Clemens Hall, Bastian Rauch, Uwe Bauder, Manfred Aigner, Predictive Capability Assessment of Probabilistic Machine Learning Models for Density Prediction of Conventional and Synthetic Jet Fuels, Energy & Fuels 35 (2021) 2520–2530.

http://dx.doi.org/10.1021/acs.energyfuels.0c03779

Predictive Capability Assessment of Probabilistic Machine Learning Models for Density Prediction of Conventional and Synthetic Jet Fuels

Clemens Hall, Bastian Rauch, Uwe Bauder, Patrick Le Clercq, Manfred Aigner

DLR, German Aerospace Center, Institute of Combustion Technology, MAT, 70569 Stuttgart, Germany,

Corresponding Author Email: Clemens.hall@dlr.de

KEYWORDS

Aviation Fuel, Synthetic Fuel, Machine Learning, Modelling, Uncertainty Quantification, Predictive Capability, Gaussian Process Regression

ABSTRACT

Machine Learning (ML) models are increasingly applied in the field of jet fuel property predictions due to their ability of modelling a high number of complex composition-property relationships directly on measurement data. Their applicability is still limited as for safety relevant use cases like synthetic fuel approval or aircraft design consequences of prediction errors might be too severe to be acceptable. For Machine Learning algorithms the predictive capability strongly depends on the data utilized for the training for the models. Predictions for fuels that that differ from the training data might have uncertainties that need to be systematically considered. We present an approach of utilizing the probabilistic ML algorithm Gaussian Process Regression to model jet fuel density over a range of -40 to 140 °C and estimating the uncertainty that results from limited training data. To assess the influence of synthetic fuels on the predictive capability two models are studied, one trained exclusively on conventional fuels data and the other one trained on the same conventional fuels and additional synthetic fuels. The predictive capability of the models is assessed using metrics to rate the accuracy and precision of the prediction, as well as the validity and reliability of the estimated prediction interval. Results show that prediction intervals can correctly be estimated by both models and a valid estimation of the predictive capability is possible. Furthermore, the addition of synthetic fuels data drastically improves the accuracy, reduces the uncertainty, and is necessary to achieve adequate predictions for the considered hold-out fuels. The presented method is transferable and can be applied for different probabilistic models and jet fuel properties

1 INTRODUCTION

The impact of jet fuels on the aircraft and the surrounding infrastructure are complex and manifold. To ensure their safe application an extensive evaluation process, regulated by the ASTM D4054 [1], has to be executed for each potential fuel production pathway and corresponding fuel candidate. The second tier of this process tests the fit-for-purpose properties, which refer to properties inherent of a petroleum-derived fuel. These fit-for-purpose properties are not controlled by a specification in ASTM D4054, but are considered critical to engine, airframe and fuel system performance. New fuel candidates assessed in the framework of ASTM D4054 have to lie within a given range of experience to guarantee safe operation. In contrast to specification properties they are investigated over a specification range e.g. a temperature interval.

The fuel amounts necessary for the fuel approval process often exceed the production capabilities of facilities producing synthetic fuels, especially at early stages in development. This restrains the development and introduction of promising solutions for the pursuit goal of CO_2 reduction in aviation. Model-based prediction of the fuel properties for a screening based solely on the fuel composition could support the test process and development of new and optimized production pathways and accelerate their approval. Since measurement of composition is possible with volumes below 1 ml, by the use of two-dimensional gas chromatography (GCxGC), a quick and cost-efficient assessment of essential properties would be possible. However, the high safety requirements of the jet fuels demand highly accurate and reliable models, which are adequate for predicting the properties for a large variety of fuels.

As jet fuels are mixtures of hundreds of different possible hydrocarbon molecules, the fuel composition can vary, especially between conventional and synthetic fuels. *Figure 1* displays this variety of the fuel composition of a typical conventional jet fuel on the left and a composition of an exemplary synthetic fuel on the right as an extreme example. The species present in the fuel are displayed with their corresponding mass fractions and respect to the number of carbon atoms they contain. The exemplary synthetic fuel differs drastically from the conventional fuel, with a composition dominated by only a distinct number of iso-alkane molecules and some hydrocarbon families usually found in conventional fuel not being present.

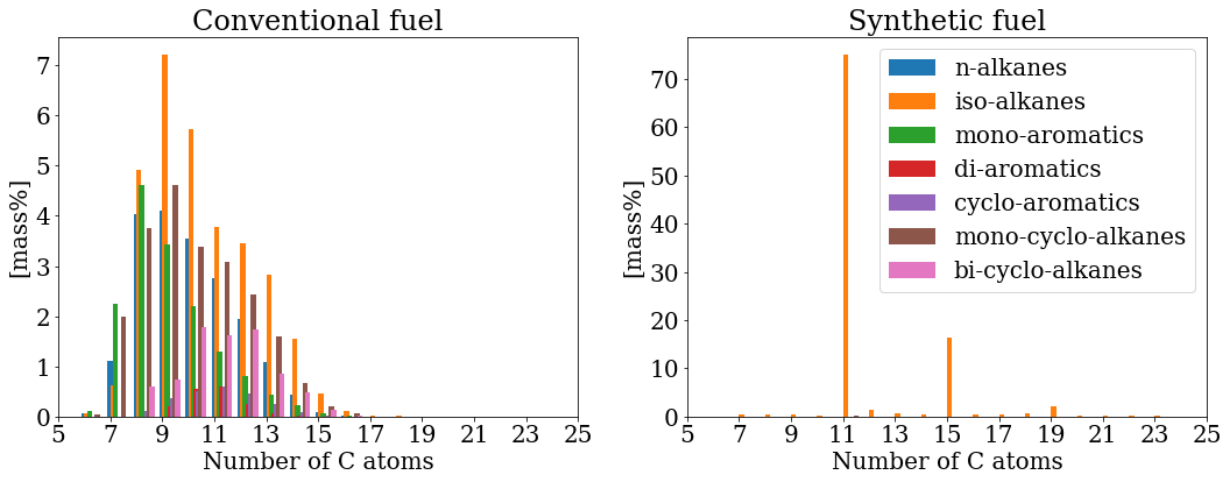

Figure 1: Exemplary fuel compositions of a conventional fuel (left) and of a synthetic fuel (right)

The large variety of possible fuel molecules and the resulting number of potential intra- and intermolecular relations, that determine the fuel properties, pose a challenge for property modelers. In recent years data based Machine Learning algorithms are increasingly applied for this modelling problem [2]. Their flexibility of relating different fuel composition representations with property data enabled different modelling approaches. They range from linking the molecular structure to properties of fuel components, the Quantitative Structure-Property Relationships (QSPR) method [3] [4] [5], to models correlating measurements of the composition directly with the properties [6] [7]. The QSPR method predicts properties for individual molecules. For jet fuels this method therefore relies on information regarding the exact molecules present in the fuel or valid representatives and a sufficiently accurate physical model for the mixing of those molecules. Therefore, the method needs an extensive database with information about the molecular structure and the division of the molecular structure in quantitative structures for correlation [4] [5]. The other method correlates measurement signals [8] [9] [10] or measurement results [6] [11] of the fuel composition directly with the properties. Thereby no assumption about representative molecules and the division of their structure and no mixing rules are required. However, the measurements in the dataset used for the development of the model have to provide correlative relation and a sufficient variance to sufficiently assign input and output data. Studies correlated two- dimensional gas chromatography measurements with measurements of jet fuel the densities at 15 °C and observed a prediction error of 0.205 % [6] and 0.107 % [11]. The GCxGC method allows a very detailed and comprehensive description of the composition of a fuel. The detected species are assigned to a matrix based on their hydrocarbon family and their number of carbon atoms. Since this evaluated GCxGC data is part of the new ASTM D4054 fast track certification process [1], it was chosen as input data for this work.

The application of those algorithms for safety relevant use cases like predictions for fuel fit-for-purpose testing requires an extensive assessment of the inter- and extrapolation behavior of the model according to

Oberkampf and Roy [12] p. 36. Furthermore, this assessment has to be related to the adequacy requirements of each specific use case the model will be applied on. The predictive capability thereby has to be estimated under consideration of potential uncertainties that might limit the application of the model for the desired use case.

