RESEARCH ARTICLE

Journal of
COMPUTATIONAL CHEMISTRY WILEY

Prediction of self-diffusion coefficients of chemically diverse pure liquids by all-atom molecular dynamics simulations

Hiromi Baba[1,2] ⬤ | Ryo Urano[3] | Tetsuro Nagai[2,4] | Susumu Okazaki[2]

[1]New Modality Research Group, Research Department, Kyoto R&D Center, Maruho Co., Ltd., Kyoto, Japan

[2]Department of Advanced Materials Science, The University of Tokyo, Chiba, Japan

[3]Research Institute for Interdisciplinary Science, Okayama University, Okayama, Japan

[4]Department of Chemistry, Faculty of Science, Fukuoka University, Nanakuma, Japan

Correspondence
Hiromi Baba, New Modality Research Group, Research Department, Kyoto R&D Center, Maruho Co., Ltd., 93 Chudoji-Awatacho, Shimogyo-Ku, Kyoto 600-8815, Japan.
Email: baba_dfq@mii.maruho.co.jp

Abstract

Molecular self-diffusion coefficients underlie various kinetic properties of the liquids involved in chemistry, physics, and pharmaceutics. In this study, 547 self-diffusion coefficients are calculated based on all-atom molecular dynamics (MD) simulations of 152 diverse pure liquids at various temperatures employing the OPLS4 force field. The calculated coefficients are compared with experimental data (424 extracted from the literature and 123 newly measured by pulsed-field gradient nuclear magnetic resonance). The calculations well agree with the experimental values. The determination coefficient and root mean square error between the observed and calculated logarithmic self-diffusion coefficients of the 547 entries are 0.931 and 0.213, respectively, demonstrating that the MD calculation can be an excellent industrial tool for predicting, for example, molecular transportation in liquids such as the diffusion of active ingredients in biological and pharmaceutical liquids. The self-diffusion coefficients collected in this study are compiled into a database for broad researches including artificial intelligence calculations.

KEYWORDS
liquid state, mean square displacement, molecular dynamics simulation, self-diffusion coefficient, water model

1 | INTRODUCTION

Liquid materials are ubiquitous in organisms, environments, and industries involving reactants, solvents, and other functional entities. The diffusion of molecules in liquids has attracted considerable attention in the fields of fluid mechanics and liquid science, materials physics, chemical engineering, and theoretical chemistry.[1–5] The self-diffusion coefficients of molecules in liquids are sensitive to the molecular structure, as well as the thermodynamic conditions such as temperature and pressure. Despite the increasing importance of the self-diffusion coefficient in theoretical and applied chemistry, successful reports including a comprehensive prediction of the self-diffusion

coefficient of chemically diverse liquids under various conditions have been quite limited.[6]

Computationally, the self-diffusion coefficients (D) of molecules in liquids have been obtained with molecular dynamics (MD) calculations by two popular methods. The first method integrates the velocity auto-correlation function over time using the Green–Kubo formula[1]

$$D = \frac{1}{3}\int_0^\infty \langle v_i(0) \cdot v_i(t)\rangle dt, \qquad (1)$$

where v_i is the velocity of a molecule i of interest at time t and the angular brackets represent the ensemble average. The second method uses the mean square displacement (MSD) as[7]

$$D = \lim_{t \to \infty} \frac{1}{6t} \langle |r_i(t) - r_i(0)|^2 \rangle, \quad (2)$$

where r_i is the position of molecule i.

Various force fields[8,9] for MD simulations have been developed and employed in the fields of biophysics,[10] chemical physics,[11] materials science,[12] and drug discovery.[13] In particular, OPLS4,[14] among the newest force fields and an improved version of OPLS3e,[15] is a promising force field with new potential parameters. Therefore, applying this force field to self-diffusion coefficients is an interesting proposition.

Experimentally, self-diffusion coefficients in liquids have been determined using nuclear magnetic resonance (NMR) with spin-echo techniques[16,17] as well as radioactive or stable isotopic tracers.[18–20] The former techniques, especially pulsed-field gradient (PFG)–NMR,[21,22] which is also referred to as pulsed-gradient spin-echo NMR, and their improvements[23,24] have been widely used to measure the self-diffusion coefficient. The NMR methods have handled a broad range of diffusion coefficients and temperature conditions since their development by Stejskal and Tanner in 1968.[25] In PFG–NMR experiments using a rectangular gradient pulse, the self-diffusion coefficient D is obtained as[26]

$$\frac{S}{S_0} = \exp \left[-\gamma^2 g^2 \delta^2 D \left(\Delta - \frac{\delta}{3} \right) \right], \quad (3)$$

where γ is the gyromagnetic ratio, which is intrinsic to the resonant nucleus, g is the pulse gradient, δ is the width of the pulse gradient, Δ is the interval between the gradient pulses, S is the observed signal on the gradient, and S_0 is the signal at zero gradient.

In this study, we have attempted to evaluate the performance of predicting self-diffusion coefficient for chemically diverse pure liquids by all-atom MD simulations. To this end, we employed the following three processes: (1) compilation of a new large dataset containing the self-diffusion coefficients of pure liquids by performing a literature survey and conducting additional PFG-NMR experiments; (2) all-atom MD simulations of the self-diffusion coefficients for all pure liquids found in our database; (3) assessment of the predictive performances of our self-diffusion coefficient calculations using reliable statistical metrics.

2 | METHODS

We calculated 547 diffusion coefficients from the MD trajectories of molecules in pure liquids and compared them with the experimental values. Among the 547 experimental data, 424 were extracted from the literature and 123 were newly measured by the PFG–NMR technique. The data were all compiled in a database constructed in the present study.

2.1 | System preparation

The initial liquid structures for the MD calculations were efficiently prepared using the following procedures. First, the two-dimensional molecular structures of liquids in our database were obtained from CAS SciFinder. Subsequently, we applied the LigPrep package with the OPLS4 force field for energy minimization in the Schrödinger Small-Molecule Discovery Suite ver. 2021-4[27] to generate the three-dimensional molecular structures of each liquid. Based on the generated molecular structures of the liquids, we constructed a cubic simulation cell for each pure liquid using the System Builder and Prepare for MD functions in the Schrödinger Materials Science suite: Polymer Package (MSS) ver. 2021-4.[27] One simulation cell contained over 1000 molecules of a pure liquid to ensure sufficient statistical convergence. The OPLS4 force field was applied to the simulation cells. For water, we prepared eight simulation cells for eight different potential models, namely, the single-point charge (SPC),[28] SPC/E,[29] the transferable intermolecular potential with three, four, and five points (TIP3P,[30] TIP4P,[31] and TIP5P,[32] respectively), TIP4P/2005,[33] TIP4P-Ew,[34] and TIP4P-D.[35] The performances of the potential models in the prediction of self-diffusion coefficients were then compared.

