PCCP
Physical Chemistry Chemical Physics

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: A. A. Roosta, N. Rezaei and H. R. Godini, *Phys. Chem. Chem. Phys.*, 2026, DOI: 10.1039/D6CP01425A.

This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/pccp

View Article Online
DOI: 10.1039/D6CP01425A

ARTICLE

Prediction of self-diffusion coefficients via a hybrid PCP-SAFT+ANN model incorporating COSMO-SAC sigma-profile descriptors

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Aliakbar Roosta*[a,b], Nima Rezaei[a], Hamid Reza Godini[b]

Reliable estimation of self-diffusion coefficients is fundamental for characterizing mass transport in fluids; however, accurate prediction remains difficult due to the strong influence of thermodynamic conditions (temperature and pressure) and molecular characteristics such as size, shape, and intermolecular forces. In this study, a hybrid predictive model is introduced, combining the PCP-SAFT equation of state with an artificial neural network (ANN) to estimate self-diffusion coefficients over a broad range of conditions. The model is developed using a dataset, comprising of 2263 experimental measurements for 67 compounds, spanning temperatures between 93.0 and 973.2 K, pressures up to 3036 bar, corresponding to self-diffusion coefficients, spanning nearly five orders of magnitude from 10^{-12} to 10^{-7} m^2 s^{-1}. For a rigorous assessment of predictive performance, the dataset was partitioned into 30% reserved for independent validation and 70% for training. The proposed model incorporates thermodynamic inputs, namely density and dimensionless form of residual entropy obtained from PCP-SAFT, together with molecular descriptors derived from COSMO-SAC sigma-profiles. The selected ANN architecture, comprising two hidden layers with 14 and 7 neurons, respectively, provides high predictive performance, achieving R^2 values of 0.9937 and 0.9763 and AARD values of 8.89% and 15.89% for the training and testing datasets. Overall, the proposed framework offers a unified reliable model for predicting diffusion behavior under diverse thermodynamic conditions.

1 Introduction

Self-diffusion coefficient is a fundamental transport property. It characterizes molecular diffusion in fluids and plays a critical role in a wide range of engineering and scientific applications, including material science,[1] separation processes,[2] reaction engineering,[3] energy systems,[4] and corrosion [5]. However, predicting self-diffusion coefficient remains as a challenging task due to strong dependence on temperature, pressure, intermolecular interactions, and molecular structure (e.g., size, shape, and functional groups), whose effects on transport behavior are difficult to be captured in a unified predictive model.[6]

Experimental measurements of self-diffusion coefficients is often time-consuming, costly, and limited to specific compounds and conditions. As a result, there is a growing need for predictive models capable of estimating diffusion coefficients over wide thermodynamic ranges and for diverse chemical families. Conventional approaches such as theoretical[7–9] and semi-empirical models[10–12] have been developed to describe the self-diffusion behavior. While such methods can provide reasonable accuracy for specific systems, they are typically restricted to the compounds and the experimental conditions studied, limiting their applicability to new or uncharacterized compounds and their predictability outside the studied conditions.

To address these limitations, recent efforts have focused on data-driven approaches, particularly machine learning techniques such as artificial neural networks (ANNs), to establish relationships between molecular structure, thermodynamic properties, and transport behavior.[13–16]

As a potential source of molecular-level input for these models, COSMO-based methods have emerged as powerful tools for describing molecular interactions through sigma-profiles, which represent the surface charge density distribution of the molecules.[17,18] These profiles can be transformed into compact molecular descriptors that capture key molecular features such as polarity, charge distribution, and hydrogen-bonding capability. When combined with machine learning models, these descriptors enable the development of predictive models that incorporate molecular-level information into the model.[19–21]

In this work, we propose a hybrid modeling approach that integrates COSMO-SAC-derived molecular descriptors with thermodynamic properties (density and dimensionless form of residual entropy) calculated from perturbed-chain polar statistical associating fluid theory (PCP-SAFT) equation of state within an ANN framework. The developed model is designed to provide accurate predictions of self-diffusion coefficients for different chemicals across a wide range of temperatures and pressures. A notable advantage of the proposed model is its

[a] Department of Separation Science, School of Engineering Science, LUT University, Lappeenranta, Finland. E-mail: aliakbar.roosta@aalto.fi
[b] Department of Energy and Mechanical Engineering, School of Engineering, Aalto University, Espoo, Finland

View Article Online
DOI: 10.1039/D6CP01425A

broad applicability across extensive operating conditions, including elevated temperatures (up to 973 K) and pressures (up to 3036 bar), conditions that are seldom covered in the existing studies. This capability is particularly relevant for practical applications where fluids are exposed to extreme conditions, such as high-temperature and high-pressure processes, including supercritical separation.[22,23]

The remainder of this paper is structured as follows. Section 2 describes the dataset compilation, the extraction of molecular descriptors from COSMO-SAC sigma-profiles, and the implementation of the PCP-SAFT equation of state, along with the design and training of the ANN model. Section 3 presents a detailed evaluation of the model performance using statistical metrics, graphical analyses, and sensitivity analysis to identify the most influential input variables. Finally, Section 4 summarizes the main findings and outlines potential directions for future work.

2 Methods and Data Collection

2.1 Self-diffusion coefficient data sources

Experimental self-diffusion coefficient data were gathered from the literature across a wide spectrum of chemical families, encompassing alkanes, cycloalkanes, alcohols, polyols, aromatics, ketones, ethers, water, olefins, nitriles, amines, and halogenated species. In total, the assembled dataset includes 2263 measurements corresponding to 67 different compounds, ensuring coverage of broad chemical diversity. The collected data also cover extensive thermodynamic conditions, with temperatures varying between 93 and 973 K, pressures ranging from 1 to 3036 bar, and self-diffusion coefficients spanning from 1.89×10^{-12} to 3.61×10^{-7} m^2 s^{-1}. Having such wide ranges of input data enables the model to be trained and evaluated across significantly different conditions affecting molecular diffusion. To enhance the reliability of the dataset, the data were obtained from multiple independent literature sources, thereby reducing potential biases arising from individual studies and better representing experimental variability. An overview of the compiled dataset is provided in Table 1, where the investigated compounds, the corresponding operating conditions, and the number of available data points are summarized. To develop and validate the ANN model, the dataset was divided into training and testing stages by also considering the chemical species classes, so that both sub-datasets include a wide range of chemical types. Specifically, 45 compounds (1586 data points) were assigned to the training set, while the remaining 24 compounds (677 data points) were reserved exclusively for testing. This strategy, corresponding to approximately 30% of the data used for validation and ensures that the model is evaluated on entirely unseen compounds, providing a stringent assessment of its predictive capability and generalization performance.

2.2 Reference Diffusion Coefficient and Dimensionless Scaling

Rosenfeld[24,25] proposed that the transport properties, when expressed in dimensionless form, can be correlated to the residual entropy (s^{res}). In the case of self-diffusion, this relationship can be written as:

$$\ln D^* = \ln\left(\frac{D}{D^{\mathrm{ref}}}\right) = f(s^{\mathrm{res}}) \quad (1)$$

where D^*, D and D^{ref} denote dimensionless, real, and reference self-diffusion coefficients, respectively. Residual entropy represents the deviation from ideal-gas behavior and reflects the reduction in configurational freedom due to intermolecular interactions, making it a compact thermodynamic descriptor of structural disorder of fluid structure and non-ideality.[26] This concept has been extensively validated for simple fluids, where diffusion, viscosity, and thermal conductivity follow quasi-universal relationships, when expressed in reduced form as functions of excess entropy.[27]

However, entropy-scaling relationships are not universally exact. Deviations have been reported for complex fluids, mixtures, strongly associating systems, and systems exhibiting thermodynamic anomalies, where the coupling between the molecular structure and transport becomes more intricate. In particular, near critical regions, at elevated densities, or in systems with strong directional interactions, local structuring and fluctuations may limit the applicability of simple entropy-based scaling.[27]

In the present work, the Rosenfeld scaling concept is employed as a physically motivated normalization framework for the self-diffusion coefficient, while the final nonlinear relationship is learned using an ANN that incorporates both thermodynamic variables and molecular descriptors.

Several formulations are suggested in the literature for defining the reference self-diffusion coefficient, including those proposed by Rosenfeld,[24] Chapman-Enskog,[28,29] and Bretonnet.[30] In addition, various empirical relationships are reported to correlate dimensionless self-diffusion (D^*) with residual entropy.[31–33] However, these correlations are typically parameterized for specific compounds, and their predictive capability is restricted to the systems studied, limiting their applicability to broader chemical systems.

In this work, a generalized predictive model is developed to estimate self-diffusion coefficients for a wide variety of compounds over extended temperature and pressure ranges. First, the Rosenfeld[24] scaling approach is adopted to define the reference for self-diffusion coefficient:

$$D^{\mathrm{ref}} = \hat{\rho}^{-\frac{1}{3}}\sqrt{\frac{RT}{M}} \quad (2)$$

View Article Online
DOI: 10.1039/D6CP01425A

ARTICLE

Table 1. Summary of experimental self-diffusion coefficient datasets employed for ANN model training and validation.

