IECM 2025 Conference
The 4th International Electronic Conference on Metabolomics
13–15 October 2025 | Online
MDPI

A New Enzymatic Distance Concept Enables Machine Learning Regression of Metabolite Chemical Representation Features to Distance in Metabolism

Grace Sun[1], Erik D. Huckvale[1], Hunter N.B. Moseley[1]
[1]Department of Molecular and Cellular Biochemistry, University of Kentucky, CC434 Roach Building, 800 Rose Street, Lexington, KY 40536
hunter.moseley@uky.edu

INTRODUCTION

Alterations in human metabolic processes can lead to changes in enzyme expression, often becoming a notable hallmark of various systemic diseases. As a result, interpreting metabolic pathways and changes becomes critical to understanding metabolic-driven pathogenesis and finding potential drug targets.

Complex techniques such as nuclear magnetic resonance (NMR) and mass spectrometry (MS) are used to detect and characterize metabolites[1]. Annotating and understanding these metabolites can interpret target regions of a pathway and help identify potential drug targets to alleviate observed disease-driven differences.

Problem: In most metabolomics datasets, it is possible for more than half of the experimentally detected metabolites to be unannotated[2]. This makes their metabolic roles unknown, hindering pathway-specific interpretation.

Solution: Develop a way to place unannotated metabolites into a metabolic network and predict that involvement only given their atomic-level features, which are detectable experimentally.

Research Goal

A distance concept could help determine the role of an unannotated metabolite in a metabolic network by finding its "distance" in relation to known metabolites. The goal is to develop a way to represent relationships between metabolites & predict those relationships given metabolites' chemical features.

NEW CONCEPT: ENZYMATIC DISTANCE

This study presents a concept called "**enzymatic distance"**, which represents the number of enzymatic steps and the mass flow separating any two metabolites in a metabolic network.

Counts of **atom mappings** and **reaction centers** between metabolites were selected as quantitative metrics to represent Enzymatic Distance:

Atom Mappings

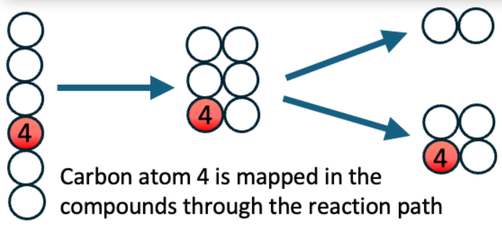

Fig. 1 Schematic to describe the atoms mapping through compounds in a reaction. Creative reference from [6].

Reaction Centers

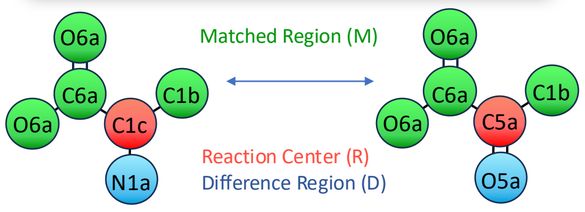

Fig. 2 RDM Pattern for RCLASS RC00006 with color-coded reaction structures. Generated using data from KEGG [7].

METHODOLOGY

KEGG: Kyoto Encyclopedia of Genes and Genomes [7]*:* *https://www.genome.jp/kegg/*

New metric and dataset creation → Metric prediction from chemical structure features → Mechanistic Interpretation of Metabolic Network

DATASET CURATION & CREATION

1. Harmonized KEGG Pull
- COMPOUND and REACTION data from KEGG[7]
- Atom mapping data generated by md_harmonize[4]

2. Create reaction paths between metabolites

Each meta RCLASS represents a path of reactions to transform a metabolite into another.

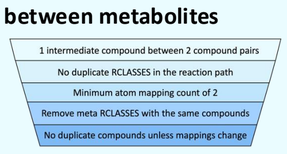

Fig. 3 Meta RCLASS creation criteria

3. Calculate enzymatic distance metrics

Atom mappings and reaction centers between two metabolites are calculated and averaged over all the meta RCLASSES for a metabolite pair

1. **Atom mapping counts**: Intersection between the mappings of two pairs of metabolite compounds, with an intermediate compound.
2. **Reaction centers**: Sum of the reaction centers of component RCLASSES in the reaction path.

4. Generate atom coloring feature vectors

Acquired feature vectors to represent the substructures of metabolite compounds from [2]. Each vector includes 16509 chemical features, with each feature entry representing the count of a specific substructure in the metabolite.

