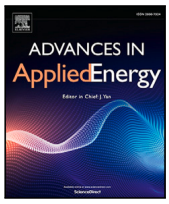
Advances in Applied Energy 19 (2025) 100226

Contents lists available at ScienceDirect

Advances in Applied Energy

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



Pathways to sustainable fuel design from a probabilistic deep learning perspective

Rodolfo S.M. Freitas [a,*], Zhihao Xing [a], Fernando A. Rochinha [b], Roger F. Cracknell [c], Daniel Mira [d], Nader Karimi [a], Xi Jiang [a]

[a] School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London, E1 4NS, UK
[b] COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[c] Shell Global Solutions, Shell Centre, London, SE1 7NA, UK
[d] Barcelona Supercomputing Center (BSC), Barcelona, 1–3, 08034, Spain

ARTICLE INFO

Keywords:
Fuel design
Fuel property prediction
Inverse design
Probabilistic deep learning
Scientific machine learning

ABSTRACT

To achieve net zero CO_2 emissions by 2050–2060, decarbonising the hard-to-abate sectors such as long-distance, heavy-duty transport is a top priority worldwide. These sectors are particularly challenging to decarbonise due to the use of high-energy-density liquid fossil fuels. In this context, designing low-carbon alternative fuels compatible with existing engines and fuel infrastructures is essential. This work presents an advanced fuel design framework to develop sustainable fuels that meet the high energy density requirements of heavy-duty vehicles. The fuel design approach is built upon a probabilistic perspective by considering a conditional generative model to predict the physicochemical properties of pure compounds and fuel blends with confidence bounds required for decision-making tasks. The probabilistic model is then integrated into an inverse design framework to design fuels with specific requirements. Finally, the fuel design framework is employed to develop new diesel fuel compositions according to the desired targets: ignition quality (cetane number) and sooting tendency (yielding sooting index). The AI-assisted fuel design approach can potentially lead to sustainable liquid fuels that are fully compatible with the existing utilisation equipment and can satisfy the requirements of different application sectors.

1. Introduction

On the journey to and beyond net zero carbon emissions, fuels will be continuously used in the hard-to-abate manufacturing sectors (e.g. cement, iron and steel, and chemicals) and transport applications where electrification is difficult [1]. Overall, global CO_2 emissions from fossil fuel combustion reached 35.8 $GtCO_2$ in 2023 [2], where the manufacturing and transportation sectors are responsible for emitting more than 50% CO_2 of the global greenhouse gas (GHG) emissions. Despite the urgent need to take action to meet climate targets, GHG emissions from fossil fuel combustion have risen year-on-year in the global transportation sector, from 5.1 $GtCO_2$ in 1990 [3] to almost 8.3 $GtCO_2$ in 2023 [4], a sharp increase of 63%. Addressing these emissions is essential to achieving international climate goals, such as the Paris Agreement's target of limiting anthropogenic warming to below 2 °C, net zero CO_2 emissions by 2050–2060, and net zero GHG emissions by 2070–2100 [5]. At the 29th Conference of the Parties (COP29) in Baku, Azerbaijan, in November 2024, decarbonising transportation was a central focus [6]. For the first time, United Nations organisations responsible for key transport sectors – the International Civil Aviation Organization (ICAO), International Maritime Organization (IMO), and the United Nations Economic Commission for Europe (UNECE) – joined forces to accelerate decarbonisation across all modes of transport. This unified effort highlights the critical role of sustainable fuels in reducing GHG emissions and achieving net zero targets.

In particular, most emissions from the transportation sector are due to long-haul transport, including ships, trucks, and planes [7]. These sectors rely heavily on fossil fuels such as diesel and kerosene. Their extraction, processing, and combustion contribute to a range of problems such as air pollution, resource depletion, and climate change [5]. In addition, future projections consistently anticipate continued growth in fuel demand, driven by urbanisation, population growth, and economic expansion, intensifying pressure on fuel resources and the environment. This is the case for fuels for road freight vehicles, which account for roughly 17 million barrels per day (mb/d), almost a fifth of global oil demand, and are projected to rise by 5 mb/d by 2050 [8]. However, sustainable alternatives, such as biofuels, are not being produced at a

* Corresponding author.
E-mail address: rodolfo.dasilvamachadodefreitas@qmul.ac.uk (R.S.M. Freitas).

https://doi.org/10.1016/j.adapen.2025.100226
Received 15 February 2025; Received in revised form 15 May 2025; Accepted 24 May 2025
Available online 11 June 2025

Nomenclature

Acronyms

AI	artificial intelligence
CFD	computational fluid dynamics
CFR	Cooperative Fuel Research
CN	cetane number
COP29	29th Conference of the Parties
CVCC	constant volume combustion chamber
DME	dimethyl ether
ECFPs	extended-connectivity fingerprints
FAMEs	fatty acid methyl esters
FT	Fischer–Tropsch
GAN	generative adversarial network
GHG	global greenhouse gas
GNN	Graph Neural Network
GPR	gaussian process regression
ICAO	International Civil Aviation Organization
IMO	International Maritime Organization
IQT	ignition quality tester
KL	Kullback–Leibler
LCA	life cycle assessment
ML	machine learning
MLP	multi-layer perceptron
Mol2Vec	molecule-to-vector fingerprints
OMEx	oxymethylene ethers
QSPR	quantitative structure–property relationship
Seq2Seq	sequence-to-sequence fingerprints
SLSQP	sequential least squares programming
SMILES	simplified molecular-input line-entry system
UNECE	United Nations Economic Commission for Europe
UQ	uncertainty quantification
VAE	variational autoencoder
YSI	yielding sooting index

Parameters and variables

β	residual penalty parameter
η	hyper-parameter vector of the encoder distribution
γ	A generic property
λ	entropy regularisation parameter
$\mathcal{L}_D$	discriminator loss
$\mathcal{L}_G$	generator loss
μ_γ	predictive distribution mean
ω	composition vector
Φ	molecular representation vector
ψ	hyper-parameter vector of the discriminator
σ_γ	predictive distribution standard deviation
θ	hyper-parameter vector of the decoder distribution
ξ	l_1 regularisation parameter
f_θ	generative model
N_s	number of samples
p	probability distribution
p_θ	generative model distribution
q_η	latent variable distribution
T_ψ	discriminator network
z	latent variable

rate that can fully meet fuel demand. In 2023, global biofuel production reached 4.2 mb/d, with projections to reach 4.8 mb/d by 2030 [9]. The demand scenarios clearly illustrate a gap that needs to be filled with scalable solutions to decarbonise the sector while meeting growing energy demands.

In this context, designing low-carbon alternative fuels derived from atmospheric carbon or waste feedstocks is becoming increasingly important for the transportation sector [10–12]. Particularly, sustainable transportation fuels, such as biofuels, hydrogen, and synthetic fuels, offer a promising path for heavy-duty vehicles [13,14]. Liquid synthetic fuels, including oxymethylene ethers (OME_x), fatty acid methyl esters (FAMEs), Fischer–Tropsch (FT) fuels, bio-based fuels, and long-chain alcohols, mimic the physicochemical properties of fossil fuels and can significantly reduce GHG emissions [15,16]. These fuels also address environmental challenges, such as air pollution and secondary aerosol formation, by minimising harmful pollutants like nitrogen oxides (NOx), sulphur oxides (SOx), and particulates (PM) [17]. Unlike conventional fuels, synthetic fuels avoid the NOx-soot trade-off [18,19] and can further reduce pollutant formation by using modern engine and exhaust after-treatment technologies [20,21]. Liquid synthetic fuels have shown promise to be a sustainable alternative to liquid fossil fuels.

Modern transportation systems increasingly rely on advanced engines optimised for efficiency and performance, necessitating fuels with specific properties, such as high energy density and tailored volatility [22]. This can ensure efficient atomisation and combustion with a direct impact on emissions [23]. However, alternative fuels, composed of complex mixtures, present challenges due to variability in composition, which is associated with their source and production process [7]. Designing novel fuel blends compatible with existing engines and infrastructure is critical for transitioning to sustainable fuels as drop-in fuels. Based on life cycle assessment (LCA), the traditional fuel design approach provides a holistic evaluation of environmental impact, energy efficiency, and economic feasibility [16,24]. In the traditional approach, fuel blends are designed by combining known pure compounds and the compositions are adjusted iteratively to meet fit-for-purpose fuel standard specifications [25]. While widely used, this empirical approach often relies on trial and error, making it inefficient for complex mixtures with non-linear behaviours between composition and properties. Additionally, because the mapping between the composition and physicochemical properties is not known a priori, creating a new blend requires extensive physical testing in engines or combustion chambers, making the process time-consuming and expensive. Furthermore, the conventional approach is restricted to a chemical space containing known compounds, rendering it impractical for newly discovered or designed chemical compounds. Therefore, more sophisticated fuel design methods are urgently needed to move towards sustainable fuel utilisation.

Artificial intelligence (AI) has become a transformative tool for addressing challenges in fuel design through data-driven approaches to predict and optimise fuel properties. Techniques such as machine learning (ML) and deep learning provide predictive capabilities that complement traditional methods. By leveraging large datasets, AI can accurately predict critical physicochemical properties of fuels, including viscosity, density, energy content, along with combustion characteristics and emissions [26]. ML and big data analytics enable mapping fuel composition to key utilisation properties, with studies predicting ignition quality [27–29] and emissions [30,31]. Deep learning models like variational autoencoders (VAEs) have also been used to generate

fuel molecules with specific properties [32]. Additionally, AI-assisted inverse fuel design frames fuel mixture development as a constrained optimisation problem, identifying compound combinations to achieve the desired properties [33]. These advancements accelerate fuel formulation screening, significantly reducing development timelines and costs.

In addition to property prediction, AI enables optimisation strategies that consider the multifaceted requirements of sustainable fuel design. AI-powered multi-objective optimisation frameworks balance competing priorities, such as maximising energy efficiency while minimising emissions. Additionally, AI models can integrate data from sources like molecular dynamics simulations, experimental measurements, and computational fluid dynamics (CFD) studies, offering holistic insights into real-world fuel performance. Integrating AI with digital tools, such as computational simulations and digital twins, further enhances decision-making and lowers time-to-solution in fuel design. Digital twins, virtual replicas of physical systems, simulate fuel behaviour in engines and supply chains, enabling rapid prototyping and validation [34]. AI augments these simulations by identifying patterns, predicting outcomes, and optimising processes in real-time, supporting the seamless adoption of sustainable fuels. By leveraging AI-driven strategies and a probabilistic deep learning perspective, this approach addresses key challenges in fuel development and advances the scalable deployment of next-generation fuels, contributing to a low-carbon transportation future.

1.1. A probabilistic data-driven approach for the design of sustainable fuels

Conventional fuel formulation is an ad hoc process that aims to find a mixture of compounds that match the desired properties, where a fuel blend is commercially acceptable as long as it meets the fuel quality standards. In general, the fuel compounds and their respective proportions within the mixture are not fixed a priori, and the fuel composition may change depending on the source, refinery standards, and production process. Accordingly, commercially available liquid fuels such as fossil diesel and kerosene can include hundreds or thousands of compounds, with their exact compositions largely unknown. In practice, additives may be added to the mixtures to enhance the properties of the fuel blends. When designing alternatives to these fuel blends, however, the empirical nature of the conventional approach does not help because the highly complex relationships between compositions and properties remain unknown. Therefore, more effective approaches to designing sustainable fuel mixtures are paramount to identify alternatives to fossil fuel blends.

Artificial intelligence, such as promising ML tools, has shown potential to revolutionise fuel design formulation by building accurate predictive models for key physicochemical properties of fuel utilisation [28,30] while also helping to screen fuel mixtures in inverse design formulations [33]. Integrated advanced AI tools can be used to leverage fuel screening, significantly dwindling development timelines and costs compared to the conventional fuel design approach. However, the lack of rigorous quantification of the predictive uncertainty of ML models does not provide the confidence needed for decision-making. Uncertainty quantification (UQ) is critical in the context of fuel design. Typically, the physicochemical properties of pure compounds and fuel mixtures are collected from different sources with variable fidelities or corrupted by complex noisy processes. Also, due to variability in the purity of the fuel compounds, the same method may produce different measurements [35]. All in all, these lead to unavoidable uncertainties. Therefore, probabilistic deep learning models capable of building predictive models with rigorous and robust uncertainty quantification are essential to providing the confidence needed for informed decision-making in the fuel design approach.

In the present work, we aim to formulate a robust probabilistic predictive model for the physicochemical properties of fuels required for a decision-making fuel design framework to explore sustainable low-emission fuels. Unlike the existing predictive machine learning models, the proposed methodology leverages data obtained from different sources with intrinsic uncertainties to build quantitative structure–property relationship (QSPR) models that produce reliable confidence intervals. Such models can be embedded into an inverse design framework to screen fuel blends matching fit-for-purpose fuel performance requirements while accounting for uncertainties in physicochemical properties.

