A new ChatGPT-empowered, easy-to-use machine learning paradigm for environmental science

Haoyuan An, Xiangyu Li, Yuming Huang, Weichao Wang, Yuehan Wu, Lin Liu, Weibo Ling, Wei Li, Hanzhu Zhao, Dawei Lu, Qian Liu, Guibin Jiang

PII: S2772-9850(24)00007-3

DOI: https://doi.org/10.1016/j.eehl.2024.01.006

Reference: EEHL 81

To appear in: *Eco-Environment & Health*

Received Date: 21 October 2023

Revised Date: 23 December 2023

Accepted Date: 2 January 2024

PERSPECTIVE

A new ChatGPT-empowered, easy-to-use machine learning paradigm for environmental science

Haoyuan An[a,b], Xiangyu Li[a], Yuming Huang[a], Weichao Wang[a], Yuehan Wu[a], Lin Liu[a], Weibo Ling[a], Wei Li[b], Hanzhu Zhao[b], Dawei Lu*[a], Qian Liu[a], Guibin Jiang[a]

[a] State Key Laboratory of Environmental Chemistry and Toxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

[b] Biomedical Engineering Institute, School of Control Science and Engineering, Shandong University, Jinan 250061, China

* Corresponding author. Email: dwlu@rcees.ac.cn

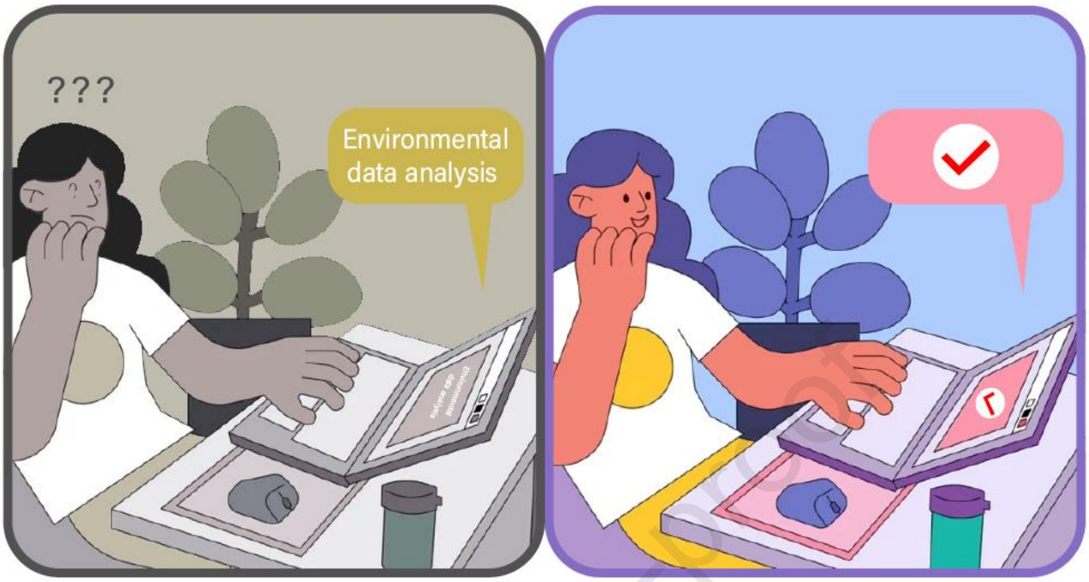
Abstract

The quantity and complexity of environmental data show exponential growth in recent years. High-quality big data analysis is critical for performing a sophisticated characterization of the complex network of environmental pollution. Machine learning (ML) has been employed as a powerful tool for decoupling the complexities of environmental big data based on its remarkable fitting ability. Yet, due to the knowledge gap across different subjects, ML concepts and algorithms have not been well-popularized among researchers in environmental sustainability. In this context, we introduce a new research paradigm—"ChatGPT + ML + Environment", providing an unprecedented chance for environmental researchers to reduce the difficulty of using ML models. For instance, each step involved in applying ML models to environmental sustainability, including data preparation, model selection and construction, model training and evaluation, and hyper-parameter optimization, can be easily performed with guidance from ChatGPT. We also discuss the challenges and limitations of using this research paradigm in the field of environmental sustainability. Furthermore, we highlight the importance of "secondary training" for future application of "ChatGPT + ML + Environment".

Keywords

Machine learning; Environmental application; ChatGPT; Secondary training

Table of Contents (TOC) Graphic



“Machine learning+Environment” “ChatGPT+Machine learning+Environment”

1. Introduction

An environmental issue usually involves multiple substances, factors, and processes, leading to the generation of environmental big data generally characterized by rich sets of input features, e.g., the data of real-time monitoring[1, 2], human activities[3-6], meteorological parameters[7-10], emission inventories[11-14], chemical composition[15, 16], environmental transportation[17, 18], and pollution exposure[19, 20]. In addition to numbers, the input formats of environmental data also include texts, graphs, and images[21]. Hence, environmental big data analysis requires more advanced approaches and powerful tools. In recent years, machine learning (ML), an emerging data mining tool for addressing the multi-dimensional/variety data[22], has triggered a revolutionary development in the field of environmental science[8, 21, 23-28]. ML is defined as "developing a model based on a set of example data, known as 'training data', to generate predictions or decisions without the need for explicit programming"[29]. ML algorithms show an excellent capacity for handling data with various input features and formats, outperforming traditional statistical tools that are often limited to data showing linear relationships with the outcomes[30-32]. It is worth noting that the dataset to be processed can be directly packaged and input into an ML model without prior knowledge of relevant features, and their patterns or trends can be identified or predicted.

