Project 033 Alternative Fuels Test Database Library (Year VI)

University of Illinois Urbana-Champaign

Project Lead Investigator

Tonghun Lee
Professor
Mechanical Science & Engineering
University of Illinois at Urbana-Champaign
1206 W. Green St.
Urbana, IL 61801
517-290-8005
tonghun@illinois.edu

University Participants

University of Illinois at Urbana-Champaign

- PI: Tonghun Lee, Professor
- FAA Award Number: 13-C-AJFE-UI-026
- Period of Performance: October 1, 2019 to September 30, 2020
- Tasks:
    1. Generation II Online Database Update and JETSCREEN Connection.
    2. Machine Learning-Based Online Analysis,

Project Funding Level

FAA funding Level: $130,000
Cost Share: Software license support from Reaction Design (ANSYS)

Investigation Team

- Tonghun Lee (Professor, University of Illinois at Urbana-Champaign): Overall research supervision.
- Isabel Anderson (Graduate Student, University of Illinois at Urbana-Champaign): Database development and Machine Learning-Based Analysis.

Project Overview

This study seeks to develop a comprehensive and foundational database of current and emerging alternative jet fuels by integrating relevant pre-existing jet fuel data into a common archive that can support scientific research, enhance operational safety, and provide guidelines for the design and certification of new jet fuels. In previous years of this project, efforts were focused on the integration and analysis of pre-existing jet fuel data from various government agencies and individual research groups. Recently, we have converted all of the compiled data to a new nonstructured query language (NoSQL) format using a JavaScript object notation (JSON) schema, thus allowing the data to be analyzed in a flexible manner using various programming languages. To this end, we have launched the second generation of our online database, which utilizes the new nonrelational database structure. This version is equipped with interactive analysis functions for users and flexible methods for plotting and downloading data. In the previous year, we have extended this effort to incorporate advanced machine learning algorithms in the analysis process. Additionally, we have worked on integrating our database with the database assembled by the European JETSCREEN program, potentially leading to a global database structure in the future. We hope that the database will one day not only serve as a comprehensive and centralized knowledge base utilized by the jet fuel research community, but will also serve as a resource that can enhance global operation efficiency and safety. Future efforts will include not only expansion of the international framework with JETSCREEN, but also efforts to potentially include

real-time data being used at the airports. With the prolific diversification of new alternative jet fuels expected in the near future, the ability to track critical fuel properties and test data from both research and operation perspectives will be highly valuable for the future of commercial aviation.

Task 1- Generation II Online Database Update and JETSCREEN Connection

University of Illinois Urbana-Champaign

Objectives

The main objective of this Task is to upgrade/debug the generation II online National Alternative Jet Fuels Test Database functions and link the database to the European JETSCREEN program. The generation II database is designed using a new architecture that allows for flexible analysis and scaling based on a NoSQL data format. This format can accommodate various data types and that can be easily accessed by any common programming language, and basic analysis functions have been built right into the web interface. Following the launch of the generation II web interface, significant effort has been made in the past year to upgrade the functionalities and address bugs based on user feedback. We have also converted much of the data to a comma-separated values (CSV) format to enable machine learning-based analysis in the future, of which more will be discussed in Task 2. The specific goals in Task 1 are as follows:

- Test and improve functionality of the generation II online web interface and database structure.
- Convert dataset from nonrelational JSON (Schema) format to CSV for machine learning-based analysis.
- Link database with the European JETSCREEN program with automatic file sharing.
- Link database with real-time airport fuels data (delayed due to COVID-19, efforts restarted as of September 2020).

Research Approach

Generation II Database Debugging and Upgrade

A beta version of the generation II database was launched online in the summer of 2019. The web interface of the generation II database is shown in Figure 1. All of the functionality of the previous database is maintained, and the security login features have been migrated from the previous version. The generation II web interface, much like generation I, is a HTML-oriented program that is built on a layer of metadata which supports search functions for the users. The tree structure that was applied to organize the data folders in the first database was also retained in this version, allowing the user to access the data in a similar manner. The main difference is that there is an additional inner core which houses the JSON files, and it is here where the test data resides. Currently, the database has grown to house over 25,000 separate fuel records.