No	Name	CAS no	No of data	T/K	P/bar	Ref.
			train data			
1	methane	74-82-8	117	93–454	1–898.3	34–41
2	ethane	74-84-0	54	136–454	43.6–978.5	34,36,37,39
3	propane	74-98-6	21	112–453	14.7–500	34,39
4	butane	106-97-8	9	150–451	50–500	34
5	pentane	109-66-0	29	174–450	1–981	34,42–44
6	heptane	142-82-5	47	186.1–360.6	1–981	34,42–44
7	octane	111-65-9	36	248.14–383.7	1–998	34,44,45
8	decane	124-18-5	39	247.86–448	1–750	34,42–45
9	undecane	1120-21-4	17	293–353	1–981	34,45
10	dodecane	112-40-3	33	268.75–434.7	1–510	34,45,46
11	tetradecane	629-59-4	23	279.36–443	1–750	34,45,46
12	pentadecane	629-62-9	16	288.16–353	1–981	34,45
13	hexadecane	544-76-3	31	292.68–472.5	1–996	34,45
14	heptadecane	629-78-7	14	303–353	1–981	34
15	octadecane	593-45-3	11	301.86–425.8	1	34
16	eicosane	112-95-8	6	323.16–443.7	1	34
17	tetracosane	646-31-1	10	322.16–423.7	1	34
18	2-methylpentane	107-83-5	5	200–308.2	1	34
19	3-methylpentane	96-14-0	8	200–313.2	1	34,47
20	2,3-dimethylbutane	79-29-8	11	175.48–453	1–500	34
21	2,2-dimethylbutane	75-83-2	18	262.37–450	1–600	34,47
22	cyclopentane	287-92-3	12	273.16–328	1–750	34
23	cyclohexane	110-82-7	84	281.7–393.2	1–900	34,44–46,48,49
24	cycloheptane	291-64-5	7	288.16–348.8	1	34
25	ethanol	64-17-5	54	173–437	1–931	34,45,50–53
26	1-propanol	71-23-8	32	212–441	1–750	34,45,53
27	1-butanol	71-36-3	11	268.16–353.2	1	34,45,54,55
28	2-pentanol	6032-29-7	46	237.1–483.1	50–500	34,56
29	3-pentanol	584-02-1	52	249.7–474.5	50–500	34,56
30	1-pentanol	71-41-0	22	213–428.6	1–500	34,56
31	1-hexanol	111-27-3	5	278.16–338.2	1	34,55
32	1-octanol	111-87-5	9	288.16–343.2	1	34
33	glycerol	56-81-5	35	296.8574–435.1	1	57–60
34	benzene	71-43-2	146	279.96–373.2	1–980.7	34,44,48–50,61–64
35	toluene	108-88-3	61	175.4286–729.2	1–997	34,61
36	*o*-terphenyl	84-15-1	16	328.16–438.2	1	65

View Article Online
DOI: 10.1039/D6CP01425A

37	acetone	67-64-1	20	182.86–323.2	1	34
38	water	7732-18-5	264	273–973.2	1–976	34,66,67
39	tetrahydrofuran	109-99-9	7	180.56–308.2	1	34
40	ethylene	74-85-1	62	123.15–348.2	20.4–810.6	34,36
41	carbon disulfide	75-15-0	10	268.2–313.2	1–811	34
42	1,2-dichloroethane	107-06-2	12	278.15–298.2	1–2795	68
43	acetonitrile	75-05-8	64	238.2–343.2	1–3036	69
			test data			
44	ammonia	7664-41-7	18	199.2–473	1–750	34,70,71
45	2-propanol	67-63-0	10	263–360	1–500	72
46	methanol	67-56-1	43	157–453	1–981	50–52,72
47	tridecane	629-50-5	18	288.2–353	1–981	34,45
48	nonane	111-84-2	38	235.5–403.2	1–990	34,44,45
49	hexane	110-54-3	72	188.5–443	1–998	34,44,47,61,72,73
50	isopentane	78-78-4	24	298–328	1–2000	74
51	bromoform	75-25-2	8	283.2–343.2	1	75
52	*N,N*-dimethylacetamide	127-19-5	35	255–468	1–2000	76
53	dimethyl ether	115-10-6	40	184.5–458	500–2000	77
54	diiodomethane	75-11-6	24	285.7–351.3	1	78
55	dichloromethane	75-09-2	36	186–406	1–2000	79
56	chloroform	67-66-3	40	217–397	1–1500	79
57	carbon tetrachloride	56-23-5	3	313.2–333.2	1	80
58	chlorotrifluoromethane	75-72-9	60	133–433	250–2000	81
59	bromotrifluoromethane	75-63-8	59	141–432	250–2000	81
60	fluorobenzene	462-06-6	13	240–360	1	82
61	iodobenzene	591-50-4	15	330–440	1	82
62	bromobenzene	108-86-1	18	250–420	1	82
63	trimethylamine	75-50-3	44	174–423	100–2000	83
64	*N,N*-dimethylformamide	68-12-2	36	243–448	1–2000	84
65	propylene glycol	57-55-6	4	304–318	1	85
66	dibromomethane	74-95-3	8	285–363	1	86
67	1,2-dibromoethane	106-93-4	11	285–400	1	86

where $\hat{\rho}$ denotes number density of molecules (m^{-3}), T (K) is temperature, M (kg $kmol^{-1}$) stand for molar mass, and R (8314.46 J K^{-1} $kmol^{-1}$) is the universal gas constant. The reference diffusion coefficient has dimensions of diffusivity. The term $\hat{\rho}^{-\frac{1}{3}}$ represents a characteristic molecular length scale, while $\sqrt{\mathrm{RT/M}}$ represents a characteristic molecular thermal velocity. Their product therefore gives units of m^2 s^{-1}, consistent with a diffusion coefficient. In this work, D^{ref} is not used as an independent model for diffusion. Instead, it is used to nondimensionalize the experimental self-diffusion coefficient. The effects of molecular structure are subsequently incorporated through PCP-SAFT-derived thermodynamic properties and COSMO-SAC molecular descriptors in the ANN framework.

Having the reference parameter, we present in the subsequent sections, the methodology to develop a generalized model for predicting the dimensionless self-diffusion coefficient (D^*), following the procedure for obtaining the actual self-diffusion coefficient (D) using the reference value D^{ref}.[24]

2.3 COSMO-SAC sigma-profile and molecular descriptor extraction

The COSMO-SAC sigma-profiles of all compounds considered in this study were obtained from the literature and used to derive the molecular descriptors for developing a correlation with the dimensionless self-diffusion coefficient (D^*).[87] Each sigma-profile describes the distribution of surface charge density (e $Å^{-2}$) over 51 discrete intervals spanning from−0.025 e $Å^{-2}$ to +0.025 e $Å^{-2}$. For each bin, COSMO-SAC provides the associated surface area related to each surface charge density, denoted as A_j (in $Å^2$). The total surface area of each molecule (A_{tot}) is obtained by summing the contributions from all bins (j):

View Article Online DOI: 10.1039/D6CP01425A

$$A_{\text{tot}} = \sum A_j(\sigma) \quad (3)$$

In addition, key molecular descriptors can be derived using sigma-moments defined by eqns (4-8):

$$M_1 = \sum A_j(\sigma)\sigma_j \quad (4)$$

$$M_2 = \sum A_j(\sigma){\sigma_j}^2 \quad (5)$$

$$M_3 = \sum A_j(\sigma){\sigma_j}^3 \quad (6)$$

$$M_1^{\text{HBD}} = \sum A_j(\sigma) max(0, -\sigma - \sigma^{\text{HB}}) \quad (7)$$

$$M_1^{\text{HBA}} = \sum A_j(\sigma) max(0, \sigma - \sigma^{\text{HB}}) \quad (8)$$

These moments characterize different aspects of the surface charge distribution: M_1 gives the net surface charge, M_2 reflects polarity, and M_3 describes asymmetry. However, during model development, M_3 was found to have a negligible influence on the prediction of self-diffusion coefficients. The sensitivity analysis on the experimental data showed that M_3 contributes only approximately 3% to the prediction of the self-diffusion coefficient, which is significantly lower than the contributions from the other descriptors considered in this work; therefore M_3 was excluded from input variables.

M_1^{HBD} and M_1^{HBA} quantify hydrogen-bond donor and acceptor strengths, respectively. The threshold value σ^{HB} = 0.008 e·Å^{-2} separates nonpolar and polar surface segments.[17,18]

Table 2. Definition of input variables used in the ANN model.

	Input variables	Description
COSMO-SAC parameters	A_{tot}	Total surface area of each molecule (Å²)
	M_1	Index of net surface charge (e)
	M_2	Index of polarity (e2 Å-2)
	M_1^{HBD}	Index of hydrogen-bond donor strength (e)
	M_1^{HBA}	Index of hydrogen-bond acceptor strength (e)
Thermodynamic properties	s^{res}/R	Dimensionless form of residual entropy
	ρ	Molar density (kmol m-3)

2.4 PCP-SAFT equation of state

Previous studies have demonstrated PCP-SAFT equation of state (EoS),[88,89] which extends the original PC-SAFT EoS[90,91] by incorporating polar interactions in addition to dispersion and association effects, thereby improving accuracy for polar compounds, is capable of accurately predicting thermodynamic properties such as density and residual entropy for a wide variety of fluids. In this work, we use PCP-SAFT to generate these properties for all compounds considered, including nonpolar, polar, and associating systems. The residual Helmholtz energy is expressed in eqn (9):[88,89]

$$a^{\text{res}} = a^{\text{hc}} + a^{\text{d}} + a^{\text{p}} + a^{\text{assoc}} \quad (9)$$

The contributions correspond to hard-chain repulsion (hc), dispersion interactions (d), polar interactions (p), and association (assoc), respectively. For non-associating, nonpolar compounds, the model requires three parameters: segment number (m), segment diameter (σ), and dispersion energy (ε). Associating fluids require two additional parameters, the association energy ($\varepsilon^{\mathcal{AB}}$) and association volume ($k^{\mathcal{AB}}$), while polar compounds are characterized using the dipole moment (μ^{D}).

The dimensionless form of residual entropy is calculated from the Helmholtz energy as:[90]

$$\frac{s^{\text{res}}(P,T)}{R} = -T\left[\left(\frac{\partial a^{\text{res}}}{\partial T}\right)_{\rho} + \frac{a^{\text{res}}}{T}\right] + \ln(Z) \quad (10)$$

where Z is the compressibility factor obtained from PCP-SAFT.

In addition to residual entropy, density (ρ) is also included as an input variable which is calculated from the EoS. The inclusion of density as an input variable is because of the dependency of molecular diffusion on fluid density.

2.5 ANN structure and training approach

The self-diffusion coefficient was modeled using a feedforward ANN implemented in MATLAB (R2025a, Neural Network Toolbox). The model combines thermodynamic variables (dimensionless form of residual entropy and density) with COSMO-SAC–derived molecular descriptors to predict diffusion behavior. The input vector consists of seven variables (Table 2), while the output corresponds to the logarithm of the dimensionless self-diffusion coefficient (D^*).

Prior to training, all inputs and outputs were linearly scaled to the range [−1, 1], which improved numerical stability and facilitated convergence. The same scaling parameters derived from the training data were applied to the testing dataset. To identify an optimal and robust network configuration, we systematically explored different architectures, activation functions, and training algorithms. Both shallow and deep network structures were examined, including single hidden layers with 10–40 neurons and two-layer configurations with varying neuron counts.

