Environmental Science & Technology

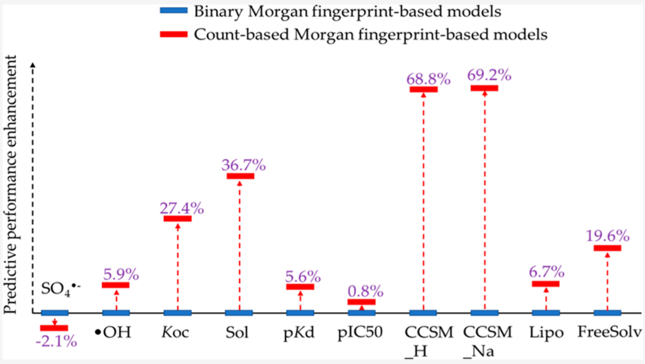
Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties

Shifa Zhong and Xiaohong Guan*

Cite This: https://doi.org/10.1021/acs.est.3c02198 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this study, we introduce the count-based Morgan fingerprint (C-MF) to represent chemical structures of contaminants and develop machine learning (ML)-based predictive models for their activities and properties. Compared with the binary Morgan fingerprint (B-MF), C-MF not only qualifies the presence or absence of an atom group but also quantifies its counts in a molecule. We employ six different ML algorithms (ridge regression, SVM, KNN, RF, XGBoost, and CatBoost) to develop models on 10 contaminant-related data sets based on C-MF and B-MF to compare them in terms of the model's predictive performance, interpretation, and applicability domain (AD). Our results show that C-MF outperforms B-MF in nine of 10 data sets in terms of model predictive performance. The advantage of C-MF over B-MF is dependent on the ML algorithm, and the performance enhancements are proportional to the difference in the chemical diversity of data sets calculated by B-MF and C-MF. Model interpretation results show that the C-MF-based model can elucidate the effect of atom group counts on the target and have a wider range of SHAP values. AD analysis shows that C-MF-based models have an AD similar to that of B-MF-based ones. Finally, we developed a "ContaminaNET" platform to deploy these C-MF-based models for free use.



KEYWORDS: *count-based Morgan fingerprint, machine learning, QSAR, water contaminants, ContaminaNET*

1. INTRODUCTION

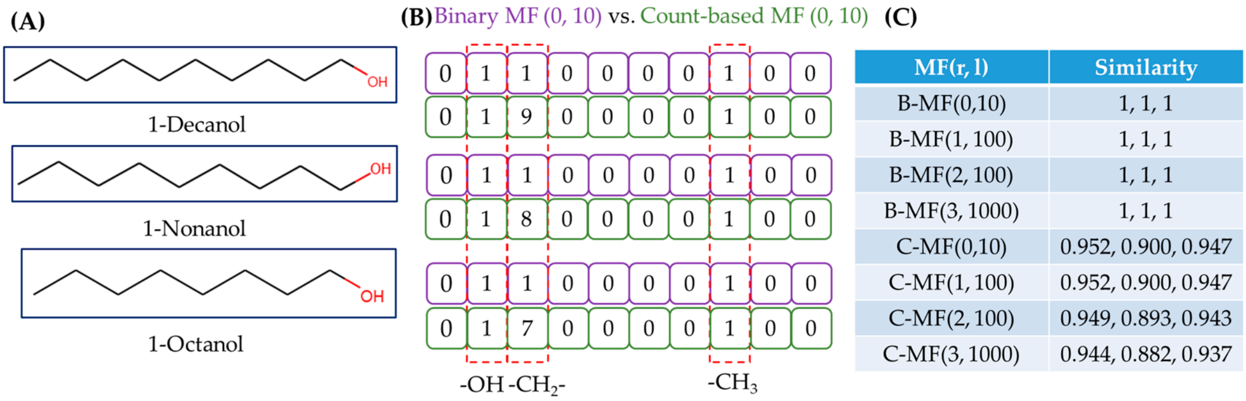
Quantitative structure−activity relationship (QSAR) models are a valuable tool for predicting the activities and properties of query chemicals, which can help to avoid time-consuming and labor-intensive experimental work required in various research fields, including chemistry,[1] physics,[2] drug discovery,[3,4] and environmental engineering.[5−7] QSAR models can also provide insights into the mechanisms underlying the relationship between chemical structures and their activities and/or properties.[8] The first step in developing QSAR models is to represent chemicals or molecules numerically, which is known as a molecular representation. Various types of molecular representations can be used, such as molecular fingerprint (MF),[8,9] molecular descriptor (MD),[10,11] molecular image (MI),[12−15] or molecular graph (MG),[3,16,17] because of the availability of powerful machine learning/deep learning tools capable of handling multiple data formats.

MF is also seen as one type of MD. Here, we specify the MF as the binary vectors (that is, containing only zeroes and ones) that encode the structural fragments or atom groups of a molecule (e.g., -OH and -COOH groups), while MD refers to the vectors that encode the bulk properties or physicochemical properties (e.g., molecular weight and charge). Compared with MD, MI, and MG, MF is generally easier to interpret and can be handled by traditional ML algorithms (e.g., random forest and ridge regression). It is also flexible to combine with other features, such as reaction conditions. For a detailed comparison between MF and other types of MD, MI, and MG, see Text S1 of the Supporting Information.

MF represents molecular structures as binary vectors with positions or bits, where the numbers in each bit indicate the presence or absence of specific atom groups in the molecule. MF can be divided into two major categories: structural key and hashed fingerprint. Structural key works by encoding a

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: March 23, 2023
Revised: June 16, 2023
Accepted: June 20, 2023



0	1	1	0	0	0	0	1	0	0
0	1	9	0	0	0	0	1	0	0
0	1	1	0	0	0	0	1	0	0
0	1	8	0	0	0	0	1	0	0
0	1	1	0	0	0	0	1	0	0
0	1	7	0	0	0	0	1	0	0

MF(r, l)	Similarity
B-MF(0,10)	1, 1, 1
B-MF(1, 100)	1, 1, 1
B-MF(2, 100)	1, 1, 1
B-MF(3, 1000)	1, 1, 1
C-MF(0,10)	0.952, 0.900, 0.947
C-MF(1, 100)	0.952, 0.900, 0.947
C-MF(2, 100)	0.949, 0.893, 0.943
C-MF(3, 1000)	0.944, 0.882, 0.937

Figure 1. Comparison between binary MF (B-MF) and count-based MF (C-MF) when representing 1-decanol, 1-nonanol, and 1-octanol. (A) Chemical structures of 1-decanol, 1-nonanol, and 1-octanol. (B) Example of B-MF and C-MF using the a radius of 0 and a length of 10. (C) Similarities calculated on the basis of B-MF and C-MF with various settings of radius and length, in which Dice similarity is used as the metric. The three similarities in each row correspond to the similarity of (1-decanol, 1-nonanol), (1-decanol, 1-octanol), and (1-nonanol, 1-octanol).

molecule's structure into binary vectors based on predefined structural features. For example, the Molecular ACCes System (MACCS) fingerprint is one of the most commonly used structural key fingerprints.[18] The MACCS fingerprint of a molecule can be obtained using the RDKit package, which uses a library of 166 predefined structural features (or keys) to generate a binary vector with a length of 166. However, MACCS may have limitations in encoding the structures that are not predefined, which could lead to challenges in predicting the activities and properties of new query molecules. Moreover, each bit in the MACCS fingerprint represents groups of atom groups rather than a unique or single atom group. Additionally, some of the keys in MACCS are not well-defined, such as key 44, which is defined as "OTHER" and may pose challenges for model interpretation. For more details, see https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py.

