Document Title:	**Enhanced Ignitable Liquid and Substrate Database Functionality for Improved Casework and Research**
Author(s):	**Michael Sigman, Ph.D.**
Document Number:	**310048**
Date Received:	**January 2025**
Award Number:	**15PNIJ-22-GG-04421-SLFO**

Agency: National Institute of Justice

Award number: 15PNIJ-22-GG-04421-SLFO

Project Title: Enhanced Ignitable Liquid and Substrate Database Functionality for Improved Casework and Research

PI: Michael Sigman, PhD
Director, National Center for Forensic Science
michael.sigman@ucf.edu
407-823-3526

Submitting official:

Submission date: 12/11/2024

DUNS:

EIN:

Recipient Organization: University of Central Florida
12201 Research Parkway Suite 501., Orlando, FL 32826

Award Period: 01/01/2023 to 12/31/2024

Award Amount: $289,496

Final Summary Overview

Signature of Submitting Official:

Summary of the project

Major goals and objectives

Goal 1: develop a resource for casework whereby the Ignitable Liquids Reference Collection (ILRC) and Substrate databases share the same web interface.

Goal 2: create a large dataset of in-silico fire debris data for development and application of machine learning methods.

Research questions

- Can the three fire debris focused databases be uniformly designed while providing the necessary functionality and be secure?
- Can the databases be accessible on a single website dedicated to the databases?
- Can large datasets of in-silico fire debris be generated for machine learning?
- How well do machine learning methods trained using the in-silico fire debris predict experimental (laboratory generated) fire debris?

Research design, methods, analytical and data analysis techniques

Database Website

The Ignitable Liquids Reference Collection (ILRC), Substrate and Fire Debris databases were developed from 2000 to 2020. The ILRC database was developed in 2000 with a couple of upgrades over the twenty years. The Substrate database was developed in 2010 with one upgrade in 2014. The Fire Debris database was developed in 2020. By 2022, it was necessary to upgrade the programming software to maintain the security of the databases, most specifically the ILRC database. Since the databases were developed over a twenty-year period, there was some discontinuity between them in function and design. Development of new databases within WordPress was suggested by the University of Central Florida Information Security office. Besides providing additional security, utilizing WordPress supported the integration of the three databases within a website dedicated to fire debris analysis.

In-silico Fire Debris Datasets

Complex data such as fire debris requires large amounts of data to train machine learning methods. The question in fire debris analysis is whether the fire debris contains an ignitable liquid residue (ILR) or not. For machine learning classifiers the classifications would be ILR or no ILR which requires data with ground truth knowledge of the class condition. Fire debris collected as evidence at a scene may contain chemical compounds from several pyrolyzed materials (substrates) and if an ILR is present it will most likely be partially evaporated. Producing laboratory generated fire debris samples with these conditions is time and cost intensive. An alternative is to utilize reference material data to electronically evaporate ignitable liquid (produce ILRC data) and mix them with multiple substrate materials to produce in-silico fire debris data. Generation of in-silico fire debris data can provide large datasets for training and cross-validation of machine learning classifiers. A simple stepwise procedure of generating in-silico fire debris is described in Figure 1.

