Predicting self-diffusion coefficients of small molecular fluids using machine learning and the statistical associating fluid theory for Mie segments

Justinas Šlepavičius,[1] Alessandro Patti,[1,2,3] and Carlos Avendaño[1,a)]

[1)]*Department of Chemical Engineering, School of Engineering, The University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom*

[2)]*Department of Applied Physics, University of Granada, Fuente Nueva s/n, 18071 Granada, Spain*

[3)]*Carlos I Institute of Theoretical and Computational Physics, Fuente Nueva s/n, 18071 Granada, Spain*

(Dated: 16 December 2024)

In our previous work [J. Chem. Phys. **159**, 024127 (2023)], we applied three Machine Learning (ML) models to predict the self-diffusion coefficient of spherical particles interacting via the Mie potential. Here, we introduce an optimization approach using the so-called Statistical Associating Fluid Theory for Mie segments (SAFT-VR Mie) and available vapor-liquid equilibria data to obtain molecular parameters for both Mie and Lennard-Jones potentials to describe the diffusion coefficient of 16 molecules described as a single sphere. Our ML models utilize these molecular parameters to predict the self-diffusion of these molecules. We conduct a comparative analysis between the molecular parameters derived from our thermodynamic approach and those obtained through direct fitting of the experimental self-diffusion coefficients. Our findings indicate that the predictive accuracy remains largely unaffected by the specific repulsive and attractive exponents of the Mie potential employed, provided that the fitting of the molecular parameters is precise. The Mie parameters obtained within a thermodynamic framework exhibit a higher coefficient of determination (R^2) and absolute average relative deviation (AARD) values compared to those derived from molecular parameters derived from fitting the self-diffusion coefficient, indicating their superior precision at higher values of the self-diffusion coefficient. Despite this discrepancy, the overall precision of both methodologies remains comparable. Given the abundance of precise thermodynamic data in contrast to self-diffusion data, we advocate the thermodynamic fitting approach as the preferred method for acquiring accurate Mie coefficients, essential to predict self-diffusion coefficients with ML and semi-empirical models.

[a)]Electronic mail: carlos.avendano@manchester.ac.uk

I. INTRODUCTION

Understanding and predicting the thermophysical properties of fluids is fundamental for the design of multiple industrial and scientific applications, playing a key role in the optimization of processes, the design of advanced functional materials, as well as in energy applications[1,2]. However, while there is a large body of work to describe the thermodynamic properties of real fluids, accurate models to describe transport properties are not as common[3]. The self-diffusion coefficient D is a key transport property that measures the rate at which molecules or particles diffuse through identical counterparts. This coefficient is determined as the limit of the rate of change in mean-squared displacement (MSD) over time as

$$D = \frac{1}{2d} \lim_{t \to \infty} \frac{d\left\langle r^2 \right\rangle}{dt}, \tag{1}$$

where d is the dimensionality of the system, t is the time, and $\left\langle r^2 \right\rangle$ is the MSD, respectively. The self-diffusion coefficient D_0 of hard spheres of diameter σ and mass m in the infinite dilution regime can be obtained analytically from Chapman-Enskog kinetic theory[4] and is given by

$$D_0 = \frac{3}{8} \left(\frac{k_B T}{m\pi} \right)^{1/2} \frac{1}{\rho \sigma^2}, \tag{2}$$

where T is the temperature, ρ is the number density and k_B is the Boltzmann constant, respectively[4,5]. This expression is only valid at infinite dilution (low densities). At higher densities, the theoretical description of the self-diffusion coefficient of hard spheres is challenging due to many-body interactions[6]. Molecular dynamics (MD) simulations are frequently used to describe the transport properties of model potential and real fluids in different regimes[7,8], thereby providing essential data for model development[9–12]. For the case of the self-diffusion coefficient, many of these models are often empirical correlations that represent the self-diffusion coefficient of model potentials and real fluids as a departure between the self-diffusion at infinite dilution D_0 and the real value of the self-diffusion coefficient D[7,8,13–19]. However, these correlations require several fitting parameters to work over a wide range of conditions, and the parameters are system-dependent.

Although the hard-sphere and Lennard-Jones potentials provide excellent reference potentials for understanding the physics of fluids, they are often not accurate in representing more complicated intermolecular interactions observed in real fluids, mainly due to the lack of control in the description of both the repulsive and attractive contributions of the potential. An alternative and

more accurate representation of the interaction between molecules is the so-called Mie potential, which is given by

$$u^{\mathrm{Mie}}(r) = \mathscr{C}\varepsilon\left[\left(\frac{\sigma}{r}\right)^{\lambda_\mathrm{r}} - \left(\frac{\sigma}{r}\right)^{\lambda_\mathrm{a}}\right], \tag{3}$$

where

$$\mathscr{C} = \frac{\lambda_\mathrm{r}}{\lambda_\mathrm{r}-\lambda_\mathrm{a}}\left(\frac{\lambda_\mathrm{r}}{\lambda_\mathrm{a}}\right)^{\lambda_\mathrm{a}/(\lambda_\mathrm{r}-\lambda_\mathrm{a})}, \tag{4}$$

ε corresponds to the depth of the well of the potential, σ is the collision diameter, and λ_r and λ_a are the exponents controlling the repulsive and attractive contributions of the Mie potential, respectively. The Mie potential allows an accurate description of the properties of real substances as repulsive and attractive contributions can be modulated for different molecules, offering a much better representation of the thermophysical properties of many complex systems[20–24]. The Chapman-Enskog kinetic theory for the diffusion coefficient at infinite dilution for particles that interact via the Mie potential has also been published[25].

Recently, there have been several attempts to use machine learning (ML) to correlate and predict self-diffusion coefficients by harnessing vast datasets available in the literature. These methodologies have been applied to predict diffusion in various contexts, including hydrocarbons[26], a range of molecular and atomic fluids[27], the diffusion of organic compounds in supercritical CO_2[28] and a large selection of Mie fluids[3,29]. However, a notable drawback of these approaches is their lack of transparency in elucidating the relationships between inputs and outputs. Although ML methods offer predictive capabilities, they lack interpretability, making it challenging to discern the rationale behind their predictions. A novel ML technique that is gaining traction is Symbolic Regression (SR)[30,31], which diverges from conventional approaches by providing an empirical equation as an outcome of the training process. The SR method has gained considerable attention for its ability to generate interpretable models, however, some of these regressions lack physical interpretation, however, new advances in SR algorithms have been able to describe the correct physics of the data and have been proven useful in discovering new physics[32,33]. Recently, SR models have been reported to describe the self-diffusion coefficients of the LJ[31,34] and Mie[29] potentials.

In this study, we represent 16 molecular fluids as single spheres interacting via the Mie potential. The so-called SAFT-VR Mie equation of state (EoS), which can accurately describe the

thermodynamic properties of various Mie fluids, is used to determine the intermolecular parameters for each fluid[22]. The selection of the fluids used in this study is based on the availability of experimental transport data and the complexity of their properties. To model the fluids, we employ a fitting procedure to characterize their vapor-liquid equilibrium (VLE) and determine their corresponding Mie molecular coefficients. Subsequently, using these molecular coefficients, we predict their self-diffusion coefficients using our previously developed ML models[29]. Additionally, we explore the feasibility of calculating the Mie coefficients directly from fitting the self-diffusion coefficient. These models enable us to assess the accuracy of thermodynamic coefficients compared to those of self-diffusion coefficients and gain insight into the molecular properties influencing more precise predictions.

The structure of this paper is as follows. In Section II, we provide a detailed discussion of the simulation methods used to obtain the self-diffusion coefficient and fit the VLE. In Section III, we present the performance of the ML methods for the 16 molecular systems using both sets of molecular parameters to describe the self-diffusion coefficient. Finally, Section IV we provide a summary of the key findings and conclusions drawn from this study.