Hash fingerprint does not rely on any predefined structural features. Instead, it is generated by enumerating all possible fragments in a molecule and converting them into numeric values by a "hash" function, such as an atom-pair fingerprint (AP-MF), a topological torsion fingerprint (TP-MF), and a Morgan fingerprint. The Morgan fingerprint, also known as the extended-connectivity fingerprint (ECFP), is one of the most commonly used hash fingerprints. Rogers et al. and our recent papers have provided detailed explanations of how the Morgan fingerprint represents molecules.[11,19] In short, the Morgan fingerprint encodes the atom groups of a chemical into a binary vector with length and radius as its two parameters. The length refers to the lengths of the vector, while the radius represents the size of the atom groups. The larger the radius, the larger the atom group it can represent. Each bit in the vector represents a unique or single atom group. However, if the vector's length is too short, the same bit may represent two or several atom groups, resulting in bit collisions. The Morgan fingerprint of a molecule can be easily generated by the RDKit package, and the atom group in each bit can also be obtained by using the package, which facilitates model interpretation. The Morgan fingerprint has been widely used in many QSAR models.[20−23]

While the Morgan fingerprint is a useful representation for qualifying a molecule by identifying whether a certain atom group exists within it, it falls short in quantifying the number of such groups present. For instance, 1-decanol, 1-nonanol, and 1-octanol contain only three types of atom groups (that is, -OH, -CH$_2$-, and -CH$_3$) (Figure 1A); however, the binary Morgan fingerprints (B-MF) for these three alcohols yield the same result, as shown in Figure 1B, and the calculated similarities are all 1 (Figure 1C). Although the difference among these three alcohols lies in the number of -CH$_2$- groups they possess, B-MF fails to distinguish them because it lacks the necessary information about atom group counts. One may argue that increasing the radius of B-MF (e.g., radius = 4) to include larger atom groups would help to distinguish them. However, a larger radius often requires a longer length to avoid bit collisions when dealing with thousands of chemicals. Furthermore, it may lead to an overfitting problem by including many unnecessary atom groups. For MF, we often do not manually apply feature selection or reduction to avoid this issue (Text S2). Using B-MF can pose problems during the development and application of QSAR models. For instance, training a QSAR model based on B-MF would treat these three alcohols as identical, even though their activities and properties differ from one another. Consequently, the developed QSAR model may be less accurate. On the contrary, if a trained B-MF-based QSAR model is used to predict the activities and properties for these three alcohols, the predicted values will be the same for all three. However, in most cases, the true activities and properties of these three alcohols may be very similar but not entirely identical. Such differences cannot be captured by B-MF-based QSAR models. Hence, we conclude that including the count information about atom groups may benefit both the training and prediction performance. The count information about atom groups plays a vital role in some activities and properties. For example, CO_2 solubility in ionic liquids increases with the length of the alkyl chain because it decreases the density of ILs to let them have more free volume.[24−26] Similarly, the OH radical reacts with hydrocarbons in the air faster for those with a longer alkyl chain.[27]

To address the limitations of binary MF, we should include information about atom group quantities. Here, we introduce the count-based Morgan fingerprint (C-MF) that can count how many atom groups exist for each bit. The C-MFs for the aforementioned three alcohols are listed in Figure 1B, in which the third bit is filled with integers (>1) representing the number of -CH$_2$- groups. The similarities among these alcohols are not 1 anymore (Figure 1C), indicating that C-MFs can distinguish them. Considering the useful quantitative information C-MF contains, we surprisingly find that C-MF was seldom used in developing QSAR models.[28] Hence, in this study, we comprehensively investigate if C-MF can outperform B-MF in applications for developing QSAR models.

To this end, we developed B-MF- and C-MF-based QSAR models on 10 data sets related to water contaminants' activities and properties [that is, rate constants toward HO• and $SO_4^{\bullet-}$ radicals, solubility in water, and the soil organic carbon-normalized sorption coefficient of chemicals (K_{oc}), human peroxisome proliferator-activated receptor γ (PPARγ) binding affinity data on the dissociation constant and inhibition constant (pK_d), the concentration for the 50% maximum inhibition (pIC$_{50}$), collision cross sections for $[M + H]^+$ and $[M + H]^+$ (CCSM_H and CCSM_Na, respectively), lipophilicity (Lipo), and Free Solvation Database (FreeSolv)]. In addition to B-MF and C-MF, we also investigated the efficiency of count-based AP-MF and TT-MF[29,30] and tested six different ML algorithms, including ridge regression, support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), XGBoost, and CatBoost, to check if the efficiency of the fingerprint is ML dependent. We also compared the B-MF- and C-MF-based models in terms of model interpretation and the applicability domain. Finally, we deployed our QSAR models online for free use on the ContaminaNET platform at https://contaminanet.streamlit.app/. Overall, the results of our study suggest that C-MF is a powerful alternative to B-MF for developing QSAR models.

2. MATERIALS AND METHODS

2.1. Data Sets. Ten data sets related to water contaminants were compiled from previous studies.[2,5,21,31−33] A summary of these data sets is listed in Table 1, including the number of chemicals in each data set and their chemical diversities. The chemical diversity is measured by the average Dice similarity calculated by eq 1.[34]

Table 1. Summary of Four Data Sets Used in This Study

data set	no. of chemicals	B-MF-based diversity[a]	C-MF-based diversity[b]
$SO_4^{\bullet-}$	400	0.353	0.294
HO•	1374	0.324	0.251
K_{oc}	752	0.419	0.365
solubility	1395	0.365	0.299
pK_d	420	0.737	0.693
pIC$_{50}$	1316	0.687	0.691
CCS_MH	1076	0.487	0.397
CCS_MNa	645	0.526	0.391
Lipo	4200	0.571	0.557
FreeSolv	642	0.309	0.249

[a]The calculated average Dice similarity based on B-MF (0, 2048). [b]The calculated average Dice similarity based on C-MF (0, 2048).

$$S_{A,B} = \frac{2\sum_{j=1}^{n} X_{jA}X_{jB}}{\sum_{j=1}^{n}(X_{jA})^2 + \sum_{j=1}^{n}(X_{jB})^2} \quad (1)$$