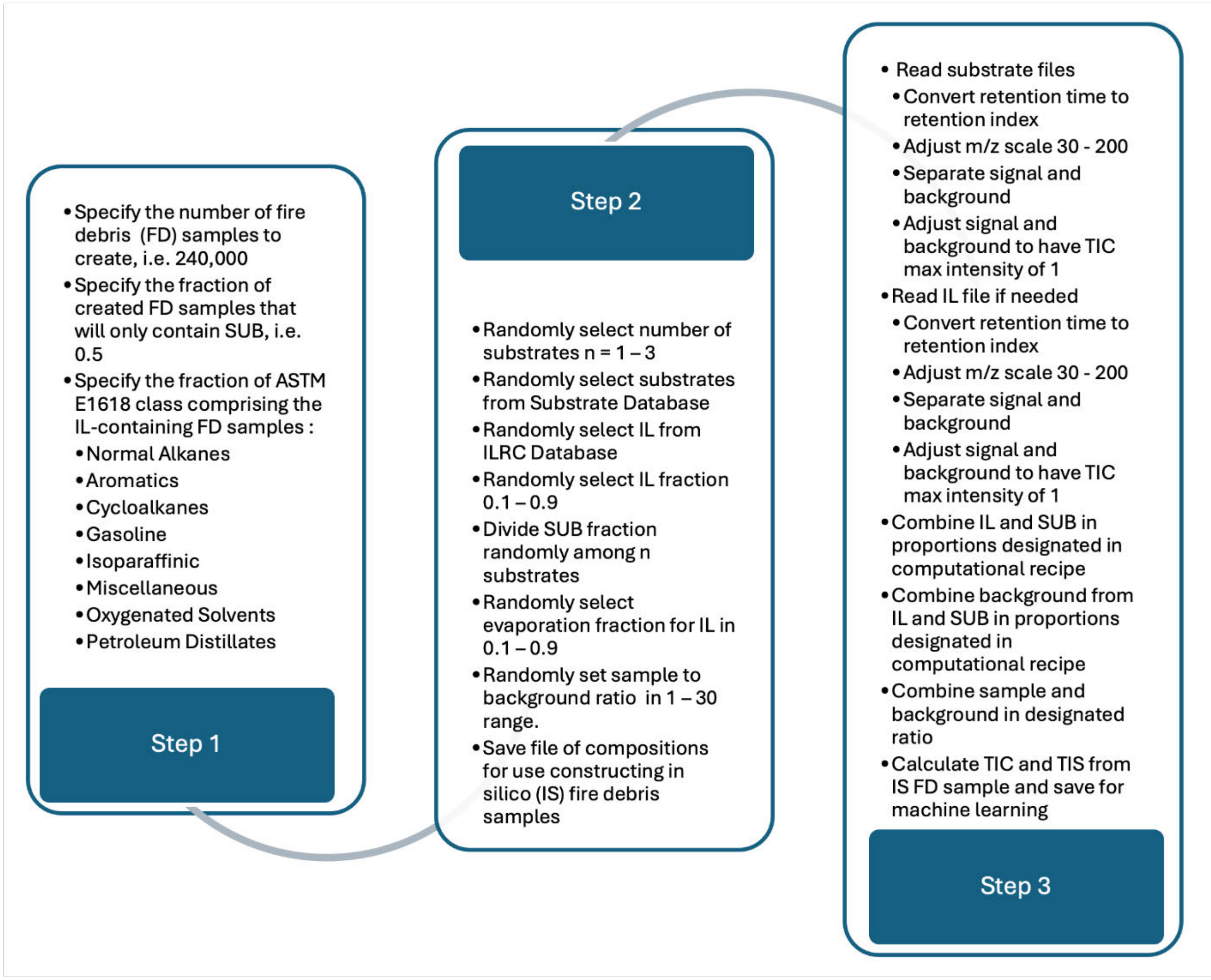


Figure 1 A simple stepwise procedure for generating in-silico fire debris gas chromatography – mass spectrometry data.

An XGBoost machine learning classifier was trained on an in-silico fire debris dataset totaling 240,000 total ion spectra (TIS) to test the dataset's suitability for machine learning. Background ions 32 m/z and 76 m/z were removed from the TIS in the dataset. From the remaining ions, a set of 28 ions representative of IL classes, as given in Table 2 of ASTM E 1618-19 were selected from the IS FD data. Ten percent of the data was held for testing. The remaining data was subjected to principal components analysis (PCA) where 14 principal components were utilized as observations in the model.

In addition, an evaluation of seven machine learning methods was performed utilizing both TIS and total ion chromatogram (TIC) data from the 240,000 in-silico fire debris datasets. The seven machine learning methods are logistic regression (LR), linear discriminant analysis (LDA), quadradic discriminant analysis (QDA), support vector machine (SVM), random forest (RF), XG boost (XGB), and neural networks (NN) to predict whether an ignitable liquid residue was present in the fire debris. From the TIS data, a set of 30 ions were selected and normalized. From the TIC data, 2660 retention indices were selected and normalized. In both datasets, 80% were used for training and 20% were used for testing. Principal components analysis (PCA) was performed on the training data with principal components accounting for 90% of the variance being used to train the machine learning models. Multiple TIC models were developed by changing the bin size of the retention indices to determine an optimal bin size.

Expected applicability of the research

A unified website provides easy access to three databases utilized in fire debris analysis and will enhance training and casework. The addition of a large in-silico fire debris dataset provides further research for the development of machine learning methods in fire debris analysis.

Participants and other collaborating organizations

Dr. Tang provided in-silico fire debris datasets for his statistics students in their course work involving machine learning.

Changes in approach from original design and reason for change, if applicable

An in-silico fire debris database was to be developed with the data constructed for the in-silico fire debris datasets. In August 2024, the ILRC, Substrate and Fire Debris databases needed to be removed from the internet due to security issues. A grant award modification was submitted and approved to redevelop the Fire Debris database (laboratory generated data) and incorporate it into the new database website rather than creating a new database of in-silico fire debris.

