

IMPERIAL COLLEGE LONDON

DOCTORAL THESIS

---

# Thermodynamically-guided machine learning for thermophysical property modeling and prediction

---

*Author:*

Gustavo Alexander CHAPARRO  
MALDONADO

*Supervisor:*

Prof. Erich A. MÜLLER

*A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy*

Molecular Systems Engineering Group  
Department of Chemical Engineering

April, 2025

To Taylor Swift, who kept me sane (and  
insane) throughout this journey

# Abstract

The modeling of thermophysical properties is central to chemical process design and optimization. These include thermodynamics (e.g., phase equilibrium, heat capacities) and transport (e.g., diffusivity, viscosity, thermal conductivity) properties. Thermophysical properties have been modeled using empirical, semi-empirical, and theoretical approaches. Despite their success, current models rely on incomplete theories fitted to limited data. To address this issue, this thesis develops a hybrid approach combining molecular simulations, theory, and artificial neural networks (ANNs) for modeling thermophysical properties. This approach is demonstrated for systems interacting via the Mie potential. Thermodynamic properties can be determined from the derivatives of the Helmholtz Free Energy surface, which is modeled using a physics-embedded ANN Equation of State (FE-ANN EoS). This EoS is formulated to satisfy known thermodynamic constraints. The FE-ANN EoS is comparable to the monomer term of the theoretically derived SAFT-VR-Mie EoS. Moreover, the FE-ANN EoS provides a flexible framework for finding the appropriate functional form of the free energy, including contributions where an analytical solution is unavailable. The FE-ANN(s) EoS extends this framework to handle both fluid and solid phases. The resulting EoS predicts the entire phase diagram and provides insights into the continuity between different aggregation states. A similar physics-informed data-driven approach is used to model transport properties. In this case, ANN models are constrained by the dilute gas limit predicted by kinetic theory. These models accurately predict transport properties across a broad phase space region. The models developed in this thesis comprehensively characterize Mie particles. These models enable testing parametrization schemes for quasi-spherical molecules. The Mie potential exhibits a trade-off between fitting thermodynamic and transport properties. Incorporating transport properties or solid-phase data yields similar molecular parameters, suggesting that force field parametrization should account for this information. This thesis demonstrates how scientific machine learning enhances thermophysical property modeling, improving accuracy and addressing the limitations of conventional methods.

# Declaration of Authorship

I, Gustavo Alexander CHAPARRO MALDONADO, declare that this thesis titled, “Thermodynamically-guided machine learning for thermophysical property modeling and prediction” and the work presented in it are my own.

I have explicitly acknowledged and appropriately referenced the work of others where applicable.

Gustavo Alexander CHAPARRO MALDONADO

April, 2025



# Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Publications

Part of this thesis has been published in the following peer-reviewed scientific articles.

1. Gustavo Chaparro and Erich A. Müller. “Development of a Helmholtz free energy equation of state for fluid and solid phases via artificial neural networks”. *Communications Physics* 7.1 (2024), p. 406. DOI: 10.1038/s42005-024-01892-3, Ref. [1]
2. Gustavo Chaparro and Erich A. Müller. “Simulation and Data-Driven Modeling of the Transport Properties of the Mie Fluid”. *The Journal of Physical Chemistry B* 128.2 (2024), pp. 551–566. DOI: 10.1021/acs.jpcc.3c06813, Ref. [2]
3. Gustavo Chaparro and Erich A. Müller. “Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid”. *The Journal of Chemical Physics* 158.18 (2023), p. 184505. DOI: 10.1063/5.0146634, Ref. [3]

The outcomes of this thesis have also been communicated in the following meetings.

1. *Multiphase parametrization of quasi-spherical Mie particles*. Oral Presentation. Iberoamerican Conference Equifase, Évora, Portugal, 2024
2. *A continuous machine-learned molecular Equation of State to describe the vapor, liquid, and solid states*. Oral Presentation. Thermodynamics Conference, Delft, Netherlands, 2024
3. *On the continuity of the equation of state of fluid and solid phases*. Oral Presentation. Symposium on Thermophysical Properties, Boulder, United States, 2024
4. *Development of thermodynamically-consistent machine-learning Equations of State: Application to the Mie fluid*. Oral Presentation. AIChE Annual Meeting, Orlando, United States, 2023
5. *Thermodynamically consistent machine-learning-based Equation of State for the Mie fluid*. Poster. PPEPPD Conference, Tarragona, Spain, 2022
6. *Square peg in a round hole: multi-property parametrization of Mie Fluids*. Poster. Iberoamerican Conference Equifase, Campinas, Brazil, 2022
7. *Fitting a square peg in a round hole: multi-property parametrization of Mie Fluids*. Short Presentation, Thermodynamics, Bath, United Kingdom, 2022

# Data Statement

Details to replicate the results of this thesis are found in the following GitHub repository: <https://github.com/gustavochm/Chaparro-PhD-Thesis>. This repository includes the following:

- Mie particle thermophysical properties database generated by molecular simulation
- Mie fluid transport properties database
- LAMMPS scripts to produce the calculation of thermophysical properties
- Trained machine-learning models (FE-ANN EoS, FE-ANN(s) EoS, and ANN-based transport property models)

# Acknowledgements

My PhD journey has been a very special time of my life. I have grown as a person and as a researcher. This journey would not have been the same without so many people who made my time in London so joyful. To all of you, my deepest thanks!

I want to start by thanking Prof. Erich A. Müller for his unconditional support during my PhD. I really value how Erich encouraged me to explore and learn about different topics, even if they were unrelated to my thesis. I also appreciate how you cared about my academic progress and personal growth beyond my PhD. I also want to acknowledge the MSE group. Amparo, Claire, and George for always maintaining a healthy environment within the group. I always felt I could count on them even though they did not directly supervise me. The MSE group is a special one. I want to thank Andrew for being the cohesive force in the group. I enjoyed all the times we explored London (and occasionally the UK) during our group walks. I have really liked being part of this group. I have too many friends to thank here. Tanuj, for our riverside walks and for being such a good friend, Thomas, for always having the most interesting stories, Ben, for setting high standards for the group, Shubhani, for joining me in crazy swifty adventures, Fabian, for all the interesting conversations about the future, Florian, for the nice brewing sessions, Ricardo, for always keeping a nice vibe, and Harry, for all the help when I moved to the UK.

Imperial has been a place to meet amazing people. Here, I have met people that I care about and admire. Thank you, Mariana, for being my office buddy. I have always enjoyed how easily we can talk about so many topics, from politics, research, and life to pure gossip. I have always trusted your honesty whenever I needed advice. Ilya, thank you for making my life in London so fun, for all the honest conversations, and for always making me feel better about myself. Miriam, thank you for being such a lovely person, for being my reference in terms of energy for social activities, and for always caring for others. You made the Sargent Centre a better place for many people, including me. Nikki, for always being up for hanging out, for being my concert buddy, and for sharing my addiction to sweets with me. Luxi, for being someone with whom I can share my swifty craziness and for always being willing to be interrupted whenever I want to chit-chat at the office. This list could get very long, but I also want to thank Lily, Fred, Alex, Kostis, Matthias, Haiting, Thanasis, Niki, Gabi, Marc, and Reewa for making my time here better.

My time in London has also been marked by trying new hobbies, with gymnastics being my favorite. I want to thank my coaches, Alex and Kasper, for teaching me so much, and so many friends—Leo, Theo, Alex, Hanafi, Nick, Sarah, and Chiara—for making the training sessions so fun and something that I will be looking forward to every week.

I also want to thank many friends from Chile who have always had my back. I want to thank Andrés and Marce for being the initial driving force behind my academic path. You two have built an incredible family at the Laboratorio de Cohesión. I know you genuinely care for me, and I definitely would not be here without your support. I also want to thank Javi and Vale for being one of my favorite people and always being there for me when needed. Finally, I want to thank my parents, Boris and Patricia, and my brothers, Boris and Andrés, for always supporting me in chasing my dreams.

Many thanks!

# Contents

|                                                                                          |           |
|------------------------------------------------------------------------------------------|-----------|
| <b>Abstract</b>                                                                          | <b>2</b>  |
| <b>Declaration of Authorship</b>                                                         | <b>3</b>  |
| <b>Copyright Declaration</b>                                                             | <b>4</b>  |
| <b>Publications</b>                                                                      | <b>5</b>  |
| <b>Data Statement</b>                                                                    | <b>6</b>  |
| <b>Acknowledgements</b>                                                                  | <b>7</b>  |
| <b>List of Figures</b>                                                                   | <b>12</b> |
| <b>List of Tables</b>                                                                    | <b>14</b> |
| <b>1 Introduction</b>                                                                    | <b>15</b> |
| 1.1 Data-driven thermodynamics . . . . .                                                 | 17        |
| 1.2 Thesis overview . . . . .                                                            | 22        |
| <b>2 Thermodynamically-consistent machine-learned equation of state</b>                  | <b>24</b> |
| 2.1 Introduction . . . . .                                                               | 24        |
| 2.2 Helmholtz free energy thermodynamics . . . . .                                       | 27        |
| 2.3 Free Energy Artificial Neural Network based Equation of State (FE-ANN EoS) . . . . . | 30        |
| 2.3.1 FE-ANN EoS formulation . . . . .                                                   | 30        |
| 2.3.2 FE-ANN EoS training . . . . .                                                      | 33        |
| 2.3.3 FE-ANN EoS hyperparameters . . . . .                                               | 35        |
| 2.4 Mie fluid thermophysical properties database . . . . .                               | 36        |
| 2.4.1 Thermophysical properties computation . . . . .                                    | 37        |
| 2.4.2 Thermophysical properties database validation . . . . .                            | 39        |
| 2.4.3 Thermophysical properties database description . . . . .                           | 41        |
| 2.5 FE-ANN EoS for the Mie fluid . . . . .                                               | 42        |
| 2.5.1 FE-ANN EoS: implementation and architecture . . . . .                              | 42        |
| 2.5.2 FE-ANN EoS: training and performance . . . . .                                     | 43        |
| 2.5.3 FE-ANN EoS: thermophysical properties modeling and prediction . . . . .            | 47        |
| 2.6 Conclusions . . . . .                                                                | 56        |

|          |                                                                                                                                        |            |
|----------|----------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>3</b> | <b>On the continuous modeling of fluid and solid states</b>                                                                            | <b>58</b>  |
| 3.1      | Introduction . . . . .                                                                                                                 | 58         |
| 3.2      | Mie particle thermophysical properties database . . . . .                                                                              | 63         |
| 3.2.1    | Mie particle solid-liquid equilibria (SLE) . . . . .                                                                                   | 63         |
| 3.2.2    | Mie particle thermophysical properties computation and validation . . . . .                                                            | 66         |
| 3.2.3    | Mie particle database description . . . . .                                                                                            | 67         |
| 3.3      | Continuous Artificial Neural Network based Equation of State for fluid and solid states (FE-ANN(s) EoS) for the Mie particle . . . . . | 67         |
| 3.3.1    | FE-ANN(s) EoS: training and performance . . . . .                                                                                      | 68         |
| 3.3.2    | FE-ANN(s) EoS: thermophysical properties modeling and prediction . . . . .                                                             | 73         |
| 3.4      | Conclusions . . . . .                                                                                                                  | 78         |
| <b>4</b> | <b>Simulation and data-driven modeling of the transport properties of the Mie fluid</b>                                                | <b>82</b>  |
| 4.1      | Introduction . . . . .                                                                                                                 | 82         |
| 4.2      | Mie fluid transport properties database . . . . .                                                                                      | 84         |
| 4.2.1    | Molecular dynamics simulations . . . . .                                                                                               | 85         |
| 4.2.2    | Transport properties database description . . . . .                                                                                    | 89         |
| 4.3      | Physically-inspired data-driven modeling of the Mie fluid transport properties . . . . .                                               | 92         |
| 4.3.1    | Transport property data-driven models formulation . . . . .                                                                            | 92         |
| 4.3.2    | Transport properties of the dilute Mie gas . . . . .                                                                                   | 96         |
| 4.3.3    | Transport property data-driven models training and architecture . . . . .                                                              | 97         |
| 4.3.4    | Transport properties data-driven models performance and predictions . . . . .                                                          | 99         |
| 4.4      | Conclusions . . . . .                                                                                                                  | 109        |
| <b>5</b> | <b>Parametrization of quasi-spherical molecules through the Mie potential</b>                                                          | <b>111</b> |
| 5.1      | Introduction . . . . .                                                                                                                 | 111        |
| 5.2      | Multi-property parametrization approaches . . . . .                                                                                    | 113        |
| 5.2.1    | Experimental data acquisition . . . . .                                                                                                | 114        |
| 5.2.2    | Thermophysical properties modeling . . . . .                                                                                           | 115        |
| 5.2.3    | Multi-start optimization approach . . . . .                                                                                            | 117        |
| 5.3      | Parametrization results for quasi-spherical molecules . . . . .                                                                        | 117        |
| 5.3.1    | Monoatomic molecules . . . . .                                                                                                         | 118        |
| 5.3.2    | Linear molecules . . . . .                                                                                                             | 120        |
| 5.3.3    | Tetrahedral molecules . . . . .                                                                                                        | 122        |
| 5.4      | Conclusions . . . . .                                                                                                                  | 124        |
| <b>6</b> | <b>Conclusions and outlook</b>                                                                                                         | <b>128</b> |
| 6.1      | Suggested future work . . . . .                                                                                                        | 129        |
| 6.1.1    | Machine learning-based SAFT EoS . . . . .                                                                                              | 129        |
| 6.1.2    | Molecular descriptors embedded into the FE-ANN EoS . . . . .                                                                           | 131        |
| 6.1.3    | Application of the FE-ANN EoS to other molecular fluids . . . . .                                                                      | 132        |
| 6.1.4    | Extension of the FE-ANN EoS to mixtures . . . . .                                                                                      | 133        |

|                                                                                                                |            |
|----------------------------------------------------------------------------------------------------------------|------------|
| <b>Bibliography</b>                                                                                            | <b>135</b> |
| <b>A Data modeling with artificial neural networks</b>                                                         | <b>158</b> |
| A.1 Artificial neural networks description . . . . .                                                           | 158        |
| A.2 Hyperparameter optimization of the physics-informed artificial neural network models                       | 162        |
| <b>B Thermophysical properties modeling of selected Mie particles</b>                                          | <b>166</b> |
| B.1 FE-ANN EoS results for the $\lambda_r=10, 18$ , and $26$ , and $\lambda_a=6$ Mie fluids. . . . .           | 167        |
| B.2 FE-ANN(s) EoS results for the $\lambda_r=10, 18$ , and $26$ , and $\lambda_a=6$ Mie particles. . . . .     | 171        |
| B.3 Transport properties results for the $\lambda_r=10, 18$ , and $26$ , and $\lambda_a=6$ Mie fluids. . . . . | 175        |
| <b>C Further parametrization results of quasi-spherical molecules</b>                                          | <b>176</b> |



# List of Figures

|      |                                                                                                                            |     |
|------|----------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1  | Approach to develop a machine-learned equation of state. . . . .                                                           | 31  |
| 2.2  | Sanity checks for computed thermophysical properties. . . . .                                                              | 40  |
| 2.3  | Mie fluid Thermophysical properties database distribution. . . . .                                                         | 42  |
| 2.4  | Squared error distribution of the FE-ANN EoS for selected first- and second-order derivative properties. . . . .           | 46  |
| 2.5  | Parity plot of thermophysical properties computed with FE-ANN EoS with respect to the reference database. . . . .          | 48  |
| 2.6  | Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                          | 49  |
| 2.7  | Second-order derivative properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                | 51  |
| 2.8  | Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                    | 52  |
| 2.9  | Critical point of ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                                       | 53  |
| 2.10 | Brown's characteristic curves and second virial coefficient of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . . | 55  |
| 3.1  | Maxwell thermodynamic surface. . . . .                                                                                     | 59  |
| 3.2  | Approach to develop a continuous equation of state for fluid and solid phases. . . . .                                     | 62  |
| 3.3  | Summary of the freeze method to compute SLE in molecular dynamic simulations. . . . .                                      | 64  |
| 3.4  | Mie particle thermophysical properties database distribution. . . . .                                                      | 69  |
| 3.5  | Parity plot of thermophysical properties computed with FE-ANN(s) EoS with respect to the reference database. . . . .       | 71  |
| 3.6  | Squared error distribution of the FE-ANN(s) EoS for selected first- and second-order derivative properties. . . . .        | 73  |
| 3.7  | Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                       | 74  |
| 3.8  | Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                 | 75  |
| 3.9  | Critical and triple points of ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                        | 77  |
| 3.10 | Isochoric heat capacity of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                  | 77  |
| 3.11 | Second-order properties of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                  | 79  |
| 3.12 | Brown's characteristic curves and virial coefficients of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .    | 80  |
| 4.1  | Mie fluid transport properties database distribution. . . . .                                                              | 91  |
| 4.2  | Transport data simulation data of the Lennard-Jones fluid ( $\lambda_r = 12$ and $\lambda_a = 6$ ). . . . .                | 91  |
| 4.3  | Finite-size effect correction to the self-diffusivity. . . . .                                                             | 92  |
| 4.4  | Self diffusivity of the Lennard-Jones fluid ( $\lambda_r = 12$ and $\lambda_a = 6$ ). . . . .                              | 94  |
| 4.5  | Shear viscosity and thermal conductivity of the Lennard-Jones fluid ( $\lambda_r = 12$ and $\lambda_a = 6$ ). . . . .      | 95  |
| 4.6  | Transport properties at zero density of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids . . . . .                       | 102 |

|      |                                                                                                                                                              |     |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.7  | Parity plots and error distribution of transport property models based on ANNs ( $\rho^* D^*$ , $\ln \eta^*$ and $\ln \kappa^*$ ). . . . .                   | 103 |
| 4.8  | Parity plots and error distribution of transport property models based on ANNs ( $\rho^* D^{*,res}$ , $\ln \eta^{*,res}$ and $\ln \kappa^{*,res}$ ). . . . . | 104 |
| 4.9  | Transport properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                                                | 106 |
| 4.10 | Predicted transport property models against published molecular dynamics data. . .                                                                           | 108 |
| 5.1  | Critical and triple temperature ratio of selected quasi-spherical molecules. . . . .                                                                         | 118 |
| 5.2  | Selected thermophysical properties of Argon. . . . .                                                                                                         | 121 |
| 5.3  | Selected thermophysical properties of Nitrogen. . . . .                                                                                                      | 123 |
| 5.4  | Selected thermophysical properties of Methane. . . . .                                                                                                       | 125 |
| 5.5  | Selected thermophysical properties of Tetrafluoromethane. . . . .                                                                                            | 126 |
| A.1  | Schematic representation of an Artificial Neural Network (ANN). . . . .                                                                                      | 158 |
| A.2  | Hyperparameter optimization results of the FE-ANN EoS developed in Chapter 2. . .                                                                            | 163 |
| A.3  | Hyperparameter optimization results of the FE-ANN(s) EoS developed in Chapter 3. .                                                                           | 164 |
| A.4  | Hyperparameter optimization results of the self-diffusivity ANN-based model developed in Chapter 4. . . . .                                                  | 164 |
| A.5  | Hyperparameter optimization results of the shear viscosity ANN-based model developed in Chapter 4. . . . .                                                   | 165 |
| A.6  | Hyperparameter optimization results of the thermal conductivity ANN-based model developed in Chapter 4. . . . .                                              | 165 |
| B.1  | Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                                                            | 167 |
| B.2  | Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                                                      | 168 |
| B.3  | Second-order derivative properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . .                                                    | 169 |
| B.4  | Brown's characteristic curves and virial coefficients of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. .                                               | 170 |
| B.5  | Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                                                         | 171 |
| B.6  | Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                                                   | 172 |
| B.7  | Second-order properties of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. . . . .                                                                    | 173 |
| B.8  | Brown's characteristic curves and virial coefficients of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles. .                                            | 174 |
| B.9  | Transport properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. . . . .                                                                | 175 |
| C.1  | Selected thermophysical properties of Krypton. . . . .                                                                                                       | 177 |
| C.2  | Selected thermophysical properties of Xenon. . . . .                                                                                                         | 178 |
| C.3  | Selected thermophysical properties of Carbon monoxide. . . . .                                                                                               | 179 |

