Energy Fuels **2010,** *24,* 5560–5572 · DOI:10.1021/ef100840q
Published on Web 09/09/2010

energy&fuels
article

Use of Genetic Algorithms To Improve Partial Least Squares Fuel Property and Synthetic Fuel Modeling from Near-Infrared Spectra

Jeffrey A. Cramer,* Robert E. Morris, and Susan L. Rose-Pehrsson

United States Naval Research Laboratory, 4555 Overlook Avenue, Southwest, Washington, District of Columbia 20375

Received July 1, 2010. Revised Manuscript Received August 24, 2010

Partial least-squares (PLS) regression models can be constructed from near-infrared (NIR) spectroscopic data to predict critical specification properties of jet and diesel fuels for quality surveillance. This same approach has also been used to identify Fischer−Tropsch (FT) synthetic fuels and predict their quantities in blends with jet or diesel petroleum-derived fuels at concentrations as low as 15% with a sequential modeling approach. The present work expands on these previous results using genetic algorithms to select the maximally useful subset of all available training data for both fuel property modeling and FT fuel identification purposes. Prediction improvements primarily stem from the ability of the automated genetic algorithm and *F*-test algorithm to select larger numbers of latent variables without overfitting, which is not possible by simply modifying the *F*-test threshold. These improved data subsets provide significant improvements in the precision of PLS property modeling, as well as the lower detection limit of 10% for FT fuels in blends with conventional petroleum-derived fuels. This provides the means to extend the development of predictive PLS models to detect, quantify, and predict many specification properties of FT fuels, neat and in blends with petroleum fuels. Additionally, the use of genetic algorithms now makes the detection and discrimination of FT fuels in blends with multiple FT fuels feasible with the same 10% limit of detection.

Introduction

The Naval Research Laboratory has been engaged in the development of sensor-based technology to replace the Combined Contaminated Fuel Detector (CCFD), used to perform fuel-quality surveillance onboard Navy ships. The Navy Fuel Property Monitor (NFPM)[1] derives predictions of various critical specification fuel properties using partial least-squares (PLS) modeling of near-infrared (NIR) spectra. This approach offers significant advantages by reducing the time and manpower currently required to measure specification properties[2] and will also improve shipboard safety by minimizing the manual transport of fuel aboard Naval vessels. A spectroscopic analyzer can perform the analysis *in situ* without altering the fuel and will also not require any supplies or generate any disposable waste. As synthetic fuels and fuels derived from biomass are introduced into the Navy fuel supply system, the NFPM will be expected to perform these functions in the presence of alternative fuels, such as Fischer−Tropsch (FT) synthetic fuels.[3] While it is anticipated that synthetic fuels will initially be deployed as 50% blends with conventional petroleum fuels, standard fuel handling and storage practices lead to the comingling and blending of different fuels of a similar grade. Therefore, one of the design goals of the NFPM is to have the capability to successfully analyze fuels with synthetic fuel contents ranging from 0 to 100%. This laboratory has previously shown[4] that PLS property predictions from models created from conventional fuels can be used with blends containing FT fuel by applying corrections derived from the FT fuel identity and content. The precision of these corrected property values predicted in FT fuels and blends is on par with those derived from NIR spectra of conventional fuels. Both the fuel properties themselves and the FT fuel content are calculated from the NIR spectrum using a PLS-based modeling procedure.

Improvements to the fuel property modeling algorithms are continuously pursued to improve the performance of the NFPM. While the fuel properties are modeled with property-specific PLS models, current methods to quantify FT fuel content in a blend require successive testing with model pairs that are specific for each different FT fuel and each FT fuel must be evaluated separately because of the impact of processing and blending on the chemical constituencies of FT fuels from different refineries. The detection procedure makes use of PLS model pairs consisting of one identification model and one quantification model. The identification models are constructed with a training set consisting of all of the neat petrochemical fuel samples available to our laboratory as well as those samples possessing an appropriate FT fuel content. The primary use of identification models is to simply determine whether or not a FT fuel is present. While these models are well-suited to this task because of the variety of calibration data, they also tend to produce relatively inaccurate quantification results, hence, the need for separate quantification models. The quantification models, in contrast, are constructed

*To whom correspondence should be addressed. Telephone: 202-404-3419. Fax: 202-767-1716. E-mail: jeffrey.cramer@nrl.navy.mil.

(1) Morris, R. E.; Hammond, M. H.; Cramer, J. A.; Johnson, K. J.; Giordano, B. C.; Kramer, K. E.; Rose-Pehrsson, S. L. *Energy Fuels* **2009**, *23*, 1610–1618.

(2) American Society for Testing and Materials (ASTM). *Annual Book of ASTM Standards*; ASTM: West Conshohocken, PA, 1997.

(3) Lee, S.; Speight, J. G.; Loyolka, S. K. *Handbook of Alternative Fuel Technologies*; CRC Press (Taylor and Francis Group): Boca Raton, FL, 2007; pp 160−167.

(4) Cramer, J. A.; Morris, R. E.; Giordano, B. C.; Rose-Pehrsson, S. L. *Energy Fuels* **2009**, *23*, 894–902.

with only those blended samples containing the specific FT fuel being sought, including the pure FT sample. The quantification models provide much more accurate predictions of FT content for those blends actually containing FT fuels but also produce unduly high FT content predictions for blends containing no FT fuel, as would be expected for a model with no neat, non-FT fuels included in its calibration data. Both of these model types produce quantitative results, and an unknown sample must be identified as a FT fuel or a blend containing a FT fuel by both the identification and quantification model in a model pair before it is reported as containing a quantity of FT fuel as predicted by the actual quantification model. This approach tends to decrease the likelihood of reporting false-positive (FP) results while providing quantitative predictions that are as accurate as possible. Furthermore, the model pairs are sequentially applied to enable each model pair to be focused more closely on the FT fuel being sought. This sequential application requires each model pair after the first pair to exclude those samples containing previously predicted FT fuels. This operation functions as a filtering step that allows sequential model pairs to more accurately model FT fuels of immediate concern without having to account for those FT fuels that have already been detected. While this filtering is both practical and effective, it provides an inherent limitation to the analysis, because once a fuel sample is determined to contain a certain FT fuel, no other FT fuel can be reliably identified because subsequent model pairs do not take previously detectable FT fuels into account.

Despite the inherent intermodel dependence in this FT assessment scheme, the lower limit of detection, with no FP results for all FT fuels individually, is 15%. This limit of detection requires a threshold FT content that must be surpassed by the FT fuel content prediction of any given model before the model is allowed to report the presence of a FT fuel at all. In other words, if the FT content of a sample is predicted to be below this threshold value in one of the models in an identification and quantification model pair, then the model pair as a whole reports a 0% FT fuel content, thereby further discriminating against FP results. This threshold value, referred to in the previous[4] and present work as a "cutoff" value, is set to 9% in our present hardware to achieve the aforementioned 15% limit of detection. It should be noted here that cutoff values are defined as the minimum whole-number percent values that still allow for no FP FT content results, which are highly undesirable in the context of an automated fuel analysis tool, such as the NFPM. Limits of detection for predicted FT content, meanwhile, are defined as the minimum percent of FT fuel content found in the available FT fuel blends that can be reliably predicted by a given model pair.

Figure 1 depicts the predicted FT contents from a series of known (calibrated) blends with both an identification and quantification model. The deviation from the 1:1 correspondence seen with samples possessing lower FT contents yields a lower limit of detection of 15% when the cutoff value is 9%. It should be noted that this FT model pair is the poorest performer among those used in our standard detection scheme, hence, the fact that the cutoff and limit of detection are identical to those found in the overall scheme. In this particular model pair, a sample must be determined to possess a FT content above 9% by both models before the model pair will report the prediction of the quantification model as the final predicted FT fuel content and the sample with the lowest concentration that is reported under these conditions possesses a known FT content of 15%.

