Fuel 329 (2022) 125415

Contents lists available at ScienceDirect

Fuel

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

Full length article

Towards predicting liquid fuel physicochemical properties using molecular dynamics guided machine learning models

Rodolfo S.M. Freitas [a], Ágatha P.F. Lima [a], Cheng Chen [b], Fernando A. Rochinha [a], Daniel Mira [c], Xi Jiang [b,*]

[a] COPPE, Federal University of Rio de Janeiro, Rio de Janeiro 21941-598, Brazil
[b] School of Engineering & Materials Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK
[c] Barcelona Supercomputing Center (BSC), Barcelona, Spain

ARTICLE INFO

Keywords:
Fuel properties
Molecular dynamics
Deep learning
Machine learning models

ABSTRACT

Accurate determination of fuel properties of complex mixtures over a wide range of pressure and temperature conditions is essential to utilizing alternative fuels. The present work aims to construct cheap-to-compute machine learning (ML) models to act as closure equations for predicting the physical properties of alternative fuels. Those models can be trained using the database from MD simulations and/or experimental measurements in a data-fusion-fidelity approach. Here, Gaussian Process (GP) and probabilistic generative models are adopted. GP is a popular non-parametric Bayesian approach to build surrogate models mainly due to its capacity to handle the aleatory and epistemic uncertainties. Generative models have shown the ability of deep neural networks employed with the same intent. In this work, ML analysis is focused on two particular properties, the fuel density and diffusion, but it can also be extended to other physicochemical properties. This study explores the versatility of the ML models to handle multi-fidelity data. The results show that ML models can predict accurately the fuel properties of a wide range of pressure and temperature conditions.

1. Introduction

Fossil fuels have been playing a major role in energy supply and liquid fossil fuels have dominated the energy use in transport, which will continue to be so for many decades to come, especially for sectors that are difficult to decarbonize [1,2]. With the pressing needs of decarbonization and sustainable energy utilization, renewable fuels and biofuels are becoming increasingly important [3,4]. For instance, synthetic fuels like Oxymethylene Dimethyl Ethers (OMEs) have shown high potential for low-carbon transport applications due to their capacity to avoid soot formation [5]. However, the physicochemical properties of these fuels must be known for their rapid integration into current infrastructures for storage, transport and direct injection in combustion engines. This represents a significant challenge, due to the fact that practical fuels are often composed by complex mixtures and vary widely in their chemical compositions depending on the production source and process [3]. For example, petroleum diesel is a complex mixture involving molecules with carbon chains that typically contain between 9 and 25 carbon atoms per molecule. To simplify the complex chemical compositions of these fuels, surrogate models have been used to represent the chemical composition and combustion characteristics in practical applications [6,7]. In addition, modern combustion engines have to operate at high pressure conditions in order to improve the energy conversion efficiency. Fuel properties at extreme conditions such as high pressure and high temperature conditions, are very difficult to measure and predict [5], leading to an additional challenge.

Accurate determination of fuel properties of complex mixtures over a wide range of pressure and temperature conditions is essential to adapt the system operation to alternative fuels. In recent years, molecular dynamics (MD) simulations have been used to predict the physicochemical properties of practical fuels including transport properties at supercritical conditions [8]. By using equilibrium molecular dynamics (EMD) and nonequilibrium molecular dynamics (NEMD), Yang et al. [9,10] predicted the viscosity and thermal conductivity of alkanes (n-decane, n-undecane and n-dodecane). Kondratyuk et al. [11–13] performed a serial of MD simulation to study the viscosity of hydrocarbons (1-methylnaphthalene, methylcyclohexane and 2,2,4-trimethylhexane) in high pressure conditions up to 1000 MPa. Caleman et al. [14] tested the capacity of existing force fields on prediction of properties (density, enthalpy of vaporization, surface tension and heat capacity etc.) of

[*] Corresponding author.
E-mail address: xi.jiang@qmul.ac.uk (X. Jiang).

https://doi.org/10.1016/j.fuel.2022.125415
Received 11 June 2022; Received in revised form 13 July 2022; Accepted 24 July 2022
Available online 5 August 2022

Nomenclature

Abbreviations

ANN	Artificial neural network
CFD	Computational Dluid Dynamics
CN	Cetane number
EMD	Equilibrium Molecular Dynamics
EoS	Equation of State
FAME	Fatty acid methyl ester
GANs	Generative Adversarial Networks
GP	Gaussian Process
MD	Molecular dynamics
ML	Machine learning
MLPNNs	Multilayer Perceptron Neural Networks
NARGP	Nonlinear autoregressive multifidelity Gaussian Process
NEMD	Nonequilibrium Molecular Dynamics
NIST	National Institute of Standards and Technology
OMEs	Oxymethylene Dimethyl Ethers
TraPPE	Transferable Potential for Phase Equilibria
VAE	Variational auto-encoders

Greek letters

β	Residual penalty parameter
$\boldsymbol{\theta}$	A vector of hyper-parameters
γ	A generic property
λ	Entropy regularization parameter
μ	Expected value
ϕ	A vector of parameters
ρ	Density
σ	Standard deviation
ξ	A potential noisy

Latin letters

$\mathbf{x}, \mathbf{y}$	Input and output vectors
cv	Coefficient of variation
C	Number of atoms of carbon
D	Diffusion coefficient
f	Gaussian function
g	Mapping function
K	Covariance matrix
k	A kernel function
l	Correlation length
n	Dimension of the input and output
N_s	Number of samples
P	Pressure
p	Probability distribution
P_c	Critical pressure
T	Temperature
t	Time
T_c	Critical temperature
z	Latent variable

organic liquids. Although MD simulations provide molecular details that can be potentially used to accurately predict fuel properties, they are generally expensive in terms of computational costs (CPU time and memory). In addition, MD predictions also need to be validated against experimental measurements, which can be even more costly especially at extreme conditions. Accordingly, it is not feasible to establish complete and detailed fuel property databases consisting of a wide range of pressure and temperature conditions using either MD simulations or experiments.

Machine learning has great potentials to discover the relation between inputs and outputs in a thermodynamic system directly from the data of complex systems [15] and for predicting the properties of materials based on their composition [16]. ML can be a powerful tool to predict fuel properties from chemical compositions of the fuel mixture and/or chemical structures of the fuel molecules. Several works have been devoted to designing ML models capable of predicting complex fuels properties from experimental data. In this regard, ML models obtained accurate predictions of cetane number (CN) compared to experimental data [17–19]. A satisfactory ML approach for modeling the CN of biodiesel based on four operating conditions given by iodine volume (IV), carbon number, double bounds, and saponification value was proposed [20]. Recently, an artificial neural network (ANN) was applied to predict and identify the underlying links between the fuel properties and the octane number (ON) [21]. Moreover, ML models were tuned with evolutionary algorithms to predict the CN of biodiesel as a function of its fatty acid methyl ester (FAME) profile [22,23]. The predictability, i.e. the ability to predict, of the ML approaches also can be improved by using different optimization algorithms for the training and/or hyperparameter search such as teaching–learning based optimization (TLBO), backpropagation, Quasi-Newton and particle swarm optimization (PSO) [24–26]. Also, ML models have been used for modeling the kinematic viscosity of diesel-derived fuels as a function of their FAMEs profiles [27–29]. In the last years, Multilayer Perceptron Neural Networks (MLPNNs) have been successfully built to estimate the physicochemical characteristics of biodiesel [30–33] combining different parameters of model inputs. Furthermore, ML models based on state variables such as temperature and pressure showed high potential to obtain physicochemical properties of biodiesel/diesel fuels more accurately [34–36]. In particular, ML models have been developed to predict thermodynamic properties such as critical pressure and temperature, vapor pressures, and densities of pure fluids [37]. Moreover, approaches combining MD simulations and ML have been applied to modeling the diffusion of pure liquids [38,39]. Following the same context, a ML approach based on support vector regression (SVR) was proposed by [40] for predicting the PVT properties of pure fluids (H_2O, CO_2, and H_2) and their mixtures, where the training database is provided by the National Institute of Standards and Technology (NIST) and MD simulations. Also, an ML approach was proposed to assess the macroscopic Engine Combustion Network (ECN) Spray-A characteristics and predictions of fluid properties for the thermodynamic states found in such conditions [41]. Yet, from our knowledge, little work has been dedicated towards exploring the thermodynamic properties of practical fuels combining MD simulations and ML models. ML can be a powerful tool to predict fundamental fuel properties directly from the chemical compositions of the fuel mixture by using databases from MD simulations or available experimental measurements.

The aim of the study was to demonstrate and validate a ML-MD methodology to predict fundamental properties of liquid fuels. In this approach, the ML models are built from data provided by MD simulations, while a combination of MD and NIST data is used for model assessment and validation. This study is the first attempt of using ML models with Gaussian process regression [42] and probabilistic conditional generative learning [43,44] for the property predictions of single-compounds. The ML analysis is focused on fuel density in this study as one of fundamental properties of liquid fuels, though it can easily be extended to other physicochemical properties of relevance for practical applications like diffusion coefficient, viscosity, conductivity or surface tension.

The rest of the paper is organized as follows. Section 2 presents the ML models and the molecular dynamics simulation methodology. Section 3 describes the ML results for typical fuel surrogates of diesel. Finally, Section 4 concludes the study with recommendation for further investigations.

2. Methodology: Building machine learning models to describe physicochemical properties

In order to reduce energy consumption and pollutant formation, supercritical combustion has been increasingly explored in the context of high pressure internal combustion engines and rocket engines [45]. Specifically, in supercritical conditions, the devices operate with pressures and temperatures higher than the critical values, which implies that physicochemical properties of fluids are quite different from those at liquid conditions [46]. In such scenarios, the design of devices become more complex, specially due to limitations of replicating flow and combustion in controlled laboratory environments. In order to cope with these challenges, computational models can provide adequate tools for obtaining more accurate predictions of state variables and increase cycle performance in transcritical conditions.

From a computational fluid dynamics (CFD) perspective, combustion models are built upon the combination of solid and reliable physico/chemical principles with closure models, typically describing physicochemical properties of the fuels and their mixtures using approaches that normally entail uncertainties. The use of numerical simulations for practical applications encompass a wide range of conditions, resulting in different fundamental problems depending on the nozzle geometry, engine architecture or thermodynamic conditions. A good example is the database from the Engine Combustion Network [47] for which different sprays for diesel- and gasoline-like conditions are investigated. For instance, pressure can go from sub-atmospheric to 2000 bar, and temperatures from cold to highly preheated conditions. In that context, having accurate values for macroscopic fuel characteristics and properties over such wide variety of spatial and time scales is one of the main challenges for physically-driven methods. That is particularly more dramatic for modern compounds depicting complex chemical compositions, and simplified surrogate fuels [48] are employed to estimate the properties of the original compounds. That allows the systematic use of controlled experiments and, also, Molecular Dynamics simulations [49,50]. Indeed, here our focus lies on using ML models to leverage such type of simulations when obtaining liquid fuel physicochemical properties. Those properties are generally expressed as functions of local thermodynamic conditions like pressure and temperature, which motivate to refer to closure models such as the Equations of State (EoS). In general, the EoS is embedded in complex CFD simulations resulting in divergence or numerical oscillations when used with traditional methods based on tabular and interpolation schemes [51]. It is worth to remark that we are seeking for models capable of describing physicochemical properties over a wide range of flow conditions and we expected to observe abrupt changes around critical conditions.

