Chemical Engineering Science 274 (2023) 118686

Contents lists available at ScienceDirect

Chemical Engineering Science

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

Deep generative fuel design in low data regimes via multi-objective imitation

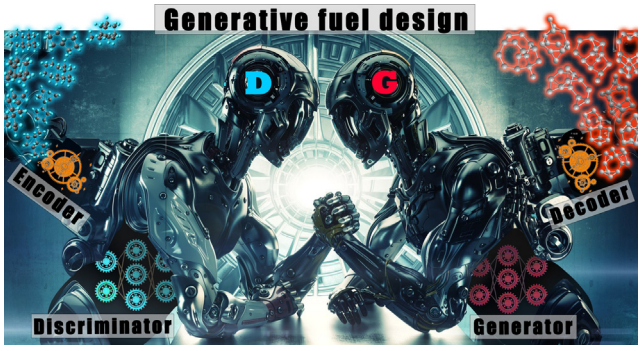
Yifan Liu [a], Runze Liu [a], Jinyu Duan [a], Li Wang [a,b,c], Xiangwen Zhang [a,b,c], Guozhu Li [a,b,c,*]

[a] Key Laboratory for Green Chemical Technology of Ministry of Education, School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China
[b] Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, China
[c] Haihe Laboratory of Sustainable Chemical Transformations, Tianjin 300192, China

HIGHLIGHTS

- A deep generative model is established to design desired fuel molecules.
- Fuel-relevant chemical space is enriched automatically in low data regimes.
- Multi-objective imitation on target fuel is realized by in-depth optimization.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 July 2022
Received in revised form 10 March 2023
Accepted 21 March 2023
Available online 24 March 2023

Keywords:
Deep learning
Generative model
Fuel
Variational autoencoder
Generative adversarial network
Ensemble learning

ABSTRACT

Commercial fuel discovery faces a constantly decreasing return of investment due to due to increasingly tight environmental criteria and reducing potential uses for each new fuel. In this paper, a deep generative model, termed *Latent Interspace Generative Adversarial Network with a Domain of Stacking* (LIGANDS), has been established to screen desired fuel molecules in the large chemical space without setting design rules manually. A variational autoencoder, a generative adversarial network and a stacking model are well integrated in LIGANDS through model convergence. Given only the structures of 255 typical high-energy–density fuels in low data regimes, LIGANDS generated 3461 new fuel molecules with similar property distribution and improved energy performance as the qualified candidates of next-generation fuels. To expand and enrich the fuel-relevant chemical space with innovative molecular entities on demand, in-depth multi-objective imitation on the key properties of target fuel is realized by LIGANDS through optimizing generative molecular structures and their distribution.

1. Introduction

The goal of fuel design is to identify novel structures and optimal composition that can endow desired properties for the discovery of new fuels (Lu et al., 2011; Yalamanchi et al., 2022; Yue et al., 2016; Zhang and Jia, 2020; Zhang et al., 2018). Classification of hydrocarbon structures and regression of molecular properties are simple and efficient strategies for fuel design. Many novel mod-

* Corresponding author at: Key Laboratory for Green Chemical Technology of Ministry of Education, School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China.
E-mail address: gzli@tju.edu.cn (G. Li).

https://doi.org/10.1016/j.ces.2023.118686
0009-2509/

els were established, e.g., group contribution method (Marrero and Gani, 2001; Osmont et al., 2006), DFT calculation (Ramakrishnan et al., 2014; Wheeler et al., 2009) and various machine learning algorithms (Guo et al., 2017; Han et al., 2021; Hou et al., 2018; Lehn et al., 2020; Li et al., 2020; Liu et al., 2022a; Liu et al., 2022c; Schweidtmann et al., 2020), for predicting fuel properties based on their structures and subsequently directing fuel discovery. In addition to the diversity of fuel molecular structure, the composition of a fuel is also very complex and with great importance. Thousands of hydrocarbon substances are usually contained in a typical liquid fuel, which further complicates fuel design. Many quantitative composition-property relationships were established based on a fixed or restricted composition (Heyne et al., 2022; Shi et al., 2017). Research and development of the fuels in commission were greatly accelerated by these robust methods.

Currently, commercial fuel discovery faces constantly increased difficulty and a decreasing return of investment. Because the required criteria for a new fuel are increasing but the number of target engines is shrinking, more rational and efficient design of next-generation fuels is urgently demanded. Both the structures and composition that maximize the quantitative desiderata should be searched, which can be viewed as a multi-objective optimization problem. However, the chemical space of hydrocarbon fuels is discrete, large, and unstructured (Kirkpatrick and Ellis, 2004). For instance, the number of hydrocarbon compounds possessing carbon atoms of $\leq$17 is around 10^9 (Reymond, 2015), which makes the optimization extremely challenging. We recently proposed a deep learning architecture of variational autoencoder (VAE) for hydrocarbon molecules (Liu et al., 2022b), based on which the discrete molecular structure can be manipulated and optimized on demand. To obtain the globally optimal solution for fuel design, a deep generative model based on the VAE should be developed to navigate the whole chemical space of hydrocarbon, including the set of all possible compounds and their distribution. It is anticipated that both generative molecular structures and their distribution are effectively and automatically optimized by deep learning for the rational design of a complex fuel system possessing multiple desired properties.

Herein, we report an integrated deep generative model based on VAE, generative adversarial network (GAN) and ensemble learning for *de novo* fuel design. The applicability of the deep learning framework is shown to design new hydrocarbon molecules that imitate target fuel in depth, which will be an important inspiration source for the discovery of next-generation fuels. Multi-objective imitation on the structures and their distribution of given fuels was achieved during model training for generating new qualified fuel molecules. The chemical libraries of targeted new hydrocarbon compounds optimized for multiple desired fuel properties can be automatically generated by our deep learning framework.

2. Methods

A deep generative model, termed Latent Interspace Generative Adversarial Network with a Domain of Stacking (LIGANDS), has been established and trained for *de novo* design of hydrocarbon fuels with desired properties. As shown in Scheme 1, LIGANDS integrated a variational autoencoder (VAE), a generative adversarial network (GAN) and a stacking model for deep generative fuel design. The VAE was used to convert discrete hydrocarbon molecules to and from a real-valued multidimensional continuous vectors, termed Continuous Operable Molecular Entry Specification (COMES). In the latent space built by the VAE, a GAN was established and trained to achieve a Nash equilibrium between the generator and the discriminator for generative deep learning of target fuels. Input with COMES in the latent space, the stacking model can predict corresponding fuel properties accurately based on ensemble learning. *TensorFlow* (Abadi et al., 2016) was used to construct the framework of LIGANDS, and the pseudocodes of the model is presented in Algorithm 1. The notations in this work are described in Table S1.

The VAE model was first pre-trained on a database containing 319,893 hydrocarbon structures (GDB-13C) for mapping hydrocarbon structures to latent vectors. The stacking model was trained on a self-built database containing 739 fuel molecules and their properties. To train the full LIGANDS model, several typical high-energy–density fuels in the training set was converted to latent vectors by the encoder of the VAE. These vectors were used as the true data for discriminator learning of GAN. A set of random vectors sampled from Gaussian distribution were input into the generator of GAN for generating fake data in the latent space. Fuel properties of the newly generated fuel molecules were calculated by the stacking model using COMES as input, which were monitored in each epoch during model training. Once the LIGANDS training was finished, the generator of GAN was sampled several times, and the obtained latent vectors were decoded to corresponding Simplified Molecular Input Line Entry System (SMILES) strings by the decoder. The properties of the as-generated fuel molecules were also predicted by the stacking model.

Algorithm 1 Training procedure of the LIGANDS model (see Table S1 in Supplementary Material for nomenclature list).

Algorithm 1 Traning procedure of the LIGANDS model.

