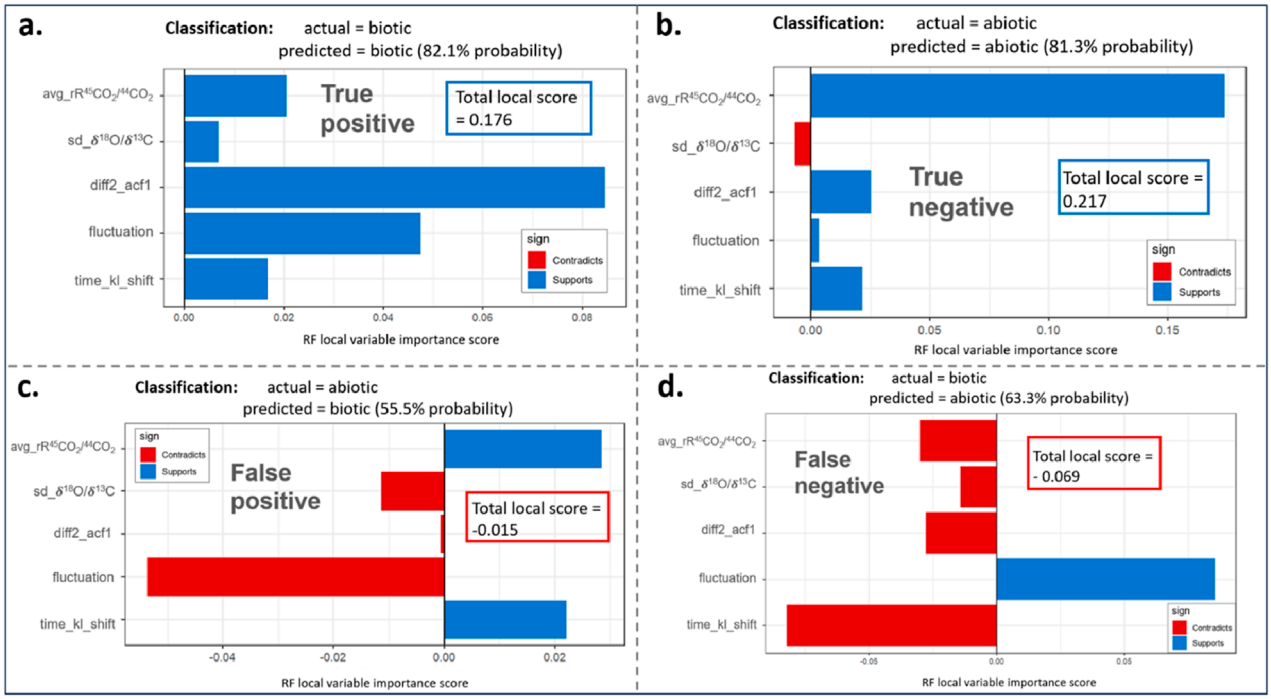

FIGURE 11
Local-RF variable importance scores for the cases analyzed by local-NPDR (Figure 10). The variables are listed according to global-NPDR importance scores. "Contradicts" (red bars) means the local-RF score is negative (unimportant for classification) for a variable that is globally important, 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.

(see Supplementary Figure S7). For the biosignature data, we use a local-NPDR score threshold of 25 (see Supplementary Figure S8). 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 2). 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 2). 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 quarantines 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 Section 4. 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.

4 Discussion

One minimal way of providing explainability for ML tools is to provide a set of variables or features that are important for predicting the outcome of interest (Montavon et al., 2018; Roscher et al., 2020), essentially feature selection (*e.g.*, global-NPDR). If ML model prediction is to provide a deeper level of explainability, analyses beyond global feature selection are needed. Previously, we provided tools for global feature selection, for network visualization of how those features work together to affect model predictions (Lareau et al., 2015; Le et al., 2020), and applied local-RF variable importance to assess the relative likelihood that an individual ML prediction is true or false (Clough et al., 2025).

