Fuel 310 (2022) 122274

Contents lists available at ScienceDirect

Fuel

journal homepage: www.elsevier.com/locate/fuel

Full Length Article

Multi-output chemometrics model for gasoline compounding

Harbil Bediaga [a], María Isabel Moreno [b], Sonia Arrasate [b], José Luis Vilas [a], Lucía Orbe [c], Elías Unzueta [c], Juan Pérez Mercader [d], Humberto González-Díaz [b,e,f,*]

[a] Department of Physical Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain
[b] Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain
[c] Petronor Innovación S.L., 48550 Muskiz, Spain
[d] Santa Fe Institute, Santa Fe, NM 85452, USA
[e] Basque Center for Biophysics, CSIC-UPV/EHU, 48940 Leioa, Spain
[f] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

ARTICLE INFO

Keywords:
Fuel
Gasoline
Blend design
Chemometrics
Multi-output Machine Learning
Perturbation Theory
Discriminant Analysis
Artificial Neural Networks

ABSTRACT

Computational models may help to reduce research cost by predicting properties of alternative blends. Nowadays, most efforts focus on prediction of a few properties for sets of gasoline samples. However, there are no reports of models able for classification of gasoline samples with multiple output properties measured in real life refinery plants. In this work, Information Fusion (IF), Perturbation Theory (PT), and Machine Learning (ML) algorithm (IFPTML) was used to model real production data with >230,000 outcomes gathered from a petroleum refinery plant. IF-pre-processing phase assembled the working dataset with 44 physicochemical output properties *vs*. 574 input variables of 4 production lines distributed in 26 data blocks including 14 different streams and 23 operations carried out in the plant. PT-calculation phase quantifies the effect of perturbations (deviations) in all input variables using PT Operators. Last, in ML-analysis phase involved Linear Discriminant Analysis (LDA) and Artificial Neural Networks (ANN) models training. IFPTML-LDA model presented AUROC = 0.936 with overall Sensitivity Sn and Specificity Sp ≈ 84–91% for training and validation sets. In internal control experiment we obtained an IFPTML-FT-NIR model with similar Sn and Sp ≈ 86–97%, for >25,000 values of 16 properties measured FT-NIR technique; demonstrating the robustness of the algorithm to changes on the experimental techniques used. This model could be useful for the design of new alternatives blends (biofuels, refuse-derived fuels, *etc*.) with lower environmental impact.

1. Introduction

Worldwide energy demand is increasing and pollution problems grow with it. This problem creates the need to find a sustainable and renewable way to meet the demand. There are several options to find a way out of this problem. The composition, parameters, and standard limits established for gasoline properties may vary from region to region. The different types of gasoline and blending are regulated by agencies such as American Society for Testing and Materials (ASTM), European committee for Standardization (CEN), Bureau of Indian Standards (BIS), *etc* [1].

In this context, we can try to carry out gasoline compounding with alternative fuel sources. For instance, the use of gasoline blends together with methanol has many advantages compared to other liquid fuels since it can reduce NO_x and CO_2 emissions, increase the octane number and steam pressure, *etc* [2]. It has several disadvantages such as being toxic and corrosive in some materials [3]. Biofuels [4] obtained from processes of conversion of biomass or waste from plants, animals and/or agriculture has become an interesting alternative. Particularly, ethanol-reach alternative biofuels have a Research Octane Number (R.O.N.) of 108–109 and a Motor Octane Number (M.O.N.) of 91. These values are greater than the required for gasoline (R.O.N. equal or greater than 95 and M.O.N. equal or greater than 85) [6]. A third alternative, is the use of refuse-derived fuel [5] produced from sundry types of waste and by-products as non-recyclable plastics, labels and other corrugate material, is another alternative for gasoline compounding.

Taking everything into account, the discovery, characterization, optimization, and introduction to production lines of these alternatives for gasoline compounding is a long and costly process. It involves the consideration of a high number of combinations of gasoline blend

* Corresponding author at: Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain.
E-mail addresses: lucia.orbe@repsol.com (L. Orbe), humberto.gonzalezdiaz@ehu.es (H. González-Díaz).

https://doi.org/10.1016/j.fuel.2021.122274
Received 26 January 2021; Received in revised form 30 September 2021; Accepted 9 October 2021
Available online 20 October 2021

components, sources, additives, *etc*. *vs*. multiple input/output properties to be tested. In addition, many industrial platforms are prepared for production process but not to test new blending alternatives to gasoline compounding. This is why predictive models are needed for the design of new alternative mixtures for gasoline compounding [7,8]. In this context, there are reports of computational models useful to determine the composition of gasoline from near infrared spectrum data [9–11]. In fact, many groups have successfully employed modern Machine Learning (ML) algorithms to predict target composition-properties for the discovery of new materials. For instance, ML models have been used to predict the density, compressibility, expansivity [12], or heating value [13] of blends. ML models have been used also for screening molecules for the design of next generation fuels [14]. In addition, Information Fusion (IF) strategies are being used to process information from different sources in order to train ML models [15–20]. Alternatively, ML algorithms can be employed for the classification of different samples in classes instead of predicting the numerical values of their physicochemical properties. Linear Discriminant Analysis (LDA) [21] is one of the most used linear technique for classification. However, Artificial Neural Networks (ANN) [22] and Support Vector Machine (SVM) [23] are among the most used non-linear ML methods [24]. In recent works we successfully introduced the IFPTML (IF + PT + ML) algorithm, which includes a Perturbation Theory (PT) step to the IF and ML phases. The IFPTML models have been applied to multiple problems in Organic Chemistry, Medicinal Chemistry, Nanotechnology, Biofuel design, Materials science, *etc*. [25–33]. In any case, there are not been reports of IFPTML models for alternative fuel blend modelling.

In this paper, we work with a database created from the beginning based on real analysis of a refinery. This model is not intended to analyze or model the process of the plant. The aim of this work is the generation of a model for the synthesis of new gasoline blends, trying to get the raw materials from natural or green sources or waste. For this purpose, an IFPTML model has been obtained based on real analysis data from a refinery placed in Basque Country, North Spain. First, a database has been created with the analysis of different streams and operations carried out in the plant. In so doing, we carried out an IF process to assemble a large working data set with 270,918 cases (property outcomes for different samples). Next, we are going to train alternative IFPTML models with different ML techniques such as LDA, ANN, and SVM. We also show the results of internal control experiments using Fourier Transform Near Infrared (FT-NIR) spectroscopy. This kind of models could become a useful tool towards the design of new biofuels, Refuse-derived fuels, *etc*. with lower environmental impact.

2. Materials AND methods

2.1. IF Pre-processing and data healing phase

2.1.1. Data set compilation

The first step was the compilation of information released by a petroleum refinery sited in the Basque Country. The workflow diagram of the refinery is depicted in Fig. 1. In this figure, squares represent streams, dots operations, and ellipses crude oil streams. We obtained a total of 657 samples and carry out 5758 measurements of at least 1 out of 574 input properties and at least 1 out 44 possible output properties. This led to the formation of 26 blocks of information (datasets) including 574 input properties (upstream samples) and up to 44 output properties (downstream blending samples) from 14 different streams and 23 operations included in 4 different production lines. After that, we carry out the IF process to assemble all the initial data blocks into a large working data set with 270,918 cases (property outcomes for different samples), see next sections. Please, take into consideration that one upstream sample may be involved in many different operations and streams and ultimately in different lines of production (bifurcations) of the process. This give place to a formation of a very large dataset with a notably higher number of final cases $n_{total} = 270{,}918 \gg n_{original} = 5758$ original cases. The data set built follows the order of streams in Fig. 1. The data set is composed of a number of input variables and output variables to build a *multi-label* (multi-*output*) classification model.