Outcomes

Activities/accomplishments

The three NCFS databases focusing on fire debris analysis were redeveloped and are now accessible from a singular website. Redevelopment of the databases included new programming, design, and additional security features while maintaining the functionality users expect. Another accomplishment was completing research on the development of in-silico fire debris data. The results are four datasets each containing 60,000 records of in-silico fire debris. An additional dataset is laboratory generated fire debris from the Fire Debris database. These datasets are available on the Database website for use in machine learning.

Results and findings

Database Website

The database website is accessible at https://ilrc.ucf.edu/. This website is hosted by the National Center for Forensic Science at the University of Central Florida. The Homepage provides information about the databases, what is new, and their history as shown in figure 2. **ILRC-Substrate-Fire Debris** contains several pages providing information about each database, how to order ignitable liquid reference materials, and contact information.

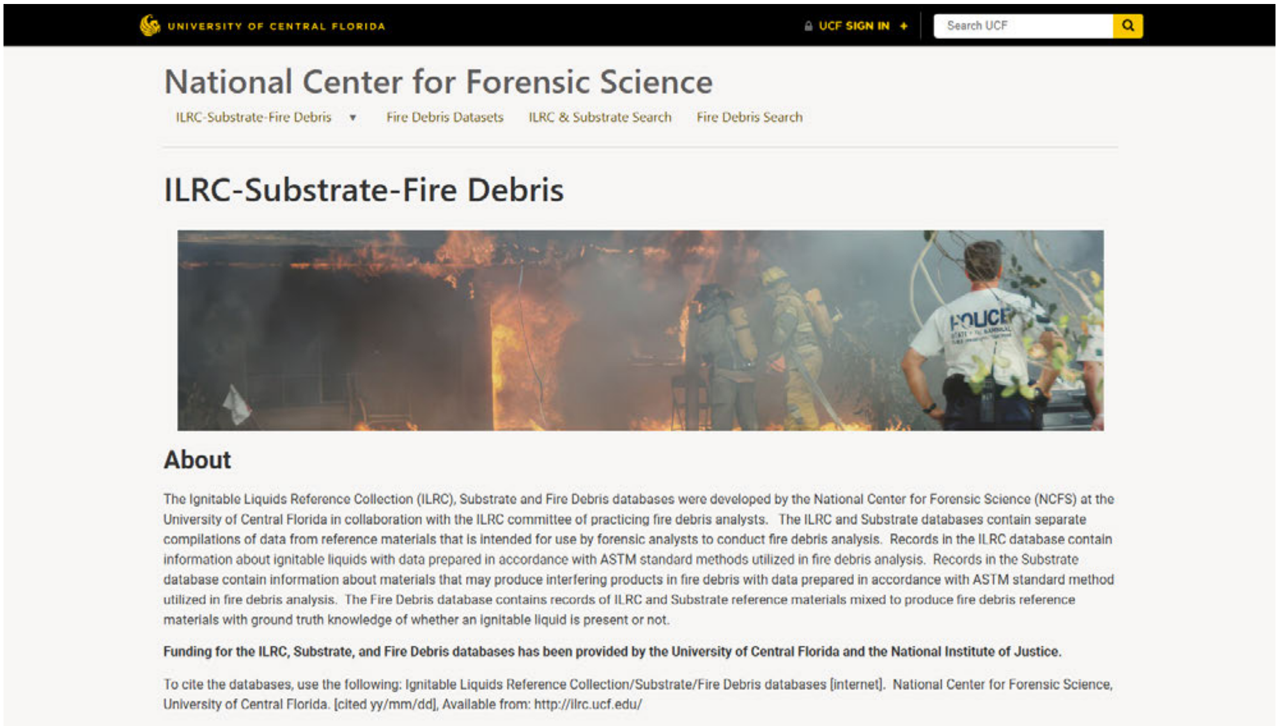


Figure 2 Homepage of new ILRC-Substrate-Fire Debris database website.

Fire Debris Datasets page contains information about the development of in-silico fire debris data, download links to both the in-silico and the laboratory generated dataset from the Fire Debris database, and training videos. Each dataset contains three files: sample information, total ion chromatogram data, and total ion spectra data as shown in figure 3.

