Improved Environmental Chemistry Property Prediction of Molecules with Graph Machine Learning

Shang Zhu,[†] Bichlien H. Nguyen,[‡] Yingce Xia,[‡] Kali Frost,[‡] Shufang Xie,[‡] Venkatasubramanian Viswanathan,[†] and Jake Smith[*,‡]

[†]*Department of Mechanical Engineering,*
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA
[‡]*Microsoft Research, AI4Science.*

E-mail: jakesmith@microsoft.com

Abstract

Rapid prediction of environmental chemistry properties is critical towards the green and sustainable development of chemical industry and drug discovery. Machine learning methods can be applied to learn the relations between chemical structures and their environmental impact. Graph machine learning, by learning the representations directly from molecular graphs, may enable better predictive power than conventional feature-based models. In this work, we leveraged graph neural networks to predict environmental chemistry properties of molecules. To systematically evaluate the model performance, we selected a representative list of datasets, ranging from solubility to reactivity, and compare directly to commonly used methods. We found that the graph model achieved near state-of-the-art accuracy for all tasks and, for several, improved the accuracy by a large margin over conventional models that rely on human-designed chemical features. This demonstrates that graph machine learning can be a powerful tool to do representation learning for environmental chemistry. Further, we compared the data efficiency of conventional feature-based models and graph neural networks, providing guidance for model selection dependent on the size of datasets and feature requirements.

INTRODUCTION

A recent focus of the chemical industry is the reduction of its environmental footprint.[1] Proposed routes to this goal include the adoption of green chemistry frameworks that minimize the impact of chemical synthesis and manufacturing at scale and alter process designs to use chemicals with minimal carbon intensity and toxicological risk.[2] Successful application of such a framework requires rapid and accurate assessments of the environmentally relevant properties of prospective chemical components — a task to which machine learning (ML) techniques are particularly well-suited.[3–8]

Machine learning algorithms have proven a useful augmentation to traditional data analytics techniques in the evaluation of environmental impacts of chemical processes. For example, Zhang and Zhang employed a deep-neural-network regression for the prediction of aqueous solubilities of persistent, bioaccumulative, and toxic chemicals.[9] Dawson et al. approximated the intrinsic metabolic clearance rate and plasma bound fraction of toxic chemicals using random forest regression for their application in the toxicokinetic modeling.[10] Zhong et al. trained an ensemble regression model for the prediction of reactivity of organic contaminants toward a variety of oxidants.[11] Other successful applications include the identification of endocrine-disrupting chemicals[12] and direct modeling of environmental impacts from chemical production.[13] These and other use cases demonstrate the broad applicability of machine learning techniques to problems in environmental engineering.[14]

Common across the existing literature is the use of chemical features to produce a flattened, vector representation of the complex geometry of an organic molecule. We denote ML models that take this approach as "feature-based" models, as they rely on explicit featurization of the molecular structure to construct an input representation. Chemical features have a long history of use in cheminformatics applications and may be broadly classified into two families: molecular descriptors and fingerprints.[15] Molecular descriptors may be understood to abstract molecular structural information into summary statistics, such as molecular weight, polarizability, or numbers of heteroatoms. They have the advantage of being relatively intuitively understood; however, they fail to fully capture the information contained in the molecular structure, and the selection of appropriate molecular descriptors for a given prediction task is often nontrivial. Common examples of descriptor-based features include PaDEL descriptors,[16] Mordred descriptors,[17] and MACCS descriptors.[18] The second class of chemical features, molecular fingerprints, explicitly encode the presence and local environment of functional groups into a feature vector. An example is extended-connectivity fingerprints (ECFP).[19] The use of molecular fingerprints provides a more direct representation of the molecular structure and simplifies feature selection at the cost of some interpretability relative to molecular features.