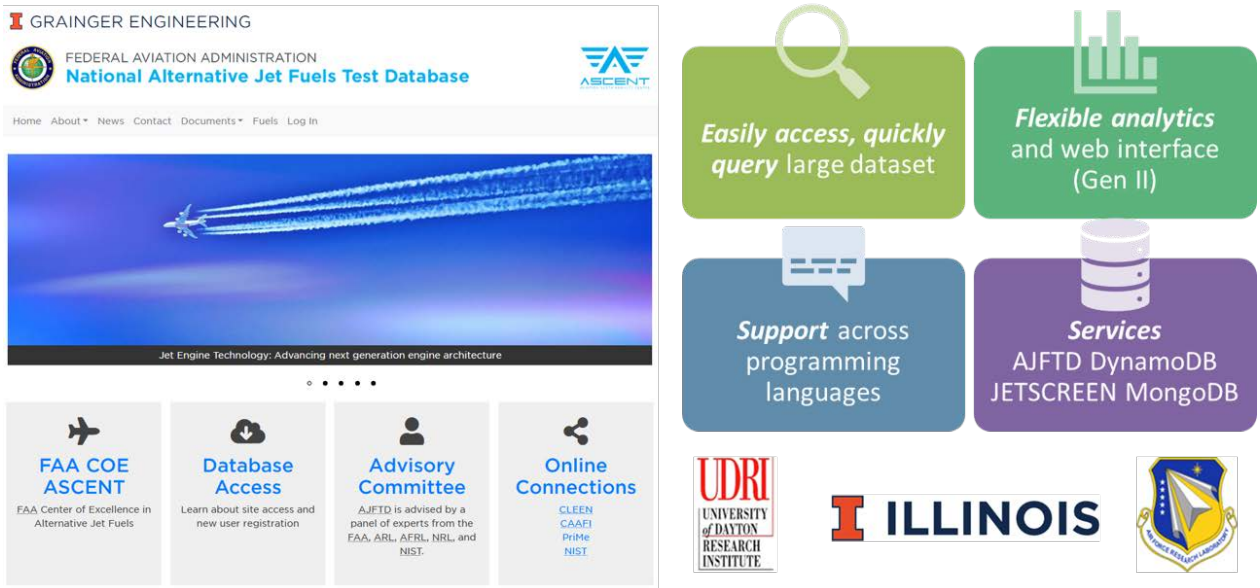

Figure 1. Generation II National Alternative Jet Fuels Test Database web interface. (altjetfuels.illinois.edu)

The catalogue of data currently available in the database is primarily assembled from four separate sources. The fuels with POSF (Air Force Research Laboratory (AFRL) fuel database code) number designations were added from the internal database maintained by the AFRL at the Wright Patterson Air Force Base. The second dataset was obtained from the PQIS reports of the Naval Air Systems Command (NAVAIR) and corresponds to a compilation of fuel data geared primarily towards government use. The third set was provided by Metron Aviation, who compiled fuel properties from samples collected at airports through a previous ASCENT project. The dataset resulting from this study proved valuable by providing a landscape of fuels currently used in commercial aviation and will guide our future efforts focused on capturing this type of data in real time. The final dataset was obtained from the National Jet Fuel Combustion Program (NJFCP) within ASCENT.