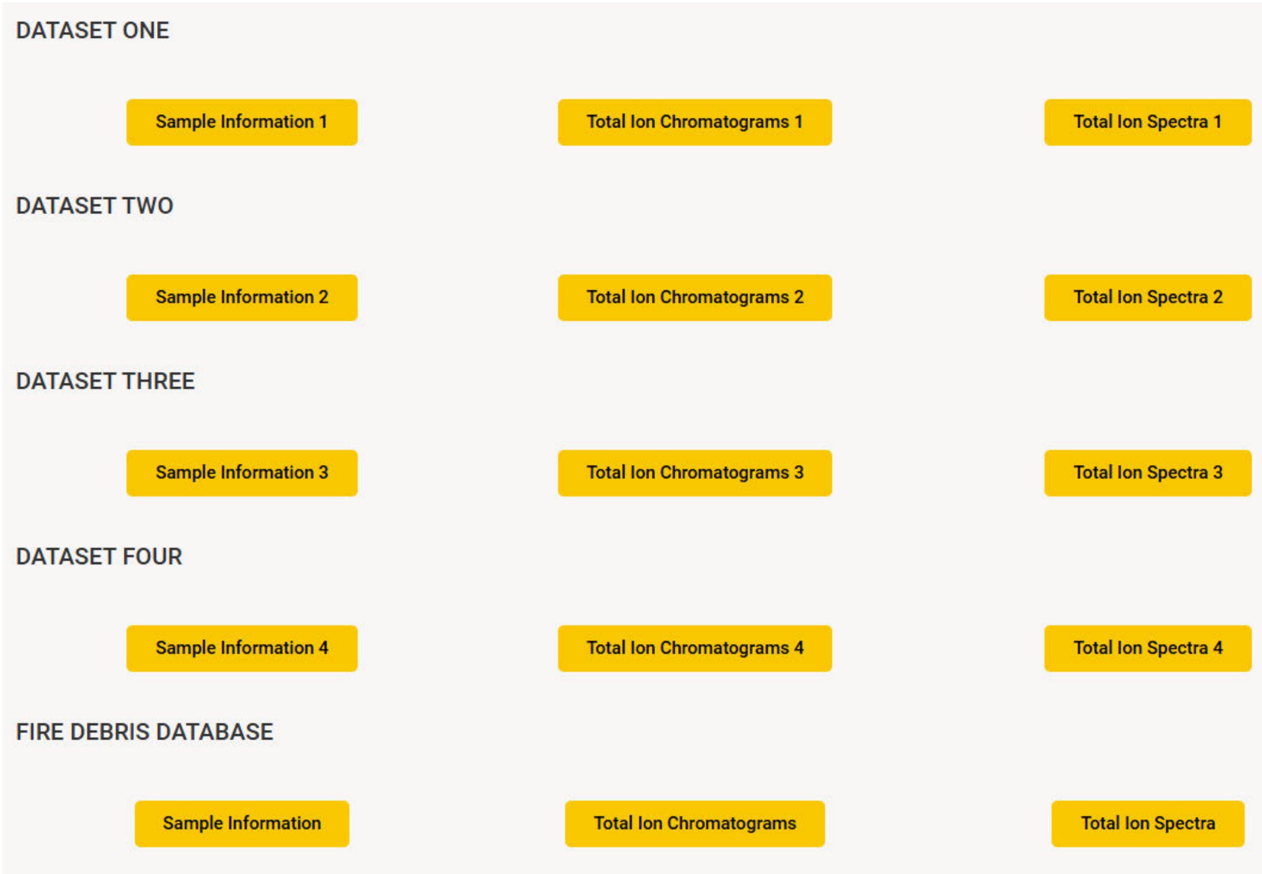


Figure 3 View of Fire Debris Datasets webpage where users can download large fire debris datasets.

ILRC & Substrate Search directs the user to the ILRC and Substrate databases where they can search each database and have the record results displayed side by side as shown in figure 4.