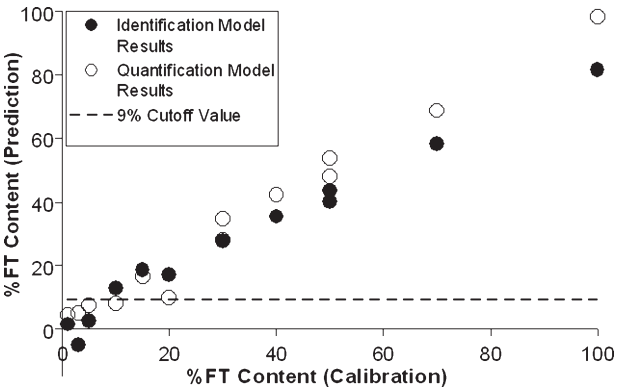

Figure 1. Sequential modeling results of a gas-to-liquid (GTL) FT JP-5 blended with petroleum JP-5 (both identification and quantification model results shown). Results for samples containing no FT fuel (i.e., 0% calibration FT content) were not plotted to clarify the figure.

It is obviously desirable to lower this detection limit, because many of the properties predicted in blends containing less than 15% FT fuel require corrections that are based on the FT content. Furthermore, as stated above, comingling of fuels within the fuel handling system requires that the modeling strategy must be capable of detecting individual FT fuels in the presence of other FT fuels, which is not supported in the current strategy. In other words, although a version of FT fuel modeling is presently used, this version is constrained to detect fuels in a specific order because of the aforementioned intermodel dependence, and such a constrained "in-order" modeling strategy would not be as useful as an unconstrained, "out-of-order" strategy with no such dependence.

The studies described herein were undertaken in an effort to improve the precision of the fuel property models, to reduce the lower limits of detection of FT fuels, and to accommodate multiple FT fuels in the same blend.

Genetic Algorithms (GAs). Previous work has shown that deliberate data selection can provide significant improvements in spectroscopic models. The vast majority of this work is based on truncating the multivariate spectroscopic variables in such a way that irrelevant or superfluous linear relationships that can interfere with the construction of useful models are minimized.[5−19] A great deal of this wavelength

(5) Jouan-Rimbaud, D.; Walczak, B.; Massart, D. L.; Last, I. R.; Prebble, K. A. *Anal. Chim. Acta* **1995**, *304*, 285–295.
(6) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandenginste, B.; Sterna, C. *Anal. Chem.* **1996**, *68*, 3851–3858.
(7) Spiegelman, C. H.; McShane, M. J.; Goetz, M. J.; Motamedi, M.; Yue, Q. L.; Cote, G. L. *Anal. Chem.* **1998**, *70*, 35–44.
(8) Horchner, U.; Kalivas, J. H. *Anal. Chim. Acta* **1995**, *311*, 1–13.
(9) Carlson, P.; Danielssona, M.; Dejardin, M.; Jon-And, K.; Sjolin, J. *Nucl. Instrum. Methods Phys. Res., Sect. A* **1996**, *381*, 152–156.
(10) Swierenga, H.; de Groot, P. J.; de Weijer, A. P.; Derksen, M. W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 237–248.
(11) Aarts, E.; Korst, J. *Simulated Annealing and Boltzmann Machines*; John Wiley and Sons, Ltd.: Chichester, U.K., 1989; pp 13−31.
(12) Todeschini, R.; Galvagni, D.; Vèlchez, J. L.; del Olmo, M.; Navas, N. *Trends Anal. Chem.* **1999**, *18*, 93–98.
(13) Capitan-Vallvey, L. F.; Navas, N.; del Olmo, M.; Consonni, V.; Todeschini, R. *Talanta* **2000**, *52*, 1069–1079.
(14) Forbes, N. *Imitation of Life: How Biology Is Inspiring Computing*; MIT Press: Cambridge, MA, 2004; pp 1−11.
(15) Brown, P. J.; Vannucci, M.; Fearn, T. *J. Chemom.* **1998**, *12*, 173–182.
(16) Hageman, J. A.; Streppel, M.; Wehrens, R.; Buydens, L. M. C. *J. Chemom.* **2003**, *17*, 427–437.
(17) Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. *Appl. Spectrosc.* **2000**, *54*, 413–419.

selection has been performed with GAs.[20−24] We have previously investigated wavelength selection as a means to gain additional improvements in the precision of PLS fuel models constructed from NIR spectra.[25] However, although the regression vectors for the PLS models indicated that only certain portions of the spectrum were being used for various fuel property models, simple wavelength-based variable selection with NIR spectra acquired with the NFPM did not significantly improve either fuel property predictions or FT fuel detection beyond what had already been achieved. In addition to the deliberate selection of spectroscopic wavelength data using GAs, it is also possible to deliberately select the samples from which a PLS model is constructed, with the same goal of reducing interfering linear variance. This type of deliberate sample selection was performed roughly a decade ago based on the use of GAs.[26] However, the goal of this previous GA sample selection work was to select representative subsets and not to minimize the limit of detection, as is the goal of the present work.

The theory behind GAs can be summed up in Darwinian terms, as "survival of the fittest", and this type of approach can be applied to the systematic selection of subsets of fuel samples for optimizing subsequent PLS models. Instead of using all available samples to produce PLS models, only a subset of all available samples is used. This sample subset, in the context of GAs, can be regarded as an "individual" with a specific "genetic code" that is represented by its constituent training samples. The precision of the resultant PLS property prediction model constructed from an "individual" would thus constitute the "fitness" of said individual because the quality of the final model is the criterion upon which the assessment of the GA is based. In an iterative routine, a population of multiple individuals is used and those with the best fitness, i.e., most useful PLS models resulting from their "genetic code" samples, are those that survive, with the rest being eliminated. The population of individuals that survives this process is then used to produce new individuals to replace the eliminated members, which is accomplished by randomly recombining their constituent samples (i.e., "genetic code") into new "individual" sample populations. This usually includes a small chance of random mutation to ensure that as many samples as possible are evaluated in the course of performing the GAs. This process is then repeated until each of the sample populations are identical, at which point, the data subsets are optimized with respect to "fitness" for model construction purposes.

To account for potential local minimum results, GA procedures are generally performed in replicate. It should also be noted that, to maintain maximum model utility, all PLS models produced by the GA-selected sample subsets should still be evaluated for fitness using all available training data regardless of whether or not those training data were used to construct the PLS model.

This sample selection procedure is performed to improve the fuel property modeling by selecting a unique training sample subset for each property. It is also applied to the FT modeling but only to detect and identify the particular FT fuel present and not to quantify the FT fuel content in blends with petroleum-derived fuels. In other words, GAs are only used to refine the identification models and not the quantification models. Because relatively few of the identification model samples contain FT fuels, the identification modeling is a good candidate for GA-based sample selection; because there are relatively few samples that actually contain FT fuel, the quantification modeling is a comparatively poor candidate. Explanations of the FT fuel modeling populations and their FT fuel content information can be found in the following section.

Experimental Section

Fuels. A group of 708 fuel samples, collected worldwide, was used for the fuel property modeling seen in the present study. These 708 samples do not include the FT fuel and blend samples that will be described shortly, because it was previously determined that fuel property models produce more accurate models for FT fuels and blends when the property predictions made for these samples are mathematically adjusted on the basis of predicted FT fuel content.[4] Because many of the fuel constituents that influence the critical fuel properties of interest are different for different fuel types, the precision of the predictive PLS models was improved by segregating the fuel samples into two broad categories (jet and diesel), with no cross-analyses between the two populations. The jet fuel training set consisted of 64 JP-5, 195 JP-8, 101 Jet A, and 26 Jet A-1 fuels. The diesel fuel training set consisted of 226 F-76 Naval distillate samples, 81 marine gas oil (MGO) samples, and 15 ultra-low sulfur diesel (ULSD) samples. All of the fuel samples were provided with property analyses, and all were in compliance with the applicable specifications. The property values modeled in this study for each class of fuel (jet and diesel) along with the American Society for Testing and Materials (ASTM) methods used are shown in Table 1. The property analyses were, in turn, used as the calibration data for the fuel property model constructions. It should be noted here that, because of differences in fuel property monitoring requirements when samples are collected from different sources, not every sample in the training set has an associated ASTM value available for every desired fuel property. This is the reason that the number of samples available for each fuel property does not equal 708 and the reason that some fuel properties are represented by much fewer samples. However, every sample that has a usable ASTM value is included in every given model construction.

