OPEN ACCESS

Geochemistry REPORT

The Innovation Geoscience

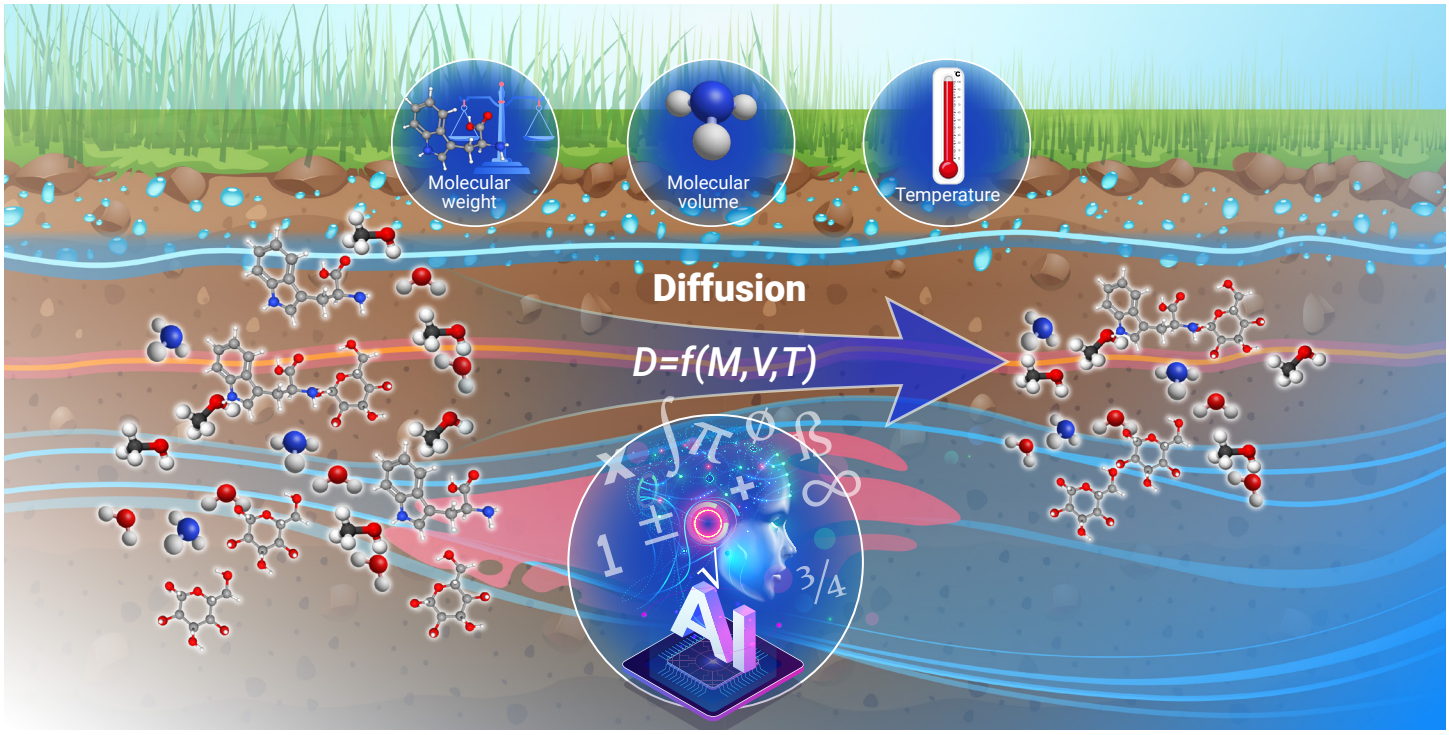
Discovery of new formula accurately determining organic compounds aqueous diffusion coefficients by symbolic regression

Min Han[1,2,3] **and Biao Jin**[1,2,3,*]

*Correspondence: jinbiao@gig.ac.cn

Received: March 20, 2025; Accepted: March 4, 2026; Published Online: March 11, 2026; https://doi.org/10.59717/j.xinn-geo.2026.100208

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- A novel formula for determining organic compounds aqueous diffusion coefficients was discovered.
- The proposed formula achieved a higher accuracy compared to the currently available empirical equations.
- The insights from this study are shedding light on understanding of organic compounds diffusion in water.

REPORT Geochemistry

www.the-innovation.org/geoscience

Discovery of new formula accurately determining organic compounds aqueous diffusion coefficients by symbolic regression

Min Han[1,2,3] **and Biao Jin**[1,2,3,*]

[1]State Key Laboratory of Advanced Environmental Technology, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China

[2]University of Chinese Academy of Sciences, Beijing 10069, China

[3]Guangdong Provincial Key Laboratory of Environmental Protection and Resources Utilization, Guangzhou 510640, China

*Correspondence: jinbiao@gig.ac.cn

Received: March 20, 2025; Accepted: March 4, 2026; Published Online: March 11, 2026; https://doi.org/10.59717/j.xinn-geo.2026.100208

Citation: Han M. and Jin B. (2026). Discovery of new formula accurately determining organic compounds aqueous diffusion coefficients by symbolic regression. *The Innovation Geoscience* **4**:100208.

Aqueous diffusion coefficients of organic compounds are important parameters for many studies in geosciences and relevant applications, in particular, for modelling solute transport in groundwater. Accurately determining organic compounds aqueous diffusion coefficients still remains challenging due to lack of high-quality experimental data as well as reliable predictive tools. Based on symbolic regression incorporating theoretical constraints derived from the Stokes-Einstein hydrodynamics and Wilke-Chang theory, our study discovered a symbolic formula, which is mathematically linking the diffusion coefficients with water temperature, molecular weight, molecular volume and viscosity of water. The proposed formula achieved a preferential accuracy for predicting different organic compounds diffusion coefficients in water comparing with the currently available empirical equations. The complex interplay of molecular weight, molecular volume and water temperatures on aqueous diffusion coefficients of organic compounds is substantiated by causal inference and controlled test. The insights from this study are shedding light on better understanding of organic compounds diffusion in water.

INTRODUCTION

Diffusion of organic compounds in water is one of the most important processes in natural aquatic environments.[1-3] This process becomes the dominant transport mechanism in low-permeability geologic media—such as silts, clays, and fractured rock—where groundwater advection is negligible.[3] Determination of organic compounds' aqueous diffusion coefficients has become the subject of many studies in geoscience.[4-11] For instance, diffusion coefficients serve as critical inputs for modelling the transport of organic pollutants in groundwater systems, enabling accurate risk assessments and reasonable remediation designs for contaminated sites. Additionally, diffusion coefficients are critical parameters in predicting the uptake kinetics of organic pollutants in passive samplers, which has become widely deployed tools for monitoring trace organic contaminants across diverse water environments, including sediments and porewaters.[12,13] Despite the important role of organic compounds' aqueous diffusion coefficients as well as the lack of reliable experimental data, prediction of this parameter still remains challenging for organic compounds, due to the fact that diffusion processes involve complex interplay between different molecular structures and properties, and also substantially vary with water temperatures. These complexities notably restrict the examination and understanding of structure-property relationships of different organic compounds diffusion coefficients.

To determine diffusion coefficients of organic compounds in water, different empirical equations have been proposed, directly correlating molecules physical or chemical properties, including molecular weight, molar volume and temperature, to their diffusion coefficients.[14-21] However, the currently available equations are proved only applicable for precise estimation of the diffusion coefficients for a limited number of organic compounds, mainly owing to the following shortcomings: (i) the reliable data on organic compounds physical or chemical properties required for the empirical equations are often unavailable; (ii) the available equations are derived based on a limited number of experimental data, covering a narrow spectrum of organic compounds.[14,22-24] Notably, no empirical equations are simultaneously correlating molecular weight, molecular volume and water temperatures with organic compounds aqueous diffusion coefficients.