ENZYMATIC DISTANCE PREDICTION

1. Build machine learning training/testing dataset

Inputs: Pairs of metabolite atom color feature vectors

Outputs: Enzymatic distance metrics

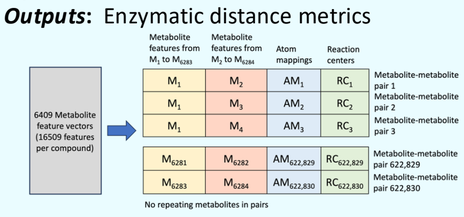

Fig. 4 Diagram of the dataset of metabolite-metabolite features pairs with their enzymatic distance metrics. Creative reference from [2].

2. Develop models for enzymatic distance prediction

Compare univariate & multivariate MLP regression models to predict enzymatic distance metrics

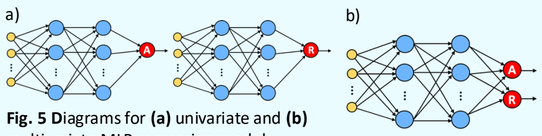

Fig. 5 Diagrams for **(a)** univariate and **(b)** multivariate MLP regression models.

Model Details:
- *Multi-Layer Perceptron model – PyTorch package*
- *Trained with Mean Squared Error loss*
- *Hyperparameters: 100 epochs, 3 layers*

Model Evaluation Criteria: Model evaluated with R-squared (R^2) (evaluation of variance representation) over 30 5-fold cross-validation iterations using a single test fold from each iteration

3. Evaluate metabolite chemical feature importance

Used the Python SHAP library[5] for MLP model feature attribution to determine important substructures in metabolites for predicting enzymatic distance.

RESULTS

Enzymatic distances were found between 622,830 viable pairs of metabolites, which were generated from 8,312,131 reaction paths.

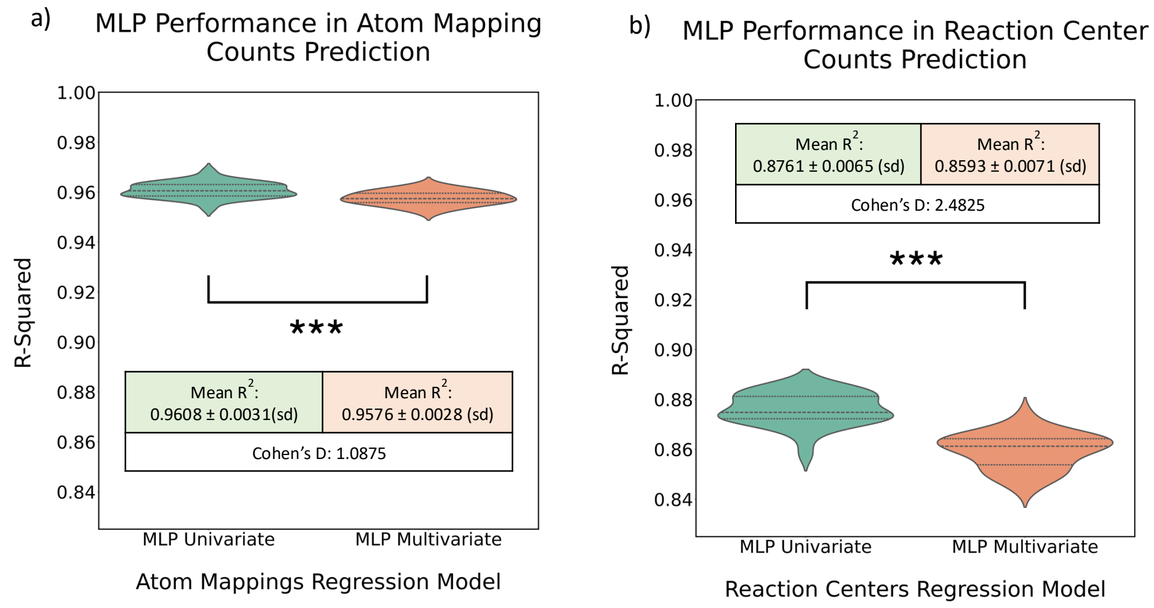

Fig. 6 Violin plots of the R-squared values (prediction performance) of **(a)** atom mapping predictions and **(b)** reaction center predictions over 30 cross-validation iterations, depending on the number of variables predicted. The univariate MLP model is colored in green and the multivariate MLP model is colored in orange. The R-squared performance metrics of each model are colored the same. Evaluated with two-sample equal variance t-tests ($p < 0.001$ [***]). Created with Matplotlib.