With recent advances in graph machine learning, direct graph representation of molecular structures, where nodes represent atoms and edges represent chemical bonds, has become a viable alternative to chemical descriptors.[15] Following this approach, Duvenaud et al. created data-driven features, NeuralFPs, by applying convolution operations directly to molecular graphs and showed the resulting representation to be better performing than ECFP features.[20] Subsequent work has solidified these results, with graph neural networks achieving state-of-the-art accuracy for a variety of molecular machine learning tasks.[21,22] Recently, ring-enhanced graph neural network (`O-GNN`[23]) has been reported as an advancement to existing graph-machine-learning methods, by explicitly encoding information on rings into graph neural networks. It shows state-of-the-art accuracy on molecular property prediction benchmarks.

In this work, we systematically evaluated the predictive power of graph machine learning methods and compared them with feature-based models that rely on chemical features. A list of molecular property prediction tasks were selected in the environmental chemistry domain. Results from four sets of models have been reported: ECFP-based models, NeuralFP-based graph models, `NeuralFP`, feature-based models built on other types of chemical features, as well as one of the state-of-the-art graph models, `O-GNN`. We found that the state-of-the-art graph machine learning models outperformed or were at least on par with the feature-based methods in all tasks. To support these results, we conducted a data-efficiency analysis to provide guidance on when graph models are advantaged over feature-based approaches and examine the correlation of residual errors across both methods. We found the graph machine learning architecture an exemplary tool for molecular property prediction tasks on datasets exceeding 1000 observations and competitive with conventional feature-based models down to several hundred observations. The state-of-the-art graph machine learning methods provide a rapid and accurate approach for environmental chemistry property prediction.

MATERIALS AND METHODS

Data Collection

We identified a series of molecular property prediction tasks with associated datasets reported in the recent literature, ranging from solubility to metabolic susceptibility to reactivity, on which to assess the performance of `O-GNN` relative to the literature-reported model. We provide an overview of the selected datasets, baseline accuracy and our new results in Table 1 of the results section. In the first task, $ESOL$, the model was asked to predict the aqueous solubility of a series of small molecules. The $ESOL$ dataset is composed of 1144 structures paired with experimentally measured aqueous solubilities reported in logarithm-transformed units of mol/L.[24] In the reporting publication,[24] the $ESOL$ dataset was fit using molecular descriptors and linear regression, which identified a high dependence of aqueous solubilities on both the calculated octanol-water partition coefficient $logP_{octanol}$ and the proportion of heavy atoms in aromatic systems. Recently, Zhang and Zhang demonstrated improved accu-

Table 1: Selected Datasets for Environmental Chemistry[a]

Task	Property	Size	Baseline Accuracy[a]	Baseline Model
ESOL	Small Molecule Solubility in Water	1144	0.62 (0.04)	PaDEL-DNN
BCF	Bioconcentration Factor	1056	0.67 (0.04)	PaDEL-DNN
Clint	Intrinsic Metabolic Clearance Rate	4422	0.86 (0.05)	Descriptor-based Features + Random Forest Regression[b]
O_3-react	Chemical Reactivity with O_3 Oxidants	759	2.06	Fingerprint-based Features + Ridge-Regression
SO_4-react	Chemical Reactivity with $SO_4^{\bullet -}$ Oxidants	568	0.64	Descriptor-based Features + Random Forest Regression

[a] Baseline accuracy is reported in root-mean-square-error of the testing dataset ($RMSE_{test}$), where the numbers outside and inside the parenthesis are the mean and standard deviation values obtained from cross-validation. The splits in $\mathrm{O_3} - react$ and $\mathrm{SO_4} - react$ are given in the literature,[11] so no cross-validation is conducted. The units of ESOL and Clint are $ln(mol/L)$ and $ln(\mu L/min/10^6)$, while others are non-dimensional properties.
[b] This baseline result is created by this work.