where X_{jA} represents the jth feature of molecule A and X_{jB} represents the jth feature of molecule B. Tanimoto similarity is not used here because C-MF is not applicable.[31] Apart from the pIC$_{50}$ data set, the diversity calculated from C-MF is higher than that from B-MF for others, with lower average similarity indicating greater diversity. This is expected because C-MF can distinguish molecules containing the same types of atom groups but with different counts. The chemical diversity of the 10 data sets decreased in the following order: FreeSolv > HO• > $SO_4^{\bullet-}$ > solubility > K_{oc} > CCSM_Na > CCSM_H > pIC$_{50}$ > pK_d. All data sets contain the SMILES of chemicals and their activities and properties, and we removed inorganic salts from SMILES only if they existed without further preprocessing, as previous studies had curated the data set well. More details about data collection and preprocessing can be found in related studies.[2,5,21,31−33] We selected these 10 data sets without any specific criteria but with the aim of enriching their diversities in terms of the number of chemicals, molecular diversity, and targets. This approach allows us to present our findings in a more generalizable manner. Further details are provided in Text S3. $k_{SO_4^{\bullet-}}$ and $k_{\bullet OH}$ are second-order rate constants of contaminants toward oxidants of HO• and $SO_4^{\bullet-}$ radicals, respectively.[11] They are important parameters for optimizing advanced oxidation processes. pH often affects the distribution of dissociated species of contaminants, which, in turn, affects the measurements of rate constants. pH is not included in the source data sets of $k_{SO_4^{\bullet-}}$ and $k_{\bullet OH}$, which may affect the model performance.[11] Here, we focused on comparing the C-MF-based models with the B-MF-based models in terms of model performance on the same data set. K_{oc} measures how easily the organic chemicals adsorb on the soils and sediments, which is important for determining their ultimate environmental fate and transport.[32] The solubility of contaminants in water determines their bioaccessibility/bioavailability and their fate and transport in aquatic environments.[31] pK_d and pIC$_{50}$ measure the binding affinity of PPARγ, a ligand-activated nuclear receptor involved in multifaceted physiology and chemically induced endocrine disruption.[5] CCS, derived from ion mobility separation (IMS), is a physicochemical property of ions and is related to the chemical structure and three-dimensional conformation of the molecules.[33] FreeSolv provides the experimental hydration free energy of small molecules in water,[35] while Lipo offers experimental results of the octanol/water distribution coefficient (log D at pH 7.4) of 4200 compounds.[36] Developing QSAR models that accurately predict these activities and properties for contaminants is significantly important.

2.2. Fingerprint Generation. All of the types of fingerprints of a molecule were generated from its SMILES by the RDKit package in Python. B-MF and C-MF were obtained using the "AllChem.GetMorganFingerprintAsBitVect" and "AllChem.GetHashedMorganFingerprint" commands, respectively. Count-based AP-MF and TT-MF were generated using the "AllChem.GetHashedAtomPairFingerprint" and "AllChem.GetHashedTopologicalTorsionFingerprint" commands, respectively. Note that the RDKit package also provides binary count-based AP and TT fingerprints that can be obtained using the "AllChem.GetHashedAtomPairFingerprintAsBitVect" and "AllChem.GetHashedTopologicalTorsionFingerprintAsBitVect" commands. However, unlike B-MF and C-MF, the non-zero bits in binary and count-based AP and TT fingerprints are not one to one (as shown in Figure S1). These binary fingerprints are simulating count fingerprints, which includes the count information by using multiple bits (four bits in RDKit) to record the count of each feature, similar to one-hot encoding. For example, the four bits of [1000] and [0100] represent a molecule containing one and

https://doi.org/10.1021/acs.est.3c02198

two of the same atom groups, respectively. Simulated count fingerprints of AP-MF and TT-MF are thus 4 times larger than the normal count-based fingerprint. Therefore, we did not use them for AP-MF and TT-MF. Moreover, RDKit does not provide interpretation methods for AP-MF and TT-MF to show what the atom groups represent and what each bit represents. Although we can still apply the model interpretation method to the AP-MF- and TT-MF-based models to show how each bit affects them, we do not have the atom group information for each bit so that how atom groups affect the models remains unknown.

2.3. Model Development, Interpretation, Applicability Domain Analysis, and Deployment. For all of the data sets, the development of the ML model followed a standardized procedure consisting of the following steps.

(1) The data set was randomly split into training and test sets with an 80:20 ratio. The training set was used to train the ML models, while the test set was reserved for evaluating the generalization ability of the models. The test set is not used during the ML model development. The performance of the model was evaluated using the root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2), with lower RMSE and MAE values and higher R^2 values indicating better model performance.

(2) Six ML algorithms, i.e., ridge regression, SVM, KNN, RF,[37] XGBoost,[38] and CatBoost,[39] were used to test if the efficiency of the fingerprint type is dependent on the ML algorithm. The candidate hyperparameters for each ML are listed in Table S1. The length of the fingerprint (i.e., the number of features) may be larger than the data size, which will lead to the "curse of dimensionality". The algorithms we applied here can conduct feature selection automatically through tuning the hyperparameters that control overfitting; that is, not all of the features were used during training. In other words, we overcome the "curse of dimensionality" by feature selection. More discussion is presented in Text S4. Similar to our previous studies, we employed Bayesian optimization algorithms[40] and 5-fold cross-validation on the training set to tune the hyperparameters. Maximizing average validation performance is the target for Bayesian optimization. Bayesian optimization will iteratively select the next candidate hyperparameters based on the last selected ones. If the validation performance does not increase over 300 iterations during the Bayesian optimization process, we will stop optimization. To better control the overfitting, we plotted the training and validation performances to determine the optimum hyperparameters (more details are discussed in section 3.1). After obtaining optimum hyperparameters, we retrained ML models on the whole training set (without cross-validation anymore) to obtain the final ML models. The generalization ability or predictive ability of the obtained ML models was evaluated on the test set.

(3) The data were randomly split five times using different seeds in Python to obtain five groups of training and test sets, which ensure that the results are not accidental. All of the models were developed and tested on these same five groups of training and test sets. The model performance is thus the average predictive training and test performance with a standard deviation.

(4) Following our previous studies,[11] we interpreted these ML models using the SHAP method to check the differences between B-MF and C-MF.[41]

(5) Following our previous studies,[8] we chose these ML models trained on the same training set to determine their applicability domain by the similarity method. Specifically, we compared the Dice similarity between the test set and the training set and found a threshold similarity that can exclude minimum molecules in the test set and reach the minimum recalculated $\mathrm{RMSE}_{\mathrm{test}}$. This threshold similarity is the AD of the model. If the query chemicals are below this threshold similarity, they are outside the AD and the model cannot reliably predict them. Otherwise, they are within the AD and the model can be used.[37]

(6) Finally, we deployed our models on the ContaminaNET platform, which is available at https://contaminanet.streamlit.app/.

3. RESULTS AND DISCUSSION

3.1. Hyperparameter Tuning and Overfitting Control. The hyperparameters that can enable ML models to achieve the minimum average RMSE for validation sets ($\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$) (if cross-validation is used) are often determined as the optimum hyperparameters for the model. However, we found that this is not always true and that it sometimes can lead to overfitting problems. Figure 2 shows the results of hyperparameter tuning by Bayesian optimization, taking the solubility data set and CatBoost algorithm as examples. During the optimization process, more than 500 groups of hyperparameters were selected by the Bayesian optimization algorithm and then tested to obtain more than 500 groups of average RMSE for training sets ($\mathrm{RMSE}_{\mathrm{CV\text{-}Train}}$) and $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$. As shown in Figure 2A, with a decrease in $\mathrm{RMSE}_{\mathrm{CV\text{-}Train}}$, $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$ gradually decreased initially and then leveled out or even increased finally. The black dotted line indicates the position of the minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$. Figure 2B shows the enlarged pattern around the minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$. C1 marks the hyperparameters that achieved the lowest minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$, while C2 marks the hyperparameters that achieved the third lowest minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$. The degree of decline trend for $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$ before position C2 was much higher than that after position C2 (the orange dotted line). More details are listed in Table 2, in which decreasing $\mathrm{RMSE}_{\mathrm{CV\text{-}Train}}$ from

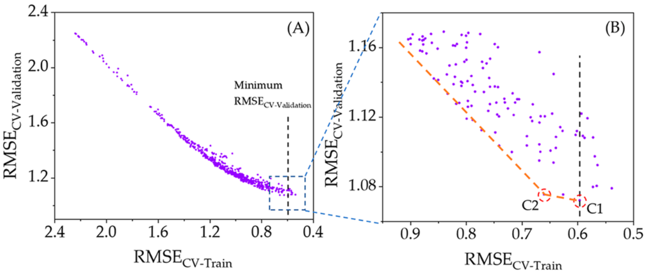


Figure 2. Plots of $\mathrm{RMSE}_{\mathrm{CV\text{-}Train}}$ vs $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$ from the Bayesian optimization results in which each dot represents a group of hyperparameters. (A) Global pattern. (B) The enlarged pattern is located at the minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Train}}$. The solubility data set and CatBoost algorithm are used as examples. The black dotted line is the position of the minimum $\mathrm{RMSE}_{\mathrm{CV\text{-}Validation}}$.