# List of Tables

|     |                                                                                                                                               |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1 | Definition of physical quantities in reduced units. . . . .                                                                                   | 36  |
| 2.2 | Mie fluid database description. . . . .                                                                                                       | 41  |
| 2.3 | Comparison of Mean Squared Error (MSE) of selected trained FE-ANN EoS. . . . .                                                                | 45  |
| 3.1 | Initial guesses for the coexistence step in the freeze method. . . . .                                                                        | 65  |
| 3.2 | Mie particle database description. . . . .                                                                                                    | 68  |
| 3.3 | Squared error distribution of selected thermophysical properties as obtained from the FE-ANN(s) EoS. . . . .                                  | 72  |
| 4.1 | Definition of physical quantities in reduced units. . . . .                                                                                   | 85  |
| 4.2 | Mie fluid transport properties database description. . . . .                                                                                  | 90  |
| 4.3 | Coefficients $a_{ij}^{(k,k)}$ needed to evaluate the collision integral of $(\lambda_r, \lambda_a = 6)$ Mie fluids. . . . .                   | 97  |
| 4.4 | Transport properties Mean Absolute Percentage Error (MAPE) from transport property models based on artificial neural networks (ANNs). . . . . | 101 |
| 5.1 | Parametrization approaches to obtain the Mie potential parameters. . . . .                                                                    | 114 |
| 5.2 | Optimization details to fit the Mie potential molecular parameters. . . . .                                                                   | 117 |
| 5.3 | Optimized Mie molecular parameters for selected quasi-spherical molecules. . . . .                                                            | 119 |
| A.1 | Common activation functions used in ANNs. . . . .                                                                                             | 160 |

## Chapter 1

# Introduction

Thermophysical properties are of great importance for design and engineering purposes [4–6]. Thermophysical properties include thermodynamic (volumetric, thermal, and equilibrium) and transport (diffusivity, shear viscosity, and thermal conductivity) properties. Volumetric properties (e.g., density and pressure) are basic thermophysical data required for process equipment sizing. Thermal properties (e.g., heat capacity and enthalpy) are key for the design of any process that involves a temperature change. Moreover, they allow for the assessment of key processes to achieve energy mitigation goals, such as organic Rankine cycles [7] and heat pumps [8]. Additionally, an entropy analysis, also known as exergy analysis, allows for the optimization of the global efficiency of a chemical process [9]. Equilibrium properties refer to the coexistence of multiple phases and are essential for the design of separation equipment, such as flash separators, distillation columns, liquid-liquid extractors, and absorption columns [6]. Equilibrium data are also necessary to avoid cavitation in pumps and condensation in compressors. Transport properties, such as the diffusion coefficient, are essential for designing separation equipment. Similarly, the shear viscosity is needed for piping and pump sizing [10], while the thermal conductivity is required for heat exchanger design [11]. For these reasons, thermophysical property data have been collected from several sources, such as the U.S. National Institute of Standards and Technology (NIST) [12] and the DECHEMA consortium [13]. Furthermore, accurately modeling these properties allows for assessing the feasibility of a chemical process and, in the best of cases, optimizing it.

The correct modeling of thermophysical properties is essential for design purposes [14, 15]. The thermophysical property modeling approaches can be classified as empirical, semi-empirical, and theoretically based. The empirical approach fits a given property dataset to a simple, non-physical mathematical formula. NIST has widely used the empirical approach to develop technical equations of state (EoSs) [16], for example, for water (IAPWS-97) [17], natural gas mixtures (GERG-2008) [18], and propane [19]. These EoSs explicitly model the Helmholtz free energy, from which volumetric, thermal, and equilibrium properties can be derived. Additionally, the empirical approach is widely used in databases such as the NIST TDE database [12] and the DIPPR 801 project [13] to interpolate the thermophysical properties of pure fluids. The empirical approach offers high accuracy at the cost of a lack of generalization. Therefore, an empirical model should only be used within its fitted range.

The semi-empirical modeling approach consists of a simplified theory fitted to reproduce experimental data. The semi-empirical approach has been widely used for phase equilibrium calculations where

independent EoSs are used for the vapor and liquid phases. In this context, the vapor phase is commonly considered an ideal gas, and the liquid phase nonidealities are obtained from activity coefficient models such as the NRTL [20], Wilson [21], UNIQUAC [22], and UNIFAC [23] models. These activity coefficient models are pressure-independent and do not predict an excess volume. The ideal gas condition and neglective pressure dependency of activity coefficients are valid at low to moderate pressures, which restricts the applicability of this modeling approach. The semi-empirical approach has also been used to model transport properties. The free volume theory [24] and friction theory [25] are examples of semi-empirical models to predict the shear viscosity. Additionally, entropy scaling [26] has been suggested as a link between transport and thermodynamic properties for dense phases. These theories rely on thermophysical data from an EoS and multiple component-specific adjustable parameters. Semi-empirical models have clear limitations to their range of applicability. However, they are still based on a theory, therefore they tend to generalize better than empirical models.

Lastly, theoretically-based models generalize better than the previously mentioned approaches but require a more rigorous formulation, where the thermophysical properties of a molecular model are obtained using statistical mechanics. The statistical mechanics solution of a molecular model can be translated into a mathematical formulation using analytical approximations or by directly using the molecular simulation results. Theoretically-based Equations of State (EoSs) are the most common models for obtaining volumetric, thermal, and equilibrium properties of fluids. The ideal gas law is one of the few systems for which the rigorous statistical mechanics of non-interacting particles leads to a closed-form equation [27]. This simple molecular model is valid for low pressures. A more real representation considers that molecules have volume and exhibit dispersive forces. Under these assumptions, van der Waals [28] developed a simple continuous EoS to describe dilute (vapor) and dense (liquid) phases. The van der Waals EoS and its modifications [29, 30] have been widely adopted due to their simple mathematical form and flexibility in modeling from simple fluids to complex mixtures [31]. The main limitations of this kind of EoS are their inability to accurately predict liquid densities [32] and thermal-derived properties [33]. Some of these flaws have been addressed through empirical modifications [34].

The Statistical Association Fluid Theory (SAFT) EoSs [35], such as PC-SAFT [36] and SAFT-VR-Mie [37], are theoretically based EoSs that perform well for both volumetric and thermal properties. The SAFT formalism considers an underlying molecular model and can explicitly account for molecular association, providing a framework to account for hydrogen bonding (partial association) and chain formation (complete association). Moreover, the SAFT EoSs are based on perturbation theory, and extra nonideal contributions to the free energy of a fluid can be added as needed. In this context, the free energy of a molecule is initially described by the kinetic (ideal) contribution and dispersion (nonbonded) interactions. Other contributions can be considered depending on the nature of the molecule. For example, non-spherical molecules, such as alkanes, can consider a chain contribution term. Additionally, molecules capable of forming hydrogen bonds, such as alcohols and water, consider an association contribution to the free energy. This formalism can also consider other sources of nonidealities. Examples of these are dipolar [38], quadrupolar [39], and ion-ion interactions [40]. Despite their flexibility and accuracy, SAFT EoSs are still not widely adopted. This is mainly due to

their complex numerical implementation, the increased computational cost of using them compared to classical EoSs, and because there is general confusion by non-experts associated with the multiple variants of SAFT models [14, 41, 42]. These accessibility problems have been partially addressed by open-source packages that facilitate the direct use of these EoSs [43–46].

Transport properties have also been modeled using a theoretically-based approach. The Chapman-Enskog theory allows computing the transport properties of a dilute gas given an interaction potential. [27] In this theory, a transport property is related to an expansion based on collision integrals. It is estimated that the first-order approximation of this expansion can predict a transport property with a 1% error. The computation of the collision integrals is numerically demanding. For this reason, their solution has been correlated for the Lennard-Jones [27] and Mie fluid [47]. The Chapman-Enskog theory is valid only close to the zero density limit and correctly predicts the temperature dependence of a transport property. The revised Enskog theory extends the Chapman-Enskog theory to denser states by considering structural information of the molecular fluid. This theory has been successfully coupled with the SAFT-VR-Mie EoS to predict transport properties of the Mie fluid using parameters fitted only to equilibrium properties [48]. Despite these promising results, the accuracy and range of validity of this extended theory have not been verified against molecular simulation results.

Although numerous approaches exist for modeling thermophysical properties, an industrial survey conducted by Kontogeorgis et al. [14] revealed a need for more accurate models, particularly for complex systems such as polymers, hydrogen-bonding fluids, and charged species. These models should accurately describe experimental data, extrapolate to unknown molecular systems, and be computationally efficient. Based on that, machine learning algorithms provide a suitable alternative for correlating and predicting thermophysical properties [49, 50]. Given that reliable data are available, machine learning models can bypass the need for a rigorous statistical mechanics solution to predict thermophysical properties. These algorithms can serve as universal function approximators for learning structure-property relationships in molecular systems. However, this advantage comes with the risk of overfitting, limited generalization, and producing thermophysical property models that lack physical meaning. This thesis aims to develop machine-learning methods for predicting thermophysical properties while ensuring thermodynamic consistency.

The rest of this chapter is structured as follows. Section 1.1 discusses recent developments in machine learning methods for data-driven thermodynamics. Based on that, an overview of this thesis is provided in Section 1.2.

## 1.1 Data-driven thermodynamics

Machine learning models have been used extensively for molecular modeling purposes. Machine learning methods provide a flexible mathematical framework for finding patterns in data [51]. For example, artificial neural networks (ANNs) [52] are universal function approximators [53]. This characteristic makes machine learning algorithms suitable to correlate data, such as thermophysical properties [50, 54], molecular interactions (force fields) to perform molecular simulations [55] and molecular design [56]. There are two main approaches to using machine learning for molecular

modeling: a purely data-driven approach and a hybrid physics-informed data-driven approach. This section will discuss contributions using these approaches.

In the purely data-driven approach, a machine learning model is used to correlate a single target property. As a proof of concept and for education purposes, Joss and Müller [57] correlated the normal boiling points of over 6000 species using an ANN. Despite the simplicity of this task, it was recognized that selecting the proper molecular descriptors and regularizing the data improves the performance of the model. Similarly, Zhu and Müller [58] used ANNs and Gaussian process regression to correlate several thermophysical properties (phase equilibria, critical point, and supercritical density) of the Mie fluid. An equation of state based on the Helmholtz free energy requires iterative calculations based on the free energy derivatives to predict the same properties. The conjunction of the developed machine learning models directly predicts the target thermophysical property and bypasses the need for these iterative calculations. Similar to the work of Joss and Müller [57], it is observed that manipulating the features and outputs of the ANN can improve the accuracy of the models.

Machine learning algorithms are not restricted to experimental data. Gong et al. [59] generated pseudo-experimental data of alkanes using high-throughput molecular dynamics simulations. They were able to predict the density, internal energy, and isobaric heat capacity by using an artificial neural network that relies on the molecular fingerprints and the state conditions. Using molecular fingerprints enables trained artificial neural networks to predict the thermophysical properties of previously unseen alkanes. It must be noted that the quality of the predicted results follows the accuracy of the chosen force field to generate the reference thermophysical data.

The work of Gong et al. [59] showcases how considering molecular information allows the machine learning algorithms to model unseen molecules. In this context, a purely data-driven approach has been used to develop group contribution methods. In this context, molecules must be represented by a computer-friendly representation. Examples of these are the simplified molecular input line entry system (SMILES) [60] and the self-referencing embedded string (SELFIES) [61] notations, which allow representing molecular structures using strings. These nomenclatures encode information about the chemical composition (main atomic groups) and structure (e.g., bonds, branches, and rings). Alshehri et al. [62] used group contribution information obtained from the SMILES notation to correlate over 25 thermophysical properties, such as the critical point, melting point, and solubility in water using a Gaussian process regression. Group contribution methods can also benefit from graph neural networks (GNNs) for predicting thermophysical properties [63]. A GNN can directly handle molecular structures where each atom is a node that contains relevant information, such as atom type, number of bonds, and hybridization state. The nodes are then connected and allowed to exchange information. Based on this information, GNN can produce a numerical fingerprint of a molecule that is used to model thermophysical properties. These machine-learning-based group contribution methods explicitly consider the molecular structure and molecular environment of each functional group. Contrarily, classical group contribution methods, such as the one from Joback and Reid [64] or the SAFT- $\gamma$ -Mie EoS [65], only consider the molecular composition without explicitly considering the connectivity between groups. Group contribution methods based on machine learning can consider higher-order interactions between groups, and consequently, they hold the potential to outperform classical group contribution methods.

The described studies use a purely data-driven approach, where the machine learning algorithm directly learns from the data without relying on physical intuition. This purely data-driven approach has demonstrated good performance within the domain in which the algorithm was trained. However, in most cases, it lacks predictability because the machine learning model cannot go beyond the data itself. On top of that, in this black-box modeling approach, the physicality of these models is not fully understood. Concerning this last point, both academia [49, 50] and industry [14, 15] agree that models for the prediction of thermophysical properties should include a physical-inspired background that allows for validating the results. Moreover, a physics-inspired background can facilitate sensible extrapolation to unseen phenomena. These hybrid physics-informed, data-driven models are the focus of the rest of this thesis.

In this context, several approaches to leverage the flexibility of machine learning models while retaining the physicality of the predictions have been proposed. Three approaches will be covered here: machine learning to model interactions at the atom level, machine learning methods to predict parameters of existing thermodynamic models, and a physics-embedded machine learning model to predict thermophysical properties directly. In the first approach, the ultimate goal is to find an accurate representation of the interactions between atoms using a machine learning model [55]. Machine learning force fields are a promising alternative to classical force fields that rely on semi-empirical mathematical expressions to describe inter and intramolecular interactions [66]. Classical force fields explicitly consider nonbonded, bonded, angular, torsion, and electrostatic interactions. However, these interactions are often described using simple mathematical expressions that do not represent the exact underlying physics or the molecular environment of the interactions. Machine learning force fields are primarily trained on rigorous *ab initio* reference data. These force fields aim to provide *ab initio* accuracy to molecular dynamics simulations at a fraction of the computational cost. Moreover, machine learning force fields rely on descriptors that account for the molecular environment of each atom [67]. These descriptors are formulated following physical constraints, e.g., invariances of the system. Additionally, they provide a pathway to account for multi-body interactions. Machine learning potentials provide a detailed resolution of phenomena at the molecular level [68, 69]. Finally, in this approach, the thermophysical properties of a molecular system are then obtained using statistical thermodynamics [66, 70]. The use of machine learning force fields alongside molecular simulations provides valuable insights at the molecular level. However, the high computational cost of molecular simulations makes this approach unfeasible for design and engineering purposes.

The second approach relies on an already defined thermophysical properties model, such as an equation of state or activity coefficient model. This approach uses a machine learning model to predict the parameters of a model. The matrix completion method (MCM) is a prime example of this approach. In this method, there is a matrix of interactions with missing values. These missing values are then predicted based on the known values. Thermodynamic models mainly rely on pair interactions to predict the behavior of mixtures [71]. Therefore, the MCM is a suitable method to fill the gaps of missing interactions. This approach has been used to complete the matrix of activity coefficients at infinite dilution [72, 73]. This information is the heart of the parametrization of activity coefficient models, which then can be used to predict phase equilibria. The MCM has also been used to predict Henry's constants [74] that are necessary to predict the solubility of gases in solvents. Finally, the



MCM has been used to complete the interaction parameters matrix of the UNIQUAC model. The predicted interaction parameters for this model outperform the modified UNIFAC activity coefficient model [75]. The MCM offers an accessible pathway to predict the behavior of mixtures where no experimental data is available to parametrize a thermodynamic model.

The MCM is suitable for completing square matrices with missing information. This methodology is naturally applicable for predicting binary interaction parameters in the context of thermodynamic models. Other thermodynamic models require component-specific parameters. For example, Matsukawa et al. [76] used a hybrid approach to predict the molecular parameters of PC-SAFT EoS [36]. This approach relies on group contribution information that was then correlated to the molecular parameters using an artificial neural network. The proposed approach provides a suitable framework to model molecules without enough experimental data to parametrize the EoS. The group contribution method used in this work relies only on the chemical composition and does not consider the molecular structure. Therefore, it is not suitable to discriminate between isomers.

