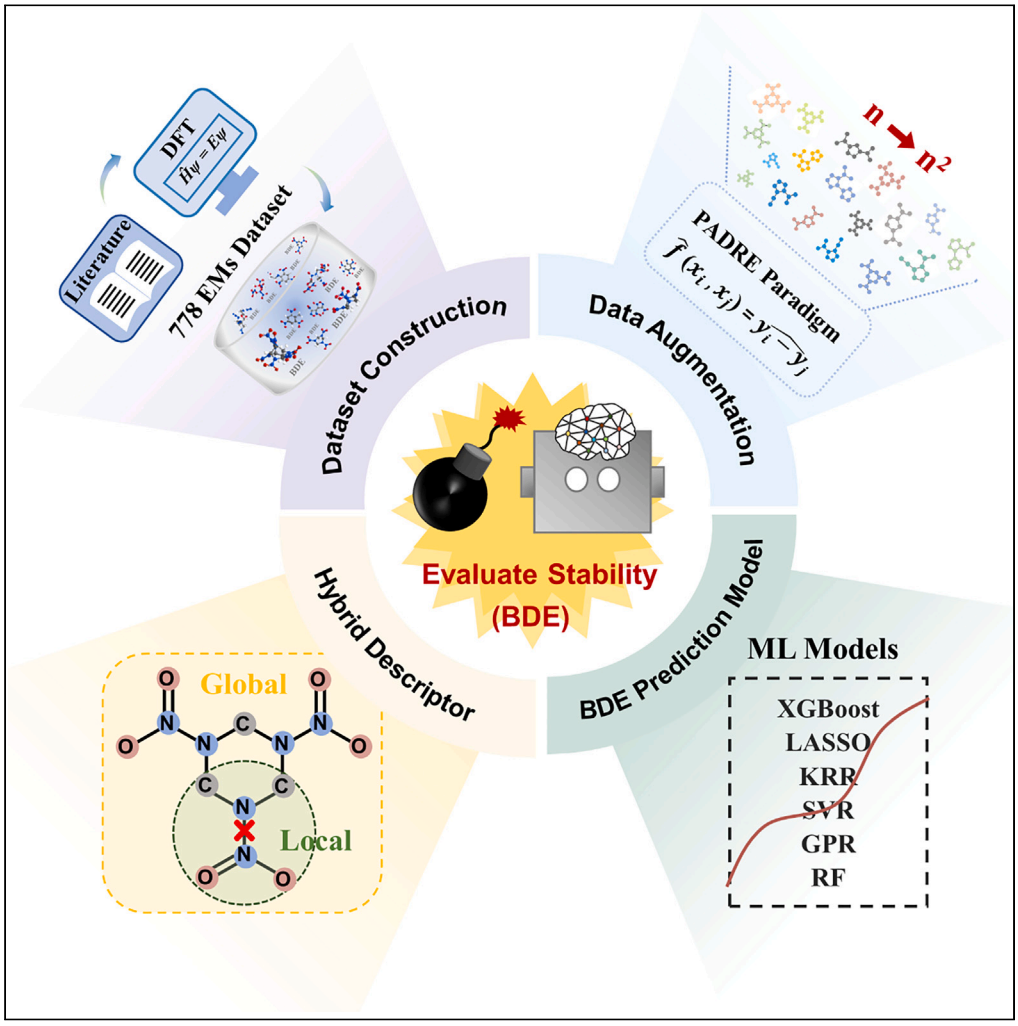
iScience

CellPress
OPEN ACCESS

Article

Exploring an accurate machine learning model to quickly estimate stability of diverse energetic materials



Qiaolin Gou, Jing Liu, Haoming Su, Yanzhi Guo, Jiayi Chen, Xueyan Zhao, Xuemei Pu

xmpuscu@scu.edu.cn

Highlights

A representative energetic dataset is constructed by real and diverse explosives

The hybrid strategy couples local bond into global structure features

Pairwise difference regression augments dataset with reducing systematic errors

A highly accurate XGBoost prediction model is obtained to evaluate EM stability

Gou et al., iScience 27, 109452
April 19, 2024
https://doi.org/10.1016/j.isci.2024.109452

iScience

CellPress
OPEN ACCESS

Article

Exploring an accurate machine learning model to quickly estimate stability of diverse energetic materials

Qiaolin Gou,[1] Jing Liu,[1] Haoming Su,[1] Yanzhi Guo,[1] Jiayi Chen,[1] Xueyan Zhao,[2] and Xuemei Pu[1,3,*]

SUMMARY

High energy and low sensitivity have been the focus of developing new energetic materials (EMs). However, there has been a lack of a quick and accurate method for evaluating the stability of diverse EMs. Here, we develop a machine learning prediction model with high accuracy for bond dissociation energy (BDE) of EMs. A reliable and representative BDE dataset of EMs is constructed by collecting 778 experimental energetic compounds and quantum mechanics calculation. To sufficiently characterize the BDE of EMs, a hybrid feature representation is proposed by coupling the local target bond into the global structure characteristics. To alleviate the limitation of the low dataset, pairwise difference regression is utilized as a data augmentation with the advantage of reducing systematic errors and improving diversity. Benefiting from these improvements, the XGBoost model achieves the best prediction accuracy with R^2 of 0.98 and MAE of 8.8 kJ mol^{-1}, significantly outperforming competitive models.

INTRODUCTION

Energetic materials (EMs) are a special group of metastable substances, including explosives, pyrotechnics, and propellants, which can release large amounts of chemical energy under certain external stimuli and play important roles in military, aerospace, and civilian fields.[1,2] Besides the high energy requirement to EMs, stability is also one of the most concerned properties, as it involves safety during the process of production, storage, and transportation.[3] The sensitivity of EMs is used as a measure of their stability, yet its measurement is closely dependent on complex experimental conditions, thus generally exhibiting large uncertainty and low reproducibility,[4,5] in turn leading to difficulty in analyzing and evaluating the stability of EMs. For the energetic molecules, it was revealed that the sensitivities are strongly correlated with the dissociation energy of the weakest bond X-NO_2 (X = C, N, O),[6–8] thus bond dissociation energy (BDE) can act as an alternative indicator to evaluate the stability of EMs. In general, the energetic molecule is considered as stable if its BDE is higher than 80 kJ mol^{-1} while BDE is required to be higher than 120 kJ mol^{-1} in practical application.[9]

Although BDEs can be measured experimentally by photoelectron spectroscopy,[10] mass spectrometry,[11] and other methods, the experimental measurement is also a complex and time-consuming task. Up to date, the number of BDEs measured by experiments is only one ten-thousandth of the registered molecules, in which most molecules have heavy atoms below ten,[12,13] and their BDEs mainly involve R-H bonds, rather than the trigger bonds X-NO_2 associated with EMs. Fortunately, it has been confirmed that BDEs calculated by quantum mechanics (QM) methods almost achieve comparable accuracy to experiments,[14,15] providing an alternative way to obtain BDEs. Despite the high accuracy of QM, the high requirement to computational cost limits it to explore a vast unknown chemical space. Therefore, there remains an urgent need for developing quick and accurate methods to estimate BDEs such that can quickly evaluate the stability of new EMs, in turn advancing the development of novel EMs with high comprehensive performance. Data-driven machine learning (ML) stands out, as it can mine the structure-property relationship underlying complex data such that it can intelligently navigate. MLs have exhibited great success in chemistry, material, and medicine fields, involving property prediction,[16] reaction mechanism,[17] and molecule generation,[18] etc. As is known, ML is a data-driven technique. Unfortunately, compared to other fields like images and drugs, EMs have been in a state of data scarcity for a long time due to the long-term experimental and high risk, which limit the application of ML. However, some exploits also introduced ML to construct prediction models mainly focusing on high energy properties of EMs like density,[19] formation enthalpy,[20] detonation velocity,[21] and detonation pressure.[22] Unfortunately, the ML-based prediction models for the stability of energetic molecules have been lacked. Wang et al.[23] developed some ML models to predict several EM properties including BDEs, based on 4679 derivatives of a single benzene ring and molecular descriptors, in which the best performance are R^2 of 0.775 for BDEs of C-NO_2 bonds and 0.276 for BDE_S of N-NO_2 bonds, thus remaining large space to be

[1]College of Chemistry, Sichuan University, Chengdu 610064, China
[2]Institute of Chemical Materials, China Academy of Engineering Physics, Mianyang 621900, China
[3]Lead contact
*Correspondence: xmpuscu@scu.edu.cn
https://doi.org/10.1016/j.isci.2024.109452

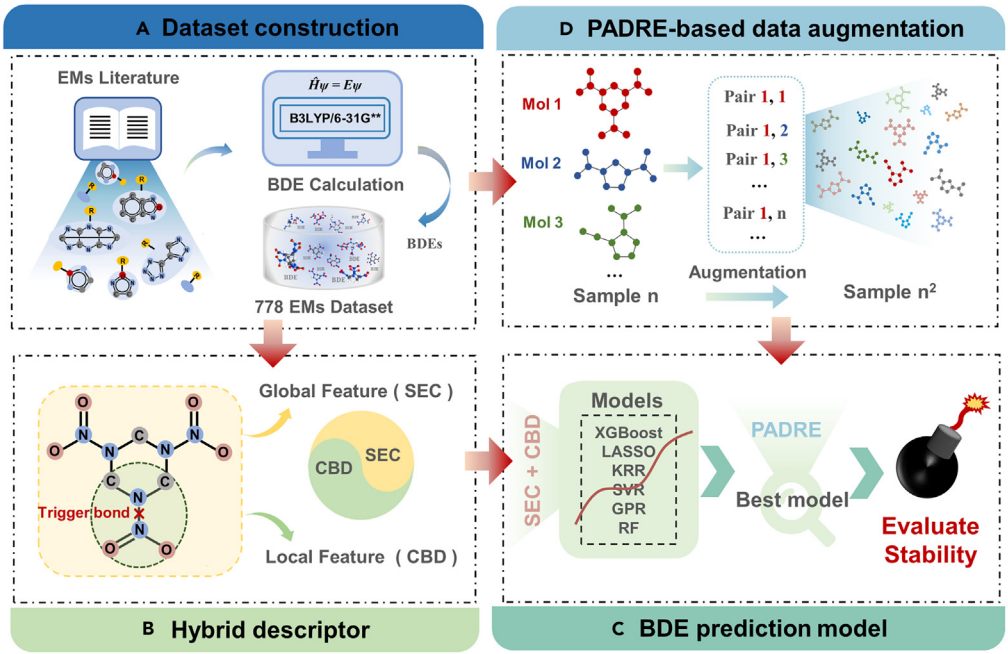


Figure 1. The workflow of constructing the BDE prediction model of EMs

The workflow mainly involves four modules: (A) EM dataset construction; (B) Construction of hybrid descriptor by combining SEC and CBD. The global structure descriptor (SEC) consists of Sum Over Bonds (SOB), Electrotopological state Fingerprints (E-state), and Custom Descriptor Set (CDS). The local structure feature (CBD) is Chemical bond descriptor; (C) Prediction model determination; (D) Pairwise difference regression (PADRE) data augmentation.

improved. In addition, the dataset that only involved benzene derivatives would limit the extrapolation ability of the model to diverse EMs in practical applications.