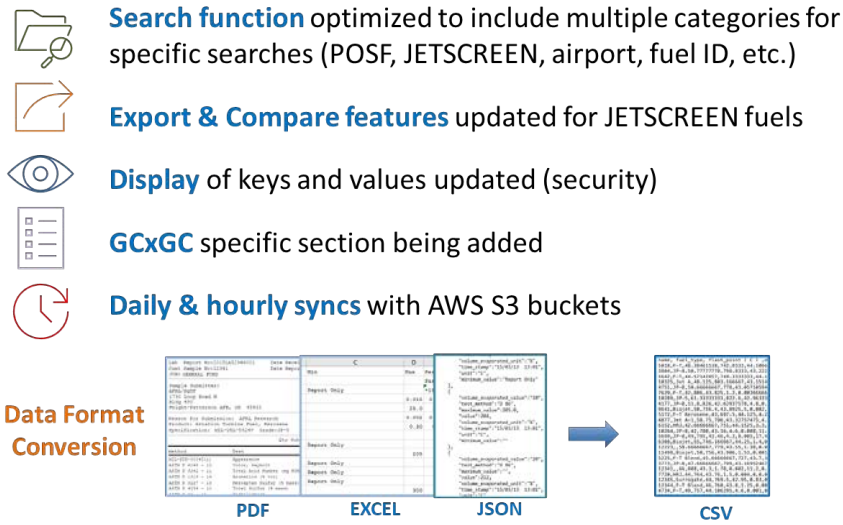

Figure 2. Modifications to the Generation-II Database and conversion of data to CSV for machine learning.

Following the launch of the generation II database (summer 2019), significant effort was required to fix bugs and upgrade various aspects of the database. Some of the key changes to the database are listed in Figure 2 and summarized below.

- During the integration of the database with JETSCREEN, we modified the labeling structure to ensure that files were coded as JETSCREEN data and separately searchable. Similar to how the Metron data is labeled according to the airport the data was retrieved from, JETSCREEN was added as a separate search label also. This search filter can be integrated with additional search filters to allow users to view tests for a specific fuel type from the JETSCREEN group if so desired. The search page was also updated to include options to search by POSF number and GCxGC data. Searching by POSF number allows the user to find not only the specific POSF fuel they are looking for, but also any fuels that include it in a blend. Searching "GCxGC" in this search bar also returns any files that contain "GCxGC" in the fuel description. The user can also combine this search with a "Search by Fuel Type" or "Search by Airport" to narrow the results. This function is still being optimized based on user experience.

- After the search page was updated to include JETSCREEN files, the Export and Compare features required updates as well. Although we had worked with JETSCREEN extensively to create a standard JSON format, the files generated from the two camps had minor differences which caused the current code on the database to fail at times. Slight differences between the JETSCREEN files themselves were also causing errors not only for comparing the data with FAA files, but also with other JETSCREEN files. The Export and Compare features were updated to work around these issues and to support the comparing of all test files on the database.

- The display of data on the database was also changed to allow for more privacy and security. Authors of the files (which included student names) were removed from the JETSCREEN display. The sharing function, which was put in place to share selected FAA data with JETSCREEN via Amazon Web Services (AWS), was set to display for admin accounts only. These

shared FAA files sync with the AWS S3 bucket every hour. The JETSCREEN bucket on AWS is checked each day for new files, which are then downloaded to the website (more detail will be provided in Figure 3).

- Effort was made to convert the JSON format on the database to a CSV format for select files so that we could utilize machine learning-based analysis, which will be addressed further in Task 2. The actual files that are being stored will utilize the NoSQL JSON format, which is more conducive to maintaining a flexible database. However, certain parts of the data that are to be analyzed using machine learning will need to be converted to CSV format for which multiple Python based machine learning scripts are available. In the future, there may need to be a process to automate this conversion in real-time for when it is needed.

Integration of Database with JETSCREEN

During the past year, we have made significant progress in integrating our database with the European JETSCREEN program. The JETSCREEN program was initiated to provide fuel producers, air framers, and aero-engine and fuel system original equipment manufacturers (OEMs) with knowledge-based screening tools for fuels and also have a similar database that could be linked with ours. We first started discussing a potential merger with the JETSCREEN database in 2018, after which we started methodically synchronizing the data structure so that a merger could be possible. After much beta testing, the two databases were first linked in Spring of 2020, and the data sharing process is shown in Figure 3.