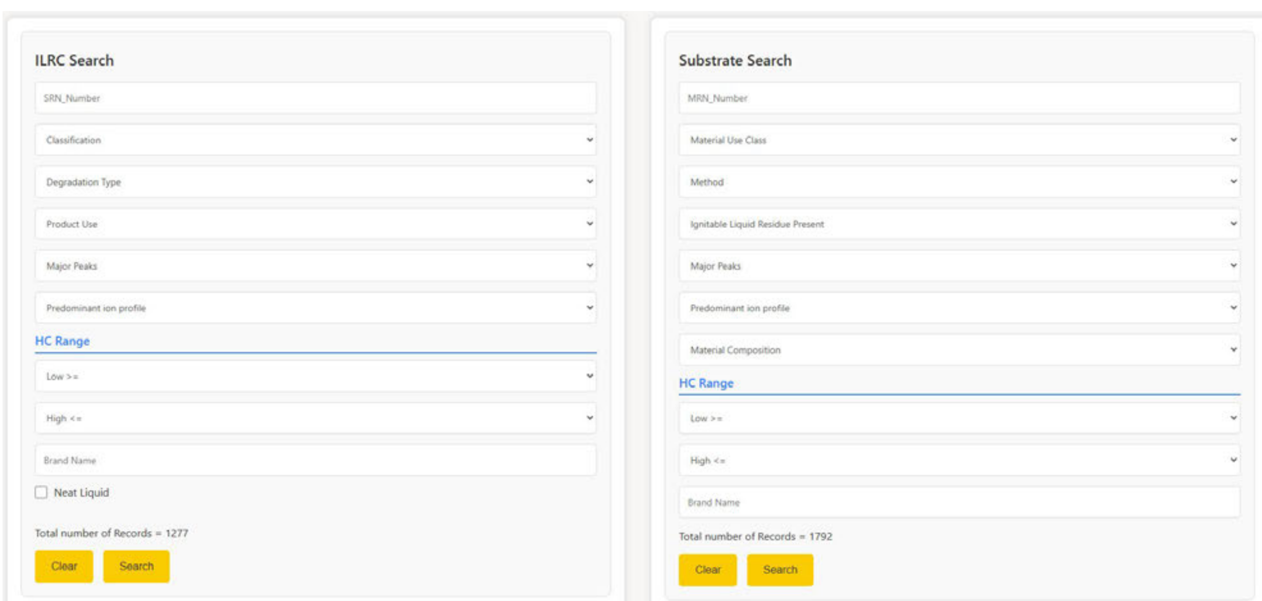


Figure 4 View of the ILRC-Substrate databases search page where users can select parameters for the search of the ILRC and Substrate databases.

Search returns records meeting the criteria selected by the user as shown in figure 5. The total number of records returned is displayed.

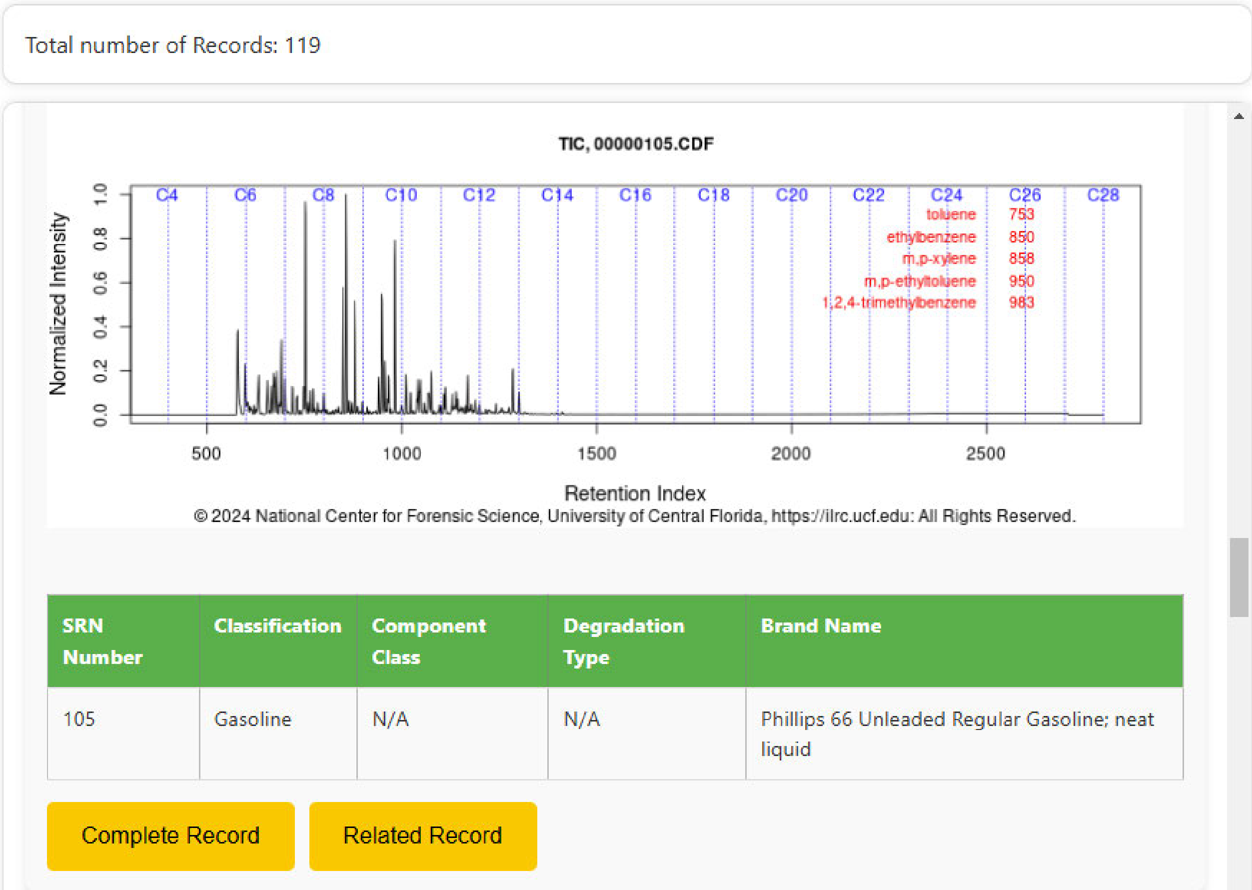
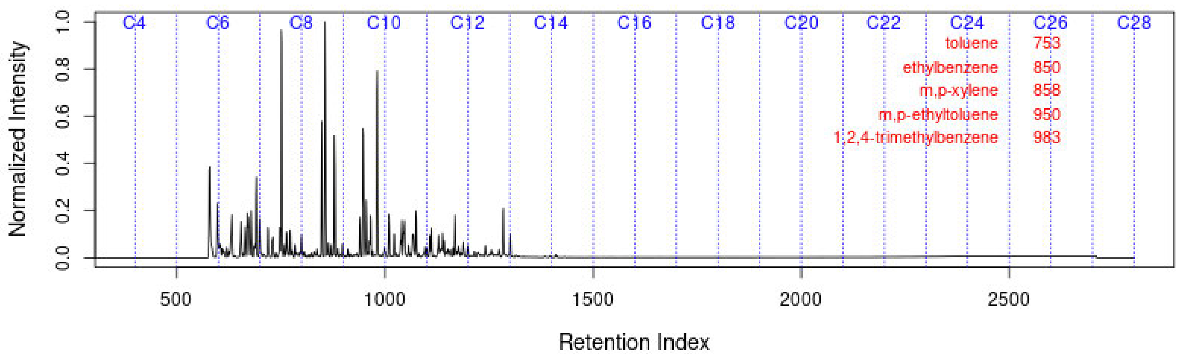