We built two different ML models, namely Gaussian Processes (GP) [42] and a probabilistic conditional generative approach [44]. We train both in a supervised learning fashion using data produced with expensive MD simulations. Therefore, we rely on their ability to learn from a small amount of data and their capacity of extrapolation. Moreover, we also want to take into consideration the unavoidable uncertainties arising from limited information (epistemic) and from noisy data (aleatoric).

GPs have become popular due to its success on being a proxy for physics-based high-fidelity models in different applications [52–57]. Another well proved ML approach are the so called generative models that explore existing low-dimensional structures capable of explaining high-dimensional data introducing probabilistic latent variables.

In the remainder of this chapter, we present a brief description of both ML models for a generic property $\gamma(P,T)$ function of pressure and temperature, along with the corresponding training algorithms. For the training of the models, we assume the availability of, potentially expensive, dataset comprising input/output pairs $\{(P,T)_i, \boldsymbol{\gamma}_i \quad i = 1,\dots,n\}$ generated by an implicit mapping g characterizing the macroscopic thermodynamic relation between the property and the state variables:

$$\gamma = g(P,T;\xi). \tag{1}$$

The role of g here is played by upscaling MD simulations or, to a less extent, by experimental available data. The vector ξ denotes potential noisy and is often considered a random. In order to keep a compact notation, we refer to the above dataset as $\mathcal{D} = (\mathbf{x}, \mathbf{y})$, with $\mathbf{x} \in \mathbb{R}^{2n}$ and $\mathbf{y} \in \mathbb{R}^{n}$ vectors containing inputs and outputs. We intentionally do not use the word surrogate to designate any of the two ML models to avoid misleadings. In the combustion technical literature, it is employed to refer to compounds with simpler compositions to replace complex fuels in experimental or numerical analysis.

2.1. Gaussian process regression

A GP is an infinite collection of random variables, in which any finite number of such variables depict a joint Gaussian distribution [42]. In line with Bayesian estimation, to approximate g we assign a GP zero mean prior $f(\mathbf{x})$, i.e., $f \sim GP(f|\mathbf{0}, k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$, where k is a kernel parametrized by a vector of hyper-parameters $\boldsymbol{\theta}$ to be learned from $\mathcal{D}$ and engenders a symmetric positive-definite $n \times n$ covariance matrix $K_{ij} = k(x_i, x_j; \boldsymbol{\theta})$. Instead of choosing the squared exponential form of the kernel as usual [42], here, we test some forms of covariance matrix belonging to the Matern family. More specifically, we employ the Mayern 3/2 covariance matrix given as

$$k(\mathbf{r}) = \sigma^2 \left(1 + \sqrt{6}\frac{|r|}{l}\right) \exp\left(-\sqrt{6}\frac{|r|}{l}\right) \tag{2}$$

with $\mathbf{r} = \mathbf{x} - \mathbf{x}'$ denoting the distance between different inputs. The hyper-parameters are the standard deviation σ, and the correlation lengths $\mathbf{l} = \{l_1, l_2, \dots, l_{n_k}\}$, and n_k denotes the dimension of input $\mathbf{r}$. Hence, the hyper-parameters vector reduces to $\boldsymbol{\theta} = \{\mathbf{l}, \boldsymbol{\sigma}\}$.

We do not follow a fully Bayesian approach, and obtain the vector of hyper-parameters $\boldsymbol{\theta}$ by maximizing the marginal log-likelihood of the model, i.e.

$$\log p(\gamma|\mathbf{x}, \theta) = -\frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\gamma^T\mathbf{K}^{-1}\gamma - \frac{n}{2}\log 2\pi. \tag{3}$$

using a conjugate gradient descend method.

The final goal of the regression is obtaining a predictive model for γ, which means to compute its value for an untested state $\mathbf{x}_*$ [53]

$$\mu_*(\mathbf{x}_*) = k_{*n}\mathbf{K}^{-1}\mathbf{y} \tag{4}$$

and

$$\sigma_*^2(\mathbf{x}_*) = k_{**} - k_{*n}\mathbf{K}^{-1}k_{*n}^T \tag{5}$$

where $k_{*n} = [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_n)]$ and $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$. The predictions are computed using the posterior mean μ_*, and the uncertainty associated with that predictions is quantified through the posterior variance σ_*^2. It is worth to mention that in absence of noisy in the training data, the later represents epistemic uncertainty due to lack of data.

2.2. Probabilistic conditional generative model

Now, we explore a probabilistic conditional generative approach [43,44], that integrates variational auto-encoders (VAE) [58] and generative adversarial networks (GANs) [59]. Moreover, it employs a probabilistic perspective that enables to take into consideration noisy and limited data from the beginning. It is also capable of dealing with high-dimensionality of inputs and outputs, what is not explored here due to the specific aspects of our needs.

The final goal is to build probabilistic neural networks that follow a conditional probability density function $p(\gamma|(P,T), \mathcal{D})$ learnt from the data. So, the surrogate model can deploy accurate values for the property γ by estimating the expectation $\mathbb{E}(\gamma|(P,T), \mathcal{D})$, and also, to quantify the uncertainty associated with that prediction in CFD calculations.

The main ingredient for this approach is the introduction of a vector of latent random variables aiming at seeking for a hidden low dimensional structure for explaining the data structure. In a formal abstract perspective, such latent variables allow us to express the conditional probability associate to the data $\mathcal{D}$, not included in the expression to keep the notion clear, $p(\gamma|P,T)$, as an infinite mixture model through

$$p(\gamma|P,T) = \int p(\gamma,\mathbf{z}|P,T)\ d\mathbf{z} = \int p(\gamma|P,T,\mathbf{z})\ p(\mathbf{z}|P,T)\ d\mathbf{z} \tag{6}$$

where $p(\mathbf{z}|p,T)$ is a prior distribution on the latent variables. The above hierarchical mathematical ansatz, despite being very elegant and rigorous, has to be approximated [44], where a regularized adversarial inference framework is proposed and detailed. The final result is a generator model $\gamma = f_\phi(p,T,\mathbf{z})$ parametrized by vector ϕ, like trained deep neural networks. In conjunction with $p(\mathbf{z})$, the statistics of γ can be characterized. More specifically, we can compute its low order statistics via Monte Carlo sampling. It is important to remark that the predictions with the identified probabilistic generator, that, in present context, plays the role of a proxy for obtaining macroscopic thermodynamic properties of mixtures for pressures and temperatures not contained in $\mathcal{D}$, is negligible when compared to MD simulations. The mean and variance of the predictive distribution at a new point (p^*,T^*) are computed as

$$\mu_\gamma(P^*,T^*) = \mathbb{E}[\gamma|P^*,T^*,\mathbf{z}] \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\left[f_\phi(P^*,T^*,\mathbf{z}_i)\right] \tag{7}$$

$$\sigma_\gamma^2(P^*,T^*) = \mathbb{V}\mathrm{ar}[\gamma|P^*,T^*,\mathbf{z}] \approx \frac{1}{N_s}\sum_{i=1}^{N_s}\left[f_\phi(P^*,T^*,\mathbf{z}_i) - \mu_{\mathrm{y}}(P^*,T^*)\right]^2, \tag{8}$$

where $\mathbf{z}_i \sim p(\mathbf{z})$, $i = 1,\dots,N_s$, and N_s corresponds to the total number of samples.

At this point, it is important to clarify that the predictive uncertainty encoded in $\mathbf{z}$ is due to noise in the Molecular Dynamics computations originated by numerical approximations and to the potential small amount of data employed in the training process. Therefore, it encapsulates aleatoric and epistemic uncertainties.

Later, we explore the versatility of the probabilistic ML model employing the fusion of data produced by MD with experimental data obtained for supercritical behavior of the mixture.

2.3. Physicochemical properties prediction in EMD simulation

In this study, all MD simulations are performed in Gromacs package [60] with Transferable Potentials for Phase Equilibria (TraPPE) force field [61]. United-atom molecular description is used in order to reduce the computational cost. Before simulation, 1000 molecules are distributed in a box with relatively large edge length of 14 nm to avoid atom's overlap. After energy minimization, a 2 ns simulation is performed with time setup of 1fs in isobaric–isothermal NPT (fix the number of atoms, pressure and temperature of the system) ensemble by using Parrinello–Rahman method [62] to maintain the pressure. Then 1 ns NVT (fix the number of atoms, volume and temperature of the system) simulation is followed for production run. The temperature is controlled by velocity rescale. The fixed bond length in TraPPE force field is achieved by using LINCS algorithm [63]. The density and diffusion is calculated in NVT simulation.

The diffusion coefficient (D) can be obtained from the linear fittings of mean square displacement (MSD) of molecules:

$$MSD(t) = \langle|\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2\rangle \tag{9}$$

$$D(t) = \frac{1}{6}\frac{d}{dt}\langle|\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2\rangle \tag{10}$$

where $\mathbf{r}_i(t)$ is the position of the ith particle at time t, angle bracket indicates the ensemble average over all the particles in the system.

Table 1
Gaussian process training accuracy.

Train data	L_{2-MRE}	R^2-score
10%	6.2805×10^{-2}	0.8538
50%	4.7438×10^{-2}	0.9976
100%	2.7272×10^{-2}	0.9991

Table 2
Generative model training accuracy.

Train data	L_{2-MRE}	R^2-score
10%	4.9316×10^{-2}	0.9359
50%	2.8989×10^{-3}	0.9983
100%	2.1409×10^{-3}	0.9990

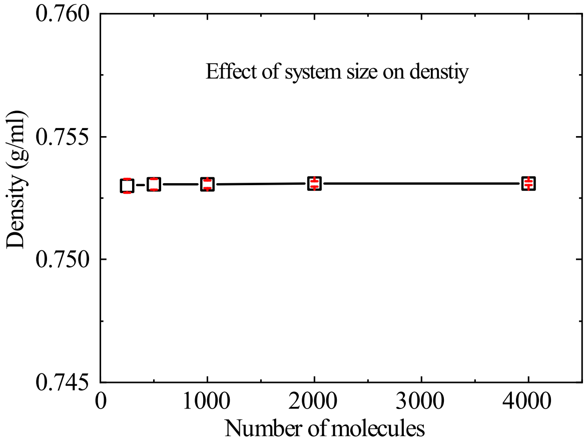


Fig. 1. Effect of the system size on density prediction.

The number of fuel molecules and simulation time in our simulation is setup according to previous studies. For example, Yang et al. [64] used 250 molecules with 2 ns simulation time in transport property prediction of n-alkanes, and Kondratyuk et al. [65] used 1000 molecules in modeling branched alkanes running in EMD simulation of 1 ns. Fig. 1 depicts the effect of the system size on the n-dodecane density prediction. As we can see 1000 molecules are sufficient to achieve convergence of the density prediction at an affordable computational cost.

