Environmental Science & Technology

pubs.acs.org/est

Article

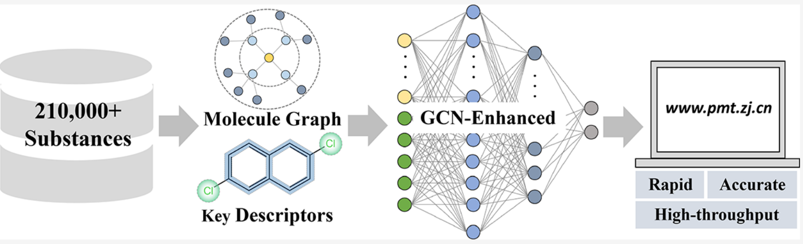
Graph Convolutional Network-Enhanced Model for Screening Persistent, Mobile, and Toxic and Very Persistent and Very Mobile Substances

Qiming Zhao, Yuting Zheng, Yu Qiu, Yang Yu, Meiling Huang, Yiqu Wu, Xiyu Chen, Yizhou Huang, Shixuan Cui, and Shulin Zhuang*

Cite This: *Environ. Sci. Technol.* 2024, 58, 6149−6157

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The global management for persistent, mobile, and toxic (PMT) and very persistent and very mobile (vPvM) substances has been further strengthened with the rapid increase of emerging contaminants. The development of a ready-to-use and publicly available tool for the high-throughput screening of PMT/vPvM substances is thus urgently needed. However, the current model building with the coupling of conventional algorithms, small-scale data set, and simplistic features hinders the development of a robust model for screening PMT/vPvM with wide application domains. Here, we construct a graph convolutional network (GCN)-enhanced model with feature fusion of a molecular graph and molecular descriptors to effectively utilize the significant correlation between critical descriptors and PMT/vPvM substances. The model is built with 213,084 substances following the latest PMT classification criteria. The application domains of the GCN-enhanced model assessed by kernel density estimation demonstrate the high suitability for high-throughput screening PMT/vPvM substances with both a high accuracy rate (86.6%) and a low false-negative rate (6.8%). An online server named PMT/vPvM profiler is further developed with a user-friendly web interface (http://www.pmt.zj.cn/). Our study facilitates a more efficient evaluation of PMT/vPvM substances with a globally accessible screening platform.

KEYWORDS: *PMT/vPvM, high-throughput screening, deep learning, feature fusion, online server*

INTRODUCTION

The persistent, mobile, and toxic (PMT) and very persistent and very mobile (vPvM) substances can persist for a long time and spread in the urban water circulation system, posing threats to sustainable, safe, and healthy drinking water.[1−3] With the increasing international attention on PMT/vPvM substances, they are regarded with an equal concern level as persistent, bioaccumulative, and toxic (PBT) and very persistent and very bioaccumulative (vPvB) substances by the German Environment Agency.[4−7] The definition of PMT/vPvM substances was updated in 2022 by the European Union classification, labeling, and packaging (CLP) to match the future management of chemicals.[8] The current identification methods for PMT/vPvM substances by traditional experimental analysis remains a huge challenge due to the incapability of the screening emerging contaminants in large scale.[9,10] An accurate and robust ready-to-use tool for high-throughput screening of PMT/vPvM substances is urgently needed.[11,12]

Machine learning and deep learning-driven models have emerged as efficient screening methods alternative to the experimental determination of PMT/vPvM and PBT/vPvB substances.[13−16] Considering the construction of certain models by conventional algorithms using simplistic features selection and generally small-scale data set, models with broad applicability domains, more interpretability, higher accuracies, and lower false negative rate (FNR) are essential for the promising practice of PMT/vPvM screening. Taking advantage of the multifusion approach in the characterization of chemical structures, models with better performance and more

Received: February 1, 2024
Revised: March 16, 2024
Accepted: March 19, 2024
Published: April 1, 2024

ACS Publications

https://doi.org/10.1021/acs.est.4c01201

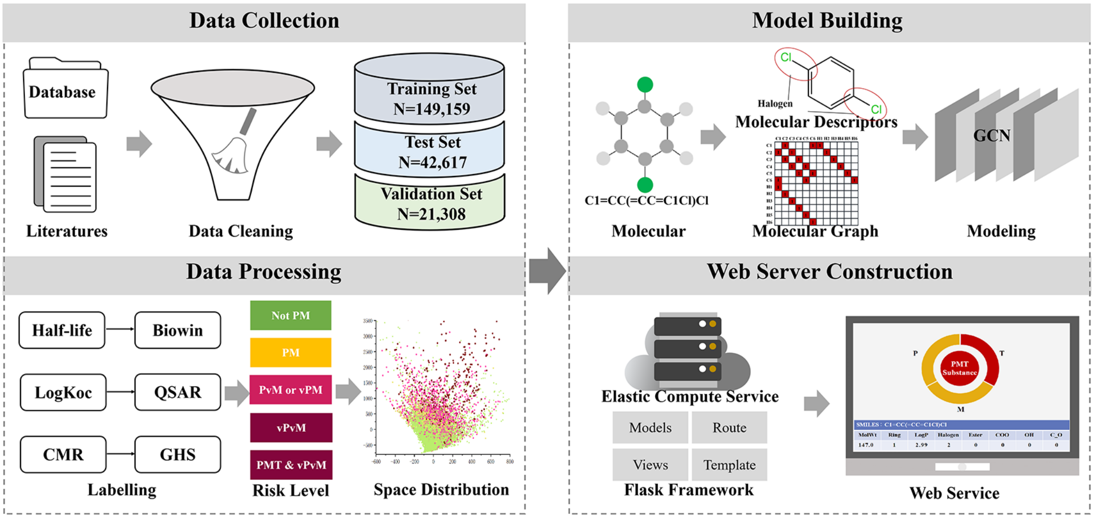


Figure 1. Scheme of data set construction, model building, and web service establishing. The information on substances was retrieved from various sources and integrated into the PMT data set. Four colors, dark red, pink, yellow, and green, were used to represent the risk levels of substances, which were projected on the chemical space following the latest criteria. The processed data set was employed to build the GCN-enhanced model with the feature fusion of critical descriptors and molecular graph. The PMT screening web server was constructed by Flask framework and Elastic Compute Service based on the GCN-enhanced model.