II. METHODS

In this work, we explore the idea of simultaneously predicting both thermodynamic and dynamic properties across a diverse set of 16 fluids, which are represented as a single sphere. These molecules are classified into five distinct categories: light hydrocarbons and methanol (CH_4, C_2H_6, C_2H_4, CH_3OH), noble gases (Ar, Kr, Xe), halogenated methanes (CHF_3, CH_3F, CF_4, CF_3Cl), hydrogen isotopes (H_2, D_2) and molecular fluids (CO_2, SF_6, NH_3). These fluid selections represent a broad spectrum of substances exhibiting quasispherical geometry, diverse complex interactions, and anisotropic shapes. In addition, they boast a wealth of readily available transport data[27], facilitating a comprehensive exploration of the properties crucial for accurate property prediction. In this work, the intermolecular interactions of these fluids are described using the Mie potential (Equation 3).

The Mie intermolecular parameters used to represent each molecular system are obtained by fitting the so-called SAFT-VR Mie (EoS)[22] to available VLE data. The parameterization is done through the minimization of two objective functions. The first objective function F_1 corresponds to the standard residuals of the vapor pressure and liquid densities[21,22,35,36] and is given by

$$F_1(\sigma,\varepsilon,\lambda_r,\lambda_a) = \sum_{i=1}^{N_p}\left(\frac{P^{sat}(T_i;\sigma,\varepsilon,\lambda_r,\lambda_a) - P^{sat,exp}(T_i)}{P^{sat,exp}(T_i)}\right)^2 + \\ + \sum_{i=1}^{N_p}\left(\frac{\rho^{L}(T_i;\sigma,\varepsilon,\lambda_r,\lambda_a) - \rho^{L,exp}(T_i)}{\rho^{L,exp}(T_i)}\right)^2, \tag{5}$$

where N_p is the number of experimental points, T_i is the absolute temperature at point i, $P^{sat,exp}$ and P^{sat} correspond to the experimental and theoretical (SAFT-VR Mie EoS) values of the vapour, respectively, while $\rho^{L,exp}$ and ρ^{L} correspond to the experimental and theoretical values of the liquid density, respectively. It has been shown that an improvement in the representation of second-order derivative properties and the critical region is achieved using additional properties during parameter estimation. Therefore, the vapor pressure has also been considered in the second objective function F_2, and is given by

$$F_2(\sigma,\varepsilon,\lambda_r,\lambda_a) = F_1 + \sum_{i=1}^{N_p}\left(\frac{\rho^{V}(T_i;\sigma,\varepsilon,\lambda_r,\lambda_a) - \rho^{V,exp}(T_i)}{\rho^{V,exp}(T_i)}\right)^2, \tag{6}$$

where $\rho^{V,exp}$ and ρ^{V} correspond to the experimental and theoretical values of the vapor density. This additional objective function allows one to investigate the importance of the vapor density in the VLE fitting and self-diffusion coefficient prediction. The molecular parameters of the Mie potential are obtained by minimizing the two objective functions using the Nelder-Mead algorithm[29]. All experimental VLE properties used for optimization of the molecular parameters are taken from the NIST Chemistry WebBook[37].

In addition to employing this pair of objective functions to predict variables for the Mie potential, we utilize identical objective functions to compute the VLE fitting for the LJ potential by setting $\lambda_r = 12$ and $\lambda_a = 6$. Consequently, these fittings enable the derivation of four molecular models for each fluid under investigation, utilizing both the Mie and the LJ potentials. Using the ML algorithms trained on the complete dataset from our prior study, we obtain the self-diffusion coefficient for each model. These predictions are made solely based on the temperature T, density ρ, and the so-called cohesive parameter α, being the cohesive parameter of the Mie potential. In particular, α is defined as[38]

$$\alpha \equiv \mathscr{C}\left(\frac{1}{\lambda_a - 3} - \frac{1}{\lambda_r - 3}\right). \tag{7}$$

Once the molecular parameters of the Mie potential for every substance are determined using the two objectives functions given by Equations 5 and 6 using the SAFT-VR Mie EoS[21,22], the

ML models developed in our previous work[29] are used to predict the self-diffusion coefficient. The details of these ML models are presented in the Supplementary Information (SI). These ML models have been constructed using the Artificial Neural Network (ANN), k-Nearest Neighbors (KNN) and Symbolic Regression (SR) algorithms. On the one hand, ANN and KNN have been trained with accurate molecular dynamics simulation results for the diffusion coefficient D for different state conditions and different attractive and repulsive exponents, and only require as features (inputs) the reduced temperature $T^* = k_B T/\varepsilon$, the reduced density $\rho^* = \rho\sigma^3$, and the cohesive parameter α, while the output is obtained as the ratio D/D_0. On the other hand, the SR method generates a correlation that facilitates the computation of the diffusion coefficient. Specifically, the correlation derived from SR is represented as[29]:

$$\frac{D}{D_0} = \exp\left(-\rho^*/T^*\right)\exp\left(-|0.273\log(\log(1/\rho^*))|\right). \tag{8}$$

It is important to stress that while this correlation lacks any physical meaning, it is convenient from the interpretability point of view and is applicable over a wide range of conditions in vapour, liquids and supercritical states. To numerically evaluate the performance of the models, we use the coefficient of determination R^2 as well as the absolute average relative deviation (AARD), defined as

$$\text{AARD} = \frac{1}{n}\sum_{i=1}^{n}\frac{|(y_i - \hat{y}_i)|}{y_i} \times 100\%, \tag{9}$$

where n denotes the total number of samples, and y_i and $\hat{y}_i$ represent the experimental and predicted values for the i-th sample of property y, respectively. The reference experimental values for the self-diffusion coefficient are taken from the work of Allers *et al.*[27]. These evaluation metrics are used due to their sensitivity to different types of errors: while R^2 is particularly responsive to variations in larger values, the AARD is more sensitive to errors occurring in smaller values.

Finally, we have also optimized the intermolecular parameters of the Mie potential by using direct experimental values of the diffusion coefficient obtained from the literature to refine the model for characterizing the self-diffusion coefficient. To this end, we minimize an objective function representing the residual of the diffusion coefficient, defined as:

$$F_D(\sigma,\varepsilon,\lambda_r,\lambda_a) = \left[\left(\sum_{i=1}^{N}\frac{D(T_i,\rho;\sigma,\varepsilon,\lambda_r,\lambda_a) - D^{\exp}(T_i)}{D^{\exp}(T_i)}\right)^2\right], \tag{10}$$

where D^{exp} and D correspond to the experimental and theoretical values (molecular dynamics) of the self-diffusion coefficient. To explore the effect of the cohesive parameter on the accuracy of the prediction and to investigate the consistency between the thermodynamic and diffusion approaches, we optimize the above function for all values of the cohesive parameter α in the range $0.25 \leq \alpha \leq 1$ in steps of $\Delta\alpha = 0.05$.

III. RESULTS

A. Fitting of molecular parameters

The molecular parameters ε and σ of both Mie and LJ potentials have been obtained by fitting experimental vapor-liquid equilibria using the SAFT-VR Mie EoS. This fitting procedure employs the two objective functions described in the previous section (cf. Eqs. 5 and 6), and the results are presented in Figures 1 and 2. The Mie parameters derived from the objective function F_1^{Mie} are detailed in Table I, while the results for other parameters can be found in the SI.

Before assessing the accuracy of the models in representing the VLE data, we discuss the Mie coefficients obtained. As discussed in the preceding section, the Mie potential represents a more versatile form of the LJ potential, where the LJ potential can be viewed as a special case of the Mie potential with $\lambda_r = 12$ and $\lambda_a = 6$. The selection of $\lambda_r = 12$ was somewhat arbitrary, although it aligns well with the Pauli exclusion principle. In contrast, $\lambda_a = 6$ is theoretically derived from London forces[39]. In addition, setting $\lambda_r = 2\lambda_a$ helps reduce the computational intricacies associated with the potential. It should be noted that the value of λ_a can increase to greater numbers if there are quadrupole interactions within the molecule[40]. For most of the fluids studied, the Mie parameters obtained include $\lambda_a \approx 6$ for molecules without quadrupoles, while $\lambda_a \approx 10$ is observed for molecules possessing quadrupoles, such as CO_2 and SF_6. This observation validates our derived potential parameters in comparison to theoretical expectations.