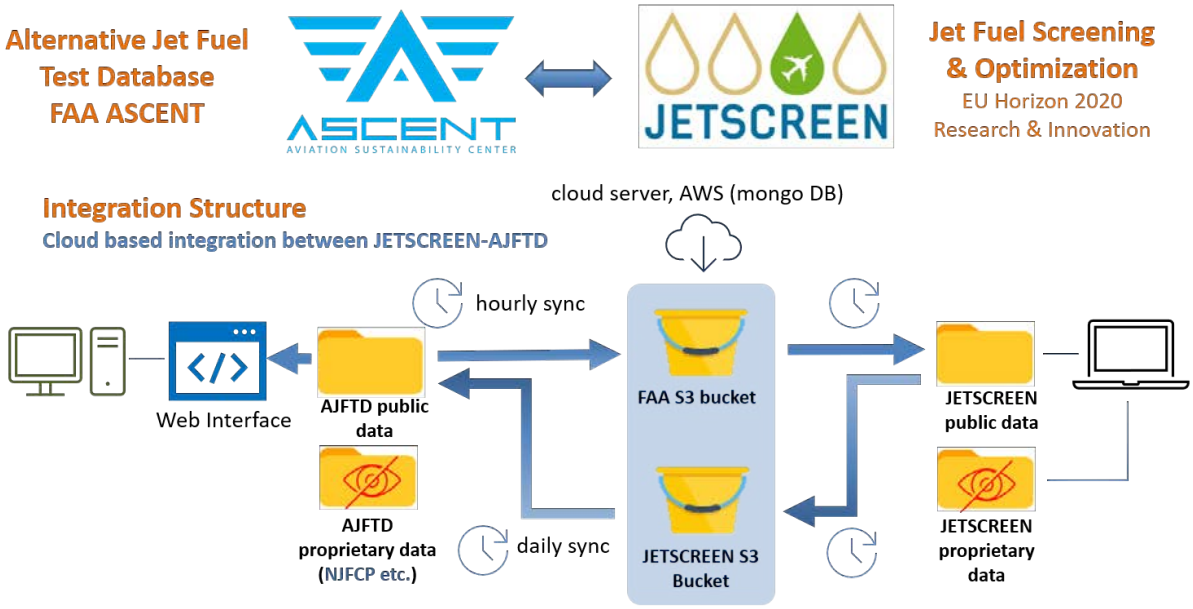

Figure 3. Database integration schematic with JETSCREEN.

As shown, the JETSCREEN and FAA databases are joined by a common cloud storage. AWS was selected as the server to store the shared data, mainly due to their affiliation with the University of Illinois. S3 buckets (Amazon database structure) were created for both FAA and JETSCREEN to share their JSON files. Each can pull files from the other's folder, but read and write access is only granted for the owners of the bucket. The FAA data is shared to its S3 bucket via altjetfuels.illinois.edu. All public FAA data on the website will have an option to be shared with JETSCREEN, which can be toggled by administrators. The website syncs hourly with the bucket to upload newly shared data. No proprietary data is shared to the FAA S3 bucket. Any files uploaded to the FAA bucket can be viewed and downloaded by JETSCREEN. For downloading new JETSFCREEN data to the website, a script runs daily to check JETSCREEN's S3 bucket for newly shared data. Any new files are then downloaded to our local database and can be consumed by the users. We note that the actual interface of the database will be left to each entity to decide. For us, we have adopted an open web interface, whereas JETSCREEN has a proprietary software with no direct web access currently.

We feel that this is a monumental first step in linking many other fuel databases across the globe in the future. From this joint effort between FAA and JETSCREEN, we hope to establish a foundation which can help to both monitor and evaluate fuels used around the international airspace in the future. As new fuels are integrated into the global supply chain, a means to keep track of their properties will become critical. Such an interconnected database will ensure that we are able to provide both the information needed for research and certification of new fuels, but also to uphold quality standards and demonstrated the feasibility of sustainable aviation fuels (SAFs) in our future. The database integration impacts are outlined in Figure 4.

