# Machine Learning-Aided High-Throughput Virtual Screening of Novel Fused Nitrogen-Rich Heterocyclic Energetic Compounds

## Supporting Information

Jing Yang[1,2*], Lijun Zhao[1], Luyang Zhang[1], Zhongyang Li[3], Minbei Li[4], Wenhan Yu[5], Luting Xiao[6], Junfang Jia[1], Junxia Guan [1], Yuhao Shi[7*]

[1]Department of Chemistry, Tangshan Normal University, Tangshan 063000, China;

[2]Hebei Key Laboratory of Degradable Polymers;

[3]Horton High School, Wolfville, Nova Scotia B4P 2R2, Canada;

[4]Department of Social Sciences, University of Toronto Mississauga 3359 Mississauga Road, Mississauga, Ontario L5L 1C6, Canada;

[5]Faculty of Science, University of Alberta, Edmonton, AB T6G 2R3, Canada;

[6]Department of Inorganic Non-metallic Materials Engineering, School of Materials Science and Engineering, Hebei University of Technology, Tianjin 300401, China;

[7]Hubei Key Laboratory of Drug Synthesis and Optimization, Jingchu University of Technology, Jingmen 448000, China.

***Corresponding Author:**

[*]Jing Yang E-mail: yjlzddove@163.com

[*]Yuhao Shi E-mail: 202411010@jcut.edu.cn

**LIST OF SUPPORTING INFORMATION**

## 1. Datasets and Molecular Representation

The dataset for model training and validation was curated from the public EM Database v1.0(https://doi.org/10.1016/j.enmf.2023.09.002), containing 4007 unique energetic compounds with experimentally/DFT-validated property data. All molecular structures were standardized using canonical SMILES, and statistical outliers were removed via the interquartile range (IQR) method to ensure dataset quality.

Molecular features were encoded as 1024-bit Morgan fingerprints (radius=2) computed with RDKit (v2023.09), which capture key structural characteristics (e.g., functional groups, bond topology) relevant to energetic material performance and ensure computational efficiency for high-throughput screening.

**Table S1a. Comparison of Performance Metrics for Different Machine Learning Algorithms on the Test Set for Six Key Energetic Properties[a]**

| Target | Model | $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| Density | BayesianRidge | 0.9302 | 0.8361 | 0.0444 |
| Detonation velocity | BayesianRidge | 0.9553 | 0.8834 | 361.5509 |
| Detonation pressure | BayesianRidge | 0.9625 | 0.8999 | 2.0151 |
| Heat of detonation | BayesianRidge | 0.9268 | 0.842 | 445.4382 |
| Detonation volume | BayesianRidge | 0.8774 | 0.6715 | 40.2061 |
| Solid phase enthalpies of formation | BayesianRidge | 0.9079 | 0.817 | 97.3135 |
| Density | KNN | 0.9491 | 0.8138 | 0.0388 |
| Detonation velocity | KNN | 0.9677 | 0.8565 | 329.7352 |
| Detonation pressure | KNN | 0.9711 | 0.8788 | 1.7858 |
| Heat of detonation | KNN | 0.9481 | 0.8033 | 393.7999 |
| Detonation volume | KNN | 0.9156 | 0.6098 | 35.9736 |
| Solid phase enthalpies of formation | KNN | 0.9312 | 0.7988 | 78.7433 |
| Density | SVM | 0.8208 | 0.7661 | 0.0704 |

[a] Note: CV (Cross-Validation) refers to 5-fold cross-validation. MAE stands for Mean Absolute Error, RMSE for Root Mean Squared Error, and r for the Correlation Coefficient. CV Mean represents the mean value from cross-validation, CV Std represents the standard deviation from cross-validation, CV Mean RMSE denotes the mean value of the cross-validated root mean square error, CV Mean $R^2$ denotes the mean value of the cross-validated coefficient of determination ($R^2$), and CV Std $R^2$ denotes the standard deviation of the cross-validated $R^2$.

