Interpretable Machine Learning for Molecular Biosignatures: a Novel Single-Sample Feature Importance Method That Is Sensitive To Statistical Interactions

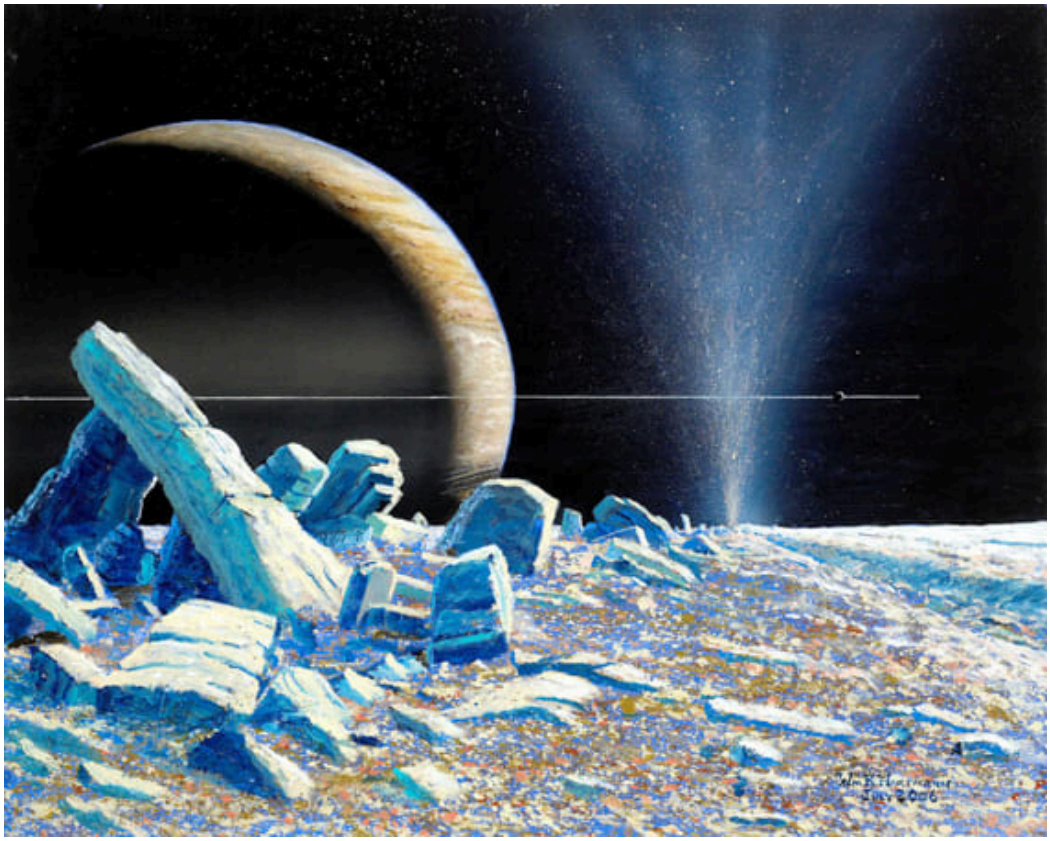
Lily A. Clough, Victoria Da Poian, Bethany P. Theiling, Brett A. McKinney

The University of Tulsa, Aurora Engineering, NASA-Goddard Space Flight Center, Microtel LLC, Johns Hopkins University

PRESENTED AT:

DETECTING ISOTOPIC BIOSIGNATURES FOR OCEAN WORLDS

Ocean Worlds (OWs) like Europa and Enceladus are prime targets for the *in situ* detection of microbial molecular biosignatures via mass spectrometry.

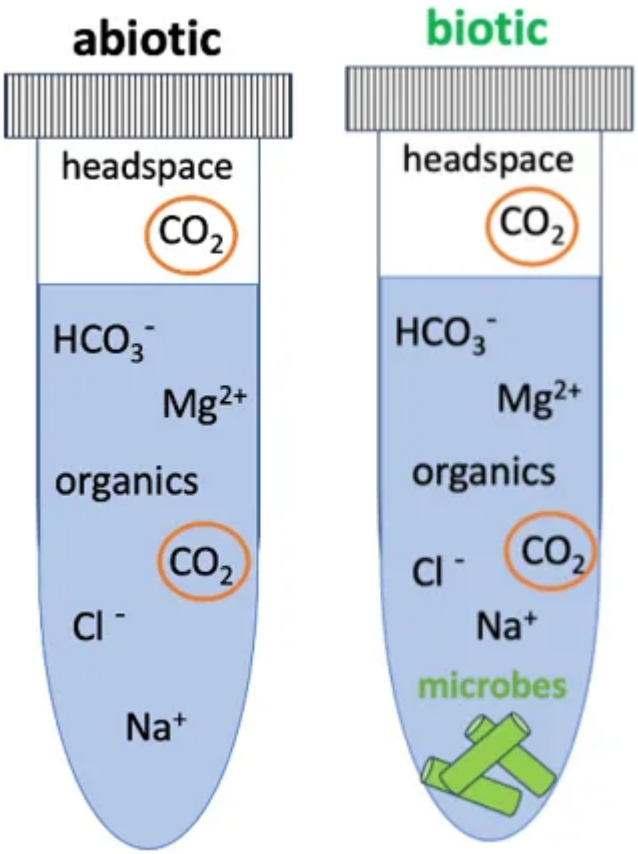


Art: William Hartmann

However, future OW astrobiology missions face challenges such as limited bandwidth and communciation, as well as extreme environmental conditions (*i.e.*, radiation). These challenges can be mitigated by employing onboard **science autonomy** capabilities (Da Poian et al. 2022, 2024; Theiling et al. 2023), such as data prioritization using statistical and **machine learning (ML)** methods.