In recent years, several reviews have summarized the current state of ML applications in environmental research. In 2021, Zhong et al. reported the working principles of ML algorithms and presented their specific applications in environmental pollution research, including predicting the pollution trends of atmospheric fine particulate matter ($PM_{2.5}$), predicting the future water availability, data processing from different water facilities, predicting sludge bulking in wastewater treatment plants, and identifying the Endocrine Disrupting Chemicals (EDCs)[21]. In 2022, Liu et al. summarized the new gains in using ML algorithms to study environmental issues, and highlighted their applications in estimating the health outcome of exposure[22]. Furthermore, they illustrated the importance of balancing the performance and interpretability of ML models in environmental research. Since 2022, the environmental scenarios of applying ML algorithms have been further expanded. For instance, ML algorithms have been widely used for improving the efficiency of environmental monitoring and policy-making[27], accounting carbon budget[33, 34], decoupling the meteorological impact on air pollution[9, 35], screening the new pollutants from a tremendous number of chemicals[36], predicting the health benefits through reducing pollution[37-42], identifying the impactors affecting the food chain or ecosystem[43, 44], etc. Example ML algorithms used in environmental research include recurrent neural network (RNN)[45], convolutional neural network (CNN)[46], decision tree[47], support vector machine (SVM)[48, 49], random forest (RF)[8, 10], and artificial/deep neural network[22]. Most of these ML models used in environmental research are well-developed, and their concepts, principles, and example codes are publicly shared. Despite that, environmental researchers with less experience in AI techniques still face challenges in appropriate applications of ML algorithms, e.g., misuse of cross-validation to the entire data set[21], or confusion between the validation set and test set[50]. Hence, they usually seek collaborations with researchers in the field of computing, ensuring a correct application of ML algorithms. Yet, some critical parameters for proper ML application, e.g., feature description and hyper-parameter tuning, should be drawn upon domain expertise, rather than only AI techniques[21].

ChatGPT, as a state-of-the-art version of the dialogue-based model, was launched in November 2022 and will probably simplify ML usage in environmental research[51]. Specifically, ChatGPT has been trained on a large corpus of billions of text data, and is embedded with human feedback reinforcement learning and manually supervised fine-tuning[52-55]. This enables it to naturally understand and generate the text like a human[56]. Moreover, the human-like text ability makes it an indispensable tool for handling a variety of language-based tasks, e.g., providing exampled codes of ML models and connecting up-/down-stream sections in the full-chain study mentioned above. Thus, for environmental researchers with less knowledge of ML algorithms, ChatGPT might reduce the threshold of using ML for environmental big data analysis.

Here, we present a novel research paradigm—"ChatGPT + ML + Environment" and highlight its potential in popularizing ML in the field of environmental science. We also discuss the challenges and limitations remaining in this technique. Considering the current version of ChatGPT-3.5 is mainly performed based on a general database, we give our perspectives on its performance improvement by "secondary training" with some professional databases.

Furthermore, we also discuss the possibility of coupling ChatGPT with other AI techniques, e.g., intelligent robots and console algorithms. This training provides a chance for generating an integration solution in the full-chain study of environmental sustainability.

2. A new paradigm of "ChatGPT + ML + Environment"

The workflow of ML models used in environmental research can generally be decomposed into data preparation, model selection and construction, model training and evaluation, hyper-parameter optimization, and output[57]. Note: hyper-parameter optimization means improving the performance and accuracy of the model by adjusting the hyper-parameters (parameters that cannot be learned by the model itself and require to be manually set) in the algorithm[57]. As shown in Fig. 1 and Supplementary discussion, the specific concepts, common errors, features, and example codes of solutions can be obtained by consulting ChatGPT. Therefore, the paradigm of "ChatGPT+ ML + Environment" is a promising tool that provides an unprecedented chance for inexperienced environmental researchers to address complex data analysis.