2.2 | Equilibration calculation

To attain thermal equilibrium of the calculation systems, we carried out the following four-stage process before computing the self-diffusion coefficients: (1) Brownian dynamics[37] at 10 K for 100 ps with a 1 fs time step; (2) MD calculation in the canonical (NVT) ensemble using the Langevin thermostat[1] at 10 K for 100 ps with a 2 fs time step; (3) MD calculation in the canonical ensemble using the Langevin thermostat at the temperature of interest for 100 ps with a 2 fs time step; (4) MD calculation in the isothermal–isobaric (NPT) ensemble using the Nose–Hoover thermostat[38] and the Martyna–Tobias–Klein barostat[39] at the temperature of interest and 1.01325 bar for 20 ns with a 2 fs time step. The electrostatic interaction was calculated by the u-series algorithm[40] developed by Shaw and coworkers, and the cutoff radius of the short-range interaction was set to 9.0 Å. Other tunable parameters were used as default values for simulations and all simulations were performed with periodic boundary conditions in the Desmond/G package ver. 6.8.135.[41,42]

2.3 | Calculation of the diffusion coefficients

Production MD runs of the self-diffusion coefficients in the NPT ensemble were started from the equilibrated simulation systems and continued at the temperature of interest. Theoretically, the self-diffusion coefficient is accurately calculated based on Equation (2) and practically we can estimate self-diffusion coefficient from the slope of the MSD versus sufficient long lag time τ.[43] In this study, we

divided the database into the two sets of highly diffusive samples (with logarithmic experimental diffusion coefficients [m²/s] larger than −9.5) and lowly diffusive samples (the rest of the entire dataset) to set an appropriate simulation time of 40 ns for the former and 150 ns for the latter. The pressure, temperature, and other settings were kept the same as described in preprocess (4) above. The MSD of a molecule's center-of-mass was calculated from the trajectories as a function of at 4 ps intervals and the MSDs of all molecules in the simulation system were averaged using the Diffusion Coefficient tools in the MSS. Subsequently, the self-diffusion coefficient was calculated as one-sixth of the slope of the averaged MSD versus lag time ranging from 12 to 20 ns for the highly diffusive samples and from 45 to 75 ns for the lowly diffusive samples, which was linearly regressed by the least-squares technique using the lm function in the R environment ver. 4.1.1.[44] To assess the adequacy of the calculations of self-diffusion coefficients, we evaluated the determination coefficients of the linear regressions and diffusion coefficient transitions of the system every 500 ps lag time.

The 547 logarithmic values of self-diffusion coefficients predicted in the present MD calculations were statistically compared with the experimental data by four metrics, namely, the determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), and concordance correlation coefficient (CCC).[45] The four statistics are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - y_i^{\mathrm{calc}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - \widehat{y}^{\mathrm{obs}}\right)^2}, \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - y_i^{\mathrm{calc}}\right)^2}{n}}, \tag{5}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|y_i^{\mathrm{obs}} - y_i^{\mathrm{calc}}\right|}{n}, \tag{6}$$

$$\mathrm{CCC} = \frac{2\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - \widehat{y}^{\mathrm{obs}}\right)\left(y_i^{\mathrm{calc}} - \widehat{y}^{\mathrm{calc}}\right)}{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - \widehat{y}^{\mathrm{obs}}\right)^2 + \sum_{i=1}^{n}\left(y_i^{\mathrm{calc}} - \widehat{y}^{\mathrm{calc}}\right)^2 + n\left(\widehat{y}^{\mathrm{obs}} - \widehat{y}^{\mathrm{calc}}\right)^2}. \tag{7}$$

where y_i^{obs}, y_i^{calc}, $\widehat{y}^{\mathrm{obs}}$, and $\widehat{y}^{\mathrm{calc}}$ are the observed, calculated, mean of the observed, and mean of the calculated logarithmic self-diffusion coefficients, respectively, and n is the number of samples. All metrics were calculated in the R environment ver. 4.1.1.[44] The experimental data were collected as described in the next section.

2.4 | Self-diffusion coefficients in the database

From the literature, we extracted 424 self-diffusion coefficients of 81 chemically diverse pure liquids.[46–62] This compilation process excluded self-diffusion data of stereoisomeric mixtures. To quantitatively evaluate the impact of temperature on self-diffusion, we

collected the self-diffusion coefficients at different temperatures for most of the individual liquids.

To expand the chemical diversity of our database, we conducted PFG-NMR spectroscopy experiments of 75 pure liquids at temperatures ranging from 278.15 to 328.15 K using the bipolar gradient stimulated echo pulse sequence.[24] All measurements were made on protons at 395.88 MHz. Applying Equation (3) to the experimental data, we newly obtained 123 self-diffusion coefficients. The details of materials, instrumental information, and parameter settings of the experiments are given in Tables S1 and S2. Note that although the four liquids, acetone, acetonitrile, ethanol, and tetrahydrofuran, are duplicated in the literature and our additional observations, the diffusion temperature for each liquid is different in the two cases. Thus, we integrated our new observations with the literature data to construct a database embracing 547 self-diffusion coefficients for 152 pure liquids. Our database includes the liquid names, CAS registry numbers, diffusion temperatures, logarithmic values of the experimental and calculated self-diffusion coefficients (unit: m²/s), the simulation-box sizes after equilibration calculations, the number of molecules in the simulation boxes and data references. The database is provided in Table S3.