Planetary mass spectrometry data (Chou et al. 2021) is a good candidate for the application of science autonomy methods for biosignature detection and OW seawater chemistry characterization. In particular, **Isotope Ratio Mass Spectrometry (IRMS)** is promising for the detection for isotopic biosignatures since it is known that biotic processes may produce large isotope fractionations (Park and Epstein 1960, Vogel 1980). However, geochemically complex abiotic processes may mimic biogenic isotope fractionations, prompting the need for well-designed experimental data.

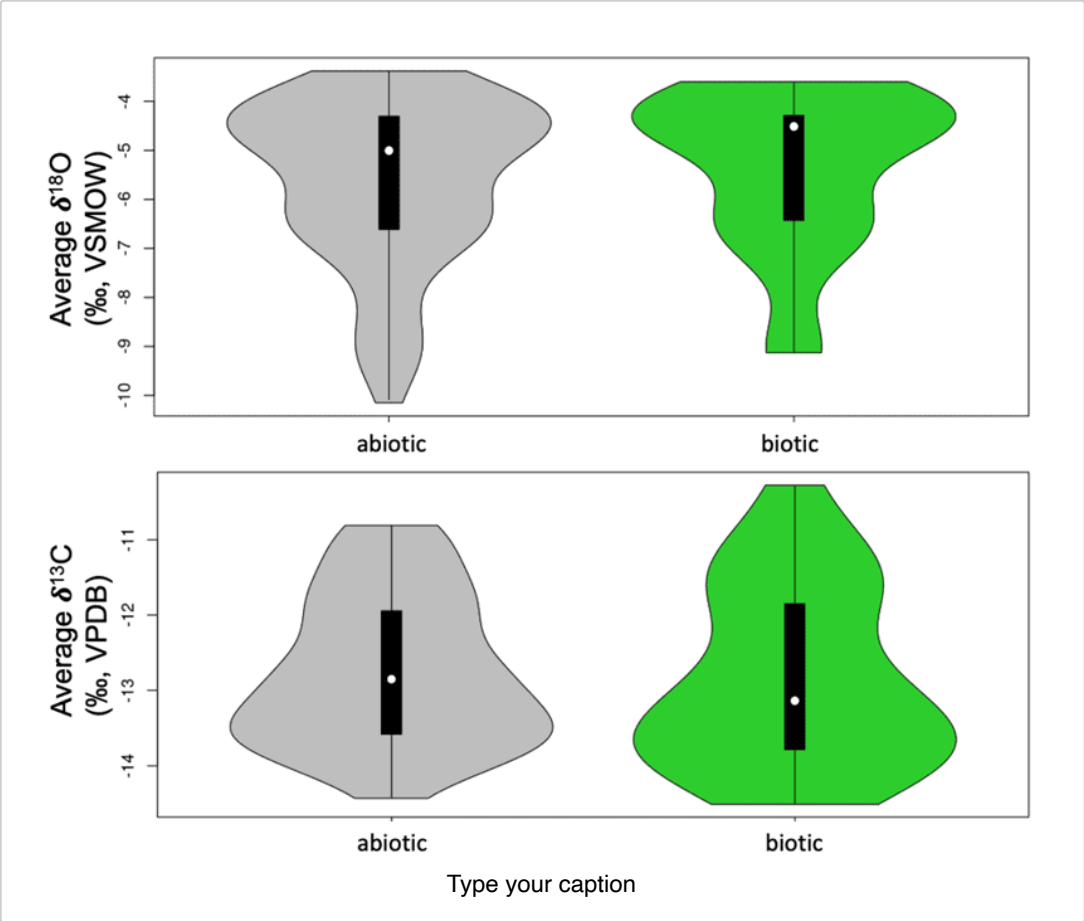
We therefore use laboratory-generated analogue OW brines to collect volatile CO_2 IRMS data. Brines are composed of measured and hypothesized OW seawater salts (Waite et al. 2006, Zolotov and Shock 2001) and span a wide range of pH values and ionic strengths, from 3.5-9.0 and 0.00-15.46 M, respectively. The preparation and geochemistry of these brines is described in Theiling 2021. Some brines are inocculated with microbes (a known sulfate reducer and unknown heterogeneous mixtures). Abiotic brines are geochemically complex and contain non-biogenic organics.



Laboratory-generated OW analogue brines are equilibrated with 0.3% CO_2 and some are inocculated with microbes. These experiments are used to develop isotopic biosignature detection methods.

This IRMS data is processed (see Quality Control and past AGU posters under Future Work, Resources, and More Information) and used to develop methods that can detect isotopic biosignatures from volatile CO_2 isotopologues.

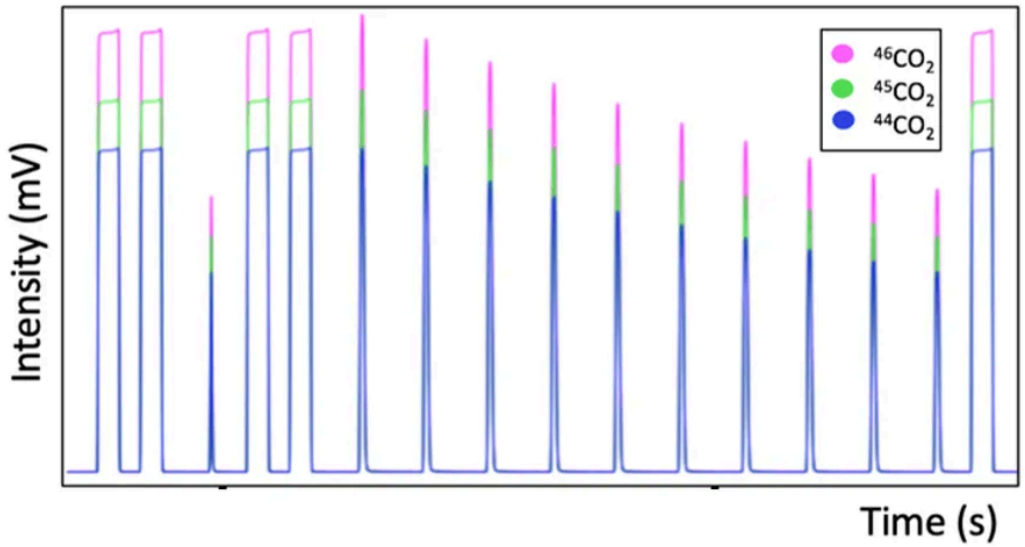
For our dataset, which contains biotic samples and geochemically complex abiotic samples, statistical and ML methods cannot discern between biotic and abiotic samples based on isotope fractionations alone.



