AIP Advances

Machine learning symbolic equations for diffusion with physics-based descriptions

Cite as: AIP Advances 12, 025004 (2022); https://doi.org/10.1063/5.0082147
Submitted: 22 December 2021 • Accepted: 08 January 2022 • Published Online: 01 February 2022

Konstantinos Papastamatiou, Filippos Sofos and Theodoros E. Karakasidis

Machine learning symbolic equations for diffusion with physics-based descriptions

Cite as: AIP Advances 12, 025004 (2022); doi: 10.1063/5.0082147
Submitted: 22 December 2021 • Accepted: 8 January 2022 •
Published Online: 1 February 2022

Konstantinos Papastamatiou, Filippos Sofos,[a] and Theodoros E. Karakasidis

AFFILIATIONS

Condensed Matter Physics Laboratory, Department of Physics, University of Thessaly, 35100 Lamia, Greece

[a]Author to whom correspondence should be addressed: fsofos@uth.gr

ABSTRACT

This work incorporates symbolic regression to propose simple and accurate expressions that fit to material datasets. The incorporation of symbolic regression in physical sciences opens the way to replace "black-box" machine learning techniques with representations that carry the physical meaning and can reveal the underlying mechanism in a purely data-driven approach. The application here is the extraction of analytical equations for the self-diffusion coefficient of the Lennard-Jones fluid by exploiting widely incorporating data from the literature. We propose symbolic formulas of low complexity and error that achieve better or comparable results to well-known microscopic and empirical expressions. Results refer to the material state space both as a whole and in distinct gas, liquid, and supercritical regions.

I. INTRODUCTION

Symbolic Regression (SR) (very high order and advanced mathematics), the Machine Learning (ML) method used in the present research, tries to unveil hidden phenomena occurring at the nanoscale by incorporating simulation data. Based on the formulation derived from genetic programming,[1] SR can yield a mathematical expression that fits to a given dataset. After a period of idleness, the concept of SR has been reintroduced at times where the increased computational power and the application of ML algorithms in physical sciences have reached a mature level, allowing the wide applicability of this method.[2]

Following the vast deployment of material databases in the last decade, with several simulation and experimental data being available to the scientific community, statistical and data science has given a boost in material-related fields[3,4] by incorporating ML techniques.[5] More specifically, ML has been exploited to accelerate quantum mechanical (QM), molecular dynamics (MD) and continuum simulations, the construction of coarse-grained models, and the solution of physics-based partial difference equations.[6–10] ML can capture data behavior and reach predictions at only a fraction of the initial computational cost with comparable accuracy.[11] Nonetheless, the construction of a ML model is often an ambiguous task, since its inherent "black box" nature would make predictions difficult to rationalize and reveal the physical meaning behind the data.

This is where SR finds a field of applicability, as it can produce an analytic expression instead of a series of predictions[12,13] and, most importantly, without any prior assumption. The symbolic expression is derived from a potentially infinite search space that can contain any mathematical symbol or function. It makes no *a priori* linear or any other kind of extrapolation, as, for example, a common statistics method would do. The SR-derived expression can provide universality, generalization, and expand the predictions beyond the given domain without further recalibration. It also addresses the limited data point problem, as it is capable of producing results with fewer data points.[12] Symbolic regression usually employs neural networks architectures to boost computations and achieve higher prediction accuracy.[14–17]

In this work, SR is employed for extracting analytical expressions to describe the self-diffusion coefficient of the Lennard–Jones (LJ) fluid as a function of other physical properties, such as density and temperature. Taking in mind that experimental determination for diffusion coefficients is hard to achieve and, in parallel, computational complexity may be prohibitive for numerical determination,[18] empirical relations[19] and ML-derived models have also been proposed.[20] The diffusion coefficient controls the mechanism of mass transport in materials, and it would be of interest to propose a

generalizable model capable of predicting its value in various input conditions, overcoming long simulations and expensive experimental procedures.

On the other hand, generalization is hard to achieve in all fluid states (gas, liquid, high-dense liquid, and supercritical) and distinct investigation between states would be preferable. It has been observed that for low-density fluids, diffusion is facilitated, while at higher densities, fluid particles "feel" the confinement from their neighbors and spend more time oscillating around their initial sites before diffusing.[21] Furthermore, fluid behavior in the supercritical (SC) state and near the critical region, where density fluctuations become large, is still an open issue. The distinction of a SC fluid in gas- and liquid-like regions is a non-trivial task.[22] The formation of clusters around a solute molecule has been observed, restricting fluid thermal motion, and this is the main reason for the anomalous diffusion behavior at the supercritical region.[23] Experimental results have also shown that diffusion coefficients approach zero values at the vicinity of the critical point.[24] Nevertheless, simulation and experimental results are often contradicting whether an anomalous behavior of motion exists near the critical region or not.[25]

To overcome experimental and simulation restrictions, machine learning algorithms have been currently applied to develop models for the prediction of diffusion coefficients in supercritical systems[26] and across all fluid states.[27] It is a fact that ML could reveal the diffusion coefficient behavior at the microscopic level; however, a black-box consideration without a physical-based explanation would be incomplete. In Secs. II–IV, we describe a SR procedure to acquire an analytical prediction of the LJ fluid diffusion coefficient based entirely on literature data concerning diffusion data obtained at gas, liquid, and supercritical (SC) states. We propose general symbolic expressions that capture fluid behavior on the entire dataset, and, in a broader sense, we extend our calculations specifically for every fluid state, focusing on the special characteristics of every region. The selection of the equations, apart from decreased error metrics and low complexity, has been made on a physics-based analysis, aiming to bind mathematics with hidden aspects of material behavior.

II. METHODS

A. Symbolic regression flowchart

Figure 1 presents a general view of the process followed in this study. The diffusion data enter the calculation flow in two ways: all-region dataset corresponding to all fluid states and three different datasets corresponding to the gas, liquid, and SC states [Fig. 1(a)]. Symbolic regression computational cells are working in parallel and provide raw symbolic expressions that fit on the input data (see the supplementary material). There are a huge unordered number of equations that need to be cleared out before proceeding. This is done through a decision step, which involves a physics-based explanation of the results, and the final equations are selected, in terms of low complexity and minimum Mean Squared Error (MSE), as will be shown later.

The SR cell derives a mathematical expression selected from a wide function space and is tested on the input dataset. Two simple examples are given here in Fig. 1(b). After N computational steps, the algorithm selects those expressions that exhibit lower MSE along with reduced complexity. The term complexity is directly proportional to the number of calculation nodes in the tree-like structure. An expression of increased complexity is more prone to overfitting, so the choice of simple and accurate expressions is preferable. To gain insight on the physics-based decision step, we also present an example of diffusion behavior in fluids in Fig. 1(c), where the mechanism of fluid atom jump in a new position in a dense and a sparse environment, respectively, is shown. It is harder for an atom to attain a new position (diffuse) in a dense fluid due to barriers induced by neighboring particles, while diffusion is facilitated in sparse fluids where neighboring particles are less.