Three activation functions of tansig, logsig, and poslin were evaluated for the hidden layers, while a linear activation was consistently used in the output layer to do the regression task. Training was conducted using multiple optimization algorithms, including Levenberg–Marquardt, scaled conjugate gradient, and Bayesian regularization. Model performances using training and independent testing datasets were assessed through statistical metrics such as the coefficient of determination (R^2),

View Article Online
DOI: 10.1039/D6CP01425A

mean absolute error (MAE), average absolute relative deviation (AARD%), and maximum absolute relative deviation. In addition, we analyzed parity plots, error distributions, and error trends across diffusion ranges, to evaluate the accuracy and robustness of the model.

3 Results and Discussion

In this section, the development and validation of the hybrid model, combining PCP-SAFT and an artificial neural network (ANN), are presented for predicting the self-diffusion coefficient. The predictive capability of the model is systematically evaluated through comparison with experimental data, along with statistical performance indicators. In addition, a sensitivity analysis is conducted to determine the relative importance of the input variables.

3.1 Model development and evaluation

A range of ANN architectures was initially constructed and trained. The resulting models were then ranked based on their predictive performance on the test dataset, using statistical criteria such as the coefficient of determination (R^2), mean absolute error (MAE), and average absolute relative deviation (AARD). The architecture corresponding to the best overall performance was selected for further analysis.

To justify the selected ANN topology, several network architectures were evaluated, and some of them are summarized in Table 3. Networks with fewer neurons, such as 8–4 and 10–5, showed higher AARD values for both training and testing datasets, indicating insufficient flexibility to capture the nonlinear relationship between the input descriptors and the dimensionless self-diffusion coefficient. Increasing the number of neurons improved the training performance; however, architectures larger than 14–7 led to reduced testing accuracy despite lower training errors, suggesting the onset of overfitting. The 14–7 architecture provided the best compromise between model complexity and generalization capability, achieving high accuracy for the training data while maintaining the lowest testing AARD among the evaluated structures. Therefore, this topology was selected as the final ANN configuration.

In addition, to evaluate the statistical robustness and generalization capability of the 14–7 architecture model, a repeated compound-wise validation procedure was performed. In this approach, the dataset was randomly partitioned multiple times at the compound level, ensuring that each testing set contained entirely unseen chemical species. The ANN model was retrained for each split, and the resulting performance metrics were statistically analyzed.

As shown in Table 4, the variation in model performance across different splits is relatively small. The AARD and R^2 values for both training and testing datasets exhibit limited standard deviations, and the corresponding 95% confidence intervals remain narrow. This indicates that the predictive performance of the model is not sensitive to the specific selection of compounds in the training or testing sets. Overall, this repeated validation analysis provides strong evidence that the hybrid PCP-SAFT+ANN model exhibits both robustness and generalizability.

Table 3. Effect of ANN architecture on model performance.

Hidden-layer neurons	Train AARD%	Test AARD%	Comment
8–4	15.36	21.84	Underfitting
10–5	11.92	19.96	Improved but less accurate
12–6	9.84	17.51	Good performance
14–7	8.89	15.89	Selected model
16–8	7.45	18.54	Slight overfitting
20–10	6.43	22.11	Overfitting

Table 4. Statistical robustness of the model based on repeated compound-wise data splits, reported as mean values ± 95% confidence intervals

Dataset	Training	Testing
R^2	0.9937 ± 0.0038	0.9763 ± 0.0131
ARAD	8.89 ± 1.07	15.89 ± 2.08

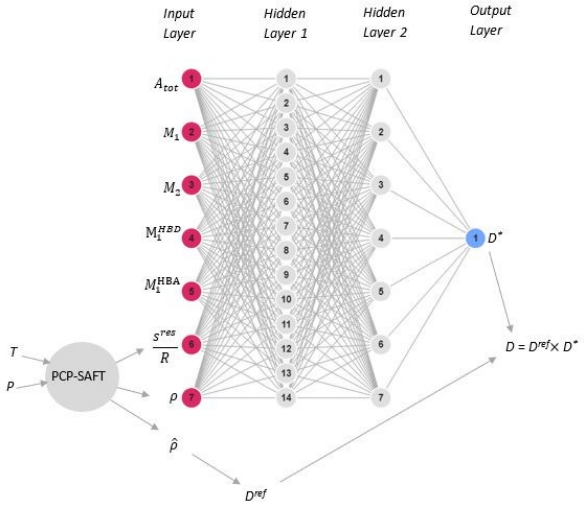

Fig. 1 Schematic representation of the hybrid PCP-SAFT+ANN framework used for self-diffusion coefficient prediction.

Table 5 presents statistical evaluation of the 14-7 architecture model using training and independent testing datasets. The results clearly demonstrate the high predictive accuracy of the proposed hybrid framework. For the training dataset (1586 data points), the model achieves a coefficient of determination of R^2 = 0.9937 along with an MAE of 9.23×10^{-10} m^2 s^{-1} and an AARD of 8.89%. These low error values, combined with the high R^2, indicate that the ANN successfully captures the complex and nonlinear dependence of the self-diffusion coefficient on the selected input variables. The predictive performance remains robust when evaluated against the independent testing dataset (677 data points). In this case, the model yields R^2 = 0.9763, MAE = 4.32×10^{-10} m^2 s^{-1}, and AARD = 15.89%. Considering the full dataset (2263 data points), the overall performance remains consistently high, with R^2 = 0.9839. The model maintains reliable accuracy across a wide range of self-diffusion coefficients, spanning approximately five orders of magnitude

(1.89×10^{-12}–3.61×10^{-7} m^2 s^{-1}). This broad applicability demonstrates the robustness of the hybrid PCP-SAFT+ANN approach for predicting self-diffusion behavior in diverse chemical systems.

Table 5. Statistical evaluation of the 14–7 architecture model for self-diffusion coefficient prediction using training and testing datasets.

Set	No data	R^2	MAE (m^2 s^{-1})	AARD%	Max ARD%
Train	1586	0.9937	9.23×10^{-10}	8.89	54.38
Test	677	0.9763	4.32×10^{-10}	15.89	61.14
Total	2263	0.9839	7.76×10^{-10}	10.98	61.14

Fig. 1 shows the structure of the proposed hybrid framework, which integrates thermodynamic information from the PCP-SAFT EoS with a data-driven ANN model. The input layer consists of seven neurons representing key descriptors, including COSMO-SAC-derived molecular parameters, dimensionless residual entropy, and density obtained from PCP-SAFT calculations. These inputs provide both molecular-level and thermodynamic information, enabling a physically informed prediction. The proposed ANN includes two hidden layers with 14 and 7 neurons, respectively. This configuration was found to be sufficiently flexible to capture the nonlinear interactions between the descriptors without overfitting the data. The output layer contains a single neuron that predicts the dimensionless self-diffusion coefficient (D^*). The complete set of model parameters and implementation details are provided in the supplementary material.

Fig. 2 presents the parity plot comparing the predicted and experimental values of the self-diffusion coefficient (D) for both the training and testing datasets. The close alignment of the data points along the diagonal reference line ($y = x$) indicates a high level of agreement between the model predictions and experimental measurements. The model maintains a strong predictive accuracy over a broad range of self-diffusion coefficients, spanning approximately five orders of magnitude (10^{-12} to 10^{-7} m^2 s^{-1}). This wide coverage highlights the capability of the hybrid PCP-SAFT+ANN framework to reliably capture diffusion behavior for systems with significantly different transport characteristics. Data points corresponding to the training set (blue circles) are densely distributed around the parity line, confirming that the model has effectively learned the underlying nonlinear relationships between the input descriptors and the target property. More importantly, the testing dataset (green triangles), which includes compounds not involved in the training process, also follows the parity line closely. This demonstrates that the model retains its high predictive accuracy when applied to unseen data. No noticeable systematic bias or deviation is observed across the entire range of D values, further supporting the robustness and stability of the selected network architecture. Overall, the results confirm that the proposed hybrid model provides reliable and consistent predictions of self-diffusion coefficients across diverse chemical systems and thermodynamic conditions.

Fig. 3 shows the distribution of relative error residuals for both the training and testing datasets as a function of the experimental self-diffusion coefficient. The error profiles

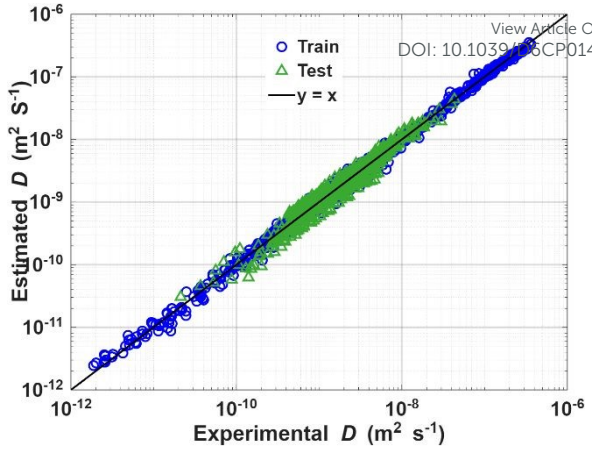

Fig. 2 Parity plot comparing estimated and experimental self-diffusion coefficients for the training (1586 points) and testing (677 points) datasets.

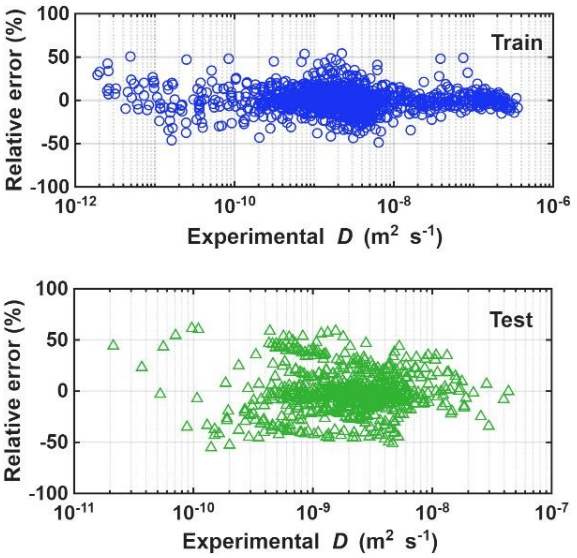

Fig. 3 Relative prediction error (%) as a function of experimental self-diffusion coefficient for the training (top) and testing (bottom) datasets.