Type your caption

More descriptive variables, or **features**, are needed for biosignature prediction. Mathematical feature extraction or construction mines experimental data to find predictors for ML.

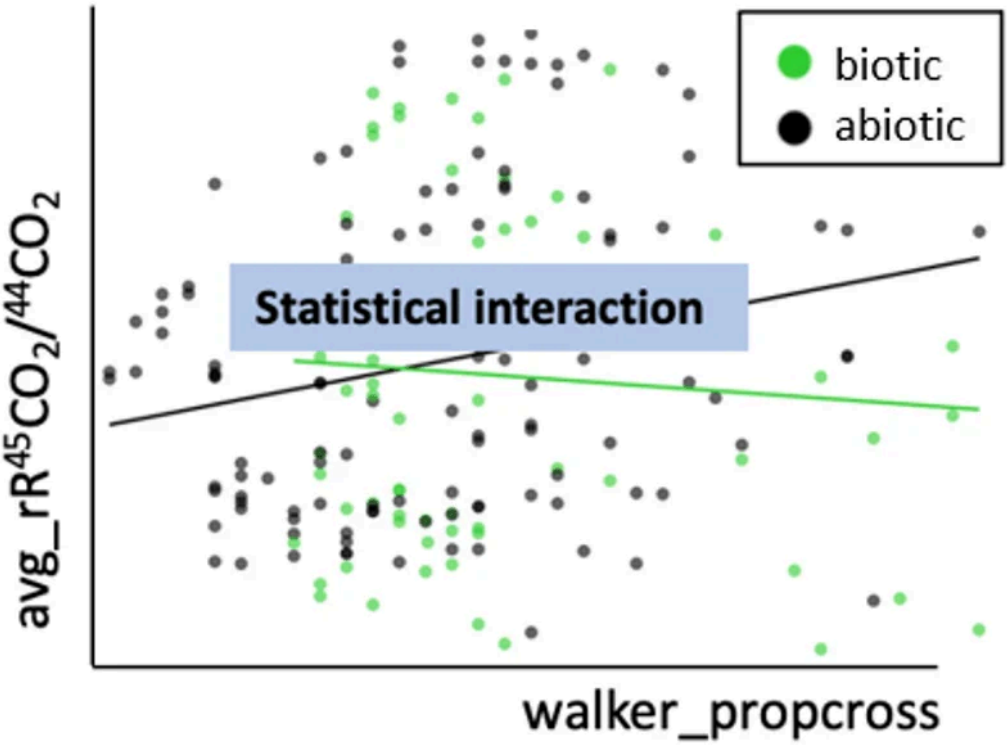
Features in our volatile CO_2 dataset are based on IRMS intrument output (Kopf et al. 2021) and a time-series analysis (Fulcher et al. 2013) of the chromatograms (intensity versus time plots).



A typical volatile CO_2 IRMS chromatogram, with signal intensity on the *y*-axis and time on the *x*-axis. Our experiments are performed over 900 seconds and contain reference gas peaks (rectangular peaks) and sample peaks (pointed peaks). Reference gas is used for calibration and calculation of sample isotope ratios.

THE IMPORTANCE OF STATISTICAL INTERACTIONS FOR BIOSIGNATURE DETECTION

Statistical and machine learning (ML) methods for biosignatures should be able to account for **statistical interactions**, or changes in correlation between variables based on *class*.



A statistical interaction in our IRMS dataset occurs between a feature representing $^{45}CO_2/^{44}CO_2$ detected by the instrument (y-axis) and a time-series feature, walker_propcross (x-axis), that simulates a hypothetical walker moving through the time-series (chromatogram) in response to the amplitude.

Mathematical extraction of features and a small sample size (<200 samples) result in a potentially **high dimensiona**l variable space, where the number of observations (samples) is comparable or less than the number of predictors (variables or features). This can have an adverse effect on statistical and ML methods, and result in poor prediction accuracies and difficult to understand predictions.

Feature selection can be used to identify important features for a model and reduce high-dimensional feature spaces to be more understandable (Guyon and Elisseeff 2003). However, many feature spaces created from real data may contain complex statistical interactions.