2.1.2. Output values healing

The original dataset released by the refinery contained a total of 62 output variables $v_{k,i}(t_j)$ of different types k^{th} (R.O.N., M.O.N., Temperatures, Density, *etc*.) measured at the end of the process. In order to have valid values when generating the model, we cleaned the data by removing those data that do not have a numerical outcome and empty variables. We have fond also some variables with same property measured but different units: Reid steam pressure (kPa and psi), density (kg/m^3 and g/cm^3).

2.1.3. Input variables healing

Our dataset contains multiple output variables $v_{k,i}(t_j)$ and input variables $V_{k,i}(s_j, o_j, t_j)$. The output variables $v_{k,i}(t_j)$ are real numerical parameters used to quantify the value of the of the k^{th} physicochemical property (R.O.N., M.O.N., Temperatures, Density, *etc*.) for the i^{th} product sample obtained at time t_j. The input variables $V_{k,i}(s_j, o_j, t_j)$ are used to quantify the value of the k^{th} property for the i^{th} sample obtained/

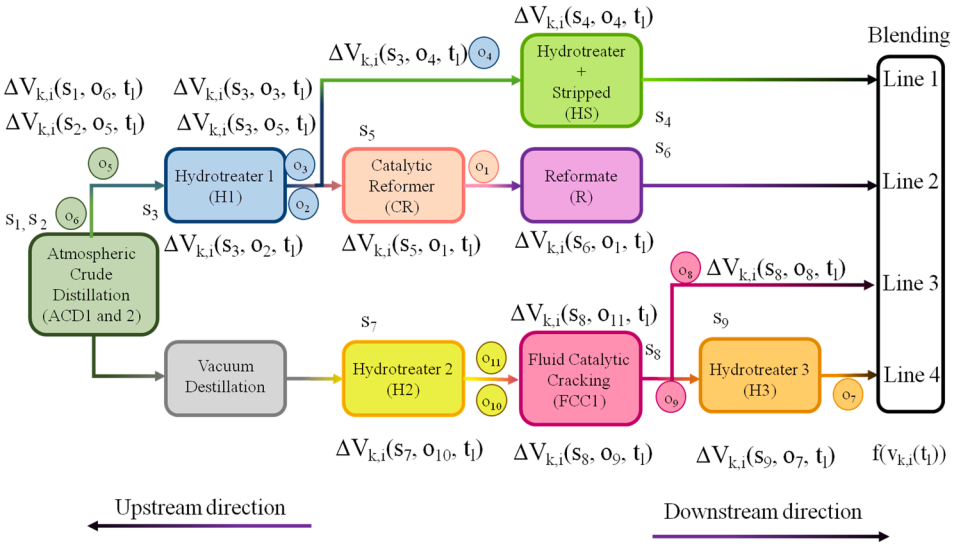


Fig. 1. Refinery general workflow.

measured from the stream s_j and operation o_j at time t_j. The output $v_{k,i}(t_j)$ and input $V_{k,i}(s_j, o_j, t_j)$ variables quantify essentially the same properties. However, for the case of input variables $V_{k,i}(s_j, o_j, t_j)$ the property have been measured in other operations and at other time t_j up stream not in the end of the process as is the case for output variables $v_{k,i}(t_j)$. In the case of missing values $V_{k,i}(s_j, o_j, t_j)$ for a specific sample at a given sampling time t_j, we assigned the average value $<V_{k,i}(s_j, t_j)>_{oj}$ of the variable for all samples from the same operation (o) in different times t_j. In the case that we cannot calculate $<V_{k,i}(s_j, t_j)>_{oj}$ because there are no measures of $V_{k,i}(s_j, o_j, t_j)$ for the operation o_j, we assigned the $<V_{k,i}(s_j, t_j)>_{o'j}$. This is the average value of the equivalent operation (o'_j) of the same stream but released on a different data package. For instance, if we have not data for the density $V_{k,i}(s_j, o_j, t_j) = d_i$(Catalytic reformer, Reformate, 1) of one sample at time $t_j = 1$ in the stream for the Catalytic reformer in Reformate operation we used the average $<d_i$(Catalytic reformer, 1)$>_{Reformate}$. In the case we cannot calculate $< d_i$(Catalytic reformer 1, 1)$>_{Reformate}$ we used the average $<d_i$(Catalytic reformer 2, 1)$>_{Reformate}$. Last, if we cannot calculate neither $<d_i$(Catalytic reformer 1, 1)$>_{Reformate}$ nor $< d_i$(Catalytic reformer 2, 1)$>_{Reformate}$ then we eliminated the variable d_i(Catalytic reformer 1, Reformate, t_j) from the dataset. After healing/replacing missing data for all these input variables $V_{k,i}(s_j, o_j, t_j)$ in each one of the streams, we analyzed their variability. Those that remained constant or had a very small variability were deleted. In total, we eliminated 237 variables out of 744 original variables. The existence of missing data is explained due to the high cost of a constant sampling of all the 744 variables several times per day.

For the sake of simplicity we grouped and processed all the information into data matrices or blocks $DB_m(s_j, o_j, t_j)$. Each $DB_m(s_j, o_j, t_j)$ contains the values for all input variables $V_{k,i}(s_j, o_j, t_j)$ of operation. It means that variables $V_{k,i}(s_j, o_j, t_j)$ included in the m^{th} data block $DB_m(s_j, o_j, t_j)$ have the same values s_j, o_j, and t_j of this $DB_m(s_j, o_j, t_j)$. After healing process the number of variables that remain in each $DB(s_j, o_j, t_j)$ are the following. The first data block $DB_1(s_1, o_6, t_j)$ includes the variables 21 of stream s_1 = Atmospheric Crude Destillator 1 (ACD1) and the operation o_6 = Straight Naphtha (StN). The second data block $DB_2(s_2, o_5, t_j)$ includes 48 variables of the stream s_2 = Atmospheric Crude Destillator 2 (ACD2) and operation o_5 = Stabilized Naphtha 2 (StN2). The other data blocks and their number of variables appear in Table 1. See also Fig. 2 to see the correspondence between the data blocks $DB(s_j, o_j, t_j)$ and their respective streams sj and oj in the productive process. This make a total of 507 input variables $V_{k,i}(s_j, o_j, t_j)$ retained for IFPTML analysis distributed into 14 data blocks $DB_m(s_j, o_j, t_j)$ involving 14 operations and 9 process streams s_j, see details in Table 1. See also the points of sampling/measuring for these variables on the process diagram in Fig. 1.

2.1.4. IF process and data set design

In order to design the final working data set we carried out an IF process with all the values of the output $v_{k,i}(t_j)$ and input measured variables $V_{m,i}(s_j, o_j, t_j)$ from the 14 data blocks $DB(s_j, o_j, t_j)$ released by the refinery. In our IF process we fused two data data blocks $DB(s_j, o_j, t_j)$ and $DB(s_{j+1}, o_j, t_{j+1})$ as follows. We fused $DB(s_j, o_j, t_j)$ and $DB(s_{j+1}, o_j, t_{j+1})$ in horizontal from left to right if they belong to the same production line, are consecutive in the production flow, and the operation of the first block was previous to the second block $t_j < t_{j+1}$. Consequently, all data blocks $DB(s_j, o_j, t_j)$ fused in horizontal correspond to continuous batch operations performed in series downstream, see Table 2. On the other hand, blocks $DB(s_j, o_j, t_j)$ fused in vertical each other correspond to operations performed in parallel in different lines of the process. Please, compare Fig. 2 *vs.* Fig. 3 to see the mapping of $DB(s_j, o_j, t_j)$ allocation from flow diagram to data sheet during IF process. In order to facilitate understanding of the IF process we used the same operation $DB(s_j, o_j, t_j)$ letter code and operation color code in both figures. We used the DBrowser for SQLite DB4S 3.11.2 to create the database [34]. It is a tool for creating databases based on smaller tables in an IF like process. To generate this database, first, a data file for each of the operations in the different streams must be created. In one line of our working dataset we put one value of the output $v_{k,i}(t_j)$ variable and all the corresponding values of the input $V_{k,i}(o_j, s_j, t_j)$ variables for a given pathway