interpretability can be developed with multidimensional feature fusion to maximize the potential of different data dimensions for more accurate analysis.[17,18]

In this study, we constructed PMT/vPvM screening models by the graph convolutional network (GCN)-enhanced algorithm with feature fusion of molecular graphs and critical descriptors. The key substructures of substances retrieved from a large curated data set were analyzed by the occlusion sensitivity method with the visualization of atomic contribution to enhance model interpretability. The application domains were defined with the kernel density estimation (KDE) approach. The constructed model was integrated into a publicly accessible web server using flask framework and elastic compute service. Our study facilitates the high-throughput and high-precision screening of PMT/vPvM substances.[19−21]

■ MATERIALS AND METHODS

Curation of the Data Set Following the Latest Classification Criteria. To construct a large-scale data set with a broad chemical space, we collected massive substances from various databases and data sets covering PubChem, persistent organic pollutants list, PBT data from the European Chemicals Agency, comparative toxicogenomics database, carcinogenic, mutagenic, and toxic to reproduction data and substances of very high concern data (Figure 1).[22−25] The collected information contained canonical SMILES, half-time, biodegradability, Koc value, Kow value, bioconcentration factor value, and globally harmonized system of classification and labeling of chemicals classification for each substance (Table S1 and Text S1). The duplicates, mixtures, inorganics, salts, and noncovalent substances were removed after data cleaning. The remaining substances were integrated into the PMT data set and labeled as positive or negative for P, vP, M, vM, and T following the newest classification criteria of CLP (EU 2023/707) (Table S2). These criteria increased the mobility threshold compared with the old standard for better management of chemicals in the future. The risk level of each substance was determined according to the labels. The hold-out method was employed to randomly split PMT data set into training, test, and validation sets with a ratio of 7:2:1 for the model building, evaluation, and cross-validation. The chemical spatial distribution of the data set was dimensionally reduced by the t-distributed stochastic neighbor embedding (t-SNE) method and visualized in two-dimensional coordinate space.[26] The abbreviations of this article are summarized in Table S3.

Feature Engineering for Molecular Structure Representation. The feature engineering of transforming raw data into structural representations is essential for model construction. To extract the representative molecular features, six node features (type of atoms, hybridization, formal charge, degree of atom, number of radical electrons, and implicit valence of the atom), four bond features (type of bonds, whether the bond is conjugate, in ring, and aromatic), and 208 molecular descriptors for each substance were calculated.[27] The atom features u and bond features v, representing the local feature of substances, were encoded as molecular graphs $G = (u,v)$ using Deep Graph Library package (https://www.dgl.ai).[28] The molecular descriptors were calculated by the RDKit package (https://www.rdkit.org).[29]

Utilization of Molecular Descriptors in GCN-Enhanced Construction. To fully leverage the strong correlation between molecular descriptors and characteristics of PMT/vPvM substances, the top five important molecular descriptors for PMT substances selected by the SHAP method, and the node features and bond features were integrated to enhance the performance of GCN-enhanced model (Table S4).[30] The model propagated the integrated feature matrix from the input layer to hidden layers and output layer with information aggregation and results presentation (Figure 2).[18] In the convolution process, the integrated feature matrixes were iteratively updated by the approximated spectral graph convolution according to the following equations

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)})$$

$$\tilde{A} = A + I$$

where $H^{(l)}$ represents the characteristic matrixes of substances in the l-th layer, σ represents the activation function ReLU. $\tilde{D}$,

https://doi.org/10.1021/acs.est.4c01201

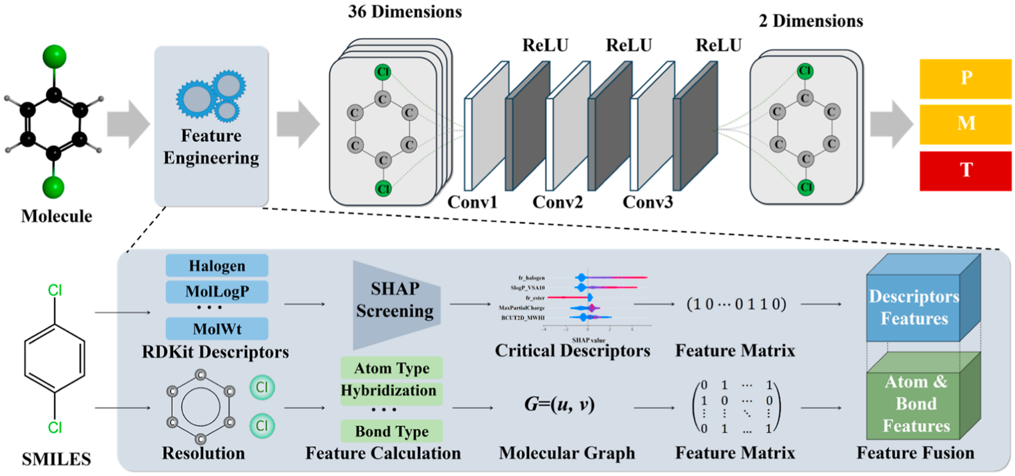


Figure 2. Structure of the GCN-enhanced model. The feature engineering of the molecule was divided into two processes. The top five critical descriptors were screened using the SHAP method in the first process. The atom and bond features were calculated and further converted to molecular graph in the second process. The feature matrixes calculated by two processes were integrated into a 36-dimension feature matrix and inputted into the convolution process. The prediction results were presented based on the output of the convolutional layer.

