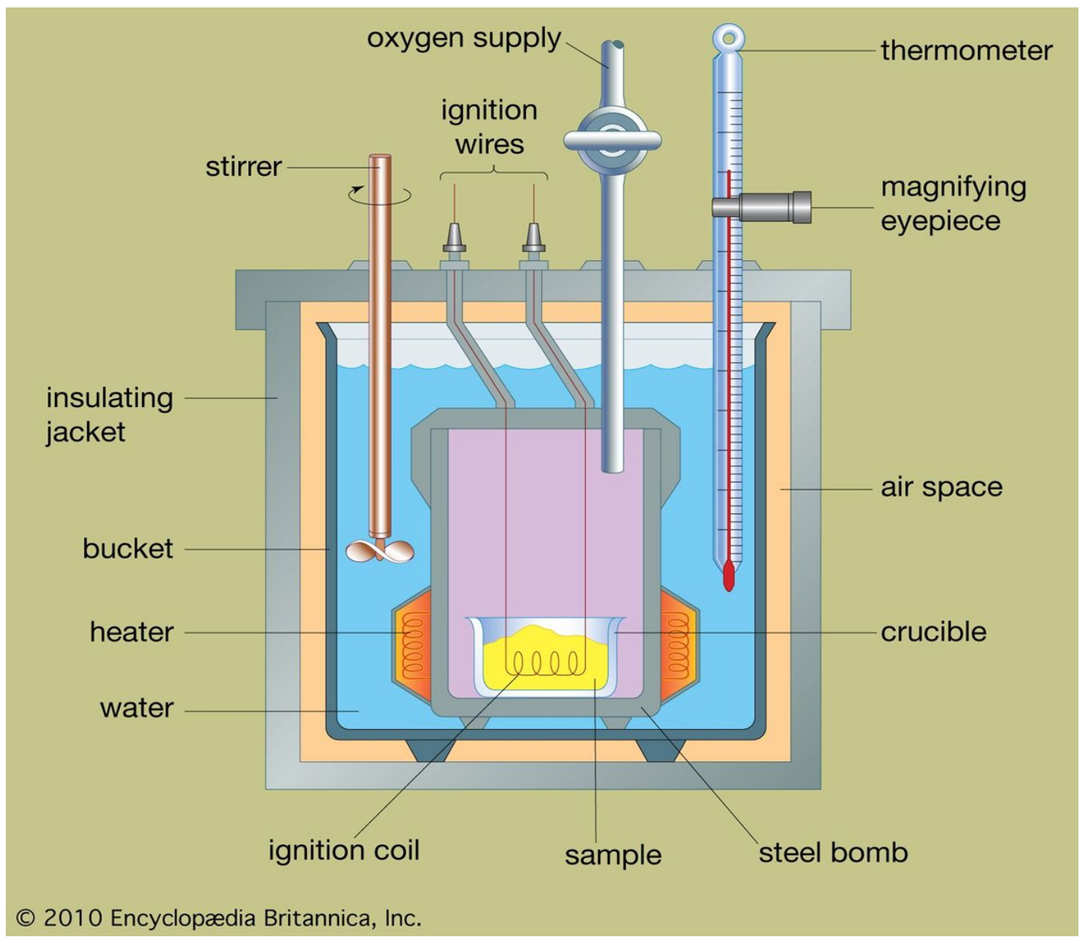
Predicting Heat of Combustion from SMILES with Machine Learning

Braden Garrett M.S. Candidate, Jingpeng Tang, Ph.D., Xi Chen, Ph.D., Department of Computer Science

Project Summary

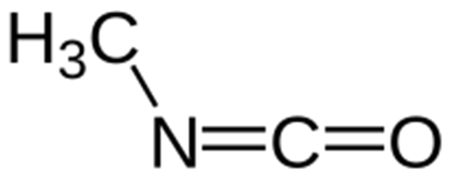
Accurately estimating the heat of combustion is important across energy, materials, and chemical applications. Traditional methods like bomb calorimetry are slow, expensive, and hazardous. This project explores using machine learning to predict combustion energy from molecules using a format called Simplified Molecular-Input Line Entry System (SMILES). A variety of machine learning models were trained on molecular data using SMILES-based embeddings. Early results showed moderate accuracy, which improved significantly after augmenting the dataset with Neural Network showing the best accuracy. The approach shows potential for replacing or supplementing experimental methods. Ongoing work includes expanding data sources and optimizing model performance.

Bomb Calorimetry



SMILES

Methyl isocyanate (MIC) CN=C=O

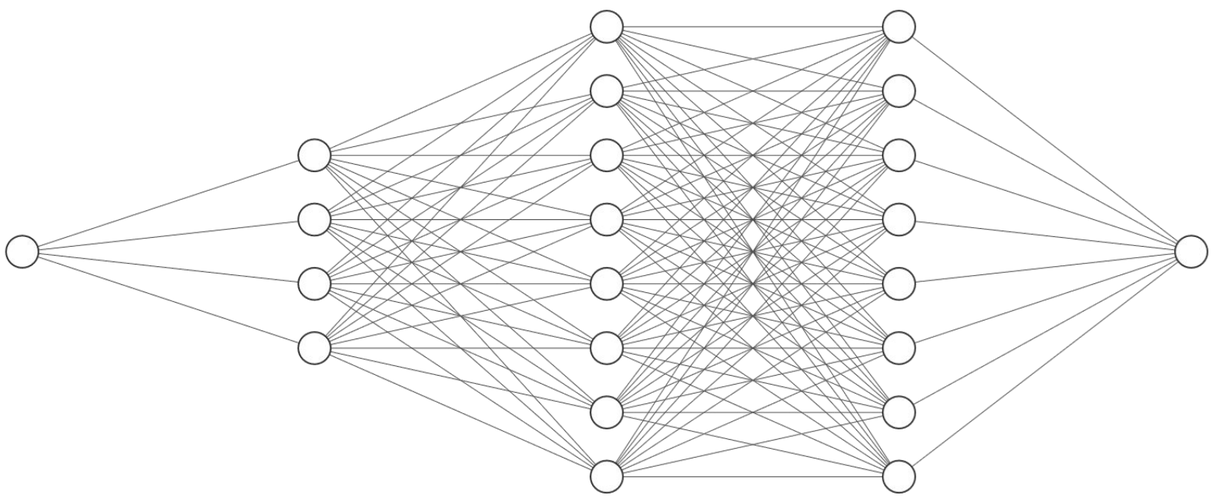


Training

Model Type	R^2	MSE
Random Forest	0.09610	7,353,301
XGBoost	-0.04456	12,084,850
SVM	0.17999	9,525,979
CNN	0.522	9,970,441

SMILES	Prediction	Target	Difference
Most Accurate			
NC	-1088.39	-1086.81	1.58
N#CC#N	-1093.25	-1097.07	3.81
C(CN)C=O	4164.06	4159.50	4.56
Least Accurate			
CC1C2CCC1CC(C2)…	1067.86	4863.73	3795.86
C1=NC(=C(N1)C(=O…	1041.53	-1910.62	2952.15
C1=CC=C2C(=C1)C3…	7133.93	10040.60	2906.6

Model Architecture



Testing Results

Test Mean Average Error	Test R^2
1589.12	0.296