



Universiteit
Leiden
The Netherlands

Addressing the data scarcity problem in ecotoxicology via small data machine learning methods

Wang, Y.; Dong, J.; Zhou, Y.; Cheng, Y.; Zhao, X.; Peijnenburg, W.J.G.M.; ... ; Wu, F.

Citation

Wang, Y., Dong, J., Zhou, Y., Cheng, Y., Zhao, X., Peijnenburg, W. J. G. M., ... Wu, F. (2025). Addressing the data scarcity problem in ecotoxicology via small data machine learning methods. *Environmental Science And Technology*, 59(12), 5867-5871.
doi:10.1021/acs.est.5c00510

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/4212177>

Note: To cite this publication please use the final published version (if applicable).

Addressing the Data Scarcity Problem in Ecotoxicology via Small Data Machine Learning Methods

Ying Wang, Jinchu Dong, Yunchi Zhou, Yinghao Cheng, Xiaoli Zhao, Willie J. G. M. Peijnenburg, Martina G. Vijver, Kenneth M. Y. Leung, Wenhong Fan,^{*} and Fengchang Wu^{*}



Cite This: <https://doi.org/10.1021/acs.est.5c00510>



Read Online

ACCESS |

Metrics & More

Article Recommendations

SCIENTIFIC
OPINION
NON-PEER
REVIEWED



KEYWORDS: *ecotoxicity, small data machine learning (SDML), prediction, data augmentation, modeling workflow, artificial intelligence*

Machine learning (ML) has been extensively employed in the field of ecotoxicological research, with a notable application in predicting ecotoxicity. Typically, researchers employ chemical substance property data, environmental exposure data, and species information to predict the potential toxicity of pollutants through specific ML models and algorithms.¹ However, it is difficult to gather systematic collected toxicological data as they are costly and time-consuming to generate, while endless combinations of pollutant species with exposure media properties can be made. This leads to unbalanced experimental data, which often results in data sets with varying quality that are unsuitable for modeling. In fact, many modeling studies in this field lack a solid database.² In the process of applying ML for toxicity prediction, limited data and sparse learnable features are common challenges. When the available data cannot meet the requirements for model construction, this situation is termed a “small data” problem. This issue can be understood from two

perspectives. First, the sample size fails to meet the minimum data requirements for ML modeling; for instance, previous studies suggested such a minimum threshold as 100 samples.^{3,4} Second, the sample size fails to satisfy the specific needs of the modeling task. For example, when the ratio of the number of samples to the number of features is too low, the model is prone to overfitting. In such cases, even thousands of samples may still be considered small data for models with hundreds of features. In our previous literature review, most studies that achieve favorable predictions maintain a sample data-to-features ratio above 20.^{5,6} Hence, we suggest that a ratio