In particular, we explore the use of a probabilistic conditional generative model [36] that enables an end-to-end building of predictive models to relate molecular representations with macroscopic properties, using data that may be noisy and sparse. This approach outperformed traditional Gaussian Process Regression (GPR) for estimating the physicochemical properties of pure compound fuels [37]. Although GPR has excellent distance-awareness and has been successfully applied as a proxy for many physics-based models [38,39], it may suffer from a lack of scalability if employed to high-dimensional molecular representations, rendering this approach unfeasible for the fuel design approach aimed here. We propose a conditional generative model taking a molecular representation (structural features) using a simplified molecular-input line-entry system (SMILES) representation of the compounds as inputs. It is worth highlighting that such molecular representation is similar to a latent representation, an embedded representation that captures key structural features of the fuel molecule, that enables pure compounds and fuel mixtures to be used in downstream learning tasks. In this regard, we tested different molecular representations to evaluate their impact on model predictability.

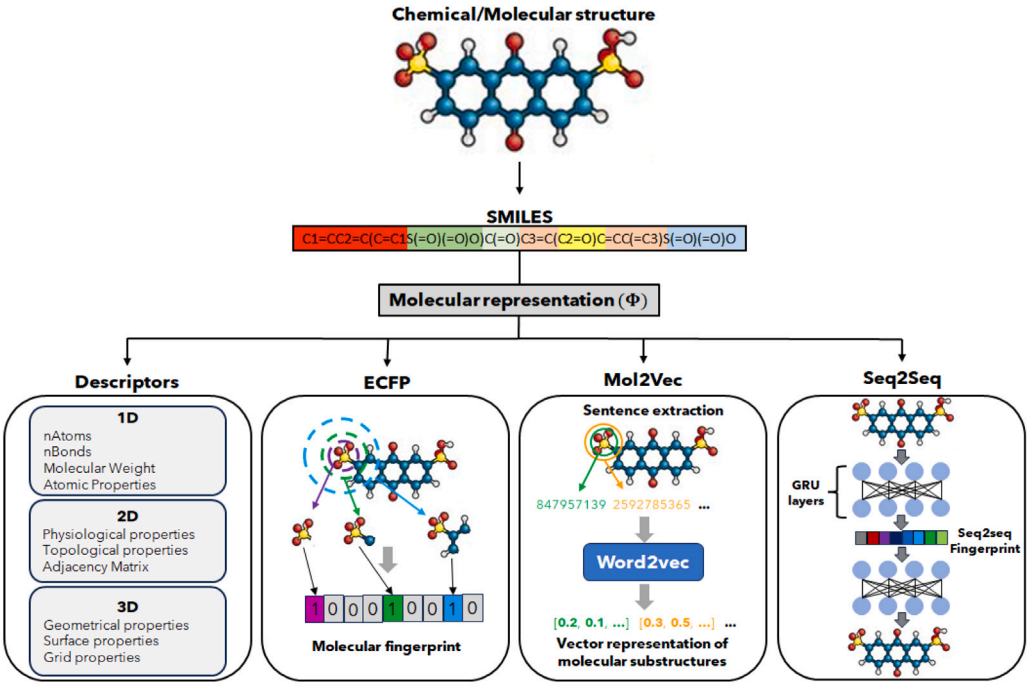
Fig. 1 illustrates a schematic of the proposed fuel design methodology. The pathways for the proposed fuel design approach consist of (i) collecting data on critical physicochemical properties of fuels, (ii) using the conditional generative model to build a detailed quantitative mapping between the chemical composition of the fuel and the key properties of fuel utilisation. The fuel chemical structures are represented using four different molecular representations: Descriptors, extended-connectivity fingerprints (ECFPs), molecule-to-vector (Mol2Vec) fingerprints, and sequence-to-sequence (Seq2Seq) fingerprints, as shown in Fig. 1(a). Such molecular representations can be used as input to characterise the molecular structure of fuel compounds in ML models, and here the goal is to evaluate how various computational representations of molecules influence the predictability of the generative model. Further details on the different molecular representations are provided in the next section. Finally, (iii) after training the conditional generative model that maps the molecular structure of the fuel to key combustion characteristics, we leverage it in a formal inverse design framework to screen novel fuel blends that match the required performance properties with quantified uncertainties compatible with fuel standards and components of current engines and infrastructure. These pathways allow the rapid integration of alternative liquid fuels into the current transport system, moving towards sustainable fuel utilisation.

The remainder of this paper is organised as follows. A detailed description of the fuel design methodology is presented in Section 2. Then, the databases of key physicochemical properties are described in Section 3. The performance of the conditional generative model to build predictive models and a case study to screen new fuel compositions that mimic diesel-like fuel are presented in Section 4. Finally, the main conclusions are presented in Section 5.

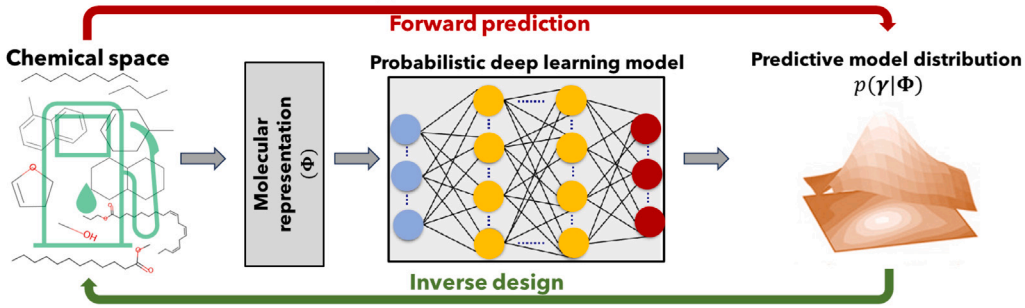
2. Methodology

This section outlines the framework for fuel design aimed at developing sustainable fuels that meet the specific requirements of heavy-duty vehicles. The main steps in the fuel design framework include:

1. Selection of a molecular representation Φ, which allows the molecular structure of fuels to be represented as a set of features useful for constructing QSPR models;



(a) Molecular representation



(b) Fuel design methodology

Fig. 1. Schematic overview of the proposed fuel design methodology. The end-to-end fuel design methodology consists of: (a) Different ways of representing a molecule in cheminformatics, and (b) a robust predictive model with quantified uncertainty (Forward Prediction) and a formal framework to develop new fuel mixtures (Inverse design).

2. Develop a robust conditional generative model that establishes a compositional mapping between key fuel properties and component chemical structures, while ensuring thorough and reliable uncertainty quantification;
3. Build an optimisation framework to screen optimal fuel compositions that satisfy performance and sustainability requirements.

The step-by-step integration of these stages enables the development of an innovative end-to-end AI-assisted framework that ensures a streamlined and scalable pipeline for the development and optimisation of fuels to meet specific performance, environmental, and economic requirements while significantly reducing development timelines and costs.

2.1. Molecular representation

The computational representation of molecules remains an open research dilemma in chemical informatics [40]. Many molecular representations exist, including graphs, text, binary vectors, or learned embedded continuous feature vectors. These representations can be fed into machine learning models to perform downstream tasks, including classification, regression, and inverse design [41]. In the present work, the molecular representations Φ are generated from the SMILES, a formal 1D text representation of molecules derived from molecular graph theory [42]. In addition, the molecular representation of a fuel mixture is defined as a linear combination of molecular representations weighted by their respective compositions, $\Phi(\omega) = \sum_{i=1}^{n} \omega_i \Phi_i$. To keep a compact notation, we refer to Φ for the molecular representation of a pure compound and a fuel mixture henceforth.

In this context, to evaluate the impact of the molecular representation on the predictability of the QSPR models, we tested four different computational representations of molecules.

Descriptors. Molecular descriptors reflect complex structural patterns such as the number of atoms in the molecule, molecular shape, size, as well as atomistic properties. Such a molecular representation has shown promise for constructing accurate and interpretable predictive models for key properties of fuel utilisation [28]. The open-source descriptor-calculator Mordred software [43] is used to calculate the descriptors.

ECFPs. Extended-connectivity fingerprints are a class of topological fingerprints for molecular representation that encode the presence of specific molecular features in fixed-length binary vectors [44]. Here, the ECFPs are obtained using the DeepChem package [45] with a fingerprint radius of 2 and a vector length of 2048 bits.

Mol2Vec. Molecule-to-vector is an unsupervised learning approach to generate an embedded representation of molecular substructures based on the natural language approach Word2Vec [46]. Once the Mol2vec

fingerprints are learned, they can be fed into supervised machine-learning models for predictive tasks. A default model with a fingerprint radius of two and embedded size of 300, trained on 19.9 million compounds [47] is used to generate the molecular representation of the present work.

Seq2Seq. Sequence-to-sequence is an unsupervised molecular embedding learning approach to design fixed-length vector molecular representation [48]. In particular, an auto-encoder network is proposed to map the molecule into an embedding representation, in which the encoder and decoder are built using multi-layered Gated Recurrent Unit (GRU) networks. As a result, the embedding vector represents a hidden state of the molecule containing rigorous information to perform further tasks. Here, a pre-trained model, using the Molecular Sets (MOSES) dataset [49], with encoder and decoder networks containing 2 recurrent layers and hidden sizes equal to 256 and a feature vector length of 512 is used to yield the molecular representation.

The molecular representations are provided as input to the machine learning model. Typically, such representations present a high dimensionality, that is, $\boldsymbol{\Phi} \in \mathbb{R}^N$, where N is very large. In the next section, we explore a generative methodology that explores existing low-dimensional structures capable of explaining high-dimensional molecular representation by introducing probabilistic latent variables.

2.2. Probabilistic conditional generative model

Building robust predictive models for the physicochemical properties of alternative fuels is a challenge. Commonly, we are dealing with data collected from different levels of fidelity or corrupted by complex noisy processes. Here, we explore a probabilistic conditional generative methodology [36], that integrates variational autoencoders (VAEs) [50] and generative adversarial networks (GANs) [51]. Here, unlike generative models found in literature, which learn to generate data from a distribution without guidance, the conditional generative model learns to generate physicochemical properties conditioned on the molecular structure of fuels, levels of fidelity, and uncertainties. Furthermore, it employs a probabilistic perspective that enables the end-to-end training of predictive models from limited and noisy data and can also deal with the high dimensionality of the molecular representation.

In this context, the focal point is to build probabilistic predictive models that follow a conditional probability density function $p(\gamma|\boldsymbol{\Phi})$ capturing the statistical dependence between the physicochemical property and the molecular representation and employing the available data $\mathcal{D}$. Therefore, the predictive model can provide accurate estimates for the physicochemical property γ by calculating the expectation $\mathbb{E}[\gamma|\boldsymbol{\Phi}, \mathcal{D}]$, and also, to quantify the uncertainty associated with the measurements of the physicochemical properties, $\mathbb{V}\text{ar}[\gamma|\boldsymbol{\Phi}, \mathcal{D}]$. The main component of this approach is the introduction of latent random variables to identify a hidden low-dimensional representation of the data structure. Such latent variables allow us to represent the conditional probability $p(\gamma|\boldsymbol{\Phi})$ as an infinite mixture model,

$$p(\gamma|\boldsymbol{\Phi}) = \int p(\gamma, z|\boldsymbol{\Phi})\, dz = \int p(\gamma|\boldsymbol{\Phi}, z)\; p(z|\boldsymbol{\Phi})\; dz \tag{1}$$

where $p(z|\boldsymbol{\Phi})$ is a prior distribution on the latent variables. The above hierarchical mathematical ansatz postulates that the physicochemical properties γ can be generated by a transformation of the molecular representation $\boldsymbol{\Phi}$ and latent variables z, $\gamma = f_\theta(\boldsymbol{\Phi}, z)$. The final conditional generative model, $\gamma = f_\theta(\boldsymbol{\Phi}, z)$, can be learned from a regularised adversarial learning framework [36].

Adversarial learning framework. The main idea is to merge a VAE and a GAN to approximate the true underlying conditional distribution $p(\gamma|\boldsymbol{\Phi})$ and intractable posterior of the latent variables $p(z|\gamma, \boldsymbol{\Phi})$ with parametrised approximating distributions $p_\theta(\gamma|\boldsymbol{\Phi})$ and $q_\eta(z|\boldsymbol{\Phi}, \gamma)$, where these distributions are modelled using deep neural networks with parameters $\{\theta, \eta\}$, as shown in Fig. 2. The VAE allows modelling the posterior distribution of the latent space $q_\eta(z|\boldsymbol{\Phi}, \gamma)$. At the same time, the physicochemical property is generated by pushing the molecular representation $\boldsymbol{\Phi}$ and samples from the latent space distribution $p(z)$ through the generator network $\gamma = f_\theta(\boldsymbol{\Phi}, z)$. Finally, the discriminator network T_ψ, with model parameters ψ, is employed to classify samples generated between the conditional generative model distribution $p_\theta(\boldsymbol{\Phi}, \gamma)$ and the observed data distribution $q(\boldsymbol{\Phi}, \gamma)$. This approach resembles the GANs methodology, which defines a zero-sum game between the generator and the discriminator

$$\mathcal{L}_\mathcal{D}(\psi) = \mathbb{E}_{q(\boldsymbol{\Phi})p(z)}[\log \sigma(T_\psi(\boldsymbol{\Phi}, f_\theta(\boldsymbol{\Phi}, z)))] + \mathbb{E}_{q(\boldsymbol{\Phi},\gamma)}[\log(1 - \sigma(T_\psi(\boldsymbol{\Phi}, \gamma)))] \tag{2}$$

$$\begin{aligned}\mathcal{L}_\mathcal{G}(\theta, \eta) = \mathbb{E}_{q(\boldsymbol{\Phi},\gamma)p(z)}[T_\psi(\boldsymbol{\Phi}, f_\theta(\boldsymbol{\Phi}, z)) + (1 - \lambda)\log(q_\eta(z|\boldsymbol{\Phi}, f_\theta(\boldsymbol{\Phi}, z)))\\ + \beta\|f_\theta(\boldsymbol{\Phi}, z) - \gamma\|^2],\end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function and β is a residual penalty parameter that encourages a good fit to the measured data for supervised tasks, i.e., $\|f_\theta(\boldsymbol{\Phi}, z) - \gamma\|^2$. Also, $\lambda \geq 1$ represents an entropy regularisation parameter for controlling and mitigating the effect of mode collapse [52]. Here, the binary cross-entropy discriminator loss ($\mathcal{L}_\mathcal{D}$) aims to improve the ability of the discriminator T_ψ to distinguish *fake* samples produced by the generative model $f_\theta(\boldsymbol{\Phi}, z)$ and *true* samples from the observed data $q(\boldsymbol{\Phi}, \gamma)$, while the generator loss ($\mathcal{L}_\mathcal{G}$) aims at leveraging the ability of the generative model to *fool* the discriminator. Also, the encoder network $q_\eta(z|\boldsymbol{\Phi}, f_\theta(\boldsymbol{\Phi}, z))$ provides a variational approximation for the posterior of the latent variables at the end of the learning procedure.