Despite the absence of ML-based BDE prediction models for EMs, there are several works that developed machine learning models for BDEs of non-energetic systems.[24–26] These works exhibited high accuracy in predicting BDE for the non-energetic systems, largely benefiting from the large size data available. However, these models from non-energetic systems perform poorly when applied to predict the BDEs of EMs (see Results section for more details). The reason should be attributed to the absence of energetic molecules in these datasets and their feature representation without focusing on the structure characteristics of energetic molecules, thus resulting in high accuracy on small organic molecules but low accuracy in predicting EMs. Overall, there is still an unmet need to develop a general prediction model with high accuracy for BDEs of diverse EMs such that can provide a fast and reliable prediction tool for facilitating the design of new EMs with high performance.

Inspired by the urgent need, we developed an accurate machine learning model for BDEs of the trigger bond of EMs based on three key factors of machine learning (data, feature, and model), as illustrated by Figure 1. As accepted, data is the first key component for the performance of ML models due to the data-driven nature of ML. Our main objective is to develop an ML-based BDE model with high accuracy for diverse energetic molecules associated with explosives, pyrotechnics, and propellants. The energetic molecules are significantly different from common organic molecules in structures, for example, energetic skeleton and energetic substituents with more -NO_2 and high N contents, as shown in Figure 1. Theoretically, the model trained on common organic molecules hardly learns sufficient energetic structure characteristics, which would lead to poor performance in predicting the BDE of EMs. Unfortunately, the BDE dataset for real energetic molecules has been lacking. To address this limitation, we construct a new dataset of energetic molecules by collecting 778 synthesized explosives containing C, H, O, and N elements from over 400 literature and calculate their BDEs by QM method (vide Figure 1A). Consequently, a reliable and diverse EM dataset with BDE labels is constructed. Despite constructing a new EM dataset with 778 samples, its size is still limited. For the small size dataset, the feature representation of samples is particularly important for the performance of ML. In order to sufficiently capture the structural features associated with BDE of EMs, we propose a hybrid descriptor strategy. This strategy couples the local feature of the target bond with the global structure features reflecting the inherent energy characteristics of the sample, based on the nature of BDE and the energetic characteristics of the sample, as illustrated by Figure 1B. With the hybrid feature representation, we compared six machine learning models (vide Figure 1C) to determine an appropriate model architecture such that can sufficiently learn the relationship between the structure and the property. After that, we further introduce a novel data augmentation strategy inferred from pairwise difference regression (PADRE) to increase the data size and diversity such that enhances the model's robustness, as illustrated by Figure 1D. The PADRE-based augmentation strategy generates new samples by introducing differences between pairs of feature vectors and labels, which can help to reduce systematic errors such that are beneficial to improve prediction accuracy.

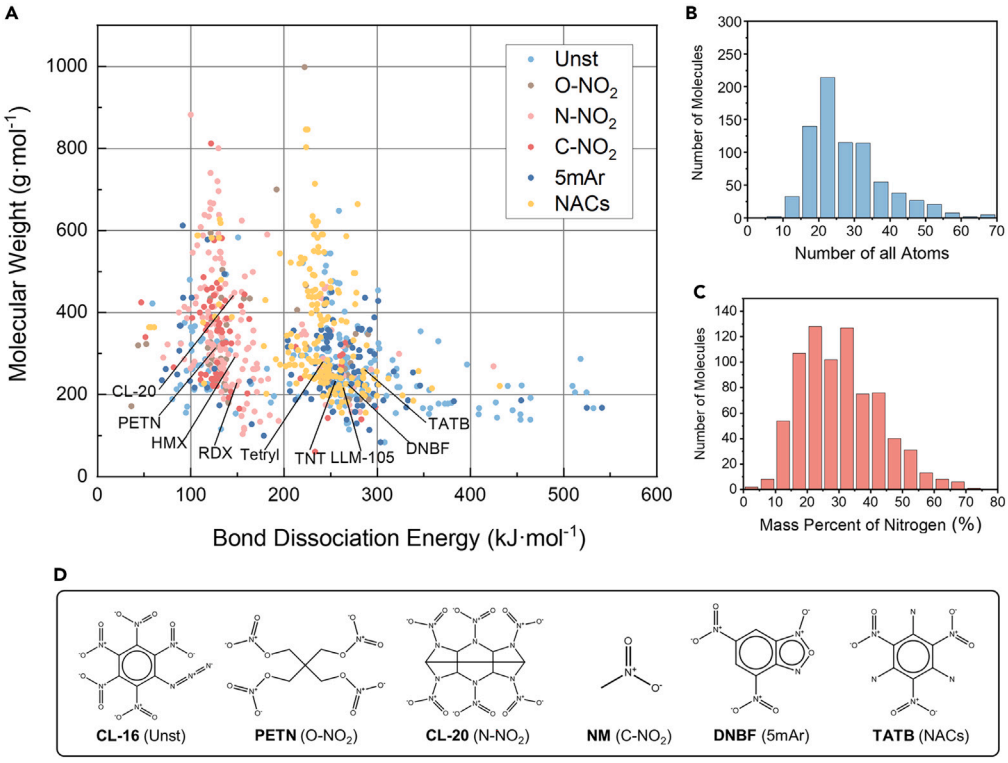


Figure 2. Statistical analyses of the EM dataset constructed

(A) The distribution of BDE and molecular weight. The 778 energetic compounds are classified into six categories following Mathieu's work,[27] which are highlighted by different colors. Unst denotes a generic category including unstable explosophores like azides, difluoroamines, diazo phenols, tetrazoles, and triazole compounds. O-NO_2, N-NO_2, and C-NO_2 represent categories containing O-NO_2 group, N-NO_2 group, and aliphatic C-NO_2 groups, respectively. 5mAr denotes a category containing five-membered aromatics rings. NACs denotes all remaining compounds only containing a single kind of explosophore like nitro groups bonded to carbon atoms in aromatic 6-membered ring. The statistical distribution of (B) atom number, and (C) mass percent of nitrogen for the 778 energetic molecules. (D) Chemical structures of representative molecules from the six categories of the dataset.

RESULTS

As illustrated by Figure 1, our computation framework mainly includes four modules: (1) a real and diverse EM dataset construction; (2) feature representation based on a hybrid strategy; (3) Comparison of different machine learning algorithms to determine the BDE prediction model; (4) PADRE-based data augmentation to further improve the ML performance. Finally, we also compare our model with several competitive BDE models in order to evaluate the advantage of our model in predicting diverse EMs.

Construction of a representative and diverse energetic dataset

Data scarcity is a primary challenge for the application of ML in EMs. Besides the data size, the data representativeness and quality are crucial for the ML model's performance. Thus, we hope to construct a reliable and representative energetic molecule dataset. To this end, we manually collected 778 synthesized CHON-containing energetic molecules from 427 published literature. With the new and real energetic molecule dataset, we optimized their structures and calculated their weakest bond dissociation energies at the level of B3LYP/6-31G** (see Method details section for calculation details). The 778 energetic molecules, including their SMILES, calculated BDEs, the atom numbering of the trigger bond, and related 427 references are listed in "Data S1".

To evaluate the representativeness of the dataset, we performed statistical analyses on the distribution of bond dissociation energies, the molecular weight, the total number of atoms, the number of C, H, O, and N atoms, and the mass percent of nitrogen. As shown in Figure 2, the molecular weights range from 61.04 g mol^{-1} to 998.48 g mol^{-1} and the number of total atoms in these molecules ranges from 7 to 87, and the mass percent of nitrogen in each molecule is in the range of 4.17%–72.73%. The counts of C, H, O, and N atoms in each molecule are presented in Figure 3. The distribution of BDE values of 778 samples ranges from 50 kJ mol^{-1} to 550 kJ mol^{-1}. Additionally, we analyzed the trigger bond types and their counts. The 778 EMs involve ten types of trigger bonds, including C-NO_2, N-NO_2, O-NO_2, C-N_3, C-NH_2, N-NH_2, C-OH, N-OH, C-C, and C-N. Table 1 shows their counts. 593 energetic molecules contain the trigger bond of C–NO_2 type and their BDEs range from 69.2 kJ mol^{-1} to 436.6 kJ mol^{-1}. BDEs of 82 molecules containing the N–NO_2 type are in the range of 58.9 kJ mol^{-1} to 194.3 kJ mol^{-1}. BDEs of 25 molecules with the trigger bond type of O–NO_2 change from 120.9 kJ mol^{-1} to 140.6 kJ mol^{-1}. In addition, following Mathieu's works,[28,29] we classified the molecules based on different nitro types of the trigger bonds to probe

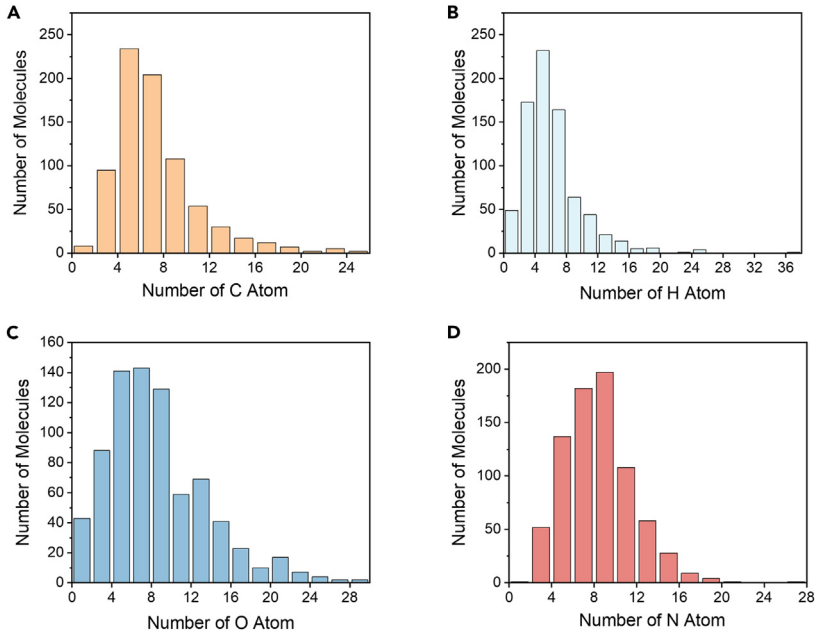


Figure 3. The statistics data of the EM dataset

The statistics numbers of (A) C atoms, (B) H atoms, (C) O atoms, and (D) N atoms contained in the molecules.

whether there is a certain relationship between the BDE values and the nitro types. Figure S1 further shows the chemical structures of representative molecules with different nitro types and their BDE values. It can be seen that BDEs of the different molecules with the same nitro type show great differences. These observations indicate that BDE is not only associated with the trigger bond type but also heavily dependent on the global chemical environment around the trigger bond. As is known, the structures of energetic molecules are diverse and complicated. Assigning a fixed BDE to each nitro type, as done in Mathieu's work, fails to account for the role of the bond environment on the impact sensitivity. It is also why we introduce the ML method to mine the complex relationship between BDEs and the molecular structures.