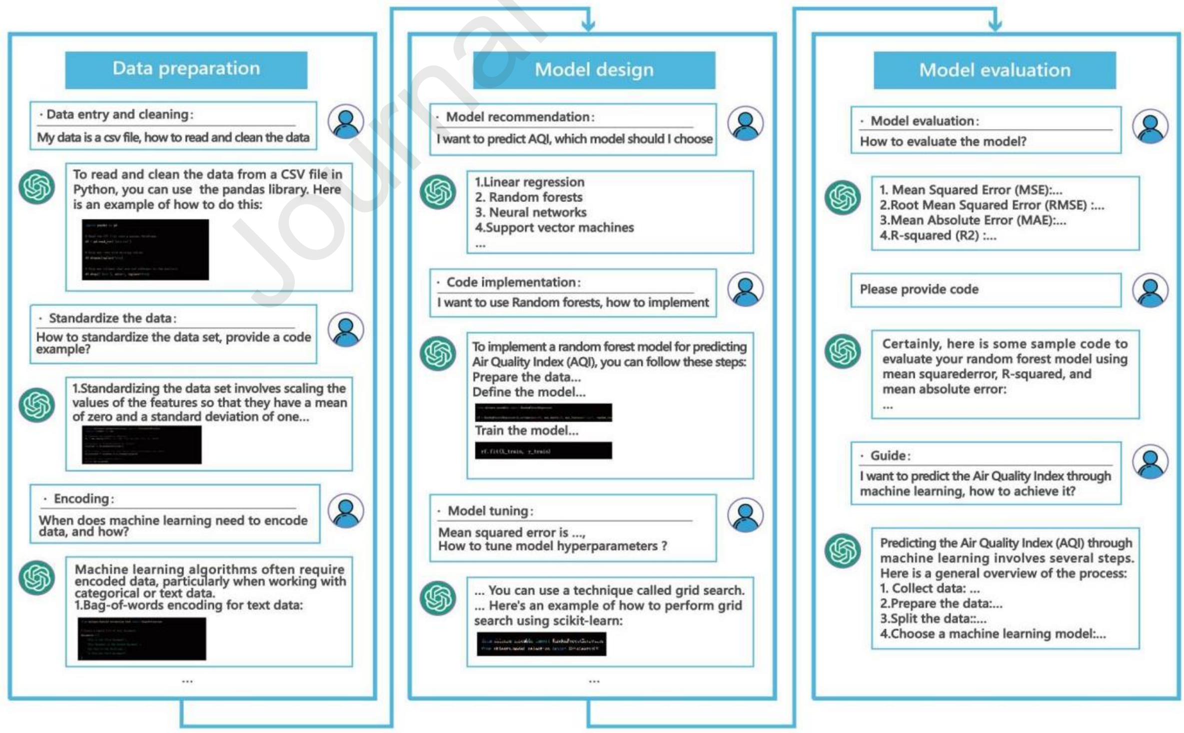


Figure 1. Schematic overview of "ChatGPT + ML + Environment". The workflow of using ML in environmental research can be roughly decomposed into data preparation, model design, and model evaluation. The dialog boxes show examples of how ChatGPT makes ML

algorithms to be easy used in environmental research.

2.1 Data preparation

The raw data of environmental analysis and monitoring usually contain a large amount of "noise" and irrelevant information, as well as incorrect, missing, or duplicate results. Moreover, some types of environmental data cannot be read by the ML model. Although some data can be directly inputted into the model, their uneven distribution also leads to unstable model training and slow model convergence. Therefore, to ensure the smooth running of ML models in environmental research, the first step is to perform data preparation of environmental big data by using some algorithms, e.g., Python's Pandas library and Scikit-Learn library[57]. Specifically, we can inquire with ChatGPT about the data preparation methods and their functions, and choose an appropriate one according to the specific formats and features of raw data (Fig. 1). Alternatively, we can also enter ChatGPT with our available data storage formats, and then guide it to provide appropriate data preparation methods (Fig. 1). Furthermore, ChatGPT can also generate the code examples for operating data preparation.

To further test the reliability of this method, we performed an example procedure of data preparation in Air Quality Index (AQI) prediction[58]. Specifically, we inputted "My data is a csv file, the columns are 'date, $PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, O_3, AQI', the date column does not need to be entered into the model, the remaining columns may be partially missing, how to read the file, perform data cleaning and divide it into a training set and a validation set?" into the ChatGPT. As shown in Supplementary discussion, ChatGPT directly provided annotated codes and their description. However, ChatGPT seemed to ignore that "the date column does not need to be entered into the model." Then, a further instruction, "I don't need the data in the date column," was entered into the ChatGPT, which provided a complete set of code and explanation. Hence, ChatGPT can help inexperienced environmental researchers achieve data preparation of complex environmental data.

2.2 Model selection and construction

As aforementioned, ML models have been widely used for environmental big data analysis, including classification, data fitting, clustering analysis, association analysis, and anomaly detection[21]. Theoretically, there are multiple ML models available that can be used to resolve the same type of task in data analysis. Yet, the model capacity, training speed, and functional focus of these ML models are different. Thus, a sophisticated analysis of the fundamentals and functional differences of the numerous models is essential for model selection. ChatGPT provides an effective solution for selecting an appropriate ML model. Specifically, we can learn about the patterns, basics and fundamentals, functional focuses, advantages, and disadvantages of the intentional-required models by inquiring with ChatGPT. It is worth noting that using ChatGPT to select an ML model only requires a few short conversations, saving considerable time compared with manual research and investigation.

Considering that different ML models have their own frameworks, the data to be processed should be optimized to achieve the requirements of the selected ML's framework. For example, if a convolutional neural network (CNN) is chosen to perform AQI prediction (Supplementary discussion), bootstrap instructions can be given to ChatGPT, such as "I want to achieve AQI prediction through a one-dimensional convolutional neural network based on the pytorch framework". Then, ChatGPT would present guidelines for converting the pending data into a readable format for Data Loader. Moreover, a complete set of "sample code" for the selected model construction can also be provided by ChatGPT (Supplementary discussion). After a slight optimization, we can easily build the selected ML model. Hyper-parameters selection, an important factor for proper model building, directly affects the capacity, convergence speed, and performance of the ML model. Particularly, some hyper-parameters (e.g., the depth of trees in the RF model) are not fixed options, which should be set with a comprehensive account of the number of input data features, data volume, data distribution, and application scenario, etc[21]. Considering that hyper-parameters selection is a dilemma that involves the knowledge of AI and environmental science, inexperienced environmental researchers can seek solutions with the support of ChatGPT. Although ChatGPT might not provide optimum parameter settings, it can provide the detailed meaning of each hyper-parameter and advanced methods (e.g., grid search) for proper selection. Thus, ChatGPT can guide the ML model building in the field of environmental science.