First, we discuss the thermodynamic fittings of light hydrocarbons and methanol, as shown in the upper rows of Figures 1 and 2. The VLE of methane shows excellent agreement with both LJ and Mie potentials using the two objective functions F_1 and F_2. This agreement extends across the entire tested range of temperatures and densities, reflecting a robust representation of VLE and saturation pressure. Moreover, the Mie coefficients obtained align closely with the results reported by Lafitte *et al.*[22]. For ethane and ethene, the Mie potentials offer robust representa-

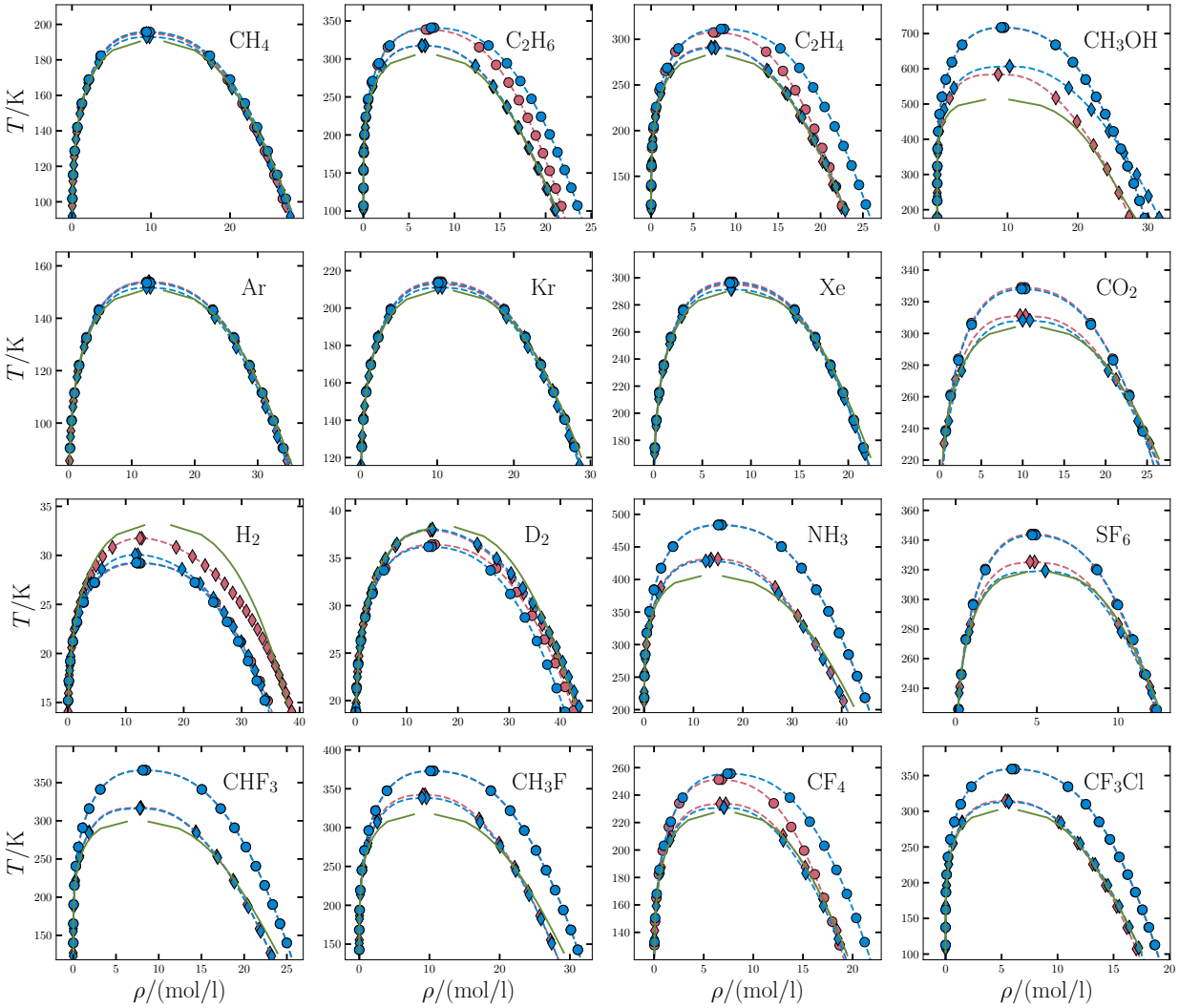

FIG. 1. Results for the saturation densities as function of the temperature using the SAFT-VR Mie EoS. Circles denote the results obtained with the LJ potential, while diamonds represent those obtained with the Mie potential. Magenta symbols correspond to the potentials derived using the objective function F_1, whereas blue symbols denote those obtained with F_2. Green solid curves represent the VLE data obtained from the NIST Chemistry WebBook[37], while the dashed curves denote the complete VLE obtained from the SAFT-VR Mie EoS.

tion, while the LJ potentials tend to overpredict pressures and critical points. This discrepancy is expected considering that the LJ potential was initially tailored as approximations for noble gases, whereas the more intricate structures and interactions of ethane and ethene require a more nuanced approach. The excellent performance of the Mie potential to capture thermodynamic properties is noteworthy, despite modeling molecules as single spheres. In contrast, a single-site

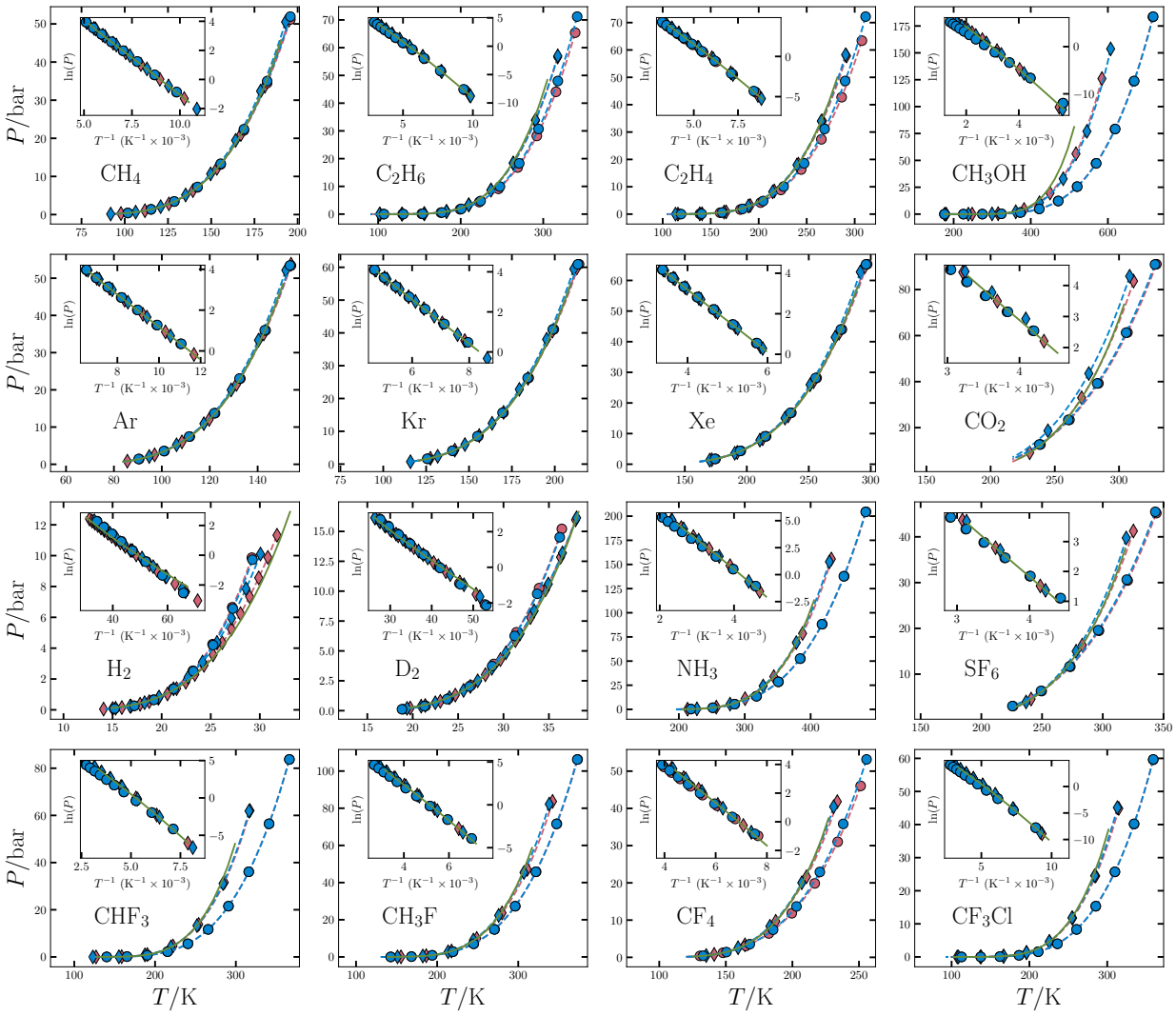

FIG. 2. Results for the vapour-pressure using the SAFT-VR Mie EoS. Circles denote the results obtained with the LJ potential, while diamonds represent those obtained with the Mie potential. Magenta symbols correspond to the potentials derived using F_1, whereas blue symbols denote those obtained with F_2. Green lines represent the vapour pressure data obtained from the NIST Chemistry WebBook[37]. The inset contains a plot of $\ln P$ against $1/T$ to highlight the results are low temperatures.