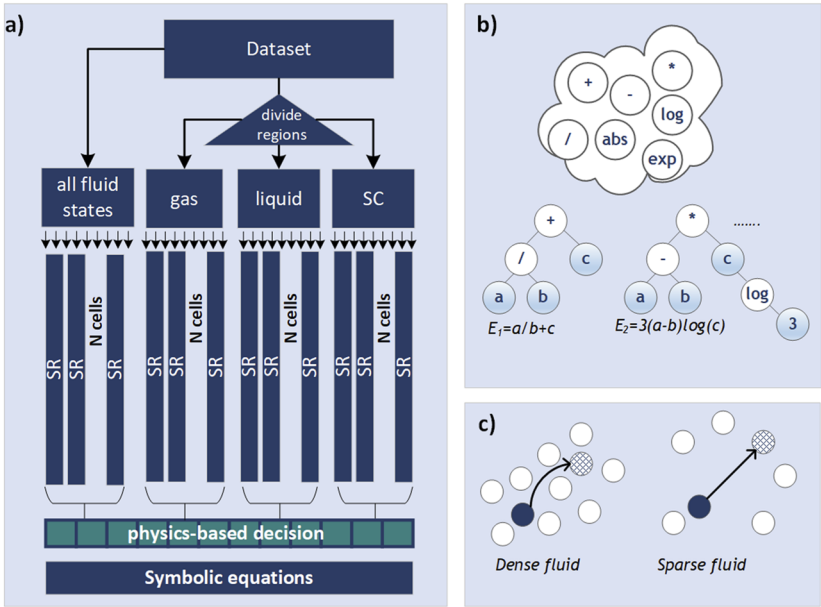

FIG. 1. (a) The proposed symbolic regression procedure, (b) the SR cell, and (c) fluid atom jump in a new position in a dense and a sparse environment, respectively.

B. Molecular dynamics

The molecular dynamics (MD) simulation method has been incorporated for the calculation of the diffusion coefficients throughout the input dataset. MD involves atomic interactions in a LJ fluid, which are commonly described by the potential Φ_{LJ}. Φ_{LJ} is a function of the distance r between the particles, and it is defined by[28]

$$\Phi_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{1}$$

The LJ system is described by dimensionless variables scaled by σ and ε, where ε is the depth of the potential well and σ is the inter-atomic separation where the potential energy is zero. This set of variables includes temperature T^*, density ρ^*, distance r^*, diffusion D^*, and time t^*, which are defined in the following equations:

$$T^* = \frac{k_B T}{\varepsilon}, \quad \rho^* = \frac{N\sigma^3}{V}, \quad r^* = \frac{r}{\sigma}, \quad D^* = \frac{D\sqrt{m/\varepsilon}}{\sigma}, \quad t^* = \frac{t\sqrt{m/\varepsilon}}{\sigma}, \tag{2}$$

where k_B is the Boltzmann constant, σ and ε are the LJ parameters, N is the number of particles, V is the volume of the system, and m is the mass of the atom in pure system.

The dataset incorporated here is obtained from Ref. 20 and consists of 927 data points from 17 literature sources. Significant dataset properties are summarized in Fig. 2. More details on the dataset can be found in the supplementary material. The phase diagram [Fig. 2(a)] covers a wide range of the fluid region, from gas to the high-density liquid. D^* is inversely proportional to ρ^* [Fig. 2(d)] for constant T^* and linearly proportional to T^* [Fig. 2(c)] for constant ρ^*. From the correlation diagram in Fig. 2(b), it is evident that the two inputs (T^* and ρ^*) are not significantly correlated, but there is a negative correlation between D^* and ρ^*, which represents the physical tendency of reduced atomic mobility as density increases.

To justify the use of SR to distill an analytical formula, we have to note that our problem is low dimensional from $R^2 \rightarrow R$ and both features and labels are dimensionless. It has been argued that low-dimensional problems are better fitted to SR techniques since this diminishes the search space and allows faster algorithm execution.[31] In search of a simple formula, SR could address