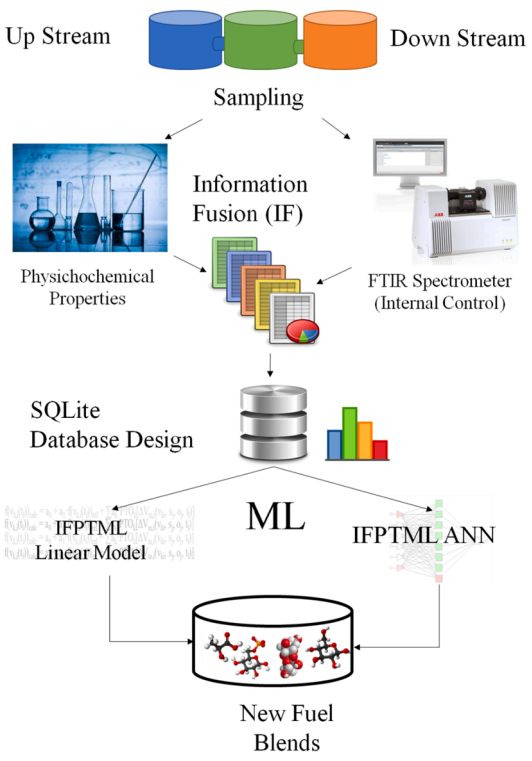


Fig. 2. Workflow used to develop the model.

Table 1
Data blocks, streams, operations and input variables count after healing.

DB (m)	DB Symbol	s_j	Stream Name	Stream Code	o_j	Oper. Name	Oper. Code	Var. Count
1	$DB_1(s_1, o_6, t_j)$	s_1	Atmospheric Crude Destillator 1	ACD1	o_6	Straight Naphtha	StN	21
2	$DB_2(s_2, o_5, t_j)$	s_2	Atmospheric Crude Destillator 2	ACD2	o_5	Stabilized Naphtha 2	StN2	48
3	$DB_3(s_3, o_5, t_j)$	s_3	Hydrotreater 1	H1	o_5	Stabilized Naphtha 1	StN1	25
4	$DB_4(s_3, o_4, t_j)$	s_3	Hydrotreater 1	H1	o_4	Hydrogenated High Naphtha 2	HHiN2	62
5	$DB_5(s_3, o_3, t_j)$				o_3	Hydrogenated Heavy Naphtha 2	HHeN2	26
6	$DB_6(s_3, o_2, t_j)$				o_2	Hydrogenated Heavy Naphtha 1	HHeN1	56
7	$DB_7(s_4, o_4, t_j)$	s_4	Hydrotrater + Stripped	HS	o_4	Hydrogenated High Naphtha 1	HHiN1	25
8	$DB_8(s_5, o_1, t_j)$	s_5	Catalytic Reformer	CR	o_1	Reformate	R	43
9	$DB_9(s_6, o_1, t_j)$	s_6	Reformate (Low Benzene Content)	R				40
10	$DB_{10}(s_7, o_{10}, t_j)$	s_7	Hydrotrater 2	H2	o_{10}	Cracked Gasoline Precursor	CGP	32
11	$DB_{11}(s_8, o_8, t_j)$	s_8	Fluid Catalytic Cracking 1	FCC1	o_8	Cracked High Gasoline	CHG	31
12	$DB_{12}(s_8, o_9, t_j)$				o_9	Cracked Intermediate Gasoline 2	CIG2	33
13	$DB_{13}(s_8, o_{11}, t_j)$				o_{11}	Cracked Intermediate Gasoline 1	CIG1	35
14	$DB_{14}(s_9, o_7, t_j)$	s_9	Hydrotrater 3	H3	o_7	Hydrogenated Cracked Gasoline	HCG	30

Table 2
IF process schematic illustration.

Input Variables $V_{k,i}(s_j, o_j, t_l)$ Donwstream Direction ==>									Set	$f(v_{k,i}(t_l))_{ref}$	Line	$f(v_{k,i}(t_l))_{obs}$
0	0	0	0	DB_7	DB_4	DB_3	DB_2	DB_1	t	x	1	x
0	0	0	0	DB_7	DB_4	DB_3	DB_2	DB_1	v	x	1	x
DB_9	DB_8	DB_6	DB_5	0	0	DB_3	DB_2	DB_1	t	x	2	x
DB_9	DB_8	DB_6	DB_5	0	0	DB_3	DB_2	DB_1	v	x	2	x
DB_{14}	DB_{13}	DB_{12}	DB_{10}	0	0	0	0	0	t	x	3	x
DB_{14}	DB_{13}	DB_{12}	DB_{10}	0	0	0	0	0	v	x	3	x
DB_{11}	0	0	DB_{10}	0	0	0	0	0	t	x	4	x
DB_{11}	0	0	DB_{10}	0	0	0	0	0	v	x	4	x

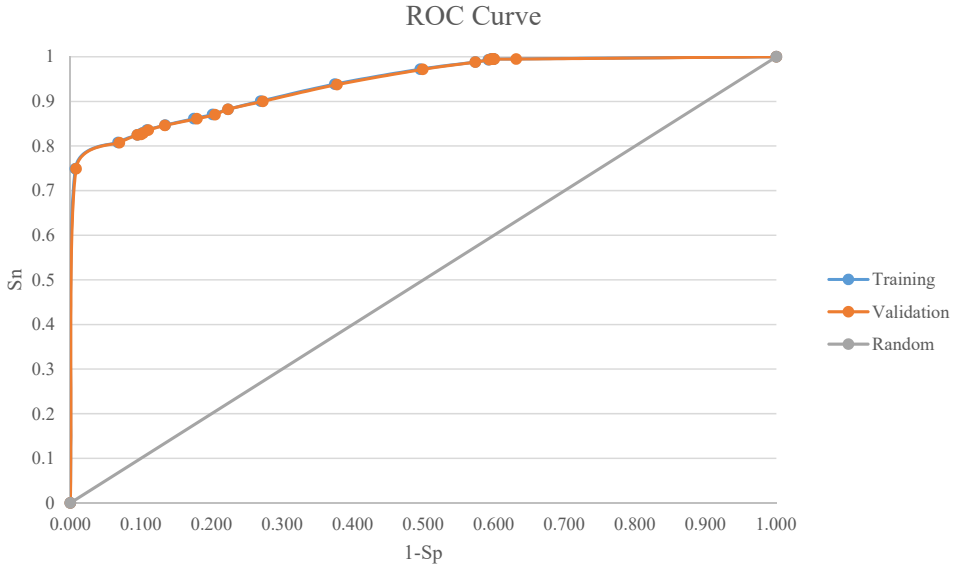


Fig. 3. ROC Curve of Equation 8 model.

downstream from the crude sample to the end of the process. In so doing, we have taken into account that streams from ACD can go either way by the upper path to the Blending and can go through any of the processes that appear in each stream. In addition, crude stream could reach to the Blending stream along the lower path.

