Potential application of machine-learning-based quantum chemical methods in environmental chemistry

Xia, D.; Chen, J.; Fu, Z.; Xu, T.; Wang, Z.; Wenjia, W.; ... ; Peijnenburg, W.J.G.M.

Environmental Science & Technology

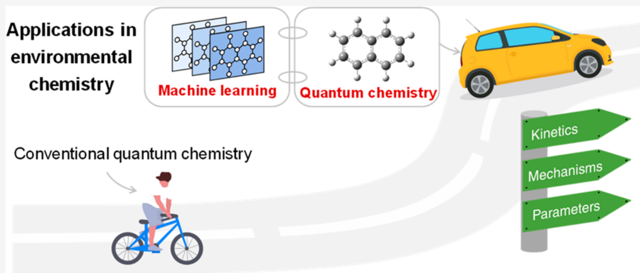
pubs.acs.org/est

Perspective

Potential Application of Machine-Learning-Based Quantum Chemical Methods in Environmental Chemistry

Deming Xia, Jingwen Chen,* Zhiqiang Fu, Tong Xu, Zhongyu Wang, Wenjia Liu, Hong-bin Xie, and Willie J. G. M. Peijnenburg

Cite This: https://doi.org/10.1021/acs.est.1c05970



ABSTRACT: It is an important topic in environmental sciences to understand the behavior and toxicology of chemical pollutants. Quantum chemical methodologies have served as useful tools for probing behavior and toxicology of chemical pollutants in recent decades. In recent years, machine learning (ML) techniques have brought revolutionary developments to the field of quantum chemistry, which may be beneficial for investigating environmental behavior and toxicology of chemical pollutants. However, the ML-based quantum chemical methods (ML-QCMs) have only scarcely been used in environmental chemical studies so far. To promote applications of the promising methods, this Perspective summarizes recent progress in the ML-QCMs and focuses on their potential applications in environmental chemical studies that could hardly be achieved by the conventional quantum chemical methods. Potential applications and challenges of the ML-QCMs in predicting degradation networks of chemical pollutants, searching global minima for atmospheric nanoclusters, discovering heterogeneous or photochemical transformation pathways of pollutants, as well as predicting environmentally relevant end points with wave functions as descriptors are introduced and discussed.

KEYWORDS: *machine learning, quantum chemistry, environmental process, environmental computational toxicology, chemicals management*

1. INTRODUCTION

Over 350 000 chemicals and their mixtures have been registered for utilization in the global market.[1] These chemicals can be released into the environment and become pollutants threatening human and ecosystem health.[2] It is a prerequisite for preventing pollution of these chemicals that their environmental behavior and toxicological effects to humans and ecological species be understood (Figure 1). Due to the wide diversity of the chemical composition of pollutants, of the environmental media under different conditions, and of the different biological systems, it is time-consuming, expensive, and also impossible to empirically determine all parameters required for quantifying the environmental fate and toxicological effects of all chemical pollutants.[2,3] Prediction based on quantum chemical methods (QCMs) that solve the Schrödinger equation (or its variants) to obtain parameters for environmental behavior and toxicological effects of chemical pollutants is becoming an appealing alternative.[4]

However, it is also time-consuming to directly solve the Schrödinger equation.[5] To overcome this obstacle, various alternative QCMs have been developed, such as the density functional theory (DFT) that replaces the Schrödinger equation with some easily solved equations for electron densities.[6,7] Even so, the DFT method is still too costly in time to be applied to relatively large systems.[5]

In recent years, machine learning (ML) has gained the increasing interest of quantum chemists. ML-based QCMs (ML-QCMs) were even considered as the next big leap in the evolution of computational chemistry, similar to the development of DFT (the 1998 Nobel Prize in Chemistry) and hybrid quantum-mechanical/molecular-mechanical (the 2013 Nobel Prize in Chemistry) methods.[8] ML can significantly improve the speed of quantum chemical calculations with a negligible loss in accuracy.[9−11] One important reason lies in that, at least in theory, ML models can learn any input−output relations even as complex as the Schrödinger equations.[8−11] Therefore, the Schrödinger equations can be in turn replaced by the ML models that can be comparatively easily solved.[8−11]

To date, only a few studies in the field of environmental chemistry have applied the ML-QCMs, although ML has been

Received: September 3, 2021

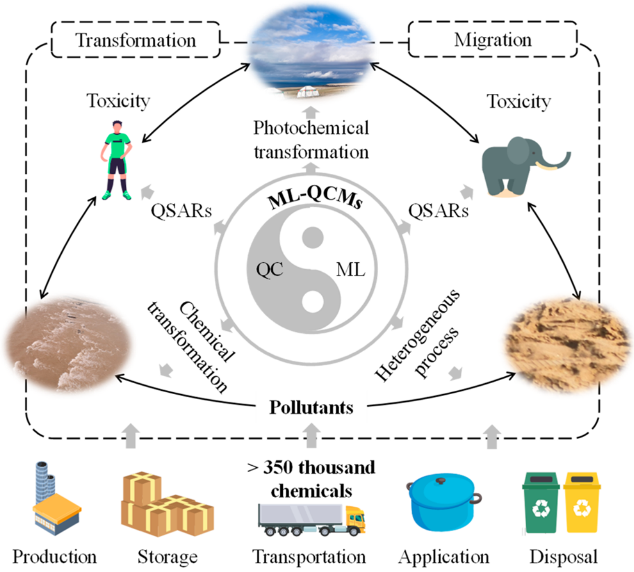


Figure 1. Role of machine-learning-based quantum chemical methods (ML-QCMs) in environmental chemical studies (QSAR, quantitative structure−activity relationship; QC, quantum chemistry; ML, machine learning).

adopted in environmental studies (e.g., prediction models on particulate matter concentrations and water resource availability[12]). As several recent reviews elaborated the methodology of the ML-QCMs,[8−11] this Perspective briefly summarizes the methodology and focuses on its potential applications (Figure 1) as well as challenges that lie ahead in environmental chemical investigations.