In recent years, artificial intelligence (AI) is being increasingly integrated into scientific discovery to augment and accelerate novel findings.[25] The advances in AI make it possible to automatically discover laws of nature in symbolic form.[26-30] Specifically, symbolic regression provides a systematic and interpretable method for identifying equations that optimally fit the provided data.[26,31] This technique plays a crucial role in generating hypotheses and discovering mathematical relationships.[32] For instance, by taking advantage of symbolic regression, AI can re-discover Kepler's third law.[29] Moreover, causality is essential to scientific investigation, forming the basis for comprehending interactions among variables within physical systems.[33] The causal inference enables the extraction of causal relationships, thereby facilitating the evaluation of reliability of proposed hypotheses. We believe that the above-mentioned approaches are posing a great potential to discover a precise, physically interpretable, and causality-based symbolic law to quantify diffusion coefficients of organic compounds in water.

Here, we combined experimental data, deep learning model, symbolic regression, and causal inference to establish a general theory of organic compounds diffusion coefficients in water. Specifically, a formula was developed by using a symbolic regression algorithm, expressing the mathematical relationship between molecular weight, molecular volume and temperatures with organic compounds diffusion coefficients in water. The proposed formula achieved a higher accuracy compared with the results determined by the currently available empirical equations. Furthermore, reliability of the obtained equation was verified based on causal inference.

MATERIALS AND METHODS

Data preparation

The experimental diffusion coefficients were sourced from the Institute for Scientific Information Web of Science, with datasets collected prior to 9 May 2023. An initial search yielded 7,285 literatures using the equation: topical subject (TS) = Diffusion coefficient AND TS = organic AND TS = water. From these, 64 studies were meticulously selected based on their adherence to three criteria: (1) availability of the full text; (2) the solute being organic compounds and the solvent being water; and (3) provision of the experimental temperature. Detailed information pertaining to these 64 studies is presented in Table S1. From literature survey, we constructed an extensive experimental dataset of 20,728 diffusion coefficients covering 6,913 compounds (Table S2). The SMILES codes for these compounds were obtained from PubChem. Furthermore, a total of 3,007 MDs were computed utilizing the alvaDesc software (Version 2.0.12). No conformer or tautomer averaging was performed during descriptor calculation. Of the 6,913 unique chemical structures, aliphatic compounds constituted the majority at 71.7%. Furthermore, carbocycles and heterocycles were found in smaller fractions, 22.6% and 5.7%, respectively. In addition, the molecular weights of the organic compounds in our dataset varied between 16 and 1,405 g mol^{-1} (Figure S1), suggesting that our dataset predominantly consists of small to medium-sized organic molecules. The whole dataset was randomly divided into training set and test set with ratio of 9:1, according to their unique chemical structures. Furthermore, previous studies have employed model-predicted data combined with experimental data as training labels for machine learning models and symbolic regression models.[27,34] However, the effects of incorporating model-predicted or synthetic data on predictive performance remain unclear. To expand the data volume and test impact of

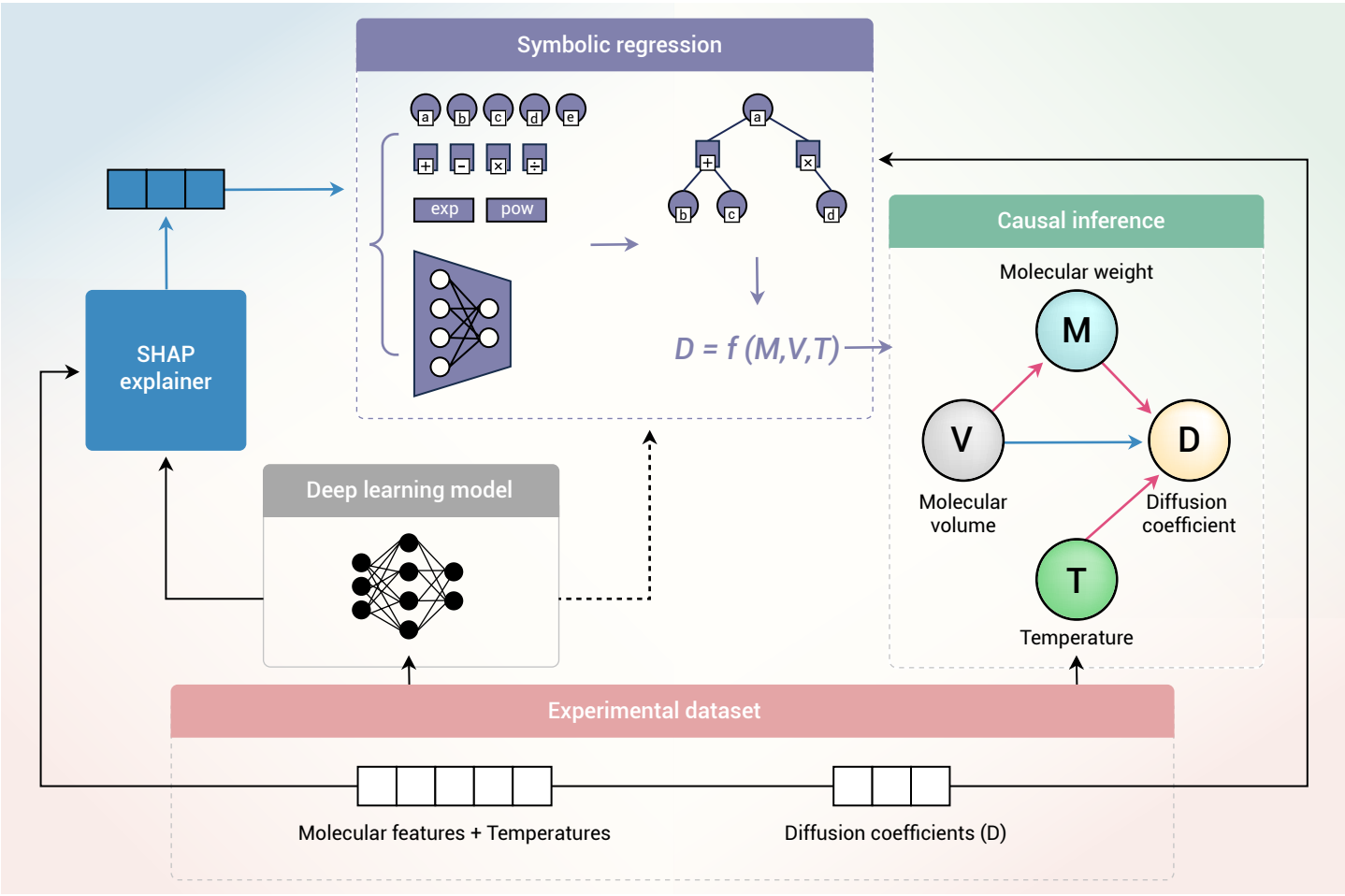


Figure 1. Overview for the approach A mathematical framework for the discovery of symbolic formula determining organic compounds diffusion in water, which combines deep learning model, SHapley Additive exPlanations (SHAP) method, symbolic regression, and causal inference. Firstly, an experimental dataset and a mixed dataset (i.e. mixing experimental data and synthetic data) was constructed. Secondly, the primary features were selected based on a deep learning model utilizing SHAP method. Thirdly, symbolic laws were extracted through symbolic regression based on experimental dataset and mixed dataset, respectively. Lastly, causal inference was utilized to validate the optimal equation.

incorporating synthetic data on model performance, we constructed a mixed dataset using our deep learning model.[35] Specifically, the deep learning model was utilized to interpolate diffusion coefficients at different temperatures and combined with the experimental training dataset to construct the mixed dataset.[27] The resulting mixed dataset contained about 110,000 data points. The detailed description of the deep learning model was presented in Text S1. In order to further assess the impact of synthetic data on model performance, sensitivity analyses were conducted by progressively increasing the proportion of model-generated data (0%, 25%, 50%, 75%, and 100%) incorporated into the experimental training dataset.

Symbolic law extraction

To obtain governing equations to quantify diffusion coefficients, we employed two strategies: a completely data-driven symbolic regression approach and a physics-informed hybrid model. For the data-driven approach, the candidate variables of symbolic models were selected before fitting with symbolic regression.[36] As a typical combination optimization problem, symbolic regression follows a brute force procedure. Therefore, the search space of the symbolic equation grows exponentially with the increasing number of variables. As mentioned before, the optimal features of the deep learning model showed a high dimensionality with counts of 198. To reduce the search space of symbolic models, we utilized the top 20 important features identified by SHapley Additive exPlanation (SHAP) method as the candidate variables of symbolic models. Details of descriptor computation and feature selection were provided in Text S2 and Table S3. Finally, symbolic regression was used to fit analytical expressions based on the datasets and chosen candidate variables (Figure 1). In a manner analogous to Occam's razor, the "fitness" of each expression is determined by its simplicity and accuracy. Initially, a variety of candidate analytical expressions are presented at varying complexity levels. In this context, complexity is quantified as the sum of operators, constants, and input variables. The simplicity and accuracy of each equation is assessed using complexity and mean square error (MSE), respectively. The equations with lower MSE and complexities are selected as the optimal equations. The symbolic models were fitted using a fixed set of operators, including "+", "×", "pow", "exp" and "inv(x) = 1/x". In order to enhance the interpretability of equations, the operators beside"+" and "×" are prohibited from being nested.[28]

To strengthen the physical insight of symbolic laws, we developed a hybrid model by integrating the theoretical core of the Wilke-Chang equation ($W = T/(V^{0.6}\cdot\eta)$) as a predefined feature within the symbolic regression framework. This term explicitly encodes the Stokes-Einstein relationship between diffusion, temperature, molecular size, and solvent viscosity (η), while allowing data-driven refinement through nonlinear extensions. The regression process retained the same operator constraints but introduced W as a fixed input variable alongside SHAP-selected features like molecular weight (M), enabling the discovery of corrective terms that augment the physical foundation with empirically observed nonlinearities. It is noted that the molecular volume used in W is McGowan volume.[37] The detailed information about McGowan volume was presented in Text S2.

The two aforementioned strategies were applied to construct symbolic models using experimental and mixed datasets, respectively. Model prediction quality was evaluated through comparison with experimental data using the following metrics: squared Pearson correlation coefficient (R^2), root-mean-square error (RMSE), mean absolute error (MAE) and mean square error (MSE). The detailed description of these metrics was provided in Text S3. In order to verify the effectiveness of the obtained equations, causal inference

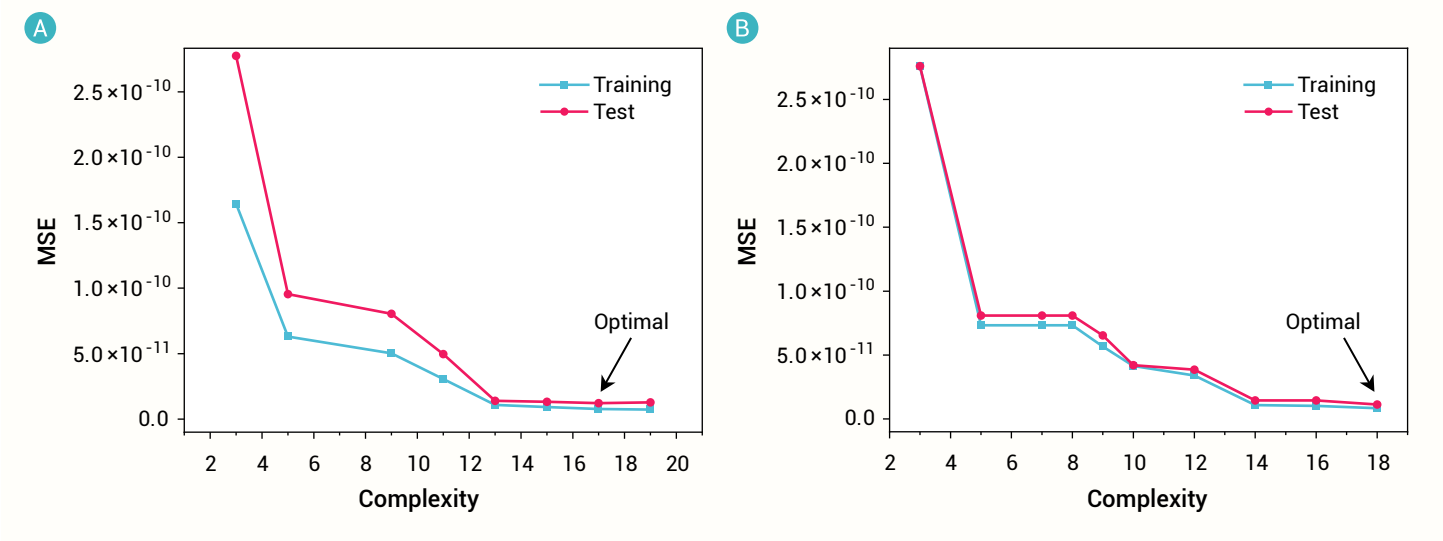


Figure 2. Training and test mean square error (MSE) against the feature complexity for the data-driven models based on mixed dataset (A) and experimental dataset (B).

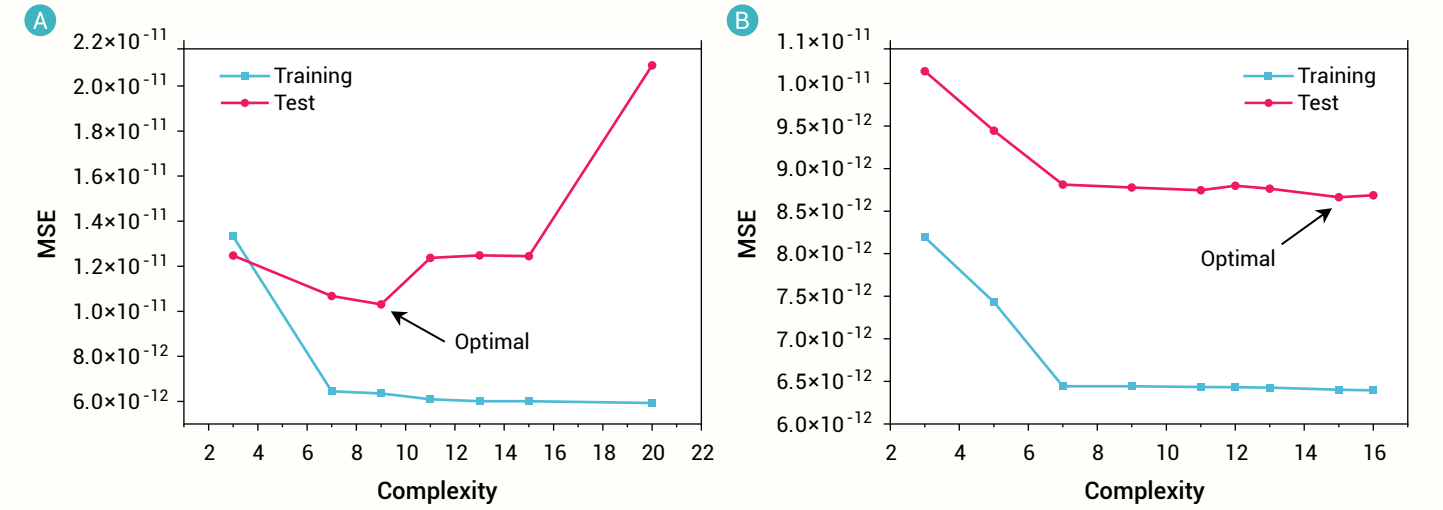


Figure 3. Training and test mean square error (MSE) against the feature complexity for the hybrid models based on mixed dataset (A) and experimental dataset (B).

was performed to identify the causal relationships between the experimental diffusion coefficients and the variables in the proposed symbolic equation (Text S4).

