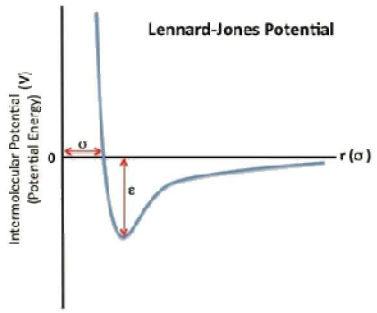
SAND2020-8122C

Using Machine Learning to Predict Self-Diffusion in Lennard Jones Fluids



ACS National Meeting
San Francisco, CA
August 17, 2020

Division/Committee: Chemical Information

**Due to the COVID-19 pandemic this presentation was pre-recorded and will be broadcast*

PRESENTED BY

Joshua P. Allers

Organic Materials Science Department
Sandia National Laboratories
Albuquerque, N

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

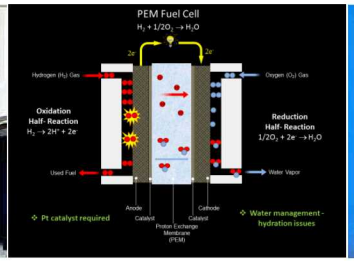
LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525

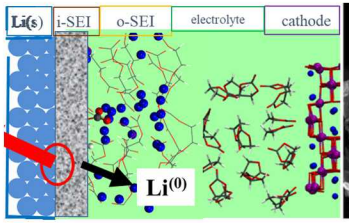
SAND2020-xxxx

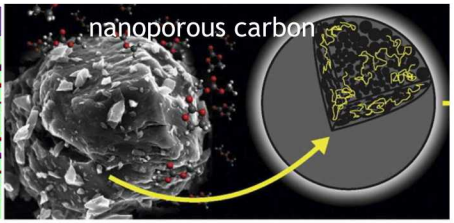
Diffusion of Mixtures Absorbed into Materials: Interest at Sandia National Laboratories

H_2O/MeOH Fuel Cells



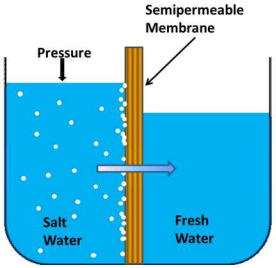
Batteries

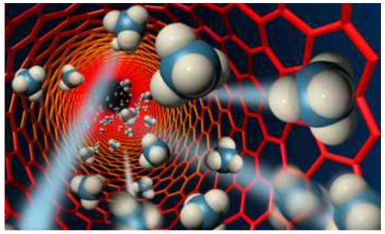




Separation Membranes

Robert Service, Science 2006





LLNL, Carbon Nanotubes

- Understanding the diffusion of chemical species in porous materials critical for the design and performance optimization for different materials.

Current Methods for Calculating Diffusion

Calculating diffusion experimentally

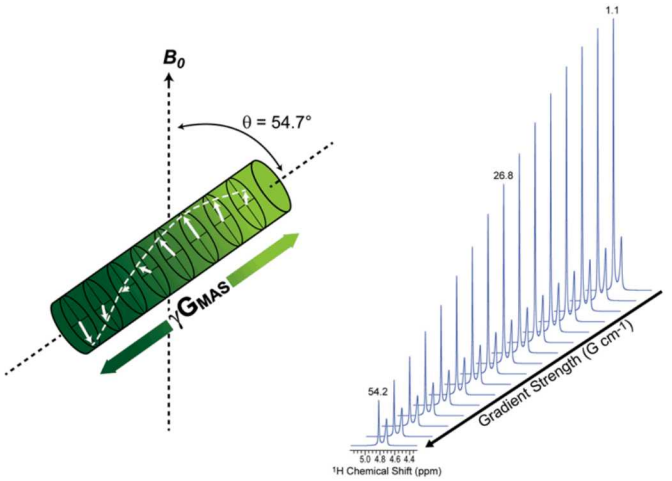
- NMR diffusometry, absorption, CT, etc.
- expensive and time-consuming

Calculating diffusion computationally

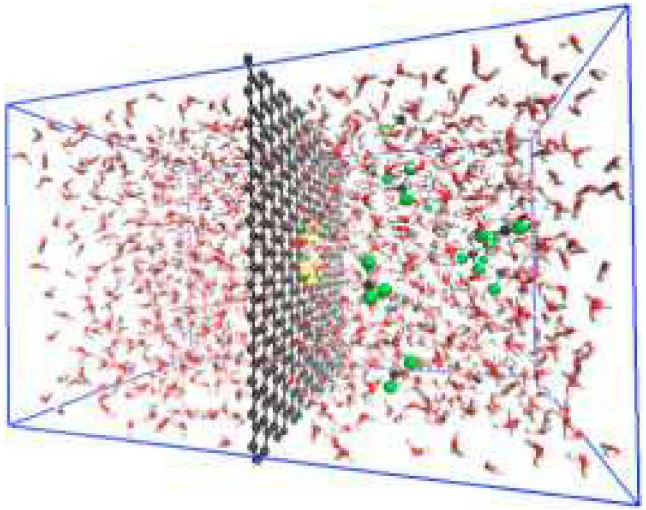
- Molecular dynamics and DFT
- Accurate, but time-consuming
- Require large amounts of computation

Consider studying 10 different compounds

- Forming all possible binary mixtures
- Testing 10 different compositions
- 450 experiments or simulations



Todd Alam and Janelle Jenkins. Advanced Aspects of Spectroscopy. **2012.**



Jafar Azamat *et. al.* Chemical Engineering Science. **2015.**

Challenge - Universal Model for Diffusion

Maxwell Stephen (MS) Model
(dominant diffusion model)

$$\frac{1}{D_{\mathrm{i,self,s}}} = \frac{1}{D_{i,s}} + \frac{x_i}{D_{ii}} + \frac{x_j}{D_{ij}}$$

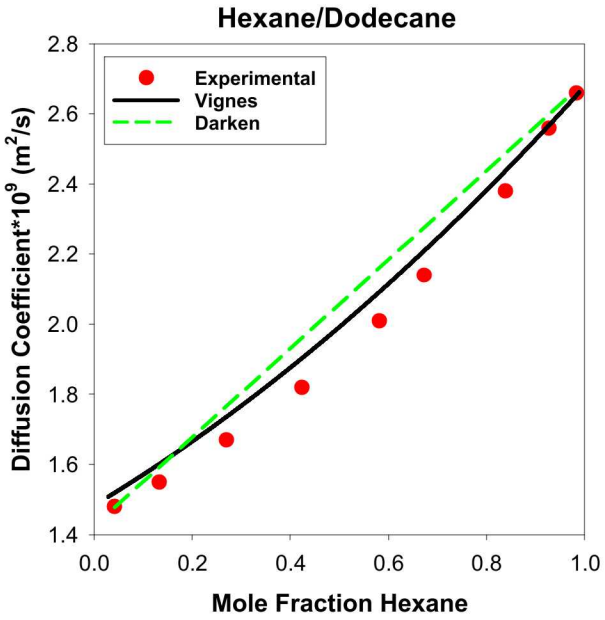
surface self exchange

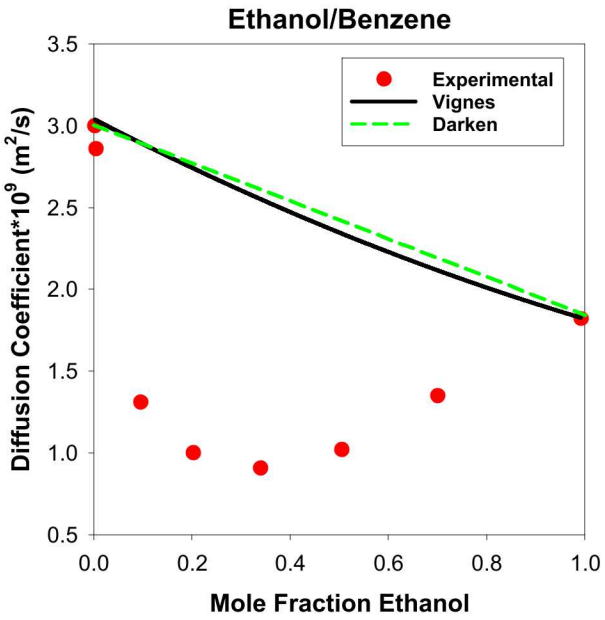
Darken Relationship
(semi-empirical - linear)

$$D_{ij} = D_i D_j \sum_{k=1}^{N} \frac{x_k}{D_k}$$

Vignes Model
(semi-empirical- power law)

$$D_{ij} = \left(D_{ij}^{x_j \to 1}\right)^{x_i} \left(D_{ij}^{x_j \to 1}\right)^{x_j} \prod_{k,k\neq i,j}^{i=N} \left(D_{ij}^{x_k \to 1}\right)^{x_k}$$



