Graph Machine Learning for Design of High-Octane Fuels

Jan G. Rittig[a*], Martin Ritzert[b*], Artur M. Schweidtmann[c], Stefanie Winkler[d], Jana M. Weber[e], Philipp Morsch[f], K. Alexander Heufer[f], Martin Grohe[d], Alexander Mitsos[h,a,g], Manuel Dahmen[g†]

[a] RWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen 52074, Germany
[b] Aarhus University, Department of Computer Science, Aarhus, Denmark
[c] Delft University of Technology, Department of Chemical Engineering, Delft 2629 HZ, The Netherlands
[d] RWTH Aachen University, Chair of Computer Science 7, Aachen 52074, Germany
[e] Delft University of Technology, Delft Bioinformatics Lab, Intelligent Systems, 2628 XE Delft, The Netherlands
[f] RWTH Aachen University, Chair of High Pressure Gas Dynamics, Aachen 52074, Germany
[g] Forschungszentrum Jülich GmbH, Institute for Energy and Climate Research IEK-10: Energy Systems Engineering, Jülich 52425, Germany
[h] JARA-ENERGY, Aachen 52056, Germany

arXiv:2206.00619v2 [cs.LG] 14 Oct 2022

Abstract

Fuels with high-knock resistance enable modern spark-ignition engines to achieve high efficiency and thus low CO_2 emissions. Identification of molecules with desired autoignition properties indicated by a high research octane number and a high octane sensitivity is therefore of great practical relevance and can be supported by computer-aided molecular design (CAMD). Recent developments in the field of graph machine learning (graph-ML) provide novel, promising tools for CAMD. We propose a modular graph-ML CAMD framework that integrates generative graph-ML models with graph neural networks and optimization, enabling the design of molecules with desired ignition properties in a continuous molecular space. In particular, we explore the potential of Bayesian optimization and genetic algorithms in combination with generative graph-ML models. The graph-ML CAMD framework successfully identifies well-established high-octane components. It also suggests new candidates, one of which we experimentally investigate and use to illustrate the need for further auto-ignition training data.

1 Introduction

With a share of 23% of total CO_2 emissions, transportation is a major CO_2 emission source (IEA, 2020). Replacing fossil fuels with renewable alternatives may provide a path towards carbon neutrality for the transportation sector and is investigated actively (Dahmen and Marquardt, 2016; Leitner et al., 2017; Gschwend et al., 2019; König et al., 2021). An important step towards renewable fuels is the search for suitable gasoline substitutes for use in advanced high compression, turbocharged *spark-ignition* (SI) engines. A property of paramount importance for a renewable SI engine fuel is knock resistance, traditionally indicated by the *research octane number* (RON) (American Society for Testing and Materials, 2018), the *motor octane number* (MON) (American Society for Testing and Materials, 2019), and more recently the *octane sensitivity* (OS), i.e., the difference between RON and MON values. The weighted sum of RON and OS is referred to as the *octane index* (OI) (Kalghatgi, 2001). For modern SI engines, fuels with both high RON and high OS, hence high OI, are desired as they enable engine operation at conditions associated with particularly high efficiency (Kalghatgi, 2005; Bell, 2010; Mittal and Heywood, 2008; Amer et al., 2012; Kalghatgi, 2014; Szybist and Splitter, 2017; Abdul-Manan et al., 2018). To boost the OI of a fuel, chemical species with high RON and high OS such as ethanol and MTBE can be added (Demirbas et al., 2015; Badia et al., 2021). Identification of further molecules providing octane boosting is of great practical relevance and is studied actively, e.g., see (Badia et al., 2021; Li et al., 2022). Herein, we aim to identify such promising candidates exhibiting both high RON and high OS by *computer-aided molecular design* (CAMD). In particular, we investigate the role of novel methods from the domain of *graph machine learning* (graph-ML).

Traditionally, the search for molecules with desired properties for a given application has been mostly guided by human experts and experimentation. CAMD can enhance this process by utilizing

* Equally contributed
† Corresponding author, E-mail: m.dahmen@fz-juelich.de

computational methods to efficiently pre-screen a large number of molecular structures so that experiments can be dedicated to the most promising candidates. A wide variety of methods and tools for CAMD has been proposed over the last decades; we refer the interested reader to review articles for a detailed CAMD overview (Joback, 1989; Achenie et al., 2003; Gani, 2004; Zhang et al., 2015; Ng et al., 2015; Austin et al., 2016; Alshehri et al., 2020). Generally, the CAMD process incorporates the computational generation of candidate structures and the model-based prediction of their physico-chemical properties. Well-established approaches for the generation of candidate structures include formulating optimization problems in which structural groups are pieced together to form molecules (Samudra and Sahinidis, 2013; Austin et al., 2016), exhaustive generation of molecular structures in a sequential generate-and-test manner (Harper et al., 1999), and utilizing evolutionary theory to evolve molecular structures (Douguet et al., 2005). For predicting application-relevant properties of the formed candidate structures, CAMD typically employs quantitative structure-property relationships (QSPRs) (Katritzky et al., 2010). QSPRs first describe the molecular structure by so-called molecular descriptors, e.g., atom counts, and secondly map those descriptors to a property of interest by linear or nonlinear models. Today, nonlinear ML models such as feedforward neural networks or random forests are often utilized in this regression step (Mitchell, 2014; Lo et al., 2018; Muratov et al., 2020).

For classical CAMD, a broad range of applications (Alshehri et al., 2020) can be found in the process systems engineering (PSE) literature, covering the design of single molecules (e.g., ionic liquids (Karunanithi and Mehrkesh, 2013), polymers (Zhang et al., 2015)), the design of mixtures (Austin et al., 2017, 2018; Liu et al., 2019), as well as integrated product and process design (Lampe et al., 2015; Schilling et al., 2017). Classical CAMD techniques have also been applied extensively in the context of fuel design (Dahmen and Marquardt, 2016; Hechinger et al., 2010; Hoppe et al., 2016b; Whitmore et al., 2016; McCormick et al., 2017; Lunderman et al., 2018). For example, in two previous articles (Dahmen et al., 2012; Hoppe et al., 2016b), we used enumeration-based generation of oxygenated hydrocarbons and subsequently screened the obtained molecules via QSPR models with respect to engine-relevant properties. We previously also developed a generate-and-test approach where molecular candidates are generated by iteratively refunctionalizing bioderived intermediates based on pre-defined transformation rules (Dahmen and Marquardt, 2016). Also, Cai et al. (Cai et al., 2021) proposed a gasoline design model that employs rule-based transformation of molecules in combination with QSPR for property prediction to identify molecules with desired fuel properties such as high RON.

ML has recently been utilized for molecular structure generation by means of generative ML models, leading to novel, fully *ML-based CAMD* approaches (Alshehri et al., 2020; Elton et al., 2019). In generative ML for molecules, two main directions can be distinguished: String-based approaches, e.g., based on SMILES strings (Weininger, 1988), and graph-based approaches, the latter directly working on the molecular graph. For both directions, a range of models has been developed such as recurrent neural networks (RNNs), variational or adversarial autoencoders (VAEs/AAEs), generative adversarial networks (GANs), and reinforcement learning (RL) (Elton et al., 2019; Faez et al., 2021). The goal of such generative ML techniques is the unsupervised learning from a data set of molecular structures to generate new, chemically feasible structures that were not seen during training, thereby designing molecules. Specifically, generative ML models typically learn to encode molecules into a continuous space, the so-called latent space, and then decode samples from the latent space back to molecular structures. The continuous latent space is assumed to capture chemical information about molecules and embed molecules with similar structure or even similar properties close to each other (Winter et al., 2019a). Depending on the model architecture, ML-based CAMD typically relies either on strategic sampling of molecules from the latent space of the generative model using optimization strategies, e.g., with VAEs (Sanchez-Lengeling and Aspuru-Guzik, 2018; Jin et al., 2018; Kajino, 2019), or on direct generation of molecules with desired properties, e.g., by GANs (Guimaraes et al., 2018; De Cao and Kipf, 2018) or RL (You et al., 2018; Zhou et al., 2019). In contrast to classical CAMD, generative models in ML-based CAMD replace discrete molecule representations such as combinations of structural groups, molecular graphs, or SMILES strings with a continuous representation, thus enabling the use of continuous optimization approaches for molecular design (Coley et al., 2020).

ML has also recently enabled end-to-end learn-

ing of physico-chemical properties from molecular structure by means of graph neural networks (GNNs) (Sperduti and Starita, 1997; Gori et al., 2005; Scarselli et al., 2009). GNNs are graph-ML architectures that directly operate on the underlying graph structure of a molecule and thus circumvent the need for selecting meaningful molecular descriptors, a step that is inherent to all QSPR/QSAR approaches. Instead, GNNs enable a data-driven end-to-end learning framework for molecular property prediction.

Up to now, fully ML-driven CAMD has mainly focused on drug design (Elton et al., 2019; Xia et al., 2019; Xiong et al., 2021; Gaudelet et al., 2021). A particular reason might be the availability of large training data sets and the incorporation of multiple drug design targets such as logP and drug-likeness in benchmarking platforms such as MOSES (Polykovskiy et al., 2020) and GuacaMol (Brown et al., 2019). Such ML-driven CAMD approaches often combine molecule generation and property prediction (e.g., VAEs (Jin et al., 2018; Kajino, 2019)), and sometimes optimization (e.g., GANs (Guimaraes et al., 2018; De Cao and Kipf, 2018) or RL (Zhou et al., 2019)) in a single ML model which needs to be retrained once the design target property changes and typically requires large property data sets for training.

In contrast to drug design, PSE applications, in particular model-based fuel design, often take place in a data-scarce environment, making ML-based CAMD challenging. In fact, there is only one very recent study using generative ML for fuel design: Liu et al. (Liu et al., 2022) employed a string-based VAE to generate a large database of non-oxygenated hydrocarbons for subsequent screening of candidates with respect to fuel properties, followed by sampling further candidates from the most promising regions of the VAE's latent space. However, ML-driven CAMD has not yet been utilized for fuel design focusing on high SI engine efficiency including oxygenated hydrocarbons. Moreover, graph-ML approaches have not yet been applied to computer-aided fuel design.

In the present contribution, we propose a modular graph-ML CAMD framework[1] that integrates state-of-the-art graph-based ML methods and tools from the ML and drug design community and apply our framework to computer-aided design of high-octane fuel components for SI engines. Our framework is depicted in Figure 1 and consists of three distinct modules: (1) molecule generation by generative graph-ML models that learn a continuous molecular space from which new molecules can be generated; (2) property prediction through our recently published GNN model for fuel ignition quality prediction (Schweidtmann et al., 2020a); (3) optimization for strategic sampling from the continuous space of the generative graph-ML models to identify vectors that correspond to molecules with high predicted RON and OS values. Our framework has a modular architecture requiring minimal changes to the model structures if an additional property shall be targeted, i.e., only a new property model needs to be trained and added, but the molecule generation and optimization modules do not need to be altered. Thus, the modular setup enhances reusability and therefore reduces the training effort compared to a single ML model approach, as indicated by Winter et al. (Winter et al., 2019b).

We explore three different generative graph-ML models and two different optimization strategies. Importantly, we propose an applicability domain approach for GNN-based property prediction that allows us to focus the design process on molecules that presumably come with reliable predictions. We analyze the influence of the different ML methods on the structure and properties of the resulting molecules and compile a list of most promising high-octane fuel candidates. Finally, we perform an experimental investigation of one selected high-octane fuel candidate that emphasizes the importance of experimental validation of CAMD results and discuss potential pitfalls of the fully data-driven approach, particularly in a data-scarce environment.

The article is structured as follows: In the section "Preliminaries of graph machine learning", we briefly introduce the main principles behind graph-ML for molecules with regard to both molecule generation and property prediction. In the subsequent section, we present the modular graph-ML CAMD framework for design of high-octane fuels. The application of the framework in the section "Results and discussion" includes a comparative analysis of the candidates obtained with different graph-ML modules and the experimental investigation of one particular candidate. The last section concludes our work.

2 Preliminaries of graph machine learning

Graph-ML relies on a graph representation of molecules that can be utilized for generating molec-

[1] Code is openly available, see (Rittig et al., 2022)

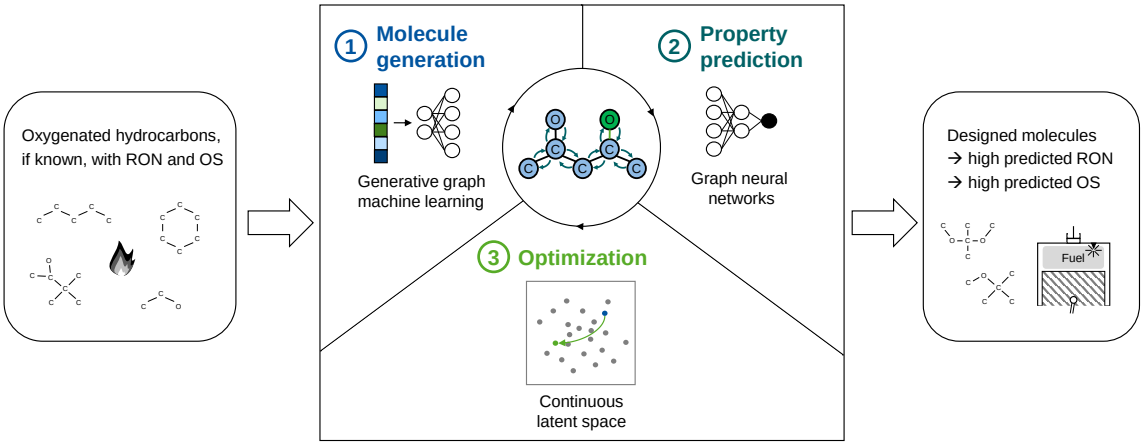


Fig. 1. Schematic overview of the modular graph-ML CAMD framework for identification of high-octane fuels.

ular structures from a continuous space and for property prediction, as we briefly describe in the following. The interested reader is referred to (Hamilton et al., 2017; Wu et al., 2021; Bronstein et al., 2021) for further details on graph-ML.

2.1 Molecular graph

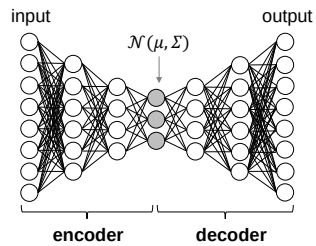
The *molecular graph* of a molecule is an undirected graph $G_{\text{mol}} = \{V, F_v, E, F_e\}$; the nodes V represent the atoms; pairs of atoms $u, v \in V$ that share a bond are connected by edges $(u, v) \in E$. Additional features of nodes (e.g., type of atom, degree of hybridization) are stored in F_v, while additional features of edges (e.g., bond length or type) are stored in F_e.

2.2 Generative models

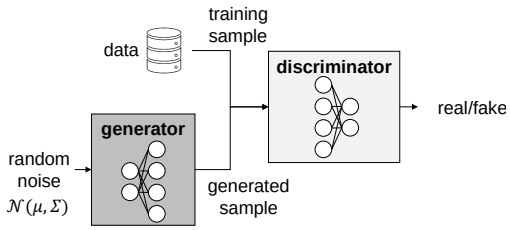
Generative ML, the unsupervised learning from input data to generate new data that is similar to the provided data, allows to perform fully data-driven molecule generation and is an active research area (Elton et al., 2019; Faez et al., 2021; Gaudelet et al., 2021; Atz et al., 2021). Various works have developed string-based ML models in order to generate molecules with optimal properties based on SMILES (Kadurin et al., 2017; Gómez-Bombarelli et al., 2018; Mario Krenn et al., 2020; Blaschke et al., 2018; Lim et al., 2018; Bjerrum and Sattarov, 2018; Prykhodko et al., 2019; Griffiths and Hernández-Lobato, 2020), InChI (Winter et al., 2019a), or SELFIES (Mario Krenn et al., 2020), the latter being a more robust string representation of molecules. In contrast, graph-ML directly works on the molecular graph which is arguably the more natural representation of a molecule and provides permutation invariance (Mercado et al., 2021), i.e., there is exactly one molecular graph for each molecule (neglecting steric effects). In this paper, we focus on two frequently employed generative graph-ML approaches (Elton et al., 2019; Faez et al., 2021; Gaudelet et al., 2021): VAEs and GANs. Both methods construct a latent space where molecules are encoded as high-dimensional continuous vectors, referred to as latent vectors (LVs), which we denote as $\mathbf{h}_{\text{LV}} \in \mathbb{R}^n$ with the dimension n being a hyperparameter. We denote the encoding of a molecular graph into the latent space as a function

$$e_{\text{GEN}} \colon G_{\text{mol}} \mapsto \mathbf{h}_{\text{LV}}. \tag{1}$$

Autoencoders, and specifically VAEs, are a class of neural network architectures that employs an hourglass shape (cf. Figure 2a). They are trained to reproduce the input data at the output layer, a non-trivial task as the information has to be moved through some narrow layers in the middle of the network, that is, the hourglass shape forces VAEs to learn $\mathbf{h}_{\text{LV}}$ as a low-dimensional representation of the input data at the most narrow layer. The left part of the network (from input to the latent vector) is called the *encoder* and the right part (from the latent vector to the output) is referred to as the *decoder*. The main difference between a standard autoencoder and a variational autoencoder (VAE) is that the latter assumes an underlying distribution for the data that it tries to learn in the latent vector space, e.g., a multivariate Gaussian distri-



(a) VAE



(b) GAN

Fig. 2. Schematic structure of (a) VAEs and (b) GANs.

bution $\mathbf{h}_{\mathrm{LV}} \sim \mathcal{N}(\mu, \Sigma)$ with parameters μ and Σ. VAEs can therefore be used to generate new data from presumably the same distribution as the input data. In the molecular context, VAEs map discrete molecule representations such as graphs to a continuous distribution from which new molecules can be sampled.

GANs generate objects from a latent representation in a different manner (cf. Figure 2b). Instead of trying to reproduce an input sample, a GAN consists of two neural networks, a *generator* and a *discriminator*, where the discriminator is trained to distinguish between output data produced by the generator and real data, i.e., the training samples. The generator thus learns to produce output data that resembles a given training data based on random input vectors $\mathbf{h}_{\mathrm{LV}}$ that are, for example, sampled from a Gaussian distribution, i.e., $\mathbf{h}_{\mathrm{LV}} \sim \mathcal{N}(\mu, \Sigma)$. In a GAN, the latent space therefore corresponds to the input space of the generator. We denote the decoding of the latent vector $\mathbf{h}_{\mathrm{LV}}$ to the molecular graph in case of both generators, VAE and GAN, with the function

$$d_{\mathrm{GEN}} \colon \mathbf{h}_{\mathrm{LV}} \mapsto G_{\mathrm{mol}}. \tag{2}$$

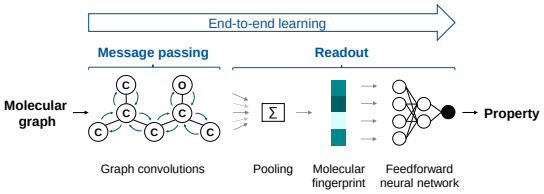


Fig. 3. Schematic structure of a graph neural network for molecular property prediction.

2.3 Graph-based property prediction

A GNN (Gori et al., 2005; Scarselli et al., 2009) is a type of neural network that operates directly on the graph structure and thus enables end-to-end learning in molecular property prediction. Thereby, GNNs avoid the need for the often subjective manual selection process of molecular descriptors in QSPR/QSAR modeling that requires intuition and experience of the modeler.

GNNs for molecular property prediction are typically structured into two parts, a message passing phase and a readout phase (Gilmer et al., 2017; Coley et al., 2017) (cf. Figure 3). In the message passing phase, structural information is extracted from a local neighborhood of atoms by means of graph convolutions. In each graph convolution, every node sends a message to all its neighbors and thus also receives a message from each of its neighbors. The node uses the received messages, typically in form of a weighted sum, to update its current state (e.g., in GCN (Hamilton et al., 2017) and GAT (Velickovic et al., 2018)). The update of the state $\mathbf{h}_v^l$ of a node v in a graph convolutional layer l can then be written as

$$\mathbf{h}_v^{l+1} = \sigma_{\mathrm{ReLU}} \left(\mathbf{h}_v^l \mathbf{W}_1 + \sum_{u \in N(v)} \mathbf{h}_u^l \mathbf{W}_2 \right), \tag{3}$$

where $\mathbf{W}_1, \mathbf{W}_2$ are trainable weight matrices, $N(v)$ is the one-hop neighborhood of v, and σ_{ReLU} denotes the elementwise application of the ReLU activation function. Many different update functions have been proposed in the last years, see, e.g., (Wu et al., 2021; Zhou et al., 2020; Zhang et al., 2022), to advance the basic Equation (3) into a more powerful model for extracting information from the graph during message passing (Xu et al., 2018). For instance, inter-atomic distances and angles between atom pairs (Schütt et al., 2018; Unke and Meuwly, 2019; Klicpera et al., 2020; Zhang et al., 2020) are commonly considered. Higher-order GNNs (Mor-

ris et al., 2019; Flam-Shepherd et al., 2021) and approaches where the information exchange is also based on individual edges (Yang et al., 2019) constitute further extensions to the basic GNN approach.

Subsequent to the message passing phase, a GNN employs a readout phase, where the molecular structure information that is stored in the nodes is aggregated into a single vector for the complete molecule, the so-called *molecular fingerprint* $\mathbf{h}_{\mathrm{FP}}$. This aggregation, also called pooling, is typically performed by summing up the states of all nodes in the molecular graph after the last graph convolutional layer L, i.e., $\mathbf{h}_{\mathrm{FP}} = \sum_{v \in V} \mathbf{h}_v^L$. We denote the GNN encoding of the molecular graph into the molecular fingerprint with the function

$$g_{\mathrm{GNN}} \colon G_{\mathrm{mol}} \mapsto \mathbf{h}_{\mathrm{FP}}. \tag{4}$$

Note that although the molecular fingerprint $\mathbf{h}_{\mathrm{FP}}$ in a GNN and the latent vector $\mathbf{h}_{\mathrm{LV}}$ in a generative ML model both represent a molecule in a continuous space, they are not related. In the GNN, the molecular fingerprint $\mathbf{h}_{\mathrm{FP}}$ is passed through a feedforward neural network (cf. Figure 3) to yield the property prediction $\hat{p} = MLP(\mathbf{h}_{\mathrm{FP}})$. Here, a multi-layer perceptron (MLP) is one of the simplest feedforward neural architectures and most frequently employed. We denote the entire end-to-end prediction process of a GNN as a function f_{GNN} that maps the molecular graphs to a property prediction, i.e.,

$$f_{\mathrm{GNN}} \colon G_{\mathrm{mol}} \mapsto \hat{p}. \tag{5}$$

3 Graph-ML CAMD framework for high-octane fuels

In this section, we propose a fully data-driven, modular graph-ML CAMD framework for identification of high-octane fuels. The framework utilizes recent methods from the field of generative graph-ML and GNNs to design molecules with high-knock resistance for modern SI engines. Specifically, we set out to maximize the sum of RON and OS, hence the OI, as high-efficiency SI engines require both a high RON and a high OS (Kalghatgi, 2005; Bell, 2010; Mittal and Heywood, 2008; Amer et al., 2012; Kalghatgi, 2014; Szybist and Splitter, 2017; Abdul-Manan et al., 2018). We show a high-level overview of our framework in Figure 1 and provide a detailed framework overview including our choices for algorithms and models in the three different modules in Figure 4. We combine the three modules to form an iterative molecular design loop: The *optimization module* proposes initial latent vectors from a continuous space, $\mathbf{h}_{\mathrm{LV}}$, that are translated to corresponding molecules by the *molecule generation module*, cf. Equation (2). Then, the *property prediction module* performs the property evaluation, cf. Equation (5), and based on the property predictions, the optimization algorithm suggests new latent vectors to be tested. This iterative procedure is repeated until a pre-defined stopping criterion is met, e.g., a certain number of molecules has been evaluated.

An important observation with the graph-ML CAMD framework though is that not all molecules come with physically reasonable predictions. For instance, we have observed a molecule with predicted OS > 400 and negative RON and negative MON[2]. In fact, the optimization often exploits weak spots of the GNN prediction model. Those weak spots typically appear for molecules that are strongly dissimilar from the molecules used for training the GNN. To focus on molecules with more reasonable property predictions, we extend the iterative design loop by an *applicability domain* (AD) for the GNN property prediction model. To this end, we build upon the AD approach from our previous study (Schweidtmann et al., 2021b) where we proposed to use a one-class classification model to identify the AD of feedforward NNs. The classification model learns from the data on which the NN is trained to decide if a new data point is similar to the training data and thus considered within the input domain for which the NN presumably provides reliable predictions. To transfer the AD approach to GNNs, we apply the classification model to the molecular fingerprint that serves as input to the MLP part of the GNN (cf. Subsection "Graph-based property prediction"). If the AD is included, GNN predictions considered unreliable by the AD are ignored and instead a penalty value (-1000) is returned to the optimization approach so that the corresponding molecules are assigned a low objective value.

The design loop runs can be formulated as an optimization problem that aims to find the molecules with the highest predicted value of a certain target property $\hat{p}$ of interest, i.e.,

$$\begin{aligned} \max_{\mathbf{h}_{\mathrm{LV}}} \quad & \hat{p} \\ \text{s. t.} \quad & G_{\mathrm{mol}} = d_{\mathrm{GEN}}(\mathbf{h}_{\mathrm{LV}}), \\ & \hat{p} = f_{\mathrm{GNN}}(G_{\mathrm{mol}}), \\ & \mathbf{h}_{\mathrm{FP}} = g_{\mathrm{GNN}}(G_{\mathrm{mol}}), \\ & AD(\mathbf{h}_{\mathrm{FP}}) \geq 0, \end{aligned} \tag{6}$$

[2]Found by GA when optimizing for OS only

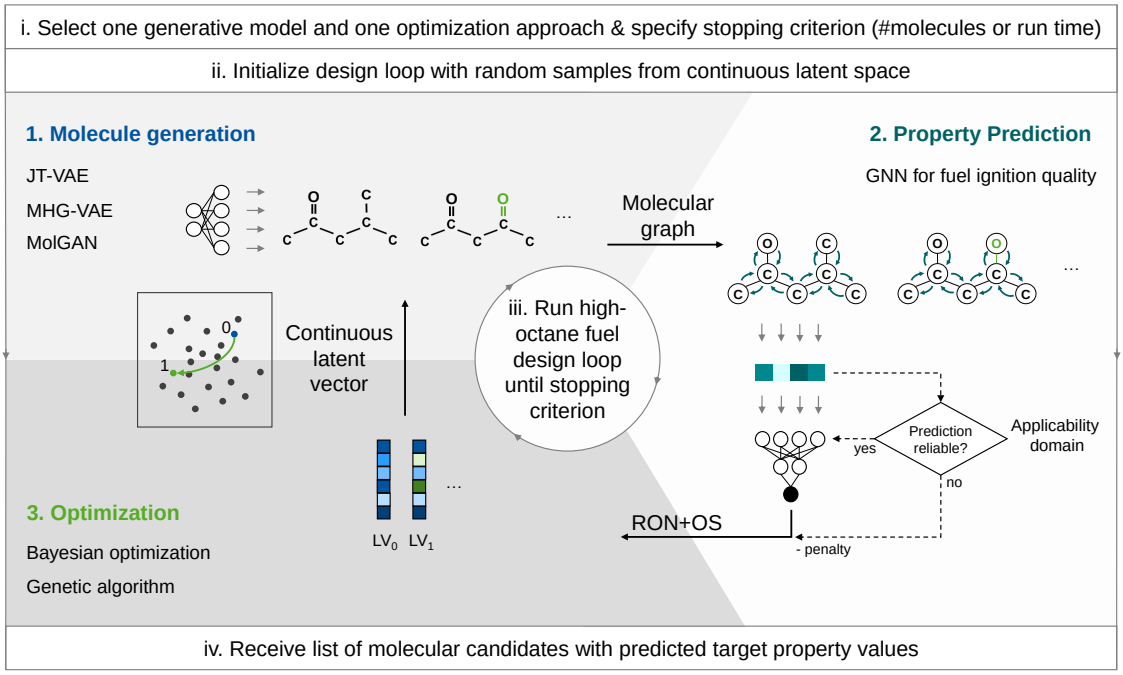


Fig. 4. Detailed overview of the modular graph-ML CAMD framework for identification of high-octane fuels including methods for the individual modules.

whereby the constraint with $AD(\mathbf{h}_{\mathrm{FP}}) \geq 0$ denotes a positive decision by the AD model.

Due to the high dimensionality of the search space that corresponds to the latent space of the generator models (see Equation (6)), deterministic global optimization is too computationally costly and practically impossible with current methods (cf. Subsection "Optimization" below). Instead, we employ black-box optimization approaches that direct a heuristic search towards molecules with high $\hat{p}$. Note that uncertainties in the prediction model prohibit a strict ranking of molecular candidates with similar $\hat{p}$ values. Practically, we therefore compile a list of molecules sampled by the optimizer and perform an investigation of the top candidates, i.e., the molecules with the highest $\hat{p}$ values. Having multiple top candidates, also allows to take additional desired properties into account in later investigations, e.g., availability for procurement and low production costs.

In the following, we briefly describe the three generative graph-ML models used in this paper for the generation module, the GNN model used for the property prediction module, the two optimization algorithms used in the optimization module, and our AD approach.

3.1 Molecule generation

We consider two graph VAE models as generators: The Junction-Tree VAE by Jin et al. (Jin et al., 2018) (JT-VAE) and the Molecular Hypergraph Grammar VAE by Kajino (Kajino, 2019) (MHG-VAE). Furthermore, we employ MolGAN, a GAN for molecular graphs published by De Cao and Kipf (De Cao and Kipf, 2018). Those three models have close to 100% chemical validity, i.e., almost 100% of the generated molecules are chemically feasible (Jin et al., 2018; Kajino, 2019; De Cao and Kipf, 2018), a feature that earlier generative methods struggled with, cf. (Kusner et al., 2017; Dai et al., 2018). Apart from achieving high validity, the three models have strong conceptual differences, presumably leading to molecules with somewhat different characteristics.

The JT-VAE (Jin et al., 2018) utilizes two graph representations of a molecule in parallel: The molecular graph and its associated junction tree, which is a contracted cycle-free graph generated by merging cycles of atoms into a single node. For encoding, the JT-VAE learns molecular structure information, represented as high-dimensional vectors, from the molecular graph and the junction tree

through graph convolutions (cf. Section "Preliminaries of graph machine learning"). For decoding, first, the junction tree's latent vector is decoded resulting in the general molecular structure. Then, the molecular graph's latent vector is decoded to determine the characteristics of the nodes within the junction tree, i.e., (re)generating the local structure of the molecule. Jin et al. report a molecule reconstruction rate of 76.7% and 100% chemical validity of the decoded molecules (Jin et al., 2018).

The MHG-VAE (Kajino, 2019) generates a graph grammar from the given training molecules which is used for the reconstruction of molecules. In this automatically generated graph grammar, terminal symbols can refer to either single atoms or complete functional groups and the rules of the grammar describe how such atoms of partial molecules can be combined into a chemically valid molecule. During the generation of the grammar, MHG-VAE ensures that the grammar accounts for chemical feasibility constraints such as valency rules, explaining the validity of 100%.

MolGAN (De Cao and Kipf, 2018) only partially relies on graphs. Its adaptation to our case of high-octane fuel design is illustrated in Fig. 5. The generator tries to directly predict a molecular graph's adjacency matrix with corresponding atom and bond features by using an MLP with a fixed output size, i.e., the maximal size of a molecule that can be predicted by MolGAN is bounded. On the other hand, the discriminator is a GNN. One conceptual difference to the VAEs is that MolGAN is able to focus the generation on molecules with desirable properties by using a 'reward network', i.e., a third network that encourages the generator to output molecules with high RON and OS. We use our GNN model (Schweidtmann et al., 2020a) to provide RON and OS predictions such that, in contrast to the VAEs, the training of MolGAN partially depends on the property prediction module. De Cao & Kipf state that while MolGAN generates novel molecules with desirable properties and almost 100% chemical validity, it also outputs many duplicates with only about one in ten molecules being unique (De Cao and Kipf, 2018).

3.2 Property prediction

We recently developed a GNN for predicting the RON, MON, and the derived cetane number (DCN) of a wide range of oxygenated and non-oxygenated hydrocarbons (Schweidtmann et al., 2020a), e.g., (cyclo-) alkanes, (cyclo-) alkenes, alcohols, esters, ethers, aromatics, and ketones. The model architecture is based on higher-order GNNs (Morris et al., 2019) and additionally leverages the increased stability and accuracy of ensemble methods (Breiman, 1996a,b), i.e., the final property prediction is the average of multiple higher-order GNN predictions. Further, our GNN incorporates multi-task learning (Caruana, 1997; Ruder, 2017) as it was trained on RON, MON, and DCN values simultaneously allowing the model to capture and exploit correlations between octane and cetane numbers.

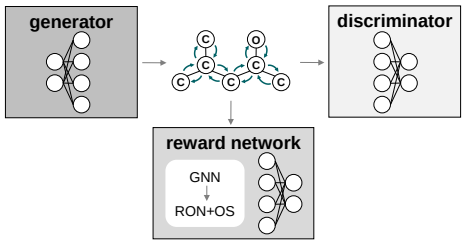


Fig. 5. Adapted MolGAN for high-octane fuels, modified from (De Cao and Kipf, 2018). The reward network is coupled with our GNN (Schweidtmann et al., 2020a) for predicting RON and OS values.

As described in detail in (Schweidtmann et al., 2020a), we compiled a data set comprising 335 RON, 318 MON, and 236 DCN values for 505 unique molecules in total to train the GNN. 85% of the data was used for training and validation, and 15% was used for testing. Note that for most molecules, both RON and MON values and thus OS were available. The mean absolute prediction error of the GNN model was 4.5 on the RON test set and 4.4 on the MON test set, indicating an overall high prediction quality on par with state-of-the-art QSPR- and ML-based RON and MON prediction models, cf. (Schweidtmann et al., 2020a). The test sets also contain few outliers: Six predictions for RON and seven predictions for MON have a deviation > 10, which we attribute, similarly to vom Lehn et al (Vom Lehn et al., 2020), to some of these molecules having unique characteristics that are not well represented in the training data, a relatively small number of data points available with low RON and MON values, and potential disruptive factors in experimental data assembled from different sources.

3.3 Optimization

To sample molecules with high RON and OS from the latent space of the generative models, we employ numerical optimization using the $\mathrm{RON} + \mathrm{OS}$ score

predicted by the GNN model as objective function. Specifically, we seek to maximize $\hat{p} = \mathrm{RON} + \mathrm{OS} = 2 \cdot \mathrm{RON} - \mathrm{MON}$ (cf. Equation (6)). We explore two derivative-free stochastic global optimization methods to perform the molecule sampling: A Bayesian optimization algorithm and a genetic algorithm.

Bayesian optimization (BO) is a probabilistic approach for global optimization (Shahriari et al., 2016) commonly used for optimization of black-box models that are costly to evaluate. Usage of BO is well-established in ML-based CAMD, see, e.g., (Jin et al., 2018; Kajino, 2019; Gómez-Bombarelli et al., 2018), as well as in chemical engineering applications, e.g., the design of experiments in automated reaction platforms (Schweidtmann et al., 2018; Felton et al., 2021; Florian Häse et al., 2021). BO uses a surrogate model, typically a Gaussian process (GP), to map the input variables to the objective. Based on the surrogate model, an acquisition function locates input variable values that have a high potential of maximizing the objective by accounting for both exploitation and exploration. For running BO, the GP is initialized with a set of feasible points. Then, the following steps are repeated until a termination criterion is reached: The acquisition function is optimized to determine the next sampling points, the sampling points are evaluated with respect to the objective function, and the objective values are used to refine the surrogate model. Note that different optimization algorithms can be used for maximizing the acquisition function, cf. (Shahriari et al., 2016).

A *genetic algorithm* (GA) is a meta-heuristic, population-based approach for global optimization that is inspired from evolutionary theory (Holland, 1992; Whitley, 1994). It is typically applied to optimization problems with cheap and fast evaluations of the objective function. In GAs, a set of feasible points is called population. Each feasible point has genes corresponding to specific values for the input variables of the optimization problem and constitutes a fitness related to the objective value. To solve an optimization problem, an initial population evolves in an iterative manner over multiple generations by promoting points with high fitness and using evolutionary heuristics, e.g., combining genes of high fitness points, to replace points with low fitness. We choose the fitness to be $\mathrm{RON} + \mathrm{OS}$ to directly optimize for high-octane ratings.

A major challenge in ML-based CAMD is the high dimensionality of the generators' latent space which typically requires a large number of sampling points for optimization, e.g., in case of our generative models, we have latent space dimensionalities of 56 (JT-VAE) (Jin et al., 2018), 72 (MHG-VAE) (Kajino, 2019), and 32 (MolGAN) (De Cao and Kipf, 2018). BO, however, employs a GP as surrogate model that in standard form has cubic scaling in complexity with respect to the number of sampling data points. Following the strategy by Kajino (Kajino, 2019), we thus use PCA to reduce the dimensions of both the JT-VAE and the MHG-VAE before performing BO. Since the execution time of the evolutionary-based heuristics in the GA does not suffer from a high number of sampling points, we run the GA without dimensionality reduction. Note that the effects of PCA-based dimensionality reduction on the obtained molecules as well as the use of other mitigation strategies, such as reduction of the latent dimension within the generator or modification of BO for high-dimensional problems, see, e.g., (Snoek et al., 2015; Shahriari et al., 2016; Wang et al., 2016, 2018; Kirschner et al., 2019), are beyond the scope of this work.

3.4 Applicability domain

The AD of a model is a well-established concept in QSPR/QSAR modeling and is based on the general assumption that the prediction model would provide most reliable predictions for molecules that are similar to the ones seen during training (Tropsha et al., 2003; Jaworska et al., 2005; Gramatica, 2007; Weaver and Gleeson, 2008). Molecular similarity is usually assessed by means of a distance metric, e.g., the Euclidean distance between the descriptor values of two molecules (Jaworska et al., 2005; Sheridan et al., 2004). For molecular property prediction with GNNs, determination of the AD is largely unexplored. Only very recently first approaches to quantify the AD of GNNs based on uncertainty quantification methods were proposed (Hirschfeld et al., 2020; Soleimany et al., 2021; Nigam et al., 2021). Conceptually, defining the AD of a GNN requires handling the varying input sizes of molecular graphs and measuring the degree of similarity between different graphs. In this work, we address these challenges by extending our recently developed AD approach based on one-class support vector machines (SVMs) (Schweidtmann et al., 2021b) to GNNs. A one-class SVM is a ML model that can be used to identify outliers by classifying whether an input is similar or dissimilar to the training data. We train one-class SVMs on the molecular fingerprint of the GNN (cf. Figure 3) to determine the GNN's AD. We then restrict our molecular design loop to molecules which

are accepted by the SVM (cf. Equation (6)) which formally means $AD(\cdot) = SVM_{\text{AD}}(\mathbf{h}_{\text{FP,train}}) \geq 0$ where $\mathbf{h}_{\text{FP,train}}$ is the molecular fingerprint computed by the GNN and SVM_{AD} denotes the trained SVM. The underlying idea for the AD is that the GNN computes similar molecular fingerprints whenever two molecules are structurally similar. Since our prediction model is an ensemble of multiple GNNs, we train one SVM for each GNN model in the ensemble and apply a majority vote. That is, each SVM j evaluates $SVM_{\text{AD},j}(\mathbf{h}_{\text{FP}})$ and returns 1 if the molecule lies within the AD or -1 if not. Subsequently, we sum up the votes to decide if the prediction of the GNN ensemble (EL) for a new molecules is classified as reliable, i.e., $SVM_{\text{AD-EL}}(\mathbf{h}_{\text{FP}}) = \sum_j SVM_{\text{AD},j}(\mathbf{h}_{\text{FP}}) \overset{!}{>} 0$. Note that further details on the AD are described in the ESI.

3.5 Implementation and hyperparameters

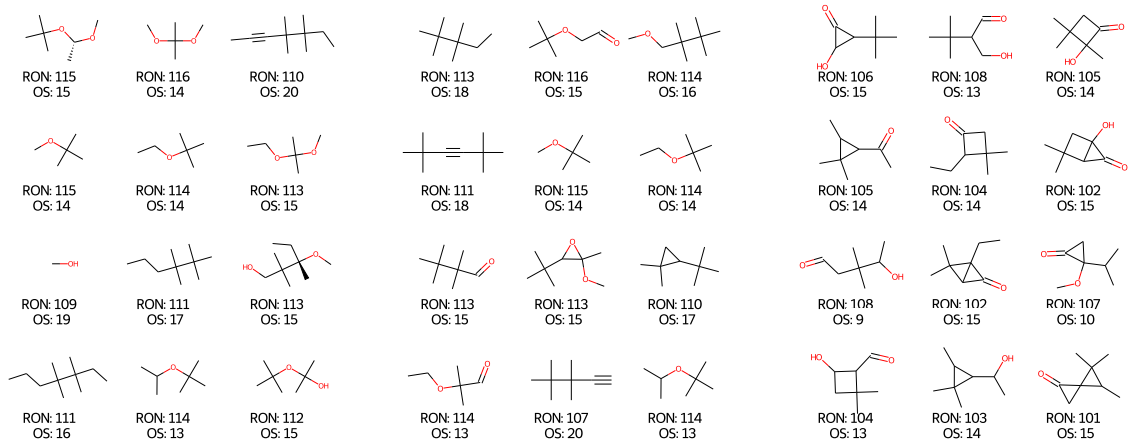
We implement our graph-ML CAMD framework in Python with the cheminformatic package RDKit (Landrum, 2021) and the ML frameworks pytorch and tensorflow, accounting for the different implementations of the generators, and provide our code open-source, see (Rittig et al., 2022). Moreover, we follow the implementation of the MHG-VAE by Kajino (Kajino, 2019) and use Luigi (The Luigi authors, 2012) to automate computational experiments. For the three generators, JT-VAE (Jin et al., 2018), MHG-VAE (Kajino, 2019), and MolGAN (De Cao and Kipf, 2018), we use the original implementations and hyperparameters as provided in the respective study and code repository and only extend the code to work in our framework. We train the molecule generation models on all HCO-molecules in the QM9 data set (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014), i.e., all molecules within QM9 that contain exclusively hydrogen, carbon, and oxygen atoms. QM9 contains approximately 50,000 HCO-molecules from various molecular classes. We use the original implementation and model parameters of our GNN (Schweidtmann et al., 2020a) which is based on pytorch-geometric (Fey and Lenssen, 2019). The SVMs for the AD are implemented with scikit-learn (Pedregosa et al., 2011) building on our AD study (Schweidtmann et al., 2021b). For BO, we use GPyOpt (The GPyOpt authors, 2016). Note that we did not attempt deterministic global optimization of the acquisition function within the BO, e.g., by using our tool MeLOn (Schweidtmann and Mitsos, 2019; Schweidtmann et al., 2021a), due to the high dimensionality (cf. Subsection "Optimization") and associated high computational cost. Thus, we use the local optimization algorithm L-BFGS (Liu and Nocedal, 1989) implemented in GPyOpt (The GPyOpt authors, 2016). As GA, we use the python package geneticalgorithm (Solgi, 2020). For both BO and GA, we apply default settings. We follow the study of MHG-VAE by Kajino (Kajino, 2019) and reduce the dimensionality of the latent space within the VAEs by means of PCA aiming for an explained variance ratio of 99.9% (JT-VAE: from 56 to 41, MHG-VAE: from 72 to 38) before performing BO. Further details on the hyperparameter choice can be found in the ESI. We run all computations on the HPC-cluster (CLAIX-2018) of RWTH Aachen University using one Supermicro 1029GQ-TVRT-01 node of an Intel Platinum 8160 core with 192 GB RAM, of which we used at most 8 GB, plus one NVIDIA Volta V100-SXM2 16 GB GPU. For reproducibility, we fixed random seeds for training the models and running the design loop that we provide with our code.

4 Results and discussion

We first present the computational results of our graph-based CAMD of high-octane fuels and then provide a discussion of the top candidates to demonstrate both strengths and potential weaknesses of the fully data-driven design approach.

4.1 CAMD results

We test all combinations of the three generator models (JT-VAE, MHG-VAE, and MolGAN) and the two optimization approaches (BO and GA) as well as two different stopping criteria (SC), i.e., a limit on the number of candidate molecules generated ($\text{SC}_{\#\text{molecs}}$) and an upper limit on the wall-clock run time (SC_{time}). For $\text{SC}_{\#\text{molecs}}$, we consider both the number of unique molecules (1,000) and the total number of molecules (2,000) generated, as the number of duplicates can otherwise cause an unlimited run time. In the $\text{SC}_{\#\text{molecs}}$ setting, the design loop will typically run for 0.5 to 8 hours. The run time limit in SC_{time} is set to 12 hours to investigate the effects of keeping the design loop running for a longer time. Furthermore, we distinguish between runs with and without the AD. All design loop runs are run five times (initialized with different random seeds) and the results are aggregated.



(a) JT-VAE **(b)** MHG-VAE **(c)** MolGAN

Fig. 6. Top 12 candidates identified by the three different generator models with stopping criterion $SC_{\#molecs}$ (max. 1000 unique molecules or max. 2000 total molecules) and applicability domain. RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020a).

The top 12 molecules identified with $SC_{\#molecs}$ and active AD for the respective generators are shown in Figure 6 together with the predicted RON and OS values. The results demonstrate that the generators successfully propose molecules with high predicted RON and OS. Moreover, the top molecules are from a variety of different molecular classes, e.g., ethers, alcohols, and ketones, some of which are known to contain promising SI engine fuel candidates (Dahmen and Marquardt, 2016). The majority of molecules has at least one oxygen atom. Almost all top molecules generated by MolGAN include a cyclic structure, often associated with a cyclopropane feature, which we attribute to the high RON and OS for components with a cyclopropane substructure in the training set of the GNN model (Schweidtmann et al., 2020b). Most top molecules generated by the two VAE models include strongly branched non-cyclic components, often in combination with one or two oxygen atoms, which are also known for high RON and OS values. Both VAE models generate the popular octane enhancers MTBE and ETBE, and some related small, branched ether structures. The JT-VAE also identifies ethanol, the prototype biofuel for SI engines.

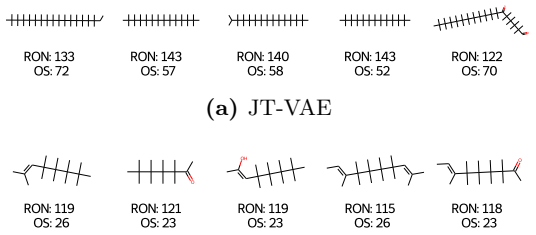
Table 1 shows the statistics of all the runs with and without the AD, whereby each entry corresponds to the aggregated results over five runs. Both the maximum and the mean predicted RON + OS are typically lower if the AD is used. In most cases, also the total number of molecules generated is lower if the AD is considered. The observation that the AD often reduces the exploration performance is expected and in fact intended as the AD prohibits the generators from exploring structures that are far from the training data by strongly extrapolating the GNN model. We want to emphasize that we find the generators to mainly produce chemically valid molecules. Otherwise, e.g., MolGAN sometimes generates disconnected substructures, the generated molecule is dropped so that effectively no chemically invalid structures are provided to the GNN and AD. Note that generated molecules, which are considered highly dissimilar to the training molecules by the AD, can still be chemically valid. We show examples of such chemically valid molecules well outside the GNN's AD in Figure 7, where the top candidates identified by the two VAEs with SC_{time} are depicted; we refer to the ESI for further examples.

When visually inspecting the top molecules from the design runs without AD, we find that the obtained molecules are typically huge, strongly branched hydrocarbons, e.g., with up to almost 50 carbon atoms. As such compounds are presumably solid at room temperature, they are not suitable as fuels. To avoid solid molecules, a constraint on the melting point could be included in the design loop. However, the melting point can only serve as a rough proxy for the suitability of a compound as an octane booster, since miscibility and volatility also depend on the composition of the base fuel and the

Tab. 1. Results of optimization over 5 runs each. A molecule is considered promising if both RON > 110 and OS > 10. Runs with applicability domain are indicated by +AD.

predicted RON+OS		JT-VAE				MHG-VAE				MolGAN			
		BO	BO+AD	GA	GA+AD	BO	BO+AD	GA	GA+AD	BO	BO+AD	GA	GA+AD
$SC_{\#molecs}$ (1000 unique molecules, 2000 total) #runs: 5	max	205	130	129	130	138	129	136	131	121	121	121	121
	mean top 20	181	125	125	126	131	125	132	128	110	111	116	116
	# unique mol.	2390	1347	3472	3712	4671	4308	4683	4427	21	21	46	46
	# promising mol.	117	10	15	19	45	9	52	30	0	0	0	0
SC_{time} (12h run time) #runs: 5	max	205	130	187	131	138	129	145	131	121	121	121	121
	mean top 20	183	126	180	130	133	126	140	129	111	112	118	118
	# unique mol.	2996	1935	109830	80818	6710	7081	55255	46989	22	23	193	172
	# promising mol.	140	12	2096	376	104	15	678	142	0	0	0	0

blending ratio (McCormick et al., 2017; König et al., 2020). Some of the proposed large molecules might be soluble in a fuel blend, which could be evaluated in further investigations of mixture properties, but is beyond the scope of this work. Furthermore, the RON and OS predictions for the molecules identified with the JT-VAE without AD (cf. Figure 7a) are visibly higher than the maximum RON (of 120 for 1,3,5-trimethylbenzene (Schweidtmann et al., 2020a; Derfer et al., 1958)) and the maximum OS (of 36 for 1,4-cyclohexadiene (Schweidtmann et al., 2020a; Derfer et al., 1958)) of the data used to train the GNN prediction model, indicating strong extrapolation. In the following, we therefore present and discuss only those results that have been obtained with the AD.



(a) JT-VAE

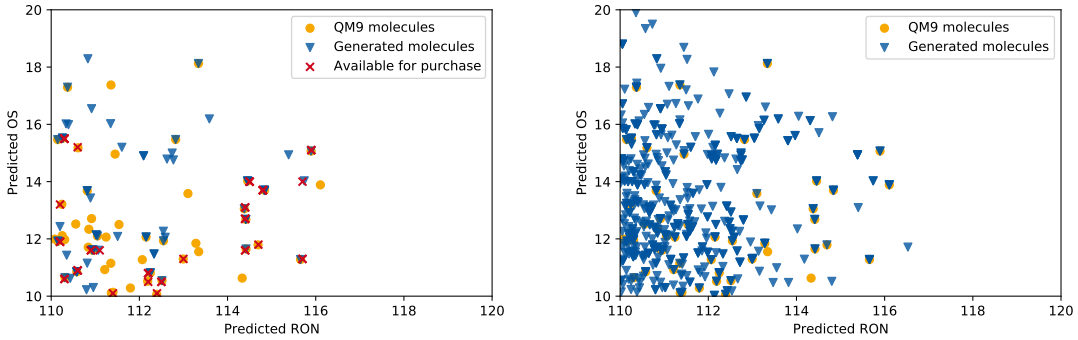
(b) MHG-VAE

Fig. 7. Top 5 candidates identified by the two VAE generator models with stopping criterion SC_{time} (12 run time) and without applicability domain. All RON and OS values are GNN predictions (Schweidtmann et al., 2020a).

We observe that the VAE generators predict molecules with a maximum RON + OS of about 130 while MolGAN achieves a maximum of only 121 (cf. Table 1). The maximum RON + OS values of slightly above 130 for the two VAE models are in good agreement with known high-octane fuels such as MTBE with its experimentally validated RON + OS of 135. The encouraging performance of both VAE generators thus shows the general feasibility of our graph-ML CAMD framework utilizing the SVM-based AD.

To further compare the different generator and optimization combinations, we analyze the number of distinct molecules generated as well as the number of molecules with promising ignition properties, i.e., the molecules with both a predicted RON > 110 and a predicted OS > 10. Both VAEs find a large number of distinct molecules irrespective of the employed stopping criteria (cf. Table 1). Specifically for $SC_{\#molecs}$, both VAEs generate more than 3,500 unique molecules out of 5,000 maximally possible unique molecules (1,000 unique molecules each over 5 runs). This means that not only do the VAEs find a large number of distinct molecules in each run, but the identified molecules also vary greatly between different runs, thus leading to an overall small number of duplicates. In contrast, MolGAN mainly generates duplicates of which none are considered promising (cf. Table 1). Comparing the results for $SC_{\#molecs}$ and SC_{time} (cf. Table 1), it can be seen that the VAE-GA combinations significantly increase the number of both explored and promising candidates with longer run time. Apparently, this observation does not extend to BO, with one possible explanation being that BO becomes inherently slower as more data points are added to the surrogate model, thereby reducing the number of predictions per time, whereas the corresponding rate remains unchanged in the GA (cf. Subsection "Optimization").

The predicted RON and OS values of all promising molecules obtained with the two stopping criteria are shown in Figure 8. We also highlight those molecules identified in the $SC_{\#molecs}$ setting that are commercially available at chemical suppliers. Commercial availability was assessed by a manual search on Sigma-Aldrich (Merck KGaA, 2021) and Chemspider (Royal Society of Chemistry, 2021) websites without imposing a price limit but only including those molecules with an explicitly stated



(a) $SC_{\#molecs}$ (1000 unique molecules, 2000 total), all generators, with AD

(b) SC_{time} (12h run time), all generators, with AD

Fig. 8. Promising candidates (predicted RON > 110 and OS > 10). Commercially availability (red crosses) determined by manual search on Sigma-Aldrich and Chemspider websites (Merck KGaA, 2021; Royal Society of Chemistry, 2021).

price; we did not search for the lowest price on different websites. For SC_{time}, Figure 8b, the effort for a manual search was considered disproportional due to the high number of promising candidates. We further indicate molecules with high predicted RON + OS in the QM9 database (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) that is used for training the generative models; additional QM9 statistics are provided in the ESI. Figure 8 demonstrates that the graph-ML CAMD framework is able to generate molecules with high predicted RON and high predicted OS that are not in the QM9 database. This observation is emphasized in case of SC_{time} (cf. Figure 8b). The capabilities of the generator models to generalize therefore allow to explore novel molecules for further investigation.

4.2 Discussion of top candidates

In the discussion of the top molecules, we restrict our analysis to the promising molecules (RON > 110 and OS > 10) generated using $SC_{\#molecs}$, as the number of molecules generated with SC_{time} is very large; we refer to the ESI for a detailed list of all generated promising molecules. The top molecules that are also commercially available are illustrated in Table 2, including RON and OS predictions, literature values for RON and OS (where available), price category, and the respective combinations of generator and optimizer that identified the molecule.

4.2.1 Promising classes of molecules

We find both pure hydrocarbons and oxygenated hydrocarbons (cf. Table 2), molecules already in use as octane boosters and molecules that constitute interesting candidates for further experimental investigation. The two identified alkanes, ethane and cyclopropane, are gaseous under ambient conditions, whereas the one aromatic hydrocarbon, 3-ethyltoluene, is liquid. The known RON + OS scores from literature for ethane and 3-ethyltoluene of 122 and 124, respectively, are in good agreement with the GNN predictions. We want to emphasize that gaseous compounds, such as ethane and cyclopropane, are difficult to implement as octane boosters. To prevent gases within the candidate list, one could include boiling point constraints in the design loop. However, the normal boiling point is, similar to the melting point discussed at the beginning of this section, only a rough preselection criterion, since the miscibility and volatility of a potential octane booster in a fuel blend strongly depend on the overall blend composition. Next to alkanes, three ethers are identified, including methyl tert-butyl ether (MTBE) and ethyl tert-butyl ether (ETBE) that are used as octane boosters in practical applications (Demirbas et al., 2015; Badia et al., 2021). Their experimentally RON + OS scores of 135 and 134 (Leppard, 1991; Kubic et al., 2017) are slightly higher than the predicted scores. Furthermore, molecules from the class of aldehydes are identified. It has been found, however, that the formation of aldehydes during the combustion process of

Tab. 2. All 16 commercially available molecules with predicted RON > 110 and OS > 10 (identified in $SC_{\#molecs}$ setting and active applicability domain). RON and OS data available in the literature are stated in parentheses. Prices are categorized based on data from different chemical suppliers (Merck KGaA, 2021; Chemspace US Inc., 2021; Enamine Ltd, 2021; SynQuest Laboratories, Inc., 2021): $\leq$ 1000\$/l (low), > 1000 \$/l and $\leq$ 10 000 \$/l (medium), > 10 000 \$/l (high).

Class	Structure	SMILES	RON	OS	Price category	Generator (Optimizer)
alkanes		C1CC1 cyclopropane	110	16	medium	JT (BO, GA), MHG (BO, GA)
		CC ethane	110 (111 (Derfer et al., 1958))	12 (11 (Derfer et al., 1958))	low	JT (BO, GA)
aromatics		CCc1cccc(C)c1 3-ethyltoluene	110 (112 (Derfer et al., 1958))	11 (12 (Derfer et al., 1958))	high	JT (GA)
ethers		COC(C)(C)C MTBE	115 (118 (Leppard, 1991))	14 (17 (Leppard, 1991))	low	JT (BO, GA), MHG (BO, GA)
		CCOC(C)(C)C ETBE	114 (118 (Kubic et al., 2017))	14 (16 (Kubic et al., 2017))	medium	JT (BO, GA), MHG (GA)
		CC(C)OC(C)(C)C tert-butyl isopropyl ether	114	13	high	JT (GA), MHG (GA)
aldehydes		CC(C=O)C(C)(C)C 2,3,3-trimethylbutanal	111	12	high	MHG (GA)
		CC(C)(C)C=O trimethylacetaldehyde	111	11	medium	JT (GA), MHG (BO)
polyfunctional (aldehyde + ether)		CC(C)(C)OCC=O tert-butoxyacetaldehyde	116	15	high	MHG (GA)
		CCOC(C)(C)C=O 2-ethoxy-2-methylpropanal	114	13	high	MHG (GA)
		COC(C)(C)C=O 2-methoxy-2-methylpropanal	116	11	high	MHG (GA)
		CC(C)OC(C)(C)C=O 2-methyl-2-propan-2-yloxypropanal	114	12	high	MHG (GA)
		COC(C)C=O 2-methoxypropanal	112	11	high	MHG (BO, GA)
polyfunctional (ketone + ether)		COC(C)(C)C(C)=O 3-methoxy-3-methyl-2-butanone	113	11	high	MHG (GA)
		COC(C)C(=O)C(C)(C)C 4-methoxy-2,2-dimethylpentan-3-one	111	12	high	MHG (GA)
acetals		COC(C)(C)OC 2,2-dimethoxypropane	116	14	low	JT (GA)

high-octane, oxygenated hydrocarbons results in increased exhaust emissions (Magnusson and Nilsson, 2011), indicating a lower suitability of aldehydes as fuels. Polyfunctional molecules with an aldehyde and an ether group are generated as well, which also entail the problem of aldehyde emissions. Further polyfunctional molecules containing an ether group and a ketone group are generated, with ketones being prominent high-octane fuels (Hoppe et al., 2016a; Hechinger, 2014). Most of the molecules containing an ether, a ketone, and/or an aldehyde functionality have a compact, branched structure with similarities to MTBE and ETBE, making them interesting high-octane fuel candidates; however, they also have a high price, hindering experimental investigation.

The last top candidate in Table 2, namely 2,2-dimethoxypropane (2,2-DMP), belongs to the class of acetales. It is a compact structure similar to ETBE, with the difference being that one carbon atom is replaced by a second oxygen atom. 2,2-DMP also has a low price, making it an attractive target for experimental investigation. A DCN measurement of 31 is known from literature (Yanowitz et al., 2017) which, however, is not suggestive of a very high RON, as molecules with RON > 110 typically correspond to DCN values below 10, cf. (Dahmen and Marquardt, 2016; Perez and Boehman, 2012). Our high RON + OS prediction (cf. Table 2), however, is consistent with the RON + OS value of 143 stated in a recent study by Li et al. (Li et al., 2022) who used a ML-QSPR prediction model combining both ML and a group contribution approach. Another ML-based QSPR model for RON and OS recently developed by vom Lehm et al. (Vom Lehn et al., 2020) likewise predicts a high RON + OS value of 156.

4.2.2 Comparison to previous fuel design studies

Our commercially available top candidates (cf. Table 2) generally match the molecular classes identified in previous fuel design/screening studies for SI engine fuels, e.g., in (Dahmen and Marquardt, 2016; Li et al., 2022; Hoppe et al., 2016b; Vom Lehn et al., 2021). Specifically, prominent molecular classes from previous studies include the herein identified groups of ethers (Dahmen and Marquardt, 2016; Li et al., 2022; Vom Lehn et al., 2021), ketones (Dahmen and Marquardt, 2016; Li et al., 2022; Hoppe et al., 2016b; Vom Lehn et al., 2021), aromatics (Vom Lehn et al., 2021), aldehydes (Li et al., 2022; Hoppe et al., 2016b), alkanes (Vom Lehn et al., 2021), and acetals (Li et al., 2022). Interestingly, our top candidates do not include any esters, alcohols, and furans that have often been identified in the literature (Dahmen and Marquardt, 2016; Li et al., 2022; Vom Lehn et al., 2021). When inspecting all molecules generated in our design loop runs with $SC_{\#molecs}$ and with AD, we indeed find esters (e.g., methyl acetate), alcohols (e.g., ethanol and methanol), as well as furans (e.g., 2-methylfuran). However, these are not considered top candidates as predicted OS is below 10 for most esters and predicted RON is slightly below 110 in case of furans and alcohols. Such RON and OS predictions are generally in accordance with the literature values for representative molecules of these classes, cf. (McCormick et al., 2017; Schweidtmann et al., 2020a; Derfer et al., 1958; Naegeli et al., 1989; Yanowitz et al., 2011).

The polyfunctional molecules identified in our study are hardly discussed in the literature. It should be noted that the availability of experimental RON and MON values for polyfunctional molecules is very limited, indicating a high uncertainty in the GNN predictions.

The generated top candidate of acetals, 2,2-DMP, has also been identified in the fuel screening by Li et al. (Li et al., 2022) and will be investigated experimentally in the following.

4.2.3 Experimental assessment of 2,2-DMP

Experimental investigation of 2,2-DMP was conducted in dedicated test engines according to the DIN EN ISO 5164 (DIN EN ISO 5164:2014-10, 2014) and DIN EN ISO 5163 standards (DIN EN 5163:2014-10, 2014), respectively, by an external company. Measurement of RON and MON of pure 2,2-DMP, however, could not be performed. Instead, blends of 2,2-DMP with 90 %, 80 %, and 60 % (v/v) of gasoline were investigated. The extrapolation to pure component values yielded a RON of 91.75 (+/- 0.25) and a MON of 87.27 (+/- 0.3), hence a RON + OS score of about 96, indicating a strong misprediction by our GNN model as well as the models by Li et al. (Li et al., 2022) and by vom Lehn et al. (Vom Lehn et al., 2020). To further clarify the ignition properties of 2,2-DMP, we experimentally measured ignition delay times (IDT) in a rapid compression machine (RCM) (Lee et al., 2012; Ramalingam et al., 2017) and compared the chemical reactivity of 2,2-DMP to that of a typical RON95 E10 pump station fuel. IDT measurements for 2,2-DMP were performed at an end-of-compression pressure of 20 bar for a stoichiometric

mixture and with an argon-to-oxygen dilution ratio of 3.762 in the temperature range of 733 to 940 K. Details on the RCM measurements can be found in the ESI. The ignition took place via a two-stage process in the investigated temperature regime indicating strong low-temperature chemistry, cf. Figure 9, not representative for a high-octane fuel. Compared to the RON95 E10 fuel, 2,2-DMP shows a distinctively higher reactivity between 740 and 870 K pointing towards a lower knock resistance and thus RON value. The RCM results suggest a slightly worse knock resistance of 2,2-DMP compared to RON95 E10 pump station fuel, supporting the extrapolated RON and MON measurements.

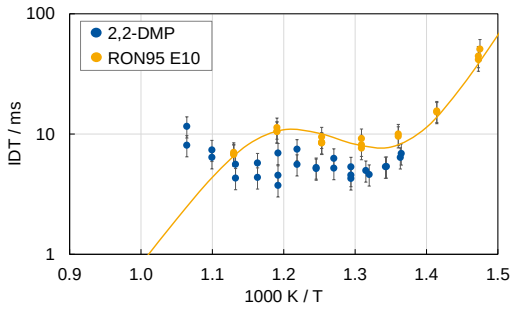


Fig. 9. Measured ignition delay time in a rapid compression machine for 2,2-dimethoxypropane and a commercially available RON95 E10 pump station fuel. The error bars indicate ± 20 % scatter of the measured ignition delay time. The yellow line corresponds to a three-fold Arrhenius model fit to the RON95 E10 ignition delay times (Weisser, 2001).

The case of 2,2-DMP shows the potential weaknesses of a fully data-driven approach in a data-scarce environment. We account the large model prediction error of our GNN as well as those of the models by Li et al. (Li et al., 2022) and by vom Lehn et al. (Vom Lehn et al., 2020) to the comparatively little training data available for RON and MON modeling. Specifically, our RON and MON training database includes just five ethers, a single acetal (not 2,2-DMP), no aldehydes, eight ketones, and only two molecules with more than one type of oxygen functionality (cf. (Schweidtmann et al., 2020a)). Similar data limitations apply to the other RON and MON prediction models (Li et al., 2022; Vom Lehn et al., 2020), explaining their similarly bad predictions in case of 2,2-DMP. Furthermore, we want to stress the fact that no RON and MON values for aldehydes are included in the training data, so our GNN may not sufficiently distinguish between aldehydes and ketones. The RON + OS predictions of the identified molecules with an aldehyde group are therefore considered subject to large uncertainty. In the case of 2,2-DMP, a DCN data point was available and used in the training of our multi-task GNN for simultaneous RON, MON, and DCN prediction (cf. Subsection "Property prediction"). As expected, our AD approach based on majority voting (cf. Subsection "Applicability domain") considers 2,2-DMP within the region of reliable predictions as it was part of the training data. Yet, only 31 out of 40 SVMs voted for 2,2-DMP. Increasing the AD consensus level, e.g., 80% instead of 50%, may provide some protection against such strong mispredictions, at the cost of a smaller search space. A systematic investigation of the relationship between the AD consensus level and the prediction accuracy for molecules proposed by the design loop, however, is beyond the scope of this work. The weak spots of prediction models for fuel ignition quality remain a huge challenge for model-based fuel design, even when utilizing state-of-the-art ML (Schweidtmann et al., 2020a; Vom Lehn et al., 2020) and an applicability domain. Therefore, acquiring more training data is absolutely crucial.

5 Conclusion

We propose a fully data-driven CAMD approach based on recent methods from graph-ML for the identification of molecules with desired ignition characteristics for modern SI engines. Our graph-ML CAMD framework utilizes a representation of molecules as graphs and incorporates three modules for building a molecular design loop: (1) molecule generation from a continuous molecular space with generative graph-ML, (2) molecular property prediction through GNNs, and (3) optimization for strategic sampling from the continuous molecular space to find molecules with high predicted RON + OS. The modular structure enables the exploration of different ML models in combination with different optimization approaches. We additionally present a novel approach to identify the applicability domain (AD) of GNN models for molecular property prediction. By predicting promising high-octane fuel molecules in a fully data-driven fashion, our study exemplifies how recent developments in ML can be utilized for CAMD and its automation.

The top molecular candidates identified with our graph-ML CAMD framework are from well-known molecular classes for high-octane fuels, e.g., ethers and ketones, and include both well-established components like MTBE and ETBE as well as new promising candidates for further experimental investigation. The comparison of different generative graph-ML models, namely JT-VAE (Jin et al., 2018), MHG-VAE (Kajino, 2019), and MolGAN (De Cao and Kipf, 2018), in combination with different optimization approaches, BO and GA, shows that the choice of the generative model and optimization strategy influences the number and type of identified candidate molecules. Both VAEs provide a diverse continuous molecular space with a large number of potential molecules, while MolGAN generates a comparatively low number of candidates and yields lower target property values. We conclude that the GA is well suited for exploring large portions of the continuous molecular space of the generative models, especially when working with high dimensions where BO struggles but still finds some promising candidates. Our AD approach additionally enables us to focus the exploration on candidates with presumably more accurate predictions. The experimental investigation of one candidate within the AD, namely 2,2-dimethoxypropane, shows lower RON and OS values than predicted by our GNN model, demonstrating the limitations of CAMD in a comparatively data-scarce environment. We thereby highlight the importance of experimental validation to fuel design and the need for further RON and OS training data. Furthermore, the correlation between the AD threshold, i.e., the consensus level, and the prediction accuracy for molecules proposed by the design loop should be investigated.

Future work could include additional physical and chemical properties in the design, e.g., melting point, boiling point, vapor pressure, toxicity, or viscosity, similar to previous studies (Dahmen and Marquardt, 2016; Li et al., 2022; Hoppe et al., 2016b; Vom Lehn et al., 2021). The framework, in principle, is not bound to fuel design as application but could also be applied to other CAMD applications such as drug discovery, design of catalysts, pesticides, etc.

Declaration of Competing Interest

We have no conflict of competing interest.

Data Availability Statement

The data that support the findings of this study are openly available in our GitLab repository "Graph machine learning for design of high-octane fuels" at `https://git.rwth-aachen.de/avt-svt/public/graph_ML_fuel_design`, reference number (Rittig et al., 2022).

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 466417970 – within the Priority Programme "SPP 2331: Machine Learning in Chemical Engineering". It was also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - Cluster of Excellence 2186 "The Fuel Science Center". This work was also performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE). Further, this work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the GRK 2236 UnRAVel. Simulations were performed with computing resources granted by RWTH Aachen University under project "rwth0664". The authors thank Florian vom Lehn for providing RON and OS predictions for 2,2-dimethoxypropane by his model. MD received funding from the Helmholtz Association of German Research Centres.

Authors contributions

Jan G. Rittig: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Funding acquisition.

Martin Ritzert: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization.

Artur M. Schweidtmann: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Funding acquisition.

Stefanie Winkler: Methodology, Validation, Software, Formal analysis, Data curation, Writing - Review & Editing.

Jana M. Weber: Methodology, Writing - Review & Editing.

Philipp Morsch: Investigation, Writing - Original Draft, Writing - Review & Editing.

K. Alexander Heufer: Writing - Review & Editing, Supervision, Funding acquisition.

Martin Grohe: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition.

Alexander Mitsos: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition.

Manuel Dahmen: Conceptualization, Formal analysis, Writing - Review & Editing, Supervision.

Bibliography

Abdul-Manan, A. F. N., Kalghatgi, G., and Babiker, H. (2018). Exploring alternative octane specification methods for improved gasoline knock resistance in spark-ignition engines. *Frontiers in Mechanical Engineering*, 4:259.

Achenie, L. E. K., Gani, R., and Venkatasubramanian, V. (2003). *Computer aided molecular design: Theory and practice*, volume 12 of *Computer-aided chemical engineering*. Elsevier, Amsterdam and Boston, 1st ed. edition.

Alshehri, A. S., Gani, R., and You, F. (2020). Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions. *Computers & Chemical Engineering*, 141:107005.

Amer, A., Babiker, H., Chang, J., Kalghatgi, G., Adomeit, P., Brassat, A., and Günther, M. (2012). Fuel effects on knock in a highly boosted direct injection spark ignition engine. *SAE International Journal of Fuels and Lubricants*, 5(3):1048–1065.

American Society for Testing and Materials (2018). ASTM D 2699: Standard test method for research octane number of spark-ignition engine fuel.

American Society for Testing and Materials (2019). ASTM D 2700: Standard test method for motor octane number of spark-ignition engine fuel.

Atz, K., Grisoni, F., and Schneider, G. (2021). Geometric deep learning on molecular representations. arXiv preprint arXiv:2107.12375v4.

Austin, N. D., Sahinidis, N. V., Konstantinov, I. A., and Trahan, D. W. (2018). COSMO-based computer-aided molecular/mixture design: A focus on reaction solvents. *AIChE Journal*, 64(1):104–122.

Austin, N. D., Sahinidis, N. V., and Trahan, D. W. (2016). Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chemical Engineering Research and Design*, 116:2–26.

Austin, N. D., Sahinidis, N. V., and Trahan, D. W. (2017). A COSMO-based approach to computer-aided mixture design. *Chemical Engineering Science*, 159(2):93–105.

Badia, J. H., Ramírez, E., Bringué, R., Cunill, F., and Delgado, J. (2021). New octane booster molecules for modern gasoline composition. *Energy & Fuels*, 35(14):10949–10997.

Bell, A. (2010). Modern SI engine control parameter responses and altitude effects with fuels of varying octane sensitivity. SAE Technical Paper 2010-01-1454.

Bjerrum, E. J. and Sattarov, B. (2018). Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules*, 8(4):1–17.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. (2018). Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, 37(1-2):1700123.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (1996b). Stacked regressions. *Machine Learning*, 24(1):49–64.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478v2.

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108.

Cai, G., Liu, z., and Zhang, L. (2021). Transformation rule-based molecular evolution for automatic gasoline molecule design. Authorea preprint doi:10.22541/au.163976199.99452600/v1.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Chemspace US Inc. (2021). Chemspace [online]. https://chem-space.com. Accessed on 08-09-2021.

Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S., and Jensen, K. F. (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of Chemical Information and Modeling*, 57(8):1757–1772.

Coley, C. W., Eyke, N. S., and Jensen, K. F. (2020). Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie (International ed. in English)*, 59(51):22858–22893.

Dahmen, M., Hechinger, M., Villeda, J. V., and Marquardt, W. (2012). Towards model-based identification of biofuels for compression ignition engines. *SAE International Journal of Fuels and Lubricants*, 5(3):990–1003.

Dahmen, M. and Marquardt, W. (2016). Model-based design of tailor-made biofuels. *Energy & Fuels*, 30(2):1109–1134.

Dai, H., Tian, Y., Dai, B., Skiena, S., and Le Song (2018). Syntax-directed variational autoencoder for structured data. arXiv preprint arXiv:1802.08786.

De Cao, N. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973.

Demirbas, A., Balubaid, M. A., Basahel, A. M., Ahmad, W., and Sheikh, M. H. (2015). Octane rating of gasoline and octane booster additives. *Petroleum Science and Technology*, 33(11):1190–1197.

Derfer, J. M., Boord, C. E., Burk, F. C., Hess, R. E., Lovell, W. G., Randall, R. A., and Sabina, J. R. (1958). *Knocking Characteristics of Pure Hydrocarbons*. ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959.

DIN EN 5163:2014-10 (2014). Petroleum products - Determination of knock characteristics of motor and aviation fuels - Motor method (ISO 5163:2014). German version.

DIN EN ISO 5164:2014-10 (2014). Petroleum products - Determination of knock characteristics of motor fuels - Research method (ISO 5164:2014). German version.

Douguet, D., Munier-Lehmann, H., Labesse, G., and Pochet, S. (2005). LEA3D: A computer-aided ligand design for structure-based drug design. *Journal of Medicinal Chemistry*, 48(7):2457–2468.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., and Chung, P. W. (2019). Deep learning for molecular design - A review of the state of the art. *Molecular Systems Design and Engineering*, 4(4):828–849.

Enamine Ltd (2021). EnamineStore [online]. https://www.enaminestore.com. Accessed on 08-09-2021.

Faez, F., Ommi, Y., Baghshah, M. S., and Rabiee, H. R. (2021). Deep graph generators: A survey. *IEEE Access*, 9:106675–106702.

Felton, K. C., Rittig, J. G., and Lapkin, A. A. (2021). Summit: Benchmarking machine learning methods for reaction optimisation. *Chemistry–Methods*, 1(2):116–122.

Fey, M. and Lenssen, J. E. (2019). Fast graph representation learning with PyTorch Geometric. arXiv preprint arXiv:1903.02428v3.

Flam-Shepherd, D., Wu, T. C., Friederich, P., and Aspuru-Guzik, A. (2021). Neural message passing on high order paths. *Machine Learning: Science and Technology*, 2(4):045009.

Florian Häse, Matteo Aldeghi, Riley J Hickman, Loïc M Roch, Melodie Christensen, Elena Liles, Jason E Hein, and Alán Aspuru-Guzik (2021). Olympus: A benchmarking framework for noisy optimization and experiment planning. *Machine Learning: Science and Technology*, 2(3):035021.

Gani, R. (2004). Chemical product design: challenges and opportunities. *Computers & Chemical Engineering*, 28(12):2441–2457.

Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., Hayter, J. B. R., Vickers, R., Roberts, C., Tang, J., Roblin, D., Blundell, T. L., Bronstein, M. M., and Taylor-King, J. P. (2021). Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6).

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. *34th International Conference on Machine Learning, ICML 2017*, 3:2053–2070.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276.

Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for earning in graph domains. *Proceedings of the International Joint Conference on Neural Networks*, 2:729–734.

Gramatica, P. (2007). Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, 26(5):694–701.

Griffiths, R. R. and Hernández-Lobato, J. M. (2020). Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11(2):577–586.

Gschwend, D., Soltic, P., Wokaun, A., and Vogel, F. (2019). Review and performance evaluation of fifty alternative liquid fuels for spark-ignition engines. *Energy & Fuels*, 33(3):2186–2196.

Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. (2018). Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv preprint arXiv:1705.10843v3.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 1024–1034.

Harper, P. M., Gani, R., Kolar, P., and Ishikawa, T. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, 158-160(6/7):337–347.

Hechinger, M. (2014). *Model-based identification of promising biofuel candidates for spark-ignited engines*. PhD thesis, RWTH Aachen University, Germany.

Hechinger, M., Voll, A., and Marquardt, W. (2010). Towards an integrated design of biofuels and their production pathways. *Computers & Chemical Engineering*, 34(12):1909–1918.

Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., and Coley, C. W. (2020). Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780.

Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267(1):66–73.

Hoppe, F., Burke, U., Thewes, M., Heufer, A., Kremer, F., and Pischinger, S. (2016a). Tailor-made fuels from biomass: Potentials of 2-butanone and 2-methylfuran in direct injection spark ignition engines. *Fuel*, 167(3):106–117.

Hoppe, F., Heuser, B., Thewes, M., Kremer, F., Pischinger, S., Dahmen, M., Hechinger, M., and Marquardt, W. (2016b). Tailor-made fuels for future engine concepts. *International Journal of Engine Research*, 17(1):16–27.

IEA (2020). The role of CCUS in low-carbon power systems. Technical report. https://www.iea.org/reports/the-role-of-ccus-in-low-carbon-power-systems. Accessed on 08-09-2021.

Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. *Alternatives to Laboratory Animals: ATLA*, 33(5):445–459.

Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. *35th International Conference on Machine Learning, ICML 2018*, 5:3632–3648.

Joback, K. G. (1989). *Designing molecules possessing desired physical property values*. PhD thesis, Massachusetts Institute of Technology, USA.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A., and Zhavoronkov, A. (2017). DruGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, 14(9):3098–3104.

Kajino, H. (2019). Molecular hypergraph grammar with its application to molecular optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3183–3191. PMLR.

Kalghatgi, G. (2001). Fuel anti-knock quality - part i. engine studies. *SAE Transactions*, 110:1993–2004.

Kalghatgi, G. T. (2005). Auto-ignition quality of practical fuels and implications for fuel requirements of future SI and HCCI engines. SAE Technical Paper 2005-01-0239.

Kalghatgi, G. T. (2014). The outlook for fuels for internal combustion engines. *International Journal of Engine Research*, 15(4):383–398.

Karunanithi, A. T. and Mehrkesh, A. (2013). Computer-aided design of tailor-made ionic liquids. *AIChE Journal*, 59(12):4627–4640.

Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., and Dobchev, D. A. (2010). Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction. *Chemical Reviews*, 110(10):5714–5789.

Kirschner, J., Mutny, M., Hiller, N., Ischebeck, R., and Krause, A. (2019). Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3429–3438. PMLR.

Klicpera, J., Groß, J., and Günnemann, S. (2020). Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123.

König, A., Neidhardt, L., Viell, J., Mitsos, A., and Dahmen, M. (2020). Integrated design of processes and products: Optimal renewable fuels. *Computers & Chemical Engineering*, 134:106712.

König, A., Siska, M., Schweidtmann, A. M., Rittig, J. G., Viell, J., Mitsos, A., and Dahmen, M. (2021). Designing production-optimal alternative fuels for conventional, flexible-fuel, and ultra-high efficiency engines. *Chemical Engineering Science*, 237:116562.

Kubic, W. L., Jenkins, R. W., Moore, C. M., Semelsberger, T. A., and Sutton, A. D. (2017). Artificial neural network based group contribution method for estimating cetane and octane numbers of hydrocarbons and oxygenated organic compounds. *Industrial & Engineering Chemistry Research*, 56(42):12236–12245.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1945–1954. PMLR.

Lampe, M., Stavrou, M., Schilling, J., Sauer, E., Gross, J., and Bardow, A. (2015). Computer-aided molecular design in the continuous-molecular targeting framework using group-contribution PC-SAFT. *Computers & Chemical Engineering*, 81(4):278–287.

Landrum, G. (2021). RDKit: Open-source cheminformatics software. http://www.rdkit.org. Accessed on 08-09-2021.

Lee, C., Vranckx, S., Heufer, K. A., Khomik, S. V., Uygun, Y., Olivier, H., and Fernandez, R. X. (2012). On the chemical kinetics of ethanol oxidation: Shock tube, rapid compression machine and detailed modeling study. *Zeitschrift für Physikalische Chemie*, 226(1):1–28.

Leitner, W., Klankermayer, J., Pischinger, S., Pitsch, H., and Kohse-Höinghaus, K. (2017). Advanced biofuels and beyond: Chemistry solutions for propulsion and production. *Angewandte Chemie (International ed. in English)*, 56(20):5412–5452.

Leppard, W. R. (1991). The autoignition chemistries of octane-enhancing ethers and cyclic ethers: A motored engine study. *SAE Transactions*, pages 589–604.

Li, R., Herreros, J. M., Tsolakis, A., and Yang, W. (2022). Integrated machine learning-quantitative structure property relationship (ML-QSPR) and chemical kinetics for high throughput fuel screening toward internal combustion engine. *Fuel*, 307(1):121908.

Lim, J., Ryu, S., Kim, J. W., and Kim, W. Y. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, 10(1):31.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528.

Liu, Q., Zhang, L., Liu, L., Du, J., Tula, A. K., Eden, M., and Gani, R. (2019). OptCAMD: An optimization-based framework and tool for molecular and mixture product design. *Computers & Chemical Engineering*, 124(1):285–301.

Liu, R., Liu, R., Liu, Y., Wang, L., Zhang, X., and Li, G. (2022). Design of fuel molecules based on variational autoencoder. *Fuel*, 316(2):123426.

Lo, Y.-C., Rensi, S. E., Torng, W., and Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546.

Lunderman, S., Fioroni, G. M., McCormick, R. L., Nimlos, M. R., Rahimi, M. J., and Grout, R. W. (2018). Screening fuels for autoignition with small-volume experiments and gaussian process classification. *Energy & Fuels*, 32(9):9581–9591.

Magnusson, R. and Nilsson, C. (2011). The influence of oxygenated fuels on emissions of aldehydes and ketones from a two-stroke spark ignition engine. *Fuel*, 90(3):1145–1154.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

McCormick, R. L., Fioroni, G., Fouts, L., Christensen, E., Yanowitz, J., Polikarpov, E., Albrecht, K., Gaspar, D. J., Gladden, J., and George, A. (2017). Selection criteria and screening of potential biomass-derived streams as fuel blendstocks for advanced spark-ignition engines. *SAE International Journal of Fuels and Lubricants*, 10(2):442–460.

Mercado, R., Rastemo, T., Lindelöf, E., Klambauer, G., Engkvist, O., Chen, H., and Jannik Bjerrum, E. (2021). Graph networks for molecular design. *Machine Learning: Science and Technology*, 2(2):025023.

Merck KGaA (2021). Sigma-Aldrich [online]. https://www.sigmaaldrich.com. Accessed on 08-09-2021.

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, 4(5):468–481.

Mittal, V. and Heywood, J. B. (2008). The relevance of fuel RON and MON to knock onset in modern SI engines. SAE Technical Paper 2008-01-2414.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2019). Weisfeiler and leman go neural: Higher-order graph neural networks. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 4602–4609.

Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtarolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., and Tropsha, A. (2020). QSAR without borders. *Chemical Society reviews*, 49(11):3525–3564.

Naegeli, D. W., Yost, D. M., Moulton, D. S., Owens, E. C., and Chui, G. K. (1989). The measurement of octane numbers for methanol and reference fuels blends. *SAE Transactions*, pages 712–722.

Ng, L. Y., Chong, F. K., and Chemmangattuvalappil, N. G. (2015). Challenges and opportunities in computer-aided molecular design. *Computers & Chemical Engineering*, 81:115–129.

Nigam, A., Pollice, R., Hurley, M. F. D., Hickman, R. J., Aldeghi, M., Yoshikawa, N., Chithrananda, S., Voelz, V. A., and Aspuru-Guzik, A. (2021). Assigning confidence to molecular property prediction. *Expert Opinion on Drug Discovery*, 16(9):1009–1023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perez, P. L. and Boehman, A. L. (2012). Experimental investigation of the autoignition behavior of surrogate gasoline fuels in a constant-volume combustion bomb apparatus and its relevance to HCCI combustion. *Energy & Fuels*, 26(10):6106–6117.

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., and Zhavoronkov, A. (2020). Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644.

Prykhodko, O., Johansson, S. V., Kotsias, P. C., Arús-Pous, J., Bjerrum, E. J., Engkvist, O., and Chen, H. (2019). A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11(1):1–13.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022.

Ramalingam, A., Zhang, K., Dhongde, A., Virnich, L., Sankhla, H., Curran, H., and Heufer, A. (2017). An RCM experimental and modeling study on CH4 and CH4/C2H6 oxidation at pressures up to 160 bar. *Fuel*, 206:325–333.

Rittig, J. G., Ritzert, M., Schweidtmann, A. M., Winkler, S., Weber, J. M., Morsch, P., Heufer, K. A., Grohe, M., Mitsos, A., and Dahmen, M. (2022). Open-source graph machine learning for design of high-octane fuels. https://git.rwth-aachen.de/avt-svt/public/graph_ML_fuel_design. Accessed on 03-06-2022.

Royal Society of Chemistry (2021). Chemspider [online]. http://www.chemspider.com. Accessed on 08-09-2021.

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Samudra, A. P. and Sahinidis, N. V. (2013). Optimization-based framework for computer-aided molecular design. *AIChE Journal*, 59(10):3686–3701.

Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). Computational capabilities of graph neural networks. *IEEE Transactions on Neural Networks*, 20(1):81–102.

Schilling, J., Gross, J., and Bardow, A. (2017). Integrated design of ORC process and working fluid using process flowsheeting software and PC-SAFT. *Energy Procedia*, 129(10):129–136.

Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A., and Müller, K. R. (2018). SchNet - A deep learning architecture for molecules and materials. *Journal of Chemical Physics*, 148(24):1–11.

Schweidtmann, A. M., Bongartz, D., Grothe, D., Kerkenhoff, T., Lin, X., Najman, J., and Mitsos, A. (2021a). Deterministic global optimization with gaussian processes embedded. *Mathematical Programming Computation*, 13(3):553–581.

Schweidtmann, A. M., Clayton, A. D., Holmes, N., Bradford, E., Bourne, R. A., and Lapkin, A. A. (2018). Machine learning meets continuous flow chemistry: Automated optimization towards the pareto front of multiple objectives. *Chemical Engineering Journal*, 352(1–9):277–282.

Schweidtmann, A. M. and Mitsos, A. (2019). Deterministic global optimization with artificial neural networks embedded. *Journal of Optimization Theory and Applications*, 180(3):925–948.

Schweidtmann, A. M., Rittig, J. G., König, A., Grohe, M., Mitsos, A., and Dahmen, M. (2020a). Graph neural networks for prediction of fuel ignition quality. *Energy & Fuels*, 34(9):11395–11407.

Schweidtmann, A. M., Rittig, J. G., König, A., Grohe, M., Mitsos, A., and Dahmen, M. (2020b). Supporting information: Graph neural networks for prediction of fuel ignition quality. *Energy & Fuels*, 34(9):11395–11407.

Schweidtmann, A. M., Weber, J. M., Wende, C., Netze, L., and Mitsos, A. (2021b). Obey validity limits of data-driven models through topological data analysis and one-class classification. *Optimization and Engineering*, pages 1–22.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences*, 44(6):1912–1928.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., and Adams, R. (2015). Scalable Bayesian optimization using deep neural networks. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2171–2180, Lille, France. PMLR.

Soleimany, A. P., Amini, A., Goldman, S., Rus, D., Bhatia, S. N., and Coley, C. W. (2021). Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367.

Solgi, R. M. (2020). geneticalgorithm (python package). https://github.com/rmsolgi/geneticalgorithm. Accessed on 08-09-2021.

Sperduti, A. and Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735.

SynQuest Laboratories, Inc. (2021). SynQuest Labs [online]. http://synquestlabs.com. Accessed on 08-09-2021.

Szybist, J. P. and Splitter, D. A. (2017). Pressure and temperature effects on fuels with varying octane sensitivity at high load in SI engines. *Combustion and Flame*, 177:49–66.

The GPyOpt authors (2016). GPyOpt: A bayesian optimization framework in python. http://github.com/SheffieldML/GPyOpt. Accessed on 08-09-2021.

The Luigi authors (2012). Luigi (python package). https://github.com/spotify/luigi. Accessed on 08-09-2021.

Tropsha, A., Gramatica, P., and Gombar, V. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science*, 22(1):69–77.

Unke, O. T. and Meuwly, M. (2019). PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15(6):3678–3693.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks. 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

Vom Lehn, F., Brosius, B., Broda, R., Cai, L., and Pitsch, H. (2020). Using machine learning with target-specific feature sets for structure-property relationship modeling of octane numbers and octane sensitivity. *Fuel*, 281:118772.

Vom Lehn, F., Cai, L., Tripathi, R., Broda, R., and Pitsch, H. (2021). A property database of fuel compounds with emphasis on spark-ignition engine applications. *Applications in Energy and Combustion Science*, 5(12):100018.

Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. (2018). Batched large-scale Bayesian optimization in high-dimensional spaces. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 745–754. PMLR.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Feitas, N. (2016). Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387.

Weaver, S. and Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics & Modelling*, 26(8):1315–1326.

Weininger, D. (1988). SMILES, A chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36.

Weisser, G. A. (2001). *Modelling of combustion and nitric oxide formation for medium-speed DI diesel engines: A comparative evaluation of zero- and three-dimensional approaches*. PhD thesis, ETH Zurich, Switzerland.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4(2).

Whitmore, L. S., Davis, R. W., McCormick, R. L., Gladden, J. M., Simmons, B. A., George, A., and Hudson, C. M. (2016). BioCompoundML: A general biofuel property screening tool for biological molecules using random forest classifiers. *Energy & Fuels*, 30(10):8410–8418.

Winter, R., Montanari, F., Noé, F., and Clevert, D.-A. (2019a). Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science*, 10(6):1692–1701.

Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., and Clevert, D.-A. (2019b). Efficient multi-objective molecular optimization in

a continuous latent space. *Chemical Science*, 10(34):8016–8024.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Xia, X., Hu, J., Wang, Y., Zhang, L., and Liu, Z. (2019). Graph-based generative models for de novo drug design. *Drug Discovery Today. Technologies*, 32-33:45–53.

Xiong, J., Xiong, Z., Chen, K., Jiang, H., and Zheng, M. (2021). Graph neural networks for automated de novo drug design. *Drug Discovery Today*, 26(6):1382–1393.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? arXiv preprint arXiv:1810.00826v3.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388.

Yanowitz, J., Christensen, E., and McCormick, R. L. (2011). Utilization of renewable oxygenates as gasoline blending components (technical report nrel/tp-5400-50791). Technical report, National Renewable Energy Laboratory (NREL), Golden, CO, United States.

Yanowitz, J., Ratcliff, M. A., McCormick, R. L., Taylor, J. D., and Murphy, M. J. (2017). Compendium of experimental cetane numbers (Technical Report NREL/TP-5400-67585). Technical report, National Renewable Energy Laboratory (NREL), Golden, CO, United States.

You, J., Liu, B., Ying, R., Pande, V., and Leskovec, J. (2018). Graph convolutional policy network for goal-directed molecular graph generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6412–6422, Red Hook, NY, USA. Curran Associates Inc.

Zhang, L., Cignitti, S., and Gani, R. (2015). Generic mathematical programming formulation and solution for computer-aided molecular design. *Computers & Chemical Engineering*, 78:79–84.

Zhang, S., Liu, Y., and Xie, L. (2020). Molecular mechanics-driven graph neural network with multiplex graph for molecular structures. arXiv preprint arXiv:2011.07457.

Zhang, Z., Cui, P., and Zhu, W. (2022). Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1(1):57–81.

Zhou, Z., Kearnes, S., Li, L., Zare, R. N., and Riley, P. (2019). Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752.

Electronic Supporting Information: Graph Machine Learning for Design of High-Octane Fuels

Jan G. Rittig[a*], Martin Ritzert[b*], Artur M. Schweidtmann[c], Stefanie Winkler[d], Jana M. Weber[e], Philipp Morsch[f], K. Alexander Heufer[f], Martin Grohe[d], Alexander Mitsos[h,a,g], Manuel Dahmen[g†]

[a] RWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen 52074, Germany
[b] Aarhus University, Department of Computer Science, Aarhus, Denmark
[c] Delft University of Technology, Department of Chemical Engineering, Delft 2629 HZ, The Netherlands
[d] RWTH Aachen University, Chair of Computer Science 7, Aachen 52074, Germany
[e] Delft University of Technology, Delft Bioinformatics Lab, Intelligent Systems, 2628 XE Delft, The Netherlands
[f] RWTH Aachen University, Chair of High Pressure Gas Dynamics, Aachen 52074, Germany
[g] Forschungszentrum Jülich GmbH, Institute for Energy and Climate Research IEK-10: Energy Systems Engineering, Jülich 52425, Germany
[h] JARA-ENERGY, Aachen 52056, Germany

arXiv:2206.00619v2 [cs.LG] 14 Oct 2022

1 Applicability domain for graph neural networks

We apply our recently published applicability domain (AD) approach (Schweidtmann et al., 2021) to graph neural networks (GNNs). Specifically, we determine the AD of a GNN in the GNN's molecular fingerprint space by using a one-class support vector machine (SVM) to identify those molecular fingerprints that correspond to molecules for which our GNN tool presumably provides reliable predictions. Using the molecular fingerprint vectors means that the SVMs are trained on fixed-size inputs, circumventing the need for handling the varying input size of molecular graphs. Note that the dimension of the molecular fingerprint vector is a hyperparameter of the GNN and thus fixed before the training of the GNN.

For the training step of the AD, we extract the molecular fingerprints of the training molecules of the GNN and use those fingerprints as the training set for the one-class SVM. We use an one-class SVM to determine the AD since we only have positive training samples, i.e., the fingerprints of the molecules we used for training the GNN. After training, the SVM classifies predictions as non-reliable for every molecule that is not similar to the training data. Technically, we select a linear classifier SVM_{AD} with $SVM_{\mathrm{AD}}(\mathbf{h}_{\mathrm{FP,train}}) \geq 0$ where $\mathbf{h}_{\mathrm{FP,train}}$ is a molecular fingerprint belonging to a molecule with its graph $G_{\mathrm{mol,train}}$ seen during training, i.e., $\mathbf{h}_{\mathrm{FP,train}} = g_{\mathrm{GNN}}(G_{\mathrm{mol,train}})$. After training, a property prediction for a molecule is classified as reliable or non-reliable by first computing the corresponding molecular fingerprint through the GNN and then evaluating this fingerprint with the SVM. A molecule is classified as reliable if its molecular fingerprint $\mathbf{h}_{\mathrm{FP}}$ results in a non-negative value, i.e., $SVM_{\mathrm{AD}}(\mathbf{h}_{\mathrm{FP}}) \geq 0$.

Since our GNN is based on ensemble learning (EL), i.e., the predictions of 40 GNNs are averaged to get the final prediction, we train 40 one-class SVMs in total; one SVM for each GNN. We then apply a majority vote to determine if a prediction lies within the AD or not (cf. Subsection "Applicability domain" in the main text).

2 Hyperparameters of graph-ML CAMD framework

Generator and prediction models

For all three generator models, i.e., JT-VAE (Jin et al., 2018), MHG-VAE (Kajino, 2019), MolGAN (De Cao and Kipf, 2018), and the GNN prediction model for fuel ignition quality (Schweidtmann et al., 2020), we use the hyperparameter configuration as provided in the original publications and corresponding code repositories.

Bayesian optimization

For BO, we follow the optimization procedure of the MHG-VAE study by Kajino (Kajino, 2019) and select the default parameters of GPyOpt (The GPyOpt authors, 2016) using Gaussian Process models with the Matern 5/2 kernel as a surrogate model. The Gaussian Process surrogate models are initialized with 10 molecules randomly sampled from QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014). We apply expected improvement for the acquisition function and optimize it with L-

* Equally contributed
† Corresponding author, E-mail: m.dahmen@fz-juelich.de

BFGS (Liu and Nocedal, 1989). In contrast to the MHG study (Kajino, 2019), we use Thompson sampling (Thompson, 1933; Russo et al., 2018) with a batch size of 10 to support exploration. For the two VAEs, we additionally apply principal component analysis (PCA) from scikit-learn (Pedregosa et al., 2011) trained on the latent vectors of all QM9 molecules, reducing the dimensionality for the MHG-VAE from 72 to 41 and for the JT-VAE from 56 to 38 which in both cases maintains an explained variance ratio of 99.9%.

Genetic algorithm

When using GA for optimization, we apply the default parameters of the package *geneticalgorithm* (Solgi, 2020), i.e., mutation and crossover probability of 0.1 and 0.5, respectively, elite ratio of 0.01 and parents portion of 0.3. However, we reduce the population size from 100 to 50 to increase the number of evolutionary steps per unit computational time and thus exploration.

Applicability domain

Our implementation of the SVMs for determining the applicability domain (AD) of graph neural networks follows the implementation of Schweidtmann et al. (Schweidtmann et al., 2021), i.e., we use the class *OneClassSVM* of the Python package scikit-learn (Pedregosa et al., 2011). We apply the default hyperparameter settings, except for the kernel coefficient γ used for radial basis funcion kernel of the SVM and the parameter ν that determines the maximal fraction of training points classified as outliers and a minimum fraction of support vectors. We determine γ and ν through a grid search with $\gamma \in \{0.5, 0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001, \text{scale}\}$ and $\nu \in \{0.5, 0.1, 0.05, 0.01\}$. Note that when using $\gamma = \text{scale}$ (default in sklearn), the value for γ is automatically selected for each one-class SVM by scikit-learn by multiplying the inverse of the fingerprint dimension with the inverse of the variance in all training molecular fingerprints. To identify γ and ν, we gradually decrease γ while testing different values for ν until the number of support vectors does not decrease much anymore (Dreiseitl et al., 2010; Schweidtmann et al., 2021). We find the default scaling option within skicit-learn for γ to work well and values for ν below 0.05 to not decrease the number of support vectors significantly and thus select $\gamma = 0.00086$ and $\nu = 0.05$ values as final hyperparameters.

Fuel design loop runs

For all design loop runs, we set the lower and upper bounds of the search space for the optimization to the minimum and the maximum entries of the latent vectors of the generative models, respectively, expanded by 20% of the difference between maximum and minimum to allow some extrapolation. The minimum and maximum entries of the latent vectors in the generative models are determined based on the about 50,000 HCO-molecules within the QM9 data set (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) for the two VAEs, JT-VAE and MHG-VAE, and 50,000 samples from a standard normal distribution for MolGAN (cf. (De Cao and Kipf, 2018)). In addition, we set a maximum time limit of 10 seconds for decoding a latent vector to a molecular graph since we experienced rare cases of very long decoding times with the JT-VAE. The time limit of 10 seconds corresponds to about 10 times the maximum decoding time of 95 % of the molecules in QM9 with the JT-VAE, which we found to have the longest decoding times among the generators.

3 QM9 statistics

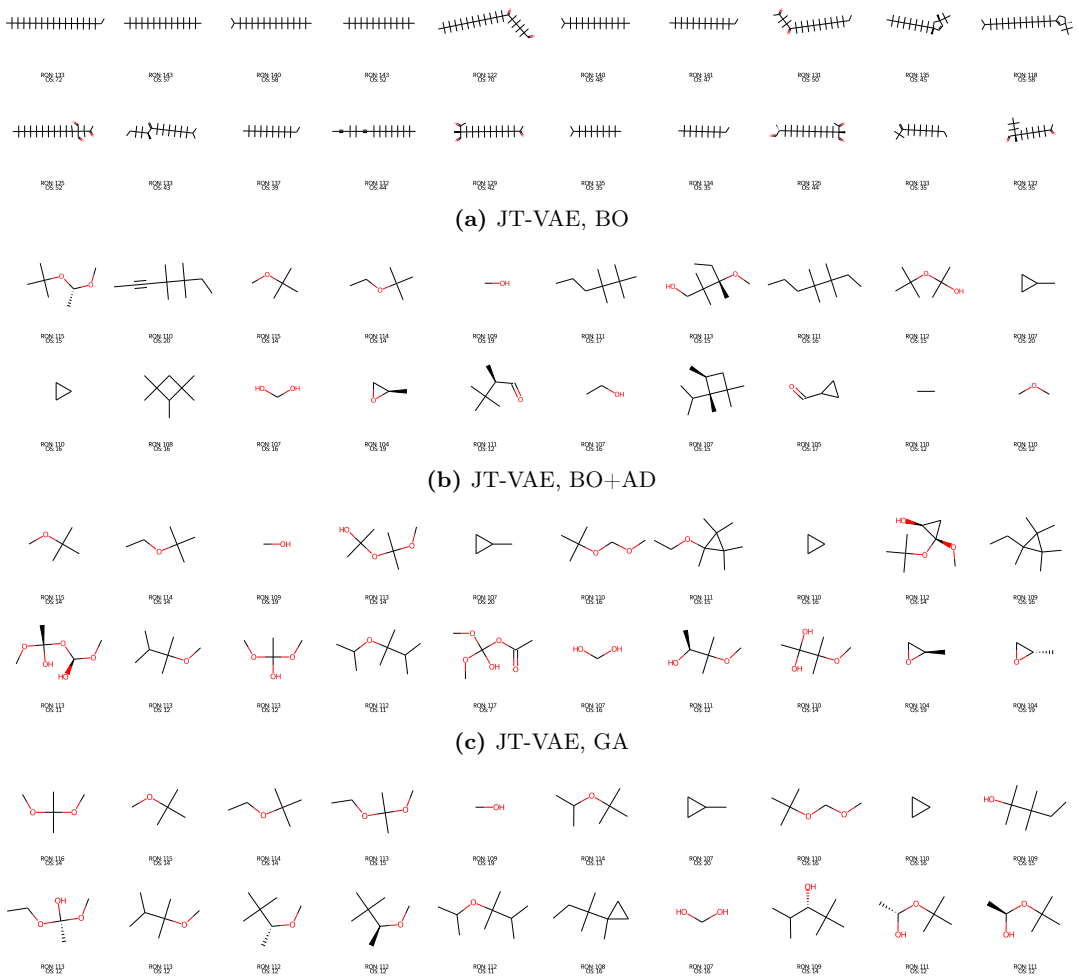
The QM9 (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) molecule database that we use for the training of the generative models also includes molecules with high predicted RON and OS (cf. Table 1). The maximum RON + OS score in QM9 is 135 and the mean score of the top 20 molecules is 129. When applying the AD, the GNN predictions for about 3,250 out of the total 50,150 hydrocarbons within QM9 are classified as unreliable and are thus omitted, yielding a reduced maximum score of 131 and mean score of 128.

predicted RON+OS	QM9	QM9+AD
max	135	131
mean top 20	129	128
# unique mol.	50150	46902
# promising mol.	63	51

Tab. 1. Promising molecules (both RON > 110 and OS > 10) in the QM9 data set with and without considering the applicability domain (AD).

4 Detailed CAMD results

Figures 1-3 show the top 20 molecules with regard to predicted RON + OS that are identified in the fuel design loop runs with the three generator models (JT-VAE (Jin et al., 2018), MHG-VAE (Kajino, 2019), MolGAN (De Cao and Kipf, 2018)) and the two optimization approaches (BO and GA) with and without AD and with stopping criterion $SC_{\#molecs}$ (max. 1000 unique molecules, max. 2000 total molecules). Figures 4-6 show the top 20 molecules with regard to predicted RON + OS that are identified in the fuel design loop runs with the respective generator and optimization approaches with and without AD and with stopping criterion SC_{time} (12 hours run time). Figure 7 illustrates the top 20 molecules with regard to predicted RON + OS in the QM9 dataset (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014), for both with and without AD.



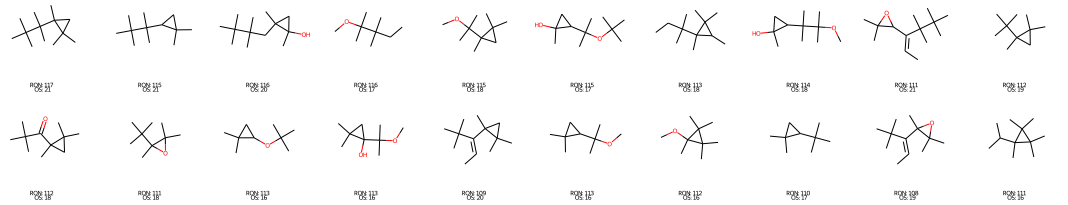
(a) JT-VAE, BO

(b) JT-VAE, BO+AD

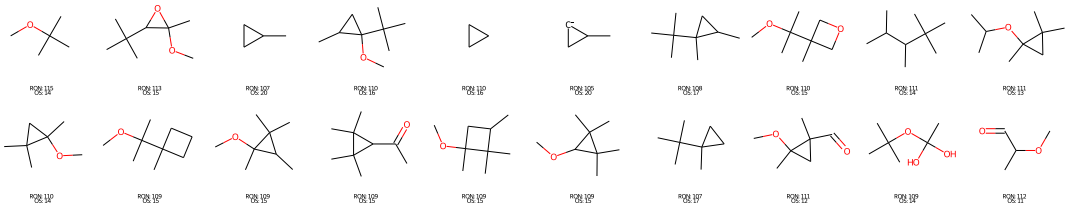
(c) JT-VAE, GA

(d) JT-VAE, GA+AD

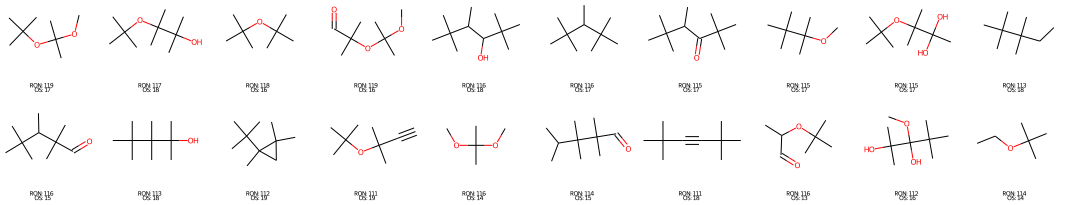
Fig. 1. Top 20 candidates identified in fuel design loop runs with the JT-VAE (Jin et al., 2018) model, with BO and GA, without and with applicability domain, and with stopping criterion $SC_{\#molecs}$ (max. 1000 unique molecules, max. 2000 total molecules). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



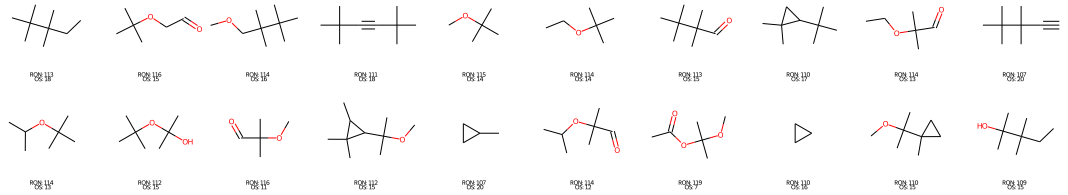
(a) MHG-VAE, BO



(b) MHG-VAE, BO+AD

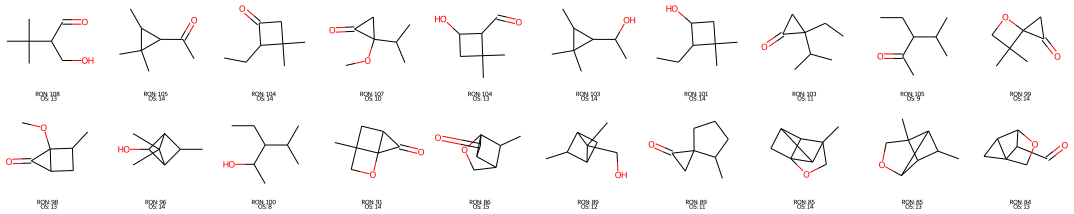


(c) MHG-VAE, GA

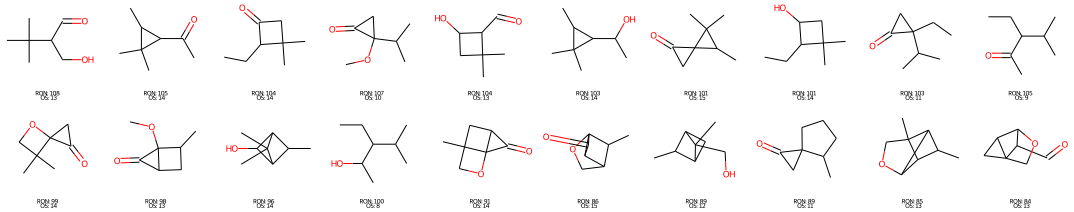


(d) MHG-VAE, GA+AD

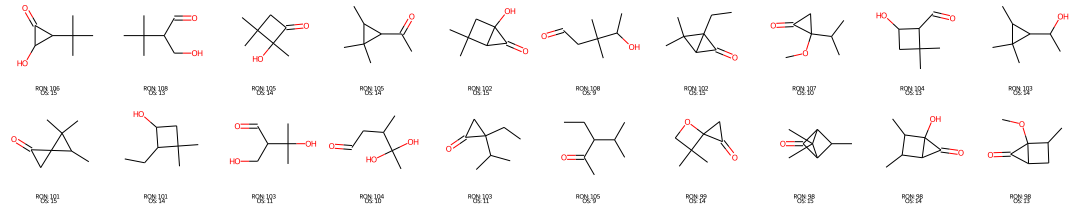
Fig. 2. Top 20 candidates identified in fuel design loop runs with the MHG-VAE (Kajino, 2019) model, with BO and GA, without and with applicability domain, and with stopping criterion $SC_{\#molecs}$ (max. 1000 unique molecules, max. 2000 total molecules). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



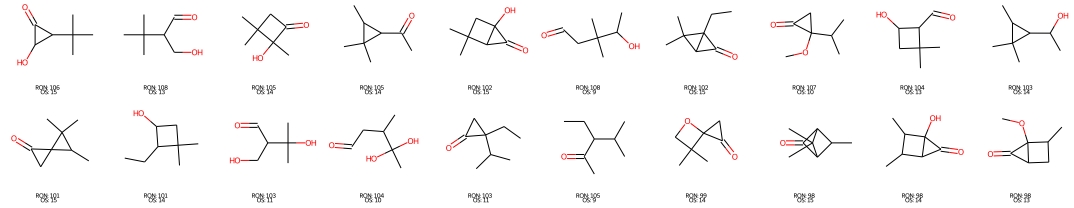
(a) MolGAN, BO



(b) MolGAN, BO+AD

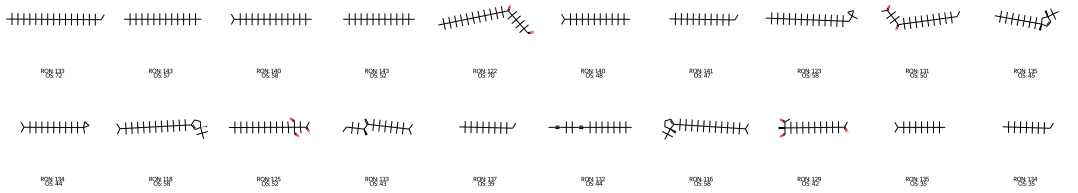


(c) MolGAN, GA

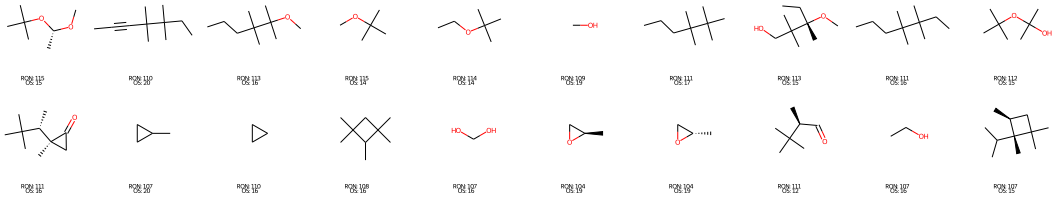


(d) MolGAN, GA+AD

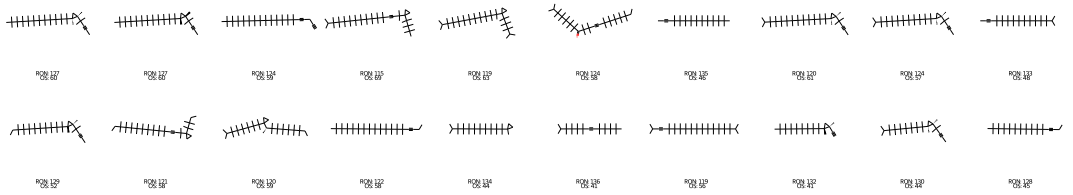
Fig. 3. Top 20 candidates identified in fuel design loop runs with the MolGAN (De Cao and Kipf, 2018) model, with BO and GA, without and with applicability domain, and with stopping criterion $SC_{\#molecs}$ (max. 1000 unique molecules, max. 2000 total molecules). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



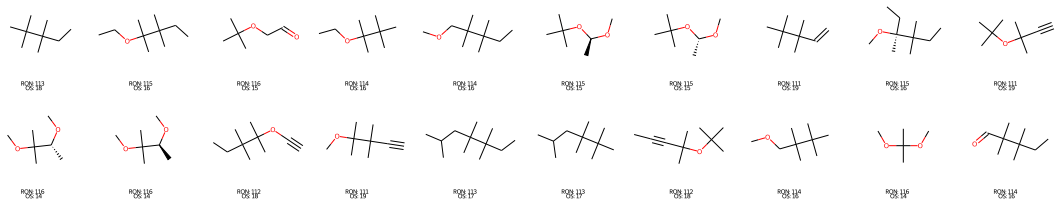
(a) JT-VAE, BO



(b) JT-VAE, BO+AD

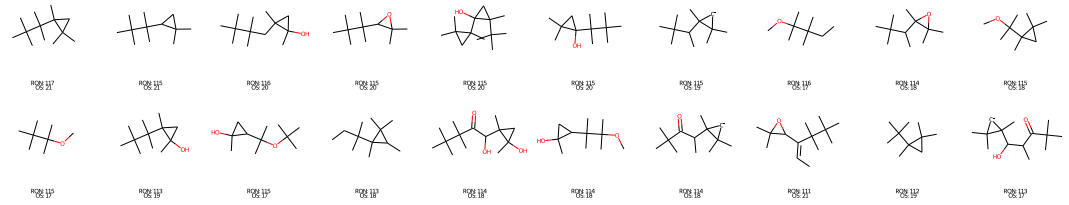


(c) JT-VAE, GA

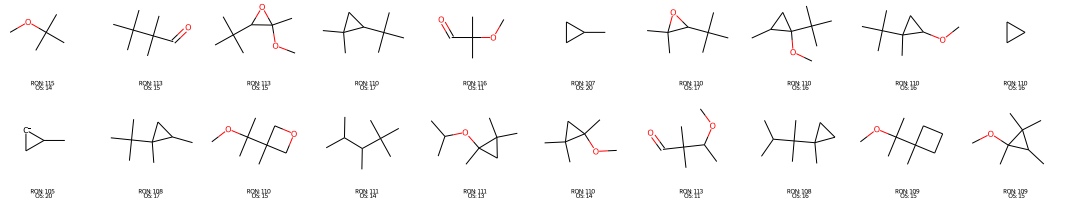


(d) JT-VAE, GA+AD

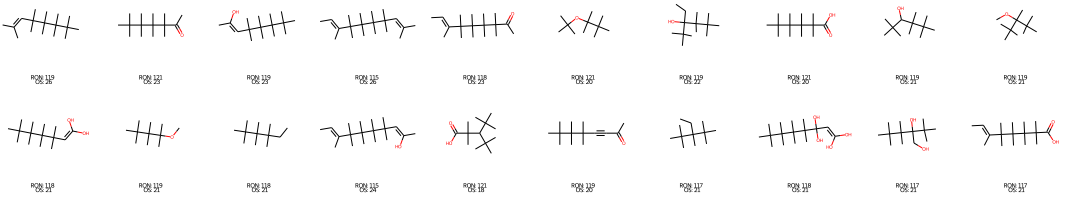
Fig. 4. Top 20 candidates identified in fuel design loop runs with the JT-VAE (Jin et al., 2018) model, with BO and GA, without and with applicability domain, and with stopping criterion SC_{time} (12 hours run time). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



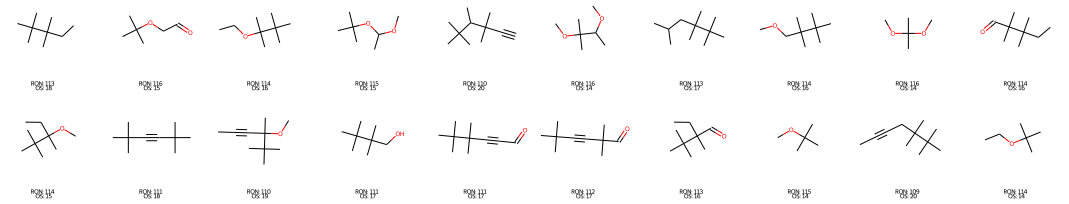
(a) MHG-VAE, BO



(b) MHG-VAE, BO+AD

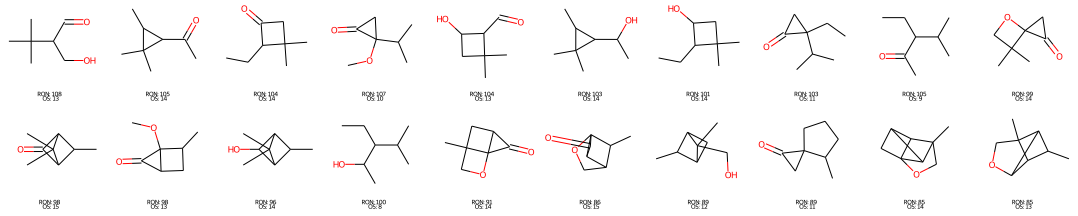


(c) MHG-VAE, GA

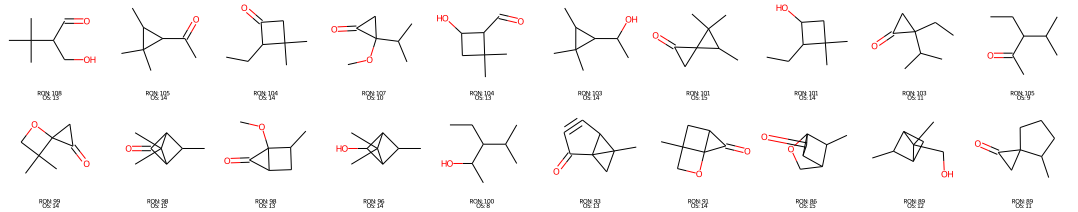


(d) MHG-VAE, GA+AD

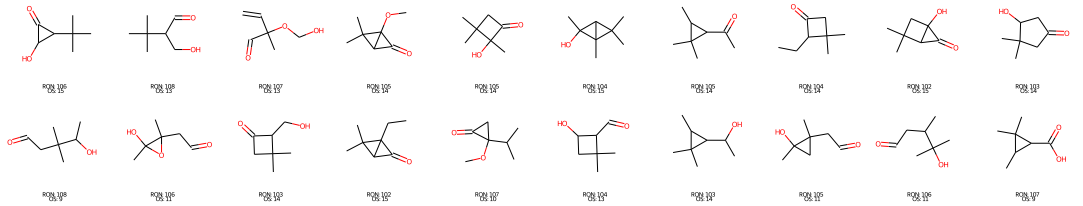
Fig. 5. Top 20 candidates identified in fuel design loop runs with the MHG-VAE (Kajino, 2019) model, with BO and GA, without and with applicability domain, and with stopping criterion SC_{time} (12 hours run time). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



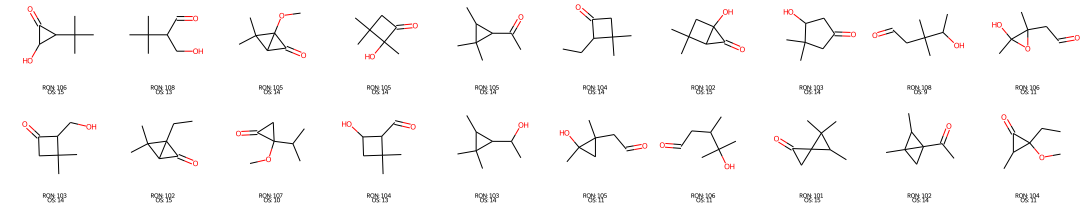
(a) MolGAN, BO



(b) MolGAN, BO+AD

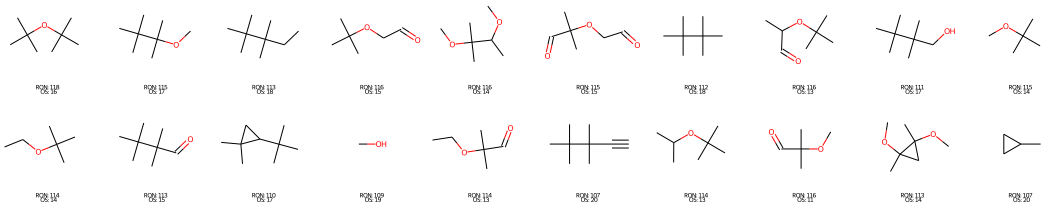


(c) MolGAN, GA

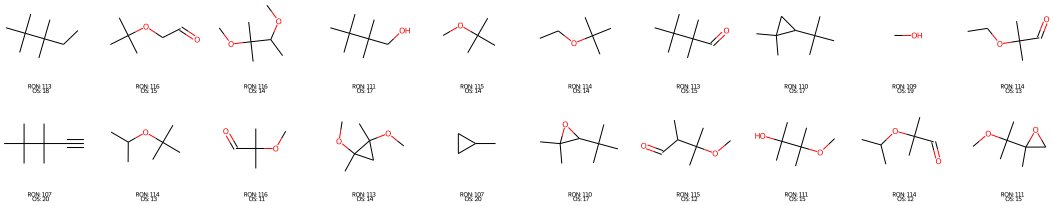


(d) MolGAN, GA+AD

Fig. 6. Top 20 candidates identified in fuel design loop runs with the MolGAN (De Cao and Kipf, 2018) model, with BO and GA, without and with applicability domain, and with stopping criterion SC_{time} (12 hours run time). RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).



(a) QM9



(b) QM9+AD

Fig. 7. Top 20 molecular candidates with regard to predicted RON + OS within the QM9 data set (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014) without and with applicability domain. RON and OS values are predicted by the graph neural network (Schweidtmann et al., 2020).

5 Rapid compression machine screening

Ignition delay times were measured in our rapid compression machine (RCM) at RWTH Aachen University. The RCM has been presented in detail by Lee et al. (Lee et al., 2012) and Ramalingam et al. (Ramalingam et al., 2017). Briefly, the RCM has a stainless steel reaction chamber with an inner diameter of 50 mm and a total stroke length of 250 mm. Both the manifold system with the two mixture storage tanks and the reaction chamber were heated to 75 °C during the experimental series to avoid local condensation issues. A Kistler pressure transducer (6125C-U20) was used to determine the end-of-compression (EOC) pressure and to detect the ignition event, which was defined by the maximum pressure gradient with respect to time. The time interval between these events was defined as the ignition delay time, cf. Figure 8. In order to change the compression ratio and thus EOC temperature, which was calculated under the assumption of an isentropic compression inside the core gas of the reactor, a moveable endwall was used. The measurement uncertainties for the EOC conditions have been estimated to be ± 5 K for the EOC temperature and ± 0.15 bar for the EOC pressure, respectively (Ramalingam et al., 2017). The EOC pressures of 20 and 40 bar were set by adjusting the initial pressure in the reaction chamber accordingly. A dilution of 3.762 was achieved by diluting a stoichiometric mixture of 2,2-dimethoxypropane (Alfa Aesar, 99.7 % purity) and oxygen (grade 5.0) with argon (grade 5.0) ensuring a constant mixture composition for the measurable temperature regime. The expected scatter of ± 20 % (Ramalingam et al., 2017) is indicated by the error bars in the illustration of the experimental results.

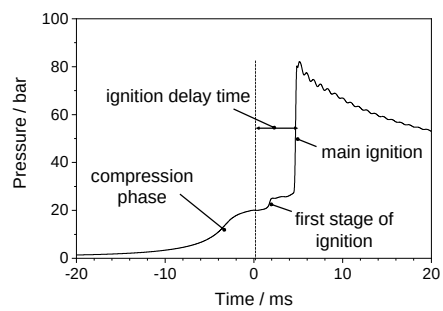


Fig. 8. Exemplary pressure trace of a rapid compression machine experiment with a two-stage ignition. Time 0 corresponds to the end of compression.

Bibliography

De Cao, N. and Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973.

Dreiseitl, S., Osl, M., Scheibböck, C., and Binder, M. (2010). Outlier detection with one-class SVMs: An application to melanoma prognosis. *AMIA Annual Symposium Proceedings*, 2010:172–176.

Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. *35th International Conference on Machine Learning, ICML 2018*, 5:3632–3648.

Kajino, H. (2019). Molecular hypergraph grammar with its application to molecular optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3183–3191. PMLR.

Lee, C., Vranckx, S., Heufer, K. A., Khomik, S. V., Uygun, Y., Olivier, H., and Fernandez, R. X. (2012). On the chemical kinetics of ethanol oxidation: Shock tube, rapid compression machine and detailed modeling study. *Zeitschrift für Physikalische Chemie*, 226(1):1–28.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022.

Ramalingam, A., Zhang, K., Dhongde, A., Virnich, L., Sankhla, H., Curran, H., and Heufer, A. (2017). An RCM experimental and modeling study on CH4 and CH4/C2H6 oxidation at pressures up to 160 bar. *Fuel*, 206:325–333.

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875.

Russo, D. J., van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96.

Schweidtmann, A. M., Rittig, J. G., König, A., Grohe, M., Mitsos, A., and Dahmen, M. (2020). Graph neural networks for prediction of fuel ignition quality. *Energy & Fuels*, 34(9):11395–11407.

Schweidtmann, A. M., Weber, J. M., Wende, C., Netze, L., and Mitsos, A. (2021). Obey validity limits of data-driven models through topological data analysis and one-class classification. *Optimization and Engineering*, pages 1–22.

Solgi, R. M. (2020). geneticalgorithm (python package). https://github.com/rmsolgi/geneticalgorithm. Accessed on 08-09-2021.

The GPyOpt authors (2016). GPyOpt: A bayesian optimization framework in python. http://github.com/SheffieldML/GPyOpt. Accessed on 08-09-2021.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285.