Input: Database of SMILES $\mathcal{D} = \left\{x^{(i)}, y_i\right\}_{i=1}^{N} (y_i \in \mathcal{Y})$, all of which are high energy density fuel molecules.

Output: Novel structures with properties, which are similar to the molecules in the database.

Step 1: Train a probabilistic encoder and decoder.

$\phi, \theta \leftarrow$ Initialize parameters

repeat

for $i = 1, 2, \ldots, N$ **do**

Draw $\mathcal{L}$ samples from $\epsilon \sim \mathcal{N}(0, 1)$

$$z^{(i,l)} = h_\phi\left(\epsilon(i), x^{(i)}\right) \qquad i = 1, \ldots, N$$

end for

$$E = \sum_{i=1}^{N} -D_{KL}\left(q_\phi\left(z|x^{(i)}\right) \| p_\theta(z)\right) + \frac{1}{L}\sum_{l=1}^{L}\left(\log p_\theta\left(x^{(i)}|z^{(i,l)}\right)\right)$$

$\phi, \theta \leftarrow$ Update parameters using gradients of E (e.g. Stochastic Gradient Descent)

until convergence of parameters ϕ, θ

Step 2: Use encoder to convert SMILES $x^{(i)}$ from database $\mathcal{D}$ to COMES $\boldsymbol{x}_i (\boldsymbol{x}_i \in \mathbb{R}^n)$.

Step 3: Learn base-level regressors

for $t = 1, 2, \ldots, T$ **do**

Learn a based regressor h_t based on $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} (\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \mathcal{Y})$

end for

Step 4: Construct new data set of predictions

for $i = 1, 2, \ldots, N$ **do**

Construct a new data set $D_h = \{h(\boldsymbol{x}_i), y_i\}'$ from D, where $h(\boldsymbol{x}_i) = \{h_1(\boldsymbol{x}_i), ..., h_T(\boldsymbol{x}_i)\}$

end for

Step 5: Learn a meta-regressor

Learn a new regressor h' based on the newly constructed data set D_h

Step 6: Build an adversarial network to generate novel COMES $\boldsymbol{x}_i' (\boldsymbol{x}_i' \in \mathbb{R}^n)$.

for number of training iterations **do**

Sample minibatch of m noise samples $\left\{z^{(1)}, \ldots, z^{(m)}\right\}$ from noise prior $p_g(z)$.

Sample minibatch of m examples $\{\boldsymbol{x}_1 \ldots\ldots \boldsymbol{x}_m\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.

Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d}\frac{1}{m}\sum_{i=1}^{m}\left[\log D(\boldsymbol{x}_i) + \log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right].$$

Sample minibatch of m noise samples $\left\{z^{(1)} \ldots, z^{(m)}\right\}$ from noise prior $p_g(z)$.

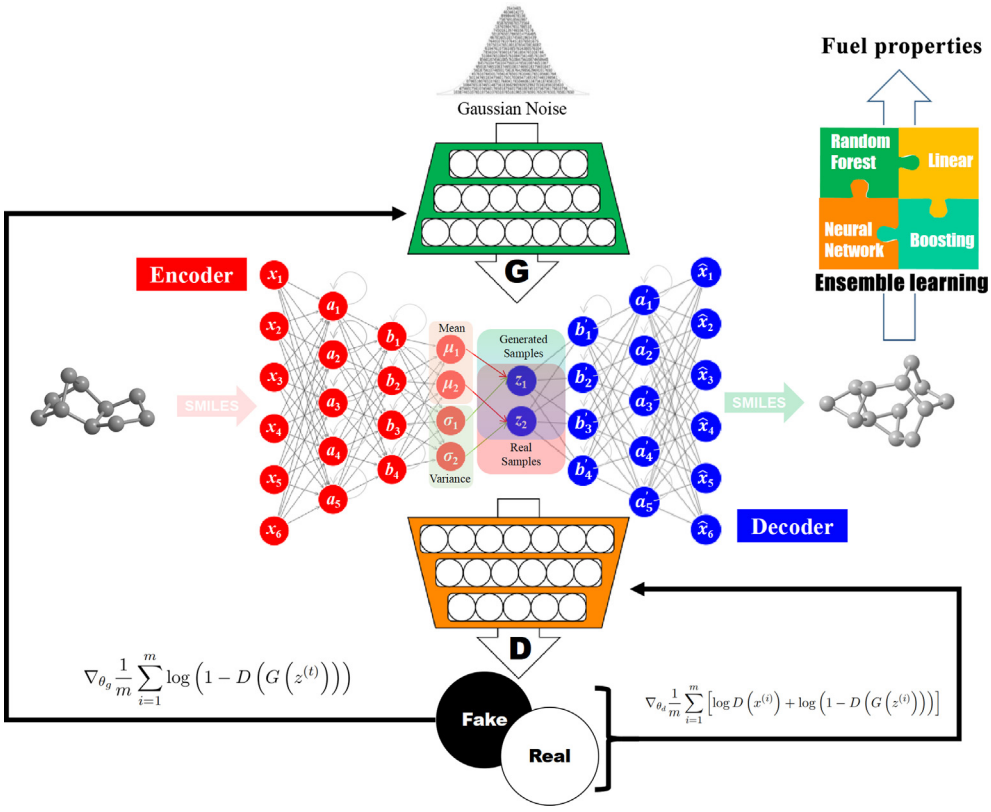
Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g}\frac{1}{m}\sum_{i=1}^{m}\log\left(1 - D\left(G\left(z^{(t)}\right)\right)\right).$$

end for

Step 7: Use decoder to convert COMES from generator to SMILES $x^{(i)'}$ and meta-regressor to predict its properties.

return $\mathcal{D}' = \left\{x^{(i)'}, H_i(\boldsymbol{x}_i')\right\}_{i=1}^{M} \{H_i(\boldsymbol{x}_i') = h'(h_1(\boldsymbol{x}_i'), \ldots, h_T(\boldsymbol{x}_i'))\}$



Scheme 1. The workflow of deep generative algorithm of LIGANDS by integrating a VAE with encoder and decoder, a GAN with generator and discriminator and a stacking model with 8 base learners for *de novo* fuel design.

2.1. Variational autoencoder (VAE)

A VAE was built and trained to establish a rational latent space of hydrocarbon compounds, in which any point can be decoded to a reasonable molecule as shown in Scheme S1. Based on deep learning, the as-trained VAE can convert a hydrocarbon molecule (discrete representation) to and from a continuous multidimensional vector, which is similar to our previous work (Liu et al., 2022b) and that reported by Aspuru-Guzik group (Gómez-Bombarelli et al., 2018). The continuous multidimensional vectors, COMES, were used as the descriptor of hydrocarbon molecules.

GDB-13 dataset (Blum and Reymond, 2009) contains the structures of 977,468,314 organic small molecules. Each molecule contains several of C, H, N, O, S, Cl atoms with a total atom number of $\leq$ 13. We screened saturated hydrocarbon molecules from GDB-13 by a piece of python code. This process ran on a personal computer (CPU: Intel core i5-8250U, memory: 12 GB) in 30.05 s. Finally, 319,893 molecular structures containing only C and H elements without any unsaturated bonds (double and triple bonds) were obtained, and the database was denoted as GDB-13C. The VAE was trained on the GDB-13C database (Blum and Reymond, 2009). As shown in Scheme S1, a hydrocarbon molecule is reversibly converted to a unique vector of 192 dimensions.

Convolutional networks previously showed improved performance for string encoding (Nal Kalchbrenner and Blunsom, 2014). Herein, due to the presence of many repetitive chemical substructures (e.g., functional groups and different cycles), the string mostly contains several translationally invariant substrings. Therefore, recurrent neural networks (RNNs) were employed to encode a string of characters into a vector. A pair of an encoder RNN and a decoder RNN can perform sequence-to-sequence learning (Sutskever et al., 2014). In the encoding of SMILES-based text, the subset containing 35 characters was encoded into a vector with a length of up to 120 characters. Spaces were padded in the short strings to reach the same length. In the training, the SMILES strings were canonicalized to avoid undesirably equivalent SMILES representation. In the deep network of the VAE, the encoder contained three 1D convolutional layers and one fully connected layer (the width was 196). The decoder has three layers of gated recurrent unit (GRU) networks (Chung et al., 2014), in which the hidden dimension was 488.

2.2. Fuel property prediction by the stacking model

The small dataset used for training the stacking model was established in our previous work (Liu et al., 2022c), which contains 739 hydrocarbon molecules and their properties. In the database, 342 samples and their fuel properties were previously collected by our group (Li et al., 2020), and the other molecules and corresponding properties were obtained from American chemical abstracts (CA). Six key fuel properties were investigated, including specific impulse (I_{sp}), density at 25 ℃ (ρ), boiling point (T_b), flash points at atmospheric pressure in air (FP), the net heat of combustion ($NHOC$) and melting point (T_m), which should be given priority in fuel applications. In our database, part of the fuel properties of some hydrocarbon compounds are missing. Energy properties (e.g., specific impulse, heat value) were determined via structure optimization at the DFT level of B3LYP/6-31G(d, p) by using Gaussian 09. Some physicochemical properties (e.g., density, boiling point, flash point) were calculated by the group-contribution method due to its high accuracy, good reliability and low time-consumption (Marrero and Gani, 2001; Osmont et al., 2008; Osmont et al., 2006). For training different fuel properties, corre-

sponding numbers of the samples are summarized in Table S2. The database mentioned above was used as the training set.

The vectors of COMES generated by VAE were used as the input, which were directly fed into the learning model for predicting various fuel properties. As an ensemble learning technique, stacking was employed to combine multiple regression models in a *meta*-regressor (Tang et al., 2015; Wolpert, 1992). The dataset was randomly split into two parts, the testing set (20%) and the training set (80%). In the procedure of stacking, cross-validation was conducted for preparing the input data for the second-level regressor to suppress overfitting as shown in Scheme S2. In 5 successive rounds, the dataset has been split into 5 folds. In each round, the first-level regressors were trained on 4 folds, and then applied to the remaining 1 fold. The first-level predictions were well stacked, and then input to the second-level regressor. After model training, the stacked first-level regressors were fit to the whole dataset. Thus, the error between the predicted value and the real value was determined. Finally, the best stacking model can be obtained via comparison.

2.3. Generative adversarial network (GAN)

In order to realize the function of generative deep learning, a GAN was built based on the as-developed VAE of hydrocarbon molecules. A generator and a discriminator were included in the GAN. In our model, the discriminator constantly improved itself to judge whether the new molecule generated by the generator is a qualified fuel based on deep learning in a database of target fuel structures. And the generator was trained to generate qualified new molecules to confuse false with true in the discriminator. Herein, the structures of 255 typical known high-energy–density fuels were used as the training set.

In GAN, the model of discriminator (D) and the model of generator (G) were built and jointly trained. Both models were multilayer perceptrons (MLP) to make sure that the adversarial modeling framework can be most straightforward to apply. The MLP had 2 layers with 1024 neurons and 384 neurons in the first layer and the second layer, respectively. The G model has a parameter of θ. An optimal θ makes the probability distribution of the sample generated by model G as close as possible to the probability distribution of the real data ($\mathbb{P}$). The target is to minimize the measure of the difference between the two as shown in equation (1).

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ D(\mathbb{P}, \mathbb{Q}_\theta) \tag{1}$$

The D model possesses a parameter of ϕ. A value function (V) was defined to determine the optimal ϕ by maximizing V. The value function was defined in equation (2).

$$V(\mathbb{P}, \mathbb{Q}_\theta, D_\phi) = \mathbb{E}_{\mathbb{P}}\left[\log D_\phi(x)\right] + \mathbb{E}_{h(z)}\left[\log\left(1 - D_\phi(G(z))\right)\right] \tag{2}$$

The value function was calculated based on cross entropy, in which the first item corresponds to the real samples and the second item corresponds to the generated samples. When more real samples were classified as true (with the value of 1) and more generated samples were classified as false (with the value of 0) by D, the value function of D has the higher value. G and D were jointly trained in GAN, and their abilities were constantly improved in the confrontation. It can be described as an optimization problem as depicted in equation (3).

$$\left(\hat{\theta}, \hat{\phi}\right) = \underset{\theta}{\operatorname{argmin}}\ \underset{\phi}{\operatorname{argmax}}\ V(\mathbb{P}, \mathbb{Q}_\theta, D_\phi) \tag{3}$$

When Nash equilibrium between G and D is achieved, the convergence of the GAN algorithm for the design of target fuels is realized. In such condition, the generator can convert a randomly input vector into a new hydrocarbon molecule possessing requisite fuel properties. The discriminator can effectively distinguish whether a molecule is qualified as the target fuel.

3. Results and discussion

A deep generative algorithm, termed Latent Interspace Generative Adversarial Network with a Domain of Stacking (LIGANDS), has been established as shown in Scheme 1. The training procedure of LIGANDS was described in the section of Methods. Briefly, a latent space of target hydrocarbon molecules was built by the variational autoencoder (VAE), in which a real-valued multidimensional vector (COMES) can be reversibly converted to a reasonable hydrocarbon molecule. In the latent space, a GAN and a stacking model were built and work collaboratively for generating new target fuel molecules and predicting their properties, respectively. After the LIGANDS algorithm was trained and converged on the three databases (hydrocarbon structure database, target fuel structure database and fuel property database), new fuel molecules with multiple desired properties have been generated automatically by LIGANDS.

3.1. Representation of hydrocarbon molecules in latent space by VAE

Firstly, a VAE has been built and trained on a database of hydrocarbon structures (GDB-13C) to set up a rational latent space of hydrocarbons as illustrated in Scheme S1. RNNs exhibited superior capability of learning the context information of SMILES, which is beneficial to reversible coding and decoding. Therefore, RNNs were employed in the encoder and decoder. Given a discrete hydrocarbon molecule, the encoder of VAE converts the SMILES string to a continuous vector in the multidimensional latent space, which is a versatile molecular representation. For a point in the latent space, the network of the decoder can produce a SMILES string, and a hydrocarbon molecule with a reasonable structure is generated.

We examined the ability of the VAE for encoding and decoding hydrocarbon molecules. The statistical data show that 99% of the molecular structures can be well maintained after coding and decoding, indicating the good robustness of the model. The performance of the as-built latent space to capture structural features of the hydrocarbon molecules has also been evaluated. Fig. 1a shows a kernel density estimate (KDE) of each dimension when encoding 319,893 hydrocarbon molecules from GDB-13C. Along each dimension of the latent space, the distribution of encoded molecules can be illustrated by KDE. As shown in Fig. 1a, slightly different means and standard deviations are depicted by the distribution of data points in each individual dimension. Overall, normal distributions have been achieved as enforced by the variational regularizer.

Fig. 1b shows the sampling results around decalin. Fig. 1c displays some molecules that are close to 1-methyladamantane in the latent space. The numbers labelled near the decoded molecules are the distance between the molecule and the initial one (decalin or 1-methyladamantane) in the latent space. New hydrocarbon molecules with reasonable structures can be sampled in the as-built latent space. By increasing the distance in the latent space, the decoded new structures become less similar to the original ones. Fig. 1d shows the spherical interpolation between two typical fuel molecules, decalin and JP-10. The sampling points of spherical interpolation are distributed on a high-dimensional sphere centered on the targeted molecule in the latent space, which ensures high-efficiency sampling with a diversity of collection results. Smooth transitions of molecular structure were realized in the latent space. In the continuous latent space, interpolation of known molecules is allowed by following the shortest Euclidean path in the interspace, and both the stacking model and the GAN of LIGANDS will work hard for generative fuel design.

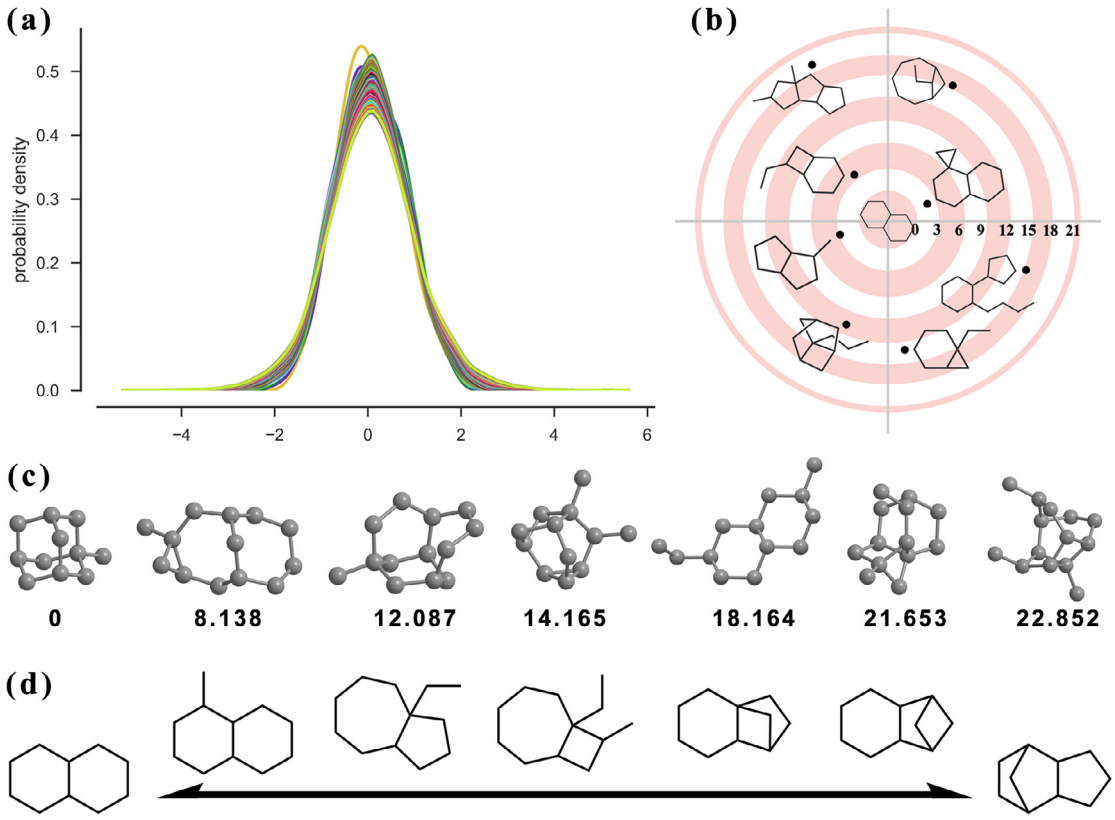


Fig. 1. Sampling results from the VAE. (a) KDE of each latent dimension of the encoder. (b) Distribution of the as-sampled molecules near decalin in the latent space. The distance of a molecule from the original query was expressed by the radius to the center of the circle. (c) Some molecules that were sampled near the location of 1-methyladamantane in the latent space. The values below the molecules are the distance in the latent space from the decoded molecule to 1-methyladamantane. H atoms were omitted for clear observation. (d) Slerp interpolation between decalin and JP-10 in the latent space by six steps of equal distance.

3.2. Predicting fuel properties by ensemble learning

In order to collect sufficient data for efficient machine learning, multiple sources of fuel properties were combined, including experimental data, DFT calculation results, the data calculated by group contribution method, and the data predicted by software. Due to the presence of some systematic differences among various data from different sources, high robustness of the model is required, which can be effectively achieved by ensemble learning (Liu et al., 2022c; Zhou, 2012). Ensemble learning was implemented in LIGANDS for predicting fuel properties as illustrated in Scheme 1. In this stacking model, multiple base learners work cooperatively, including support bagging, vector machine (SVM), extremely randomized trees, random forest, light gradient boosting machine, voting, histogram-based gradient boosting, and linear, as shown in Table S3. The final regressor was set to be the simple and efficient learner of linear. The other learners were optimally stacked in the first-level regressor. The errors (MAE and R^2) for predicting different fuel properties by the stacking model are summarized in Table 1. The stacking model exhibited superior accuracy for calculating all six properties of fuel, especially the energy property (R^2 > 94%), compared with the common single learners (Hou et al., 2018; Li et al., 2020).

Table 1
The errors for predicting different fuel properties by the stacking model.

Property	MAE	R^2
T_m	5.8	0.98
T_b	11.6	0.88
FP	2.5	0.99
ρ	0.04	0.96
NHOC	0.16	0.94
I_{sp}	0.70	0.94

3.3. Generating target fuel molecules by GAN

A usable GAN should be a convergent algorithm (Goodfellow et al., 2014). When the discriminator reaches to perfection by optimizing *D* at each step, gradient descent of $KL[q_\theta \| p]$ should be technically achieved with respect to θ. However, in practice, most GANs are tending to be highly unstable. Based on Eq. (1), sufficient gradient can not be provided for *G* to learn well. In the early learning, *G* performs poor, *D* can reject the newly generated samples with high confidence. Because these samples are clearly different from the original target data. The Nash equilibrium is hardly realized, the GAN generally underfits. Our GAN in the latent space of hydrocarbon molecules was also hardly trained to reach convergence. Initially, the generator gave degenerate distributions, and the support don't generally overlap with the distribution of the true high-energy–density fuels. The overfitting of the discriminator prevents the evolution of the generator, leading to unstable behaviors of the GAN.

We tried many methods to help the GAN converge (Karras et al., 2018; Salimans et al., 2016). For our system, two strategies of adding noises was found to be beneficial to the convergence of the GAN, in which noises were added into the label (Salimans et al., 2016) and both real and synthetic data (Sønderby et al., 2016). The noises make the discriminator's job more complex, which prevents the extreme inference behavior of the discriminator. Thus,

the discriminator gives small gradient signal to the generator, and reinforce correct behavior of the generator in the training.

3.3.1. One-sided label smoothing

The technique of label smoothing replaces the target of 0 or 1 for a discriminator with a smoothed value, e.g., 0.1 or 0.9 (Salimans et al., 2016; Szegedy et al., 2015). Vulnerability of the model (neural network) to adversarial examples can be reduced (Warde-Farley and Goodfellow, 2016). The positive and negative classification targets are replaced by (1-α) and β, respectively. p_{data} is the unknown data's probability distribution, and p_{model} is the estimate of the distribution by the generator. The optimal discriminator can be described in equation (4).

$$D(x) = \frac{(1-\alpha)p_{\text{data}}(x) + \beta p_{\text{model}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)} \tag{4}$$

In some areas, p_{model} is large and p_{data} is approximately zero. In such condition, erroneous data from p_{model} have little or no incentive to move nearer to the training data. Thus, it becomes problematic to present p_{model} in the numerator. Therefore, the negative labels were still set to 0, the positive labels were smoothed to (1-α). The label targets of the real samples were replaced by a value of slightly less than 1, such as 0.9, which prevents the extreme inference behavior of the discriminator. One-sided label smoothing prevents the extreme inference behavior of the discriminator and encourages it to estimate soft probability. In such condition, the discriminator does not reduce its classification accuracy, but only reduce its confidence in the true category.

3.3.2. Instance noise

Another origin of the instability of our GAN is the fact that p and q_θ are concentrated distributions, and the support does not overlap. The distribution of high-energy–density fuels (p) is assumed to concentrate around or even on a low-dimensional manifold. Herein, q_θ is manifold-like and degenerate by construction. Before the convergence is reached, p and q_θ were separated by several Ds perfectly, which violates a condition for the convergence proof. This problem has been remedied by adding instance noise to both newly generated molecules and the true samples of high-energy–density fuels. Then, the divergence can be minimized as shown in equation (5).

$$d_\sigma(q_\theta, p) = KL[p_\sigma * q_\theta \| p_\sigma * p] \tag{5}$$

where $p_\sigma * q_\theta$ denotes the convolution of q_θ with a noise distribution of p_σ. The trick of instance noise is related to the strategy of one-sided label noise, but no bias is introduced in the optimal discriminator. During training, the noise level is annealed, which allows us to safely optimize D in each iteration until convergence. Finally, the generator was well trained and the discriminator was unable to distinguish the artificial data from the real data. Then, the convergence of our GAN model was achieved.

The loss changes of the generator and the discriminator during training are shown in Fig. 2. The loss of discriminator dropped quickly in the initial 100 epochs, and then decreased slowly and oscillated slightly. In comparison, the loss of generator also rapidly declined in the first 100 epochs, which was well maintained for another 500 epochs and quickly increased with large fluctuations by further increasing the epochs. As a form of regularisation, early stopping at 600 epoch prevented the discriminator from overfitting. Some fluctuations were present because each batch of molecules was randomly sampled from the latent space by the generator during the training process. On the whole, both the generator and the discriminator finally converged. The structures of the new molecules generated by the generator gradually approximated the batch of the target fuel. Both the discrimination and generation of high-energy–density fuels with high precision have been realized by LIGANDS.

3.4. The performance of LIGANDS for de novo design of high-energy–density fuels

Fig. 3 displays the original samples and the new samples in high-dimensional vector space during training. Principal component analysis (PCA) dimensionality reduction was applied. The blue points indicate the high-energy–density fuels in the training set (real data). The green points display the new molecules produced by the LIGANDS model (generated data). At interaction 100, the green dots gathered in a remote place away from the blue dots. It indicates that the molecules generated by LIGANDS can not well simulate the real samples. With the increase of iterations, the green dots gradually distributed in between the blue dots, illustrating the convergence of the model during training. At interaction

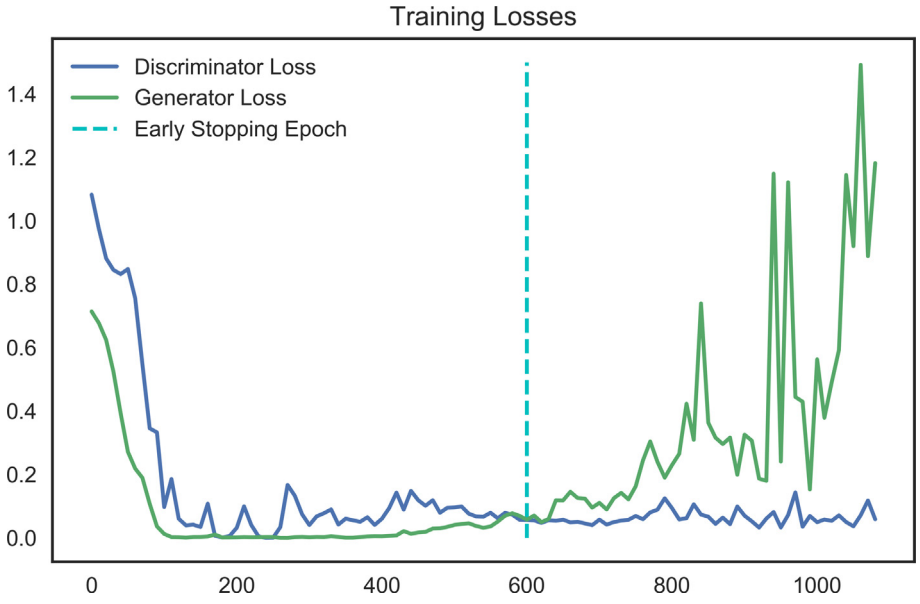


Fig. 2. The changes of loss values of the generator and the discriminator with increasing epoch.

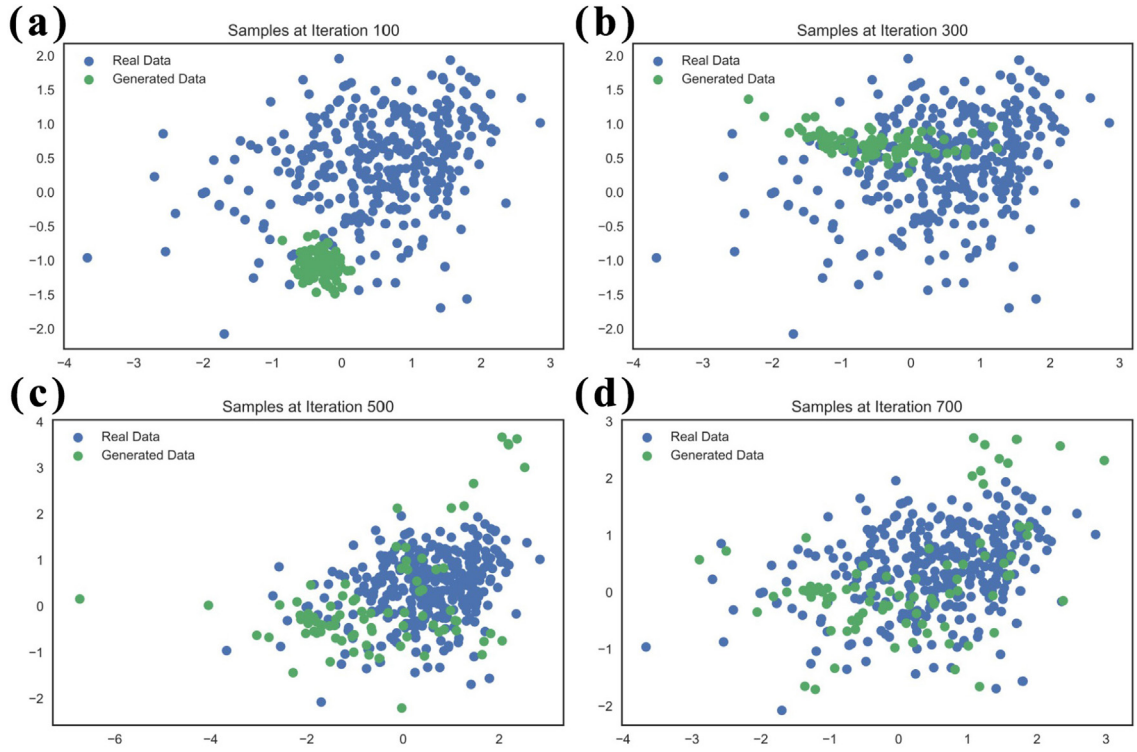


Fig. 3. Convergence of the LIGANDS model. Changes between the original samples and the newly generated data during training are shown. The vector representations of the fuel molecules were reduced to 2 dimensions by PCA for the molecules in the training set and the newly generated molecules.

700, the distribution of green dots and blue dots can not be distinguished, demonstrating the good generating ability of LIGANDS for *de novo* design of target fuel molecules.

Generative deep learning on hydrocarbon fuels is very difficult due to the discrete and complex molecular structures (Zhang et al., 2016). For the SMILES string of a fuel molecule, even one character is altered, the whole will be changed from a legal structure to an illegal structure, or to another molecule. Based on a partially generated sequence, it is very difficult to judge the score of the fully generated sequence. Above problem has been solved by the proper integration of GAN with VAE. Based on robust coding and decoding by the VAE, local information reconstruction is efficiently realized to generate new fuel molecules approaching the target.

$$\nabla_\theta \left[KL\left(\mathbb{P}_{g_\theta} \| \mathbb{P}_r\right) - 2JS\left(\mathbb{P}_{g_\theta} \| \mathbb{P}_r\right) \right]\big|_{\theta=\theta_0} \tag{6}$$

Value function of the generator can be described by equation (6) (Arjovsky and Bottou, 2017). In our LIGANDS model, the generator samples a vector from a low-dimensional (64 dimensions) random distribution, and then generates a high-dimensional (192 dimensions) sample through a neural network. The support of P_r and P_g is a low-dimensional manifold in a high-dimensional space. Therefore, the probability is 1 that the measure of the overlapping part between P_r and P_g is 0. Under such condition, the *JS* divergence is a fixed value (log2) without any gradient (Arjovsky and Bottou, 2017). As depicted in equation (6), the optimization of the generator is to minimize the *KL* divergence between two distributions. The problem is that *KL* divergence is not a symmetrical measure. When $P_g(x) \to 0$ and $P_r(x) \to 1$, $P_g(x) log \frac{P_g(x)}{P_r(x)} \to 0$, the contribution to $KL(P_g \| P_r)$ tends to zero. If the generator fails to generate real samples, the punishment is small. When $P_g(x) \to 1$ and $P_r(x) \to 0$, $P_g(x) log \frac{P_g(x)}{P_r(x)} \to +\infty$, the contribution to $KL(P_g \| P_r)$ tends to positive infinity. If the generator generates an unrealistic sample, the punishment is huge. The punishment for the above two errors of the generator is different. In such situation, the generator would rather generate some repeated but safe samples than give diversified ones. This will give rise to a serious decline in the diversity of generated fuel molecules, that is, collapse mode (Arjovsky et al., 2017). In addition, because of the low diversity, the discriminator is prone to overfitting, which leads to the limited ability to identify only part of the true samples.

Herein, the addition of instance noise (Sønderby et al., 2016) solved the problem that *JS* divergence has no gradient and model cannot be optimized. Instance noise made the low dimensional manifolds mapped by neural networks diffuse to the entire high dimensional space, forcing them to produce an unneglectable overlap. Then, *JS* divergence generated meaningful gradients to draw two low dimensional manifolds closer until they almost coincide. The collapse mode was avoided by one-sided label smoothing (Salimans et al., 2016), which helped the discriminator better counterwork the generator in training LIGANDS. It does not encourage the discriminator to select the incorrect class in the training set, but reduces the confidence in the correct class. Compared with a regularizer, one-sided label smoothing effectively avoided the misclassification problem caused by high regularization coefficient. In order to make the generation results of the generator more stable, in addition to improving its diversity, we also applied different update rules for the generator and discriminator (Heusel et al., 2017). The generator and the discriminator were updated at the same frequency, but trained with different learning rates (a ratio of 1:3) to stabilize the gradient of the generator. Thus, efficient generative deep learning was achieved for fuel design. In the as-developed LIGANDS model, discriminator learned comprehensive fuel characteristics and eliminated overconfidence in classification to properly guide the generator. Meanwhile, the generator was endowed with high diversity, good stability and high accuracy for generating new target fuel molecules.

Given 255 hydrocarbon structures as the input, LIGANDS finally generated 3461 qualified new fuel molecules. The statistical results of carbon number and unsaturation for the 3461 new hydrocarbon molecules are shown in Figs. S1 and S2, respectively. The carbon numbers of these molecules are mainly distributed from 11 to 15, of which 48.9% contain 13 carbon atoms. The unsaturation of most molecules are 2–5, of which 91.8% are 3–4.

The fuel properties of those hydrocarbon molecules were predicted by the stacking model. The violin plots in Fig. 4 compare the property distributions of new molecules generated by LIGANDS at different epochs and those of the original fuel molecules. Numerical characteristics of the property data for the generated molecules are clearly visualized. The changes of density and mass calorific value of the generated fuel molecules with the increase of epoch during training are shown in Figs. S3 and S4, respectively. The density, specific impulse and heat value of the as-generated hydrocarbon molecules were constantly increased during training for pursuing high energy density. Meanwhile, the flash point of those molecules was kept at a high value of around 90 ℃. The probability density curves of the properties for different batches of the generated molecules were significantly changed during training. At a middle epoch of 500, the property distributions of the newly generated molecules were similar to those of the original high-energy–density fuels. And the median values of density, heat value, specific impulse and flash point for the new molecules were comparable to or even higher than those of the original ones. The results demonstrate the strong deep learning capability of LIGANDS for generative fuel design.

Fig. 5 displays 16 new molecules with distinctive structures generated by LIGANDS. H atoms were omitted in the ball-and-stick model for a better observation of the spatial structure of the carbon skeleton. Those hydrocarbon structures are new, unique and complex as designed by LIGANDS. Typical fuel properties of these molecules, including density (ρ), flash point (FP), melting point (T_m), specific impulse (I_{sp}), boiling point (T_b), the

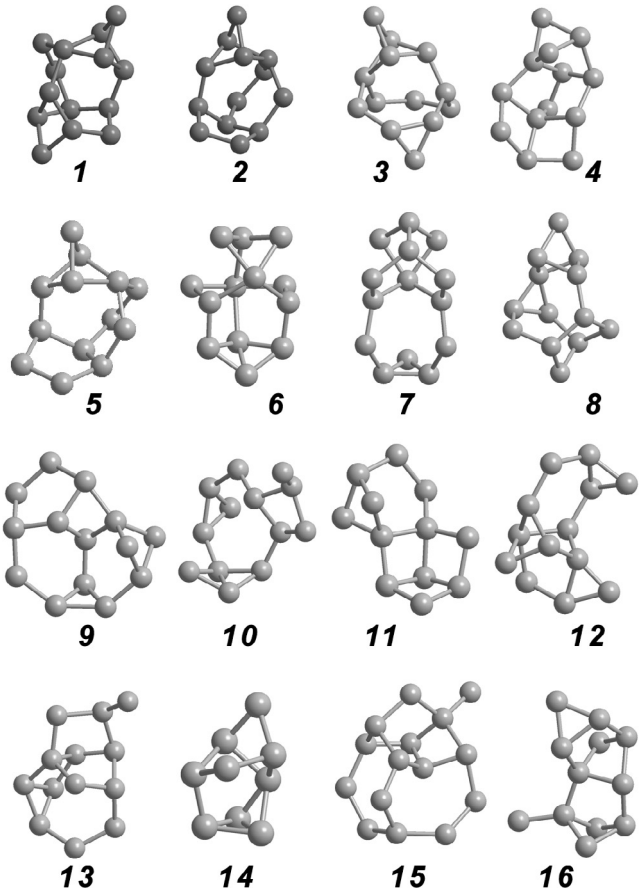


Fig. 5. Molecular structures of 16 new hydrocarbon molecules generated by LIGANDS. H atoms were omitted for clear observation.

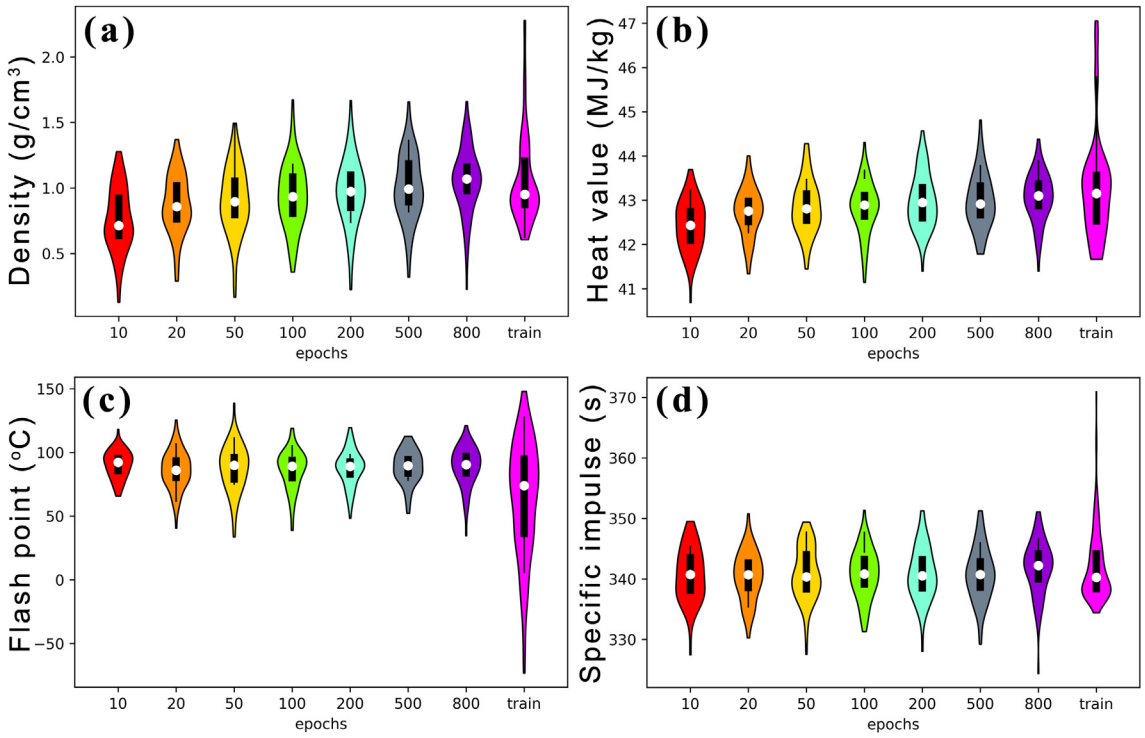


Fig. 4. Violin plots of various fuel properties for the molecules from the original dataset and the newly generated ones by LIGANDS. The highest and lowest points of the outline represent the maximum and minimum values. The side shape represents the cumulative probability density distribution. The inner small black bar represents the 1/4 quantile, and the white dot in the middle represents the median. The inner black line represents the 95% confidence interval.

net heat of combustion (*NHOC*), and SA score were predicted by the stacking model in LIGANDS. For comparison, the methods of group contribution and DFT were also employed to calculate the values of ρ, T_m, *FP*, T_b, *NHOC* and I_{sp} for each hydrocarbon molecule.

In Fig. 6, the values (ρ, T_m, *FP*, T_b, *NHOC* and I_{sp} of the 16 molecules) calculated by group contribution and DFT methods (red circles) and predicted by ensemble learning (black crosses) are compared. Compared with our previous regression results by a single neural network (Li et al., 2020), the accuracy for predicting the properties of unlabeled hydrocarbon molecules have been greatly improved by ensemble learning. The results verify that our stacking model possesses a good generalization ability. It can be concluded that LIGANDS not only automatically generates desirable new fuel molecules, but also accurately predicts their fuel properties.

The new molecules generated by LIGANDS possess certain outstanding fuel properties (high density, high flash point, and large specific impulse), meanwhile their other properties are comparable to those of traditional quadricyclane (QC) and JP-10. The synthetic accessibility (SA) scores of the new hydrocarbon molecules were also predicted by LIGANDS, and the results are summarized in Fig. S5. The SA scores of QC and JP-10 are also shown for easy comparison. However, all new molecules generated by LIGANDS exhibit higher SA scores than those of QC and JP-10. It indicates that the new molecules have lower synthetic accessibility. The synthesis of these molecules will be more complex. In the future, syn-

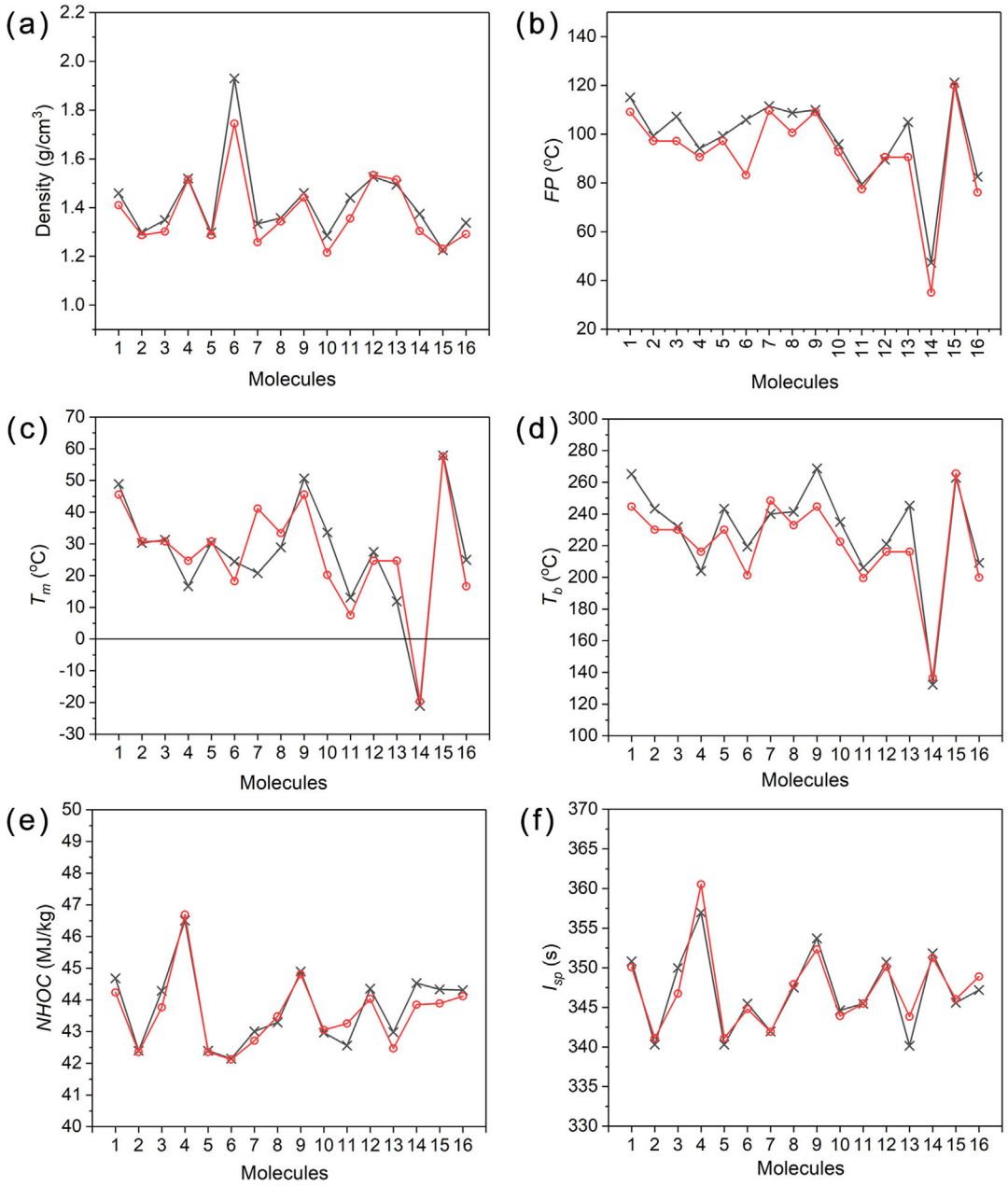


Fig. 6. Comparison of the fuel properties calculated by DFT and group contribution (○ red circles) and predicted by machine learning (×black crosses). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

thetic accessibility should be considered by the model for a more robust fuel design.