In addition to the diesel fuels described above, another 60 sample diesel data set consisting of 52 MGO samples and 8 F-76 samples was available for this study. These samples were not included in the fuel property prediction work because their reported ASTM fuel property values were known to be significantly less precise than those reported for the primary 322 diesel samples and, thus, were deemed unsuitable for use as diesel fuel training data. This population of 60 outliers was instead used to evaluate the ability of GAs to eliminate poorly correlated samples from the calibration fuel set.

The FT training data were comprised of not only the aforementioned 708 samples but also two synthetic JP-5 fuels and two synthetic diesel fuels. One FT JP-5 fuel was produced by a coal-to-liquid (CTL) process, and one FT JP-5 fuel was produced from a natural gas-to-liquid (GTL) process. The FT diesel fuel sample set, in contrast, consisted of one CTL product and a GTL FT diesel fuel derived from ethylene that is simply referred

(18) Jiang, J.-H.; Berry, R. J.; Siesler, H. W.; Ozaki, Y. *Anal. Chem.* **2002**, *74*, 3555–3565.
(19) Du, Y. P.; Liang, Y. Z.; Jiang, J. H.; Berry, R. J.; Ozaki, Y. *Anal. Chim. Acta* **2004**, *510*, 183–191.
(20) Lucasius, C. B.; Beckers, M. L. M.; Kateman, G. *Anal. Chim. Acta* **1994**, *286*, 135–153.
(21) Arcos, M. J.; Ortiz, M. C.; Villahoza, D.; Sarabia, L. A. *Anal. Chim. Acta* **1997**, *339*, 63–77.
(22) Ghasemi, J.; Niazi, A.; Leardi, R. *Talanta* **2003**, *59*, 311–317.
(23) Abdollahi, H.; Bagheri, L. *Anal. Chim. Acta* **2004**, *514*, 211–218.
(24) Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*; Springer-Verlag: Berlin, Germany, 1992; pp 13−22.
(25) Cramer, J. A.; Kramer, K. E.; Johnson, K. J.; Morris, R. E.; Rose-Pehrsson, S. L. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 13–21.
(26) Tominaga, Y. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 157–163.

Table 1. ASTM Methods Used To Measure Critical Fuel Properties Modeled

jet fuel property	units	ASTM method
flash point	°C	D 93
density at 15 °C	kg/m^3	D 4052
viscosity at −20 °C	cSt	D 445
fuel system icing inhibitor (FSII)	vol %	D 5006
freeze point	°C	D 5972
aromatics	vol %	D 1319
saturates	vol %	D 1319
distillation (IBP)	°C	D 86
distillation (10%)	°C	D 86
distillation (50%)	°C	D 86
distillation (90%)	°C	D 86
distillation (EP)	°C	D 86
diesel fuel property	**units**	**ASTM method**
flash point	°C	D 93
density at 15 °C	kg/m^3	D 4052
viscosity at 40 °C	cSt	D 445
aromatics	vol %	GC/MS
cetane index	N/A	D 976
cloud point	°C	D 5773
pour point	°C	D 5949
distillation (IBP)	°C	D 86
distillation (10%)	°C	D 86
distillation (50%)	°C	D 86
distillation (90%)	°C	D 86
distillation (EP)	°C	D 86

to as "synthetic" in the following text and tables. A total of 78 blends were prepared by combining various amounts of these FT fuels with 39 randomly selected jet fuels and 39 randomly selected diesel fuels. The percent content of each FT fuel in each blend was either reported to us when the blend was collected from an external source or known by virtue of the ratio in which the blend was created in-house. Although the available neat FT fuels were produced as either jet or diesel fuels, there is some reason to suspect that there may be instances where FT jet fuels could become comingled with conventional diesel fuels and vice versa. One likely situation where this could occur in the Navy is when JP-5 jet fuel is downgraded and blended with shipboard diesel fuel for ship propulsion. This invariably occurs when an offloaded JP-5 fuel fails to meet thermal stability requirements because of the acquisition of copper after contact with copper-bearing shipboard fuel system components. Therefore, each of the neat FT fuels was evaluated in both the jet- and diesel-specific FT content models. FT blends were classified as either a jet or diesel fuel based on the classification of the petroleum fuel component of the blend. Because blends of FT jet in diesel and FT diesel in jet fuels were not primary concerns of the present work, training blends of these types were only analyzed to a minimum concentration of 30% and not included in the overall determination of the practical overall lower limit for FT fuel detection.

Spectra. NIR absorbance spectra were collected from the training set samples with the NFPM as described elsewhere.[1] The NFPM uses a NIR spectrometer (Bruker Optics, Inc., Billerica, MA) with a feedback-stabilized high-intensity tungsten halide lamp source and fiber-optic transflectance dip probe. The spectrometer employs a thermoelectrically cooled 512-element GaAs detector array, and the portion of each spectrum corresponding to the range of 1000−1600 nm was interpolated to provide 601 data points at a 1 nm resolution over this range. Spectra were sampled at a rate of 500 ms, and data acquisition and spectral preprocessing were performed with an Avantech TPC-1070 touch-screen computer with a 1 GHz Intel Celeron processor, running MS Windows XP Professional SP3. Data acquisition and processing were performed in the NFPM with software written in-house compiled from LabVIEW 8.5 (National Instruments Corporation, Austin, TX). Spectral data were referenced to an air background, baseline-corrected to points at 1000 and 1569 nm to eliminate slopes caused by optical dispersion, normalized to unit length, and mean-centered prior to all PLS model constructions.

Chemometric Analysis. PLS regression was performed to correlate the NIR spectra of the fuel samples to their known fuel property values and their known FT content. The numerical spectral, fuel property, and FT content data were imported into MATLAB R2008a (MathWorks, Inc., Natick, MA), where the spectra were assembled into matrices in which each row represented the NIR spectrum of a different fuel sample. PLS algorithms were developed with functionality provided by the PLS_Toolbox for MATLAB, version 4.2 (Eigenvector Research, Inc., Wenatchee, WA). Results of the PLS modeling of fuel properties and FT fuel contents were initially evaluated in terms of root-mean-square error of cross-validation (RMSECV) results obtained from leave-one-out cross-validation,[27] in which the predicted value of each sample in a given model is based on a submodel built from every other sample except the sample being given a prediction value, a technique which indirectly ascertains the performance of a given model with uncalibrated data. The number of constituent latent variables (LVs), the underlying linear factors to which the training data are converted for calibration purposes, for all percent content prediction models were chosen automatically using a statistic called the F test.[28−30] The F test was applied to the cross-validation results of the PLS fuel modeling[4,25,31] with an 85% confidence interval, unless otherwise noted, using a maximum of 10 LVs to maintain model versatility and utility in the presence of uncalibrated data. The F test, by limiting the number of LVs, protects against models that are too biased toward a specific set of training data, which would make the models themselves improperly biased against and less effective when predicting the fuel properties of uncalibrated data. A thorough analysis of this type of model bias, known as overfitting, can be found in previous work.[31]

Once the initial RMSECV-based evaluation was completed, each model was rebuilt to the parameters indicated by the cross-validation and all possible calibration data were re-introduced into the model to obtain the actual prediction values reported in the present work, hence, the reason that root-mean-square error of prediction (RMSEP) values, as opposed to RMSECV values, are reported as both intermediate and final results. It should be noted that the use of these RMSEP results was decided upon to provide consistent measures of model quality regardless of whether or not the samples used to produce the quality measures were included in the GA-augmented model construction process. The FT content of any given fuel sample, which was known at the time of sample collection as described previously, was expressed as a volume percent with a value ranging from 0 to 100.

GA-Augmented Modeling. For the fuel property modeling, the GAs were created to find final populations of sample sizes of 8, 16, 32, 64, 128, or 256, depending upon which would be the greatest number of samples that would still be less than 90% of the total number of available samples for a given fuel property. This sampling arbitrarily ensures the removal of at least 10% of the potential training data. The fitness condition evaluated during the course of these GA sample selections was the minimization of the RMSEP obtained from the property model.