Table 2. $RMSE_{CV\text{-}Train}$ and $RMSE_{CV\text{-}Validation}$ Based on C1 and C2 Hyperparameters and the Performance of the Corresponding Developed Models on Training and Test Sets

hyperparameter	$RMSE_{CV\text{-}Train}$	$RMSE_{CV\text{-}Validation}$	$RMSE_{train}$	$RMSE_{test}$	R^2_{train}	R^2_{test}
C1	0.596748	1.072243	0.640	1.14	0.92	0.78
C2	0.660338	1.075597	0.686	1.09	0.91	0.80

Table 3. Optimal Fingerprint Type for Each Data Set and ML Algorithm and the Corresponding Model Performance

data set	ML algorithm	fingerprint type	$RMSE_{train}$	MAE_{train}	R^2_{train}	$RMSE_{test}$	MAE_{test}	R^2_{test}
$SO_4^{\bullet-}$	CatBoost	B-MF	0.069 ± 0.017	0.051 ± 0.010	0.821 ± 0.089	0.112 ± 0.013	0.078 ± 0.005	0.514 ± 0.092
HO•	CatBoost	C-MF	0.055 ± 0.003	0.039 ± 0.003	0.856 ± 0.016	0.089 ± 0.007	0.060 ± 0.003	0.640 ± 0.041
K_{oc}	ridge	C-MF	0.332 ± 0.024	0.257 ± 0.020	0.925 ± 0.009	0.507 ± 0.019	0.387 ± 0.019	0.830 ± 0.022
solubility	CatBoost	C-MF	0.389 ± 0.009	0.299 ± 0.009	0.971 ± 0.003	0.692 ± 0.033	0.524 ± 0.020	0.909 ± 0.009
pK_d	CatBoost	C-MF	0.413 ± 0.064	0.333 ± 0.056	0.868 ± 0.042	0.729 ± 0.019	0.582 ± 0.020	0.577 ± 0.064
pIC_{50}	CatBoost	C-MF	0.372 ± 0.010	0.290 ± 0.009	0.889 ± 0.007	0.636 ± 0.025	0.488 ± 0.013	0.675 ± 0.023
CCS_MH	ridge	C-MF	3.070 ± 0.109	2.255 ± 0.093	0.992 ± 0.001	5.285 ± 0.458	3.638 ± 0.195	0.976 ± 0.004
CCS_MNa	ridge	C-MF	2.717 ± 0.511	2.058 ± 0.397	0.992 ± 0.003	6.480 ± 1.095	4.598 ± 0.602	0.952 ± 0.021
Lipo	ridge	C-MF	0.397 ± 0.028	0.301 ± 0.021	0.891 ± 0.015	0.683 ± 0.020	0.499 ± 0.018	0.676 ± 0.018
FreeSolv	ridge	C-MF	0.489 ± 0.044	0.356 ± 0.036	0.984 ± 0.004	1.230 ± 0.121	0.764 ± 0.044	0.893 ± 0.027

Table 4. Enhancements of C-MF-Based Models Relative to B-MF-Based Models in Terms of $RMSE_{test}$

data set	$SO_4^{\bullet-}$	HO•	K_{oc}	solubility	pK_d	pIC_{50}	CCS_MH	CCS_MNa	Lipo	FreeSolv
enhancement (%)	−2.1	5.9	27.4	36.7	5.6	0.8	68.8	69.2	6.7	19.6

0.660338 to 0.596748 merely enabled $RMSE_{CV\text{-}Validation}$ to decrease from 1.075597 to 1.072243. Such a decrease in $RMSE_{CV\text{-}Validation}$ is marginal compared with the decrease in $RMSE_{CV\text{-}Train}$. In other words, the ML model fitted more noise or irrelevant information from C2 to C1. Hence, C1 should have a more serious overfitting problem than C2. As demonstrated in Table 2, CatBoost trained with C1 hyperparameters achieved a lower RMSE on the training set ($RMSE_{train}$) but a higher RMSE on the test set ($RMSE_{test}$) than that trained with C2 hyperparameters. Hence, across this study, we plotted the hyperparameter tuning results to select the hyperparameters to control overfitting rather than directly selecting the hyperparameters that achieve a minimum $RMSE_{CV\text{-}Validation}$.

3.2. Comparison of C-MF to Other Types of MF in Terms of Model Predictive Performance.

We summarized the performance of the models on the training set ($RMSE_{train}$, MAE_{train}, and R^2_{train}) and the test set ($RMSE_{test}$, MAE_{test}, and R^2_{test}) for each data set, fingerprint type, and ML algorithm in Table S2. $RMSE_{test}$, MAE_{test}, and R^2_{test} measure the model's generalization ability. Typically, $RMSE_{train}$ and MAE_{train} are lower than $RMSE_{test}$ and MAE_{test}, respectively, while R^2_{train} is higher than R^2_{test} because the model is trained on the training set. Thus, the weights or parameters of the model are more optimized for the training set than for the test set. The difference between $RMSE_{train}$ and $RMSE_{test}$ indicates the overfitting trend, where a larger difference implies more serious overfitting. The overfitting trend can also be represented by the difference between MAE_{train} and MAE_{test} or between R^2_{train} and R^2_{test}. To eliminate the effect of randomness during data splitting, we repeated the data splitting process five times and calculated the average values and standard deviations for $RMSE_{train}$, MAE_{train}, R^2_{train}, $RMSE_{test}$, MAE_{test}, and R^2_{test}. Further details of the model performance for each data splitting are listed in the Excel file in the Supporting Information. To demonstrate these five groups are representative of the characteristics of the entire data set, we used the Uniform Manifold Approximation and Projection (UMAP) algorithm to visualize the molecules in a two-dimensional plot (Figure S2), taking the K_{oc} data set as an example. Figure S2A shows the distribution of molecules in the K_{oc} data set, in which several "islands" can be identified (red circle). The molecules in these five training and test sets are uniformly sampled from the space of all molecules in the K_{oc} data set, indicating that they are representative of the data set's characteristics.

We checked the optimal fingerprint type for each data set based on $RMSE_{test}$. AP-MF is the optimal fingerprint for $SO_4^{\bullet-}$ and pK_d data sets, while C-MF is the optimal choice for the other eight data sets (Table S2). This suggests that including the counts of atom groups in MF (i.e., C-MF) can improve model performance, although this may vary depending on the data set. However, we did not discuss the applications of AP-MF and TT-MF because they are not interpretable, as explained in section 2.2. Nonetheless, if model interpretability is not a concern, then AP-MF or TT-MF could also be considered. In our subsequent analysis, we will focus on only comparing B-MF and C-MF.