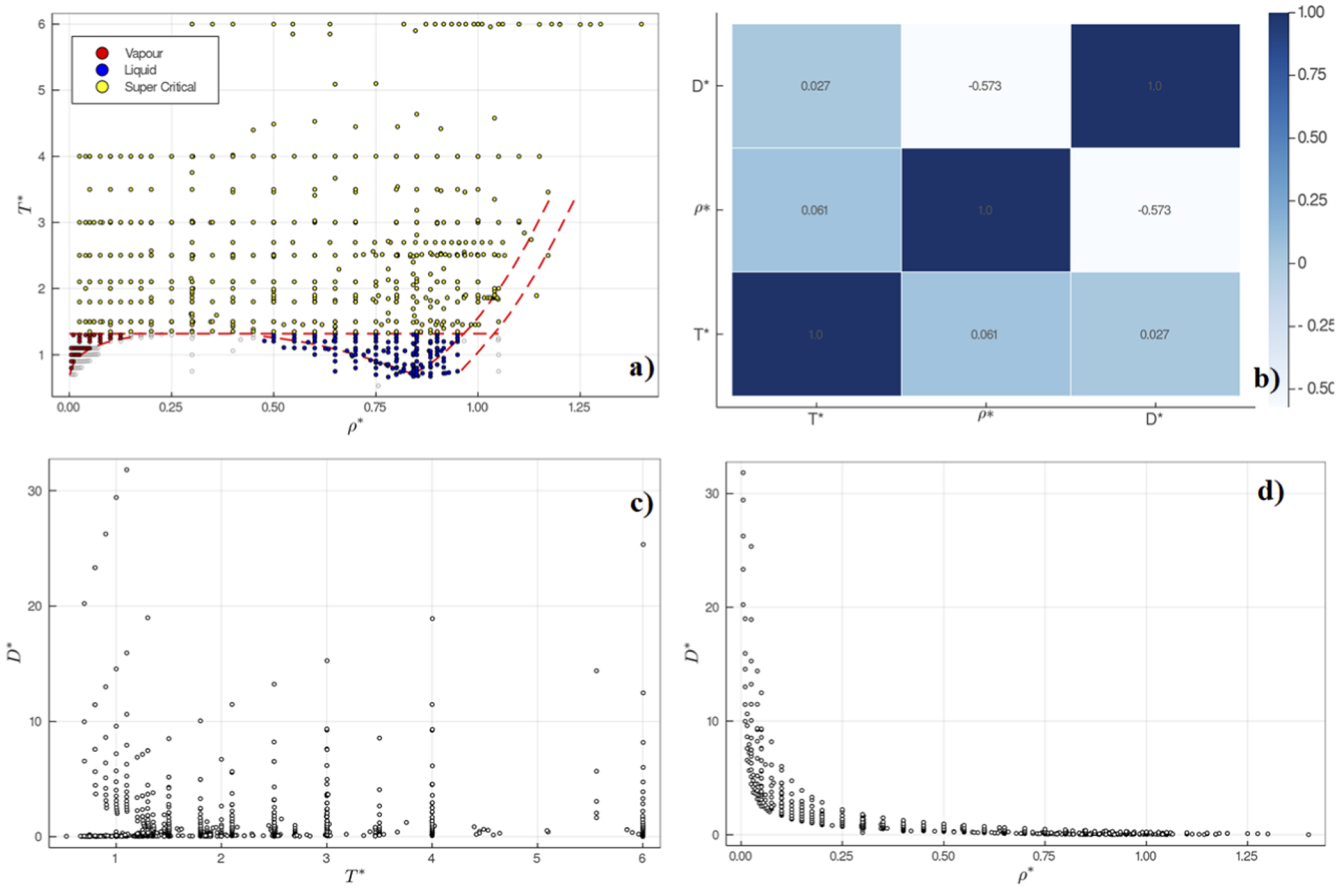

FIG. 2. Characteristics of the dataset employed. (a) The $\rho^* - T^*$ phase diagram. The red lines denote the Vapor–Liquid equilibrium (VLE),[29] the freezing line (FL),[30] and the horizontal line is the critical temperature line, T_c. Distinction in gas, liquid, and supercritical states is depicted. (b) Correlation diagram between the two inputs, $\rho^* - T^*$, and the output D^*, (c) partial plot of D^* vs T^*, and (d) partial plot of D^* vs ρ^*.

overfitting issues and this is especially significant in physical sciences and material-related applications.[32] It is a fact that laws of physics are generally mathematically simple and physical explanation can be made easier.[33,34]

The main objective in this work is to suggest a simpler formula for the diffusion coefficient calculation, without having to incorporate the microscopic state of the system, as expressed by the Green–Kubo equation [Eq. (3)], Einstein's relation [Eq. (4)],[18] or complex empirical formulas [Eq. (5)],[19]

$$D = \frac{1}{3(N-1)} \sum_{i=1}^{N} \int_0^{\infty} \langle v_i(t) \cdot v_{i(t_0)} \rangle dt, \tag{3}$$

$$D = \lim_{t\to\infty} \frac{1}{6(N-1)} \sum_{i=1}^{N} \frac{d}{dt} \langle \left[r_i(t) - r_{i(t_0)} \right]^2 \rangle, \tag{4}$$

$$D^* = \frac{3}{8\rho^*} \sqrt{\frac{T^*}{\pi}} \times A \times B, \tag{5}$$

where

$$A = \left(1 - \frac{\rho^*}{\alpha (T^*)^b}\right) \left[1 + (\rho^*)^c \left(\frac{P1(\rho^* - 1)}{P2(\rho^* - 1) + (T^*)^{P3+P4\rho^*}}\right)\right], \tag{6}$$

$$B = \exp\left(-\frac{\rho^*}{2T^*}\right), \tag{7}$$

and $a, b, c, P1, P2, P3, P4$, a set of parameters given in Ref. 19.

Symbolic regression libraries (usually written in Python or Julia) are freely available in the literature.[35–40] In this work, the package created by Cranmer *et al.*[15] has been exploited, with the criterion of having achieved minimum mean squared error (MSE) and complexity (Comp.). The loss function is also calculated by the Logistic Distance Loss function, $L(r)$, which is strictly convex, and Lipschitz continuous

$$L(r) = -\ln \frac{4e^r}{(1+e^r)^2}. \tag{8}$$

The term complexity refers on the number of nodes used in the binary tree that constitute the symbolic expression. The complexity spans over the range $1 \le \text{Comp.} \le 20$, i.e., the expression ranges from a simple constant to a 20-node expression. The SR algorithm accepts a set of basic operators, $A = \{+, -, *, /, pow\}$, and functions, $B = \{\exp, \log\_abs, sqrt\_abs\}$, as inputs. Figure 1(b) presents an example of how the SR creates a node structure for two simple functions.

III. RESULTS

A. Symbolic equations across all fluid states

1. Selection of an alternative diffusion coefficient equation

Diffusion coefficient data corresponding to all fluid states have entered the symbolic regression process [Fig. 1(a)] and output an expression depending only on the reduced quantities of density (ρ^*) and temperature (T^*). As the SR problem is a multi-objective optimization problem, there is no unique solution rather than a set of solutions forming a line called Pareto front. The formulas and their parameters, and metrics such as complexity (Comp.), coefficient of determination squared (R^2), mean squared error (MSE), root MSE (RMSE), mean absolute error (MAE), average absolute deviation (AAD), and the correlation coefficient (Corr.) for all data points are presented in Table I. Training and testing datasets have been acquired by the initial dataset by a partitioning factor of 80% and 20%, respectively, shuffled and split randomly.

To obtain the symbolic expression, the SR algorithm is applied on the diffusion dataset and 20 different equations of increasing complexity are extracted. This process is repeated for about 100 times (100 parallel runs) to avoid overfitting and randomness. A different set of equations is obtained for each independent algorithm execution (see all proposed equations in the supplementary material). However, this process has revealed a pattern; there are some equations that keep on appearing on the results, despite the tree-like randomness and the mutation of the method. It has been observed that three formulas are continuously appearing in the results, F1, F2, and F3, as presented in Table I. Their selection is based on the fact that they have appeared more than one time, and they are interpretable in physical terms.

The first formula (F1) appeared in the Pareto front with low complexity (this equation ranked fifth in the complexity ordering) shows that D^* is directly proportional to T^* and inversely proportional to ρ^*. Furthermore, the value of a reveals that temperature has only a small contribution to diffusion. An obvious disadvantage

TABLE I. All-region SR-extracted formulas, F1, F2, and F3, for the self-diffusion coefficient. The average values of Comp., R^2, MSE, RMSE, MAE, AAD, and Corr. are obtained by both training and testing sets.