2.5 | Database characterization

We examined the distribution of the logarithmic self-diffusion coefficients and the number of atoms of the constituent molecules of liquids in our database. To chemically characterize our database, we generated 5290 molecular descriptors consisting of one-, two-, and three-dimensional descriptors[63] and various molecular properties, based on three-dimensional neutral structures (see Section 2.1) for all the 152 pure liquids utilizing alvaDesc ver. 2.0.10.[64] After eliminating the constant and erroneous descriptors, 2119 descriptors remained in the list. Among this molecular descriptor pool, we selected five descriptors that are frequently used as chemical characteristics—molecular weight, octanol–water partition coefficient (ALogP),[65] number of rotatable bonds, topological polar surface area, and number of hydrogen bonds—and determined their statistical metrics to evaluate the chemical diversity of our database. We then evaluated the complementary relation between the experimentally observed liquids in this study and the liquids reported in the literature. To this end, we visually compared the three-dimensional chemical spaces of the corresponding liquids by the t-distributed stochastic neighbor embedding (t-SNE) technique.[66] The t-SNE analysis was executed by applying the tsne function in the tsne R package ver. 0.1-3[67] to the scaled values of the 2119 descriptors of the liquids. These statistical analyses were implemented in the R environment ver. 4.1.1.[44]

3 | RESULTS AND DISCUSSION

3.1 | Chemical diversity of the materials

To evaluate the performance of computational methods for predicting physico-chemical or thermodynamic properties of materials, a

sufficient chemical diversity of the test samples is essential. The database compiled in this study comprises 547 self-diffusion coefficients of 152 chemically diverse pure liquids and is much larger than that of previous work.[6] Figure 1 shows the distributions of the logarithmic self-diffusion coefficients and the sizes of the constituent molecules of the pure liquids in our database. As shown in Figure 1A, the logarithmic self-diffusion coefficient was widely distributed from extremely low (under −12.0) to high (over −8.5) values and the number of atoms of the molecules of liquids in our database also widely ranged from very small (under 10) to comparatively large (over 50) values, as shown in Figure 1B. These results suggest that our database is sufficiently diverse for assessing the predictive ability of MD for self-diffusion coefficients. Actually, liquids are composed of a wide range of chemical structures, such as alcohols (monohydric and polyhydric), hydrocarbons (acyclic, alicyclic, aromatic, and halogenated), short-, medium-, and long-chain saturated fatty acids (n-alkanoic acids), amides, esters, carbonates, ketones, nitriles, amines, thiols, sulfides, ethers, silanes, nitro compounds, and other polar liquids. Note that many of the liquids in our database are often used as thermal energy storage media, pharmaceutical excipients and solvents in electronics, pharmaceutical, and chemical industries.[68–72] The diversity of our database is also demonstrated by the statistics summarized in Table 1, including the molecular weight, octanol–water partition coefficient, number of rotatable bonds, topological polar surface area, and number of hydrogen bonds. Clearly, the liquids in our database exhibit a wide range of chemical characteristics.

In this study, we experimentally determined the self-diffusion coefficients of 75 pure liquids by PFG-NMR techniques and 71 of them were not retrieved from the literature. The expansion of the chemical space by our additional data is graphically demonstrated by t-SNE techniques as a three-dimensional scatter plot and its projections to two-dimensional planes in Figure 2. As shown in the figure, the red data points representing the 71 additional liquids occupy a different space from the black data points corresponding to the literature liquids, which means that the additional liquids efficiently complement the liquids presented in the literature and provide additional chemical diversity to the database. These results suggest that our database covers a sufficient chemical space for assessing the predictive performance of MD for self-diffusion coefficients.

3.2 | Adequacy of calculating diffusion coefficients

The determination coefficient (R^2) of the linear regression of the averaged MSD as a function of lag time (τ) critically determine the

FIGURE 1 Distributions of (A) logarithmic self-diffusion coefficients and (B) number of atoms of the constituent molecules of liquids contained in the database. The red, blue, and green bars represent the experimentally observed liquids in the present study, the liquids collected from the literature, and the liquids in both the literature and our observations, respectively.
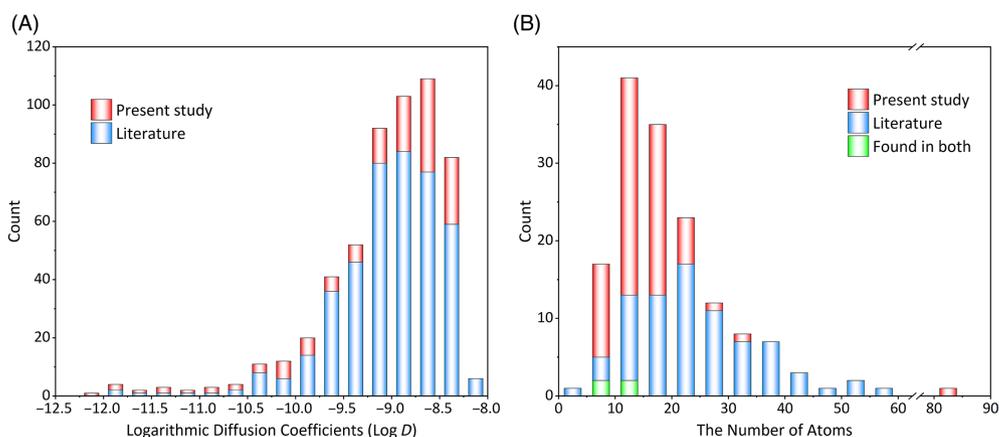
TABLE 1 Chemical characteristics of the liquids in our database

Metrics	Log D[a]	MW[b]	A Log P[c]	RBN[d]	TPSA[e]	HBA[f]	HBD[g]
Samples	547	152	152	152	152	152	152
Max	−8.20	410.8	11.33	16	60.69	4	3
Median	−8.94	100.1	1.81	1	20.23	1	0
Min	−12.21	18.0	−1.41	0	0	0	0
SD[h]	−0.65	54.1	1.91	3.76	15.59	1.04	0.63

[a]Logarithmic self-diffusion coefficient.
[b]Molecular weight.
[c]Ghose–Crippen octanol–water partition coefficient.
[d]Number of rotatable bonds.
[e]Topological polar surface area.
[f]Number of hydrogen bond acceptors.
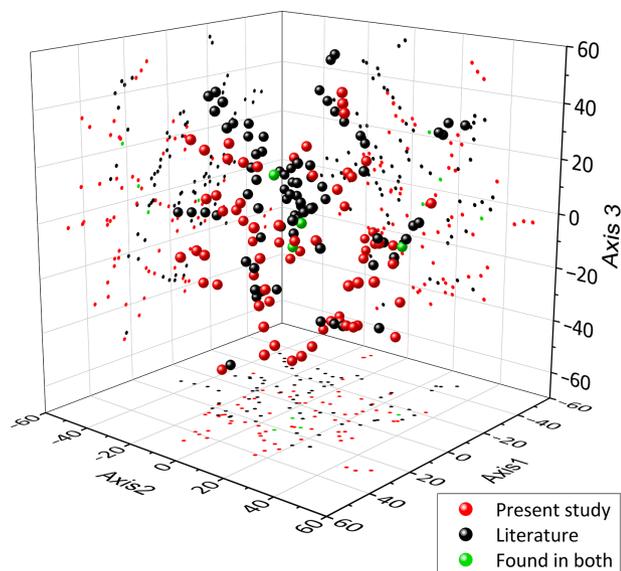[g]Number of hydrogen bond donors.
[h]Standard deviation.