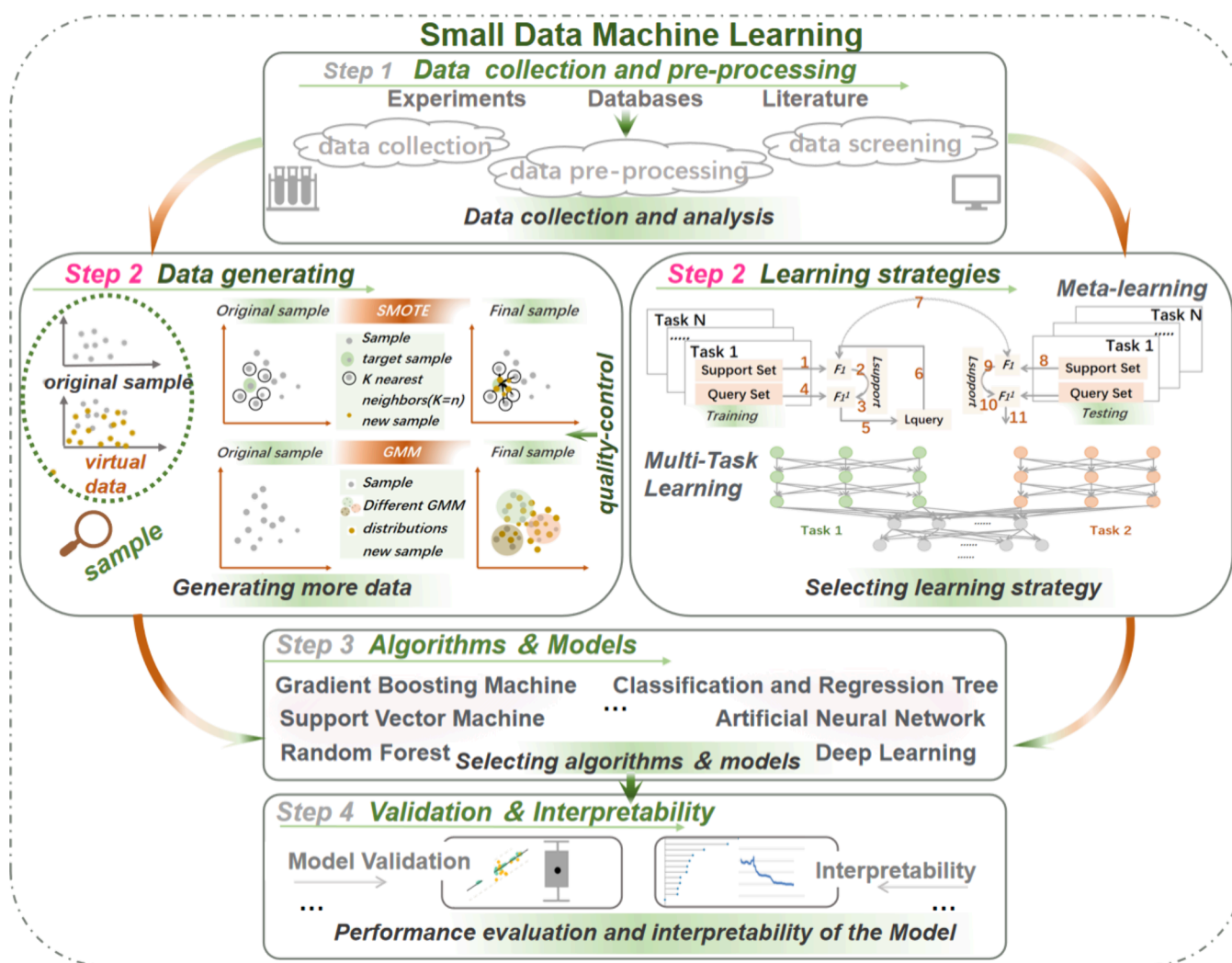


Figure 1. Workflow for the application of SDML methods in ecotoxicology.

below 20 should be considered a sign of a small data scenario. Such limited data sets not only curtail the predictive performance of ML models but also diminish the representativeness and persuasiveness of the models.

Acknowledging the difficulty of the acquisition of reliable toxicity data, there is an urgent need to develop a specific ML-based modeling process for small data samples, with the aim of improving data utilization and improving the predictive performance of models. A rapid and large-scale prediction of the ecotoxicity of chemicals and/or nanomaterials is crucial for reducing the time and cost of risk assessment, as well as minimizing animal experiments and selecting optimal testing strategies. This is also considered as a key direction in 21st century toxicology. Therefore, we believe that small data ML (SDML) methods are especially valuable when the available research data set is inadequate for applying conventional ML models. The outcomes of SDML models can also identify data gaps and help refine the plan for conducting supplementary ecotoxicity tests, thereby optimizing the allocation of limited resources.

In this Viewpoint, we discuss the implementation principles, modeling processes, and challenges of SDML methods and elucidate the application of SDML methods in ecotoxicology for advancing science and technology.

SMALL DATA ML IMPLEMENTATION

The one-shot learning framework was first proposed by Li et al.⁷ in 2003, and it is generally accepted that the quality and quantity of data, as well as the algorithms used for modeling, are the determining factors for the performance of SDML models. Therefore, the development of SDML methods in ecotoxicology applications can be achieved in two ways. One is to solve the problem of limited toxicity data through increasing the amount of data as much as possible by data augmentation, and the other is to solve the problem of possible model overfitting through choosing the learning strategy and ML algorithms suitable for the small toxicology data set (and, at the same time, avoiding overfitting to increase the accuracy of the prediction model and its generalization ability).