For Machine Learning algorithms a major reason for uncertainties in the prediction is their strong dependence on the utilized training data [13]. Limited training data also limits the predictive capability and extrapolation outside of the training range results in uncertain and potentially false predictions that are unacceptable for the safety relevant use case. An estimation of the uncertainty of the prediction is therefore needed to illustrate uncertainty as a result of limited training data, which can then be used for the assessment of the predictive capability of the model. In Machine Learning, probabilistic algorithms can be used to estimate this uncertainty. In contrast to deterministic algorithms they do not only return the one value with the highest probability, but rather a distribution of possible values. The distribution corresponds with the certainty of the prediction based on the training data [14] [15]. For inputs that are "similar" to examples in the training data, the distribution is narrow and the prediction certain, for unfamiliar inputs the distribution broadens. In a recent study, Rocha and Sheen utilized this concept using a bootstrap approach for the uncertainty estimation of predictions for density and viscosity of fuels at specific temperatures [16]. The observed uncertainties described variability of measurements and provided additional trustworthiness of the prediction on the basis of the utilized data. We extend this concept using a Gaussian Process Regressor model (GPR) that predicts the uncertainty distribution as part of the prediction and utilize the estimated uncertainty as part of the predictive capability assessment. The prediction intervals of the distribution are used to estimate validity and reliability of the prediction, as well as the expressed precision. Together with the accuracy we estimate the predictive capability using a quantitative metric for each of the aspects. These metrics are put into relation of the requirements that we derived from the ASTM D4054 standard practice for fit-for-purpose properties [15] to rate if the model predicts adequate results for the intended use case.

If the calculated metrics comply with defined limits that we derived as adequacy limits for the fit-for-purpose prediction, the results are considered adequate for the fit-for-purpose property prediction. To investigate the influence of the training data on the predictive capability of the model, the assessment is executed on two models, trained on different datasets: One trained exclusively on data from 54 conventional fuels that were recorded in the CRC World fuel survey of 2006 [17]. The other one trained on additional 12 synthetic fuels. The predictive capability is thereby inferred using the results of the cross-validation of the two datasets. The findings of the predictive capability assessment are tested on the adequacy of the prediction of three hold-out fuels that are chosen as representative exemplary cases for the application domain; one conventional Jet A-1 fuel and two synthetic fuels. The conventional Jet A-1 fuel represents a typical crude oil based fuel. For the synthetic fuels we chose a fuel produced by the alcohol to jet process

(ATJ) consisting mainly of iso-alkanes and fuel produced from hydro processed esters and fatty acids (HEFA).

2 CHARACTERIZATION OF FUEL DATASET

For this work fuel composition data is correlated with the jet fuel density, a temperature dependent property essential for the fuel certification after ASTM D4054 [1]. This temperature dependent behavior is critical to engine, airframe and fuel system performance and design. For the fit-for-purpose testing after ASTM D4054, Tier 2 the density behavior is therefore investigated over a temperature range from -40 to 140 °C. Since jet fuels are complex mixtures, composed of hundreds of different hydrocarbon molecules, the composition can vary depending on feedstock and production pathway. The exact identification of each molecule is presently not possible or too expensive [11]. Two-dimensional gas chromatography (GCxGC) is a comprehensive analytical way of determining fuel compositions [11] and has been adopted into the ASTM D4054 fast track certification process. The detected species are assigned to a matrix bin based on their hydrocarbon family and their number of carbon atoms. For this work hydrocarbon molecules in a range of 1 to 25 carbon atoms from the seven families n-alkanes, iso-alkanes, mono-cyclo-alkanes, bi-cyclo-alkanes, mono-aromatics, cyclo-aromatics, and di-aromatics are considered. Information about the molecular formula of the families is given in the supporting material. Tri-cyclo-alkanes were lumped into the group of bi-cyclo-alkanes. Since the fraction of tri-cyclo-alkanes is marginal (in all cases below 1 mass%) resulting error can be neglected.

2.1 Training and validation data

The performance of a Machine Learning algorithm is highly determined by the dataset that it was trained and tested on. Number and variance of the datasets over the range of interest strongly influence the prediction capability of the model. In the scope of this work we investigate this influence using two datasets: One solely composed out of conventional fuels form the CRC world fuel survey 2006 (CRC fuels) [17] with 54 individual fuels and 335 data points, and a second dataset that additionally contains 12 synthetic fuels with 31 data points from the DLR jet fuel database (CRC & Syn. fuels). Multiple datapoints at different temperatures were provided to give the algorithm a sufficient database to correctly fit and capture the physical trend of the density over the temperature. Jet fuels with the fuel type Jet A, Jet A-1, JP-8, JP-5 and TS-1 are classified as conventional fuels; fuels with other fuel types as well as blends are classified as synthetic fuels

As visual representation of the number and variance of the two datasets, *Figure 2* shows the box plots for the mass fraction of each fuel family for the CRC fuels in blue and CRC & Syn. fuels in black. The box plot displays the median, the first and third quartile, as well as the minimum and maximum values of the

data. The fractions of the individual fuel are displayed as scatter dots, conventional fuels are thereby displayed in light blue, whereas the synthetic fuels are displayed in green. Continuous lines show the fractions of the three hold-out fuels, which will be discussed later on. The comparison of the box plots of the two datasets shows a great similarity between the distributions of the datasets. The presence of the synthetic fuels in the CRC & Syn. fuels dataset does not alter the overall distribution, shown by the boxplots, significantly. In contrast to the compositions of the synthetic fuels, which differ strongly from the mean of the conventional fuels datasets: the synthetic fuels are therefore considered outliers of the boxplots. This is visualized by the individual scatter plots of the synthetic fuels, which lie outside the box plot range. The positions of the lines of the hold-out fuels relative to the scatter dots illustrate their similarity to conventional and synthetic fuels.

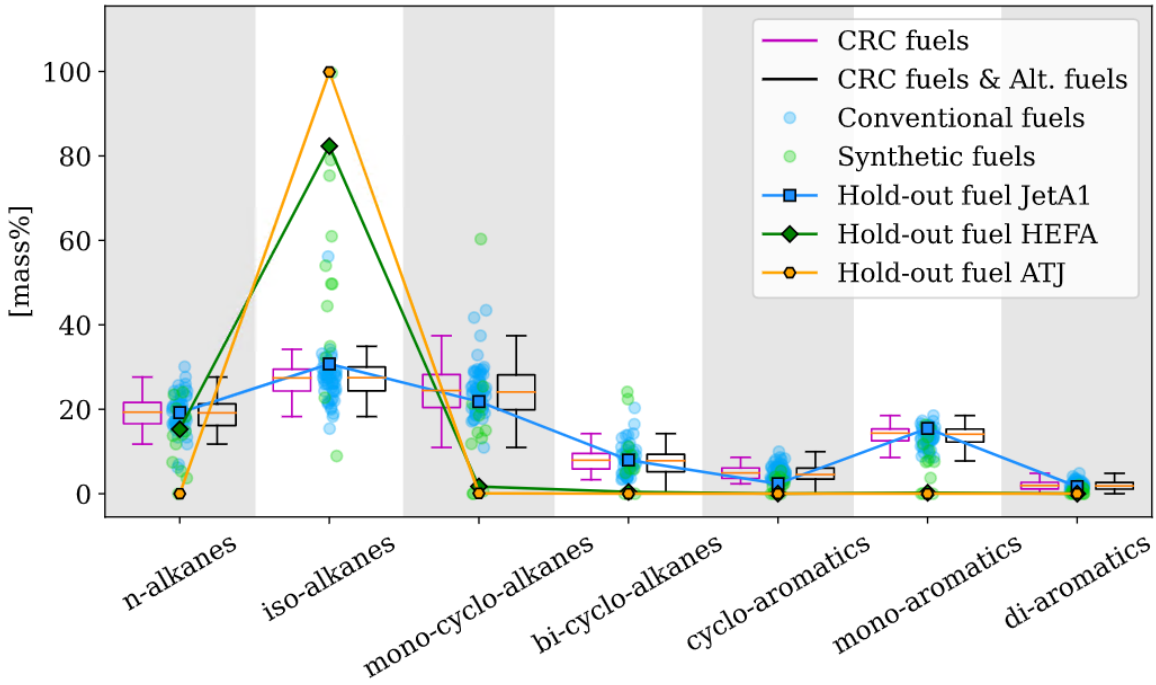

Figure 2: Comparison of test fuels as lines and training data as boxplot with respect to considered fuels families

Since the modelled fuel property does not only depend on the composition but also varies with temperature, training data is needed for the temperature range of interest. *Figure 3* shows the number of data points available for the conventional and synthetic fuels of the datasets for a given temperature with bins of 5 K. The plots illustrate the unequal distribution of available data. Again, the fraction of conventional fuels is colored in light blue and the fraction of the synthetic jet fuels in green. The plots show, that most of the data is available at very distinct temperature values. This is due to the data origin; the great extent of the data originates from technical reports or certificates, where data is needed at specific temperatures to ensure

safe operability of the fuel. The influence of the data distribution with respect to the temperature and the influence of this distribution on the uncertainties will be investigated in this work.