Additionally, as reflected by Figure 2, the 778 energetic molecules include diverse categories involving different energetic substituents and rings,[27,30] also including classic explosives like TNT, TATB, TNB, RDX and CL-20, rather than only containing -NO_2 group compounds derived from computational combination or collected from Cambridge Structural Database (CSD),[23,31] These statistical results indicate the diversity of EMs for either the structures or the BDE values, thus being representative.

Hybrid feature representation based on the local chemical bond and the global molecule characteristics

As outlined above, the feature representation is a crucial factor for machine learning, especially for the small size dataset. For low data available, the hand-crafted features based on chemical intuition and domain knowledge are generally considered as the best choice. As is known, energetic molecules are generally composed of energetic rings and energetic substituents, thus significantly different from common organic molecules in structures. Energetic rings refer to high-nitrogen heterocycles, mainly including heterocyclic five and six-membered rings like triazole, tetrazole, furazan, pyrazole, oxazole, triazine, and tetrazine. These skeletons can increase the nitrogen content such that can improve the energy content, density, oxygen balance, and detonation performance of EMs.[32,33] Energetic substituents mainly include –NO_2, –NH_2, –ONO_2, –N_3, –$NHNO_2$, and –$C(NO_2)_3$, which are covalently or ionically attached to the energetic rings, where O atoms can oxidize C, N, and H atoms to generate CO/CO_2, NO/NO_2, and OH/H_2O to release chemical energy stored in the explosive.[34] As revealed above, the global chemical environment has a great impact on BDE, besides the trigger bond. The chemical environment involves the global molecular structure, including the energetic skeleton and energetic substituents. Consequently, sufficient feature representation for BDEs of EMs should reflect this structure information. Thus, we propose a hybrid feature strategy by coupling the local feature of the trigger bond into the global molecule representation, as illustrated in Figure 4.

Table 1. The counts of trigger bond types of 778 energetic molecules

Bond Type	C-NO_2	N-NO_2	O-NO_2	C-N_3	C-NH_2	N-NH_2	C-OH	N-OH	C-C	C-N
Counts	593	82	25	26	7	9	1	1	16	18

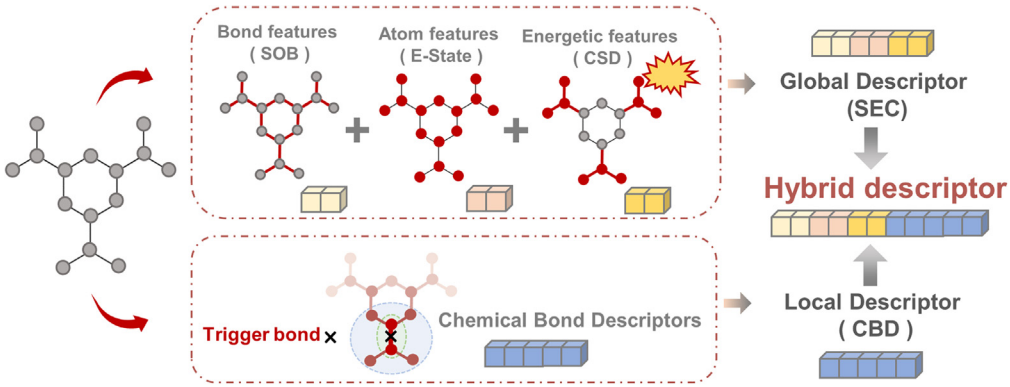


Figure 4. Illustration of the hybrid feature representation

The global descriptor (SEC) consists of Sum Over Bonds (SOB), Electrotopological state Fingerprints (E-state), and Custom Descriptor Set (CDS), which reflect features involving the bonds, atoms, and energetic groups. Chemical bond descriptor (CBD) focuses on the local features involving the target bond and its environment.

To well reflect the global structure features associated with the energetic characteristics, we adopted a fusing descriptor (labeled as SEC) by combining sum over bond (SOB), electrotopological state fingerprint (E-state), and custom descriptor set (CDS), as shown in Figure 4. SOB characterizes different bond types. E-state encodes information about each atom, and CDS contains features of energetic groups. Their combination can characterize the global structure information of energetic molecules from different structural views, thus being found to perform well in predicting some properties of EMs like heat of formation, heat of explosion, detonation velocity, and detonation pressure.[35,36] Herein, we compared the SEC fusing descriptor with the three single type descriptors (SOB, E-state, CDS) and three paired combination descriptors (SOB + E-state, SOB + CDS, and E-state + CDS).

To evaluate their performance on the BDE of EMs, we selected the XGBoost algorithm coupled with different descriptors to predict BDEs of the EM dataset. The reason to use the XGBoost algorithm is due to its built-in regularization and various available hyperparameters that can avoid overfitting and improve its performance. In addition, it was reported that the tree-based ensemble model exhibits good performance in processing small size data with high-dimensional.[37] As shown in Table 2, the SEC descriptor coupled with the XGBoost algorithm shows higher R^2 than the descriptors from the three single types and three paired combinations, showing its advantage in characterizing the global molecule structure. However, the independent test set only exhibits an R^2 of 0.78 for the SEC descriptor, not achieving the expected performance like some other energetic properties reported. The result implies that the SEC descriptor is limited in characterizing the structure feature associated with BDE, which should be attributed to the absence of information reinforcing the trigger bond.

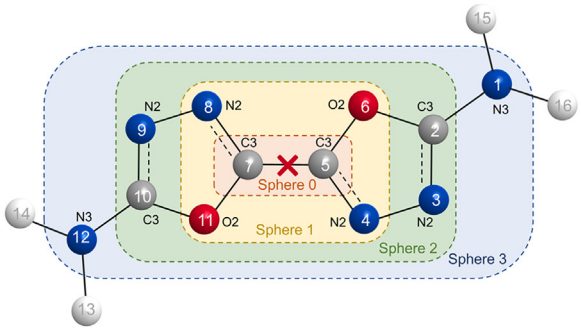
Table 2. Comparison of prediction performance between different descriptors combined with XGBoost algorithm

	Train[a]			Test[a]		
Descriptors	R^2	MAE	RMSE	R^2	MAE	RMSE
SEC[b]	0.93 ± 0.02	15.0 ± 2.1	22.4 ± 3.2	0.78 ± 0.07	25.0 ± 2.6	42.1 ± 5.7
CBD	0.93 ± 0.01	9.1 ± 1.2	20.7 ± 1.9	0.87 ± 0.04	15.6 ± 1.3	30.3 ± 5.3
SEC + CBD	**0.98±0.007**	**7.2±0.9**	**13.1±1.5**	**0.92±0.03**	**14.7±1.2**	**24.3±4.0**
SOB	0.90 ± 0.02	18.2 ± 1.8	26.8 ± 2.4	0.65 ± 0.05	31.2 ± 2.4	48.6 ± 4.6
E-state	0.83 ± 0.03	22.2 ± 2.5	34.6 ± 3.4	0.67 ± 0.07	29.8 ± 2.8	47.7 ± 5.3
CDS	0.89 ± 0.03	19.4 ± 3.2	27.2 ± 4.2	0.64 ± 0.08	31.2 ± 3.7	49.0 ± 5.8
SOB + E-state	0.90 ± 0.04	17.7 ± 3.1	26.4 ± 4.8	0.67 ± 0.07	28.8 ± 3.0	46.9 ± 6.1
SOB + CDS	0.93 ± 0.01	15.5 ± 1.4	22.3 ± 2.0	0.69 ± 0.08	28.0 ± 3.1	45.7 ± 6.0
E-state + CDS	0.92 ± 0.02	16.3 ± 1.8	24.4 ± 2.8	0.73 ± 0.08	25.3 ± 3.3	42.2 ± 6.6
CMs vec	0.84 ± 0.09	24.2 ± 7.9	32.5 ± 9.9	0.32 ± 0.06	51.6 ± 3.3	69.9 ± 4.3
CMs eigs	0.90 ± 0.05	19.3 ± 4.5	26.1 ± 5.7	0.43 ± 0.07	46.4 ± 3.3	64.0 ± 4.7
BOB	0.96 ± 0.01	11.1 ± 1.2	15.9 ± 1.8	0.69 ± 0.07	29.1 ± 3.2	45.1 ± 5.4
Summed BOB	0.90 ± 0.05	19.3 ± 4.3	26.7 ± 5.8	0.55 ± 0.07	38.1 ± 2.9	54.6 ± 4.2
ACSF	0.32 ± 0.56	69.7 ± 13.9	92.7 ± 17.6	−0.03 ± 0.12	74.5 ± 4.3	102.8 ± 17.2

[a]MAEs and RMSEs are reported in $kJ \cdot mol^{-1}$.

[b]SEC is a combination of SOB, E-state, and CDS descriptors.

A



B

Example of point descriptors for bond 5-7

	C2	C3	C4	H1	N1	N2	N3	N4	O1	O2
Sphere 0	0	2	0	0	0	0	0	0	0	0
Sphere 1	0	0	0	0	0	2	0	0	0	2
Sphere 2	0	2	0	0	0	2	0	0	0	0
Sphere 3	0	0	0	0	0	0	2	0	0	0

Example of pair descriptors for bond 5-7

	C-C	C-N	C-O	C-H	N-N	N-O	N-H	O-O	O-H
Sphere = 1, Distance = 1	0	2	2	0	0	0	0	0	0
Sphere = 2, Distance = 1	0	2	2	0	2	0	0	0	0
Sphere = 3, Distance = 1	0	2	0	0	0	0	0	0	0
Sphere = 1, Distance = 2	0	2	2	0	0	0	0	0	0
Sphere = 2, Distance = 2	0	2	2	0	0	0	0	0	0

Figure 5. Examples of chemical bond descriptors

(A) Numbering of all atoms in the molecule. The gray, blue, red, and white spheres represent C, N, O, and H atoms, respectively, and then labeling each atom according to its element and number of connected atoms, and classifying the sphere based on the distance between the atom and the target bond, which are shown by different colors.

(B) Constructing the point descriptors and pair descriptors.

