

# RON prediction models for the New Fuels and Vehicles Systems

Leanne Whitmore  
Corey M. Hudson  
- Sandia National Labs



# RON Prediction



Goal: To engineer and distribute a high-quality **fully open** software, using **publicly available** resources to predict **fuel properties**.

Stretch: Allow internal (closed source) datasets and tools to be added to the prediction framework.

# Available Training Datasets



## 152 RON Compounds

- Collected from Al-Fahemi et. al (2014), ASTM (1958), Balaban et al. (1992) and Bluock et al. (1995)

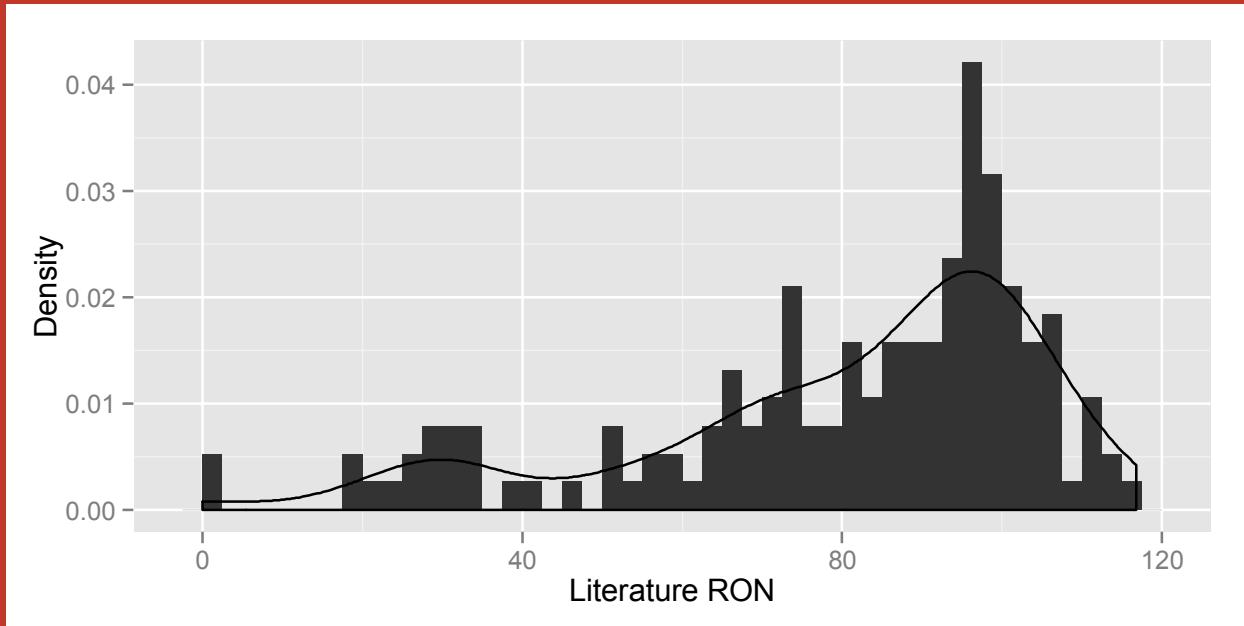
## Available Features (Public)

- NCBI PubChem (881 structural features)
- NCBI Experimental (~20 common features)
- PaDEL-Descriptor (1875 QSAR Descriptors)

