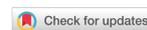


Research Article

Open Access



A new simple and efficient molecular descriptor for the fast and accurate prediction of log P

Xiaojian Zeng[#], Xin Ye[#], Donghua Liu, Ningyi Cui, Xiaopeng Li, Yufan Bao, Yecheng Zhou^{*}

Guangzhou Key Laboratory of Flexible Electronic Materials and Wearable Devices, School of Materials Science & Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China.

[#]Authors contributed equally.

^{*}**Correspondence to:** Prof. Yecheng Zhou, Guangzhou Key Laboratory of Flexible Electronic Materials and Wearable Devices, School of Materials Science & Engineering, Sun Yat-sen University, No.132 East Outer Ring Road, Guangzhou 510006, Guangdong, China. E-mail: zhouych29@mail.sysu.edu.cn

How to cite this article: Zeng, X.; Ye, X.; Liu, D.; Cui, N.; Li, X.; Bao, Y.; Zhou, Y. A new simple and efficient molecular descriptor for the fast and accurate prediction of log P. *J. Mater. Inf.* 2025, 5, 4. <https://dx.doi.org/10.20517/jmi.2024.61>

Received: 21 Oct 2024 **First Decision:** 26 Nov 2024 **Revised:** 14 Dec 2024 **Accepted:** 26 Dec 2024 **Published:** 16 Jan 2025

Academic Editors: Lingyan Feng, Xingjun Liu, Rika Kobayashi **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

The partition coefficient (log P) is a critical parameter that measures the balance between hydrophilicity and lipophilicity of molecules, playing a key role in molecular material design and drug development. Developing accurate, efficient, and computationally simple models for log P prediction is essential for advancing drug discovery and materials science. In this study, we introduce the optimized 3D molecular representation of structures based on electron diffraction descriptor (opt3DM) into machine learning (ML) frameworks, achieving highly accurate log P predictions. By fine-tuning key parameters, the scale factor (s_l) and descriptor dimension (N_s), we identified the optimal values of $s_l = 0.5$ and $N_s = 500$. Among various ML algorithms tested, automatic relevance determination (ARD) regression, Ridge regression, and Bayesian Ridge regression demonstrated superior predictive performance. These optimized models outperformed the OPEn structure-activity/property relationship app (OPERA) model on the M-dataset and also delivered competitive results in the SAMPL6 and SAMPL9 challenges. Our findings not only establish a robust, fast, and precise approach for log P prediction, but also highlight the potential of opt3DM as a powerful tool for molecular representation. This work lays a foundation for broader applications in molecular material design and drug development.

Keywords: Molecular descriptor, machine learning, partition coefficient, optimized 3D MoRSE descriptor, SAMPL6, SAMPL9



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



INTRODUCTION

The log P value, which denotes the base-10 logarithm of the partition coefficient between two immiscible phases, is a critical physicochemical property that provides essential information about a molecule's pharmacokinetics, toxicity, and bioavailability, making it a key factor in drug design. Accurate prediction of log P aids in optimizing the absorption, distribution, metabolism, and excretion properties of drug candidates, thereby reducing costs and accelerating the drug discovery process. For example, log P is part of Lipinski's Rule of Five, which serves as a significant guide in oral drug development. According to the rule, the calculated n-octanol-water log P value will be less than 5^[1]. Additionally, log P significantly affects the solubility and miscibility of solution-processed organic materials, such as organic photovoltaic materials and organic light-emitting diodes, which are crucial for the morphology and performance of organic active layers^[2,3]. Therefore, fast and accurate log P estimation is highly demanded.

Recent advances in computational chemistry have led to the development of several sophisticated methods aimed at accurately predicting the log P values of molecules. These methods range from empirical quantitative structure-property relationship (QSPR) models, which rely on historical data and molecular descriptors, to more complex physics-based approaches, such as molecular dynamics (MD) simulations^[4-6] and quantum chemical (QC) calculations^[7-9]. Earlier QSPR approaches were widely used in the prediction of log P in the 2000s. However, recent years have seen improved predictive performance with more sophisticated, physics-based approaches, as evidenced in the SAMPL6 and SAMPL9 challenges. For example, In October 2019, Procacci *et al.* utilized the CHARMM general force field (CGenFF) in a nonequilibrium alchemical approach in the SAMPL6 challenge, achieving a root mean square error (RMSE) of 0.82^[4]. Nikitin employed a Toukan-Rahman water, resulting in an RMSE of 0.75^[5]. Tielker *et al.* applied the embedded cluster reference interaction site model (EC-RISM) solvation model with quantum-chemical (QC) calculations, attaining an RMSE of 0.47^[7]. Guan *et al.* also made strides with the solvation model density (SMD) solvation model, achieving an RMSE of 0.49^[8]. Loschen *et al.* used the conductor-like screening model for realistic solvation (COSMO-RS) approach, achieving a lower RMSE of 0.38^[9]. However, the COSMO-RS method encountered difficulties in predicting the log P values of molecules exhibiting dimerization effects, as it tends to overestimate their hydrophilicity, thereby diminishing the accuracy of the predictions. To enhance the predictive accuracy, additional data and computational efforts are necessary. This indicates that complex approaches are not invariably precise.

In 2020, machine learning (ML) made significant strides in the SAMPL6 challenge. Prasad *et al.* developed a deep learning approach that achieved an RMSE of 0.61^[10], which used the extended-connectivity finger printing (ECFP) to make a vector space representation of the molecule as the input to the neural network. Meanwhile, Lui *et al.* presented a ML-QSPR model that further excelled with an even lower RMSE of 0.49^[11], outperforming all MD methods. Later, in 2021, Ulrich *et al.* introduced a new deep learning approach that surpassed the best results of the COSMO-RS method in the SAMPL6 challenge, with an RMSE of 0.33^[12]. Their approach utilized data augmentation by taking potential tautomers of chemicals. Continuing this trend, in the SAMPL9 challenge, two ML approaches by Zamora *et al.* claimed the top spots, with RMSEs of 0.84 and 0.86^[13], using their meticulously selected high-quality dataset. Topological indices and graph algorithms have been rapidly developed in recent years. The optimization of topological indices and graph algorithms will promote the predictive capability of deep learning models and lead to accurate prediction^[14]. For example, these models with graph algorithms were implemented to predict log P. Nevolianis *et al.* implemented the directed-message passing neural network (D-MPNN) to predict the log P values of molecules in the SAMPL9 data, achieving an RMSE of 1.02^[15]. However, despite its greater model complexity and longer training times, the predictive performance did not significantly improve and the GNN did not demonstrate significantly superior predictive performance over other approaches.

