Cell Classification Using Single Cell Mass Spectrometry Through Interpretable Machine Learning

Yuxuan Richard Xie[1,3], **Daniel C. Castro**[3], **Sara E. Bell**[2,3], **Stanislav Rubakhin**[2,3], **Jonathan V. Sweedler**[1,2,3]

[1]Department of Bioengineering, [2]Department of Chemistry, [3]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign

OVERVIEWS

- Multicellular organisms are made up of individual cells with distinct chemistries, morphologies and functions. Differentiation of previously defined cell types at the single cell level based on their chemical profiles is poorly understood.
- Various single cell MALDI-MS datasets containing different cell groups were generated to train machine learning models for cell classification given single cell spectra.
- Models were subjected to a pipeline to enable global and local interpretations of chemical features. Top features for the classification tasks were selected, facilitating downstream analysis and validation.

INTRODUCTION

- Recent advancements in mass spectrometry such as matrix-assisted laser desorption/ ionization (MALDI) MS enabled high-throughput chemical analysis of dissociated single cells from animal models, capable of resolving hundreds of biomolecules in a single mass spectrum.[1]
- However, few efforts are made to improve the informatic counterpart for better data interpretation and information mining.
- Here, a single cell classification workflow[2] using mass spectrometry is introduced and demonstrated, providing new ways to interpret the data and to gain biological insights.

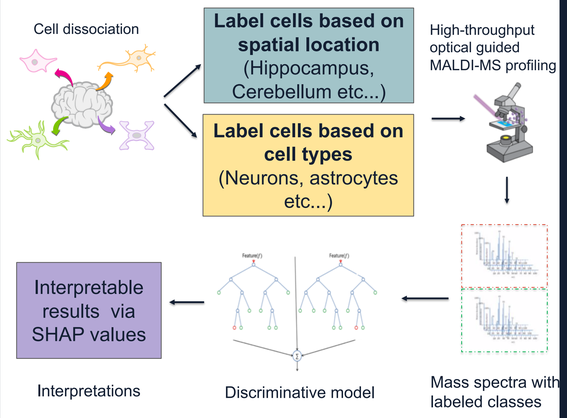
AIMS

- Preparing rodent single cell dissociates with prior information such as their canonical cell types or cell origins.
- Classifying cell types given single cell mass spectra using machine learning models.
- Interpreting the trained classification models using SHAP.[3]
- Ranking *m/z* features for feature selection.
- Classifying large number of unlabeled single cell spectra from rodent cerebellar dissociates[4] into neurons and astrocytes using models trained on labeled data.[5]

METHODS

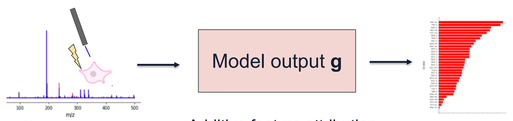
Cell classification workflow

- Single cells were dissociated from tissue. Cells were labeled based on either canonical cell types (e.g. astrocytes vs. neurons) or anatomical brain regions (hippocampus vs. cerebellum).
- Discriminative model (gradient boost tree) were trained to predict the cell classes given mass spectra.



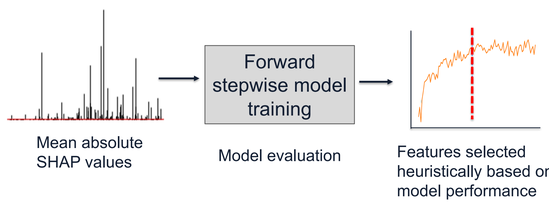
Model interpretation

- The trained model was then subjected to interpretations for local feature explanations. SHAP values are computed for each single cell spectra



Feature selection

- Mean absolute SHAP values used as the ranking metric to perform feature selection. Models were trained iteratively by incremental numbers of ranked features as input variables.



RESULTS

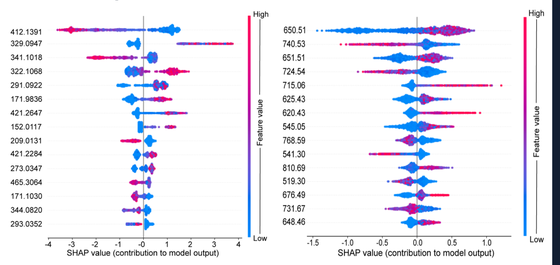
Model performance:

- Three datasets were used to demonstrate the cell classification workflow.
- Data were split into training (70%) and test set (30%).