Programming and computational approaches

All models were programmed using Python 3.8.8. The deep learning models were established using Pytorch package. The SHAP package was used to select the important features.[38] The package PySR was utilized to discover symbolic laws of diffusion coefficient.[39] The causal inference was implemented by using EconML 0.13.1 package and DoWhy 0.7.1 package.[40] The symbolic law extraction and causal inference were accomplished using a Tianhe-2 Supercomputer.

RESULTS

Figure 2A shows the performance of data-driven symbolic models trained on mixed dataset. Further increasing the complexity would slightly enhance training accuracy, but at the cost of interpretability and generalizability (see Figure 2A). Specifically, the optimal model exhibited excellent performance (R^2 = 0.967) on the test set (see Table S4). Interestingly, the optimal model has the temperature, molecular weight and molecular volume naturally emerge as terms, with the models' form as:

$$D = \frac{1}{(e^{2.2768593 \cdot e^{2} \cdot 0.995762^{T}} - 0.061041337 \cdot M) \cdot (V + 38.09467)} \quad (1)$$

where D ($cm^2 \cdot s^{-1}$) is the diffusion coefficient, T (K) is the temperature, M ($g \cdot mol^{-1}$) is the molecular weight of diffusing compound and V ($cm^3 \cdot mol^{-1}$) is the molecular volume (McGowan volume) of diffusing compound.

In addition, data-driven models were trained using only the experimental dataset (Figure 2B & Table S5). The optimal model obtained from the experimental dataset demonstrated a slightly better performance (R^2 = 0.969) compared to its mixed-dataset counterpart (R^2 = 0.967), with a functional form characterized by:

$$D = \left(e^{\frac{T}{0.0800034 \cdot V + 59.35025}} + T - 0.18765856\right) \cdot 1.5495097 \cdot 10^{-7} - 4.5314013 \cdot 10^{-5} \quad (2)$$

where D ($cm^2 \cdot s^{-1}$) is the diffusion coefficient, T (K) is the temperature and V ($cm^3 \cdot mol^{-1}$) is the molecular volume (McGowan volume) of diffusing compound.

Furthermore, the hybrid model was trained using the mixed dataset, and the optimal formula was selected based on complexity and accuracy (see Figure 3A and Table S6). When the complexity of equation exceeds 9, the MSE on the training set decreases while the MSE on the test set increases. This indicates that overfitting begins when the complexity of formula exceeds 9. Therefore, the formula with a complexity of 9 is chosen as the optimal

formula:

$$D = 1.3578 \cdot 10^{-9} \cdot (W \cdot (W + T) + M) \tag{3a}$$

$$W = \frac{T}{V^{0.6} \cdot \eta} \tag{3b}$$

where D ($cm^2 \cdot s^{-1}$) is the diffusion coefficient, T (K) is the temperature, M ($g \cdot mol^{-1}$) is the molecular weight of diffusing compound, η ($kg \cdot ms^{-1}$) refers to the dynamic viscosity of the solvent and V ($cm^3 \cdot mol^{-1}$) is the molecular volume (McGowan volume) of diffusing compound.

Concerning the hybrid models trained exclusively using experimental data, we found that the model with a feature complexity of 15 is optimal by approaching the trade-off between predictive accuracy and model complexity (see Figure 3B and Table S7). The corresponding equation demonstrated superior performance (R^2 = 0.976) on the test set and is mathematically expressed as:

$$D = \left(\frac{2.884 \cdot 10^{-10} \cdot W}{\frac{1}{M} + 0.2028} + 5.2329 \cdot 10^{-7} \right) \cdot W \tag{4a}$$

$$W = \frac{T}{V^{0.6} \cdot \eta} \tag{4b}$$

with identical variable definitions as eq. 3.

Figure 4 compares the performance of optimal models developed under different construction strategies and training datasets. Specifically, the hybrid model trained on experimental dataset (eq. 4) performed the best (R^2 = 0.976, RMSE = 2.94×10^{-6} and MAE = 1.32×10^{-6}), whereas the data-driven model trained on mixed dataset (eq. 1) exhibited the lowest accuracy (R^2 = 0.967, RMSE = 3.49×10^{-6} and MAE = 2.32×10^{-6}). The sensitivity analysis of the optimal hybrid approach revealed progressive performance degradation (R^2 decreasing from 0.976 to 0.972) with increasing synthetic data (see Figures S2-S4 & 5). These findings confirm that eq. 4 achieved the highest predictive accuracy among all the models and thus is optimal. The performance of this selected equation was further compared with that of different empirical equations, including Wilke and Chang (eq. 5), Scheibel (eq. 6) , Hayduk and Laudie (eq. 7), Worch (eq. 8) and SEGWE (eq. 9).[16-21] The mathematical expressions of these equations are given:

$$D_i = \frac{7.4 \times 10^{-8} \cdot (\chi M_S)^{\frac{1}{2}} T}{\eta * V_i^{0.6}} \tag{5}$$

$$D_i = \frac{8.2 \cdot 10^{-8} T}{\eta V_i^{1/3}} \left[1 + \left(\frac{3V_s}{V_i} \right)^{2/3} \right] \tag{6}$$

$$D_i = \frac{13.36 \cdot 10^{-5}}{\eta^{1.14} V_i^{0.589}} \tag{7}$$

$$D_i = \frac{3.595 \cdot 10^{-7} T}{\eta m_i^{0.53}} \tag{8}$$

$$D_i = \frac{k_B T \left(\frac{3\alpha}{2} + \frac{1}{1+\alpha} \right) \cdot 10^4}{6\pi\eta \sqrt[3]{\frac{3m_i}{4\pi\rho_{eff} N_A}}} \tag{9a}$$

$$\alpha = \sqrt[3]{\frac{M_S}{m_i}} \tag{9b}$$

where D_i ($cm^2 \cdot s^{-1}$) is the diffusion coefficient of diffusing compound i, χ is the empirical association parameter of the solvent, M_S is the molecular mass of solvent, T (K) is the temperature, η ($kg \cdot ms^{-1}$) refers to the dynamic viscosity of the solvent, m_i ($g \cdot mol^{-1}$) is the molecular mass of diffusing compound i, V_i ($cm^3 \cdot mol^{-1}$) is the molar volume of diffusing compound i, V_s ($cm^3 \cdot mol^{-1}$) is the molar volume of solvent, k_B ($J \cdot K^{-1}$) is the Boltzmann's constant, N_A is the Avogadro number, ρ_{eff} (619 $kg \cdot m^{-3}$) is the effective density. Comparing with the equation proposed in this study, the equations defined by Wilke and Chang, Scheibel, Hayduk and Laudie, Worch and SEGWE returned the lower R^2 values of 0.912, 0.878, 0.920, 0.938 and 0.809 on the test set, respectively. Additionally, the RMSE (> 4.75×10^{-6}) and MAE (> 2.48×10^{-6}) values of the above equations were higher than the ones of our proposed equation (2.94×10^{-6} and 1.32×10^{-6}). This suggested that the proposed equation is more accurate for predicting organic compounds diffusion coefficients in water.