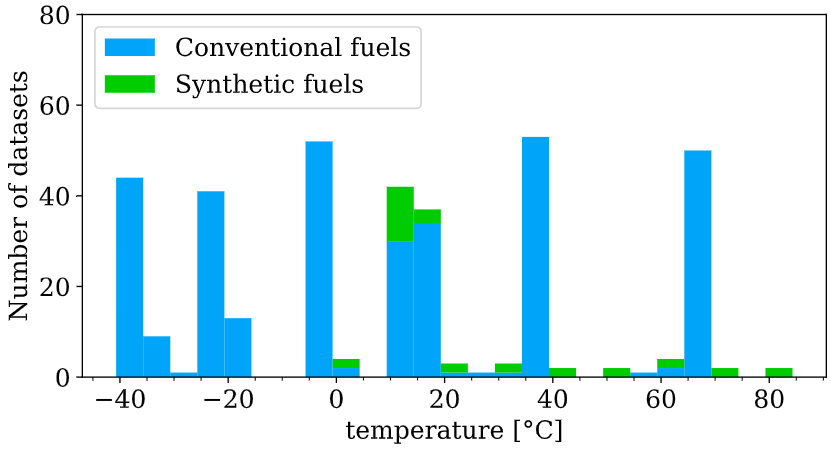

Figure 3: Bar chart of available datasets for density in the DLR database with respect to the temperature

2.2 Hold-out data

The three fuels chosen as unseen hold-out data differ significantly in their compositions. They were chosen to investigate the predictive capability of the trained algorithms over a broad range of possible jet fuels that represent different cases in the application domain. *Table 1* tabulates the hold-out fuels with the summed-up mass fractions (in mass%) of the considered families.

Fuel	n-alkanes	iso-alkanes	mono-cyclo-alkanes	bi-cyclo-alkanes	mono-aromatics	cyclo-aromatics	di-aromatics
Jet A-1	19.2	30.7	21.8	8	15.5	2.4	1.7
ATJ	0	99.9	0.1	0	0	0	0
HEFA	15.265	82.32	1.721	0.413	0	0.238	0.04

Table 1: Table of considered test fuels with their composition summed up to the corresponding families (in mass%)

The Jet A-1 fuel has a typical composition of a crude oil-based fuel, containing significant mass fractions for all considered hydrocarbon families. The synthetic fuels differ especially with respect to their mass fractions for iso-alkanes and hydrocarbon families containing cyclic hydrocarbon molecules. The ATJ fuel is almost exclusively composed of heavily branched iso-alkanes. The distinguishing composition is due to the feedstock and the production pathways of the synthetic jet fuels.

A visual comparison of the hold-out fuels with the utilized training and validation data is shown in *Figure 2*. As a typical representative of the conventional fuels, the Jet A-1 lies close to the median of the box plots, whereas the synthetic hold-out fuels lie at the extremes close to the synthetic fuel outliers of the CRC & Syn. fuels dataset. The influence of the composition of the dataset and the presence of synthetic fuels on the prediction capability will be reviewed in this work.

3 PROBABILISTIC REGRESSION MODEL

As stated earlier, for this work we use the GPR [14], a proven probabilistic algorithm for regression problems with high dimensional data input. Probabilistic algorithms do not only return one deterministic value as prediction but rather a distribution of possible values or parameters that define a distribution [14]. In the training the model is fitted using the training data set so its predicted distribution describes underlying noise e.g. due to different measurements for the same fuel, underfitting of the model and prediction uncertainties due to lack of data. These aspects induce uncertainty in the model fit which prevents one definite prediction. The distribution therefore corresponds with the uncertainty of the model fit [14] [15]. The distribution is described by the value with the highest likelihood, the expectation and an upper and lower prediction interval (PI) that describe the range where a set percentage of values are expected, based on the training data. For data that was not seen in training, the validity of this distribution has to be assessed. We review the validity of the predictions as part of the predictive capability.

3.1 Gaussian Process Regression

The GPR can be interpreted as an algorithm that approximates a distribution of regression functions $f(x)$ for training data $D = \{(x_i, y_i) | i = 1, \dots n\}$ composed of features X and labels Y with their corresponding probabilities [14]. Thereby not only the function with the best fit or highest probability, the mean function, is given as a result, but also the corresponding standard deviations. The parameters mean value and standard deviation define a Gaussian distribution for each predicted value. The model is fitted to the training data, so the distribution captures its noise and describes uncertainty of the prediction [14]. For this work this distribution and the corresponding quantiles are interpreted as prediction intervals. The validity of this interpretation will be reviewed with metrics as part of the predictive capability assessment. The distribution

is determined by a so called covariance function $k(x, x')$ or kernel, which characterizes the correlation between two different points in the process x and x'. This kernel can be composed out of different sub-kernels. In the scope of this work a weighted linear combination of three different kernels, with different weightings are used, which are listed below.:

Radial-basis kernel: $$k(x, x') = \exp\left(-\frac{(x - x')^2}{l_{RBF}}\right)$$ *Equation 1*

White kernel: $$k(x, x') = \delta,\ 0 \text{ if } x = x'$$ *Equation 2*

Rational Quadratic kernel: $$k(x, x') = \left(1 + \frac{(x - x')^2}{2\alpha * {l_{RQ}}^2}\right)^{-\alpha}$$ *Equation 3*

The Radial-basis kernel is a universal kernel for GPR applicable for most of the non-periodic regression problems. The White kernel deals with potential noise of data. The Rational Quadratic kernel is selected due to its characteristic of modeling smooth functions that do not vary too quickly. This kernel is equivalent of adding multiple Radial-basis kernels together. The weighted sum of all three kernels creates a covariance function that is able to cover the complexity of high-dimensional problems, handle the noisiness problem and return a smooth function as to be expected from physical behaviors. The individual weights are adjusted by hyperparameter optimization using a Bayesian optimization approach and a Gaussian Process with a Radial-basis kernel to estimate the optimized values. As loss function we used the expectation maximization algorithm [14]. The optimization was carried out for 30 iterations. To the diagonal of the kernel matrix a value $alpha$ is added, that is adjusted during training and additionally captures noise in the training data. The parameters $l_{RBF}, l_{RQ}, \delta,\ \alpha$ and *alpha* are parameters which have to be adjusted during the training of the GPR. This task is typically accomplished by an optimizer, maximizing a likelihood function [14]. In this work, the "L-BFGS-B" optimizer [18] and the log marginal likelihood [14] are used. The process is initialized by a distribution of random functions. After the determination of the parameters in the optimization loop, the predictions of test data are compared with the actual test values. Based on their accordance, the covariance function is adjusted via Bayes theorem. The prediction of values for regions, where no training data is available, is possible via the assumption of a multidimensional Gaussian distribution N as shown in *Equation 4*. The unknown (subscript u) value y_u from the test set can be calculated with known values (subscript k) from the trainings set via *Equation 5*. Therefore the mean function μ and the covariance matrix, which is composed out of the matrices of the covariance for the trainings set K_{kk}, the trainings and the test set K_{ku}, as well as the test set K_{uu} itself, are used. This allows the calculation of the mean function and the corresponding variance for the unknown location.

$$\begin{pmatrix} Y_k \\ y_u \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_k \\ \mu_u \end{pmatrix}, \begin{pmatrix} K_{kk} & {K_{ku}}^T \\ K_{ku} & K_{uu} \end{pmatrix}\right) \qquad \textit{Equation 4}$$

$$y_u|Y_k \sim N\left(K_{ku}{K_{kk}}^{-1}Y_k, K_{uu} - K_{ku}{K_{kk}}^{-1}{K_{ku}}^T\right) \qquad \textit{Equation 5}$$

3.2 Training and validation metrics

For the training and validation of the model we introduce a metric that quantifies accuracy of the prediction: the Root Mean Squared Error (RMSE). It calculates the square root of the quadratic mean of all deviations from the predicted values y_{pred} and the measured values y_{test}, see *Equation 6*. Compared to a mean averaged error, large individual deviations are valued more strongly. The RMSE is utilized to rate the general fit of the model and observe potential overprediction. The metric is not part of the predictive capability assessment, which is presented in the following section.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{Test}} \left(y_{pred,i} - y_{test,i}\right)^2}{n_{Test}}} \qquad \textit{Equation 6}$$