Figure 5 View of search return results where the user selected the Gasoline classification parameter in the ILRC database search. Included is the total number of records that were returned.

Complete Record displays the total ion chromatogram, information about the reference material (ignitable liquid or substrate), and download links to relevant data as shown in figure 6.



Record Details

Field	Value
Classification	Gasoline
Component Class	N/A
Degradation type	N/A
Extent of Degradation	N/A
Carbon Low	C 6
Carbon High	C 13
Product Use	Gasoline
Profile	Aromatics
Brand Name	Phillips 66 Unleaded Regular Gasoline; neat liquid
Major Peaks	toluene m,p-xylene 1,2,4-trimethylbenzene ethylbenzene m,p-ethyltoluene

Chromatograms Data Files SDS

Figure 6 A view of the Complete Record of SRN 105 that is one of the records returned in the Gasoline classification search.

Chromatograms documents include the total ion chromatogram, extracted ion chromatograms, total ion spectrum, and a bar chart demonstrating the relative concentration of the five compound classes. **Data Files** downloads the common data format (cdf) files of reference material that can be imported into any gas chromatograph mass spectrometer data analysis software. **SDS** opens the safety data sheet for the reference ignitable liquid. **Related Records** opens a new tab displaying the total ion chromatograms of records that demonstrate the various degradation conditions of the reference material.

Fire Debris Search directs the user to the Fire Debris database where users can search laboratory generated fire debris records as shown in figure 7.