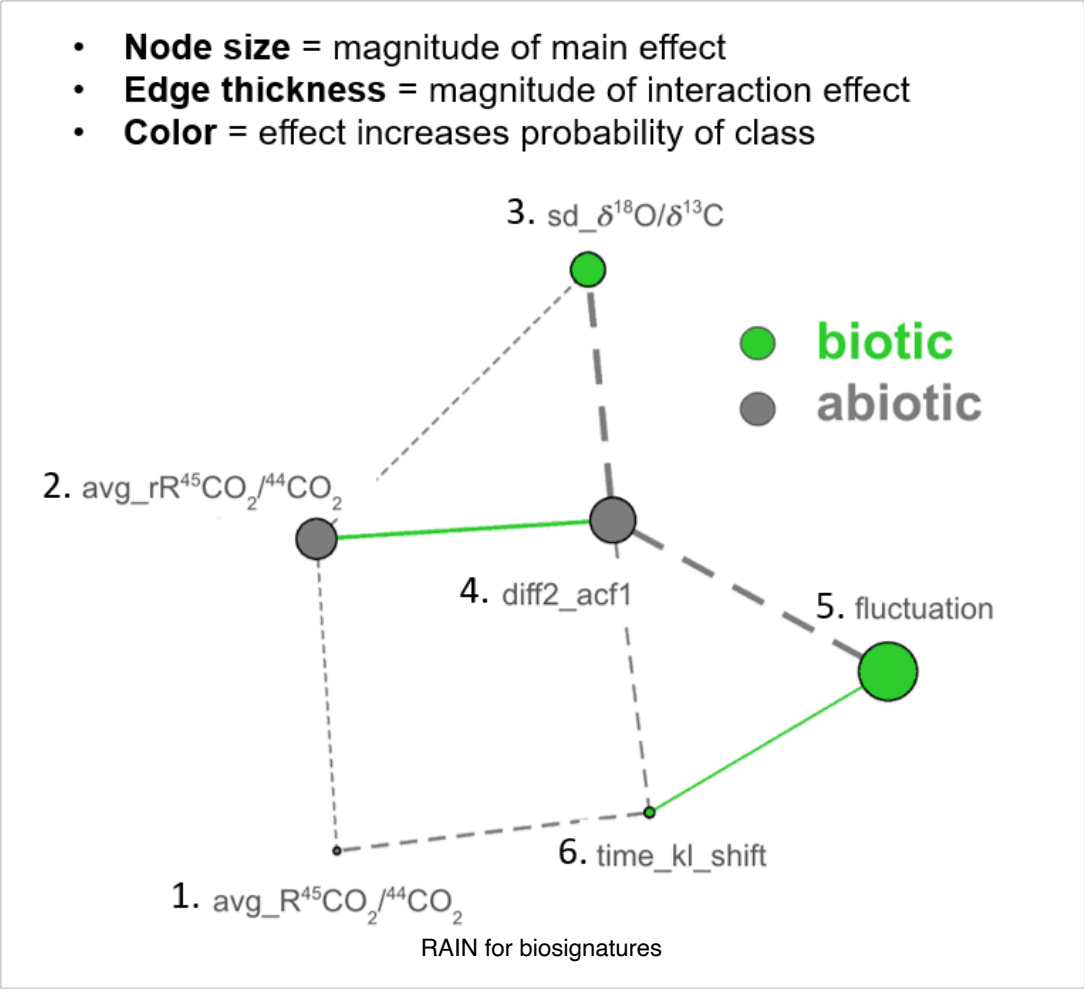
Although statistical interactions can be difficult to detect (Wright et al. 2026, McKinney et al. 2009) they can be informative for prediction, especially when variables do not have a significant **main effect** on prediction, as is the case with our volatile CO_2 IRMS data (see violin plot in the panel above).

Feature selection methods should therefore be able to detect complex statistical interactions between features as well as main effects.

Previously, we developed an ML feature selection method that is sensitive to statistical interactions called **Nearest-neighbors Projected Distance Regression (NPDR),** designed for and demonstrated on bioinformatics data (*e.g.*, gene-expression data) (Le et al. 2020). Recently, we applied this method to OW analogue data to identify important predictors for biosignatures and seawater chemistry in our volatile CO_2 IRMS dataset. We included the use of a novel distance matrix that is non-isotropic for the construction of nearest-neighbors, the unsupervised Random Forest proximity distance (Shi and Horvath 2006).

PREVIOUS WORK: INTERPRETABLE HIGH-ACCURACY MACHINE LEARNING MODELS FOR BIOSIGNATURES AND SEAWATER CHEMISTRY

Regression-based Association-Interaction Interaction Networks (RAINs) visualize variable effects and aid in model prediction interpretation.