Here we present a new local feature importance tool, local-NPDR, by extending global-NPDR variable importance to compute importance scores in 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

TABLE 2 False prediction diagnosis rates for local-RF and local-NPDR variable importance methods for four datasets using either a RF prediction probability <75% or a 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 1	71.70%	52.90%	61.5%/25.0%	70.60%	61.5%/100.0%
Simulated 2	78.30%	76.90%	75.0%/100.0%	61.50%	58.3%/100.0%
Simulated 3	80.00%	75.00%	72.7%/77.8%	90.00%	81.8%/100.0%
Biosignature	91.20%	66.70%	100.0%/50.0%	66.70%	100.0%/50.0%

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 (Clough et al., 2025); 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 and 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. 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 (Supplementary Figure S2). The centrality scores of the variables in this interaction network give the cumulative effects of the interactions and main effects of each variable (Supplementary Table S1) (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.

The context of the interaction network and the local-NPDR score distributions for each prediction type can be used to understand which variables are important for each prediction type: FP, FN, TP, FN (Supplementary Figure S3). This means the relative reliability of each feature can be understood and used to make a final decision about the likelihood of a particular sample being correctly or incorrectly predicted, providing another level of explanation with implications for astrobiology missions seeking isotopic biosignatures. For example, the top-ranked global-NPDR feature, $avg\_rR^{45}CO_2/^{44}CO_2$ has a large interaction network centrality, with a moderate main effect while participating in two large-magnitude interactions (see Supplementary Figure S2, node 1). From local-NPDR feature importance analysis on the biosignature training data, this variable is an important indicator for TPs and FPs, but a poor indicator of FNs and TNs (see Supplementary Figure S2). This means for a sample labeled *biotic*, a large positive local-NPDR score for $avg\_rR^{45}CO_2/^{44}CO_2$ indicates the prediction likely represents a true biosignature, and a negative or a low-magnitude score means a likely false biosignature prediction. However, this variable on average has negative scores for TN predictions while being on average positive for FNs. This contradictory behavior indicates that it is important to consider the interactions this variable participates in rather assume a large main effect is driving classification, since it is a good predictor for some biotic and abiotic samples but not all.

If a particular sample is predicted to be *abiotic*, the local-NPDR score for $avg\_rR^{45}CO_2/^{44}CO_2$ should be considered in the context of other local variable importance scores. One of the largest statistical interactions that $avg\_rR^{45}CO_2/^{44}CO_2$ participates in is with *diff2_acf1*, the third-ranked global NPDR feature. This feature is important for diagnosing FNs and TPs, and unimportant for FPs (Supplementary Figure S2), meaning that the sign of *diff2_acf1* can help determine if a sample labeled *abiotic* is likely to be a FN or a TN–if a sample labeled *abiotic* has a negative local-NPDR score for *diff2_acf1* and a large positive score for $avg\_rR^{45}CO_2/^{44}CO_2$, it is likely that the prediction is false, despite the fact that the local score for the top-ranked global-NPDR feature is large and positive, and may dominate the TLS. This understanding may result in the acceptance of more true predictions and the rejection of more false predictions in our applied analysis in Section 3.3 if it were to be encoded in the algorithm to accept or quarantine individual samples. This analysis illustrates the complexity of an extremely small variable space in terms of variable interactions and main effects and how they work together to inform certain prediction types. Understanding and appreciating this nuance can enable increased explanation and scrutiny for individual sample predictions in biosignature classification.

We compared local-NPDR with widely-used local explainers: local-RF, LIME and treeSHAP. Local-NPDR and local-RF showed a statistically significant higher TLS for true versus false prediction samples for all simulated datasets (Figures 4 and 7). LIME and treeSHAP TLS were also higher in the true prediction samples, but there was more overlap between true and false distributions and some of the differences were not statistically significant (Figure 9).

Despite having the largest sample size and being class balanced, the treeSHAP TLS difference was not significant for Simulation 3 (Figure 9f). This could be due to the stronger interaction effects in Simulation 3 compared to the main effects. Local-NPDR had the lowest P-value of its results for Simulation 3 (Figure 4c) while this simulation did not yield the lowest P-value for local-RF (Figure 7c). NPDR has been shown to have good power to detect interaction effects whereas RF has some limitations when detecting interactions. However, feature selection prior to running local-RF helps its sensitivity to interactions. The LIME TLS difference was not significant for Simulation 2 (Figure 9b), which is imbalanced and has a lower sample size than Simulation 3.

The simulated data enabled us to compare the differential effects of class imbalance, sample size, and the relative magnitudes of main/interaction effects for variables, and we can see through this analysis that it is a combination of interaction effects and class imbalance that affects 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.

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, the nature of the model could change 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.