We listed the optimal fingerprint type for each data set and each ML algorithm in Table 3. C-MF outperformed B-MF in nine of 10 data sets in terms of $RMSE_{test}$. For the $SO_4^{\bullet-}$ data set, there is only a marginal difference in performance between the C-MF-based model ($RMSE_{test}$ = 0.115) and the B-MF-based model ($RMSE_{test}$ = 0.112) in terms of $RMSE_{test}$ (Table S2). Upon comparison in terms of MAE_{test}, the C-MF-based model (MAE_{test} = 0.077) outperforms the B-MF-based model (MAE_{test} = 0.078), but again, the difference is still marginal. Therefore, we can conclude that both C-MF and B-MF are equally suitable for developing models for this data set. However, the advantage of C-MF over B-MF varied with the ML algorithm used. For example, C-MF underperforms B-MF in terms of $RMSE_{test}$ for the pIC_{50} data set with ridge regression, the HO• data set with SVM, and the pK_d data set with SVM (Table S2). This could be due to the increased complexity of QSAR models caused by C-MF's extra information on atom counts, which may not be adequately

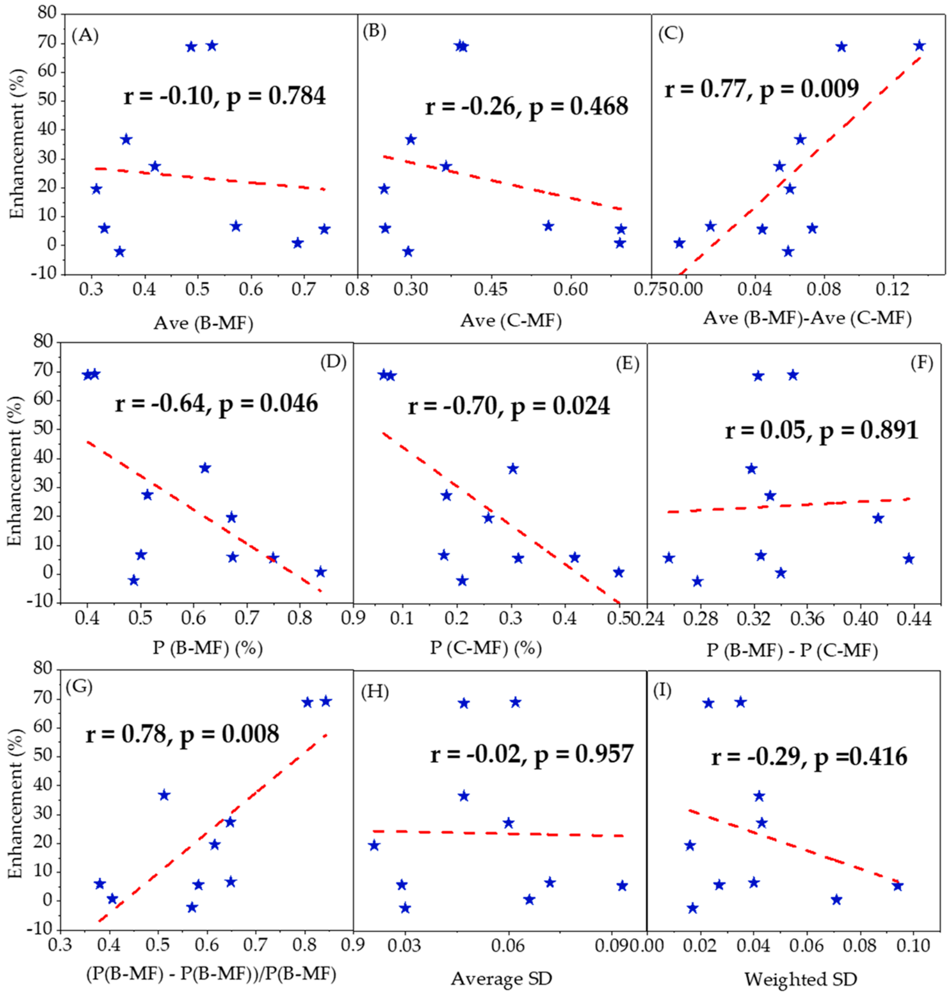


Figure 3. Correlation plots between performance enhancement and different metrics: (A) B-MF-based diversity [Ave (B-MF)], (B) C-MF-based diversity [Ave (C-MF)], (C) Ave (B-MF) − Ave (C-MF), (D) percentage of molecules that cannot be distinguished by B-MF [*P* (B-MF) (%)], (E) percentage of molecules that cannot be distinguished by C-MF [*P* (C-MF) (%)], (F) *P* (B-MF) (%) − *P* (C-MF) (%), (G) [*P* (B-MF) (%) − *P* (C-MF) (%)]/[*P* (B-MF) (%)], (H) average standard deviation (SD), and (I) weighted SD. The *p* value is calculated by the equation $p = 2 \times P(T > t)$, where $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (*n* is the number of samples, and *r* is the Pearson correlation coefficient). A *p* of <0.05 indicates the Pearson correlation coefficient is statistically significant. However, it should be noted that the correlation analysis based on 10 samples may not be adequate.

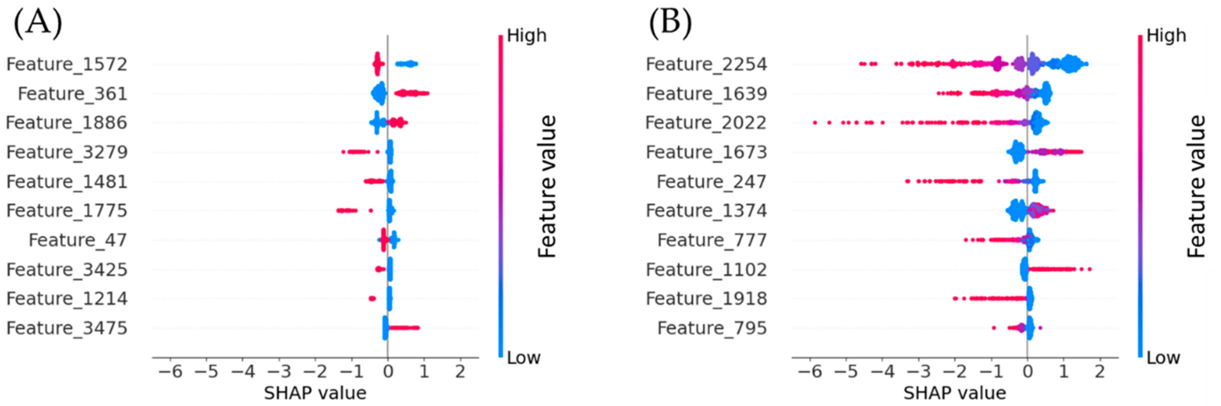


Figure 4. SHAP plots of the (A) B-MF-based model and (B) C-MF-based model for the solubility data set. We set the same range for *x*-axis values for each data set, which is more convenient for comparing the range of Shapley values. The optimum radius and length for B-MF are 2 and 3622, respectively, while those for C-MF were 0 and 2894, respectively, both of which were obtained by Bayesian optimization. Feature_# represents the position of a vector, that is, a specific atom group. Different features in different settings of radius and length may represent the same atom group. For example, Feature_1572 in B-MF and Feature_2254 in C-MF represent the same aromatic carbon. The atom groups represented by each bit in MFs for each data set can be obtained via https://contaminanet.streamlit.app/.