Once the previous working data set was constructed we defined the objective function $f(v_{k,i}(t_j))_{obs}$. This function gets the values $f(v_{k,i}(t_j))_{obs} = 1$ when the k^{th} output physicochemical property $v_{k,i}(t_j)$ (R.O.N., M.O.N., Density, *etc.*) of the i^{th} sample reach a value within the desired range at time t_j. The function gets the values $f(v_{k,i}(t_j))_{obs} = 0$ otherwise. The desired limits of one variable are given by production guidelines specified under UNE EN 228:2012 + A1:2017 standard. This European Union (EU) standard specifies the requirements and test methods for the unleaded gasoline that is distributed and marketed. This include the top $v_{ki}(t_j)_{max}$ and bottom $v_{ki}(t_j)_{min}$ limits for the physiochemical properties $v_{ki}(t_j)$ of gasoline blends.[35] Consequently, $f(v_{ki}(t_j))_{obs} = 1$ if $v_{ki}(t_j)_{min} < v_{ki}(t_j) < v_{ki}(t_j)_{max}$, $f(v_{ki}(t_j))_{obs} = 0$ otherwise. In the case of physicochemical property $v_{k,i}(t_j)$ interesting for the production process that do not have specified limits in the guidelines, the following heuristic has been used. For these cases the superior/inferior limits are calculated as $< v_{k,i}(t_j) > \pm 3{\cdot}\sigma_{k,j}$. Specifically, $v_{ki}(t_j)_{max} = <v_{k,i}(t_j)> + 3{\cdot}\sigma_{k,j}$ and $v_{ki}(t_j)_{min} = <v_{k,i}(t_j)> - 3{\cdot}\sigma_{k,j}$. It represents a deviation from the $< v_{k,i}(t_j) >$ value expressed in 3-times the value of the standard deviation $\sigma_{k,j}$. This comes from a rearrangement of the formula used to calculate standardized values with z-scores.[24] PT data pre-processing phase

2.1.5. PT Operators calculation

After healing the input variables, we transformed them into Perturbation-Theory Operators (PTO). These PTOs have the notation $PTO_k(v_k, s_j, t_j)$ for this problem. The values used here as input are Moving Averages (MA) of the original input variable $PTO_k(o_j, s_j, t_j) = \Delta V_{k,i}(o_j, s_j, t_j)$. The MAs depend on the experimental values of each input parameter measured $V_{m,i}(s_j, t_j)$ in the operation o_j of the stream s_j at the

sampling time t_j. In this way, the deviation of each input value $V_{k,i}(o_j, s_j, t_j)$ with respect to the average $< V_{k,i}(t_j)>_{oj,sj}$ (expected value) of this value for all the samples with the same output parameter $v_{k,i}$ and the same stream s_j was calculated as follows (Equation 1).

$$\Delta V_{k,i}(o_j, s_j, t_j) = V_{k,i}(o_j, s_j, t_j) - <V_{k,i}(t_j)>_{oj,sj} \quad (1)$$

2.1.6. Function of reference

The reference function $f(v_{k,i}(t_j))_{ref}$ is the first variable that enters the model. It is equal to the expected probability $p[f(v_{k,i}(t_j)) = 1]_{expt}$ with which the parameter $v_{k,i}(t_j)$ of a sample at the end of the process falls within the specified limits, $f(v_{k,i}(t_j))_{obs} = 1$. It depends on the number of samples $n(f(v_{k,i}(t_j))_{obs} = 1)$ that fall within the limits and on the total number of samples $n(f(v_{k,i}(t_j))_{obs})$ with parameter $f(v_{k,i}(t_j))$. It is obtained as follows in Equation 2.

$$f(v_{k,i}(t_j))_{ref} = p[f(v_{k,i}(t_j)) = 1]_{expt} = n(f(v_{k,i}(t_j))_{obs} = 1)/\ n(f(v_{k,i}(t_j))_{obs}) \quad \textbf{(2)}$$

2.2. ML training and validation phase

2.2.1. IFPTML linear and non-linear models

Firstly, linear IFPTML models will be sought using Linear Discriminant Analysis (LDA).[24] The values of objective function $f(v_{k,i}(t_j))_{obs}$ (function to be fitted) were used to train the model together with the values of function of reference $f(v_{k,i}(t_j))_{ref}$ and the operators $PTO_k[\Delta V_{k,i}(s_j, t_j)]$ as input variables. In addition, we are going to run Artificial Neural Network (ANN) algorithms to seek PTML-ANN models. The PTML-ANN models using Linear Neural Network (LNN) topology are similar to PTML-LDA models (both present a linear form).[24] The PTML-ANN models using Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) topologies are non-linear. In most cases we are going to use Specificity (Sp), Sensitivity (Sn), Fisher ratio (F), p-level, and/or Area Under Receiver Operating Characteristic (AUROC) curve to measure the performance of the model. In Fig. 2 we summarize all the steps given to seek this kind of IFPTML models. This model takes as a starting point the reference function $f(v_{k,i}(t_j))_{ref}$ and adds the effect of deviations (perturbations) measured by the $PTO_k[\Delta V_{k,i}(s_j, t_j)]$ input variables. See the general equation of the linear models:

$$f(v_{k,i}(t_j))_{calc} = a_0 + a_1 \cdot f(v_{k,i}(t_j))_{ref} + \sum a_k \cdot \Delta V_{k,i}(o_j, s_j, t_j) \quad \textbf{(3)}$$

2.2.2. Output variable and posterior probabilities calculation

The output variable $f(v_{k,i}(t_j))_{cal}$ is a real-valued scoring function related to the probability $p[f(v_{k,i}(t_j))_{pred} = 1]$ with which the experimental value of the i^{th} property measured on the j^{th} stream in l^{th} time is predicted $f(v_{k,i}(t_j))_{calc} = 1$ to fall within the specified limits. The output variable $f(v_{k,i}(t_j))_{calc}$ was discretized to obtain the predicted classification of each case ($f(v_{k,i}(t_j))_{pred} = 1$) or not ($f(v_{ij})_{pred} = 0$), see next section. This Boolean variable $f(v_{k,i}(t_j))_{calc}$ should be compared to the observed classification $f(v_{k,i}(t_j))_{obs}$ for each sample in order to calculate the Specificity (Sp) and Sensitivity (Sn) [36,37] of the model. On the other hand, the function $f(v_{k,i}(t_j))_{ref}$ is the first input variable.

2.2.3. Posterior probabilities calculation

When the IFPTML model is generated (by LDA, SVM, ANN, etc.), we can obtain the values of $f(v_{k,i}(t_j))_{calc}$. However, $f(v_{k,i}(t_j))_{calc}$ is a real value variable (scoring function) and we need to discretize it in order to classify the samples as $f(v_{k,i}(t_j))_{pred} = 1$ or $f(v_{k,i}(t_j))_{pred} = 0$. In so doing, we used a sigmoidal function to calculate the probabilities *a posteriori* $p[f(v_{k,i}(t_j))_{pred} = 1]$ with which a sample is classified as $f(v_{k,i}(t_j))_{pred} = 1$. This sigmoid function uses as input *a priori* probabilities (π_0 and π_1). These prior probabilities are used in the Bayesian method to calculate the posterior probabilities. The equation of the sigmoid function of the posterior probabilities is the following [24].

$$p[f(v_{k,i}(t_j))_{pred} = 1] = 1/[1 + (\pi_0/\pi_1)\cdot\exp(-f(v_{k,i}(t_j))_{calc}) \quad \textbf{(4)}$$

2.2.4. FT-NIR spectra acquisition

For internal control study a total of 28,876 samples have been measured in the refinery quality control lab using FT-NIR technique. The spectra of these samples were recorded as follow. Firstly, thermostatic bath was turned on regulated at 25 °C and connected the FT-NIR spectrometer to run the standard samples of reference. All Gasoline samples were keep a low temperatures and the cell was washed repeated times with samples avoiding either loss by vaporization or the formation of bubbles. FT-NIR spectra were recorded in the range 4000 cm^{-1} to 4800 cm^{-1}. The FT-NIR spectrometer used was ABB MB3600. This spectrometer is able to comparing recorded spectra *vs*. an internal database of standard samples. Specifically, for this study all results were recorded and expressed according to the respective EU Standards: ASTMD2899 for RON, ASTMD2700 for MON, ASTMD5191 for Reid's Vapor Pressure (RVP), ASTMD4052 for Density, EN22854 for Benzene, EN22854 for Aromatics, EN22854 for Olefins, ASTMD88 for Evaporated (70, 100, and 150 °C), EN13132 for MTBE, and ETBE, *etc* [35].