To address the limitation of SEC, we introduced a chemical bond descriptor scheme proposed by Qu et al.[38] to simultaneously focus on the trigger bond in EMs. The chemical bond descriptor (CBD) scheme encodes a bond based on atom types and atom type pairs counts in the target bond and its neighborhood by a sphere split, which can characterize the bond and its local environment. Specifically, after determining the target bond (the red 'cross' in Figure 5A), we number all atoms in the molecule (the white digit on the sphere in Figure 5A). Then, we label each atom according to its element type and the number of connected atoms. For example, atom 5 in Figure 5A denotes the carbon element connecting to atoms 4, 6, and 7, so it is defined as C3. There are ten atom types in the molecule: C2, C3, C4, H1, N1, N2, N3, N4, O1, and O2. After defining the atom types, different spheres are classified in terms of the distance between the atom and the target bond. Atoms 5 and 7 constituting the target bond belong to sphere 0, and atoms 4, 6, 8, and 11 directly connecting to the atoms of the target bond belong to sphere 1, and so on. In the chemical bond descriptor, only up to 4 spheres (from 0 to 3) environment are taken into account for each molecule to achieve a balance between the descriptor length and the model accuracy. After defining the atom type and sphere, three types of descriptors are calculated: point descriptors, pair descriptors, and fragment point descriptors (See Method details section for more details). Figure 5B shows examples of point descriptors and pair descriptors calculated. Finally, the feature with 98-dimension is obtained for each sample.

We also tested XGBoost only coupled with the chemical bond descriptor (CBD) to predict BDE, as shown in Table 2. It can be seen that the independent test set achieves an R^2 of 0.87, showcasing the role of CBD in characterizing the structure feature associated with BDE. However, the chemical bond descriptor mainly focuses on the structural features of the local environment of the broken bond, which is limited in reflecting the global structure of the compound and its energetic characteristics. Thus, we combine CBD and SEC to comprehensively characterize the structural features associated with BDE of the energetic molecules. As expected, the R^2 value is improved to be 0.92 by the hybrid feature (SEC+CBD) on the independent test set (vide Table 2), significantly higher than those from individual SEC and CBD, confirming the effectiveness of the hybrid descriptor in characterizing the structural feature for BDE of EMs.

To more sufficiently evaluate the advantage of the new hybrid descriptor proposed by us, we also compared it with some classic molecular descriptors, such as the Coulomb matrices (including CMs vec and CMs eigs), Bag of Bonds (including BOB and Summed BOB), and atom-centered symmetry function (ACSF). These descriptors are widely used and have achieved good predictive performance in the field of materials science and chemistry.[39,40] The Coulomb matrix characterizes the atomic energies, interatomic distance, and nuclear charges. The bag of bonds reflects a collective energy expression based on different types of bonds. The atom-centered symmetry function characterizes molecular structure based on the local environment near each atom. More introductions about these descriptors can be seen in Method details section. The comparison results are also listed in Table 2. It can be seen that the performance of the five descriptors is significantly different. BOB exhibits better performance than the Coulomb matrix and ACSF. It may be due to the reason that the Coulomb matrix and ACSF do not involve the bond representation while BOB involves different bond types of the molecule, thus including more correlated information with BDEs. In particular, the ACSF descriptor exhibits the worst performance (R^2_{test} = -0.03), which is also attributed to its BDE-independent high dimensionality representation and the small size of our energetic dataset. Summed BOB (R^2_{test} = 0.55) presents lower performance than traditional BOB (R^2_{test} = 0.69), which should be due to the loss in the structure information of different bonds resulted from the summation operation. The results further indicate that the feature representation should sufficiently consider the nature of the target property and the data size, rather than simply adopt existing descriptors.

Comparison of different ML algorithms

To determine an appropriate machine learning to capture the causality between the structure and BDE, we tested several ML algorithms. Given the limited sample size that is suitable for traditional MLs, we preferred to consider XGBoost and five other traditional MLs, such as LASSO, KRR, SVR, GPR, and RF, which exhibited good performance in learning some structure-property relationships for the small size

Table 3. Comparison of prediction performance between six regression algorithms coupled with the hybrid descriptor (SEC+CBD)

	Train[a]			Test[a]		
Models	R^2	MAE	RMSE	R^2	MAE	RMSE
LASSO	0.89 ± 0.01	16.1 ± 1.0	27.5 ± 1.8	0.87 ± 0.05	18.1 ± 1.7	30.1 ± 5.7
KRR	0.95 ± 0.03	11.4 ± 2.0	19.1 ± 5.1	0.89 ± 0.03	17.8 ± 1.2	30.0 ± 3.9
SVR	0.88 ± 0.01	16.1 ± 1.2	29.5 ± 1.6	0.85 ± 0.06	18.3 ± 2.4	32.1 ± 6.9
GPR	0.96 ± 0.03	9.7 ± 4.3	16.1 ± 7.2	0.87 ± 0.03	18.5 ± 1.2	30.7 ± 3.5
RF	0.99 ± 0.001	5.7 ± 0.2	9.8 ± 0.5	0.92 ± 0.02	15.6 ± 1.4	25.6 ± 3.1
XGBoost	0.98 ± 0.007	7.2 ± 0.9	13.1 ± 1.5	0.92 ± 0.03	14.7 ± 1.2	24.3 ± 4.0

[a]MAEs and RMSEs are reported in kJ·mol^{-1}.

datasets.[41–43] Table 3 depicts the performance of the six models coupled with the hybrid descriptor (SEC+CBD) for the training and test sets. It can be seen that the two ensemble models RF and XGBoost exhibit the highest R^2 value of 0.92. RF adopts the Bagging algorithm, which can improve the model performance by integrating models trained on different subsets of the same dataset, thus being widely used in classification and regression problems. XGBoost is a Boosting algorithm which adds a regularization term to the loss function based on Gradient Boosting Decision Trees (GBDT) to avoid overfitting, thus can enhance the generalization ability of the model and effectively handling sparse or missing data. Our result further supports that the ensemble learning algorithms like RF and XGBoost, which combine several weak learners into a strong learner, perform better than the single learner-based ML models like LASSO, KRR, SVR, and GPR. However, compared to RF, XGBoost presents lower MAE and RMSE as well as remarkably quick speed, which can accelerate the training process by up to over 20 times at the same accuracy. Thus, it is determined that XGBoost is more appropriate to serve as the regression model for mining the relationship between the structure and BDEs of EMs than RF.

Further improvement in the predictive performance by PADRE-based data augmentation

Despite of good performance obtained by XGBoost coupled with the hybrid descriptor (SEC+CBD), the dataset of 778 unique energetic samples is still a small data size while machine learning is a data-driven technique. To alleviate the limitation of the small size dataset, data augmentation strategies have been introduced in the ML application, which can generate new samples from available data via different ways, for example, SMILES enumeration,[30] oversampling,[44] and molecular graph augmentation.[45] Here, we introduced the pairwise difference regression (PADRE) paradigm proposed by Tynes et al.[46] as a data augmentation way. PADRE is a machine learning meta-algorithm that utilizes chemical intuition by taking advantage of differences between chemical conditions rather than their absolute values to generate more reliable results. The core concept of PADRE is to use the features from two samples simultaneously and predict the difference between their regression targets. Tynes's results showed that PADRE can reliably improve the model performance of the RF algorithm across five regression tasks on four metal-ligand complex datasets. Theoretically, the differences generated by comparing two samples can be considered as a new sample with the advantage of reducing systematic errors resulted from any single sample measurement, thus we can use it as a novel data augmentation strategy. As illustrated by Figure 6, for the training set, we paired two samples in the training set and calculated the difference in one-dimensional descriptors between the two samples. Then, we concatenated the difference descriptor with the original descriptor. The input is the concatenated descriptor and the output is the difference of the label. As a result, the training set can be augmented from n to n^2 (here, 622 to 622^2). For the test set (156 samples), we paired each sample in the training set with the unknown sample μ in the test set. Then, we calculated the pairwise feature of each known sample paired with unknown sample μ. The trained model was used to generate n difference predictions. The set of difference predictions was then converted by the addition of known label of the training set such that obtain n predictions, which can be regarded as a distribution. The mean value of the distribution is the prediction value for the unknown sample μ.

With the data augmentation, the 148-dimension features are increased to three times (444-dimension features), and the number of training sets increases from the original 622 to 622,[2] which is equivalent to the number of all possible combinations between samples. The increase in the sample number in the training size is much more than the increase in the descriptor dimension, thus alleviating the problem of data scarcity to some extent and reducing the risk of overfitting. Figure 7 shows a comparison in the predictive performance between with the augmentation and without one for XGBoost. Although the predictive performance on the training set is not improved, the model's performance on the independent test set is remarkably improved. The R^2 value of the test set is boosted by the data augmentation from 0.92 to 0.98, and MAE and RMSE are significantly reduced to 8.8 kJ mol^{-1} and 12.9 kJ mol^{-1}, showcasing the advantage of the data augmentation strategy. In addition, we probe the impact of the PADRE augmentation on the performance of other ML algorithms under study. For RF, the PADRE augmentation also greatly improves the prediction performance, in which R^2_{test} increases from 0.92 to 0.97, similar to XGBoost. The result further confirms the effectiveness of PADRE-based data augmentation in improving the ML prediction performance for the small size dataset. However, RF is much slower than XGBoost at the similar accuracy. Thus, XGBoost is still a better choice for the large dataset. While for other machine learning algorithms like LASSO, KRR, SVR, and GPR, they are failed to give results at our computer resource after using PADRE.

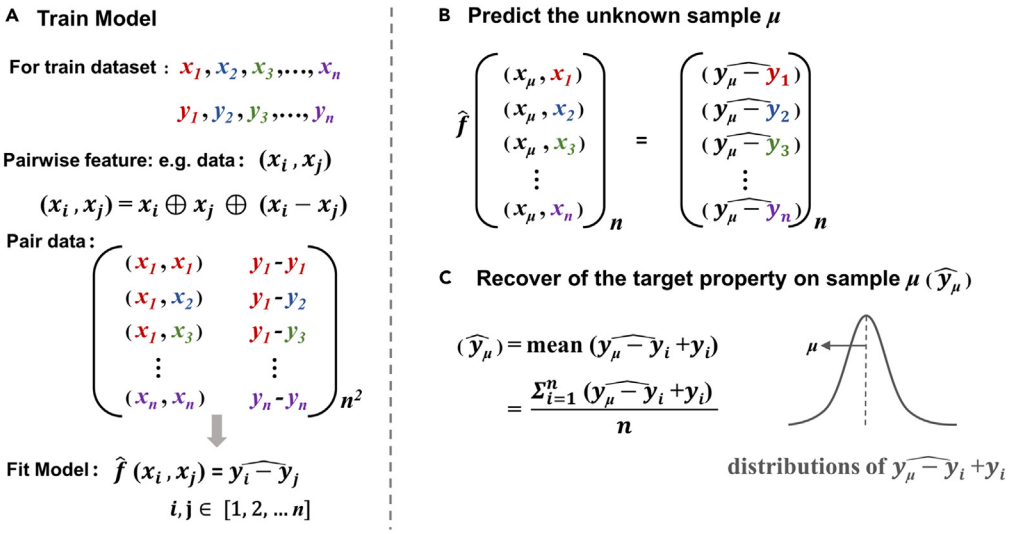


Figure 6. Illustration of PADRE-based data augmentation

(A) Process of model training. A pairwise training set is constructed and the model ($\hat{f}$) is trained on it to predict the label differences ($y_i - y_j$) by the pairwise feature s (x_i, x_j).