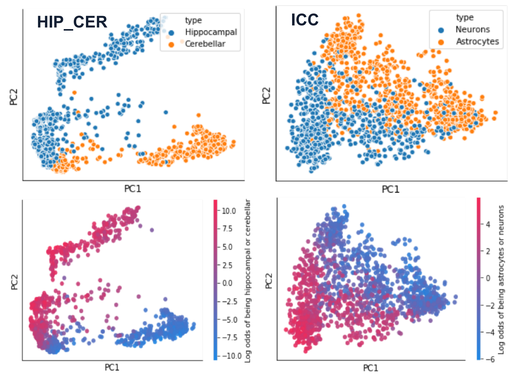
Dataset	Accuracy	AUC
HIP_CER	0.963	0.995
ICC	0.774	0.833
SIMS	0.995	0.996

- *HIP_CER dataset (1201 cells in total, acquired on 7T FT-ICR mass spectrometer),*
- *ICC dataset (neurons vs. astrocytes, 1544 cells, acquired on TOF)*
- *SIMS dataset (Cerebellar neurons vs. Dorsal Root Ganglia, 1542 cells in total, acquired by SIMS)*

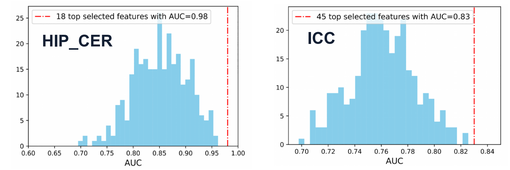
Model interpretation via SHAP:



Summary plots of the top contributing mass features. Color represents normalized mass peak intensity, and positive and negative SHAP values indicate the feature contribution to the predicted probability of the cell being a specific class.

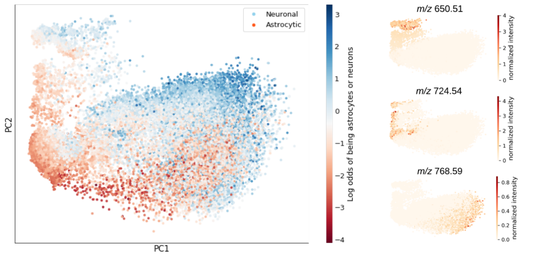


Performing PCA on SHAP values shows the model decisions in low-dimensional space. Non-contributing features have zero or very small SHAP values, whereas contributing features relevant to classification tasks have larger SHAP values. The variations in the feature intensity only affects the outcome of PCA if they impact the model output.



Distribution of scores (AUC) from models trained on randomly selected features and the models trained on SHAP selected features (red dash line). 18 features were selected for the HIP_CER dataset and 45 features were selected for the ICC dataset.

RESULTS (cont.)



Model trained on ICC dataset (neuron vs astrocyte) was used to classify 30,000 unlabeled single cells. The SHAP values reveal neuronal cells and astrocytic cells-specific knowledge. Top selected features for each cell class agree with previous study and are well-localized with model predictions.

CONCLUSIONS

- We have developed a workflow for cell classification through single cell mass spectrometry analysis and interpretable machine learning.
- SHAP allows both global and local interpretations of single cell spectra toward predicting cell groups of interests.
- The top-identified mass features are consistent with previous studies.
- Further validation is still required for confident identification of these molecules.

REFERENCES

1. Comi, T. J.; Neumann, E. K.; Do, T. D.; Sweedler, J. V., J. Am. Soc. Mass Spectrom. 2017, 28, 1919–1928.
2. Xie, Y.R.; Castro, D.C.; Bell, S.E; Rubakhin, S.S.; Sweedler, J.V., Anal. Chem. (submitted)
3. Lundberg, S. M.; Lee, S.-I. In Advances in Neural Information Processing Systems 30; 2017; pp 4765–4774.
4. Neumann, E. K.; Ellis, J. F.; Triplett, A. E.; Rubakhin, S. S.; Sweedler, J. V., Anal. Chem. 2019, 91, 7871–7878.
5. Neumann, E. K.; Comi, T. J.; Rubakhin, S. S.; Sweedler, J. V., Angew. Chem. Int. Ed. 2019, 58, 5910–5914.

ACKNOWLEDGMENTS

This project was supported by the National Institute on Drug Abuse and the National Human Genome Research Institute.

ILLINOIS