3. Results and discussion

3.1. IFPTML-LDA linear models

We decided to train our model using an Expert-Guided Selection (EGS) strategy. The aim was to seek a model taking into account as many streams and production lines as possible but using few variables as possible and without losing Sp and Sn. In so doing, used a Forward Stepwise (FSW) variable selection strategy to select the more influential variables and complemented them with criteria of the experts in our team [38]. The equation of the best model found written upstream from left to write is the following, Equation 5:

$$f(v_{k,i}(t_j))_{calc} = -9.381 + 16.898\cdot f(v_{k,i}(t_j))_{ref} - 0.418\cdot \Delta V_{5,i}(s_1, o_6, t_j) - 0.094\cdot\Delta V_{8,i}(s_1, o_6, t_j) + 0.353\cdot\Delta V_{30,i}(s_2, o_5, t_j) + 0.409\cdot\Delta V_{8,i}(s_2, o_5, t_j) - 0.022\cdot\Delta V_{1,i}(s_3, o_3, t_j) - 2.689\cdot\Delta V_{13,i}(s_4, o_4, t_j) - 0.049\cdot\Delta V_{5,i}(s_5, o_1, t_j) + 17.153\cdot\Delta V_{9,i}(s_5, o_1, t_j) + 0.025\cdot\Delta V_{3,i}(s_6, o_1, t_j) - 0.575\cdot\Delta V_{26,i}(s_6, o_1, t_j) + 0.081\cdot\Delta V_{4,i}(s_7, o_{10}, t_j) - 0.028\cdot\Delta V_{2,i}(s_7, o_{10}, t_j) + 0.039\cdot\Delta V_{30,i}(s_8, o_8, t_j) - 0.001\cdot\Delta V_{3,i}(s_9, o_7, t_j) \quad \textbf{(5)}$$

This IFPMTL-EGS model has only 14 $\Delta V_{k,i}(o_j, s_j, t_j)$ variables plus the function of reference and the independent term. The details about the input variables (names, data block, etc.) and he exact values of the coefficients of the new EGS model are shown in Table 3. The classification matrix (Table 3) shows that the values of Sn and Sp ≈ 84–91% are high for this kind of model. The EGS model take into account the most important variables of each of the operation's streams overcoming the pitfall of the FSW strategy. Consequently, we have input variables $\Delta V_{k,i}(o_j, s_j, t_j)$ measured at almost all points of the process. This is very important if we want to maximize the possibility of introducing alternative raw materials blending mixtures from other sources in different points of the process. In fact, the model takes into account the 4 lines of production, 9 streams, and 9 operations directly. The model also takes into account other streams and operations in an indirectly manner. It happens because when you include a variable for one stream/operation upstream it shall affect in turn all streams/operations downstream in the production line, see Fig. 1.

The relatively low number of variables that enter makes the model easier to use and with less risk of over fitting and co-linearity. In any case, in order to check on the collinearity between the variables, we

Table 3
IFPTML-LDA models results.

Classes			$f(v_{k,i}(t_j))_{pred}$		
$f(v_{k,i}(t_j))_{obs}$	Stat.	(%)	n_j	0	1
Training set					
0	Sp	91.1	25,189	22,940	2249
1	Sn	84.8	151,108	23,013	128,095
Validation set					
0	Sp	91.1	8405	7660	745
1	Sn	84.7	50,339	7700	42,639

generated a correlation matrix with R^2 for every two variables in the EGS model. We can see that almost all variable pairs have very low $R^2 < 0.1$. This demonstrates a very low risk of pair-wise co-linearity among the variables of the model. Consequently, the coefficients of the variables should express their real contribution to the output variable because the lack of collinearity masking effects among them [39]. There is only one case in which $R^2 = 0.42$ with p-level < 0.05 indicating certain collinearity between this pair of variables. These are variables measuring 10% collected from the sample and the amount in volume of toluene are related to each other. We also analyzed the AUROC values. This is a widely used technique because it is very visual to evaluate the performance of the classifier model [40]. Consequently, we also obtained values of AUROC = 0.936; which are notably higher than AUROC = 0.5 (value for a random classifier), Fig. 3. This indicates that this model significantly discriminate between acceptable and unacceptable fuel samples being a not random classifier [41]. Consequently, the model could be useful to design alternative fuel blends for gasoline compounding. In so doing, the following steps are necessary. Firstly, the user have to obtain experimentally samples of new alternative mixtures from biofuel, refuse-derived fuels, etc. Next, it is necessary to measure experimentally the input properties of these samples. Last, we introduce the values in the respective terms of the equation of the model (according to the operation or stream desired) to obtain the predicted probability of success.

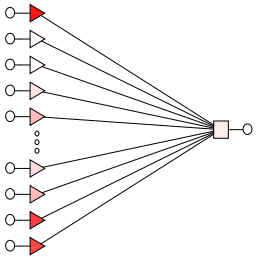
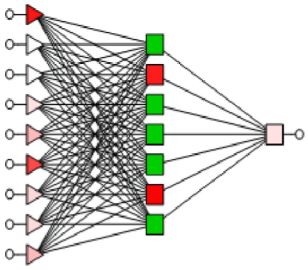
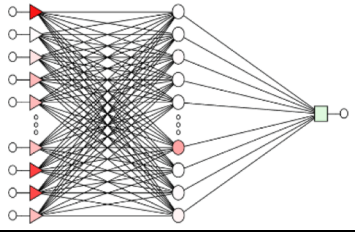
3.2. IFPTML-ANN non-linear models

After analyzing the data with linear equations, more complex models were generated to analyze if the prediction can be improved. Several types of ANN were analyzed in order to obtain the best results. We used the variables of the EGS model. Three types of ANN were analyzed in this work: the LNN [24], MLP [42], and RBF [43] topologies. They are more complicated models to generate and use, but in some cases better results are obtained. In this case, very similar results were obtained with the RBF and MLP models, but they are much more complex to use than the EGS model and the prediction improvement is not enough to opt for these ANN, see Table 4. It is interesting that the IFPTML-LNN linear model is similar to the IFPMTL-LDA model (also linear) but the values of Sp in the LDA model are 10% higher. It could be because the LDA is a more flexible Bayesian model allowing to adjust prior probabilities (π_0/π_1) [24]. In conclusion, the results obtained with the IFPTML-ANN models did not outperform the optimized IFPTML-LDA model. This confirms our decision on using a linear instead of a non-linear model.

3.3. IFPTML-FT-NIR internal control study

FT-NIR methodology has been used before for Chemometrics studies of gasoline fuel blends.[44,45] In most studies using FT-NIR it is used as the main technique for the characterized the samples. In our case we

Table 4
IFPTML-ANN models results.

Profile			Training				Validation		
	AUROC	$f(v_{k,i}(t_j))$	0[a]	1[a]	(%)	Par.	(%)	0[a]	1[a]
LNN 15:15-1:1	0.898	0[b]	24,523	28,229	81.3	Sp	81.1	8172	9404
		1[b]	5636	122,879	81.3	Sn	81.3	1904	40,935
									
MLP 9:9-7-1:1	0.900	0[b]	24,720	5439	82.0	Sp	81.9	8249	9081
		1[b]	5439	123,913	95.8	Sn	82.0	1827	41,258
									
RBF 11:11-388-1:1	0.900	0[b]	24,725	27,058	82.0	Sp	81.8	8246	9044
		1[b]	5434	124,050	82.1	Sn	82.0	1830	41,295
									

[a] $f(v_{k,i}(t_j))_{pred} = 0$ or 1 predicted values.
[b] $f(v_{k,i}(t_j))_{obs} = 0$ or 1 observed values.

used FT-NIR methodology to run an internal experimental control test of the robustness of IFPTML methodology based on physicochemical properties (PCP). The aim of this study was not the comparison of indirect FT-NIR *vs*. direct PCP measurement techniques *per se*. The objective was to analyze the robustness of the IFPTML model predictions using the two types of inputs (direct PCP measurements *vs*. FT-NIR inference). Different researchers have used different IR ranges in this kind of studies, *e.g.*; Balabin *et al.* [46] used 14,000–8000 cm^{-1}, da Silva *et al.* used 10,000–4000 cm^{-1}, but Al-Ghouti *et al.* [47] used peaks in the range 400–3500 cm^{-1}. In our case, following spectrometer calibration specifications FT-NIR spectra were recorded in the refinery quality control laboratory in the range 4800–4000 cm^{-1}. In Fig. 4 we depict two examples of FT-NIR spectra recorded for a reformatted naphtha sample (upstream) and gasoline sample (downstream).