4 PREDICTIVE CAPABILITY METRICS

For the assessment of the predictive capability of the model and the assessment of the adequacy of the prediction for the hold-out fuels, we introduce three metrics that measure (1) the accuracy of the prediction, (2) the validity of the PI and (3) the precision of the prediction. These metrics consider the deviation of the predictions for the measurements by the accuracy, the validity of distribution by the validity of the PI and the reasonableness of the distribution measuring the width of the PI as precision. To measure the accuracy of the predictions in a way that it can be related to the prediction task, we use the Mean Relative Error (MRE), see *Equation 7*. The MRE calculates the average relative deviation of the predicted value $y_{pred,i}$ from the actual test value $y_{test,i}$. If the MRE is smaller than a predefined percentage (accuracy limit), the results are on average assumed to be accurate enough for the prediction task.

$$MRE = \frac{1}{n_{Test}} \sum_{i=1}^{n_{Test}} \frac{\left(y_{pred,i} - y_{test,i}\right)}{y_{test,i}} * 100\ \% \qquad \textit{Equation 7}$$

To assess the validity (2) of the PI and precision (3), we use two metrics from the uncertainty quantification: The Prediction Interval Coverage Probability (PICP), see *Equation 8*, and the Normalized Mean Prediction Interval Width (NMPIW) [19], see *Equation 9*.

$$PICP = \frac{1}{n_{Test}} \sum_{i=1}^{n_{Test}} c_i * 100\ \% \qquad \textit{Equation 8}$$

The PICP calculates the average probability that a measured test value $y_{test,i}$ lies inside predicted lower limit ${y_{PI}}^{l}$ and upper limit ${y_{PI}}^{u}$ of the predicted PI. The constant c_i is a boolean value; it is 1 if ${y_{PI}}^{l} < y_{test} < {y_{PI}}^{u}$ and 0 otherwise. The PICP is the sum of all test values divided by the number of test values n_{Test}. With the PICP an assessment of the validity of the PI is possible. If the PICP and the set percentage of the PI are comparable, the values lie on average inside the PI and the PI is considered to be valid. If this is true for training and testing the PI are furthermore considered reliable. The NMPIW in contrast measures the relative size of the PI, by equating the mean width over all predicted values, normalized by a reference width [19] [20]. If the width of the PI is on average wider than the reference width, the NMPIW exceeds 100 %.

$$\mathrm{NMPIW} = \frac{1}{n_{Test}} \sum_{i=1}^{n_{Test}} \frac{{y_{PI,i}}^{u} - {y_{PI,i}}^{l}}{{\Delta y_{CRC}}^{max-min}(T)} * 100\ \% \qquad \textit{Equation 9}$$

For this work, the reference width was set to the difference between the minimum and maximum values of conventional fuels recorded in the CRC world fuel survey of 2006 [17] (${\Delta y_{CRC}}^{max-min}$). The CRC values are utilized for the screening of new synthetic jet fuels. To guarantee safe operation, fuel candidates have to lie within this range of experience. Since the CRC world fuel survey does not contain values for the whole temperature range of -40 to 140 °C, curves were fitted to the available measurement data. For the density a first order polynomial was fitted to the available values at 15 and 65°C. The fitted curves for the maximum and minimum values are plotted in *Figure 4*.

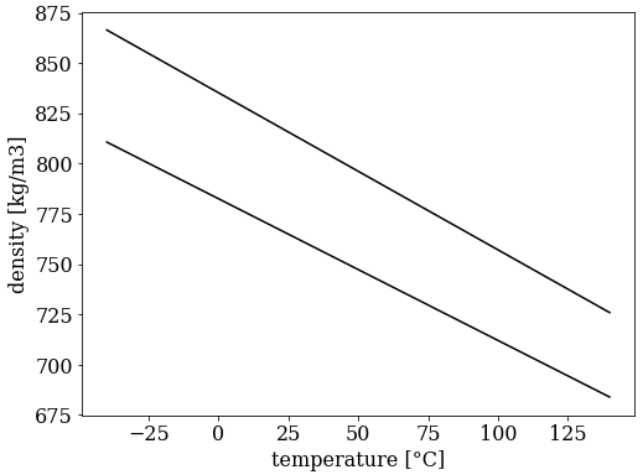

Figure 4: Fitted curves of the minimum and maximum values for density recorded by the CRC world fuel survey 2006

The plots show, that the difference between the minimum and maximum values varies with temperature. Therefore, the $\Delta {y_{CRC}}^{max-min}$ was calculated for each temperature using the fitting curves shown in *Figure 4*. To consider the PI over the whole range of interest, the NMPIW was calculated not only for the predictions of the measured values but also for the predictions over the complete range of -40 to 140 °C in 5 K steps for each fuel in cross-validation and for the hold-out fuels.

5 RESULTS

5.1 Cross-validation

The ability of the GPR to model the fuel density on the basis of the GCxGC input is assessed using a tenfold cross-validation. In this process the GPR is trained using 10 different random copies of the two datasets with a 75 % fraction for training and 25 % for testing. To avoid local optima in the training of the GPR, the optimization of the GPR was repeated with 10 random starting positions for the model parameters for each training. The predictions from the cross-validation and for the hold-out fuels are plotted over the measured values in unity plots in *Figure 5* for the model trained only on conventional fuels from the CRC world fuel survey (CRC fuels model) and *Figure 6* for the model that was trained on the dataset that additionally contained synthetic fuels dataset (CRC & Syn. fuels model). If prediction and measured value are in perfect accordance with each other, the marker lies on the unity line, which is indicated in black. As additional reference for the accuracy, grey dashed lines are plotted to illustrate a 1 % deviation range from the unity line. The predictions of conventional fuels are displayed in blue, predictions for synthetic fuels in green and orange. For the cross-validation the results for the training are shown on the left of the figures, whereas the

results of the testing are shown on the right. The results for the prediction of the hold-out fuels are displayed on top of the testing predictions on the right. The PI are displayed as individual error bars. The validity of the PI is reviewed as part of the predictive capability assessment later on. For the comparison of the model results for the two datasets, the mean RMSE of the folds are tabulated in *Table 2*.

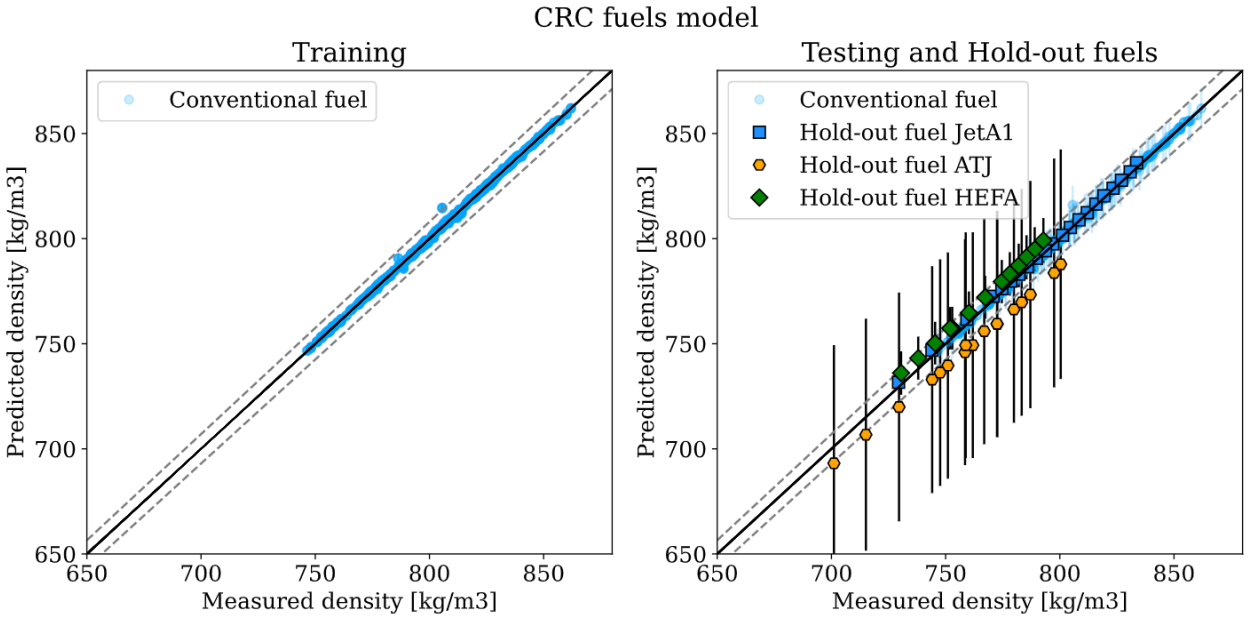

Figure 5: CRC fuels model unity plots of tenfold cross validation with 1% deviation lines for training data (left) and testing and hold-out data (right)