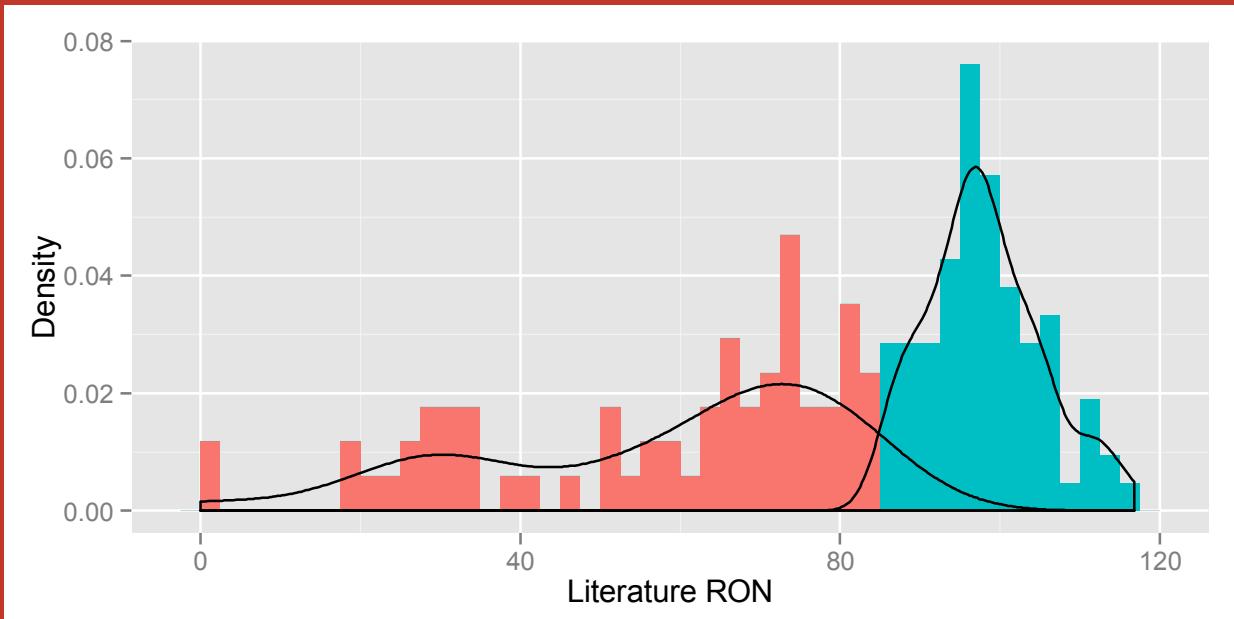
## Private

- Collected from licensed software or data stores (ACD, EPI, etc.)

# Literature RON Distribution



# RON for Compounds of Interest



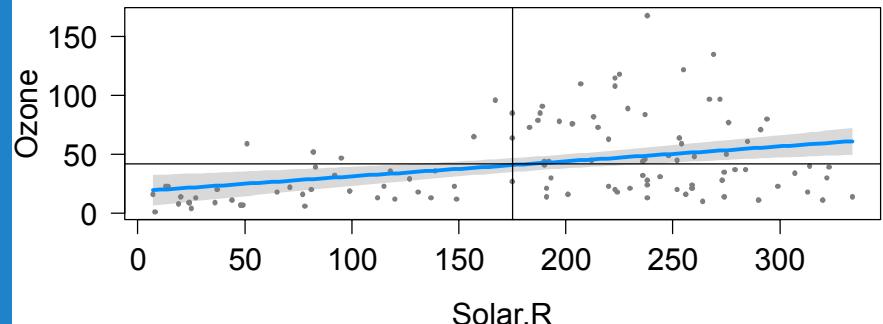
# Classification vs. Regression

- **Machine Learning Classification** dramatically decreases the problem of overprediction.
- **Reason** information content.



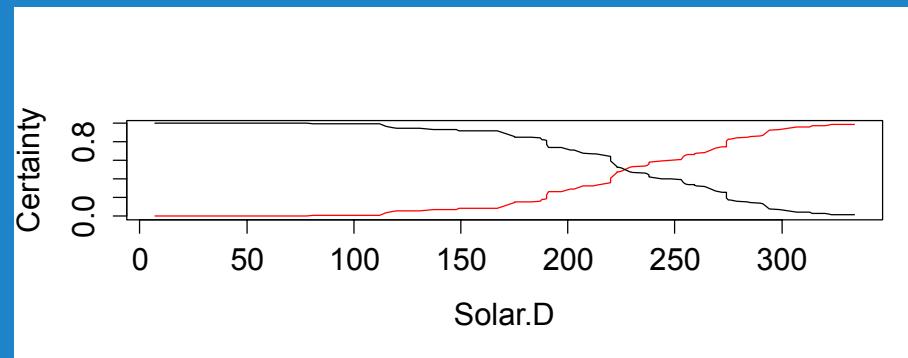
# Regression Uncertainty

Regression residuals are biased toward the highest/lowest values



# Classification Uncertainty

Classification residuals  
are biased toward the  
median values



# RON Classification

- For the purposes of screening, RON is a classification problem.
- High RON Chemicals are useful for drop-in blendstocks in ignition engines.