Machine learning models can process string information through the attention mechanism of transformers [77]. Transformers are the fundamental structure behind large language models (LLMs) [78], such as ChatGPT [79]. The Smiles-to-Property-Transformer (SPT) framework has been proposed to process SMILES strings [60] to predict activity coefficients at infinite dilution [80], parameters of the NRTL model [81] and parameters of the PC-SAFT EoS [82]. The SPT framework decodes the structure of a molecule based on its SMILES representation. The SMILES representation implicitly considers the molecular environment of each functional group in the molecule. This characteristic makes the SPT particularly well-suited for distinguishing between isomers. This state-of-the-art approach provides a suitable framework to predict the molecular parameters of a thermodynamic model. Moreover, the authors showed that the predicted molecular parameters correctly describe seen and unseen molecular systems.

Machine learning has been successfully used to predict the parameters of thermodynamic models. Additionally, the inherent group contribution nature of this machine learning approach makes it compatible with molecular design algorithms [56]. However, the main drawback of this approach is that the accuracy of the predictions is limited to the capabilities of the chosen thermodynamic model. Most thermodynamic models are based on simplified theories to represent a molecular model, which might be unsuitable for modeling multiple target thermophysical properties [31]. The limitations of current thermodynamic models can be addressed by directly embedding physical knowledge into a machine-learning model.

The third hybrid modeling approach considers embedding physical knowledge into a machine learning model to predict thermophysical properties. For example, Craven et al. [83] modeled the radial distribution function (RDF) of the Lennard-Jones fluid using a multivariate function decomposition approach. The RDF provides a link between structural and macroscopic properties such as the pressure or the internal energy [27]. Other thermophysical properties can be computed using standard thermodynamic relationships, see Section 2.2. This approach is a clear example of how a hybrid theoretical and data-driven approach successfully describes the thermodynamic behavior of the Lennard-Jones fluid.

A similar hybrid approach has been used to produce thermodynamically consistent activity coefficient models [84, 85]. Rittig et al. [84] used a graph neural network to predict the activity coefficient of binary mixtures, and the Gibbs-Duhem equation is used as a penalty in the loss function [86] to enforce a consistent thermodynamic behavior. This penalty is a soft constraint on the graph neural network. Therefore, there is no guarantee that it will be exactly fulfilled. The authors show that this constraint is not fulfilled over the entire dataset. Contrarily, Specht et al. [85] proposed a Hard-constrained Neural Network for consistent Activity coefficient prediction (HANNA). HANNA explicitly models the Gibbs excess free energy. Activity coefficients are obtained as a partial molar property of the Gibbs free energy. In this case, the molecular fingerprints are obtained from a transformer [77]. Additionally, this work considers further constraints for the activity coefficients. For example, activity coefficients equal to one for pure species, tend to be one if the molecules are similar, and the order of the mixture should not affect the predicted activity coefficient. These constraints are embedded in the model formulation and ensure the thermodynamic consistency of the results. The benefits of this physics-informed data-driven approach arise naturally, as HANNA exhibits better performance than the UNIFAC activity coefficient model. Finally, the HANNA framework has yet to be extended to multicomponent mixtures.

Finally, Rosenberger et al. [87] proposed using an artificial neural network to implicitly learn the Helmholtz free energy of a fluid (FE-ANN EoS) through its derivative properties. This idea was tested for the van der Waals and the Lennard-Jones fluids. The authors trained the model using the pressure, chemical potential, and internal energy data. These properties are directly available from automatic differentiation [88]. The main advantage of this approach is that Maxwell's relations are exactly fulfilled. Hence, the trained FE-ANN EoS is thermodynamically-consistent. Additionally, the trained Helmholtz free energy allows computing of thermophysical properties for which the EoS has not been trained. For example, heat capacities, speed of sound, and isothermal compressibility are related to the second-order derivatives of the Helmholtz free energy. This approach is suitable for generating a "digital Equation of State". However, the publication of Rosenberger et al. [87] does not consider relevant physical limits and constraints that require further research. For example, the ideal gas limit is not mentioned. This limit is well known from a theoretical perspective [27]. It is unclear if the free energy model can represent the state of fluids at zero density. Additionally, the internal thermodynamic consistency of the model determined by Brown's characteristic curves was not addressed. These curves represent conditions where an equation of state predicts the behavior of an ideal gas. These curves must follow the behavior described by Deiters and De Reuck [89]. Any deviation from the expected behavior showcases the thermodynamic inconsistency of an equation of state. For these reasons, this PhD aims to build on the work Rosenberger et al. [87] but to address the shortcomings of the framework.

The benefits of embedding a theoretical background in a data-driven model are clear. These hybrid models exploit the flexibility of the machine learning algorithms while fulfilling well-established physical constraints. This modeling approach holds the potential to overcome the limitations of current theoretically-based models to handle complex behavior. For this reason, this modeling approach is at the heart of this PhD thesis. This project aims to explore further the development of a physics-informed machine-learning free energy model to address the limitations of the FE-ANN EoS proposed

by Rosenberger et al. [87]. Furthermore, this hybrid approach can also be used to model transport properties (i.e., diffusivity, viscosity, and thermal conductivity). No rigorous theoretically-based framework exists to model these properties over the entire phase space. In this case, a physics-informed data-driven modeling approach is convenient.

## 1.2 Thesis overview

This thesis focuses on embedding physical knowledge into machine learning algorithms to model and predict thermophysical properties. The development of these models will be showcased for systems interacting through a truncated Mie potential [90, 91].

$$\mathcal{U}^{\text{Mie}} = \begin{cases} \mathcal{C}^{\text{Mie}} \epsilon \left[ (\sigma/r)^{\lambda_r} - (\sigma/r)^{\lambda_a} \right] & \text{if } r \leq 5\sigma \\ 0 & \text{if } r > 5\sigma \end{cases} \quad (1.1)$$

$$\mathcal{C}^{\text{Mie}} = \frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}$$

Here,  $\mathcal{U}^{\text{Mie}}$  is the interaction energy between two Mie particles,  $\epsilon$  is the interaction energy well depth,  $\sigma$  is the length scale, which is loosely related to the effective particle diameter, and  $r$  is the center-to-center distance between two particles. Finally,  $\lambda_r$  and  $\lambda_a$  are the repulsive and attractive exponents. These exponents control the steepness of these interactions. It has been shown that the exponents of the Mie potential are conformal [92]. The conformality implies that multiple repulsive and attractive exponent combinations lead to the same fluid macroscopic equilibrium properties, such as the vapor liquid equilibria. Therefore, the interplay between the repulsive and attractive exponents becomes relevant to describe the Mie potential. For simplicity, the attractive exponent is commonly set to 6 in agreement with the London theory for dispersion forces [93], resulting in what is sometimes referred to as the  $(\lambda_r, 6)$  Mie potential [91]. In this case, the Mie potential is fully described by the  $\epsilon$ ,  $\sigma$ , and  $\lambda_r$  parameters. Finally, the Mie potential reduces to the Lennard-Jones potential when the repulsive exponent is set to 12.

The Mie potential has gained attention due to its flexibility in modeling the thermophysical properties of industrially relevant fluids [94–96]. For example, the Mie potential has been used as a model to represent the fluid phase equilibria of simple [97] and asymmetric [98] mixtures, biofuel blends [99, 100], active pharmaceutical ingredients (API) [101, 102], cryogens with quantum effects [95, 103] and polymers [104, 105]. Additionally, the Mie potential has been used to model the self-assembly of liquid crystals [106], superspreading of surfactants [107], asphaltene aggregation [108], and interfacial tensions [43, 109]. Furthermore, it has been used as a model to describe the transport properties, such as bulk [110] and shear viscosities [111, 112] and diffusion coefficients [48] of complex fluids.

For design and engineering purposes [14, 41], it is necessary to obtain accurate data for *both* thermodynamic properties (e.g., vapor pressure, phase equilibria, heat capacities) and transport properties (e.g., diffusivity, shear viscosity, thermal conductivity) of a fluid (or pseudo-fluid in this case) over a wide range of conditions. These properties can be directly computed from molecular simulations. However, the computational cost and speed at which molecular simulations can be integrated prohibit

its use for practical engineering purposes. It is preferred to rely on fast-to-evaluate mathematical models that can accurately predict these properties. Thermodynamic properties can be appropriately obtained from molecular-based Equations of State (EoS) and empirical correlations. In the case of the Mie fluid, several EoSs are available to represent its thermodynamic properties. Among them are the SAFT-VR-Mie EoS [37] and its group contribution version, the SAFT- $\gamma$ -Mie EoS [65], which are based on perturbation theory. Reimer et al. [113] proposed an EoS for the Mie fluid based on the uv-theory [114]. An empirical EoS proposed by Pohl et al. [115] also models Mie fluids accurately. The transport properties of the Mie fluid have been modeled using revised Enskog theory [48]. The existing thermophysical property models will also allow a comparison of the developed physics-informed machine-learning models against traditional methods.

Accordingly, the contributions of this thesis are divided into four chapters. Chapter 2 extends the scope of the FE-ANN EoS proposed by Rosenberger et al. [87] to address the ideal gas limit analytically. Additionally, further constraints to fulfill stability constraints [116] are incorporated into the model training. Due to its physics-inspired formulation, the FE-ANN EoS can predict unseen behavior, such as phase equilibria. The FE-ANN is trained directly using molecular dynamics simulations, and its performance is par with the monomer term of the SAFT-VR-Mie EoS [37]. Chapter 3 then exploits the flexibility of the FE-ANN EoS framework to handle both fluid and solid states with a single equation of state. The FE-ANN(s) EoS (an extra "s" is added to emphasize that this EoS handles the solid phase) can predict interesting behavior, such as the entire phase diagram and the triple point. This EoS is the first to predict such behavior. Moreover, the FE-ANN(s) EoS fulfills rigorous thermodynamic consistency tests. Chapter 4 models transport properties, including self-diffusivity, shear viscosity, and thermal conductivity. These models are directly trained using molecular simulation data, address the dilute gas limit as predicted by the Chapman-Enskog theory [117], and incorporate constraints to ensure a correct density dependence. The produced transport property models are accurate and well-behaved in the fluid phase space. Finally, the results from Chapters 2, 3, and 4 provide a comprehensive description of the thermophysical properties of the Mie particle. In Chapter 5, these machine-learning models are used to test different parametrization schemes of the Mie potential for quasi-spherical molecules. The limitations of the Mie potential to simultaneously target fluid, solid, and transport properties data are clear. Interestingly, due to their dependence on repulsive interactions, data on solid and transport properties target similar molecular parameters. In summary, this thesis lays the foundations for developing physics-embedded machine-learning models for modeling and predicting thermophysical properties. These findings, along with suggestions for future work, are presented in Chapter 6.

## Chapter 2

# Thermodynamically-consistent machine-learned equation of state

## 2.1 Introduction

The thesis of Johannes D. van der Waals [28] provoked a transformational change in the way thermophysical properties of fluids were understood and to be looked at in the future. At the time, the experiments of Thomas Andrews with carbon dioxide in 1869 and those of Kamerlingh Onnes (his colleague) with helium in 1908 provided insights into the pressure-volume-temperature relationship in both liquids and gases as well as the interpretation of a critical condition that limits the coexistence of these phases. After a rigorous theory-based procedure, in which van der Waals considered that molecules exhibit both attraction and repulsion, he derived a relatively simple mathematical function (a polynomial) capable of qualitatively fitting experimental data for simple fluids. The general functional form of pressure as a function of density and temperature (i.e.,  $P = f(T, \rho)$ ), later known as a "cubic" equation of state (EoS), has guided over a century of research [31] in the search for a general mathematical function (or formulation) that relates the macroscopic thermodynamic properties of fluids to each other (e.g., the pressure-volume-temperature relationship). Until the van der Waals era, the search for this functionality had been exclusively driven by empiricism [118], and it is now widely accepted that his work inspired a whole new research field.

Since the beginning of the 20<sup>th</sup> century, the field has focused on finding a systematic route to obtain property-structure relationships by recognizing the molecular nature of matter and incorporating the theoretical developments from statistical mechanics. An early example of this was the elegant derivation of the ideal gas law by integrating the canonical partition function for a system of non-interacting particles. Subsequently, the application of a Sutherland potential to describe inter-particle interactions led to the re-derivation of the van der Waals EoS. Similarly, the Carnahan-Starling hard-sphere EoS [119], which was originally derived semi-empirically, was recovered from integral equation theory. The pathway developed over the past century begins with a conceptual description of a molecular model (i.e., interaction potential). Provided that the molecular model was simple enough, and after a series of (usually) long and convoluted theoretical steps and approximations, one would ideally be able to develop an analytical theory capable of describing complex macroscopic observables. Examples of this pathway to develop EoSs have been broadly described in open literature [120–122].

Despite the practical success of this pathway, in most cases, there is a divorce between the parameters that define the molecular model and their appearance in the theory. This discrepancy arises from the necessary approximations and simplifications made when deriving the closed-form equations. In essence, this pathway provides a flexible mathematical formulation that can be used for practical engineering and design purposes upon fitting it to experimental data. Only a few exceptions exist where the parameters that define the conceptual molecular model match those in the final theory while carrying the same information [37]. In this case, a clear connection between macroscopical properties and molecular parameters can be made. This subtle point is a crucial trait for an EoS, as it allows direct back-mapping of intermolecular potentials from experimental data. This desired characteristic in an EoS is one of the most promising avenues for coarse-graining force fields [123]. The theoretical pathway for determining equations of state (EoSs) of fluids has been long, convoluted, and often unsatisfactory, as evidenced by the fact that there is no universally accepted EoS [31] even for the simplest of fluids [42]. A more direct approach is to rely on empirical methods (e.g., by least-squares minimization with respect to experimental data) to correlate fluid phase properties. A prime example are the technical EoSs [16–19] used by the U.S. National Institute of Standards and Technology (NIST) [12]. Although very effective and accurate, their applicability is limited to a few well-studied fluids. This approach can exhibit unphysical behavior and is suboptimal as it faces hurdles due to the scarcity of high-quality experimental data.

This chapter explores a new paradigm for predicting thermodynamic properties that simultaneously addresses the above-mentioned issues by spawning a potentially unlimited amount of raw data *in silico* to train directly machine learning models. This approach mimics the conceptual process currently used in developing thermodynamic theories, in which a mathematical representation is constructed to map the thermophysical properties of a given molecular model. This approach relies on the statistical mechanics solution of the molecular model, which is obtained numerically from molecular simulations. However, the proposed approach diverges from the current paradigm by replacing the lengthy, burdensome, and (sometimes) uncertain theoretically-based approach with a physics-embedded machine learning model. This machine learning model will be supplemented with physically sound constraints to ensure correct thermodynamic behavior. Moreover, due to its nature, the machine learning model directly learns from the thermophysical property data and does not make any simplifying assumptions. *In effect*, the machine learning model "learns" the statistical mechanics' solution and provides discrete point-wise information on the link between molecular parameters that define molecules and the corresponding macroscopical observables. Similar to theoretically-based EoSs, the link to experimental data can be established later by reverse-calculating and optimizing (fitting) the force field parameters. In essence, this approach generates a "digital equation of state" that links the "genotype" of the molecule (its molecular description) with its "phenotype" (its thermodynamic observables).

As described in Chapter 1, machine learning has been extensively used for modeling thermophysical properties. For example, Rosenberger et al. [87] proposed using an Artificial Neural Network (ANN) to "learn" the Helmholtz free energy of a fluid through its first-derivative properties. This idea was successfully tested on the van der Waals and Lennard-Jones fluids. The main advantage of this approach is that all thermophysical properties (e.g., the pressure and internal energy) are obtained from the same Helmholtz free energy surface, which guarantees that Maxwell's relations

are fulfilled. This constraint is a requirement for ensuring the thermodynamic consistency of an EoS. Another advantage of this approach is that the Helmholtz free energy enables the computation of thermophysical properties for which the model has not been trained. Additionally, the description of the Helmholtz free energy can then be used to study interfaces through Density Gradient Theory (DGT) [43, 109, 124]. These characteristics make this framework suitable for generating a "digital equation of state".

This chapter builds upon the work of Rosenberger et al. [87] by enforcing well-accepted thermodynamic limits, such as the ideal gas law, along with additional thermodynamically inspired constraints related to the thermal and mechanical stability of a phase [116, 125] into the model formulation and training. The content of this chapter is mainly based on an already published article (Gustavo Chaparro and Erich A. Müller. "Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid". *The Journal of Chemical Physics* 158.18 (2023), p. 184505. DOI: 10.1063/5.0146634, Ref. [3]). However, the results shown here have been updated by training the EoS on a more extensive database and incorporating extra thermodynamically inspired constraints. The original publication considered only a constraint to enforce the criteria of thermal stability. The results presented here additionally consider constraints to ensure mechanical stability and preclude crossing pressure isotherms.

As a proof of concept, this chapter focuses on generating a digital equation of state using classical molecular simulation data of fluid systems interacting through a truncated Mie potential. This force field allows the representation of the vapor-liquid behavior of an extensive range of molecules [97, 126], as well as the interfacial properties of complex systems [98, 109, 127, 128]. Moreover, the Mie potential has also been used to describe other complex phenomena, such as the phase equilibria of fluids with quantum effects [95, 103], polymer systems [105, 129], asphaltene aggregation [108], self-assembly of liquid crystals [106], among others. Additionally, there exist both theoretically-based [37, 113] and semi-empirical [115] EoSs for the Mie fluid that can be used to assess the performance of the "digital" EoS. The Mie potential,  $U^{\text{Mie}}$ , is described as follows.

$$U^{\text{Mie}} = \begin{cases} C^{\text{Mie}} \epsilon \left[ (\sigma/r)^{\lambda_r} - (\sigma/r)^{\lambda_a} \right] & \text{if } r \leq 5\sigma \\ 0 & \text{if } r > 5\sigma \end{cases} \quad (1.1)$$

$$C^{\text{Mie}} = \frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}$$

The Mie potential is characterized by an energy parameter ( $\epsilon$ ), a size parameter ( $\sigma$ ), and the repulsive and attractive exponents ( $\lambda_r$  and  $\lambda_a$ ). Due to the inherent conformality of the potential [92], the attractive exponent ( $\lambda_a$ ) can be taken to be 6 [93] without lack of generality. Hence, only three parameters ( $\epsilon$ ,  $\sigma$ ,  $\lambda_r$ ) are required to describe a fluid using the Mie potential. The Mie potential simplifies to the well-known Lennard-Jones potential when  $\lambda_r = 12$  and  $\lambda_a = 6$ .

