Chinese Journal of Chemical Engineering 90 (2026) 261–271

Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

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

Full Length Article

Machine learning for adsorption-related parameters prediction of electronic specialty gases: DFT-based dataset construction and balanced data augmentation

Zhikang Wu [1], Ying Wu [1], Guang Miao [1], Runze Chen [2], Lingjun Ma [2], Hongxia Xi [1], Jing Xiao [1,*]

[1] School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou 510640, China
[2] Peric Special Gases Co., Ltd., Handan 057550, China

ARTICLE INFO

Article history:
Received 15 July 2025
Received in revised form
22 September 2025
Accepted 22 September 2025
Available online 18 October 2025

Keywords:
Molecular property database
Small sample machine learning
Data augmentation
Molecular property prediction
Adsorption

ABSTRACT

Electronic specialty gases play vital roles in key chip manufacturing processes like lithography, etching, deposition and cleaning. While their ultra-high purity (≥99.999%) creates challenging separation requirements, insufficient physicochemical data has hindered adsorbent development. To bridge this gap, we constructed a multidimensional database covering 101 semiconductor-related molecules with 19 physical parameters, and developed a Bayesian regression-based collaborative prediction model demonstrating high accuracy (R^2 = 0.95–0.97) on test sets. We further constructed the balanced data-augmented Transformer-based molecular property prediction (BD-TMPP) model to address the overfitting problem in small-sample learning. This model achieves the end-to-end prediction of molecular quadrupole moment (R^2 = 0.99), and polarizability (R^2 = 0.98) *via* the capture of interatomic spatial correlations. Compared with traditional density functional theory calculations, the model achieves a five-orders-of-magnitude improvement in computational efficiency while maintaining accuracy, demonstrating a successful application of the "structure-property relationship" theory in chemical machine learning.

1. Introduction

The fast reasoning capabilities of artificial intelligence (AI) models rely on high-performance computing chips, the production of which is heavily dependent on electronic specialty gases (ESGs) [1]. These gases play vital roles in key wafer fabrication processes such as etching, and doping. Consequently, an increasing number of researchers have focused on investigating the synthesis, separation, and purification of ESG, particularly fluorinated compounds [2,3]. As semiconductor manufacturing processes require increasingly stringent gas purity standards and drive distillation energy consumption higher, adsorption separation technology is gaining prominence owing to its lower operational energy demands [4–8].

In gas and liquid adsorption processes, molecular physical parameters have a significant impact on adsorption performance. The size of the molecules affects their diffusion rate and adsorption capacity within the pores of the adsorbent [9,10]. Huang *et al.* [11] designed polydopamine-derived carbonaceous adsorbents tailored to the kinetic diameters of N_2/CH_4 to realize the inverse adsorption separation of this system. Gas polarizability modulates adsorption behaviors through its impact on intermolecular interactions [12]. Moreover, the electric field properties of the adsorbent surface, which arise from polar functional groups or delocalized π-electrons, promote the generation of larger transient dipoles in gas molecules, thereby enhancing adsorption performance. Given that C_3H_8 exhibits significantly greater polarizability than CH_3F, it can be selectively adsorbed by glucose-derived carbon molecular sieve (CMS-T) under ultra-low pressure conditions, enabling effective removal of trace C_3H_8 from CH_3F streams. This process achieves CH_3F product gas purity of 7 N (99.99999%) [13]. Furthermore, quadrupole moments in

* Corresponding author.
E-mail address: cejingxiao@scut.edu.cn (J. Xiao).

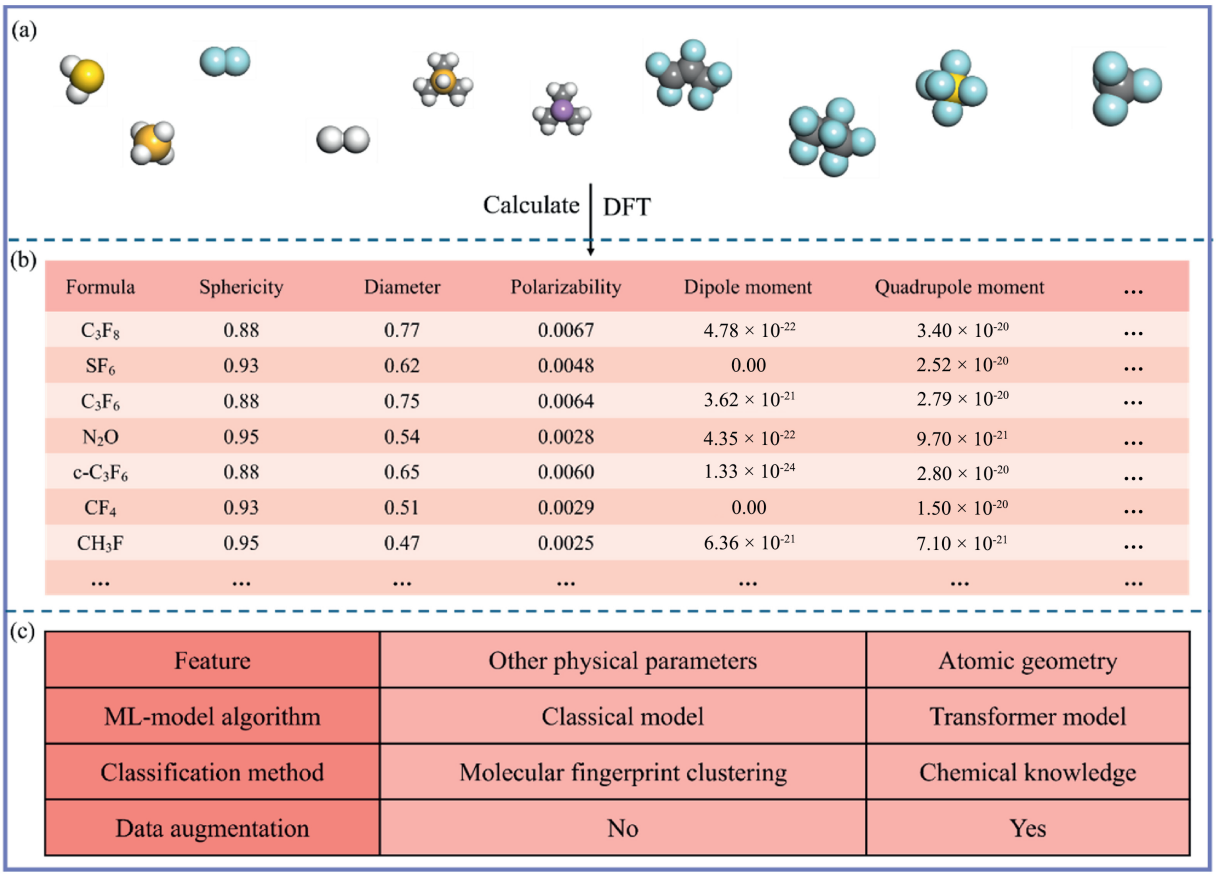
https://doi.org/10.1016/j.cjche.2025.09.012

adsorbates make the molecular surface interactions more complex, affecting adsorption performance *via* strengthened intermolecular forces.

An in-depth study of molecular physical parameters has important theoretical and practical significance for the design and development of adsorbents [14]. However, the adsorption-related property parameters of molecules relevant to chip production are still not fully characterized and a comprehensive database has not yet been established [15–18]. This absence of reliable datasets not only limits the optimization of adsorbents but also hinders technological advancements in ESG separation. The systematic investigation and collection of these molecular properties will fill critical data gaps, provide a database for the development of high-performance adsorbents, and ultimately drive innovation in specialty gas production technologies. Notably, these parameters typically need to be obtained by high-accuracy density functional theory (DFT) calculations, which makes the process computationally intensive and resource-demanding.

In recent years, machine learning (ML) has played an important role [19] in material property prediction [20,21], reaction pathway optimization [22] and intelligent process control [23–25]. In addition, the application of ML in molecular property prediction has become a key tool for accelerating the design of drug molecules [26,27]. Therefore, it is increasingly important to develop highly accurate, strongly generalizable small molecule machine learning (SMML) models of predicting property parameters from molecular structure descriptors [28,29]. Kretschmer *et al.* [30] identified the problem of datasets coverage bias prevalent in SMML and proposed a distance metric based on maximum common edge subgraph (MCES) to assess and improve dataset representativeness, thus providing modeling capabilities. Wan *et al.* [31] constructed a multi-channel pre-training architecture to improve the accuracy of structural representations through a hierarchical integration of molecular topology robustness and generalization capabilities, providing guidance for drug development. However, these ML paradigms require large amounts of high-quality data [32], but it is difficult to accumulate large-scale standardized datasets for experimental chemistry research. Pi *et al.* [33] used datasets constructed from molecular dynamics simulations for ML model training, and the trained models were used to screen high-performance ionic liquids (ILs) for green chemistry processes. This simulation-driven strategy for addressing experimental data scarcity has applicability in diverse chemical topics [22,24,34–37]. Nevertheless, these ML models are still limited by the need for large amounts of data.