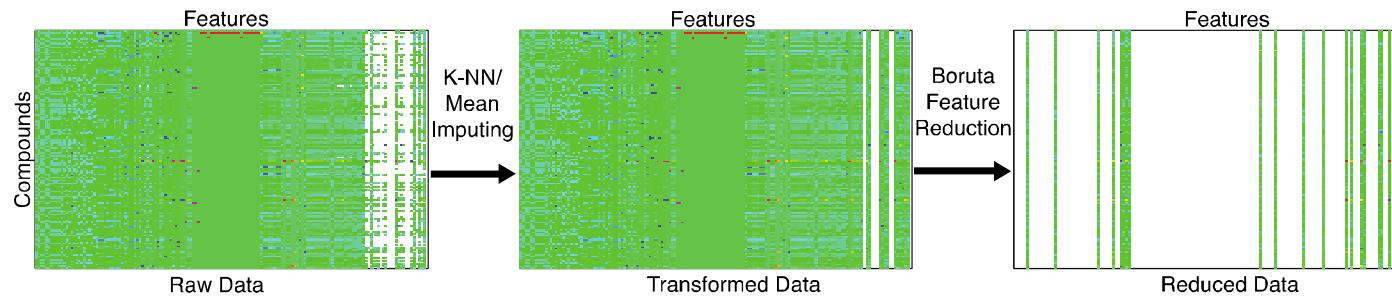


# Machine Learning Methodology

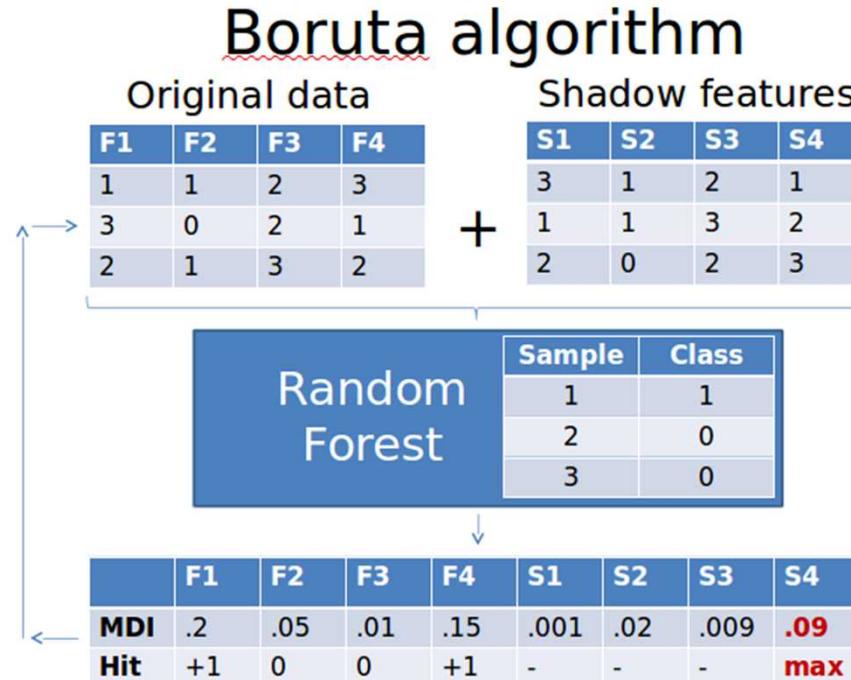


- **Random Forests**
  - Classification
  - Fast
  - Scalable
  - Robust
- **Tanimoto**
  - Clustering
  - General
  - Agnostic to Imputed Features

# Procedures for Feature Selection/Reduction



# Boruta Feature Selection



## Scale of Feature Reduce

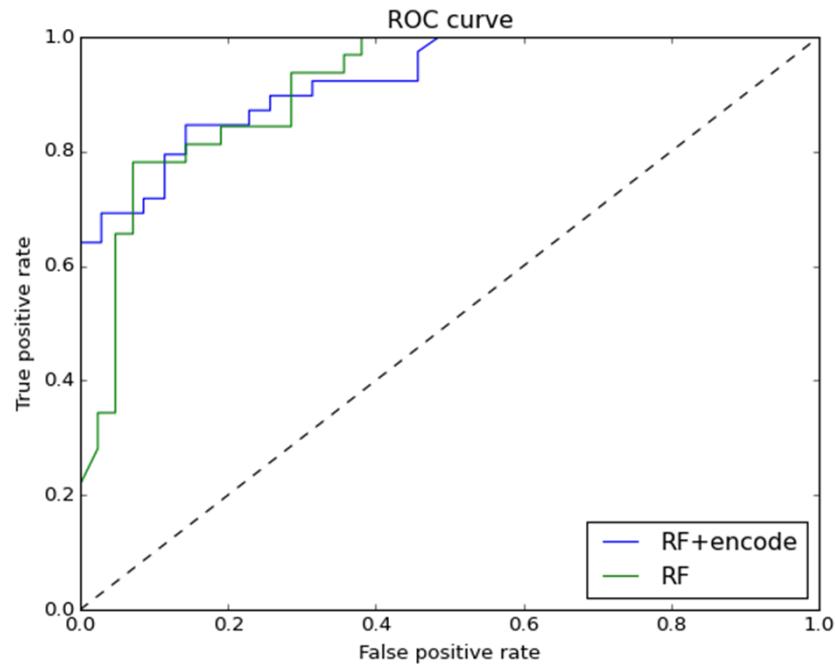
- 926 original features  
(experimental/NCBI/ACD-EPI)
- 147 variable features
- 19 after KNN-Imputation, followed by Boruta  
Features Selection
- Empirical estimates of accuracy improved 4%,  
precision improved 5%

# Performance of classifier

- 100 sub-sampled cross-validations (with 50% leave out)

Metric	Mean value	Std. dev
Accuracy	0.84	0.08
Precision	0.85	0.14
Sensitivity	0.83	0.17
Receiver Operator Characteristic (AUC)	0.93	0.06