The rest of this chapter is structured as follows. Section 2.2 explains relevant Helmholtz free energy thermodynamic relationships. Section 2.3 describes the general framework and pipeline to generate a physics-embedded machine learning equation of state. Section 2.4 describes how the reference database of volumetric thermophysical properties of the Mie fluid was generated. In Section 2.5,

the digital equation of state for the Mie fluid is generated, and its use and capabilities are discussed. Finally, the main conclusions of this chapter are drawn in Section 2.6.

## 2.2 Helmholtz free energy thermodynamics

Modern equations of state, such as those formulated under the Statistical Fluid Associating Theory (SAFT) [35–37, 130], define the intensive Helmholtz free energy ( $A$ ) as a function of the molar density ( $\rho$ ), and absolute temperature ( $T$ ). The Helmholtz free energy is a convenient state function whose derivatives allow the reconstruction of other thermodynamic functions and relevant thermophysical properties [131]. The intensive entropy ( $S$ ) and pressure ( $P$ ) are related to the first-order derivatives of the Helmholtz free energy, as shown in Eqs. (2.1).

$$S = - \left( \frac{\partial A}{\partial T} \right)_{\rho} \quad (2.1a)$$

$$P = \rho^2 \left( \frac{\partial A}{\partial \rho} \right)_T \quad (2.1b)$$

Similarly, other state functions can be derived using Eqs. (2.1) and standard thermodynamic relationships. For example, the intensive internal energy ( $U$ ), intensive enthalpy ( $H$ ), and intensive Gibbs free energy ( $G$ ) are obtained as follows.

$$U = A + TS = A - T \left( \frac{\partial A}{\partial T} \right)_{\rho} \quad (2.2a)$$

$$H = U + \frac{P}{\rho} = A - T \left( \frac{\partial A}{\partial T} \right)_{\rho} + \rho \left( \frac{\partial A}{\partial \rho} \right)_T \quad (2.2b)$$

$$G = A + \frac{P}{\rho} = A + \rho \left( \frac{\partial A}{\partial \rho} \right)_T \quad (2.2c)$$

For pure systems, the Gibbs free energy, as defined in Eq. (2.2c), is equivalent to the chemical potential ( $\mu$ ).

The second-order derivatives of the Helmholtz free energy are related to experimentally measurable thermophysical properties. Among them are the intensive isochoric heat capacity ( $C_V$ ), the thermal expansion coefficient ( $\alpha_P$ ), the isothermal compressibility ( $\kappa_T$ ), the thermal pressure coefficient ( $\gamma_V$ ), the intensive isobaric heat capacity ( $C_P$ ), and the Joule-Thomson coefficient ( $\mu_{JT}$ ), whose definitions



are given below in Eqs. (2.3).

$$C_V = \left( \frac{\partial U}{\partial T} \right)_\rho = -T \left( \frac{\partial^2 A}{\partial T^2} \right)_\rho \quad (2.3a)$$

$$\alpha_P = \frac{-1}{\rho} \left( \frac{\partial \rho}{\partial T} \right)_P = \left( \frac{\partial P}{\partial T} \right)_\rho / \left[ \rho \left( \frac{\partial P}{\partial \rho} \right)_T \right] \quad (2.3b)$$

$$\kappa_T = \frac{1}{\rho} \left( \frac{\partial \rho}{\partial P} \right)_T = \left[ \rho \left( \frac{\partial P}{\partial \rho} \right)_T \right]^{-1} \quad (2.3c)$$

$$\gamma_V = \left( \frac{\partial P}{\partial T} \right)_\rho = \frac{\alpha_P}{\kappa_T} \quad (2.3d)$$

$$C_P = \left( \frac{\partial H}{\partial T} \right)_P = C_V + \frac{T}{\rho} \frac{\alpha_P^2}{\kappa_T} \quad (2.3e)$$

$$\mu_{JT} = \left( \frac{\partial T}{\partial P} \right)_H = \frac{1}{\rho C_P} [T \alpha_P - 1] \quad (2.3f)$$

As shown above, the Helmholtz free energy is a convenient state function that allows the reconstruction of other state functions and thermophysical properties. It is advantageous to consider the total Helmholtz free energy as the sum of an ideal ( $A^{\text{id}}$ ) and a residual ( $A^{\text{res}}$ ) contribution.

$$A(\rho, T) = A^{\text{id}}(\rho, T) + A^{\text{res}}(\rho, T) \quad (2.4)$$

On the one hand, the ideal contribution is well-defined from statistical thermodynamics [27] and represents the kinetic contribution of non-interacting particles to the free energy. On the other hand, the residual contribution carries the information (contribution) of the intra- and inter-molecular interactions present in the system. In the context of perturbation theory, different contributions to the residual Helmholtz free energy can be considered. Examples of these contributions include the formation of hydrogen bonding [35], dipolar [38], quadrupolar [39] and inductions [132] terms.

Regardless of which non-ideal terms are considered in the residual Helmholtz free energy, these contributions must satisfy some fundamental physical constraints. The residual Helmholtz free energy must vanish in the zero-density (ideal gas) limit. Additionally, the *criteria of thermal and mechanical stability* require the Helmholtz free energy to be concave in temperature and convex with respect to density for stable phases. The criterion of thermal stability states that the internal energy is a monotonically increasing function of temperature (i.e., positive isochoric heat capacity), and the criterion of mechanical stability requires the pressure of stable phases to increase with respect to the density (i.e., positive isothermal compressibility). [116, 125] These last two constraints are naturally satisfied in most theoretical formulations. However, there is no guarantee that these will be fulfilled for data-driven-based models. Therefore, they must be verified or incorporated during the model development process.

Finally, the residual contribution to the Helmholtz free energy leads to important non-ideal behavior of real molecules. For example, virial coefficients, the critical point, and phase equilibria are a direct consequence of these non-ideal interactions.

- Virial coefficients

The second ( $B_2$ ) and third ( $B_3$ ) virial coefficients can be obtained from the residual Helmholtz free energy as follows.

$$B_2 = \lim_{\rho \rightarrow 0} \left( \frac{\partial Z}{\partial \rho} \right)_T = \frac{1}{RT} \left( \frac{\partial A^{\text{res}}}{\partial \rho} \right)_T \bigg|_{\rho=0} \quad (2.5a)$$

$$B_3 = \lim_{\rho \rightarrow 0} \left( \frac{\partial^2 Z}{\partial \rho^2} \right)_T = \frac{1}{RT} \left( \frac{\partial^2 A^{\text{res}}}{\partial \rho^2} \right)_T \bigg|_{\rho=0} \quad (2.5b)$$

In Eqs. (2.5),  $Z = P/(RT\rho)$  is the compressibility factor. From a theoretical perspective, the virial coefficients are relevant due to their roots and interpretation from statistical thermodynamics [27, 133]. The second virial coefficient is related to two-body interactions, and its sign is associated with the predominance of attractive ( $B_2 < 0$ ) or repulsive ( $B_2 > 0$ ) interactions. Similarly, the third virial coefficient is related to interactions between clusters of three particles, which are relevant even for systems interacting through pairwise potentials. The virial coefficients are also relevant to the internal consistency of an equation of state at the low-pressure regime [89]. Finally, the virial coefficients of a wide range of fluids have been measured and reported [12], allowing the comparison of different theories to experimental data.

- Critical point

The critical point of a pure substance is an inflection point in a pressure-density ( $P - \rho$ ) diagram. Mathematically speaking, the critical point is defined as the point where the first- and second-density derivatives of the pressure vanish, as shown below.

$$\left( \frac{\partial P}{\partial \rho} \right)_T = 2\rho \left( \frac{\partial A}{\partial \rho} \right)_T + \rho^2 \left( \frac{\partial^2 A}{\partial \rho^2} \right)_T = 0 \quad (2.6a)$$

$$\left( \frac{\partial^2 P}{\partial \rho^2} \right)_T = 2 \left( \frac{\partial A}{\partial \rho} \right)_T + 4\rho \left( \frac{\partial^2 A}{\partial \rho^2} \right)_T + \rho^2 \left( \frac{\partial^3 A}{\partial \rho^3} \right)_T = 0 \quad (2.6b)$$

The set of Eqs. (2.6) is solved to find the critical density ( $\rho_c$ ) and critical temperature ( $T_c$ ). The critical pressure ( $P_c$ ) is then computed by evaluating Eq. (2.1b) at the found solution.

- Phase equilibria

Multiple phases can coexist at a given temperature and pressure if they minimize the Gibbs free energy of the system [131]. For the case of pure substances, this optimization problem can be reformulated into a system of equations that involve a thermal, mechanical, and chemical equilibrium between two different phases (I and II).

$$T_I = T_{II} \quad (2.7a)$$

$$P_I = P_{II} \quad (2.7b)$$

$$\mu_I = \mu_{II} \quad (2.7c)$$

This system of equations, Eqs. (2.7), is solved for the densities of coexisting phases ( $\rho_I$  and  $\rho_{II}$ ) at a given temperature. In the case of Vapor-Liquid Equilibria (VLE), this temperature must be

subcritical (i.e.,  $T_I = T_{II} < T_c$ ). The pressure and chemical potential are obtained by evaluating Eqs. (2.1b) and (2.2c), respectively. Numerical methods for efficiently solving these equations have been discussed elsewhere [131, 134].

## 2.3 Free Energy Artificial Neural Network based Equation of State (FE-ANN EoS)

As described previously in Section 2.2, the total Helmholtz free energy is a convenient state function to reconstruct other state functions, obtain relevant thermophysical properties, and characterize the phase diagram. Rosenberger et al. [87] proposed a methodology to obtain thermodynamically-consistent Equations of State that model the Helmholtz Free Energy using Artificial Neural Networks (FE-ANN EoS). This thesis follows up on this idea, but it considers extra aspects and constraints to ensure the physicality of the obtained equation of state [3].

A summary of the pipeline for generating equations of state using the FE-ANN EoS is shown in Figure 2.1. The first step, shown in Figure 2.1 (a), requires generating a thermophysical properties database (PVT data) to train the FE-ANN EoS. This database can consist of experimental or synthetic (computer-simulated) data. In this work, the latter approach is preferred as it allows the generation of a potentially extensive database as needed. The database can include first-order properties, such as the pressure ( $P$ ), second virial coefficient ( $B_2$ ), and internal energy ( $U$ ), as well as second-order thermophysical properties, such as the isochoric heat capacity ( $C_V$ ), thermal pressure coefficient ( $\gamma_V$ ), isothermal compressibility ( $\kappa_T$ ), thermal expansion coefficient ( $\alpha_P$ ), isobaric heat capacity ( $C_P$ ), the Joule-Thomson coefficient ( $\mu_{JT}$ ) and third virial coefficient ( $B_3$ ). Details about the database generation will be covered in Section 2.4. In the second step, Figure 2.1 (b), the generated database is used to train the FE-ANN EoS. The FE-ANN EoS explicitly models the residual Helmholtz free energy, whose derivatives are related to the available thermophysical properties as shown in Section 2.2. As will be shown later in Eq. 2.9, the FE-ANN EoS fulfills by definition the ideal gas law [3]. Additionally, as the thermophysical properties are obtained directly from the Helmholtz free energy derivatives, the FE-ANN EoS exactly fulfills Maxwell's relationships [87]. The training also considers extra constraints to ensure the FE-ANN EoS is physically sound. Details about the FE-ANN EoS formulation and training will be covered in Sections 2.3.1 and 2.3.2. In the third and final step, shown in Figure 2.1 (c), the trained FE-ANN EoS can be used to model thermophysical properties. More importantly, due to the physics-inspired formulation, the FE-ANN EoS can be used to predict unseen behavior, such as phase equilibria.

### 2.3.1 FE-ANN EoS formulation

The Helmholtz free energy can be constructed as the sum of an ideal and a residual contribution, see Eq. (2.4). From statistical thermodynamics, the ideal contribution is found to be a function of the logarithm of the density ( $A^{\text{id}} = A^{\text{id}}[\ln \rho]$ ) [27]. This functionality diverges for data points close to the zero density limit ( $\rho \rightarrow 0$ ), which entails a numerical inconvenience when learning the Helmholtz free energy using data-driven models. Nonetheless, this contribution and its derivatives are well-defined by the ideal gas law and can be accounted for analytically. For example, the ideal contribution ( $A^{\text{id}}$ ),

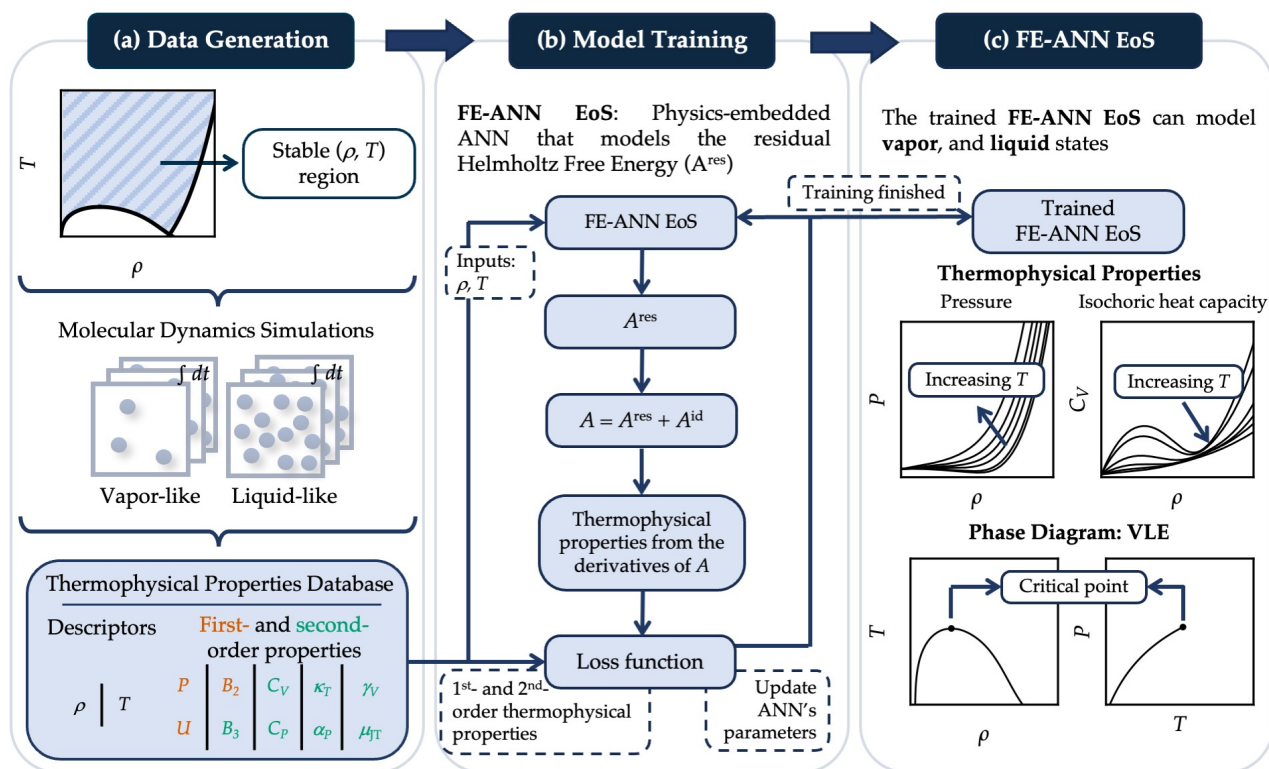


FIGURE 2.1: **Approach to develop a machine-learned equation of state.** The approach consists of three main steps. **(a)** First, molecular dynamics simulations are run to produce a thermophysical properties database. The database can include first-order properties such as the pressure ( $P$ ), second-virial coefficient ( $B_2$ ), and internal energy ( $U$ ), as well as second-order thermophysical properties, such as the isochoric heat capacity ( $C_V$ ), thermal pressure coefficient ( $\gamma_V$ ), isothermal compressibility ( $\kappa_T$ ), thermal expansion coefficient ( $\alpha_P$ ), isobaric heat capacity ( $C_P$ ), the Joule-Thomson coefficient ( $\mu_{JT}$ ) and third virial coefficient ( $B_3$ ). **(b)** Secondly, this database is used to train the FE-ANN EoS, which is an artificial neural network-based equation of state that models the residual Helmholtz free energy ( $A^{res}$ ). The residual Helmholtz free energy is trained based only on its first- and second-order derivatives. **(c)** Finally, the trained FE-ANN EoS can be used to model the thermophysical properties of vapor and liquid states, as well as phase equilibria (VLE).

the ideal pressure ( $P^{\text{id}}$ ), and the ideal internal energy ( $U^{\text{id}}$ ) are computed as follows.

$$A^{\text{id}} = RT [\ln(\Lambda^3 \rho) - 1] \quad (2.8a)$$

$$P^{\text{id}} = \rho RT \quad (2.8b)$$

$$U^{\text{id}} = C_V^{\text{id}} T = \frac{3}{2} RT \quad (2.8c)$$

As shown in Eq. (2.8a), the ideal contribution depends on the ideal gas constant ( $R$ ), temperature ( $T$ ), density ( $\rho$ ) and the de Broglie volume ( $\Lambda^3$ , where  $\Lambda \propto 1/\sqrt{T}$ ). The well-known ideal gas equation, Eq. (2.8b) is recovered by applying Eq. (2.1b) to Eq. (2.8a) (i.e.,  $P = \rho^2(\partial A/\partial \rho)_T$ ). Similarly, the ideal gas internal energy follows the equipartition theorem and is obtained by applying Eq. (2.2a) to Eq. (2.8a) (i.e.,  $U = A - T(\partial A/\partial T)_\rho$ ). Other thermophysical properties from the ideal contribution are obtained by applying the thermodynamic relations described in Eqs. (2.1), (2.2) and (2.3).