As mentioned in Section 3, the practice of diagnosing false samples using a TLS method requires appropriate thresholds for the RF probability and the TLS to flag samples (see Table 2). While the probability and TLS thresholds used in our comparative analysis are based on properties specific to our training data, both local-NPDR and local-RF can successfully diagnose false predictions using various TLS and probability thresholds. Different TLS and probability thresholds result in different numbers of samples being quarantined, some of which will be true predictions that have a low TLS and/or low prediction probability. Future work will incorporate data-driven statistical thresholds such as two-mode Gaussian mixture modeling, where the two modes represent true and false predictions. 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 minimize samples being flagged in the training data (Supplementary Figures S6–S8). Regardless of the thresholding method, the acceptable risk for false predictions will be user and application dependent. For example, in terms of biosignature detection for 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 potentially as false than local-NPDR, 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 (in our example datasets) 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 performance of the local-RF false prediction diagnosis method. The LASSO version of NPDR provides a feature selection threshold that helps local-RF determine local variable importance and therefore to detect false predictions.

Global-NPDR feature selection also allows the RF classifiers to be computationally more lightweight than if the classifier were required to use the full variable space. In a case where 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 onboard 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, like global-NPDR with LASSO penalty. This approach is sensitive to statistical interactions in high-dimensional datasets, making it a natural choice in complex real datasets such as 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.

5 Conclusion

One of the primary goals of science is to explain *how* and *why*, however black box ML models in scientific applications are antithetical to this goal. We develop a new ML explanation tool, local-NPDR, and test it on three simulated datasets and one real dataset. Two of the simulated datasets are class imbalanced and one has decreased main effects. The real biosignature-analog dataset has a small sample size, is similarly imbalanced, and is known to have lower variable main effects. Our local-NPDR ML tool can be used to help explain why a single sample is 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 prediction (biotic or abiotic) to be true or false. In conjunction with single sample RF prediction probability, local-NPDR can be a useful tool for future astrobiology missions. 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.

Local-NPDR variable importance has the potential advantage of being calculated independently of the classifier. The ability to apply a method that is independent from (agnostic to) the ML classifier to diagnose false predictions can be an advantage as it avoids biases in explainability caused by using the local importance method native to the classifier. If local-NPDR and local-RF were used together, the likelihood of successfully quarantining most if not all falsely predicted samples significantly increases. However, potential limitations of all methods should be considered in high-risk applications. If new samples are very different from samples used to train the models, numeric instability in the local regression may result from distances that are too large between neighboring samples. While this is a potential limitation if not considered, with awareness it can be turned into an indication that the model being used is no longer appropriate, an important conclusion in any high-risk domain.

Being independent of the classifier could also be a potential limitation, since it will not be a perfect explainer for the classifier. In other words, it may be preferrable to have a local explainer that explains the mechanisms of a particular classifier. We limit this by training the RF classifier with global-NPDR selected features, and then 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. Another way to link local-NDPR to a specific classifier could be to use nested cross-validation feature selection (Parvandeh et al., 2020). 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 while local-RF was less susceptible. 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 2).

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. Another potential 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.

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 (quantum mechanics) (Gamow, 1985). 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 ocean world (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 explainability 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. Our ML tools take this step in providing explainability.

Data availability statement

The data and software supporting the analysis in this study are freely available at https://github.com/insilico/OceanWorldsLocalNPDR.

Author contributions

LC: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing. JM: Data curation, Writing – review and editing. LS: Data curation, Writing – review and editing. VD: Writing – review and editing. BT: Data curation, Funding acquisition, Writing – review and editing. BM: Conceptualization, Funding acquisition, Methodology, Software, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors would like to thank Jingyi Chen and the Department of Geosciences at the University of Tulsa for institutional support and Steven Bates in the Stable 1sotope Laboratory at NASA Goddard Space Flight Center (GSFC) for assistance in understanding Isodat© output. Data collection was funded through a NASA Oklahoma Established Program to Stimulate Competitive Research (EPSCoR) Research Initiation Grant (PI: BT). Data analysis and ML efforts were funded primarily through a NASA GSFC Internal Scientist Funding Model (ISFM) Fundamental Laboratory Research (FLaRe) award "Machine Learning of Ocean Worlds Laboratory Analog Seawater Volatiles" (PI: BT) and NASA Oklahoma EPSCoR Infrastructure Development, "Machine Learning Ocean World Biosignature Detection from Mass Spec" (PI: BMcK). NASA Oklahoma Established Program to Stimulate Competitive Research (EPSCoR) Infrastructure Development, "Biosignature Detection of Solar System Ocean Worlds using Science-Guided Machine Learning." (PI: BMcK). Grant 80NSSC24M0109.