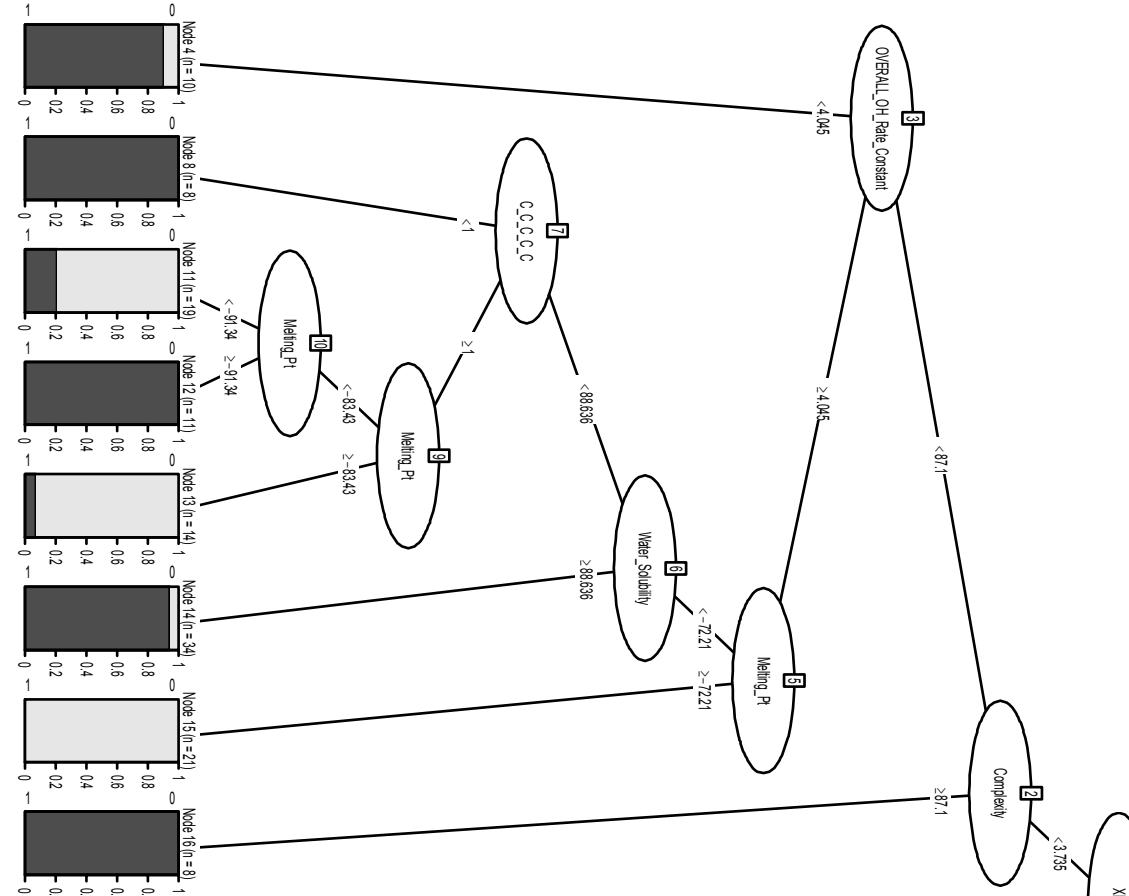
# ROC Curve



# Most Heavily Weighted Features in Random Forest Classifier

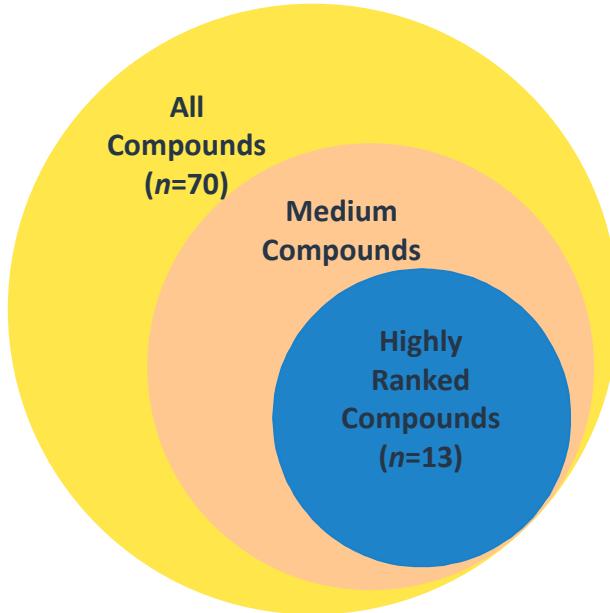
Features	Weight	Type
XLogP3 (Lipidocity)	0.1277	Physical
Log KOA (Air Partitioning)	0.0797	Physical
SMARTS Pattern: C-C-C-C-C-C	0.0781	Structural
Auto-Ignition	0.0772	Physical
Water Solubility	0.0767	Physical
Melting Point	0.0403	Structural
Boiling Point	0.0392	Physical
Surface Tension	0.0324	Physical
OH Rate Constant	0.0288	Physical
Complexity	0.0283	Physical