In the context of a data-driven EoS, one option is to let the model "learn" the ideal gas limit, Eqs. (2.8). However, even if the model gets close to this limit, it is not guaranteed to reach it exactly. There is no added benefit to re-discover the ideal gas equation. For this reason, in this thesis, only the residual Helmholtz free energy ( $A^{\text{res}}$ ) is modeled using a data-driven model. The following two formulations that analytically account for the appropriate low-density limit are considered.

$$A^{\text{res}} = \underbrace{\text{ANN}(\chi, \rho, T) - \text{ANN}(\chi, \rho = 0, T)}_{\text{Model to be trained}} \quad (2.9a)$$

or,

$$A^{\text{res}} = \underbrace{\text{ANN}(\chi, \rho, \frac{1}{T}) - \text{ANN}(\chi, \rho = 0, \frac{1}{T})}_{\text{Model to be trained}} \quad (2.9b)$$

In Eqs. (2.9),  $A^{\text{res}}$  is the residual Helmholtz free energy,  $\rho$  is the density,  $T$  is the temperature, and  $\chi$  are the descriptors of the molecule (or pseudo-molecule) of interest, e.g., the molecular parameters of an interaction potential. Finally, ANN refers to an artificial neural network, which acts as universal approximators [53] to learn the residual free energy surface. Mathematically speaking, an  $N$ -layered ANN consists of a composition of  $N$  linear transformations followed by a non-linear transformation, a.k.a. activation function, as shown below.

$$\mathbf{y} = \text{ANN}(\mathbf{x}) = f_N(\mathbf{W}_N \dots f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \dots + \mathbf{b}_N) \quad (2.10)$$

In Eq. (2.10),  $\mathbf{x}$  and  $\mathbf{y}$  are the input and output features of the ANN, respectively. In this equation, each layer considers a linear transformation  $\mathbf{W}_i \mathbf{x}_{i-1} + \mathbf{b}_i$ , in which  $\mathbf{W}_i$  is the weight matrix and  $\mathbf{b}_i$  is the bias vector of the  $i$ -th layer of the ANN. The weight matrix is a  $(m \times n)$  dimensional matrix, where  $m$  is the number of neurons in the layers, and  $n$  is the dimension of the input vector. The bias vector is of dimension  $m$ . The linear transformation is followed by an activation function,  $f_i$ . The role of the activation function in each layer is considerable. This function adds non-linearity to the ANN, providing flexibility to model complex data behavior [135]. Further details about ANNs can be found in Appendix A.

The proposed residual Helmholtz free energy formulations evaluate the same artificial neural network

at two different densities: the given density  $\rho$  and at  $\rho = 0$ . This procedure forces a zero residual Helmholtz energy at the zero density limit, fulfilling the ideal gas limit. As shown by Rosenberger et al. [87], Maxwell's relationships are fulfilled by default by modeling the Helmholtz free energy in this fashion. These two physical constraints are rooted in the FE-ANN EoS formulation and are fulfilled even before training the model.

The two FE-ANN EoS formulations shown in Eqs. (2.9) differ in how the ANN interprets the temperature. The linear temperature scale, as shown in Eq. (2.9a), is a natural interpretable variable for humans. However, there are two main reasons for using the inverse temperature, as proposed in Eq. (2.9b). The first reason is based on a theoretical justification. In statistical thermodynamics, the partition function, which is directly related to the free energy of a system, is directly related to the temperature in the form of a Boltzmann's factor ( $\beta = 1/k_b T$ ) [27, 66]. This dependence implies that the Helmholtz free energy can be considered a natural function of the inverse of the temperature rather than the temperature itself. This functionality is also seen in high-energy perturbation theory expansion, such as the one seen in the development of molecular-inspired equations of state, i.e. SAFT-models [35–37, 130], as well as in technical equations of state [16, 115, 136]. The second operational reason for using the temperature inverse is a numerical convenience. As shown later in Figure 2.3 (b), higher temperature points move closer when inverting the linear temperature scale. This numerical convenience eases the interpolating efforts of the machine learning model.

### 2.3.2 FE-ANN EoS training

The FE-ANN EoS, Eq. (2.9), considers several trainable parameters (i.e., weights and biases) that can be optimized to model thermophysical properties. In principle, this EoS can be trained using either experimental or simulated data. For practical reasons, this contribution focuses on training an FE-ANN EoS with computer-generated data, as this approach allows data production as needed.

The FE-ANN EoS explicitly models the Helmholtz free energy. However, it is recognized that this free energy data cannot be easily obtained from computational simulation methods and that non-trivial methodologies such as particle insertions and umbrella sampling [70] are needed to compute it. Instead, thermophysical properties related to the derivatives of the Helmholtz free energy are accessible from computational methods. It is suggested to train the FE-ANN EoS using first- and second-order derivative properties. First-order derivative properties are related to important phenomena like phase equilibria, while second-order derivative properties are related to the thermodynamic consistency of the model. The training data then can include thermophysical properties, such as pressure ( $P$ ), internal energy ( $U$ ), isochoric heat capacity ( $C_V$ ), thermal pressure coefficient ( $\gamma_V$ ), isothermal compressibility ( $\kappa_T$ ), thermal expansion coefficient ( $\alpha_P$ ), the isobaric heat capacity ( $C_P$ ) and Joule-Thomson coefficient ( $\mu_{JT}$ ) which can be directly obtained from molecular simulations [66, 70]. Additionally, the training data set can include the second ( $B_2$ ) and third ( $B_3$ ) virial coefficients that can be obtained by numerical integration given an interaction potential [27]. These properties are computed from the FE-ANN EoS using the definitions described in Section 2.2, see Eqs. (2.1), (2.2), (2.3), and (2.5).

There are some numerical aspects to be considered when performing this multi-objective training. First, the pressure of the system can vary several orders of magnitude depending on the density and

temperature. This behavior would imply that an optimizer based on the mean squared error (MSE) would be biased towards high-pressure data points. Therefore, often low-pressure (vapor-like) regions would be misrepresented. Using the compressibility factor ( $Z = P/\rho RT$ ) as a pressure indicator is more numerically convenient, as this better behaves for both low and high-density regions [58]. Based on the same argument, the isothermal compressibility ( $\kappa_T$ ) is orders of magnitude higher for low-density data points than for high-density data points. For this reason, the product  $\rho\kappa_T$  is preferred for training instead. Finally, the isobaric heat capacity can change considerably near unstable phases. For this reason, the adiabatic index ( $\gamma = C_P/C_V$ ) is more convenient for training purposes.

The parameters of the FE-ANN EoS are trained to minimize a loss function ( $\mathcal{L}$ ) that considers two contributions, as shown below.

$$\mathcal{L} = \mathcal{L}^{\text{MSE}} + \mathcal{L}^{\text{pen}} \quad (2.11)$$

The first contribution,  $\mathcal{L}^{\text{MSE}}$ , corresponds to a weighted Mean Squared Error (MSE) loss with respect to different target thermophysical properties.

$$\mathcal{L}^{\text{MSE}} = \sum_i \frac{w_i}{\sigma_i^2 N_{P,i}} \sum_j^{N_{P,i}} \left( y_{i,j}^{\text{pred}} - y_{i,j}^{\text{true}} \right)^2 \quad (2.12)$$

In Eq. (2.12),  $\sigma_i^2$ ,  $w_i$ , and  $N_{P,i}$  are the variance, relative weight, and number of data points of the thermophysical property "i". The  $w_i/\sigma_i^2$  ratio allows defining the practical relative importance of the property "i" in the loss function. The superscript "pred" and "true" refer to the value predicted by the FE-ANN EoS and the value available from a database, respectively. This loss function can consider first- and second-order derivative properties as discussed above.

The second contribution to the loss function,  $\mathcal{L}^{\text{pen}}$ , is a penalty function to impose a correct and consistent thermodynamic behavior.

$$\mathcal{L}^{\text{pen}} = \mathcal{L}^{\text{pen,I}} + \mathcal{L}^{\text{pen,II}} + \mathcal{L}^{\text{pen,III}} \quad (2.13a)$$

$$\mathcal{L}^{\text{pen,I}} = \frac{w_{\text{I}}}{\sigma_{C_V}^2 N_{P,\text{I}}} \sum_j^{N_{P,\text{I}}} \min \left[ 0, C_{V,j}^{\text{pred}} \right]^2 \quad (2.13b)$$

$$\mathcal{L}^{\text{pen,II}} = \frac{w_{\text{II}}}{\sigma_{\kappa_T}^2 N_{P,\text{II}}} \sum_j^{N_{P,\text{II}}} \min \left[ 0, \rho \kappa_{T,j}^{\text{pred}} \right]^2 \quad (2.13c)$$

$$\mathcal{L}^{\text{pen,III}} = \frac{w_{\text{III}}}{\sigma_{\gamma_V}^2 N_{P,\text{III}}} \sum_j^{N_{P,\text{III}}} \min \left[ 0, \gamma_{V,j}^{\text{pred}} \right]^2 \quad (2.13d)$$

The penalty loss function, Eq. (2.13a), comprises different physically-inspired penalties.  $\mathcal{L}^{\text{pen,I}}$  and  $\mathcal{L}^{\text{pen,II}}$ , Eqs. (2.13b) and (2.13c), preclude the prediction of negative isochoric heat capacities and isothermal compressibility, as required by the *criteria of thermal and mechanical stability* discussed in Section 2.2. Similarly,  $\mathcal{L}^{\text{pen,III}}$ , Eq. (2.13d), prevents undesired crossing pressure isotherms. These penalty functions only contribute to the loss function when the FE-ANN EoS predicts an unphysical value. Similar to the MSE loss function, Eq. (2.12), the penalty functions are weighted by the ratio of a relative importance ( $w_i$ ) and the variance of the property ( $\sigma_i^2$ ). Optionally, these constraints can also be

artificially added in regions where no data is available, for example, Eqs. (2.13b) and (2.13d) could be applied in the region inside the binodal curve. These additional constraints could be useful for enforcing thermodynamically correct behavior. Lastly, it must be noted that these penalty functions are soft constraints, and their value must be checked at the end of the training stage.

The trainable parameters of the FE-ANN EoS are updated to minimize the total loss function  $\mathcal{L}$ .

$$\underset{\mathbf{w}_i, \mathbf{b}_i}{\text{minimize}} \mathcal{L} = \mathcal{L}^{\text{MSE}} + \mathcal{L}^{\text{pen,I}} + \mathcal{L}^{\text{pen,II}} + \mathcal{L}^{\text{pen,III}} \quad (2.14)$$

In principle, this minimization problem could be solved by any optimization method. However, gradient descent methods are preferred due to the high dimensionality of the problem and the elevated computational cost associated with higher-order derivatives. A gradient descent step is shown below.

$$\theta_{j+1} = \theta_j - \alpha \frac{\partial \mathcal{L}}{\partial \theta_j} \quad (2.15)$$

In Eq. (2.15),  $\theta_i$  refers to a parameter of the ANN (i.e., weight or bias) at step  $j$ , and  $\alpha$  is the learning rate. The derivatives of the loss function with respect to these parameters ( $\partial \mathcal{L} / \partial \theta_i$ ) are obtained using backpropagation [88]. This method computes derivatives at computer precision, and it is available in modern machine learning libraries, such as TensorFlow [137] and JAX [138]. The minimization step has been modified to include information about previous iterations (a.k.a. momentum). The Adam [139] optimizer is an example of an optimization algorithm with momentum. Finally, due to constraints set by the computer memory, the training of the ANN is usually performed with mini-batches of the data. The number of epochs is the number of times the optimizer passes through the entire data set. The batch size and the number of epochs are training hyperparameters of the model.

### 2.3.3 FE-ANN EoS hyperparameters

As mentioned throughout this section, the performance of the FE-ANN EoS is highly dependent on several hyperparameters. They include the architecture of the ANN (i.e., the number of layers and number of neurons per layer) and the activation function used in each layer. Additionally, the training hyperparameters include the learning parameter, the batch size, the number of epochs, the learning rate, and the optimizer algorithm. These hyperparameters must be tuned. This task is not trivial and usually relies on black-box optimization methods. For a complete discussion on hyperparameter optimization, the reader is referred to the review of Yang and Shami [140].

In the case of the FE-ANN EoS, there are some recommendations to mitigate the complexity of the hyperparameter optimization. The FE-ANN EoS must be differentiable. Hence, it is recommended to use a differentiable activation function, such as the tanh activation function ( $f(x) = \tanh(x)$ ), for the hidden layers. For simplicity, all hidden layers can use the same number of neurons. Finally, the output layer does not require a bias and must return a single real value. The linear activation function ( $f(x) = x$ ) is recommended for this layer. The batch size is constrained by computer memory, and it is recommended to use a power of two (e.g., 32, 64, or 128). The learning rate is the most relevant hyperparameter during training. It is recommended to use the largest value that leads to a numerically stable optimization. It is customary to use a value of the power of ten (e.g.,  $10^{-3}$ ,  $10^{-4}$ , or



$10^{-5}$ ). Finally, the number of training epochs must be selected such that the loss function reaches a plateau, indicating that further training does not significantly improve the model performance.

## 2.4 Mie fluid thermophysical properties database

The FE-ANN EoS must be trained using a thermophysical properties database. This section focuses on computing a thermophysical properties database of fluid systems interacting through a truncated Mie potential, Eq. (1.1). This database is used to exemplify the use of the FE-ANN EoS framework in Section 2.5.

The database presented here reports thermophysical properties in reduced units, which is equivalent to setting to unity the Boltzmann constant ( $k_B$ ), the monomer mass ( $M$ ), the potential well-depth ( $\epsilon$ ), and the characteristic size parameter ( $\sigma$ ). It must be noted that when using reduced units, the Mie fluid is fully characterized by the repulsive and attractive exponents. The definitions of some relevant thermophysical properties in reduced units are given in Table 2.1. Thermophysical properties in reduced units are indicated by a "\*" superscript.

TABLE 2.1: Definition of physical quantities in reduced units.

| Quantity                      | Symbol            | Units                             | Dimensionless Definition <sup>a, b, c</sup>                |
|-------------------------------|-------------------|-----------------------------------|------------------------------------------------------------|
| Time                          | $\tau$            | s                                 | $\tau^* = \tau / \sigma \sqrt{M / N_{\text{Av}} \epsilon}$ |
| Distance                      | $r$               | m                                 | $r^* = r / \sigma$                                         |
| Mass                          | $m$               | $\text{kg mol}^{-1}$              | $m^* = m / M$                                              |
| Density                       | $\rho$            | $\text{mol m}^{-3}$               | $\rho^* = \rho N_{\text{Av}} \sigma^3$                     |
| Temperature                   | $T$               | K                                 | $T^* = T k_B / \epsilon$                                   |
| Pressure                      | $P$               | Pa                                | $P^* = P \sigma^3 / \epsilon$                              |
| Entropy                       | $S$               | $\text{J K}^{-1} \text{mol}^{-1}$ | $S^* = S / N_{\text{Av}} k_B$                              |
| Helmholtz free energy         | $A$               | $\text{J mol}^{-1}$               | $A^* = A / N_{\text{Av}} \epsilon$                         |
| Internal energy               | $U$               | $\text{J mol}^{-1}$               | $U^* = U / N_{\text{Av}} \epsilon$                         |
| Enthalpy                      | $H$               | $\text{J mol}^{-1}$               | $H^* = H / N_{\text{Av}} \epsilon$                         |
| Gibbs free energy             | $G$               | $\text{J mol}^{-1}$               | $G^* = G / N_{\text{Av}} \epsilon$                         |
| Isochoric heat capacity       | $C_V$             | $\text{J K}^{-1} \text{mol}^{-1}$ | $C_V^* = C_V / N_{\text{Av}} k_B$                          |
| Thermal expansion coefficient | $\alpha_P$        | $\text{K}^{-1}$                   | $\alpha_P^* = \alpha_P \epsilon / k_B$                     |
| Isothermal compressibility    | $\kappa_T$        | $\text{Pa}^{-1}$                  | $\kappa_T^* = \kappa_T \epsilon / \sigma^3$                |
| Thermal pressure coefficient  | $\gamma_V$        | $\text{Pa K}^{-1}$                | $\gamma_V^* = \gamma_V \sigma^3 / k_B$                     |
| Isobaric heat capacity        | $C_P$             | $\text{J K}^{-1} \text{mol}^{-1}$ | $C_P^* = C_P / N_{\text{Av}} k_B$                          |
| Joule-Thomson coefficient     | $\mu_{\text{JT}}$ | $\text{K Pa}^{-1}$                | $\mu_{\text{JT}}^* = \mu_{\text{JT}} k_B / \sigma^3$       |
| Second virial coefficient     | $B_2$             | $\text{m}^3 \text{mol}^{-1}$      | $B_2^* = B_2 / N_{\text{Av}} \sigma^3$                     |
| Third virial coefficient      | $B_3$             | $\text{m}^6 \text{mol}^{-2}$      | $B_3^* = B_3 / N_{\text{Av}}^2 \sigma^6$                   |

<sup>a</sup>Superscript "\*" refers to a reduced property. <sup>b</sup> $N_{\text{Av}} = 6.02214076 \cdot 10^{23} \text{ mol}^{-1}$  is the Avogadro constant [141],  $k_B = 1.380649 \cdot 10^{-23} \text{ J K}^{-1}$  is the Boltzmann constant [141] and  $M$  is the molar mass of the Mie particle. <sup>c</sup> $\sigma$  is the shape parameter in meters, and  $\epsilon$  is the potential well depth of the Mie potential in Joules. See Eq. (1.1).

### 2.4.1 Thermophysical properties computation

The thermophysical properties database considers Mie fluids with a repulsive exponent ( $\lambda_r$ ) ranging from 7 to 34. The attractive exponent ( $\lambda_a$ ) is set to 6 to account for London dispersion forces [93]. The database consists of computer-generated thermophysical properties data. The database constructed in this thesis includes virial coefficients obtained by numerical integration [27] and first- and second-order derivatives obtained using high-throughput Molecular Dynamics (MD) simulations.