Generally, the accuracy of log P prediction, no matter which methods and how complex models were used, remains limited. In this work, we were trying to develop the optimized 3D molecule representation of structure based on electron diffraction descriptor (opt3DM) descriptors by implementing a scale factor (s_L) to realize highly accurate log P prediction. After optimizing the s_L and dimension, the accuracies of our models are comparable, and even higher than QC- and MD-based models. By training ML models on the M-dataset, we achieved the lowest RMSE of 0.31 on the SAMPL6 data among all approaches. This work not only achieved fast and accurate methods for log P prediction, but also demonstrated that the simple ML-based models can realize better evaluations than more advanced ML models and QC and MD simulations by using only efficient opt3DM descriptors.

METHODS

Dataset

In 2018, Mansouri *et al.* developed the OPEN structure-activity/property relationship app (OPERA) models based on their dataset^[16]. We used their data as the initial training and testing sets, referring to it as M-dataset. The M-dataset was derived from the available PHYSPOROP physicochemical property and environmental fate datasets, with extensive curation conducted to ensure data quality. It provided log P values and chemical structure formats for 14,050 molecules, along with identifiers such as simplified molecular input line entry specification (SMILES) and structure data file (SDF). In this work, SMILES of 13,963 molecules and their experimental log P values were selected to create the descriptors. Finally, descriptors of 13,952 molecules were calculated using homemade code based on the RDKit library.

The datasets for prediction were obtained from the SAMPL6 challenge and the SAMPL9 challenge. The SAMPL6 challenge provided a set of 11 drug-like molecules in SMILES string format and their experimental values, and the SAMPL9 challenge provided 16 molecules [Figure 1], which had experimental log P values ranging from -1.37 log P units to 4.92 log P units. With the SMILES provided, the descriptors used for prediction were created in the same way as the ones used for training and testing.

Descriptors

The concept of 3D molecular representation of structures based on electron diffraction (3D-MoRSE) descriptors was initially introduced by Schuur, Selzer, and Gasteiger in 1996^[17,18]. These descriptors are expressed as $I(s) = \sum_{i=2}^N \sum_{j=1}^{i-1} A_i A_j f(s, r_{ij})$, where $f(s, r_{ij}) = \frac{\sin sr_{ij}}{sr_{ij}}$ serves as the core function, where s is the scattering parameter ranging from 0 to 31 Å⁻¹, r_{ij} is the distance between i th and j th atoms, N is the total number of atoms, and A_i and A_j are the atomic weights, which can be unweighted or represent various atomic properties. The atomic weights used in this work are shown in Table 1.

Our previous works have shown that the ML approach with intermolecular 3D-MoRSE descriptors could realize very high electronic coupling prediction between molecules^[19] and the opt3DM descriptors could act as efficient descriptors for developing interface modifiers in perovskite solar cells^[20]. Based on these works, we here optimized 3D-MoRSE descriptors for the log P prediction. A s_L was added to adjust $f(s, r_{ij})$, which is defined as $f(s, r_{ij}) = \frac{\sin s \times s_L \times r_{ij}}{s \times s_L \times r_{ij}}$. By meticulously adjusting the coefficients ($s \times s_L$) and the dimension (N_s , range of s), this updated descriptor set has demonstrated the potential to surpass all currently available descriptors in terms of prediction accuracy. As found, the best prediction results were calculated when $s_L = 0.5$ and $N_s = 500$.

ML model

The ML algorithms used in this work were implemented with the scikit-learn library. Our ML models consisted of a feature selector and a fitting regressor. The SelectFromModel, a feature selector from the scikit-learn library, served as the selector in the ML model. Various algorithms, including automatic

Table 1. Atomic weights used in the optimized 3D MoRSE descriptors

Abbreviation	Illustration
morU	1.0
morM	Atomic mass
morV	Atomic van der Waals volume
morP	Atomic polarizability
morE	Atomic Sanderson electronegativity
morC	Atomic charge
morIP	Atomic ionization potential
morIS	Atomic intrinsic state
morRC	Atomic covalent radius

3D MoRSE: 3D molecular representation of structures based on electron diffraction.

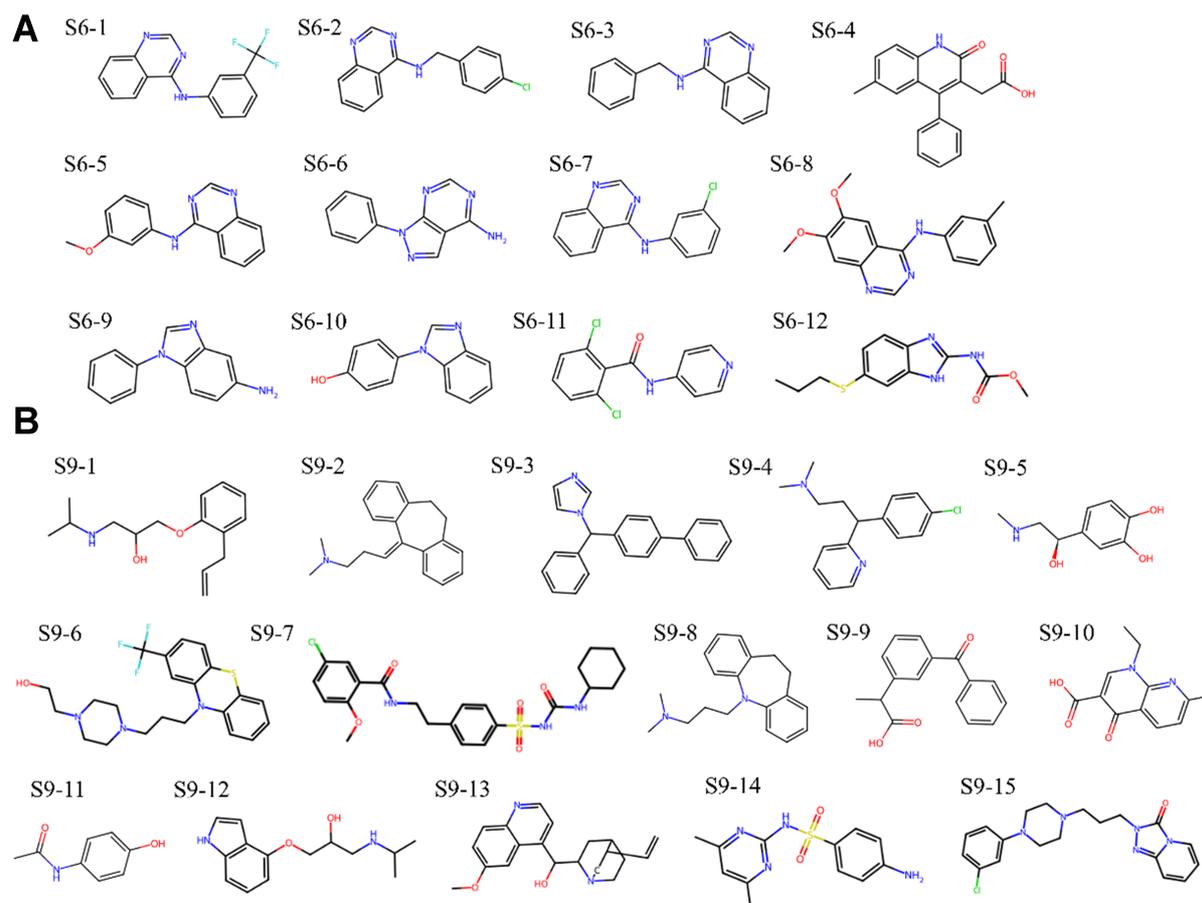


Figure 1. (A) Eleven molecules in the SAMPL6 log P challenge; (B) Sixteen molecules in the SAMPL9 log P challenge.