The FT detection strategy inherently allows for more inter-analysis regularity and a secondary analysis goal besides basic

(27) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998; pp 93−94.
(28) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193–1202.
(29) Thomas, E. V. *J. Chemom.* **2003**, *17*, 653–659.
(30) Lin, W. Q.; Jiang, J. H.; Shen, Q.; Shen, G. L.; Yu, R. Q. *J. Chem. Inf. Model.* **2005**, *45*, 486–493.
(31) Kramer, K. E.; Johnson, K. J.; Cramer, J. A.; Morris, R. E.; Rose-Pehrsson, S. L. *Energy Fuels* **2008**, *22*, 523–534.

Table 2. Comparison of Fuel Property Modeling Results from PLS Models Derived from Entire Fuel Training Sets to Models Derived from GA-Selected Training Sets (n = Number of Samples)

	entire training sets				GA-selected training sets			
jet property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	352	0.77	3.39	8	256	0.83	2.92	10
density at 15 °C	130	0.97	0.002	7	64	0.98	0.001	9
viscosity at −20 °C	50	0.87	0.27	5	32	0.94	0.18	7
FSII	276	0.90	0.007	9	128	0.90	0.007	9
freeze point	376	0.61	2.34	7	256	0.68	2.12	9
aromatics	50	0.93	0.68	6	32	0.96	0.54	8
saturates	42	0.94	0.48	6	32	0.97	0.36	9
distillation (IBP)	279	0.79	5.7	8	128	0.82	5.4	10
distillation (10%)	279	0.91	3.1	9	128	0.93	2.9	10
distillation (50%)	279	0.91	2.7	8	128	0.92	2.5	9
distillation (90%)	279	0.60	4.4	7	128	0.67	3.9	9
distillation (EP)	279	0.56	5.5	7	128	0.64	5.0	8
	entire training sets				GA-selected training sets			
diesel property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	271	0.51	5.56	7	128	0.59	5.16	9
density at 15 °C	306	0.97	0.002	9	256	0.98	0.002	10
viscosity at 40 °C	276	0.89	0.14	9	128	0.92	0.12	9
aromatics	13	0.50	8.77	1	8	0.71	7.52	4
cetane index	281	0.93	0.98	4	128	0.94	0.90	7
cloud point	238	0.66	3.14	5	128	0.71	2.92	7
pour point	157	0.57	3.28	5	128	0.60	3.13	5
distillation (IBP)	203	0.11	11.7	1	128	0.53	8.7	10
distillation (10%)	222	0.74	6.9	7	128	0.80	6.1	8
distillation (50%)	222	0.84	4.6	8	128	0.88	3.9	9
distillation (90%)	298	0.51	7.1	6	256	0.59	6.5	7
distillation (EP)	297	0.41	8.5	5	256	0.50	7.9	7

accuracy, namely, a lower limit of detection. Therefore, more leeway was available to properly evaluate FT analysis sample population sizes and fitness conditions. The genetic models were created to find final populations of sample sizes of 64, 128, and 256 for each FT fuel blend ratio with petroleum fuel. Three fitness conditions were evaluated during the course of the FT GA sample selections for each FT fuel blend: (1) the minimization of the maximum FP result obtained from the identification model, (2) the minimum RMSEP obtained from the identification model, and (3) the minimum of the sum of these two values.

As described previously, models constructed through the use of GAs should still be evaluated in terms of how well they predict the fuel properties and FT contents of all possible fuels and not just those included in their construction. Therefore, the fitness tests used during GA model constructions could not be based on RMSECV values, which only exist for those samples used to construct any given PLS model. Instead, RMSEP values, recollected for all of the available samples that could have been used in each model, are used as fitness tests. For all GAs, regardless of final sample sizes or fitness conditions, after each population of "individuals" was created and initially evaluated for fitness, half of the population displaying below-median fitness was eliminated and the rest were paired and subsequently used to produce new "individuals" for the next iteration. This was performed by randomly combining the samples of each pair using the standard one-point crossover methodology,[32] where the sample data of each individual were divided into two randomly sized pieces and recombined with individuals with similarly divided data. Random mutations were introduced whenever the sample combination derived from two fit "individuals" yielded identical samples. This ensures not only population diversity but also that the final sample set has the desired number of samples. The repeated fitness evaluations were designed to artificially stop after 100 iterations to reduce the computational time, and each GA fitness evaluation was repeated 3 times to avoid local minimum results, as described previously.

(32) Whitley, D. *Stat. Comp.* **1994**, *4*, 65–85.

Results and Discussion

Fuel Property Modeling. Table 2 shows the model prediction results of all available samples after using standard (i.e., all available) and GA-selected sample populations as training data. The results indicate that the application of GAs lowers the resulting RMSEP results and increases the correlations of predicted versus measured property values, as measured by the linear correlation coefficients (R^2) for each jet and diesel fuel property. It is also evident from the results in Table 2 that GA refinement of the training sets yielded significant improvements in the precision of the predicted properties. Note that for 22 of 25 properties, the F-test procedure that was used to determine the model size, i.e., the number of LVs that are used in a particular PLS model, chose a larger number of LVs, which, in itself, can improve correlations between measured and predicted values. It is recognized that incorporating too much of the linear variance in the model by using more LVs can result in a type of model bias known as overfitting, in which the model is overly specified toward its training data and, thus, rendered less effective with respect to non-training data. To obtain PLS fuel property models that can operate with new fuels that are not part of the training set, a balance must be struck between including sufficient detail to define the relationship between the data and the property of interest and avoiding overfitting. This is referred to as robustness in the present context. Therefore, the key issue for the successful implementation of GAs to improve fuel property modeling precision will be whether more LVs can be used with GA-selected training data sets, without overfitting and, thus, maintaining robustness for use with uncalibrated fuel samples.

To evaluate standard and GA model sizes in terms of robustness, the standard populations of samples were modeled using the GA-selected number of LVs and vice versa.

Table 3. Comparison of Fuel Property Modeling Results from PLS Models Derived from Entire Fuel Training Sets and GA-Selected Number of LVs to Models Derived from GA-Selected Training Sets and LVs Selected from All Samples (n = Number of Samples)

	models from all samples + GA-selected LVs				models from GA-selected samples + LVs from all samples			
jet property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	352	0.81	3.09	10	256	0.77	3.37	8
density at 15 °C	130	0.98	0.001	9	64	0.97	0.002	7
viscosity at −20 °C	50	0.92	0.22	7	32	0.86	0.29	5
FSII	276	0.91	0.007	9	128	0.90	0.007	9
freeze point	376	0.96	0.50	8	256	0.61	2.37	7
aromatics	50	0.98	0.31	9	32	0.89	0.86	6
saturates	42	0.46	1.75	3	32	0.94	0.47	6
distillation (IBP)	279	0.84	5.0	10	128	0.76	6.2	8
distillation (10%)	279	0.93	2.9	10	128	0.90	3.4	9
distillation (50%)	279	0.92	2.6	9	128	0.91	2.8	8
distillation (90%)	279	0.67	3.9	9	128	0.59	4.4	7
distillation (EP)	279	0.60	5.3	8	128	0.56	5.5	7
	models from all samples + GA-selected LVs				models from GA-selected samples + LVs from all samples			
diesel property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	271	0.60	5.03	9	128	0.47	5.85	7
density at 15 °C	306	0.98	0.001	10	256	0.97	0.002	9
viscosity at 40 °C	276	0.89	0.14	9	128	0.92	0.12	9
aromatics	13	0.76	6.12	4	8	0.50	9.36	1
cetane index	281	0.94	0.88	7	128	0.93	0.99	4
cloud point	238	0.71	2.91	7	128	0.64	3.24	5
pour point	157	0.57	3.28	5	128	0.60	3.13	5
distillation (IBP)	203	0.55	8.3	10	128	0.11	11.7	1
distillation (10%)	222	0.77	6.5	8	128	0.75	6.8	7
distillation (50%)	222	0.86	4.2	9	128	0.86	4.3	8
distillation (90%)	298	0.56	6.7	7	256	0.49	7.3	6
distillation (EP)	297	0.49	7.9	7	256	0.38	8.8	5

The results for this operation are shown in Table 3. Comparison of Tables 2 and 3 indicates that, in each of the 22 cases, where the number of LVs was increased after the use of GAs, a lower RMSEP value and a higher R^2 value were obtained with respect to the results from either sample population without GAs. In addition, it is apparent that the number of samples, when using the GA-selected number of LVs, was much less critical for the improvement of net accuracy than the number of LVs themselves. The overall linear correlations between measured and predicted properties (R^2) improved upon use of the GA-selected samples in 11 cases, became worse in 6 cases, and showed insufficient change to determine either event in 8 cases. The RMSEP values, meanwhile, improved with the use of the GA-selected samples in 13 cases but were worse in 11 cases, with 1 result showing insufficient change to determine either event. In addition, when using the standard number of LVs for the GA-selected samples, the modeling results showed a net loss from those using the standard number of samples with the same number of LVs.