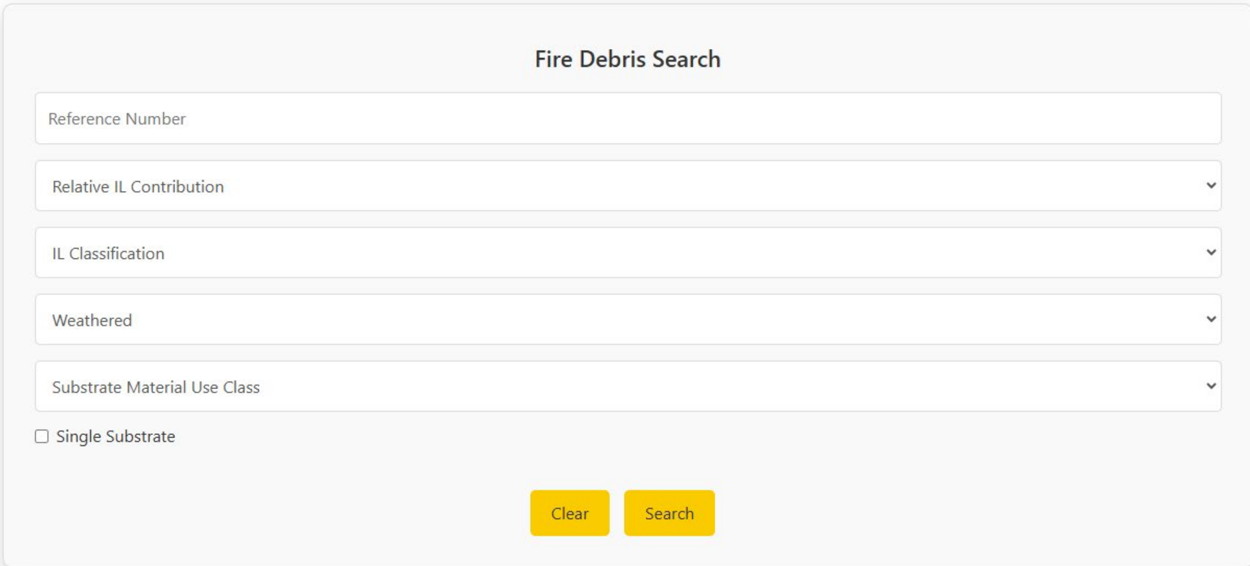


Figure 7 A view of the Fire Debris database search page.

Search returns records meeting the criteria selected by the user. The total number of records returned is displayed. Each record displayed in the return contains all information and download links to Chromatograms and Data files as shown in figure 8.

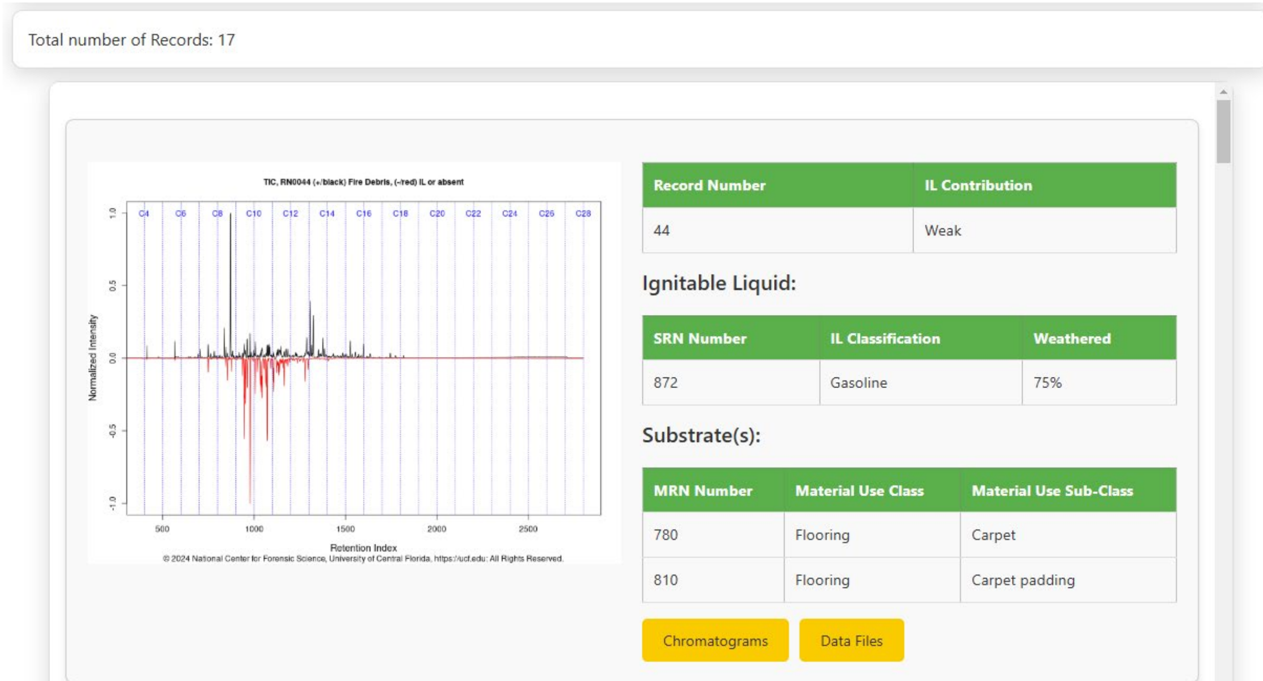


Figure 8 A view of record RN 44 from the Fire Debris database where the user selected Gasoline for the ignitable liquids classification and carpet for the substrate material class for the search. Included is the total number of records that were returned.

In-silico Fire Debris Datasets

A major process in the generation of in-silico fire debris data is the digital evaporation of the ignitable liquid. In this research, the methodology followed was developed by Smith and coworkers. Results from a comparison of digital and experimental evaporation (weathering) established the fitness for utilizing this methodology in generating in-silico fire debris data. The comparisons were performed by calculating the Pearson product moment correlation coefficient (PPMC) between the digital and experimental data. The best matches are represented by a large PPMC. The range of PPMC was 0.7 to 0.98 and considered sufficient for incorporation into the in-silico data.