relevance determination (ARD) regression, Bayesian Ridge, and Ridge, were employed in both the feature selection and regression stages. These algorithms were trained on the M-dataset to identify the most suitable ones for the ML model, as illustrated in the subsequent sections. The selected ML model was used for the prediction of the SAMPL6 and the SAMPL9 datasets. The mean absolute error (MAE), RMSE, and coefficients of determination (R^2) serve as evaluation metrics, which are defined as

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

Where n represents the number of molecules, and y_i and \hat{y}_i are the experimental values and the predicted values of log P, respectively.

RESULTS AND DISCUSSION

Prediction on the M-dataset

The M-dataset, created by Mansouri *et al.*, is a dataset providing information for predictions on physicochemical properties and other fields and it is available on GitHub^[21]. The log P distribution is shown in [Supplementary Figure 1](#). Mansouri *et al.* utilized genetic algorithms (GA) to select features and employed a k-nearest neighbors (KNN) approach for model fitting and prediction^[16]. Their GA-KNN model achieved an RMSE of 0.78 and an R^2 of 0.86 in the test set.

To achieve the best prediction on the OPERA dataset, selecting descriptors and algorithms is crucial. By introducing the s_L , the range and sensitivity of the coefficients can be adjusted. The s_L serves as a parameter that adjusts the granularity at which the descriptors operate. A lower s_L increases the sensitivity of the descriptors, allowing them to detect finer variations in the molecular stackings, while a higher s_L might simplify the descriptor, potentially overlooking critical details. Following our previous work^[20], we have implemented algorithms referred to in the Methods section and tested the s_L from 0.03 to 3.0 and the N_s up to 500 on the M-dataset with 90% for training and 10% for testing [[Supplementary Tables 1 and 2](#)]. [Figure 2A](#) shows the s_L -dependent RMSEs of various algorithms with descriptors calculated with 100 dimensions ($N_s = 100$). The Ridge, Bayesian Ridge, and ARD regression achieved lower RMSEs among the algorithms, in which we found that the optimal s_L is between 0.1-1.0. As shown in [Figure 2B](#), the result showcased that feature selection did not affect the selection of s_L . Second, we increased the dimension to 500 and screened s_L in the optimal range with 10 tests [[Table 2](#)]. The best prediction result (with the smallest average RMSE for independent ten predictions) was achieved when $s_L = 0.50$ and $N_s = 500$.

To abstract as much information as possible from molecules, a large N_s is required. However, the large N_s increases the dimension, and also includes some reductant information. Therefore, a feature selection was necessary. We employed the SelectFromModel module (Meta-transformer for selecting features based on importance weights, which is a module in the scikit-learn library) with ARD regression as the estimator for feature selection. In this process, we used 90% of the dataset to select features and to train the ML model, as 10% of the dataset was set for testing. The dataset was split randomly with a random state ranging from 0 to 20 in order to obtain robust and reliable results. Moreover, a threshold was set for finding the optimal dimension and we found the optimal threshold was 0.1 [[Supplementary Tables 3-8](#)]. However, noticing that the reduction of feature dimension may enhance the generalized capability, features selected at a threshold of 0.05 will be considered. To test the average performance of these descriptors, we employed these descriptors to fit the model and predict the test set based on those divided datasets. The average performance of descriptors selected in various thresholds and the dimensions of these descriptors are shown

Table 2. The RMSEs of prediction for the test set of M-dataset with various algorithms and optimized 3D-MoRSE descriptors calculated by selected s_L and N_s

Parameter	Regressor	Average RMSE (lowest, highest)
$s_L = 0.1, N_s = 500$	ARD	0.73 (0.68, 0.74)
	Ridge	0.77 (0.71, 0.88)
	Bayesian Ridge	0.73 (0.69, 0.76)
$s_L = 0.2, N_s = 500$	ARD	0.72 (0.67, 0.76)
	Ridge	0.74 (0.69, 0.78)
	Bayesian Ridge	0.71 (0.68, 0.75)
$s_L = 0.5, N_s = 500$	ARD	0.72 (0.67, 0.78)
	Ridge	0.72 (0.67, 0.79)
	Bayesian Ridge	0.71 (0.67, 0.77)
$s_L = 1.0, N_s = 500$	ARD	0.81 (0.77, 0.92)
	Ridge	0.79 (0.71, 0.88)
	Bayesian Ridge	0.77 (0.72, 0.87)

The test size was 0.1. RMSEs: Root mean square errors; 3D MoRSE: 3D molecular representation of structures based on electron diffraction; ARD: automatic relevance determination.

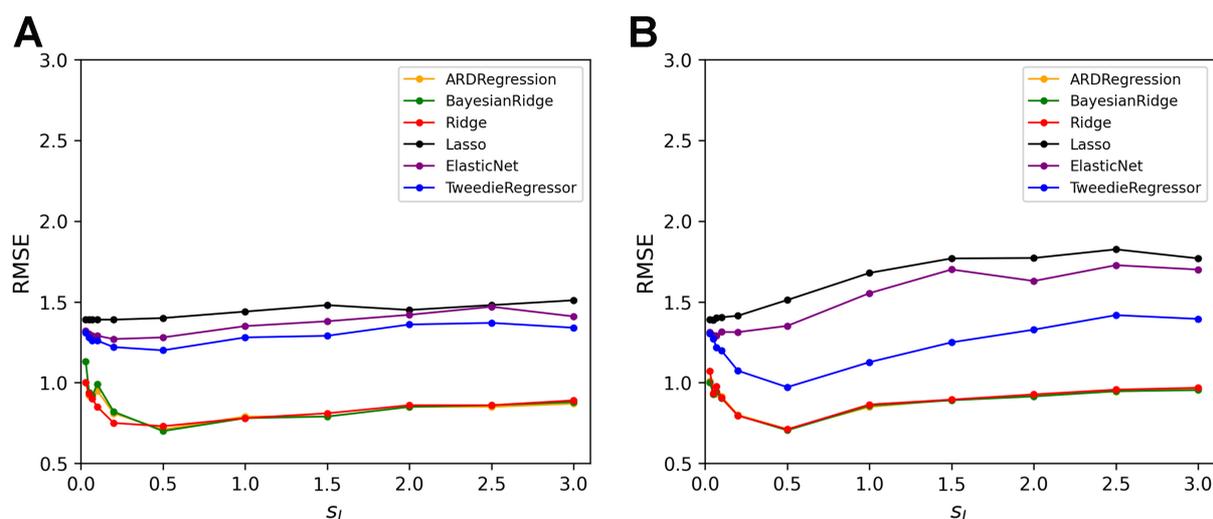


Figure 2. The RMSEs of predictions for the test set of M-dataset with various algorithms and optimized 3D-MoRSE descriptors calculated by various s_L . (A) The s_L -dependent RMSE without descriptor selection; (B) The s_L -dependent RMSE with descriptor selection. The testing ratio was 0.1. The descriptors were selected using ARD regression. RMSEs: Root mean square errors; 3D-MoRSE: 3D molecular representation of structures based on electron diffraction; ARD: automatic relevance determination.

in Figure 3.