These results imply that simply choosing less restrictive *F*-test confidence intervals and, consequently, selecting a greater number of LVs for resulting models would provide similar model improvements to those seen with the use of GA-selected training samples. An attempt to reproduce the improved LV selection in this fashion can be seen in Table 4. These results, in contrast, indicate that the LV selection process is not replicable through this straightforward parameter modification. Even in the best-case scenario, when the 60% confidence interval is used, the desired number of LVs is selected only a little over half of the time. To investigate this seeming discrepancy, the number of LVs selected must be interpreted in terms of how the GAs actually affects the LV selection.

A closer examination of the cumulative predicted residual sum of squares (PRESS) results obtained through cross-validation is necessary to understand why a larger number of LVs is chosen when using GA-selected training sample subsets and how these results are distinct from those produced by the standard sample subsets, regardless of the *F*-test confidence interval chosen. Figure 2 shows the cumulative PRESS results of four data sets (flash point and density, each for both jet and diesel fuels). Note that, although the sample subsets resulting from GAs all change the resulting cumulative PRESS profiles differently, what they all have in common is that the changes are more apparent with models derived with a smaller number of LVs. It should be noted here that this illustrates why a larger, GA-derived number of LVs can be used while modeling the total sample population with no net ill effects, because the errors using either sample population become more comparable as the number of LVs increases.

The *F* test is used to determine the smallest number of LVs that still produces cumulative PRESS results that are statistically indistinguishable from the number of LVs that truly produces the smallest cumulative PRESS result. The algorithm used to perform the *F* test[31] uses a ratio of the cumulative PRESS values resulting from all possible numbers of LVs over the minimum available cumulative PRESS value. When the GA-selected sample subsets produce cumulative PRESS trends that show greater improvements with a lower number of LVs, the net effect on the relevant PRESS ratio trend is a "flattening" of said trend. This increases the relevance of the modeling contributions of less prominent LVs and allows, in turn, the selection of a larger number of LVs using the *F* test.

However, this still does not explain why the cumulative PRESS results improve more dramatically when using smaller models with a lower number of LVs, as opposed to large models with more LVs, and consequently, why the use of GAs in this fashion allows for the confident selection of a

Table 4. Automated LV Determinations for All Jet and Diesel Fuel Properties Using All Available Training Samples and Various *F*-Test Confidence Intervals with a Maximum of 10 LVs[a]

	F test							
jet fuel properties	85%	80%	75%	70%	65%	60%	55%	50%
flash point	8	8	8	9	9	***10***	***10***	***10***
density at 15 °C	7	7	8	8	8	***9***	***9***	10
viscosity at −20 °C	5	5	5	6	6	***7***	8	9
FSII content	***9***	***9***	***9***	***9***	***9***	***9***	***9***	10
freeze point	7	7	8	8	8	***9***	***9***	***9***
aromatics	6	6	7	7	7	7	7	***8***
saturates	6	7	7	7	7	***9***	***9***	***9***
distillation (initial point)	8	9	9	9	***10***	***10***	***10***	***10***
distillation (10% point)	9	9	9	***10***	***10***	***10***	***10***	***10***
distillation (50% point)	8	***9***	***9***	***9***	***9***	10	10	10
distillation (90% point)	7	8	8	9	9	9	10	10
distillation (end point)	7	7	***8***	***8***	***8***	9	9	10

	F test							
diesel fuel properties	85%	80%	75%	70%	65%	60%	55%	50%
flash point	7	8	8	8	8	9	9	10
density at 15 °C	9	9	9	9	9	9	***10***	***10***
viscosity at 40 °C	***9***	***9***	10	10	10	10	10	10
aromatics	1	1	1	1	1	1	1	3
cetane index	4	4	4	4	5	5	5	6
cloud point	5	5	5	6	6	6	***7***	7
pour point	***5***	***5***	***5***	***5***	***5***	6	7	7
distillation (initial point)	1	1	6	6	6	9	9	10
distillation (10% point)	7	7	***8***	***8***	***8***	***8***	***8***	10
distillation (50% point)	8	***9***	***9***	***9***	10	10	10	10
distillation (90% point)	6	6	*7*	*7*	*7*	*7*	8	9
distillation (end point)	5	5	6	6	*7*	*7*	*7*	8
percent target LVs (%)	12	20	28	36	44	56	52	40

[a] Target LV values, found during the performance of the GA, are shown in bold italics.

larger number of LVs. To determine this, the outlier diesel fuel population, described above in the Experimental Section, was added to the standard fuel sample set prior to performing the GA sample selection. As can be seen in Table 5, the GA-based selection scheme tended to remove the ill-conditioned sample population preferentially. In the case of all available diesel fuel properties, GAs removed at least half of the data known to be unsuitable for modeling fuel properties from the final selected training data population.

Thus, the true utility of the GA is to remove the most deviant samples, thereby allowing the models to be constructed from a training data set that is more highly correlated with the property of interest. This would most acutely affect models using a smaller number of LVs, because they focus on the more obvious underlying linear variances. A major benefit of this approach is that it allows for improvements in prediction precision by confidently producing models with a larger number of LVs, without overfitting to any greater extent than what would occur with models derived from a smaller number of LVs without GA refinement. Because the purpose of the *F* test is to systematically determine the largest number of LVs that a model can support without risking overfitting, the primary utility of GAs is the removal of those samples that would most likely hinder the optimal performance of the *F* test. Furthermore, the use of GAs to perform this removal simplifies and automates the operation. In combination with the automated *F* test, the use of GAs to produce model improvements results in a non-expert operation that can be performed without any *a priori* knowledge of what types of samples should be removed.

On the basis of the results discussed thus far, it could be concluded that, once the GA-augmented modeling has concluded what number of LVs to use, then the actual samples selected through the use of GAs to obtain this number can be discarded in favor of the entire sample population. Comparisons can be made between the "GA-selected training sets" results in Table 2 and the "models from all samples + GA-selected LVs" results in Table 3 to assess the consequences of doing this. With respect to RMSEP, the decision on whether to use the entire data set or the GA-selected subset with the GA-selected number of LVs is completely ambiguous, because 10 properties show lower RMSEP results in one circumstance, 10 are lower in the other circumstance, and 5 results show no change in RMSEP regardless of data set used. When the R^2 values are compared, however, a slight advantage is seen when using the GA-selected samples with the GA-selected LVs because 10 of the results show a higher degree of linearity than when using all of the samples. This is as opposed to the 7 results that are more linear when using all of the samples and the 8 R^2 results that are not significantly affected. While the model improvements that can be had using the GA-selected samples are rather modest compared to those obtained by increasing the number of LVs alone, they are quantifiable enough that they should not be discarded. Therefore, for the remainder of the present work, GA-selected samples will be used in concert with the GA-selected number of LVs.