Data Augmentation Methods. The principal tools of data augmentation for generating new data include the synthetic minority oversampling technique (SMOTE), Gaussian mixture model (GMM), and other methods based on data distribution of the original data set. Newly generated virtual data need to be quality controlled. Quality control of virtual data is crucial in ensuring that augmented data do not introduce noise or bias. This quality control can be performed in two stages. (1) In pre-virtual data generation, rules should be established to restrict the scope of data augmentation to

ensure that the distribution of virtual data closely resembles that of real data (e.g., feature distribution and category balance) and at the same time adheres to physical laws and the logic of toxicology, integrating logical constraints into the model. For instance, the exposure time of pollutants to aquatic organisms cannot be negative. (2) For post-virtual data generation, the credibility of the virtual data should be assessed by pretrained models to detect anomalies in the generated data. For example, the generated data are considered valid if the difference between their predicted and actual outputs is less than 10%.⁸ The newly generated data will then be included in modeling (Figure 1).

For a case study, in our previous work we established a database of the toxicity of metallic and metal oxide nanoparticles to multiple species.⁹ We selected 131 toxicity data points of metallic nanoparticles for *Daphnia magna* from the database and expanded this data set by 20-fold using the SMOTE method. Based on this data set, we employed five ML models (support vector machine (SVM), random forest (RF), classification and regression tree (CART), gradient boosting machine (GBM), and artificial neural network (ANN)) to predict the toxicity of nanometals to *D. magna*. The result showed that the ANN model achieved the best performance ($R^2 = 0.99$; RMSE = 0.50), outperforming the modeling results for the original data set ($R^2 = 0.77$; RMSE = 0.51). Additionally, we attempted the same approach using the GMM method and found that the GBM model yielded the best fit after augmentation by 20 times ($R^2 = 0.98$; RMSE = 0.60), also surpassing the modeling results of the original data set.

Selection of Learning Strategies. Learning strategies suitable for small data scenarios are specifically meta-learning and multi-task learning. Meta-learning is used to design a cross-task optimization strategy that can quickly converge to the best model parameters. Meta-learning can extract toxicological data features from the data and select better parameters or models by learning how to learn from multiple related tasks. Effectively, only a small amount of data are used to quickly train a reliable model and ensure the accuracy. Multi-task learning can learn from multiple related tasks while exploiting the shared information between tasks to improve prediction performance. An example of how to execute these learning strategies can be found in ref 10. Schlender et al. executed different meta-learning techniques and showed that a multi-task RF model performed optimally with scarce data and was able to accurately predict toxicity for multiple aquatic species.

MODELING PROCESS FOR SMALL DATA ML

Based on the above discussion and characteristics of toxicological data, we propose a modeling workflow for the application of SDML methods in ecotoxicology at the data and model levels (Figure 1):

Step 1. Data collection and analysis

Acquire quality ecotoxicological data, physicochemical properties of chemical and/nanomaterial substances, and species information through toxicity experiments, online databases, and pertinent literature and preprocess the data by the Findability, Accessibility, Interoperability, and Reuse (FAIR) Principles or the Klimisch system to scoring methodologies for data quality.¹¹

Step 2. Choosing the right data augmentation method or learning strategies

Choose the most suited ML methods for small data sets based on the research objectives (classification or regression). In the case of performing regression tasks, utilizing SMOTE and GMM to generate virtual data is recommended, followed by filtering based on established data selection principles. For classification problems, strategies emphasizing enhanced learning capabilities across multiple tasks, such as meta-learning and multi-task learning, are advised methods to optimize parameters.

Step 3. ML algorithms selection and modeling

Select ML modeling algorithms that are more suitable for small data samples and can quickly capture data characteristics, such as SVM, RF, GBM, decision trees, and extreme gradient boosting (XGBoost).

Step 4. Model performance evaluation and interpretability

Model performance and credibility are recommended to be tested using internal and external validation methods, and the use of the mean square error (MSE) increase, SHapley Additive exPlanation (SHAP), partial dependence plot (PDP), and local interpretable model-agnostic explanation (LIME) ensures that the model is satisfactorily interpretable.

This workflow has the potential for the prediction of ecotoxicity of the infinite combinations of chemicals/nanomaterials and species as *in vivo* testing is tedious and virtually undoable.

CHALLENGES FOR SMALL DATA ML

Facing the dilemma of data scarcity in ecotoxicology, the use of SDML methods can solve problems of severe performance degradation of traditional ML methods when scattered scarce data are available.