A total of 28,876 values for 16 different PCP; *e.g.*, Benzene (%V) and Aromatics (%V) of multiple samples were inferred from the FT-NIR spectra using the software associated to the ABB MB3600 spectrometer. These experimental values and their respective input variables were distributed as follow: 21,652 outcomes for training and 7224 outcomes for validation. After that, alternative values of input variables $\Delta V_{k,i}(o_j, s_j, t_j)_{IR}$ were calculated using FT-NIR-inferred properties instead of classic PCP properties $\Delta V_{k,i}(o_j, s_j, t_j)_{PCP}$. Both kinds of values were used above to train and validate the IFPTML model altogether. We depict in Table 5 the classification matrix for the control IFPTML-FT-NIR model. This model was generated using fully balanced prior probabilities of $\pi_1 = \pi_0 = 0.5$. We can note that the Sp and Sn ≈ 86–97%values for FT-NIR predictions are very high and mostly in the same range, only slightly higher than, Sp and Sn ≈ 84–91% of PCP model.

In addition, there are only few Sign Inversions (SI) on the coefficients when we change use FT-NIR instead of classic procedures, see Table 6. Fortunately, the magnitude of sign change (ΔSI) is small when happening. For instance, Naphthene (%V) have a change of only 0.87. The only variable with a relatively higher change in coefficient was Paraffin (%V) but this does not affect the overall Sp and Sn values of the model. We can conclude that both models IFPTML-PCP and IFPTML-FT-NIR are coincident and the IFPTML algorithm is robust to the change of technique from classic procedures to FT-NIR.

Table 5
IFPTML-LDA FT-NIR models results.

Classes			$f(v_{k,i}(t_j))_{pred}$		
$f(v_{k,i}(t_j))_{obs}$	Stat.	(%)	n_j	0	1
Training set					
0	Sp	97.4	230	224	6
1	Sn	86.7	21,422	2857	18,565
Validation set					
0	Sp	94.9	78	74	4
1	Sn	86.9	7146	937	6209

3.4. IFPTML model compared to other ML models

In closing, we would like to present a comparison of our IFPTML models to other ML models reported in the literature for similar problems. In general, the models compared do not use the same dataset or solve exactly the same problem. As result, the present comparison is not focused on statistical parameters but in terms of range of applicability of the model. In Table 7 we summarized the results obtained with different models in this work and in the literature. The criteria used to select the models was the following. The dataset details and model performance statistics are public and the model applies to at least one Gasoline/Fuel property. As it can be seen in the table all models use relatively small datasets (<1500 cases) *vs*. >230 K cases of our model. All models apply to only one physicochemical property. In the case that more than one output PCPs are predicted in the same paper they need to fit more than one model. In addition, all papers focus on few (1–3) output properties by sample *vs*. up to 44 output properties by sample of our model. In addition, all works use as input variables PCPs or FT-NIR to predict the output PCPs. Our paper is the only one reporting a comparative study (internal control) with both PCPs *vs*. FT-NIR as alternative input variables.

4. Conclusions

Optimization of current or alternative fuel blends to be introduced in

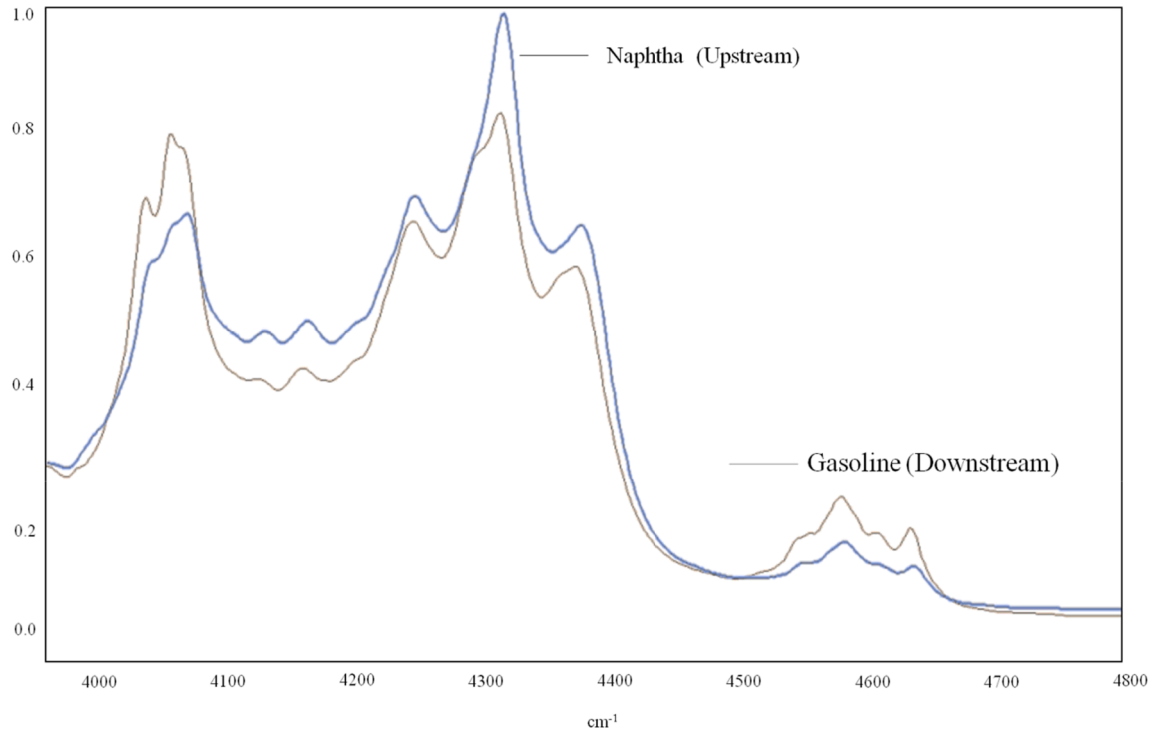


Fig. 4. FT-NIR control spectra for naphtha (upstream) and gasoline blend (downstream).

Table 6
IFPTML-LDA PCP *vs.* FT-NIR model coefficients.