model fails in representing the VLE of methanol over the entire temperature range regardless of the method tested, with only the Mie potential model using the objective function F_1 showing a good representation at low temperatures. This outcome is unsurprising, given the multitude of interactions inherent in methanol that are not adequately represented in our model, notably the intricate directional hydrogen bonding between oxygen and hydrogen atoms and the polar nature of the molecule. Additionally, methanol's highly nonspherical molecular shape poses challenges

TABLE I. Summary of the Mie intermolecular parameters obtained by the F_1^{Mie} and F_2^{Mie} objective functions. The results are obtained by the use of the Nelder-Mead algorithm fitting the SAFT Mie EoS to VLE data obtained from NIST.

	F_1^{Mie}					F_2^{Mie}				
Fluid	σ/Å	(ε/k_B)/K	λ_r	λ_a	α	σ/Å	(ε/k_B)/K	λ_r	λ_a	α
CH_4	3.748	154.14	12.66	6.03	0.85	3.754	160.53	14.15	5.98	0.80
C_2H_6	4.291	298.83	18.33	6.00	0.69	4.281	299.88	18.15	6.04	0.68
C_2H_4	4.147	278.78	18.94	6.02	0.67	4.136	281.30	19.70	6.00	0.66
CH_3OH	4.082	747.77	27.64	7.02	0.44	3.879	683.33	33.19	5.93	0.55
Ar	3.405	126.54	10.13	7.06	0.80	3.429	126.67	14.23	5.99	0.80
Kr	3.655	158.01	14.98	5.40	0.93	3.663	176.34	14.28	5.98	0.80
Xe	3.981	228.14	14.12	5.67	0.88	3.996	224.93	17.48	5.28	0.89
CHF_3	4.224	352.82	31.47	5.93	0.56	4.222	357.90	29.99	6.09	0.54
CH_3F	3.888	346.16	22.55	5.97	0.63	3.902	346.56	25.54	5.84	0.62
CF_4	4.344	256.87	28.85	6.00	0.56	4.380	265.77	29.74	6.20	0.53
CF_3Cl	4.668	329.43	24.82	5.98	0.60	4.652	324.17	28.59	5.73	0.61
H_2	3.409	10.33	5.69	5.67	2.15	3.440	16.40	7.92	5.93	1.30
D_2	3.196	18.40	6.68	6.25	1.47	3.190	17.29	6.46	6.15	1.57
CO_2	3.818	449.44	18.24	10.03	0.35	3.741	353.55	23.00	6.66	0.52
NH_3	3.443	485.01	30.41	6.04	0.54	3.442	558.57	17.92	8.64	0.42
SF_6	4.902	468.11	18.06	10.04	0.36	4.956	458.31	31.62	7.71	0.37

for accurate modeling compared to other hydrocarbons examined.

The noble gases, namely Ar, Kr, Xe, depicted in the three leftmost subplots of the second row in Figures 1 and 2, demonstrate excellent agreement in all approaches. Although the critical point is overpredicted by approximately 5% in all fitting models, the overall fitting performance is very good. This outcome was anticipated because noble gases exhibit a spherical symmetry and lack additional intermolecular forces apart from London dispersion forces. These forces are effectively captured by the LJ potential and its extension, the Mie potential. The calculated LJ coefficients ε and σ, as shown in the SI, are closely aligned with those reported by Dufal *et al.*[41].

The VLEs of halogenated methanes are not well represented by the Mie potential, and we ob-

serve even worse performance using the LJ potential, as shown in the bottom rows of Figures 1 and 2. Among halogenated methanes, CF_4 and CF_3Cl are the molecules best represented by the Mie potential. This observation may be attributed to the absence of hydrogen bonding between halogen atoms compared to that of hydrofluorocarbons, rendering these fluids to be more spherosymmetric. An intriguing observation pertains to the variations in CF_4 using the LJ models, where F_1^{LJ} predicts ρ_{L} with notably higher accuracy; the inclusion of the term ρ_{v} in the objective function appears to have adversely impacted the accuracy of the VLE in the liquid branch. In contrast, the hydrofluorocarbons are described accurately using the Mie potentials. However, owing to the increased intermolecular forces, the critical point tends to be overpredicted to a greater extent compared to molecules devoid of such forces.

The modeling of hydrogen isotopes varies due to their quantum mechanical effects, as depicted in the two left sub-figures in the third rows of Figures 1 and 2. The LJ models demonstrate poor performance, notably underpredicting the VLE and critical points. This discrepancy suggests that the potentials for hydrogen molecules are softer than the LJ potential, with $\alpha_{\mathrm{H_2}} > 1.3$ and $\alpha_{\mathrm{D_2}} > 1.4$, compared to $\alpha_{\mathrm{LJ}} = 0.89$. The coefficients obtained differ significantly from those reported by Aasen *et al.*, which could be attributed to fixed values λ_r and λ_a[42], as well as the limitations of our minimization algorithm and the large influence of quantum mechanical effects on H_2. For H_2, the best model is obtained using the parameters derived from the objective function F_1^{Mie}, where the increased importance of the liquid branch and the saturation pressure result in a closer approximation. Deuterium, being heavier and less influenced by quantum-mechanical effects, is equally well approximated by both sets of parameters using the Mie potential. The molecular parameters for D_2 show better agreement with their counterparts in Aasen's work with $\mathrm{AARD}_{\varepsilon}\ (\mathrm{H_2}, \mathrm{D_2}) = 73.6\%, 15.2\%$ and $\mathrm{AARD}_{\sigma}\ (\mathrm{H_2}, \mathrm{D_2}) = 4.4\%, 1.3\%$.

The molecular fluids, despite their highly diverse compositions, exhibit remarkably similar modeling results. CO_2 and SF_6, depicted in the right-hand subplots of rows two and three of Figures 1 and 2, are very well modeled using both Mie potentials. In particular, the use of F_2^{Mie} appears to yield a slightly more accurate critical point for these fluids, although the differences are marginal. On the other hand, NH_3 is less satisfactorily modeled, with an over-predicted critical point. This discrepancy could potentially be attributed to hydrogen bonding between ammonia molecules, as illustrated in the third subfigure in the third row of Figures 1 and 2. The LJ potential does not represent these fluids well, displaying a substantial overprediction of the critical point for CO_2 and SF_6. Moreover, the LJ model for NH_3 does not accurately capture both the critical point

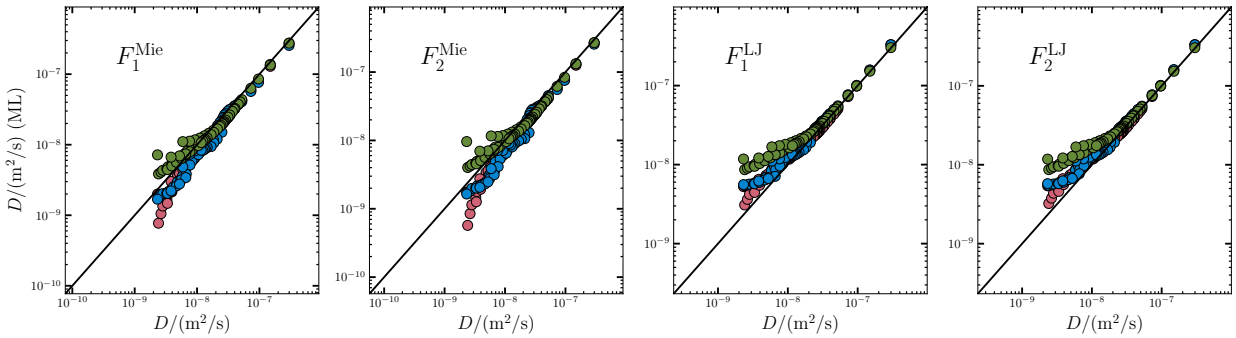

FIG. 3. Parity plots for the self-diffusion coefficients of SF_6 using the 4 model potentials. The ANN model is presented by magenta circles, KNN by blue circles, and SR by green circles.

and the liquid branch of the VLE.