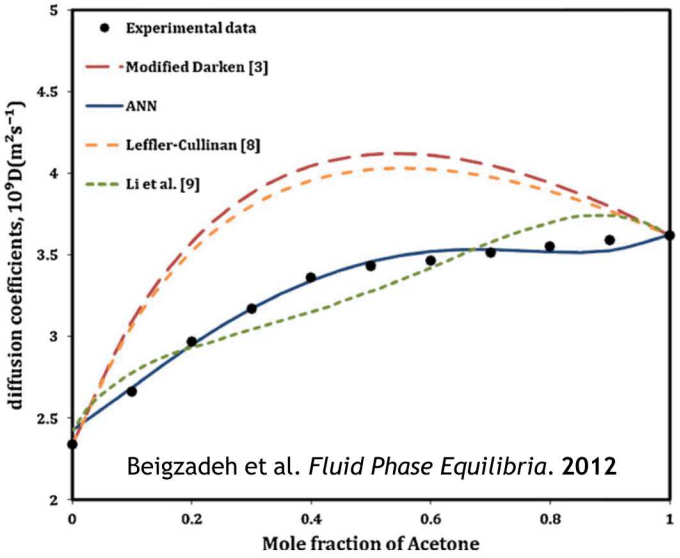


Goal: Develop a generalized model that can accurately predict diffusion in multi-component mixtures

Potential of Machine Learning

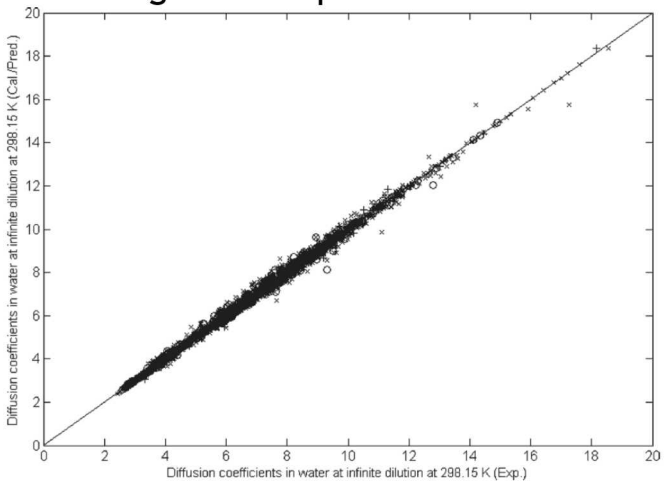
- Rapidly growing field in materials science
- Powerful tools for prediction
  - Random Forests
  - Artificial Neural Networks (ANN)
  - Symbolic Regression
  - Genetic Engineering
- Already showing promise in literature

acetone-chloroform mixture



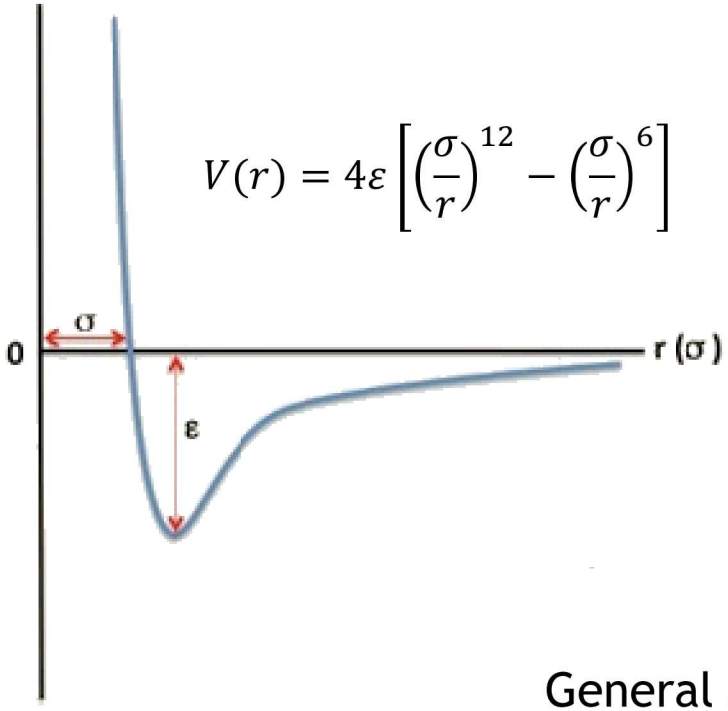
Beigzadeh et al. *Fluid Phase Equilibria*. **2012**

organic compounds in water



Gharagheizi et al. *Journal of Chemical Engineering Data*. **2011**

Start Simple – Lennard Jones (LJ)

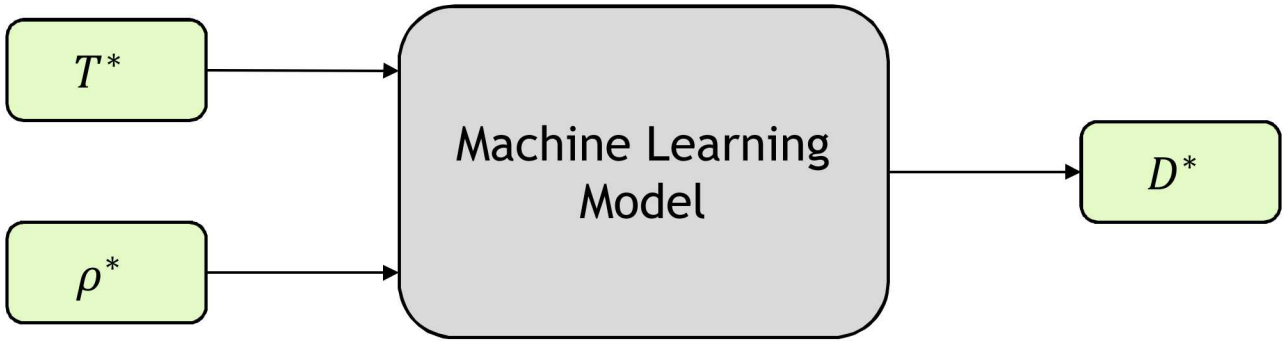


$$V(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

$$T^* = \frac{Tk}{\varepsilon} \qquad \rho^* = \frac{N\sigma^3}{V} \qquad D^* = D\frac{\sqrt{\frac{m}{\varepsilon}}}{\sigma}$$