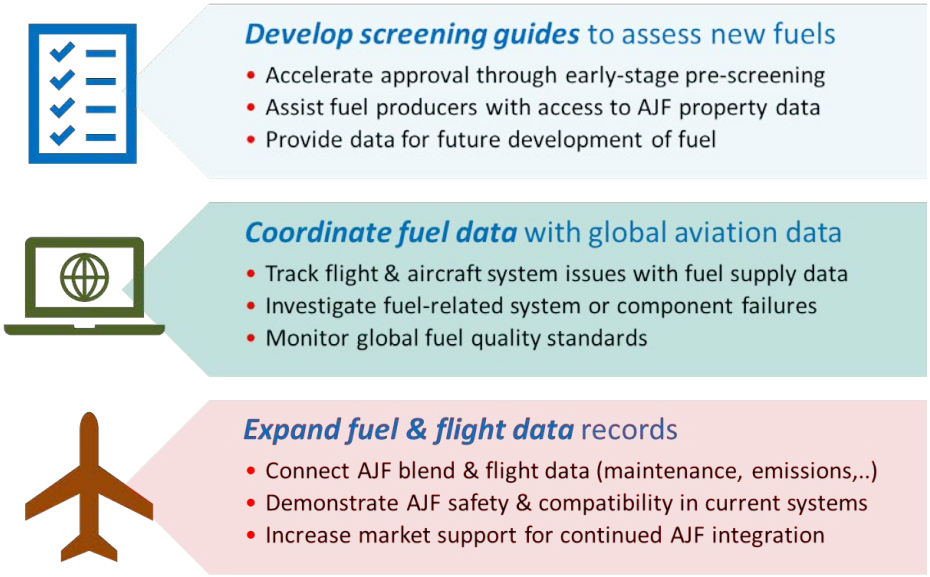

Figure 4. Database integration impacts.

Milestones

3 months

- Initiation of debugging and optimization of data structure in the generation II database.
- Collaboration with JETSCREEN to standardize the data format for a potential merger in the near future.

6 months

- Completion of most debugging in the generation II database and further improvements to online analysis tools.
- Writing of scripts for integration of database with JETSCREEN on cloud server and selection of provider (Amazon).
- Conversion of data from NoSQL to additional CSV format.

9 months

- Launch of the joint cloud server with JETSCREEN and turning on of data sharing scripts.
- Continued discussions with JETSCREEN to optimize data sharing protocols and set security protocols.

12 months

- Preliminary tests with JETSCREEN for optimization of file sharing and identification of problems.
- Modification of Compare, Export, and other functions to accommodate integration of JETSCREEN data.

Major Accomplishments

Launching of the Integrated Database with JETSCREEN

We have finally merged our FAA database with the JETSCREEN database at https://altjetfuels.illinois.edu/. The database is linked with JETSCREEN through a cloud archive (AWS) where FAA and JETSCREEN data are kept in separate secure online folders. Data is automatically shared and downloaded to the local servers for both us and JETSCREEN several times each day through the execution of customized scripts. On our side, new JETSCREEN data will be pulled into our local database each

day and available for processing through our regular analysis tools. This new database structure can be a foundation for a global database in the future, contributing to the development and certification of new fuels in the global aviation pipeline as well as monitoring fuel quality and safety concerns.

Modifications to the Generation II Online Database and Conversion of Data to CSV
Upon launching of the generation II database and migrating more than 25,000 fuel records in a NoSQL JSON format, the web interface and analysis tools have been rigorously tested and debugged. These improvements will continue in the future as new data are added and new analysis techniques are developed. In anticipation of using machine learning-based analysis in the future, parts of the database have been converted to a CSV format so that Python-based machine learning scripts could be utilized. Significant upgrades to the search functions have been carried out so that users can search for specific fuel type and other cross-referenced properties directly online.

Publications
N/A

Outreach Efforts
Database made accessible through https://altjetfuels.illinois.edu/

Awards
N/A

Student Involvement
This project was primarily conducted by one graduate student (Isabel Anderson).

Plans for Next Period
In the next period, we intend to optimize the integration with JETSCREEN as well as develop common online analysis tools based on "big data" analysis. New discussions regarding management of proprietary data will need to take place with JETSCREEN. Most importantly, we will need to engage with airports in the U.S. to capture real-time fuel data and determine the feasibility of integrating it into our current database.