Cross-validation results from the XGBoost machine learning classifier model using 28 TIS ions had an area under the curve (AUC) for the receiver operating characteristics (ROC) curve of 0.965. This result indicates the data performs well as a training dataset for XGBoost models. A dataset of laboratory generated fire debris data from the Fire Debris data was also predicted as to whether an ignitable liquid was present utilizing the same models. The ROC AUC for the laboratory generated fire debris (experimental data from Fire Debris database) was 0.85.

ROC curves were created from the predicted probability of class IL membership and the sample ground truths using seven machine learning methods. The area under the curve (AUC) from the testing of TIS in-silico fire debris data and TIS experimental fire debris data (Fire Debris database) for each of the seven machine learning models are in table 1.

Method	In-silico TIS	Experimental TIS
LR	0.86	0.83
LDA	0.86	0.83
QDA	0.87	0.79
SVM	0.95	0.84
RF	0.90	0.82
XGB	0.95	0.84
NN	0.96	0.86

Table 1 ROC AUC results from seven machine learning methods for test in-silico and experimental TIS.

The optimal binning size for the TIC retention indices was 25. The ROC AUC results for the testing data of in-silico TIC at various bin sizes are in table 2.

Method	Bin n = 0	Bin n = 10	Bin n = 25	Bin n = 50
LR	0.54	0.65	0.62	0.63
LDA	0.54	0.66	0.62	0.62
QDA	0.63	0.71	0.67	0.67

SVM	0.60	0.78	0.78	079
RF	0.64	0.73	0.79	0.63
XGB	0.66	0.72	0.68	0.73
NN	0.59	0.71	0.77	0.71

Table 2 ROC AUC results from seven machine learning methods for test in-silico TIC data utilizing various sized bins of retention indices.

The results from modeling seven machine learning methods on in-silico fire debris data demonstrate that better results are obtained utilizing the TIS rather than the TIC data. Large datasets of in-silico fire debris data are valuable in training machine learning methods that can obtain ROC AUC above 0.85. Although lower ROC AUC predictions of experimental fire debris data were obtained, they are greater than 0.79.

Limitations

Database Website

A couple of functions in the previous database programming were eliminated in the new programming. This was necessary to ensure the security of the new databases.

In-silico Fire Debris Datasets

Fire debris is formed within a fire and disturbed by firefighting efforts creating complexity in fire scene evidence. The Insilco and laboratory generated fire debris data may not be as complex as fire scene evidence. The advantage of the datasets is knowing the ground truth as to whether an ignitable liquid is present or not. Currently, there is no process for determining the ground truth (ignitable liquid present or not) of fire scene evidence.

Artifacts

List of products (e.g. publications, conference papers, technologies, websites, databases), including location of these products on the internet or in other archives or databases

Databases

ILRC-Substrate-Fire Debris Database Website https://ilrc.ucf.edu/

Publications

Sigman, M. E., Williams, M. R., Tang, L., Booppasiri, S., Prakash, N., In Silico Created Fire Debris Data for Machine Learning, Forensic Chemistry, submitted October 2024, https://doi.org/10.1016/j.forc.2024.100633

Tang, L., Booppasiri, S., Sigman, M. E., Williams, M. R., Evaluating Machine Learning Methods on a Large-Scale of In Silico Fire Debris Data, Forensic Chemistry, submitted November 2024, under review

Presentations

Sigman, M. E., Williams, M. R., Tang, L., Booppasiri, S., In-Silico Fire Debris Data for Machine Learning Education and Forensic Applications, University of Central Florida, Department of Statistics, November 3, 2023.

Tang, L., Sigman, M. E., Williams, M. R., Comparison of Computational and Laboratory Generated Fire Debris Data, American Academy of Forensic Science, Denver, Colorado, February 19 – 24, 2024.

Prakash, N., Sigman, M. E., Williams, M. R., Tang, L., Enhanced Ignitable Liquid and Substrate Database Functionality for Improved Casework and Research, American Academy of Forensic Science, Denver, Colorado, February 19 – 24, 2024.

Datasets generated (broad descriptions will suffice)

Four datasets of 60,000 in-silico fire debris data are available for download. Each dataset contains three files: sample information, total ion chromatograms, and total ion spectra. The files are in zipped (zip) comma separated value (csv) file formats. The datasets are found on the ILRC-Substrate-Fire Debris Databases website https://ilrc.ucf.edu/insilico-fire-debris-datasets/

Dissemination activities

A Database website containing three databases and five fire debris datasets are available on https://ilrc.ucf.edu/. Two manuscripts will be published describing the generation of in-silico fire debris and machine learning methods utilizing the in-silico and laboratory generated fire debris datasets.