A, and I are degree matrix, adjacency matrix, and identity matrix, respectively. The feature matrixes of node in l layer $H^{(l)}$ were updated to next layer $H^{(l+1)}$ through aggregating features of adjacent nodes during the graph convolution process. Based on the output of the convolutional layer, the category of each substance was classified by the fully connected layer.

Model Training and Optimization with Multiple Algorithms. For the optimization of the GCN-enhanced model, the learning rate, batch size, and maximum epoch were set to 0.001, 200, and 300, respectively. The optimizer Adam was adopted to optimize the performance based on the cross entropy loss and the stochastic gradient descent algorithm.[31] During the training process, the performance of the GCN-enhanced model on the test set was recorded in each epoch and was projected on the learning curve to determine the optimal epoch numbers and avoid overfitting.[32] To further verify the performance of the GCN-enhanced model in the prediction of PMT/vPvM substances, five conventional algorithms [support vector machine (SVM), k-nearest neighbors (KNN), light gradient boosting machine (LightGBM), logistic regression (LR), and GaussianNB], and the GCN without integration of critical descriptors were trained as benchmarks for model comparison. The hyper-parameters of each model were determined using the grid search method to ensure optimization of the model performance.

Model Comparison by Evaluation Metrics. The performance of the constructed models was comprehensively compared by evaluation metrics. Three metrics including accuracy, sensitivity, and F1-score were employed to evaluate the performance of models. The parameters were defined as follows

$$\text{Accuracy} = \frac{N_{\text{TP}} + N_{\text{TN}}}{N_{\text{TP}} + N_{\text{TN}} + N_{\text{FP}} + N_{\text{FN}}}$$

$$\text{Sensitivity} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}$$

$$F1\text{-score} = \frac{2N_{\text{TP}}}{2N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}}}$$

where N_{TP}, N_{TN}, $N_{\text{FP},}$ and N_{FN} represent the number of true positive, true negative, false positive, and false negative, respectively. Accuracy and F1-score represent the overall correctness of the model, and sensitivity measures the ability of the model to correctly identify the positive substances. Considering the importance of FNR for contaminants identification, a sensitivity equal to 1-FNR was regarded as the primary evaluation metric in model evaluation.

Interpretability of Models by the Occlusion Sensitivity Method. To provide insights into the operation mechanism of trained models, the occlusion sensitivity method was employed to identify the key substructures of substances.[33,34] Each single atom was removed from the molecule to analyze the effect of atomic absence on the prediction results. The property difference between the molecular fragment and the complete molecule was further compared to quantify the atomic contribution. The atomic contribution to the prediction results was visualized with red and blue colors, revealing the key role of critical substructures.

Characterization of Application Domains by Gaussian Function Based-KDE. Density-based application domains were used to determine the reliable prediction regions in chemical space.[35] To define the application domains of model, the KDE method was employed, which involves the following formulas

$$f_h(x) = \frac{1}{nh} \times \sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)$$

$$k(x, x') = \frac{1}{\sigma\sqrt{2\pi}} e^{-x-x\prime^{2}/2\sigma^{2}}$$

$$h = \left(\frac{4\sigma^5}{3n}\right)^{1/5}$$

where σ represents the standard deviation of x, n represents the number of x, $f_h(x)$ represents the density of x in chemical space. The $f_h(x)$ of the training set points in chemical space can be projected onto a kernel density distribution plot. The thresholds of 5% and 10% were used to balance the model coverage (capturing an appropriate number of data points) and

https://doi.org/10.1021/acs.est.4c01201

prediction reliability (limiting the inclusion of outliers or noisy regions).

Implementation of an Online Server. To provide a global application of the constructed model, we established an online server named PMT/vPvM profiler. The server was developed using the Flask package (https://flask.palletsprojects.com) following reported protocols.[36] The persistence, mobility, and toxicity prediction models were integrated into the server with the insertion of each model into the Flask framework. The user-friendly interface was designed using the Hypertext Markup Language 5, Cascading Style Sheets (CSS), and JavaScript. CSS was used for styling and formatting webpage appearance, and JavaScript was used to implement interactive features. To ensure public accessibility and availability, the PMT/vPvM profiler was hosted on an elastic cloud service (4GiB, dual core) and deployed on port 80.[37,38]

Hardware and Software. The Scikit-learn package (v1.2.0) (https://scikit-learn.org/stable/) was adapted to build conventional models. The GCN and GCN-enhanced models employed in the PMT/vPvM profiler were established using the PyTorch package (v1.9.1) (https://pytorch.org/). The Scipy package (v1.7.3) (https://docs.scipy.org/) was used to analyze the application domains of models. To support its functionality, data processing and model construction were performed on a workstation computer (UltraLAB GX630 M with 204 RTX3090 GPUs).

RESULTS AND DISCUSSION

Large-Scale Data Set with Broad Chemical Space and Rich Diversity. We have curated a large-scale PMT/vPvM data set consisting of 213,084 substances following the latest classification criteria for PMT/vPvM substances from CLP. The distribution of positive and negative substances was calculated at a ratio ranging from 1:5 to 3:1 (Table S5). Based on the assigned labels classified by the latest criteria, the risk of substances was divided into four levels: Not PM, PM, vPM or PvM, PMT, and vPvM (Table S6). The adoption of the newest criteria aligns with the international trend of redefining the definition of PMT/vPvM substances, ensuring the availability of the curated PMT/vPvM data set. To the best of our knowledge, this is the largest data set used for PMT model construction until now, which encompassed most of expert-evaluated lists about PMT/vPvM substances.[13,14]

For further exploration of the data set diversity, all substances were dimensionally reduced by t-SNE and the chemical space was further projected on a two-dimensional graph with four colors representing four risk levels (Figure S1). Each point symbolized a substance, and the color of the data points presented a clear regional distribution in the chemical space, which proved that the PMT risk of the substance was highly correlated with molecular structure. The whole data points covered a large area in chemical space with a large span on the axis, demonstrating the rich structure diversity and the strong representation of PMT/vPvM data set.[39]