Task 2 – Machine Learning-Based Analysis
University of Illinois Urbana-Champaign

Objectives
The main objective of this Task was to develop advanced analysis methods based on machine learning algorithms for analysis of the data in the alternative jet fuel database. The effort is inspired by the notion that the intricate relations between properties of fuels and their chemical signatures are critical, but maybe beyond the complexity that can be addressed with routine, classical, regression-based analysis. The effort is ever more important when new analysis techniques such as GCxGC can provide large amounts of data that are difficult to process using simple analytical algorithms. Machine learning can provide the means for the most advanced analysis to be applied to our current data and will prove to be even more powerful as the size of the data grows in the future. This effort was also established through a series of discussions with the JETSCREEN team and both programs will devote considerable effort to this cause. The major goals of this Task are as follows:

- Identify best machine learning-based approach for the jet fuel data.
- Identify the best data format for implementation of machine learning algorithms.
- Carry out binary regression analysis of key jet fuel properties.
- Conduct prediction analysis of jet fuel properties based on classical machine learning.
- Conduct prediction analysis of jet fuel properties based on neural network (deep learning) machine learning.

Research Approach
Classical Regression-Based Analysis
The need to utilize basic regression analysis came from joint collaborative work with JETSCREEN during a phase when we were trying to synchronize the data scripts to merge the two databases. A key insight was that many properties of the fuels are correlated and having a strong understanding of such correlations would help reduce the number of testing procedures

for certification. Naturally, our interests extend to understanding if this type of correlation extends to alternative fuel sources with different chemical makeups. On the JETSCREEN side, a separate effort was also initiated in looking into the correlation between different fuel properties; the results were presented at the International Association for Stability, Handling and Use of Liquid Fuels (IASH) meeting in 2019.

On our side, we approached this in two steps. The first was to carry out an extensive binary correlation to understand the relationship between all the major properties that we had access to. The second step was to utilize the correlation information to test out classical regression-based machine learning algorithms to see if we could predict certain properties based on other properties through a training algorithm. The two approaches are shown in Figure 5.

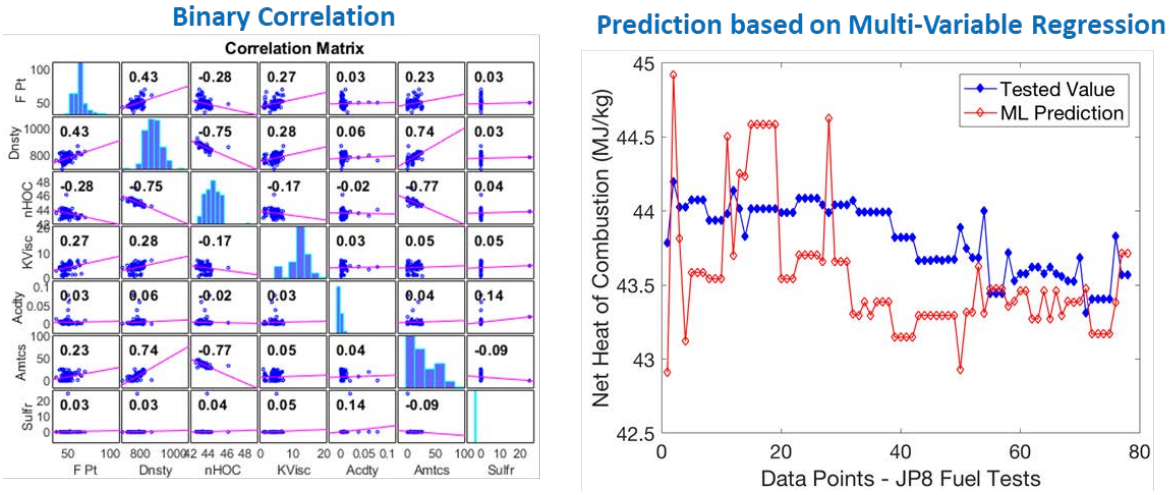

Figure 5. Classical regression-based machine learning.

In order to extract the relationships between the jet fuel properties, the Pearson Coefficients for several properties were investigated using MATLAB's Corrplot tool. The binary correlation matrix in Figure 5 shows the Pearson Coefficients for binary relationships between flash point, density, net heat of combustion, kinematic viscosity, acidity, aromatics, and sulfur content. After correlations were confirmed, the data was transferred to Microsoft's ML.Net program to make predictions. ML.Net is a machine learning platform developed by Microsoft to integrate multiple machine learning algorithms and find the best performing model for the input data. This reduces the computation time for testing various algorithms by hand and provides the user with the reassurance of knowing they are using one of the best models for their data from the extensive Microsoft archive. Our database provides the extensive training set to ensure accurate models are created.

