Chinese Journal of Chemical Engineering 84 (2025) 266–273

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Full Length Article

Molecular design of high energy density fuels from coal-to-liquids

Haowei Li [1], Bingzhu Min [1], Yaling Gong [1], Linsheng Li [1], Xingbao Wang [1, *], Yimeng Zhu [2], Wenying Li [1, *]

[1] State Key Laboratory of Clean and Efficient Coal Utilization, Taiyuan University of Technology, Taiyuan 030024, China
[2] Kyland Technology Co., Ltd., Beijing 100144, China

ARTICLE INFO

Article history:
Received 14 January 2025
Received in revised form
27 May 2025
Accepted 2 June 2025
Available online 14 June 2025

Keywords:
Coal-based liquid fuel
Structure-property relationship
Convolutional neural network
Variational autoencoder

ABSTRACT

Direct coal liquefaction products offer a considerable quantity of cycloalkanes, which are the valuable candidates for making the high energy density fuels. The creation of such fuels depends on designing molecular structures and calculating their properties, which can be expedited with computer-aided techniques. In this study, a dataset containing 367 fuel molecules was constructed based on the analysis of direct coal liquefied oil. Three convolutional neural network property prediction models have been created based on molecular structure-physical and chemical property data from the library. All the models have good fitting ability with R^2 values above 0.97. Then, a variational autoencoder generation model has been established using the molecular structures from the library, focusing on the structure of saturated cycloalkanes. The structure-property prediction model was then applied to the newly generated molecules, assessing their density, volumetric calorific value, and melting point. As a result, 70000 novel molecular structures were generated, and 25 molecular structures meeting the criteria for high energy density fuels were identified. The established variational autoencoder model in this study effectively assimilates the structural information from the sample set and autonomously generates novel high energy density fuels, which is difficult to achieve in traditional experimental methods.

1. Introduction

High energy density fuels (HEDFs) possess several key advantages, including high density, high calorific value, high flash point, and low melting point, the properties of representative HEDFs are shown in Table 1. The performance of HEDFs plays a crucial role in determining the range, speed, and payload capacity of aircraft [1]. As the aviation transportation sector evolves, the development of higher-performing and more easily accessible HEDFs is becoming increasingly essential. Direct coal liquefaction oil is produced through a three-stage hydrogenation process with the aid of a catalyst, which preserves the cyclic molecular structure of coal within the oil [2]. The hydrocarbon profile of the jet fuel fraction obtained from direct coal liquefaction oil closely mirrors that of the traditional jet fuel JP-900 [3]. Consequently, coal-based HEDFs research shows great potential. Developing new fuels requires addressing two critical factors: the calculation of diverse properties and the design of novel molecules.

Quickly and accurately determining the physical and chemical properties of fuel compounds is fundamental for guiding molecular design [8]. With the rapid advancement of computer technology, machine learning techniques for predicting compound properties have gained significant attention in research [9]. Casey *et al.* [10] utilized spatial points to represent the 3D structure of molecules and applied a convolutional neural network (CNN) to predict eight key properties of high-energy molecules. All models showed a determination coefficient exceeding 0.95. Yang *et al.* [11] employed graph neural networks (GNN) combined with molecular topological structures to establish an accurate relationship between molecular structure and density. This approach not only provides high computational accuracy but also reduces computation time. In comparison to the QSPR model based on feature engineering [12], machine learning is a data-driven approach that can automatically extract key features from large datasets of known fuel property data, achieving high-precision property predictions for new fuels without the need for an in-depth understanding of the fuel molecule structure [13]. This significantly supports the development and optimization of

This article is part of a special issue entitled: AI for Chemical Engineering published in Chinese Journal of Chemical Engineering.

* Corresponding authors.
E-mail addresses: wangxingbao@tyut.edu.cn (X. Wang), ying@tyut.edu.cn (W. Li).

https://doi.org/10.1016/j.cjche.2025.06.005

Table 1
Properties of representative HEDFs.

HEDFs	Density/g·cm^{-3}	Volumetric calorific value/MJ·L^{-1}	Melting point/°C	Ref.
JP-10	0.94	39.60	<−79.15	[3]
JP-900	0.84	35.95	<−65.00	[4]
RJ-6	1.02	42.60	<−20.85	[5]
RP-3	0.79	34.21	–	[6]
Jet-A	0.81	34.83	<−47.00	[7]

hydrocarbon fuels. Neural networks, particularly, have become widely applied in numerical analysis and data processing [14–16].

Traditional fuel molecule design primarily relied on existing HEDFs molecules, where the skeletal structure was manually altered [17] and functional groups were added or removed [18]. However, the vast search space for fuel compounds made manual design methods highly uncertain. The currently popular high-throughput screening (HTS) method involves large-scale property prediction within a fixed chemical library to identify promising candidates [19]. However, the effectiveness of HTS is dependent on the quality of the chemical library created. Genetic methods generally require the manual setting of rules for mutation and hybridization [20], and traditional discrete optimization techniques face limitations in search efficiency. An alternative approach to overcome these challenges is to directly design the desired molecules, Wen *et al.* [21] employed HTS to evaluate the ZINC20 database, followed by structural retrieval and screening to design HEDFs molecule. Some researchers have proposed extracting potential data from molecular datasets such as PubChem [22] and ZINC [23] to generate target molecules through inverse QSPR/QSAR [24,25]. Also, some researchers have developed data-driven methods to design molecules [26], and this approach can significantly expand the chemical space for molecular screening and markedly improving design efficiency. With advancements in deep learning technology, powerful probabilistic generative models have been developed. These models can uncover design knowledge hidden within molecular datasets and autonomously generate entirely new molecular structures in a systematic and continuous manner. Recently, a variety of deep learning-based generative models for molecular design have been successfully introduced [27,28]. In 2017, the variational autoencoder (VAE) was first applied to molecular generation [29]. This model can transform complex data into low-dimensional, continuous representations, allowing for the automatic generation of new molecules by adjusting vector representations and performing decoding. GNN [30] have also been applied to molecular generation, but require significantly more computational resources for training compared to VAE.

Direct coal liquefaction oil exhibit excellent thermal stability and high flash point characteristics. Further improvement of their density (ρ), volumetric calorific value ($NHOC_V$), and melting point (T_m) would enhance overall fuel performance. This study utilizes deep learning for the design of coal-based HEDFs molecules. A VAE model tailored for molecules design of HEDFs, along with three CNN-based structure-property prediction models, are developed. The VAE model is capable of generating novel fuel molecules, while the CNN models efficiently predict various properties based on molecular structure. The R^2 values for both the training and test sets exceed 0.97, indicating high accuracy. The CNN models were then applied to the new molecules generated by the VAE to calculate their ρ, $NHOC_V$, and T_m. As a result, 25 molecular structures that satisfy HEDFs requirements were identified.

2. Theoretical and Experimental Methods

2.1. Establishment of structure-property dataset of coal liquefaction products

Qualitative analysis of direct coal liquefaction oil was carried out. As shown in Fig. 1(a), the primary components of direct coal liquefied oil include normal alkanes, cycloalkanes, and adamantane. The structure of these compounds was modified by incorporating functional groups such as methyl, ethyl, and cyclopropane, leading to the generation of 367 coal-based hydrocarbon molecules. Fig. 1(b) and (c) presents the distribution of carbon atom count and unsaturation degree for the 367 molecules, illustrating the diversity of molecular structures within the self-constructed dataset. Using Gaussian 16 software package [31] and the M06-2X/6−31+G(d,p) method, these molecular structures were optimized. The ρ, $NHOC_V$, and T_m were subsequently calculated using the DFT method and group contribution method [32]. The T_m was determined using the group contribution method, while ρ and $NHOC_V$ at 298 K were derived from energy properties. The total energy calculated by DFT was used to calculate the combustion enthalpy of the fuel [33]. The optimized wave function files were processed in the Multiwfn [34] molecular surface analysis module. Considering that most hydrocarbon fuels exist in a liquid state at room temperature, the electron density was set to 0.002 to obtain the molecular van der Waals volume. A linear relationship between molecular and experimental volume was used, allowing for the calculation of the experimental volume of unknown hydrocarbons using van der Waals volume regression.

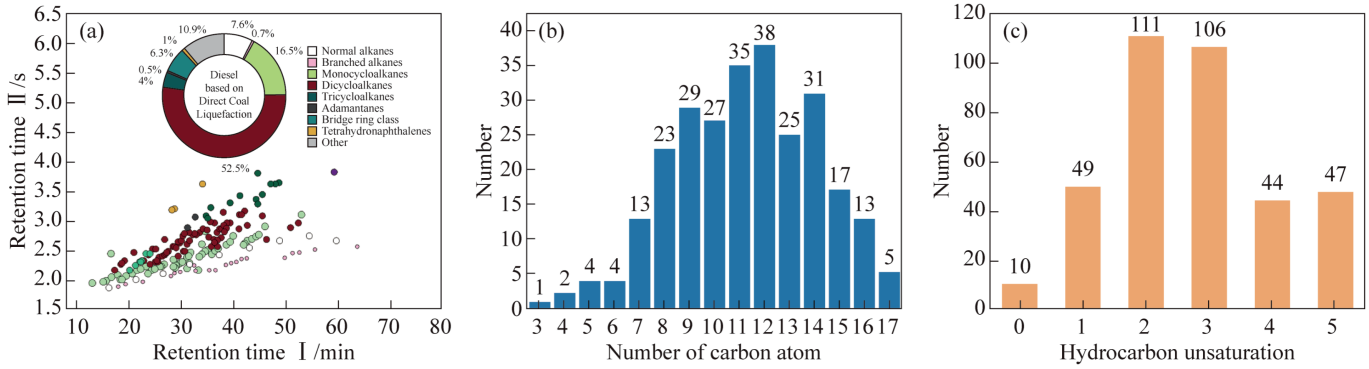

Fig. 1. (a) GC × GC chromatogram diagram and composition diagram of oil based on direct coal liquefaction; carbon number distribution (b) and hydrocarbon unsaturation (c) for the 367 molecules in the dataset.

Fuel density was calculated based on the volume and relative molecular mass, and the calorific value was then derived from the density and combustion enthalpy. The theoretical properties of 11 experimentally synthesized fuel molecules were calculated and compared with experimental data. The results show that the theoretical values align well with experimental trends, confirming the reliability of the calculation method in predicting molecular fuel properties.

2.2. Representation of data

It is crucial to transform the molecular structure into a format suitable for machine learning processing. To efficiently represent the molecular structure, the simplified molecular input line entry system (SMILES) is employed [35]. SMILES uses character strings to depict the graph structure of molecules. Atoms are denoted by their respective atomic symbols, while hydrogen atoms and carbon-carbon single bonds are omitted. Ring bonds are opened, and the disconnected atoms are numbered similarly. RDKit [36] is used to standardize the SMILES representations, ensuring uniqueness in the expression of each molecular structure. Since SMILES is a sequence-based representation, it can be easily converted into data for deep learning applications. The one-hot encoding process is depicted in Fig. 2. First, all SMILES strings are processed to form a unique charset. Then, each SMILES string is traversed, and an array of zeros with the same length as the charset is created. The position of each character within the charset is found, and the corresponding position in the array is set to 1. This results in the one-hot encoded representation of the standardized SMILES.

2.3. Neural network prediction model of coal-based fuel properties

Using the self-constructed coal-based hydrocarbon structure-property dataset as the training dataset, a CNN model is established to predict the ρ, $NHOC_V$, and T_m of the fuel. The training set is divided with a 4:1 ratio relative to the test set. The one-hot matrix derived from SMILES includes the connectivity patterns of chemical compounds. The CNN model extracts structural information through its convolutional layers, then progressively processes these features via pooling layers before ultimately determining compound properties through fully connected layers. The first convolution layer has 1 input channel and 16 output channels, while the second convolution layer has 16 input channels and 32 output channels. Both convolution layers use a kernel size of 3, a stride of 1, and padding of 1. The pooling window size is set to 2, with a step size of 2. The first fully connected layer has an input dimension of 32 × 155 and an output dimension of 128, while the second fully connected layer has an input dimension of 128 and an output dimension of 8. Detailed architectural parameters of the CNN are provided in supplemental files Table S5, Supplement Material. The output layer has a dimension of 1. Other parameters of the network are optimized through a hyperparameter tuning process. The CNN utilizes the Huber loss function, which combines the benefits of mean square error and absolute error. The Huber loss function is defined as follows:

$$\text{Huber loss}(x) = \begin{cases} \frac{1}{2}x^2, \text{if} |x| \le \Delta \\ \Delta\left(|x| - \frac{1}{2}\Delta\right), \text{otherwise} \end{cases}$$

During training, the AdamW optimization algorithm is applied. This algorithm merges the Adam optimizer with weight decay technology. In AdamW, weight decay is applied first, followed by updates of the remaining gradients using the Adam optimizer. This strategy aims to enhance model robustness by incorporating weight decay while maintaining the flexibility and effectiveness of the optimizer.

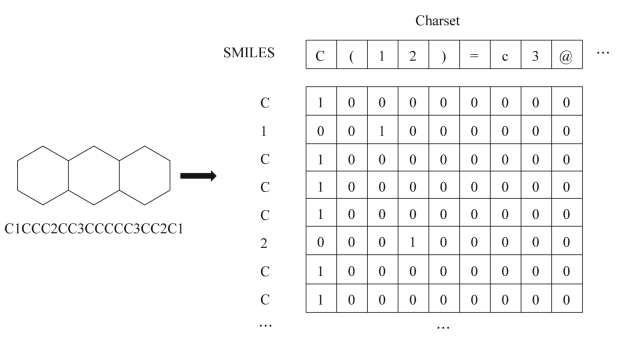

Fig. 2. Schematic of converting SMILES into one-hot coding.

2.4. VAE generation model of coal-based HEDFs

The dataset used in this study consists of a molecular structure dataset of coal liquefaction saturated hydrocarbons, with SMILES serving as the molecular structure descriptor. SMILES is chosen because it easily converts into molecular structures and supports reverse decoding. Given the small size of the self-built dataset, deep learning models may exhibit instability and significant errors. To mitigate this, a subset of the GDB-13 dataset (https://gdb.unibe.ch/downloads/), containing 818477 cyclic hydrocarbon molecules with spatial structure data provided as standard SMILES, is incorporated. PyTorch is used to convert the SMILES strings from both datasets into one-hot encoded vectors, which are then input into the VAE model.

The VAE's encoder and decoder are both composed of fully connected neural networks with three layers, and ReLU activation functions are applied between layers. Detailed architectural parameters of the VAE are provided in supplemental files Table S6. The SMILES representations are mapped by the VAE encoder into a latent space — a low-dimensional continuous space that encodes abstract features of the SMILES strings, specifically their distributional mean and variance. The VAE decoder can decode any point in this latent space to generate novel molecular structures distinct from the original molecules. Through the encoder, we obtained the mean and variance parameters for 350 reference molecules. Leveraging the continuous nature of the latent space, the decoder performs decoding operations in the vicinity of these 350 molecular representations. Here, the mean vectors capture the essential molecular characteristics, while the variance parameters control structural diversity. By systematically modulating these variance parameters during decoding, we generate new molecules. This structure is effective for learning and representing complex data. The ADAM optimization algorithm is employed during training, with the learning rate set to 0.0001. The batch size is set to 5 for training and 2 for testing. The model is developed within an Anaconda 3 environment running on a Windows 10 system, using Python 3.9. The deep learning framework used is PyTorch (version 2.6.0), and the RDKit library (version 2020.03.2) is used for chemoinformatics processing.

The objective of the fuel molecule generation model is to learn the probability distribution over the training set, extract features, and generate new molecules with HEDFs-like structures and chemical properties. The quality, diversity, and chemical space coverage of the generated molecules are assessed using RDKit software.

Molecular similarity is typically assessed using the Tanimoto coefficient, which compares the overlap in molecular fingerprints. SMILES were processed using RDKit to compute Morgan fingerprints (radius = 21024 bits), which were subsequently used to calculate Tanimoto similarity coefficients.

$$\text{Tanimoto}(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

Validity checks whether the generated SMILES string corresponds to a legal molecular structure. Uniqueness indicates that the generated SMILES string has not been repeated from previous model outputs. Novelty refers to SMILES strings that are both unique and absent from the training dataset.

3. Results and Discussion

3.1. Prediction accuracy of CNN

The machine learning approach was used to train a self-constructed dataset containing 367 hydrocarbon molecules along with their physicochemical properties. Fig. 3 demonstrates how well the CNN predicts three target properties: ρ, $NHOC_V$, and T_m. The scatter plots of predicted versus actual values for both the training and test sets show that data points close to the diagonal line ($y = x$) indicate accurate predictions. This observation suggests that the CNN performs well across both sets. The determination coefficients R^2 for the predictions of ρ, $NHOC_V$, and T_m on the training set are 0.993, 0.997, and 0.990, respectively. The linear correlation of these properties exceeds 0.97 for both the training and test sets, and the Mean Squared Error (MSE) is low, indicating a high model fit for each property, with strong results in both training and testing phases and no indication of overfitting. The performance of the CNN's predictive models is consistent across both the training and test sets. This suggests that even when trained on the relatively small dataset of 367 molecules, the model does not face overfitting issues. The CNN's lower reliance on feature quantity, coupled with its fast computational speed and high accuracy, enables it to predict target molecule properties based on molecular structures quickly, making it effective for determining if new compounds meet HEDFs standards.

3.2. Performance of VAE to generate new molecules

3.2.1. Similarity between new molecules generated by VAE and concentrated molecules of coal-based fuel

Using the trained model for molecular generation, 350 cyclic hydrocarbon compounds from the self-built coal-based hydrocarbon molecular structure library were input into the encoder. Without specifying any target property values, the VAE model generated 200 molecules for each compound, totaling 70000 molecules, from which effective molecules were selected. The VAE model generated 13723 molecules that were already present in the GDB-13 database. After removing these duplicates along with molecules identical to the original 350 reference compounds, a total of 15277 unique and non-redundant novel molecules were successfully produced. According to VAE principles, the latent space is continuous, where each molecule corresponds to a distribution with a mean determined by its structural characteristics. When the decoding variance is set to 0, the distribution of molecules in the latent space follows a standard normal distribution, making the VAE model function similarly to an Autoencoder (AE). In this case, the model can only reproduce the input SMILES strings, lacking generalization, which results in the generated molecules being identical to those in the input set and not producing new structures. To overcome this, the decoding variance was varied. Six different variance values (0, 1, 2, 3, 4, and 5) were used for decoding. The similarity between the newly generated molecules and the original molecules in the input set was evaluated using the Tanimoto coefficient. As shown in Fig. 4, when the decoding variance is set to 0, the similarity between the generated and original molecules is very high, often exceeding 0.8. This occurs because the distribution range of the input molecules

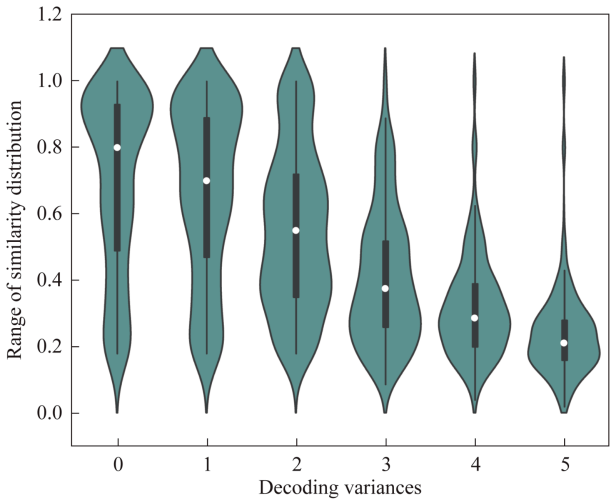

Fig. 4. Similarity distribution between new molecules generated by VAE model and original molecules under different decoding variances.

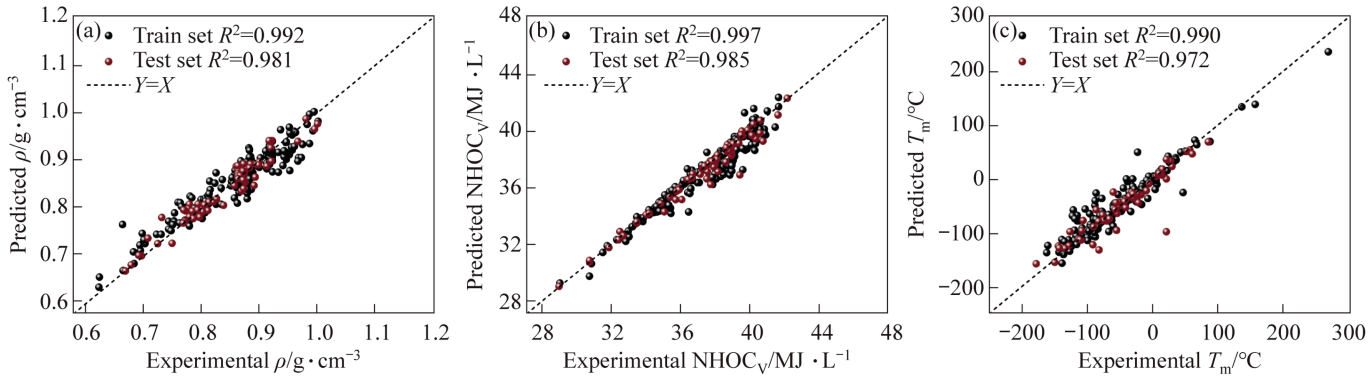

Fig. 3. Predictions of ρ (a), $NHOC_V$ (b), and T_m (c) of coal-based hydrocarbon fuel by CNN via train and test set.

in the latent space is narrow, so most of the molecules are sampled close together. As the variance increases, the similarity gradually decreases. This happens because a higher variance expands the distribution range of the input molecules in the latent space, increasing the sampling area. As a result, most generated molecules are decoded from positions farther away from the original ones, lowering their similarity to the input molecules. However, it is also possible for the similarity to increase with other molecules from the input set. Therefore, similarity is just one aspect used to assess the model's generative capacity.

3.2.2. Variation of effective fuel molecules generated by VAE with decoding variance

Fig. 5 assesses the model's generative performance by using specific metrics for validity, uniqueness, and novelty to determine the most appropriate decoding variance. From Fig. 5, it is evident that when the variance is set to 0, the validity of the generated molecules is at its highest, while the uniqueness is at its lowest. This happens because, at this point, the model mainly reproduces the input molecules with minimal changes, and since the input SMILES are all valid, the validity is high. Uniqueness, however, refers to the distinctness of the molecules after removing duplicates, and since the generated molecules are nearly identical to the originals, the uniqueness is low. The decoding process ratio is set at 1:200, meaning every 200 generated SMILES are almost identical to the original, resulting in reduced uniqueness. As the variance increases, the validity of the generated molecules declines. This occurs because the VAE model transforms the discrete representations of molecules into real vectors through the encoder, which are then decoded back into discrete representations to generate new molecules. However, some of these decoded representations do not correspond to real molecules, reducing the validity of the generated results. Uniqueness initially increases and then decreases with higher variance. When the variance is 0, the model simply replicates the training set, but with larger variance, the model loses the ability to map features correctly, producing many repetitive and unrelated SMILES. Similarly, novelty first increases and then decreases as variance rises, for reasons akin to those observed in uniqueness. The novelty of the generated molecules reaches its peak when the sampling variance is 3, so the variance is set to 3 for subsequent decoding.

In this work (decoding variance = 3: 0.396), the generated molecular validity demonstrates competitive advantages compared to other generative models (GrammarVAE: 0.310 [37], ChemVAE: 0.170 [29], GraphVAE: 0.140 [38]). Although outperformed by the GNN (0.753 [30]), our approach achieves favorable results with significantly reduced computational resources (Nvidia GeForce RTX 3080 10 GB *vs.* GNN's Nvidia Tesla V100 32 GB [30]). This resource efficiency proves particularly advantageous for high-throughput screening of HEDFs.

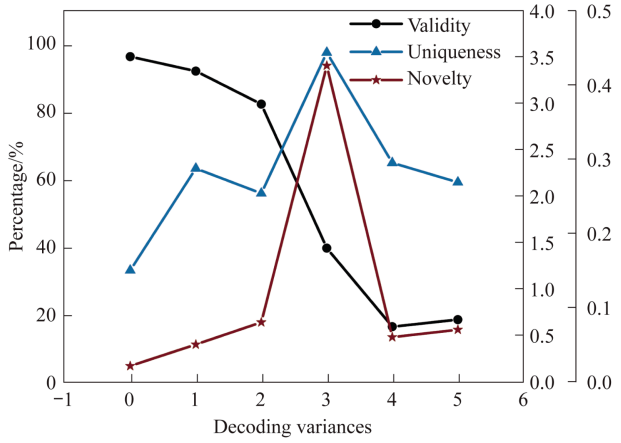

Fig. 5. Validity, uniqueness and novelty of new molecules generated by VAE model under different decoding variances. (The greater the variance, the farther the decoding point is from the original molecule in the hidden space).

3.2.3. Distribution of physical and chemical properties of new molecules generated by VAE

The trained CNN prediction model was used to calculate the physicochemical properties of the novel molecules generated by the VAE model. In Fig. 6, the red and blue bar charts illustrate the distribution of physicochemical properties in the generated set and the training set, respectively. Fig. 6(a) compares the ρ distribution between the two datasets, Fig. 6(b) compares the distribution of $NHOC_V$, and Fig. 6(c) compares the distribution of T_m. The figures show that the distributions of the three physicochemical properties in the generated set align closely with those in the training set across most areas, indicating that the model has effectively captured the distribution of physicochemical properties from the training set and exhibits good reproducibility. From Fig. 6, it can be seen that the T_m of the molecules in the generated set predominantly fall between −50 °C and 0 °C, the ρ are concentrated between 0.85 and 0.95 $g \cdot cm^{-3}$, and the $NHOC_V$ are mostly in the range of 38 to 42 $MJ \cdot L^{-1}$. This suggests that the generated molecules maintain the physicochemical property characteristics typical of polycyclic alkanes found in coal liquefaction products.

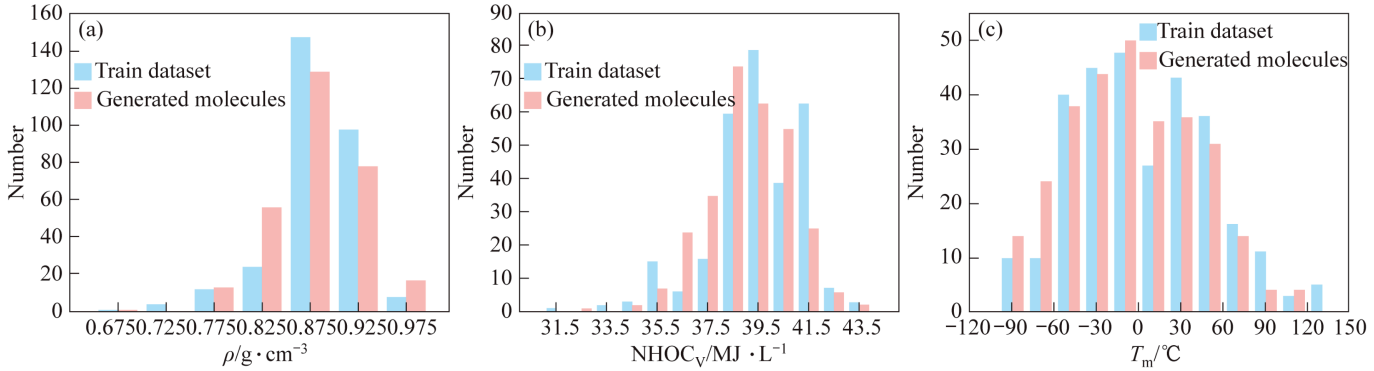

Fig. 6. Distribution of physical and chemical properties of generated molecules and original molecules: (a) ρ, (b) $NHOC_V$, and (c) T_m.

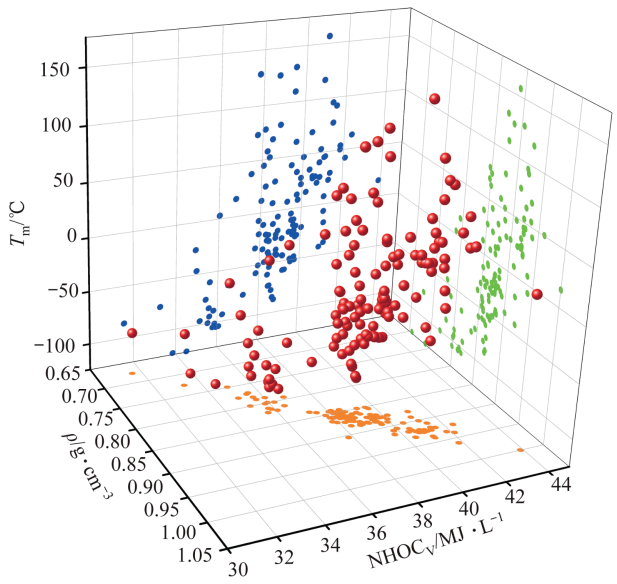

Fig. 7. The values of ρ (x axis), $NHOC_V$ (y axis), and T_m (z axis) for the molecules of generated set.

3.3. Screening HEDFs from the generated fuel molecules

The generated molecules were screened based on three primary performance criteria for high energy density fuels: $\rho > 0.85$ g·cm^{-3}, $NHOC_V > 38$ MJ·L^{-1}, and $T_m < -50$ °C. A three-dimensional plot of these properties for each generated molecule is shown in Fig. 7. The graph represents the performance of all molecules in the generated set across the three properties of ρ, $NHOC_V$, and T_m. By examining the distribution of the data points in the three-dimensional space, potential high-performing molecules can be easily identified, while those with poor properties can be quickly excluded. The molecules were then sorted according to the three criteria, and the ones that met the screening requirements were selected. To validate the accuracy of the CNN model, we performed additional verification through DFT calculations on these molecules. The results demonstrate that the CNN model achieves reliable predictions for the newly generated molecules, as detailed in Supplementary Table S8. The structures of these molecules are presented in Fig. 8. The analysis of the selected structures showed that among the 25 molecules, 12 contain a three-membered ring, 9 contain a four-membered ring, 8 contain a five-membered ring, 18 have a bicyclic structure, 4 have a tricyclic structure, and 3 have a bridged ring structure. These structures are likely to meet the HEDFs requirements due to the high-tension nature of three- and four-membered rings, which release significant energy when their carbon-carbon bonds break. Additionally, the bridged ring structures share similarities with JP-10. It can be concluded that incorporating three-, four-, and five-membered rings, as well as bridged ring structures, could be a promising strategy for designing new HEDFs or improving the performance of existing hydrocarbon fuels.

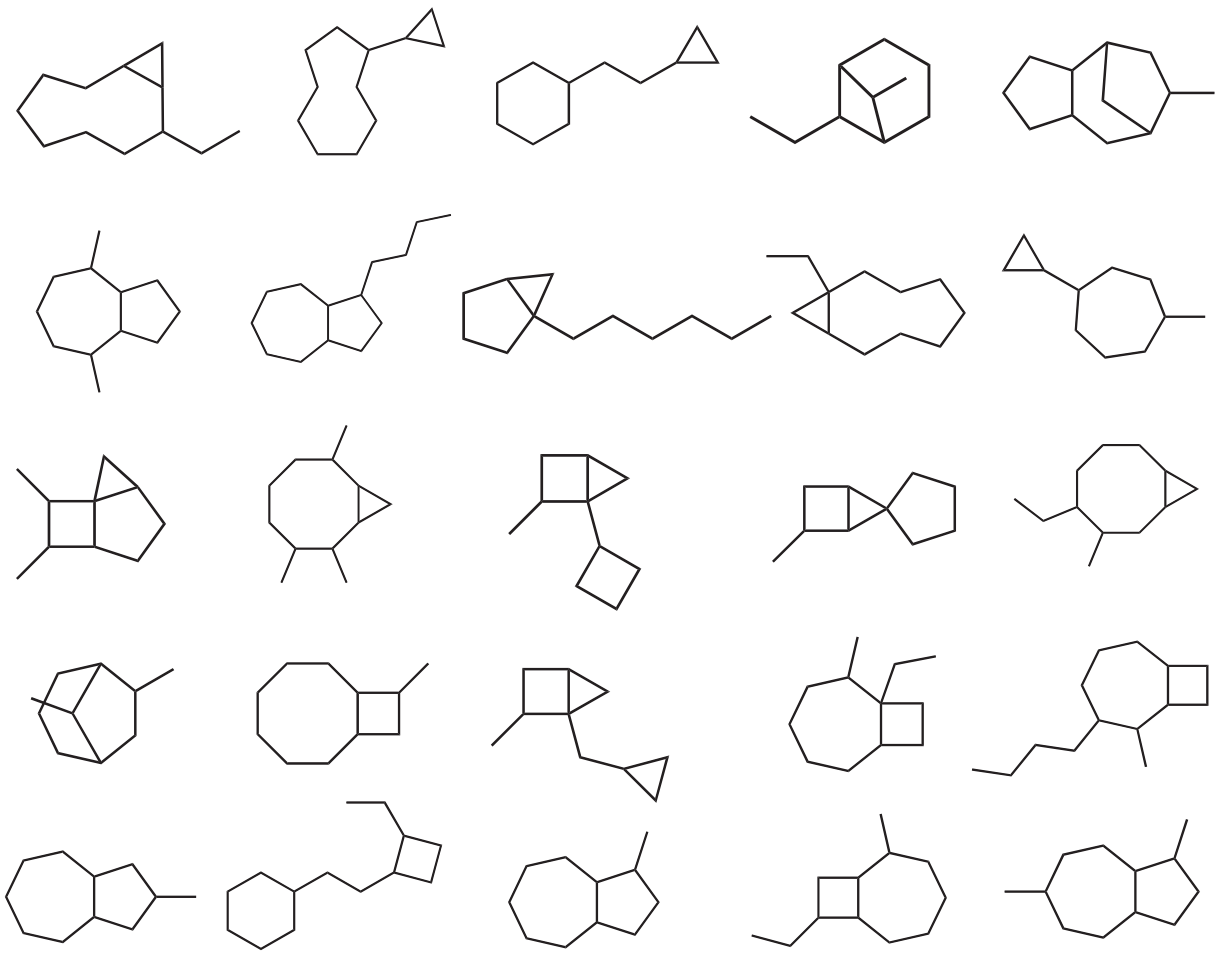

Fig. 8. Molecular structures and their corresponding properties that meet the requirements of high energy density fuels in generated set.

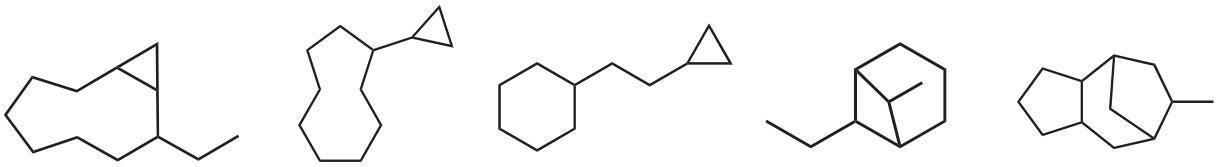

Fig. 9. Structures can be synthesized theoretically in 25 molecules that meet the requirements of HEDFs.

By analyzing the structures of the fuel molecules and their corresponding properties in Fig. 8, it is clear that bicyclic saturated hydrocarbons with more than 10 carbon atoms meet the requirements for $\rho > 0.85$ g cm^{-3} and NHOC$_V$ > 38 MJ·L^{-1}. Whether these molecules pass the screening mainly depends on their T_m. The presence of tricyclic, tetracyclic, and heptacyclic rings, as well as the $-CH_2$ group, tends to lower the T_m. This is because the T_m of heptane is inherently lower compared to other cycloalkanes, and the $-CH_2$ group has a negative contribution to the T_m in the group contribution method, thus reducing the T_m. Furthermore, bicyclic structures involving cyclopropane and cyclobutane combined with most cycloalkanes also result in lower T_m. These results demonstrate that the VAE model has learned the distinctive characteristics of the input structures and has generated combinations that align with the HEDFs requirements. The first five structures in Fig. 9 are synthesizable, with molecules containing tricyclic and bridged ring structures able to be synthesized through cyclopropanation reactions [39] and cycloaddition reactions [40].

At present, research on using machine learning for designing new HEDFs remains limited. Li *et al.* [41] applied high-throughput screening methods to calculate the properties of compounds within existing structural datasets, identifying 28 new hydrocarbon molecules with high density, high specific impulse, high mass calorific value, and low freezing points. Half of these molecules contain four carbon rings, while the other half feature five carbon rings. The most common ring type in these molecules is the pentagonal ring, followed by triangular, square, and hexagonal rings. These molecules are notable for their numerous rings and strong spatial structures. However, their practical existence is unlikely, and their synthesis would be challenging. In contrast, the molecules generated in this study primarily consist of two or three rings, and many of the molecules that are easier to synthesize have already been identified. The findings from this study can provide useful insights for developing new coal-based HEDFs.

4. Conclusions

In this study, a method for automatically generating coal-based high energy density fuel molecules was developed. Firstly, the combustion enthalpy, calorific value, density, and melting point of direct coal liquefaction saturated hydrocarbon products were accurately calculated using DFT methods and the group contribution method. The errors between the calculated and the experimental values of the physical and chemical properties of high energy density fuels are relatively small. Secondly, taking the hydrocarbon compounds in direct coal liquefaction oil as the basic structure, and using the developed methods, the fuel properties of all compounds in the dataset were calculated. An independent molecular structure dataset dominated by cycloalkanes was constructed. Subsequently, using the dataset as input samples, three convolutional neural network prediction models were built. These models enabled more accurate prediction of molecular density, freezing point, and volume calorific value directly based on the molecular structure. A variational autoencoder generative model based on saturated cycloalkanes structures was established, resulting in the generation of 70000 novel coal-based fuel molecules. The generated molecules were screened based on the high-energy density fuels criteria: $\rho > 0.85$ g cm^{-3}, NHOC$_V$ > 38 MJ·L^{-1}, and $T_m < -50$ °C. 25 molecular structures that met the requirements were identified, with five of them being synthesizable. These results confirm that the variational autoencoder model can effectively learn the structural characteristics of high-energy density fuels molecules in the training set and generate new high-energy density fuels molecules autonomously.

CRediT Authorship Contribution Statement

Hao-Wei Li: Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. Bing-Zhu Min: Writing – original draft, Methodology, Investigation, Data curation. Ya-Ling Gong: Visualization, Methodology, Conceptualization. Lin-Sheng Li: Validation, Investigation, Data curation. Xing-Bao Wang: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. Yimeng Zhu: Supervision, Methodology, Conceptualization. Wen-Ying Li: Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

The authors are very grateful for the National Natural Science Foundation of China (22178243 and 22038008).

Supplementary Material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cjche.2025.06.005.

References

[1] H. Yang, Z.J. Yang, Q.F. Yang, X.M. Wei, Y.Q. Yuan, L.L. Wang, Y.F. Hu, J.J. Ding, Simple and high-precision DFT-QSPR prediction of enthalpy of combustion for sesquiterpenoid high-energy–density fuels, *Fuel* 332 (2023) 126157.
[2] J. Stihle, D. Uzio, C. Lorentz, N. Charon, J. Ponthus, C. Geantet, Detailed characterization of coal-derived liquids from direct coal liquefaction on supported catalysts, *Fuel* 95 (2012) 79–87.
[3] H.S. Chung, C.S.H. Chen, R.A. Kremer, J.R. Boulton, G.W. Burdette, Recent developments in high-energy density liquid hydrocarbon fuels, *Energy Fuel.* 13 (3) (1999) 641–649.
[4] L.M. Balster, E. Corporan, M.J. DeWitt, J.T. Edwards, J.S. Ervin, J.L. Graham, S.Y. Lee, S. Pal, D.K. Phelps, L.R. Rudnick, R.J. Santoro, H.H. Schobert, L.M. Shafer, R. C. Striebich, Z.J. West, G.R. Wilson, R. Woodward, S. Zabarnick, Development of an advanced, thermally stable, coal-based jet fuel, *Fuel Process. Technol.* 89 (4) (2008) 364–378.
[5] G.W. Burdette, H.R. Lander, J.R. McCoy, High-energy fuels for cruise missiles, *J. Energy* 2 (5) (1978) 289–292.

[6] L.Y. Huang, S. Huang, Y.B. Mao, B. Wang, Q. Zhu, R.P. Jiang, An experimental study on the laminar burning velocities of RP-3 kerosene and its surrogate fuel at elevated pressures and temperatures, *Fuel* 331 (2023) 125844.
[7] J. Yang, Z. Xin, Q. He, K. Corscadden, H.B. Niu, An overview on performance characteristics of bio-jet fuels, *Fuel* 237 (2019) 916–936.
[8] J.J. Zou, X.W. Zhang, J. Kong, L. Wang, Hydrogenation of dicyclopentadiene over amorphous nickel alloy catalyst SRNA-4, *Fuel* 87 (17–18) (2008) 3655–3659.
[9] J. Zhang, Q. Wang, W.F. Shen, Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library, *Chin. J. Chem. Eng.* 52 (2022) 115–125.
[10] A.D. Casey, S.F. Son, I. Bilionis, B.C. Barnes, Prediction of energetic material properties from electronic structure using 3D convolutional neural networks, *J. Chem. Inf. Model.* 60 (10) (2020) 4457–4473.
[11] C.M. Yang, J. Chen, R.W. Wang, M. Zhang, C.Y. Zhang, J. Liu, Density prediction models for energetic compounds merely using molecular topology, *J. Chem. Inf. Model.* 61 (6) (2021) 2582–2593.
[12] R. Liu, Y.C. Tang, J. Tian, J. Huang, C.Y. Zhang, L.Y. Wang, J. Liu, QSPR models for sublimation enthalpy of energetic compounds, *Chem. Eng. J.* 474 (2023) 145725.
[13] R.Z. Li, J.M. Herreros, A. Tsolakis, W.Z. Yang, Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types, *Fuel* 304 (2021) 121437.
[14] S.W. Xiong, L. Zhou, Y.Y. Dai, X. Ji, Attention-based long short-term memory fully convolutional network for chemical process fault diagnosis, *Chin. J. Chem. Eng.* 56 (2023) 1–14.
[15] F. Hou, Z.Y. Wu, Z. Hu, Z.R. Xiao, L. Wang, X.W. Zhang, G.Z. Li, Comparison study on the prediction of multiple molecular properties by various neural networks, *J. Phys. Chem. A* 122 (46) (2018) 9128–9134.
[16] X.B. Zhang, A. Rajendran, X.B. Wang, W.Y. Li, Solubility study of hydrogen in direct coal liquefaction solvent based on quantitative structure–property relationships model, *Chin. J. Chem. Eng.* 64 (2023) 250–258.
[17] M.S. Manna, C.K. Das, S. Ghanta, Design of C-H-N-O based new hetero-cyclic high energy density molecules: a theoretical survey, *Struct. Chem.* 32 (3) (2021) 1095–1104.
[18] J.P. Liu, L.L. Liu, X.B. Liu, Development of high-energy-density materials, *Sci. China Technol. Sci.* 63 (2) (2020) 195–213.
[19] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid, *J. Phys. Chem. Lett.* 2 (17) (2011) 2241–2251.
[20] Y. Kwon, S. Kang, Y.S. Choi, I. Kim, Evolutionary design of molecules based on deep learning and a genetic algorithm, *Sci. Rep.* 11 (1) (2021) 17304.
[21] L.Y. Wen, S.Q. Shan, W.P. Lai, J.W. Shi, M.T. Li, Y.Z. Liu, M.C. Liu, Z.H. Zhou, Accelerating the design of high-energy-density hydrocarbon fuels by learning from the data, *Molecules* 28 (21) (2023) 7361.
[22] S. Kim, J. Chen, T.J. Cheng, A. Gindulyte, J. He, S.Q. He, Q.L. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.* 47 (D1) (2019) D1102–D1109.
[23] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, ZINC: a free tool to discover chemistry for biology, *J. Chem. Inf. Model.* 52 (7) (2012) 1757–1768.
[24] T. Miyao, H. Kaneko, K. Funatsu, Inverse QSPR/QSAR analysis for chemical structure generation (from y to x), *J. Chem. Inf. Model.* 56 (2) (2016) 286–299.
[25] T. Miyao, M. Arakawa, K. Funatsu, Exhaustive structure generation for inverse-QSPR/QSAR, *Mol. Inform.* 29 (1–2) (2010) 111–125.
[26] L.Y. Wen, Y.L. Wang, Y.Z. Liu, Data-driven combinatorial design of highly energetic materials, *Acc. Mater. Res.* 6 (1) (2025) 64–76.
[27] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
[28] M. Mukaidaisi, A. Vu, K. Grantham, A. Tchagang, Y.F. Li, Multi-objective drug design based on graph-fragment molecular representation and deep evolutionary learning, *Front. Pharmacol.* 13 (2022) 920747.
[29] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, *ACS Cent. Sci.* 4 (2) (2018) 268–276.
[30] P. Bongini, M. Bianchini, F. Scarselli, Molecular generative graph neural networks for drug discovery, *Neurocomputing* 450 (2021) 242–252.
[31] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery, J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Gaussian 16, Revision C.01, Gaussian, Inc., Walling ford CT, 2019.
[32] J. Marrero, R. Gani, Group-contribution based estimation of pure component properties, *Fluid Phase Equilib.* 183 (2001) 183–208.
[33] S. Wang, Z.H. Cui, C.Y. Yu, T.W. Tan, Computational assessment of the molecular structure and properties for high energy density fuel, *J. Phys. Chem. A* 124 (33) (2020) 6660–6666.
[34] T. Lu, F.W. Chen, Multiwfn: a multifunctional wavefunction analyzer, *J. Comput. Chem.* 33 (5) (2012) 580–592.
[35] Y.X. Wei, L.Y. Shan, T. Qiu, D.N. Lu, Z. Liu, Machine learning-assisted retrosynthesis planning: current status and future prospects, *Chin. J. Chem. Eng.* 77 (2025) 273–292.
[36] G. Landrum, RDKit: open-Source cheminformatics, Software (2016). https://www.rdkit.org.
[37] J.K. Matt, P. Brooks, J.M. Hernández-Lobato, Grammar variational autoencoder, Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017.
[38] M. Simonovsky, N. Komodakis, GraphVAE: towards generation of small graphs using variational autoencoders. Artificial Neural Networks and Machine Learning – ICANN 2018, Springer International Publishing, 2018, pp. 12–422.
[39] J.D. Woodroffe, D.V. Lupton, M.D. Garrison, E.M. Nagel, M.J. Siirila, B.G. Harvey, Synthesis and fuel properties of high-energy density cyclopropanated monoterpenes, *Fuel Process. Technol.* 222 (2021) 106952.
[40] S.M. Cho, J.C. Kim, J. Kim, Y.M. Cho, H.W. Kwak, B. Koo, I.G. Choi, Alkyl bicyclo [2.2.2] octanes as high-energy-density bio-aviation fuel, *Fuel Process. Technol.* 254 (2024) 108047.
[41] G.Z. Li, Z. Hu, F. Hou, X.Y. Li, L. Wang, X.W. Zhang, Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels, *Fuel* 265 (2020) 116968.