	Formula	Comp.	R^2	MSE	RMSE	MAE	AAD	Corr.
F1	$\hat{D}^* = \alpha \cdot \frac{T^*}{\rho^*}$ $0.10 \le a \le 0.15$	5–6	0.970 9	0.2738	0.5233	0.2771	182.669	0.9896
F2	$\hat{D}^* = -\alpha_1 \cdot T^* + \alpha_2 \cdot \frac{T^* \cdot \alpha_3^{T^*}}{\rho^*}$ $a_1 = 0.070\,497\,77, \alpha_2 = 0.153\,970\,642, \alpha_3 = 0.941\,462\,04$	13–16	0.999 38	0.006 04	0.05704	0.05704	60.453	0.9997
F3	$\hat{D}^* = \frac{\alpha_1 (T^* - a_2)}{\rho^* ((\rho^* + a_3)(T^* + \rho^*) - \rho^* + \alpha_4)}$ $\alpha_1 = 0.491\,475, \alpha_2 = 0.058\,818\,2, \alpha_3 = 0.257\,741, \alpha_4 = 2.948\,75$	15–19	0.999 2	0.007 32	0.05946	0.05946	93.6713	0.99964

of F1 is the high error achieved for all the metrics used, although R^2 reaches a satisfying 97%. An inherent advantage, though, is that a simple formula usually achieves no overfitting, since the accuracy is balanced with complexity.[41]

The second formula (F2) is ranked between 13 and 16 in the complexity ordering results. It is observed that, as complexity increases, all error metrics have improved compared to F1. By allowing larger complexity, we acquire F3, where statistics are greatly

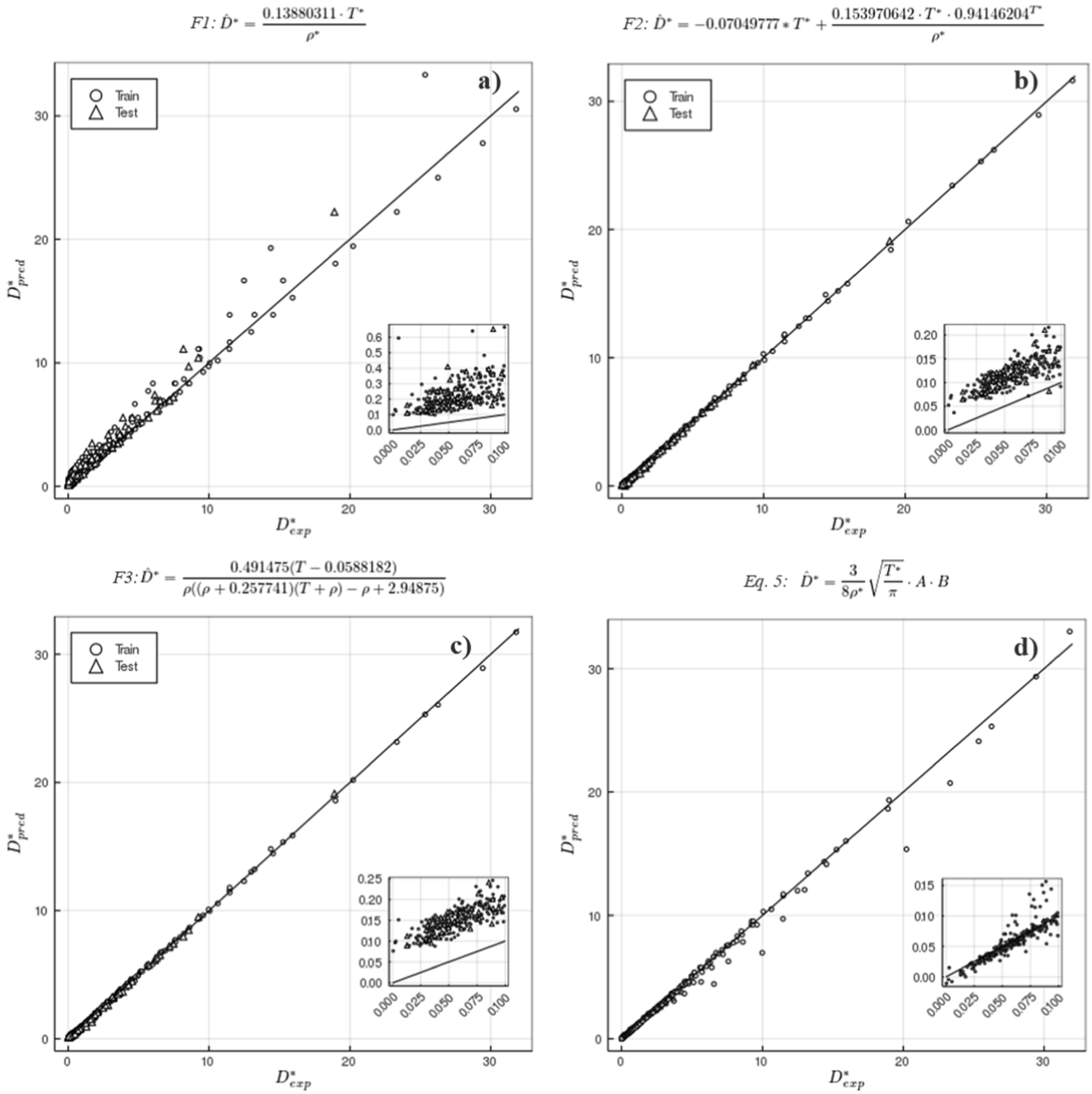

FIG. 3. Observed vs equation-predicted values for the proposed diffusion formulas. Data points distinguish between training and testing datasets. The 45° identity line denotes the perfect match. (a) F1 fitted on the available data achieves satisfying prediction metrics, (b) better prediction achieved for expression F2, (c) even better prediction for the increased complexity F3 expression, and (d) fitting on the empirical Eq. (5).[19]

TABLE II. Comparison with existing methods. The proposed SR-extracted expressions are compared with literature results from machine learning methods[20] and empirical relations.[19]

Model	MSE	Std. dev.	Avg. corr.
F1	0.273 8	±0.3	0.989 6
F2	0.006 04	±0.063	0.999 7
F3	0.007 32	±0.059	0.999 64
RF	0.18	±0.21	0.996 24
ANN[20]	0.001 2	±0.001	0.999 94
Equation (5) [19]	0.067	±0.077	0.996 85

improved compared to F1, but they are comparable to F2, having only a slightly better R^2 value. Nevertheless, complexity is still small compared to the empirical Eq. (5), which has been widely accepted for self-diffusion calculations in the past.

The identity plots of Fig. 3 show that the accuracy increases as the symbolic complexity increases. The simple form of F1 fitted on the available data achieves satisfying prediction metrics [Fig. 3(a)]. Inconsistencies are observed for lower diffusion values, as shown in the inset. For F2 and F3, the observed and predicted values have shown a perfect fit for both the training and the testing sets, but deviations still exist at lower D^* values [Figs. 3(b) and 3(c)]. The identity plot for Eq. (5) is also shown for comparison in Fig. 3(d). Minor deviations from the identity line are spread throughout all data points, as shown in the inset. It is evident that at least the proposed F2 and F3 perform better than Eq. (5) in the applied diffusion dataset.