The model parameters $\{\theta, \eta, \psi\}$ of the discriminator (T_ψ), encoder (q_η) and generator (f_θ) networks are estimated using the stochastic optimisation method by alternatively minimising the generator loss and maximising the discriminator loss as

$$\min_{\theta,\eta} \max_{\psi}\; \mathcal{L}_\mathcal{G}(\theta, \eta),\; \mathcal{L}_\mathcal{D}(\psi) \tag{4}$$

and a detailed overview of the training process is provided in Algorithm 1.

Algorithm 1 Training of conditional generative model.

Require: Dataset $D = \{\boldsymbol{\Phi}, \gamma\}$ and number of training iterations N_{iter}.
$it = 0$
while $it < N_{iter}$ **do**
  Sample a random batch of n samples $\{z_1, \cdots, z_n\}$ from prior distribution of the latent variable $p(z)$
  Sample a random batch of n training pairs $\{\boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_n\}$ and $\{\gamma_1, \cdots, \gamma_n\}$.
  Discriminator
  Compute the discriminator loss:
  $\mathcal{L}_\mathcal{D}(\psi) = \frac{1}{n}\sum_{i=1}^{n}[\log \sigma(T_\psi(\boldsymbol{\Phi}_i, f_\theta(\boldsymbol{\Phi}_i, z_i))) + \log(1 - \sigma(T_\psi(\boldsymbol{\Phi}_i, \gamma_i)))]$
  Update the discriminator network by maximising $\mathcal{L}_\mathcal{D}(\psi)$ using $\nabla_\psi \mathcal{L}_\mathcal{D}(\psi)$.
  Generator
  Compute the generator loss:
  $\mathcal{L}_\mathcal{G}(\theta, \eta) = \frac{1}{n}\sum_{i=1}^{n}[T_\psi(\boldsymbol{\Phi}_i, f_\theta(\boldsymbol{\Phi}_i, z_i)) + (1 - \lambda)\log(q_\eta(z_i|\boldsymbol{\Phi}_i, f_\theta(\boldsymbol{\Phi}_i, z_i))) + \beta\|f_\theta(\boldsymbol{\Phi}_i, z_i) - \gamma_i\|^2]$
  Update the generator network by minimising $\mathcal{L}_\mathcal{G}(\theta, \eta)$ using $\nabla_{\theta,\eta} \mathcal{L}_\mathcal{G}(\theta, \eta)$.
  Next iteration: it = it + 1
end while

Training and hyperparameters. For consistency, we departed from the architecture proposed and validated by Yang and Perdikaris [36]. We combined the try-and-error and Ray Tune tuning library [53] to search for the hyperparameters that provide the smallest mean-squared-error between the conditional model and the observed data to build the conditional generative model. In particular, the conditional generative model consists of feed-forward neural networks with 3 hidden layers and 100 neurons per layer for the encoder and generator. In comparison, the discriminator has two hidden layers and 100 neurons per layer. All hidden layers use a rectified linear unit activation function.

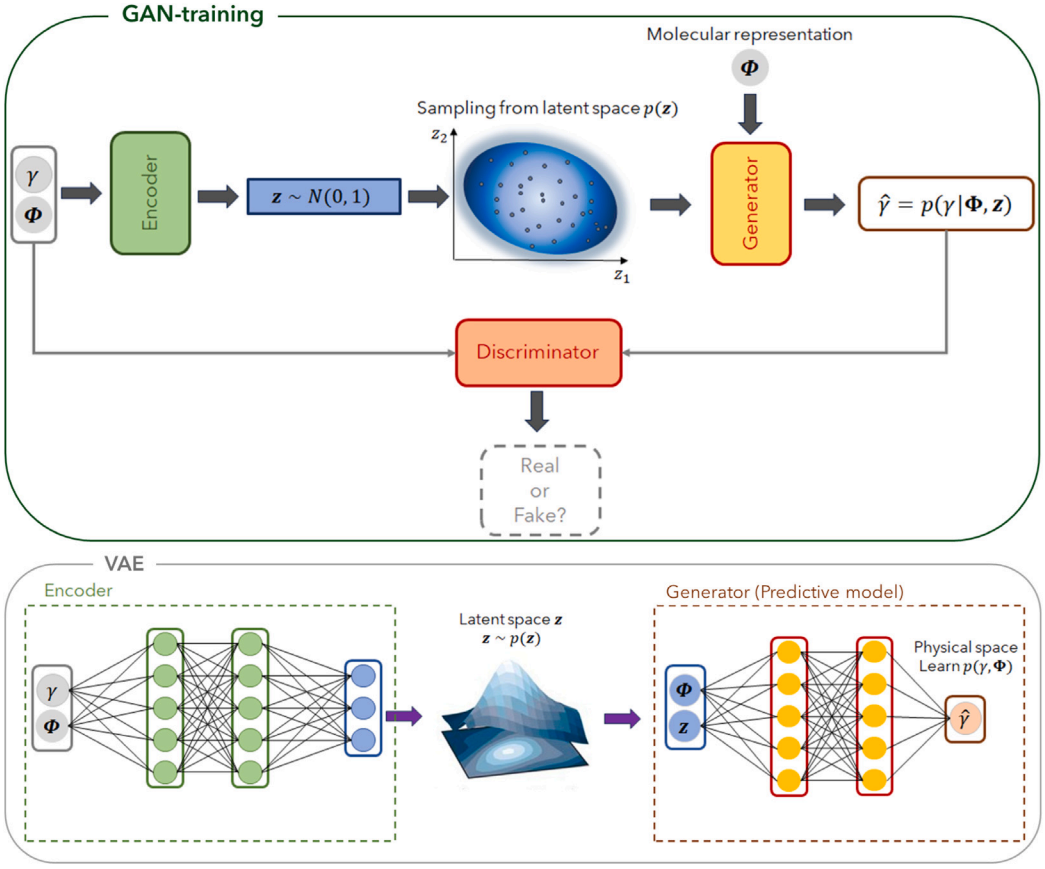


Fig. 2. Schematic illustration of the conditional generative model. The encoder network approximates the posterior for the latent variables $q_\eta(z|\boldsymbol{\Phi},\gamma)$. The generator network approximates the posterior for the generative model $p_\theta(\gamma|\boldsymbol{\Phi},z)$ combining the molecular representation $\boldsymbol{\Phi}$ and the latent variables z. The discriminator network T_ψ classifies the samples from the generative model distribution $p_\theta(\boldsymbol{\Phi},\gamma)$ and the measured data distribution $q(\boldsymbol{\Phi},\gamma)$.

Moreover, we employed a one-dimensional latent space with a standard normal prior distribution, $p(z) \sim \mathcal{N}(0,1)$. During model training, the entropy regularisation and the residual penalty parameters are fixed at $\beta = 1.5$ and $\lambda = 0.5$, respectively. The model is trained for 30,000 stochastic gradient descent steps using the Adam optimiser [54] and a 1×10^{-3} learning rate. Moreover, a multistep reduction schedule is applied to the learning rate for the encoder and generator networks. The conditional generative model is implemented in PyTorch [55], and computations are performed on a single NVIDIA GeForce RTX 4090 GPU card.

Predictive model. Once trained, we have a probabilistic conditional generative model $\gamma = f_\theta(\boldsymbol{\Phi}, z)$ that maps the physicochemical properties of fuels from their molecular structures. The statistical moments of γ can be characterised by sampling latent variables from the prior distribution $p(z)$. In particular, the low-order statistics can be efficiently computed via Monte Carlo sampling. The mean and variance of the predictive distribution for a novel fuel mixture $\boldsymbol{\Phi}(\boldsymbol{\omega}^*)$ are computed as

$$\mu_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}^*)) = \mathbb{E}[\gamma|\boldsymbol{\Phi}(\boldsymbol{\omega}^*), z] \approx \frac{1}{N_s}\sum_{i=1}^{N_s} f_\theta(\boldsymbol{\Phi}(\boldsymbol{\omega}^*), z_i) \tag{5}$$

$$\sigma^2_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}^*)) = \mathbb{V}\mathrm{ar}[\gamma|\boldsymbol{\Phi}(\boldsymbol{\omega}^*), z] \approx \frac{1}{N_s}\sum_{i=1}^{N_s} \left[f_\theta(\boldsymbol{\Phi}(\boldsymbol{\omega}^*), z_i) - \mu_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}^*))\right]^2 \tag{6}$$

where $z_i \sim p(z)$, $i = 1, \ldots, N_s$, and N_s corresponds to the total number of samples from the latent variables prior $p(z)$ propagating through the generator network.

2.3. Inverse design

In Section 2.2, we developed a deep learning generative model for characterising the dependence of physicochemical properties on the fuel molecular composition. Here, we introduce a formal framework for designing new chemical compositions that will lead to fuels achieving the required performance indicators.

Given the stochastic nature of the fuel design approach, the inverse framework can be formulated as an optimisation problem that seeks to find the optimal compositions that maximise the probability of achieving the target properties:

$$\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}\in\mathbb{R}^n} p(\gamma|\boldsymbol{\Phi}(\boldsymbol{\omega})) \tag{7}$$

where, resorting to the previous section, $p(\gamma|\boldsymbol{\Phi}(\boldsymbol{\omega})) = \int f_\theta(\boldsymbol{\Phi}(\boldsymbol{\omega}), z)p(z)dz$. Therefore, the above objective function and its derivatives can be computed using Monte Carlo sampling, exploring the easiness of computation provided by the pre-trained generator network. However, estimating the model coefficients using Bayesian optimisation can be a computational burden since we are dealing with a statistical model in a high-dimensional space (a large number of model coefficients) that may require a significant amount of Markov chain Monte Carlo samples to converge. Therefore, we recover a more straightforward optimisation framework.

Here, the aim is to find an ensemble of optimal fuel compositions fulfilling design requirements. Since we expect to obtain simple fuel mixtures, sparse in the high-dimensional feature space, we will regress the sparse vector of model coefficients $\boldsymbol{\omega}$ employing the least absolute shrinkage and selection operator method [56], which is an l_1-regularised regression that encourages sparsity of model coefficients $\boldsymbol{\omega}$.

We now turn to a specific formulation within the above mathematical framework. Due to the nature of the performance indicator, we employ a minimisation instead of a maximisation to seek a fuel composition that meets (on average) a value for the physicochemical property $\bar{\gamma}$ as a target of the inverse design. Indeed, as γ is characterised as a random variable by the generative model, we employ the expected value of the predictive distribution, $\mu_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}))$, defined in Eq. (5) and introduce the following optimisation problem:

$$\begin{aligned} &\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}\in\mathbb{R}^n} \|\bar{\gamma} - \mu_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}))\|_2^2 + \xi\|\boldsymbol{\omega}\|_1 \\ &\text{subject to: } \mathbf{c}(\boldsymbol{\omega}) = 1, \\ &\boldsymbol{\omega}^{lb} \le \boldsymbol{\omega} \le \boldsymbol{\omega}^{ub}. \end{aligned} \tag{8}$$

where ξ is a regularisation parameter that prescribes the amount of sparsity constraint in the model coefficients, resulting in fuel mixtures with few components. We specify the values of $\xi \in [0.1, 0.5, 1.0]$, which allows us to generate fuel mixtures with different numbers of compounds. In addition, we use a norm for the vectors above to include the possibility of optimising the composition, aiming at multiple different properties as targets.

Moreover, we are dealing with a constrained minimisation problem, where $\mathbf{c}(\boldsymbol{\omega})$ is a linear function that constrains the volumetric weight of the fuel compositions to be equal to 1, i.e., $\sum_{i=1}^{n} \omega_i = 1$. Moreover, $\boldsymbol{\omega}^{lb}$ and $\boldsymbol{\omega}^{ub}$ are the lower and upper bounds constraints for the composition vector, i.e., $0 \le \boldsymbol{\omega} \le 1$. The sequential least squares programming (SLSQP) method is used in the present work, where the objective function is minimised using a quadratic approximation, and the constraints are linearised and handled via Lagrange multipliers. The optimisation algorithm is implemented via the SciPy package for Python [57], in which the constraint functions are defined as dictionaries. Moreover, we use the automatic differentiation module (`torch.autograd`) from PyTorch [55] to efficiently estimate the Jacobian matrix $(\partial f_\theta(\boldsymbol{\Phi}(\boldsymbol{\omega}), z)/\partial\boldsymbol{\omega})$ to drive the search space.