The second and third virial coefficients are obtained numerically for systems interacting through a pairwise interaction potential. This calculation relies on the theoretical-based definition of the virial coefficients from cluster integrals [121] as shown below.

$$B_2^* = -\frac{1}{2} \int f(r_{12}^*) d\mathbf{r}_2^* \quad (2.16a)$$

$$B_3^* = -\frac{1}{3} \int f(r_{12}^*) f(r_{13}^*) f(r_{23}^*) d\mathbf{r}_2^* d\mathbf{r}_3^* \quad (2.16b)$$

Where,  $f(r_{ij}^*) = \left[ \exp \left( -\mathcal{U}^*(r_{ij}^*)/T^* \right) - 1 \right]$  is the Mayer function between particles  $i$  and  $j$  given the reduced interaction potential ( $\mathcal{U}^* = \mathcal{U}/\epsilon$ ) and  $d\mathbf{r}_i^*$  is the volume element where particle  $i$  is located.

These two integrals, Eqs. (2.16), can be manipulated and simplified for systems interacting through spherically symmetric potentials, such as the Mie potential. On the one hand, the third virial coefficient integral, Eq. (2.16b), can be solved using bipolar coordinates [142] or Fast-Fourier-Transforms [143]. On the other hand, the second virial coefficient, Eq. (2.16a), can be simplified into Eq. (2.17) [27].

$$B_2^* = -2\pi \int_0^\infty f(r^*) r^{*2} dr^* \quad (2.17)$$

This integral, Eq. (2.17), has been solved analytically for simple systems, such as the hard-sphere [27], square-well [144], and Mie [145] potentials. Alternatively, it can be solved numerically for other interaction potentials.

In this work, the virial coefficients of systems interacting via the Mie potential are obtained numerically for reduced temperatures ( $T^*$ ) ranging from 0.65 to 39.62. The second virial coefficient is obtained by numerically integrating Eq. (2.17) using the quad function from the `scipy.integrate` Python library [146]. The third virial coefficient is efficiently computed using Fast-Fourier-Transforms as suggested by Schultz [143]. The second and third virial coefficients are computed to meet a  $10^{-8}$  numerical tolerance error.

Other first-order ( $P^*$  and  $U^*$ ) and second-order ( $C_V^*$ ,  $\gamma_V^*$ ,  $C_V^*$ ,  $\alpha_P^*$ , and  $\kappa_T^*$ ) derivative properties of the Mie fluid are generated using high-throughput Molecular Dynamics (MD) simulations. The fundamentals of molecular simulations have been thoroughly covered in the books of Allen and Tildesley [66] and Frenkel and Smit [70]. Therefore, details about this methodology are not covered in this thesis. All the MD simulations performed in this chapter use the open-source Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [147] (stable release June 2022) code.

The database generation requires knowledge (or intuition) of the phase boundaries (i.e., vapor-liquid equilibria and solid-liquid equilibria) to avoid running simulations at unstable conditions that will

likely fail. Vapor-Liquid Equilibrium (VLE) is computed using the temperature quench method [148] following the guidelines of Müller et al. [149]. These simulations are conducted entirely in the canonical ensemble (NVT) using a time step of  $\Delta\tau^* = 0.002\tau^*$  and the Nosé-Hoover thermostat [150] with a time constant of  $100\Delta\tau^*$ . The non-bonded interactions are cut off at  $r_c = 5\sigma$  radius, and no long-range corrections are applied. These simulations consist of 6000 Mie beads distributed in an elongated simulation box ( $L_x = L_z$  and  $L_z = 6L_x$ ). Initially, the simulation is quenched at high temperature for  $10^5$  time steps. The simulation is then equilibrated for  $5 \cdot 10^6$  time steps at a desired coexistence temperature. Finally, the production stage integrates the simulation at the same coexistence temperature for further  $10^7$  time steps. During the production stage, the density and pressure profiles are accumulated every 1000 time steps. The equilibrium densities are obtained by averaging the density profile, and the equilibrium pressure is averaged from the normal component of the pressure tensor ( $P_{zz}^*$ ). Further details about this methodology are available elsewhere [149]. For the Solid-Liquid Equilibria (SLE) boundaries, Khrapak et al. [151] have correlated the freezing line of  $(\lambda_r, \lambda_a = 6)$  Mie particles, as shown below.

$$T_{\text{SLE}}^* = L(\lambda_r) \left( \frac{\lambda_r}{\lambda_r - 6} \right) \left( \frac{\lambda_r}{6} \right)^{6/(\lambda_r - 6)} \left[ \lambda_r(\lambda_r + 1)\rho_1^{*\lambda_r/3} - 42\rho_1^{*2} \right] \quad (2.18)$$

$$L(\lambda_r) = 212 + 15\lambda_r \exp(-10/\lambda_r)$$

This correlation allows estimating the freezing temperature ( $T_{\text{SLE}}^*$ ) given the equilibrium liquid density ( $\rho_1^*$ ).

Knowing the phase boundaries allows to set up molecular simulations for stable single-phase conditions. The simulated phase space covers reduced densities ( $\rho^*$ ) ranging from  $10^{-6}$  to 1.3 and reduced temperatures ( $T^*$ ) from 0.6 to 10.0. The thermophysical properties of each state point are obtained from a single simulation replica. The simulations are performed in the canonical (NVT) and isobaric-isothermal (NPT) ensembles to produce thermophysical properties data. The equilibration and production times of each simulation were chosen to produce converged thermophysical properties for most explored conditions of the phase space. Tentatively incorrect simulation results are identified and disregarded, as explained in Section 2.4.2.

The pressure ( $P^*$ ), internal energy ( $U^*$ ), isochoric heat capacity ( $C_V^*$ ) and thermal pressure coefficient ( $\gamma_V^*$ ) are obtained from molecular dynamics simulations in the canonical (NVT) ensemble. The simulation setup consists of a cubic simulation box with 4096 Mie beads. This relatively high number of beads is chosen to minimize any finite-size effects on the results. All the simulations consider a cutoff radius of  $r_c = 5\sigma$  with no long-range corrections for pairwise interactions. The simulations are integrated using a timestep of  $\Delta\tau^* = 0.002\tau^*$ . Firstly, the simulations are run for  $3.2 \cdot 10^6$  timesteps in the NVT ensemble using the Nosé-Hoover thermostat [150] with a time constant equal to  $100\Delta\tau^*$ . The statistics of the system are accumulated every 1000 time steps. The first  $2 \cdot 10^5$  timesteps are disregarded to ensure equilibration of the system. The pressure of the system ( $P^*$ ) is computed directly from virial expression [152] whereas its internal energy ( $U^*$ ) is directly obtained from the averaged Hamiltonian. Second-order derivative properties, such as the isochoric heat capacity ( $C_V^*$ ) and thermal

pressure coefficient ( $\gamma_V^*$ ), are computed in the NVT ensemble as follows [66].

$$C_V^* = \frac{1}{NT^{*2}} \langle [\mathcal{V}^* - \langle \mathcal{V}^* \rangle]^2 \rangle_{NVT} + \frac{3}{2} \quad (2.19a)$$

$$\gamma_V^* = \frac{1}{T^{*2}} \langle [\mathcal{V}^* - \langle \mathcal{V}^* \rangle] [\mathcal{P}^* - \langle \mathcal{P}^* \rangle] \rangle_{NVT} + \rho^* \quad (2.19b)$$

In Eqs. (2.19),  $\langle \dots \rangle$  corresponds to the ensemble average,  $\mathcal{V}^*$  refers to the extensive potential energy,  $\mathcal{P}^*$  is the instantaneous pressure and  $N$  is the number of particles in the system.

After the production stage in the canonical ensemble is finished, the system is run for further  $3.2 \cdot 10^6$  timesteps in the isothermal-isobaric (NPT) ensemble using the Nosé-Hoover thermostat-barostat [153]. A time constant of  $100\Delta\tau^*$  and  $1000\Delta\tau^*$  are used for the thermostat and barostat, respectively. Similarly, as for the NVT simulation stage, statistics of the system are accumulated every 1000 time steps, and the first  $2 \cdot 10^5$  time steps are ignored to ensure averages over an equilibrated system. The isobaric heat capacity ( $C_P^*$ ), the thermal expansion coefficient ( $\alpha_P^*$ ), and the isothermal compressibility ( $\kappa_T^*$ ) are then computed by using the fluctuations of the system in the NPT ensemble [66].

$$C_P^* = \frac{1}{NT^{*2}} \langle [\mathcal{H}^* - \langle \mathcal{H}^* \rangle]^2 \rangle_{NPT} \quad (2.20a)$$

$$\alpha_P^* = \frac{1}{\langle V^* \rangle T^{*2}} \langle [V^* - \langle V^* \rangle] [\mathcal{H}^* - \langle \mathcal{H}^* \rangle] \rangle_{NPT} \quad (2.20b)$$

$$\kappa_T^* = \frac{1}{\langle V^* \rangle T^*} \langle [V^* - \langle V^* \rangle]^2 \rangle_{NPT} \quad (2.20c)$$

Here, in Eqs. (2.20),  $\mathcal{H}^*$  refers to the extensive instantaneous enthalpy at constant pressure, and  $V^*$  is the instantaneous system volume. Some other thermophysical properties are obtained from the NPT simulation using the definitions given in Eqs. (2.3). For example, the isochoric heat capacity can be obtained by using Eq. (2.3e), the thermal pressure coefficient is directly computed from Eq. (2.3d), and the Joule-Thomson coefficient is obtained using Eq. (2.3f).

In summary, the following thermophysical properties are computed for a single data point in molecular dynamics simulation. First-order derivative properties (i.e.,  $P^*$  and  $U^*$ ) are directly averaged from simulations in the NVT ensemble. Similarly, second-order properties ( $C_V^*$ ,  $\gamma_V^*$ ,  $C_P^*$ ,  $\alpha_P^*$ , and  $\kappa_T^*$ ) are computed using the sampled temperature, pressure, potential energy and system volume alongside Eqs. (2.19) and (2.20). Finally, the Joule-Thomson coefficient is obtained using Eq. (2.3f). The second-order properties are computed at increasing time intervals during the last  $1/3$  of the sampled simulation data, and their average value is reported. The deviations of the computed thermophysical properties are also computed. The standard deviations of the pressure and internal energy are directly computed from the sampled data. In contrast, the standard deviation of second-order properties is computed from the sampled values at different time intervals.

#### 2.4.2 Thermophysical properties database validation

The correctness of the molecular simulation results is checked by four sanity tests. If a computed thermophysical property of a data point fails to pass any of these tests, that value is disregarded.

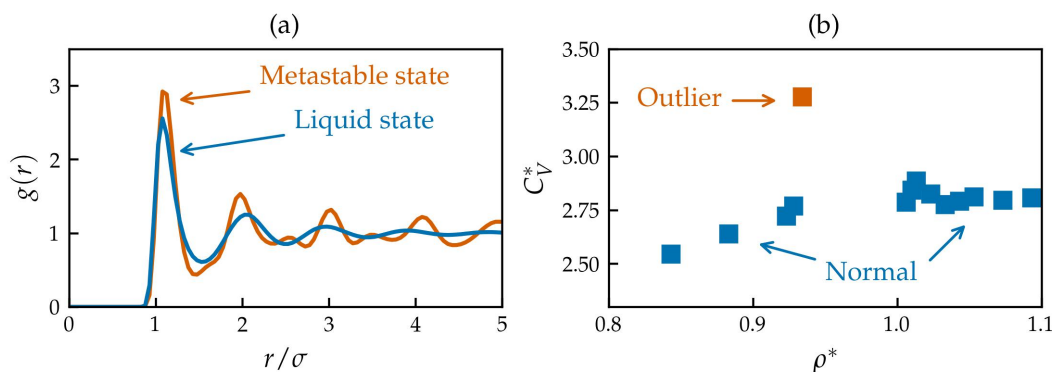


FIGURE 2.2: **Sanity checks for computed thermophysical properties.** (a) Radial distribution function ( $g(r)$ ) test. Simulations in the liquid state have a well-defined  $g(r)$  as shown by the blue solid line. If the liquid simulation entered a glassy (metastable) state, the  $g(r)$  function will have extra peaks as shown by the orange solid line. Data points in a metastable regime are disregarded. (b) Outlier detection test. Second-order derivative properties are obtained by analysis of fluctuations of a molecular simulation, requiring a long simulation to converge properly. Possible outliers are automatically detected using the LocalOutlierFactor method [155] implemented in Scikit-Learn [156]. For example, blue data points are classified as normal, and the orange data point is identified as an outlier by this algorithm. Outliers are disregarded from the database.

The first sanity test verifies the consistency between the results obtained from different ensembles. For example, the density, pressure, and internal energy averages from the NVT and NPT ensembles can be compared. The averaged values from each ensemble must be similar, given a tolerance. A data point passes this sanity test if the averaged value in one ensemble falls within the 95% confidence interval of the other ensemble.

The second sanity test checks the deviations of the computed thermophysical properties. Thermophysical properties with high deviations are a sign of simulation that is not correctly equilibrated. This problem is particularly relevant to second-order properties, which are computed from the fluctuations of the simulation results. A thermophysical property passes this test if the relative standard deviation ( $100|\sigma_x / \langle x \rangle|$ ) is below 5%.

The third sanity test checks if a fluid data point entered a glassy (metastable) state. This undesired behavior can occur in the vicinity of the SLE boundaries. The structure of the simulation is checked by the nature of the radial distribution function (RDF,  $g(r)$ ) [66]. As shown in Figure 2.2 (a), the RDF of a glassy data point (orange line) exhibits a higher number of peaks and some periodicity when compared to a data point in a liquid state (blue line). A fluid data point passes this test if it has four or fewer local minima in its RDF. This test was automated by analyzing the computed RDF using standard numpy [154] functions.

The fourth and final sanity test searches for possible outliers in the database. As exemplified in Figure 2.2 (b), some computed thermophysical properties might not have converged properly in the simulation. These outliers are automatically detected using the LocalOutlierFactor method [155] implemented in Scikit-Learn [156]. This outlier detection method provides scores that measure the "outlierness" of a data point. A regular data point is expected to have a score close to 1, while an outlier will have a "big" score. In the original publication, Breunig et al. [155] suggested a value of 1.5 as a threshold to identify outliers. However, this value might be unsuitable for some data sets. For this reason, the threshold for filtering outliers is determined by fitting the total population of scores

to a probability distribution. As most of the data points are normal, the mean of the distribution is close to 1. Then, the threshold is obtained from the upper bound of the 95% confidence interval of the distribution [157]. An adequate distribution that fits the scores is found using the `fitter` [158] Python package.

### 2.4.3 Thermophysical properties database description

The Mie fluid thermophysical properties database considers systems with repulsive exponent  $\lambda_r \in [7, 34]$  and  $\lambda_a = 6$ . The database considers VLE simulation results and PVT properties. The PVT database includes first-order ( $P^*$ ,  $U^*$  and  $B_2^*$ ) and second-order ( $C_V^*$ ,  $\gamma_V^*$ ,  $C_P^*$ ,  $\alpha_P^*$ ,  $\kappa_T^*$ ,  $\mu_{JT}^*$  and  $B_3^*$ ) derivative properties. As indicated by the "\*" superscript, these thermophysical properties are reported in reduced units (see Table 2.1). A complete database description, including the ranges and number of data points for each thermophysical property, is given in Table 2.2. A visual representation of the

TABLE 2.2: Mie fluid database description.<sup>a</sup>

| Descriptor                             | Symbol         | # of data points | Min value | Max value |
|----------------------------------------|----------------|------------------|-----------|-----------|
| Repulsive exponent <sup>b</sup>        | $\lambda_r$    | —                | 7         | 34        |
| Attractive exponent <sup>b</sup>       | $\lambda_a$    | —                | 6         | 6         |
| van der Waals parameter <sup>c</sup>   | $\alpha_{vdw}$ | —                | 0.53      | 1.47      |
| Density                                | $\rho^*$       | —                | 1.00e-06  | 1.31      |
| Temperature <sup>d</sup>               | $T^*$          | —                | 0.60      | 10.06     |
| Thermophysical Property                |                |                  |           |           |
| Pressure                               | $P^*$          | 21679            | 7.80e-07  | 1.60e+02  |
| Compressibility factor <sup>e</sup>    | $Z$            | 21679            | 1.15e-04  | 13.18     |
| Internal energy                        | $U^*$          | 21679            | -7.82     | 22.70     |
| Second virial coefficient <sup>f</sup> | $B_2^*$        | 3485             | -9.91     | 1.57      |
| Third virial coefficient <sup>f</sup>  | $B_3^*$        | 3462             | -3.61     | 2.60      |
| Isochoric heat capacity                | $C_V^*$        | 18435            | 1.50      | 3.16      |
| Isobaric heat capacity                 | $C_P^*$        | 16427            | 2.50      | 23.98     |
| Adiabatic index <sup>g</sup>           | $\gamma$       | 16235            | 1.21      | 9.50      |
| Thermal pressure coefficient           | $\gamma_V^*$   | 18483            | 1.00e-06  | 11.31     |
| Thermal expansion coefficient          | $\alpha_P^*$   | 17807            | 1.52e-02  | 8.52      |
| Isothermal compressibility             | $\kappa_T^*$   | 17775            | 1.46e-03  | 1.31e+06  |
| Joule-Thomson coefficient              | $\mu_{JT}^*$   | 11972            | -0.50     | 11.85     |

<sup>a</sup> Descriptors and thermophysical properties are reported in reduced units. See Table 2.1. <sup>b</sup> Repulsive and attractive exponent of the Mie potential, Eq. (1.1). <sup>c</sup> van der Waals parameter is a descriptor of a Mie particle. See Eq. (2.21). <sup>d</sup> Temperature range used in molecular dynamics simulations. <sup>e</sup> Compressibility factor,  $Z = P^*/\rho^*T^*$ . <sup>f</sup> Virial coefficients obtained in the temperature range of 0.65 to 39.62. <sup>g</sup> Adiabatic index (heat capacity ratio),  $\gamma = C_P^*/C_V^*$ .

distribution of this database is provided in Figure 2.3. As shown in Figure 2.3 (a), most data points are in the range of temperatures up to  $T^* = 2.0$ . This region was purposely saturated, as the Helmholtz free energy is expected to change the most due to the presence of phase instability and the transition from subcritical to supercritical behavior. Higher temperatures up to  $T^* = 10.0$  were included sparsely. The density distribution of the database is even except in two regions. The database has fewer data

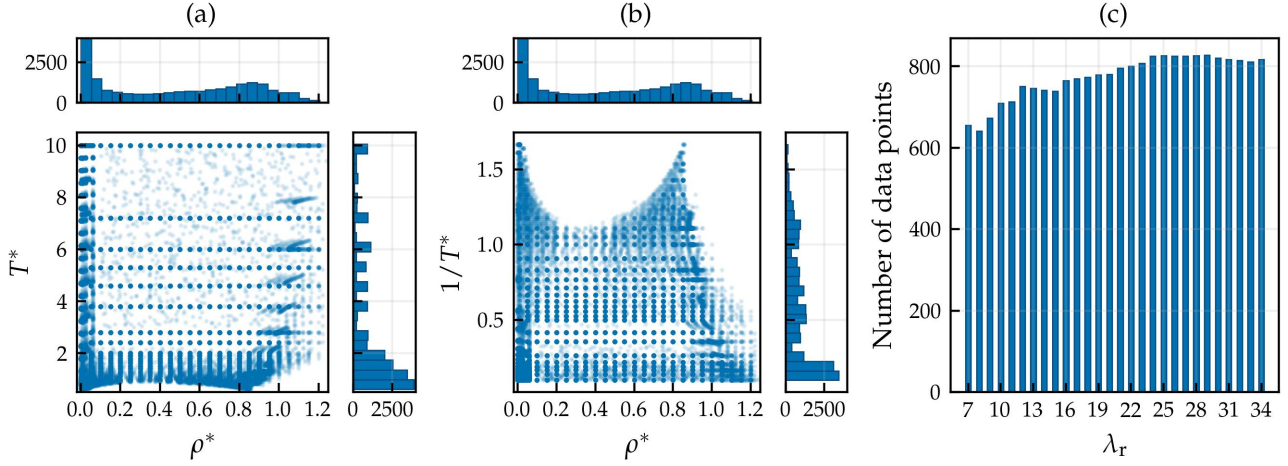


FIGURE 2.3: **Mie fluid Thermophysical properties database distribution** (a) Density-Temperature ( $\rho^* - T^*$ ) distribution. (b) Density-Temperature inverse ( $\rho^* - 1/T^*$ ) distribution. (c) Repulsive exponents ( $\lambda_r$ ) distribution.

points at high densities and low temperatures because the Mie monomer is likely to freeze under these conditions. Contrarily, the database is saturated with low-density data points. These low-density data points were required at the early stages of this project when an analytical solution to address the ideal limit gas was not available. From Figure 2.3 (b), it is observed that in contrast to Figure 2.3 (a), the data points at high and low temperatures are closer when deforming (inverting) the linear temperature scale. The inverse temperature scale is better distributed. This numerical convenience eases the interpolating efforts of the FE-ANN EoS. Finally, it can be observed from Figure 2.3 (c) that the database considers 600 to 800 data points per repulsive exponent.