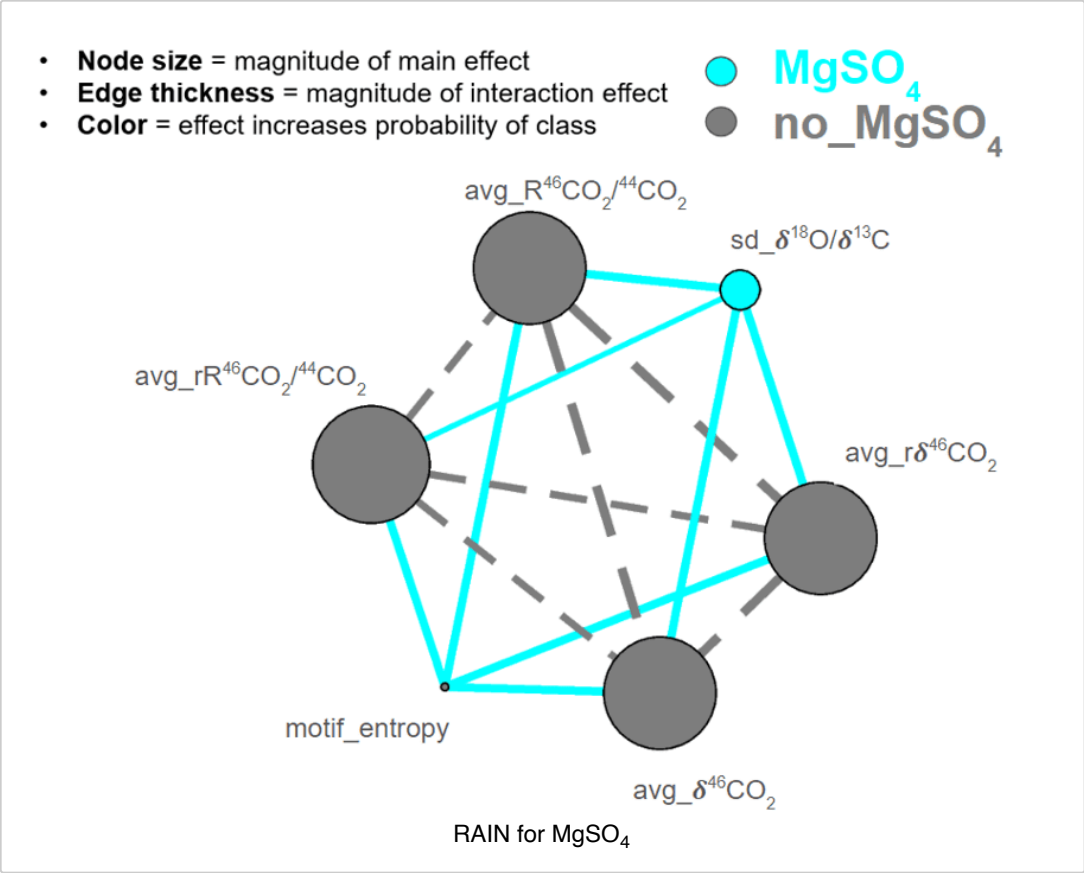
RAIN for biosignatures

We use NPDR feature selection (see previous panel) to identify a handful of features to predict biosignatures. We then apply RAIN analysis to the selected feaures to visualize the variable effects on classification.

In a RAIN, nodes indicate the magnitude of the variable's main effect, and edges indicate the magnitude of the interaction effect. The colors illustrate the probability direction for classification the effect has.

For example, to predict biosignatures, the variables "fluctuation" and "diff2_acf1" have moderately large main effects with different effects: for "diff2_acf1", the effect increases the probability of an abiotic classification, for "fluctuation", the main effect incrfeases the probability of biotic prediction. There is a strong interaction between these variables that increases the probability of an abiotic prediction.

We also use our ML methods for seawater chemistry prediction, including pH and ionic strength (see past AGU posters, Future Work, Resources, and More Information).



RAIN for $MgSO_4$

For predictions, we use **Random Forest (RF)** (Breiman 2001), a decision-tree based algorithm, in the NPDR selected feature spaces for outcomes of interest, such as biotic class, or $MgSO_4$ content.

Despite our small sample size, the NPDR features yield high RF model prediction test accuracies (> 88%) for biosignatures in repeated random splits.

Train data	Abiotic	Biotic
Number of samples	89	51

Test data	Abiotic	Biotic
Number of samples	22	12

Test accuracy = 88.2%	Abiotic	Biotic
Abiotic	20	2
Biotic	2	10

Left: training and testing sample sizes. **Right:** Classification table (or confusion matrix) for biosignatures. Rows represent true labels (there are 12 total biotic samples in the test data) and columns represent predictions (20 abiotic samples are correctly classified and two are inocrrectly predicted to be biotic).

We observe even higher accuracies for the prediction of salt components such as $MgSO_4$.

Train data	$MgSO_4$	no_$MgSO_4$
Samples	81	209

Test data	$MgSO_4$	no_$MgSO_4$
Samples	18	53

Test accuracy = 98.6%	$MgSO_4$	no_$MgSO_4$
Abiotic	17	1
Biotic	0	53

Left: training and testing sample sizes. **Right:** Classification table (or confusion matrix) for $MgSO_4$. Rows represent true labels (there are 18 total $MgSO_4$ samples in the test data) and columns represent predictions (53 $MgSO_4$ samples are correctly classified and one is inocrrectly predicted).

For false prediction diagnostics, we previously used Local Interpretable Model Eplainer (LIME) (Ribeiro et al. 2016) and RF local variable importance scores. LIME uses a local linear model and will not be able to account for interactions, and RF local importance requires re-training the model so that test samples can be OOB (out-of-bag), the random sampling method used to build trees and calculate variable importance (Breiman 2001).

We propose a new local feature importance method to diagnose false predictions based on NPDR, called **single-sample NPDR (ssNPDR).**

NOVEL SINGLE-SAMPLE MACHINE LEARNING ANALYSIS FOR FALSE PREDICTION DIAGNOSTICS

single-sample Nearest-neighbors Projected Distance Regression (ssNPDR) expands NPDR to enable local analysis.

$$\vec{\beta}^{global} = \min_{\beta_o,\vec{\beta}} \left(\sum_{i=1}^{m} \sum_{j \in N_k(i)} \mathcal{L}\left(\delta_{ij}(y); \beta_0 + \vec{\beta} \cdot \vec{d}_{ij}(A)\right) + \lambda\left(\alpha\|\vec{\beta}\|_1 + (1-\alpha)\|\vec{\beta}\|_2\right) \right)$$

Contrastive Loss

Lasso and Ridge Penalties

NPDR Scores

Using All Neighbors

Projected distances between neighbors for all attributes A

$$\vec{\beta}_i^{local} = \min_{\beta_o,\vec{\beta}} \left(\sum_{j \in N_k(i)} \mathcal{L}\left(\delta_{ij}(y); \beta_0 + \vec{\beta} \cdot \vec{d}_{ij}(A)\right) + \lambda\left(\alpha\|\vec{\beta}\|_1 + (1-\alpha)\|\vec{\beta}\|_2\right) \right)$$