To have a direct comparison with the work of Mansouri *et al.*, we used the same dataset and training and testing splitting ratio of 0.75:0.25^[16]. To ensure robustness, the dataset was partitioned into training and test subsets 20 times with different random states, resulting in 20 distinct tests [Supplementary Table 9]. This iterative process mitigates the risk of coincidental outcomes and substantiates the model's dependability. The best and the worst predictions have not much difference in prediction accuracy, as shown in Figure 4, demonstrating that our models are very stable and robust. Table 3 lists the performance of the OPERA model and our models. The OPERA model predicts on test set for one time and achieves an RMSE of 0.78, while our models demonstrate an average RMSE of 0.68. Collectively, these results indicate that our model

Table 3. Comparison of the performance of our model and the OPERA model

Parameter	Regressor	RMSE (lowest)	R ² (highest)	Note
The OPERA model	KNN	0.78	0.86	One simulation
$s_L = 0.5, N_s = 500$	ARD	0.68 (0.66)	0.86 (0.87)	Average of 20 simulations
$s_L = 0.5, N_s = 500$	Ridge	0.68 (0.66)	0.86 (0.87)	Average of 20 simulations
$s_L = 0.5, N_s = 500$	Bayesian Ridge	0.68 (0.66)	0.86 (0.87)	Average of 20 simulations

Both results of the best random state and the average results of all random states are provided. The test size was 0.25. Descriptors are selected from the SelectFromModel module with Thresholds = 0.05 using the ARD regression as the estimator. OPERA: OPEn structure-activity/property relationship app; RMSE: root mean square error; R²: coefficients of determination; KNN: k-nearest neighbor; ARD: automatic relevance determination.

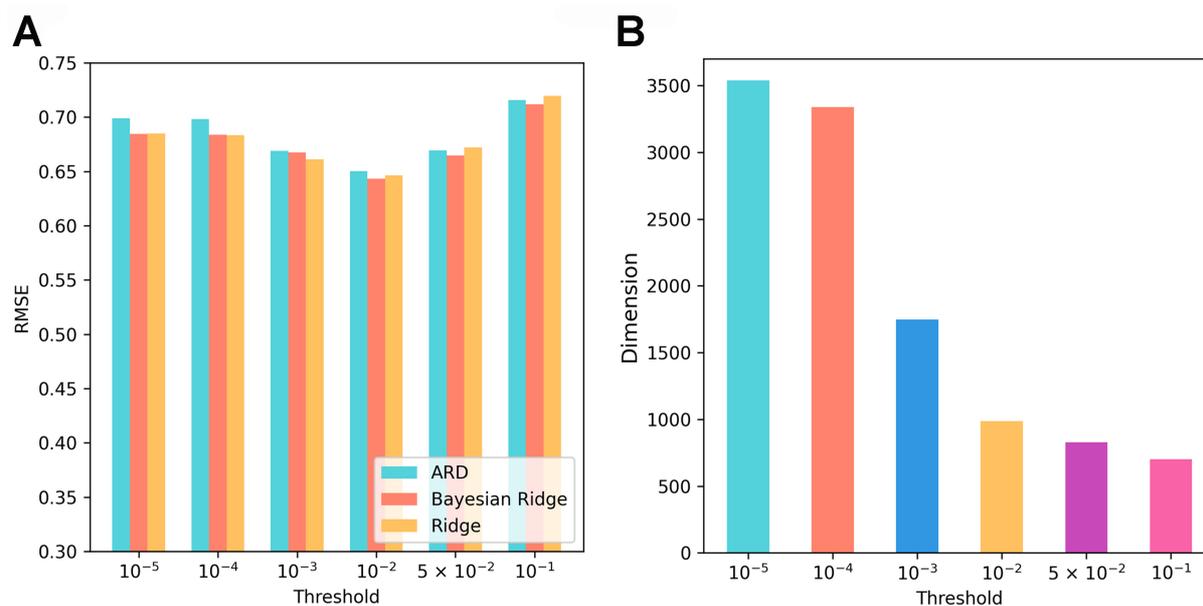


Figure 3. The RMSEs and feature dimensions of models using descriptors selected with various thresholds. (A) The RMSEs of models fitted with various algorithms. The optimal threshold was around 0.001 and 0.05; (B) The feature dimension decreased rapidly with the increased threshold. RMSEs: Root mean square errors.

demonstrated superior performance compared to the OPERA model.

Predictions on the SAMPL6 dataset and the SAMPL9 dataset

The SAMPL6 and the SAMPL9 challenges are the two most famous log P challenges, having received over 100 submissions from various groups. To verify the reliability of our models, we utilized the selected descriptors and models trained on the M-dataset to predict the log P in the SAMPL6 and SAMPL9 challenges.

We employed the ARD regression, Bayesian Ridge Regression and Ridge Regression to train our models, which achieved average RMSEs of 0.309, 0.326 and 0.329 [Supplementary Tables 10 and 11], respectively. Notably, the minimum values achieved was an RMSE of 0.29 [Figure 5A]. The comparative performance of our models and other state-of-the-art models on the SAMPL6 dataset is detailed in Supplementary Table 10 and Figure 6A. The other models encompassed various methodologies, including deep learning, QC calculations, MD simulations, and other complex approaches. Our models outperformed most of the challengers. Among them, only the deep neural network with tautomer (DNNtaut) and the deep neural