High Sensitivity of GCN-Enhanced Algorithms with Multifeature Integration. Five conventional algorithms, original GCN without feature fusion and GCN-enhanced algorithm, were trained using the curated PMT/vPvM data set. To ensure the optimal configuration of conventional algorithms, the algorithmic hyperparameters of each algorithm were determined by the grid search method (Table S7). The LightGBM performed better in the vP model with a sensitivity of 0.967, and SVM has a better performance in the P, M, vM, and T models with sensitivity reaching 0.896, 0.908, 0.922, and 0.957, respectively (Table S8). The learning curves of GCN and GCN-enhanced models depicted by Accuracy, sensitivity, F1-score, and loss values were employed to determine the optimal hyperparameters of models. For the GCN-enhanced P model, the loss values decreased significantly from 0.506 to near 0.23, showing the improvement of model performance with the increase of iteration (Figure 3A). Accuracy and F1-score both increased gradually, while sensitivity initially decreased and then steadily rose. The GCN-enhanced model under the epoch of 231 was chosen as the optimistic model with sensitivity, accuracy, and F1-score reaching 0.935, 0.922, and 0.889, respectively. The remaining models adopted a similar principle to determine the optimistic parameters (Figure S2).

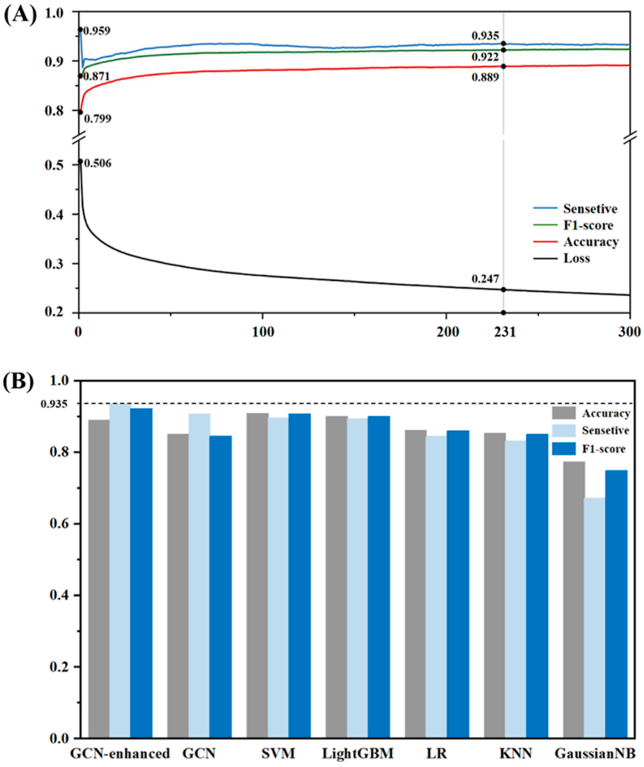


Figure 3. (A) Learning curve of the GCN-enhanced model with the variation trend of sensitivity, accuracy, F1-score, and loss function value. (B) Comparison of performance between GCN-enhanced, GCN, SVM, LightGBM, LR, KNN, and GaussianNB algorithms indicated by the metrics of accuracy, sensitivity, and F1-score.

Compared with five conventional algorithms and the GCN model for persistence prediction, the GCN-enhanced model performed better with the highest sensitivity at 0.935, which was significantly higher than the closest competitor GCN at 0.907 (Figure 3B). Similarly, the GCN-enhanced model demonstrated its superior performance in predicting vP, M, vM, and T with sensitivity reaching 0.978, 0.951, 0.962, and 0.932, effectively minimizing the risk of overlooking hazardous PMT/vPvM substances. The enhanced performance of the GCN-enhanced model suggested that the integration of critical descriptors and molecular graph strengthened the predictive ability for PMT/vPvM substances. Considering the limitation of GCN in preserving local structure information during the

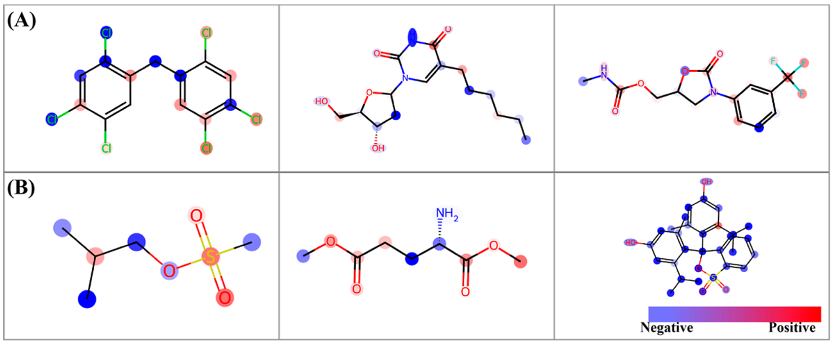


Figure 4. Atomic contribution plot of substances. (A) Influence of the GCN-enhanced model in the prediction of persistent. (B) Influence of the GCN-enhanced model in the prediction of very persistent. The atoms depicted in the red circle and the blue circle represent the positive influence and the negative influence, respectively.