To illustrate how to select the most appropriate ML mode, we performed an exampled case of the Shannon index (a critical indicator for measuring biodiversity) prediction with the parameters of nanoparticles (e.g., type, shape, size, potential) and relevant environmental factors (e.g., temperature, pH, soil depth). For instance, we performed an original prediction with linear regression based on this ChatGPT-empowered system. Then, "Can any other model be used to achieve this prediction? Output the performance of each model and select the best one." was inputted into the ChatGPT-empowered system. As shown in Supplementary discussion, the ChatGPT-empowered system provided the codes of linear regression, random forest, and xgb tree models, and output the name and RMSE (Root Mean Square Error) of the most suitable model. Moreover, the ChatGPT-empowered system can provide codes of cross-validation to evaluate the performance of these models. It can also search the most suitable parameters on the internet automatically. For the whole process, we merely provided the output and error message from the last step for ChatGPT, which then generated the subsequent codes of correction and implementation automatically.

2.3 Model training, performance, and hyper-parameter optimization

ChatGPT can further guide the training, performance evaluation, and hyper-parameter optimization of the ML models used in environmental research. For traditional ML models like RF and SVM, most of their codes used for model training are with fixed structures[21, 22]. The corresponding statements and structures can usually be found by ChatGPT in the database of code examples. For instance, the training procedure of the RF model for air quality (AQI) prediction from emissions was smoothly performed with guidance from ChatGPT (Supplementary discussion). With regard to deep learning models, to reduce running problems (e.g., convergence difficulties and declining model generalization ability), the parameters, including learning rate, optimizer, and learning rate decay, are required to be set prior to training[22]. Taking an example of AQI prediction by using CNN (Supplementary discussion), the parameters including adam optimizer, learning rate (0.001), and mean squared error loss were successfully set guided by ChatGPT. Moreover, to further optimize the training process, the procedures of gradient descent and backpropagation, and the codes for learning rate decay were also provided by ChatGPT.

Model performance is critical for ML applications, determining the reliability of prediction[57]. Although there are many ways to evaluate an ML model's performance, some evaluation parameters involve computer terminology and are difficult to understand for environmental researchers. ChatGPT can provide formulas, meanings, and examples of application scenarios of the various evaluation parameters for users to understand and select appropriate evaluation methods. Specifically, we can obtain the "Mean Squared Error," "Root Mean Squared Error," "Mean Absolute Error," and "R-squared" of the models used in AQI predictions via inquiring with ChatGPT (Fig.1, Supplementary discussion). More importantly, the implementation codes for model evaluation can be accessed directly from the package provided by ChatGPT. Furthermore, tuning hyper-parameters is usually required to further improve the model performance. Similar to hyper-parameters selection (*see section 2.2*), we can obtain specific tuning codes of the selected model, and find the optimum hyper-parameters by ChatGPT.

The aforementioned applications mainly tend to directly use or make slight modifications to the existing code structures. In these applications, ChatGPT can provide clear and concise code examples, preventing us from spending tremendous time studying the user manual of various ML models. This is of extreme importance for those with less knowledge in ML programming, as it can greatly reduce the interference and misdirection caused by complex codes. Additionally, ChatGPT can provide code interpretation and error-checking assistance, enabling us to quickly grasp the logical framework of a code segment and apply it to environmental studies. To facilitate understanding, the whole process of application examples based on the paradigm of "ChatGPT + ML + Environment" has been successfully performed, as detailed in Supplementary discussion.

3 Advancement and challenges

In addition to the aforementioned text data processing, the ChatGPT-empowered system also shows advantages in processing complex data. For instance, it can be used to predict the toxicology of chemicals based on their physical-chemical properties dataset (see Supplementary discussion). The used dataset consists of 210 features, including a series of specific chemical descriptors (e.g., molecular structure, chemical name, source, and CAS number), a range of refined molecular properties (e.g., polar surface area, adsorption properties, the quantity, state, and size of atoms and functional groups), and some important physicochemical properties (e.g., solubility, lipophilicity, and surface area). Considering that the dataset is a mixture of both useful and irrelevant information, including numerical and character-based data, we initially used the ChatGPT-3.5 to generate the code of a random forest model, yielding an RMSE of 1.39. To address the possible limitations of ChatGPT-3.5 missing some contextual information in complex datasets, we further performed this prediction by using the ChatGPT4.0-empowered system. As shown in Supplementary discussion, the RMSE is 0.67 with an R-squared (R^2) of 0.57, which demonstrates the potential of the ChatGPT-empowered system in addressing complex ML tasks.