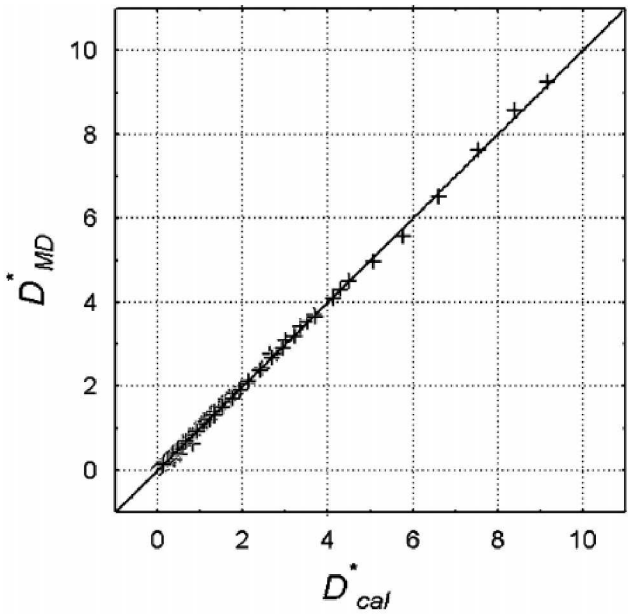
- Explained by two parameters
  - σ - Distance where potential becomes zero
  - ε - Well depth and measure of attractive force
- Allows model development over multiple phases

General Model Architecture



Lennard Jones Molecular Dynamics (MD)

- Many MD simulations have been performed for LJ systems
- Zhu's empirical equation is one of the better models
  - Has 8 adjustable parameters
- Will act as the "gold standard" for comparison



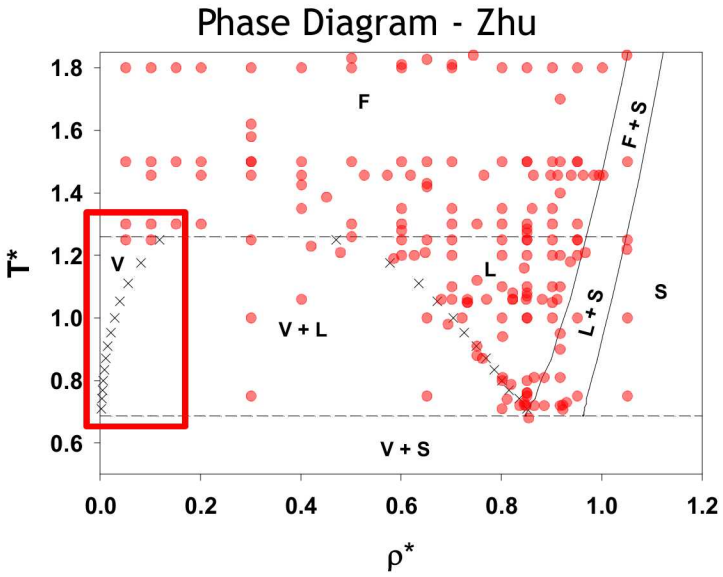
$$D^*_{cal} = \frac{3\sqrt{T^*}}{8\rho^*\sqrt{\pi}} A \times B$$

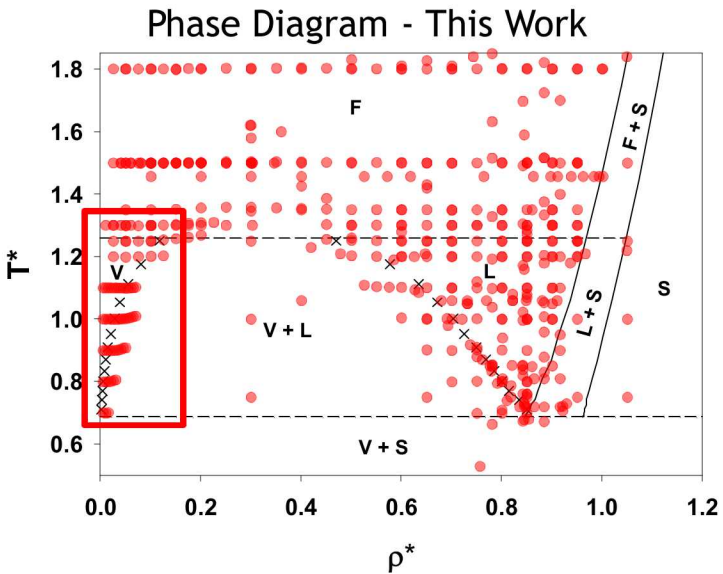
$$A = \left(1 - \frac{\rho^*}{a(T^*)^b}\right)\left[1 + (\rho^*)^c\left(\frac{P1(\rho^* - 1)}{P2(\rho^* - 1) + (T^*)^{(P3+P4\rho^*)}} + P5\right)\right]$$

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

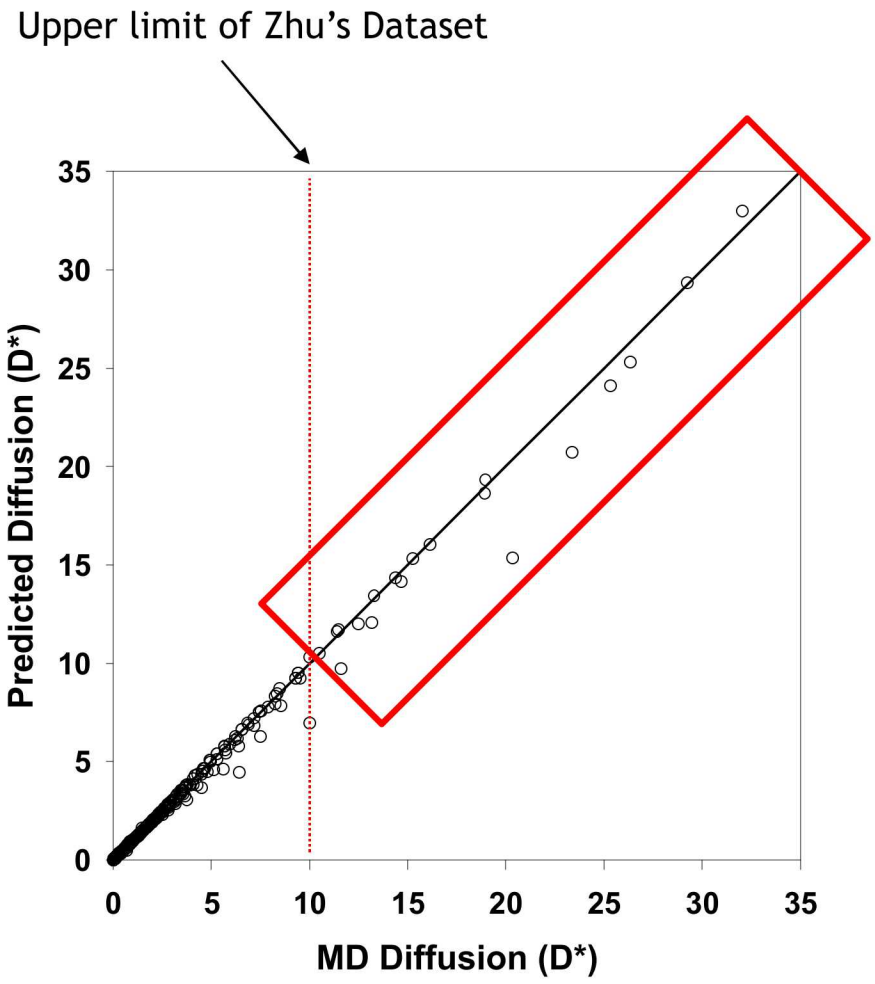
Yu Zhu, Xiaohua Lu, Jian Zhou, Yanru Wang, Jun Shi, "Prediction of diffusion coefficients for gas, liquid and supercritical fluid: application to pure real fluids and infinite binary solutions based on the simulation of Lennard-Jones fluid". *Fluid Phase Equilibria*. **2002**. https://doi.org/10.1016/S0378-3812(01)00669-0