In Table II, a comparison is made between F1 and F3 statistics with results acquired from the empirical Eq. (5). We also extend the comparison to other ML-based methods, such as a simple random forest (RF) and a complex artificial neural network (ANN) with two hidden layers,[20] for metrics, such as the MSE, the standard deviation, and the average correlation coefficient. More specifically, the results suggest that F2 and F3 perform better than the empirical Eq. (5) and the simple RF approach in all statistics while keeping lower complexity. F2 and F3 performance are comparable to the complex ANN.[20]

From a statistical point of view, it is evident that symbolic regression is capable of providing accurate expressions with controllable complexity, which performs equally good or better than traditional methods. The superiority of the method, however, is the fact that all formulas proposed can reveal the physical meaning of the property.

The behavior of diffusion on fluid density in the proposed formulas F1 and F3 follows the rule as $\rho^* \to 0, D^* \to \infty$. This result is consistent with the expected physical behavior, since, as density approaches zero (near gas state), diffusion increases. Similar behavior has also been found for F2, as $\rho^* \to 0, D^* \to \infty$, i.e., the diffusion tends to infinity. It has to be emphasized, though, that a zero-density fluid has no physical meaning. Moreover, for all proposed equations, in the extreme case where $T \to 0$, we obtain $D^* \to 0$ (for $\rho^* \neq 0$) as expected, since at zero temperature, there is no particle movement.

2. Physical correspondence

Let us turn our attention now to each proposed equation performance on the original dataset for various density (ρ^*) values. In Fig. 4, data portions (T^*, D^*) of the original dataset for various ρ^* values are plotted, along with the corresponding equations F1–F3. As far as low densities ($0.005 < \rho^* < 0.025$) are concerned, all three equations present very good agreement with simulation

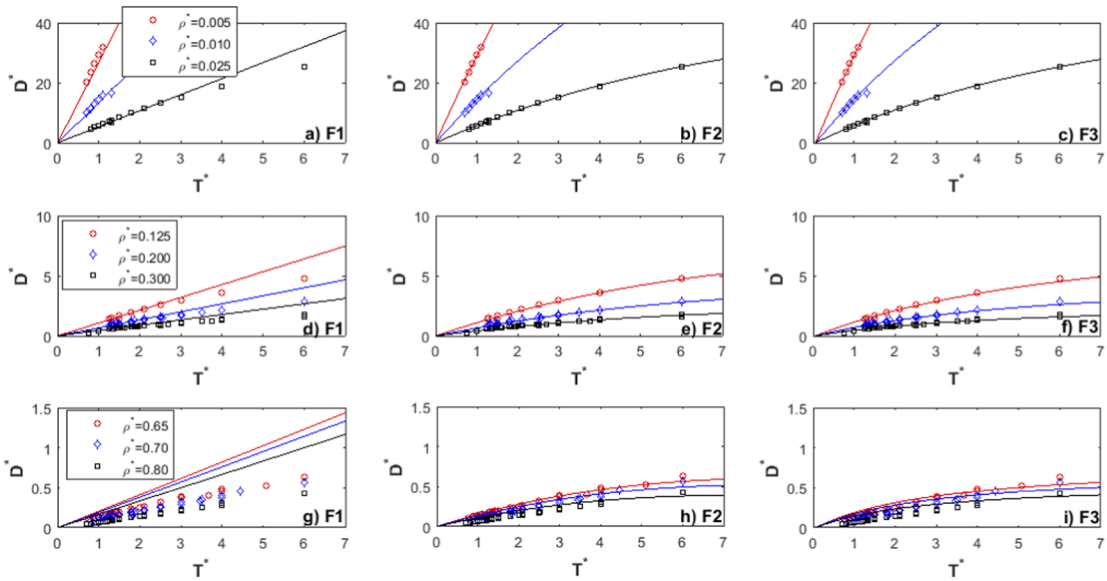

FIG. 4. All-region dataset points (T^*, D^*) for various ρ^* values. In low-density conditions, fitted on the (a) F1 equation, (b) F2 equation, and (c) F3 equation. In medium-density conditions, fitted on the (d) F1 equation, (e) F2 equation, and (f) F3 equation. In high-density conditions, fitted on the (g) F1 equation, (h) F2 equation, and (i) F3 equation. Data points are taken from the respective region of the dataset and straight lines refer to equation values for the respective ρ^* range.

data, as we can see from Figs. 4(a)–4(c). In the case of medium-range densities ($0.125 < \rho^* < 0.3$), F2 and F3 have also shown good agreement, but, on the other hand, F1 presents large deviations, especially at high temperatures [Figs. 4(d)–4(f)]. This is an indication that there are more complex mechanisms that take place and contribute to diffusion coefficient behavior, which are not being represented by the simple form of F1. As far as higher densities are concerned ($0.65 < \rho^* < 0.80$), in Figs. 4(g)–4(i), it is obvious that F1 has seriously over-estimated the diffusion coefficient values on the simulation data. Nevertheless, F2 and F3 perform better than F1, with F2 performing even better than F3. To summarize, one could say that the self-diffusion coefficient of the LJ fluid is accurately reproduced by the proposed equation F2 along various density and temperature values for all fluid states.

From physical grounds, it is anticipated that for low-density conditions, where fluid particles have more space to move, diffusion is enhanced, and this is further facilitated at higher temperatures. Therefore, all proposed equations reproduce the diffusion dataset satisfyingly. However, as density increases and the solid phase is approached, fluid particles "feel" the confinement effect, and, thus, they spend more time oscillating around sites of residence before being allowed to diffuse,[21] notwithstanding the fact that temperature provides energy to enhance diffusion jumps.

These two different mechanisms are reproduced by the two terms in equation F2. Increased density leads to faster "collision" between atoms since they have less space to move freely, while, in contradistinction, the increased temperature, although it provides more energy for diffusion jumps, it also leads to larger vibration amplitude, blocking these jumps. Moreover, diffusion events diminish, and vibrational motions lead to faster loss of correlations. This has also been observed in Ref. 42, as an effect of higher temperatures. In other words, equation F2 represents the two aforementioned mechanisms and further takes into account the complex dependence on temperature through the term $\alpha_3^{T^*}$ with values $\alpha_3^{T^*} < 1$. These two different behaviors are also observed in Ref. 19.

Moreover, it is worth mentioning that the fluid density effect on the self-diffusion coefficient is stronger than temperature and this is expressed by all three proposed symbolic formulas, a result that is also consistent with our prior knowledge.[43]

To gain insight into the hidden physical behavior, we also argue on possible modes of fluid atom movement under specific conditions [as illustrated in Fig. 1(c)]. At high densities, atoms are bound to frequent "collisions" since they are more densely localized, resulting in smaller jumps than they would do at lower densities. In addition, they perform a kind of vibrating motion before jumping to the new location. However, temperature increase has a two-fold effect; more frequent jumps and increased vibrations of those atoms that do not diffuse. As a consequence, vibration increase would minimize the possibility of the remaining atoms jumping, and it is believed that the negative term in F2 ($-\alpha_1 \cdot T^*$) epitomizes such an effect. On the other hand, the second term ($\alpha_2 \cdot \frac{T^* \cdot \alpha_3^{T^*}}{\rho^*}$) represents the inverse dependence of density and temperature and, in parallel, reveals a complex temperature behavior through the linear (T^*) and the nonlinear ($\alpha_3^{T^*}$) terms. As fluid density is reduced, it is easier for atoms to move, there are fewer "collisions," and an increase in temperature favors more diffusive jumps and less vibrational behavior compared to the high-density case. It should also be pointed out that the negative term in F2 is relatively small for low temperatures at low densities.