NPDR Scores for one sample i

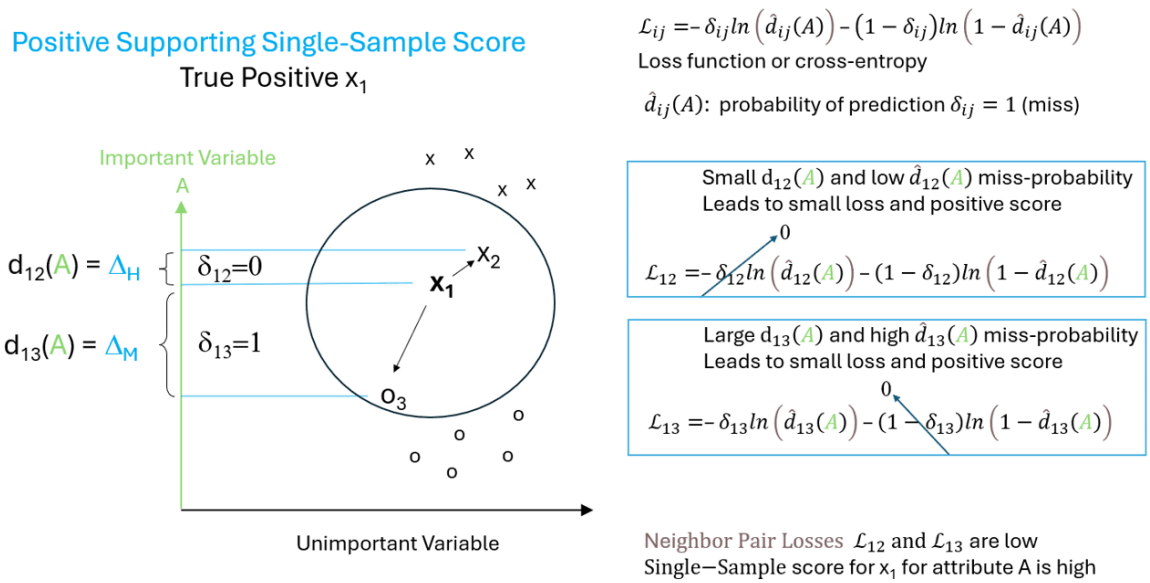
Only using neighbors for sample i

Contrastive Loss for a pair neighbors

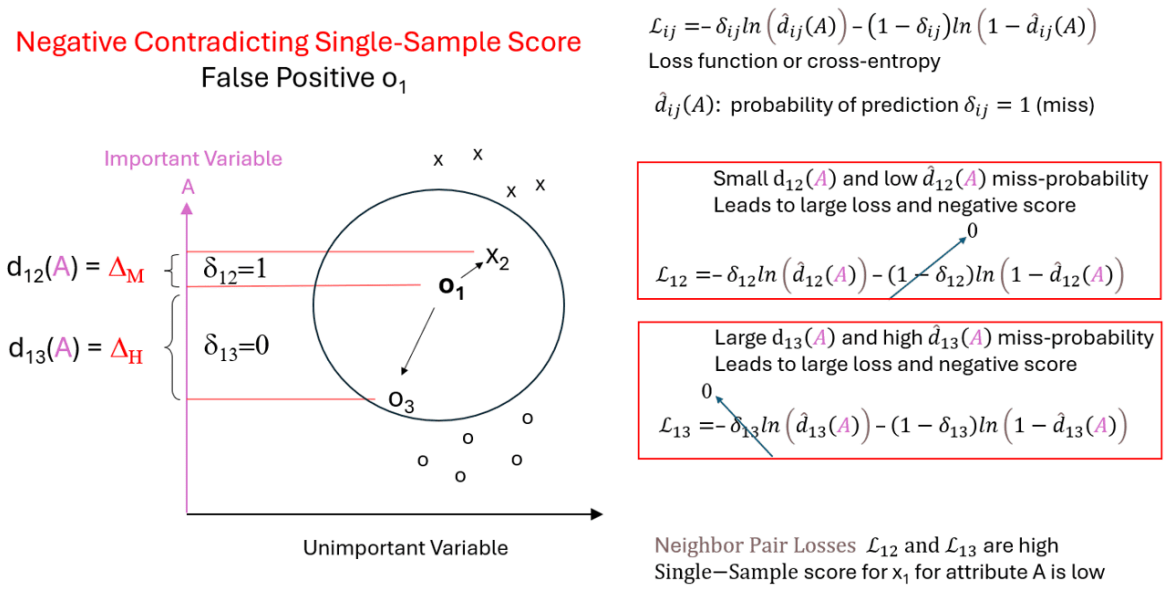
$$\mathcal{L}_{ij} = -\delta_{ij} ln\left(\hat{d}_{ij}(A)\right) - (1-\delta_{ij}) ln\left(1-\hat{d}_{ij}(A)\right)$$

$\hat{d}_{ij}(A)$: probability of predicting $\delta_{ij} = 1$ (miss)

Single-Sample NPDR Variable Importance. The top equation shows the NPDR objective function for finding variable weights using all neighbors. For single-sample NPDR (second equation), one sample i is fixed and the objective function sums the loss over all neighbor samples of i. Global and ssNPDR can use a LASSO α=1 or Ridge pentaly α=0 (Hesterberg et al. 2008, Zou and Hastie 2005). For reference, we show the loss for a pair of neighbors (Lij, bottom equation), where $\hat{d}_{ij}$ is the contrastive probability for the projected dstance to predict a class mismatch between neighbors, and δ_{ij} is the contrastive outcome between neighbors (1 if neighbors have a different class and 0 if the same class).



