Supplementary Information for: 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: 8 December 2024)

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

I. DIPOLE AND QUADRUPOLE MOMENTS OF SELECTED SUBSTANCES

TABLE S1. Gas-phase dipole and quadrupole moments for the molecules studied in this work. The data has been taken from the NIST 101 Computational Chemistry Comparison and Benchmark Database[1].

Molecule	Dipole Moment	Quadrupole moment		
	μ	Q_{xx}	Q_{yy}	Q_{zz}
CH_4	-	-	-	-
C_2H_6	-	0.336	0.336	-0.673
C_2H_4	-	-3.160	1.480	1.670
CH_3OH	1.672	-	-	-
Ar	-	-	-	-
Kr	-	-	-	-
Xe	-	-	-	-
CHF_3	1.645	-1.800	-1.800	3.600
CH_3F	1.847	0.700	0.700	-1.400
CF_4	-	-	-	-
CF_3Cl	0.500	-	-	-
H_2	-	-0.260	-0.260	0.520
D_2	-	-	-	-
CO_2	-	2.139	2.139	-4.278
NH_3	1.476	1.160	1.160	-2.320
SF_6	-	-	-	-

II. THERMODYNAMIC PARAMETERS OBTAINED FROM THERMODYNAMIC FITTING

Tables S2, S3, S4, S5 and S6 present all of the obtained Mie intermolecular parameters obtained by the thermodynamic fitting. For the majority of the fluids F_1 and F_2 produced very similar values of α. The the values of σ, ε, λ_r and λ_a for both of the F^{LJ} objective functions, while they are dissimilar for the F^{Mie} objective function.

TABLE S2. Summary of the Mie intermolecular parameters for light hydrocarbons and alcohols obtained by all objective functions studied.

Molecule	Objective Function	σ (Å)	ε (k_{B}/K)	λ_r	λ_a	α
CH_4	F_1^{Mie}	3.748	154.14	12.66	6.03	0.85
	F_2^{Mie}	3.754	160.53	14.15	5.98	0.80
	F_1^{LJ}	3.745	149.19	12.00	6.00	0.89
	F_2^{LJ}	3.727	149.21	12.00	6.00	0.89
C_2H_6	F_1^{Mie}	4.291	298.83	18.33	6.00	0.69
	F_2^{Mie}	4.281	299.88	18.15	6.04	0.68
	F_1^{LJ}	4.142	258.10	12.00	6.00	0.89
	F_2^{LJ}	4.036	259.95	12.00	6.00	0.89
C_2H_4	F_1^{Mie}	4.147	278.78	18.94	6.02	0.67
	F_2^{Mie}	4.136	281.30	19.70	6.00	0.66
	F_1^{LJ}	4.053	234.34	12.00	6.00	0.89
	F_2^{LJ}	3.897	237.31	12.00	6.00	0.89
CH_3OH	F_1^{Mie}	4.082	747.77	27.64	7.02	0.44
	F_2^{Mie}	3.879	683.33	33.19	5.93	0.55
	F_1^{LJ}	3.771	546.22	12.00	6.00	0.89
	F_2^{LJ}	3.771	546.84	12.00	6.00	0.89

TABLE S3. Summary of the Mie intermolecular parameters for noble gases obtained by all objective functions studied.

Molecule	Objective Function	σ (Å)	ε (k_B/K)	λ_r	λ_a	α
Ar	F_1^{Mie}	3.405	126.54	10.13	7.06	0.80
	F_2^{Mie}	3.429	126.67	14.23	5.99	0.80
	F_1^{LJ}	3.408	117.29	12.00	6.00	0.89
	F_2^{LJ}	3.404	116.97	12.00	6.00	0.89
Kr	F_1^{Mie}	3.655	158.01	14.98	5.40	0.93
	F_2^{Mie}	3.663	176.34	14.28	5.98	0.80
	F_1^{LJ}	3.640	163.23	12.00	6.00	0.89
	F_2^{LJ}	3.637	162.70	12.00	6.00	0.89
Xe	F_1^{Mie}	3.981	228.14	14.12	5.67	0.88
	F_2^{Mie}	3.996	224.93	17.48	5.28	0.89
	F_1^{LJ}	3.968	226.44	12.00	6.00	0.89
	F_2^{LJ}	3.964	225.71	12.00	6.00	0.89

TABLE S4. Summary of the Mie intermolecular parameters for halogenated methanes obtained by all objective functions studied.

