This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2018-7883C

Exceptional service in the national interest

Sandia National Laboratories

Discovering Novel Biofuels Using Machine Learning Software BiocompoundML

Michelle Tse, Leanne Whitmore & Corey Hudson

Abstract

Cars, planes, and trains are an integral part of our lives—however, they are fueled by petroleum which is bad for the environment, is becoming increasingly expensive, and the amount of which will be heavily depleted within the next few decades. There is an increasing trend of hybrid and electric cars; however, hybrid cars still require gasoline, and electric cars are not currently long-term viable options due to the short life of lithium ion batteries, typically low ranges (miles traveled on one charge), and long charging times. As such, we are attempting to find a suitable alternative fuel—a biofuel—which can reduce the amount of petroleum. Using the machine learning techniques we are screening molecules to find ones with desirable fuel properties: molecules with a high RON (research octane number, i.e. the anti-knock performance), high MON (motor octane number, normally for engines that run at higher speeds), and low surface tension. Currently there is no easy method to predict or measure RON, MON, and surface tension values that is not expensive or time consuming. Therefore, to rapidly predict these properties for a large set of chemical compounds we are using BioCompoundML which was developed at Sandia National Laboratories. BioCompoundML uses random forest classifiers to classify whether a compound has a high or low RON, MON or surface tension values. Cheminformatic features such as chemical fingerprints, melting point and density are used as features. Training data (compounds with known RON, MON or surface tension values) is necessary to predict the following properties for other compounds. As such, we have pulled data from the fuel property databases to obtain the necessary training data for BioCompoundML. Then we predicted RON, MON, and surface tension properties for all compounds in the MetaCyc database. By identifying compounds that have high RON, MON and low surface tension values we can help further development of new biofuels.