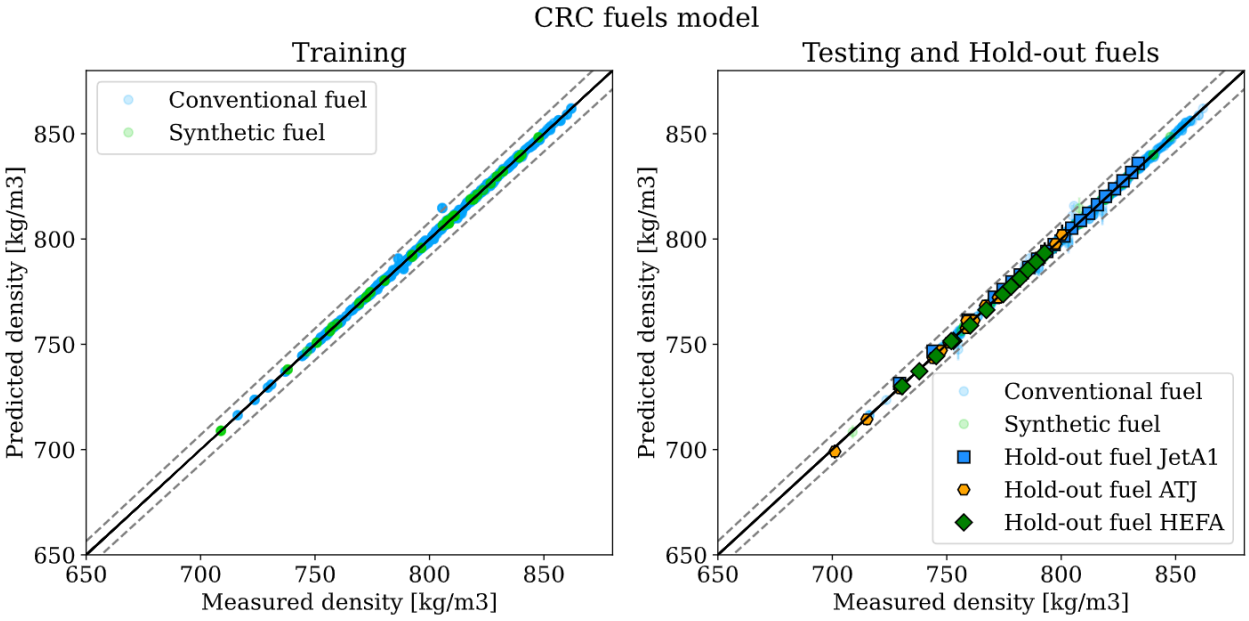

Figure 6: CRC & Syn. fuels model unity plots of tenfold cross validation with 1% deviation lines for training data (left) and testing and hold-out data (right)

	CRC fuels model		CRC & Syn. fuels model	
	RMSE [kg/m^3] Training	RMSE [kg/m^3] Testing	RMSE [kg/m^3] Training	RMSE [kg/m^3] Testing
Complete dataset	0.76	0.93	0.8	0.98
Conventional fuels	*0.76*	*0.93*	*0.82*	*0.98*
Synthetic fuels	-	-	0.47	0.98

Table 2: Mean RMSE of cross-validation results for the CRC fuels and CRC & Syn. fuels model

The predictions in *Figure 5* and *Figure 6* illustrate, that GPR is able to model the density on the basis of the GCxGC data from the chosen dataset. The predicted values of both conventional and synthetic fuels follow the unity line in the unity plots, which indicates no systemic errors. The majority of the density predictions lie inside the 1 % deviation region from the unity line. For the CRC fuels model 99.7 % of the predicted values lie inside the deviation lines for the training and 99.7 % for the testing. For the CRC & Syn. fuels model 99.6 % of the predicted values of the conventional and 100 % of the synthetic fuels lie inside the deviation lines for the training and 99.8 % of the conventional fuels and 100 % of the synthetic fuels for the testing. For comparison of the predictions of the hold-off fuels, the results of the CRC fuels model show stronger deviations and larger PI, indicating more uncertainty for the CRC fuels model in *Figure 5*. This will be discussed in more detail in the adequacy assessment of the hold-out fuels.

The mean RMSE in *Table 2* show slight differences in training and testing for both datasets, but the values have the same magnitude and therefore show no overfitting. The mean RMSE of both two datasets are almost identical, therefore the additional fraction of synthetic fuels does not influence the overall modelling. The comparison of the mean RMSE for the two fuel categories shows, that the GPR is able to model both conventional and synthetic fuels. For the synthetic fuels in the CRC & Syn. fuels dataset the mean RMSE of the testing is almost twice the value for the training, this might be due to the lower fraction of synthetic fuels, compared to the conventional fuels. Since the CRC fuels dataset did not contain synthetic fuels, a respective RMSE was not calculated.

5.2 Predictive capability assessment on cross-validation results

The results of the cross-validation show, that the GPR algorithm is able to model the density without a systemic error. The questions that remain are the question about its predictive capability for the intended use case and the question about the influence of presence of synthetic fuels in the training data. Therefore, the results of the cross-validation are assessed with respect to the accuracy, the validity of the uncertainty and the precision, using the three introduced metrics MRE, PICP and NMPIW. The results of the training set represent the interpolation and the results of the testing set represent the extrapolation. For the review of the adequacy of the results use-case specific thresholds are set to the three introduced metrics. As accuracy threshold a MRE of 1% is set. To review the validity of the distribution 95 % of the predicted values have to lie inside the 95 % PI, which is expressed by a PICP of 95 %. If the PI are valid both in training and in testing, the model is considered reliable. As threshold for the precision a NMPIW of 30 % is set. Results are therefore defined to be adequate for the fit-for-purpose use case:

1. If the predictions on average do not deviate more than 1 % from measurements
2. If 95 % of all the measurements are contained by the 95 % PI for training and testing
3. If the average width of PI is not wider than 30 % of the reference width.

MRE and PICP can only be calculated if actual validation data exist. However, the NMPIW does not depend on available data and can be calculated for every property prediction. It was therefore predicted for each fuel over the complete temperature interval of -40 to 140 °C. For prediction tasks of unknown fuels accuracy and reliability have therefore been taken from previous validation task. The calculated metrics are shown in *Table 3* for the CRC fuels model and in *Table 4* for the CRC & Syn. fuels model.

	CRC fuels model					
	MRE [%]		PICP [%]		NMPIW [%]	
	Training	Testing	Training	Testing	Training	Testing
Complete dataset	0.05	0.07	98.6	98.0	14.8	15
Conventional fuels	0.05	0.07	98.6	98.0	14.8	15
Synthetic fuels	-	-	-	-	-	-

Table 3: Mean MRE, PICP and NMWIP CRC fuels model

	CRC & Syn. fuels model					
	MRE [%]		PICP [%]		NMPIW [%]	
	Training	Testing	Training	Testing	Training	Testing
Complete dataset	0.05	0.07	98.5	97.4	11.1	11.4
Conventional fuels	0.06	0.07	98.3	97.3	11.1	11.4
Synthetic fuels	0.04	0.07	100	98.7	11.2	11.9

Table 4: Mean MRE, PICP and NMWIP for CRC & Syn. fuels model

The results show, that all calculated metrics for the models comply with the set thresholds for conventional fuels and for the CRC & Syn. fuels model also for synthetic fuels. The models therefore have the predictive capability to predict the density of jet fuels for the fit-for-purpose use case. Likewise to the RMSE of the cross-validation the values of the MRE and the PICP differ slightly, but are compliant for both cases. For the NMPIW the values stay almost constant for training and testing illustrating an average precision of less than half of the requirement. The comparison of the metrics for the two models shows an increase of the average precision (lower NMPIW) for the CRC & Syn. fuels model. This can be probably attributed to the increased number of fuels for training.