Table 5. Comparison of ADs for B-MF-Based and C-MF-Based Models for All Data Sets

data set	B-MF AD	C-MF AD
$SO_4^{\bullet-}$	0.055	0.040
HO•	0.020	0.010
K_{oc}	0.032	0.035
solubility	0.020	0.060
pK_d	0.248	0.176
pIC_{50}	0.123	0.133
CCS_MH	0.067	0.093
CCS_MNa	0.101	0.080
Lipo	0.124	0.139
FreeSolv	0.031	0.031

captured by ridge regression and SVM, resulting in underfitting. Alternatively, their capacity to control overfitting may be insufficient, resulting in overfitting. Ridge regression and CatBoost were identified as the optimal ML algorithms for all data sets, with Ridge regression being preferred for the K_{oc}, CCSM_H, CCSM_Na, Lipo, and FreeSolv data sets, while CatBoost was used for the others.

We then calculated the improvement in model performance measured by RMSE using eq 2, and the results are presented in Table 4. The performance improvement achieved with C-MF varied across different data sets, with the CCS_MNa data set showing the largest improvement and $SO_4^{\bullet-}$ showing the smallest.

$$\text{enhancement}\,(\%) = \frac{\text{RMSE}_{\text{test}}(\text{B-MF}) - \text{RMSE}_{\text{test}}(\text{C-MF})}{\text{RMSE}_{\text{test}}(\text{B-MF})} \times 100\% \qquad (2)$$

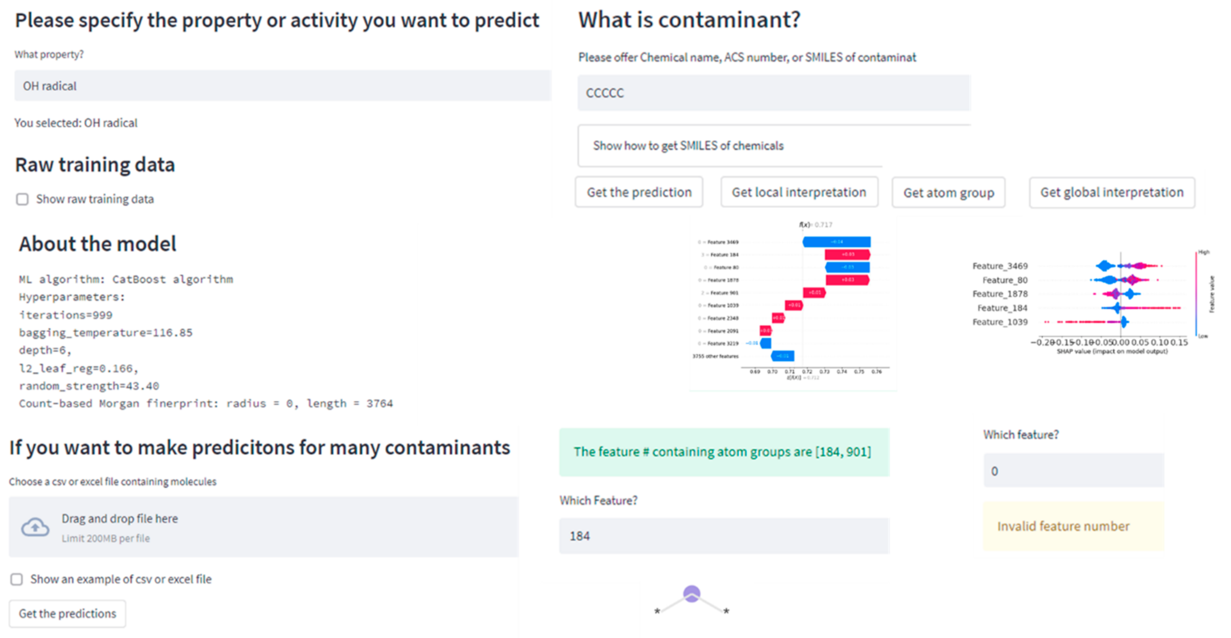
The various improvements that C-MF provides to the model performances offer us an opportunity to investigate when C-MF is superior to B-MF in developing models, which can guide us to choose the appropriate molecular fingerprint for other data sets. We employed nine metrics to correlate them with the performance enhancements, namely, B-MF-based diversity [Ave (B-MF)], C-MF-based diversity [Ave (C-MF)], Ave (B-MF) − Ave (C-MF), the percentage of molecules that cannot be distinguished by B-MF [*P* (B-MF) (%)], the percentage of molecules that cannot be distinguished by C-MF [*P* (C-MF) (%)], *P* (B-MF) (%) − *P* (C-MF) (%), [*P* (B-MF) (%) − *P* (C-MF) (%)]/[*P* (B-MF) (%)], average standard deviation (SD), and weighted SD. B-MF-based diversity [Ave (B-MF)] and C-MF-based diversity are calculated on the basis of eq 1. The percentage of molecules that cannot be distinguished by B-MF [*P* (B-MF) (%)] and the percentage of molecules that cannot be distinguished by C-MF [*P* (C-MF) (%)] are calculated by eq 3

$$\text{percentage}\,(\%) = \frac{\sum_{i=0}^{n} m_i}{N} \qquad (3)$$

where m_i is the number of chemicals with a B-MF-based similarity of 1 in each group, n is the number of groups (an example is shown in Table S3), and N is the total number of chemicals. SD refers to the SD of their activities and properties. Because we have different groups of chemicals that have a similarity of 1 with each other (Table S3), each group will offer us a standard deviation. We then took the average value for all of these groups as the average SD determined by eq 4.

$$\text{average SD} = \frac{\sum_{i=0}^{n} \text{SD}_i}{n} \qquad (4)$$

where SD_i represents the standard deviation of each group. It should be noted that because the range of activities and properties for each data set is different, we scaled them to the same range of 0−1 to make the standard deviation comparable before calculating them. The average SD assigns the same importance to each group of chemicals, but the number of chemicals in each group is different (Table S3, 10 chemicals in group 1 and eight chemicals in group 2). We then developed a weighted SD that assigns the contributions of each group on



Please specify the property or activity you want to predict

What property?

OH radical

You selected: OH radical

Raw training data

□ Show raw training data

About the model

```
ML algorithm: CatBoost algorithm
Hyperparameters:
iterations=999
bagging_temperature=116.85
depth=6,
l2_leaf_reg=0.166,
random_strength=43.40
Count-based Morgan finerprint: radius = 0, length = 3764
```

If you want to make predicitons for many contaminants

Choose a csv or excel file containing molecules

Drag and drop file here
Limit 200MB per file

□ Show an example of csv or excel file

What is contaminant?

Please offer Chemical name, ACS number, or SMILES of contaminat

CCCCC

The feature # containing atom groups are [184, 901]

Which Feature?

184

Which feature?