DISCUSSION

Comparison of data-driven model and hybrid model

The purely data-driven equation, derived via symbolic regression without prior physical constraints, demonstrated robust predictive accuracy across the test datasets (R^2 = 0.967-0.969). But the use of mathematically complex functions (such as nested exponentials in temperature term) and empirically adjusted coefficients makes it physically uninterpretable. While such equation works well for interpolation within the training domain, it lacks a clear connection to the established diffusion theory (e.g., Arrhenius-type temperature dependence or Stokes-Einstein hydrodynamics), limiting its interpretability and broader applicability. In contrast, the hybrid model not only retains physical transparency through explicit encoding of Stokes-Einstein hydrodynamics (via W) but also outperforms the data-driven model. This enhancement arises from the incorporation of domain-specific knowledge. By introducing the Wilke-Chang term, the model is better equipped to accurately capture the temperature dependency of diffusion coefficients. Significantly, the hybrid framework illustrates that the incorporation of physical priors does not need to compromise empirical accuracy. Instead, it can augment it by constraining the symbolic search space to physiochemically meaningful relationships.

Experimental data versus synthetic data

In comparison, symbolic models trained exclusively on experimental data demonstrated marginally superior performance. Specifically, both data-driven and hybrid methods achieved slightly higher accuracy using experimental datasets (eq. 2: R^2 = 0.969; eq. 4: R^2 = 0.976) than mixed datasets (eq. 1: R^2 = 0.967; eq. 3: R^2 = 0.972). Furthermore, sensitivity analysis of the hybrid model revealed gradual performance degradation with increasing synthetic data. The performance degradation likely stemmed from two mechanisms: amplification of inherited parent model bias during training, and distortion of descriptor-target associations caused by noise in synthetic data. These findings indicated that data quality may hold greater significance than data quantity in model development. Therefore, high-quality experimental data should be prioritized for model training and testing.

Analysis of model structure uncertainty

To quantify the model-form uncertainty and verify the reliability of the optimal model, expressions and performance metrics of top-performing models across diverse construction strategies and training datasets were compared (Table S8). In general, these models exhibited consistent structural patterns. Specifically, six out of seven models consistently include temperature, molecular volume, molecular weight, and viscosity as predictive variables. In these six models, temperature and molecular weight demonstrated positive correlations with diffusion coefficients, whereas molecular volume and viscosity showed negative impacts. To further assess the robustness of relationships derived from eq. 4, we analysed the top-five hybrid models trained exclusively on experimental datasets (Table S7). Notably, four of these five models included the identical descriptors (temperature, molecular volume, molecular weight, and viscosity), with functional forms maintaining the relationships observed in the optimal model. This consistency across multiple high-performing models confirms the structural stability of the inferred physicochemical relationships in the optimal model.

Nature of organic compounds aqueous diffusion coefficients

Similar to the above-mentioned empirical equations, the temperature and molecular volume were positively and negatively related to diffusion coefficient in our equation, respectively. Specifically, temperature governs diffusion through dual mechanisms: (1) direct thermal activation, where higher

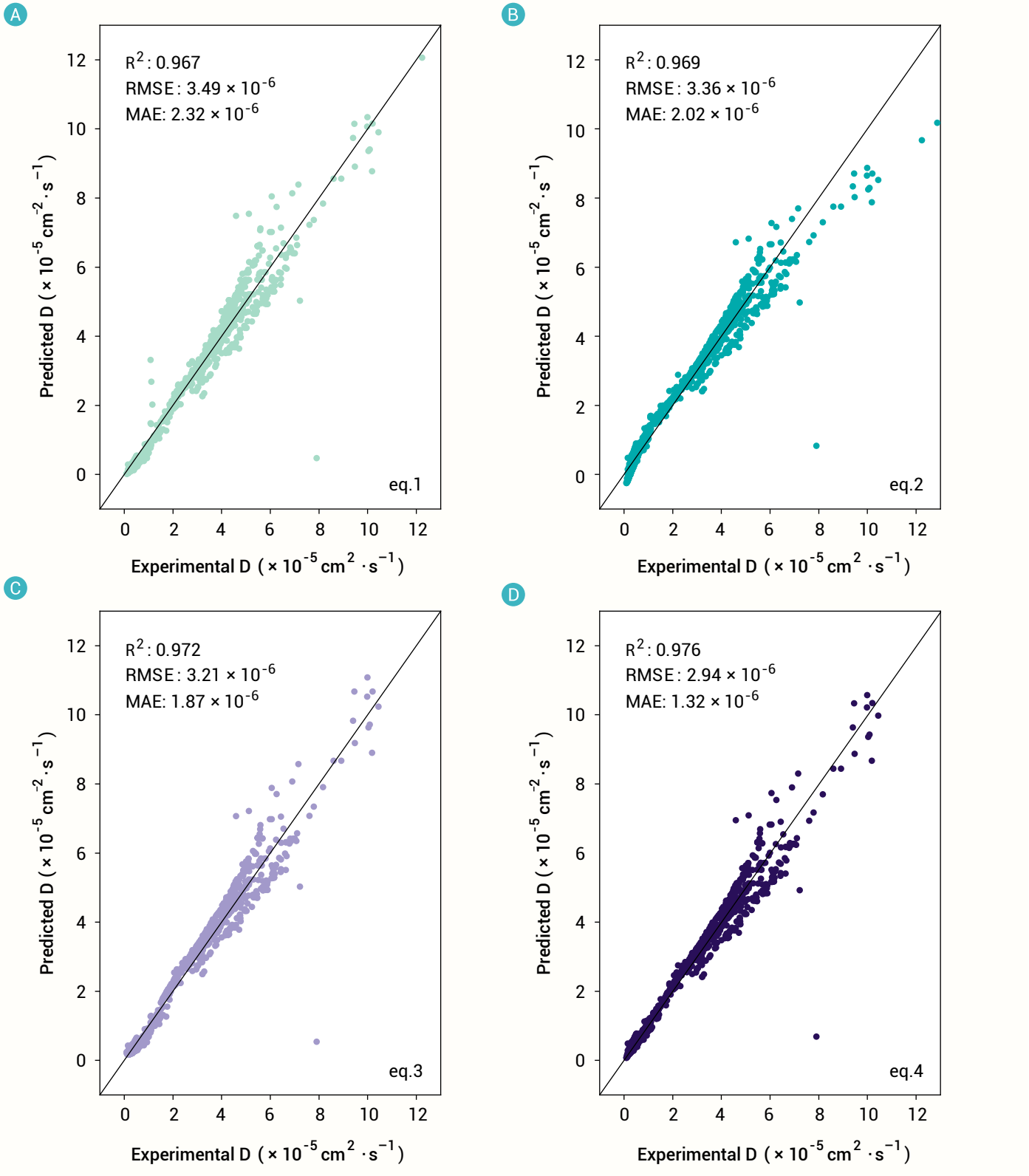


Figure 4. Test accuracy of optimal models (A) Data-driven model trained on mixed dataset (eq. 1), (B) Data-driven model trained on experimental dataset (eq. 2), (C) Hybrid model trained on mixed dataset (eq. 3), (D) Hybrid model trained on experimental dataset (eq. 4).

temperature accelerates molecular motion, reflected in the linear temperature-dependence of the Wilke-Chang term; and (2) indirect viscosity (η) modulation, as η decreases exponentially with temperature, amplifying diffusion coefficients. Molecular volume primarily imposes hydrodynamic drag, with the $V^{-0.6}$ term in W aligning with Stokes-Einstein scaling to describe size-dependent retardation. Nevertheless, there are two differences between our equation and the previous empirical equations. Firstly, the molecular-related term in previous equations usually only included molecular weight or molecular volume of diffusing compounds (Table 1 and eq. 5-9). For instance, molecular-related terms in the equation of Worch and SEGWE only included molecular weight (eq. 8-9), and the empirical equations of Wilke and Chang, Scheibel and Hayduk and Laudie only consider molecular volume (eq. 5-7).