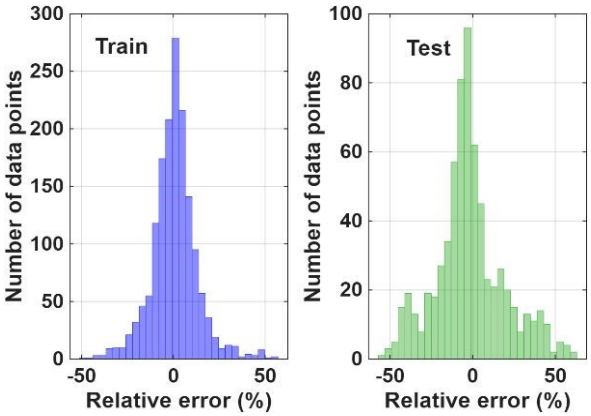

Fig. 4 Histograms of relative prediction errors (%) for the self-diffusion coefficient for the training (left) and testing (right) datasets.

provide additional insight into the consistency and reliability of the developed model. For the training dataset, the relative errors are predominantly distributed around zero across the entire range of diffusion coefficients. The narrow spread of the

data indicates that the model achieves high accuracy with minimal dispersion, confirming its ability to represent the underlying relationships within the training data. In the case of the testing dataset, a slightly broader distribution of errors is observed, which is expected for data not included during the model calibration. Nevertheless, the errors are randomly distributed around zero, and no systematic overprediction or underprediction trends are observed. This relatively low error and its random distribution indicate that the model preserves its predictive capability when applied to unseen compounds. Importantly, the absence of any noticeable bias or trend in error magnitude with respect to the diffusion coefficient suggests that the model performance is reliable across the investigated range. Overall, the results demonstrate that the hybrid PCP-SAFT+ANN model delivers unbiased and robust predictions, with no indication of overfitting and with strong generalization across diverse chemical systems and operating conditions.

Fig. 4 presents the histograms of relative prediction errors for both the training and testing datasets, providing a statistical perspective on the model accuracy and error distribution. For the training dataset, the error distribution is sharply centered around zero, with 70% falling to within ±10%, and over 90% within ±20%. This narrow and symmetric distribution confirms the high prediction accuracy of the model and demonstrates its ability to accurately represent nonlinear dependence of the self-diffusion coefficient with the selected descriptors. Only a small fraction of the data exhibits larger deviations, which are mainly associated with conditions at very low diffusion coefficients, where sensitivity to input parameters and experimental uncertainty are typically higher. For the testing dataset, the error distribution remains centered close to zero, indicating that the model predictions are essentially unbiased for unseen compounds. Despite a slightly broader spread compared to the training data, 50% of error residuals are located within ±10% and 70% within ±20%, that are practically acceptable error bounds. The symmetric shape of the histogram further confirms lack of skew relating overprediction or underprediction. Overall, these quantitative error distributions demonstrate that the hybrid PCP-SAFT+ANN model achieves both high accuracy and strong generalization capability, maintaining reliable performance across a wide range of self-diffusion coefficients and thermodynamic conditions.

Fig. 5 presents the variation of the average absolute relative deviation (AARD%) across different ranges of self-diffusion coefficient for both the training and testing datasets. Each interval corresponds to one order of magnitude of *D*, enabling a consistent assessment of model performance across the entire diffusion range. For the training dataset, the AARD shows a clear decreasing trend with increasing self-diffusion coefficient. The highest deviations are observed at the lowest diffusion range (10^{-12}–10^{-11} m^2 s^{-1}); the AARD steadily declines as *D* increases, reaching its minimum values at the highest diffusion ranges. The increase in the relative deviation observed at very low diffusion coefficients can be attributed to both numerical sensitivity and experimental uncertainty. From a mathematical standpoint, the AARD involves normalization by the experimental value; therefore, when *D* is very small, even

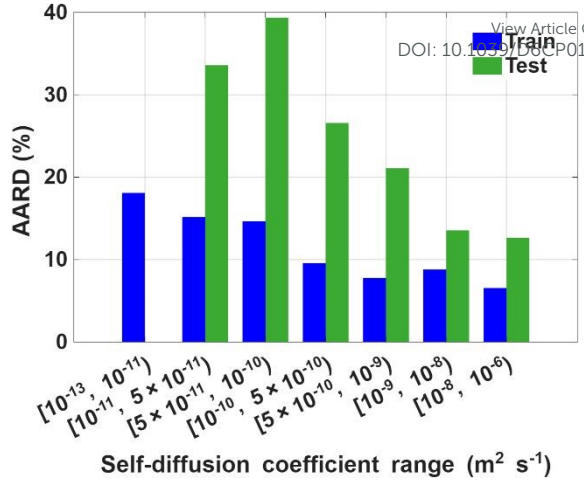

Fig. 5 Variation of AARD (%) of the ANN model in estimating self-diffusion coefficient intervals for both training and test datasets.

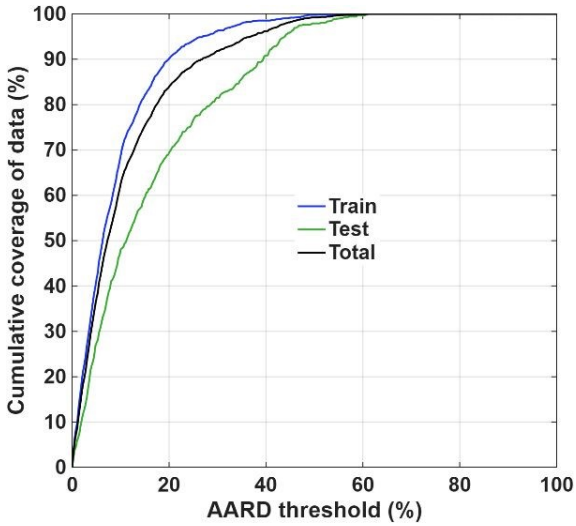

Fig. 6 Cumulative fraction of data within a given AARD threshold for the training, testing, and overall datasets.

minor absolute differences between the predicted and experimental values can result in larger relative errors.

In addition, low-diffusivity conditions typically correspond to high-density or highly structured fluid states, where molecular mobility is significantly restricted. Under such conditions, experimental measurements of self-diffusion coefficients are inherently more challenging and may be associated with higher uncertainty due to limitations in measurement techniques and sensitivity to temperature and pressure control.

From a modeling perspective, the calculation of residual entropy using PCP-SAFT also becomes more sensitive under these conditions.

Despite these challenges, the model maintains consistently low absolute accuracy across the full range of diffusion coefficients, and the observed increase in relative deviation at very low values is primarily a consequence of normalization effects and data sensitivity rather than a systematic limitation of the proposed framework. A similar trend is observed for the testing dataset, although with higher deviations as expected for unseen data. Overall, Fig. 5 confirms that the hybrid PCP-SAFT+ANN

Journal Name ARTICLE

View Article Online
DOI: 10.1039/D6CP01425A

model provides reliable predictions of the self-diffusion coefficient across six orders of magnitude range.

Fig. 6 presents the cumulative coverage curves for the training, testing, and overall datasets as a function of the AARD threshold. This representation provides a comprehensive assessment of the predictive reliability of the hybrid PCP-SAFT+ANN model across the full range of self-diffusion coefficients. For the training dataset, the curve increases sharply at low AARD thresholds, indicating that a large fraction of the data is predicted with high accuracy. 90% of the training data fall within 20% AARD, and complete coverage is achieved below about 52% AARD. This steep rise confirms the strong fitting capability of the model. For the testing dataset, the cumulative curve exhibits a more gradual increase, reflecting a greater variability as expected for unseen data. Nevertheless, the model still demonstrates solid predictive performance, with 70% of the data within 20% AARD and over 90% within 40% AARD. This behavior highlights the ability of the model to generalize effectively across different compounds and thermodynamic conditions. The curve corresponding to the full dataset lies between the training and testing curves, as expected, and reflects the overall predictive performance of the model. All three curves approach 100% coverage at AARD values below about 62%, indicating that even the largest deviations remain within acceptable bounds. Overall, this cumulative analysis confirms that the developed hybrid model provides reliable and consistent predictions of self-diffusion coefficient, with a high proportion of results falling within practically acceptable error limits for both known and unseen systems.

Table 6 presents the predictive performance of the hybrid PCP-SAFT+ANN model for different classes of compounds, categorized based on their intermolecular interaction type. This classification enables a more detailed assessment of the model capability across fluids with fundamentally different interaction mechanisms, including dispersion-dominated (nonpolar), dipolar (polar non-associating), and hydrogen-bonding (associating) systems.

For nonpolar compounds, the model achieves a high level of accuracy, with an R^2 value of 0.9931 and an AARD of 8.81%. These systems are primarily governed by dispersion interactions, which are well described by PCP-SAFT. As a result, the residual entropy and density provide a consistent representation of the thermodynamic state, enabling the ANN to accurately capture the diffusion behavior across a wide range of conditions.

Similarly, associating compounds exhibit excellent predictive performance, with the highest R^2value of 0.9959 and an AARD of 9.87%. This indicates that the hybrid framework successfully captures the effect of hydrogen-bonding interactions. The inclusion of association terms in PCP-SAFT, together with hydrogen-bond-related descriptors derived from COSMO-SAC, provides sufficient information to the ANN to capture the additional complexities introduced by specific intermolecular interactions.

In contrast, the performance for polar non-associating compounds is comparatively lower, with an R^2 value of 0.9343 and an AARD of 16.57%. This can be attributed to the more complex behavior of dipolar interactions, which are generally weaker and more sensitive to molecular orientation compared to hydrogen bonding. In such systems, the relationship between the thermodynamic properties and diffusion behavior is less direct, leading to increased variability in the data and a slightly reduced predictive accuracy. In addition, polar compounds often exhibit a wider range of molecular structures and dipole moments, which may not be fully captured by the selected descriptors.

Despite these differences, the model maintains reasonable accuracy across all categories, demonstrating its robustness and general applicability. The maximum absolute relative deviation (Max ARD) remains within a similar range for all groups, indicating that extreme deviations are not systematically associated with any specific class of compounds.

Table 6. Statistical evaluation of the hybrid PCP-SAFT+ANN model for different intermolecular interaction classes, including nonpolar, polar and associating compounds.

Set	No component	No data	R^2	MAE ($m^2 s^{-1}$)	AARD %	Max ARD%
Nonpolar	34	1108	0.9931	1.15×10^{-9}	8.81	54.38
Polar non-associating	19	550	0.9343	4.57×10^{-10}	16.57	58.72
Associating	14	605	0.9959	3.81×10^{-10}	9.87	61.14

3.2 Model capability to capture the trends in self-diffusion coefficient with temperature and pressure

Fig. 7 illustrates the pressure and temperature dependency of the self-diffusion coefficient for four representative compounds (benzene, toluene, cyclohexane, and propane) selected from the training dataset. For each system, the self-diffusion coefficient is evaluated at multiple temperature levels and over a wide range of pressures to assess the capability of the hybrid PCP-SAFT+ANN model in capturing coupled thermodynamic effects. As observed in Fig. 7, the predicted values closely follow the experimental data across all investigated conditions. The model successfully reproduces the expected physical trend, namely the decrease of the self-diffusion coefficient with increasing pressure at a given temperature. In addition, the model accurately captures the influence of temperature, where higher temperatures lead to increased diffusion coefficients due to enhanced molecular motion. The predicted curves remain smooth and physically consistent over the entire pressure range, indicating that the model provides stable and continuous predictions. The excellent agreement between predicted and experimental data, combined with the smooth curves, demonstrates the strong interpolation capability of the ANN within the training domain. Overall, the results confirm that the hybrid model effectively captures the combined effects of pressure and temperature on self-diffusion behavior.

View Article Online
DOI: 10.1039/D6CP01425A

ARTICLE

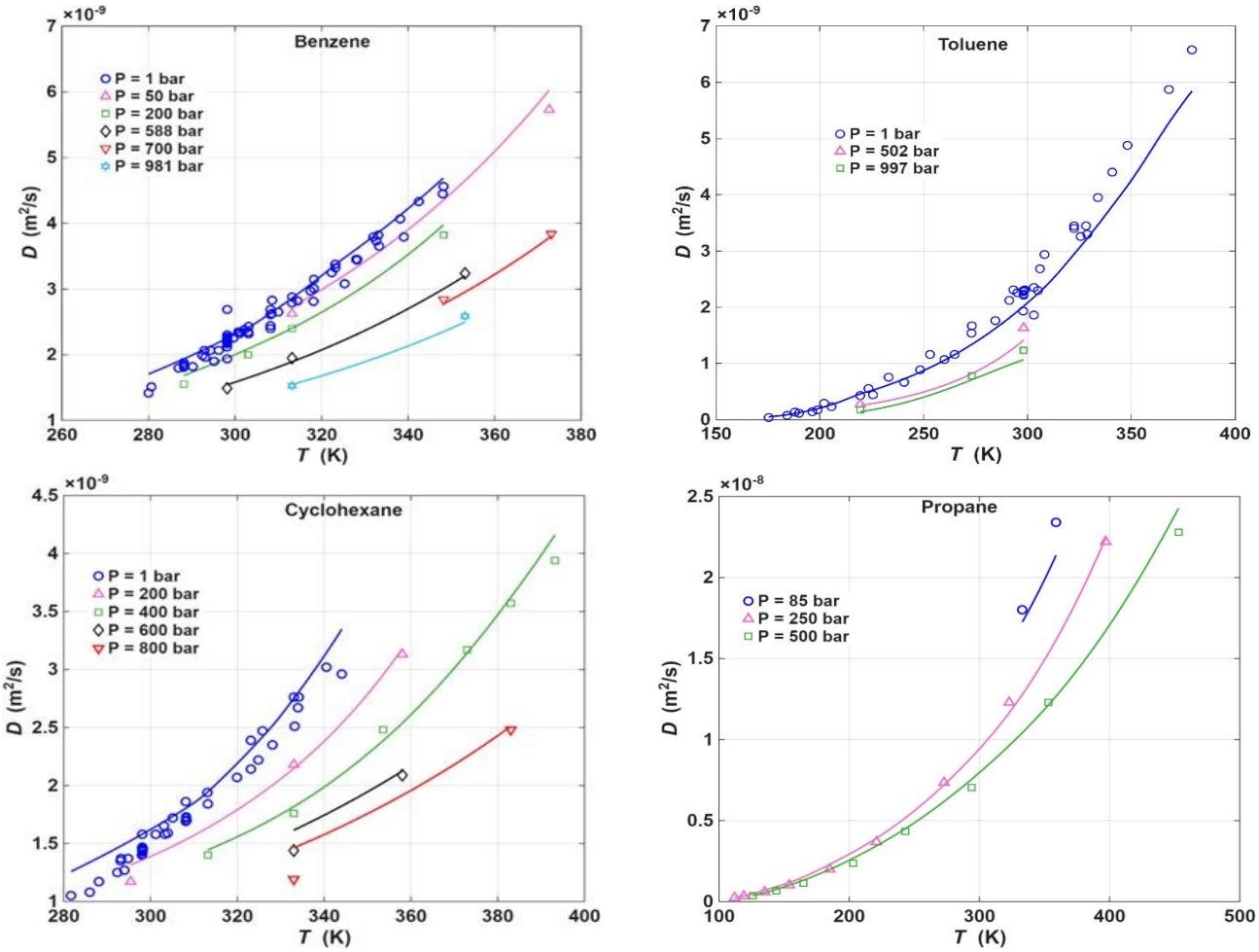

Fig. 7 Comparison of estimated and experimental self-diffusion coefficients for four representative compounds from the training dataset across a wide range of temperatures and pressures. Lines represent predictions, while symbols denote literature data: benzene,[34,44,48–50,61–64] toluene,[34,61] cyclohexane,[34,44–46,48,49] and propane.[34,39]

Fig. 8 shows the predictive performances of the developed hybrid PCP-SAFT+ANN model for four representative compounds (ammonia, methanol, isopentane, and chloroform) that were not included in the training dataset. The comparison between predicted values and reported experimental data validate excellent model's generalization capability across different chemical systems and thermodynamic conditions. As can be seen in Fig. 8, the ANN predictions closely follow the reported data over the full range of temperatures and pressures. The model accurately reproduces both the magnitude and the variation of self-diffusion coefficient, with predicted curves closely following the experimental data points. The agreement is consistent across different pressure levels, and the model successfully captures the separation between isobars at each temperature. The deviations between the predicted and experimental values are generally small and show no systematic trend. Note that all systems presented in Fig. 8 were not included in the training dataset, making this validation particularly rigorous. Overall, Fig. 8 demonstrates that the hybrid PCP-SAFT+ANN model provides accurate and reliable predictions for unseen compounds over a wide range of thermodynamic conditions, highlighting its strong generalization capability and suitability for practical applications.

3.3 Relative importance analysis of input variables

To quantify the relative importances of the input variables in predicting the self-diffusion coefficient, a correlation-based sensitivity analysis was performed. This approach is particularly useful for ANN models, where the contribution of individual

View Article Online
DOI: 10.1039/D6CP01425A

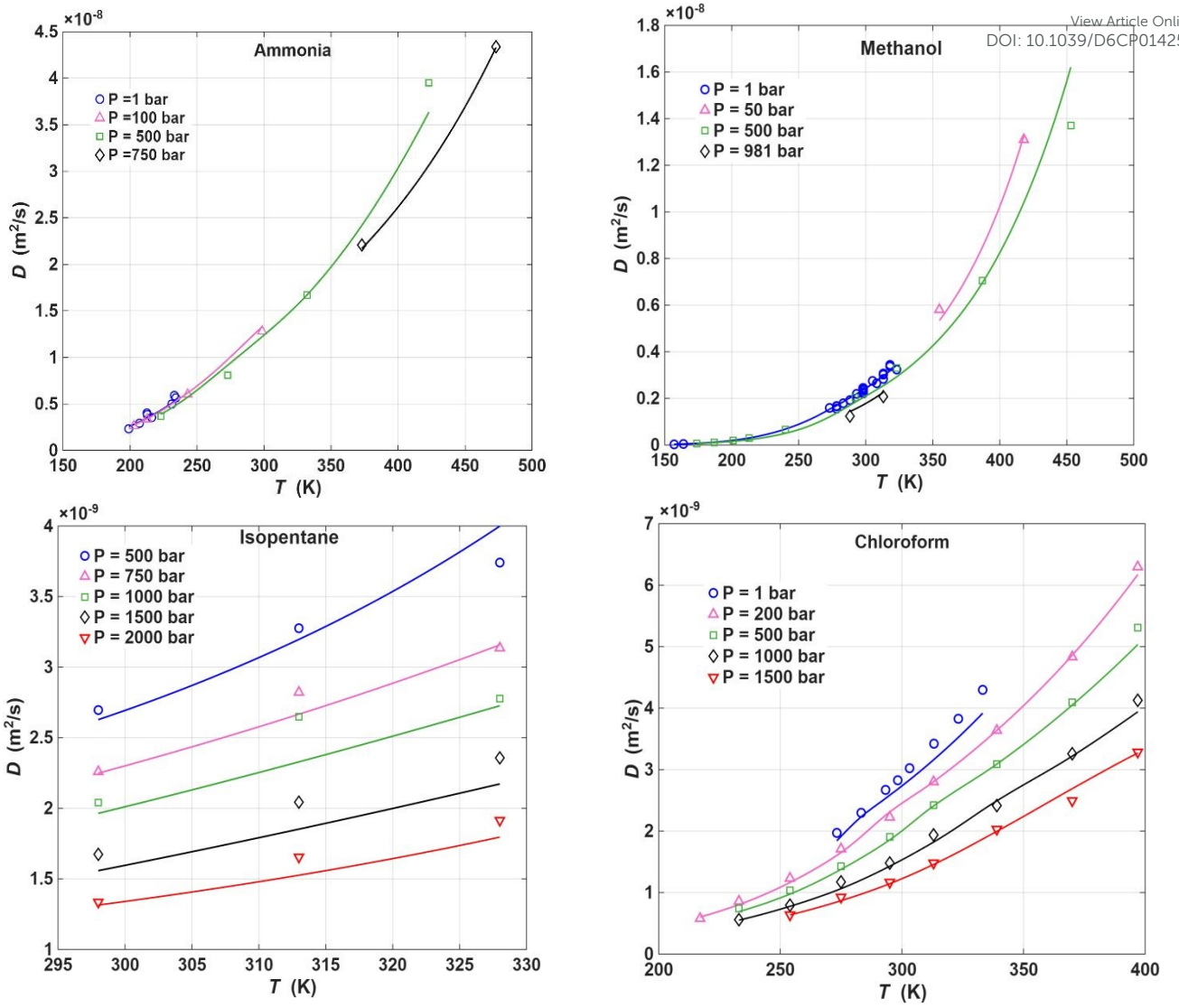

Fig. 8 Comparison of estimated and experimental self-diffusion coefficients for four representative compounds from the test dataset across a wide range of temperatures and pressures: ammonia,[34,70,71] methanol,[30–32,72] isopentane,[74] and chloroform.[75]