Herein, in order to address the lack of physical parameters associated with the aforementioned molecules, we obtained adsorption-related property parameters *via* DFT calculations for the target molecules. SMML models were subsequently trained using this limited dataset. As illustrated in Fig. 1, this study systematically evaluates the predictive capabilities of classical ML models and Transformer-based models. Specifically, the classical ML models employed the remaining physical parameters (excluding target parameters) as features, assisted by molecular fingerprints. Instead, Transformer-based models employed atomic geometric position coding as features, using chemical knowledge



Formula	Sphericity	Diameter	Polarizability	Dipole moment	Quadrupole moment	...
C_3F_8	0.88	0.77	0.0067	4.78×10^{-22}	3.40×10^{-20}	...
SF_6	0.93	0.62	0.0048	0.00	2.52×10^{-20}	...
C_3F_6	0.88	0.75	0.0064	3.62×10^{-21}	2.79×10^{-20}	...
N_2O	0.95	0.54	0.0028	4.35×10^{-22}	9.70×10^{-21}	...
c-C_3F_6	0.88	0.65	0.0060	1.33×10^{-24}	2.80×10^{-20}	...
CF_4	0.93	0.51	0.0029	0.00	1.50×10^{-20}	...
CH_3F	0.95	0.47	0.0025	6.36×10^{-21}	7.10×10^{-21}	...
...	...	...	...	...	...	...

Feature	Other physical parameters	Atomic geometry
ML-model algorithm	Classical model	Transformer model
Classification method	Molecular fingerprint clustering	Chemical knowledge
Data augmentation	No	Yes

Fig. 1. The conceptual framework of this work: (a) some examples of semiconductor-related molecules; (b) example of a multidimensional database of physical property parameters of semiconductor-related molecules; (c) the two types of machine learning models involved in this study.

as a classification criterion. It is worth emphasizing that this work applied data augmentation methodologies commonly used in computer vision [38,39] which remain understudied for SMML.

2. Methods

2.1. Database of molecular property parameters

In this study, 101 molecules related to chip production were systematically calculated and analyzed theoretically. First, the geometrical configurations of all molecules were optimized by DFT calculations using the Gaussian 16 package, with the B3LYP exchange-correlation functional and the def2-svpd basis set [40–43]. The quantum chemical descriptors, including molecular polarizabilities, electric dipole moments, electric quadrupole moments, and highest occupied molecular orbital–lowest unoccupied molecular orbital energy gap (Homo-Lumo), were extracted directly from the output files. Meanwhile, the obtained molecular wave function data were analyzed in depth using Multiwfn software [44,45] to obtain more complex electronic structures and geometrical features. The complete dataset is systematically presented through Tables S1–S3 in the Supplementary Material. Table 1 shows the definitions of 19 parameters in the dataset and their relevance to the adsorption process.

The Spearman rank correlation coefficient (SRCC) analysis method is particularly suitable for revealing nonlinear interrelationships and overcomes the limitations of the traditional Pearson correlation coefficient (PCC), which only reflects linear relationships. And the Pandas library in Python was employed to compute and generate both SRCC and PCC matrices. Subsequently, the corresponding heatmaps were generated through the Seaborn library [48] to enable visual data presentation.

2.2. Molecular fingerprint and Tanimoto similarity

Molecular fingerprinting is an algorithmically generated numerical characterization of molecular structure. In the field of cheminformatics, the Molecular ACCess System (MACCS) fingerprint is widely recognized as a standard characterization method. In this study, the MACCS fingerprint was generated using the RDKit toolkit for all molecules in the database [49,50]. And the Tanimoto similarity coefficient can be used to calculate the proportion of shared chemical substructures between pairs of molecules, with values bounded between 0 and 1, where 0 denotes no shared substructures and 1 indicates complete substructural overlap [51,52]. This metric is mathematically expressed as:

$$T = \frac{|A \cap B|}{|A \cup B|}$$

where the numerator characterizes the number of substructures common to the two molecules A and B, and the denominator is the total number of substructures contained in the two molecules. These substructures were determined by the MACCS fingerprint conversion algorithm.

The cross-category molecular similarity evaluation was performed using the mean value method, which calculates the arithmetic mean of the Tanimoto similarity index between all cross-category molecular pairs.

2.3. Machine learning

In this study, four types of classical ML models, Random Forest, Decision Tree, Gradient Boosting Tree and Bayesian, were

Table 1
The definitions of 19 parameters in the dataset and their relevance to the adsorption process.

	Definitions	Example (C_2H_4)	Adsorption relevance
M/g·mol^{-1}	Molecular mass	28.05	Related to intermolecular interactions and thermodynamic stability [9–11,14].
Boiling temperature/K	Boiling temperature	170.15	
X/nm	Three-dimensional size, X	0.43	Closely related to the accessibility of molecules within the pores and determines the steric hindrance effect in adsorption [9–11].
Y/nm	Three-dimensional size, Y	0.49	
Z/nm	Three-dimensional size, Z	0.34	
Shape factor	Min (X, Y, Z)/Max (X, Y, Z)	0.70	
Sphericity	The similarity between the molecular shape and the spherical	0.96	
Diameter/nm	Diameter of the circumscribed sphere	0.55	
Polarizability/nm^3	The degree a particle's charge distribution distorts in an electric field	0.004	Influences the distribution of charges on the surface of the adsorbent and the polar sites [13].
Dipole moment/C·nm	Measure of separation of positive and negative charges in a molecular	0.00	
Quadrupole moment/C·nm^2	Measure of deviation of charge distribution from spherical symmetry	7.86×10^{-21}	
Homo-Lumo/eV	Energy gap between highest occupied and lowest unoccupied molecular orbitals	−7.45	Reflects molecular reaction activity and electron transfer ability [46].
Ionization energy/eV	Energy required to remove an electron from an molecule	10.35	
D-band center/eV	Central energy of the D-band in electron spectroscopy	1.67	
Median	The average value of molecular charge distribution	0.11	Influences adsorption through electrostatic interaction forces [47].
Min	The maximum value of molecular charge distribution	−0.22	
Max	The minimum value of molecular charge distribution	0.11	
Q1	The first quartile of the molecular charge distribution	−0.14	
Q3	The third quartile of the molecular charge distribution	0.11	

developed through the scikit-learn ML framework [53]. In feature engineering, the 18 initial features were first ranked and selected using recursive feature elimination (RFE) (Tables S6–S9) and then normalized by the Min–Max approach. For model optimization, three-fold cross-validation was performed using grid search (with the parameter space detailed in Supplementary Material) to obtain the optimal combination of hyperparameters based on the validation performance and build the final model. In order to improve the model's generalization ability, Tanimoto similarity-guided stratified sampling was used to divide the test set: MACCS fingerprints were clustered using the K-Means algorithm; 20% of each category was randomly sampled to ensure adequate representation of chemical spatial diversity. The t-distributed stochastic neighbor embedding (t-SNE) technique was employed to visualize clustering patterns. Finally, feature importance analysis was performed using the SHapley Additive exPlanations (SHAP) [54] method, which systematically describes feature contributions and their directional impact on model predictions.

As shown in Fig. 2, when constructing the Transformer-based ML model [55], the feature matrix is based on the atomic 3D spatial coordinates. The feature matrix was padded to 40 × 4 dimensions by padding with zeros, where 40 represents the maximum number of atoms and the 4 columns contain the 3D coordinates and atom type encoding. The test set was divided using a stratified sampling approach based on chemical knowledge categorization, with 20% of molecules in each category randomly selected to maintain categorical balance. To improve the model's generalizability, molecular geometric transformations were used as data augmentation strategies. The model is trained using the Adam optimizer with mean square error (MSE) as the loss function for parameter optimization. And the model's architecture (Fig. S5) is as follows:

1) a multi-head self-attention mechanism layer for capturing inter-atom interactions;
2) a normalization layer module to optimize the feature distribution;
3) a global average pooling layer to achieve feature dimensionality reduction;
4) a regression header consisting of two fully connected layers.

The coefficient of determination (R^2), which is defined by the following formula, is used as a performance evaluation measure for regression ML models:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

where y_i represents the true value, and $\hat{y}_i$ represents the predicted value, and $\bar{y}_i$ is the mean of true value, and n indicates the sample size. This statistic evaluates regression model efficacy by quantifying the proportion of the variance in the dependent variable that is predictable from the independent variables, with values ranging from 0 to 1. All codes and datasets used in this work have been made publicly available on GitHub (https://github.com/WJ-19981/SMML).