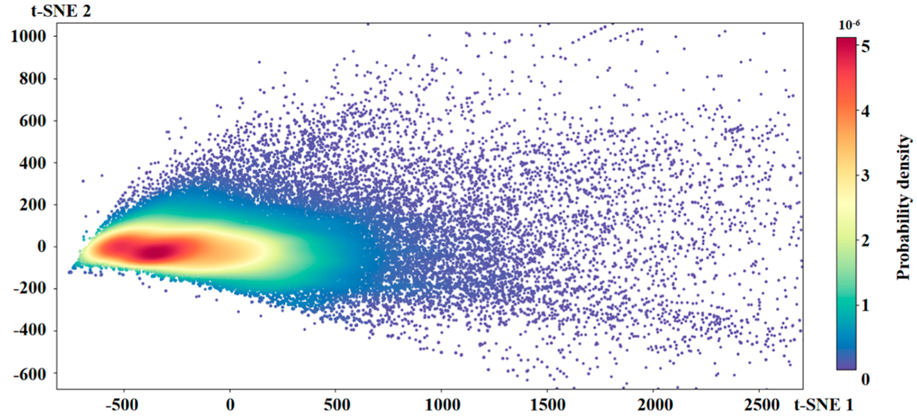


Figure 5. Application domains density plot for the training set with 149,159 substances. The chemical space of the training set was calculated based on RDKit descriptors and dimensionally reduced by t-SNE. The application domains density was defined by the KDE method and projected on the 2D chemical space. The color scale represented the probability density.

propagation process, incorporating key descriptors highly related to the PMT substance can make up the missing part information and significantly improve the representation of local structural details. With the fusion of key descriptors, the GCN-enhanced model provided a more efficient analysis for the characteristics of PMT/vPvM substances, achieving a better predictive performance.[40]

The GCN-enhanced model was further compared to previously established PMT models. The sensitivity and accuracy of GCN-enhanced model are significantly higher than previous models constructed by conventional algorithms and simplistic features.[41] Ascribed to the combination of the large-scale data set and multifeature fusion of deep learning, our GCN-enhanced model has advantages over previous model with the higher suitability for more accurate predictions, in line with the future trend of global assessment and management on PMT/vPvM substances.

Key Substructures Identified for Model Interpretability. Machine learning models are commonly perceived as a black box for its unclear prediction mechanism and poor interpretability. To explore the key substructure and attain the interpretability of models, the atomic contributions for the prediction of PMT/vPvM substances were visualized on the molecular structure diagram. Six randomly selected substances from the test set were specifically examined to understand how key substructures influenced P (Figure 4A) and vP prediction (Figure 4B). The colors on the diagrams represent the detailed mechanisms between substructures and properties of substances. The red areas surrounding the chlorine, bromine, and iodine atoms indicated that halogens contributed to the persistence of substances. The result was in consistence with the property of POPs and PFAS, which have a long degradation half-life for halogen atoms, proving the reasonability of the P model.[42,43] The key substructures related to mobility and toxicity were also visualized with the atomic contributions of substances (Figure S3). The carboxyl, hydroxyl, amidogen, and carbonyl group were marked with red in M and vM prediction, demonstrating the positive impact on the mobility, consistent with their inherent hydrophilic properties.

The halogen was covered with red in the T prediction, in line with the high toxicity of organic halides. The consistency of key substructures with reported research validated the rationale of GCN-enhanced models and consolidated the reliability of the prediction.[44]

Broad and Accurate Application Domains for PMT/vPvM Screening. The training set consisting of 149,159 substances was visualized in the chemical space by projecting the data points onto a kernel density distribution plot with color representing the density of the data (Figure 5). A majority of substances were concentrated in the area surrounded by yellow, indicating that the distribution of training sets in chemical space was a central regional concentration. As indicated by application domains thresholds at 10% and 5%, 2331 substances under the 5% threshold and 4483 substances under the 10% threshold were outside of the

Table 1. Sensitivity of Five Models Under Different Thresholds and the Number of Substances Outside the Application Domains[a]

threshold	substances outside application domains	proportion	sensitivity				
			P	vP	M	vM	T
5%	2331	1.6%	94.2	97.7	94.9	97.3	93.6
10%	4483	2.99%	94.5	97.9	95.3	97.8	93.9

[a] Proportion: The ratio of the number of substances outside the domain to the total number of substances in the external verification set. P, vP, M, vM, and T: the GCN-enhanced models for the prediction of persistent, very persistent, mobile, very mobile, and toxic, respectively.

application domains (Table 1). To balance the application domains and prediction accuracy, we chose 10% as the threshold for obtaining the optimal model. The ratio of the substances in the application domains to substances outside the application domains was within the appropriate range, and the number of substances in the application domains accounted for a very large proportion higher than 97%, showing extremely broad application domains of our models.

The PMT models developed previously were accurate, especially in a narrow application range due to its small-scale data set and relatively limited application domains. The application domains of our model were characterized by the largest PMT data set and calculated with the strictest defining approach Gaussian KDE, a method with more accurate and rigorous definition compared with traditional distances method.[45] The value of $f_h(x)$ at any point in the chemical space accurately represented the density of the data set, effectively identifying the application range of model.[46,47] The employment of Gaussian KDE in application domains enabled our models to achieve a broad and precise application scope.

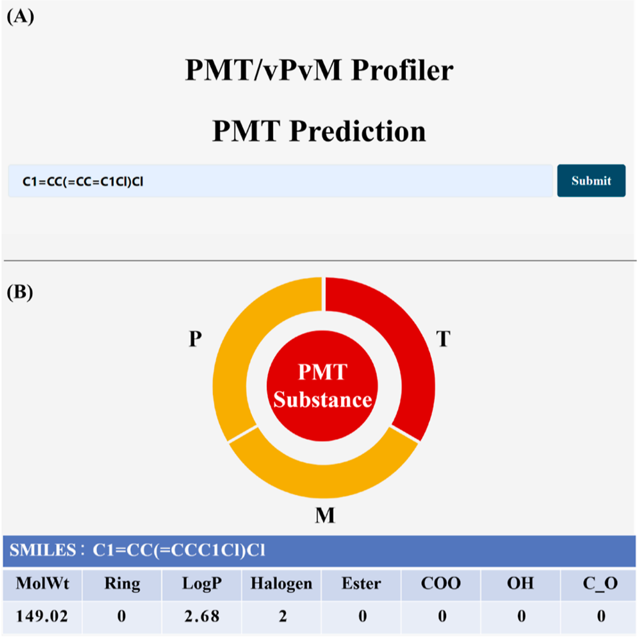
Excellent Generalization Capability of Established Models. The validation set containing 21,308 substances without overlap between the training set was employed to evaluate the generalization capability of GCN-enhanced models (Table 2). The distribution of the validation and training set was projected on chemical space, where part of validation substances were uncovered by the training set with large structural differences, providing a robust test for model generalization capabilities (Figure S4). The GCN-enhanced model performed excellent with a sensitivity over 90.5% and an accuracy over 88.8%, particularly for the prediction of vP where the sensitivity reached as high as 97.6%. Since the training set contained a sufficiently large number of substances with the encompassment of most compound characteristics, the GCN-enhanced model exhibited a similar excellent performance on both test and validation set. The robust performance of the GCN-enhanced model demonstrated the good generalization capability, facilitating the accurate screening of large quantities of emerging contaminants with diverse structures.[48]