Introduction

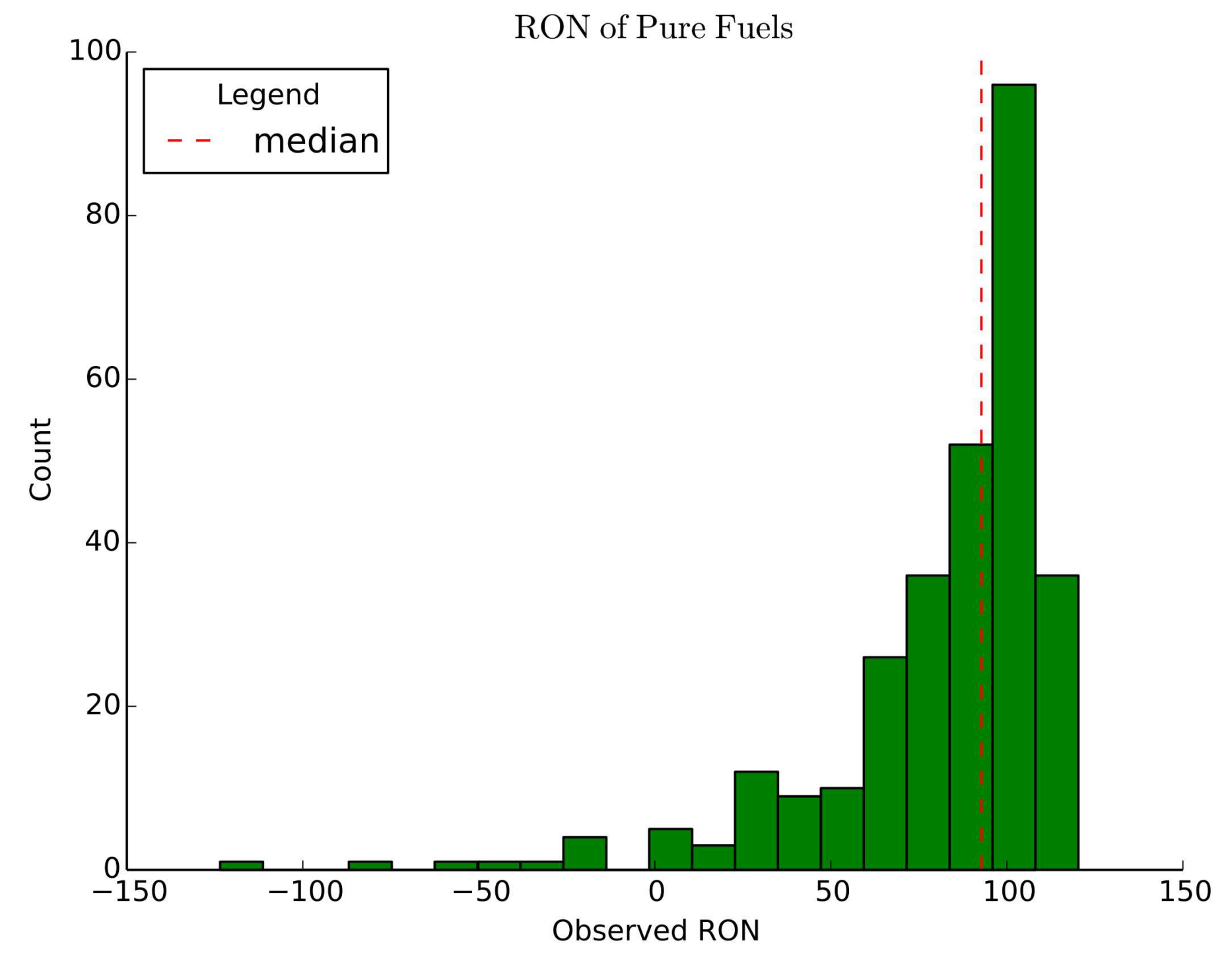
RON, MON, and surface tension are the three fuel properties we are primarily focusing on. RON, the research octane number, indicates a compound's ability to resist ignition under pressure and subsequently a compound's resistance to engine knocking. Knocking is pre-ignition of fuel in the engine and ultimately results in engine inefficiency and can also damage the engine. The higher the RON value indicates better anti-knock performance and is normally a measure of engines that run at slower speeds. On the other hand, MON (the motor octane number) is a measure of a compound's resistance to knocking in engines that run at higher speeds. Similarly, to RON, a higher MON value indicates a more robust resistance to ignition under pressure and is also desired as a high number. Surface tension is another property important for determining if a compound will be a good candidate for a biofuel. Having a lower surface tension allows the fuel to be distributed to the compression cylinder in the engine evenly. Using BioCompoundML we can rapidly predict these properties for a large number of compounds and therefore quickly identify compounds that would make a good fuel additive.

