Unlocking Molecular Insights: A Machine Learning-Driven Pipeline for Metabolomics

Christian Ayala-Ortiz[1], Sumudu Rajakaruna[1], Ernesto Delamaza[2], Dalal Alharthi[2], **Malak M. Tfaily[1]**

[1]Department of Environmental Science, University of Arizona, Tucson, Arizona, 85721, [2] College of Applied Science and Technology, University of Arizona, Sierra Vista, Arizona, 85635

Introduction

- Metabolites are usually at the end of complex biochemical cascades allowing for a better understanding of phenotypic responses [1].
- Metabolite annotation is crucial for extracting biological inferences from metabolomics datasets [2]. However, their complexity and heterogeneity make them challenging to study [3].
- True metabolite identification requires the use of in-house libraries to allow for the comparison of retention times, mass and fragmentation spectra [4]. However available standards are limited and not always available for all researchers [5].
- Machine learning approaches can help in this issue by providing means to "predict" the RT time of compounds with known structure that have been reported in databases [6].

Pipeline workflow

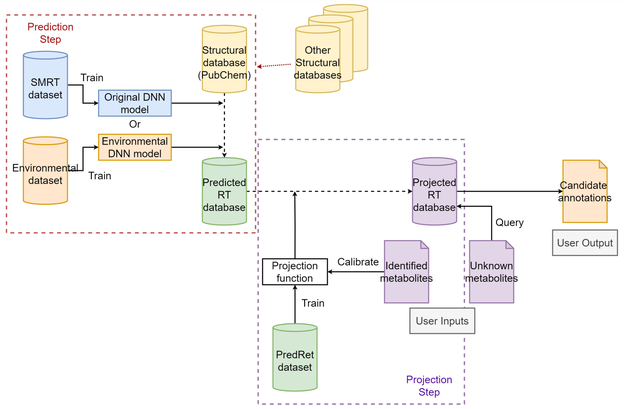

Figure 1. Workflow of the CMM-RT pipeline [6] showing the different datasets that are used in each of the two main steps. A more focused deep neural network (DNN) model was created using an environmentally relevant dataset.

Results

- Hand selecting features to use as known compounds for the projection step can help to better calibrate the projection functions (Figure 2).

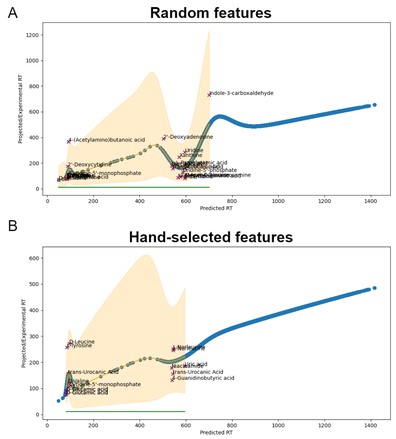

Figure 2. Projection functions using A) random selected features and B) hand-selected features. The projection function is smoother in the second case

- Annotations from the environmental model were more accurate (>90%) and with projected RT closer to the experimental RT

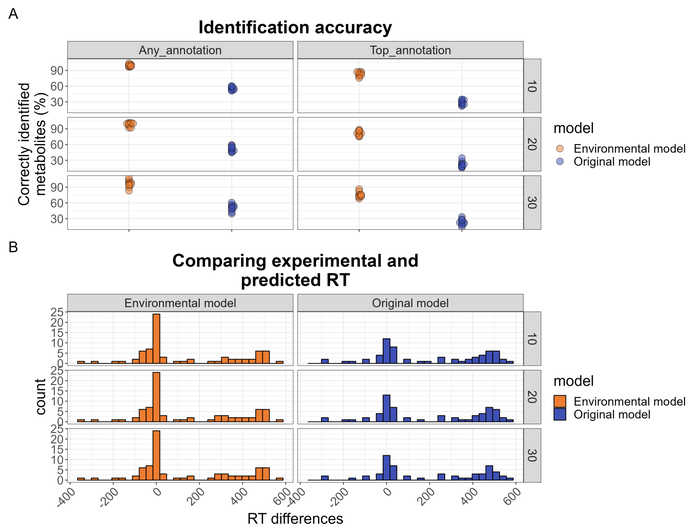

Figure 3. A) Metabolite annotation accuracy for the two models in the two datasets using 10, 20 and 30 random features (repeated 20 times each). B) Differences between projected and experimental retention time (RT)

Results (Cont.)

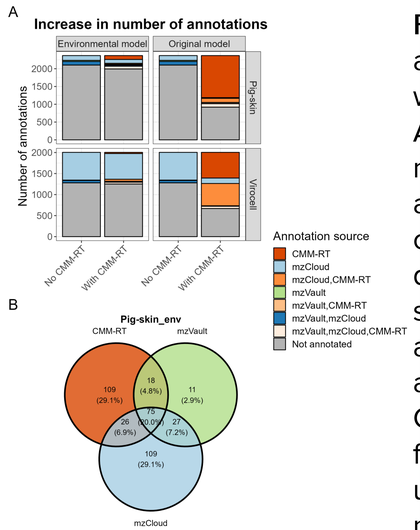

Figure 4. A) Number of annotated metabolites with different methods. A larger number of metabolites are annotated with the original model. B) Venn diagram showing the source of the annotation, most annotations from the CMM-RT pipeline are for previously unannotated metabolites

Conclusions and Future Work

- The use of machine learning approaches for the prediction of retention times is a valuable tool to increase the number of annotated metabolites that can be retrieved from a particular system.
- The use of a more "focused" retention time prediction model can allow to produce more accurate annotations, however, there is a tradeoff with the number of annotated metabolites when compared to a more "broader" model.
- The scoring system for the candidate annotations needs further improvement to help in the selection of the "correct" metabolite annotation.
- A web app is being simultaneously developed to share the tool with more members of the metabolomics community

REFERENCES

1. Fiehn, O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155–171 (2002).
2. Viant, M. R., Kurland, I. J., Jones, M. R. & Dunn, W. B. How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* **36**, 64–69 (2017).
3. Cooper, W. T. *et al.* A History of Molecular Level Analysis of Natural Organic Matter by FTICR Mass Spectrometry and The Paradigm Shift in Organic Geochemistry. *Mass Spectrom. Rev.* **41**, 215–239 (2022).
4. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
5. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted Metabolomics Strategies-Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).
6. García, C. A., Gil-de-la-Fuente, A., Barbas, C. & Otero, A. Probabilistic metabolite annotation using retention time prediction and meta-learned projections. *J. Cheminform.* **14**, 33 (2022).

Get this poster

ACKNOWLEDGEMENTS

- We thank members of the Tfaily Lab Sumudu Rajakaruna, Viviana Freire-Zapata and John Bourannis for providing access to the data
- **Funding:** Bio5 Grant