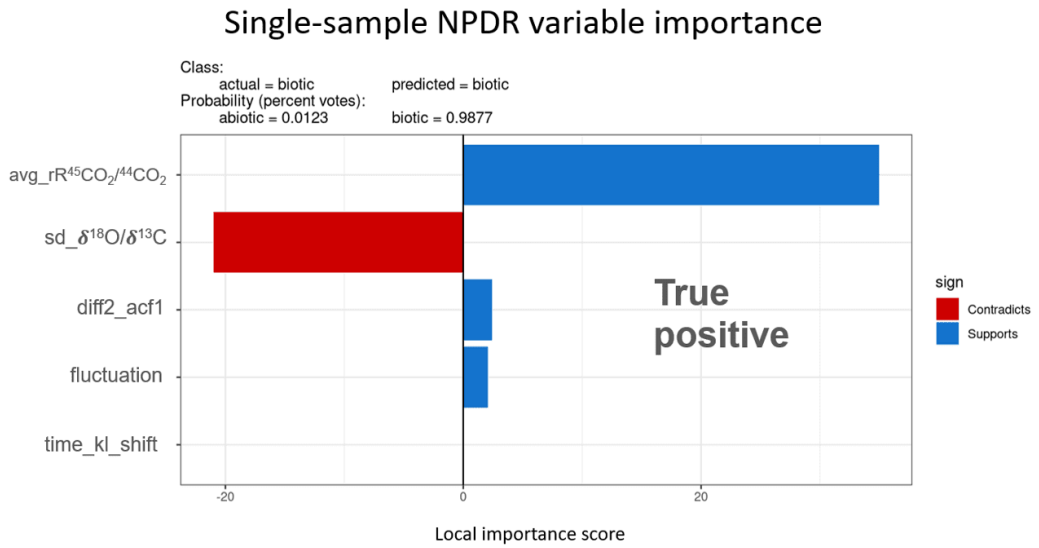
True prediction. The variable A is an important variable in the sense that it discriminates between the x-samples and o-samples. The single-sample NPDR score is illustrated for the true-positive sample x1 using its nearest neighbors x2 (same class) and o3 (different class). The contrastive loss for a pair of neighbors (L_{12}) is low when they are close to each other (small d_{12}) and their mismatch-delta is 0 ($\delta_{12} = 0$). The neighbor loss L_{13} is also low because the pair of samples are far apart (large d_{13}) and their mismatch delta equals 1 ($\delta_{12} = 1$). These low losses lead to a high single-sample NPDR importance score for true positive x_1 for the important variable A.



False prediction. Same data as previous figure, but sample 1 is now incrorreclty labeled or predicted to be class o1 instead of x1 (false classification), resulting in high contrastive neighbor losses L_{12} and L_{13} and a negative single-sample NPDR score for variable A, which should have positive score because it discriminates between x's and o's.

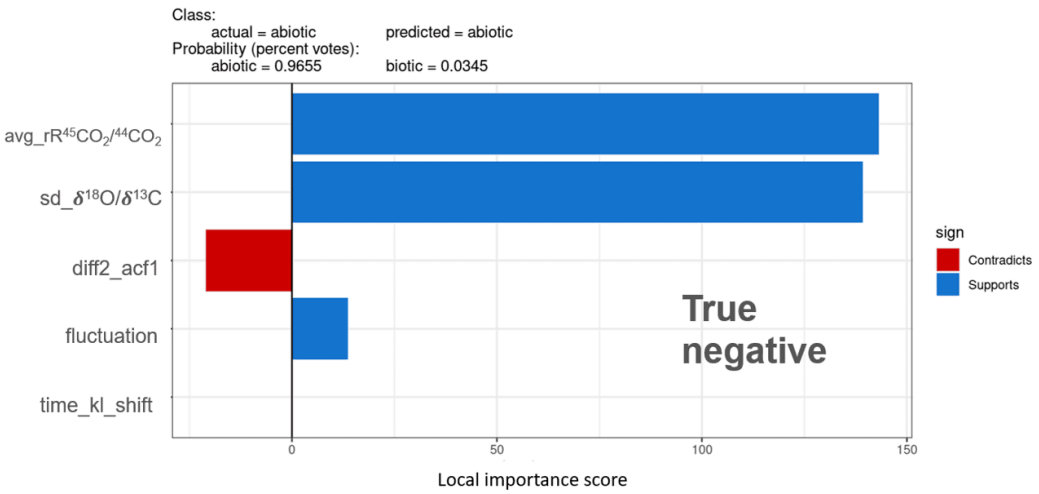
RESULTS AND CONCLUSIONS: DIAGNOSING FALSE PREDICTIONS

ssNPDR can be used to diagnose false biosignature predictions and boost confidence in positive detections.

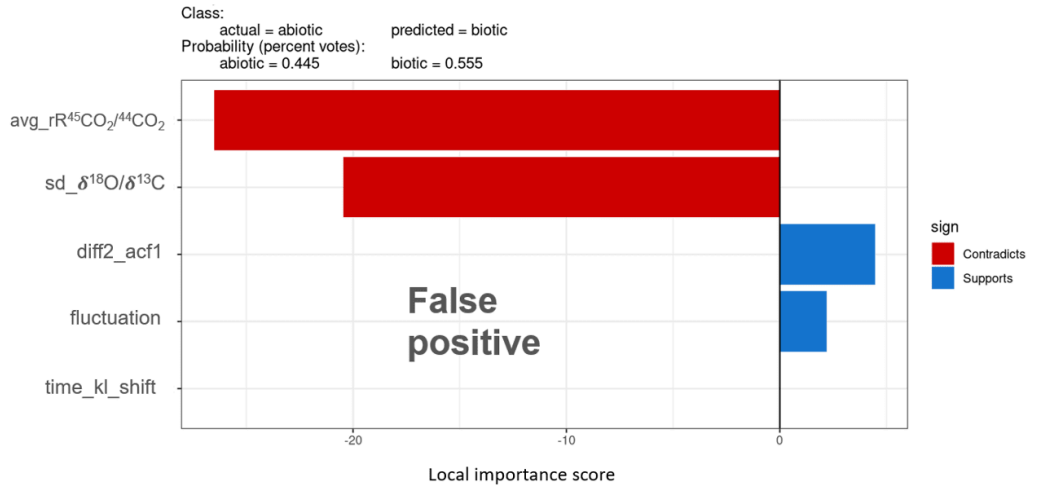


A known true positive prediction (correct biotic prediction) shows a high-magnitude blue bar for the top NPDR-selected feature for biosignatures. Additionally, this prediction has a high class probability of 98.8%, provided by the RF model.