In addition, the inverse design consists of a highly non-convex optimisation process, which means there is no optimal minimum, i.e., we cannot guarantee that $\boldsymbol{\omega}^*$ is an optimal fuel mixture. Therefore, we run the optimisation framework j times, with different initialisation of the mixture vector $\boldsymbol{\omega}_0$ using a Dirichlet distribution with concentration parameters equal to 1.0, to find an ensemble of optimal mixture compositions such as the mean squared error is lower than a threshold, $\|\gamma - \mu_\gamma(\boldsymbol{\Phi}(\boldsymbol{\omega}))\|_2^2 < \epsilon$.

3. Databases

The databases play a critical and unique role in this study, providing reliable information that can be used to construct predictive models and establish a compositional mapping between fuel properties and component chemical structures using AI tools. Here, key physical properties for fuel utilisation are curated to build robust probabilistic predictive models for the physicochemical properties of fuels. These predictive models can map the fuel properties from their chemical structures, enabling their integration into a decision-making fuel design framework that explores sustainable fuels with specific requirements.

In terms of diesel fuel utilisation, we curated two key properties focused on maximising efficiency and performance while improving sustainability (combustion emissions): cetane number (CN), also called derived cetane number D(CN), and the yielding sooting index (YSI). The formers are key indicators of the ignition quality of diesel-like fuels. The main difference is that CN is usually measured on a Cooperative Fuel Research (CFR) engine, while D(CN) is measured on an ignition quality tester (IQT) using a constant-volume combustion chamber (CVCC) [35]. The YSI measures the tendency of fuels to produce soot particulates, which is one of the major issues for diesel utilisation in mobility applications.

The databases for physicochemical properties (CN and YSI) were compiled from various literature sources for this study. Cetane numbers for pure compounds were curated from the report by Murphy et al. [35]. Here, the database consists of experimental CNs of hydrocarbons divided into n-alkanes, iso-alkanes, cycloalkanes, alkenes, and aromatics, and oxygenates divided into ketones, ethers, esters, acids, and furans. Also, additional data for OME_x is collected from Pélerin et al. [58]. OME_x is a cleaner alternative fuel to compression-ignition diesel engines. In particular, OME_x is a class of dimethyl ether (DME) derivatives that can be produced from a range of waste feedstocks and biomass, thereby preventing new fossil carbon from entering the supply chain [59]. In terms of numbers, the final database consists of 708 measurements for a total of 475 different compounds. Fig. 3 shows an overview of the CN database for pure compounds, including the percentage of the molecular classes used in the present work. The pure compound CNs are measured using six different methods, including ASTM D613, ASTM D6890, ASTM D7170, other ignition delay method, blend method, and an unknown method [35]. It is worth mentioning that these methods present different levels of fidelity. The latter three methods provide lower-fidelity measurements. These methods are less accurate and may be influenced by varying methodologies and uncertain correlations [35]. Data collected from ASTM methods may be considered as high-fidelity measurements since these methods are well-documented and consistently implemented. However, since the measurements were made by different laboratories on different samples with varying purities of the compounds, the CN values are prone to bearing uncertainties. Fig. 4 compares CN measurements for randomly selected compounds. The average absolute difference between the CN measured by the ASTM D613 and the other methods may vary significantly, from 2.5 to 15.7 [60].

Moreover, we extend the database by adding a series of CN measurements for mixtures. Here, the mixture database consists of 624 mixtures. In particular, the database is constructed using the 572 mixtures considered by Creton et al. [61]. The mixtures have 2 to 9 compounds, where 96% of the compounds are hydrocarbons and 4% are oxygenates. Moreover, the collection of CN of various biodiesel mixtures [62] was added to the mixture database. The complete mixture database containing CN measurements for the 624 mixtures is available in the *Supplementary Information*.

Yielding Sooting Index measures the chemical propensity of pure compounds or fuel mixtures to produce soot particles in a combustion environment. The sooting tendency of fuels is a valuable physicochemical property for predicting harmful emissions from practical combustion engines. The present work uses the experimentally measured YSI database [63]. The database is constructed using two mutually-incompatible sub-databases: a *low-scale* database for low-sooting tendency pure compounds (alkanes, cycloalkanes, and oxygenates), and a *high-scale* database for high-sooting tendency pure compounds (aromatics). A unified sooting tendency database with a substantial number of pure compounds (≥400) was produced [64]. In particular, a colour-ratio pyrometry diagnostic was implemented to measure the compounds from both databases and merge them into a single *unified scale*. However, such a diagnostic leads to unavoidable uncertainties in the YSI database, returning experimental uncertainties greater than 2 to around 100 YSI units. Fig. 5 provides an overview of the YSI database, including the experimental error associated with different molecular classes. In addition, we leverage the YSI database by collecting measurements for mixtures [33]. YSI measurements for mixtures are scarce and only 40 measurements are available representing gasoline, diesel, and surrogate fuels. The measurements were converted to a *unified scale* using the orthogonal distance regression model [64].

The CN and YSI databases are pre-processed using the standard scaler, which removes the mean and divides the values by the standard deviation. Next, 80% of data points are selected randomly to train the ML model. The remaining 20% are used to test the ability of the proposed model to be generalised to unseen fuel molecules. Further database information is available at *Supplementary Information*.

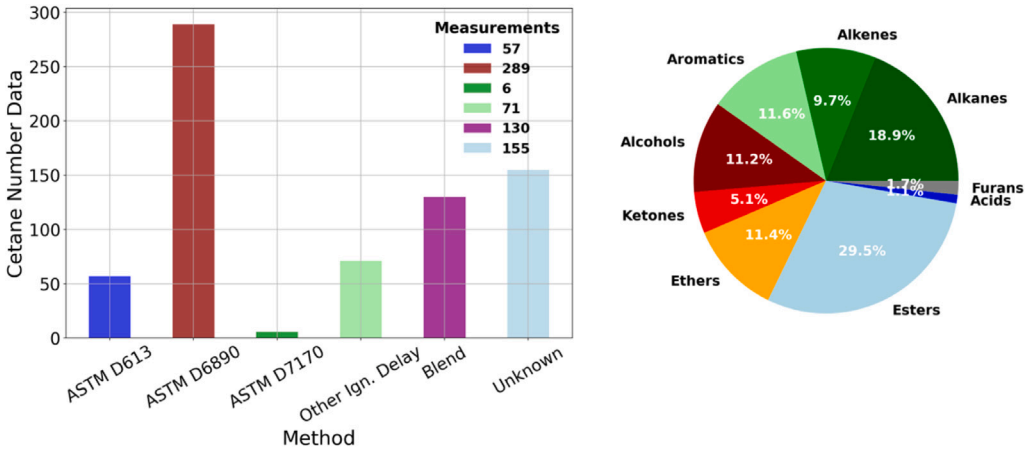


Fig. 3. Distribution of the cetane number in the database. Hydrocarbons represent approximately 40.2% of the compounds and oxygenates 59.8%. The number of measurements for each method in the entire database is provided in the legend of the figure.

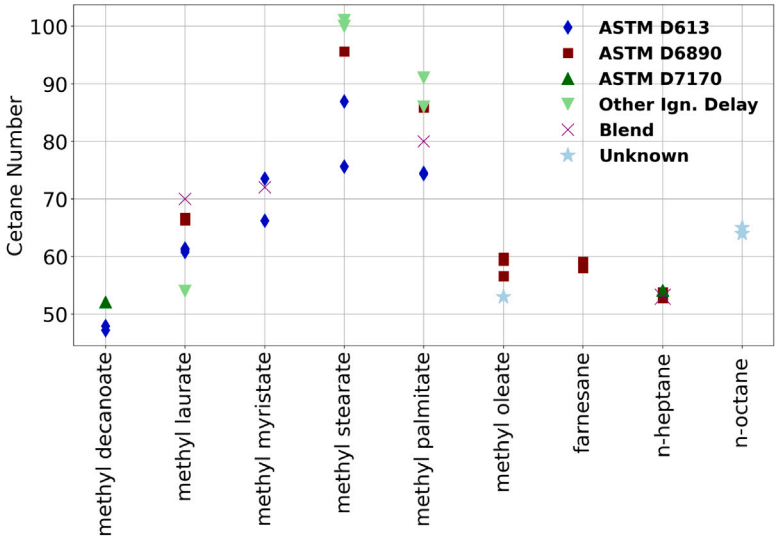


Fig. 4. Comparison of CN measurements from different methods for randomly selected pure compounds.

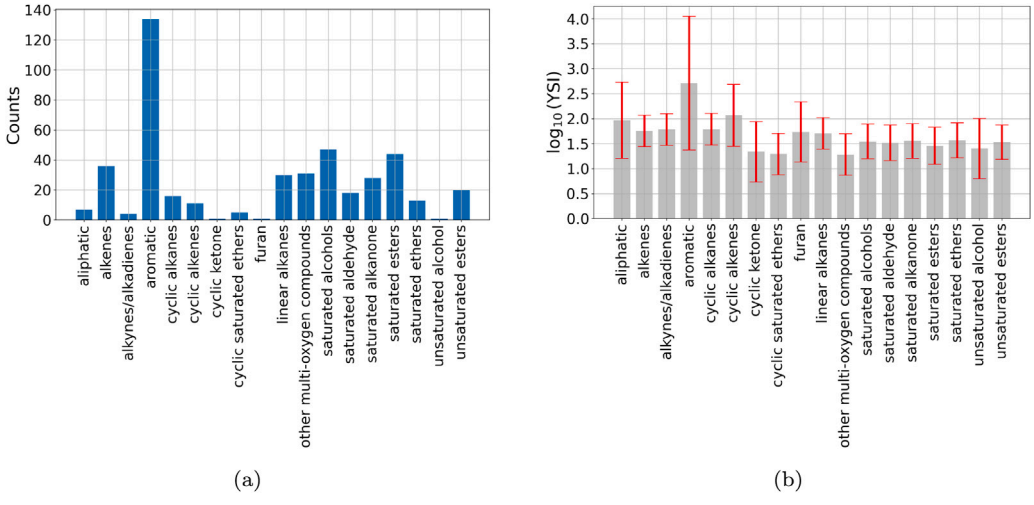


Fig. 5. Distribution of the yielding sooting index database on (a), and (b) average YSI values on a logarithmic scale, including the average error associated with the molecular classes.

4. Results and discussions

In this section, we report the robustness of the proposed fuel design methodology. Our analysis is organised as follows: We first assess the performance of the conditional generative model to build predictive models of the physicochemical properties with quantified uncertainty (see Section 2.2). Then, this model is embedded in a robust optimisation framework that leads to fuel composition attaining the required performance criteria. In particular, we employ the proposed fuel design methodology to screen new fuel compositions that mimic diesel-like fuel, which has great relevance to heavy-duty transportation. Nonetheless, the proposed pathways can be extended to design alternatives to any derived fossil fuels, such as gasoline or aviation fuels, paving the way for the defossilisation of the transportation sector. The code and data accompanying this manuscript are made publicly available at https://github.com/RodolfosmFreitas/FuelDesign.

In this work, the quality of the conditional probabilistic model is evaluated using the coefficient of determination (R^2-score) and normalised mean absolute error (MAE). The definitions are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\gamma_i - \hat{\gamma}_i)^2}{\sum_{i=1}^{N}(\gamma_i - \overline{\gamma})^2} \tag{9}$$

$$\text{nMAE} = \frac{\sum_{i=1}^{N}\left|\gamma_i - \hat{\gamma}_i\right| / N}{\overline{\gamma}} \tag{10}$$

where N is the number of samples (molecules), γ_i is the measured physicochemical property, $\hat{\gamma}_i$ is predicted value, and $\overline{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i$ is the mean over the samples.

4.1. Performance of the deep generative model

Molecular representation. Here, to evaluate the impact of the molecular representation on the predictability of the conditional generative model, four different representations were chosen: Descriptors (molecular descriptors), Extended-Connectivity Fingerprints (ECFPs), Molecule-to-Vector (Mol2Vec) fingerprints, and Sequence-to-Sequence (Seq2Seq) fingerprints. Fig. 6 illustrates the convergence history of accuracy metrics on the test data throughout the training process iterations for the four representations. The conditional generative model is trained on a NVIDIA GeForce RTX 4090 GPU, requiring approximately 60–150 s to train for 30,000 iterations. It is observed that the accuracy metrics begin to stabilise after 15,000 iterations for all molecular representations. The figure demonstrates that the Mol2Vec representation outperforms other representations. Molecular descriptors showed slightly lower performance compared to Mol2Vec. ECFP and Seq2Seq depicted unsatisfactory performance. Furthermore, Fig. 7 shows parity plots between the measured and the predicted CN for the four molecular representations used to build the generative models. The mean values of the generative models are used to construct the parity plots. As we can see, the predictions of the generative model using Mol2Vec representation lie near the diagonal line, suggesting that the model captures the relationship between the molecular structure and CNs satisfactorily well. The other molecular representations present worse predictions. Such behaviour is further confirmed in the performance metrics shown in Figs. 7(e)–(f). We can observe that the Descriptors, Seq2Seq, and Mol2Vec models return high scores ($R^2 \geq 0.95$) for the train data. However, the Descriptor model yields unsatisfactory predictions on the test data ($R^2 < 0.9$). Also, the results indicate that the ECFP and Seq2Seq models are overfitting, typical for machine learning models developed on small datasets. Moreover, we can verify that the Mol2Vec-based model yields good predictions on the test data ($R^2 \geq 0.9$). Therefore, we top-ranked the Mol2Vec model for predicting the physicochemical properties of fuels since the aim is to predict the physicochemical properties of unseen fuel molecules satisfactorily.