However, ChatGPT, one of the first human-like language models, still faces challenges and limitations in environmental applications. For instance, 1) Honest use. Most of ChatGPT's output is difficult to distinguish from the text written by humans. Recently, ChatGPT was directly listed as the author of several publications, which has triggered a widespread discussion among the academic community[53-55]. Indeed, the use of ChatGPT must strictly adhere to academic ethics and standards. To popularize the applications of public-shared tools (i.e., ML) in the field of environmental science, the details of ChatGPT usage should be clearly disclosed in the publications. Furthermore, for better regulation, the usage record can be documented accurately with the time stamp in blockchain technique. 2) Model development. The training of ChatGPT is still based on a large amount of existing data. Therefore, ChatGPT can provide code examples for the well-developed ML models used in environmental research but fails to develop new models. As shown in Supplementary discussion, the ChatGPT-empowered system can perform almost all ML tasks in environmental science. Yet, it is still a probability-based AI model[51]. Its responses are the results of analyzing a large amount of training data, lacking thought of the context and background information. Therefore, it may not understand why we perform these analyses, and hence the whole data processing strategy should be designed by the researchers. Moreover, ChatGPT would be unaware of the parameter errors existing in its generated codes, which can only be found when the codes are actually executed. 3) Professional database. The current ChatGPT database is limited to general data prior to 2021[51, 53], lacking a professional dataset of environmental sustainability. This may result in suboptimal performance in solving environmental problems. Therefore, the ChatGPT-empowered plug-in can be embedded into the professional system of environmental research to promptly provide ML applications. Additionally, to obtain high-quality big data analysis, some environmental data are encouraged to be open to the public.

4. Discussion

Although ML is a powerful tool for addressing complex environmental problems, it can be a challenging task for environmental scientists without AI research backgrounds. Integrating ChatGPT can provide effective solutions, including the concepts, principles and exampled codes, for ML applications. For environmental researchers with no prior knowledge, it can help them to perform ML analysis smoothly; for scientists with some AI knowledge, this process will improve their efficiency by saving their time to edit the codes. Notably, almost all programming tools or languages like Python and R can be used to build the ChatGPT-based process. In addition to environmental science, this process will extend ML application to other fields, e.g., industrial, biology, and geochemistry. Furthermore, it is noted that other Generative Pre-trained Transformer-based tools like Claude and Bard have similar effects as the ChatGPT[51], reducing the threshold of environmental application of ML. With the development of generative models and AI technologies, the application of the "ChatGPT + ML + Environment" research paradigm will be further expanded. For instance, the processed data will not be limited to text, and graphic data might be understood and processed as the ChatGPT evolves [53]. In the future, these techniques, used correctly in accordance with academic ethics and usage guidelines, would provide excitement for solving complex environmental problems:

1) Enhancing "secondary training" based on professional datasets. As shown in Fig. 2, the first step involves choosing a certain type of environmental case (e.g., environmental monitoring, source tracing, and policy making) and introducing a specific professional dataset. Moreover, a standard description file of the professional dataset, including dataset format, data types, additional data description, number of data entries, and dataset content description, should be set for the system of "ChatGPT + ML + Environment." This step will help ChatGPT to learn about the overview of the dataset. Afterward, a "secondary training" model, including the framework of data processing, the code for data preparation, model construction, and performance evaluation, would be built for the professional dataset. The detailed implementation procedures are similar to that mentioned in *Section 2*. Through further training or optimization, the "secondary training" model would show a capacity to provide effective and quick solutions for such environmental problems, especially for some emergency events.

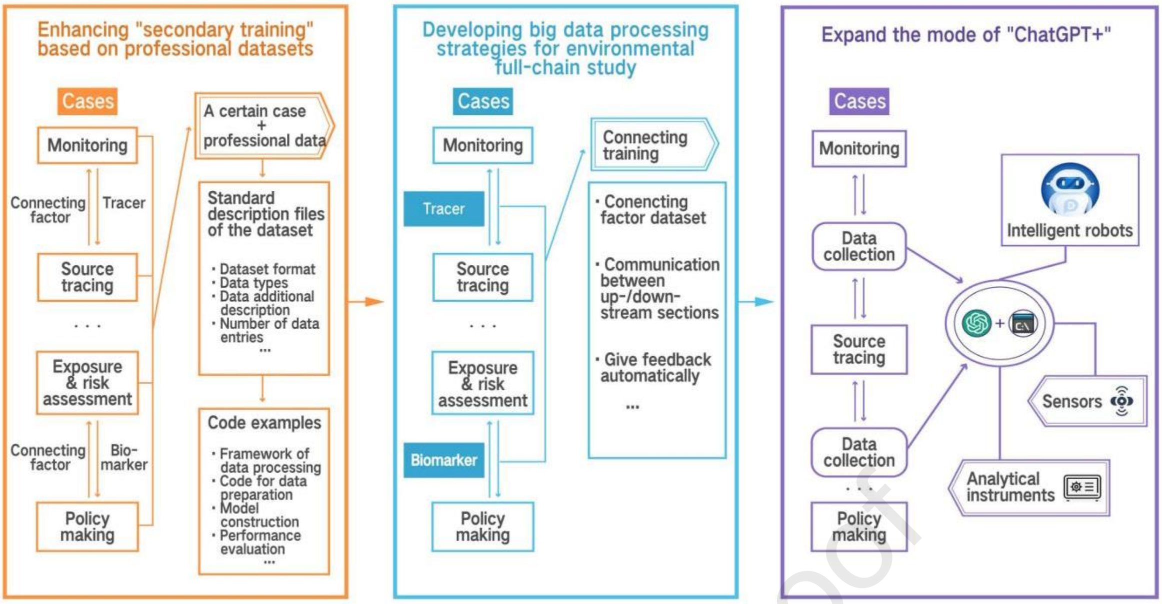


Figure 2. The conceptual mode of "ChatGPT + ML + Environment" in future environmental research. The left box shows the secondary training by introducing environmental professional dataset. The middle box mainly shows the potential in connecting the up-/down-stream tasks of data analysis in the full chain study of environmental sustainability. The right box mainly gives a perspective on coupling data processing with data collection via using an integration of ChatGPT, control algorithms, ML, and robots, etc.