3. Results and Discussion

3.1. Database analysis

ESG play an important role in chip production, which mainly include functional reaction gases and auxiliary process gases. For example, fluorinated gases mainly play a role in etching, and the high-purity inert gas N_2 is used as a carrier gas. In this study, data on the properties of 101 molecules associated with chip production were collected. All molecules were categorized into five distinct groups (Table S4) based on their structural and chemical properties. Tanimoto similarity analysis (Fig. 3(a)) revealed that Category IV demonstrated the strongest similarity to Category II, suggesting shared molecular features of halogen elements, particularly fluorine, between these groups. The low similarity coefficients between other category pairs confirm the validity of the categorization scheme.

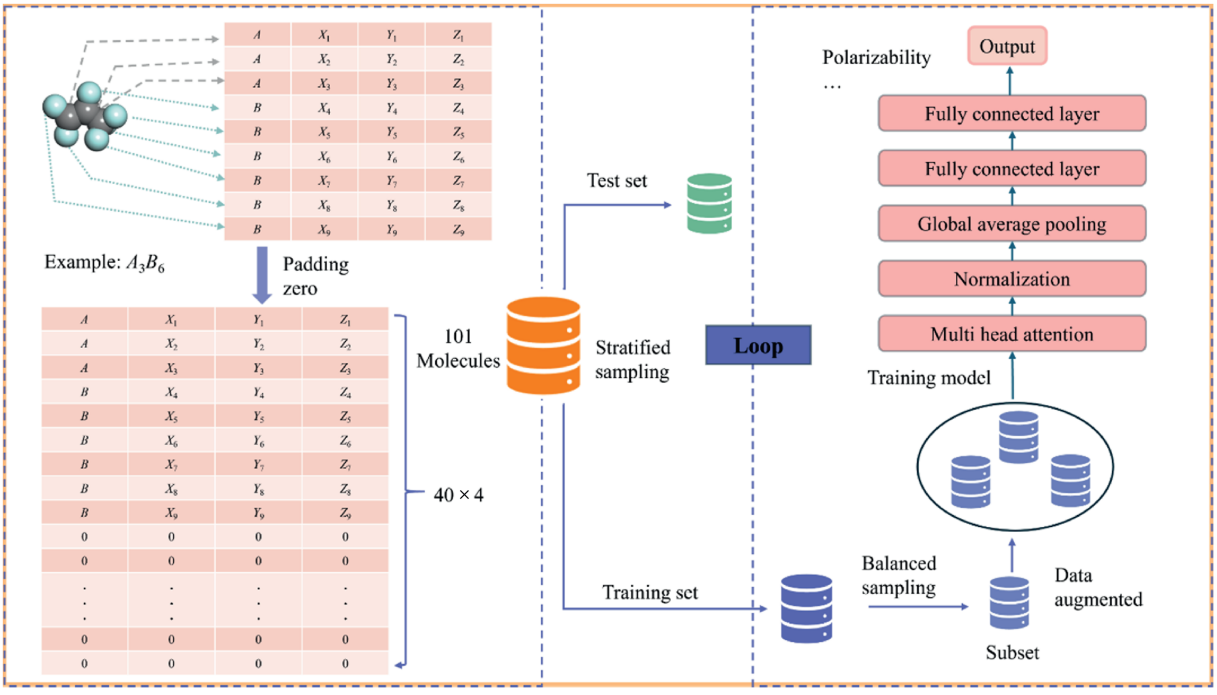


Fig. 2. The process of training the Transformer-based molecular property prediction model using balanced data-augmented (BD-TMPP).

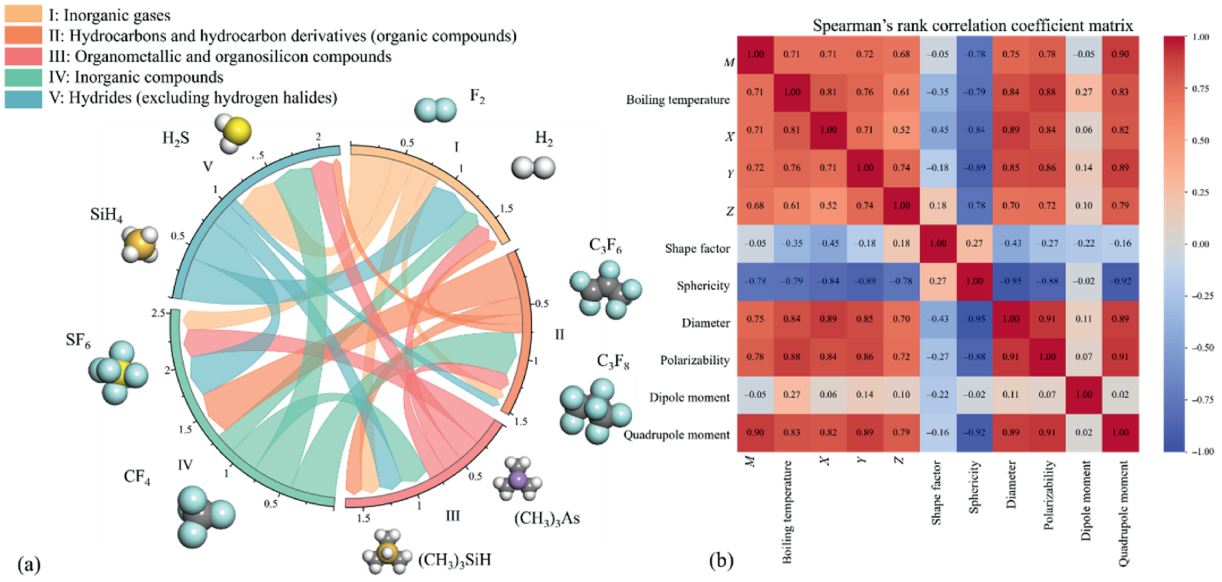


Fig. 3. (a) Mean Tanimoto similarity between different categories of molecules (the thickness of the chords representing the degree of similarity); (b) Spearman's rank correlation coefficient matrix between parameters of molecular physical properties (1 indicates positive correlation and −1 indicates negative correlation).