# Visualizing a Single Decision Tree



# Ranking Compounds of Interest

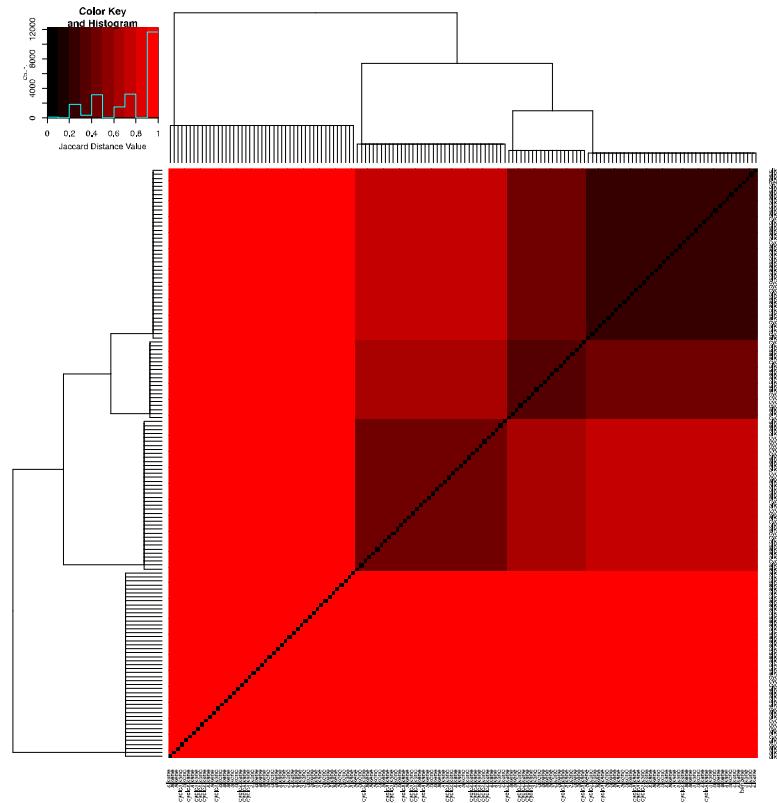
- Initial Ranking for 70 Compounds of Interest



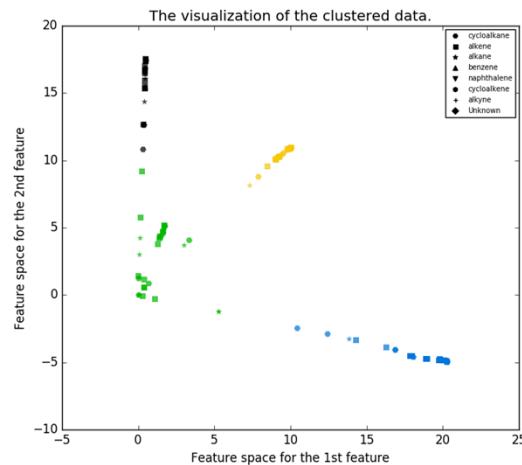
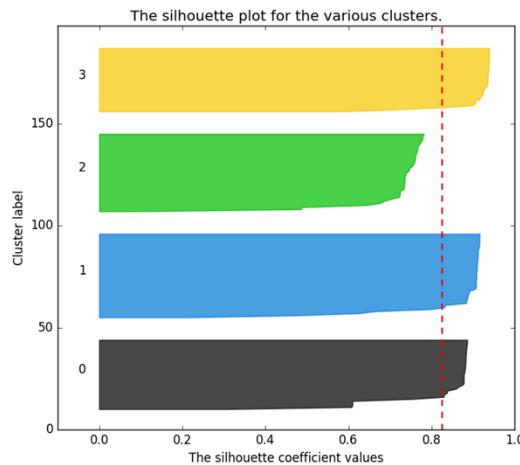
Compounds predicted to have RON > 85 in order of probability in class

1. Isooctane	8. Isoprenol
2. Methylcyclopentane	9. Isobutanol
3. Ethanol	10. 3 methyl 1 butanol
4. Methyl butyrate	11. Butyl acetate
5. Ethyl isobutyrate	12. Toluene
6. Methyl 2-methylbutyrate	13. Isoamyl acetate
7. 2-methyl 2 butanol	