2. MACHINE-LEARNING-BASED QUANTUM CHEMICAL METHOD

As Dirac mentioned in 1929: "*It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.*[13]" It is a dream of quantum chemists to develop methods to accurately describe many-body systems with computational costs as low as possible. In the past decade, new tools from the rapidly developing field of ML have emerged to significantly impact the development of approximate methods for complex many-body systems, by passing or assisting the direct solution of the many-body Schrödinger equations.[8−11,14]

There are generally two types of philosophy to build ML-based quantum chemical models: supervised-learning (type-I) and unsupervised learning (type-II)-based methods. Recently, some pioneering works[15,16] also employed reinforcement-learning-based methods (type-III) to enhance sampling for molecular dynamics simulations. However, the type-III method has not been directly adopted for predicting chemical end points (e.g., molecular orbitals and wave functions) so far.[17,18] This Perspective mainly focuses on the type-I and type-II methods. Some potential applications of the type-III method are also briefly discussed.

As shown in Figure 2, for the type-I method, "machines" can be trained based on given end points and inputs using the supervised learning algorithms. Existing type-I models can also be roughly distinguished into two types according to their architectures: descriptor-based models or end-to-end models.

The descriptor-based type-I method is similar to the quantitative structure activity relationship (QSAR) methodology.[19−21] The end points for the type-I methods are usually basic molecular properties (e.g., electronic energies, electron densities, and molecular orbitals), while the end points for QSARs in environmental chemical studies are in general more complex (e.g., various partition coefficients, protein binding constants, reaction rate coefficients and toxicities).[19−21] The type-I models were conventionally trained using algorithms such as neural networks (NN), support vector machines, the Gaussian process for regression, and kernel ridge regression.[22−31] An advantage of these descriptor-based models is that the computational complexity is low and generally linearly scaled with regard to data quantity.[18]

The end-to-end type-I models directly connect chemical structures represented by selected architecture (e.g., SchNet[32] and PaiNN[33] architectures) with concerned end points. No additional descriptor is required by the end-to-end models. The computational complexity of the end-to-end type-I

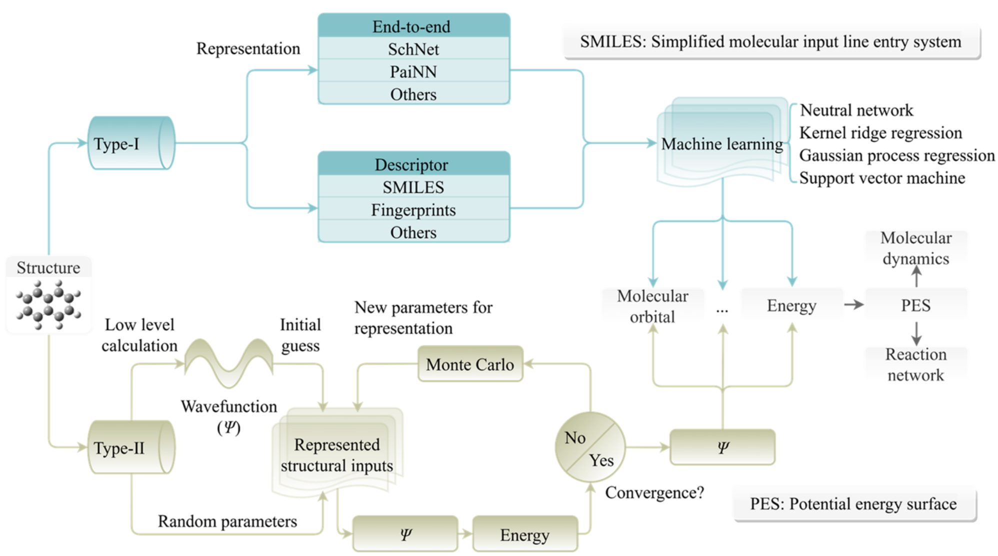


Figure 2. Type-I and type-II methods of ML-based quantum chemical models.

https://doi.org/10.1021/acs.est.1c05970
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