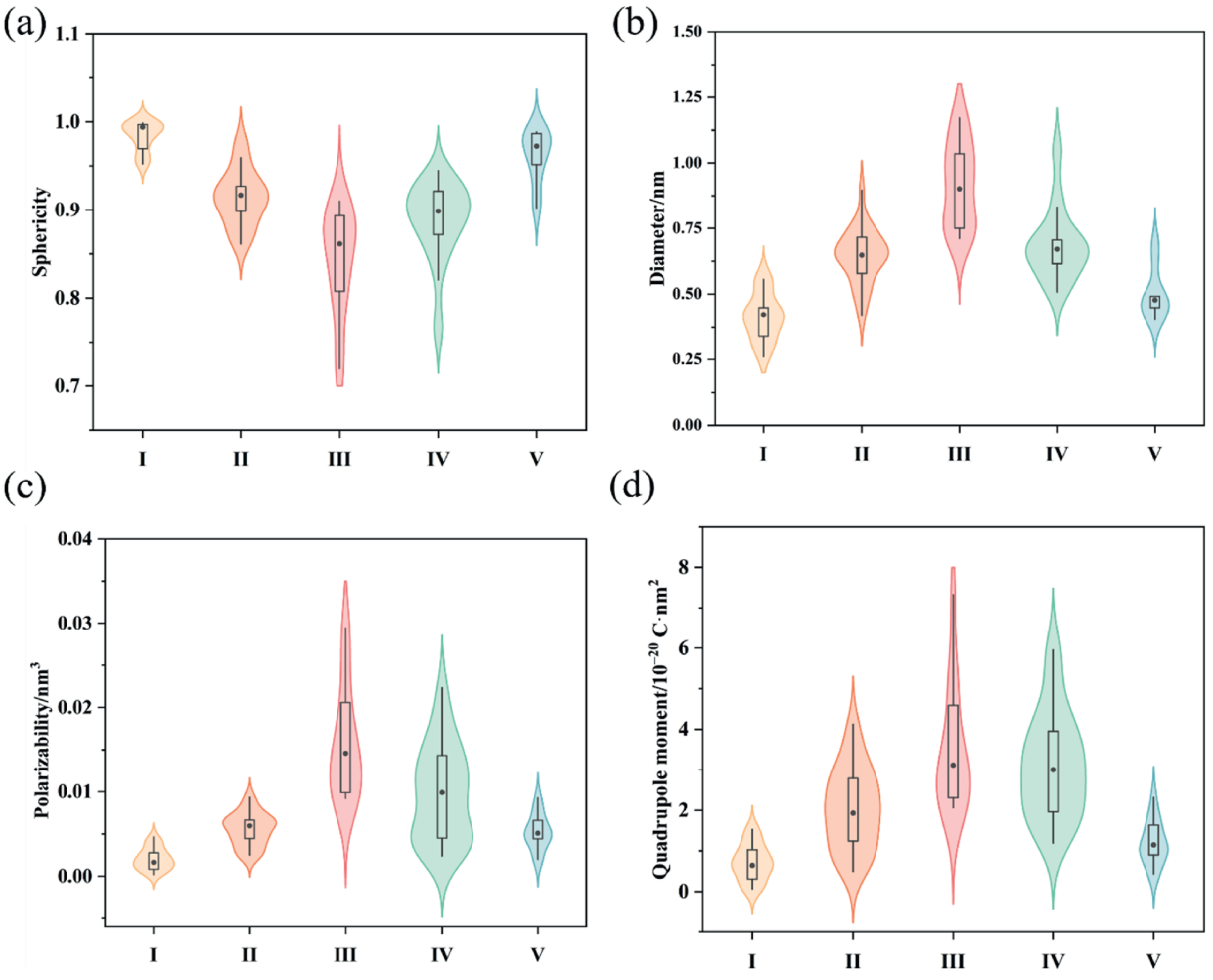


Fig. 4. Violin plots of molecular parameters related to adsorption: (a) sphericity; (b) diameter; (c) polarizability; (d) quadrupole moment. I: inorganic gases; II: hydrocarbons and hydrocarbon derivatives; III: organometallic and silicon compounds; IV: inorganic compounds; V: hydrides (excluding hydrogen halides).

For all molecules, 19 parameters were obtained through data collection and DFT calculations, thereby establishing a comprehensive database (Tables S1–S3). In this study, the B3LYP functional and def2-SVP basis set were employed. This choice was mainly due to the enormous number of molecules under consideration, as using a more complete basis set (triplet ζ or quartet ζ) would significantly increase the computational cost. Inevitably, this selection introduces errors due to incomplete basis set errors and functional approximation errors. These errors also propagate to the ML model, resulting in deviations between the model's predicted values and the actual values. Table S5 presents the comparison of the reported polarizability values of some molecules with the values calculated through DFT, and the results show that the relative error between the DFT-calculated values and the experimental values is small (typically less than 10%), which indicates the accuracy of the calculations in this study.

Correlation analysis of 19 physical parameters using SRCC and PCC (complete dataset in Figs. S1–S2) revealed statistically significant correlations. Notably in Fig. 3(b), sphericity exhibited exceptionally strong correlations ($|\rho| > 0.9$) with molecular diameter and quadrupole moment, suggesting intrinsic geometric–electronic couplings. And molecular diameter exhibited strong correlations ($|\rho| > 0.8$) with both polarizability and quadrupole moment, whereas the polarizability-quadrupole moment relationship displayed a comparably strong correlation strength ($|\rho| = 0.92$). Furthermore, molecular size descriptors (*X*, *Y*, *Z*), boiling point, and molecular weight exhibited significant correlations ($|\rho| = 0.64–0.95$) with the parameters cluster. These findings suggest that there may be a mathematical derivation or physical property of the correlation between some of the physical parameters.

Fig. 4 depicts comparative violin plots analyzing four molecular parameters across classification Categories I–V using kernel density estimation. As shown in Fig. 4(a), Category I predominantly exhibits high sphericity values (>0.9), implying near-ideal spherical geometry characteristic of closed-shell inorganic gas molecules where surface energy minimization prevails. The polarizability distributions in Fig. 4(c) show that Categories I and II maintain low and narrowly distributed values, consistent with molecular systems where rigid inorganic frameworks restrict electron delocalization. Fig. 4(d), the distribution of quadrupole moments for Category V is more concentrated and has a smaller value, indicating a highly symmetric and consistent of the charge distribution. And Category III displays the widest distribution, highlighting the structural flexibility of organometallic compounds where d-orbital hybridization enables multifarious electronic configurations. In contrast, Category V exhibits parametric consistency across all parameters, which suggests strict stoichiometric constraints in hydrogen-bonded networks.

These molecular parameters provide valuable insights into the mechanisms of adsorbate-adsorbent interactions. The low polarizability of Categories I and II suggests limited induced dipole interactions with the adsorbent surface, favoring van der Waals-dominated adsorption over stronger electronic coupling. In order to adsorb trace C_3H_8 from CH_3F, Huang *et al.* [13] took advantage of the strong van der Waals forces between C_3H_8 and CMS-600 to achieve faster adsorption of C_3H_8 on CMS-600. Category V exhibit reduced sphericity and increased polarity compared to Group I. Consequently, polar zeolite molecular sieves with appropriately matched pore sizes should be selected. In the adsorption study of H_2S, the adsorption capacity of NaY zeolite reached 6 mmol·kg^{-1}. Although the adsorption isotherms indicated that the adsorption of H_2S on NaY was dominated by physisorption, the heat of adsorption reached 40 kJ·mol^{-1} [56]. Categories III and IV exhibit broad distributions in molecular diameter, sphericity, polarizability, and quadrupole moments. Therefore, adsorbent selection requires precise matching with the specific physicochemical properties of the target molecules to optimize adsorption efficiency. Kim *et al.* [57] modulated the interaction between SF_6 and UiO-66-X surfaces by introducing different polar functional groups, resulting in high SF_6/N_2 selectivity (200).

3.2. Machine learning

Four key molecular physical property parameters served as independent prediction targets within Bayesian ridge regression framework, implemented through the schematic in Fig. 5(a). Cluster analysis was performed with MACCS fingerprints, and the clustering results were visualized after dimensionality reduction by t-SNE (Fig. 5(b)), showing significant group separation, confirming the discriminative ability of the fingerprints in chemical space delineation. To enhance model generalization, cluster-

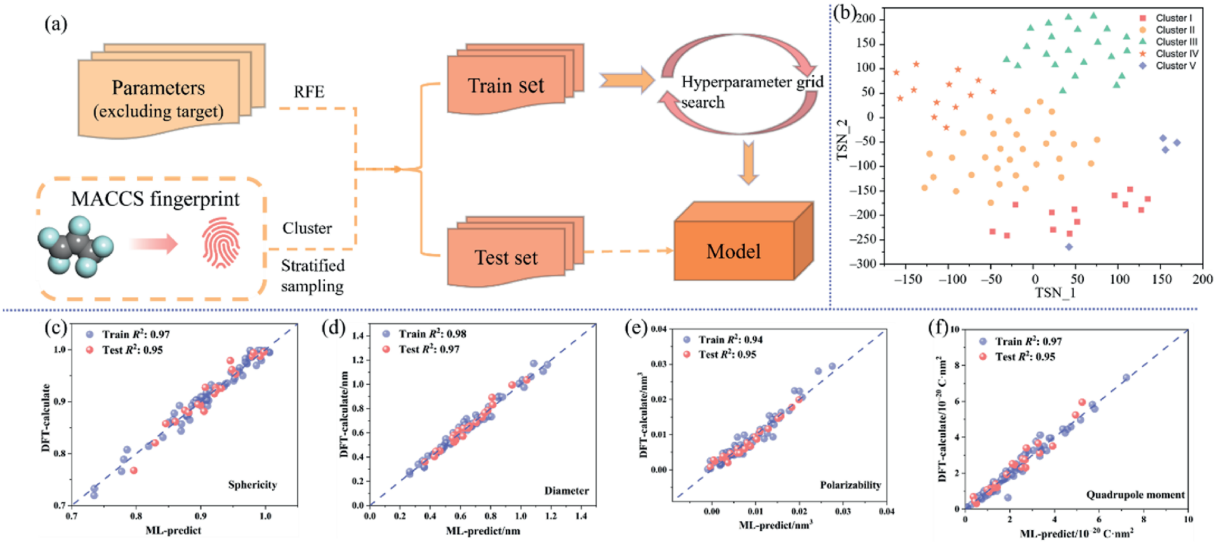


Fig. 5. (a) The training process of the ML-model. (b) The clusters obtained from the K-means algorithm. (c) Model prediction accuracy of sphericity. (d) Model prediction accuracy of diameter. (e) Model prediction accuracy of polarizability. (f) Model prediction accuracy of quadrupole moment.

stratified sampling was applied during test set construction, ensuring balanced representation across categories.

Fig. 5(c)–(f) present the Bayesian model's performance in predicting the four physicochemical properties. The model achieved coefficients of determination (R^2) of 0.97 (sphericity), 0.98 (diameter), 0.94 (polarizability), and 0.97 (quadrupole moment) on the training set, with corresponding values of 0.95, 0.97, 0.95, and 0.95 on the test set, indicating robust predictive performance without significant overfitting. These findings demonstrate that the Bayesian model exhibits strong generalization capacity and effectively extracts relevant feature information from the dataset.

Aiming at the problem of chance in model performance that may be caused by the limited sample size, ten independent random resampling experiments were used to evaluate model stability. As shown in Fig. S3, the R^2 distributions of the four ML models demonstrate distinct distribution characteristics between the training and test sets. The Random Forest, Decision Tree, and Gradient Boosting Tree models exhibit highly concentrated R^2 value distributions across all features (particularly sphericity and diameter) in the training set, indicating stable training processes. However, in the test set, polarizability and quadrupole moment exhibited substantial dispersion in their R^2 distributions, which suggests limited generalization capabilities of these models for such targets. Notably, the Bayesian regression model showed significantly tighter R^2 distribution clustering in the test set compared to that of other models, indicating superior generalization performance. This means that only with the features listed in Tables S6–S9 can predictions of sphericity, diameter, polarizability, and quadrupole moments be made by the Bayesian model without the need for expensive DFT calculations. And the model does not encounter the convergence problem that often occurs in DFT calculations, thereby avoiding excessive computational time caused by such failures. However, some features in Tables S6–S9 still need to be obtained through DFT computation. In order to completely eliminate the reliance on DFT calculations, the atomic geometric coordinates are instead used as features to train the new model.

Following the process outlined in Fig. 2, the balanced data-augmented Transformer-based molecular property prediction (BD-TMPP) model was constructed. Based on the process in Fig. S6 (a) and a Bayesian model (Bayesian-MF) was developed as a comparative benchmark. Additionally, the direct data-augmented Transformer-based model (DD-TMPP) was developed following the process outlined in Fig. S6(b). The three models used two different structural descriptors as input features: the Bayesian-MF model utilized molecular fingerprints, whereas both BD-TMPP and DD-TMPP leveraged atomic geometric coordinates. Notably, their data augmentation mechanisms were quite different: BD-TMPP extracted the training subset through hierarchical sampling based on molecular categorization before data augmentation, while DD-TMPP employed the complete training set for global augmentation. Given the limited dataset size, rigorous evaluation was conducted by performing 10 independent test set divisions with different random seeds to assess model performance stability.

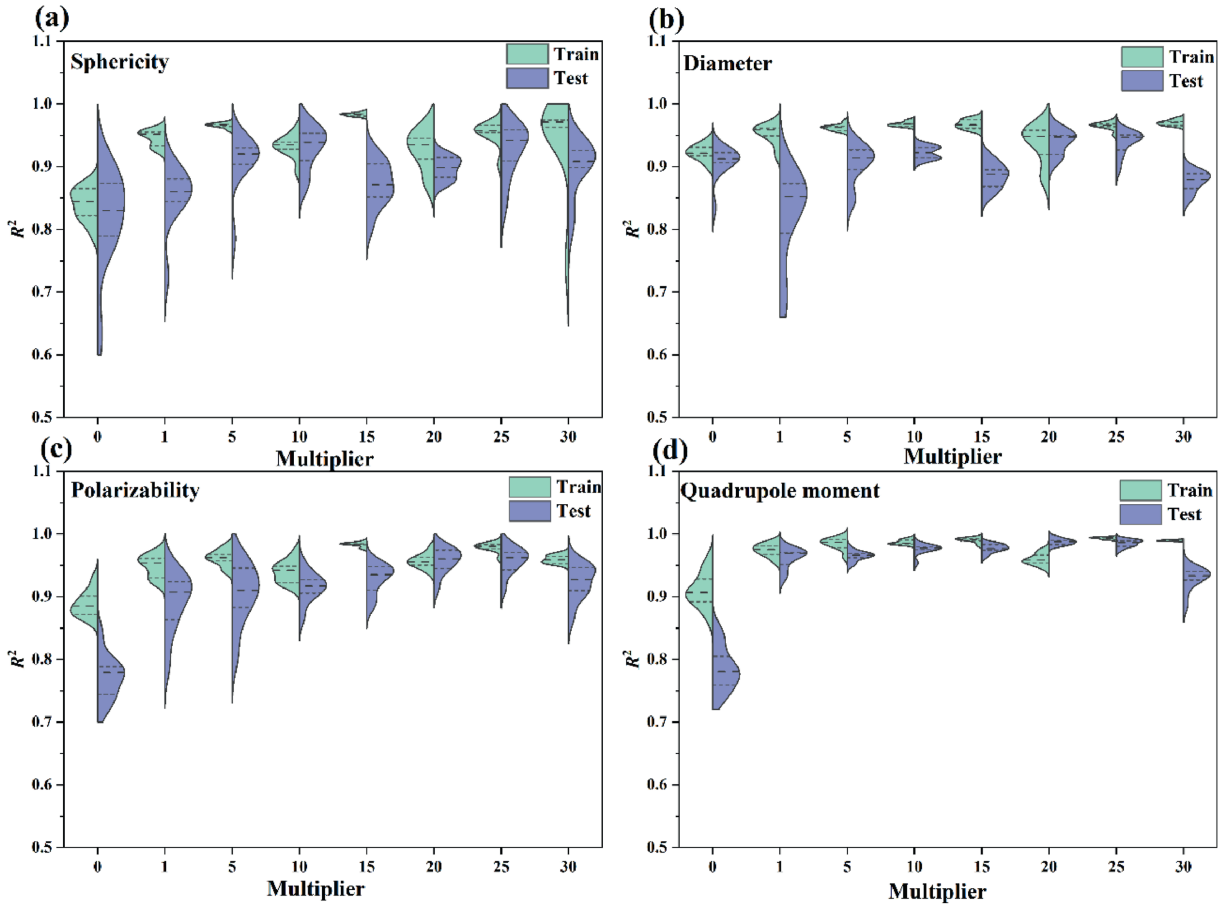


Fig. 6. Split violin plots of the R^2 distribution of the regression coefficient for different data enhancement multiplier (left side: training set; right side: test set): (a) sphericity; (b) diameter; (c) polarizability; (d) quadrupole moment.

As illustrated in Fig. S7, the distribution characteristics of R^2 across the three models reflect the impact of distinct modeling and data augmentation strategies on predictive generalization capabilities. The Bayesian-MF model shows significant overfitting across all four prediction targets. This limitation stems from the intrinsic characteristics of molecular fingerprint data: while its high-dimensional array effectively characterizes molecular substructure mapping relationships, it restricts the implementation of data augmentation operations such as translational transformations. As demonstrated by principal component analysis (PCA) of dimensionality-reduction analyses in Figs. S8 and S9, the PCA approach fails to significantly alleviate the model's overfitting, highlighting the limitations of molecular fingerprints as structural descriptors. In contrast, the Transformer-based BD-TMPP and DD-TMPP models, which employed atomic geometric coordinate matrices as structural descriptors, demonstrate superior generalization performance. This enhanced performance originates from: (1) the multi-head attention mechanism's ability to effectively learn interatomic interactions; (2) the utilization of atomic geometric coordinates as structural descriptors, which enable data augmentation to mitigate overfitting in small-sample learning scenarios. It is worth emphasizing that, BD-TMPP surpasses DD-TMPP in generalization performance through its balanced data augmentation strategy, which facilitates equitable learning of feature representations across molecular classes during iterative training cycles.

Due to the better generalization performance of the BD-TMPP model, this study systematically examined the effects of data augmentation magnitude (Fig. 6) and hyperparameter configurations (Fig. S10) on predictive performance. Fig. 6 demonstrates that model instability manifests as broad R^2 distribution ranges (0.60−0.99) without data augmentation, accompanied by overfitting in polarizability and quadrupole moment predictions. Within the range of up to 20× augmentation, the additional samples provide more equivalent geometric perspectives, which help the model learn invariance to translation, rotation, and symmetry, thus improving performance. When augmentation magnitude reaches 20×, R^2 values demonstrate marked concentration with concomitant enhancement of generalization metrics. The performance of the model deteriorates when the multiplier exceeds 20×. This may be due to the limitation of the model complexity; additionally, excessive augmentation may introduce numerical noise (making it more difficult for model training to converge) and even shift the overall data distribution, thereby reducing generalization ability. Therefore, 20 times is empirically established as the optimal magnitude. The hyperparameter optimization tests (Fig. S10) revealed the following effects. The number of attention heads exhibits a non-linear relationship with model accuracy, achieving peak performance at *num_heads* = 8, beyond which overfitting intensifies. And batch size demonstrates a marked improvement at *batch_size* = 64 (R^2 increases from 0.65 to 0.93), with diminishing returns beyond this threshold. In addition, validation set partition ratio inversely correlates with model performance, reaching optimal generalization at *validation_split* = 0.1. Through systematic experimentation, the optimal hyperparameter configuration was obtained, as shown in Table S14.

As demonstrated in Fig. S11, the BD-TMPP model achieves exceptional predictive performance, with consistency between training and test sets across four molecular parameters: sphericity

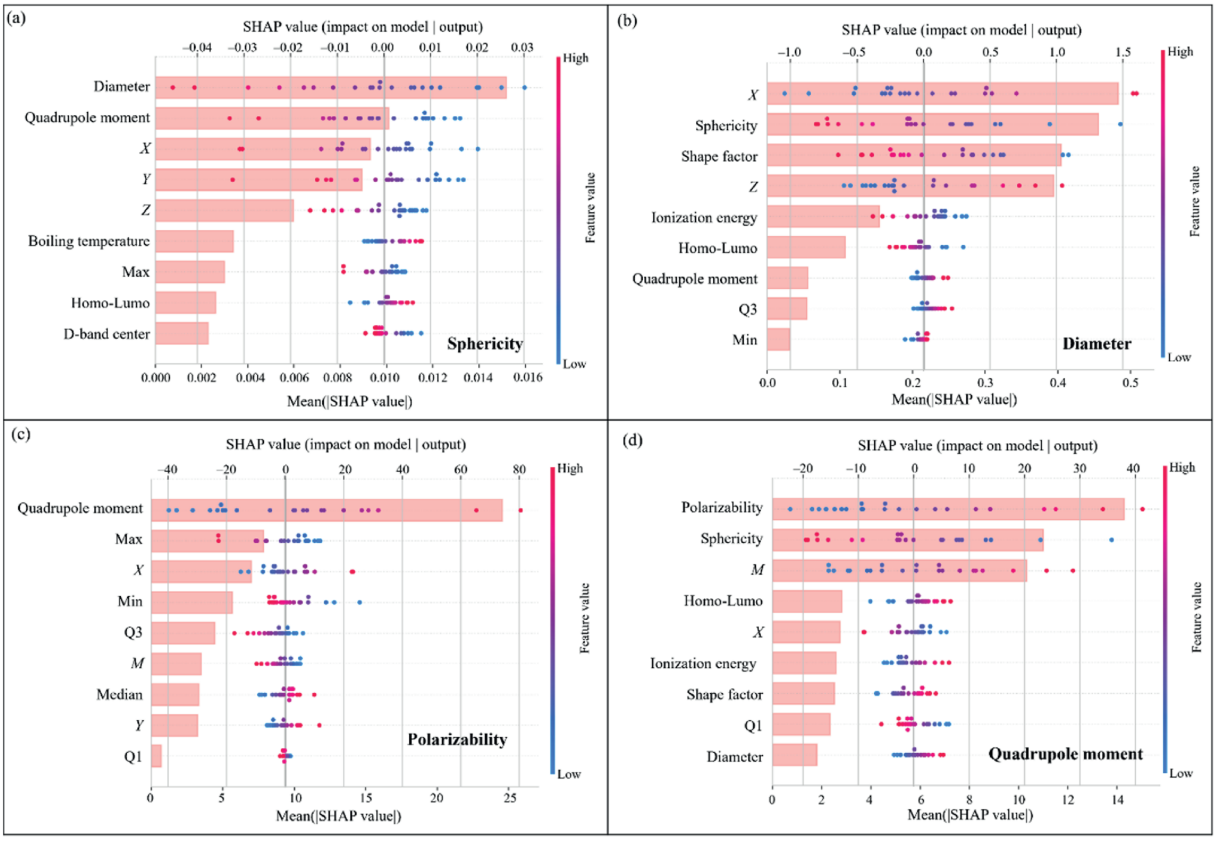


Fig. 7. Analysis of the feature importance and SHAP values within Bayesian (feature importance rankings: bar and directional correlation analysis: dot plots (red = positive, blue = negative)): (a) sphericity; (b) diameter; (c) polarizability; (d) quadrupole moment.

(R^2 = 0.91/0.92), diameter (R^2 = 0.94/0.95), polarizability (R^2 = 0.96/0.98), and quadrupole moment (R^2 = 0.98/0.99). Table 2 shows that the BD-TMPP model predicts the property parameters of all molecules in just 0.641 s, and this speed is 5 orders of magnitude faster than the time required for DFT calculations. This demonstrates that atomic coordinates, when used as structural descriptors, not only maintain accuracy but also far surpass DFT calculations in computational efficiency.

In summary, both the Bayesian regression model (utilizing molecular property descriptors) and the BD-TMPP model (utilizing molecular structural descriptors) provide effective predictions for sphericity, diameter, polarizability, and quadrupole moments. Fig. 7 presents the SHAP-based feature attribution analysis results for the Bayesian regression model. The SHAP value distributions in the quadrupole moment parameter space (Fig. 7(a), (c)) indicate that quadrupole moment enhancement inversely correlates with sphericity while directly correlating with polarizability enhancement, consistent with quantum chemical theory predictions. Meanwhile, the inverse relationship between molecular diameter and sphericity (Fig. 7(a), (c)) elucidates the physical origin of diminished geometric symmetry stemming from elevated molecular length-to-diameter ratios. Furthermore, polarizability positively correlates with quadrupole moments (Fig. 7(d)). Feature importance reveals significant multidimensional interdependencies among sphericity, diameter, polarizability, and quadrupole moment, which are reciprocally corroborated by Spearman rank correlation analysis (Fig. 3(b)). The replacement importance analysis in Fig. S4 likewise yields similar results, which demonstrate the ability of Bayesian models to effectively learn the intrinsic physical relationships of parameters in a database.

Table 2
Computation time of three methods (time to compute all samples; device configuration: CPU: i5-13600KF; GPU: RTX 4060; RAM: 32 GB).

	DFT	Bayesian	BD-TMPP
Time/s	23941	0.048	0.641

Fig. 8 shows the graphical representation of data transformation and the score matrix of the attention layer in the model to elucidate the prediction mechanism of the BD-TMPP model. Starting from the original atomic coordinates as structural descriptors and finally obtaining the property parameters of a molecule is closely related to the chemical principle of "structure-property relationship". In addition, the row-column correspondence of the matrix of attention scores reveals that BD-TMPP achieves end-to-end molecular property prediction through the capture of interatomic spatial correlations. Meanwhile, this core principle based on the spatial relationships between atoms

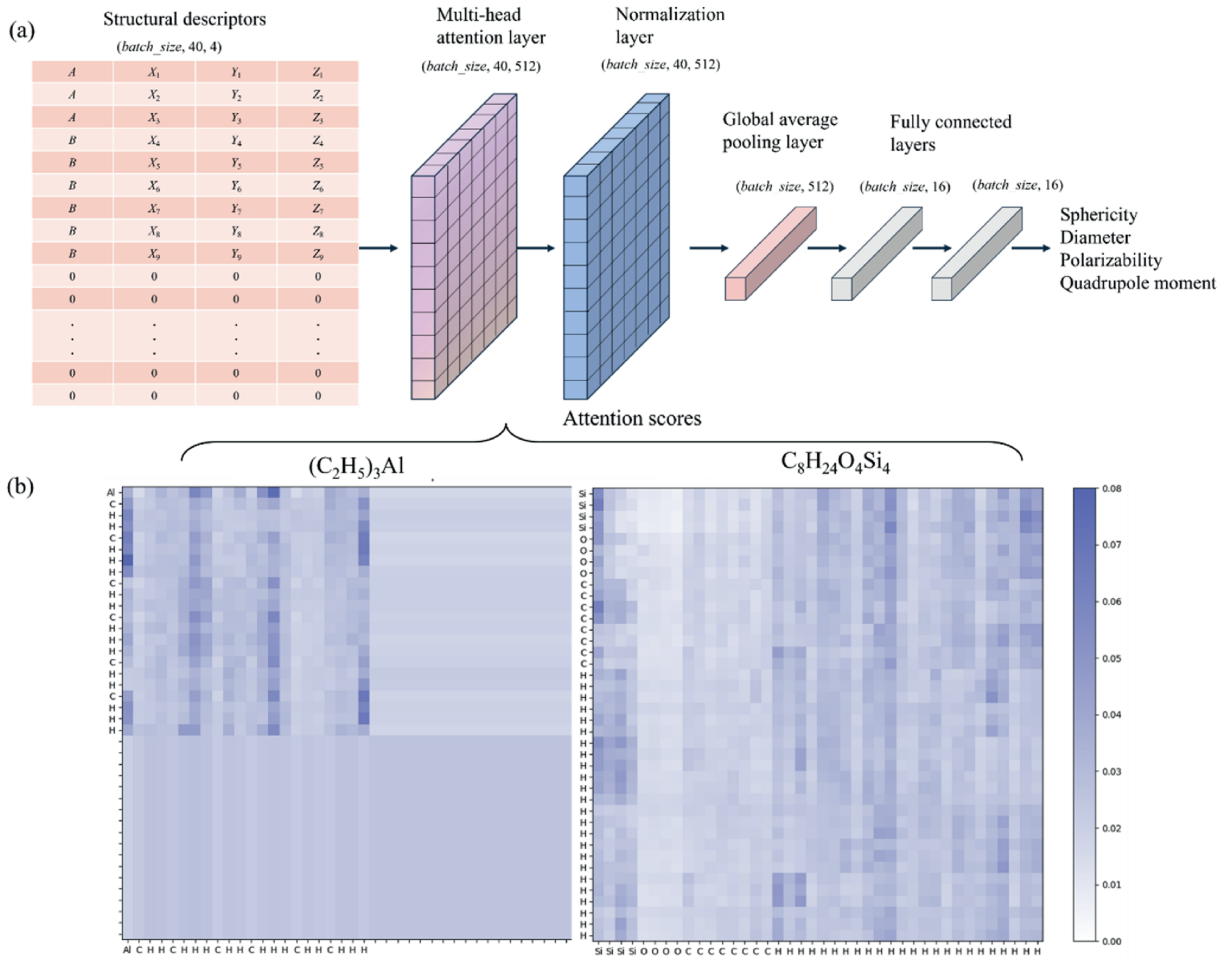


Fig. 8. (a) Diagram of data transformation in the BD-TMPP model. (b) Attention scores of different samples in BD-TMPPP model. The coordinate of each score represents the degree of attention of the atoms in the row to the atoms in the column.

imparts permutation invariance to the model with respect to the input atomic order. Table S15 shows that the prediction results of this model remain largely consistent whether the atomic sequence is shuffled or not. The multi-head attention layer captures interactions between atoms regardless of their order, as attention mechanisms compute pairwise similarities that are inherently permutation-equivariant at the intermediate level. The subsequent global average pooling layer aggregates the attention outputs across the sequence dimension, effectively rendering the final representation permutation-invariant by averaging features over all atoms [58,59]. Furthermore, the model demonstrates molecular scale-adaptive attention allocation that mitigates zero-padding interference in parameter estimation *via* dynamic masking, thereby directly mapping atomic-level structural features to macroscopic molecular properties, in line with the fundamental "structure-property relationship" principle.

4. Conclusions

This study developed a multidimensional database covering 101 semiconductor-critical molecules (19 physical parameters), addressing key data gaps in ESG research and advancing adsorbent development. Comparative analysis of four molecular parameters across taxonomic groups revealed characteristic variations that guided adsorbent selection strategies. Using Bayesian regression, we achieved multivariate predictions of target parameters from molecular properties (test-set $R^2 = 0.95–0.97$). SHAP analysis uncovered intrinsic property correlations among critical molecular features. Further development of the BD-TMPP model enabled end-to-end predictions of quadrupole moments ($R^2 = 0.99$) and polarizability ($R^2 = 0.98$) *via* the capture of interatomic spatial correlations. Compared with traditional DFT calculations, the model achieves a five-orders-of-magnitude improvement in computational efficiency while maintaining accuracy, demonstrating a successful application of the "structure-property relationship" theory in chemical ML. This work will reduce the exploration of time required for ESG purification studies and offer significant assistance in advancing the adsorption studies of ESG in porous materials.

CRediT Authorship Contribution Statement

Zhikang Wu: Writing – original draft, Validation, Methodology, Investigation. Ying Wu: Writing – review & editing, Software, Methodology. Guang Miao: Writing – review & editing. Runze Chen: Investigation. Lingjun Ma: Investigation. Hongxia Xi: Writing – review & editing. Jing Xiao: Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

We gratefully acknowledge the support from the National Natural Science Foundation of China (U24A20532 and 22278146), Guangdong Basic and Applied Basic Research Team Fund (2024B1515040016), and Fundamental Research Funds for the Central Universities.

Supplementary Material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cjche.2025.09.012.

References

[1] M.B. Chang, J.S. Chang, Abatement of PFCs from semiconductor manufacturing processes by nonthermal plasma technologies: a critical review, *Ind. Eng. Chem. Res.* 45 (12) (2006) 4101–4109.
[2] W.X. Zhang, Y.H. Li, Y. Wu, Y. Fu, S.H. Chen, Z.H. Zhang, S.J. He, T. Yan, H.P. Ma, Fluorinated porous organic polymers for efficient recovery perfluorinated electronic specialty gas from exhaust gas of plasma etching, *Sep. Purif. Technol.* 287 (2022) 120561.
[3] M.Z. Zheng, W.J. Xue, T.A. Yan, Z.F. Jiang, Z. Fang, H.L. Huang, C.L. Zhong, Fluorinated MOF-based hexafluoropropylene nanotrap for highly efficient purification of octafluoropropane electronic specialty gas, *Angew. Chem. Int. Ed.* 63 (15) (2024) e202401770.
[4] W.G. Cui, T.L. Hu, X.H. Bu, Metal–organic framework materials for the separation and purification of light hydrocarbons, *Adv. Mater.* 32 (3) (2020) e1806445.
[5] K.E. Lamb, M.D. Dolan, D.F. Kennedy, Ammonia for hydrogen storage; a review of catalytic ammonia decomposition and hydrogen separation and purification, *Int. J. Hydrog. Energy* 44 (7) (2019) 3580–3593.
[6] R.B. Lin, S.C. Xiang, H.B. Xing, W. Zhou, B.L. Chen, Exploration of porous metal–organic frameworks for gas separation and purification, *Coord. Chem. Rev.* 378 (2019) 87–103.
[7] X.Q. Li, K. Chen, R.L. Guo, Z. Wei, Ionic liquids functionalized MOFs for adsorption, *Chem. Rev.* 123 (16) (2023) 10432–10467.
[8] X. Peng, R.G. Pan, X. Li, W.M. Zhong, F. Qian, Molecular descriptor-assisted interpretable machine learning: a scheme for guiding the synthesis of zeolites with target structures, *Chem. Eng. Sci.* 308 (2025) 121378.
[9] X. Zhao, Y.X. Wang, D.S. Li, X.H. Bu, P.Y. Feng, Metal–organic frameworks for separation, *Adv. Mater.* 30 (37) (2018) 1705189.
[10] S.A. Chen, W.L. Wu, Z.Y. Niu, D.Q. Kong, W.B. Li, Z.L. Tang, D.H. Zhang, High adsorption selectivity of activated carbon and carbon molecular sieve boosting CO_2/N_2 and CH_4/N_2 separation, *Chin. J. Chem. Eng.* 67 (2024) 282–297.
[11] J.W. Huang, C.T. Yang, X.Y. Zhou, X.X. Li, Z.L. Du, L. Zhu, H. Yin, G. Miao, J. Xiao, Sub-nanopore orifice control on carbonaceous adsorbent boosting N_2/CH_4 inverse separation with ultra-high selectivity, *Carbon* 233 (2025) 119922.
[12] C.Q. Su, W.T. Jiang, Y. Guo, G.D. Yi, Z.X. Li, H. Li, Rational molecular design of P-doped porous carbon material for the VOCs adsorption, *Chin. J. Chem. Eng.* 79 (2025) 155–163.
[13] J.W. Huang, J.J. Peng, X. Wei, S.J. Du, C.T. Yang, J. Xiao, Synergetic thermodynamic/kinetic separation of C_3H_8/CH_3F on carbon adsorbents for ultrapure fluoromethane electronic gas, *AIChE J.* 69 (5) (2023) e18027.
[14] J.R. Li, R.J. Kuppler, H.C. Zhou, Selective gas adsorption and separation in metal–organic frameworks, *Chem. Soc. Rev.* 38 (5) (2009) 1477–1504.
[15] R.M. Barrer, Molecular sieves, *Nature* 249 (5459) (1974) 783.
[16] B.E. Poling, J.M. Prausnitz, J.P. O'Connell, Properties of Gases and Liquids, fifth ed., McGraw-ill Education, New York (2001).
[17] S. Sircar, Basic research needs for design of adsorptive gas separation processes, *Ind. Eng. Chem. Res.* 45 (16) (2006) 5435–5448.
[18] D.R. Lide, CRC Handbook of Chemistry and Physics, 88th Edition, Taylor & Francis, London (2007).
[19] J.A. Keith, V. Vassilev-Galindo, B.Q. Cheng, S. Chmiela, M. Gastegger, K.R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems, *Chem. Rev.* 121 (16) (2021) 9816–9872.
[20] J.B. Hu, J.Y. Cui, B. Gao, L.F. Yang, Q. Ding, Y.J. Li, Y.M. Mo, H.J. Chen, X.L. Cui, H.B. Xing, Machine-learning-assisted exploration of anion-pillared metal organic frameworks for gas separation, *Matter* 5 (11) (2022) 3901–3911.
[21] J.Y. Cui, F. Wu, W. Zhang, L.F. Yang, J.B. Hu, Y. Fang, P. Ye, Q. Zhang, X. Suo, Y.M. Mo, X.L. Cui, H.J. Chen, H.B. Xing, Direct prediction of gas adsorption *via* spatial atom interaction learning, *Nat. Commun.* 14 (1) (2023) 7043.
[22] W.Y. Zhou, H.S. Feng, S.H. Zhou, M.X. Wang, Y.P. Chen, C.Y. Lu, H. Yuan, J. Yang, Q. Li, L.X. Tan, L.C. Dong, Y.W. Zhang, Designing and screening single-atom alloy catalysts for CO_2 reduction to CH_3OH *via* DFT and machine learning, *AIChE J.* 71 (3) (2025) e18678.
[23] N.D. Vo, D.H. Oh, S.H. Hong, M. Oh, C.H. Lee, Combined approach using mathematical modelling and artificial neural network for chemical industries: steam methane reformer, *Appl. Energy* 255 (2019) 113809.
[24] F. Ye, S. Ma, L. Tong, J.S. Xiao, P. Bénard, R. Chahine, Artificial neural network based optimization for hydrogen purification performance of pressure swing adsorption, *Int. J. Hydrog. Energy* 44 (11) (2019) 5334–5344.
[25] L.M.C. Oliveira, H. Koivisto, I.G.I. Iwakiri, J.M. Loureiro, A.M. Ribeiro, I.B.R. Nogueira, Modelling of a pressure swing adsorption unit by deep learning and artificial intelligence tools, *Chem. Eng. Sci.* 224 (2020) 115801.

[26] Q.M. Pu, Y.H. Li, H. Zhang, H.D. Yao, B. Zhang, B.J. Hou, L. Li, Y.L. Zhao, L.N. Zhao, Screen efficiency comparisons of decision tree and neural network algorithms in machine learning assisted drug design, *Sci. China Chem.* 62 (4) (2019) 506–514.
[27] J. Peña-Guerrero, P.A. Nguewa, A.T. García-Sosa, Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases, *Wires Comput. Mol. Sci.* 11 (5) (2021) e1513.
[28] H. Liu, H.W. Xu, W.G. Zhu, Y. Zhou, K. Xue, Z.Y. Zhu, Y.L. Wang, J.G. Qi, Prediction of the viscosity of green deep eutectic solvents by constructing ensemble model based on machine learning, *Chem. Eng. Sci.* 304 (2025) 120987.
[29] M.R. Youcefi, F.M. Alqahtani, M. Nait Amar, H. Djema, M. Ghasemi, An interpretable and explainable deep learning model for predicting hydrogen solubility in diverse chemicals, *Chem. Eng. Sci.* 304 (2025) 121048.
[30] F. Kretschmer, J. Seipp, M. Ludwig, G.W. Klau, S. Böcker, Coverage bias in small molecule machine learning, *Nat. Commun.* 16 (1) (2025) 554.
[31] Y. Wan, J.L. Wu, T.J. Hou, C.Y. Hsieh, X.W. Jia, Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation, *Nat. Commun.* 16 (1) (2025) 413.
[32] F. Wang, Z.Y. Bi, L.F. Ding, Q.Y. Yang, Large-scale computational screening of metal–organic frameworks for D_2/H_2 separation, *Chin. J. Chem. Eng.* 54 (2023) 323–330.
[33] X.Y. Pi, J.F. Lu, S.M. Li, J.L. Zhang, Y.L. Wang, H.Y. He, Computer-aided ionic liquid design for green chemical processes based on molecular simulation and artificial intelligence, *Sep. Purif. Technol.* 361 (2025) 131585.
[34] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, *Nature* 559 (7715) (2018) 547–555.
[35] W.P. Walters, R. Barzilay, Applications of deep learning in molecule generation and molecular property prediction, *Acc. Chem. Res.* 54 (2) (2021) 263–270.
[36] X.X. Yu, Y.H. Shen, Z.B. Guan, D.H. Zhang, Z.L. Tang, W.B. Li, Multi-objective optimization of ANN-based PSA model for hydrogen purification from steam-methane reforming gas, *Int. J. Hydrog. Energy* 46 (21) (2021) 11740–11755.
[37] J.Q. Wang, J.P. Liu, H.S. Wang, M.S. Zhou, G.L. Ke, L.F. Zhang, J.Z. Wu, Z.F. Gao, D.N. Lu, A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks, *Nat. Commun.* 15 (1) (2024) 1904.
[38] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, *J. Big Data* 6 (1) (2019) 60.
[39] J.J. Su, X.J. Yu, X.R. Wang, Z.J. Wang, G.Q. Chao, Enhanced transfer learning with data augmentation, *Eng. Appl. Artif. Intell.* 129 (2024) 107602.
[40] D. Andrae, U. Häußermann, M. Dolg, H. Stoll, H. Preuß, Energy-adjusted*ab initio* pseudopotentials for the second and third row transition elements, *Theor. Chim. Acta.* 77 (2) (1990) 123–141.
[41] K.A. Peterson, D. Figgen, E. Goll, H. Stoll, M. Dolg, Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-*d* group 16–18 elements, *J. Chem. Phys.* 119 (21) (2003) 11113–11123.
[42] F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy, *Phys. Chem. Chem. Phys.* 7 (18) (2005) 3297–3305.
[43] D. Rappoport, F. Furche, Property-optimized Gaussian basis sets for molecular response calculations, *J. Chem. Phys.* 133 (13) (2010) 134105.
[44] T. Lu, F.W. Chen, Multiwfn: a multifunctional wavefunction analyzer, *J. Comput. Chem.* 33 (5) (2012) 580–592.
[45] T. Lu, A comprehensive electron wavefunction analysis toolbox for chemists, Multiwfn, *J. Chem. Phys.* 161 (8) (2024) 082503.
[46] C. Zhao, J. Zhang, W.J. Zhang, Y. Yang, D.G. Guo, H.J. Zhang, L. Liu, Reveal the main factors and adsorption behavior influencing the adsorption of pollutants on natural mineral adsorbents: based on machine learning modeling and DFT calculation, *Sep. Purif. Technol.* 331 (2024) 125706.
[47] Z.Y. Yang, Z.Z. Chen, H.J. Gong, X.S. Wang, Copper oxide modified activated carbon for enhanced adsorption performance of siloxane: an experimental and DFT study, *Appl. Surf. Sci.* 601 (2022) 154200.
[48] M. Waskom, Seaborn: statistical data visualization, *J. Open Source Softw.* 6 (60) (2021) 3021.
[49] O. Méndez-Lucio, B. Baillif, D.A. Clevert, D. Rouquié, J. Wichard, *De novo* generation of hit-like molecules from gene expression signatures using artificial intelligence, *Nat. Commun.* 11 (1) (2020) 10.
[50] C.L. Xie, X.X. Zhuang, Z.M. Niu, R.X. Ai, S. Lautrup, S.J. Zheng, Y.H. Jiang, R.Y. Han, T.S. Gupta, S.Q. Cao, M.J. Lagartos-Donate, C.Z. Cai, L.M. Xie, D. Caponio, W.W. Wang, T. Schmauck-Medina, J.Y. Zhang, H.L. Wang, G.F. Lou, X.L. Xiao, W.H. Zheng, K. Palikaras, G. Yang, K.A. Caldwell, G.A. Caldwell, H.M. Shen, H. Nilsen, J.H. Lu, E.F. Fang, Amelioration of Alzheimer's disease pathology by mitophagy inducers identified *via* machine learning and a cross-species workflow, *Nat. Biomed. Eng.* 6 (1) (2022) 76–93.
[51] P. Willett, Similarity-based virtual screening using 2D fingerprints, *Drug Discov. Today* 11 (23–24) (2006) 1046–1053.
[52] P.R. Haddad, M. Taraji, R. Szücs, Prediction of analyte retention time in liquid chromatography, *Anal. Chem.* 93 (1) (2021) 228–256.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
[54] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Curran Associates, Inc., Red, Hook, NY, (2017) 4678–4777.
[55] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C.W. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ACL, Stroudsburg, PA, (2020) 38–45.
[56] L.H. de Oliveira, J.G. Meneguin, M.V. Pereira, E.A. da Silva, W.M. Grava, J.F. do Nascimento, P.A. Arroyo, H_2S adsorption on NaY zeolite, *Microporous Mesoporous Mater.* 284 (2019) 247–257.
[57] M.B. Kim, K.M. Kim, T.H. Kim, T.U. Yoon, E.J. Kim, J.H. Kim, Y.S. Bae, Highly selective adsorption of SF_6 over N_2 in a bromine-functionalized zirconium-based metal-organic framework, *Chem. Eng. J.* 339 (2018) 223–229.
[58] J. Lee, Y. Lee, J. Kim, A.R. Kosiorek, S. Choi, Y.W. Teh, Set transformer: a framework for attention-based permutation-invariant neural networks, *arXiv (2019) arXiv: 1810.00825.*
[59] D. Buterez, J.P. Janet, D. Oglic, P. Liò, An end-to-end attention-based approach for learning on graphs, *Nat. Commun.* 16 (1) (2025) 5244.