Molecule	Objective Function	σ (Å)	ε (k_B/K)	λ_r	λ_a	α
CHF_3	F_1^{Mie}	4.224	352.82	31.47	5.93	0.56
	F_2^{Mie}	4.222	357.90	29.99	6.09	0.54
	F_1^{LJ}	3.916	279.16	12.00	6.00	0.89
	F_2^{LJ}	3.915	279.06	12.00	6.00	0.89
CH_3F	F_1^{Mie}	3.888	346.16	22.55	5.97	0.63
	F_2^{Mie}	3.902	346.56	25.54	5.84	0.62
	F_1^{LJ}	3.639	284.46	12.00	6.00	0.89
	F_2^{LJ}	3.639	284.10	12.00	6.00	0.89
CF_4	F_1^{Mie}	4.344	256.87	28.85	6.00	0.56
	F_2^{Mie}	4.380	265.77	29.74	6.20	0.53
	F_1^{LJ}	4.216	191.49	12.00	6.00	0.89
	F_2^{LJ}	4.042	194.87	12.00	6.00	0.89
CF_3Cl	F_1^{Mie}	4.668	329.43	24.82	5.98	0.60
	F_2^{Mie}	4.652	324.17	28.59	5.73	0.61
	F_1^{LJ}	4.358	273.68	12.00	6.00	0.89
	F_2^{LJ}	4.355	273.69	12.00	6.00	0.89

TABLE S5. Summary of the Mie intermolecular parameters for hydrogen isotopes obtained by all objective functions studied.

Molecule	Objective Function	σ (Å)	ε (k_B/K)	λ_r	λ_a	α
H_2	F_1^{Mie}	3.409	10.33	5.69	5.67	2.15
	F_2^{Mie}	3.440	16.40	7.92	5.93	1.30
	F_1^{LJ}	3.437	22.24	12.00	6.00	0.89
	F_2^{LJ}	3.455	22.29	12.00	6.00	0.89
D_2	F_1^{Mie}	3.196	18.40	6.68	6.25	1.47
	F_2^{Mie}	3.190	17.29	6.46	6.15	1.57
	F_1^{LJ}	3.203	27.76	12.00	6.00	0.89
	F_2^{LJ}	3.247	27.58	12.00	6.00	0.89

TABLE S6. Summary of the Mie intermolecular parameters for molecular fluids obtained by all objective functions studied.

Molecule	Objective Function	σ (Å)	ε (k_B/K)	λ_r	λ_a	α
CO_2	F_1^{Mie}	3.818	449.44	18.24	10.03	0.35
	F_2^{Mie}	3.741	353.55	23.00	6.66	0.52
	F_1^{LJ}	3.675	250.85	12.00	6.00	0.89
	F_2^{LJ}	3.672	249.98	12.00	6.00	0.89
NH_3	F_1^{Mie}	3.443	485.01	30.41	6.04	0.54
	F_2^{Mie}	3.442	558.57	17.92	8.64	0.42
	F_1^{LJ}	3.191	368.98	12.00	6.00	0.89
	F_2^{LJ}	3.190	368.52	12.00	6.00	0.89
SF_6	F_1^{Mie}	4.902	468.11	18.06	10.04	0.36
	F_2^{Mie}	4.956	458.31	31.62	7.71	0.37
	F_1^{LJ}	4.711	262.25	12.00	6.00	0.89
	F_2^{LJ}	4.697	261.62	12.00	6.00	0.89

III. MACHINE-LEARNING MODELS

In this Section, we describe the details of the machine-learning algorithms used in this work to correlate the self-diffusion coefficient. These algorithms correspond to k-nearest neighbours (KNN), artificial neural network (ANN), and symbolic regression (SR). For all the algorithms, 80% of the data obtained from MD simulations from our previous work[2] are used to train the model and 20% to test the model. The MD data is first normalised with respect to the diffusion coefficient of hard spheres at infinite delusion D_0, follow by a normalisation in the interval [-1,1] for the training of both ANN and KNN, respectively. The calculation of method-training accuracy is done using 10-fold cross-validation (CV10)[3]. Both ANN and KNN methods have been implemented in Python 3 using the *scikit-learn* library version 1.2.2[4], while SR is implemented from the *gplearn* library version 0.4.1 is used[5].

The ANN model consists of a single hidden layer consisting of 28 nodes using reLU activation function, the training is done for 1000 epochs and the *lbfgs* solver is used for back-propagation. For our KNN model, the hyperparameters that provide the best performance are k= 4 (number of neighbours), the neighbouring points weighted by distance, and the power parameter of 4 in the Minkowski metric. The hyperparameter space for SR corresponds to a population size of 5000, 50 generations, the AARD metric and a parsimony coefficient of 0.3. The operation set used in this work comprises additions, subtractions, multiplications, divisions, exponential, square roots of the absolute values, and the natural logarithm of the absolute values.

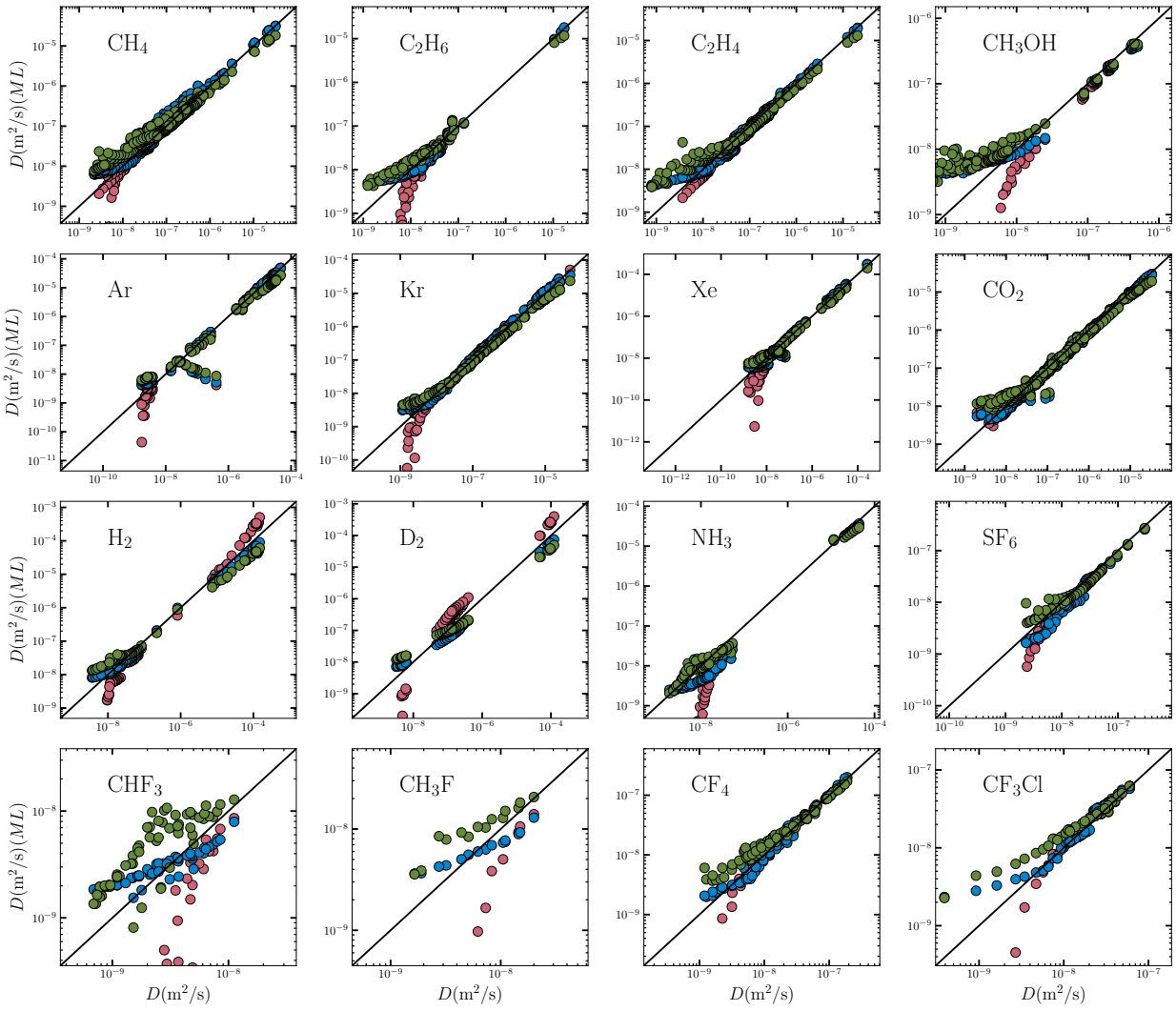

FIG. S1. The parity plots for predicting the self-diffusion coefficients of the fluids studied, modelled by F_2^{Mie}. The ANN model is presented by magenta circles, KNN - blue circles and SR - green circles.

IV. PREDICTIONS OF THE SELF-DIFFUSION COEFFICIENT USING THE MOLECULAR PARAMETERS OBTAINED FROM THE THERMODYNAMIC FITTING

Figures S1, S2 and S3 present the parity plots for all fluids from obtained thermodynamic models. All figures show very similar predictive ability in high D regions, with the main differences found at lower values of D. This corresponds to finding presented in Tables S7, S8, S9, S10 and S8, where the AARD varies highly between the methods, but R^2 is relatively constant and high, with $R^2 > 0.99$ for the large majority of fluids and methods.

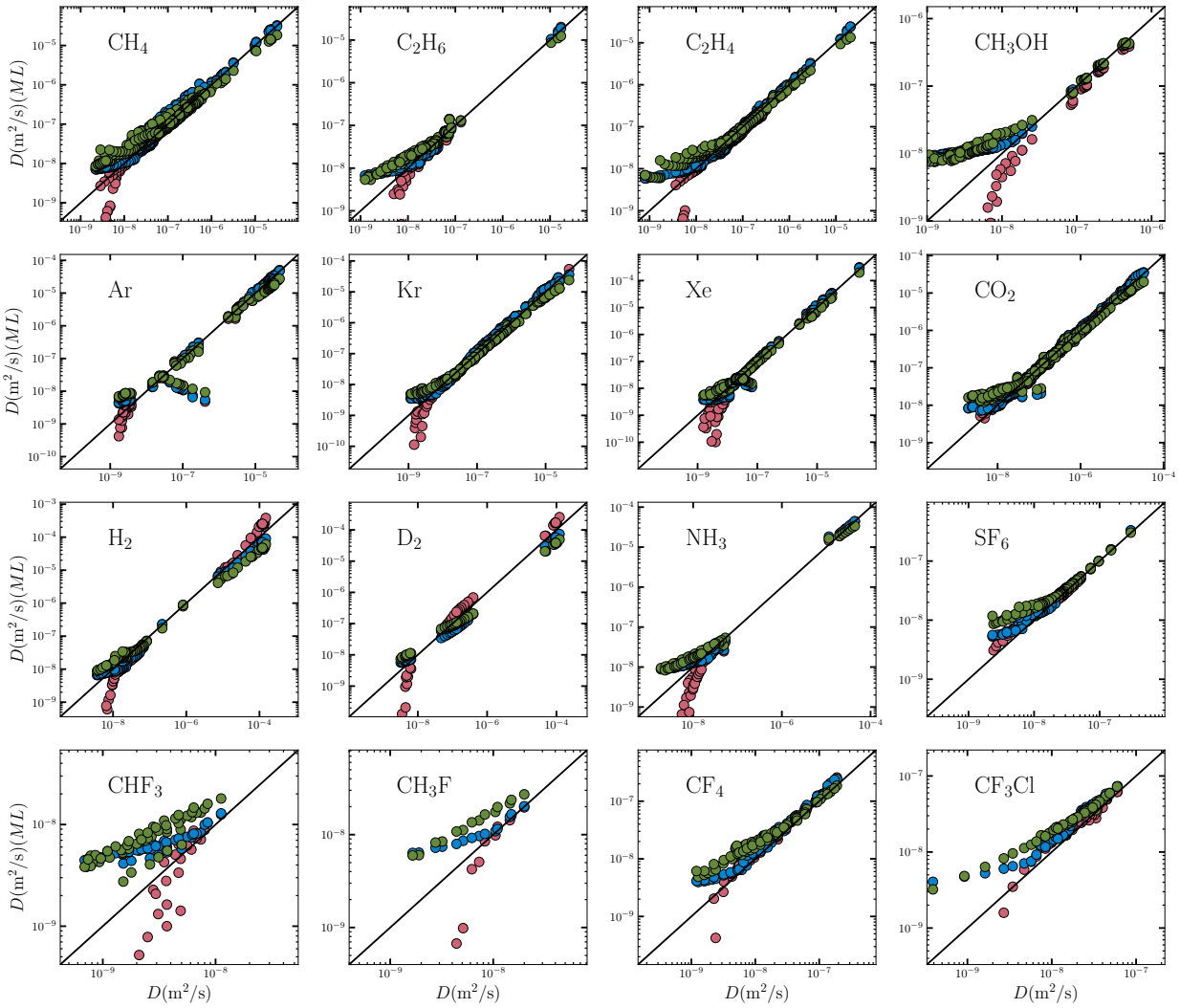

FIG. S2. The parity plots for predicting the self-diffusion coefficients of the fluids studied, modelled by F_1^{LJ}. The ANN model is presented by magenta circles, KNN - blue circles and SR - green circles.

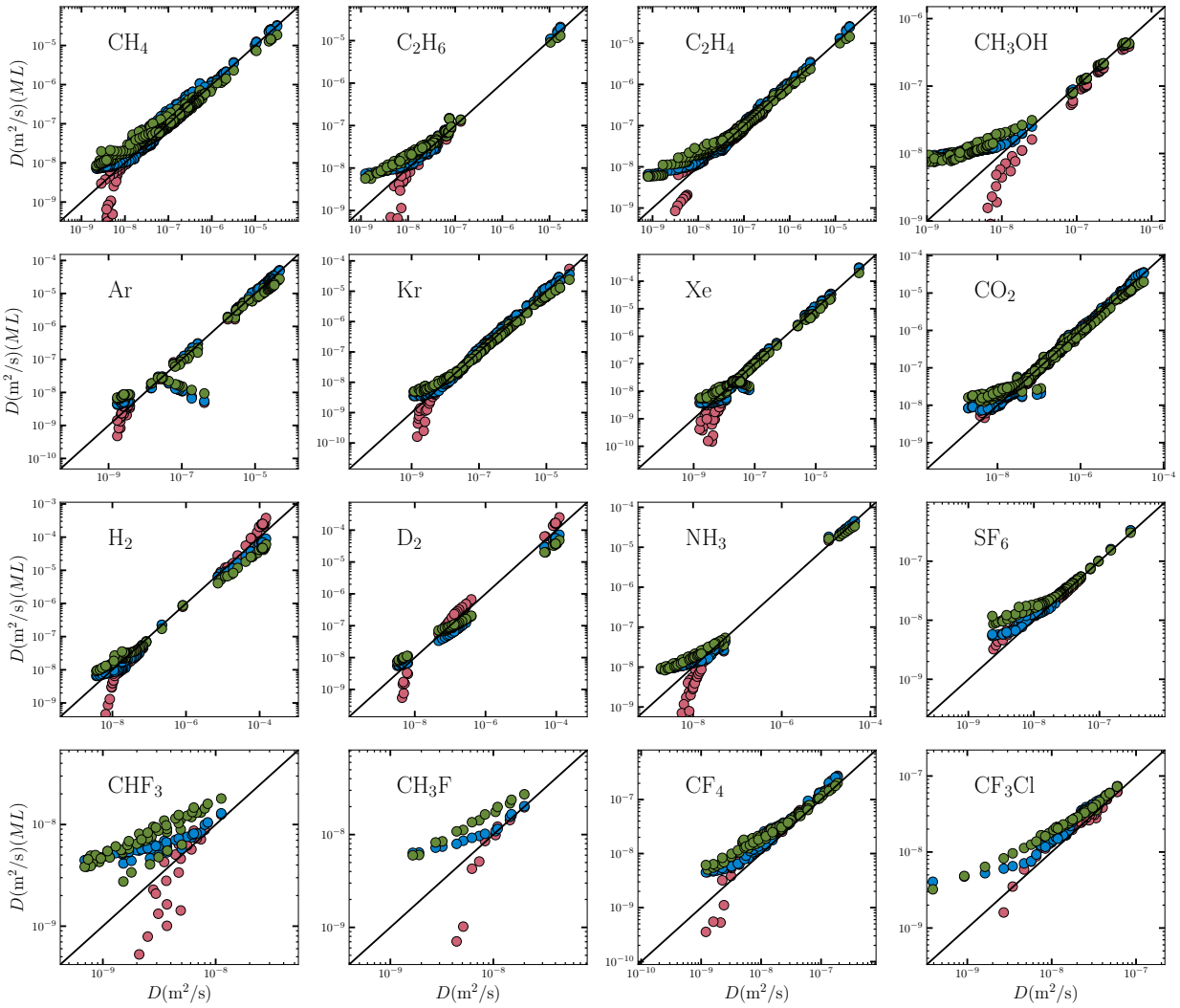

FIG. S3. The parity plots for predicting the self-diffusion coefficients of the fluids studied, modelled by F_2^{LJ}. The ANN model is presented by magenta circles, KNN - blue circles and SR - green circles.

TABLE S7. Summary of the AARD and R^2 descriptors for hydrocarbons predicted by the different ML methods applied in this work using the model obtained all objective functions studied.

Fluid	Potential Function	KNN		ANN		SR	
		AARD	R^2	AARD	R^2	AARD	R^2
CH_4	F_1^{Mie}	26.98%	0.9977	25.16%	0.9978	47.94%	0.9960
	F_2^{Mie}	25.81%	0.9980	25.53%	0.9985	45.86%	0.9960
	F_1^{LJ}	27.39%	0.9979	24.99%	0.9979	49.70%	0.9960
	F_2^{LJ}	27.85%	0.9979	24.48%	0.9979	49.31%	0.9960
C_2H_6	F_1^{Mie}	38.51%	0.9955	73.15%	0.9992	54.00%	0.9981
	F_2^{Mie}	38.59%	0.9956	71.94%	0.9992	52.32%	0.9981
	F_1^{LJ}	50.98%	0.9963	60.23%	0.9952	74.71%	0.9981
	F_2^{LJ}	59.73%	0.9966	55.52%	0.9954	81.96%	0.9981
C_2H_4	F_1^{Mie}	28.99%	0.9989	38.76%	0.9997	43.24%	0.9963
	F_2^{Mie}	28.15%	0.9994	37.65%	0.9997	44.37%	0.9963
	F_1^{LJ}	31.96%	0.9937	38.11%	0.9960	48.77%	0.9963
	F_2^{LJ}	35.03%	0.9968	33.64%	0.9958	44.50%	0.9963
CH_3OH	F_1^{Mie}	120.08%	0.9951	532.22%	0.9960	93.79%	0.9939
	F_2^{Mie}	270.65%	0.9958	509.83%	0.9958	217.65%	0.9955
	F_1^{LJ}	538.43%	0.9962	690.07%	0.9953	385.20%	0.9955
	F_2^{LJ}	538.68%	0.9962	690.98%	0.9953	384.16%	0.9955

TABLE S8. Summary of the AARD and R^2 descriptors for noble gases predicted by the different ML methods applied in this work using the model obtained all objective functions studied.

Fluid	Potential Function	KNN		ANN		SR	
		AARD	R^2	AARD	R^2	AARD	R^2
Ar	F_1^{Mie}	36.41%	0.9903	20.62%	0.9984	73.92%	0.9861
	F_2^{Mie}	35.35%	0.9903	21.15%	0.9984	73.59%	0.9861
	F_1^{LJ}	38.63%	0.9851	20.61%	0.9982	79.84%	0.9861
	F_2^{LJ}	38.82%	0.9851	20.68%	0.9982	80.10%	0.9861
Kr	F_1^{Mie}	25.94%	0.9633	23.52%	0.9897	48.53%	0.9773
	F_2^{Mie}	23.60%	0.9659	21.28%	0.9914	43.16%	0.9773
	F_1^{LJ}	26.81%	0.9599	22.70%	0.9896	46.73%	0.9773
	F_2^{LJ}	26.99%	0.9600	22.55%	0.9897	46.86%	0.9773
Xe	F_1^{Mie}	20.31%	0.9997	23.75%	0.9997	37.68%	0.9999
	F_2^{Mie}	20.43%	0.9997	24.37%	0.9997	38.31%	0.9999
	F_1^{LJ}	20.36%	0.9997	23.52%	0.9997	38.15%	0.9999
	F_2^{LJ}	20.51%	0.9997	23.30%	0.9997	38.34%	0.9999

TABLE S9. Summary of the AARD and R^2 descriptors for chlorofluorocarbons predicted by the different ML methods applied in this work using the model obtained all objective functions studied.

Fluid	Potential Function	KNN		ANN		SR	
		AARD	R^2	AARD	R^2	AARD	R^2
CHF_3	F_1^{Mie}	55.23%	0.9148	204.51%	0.8886	116.98%	0.6404
	F_2^{Mie}	50.12%	0.9040	198.89%	0.8881	112.85%	0.6438
	F_1^{LJ}	187.98%	0.9071	184.93%	0.8618	226.80%	0.8476
	F_2^{LJ}	188.02%	0.9070	184.70%	0.8618	226.99%	0.8476
CH_3F	F_1^{Mie}	40.74%	0.9756	125.47%	0.9851	89.90%	0.7925
	F_2^{Mie}	38.12%	0.9784	125.57%	0.9852	71.97%	0.9387
	F_1^{LJ}	78.29%	0.9707	88.79%	0.9699	122.06%	0.9764
	F_2^{LJ}	78.37%	0.9706	88.36%	0.9697	122.44%	0.9764
CF_4	F_1^{Mie}	21.69%	0.9923	29.37%	0.9948	49.15%	0.9959
	F_2^{Mie}	17.83%	0.9912	27.42%	0.9946	47.37%	0.9962
	F_1^{LJ}	45.80%	0.9876	39.93%	0.9917	82.02%	0.9952
	F_2^{LJ}	66.35%	0.9838	53.10%	0.9903	93.45%	0.9951
CF_3Cl	F_1^{Mie}	23.89%	0.9664	32.77%	0.9824	42.81%	0.9883
	F_2^{Mie}	25.52%	0.9641	33.68%	0.9815	44.93%	0.9884
	F_1^{LJ}	62.40%	0.9809	43.62%	0.9699	80.66%	0.9882
	F_2^{LJ}	62.51%	0.9808	43.77%	0.9699	80.91%	0.9882

TABLE S10. Summary of the AARD and R^2 descriptors for hydrogens predicted by the different ML methods applied in this work using the model obtained all objective functions studied.

Fluid	Potential Function	KNN		ANN		SR	
		AARD	R^2	AARD	R^2	AARD	R^2
H_2	F_1^{Mie}	40.08%	0.9928	117.26%	0.9753	105.93%	0.9924
	F_2^{Mie}	32.12%	0.9928	90.34%	0.9767	78.26%	0.9924
	F_1^{LJ}	31.49%	0.9927	65.02%	0.9777	55.21%	0.9924
	F_2^{LJ}	31.82%	0.9927	66.18%	0.9777	56.81%	0.9924
D_2	F_1^{Mie}	48.80%	0.9980	125.67%	0.9873	57.05%	0.9979
	F_2^{Mie}	49.21%	0.9980	138.40%	0.9872	59.15%	0.9979
	F_1^{LJ}	43.23%	0.9980	54.46%	0.9881	41.87%	0.9979
	F_2^{LJ}	43.79%	0.9980	52.59%	0.9881	42.78%	0.9979

TABLE S11. Summary of the AARD and R^2 descriptors for molecular fluids predicted by the different ML methods applied in this work using the model obtained all objective functions studied.

Fluid	Potential Function	KNN		ANN		SR	
		AARD	R^2	AARD	R^2	AARD	R^2
CO_2	F_1^{Mie}	25.88%	0.9957	25.28%	0.9974	29.81%	0.9856
	F_2^{Mie}	21.26%	0.9963	20.78%	0.9973	32.92%	0.9856
	F_1^{LJ}	20.11%	0.9911	20.23%	0.9906	43.36%	0.9856
	F_2^{LJ}	20.29%	0.9920	20.38%	0.9908	43.49%	0.9856
NH_3	F_1^{Mie}	33.63%	0.9933	136.57%	0.9980	47.71%	0.9865
	F_2^{Mie}	38.69%	0.9917	128.12%	0.9965	24.99%	0.9865
	F_1^{LJ}	88.11%	0.9943	110.21%	0.9982	115.14%	0.9865
	F_2^{LJ}	88.15%	0.9945	109.64%	0.9982	115.51%	0.9865
SF_6	F_1^{Mie}	25.37%	0.9951	25.54%	0.9979	23.95%	0.9948
	F_2^{Mie}	28.65%	0.9950	28.96%	0.9980	26.86%	0.9940
	F_1^{LJ}	28.24%	0.9981	20.44%	0.9958	75.27%	0.9967
	F_2^{LJ}	29.32%	0.9982	21.66%	0.9958	76.35%	0.9967

V. MOLECULAR PARAMETERS OBTAINED FROM FITTING THE SELF-DIFFUSION COEFFICIENT

Figure S4 presents all of the obtained combinations of Mie parameters in the range $0.25 < \alpha < 1$ using the self-diffusion coefficient based objective function. The plots additionally present the values obtained from the thermodynamic objective functions.

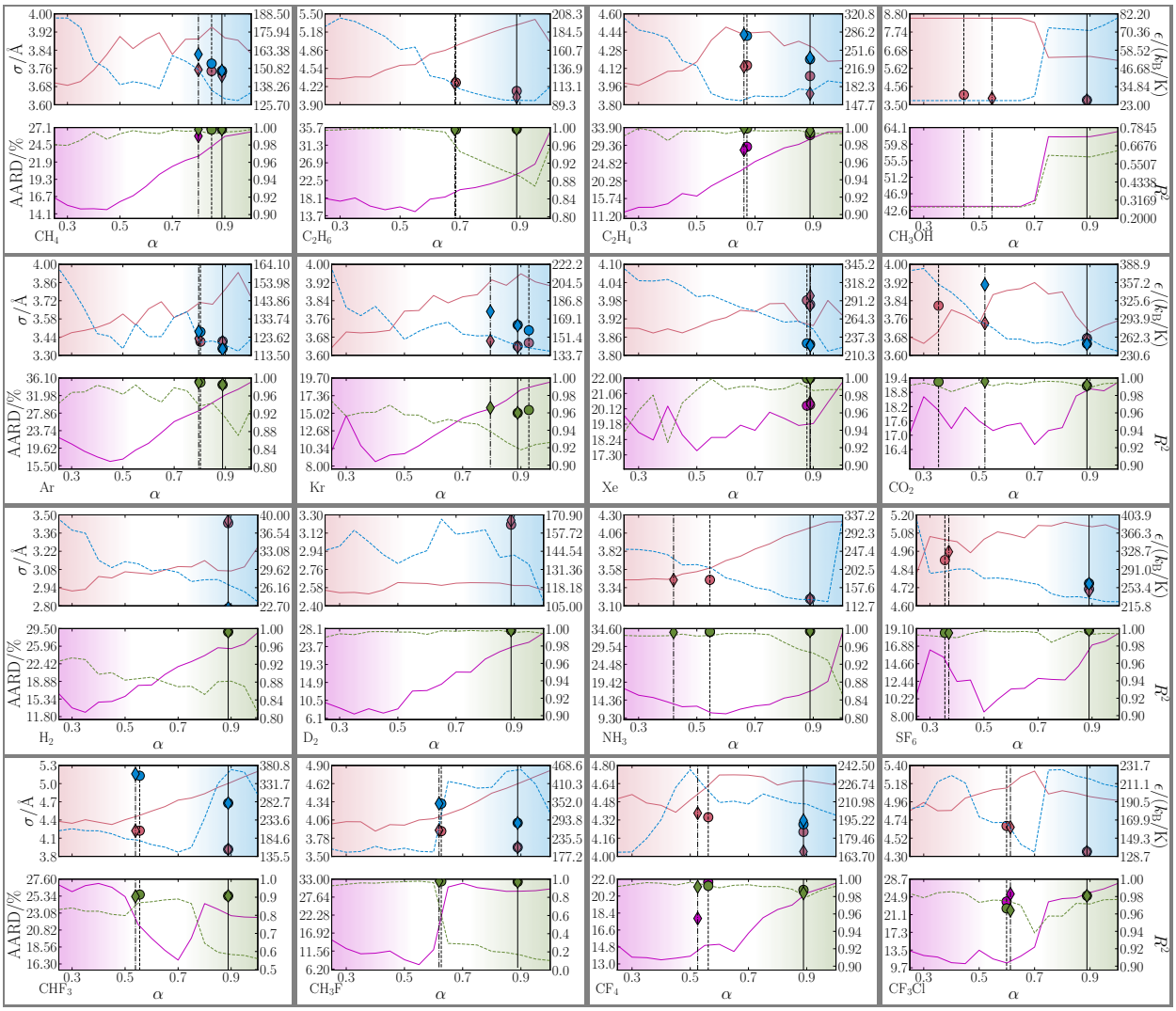

FIG. S4. The results of fitting the self-diffusion coefficients at different values of α to obtain the best Mie intermolecular parameters σ and ε. The obtained intermolecular parameters (top) and errors (bottom) are presented for each fluid. σ is presented in red, ε - blue, AARD - magenta and R^2 - green. Additionally, the α of the Lennard-Jones potential (solid black line) and the potentials used in the previous section - F_1^{Mie} (dashed line) and F_2^{Mie} (dash-dot line). The intermolecular parameter and error values obtained from the previous section are also presented for F_1 (circles) and F_2 (diamonds).

REFERENCES

[1]R. Johnson, en"Computational chemistry comparison and benchmark database, nist standard reference database 101," (2002).

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

[3]M. Kuhn and K. Johnson, *Applied Predictive Modeling* (Springer New York, 2013).

[4]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[5]T. Stephens, "Gplearn: Genetic programming in python, with a scikit-learn inspired API," (2015).