B. Regional symbolic equations for gas, liquid, and supercritical fluid

1. Selection of regional diffusion coefficient equations

It is a fact that fluid data points investigated in Sec. III A cover a range from vapor to liquid, vapor–liquid coexistence, and dense liquid. Notwithstanding the accuracy of the equations extracted that seem to capture the $\rho^* - T^*$ effect on D^* in all fluid states, we further examine the application of SR in distinct fluid regions. Data separation follows the proposed scheme in Ref. 29.

For the region denoted as gas (or vapor), Fig. 5(a), formula F4 obtained is similar to F1, with parameter $a = 0.145\,505\,79$. The difference here is that all error metrics have been greatly improved

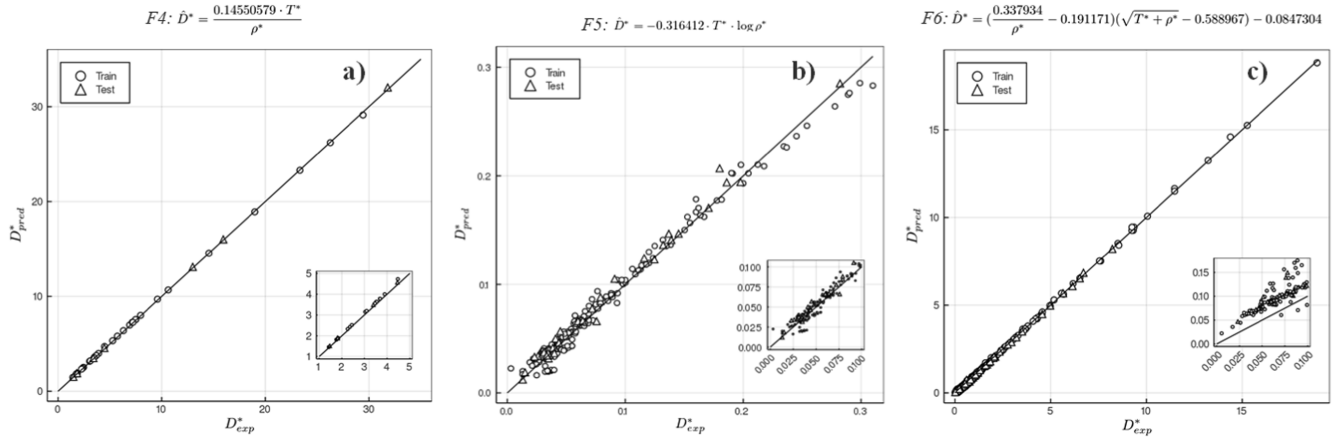

FIG. 5. Observed vs equation-predicted values for the proposed regional diffusion formulas. Data points distinguish between training and testing datasets. The 45° identity line denotes the perfect match. (a) F4 fitted on the LJ gas data and (b) F5 fitted on liquid data. Deviations shown are expressed by the AAD value (see Table III) and (c) F6 fit for the supercritical fluid.

TABLE III. Regional SR-extracted formulas for the self-diffusion coefficient. A different expression corresponds to gas, liquid, and supercritical regions. The values of Comp., R^2, MSE, RMSE, MAE, AAD, and Corr. are the average values obtained by both training and testing sets.

	Formula	State	Comp.	R^2	MSE	RMSE	MAE	AAD	Corr.
F4	$\hat{D}^* = \alpha \cdot \frac{T^*}{\rho^*}$ $a = 0.145\,505\,79$	Gas	5	0.9998	0.0118	0.1086	0.0939	2.120	1.0
F5	$\hat{D}^* = -a \cdot T^* \cdot \log \rho^*$ $a = 0.316\,412$	Liquid	6	0.9857	0.0001	0.0078	0.0061	14.479	0.9932
F6	$\hat{D}^* = \left(\frac{a_1}{\rho^*} - a_2\right)\left(\sqrt{T^* + \rho^*} - a_3\right) - a_4$ $a_1 = 0.337\,934$, $a_2 = 0.191\,171$, $a_3 = 0.588\,967$, $a_4 = 0.084\,730\,4$	SC	14	0.9996	0.0016	0.0396	0.0300	16.391	0.9998

(see Table III) compared to the all-region analysis and the respective identity plot shows a perfect fit [Fig. 5(a)]. F4 is finely fitted on the LJ gas data, even for lower diffusion values, as shown in the inset. Diffusion at gas state is generally higher than other states and is inversely dependent on fluid density, ρ^*, and linearly dependent on T^*.

A low-complexity formula F5 has been also extracted for the liquid/dense liquid region. Diffusion presents a decaying logarithmic behavior as ρ^* increases, along with a small linear dependence on T^*. Both F4 and F5 present decaying behavior as ρ^* increases. As $\rho^* \to 0$ (gas state), diffusion fits on the ratio $1/\rho^*$, and F4 successfully reproduces the values of the calculated diffusion coefficients. As $\rho^* \to 1$ (liquid state), diffusion is better being captured by $\log(\rho^*)$, and F5 is the best choice for the LJ liquid [Fig. 5(b)]. In terms of R^2, F5 gives a training/testing average of 0.9857, while for F4, it is $R^2 = 0.9998$. These values, although both satisfyingly high, may differ for various reasons. One obvious reason is the fact that D^* ranges from about $1.42 \to 31.81$ in F4 and from about $0.003 \to 0.310$ in F5. Thus, data points corresponding to the liquid state may be prone to noise and minor inconsistencies may exist.

As far as the supercritical region is concerned, equation F6 is of increased complexity compared to F4 and F5 (Table III), though smaller than the all-region formulas F2 and F3, since it has to capture more complex phenomena that characterize a SC fluid. Gas/liquid coexistence and near-critical point phenomena affect diffusion coefficient values. Taking a closer look at the terms that constitute F6, it is observed that the SR algorithm has also provided the ratio $1/\rho^*$, as in F4 and Eq. (5) [19] cases. Similar decreasing behavior for D^*, as stated before, is obtained by the $-\log(\rho^*)$ term in F5. This is evidence that, apart from other mechanisms, diffusion is always decreased in high-density fluids, and, from a computational point of view, the SR algorithm adheres to physical intuition. After all, the interesting point here is that F6 succeeds in reproducing accurate data output to the applied dataset, as shown in Fig. 5(c).