Expanded Diffusion Dataset



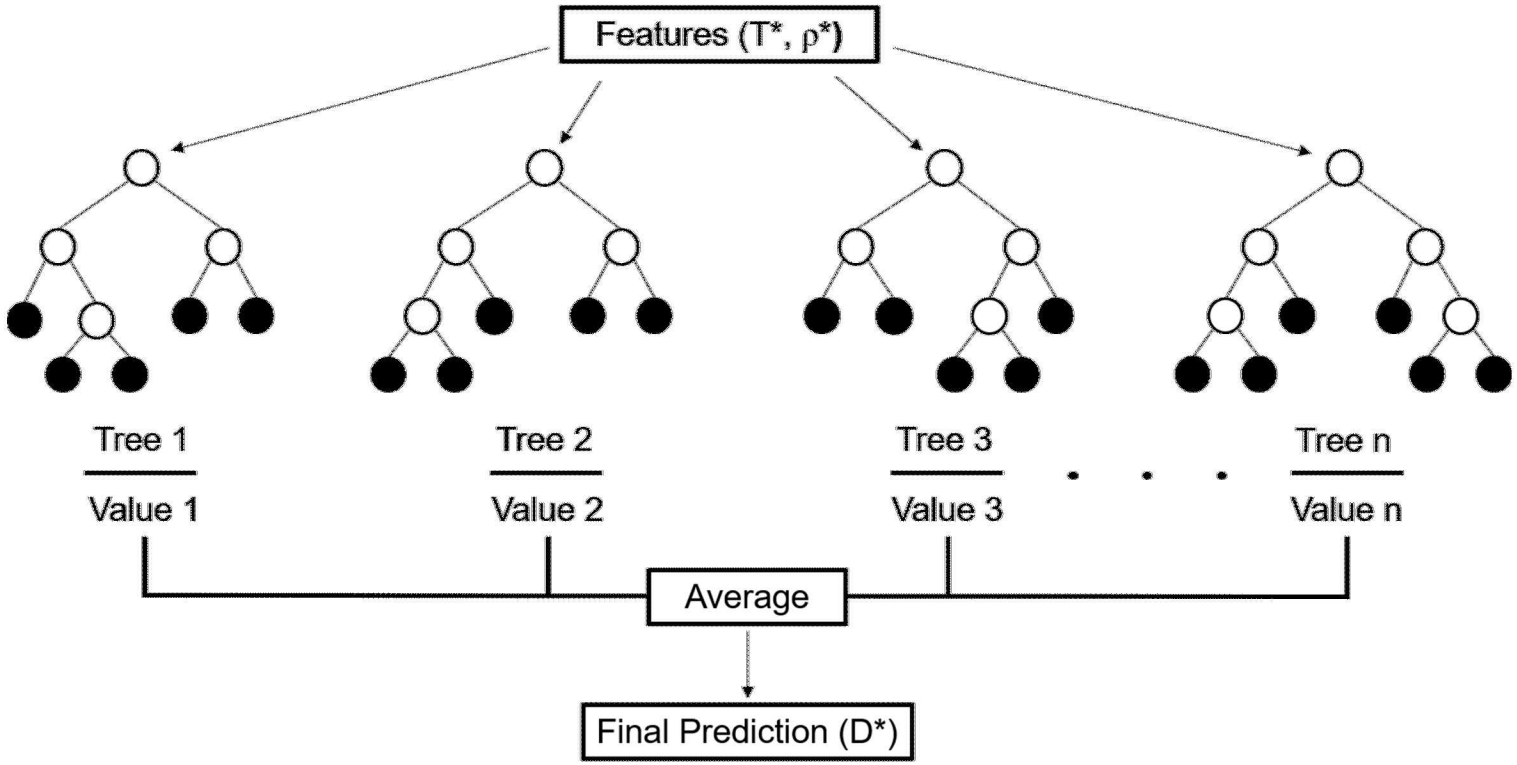


V - vapor
L - liquid
F - supercritical
S - solid



Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, Todd M.
"Machine learning prediction of self-diffusion in Lennard Jones fl
Chem. Phys. 153, 034102 (2020).

Random Forest Architecture

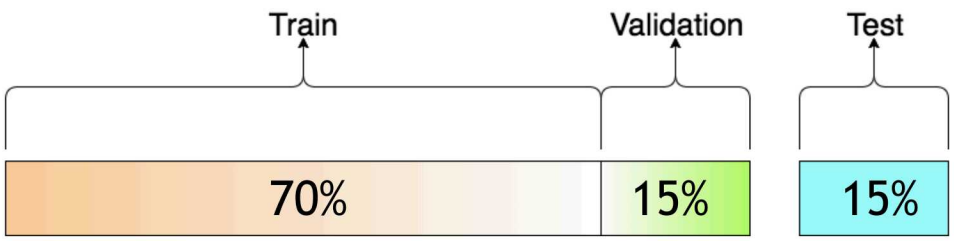


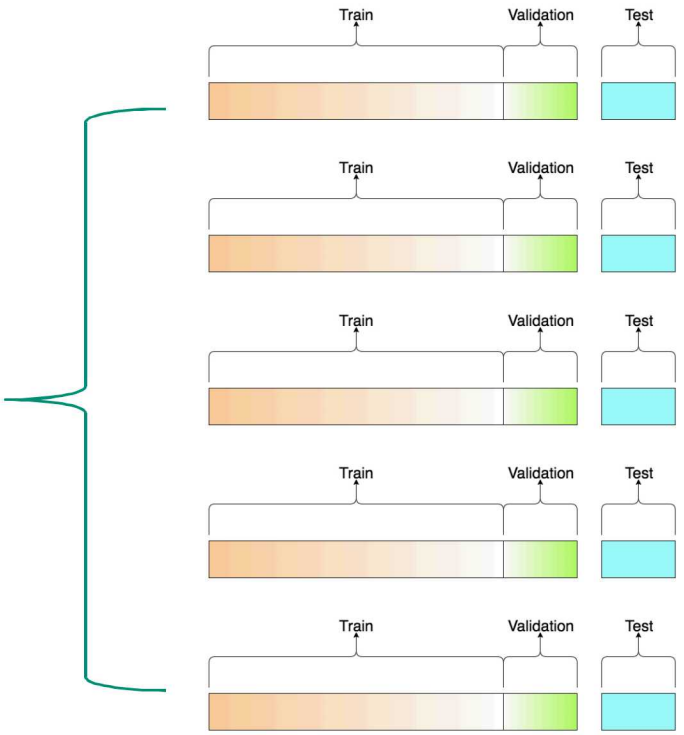
- Software: Python Scikit-Learn
- Optimal number of trees: 235

Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, Todd M. Alam. "Machine learning prediction of self-diffusion in Lennard Jones fluids." *J Chem. Phys.* 153, 034102 (2020).

Performance Assessed with Cross Validation (CV)

