Using machine learning to model promiscuous activity of thiamine diphosphate-dependent carboligases and side reactions in the *E. coli* metabolome

Tracey Dinh[1]*(tracey.dinh@u.northwestern.edu), Bradley W. Biggs[1], Matthew T. Robey[2], Catherine Majors[1], Lindsay Caesar[2], Neil L. Kelleher[2,3,4], Paul M. Thomas[4], Linda J. Broadbelt[1], **Keith E.J. Tyo[1]**

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL; [2]Department of Molecular Biosciences, Northwestern University, Evanston, IL; [3]Department of Chemistry, Northwestern University, Evanston, IL; and [4]Proteomics Center of Excellence, Northwestern University, Evanston, IL

https://pamspublic.science.energy.gov/CCBond

Project Goals: The goal of this project is to develop a predictive machine learning model that will elucidate the catalytic landscape of thiamine-diphosphate dependent carboligase enzymes across the chemical space of α-ketoacid substrates. This model will be used to identify enzymes for biosynthetic applications and predict potential effects of selected enzymes on the *E. coli* metabolome.

Abstract: Increasing recognition of non-canonical enzyme activity has revealed potential problems for heterologous expression; however, understanding the potential cell burden due to promiscuous enzyme activity remains a challenge. Toward this end, our team seeks to develop cheminformatics tools that predict enzyme substrate promiscuity and predict the resulting metabolomic consequences. Specifically, we focus on the biological activity of a family of thiamine-diphosphate dependent enzymes capable of catalyzing the condensation of α-ketoacids and aldehydes. Molecular fingerprints allow us to characterize the chemical space of α-ketoacid substrates, from which we sample and screen for enzyme activity in high-throughput. Using an active learning approach, we plan to train a support vector machine learning model on this bioactivity data. Robust models for each carboligase enzyme will not only allow us to identify novel biosynthetic reactions but also apply predicted promiscuous activity to genome-scale models of host organisms such as *E. coli*. We plan to carry out flux balance analysis to characterize effects of side reactions and select enzymes with minimal cell burden.

This work is supported by DOE grant DE-SC0019339.