3. Results and discussion

Here, we demonstrate the performance of the proposed methodology. Despite alternative fuels can be very complex mixtures consisting of hundreds of compounds, we consider single-component alkanes C_nH_{2n+2}, so reliable data for model assessment and validation can be used. In general, realistic fuels are usually described by surrogate models [8] because of availability of validated chemical mechanisms and experimental measurements. The data to train our ML models consist of properties of a family of alkanes, ranging from normal to supercritical conditions. More specifically, we construct ML models to characterize density dependency on some operational conditions in which data is not available. As mentioned before, in order to take into consideration unavoidable uncertainties, we approximate the conditional probability $p(\gamma|\mathbf{x},\boldsymbol{\theta})$, with $\mathbf{x}$ being the input vector with components pressure p, temperature T and chemical composition. Moreover, it is worth mentioning here that for simplicity we consider as the input that characterizes the chemical compositions the number of atoms of carbon C in the molecule of the pure compounds, a categorical variable. However, parameters from the EMD used to characterize the physicochemical properties of the fuel molecule can be used. Also, for the GP learning model, the hyper-parameters vector reduces to $\boldsymbol{\theta} = \{\mathbf{l},\boldsymbol{\sigma}\}$, and for the

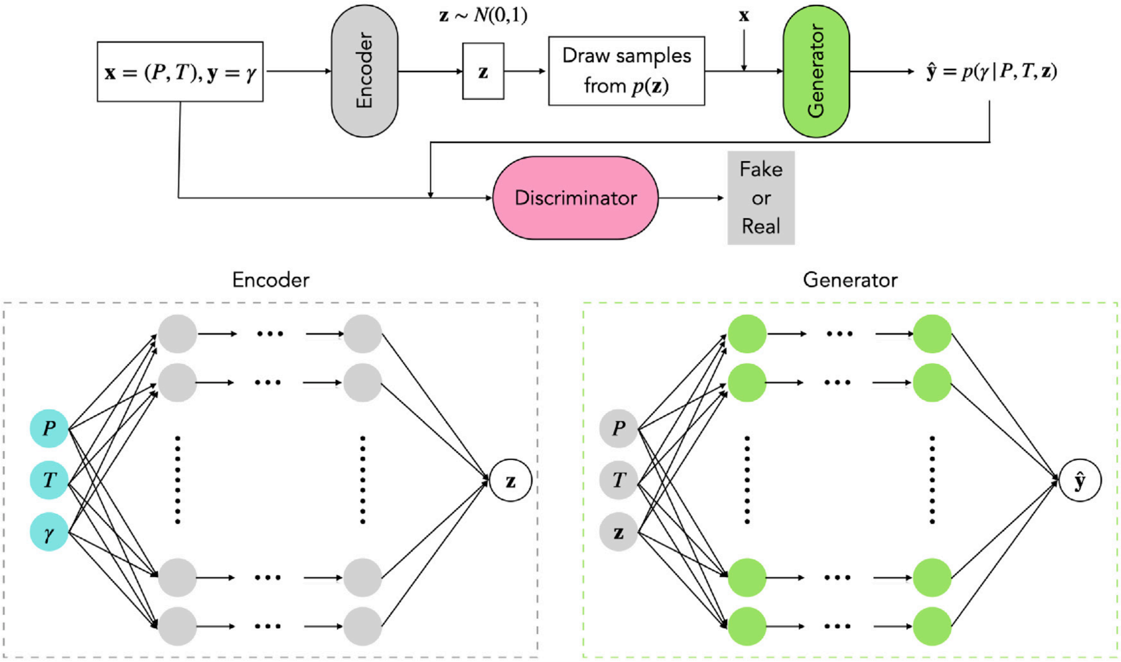
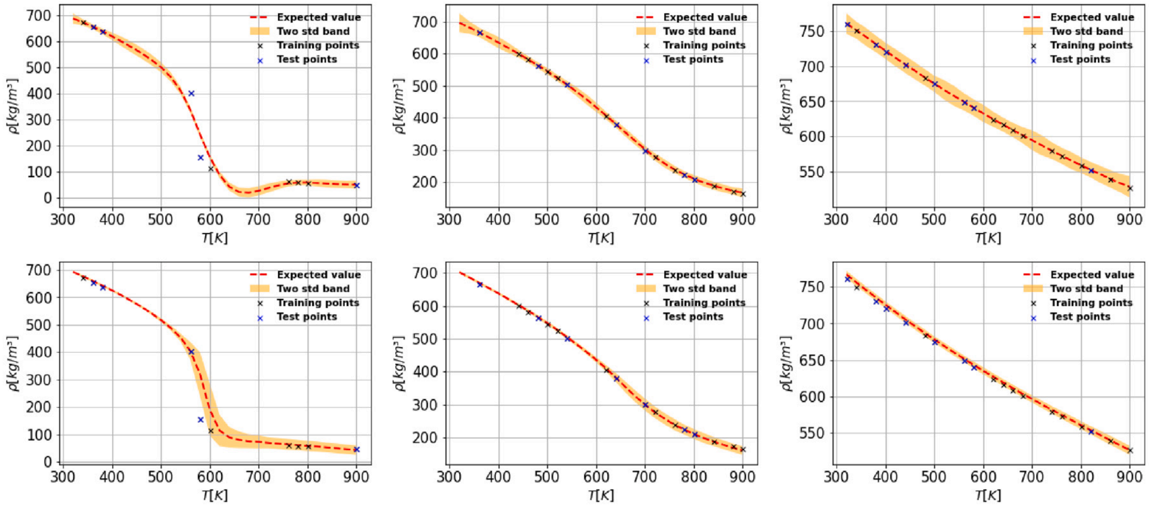

$\mathbf{x} = (P, T), \mathbf{y} = \gamma$
Encoder
$\mathbf{z} \sim N(0,1)$
z
Draw samples from $p(\mathbf{z})$
x
Generator
$\hat{\mathbf{y}} = p(\gamma | P, T, \mathbf{z})$
Discriminator
Fake or Real
Encoder
P
T
γ
z
Generator
P
T
z
ŷ

Fig. 2. Schematic view of the conditional generative model.



(a) 3 MPa

(b) 10 MPa

(c) 100 MPa

Fig. 3. n-Octane predictions with the GP (top) and probabilistic conditional generative models (bottom) at the pressures 3, 10, and 100 MPa.

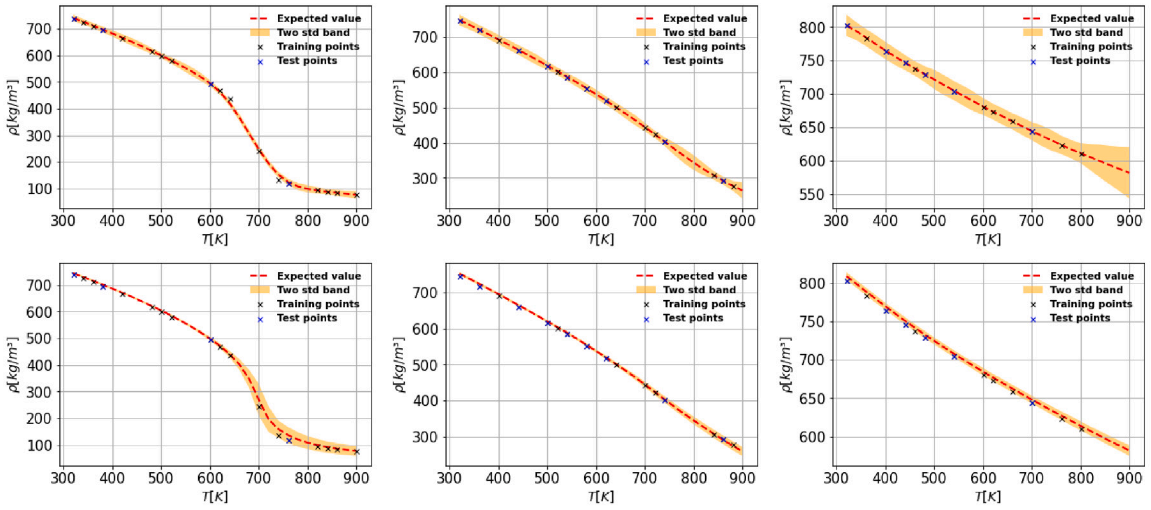
generative model θ represents the vector of parameters of the deep neural networks ϕ. The latent variable z is embedded in the input vector $\mathbf{x}$. We employ a one-dimensional latent space with a standard normal prior, $p(z) \sim \mathcal{N}(0,1)$.

The pure compounds considered are n-octane, n-nonane, n-decane, n-dodecane, and n-hexadecane, operating from high-pressure nozzle to supercritical chamber environment conditions. The dataset used to build the ML models consists of 1200 density values. Specifically, the there are 240 values of the density for each compound, computed at a regular temperature grid within $T \in [320, 900]$ K, varying by 20 K, and at the specific pressures values: $P = \{3, 4, 6, 8, 10, 20, 100, 150\}$ MPa. It is worth remarking that in this dataset we included density values for supercritical regions, more specifically values above the critical temperature (T_c) of the compounds, being the critical values for n-octane (T_c = 569.32 K), n-nonane (T_c = 594.55 K), n-decane (T_c = 617.7 K), n-dodecane (T_c = 658.1 K), and n-hexadecane (T_c = 722 K), which replicate engine-like conditions.

In the learning process, 80% of the data points are selected randomly to training the ML models. The remaining 20% are used to validating them. Moreover, the training data set is organized in three subsets with 10%, 50%, and 100% of data available to train the models. The aim here is to evaluate the convergence and impacts of constructing the ML models in a small data regime. Accuracy is measured using the distance between the expected values predicted with the ML models and the predictions computed with the MD simulations. We check this accuracy computing the L_2 mean relative error (L_{2-MRE})

$$L_{2-MRE} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\rho_i - \hat{\rho}_i}{\rho_i} \right)^2 \tag{11}$$

where ρ_i is the density computed with MD simulations, $\hat{\rho}_i$ is the expected ML output and N is the number of test samples. Also, we

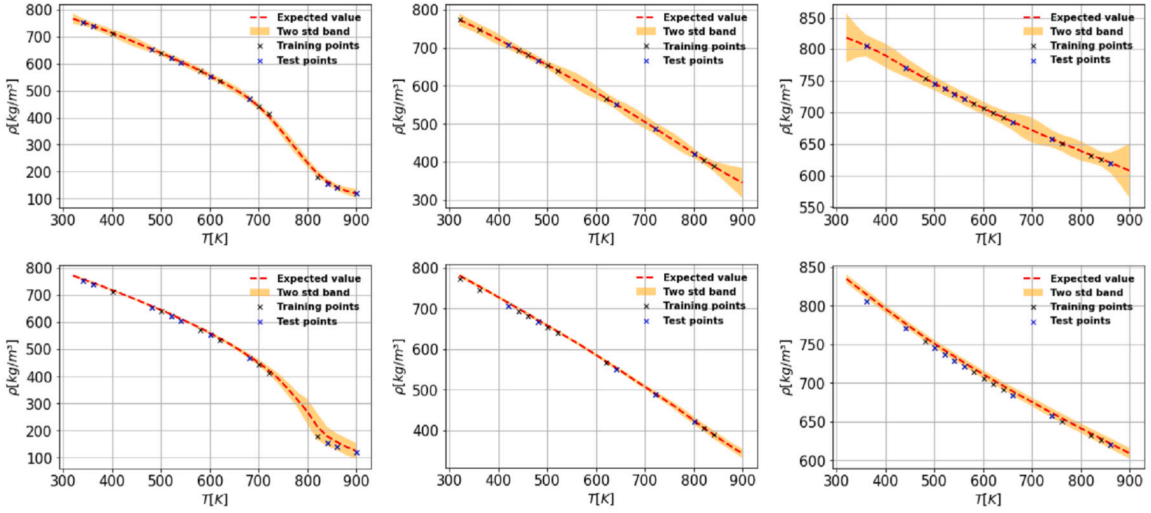


(a) 3 MPa

(b) 10 MPa

(c) 100 MPa

Fig. 4. n-Dodecane predictions with the GP machine learning model (top) and conditional generative machine learning model (bottom) at the pressures 3, 10, and 100 MPa.