2. Physical correspondence—Regional

In Fig. 6, data portions (T^* and D^*) of the original dataset for various ρ^* values are plotted, along with the corresponding equations F4–F6. For the gas state data region [Fig. 6(a)], where density is low (quite smaller than unity), F4 presents fine agreement with simulation data. Diffusion is greatly facilitated as density reduces and this is represented by the dependence on $1/\rho^*$. There is also a direct diffusion coefficient increase when the temperature rises. In the aftermath, this has a primitive effect on fluid kinetic energy, which, in turn, contributes to the diffusion process. Therefore, equation F4 involves D^* being linearly dependent on T^* for constant density values.

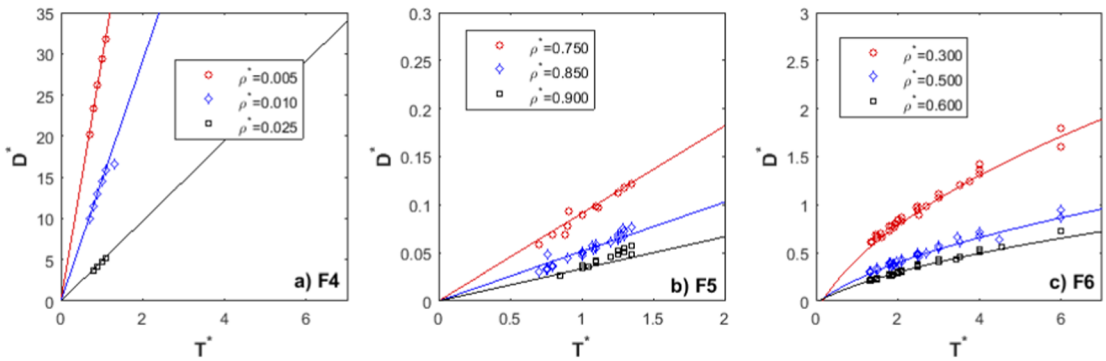

FIG. 6. Regional dataset points (T^*, D^*) for various ρ^* values: (a) in low-density conditions (gas state), fitted on the F4 equation, (b) in high-density conditions (liquid/dense liquid), fitted on the F5 equation, and (c) in medium-density conditions, above the critical temperature T_c (SC state), fitted on the F6 equation. Data points are taken from the respective region of the dataset, and straight lines refer to equation values for the respective ρ^* range.

The liquid/dense liquid region, as described by F5, corresponds to significantly lower D^* values compared to the gas state. Temperature (as it remains below the critical value T_c) has a linear contribution to diffusion since it corresponds to increased kinetic energy. However, fluid density seems to be the dominant effect on D^* since the increased density reduces atomic mobility through atom jumps [see Fig. 1(c)]. As ρ^* increases, there is significant reduction on D^* and this is successfully reflected by the logarithmic term $-\log(\rho^*)$ in F5.

The complex nature of the supercritical fluid is captured by equation F6 [Fig. 6(c)]. We expect no clear distinction between liquid and gas phases in supercritical states, resulting in relatively high diffusion coefficients (lower than the gas state but significantly higher than the liquid state) at high temperatures and pressures. Thus, in F6, we obtain a complex dependence on ρ^*, resulting from the inverse term $1/\rho^*$ which might lead to higher D^* when fluid density decreases, and an even more complex dependence from the term ρ^* inside the square root $\sqrt{T^* + \rho^*}$ (along with T^*). Fluid temperature affects D^* in a more straightforward way, though not as important, as T^* appears only inside the $\sqrt{T^* + \rho^*}$ term.

IV. DISCUSSION

Distilling natural laws from a simulation/experimental dataset, from a near-infinite parameter space, is made possible through symbolic regression. This would have a great impact on material and relevant physical sciences, where research is defined by accurately harnessing a multiple parametric environments. A symbolic expression would reveal hidden material properties and provide the pathway to discover new. In Secs. II and III, we applied SR to a diffusion coefficient simulations dataset and proposed analytical equations that can accurately reproduce the output.

The symbolic regression methodology incorporated here goes through a large number of realizations performed in parallel [Fig. 1(a)], and this can be considered an essential step toward uncertainty quantification.[44] A proposed equation has been selected from a pool of mathematical expressions, which are nearly or exactly similar to each other (see the output equations in the supplementary material). The final form of the proposed equation that represents the diffusion coefficient for the LJ fluid is derived in a way that mimics the principle of small variations. Small variations constitute a neighborhood of a potential solution, and the solution is searched for within this neighborhood.[45]

If the fluid state space is considered uniform, keeping in mind all induced advantages and disadvantages that come from a universal expression, the suggested formulas (F1–F3) are relatively simple, they achieve different error metrics while balancing with complexity, and can reduce the computational time needed to calculate the self-diffusion of LJ fluids. Comparisons made with current and well-posed literature results have guaranteed the validity of this method. To epitomize the results, the proposed equation F2 seems to be on firm ground since it can reveal hidden information on the underlying physical mechanisms. It reproduces fluid behavior, where diffusion is enhanced at low-density conditions and is further facilitated at higher temperatures.

However, considering the state space as a uniform means that the three main states of fluids (gas, liquid, and supercritical) and their possible combinations (e.g., gas–liquid or liquid–crystal coexistence) are reproduced by one equation that embeds all the differences in fluid behavior. Thus, further investigation has been made to examine possible deviations from the extracted expressions, by providing a respective expression for the fluids in its gas, liquid, and SC states.

Diffusion at the gas state achieves larger values, in general, compared to liquid or near-solid states. A simple equation of the form $F4 = f\left(T^*/\rho^*\right)$ reproduces the mechanism of gas diffusion accurately. The inherent advantage of this approach is the low complexity value, along with small error metrics (R^2, MSE, MAE, and AAD). For the diffusion coefficient values in LJ liquids, the proposed formula gives $F4 = f\left(-T^* \log \rho^*\right)$. This formula has also captured the fact that D^* is decreased as ρ^* increases. The SC region has given a more complex expression, which, if the multi-factorial region's behavior is taken into account, F6 performs well, with reduced error metrics.

V. CONCLUSIONS

In conclusion, the symbolic formulas proposed to provide a new direction to a well-studied physical problem and can be incorporated to obtain models that do not only extrapolate results in other conditions but, more importantly, derive equations from scratch that accurately represent the underlying physical processes. Results shown here for the LJ fluid will be upscaled to real materials, where more parameters might enter the calculations. We believe that establishing a mechanism for discovering new functionals with improved accuracy using only a set of data points, even without prior assumption on the form of the relations, will be the next focal point of investigation in physical and material sciences.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional details about dataset statistics, details on training and testing datasets, LJ fluid region division, and raw equations extracted from the symbolic regression algorithm.