After specifying which properties to predict, ML.net was given a training set containing approximately 9,000 fuel tests to find the highest performing regression model for each predicted property. ML.net tests about 30 varying algorithms for each property and returns the top one. Figure 5 shows the results from prediction net heat of combustion for JP8 fuel type using Fast Tree Regression. This test set had a nominal error percentage of ~1.8%. Fast Tree Regression is a multiple additive regression tree gradient boosting algorithm. It goes through multiple regression trees and calculates the error for each step and corrects for in the next step. This algorithm was found by ML.NET to be the top performing model for this data.

Deep Learning (Neural Network)-based Data Analysis

While it is recognized that estimating various properties of fuels will be helpful for investigating alternative jet fuels and even for drop-in fuels, the problem itself is complex. Traditionally, a mathematical description of the physical relationships was used to predict properties. This process requires a study of the underlying physicochemical characteristics and the use of classical correlations to determine the outcome. Jet fuels, however, are highly complex mixtures with thousands of hydrocarbon species as opposed to single-component fuels. This complexity obscures the direct implementation of simplified laws, and simple correlations are difficult to find. Over the years, tremendous effort has been made to correlate and estimate various properties based on knowledge of other properties which were generally measured. The work here strives to push this effort to the next level where nonlinear and complex relations can be more accurately modeled. The

proposed methods in this section are based on predictor selection and multivariate linear regressions through an artificial neural network. In so doing, we will strive to build a foundation to enhance the accuracy of the predictions to within 1~2 % for the most well correlated properties.

To improve the accuracy and performance of an artificial neural network and regression analysis, we have employed deep learning methodologies that can be used to learn complex prediction models between input and output variables by utilizing multi-layer neural networks with multiple hidden layers. A benefit of deep learning is that features of the data are connected mathematically and statistically, which enables the representation of intricate nonlinear functions. The correlation used in deep learning has proven to be a reliable property model for nonlinear prediction, even though it is typically shown as a black box.

The prototype of our deep learning-based machine learning approach is shown in Figure. 6. Similar to classical regression process learning in Figure 5, one of the properties among flash point, density, net heat of combustion, kinematic viscosity, acidity, aromatics, and sulfur content have been labeled as output and the other properties are labeled as input. Values are normalized with mean 0 and standard deviation 1 to prevent distortion due to differences in the range of values. A PyTorch (Facebook machine learning code library)-based, fully connected deep learning network was built and trained with 138 fuels' datasets by dividing 100 batches over 7000 epochs of iteration. To ensure nonlinearity in the correlation and prevent vanishing gradient problems for training, LeakyReLU was selected as the activation function. In order to improve calculation without overfitting, four different sets of hidden layer compositions were tested.