racy on this task using molecular descriptors, PaDEL features, with a deep neural network (PaDEL-DNN), and we included their achieved $RMSE_{test}$ of 0.62 as the baseline in Table 1.[9] The second task, BCF, required to predict a bioconcentration factor for the accumulation of a series of small molecules in fish. The BCF dataset covers 1056 molecules, and includes both molecular structures and bioconcentration factors reported as a the logrithm-transformed ratio between concentration in the organism and in the containing water at steady-state.[25] Zhang and Zhang also applied the PaDEL-DNN method to this task, achieving an $RMSE_{test}$ of 0.67. In the third task, $Clint$, the model was developed to predict the rate of intrinsic metabolic clearance (Cl_{int}) of a series of small molecules, an important parameter for toxicokinetic modeling.[10] Dawson et al. assembled experimental measurements of Cl_{int} by hepatic cells and microsome assays from the ChEMBL and ToxCast databases, which were standardized into the unit of $\mu L/min/10^6$ cells. While they utilized this dataset to train a classifier, we framed a regression problem for consistency with the remainder of the tasks and trained a random forest model with Mordred descriptors, in order to predict the logrithm-transformed Cl_{int} to serve as the baseline model in Table 2.[17] The last two tasks, O_3-react and SO_4-react, asked the model to predict the reactivity of organic contaminants to two oxidants, O_3 (ozone), and $SO_4^{\bullet -}$.[11] To construct the associated datasets, Zhong et al. collected reactivity data from the literature, curating a total of 759 and 557 data points in O_3-react and SO_4-react, respectively.[11] The logarithm-transformed reaction rate constants $log(k)$ were reported alongside reaction conditions.[11] $ECFP$ fingerprints and molecular descriptors were benchmarked in combination with multiple machine learning algorithms, with the best performing models ultimately obtaining an $RMSE_{test}$ of 2.06 on the O_3-react task and an $RMSE_{test}$ of 0.64 on the SO_4-react task.

Graph-based Machine Learning

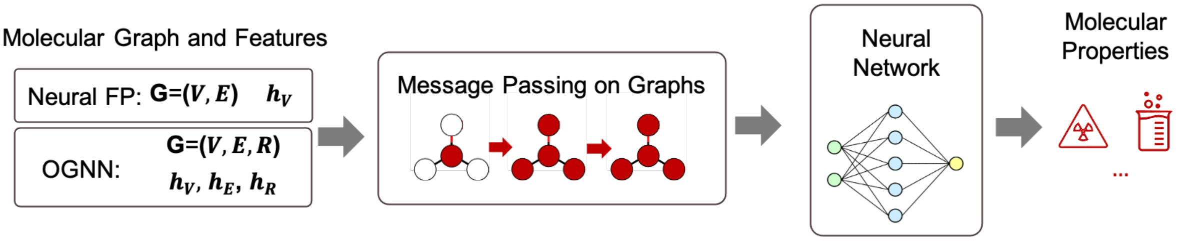
Molecular Graph and Features
Neural FP: G=(V, E) h_V
OGNN: G=(V, E, R) h_V, h_E, h_R
Message Passing on Graphs
Neural Network
Molecular Properties
...

Figure 1: Model Architecture for Graph Neural Networks. The model starts with molecular graph $\mathbf{G}$ and features and then applies graph convolutions iteratively on those features to get graph-level features. The graph-level features are transformed by feed-forward neural networks to predict the environmental properties. For the NeuralFP-based graph model, the graph only has atom features h_V, $\mathbf{G}$=($\mathbf{V}$, $\mathbf{E}$). For `O-GNN`, the graph covers atom, bond, and ring-level features (h_V, h_E, h_R) and $\mathbf{G}$=($\mathbf{V}$, $\mathbf{E}$, $\mathbf{R}$)

We briefly introduce the graph neural networks leveraged in this work, NeuralFP-based graph model, `NeuralFP`, and `O-GNN`. `NeuralFP` generalizes ECFP features by applying convolution operations directly on graphs, while `O-GNN` further adds the explicit encoding of ring structures, along with features of bonds and rings in the graph convolution steps. A more in-depth theoretical analysis of graph machine learning approach may be found in literature.[20,23] We summarize the architectures of `NeuralFP` and `O-GNN` in Figure 1. Mathematically, for graph machine learning methods, we define the molecular graph G as $G = (\mathbf{V}, \mathbf{E}, \mathbf{R})$, where