| | | | | |
|---|---|---|---|---|
| Detonation velocity | SVM | 0.0495 | 0.0244 | 1625.684 |
| Detonation pressure | SVM | 0.8748 | 0.8555 | 2.6101 |
| Heat of detonation | SVM | 0.0338 | 0.0393 | 1516.7929 |
| Detonation volume | SVM | 0.1975 | 0.1936 | 71.9166 |
| Solid phase enthalpies of formation | SVM | 0.1245 | 0.1435 | 257.6866 |
| Density | XGBoost | 0.9033 | 0.8088 | 0.0529 |
| Detonation velocity | XGBoost | 0.9401 | 0.8847 | 399.908 |
| Detonation pressure | XGBoost | 0.9461 | 0.9022 | 2.2252 |
| Heat of detonation | XGBoost | 0.9001 | 0.84 | 496.1766 |
| Detonation volume | XGBoost | 0.8403 | 0.6536 | 45.1243 |
| Solid phase enthalpies of formation | XGBoost | 0.8964 | 0.8209 | 102.0231 |
| Density | DecisionTree | 0.9505 | 0.8004 | 0.0396 |
| Detonation velocity | DecisionTree | 0.9688 | 0.876 | 297.8253 |
| Detonation pressure | DecisionTree | 0.9722 | 0.8755 | 1.7284 |
| Heat of detonation | DecisionTree | 0.9491 | 0.8551 | 334.9243 |
| Detonation volume | DecisionTree | 0.9168 | 0.6564 | 32.429 |
| Solid phase enthalpies of formation | DecisionTree | 0.9337 | 0.8073 | 77.7723 |
| Density | RandomForest | 0.9436 | 0.8464 | 0.0394 |
| Detonation velocity | RandomForest | 0.964 | 0.8946 | 317.7464 |
| Detonation pressure | RandomForest | 0.9682 | 0.9117 | 1.7445 |
| Heat of detonation | RandomForest | 0.9395 | 0.8619 | 394.1393 |
| Detonation volume | RandomForest | 0.8956 | 0.6696 | 37.2095 |
| Solid phase enthalpies of formation | RandomForest | 0.9197 | 0.8339 | 83.7347 |

**Table S1b. Cross-Validation Performance Metrics of Different Machine Learning Algorithms**

| r | CV Mean $R^2$ | CV Std $R^2$ | CV Mean RMSE |
|---|---|---|---|
| 0.0729 | 0.9146 | 0.8585 | 0.0197 |
| 614.618 | 0.9399 | 0.9011 | 0.0162 |
| 3.4932 | 0.9487 | 0.914 | 0.0152 |
| 698.2241 | 0.918 | 0.8344 | 0.0271 |

| | | | |
|---|---|---|---|
| 64.4399 | 0.8204 | 0.7105 | 0.0233 |
| 155.5199 | 0.9044 | 0.793 | 0.0313 |
| 0.0777 | 0.9052 | 0.8313 | 0.0367 |
| 681.7298 | 0.9276 | 0.8769 | 0.0263 |
| 3.8434 | 0.9396 | 0.8913 | 0.0259 |
| 778.8837 | 0.8996 | 0.7974 | 0.037 |
| 70.2235 | 0.789 | 0.6667 | 0.0291 |
| 163.0948 | 0.8955 | 0.7629 | 0.0376 |
| 0.0871 | 0.889 | 0.7821 | 0.0166 |
| 1777.5349 | 0.8435 | 0.043 | 0.0098 |
| 4.1968 | 0.9263 | 0.8501 | 0.0083 |
| 1721.4828 | 0.7538 | 0.0228 | 0.0107 |
| 100.9548 | 0.5816 | 0.1841 | 0.0135 |
| 336.4868 | 0.759 | 0.122 | 0.0122 |
| 0.0787 | 0.8994 | 0.8265 | 0.019 |
| 611.029 | 0.9408 | 0.8881 | 0.0147 |
| 3.4527 | 0.9501 | 0.8982 | 0.0151 |
| 702.4384 | 0.9183 | 0.8154 | 0.0199 |
| 66.1723 | 0.81 | 0.6803 | 0.0267 |
| 153.8565 | 0.9069 | 0.7893 | 0.0305 |
| 0.0804 | 0.8993 | 0.7859 | 0.0237 |
| 633.6406 | 0.9375 | 0.833 | 0.0256 |
| 3.8959 | 0.9374 | 0.8468 | 0.0238 |
| 668.4563 | 0.9259 | 0.7562 | 0.0356 |
| 65.9049 | 0.8172 | 0.5923 | 0.0213 |
| 159.5951 | 0.9006 | 0.7309 | 0.0373 |
| 0.0705 | 0.9204 | 0.8686 | 0.0219 |
| 584.3288 | 0.9463 | 0.9126 | 0.0176 |
| 3.2815 | 0.9552 | 0.9246 | 0.0139 |

| | | | |
|---|---|---|---|
| 652.8097 | 0.93 | 0.854 | 0.0284 |
| 64.6261 | 0.8201 | 0.7065 | 0.0169 |
| 148.1975 | 0.915 | 0.8042 | 0.0395 |

**2. Machine Learning Models and Training Details (SM2)**

**2.1 Feature Descriptor**

All molecular structures were converted to 1024-bit Morgan fingerprints (radius=2) using RDKit. This fingerprinting method was selected for its balance of structural information capture and computational efficiency, which is critical for high-throughput screening of 10,000 molecular derivatives. No additional feature scaling was required for the selected ML algorithms (RandomForest, XGBoost, DecisionTree, KNN).