Table 2. Accuracy, Sensitivity, F1-Score, and FNR of GCN-Enhanced Models Calculated With the Validation Set[a]

	P	vP	M	vM	T[a]
Accuracy (95% CI), %	88.8	96.6	90.8	93.8	86.5
Sensitivity (95% CI), %	93.9	97.6	94.6	90.7	90.5
F1-score (95% CI), %	92.2	96.3	93.1	91.2	90.1
FNR (95% CI), %	6.1	2.4	5.4	9.3	9.5

[a] P, vP, M, vM, and T: the GCN-enhanced models for the prediction of persistent, very persistent, mobile, very mobile, and toxic, respectively.

Intuitive Interface of the PMT/vPvM Profiler. The PMT/vPvM Profiler was designed with an intuitive interface and user-friendly online service (Figure 6), which is globally accessible at http://www.pmt.zj.cn/. The architecture of the PMT/vPvM Profiler was composed by a data input interface using SMILES representation of substances as input, a backend based on the Flask framework, and a prediction result page for the visualization of persistence, mobility, toxicity classification, and corresponding risk levels of substances. To provide insight into molecular structure characteristics, the critical structural information was calculated and listed in the prediction results page. To meet with the need for high-throughput screening, the file upload and results download function were developed for task submission and output download (Figure S5). After uploading the input file, the PMT/vPvM profiler extracted the SMILES of substances in batches, consequently computed the PMT properties with related confidence coefficient, and finally provided prediction results. The PMT/vPvM profiler provides a simple and efficient operation process, enabling the rapid, accurate, and high-throughput screening of PMT/vPvM substances.



C1=CC(=CC=C1Cl)Cl

SMILES：C1=CC(=CCC1Cl)Cl

MolWt	Ring	LogP	Halogen	Ester	COO	OH	C_O
149.02	0	2.68	2	0	0	0	0

Figure 6. Input and output interface for the PMT/vPvM profiler. (A) SMILES of substances was designed as the input information. (B) Output interface exhibited the molecular skeletal formula diagram, prediction results summary, and substances' critical structure information.

ENVIRONMENTAL IMPLICATIONS

The increasing international attention for PMT/vPvM substances urgently calls for efficient tools for the high-throughput screening of PMT/vPvM substances. We implemented a PMT/vPvM profiler, the first online server for screening PMT/vPvM substances, using the GCN-enhanced model trained with the currently largest PMT/vPvM data set. The GCN-enhanced model operated with wide application domains accurately defined by the Gaussian KDE method and high interpretability evaluated by atomic contributions. The integration of molecular graphs and key molecular descriptors contributed to the state-of-the-art performance of the GCN-enhanced model. Considering that PMT substances can exist in the aqueous environment in the ionic form, the construction of prediction models suitable for ionic compounds is urgently needed. Overall, the PMT/vPvM profiler provides an efficient and robust tool for rapid screening PMT/vPvM substances, beneficial for the risk assessment of emerging contaminants.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.4c01201.

> Collection and process of toxicity information, information on data set and collected data, the specific criteria of PMT/vPvM substances in the newest classification criteria of CLP (EU 2023/707), acronym glossary, top five critical RDKit descriptors for the prediction of persistent, very persistent, mobile, very mobile and toxic, the number of positive and negative substances in the PMT/vPvM data set, the number of substances in four risk levels, optimal hyperparameters of five conventional algorithms, the model performance for the prediction of PMT/vPvM substances, the chemical space distributions of PMT/vPvM data set, the learning curve of GCN-enhanced model for the prediction of PMT/vPvM substances, atomic contributions of substances influencing the prediction of GCN-enhanced model, the chemical space distributions of training set and validation set, and the template for input file and result file PDF

AUTHOR INFORMATION

Corresponding Author

Shulin Zhuang − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China;* orcid.org/0000-0002-7774-7239; Email: shulin@zju.edu.cn

Authors

Qiming Zhao − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Yuting Zheng − *Solid Waste and Chemicals Management Center, Ministry of Ecology and Environment of the People's Republic of China, Beijing 100029, China*

Yu Qiu − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Yang Yu − *Solid Waste and Chemicals Management Center, Ministry of Ecology and Environment of the People's Republic of China, Beijing 100029, China*

Meiling Huang − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Yiqu Wu − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Xiyu Chen − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Yizhou Huang − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Shixuan Cui − *College of Environmental and Resource Sciences, and Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310058, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.4c01201

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors appreciated the financial support from the National Key Research & Developmental Program of China (no. 2023YFD1701302), the National Natural Science Foundation of China (no. 22136001), and the Zhejiang Provincial Natural Science Foundation of China (no. LZ22B070001).

REFERENCES

(1) Hale, S. E.; Arp, H. P. H.; Schliebner, I.; Neumann, M. What's in a name: persistent, mobile, and toxic (PMT) and very persistent and very mobile (vPvM) substances. *Environ. Sci. Technol.* **2020**, *54* (23), 14790−14792.