## 2.5 FE-ANN EoS for the Mie fluid

### 2.5.1 FE-ANN EoS: implementation and architecture

The Mie fluid is characterized by its potential well depth ( $\epsilon$ ), its length scale ( $\sigma$ ), and their repulsive ( $\lambda_r$ ) and attractive ( $\lambda_a$ ) exponents. The energy and length scale dependence can be omitted when using reduced units (see Table 2.1). Additionally, it has been shown that the attractive and repulsive exponents are conformal for fluid equilibrium properties [92]. The conformality of the Mie fluid implies that multiple combinations of repulsive and attractive exponents with the same van der Waals parameter ( $\alpha_{vdw}$ ) lead to a similar macroscopic description of the vapor-liquid phase envelope. Hence, a Mie fluid can be characterized by its van der Waals parameter defined below.

$$\alpha_{vdw} = - \int_1^\infty \mathcal{U}^{*,Mie} r^{*2} dr^* = \underbrace{\frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}}_{C^{Mie}} \left[ \frac{1}{\lambda_a - 3} - \frac{1}{\lambda_r - 3} \right] \quad (2.21)$$

The  $\alpha_{vdw}$  parameter measures the balance between attractive and repulsive interaction in the molecular system. This parameter varies from 0.53 ( $\lambda_r = 34$  and  $\lambda_a = 6$ ) to 1.47 ( $\lambda_r = 7$  and  $\lambda_a = 6$ ) for the Mie fluids available in the computed thermophysical properties database. This parameter is chosen as the molecular descriptor ( $\chi$ ) needed for the FE-ANN EoS, Eq. (2.9). Hence, the FE-ANN EoS for

the Mie fluid is a function of the van der Waals parameter ( $\alpha_{\text{vdw}}$ ), the reduced density ( $\rho^*$ ), and the temperature (as  $T^*$  or  $1/T^*$ ).

$$A^{*,\text{res}} = \text{ANN}(\alpha_{\text{vdw}}, \rho^*, T^*) - \text{ANN}(\alpha_{\text{vdw}}, \rho^* = 0, T^*) \quad (2.22a)$$

or,

$$A^{*,\text{res}} = \text{ANN}(\alpha_{\text{vdw}}, \rho^*, \frac{1}{T^*}) - \text{ANN}(\alpha_{\text{vdw}}, \rho^* = 0, \frac{1}{T^*}) \quad (2.22b)$$

The hyperparameters of the ANNs in Eqs. (2.22) must be tuned to provide satisfactory results. The number of hidden layers and their size (number of neurons) are optimized using the Tree-structured Parzen Estimator [159] method implemented in the black-box optimization library Optuna [160]. In order to reduce the complexity of the optimization, all hidden layers are forced to have the same number of neurons. Additionally, to ensure the Helmholtz function is differentiable, the hidden layers use the tanh activation function ( $f(x) = \tanh(x)$ ). The ANNs used in this work consist of 3 hidden layers of 30 neurons each. The output layer consists of a layer with a single neuron using the linear activation function ( $f(x) = x$ ). Other hyperparameters are set to standard choices [135]. For example, the batch size is set to 32, and the Adam optimizer is used for training [139]. The learning rate and number of training epochs are hand-tuned to obtain stable training and a converged ANN. Further details about the hyperparameters of the FE-ANN EoS can be found in Section A.2 of Appendix A.

The FE-ANN EoS is implemented using JAX [138] and Flax [161] and is trained using Optax, which is part of the DeepMind JAX Ecosystem [162].

### 2.5.2 FE-ANN EoS: training and performance

The produced Mie fluid database is used to train different versions of the FE-ANN EoS. The tested training approaches consider assessing the impact of using  $(\alpha_{\text{vdw}}, \rho^*, T^*)$  or  $(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  as input variables to the ANN as shown in Eqs. (2.22). Additionally, it was noticed that during training, there was a compromise with respect to which of the target properties the FE-ANN EoS would attempt to optimize better. The impact of training on first-order ( $Z$ ,  $U^*$  and  $B_2^*$ ) and second-order ( $C_V^*$ ,  $\gamma_V^*$ ,  $\rho^* \kappa_T^*$ ,  $\alpha_p^*$ ,  $\gamma$ ,  $\mu_T^*$ , and  $B_3^*$ ) is also assessed by controlling the relative weight ( $w_i$ ) of these properties in the loss function, Eq. (2.11).

The FE-ANN EoS is trained using 90% of the thermophysical properties database, while the remaining 10% is left for testing purposes. The chosen split offers sufficient coverage of the phase space in both training and testing datasets. The test dataset allows for assessing the performance of the FE-ANN EoS on unseen data. If the errors are considerably lower for the training data compared to the testing data, it is a sign that the model is overfitting the noise of the data and will not generalize well. The training and testing use a batch size of 32. The recommended training procedure for the FE-ANN EoS consists of two main steps. Firstly, the FE-ANN EoS is trained using only first-order properties for 1000 epochs using a learning rate  $\alpha = 10^{-4}$ . This first stage ensures that the parameters of the ANN are quickly trained to overfit these properties. The second stage aims to correct second-order properties and potential unphysical behaviors. This stage considers first- and second-order properties as well as the penalty functions, Eqs. (2.13). The inclusion of second-order properties in the loss function compromises the numerical stability of the training, as small changes in the trainable parameters cause



significant changes in the predicted derivative data. For this reason, in this second stage, the FE-ANN EoS is trained using a relatively low learning rate  $\alpha = 10^{-6}$  to a total of 20000 epochs. An appropriate balance between these properties can be achieved by employing relative weights ( $w_i$ ) in the loss function. The relative weight for first-order properties and the penalty functions are set to 1. The relative weight for second-order derivative properties can vary from 0, where the model ignores the second-order properties, up to 1, where these properties are equally important as first-order properties. This chapter considers the following relative weights for second-order properties: 0, 0.01, 0.05, and 1.

The different trained FE-ANN EoSs deviation results are summarized in Table 2.3. The results for FE-ANN EoS that use  $(\alpha_{\text{vdw}}, \rho^*, T^*)$  as input feature, Eq. (2.22a), are shown in the top half of this table, while the deviations of the FE-ANN EoS that use  $(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  as input features, Eq. (2.22b), are shown in bottom half of the table. Despite which input features are used in the FE-ANN EoS, the effect of the second-order properties is similar in the EoS. When the second-order properties are not considered in the model training, i.e., the relative weight is fixed to  $w_i = 0$ , the EoSs accurately describe the compressibility factor, the internal energy, and the second virial coefficient. However, this choice does not guarantee a good performance in the description of second-order properties, which are, in this case, predicted. In particular, properties related to cross second-order derivatives, such as the thermal expansion coefficient ( $\alpha_p^*$ ), adiabatic index ( $\gamma$ ), and Joule-Thomson coefficient ( $\mu_{\text{JT}}^*$ ). Similarly, the third virial coefficient ( $B_3^*$ ) is not correctly modeled unless the FE-ANN EoS has seen it during the training. The higher errors are especially noticeable compared to the other EoSs that include second-order properties in the training. At the other extreme, when the relative weight for the second-order properties is high (i.e.,  $w_i = 1$ ), there is a clear overfit of these properties that comes at the cost of high deviations for the first-order properties. This unsatisfactory result is a consequence of the loss function being biased toward these properties and attempting to fit the second derivatives independent of the pressure and internal energy values. This extreme is counterproductive. Second-order properties are obtained by analyzing the fluctuations of the molecular simulation and are inherently noisier than the first-order properties.

The relative weight for the second-order properties is a design choice. First-order derivative properties should be as accurate as possible because they relate to important phase behavior, like phase equilibria. Similarly, second-order properties are related to the thermodynamic consistency of the model. The relative weight for the second-order properties should be enough to account for this information but not excessively large to bias the optimizer towards these properties. It is accepted that the relative weights in between (i.e.,  $w_i = 0.01$  or 0.05) produce a more balanced EoS that captures both first- and second-order derivative properties well.

This balance is found, for example, in the FE-ANN EoSs, trained using  $w_i = 0.05$ . From Table 2.3, it is noticed that both choices of input variables for the temperature ( $T^*$  or  $1/T^*$ ) provide similar deviations for the target thermophysical properties, with the model using  $1/T^*$  showing slightly smaller errors. An inspection of how deviations of these two models are distributed in the density-temperature projection is shown in Figure 2.4. The selected thermophysical properties allow assessing the behavior of first ( $Z$  and  $U^*$ ), second ( $C_V^*$  and  $\rho^* \kappa_T^*$ ), and cross ( $\gamma_V^*$ ) derivatives of the Helmholtz free energy surface. It can be observed that, in general, both FE-ANN EoS formulations, Eqs. (2.22a) and (2.22b) behave similarly with respect to the reference data. The dark purple data points show that

TABLE 2.3: Comparison of Mean Squared Error (MSE) of selected trained FE-ANN EoS.<sup>a, b</sup>

| Input Features <sup>d</sup>          | $w_i^e$ |       | $100 \times \text{MSE}^c$ |             |             |                     |              |              |             |              |                |             |
|--------------------------------------|---------|-------|---------------------------|-------------|-------------|---------------------|--------------|--------------|-------------|--------------|----------------|-------------|
|                                      |         |       | Z                         | $U^*$       | $C_V^*$     | $\rho^* \kappa_T^*$ | $\gamma_V^*$ | $\alpha_P^*$ | $\gamma$    | $\mu_{JT}^*$ | $B_2^*$        | $B_3^*$     |
| $\alpha_{\text{vdw}}, \rho^*, T^*$   | 0.0     | Train | <b>0.11</b>               | <b>0.07</b> | 0.91        | 0.10                | 0.79         | 0.63         | 0.88        | 2.83         | <b>2.6e-04</b> | 96.60       |
|                                      |         | Test  | <b>0.51</b>               | <b>0.35</b> | 1.25        | 0.15                | 0.67         | 1.17         | 0.98        | 5.15         | <b>2.2e-04</b> | 98.16       |
|                                      | 0.01    | Train | 0.12                      | 0.08        | 0.29        | 0.09                | <b>0.73</b>  | 0.53         | 0.53        | 2.36         | 0.01           | 0.05        |
|                                      |         | Test  | 0.52                      | 0.37        | 0.44        | 0.14                | <b>0.63</b>  | 0.89         | 0.59        | 4.06         | 0.02           | 0.05        |
|                                      | 0.05    | Train | 0.14                      | 0.09        | 0.23        | 0.07                | 0.74         | 0.43         | <b>0.50</b> | 2.35         | 0.03           | <b>0.03</b> |
|                                      |         | Test  | 0.54                      | 0.38        | 0.36        | 0.12                | 0.65         | 0.73         | 0.56        | <b>3.88</b>  | 0.04           | <b>0.03</b> |
|                                      | 1.0     | Train | 0.41                      | 0.15        | <b>0.18</b> | <b>0.06</b>         | 0.79         | <b>0.32</b>  | 0.53        | <b>2.35</b>  | 0.20           | 0.06        |
|                                      |         | Test  | 0.78                      | 0.44        | <b>0.28</b> | <b>0.08</b>         | 0.72         | <b>0.52</b>  | <b>0.56</b> | 3.90         | 0.19           | 0.04        |
|                                      | 0.0     | Train | <b>0.11</b>               | <b>0.06</b> | 1.29        | 0.08                | 0.98         | 0.47         | 2.07        | 2.76         | <b>3.6e-04</b> | 539.89      |
|                                      |         | Test  | <b>0.50</b>               | <b>0.35</b> | 1.53        | 0.10                | 0.98         | 0.90         | 2.42        | 4.92         | <b>3.8e-04</b> | 642.67      |
| $\alpha_{\text{vdw}}, \rho^*, 1/T^*$ | 0.01    | Train | 0.12                      | 0.07        | 0.23        | 0.06                | 0.77         | 0.34         | 0.55        | 2.36         | 0.01           | 0.06        |
|                                      |         | Test  | 0.51                      | 0.35        | 0.35        | 0.08                | 0.68         | 0.59         | 0.59        | 3.96         | 0.01           | 0.04        |
|                                      | 0.05    | Train | 0.13                      | 0.08        | 0.19        | 0.06                | 0.71         | 0.32         | 0.50        | 2.34         | 0.04           | 0.04        |
|                                      |         | Test  | 0.52                      | 0.37        | 0.30        | <b>0.08</b>         | 0.63         | <b>0.54</b>  | 0.55        | 3.86         | 0.04           | 0.03        |
|                                      | 1.0     | Train | 0.22                      | 0.15        | <b>0.17</b> | <b>0.06</b>         | <b>0.70</b>  | <b>0.31</b>  | <b>0.50</b> | <b>2.34</b>  | 0.16           | <b>0.03</b> |
|                                      |         | Test  | 0.60                      | 0.43        | <b>0.28</b> | 0.08                | <b>0.60</b>  | 0.55         | <b>0.54</b> | <b>3.83</b>  | 0.16           | <b>0.02</b> |

<sup>a</sup> All the trained FE-ANN EoS consists of a neural network with three hidden layers, 30 neurons each, and the tanh activation function. A 90/10 split is used for training and testing purposes. The training is performed using a batch size of 32. All EoSs were trained initially for 1000 epochs considering only first-order properties and a learning rate of  $\alpha = 10^{-4}$ . Then, the EoSs were trained to a total of 20000 epochs, considering both first and second-order properties using a learning rate of  $\alpha = 10^{-6}$ . The ADAM optimizer [139] was used for training all the models. <sup>b</sup> The smallest MSE deviations for the test and train datasets are highlighted in **bold**. <sup>c</sup> The Mean Squared Error (MSE) is obtained as  $100 \times \text{MSE} = 100/N_P \sum_j^{N_P} (y_j^{\text{pred}} - y_j^{\text{true}})^2$ . <sup>d</sup> Inputs features to the FE-ANN EoS as defined in Eqs. (2.22a) and (2.22b). <sup>e</sup>  $w_i$  is the relative weight used for the second-order derivative properties in the loss function, Eq. (2.12). Relative weights of first-order properties and penalties are set to 1.