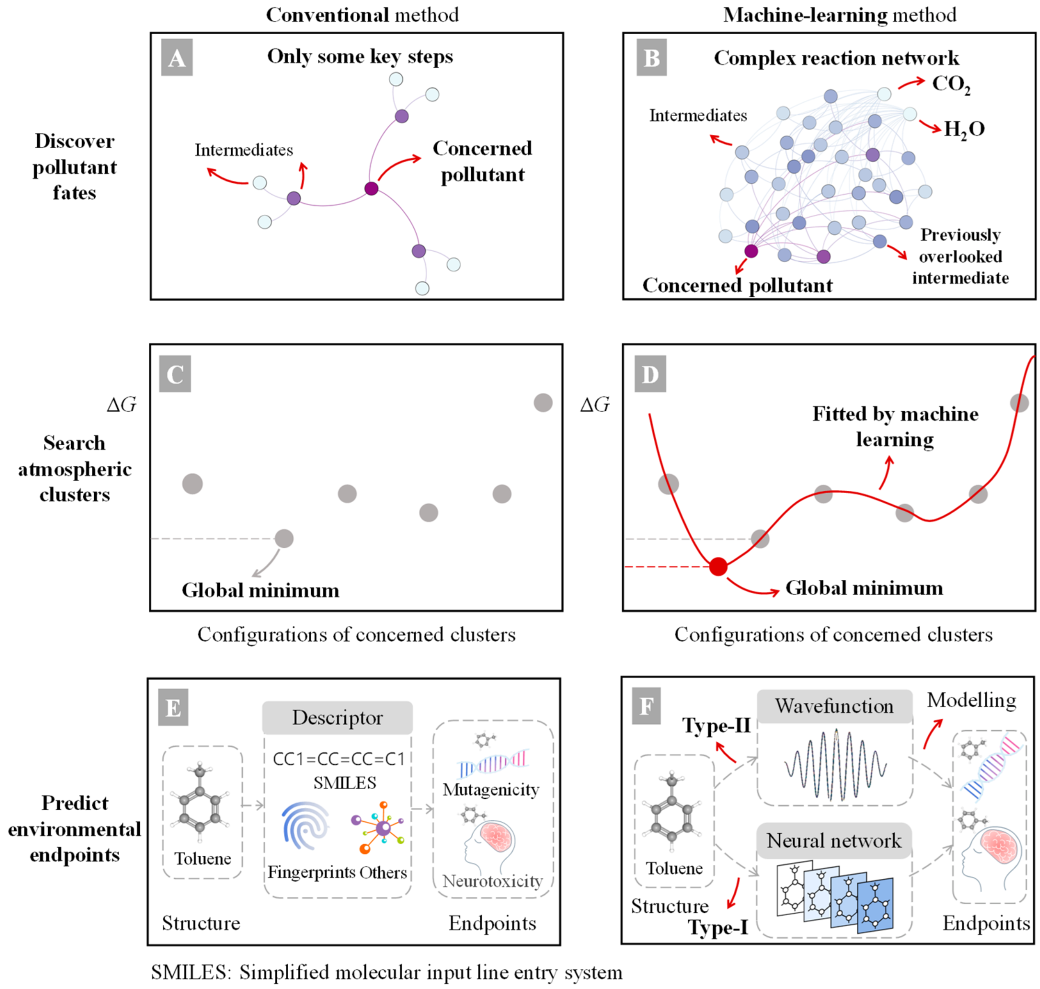


Figure 3. Conventional (left panel) and machine-learning-based (right panel) quantum chemical methods for probing behavior of chemical pollutants initiated by radicals (A and B), for obtaining the global minimum of concerned clusters during new particle formation processes (C and D), and for predicting environmental chemical end points of concerned pollutants/chemicals (E and F), where in F, the mapping relationship between structures and concerned end points can be built using neural networks like the end-to-end type-I method, and the wave functions predicted using the type-II method can be employed as "descriptors" to predict the end points.

method is roughly proportional to N^2 where N stands for data quantity.[18]

The type-I approach entirely avoids the solution of Schrödinger equations, at the price of data sets obtained by, for instance, the DFT methods.[6,7,34] Databases such as QM7,[35] QM9,[36] QM-sym,[37] PubChemQC,[38] and recently developed ORD[39] can be adopted in the type-I modeling. Most existing ML-based quantum chemical models were constructed in this way.

In the type-II method, the wave functions (or electron densities) that describe the probability of a particle's quantum states can in most cases be directly predicted by a NN with some trainable parameters.[40−42] In each step of the training, a new group of the parameters is autogenerated by the Monte Carlo method.[43−47] The corresponding energies of the wave functions characterized by the parameters can then be calculated. The training processes are not stopped until the energy change between two steps of the calculations reaches prespecified convergence criteria. The wave functions and energies for the last training step are outputs. Carleo and Troyer[41] adopted this type of thinking to establish mapping from spin configurations represented by NN parameters to corresponding wave functions via a restricted Boltzmann machine (a type of unsupervised NN). The initial NN parameters can be either guessed based on some quantum chemical methods (e.g., the Hartree−Fock and DFT) or randomly generated ones.[41,42]

As all data except for the initial guesses are automatically generated by the Monte Carlo method, the type-II method requires no pre-existing data on concerned end points. Another advantage of the method is that the models can provide wave functions (or electron densities) of investigated systems that contain all information for the systems at the simulated quantum states (e.g., ground states or first excited states).[11] The type-II method could provide more information for solving many problems and allow for predicting other concerned properties. A current limitation is that the type-II method can only be used for small molecules (typically within 30 electrons[42]) due to computational costs.

On the basis of the energies obtained via the ML-QCMs, potential energy surfaces (PES), ML potentials for molecular dynamics simulations, and reaction networks can be constructed.[48,49] For example, a reaction network for methane combustion was built by Zeng et al.[50] with the ML-QCMs. The PES, reaction networks, chemical properties, and wave functions predicted by the ML-QCMs are associated with the behavior and toxicology of environmental chemicals, which are discussed as follows.

3. POTENTIAL APPLICATIONS IN ENVIRONMENTAL CHEMISTRY

3.1. Chemical Transformation. Chemical transformation of pollutants is a classical topic in environmental chemistry.[51] QCMs were adopted to calculate the PES, formation free energies, reaction rate coefficients, and product branching ratios for transformation of pollutants.[52−54] Nevertheless, it is very computationally expensive for accurate calculations of chemical reactions involving large molecules. The ML-QCMs may reduce the costs and expand applications of the QCMs in environmental chemical studies.[55]

3.1.1. Chemical Reactions Involving Radicals. Reactions with radicals (e.g., hydroxyl and halogen radicals) are important removal pathways of chemical pollutants and have been widely investigated with the conventional QCMs.[56,57] The ML-QCMs may provide an efficient way to accurately obtain energies and a clever way to construct PES under the framework of the type-I or type-II methods (Figure 2).