(2) Lau, S. S.; Bokenkamp, K.; Tecza, A.; Wagner, E. D.; Plewa, M. J.; Mitch, W. A. Toxicological assessment of potable reuse and conventional drinking waters. *Nat. Sustainability* **2022**, *6* (1), 39−46.

(3) Hale, S. E.; Neumann, M.; Schliebner, I.; Schulze, J.; Averbeck, F. S.; Castell-Exner, C.; Collard, M.; Drmac, D.; Hartmann, J.; Hofman-Caris, R.; Hollender, J.; de Jonge, M.; Kullick, T.; Lennquist, A.; Letzel, T.; Nödler, K.; Pawlowski, S.; Reineke, N.; Rorije, E.; Scheurer, M.; Sigmund, G.; Timmer, H.; Trier, X.; Verbruggen, E.; Arp, H. P. H. Getting in control of persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances to protect water resources: strategies from diverse perspectives. *Environ. Sci. Eur.* **2022**, *34* (1), 22.

(4) Jin, B.; Huang, C.; Yu, Y.; Zhang, G.; Arp, H. P. H. The need to adopt an international pmt strategy to protect drinking water resources. *Environ. Sci. Technol.* **2020**, *54* (19), 11651−11653.

(5) Neumann, M. S. M.; Sättler, D.; Oltmanns, J.; Vierke, L.; Kalberlah, F., In a proposal for a chemical assessment concept for the protection of raw water resources under REACH. In *Extended Abstract for the Oral Presentation at the 25th SETAC Annual Meeting*, 2015, https://www.umweltbundesamt.de/sites/default/files/medien/362/dokumente/20150505_neumann_setac_europe_2015_extended_abstract.pdf. (accessed on May 10, 2023).

(6) Neumann, M.; Schliebner, I., Protecting the sources of our drinking water: The criteria for identifying persistent, mobile, and toxic (PMT) substances and very persistent, and very mobile (vPvM) substances under the EU chemical legislation REACH, UBA Texte 127/2019; German Environmental Agency (UBA): Dessau-Rosslau, Germany. 2019, ISSN: 1862−4804.

(7) Hale, S. E.; Arp, H. P. H.; Schliebner, I.; Neumann, M. Persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances pose an equivalent level of concern to persistent, bioaccumulative and toxic (PBT) and very persistent and very bioaccumulative (vPvB) substances under REACH. *Environ. Sci. Eur.* **2020**, *32* (1), 155.

(8) European Commission. Amending regulation (ec) no 1272/2008 as regards hazard classes and criteria for the classification, labelling and packaging of substances and mixtures, 2022. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32008R1272. (accessed on April 17, 2023).

(9) Zhang, S. X.; Chen, J. Z.; Wang, Z. Y.; Chen, C. K.; Chen, A. N.; Jing, Q. A.; Liu, J. G. Dynamic source distribution and emission inventory of a persistent, mobile, and toxic (PMT) substance, Melamine, in China. *Environ. Sci. Technol.* **2023**, *57* (39), 14694−14706.

(10) European Commission. Chemicals strategy for sustainability towards a toxic-free environment, 2020. https://ec.europa.eu/environment/pdf/chemicals/2020/10/Strategy.pdf. (accessed on May 15, 2023).

(11) Huang, C.; Jin, B.; Han, M.; Yu, Y.; Zhang, G.; Arp, H. P. H. The distribution of persistent, mobile and toxic (PMT) pharmaceuticals and personal care products monitored across Chinese water resources. *J. Hazard. Mater. Lett.* **2021**, *2*, 100026.

(12) Xu, J.; Liu, J. Managing the risks of new pollutants in China: the perspective of policy integration. *Environ. Health* **2023**, *1* (6), 360−366.

(13) Zhao, Q. M.; Yu, Y.; Gao, Y. C.; Shen, L. L.; Cui, S. X.; Gou, Y. Y.; Zhang, C. L.; Zhuang, S. L.; Jiang, G. B. Machine learning-based models with high accuracy and broad applicability domains for screening PMT/vPvM substances. *Environ. Sci. Technol.* **2022**, *56* (24), 17880−17889.

(14) Han, M.; Jin, B.; Liang, J.; Huang, C.; Arp, H. P. H. Developing machine learning approaches to identify candidate persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances based on molecular structure. *Water Res.* **2023**, *244*, 120470.

(15) Wang, H. B.; Wang, Z. Y.; Chen, J. W.; Liu, W. J. Graph attention network model with defined applicability domains for screening PBT chemicals. *Environ. Sci. Technol.* **2022**, *56* (10), 6774−6785.

(16) Wu, G.; Zhu, F.; Zhang, X.; Ren, H.; Wang, Y.; Geng, J.; Liu, H. PBT assessment of chemicals detected in effluent of wastewater treatment plants by suspected screening analysis. *Environ. Res.* **2023**, *237*, 116892.

(17) Song, W. W.; Li, S. T.; Fang, L. Y.; Lu, T. Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote* **2018**, *56* (6), 3173−3184.

(18) Liu, Q. C.; Xiao, L.; Yang, J. X.; Wei, Z. H. CNN-Enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification. *IEEE Trans. Geosci. Remote* **2021**, *59* (10), 8657−8671.

(19) Tsai, W. P.; Feng, D. P.; Pan, M.; Beck, H.; Lawson, K.; Yang, Y.; Liu, J. T.; Shen, C. P. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nat. Commun.* **2021**, *12* (1), 5988.

(20) Bzdok, D.; Nichols, T. E.; Smith, S. M. Towards algorithmic analytics for large-scale datasets. *Nat. Mach. Intell.* **2019**, *1* (7), 296−306.

(21) Yan, X. L.; Sedykh, A.; Wang, W. Y.; Yan, B.; Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* **2020**, *11* (1), 2519.