Comparison with state-of-the-art models. To verify the performance of the proposed model against state-of-the-art predictive models of the physicochemical properties of the fuel, we compare the predictions of the present model with other machine learning models available in the literature. In particular, the Graph Neural Network (GNN) used to predict fuel ignition quality, including derived cetane number D(CN) [27], is employed. Specifically, GNN represents the fuel molecular structure as graphs and learns the physicochemical properties using the molecular graphs as inputs. Such an approach eliminates the need for a molecular representation since the molecular characteristics are learned through graph convolutions, and the molecular graphs are directly mapped to the properties. We do not claim this represents a fair comparison, as the models were built on different train/test datasets. Still, this analysis can be used to demonstrate the performance of the current methodology for building accurate models. Table 1 compares the CN predictions of fuel compounds representing different molecular classes with different molecular structures made by the GNN model, predicted by the generative model, and the measured values. As we can see, the proposed generative model outperforms the state-of-the-art GNN model, returning an MAE of 0.06.

Table 1
Comparison between the CN predictions made by the GNN model [27] and our Mol2Vec-based generative model.

Pure compounds	Measured [35]	GNN model[a]	Generative model[b]
1,2-dimethylbenzene	8.3	7.4	8.3
furan	7.0	8.2	6.98
methyl erucate	74.2	75.9	74.12
2-heptanol	25.0	24.1	24.96
4-ethyl guaiacol	19.6	17.4	19.64
1,3,5-triisopropylbenzene	2.8	11.5	2.99
ocimene	28	20.1	28.09
1,4-dimethylbenzene	−4, 6.2	6.8	6.27
6-undecanone	49.0	59.2	49.04
4-nonanone	43.0	41.3	43.06
δ-undecalactone	48.6	47.1	48.66
4-methoxybenzaldehyde	25.8	12.5	25.83
geraniol	16.5, 22.0	19.4	19.2
dodecyl vinyl ether	101.7	96.0	101.75
mean absolute error (MAE)	–	**4.05**	**0.06**

[a] The GNN model was built using the average values, although multiple values have been reported for pure compounds.
[b] Expected values from the generative model.

Uncertainty Quantification. The generative model allows rigorous and robust quantification of the predictive uncertainty of physicochemical properties, providing the confidence needed for informed decision-making in the fuel design methodology [26]. Fig. 8 shows predictions of the CNs for different pure compounds with a 95% confidence interval ($\mu \pm 2\sigma$). It is worth noting that the pure compounds represent different molecular classes found in diesel blends and the ability to accurately predict the ignition quality of these fuels is important for the fuel design framework. Moreover, the figure illustrates the effects of the isomeric structures on the physicochemical properties of fuels. We can observe the generative model yields robust predictions with uncertainty bounds that capture the CN for different chemical structures with multiple measurements. At this point, it is worth mentioning that the generative model encompasses uncertainties due to noisy data and epistemic uncertainties resulting from the small amount of data used in the training process. Moreover, we can observe that the predictive uncertainty reaches lower values for the compounds in which the error of the model predictions is also smaller, as shown in Fig. 8(b). Such behaviour indicates that the generative model can provide calibrated and conservative uncertainties. We also report the marginal probability distribution of the estimated uncertainty and the predictive error (Fig. 8(c)).

Transfer learning. Although the YSI database is derived from different sources compared to the CN database, we believe that these physicochemical properties may contain some intrinsic relationships and knowledge learned from CN can be transferred to YSI, in a transfer

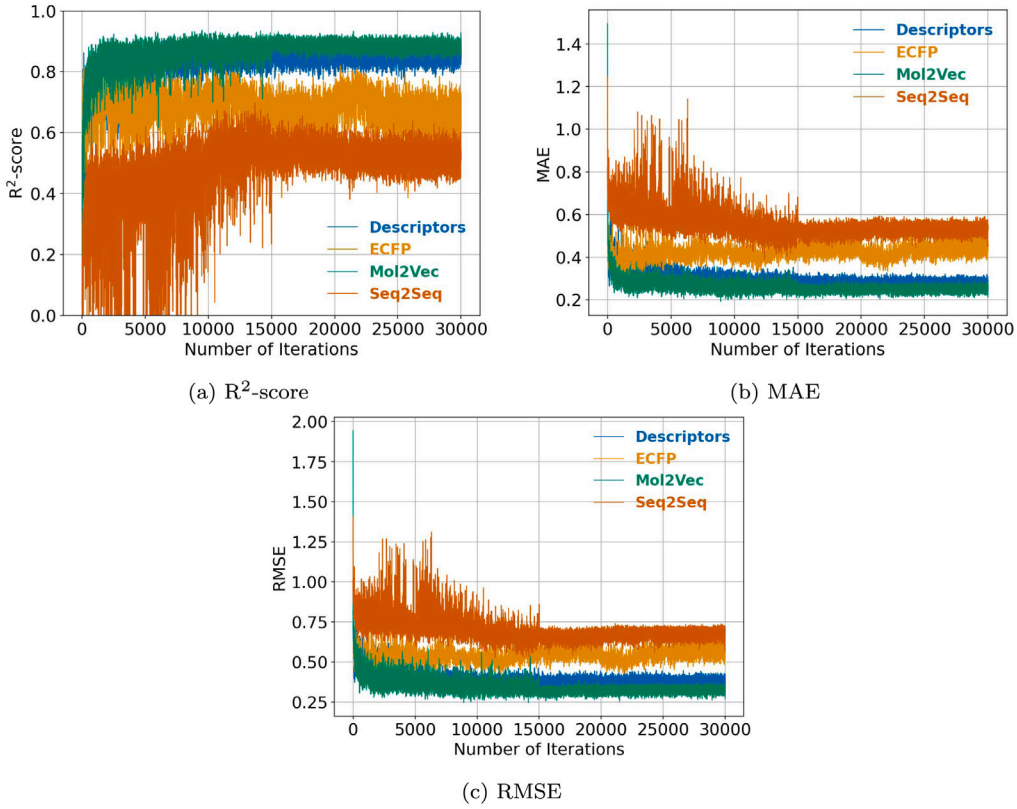


Fig. 6. Convergence history of accuracy metrics with the number of iterations for the conditional generative model using four different molecular representations. The metrics are calculated based on the test data. The root-mean-square error (RMSE) is a metric frequently used to monitor the convergence of both training and test errors during the training process. RMSE = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$.

learning perspective [65]. In particular, the parameters of the pre-trained model for CN are used to initialise the parameters of the YSI model. Fig. 9(a) shows the convergence history of the training process as a function of the number of iterations. We can verify that after a few iterations of the gradient descent optimiser, the predictive model for YSI achieves convergence, and only 1000 iterations are required for a good performance. This is further confirmed by the parity plot between the measured and the predicted YSI, returning high-performance scores (R^2=0.948) in the test set, as shown in Fig. 9(b). Additionally, the generative model returns robust predictions with uncertainty bounds able to capture the YSI of different molecular classes (see Fig. 9(c)). The performance of the proposed model compared with a multi-layer perceptron (MLP) model available in the literature [30] is presented in Table 2. Our proposed generative model outperforms the MLP model, yielding an MAE equal to 2.61. Transfer learning is particularly relevant when dealing with a small data regime, but it also greatly improves learning performance by avoiding costly training efforts.

Predicting physicochemical properties of fuel mixtures. Our findings showed that the conditional generative model yields robust predictions of pure fuel compounds. Nonetheless, from a fuel design perspective aiming to develop fuel mixtures with desired properties, it is paramount that the proposed model can be generalised to predict the physicochemical properties of complex fuel blends. Pushing towards that goal, we enriched the databases by including physicochemical measurements of fuel blends (see details in Section 3). This allows the model to be generalised to predict the properties of fuel mixtures and helps screen fuel compositions on demand based on multi-objective optimisation in an inverse fuel design approach. Fig. 10 presents the parity plot of measured and predicted properties over the test set. We can see that for both properties, CN and YSI, the conditional generative model can yield accurate predictions, returning an R^2-score higher than 0.9. In the present study, we consider 0.9 a good target. Here, it is worth mentioning that the molecular representation of a fuel mixture is assumed to be a linear combination of the representation of pure compounds weighted by volumetric composition. To clarify, we do not assert that the physicochemical properties of fuel mixtures follow a volumetric weighting of their compounds, as many non-linear effects might exist. However, such a definition has been used to construct predictive models of fuel mixtures [33,61]. Thus, due to the lack of better alternatives to describe the molecular representation of fuel mixtures, such an approximation was used in the present study.

Table 2
Comparison between the YSI predictions made by the MLP model [30] and our Mol2Vec-based generative model.

Fuel	Measured YSI	MLP[a,b]	Generative model[c]
cyclohexene	45.6	48.81 ± 9.11	54.25
2,4,4-trimethyl-1-pentene	68.5	74.06 ± 6.97	68.34
2-methyl-2-butene	43.5	33.10 ± 4.90	44.51
methylcyclopentane	50.3	42.90 ± 10.45	48.33
1-methylcyclopentene	96.5	63.66 ± 11.13	96.47
2,4,4-trimethyl-2-pentene	89.3	64.02 ± 8.02	86.64
cycloheptene	78.847	46.24 ± 8.64	70.35
2,3,4-trimethyl-2-pentene	87.0	71.87 ± 8.47	87.24
mean absolute error (MAE)	–	21.56	2.61

[a] The YSI values represent the mean ± 95% confidence interval from 25 randomised replications of the MLP model [30].
[b] The MLP model was constructed based purely on the *low-scale* database (see the details in Databases). For fair analysis, the accuracy metric of the MLP model was computed based on the YSI values from *low-scale* database.
[c] Expected values from the generative model.

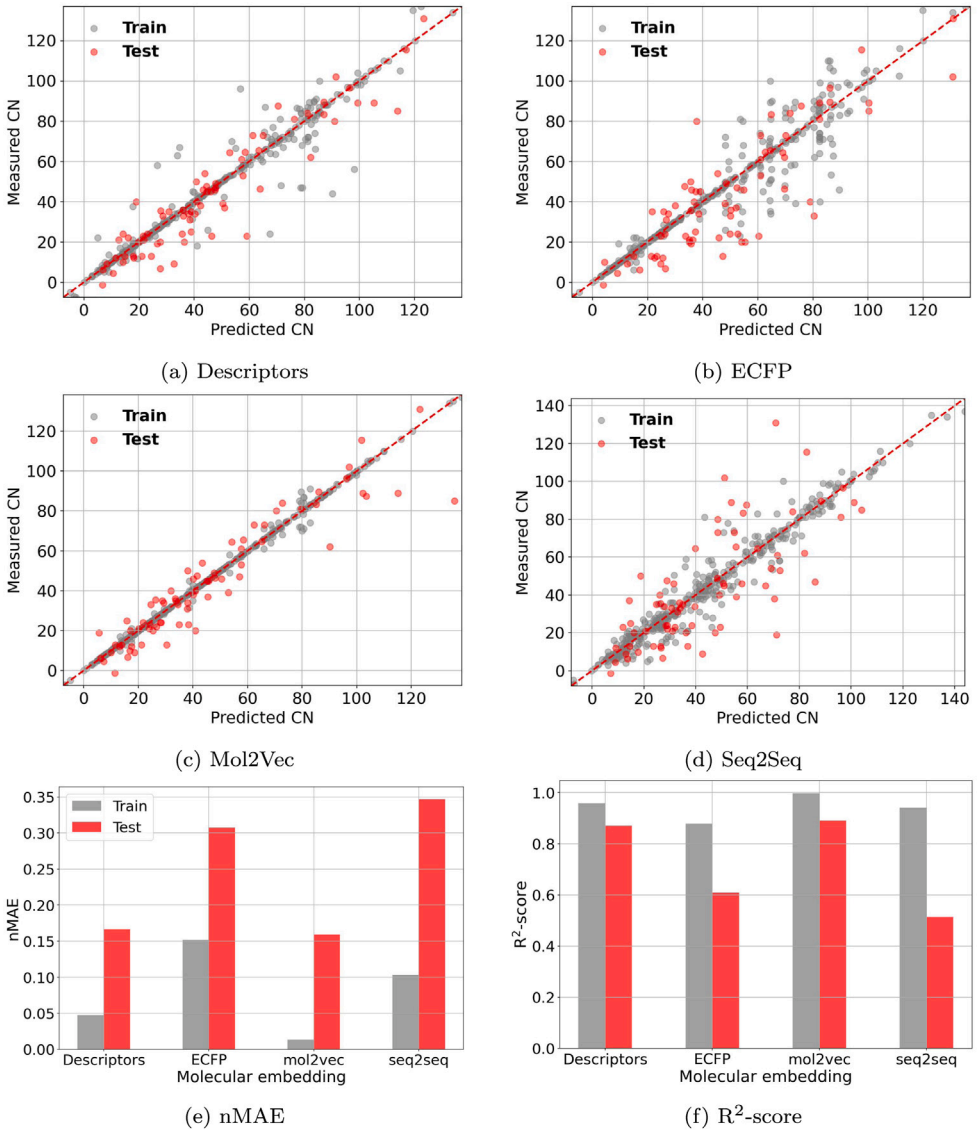


Fig. 7. Impact of the molecular representation on the predictability of the generative model. Parity plots between measured and predicted cetane number for (a) Descriptors, (b) ECFP, (c) Mol2Vec and (d) Seq2Seq molecular representations. The mean values of the generative models are used to construct the parity plots. (e) and (f) show the accuracy metrics for the models: Normalised mean absolute error (nMAE) and coefficient of determination (R^2-score).