B. Prediction of self-diffusion coefficients

We now compare the prediction of the self-diffusion coefficient in real fluids through two primary avenues. First, we assess the self-diffusion coefficient of each fluid using all four different thermodynamic models. Specifically, we evaluate the performance of the ML models in predicting the self-diffusion coefficient of SF_6 using the four molecular parameters obtained using the Mie and LJ potential models. We chose SF_6 for this comparison due to the significant differences between the LJ and Mie potentials, as well as the notable observable effect on the parity plots. The parity plots for SF_6 are presented in Figure 3, while the parity plots for the 16 molecules obtained from the objective function F_1^{Mie} are shown in Figure 4. More detailed figures for the remaining fluids are provided in the SI.

In the parity plots shown in Figure 3, one can observe that all four sets of parameters obtained via the SAFT-VR Mie EoS offer a good representation of the self-diffusion coefficient of SF_6 using the three ML methods developed in our previous work, particularly in the region of high values of the self-diffusion coefficient. The performance of these models can also be observed in the AARD and R^2 values presented in Table II, where all models show $R^2 \geq 0.994$. This outcome is not entirely surprising, given that the model obtained via SR indicates that the self-diffusion coefficient does not explicitly depend on α, suggesting that the value of α may not be essential to achieve a good model for the self-diffusion coefficient when fluid interactions are described using the Mie potentials. This hypothesis is supported by the substantial increase in AARD for

TABLE II. Summary of the AARD and R^2 descriptors of different ML methods applied in this work to predict the self-diffusion coefficient of SF_6 using the four objective functions. The reference self-diffusion coefficient data used is obtained from Allers *et al.*[27]

	ANN		KNN		SR	
Objective Function	AARD	R^2	AARD	R^2	AARD	R^2
F_1^{Mie}	25.54%	0.9979	25.37%	0.9951	23.95%	0.9948
F_2^{Mie}	28.96%	0.9980	28.65%	0.9950	26.86%	0.9940
F_1^{LJ}	20.44%	0.9958	28.24%	0.9981	75.27%	0.9967
F_2^{LJ}	21.66%	0.9958	29.32%	0.9982	76.35%	0.9967

F^{LJ} functions obtained by SR. Furthermore, the significance of α becomes more apparent in the region of high liquid densities, which coincides with deviations observed along the middle of the parity line. These deviations are attributed to the strong correlation due to the intermolecular forces in the liquid state where low values of the self-diffusion coefficient are observed. With respect to these high-density states, the accuracy of the predictions varies depending on the potentials used. Specifically, ANN demonstrates better prediction performance for LJ potentials, while KNN and SR prove to be superior predictors when the Mie potential is used.

The results for the self-diffusion coefficient for all molecules using the molecular parameters obtained using the F_1^{Mie} objective function are shown in Figure 4 and the corresponding values of AARD are presented in Table III. In Figure 4, we can observe that for methane, ethane, and ethene, all models exhibit well-predicted self-diffusion coefficients in the region of high values of D. However, for the region of low values of D, the parity plots deviate from the diagonal, a phenomenon also observed in our previous work[29]. Specifically, ANN tends to underestimate the self-diffusion coefficient, while SR and KNN tend to overestimate it. For methane, this high-density region is relatively small due to the compact shape of the VLE, resulting in more points closer to the parity line. Interestingly, in the parity plot for C_2H_4, ANN predicts the value of D very accurately and does not show such a pronounced decaying tail. In contrast, the predictions of the self-diffusion coefficients of CH_3OH form a cloud around the parity line. It is evident that at intermediate values of D, the prediction lines for ANN and KNN fall below the parity line before diverging from it. This highlights the fact that a single Mie potential does not adequately capture the complex interactions of methanol.

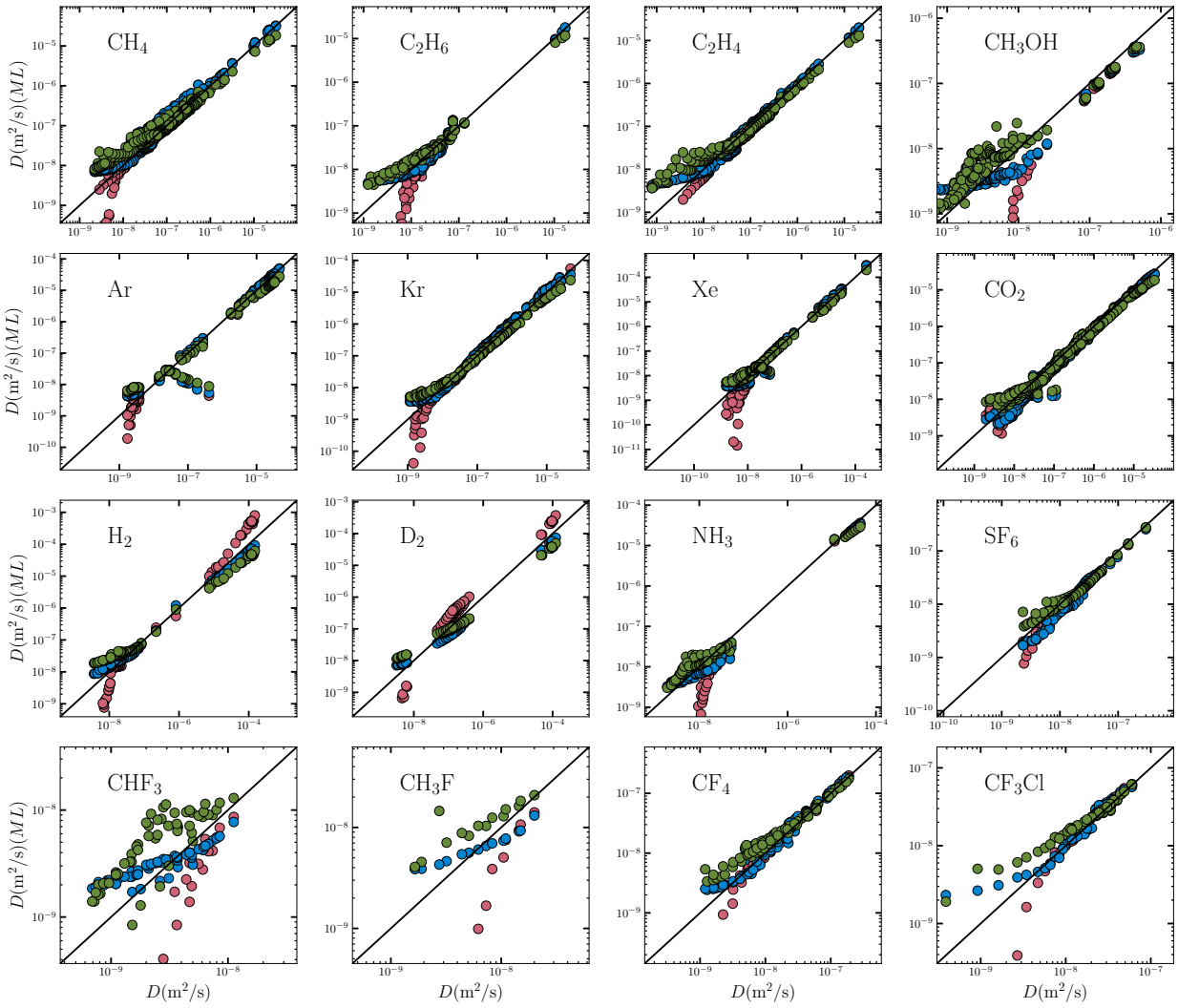

FIG. 4. Parity plots for the self-diffusion coefficients of the fluids studied in this work, modelled with the Mie potential with molecular parameters obtained by the F_1^{Mie} objective function. The ANN model is presented by magenta circles, KNN by blue circles, and SR by green circles. Self-diffusion coefficient data were obtained from Allers *et al.*[27].