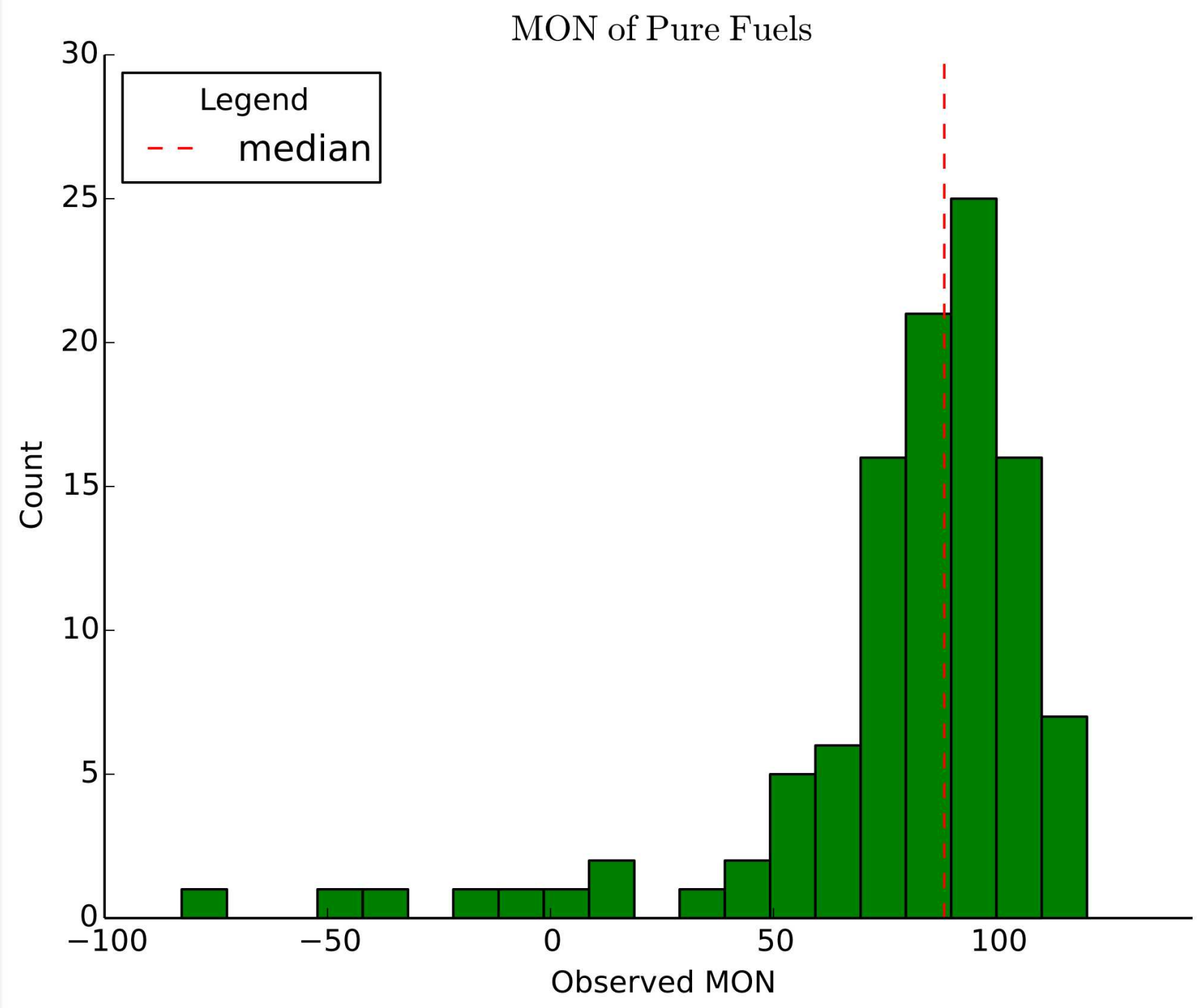
BioCompoundML Workflow:

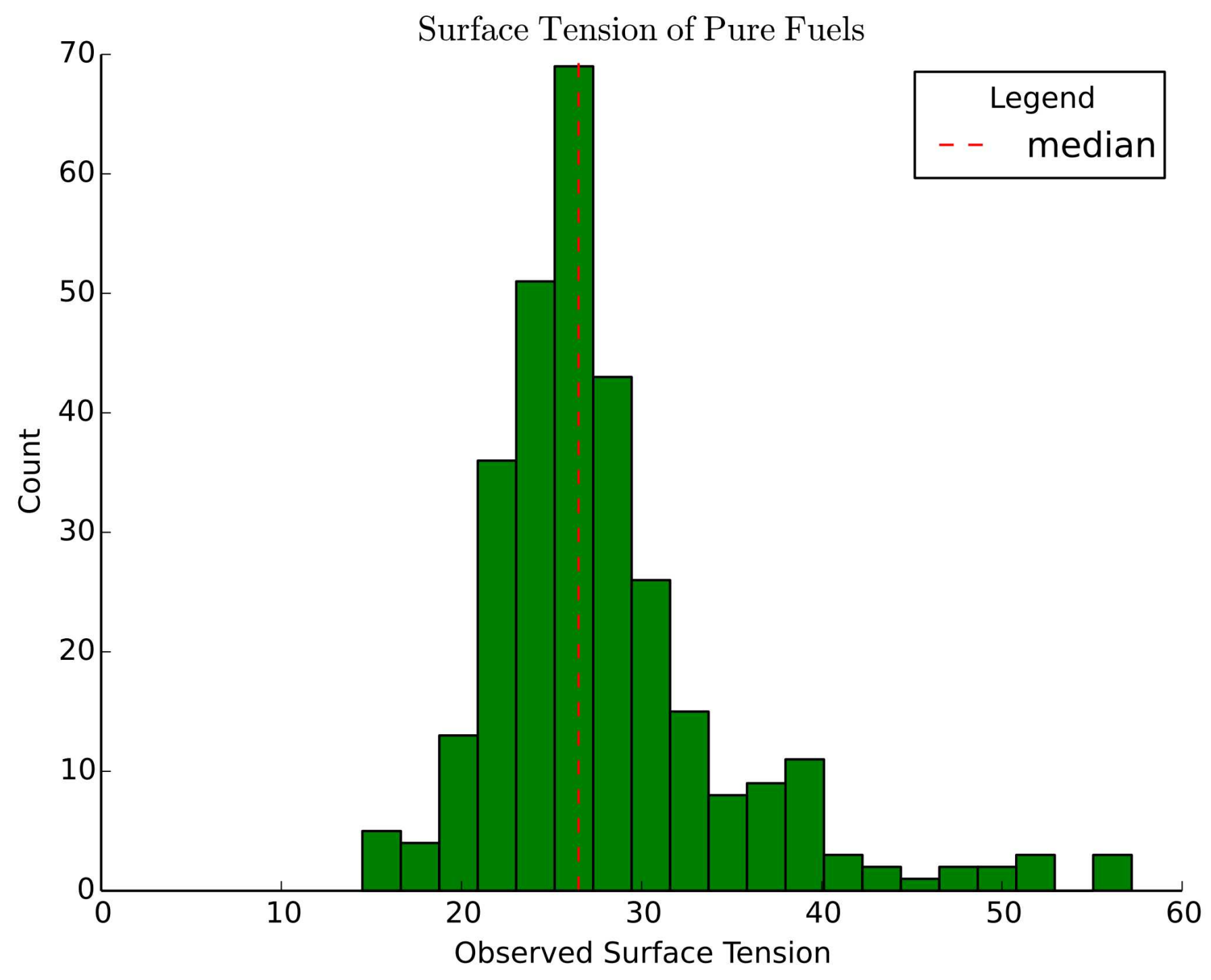
1. Run the (training) data through BioCompoundML[3].
2. BioCompoundML uses random forest classification to create a training model.
3. Run testing compounds through BioCompoundML using the model to find the probability of the compound is classified as having a high (above training data median property value) or low property value (below training data median property value).

Retrieving Training Data

Training data is retrieved from the Co-Optimization of Fuels & Engines Project[1] fuel database; the data (such as RON, MON, and surface tension) is extracted from the databases and saved as tab delimited files.