Procc.	Data	Variable	Variable	$f(v_{k,i}(t_j))_{calc}$ [a]			
Line	Block	Name	Notation	PCP	FT-NIR	SI	ΔSI
–	–	Indep. term	a_0	−9.3812	−91.8575	No	–
–	–	Ref. function	$f(v_{k,i}(t_j))_{ref}$	16.9891	100.1441	No	–
1,2	$DB_1(s_1, o_6, t_j)$	Benzene (%V)	$\Delta V_{05,i}(s_1, o_6, t_j)$	−0.4184	−1.3746	No	–
		Naphthene (%V)	$\Delta V_{08,i}(s_1, o_6, t_j)$	−0.0944	0.7721	Yes	0.87
1,2	$DB_2(s_2, o_2, t_j)$	RVP (psi)	$\Delta V_{30,k}(s_2, o_5, t_j)$	0.3533	0.8362	No	–
		Naphthene (%V)	$\Delta V_{08,i}(s_2, o_5, t_j)$	0.4095	1.6462	No	–
1,2	$DB_5(s_3, o_3, t_j)$	Collected10% (°C)	$\Delta V_{01,i}(s_3, o_3, t_j)$	−0.0217	0.2188	Yes	0.24
1	$DB_7(s_4, o_4, t_j)$	Olefin (%V)	$\Delta V_{13,i}(s_4, o_4, t_j)$	−2.6887	5.4905	Yes	8.18
2	$DB_8(s_5, o_1, t_j)$	Benzene (%V)	$\Delta V_{05,i}(s_5, o_1, t_j)$	−0.0492	−0.2618	No	−0.21
		Paraffin (%V)	$\Delta V_{09,i}(s_5, o_1, t_j)$	17.1532	0	Yes	−17.15
2	$DB_9(s_6, o_1, t_j)$	Collected90% (°C)	$\Delta V_{03,i}(s_6, o_1, t_j)$	0.0253	0.0481	No	–
		Toluene (%V)	$\Delta V_{26,i}(\sigma_6, o_1, t_j)$	−0.5747	−0.2433	No	–
3,4	$DB_{10}(s_7, o_{10}, t_j)$	Aromatic (%V)	$\Delta V_{04,i}(s_7, o_{10}, t_j)$	0.0809	−0.5054	Yes	−0.59
		Collected50% (°C)	$\Delta V_{02,i}(s_7, o_{10}, t_j)$	−0.0278	−0.0197	No	–
3	$DB_{11}(s_8, o_8, t_j)$	RVP(psi)	$\Delta V_{30,i}(s_8, o_8, t_j)$	0.0387	−0.3453	Yes	−0.38
4	$DB_{14}(s_9, o_9, t_j)$	Collected90% (°C)	$\Delta V_{03,i}(s_9, o_7, t_j)$	−0.0005	−0.0459	No	–

[a] IFPTML model coefficients, PCP = coefficients obtained using classic analytical techniques to determine Physico-Chemical Properties (PCP) of gasoline blends downstream, FT-NIR = coefficients obtained using FT-NIR as internal control technique, SI = Signal Inversion of the coefficient due to PCP to FT-NIR change, ΔSI = magnitude of coefficient SI switching.

Table 7
IFPTML models compared to other ML models from literature.

Mod.	ML[a]	IV[b]	MO[c]	RGB[d]	IFRP[e]	OP[f]	Cases	Stat.	Val.	Ref.
IF PT	LDA	PCP	Yes	Yes	Yes	44	>230 K	Sp Sn	84–91 %	This work
ML	LDA	FT NIR	Yes	Yes	Yes	16	>25 K	Sp Sn	86–97%	This work
	ANN	PCP	Yes	Yes	Yes	44	>230 K	Sp Sn	81–82%	This work
ML	PLS	FT NIR	No	Yes	No	2	259	R^2	0.83	[44]
	PF	PCP	No	Yes	No	3	273	AAD	6.5	[48]
	CPLS	PCP	No	Yes	No	1	1471	SE	1.01	[49]
	RTO	PCP	No	Yes	Yes	1	–	–	–	[50]
	ANN	PCP	No	Yes	No	1	173	R^2	0.98	[51]
	PLS	FT NIR	No	Yes	No	1	16	RMSE	–	[52]
	RS	FT NIR	No	Yes	Yes	1	103	RMSE	–	[53]

[a] ML = Machine Learning Technique, NM = Nelson's Method, LDA = Linear Discriminant Analysis, ANN = Artificial Neural Network, PLS = Partial Least Squares.
[b] IV = Input Variables, PCP = Physicochemical properties, FT-NIR = Fourier Transform Infra-Red spectra.
[c] MO = Multi-Output model.
[d] RGB = Real Gasoline Blends.
[e] IFRP = Inputs from Full Refinery Plant process. RTO = Real Time Optimization, AAD = Average absolute deviation, SE = Standard Deviation, RS = Reverse Standardization.
[f] OP = Output Parameters.

real refinery plants may benefit from ML models able to predict multiple output properties along time series. IFPMTL methodology is able to predict with high overall Sp and Sn the levels of 45 output properties. The methodology is flexible enough to take into consideration inputs from multiple lines, streams, and operations in a refinery. EGS heuristic based on expert criteria combined with automatic feature selection techniques give the better results in this cases. The use of non-linear models like ANNs does not improve the results confirming the validity of the linear hypothesis. Using FT-NIR experiments as internal control demonstrated the robustness of the IFPTML to changes on the experimental technique used to determine the output variables. The present IFPTML opens a gate to the multi-output computational optimization of current and new alternative fuel blends in gasoline production.

CRediT authorship contribution statement

Harbil Bediaga: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **María Isabel Moreno:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sonia Arrasate:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **José Luis Vilas:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Lucía Orbe:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Elías Unzueta:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft,

Writing – review & editing. **Juan Pérez Mercader:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Humberto González-Díaz:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The SPRI ELKARTEK program grant (KK-2019/00037) funded the PhD scholarship of H. Bediaga. M.I. Moreno, S. Arrasate, J.L Vilas, declares no competing interests, they are professors working on KK-2019/00037 hired by UPV/EHU and do not have contractual relationships with PETRONOR S.A. L. Orbe, and E. Unzueta, are hired by Petronor Innovación S.L. a research unit of PETRONOR S.A., this is declared on affiliations and no have other competing interests. J. Pérez Mercader has a contract as consultant with PETRONOR S.A. and no have other conflict interests.

Acknowledgments

The authors acknowledge financial support from Basque government SPRI ELKARTEK program grant (KK-2019/00037). The authors are very especially grateful for the scientific lobbing and networking activity developed by F. Temprano towards the genesis of this project. The authors greatly appreciate and thanks expert consulting services from F. Temprano, J.I. Goicolea, and J.P. Gómez-Martin. It included state-of-art discussions and technical recommendations made on brain-storm meetings sponsored by the present grant. The authors also acknowledge partial financial support from research grants of Ministry of Economy and Competitiveness, MINECO, Spain (FEDER CTQ2016-74881-P) and Basque Government (Eusko Jaurlaritza) consolidation groups grant (IT1045-16). G.D.H. personally acknowledges the support of IKERBASQUE, Basque Foundation for Science. J.P.-M. thanks Petronor S.A. for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fuel.2021.122274.