As any method is experimental or computational, SDML models also have precautions in their use. The following aspects should be acknowledged:

- (1) SDML models are more sensitive to anomalous data (which is more common in toxicological trials), and one must thus be more judicious in data processing when toxicological data are collected in databases or from the literature.
- (2) When the task is highly variable or the distribution of the data is changed, SDML models often have difficulty in learning and predicting effectively across different domains. SDML methods are therefore more suitable for prediction tasks with a single goal and low complexity.
- (3) SDML methods pose significant challenges in terms of scalability for large-scale applications due to their high model complexity and computational costs. Just as in case of the Ecological Structure–Activity Relationship (ECOSAR) predictive model of the U.S. Environmental Protection Agency, the result generated from SDML models should be primarily used for screening-level assessments of ecological risk due to the insufficiency of quality experimental data, while professional judgment is required to determine the appropriateness and applicability of the predictions from the SDML models.

Computer scientists recognize the problems of SDML methods and are currently working on improved strategies. First, the toxicological data collection and screening stages must be performed with extreme precision to ensure a model's accuracy and minimize the influence of outliers, such as eliminating attributes that cannot represent the sample distribution, performing multiple preprocessing steps, and selecting features. Second, integrating methods for generalization can be performed to enable the models to quickly adapt to new tasks.^{12,13} These methods include transfer learning, ensemble learning, and imbalanced learning. Third, data sharing, multimodal methods, and combining multiple algorithms and models should be encouraged to enhance model scalability.^{14,15} We believe that the future addition of advanced frameworks such as active learning and reinforcement learning will break through the limitations of SDML methods and provide new perspectives on the application of generative artificial intelligence (AI) to advance ecotoxicological research.

AUTHOR INFORMATION

Corresponding Authors

Wenhong Fan — School of Materials Science and Engineering, Beihang University, Beijing 100191, China; [orcid.org/0000-0001-5047-1714](#); Phone: +86-10-61716810; Email: fanwh@buaa.edu.cn

Fengchang Wu — State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China; [orcid.org/0000-0001-8595-2243](#); Phone: +86-10-84931804; Email: wufengchang@vip.skleg.cn

Authors

Ying Wang — School of Materials Science and Engineering, Beihang University, Beijing 100191, China; [orcid.org/0000-0002-1424-0422](#)

Jinchu Dong — School of Materials Science and Engineering, Beihang University, Beijing 100191, China

Yunchi Zhou — School of Materials Science and Engineering, Beihang University, Beijing 100191, China; [orcid.org/0000-0001-6603-2177](#)

Yinghao Cheng — School of Materials Science and Engineering, Beihang University, Beijing 100191, China; Nuclear and Radiation Safety Center, Beijing 100082, China

Xiaoli Zhao — State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China; [orcid.org/0000-0002-5671-389X](#)

Willie J. G. M. Peijnenburg — Institute of Environmental Science, Leiden University, Leiden 2300 RA, The Netherlands; National Institute of Public Health and the Environment, Center for Safety of Products and Substances, Bilthoven 3720BA, The Netherlands; [orcid.org/0000-0003-2958-9149](#)

Martina G. Vijver — Institute of Environmental Science, Leiden University, Leiden 2300 RA, The Netherlands

Kenneth M. Y. Leung — State Key Laboratory of Marine Pollution, Department of Chemistry and School of Energy and Environment, City University of Hong Kong, Hong Kong 999077, China; [orcid.org/0000-0002-2164-4281](#)

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.est.5c00510>

Notes

The authors declare no competing financial interest.

Biographies



Dr. Fengchang Wu is the director of the State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences (CRAES). His research interests mainly cover environmental toxicology and chemistry by multi-disciplinary approaches, including environmental biogeochemistry, chemistry, toxicology, and risk assessment. He has made remarkable contributions to the protection of the aquatic ecosystem in China through his advancement of environmental science, standards, technologies, and engineering. Dr. Wu was elected as an academican of the Chinese Academy of Engineering in 2017 in recognition of his work in the development of the environmental criteria/standards system in China.



Dr. Wenhong Fan is a Professor in the School of Materials Science and Engineering at Beihang University. Her research focuses on the environmental behavior, ecological effects, and remediation of metal and nanoparticle pollutants in aqueous environments, as well as the exposure, toxicity, and risk assessment of emerging pollutants. To date, she has published more than 150 international peer-reviewed papers. She serves as an Associate Editor of *Aquatic Toxicology* and an Editor of *Bulletin of Environmental Contamination and Toxicology*. She was awarded a 2022 Carbon Research Best Paper and the First Class of the Beijing Science and Technology Award from the Beijing Government.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2022YFC3204800), the National Natural Science Foundation of China (42177240, 42330710, and 42430714), the Beijing Natural Science Foundation (8242033), and the Fundamental Research Funds for the Central Universities.

M.G.V. received funding from European Union's ERC-consolidator Grant Agreement 101002123.

REFERENCES

- (1) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671–17689.
- (2) Miller, T. H.; Gallidabino, M. D.; MacRae, J. I.; Hogstrand, C.; Bury, N. R.; Barron, L. P.; Snape, J. R.; Owen, S. F. Machine Learning for Environmental Toxicology: A Call for Integration and Innovation. *Environ. Sci. Technol.* **2018**, *52* (22), 12953–12955.
- (3) Turner, B. O.; Paul, E. J.; Miller, M. B.; Barbey, A. K. Small Sample Sizes Reduce the Replicability of Task-Based fMRI Studies. *Communications Biology* **2018**, *1* (1), 62.
- (4) Faraway, J. J.; Augustin, N. H. When Small Data Beats Big Data. *Statistics & Probability Letters* **2018**, *136*, 142–145.
- (5) Zhou, Y.; Wang, Y.; Peijnenburg, W.; Vijver, M. G.; Balraadsing, S.; Dong, Z.; Zhao, X.; Leung, K. M. Y.; Mortensen, H. M.; Wang, Z.; Lynch, I.; Afantitis, A.; Mu, Y.; Wu, F.; Fan, W. Application of Machine Learning in Nanotoxicology: A Critical Review and Perspective. *Environ. Sci. Technol.* **2024**, *58* (34), 14973–14993.
- (6) Xu, P.; Ji, X.; Li, M.; Lu, W. Small Data Machine Learning in Materials Science. *npj Computational Materials* **2023**, *9* (1), 42.
- (7) Li, F.-F.; Fergus, Perona. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*; 2003; Vol. 2, pp 1134–1141.
- (8) Li, D.-C.; Wen, I.-H. A Genetic Algorithm-Based Virtual Sample Generation Technique to Improve Small Data Set Learning. *Neurocomputing* **2014**, *143*, 222–230.
- (9) Zhou, Y.; Wang, Y.; Peijnenburg, W.; Vijver, M. G.; Balraadsing, S.; Fan, W. Using Machine Learning to Predict Adverse Effects of Metallic Nanomaterials to Various Aquatic Organisms. *Environ. Sci. Technol.* **2023**, *57* (46), 17786–17795.
- (10) Schlender, T.; Viljanen, M.; van Rijn, J. N.; Mohr, F.; Peijnenburg, W. J. G. M.; Hoos, H. H.; Rorije, E.; Wong, A. The Bigger Fish: A Comparison of Meta-Learning QSAR Models on Low-Resourced Aquatic Toxicity Regression Tasks. *Environ. Sci. Technol.* **2023**, *57* (46), 17818–17830.
- (11) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* **2016**, *3* (1), No. 160018.
- (12) Huang, S.; Xia, J.; Wang, Y.; Lei, J.; Wang, G. Water Quality Prediction Based on Sparse Dataset Using Enhanced Machine Learning. *Environmental Science and Ecotechnology* **2024**, *20*, No. 100402.
- (13) Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of Classification Methods on Unbalanced Data Sets. *IEEE Access* **2021**, *9*, 64606–64628.
- (14) Heacock, M. L.; Lopez, A. R.; Amolegbe, S. M.; Carlin, D. J.; Henry, H. F.; Trottier, B. A.; Velasco, M. L.; Suk, W. A. Enhancing Data Integration, Interoperability, and Reuse to Address Complex and Emerging Environmental Health Problems. *Environ. Sci. Technol.* **2022**, *56* (12), 7544–7552.
- (15) Liu, W.; Chen, J.; Wang, H.; Fu, Z.; Peijnenburg, W. J. G. M.; Hong, H. Perspectives on Advancing Multimodal Learning in Environmental Science and Engineering Studies. *Environ. Sci. Technol.* **2024**, *58* (38), 16690–16703.