As shown in Figure 3A and B, the ML-QCMs can discover previously overlooked intermediates during degradation of pollutants. Stocker et al.[55] recently established an ML database consisting of thermodynamics data for 10 712 molecules and over 20 000 elementary reactions. On the basis of the database, a kernel ridge regression model was trained to predict reaction energies for the elementary reactions during CH_4 combustion processes following the type-I routine.[55] By analyzing the predicted reaction energies, Criegee intermediates overlooked previously were found to be involved in the CH_4 combustion processes.[55]

The more complex the simulated system, the longer the simulation time that is needed. The advantages in computing speed of the ML-QCMs can enable a large system with more specific species/conditions to be simulated.[58−60] Hence, some complex systems that can hardly be simulated by the conventional QCMs can also be simulated with the ML-QCMs.

3.1.2. New Particle Formation. Atmospheric new particle formations (NPF), including nucleation and subsequent growth, are significant sources for atmospheric particles and, in turn, affect global climate, local air quality, and public health.[61] Quantum chemical calculations can be employed to investigate the process of gas-phase precursors (e.g., H_2SO_4, iodine, and NH_3) to form small nanoclusters.[62−64]

However, there are two main obstacles for the calculation. First, it is still difficult to search for the global minimum configuration for a given cluster [e.g., $(H_2SO_4)_3(NH_3)_3$], although several techniques (e.g., the Artificial Bees Colony Algorithm in the ABCluster software[65] and some scripts for randomly generating configurations[62]) were proposed. Second, the accurate state-of-the-art QCMs [i.e., the CCSD(T) method[34]] can only be employed for very small clusters (typically within 30 atoms),[34] due to limits in computational capacities. This implies that the overall transition from gas-phase vapors, via small clusters, to large particles can hardly be captured using the conventional methods.[66]

The ML-QCMs may overcome the limitations. As can be seen from Figure 3C and D, ML can be employed to discover the global minimum of concerned clusters via mapping relationships between energies and different configurations for clusters with the same compositions. Once the relationships are determined, PES for the clusters can be constructed, and subsequently the global minimum can be obtained. In addition, the type-I models can be constructed by a fragment approach,[66] in which individual atomic energies in clusters are trained with experimental or high-level quantum chemical values. With the model, energies for larger clusters can be predicted by summarizing the energies of all the simulated atoms. Although extensive studies are still required, the ML-QCMs have promises in solving the puzzle of atmospheric NPF.

3.1.3. Heterogeneous Reaction of Pollutants. Heterogeneous transformations of chemical pollutants were conventionally investigated using molecular dynamics methods, such as *ab initio* molecular dynamics (AIMD) and hybrid quantum-mechanical/molecular-mechanical molecular dynamics simulations.[67−69] These methods were limited to reactions with low energy barriers (typically <2 kcal·mol^{-1}), as reactions with higher energy barriers require an extremely long time to reach transition states.[70]

The ML-QCMs can overcome the above difficulties via two different ways. First, trained ML models can be used for calculating forces and energies of a given system and, thus, can be integrated into molecular dynamics to accelerate the simulation. As the ML-based models are constructed for predicting interactions between atoms rather than molecules, the models can describe chemical reactions. This type of thinking was successfully adopted by Galib and Limmer[71] to investigate heterogeneous hydrolysis of N_2O_5 at air−water interfaces. They built a deep-learning model to predict forces and energies based on the structures of N_2O_5 and the air−water interfaces.[71] The AIMD simulation was significantly accelerated, as the new model was adopted to replace the DFT part (the most time-consuming part) required by the conventional AIMD.[71] They found that compared with transfer into the bulk of water, hydrolysis at the interface is faster.[71] A similar route can be adopted to explore other heterogeneous processes of various chemical pollutants.

Second, with the type-III method [e.g., targeted adversarial learning optimized sampling (TALOS)[15,16] methods], transition states and other rare events can be searched easily. Therefore, the time cost for obtaining the transition states can be reduced. For example, the TALOS method was successfully employed to explore the reaction between Cl^- and CH_3Cl in aqueous phases.[15] The TALOS method only took ca. 250 ps to observe the deserved reaction. However, no transition state was observed using some conventional enhanced sampling techniques (e.g., metadynamics and replica-exchange simulations).[15] As the main difference in the simulations between heterogeneous and aqueous phases lies in the modeling rather than the sampling aspect,[71] the type-III methods can also be employed to unveil the mysteries hidden in environmental interfaces.

3.2. Photochemical Transformation. Environmental photochemical transformations, especially direct photolysis and indirect photolysis with sensitizers, are important removal pathways of organic pollutants.[51] The conventional QCMs were adopted to calculate electronic adsorption spectra, excitation energies, photophysical processes (e.g., phosphorescence, fluorescence, and intersystem crossing), and photochemical transformations.[11] ML models entered into the field of electronically excited states relatively late, and it seems that this research field is developing at a slower pace, compared

https://doi.org/10.1021/acs.est.1c05970
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

with the exploding field of ML for characterizing ground states.[11,72]

Even so, several potential applications in environmental photochemistry can be foreseen. The PES of excited-state pollutants can be fitted smoothly with the ML methods. For example, Williams et al.[73] incorporated artificial neural networks into diabatization by ansatz and fitted the diabatic PES of excited-state $\cdot NO_3$. Similar methods can be used to investigate other chemical pollutants.