2) Developing big data processing strategies for full-chain environmental study. An environmental event usually involves the coupling of multiple substances, factors, and processes across various scales, requiring a comprehensive research route covering "monitoring—source tracing—environmental behavior and transformation—exposure and risk assessment—policy making." Each of them can generate different datasets (Fig. 2). These datasets might have become "data islands" due to a lack of proper data analysis techniques, hampering the proposal of a systematic solution for real environmental problems[22]. Identifying the connection factors and developing an intelligent data processing system is critical for achieving full-chain environmental study. For instance, we would first establish a dataset composed of connection factors (Fig. 2), e.g., tracers, transformation reactions, biomarkers, and policy implementation date. The specific communication instructions for connecting up-/down-stream sections would be well-trained by ChatGPT with its human-like text ability[54]. In this way, the ML-based data processing in a down-stream section can be

operated automatically after receiving the output from the up-stream section. Alternatively, they can provide feedback of the output to the up-stream section, guiding its optimization. Thus, the integration of ChatGPT and ML algorithms is a promising tool for future full-chain environmental research.

3) Expanding the application mode of "ChatGPT +". The integration of ChatGPT and ML significantly improves the processing capacity of environmental big data, promoting the rapid development of environmental science. For instance, the current environmental monitoring system is capable of continuously collecting real-time environmental data and outputting brief reports[48, 58]. Such operations are tasks consisting of specific sequences of steps, where the execution of each task is based on previously normalized instructions. However, these tasks pose challenges in terms of generating predictions, making decision, and developing smart feedback to optimize the next step of data collection. In the future, the "ChatGPT + ML" mode can be further expanded by combining with other intelligent techniques like intelligent robots and control algorithms. Specifically, multiple environmental data collection devices (e.g., intelligent robots, sensors, and analytical instruments) and their carriers would be connected by the "ChatGPT + ML" system integrated with computer control algorithms (Fig. 2). This will integrate static environmental big data processing with dynamic environmental analysis, providing a novel tool for future environmental research, especially for some environmental monitoring under extreme conditions.

Declaration of competing interests

The authors declare no competing interests.

Acknowledgments

This work was financially supported by the National Key R&D Program of China (No. 2023YFF0614200), National Natural Science Foundation of China (Nos. 22222610, 22376202, 22193051), and the Chinese Academy of Sciences (Nos. ZDBS-LY-DQC030, YSBR-086). D. L. acknowledges the support from the Youth Innovation Promotion Association of CAS.

References

[1] G. Geng, Q. Xiao, S. Liu, X. Liu, J. Cheng, Y. Zheng, et al., Tracking air pollution in China: Near real-time $PM_{2.5}$ retrievals from multisource data fusion, Environ. Sci. Technol. 55 (2021) 12106.

[2] H. Messer, A. Zinevich, P. Alpert, Environmental monitoring by wireless communication networks, Science 312 (2006) 713.

[3] D.G. Streets, H.M. Horowitz, D.J. Jacob, Z. Lu, L. Levin, A.F.H. ter Schure, et al., Total mercury released to the environment by human activities, Environ. Sci. Technol. 51 (2017) 5969.

[4] A. Pruden, M. Arabi, H.N. Storteboom, Correlation between upstream human activities and riverine antibiotic resistance genes, Environ. Sci. Technol. 46 (2012) 11541.

[5] D.G. Streets, M.K. Devane, Z. Lu, T.C. Bond, E.M. Sunderland, D.J. Jacob, All-time releases of mercury to the atmosphere from human activities, Environ. Sci. Technol. 45 (2011) 10485.

[6] A.R. Ferro, R.J. Kopperud, L.M. Hildemann, Source strengths for indoor human activities that resuspend particulate matter, Environ. Sci. Technol. 38 (2004) 1759.

[7] J. Klánová, P. Èupr, J. Kohoutek, T. Harner, Assessing the influence of meteorological parameters on the performance of polyurethane foam-based passive air samplers, Environ. Sci. Technol. 42 (2008) 550.

[8] Z. Shi, C. Song, B. Liu, G. Lu, J. Xu, T. Van Vu, et al., Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns, Sci. Adv. 7 (2021) eabd6696.

[9] P. Zuo, Y. Huang, P. Liu, J. Zhang, H. Yang, L. Liu, et al., Stable iron isotopic signature reveals multiple sources of magnetic particulate matter in the 2021 Beijing sandstorms, Environ. Sci. Technol. Lett. 9 (2022) 299.