inputs cannot be directly inferred from the internal network structure. In this study, the Pearson correlation coefficient between each input variable and the predicted self-diffusion coefficient was calculated for the entire dataset. The absolute values of diffusion coefficients were then normalized by taking the sum of all absolute correlations to obtain the relative sensitivity for each input. This normalization enables a direct comparison between the contributions from each variable to the model output. The results are summarized in Fig. 9, where the bar heights indicate their relative influence on the predicted self-diffusion coefficient. As shown in Fig. 9, the most influential variable is dimensionless residual entropy $\left(\frac{s^{\mathrm{res}}}{R}\right)$, with a relative contribution of approximately 35%, indicating that intermolecular interactions play the most dominant role in determining diffusion behavior. This is followed by A_{tot} (17%); the first sigma-profile moment, M_1 (14%); and the second sigma-profile moment, M_2 (12%); all of which make notable contribution to the model prediction. The hydrogen-bond donor descriptor ($M_1{}^{\mathrm{HBD}}$) also shows a meaningful effect, with a relative importance of approximately 9%, while density (ρ) contributes around 8%, and the hydrogen-bond acceptor descriptor ($M_1{}^{\mathrm{HBA}}$) exhibits the smallest contribution, at approximately 5%. These results highlight that thermodynamic properties and molecular surface characteristics play key roles in governing the molecular diffusion. Moderate contributions are observed for M_1 (14%) and M_2 (12%), suggesting that molecular surface descriptors still play relevant, though secondary roles in influencing diffusion. Overall, the sensitivity analysis indicates that the ANN model relies primarily on residual entropy, while molecular descriptors also contribute significantly.

Table 7 presents the Pearson correlation matrix for the seven input variables used in the ANN model. The results show that most descriptors exhibit weak to moderate correlations, while some COSMO-SAC-derived descriptors show stronger correlations with each other. For example, M_2 is strongly correlated with M_1^{HBD} and M_1^{HBA}, with correlation coefficients of 0.80 and 0.68, respectively. This is expected because these descriptors are all derived from the sigma-profile and represent

View Article Online
DOI: 10.1039/D6CP01125A

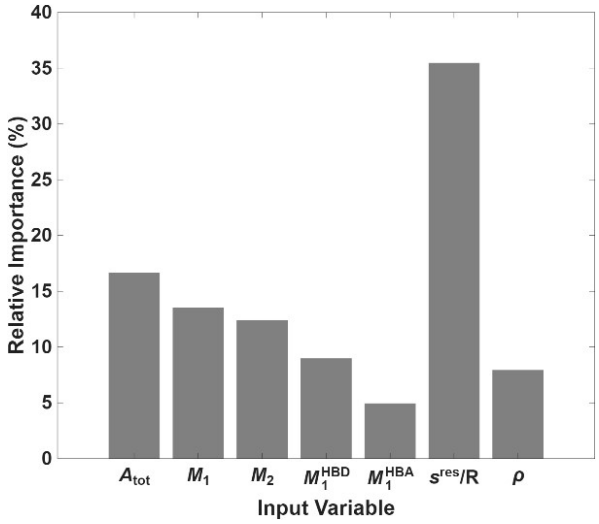

Fig. 9 Contribution of input variables to the ANN predictions of self-diffusion coefficients, evaluated via permutation-based sensitivity analysis.

related aspects of molecular surface charge distribution and hydrogen-bonding tendency.

A strong correlation is also observed between $M_1^{\mathrm{HBA}}$and density, with a correlation coefficient of 0.77. This suggests that compounds with stronger hydrogen-bond acceptor characteristics in the present dataset tend to be associated with higher-density conditions or molecular classes. However, having correlations among the inputs does not necessarily imply redundancy, because for example, density is a thermodynamic-state variable, while $M_1^{\mathrm{HBA}}$is a molecular descriptor.

Notably, the dimensionless residual entropy, s^{res}/R, shows weak to moderate correlations with most COSMO-SAC descriptors. Its strongest correlations are with A_{tot} and M_1, with coefficients of −0.59 and 0.53, respectively, while its correlations with M_2, M_1^{HBD}, M_1^{HBA}, and density are relatively weak.

To further evaluate the contribution of residual entropy to the predictive capability of the model, an ablation study was performed in which the dimensionless residual entropy (s^{res}/R) was removed from the input set and the ANN was retrained using the same training/testing protocol.

Table 7. Pearson correlation coefficients between the input variables

	A_{tot}	M_1	M_2	M_1^{HBD}	M_1^{HBA}	s^{res}/R	ρ
A_{tot}	1.00	-0.68	-0.33	-0.35	-0.40	-0.59	-0.53
M_1	-0.68	1.00	0.12	0.06	0.01	0.53	0.12
M_2	-0.33	0.12	1.00	0.80	0.68	-0.23	0.55
M_1^{HBD}	-0.35	0.06	0.80	1.00	0.64	-0.20	0.54
M_1^{HBA}	-0.40	0.01	0.68	0.64	1.00	-0.09	0.77
s^{res}/R	-0.59	0.53	-0.23	-0.20	-0.09	1.00	0.12
ρ	-0.53	0.12	0.55	0.54	0.77	0.12	1.00

As seen in Table 8, the resulting model exhibited a noticeable deterioration in predictive performance, particularly for the independent testing dataset, with increased AARD values and reduced R^2. This result confirms that residual entropy provides essential thermodynamic information that cannot be fully replaced by the remaining molecular descriptors and density alone.

Table 8. Evaluation of the contribution of residual entropy to the predictive capability of the model.

Model	Train		Test	
	R^2	AARD%	R^2	AARD%
With s^{res}/R	0.9937	8.89	0.9763	15.89
Without s^{res}/R	0.9491	29.60	0.8390	51.69

4 Conclusions

In this work, a hybrid modeling framework integrating the PCP-SAFT model with an ANN was developed for the prediction of self-diffusion coefficient across a wide range of compounds and thermodynamic conditions. The model was trained and evaluated using a comprehensive dataset comprising 2263 experimental self-diffusion coefficient data points for 67 compounds from multiple chemical families, covering temperatures from 93.0 to 973.2 K and pressures up to 3036 bar. The considered dataset spans from 1.89×10^{-12} to 3.61×10^{-7} m^2 s^{-1}, corresponding to approximately five orders of magnitude, ensuring a broad representation of molecular transport behavior.

Molecular descriptors derived from COSMO-SAC sigma-profiles, including total surface area, polarity-related parameters, and hydrogen-bonding contributions, were employed alongside thermodynamic properties (dimensionless form of residual entropy and density) calculated using PCP-SAFT. These inputs enabled the ANN model to capture the underlying relationships governing diffusion behavior.

To ensure the model robustness and generalization capability, the dataset was partitioned on a compound basis, with 45 compounds (1586 data points) used for training and 24 compounds (677 data points) reserved exclusively for independent testing. A systematic exploration of ANN architectures, activation functions, and training algorithms was conducted to identify the optimal model configuration. The selected architecture, consisting of two hidden layers with 14 and 7 neurons, demonstrated excellent predictive performance, achieving R^2 values of 0.9937 and 0.9763 for the training and testing datasets, respectively, along with AARD values of 8.89% and 15.89%. The largest relative deviations were observed at very low diffusion coefficients, where small absolute differences lead to amplified percentage errors.

Sensitivity analysis based on Pearson correlation coefficients indicates that the dimensionless residual entropy is the most influential variable in predicting self-diffusion coefficients, followed by total surface area and polarity-related descriptors. In contrast, hydrogen-bonding descriptors and density exhibit comparatively lower contributions. This highlights the dominant role of thermodynamic state representation, particularly residual entropy, in governing diffusion behavior. Such a finding underscores the effectiveness of integrating PCP-

SAFT-derived thermodynamic properties with molecular descriptors, providing a robust basis for predictive modeling. The strong predictive capability of the model highlights the effectiveness of integrating physically grounded thermodynamic inputs with data-driven approaches. The proposed PCP-SAFT+ANN framework provides a reliable and generalizable tool for estimating self-diffusion coefficients across a wide range of compounds and conditions. This approach can support process design, transport modeling, and simulation tasks where accurate diffusion data are required. Future work may extend this framework to multicomponent systems and incorporate it into process simulation platforms.

Author contributions

Aliakbar Roosta: conceptualization, data collection, programming, analysis, writing – review and editing. Nima Rezaei: conceptualization, supervision, writing – review and editing. Hamid Godini: conceptualization, supervision, writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

All supplementary materials related to this study are accessible online to support the reproducibility and promote transparency of the proposed model. These materials include:

- dimensionless_self_diffusion_ANN_2026.mat — Contains the trained ANN model, input standardization parameters, and model configuration metadata.
- dimensionless_self_diffusion_Predictor.m — MATLAB script that allows users to predict the self-diffusion coefficient of chemicals by entering COSMO-SAC–derived molecular descriptors, dimensionless form of residual entropy, and molar density. The script uses the trained ANN model.
- COSMO_SAC_derived_molecular_descriptors.xlsx — Contains the COSMO-SAC–derived molecular descriptors for 69 chemicals.

Acknowledgements

We acknowledge the funding received from the Research Council of Finland Academy Project Funding (3545438) that enabled this research.

Notes and references

1 B. A. Johnson, A. T. Castner, H. Agarwala and S. Ott, Beyond diffusion: ion and electron migration contribute to charge transport in redox-conducting metal–organic frameworks, *Chem. Sci.*, 2025, **16**, 5214–5222.

2 M. Mohammadi, M. Zirrahi and H. Hassanzadeh, An Analytical Model for Estimation of the Self-Diffusion Coefficient and Adsorption Kinetics of Surfactants Using Dynamic Interfacial Tension Measurements, *J. Phys. Chem. B*, 2020, **124**, 3206–3213.

3 D. S. Grebenkov and D. Krapf, Steady-state reaction rate of diffusion-controlled reactions in sheets, *J. Chem. Phys.*, 2018, 149, 064117. DOI:10.1063/1.5041074.

4 A. Szczęsna-Chrzan, M. Vogler, P. Yan, G. Z. Żukowska, C. Wölke, A. Ostrowska, S. Szymańska, M. Marcinek, M. Winter, I. Cekic-Laskovic, W. Wieczorek and H. S. Stein, Ionic conductivity, viscosity, and self-diffusion coefficients of novel imidazole salts for lithium-ion battery electrolytes, *J. Mater. Chem. A Mater.*, 2023, **11**, 13483–13492.

5 I. B. Obot, A. A. Bahraq and A. H. Alamri, Density functional theory and molecular dynamics simulation of the corrosive particle diffusion in pyrimidine and its derivatives films, *Comput. Mater. Sci.*, 2022, **210**, 111428.

6 B. Zhang, X. Li, J. Zhang, J. Wang and H. Jin, Study on the self-diffusion coefficients of binary mixtures of supercritical water and H2, CO, CO2, CH4 confined in carbon nanotubes, *Water Res.*, 2025, **283**, 123856.

7 J. Busch and D. Paschek, An OrthoBoXY-method for various alternative box geometries, *Physical Chemistry Chemical Physics*, 2024, **26**, 2907–2914.

8 M. A. Hunter, B. Demir, C. F. Petersen and D. J. Searles, New Framework for Computing a General Local Self-Diffusion Coefficient Using Statistical Mechanics, *J. Chem. Theory Comput.*, 2022, **18**, 3357–3363.

9 P. Ghesquière, T. Mineva, D. Talbi, P. Theulé, J. A. Noble and T. Chiavassa, Diffusion of molecules in the bulk of a low density amorphous ice from molecular dynamics simulations, *Physical Chemistry Chemical Physics*, 2015, **17**, 11455–11468.

10 R. Kokubu, S. Inasawa and H. Ohashi, Validation of Shell-like Free Volume Model for Self-Diffusion Coefficients in Polymer–Solvent System with Practical Parameter Determination Method, *Ind. Eng. Chem. Res.*, 2025, **64**, 4596–4603.

11 Z. Zuo, X. Lu and X. Ji, Modeling Self-Diffusion Coefficient and Viscosity of Chain-like Fluids Based on ePC-SAFT, *J. Chem. Eng. Data*, 2024, **69**, 348–362.

12 Y. Wei, Z. Dai, Y. Dong, A. Filippov, X. Ji, A. Laaksonen, F. U. Shah, R. An and H. Fuchs, Molecular interactions of ionic liquids with SiO_2 surfaces determined from colloid probe

atomic force microscopy, *Physical Chemistry Chemical Physics*, 2022, **24**, 12808–12815.

13 F. Zeng, R. Wan, Y. Xiao, F. Song, C. Peng and H. Liu, Predicting the Self-Diffusion Coefficient of Liquids Based on Backpropagation Artificial Neural Network: A Quantitative Structure–Property Relationship Study, *Ind. Eng. Chem. Res.*, 2022, **61**, 17697–17706.

14 J. P. Allers, J. A. Harvey, F. H. Garzon and T. M. Alam, Machine learning prediction of self-diffusion in Lennard-Jones fluids, *J. Chem. Phys.*, 2020, **153**, 034102 DOI:10.1063/5.0011512.

15 C. J. Leverant, J. A. Greathouse, J. A. Harvey and T. M. Alam, Machine Learning Predictions of Simulated Self-Diffusion Coefficients for Bulk and Confined Pure Liquids, *J. Chem. Theory Comput.*, 2023, **19**, 3054–3062.

16 J. P. Allers, F. H. Garzon and T. M. Alam, Artificial neural network prediction of self-diffusion in pure compounds over multiple phase regimes, *Physical Chemistry Chemical Physics*, 2021, **23**, 4615–4623.

17 Andreas. Klamt, *COSMO-RS : from quantum chemistry to fluid phase thermodynamics and drug design*, Elsevier, 2005.

18 A. Klamt, The COSMO and COSMO-RS solvation models, *WIREs Computational Molecular Science*, 2011, **1**, 699–709.

19 T. Nevolianis, R. A. Ahmed, A. Hellweg, M. Diedenhofen and K. Leonhard, Blind prediction of toluene/water partition coefficients using COSMO-RS: results from the SAMPL9 challenge, *Physical Chemistry Chemical Physics*, 2023, **25**, 31683–31691.

20 G. Chen, Z. Song and Z. Qi, Transformer-convolutional neural network for surface charge density profile prediction: Enabling high-throughput solvent screening with COSMO-SAC, *Chem. Eng. Sci.*, 2021, **246**, 117002.

21 N. Mac Fhionnlaoich, J. Zeglinski, M. Simon, B. Wood, S. Davin and B. Glennon, A hybrid approach to aqueous solubility prediction using COSMO-RS and machine learning, *Chemical Engineering Research and Design*, 2024, **209**, 67–71.

22 J. J. Suárez, I. Medina and J. L. Bueno, Diffusion coefficients in supercritical fluids: available data and graphical correlations, *Fluid Phase Equilib.*, 1998, **153**, 167–212.

23 P. N. Bartlett, D. A. Cook, M. W. George, A. L. Hector, J. Ke, W. Levason, G. Reid, D. C. Smith and W. Zhang, Electrodeposition from supercritical fluids, *Physical Chemistry Chemical Physics*, 2014, **16**, 9202.

24 Y. Rosenfeld, Relation between the transport coefficients and the internal entropy of simple systems, *Phys. Rev. A (Coll. Park).*, 1977, **15**, 2545–2549.

25 Y. Rosenfeld, A quasi-universal scaling law for atomic transport in simple fluids, *Journal of Physics: Condensed Matter*, 1999, **11**, 5415–5427.

26 I. H. Bell, J. C. Dyre and T. S. Ingebrigtsen, Excess-entropy scaling in supercooled binary mixtures, *Nat. Commun.*, 2020, **11**, 4300.

27 J. C. Dyre, Perspective: Excess-entropy scaling, *J. Chem. Phys.*, 2018, **149**, 210901 DOI:10.1063/1.5055064.

28 S. Chapman and T. G. Cowling, *The Mathematical Theory Of Nonuniform Gases*, Cambridge At The University Press, 3rd edn., 1970.

29 J. O. Hirschfelder, C. F. Curtiss and R. B. Bird, *The Molecular Theory of Gases and Liquids*, John Wiley & Sons, Inc, New York, 1964.

30 J.-L. Bretonnet, Self-diffusion coefficient of dense fluids from the pair correlation function, *J. Chem. Phys.*, 2002, **117**, 9370–9373.

31 M. Hopp, J. Mele and J. Gross, Self-Diffusion Coefficients from Entropy Scaling Using the PCP-SAFT Equation of State, *Ind. Eng. Chem. Res.*, 2018, **57**, 12942–12950.

32 A. Dehlouz, J.-N. Jaubert, G. Galliero, M. Bonnissel and R. Privat, Entropy Scaling-Based Correlation for Estimating the Self-Diffusion Coefficients of Pure Fluids, *Ind. Eng. Chem. Res.*, 2022, **61**, 14033–14050.

33 F. J. Carmona Esteva, Y. Zhang, K. Duncheskie, E. J. Maginn and Y. J. Colón, Excess entropy scaling explains the enhanced dynamics of the ionic liquid 1-ethyl-3-methylimidazolium chloride in external electric fields, *Physical Chemistry Chemical Physics*, 2026, **28**, 353–364.

34 O. Suárez-Iglesias, I. Medina, M. de los Á. Sanz, C. Pizarro and J. L. Bueno, Self-Diffusion in Molecular Fluids and Noble Gases: Available Data, *J. Chem. Eng. Data*, 2015, **60**, 2757–2817.

35 E. B. Winn, The Temperature Dependence of the Self-Diffusion Coefficients of Argon, Neon, Nitrogen, Oxygen, Carbon Dioxide, and Methane, *Physical Review*, 1950, **80**, 1024–1027.

36 A. Boushehri, J. Bzowski, J. Kestin and E. A. Mason, Equilibrium and Transport Properties of Eleven Polyatomic Gases At Low Density, *J. Phys. Chem. Ref. Data*, 1987, **16**, 445–466.

37 C. R. Mueller and R. W. Cahill, Mass Spectrometric Measurement of Diffusion Coefficients, *J. Chem. Phys.*, 1964, **40**, 651–654.

38 H. F. Vugts, A. J. H. Boerboom and J. Los, Diffusion coefficients of isotopic methane mixtures and of methane-rare-gas mixtures, *Physica*, 1971, **51**, 311–318.

39 A. Greiner-Schmid, S. Wappmann, M. Has and H.-D. Lüdemann, Self-diffusion in the compressed fluid lower alkanes: Methane, ethane, and propane, *J. Chem. Phys.*, 1991, **94**, 5643–5649.

40 K. R. Harris, The density dependence of the self-diffusion coefficient of methane at −50°, 25° and 50°C, *Physica A: Statistical Mechanics and its Applications*, 1978, **94**, 448–464.

41 K. R. Harris and N. J. Trappeniers, The density dependence of the self-diffusion coefficient of liquid methane, *Physica A: Statistical Mechanics and its Applications*, 1980, **104**, 262–280.

42 M. Iwahashi, Y. Yamaguchi, Y. Ogura and M. Suzuki, Dynamical Structures of Normal Alkanes, Alcohols, and Fatty Acids in the Liquid State as Determined by Viscosity, Self-Diffusion Coefficient, Infrared Spectra, and 13CNMR Spin-Lattice Relaxation Time Measurements, *Bull. Chem. Soc. Jpn.*, 1990, **63**, 2154–2158.

43 D. C. Douglass and D. W. McCall, Diffusion in Paraffin Hydrocarbons, *J. Phys. Chem.*, 1958, **62**, 1102–1107.

44 D. W. McCall, D. C. Douglass and E. W. Anderson, Diffusion in Liquids, *J. Chem. Phys.*, 1959, **31**, 1555–1557.

45 P. S. Tofts, D. Lloyd, C. A. Clark, G. J. Barker, G. J. M. Parker, P. McConville, C. Baldock and J. M. Pope, Test liquids for quantitative MRI measurements of self-diffusion coefficient in vivo, *Magn. Reson. Med.*, 2000, **43**, 368–374.

46 M. Holz, S. R. Heil and A. Sacco, Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements, *Physical Chemistry Chemical Physics*, 2000, **2**, 4740–4742.