4.2. Inverse fuel design

In this section, we leverage the conditional generative model in an inverse design framework introduced in Section 2.3. The aim is to find possible fuel mixtures matching target combustion characteristics for sustainable fuel utilisation. In particular, the expected values of the physicochemical properties are targeted, since we are dealing with a probabilistic model. Specifically, the inverse framework was performed for the China stage VI #0 diesel fuel as the target, with CN=56.5 [66], a typical diesel fuel designed aimed at the Chinese national stage VI emission standard. It is worth noting that such a high-cetane fuel has been used to illustrate the fuel design framework; however, the optimisation framework can be extended to any other cetane requirements. Moreover, to our knowledge, there is no indication of an optimal YSI for diesel fuels. In particular, traditional diesel fuels have an approximate value of ~235. Although fuel design is a multi-task constrained optimisation, we only consider CN as the target and leverage the trained predictive model for YSI to rank the designed fuel mixtures from the lowest to the highest sooting tendency based on the mean value produced by the generative model. The details regarding constraints, bounds, and search chemical space are elaborated below.

Fuel palette. A critical component in the fuel design formulation is to define the chemical space search, also called the fuel palette [67]. Liquid fuel compounds are very complex, and each compound in the mixture may affect engine performance and emissions. Moreover, the fuel palette must be able to represent the types of compounds found in the target fuels. In this context, Farrell et al. [68] systematically summarised the importance of each compound in the design of diesel surrogate fuels. For instance, they included n-decane, n-hexadecane, iso-octane, methylcyclohexane, toluene, heptamethylnonane, n-decylbenzene, and 1-methylnaphthalene as important diesel compounds. In the current study, we adopt a fuel palette proposed by Qian et al. [69], which comprises compounds representing the different hydrocarbon classes. In particular, they chose 14 hydrocarbon compounds based on two criteria:

- The pure compounds are available in high purity that can be fed on engine test conditions to evaluate the impact of different compounds on performance and emissions.
- Availability of chemical oxidation/pyrolysis mechanisms of compounds for subsequent application to combustion modelling.

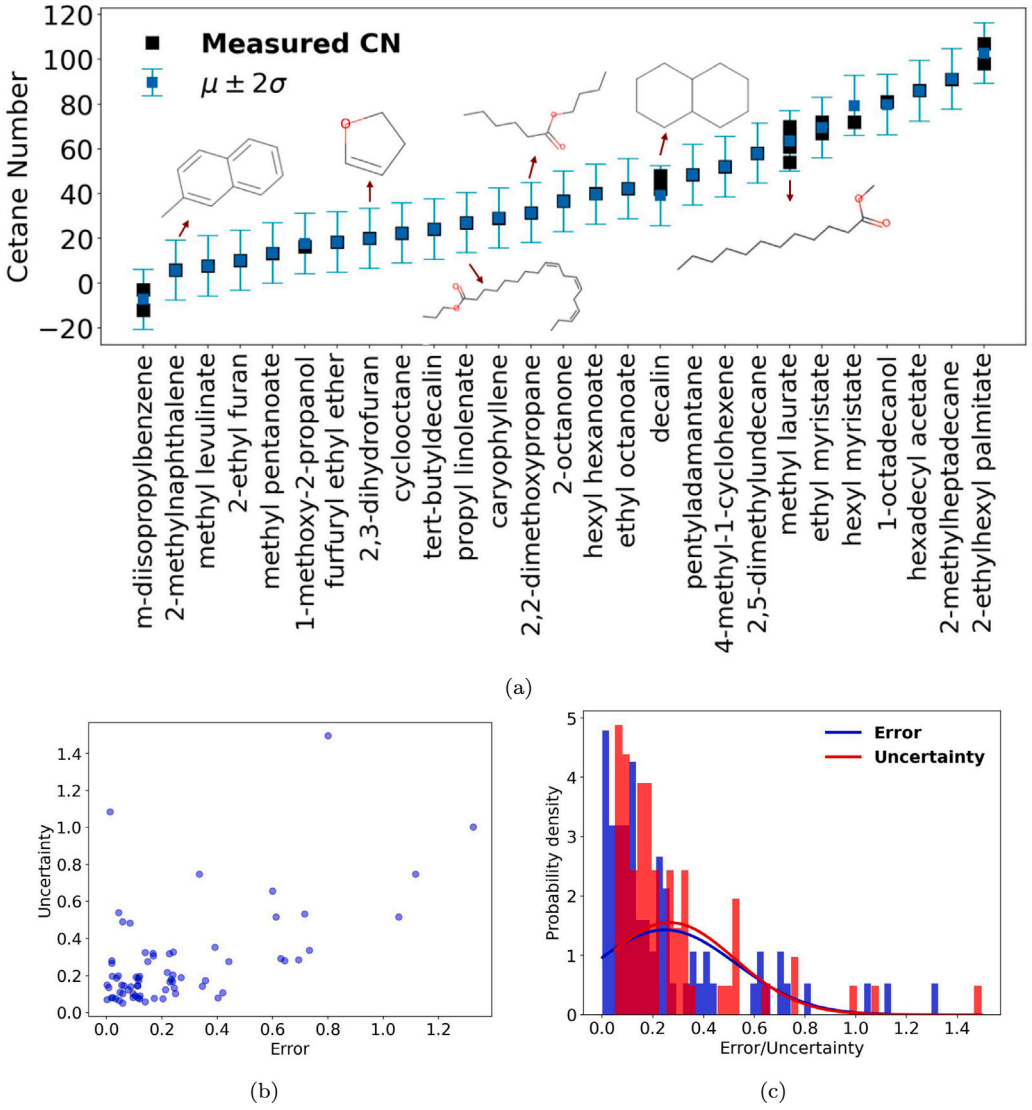


Fig. 8. Uncertainty quantification. (a) Conditional generative predictive mean and confidence interval, (b) Predictive error ($|\bar{\gamma} - \mu_{\gamma}(\Phi)|/|\bar{\gamma}|$) versus uncertainty quantification ($|\sigma_{\gamma}(\Phi)|/|\bar{\gamma}|$), and (c) Marginal distributions of the predictive error (blue) and predictive uncertainty (red). The predictive error and uncertainty are computed using the test set.

In addition, alcohols [70,71], OME_x [14,66], and FAMEs [72,73] are currently the primary oxygenated fuels blended into diesel to mitigate harmful emissions. Therefore, molecules representing these oxygenate classes were included in the surrogate palette. The pure compounds adopted in the inverse fuel design are listed in Table S1 in *Supplementary Information*.

Targets and constraints. To design an ensemble of optimal diesel mixtures we adopted the following targets and constraints:

- The reference target fuel adopted is based on the cetane number of China VI **#0** diesel [66], $\bar{\gamma} = 56.5$.
- Typically, diesel fuels are composed of 50%–65% alkanes (mostly n-alkanes), 20%–30% cycloalkanes and 10%–30% aromatics [68]. Therefore, we screened fuel mixtures respecting this proportion.
- In general, oxygenated diesel blends are composed of around 10%, 20%, and 30% of oxygenated compounds [74,75]. Hence, we selected fuel blends whose oxygenated proportion is less than 30%.
- The fuel mixtures are ranked based on the average sooting tendency $\mu_{YSI}(\Phi)$, thanks to the predictive model built in the previous section.

The present study screened four diesel blend categories in the inverse fuel design framework. First, diesel blends containing hydrocarbons divided into n-alkanes, iso-alkanes, cycloalkanes, and aromatics were designed. This is a natural choice since diesel fuels are primarily composed of hydrocarbons. Afterwards, fuel blends containing additives were developed. In particular, we consider possible additives for diesel fuels: alcohols, ethers (basically OME_x including dimethyl ether OME_0), and FAMEs. This approach allows uncovering effective additives for alternative diesel fuels.

From our results, a total of 157 diesel blends with 7–19 compounds were designed, where 43 diesel blends are composed of hydrocarbons containing 23%–57% n-alkanes, 3%–40% iso-alkanes, 20%–30% cycloalkanes, and 9%–28% aromatics. In addition, 24 mixtures containing 3%–28% alcohols, 30 mixtures containing 10%–29% OME_x, and 60 mixtures containing 7%–30% FAMEs as additives were reported. Also,

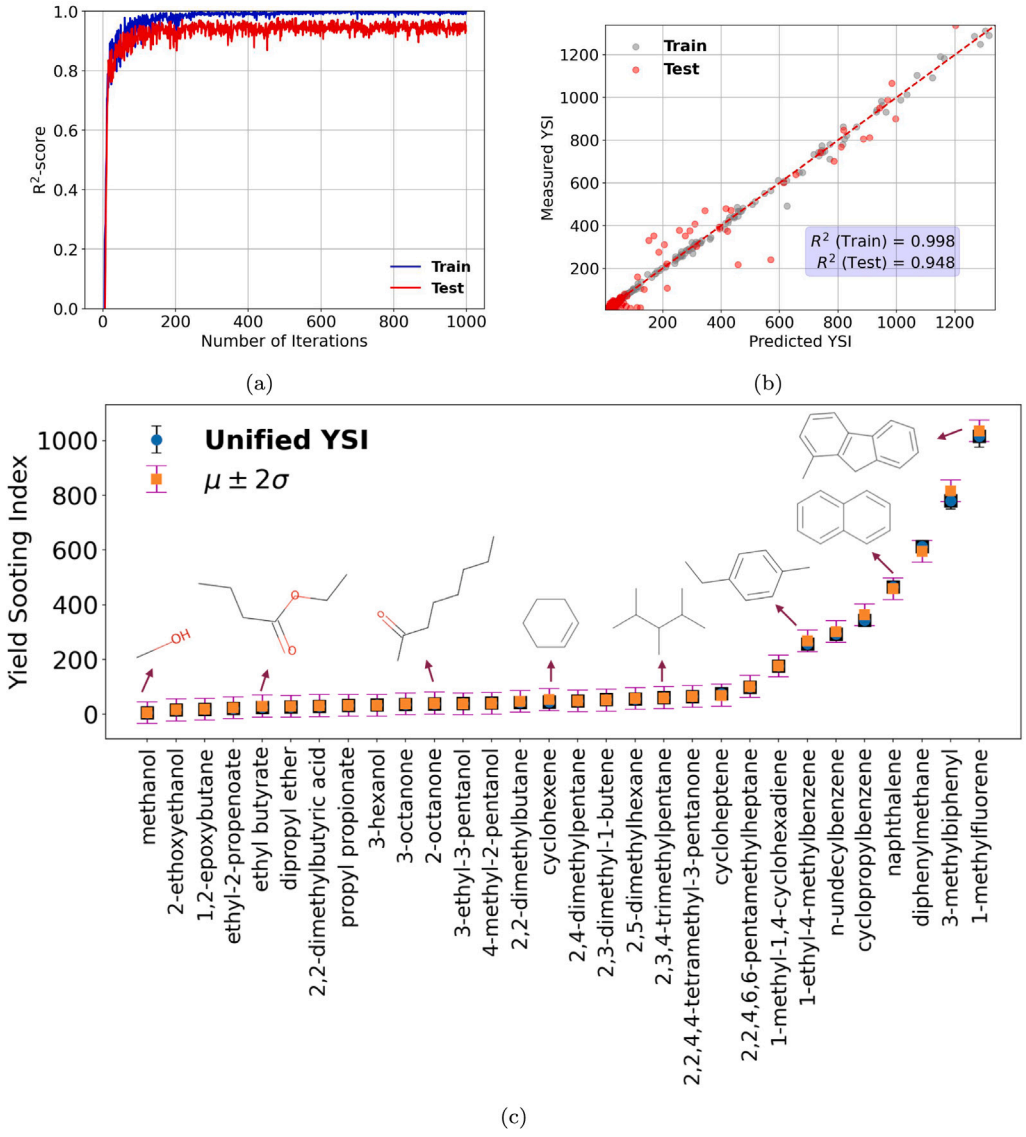


Fig. 9. Performance of the conditional model on predicting sooting tendency. (a) Convergence of the R^2 in the training and test sets, (b) parity plot between measured and predicted YSI, and (c) conditional generative predictive mean and confidence interval for the YSI model.

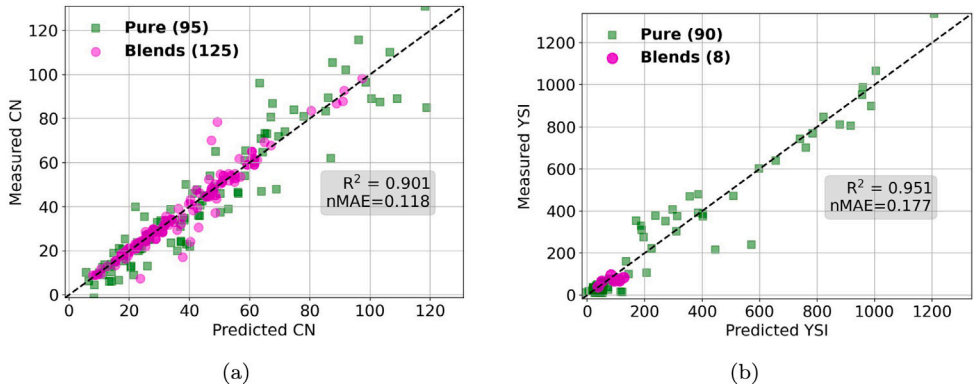


Fig. 10. Performance of the conditional model in predicting physicochemical properties of fuel mixtures. (a) Parity plot between measured and predicted cetane number and (b) parity plot between the measured and predicted sooting tendency. Green squares represent pure compounds and purple circles represent fuel mixtures.

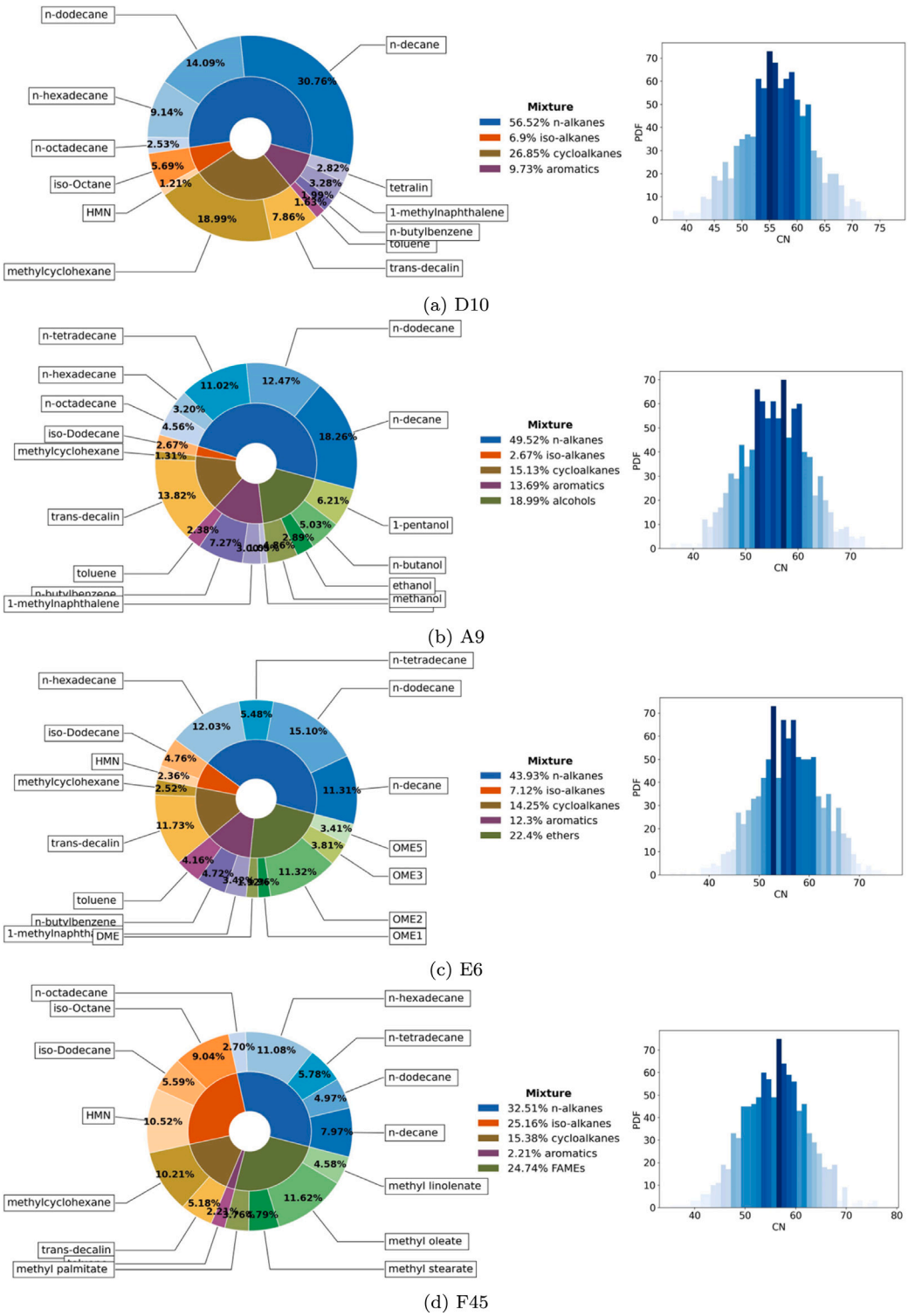


Fig. 11. Pie charts composition and CN marginal distributions of the top-ranked fuel blends based on YSI for the different diesel categories. (a) blend D10 containing only hydrocarbons, (b) blend A9 containing alcohols as additive, (c) blend E6 containing ethers as additive, and (d) blend F45 containing FAMEs as additive.

for all developed blends the density, flash point (FP), boiling point (BP), and lower heating value (LHV) were estimated using linear blending rules from data for pure compounds. Albeit simple, mixing rules are often used to describe fuel properties due to the lack of better alternatives [76]. The physicochemical properties of the four top-ranked blends based on the sooting tendency for the different diesel categories are shown in Table 3. Blends labelled D represent fuels based solely on hydrocarbons, and A, E, and F contain alcohols, ethers, and FAMEs as additives, respectively. Here, we also show the expected values of CN and YSI predicted by the generative model. We can see that the designed blends exhibit properties similar to typical diesel fuels (EN590 and China VI 0#). Moreover, we can verify the ability of the present

Table 3
Physicochemical properties of designed and selected reference diesel fuels.

Fuel	MW [g/mol]	Density[a] [g/cm^3]	FP [°C]	BP [°C]	LHV [MJ/kg]	CN	YSI
EN590[b]	~ 200	0.820–0.845	> 55	~ 170.0	42.6	> 51	~ 235[d]
VI 0#[c]	–	0.820–0.845	> 60	188.0	42.5	> 51	~ 235[d]
D8	154.99	0.796	57.73	193.67	43.11	56.59	59.96
D10	146.23	0.779	50.14	180.56	43.45	56.37	59.02
D12	147.14	0.787	51.94	179.03	43.32	56.48	60.88
D33	156.76	0.805	65.85	200.89	43.05	55.87	61.08
A3	148.49	0.802	58.58	191.92	41.23	56.18	60.07
A9	142.59	0.801	61.02	188.49	40.64	55.73	57.59
A21	152.99	0.801	58.93	196.99	41.17	56.15	60.03
A23	146.60	0.808	61.18	191.84	39.63	55.89	59.72
E2	166.76	0.844	70.68	204.59	38.25	56.59	47.34
E6	153.37	0.839	61.85	187.99	38.26	55.89	44.14
E16	158.62	0.836	59.77	189.83	38.13	56.07	45.03
E24	164.74	0.854	64.59	198.86	37.23	57.09	47.14
F19	176.47	0.783	65.05	191.88	42.84	56.49	63.47
F35	197.96	0.796	85.74	208.82	42.67	55.84	65.42
F36	179.35	0.804	77.77	193.27	42.92	56.10	64.42
F45	198.43	0.795	85.73	195.63	42.28	55.79	60.51

[a] at 20 °C.
[b] Properties collected from Pitsch et al. [15].
[c] Properties collected from www.theicct.org [77].
[d] Approximate value of traditional diesel fuels.

approach to develop high-performance fuels with low emissions, minimising the emissions by more than 72% compared to traditional diesel fuels. It is worth commenting on the potential of OME_x as a renewable drop-in fuel for compression ignition engines [15]. Due to the absence of C–C bonds, OME_x combustion produces far less particulate matter as a cleaner alternative to conventional diesel fuel [14]. Additional information on all designed mixtures is provided in the Supplementary Information.

Fig. 11 shows the pie chart with the detailed composition of the top-ranked mixtures of the different diesel categories. As we can observe, the blends are mainly composed of n-alkanes and iso-alkanes (50%–65%) with lower proportions of cycloalkanes and aromatics. Also, additives account for about 20% in the fuel blends. Furthermore, the marginal distributions of the cetane number are provided for each blend. The results illustrate the ability of the generative model to provide the confidence needed in the fuel design framework, yielding robust predictive models. The marginal distributions of the YSI for the respective blends are provided in Fig. S1 in *Supplementary Information*.

5. Conclusions

We have reported an AI-assisted fuel design methodology from a probabilistic perspective. The proposed methodology comprises a conditional generative model capable of building robust predictive models that are paramount to providing the confidence needed for informed decision-making in the fuel design framework. The conditional model was embedded in an inverse design framework to design alternative diesel fuel blends. The approach's effectiveness was tested by developing new diesel blends based on several hydrocarbon classes and fuel additives, such as alcohols, ethers, and FAMEs. The inverse design was formulated considering CN as the target, and the designed blends were ranked from the lowest to the highest sooting tendency.

Overall, the generative model was revealed to be a powerful tool for accurately predicting fuel physicochemical properties (CN and YSI), returning an R^2-score greater than 0.9. Moreover, the generative model yields calibrated and conservative uncertainties for different pure compounds and fuel blends with multiple measurements, providing the confidence needed for decision-making in the inverse fuel design. From our presented AI-assisted fuel design approach, 157 diesel blends were developed with combustion characteristics similar to typical diesel fuels. We also verified the ability of the AI approach to design high-performance fuels with low emissions, dwelling emissions by more than 72% compared to traditional diesel fuel. In addition, we illustrate the potential for improving the performance of the fuels by including additives in the diesel compositions.

It is worth mentioning that the state-of-the-art AI approach was used to build predictive models for CN and YSI, and such a framework can be extended to any key physicochemical property for fuel design. However, the lack of data representing fuel blends remains the bottleneck of the AI-assisted fuel design approach. A strategy to deal with data scarcity is to enrich the dataset using measurements provided by physics-based numerical models, such as molecular dynamics simulations. Also, open-access data sharing will be of great value in moving towards the generation of digital tools that can assist in the design of alternative fuels.

Furthermore, the approach discussed in this paper opens several avenues for future work. Although the proposed methodology has been used to design new blends, inverse design can also be developed to provide new fuel molecules. In this regard, reactive pathways and energy transfer during the combustion of the new molecules can be investigated from reactive molecular dynamics using deep learning potentials learned from density functional theory data. Moreover, research efforts should focus on constructing data-driven surrogate models of practical combustion engines, this will allow for bridging the gap between property optimisation and combustion performance. Indeed, these data-driven surrogate models enable rapid assessment of fuel behaviour in engines and supply chains, allowing prompt prototyping and validation. Finally, a posteriori assessment of the designed fuel blends is important. Experimental apparatus and numerical modelling can validate the developed fuel blends, providing insights into fuel performance.

CRediT authorship contribution statement

Rodolfo S.M. Freitas: Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Zhihao Xing:** Writing – review & editing, Visualization, Investigation, Data curation. **Fernando A. Rochinha:** Writing – review & editing, Visualization, Supervision, Data curation. **Roger F. Cracknell:** Writing – review & editing, Visualization, Supervision, Investigation, Data curation, Conceptualization. **Daniel Mira:** Writing – review & editing, Visualization, Validation, Supervision, Investigation, Data curation, Conceptualization. **Nader Karimi:** Writing – review & editing, Visualization, Validation, Data curation, Conceptualization. **Xi Jiang:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

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.

Acknowledgement

This study is supported by the UK Physical Sciences Research Council under Grant No. EP/X019551/1.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.adapen.2025.100226.

Data availability

Data will be made available on request.

References

[1] Martin J, Dimanchev E, Neumann A. Carbon abatement costs for renewable fuels in hard-to-abate transport sectors. Adv Appl Energy 2023;12:100156. http://dx.doi.org/10.1016/j.adapen.2023.100156.
[2] Liu Z, Deng Z, Davis SJ, Ciais P. Global carbon emissions in 2023. Nat Rev Earth & Environ 2024;5(4):253–4. http://dx.doi.org/10.1038/s43017-024-00532-2.
[3] World energy outlook 2023. 2023, International Energy Agency, 2023.
[4] Carbon monitor project. 2023, Available from: https://carbonmonitor.org/. [visited on 11 May 2025].
[5] Summary for policymakers: Climate change 2022 - Mitigation of climate change: Working Group III contribution to the sixth assessment report of the intergovernmental panel on climate change. Cambridge University Press; 2022, p. 3–48.
[6] COP29: UNECE advocates for energy transition financing, environmental cooperation, decarbonization of transport, and local climate action, Available from: https://unece.org/climate-change/press/cop29-unece-advocates-energy-transition-financing-environmental-cooperation.
[7] Hutchings G, Davidson M, Atkins P, Collier P, Jackson N, Morton A, et al. Sustainable synthetic carbon based fuels for transport: Policy Briefing. United Kingdom: The Royal Society; 2019.
[8] IEA. The Future of Trucks: Implications for energy and the environment. Paris: IEA; 2017, http://dx.doi.org/10.1787/9789264279452-en.
[9] Biofuel production forecast by product type 2030. In: Statista. 2023, Available from: https://www.statista.com/statistics/1440696/biofuel-production-forecast-by-product-type/. [visited on 12 May 2025].
[10] Jeswani HK, Chilvers A, Azapagic A. Environmental sustainability of biofuels: a review. Proc R Soc A: Math Phys Eng Sci 2020;476(2243):20200351. http://dx.doi.org/10.1098/rspa.2020.0351, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2020.0351.
[11] Ammonia: Zero-carbon fertiliser, fuel and energy store. In: Policy briefing, Royal Society; 2020.
[12] Net Zero Aviation Fuels: resource requirements and environmental impacts policy briefing. The Royal Society; 2023.
[13] Leitner W, Klankermayer J, Pischinger S, Pitsch H, Kohse-Höinghaus K. Advanced biofuels and beyond: Chemistry solutions for propulsion and production. Angew Chem Int Ed 2017;56(20):5412–52. http://dx.doi.org/10.1002/anie.201607257, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201607257.
[14] Styring P, Dowson GRM, Tozer IO. Synthetic fuels based on dimethyl ether as a future non-fossil fuel for road transport from sustainable feedstocks. Front Energy Res 2021;9. http://dx.doi.org/10.3389/fenrg.2021.663331.
[15] Pitsch H, Goeb D, Cai L, Willems W. Potential of oxymethylene ethers as renewable diesel substitute. Prog Energy Combust Sci 2024;104:101173. http://dx.doi.org/10.1016/j.pecs.2024.101173.
[16] Voelker S, Groll N, Bachmann M, Mueller L, Neumann M, Kossioris T, et al. Towards carbon-neutral and clean propulsion in heavy-duty transportation with hydroformylated Fischer–Tropsch fuels. Nat Energy 2024. http://dx.doi.org/10.1038/s41560-024-01581-z.
[17] Ram V, Salkuti SR. An overview of major synthetic fuels. Energies 2023;16(6).
[18] Putrasari Y, Lim O. Dimethyl ether as the next generation fuel to control nitrogen oxides and particulate matter emissions from internal combustion engines: A review. ACS Omega 2022;7(1):32–7.
[19] Holt D. Alternative diesel fuels. In: Progress in technology series, Society of Automotive Engineers; 2004.
[20] Martinovic F, Castoldi L, Deorsola FA. Aftertreatment technologies for diesel engines: An overview of the combined systems. Catalysts 2021;11(6). http://dx.doi.org/10.3390/catal11060653.
[21] Saarikoski S, Järvinen A, Markkula L, Aurela M, Kuittinen N, Hoivala J, et al. Towards zero pollution vehicles by advanced fuels and exhaust aftertreatment technologies. Environ Pollut 2024;347:123665. http://dx.doi.org/10.1016/j.envpol.2024.123665.
[22] Lu X, Han D, Huang Z. Fuel design and management for the control of advanced compression-ignition combustion modes. Prog Energy Combust Sci 2011;37(6):741–83. http://dx.doi.org/10.1016/j.pecs.2011.03.003.
[23] Bessonette PW, Schleyer CH, Duffy KP, Hardy WL, Liechty MP. Effects of fuel property changes on heavy-duty HCCI combustion. SAE Trans 2007;116:242–54, .
[24] Liu F, Shafique M, Luo X. Literature review on life cycle assessment of transportation alternative fuels. Environ Technol Innov 2023;32:103343. http://dx.doi.org/10.1016/j.eti.2023.103343.
[25] Dryer FL. Chemical kinetic and combustion characteristics of transportation fuels. Proc Combust Inst 2015;35(1):117–44. http://dx.doi.org/10.1016/j.proci.2014.09.008, URL https://www.sciencedirect.com/science/article/pii/S1540748914004258.
[26] Sarathy SM, Eraqi BA. Artificial intelligence for novel fuel design. Proc Combust Inst 2024;40(1):105630. http://dx.doi.org/10.1016/j.proci.2024.105630.
[27] Schweidtmann AM, Rittig JG, König A, Grohe M, Mitsos A, Dahmen M. Graph neural networks for prediction of fuel ignition quality. Energy & Fuels 2020;34(9):11395–407. http://dx.doi.org/10.1021/acs.energyfuels.0c01533.
[28] Freitas RS, Jiang X. Descriptors-based machine-learning prediction of cetane number using quantitative structure–property relationship. Energy AI 2024;17:100385. http://dx.doi.org/10.1016/j.egyai.2024.100385.
[29] Bounaceur R, Barthélemy N, Delort N, Herbinet O, Battin-Leclerc F. A multimodal learning model based on a QSPR approach for the estimation of RON, MON and CN, for any C, H, O hydrocarbons. Fuel 2025;381:133438. http://dx.doi.org/10.1016/j.fuel.2024.133438.
[30] St. John PC, Kairys P, Das DD, McEnally CS, Pfefferle LD, Robichaud DJ, et al. A quantitative model for the prediction of sooting tendency from molecular structure. Energy & Fuels 2017;31(9):9983–90. http://dx.doi.org/10.1021/acs.energyfuels.7b00616.
[31] Ahmed Qasem MA, Al-Mutairi EM, Abdul Jameel AG. Smoke point prediction of oxygenated fuels using neural networks. Fuel 2023;332:126026. http://dx.doi.org/10.1016/j.fuel.2022.126026.
[32] Liu R, Liu R, Liu Y, Wang L, Zhang X, Li G. Design of fuel molecules based on variational autoencoder. Fuel 2022;316:123426. http://dx.doi.org/10.1016/j.fuel.2022.123426.
[33] Kuzhagaliyeva N, Horváth S, Williams J, Nicolle A, Sarathy SM. Artificial intelligence-driven design of fuel mixtures. Commun Chem 2022;5(1):111. http://dx.doi.org/10.1038/s42004-022-00722-3.
[34] Parente A, Swaminathan N. Data-driven models and digital twins for sustainable combustion technologies. IScience 2024;27(4). http://dx.doi.org/10.1016/j.isci.2024.109349.
[35] Yanowitz J, Ratcliff MA, McCormick RL, Taylor JD, Murphy MJ. Compendium of experimental cetane numbers. 2017, http://dx.doi.org/10.2172/1345058.
[36] 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.
[37] Freitas RSM, Lima ÁP, Chen C, Rochinha FA, Mira D, Jiang X. Towards predicting liquid fuel physicochemical properties using molecular dynamics guided machine learning models. Fuel 2022;329:125415. http://dx.doi.org/10.1016/j.fuel.2022.125415.
[38] 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.
[39] Perdikaris P, Raissi M, Damianou A, Lawrence ND, Karniadakis GE. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proc R Soc A: Math Phys Eng Sci 2017;473(2198):20160751. http://dx.doi.org/10.1098/rspa.2016.0751.
[40] Boulougouri M, Vandergheynst P, Probst D. Molecular set representation learning. Nat Mach Intell 2024;6(7):754–63. http://dx.doi.org/10.1038/s42256-024-00856-0.
[41] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science 2018;361(6400):360–5. http://dx.doi.org/10.1126/science.aat2663, arXiv:https://www.science.org/doi/pdf/10.1126/science.aat2663.
[42] Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci 1989;29(2):97–101. http://dx.doi.org/10.1021/ci00062a008.
[43] Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminformatics 2018;10(1):4.
[44] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54. http://dx.doi.org/10.1021/ci100050t, PMID: 20426451.
[45] Ramsundar B, Eastman P, Walters P, Pande V, Leswing K, Wu Z. Deep learning for the life sciences. O'Reilly Media; 2019.

[46] Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised machine learning approach with chemical intuition. J Chem Inf Model 2018;58(1):27–35. http://dx.doi.org/10.1021/acs.jcim.7b00616, PMID: 29268609.
[47] https://github.com/samoturk/mol2vec.
[48] Xu Z, Wang S, Zhu F, Huang J. Seq2seq fingerprint: An unsupervised deep molecular embedding for drug discovery. In: Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics. 2017.
[49] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol 2020.
[50] Kingma DP, Welling M. Auto-encoding variational Bayes. 2014, arXiv:1312.6114.
[51] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. 2014, arXiv:1406.2661.
[52] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proceedings of the 30th international conference on neural information processing systems. Red Hook, NY, USA: Curran Associates Inc.; 2016, p. 2234—-2242.
[53] Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A research platform for distributed model selection and training. 2018, arXiv preprint arXiv:1807.05118.
[54] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2017, arXiv:1412.6980.
[55] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems, vol. 32. Curran Associates, Inc.; 2019.
[56] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 2022;58(1):267–88.
[57] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020;17:261–72. http://dx.doi.org/10.1038/s41592-019-0686-2.
[58] 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.
[59] Omari A, Heuser B, Pischinger S. Potential of oxymethylenether-diesel blends for ultra-low emission engines. Fuel 2017;209:232–7. http://dx.doi.org/10.1016/j.fuel.2017.07.107.
[60] Yanowitz J, Ratcliff MA, McCormick RL, Taylor JD, Murphy MJ. Compendium of experimental cetane numbers. 2014, http://dx.doi.org/10.2172/1150177.
[61] Creton B, Brassart N, Herbaut A, Matrat M. Numerical approaches to determine cetane number of hydrocarbons and oxygenated compounds, mixtures, and their blends. Energy & Fuels 2024;38(16):15652–61. http://dx.doi.org/10.1021/acs.energyfuels.4c03007.
[62] Mishra S, Anand K, Mehta PS. Predicting the cetane number of biodiesel fuels from their fatty acid methyl ester composition. Energy & Fuels 2016;30(12):10425–34. http://dx.doi.org/10.1021/acs.energyfuels.6b01343.
[63] McEnally CS, Das DD, Pfefferle LD. Yield sooting index database volume 2: Sooting tendencies of a wide range of fuel compounds on a unified scale. 2017, http://dx.doi.org/10.7910/DVN/7HGFT8, Harvard Dataverse.
[64] Das DD, St. John PC, McEnally CS, Kim S, Pfefferle LD. Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale. Combust Flame 2018;190:349–64. http://dx.doi.org/10.1016/j.combustflame.2017.12.005.
[65] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010;22(10):1345–59. http://dx.doi.org/10.1109/TKDE.2009.191.
[66] Zhao Y, Geng C, Weibo E, Li X, Cheng P, Niu T. Experimental study on the effects of blending PODEn on performance, combustion and emission characteristics of heavy-duty diesel engines meeting China VI emission standard. Sci Rep 2021;11(1):9514. http://dx.doi.org/10.1038/s41598-021-89057-y.
[67] Kukkadapu G, Whitesides R, Wang M, Wagnon SW, Mehl M, Westbrook CK, et al. Development of a diesel surrogate for improved autoignition prediction: Methodology and detailed chemical kinetic modeling. Appl Energy Combust Sci 2023;16:100216. http://dx.doi.org/10.1016/j.jaecs.2023.100216.
[68] Farrell JT, Cernansky NP, Dryer FL, Law CK, Friend DG, Hergart CA, et al. Development of an experimental database and kinetic models for surrogate diesel fuels. In: SAE world congress & exhibition. SAE International; 2007, http://dx.doi.org/10.4271/2007-01-0201.
[69] Qian Y, Yu L, Li Z, Zhang Y, Xu L, Zhou Q, et al. A new methodology for diesel surrogate fuel formulation: Bridging fuel fundamental properties and real engine combustion characteristics. Energy 2018;148:424–47. http://dx.doi.org/10.1016/j.energy.2018.01.181.
[70] Babu V, Murthy KM, Rao GAP. Butanol and pentanol: The promising biofuels for CI engines – A review. Renew Sustain Energy Rev 2017;78:1068–88. http://dx.doi.org/10.1016/j.rser.2017.05.038.
[71] Çelebi Y, Aydın H. An overview on the light alcohol fuels in diesel engines. Fuel 2019;236:890–911. http://dx.doi.org/10.1016/j.fuel.2018.08.138.
[72] Geller DP, Goodrum JW. Effects of specific fatty acid methyl esters on diesel fuel lubricity. Fuel 2004;83(17):2351–6. http://dx.doi.org/10.1016/j.fuel.2004.06.004.
[73] Bukkarapu KR, Krishnasamy A. A critical review on available models to predict engine fuel properties of biodiesel. Renew Sustain Energy Rev 2022;155:111925. http://dx.doi.org/10.1016/j.rser.2021.111925.
[74] Kalil Rahiman M, Santhoshkumar S, Subramaniam D, Avinash A, Pugazhendhi A. Effects of oxygenated fuel pertaining to fuel analysis on diesel engine combustion and emission characteristics. Energy 2022;239:122373. http://dx.doi.org/10.1016/j.energy.2021.122373.
[75] Wang J, Wu F, Xiao J, Shuai S. Oxygenated blend design and its effects on reducing diesel particulate emissions. Fuel 2009;88(10):2037–45. http://dx.doi.org/10.1016/j.fuel.2009.02.045.
[76] Dahmen M, Marquardt W. Model-based formulation of biofuel blends by simultaneous product and pathway design. Energy & Fuels 2017;31(4):4096–121. http://dx.doi.org/10.1021/acs.energyfuels.7b00118.
[77] The International Council on Clean Transportation. Early adoption of China VI vehicle fuel standards in Jing-Jin-Ji and surrounding areas. 2018, Available from: https://theicct.org/publication/early-adoption-of-china-vi-vehicle-fuel-standards-in-jing-jin-ji-and-surrounding-areas/.