[10] P. Zuo, Z. Zong, B. Zheng, J. Bi, Q. Zhang, W. Li, et al., New insights into unexpected severe $PM_{2.5}$ pollution during the SARS and COVID-19 pandemic periods in Beijing, Environ. Sci. Technol. 56 (2022) 155.

[11] J. Hao, H. Tian, Y. Lu, Emission inventories of NOx from commercial energy consumption in China, 1995−1998, Environ. Sci. Technol. 36 (2002) 552.

[12] K. Breivik, V. Vestreng, O. Rozovskaya, J.M. Pacyna, Atmospheric emissions of some POPs in Europe: a discussion of existing inventories and data needs, Environ. Sci. Policy 9 (2006) 663.

[13] X. Lu, X. Ye, M. Zhou, Y. Zhao, H. Weng, H. Kong, et al., The underappreciated role of agricultural soil nitrogen oxide emissions in ozone pollution regulation in North China, Nat. Commun. 12 (2021) 5021.

[14] M. Li, H. Liu, G. Geng, C. Hong, F. Liu, Y. Song, et al., Anthropogenic emission inventories in China: a review, Natl. Sci. Rev. 4 (2017) 834.

[15] X. Feng, H. Sun, X. Liu, B. Zhu, W. Liang, T. Ruan, et al., Occurrence and ecological impact of

chemical mixtures in a semiclosed sea by suspect screening analysis, Environ. Sci. Technol. 56 (2022) 10681.

[16] S. Breinlinger, T.J. Phillips, B.N. Haram, J. Mareš, J.A. Martínez Yerena, P. Hrouzek, et al., Hunting the eagle killer: A cyanobacterial neurotoxin causes vacuolar myelinopathy, Science 371 (2021) eaax9050.

[17] Z. Tian, H. Zhao, K.T. Peter, M. Gonzalez, J. Wetzel, C. Wu, et al., A ubiquitous tire rubber-derived chemical induces acute mortality in coho salmon, Science 371 (2021) 185.

[18] Y. Yin, Y. Li, C. Tai, Y. Cai, G. Jiang, Fumigant methyl iodide can methylate inorganic mercury species in natural waters, Nat. Commun. 5 (2014) 4633.

[19] D. Lu, Q. Luo, R. Chen, Y. Zhuansun, J. Jiang, W. Wang, et al., Chemical multi-fingerprinting of exogenous ultrafine particles in human serum and pleural effusion, Nat. Commun. 11 (2020) 2567.

[20] R. Vermeulen, E.L. Schymanski, A.-L. Barabási, G.W. Miller, The exposome and health: Where chemistry meets biology, Science 367 (2020) 392.

[21] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, et al., Machine learning: New ideas and tools in environmental science and engineering, Environ. Sci. Technol. 55 (2021) 12741.

[22] X. Liu, D. Lu, A. Zhang, Q. Liu, G. Jiang, Data-driven machine learning in environmental pollution: Gains and problems, Environ. Sci. Technol. 56 (2022) 2124.

[23] Z. Cao, J. Zhou, M. Li, J. Huang, D. Dou, Urbanites' mental health undermined by air pollution, Nat. Sustain. (2023).

[24] W. Li, W.-Y. Guo, M. Pasgaard, Z. Niu, L. Wang, F. Chen, et al., Human fingerprint on structural density of forests globally, Nat. Sustain. (2023).

[25] M. Toetzke, N. Banholzer, S. Feuerriegel, Monitoring global development aid with machine learning, Nat. Sustain. 5 (2022) 533.

[26] Z. Mehrabi, M.J. McDowell, V. Ricciardi, C. Levers, J.D. Martinez, N. Mehrabi, et al., The global divide in data-driven farming, Nat. Sustain. 4 (2021) 154.

[27] M. Hino, E. Benami, N. Brooks, Machine learning for environmental monitoring, Nat. Sustain. 1 (2018) 583.

[28] M. Callaghan, C.-F. Schleussner, S. Nath, Q. Lejeune, T.R. Knutson, M. Reichstein, et al., Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies, Nat. Clim. Change 11 (2021) 966.

[29] J.R. Koza, F.H. Bennett, D. Andre, M.A. Keane, in: J.S. Gero and F. Sudweeks (Eds.), Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming, Springer Netherlands. Dordrecht, 1996, pp. 151.

[30] D. Seng, Q. Zhang, X. Zhang, G. Chen, X. Chen, Spatiotemporal prediction of air quality based on LSTM neural network, Alex. Eng. J. 60 (2021) 2021.

[31] Y. Zhao, L. Wang, J. Luo, T. Huang, S. Tao, J. Liu, et al., Deep learning prediction of polycyclic aromatic hydrocarbons in the high arctic, Environ. Sci. Technol. 53 (2019) 13238.

[32] R. Janarthanan, P. Partheeban, K. Somasundaram, P. Navin Elamparithi, A deep learning approach for prediction of air quality index in a metropolitan city, Sustain. Cities Soc. 67 (2021) 102720.

[33] M. Mugabowindekwe, M. Brandt, J. Chave, F. Reiner, D.L. Skole, A. Kariryaa, et al., Nation-wide mapping of tree-level aboveground carbon stocks in Rwanda, Nat. Clim. Change 13 (2023) 91.

[34] Z. Ban, X. Hu, J. Li, Tipping points of marine phytoplankton to multiple environmental stressors, Nat. Clim. Change 12 (2022) 1045.