FIGURE 2 Three-dimensional chemical space by *t*-SNE analysis. The red, black, and green data points represent the experimentally observed liquids in the present study, the liquids collected from the literature, and the liquids in both the literature and our observations, respectively.

TABLE 2 Statistics of the calculated mean square displacements in the MD calculations

Metrics	Values
Samples	610
Max	1.0000
Median	0.99999
Min	0.99964
Standard deviation	0.00003

FIGURE 3 Plots of MSD versus lag time (τ): (A) acetonitrile at 308.15 K, (B) (*R*)-1,2,4-butanetriol at 288.15 K

adequacy of the analyses, including the system preparation, simulation preprocesses, and sampling simulation. Table 2 summarizes the statistics of the R^2 values for all 610 regressions, corresponding to $538 + 9 \times 8$ sampling simulations, where nine water-diffusion entries of all the 547 entries in our database were applied to eight different water models. Table 2 clearly shows that the plots of MSD versus prescribed τ range (12–20 ns for highly diffusive liquids or 45–75 ns for lowly diffusive liquids) were almost completely linear, as all the R^2 values were greater than 0.9996, indicating that our simulations statistically converged to the diffusional region. Based on the molecular behavior, plots of MSD versus lag time can take two major modes, that is, the ballistic and the diffusion regimes.[73] The former is a collision-free mode that is found at earlier simulation times where the MSD plots are nonlinear. The latter is a typical random-walk (diffusion) mode based on molecular collisions, which gives rise to linear MSD plots. Two examples of MSD plots are presented in Figure 3. Figure 3A plots the results of acetonitrile at 308.15 K, which yielded the maximum calculated self-diffusion coefficient in our database. In contrast, Figure 3B plots the results of (*R*)-1,2,4-butanetriol at 288.15 K, which shows the minimum self-diffusion coefficient. The MSD versus τ plot of (*R*)-1,2,4-butanetriol with the small self-diffusion coefficient was clearly nonlinear at small τ. The nonlinearity survived longer for the sample with low diffusion coefficient than for the sample with high diffusion coefficient. Actually, the diffusion coefficients of (*R*)-1,-2,4-butanetriol per 500 ps exhibited slower convergence than those of acetonitrile, although transitions of diffusion coefficients of both liquids were in the plateau region in a prescribed τ range (12–20 ns for acetonitrile or 45–75 ns for (*R*)-1,2,4-butanetriol) as shown in Figure 4. The same tendency is seen in all calculations on the samples in our entire database. These results indicate that our MSD calculation time, which is 12–20 ns
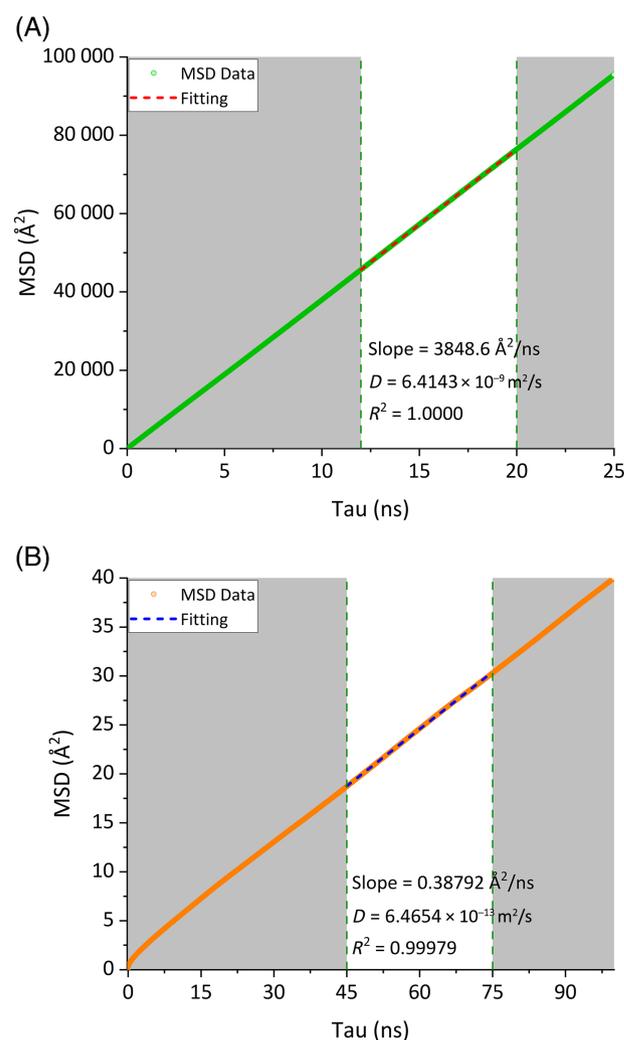
for highly diffusive liquids or 45–75 ns for lowly diffusive liquids, is appropriate for analyzing the self-diffusion coefficients.