**Table S2. Property Comparison of Machine Models**

| Target Property | Model | $R^2$ | Pearson r | MAE |
|---|---|---|---|---|
| Density[g/cm$^3$] | Random Forest (RF) | 0.9941 | 0.9971 | 0.0065 |
| Density[g/cm$^3$] | XGBoost | 0.9957 | 0.9978 | 0.0055 |
| Density[g/cm$^3$] | SVR | 0.9479 | 0.9815 | 0.0317 |
| Density[g/cm$^3$] | Decision Tree (DT) | 0.9908 | 0.9954 | 0.0057 |
| Density[g/cm$^3$] | KNN | 0.9787 | 0.9893 | 0.0158 |
| Density[g/cm$^3$] | Ridge Regression | 0.9707 | 0.9853 | 0.0196 |
| Detonation velocity[m/s] | Random Forest (RF) | 0.9999 | 0.9999 | 7.9386 |
| Detonation velocity[m/s] | XGBoost | 0.9999 | 1 | 7.4975 |
| Detonation velocity[m/s] | SVR | 0.4914 | 0.9415 | 1065.6414 |
| Detonation velocity[m/s] | Decision Tree (DT) | 0.9998 | 0.9999 | 7.124 |
| Detonation velocity[m/s] | KNN | 0.9963 | 0.9982 | 55.6748 |
| Detonation velocity[m/s] | Ridge Regression | 0.9915 | 0.9958 | 104.1285 |
| Detonation pressure[GPa] | Random Forest (RF) | 0.9998 | 0.9999 | 0.0436 |
| Detonation pressure[GPa] | XGBoost | 0.9999 | 1 | 0.0382 |
| Detonation pressure[GPa] | SVR | 0.9978 | 0.9989 | 0.1952 |
| Detonation pressure[GPa] | Decision Tree (DT) | 0.9998 | 0.9999 | 0.0406 |
| Detonation pressure[GPa] | KNN | 0.9973 | 0.9986 | 0.277 |

| | | | | |
|---|---|---|---|---|
| Detonation pressure[GPa] | Ridge Regression | 0.9858 | 0.9929 | 1.051 |
| Heat of detonation [kJ/mol] | Random Forest (RF) | 0.9916 | 0.9958 | 53.2841 |
| Heat of detonation [kJ/mol] | XGBoost | 0.9931 | 0.9966 | 44.5065 |
| Heat of detonation [kJ/mol] | SVR | 0.3975 | 0.8503 | 1116.645 |
| Heat of detonation [kJ/mol] | Decision Tree (DT) | 0.9924 | 0.9962 | 45.1386 |
| Heat of detonation [kJ/mol] | KNN | 0.9819 | 0.9912 | 110.7338 |
| Heat of detonation [kJ/mol] | Ridge Regression | 0.9751 | 0.9875 | 163.4007 |
| Detonation volume[L] | Random Forest (RF) | 0.9651 | 0.9827 | 9.3565 |
| Detonation volume[L] | XGBoost | 0.9713 | 0.9855 | 8.0921 |
| Detonation volume[L] | SVR | 0.3787 | 0.7271 | 62.0451 |
| Detonation volume[L] | Decision Tree (DT) | 0.9446 | 0.9725 | 8.153 |
| Detonation volume[L] | KNN | 0.9124 | 0.9556 | 18.1696 |
| Detonation volume[L] | Ridge Regression | 0.8513 | 0.9234 | 28.9248 |
| Solid phase enthalpies of formation[kJ/mol] | Random Forest (RF) | 0.8981 | 0.9486 | 49.2374 |
| Solid phase enthalpies of formation[kJ/mol] | XGBoost | 0.898 | 0.9482 | 45.6588 |
| Solid phase enthalpies of formation[kJ/mol] | SVR | 0.3588 | 0.6398 | 247.8064 |
| Solid phase enthalpies of formation[kJ/mol] | Decision Tree (DT) | 0.8885 | 0.9428 | 39.8381 |
| Solid phase enthalpies of formation[kJ/mol] | KNN | 0.7968 | 0.8927 | 89.4692 |

| | | | | |
|---|---|---|---|---|
| Solid phase enthalpies of formation[kJ/mol] | Ridge Regression | 0.6046 | 0.7781 | 223.4964 |

**2.2 DFT Calculation Methodology and Formulas for High-Throughput Screening**

**1.Geometry Optimization and Frequency Analysis**

○Method: B3LYP/6-311+G(d,p)

○Task: opt freq

○Purpose: Determination of equilibrium geometry, vibrational frequencies, and thermal corrections

**2.High-Precision Single-Point Energy Calculation**

○Method: *M062X* /6-311+G(d,p)

○Task: SP(Single Point)

○Purpose: Calculate accurate electronic energy for thermochemical property determination (consistent with reference molecule data).

**3.Molecular Volume Calculation**

○Method: B3LYP/6-311+G(d,p)

○Task: Volume

○Purpose: Determine molar molecular volume ($V_m$) for crystal density estimation.

All DFT calculations included default convergence criteria and no imaginary frequencies (confirming stable molecular geometries).

**3. Reference Molecules Database**

Thermochemical data for reference molecules (used in isodesmic reaction calculations) were obtained from high-precision M062X /6-311+G(d,p) calculations (consistent with target molecule calculations) and validated against experimental literature values (NIST Chemistry WebBook).

**Table S3. Thermochemical Data of Reference Molecules[b]**

| Molecule | Formula | Molecular Weight (g/mol) | Esp(kJ/mol)) | G298 (kJ/mol) | $\Delta fH^{\circ}_{298}$ (kJ/mol) |
|---|---|---|---|---|---|

[b] Note: $E_{SP}$ denotes electronic single-point energy at the M062X/6-311+G(d,p) level; G298 denotes Gibbs free energy at 298 K including thermal corrections.

| | | | | | |
|---|---|---|---|---|---|
| $H_2$ | $H_2$ | 2.01588 | -3067.64 | -3076.29 | 0.0 |
| $H_2O$ | $H_2O$ | 18.01528 | -200642.85 | -199464.36 | -241.826 |
| $CO_2$ | $CO_2$ | 44.0095 | -495103.04 | -492369.13 | -393.51 |
| $NH_3$ | $NH_3$ | 17.03052 | -148414.06 | -147585.28 | -45.9 |
| $N_2$ | $N_2$ | 28.0134 | -287549.71 | -285931.95 | 0.0 |
| HF | HF | 20.00634 | -263671.44 | -263495.79 | -273.3 |

## 4. Theoretical Framework for Property Calculation (SM4)

### 4.1. Crystal Density Calculation

The crystal density ($\rho$) was calculated from the molecular volume ($V$) obtained from the B3LYP/6-31G(d) volume calculation:

$$\rho = \frac{m}{V} \times 0.6022 \quad (1)$$

○$\rho$ = crystal density (g.cm$^{-3}$)

○$M$= molecular weight (g.mol$^{-3}$)

○$V_m$= molecular volume (cm$^3$.mol$^{-3}$)

○0.6022 = conversion factor $\left(\frac{N_A}{10^{24}}\right)$( with $N_A$ is Avogadro's constant)

### 4.2. Solid-State Heat of Formation Calculation

The solid-state standard enthalpy of formation ($\Delta f H^{\circ}_{298,solid}$) was determined using the isodesmic reaction method:

### 4.3.Isodesmic Reaction Principle

$$\Delta H_f^{\circ}(\text{target, gas, 298K}) = \sum \nu_i \Delta H_{f,i}^{\circ}(\text{reference}_i, 298\text{K}) - \Delta H_{\text{rxn}}(298\text{K})$$

$$\Delta H_{\text{rxn}}(298\text{K}) = \Delta E_{\text{elec}} + \Delta\text{ZPE} + \Delta H_{\text{corr}} \quad (2)$$

○$\nu_i$= stoichiometric coefficient

○$\Delta E_{\text{elec}}$= electronic energy change

○$\Delta$ZPE = zero-point energy correction

○$\Delta H_{\text{corr}}$= thermal enthalpy correction

### 4.4.Gas-to-Solid Correction

$$\Delta H_f^{\circ}(\text{target, solid, 298K}) = \Delta H_f^{\circ}(\text{target, gas, 298K}) - \Delta H_{\text{sub}} \quad (3)$$

where $\Delta Hs_{ub}$ represents the sublimation enthalpy (typically 20-40 kJ/mol for molecular crystals)

## 5. Detonation Performance Parameters and Calculations (SM5)

## 5.1 Kamlet-Jacobs Semi-Empirical Equations

For CHNO-based energetic materials, detonation parameters were calculated using the Kamlet-Jacobs semi-empirical correlations:

## Key Equations

## Detonation Velocity:

$$D = 1.01 \times \left( N \times \bar{M}^{\frac{1}{2}} \times Q^{\frac{1}{2}} \right)^{\frac{1}{2}} \times (1 + 1.30\rho) \tag{4}$$

## Detonation Pressure:

$$P = 0.131 \times \rho^2 \times N \times \bar{M}^{\frac{1}{2}} \times Q^{\frac{1}{2}} \tag{5}$$

## Heat of Detonation:

$$Q_{det} = \frac{\left| \sum \Delta H_f(products) - \Delta H_f(explosive) \right|}{M} \times 1000 \tag{6}$$

## Specific Volume:

$$V_0 = \frac{n_{gas} \times 22.4 \times 1000}{M} \tag{7}$$

○$D$= detonation velocity (km/s)

○$P$= detonation pressure (GPa)

○$Q$= heat of detonation (kJ/mol)

○$N$= moles of gaseous products per gram (mol.g$^{-1}$)

○$\bar{M}$ = average molecular weight of gaseous products (g/mol)

○$\rho$ = crystal density (g.cm$^{-3}$)

○$n_{gas}$= total moles of gaseous products

○$V_0$= specific volume (L.kg$^{-1}$)

## 5.2. Product Atom Conservation

For general formula $C_aH_bN_cO_d$, balance the reaction:

$$C_aH_bN_cO_d \rightarrow xCO_2 + yCO + zH_2O + wN_2 + uO_2 + vH_2 \tag{8}$$

subject to atom conservation:

$$\begin{cases} a = x + y \\ b = 2z + 2v \\ c = 2w \\ d = 2x + y + z + 2u \end{cases} \tag{9}$$

For general formula $C_aH_bN_cO_dF\,e$, balance the reaction:

$$C_aH_bN_cO_dF_e \rightarrow xCO_2 + yCO + zH_2O + wN_2 + uO_2 + vH_2 + rHF \quad (10)$$

subject to atom conservation:

$$\begin{cases} a = x + y \\ b = 2z + 2v + r \\ c = 2w \\ d = 2x + y + z + 2u \\ e = r \end{cases} \quad (11)$$

## 5.3. Energy Unit Conversions

$$1\text{kJ.mol} = 0.239006\text{kcal.mol}^{-1}$$

$$1\text{kcal.mol} = 4.184\text{kJ/mol}$$

$$1\text{eV} = 96.4853\text{kJ/mol}$$

This computational protocol follows established methodologies for energetic materials characterization. Results should be validated against experimental data when available.

**Table S4.a. DFT Calculated vs. ML Predicted Key Detonation Properties (Top 20 Candidates)**

| Candidate | Density [g/cm$^3$] | Detonation velocity [m/s] | Detonation pressure [GPa] |
|---|---|---|---|
| M1 | 1.602 | 7640 | 23.3 |
| M2 | 1.029 | 5910 | 13.4 |
| M3 | 1.881 | 7900 | 28.5 |
| M4 | 1.402 | 6710 | 17 |
| M5 | 1.843 | 8470 | 30.5 |
| M6 | 1.846 | 9050 | 35.3 |
| M7 | 2.072 | 9080 | 44.0 |
| M8 | 1.834 | 8470 | 31.3 |
| M9 | 1.95 | 8910 | 36.8 |
| M10 | 2.451 | 9820 | 51.5 |
| M11 | 1.548 | 7840 | 26.6 |
| M12 | 1.541 | 7140 | 20.8 |
| M13 | 1.955 | 8360 | 32.0 |
| M14 | 1.878 | 9130 | 37.4 |
| M15 | 2.27 | 9450 | 44.7 |
| M16 | 1.506 | 7460 | 22.9 |

| | | | |
|---|---|---|---|
| M17 | 2.194 | 9120 | 39.8 |
| M18 | 2.055 | 9370 | 44.8 |
| M19 | 1.966 | 8400 | 33.5 |
| M20 | 2.014 | 8400 | 33.5 |

**Table S4.b. DFT Calculated vs. ML Predicted Key Detonation Properties (Top 20 Candidates)**

| Candidate | Detonation volume [L/kg] | Heat of detonation [kJ/mol] | Solid phase enthalpies of formation[kJ/mol] |
|---|---|---|---|
| M1 | 803 | 3582 | 250 |
| M2 | 803 | 3582 | 275 |
| M3 | 777 | 3649.4 | 275 |
| M4 | 803 | 3582 | 250 |
| M5 | 1415 | 3779 | 267 |
| M6 | 1054 | 4040 | 321 |
| M7 | 964 | 3774 | 275 |
| M8 | 985 | 4271 | 225 |
| M9 | 771 | 4176 | 350 |
| M10 | 1244 | 3561 | 333 |
| M11 | 757 | 3908 | 166 |
| M12 | 806 | 3387 | 156 |
| M13 | 755 | 3907 | 100 |
| M14 | 1018 | 4153 | 335 |
| M15 | 955 | 4212 | 325 |
| M16 | 1005 | 3610 | 168 |
| M17 | 614 | 4104 | 350 |
| M18 | 647 | 4126 | 380 |
| M19 | 815 | 3730 | 200 |
| M20 | 815 | 3730 | 265 |

**6. Supplementary Figure (SM6)**

**Fig.S1** Machine Learning Model-Predicted Energetic Properties for 10,000 Fused Nitrogen-Rich

Heterocyclic Derivatives

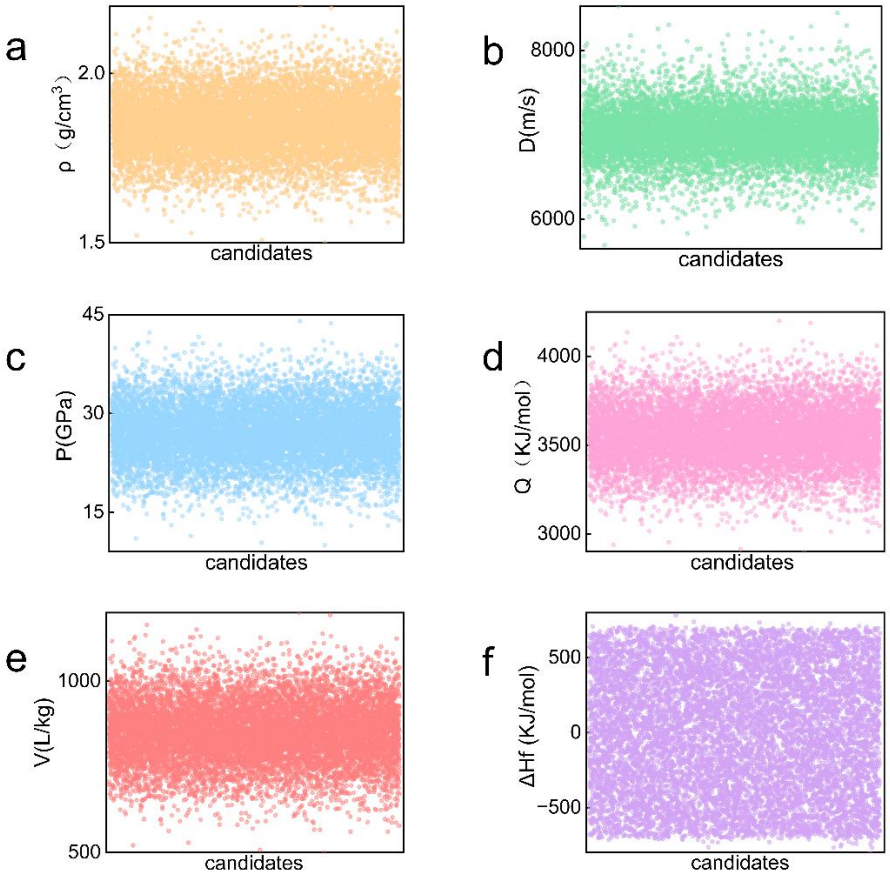


**Fig.S1** The density is denoted by ($\rho$), detonation velocity by ($D$), detonation pressure by ($P$), detonation volume by ($V$), heat of detonation by ($Q$), and solid-state enthalpy of formation by ($\Delta fH_m°$) for the energetic compounds investigated in this work.