All noble gases are predicted with remarkable accuracy. Two notable features in the parity plots warrant discussion: the underpredictions of ANN at low D and the straight perpendicular line visible in the parity plot for argon. The under-predictions predicted by ANN are intriguing, particularly because this phenomenon is not observed in the equivalent branch of KNN and SR. Under-prediction may be more noticeable on a logarithmic scale, but further investigation is required to validate this claim. The perpendicular line observed in the argon parity plot may be due to issues with the experimental data. The temperatures, densities, and phases used to derive

TABLE III. Summary of the AARD and R^2 descriptors for all fluids predicted by the different ML methods applied in this work using the model obtained from the F_1^{Mie} objective function for the self-diffusion coefficients collected by Allers *et al.*[27].

Fluid	ANN		KNN		SR	
	AARD	R^2	AARD	R^2	AARD	R^2
CH_4	25.16%	0.9978	26.98%	0.9977	47.94%	0.9960
C_2H_6	73.15%	0.9992	38.51%	0.9955	54.00%	0.9981
C_2H_4	38.76%	0.9997	28.99%	0.9989	43.24%	0.9963
CH_3OH	532.22%	0.9960	120.08%	0.9951	93.79%	0.9939
Ar	20.62%	0.9984	36.41%	0.9903	73.92%	0.9861
Kr	23.52%	0.9897	25.94%	0.9633	48.53%	0.9773
Xe	23.75%	0.9997	20.31%	0.9997	37.68%	0.9999
CO_2	25.28%	0.9974	25.88%	0.9957	29.81%	0.9856
H_2	117.26%	0.9753	40.08%	0.9928	105.93%	0.9924
D_2	125.67%	0.9873	48.80%	0.9980	57.05%	0.9979
NH_3	136.57%	0.9980	33.63%	0.9933	47.71%	0.9865
SF_6	25.54%	0.9979	25.37%	0.9951	23.95%	0.9948
CHF_3	204.51%	0.8886	55.23%	0.9148	116.98%	0.6404
CH_3F	125.47%	0.9851	40.74%	0.9756	89.90%	0.7925
CF_4	29.37%	0.9948	21.69%	0.9923	49.15%	0.9959
CF_3Cl	32.77%	0.9824	23.89%	0.9664	42.81%	0.9883

these D values probably do not correspond to the correct phase under experimental conditions, as verified by the NIST database.

CF_4 is predicted exceptionally well by all three ML methods. The deviations at low D are minimal, and the parity is maintained consistently along the parity line. Similarly, CF_3Cl is well predicted, although the deviations at low D are slightly more pronounced. In contrast, CH_3F and CHF_3 are poorly predicted by the three ML methods. The parity line is surrounded by a cloud of points, indicating a very low accuracy. However, it should be noted that there is no experimental data for the self-diffusion coefficient at high values available for these fluids. Therefore, direct comparisons of R^2 and AARD to other fluids may not be entirely appropriate. Compared with the

exact range of self-diffusion coefficients for other fluids, the observed deviations are consistent with the expected deviations at those values of D.

In contrast to the fit of the thermodynamic properties, the self-diffusion coefficients for H_2 and D_2 are well predicted. In particular, there is a discernible overprediction by the ANN model at high values of D, a deviation not observed in previous parity plots or other models. Intriguingly, no notable deviations are observed in the other plots at high values of D. This discrepancy may be due to the high values of D for hydrogen and deuterium that fall outside the training region for ML methods. This suggests that the SR and KNN models may be better suited for extrapolation. However, despite these deviations, it is crucial to emphasize that these accurate transport predictions were obtained using a thermodynamic model that failed to accurately model the thermodynamic properties. Interestingly, this discrepancy did not translate into predictions of the self-diffusion coefficient, as the values of D were accurately modeled.

Finally, the self-diffusion coefficients of molecular fluids are predicted quite accurately. The most significant deviations are observed for low values of D for NH_3, where ANN exhibits considerable deviation. However, the rest of the parity line for NH_3, as well as the parity lines for CO_2 and SF_6, are well predicted by the three ML methods. Although there is a slight increase in deviations towards the region of low values of D, these discrepancies are overall minor.

C. Fitting the self-diffusion coefficient

Having obtained the Mie molecular parameters from the thermodynamics of the molecular fluids, we investigated whether there exist better coefficients that could produce more accurate self-diffusion coefficients. As detailed in Section II, we calculated the Mie molecular parameters that would best fit the self-diffusion coefficients for each fluid within the range of $0.25 \leq \alpha \leq 1$. The resulting values of ε and σ, along with their associated errors using the KNN algorithm (AARD and R^2), are presented for SF_6 in Figure 5. The plots and values for all the fluids are provided in the SI.

The results of this investigation reveal surprising disparities between the values of σ and ε obtained from the thermodynamic properties and those obtained from fitting to the self-diffusion coefficient. Although the values obtained from thermodynamics generally fall within the same range as those obtained by fitting the self-diffusion coefficient, there are notable differences. In many cases, the difference between the values exceeds half of the range presented in σ, with a

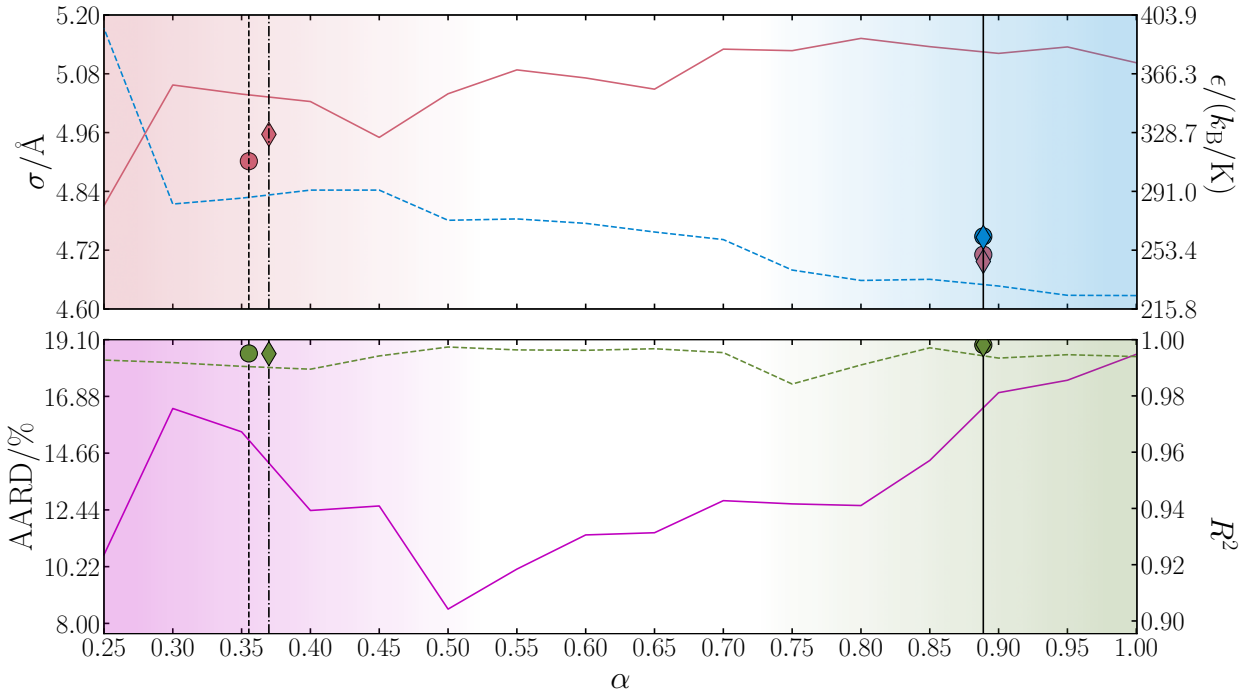

FIG. 5. The results of fitting the self-diffusion coefficients at different values of α using the ANN algorithm to obtain the optimal Mie intermolecular parameters σ and ε for SF_6. The obtained intermolecular parameters are depicted in the top plot, where σ is represented in red and ε in blue. Additionally, the average absolute relative deviation (AARD) is depicted in purple, and the coefficient of determination (R^2) is shown in green. In the bottom plot, the errors corresponding to σ and ε are illustrated. The solid black line represents the α values of the LJ potential, while the dashed and dash-dot vertical lines represent the potentials used in the previous section, namely F_1^{Mie} and $F^{\mathrm{Mie}}2$, respectively. Furthermore, the intermolecular parameter and error values obtained from the previous section are also presented for $F1$ (circles) and F_2 (diamonds).

similar behavior observed for ε, albeit to a lesser extent. Interestingly, within the range of α investigated, most fluids exhibit very low error variation. For most potentials, the AARD range is much smaller than the lowest AARD value and R^2 exceeds 0.9. It is noteworthy that the values obtained from thermodynamic fitting appear to have a larger AARD and R^2 compared to those obtained in this investigation. This suggests that the thermodynamic method exhibits better accuracy at high values of D but is less accurate for low values of D.

Let us first focus on the results for the light hydrocarbons and methanol in detail, which are presented in the SI. Methane exhibits an excellent representation throughout the α range, with a

consistently high value of R^2. The best fitting is observed around $\alpha \approx 0.3$, coinciding with the minimum AARD. In particular, this value of α is significantly harder than the thermodynamically obtained value, where $\alpha > 0.8$. For C_2H_6, an effective representation is achieved with potentials that are harder than the one obtained at $\alpha \approx 0.67$. At higher values of α, the best coefficients in terms of R^2 are obtained from the thermodynamic fit, suggesting better approximations for high D. The self-diffusion coefficient of C_2H_4 is consistently predicted well across the studied range, with AARD $\approx 30\%$ and $R^2 > 0.98$. There is a notable decrease at $\alpha = 0.4$, leading to a drastic reduction in AARD. However, methanol is not well predicted in the range by either metric, with $R^2 < 0.6$ and AARD exceeding 40%. Unlike other fluids studied, thermodynamic methods provide better predictions for the high D range. Noble gases exhibit high-accuracy predictions for most of the α range. For argon and krypton, optimal predictions are found around $\alpha \approx 0.4$, with the AARD increasing for higher values of α. Similarly, the predictions for xenon are similar within the range $0.5 < \alpha < 0.9$, with the best predictions achieved at $\alpha = 0.5$. Interestingly, all the optimal predictions for noble fluids are observed at values of α much lower than those predicted thermodynamically, typically falling within the range of $0.8 < \alpha < 0.9$.

The precision of the predictions for the halogenated methanes varies widely. The best predicted fluids, such as CF_4 and CF_3Cl, as previously observed, exhibit better accuracy when fitting the self-diffusion coefficient rather than relying on thermodynamic values. This is evidenced by an increase in R^2 and a significant decrease in AARD. For both of these fluids, the best results are found in the range $\alpha < 0.65$. In contrast, hydrofluorocarbons (CHF_3 and CH_3F) are predicted very poorly for many values of α, with the best predictions falling within the ranges of $0.55 < \alpha < 0.8$ for CHF_3 and $\alpha > 0.6$ for CH_3F. In particular, for CH_3F at $\alpha = 0.6$, there is an apparent transition from a good predictive region to a poor predictive region, coinciding with a transition in the calculated value of ε. This suggests that the algorithm may have converged to a poor local minimum at that value of ε, resulting in poor predictions. Another explanation could be the absence of high D state points for these fluids in the dataset, which might affect the predictions.

The first notable observation with respect to hydrogen and deuterium is the absence of observable best-fit lines from F_1 and F_2. This is attributed to the values obtained of α, where $\alpha_{H_2} \geq 1.3$ and $\alpha_{D_2} > 1.45$. Despite the high predicted values of α, it is surprising that for both fluids, the best predictions were found at $\alpha \approx 0.4$. H_2 exhibits good predictions within the studied range, with $R^2 > 0.9$ and AARD$< 25\%$. On the other hand, the predictions for D_2 show a higher R^2 across the entire range studied and a lower AARD by approximately 4%. SF_6 emerges as the molecular

fluid predicted best and ranks among the top two predicted fluids overall, alongside krypton. The optimal prediction occurs at $\alpha = 0.5$, a value higher than the thermodynamically derived α. This deviation is unique among the fluids studied, possibly due to the lower thermodynamically fitted values of α for molecular fluids compared to other fluids, except methanol. CO_2 demonstrates the best predictions when $\alpha = 0.7$, while NH_3 exhibits a stable prediction range of $0.4 < \alpha < 0.7$, with the best predictions found at $\alpha = 0.6$.

IV. CONCLUSION

In summary, we have derived the molecular parameter for the Mie potential for 16 fluids by fitting their thermodynamic properties, resulting in four different models. We conducted a comparative analysis to assess the efficacy of the Mie and LJ potentials in modeling both the thermodynamic and transport properties. Moreover, we examined two objective functions, one with and one without a vapor density term, to gauge their impact on the fitting process. Our investigation revealed that Mie potentials outperformed LJ potentials in fitting VLE and vapor pressure data. This superiority can be attributed to the greater flexibility of Mie potentials, which allow for more subtle adjustments to interparticle interactions. Across most fluids, the Mie potentials obtained exhibited remarkable similarity, with exceptions noted for CH_3OH and H_2, where the objective function of F_1 produced a superior fit. In particular, the inclusion of vapor densities in the objective function F_2 improved the fit for the vapor branch but led to some compromise in the fit for the liquid branch.

By using the acquired Mie parameters, we applied our preexisting ML models to predict the self-diffusion coefficients of the fluids. The predictions were notably accurate for self-diffusion coefficients greater than $10^{-7}\mathrm{m}^2\,\mathrm{s}^{-1}$, across various methods and coefficients. This reiterates our previous findings, suggesting that at higher values of the self-diffusion coefficient, the specific form of the potential has minimal impact on the accuracy of the prediction. However, at lower values of the self-diffusion coefficient, stronger intermolecular forces come into play, which may not be fully captured by the Mie potential or the spherical shape assumptions employed in our training data.

The similarity in the predictive performance across all potentials prompted us to explore the relationship between the cohesion parameter α and the predictive precision of the KNN algorithm for self-diffusion coefficients. In contrast to our previous approach, we fitted the Mie parameters

directly to the self-diffusion coefficient. Interestingly, we observed that for most of the fluids studied, the predictive precision of D remained consistent throughout the entire range of α investigated ($0.25 < \alpha < 1$). Surprisingly, the differences in predictive accuracy were not correlated with the α values obtained from the thermodynamic fitting. Strikingly, we found that the thermodynamically fitted potentials exhibited higher R^2 and AARD compared to the self-diffusion coefficient fitted potentials. This suggests that the fitting method used for the self-diffusion coefficient placed more emphasis on improving the accuracy for low values of D, while the thermodynamic fitting method provided better approximations for high values of D.

The disparities in errors between both fitting methods were noticeable, but not significant. Although each method exhibited different strengths, only the thermodynamically fitted coefficients yielded a satisfactory approximation of the VLE. This was evident from the differences observed between the obtained Mie and LJ coefficients for VLE fitting, where the accuracy of the cohesion parameter played a crucial role. In contrast, all four sets of parameters resulted in equally accurate predictions of transport properties. Given that the thermodynamic fitting method yielded accurate representations of both transport and thermodynamic properties, we argue that this method is superior for deriving Mie coefficients for small molecular fluids. Using the NIST database and the ML methods outlined here, we propose a reliable and straightforward approach for estimating self-diffusion coefficients in small molecular fluids devoid of hydrogen bonding.

The overall accuracy of the ML models appears to be independent of the selection of Mie coefficients for the fluids studied. This observation is similar to our previous finding[29], where we observed that Mie potentials with the same value of the cohesive parameter α exhibit not only the same VLE envelope but also the same values of the self-diffusion coefficient. This points to a similar three-parameter corresponding states model for the self-diffusion coefficient found for VLE[38]. Looking ahead, further improvements in accuracy could be achieved by applying a more refined model to the obtained Mie coefficients, leading to enhanced predictions of the self-diffusion coefficients. One potential avenue for improvement involves training the model on a larger dataset comprising higher-density state points. This would offer deeper insights into the transport behavior of dense Mie fluids. Such insights could prove invaluable for analyzing the transport properties of real fluids under similar conditions, thereby enabling further refinement of the models.

V. SUPPLEMENTARY INFORMATION

In the SI, we report the list of molecular parameters for the Mie and Lennard-Jones potentials using the SAFT-VR EoS and the two optimization strategies discussed in Section II, the results for the prediction of the self-diffusion coefficient using the molecular parameters from the thermodynamic fitting, and the list of molecular parameters obtained from fitting available data for the self-diffusion coefficient.

VI. ACKNOWLEDGEMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through an Industrial Cooperative Award in Science & Technology (ICASE) co-funded by IBM, project ID 2327699 - EP/T517689/1. The authors acknowledge the assistance given by Research IT and the use of the Computational Shared Facility at The University of Manchester. This work was also supported by the Hartree National Centre for Digital Innovation, a collaboration between STFC and IBM. AP acknowledges grant PID2022-136540NB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF *A way of making Europe*; grant P21_00015 funded by Junta de Andalucía - Consejería de Universidad, Investigación e Innovación; and a *María Zambrano Senior* fellowship, funded by the NextGenerationEU/PRTR program.

REFERENCES

[1]E. Hendriks, G. M. Kontogeorgis, R. Dohrn, J.-C. de Hemptinne, I. G. Economou, L. F. Žilnik, and V. Vesovic, Industrial & Engineering Chemistry Research **49**, 11131–11141 (2010).

[2]G. M. Kontogeorgis, R. Dohrn, I. G. Economou, J.-C. de Hemptinne, A. ten Kate, S. Kuitunen, M. Mooijer, L. F. Žilnik, and V. Vesovic, Industrial & Engineering Chemistry Research **60**, 4987–5013 (2021).

[3]G. Chaparro and E. A. Müller, The Journal of Physical Chemistry B **128**, 551–566 (2024).

[4]S. Chapman and T. G. Cowling, The Mathematical Gazette **38**, 63 (1952).

[5]D. Chandler, The Journal of Chemical Physics **62**, 1358–1363 (1975).

[6]C. Silva and H. Liu, "Modelling of transport properties of hard sphere fluids and related systems, and its applications," in *Theory and Simulation of Hard-Sphere Fluids and Related Systems*, edited by Á. Mulero (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008) pp. 383–492.

[7]R. J. Speedy, Molecular Physics **62**, 509–515 (1987).

[8]R. Speedy, F. Prielmeier, T. Vardag, E. Lang, and H.-D. Lüdemann, Molecular Physics **66**, 577–590 (1989).

[9]M. Schaible, Critical Reviews in Solid State and Materials Sciences **24**, 265–323 (1999).

[10]K. Liew, X. He, and C. Wong, Acta Materialia **52**, 2521–2527 (2004).

[11]Z. Makrodimitri, A. Heller, T. M. Koller, M. H. Rausch, M. S. Fleys, A. R. Bos, G. P. van der Laan, A. P. Froba, and I. G. Economou, The Journal of Chemical Thermodynamics **91**, 101–107 (2015).

[12]L. Deng and J. Du, Journal of Non-Crystalline Solids **453**, 177–194 (2016).

[13]E. Ruckenstein and H. Liu, Industrial & Engineering Chemistry Research **36**, 3927–3936 (1997).

[14]G. V. Kharlamov and S. V. Zhilkin, Journal of Physics: Conference Series **899**, 052009 (2017).

[15]Y. Zhu, X. Lu, J. Zhou, Y. Wang, and J. Shi, Fluid Phase Equilibria **194–197**, 1141–1159 (2002).

[16]R. L. Rowley and M. M. Painter, International Journal of Thermophysics **18**, 1109–1121 (1997).

[17]N. M. Blagoveshchenskii, A. G. Novikov, and V. V. Savostin, Physics of the Solid State **56**, 120–124 (2014).

[18]N. Blagoveshchenskii, A. Novikov, A. Puchkov, V. Savostin, and O. Sobolev, EPJ Web of Conferences **83**, 02018 (2015).

[19]F. Demmel, D. Szubrin, W.-C. Pilgrim, and C. Morkel, Physical Review B **84**, 014307 (2011).

[20]J. J. Potoff and D. A. Bernard-Brunel, The Journal of Physical Chemistry B **113**, 14725–14731 (2009).

[21]C. Avendaño, T. Lafitte, A. Galindo, C. S. Adjiman, G. Jackson, and E. A. Müller, The Journal of Physical Chemistry B **115**, 11154–11169 (2011).

[22]T. Lafitte, A. Apostolakou, C. Avendaño, A. Galindo, C. S. Adjiman, E. A. Müller, and G. Jackson, The Journal of Chemical Physics **139**, 154504 (2013).

[23]C. Avendaño, T. Lafitte, C. S. Adjiman, A. Galindo, E. A. Müller, and G. Jackson, The Journal of Physical Chemistry B (2013), 10.1021/jp306442b.

[24]V. Papaioannou, T. Lafitte, C. Avendaño, C. S. Adjiman, G. Jackson, E. A. Müller, and A. Galindo, The Journal of Chemical Physics **140** (2014), 10.1063/1.4851455.

[25]V. G. Jervell and Wilhelmsen, The Journal of Chemical Physics **158** (2023), 10.1063/5.0149865.

[26]N. Melzi, L. Khaouane, S. Hanini, M. Laidi, Y. Ammi, and H. Zentou, Journal of Applied Mechanics and Technical Physics **61**, 207–216 (2020).

[27]J. P. Allers, F. H. Garzon, and T. M. Alam, Physical Chemistry Chemical Physics **23**, 4615–4623 (2021).

[28]J. P. Aniceto, B. Zêzere, and C. M. Silva, Journal of Molecular Liquids **326**, 115281 (2021).

[29]J. Šlepavičius, A. Patti, J. L. McDonagh, and C. Avendaño, The Journal of Chemical Physics **159**, 024127 (2023).

[30]J. Koza, Statistics and Computing **4**, 87 (1994).

[31]K. Papastamatiou, F. Sofos, and T. E. Karakasidis, AIP Advances **12**, 025004 (2022).

[32]S.-M. Udrescu and M. Tegmark, Science Advances **6**, eaay2631 (2020).

[33]M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, Z. Yao, and A. Aspuru-Guzik, Nature Reviews Physics **4**, 761–769 (2022).

[34]T. M. Alam, J. P. Allers, C. J. Leverant, and J. A. Harvey, The Journal of Chemical Physics **157**, 014503 (2022).

[35]S. H. Huang and M. Radosz, Industrial & Engineering Chemistry Research **29**, 2284–2294 (1990).

[36]A. Pakravesh, F. Zarei, and H. Zarei, Fluid Phase Equilibria **538**, 113024 (2021).

[37]D. Siderius, "NIST standard reference simulation website - SRD 173," (2017).

[38]N. Ramrattan, C. Avendaño, E. Müller, and A. Galindo, Molecular Physics **113**, 932–947 (2015).

[39]F. London, Transactions of the Faraday Society **33**, 8b (1937).

[40]H. Margenau, Reviews of Modern Physics **11**, 1–35 (1939).

[41]S. Dufal, T. Lafitte, A. Galindo, G. Jackson, and A. J. Haslam, AIChE Journal **61**, 2891–2912 (2015).

[42]A. Aasen, M. Hammer, A. Ervik, E. A. Müller, and O. Wilhelmsen, The Journal of Chemical Physics **151**, 064508 (2019).