The time-cost of photodynamics simulations can also be reduced based on the ML fitted PES.[72] For example, with ML, a 10 ns photodynamics simulation for the *cis*−*trans* isomerization reaction of *trans*-hexafluoro-2-butene was performed in just 2 days, in contrast to ca. 58 years with the conventional methods.[74] The ML-based photodynamics simulation method can also be employed to investigate photochemical and photophysical processes of some small molecules such as NO_2 and phenol. However, the current computing power may not be sufficient for simulating macromolecules such as polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and organophosphorus flame retardants. Therefore, faster methods still need to be developed.

3.3. Sound Management on Chemicals. Pollution of synthetic chemicals is a serious and growing global problem.[75] Sound management of chemicals requires data on their physicochemical properties and environmental behavioral and toxicological parameters (e.g., octanol−water partition coefficients, degradation rate coefficients, carcinogenicity, and mutagenicity).[75] QSARs within the framework of environmental computational toxicology[2,76] can serve as a core tool for filling data gaps.

As shown in Figure 3E, SMILES, fingerprints, and other descriptors/features are extracted based on molecular structures as inputs for conventional QSAR models. In contrast, wave functions calculated via the type-II method and/or 3D structures characterized by the NN methods can also be used for constructing the QSAR models (Figure 3F). The ML-QCMs can be employed in constructing QSARs in two aspects:

(1) *Complete description of molecules.* As wave functions can be predicted by the type-II method, they can in turn be employed to predict properties of chemicals directly. Compared with conventional molecular descriptors that can only partially describe molecular characteristics, the wave functions contain all information for a certain electronic state (e.g., the ground state, the first excited state) molecule and can be better descriptors in QSARs. In other words, molecular structural information on chemicals cannot be lost using wave functions as inputs. Further studies can be performed in this aspect.

(2) *Direct mapping "structure" and "activity".* QSARs pursue "structure" and "activity" relationships but usually do not directly employ 3D structures of molecules as inputs. Alternatively, most QSARs map "molecular characteristics" (e.g., energies of the highest occupied molecular orbitals) and "activity". As aforementioned, the type-I and type-II methods can directly adopt molecular 3D structures as inputs via, for instance, graph NN representations (Figure 3F) and predict molecular properties or toxicities. Recently, Wang et al.[24] developed a new framework named SepPCNET to represent 3D molecular structures and adopted the framework to predict estrogen receptor activities of chemicals. The prediction accuracy of the model constructed under the SepPCNET framework was higher than in the case of using the conventional routines by 5−14%.[24] It is expected that further studies blend the ideas of the conventional QSARs and the ML-based quantum chemistry, which will be conducive to the development of environmental computational toxicology.

4. CHALLENGE

To date, the ML-QCMs have been rarely employed in environmental chemical studies. The following challenges should be solved to promote the use of the ML-QCMs:

(1) *Different philosophies.* Quantum chemists prefer to use simple model molecules to elucidate mechanisms or computational methods, whereas environmental chemists are committed to using the methods for probing behavior and toxicology of environmental chemicals that are always significantly larger than the chemical model molecules. Therefore, new ML-QCMs that are more suitable for environmental macromolecules should be developed.

(2) *Unfamiliarity.* The application of quantum chemistry in environmental chemical studies can be dated back to 1970s.[77] However, the rocketing development of the ML-QCMs emerged only in recent years. To popularize the ML-QCMs, some package models that hide the complex principles of ML and QCMs should be developed as initial tutorials for environmental chemists with limited backgrounds of quantum chemistry and/or ML.

(3) *GPU prices.* The speed of training ML models relies heavily on GPU performance. However, there are few cost-effective GPUs in circulation, and the price of GPUs is falsely higher than their ex-factory price by >200%, as a result of the GPUs being acquired by the mine owners of virtual currency (such as bitcoin).[78] Although some countries, such as China, have cracked down on bitcoin mining, the prices of cutting-edge GPUs are also expensive,[78] limiting popularization of the ML-QCMs.

(4) *Complex environmental conditions*. All models are wrong, but some are useful.[79] Model complexity should only be increased when necessary. Building simulation models for complex systems is an ever-lasting challenge for quantum chemists. Due to limitations in computational capacity, simulation of a big system as complex as the "real" world that contains all relevant elements is almost impossible. The ML-QCMs can reduce the computational costs and simulate a larger system that considers more essential elements. Nevertheless, further investigations are needed to clarify what key elements should be considered in the ML-QCM simulation and how to consider the additional factors with the ML-QCMs.

(5) *Methodological dilemmas.* (a) *Applicability domain characterization.* In the conventional QSAR models, applicability domains are characterized by various range-based, probability-density-based, and distance (including leverage)/similarity-based methods.[2,80,81] These methods have been proved to be useful for the models aiming at predicting typical environmentally relevant end points (e.g., physicochemical properties, environmental behavioral, and toxicological parameters) and using topological indices and fingerprints as descriptors.[2,81] However, it is unclear whether the methods can also be adopted for the models developed with the ML-QCMs. Further studies are needed to examine the effectiveness of the applicability domain characterization methods. (b) *Model evaluation.* Criticism on ML-based quantum chemical models often arises from the fact that assessment or validation

https://doi.org/10.1021/acs.est.1c05970
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

of the models can be very tricky.[18] Mean absolute errors and root mean squared errors were conventionally adopted to characterize the accuracy of ML-based quantum chemical models.[18] Some other indicators commonly used in QSARs, such as determination coefficients and leave-one-out cross-validated determination coefficients, can also be adopted.[2,80] Overfitting exists widely in some ML models.[12] There is still no universal method to avoid overfitting. It seems important to select representative and high-quality data to construct a training set, a validation set, and a test set, to avoid overfitting and ensure model quality. However, a perfect data set cannot be always guaranteed. Therefore, it is still a challenge to construct models with defective data sets. (c) *Model interpretation.* Many ML-based quantum chemical models are constructed via data-driven methods, leading to low interpretability for humans. To improve the interpretability, a useful way is to divide a NN model into several small blocks with clear physical/chemical meanings. Hermann et al.[42] combined several NNs with physical meanings (e.g., Jastrow factors and backflow functions) to predict correlation energies that characterize the Coulomb interactions between electrons and found that the backflow functions are important for predicting the energies. Other ways to improve model interpretability are also worthy of investigation.