The absence of overfitting is already indicated by the fact that the models constructed from the GA-selected training samples optimized the modeling results obtained from using all available training samples. However, an additional test of the utility of the technique with uncalibrated fuels should be performed to confirm this assessment. Such a test will consist of models constructed and refined from only a subset of the available samples and evaluations of these models using only the samples that were not included in the calibration. In this case, the data, which are continuously collected in our laboratory as a sequentially numbered series, are divided into even- and odd-numbered components. As the samples arrive in batches that are unpredictable in both point of origin and content, the separation of the data into even and odd samples effectively randomizes the two populations. The even-numbered data are used to produce models with and without the benefit of GAs, and the odd-numbered data are then evaluated in terms of property predictions using these models. Table 6 shows the results of this operation. It should first be noted that the RMSEP values produced using the even-numbered data models on the odd-numbered data are, in almost all cases, worse than those produced when constructing and validating models using the same data, as can be seen by comparing the results of Table 6 to those found in Table 2. As has been noted previously by our laboratory,[33] the use of smaller data sets can reduce the comprehensiveness of resulting models. However, note also that Table 6 shows that the application of GAs to model construction improves the prediction results obtained from the uncalibrated, odd-numbered data in the case of all jet fuel properties and a majority of diesel fuel properties. In the case of the five underperforming diesel modeling results, three of the GA

(33) Cramer, J. A.; Kramer, K. E.; Johnson, K. J.; Hammond, M. H.; Rose-Pehrsson, S. L.; Morris, R. E. *Proceedings of the Annual Conference of the International Association on the Stability, Handling, and Use of Liquid Fuels* (*IASH*); Tucson, AZ, Oct 2007.

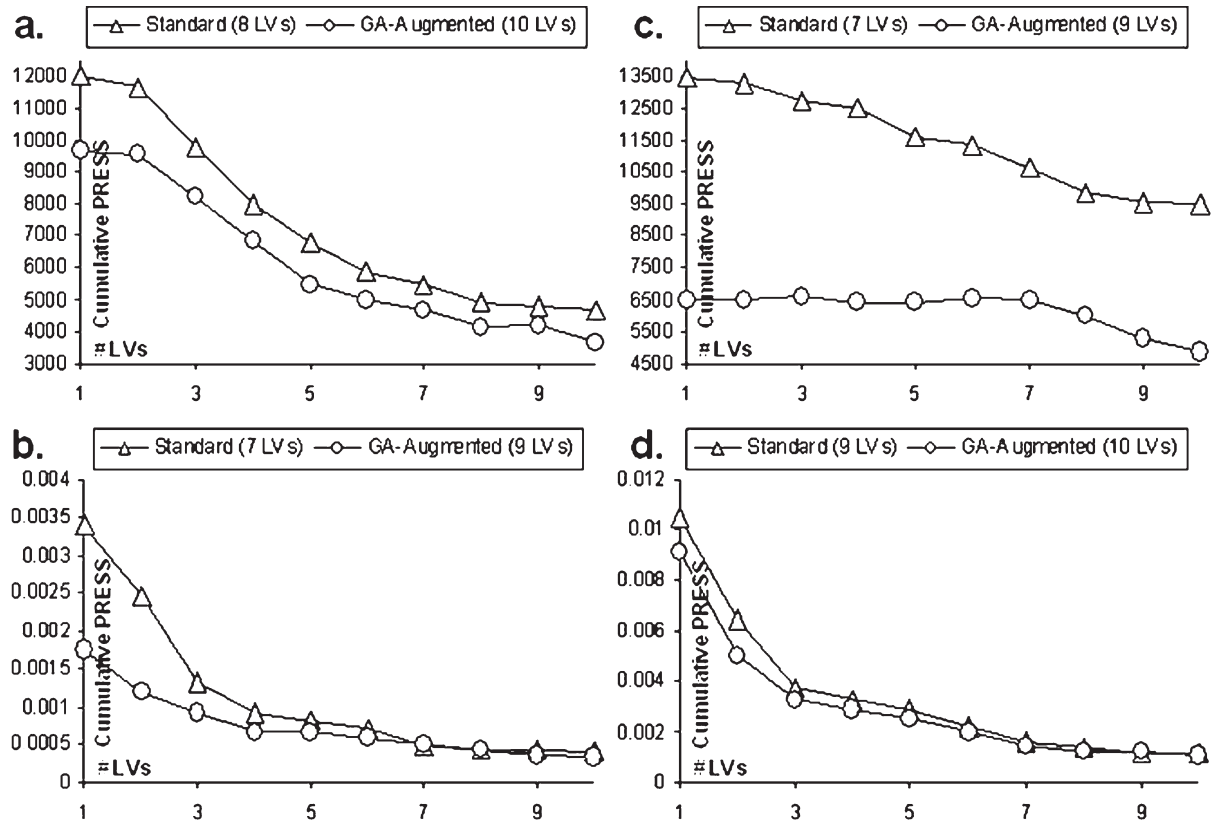

Figure 2. Cumulative PRESS plots and the number of LVs chosen with an 85% confidence interval and a maximum of 10 LVs manually defined for a *F*-test, evaluating the following properties: (a) flash point for jet fuel, (b) density for jet fuel, (c) flash point for diesel fuel, and (d) density for diesel fuel.

Table 5. GA Sample Selection Results after the Addition of a Known Outlier Population to the Standard Diesel Fuel Population

diesel fuel property	number of additional samples from outliers	number of samples removed through GA	number of outliers removed through GA	percentage of outliers removed by GA (%)
flash point	59	74	49	83
density at 15 °C	60	110	40	67
viscosity at 40 °C	57	77	46	81
cetane index	57	82	42	74
cloud point	7	117	5	71
pour point	51	80	32	63
distillation (initial point)	9	84	6	67
distillation (10% point)	8	102	5	63
distillation (50% point)	9	103	5	56
distillation (90% point)	13	55	9	69
distillation (end point)	13	54	9	69

RMSEP values are greater than the non-GA values by less than 5% and one model is based only on four fuels, which may be insufficient to truly gauge model quality. The remaining underperforming result is predicting the IBP distillation of a diesel fuel as typically assessed through ASTM D 86. While this single result serves as an important cautionary note with respect to GA-augmented model construction, it cannot reasonably be concluded from said result that the use of GAs to improve fuel property models based on all of the available data is an inadequate analysis strategy. It should finally be noted that the prediction of R^2 values decreases upon applying GAs to model construction in 3 of 25 cases; however, 2 of these cases result in decreases of only 0.02, and the third case, which is again a measure of diesel IBP distillation prediction quality, shows a decrease from 0.09 to 0.05. As with the RMSEP results, these minor decreases in prediction linearity do not negate the net model improvements that can be obtained using GAs.

FT Fuel Modeling. During the FT fuel trials, two new sets of training sample populations were obtained using GAs: one for use in the constrained FT detection methodology described previously[4] and summarized above, where each FT fuel is detected in a specified order, and a second set that is unconstrained and assumes that the individual model pairs will be used in no particular order. The FT fuel modeling that is constrained to detect fuels in a specific order because of a built-in order of operation will be referred to as an "in-order" modeling strategy, as opposed to an "out-of-order" strategy that would possess no such order of operation. The "in-order" results thus reported were used to directly lower the limits of detection in the ordered methodology that is already in use in our equipment, while the "out-of-order" results were used to explore the possibility of detecting each FT fuel in the potential presence of every other FT fuel, so that FT fuel concentrations from different sources could be quantified simultaneously. To avoid confusion regarding the content of