47 D. W. McCall, D. C. Douglass and E. W. Anderson, Self-Diffusion in Liquids: Paraffin Hydrocarbons, *Phys. Fluids*, 1959, **2**, 87–91.

48 R. Freer and J. N. Sherwood, Diffusion in organic liquids. Part 1.—Appraisal of a gel sectioning technique and its application to self-diffusion in benzene and cyclohexane, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1980, **76**, 1021.

49 R. Freer and J. N. Sherwood, Diffusion in organic liquids. Part 2.—Isotope-mass effects in self-diffusion in benzene and cyclohexane, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1980, **76**, 1030.

50 P. A. Johnson and A. L. Babb, Self-diffusion in.Liquids. 1. Concentration Dependence in Ideal and Non-ideal Binary Solutions, *J. Phys. Chem.*, 1956, **60**, 14–19.

51 G. Guevara-Carrion, C. Nieto-Draghi, J. Vrabec and H. Hasse, Prediction of Transport Properties by Molecular Simulation: Methanol and Ethanol and Their Mixture, *J. Phys. Chem. B*, 2008, **112**, 16664–16674.

52 N. Karger, T. Vardag and H.-D. Lüdemann, Temperature dependence of self-diffusion in compressed monohydric alcohols, *J. Chem. Phys.*, 1990, **93**, 3437–3444.

53 S. Meckl and M. D. Zeidler, Self-diffusion measurements of ethanol and propanol, *Mol. Phys.*, 1988, **63**, 85–95.

54 X. Chen, R. Hu, H. Feng, L. Chen and H.-D. Lüdemann, Intradiffusion, Density, and Viscosity Studies in Binary Liquid Systems of Acetylacetone + Alkanols at 303.15 K, *J. Chem. Eng. Data*, 2012, **57**, 2401–2408.

55 K. P. Das, A. Ceglie and B. Lindman, Microstructure of formamide microemulsions from NMR self-diffusion measurements, *J. Phys. Chem.*, 1987, **91**, 2938–2946.

56 N. Karger, S. Wappmann, N. Shaker-Gaafar and H.-D. Lüdemann, The p, T - dependence of self diffusion in liquid 1-, 2- and 3-pentanol, *J. Mol. Liq.*, 1995, **64**, 211–219.

57 M. I. Hrovat and C. G. Wade, NMR pulsed-gradient diffusion measurements. I. Spin-echo stability and gradient calibration, *Journal of Magnetic Resonance (1969)*, 1981, **44**, 62–75.

58 E. O. Stejskal and J. E. Tanner, Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient, *J. Chem. Phys.*, 1965, **42**, 288–292.

59 D. J. Tomlinson, Temperature dependent self-diffusion coefficient measurements of glycerol by the pulsed N.M.R. technique, *Mol. Phys.*, 1973, **25**, 735–738.

60 I. Chang and H. Sillescu, Heterogeneity at the Glass Transition: Translational and Rotational Self-Diffusion, *J. Phys. Chem. B*, 1997, **101**, 8794–8801.

61 M. A. Awan and J. H. Dymond, Transport Properties of Nonelectrolyte Liquid Mixtures. XI. Mutual Diffusion Coefficients for Toluene+n-Hexane and Toluene+Acetonitrile at Temperatures from 273 to 348 K

and at Pressures up to 25 MPa, *Int. J. Thermophys.*, 2001, **22**, 679–700.

62 H. J. Parkhurst and J. Jonas, Dense liquids. I. The effect of density and temperature on self-diffusion of tetramethylsilane and benzene- *d* 6, *J. Chem. Phys.*, 1975, **63**, 2698–2704.

63 M. A. McCool, A. F. Collings and L. A. Woolf, Pressure and temperature dependence of the self-diffusion of benzene, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1972, **68**, 1489.

64 A. F. Collings and L. A. Woolf, Self-diffusion in benzene under pressure, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1975, **71**, 2296.

65 M. K. Mapes, S. F. Swallen and M. D. Ediger, Self-Diffusion of Supercooled *o* -Terphenyl near the Glass Transition Temperature, *J. Phys. Chem. B*, 2006, **110**, 507–511.

66 H. Walderhaug and K. D. Knudsen, Aqueous Mixtures of a Trisiloxane Surfactant and Oil Studied by SANS and NMR Self-diffusion: Effect of Temperature and Oil Concentration, *J. Solution Chem.*, 2012, **41**, 367–379.

67 J. H. Wang, Self-Diffusion Coefficients of Water, *J. Phys. Chem.*, 1965, **69**, 4412–4412.

68 R. Malhotra, W. E. Price, L. A. Woolf and A. J. Easteal, Thermodynamic and transport properties of 1, 2-dichloroethane, *Int. J. Thermophys.*, 1990, **11**, 835–861.

69 R. L. Hurle and L. A. Woolf, Self-diffusion in liquid acetonitrile under pressure, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1982, **78**, 2233.

70 D. W. McCall, D. C. Douglass and E. W. Anderson, Self-Diffusion in Liquid Ammonia, *Phys. Fluids*, 1961, **4**, 1317–1318.

71 D. E. O'Reilly, E. M. Peterson and C. E. Scheie, Self-diffusion in liquid ammonia and deuteroammonia, *J. Chem. Phys.*, 1973, **58**, 4072–4075.

72 C. D'Agostino, M. D. Mantle, L. F. Gladden and G. D. Moggridge, Prediction of binary diffusion coefficients in non-ideal mixtures from NMR data: Hexane–nitrobenzene near its consolute point, *Chem. Eng. Sci.*, 2011, **66**, 3898–3906.

73 K. R. Harris, Temperature and density dependence of the self-diffusion coefficient of n-hexane from 223 to 333 K and up to 400 MPa, *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 1982, **78**, 2265.

74 A. ENNINGHORST, Density dependence of self-diffusion in liquid pentanes and pentane mixtures, *Mol. Phys.*, 1996, **88**, 437–452.

75 H. S. Sandhu, Coefficient of self-diffusion in liquids using pulsed NMR techniques, *Journal of Magnetic Resonance (1969)*, 1975, **17**, 34–40.

76 L. Chen, T. Groß and H.-D. Lüdemann, T,p-Dependence of Self-Diffusion in the Lower N-methylsubstituted Amides, *Zeitschrift für Physikalische Chemie*, 2000, **214**, 239. DOI:10.1524/zpch.2000.214.2.239.

77 A. Heinrich-Schramm, W. E. Price and H.-D. Lüdemann, Self-Diffusion in Compressed Dimethylether: The Influence of Dipole-Dipole Interaction and Hydrogen Bonding Upon Translational Diffusivity in Simple Fluids, *Zeitschrift für Naturforschung A*, 1995, **50**, 145–148.

78 H. S. Sandhu, Self-diffusion Measurements in Pure Liquids using Spin Echoes, *Can. J. Phys.*, 1971, **49**, 1069–1072.

79 F. X. Prielmeier and H.-D. Lüdemann, Self diffusion in compressed liquid chloromethane, dichloromethane and trichloromethane, *Mol. Phys.*, 1986, **58**, 593–604.

80 R. E. Rathbun and A. L. Babb, SELF-DIFFUSION IN LIQUIDS. III. TEMPERATURE DEPENDENCE IN PURE LIQUIDS [1], *J. Phys. Chem.*, 1961, **65**, 1072–1074.

81 M. Has and H.-D. Lüdemann, Self Diffusion in Compressed Fluid CF_3 Brand CF_3 Cl, *Zeitschrift für Naturforschung A*, 1989, **44**, 1210–1214.

82 H. Ertl and F. A. L. Dullien, Self-diffusion and viscosity of some liquids as a function of temperature, *AIChE Journal*, 1973, **19**, 1215–1223.

83 L. Chen, T. Groß and H.-D. Lüdemann, The density dependence of self-diffusion in some simple amines, *Physical Chemistry Chemical Physics*, 1999, **1**, 3503–3508.

84 M. Holz, X. Mao, D. Seiferling and A. Sacco, Experimental study of dynamic isotope effects in molecular liquids: Detection of translation-rotation coupling, *J. Chem. Phys.*, 1996, **104**, 669–679.

85 M. N. Rodnikova, Z. Sh. Idiyatullin and I. A. Solonina, Mobility of molecules of liquid diols in the temperature range of 303–318 K, *Russian Journal of Physical Chemistry A*, 2014, **88**, 1442–1444.

86 M. Kempka, B. Peplińska and Z. Pajak, Anisotropy of Translational Diffusion in Liquid α,ω-Dibromoalkanes,

View Article Online
DOI: 10.1039/D6CP01425A

Berichte der Bunsengesellschaft für physikalische Chemie, 1988, **92**, 686–689.

87 R. Fingerhut, W.-L. Chen, A. Schedemann, W. Cordes, J. Rarey, C.-M. Hsieh, J. Vrabec and S.-T. Lin, Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid-Phase Equilibria, *Ind. Eng. Chem. Res.*, 2017, **56**, 9868–9884.

88 J. Gross, An equation-of-state contribution for polar components: Quadrupolar molecules, *AIChE Journal*, 2005, **51**, 2556–2568.

89 J. Gross and J. Vrabec, An equation-of-state contribution for polar components: Dipolar molecules, *AIChE Journal*, 2006, **52**, 1194–1204.

90 J. Gross and G. Sadowski, Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules, *Ind. Eng. Chem. Res.*, 2001, **40**, 1244–1260.

91 J. Gross and G. Sadowski, Application of the Perturbed-Chain SAFT Equation of State to Associating Systems, *Ind. Eng. Chem. Res.*, 2002, **41**, 5510–5515.

View Article Online
DOI: 10.1039/D6CP01425A

The experimental data supporting the findings of this study are available within the published literature and have been appropriately cited and reported throughout this manuscript. No new experimental data were generated for this study.

Supplementary materials related to this study are accessible online to support the reproducibility and promote transparency of the proposed model. These materials include:

- dimensionless_self_diffusion_ANN_2026.mat — Contains the trained ANN model, input standardization parameters, and model configuration metadata.
- dimensionless_self_diffusion_Predictor.m — MATLAB script that allows users to predict the self-diffusion coefficient of chemicals by entering COSMO-SAC–derived molecular descriptors, dimensionless form of residual entropy, and molar density. The script uses the trained ANN model.
- COSMO_SAC_derived_molecular_descriptors.xlsx — Contains the COSMO-SAC–derived molecular descriptors for 69 chemicals.