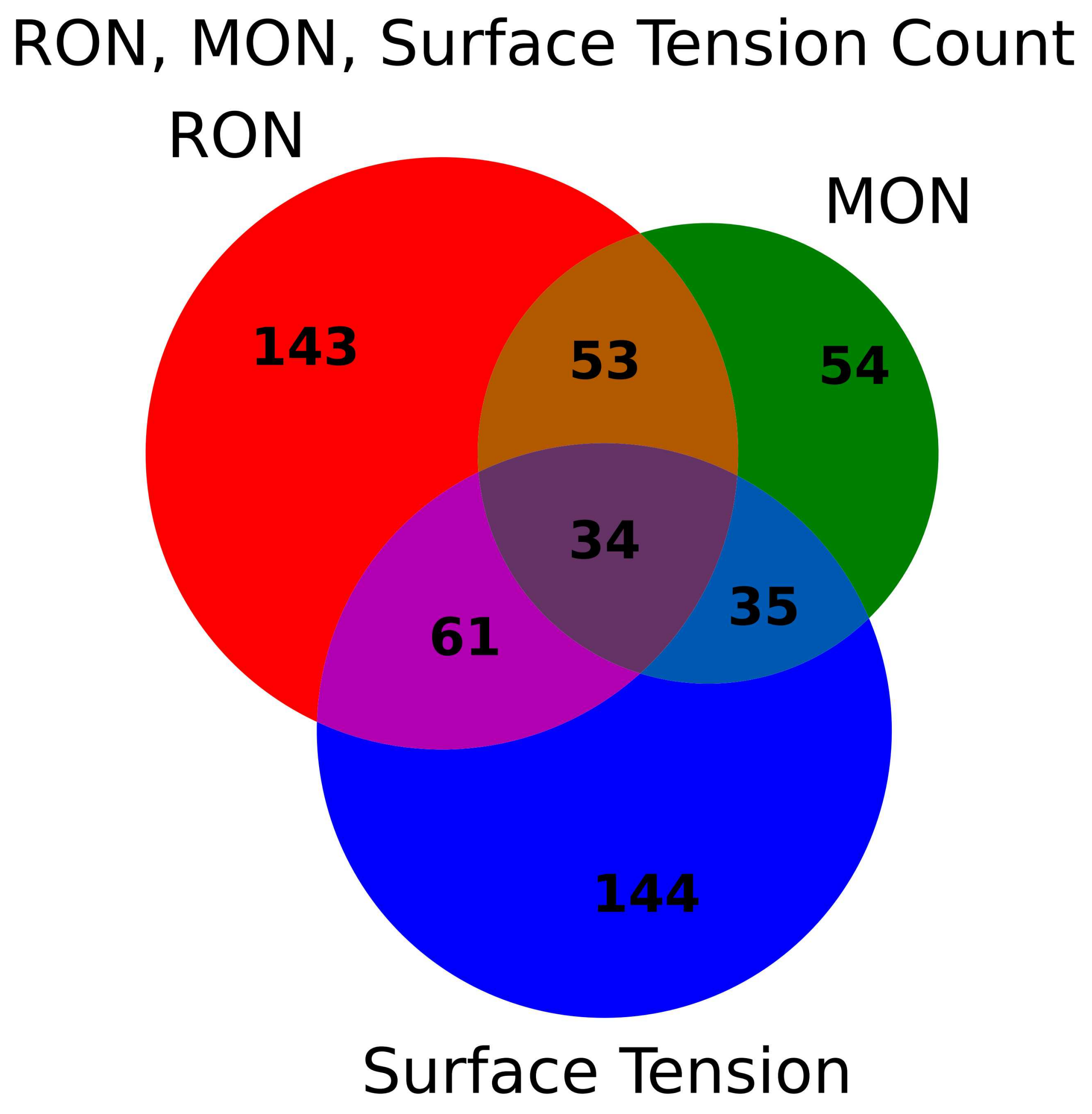






Predicting Fuel Properties

Testing compound data is retrieved from the MetaCyc[2] database, then the properties of the compounds are predicted by running the data through BioCompoundML.



Conclusions

In total, there are 34 compounds with a high value in RON and MON, and a low value in surface tension, making them good fuel additive candidates; a few of these compounds are 2-butanol, 2-pentanol, diethyl ketone, furan, and methyl isobutyl ketone. BioCompoundML does not predict fuel properties of compounds with 100% accuracy, however, the accuracy is high by using training models (from training data). Furthermore, more fuel property/compound data is needed.

References

1. *FileMaker WebDirect*, https://fuelsdb.nrel.gov/fmi/webd/FuelEngineCoOptimization.
2. *MetaCyc Metabolic Pathway Database*. https://metacyc.org/.
3. Whitmore, Leanne S., et al. "*BioCompoundML: a general biofuel property screening tool for biological molecules using Random Forest Classifiers." Energy & Fuels 30.10 (2016): 8410-8418.*

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.