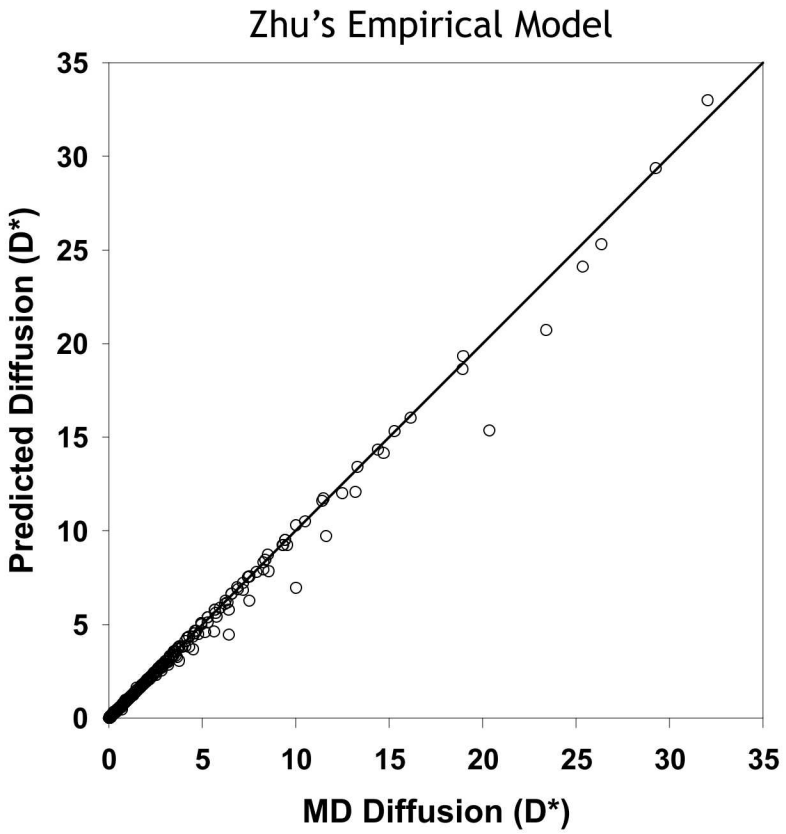
- Splits dataset in even train/validate/test sets
  - ML models train on the 70%
  - Validation set used to adjust hyperparameters (model parameters)
  - Test set gave final predictive power
- 5 different splits of the data
  - Randomly divided
- Performance metrics:
  - Mean Square Error (MSE)
  - Correlation Coefficient (R)

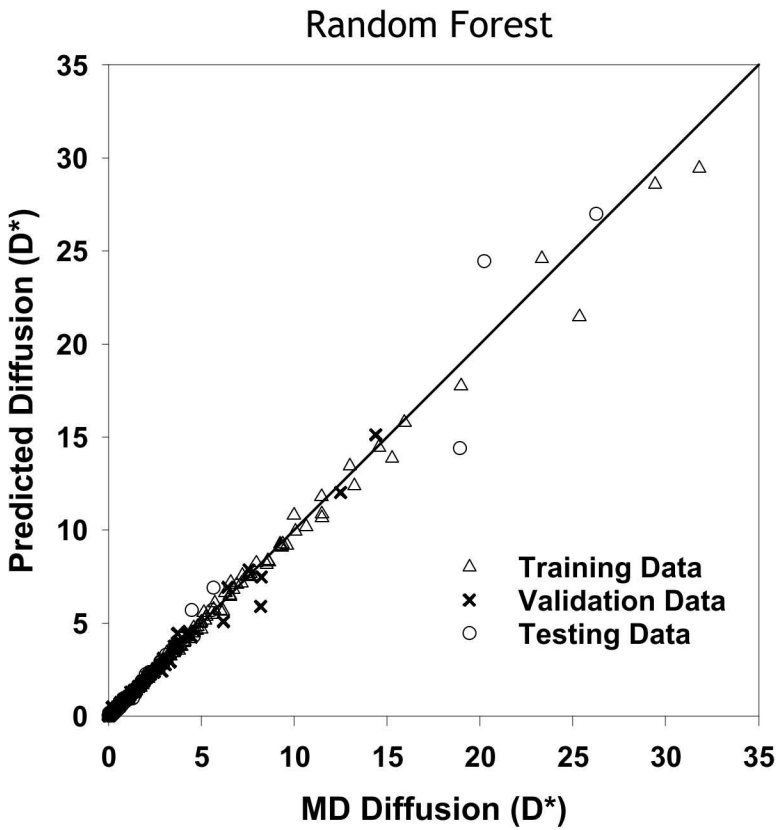




towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Performance of Random Forest

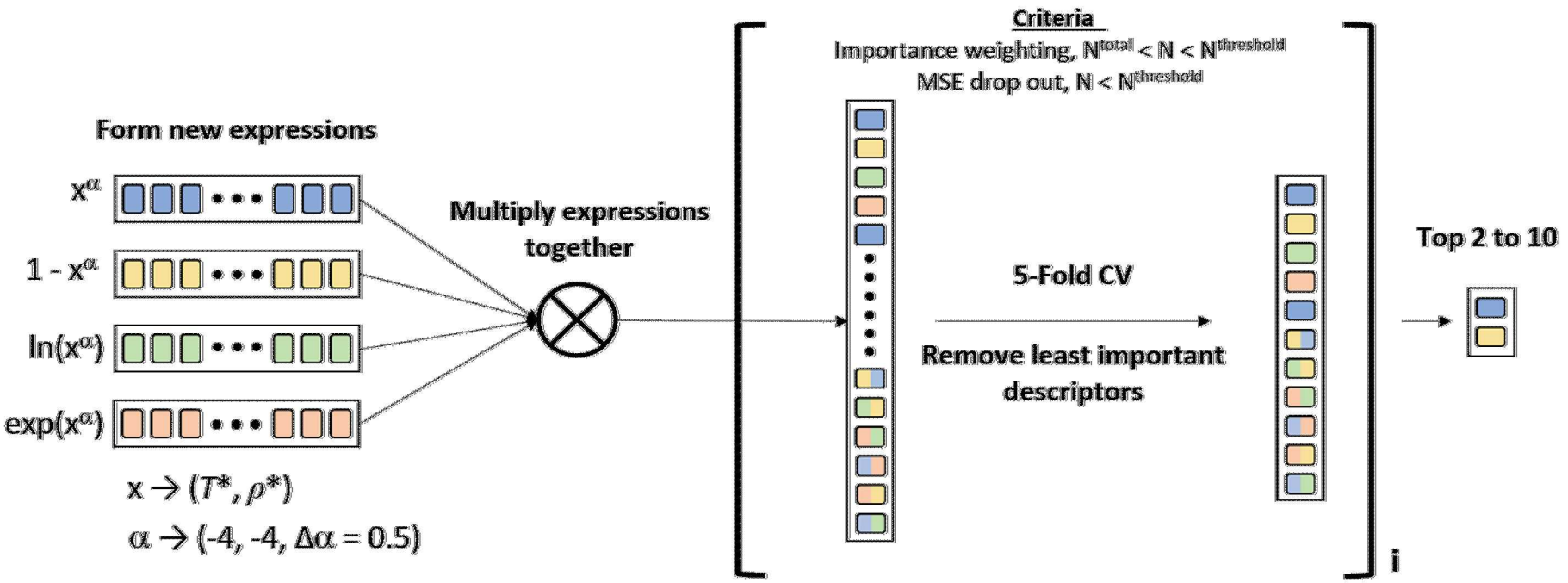




Model	Zhu	RF
Mean MSE	0.067	0.18
Std. Dev.	± 0.077	± 0.21
Mean R	0.99685	0.99624

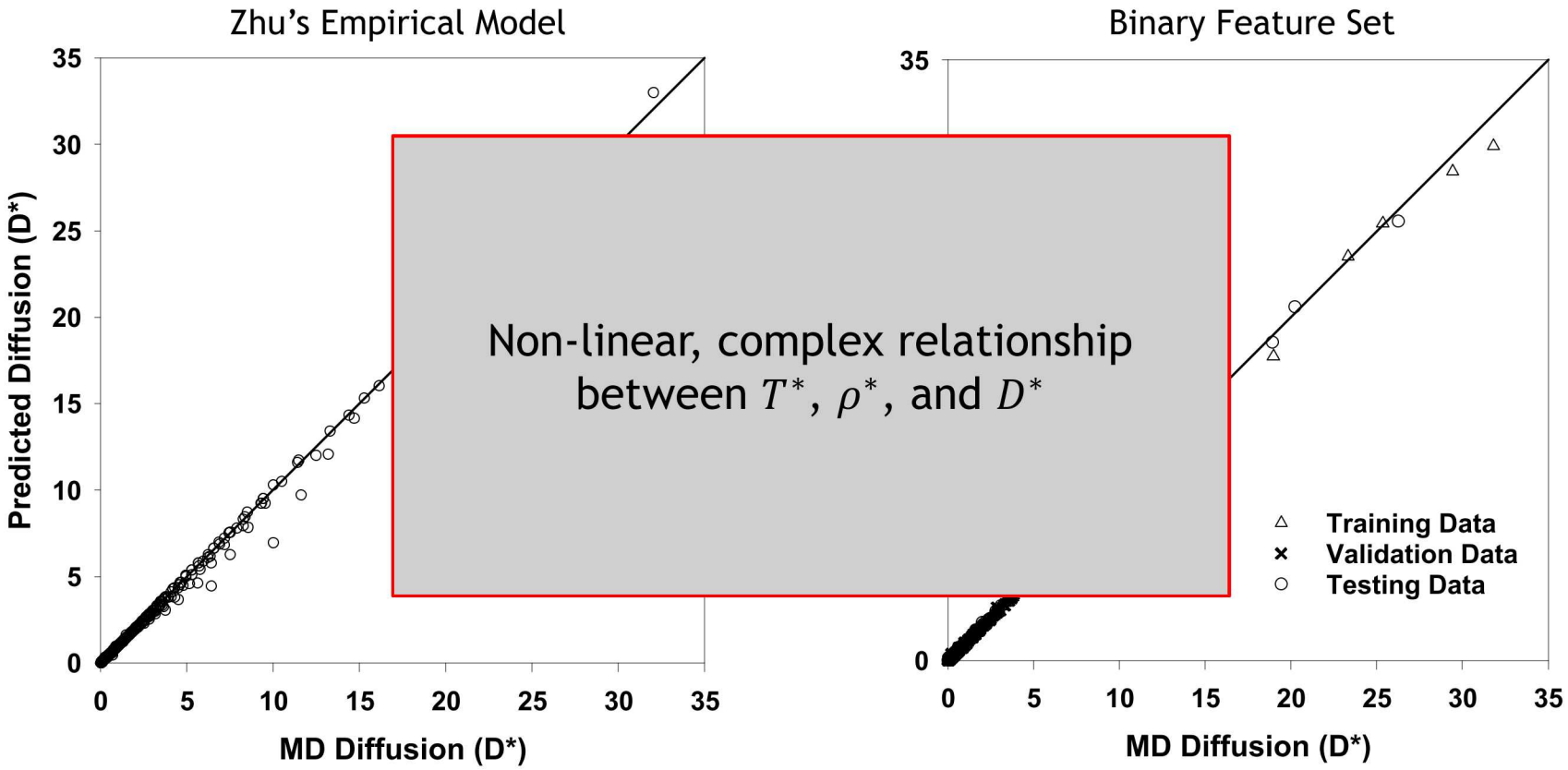
Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, Todd M. Alam. "Machine learning prediction of self-diffusion in Lennard Jones fluids." *J Chem. Phys.* 153, 034102 (2020).

Feature Engineering Process



Solo	Binary	Self-Binary
$f(T^*)$ **or** $f(\rho^*)$	$f(T^*) * g(\rho^*)$	$f(T^*) * g(T^*)$
128 Features	4224 Features	7296 Features
e.g. $\ln(T^*)$	e.g. $\ln(T^*) * \exp(\rho^*)$	e.g. $\ln(T^*) * \exp(T^*)$

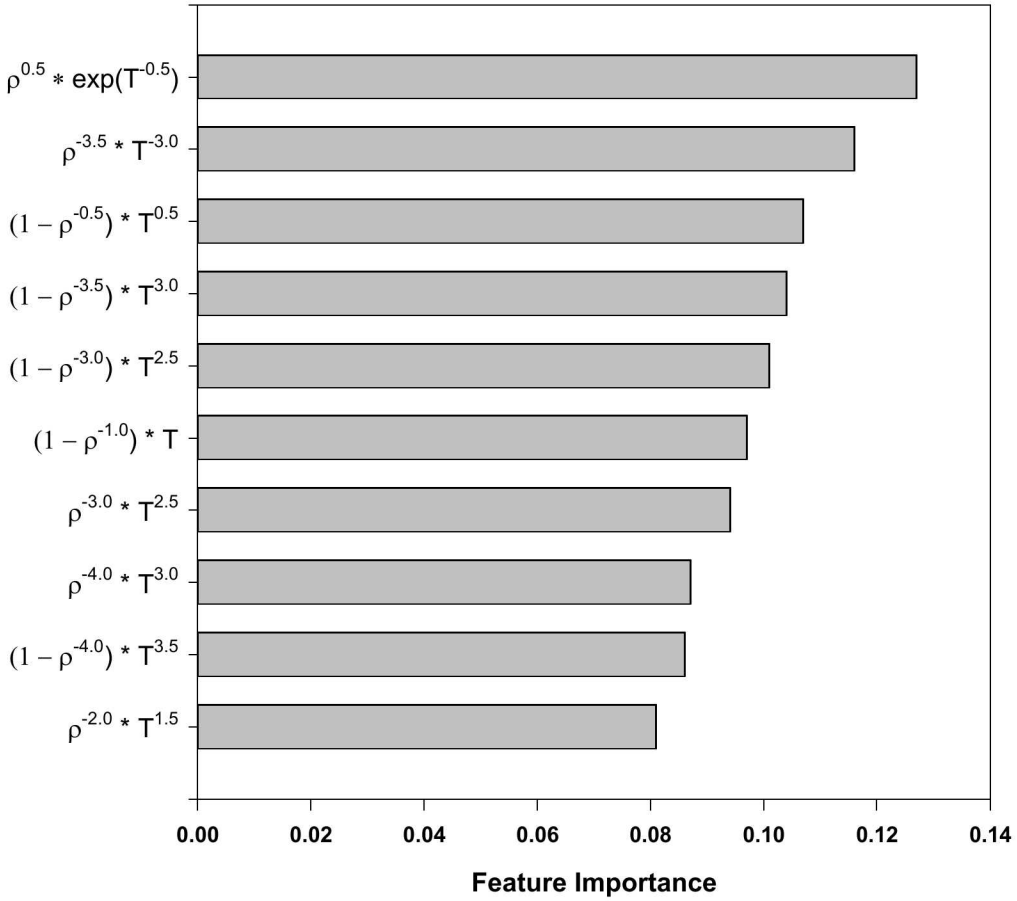
Random Forest with Feature Engineering



Model	Zhu	RF	Solo	Binary	Self-Binary
Mean MSE	0.067	0.18	0.15	0.052	0.047
Std. Dev.	± 0.077	± 0.21	± 0.18	± 0.047	± 0.072
Mean R	0.99685	0.99624	0.99642	0.99913	0.99912

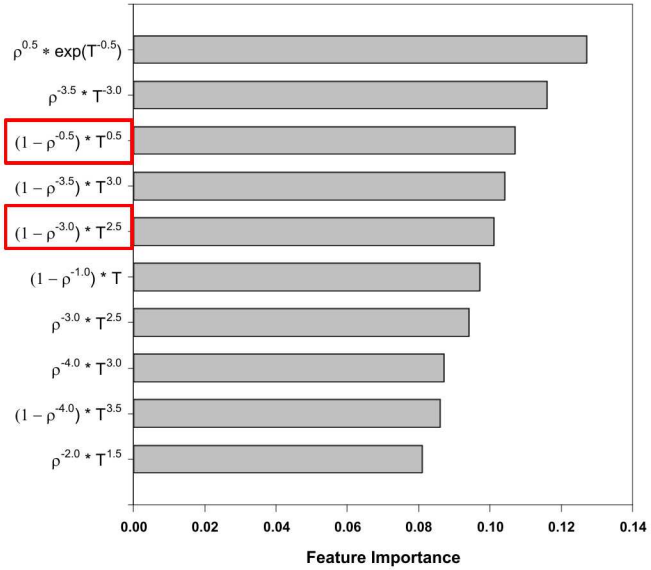
Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, Todd M. Alam. "Machine learning prediction of self-diffusion in Lennard Jones fluids." *J Chem. Phys.* 153, 034102 (2020).

Most Important Features – Binary Feature Set



Observed trend: non-linear functions of T^* and ρ^* consistently show up as the most important features

Most Important Features – Binary Feature Set



Continued iteration →

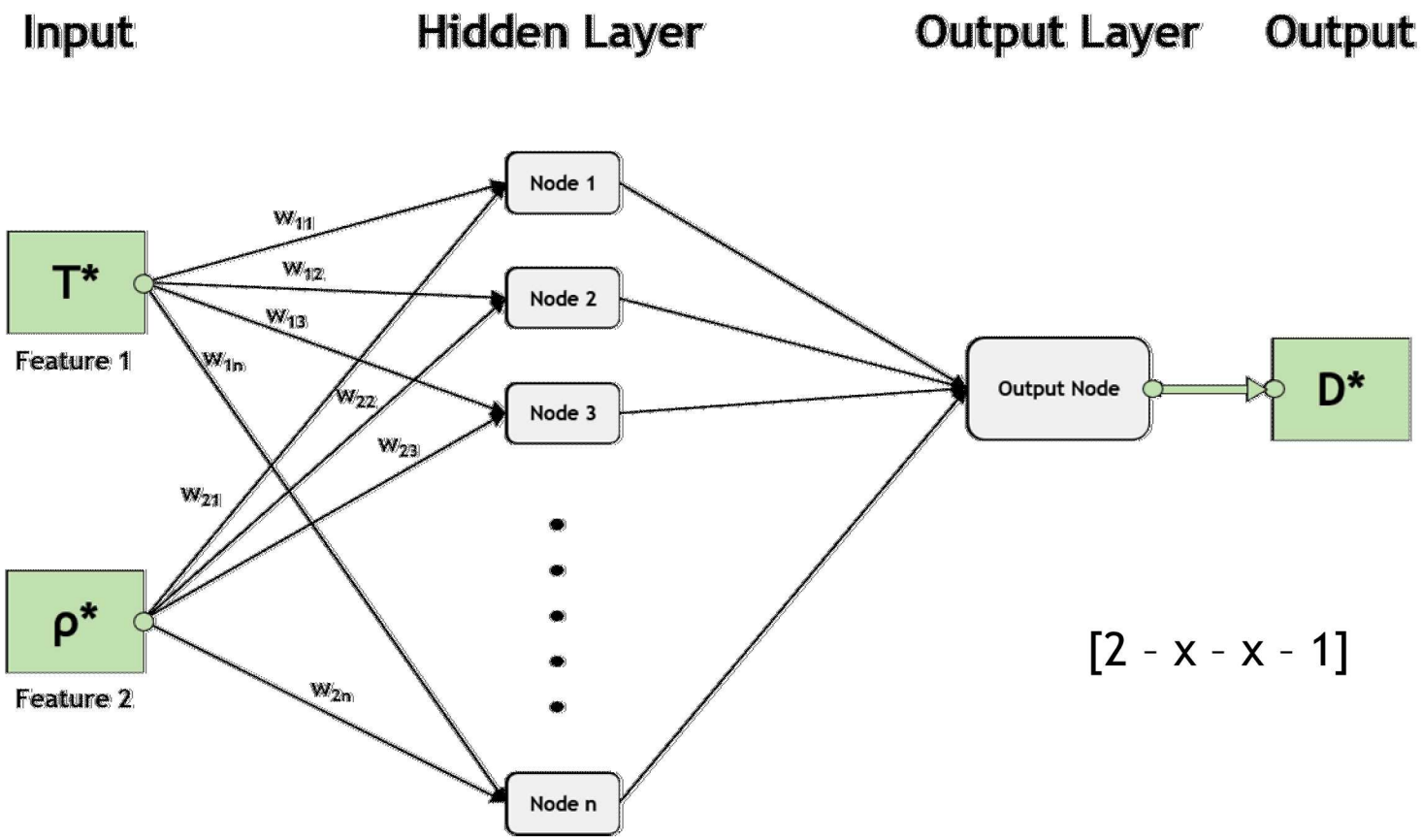
Top 2 Features

$$(1 - \rho^{-0.5}) * T^{0.5}$$

$$(1 - \rho^{-3.0}) * T^{2.5}$$

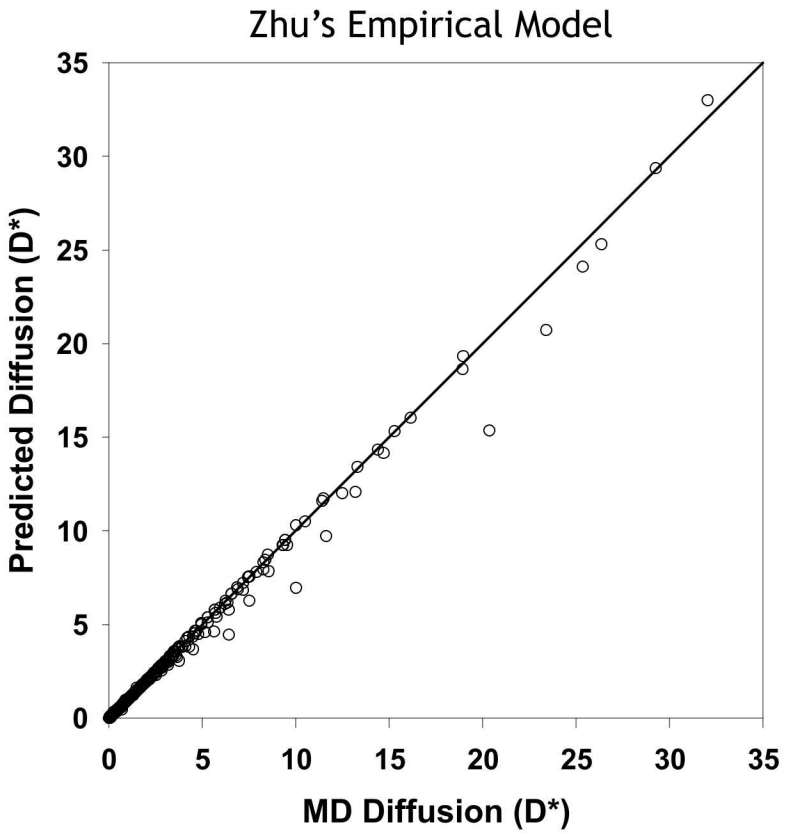
3rd and 5th ranked features become the top 2

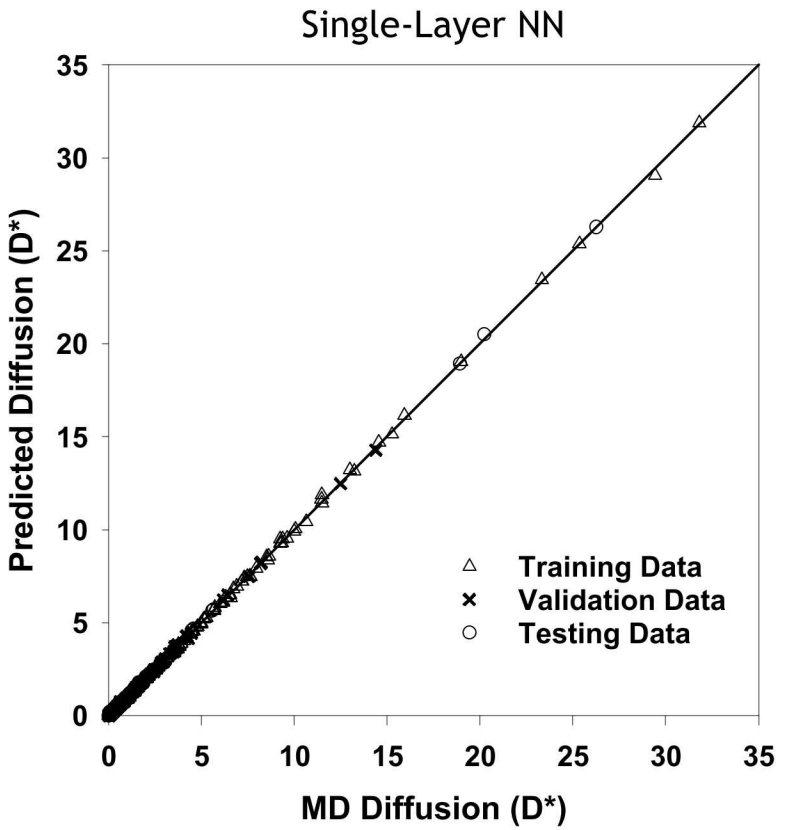
Artificial Neural Network (ANN) Architecture



- Software: MATLAB Deep Learning Toolbox

Neural Network Performance





Model	ANN(2-14-12-1)	ANN(2-14-1)	RF-FE	Zhu	RF
Mean MSE	0.0012	0.0016	0.052	0.067	0.18
Std. Dev.	± 0.0009	± 0.001	0.047	± 0.077	± 0.21
Mean R	0.99994	0.99991	0.99913	0.99685	0.99624

Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, Todd M. Alam. "Machine learning prediction of self-diffusion in Lennard Jones fluids." *J Chem. Phys.* 153, 034102 (2020).

Conclusions

- Machine learning improved upon existing empirical relationships in LJ system
- Random Forest with T^* and ρ^* performed worse than Zhu's equation
- Employed feature engineering, which improved predictions
  - On par with Zhu's equation
  - Pointed to non-linear, complex relationships
- Artificial Neural Networks provide the best predictions
  - Will be used going forward with real systems

Ongoing Work

- Machine Learning extended to sets of binary LJ fluids
- Artificial Neural Networks assessed on a dataset of pure solutions
  - Shows excellent performance over multiple compounds and phases

Acknowledgements

Todd Alam (SNL Advisor)

Fernando Garzon (UNM Advisor)

Jacob Harvey (SNL)

Lok-Kun Tsui (SNL, UNM)

Brennan Walder (SNL)

Jeff Greathouse (SNL)

Chad Priest (SNL)

Calen Leverant (SNL)

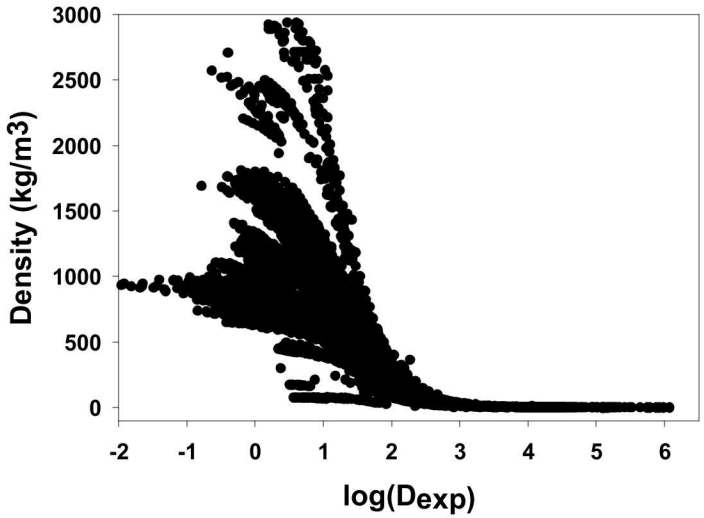
Thank you for you attention - Questions?

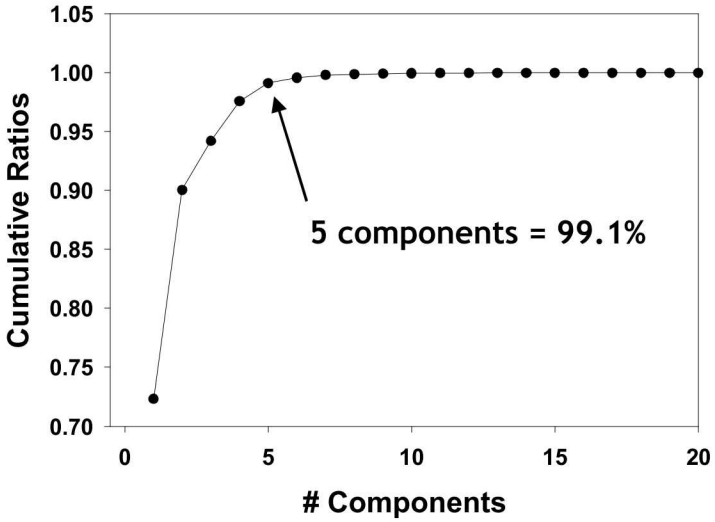
Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This presentation describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Project funding through the Sandia LDRD program

Additional Slides

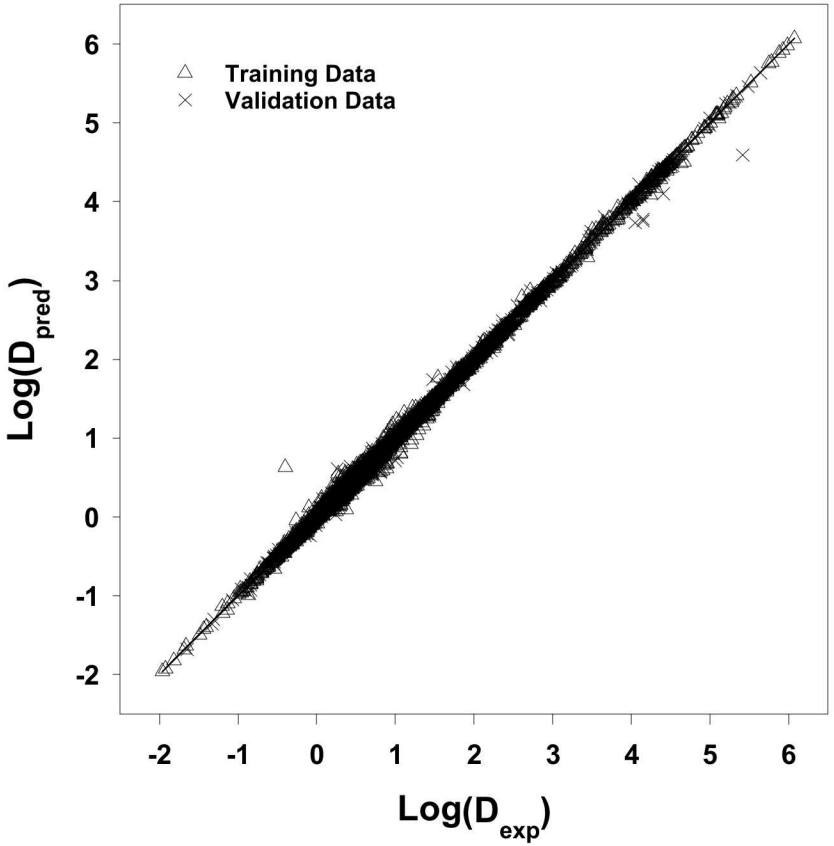
Current Work – Pure Solutions Diffusion





- 6252 points from a wide range of compounds
  - Multiple phases present
- 24 features were collected including:
  - Critical properties
  - Experimental properties
  - Phase information
  - Structural information
- Principal Component Analysis (PCA) used to reduce the dimensionality

Current Work – Pure Solution Diffusion



- Models are robust and generalized over a 5-fold cross-validation
- Shows good predictions over multiple phases

Joshua P. Allers, Fernando H. Garzon, Todd M. Alam. *In Progress*.