Table 6. Effects of GA Sample Selection When Using Only Even-Numbered Training Samples To Predict the Fuel Properties of Only Odd-Numbered Validation Samples (n = Number of Samples)

	entire training sets				GA-selected training sets			
jet property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	176	0.70	3.92	7	128	0.72	3.85	10
density at 15 °C	61	0.94	0.002	7	32	0.94	0.002	9
viscosity at −20 °C	24	0.67	0.42	4	16	0.76	0.35	6
FSII	136	0.87	0.009	9	64	0.85	0.009	8
freeze point	188	0.53	2.54	6	128	0.60	2.42	8
aromatics	24	0.10	2.63	1	16	0.89	1.18	6
saturates	21	0.69	1.19	5	16	0.72	1.14	5
distillation (IBP)	139	0.71	6.7	7	64	0.73	6.5	8
distillation (10%)	139	0.88	3.8	7	64	0.90	3.5	9
distillation (50%)	139	0.86	3.4	7	64	0.87	3.2	9
distillation (90%)	139	0.35	5.5	5	64	0.48	5.2	8
distillation (EP)	139	0.36	6.7	5	64	0.45	6.5	9
diesel property	n	R^2	RMSEP	LVs	n	R^2	RMSEP	LVs
flash point	133	0.24	7.19	2	64	0.43	6.34	8
density at 15 °C	153	0.95	0.002	9	128	0.96	0.002	10
viscosity at 40 °C	137	0.78	0.20	8	64	0.83	0.18	9
aromatics	7	0.24	12.88	1	4	0.29	14.12	2
cetane index	140	0.91	1.04	4	64	0.91	1.05	7
cloud point	116	0.66	3.22	4	64	0.64	3.37	10
pour point	80	0.08	5.24	1	64	0.41	4.10	7
distillation (IBP)	99	0.09	12.1	1	64	0.05	15.1	8
distillation (10%)	109	0.54	9.1	5	64	0.58	8.9	9
distillation (50%)	109	0.56	8.1	5	64	0.66	7.0	9
distillation (90%)	147	0.39	8.0	4	128	0.49	7.4	8
distillation (EP)	147	0.34	8.6	3	128	0.43	8.2	7

Table 7. Fuel Types That Were Used To Calibrate Each PLS Model, Including a Breakdown of Identification and Quantification Modeling

included in model training data?	neat petrochemical fuels	synthetic diesel fuel and blends	CTL diesel fuel and blends	CTL jet fuel and blends	GTL jet fuel and blends
all property models	yes	no	no	no	no
		In-Order FT Models			
synthetic diesel (identification)	yes	yes	yes	yes	yes
synthetic diesel (quantification)	no	yes	no	no	no
CTL diesel (identification)	yes	no	yes	yes	yes
CTL diesel (quantification)	no	no	yes	no	no
CTL jet (identification)	yes	no	no	yes	yes
CTL jet (quantification)	no	no	no	yes	no
GTL jet (identification)	yes	no	no	no	yes
GTL jet (quantification)	no	no	no	no	yes
		Out-of-Order FT Models			
synthetic diesel (identification)	yes	yes	yes	yes	yes
synthetic diesel (quantification)	no	yes	no	no	no
CTL diesel (identification)	yes	yes	yes	yes	yes
CTL diesel (quantification)	no	no	yes	no	no
CTL jet (identification)	yes	yes	yes	yes	yes
CTL jet (quantification)	no	no	no	yes	no
GTL jet (identification)	yes	yes	yes	yes	yes
GTL jet (quantification)	no	no	no	no	yes

each in-order and out-of-order FT fuel model, Table 7 has been included. It should be noted here that this table also lists the identification and quantification model pairs in the order in which FT fuels are identified and quantified in the standard in-order detection scheme.

The new fuel populations produced through the use of GAs were used instead of the full sample populations in the FT identification models, including the model pairing and order-of-detection schemes. As mentioned previously, the quantification model sample populations, being much smaller, were not altered. Results were collected when using the identification model only and using the identification and quantification models together as they would be used together in the detection scheme described previously. In other words, the "identification + quantification models" results in the tables to be described show limits of detection with and without the use of the unchanging quantification model as an additional filter. Biodiesel samples were included in the diesel fuel analyses.

In-Order FT Modeling. Table 8 contains the following in-order results with and without GA-selected results. The minimum cutoff is the lowest threshold detection value that prevented FP results from being obtained at the end of the analysis. A comparison of the minimum cutoff values obtained before and after GA data reduction shows that the results are improved with the use of GA-selected samples. The minimum detectable FT content is the lowest FT content sample that could be reliably detected using the assigned

Table 8. Impact of GAs on In-Order Sequential FT Fuel Detection in Blends with Jet and Diesel Fuels

Jet Fuel Blends						
identification model	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	5	30	2	30	256	minimum FP
CTL diesel	14	30	5	30	256	minimum FP + RMSEP
CTL jet	6	8	4	5	256	minimum FP
GTL jet	9	10	8	9	256	minimum FP
identification + quantification models	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	5	30	1	30	256	minimum FP
CTL diesel	7	30	5	30	256	minimum FP + RMSEP
CTL jet	6	7	4	5	256	minimum FP
GTL jet	9	10	8	9	256	minimum FP
Diesel Fuel Blends						
identification model	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	8	9	1	3	256	minimum FP
CTL diesel	19	40	2	7	256	minimum FP + RMSEP
CTL jet	13	30	5	30	128	minimum FP + RMSEP
GTL jet	8	30	6	30	128	minimum FP
identification + quantification models	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	8	9	1	3	256	minimum FP
CTL diesel	19	40	2	10	256	minimum FP + RMSEP
CTL jet	6	30	4	30	128	minimum FP + RMSEP
GTL jet	8	30	6	30	128	minimum FP

cutoff value. Results are shown for models produced before and after the GA data reduction strategy was applied. In most cases, the minimum detectable FT content was the FT content in the sample that was immediately greater than the minimum cutoff value, which is indicative of an acceptable final model quality. The primary exception to this is in the detection of CTL diesel fuel in petroleum diesel fuel, which required a 2% cutoff value but left samples containing between 2 and 7% CTL FT fuel unreliably detected in the identification model (see Figure 3 for details and also note the similarity to the situation found in Figure 1). This is further complicated by the fact that the quantification model in this detection step has a difficulty with a 9% FT sample, impairing the lower limit of detection even further. The GA sample number is the number of samples used in the most useful GA results. Values of 64, 128, and 256 were evaluated, with larger numbers generally producing the best results and 64 samples being determined to be inadequate in all cases. The fitness of any given sample population was evaluated by the three methods described previously: (1) minimization of FP, (2) minimum RMSEP, and (3) minimum sum of these two values. The most useful evaluation strategy for each method is given in the GA fitness evaluator column. Note that the minimization of the RMSEP result alone did not produce the lowest limit of detection in any of the in-order FT fuel analyses. This is reasonable, because the other two analysis options address limits of detection more directly.

The RMSEP results of the indicated sample populations, calculated from all available samples or only those that contain FT fuel, for both the identification and quantification models are shown in Table 9. It is interesting to note that the use of GAs tends to produce models that result in higher RMSEP values than their standard counterparts. For instance, when assessing all of the data, six of the eight identification models are seemingly less accurate after the use of GAs. When assessing the FT fuel population only, this number decreases to five of eight models but remains a majority. Keep in mind, however, that higher RMSEP results do not necessarily result in a less effective overall analysis because the use of the quantification models serves as a confirmation step for all identification model results. On a related note, Table 9 indicates that the RMSEP values from the GA-augmented identification models with respect to FT-bearing fuels were never as low as those produced for the same samples using the entire training set, further indicating that the quantification models are still necessary in the overall detection scheme.

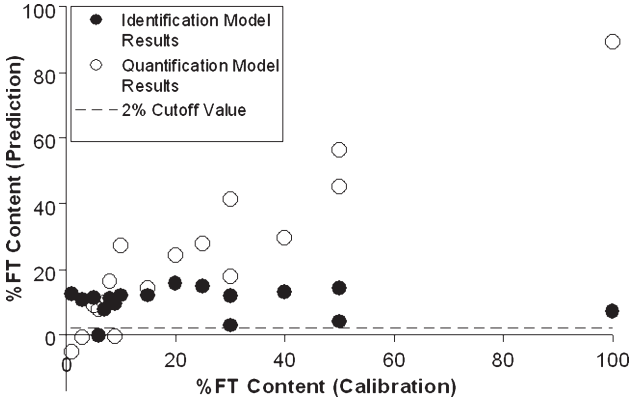

Figure 3. Modeling results of detecting CTL FT diesel fuel in petroleum diesel fuel (both identification and quantification model results shown). Results for samples containing no FT fuel (i.e., 0% calibration FT content) were not plotted to clarify the figure.