5.3 Property prediction for the hold-out fuels

After the cross-validation, the models are trained on the full dataset and used to predict the density of the three hold-out fuels to test the predictive capability assessment of the cross-validation. The density is predicted over the temperature range from -40 to 140 °C in 5 K steps. *Figure 7*, *Figure 8* and *Figure 9* display the results for all three hold-out fuels, for the CRC fuels model on the left and for CRC & Syn. fuels model on the right. The upper parts of the graphs show the actual density prediction. The lower part of the graphs show the Normalized Prediction Interval Width (NPIW) as blue line, a local variant of the NMPIW that assesses the PI of a single prediction over the temperature, with set NMPIW adequacy limit of 30 % displayed as orange line. The individual 95 % PI are displayed as a continuous band around the best estimate, which is the mean value for the GPR.. As an additional reference, the minimum and the maximum values of conventional fuels in the CRC world fuel survey from *Figure 4* are plotted as dashed grey lines. The range between the lines indicates the range of experience and serves as reference width for the precision

of the predictions. The lower part of the graph allows a grading of the local precision compared to the set threshold. If the NPIW lies below the 30 % line, the PI is compliant with the threshold at this temperature.

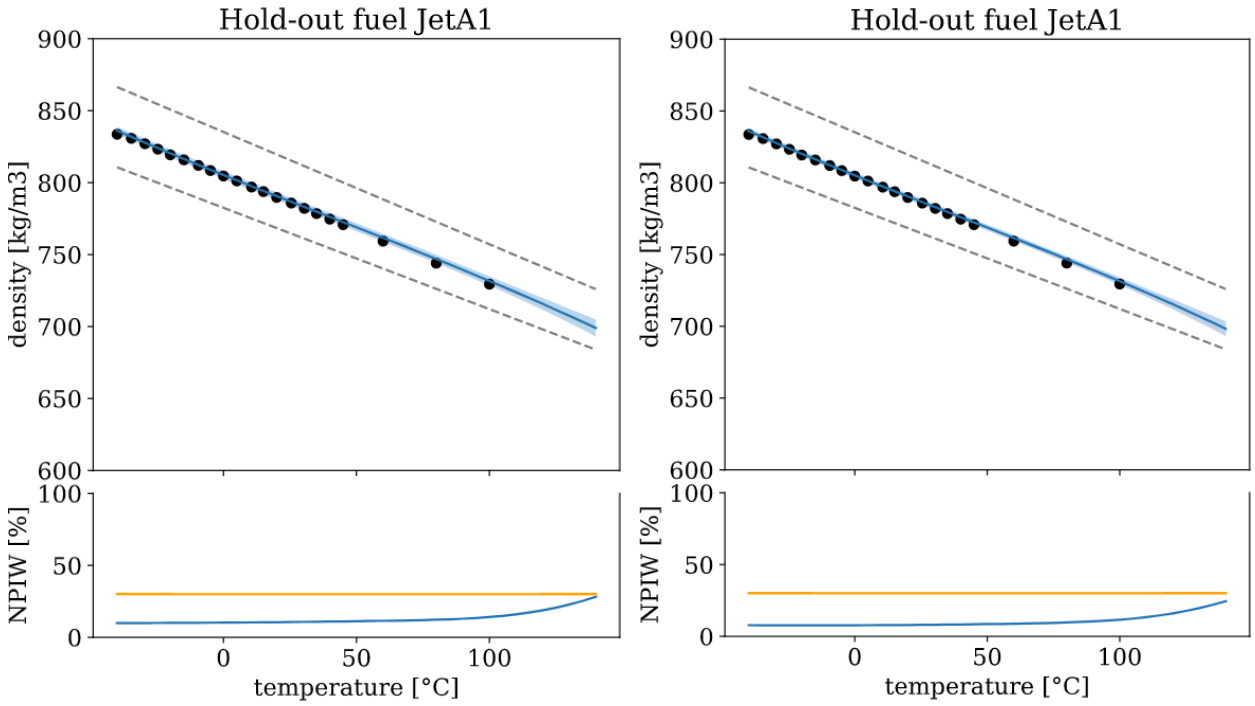

Figure 7: Prediction of density for conventional hold-out Jet A-1 fuel over a temperature range of -40 to 140 °C of CRC fuels model (left) and CRC & Syn. fuels model (right)

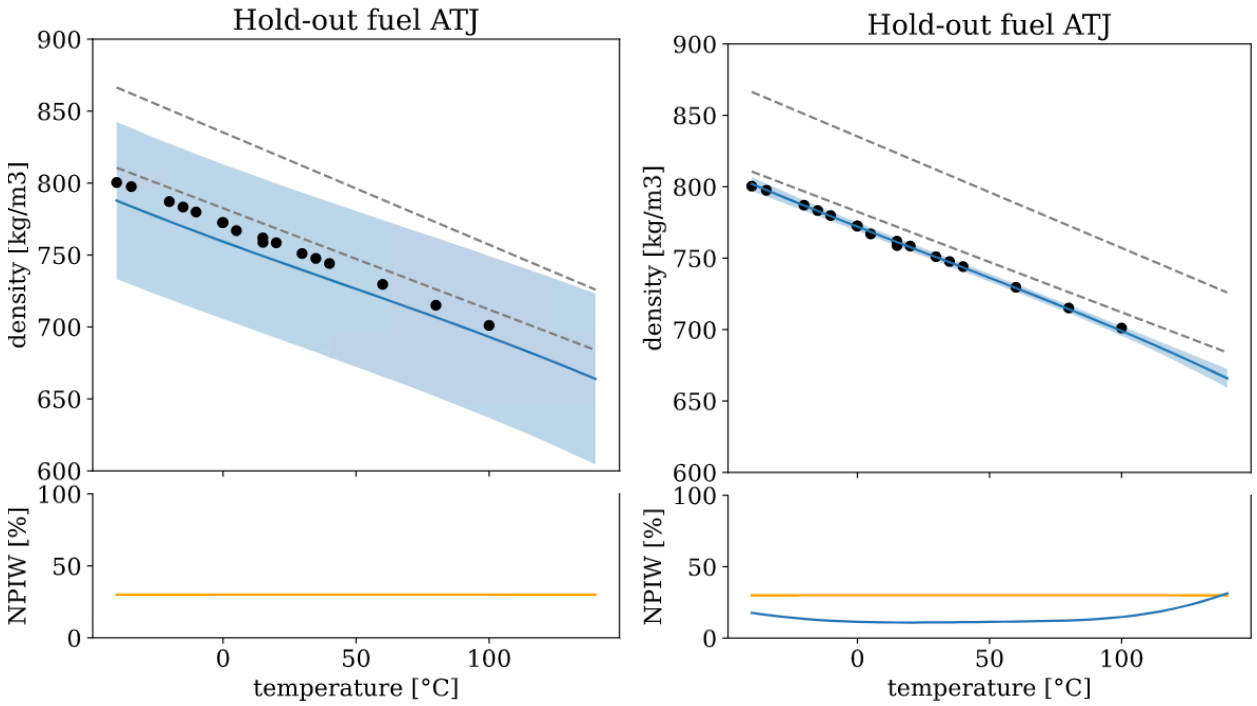

Figure 8: Prediction of density for synthetic hold-out ATJ fuel over a temperature range of -40 to 140 °C of CRC fuels model (left) and CRC & Syn. fuels model (right)