■ AUTHOR INFORMATION

Corresponding Author

Jingwen Chen − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China;* orcid.org/0000-0002-5756-3336; Email: jwchen@dlut.edu.cn

Authors

Deming Xia − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Zhiqiang Fu − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Tong Xu − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Zhongyu Wang − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Wenjia Liu − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China*

Hong-bin Xie − *Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China;* orcid.org/0000-0002-9119-9785

Willie J. G. M. Peijnenburg − *Institute of Environmental Sciences (CML), Leiden University, Leiden 2300 RA, The Netherlands; Centre for Safety of Substances and Products, National Institute of Public Health and the Environment (RIVM), Bilthoven 3720 BA, The Netherlands;* orcid.org/0000-0003-2958-9149

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c05970

Notes

The authors declare no competing financial interest.

Biography

Jingwen Chen is a professor (since 2001) at Dalian University of Technology. After getting his Ph.D. from the Nanjing University in 1997, he performed postdoctoral studies at the German Research Center for Environmental Health as an Alexander von Humboldtian. His research interests cover environmental computational toxicology and ecological chemistry (environmental photochemistry). He has published over 320 peer-reviewed papers, reviews, and book chapters and three books (two monographs and one textbook). He served as a member of Teaching Steering Committee on Environmental Science and Engineering of China Ministry of Education, an Associate Editor for *ACS Sustain. Chem. Eng.*, and a member of editorial boards of several journals. In 2013, he won the National Science Fund for Distinguished Young Scholars and was appointed as chair professor of the "Cheung Kong Scholars Program" by the China Ministry of Education.

■ ACKNOWLEDGMENTS

This study was supported by the National Key R&D Program of China (2018YFE0110700), the National Natural Science Foundation of China (22136001), and the MIGRATION project that is sponsored within the Framework of the Strategic Research Programme of RIVM. D.X. acknowledges Haobo Wang from Dalian University of Technology for insightful discussions.

■ REFERENCES

(1) Wang, Z. Y.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. *Environ. Sci. Technol.* **2020**, *54* (5), 2575−2584.
(2) Wang, Z. Y.; Chen, J. W.; Hong, H. X. Developing QSAR Models with Defined Applicability Domains on PPARγ Binding Affinity Using Large Data Sets and Machine Learning Algorithms. *Environ. Sci. Technol.* **2021**, *55* (10), 6857−6866.
(3) Ciallella, H. L.; Zhu, H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chem. Res, Toxicol.* **2019**, *32* (4), 536−547.
(4) Csaszar, A. G.; Fabri, C.; Szidarovszky, T.; Matyus, E.; Furtenbacher, T.; Czako, G. The Fourth Age of Quantum Chemistry: Molecules in Motion. *Phys. Chem. Chem. Phys.* **2012**, *14*, 1085.
(5) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(6) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.
(7) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.
(8) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11* (6), 2336−2347.
(9) Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **2016**, *4* (5), 053208.
(10) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121* (16), 10218−10239.
(11) Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2021**, *121* (16), 9873−9926.
(12) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55* (19), 12741−12754.
(13) Dirac, P. A. M. Quantum Mechanics of Many-Electron Systems. *Proc. R. Soc. London A* **1929**, *123*, 714−733.
(14) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K. R. Machine Learning Force Fields. *Chem. Rev.* **2021**, *121* (16), 10142−10186.
(15) Zhang, J.; Yang, Y. I.; Noé, F. Targeted Adversarial Learning Optimized Sampling. *J. Phys. Chem. Lett.* **2019**, *10* (19), 5791−5797.
(16) Zhang, J.; Lei, Y.-K.; Yang, Y. I.; Gao, Y. Q. Deep learning for variational multiscale molecular modeling. *J. Chem. Phys.* **2020**, *153*, 174115.
(17) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2* (5), 359−366.
(18) Zhang, J.; Lei, Y.-K.; Zhang, Z.; Chang, J.; Li, M.; Han, X.; Yang, L.; Yang, Y. I.; Gao, Y. Q. A Perspective on Deep Learning for Molecular Modeling and Simulations. *J. Phys. Chem. A* **2020**, *124* (34), 6745−6763.
(19) Xu, T.; Chen, J. W.; Chen, X.; Xie, H. J.; Wang, Z. Y.; Xia, D. M.; Tang, W. H.; Xie, H.-B. Prediction Models on pK_a and Base-Catalyzed Hydrolysis Kinetics of Parabens: Experimental and Quantum Chemical Studies. *Environ. Sci. Technol.* **2021**, *55* (9), 6022−6031.
(20) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials From Deep Learning. *Nature* **2020**, *577*, 706−710.
(21) Liu, H.; Wei, M.; Yang, X.; Yin, C.; He, X. Development of TLSER model and QSAR model for predicting partition coefficients of hydrophobic organic chemicals between low density polyethylene film and water. *Sci. Total Environ.* **2017**, *574*, 1371−1378.
(22) Ye, S. Q.; Liang, J. C.; Liu, R. L.; Zhu, X. Symmetrical Graph Neural Network for Quantum Chemistry with Dual Real and Momenta Space. *J. Phys. Chem. A* **2020**, *124* (34), 6945−6953.
(23) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K. R. SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15* (1), 448−455.
(24) Wang, L. G.; Zhao, L.; Liu, X.; Fu, J.; Zhang, A. Q. SepPCNET: Deeping Learning on a 3D Surface Electrostatic Potential Point Cloud for Enhanced Toxicity Classification and Its Application to Suspected Environmental Estrogens. *Environ. Sci. Technol.* **2021**, *55* (14), 9958−9967.
(25) Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828−12840.
(26) Balabin, R. M.; Lomakina, E. I. Support vector machine regression (LS-SVM)-an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* **2011**, *13*, 11710−11718.
(27) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(28) Nguyen, T. T.; Szekely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Gotz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.
(29) Kamath, A.; Vargas-Hernandez, R. A.; Krems, R. V.; Carrington, T., Jr; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **2018**, *148*, 241702.
(30) Dral, P. O. Mlatom: a program package for quantum chemical research assisted by machine learning. *J. Comput. Chem.* **2019**, *40*, 2339−2347.
(31) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148* (24), 241718.
(32) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.
(33) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Proc. Mach. Learn. Res.* **2021**, *139*, 9377−9388.
(34) Bartlett, R. J.; Musiał, M. Coupled-Cluster Theory in Quantum Chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291−352.
(35) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.
(36) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(37) Liang, J.; Xu, Y.; Liu, R.; Zhu, X. QM-sym, a symmetrized quantum chemistry database of 135 kilo molecules. *Sci. Data* **2019**, *6*, 213.
(38) Nakata, M.; Shimazaki, T.; Hashimoto, M.; Maeda, T. PubChemQC PM6: Data Sets of 221 Million Molecules with Optimized Molecular Geometries and Electronic Properties. *J. Chem. Inf. Model.* **2020**, *60* (12), 5891−5899.
(39) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820−18826.
(40) Choo, K.; Mezzacapo, A.; Carleo, G. Fermionic neural-network states for ab-initio electronic structure. *Nat. Commun.* **2020**, *11*, 2368.