$\mathbf{V}$, $\mathbf{E}$ and $\mathbf{R}$ are the atom, bond and ring set, respectively. Atom, bond, and ring features are specified by $\mathbf{h_V}$ (atom type, chirality, degree number, etc.), $\mathbf{h_E}$ (bond type, stereochemistry, conjugated type) and $\mathbf{h_R}$ (a concatenation of atom and bond features that are involved in the rings). For `NeuralFP`, only $G = (\mathbf{V}, \mathbf{E})$ and atom features $\mathbf{h_V}$ are utilized in the iterative message passing (graph convolution) step, where involved features are updated by message-passing layers that merge information from the neighborhood of a central node. After pooling of message-passed node features, we obtain a graph-level molecular feature. The molecular properties are then obtained by transformation with a feed-forward neural network on the graph-level features. Unlike `NeuralFP`, `O-GNN` further encodes the edge and ring features, $(\mathbf{R}, \mathbf{h_E}, \mathbf{h_R})$, inside the neural network. For studies on the graph-based models, `NeuralFP` model was implemented in `DeepChem`,[26] and `O-GNN` was implemented in `PyTorch` as previously reported.[23] A consistent 5-fold cross-validation split was defined for each task.

As only summary statistics were available in the literature reports of the baseline models, we trained a feature-based model to serve as a surrogate for the direct comparison of predictions. In each case, we reported two sets of results for feature-based models. First, to compare with `NeuralFP`, we paired ECFP features with various machine learning algorithms (random forests, gradient boosting, support vector machines, neural networks) and reported the lowest $RMSE_{test}$. Further, we obtained an optimized feature-based model with a combinatorial search of molecular features (ECFP, Mordred, MACCS) and machine learning algorithms, where the best performing model was measured by $RMSE_{test}$ to represent the feature-based methods. A consistent 5-fold cross-validation split was defined for each task. Additional details on feature generation and model selection may be found in the Supporting Information (SI).

RESULTS AND DISCUSSION

In Table 2, we report observed performances of the two feature-based models and two graph models. Consistent with the previous publication,[20] `NeuralFP` yielded better predicted values than the ECFP-based model for most tasks. However, the feature-based model using Mordred descriptors significantly outperformed both the ECFP-based model and `NeuralFP` graph-based model. For example, with Mordred descriptors, an $RMSE_{test}$ of 0.61 was observed for the $ESOL$ task, 48.7% and 24.7% lower than the ECFP-based model and `NeuralFP`, respectively.

Table 2: Overview of Collected Datasets, Model Performances of Graph Models versus Feature-based Models[a]

	ESOL	BCF	Clint	O_3-react	SO_4-react
Property	Solubility	Bioconcentration	Intrinsic Clearance	Reactivity	Reactivity
Size	1128	1034	4422	759	557
ECFP	1.19 (0.06)	0.85 (0.05)	0.91 (0.09)	2.26	0.74
`NeuralFP`	0.81 (0.01)	0.79 (0.05)	0.71 (0.04)	2.12	0.90
Best Feature-based	0.61 (0.04)	0.67 (0.05)	0.86 (0.05)	**2.05**	**0.60**
`O-GNN`	**0.36 (0.03)**	**0.40 (0.08)**	**0.34 (0.03)**	2.07	0.66

[a] Performance reported in the format of $RMSE_{test}$ after 5-fold cross-validation, except that the two reactivity datasets were trained with the splits following literature.[11] The most accurate model is highlighted by bolding.

To further explore the potential of graph machine learning for these tasks, we leverage the representation power of ring-enhanced graph neural networks, `O-GNN`. With `O-GNN`, we observed a substantial improvement in prediction accuracy on tasks $ESOL$, BCF and $Clint$, relative to the best-performing feature-based models. This improvement may be attributed to the increased capacity of the `O-GNN` architecture to capture information related to the molecular structures relative to the molecular descriptors or fingerprints employed in the baseline models.[15] On the O_3-react and SO_4-react tasks, the performance of `O-GNN` was found to be comparable to best-performing feature-based models, without the substantial gains in $RMSE_{test}$ observed on the other tasks. One plausible explanation is that, the datasets for tasks O_3-react and SO_4-react contained fewer observations than those for the other tasks. We hypothesized that the O-GNN architecture may require model training on a larger dataset to achieve optimal predictive performance than the feature-based model architectures.

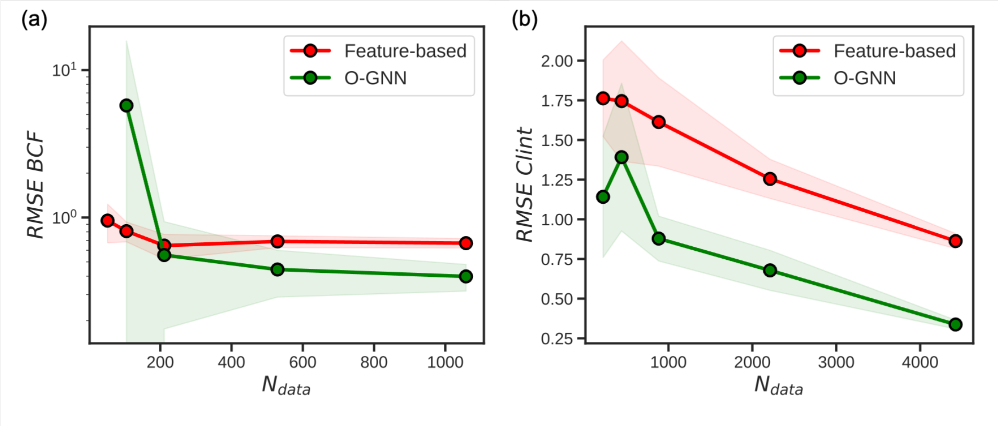

Figure 2: Comparing Learning Curves of Feature-based Models and O-GNN for (1) the *BCF* Task and (b) the *Clint* Task. The X-axis is the number of input data points for training, while the Y-axis is $RMSE_{test}$, reported by its mean (the line) and standard deviations (the colored area around the line) after cross-validation. The red curve is from feature-based models and the green curve is from O-GNN results.

Here and going forward, we will compare the best-performing graph machine learning methods, i.e. O-GNN, and the best-performing feature-based methods. We denoted them as O-GNN and 'feature-based' models, respectively, since they are the most desirable options for the two categories of experimented molecular machine learning methods.To test our hypothesis on the data size, we conducted a data-efficiency experiment, in which a series of models were trained on randomly sampled subsets of the *BCF* and *Clint* datasets utilizing both O-GNN and a feature-based architecture. The performance of each model was evaluated against varying training set size, by 5-fold cross-validation, to give learning curves (Figure 2).[27]These learning curves are data-efficiency experiments that could provide insight into the relative performance of O-GNN on the data-limited O_3-react and SO_4-react tasks. Although the O-GNN models are substantially advantaged over the feature-based models when trained on the full-sized *BCF* and *Clint* datasets, the loss reduction is less substantial as we decreased the training data size, as shown in Figure 2. In both cases, the performance of the O-GNN model becomes comparable to that of the feature-based model as the training dataset drops below approximately 1000 observations, in line with the size of the O_3-react (759) and SO_4-react (557) datasets. At the extreme, the feature-based model outperforms the O-GNN model on the *BCF* task when the training dataset drops below approximately 100 observations. This behavior may be attributed to the contributions of chemistry knowledge introduced with the use of human-designed molecular features, and suggests that a feature-based model may become a more appropriate choice on data-limited tasks.

Having established a high-level understanding of which molecular property prediction tasks O-GNN models might be expected to outperform feature-based models, we next sought to identify potential systematic trends in the models' predictions that might explain the improved performance of the O-GNN model on the *ESOL*, *BCF*, and *Clint* tasks. To this end, we drew parity plots covering model predictions on the test dataset for the *Clint* task (Figure 3a). The predicted values from each model exhibit the expected linear correlation to the true values without notable systematic deviations. This result suggests that the superior performance of the O-GNN model is attributable to a general improvement in molecular representation, as opposed to an ability to capture novel molecular features. Further corroborating, a linear trend was observed between the residual errors of the two models (Pearson correlation coefficient, $r = 0.40$), indicating that the two models generally overestimate or underestimate the $\mathrm{Cl}_{\mathrm{int}}$ of the same molecules (Figure 3b).

Finally, for *Clint* task, we directly compared the learned molecular representations of the O-GNN model to the molecular features (Modred) utilized in our surrogate feature-based model, considering the ability of each to distinguish molecules by $\mathrm{Cl}_{\mathrm{int}}$. Principle component analysis (PCA) was used to map O-GNN-derived or Mordred feature vector representations of each molecule in the *Clint* test dataset into a 2-dimensional chemical space and the results are plotted in Figure 3c-d. We scaled each dimension of Mordred feature vector to zero mean and unit variance since the chemistry information it encodes may intrinsically follow distinct distributions. Graph neural networks like O-GNN transform the discrete atom, bond, and ring

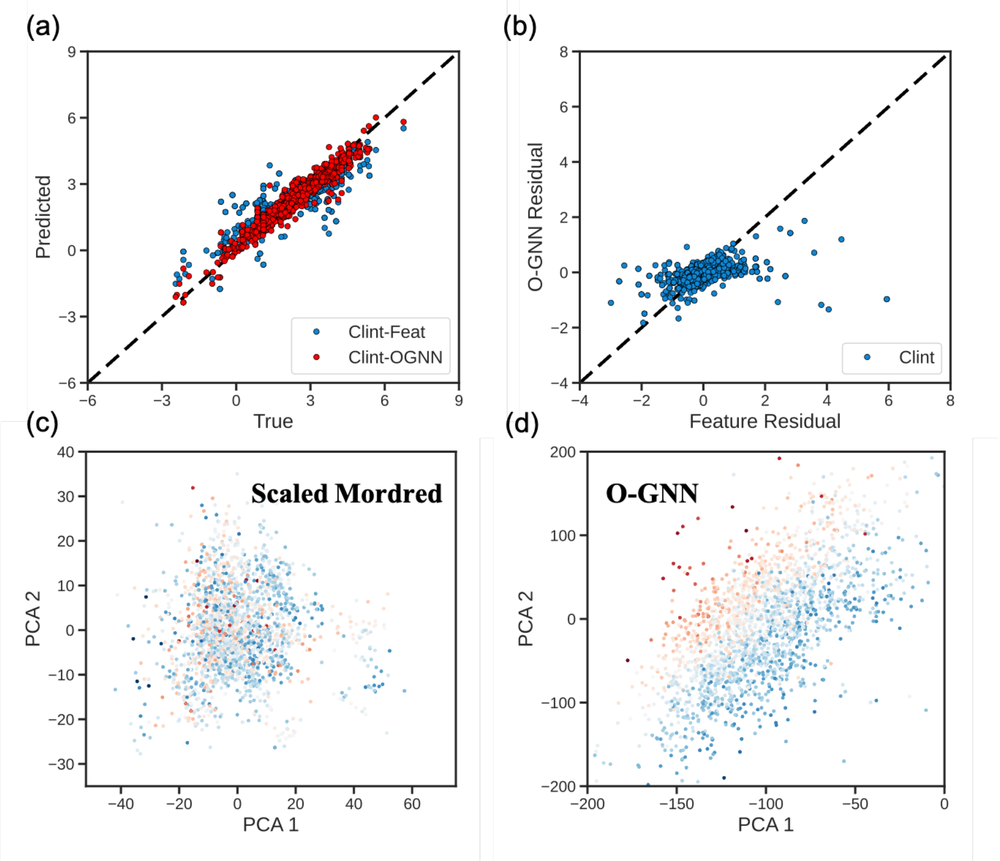

Figure 3: Detailed Analysis of the *Clint* task (a) Parity Plot. The black line represents complete agreement of the predicted and true values. (b) Prediction Residual Plot (predicted values minus true values). X-axis is the residual values of feature-based models while Y-axis is for `O-GNN`. (c-d) PCA Plots for (c) Scaled Mordred Features and (d) `O-GNN`-extracted Features. A window with PCA1 and PCA2 in [-200, 0] and [-200, 200] is shown for visualization purposes. Each dot is color-coded by their clearance values.

features that make up a molecule into a continuous latent representation. In Figure 3c-d, we observed that the first two PCA features are sufficient to cleanly arrange the `O-GNN`-encoded molecules by $\mathrm{Cl}_{\mathrm{int}}$ while the Mordred-encoded molecules remain poorly distinguished.

Conclusions

In this work, we investigated the predictive power of graph machine learning and feature-based models in order to estimate environmental properties of chemicals. We first observed that although `NeuralFP` may outperform ECFP-based models, the best feature-based model may be more desirable when appropriate chemical features are selected, e.g. Mordred for solubility-related prediction tasks. We therefore recommended the best-performing feature-based models as a new baseline. Compared with baseline feature-based approaches, O-GNN achieved state-of-the-art predictive accuracy on all tested tasks of solubility, bioconcentration, metabolism, and contaminant reactivity. By analyzing the data efficiency of the baseline and graph neural networks, we can conclude that O-GNN outperforms the baseline significantly when an enough amount of data is provided, while conventional approaches reduce the prediction error in the low-data regime. Lastly, we thoroughly evaluated the model predictions from the two approaches based on parity plots, residual analysis and the PCA plots of Mordred descriptors and O-GNN-extracted features. O-GNN demonstrated a higher predictive power by distinguishing the environmental properties, e.g. $\mathrm{Cl_{int}}$, by molecule structures. We envision future works can be conducted as follows. In the low data regime, emerging ML methods may offer additional improvement, including multitask learning,[11] transfer learning,[28,29] one-shot learning,[30] and self-supervised learning.[31,32] Where more data is available, modern graph machine learning models outperform the more commonly used ECFP fingerprint and feature-based models and should be the method of choice where prediction accuracy is prioritized.

Conflicts of interest

S.Z., B.H.N, Y.X., K.F., S.X., and J.A.S. were employed by Microsoft for portions of the work.

Acknowledgement

We thank Tian Xie (MSR) and Karin Strauss (MSR) for helpful discussions. S. Z. and V. V. acknowledge support from the Extreme Science and Engineering Discovery Environment (XSEDE) for providing computational resources through Award No. TG-CTS180061.

References

(1) National Academies of Sciences, E.; Medicine, *The Importance of Chemical Research to the U.S. Economy*; The National Academies Press: Washington, DC, 2022.

(2) Ganesh, K. N.; Zhang, D.; Miller, S. J.; Rossen, K.; Chirik, P. J.; Kozlowski, M. C.; Zimmerman, J. B.; Brooks, B. W.; Savage, P. E.; Allen, D. T.; Voutchkova-Kostal, A. M. Green Chemistry: A Framework for a Sustainable Future. *Environmental Science & Technology* **2021**, *55*, 8459–8463.

(3) Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chem.* **2009**, *11*, 1826–1831.

(4) Wang, Z.; Su, Y.; Jin, S.; Shen, W.; Ren, J.; Zhang, X.; Clark, J. H. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chem.* **2020**, *22*, 3867–3876.

(5) Mohan, M.; Demerdash, O.; Simmons, B. A.; Smith, J. C.; K. Kidder, M.; Singh, S. Accurate prediction of carbon dioxide capture by deep eutectic solvents using quantum chemistry and a neural network. *Green Chem.* **2023**, *25*, 3475–3492.