3.3 | Comparison of water models

Water is a ubiquitous liquid. This simple, two-element, three-atom molecule has been extensively studied over 50 years. The anomalous
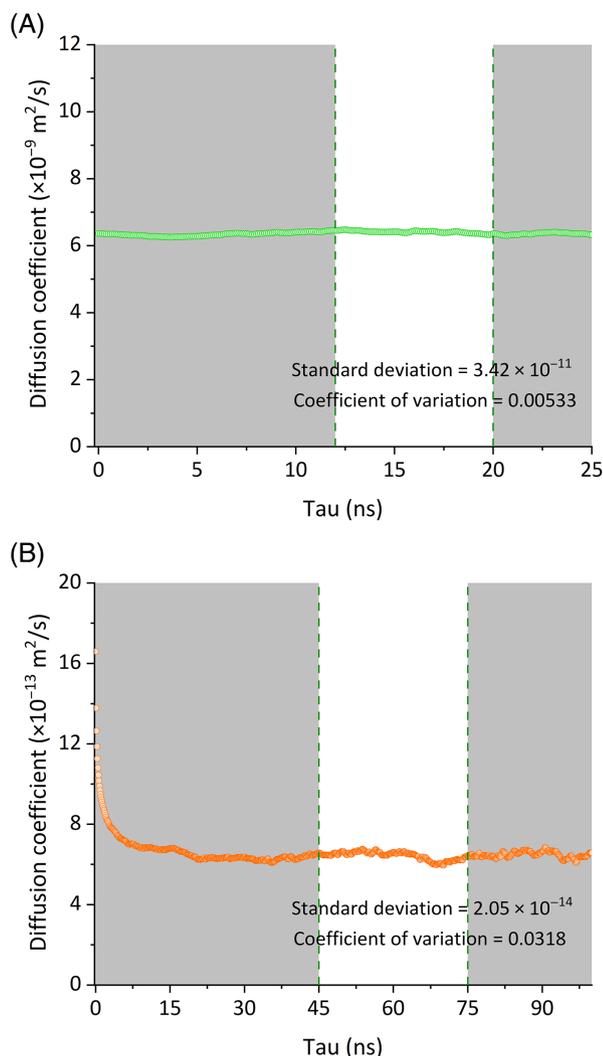
(A)

Standard deviation = 3.42 × 10⁻¹¹
Coefficient of variation = 0.00533

(B)

Standard deviation = 2.05 × 10⁻¹⁴
Coefficient of variation = 0.0318

FIGURE 4 Transitions of diffusion coefficients every 500 ps lag time (τ): (A) acetonitrile at 308.15 K, (B) (R)-1,2,4-butanetriol at 288.15 K

FIGURE 5 Comparison of calculated and experimental self-diffusion coefficients as a function of temperature for the eight water models

TABLE 3 Performances of the six water models in predicting self-diffusion coefficients

Model	R^2	RMSE	MAE	CCC
SPC	0.998	0.260	0.257	0.254
SPC/E	0.998	0.066	0.063	0.863
TIP3P	0.995	0.375	0.371	0.123
TIP4P	0.998	0.207	0.204	0.373
TIP4P/2005	0.999	0.012	0.009	0.996
TIP4P-D	0.999	0.041	0.039	0.947
TIP4P-Ew	0.998	0.052	0.050	0.918
TIP5P	0.996	0.137	0.134	0.675

R^2	RMSE	MAE	CCC
0.931	0.213	0.171	0.947

FIGURE 6 Self-diffusion coefficients calculated by the present all-atom MD calculations versus the experimental data for 547 entries in our database

physical properties[74] of water have been simulated in various water models.[75] Each model reasonably reproduces particular properties of water, but has its own limitations. Therefore, we compared the water self-diffusion coefficients obtained by eight popular water models: SPC, SPC/E, TIP*n*P ($n = 3, 4, 5$), TIP4P/2005, TIP4P-Ew, and TIP4P-D, and assessed the predictive ability of each model. Figure 5 shows the plots of the self-diffusion coefficients of water calculated by all-atom MD simulations as a function of temperature ranging from 288.15 to 329.15 K. The experimentally observed values[54] are also given for comparison. Clearly, the prediction errors depended on the model, but all models except TIP5P obtained a common temperature dependence with similar slopes. However, the difference in the self-diffusion coefficient calculated by TIP5P and the experimental ones increased further with increasing temperature. We observed that the descending order of self-diffusion coefficient around 300 K is TIP3P > SPC > TIP4P > TIP5P > SPC/E > TIP4P-Ew > TIP4P/2005 > experimental > TIP4P-D, which is in quite agreement with existing studies.[35,76,77] The TIP3P, SPC, TIP4P, and TIP5P models
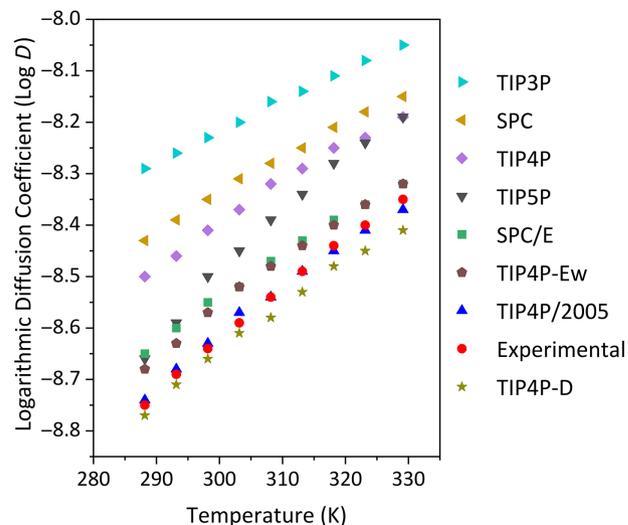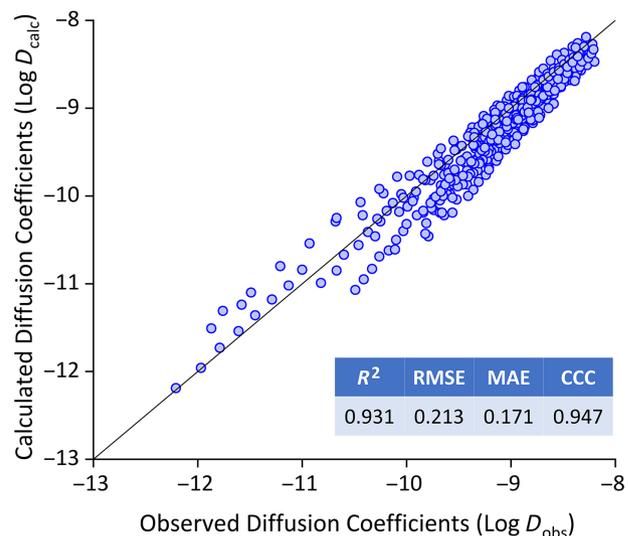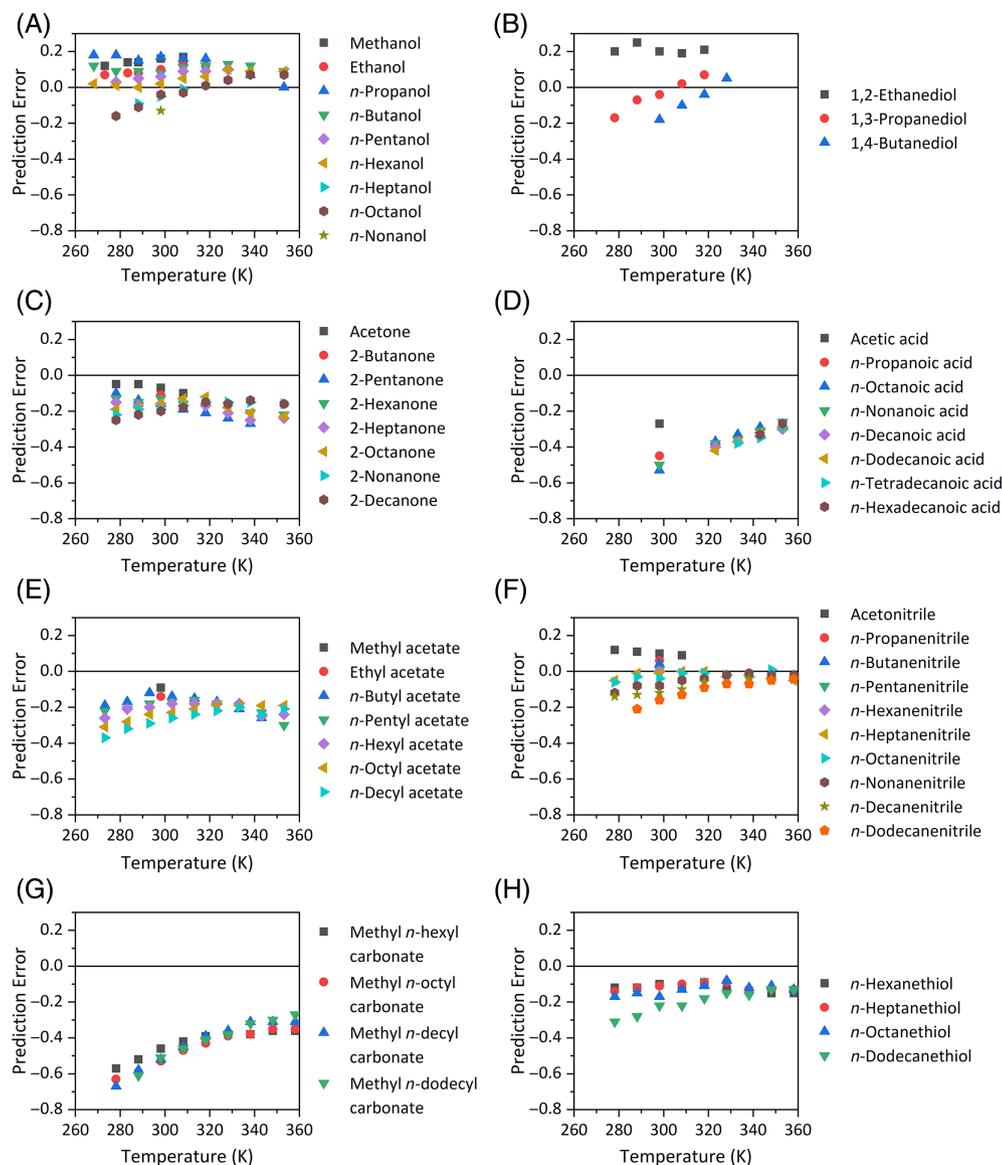
FIGURE 7 Prediction error versus temperature for various homologous series: (A) *n*-alcohols, (B) terminal diols, (C) 2-alkanones, (D) *n*-alkanoic acids, (E) *n*-alkyl acetates, (F) *n*-nitriles, (G) methyl *n*-alkyl carbonates, and (H) *n*-thiols

substantially overestimated the experimental self-diffusion coefficient, whereas the remaining models (TIP4P/2005, TIP4P-Ew, SPC/E, and TIP4P-D) reproduced the experimental data quite well over the entire temperature range (with TIP4P/2005 being slightly better than the others; see Figure 5).

The predictive statistics, namely, R^2, RMSE, MAE, and CCC, between the experimental and calculated self-diffusion coefficients for eight water models are summarized in Table 3. The TIP4P-2005 achieved the best RMSE, MAE, and CCC values, indicating that this water potential is the most favorable model for estimating the self-diffusion coefficient within the temperature range of this study. Meanwhile, the R^2 values were almost identical in all models (>0.995), reflecting the high correlation between the experimental and calculated self-diffusion coefficients versus temperature. Consequently, we adopted TIP4P-2005 as the water model in the following discussion of the total predictive performances of all-atom MDs for self-diffusion coefficient.

3.4 | Predictive performance of all-atom MDs

Figure 6 presents the predictive performances of all-atom MD simulations for the 547 self-diffusion coefficients of 152 pure liquids in our database, which is among the largest ones tested by all-atom MD. The statistical metrics R^2, RMSE, MAE, and CCC between the experimental and calculated self-diffusion coefficients are also shown. These statistics demonstrate the excellent performance of the MD predictions. That is, our all-atom MD simulations can strongly predict a wide range of self-diffusion coefficients of chemically diverse liquids spanning over four common logarithmic units. Note that no anisotropic diffusions were observed in all our simulations because coefficients of variation for calculated diffusion coefficients along the three axes (*x*, *y*, and *z*) of the 547 samples ranged from 2.10×10^{-3} to 1.18×10^{-1}.

In our database, there are homologous compounds series, for example, sequential nine normal alcohols containing one to nine carbon atoms (methanol to *n*-nonanol), in several chemical groups. To

understand the prediction trends of the self-diffusion coefficients obtained in our MD simulations, we calculated the prediction error as the difference between the experimental and calculated self-diffusion coefficients. The results of eight homologous series—n-alcohols, terminal diols, 2-alkanones, n-alkanoic acids, n-alkyl acetates, n-nitriles, methyl n-alkyl carbonates, and n-thiols—are shown in Figure 7.

As shown in the figure, carbon homologation tended to underestimate the self-diffusion coefficients at lower temperatures in the MD simulations. The prediction-error curves differed among the liquids, even among those within the same homologous series, and could be typically classified into three trend types: growing asymptotically, changing linearly, and adopting a mountain-like shape. These trends were typically found for 2-decanone, 2-pentanone, and 2-octanone, respectively. Furthermore, the prediction errors of methyl n-alkyl carbonates and n-alkanoic acids were larger than other liquids with similar molecular weights containing other chemical groups: at the same temperature, the absolute prediction errors for n-hexyl methyl carbonates (molecular weight = 160.24) were more than 3.8 times those of n-decanenitrile (153.30) and more than 2.2 times those of 2-decanone (156.30). Additionally, three n-alkanoic acids (n-octanoic acid, n-decanoic acid, and n-dodecanoic acid) share a constitutional isomeric relationship with three n-alkyl acetates (n-hexyl acetate, n-octyl acetate, and n-decyl acetate, respectively). However, in the 323–353 K range, the absolute prediction errors in the self-diffusion coefficients of the former acids were 1.2–2.1 times larger than those of the latter acetates. In particular, n-alkanoic acids have been reported to exhibit a characteristic dynamical behavior that can lead to rod-like dimers and aggregation even in the liquid state.[53] Furthermore, the formation of dimers or higher-order molecular structures and the resultant dynamic properties of liquid n-alkanoic acids depend on the conformational geometries of the acid molecules.[69]

To further improve the predictive performance, an elaborate molecular modeling and a sophisticated re-parametrization of the force field, for example, by quantum-chemical calculations[36] may be required.

4 | CONCLUSIONS

We compiled 547 self-diffusion coefficients of 152 chemically diverse pure liquids into a new database, which comprises 424 literature values and 123 newly measured values from PFG-NMR experiments. Our integrated database of self-diffusion coefficients, embracing wide chemical characteristics, can be used not only for the assessment of the predictive ability of the MD calculations but also for the extensive researches including artificial intelligence calculations.

All-atom MD calculations employing the OPLS4 force field exhibited excellent predictive performances for entire samples in our database, with R^2 and RMSE values between the logarithmic experimental and calculated self-diffusion coefficients were 0.931 and 0.213, respectively. Furthermore, our results revealed that both TIP4P/2005 achieved the best performance of reproduction of self-diffusion coefficient among eight common water models in the temperature range from 288.15 to 329.15 K.

This fully computational all-atom MD prediction can be practical tools for the analysis of self-diffusion coefficient and provide the basis of industrial trials, for instance, the diffusion of active pharmaceutical ingredients in various biological and pharmaceutical liquids.

CONFLICT OF INTEREST

Hiromi Baba is an employee of Maruho Co., Ltd. Tetsuro Nagai is a consultant for Maruho Co., Ltd. Ryo Urano and Susumu Okazaki have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Hiromi Baba https://orcid.org/0000-0001-6923-7434

REFERENCES

[1] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, 2nd ed., Oxford University Press, New York **2017**.

[2] K. Ohkubo, K. Yanagisawa, A. Kamimura, K. Fujii, *J. Phys. Chem. B* **2020**, *124*, 3784.

[3] N. Meyer, J. F. Wax, H. Xu, *J. Chem. Phys.* **2018**, *148*, 234506.

[4] H. K. Chilukoti, G. Kikugawa, T. Ohara, *J. Phys. Chem. B* **2016**, *120*, 7207.

[5] N. de Sousa, J. J. Sáenz, F. Scheffold, A. García-Martín, L. S. Froufe-Pérez, *J. Phys. Condens. Matter* **2016**, *28*, 135101.

[6] J. Wang, T. Hou, *J. Comput. Chem.* **2011**, *32*, 3505.

[7] A. Einstein, *Ann. Phys.* **1905**, *17*, 549.

[8] P. S. Nerenberg, T. Head-Gordon, *Curr. Opin. Struct. Biol.* **2018**, *49*, 129.

[9] J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg, B. H. Morrow, *Appl. Phys. Rev.* **2018**, *5*, 031104.

[10] S. A. Hollingsworth, R. O. Dror, *Neuron* **2018**, *99*, 1129.

[11] D. Saric, M. Kohns, J. Vrabec, *J. Phys. Chem.* **2020**, *152*, 164502.

[12] D. Bedrov, J. P. Piquemal, O. Borodin, A. D. MacKerell Jr., B. Roux, C. Schröder, *Chem. Rev.* **2019**, *119*, 7940.

[13] A. Ganesan, M. L. Coote, K. Barakat, *Drug Discov. Today* **2017**, *22*, 249.

[14] C. Lu, C. Wu, D. Ghoreishi, W. Chen, L. Wang, W. Damm, G. A. Ross, M. K. Dahlgren, E. Russell, C. D. Von Bargen, R. Abel, R. A. Friesner, E. D. Harder, *J. Chem. Theory Comput.* **2021**, *17*, 4291.

[15] K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner, E. D. Harder, *J. Chem. Theory Comput*. **2019**, *15*, 1863.

[16] E. L. Hahn, *Phys. Rev.* **1950**, *80*, 580.

[17] H. Y. Carr, E. M. Purcell, *Phys. Rev.* **1954**, *94*, 630.

[18] A. F. Collings, D. C. Hall, M. A. McCool, L. A. Woolf, *J. Phys. E: Sci. Instrum.* **1971**, *4*, 1019.

[19] A. J. Easteal, W. E. Price, L. A. Woolf, *J. Chem. Soc. Faraday Trans.* **1989**, *1*(85), 1091.

[20] K. C. Pratt, W. A. Wakeham, *J. Chem. Soc., Faraday Trans* **1977**, *2*, 73.

[21] W. S. Price, *Concepts Magn. Reson.* **1997**, *9*, 299.

[22] W. S. Price, *Concepts Magn. Reson.* **1997**, *10*, 197.

[23] R. Evans, *Prog. Nucl. Magn. Reson. Spectrosc.* **2020**, *117*, 33.

[24] D. H. Wu, A. D. Chen, C. S. J. Johnson, *J. Magn. Reson. Ser. A* **1995**, *115*, 260.

[25] E. O. Stejskal, J. E. Tanner, *J. Chem. Phys.* **1965**, *42*, 288.

[26] W. S. Price, in *Experimental Thermodynamics Volume IX: Advances in Transport Properties of Fluids* (Eds: M. J. Assael, A. R. H. Goodwin, V. Vesovic, W. A. Wakeham), Royal Society of Chemistry, London **2014**, p. 75.

[27] Schrödinger, L. L. C., https://www.schrodinger.com/. (accessed January 2022)

[28] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, J. Hermans, in *Intermolecular Forces* (Ed: B. Pullman), D. Reidel Publishing, Dordrecht **1981**, p. 331.

[29] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem.* **1987**, *91*, 6269.

