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

Sun, Grace (School: Paul Laurence Dunbar High School)

Understanding and reconstructing pathways in metabolism can help identify biomarkers and metabolic changes associated with systemic diseases, such as cancer and cardiovascular diseases. However, in most metabolomics datasets, more than half of the experimentally detected metabolites can be unannotated, making their pathway-specific involvement largely unknown. This study suggests that a distance concept can help place an unannotated metabolite into specific parts of a metabolic network by finding its “distance” in relation to known metabolites. The new concept of “Enzymatic Distance” was created to quantify the separation of any two metabolites in metabolism in terms of enzymatic steps and the flow of mass between them. Enzymatic Distance was scored by two metrics: counts of shared atom mappings and reaction center atoms averaged over the reaction paths separating two compounds. Then, univariate and multivariate multi-layered perception regression models were trained and evaluated over 100 epochs and 30 cross-validation iterations to predict enzymatic distance metrics between target metabolites given only their chemical structural features. While atom mappings were robustly predicted by univariate ($R^2=0.961$) and multivariate ($R^2=0.958$) regression, reaction centers were predicted significantly less accurately by univariate ($R^2=0.876$) and multivariate ($R^2=0.859$) regression. The univariate regression model predicted both metrics more accurately than the multivariate regression ($p\text{-value}<0.001$), as predicting both metrics separately renders higher prediction accuracy. Quantifying and predicting the relationships between known and unknown metabolites helps interpret unannotated metabolites in the human body, aiding the detection and interpretation of metabolic diseases.

Awards Won: