Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

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

Research
Materials Genome Engineering—Article

Machine Learning-Assisted High-Throughput Virtual Screening for On-Demand Customization of Advanced Energetic Materials

Siwei Song, Yi Wang *, Fang Chen, Mi Yan, Qinghua Zhang *

Institute of Chemical Materials, China Academy of Engineering Physics, Mianyang 621900, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Energetic materials
Machine learning
High-throughput virtual screening
Molecular properties
Synthesis

ABSTRACT

Finding energetic materials with tailored properties is always a significant challenge due to low research efficiency in trial and error. Herein, a methodology combining domain knowledge, a machine learning algorithm, and experiments is presented for accelerating the discovery of novel energetic materials. A high-throughput virtual screening (HTVS) system integrating on-demand molecular generation and machine learning models covering the prediction of molecular properties and crystal packing mode scoring is established. With the proposed HTVS system, candidate molecules with promising properties and a desirable crystal packing mode are rapidly targeted from the generated molecular space containing 25 112 molecules. Furthermore, a study of the crystal structure and properties shows that the good comprehensive performances of the target molecule are in agreement with the predicted results, thus verifying the effectiveness of the proposed methodology. This work demonstrates a new research paradigm for discovering novel energetic materials and can be extended to other organic materials without manifest obstacles.

1. Introduction

Energetic materials are a class of special reactive substances that can release enormous amounts of energy through intense redox reactions under certain external stimuli. Such materials have substantially contributed to the progress and prosperity of humankind since the discovery of black powder in ancient China more than 2000 years ago [1,2]. In the use of advanced energetic materials, energy, sensitivity, and thermostability are the three properties of most concern [3–6]. However, a relationship of mutual contradiction and restriction is always present between energy, sensitivity, and thermostability. In general, high energy of energetic materials is always accompanied by increased mechanical sensitivity and decreased thermostability. Therefore, it remains a great challenge to develop new energetic materials that simultaneously possess high energy, low sensitivity, and good thermostability.

Empirical models for guiding the design of energetic materials have been developed, such as the Kamlet–Jacobs equation for predicting detonation properties and the nitro charge method for predicting mechanical sensitivity [7,8]. Nevertheless, these empirical models are seldom used for the large-scale prescreening of energetic materials before experimental synthesis, because the time-consuming quantum calculations are unaffordable and the inference capabilities are indeterminable. For a long time, the discovery of new energetic materials has relied heavily on scientific intuition through experiments and traditional trial-and-error process [9], which suffer from low efficiency and high uncertainty [10].

With the coming of the big data era, the research paradigm for energetic materials has undergone profound changes [11,12]. Compared with empirical models, machine learning models usually have various advantages in terms of accuracy, generalization, and the capacity to cope with nonlinear problems [13], and are therefore widely used in various fields of material science [14–22]. Herein, we demonstrate a machine learning-assisted high-throughput virtual screening (HTVS) system for the accelerated discovery of new energetic materials with well-balanced energy–safety properties. This HTVS system integrates machine learning models with high-throughput molecule generation and helps to rapidly filter out promising target molecules from 25 112 generated molecular structures. The screened compounds also have a relatively high possibility of possessing a graphite-like crystal

* Corresponding authors.
E-mail addresses: ywang0521@caep.cn (Y. Wang), qinghuazhang@caep.cn (Q. Zhang).

https://doi.org/10.1016/j.eng.2022.01.008

Please cite this article as: S. Song, Y. Wang, F. Chen et al., Machine Learning-Assisted High-Throughput Virtual Screening for On-Demand Customization of Advanced Energetic Materials, Engineering, https://doi.org/10.1016/j.eng.2022.01.008

structure, since this specific crystal packing mode generally demonstrates better energy–safety characteristics. After further evaluation of synthetic feasibility, a promising fused [5,6]bi-heterocyclic backbone-based energetic material—namely 7,8-dinitropyrazolo[1,5-*a*][1,3,5]triazine-2,4-diamine (herein referred to as ICM-104)—was synthesized through three-step reactions. A study of the properties of the synthesized material revealed that this new energetic material has good comprehensive properties, including high energy, low sensitivity, and good thermostability. These findings demonstrate the effectiveness of the proposed HTVS system, as well as the great potential of machine learning in designing high-performance energetic materials.

2. Methods

2.1. Data preparation and augmentation

More than 1000 pieces of data on energetic materials were gathered from the literature from the past few decades in order to train the property regression models. The dataset contained molecules with various structures and covered aliphatic, aromatic, monocyclic, and polycyclic compounds (see Dataset1 in Appendix A for detailed samples and data sources). More features about the dataset, such as data distribution, are provided in Fig. S1 in Appendix A. Before training the regression model, all data were randomly split into training data and testing data in a ratio of 80:20. The training data was further split into a training set and a validation set for the five-fold cross-validation of the training models and tuning hyperparameters. That is, the validation set comprised five sections, each of which was used once for validation, while the remaining four sections were used as the training set. The final test scores were calculated on the hold-out testing data, which had not been used in the training process.

To train the classification model, we prepared 365 entries labeled "0" (indicating not graphite-like) and 22 entries labeled "1" (indicating graphite-like) from the Cambridge Crystallographic Data Centre (CCDC) (see Dataset2 in Appendix A). Obviously, the amount of data was too small and not suitable for deep learning. Therefore, we augmented the data using the simplified molecular input line entry specification (SMILES) enumeration trick, which generated multiple different SMILES strings that represented the same molecule. SMILES enumeration, as proposed by Arús-Pous et al. [23] and Bjerrum [24], is a novel data-augmentation technology for molecular deep learning. The SMILES labelled as "0" and "1" were enlarged by 10-fold and 30-fold, respectively. After augmentation, the total sample size was enlarged to more than 4000. When training the convolutional neural network (CNN) and long short-term memory (LSTM) model, 400 samples were held back to evaluate the performance of the proposed model.

2.2. Feature and model

Features (i.e., molecular descriptors) including custom descriptors and electro-topological fingerprints were extracted using the RDKit library. Property models were trained by means of a kernel ridge regression (KRR) algorithm, which was implemented in the Scikit-learn package. In the KRR algorithm, the prediction value (y^*) can be expressed as the weighted average (α_i) of the inner product between the new sample (x^*) and the training samples (x), given a kernel function (k) (Eq. (1)). Therefore, the learning process involves calculating the coefficient matrix (α, and α_i is the i-th entry of α) using Eq. (2), in which X ,Y, λ, and I are the sample matrix, label matrix, regularization parameter, and identity matrix, respectively. Hyperparameters including kernel function were tuned using the grid-search method and five-fold cross-validation. The coefficient of determination (R^2) was chosen as the refit score (Eq. (3)), where $\bar{y}$ is mean value. The mean absolute error (MAE) was used to evaluate the model performance, and is given by Eq. (4). In all equations, i and N refer to i-th sample and total number of samples.

$$y^* = \sum_{i=0}^{N-1} \alpha_i k(x^*, x_i) \tag{1}$$

$$\alpha \triangleq \left[k\left(X, X^T\right) + \lambda I\right]^{-1} Y \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=0}^{N-1}\left(y_i - y_i^*\right)}{\sum_{i=0}^{N-1}\left(y_i - \bar{y}\right)} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=0}^{N-1}\left|y_i - y_i^*\right| \tag{4}$$

The CNN and LSTM for the classification model were constructed using the Pytorch package. To prepare the inputs, a dictionary was automatically abstracted from the SMILES of the whole dataset. The details of the dictionary were as follows: ['N', 'c', '1', 'n', '(', ')', '[', '+', ']', '=', 'O', '-', 'o', '2', '#', 'C', '3', 'H', '/', '\\', '4', '5', 'None'] (none for padding). Accordingly, the SMILES string was transformed into a two-dimensional (2D) array with a size of [120, 23]. For the LSTM model, the length limit of the SMILES was 120, and the allowed characters were identical to those of the dictionary. Furthermore, the CNN contained two 2D convolutional layers and three full-connection layers. The 2D convolutional layers had filter sizes of 16 and 32, whereas the kernel sizes were 7 each. The max pooling layer had a kernel size of 2. The full-connection layers had widths of 800, 100, and 2, respectively. A rectified linear unit (ReLU) was used for the activation function. The LSTM possessed a hidden size of 64 and a layer number of 20. For both the deep learning architectures, the loss function was defined by cross-entropy, and an Adam optimizer with a learning rate of 0.001 was used to update the weights. Accuracy (defined by Eq. (5)), balanced accuracy (defined by Eq. (6)), and F_1 score (defined by Eq. (7)) were chosen as the metrics for evaluating the performance of the model, in which TP, FP, TN, and FN represent the true positives, the false positives, the true negatives, and the false negatives, respectively. To illustrate the necessity of deep learning architectures, the K-nearest neighbor (KNN) based on the descriptors was tested as the baseline. However, the SMILES enumeration was not applied to train the KNN model, since the descriptors based on different SMILES representing the same molecule were completely identical.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=0}^{N-1} 1\left(y_i = y_i^*\right) \tag{5}$$

$$\text{Balanced\_accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right) \tag{6}$$

$$F_1 \text{ score} = 2\left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

To score the possibility, the SMILES enumeration trick was used in the prediction process. Out of 20 SMILES representing the same

molecule, the proportion of molecules with graphite-like structures (p) could be obtained after classification (Eq. (10)). The above process was repeated ten times to alleviate the randomness in the SMILES enumeration, and the sum of p was taken as the final score (Eq. (11)).

$$p = \frac{\sum_{i=1}^{20} y_i^*}{20},\ y_i^* \in \{0, 1\} \tag{10}$$

$$Score = \sum_{i=1}^{10} p_i \tag{11}$$

2.3. Preparation and characterization

Although the compounds reported here have quite low sensitivity to external mechanical stimuli (e.g., impact and friction), highly corrosive concentrated sulfuric acid is used in the synthesis process. Thus, we recommend the use of safety equipment such as protective gloves, coats, face shields, and explosion-proof baffles.

2.3.1. Preparation of 4-nitro-1H-pyrazole-3,5-diamine hydrochloride

4-Nitro-1*H*-pyrazole-3,5-diamine was prepared according to a previously reported route [25]. Concentrated hydrochloric acid (3 mL) was added to a suspension of 4-nitro-1*H*-pyrazole-3,5-diamine (3 mmol, 0.429 g) in methanol (5 mL). After stirring for 10 min, the resulting light yellow solid was filtered and washed using ethyl acetate (EtOAc) to obtain 4-nitro-1*H*-pyrazole-3,5-diamine hydrochloride (a yield of 80%).

2.3.2. Preparation of 8-nitropyrazolo[1,5-a][1,3,5]triazine-2,4,7-triamine

This intermediate was prepared according to a previously reported route with slight changes [26]. First, 4-nitro-1*H*-pyrazole-3,5-diamine hydrochloride (3 mmol, 0.54 g) was suspended in ethanol (11 mL). Then, dicyandiamide (4 mmol, 0.33 g) was added to the suspension. The mixture was refluxed at 80 °C for 6 h. During the refluxing, an orange solid gradually appeared in the solution. The orange solid was filtered and recrystallized using water at 80 °C to obtain a yellow solid (8-nitropyrazolo[1,5-*a*][1,3,5]triazine-2,4,7-triamine; yield of 60%).

2.3.3. Preparation of 7,8-dinitropyrazolo[1,5-a][1,3,5]triazine-2,4-diamine (ICM-104)

In an ice-water bath, 8-nitropyrazolo[1,5-*a*][1,3,5] triazine-2,4,7-triamine (3 mmol, 0.63 g) was added to concentrated sulfuric acid (6 mL) in portions. Then, 30% hydrogen peroxide (2.5 mL) was added dropwise to the solution. After stirring at room temperature for 3 h, the reaction was quenched using crushed ice, and the solution was extracted using EtOAc. Then, EtOAc was removed using rotary evaporation. A light yellow solid was collected as the target compound (7,8-dinitropyrazolo[1,5-*a*][1,3,5]triazine-2,4-diamine (ICM-104); yield of 42%). The nuclear magnetic resonance (NMR) data for the target compound was as follows: ^{1}H NMR (dimethyl sulfoxide (DMSO)-d_6, 400 MHz) δ: 8.81 ppm (s, 1H, NH_2), 8.56 ppm (s, 1H, NH_2), 8.04 ppm (s, 1H, NH_2), 7.77 ppm (s, 1H, NH_2); ^{13}C NMR (DMSO-d_6, 100 MHz) δ: 162.41, 153.61, 150.44, 147.42, 109.47 ppm (Fig. S12 in Appendix A). The high-resolution electrospray ionization mass spectrometry (ESI-HRMS) data was as follows: ESI-HRMS: m/z calculated for $[M-H]^-$: 239.0283; found: 239.0282(1). Infrared (IR; KBr, cm^{-1}): 3483.42, 3431.90, 3333.44, 3205.61, 1684.94, 1633.17, 1605.24, 1565.96, 1523.60, 1491.91, 1453.41, 1396.89, 1340.13, 1291.72, 1242.11, 1220.57, 1091.12, 983.45, 881.85, 851.93, 807.86, 784.96, 775.28, 728.80, 714.26, 600.36, 550.32. The calculated elemental analysis was C 25.01%, H 1.68%, and N 46.66%; the found elemental analysis was C 24.67%, H 1.82%, and N 46.40%.

^{1}H and ^{13}C spectra were collected on a Bruker (USA) Avance Neo 400 NMR spectrometer operating at 400 and 100 MHz, respectively. High-resolution mass spectra (HRMS) were collected using a Shimadzu LCMS-IT-TOF™ mass spectrometer with electrospray ionization (ESI). Impact and friction sensitivity measurements were conducted using a standard BAM Fall hammer and a BAM friction tester. The heat of formation of the compound was calculated from the heat of combustion, which was measured using an oxygen bomb calorimeter. The standard detonation properties were calculated using Explo5 (version 6.02) software.

3. Results and discussion

3.1. HTVS system

The framework and components of the HTVS system are shown in Fig. 1. The HTVS system functions as follows (Fig. 1(a)). First, a large number of energetic molecules can be generated by the high-throughput molecular generation module (Fig. 1(b)). Then, the generated molecules are imported into the property predictor to undergo rapid and accurate property calculations. The property predictor contains four trained models for density, detonation velocity, detonation pressure, and decomposition temperature, using the same composite molecular descriptors set as the input (Fig. 1(c)). By virtue of this property predictor, potential molecules with relatively high energy, low sensitivity, and good thermostability can be filtered out based on the predicted properties. The preliminarily screened molecules with the desired properties are then fed into the crystal structure classifier to further evaluate the possibility of forming a graphite-like layered crystal structure. Finally, after evaluating the feasibility of synthesis, molecules with promising properties and a high probability of forming a graphite-like layered crystal structure are selected for experimental synthesis and characterization. This HTVS system can help experimental chemists customize energetic materials through the molecular generation and screening process, rather than spending a great deal of time and effort on trial and error.

3.2. Feature set and property models

Aside from the data, the feature (i.e., the molecular descriptor) is another factor that determines the accuracy of the machine learning model. Our composite feature set (CDS) is composed of two parts. The first part comprises the fingerprints related to carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and the halogen elements from the electro-topological state (E-state) fingerprint, which have been widely used to construct different models for predicting molecular properties [27–29]. However, domain knowledge can reduce the learning complexity and improve the accuracy for a specific task; therefore, we defined a custom descriptor set containing another 29 molecular descriptors (Table S2 in Appendix A). This custom descriptor set enhances the description of molecular shape and composition, such as the plane of best fit (PBF) and oxygen balance (OB), which will be helpful in learning about the properties of energetic materials. The correlation of the custom descriptors on the density data was visualized using a heat map (Fig. 2(a)). The heat map demonstrates that most of the custom descriptors are not significantly correlated, which is beneficial for the training model.

Principal component analysis (PCA) was used to visualize the capacity of our CDS for capturing the underlying model in density data [30]. When the features were combined into 45 principal components, the cumulative variance reached the value of 0.993

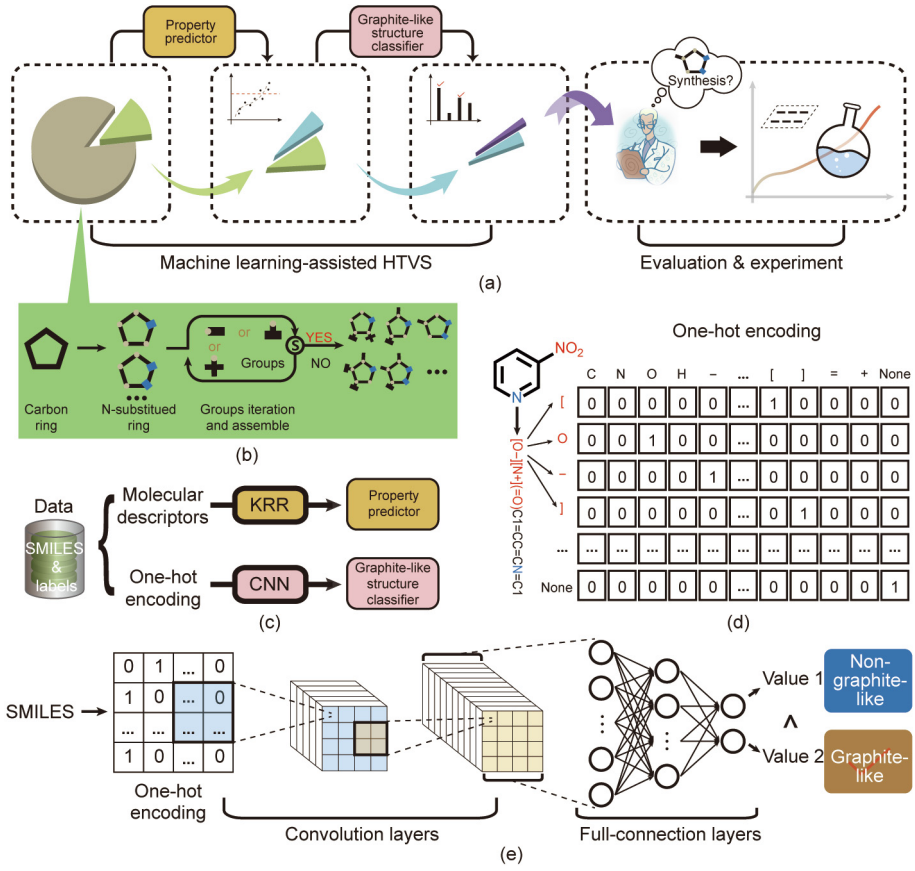


Fig. 1. Framework and components of the HTVS system. (a) Framework of machine learning-assisted HTVS; (b) schematic of molecule generation using heuristic enumeration; (c) schematic of the training of property models and the graphite-like structure classification model; (d) one-hot encoding for the input of the CNN; (e) architecture of the CNN.

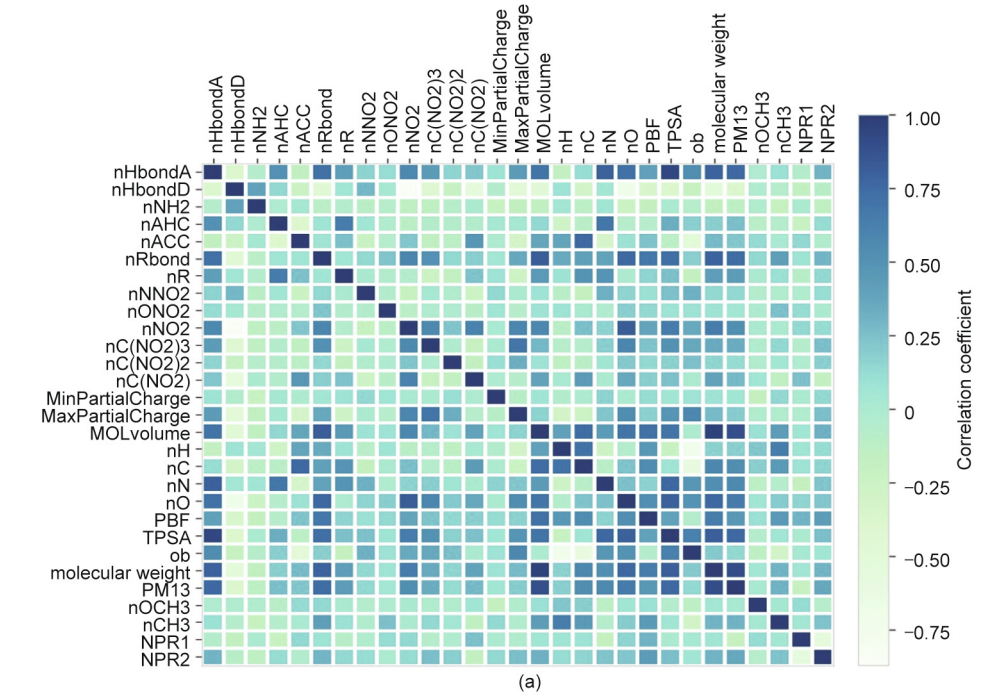
(Fig. 2(b), left). Furthermore, the most informative projections of principal components (PC14 and PC2) were visualized (Fig. 2(b)). Sample distributions with different densities were relatively concentrated and an obvious color gradient was observed, implying that the features were effective in capturing the underlying model in the density data.

After training the model using the KRR algorithm [31], we validated the performance of the model for predicting density by comparing the observed and predicted values on the training and test sets, respectively (Fig. 2(c)). We found a remarkable agreement between the observed and predicted values (Fig. 2(c)), and the deviation between them followed an almost normal distribution (Fig. 2(c), right). In the learning curves, with the increase in the training sample, both the training (red) and cross-validation (green) curves gradually approached the same asymptote (Fig. 2(d)), indicating that our model was trained well (i.e., over-fitting or under-fitting was not observed). The coefficient of determination (R^2) and MAE for the test dataset were 0.93 and 0.042 g·cm^{-3}, respectively (Fig. 2(e)). The high accuracy of the density model may stem from the large amount of data and the reasonable featurization method, which can capture both molecular and crystal characteristics to some degree. With the same composite molecular descriptor set as the input, the prediction models for the detonation velocity (D_v), detonation pressure (P), and decomposition temperature (T_d) were all trained. As shown in Fig. 2(e), the R^2 values for the test dataset of the D_v, P, and T_d models were 0.83 (MAE: 236.3 m·s^{-1}), 0.82 (MAE: 2.379 GPa), and 0.62 (MAE: 30.8 °C), respectively. (For the training and evaluation of these models, see Fig. S2 in Appendix A.) More results for the cross-validation score and training stability test are summarized in Table S3 in Appendix A. It is noticeable that, compared with past works, our models show a competitive performance in terms of accuracy, effectiveness, and comprehensiveness (Table S4 in Appendix A). Apart from the above four properties (e.g., density, detonation velocity, detonation pressure, and decomposition temperature), sensitivity is a core property for energetic materials. However, training a general model for sensitivity is still difficult, since sensitivity is correlated with multiscale factors including the electronic structure, crystal structure, and even measurement conditions. Therefore, an alternative method for tackling sensitivity prediction remains highly desired.

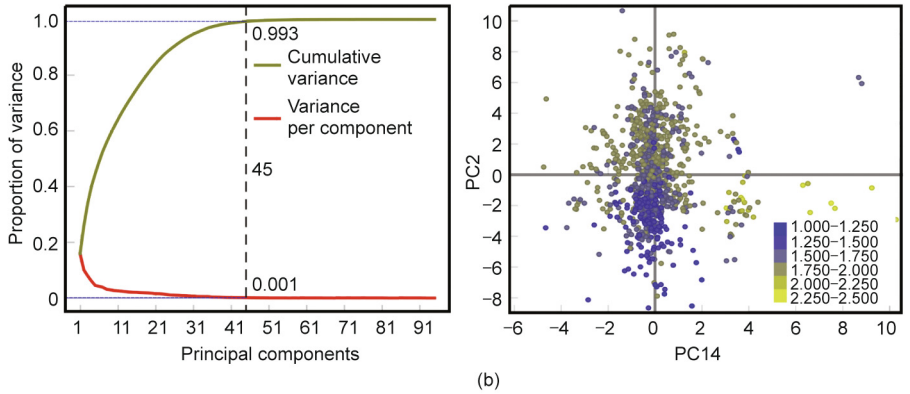
3.3. Classification model for a graphite-like layered crystal structure

To find a more reliable method for rapidly screening potential energetic molecules with low sensitivity, we tried transferring the direct prediction of impact sensitivity into a special structural identification of graphite-like layered crystal packing, as there is a widely recognized close correlation between a graphite-like layered crystal structure and low impact sensitivity in energetic materials [32–34]. The crystal structure is related to the molecular structure; in particular, certain functional groups that tend to form strong non-bonding interactions may dominate the formation of crystals. In some previous studies, a deep neural network was used for predicting the crystal structure, which inspired us to seek help from deep learning [35,36].

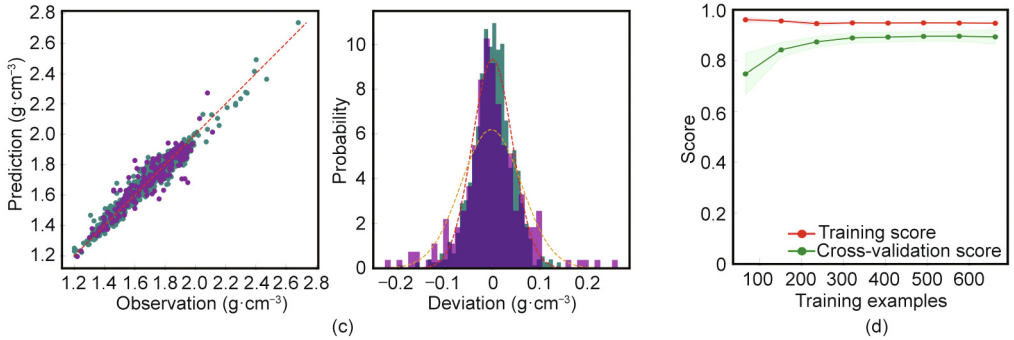
From the above considerations, a CNN and LSTM [37,38] were chosen to capture the chemical intuition that can distinguish



(a)



(b)



(c) (d)

Test score	Density	D_v	P	T_d
R^2	0.93	0.83	0.82	0.62
MAE	0.042 g·cm^{-3}	236.3 m·s^{-1}	2.379 GPa	30.8 °C

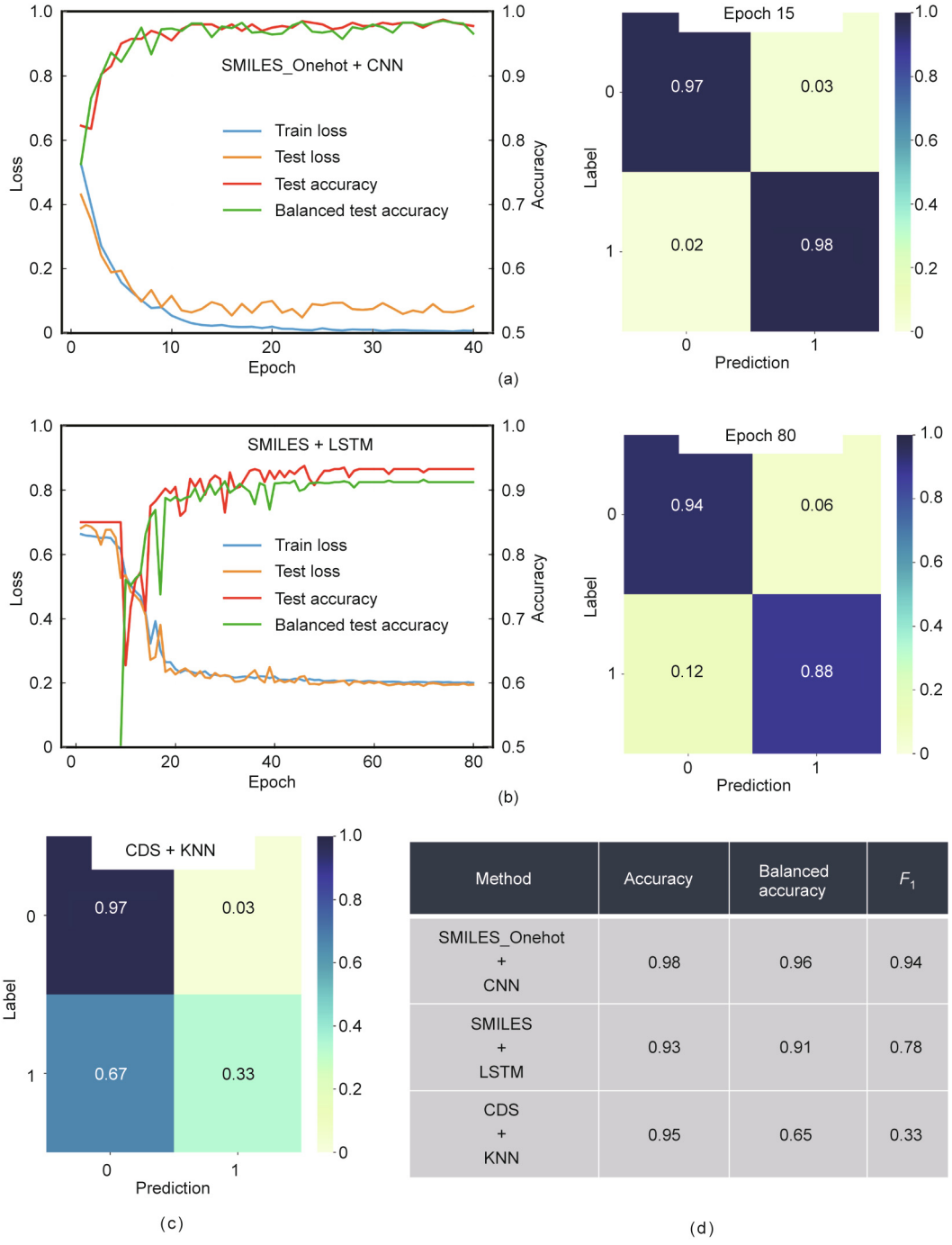
(e)

Fig. 2. Feature distribution and model evaluation of property models. (a) Custom descriptor set and its heatmap of feature distribution on density data; (b) PCA of features and scatter for most informative components on density data; (c) parity plot and deviation distribution on the training set (green) and test set (purple) of density data, where the red (orange) dashed line is the normal distribution curve of deviation for the training (test) data; (d) learning curves (training curve in red and cross-validation curve in green) of density model; (e) test scores for the four trained models (D_v: detonation velocity; P: detonation pressure; T_d: decomposition temperature).

among molecules with regards to possible graphite-like crystal structures. The CNN was trained using the one-hot encoding of molecular SMILES strings as input (Figs. 1(c, d)) [39,40], and bears a typical architecture (Fig. 1(e)). The LSTM was directly trained using SMILES as the input. In addition, a KNN model using CDS as the input (the CDS + KNN model) was trained as a baseline. A

comparison of the training process, as shown in Fig. 3, indicates that the SMILES_Onehot + CNN model was better than the SMILES + LSTM model, because the train and test losses for the former were lower, whereas the accuracy/balanced accuracy of the former model was higher than that of the latter model. The confusion matrix of the dumped model for SMILES_Onehot + CNN (epoch 15) with the lowest test loss also behaved better than that of SMILES + LSTM, since the latter had a stronger tendency to misclassify graphite-like (1) as non-graphite-like (0) molecules. In contrast, the CDS + KNN model demonstrated poor performance, especially in terms of the balanced accuracy (0.65) and confusion matrix. This phenomenon is understandable, since in the CNN and LSTM models, more information about the molecular structure (e.g., the arrangement of atoms and substituted groups, which we consider to be vital for predicting crystal packing) remained, while in the CDS + KNN model, these pieces of information were compressed during the featurization process. We also tried simpler architectures (e.g., decision tree and neural network based on CDS, as shown by the data presented in Table S5 in Appendix A), which showed that SMILES_Onehot + CNN exhibited an absolute advantage with regards to accuracy.

Finally, the SMILES_Onehot + CNN model was integrated with the SMILES enumeration trick to evaluate the possibility of potential molecules having a graphite-like crystal structure [41]. The



Method	Accuracy	Balanced accuracy	F_1
SMILES_Onehot + CNN	0.98	0.96	0.94
SMILES + LSTM	0.93	0.91	0.78
CDS + KNN	0.95	0.65	0.33

Fig. 3. Comparison among classification models. (a) Training process and confusion matrix for the SMILES_Onehot + CNN model; (b) training process and confusion matrix for the SMILES + LSTM model; (c) confusion matrix for the CDS + KNN model; (d) model metrics for the test data.

possibility value indicates the tendency of one molecule to form a graphite-like layered structure; therefore, it helps us to sort and assess these molecules from high to low likelihood. In this way, the screening step for graphite-like layered crystal structures became more robust.

3.4. High-throughput generation and screenings of energetic molecules

We used a heuristic enumeration method to generate the molecules (Fig. 1(b)) through homemade scripts (Fig. S3 in Appendix A) [42,43]. In recent years, researchers have shown increasing interest in fused heterocycle ring-based energetic materials (e.g., fused [5,5]biheterocyclic and [5,6]biheterocyclic energetic molecules). In this regard, a series of promising fused-ring energetic molecules have been reported [44–48]. Herein, we focus on energetic molecules that are constructed from a fused [5,6]biheterocyclic backbone and substituted nitro/amino groups.

Initially, the input structure for the molecular generation module contained five different [5,6]bicyclic carbon rings. After the N-substitution (from 1N to 7N) process, we obtained 355 different fused [5,6]biheterocyclic skeletons (Fig. 4(a)). Based on the acceptable time consumption for molecular generation and the feasibility of experimental synthesis, the most substituent sites in the fused [5,6]biheterocyclic skeleton were limited to four (see scatter plots in Fig. S2). Consequently, 25 112 possible fused [5,6]biheterocyclic molecules were generated, which involved the introduction of nitro/amino groups into 355 different fused [5,6]biheterocyclic skeletons after structure sanitization and deduplication. As shown in Fig. S4 in Appendix A, the generated molecules were close to the domains of applicability of the models.

The generated 25 112 energetic molecules were then fed into the property predictor to predict their properties (including density, D_v, P, and T_d) and screen them (see the predicted results in the Supplementary Data 1). The whole explored molecular space and the step-by-step screening can be visualized in color-mapped three-dimensional (3D) scatter plots (Fig. 4(b)) and doughnut charts (Fig. 4(c)). The predicted properties of the 25 112 molecules were in accordance with some common rules for energetic materials, such as a linear correlation between the density and D_v/P. A negative correlation between the density and the decomposition temperature (Fig. 4(b)) was observed. We took the density (1.80 g·cm^{-3}) of a typical energetic material (1,3,5-trinitro-1,3,5-triazinane, RDX) as the first criterion to screen. The number of molecules with a density greater than 1.80 g·cm^{-3} sharply decreased from the original 25 112 molecules to only 3141 (Fig. 4(b)). The color-mapped 3D scatter plot shows that the molecules with T_d above 280 °C (red dots) were mostly located in the region with relatively low D_v values (around 8000 m·s^{-1}). However, the molecules with D_v greater than 8800 m·s^{-1} (blue dots) were mostly located in the region with relatively low T_d values

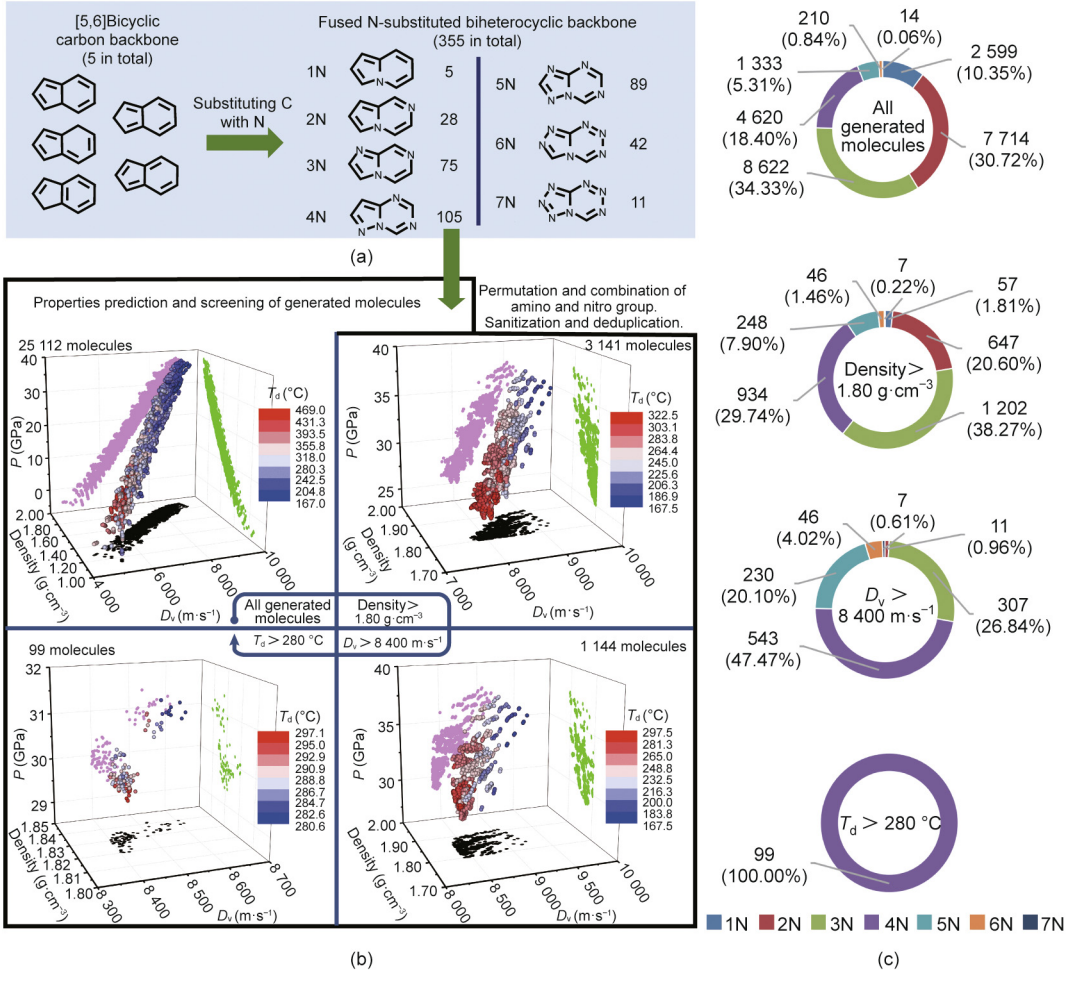


Fig. 4. Process of generating and screening the molecules. (a) Illustration of the generation process of the [5,6]biheterocyclic backbone; (b) color-mapped 3D scatter plots of the molecules in original and different screening steps (black, green, and pink dots represent the projections on the density/D_v plane, density/P plane, and D_v/P plane, respectively); (c) proportions of different nitro-substituted fused [5,6]biheterocyclic molecules in original and different screening steps.

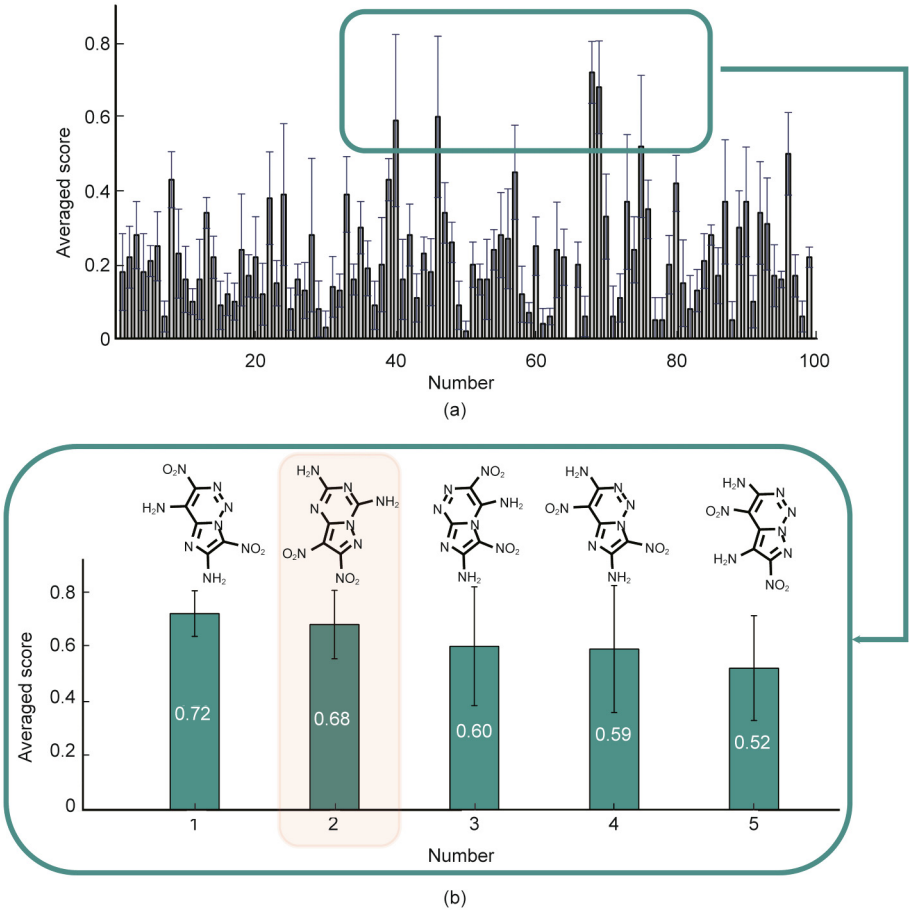


Fig. 5. Score for forming a special graphite-like layered crystal structure. (a) Average score for forming a graphite-like layered crystal structure for 99 candidates (error bars indicate the mean deviation for five predictions); (b) structures for the first five molecules sorted according to their respective averaged score.

(around 160 °C) (Fig. 4(b)). When the screening criteria for energy (D_v > 8400 m·s^{-1}) and thermostability (T_d > 280 °C) (for criteria determination, see Fig. S5 in Appendix A) were separately introduced, the number of molecules satisfying the requirements decreased from 3141 to 1144 (Fig. 4(b)); finally, they reached the value of 99 (Fig. 4(b) and Appendix A Fig. S6).

The color-filled doughnut charts clearly show the variations in the proportions of the different nitrogen-substituted [5,6]biheterocyclic molecules with the gradual introduction of the screening criteria (Fig. 4(c)). After introducing the screening criteria for density (> 1.80 g·m^{-3}) and energy (D_v > 8400 m·s^{-1}), the proportions of five- (sky blue color), six- (orange color), and seven- (navy blue color) nitrogen-atom-substituted fused [5,6]biheterocyclic molecules increased from 5.31%, 0.84%, and 0.06% to 20.10%, 4.02%, and 0.61%, respectively, which implies that a high content of nitrogen in the molecular skeleton is beneficial for increasing the energy (high density and D_v) of the fused [5,6]biheterocyclic molecules. However, a high nitrogen content will decrease the molecular thermostability, causing the decomposition temperatures to be lower than 280 °C. In contrast, the proportions of one- (blue color) and two- (red color) nitrogen-atom-substituted fused [5,6]biheterocyclic molecules decreased from 10.35% and 30.72% to 0% and 0.96%, respectively, under the screening criteria for density (> 1.80 g·cm^{-3}) and energy (D_v > 8400 m·s^{-1}), indicating the negative influence of low nitrogen content on the energy of the molecules. After screening by density (> 1.80 g·cm^{-3}) and energy (D_v > 8400 m·s^{-1}), three-nitrogen-substituted fused [5,6]biheterocyclic molecules (green color) showed a relatively high percentage (26.84%) among the filtered 1144 candidates; however, their decomposition temperatures could not satisfy the criterion for high thermostability (T_d > 280 °C), mainly because the nitrogen content for three-nitrogen-substituted fused [5,6]biheterocyclic molecules is still relatively low. The screened molecules that meet the criteria of both density and energy usually contain multiple nitro groups (around three or four) (Fig. S7 in Appendix A), although the strong electron-withdrawing effect of multiple nitro groups will reduce the molecular stability to dissatisfy the criterion of decomposition temperature (T_d > 280 °C). Overall, after three-step screening, all 99 screened molecules had four nitrogen atoms (violet color; 4N) substituted into the fused [5,6]biheterocyclic molecules, because both the nitrogen content in their molecular skeletons and the number of nitro groups were reasonable.

These 99 energetic molecules were then imported into the graphite-like structure classifier to evaluate their score for forming a special graphite-like layered crystal structure. The prediction for each molecule was repeated five times; the results are summarized in Fig. 5(a) and Dataset2. Based on the sorted averaged score, from high to low, the first five molecular structures are shown in Fig. 5 (b). After evaluating the synthetic accessibility of these five molecules (Fig. S8 in Appendix A), it was found that molecule 2 (as shown in Fig. 5(b); 7,8-dinitropyrazolo[1,5-*a*][1,3,5]triazine-2,4-diamine, herein named ICM-104) had never been reported and was synthetically feasible. Therefore, molecule 2 was selected for subsequent experimentation.

3.5. Synthesis and property studies

Encouragingly, according to the designed synthetic route, we successfully prepared the target molecule ICM-104 by three-step reactions (Section 2.3). After slow solvent evaporation of its EtOAc solution, single crystals of ICM-104 suitable for X-ray diffraction were obtained (Table S6 in Appendix A). As expected, ICM-104 had a graphite-like layered crystal stacking structure with a $P2_1/c$ space group (Fig. 6(a)). In the molecular structure, one nitro group was out of the supramolecular plane (with an angle of 66.7°), which was due to the repulsive interaction of the two adjacent nitro groups (Fig. 6(a)). The supramolecular plane of ICM-104 was constructed by the hydrogen bonds between the amino and nitrogen groups (Fig. 6(a)). This result indicates that our trained graphite-like structure classification model is helpful in identifying new energetic molecules with unique graphite-like crystal packing.

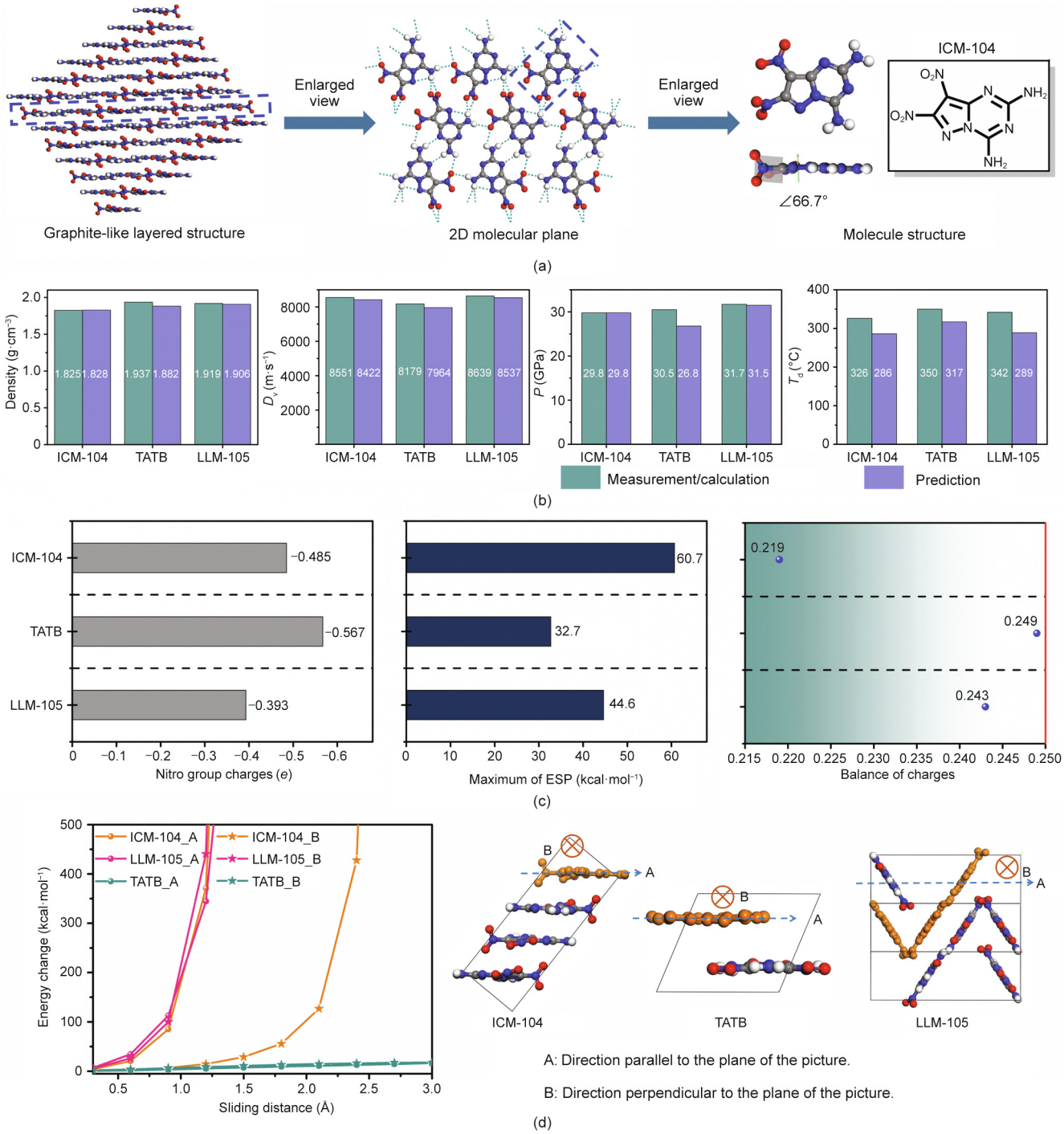


Fig. 6. Crystal structure and properties of ICM-104. (a) 3D graphite-like layered crystal stacking, 2D supramolecular plane, and molecular geometry of ICM-104; (b) comparison between the predicted and measured/calculated properties of ICM-104, 2,4,6-triamino-1,3,5-trinitrobenzene (TATB), and 2,6-diamino-3,5-dinitropyrazine-1-oxide (LLM-105) (blackish green represents the properties measured by experiments or calculated using Explo5 (v6.02), whereas lavender represents the properties predicted by the proposed machine learning models); (c) comparison of nitro group charges, maximum of electrostatic potential (ESP), and balance of charges of ICM-104, LLM-105, and TATB (1 kcal = 4.19 × 10^3 J); (d) energy change for the layer sliding of ICM-104, LLM-105, and TATB, where deep yellow indicates the chosen sliding layer.

After the structural characterizations of ICM-104, our attention turned to evaluating the practicability of the prediction models by comparing the experimental/calculated results with those predicted using the proposed models. As shown in Fig. 6(b), the predicted density, D_v, and P of ICM-104 were 1.828 g·cm^{-3}, 8422 m·s^{-1}, and 29.8 GPa (green histogram in Fig. 6(b)), respectively, all of which were close to the experimental density (1.825 g·cm^{-3}) and calculated D_v and P values (8551 m·s^{-1} and 29.8 GPa; obtained using Explo5 v6.02) (lavender histogram in Fig. 6(b)). The decomposition temperature (T_d) exhibited a clear deviation of around 40 °C between the experimental (326 °C) and predicted (286 °C) results. The main reason for this deviation is that the crystal of ICM-104 is constructed by strong intermolecular hydrogen bonds, whereas our present composite descriptor set is mostly focused on molecular level and has a weak capability for describing intermolecular interactions. The decomposition temperature of ICM-104 was impressive at 326 °C (Fig. S9 in Appendix A), which is close to those of 2,6-diamino-3,5-dinitropyrazine-1-oxide (LLM-105; 342 °C) and 2,4,6-triamino-1,3,5-trinitrobenzene (TATB; 350 °C). The non-isothermal kinetic apparent activation energy (E_a) of ICM-104, as obtained using the Kissinger and Ozawa methods, was 615 and 594 kJ·mol^{-1}, respectively (Fig. S9), indicating the excellent thermal stability of ICM-104. The high decomposition temperature of ICM-104 can be attributed to its graphite-like crystal structure, which is beneficial in providing a better thermostability and relatively stronger trigger bonds (with a bond dissociation enthalpy of 260.63 kJ·mol^{-1}) than that of LLM-105 (247.72 kJ·mol^{-1}; Fig. S10 in Appendix A) [49]. In addition, ICM-104 exhibited low impact (a measured value of 35 J) and friction (a measured value > 360 N) sensitivities. Meanwhile, our predicted results for TATB (1.882 g·cm^{-3}, 7964 m·s^{-1}, 26.8 GPa, and 317 °C) and LLM-105 (1.906 g·cm^{-3}, 8537 m·s^{-1}, 31.5 GPa, and 289 °C) were close to their measured/calculated results (Fig. 6(b)). Through detailed experimental evaluation and a comparison of properties with TATB and LLM-105 (Fig. 6(b) and Table S7 in Appendix A), it was found that ICM-104 is a promising heat-resistant insensitive energetic material.

By considering the molecular structure and crystal packing, we qualitatively elucidated the cause of the low sensitivity of ICM-104. Three molecular factors, including the nitro group charge, maximum of electrostatic potential (ESP), and balance of charges, were utilized to assess the stability of the molecules under mechanical stimulus (these factors were calculated using Gaussian 09 D.01 and Multiwfn 3.7) [50,51]. As shown in Fig. 6(c), among the three compounds, TATB undoubtedly possessed the lowest sensitivity from the molecular aspect. Comparing LLM-105 with ICM-104, the maximum of ESP and balance of charges of LLM-105 (44.6 kcal·mol^{-1} and 0.243, respectively) were better than those of ICM-104 (60.7 kcal·mol^{-1} and 0.219, respectively). Although the nitro group charge of LLM-105 (–0.393e) was a little higher than that of ICM-104 (– 0.485e) [52], the molecular structure of LLM-105 tended to be more stable than that of ICM-104. Furthermore, we calculated the energy change during the layer sliding using the force-field method, so as to evaluate the contribution of crystal packing to low sensitivity. As shown in Fig. 6(d), the intensity of the energy change followed the descending order of LLM-105 > ICM-104 ≫ TATB. Its graphite-like layered crystal structure endowed ICM-104 with a better buffering effect for external mechanical force than the wave-like crystal structure of LLM-105. However, the twisted nitro groups may induce a strong repulsive force between the layers during sliding. Therefore, the energy change of ICM-104 was still more violent than that of TATB. Based on the above analysis, it is reasonable that the mechanical sensitivities of ICM-104 were found to lie between those of LLM-105 and TATB. The distinguished comprehensive properties of ICM-104 can be further highlighted through a comparison with recently reported fused-ring compounds, as shown in Fig. S10. Our machine learning-assisted HTVS system has also been applied to the exploration of energetic melting-castable materials in our recent work [53]. Overall, our self-established machine learning-assisted HTVS system has demonstrated great potential for guiding the discovery of new energetic materials with desired structures and properties.

4. Conclusions

In this work, a machine learning-assisted HTVS system was developed and applied to guide the discovery of energetic materials. This HTVS system integrates high-throughput molecular generation and machine learning models. The high-throughput molecular generation module is responsible for the rapid and extensive generation of suitable molecular structures through heuristic enumeration. The machine learning models are composed of a property predictor and a graphite-like structure classifier. The property predictor contains four well-trained regression models (density, detonation velocity, detonation pressure, and decomposition temperature), and the structure classifier, which is derived from a CNN classification model, is in fact a possibility predictor for a graphite-like layered crystal structure. Based on this HTVS system, we rapidly targeted the promising energetic molecule ICM-104 out of a possible 25 112 [5,6]biheterocyclic molecular structures. Further experimental studies showed that ICM-104 exhibited the expected good performance, including good detonation properties (density = 1.825 g·cm^{-3}, D_v = 8551 m·s^{-1}, and P = 29.8 GPa), low sensitivity (impact sensitivity is 35 J and friction sensitivity is more than 360 N), and good thermostability (onset at 326 °C). This work demonstrates the potential of our machine learning-assisted HTVS system for quickly finding new energetic materials with promising properties. Moreover, the proposed systematic method may be expanded to the discovery of other organic functional materials.

Acknowledgments

The authors acknowledge Dr. Yuyang Wang in Jilin University for searching the synthetic routes in the Reaxys database. We thank the Science Challenge Project (TZ2018004) and the National Natural Science Foundation of China (21875228 and 21702195) for financial support.

Compliance with ethics guidelines

Siwei Song, Yi Wang, Fang Chen, Mi Yan, and Qinghua Zhang declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.01.008.

References

[1] Gao H, Shreeve JM. Azole-based energetic salts. Chem Rev 2011;111(11):7377–436.
[2] Núñez-Quintero D, Hernández-Rivera SP. Spectroscopic modeling of nitro group in explosives. In: Szu HH, editor. Proceedings Volume 6247, Independent Component Analyses, Wavelets, Unsupervised Smart Sensors, and Neural Networks IV; 2006 Apr 17–21; Orlando, FL, USA.
[3] Dippold AA, Klapötke TM. A study of dinitro-*bis*-1,2,4-triazole-1,1′-diol and derivatives: design of high-performance insensitive energetic materials by the introduction of N-oxides. J Am Chem Soc 2013;135(26):9931–8.
[4] Baxter AF, Martin I, Christe KO, Haiges R. Formamidinium nitroformate: an insensitive RDX alternative. J Am Chem Soc 2018;140(44):15089–98.

[5] Zhao G, He C, Kumar D, Hooper JP, Imler GH, Parrish DA, et al. 1,3,5-Triiodo-2,4,6-trinitrobenzene (TITNB) from benzene: balancing performance and high thermal stability of functional energetic materials. Chem Eng J 2019;378:122119.
[6] Li S, Wang Y, Qi C, Zhao X, Zhang J, Zhang S, et al. 3D energetic metal–organic frameworks: synthesis and properties of high energy materials. Angew Chem Int Ed Engl 2013;52(52):14031–5.
[7] Kamlet MJ, Jacobs SJ. Chemistry of detonations. I. A simple method for calculating detonation properties of C-H–N–O explosives. J Chem Phys 1968;48(1):23–35.
[8] Zhang C, Shu Y, Huang Y, Zhao X, Dong H. Investigation of correlation between impact sensitivities and nitro group charges in nitro compounds. J Phys Chem B 2005;109(18):8978–82.
[9] Wang Y, Liu Y, Song S, Yang Z, Qi X, Wang K, et al. Accelerating the discovery of insensitive high-energy-density materials by a materials genome approach. Nat Commun 2018;9(1):2444.
[10] Gu GH, Noh J, Kim I, Jung Y. Machine learning for renewable energy materials. J Mater Chem A 2019;7(29):17096–117.
[11] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the 'fourth paradigm' of science in materials science. APL Mater 2016;4(5):053208.
[12] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature 2018;559(7715):547–55.
[13] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev 2020;49(11):3525–64. Correction in: Chem Soc Rev 2020;49(11):3716.
[14] Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of stable lead-free hybrid organic–inorganic perovskites via machine learning. Nat Commun 2018;9(1):3405.
[15] Takahashi K, Takahashi L. Creating machine learning-driven material recipes based on crystal structure. J Phys Chem Lett 2019;10(2):283–8.
[16] Barnett JW, Bilchak CR, Wang Y, Benicewicz BC, Murdock LA, Bereau T, et al. Designing exceptional gas-separation polymer membranes using machine learning. Sci Adv 2020;6(20):eaaz4301.
[17] Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5(6):1017–26.
[18] Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, Blood-Forsythe MA, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat Mater 2016;15(10):1120–7.
[19] Oliynyk AO, Antono E, Sparks TD, Ghadbeigi L, Gaultois MW, Meredig B, et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. Chem Mater 2016;28(20):7324–31.
[20] Chen G, Shen Z, Iyer A, Ghumman UF, Tang S, Bi J, et al. Machine-learning-assisted *de novo* design of organic molecules and polymers: opportunities and challenges. Polymers (Basel) 2020;12(1):163.
[21] Elton DC, Boukouvalas Z, Butrico MS, Fuge MD, Chung PW. Applying machine learning techniques to predict the properties of energetic materials. Sci Rep 2018;8(1):9059.
[22] Kang P, Liu Z, Abou-Rachid H, Guo H. Machine-learning assisted screening of energetic materials. J Phys Chem A 2020;124(26):5341–51.
[23] Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond J-L, et al. Randomized SMILES strings improve the quality of molecular generative models. J Cheminform 2019;11(1). https://doi.org/10.1186/s13321-019-0393-0.
[24] Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. 2017. arXiv:1703.07076.
[25] Solov'eva NP, Makarov VA, Granik VG. Highly polarized enamines. Chem Heterocycl Compd 1997;33(1):78–85.
[26] Tang Y, Ma J, Imler GH, Parrish DA, Shreeve JM. Versatile functionalization of 3,5-diamino-4-nitropyrazole for promising insensitive energetic compounds. Dalton Trans 2019;48(38):14490–6.
[27] Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995;35(6):1039–45.
[28] Hall LH, Story CT. Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural net-works. J Chem Inf Comput Sci 1996;36(5):1004–14.
[29] Landrum G. RDKit: open-source cheminformatics. 2006.
[30] Abdi H, Williams LJ. Principal component analysis. WIREs Comp Stat 2010;2(4):433–59.
[31] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[32] Zhang C, Wang X, Huang H. π-Stacked interactions in explosive crystals: buffers against external mechanical stimuli. J Am Chem Soc 2008;130(26):8359–65.
[33] Zhang J, Mitchell LA, Parrish DA, Shreeve JM. Enforced layer-by-layer stacking of energetic salts towards high-performance insensitive energetic materials. J Am Chem Soc 2015;137(33):10532–5.
[34] Song S, Wang Y, Wang K, Chen F, Zhang Q. Decoding the crystal engineering of graphite-like energetic materials: from theoretical prediction to experimental verification. J Mater Chem A 2020;8(12):5975–85.
[35] Ziletti A, Kumar D, Scheffler M, Ghiringhelli LM. Insightful classification of crystal structures using deep learning. Nat Commun 2018;9(1):2775.
[36] Ryan K, Lengyel J, Shatruk M. Crystal structure prediction via deep learning. J Am Chem Soc 2018;140(32):10158–68.
[37] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[38] Gers FA, Schraudolph NN, Schmidhuber J. Learning precise timing with LSTM recurrent networks. J Mach Learn Res 2002;3:115–43.
[39] Weininger D. a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28(1):31–6.
[40] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in neural information processing systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada. p. 8026–37.
[41] Moret M, Friedrich L, Grisoni F, Merk D, Schneider G. Generative molecular design in low data regimes. Nat Mach Intell 2020;2(3):171–80.
[42] Gani R, Brignole EA. Molecular design of solvents for liquid extraction based on UNIFAC. Fluid Phase Equilib 1983;13:331–40.
[43] Sumita M, Yang X, Ishihara S, Tamura R, Tsuda K. Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies. ACS Cent Sci 2018;4(9):1126–33.
[44] Gao H, Zhang Q, Shreeve JM. Fused heterocycle-based energetic materials (2012–2019). J Mater Chem A 2020;8(8):4193–216.
[45] Chen S, Liu Y, Feng Y, Yang X, Zhang Q. 5,6-Fused bicyclic tetrazolo-pyridazine energetic materials. Chem Commun 2020;56(10):1493–6.
[46] Tsyshevsky R, Smirnov AS, Kuklja MM. Comprehensive end-to-end design of novel high energy density materials: III. Fused heterocyclic energetic compounds. J Phys Chem C 2019;123(14):8688–98.
[47] Schulze MC, Scott BL, Chavez DE. A high density pyrazolo-triazine explosive (PTX). J Mater Chem A 2015;3(35):17963–5.
[48] Yao W, Xue Y, Qian L, Yang H, Cheng G. Combination of 1,2,3-triazole and 1,2,4-triazole frameworks for new high-energy and low-sensitivity compounds. Energ Mater Front 2021;2(2):131–8.
[49] Cao Y, Lai W, Yu T, Ma Y, Liu Y, Wang B. Graphite-like packing modes facilitating high thermal stability: a comparative study in the polymorphs of planar energetic molecules. Cryst Growth Des 2021;21(6):3175–8.
[50] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 09, Revision D.01. Wallingford: Gaussian, Inc.; 2013.
[51] Lu T, Chen F. Multiwfn: a multifunctional wavefunction analyzer. J Comput Chem 2012;33(5):580–92.
[52] Mathieu D. Sensitivity of energetic materials: theoretical relationships to detonation performance and molecular structure. Ind Eng Chem Res 2017;56(29):8191–201.
[53] Song S, Chen F, Wang Y, Wang K, Yan M, Zhang Q. Accelerating the discovery of energetic melt-castable materials by a high-throughput virtual screening and experimental approach. J Matter Chem A 2021;9(38):21723–31.