Out-of-Order FT Modeling. Additional fuel training sample populations, produced through the use of GAs, were constructed under the assumption that all of the available FT fuels could be present during each of the individual FT fuel

Table 9. Impact of GAs on RMSEP of In-Order Sequential Modeling for FT Fuel Detection in Jet and Diesel Fuel Blends

Jet Fuel Blends				
identification model only	identification model, all	identification model FT, only	identification model (GA), all	identification model (GA), FT only
synthetic diesel	1.41	12.19	2.15	22.07
CTL diesel	1.81	9.09	1.78	9.62
CTL jet	1.53	1.94	1.49	2.09
GTL jet	3.04	7.58	3.31	6.57
quantification model only	quantification model, all	quantification model FT, only	quantification model (GA, no effect), all	quantification vmodel (GA, no effect) FT only
synthetic diesel	16.30	0.037	16.30	0.04
CTL diesel	16.68	1.03	16.68	1.03
CTL jet	7.18	1.35	7.18	1.35
GTL jet	10.90	6.34	10.90	6.34
Diesel Fuel Blends				
identification model only	identification model, all	identification model FT, only	identification model (GA), all	identification model (GA), FT only
synthetic diesel	1.32	4.74	3.21	15.45
CTL diesel	3.72	10.74	5.94	28.96
CTL jet	1.98	7.21	2.18	4.25
GTL jet	2.78	11.28	3.23	9.44
quantification model only	quantification model, all	quantification model FT, only	quantification model (GA, no effect), all	quantification model (GA, no effect) FT only
synthetic diesel	7.15	0.96	7.15	0.96
CTL diesel	15.15	8.10	15.15	8.10
CTL jet	16.34	0.24	16.34	0.24
GTL jet	25.69	0.79	25.69	0.79

Table 10. Comparison of the Out-of-Order Detection of Each FT Fuel in the Presence of Any Other FT Fuel When Blended with Jet and Diesel Fuels, with and without GA Selection

Jet Fuel Blends						
identification model	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	5	30	2	30	256	minimum FP
CTL diesel	30	50	7	30	128	minimum FP + RMSEP
CTL jet	8	9	4	5	256	minimum FP
GTL jet	10	15	8	9	256	RMSEP
identification + quantification models	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	5	30	1	30	256	minimum FP
CTL diesel	30	50	5	30	128	minimum FP + RMSEP
CTL jet	8	9	4	5	256	minimum FP
GTL jet	10	15	7	8	256	RMSEP
Diesel Fuel Blends						
identification model	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	8	9	1	3	256	minimum FP
CTL diesel	17	40	9	10	256	minimum FP + RMSEP
CTL jet	13	30	11	30	256	RMSEP
GTL jet	15	30	8	30	128	minimum FP
identification + quantification models	minimum cutoff (%)	minimum detectable FT content (%)	minimum cutoff (GA) (%)	minimum FT content (GA) (%)	GA sample number	GA fitness evaluator
synthetic diesel	8	9	1	3	256	minimum FP
CTL diesel	17	40	9	10	256	minimum FP + RMSEP
CTL jet	7	30	6	30	256	RMSEP
GTL jet	15	30	8	30	128	minimum FP

assessment steps and, therefore, must be accommodated by the training data. The GA-selected training samples were used instead of the full sample populations to retrain the identification models, but the individual FT fuel assessments were otherwise the same as seen above, i.e., consisting of identification and quantification model pairs. However, to allow for out-of-order modeling, the sequential elimination of each FT fuel from the overall fuel population was no longer performed to potentially detect each FT fuel in the presence of every other FT fuel.

Table 11. Impact of GA-Selected Training Set Data on the RMSEP for the Out-of-Order Detection of Each FT Fuel in the Presence of Any Other FT Fuel When Blended with Jet and Diesel Fuels

Jet Fuel Blends				
identification model only	identification model, all	identification model FT, only	identification model (GA), all	identification model (GA), FT only
synthetic diesel	1.41	12.19	2.15	22.07
CTL diesel	2.43	11.17	2.42	8.65
CTL jet	1.74	2.54	2.10	7.46
GTL jet	3.43	9.31	3.33	8.12
quantification model only	quantification model, all	quantification model FT, only	quantification model (GA, no effect), all	quantification model (GA, no effect) FT only
synthetic diesel	16.30	0.04	16.30	0.04
CTL diesel	17.58	1.03	17.58	1.03
CTL jet	7.43	1.35	7.43	1.35
GTL jet	15.04	6.34	15.04	6.34
Diesel Fuel Blends				
identification model only	identification model, all	identification model FT, only	identification model (GA), all	identification model (GA), FT only
synthetic diesel	1.32	4.74	3.21	15.45
CTL diesel	4.12	13.41	4.80	21.74
CTL jet	2.23	8.91	2.00	8.14
GTL jet	3.96	24.84	9.78	4.32
quantification model only	quantification model, all	quantification model FT, only	quantification model (GA, no effect), all	quantification model (GA, no effect) FT only
synthetic diesel	7.15	0.96	7.15	0.96
CTL diesel	17.47	8.10	17.47	8.10
CTL jet	17.58	0.24	17.58	0.24
GTL jet	27.49	0.79	27.49	0.79

The out-of-order test results are shown in Table 10. A comparison to the in-order results in Table 8 indicates that, overall, both the jet and diesel fuel out-of-order FT detection performed as well as if not slightly better than the in-order FT detection strategy that is currently being used when GAs are applied to both analyses. For instance, when the jet fuel identification + quantification limit of detection results for GTL jet from both tables are compared, it can be seen that the minimum detectable FT content improves from 9 to 8%. Also, although the diesel fuel identification + quantification limit of detection for CTL diesel increases from 7 to 10%, the fact that the cutoff value is 9%, as opposed to the 2%, shown in Table 8 and Figure 3, possibly indicates a more reliable model. Although the precision of the detection of FT jet fuel in conventional diesel fuel was decreased somewhat, this is not anticipated to be a major hindrance in practice and it also does not adversely affect the predicted 10% limit of detection derived from the in-order results. As discussed above, a larger number of model calibration samples is preferred when GA sample optimizations are performed.

Table 11 shows the same RMSEP results seen in Table 9 but for the out-of-order modeling instead of the in-order modeling. It is interesting to note that, on the basis of the RMSEP values shown in this table, the directly RMSEP-based fitness test can serve as a viable method for evaluating GA fitness in these out-of-order trials. The fact that this never occurred in the in-order trials suggests that the more complex sample populations in the out-of-order GA treatment required a more thorough model evaluation criteria than simply evaluating the minimum FP results. The overall RMSEP results obtained from the out-of-order GA trials were roughly comparable to those produced in the in-order trials. As was seen in the in-order modeling, identification model results were worse in a majority of cases, in five of eight cases with respect to all samples and four of eight cases with respect to only the FT samples. However, it should finally be noted, in directly comparing Tables 9 and 11, that the in-order trials produced RMSEP results that were roughly on par with the out-of-order trials.

Conclusions

The findings of this study illustrate the benefits of a GA-based approach to refine training data sets to improve the predictive power of both fuel property and FT fuel content PLS models. The primary advantage gained from the use of GAs in this context is the confident selection of larger numbers of latent variables without increasing the risk of overfitting. The fact that these improvements can be obtained in an automated fashion without any *a priori* knowledge of sample quality or additional sample collections is also considered a major advantage.

If the minimum combined identification and quantification cutoff values from Table 8 are used in the standard FT fuel detection scheme, as low as 5% CTL FT jet fuel could be detected in jet fuel blends, as low as 9% GTL jet fuel could be detected in jet fuel blends, as low as 3% of the ethylene-derived synthetic diesel fuel could be detected in diesel fuel blends, and as low as 10% CTL FT diesel fuel could be detected in diesel fuel blends. Therefore, the practical overall lower limit of detection for FT fuels in fuel blends after the implementation of the GA-selected sample subsets could be considered to be 10%.

The use of GA refinement of the model training data also provides the means to model and detect each FT independently

in comingled fuels containing multiple FT fuels from different sources, with the same 10% limit of detection as obtained with the single FT detection scheme using GA-selected training data. Although this multiple FT fuel modeling methodology may slightly decrease the overall precision of detecting single FT fuels in conventional fuels, it is believed that this is more than compensated for by the ability to simultaneously detect the presence of multiple FT fuels.

Acknowledgment. The authors thank the National Research Council (NRC) and the Office of Naval Research (ONR) for supporting this work.