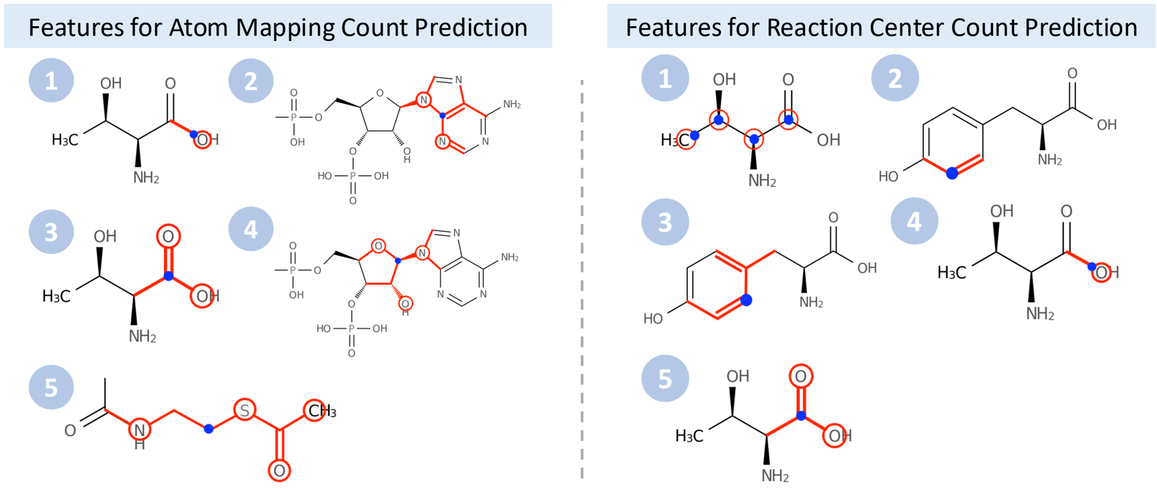

Fig. 7 Molecular visualization for the most important features (ranked) for atom mapping and reaction center count prediction. The feature is identified in red, with the central atom identified in blue. The chemical structure images are acquired from ChEBI [6].

CONCLUSIONS

This project developed the concept of "enzymatic distance", which has significant potential for improving functional interpretation of metabolomics datasets.

Atom mappings & reaction centers → new concept of enzymatic distance

Effective prediction of enzymatic distances through univariate regression

Identified chemical structure features important to enzymatic distance

- Atom mappings are conducive to enzymatic distance prediction due to their contextual similarity with atom color features, which are the model's inputs.
- The univariate regression model predicted enzymatic distance more accurately than multivariate.
- Features with oxygen & nitrogen play a significant role in defining enzymatic distance in metabolism.
- Complex substructural features (in terms of the number of bonds from the central atom and non-carbon atoms) play a significant role in atom mapping prediction.

References

[1] Sumner, L. W., *et al*. (2007). *Metabolomics*, *3*(3), 211–221. 10.1007/s11306-007-0082-2
[2] Huckvale, E. D., & Moseley, H. N. (2024). *Metabolites*, *14*(5), 266. 10.3390/metabo14050266
[3] Starke, C., & Wegner, A. (2022). *Metabolites*, *12*(2), 122. 10.3390/metabo12020122
[4] Jin, H., & Moseley, H. N. (2023). *Metabolites*, *13*(12), 1199. 10.3390/metabo13121199
[5] Lundberg S. M., Lee S. I. (2017). *Advances in Neural Information Processing Systems*, 4765–4774. 10.48550/arXiv.1705.07874
[6] Degtyarenko, K., *et al*. (2007). *Nucleic acids research*. *36*(suppl_1), D344-D350. 10.1093/nar/gkm791
[7] Kanehisa, M., *et al*. (2012). *Nucleic acids research*. 40(D1), D109-14. 10.1093/nar/gkr988

Acknowledgements

This research was funded by the National Science Foundation, grant number 2020026 (PI Moseley)