(6) Kumar, S.; Ignacz, G.; Szekely, G. Synthesis of covalent organic frameworks using sustainable solvents and machine learning. *Green Chem.* **2021**, *23*, 8932–8939.

(7) Coşgun, A.; Günay, M. E.; Yıldırım, R. Machine learning for algal biofuels: a critical review and perspective for the future. *Green Chem.* **2023**, *25*, 3354–3373.

(8) Kondo, M.; Sugizaki, A.; Khalid, M. I.; Wathsala, H. D. P.; Ishikawa, K.; Hara, S.; Takaai, T.; Washio, T.; Takizawa, S.; Sasai, H. Energy-{,} time-{,} and labor-saving synthesis of α-ketiminophosphonates: machine-learning-assisted simultaneous multiparameter screening for electrochemical oxidation. *Green Chem.* **2021**, *23*, 5825–5831.

(9) Zhang, K.; Zhang, H. Predicting Solute Descriptors for Organic Chemicals by a Deep Neural Network (DNN) Using Basic Chemical Structures and a Surrogate Metric. *Environmental Science & Technology* **2022**, *56*, 2054–2064.

(10) Dawson, D. E.; Ingle, B. L.; Phillips, K. A.; Nichols, J. W.; Wambaugh, J. F.; Tornero-Velez, R. Designing QSARs for Parameters of High-Throughput Toxicokinetic Models Using Open-Source Descriptors. *Environmental Science & Technology* **2021**, *55*, 6505–6517.

(11) Zhong, S.; Zhang, Y.; Zhang, H. Machine Learning-Assisted QSAR Models on Contaminant Reactivity Toward Four Oxidants: Combining Small Data Sets and Knowledge Transfer. *Environmental Science & Technology* **2022**, *56*, 681–692.

(12) Tan, H.; Wang, X.; Hong, H.; Benfenati, E.; Giesy, J. P.; Gini, G. C.; Kusko, R.; Zhang, X.; Yu, H.; Shi, W. Structures of Endocrine-Disrupting Chemicals Determine Binding to and Activation of the Estrogen Receptor α and Androgen Receptor. *Environmental Science & Technology* **2020**, *54*, 11424–11433.

(13) Wernet, G.; Papadokonstantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chem.* **2009**, *11*, 1826–1831.

(14) Zhong, S. et al. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science & Technology* **2021**, *55*, 12741–12754.

(15) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* **2020**, *12*, 56.

(16) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32*, 1466–1474.

(17) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018**, *10*, 4.

(18) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.

(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, PMID: 20426451.

(20) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Advances in Neural Information Processing Systems. 2015.

(21) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. Proceedings of the 34th International Conference on Machine Learning. 2017; pp 1263–1272.

(22) Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.

(23) Zhu, J.; Wu, K.; Wang, B.; Xia, Y.; Xie, S.; Meng, Q.; Wu, L.; Qin, T.; Zhou, W.; Li, H.; Liu, T.-Y. $\mathcal{O}$-GNN: incorporating ring priors into molecular modeling. The Eleventh International Conference on Learning Representations. 2023.

(24) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1000–1005.

(25) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions? *Chemosphere* **2015**, *127*, 171–179.

(26) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

(27) Viering, T.; Loog, M. The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, 1–20.

(28) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *Journal of Medicinal Chemistry* **2020**, *63*, 8683–8694, PMID: 32672961.

(29) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Central Science* **2019**, *5*, 1717–1730, PMID: 31660440.

(30) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* **2017**, *3*, 283–293, PMID: 28470045.

(31) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; WEI, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. Advances in Neural Information Processing Systems. 2020; pp 12559–12571.

(32) Zhu, J.; Xia, Y.; Qin, T.; Zhou, W.; Li, H.; Liu, T.-Y. Dual-view Molecule Pre-training. 2021; https://arxiv.org/abs/2106.10234.