Machine learning algorithms in big data analyses identify determinants of insulin gene transcription

**Abstract**

This paper will show analyzing over 490,000,000 data points to compare 10 different ML algorithms in a large (N=11,652) training dataset of human pancreatic single-cell transcriptomes to identify features (genes) associated with the presence or absence of insulin transcript(s). Prediction accuracy/sensitivity of models were tested in a separate validation dataset (N=2,913) and the performance of each ML-workflow assessed. Random Forest ML algorithm delivered high predictive power in a receiver operator characteristic (ROC) curve analysis at the highest sensitivity (0.98), compared to other algorithms. The top-10 features, (including IAPP, ADCYAP1, LDHA and SST) common to the three Ensemble ML workflows were significantly dysregulated in scRNA-seq datasets from Ire-1αβ-/- mice that demonstrate de-differentiation of pancreatic β-cells as well as in pancreatic single cells from individuals with Type 2 Diabetes. Findings here provide a direct comparison of ML workflows in big data analyses, identify key determinants of insulin transcription and provide workflows for other regulatory analyses to identify/validate novel genes/features of endocrine pancreatic gene transcription.