Conflict of interest

Author LC was employed by Aurora Engineering. Author VD was employed by Tyto Athene LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2025.1651953/full#supplementary-material

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., et al. (2023). Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. *Inf. Fusion* 99, 101805. doi:10.1016/j.inffus.2023.101805

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/A:1010933404324

Choubin, B., Jaafari, A., Henareh, J., Karimi, O., and Sajedi Hosseini, F. (2025). Explainable artificial intelligence (XAI) for interpreting predictive models and key variables in flood susceptibility. *Results Eng.* 27, 105976. doi:10.1016/j.rineng.2025.105976

Clough, L. A., Da Poian, V., Major, J. D., Seyler, L. M., McKinney, B. A., and Theiling, B. P. (2025). Interpretable machine learning biosignature detection from Ocean worlds analogue CO2 isotopologue data. *Earth Space Sci.* 12 (3), e2024EA003966. doi:10.1029/2024EA003966

Da Poian, V., Theiling, B., Lyness, E., Burtt, D., Azari, A. R., Pasterski, J., et al. (2025). Science autonomy using machine learning for astrobiology. *arXiv*. doi:10.48550/arXiv.2504.00709

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

Dawkins, B. A., and 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 Trans. Comput. Biol. Bioinforma.* 22 (01), 39–51. doi:10.1109/TCBBIO.2024.3494599

Gamow, G. (1985). *Thirty years that shook physics: the story of quantum theory*. Dover Publications, Inc.

Lareau, C. A., White, B. C., Oberg, A. L., and 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 Min.* 8 (1), 5. doi:10.1186/s13040-015-0040-x

Le, T. T., Dawkins, B. A., and 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. doi:10.1093/bioinformatics/btaa024

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable AI: a review of machine learning interpretability methods. *Entropy* 23 (1), 18. doi:10.3390/e23010018

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2 (1), 56–67. doi:10.1038/s42256-019-0138-9

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Appl. Bioinforma.* 5 (2), 77–88. doi:10.2165/00822942-200605020-00002

McKinney, B. A., Jr, J. E. C., Guo, J., and Tian, D. (2009). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLOS Genet.* 5 (3), e1000432. doi:10.1371/journal.pgen.1000432

McKinney, B. A., White, B. C., Grill, D. E., Li, P. W., Kennedy, R. B., Poland, G. A., et al. (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. doi:10.1371/journal.pone.0081527

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

Nagel, E. (1979). *The structure of science: problems in the logic of scientific explanation*. Indianapolis, Indiana: Hackett Publishing Co. Available online at: 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). doi:10.17226/25252

Navarra, G. G., Deutsch, C., Mamalakis, A., Margolskee, A., and MacGilchrist, G. (2025). Long term predictability of Southern Ocean surface nutrients using explainable neural networks. *J. Geophys. Res. Mach. Learn. Comput.* 2 (2), e2024JH000268. doi:10.1029/2024JH000268

Noordijk, B., Garcia Gomez, M. L., ten Tusscher, K. H. W. J., de Ridder, D., van Dijk, A. D. J., and Smith, R. W. (2024). The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology. *Front. Syst. Biol.* 4, 1407994. doi:10.3389/fsysb.2024.1407994

Novielli, P., Magarelli, M., Romano, D., Di Bitonto, P., Stellacci, A. M., Monaco, A., et al. (2025). Leveraging explainable AI to predict soil respiration sensitivity and its drivers for climate change mitigation. *Sci. Rep.* 15 (1), 12527. doi:10.1038/s41598-025-96216-y

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

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? explaining the predictions of any classifier. doi:10.48550/arXiv.1602.04938

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8, 42200–42216. doi:10.1109/ACCESS.2020.2976199

Theiling, B. P., Chou, L., Da Poian, V., Battler, M., Raimalwala, K., Arevalo, R., et al. (2022). Science autonomy for ocean worlds astrobiology: a perspective. *Astrobiology* 22 (8), 901–913. doi:10.1089/ast.2021.0062

Tibshirani, R. (1996). Regression shrinkage and selection *via* the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

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

Zou, H., and Hastie, T. (2005). Regularization and variable selection *via* the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x