ACKNOWLEDGMENTS

This study was supported by the Center of Research Innovation and Excellence of University of Thessaly, funded by the Special Account for Research Grants of University of Thessaly under Project CAMINOS (No. 5600.03.08.03).

AUTHOR DECLARATIONS

Conflict of Interest

I have no conflicts of interest to disclose.

Author Contributions

K.P. contributed to dataset pre-processing, programming, writing, and equation extraction. F.S. conceptualized and designed the study, retrieved the dataset, and wrote and edited the manuscript. T.E.K. discussed the results, extracted physical meaning, and reviewed the manuscript.

DATA AVAILABILITY

The dataset incorporated in this study has been downloaded from https://aip.scitation.org/doi/suppl/10.1063/5.0011512. Symbolic expressions extracted from the methodology presented are available in the supplementary material.

REFERENCES

[1] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992).
[2] S.-M. Udrescu and M. Tegmark, Sci. Adv. **6**, eaay2631 (2020).
[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, Sci. Data **3**, 160018 (2016).
[4] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, and J.-C. Zhao, npj Comput. Mater. **5**, 41 (2019).
[5] R. K. Vasudevan, K. Choudhary, A. Mehta, R. Smith, G. Kusne, F. Tavazza, L. Vlcek, M. Ziatdinov, S. V. Kalinin, and J. Hattrick-Simpers, MRS Commun. **9**, 821 (2019).
[6] M. Wang, T. Wang, P. Cai, and X. Chen, Small Methods **3**, 1900025 (2019).
[7] M. Raissi, P. Perdikaris, and G. E. Karniadakis, J. Comput. Phys. **348**, 683 (2017).
[8] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Annu. Rev. Phys. Chem. **71**, 361 (2020).
[9] K. Kontolati, D. Alix-Williams, N. M. Boffi, M. L. Falk, C. H. Rycroft, and M. D. Shields, Acta Mater. **215**, 117008 (2021).
[10] N. Asproulis and D. Drikakis, J. Comput. Theor. Nanosci. **6**, 514 (2009).
[11] M. Frank, D. Drikakis, and V. Charissis, Computation **8**, 15 (2020).
[12] Y. Wang, N. Wagner, and J. M. Rondinelli, MRS Commun. **9**, 793 (2019).
[13] B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan, and W.-J. Yin, Nat. Commun. **11**, 3513 (2020).
[14] S. Kim, P. Y. Lu, S. Mukherjee, M. Gilbert, L. Jing, V. Ceperic, and M. Soljacic, IEEE Trans. Neural Network Learn. Syst. **32**, 4166 (2020).
[15] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, in *Proceeding Advances in Neural Information Processing Systems 33 - NeurIPS 2020* (Curran Associates Inc., Vancouver, Canada, 2020), p. 17429.
[16] J. Zhong, L. Feng, W. Cai, and Y.-S. Ong, IEEE Trans. Syst. Man Cybern. Syst. **50**, 4492 (2020).
[17] S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark, in *Proceeding Advances in Neural Information Processing Systems 33 - NeurIPS 2020* (Curran Associates Inc., Vancouver, Canada, 2020), p. 4860.
[18] K. Meier, A. Laesecke, and S. Kabelac, J. Chem. Phys. **121**, 3671 (2004).
[19] Y. Zhu, X. Lu, J. Zhou, Y. Wang, and J. Shi, Fluid Phase Equilib. **194–197**, 1141 (2002).
[20] J. P. Allers, J. A. Harvey, F. H. Garzon, and T. M. Alam, J. Chem. Phys. **153**, 034102 (2020).
[21] T. E. Karakasidis and A. B. Liakopoulos, Phys. Stat. Mech. Appl. **333**, 225 (2004).
[22] V. V. Brazhkin, Y. D. Fomin, A. G. Lyapin, V. N. Ryzhov, and E. N. Tsiok, J. Phys. Chem. B **115**, 14112 (2011).
[23] H. Higashi, Y. Iwai, and Y. Arai, Ind. Eng. Chem. Res. **39**, 4567 (2000).
[24] H. Nishiumi and T. Kubota, Fluid Phase Equilib. **261**, 146 (2007).
[25] A. N. Drozdov and S. C. Tucker, J. Chem. Phys. **114**, 4912 (2001).
[26] J. P. S. Aniceto, B. Zêzere, and C. M. Silva, J. Mol. Liq. **326**, 115281 (2021).
[27] J. P. Allers, F. H. Garzon, and T. M. Alam, Phys. Chem. Chem. Phys. **23**, 4615 (2021).
[28] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 2017).
[29] S. Stephan, M. Thol, J. Vrabec, and H. Hasse, J. Chem. Inf. Model. **59**, 4248 (2019).
[30] G. C. McNeil-Watson and N. B. Wilding, J. Chem. Phys. **124**, 064504 (2006).
[31] D. Wadekar, F. Villaescusa-Navarro, S. Ho, and L. Perreault-Levasseur, arXiv:2012.00111 Astro-Ph Physics (2020).
[32] C. Loftis, K. Yuan, Y. Zhao, M. Hu, and J. Hu, J. Phys. Chem. A **125**, 435 (2021).
[33] P. A. K. Reinbold, L. M. Kageorge, M. F. Schatz, and R. O. Grigoriev, Nat. Commun. **12**, 3219 (2021).
[34] V. Tavanashad, A. Passalacqua, and S. Subramaniam, Int. J. Multiphase Flow **135**, 103533 (2021).
[35] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, J. Mach. Learn. **13**, 2171 (2012).
[36] T. Worm and K. Chiu, in *Proceeding Fifteenth Annual Conference Genetic and Evolutionary Computation Conference—GECCO 13* (ACM Press, Amsterdam, The Netherlands, 2013), p. 1021.
[37] S.-M. Udrescu, A. Tan, J. Feng, O. Neto, T. Wu, and M. Tegmark, arXiv:2006.10782 Phys. Stat (2020).
[38] E. Derner, J. Kubalík, N. Ancona, and R. Babuška, Appl. Soft Comput. **94**, 106432 (2020).
[39] S. Desai and A. Strachan, Sci. Rep. **11**, 12761 (2021).
[40] S.-M. Udrescu and M. Tegmark, Phys. Rev. E **103**, 043307 (2021).
[41] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Sci. Adv. **3**, e1602614 (2017).
[42] T. E. Karakasidis, A. Fragkou, and A. Liakopoulos, Phys. Rev. E **76**, 021120 (2007).
[43] R. J. Speedy, F. X. Prielmeier, T. Vardag, E. W. Lang, and H.-D. Lüdemann, Mol. Phys. **66**, 577 (1989).
[44] S. Chan and A. H. Elsheikh, J. Comput. Phys. **354**, 493 (2018).
[45] E. Sofronova and A. Diveev, Appl. Sci. **11**, 5081 (2021).