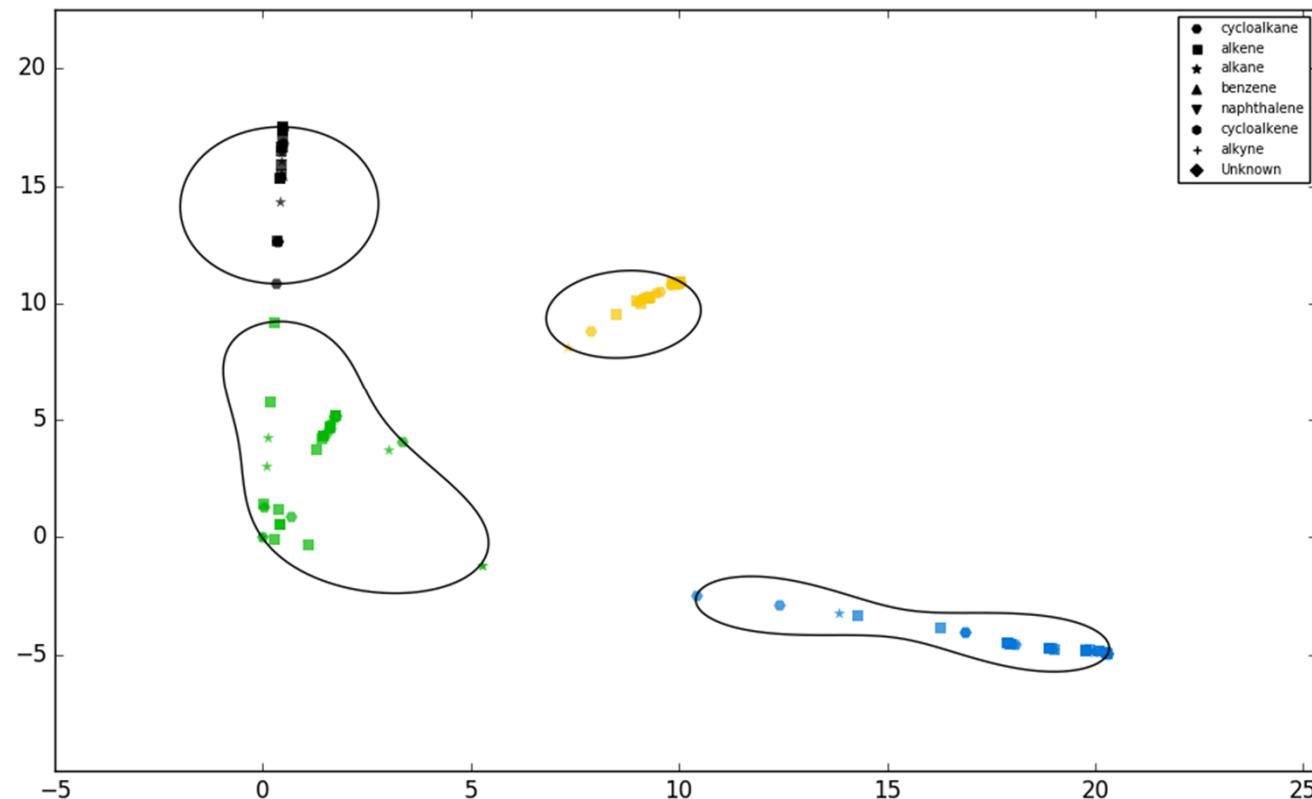
# Challenge in Generalizing Approach



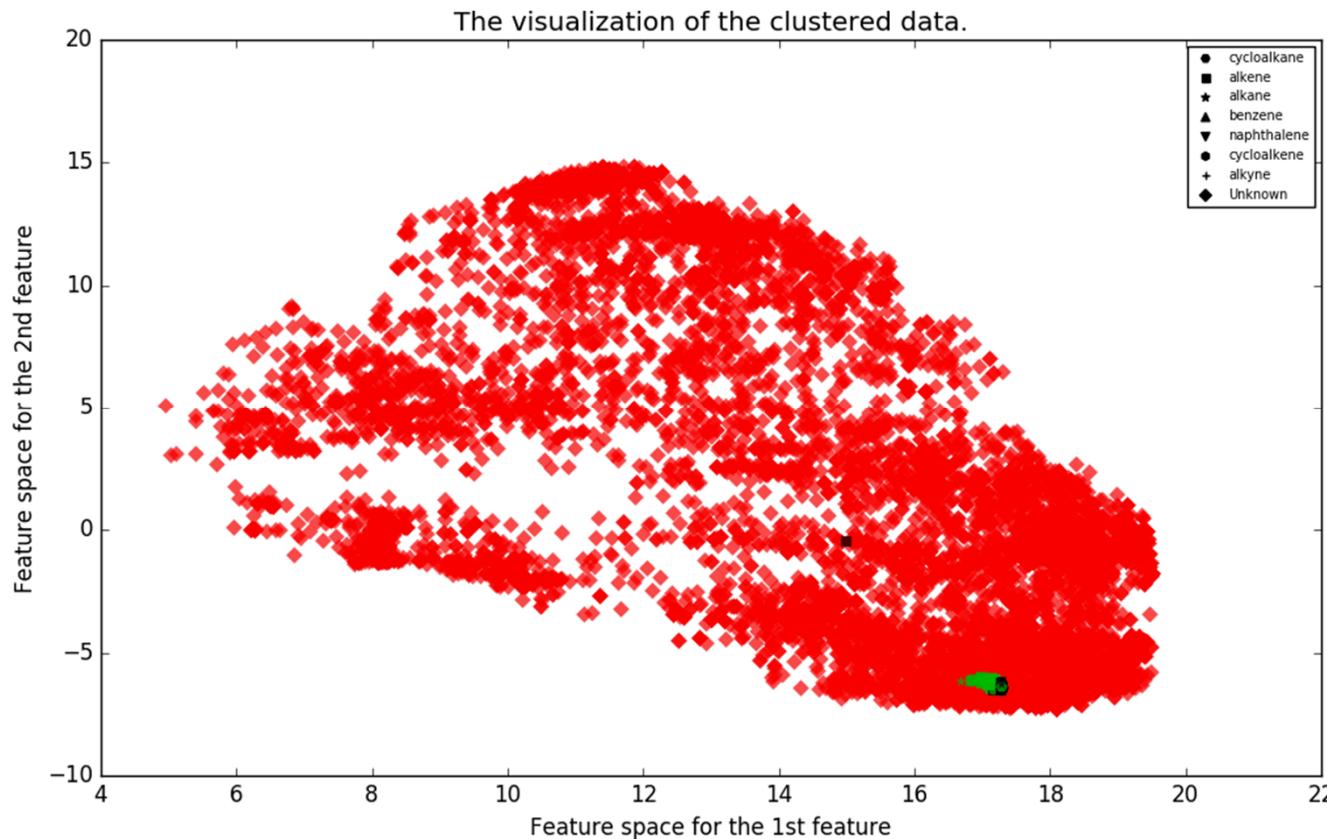
# Clustering Training Data



# Creating Clusters using Training Hydrocarbons



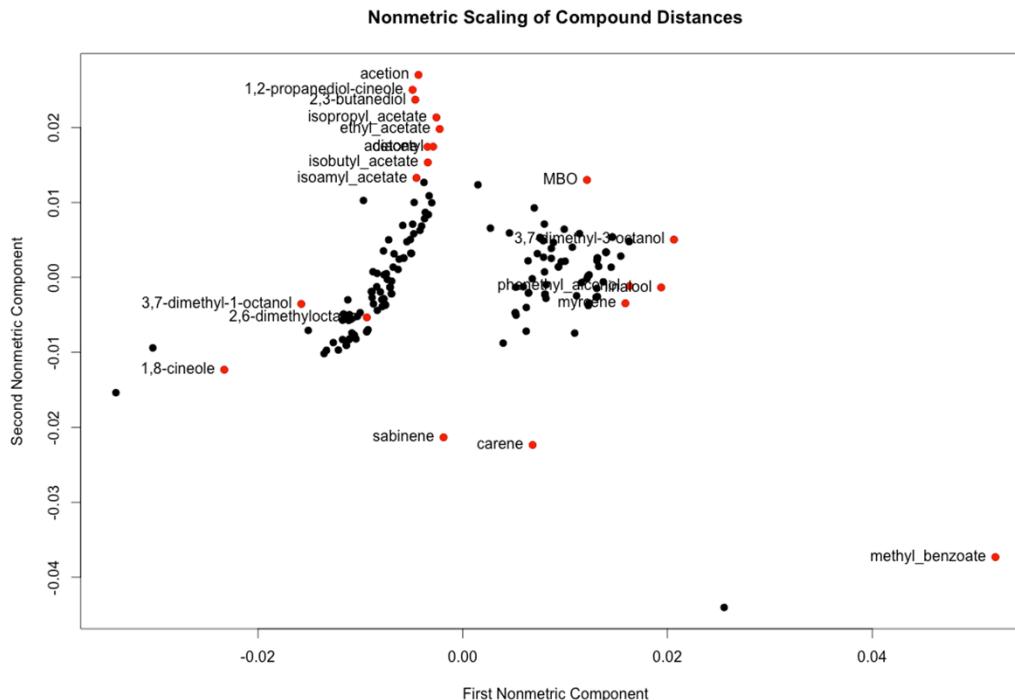
# Comparison with Total Dataset



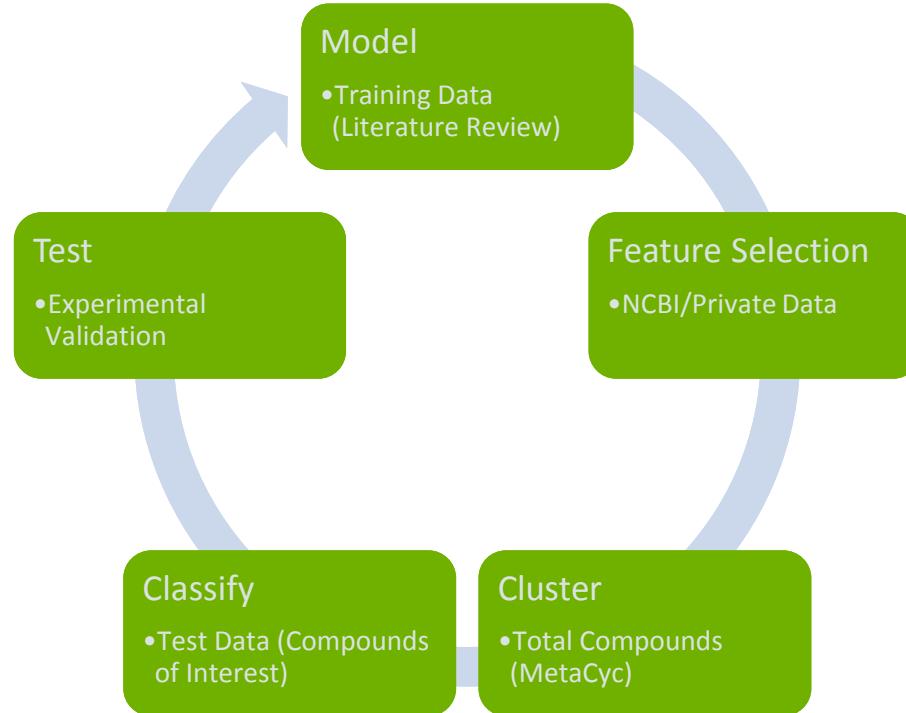
# Structure of Software



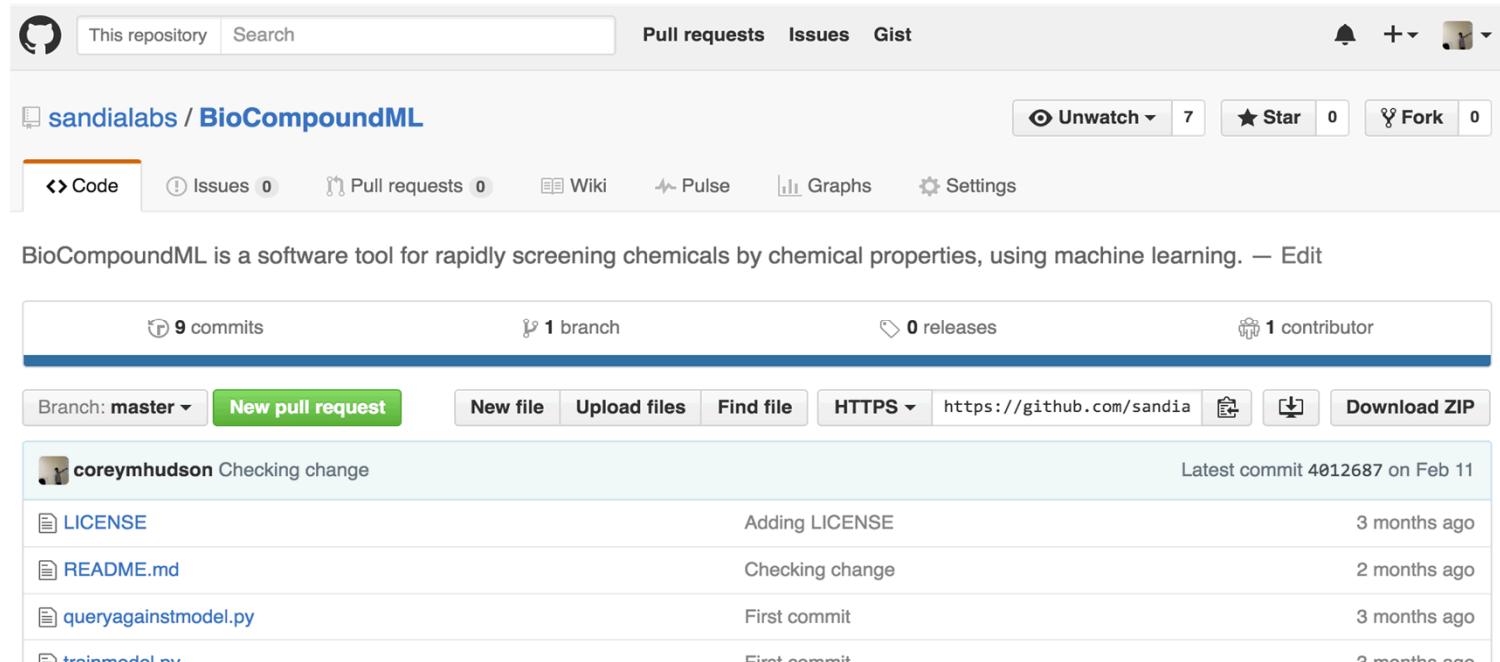
# Testing Model with Measured RON Values from SWRI



# Machine Learning Process



# Distribution of Software



This screenshot shows a GitHub repository page for `sandialabs / BioCompoundML`. The page includes a navigation bar with links for Pull requests, Issues, and Gist, and a header with a search bar, a GitHub icon, and user profile links. Below the header, there are buttons for Unwatch (7), Star (0), and Fork (0). The main content area displays the repository's description: "BioCompoundML is a software tool for rapidly screening chemicals by chemical properties, using machine learning." It also shows statistics: 9 commits, 1 branch, 0 releases, and 1 contributor. A "New pull request" button is highlighted in green. The commit history lists five commits from `coreymhudson`, including adding a LICENSE file and first commits for `queryagainstmodel.py` and `trainmodel.py`. The repository URL is <https://github.com/sandia/BioCompoundML>.

**Code** Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

BioCompoundML is a software tool for rapidly screening chemicals by chemical properties, using machine learning. — Edit

9 commits 1 branch 0 releases 1 contributor

Branch: master New pull request New file Upload files Find file HTTPS https://github.com/sandia/ README.md

coreymhudson	Checking change	Latest commit 4012687 on Feb 11
LICENSE	Adding LICENSE	3 months ago
README.md	Checking change	2 months ago
queryagainstmodel.py	First commit	3 months ago
trainmodel.py	First commit	3 months ago

## BioCompoundML

This software implements Random Forest machine learning algorithms to predict desired chemical properties given