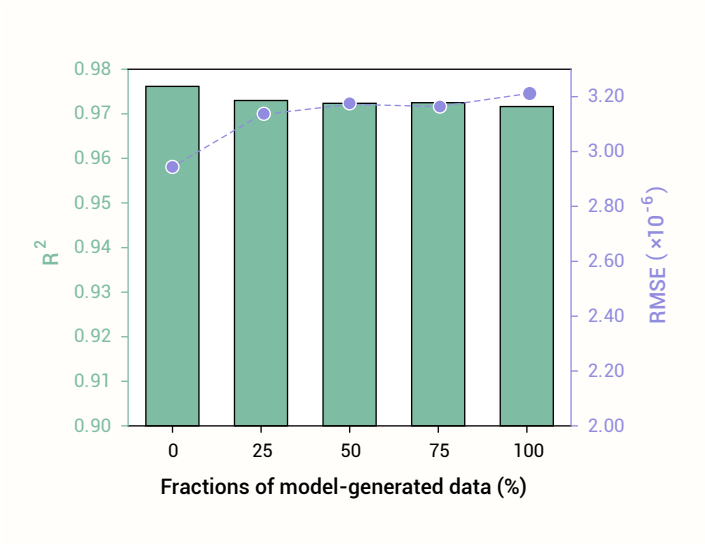


Figure 5. Sensitivity analysis of hybrid model performance to progressive synthetic data integration.

Table 1. Comparison of prediction performance for organic compounds aqueous diffusion coefficients using different equations.

Equation	[a]Variables	[b]R^2	[c]RMSE	[d]MAE
This study	M, V, T, η	0.976	2.94×10^{-6}	1.32×10^{-6}
Wilke and Chang	V, T, η	0.912	5.64×10^{-6}	2.75×10^{-6}
Scheibel	V, T, η	0.878	6.67×10^{-6}	3.41×10^{-6}
Hayduk and Laudie	V, η	0.920	5.40×10^{-6}	2.48×10^{-6}
Worch	M, T, η	0.938	4.75×10^{-6}	2.75×10^{-6}
SEGWE	M, T, η	0.809	8.325×10^{-6}	4.79×10^{-6}

[a]M ($g\cdot mol^{-1}$) is the molecular weight of diffusing compound; V ($cm^3\cdot mol^{-1}$) is the molecular volume of diffusing compound, T (K) is the temperature and η ($kg\cdot ms^{-1}$) refers to the dynamic viscosity of the solvent. [b]R^2 is squared Pearson correlation coefficient. [c]RMSE is root-mean-square error. [d]MAE is mean absolute error.

However, the molecule-related term in our optimal equation consisted of both molecular weight and molecular volume. Given the diversity of molecular structures, individual variable either molecular weight or molecular volume along might be insufficient to provide a detailed description of the molecular structures and properties. As depicted in Figure 6B, the compounds with the same molecular weight but different molecular volumes have different aqueous diffusion coefficients. For instance, the diffusion coefficients of cyclic molecules are usually greater than the ones of aliphatic molecules with the same molecular weight.[41] Based on our dataset, 125 pairs of cyclic molecules and aliphatic molecules of the same molecular weight were compared (Table S9). The results indicated that the differences on diffusion coefficients between the cyclic molecules and the aliphatic molecules could be induced by different molecular volume. Additionally, given the compounds with the identical molecular volume but different molecular weight has distinctive diffusion coefficients (Figure 6B), it is necessary to combine molecular weight and molecular volume to describe the structural information of organic compounds in the governing equation for calculation of diffusion coefficients. Secondly, molecular weight is usually negatively correlated to diffusion coefficients in the empirical equations (eq. 6 & 7). In contrast, a positive correlation between molecular weight and the diffusion coefficients was obtained in our equation (eq. 4), which contradicts the results revealed by SHAP analysis (Figure S5). The fact is that the SHAP method is good at identifying correlations, rather than capturing causal relationships.[42] In fact, the identified correlations might be inequivalent with causality due to mediation or confounding effects.[43] In order to eliminate pseudo-correlation and verify the effectiveness of our equation, causal inference was performed to identify the causal relationships between the diffusion coefficients and the variables in the proposed symbolic equation.[44] Utilizing the experimental dataset, we constructed a weighted causal diagram to illustrate the outcomes of causal effect estimations, depicted in Figure 6A. The detailed causal estimation and validation results are given in Table S10. The causal diagram showed that the causal effects of temperature, molecular volume and molecular weight on diffusion coefficient were 3.3×10^{-7}, -2×10^{-8} and 2.1×10^{-9}, respectively (Figure 6A), indicating that the water temperature and the molecular weight are positively contributing to the diffusion coefficients, but the molecular volume negatively influenced the diffusion coefficients. Moreover, the influence of temperature, molecular volume, and molecular weights on the diffusion coefficients fell in turn. The observed trends aligned with the proposed equation. As mentioned above, molecular volume could affect both molecular weight and diffusion coefficients, thereby acting as a confounding factor for these two variables (Figure 6A). Specifically, molecular volume induced a positive impact on molecular weight and a negative impact on diffusion coefficients, and thus resulting in a spurious negative relationship between molecular weight and diffusion coefficients (blue broken line in Figure 6A). Therefore, a causal relationship between molecular weight and diffusion coefficients could be resolved once the confounding impact of molecular volume was eliminated. These results suggested that the proposed equation successfully identified the causal relationship between the diffusion coefficients and molecular weight and molecular volume.

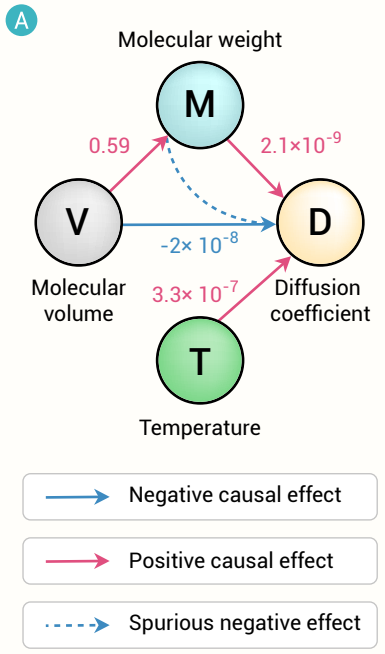
To further confirm the reliability of the proposed equation, a controlled test was performed. Concerning water temperature, diffusion coefficients of the organic compounds at different water temperatures were compared. For all the compounds in the experimental dataset, the diffusion coefficients increased at increasing water temperatures. For instance, the diffusion coefficient of Cyanogen chloride at 298.15 K is higher than the one at 274 K (Figure 6B). Also, we compared the diffusion coefficients of the organic compounds with identical molecular weight but different molecular volume at same temperature. A negative correlation between the molecular volume and diffusion coefficients was observed (Figure 6B). Moreover, at the same molecular volume and water temperature, molecules of larger molecular weight typically exhibit the higher diffusion coefficients than the ones with smaller molecular weights (Table S11). For instance, the molecular weight and diffusion coefficients of 3-iodo-1-propene are higher than the ones of butyric acid at the same water temperature as shown in Figure 6B. These results further demonstrated the validity of the proposed equation.