(41) Carleo, G.; Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **2017**, *355* (6325), 602−605.
(42) Hermann, J.; Schätzle, Z.; Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **2020**, *12* (10), 891−897.
(43) Graziano, G. Deep learning chemistry ab initio. *Nat. Rev. Chem.* **2020**, *4* (11), 564−564.
(44) Nagy, A.; Savona, V. Variational Quantum Monte Carlo Method with a Neural-Network Ansatz for Open Quantum Systems. *Phys. Rev. Lett.* **2019**, *122* (25), 250501.
(45) Hartmann, M. J.; Carleo, G. Neural-Network Approach to Dissipative Quantum Many-Body Dynamics. *Phys. Rev. Lett.* **2019**, *122* (25), 250502.
(46) Vicentini, F.; Biella, A.; Regnault, N.; Ciuti, C. Variational Neural-Network Ansatz for Steady States in Open Quantum Systems. *Phys. Rev. Lett.* **2019**, *122* (25), 250503.
(47) Yoshioka, N.; Hamazaki, R. Constructing neural stationary states for open quantum many-body systems. *Phys. Rev. B* **2019**, *99* (21), 214306.
(48) Bonati, L.; Piccini, G. M.; Parrinello, M. Deep learning the slow modes for rare events sampling. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (44), e2113533118.
(49) Schran, C.; Thiemann, F. L.; Rowe, P.; Muller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (38), e2110077118.
(50) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. H. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **2020**, *11* (1), 5713.
(51) Rene, P. S.; Gschwend, P. M.; Imboden, D. M. *Environmental Organic Chemistry*; John Wiley & Sons, Inc., 2002; pp 461−488.
(52) Ma, F. F.; Guo, X. R.; Xia, D. M.; Xie, H.-B.; Wang, Y. H.; Elm, J.; Chen, J. W.; Niu, J. F. Atmospheric Chemistry of Allylic Radicals from Isoprene: A Successive Cyclization-Driven Autoxidation Mechanism. *Environ. Sci. Technol.* **2021**, *55* (8), 4399−4409.
(53) Yu, Q.; Xie, H.-B.; Li, T. C.; Ma, F. F.; Fu, Z. H.; Wang, Z.; Li, C.; Fu, Z. Q.; Xia, D. M.; Chen, J. W. Atmospheric chemical reaction mechanism and kinetics of 1,2-bis(2,4,6-tribromophenoxy)ethane initiated by OH radical: a computational study. *Rsc Adv.* **2017**, *7* (16), 9484−9494.
(54) Li, C.; Chen, J. W.; Xie, H.-B.; Zhao, Y. H.; Xia, D. M.; Xu, T.; Li, X.; Qiao, X. L. Effects of Atmospheric Water on center dot OH-initiated Oxidation of Organophosphate Flame Retardants: A DFT Investigation on TCPP. *Environ. Sci. Technol.* **2017**, *51* (9), 5043−5051.
(55) Stocker, S.; Csanyi, G.; Reuter, K.; Margraf, J. T. Machine learning in chemical reaction space. *Nat. Commu.* **2020**, *11* (1), 5505.
(56) Ji, Y. M.; Zheng, J.; Qin, D.; Li, Y.; Gao, Y.; Yao, M.; Chen, X.; Li, G.; An, T. C.; Zhang, R. Y. OH-Initiated Oxidation of Acetylacetone: Implications for Ozone and Secondary Organic Aerosol Formation. *Environ. Sci. Technol.* **2018**, *52* (19), 11169−11177.
(57) Zhang, Q. Z.; Qu, X. H.; Wang, W. X. Mechanism of OH-Initiated Atmospheric Photooxidation of Dichlorvos: A Quantum Mechanical Study. *Environ. Sci. Technol.* **2007**, *41* (17), 6109−6116.
(58) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **2020**, *12*, 945−951.
(59) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.
(60) Schütt, K.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(61) Lee, S. H.; Gordon, H.; Yu, H.; Lehtipalo, K.; Haley, R.; Li, Y.; Zhang, R. Y. New Particle Formation in the Atmosphere: From Molecular Clusters to Global Climate. *J. Geophys. Res.: Atmos.* **2019**, *124*, 7098−7146.
(62) Xia, D. M.; Chen, J. W.; Yu, H.; Xie, H.-B.; Wang, Y.; Wang, Z. Y.; Xu, T.; Allen, D. T. Formation Mechanisms of Iodine-Ammonia Clusters in Polluted Coastal Areas Unveiled by Thermodynamics and Kinetic Simulations. *Environ. Sci. Technol.* **2020**, *54* (15), 9235−9242.