(B) Process of pairing an unknown feature vector (x_μ) with all the feature vectors (x_i) in the training set and giving a set of difference predictions ($y_\mu \widehat{-} y_i$) by using the model ($\hat{f}$).

(C) Process of predicting $\widehat{y_\mu}$ for the unknown data (x_μ). The set of difference predictions ($y_\mu \widehat{-} y_i$) is then added to the known labels (y_i) to form a distribution. The mean of this distribution (μ) is the target property of the sample μ.

Practically, some previous works[47,48] already indicated that KRR, GPR, and SVM are not suitable for the modeling of big datasets. Upon the PADRE augmentation, the number of samples increases from 622 to 622^2 for our training set.

Comparison with other competitive models

To more sufficiently verify the advantages of our model in predicting BDEs of EMs, we further make a comparison with several BDE ML works reported. Wen et al.[26] proposed a chemically inspired graph neural network model (BonDNet) and mapped the difference between the molecular graph representations of reactants and products to predict BEDs of three different datasets (PubChem, Zinc, and BDNCM), which achieved high accuracy, in particular for PubChem with MAE of 1.97 kJ mol^{-1}. However, our limited data (778 real explosives) is not suitable for deep learning training. In addition, the data size used for BonDNet is very large (290644 BDEs of 249374 CHON-containing molecules), where the PADRE augmentation strategy is not suitable. Consequently, we cannot directly use our dataset or data augmentation to train the deep learning model used in Wen's work. However, the work provided an optimized model trained on PubChem. Thus, we used the optimized BonDNet model to predict our energetic data, which achieves poor performance on our independent test set with R^2, MAE, and RMSE of 0.35, 142.87 kJ mol^{-1} and 186.08 kJ mol^{-1} respectively. Table S1 representatively shows the comparison result for ten energetic molecules with different sizes of our independent test set. It can be seen that our model achieves much better performance than BonDNet in predicting BDEs of the 10 energetic molecules. In particular, with increasing the sizes of energetic molecules, the prediction errors of BonDNet are sharply increased. As is known, the compounds from the PubChem database are all small molecules composed of C, H, O, and N atoms, and contain only ten or fewer heavy atoms while most energetic molecules have more than ten heavy atoms and unique energetic structures. Thus, BonDNet is not suitable for predicting BDE of EMs, despite its high accuracy on the common organic compounds. Practically, Similar losses of accuracy from general organic compounds to energetic ones were previously observed in the prediction of formation enthalpies of EMs.[49] It is also the main reason why we hope to exploit a BDE prediction model with high accuracy for the diverse EM field.

Nakajima et al.[24] used different types of molecular fingerprints and five machine learning methods to predict the BDE values of 716 hypervalent iodine compounds (HVIs). The Elastic Net (EN) model shows the highest accuracy with $R^2 = 0.96$ and MAE = 6.60 kJ mol^{-1}. Although the work did not provide the optimized model, it gave a brief description regarding the main parameters of the Elastic Net model construction. In terms of these parameters, we constructed an Elastic Net architecture and calculated fingerprints used in Nakajima's work, including Morgan, RDK, MACCS, and Avalon fingerprint for the 778 energetic molecules. Grid search was used to determine optimal hyper-parameters of Elastic Net in our dataset in order to compare as far as possible. Figure 8 shows the prediction accuracy of EN combined with the four types of molecular fingerprints for the independent test set. It can be seen that our hybrid descriptor coupled with XGBoost and the PADRE augmentation is significantly superior to the EN model coupled with different molecular fingerprints.

Of course, the objectives of the two ML works above did not focus on the energetic systems. The comparison with them indicates that the ML model inferred from the non-energetic dataset is indeed not suitable for the EM field. To the best of our knowledge, there are only two ML

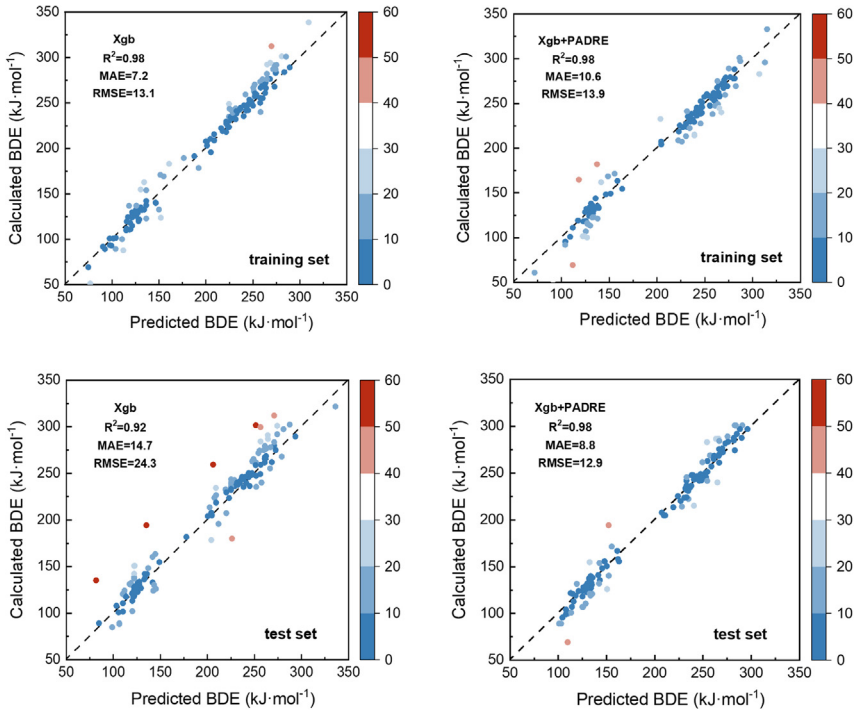


Figure 7. Ablation results of PADRE-based data augmentation

A comparison of the model performance with and without the PADRE augmentation for the training set and the independent test set. Color mapping reflects the Euclidean distance between each sample and the diagonal, which indicates the deviation of the predicted value from its calculated value. The top row displays the results for the training sets while the bottom row shows the results for the independent test sets.

works involving BDEs of EMs. In studying energetic nitrobenzene compounds with MLs, Wang et al.[23] also constructed a BDE prediction model, which used 40 substituents from Cambridge Structural Database to mono- and di-substitute benzene ring to create 4679 derivatives of a single benzene ring as a dataset. The feature descriptors used in their work included elementary percentage, oxygen balance, substituent kind and number, and type of two adjacent substituents. Their best models achieved 0.775 of R^2 for BDE of C-NO_2 (MLP) and 0.276 of R^2 for BDE of N-NO_2 (LASSO). Unfortunately, the work did not provide their ML models. Based on the main parameters of their model described in Wang's work, we constructed LASSO and MLP models for our N-NO_2 and C-NO_2 BDEs, respectively. Also, we conducted grid hyper-parameters optimization based on our dataset. Table 4 shows the comparison results. It is clear that our XGBoost model coupled with the hybrid descriptor and the PADRE data augmentation is significantly superior to the BDE strategy from Wang et al., indicating that the strategy proposed by us is more suitable for predicting BDEs of diverse EMs.

In addition, our previous work regarding the high-throughput design of EMs also involved a BDE ML-based model.[31] In the work, 2000 nitro compounds were selected from the Cambridge Structural Database (CSD) as a representative of EMs, and a graph neural network

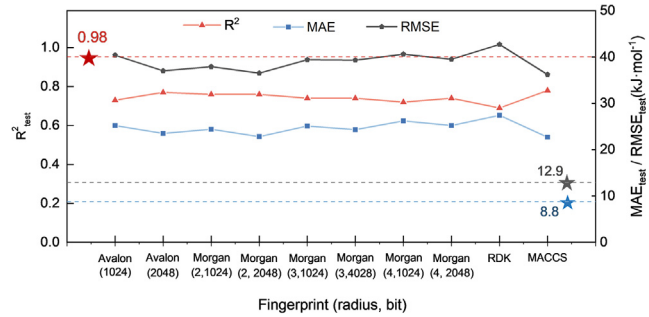


Figure 8. The prediction accuracy of EN combined with different fingerprints

A comparison of the prediction performance between our model and the Elastic Net model coupled with different fingerprints for the independent test set of EMs. The red, blue, and gray points represent the R^2, MAE, and RMSE, respectively. The dashed lines denote our model's results.

Table 4. Comparison of the predictive performance of our model with two competitive models for BDE of energetic materials

		Train			Test		
	Models[a]	R^2	MAE	RMSE	R^2	MAE	RMSE
C-NO_2	Ours	0.99	10.32	7.98	0.96	8.40	12.78
	MLP	0.76	24.26	32.26	0.65	28.67	38.96
N-NO_2	Ours	0.98	5.32	4.08	0.76	9.54	13.07
	LASSO	0.80	9.97	12.66	0.48	17.78	22.21

[a]MLP and LASSO models are derived from ref.[23]; MAEs and RMSEs are reported in kJ·mol^{-1}.

coupled with the molecular graph and the global feature was used to construct the BDE prediction model. The highest R^2 in the 10-fold cross-validation can achieve 0.928. However, given the fact that the real explosives contain diverse energetic skeletons and energetic substituents, rather than only involving -NO_2 groups, the model proposed in our previous work possibly has a risk with high uncertainty when being applied in the diverse EMs. Practically, we calculated the Murcko scaffold[50] similarity and compound similarity between the 2000 nitro compounds[31] and our 778 real energetic compounds, as shown in Figure S2. For the scaffold similarity comparison, 81.3% of scaffolds in the 2000 nitro compounds present similarity lower than 0.1 with the respect to the 778 energetic compounds. For the compound similarity, approximate 90% of the 2000 nitro compounds exhibit low similarity (<0.2) to the 778 energetic compounds. The comparisons indicate that the 2000 nitro compounds used in the previous work[31] indeed cannot represent the real energetic molecules. Not unexpectedly, when we apply the optimized BDE model from ref.[31] to predict the independent test set of the real 778 EMs, it exhibits significantly dropped performance with MAE of 34.42 kJ mol^{-1} and RMSE of 56.36 kJ mol^{-1} for the diverse EMs. The result confirms the necessity of further improving the BDE prediction model of EMs.

DISCUSSION

In the work, we developed an accurate BDE prediction model for diverse EMs mainly by improving the dataset and the feature representation. First, a representative and reliable EM dataset (real 778 energetic molecules) was constructed by extensive literature searching and high-precision DFT calculation, which can provide a reliable data resource for related studies on EMs. To more sufficiently characterize the structure features associated with BDEs of EMs, a hybrid feature representation was proposed by coupling the local feature of the target bond into the global structure representation involving the energetic feature. With the hybrid feature, we tested six ML models to determine the appropriate BDE model for EMs, in which XGBoost exhibits the best performance. In order to alleviate the limitation of the small sample size EM data and further improve the model's robustness to unseen samples, we further utilized the pairwise difference regression (PADRE), which has the advantage of reducing systematic errors from single sample measurement, to remarkably augment the data size and diversity. These technical advantages boost our model to achieve R^2 of 0.98 and MAE of 8.8 kJ mol^{-1} for the independent test set, significantly outperforming other competitive models for EMs. Overall, our observations further highlight the importance of the dataset, which should be closely associated with the target field. Furthermore, the feature representation and the model selection should be based on the chemical nature of the target property and the characteristics of the dataset (size, representativeness, and quality). Thus, this work not only provides an accurate BDE prediction tool for quickly evaluating the stability of diverse EMs, but also offers methodological guidelines (including feature representation and data augmentation) for the application of MLs in the EM field and other low data regimes.

Limitations of the study

Our model achieves high accuracy in predicting BDE of the real explosives under study, benefiting from the hybrid feature proposed and the novel PADRE-based data augmentation strategy. Despite the diversity and the representativeness of the 778 real explosives collected, the number of the explosives collected is still small size due to the experimental difficulty in synthesizing the energetic compounds, which limits the sufficient validation of the model prediction ability to more unseen samples, as ML is a data-driven technique. In the future, besides continuously collecting explosives newly reported, we will also adopt some computation ways to generate new energetic molecules, for example, combination chemistry based on energetic backbones and energetic substituent or deep learning-based molecule generation, which should be beneficial to further improving the robustness of our model.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109452.

ACKNOWLEDGMENTS

This project is supported by Sichuan International Science and Technology Innovation Cooperation Project (Grant No. 24GJHZ0431) and by the National Natural Science Foundation of China (No. 22173065).

AUTHOR CONTRIBUTIONS

Conceptualization, Q.G. and X.P.; Data curation, Q.G. and H.S.; Methodology, Q.G. and J.L.; Writing – original draft, Q.G. and J.L.; Formal analysis, Y.G. and J.C.; Visualization, X.Z.; Supervision, X.P.; Writing – review and editing, X.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 18, 2023
Revised: January 27, 2024
Accepted: March 6, 2024
Published: March 8, 2024

REFERENCES

1. Gao, H., and Shreeve, J.M. (2011). Azole-Based Energetic Salts. Chem. Rev. *111*, 7377–7436. https://doi.org/10.1021/cr200039c.
2. Huang, B., Xue, Z., Fu, X., and Yan, Q.-L. (2021). Advanced crystalline energetic materials modified by coating/intercalation techniques. Chem. Eng. J. *417*, 128044. https://doi.org/10.1016/j.cej.2020.128044.
3. Badgujar, D.M., Talawar, M.B., Asthana, S.N., and Mahulikar, P.P. (2008). Advances in science and technology of modern energetic materials: An overview. J. Hazard Mater. *151*, 289–305. https://doi.org/10.1016/j.jhazmat.2007.10.039.
4. Li, G., and Zhang, C. (2020). Review of the molecular and crystal correlations on sensitivities of energetic materials. J. Hazard Mater. *398*, 122910. https://doi.org/10.1016/j.jhazmat.2020.122910.
5. Coffey, C.S., and De Vost, V.F. (1995). Impact Testing of Explosives and Propellants. Propellants Explo. Pyrotec. *20*, 105–115. https://doi.org/10.1002/prep.19950200302.
6. Politzer, P., and Murray, J.S. (1996). Relationships between dissociation energies and electrostatic potentials of C–NO2 bonds: applications to impact sensitivities. J. Mol. Struct. *376*, 419–424. https://doi.org/10.1016/0022-2860(95)09066-5.
7. Liu, J., He, X., Xiong, Y., Nie, F., and Zhang, C. (2023). Benchmark calculations and error cancelations for bond dissociation enthalpies of X–NO2. Defence Technol. *22*, 144–155. https://doi.org/10.1016/j.dt.2021.11.014.
8. Tan, B., Long, X., Peng, R., Li, H., Jin, B., Chu, S., and Dong, H. (2010). Two important factors influencing shock sensitivity of nitro compounds: Bond dissociation energy of X–NO2 (X=C, N, O) and Mulliken charges of nitro group. J. Hazard Mater. *183*, 908–912. https://doi.org/10.1016/j.jhazmat.2010.07.115.
9. Ma, Q., Jiang, T., Zhang, X., Fan, G., Wang, J., and Huang, J. (2015). Theoretical investigations on 4,4′,5,5′-tetranitro-2,2′-1H,1′H-2,2′-biimidazole derivatives as potential nitrogen-rich high energy materials. J. Phys. Org. Chem. *28*, 31–39. https://doi.org/10.1002/poc.3395.
10. Vogelhuber, K.M., Wren, S.W., Sheps, L., and Lineberger, W.C. (2011). The C–H bond dissociation energy of furan: Photoelectron spectroscopy of the furanide anion. J. Chem. Phys. *134*, 064302. https://doi.org/10.1063/1.3548873.
11. Romanov, V., Verkerk, U.H., Siu, C.K., Hopkinson, A.C., and Siu, K.W.M. (2009). Threshold Collision-induced Dissociation Measurements Using a Ring Ion Guide as the Collision Cell in a Triple-Quadrupole Mass Spectrometer. Anal. Chem. *81*, 6805–6812. https://doi.org/10.1021/ac9009758.
12. Luo, Y.-R. (2002). Handbook of Bond Dissociation Energies in Organic Compounds (CRC press).
13. Luo, Y.-R. (2007). Comprehensive Handbook of Chemical Bond Energies (CRC press).
14. Chan, B., Collins, E., and Raghavachari, K. (2021). Applications of isodesmic-type reactions for computational thermochemistry. WIREs Comput. Mol. Sci. *11*, e1501. https://doi.org/10.1002/wcms.1501.
15. Yao, X.-Q., Hou, X.-J., Jiao, H., Xiang, H.-W., and Li, Y.-W. (2003). Accurate Calculations of Bond Dissociation Enthalpies with Density Functional Methods. J. Phys. Chem. A *107*, 9991–9996. https://doi.org/10.1021/jp0361125.
16. Feng, J., Dong, Z., Ji, Y., and Li, Y. (2023). Accelerating the Discovery of Metastable IrO2 for the Oxygen Evolution Reaction by the Self-Learning-Input Graph Neural Network. JACS Au *3*, 1131–1140. https://doi.org/10.1021/jacsau.2c00709.
17. Burés, J., and Larrosa, I. (2023). Organic reaction mechanism classification using machine learning. Nature *613*, 689–695. https://doi.org/10.1038/s41586-022-05639-4.
18. Xue, D., Gong, Y., Yang, Z., Chuai, G., Qu, S., Shen, A., Yu, J., and Liu, Q. (2019). Advances and challenges in deep generative models for *de novo* molecule generation. WIREs Comput. Mol. Sci. *9*, e1395. https://doi.org/10.1002/wcms.1395.
19. Li, M., Lai, W., Li, R., Zhou, J., Liu, Y., Yu, T., Zhang, T., Tang, H., and Li, H. (2023). Novel Random Forest Ensemble Modeling Strategy Combined with Quantitative Structure–Property Relationship for Density Prediction of Energetic Materials. ACS Omega *8*, 2752–2759. https://doi.org/10.1021/acsomega.2c07436.
20. Chen, C., Liu, D., Deng, S., Zhong, L., Chan, S.H.Y., Li, S., and Hng, H.H. (2021). Accurate machine learning models based on small dataset of energetic materials through spatial matrix featurization methods. J. Energy Chem. *63*, 364–375. https://doi.org/10.1016/j.jechem.2021.08.031.
21. Huang, X., Li, C., Tan, K., Wen, Y., Guo, F., Li, M., Huang, Y., Sun, C.Q., Gozin, M., and Zhang, L. (2021). Applying machine learning to balance performance and stability of high

energy density materials. iScience *24*, 102240. https://doi.org/10.1016/j.isci.2021.102240.

22. Song, S., Chen, F., Wang, Y., Wang, K., Yan, M., and Zhang, Q. (2021). Accelerating the discovery of energetic melt-castable materials by a high-throughput virtual screening and experimental approach. J. Mater. Chem. A Mater. *9*, 21723–21731. https://doi.org/10.1039/D1TA04441A.
23. Wang, R., Liu, J., He, X., Xie, W., and Zhang, C. (2022). Decoding hexanitrobenzene (HNB) and 1,3,5-triamino-2,4,6-trinitrobenzene (TATB) as two distinctive energetic nitrobenzene compounds by machine learning. Phys. Chem. Chem. Phys. *24*, 9875–9884. https://doi.org/10.1039/d2cp00439a.
24. Nakajima, M., and Nemoto, T. (2021). Machine learning enabling prediction of the bond dissociation enthalpy of hypervalent iodine from SMILES. Sci. Rep. *11*, 20207. https://doi.org/10.1038/s41598-021-99369-8.
25. St. John, P.C., Guan, Y., Kim, Y., Kim, S., and Paton, R.S. (2020). Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. Nat. Commun. *11*, 2328. https://doi.org/10.1038/s41467-020-16201-z.
26. Wen, M., Blau, S.M., Spotte-Smith, E.W.C., Dwaraknath, S., and Persson, K.A. (2020). BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. Chem. Sci. *12*, 1858–1868. https://doi.org/10.1039/d0sc05251e.
27. Mathieu, D. (2017). Sensitivity of Energetic Materials: Theoretical Relationships to Detonation Performance and Molecular Structure. Ind. Eng. Chem. Res. *56*, 8191–8201. https://doi.org/10.1021/acs.iecr.7b02021.
28. Mathieu, D. (2013). Toward a Physically Based Quantitative Modeling of Impact Sensitivities. J. Phys. Chem. A *117*, 2253–2259. https://doi.org/10.1021/jp311677s.
29. Mathieu, D., and Alaime, T. (2014). Predicting Impact Sensitivities of Nitro Compounds on the Basis of a Semi-empirical Rate Constant. J. Phys. Chem. A *118*, 9720–9726. https://doi.org/10.1021/jp507057r.
30. Li, C., Wang, C., Sun, M., Zeng, Y., Yuan, Y., Gou, Q., Wang, G., Guo, Y., and Pu, X. (2022). Correlated RNN Framework to Quickly Generate Molecules with Desired Properties for Energetic Materials in the Low Data Regime. J. Chem. Inf. Model. *62*, 4873–4887. https://doi.org/10.1021/acs.jcim.2c00997.
31. Liu, J., Zhao, S., Duan, B., He, X., Yang, C., Pu, X., Zhang, X., Xiao, Y., Nie, F., Qian, W., et al. (2023). High-throughput design of energetic molecules. J. Mater. Chem. A Mater. *11*, 25031–25044. https://doi.org/10.1039/D3TA05002E.
32. Yadav, A.K., Ghule, V.D., and Dharavath, S. (2021). Dianionic nitrogen-rich triazole and tetrazole-based energetic salts: synthesis and detonation performance. Mater. Chem. Front. *5*, 8352–8360. https://doi.org/10.1039/D1QM01365C.
33. Banik, S., Kumar, P., Ghule, V.D., Khanna, S., Allimuthu, D., and Dharavath, S. (2022). Facile synthesis of nitroamino-1,3,4-oxadiazole with azo linkage: a new family of high-performance and biosafe energetic materials. J. Mater. Chem. A Mater. *10*, 22803–22811. https://doi.org/10.1039/D2TA07372B.
34. Zhang, Q., and Shreeve, J.M. (2014). Energetic Ionic Liquids as Explosives and Propellant Fuels: A New Journey of Ionic Liquid Chemistry. Chem. Rev. *114*, 10527–10574. https://doi.org/10.1021/cr500364t.
35. Xie, Y., Liu, Y., Hu, R., Lin, X., Hu, J., and Pu, X. (2021). A property-oriented adaptive design framework for rapid discovery of energetic molecules based on small-scale labeled datasets. RSC Adv. *11*, 25764–25776. https://doi.org/10.1039/D1RA03715C.
36. Elton, D.C., Boukouvalas, Z., Butrico, M.S., Fuge, M.D., and Chung, P.W. (2018). Applying machine learning techniques to predict the properties of energetic materials. Sci. Rep. *8*, 9059. https://doi.org/10.1038/s41598-018-27344-x.
37. Shi, D., Zhou, F., Mu, W., Ling, C., Mu, T., Yu, G., and Li, R. (2022). Deep insights into the viscosity of deep eutectic solvents by an XGBoost-based model plus SHapley Additive exPlanation. Phys. Chem. Chem. Phys. *24*, 26029–26036. https://doi.org/10.1039/D2CP03423A.
38. Qu, X., Latino, D.A., and Aires-de-Sousa, J. (2013). A big data approach to the ultra-fast prediction of DFT-calculated bond energies. J. Cheminform. *5*, 34. https://doi.org/10.1186/1758-2946-5-34.
39. Li, S., Liu, Y., Chen, D., Jiang, Y., Nie, Z., and Pan, F. (2022). Encoding the atomic structure for machine learning in materials science. WIREs Comput. Mol. Sci. *12*, e1558. https://doi.org/10.1002/wcms.1558.
40. Musil, F., Grisafi, A., Bartók, A.P., Ortner, C., Csányi, G., and Ceriotti, M. (2021). Physics-Inspired Structural Representations for Molecules and Materials. Chem. Rev. *121*, 9759–9815. https://doi.org/10.1021/acs.chemrev.1c00021.
41. Gu, G.H., Plechac, P., and Vlachos, D.G. (2018). Thermochemistry of gas-phase and surface species via LASSO-assisted subgraph selection. React. Chem. Eng. *3*, 454–466. https://doi.org/10.1039/C7RE00210F.
42. Dou, B., Zhu, Z., Merkurjev, E., Ke, L., Chen, L., Jiang, J., Zhu, Y., Liu, J., Zhang, B., and Wei, G.-W. (2023). Machine Learning Methods for Small Data Challenges in Molecular Science. Chem. Rev. *123*, 8736–8780. https://doi.org/10.1021/acs.chemrev.3c00189.
43. Xu, P., Ji, X., Li, M., and Lu, W. (2023). Small data machine learning in materials science. npj Comput. Mater. *9*, 42. https://doi.org/10.1038/s41524-023-01000-z.
44. Hemmerich, J., Asilar, E., and Ecker, G.F. (2020). COVER: conformational oversampling as data augmentation for molecules. J. Cheminform. *12*, 18. https://doi.org/10.1186/s13321-020-00420-z.
45. Magar, R., Wang, Y., Lorsung, C., Liang, C., Ramasubramanian, H., Li, P., and Barati Farimani, A. (2022). AugLiChem: data augmentation library of chemical structures for machine learning. Mach. Learn, Sci. Technol. *3*, 045015. https://doi.org/10.1088/2632-2153/ac9c84.
46. Tynes, M., Gao, W., Burrill, D.J., Batista, E.R., Perez, D., Yang, P., and Lubbers, N. (2021). Pairwise Difference Regression: A Machine Learning Meta-algorithm for Improved Prediction and Uncertainty Quantification in Chemical Search. J. Chem. Inf. Model. *61*, 3846–3857. https://doi.org/10.1021/acs.jcim.1c00670.
47. Haghighatlari, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X., and Head-Gordon, T. (2020). Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. Chem *6*, 1527–1542. https://doi.org/10.1016/j.chempr.2020.05.014.
48. Wu, Z., Zhu, M., Kang, Y., Leung, E.L.-H., Lei, T., Shen, C., Jiang, D., Wang, Z., Cao, D., and Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief. Bioinform. *22*, bbaa321. https://doi.org/10.1093/bib/bbaa321.
49. Mathieu, D. (2022). Molecular Energies Derived from Deep Learning: Application to the Prediction of Formation Enthalpies Up to High Energy Compounds. Mol. Inform. *41*, 2100064. https://doi.org/10.1002/minf.202100064.
50. Bemis, G.W., and Murcko, M.A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. *39*, 2887–2893. https://doi.org/10.1021/jm9602928.
51. Frisch, M., Trucks, G., Schlegel, H., Scuseria, G., Robb, M., Cheeseman, J., Scalmani, G., Barone, V., Mennucci, B., and Petersson, G.A. (2009). Gaussian 09, Revision D. 01 (Gaussian, Inc.).
52. Lu, T., and Chen, F. (2012). Multiwfn: A multifunctional wavefunction analyzer. J. Comput. Chem. *33*, 580–592. https://doi.org/10.1002/jcc.22885.
53. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
54. Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilienfeld, O.A., Müller, K.R., and Tkatchenko, A. (2015). Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. J. Phys. Chem. Lett. *6*, 2326–2331. https://doi.org/10.1021/acs.jpclett.5b00831.
55. Hall, L.H., and Kier, L.B. (1995). Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. *35*, 1039–1045. https://doi.org/10.1021/ci00028a014.
56. Politzer, P., and Murray, J.S. (2014). Chapter One - Detonation Performance and Sensitivity: A Quest for Balance. In Advances in Quantum Chemistry, J.R. Sabin, ed. (Academic Press), pp. 1–30. https://doi.org/10.1016/B978-0-12-800345-9.00001-5.
57. Rupp, M., Tkatchenko, A., Müller, K.R., and von Lilienfeld, O.A. (2012). Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Phys. Rev. Lett. *108*, 058301. https://doi.org/10.1103/PhysRevLett.108.058301.
58. Behler, J. (2011). Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J. Chem. Phys. *134*, 074106. https://doi.org/10.1063/1.3553717.
59. Mani-Varnosfaderani, A., Soleimani, M., and Alizadeh, N. (2018). Least absolute shrinkage and selection operator as a multivariate calibration tool for simultaneous determination of diphenylamine and its nitro derivatives in propellants. Propellants Explo. Pyrotec. *43*, 379–389. https://doi.org/10.1002/prep.201700250.
60. Shi, Z., Yang, W., Deng, X., Cai, C., Yan, Y., Liang, H., Liu, Z., and Qiao, Z. (2020). Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. Mol. Syst. Des. Eng. *5*, 725–742. https://doi.org/10.1039/D0ME00005A.

61. Gu, B., Sheng, V.S., Wang, Z., Ho, D., Osman, S., and Li, S. (2015). Incremental learning for ν-support vector regression. Neural Netw. *67*, 140–150.

62. Higgins, K., Valleti, S.M., Ziatdinov, M., Kalinin, S.V., and Ahmadi, M. (2020). Chemical robotics enabled exploration of stability in multicomponent lead halide perovskites via machine learning. ACS Energy Lett. *5*, 3426–3436. https://doi.org/10.1021/acsenergylett.0c01749.

63. Meyer, J.G., Liu, S., Miller, I.J., Coon, J.J., and Gitter, A. (2019). Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests. J. Chem. Inf. Model. *59*, 4438–4449. https://doi.org/10.1021/acs.jcim.9b00236.

64. Zhu, X.-Y., Ran, C.-K., Wen, M., Guo, G.-L., Liu, Y., Liao, L.-L., Li, Y.-Z., Li, M.-L., and Yu, D.-G. (2021). Prediction of Multicomponent Reaction Yields Using Machine Learning. Chin. J. Chem. *39*, 3231–3237. https://doi.org/10.1002/cjoc.202100434.

65. Li, H., Cui, Y., Liu, Y., Li, W., Shi, Y., Fang, C., Li, H., Gao, T., Hu, L., and Lu, Y. (2018). Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells. IEEE Access *6*, 34118–34126.

66. Haffiez, N., Chung, T.H., Zakaria, B.S., Shahidi, M., Mezbahuddin, S., Maal-Bared, R., and Dhar, B.R. (2022). Exploration of machine learning algorithms for predicting the changes in abundance of antibiotic resistance genes in anaerobic digestion. Sci. Total Environ. *839*, 156211. https://doi.org/10.1016/j.scitotenv.2022.156211.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Dataset	This paper	https://doi.org/10.5281/zenodo.10727652
Code	This paper	https://doi.org/10.5281/zenodo.10727652
Software and algorithms		
Gaussian 09	Frisch et al.[51]	https://gaussian.com/
Multiwfn	Lu et al.[52]	http://sobereva.com/multiwfn/
Python package Scikit-learn	Pedregosa et al.[53]	https://scikit-learn.org/
PADRE	Tynes et al.[46]	https://doi.org/10.1021/acs.jcim.1c00670
BonDNet	Wen et al.[26]	https://github.com/mjwen/bondnet
A BDE GNN model	Liu et al.[31]	https://github.com/caepliujian/EM-ML

RESOURCE AVAILABILITY

Lead contact

Further information regarding the methods and the dataset should be directed to and will be fulfilled by the lead contact, Professor Xuemei Pu (xmpuscu@scu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The BDE data has been deposited at Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table.
- The codes of our model have been deposited at Zenodo and is publicly available as of the date of publication. The DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Calculation of bond dissociation energy

The 778 energetic molecules collected were first optimized at the level of B3LYP/ 6–31G**. Then, the NBO analysis was utilized to find the trigger bonds with the smallest values of the Wiberg bond index for each molecule. The bond dissociation energy of the trigger bond at 298 K and 1 atm corresponds to the enthalpy change between the molecule and the two free radicals generated by homolysis,[6] as reflected by Equations 1 and 2.

$$A - B(g) = A\cdot(g) + B\cdot(g) \quad \text{(Equation 1)}$$

$$BDE_{298}(A - B) = H_{298}(A\cdot) + H_{298}(B\cdot) - H_{298}(A - B) \quad \text{(Equation 2)}$$

where $H_{298}(A\cdot)$、 $H_{298}(B\cdot)$ and $H_{298}(A - B)$ refer to the enthalpy at 298 K of the free radicals $A\cdot$, $B\cdot$ and molecule $A - B$, respectively. The calculations of $H_{298}(A\cdot)$, $H_{298}(B\cdot)$ and $H_{298}(A - B)$ were performed at the level of B3LYP/ 6-31G**.

All of the quantum chemical calculations were carried out using the Gaussian 09 program.[51] The vibration frequencies were computed at the same level of theory to confirm that the optimized structures are the minimum on the potential energy surface. The Wiberg bond index of each bond was calculated using Multiwfn software.[52]

Feature descriptors

Feature extraction is one of the most critical steps in constructing ML models, especially for the small size dataset. The following descriptors were involved and compared in the work.

(1) Sum Over Bonds (SOB)

The SOB descriptor includes information about the occurrence of each bond type in a molecule, which is generated by counting the number of each bond type presented in each molecule after enumerating all of the bond types in the dataset.[54] In our energetic materials dataset, there are totally 22 different bond types, including N=N, N=O, C:N, C:O, C:C, C-O, C-N, C-H, C-C, C/C, N:O, N:N, C/O, C/N, N/N, O-O, C=O, C=N, H-N, N-N, N-O, and C=C. In this representation, '-' denotes single bond, '=' denotes double bond, '/' stands for directional bond, and ':' represents the aromatic bond. The count of each bond type is recorded as the molecule's SOB descriptor.

(2) Electrotopological state Fingerprints (E-state)

The Electrotopological State (E-State) fingerprint is an atomic-level molecular descriptor proposed by Hall and Kier in 1995,[55] which describes the intrinsic electronic state of each atom and the perturbations from other atoms. The E-State indices are the sum of the electrotopological indices with the same structural characteristics in the molecular structure. Specifically, the E-state indices of each non-hydrogen atom belonging to the same atom type are summed to obtain the E-state indices of each atom type in the molecule such that reflect the impact of a specific type of atomic structure on the property of matter.

Although the E-state fingerprint defines 79 atom types, only 13 are involved in our dataset. The 13 atom types cover diverse structural environments of C, N, and O atoms in 778 energetic molecules, including ≡C −, = C <, aCa −, aaCa, ≡N, >NH −, [+1], = N −, aNa, >N −, − N ≪, = O, − O −, aOa. (− for single bonds, = for double bonds, ≡ for triple bonds, a for aromatic bonds, ≪ for two double bonds or two resonance single/double bonds as in nitro group, and − > for three single bonds) Therefore, the atom types defined not only involve its atomic information, but also reflect the perturbation from the adjacent atoms. More calculation details can be found in ref.[55]

(3) Custom Descriptor Set (CDS)

High energy density of EMs mainly comes from the energetic groups involving elements like N and O.[56] Given this structural feature, the custom descriptor involves the types of N and O and the elementary composition. Here, the N and O elements are further categorized in terms of how they are incorporated into a molecule. Our dataset involves 7 types of N element and 3 types of O element: C–NO_2, N–NO_2, O–N=O, O–NO_2, C–N=N, C=N–O, C–NH_2, N–O–C, N=O, and C=O. Additionally, this descriptor set includes the number of N, C, and H atoms in the molecule, the C/N ratio, and oxygen balance, resulting in 15-dimension features.

(4) Chemical Bond Descriptor (CBD)

The chemical bond descriptor proposed by Qu et al.[38] can well characterize the target bond and its environment. It can be easily obtained, as it does not rely on quantum chemistry calculations and optimized 3D geometry. The entire process of generating the chemical bond descriptors included defining the atom types and spheres, calculating point descriptors, pair descriptors, and fragment descriptors. The atom type of each atom is defined according to its element and the number of connected atoms. Based on the distance between the atom and one atom of the target bond, the atoms can be classified into different spheres, where the distance is the number of covalent bonds on the shortest possible path. Then three kinds of descriptors are computed at different spheres. The point descriptors are counts of specific atom types in specific sphere. For each sphere, the point descriptor is an array of size M, which is the number of atom types. The pair descriptors represent the counts of atom pairs of specific atom types in specific sphere at specific distance. For each pair, one atom of the pair belongs to the specific sphere and the other atom is in the same or in a lower sphere. The distance is also defined as the number of covalent bonds between the two atoms of the pair on the shortest possible path. The fragment point descriptors describe individual fragments after the target bond is broken, which have two values (each for one fragment) in order to offer the information on the distribution of special functional structures like aromatic systems, or conjugated π systems by restricting the atoms involved in the calculation of the descriptors.

(5) Coulomb matrices (CM)

The Coulomb matrix[57] is a mathematical representation method used to describe molecular structures, which encodes structures based on the Coulombic interactions between each pair of atoms in the molecule. The Coulomb matrix M_{ij} for a given molecule can be calculated by Equation 3. The diagonal elements of the Coulomb matrix correspond to a polynomial fit of the potential energies to the nuclear charge, while the nondiagonal elements correspond to the Coulomb repulsion between different pairs of atoms in the molecule. The Coulomb matrix is invariant to the translation and rotation of the molecule, while it is not invariant under random permutations of atomic indices. To address this issue, the eigenvalues of the Coulomb matrix (CMs eigs) can be used because this representation is invariant with respect to permutations of the rows and columns of the Coulomb matrix. In this approach, the Coulomb matrix is replaced by the eigenvectors of its eigenvalues. However, using eigenvalues means a substantial reduction of the dimensionality and concomitant loss of structural information. Given this fact, here we compared the CMs eigs with the original CMs. Due to its symmetry, the original CMs are converted to a vector (called "CMs vec") by taking the elements from the diagonal and upper triangular part of the matrix in order to avoid redundancy.

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} & i = j, \\ \dfrac{Z_i Z_j}{|R_i - R_j|} & i \neq j \end{cases} \quad \text{(Equation 3)}$$

where the Z and $\boldsymbol{R}$ are the nuclear charge and Cartesian coordinate respectively.

(6) Bag of bonds (BOB)

The BOB[54] was primarily used to capture information about different bonds in a molecule, which is a distinctive variant of the Coulomb matrix and describes collective interactions between many atoms or bonds. In the BoB, the molecular Hamiltonian is first mapped to a vector composed of bags, where each bag represents a particular bond type. Then the Coulomb matrix entries are sorted into bags, and the BoB vector is obtained by concatenating these bags and adding zeros to allow for dealing with other molecules with larger bags. This representation achieves naturally invariant under molecular rotations and translations by fixing the sorting step. Summed bag of bonds (Summed BOB) is a much more efficient representation than BOB, which is the summing of each bag vector. Compared with the BOB, Summed BOB reduces the dimensionality and makes it more suitable for machine learning, but it may cause to some extent loss in the structural information. Therefore, we compared the BOB with the Summed BOB representation.

(7) Atom-centered Symmetry Functions (ACSF)

The ACSF[58] provides the local representations of each atomic environment based on both radial and angular symmetry functions. By the definition, the number of ACSFs is irrespective of the molecule size, and each ACSF is invariant with respect to the exchange of identical neighboring atoms, such that it fully satisfies the permutational invariance. In describing a given central atom, ACSF generates environmental information only up to a cutoff radius, which has the advantage of focusing the information on the chemically meaningful range of interatomic distances. Compared with the Coulomb matrix, in addition to the description of the relative distance between two atoms, the ACSF also adds an angular term to describe the position between atoms, which can more accurately describe the chemical structure of the molecule and the chemical environments of the atoms.

Model training and evaluation

In the work, six traditional machine learning algorithms were considered, which performed well in some property prediction of small size datasets, including a least absolute shrinkage and selection operator regression model (LASSO),[59] a kernel ridge regression model (KRR),[60] a support vector regression model (SVR),[61] a gaussian process regression model (GPR),[62] a random forest regression model (RF),[63] and an extreme gradient boosting regression model (XGBoost).[64]

LASSO is a linear regression algorithm while the other five models are non-linear algorithms. By using L1 regularization, LASSO sets some weights to zero and achieves feature selection, allowing the regression model to be trained on a finite dataset without severe overfitting. The KRR algorithm is a kernelized regression with an L2 regularization term, which reduces overfitting by constraining the weight parameters. It is easier for KRR to obtain coefficients in nonlinear relationships by projecting the feature vectors onto the solution space. SVR maps the input feature into a high-dimensional space, where the optimal hyperplane minimizes the total deviation of all sample points. SVR is often used to solve problems with small sample size and high-dimensional data, with strong generalization ability of the model. GPR is a non-parametric regression model based on Bayesian theory, which combines prior knowledge to make prediction. It uses the gaussian distribution functions to match the observations and shows excellent adaptability in dealing with complex regression problems. RF and XGBoost are Ensemble Learning algorithms that combine several weak learners into a strong learner,[65] mainly including the Bagging and Boosting algorithm. RF belongs to the Bagging algorithm and XGBoost is a Boosting algorithm. The machine learning models mentioned above are implemented with Python scripts by utilizing the open-source scikit-learn package.[53]

Since the data size in the work is limited, traditional K-Fold cross-validation may result in overly optimistic or overly pessimistic model performance in different folds. Therefore, we used shuffle split cross-validation to evaluate ML model performance adequately,[66] For each model, the shuffle split cross-validation was performed by randomly splitting the dataset into a training and independent test set based on an 8:2 ratio by 20 repeats, rather than one in the traditional K-Fold. For each split (also called as iteration), the model hyper-parameters were optimized using the grid search method with nested 5-fold cross-validation. After 20 iterations, the averaged evaluation metrics were used to evaluate the model performance. Mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) values were calculated by Equations 4, 5, and 6, respectively.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{(Equation 4)}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{(Equation 5)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad \text{(Equation 6)}$$

Where N is the number of samples, y_i is the real value, $\hat{y}_i$ is the predicted value, $\overline{y}$ is the mean of the real values. The value of R^2 ranges from negative infinity to 1 and 1 indicates a perfect fitting.

QUANTIFICATION AND STATISTICAL ANALYSIS

Model construction and computations were performed in the Python programming language. The evaluation indicators of the model include MAE, RMSE, and R^2. Details of all statistical analyses can be found above in the Model training and evaluation subsections of the Method details section.