(a) 3 MPa

(b) 10 MPa

(c) 100 MPa

Fig. 5. n-Hexadecane predictions with the GP machine learning model (top) and conditional generative machine learning model (bottom) at the pressures 3, 10, and 100 MPa.

compute the coefficient of determination (R^2-score) metric [66]

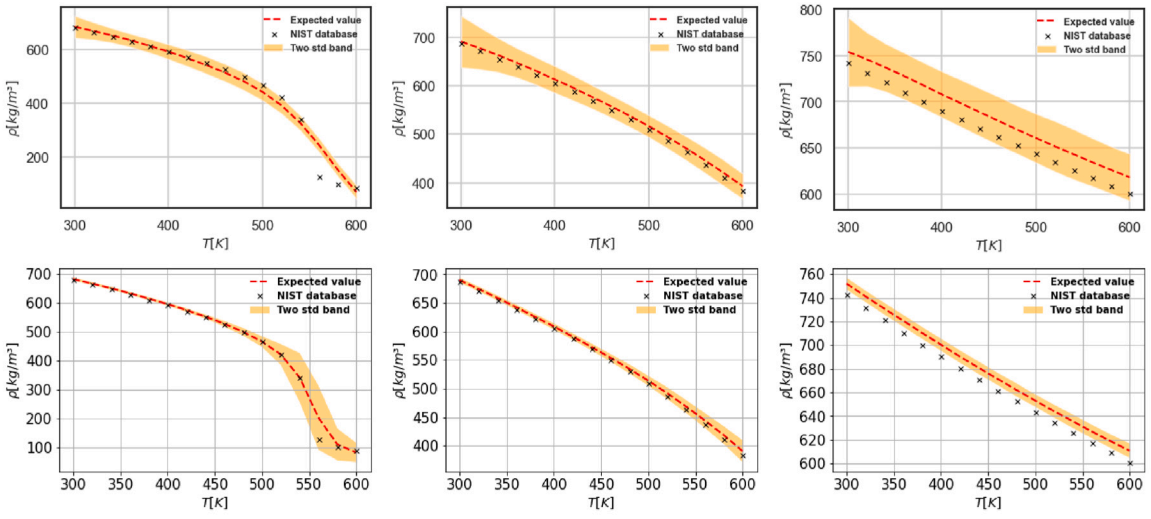
$$R^2 = 1 - \frac{\sum_{i=1}^{N} \|\rho_i - \hat{\rho}_i\|_2^2}{\sum_{i=1}^{N} \|\rho_i - \overline{\rho}\|_2^2} \tag{12}$$

where $\overline{\rho} = \frac{1}{N}\sum_{i=1}^{N} \rho_i$ is the mean density of test samples. The R^2-score metric represents the normalized error, allowing the comparison between ML models trained by different data sets, with values close to 1 corresponding to the ML models best accuracy, while L_{2-MRE} is a common metric used to check the accuracy of ML models during the optimization process.

We obtain the GP regression model of Eq. (1) via maximizing the marginal log-likelihood of Eq. (3) using the Mattern 3/2 kernel function, as that shown in Eq. (2). Also, we have used the gradient descend optimizer L-BFGS [67] using randomized restarts to ensure convergence to a global optimum. The GP learning model was implemented in GPy: Gaussian Process (GP) framework written in python [68].

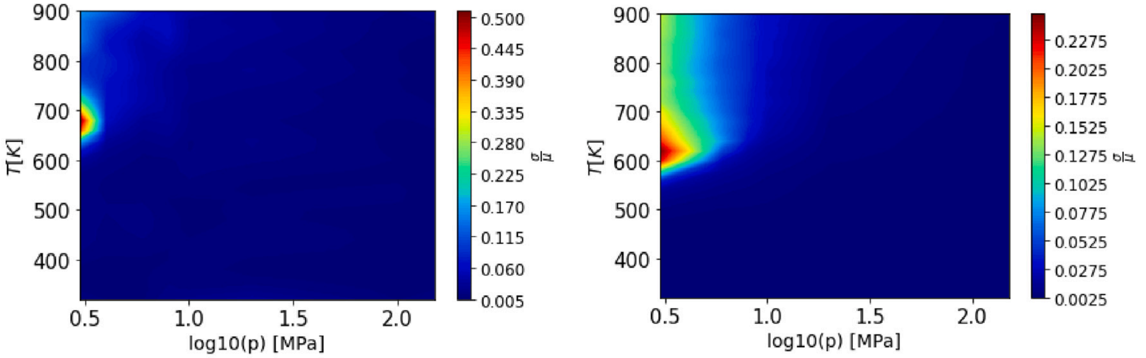
On the other hand, to construct the generative learning model, we departed from the architecture proposed and validated by Yang and Perdikaris [44]. More specifically, the conditional generative model is constructed using fully connected feed-forward architectures for the encoder and generator networks with 4 hidden layers and 100 neurons per layer, while the discriminator architecture has 2 hidden layers with 100 neurons per layer. A schematic view of the conditional generative model is depicted in Fig. 2. The neural networks are constructed by combining try-and-error and Hyperopt algorithm [69] to search for the hyperparameters that give the lowest L_{2-MRE}. All activation uses a hyperbolic tangent non-linearity. The models are trained for 50,000 stochastic gradient descent steps using the Adam optimizer [70] with a learning rate of 10^{-4}, while fixing a two-to-one ratio for the discriminator versus generator updates. Furthermore, we have also fixed the entropy regularization and the residual penalty parameters to $\lambda = 1.5$ and $\beta = 0.5$, respectively. The proposed model was implemented in TensorFlow v2.1.0 [71], and computations were performed in single precision arithmetic on a single NVIDIA GeForce RTX 2060 GPU card.

We also explore some alternatives versions of the above described ML models by proposing fusion with experimental data and the use of multi-fidelity formulations.



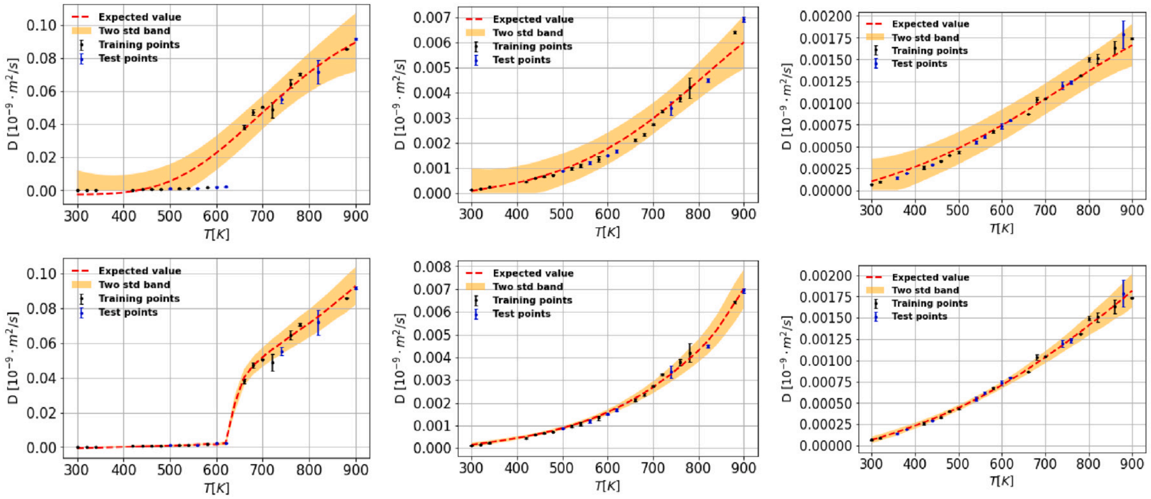
(a) 3 MPa (b) 10 MPa (c) 100 MPa

Fig. 6. n-Heptane predictions with the GP machine learning model (top) and conditional generative machine learning model (bottom) at the pressures 3, 10, and 100 MPa.



(a) GP model (b) Conditional generative model

Fig. 7. n-Octane density variability for a range of temperatures and pressures.



(a) 1 MPa (b) 10 MPa (c) 100 MPa

Fig. 8. n-Dodecane predictions of the diffusion coefficient with the GP machine learning model (top) and conditional generative machine learning model (bottom) at the pressures 1, 10, and 100 MPa.

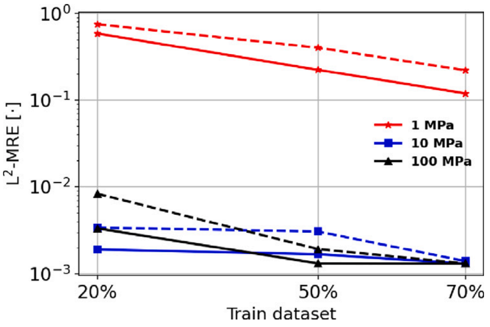


Fig. 9. Comparison of the L_2 mean relative error in different data regimes for training. Gaussian process (dashed-line) and generative model (solid-line).

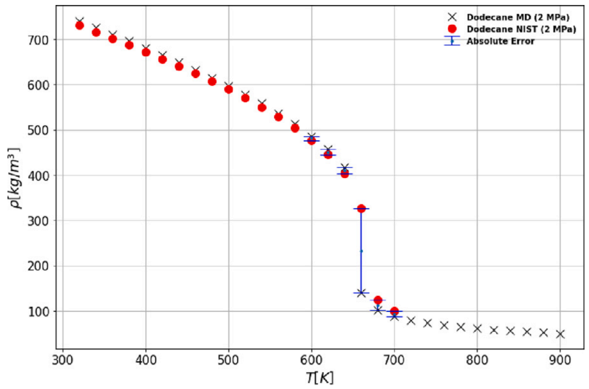


Fig. 10. Comparison between n-dodecane density predictions along with temperature at 2 MPa between MD simulations against NIST dataset.

3.1. ML results for typical fuel surrogates

Tables 1 and 2 show the coefficient of determination (R^2-score) and L_2 mean relative error, respectively, for GP and the probabilistic conditional generative models. The accuracy metrics are computed with the test samples. We observe that they are not satisfactory in the small training data scenario, with 10% of training data. R2-scores for the GP and conditional generative models in this specific training scenario are 0.8538 and 0.9359, respectively. For a data richer situation, with 50% of training data, we observe that the models return good predictions with R^2-score higher than 0.99. Also, we observe that the conditional generative model returns better predictions than the GP model in a small data scenario, with an accuracy of $L_{2-MRE} = 2.8989 \times 10^{-3}$ while the GP accuracy is $L_{2-MRE} = 4.7428 \times 10^{-2}$. Finally, with 100% of the training data, we can see that the surrogate models return excellent predictions with R^2-score very near 1.0 and mean relative errors lower than 0.03%.

As a further illustration of the performance of such approaches to predict the density, we plot its values for n-octane, n-dodecane, and n-hexadecane densities with respect to temperature for the ML models trained with 50% of the dataset, since this training scenario returns the best relation between accuracy and computational cost. Fig. 3 shows the n-octane density predictions at the pressures equal to 3, 10, and 100 MPa. We can observe that at 3 MPa the GP model fails to deliver good results around the transcritical region, while the generative model provides robust predictions with uncertainties bounds that capture the data. The predictive uncertainty of the proposed approaches reflects limited data for training the models, the epistemic uncertainty. We can also note that both models perform well at 10 and 100 MPa, wherein the density dependency on the temperature has a smooth behavior.

Also, the n-dodecane and n-hexadecane densities are depicted along with temperature in Figs. 4 and 5. We observe that the ML models return robust predictions at three different pressures. Besides, it is noted that the GP model returns larger uncertainty bounds at high pressures, specifically at density points not used in the training process.

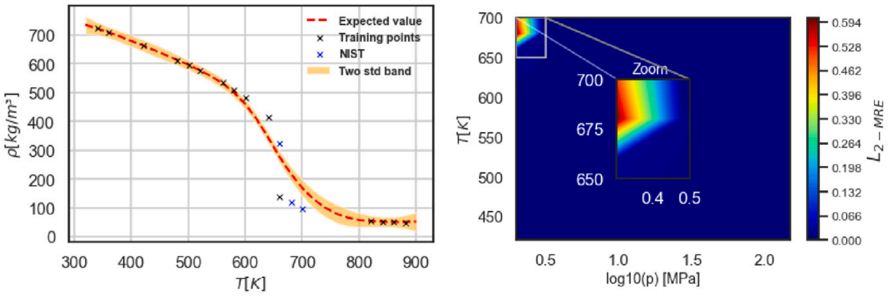
We also validate how the proposed ML technology perform in an extrapolation scenario. We validate them for the n-heptane, a fuel not used for building the models. In order to do that, instead of employing data provided by ML computations, we use an experimental database furnished by the National Institute of Standards and Technology (NIST). Fig. 6 shows that at 3 MPa and liquid condition the ML model returns good predictions of the n-heptane density behavior, with small uncertainties. However, at supercritical conditions (T_c = 540.13 K), the GP model returns density predictions far from satisfactory. Also, we note that the generative model has uncertainty bounds able to capture the thermophysical property. The L_2 mean relative error between the NIST dataset and the expected values predicted by the GP and conditional generative models are 7.1697×10^{-2} and 2.0838×10^{-2}, respectively. We can also note that at higher pressure where the density behavior is smooth, the models present better predictions, with the GP model showing larger uncertainties bounds and the generative model returns smaller uncertainty bounds. Moreover, the L_2 mean relative errors of the GP model at 10 and 100 MPa are respectively 1.8152×10^{-4} and 6.3072×10^{-4}, and for the conditional generative model the L_2 mean relative errors at the same pressures are 8.4484×10^{-5} and 2.0322×10^{-4}.

Furthermore, we use a coefficient of variation to measure the degree of uncertainty of the density predictions. It is defined as the ratio between the standard deviation σ_ρ and the mean μ_ρ of the prediction

$$\mathrm{cv}(p, T) = \frac{\sigma_\rho(p, T)}{\mu_\rho(p, T)} \tag{13}$$

Fig. 7 gives an overall picture by displaying a mapping between the operating conditions and the uncertainty on n-octane density predictions. We present an explicit quantification of the epistemic uncertainty resulting from the lack of data, which helps to understand limits of the ML models. More specifically, to make more accessible the visualization of the results, we plot this mapping for $\log_{10} p \in [0.5, 2.5]$ MPa and $T \in [320, 900]$ K with regular intervals of 20 K, allowing us to make explicit the strong dependence of the epistemic uncertainties regarding different regions of operating conditions. A critical aspect to be remarked is the higher values of cv in particular regions of the operating conditions space, especially at transcritical conditions displaying higher gradients of the property. We can note that the GP model returns a degree of uncertainty slightly large in this region. That can be mitigated by providing more training data for this specific region. Also, it is noted that variability of density provided by the conditional generative model is less pronounced at liquid regions and for high-pressure supercritical regions, which is due to the smooth density behavior resulting in a low degree of uncertainty in the predictions at these regions.

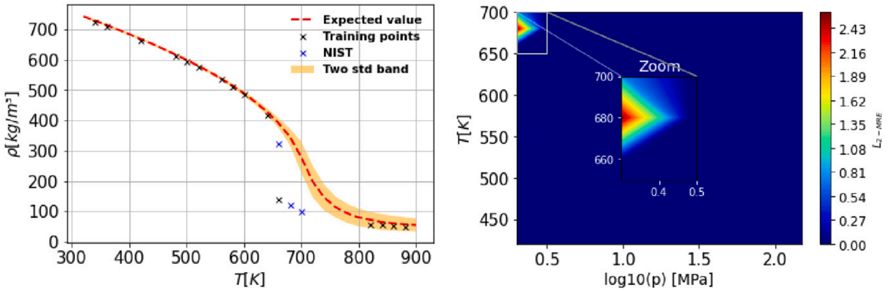
In addition, we explore the ability of ML models considering other physicochemical properties. Specifically, we extend the above approaches to predict the diffusion coefficient of the alkane compounds. The diffusion coefficient controls mass transport in combustion engines. Therefore, understanding diffusion is extremely important in order to optimize industrial processes and improve device efficiency, especially for supercritical combustion, where the physicochemical properties of fluids are quite different from those in liquid conditions. It is worth emphasizing that constructing accurate and simple predictive models overcoming costly simulations and expensive experimental procedures is crucial for describing physicochemical properties over a wide range of flow conditions. The dataset used to build the ML models consists of 1240 values of the diffusion coefficient, computed within a regular temperature grid $T \in [300, 900]$ K, varying by 20 K, and at specific pressures: $P = \{1, 2, 4, 10, 20, 40, 100, 150\}$ MPa. In the training process, 70% of the data points are selected randomly to training the ML models. The remaining 30% are used to testing them. Moreover, the training data set is organized in three subsets with 20%, 50%, and 70% of data. Fig. 8



(a) GP model

(b) L_2 relative error

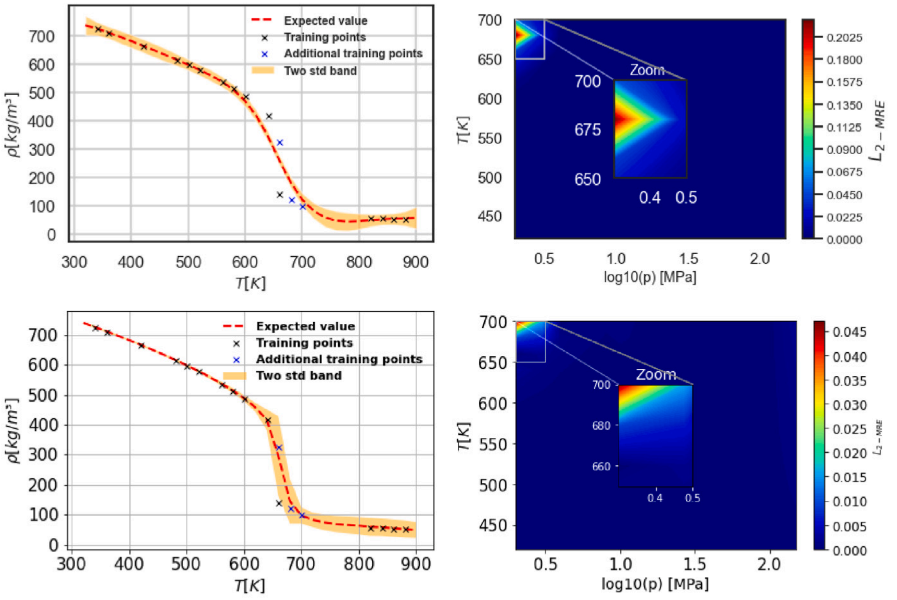
Fig. 11. n-Dodecane density predictions GP model: (a) n-Dodecane density along with temperature at 2 MPa. (b) L_2 error between the expected value predicted by the ML model against NIST.



(a) Conditional generative model

(b) L_2 relative error

Fig. 12. n-Dodecane density predictions conditional generative model: (a) n-Dodecane density along with temperature at 2 MPa. (b) L_2 error between the expected value predicted by the ML model against NIST.

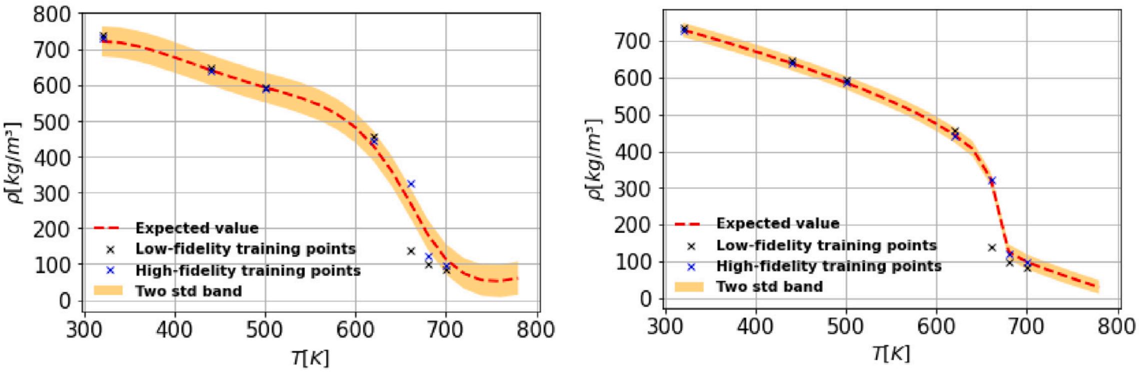


(a) 2 MPa

(b) L2 relative error

Fig. 13. n-Dodecane density predictions with the GP model (top) and conditional generative model (bottom) using the data-fusion approach with three density points from the NIST database.

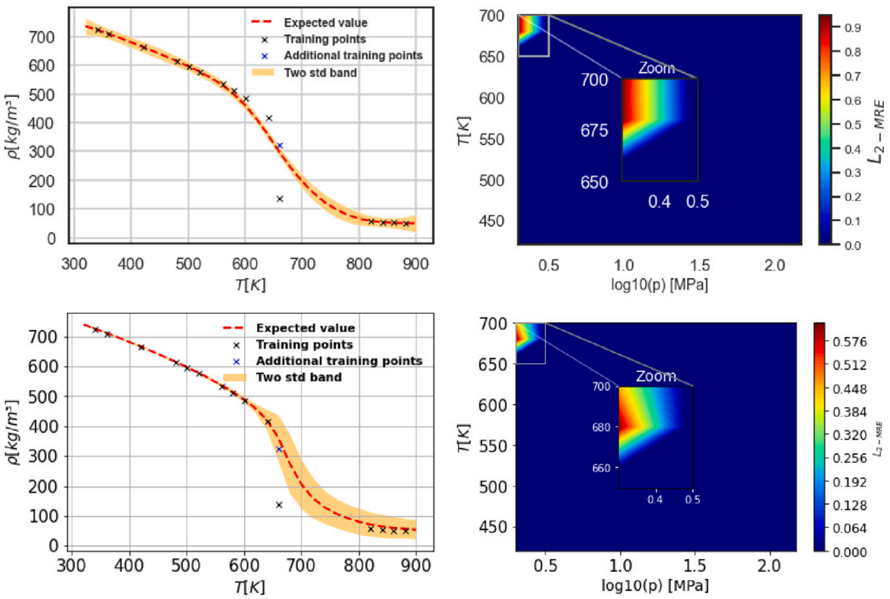
shows the n-dodecane diffusion coefficient predictions at the pressures equal to 1, 10, and 100 MPa for the ML models trained with 50% of the dataset. We observe that the ML models return robust predictions at three different pressures with GP model returns larger uncertainty bounds. We can also note that similar to density the model perform better at higher pressures, wherein the diffusion coefficient dependency on the temperature has a smooth behavior. That is further confirmed by calculating the L_2 mean relative error, where for a pressure of 1 MPa the models return worse predictions, as shown in Fig. 9. That might be explained by the fact that the physicochemical properties display higher



(a) NARGP model

(b) Conditional generative model

Fig. 14. Multi-fidelity modeling of n-dodecane diesel surrogate fuel density.



(a) 2 MPa

(b) L_2 relative error

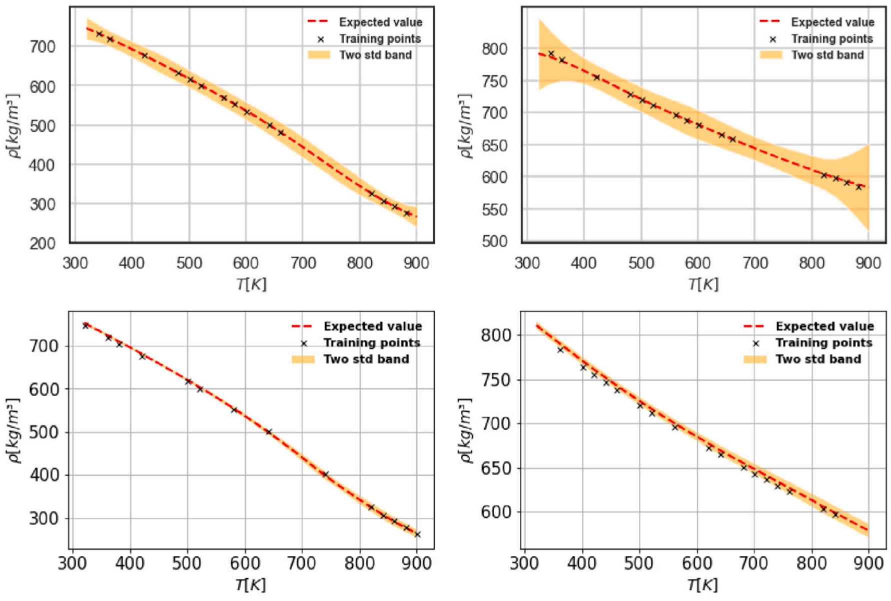
Fig. 15. n-Dodecane density predictions with the GP model (top) and conditional generative model (bottom) using the data-fusion-fidelity approach with one density point from the NIST database.

gradients near transcritical regions at lower pressures, which decreases the predictability of the models under these conditions. Also, we note that the generative model has slightly better predictions than the GP model. These results show the robustness of the proposed approaches to construct predictive models for physicochemical properties of diesel fuels.

3.2. Data-fusion machine learning models

Although MD simulation is considered to be a robust tool to predict thermodynamic properties, it returns unsatisfactory values at critical points/transcritical regions. It was shown [8] that the transport properties predictions of diesel surrogate fuels are far from satisfactory near such critical points. That is also the case with n-dodecane in that study. The results depict that EMD simulation might be unsuitable for predicting the properties at regions near the critical point. Non-equilibrium molecular dynamics simulation may leverage the results near the critical points, which is beyond the scope of the present study. Density predictions with MD simulations and NIST data at transcritical regions present considerable discrepancies, as shown in Fig. 10. More specifically, in operating conditions near the critical point of n-dodecane, critical pressure (P_c = 1.8170 MPa) and critical temperature (T_c = 658.1 K), our ML models based on the MD data fail to accurately predict the density. Figs. 11(a) and 12(a) show the density predictions at 2 MPa for GP and conditional generative model, respectively. Moreover, Figs. 11(b) and 12(b) also show that the main discrepancies between the expected values of ML models against the NIST database are into the transcritical regions.

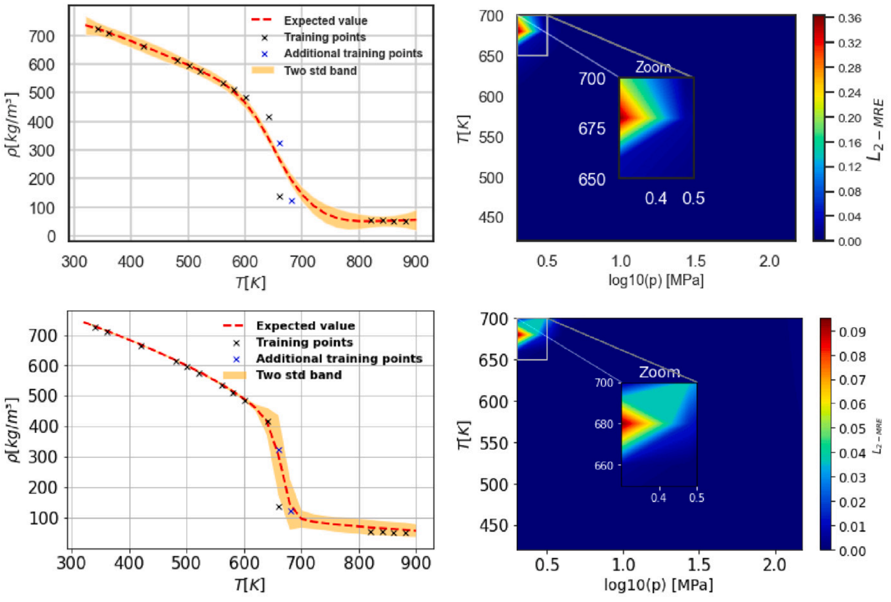
Aiming at improving the predictability of our ML models at transcritical regions, we adopt two strategies, exploring the fusion of MD simulations with experimental data. The aim here is not to compare these different strategies but to evaluate their potential. Both are formulated with the same idea, promoting the fusion of data from MD simulations and experiments datasets. In the first one, we propose a data-fusion strategy in which density points of the transcritical region provided by the NIST database are simply concatenated into the training dataset. The second differs as we propose a multi-fidelity arrangement of the data. A detailed description of both strategies is given further ahead.



(a) 10 MPa

(b) 100 MPa

Fig. 16. n-Dodecane density predictions with the GP model (top) and conditional generative model (bottom) using the data-fusion-fidelity approach with one density point from the NIST database at the pressures 10 and 100 MPa.



(a) 2 MPa

(b) L2 relative error

Fig. 17. n-Dodecane density predictions with the GP model (top) and conditional generative model (bottom) using the data-fusion-fidelity approach with two density points from the NIST database.

In the data-fusion approach, we add three density values from NIST to the original training dataset, as depicts in Fig. 13(a). Note that the fusion improves considerably the predictions of the conditional generative model with relative errors lower than 5%, while the GP model still returns relative errors not satisfactory. Further details about this data-fusion approach can be found in Appendix.

As discussed above, generating reliable data with MD simulations to be used in supervised learning might require a great computational effort. To tackle such a drawback, numerical formulations combining models displaying different levels of fidelity are frequently employed. Those multi-fidelity simulators employ, for instance, coarse grid discretizations, models based on simplified physics, or simplified iterative methods. Here, again we merge experimental data with MD simulations, restricting our approaches to two levels of fidelity.

In this new context, we propose extensions of the previous introduced ML models. We start by obtaining high-fidelity $\{\mathbf{x}_H, \gamma_H\}$ and low-fidelity $\{\mathbf{x}_L, \gamma_L\}$ input–output samples. Typically, the number of samples in the first case tends to be much smaller due to the related costs. We assign the high-fidelity score to the experimental data, according to the considerations above about the potential inaccuracy of the MD obtained computed properties for transcritical regions.

We start with our first, in this multi-fidelity context, ML model approximating the conditional probability $p(\gamma_H|\mathbf{x}_H, \gamma_L, z)$, using the generative model $\gamma_H = f_\phi(\mathbf{x}_H, \gamma_L, z)$, $z \sim p(z)$. In another words, the ML model is supposed to capture the correlation between the two level of fidelity data. Once this is achieved, we have a predictive model computing outputs for a new point $\mathbf{x}^*$: $\mathbf{y}^*_\mathbf{H} = f_\phi(\mathbf{x}^*, f_L(\mathbf{x}^*), z)$. At this point, it is worth remarking that one of the inputs is the output of the low-fidelity model, leading to a recursive scheme to obtain the predictions of the multifidelity model. In fact, here the considered low fidelity data is produced with expensive MD simulations. Therefore, in order to achieve a feasible scheme, we need to build an auxiliary, cheap to compute and accurate, proxy for the low fidelity model using the available data.

As a second approach, the one based on GPs, we employ the nonlinear autoregressive multi-fidelity GP (NARGP) regression model [53]. The main idea of the NARGP model is to extend GP modeling to capture nonlinear correlations from data generated by sources of different fidelity [72,73]. It enables the construction of probabilistic models prone to encapsulate uncertainties, built upon the recursive relation $y_H = g(x_H, f_L(x_H))$ involving low and high fidelity data, in which f_L is a GP model for the former. Moreover, we put a GP prior on g. After the training, we obtain the predictive model, which turns to be also a GP, $y_H = g(x^*, f_L(x^*))$.

To assess the above multi-fidelity ML approaches, we use an illustrative example involving data from “low-fidelity” MD simulations and “high-fidelity” NIST experimental values. For both approaches, the training dataset consists of 7 density values of n-dodecane $\rho_H(p, T_H)$ and $\rho_L(p, T_L)$, at the pressure of 2 MPa and a set of temperatures given by $T_H = T_L = \{320, 440, 500, 620, 660, 680, 700\}$ K. Note that we prioritize points located in the transcritical part, since this region presents larger discrepancies between the values predicted by MD simulations and the NIST database.

The conditional generative model is constructed using fully connected feed-forward architecture for the encoder and generator networks with 4 hidden layers and 100 neurons per layer, while the discriminator architecture has 2 hidden layers with 100 neurons per layer. All activation uses a hyperbolic tangent non-linearity. The models are trained for 20,000 stochastic gradient descent steps using the Adam optimizer [70] with a learning rate of 10^{-4}, while fixing a one-to-five ratio for the discriminator versus generator updates. Furthermore, we have fixed the entropy regularization parameter to $\lambda = 1.5$, and we also employed a one-dimensional latent space with a standard normal prior, $p(z) \sim \mathcal{N}(0, 1)$.

We train the NARGP model via maximizing the marginal log-likelihood using the Mattern 3/2 kernel function. The gradient descend optimizer L-BFGS is used considering randomized restarts to ensure convergence to a global optimum. Once the high-fidelity recursive GP is trained, we can compute the predictive posterior mean and variance at a given untested point $\mathbf{x}^*$ by sampling the probabilistic predictive model.

The main results are summarized in Fig. 14. More specifically, the results indicate that the NARGP model was able to satisfactorily reconstruct the high-fidelity data. To make this comparison quantitative, we compute the mean L_2 relative error between the expected values predicted by the generative model and the NIST data. It shows predictions with accuracy of $L_{2-MRE} = 1.4524 \times 10^{-2}$. Moreover, it returns good uncertainty bounds able to capture the high-fidelity response at the transcritical region. Also, we note a perfect agreement between the expected value provided by the probabilistic conditional generative model and the high-fidelity data, resulting in an accuracy of $L_{2-MRE} = 4.4782 \times 10^{-5}$. Finally, we observe that the multi-fidelity model returns small uncertainty bounds despite the small amount of data employed in the training process.

4. Conclusions

In this work, we propose a computational methodology based on the use of ML with Molecular Dynamics simulations to compute physicochemical properties of single compound fuels at engine-relevant conditions. The ML models have been revealed to be a powerful tool to predict accurately the fuel properties of pure compounds. Moreover, this study explores the versatility of the ML models to handle data from different sources, which can then be integrated efficiently in the context of UQ workflows with many-query tasks.

We place our contribution in the emerging area of physics-aware ML, where the final model, in many different ways, blends two main components: availability of experimental data and/or often expensive computational models relying on first principles and phenomenological closure equations, and deep learning data-driven models. Such combination allows describing physicochemical properties over a wide range of flow conditions at relatively low cost, and also offers a broad spectrum of opportunities to enhance CFD codes.

This study has shown a successful prediction of fuel physical quantities, in this case density and diffusion coefficient, that can also be extended to other physicochemical properties as well as more complex fuel molecules or multicomponent mixtures like dimethyl ethers or oxymethylene dimethyl ethers. The generation of reliable physicochemical properties of renewable fuels is an important step forward towards the generation of digital tools that can assist on the decarbonization by the use of renewable fuels.

CRediT authorship contribution statement

Rodolfo S.M. Freitas: Investigation, Software, Writing – original draft. **Ágatha P.F. Lima:** Investigation, Software. **Cheng Chen:** Investigation, Software. **Fernando A. Rochinha:** Conceptualization, Methodology, Writing – review & editing. **Daniel Mira:** Conceptualization, Methodology, Writing – review & editing. **Xi Jiang:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research leading to these results had received funding from the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) through Programa de Recursos Humanos (PRH) under the PRH 8 - Mechanical Engineering for the Efficient Use of Biofuels, grant agreement numbers F0A5.EDDE.B5C0.3BCB and 2B61.4F5C.A83B.A713.

Appendix. Data-fusion studies

In order to enhance the predictability of the ML models at transcritical regions, here we propose a data-fusion approach. Specifically, we concatenate density points of the transcritical region provided by the NIST database into the training dataset. The aim here is to improve the density predictability of our ML models, by supplying reliable information about this state variable in the specific region where MD data is scarce. Following this purpose, the first attempt is to add one density point from the NIST database. Here, we concatenate the n-dodecane density at pressure 2 MPa and temperature 660 K to the training data. By adding this point to the training set, it is verified that the ML models can recover the density at 660 K, as shown in Fig. 15. However, the L_2 relative errors between the expected values predicted by ML models and the NIST data are still considerable in transcritical regions. Also, we can note that the conditional generative model has larger uncertainty bounds at the transcritical region trying to recover density behavior due to the lack of data in this region. Furthermore, Fig. 16 shows that adding density points from NIST into the training data does not change the degree of uncertainty at other operating conditions.

As a further attempt to enhance the density predictions at the transcritical region, we now concatenate one more density point from the NIST database. More specifically, in addition to concatenating the n-dodecane density at pressure 2 MPa and temperature 660 K to the training data, we also add the n-dodecane density at 680 K. Fig. 17 shows that adding two density points from NIST data in the transcritical region slightly improves the predictions of the GP model, while the relative error remains considerable. However, we can verify that the generative model returns satisfactory predictions with L_2 relative error lower than 10% in the transcritical region. This shows the capability of the conditional generative model to enhance the predictability of the density when some pieces of information about the correct behavior of the transport property are given to the model.

Finally, to further increase the predictability of our ML models, a third attempt is proposed based on adding three density points from NIST to the training data, those being the n-dodecane densities at 2 MPa and temperatures 660, 680, and 700 K. Fig. 13 depicts that in this training scenario the density predictions of the GP model have some improvements, but the L_2 relative error is still considerable. Furthermore, we can verify that the conditional generative model returns accurate predictions, with relative errors lower than 5% in transcritical regions. Finally, we note that the generative model has uncertainty bounds able to recover the density predictions near the critical point.

References

[1] Naimoli S, Ladislaw S. Climate solutions series: Decarbonizing heavy industry. Tech. rep., Center for Strategic and International Studies (CSIS); 2020, URL http://www.jstor.org/stable/resrep26402.

[2] Mandová H, Vass T, Pales AF, Levi P, Gül T. The challenge of reaching zero emissions in heavy industry. Tech. rep., International Energy Agency (IEA); 2020, URL https://www.iea.org/articles/the-challenge-of-reaching-zero-emissions-in-heavy-industry.

[3] Omari A, Heuser B, Pischinger S. Potential of oxymethylenether-diesel blends for ultra-low emission engines. Fuel 2017;209(July):232–7. http://dx.doi.org/10.1016/j.fuel.2017.07.107.

[4] Pélerin D, Gaukel K, Härtl M, Jacob E, Wachtmeister G. Potentials to simplify the engine system using the alternative diesel fuels oxymethylene ether OME1 and OME3-6 on a heavy-duty engine. Fuel 2020;259:116231. http://dx.doi.org/10.1016/j.fuel.2019.116231, URL https://linkinghub.elsevier.com/retrieve/pii/S0016236119315856.

[5] Pastor JV, García-Oliver JM, Micó C, García-Carrero AA, Gómez A. Experimental study of the effect of hydrotreated vegetable oil and oxymethylene ethers on main spray and combustion characteristics under engine combustion network spray A conditions. Appl Sci 2020;10(16):5460.

[6] Pitz WJ, Mueller CJ. Recent progress in the development of diesel surrogate fuels. Prog Energy Combust Sci 2011;37(3):330–50. http://dx.doi.org/10.1016/j.pecs.2010.06.004, URL https://www.sciencedirect.com/science/article/pii/S0360128510000535.

[7] Lai JYW, Lin KC, Violi A. Biodiesel combustion: Advances in chemical kinetic modeling. Prog Energy Combust Sci 2011;37(1):1–14. http://dx.doi.org/10.1016/j.pecs.2010.03.001, URL https://www.sciencedirect.com/science/article/pii/S036012851000033X.

[8] Chen C, Jiang X. Transport property prediction and inhomogeneity analysis of supercritical n-dodecane by molecular dynamics simulation. Fuel 2019;244:48–60. http://dx.doi.org/10.1016/j.fuel.2019.01.181, URL https://www.sciencedirect.com/science/article/pii/S0016236119301826.

[9] Yang X, Zhang M, Gao Y, Cui J, Cao B. Molecular dynamics study on viscosities of sub/supercritical n-decane, n-undecane and n-dodecane. J Molecular Liquids 2021;335:116180.

[10] Yang X, Gao Y, Zhang M, Jiang W, Cao B. Comparison of atomic simulation methods for computing thermal conductivity of n-decane at sub/supercritical pressure. J Molecular Liquids 2021;117478.

[11] Kondratyuk ND, Pisarev VV, Ewen JP. Probing the high-pressure viscosity of hydrocarbon mixtures using molecular dynamics simulations. J Chem Phys 2020;153(15):154502.

[12] Kondratyuk N, Lenev D, Pisarev V. Transport coefficients of model lubricants up to 400 MPa from molecular dynamics. J Chem Phys 2020;152(19):191104.

[13] Kondratyuk ND, Pisarev VV. Calculation of viscosities of branched alkanes from 0.1 to 1000 MPa by molecular dynamics methods using COMPASS force field. Fluid Phase Equilib 2019;498:151–9.

[14] Caleman C, Van Maaren PJ, Hong M, Hub JS, Costa LT, Van Der Spoel D. Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. J Chem Theory Comput 2012;8(1):61–74.

[15] Freitas RSM, Rochinha FA, Mira D, Jiang X. Parametric and model uncertainties induced by reduced order chemical mechanisms for biogas combustion. Chem Eng Sci 2020;227:115949. http://dx.doi.org/10.1016/j.ces.2020.115949, URL https://www.sciencedirect.com/science/article/pii/S0009250920304814.

[16] Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput Mater 2016;2(16028). http://dx.doi.org/10.1038/npjcompumats.2016.28.

[17] Ramadhas AS, Jayaraj S, Muraleedharan C, Padmakumari K. Artificial neural networks used for the prediction of the cetane number of biodiesel. Renew Energy 2006;31(15):2524–33. http://dx.doi.org/10.1016/j.renene.2006.01.009, URL https://www.sciencedirect.com/science/article/pii/S0960148106000395.

[18] Piloto-Rodríguez R, Sánchez-Borroto Y, Lapuerta M, Goyos-Pérez L, Verhelst S. Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression. Energy Convers Manage 2013;65:255–61. http://dx.doi.org/10.1016/j.enconman.2012.07.023, URL https://www.sciencedirect.com/science/article/pii/S0196890412003093.

[19] Miraboutalebi SM, Kazemi P, Bahrami P. Fatty acid methyl ester (FAME) composition used for estimation of biodiesel cetane number employing random forest and artificial neural networks: A new approach. Fuel 2016;166:143–51. http://dx.doi.org/10.1016/j.fuel.2015.10.118, URL https://www.sciencedirect.com/science/article/pii/S0016236115011321.

[20] Faizollahzadeh Ardabili S, Najafi B, Shamshirband S. Fuzzy logic method for the prediction of cetane number using carbon number, double bounds, iodic, and saponification values of biodiesel fuels. Environ Prog Sustain Energy 2019;38(2):584–99. http://dx.doi.org/10.1002/ep.12960, URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/ep.12960.

[21] Tipler S, D'Alessio G, Van Haute Q, Parente A, Contino F, Coussement A. Predicting octane numbers relying on principal component analysis and artificial neural network. Comput Chem Eng 2022;161:107784. http://dx.doi.org/10.1016/j.compchemeng.2022.107784, URL https://www.sciencedirect.com/science/article/pii/S0098135422001259.

[22] Mostafaei M. ANFIS models for prediction of biodiesel fuels cetane number using desirability function. Fuel 2018;216:665–72. http://dx.doi.org/10.1016/j.fuel.2017.12.025, URL https://www.sciencedirect.com/science/article/pii/S0016236117315958.

[23] Bemani A, Xiong Q, Baghban A, Habibzadeh S, Mohammadi AH, Dorane-hgard MH. Modeling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. Renew Energy 2020;150:924–34. http://dx.doi.org/10.1016/j.renene.2019.12.086, URL https://www.sciencedirect.com/science/article/pii/S0960148119319585.

[24] Baghban A, Kardani MN, Mohammadi AH. Improved estimation of cetane number of fatty acid methyl esters (FAMEs) based biodiesels using TLBO-NN and PSO-NN models. Fuel 2018;232:620–31. http://dx.doi.org/10.1016/j.fuel.2018.05.166, URL https://www.sciencedirect.com/science/article/pii/S0016236118310111.

[25] Noushabadi AS, Dashti A, Raji M, Zarei A, Mohammadi AH. Estimation of cetane numbers of biodiesel and diesel oils using regression and PSO-ANFIS models. Renew Energy 2020;158:465–73. http://dx.doi.org/10.1016/j.renene.2020.04.146, URL https://www.sciencedirect.com/science/article/pii/S0960148120306844.

[26] Sánchez-Borroto Y, Piloto-Rodriguez R, Errasti M, Sierens R, Verhelst S. Prediction of cetane number and ignition delay of biodiesel using artificial neural networks. Energy Procedia 2014;57:877–85. http://dx.doi.org/10.1016/j.egypro.2014.10.297, URL https://www.sciencedirect.com/science/article/pii/S1876610214016646.

[27] Yu W, Zhao F. Prediction of critical properties of biodiesel fuels from FAMEs compositions using intelligent genetic algorithm-based back propagation neural network. Energy Sources A 2019;1–14. http://dx.doi.org/10.1080/15567036.2019.1641575.

[28] Alviso D, Artana G, Duriez T. Prediction of biodiesel physico-chemical properties from its fatty acid composition using genetic programming. Fuel 2020;264:116844. http://dx.doi.org/10.1016/j.fuel.2019.116844, URL https://www.sciencedirect.com/science/article/pii/S0016236119321982.

[29] Meng X, Jia M, Wang T. Neural network prediction of biodiesel kinematic viscosity at 313 K. Fuel 2014;121:133–40. http://dx.doi.org/10.1016/j.fuel.2013.12.029, URL https://www.sciencedirect.com/science/article/pii/S0016236113011733.

[30] Cheenkachorn K. Predicting properties of biodiesels using statistical models and artificial neural networks. 2006.

[31] Giwa SO, Adekomaya SO, Adama KO, Mukaila MO. Prediction of selected biodiesel fuel properties using artificial neural network. Front. Energy 2015;9:433–45. http://dx.doi.org/10.1007/s11708-015-0383-5.

[32] Rocabruno-Valdés CI, Ramírez-Verduzco LF, Hernández JA. Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel. Fuel 2015;147:9–17. http://dx.doi.org/10.1016/j.fuel.2015.01.024, URL https://www.sciencedirect.com/science/article/pii/S0016236115000381.

[33] de Oliveira FM, de Carvalho LS, Teixeira LSG, Fontes CH, Lima KMG, Câmara ABF, et al. Predicting cetane index, flash point, and content sulfur of diesel–biodiesel blend using an artificial neural network model. Energy Fuels 2017;31(4):3913–20. http://dx.doi.org/10.1021/acs.energyfuels.7b00282.

[34] Zhou L. Toward prediction of kinematic viscosity of biodiesel using a robust approach. Energy Sources A 2018;40(23):2895–902. http://dx.doi.org/10.1080/15567036.2018.1513099.

[35] Eryılmaz T, Yesilyurt M, Taner A, Celik ÅA. Prediction of kinematic viscosities of biodiesels derived from edible and non-edible vegetable oils by using artificial neural networks. Arab J Sci Eng 2015;40:3745–58.

[36] Eryilmaz T, Arslan M, Yesilyurt MK, Taner A. Comparison of empirical equations and artificial neural network results in terms of kinematic viscosity prediction of fuels based on hazelnut oil methyl ester. Environ Prog Sustain Energy 2016;35(6):1827–41. http://dx.doi.org/10.1002/ep.12410, URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/ep.12410.

[37] Zhu K, Müller EA. Generating a machine-learned equation of state for fluid properties. J Phys Chem B 2020;124(39):8628–39. http://dx.doi.org/10.1021/acs.jpcb.0c05806, pMID: 32870675.

[38] Leverant CJ, Harvey JA, Alam TM, Greathouse JA. Machine learning self-diffusion prediction for Lennard-Jones fluids in pores. J Phys Chem C 2021;125(46):25898–906. http://dx.doi.org/10.1021/acs.jpcc.1c08297.

[39] Allers JP, Priest CW, Greathouse JA, Alam TM. Using computationally-determined properties for machine learning prediction of self-diffusion coefficients in pure liquids. J Phys Chem B 2021;125(47):12990–3002. http://dx.doi.org/10.1021/acs.jpcb.1c07092, pMID: 34793167.

[40] Liu Y, Hong W, Cao B. Machine learning for predicting thermodynamic properties of pure fluids and their mixtures. Energy 2019;188:116091. http://dx.doi.org/10.1016/j.energy.2019.116091, URL https://www.sciencedirect.com/science/article/pii/S0360544219317864.

[41] Koukouvinis P, Rodriguez C, Hwang J, Karathanassis I, Gavaises M, Pickett L. Machine learning and transcritical sprays: A demonstration study of their potential in ECN spray-A. Int J Engine Res 2021;14680874211020292. http://dx.doi.org/10.1177/14680874211020292.

[42] Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press; 2006.

[43] Yang Y, Perdikaris P. Adversarial uncertainty quantification in physics-informed neural networks. J Comput Phys 2019;394:136–52. http://dx.doi.org/10.1016/j.jcp.2019.05.027, URL https://www.sciencedirect.com/science/article/pii/S0021999119303584.

[44] Yang Y, Perdikaris P. Conditional deep surrogate models for stochastic, high-dimensional, and multi-fidelity systems. Comput Mech 2019;64:417–34. http://dx.doi.org/10.1007/s00466-019-01718-y.

[45] Safarov J, Ashurova U, Ahmadov B, Abdullayev E, Shahverdiyev A, Hassel E. Thermophysical properties of diesel fuel over a wide range of temperatures and pressures. Fuel 2018;216:870–89. http://dx.doi.org/10.1016/j.fuel.2017.11.125, URL https://www.sciencedirect.com/science/article/pii/S0016236117315399.

[46] Pioro I, Mokry S, Draper S. Specifics of thermophysical properties and forced-convective heat transfer at critical and supercritical pressures. Rev Chem Eng 2011;27(3–4):191–214. http://dx.doi.org/10.1515/REVCE.2011.501.

[47] Pickett LM, Genzale CL, Bruneaux G, Malbec L-M, Hermant L, Christiansen C, et al. Comparison of diesel spray combustion in different high-temperature, high-pressure facilities. 2010, http://dx.doi.org/10.4271/2010-01-2106.

[48] Shen Y, Liu Y-B, Cao B-Y. C4+ surrogate models for thermophysical properties of aviation kerosene RP-3 at supercritical pressures. Energy Fuels 2021;35(9):7858–65. http://dx.doi.org/10.1021/acs.energyfuels.1c00326.

[49] Razi M, Narayan A, Kirby RM, Bedrov D. Fast predictive models based on multi-fidelity sampling of properties in molecular dynamics simulations. Comput Mater Sci 2018;152:125–33. http://dx.doi.org/10.1016/j.commatsci.2018.05.029, URL https://www.sciencedirect.com/science/article/pii/S0927025618303367.

[50] Xing WW, Shah AA, Wang P, Zhe S, Fu Q, Kirby RM. Residual Gaussian process: A tractable nonparametric Bayesian emulator for multi-fidelity simulations. Appl Math Model 2021;97:36–56. http://dx.doi.org/10.1016/j.apm.2021.03.041, URL https://www.sciencedirect.com/science/article/pii/S0307904X21001724.

[51] Koukouvinis P, Vidal-Roncero A, Rodriguez C, Gavaises M, Pickett L. High pressure/high temperature multiphase simulations of dodecane injection to nitrogen: Application on ECN spray-A. Fuel 2020;275:117871. http://dx.doi.org/10.1016/j.fuel.2020.117871, URL https://www.sciencedirect.com/science/article/pii/S001623612030867X.

[52] Alves V, Gazzaneo V, Lima FV. A machine learning-based process operability framework using Gaussian processes. Comput Chem Eng 2022;163:107835. http://dx.doi.org/10.1016/j.compchemeng.2022.107835, URL https://www.sciencedirect.com/science/article/pii/S0098135422001739.

[53] Perdikaris P, Raissi M, Damianou A, Lawrence ND, Karniadakis GE. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proc R Soc Lond Ser A Math Phys Eng Sci 2017;473(2198):20160751. http://dx.doi.org/10.1098/rspa.2016.0751, URL https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2016.0751.

[54] Su G, Peng L, Hu L. A Gaussian process-based dynamic surrogate model for complex engineering structural reliability analysis. Struct Saf 2017;68:97–109. http://dx.doi.org/10.1016/j.strusafe.2017.06.003, URL https://www.sciencedirect.com/science/article/pii/S016747301730214X.

[55] Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. Chemometr Intell Lab Syst 2007;87(1):59–71. http://dx.doi.org/10.1016/j.chemolab.2006.09.004, URL https://www.sciencedirect.com/science/article/pii/S0169743906001900.

[56] Yuan J, Wang K, Yu T, Fang M. Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression. Int J Mach Tools Manuf 2008;48(1):47–60. http://dx.doi.org/10.1016/j.ijmachtools.2007.07.011, URL https://www.sciencedirect.com/science/article/pii/S0890695507001265.

[57] Guerra GM, Freitas R, Rochinha FA. Constructing accurate phenomenological surrogate for fluid structure interaction models. In: Cavalca KL, Weber HI, editors. Proceedings of the 10th international conference on rotor dynamics – IFToMM. Cham: Springer International Publishing; 2019, p. 295–305.

[58] Kingma DP, Welling M. Auto-encoding variational Bayes. 2014, arXiv:1312.6114.

[59] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. 2014, arXiv:1406.2661.

[60] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. J Comput Chem 2005;26(16):1701–18.

[61] Martin MG, Siepmann JI. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. J Phys Chem B 1998;102(14):2569–77.

[62] Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 1981;52(12):7182–90.

[63] Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem 1997;18(12):1463–72.

[64] Yang X, Zhang M, Gao Y, Cui J, Cao B. Molecular dynamics study on viscosities of sub/supercritical n-decane, n-undecane and n-dodecane. J Molecular Liquids 2021;335:116180. http://dx.doi.org/10.1016/j.molliq.2021.116180, URL https://www.sciencedirect.com/science/article/pii/S0167732221009077.

[65] Kondratyuk N, Lenev D, Pisarev V. Transport coefficients of model lubricants up to 400 MPa from molecular dynamics. J Chem Phys 2020;152(19):191104. http://dx.doi.org/10.1063/5.0008907.

[66] Weisberg S. Applied linear regression. John Wiley & Sons, Inc.; 2005.

[67] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45(1–3):503–28.

[68] GPy. GPY: A Gaussian process framework in python. 2012, http://github.com/SheffieldML/GPy.

[69] Bergstra J, Yamins D, Cox DD. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th international conference on international conference on machine learning, vol. 28, (1):PMLR; 2013, p. 115–23.

[70] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2017, arXiv:1412.6980.

[71] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, Software available from tensorflow.org, URL http://tensorflow.org/.

[72] Gratiet LL, Garnier J. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. Int J Uncertain Quantif 2014;4(5):365–86.

[73] Kennedy MC, O'Hagan A. Predicting the output from a complex computer code when fast approximations are available. Biometrika 2000;87(1):1–13, URL http://www.jstor.org/stable/2673557.