0

Invalid feature number

Figure 5. Overview of the ContaminaNET platform.

the basis of the percentage of accounts for the total number of chemicals (eq 5).

$$\text{weighted SD} = \sum_{i=0}^{n} x_i \text{SD}_i, \text{ where } x_i = \frac{n_i}{N} \quad (5)$$

We then plotted these nine metrics against the performance enhancement, shown in Figure 3. First, the performance enhancements were not related to the structural diversity of the data set [Ave (B-MF) and Ave (C-MF)] (Figure 3A,B; *r* = −0.10, and *r* = −0.26), but it positively correlated with the difference between them [Ave (B-MF) − Ave (C-MF)] (Figure 3C; *r* = 0.77). This metric reflects the disagreement in the data set diversity calculated by B-MF and C-MF. The larger the disagreement, the better it is to use C-MF for model development. Second, it negatively correlated with the percentage of molecules that cannot be distinguished by B-MF [Figure 3D; *P* (B-MF) (%); *r* = −0.64] and C-MF [Figure 3E; *P* (C-MF) (%); *r* = −0.70]. However, it is counterintuitive when we observed that the metric of *P* (B-MF) (%) − *P* (C-MF) (%) did not correlate with the performance enhancement because this metric reflects the percentage of molecules that cannot be distinguished by B-MF but can be distinguished by C-MF. We anticipated that the more molecules that can be distinguished by C-MF, the better the C-MF-based model's performance. We then calculated how these molecules account for the molecules that cannot be distinguished by B-MF {the metric [*P* (B-MF) (%) − *P* (C-MF) (%)]/*P* (B-MF) (%)} and found that it positively correlated with the performance enhancements (Figure 3G; *r* = 0.78). The average SD and weighted SD did not show an obvious trend with the performance enhancements (Figure 3H,I; *r* = −0.02, and *r* = −0.29). This was unexpected because if their activities and properties differ greatly from each other (i.e., large standard deviation), C-MF is more efficient in distinguishing them than B-MF, which is beneficial for improving model performance. Overall, the performance enhancements that C-MF brings are positively correlated with Ave (B-MF) − Ave (C-MF) and [*P* (B-MF) (%) − *P* (C-MF) (%)]/[*P* (B-MF) (%)], which can guide us in choosing the appropriate molecular fingerprint for other data sets. By calculating these two metrics, we can easily determine whether C-MF is a better choice over B-MF.

3.3. Comparison of C-MF to B-MF in Terms of Model Interpretation.

We took the solubility data set as an example to illustrate the differences between B-MF- and C-MF-based models in terms of model interpretation. Explanations for the other nine data sets are provided in Text S5. As in our previous studies, we used the SHAP interpretation method to interpret these ML-based QSAR models, which provides information about how each atom group affects the target. Figure 4A shows the SHAP plot for the B-MF-based model, which contains only two colors: red and blue because B-MF contains only zeroes and ones. In B-MF, 1 is the highest value (red) while 0 is the lowest value (blue). Figure 4B shows the SHAP plot for the C-MF-based model, which contains a transition color from blue to red. This is because C-MF contains the counts of atom groups, and the numbers in each bit can exceed 1. In other words, 1 is no longer the highest value (0 is still the lowest value) in C-MF. Hence, we can see the effect of atom group counts on the target for C-MF, which cannot be achieved for B-MF. For example, Feature_1572 in the B-MF-based model and Feature_2254 in the C-BF-based model represent the same aromatic carbon. We only know from the SHAP plot of the B-MF-based model that the existence of aromatic carbon can decrease the solubility of chemicals. In contrast, we also know from the SHAP plot of the C-MF-based model that the increasing counts of aromatic carbons can lead to chemicals that are more insoluble in water, which is consistent with our domain knowledge or experience because aromatic rings are hydrophobic.[42] Similarly, Feature_361 in the B-MF-based model and Feature_1673 in the C-MF-based model represent the same -OH group, which is a hydrophilic group.[43] The SHAP plot of the C-MF-based model also shows that increasing the number of -OH groups in a chemical can increase its aqueous solubility, while B-MF cannot offer this information. Furthermore, the range of SHAP values in the C-MF-based SHAP plot is much larger than that in the B-MF-based SHAP plot. This may be the reason why the C-MF-based model is more accurate than the B-MF-based model because each feature has more space to dynamically affect the target. With regard to the ranking of feature importance, it is generally observed that they do not follow the same order. Moreover, it is challenging to determine which one is superior because the sizes of atom groups differ between B-MF and C-MF. The Bayesian optimization algorithm determines the radius and length of the fingerprint.

We next show how to use our developed web application to interpret the effects of atoms on contaminants' properties for each model, as shown in Figure S3. After specifying the model to interpret, we would find the details about the model and the parameters of C-MF. By generating a SHAP plot and inputting any feature number, we can identify the corresponding atom groups. For example, we selected the model for the HO● data set (Figure S3) and were interested in Feature_1039, which decreases the reaction rate constants and increases its number, which leads to a further decrease in the reaction rates. We found that Feature_1039 corresponds to electron-withdrawing -Cl atoms. Consequently, the presence of these groups is expected to decrease the reaction rate constants, and an increase in their number would further decrease the reaction rate constants, consistent with established domain knowledge. We discussed the effects of atom groups on other properties from other models in Text S5.

3.4. Comparison of C-MF to B-MF in Terms of the Applicability Domain (AD).

To compare AD for B-MF and C-MF, we select models developed on the basis of group 2 (Excel file) for all of the data sets and will deploy them on the web in the following section. It should be noted that the prediction performance of C-MF-based models is better than that of the B-MF-based ones. From the perspective of model application, we should consistently use C-MF-based models regardless of how their ADs change relative to the B-MF-based model. However, we still compare them to investigate whether they have a similar AD. Similar to our previous studies,[8] the AD of a model is determined by the threshold similarity. The lower the threshold similarity, the larger the AD of the model. The AD determination for each data set and each model is listed in Table S4, and we summarize the results in Table 5. The ADs of B-MF-based models are larger than those of C-MF-based models for the data sets of K_{oc}, pIC_{50}, Lipo, and CCSM_H, while for the other six data sets, the ADs of C-MF-based models are larger. For the following model deployment, these ADs will be used to judge if the query chemicals can be predicted by the models.

3.5. Model Deployment on the ContaminaNET Platform.

To enable researchers to easily access and use models, we developed a platform called "ContaminaNET" (https://contaminanet.streamlit.app/) (Figure 5). The ContaminaNET platform has a user-friendly interface and currently supports the predictions for the 10 activities and properties mentioned above, in which all of the models are developed on the basis of C-MF. It supports not only single predictions but also batch predictions. The user needs to supply only the CAS number, chemical name, or SMILES of the query chemical to the platform for a single prediction or supply an EXCEL or CSV file for batch predictions. It will automatically obtain SMILES (if not provided) for query chemicals, check if they are within the model's AD, and make predictions if they are within the AD. It supports global interpretation for the developed model, and users can check how each atom group affects the model prediction. Local interpretation for a single prediction is also supported, which will offer information about how the model makes predictions on the basis of the query chemical structure.

4. ENVIRONMENTAL IMPLICATIONS

Developing more accurate predictive models is crucial not only in the environmental field but also in other fields, such as drug discovery and chemistry. Morgan fingerprint, an established and widely used molecular structure representation, has been a benchmark for comparing other representations like molecular descriptors or graphs. In this study, we found that the C-MF has the potential to develop predictive models for a data set that involves chemicals. We demonstrated that C-MF is the preferred option for developing more accurate predictive models, as it more effectively represents molecular structures. B-MF cannot sensitively reflect slight or substantial changes in a molecule's properties or activities resulting from the addition or reduction of the same atom group, whereas C-MF overcomes this drawback by including information about atom group counts. SHAP interpretation revealed that C-MF-based models have a more flexible contribution of each atom group to the target, explaining why C-MF-based models generally exhibit better predictive performance. Therefore, C-MF is expected to have broad applications not only in the environmental field but also in other fields involving chemicals or molecules. Finally, the development of the "ContaminaNET" platform aims to provide a practical tool for researchers without modeling experience. The platform will continue to be updated with more advanced models or cover more activities and properties.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.3c02198.