Limitations, Outlook and Geoscientific Applicability

Despite its advancements, the hybrid model faces several limitations that warrant consideration. While the selected molecular descriptors (V, M) capture dominant diffusion controls, they incompletely represent illustrate all the factors affecting the physical properties of a molecule. In addition, the dataset in this work compiles diffusion coefficients measured by different sources and techniques, which may introduce systematic offsets.

To address these gaps, future work could pursue the following interconnected directions: (1) integrating quantum-chemical descriptors to resolve electronic effects beyond V and M; (2) implementing group-aware data handling by treating measurement techniques as grouping variables during cross-validation to address dataset heterogeneity and method bias; (3) employing extreme learning machine (ELM) ensembles separately on experimental and mixed datasets to detect model-predicted data propagation;[45] (4) testing descriptor reproducibility by grouping based on descriptor calculation methods and assessing stability across definitions; and (5) extending the HYDRUS-ML framework to generate synthetic diffusion datasets under varying conditions (porosity, tortuosity, adsorption, pH, ionic strength, and DOM) for symbolic regression mapping to effective diffusion coefficients.[46]

Nevertheless, within the applicability domain, the equation offers significant utility for geoscientific applications. One practical application of our equation is the design of passive samplers, including the widely-used diffusive gradients in thin-films (DGT). The precise diffusion coefficients predicted by our equation are crucial for accurate determination of pollutants' aqueous concentrations. Furthermore, the equation developed in this study could be



B

	M	V	T	D
3-iodo-1-propene	167.98	124.0	274	4.75×10^{-6}
Butyric acid	88.12	124.0	274	4.44×10^{-6}
Isoprene	68.13	120.8	298.15	9.62×10^{-6}
Cyclopentene	68.13	109.9	298.15	1.03×10^{-5}
Cyanogen chloride	61.47	64.1	274	7.27×10^{-6}
Cyanogen chloride	61.47	64.1	298.15	1.48×10^{-5}

Figure 6. The validation of symbolic formula (A) A causal graph with the weighted edge that indicates the estimated causal effect of different causal links. (B) Comparison of diffusion coefficients for different compounds under diffusion temperatures.

utilized to calculate the diffusion coefficients of newly identified organic pollutants, which provides essential input parameters for modelling groundwater transport of these emerging contaminants as well as for risk assessment of groundwater contamination.

CONCLUSIONS

By symbolic regression with experimental data, we developed a general theory for diffusion coefficients that can be applied to a wide range of organic compounds at varying temperatures. A concise equation comprising molecular weight, molecular volume, temperature and viscosity is developed to calculate the organic compounds aqueous diffusion coefficients at varying temperatures. Notably, the proposed equation led to more accuracy prediction of the diffusion coefficients compared with the ones determined by using the currently available empirical equations. Besides, the sensitivity test reveals that high-quality experimental data should be prioritized for model construction. Remarkably, the results revealed by causal inference suggested that the proposed equation successfully captures the causal relationship between organic compounds diffusion coefficients and the above-mentioned variables. These results highlight the importance of transparent data-driven theory in discovery of symbolic laws determining organic compounds diffusion coefficients in water. The equation developed in our study will serve as a valuable tool to predict accurate aqueous diffusion coefficients of organic compounds for different practices and applications in geosciences.

REFERENCES

1. Meng Y., Jiang J., Wu J., et al. (2024). A physics-enhanced neural network for estimating longitudinal dispersion coefficient and average solute transport velocity in porous media. *Geophys. Res. Lett.* **51**:e2024GL110683. DOI:10.1029/2024GL110683
2. Cussler E. L. (2015). Diffusion: Mass transfer in fluid systems (Cambridge University). DOI:10.1017/CBO9780511805134
3. LaBolle E. M., Fogg G. E., Eweis J. B., et al. (2008). Isotopic fractionation by diffusion in groundwater. *Water Resour. Res.* **44**:W07405. DOI:10.1029/2006WR005264
4. Ma Y., Zhu C., Ma P., et al. (2005). Studies on the diffusion coefficients of amino acids in aqueous solutions. *J. Chem. Eng. Data* **50**:1192–1196. DOI:10.1021/je049582g
5. Rana A., Renault C. and Dick J. E. (2023). Measuring liquid-into-liquid diffusion coefficients by dissolving microdroplet electroanalysis. *Anal. Chem.* **95**:18748–18753. DOI:10.1021/acs.analchem.3c03256
6. Weisz P. B. (1973). Diffusion and chemical transformation. *Science* **179**:433–440. DOI:10.1126/science.179.4072.433
7. Egger M., Riedinger N., Mogollón J. M., et al. (2018). Global diffusive fluxes of methane in marine sediments. *Nat. Geosci.* **11**:421–425. DOI:10.1038/s41561-018-0122-8
8. Jin B., Rolle M., Li T., et al. (2014). Diffusive fractionation of BTEX and chlorinated ethenes in aqueous solution: quantification of spatial isotope gradients. *Environ. Sci. Technol.* **48**:6141–6150. DOI:10.1021/es4046956
9. Lu W., Guo H., Chou I. M., et al. (2013). Determination of diffusion coefficients of carbon dioxide in water between 268 and 473K in a high-pressure capillary optical cell with in situ Raman spectroscopic measurements. *Geochim. Cosmochim. Acta* **115**:183–204. DOI:10.1016/j.gca.2013.04.010
10. Tanaka M., Takahashi Y., Yamaguchi N., et al. (2013). The difference of diffusion coefficients in water for arsenic compounds at various pH and its dominant factors implied by molecular simulations. *Geochim. Cosmochim. Acta* **105**:360–371. DOI:10.1016/j.gca.2012.12.004
11. Cao X. (2022). Diffusional isotope fractionation of singly and doubly substituted isotopologues of H_2, N_2 and O_2 during air-water gas transfer. *Geochim. Cosmochim. Acta* **332**:78–87. DOI:10.1016/j.gca.2022.06.022
12. Liu S., Jin L., Yu H., et al. (2020). Understanding and predicting the diffusivity of organic chemicals for diffusive gradients in thin-films using a QSPR model. *Sci. Total Environ.* **706**:135691. DOI:10.1016/j.scitotenv.2019.135691
13. Urík J., Paschke A. and Vrana B. (2020). Diffusion coefficients of polar organic compounds in agarose hydrogel and water and their use for estimating uptake in passive samplers. *Chemosphere* **249**:126183. DOI:10.1016/j.chemosphere.2020.126183
14. Gharagheizi F. (2012). Determination of diffusion coefficient of organic compounds in water using a simple molecular-based method. *Ind. Eng. Chem. Res.* **51**:2797–2803. DOI:10.1021/ie201944h
15. Einstein A. (1905). Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.* **322**:549–560. DOI:10.1002/andp.19053220806
16. Wilke C. R. and Chang P. (1955). Correlation of diffusion coefficients in dilute solutions. *AIChE J.* **1**:264–270. DOI:10.1002/aic.690010222
17. Scheibel E. G. (1954). Correspondence. Liquid diffusivities. Viscosity of gases. *Ind. Eng. Chem.* **46**:2007–2008. DOI:10.1021/ie50537a062
18. Hayduk W. and Laudie H. (1974). Prediction of diffusion coefficients for nonelectrolytes in dilute aqueous solutions. *AIChE J.* **20**:611–615. DOI:10.1002/aic.690200329
19. Worch E. (1993). Eine neue Gleichung zur Berechnung von Diffusionskoeffizienten gelöster Stoffe: A new equation for the calculation of diffusion coefficients for dissolved substances. *Vom Wasser* **81**:289–297. https://www.researchgate.net/publication/313202269_Eine_neue_Gleichung_zur_Berechnung_von_Diffusionskoeffizienten_geloster_Stoffe
20. Evans R., Deng Z., Rogerson A. K., et al. (2013). Quantitative interpretation of diffusion-ordered NMR spectra: Can we rationalize small molecule diffusion coefficients. *Angew. Chem. Int. Ed.* **52**:3199–3202. DOI:10.1002/anie.201207403
21. Evans R., Dal Poggetto G., Nilsson M., et al. (2018). Improving the interpretation of small molecule diffusion coefficients. *Anal. Chem.* **90**:3987–3994. DOI:10.1021/acs.analchem.7b05032