[35] Z. Zhang, B. Xu, W. Xu, F. Wang, J. Gao, Y. Li, et al., Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on $PM_{2.5}$ pollution, Environ. Res. 212 (2022) 113322.

[36] D. Xia, J. Chen, Z. Fu, T. Xu, Z. Wang, W. Liu, et al., Potential application of machine-learning-based quantum chemical methods in environmental chemistry, Environ. Sci. Technol. 56 (2022) 2115.
[37] J. Jeong, J. Choi, Artificial intelligence-based toxicity prediction of environmental chemicals: Future directions for chemical management applications, Environ. Sci. Technol. 56 (2022) 7532.
[38] L. Conibear, C.L. Reddington, B.J. Silver, Y. Chen, C. Knote, S.R. Arnold, et al., Sensitivity of air pollution exposure and disease burden to emission changes in China using machine learning emulation, GeoHealth 6 (2022) e2021GH000570.
[39] E. Isaev, B. Ajikeev, U. Shamyrkanov, K.-u. Kalnur, K. Maisalbek, R.C. Sidle, Impact of climate change and air pollution forecasting using machine learning techniques in Bishkek, Aerosol Air Qual. Res. 22 (2022) 210336.
[40] L. Zhang, X. Li, H. Chen, Z. Wu, M. Hu, M. Yao, Haze air pollution health impacts of breath-borne VOCs, Environ. Sci. Technol. 56 (2022) 8541.
[41] G.D. Thurston, L.C. Chen, M. Campen, Particle toxicity's role in air pollution, Science 375 (2022) 506.
[42] H. Tan, J. Wu, R. Zhang, C. Zhang, W. Li, Q. Chen, et al., Development, validation, and application of a human reproductive toxicity prediction model based on adverse outcome pathway, Environ. Sci. Technol. 56 (2022) 12391.
[43] C. Zhan, H. Matsumoto, Y. Liu, M. Wang, Pathways to engineering the phyllosphere microbiome for sustainable crop production, Nat. Food 3 (2022) 997.
[44] H. Meyer, E. Pebesma, Machine learning-based global maps of ecological variables and the challenge of assessing them, Nat. Commun. 13 (2022) 2208.
[45] G. Kurnaz, A.S. Demir, Prediction of SO_2 and PM_{10} air pollutants using a deep learning-based recurrent neural network: Case of industrial city Sakarya, Urban Clim. 41 (2022) 101051.
[46] V. Nikolopoulou, R. Aalizadeh, M.-C. Nika, N.S. Thomaidis, TrendProbe: Time profile analysis of emerging contaminants by LC-HRMS non-target screening and deep learning convolutional neural network, J. Hazard. Mater. 428 (2022) 128194.
[47] A. Coors, A.R. Brown, S.K. Maynard, A. Nimrod Perkins, S. Owen, C.R. Tyler, Minimizing experimental testing on ffish for legacy pharmaceuticals, Environ. Sci. Technol. 57 (2023) 1721.
[48] F. Camastra, V. Capone, A. Ciaramella, A. Riccio, A. Staiano, Prediction of environmental missing data time series by Support Vector Machine Regression and Correlation Dimension estimation, Environ. Modell. Softw. 150 (2022) 105343.
[49] X.-C. Song, N. Dreolin, E. Canellas, J. Goshawk, C. Nerin, Prediction of collision cross-section values for extractables and leachables from plastic products, Environ. Sci. Technol. 56 (2022) 9463.
[50] M. Lastra-Mejias, A. Villa-Martinez, M. Izquierdo, R. Aroca-Santos, J.C. Cancilla, J.S. Torrecilla, Combination of LEDs and cognitive modeling to quantify sheep cheese whey in watercourses, Talanta 203 (2019) 290.
[51] ChatGPT: optimizing language models for dialogue, https://openai.com/blog/chatgpt, (2022).
[52] Much to discuss in AI ethics, Nat. Mach. Intell. 4 (2022) 1055.
[53] C. Stokel-Walker, AI bot ChatGPT writes smart essays — should professors worry?, (2022) https://doi.org/10.1038/d41586.
[54] M. Hutson, Could AI help you to write your next paper?, Nature 611 (2022) 192.
[55] J.B. Eva A. M. van Dis, Willem Zuidema, Robert van Rooij, Claudi L. Bockting, ChatGPT: five priorities for research, Nature 614 (2023) 224.
[56] The AI writing on the wall, Nat. Mach. Intell. 5 (2023) 1.
[57] L. Bottou, Stochastic gradient descent tricks in Neural Networks: Tricks Trade, Berlin,

Germany:Springer, 7700 (2012).
[58] $PM_{2.5}$ Prediction Based on Random Forest Algorithm., https://github.com/StephenZheng0315/PM2.5-Prediction-Based-on-Random-Forest-Algorithm. (2023).

Highlights

- A new paradigm of "ChatGPT + Machine learning (ML) + Environment" is presented.
- The novelty and knowledge gaps of ML for decoupling the complexity of environmental big data are discussed.
- The new paradigm guided by GPT reduces the threshold of using Machine Learning in environmental research.
- The importance of "secondary training" for using "ChatGPT + ML + Environment" in the future is highlighted.