(63) Myllys, N.; Ponkkonen, T.; Passananti, M.; Elm, J.; Vehkamäki, H.; Olenius, T. Guanidine: A Highly Efficient Stabilizer in Atmospheric New-Particle Formation. *J. Phys. Chem. A* **2018**, *122* (20), 4717−4729.
(64) Myllys, N.; Olenius, T.; Kurtén, T.; Vehkamäki, H.; Riipinen, I.; Elm, J. Effect of Bisulfate, Ammonia, and Ammonium on the Clustering of Organic Acids and Sulfuric Acid. *J. Phys. Chem. A* **2017**, *121* (25), 4812−4824.
(65) Zhang, J.; Glezakou, V.-A.; Rousseau, R.; Nguyen, M.-T. NWPEsSe: An Adaptive-Learning Global Optimization Algorithm for Nanosized Cluster Systems. *J. Chem. Theory Comput.* **2020**, *16* (6), 3947−3958.
(66) Elm, J. Toward a Holistic Understanding of the Formation and Growth of Atmospheric Molecular Clusters: A Quantum Machine Learning Perspective. *J. Phys. Chem. A* **2021**, *125* (4), 895−902.
(67) Zhang, W. N.; Zhong, J.; Shi, Q. J.; Gao, L.; Ji, Y. M.; Li, G. Y.; An, T. C.; Francisco, J. S. Mechanism for Rapid Conversion of Amines to Ammonium Salts at the Air-Particle Interface. *J. Am. Chem. Soc.* **2021**, *143* (2), 1171−1178.
(68) Li, Y. W.; Shi, X. L.; Zhang, Q. Z.; Hu, J. T.; Chen, J. M.; Wang, W. X. Computational Evidence for the Detoxifying Mechanism of Epsilon Class Glutathione Transferase Toward the Insecticide DDT. *Environ. Sci. Technol.* **2014**, *48* (9), 5008−5016.
(69) Zhao, X. W.; Shi, X. L.; Ma, X. H.; Wang, J. J.; Xu, F.; Zhang, Q. Z.; Li, Y.; Teng, Z. C.; Han, Y.; Wang, Q.; Wang, W. X. Simulation Verification of Barrierless HONO Formation from the Oxidation Reaction System of NO, Cl, and Water in the Atmosphere. *Environ. Sci. Technol.* **2021**, *55* (12), 7850−7857.
(70) Xia, D.; Zhang, X.; Chen, J.; Tong, S.; Xie, H.-b.; Wang, Z.; Xu, T.; Ge, M.; Allen, D. T. Heterogeneous Formation of HONO Catalyzed by CO_2. *Environ. Sci. Technol.* **2021**, *55* (18), 12215−12222.
(71) Galib, M.; Limmer, D. T. Reactive uptake of N_2O_5 by atmospheric aerosol is dominated by interfacial processes. *Science* **2021**, *371* (6532), 921−925.
(72) Dral, P. O.; Barbatti, M. Molecular excited states through a machine learning lens. *Nat. Rev. Chem.* **2021**, *5* (6), 388−405.
(73) Williams, D. M. G.; Eisfeld, W. Neural network diabatization: A new ansatz for accurate high-dimensional coupled potential energy surfaces. *J. Chem. Phys.* **2018**, *149* (20), 204106.
(74) Li, J.; Reiser, P.; Boswell, B. R.; Eberhard, A.; Burns, N. Z.; Friederich, P.; Lopez, S. A. Automatic discovery of photoisomerization mechanisms with nanosecond machine learning photodynamics simulations. *Chem. Sci.* **2021**, *12* (14), 5302−5314.
(75) *Global Chemicals Outlook II*; United National Environment Programme, 2019.
(76) Rusyn, I.; Daston, G. P. Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. *Environ. Health Perspect.* **2010**, *118* (8), 1047−1050.
(77) Molina, M. J.; Rowland, F. S. Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone. *Nature* **1974**, *249* (5460), 810−812.
(78) Crypto-miners are probably to blame for the graphics-chip shortage. *The Economist* **2021** (accessed Aug 21, 2021).
(79) Wasserstein, R. George Box: a model statistician. *Significance* **2010**, *7* (3), 134−135.
(80) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W. D.; Veith, G.; Yang, C. H. Current status of methods for defining the applicability domain of (quantitative)

structure-activity relationships - The report and recommendations of ECVAM Workshop 52. *Altern. Lab Anim.* **2005**, *33* (2), 155−173.

(81) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Altern. Lab Anim.* **2005**, *33* (5), 445−459.

https://doi.org/10.1021/acs.est.1c05970
Environ. Sci. Technol. XXXX, XXX, XXX−XXX