Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

[Supplemental Material]

Artificial Neural Network Prediction of Self-Diffusion in Pure Compounds over Multiple Phase Regimes

Joshua P. Allers, [a] Fernando H. Garzon [b] and Todd M. Alam *[c]

Background

The following section provides some background on the existing empirical and theoretical equations for predicting the self-diffusion in pure solutions and highlight the parameters used in each. For a more in-depth discussion on the history of predicting self-diffusion see the review published by Silva and Liu on hard sphere models and the comprehensive review by Suarez-Iglesias.[1, 2]

J.H. Dymond (1974)[3]

Dymond presents the following equation for prediction of self-diffusion in liquids. The equation builds on and corrects the modified Enskog theory for hard spheres:

$$D_{DHB}^{real} = B\sqrt{T}(V - V_D) \quad \textbf{(S1)}$$

Here, *B* and V_D are adjustable parameters that are specific to each compound. Dymond proposes that the diffusion is dependent on the temperature of the system (T) and the molar volume (V). The molar mass is also wrapped up in the constant *B*. Dymond shows good agreement when predicting carbon tetrachloride and methane diffusion, but the model struggled in predicting liquid argon diffusion. Silva and Liu compared the performance of Dymond's model to 2471 experimental diffusion constants and achieved and overall AAD of 99.66%.

Lee and Thodos (1988)[4]

Lee and Thodos present the following correlation for predicting the self-diffusion at all state conditions:

$$\frac{\tau}{D_{LT}^{real}\delta_t}10^5 = exp(B_1 x^{B_3} + B_2 x^{B_4}) - 1 \quad \textbf{(S2)}$$

$$\delta_t = M^{1/2} T_t^{-1/2} V_{lt}^{-2/3} \quad \textbf{(S3)}$$

$$x = \frac{\omega}{\tau^{E_1}\omega^{E_2}} \quad \textbf{(S4)}$$

Here, δ_t is a self-diffusivity parameter, T_t is the triple-point temperature, V_{lt} is the liquid molar volume at the triple point, M is the molar mass, $\tau = T/T_t$ is the normalized temperature, x is a density-temperature variable, $\omega = \rho/\rho_{lt}$ is a normalized density, ρ_{lt} is the liquid density at the triple point, E_1 and E_2 are constants found to be 0.09 and 3.1, respectively, and B_1, B_2, B_3, and B_4 are parameters specific to each compound. The equation proposed by Lee and Thodos is dependent on the temperature, molar volume and density at the triple point along with the molar mass and system temperature. Lee and Thodos tested and fit the model on 58 molecules and 975 experimental diffusion values. The equation was also tested by Silva and Liu against their 2471 diffusion values and achieved an overall AAD of 10.13%.

Salim and Trebble (1995)[5]

Salim and Trebble extended the Lennard Jones (LJ) diffusion predicitons to real fluids by considering each molecule as a chain of LJ particles. This introduces a new parameter (N), which is specific to each molecule.

$$\frac{D_{LJC}^{real}}{D_{0,HSC}} = \left\{\left(1 - \frac{\rho\sigma_{BS}'^3}{1.09}\right)\left[1 + \left(\rho\sigma_{BS}'^3\right)^2\left(0.4 - 0.83\left(\rho\sigma_{BS}'^3\right)^2\right)\right]\right\}^N \times exp\left(-\frac{0.5\varepsilon_{LJ}/k_B}{T}\right) \quad \textbf{(S5)}$$

$$\sigma_{BS}'(T^*) = \sigma_{LJ}2^{\frac{1}{6}}\left[1 + (2T^*)^{\frac{1}{2}}\right]^{-1/6} \quad \textbf{(S6)}$$

$$D_{0,HSC} = \frac{3}{8\rho_0\sigma'^2}\left(\frac{k_BT}{\pi m'}\right)^{\frac{1}{2}} \quad \textbf{(S7)}$$

Here, $D_{0,HSC}$ is the self-diffusion at ideal gas conditions for a hard sphere chain fluid, σ_{BS}' is the effective hard sphere diameter (EHSD) using the Boltzmann criteria and Speedy form,[6] ε_{LJ} is the Lennard Jones interaction energy, σ_{LJ}' is the Lennard Jones diameter, k_B is the Boltzmann constant, N is the chain parameter describing the number of Lennard Jones molecules chained together, $T^* = k_BT/\varepsilon_{LJ}$ is the reduced temperature, ρ_0 is the density at ideal gas conditions, σ' is the ideal gas diameter and m' is the mass. By extending the Lennard Jones pair-potential to real fluids, Salim and Trebble propose that the diffusion is dependent on the inter-atomic interaction energies and the effective diameters of the molecules. The use of the chain parameter also leads to a dependence on the length of each molecule along with the usual dependence on temperature and density. Silva and Liu do not report a global AAD for this model, but Salim and Trebble report the AAD for 13 alkanes up to 154 carbons with the largest being 13.32%.

Ruckenstein and Liu (1997)[7]

Ruckenstein and Liu extend the Lennard Jones diffusion to real fluids by treating the molecules as Rough Lennard Jones (RLJ) particles. This involves multiplying the Lennard Jones diffusion constant by a new parameter that accounts for frictional losses during collisions (A_D).

$$D_{RLJ}^{real} = A_D(\omega)\frac{k_BT}{\frac{8}{3}\rho\sigma_{BAH}^2(\pi m k_BT)^{1/2}\left[\frac{g(\sigma_{BAH})}{F(\rho^*)} + \frac{0.4}{T^{*1.5}}\right]} \quad \textbf{(S8)}$$

$$\sigma_{BAH} = 1.1532\sigma_{LJ}\left[1 + (1.89757T^*)^{\frac{1}{2}}\right]^{-1/6} \quad \textbf{(S9)}$$

$$F(\rho^*) = 1 + 0.94605\rho^{*1.5} + 1.4022\rho^{*3} + 5.6898\rho^{*5} + 2.6626\rho^{*7} \quad \textbf{(S10)}$$

$$A_D(\omega) = 0.9673 - 0.2527\omega - 0.70\omega^2 \quad \textbf{(S11)}$$

$$g(\sigma) = \frac{1-0.5\eta}{(1-\eta)^3} \quad \textbf{(S12)}$$

Here, ω is the acentric factor describing how much the molecule deviates from a hard sphere, σ_{BAH} is the effective hard sphere diameter using the Boltzmann criteria and Amotz-Herschbach form,[8] $g(\sigma)$ is the pair distribution function, and $\eta = \frac{1}{6}\pi\rho^*$ is the packing fraction. Silva and Liu report an overall AAD of 7.33% over 1,822 experimental diffusion constants.

Liu, Silva, and Macedo (1998a, 1998b)[9, 10]

Liu *et al.* extends the Lennard Jones and Square-Well (SW) pair-potential to real fluids. Expressions for LJ and SW diffusion are developed by fitting molecular dynamics (MD) data. The extension into real fluids is accomplished by using the effective hard sphere diameter method. The Lennard Jones expression is shown below:

$$D_{LJ2}^{real} = \frac{21.16}{\rho\sigma_{BLSM}^2}\sqrt{\frac{1000RT}{M}}exp\left\{-\frac{0.75\rho\sigma_{BLSM}^3}{1.2588-\rho\sigma_{BLSM}^3} - \frac{0.27862(\varepsilon_{LJ}/k_B)}{T}\right\} \quad \textbf{(S13)}$$

$$\sigma_{BLSM} = \sigma_{LJ}2^{\frac{1}{6}}\left[1 + (1.3229T^*)^{\frac{1}{2}}\right]^{-1/6} \quad \textbf{(S14)}$$

Here, σ_{BLSM} is the effective hard sphere diameter using the Boltzmann criteria. Liu, Silva and Macedo (LSM) optimized the parameters found by Ben-Amotz and Herschbach to form their own expression for the EHSD. The self-diffusion is dependent on the two LJ parameters, ε_{LJ} and σ_{LJ}, along with the temperature and density of the system. The model was tested against 2,471 experimental diffusion values from 40 compounds and showed and overall AAD of 8.53%. The model struggled with hydrogen-bonding substances, so the authors developed an improved model, published later the same year (ref. 10):

$$D_{LJ4}^{real} = \frac{21.16A_D}{\rho\sigma_{LJ4}^2}\sqrt{\frac{1000RT}{M}}exp\left\{-\frac{0.75\rho\sigma_{LJ4}^3}{1.2588-\rho\sigma_{LJ4}^3} - \frac{E_D}{RT}\right\} \quad \textbf{(S15)}$$

$$\sigma_{LJ4} = \sigma 2^{\frac{1}{6}}\left[1 + \left(\frac{T}{T_D}\right)^{\frac{1}{2}}\right]^{-1/6} \quad \textbf{(S16)}$$

The updated model has 4 adjustable parameters for each molecule: A_D, a roughness factor, E_D, an energy parameter to replace ε_{LJ}, T_D, which allows optimization of the temperature-dependence in the effective diameter, and σ. The authors also define a new EHSD, σ_{LJ4}, using the new parameters. The model was tested against the same 2,471 points and achieved an overall AAD of 4.45%, improving the predictions for the hydrogen-bonding compounds.

Yu and Gao (2000)[11]

Yu and Gao utilize the LJ chain method in combination with an effective hard sphere diameter, similar to Salim and Treble. The LJ chain method introduces the parameter N, which represents the number of LJ particles in the chain:

$$D_{LJC} = \frac{D_{0,HSC}}{\frac{g(\sigma_{BAH})}{F(N,\rho^*)} + \frac{0.4}{T^*}} \quad \textbf{(S17)}$$

$$D_{0,HSC} = \frac{3}{8\rho_0\sigma'^2}\left(\frac{k_BT}{\pi m'}\right)^{\frac{1}{2}} \quad \textbf{(S18)}$$

$$F(N,\rho^*) = f(\rho^*)exp\left[-0.06356(N-1) - \left(0.05212\left(\frac{N-1}{N}\right) + 1.9709\left(\frac{N-1}{N}\right)^2\right)\rho^*\right] \quad \textbf{(S19)}$$

$$f(\rho^*) = 1 + 0.94605\rho^{*1.5} + 1.4022\rho^{*3} + 5.6898\rho^{*5} + 2.6626\rho^{*7} \quad \textbf{(S20)}$$

$$g(\sigma) = \frac{1-0.5\eta}{(1-\eta)^3} \quad \textbf{(S21)}$$

The model is heavily dependent on the density and temperature of the system and is supplemented by *N*, the effective length of the compounds. Silva and Liu compare the model to 1,081 experimental diffusion values and achieve and overall AAD of 4.72%.

Zhu, Lu, Zhou, Wang, and Shi (2002)[12]

Zhu *et. al.* initially fit an empirical equation to Lennard Jones molecular dynamics data. That equation was then extended to real fluids by using effective diameter and energy terms:

$$D_{LJ}^{*} = \frac{3}{8\sqrt{\pi}}\frac{\sqrt{T^{*}}}{\rho^{*}}\left(1-\frac{\rho^{*}}{a(T^{*})b}\right)\left[1+(\rho^{*})^{c}\left(\frac{P_1(\rho^{*}-1)}{P_2(\rho^{*}-1)+(T^{*})^{(P_3+P_4\rho^{*})}}+P_5\right)\right]\times exp\left(-\frac{\rho^{*}}{2T^{*}}\right) \quad \textbf{(S22)}$$

$$\rho^{*} = \rho\sigma_{eff}^{3} = \rho\left(\mu\sqrt[3]{\frac{\rho_c^{*}V_{cm}}{N_A}}\right)^{3} \quad \textbf{(S23)}$$

$$T^{*} = T\frac{k_B}{\varepsilon_{eff}} = T\frac{T_c^{*}}{\xi T_c} \quad \textbf{(S24)}$$

$$\mu = 1+\alpha_B\rho_r+(\beta_B+\gamma_B\rho_r)T_r \quad \textbf{(S25)}$$

$$\xi = 1+\alpha_A\rho_r+(\beta_A+\gamma_A\rho_r)T_r \quad \textbf{(S26)}$$

Here, the constants a, b, c, P_1, P_2, P_3, P_4, and P_5 were fit on the MD data while the constants α_A, β_A, γ_A, α_B, β_B, and γ_B were fit on the experimental diffusion data (see reference for values). V_{cm} is the critical molar volume, , T_c is the critical temperature, $T_c^{*} = T_c\frac{k_B}{\varepsilon_{LJ}}$ is the unitless critical temperature, $T_r = \frac{T}{T_c}$ is the reduced temperature, $\rho_r = \frac{\rho}{\rho_c}$ is the reduced density, and N_A is Avogadro's number. Rather than use the established expressions for the EHSD, Zhu *et al.* elected to predict the LJ parameters using the critical volume, temperature, and density. Zhu *et al.* tested the predictions of the model against 1,226 experimental diffusion values from 17 substances and achieve and overall AAD of 18.51%.

Summary

Each of these models suggests that certain parameters would be important input features in the development of ANNs. The temperature and density of the system are two of the more important parameters to include in a model as they appear in all models mentioned here. Many of the more recent empirical models utilize an effective hard sphere diameter to represent a real molecule rather than the traditional hard sphere representation. The LJ chain models also incorporate an effective length. These would suggest that the size/shape of the molecule would be a valuable feature for model development. On a similar note, Ruckenstein and Liu introduce the acentric factor in their model, which also aims to describe the difference between the real molecule and a hard sphere. Models proposed by Lee and Thodos and Zhu *et al.* were both developed for multiple phases and both use unique parameters. Lee and Thodos use the triple-point properties while Zhu *et al.* use the critical point properties. The fact that the Zhu model has shown good performance over liquid, gas, and supercritical states suggests that the critical properties are likely valuable in the prediction of self-diffusion and should be included as input features in ANN model development. Molar mass frequently appears and is likely necessary for improved ANN self-diffusion models. Many of the parameters mentioned were incorporated as input features in the ANN models in some form.

Computation Time and Hardware

The ANN models were developed on an AMD Ryzen 2600 6-core processor. It takes 5.3 seconds to train and test the B-ANN model, 1 minute and 11 seconds to train and test the log-ANN model, 32.8 seconds to train and test the L-ANN, 5.6 seconds to train and test the S-ANN, 3.29 seconds to train and test the G-ANN, and 40 seconds to run the combined ANN with all three phase-specific ANNs.

Table S1. Performance of log-ANN with different numbers of principal components using DB1.

	# of PCs	5	10	15	19
Train	AAD (%)	12.26	8.40	7.01	6.46
	MSE	9.89e-5	5.14e-5	3.81e-5	3.45e-5
Test	AAD (%)	20.94	9.47	7.87	7.14
	MSE	1.76e-4	5.98e-5	8.99e-5	3.97e-5
All	AAD(%)	13.56	8.56	7.14	6.56
	R^2	0.9940	0.9971	0.9975	0.9911

Table S2. Parameters for the models developed on the dataset with density (DB1, 6,625 points).

Model	B-ANN	Log-ANN	L-ANN	SC-ANN	G-ANN
Random Seed	40	75	80	99	51
Hidden Layer Nodes	4/4	18/15	19/17	20/7	8/3
Activation Function	Hyperbolic Tangent	Hyperbolic Tangent	Hyperbolic Tangent	Hyperbolic Tangent	Hyperbolic Tangent
Learn Rate	0.001	0.001	0.001	0.001	0.001
Epochs	200	305	175	100	240

Table S3. Performance metrics for the B-ANN and log-ANN when density is present (or not) in the database. DB1, which contains density, contains 6,625 points, while DB2 contains 10,569 points.

		With Density (DB1)		Without Density (DB2)	
	Model	B-ANN	Log-ANN	B-ANN	Log-ANN
Train	AAD (%)	1710.1	6.46	63,241	13.66
	MSE	8.64e-6	3.45e-5	5.70e-6	7.83e-5
Test	AAD (%)	1827.1	7.14	65,222	16.57
	MSE	2.57e-6	3.97e-5	1.54e-6	1.26e-4
All	AAD(%)	1727.6	6.56	63,538	14.09
	R^2	0.9836	0.9911	0.9861	0.8549

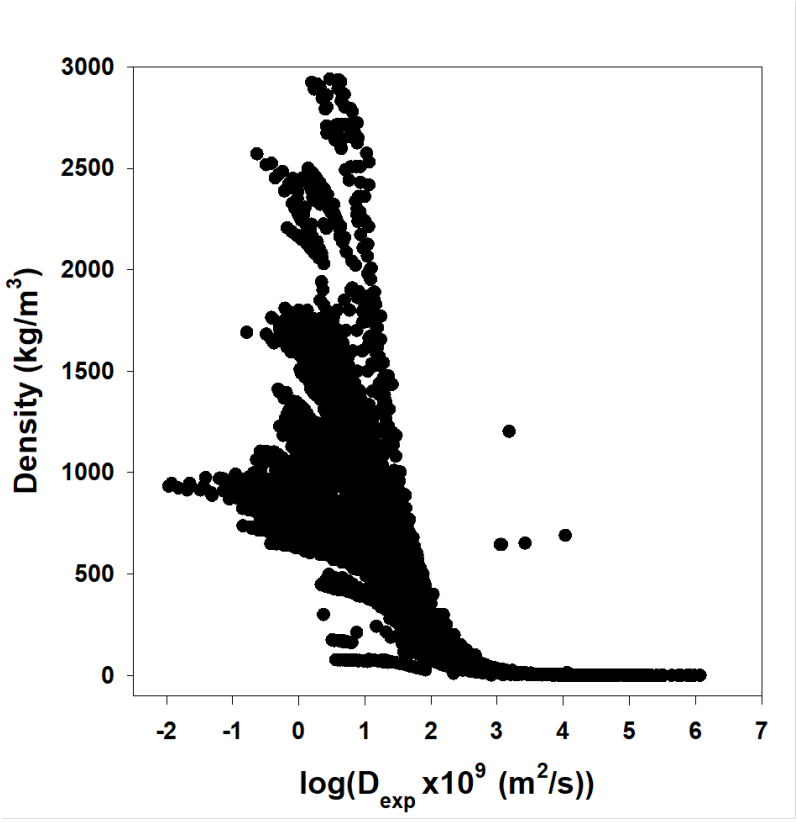

Figure S1. Density plotted versus the log10 of self-diffusion.

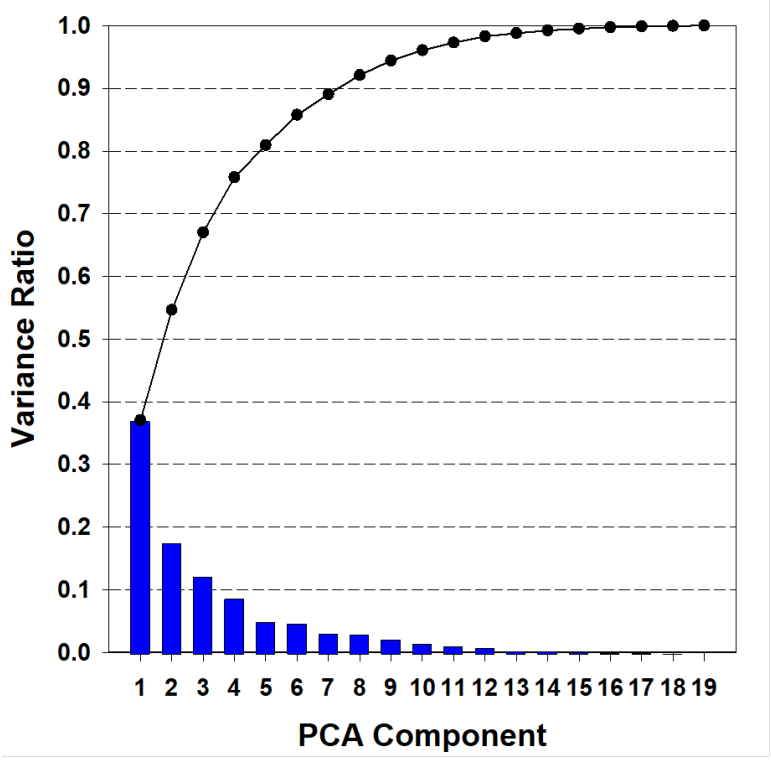

Figure S2. Individual (blue bars) and cumulative explained variance (dotted line) for principal component input features in database DB1.

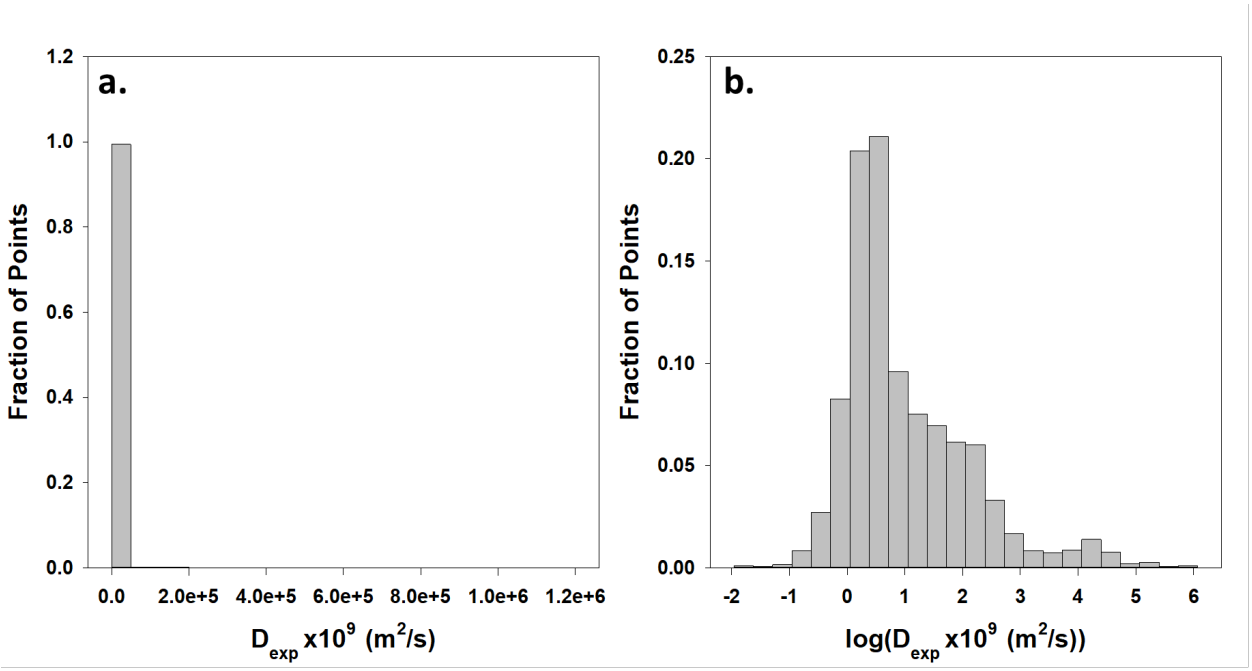

Figure S3. Distribution plots for a) the raw diffusion values and b) the log transformed diffusion values using DB1.

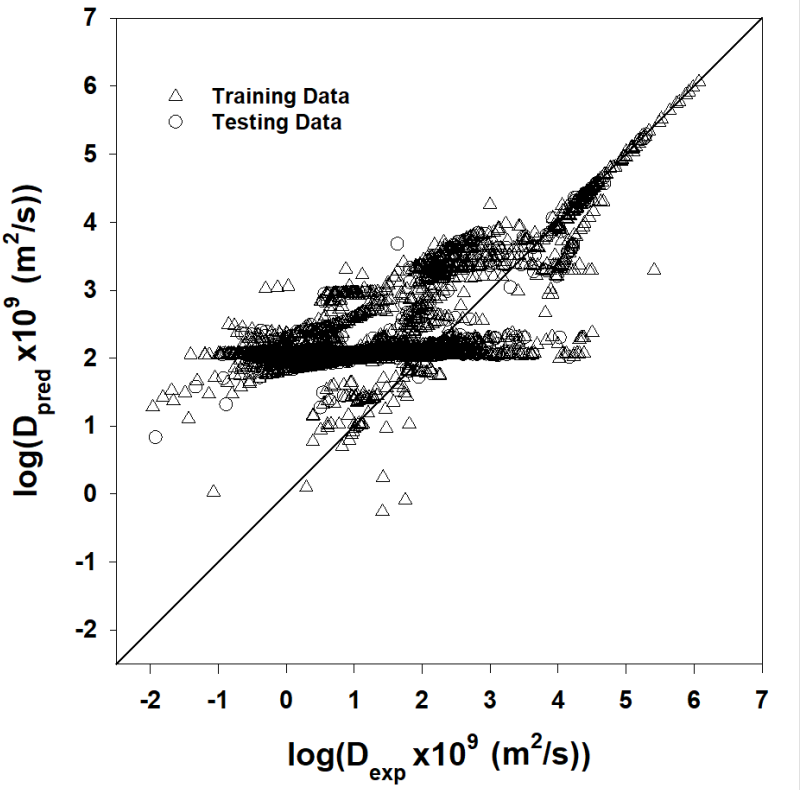

Figure S4. Correlation plot for B-ANN diffusion predictions with the log10 of diffusion taken. Compare to Figure 2 in the main text.

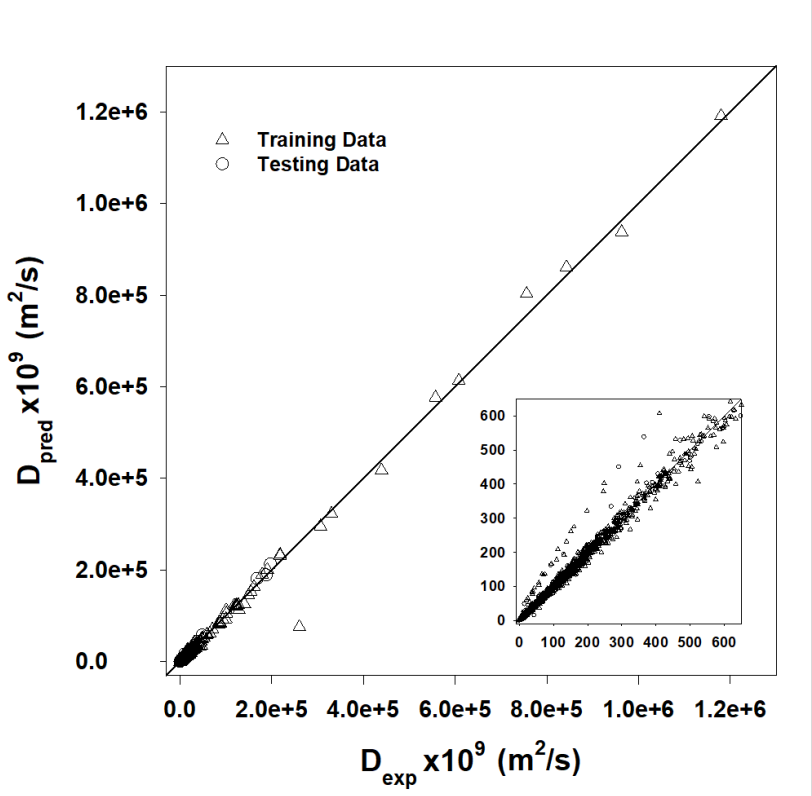

Figure S5. Correlation plot for the log-ANN diffusion predictions, with target values returned the raw scale. The inset has the same units as the main figure. Compare to Figure 1 in main text.

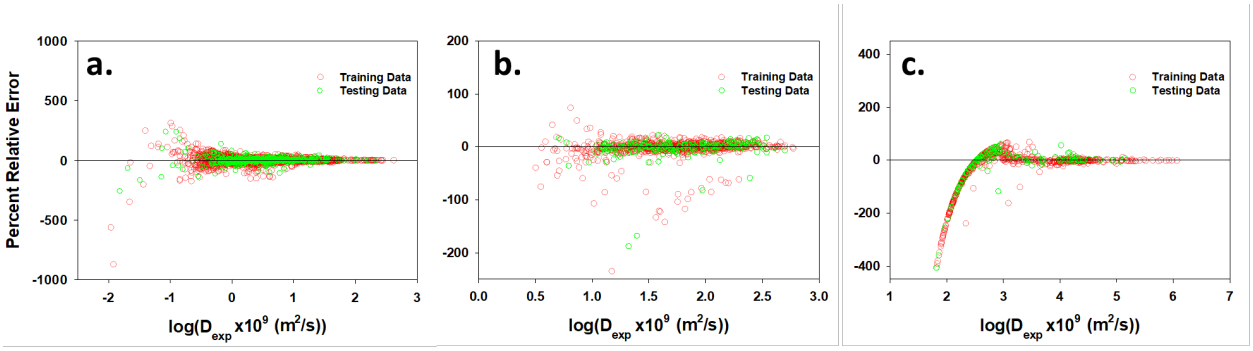

Figure S6. Relative error for self-diffusion predictions using the a) L-ANN, b) S-ANN, c) G-ANN.

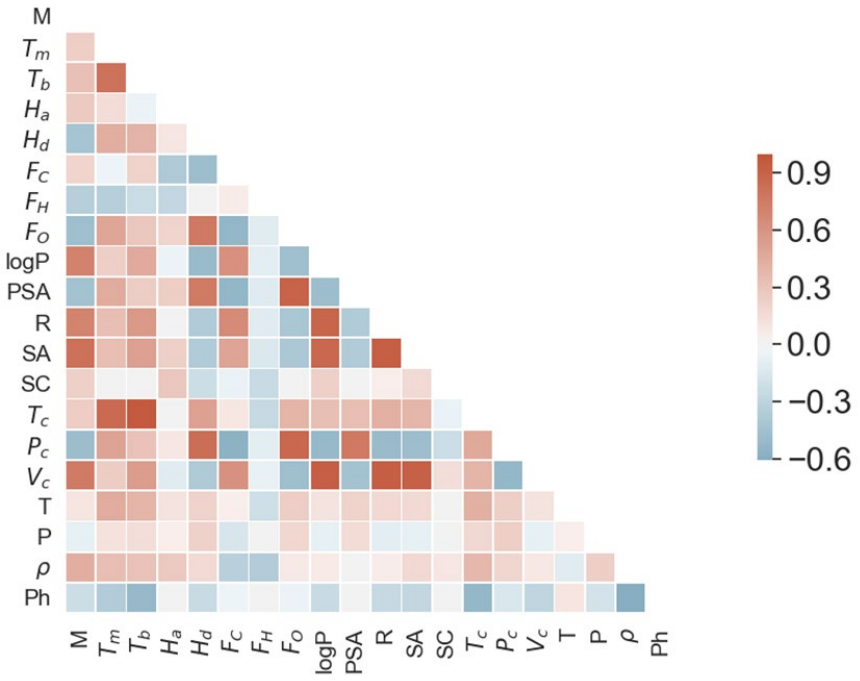

Figure S7. Correlation plot for input features used in DB1.

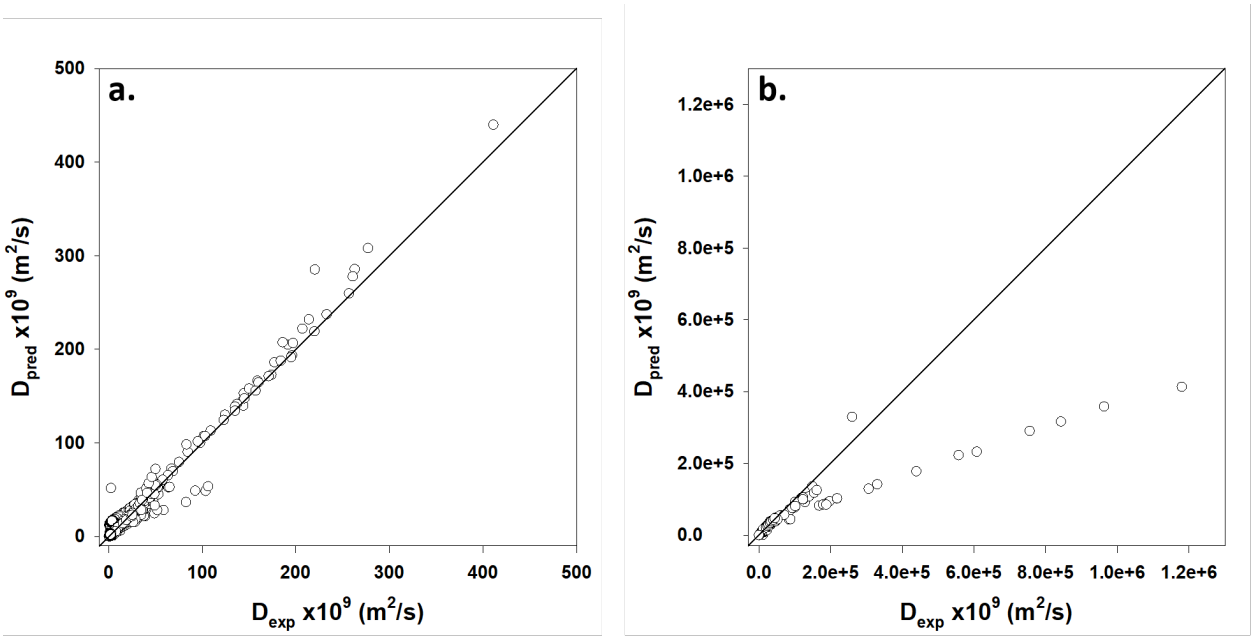

Figure S8. Correlation plot for two empirical models of self-diffusion: a) Liu *et al.* tested against only liquid values in DB1 and b) Zhu *et al.* tested against all data in DB1.

References

1. C. M. Silva and H. Liu, in *Lecture Notes in Physics* 2008, vol. 753, pp. 383-492.
2. O. Suárez-Iglesias, I. Medina, C. Pizarro and J. L. Bueno, *Chem. Eng. Sci.*, 2007, **62**, 6499-6515.
3. J. H. Dymond, *J. Chem. Phys.*, 1974, **60**, 969-973.
4. H. Lee and G. Thodos, *Ind. Eng. Chem. Res.*, 1988, **27**, 992-997.
5. P. H. Salim and M. A. Trebble, *J. Chem. Soc., Faraday Trans.*, 1995, **91**, 245-250.
6. R. J. Speedy, F. X. Prielmeier, T. Vardag, E. W. Lang and H. D. Lüdemann, *Mol. Phys.*, 1989, **66**, 577-590.
7. E. Ruckenstein and H. Liu, *Ind. Eng. Chem. Res.*, 1997, **36**, 3927-3936.
8. D. Ben-Amotz and D. R. Herschbach, *J. Phys. Chem. B*, 1990, **94**, 1038-1047.
9. H. Liu, C. M. Silva and E. A. Macedo, *Chem. Eng. Sci.*, 1998, **53**, 2403-2422.
10. C. M. Silva, H. Liu and E. A. Macedo, *Chem. Eng. Sci.*, 1998, **53**, 2423-2429.
11. Y. X. Yu and G. H. Gao, *Int. J. Thermophys.*, 2000, **21**, 57-70.
12. Y. Zhu, X. Lu, J. Zhou, Y. Wang and J. Shi, *Fluid Phase Equilib.*, 2002, **194-197**, 1141-1159.