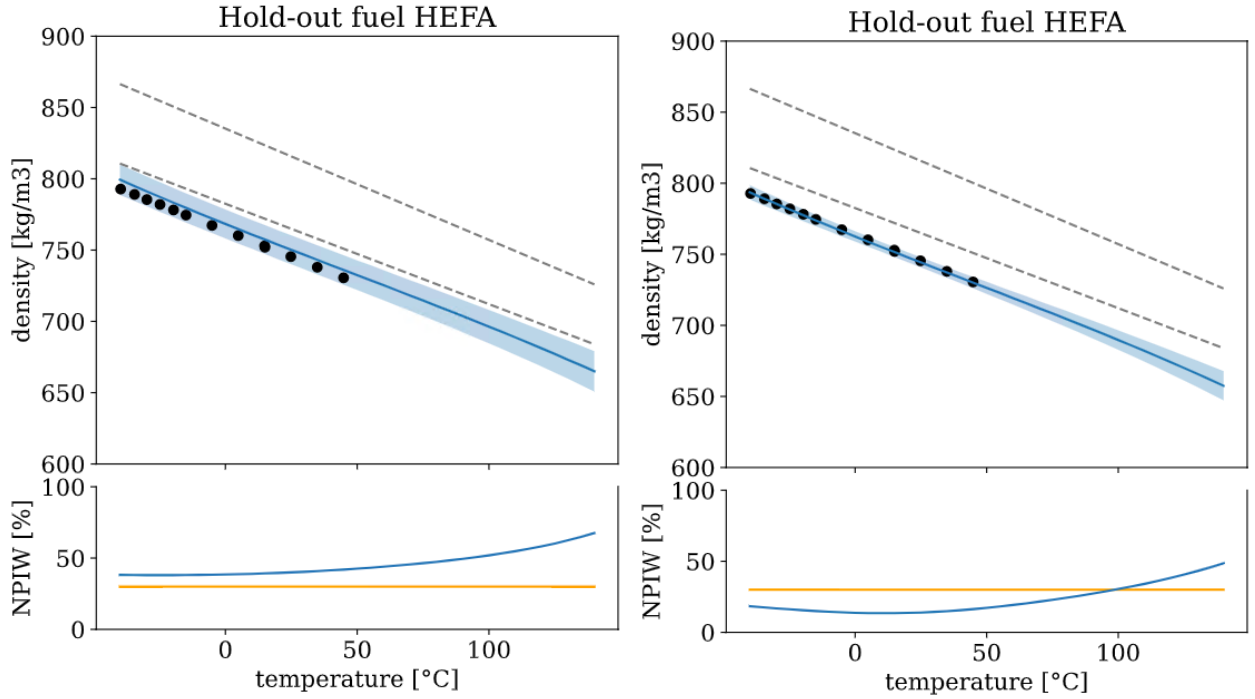

Figure 9: Prediction of density for synthetic hold-out HEFA fuel over a temperature range of -40 to 140 °C of CRC fuels model (left) and CRC & Syn. fuels model (right)

The predictions of all hold-out fuels in *Figure 7 - Figure 9* follow the overall physical trend of a decreasing density with increasing temperature. For higher temperatures above 100 °C the predictions do not follow the expected linear behavior over the temperature but show a decreasing gradient. This is also the region where the PI width widens, which indicates an increasing uncertainty. The increasing NMPIW curve in the lower part of the individual plots illustrates this clearly. The increase can be explained with the lack of data in this temperature region. A comparison with *Figure 3* illustrates that for those temperatures no values exist in the database. This results in a deviation from the linear trend, since a non-linear kernel is used for the GPR. The presence of additional data or the specific utilization of a kernel that follows the expected trend for this dimension could resolve this. The comparison of the PI and the measurements show, that the measurements are contained by the PI illustrating the reliability of the prediction.

The comparison of the predictions between the fuels and the models trained on the different datasets show significant differences. For the conventional fuel Jet A-1 the prediction is similar both for the CRC fuels model and for the CRC & Syn. fuels model. Since in both cases data of conventional fuels is present in the training data this is expected. For the two synthetic ATJ and HEFA fuels the predictions differ between the models. For the CRC fuels model the PI width is significantly larger compared to the CRC & Syn. fuels model. For the ATJ fuel predictions of the CRC fuels model, the PI width significantly exceeds the CRC reference range indicated by the NPIW curve in the lower plot not being visible in the 0-100 % range. Also the accuracy of the predictions is higher for the CRC & Syn. fuels model. Both the smaller PI width and the better accuracy can be explained by the presence of the synthetic fuels in the training data and therefore data that is more similar to the synthetic hold-out fuels. The comparison of the scatter dots with the lines of hold-out fuels in *Figure 2* illustrates the difference between the two synthetic fuels to the next conventional training fuel of the CRC fuels. For the CRC & Syn. fuels dataset similar synthetic fuels are present in the training data which results in a higher accuracy and precision. This observation demonstrates the influence of similar data and therefore data points that are close to the desired prediction. A further investigation about the influence of the similarity or distance of the training data on to the predictive capability and the three sub aspects accuracy, reliability and precision could be the next step for research for the application of Machine Learning models.

5.4 Adequacy assessment of the prediction results for the three hold-out fuels

The assessment of the predictive capability on the results of the cross-validation yield, that the models have the capability to predict the density for the use case of fit-for-purpose prediction. To review this assessment the three metrics for accuracy, validity and precision were calculated to determine the adequacy of the predictions and compare them with the findings of the assessment of the cross-validation results. The

metrics for the three hold-out fuels are listed in *Table 5* for the CRC fuels model and *Table 6* for the CRC & Syn. fuels model.

Fuel	CRC fuels model		
	MRE [%]	PICP [%]	NMPIW [%]
Jet A-1	0.1	100	13.1
ATJ	1.5	100	227.7
HEFA	0.6	100	45.85

Table 5: CRC fuels model: MRE, PICP and NMWIP for density prediction of hold-out fuels

Fuel	CRC and Syn. fuels model		
	MRE [%]	PICP [%]	NMPIW [%]
Jet A-1	0.13	100	10.48
ATJ	0.1	100	14.64
HEFA	0.09	100	22.88

Table 6: CRC & Syn. fuels model: MRE, PICP and NMWIP for density prediction of hold-out fuels

All in all the results confirm the finding of the predictive capability assessment in the cross-validation. For the predictions made with the CRC fuels model only the predictions of the conventional Jet A-1 hold-out fuel are adequate. Only this prediction would have been accepted on the basis of the calculated prediction. For the synthetic fuels, the estimated precision exceeds the limit of 30 % NMPIW. According to the adequacy criteria that we defined those predictions would have been rejected and measures need to be undertaken to improve the precision. For the ATJ fuel this would have been correct since the accuracy exceeds the limit of 1 % MRE. For the HEFA fuel the accuracy complies with the threshold, but the threshold of the precision is exceeded.

The predictions of the model that was trained on conventional and additional synthetic fuels the predictions of all three hold-out fuels comply with the threshold. Compared to the predictions of the model trained only

on conventional fuels the NMPIW is much lower and correctly indicates an adequate accuracy. The presence of synthetic fuels in the training data is therefore essential to achive the needed predictive capability. In comparison to the metrics of the predictive capability inferred by the cross-validation, the metrics of the adequacy assessment for predictions for the hold-out fuels differ. This is most probably because the average statistical truth does not hold true for the specific use case.

In conclusion, the overall assessment of the predictive capability is correct. The predictive capability assessment of the cross-validation is transferable for fuel categories that were present in the cross-validation. Furthermore, the NMPIW correctly indicates potential deviations for the hold-out fuels, for fuels that were not present in the training data. However, the a posteriori calculated MRE for the hold-out fuels is bigger than the MRE calculated during the cross-validation. An estimation of the prediction accuracy is not yet possible a priori for the prediction and further research is needed to estimate the accuracy a priori based on the cross-validation results and the utilized training data.

6 SUMMARY AND OUTLOOK

To ensure the operational safety of new synthetic jet fuels more than 100 fuel-related properties have to be critically evaluated. Machine Learning based algorithms bear the potential of modelling a wide variety of fuel properties. The question if those models can be utilized in safety-critical domains like fuel approval and aircraft design is bound to an accurate understanding of their predictive capability. It has to be comprehensible, if the model is capable to accurately and reliably predict properties for fuels that it was trained on as well as for unknown fuels under consideration of uncertainties that results from limited training data. Furthermore, this predictive capability has to be put into the relation to the requirements of the use case to determine the adequacy of the prediction.

In this study we presented an approach that utilizes a Gaussian Process Regression (GPR), a probabilistic Machine Learning model, for the prediction of jet fuel density. The probabilistic feature to predict a distribution of values that correspond to the certainty of the prediction was used to estimate uncertainty due to limitations in the training data. This information was used to infer the predictive capability, determined by three metrics that measure the accuracy and precision of the prediction, as well as the validity and reliability of the prediction intervals of the distribution.

To investigate the influence of the presence of synthetic fuels in the training data, the assessment was executed for two models: one that was exclusively trained an validated on 54 conventional fuels from the CRC world fuel survey of 2006 and one that was trained an validated with additional data from 12 synthetic fuels. The comparison allowed the investigation of the influence of the presence of synthetic fuels in the

training data on the predictive capability of the model. The predictive capability was assessed on the basis of the results from the cross-validation and tested on three unseen hold-out fuels.

The results show, that the predicted distribution of the GPR can be used to correctly estimate the uncertainty of the prediction. The determined predictive capability proved to be reliable and correctly rate the adequacy of the predictions of the hold-out fuels. The GPR showed the predictive capability to predict the density of jet fuels for the safety relevant use case of fit-for-purpose prediction, both for conventional and alternative jet fuels. However, it was observed, that the predictive capability strongly depended on the presence of similar fuels in the training data. Predictions of synthetic fuels only complied with the adequacy requirements if similar synthetic fuels were part of the training data. If this was not the case, the predicted precision exceeded the threshold which indicated an unacceptable uncertainty.