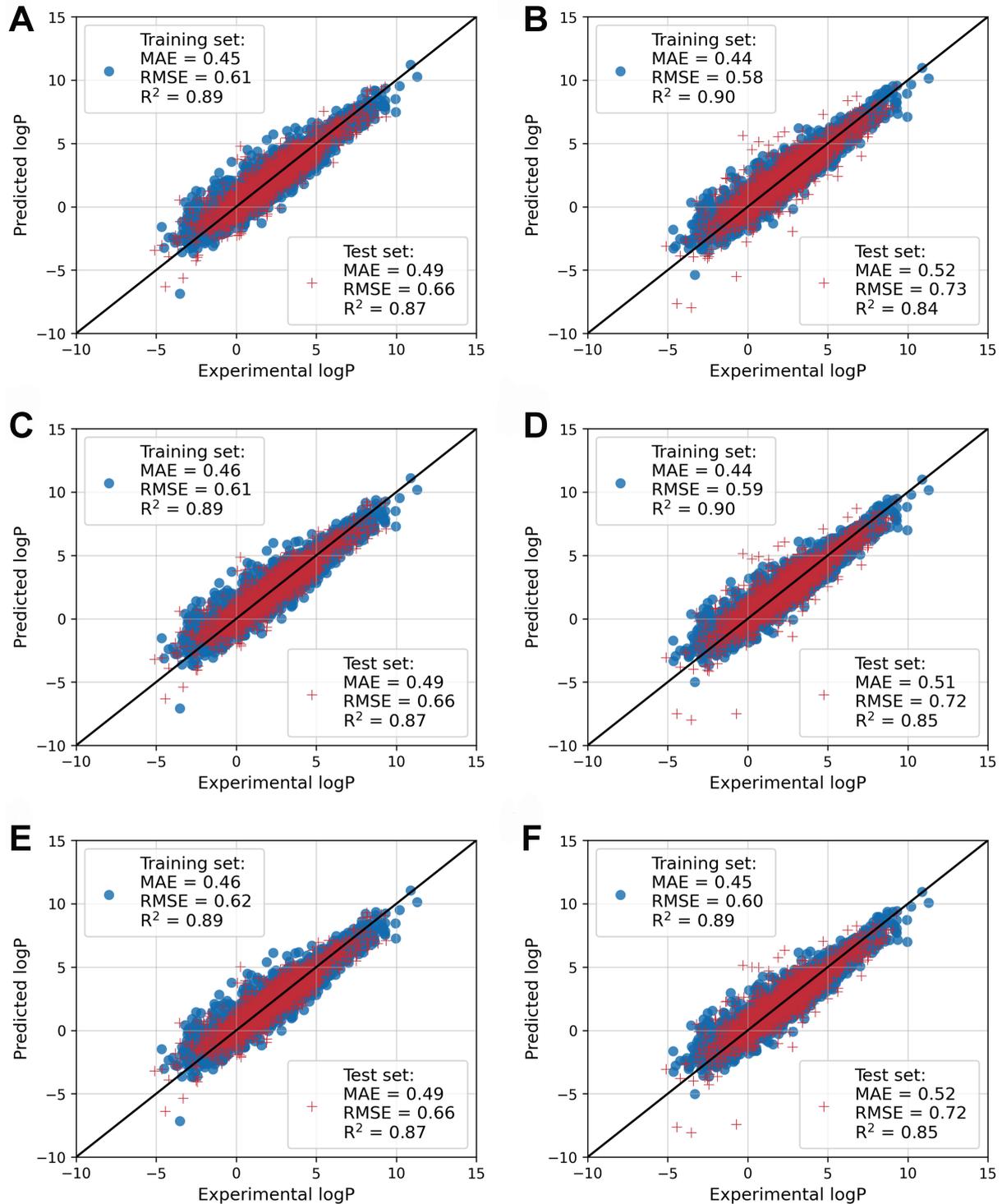


Figure 4. The experimental log P values and the ML predicted log P values. (A) The ARD regression model with the best (A) and the worst (B) performance in 20 simulations with different random states; The Bayesian Ridge regression model with the best (C) and the worst (D) performance in 20 simulations with different random states. The Ridge regression model with the best (E) and the worst (F) performance in 20 simulations with different random states. All models used the coefficient set $s_l = 0.5$ and $N_s = 500$. The test set was 0.25. ML: machine learning; ARD: automatic relevance determination.

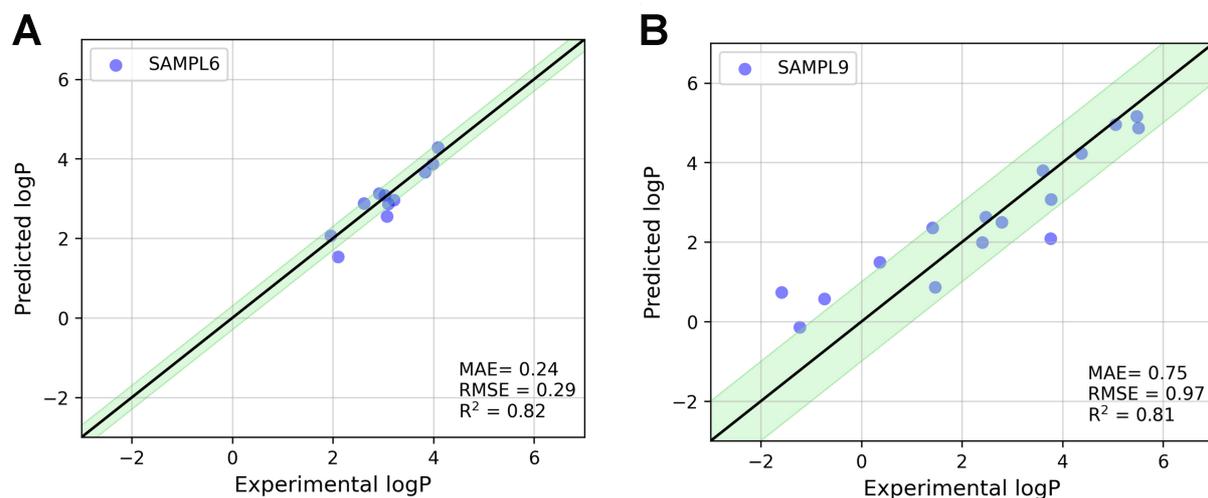


Figure 5. The experimental and predictive log P using the ARD regression model with the best random state. (A) This image showcases the prediction performance on the SAMPL6 dataset. The error range was 0.30; (B) This image showcases the prediction performance on the SAMPL9 dataset. The error range was 1.00. The MAE, RMSE, and R^2 are provided. The model was trained on the M-dataset using the coefficient set $s_L = 0.5$ and $N_s = 500$. ARD: Automatic relevance determination; MAE: mean absolute error; RMSE: root mean square error; R^2 : coefficients of determination.

network with original SMILES (DNNmono) demonstrated similar performance compared to ours. The DNNmono, a deep learning model developed by Ulrich *et al.*, reported the lowest RMSE of 0.31 and typical RMSE of 0.33 in their other models^[12]. Our models show the lowest RMSE on the SAMPL6 challenge among all models, outperform QC- and MD-based models, and also surpass deep neural network models. They offer advantages in terms of complexity, indicating faster prediction times. Additionally, these models exhibited commendable stability, which is reflected in the compact distribution of MAE, RMSE, and R^2 across 20 stochastic parameter configurations. Overall, our models provided the most accurate prediction on the SAMPL6 dataset among more than 100 submissions, emphasizing the superior efficiency of our developed opt3DM descriptors.

Our models not only demonstrated robust performance in the prediction of the M-dataset and the SAMPL6 dataset, but also showed good prediction in the SAMPL9 challenge. From 2021 to 2023, submissions to the SAMPL9 challenge showed varied results, with RMSEs ranging from 1.52 to 2.95^[15,22] [Supplementary Table 12]. Later, Nevolianis *et al.* replicated the success of the COSMO-RS approach in the SAMPL9 challenge achieving an RMSE of 1.23^[23]. Despite their sophistication, these physics-based approaches demand substantial computational resources as their model requires QC calculations. Besides, the submission by Nevolianis *et al.* to SAMPL9, using the COSMO-RS approach, required 3.7 h to predict the log P of just sixteen molecules, indicating that predictions for a larger number of molecules would be time-consuming^[23]. By using ARD regression, we achieved the lowest average RMSE of 1.01 and attained a minimum RMSE of 0.97 in a single iteration [Figure 5B and Supplementary Table 13]. The best prediction on SAMPL9 was reported by Zamora *et al.*^[13]. However, duplicate entries with SAMPL9 data were found in their training dataset^[15]. Therefore, we can conclude that our model again gives the most accurate prediction in the SAMPL9 challenge with blind test, as shown in Figure 6B. Additionally, we trained the ML models on the M-dataset consisting of the n-octanol/water log P ($\log P_{o/w}$), and realized accurate prediction on the SAMPL9 data that are all toluene/water log P ($\log P_{tol/w}$), demonstrating the generalization ability of our model in predicting log P.

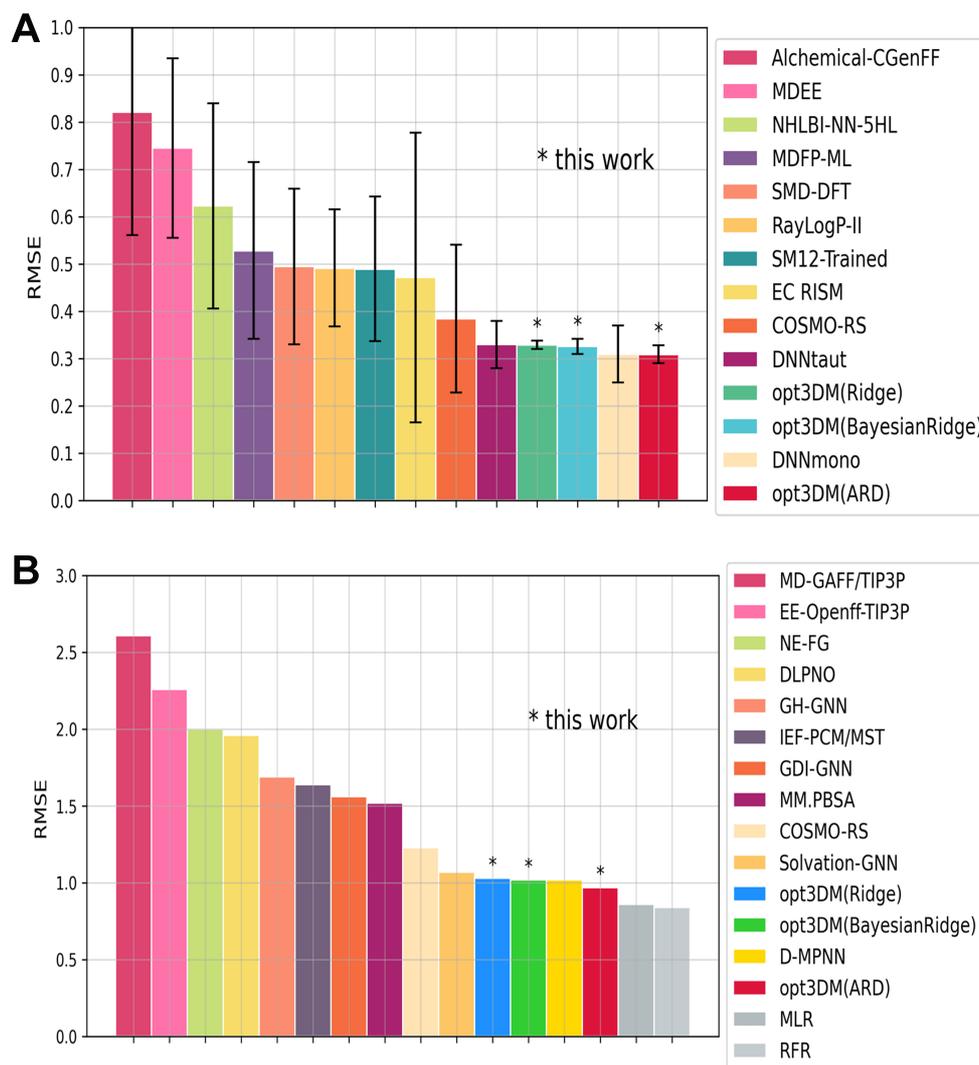


Figure 6. The RMSEs of various models and our models. (A) The RMSEs of models in SAMPL6 challenges and our models; (B) The lowest RMSEs in one iteration of models in SAMPL9 challenges and our models. RMSEs: Root mean square errors.

To identify the molecular features with great importance to the log P , we performed Shapley Additive Explanations (SHAP) analysis^[24], and the results are shown in Figure 7A and B. In previous studies, the first half of the sine wave period is pronounced. When $s_L = 0.5$, the half-sine wave period is about 3.14 Å for MorU-2, MorRC-2, and MorIP-2. They are longer than most of the chemical bonds (about 1.5 Å) in molecules. As the atomic weight is 1.0 for MorU, the descriptor value is solely determined by bond lengths. As the bond length decreases, the MorU-2 value increases. Unsaturated carbon chains, compared to alkanes, generally have stronger hydrophilicity. This is because the double or triple bonds present in unsaturated carbon chains introduce polarity, reducing the overall non-polar nature and enhancing the polarity of the molecule. In unsaturated carbon chains, the presence of double bonds leads to an uneven distribution of electron cloud density, which imparts polarity to the molecule, thereby increasing the interaction with water molecules and making these molecules more hydrophilic. In contrast, alkanes, which only have single bonds, have a more uniform electron cloud distribution and are non-polar overall, thus exhibiting stronger hydrophobicity. The more unsaturated carbon chains a molecule has, the denser the molecule is, the larger the MorU-2 value is, and the more hydrophilic the molecule becomes. It is the same

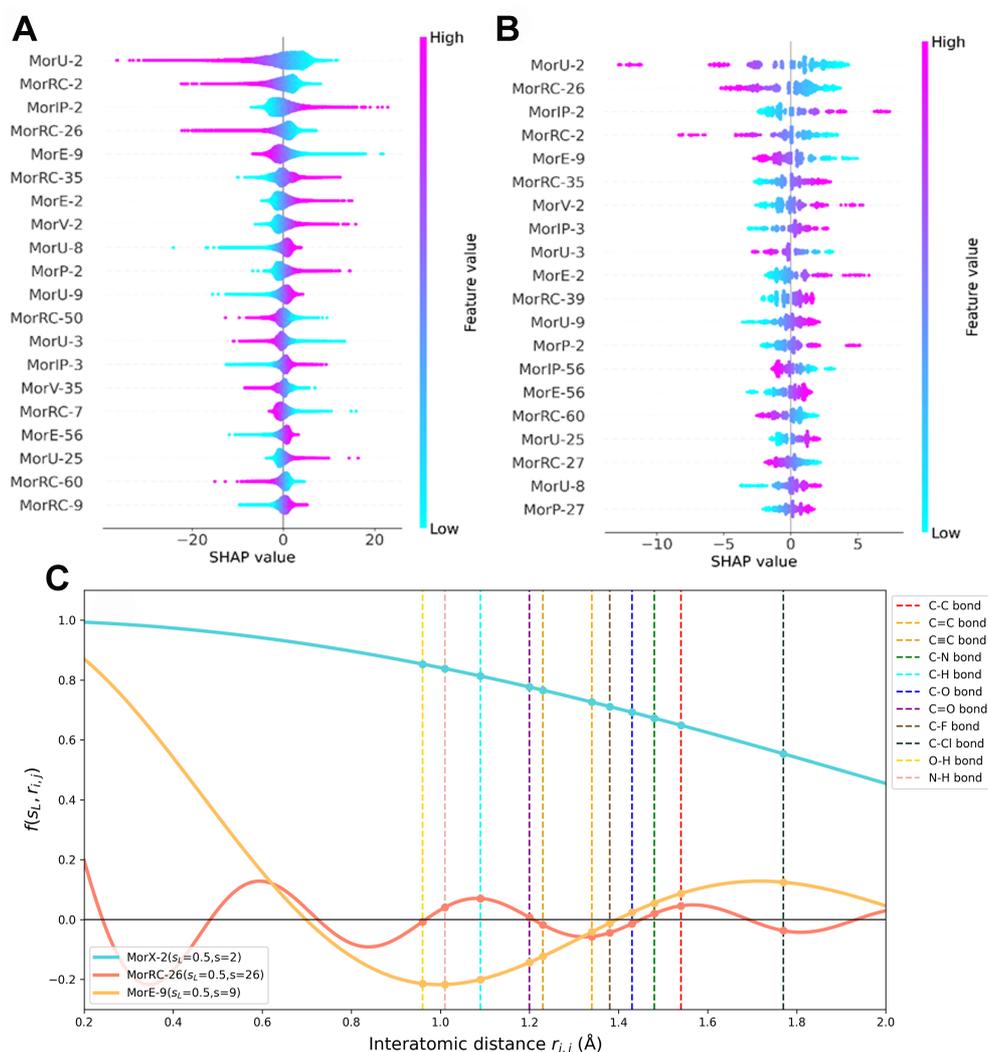


Figure 7. The SHAP values of twenty ARD regression models on (A) M-dataset, (B) SAMPL 6 and SAMPL 9 and (C) The $f(s_i, r_{i,j})$ of MorX-2 (X = U, RC, IP), MorRC-26 and MorE-9 and the function values of atoms with various distances in different chemical bonds. SHAP: Shapley Additive Explanations; ARD: automatic relevance determination.

for MorRC-2, which gives carbon a larger weight. Therefore, MorU-2 and MorRC-2 are negatively correlated with $\log P$. The MorIP-2 gives hydrogen a larger weight. A greater MorIP-2 value indicates more alkanes, leading to hydrophobicity. Therefore, MorIP-2 is positively correlated with $\log P$. For the MorE-9, the sine wave period is about 1.40 Å. As shown in Figure 7C, its value increases with the length of bonds in most cases. More alkanes cause a smaller MorE-9 value and lead to hydrophobicity. Therefore, MorIP-2 is negatively correlated with $\log P$. For MorRC-26, the sine wave period is about 0.48 Å. As shown in Figure 7C, the C–F and C–Cl bonds in halogenated hydrocarbons which are hydrophobic give a negative value on the MorRC-26 curve; The C–O and C–N bonds in hydrophilic groups have little impact on the MorRC-26 value. Therefore, MorRC-26 is negatively correlated with $\log P$.

Figure 8 illustrates the RMSE of different ML models applied to the SAMPL6 dataset, along with their complexity and computational expense. Notably, our model, being the simplest, achieved the lowest RMSE. Its minimal complexity also resulted in significantly reduced computational costs compared to other models, highlighting the efficiency of our opt3DM descriptors. These descriptors have significantly

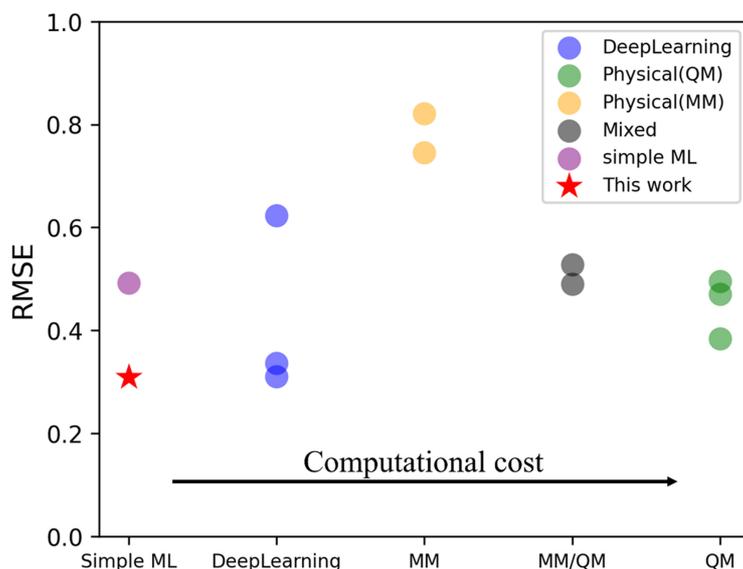


Figure 8. The RMSE of predictions, model complexity and categories of various approaches on SAMPL6 dataset. RMSE: Root mean square errors.

advanced rapid and precise log P prediction, facilitating extensive drug and material screening.

CONCLUSION

In this work, we introduced the opt3DM descriptors and employed them to construct ML models for predicting log P. These ML models were trained and tested on the M-dataset. We have tested s_l ranging from 0.03 to 3 and N_s up to 500 with various algorithms to select the most suitable parameters for the descriptors and efficient algorithms. We found that the best log P prediction was achieved when $s_l = 0.5$ and $N_s = 500$. We further refined our models by selecting features for training and optimizing the algorithms, resulting in highly accurate predictions with an average RMSE of 0.68, which surpassed the OPERA model (0.78). To further verify the accuracy of our models, we trained them on the M-dataset and subsequently used the trained models to forecast log P values of the SAMPL6 and SAMPL9 datasets. The obtained average RMSE is 0.31 for SAMPL6, which is 10% lower than the current best model (0.33) among all challengers. Our models also give the most accurate prediction for the SAMPL9 with an average RMSE of 1.01 when the training was performed on log $P_{o/w}$ datasets. It is noteworthy that while these models showed promising results, there is still room for improvement in the R^2 values. Generally, by using efficient descriptors, we are able to use simple ML methods to realize accurate log P prediction comparable to the most accurate models based on quantum chemistry calculation and more advanced ML methods. This demonstrates that the opt3DM descriptors are highly efficient in representing molecules, potentially offering high efficiency across a broad range of applications.

DECLARATIONS

Acknowledgments

The authors acknowledge the National Super-Computer Center in Guangzhou for providing computational resources on Tianhe-2.

Authors' contributions

Conceived and designed the article: Zeng, X., Zhou, Y., Ye, X.

Collected and analyzed data: Zeng, X., Cui, N., Bao, Y.
Developed programmers and performed data visualization: Zeng, X., Ye, X., Liu, D., Li, X.
Drafted manuscript: Zeng, X., Zhou, Y., Ye, X., Liu, D., Cui, N.
Reviewed and edited the manuscript: Zeng, X., Zhou, Y., Ye, X.
Supervised the project and provided financial support: Zhou, Y.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Financial support and sponsorship

This study was supported by the National Natural Science Foundation of China (No. 22103097) and funded by the Guangzhou Science and Technology Program (No. 202102020495).

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.

Copyright

© The Author(s) 2025.

REFERENCES

1. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3-25. [DOI](#)
2. Sun, J.; Li, D.; Zou, J.; et al. Accelerating the discovery of acceptor materials for organic solar cells by deep learning. *npj. Comput. Mater.* **2024**, *10*, 1367. [DOI](#)
3. Zhang, R.; Chen, H.; Wang, T.; et al. Equally high efficiencies of organic solar cells processed from different solvents reveal key factors for morphology control. *Nat Energy*. 2024. [DOI](#)
4. Procacci, P.; Guarnieri, G. SAMPL6 blind predictions of water-octanol partition coefficients using nonequilibrium alchemical approaches. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 371-84. [DOI](#) [PubMed](#)
5. Nikitin, A. Non-zero Lennard-Jones parameters for the Toukan-Rahman water model: more accurate calculations of the solvation free energy of organic substances. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 437-41. [DOI](#) [PubMed](#)
6. Ali, H. S.; Henschman, R. H. Energy-entropy multiscale cell correlation method to predict toluene-water log P in the SAMPL9 challenge. *Phys. Chem. Chem. Phys.* **2023**, *25*, 27524-31. [DOI](#) [PubMed](#) [PMC](#)
7. Tielker, N.; Tomazic, D.; Eberlein, L.; Güssregen, S.; Kast, S. M. The SAMPL6 challenge on predicting octanol-water partition coefficients from EC-RISM theory. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 453-61. [DOI](#) [PubMed](#) [PMC](#)
8. Guan, D.; Lui, R.; Matthews, S. LogP prediction performance with the SMD solvation model and the M06 density functional family for SAMPL6 blind prediction challenge molecules. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 511-22. [DOI](#) [PubMed](#)
9. Loschen, C.; Reinisch, J.; Klamt, A. COSMO-RS based predictions for the SAMPL6 logP challenge. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 385-92. [DOI](#) [PubMed](#)
10. Prasad, S.; Brooks, B. R. A deep learning approach for the blind logP prediction in SAMPL6 challenge. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 535-42. [DOI](#) [PubMed](#) [PMC](#)
11. Lui, R.; Guan, D.; Matthews, S. A comparison of molecular representations for lipophilicity quantitative structure-property relationships with results from the SAMPL6 logP Prediction Challenge. *J. Comput. Aided. Mol. Des.* **2020**, *34*, 523-34. [DOI](#)
12. Ulrich, N.; Goss, K. U.; Ebert, A. Exploring the octanol-water partition coefficient dataset using deep learning techniques and data augmentation. *Commun. Chem.* **2021**, *4*, 90. [DOI](#) [PubMed](#) [PMC](#)
13. Zamora, W. J.; Viayna, A.; Pinheiro, S.; et al. Prediction of toluene/water partition coefficients in the SAMPL9 blind challenge: assessment of machine learning and IEF-PCM/MST continuum solvation models. *Phys. Chem. Chem. Phys.* **2023**, *25*, 17952-65. [DOI](#)

14. Liu, J. B.; Wang, X.; Cao, J. The coherence and properties analysis of balanced $2p$ -ary tree networks. *IEEE. Trans. Netw. Sci. Eng.* **2024**, *11*, 4719-28. DOI
15. Nevolianis, T.; Rittig, J. G.; Mitsos, A.; Leonhard, K. Multi-fidelity graph neural networks for predicting toluene/water partition coefficients. *ChemRxiv* 2024. Available online: <https://doi.org/10.26434/chemrxiv>. (accessed 9 Jan 2024).
16. Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminform.* **2018**, *10*, 10. DOI PubMed PMC
17. Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334-44. DOI
18. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical information in 3D space. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030-7. DOI
19. Ma, J.; Du, Z.; Lei, Z.; et al. Intermolecular 3D-MoRSE descriptors for fast and accurate prediction of electronic couplings in organic semiconductors. *J. Chem. Inf. Model.* **2023**, *63*, 5089-96. DOI
20. Ye, X.; Cui, N.; Ou, W.; et al. Explainable optimized 3D-MoRSE descriptors for the power conversion efficiency prediction of molecular passivated perovskite solar cells through machine learning. *J. Mater. Chem. A.* **2024**, *12*, 26224-33. DOI
21. Mansouri, K.; Sahu, R. OPERA. Available from: <https://github.com/kmansouri/OPERA.git>. [Last accessed on 9 Jan 2025].
22. Mobley, D. L.; Amezcua, M.; Dani SAMPL9. Available from: <https://github.com/samplchallenges/SAMPL9>. [Last accessed on 9 Jan 2025].
23. Nevolianis, T.; Ahmed, R. A.; Hellweg, A.; Diedenhofen, M.; Leonhard, K. Blind prediction of toluene/water partition coefficients using COSMO-RS: results from the SAMPL9 challenge. *Phys. Chem. Chem. Phys.* **2023**, *25*, 31683-91. DOI PubMed
24. Lundberg, S.; Lee, S. I. A unified approach to interpreting model predictions. *arXiv* 2017, arXiv:1705.07874. Available online: <https://doi.org/10.48550/arXiv.1705.07874>. (accessed 9 Jan 2025).