22. Gharagheizi F., Eslamimanesh A., Mohammadi A. H., et al. (2011). Representation and prediction of molecular diffusivity of nonelectrolyte organic compounds in water at infinite dilution using the artificial neural network-group contribution method. *J. Chem. Eng. Data* **56**:1741–1750. DOI:10.1021/je101190p
23. Bellaire D., Großmann O., Münnemann K., et al. (2022). Diffusion coefficients at infinite dilution of carbon dioxide and methane in water, ethanol, cyclohexane, toluene, methanol, and acetone: A PFG-NMR and MD simulation study. *J. Chem. Thermodyn.* **166**:106691. DOI:10.1016/j.jct.2021.106691
24. Miyabe K. and Isogai R. (2011). Estimation of molecular diffusivity in liquid phase systems by the Wilke–Chang equation. *J. Chromatogr. A* **1218**:6639–6645. DOI:10.1016/j.chroma.2011.07.018
25. Wang H., Fu T., Du Y., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature* **620**:47–60. DOI:10.1038/s41586-023-06221-2
26. Schmidt M. and Lipson H. (2009). Distilling free-form natural laws from experimental data. *Science* **324**:81–85. DOI:10.1126/science.1165893
27. Li Y., Wang H., Li Y., et al. (2023). Electron transfer rules of minerals under pressure informed by machine learning. *Nat. Commun.* **14**:1815. DOI:10.1038/s41467-023-37384-1
28. Liu S., Li Q., Shen X., et al. (2024). Automated discovery of symbolic laws governing skill acquisition from naturally occurring data. *Nat. Comput. Sci.* **4**:334–345. DOI:10.1038/s43588-024-00629-0
29. Cornelio C., Dash S., Austel V., et al. (2023). Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat. Commun.* **14**:1777. DOI:10.1038/s41467-023-37236-y
30. Wang T., Hu J., Ouyang R., et al. (2024). Nature of metal-support interaction for metal catalysts on oxide supports. *Science* **386**:915–920. DOI:10.1126/science.adp6034
31. Udrescu S.-M. and Tegmark M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* **6**:eaay2631. DOI:10.1126/sciadv.aay2631
32. Li X. and Guo Y. (2025). Paradigm shifts from data-intensive science to robot scientists. *Sci. Bull.* **70**:14–18. DOI:10.1016/j.scib.2024.09.029
33. Martínez-Sánchez Á., Arranz G. and Lozano-Durán A. (2024). Decomposing causality into its synergistic, unique, and redundant components. *Nat. Commun.* **15**:9296. DOI:10.1038/s41467-024-53373-4
34. Zhao Q., Yu Y., Gao Y., et al. (2022). Machine learning-based models with high accuracy and broad applicability domains for screening PMT/vPvM substances. *Environ. Sci. Technol.* **56**:17880–17889. DOI:10.1021/acs.est.2c06155
35. Han M., Liang J., Wang Z., et al. (2025). Accurate prediction of the diffusion coefficients of organic compounds in water by multimodal learning. *J. Phys. Chem. A.* DOI:10.1021/acs.jpca.5c01881
36. Koza J. R. (1992). Genetic programming: On the programming of computers by means of natural selection (MIT Press). DOI:10.1016/0303-2647(94)90062-0
37. Zhao Y. H., Abraham M. H. and Zissimos A. M. (2003). Determination of McGowan volumes for ions and correlation with van der Waals Volumes. *J. Chem. Inf. Comput. Sci.* **43**:1848–1854. DOI:10.1021/ci0341114
38. Lundberg S. M., Erion G., Chen H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**:56–67. DOI:10.1038/s42256-019-0138-9
39. Cranmer M., Sanchez-Gonzalez A., Battaglia P., et al. (2020). Discovering symbolic models from deep learning with inductive biases. *Adv. Neural Inf. Process. Syst. Curran Associates Inc.***1462**:17429–17442. DOI:10.5555/3495724.3497186
40. Sharma A. and Kiciman E. (2020). DoWhy: An end-to-end library for causal inference. *arXiv e-prints* arXiv:2011.04216. DOI:10.48550/arXiv.2011.04216.
41. Bogdan A. R., Davies N. L. and James K. (2011). Comparison of diffusion coefficients for matched pairs of macrocyclic and linear molecules over a drug-like molecular weight range. *Org. Biomol. Chem.* **9**:7727–7733. DOI:10.1039/C1OB05996C
42. Han M., Liang J., Jin B., et al. (2024). Machine learning coupled with causal inference to identify COVID-19 related chemicals that pose a high concern to drinking water. *iScience* **2**:109012. DOI:10.1016/j.isci.2024.109012
43. Ombadi M., Nguyen P., Sorooshian S., et al. (2020). Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resour. Res.* **56**:e2020WR027251. DOI:10.1029/2020WR027251
44. Kang Q., Zhang B., Cao Y., et al. (2024). Causal prior-embedded physics-informed neural networks and a case study on metformin transport in porous media. *Water Res.* **261**:121985. DOI:10.1016/j.watres.2024.121985
45. Saberi-Movahed F., Najafzadeh M. and Mehrpooya A. (2020). Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: Training group method of data handling using extreme learning machine conceptions. *Water Resour. Manage.* **34**:529–561. DOI:10.1007/s11269-019-02463-w
46. Zhu S., Liu B., Li S., et al. (2025). Simulation and prediction of sulfamethazine migration, transformation and risk diffusion during cross-media infiltration from surface water to groundwater driven by dynamic water level: Machine learning coupled HYDRUS-GMS model. *J. Environ. Manage.* **373**:123484. DOI:10.1016/j.jenvman.2024.123484

FUNDING AND ACKNOWLEDGMENTS

This study is funded by Guangdong Major Project of Basic and Applied Basic Research (2023B0303000007) and the project of Theory of Hydrocarbon Enrichment under Multi-Spheric Interactions of the Earth (THEMSIE04010104). B. J. acknowledges the State Key Laboratory of Advanced Environmental Technology with grant NO. SKLAET2025-LH02. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

B.J. designed the study; M.H. developed symbolic models and analyzed the simulation results together with B.J.; M.H. wrote the manuscript under B.J.'s supervision. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

The data supporting this article have been included as part of the Supplemental information.

SUPPLEMENTAL INFORMATION

It can be found online at https://doi.org/10.59717/j.xinn-geo.2026.100208.