Generally, this study showed the great potential of probabilistic Machine Learning algorithms and a systematic assessment of their predictive capability for an use case with high safety requirements like the fit-for-purpose property prediction of jet fuels. The possibility of estimating the uncertainty of an individual prediction can be greatly beneficial for the risk informed decision making of a stakeholder. The possibility of quantifying the uncertainty and relating it to a use case specific reference to estimate the predictive capability can be the basis of a decision to accept or reject a prediction. This can significantly reduce the need for measurements and indicate the need for additional ones. A systematic investigation of the correlation of accuracy and precision as well as an estimation of the predictive capability limits on the basis of the cross-validation results are the next step for this modelling approach.

7 ACKNOWLEDGEMENTS

The research presented in this paper has been performed in the framework of the JETSCREEN project (JET fuel SCREENING and optimization) and has received funding from the European Union Horizon 2020 Programme under grant agreement n◦ 723525.

8 ABBREVIATIONS

CRC fuels, fuels form the CRC world fuel survey; CRC & Syn. fuels, fuels form the CRC world fuel survey and the DLR jet fuel database; GCxGC, Two-dimensional gas chromatography; GPR, Gaussian Process Regressor; PI, Prediction interval; PICP, Prediction Interval Coverage Probability; RMSE, Root Mean Squared Error; ML, Machine Learning; MRE, Mean Relative Error; NMPIW, Normalized Mean Prediction Interval Width; QSPR, Quantitative Structure-Property Relationships.

9 REFERENCES

[1] ASTM International, "ASTM D4054-19, Standard Practice for Evaluation of New Aviation Turbine Fuels and Fuel Additives," West Conshohocken, PA USA, DOI:10.1520/D4054-19, 2019.

[2] P. Vozka and G. Kilaz, "A review of aviation turbine fuel chemical composition-property relations," *Fuel,* p. doi.org/10.1016/j.fuel.2020.117391, 15 05 2020.

[3] A. R. Katritzky, M. Kuanar, S. Slavov, D. C. Hall, M. Karelson and I. Kahn, "Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction," *Chemical Reviews,* pp. 5714-5789 doi.org/10.1021/cr900238d, 23 08 2010.

[4] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. de Hemptinne, P. Ungerer, B. Rousseau and C. Adamo, "A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes," *Chemical Reviews,* pp. 13093-13164 doi.org/10.1021/acs.chemrev.5b00215, 1 December 2015.

[5] X. Shi, H. Li, Z. Song, X. Zhang and G. Liu, "Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector," *Fuel,* pp. 395-406 doi.org/10.1016/j.fuel.2017.03.073, 15 07 2017.

[6] G. Liu and L. W. e. al., "Artificial neural network approaches on composition-property relationships of jet fuels based on GC-MS," *Fuel,* pp. 2551-2559 doi.org/10.1016/j.fuel.2007.02.023, 11 2007.

[7] J. A. Cramer, R. E. Morris, B. Giordano and S. L. Rose-Pehrsson, "Partial Least-Squares Predictions of Nonpetroleum-Derived Fuel Content and Resultant Properties When Blended with Petroleum-Derived Fuels," *Energy & Fuels,* pp. 894-902, doi.org/10.1021/ef700403s, 23 January 2009.

[8] D. J. Cookson, C. P. Lloyd and B. E. Smith, "Investigation of the chemical basis of kerosene (jet fuel) specification properties," *Energy & Fuels,* pp. 438-447 DOI: 10.1021/ef00005a011, 01 09 1987.

[9] D. J. Cookson and B. E. Smith, "Calculation of jet and diesel fuel properties using carbon-

13 NMR spectroscopy," *Energy & Fuels,* pp. 152-156 doi.org/10.1021/ef00020a004, 02 04 1990.

[10] H. Yang, Z. Ring, Y. Briker, N. McLean, W. Friesen and C. Faribridge, "Neural network prediction of cetane number and density of diesel fuel from its chemical composition determined by LC and GC-MS," *Fuel,* pp. 65-74 doi.org/10.1016/S0016-2361(01)00121-1, 1 January 2002.

[11] Vozka et al., "Jet fuel density via GC × GC-FID," *Fuel,* pp. 1052-1060 doi.org/10.1016/j.fuel.2018.08.110, 30 08 2018.

[12] W. L. Oberkampf and C. J. Roy, Verification and Validation in Scientific Computing, Cambridge University Press, doi.org/10.1017/CBO9780511760396, 2010.

[13] N. Ståhl, G. Falkman, A. Karlsson and M. Gunnar, "Evaluation of Uncertainty Quantification in Deep Learning," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2020 doi.org/10.1007/978-3-030-50146-4_41.

[14] C. Rasmussen, Gaussian Processes in Machine Learning, Berlin: Springer, 2003.

[15] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation:Representing Model Uncertainty in Deep Learning," *Proceedings of the33rd International Conference on MachineLearning,* p. arXiv:1506.02142, 06 06 2016.

[16] W. F. de Carvalho Rocha and D. A. Sheen, "Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation," *Fuel,* pp. 413-422 doi.org/10.1016/j.fuel.2018.12.126, 01 05 2018.

[17] Coordinating Research Council, "CRC Report No. 647," 2006, 06.

[18] F. Pedregosa, G. Varoquaux and A. Gramfort, "scikit-learn," *Journal of Machine Learning Research,* pp. 2825-2830, 10 2011.

[19] . P. Jingyue, L. Datong and X. Peng, "Optimize the Coverage Probability of Prediction Interval for Anomaly Detection of Sensor-Based Monitoring Series," *Sensors,* p. 987 DOI: 10.3390/s18040967, 24 03 2018.

[20] T. Pearce, M. Zaki, . A. Brintrup and A. Neely, "High-Quality Prediction Intervals for Deep Learning:A Distribution-Free, Ensemble Approach," in *Proceedings of the35th*

International Conference on MachineLearning, Stockholm, Sweden, 2018.

[21] Creton, Alonso et al., "Prediction of Density and Viscosity of Biofuel Compounds Using Machine Learning Methods," *Energy and Fuels,* pp. 2412-2426 https://doi.org/10.1021/ef3001339, 20 03 2012.

[22] G. Cai, Z. Kiu, Z. Linzhou, Z. Suoqi and C. Xu, "Quantitative Structure–Property Relationship Model for Hydrocarbon Liquid Viscosity Prediction," *Energy & Fuels,* pp. 3290-3298 doi.org/10.1021/acs.energyfuels.7b04075, 07 02 2018.

[23] ASTM International, ASTM D1655-18a, Standard Specification for Aviation Turbine Fuels, West Conshohocken, PA USA: ASTM International, DOI 10.1520/D1655-18A, 2018.

[24] Jetscreen, "https://www.jetscreen-h2020.eu/," ARTTIC , [Online]. Available: https://www.jetscreen-h2020.eu/. [Accessed 03 08 2019].

[25] M. H. Hammond, R. E. Morris, J. A. Cramer, T. N. Loegel, K. J. Johnson and K. M. Myers, "Navy Fuel Composition and Screening Tool (FCAST) v.2.5," *Conference Proceedings,* p. DOI:10.21236/ada608713, 2014.

[26] A. Daw, T. R. Quinn, C. C. Cayelan, J. S. Read and A. P. Appling, "Physics-Guided Architecture (PGA) of Neural Networks for Quantifying Uncertainty in Lake Temperature Modeling," *ArXiv,* p. arXiv:1911.02682, 06 11 2019.

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Conference paper at the 3rd International Conference for Learning Representations*, San Diego arXiv:1412.6980, 2015.

[28] A. Geron, Hands-On Machine Learning, O'Reilly UK Ltd, 2017.

[29] T. Pearce and M. Zaki, "High-Quality Prediction Intervals for Deep Learning:A Distribution-Free, Ensembled Approach," in *Proceedings of the35th International Conference on MachineLearning*, Stockholm, Sweden, 2018.

[30] scikit-optimize developers, "scikit-optimize," scikit-optimize developers, [Online]. Available: https://scikit-optimize.github.io/. [Accessed 2 06 2019].

[31] J. Cramer, M. Hammond, K. Myers, T. Loegel and R. Morris, "Novel data abstraction strategy utilizing gas chromatography-mass spectrometry data for fuel property modeling,"

Energy Fuels, pp. 1781-91 doi.org/10.1021/ef4021872, 03 02 2014.

[32] ASTM International, ASTM D7566-18, Standard Specification for Aviation Turbine Fuel Containing Synthesized Hydrocarbons, West Conshohocken, PA USA: ASTM International, DOI: 10.1520/D7566-18, 2018.

1

2