References

[1] Singh D, Sharma D, Soni SL, Sharma S, Kumari D. Chemical compositions, properties, and standards for different generation biodiesels: a review. Fuel 2019;253:60–71.
[2] Schifter I, Díaz L, Sánchez-Reyna G, González-Macías C, González U, Rodríguez R. Influence of gasoline olefin and aromatic content on exhaust emissions of 15% ethanol blends. Fuel 2020;265:116950.
[3] Corach J, Sorichetti PA, Romano SD. Permittivity of gasoline/methanol blends. Application to blend composition estimation. Fuel 2019;258:116169.
[4] Paul S, Sarkar B. An exploratory analysis of biofuel under the utopian environment. Fuel 2020;262:116508.
[5] Kupka T, Mancini M, Irmer M, Weber R. Investigation of ash deposit formation during co-firing of coal with sewage sludge, saw-dust and refuse derived fuel. Fuel 2008;87:2824–37.
[6] Wang C, Prakash A, Aradi A, Cracknell R, Xu H. Significance of RON and MON to a modern DISI engine. Fuel 2017;209:172–83.
[7] Yuan H, Chen Z, Zhou Z, Yang Y, Brear MJ, Anderson JE. Formulating gasoline surrogate for emulating octane blending properties with ethanol. Fuel 2020;261:116243.
[8] Foong TM, Morganti KJ, Brear MJ, Da Silva G, Yang Y, Dryer FL. The octane numbers of ethanol blended with gasoline and its surrogates. Fuel 2014;115:727–39.
[9] Santos VHJMd, Ketzer JMM, Rodrigues LF. Classification of fuel blends using exploratory analysis with combined data from infrared spectroscopy and stable isotope analysis. Energy Fuels 2017;31:523–32.
[10] Balabin RM, Safieva RZ. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 2008;87:1096–101.
[11] Balabin RM, Safieva RZ. Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data. Fuel 2008;87:2745–52.
[12] Rokni HB, Gupta A, Moore JD, Mhugh MA, Bamgbade BA, Gavaises M. Purely predictive method for density, compressibility, and expansivity for hydrocarbon mixtures and diesel and jet fuels up to high temperatures and pressures. Fuel 2019;236:1377–90.
[13] Maksimuk Y, Antonava Z, Krouk V, Korsakova A, Kursevich V. Prediction of higher heating value based on elemental composition for lignin and other fuels. Fuel 2020;263:116727.
[14] Li G, Hu Z, Hou F, Li X, Wang L, Zhang X. Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels. Fuel 2020;265:116968.
[15] Luo RC, Chih-Chen Y, Kuo Lan S. Multisensor fusion and integration: approaches, applications, and future research directions. IEEE Sens J 2002;2:107–19.
[16] Meng T, Jing X, Yan Z, Pedrycz W. A survey on machine learning for data fusion. Information Fusion 2020;57:115–29.
[17] Willett P. Combination of similarity rankings using data fusion. J Chem Inf Model 2013;53:1–10.
[18] Whittle M, Gillet VJ, Willett P, Loesel J. Analysis of data fusion methods in virtual screening: theoretical model. J Chem Inf Model 2006;46:2193–205.
[19] Whittle M, Gillet VJ, Willett P, Loesel J. Analysis of data fusion methods in virtual screening: similarity and group fusion. J Chem Inf Model 2006;46:2206–19.
[20] Chen J, Holliday J, Bradshaw J. A machine learning approach to weighting schemes in the data fusion of similarity coefficients. J Chem Inf Model 2009;49:185–94.
[21] Skrobot VL, Castro EVR, Pereira RCC, Pasa VMD, Fortes ICP. Use of principal component analysis (PCA) and linear discriminant analysis (LDA) in gas chromatographic (GC) data in the investigation of gasoline adulteration. Energy Fuels 2007;21:3394–400.
[22] Mohamed Ismail H, Ng HK, Queck CW, Gan S. Artificial neural networks modelling of engine-out responses for a light-duty diesel engine fuelled with biodiesel blends. Appl Energy 2012;92:769–77.
[23] Balabin RM, Lomakina EI. Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst 2011;136:1703–12.
[24] Hill T, Lewicki P. STATISTICS methods and applications. A comprehensive reference for science industry and data mining. Tulsa: StatSoft; 2006.
[25] Bediaga H, Arrasate S, Gonzalez-Diaz H. PTML combinatorial model of ChEMBL compounds assays for multiple types of cancer. ACS Comb Sci 2018;20:621–32.
[26] Blay V, Yokoi T, González-Díaz H. Perturbation theory-machine learning study of zeolite materials desilication. J Chem Inf Model 2018;58(12):2414–9.
[27] Ferreira da Costa J, Silva D, Caamano O, Brea JM, Loza MI, Munteanu CR, et al. Perturbation theory/machine learning model of ChEMBL data for dopamine targets: Docking, synthesis, and assay of new l-prolyl-l-leucyl-glycinamide peptidomimetics. ACS Chem Neurosci 2018;9:2572–87.
[28] Simón-Vidal L, García-Calvo O, Oteo U, Arrasate S, Lete E, Sotomayor N, et al. Perturbation-theory and machine learning (PTML) model for high-throughput screening of parham reactions: Experimental and theoretical studies. J Chem Inf Model 2018;58(7):1384–96.
[29] Nocedo-Mena D, Cornelio C, Camacho-Corona MDR, Garza-Gonzalez E, Waksman de Torres N, Arrasate S, et al. Modeling antibacterial activity with machine learning and fusion of chemical structure information with microorganism metabolic networks. J Chem Inf Model 2019;59:1109–20.
[30] Santana R, Zuluaga R, Ganan P, Arrasate S, Onieva E, Gonzalez-Diaz H. Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. Nanoscale 2019;11:21811–23.
[31] Vasquez-Dominguez E, Armijos-Jaramillo VD, Tejera E, Gonzalez-Diaz H. Multioutput perturbation-theory machine learning (PTML) model of ChEMBL data for antiretroviral compounds. Mol Pharm 2019;16:4200–12.
[32] R. Santana, R. Zuluaga, P. Ganan, S. Arrasate, E. Onieva Caracuel, H. Gonzalez-Diaz, PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives, ACS Combinat Sci, (2020).
[33] Concu R, Dias Soeiro Cordeiro MN, Munteanu CR, González-Díaz H. PTML model of enzyme subclasses for mining the proteome of bio-fuel producing microorganisms. J Proteome Res 2019;18(7):2735–46.
[34] D.R. Hipp, SQLite in; 2020.
[35] AENOR, Automotive fuels – Unleaded petrol – Requirements and test methods. UNE-EN 228:2013+A1:2017, in: I.y.C. Ministerio de Economía (Ed.), AENOR, BOE, 2017-09-20, pp. 97534–97539.
[36] López MI, Callao MP, Ruisánchez I. A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. Anal Chim Acta 2015;891:62–72.
[37] Ellison SLR, Fearn T. Characterising the performance of qualitative analytical methods: statistics and terminology. TrAC, Trends Anal Chem 2005;24(6):468–76.
[38] Abdollahi S, Davis A, Miller JH, Feinberg AW, Zhao F. Expert-guided optimization for 3D printing of soft and liquid materials. PLoS ONE 2018;13(4):e0194890.
[39] Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018;126:1763–8.
[40] Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–74.
[41] Majnik M, Bosnić Z. ROC analysis of classifiers in machine learning: a survey. Intell Data Anal 2013;17(3):531–58.
[42] Vanneschi L, Castelli M. Multilayer perceptrons. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology. Oxford: Academic Press; 2019. p. 612–20.

[43] Tatar A, Barati A, Najafi A, Mohammadi AH. Radial basis function (RBF) network for modeling gasoline properties. Pet Sci Technol 2019;37:1306–13.
[44] Cavalcante da Silva N, Caribé de Góes Massa AR, Domingos D, Amigo JM, das Virgens Rebouças M, Pasquini C, Pimentele MF. NIR-based octane rating simulator for use in gasoline compounding processes. Fuel 2019;243:381–9.
[45] Wang S, Liu S, Zhang J, Che X, Wang Z, Kong D. Feasibility study on prediction of gasoline octane number using NIR spectroscopy combined with manifold learning and neural network. Spectrochim Acta A Mol Biomol Spectrosc 2020;228:117836.
[46] Balabin RM, Safieva RZ, Lomakina EI. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques. Anal Chim Acta 2010;671(1-2):27–35.
[47] Al-Ghouti MA, Al-Degs YS, Amer M. Determination of motor gasoline adulteration using FTIR spectroscopy and multivariate calibration. Talanta 2008;76:1105–12.
[48] Albahri TA, Riazi MR, Alqattan AA. Octane number and aniline point of petroleum fuels, in: ACS Division of Fuel Chemistry, Preprints, 2002, pp. 710–11.
[49] Ghosh P, Hickey KJ, Jaffe SB. Development of a detailed gasoline composition-based octane model. Ind Eng Chem Res 2006;45:337–45.
[50] Forbes ASJF, Vermeer PJ, Wood SS. Model-based real-time optimization of automotive gasoline blending operations. J Process Control 2000;10:43–58.
[51] Pasadakis N, Gaganis V, Foteinopoulos C. Octane number prediction for gasoline blends. Fuel Process Technol 2006;87:505–9.
[52] Peinder PD, Visser T, Petrauskas DD, Salvatori F, Soulimani F, Weckhuysen BM. Prediction of long-residue properties of potential blends from mathematically mixed infrared spectra of pure crude oils by partial least-squares regression models. Energy Fuels 2009;23:2164–8.
[53] Silva NCD, Cavalcanti CJ, Honorato FA, Amigo JM, Pimentel MF. Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters. Anal Chim Acta 2017;954:32–42.