Details for the comparison between MF and MD, MG, and MI; explanations for feature selection or reduction on the length of MF; data set selection; curse of dimensionality; model interpretation for other data sets; range of candidate hyperparameters for each ML algorithm; model performances for each data set, ML algorithm, and fingerprint type; AD determination for B-MF- and C-MF-based models for all data sets; an illustration of how to conduct model interpretation in our web app; and other related tables and figures (PDF)

Additional data (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Xiaohong Guan − *Department of Environmental Science, School of Ecological and Environmental Sciences, East China Normal University, Shanghai 200241, P. R. China;* orcid.org/0000-0001-5296-423X; Email: xhguan@des.ecnu.edu.cn

Author

Shifa Zhong − *Department of Environmental Science, School of Ecological and Environmental Sciences, East China Normal University, Shanghai 200241, P. R. China;* orcid.org/0000-0002-5822-0837

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.3c02198

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant 22076143).

■ REFERENCES

(1) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and accurate prediction of p K a values of C−H acids using graph convolutional neural networks. *J. Am. Chem. Soc.* **2019**, *141* (43), 17142−17149.

(2) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9* (2), 513−530.

(3) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021**, *13* (1), 1−23.

(4) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370−3388.

(5) Wang, Z.; Chen, J.; Hong, H. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. *Environ. Sci. Technol.* **2021**, *55* (10), 6857−6866.

(6) Roy, J.; Roy, K. Assessment of toxicity of metal oxide and hydroxide nanoparticles using the QSAR modeling approach. *Environmental Science: Nano* **2021**, *8* (11), 3395−3407.

(7) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.

(8) Zhong, S.; Zhang, K.; Wang, D.; Zhang, H. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. *Chemical Engineering Journal* **2021**, *405*, 126627.

(9) Myint, K.-Z.; Wang, L.; Tong, Q.; Xie, X.-Q. Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. *Mol. Pharmaceutics* **2012**, *9* (10), 2912−2923.

(10) Borhani, T.; Saniedanesh, M.; Bagheri, M.; Lim, J. QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Res.* **2016**, *98*, 344−353.

(11) Zhong, S.; Zhang, Y.; Zhang, H. Machine Learning-Assisted QSAR Models on Contaminant Reactivity Toward Four Oxidants:

Combining Small Data Sets and Knowledge Transfer. *Environ. Sci. Technol.* **2022**, *56* (1), 681−692.

(12) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chemical Engineering Journal* **2021**, *408*, 127998.

(13) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv* **2017**, DOI: 10.48550/arXiv.1706.06689.

(14) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. *J. Chem. Inf. Model.* **2018**, *58* (8), 1533−1543.

(15) Shi, T.; Yang, Y.; Huang, S.; Chen, L.; Kuang, Z.; Heng, Y.; Mei, H. Molecular image-based convolutional neural network for the prediction of ADMET properties. *Chemometrics and Intelligent Laboratory Systems* **2019**, *194*, 103853.

(16) Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **2022**, *4* (2), 127−134.

(17) Korolev, V.; Mitrofanov, A.; Korotcov, A.; Tkachenko, V. Graph convolutional neural networks as "general-purpose" property predictors: the universality and limits of applicability. *J. Chem. Inf. Model.* **2020**, *60* (1), 22−28.

(18) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002**, *42* (6), 1273−1280.

(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(20) Ding, Y.; Chen, M.; Guo, C.; Zhang, P.; Wang, J. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. *J. Mol. Liq.* **2021**, *326*, 115212.

(21) Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. py SiRC": Machine Learning Combined with Molecular Fingerprints to Predict the Reaction Rate Constant of the Radical-Based Oxidation Processes of Aqueous Organic Contaminants. *Environ. Sci. Technol.* **2021**, *55* (18), 12437−12448.

(22) Gao, Y.; Zhong, S.; Torralba-Sanchez, T. L.; Tratnyek, P. G.; Weber, E. J.; Chen, Y.; Zhang, H. Quantitative structure activity relationships (QSARs) and machine learning models for abiotic reduction of organic compounds by an aqueous Fe(II) complex. *Water Res.* **2021**, *192*, 116843.

(23) Zhang, K.; Zhang, H. Predicting solute descriptors for organic chemicals by a deep neural network (DNN) using basic chemical structures and a surrogate metric. *Environ. Sci. Technol.* **2022**, *56* (3), 2054−2064.

(24) Aki, S. N.; Mellein, B. R.; Saurer, E. M.; Brennecke, J. F. High-pressure phase behavior of carbon dioxide with imidazolium-based ionic liquids. *J. Phys. Chem. B* **2004**, *108* (52), 20355−20365.

(25) Yunus, N. M.; Mutalib, M. A.; Man, Z.; Bustam, M. A.; Murugesan, T. Solubility of CO2 in pyridinium based ionic liquids. *Chemical engineering journal* **2012**, *189*, 94−100.

(26) Huang, X.; Margulis, C. J.; Li, Y.; Berne, B. J. Why is the partial molar volume of CO2 so small when dissolved in a room temperature ionic liquid? Structure and dynamics of CO2 dissolved in [Bmim+][PF6-]. *J. Am. Chem. Soc.* **2005**, *127* (50), 17842−17851.

(27) DeMore, W. B.; Bayes, K. D. Rate Constants for the Reactions of Hydroxyl Radical with Several Alkanes, Cycloalkanes, and Dimethyl Ether. *J. Phys. Chem. A* **1999**, *103* (15), 2649−2654.

(28) Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Gerkin, R. C.; Aspuru-Guzik, A.; Wiltschko, A. B. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv* **2019**, DOI: 10.48550/arXiv.1910.10685.

(29) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64−73.

(30) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27* (2), 82−85.

(31) Xiao, F.; Gulliver, J. S.; Simcik, M. F. Predicting aqueous solubility of environmentally relevant compounds from molecular features: A simple but highly effective four-dimensional model based on Project to Latent Structures. *Water Res.* **2013**, *47* (14), 5362−5370.

(32) Wang, Y.; Chen, J.; Yang, X.; Lyakurwa, F.; Li, X.; Qiao, X. In silico model for predicting soil organic carbon normalized sorption coefficient (KOC) of organic chemicals. *Chemosphere* **2015**, *119*, 438−444.

(33) Song, X.-C.; Dreolin, N.; Canellas, E.; Goshawk, J.; Nerin, C. Prediction of Collision Cross-Section Values for Extractables and Leachables from Plastic Products. *Environ. Sci. Technol.* **2022**, *56* (13), 9463−9473.

(34) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7* (1), 1−13.

(35) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* **2014**, *28* (7), 711−720.

(36) Wenlock, M.; Tomkinson, N. Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds. 2015.

(37) Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25* (2), 197−227.

(38) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* **2016**, 785−794.

(39) Dorogush, A. V.; Ershov, V.; Gulin, A. CatBoost: gradient boosting with categorical features support. *arXiv* **2018**, DOI: 10.48550/arXiv.1810.11363.

(40) Pelikan, M.; Goldberg, D. E.; Cantú-Paz, E. BOA: The Bayesian optimization algorithm. In *Proceedings of the genetic and evolutionary computation conference GECCO-99*; Citeseer, 1999; Vol. *1*, pp 525−532.

(41) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017**, 4765−4774.

(42) Ritchie, T. J.; Macdonald, S. J. The impact of aromatic ring count on compound developability−are too many aromatic rings a liability in drug design? *Drug discovery today* **2009**, *14* (21−22), 1011−1020.

(43) Muromachi, S.; Kamo, R.; Abe, T.; Hiaki, T.; Takeya, S. Thermodynamic stabilization of semiclathrate hydrates by hydrophilic group. *Rsc Adv.* **2017**, *7* (22), 13590−13594.