(22) European Chemicals Agency. Candidate list of substances of very high concern for authorisation, 2023. https://www.echa.europa.eu/candidate-list-table. (accessed on May 12, 2023).

(23) European Chemicals Agency. PBT assessment list, 2023. https://echa.europa.eu/fr/pbt. (accessed on April 27, 2023).

(24) European Commission. The european parliament and of the council concerning the registration, evaluation, authorisation and restriction of chemicals (reach), as regards carcinogenic, mutagenic or reproductive toxicant (cmr) substances, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R2204. (accessed on May 2, 2022).

(25) European Commission. Protecting health and the environment from persistent organic pollutants, 2019. https://eur-lex.europa.eu/legal-content/en/LSU/?uri=CELEX:32019R1021. (accessed on May 1, 2023).

(26) Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10* (1), 5416.

(27) Shen, W. X.; Zeng, X.; Zhu, F.; Wang, Y. L.; Qin, C.; Tan, Y.; Jiang, Y. Y.; Chen, Y. Z. Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.* **2021**, *3* (4), 334−343.

(28) Lam, H. Y. I.; Pincket, R.; Han, H.; Ong, X. E.; Wang, Z. C.; Hinks, J.; Wei, Y. J.; Li, W. F.; Zheng, L. Z.; Mu, Y. G. Application of variational graph encoders as an effective generalist algorithm in computer-aided drug design. *Nat. Mach. Intell.* **2023**, *5* (7), 754−764.

(29) Scalfani, V. F.; Patel, V. D.; Fernandez, A. M. Visualizing chemical space networks with RDKit and NetworkX. *J. Cheminf.* **2022**, *14* (1), 87.

(30) Park, J.; Shim, Y.; Lee, F.; Rammohan, A.; Goyal, S.; Shim, M.; Jeong, C.; Kim, D. S. Prediction and interpretation of polymer properties using the graph convolutional network. *ACS Polym. Au* **2022**, *2* (4), 213−222.

(31) Khan, A. H.; Cao, X. W.; Li, S.; Katsikis, V. N.; Liao, L. F. BAS-ADAM: an ADAM based approach to improve the performance of beetle antennae search optimizer. *IEEE/CAA J. Autom.* **2020**, *7* (2), 461−471.

(32) Krishnan, R.; Rajpurkar, P.; Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **2022**, *6* (12), 1346−1352.

(33) Schmidt, C. W. Into the Black Box: What can machine learning offer environmental health research? *Environ. Health Perspect.* **2020**, *128* (2), 39002.

(34) Liu, X.; Guo, Y. H.; Pan, W. X.; Xue, Q.; Fu, J. J.; Qu, G. B.; Zhang, A. Q. Exogenous chemicals impact virus receptor gene transcription: insights from deep learning. *Environ. Sci. Technol.* **2023**, *57* (46), 18038−18047.

(35) Kausar, S.; Falcao, A. O. A visual approach for analysis and inference of molecular activity spaces. *J. Cheminf.* **2019**, *11* (1), 63.

(36) Vogel, P., Klooster, T., Andrikopoulos, V., Lungu, M. A low-effort analytics platform for visualizing evolving flask-based python web services. *2017 Ieee Working Working Conf Softw Vis (Vissoft 2017)* 2017109−113

(37) Xu, L.; Deng, C. Y.; Pang, B.; Zhang, X. X.; Liu, W.; Liao, G. M.; Yuan, H. T.; Cheng, P.; Li, F.; Long, Z. L.; Yan, M.; Zhao, T. T.; Xiao, Y.; Li, X. TIP: A web server for resolving tumor immunophenotype profiling. *Cancer Res.* **2018**, *78* (23), 6575−6580.

(38) Törönen, P.; Medlar, A.; Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **2018**, *46* (W1), W84−W88.

(39) Zhang, Q. C.; Yang, L. T.; Chen, Z. K.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146−157.

(40) Chaib, S.; Liu, H.; Gu, Y. F.; Yao, H. X. Deep feature fusion for vhr remote sensing scene classification. *IEEE Trans. Geosci. Remote* **2017**, *55* (8), 4775−4784.

(41) Ayantayo, A.; Kaur, A.; Kour, A.; Schmoor, X.; Shah, F.; Vickers, I.; Kearney, P.; Abdelsamea, M. M. Network intrusion detection using feature fusion with deep learning. *J. Big Data-Ger.* **2023**, *10* (1), 167.

(42) Khan, B.; Burgess, R. M.; Cantwell, M. G. Occurrence and bioaccumulation patterns of per- and polyfluoroalkyl substances (PFAS) in the marine environment. *ACS EST Water* **2023**, *3* (5), 1243−1259.

(43) Gramatica, P.; Papa, E. Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure. *Environ. Sci. Technol.* **2007**, *41* (8), 2833−2839.

(44) Cheng, W. X.; Ng, C. A. Using machine learning to classify bioactivity for 3486 per- and polyfluoroalkyl substances (PFASs) from the OECD list. *Environ. Sci. Technol.* **2019**, *53* (23), 13970−13980.

(45) Ji, P.; Zhao, N.; Hao, S. J.; Jiang, J. G. Automatic image annotation by semi-supervised manifold kernel density estimation. *Inf. Sci.* **2014**, *281*, 648−660.

(46) Chen, Y.; Yang, H. B.; Wu, Z. R.; Liu, G. X.; Tang, Y.; Li, W. H. Prediction of farnesoid x receptor disruptors with machine learning methods. *Chem. Res. Toxicol.* **2018**, *31* (11), 1128−1137.

(47) Linderman, G. C.; Rachh, M.; Hoskins, J. G.; Steinerberger, S.; Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **2019**, *16* (3), 243−245.

(48) Chen, K. J.; Chen, K. L.; Wang, Q.; He, Z. Y.; Hu, J.; He, J. L. Short-Term load forecasting with deep residual networks. *IEEE Trans. Smart Grid* **2019**, *10* (4), 3943−3952.