4. Conclusions

In summary, an artificial intelligence of generative deep learning termed LIGANDS has been devised and executed for de novo fuel design with multiple desired properties. In LIGANDS, a VAE of hydrocarbon molecules is established to reversibly represent discrete molecular structures in the mathematic form. The VAE also builds a rational latent space of hydrocarbons, in which a GAN generates new desired fuel structures and a stacking model predicts corresponding fuel properties. Based on the multi-objective imitation of the targeted fuel molecules, LIGANDS can generate well-distributed new fuel molecules with comparable and even improved properties. In our proof-of-concept study, the results of generative deep learning demonstrate that LIGANDS is efficient and robust for *de novo* design of next-generation hydrocarbon fuels. In the latent space, all competitive candidates of a given targeted fuel can be automatically and rigorously proposed by LIGANDS. Based on a specific demand, different batches of hydrocarbon molecules meeting the criteria will be efficiently generated by LIGANDS for rational fuel design in the future. In addition, the methodology of generative deep learning based on VAE, ensemble learning and GAN can be easily extended to many other fields for more broad applications, e.g., drug, protein, lubricants, additives, and explosives.

CRediT authorship contribution statement

Yifan Liu: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Writing – review & editing. **Runze Liu:** Methodology, Formal analysis, Writing – review & editing. **Jinyu Duan:** Validation, Visualization, Formal analysis. **Li Wang:** Resources, Supervision. **Xiangwen Zhang:** Supervision, Funding acquisition, Resources. **Guozhu Li:** Project administration, Supervision, Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition, Resources.

Data availability

Data will be made available on request.

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

Financial support for this work is gratefully acknowledged, including the National Natural Science Foundation of China (22178248) and the Haihe Laboratory of Sustainable Chemical Transformations. The DFT calculations (Gaussian 09) were performed on TianHe-1(A) at national supercomputer center in Tianjin.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ces.2023.118686.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467.

Arjovsky, M., Bottou, L., 2017. Towards Principled Methods for Training Generative Adversarial Networks. arXiv:1701.04862.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv:1701.07875.

Blum, L.C., Reymond, J.-L., 2009. 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. J. Am. Chem. Soc. 131, 8732–8733.

Chung, J., Gulçehre, C., Cho, K., Bengio, Y., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. http://arxiv.org/abs/1412.3555.

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., Aspuru-Guzik, A., 2018. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. 4, 268–276.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML].

Guo, Z., Lim, K.H., Chen, M., Thio, B.J.R., Loo, B.L.W., 2017. Predicting cetane numbers of hydrocarbons and oxygenates from highly accessible descriptors by using artificial neural networks. Fuel 207, 344–351.

Han, W., Sun, Z., Scholtissek, A., Hasse, C., 2021. Machine Learning of ignition delay times under dual-fuel engine conditions. Fuel 288, 119650.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500.

Heyne, J., Bell, D., Feldhausen, J., Yang, Z., Boehm, R., 2022. Towards fuel composition and properties from Two-dimensional gas chromatography with flame ionization and vacuum ultraviolet spectroscopy. Fuel 312, 122709.

Hou, F., Wu, Z., Hu, Z., Xiao, Z., Wang, L., Zhang, X., Li, G., 2018. Comparison study on the prediction of multiple molecular properties by various neural networks. Chem. A Eur. J. 122, 9128–9134.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:1710.10196 [cs.NE].

Kirkpatrick, P., Ellis, C., 2004. Chemical space. Nature 432, 823–823.

Lehn, F.V., Brosius, B., Broda, R., Cai, L., 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.

Li, G., Hu, Z., Hou, F., Li, X., Wang, L., Zhang, X., 2020. Machine learning enabled high-throughput screening of hydrocarbon molecules for the design of next generation fuels. Fuel 265, 116968.

Liu, J., Gong, S., Li, H., Liu, G., 2022a. Molecular graph-based deep learning method for predicting multiple physical properties of alternative fuel components. Fuel 313, 122712.

Liu, R., Liu, R., Liu, Y., Wang, L., Zhang, X., Li, G., 2022b. Design of fuel molecules based on variational autoencoder. Fuel 316, 123426.

Liu, R., Liu, Y., Duan, J., Hou, F., Wang, L., Zhang, X., Li, G., 2022c. Ensemble learning directed classification and regression of hydrocarbon fuels. Fuel 324, 124520.

Lu, X., Han, D., Huang, Z., 2011. Fuel design and management for the control of advanced compression-ignition combustion modes. Prog. Energy Combust. Sci. 37, 741–783.

Marrero, J., Gani, R., 2001. Group-contribution based estimation of pure component properties. Fluid Phase Equilib. 183–184, 183–208.

Nal Kalchbrenner, E.G., Phil Blunsom, 2014. A Convolutional Neural Network for Modelling Sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, pp. 655-665.

Osmont, A., Gökalp, I., Catoire, L., 2006. Evaluating missile fuels. Propellants Explos. Pyrotech. 31, 343–354.

Osmont, A., Catoire, L., Gökalp, I., 2008. Physicochemical properties and thermochemistry of propellanes. Energy Fuel 22, 2241–2257.

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

Reymond, J.-L., 2015. The chemical space project. Acc. Chem. Res. 48, 722–730.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved Techniques for Training GANs. arXiv:1606.03498 [cs.LG].

Schweidtmann, A.M., Rittig, J.G., König, A., Grohe, M., Mitsos, A., Dahmen, M., 2020. Graph neural networks for prediction of fuel ignition quality. Energy Fuel 34, 11395–11407.

Shi, X., Li, H., Song, Z., Zhang, X., Liu, G., 2017. Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector. Fuel 200, 395–406.

Sønderby, C.K., Caballero, J., Theis, L., Shi, W., Huszár, F., 2016. Amortised MAP Inference for Image Super-resolution. arXiv:1610.04490 [cs.CV].

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks, in: Proceedings of the 27th International Conference on Neural

Information Processing Systems - Volume 2. MIT Press, Montreal, Canada, pp. 3104–3112.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs.CV].
Tang, J., Alelyani, S., Liu, H., 2015. Data Classification: Algorithms and Applications. Chapman and Hall/CRC.
Warde-Farley, D., Goodfellow, I., 2016. Adversarial perturbations of deep neural networks. In: Hazan, T., Papandreou, G., Tarlow, D. (Eds.), Perturbations, Optimization, and Statistics. MIT Press.
Wheeler, S.E., Houk, K.N., Schleyer, P.V.R., Allen, W.D., 2009. A hierarchy of homodesmotic reactions for thermochemistry. J. Am. Chem. Soc. 131, 2547–2560.
Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5, 241–259.
Yalamanchi, K.K., Nicolle, A., Sarathy, S.M., 2022. Chapter 3 – artificial intelligence–enabled fuel design. In: Badra, J., Pal, P., Pei, Y., Som, S. (Eds.), Artificial Intelligence and Data Driven Optimization of Internal Combustion Engines. Elsevier, pp. 47–67.
Yue, L., Li, G., He, G., Guo, Y., Xu, L., Fang, W., 2016. Impacts of hydrogen to carbon ratio (H/C) on fundamental properties and supercritical cracking performance of hydrocarbon fuels. Chem. Eng. J. 283, 1216–1223.
Zhang, X., Jia, T., 2020. High-Energy-High Density Fuels for Advanced Propulsion: Design and Synthesis, in: Ji-Jun Zou, Xiangwen Zhang, Pan, L. (Eds.), Chromatographia, pp. 5–38.
Zhang, Y., Gan, Z., Carin, L., 2016. Generating Text via Adversarial Training, NIPS 2016, Barcelona, Spain.
Zhang, X., Pan, L., Wang, L., Zou, J.-J., 2018. Review on synthesis and properties of high-energy-density liquid fuels: hydrocarbons, nanofluids and energetic ionic liquids. Chem. Eng. Sci. 180, 95–125.
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC.