Liu *et al. J. Mater. Inf.* **2025**, *5*, 39
DOI: 10.20517/jmi.2025.10

Journal of Materials Informatics

Research Article

Open Access

Check for updates

Farthest point sampling in property designated chemical feature space as an effective strategy for enhancing the machine learning model performance for small scale chemical dataset

Yuze Liu[1,2], Lejia Wang[2,3], Weigang Zhu[1,4], Xi Yu[1,2,*]

[1]Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China.
[2]EpiScience (Shanghai) Artificial Intelligence Technology Co. Ltd, Shanghai 201210, China.
[3]Shanghai Boronmatrix Advanced Materials Technology Co. Ltd, Shanghai 201210, China.
[4]Key Laboratory of Advanced Energy Materials Chemistry (Ministry of Education), Nankai University, Tianjin 300071, China.

[*]**Correspondence to:** Prof. Xi Yu, Department of Chemistry, School of Science, Tianjin University, Building 3, Tianjin University Weijin Road Campus, No. 92 Weijin Road, Nankai District, Tianjin 300072, China. E-mail: xi.yu@tju.edu.cn

How to cite this article: Liu, Y.; Wang, L.; Zhu, W.; Yu, X. Farthest point sampling in property designated chemical feature space as an effective strategy for enhancing the machine learning model performance for small scale chemical dataset. *J. Mater. Inf.* **2025**, *5*, 39. https://dx.doi.org/10.20517/jmi.2025.10

Received: 7 Mar 2025 **First Decision:** 18 Apr 2025 **Revised:** 5 Jun 2025 **Accepted:** 6 Jun 2025 **Published:** 19 Jun 2025

Academic Editor: Rika Kobayashi **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Machine learning (ML) model development in chemistry and materials science often grapples with the challenge of small and imbalanced labeled datasets, a common limitation in experimental studies. These dataset imbalances can precipitate overfitting and diminish model generalization. Our study explores the efficacy of the farthest point sampling (FPS) strategy within targeted chemical feature spaces, demonstrating its capacity to generate well-distributed training sets and consequently enhance model performance. We rigorously evaluate this strategy across various ML models, including artificial neural networks, support vector machines, and random forests, using datasets with target physicochemical properties such as standard boiling points and enthalpy of vaporization. Our findings reveal that FPS-based models consistently surpass randomly sampled models, exhibiting superior predictive accuracy and robustness, alongside a marked reduction in overfitting. This improvement is particularly pronounced in smaller training set, attributable to increased diversity within the training data's chemical feature space. Consequently, FPS emerges as an effective and adaptable approach for achieving high-performance ML models at reduced cost by limited and biased experimental datasets typical in chemistry and materials science.

Keywords: Materials informatics, machine learning, farthest point sampling, small dataset, chemical database

INTRODUCTION

Machine learning (ML) has significantly advanced the fields of chemistry and material science[1-4], propelling the study of cheminformatics and enabling rapid structure-property prediction and design[5-12]. However, the inherent requirement on the extensive dataset for ML study raised significant challenge for practical application of ML in experimental science[13]. Often, labeled experimental chemical and material datasets are limited in size and coverage and most significantly imbalanced[14,15] due to constraints in data acquisition, including time, cost, and technical barriers. Consequently, ML models trained on these datasets, which are frequently subsampled randomly for training and testing, are prone to overfitting and exhibit diminished generalization capabilities due to the imbalanced nature of the data, where certain types of observations are disproportionately represented compared to others. The complexity of these challenges is further magnified by the high dimensionality of chemical data and the intricate nature of chemical scenarios.

To mitigate these issues, various sampling methods have been employed to achieve data balance and curtail the risk of overfitting[16-18]. Conventional methods such as oversampling and under-sampling[19,20] directly manipulate the dataset sizes to address class imbalances but may lead to information loss or induce overfitting. Stratified sampling[21], although maintaining the proportions among classes, does not necessarily enhance dataset diversity. More sophisticated approaches offer refined solutions including genetic algorithms (GA) that generate diversified samples through crossover and mutation[22], or the Most Dissimilar method, which constructs a maximally diverse sample set based on molecular fingerprints[23,24]. Active learning[25-28] optimizes models effectively by iteratively selecting samples with high informational value, while surrogate model-based[29] incremental sampling progressively improves dataset by minimizing mean squared error (MSE). Although these advanced methods are effective, their iterative processes significantly increase computational costs and require hyperparameter tuning to avoid local optima. Moreover, utilizing pretrained large-scale models[30-32], enriched with extensive chemical and materials science knowledge, enables rapid adaptation through domain-specific fine-tuning using small datasets. However, the substantial computational resources and time required for training and evaluating pretrained models limit their practical deployment in dynamic chemical scenarios that demand rapid adaptability and cost efficiency.

In this paper, we introduce an effective strategy to address the challenge, the farthest point sampling (FPS) in property designated chemical feature space (FPS-PDCFS). FPS is a sampling method tailored for high-dimensional spaces. It operates on the assumption that the distribution of structures in diversity spaces is important, which provides the prior information to guide its sampling. The method selects samples in the feature space that are furthest apart, effectively capturing the essence of the entire dataset with a minimal number of samples. Although FPS has been utilized in progressive image sampling[33], point cloud networks[34], and feature selection[35], its application chemical dataset for ML has not been explored. In this study, we employed FPS-PDCFS to partition databases with target physicochemical properties such as standard boiling points and enthalpy of vaporization (HVAP). In our study, ML models, including artificial neural networks (ANNs), support vector machines (SVMs), and random forests (RFs), utilizing FPS consistently outperformed those based on random sampling (RS), showcasing enhanced predictive accuracy, stability and markedly reduced overfitting, and a robust resilience to limited sample sizes. Our findings underscore that within a property-designated high-dimensional feature space, FPS adeptly identifies unique characteristics and preserves the diversity of the training set. Such an approach substantially elevates the efficacy of ML models in predicting molecular properties by ensuring a holistic and balanced portrayal of the chemical feature landscape. Consequently, FPS-PDCFS positions itself as a versatile sampling tool, enhancing both the quality of small, skewed chemical datasets and the predictive capability of chemical ML models.

MATERIALS AND METHODS

Physicochemical database and ML model

The thermodynamic and physical property datasets we used in the study were obtained from online databases such as the Yaws' handbook[36] and PubChem[37]. These datasets encompass structurally diverse compounds, including hydrocarbons, halogenated hydrocarbons, aromatic heterocycles, *etc.*, covering physicochemical properties such as boiling point, HVAP, and critical properties. These properties were utilized to compare the performance of FPS and RS. Additional details regarding the datasets can be found in Supplementary Section 1.

Several widely recognized ML models were selected for this study, including ANNs, SVMs, kernel ridge regression (KRR), k-nearest neighbors (KNNs), RFs and extreme gradient boosting (XGB). These methods are commonly employed in regression and classification tasks to predict material or molecular properties. The hyperparameters for training these models were optimized using Bayesian optimizing (BO), with all comparisons conducted using same training parameters; detailed parameters are listed in the Supplementary Section 2. Molecular features (descriptors) were computed using RDKit[38] and AlvaDesc[39]. Based on our previous work[40], we selected a set of interpretable molecular descriptors as input features, including structural descriptors (such as the number of hydrogen bond donors/acceptors) and topological indices. These descriptors have demonstrated stable and high predictive accuracy in tasks involving physicochemical property prediction, such as boiling point. Detailed descriptor settings used as model inputs are provided in Supplementary Section 3.

Sampling method and benchmark

RS is a straightforward process involving the selection of a random subset from the dataset. On the other hand, FPS necessitates a more structured approach, requiring execution within a pre-defined chemical feature space. We implemented FPS in a feature space constructed from molecular descriptors and subsequently used the resulting samples in ML models. This choice was informed by our discovery that FPS could augment model performance when executed within a feature space correlated with the target properties, a phenomenon we will elucidate in subsequent sections. In Figure 1A, FPS is demonstrated to sample within a chemical feature space articulated by molecular descriptors. Diverse points within this space represent various categories of molecules from the chemical database, each occupying distinct positions and exhibiting varying densities. FPS is adept at uniformly sampling across this space, ensuring a representative collection of molecules from each category. As depicted in Figure 1B, the process of FPS sampling encompasses the following steps:

(1) Randomly select an initial point from the dataset.
(2) Compute the distances from all other points to this initial point, select the farthest point as the second sampled point.
(3) For each unsampled point p_i, compute its distance $D(p_i, S)$ to the set of already sampled points S, and select the next point according to:

$$p_{next} = \underset{p_i \in U}{\operatorname{argmax}}\, D(p_i, S), \text{ where } D(p_i, S) = \min_{s \in S} \|p_i - s\|$$

Where U denotes the set of all unsampled points.
(4) Repeat step 3 until the sampling reaches the desired size. This iterative sampling procedure ensures that points in the resulting set are as distant from each other as possible.

Detailed formulations of distance metrics and pseudocode for sampling methods are provided in Supplementary Section 4, while comprehensive discussions comparing FPS with alternative sampling

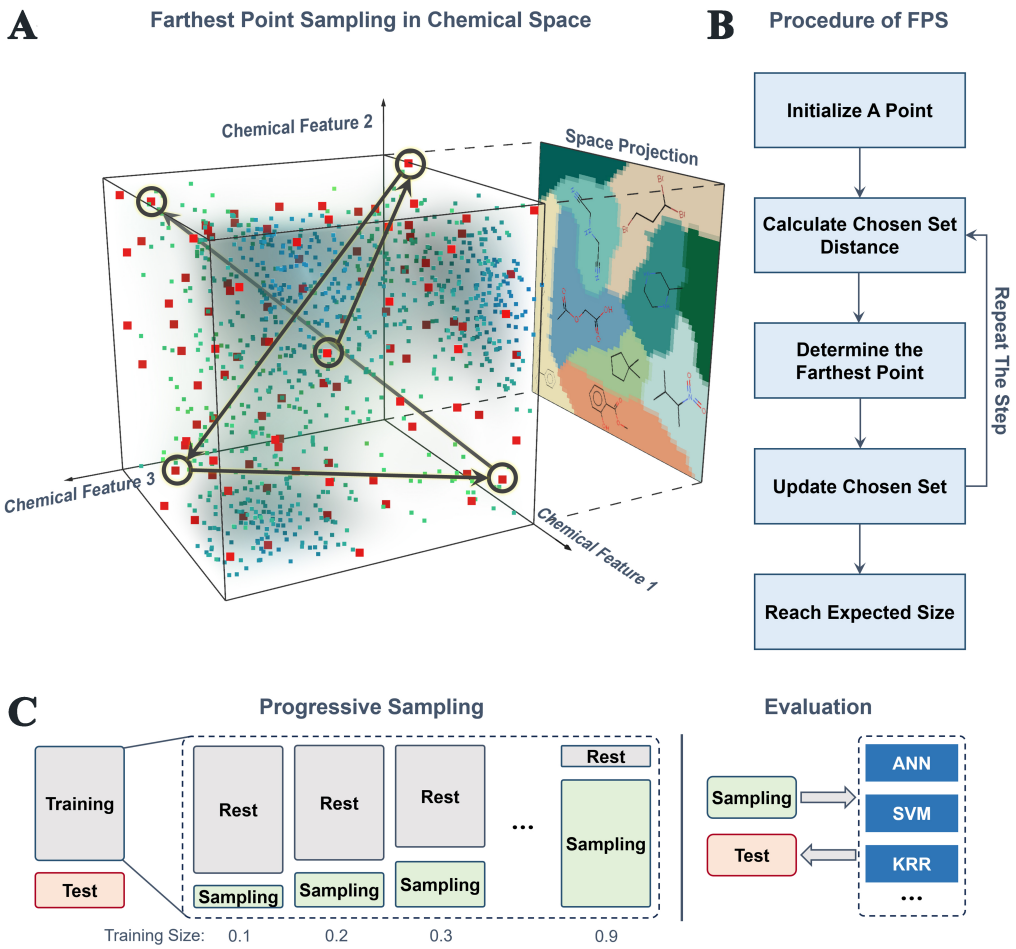

Figure 1. (A) A graphic illustration of the FPS in chemical space; (B) A Flow chart of FPS Procedure; (C) Sampling process used in this study. The initial dataset was partitioned into training and test sets randomly. The training set was then further bifurcated into a "sampling set" and a "rest set" of different ratios by sampling using different strategies. Various ML models were next trained on the sampling set and subsequently validated on the test set. FPS: Farthest point sampling; ML: machine learning.

strategies - such as Most Dissimilar and Cluster-based sampling - are presented in Supplementary Section 5.

To systematically evaluate the impact of various sampling strategies on model performance, we employed a combined approach involving five-fold cross-validation (5-fold CV) and multiple independent trials. As illustrated in Figure 1C, the initial dataset was first partitioned into a test set and a training set via cross-validation, with the test set remaining independent from any sampling procedure to ensure objective and comparable results. The training set was further subdivided by different sampling methods into sampled subsets and residual subsets, with sampling proportions progressively increasing from 0.1 to 1.0 in increments of 0.1. Multiple independent trials were conducted at each sampling proportion to enhance statistical robustness, with the MSE defined as the average of the MSE values obtained from these independent trials under the 5-fold CV setting. Only the sampled subset of data was used to train the models.

In the "Results and Discussion" section, we demonstrate that models trained using the FPS-PDCFS method on specific smaller subsets outperform those trained on the entire dataset, highlighting the notable imbalance present within specific chemical datasets.

RESULTS AND DISCUSSION

Following the procedure shown in Figure 1, we began a systematic evaluation of the FPS strategy's effectiveness, comparing it directly with the RS. ANN was chosen as the benchmark model owing to its broad usage and maturity.

Figure 2 presents a comparative analysis of the ANN model's performance, contrasting the use of FPS and RS across various training sizes within the boiling point dataset. A detailed inspection of Figure 2A and B reveals a pronounced disparity in MSE between the training and test sets under RS, particularly at smaller training sizes. This disparity signals significant overfitting, a common issue in models trained on small and imbalanced datasets[41]. With increasing training size, this overfitting tendency gradually diminishes.

In contrast, the FPS-enhanced ANN model demonstrates a remarkable alignment of MSEs for both training and test sets across a majority of training sizes, as depicted in Figure 2B. This alignment indicates a consistent reduction in overfitting, even with smaller training sets. It is noteworthy that when the training size is 0.6, both MSE and ΔMSE of the test set reach relatively low values. This indicates that the FPS-enhanced ANN model not only achieves comparable accuracy to the RS model, which utilizes the full database, but also exhibits lower risks of overfitting and stronger generalization performance. As the training size grows, the performance of FPS progressively converges with that of RS, which is expected as the sample composition of the training sets generated by FPS and RS become increasingly similar.

Moreover, it is notable that test-set MSE does not always decrease monotonically with increasing training size for either sampling strategy. This non-monotonic behavior is consistent with empirical observations in small, noisy datasets, where learning curves often display local fluctuations due to sampling variance and model-data interactions[42]. In our case, the addition of noisy or unrepresentative samples at certain training sizes can temporarily hinder generalization, causing brief performance dips that resolve as the dataset continues to grow.

Furthermore, as shown in Figure 2B, when the training sizes are 0.2, 0.3, and 0.5, the statistical results reveal that the mean ΔMSE is negative, indicating lower test-set errors than training-set errors, which is an atypical phenomenon in conventional ML scenarios. The common assumption that training errors are lower than test errors relies on the training and test sets being drawn from similar distributions. In our case, however, FPS selects training samples that differ significantly in distribution from the randomly partitioned test sets, particularly under low-data regimes. This can lead to negative ΔMSE values, where the model performs better on the test set than on the training set. A likely explanation is that FPS prioritizes diversity in chemical space, often selecting structurally distinct or outlier compounds. These are harder to fit accurately, increasing the training error. However, their broader representativeness leads to better generalization, and thus lower test error, compared to models trained on randomly sampled data.

It is noteworthy that the MSEs for both training and test sets remain closely matched even at a minimal training size of 0.1 and 0.2, with the test set consistently demonstrating significantly lower MSE, indicative of FPS's effectiveness in preventing underfitting in ANN models at these small training sizes. However, as discussed in Supplementary Section 5.2, we observed possible underfitting in both RF and XGB models at training sizes of 0.1, suggesting that certain models, may indeed experience underfitting with very small FPS-selected training subsets; therefore, the optimal training size may vary depending on the specific ML model employed. Overall, this highlights that FPS generally enables robust modeling with minimal data, significantly lowering the data requirements and associated costs in ML. Given the challenges and high expenses of chemical data collection, this efficiency is particularly valuable for applying ML in chemistry

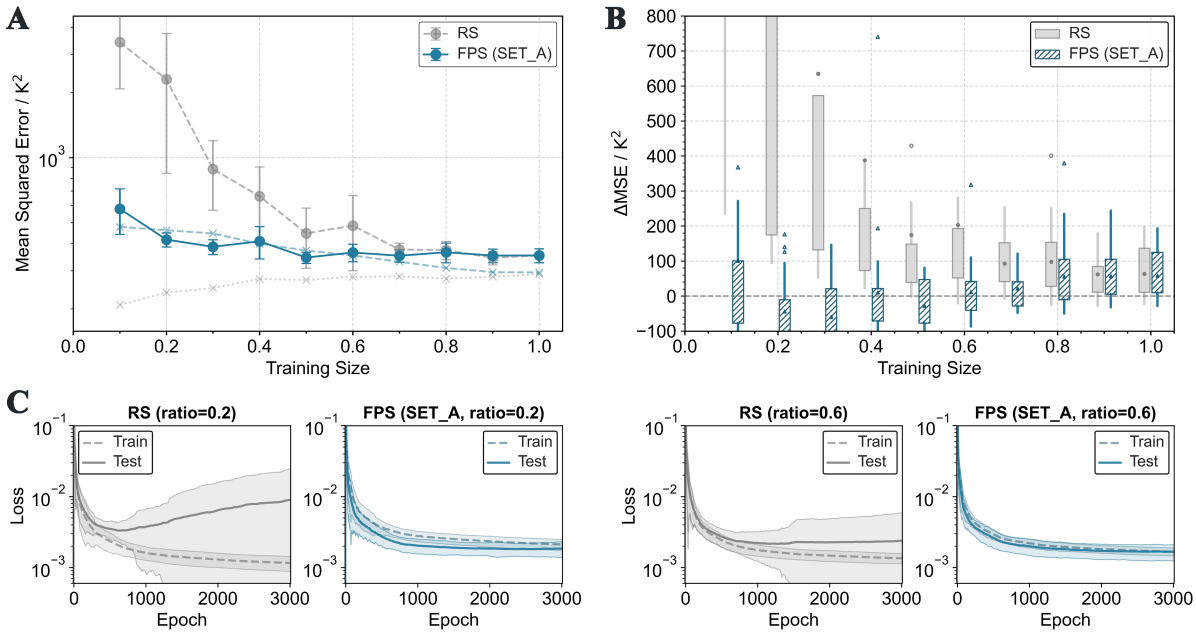

Figure 2. (A) MSE for training and test sets of ANN models built using the boiling point dataset under FPS (blue) and RS (gray) methods across different training sizes; (B) Distribution of the ΔMSE, i.e., difference between test set MSE and training set MSE, across varying training sizes. In the box plot, the box and whiskers represent the first and third quartiles, and 1.5 times the IQR, respectively. The dark dots represent the mean of ΔMSE. FPS exhibits lower ΔMSE values between training sizes of 0.2 to 0.7, indicating lower risks of overfitting; (C) Loss curves for the training and test sets of ANN models under FPS and RS at training sizes of 0.2 and 0.6. For the model by RS, larger shaded areas indicate higher variability, and higher losses in test sets suggest a greater risk of overfitting. MSE: Mean squared error; ANN: artificial neural network; FPS: farthest point sampling; RS: random sampling; IQR: interquartile range.

and materials science, thereby paving the way for more accurate and cost-effective research methodologies.

The enhanced stability, robustness and reduced overfitting attributed to the FPS strategy are further supported by the learning curve shown in Figure 2C (See Supplementary Figure 3, for comprehensive learning curves). It becomes apparent that FPS leads to greater stability and reduced overfitting in the learning process compared to RS, particularly at a training size smaller than 0.5.

Further analyses assessing FPS effectiveness across multiple ML models are detailed in Supplementary Figures 4 and 5.

Building on the earlier discussion, the successful application of FPS necessitates an appropriate feature space. To delve deeper into this aspect, we conducted a thorough evaluation of FPS's performance across a range of feature spaces. Specifically, we scrutinized three different feature space configurations, each characterized by a unique set of features:

Set A - Interpretable Descriptors, based on the physical principles linking molecular features to their boiling points, which are also the descriptors used in our ANN model, aiming to reflect the fundamental physical characteristics of the molecules.

Set B - Regression-Derived Descriptors, derived from regression analysis, specifically focusing on the correlations between descriptors and the target property, which may not be immediately apparent but are statistically significant within the dataset.

Set C - Randomly-Selected Descriptors, a random selection of descriptors from a comprehensive pool of 1600 descriptors available in alvaDesc, offering a control set with respect to Sets A and B.

The types of descriptors utilized in each configuration are summarized in Table 1. A more detailed exploration of all the descriptors included in this study is provided in Supplementary Tables 2-4.

In evaluating the effect of the feature spaces, different feature sets were employed only during the FPS, while the ANN model hyperparameters remained unchanged, still using Set A as the input.

Figure 3 illustrates model performance across different training sizes. Notably, for Set B, there is a substantial increase in MSE when the training size is only 0.1. This high MSE can be attributed to the discrepancy between the sampled Set B and the descriptors used as model inputs. With limited training data, this mismatch introduces considerable bias, which exacerbates the MSE. As the training size increases beyond 0.1, Set B's MSE decreases significantly, and aligning with the performance of Set A. Between training sizes of 0.2 to 0.7, Sets A and B demonstrate consistently low overfitting compared to RS and FPS-C. Specifically, the lowest levels of overfitting are observed from 0.2 to 0.5. However, in the interval from 0.6 to 0.7, both Sets A and B achieve an optimal balance between MSE and ΔMSE, resulting in best model performance. As the training size increases beyond 0.7, the performances of Sets A and B begin to converge with those of Sets C and RS. This convergence indicates a diminishing distinction between the sampling strategies, leading to more uniform performances across all sets. At other training sizes, both Set C and RS consistently demonstrate higher ΔMSE and MSE, underperforming compared to Sets A and B. This suggests that a randomly selected descriptor set (Set C) might lead to erroneous prior guidance, resulting in poorer performance of FPS-C relative to FPS-A and FPS-B.

At a training size of 0.3, an inflection point occurs, as demonstrated in Figure 3C. Below this training size, models using RS and FPS-C are prone to early overfitting and exhibit high variability. Figure 3C depicts the learning curves for different sampling methods at a 0.3 training size, where the RS model, in particular, shows a significant disparity between training and testing losses after just over ten training epochs. Conversely, training under FPS with Sets A and B maintains stable and low overfitting throughout all training steps, showcasing lower variability and consistent performance.

These observations underscore the nuanced capabilities of FPS in enhancing ANN model performance, particularly when applied in feature spaces that align well with target properties. Other ML models utilizing FPS across various feature spaces, along with comparisons to alternative sampling methods, are detailed in Supplementary Figure 6.

Figure 4 illustrates the MSE across different sampling methods and ML models, as represented by three heatmaps corresponding to specific training sizes: 0.2, 0.6, and 0.8. In the heatmap, the upper triangle displays the Test MSE of different models, while the lower triangle shows the Train MSE. The redder color of triangle indicates higher MSE values, whereas bluer color denotes lower MSE values. At a training size of 0.2, not all models under FPS demonstrate a Test MSE lower than RS, suggesting that the negative ΔMSE does not occur in every model. As the training size increases to 0.6, models under FPS-A and FPS-B exhibit bluer Test MSE, indicating superior performance over RS, while models under FPS-C mostly do not exhibit improvement and may perform worse than RS. By a training size of 0.8, the performance of FPS models approaches that of RS, but still displays a trend where FPS-A and FPS-B outperform RS, and FPS-C underperforms relative to RS.

Table 1. The descriptor type of chosen descriptors for FPS feature space

Descriptor type	Set A[a]	Set B[b]
Molecular properties	√	√
Functional group	√	
Constitutional indices	√	√
Topological indices	√	√
Structural parameter	√	
Information indices		√
2D autocorrelations		√
Burden eigenvalues		√
P_VSA-like descriptors		√
Pharmacophore descriptors		√
Charge descriptors	√	√
Drug-like indices		√
MDE descriptors		√
Walk and path counts		√

[a]Set A: Interpretable descriptors set. [b]Set B: Regression-derived descriptors. FPS: Farthest point sampling; MDE: molecular distance edge descriptors.

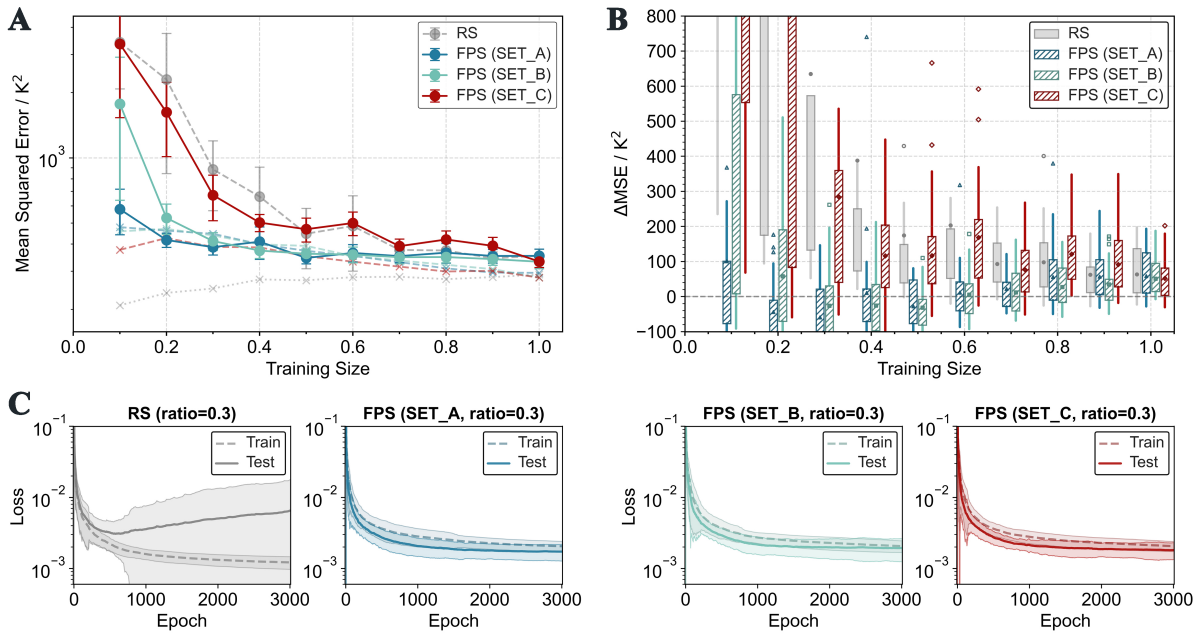

Figure 3. (A) Training and test set MSE of ANN models under FPS in interpretable space (FPS-A, blue), regression space (FPS-B, cyan), casually selected space (FPS-C, red), and RS (gray) across different training sizes; (B) Distribution of ΔMSE for different sampling methods across various training sizes, with FPS-A and FPS-B exhibiting negative ΔMSE between training sizes of 0.2 and 0.5; (C) Loss curves for training and test sets under different sampling methods at a training size of 0.3. Loss curves for FPS-A and FPS-B remain stable with increasing training epochs, indicating lower variability, with A and B showing better performance than C. MSE: Mean squared error; ANN: artificial neural network; FPS: farthest point sampling; RS: random sampling.

The heatmap further reveals that model performance is intricately associated with the model type; ANN, KRR, and KNN exhibit smaller differences between Test and Train MSE, indicating a lower risk of overfitting. In contrast, RF and XGB show a tendency to overfit across various training sizes and sampling methods. Notably, at smaller training sizes, substantial overfitting is indicated by prominent red upper

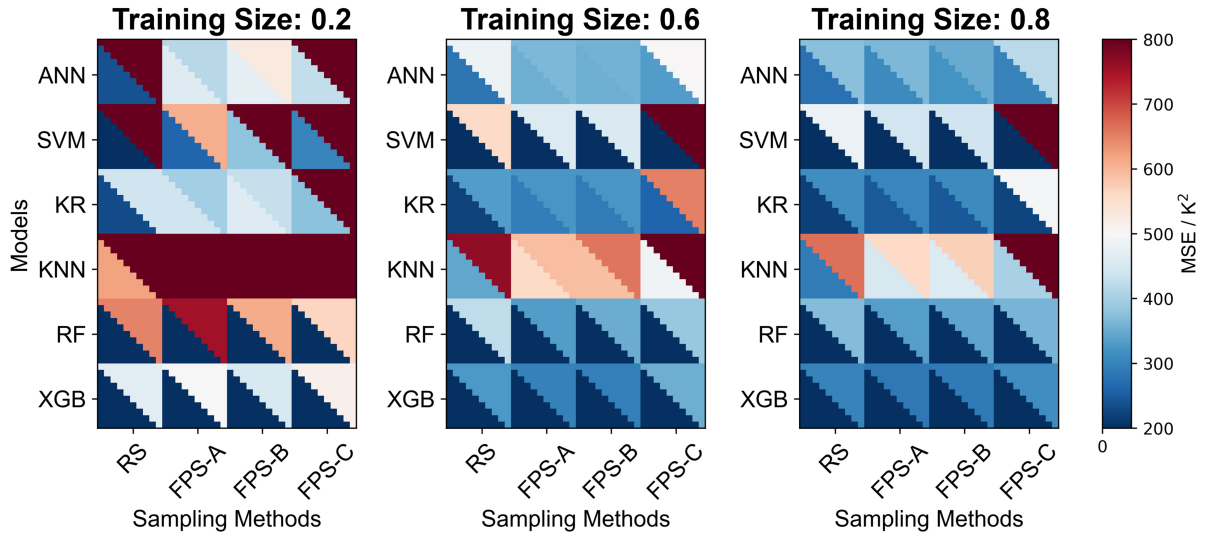

Figure 4. The MSE of training and test set under various sampling methods by different ML models at training sizes of 0.2, 0.6 and 0.8, with the upper triangle legend representing test set MSE, and the lower triangle legend representing training set MSE. The MSE of the training and test sets for FPS with Set A and B displays more similar color patterns across various ML models, indicating that less overfitting is achieved in these models by implementing the FPS strategy in the designated feature spaces. MSE: Mean squared error; ML: machine learning; FPS: farthest point sampling.

triangles for both RS and FPS-C across the models. At optimal training sizes of 0.6, models such as ANN, SVM, and RF under FPS-A and FPS-B sampling achieve a balance between prediction accuracy (MSE) and overfitting, leading to optimal performance. However, XGB shows limited sensitivity to FPS, as the partitioning strategy employed in tree-based models means that changes in the local distribution (i.e., adding or removing samples within a region) do not significantly affect the tree node splitting decisions[43,44]. Models based on FPS exhibit only marginal improvements in Test MSE compared to RS, possibly due to this characteristic of the ensemble model.

Additionally, we further explored FPS performance using other widely-used molecular representations such as Extended Connectivity Fingerprints (ECFP) and MACCS keys, alternative sampling strategies including Most Dissimilar and Cluster-based sampling [Supplementary Figures 8 and 9], and the effect of different distance metrics [Supplementary Figures 10 and 11]. Additional comparisons across these diverse feature spaces are provided in Supplementary Figures 12 and 13.

Moreover, we have expanded our study to encompass a variety of physicochemical property databases, including the HVAP, along with critical properties such as critical temperature (CT) and critical volume (CV), to further assess the effectiveness of FPS. The performance of the ANN model under FPS and RS across these four types of physicochemical property databases is elaborated in Figure 5. It is noteworthy that in the HVAP dataset, we observe a trend similar to that in the BP dataset, where FPS shows a lower test MSE compared to RS. Between training sizes of 0.3 and 0.5, the overlap of FPS's training and test set box plots is maximized, indicating that the model has the smallest ΔMSE and lower overfitting risk at these training sizes. In the critical pressure (CP) and CT databases, due to the high measurement errors of critical properties, the models show a wide range of MSE errors, yet FPS's error range remains smaller than that of RS. In the CT dataset shown in Figure 5, both FPS and RS exhibit large fluctuations in MSE mean, sometimes even exceeding the first and third quartiles, which are contributed by a particular test set in cross-validation. Under this test set, both FPS and RS struggle to deliver good predictive performance. In

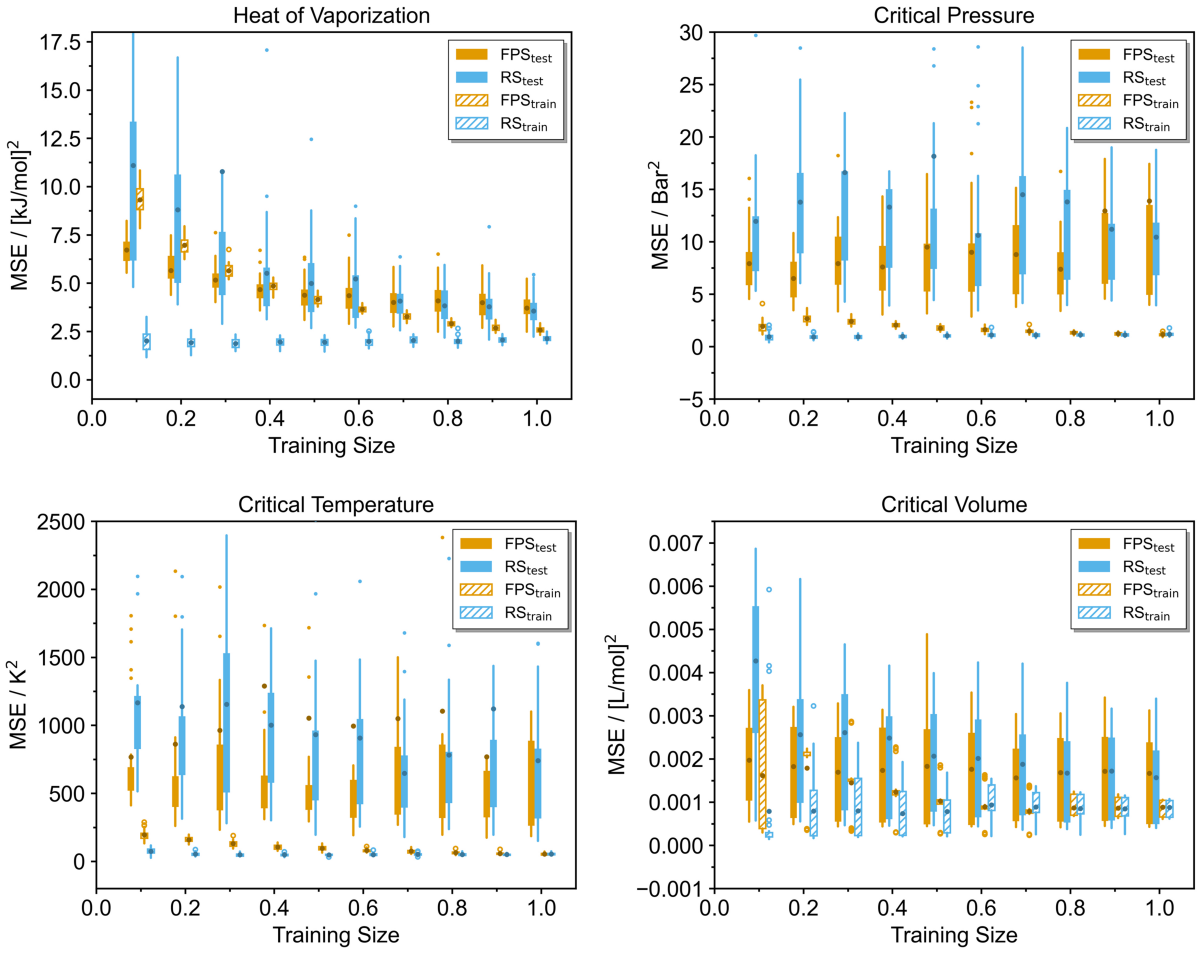

Figure 5. Comparison of MSE for training and test sets in the ANN model using FPS (yellow) and RS (blue) across various training sizes on physicochemical datasets. These datasets include the HVAP, CT, CV, and CP. MSE: Mean squared error; ANN: artificial neural network; FPS: farthest point sampling; RS: random sampling; HVAP: enthalpy of vaporization; CT: critical temperature; CV: critical volume; CP: critical pressure.

the CV database, as the most stable source of critical data, it is demonstrated that with increasing training sizes, the box plot of FPS is smaller than that of RS and gradually converges towards it. Additionally, we have evaluated the performance of other ML models on these physicochemical property databases (as detailed in Supplementary Figures 14-17). These models uniformly demonstrate that FPS offers enhanced robustness, predictive accuracy, and consistent reduction in overfitting.

To elucidate how the FPS method effectively samples the chemical space and extracts key information from chemical datasets, we constructed t-SNE[45] visualizations to compare the sampling distributions of FPS-A and RS, as illustrated in Figure 6. In these plots, gray points represent the entire boiling point dataset, while red and blue points highlight the sampled subsets selected by FPS and RS, respectively. Compared to the relatively uniform sampling of RS, FPS preferentially selects points from sparsely populated and structurally diverse regions, while limiting samples from densely clustered regions. This selective strategy significantly enhances the structural diversity of sampled data, thus improving predictive model performance. Additionally, Figure 6F further demonstrates how FPS and RS differently affect the frequency distribution of samples across k-means clusters. The distinct sampling patterns underscore the superiority of FPS in handling complex, high-dimensional datasets.

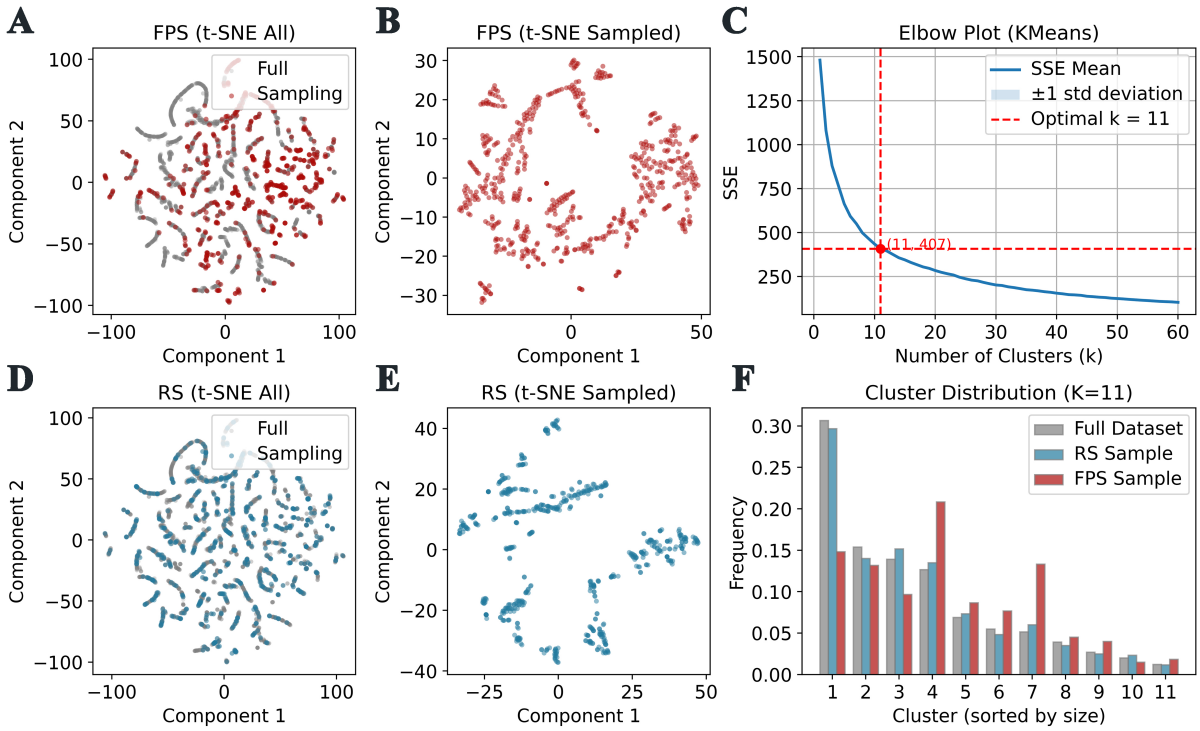

Figure 6. (A) and (B) display the t-SNE visualization of the entire dataset and the sampled subset obtained by FPS, respectively, demonstrating the distributional change after sampling; (D) and (E) correspondingly show the results of RS. Compared to RS, FPS yields a more dispersed distribution, thus enhancing diversity in the sampled feature space; (C) illustrates the k-means clustering of the full dataset, determined by the elbow method, while (F) compares the effects of RS and FPS on cluster frequency distributions. t-SNE: t-distributed Stochastic Neighbor Embedding; FPS: farthest point sampling; RS: random sampling.

CONCLUSIONS

In this study, we have applied the FPS method within a designated chemical feature space to tackle the challenges associated with small, imbalanced datasets prevalent in chemistry and material science domains. Extensive testing across various ML models and physicochemical datasets demonstrated that FPS significantly enhance training-data diversity, thereby improving predictive accuracy and reduced overfitting in nearly all tested models. A critical insight is that the effectiveness of FPS is tethered to its application within an appropriate chemical feature space aligned with the target property, and it works particularly well when applied to ANN with proper pre-knowledge descriptors tethered to the target properties.

Moreover, although it is widely recognized that uniformly sampling chemical data within a property-designated feature space is crucial for model performance, practical strategies to achieve this have not been extensively explored. A notable limitation of FPS arises when clearly defined, property-designated feature spaces are not readily available or ambiguous. In such situation, FPS may be integrated into deep learning frameworks by initially trained using such as convolutional neural network, to extract latent-space embeddings. FPS can then operate within these embeddings for targeted sampling of significant chemical data. This study is performing in our group and will be reported in a due course.

Consequently, FPS proves to be a versatile, pragmatic strategy, capable of improving both the quality of small, imbalanced chemical datasets and the predictive capability of chemical ML models, paving the way for more accurate and reliable structure-property predictions in chemistry and material science, achieved with minima datasets and reduced cost.

DECLARATIONS

Acknowledgments

VThis work was supported by Shanghai Boronmatrix Advanced Materials Technology Co. Ltd. and the Fundamental Research Funds for the Central Universities.

Authors' contributions

Made contributions to conception and design of the study, performed data analysis and interpretation: Liu, Y.

Provided advisory guidance and supervised the study: Yu, X.

Provided technical and material support: Wang, L.; Zhu, W.

Availability of data and materials

The molecular dataset used in this study, including SMILES strings and their corresponding physicochemical properties, as well as the full farthest-point sampling (FPS) code, are now publicly available in our GitHub repository (https://github.com/yuxi-TJU/Farthest-Point-Sampling-in-Chemical-Feature-Space).

In this study, we used several software tools for data processing and analysis. Below are the details of these tools:

• RDKit: RDKit is an open-source cheminformatics and machine learning tool used for handling chemical data. We utilized RDKit for processing molecular data and calculating descriptors of molecules. The latest version of RDKit can be freely obtained from its official website or GitHub page: https://www.rdkit.org/.

• scikit-learn: scikit-learn is a popular open-source machine learning library for the Python programming language. We used scikit-learn for statistical analysis of data and machine learning modeling. The latest version of scikit-learn can be found on its official website: https://scikit-learn.org/.

Financial support and sponsorship

This work was supported by ISF-NSFC Joint Scientific Research Program (22361142833), National Natural Science Foundation of China (NSFC) under Grant No. 21973069 and U21A6002, Open Project of the State Key Laboratory of Supramolecular Structure and Materials (SKLSSM2024035), Open Project of the Key Laboratory of Resource Chemistry, Ministry of Education (2024-002003) and "the Fundamental Research Funds for the Central Universities".

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Jordan, M. I.; Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **2015**, *349*, 255-60. DOI PubMed
2. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547-55. DOI PubMed
3. Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; et al. Combining machine learning and computational chemistry for predictive insights

into chemical systems. *Chem. Rev.* **2021**, *121*, 9816-72. DOI PubMed PMC
4. Shi, X.; Zhang, G.; Lu, Y.; Pang, H. Applications of machine learning in electrochemistry. *Renewables* **2023**, *1*, 668-93. DOI
5. Jiang, Y.; Yang, Z.; Guo, J.; et al. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **2021**, *12*, 5950. DOI PubMed PMC
6. Chong, Y.; Huo, Y.; Jiang, S.; et al. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120*, e2220789120. DOI PubMed PMC
7. Wang, X.; Jiang, S.; Hu, W.; et al. Quantitatively determining surface-adsorbate properties from vibrational spectroscopy with interpretable machine learning. *J. Am. Chem. Soc.* **2022**, *144*, 16069-76. DOI PubMed
8. Ren, H.; Zhang, Q.; Wang, Z.; et al. Machine learning recognition of protein secondary structures based on two-dimensional spectroscopic descriptors. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, e2202713119. DOI PubMed PMC
9. Chen, A.; Zhang, X.; Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2020**, *2*, 553-76. DOI
10. Sun, Z.; Yin, H.; Liu, K.; et al. Machine learning accelerated calculation and design of electrocatalysts for CO_2 reduction. *SmartMat* **2022**, *3*, 68-83. DOI
11. Lin, M.; Liu, X.; Xiang, Y.; et al. Unravelling the fast alkali-ion dynamics in paramagnetic battery materials combined with NMR and deep-potential molecular dynamics simulation. *Angew. Chem. Int. Ed. Engl.* **2021**, *60*, 12547-53. DOI PubMed
12. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **2018**, *361*, 360-5. DOI PubMed
13. Wang, A. Y.; Murdock, R. J.; Kauwe, S. K.; et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **2020**, *32*, 4954-65. DOI
14. Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. *npj. Comput. Mater.* **2023**, *9*, 1000. DOI
15. Dou, B.; Zhu, Z.; Merkurjev, E.; et al. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **2023**, *123*, 8736-80. DOI PubMed PMC
16. Guo, H.; Li, Y.; Shang, J.; Gu, M.; Huang, Y.; Gong, B. Learning from class-imbalanced data: review of methods and applications. *Exp. Syst. Appl.* **2017**, *73*, 220-39. DOI
17. Xu, X.; Liang, T.; Zhu, J.; Zheng, D.; Sun, T. Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing* **2019**, *328*, 5-15. DOI
18. Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol.* **1999**, *6*, 447-57. DOI PubMed
19. Pereira, T.; Abbasi, M.; Oliveira, J. L.; Ribeiro, B.; Arrais, J. Optimizing blood-brain barrier permeation through deep reinforcement learning for de novo drug design. *Bioinformatics* **2021**, *37*, i84-92. DOI PubMed PMC
20. Lu, T.; Li, H.; Li, M.; Wang, S.; Lu, W. Predicting experimental formability of hybrid organic-inorganic perovskites via imbalanced learning. *J. Phys. Chem. Lett.* **2022**, *13*, 3032-8. DOI PubMed
21. Mazouin, B.; Schöpfer, A. A.; von Lilienfeld, O. A. Selected machine learning of HOMO-LUMO gaps with improved data-efficiency. *Mater. Adv.* **2022**, *3*, 8306-16. DOI PubMed PMC
22. Akdemir, D.; Sanchez, J. I.; Jannink, J. L. Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* **2015**, *47*, 38. DOI PubMed PMC
23. Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: Theory and characteristics. *J. Cheminform.* **2021**, *13*, 32. DOI PubMed PMC
24. Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. *J. Cheminform.* **2021**, *13*, 33. DOI PubMed PMC
25. Ng, W. W. Y.; Yeung, D. S.; Cloete, I. Input sample selection for RBF neural network classification problems using sensitivity measure. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, Washington, USA. Oct 08, 2023. IEEE; 2023. pp. 2593-8. DOI
26. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733. DOI PubMed
27. Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate *ab initio* molecular dynamics. *Int. J. Quantum. Chem.* **2015**, *115*, 1074-83. DOI
28. Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924-35. DOI PubMed PMC
29. Bergström, D.; Tiger, M.; Heintz, F. Bayesian optimization for selecting training and valida-tion data for supervised machine learning. In *Proceedings of the 31st Annual Workshop of the Swedish Artificial Intelligence Society (SAIS 2019)*, Umeå, Sweden. Jun 18-19, 2019. https://www.ida.liu.se/divisions/aiics/publications/SAIS-2019-Bayesian-Optimization-Selecting.pdf. (accessed 11 Jun 2025)
30. Vaswani, A.; Shazeer, N.; Parmar, N. et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA. Curran Associates Inc.; 2017. pp. 6000-10. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. (accessed 11 Jun 2025)
31. Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **2022**, *4*, 1256-64. DOI
32. Lu, S.; Gao, Z.; He, D.; Zhang, L.; Ke, G. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-

Mol. *Nat. Commun.* **2024**, *15*, 7104. DOI PubMed PMC

33. Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y. Y. The farthest point strategy for progressive image sampling. *IEEE. Trans. Image. Process.* **1997**, *6*, 1305-15. DOI PubMed
34. Charles, R. Q.; Su, H.; Kaichun, M.; Guibas, L. J. PointNet: deep learning on point sets for 3D classification and segmentation. *2017. IEEE. Conference. on. Computer. Vision. and. Pattern. Recognition. (CVPR)*. 2017. pp. 77-85. DOI
35. Cersonsky, R. K.; Helfrecht, B. A.; Engel, E. A.; Kliavinek, S.; Ceriotti, M. Improving sample and feature selection with principal covariates regression. *Mach. Learn. Sci. Technol.* **2021**, *2*, 035038. DOI
36. Yaws, C. L. Yaws' critical property data for chemical engineers and chemists. Knovel; 2012. http://app.knovel.com/hotlink/toc/id:kpYCPDCECD/yaws-critical-property/yaws-critical-property. (accessed 11 Jun 2025).
37. PubChem. National Center for Biotechnology Information. https://pubchem.ncbi.nlm.nih.gov/. (accessed 11 Jun 2025).
38. RDKit: Open-source cheminformatics software. https://www.rdkit.org. (accessed 11 Jun 2025).
39. Mauri, A. alvaDesc: a tool to calculate and analyze molecular descriptors and fingerprints. In: Roy K, editor. Ecotoxicological QSARs. New York: Springer US; 2020. pp. 801-20. DOI
40. Liu, Y.; Li, K.; Huang, J.; Yu, X.; Hu, W. Accurate prediction of the boiling point of organic molecules by multi-component heterogeneous learning model. *Acta. Chim. Sin.* **2022**, *80*, 714-23. DOI
41. Bishop, C. M. Pattern recognition and machine learning. Springer: New York, NY; 2006. https://link.springer.com/book/9780387310732. (accessed 11 Jun 2025).
42. Viering, T.; Loog, M. The shape of learning curves: a review. *IEEE. Trans. Pattern. Anal. Mach. Intell.* **2023**, *45*, 7799-819. DOI PubMed
43. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery; 2016. pp. 785-94. DOI
44. He, H.; Garcia, E. A. Learning from Imbalanced Data. *IEEE. Trans. Knowl. Data. Eng.* **2009**, *21*, 1263-84. DOI
45. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.2008*, 2579-605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf. (accessed 11 Jun 2025)