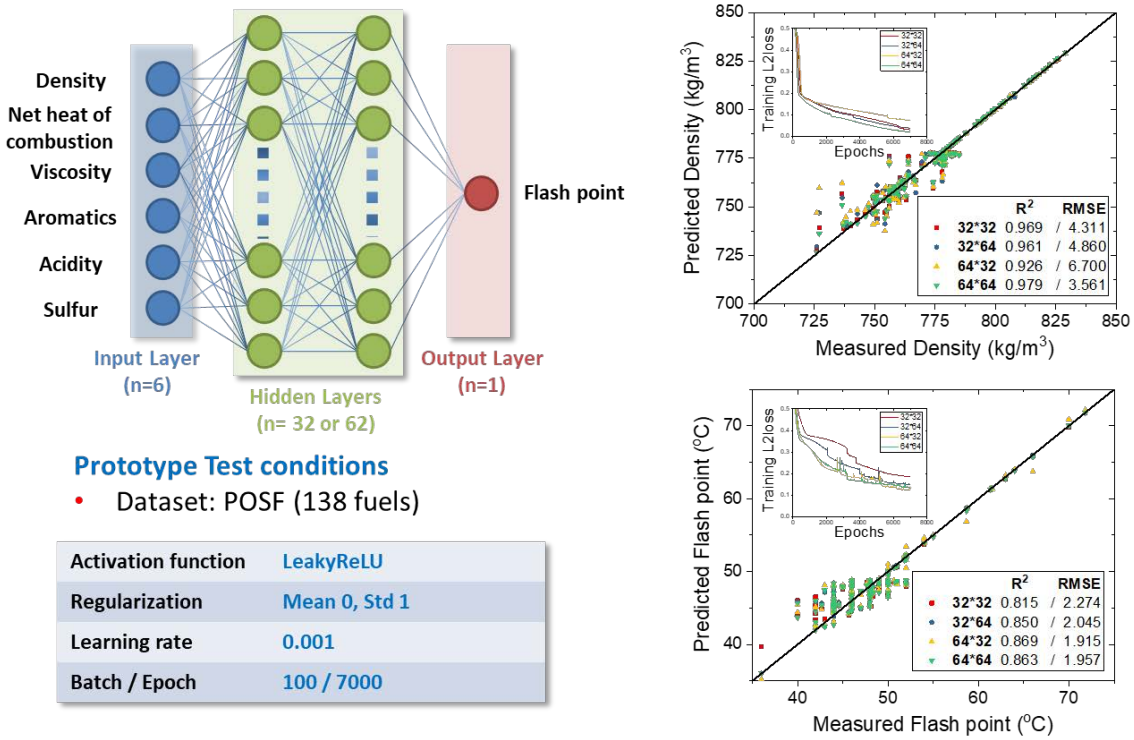

Figure 6. Neural network-based machine learning.

The right side of Figure 6 shows the results of the predictions of density and flashpoint by using other properties. The prototype predicted well, even under low numbers of test sets. In the future, it will be meaningful if this method can be tested under well-organized big data samples. Based on our preliminary studies, we have identified several aspects of the prediction method which could be improved. First, the correlation/prediction results should be explanatory and transparent. The black box model in deep learning, due to its multilayer nonlinear structure, is often criticized for being non-transparent and its predictions untraceable by the user. We feel that we should find underlying comprehensive physical laws to give

insights for guiding the machine learning algorithms. Second, we have to pay attention to prevent overfitting or underfitting to the training dataset. Wrongly trained models will give prediction errors under different sets of data. Generally, deep learning is more suitable for learning from large datasets and could potentially be more difficult to train with small datasets or datasets with skewed characteristics. We should systematically explore this limitation and present guidelines for training datasets in the future.

Milestones

3 months

- Discussion with JETSCREEN on machine learning focus and direction.
- Formalization of machine learning implementation plan (binary correlation, prediction, neural network, etc.).

6 months

- Setting up scripts and algorithms for implementation of machine learning.
- Organization of target data from the database for implementation of machine learning.
- Conversion of data from NoSQL to additional CSV format.

9 months

- Binary correlation analysis using Corrplot (MATLAB).
- Prediction of properties using ML.NET and classical regression machine learning routines.

12 months

- Coding and preliminary implementation of neural network (deep learning) algorithms.

Major Accomplishments

We have started the preliminary implementation of using advanced machine learning algorithms for analysis of data in our database. We have adopted an approach that starts from classical machine learning algorithms based on regression-based analysis with optimization of the specific mathematical routines being used. We then transitioned to more advanced neural network (deep learning)-based analysis where multiple layers of learning nodes should provide superior flexibility in terms of nonlinear computations for property predictions. The comparison between the two methods should allow us to gauge how proficient the implementation of machine learning will turn out to be when we increase the data size. We are coordinating these efforts with our collaborators in the JETSCREEN program and will work on joint publications in the near future.

Publications

N/A

Outreach Efforts

Database made accessible through https://altjetfuels.illinois.edu/

Awards

N/A

Student Involvement

This project was primarily conducted by two graduate students (Isabel Anderson and Keunsoo Kim).

Plans for Next Period

We will expand our machine learning capabilities and provide tangible performance metrics for various datasets in the database. In particular, we are strongly interested in whether machine learning can discern differences in alternative fuel blends and provide predictions that work across an entire range of conventional fuels and SAFs. This will require a focused and extensive analysis of the data in the next period of the program.