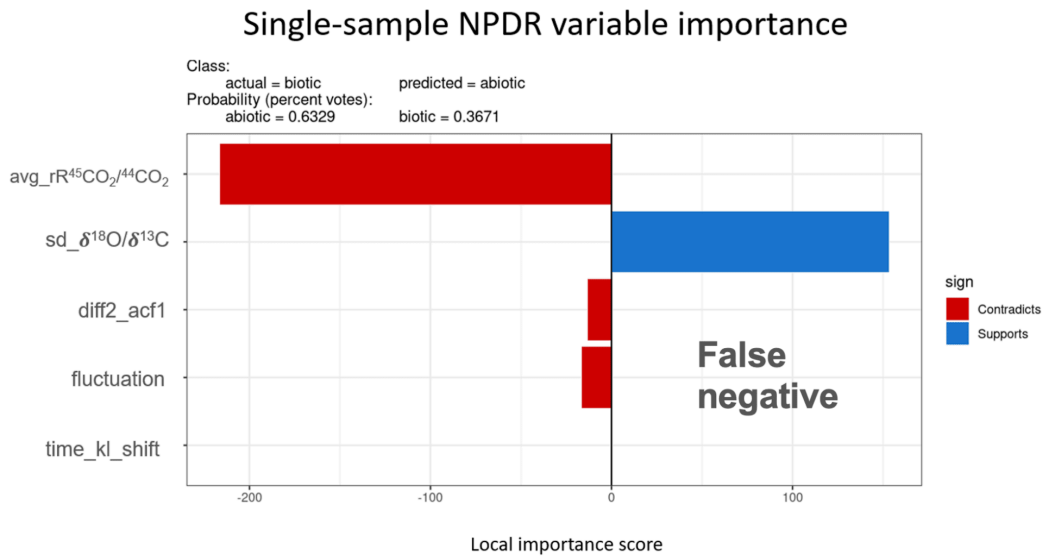
ssNPDR variable importance scores indicate whether a feature value for a sample is supporting (positive sign, blue bars) or contradicting (negative sign, red bars) the predicted class label. Signs are determined during ssNPDR and are based on the loss function. Features are shown on the *y*-axis in order of NPDR importance. The local importance scores indicate the magnitude of the feature's coefficient for the variable from the optimization of the loss function (see previous panel).



A known true negative prediction (correct abiotic prediciton) shows high-magnitude blue bars for the top two NPDR-selected feature for biosignatures. Additionally, this prediction has a high class probability of 96.6%.



A known false positive prediction (incorrect biotic prediction) shows high-magnitude red bars for the top NPDR-selected feature for biosignatures. Additionally, this prediction has a low class probability of 55.5%.



A known false negative prediction (incorrect abiotic prediction) shows a high-magnitude red bar for the top NPDR-selected feature for biosignatures. Additionally, there are two other red bars and this prediction has a relatively low class probability of 63.3%.

FUTURE WORK, RESOURCES AND MORE INFORMATION

Future Work

- Further analysis and quantification of ssNPDR performance using simulated data; NPDR library has functions for the simulation of realistic data (with correlation structure and class imbalance options)
- Analysis of false predictions for salts
- Extension to continuous outcome variables

Resources and More Information

McKinney Bioinformatics Lab:

http://insilico.utulsa.edu/ (http://insilico.utulsa.edu/)

NPDR feature selection:

https://github.com/insilico/npdr (https://github.com/insilico/npdr)

Volatile CO2 IRMS Quality Control Method:

https://github.com/insilico/QCIRMS (https://github.com/insilico/QCIRMS)

Previous AGU Posters:

AGU 2023 - Interpretable Machine Learning Models for Autonomous Characterization of Analogue Ocean World Seawater Chemistry and Biosignature Potential using Isotope Ratio Data (https://agu23.ipostersessions.com/templates/iposters/templatepdf.aspx?s=11-28-0D-AB-61-C6-D5-3D-1C-E4-8D-4F-82-E5-AE-1C)

AGU 2022 - Autonomous Astrobiology Pipeline for Biosignature Detection from Ocean World Mass Spectrometry: Data-prioritization and Machine Learning (https://agu2022fallmeeting-agu.ipostersessions.com/Default.aspx?s=DB-B7-43-A7-A8-E6-EF-7D-40-07-4E-83-25-F5-E3-DC)

TRANSCRIPT

ABSTRACT

Isotope ratio mass spectrometry (IRMS) of volatiles (e.g., CO_2) promises to be a powerful tool for potential biosignature detection for future missions to ocean worlds (OW) such as Europa and Enceladus. Machine learning (ML) methods for IRMS data could enable science autonomy by onboard prediction of seawater chemistry and biosignature presence. However, ML models are likely to be complex and involve statistical interactions between features (variables), which can make predictions seem opaque and enigmatic. For ML predictions as significant as extraterrestrial biosignatures, we must place extraordinary confidence in models. It is therefore essential that these models make interpretable predictions (*i.e.*, human-understandable) and include false-prediction diagnostics. We achieve high accuracy and interpretability in ML biosignature and seawater chemistry models for OW through a nearest-neighbors feature selection tool that detects statistical interactions between predictors, constructs interaction networks for visualization of selected features working together to make a prediction, and reports single-sample feature importance scores for false-detection diagnostics. Here we develop a novel single-sample nearest-neighbors projected distance regression (ssNPDR) feature selection method that improves upon existing single-sample algorithms through the inclusion of statistical interactions while providing false-prediction diagnostics for ML models.

REFERENCES

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

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

Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. Frontiers in Astronomy and Space Sciences, 10. https://www.frontiersin.org/articles/10.3389/fspas.2023.1134141

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

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

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

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

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

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

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier (arXiv:1602.04938). arXiv. https://doi.org/10.48550/arXiv.1602.04938

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

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

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

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

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

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

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

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

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

EVALUATIONS

#	Average Score
There are currently no completed evaluations for this presentation	