[30] W. L. Jorgensen, *J. Am. Chem. Soc.* **1981**, *103*, 335.

[31] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926.

[32] M. W. Mahoney, W. L. Jorgensen, *J. Chem. Phys.* **2000**, *112*, 8910.

[33] J. L. F. Abascal, C. Vega, *J. Chem. Phys.* **2005**, *123*, 234505.

[34] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, T. J. Head-Gordon, *J. Chem. Phys.* **2004**, *120*, 9665.

[35] S. Piana, A. G. Donchev, P. Robustelli, D. E. Shaw, *J. Phys. Chem. B* **2015**, *119*, 5113.

[36] A. Ricci, G. Ciccotti, *Mol. Phys.* **2003**, *12*, 1927.

[37] D. J. Evans, B. L. Holian, *J. Chem. Phys.* **1985**, *83*, 4069.

[38] G. J. Martyna, D. J. Tobias, M. L. Klein, *J. Chem. Phys.* **1994**, *101*, 4177.

[39] C. Predescu, A. K. Lerer, R. A. Lippert, B. Towles, J. P. Grossman, R. M. Dirks, D. E. Shaw, *J. Chem. Phys.* **2020**, *152*, 084113.

[40] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, D. E. Shaw, in *Proc. ACM/IEEE Conf. Supercomput. (SC06)*, Tampa, FL, **2006**.

[41] D. E. Shaw Research, Desmond Molecular Dynamics System and Schrödinger, Maestro-Desmond Interoperability Tools, https://www.schrodinger.com/products/desmond. (accessed January 2022)

[42] T. Giorgino, *J. Open Source Softw.* **2019**, *4*, 1698.

[43] L. I. Lin, *Biometrics* **1989**, *45*, 255.

[44] R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/. (accessed May 2022)

[45] M. Petrowsky, A. Fleshman, M. Ismail, D. T. Glatzhofer, D. N. Bopege, R. Frech, *J. Phys. Chem. B* **2012**, *116*, 10098.

[46] M. Petrowsky, R. Frech, *J. Phys. Chem. B* **2010**, *114*, 8600.

[47] Y. Kasahara, Y. Suzuki, A. Kabasawa, H. Minami, H. Matsuzawa, M. Iwahashi, *J. Oleo Sci.* **2010**, *59*, 21.

[48] M. A. Connell, P. J. Bowyer, P. A. Bone, A. L. Davis, A. G. Swanson, M. Nilsson, G. A. Morris, *J. Magn. Reson.* **2009**, *198*, 121.

[49] M. Iwahashi, Y. Kasahara, *J. Oleo Sci.* **2007**, *56*, 443.

[50] H. Kato, T. Saito, M. Nabeshima, K. Shimada, S. Kinugasa, *J. Magn. Reson.* **2006**, *180*, 266.

[51] W. S. Price, H. Ide, Y. Arata, *J. Phys. Chem. A* **2003**, *107*, 4784.

[52] M. Iwahashi, Y. Kasahara, H. Minami, H. Matsuzawa, M. Suzuki, Y. Ozaki, *J. Oleo Sci.* **2002**, *51*, 157.

[53] M. Holz, S. R. Heil, A. Sacco, *Phys. Chem. Chem. Phys.* **2000**, *2*, 4740.

[54] W. S. Price, O. Söderman, *J. Phys. Chem. A* **2000**, *104*, 5892.

[55] K. Hayamizu, Y. Aihara, S. Arai, C. G. Martinez, *J. Phys. Chem. B* **1999**, *103*, 519.

[56] M. Holz, X. Mao, D. Seiferling, A. Sacco, *J. Chem. Phys.* **1996**, *104*, 669.

[57] M. Iwahashi, Y. Yamaguchi, T. Kato, T. Horiuchi, I. Sakurai, M. Suzuki, *J. Phys. Chem.* **1991**, *95*, 445.

[58] M. Iwahashi, Y. Yamaguchi, Y. Ogura, M. Suzuki, *Bull. Chem. Soc. Jpn.* **1990**, *63*, 2154.

[59] S. Pickup, F. D. Blum, *Macromolecules* **1989**, *22*, 3961.

[60] M. I. Hrovat, C. G. Wade, *J. Magn. Reson.* **1981**, *44*, 62.

[61] M. I. Hrovat, C. G. Wade, *J. Chem. Phys.* **1980**, *73*, 2509.

[62] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*, Second, Revised and Enlarged ed., Wiley-VCH Press, Weinheim, Germany **2009**.

[63] Alvascience, alvaDesc version 2.0.10, 2022, https://www.alvascience.com. (accessed January 2022)

[64] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A* **1998**, *102*, 3762.

[65] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579.

[66] J. Donaldson, https://CRAN.R-project.org/package=tsne. (accessed May 2022)

[67] S. Kahwaji, M. B. Johnson, A. C. Kheirabadi, D. Groulx, M. A. White, *Sol. Energy Mater. Sol. Cells* **2017**, *167*, 109.

[68] J. A. Noël, L. M. LeBlanc, D. S. Patterson, L. Kreplak, M. D. Fleischauer, E. R. Johnson, M. A. White, *J. Phys. Chem. B* **2019**, *123*, 7043.

[69] J. Boonen, L. Veryser, L. Taevernier, N. Roche, K. Peremans, C. Burvenich, B. D. Spiegeleer, *J. Pharm. Anal.* **2014**, *4*, 303.

[70] K. Grodowska, A. Parczewski, *Acta Pol. Pharm.* **2010**, *67*, 3.

[71] D. J. C. Constable, C. Jimenez-Gonzalez, R. K. Henderson, *Org. Process. Res. Dev.* **2007**, *11*, 133.

[72] M. K. Riahi, I. A. Qattan, J. Hassan, D. Homouz, *AIP Adv.* **2019**, *9*, 055112.

[73] J. R. Errington, P. G. Debenedetti, *Nature* **2001**, *409*, 318.

[74] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, F. Paesani, *Chem. Rev.* **2016**, *116*, 7501.

[75] G. Guevara-Carrion, J. Vrabec, H. Hasse, *J. Chem. Phys.* **2011**, *134*, 074508.

[76] M. W. Mahoney, W. L. Jorgensen, *J. Chem. Phys.* **2001**, *114*, 363.

[77] P. Xu, E. B. Guidez, C. Bertoni, M. S. Gordon, *J. Chem. Phys.* **2018**, *148*, 090901.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.