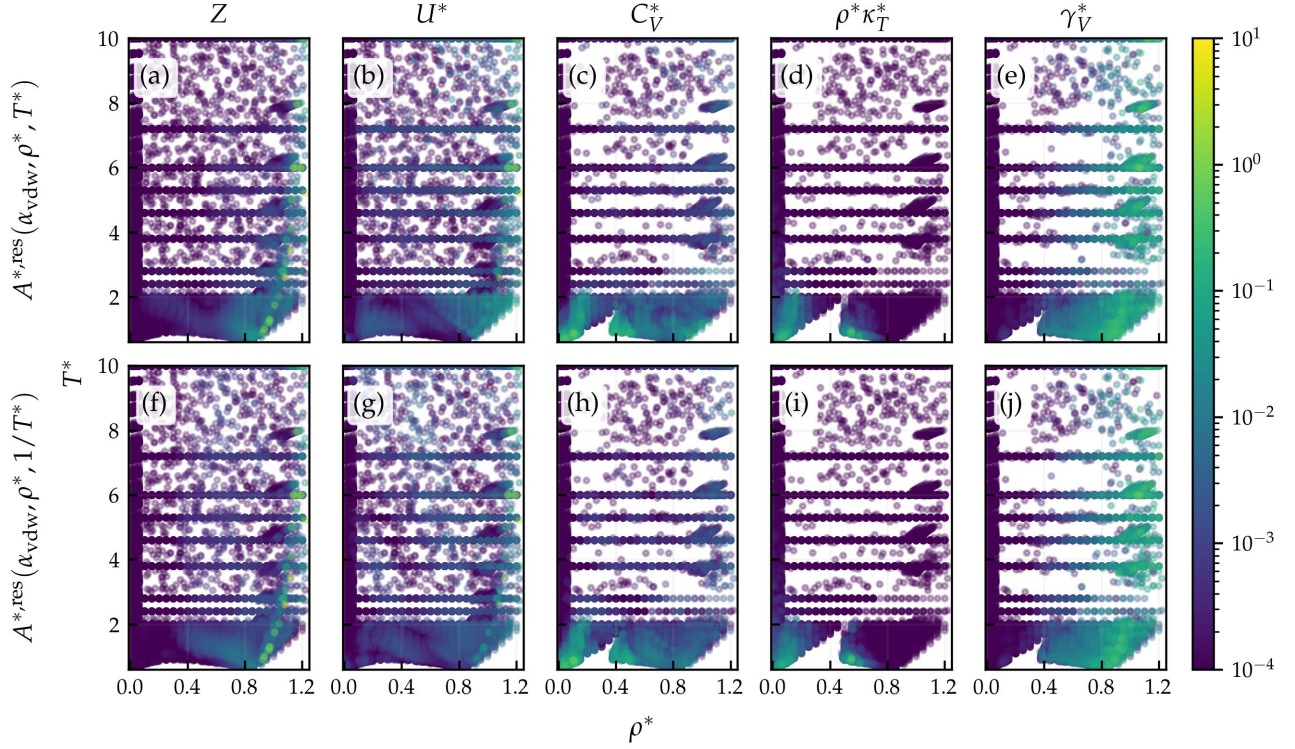


FIGURE 2.4: **Squared error distribution of the FE-ANN EoS for selected first- and second-order derivative properties.** Thermophysical properties. (a) and (f): Compressibility factor ( $Z$ ). (b) and (g): Internal energy ( $U^*$ ). (c) and (h): Isochoric heat capacity ( $C_V^*$ ). (d) and (i): Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (e) and (j): Thermal pressure coefficient ( $\gamma_V^*$ ). FE-ANN EoS models. (a) - (e) FE-ANN EoS using  $(\alpha_{\text{vdw}}, \rho^*, T^*)$  as input variables, see Eq. (2.22a). (f) - (j) FE-ANN EoS using  $(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  as input variables, see Eq. (2.22b). FE-ANN EoS trained using a relative weight  $w_i = 0.05$  for second-order properties, see Eq. (2.12). The squared error is indicated by the color bar and computed as  $(y^{\text{pred}} - y^{\text{true}})^2$ . Dark purple data points refer to low deviation, while yellow data points refer to high deviation with respect to the database.

a vast region of the phase space is modeled with minor errors. The primary deviations, indicated by the green and yellow data points, occur in two distinguishable regions. The first one occurs at low temperatures at densities that are close to the Vapor-Liquid Equilibria (VLE) phase boundaries, and the second one occurs at high densities at conditions that are likely to border the Solid-Liquid Equilibria (SLE) phase boundary. The thermophysical properties of a particle can drastically change in the vicinity of the phase boundaries, especially inside them (e.g., for metastable states). For this reason, it is still acceptable for the FE-ANN EoSs to exhibit higher deviations in those regions.

It is concluded that, for a highly populated phase space, both Eq. (2.22a) (using  $T^*$ ) and Eq. (2.22b) (using  $1/T^*$ ) exhibit a similar performance. However, it is observed that the latter formulation has some advantages. By deforming the temperature scale to  $1/T^*$ , higher temperature data points move closer, c.f. Figure 2.3 (b), this numerical artifact facilitates the interpolating efforts of the FE-ANN EoS. Moreover, as explained in Section 2.3, the use of the temperature inverse follows the natural temperature formulation in statistical thermodynamics. For these reasons, the FE-ANN EoS with  $(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  as inputs, Eq. (2.22b), and trained with  $w_i = 0.05$  is preferred to showcase the framework.

The parity plots of the FE-ANN EoS with  $(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  as inputs and  $w_i = 0.05$  are shown in

Figure 2.5. From Figure 2.5 (a), (b), and (c), it can be seen that the first-order properties obtained from the FE-ANN EoS, such as the compressibility factor, internal energy, and second virial coefficient, faithfully reproduce the reference data. For second-order derivative properties, Figure 2.5 (d) to (i), it can be observed that most of the FE-ANN EoS calculations lie in the diagonal, indicating a good agreement with the simulation data. Even though no explicit regularization technique was applied while training the FE-ANN EoS, it can be observed that both **training (blue data points)** and **testing (orange data points)** are modeled with similar deviations. This lack of overfitting is not a coincidence but a consequence of training the FE-ANN EoS based on derivative information and thermodynamically-inspired penalties. The derivative information constrains the shape of the resulting Helmholtz free energy surface and provides information about its changes with respect to temperature and density. These constraints implicitly prevent the optimizer from being biased to perform well only on the data it has seen, but instead, it forces it to find the best "well-behaved" Helmholtz free energy surface that describes the data.

### 2.5.3 FE-ANN EoS: thermophysical properties modeling and prediction

The FE-ANN EoS with  $\alpha_{vdw}$ ,  $\rho^*$  and  $1/T^*$  as input features, Eq. (2.22b), trained using  $w_i = 0.05$  for second-order properties is selected to showcase the capabilities of this framework.

Through this section, the use of the FE-ANN EoS is exemplified by describing the thermophysical properties of three selected Mie fluids, namely:  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid),  $\lambda_r = 16, \lambda_a = 6$  and  $\lambda_r = 20, \lambda_a = 6$ . Selected pressure isotherms of these Mie fluids are shown in Figure 2.6. For reference, for the Lennard-Jones fluid, Figure 2.6 (a), these isotherms include two at subcritical conditions ( $T^* = 0.8$  and  $1.0$ ), one near-critical condition isotherm ( $T^* = 1.3$ ) and two at supercritical conditions ( $T^* = 2.8$  and  $6.0$ ). This figure includes molecular simulation data and results from the FE-ANN (data-driven-based) and the SAFT-VR-Mie (theoretically-based) [37] EoSs. The FE-ANN EoS (solid lines) predicts well-behaved pressure isotherms at the trained density conditions. Both the theoretically-based and data-driven EoSs show excellent agreement with the molecular dynamics data and practically overlap. The main discrepancy between the two EoSs occurs in the high-density/low-temperature region. As suggested by the correlation of Khrapak et al. [151], and indicated by the grey-shaded area in the figure, these Mie fluids freeze under those conditions. It must be noted that both models are developed exclusively for the fluid phase. Therefore, both of the EoSs are incorrect in those conditions.

The presence of an unstable van der Waals loop at the low-pressure region must be highlighted. For clarity, this unstable behavior is shown in the zoomed area of Figure 2.6. This physically meaningful attribute of the isotherm appears as a prediction in a region of the phase space that has not been populated with data. The prediction of unstable regions suggests that the FE-ANN EoS can compute phase equilibria despite not being exposed to vapor-liquid equilibria (VLE) data. This predictive behavior is a rather considerable achievement for a model trained using only single-phase PVT data.

The FE-ANN EoS is also capable of modeling second-order derivative properties accurately. In Figure 2.7, selected isotherms of the isochoric heat capacity, isothermal compressibility, thermal pressure coefficient, thermal expansion coefficient, adiabatic index, and Joule-Thomson coefficient of the ( $\lambda_r=12$ ,

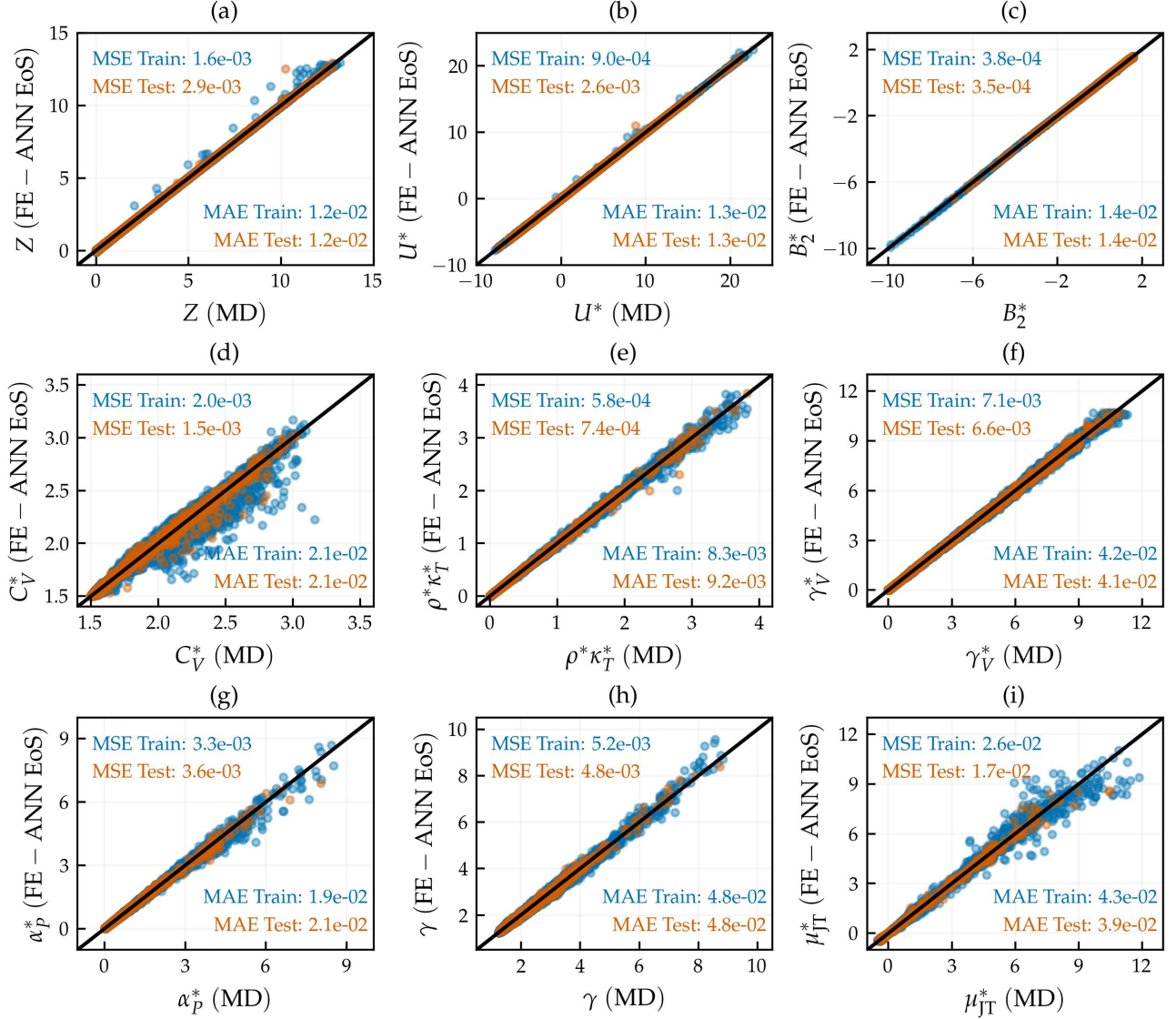


FIGURE 2.5: Parity plot of thermophysical properties computed with FE-ANN EoS with respect to the reference database. (a) Compressibility factor ( $Z$ ). (b) Internal energy ( $U^*$ ). (c) Second virial coefficient ( $B_2^*$ ). (d) Isochoric heat capacity ( $C_V^*$ ). (e) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (f) Thermal pressure coefficient ( $\gamma_V^*$ ). (g) Thermal expansion coefficient ( $\alpha_P^*$ ). (h) Adiabatic index (heat capacity ratio,  $\gamma = C_P^*/C_V^*$ ). (i) Joule-Thomson coefficient ( $\mu_{JT}^*$ ). The reference data points were obtained as explained in Section 2.4. Thermophysical properties with a "(MD)" in the x-axis label were obtained from molecular dynamics simulations. The second virial coefficient was obtained using Eq. (2.17). The thermophysical properties are compared with the results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Blue circles refer to the train data points and orange circles refer to the test data set. The reported Mean Squared Error (MSE) is obtained as  $MSE = 1/N_P \sum_j^{N_P} (y_j^{\text{pred}} - y_j^{\text{true}})^2$ . The reported Mean Absolute Error (MAE) is obtained as  $MAE = 1/N_P \sum_j^{N_P} |y_j^{\text{pred}} - y_j^{\text{true}}|$ .



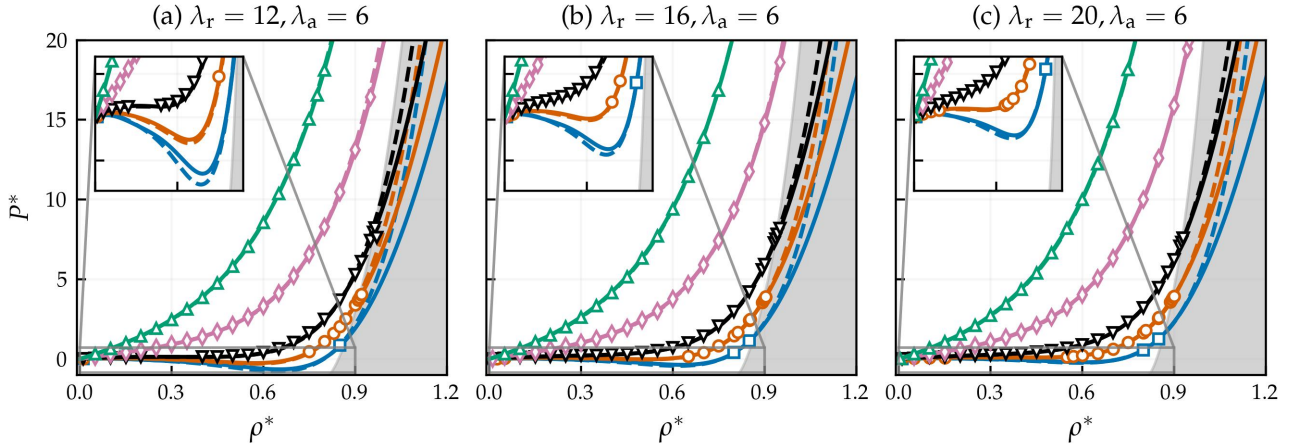


FIGURE 2.6: **Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** (a)  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid). (b)  $\lambda_r = 16, \lambda_a = 6$ . (c)  $\lambda_r = 20, \lambda_a = 6$ . Molecular simulation data. Blue squares:  $T^* = 0.8$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results. The computed isotherms from the EoSs follow the same color convention as the simulation data. The grey-shaded area represents the region where freezing is predicted by the correlation of Khrapak et al. [151], see Eq. (2.18).

16, and 20 and  $\lambda_a = 6$ ) Mie fluids are shown. Similarly, as in Figure 2.6, subcritical, near-critical, and supercritical isotherms are shown. For conciseness, the discussion will focus on Figure 2.7i, which corresponds to the Lennard-Jones fluid. This figure allows for comparing molecular simulation data and the FE-ANN EoS (solid lines) and SAFT-VR-Mie EoS (dashed lines) results. Similar to the pressure isotherms, the FE-ANN EoS and SAFT-VR-Mie EoS are in good agreement with the simulation data and predict a similar behavior for the second-order derivative properties. As expected, the main discrepancies between the two EoSs occur in regions without available data. These correspond to the region that likely lies inside the VLE phase envelope and in the solid-like region at high densities.

From Figure 2.7 (a), it can be observed that the isochoric heat capacity computed from the FE-ANN EoS is in good agreement with the computed molecular dynamics data. The heat capacities obtained from the SAFT-VR-Mie EoS and the FE-ANN EoS also differentiate at the region likely to lie inside the binodal curve. The FE-ANN EoS predicts a higher maximum for this property, which is in agreement with the available molecular dynamics data. This maximum of the isochoric capacity is well-reported from experimental [12] and simulation [3, 163] data. It is also predicted by a wide range of families of EoSs, such as cubic [28], molecular-based [37] and technical [16, 115] EoSs. For the isothermal compressibility, depicted in Figure 2.7 (b), it can be observed that both the FE-ANN EoS and the SAFT-VR-Mie EoS are in good agreement with the molecular dynamics results. The result for these two properties is directly related to the second derivatives of the Helmholtz free energy with respect to temperature and density, respectively. Hence, if the first derivatives-related properties (i.e., internal energy and pressure) are correctly modeled, these properties should behave correctly. It must be emphasized that these two thermophysical properties are related to the *criteria of thermal and mechanical stability*, and additional penalties (Eqs. (2.13b) and (2.13c)) were included to ensure that the FE-ANN EoS follows a physically correct behavior.

For thermophysical properties related to cross-second-order derivatives, Figure 2.7 (c)-(f), the FE-ANN

EoS also predicts smooth isotherms that agree with the simulation data. This result is expected from including these properties in the training. As observed in Figure 2.7 (c), the thermal pressure coefficient predicted by the FE-ANN EoS is always positive, which guarantees pressure isotherms do not cross. This behavior is required by the training data, and it is further enforced with a penalty, Eqs. (2.13d). For the isothermal compressibility, Figure 2.7 (c), thermal expansion coefficient, Figure 2.7 (d), the adiabatic index, Figure 2.7 (e), it is also observed these properties change considerably inside the binodal curve. This behavior follows the SAFT-VR-Mie EoS results, and it is a consequence of the spinodal points in the pressure isotherms (i.e.,  $(\partial P^*/\partial \rho^*)_{T^*} = 0$ ) for subcritical isotherms. Eqs. (2.3b), (2.3c) and (2.3e) can be rearranged into Eqs. (2.23) to explicitly show these properties diverge at the spinodal point due to a division by zero.

$$\alpha_P^* = \left( \frac{\partial P^*}{\partial T^*} \right)_{\rho^*} / \left[ \rho^* \left( \frac{\partial P^*}{\partial \rho^*} \right)_{T^*} \right] \quad (2.23a)$$

$$\kappa_T^* = 1 / \left[ \rho^* \left( \frac{\partial P^*}{\partial \rho^*} \right)_{T^*} \right] \quad (2.23b)$$

$$\gamma = \frac{C_P^*}{C_V^*} = 1 + \left[ T^* \left( \frac{\partial P^*}{\partial T^*} \right)_{\rho^*}^2 \right] / \left[ \rho^{*2} C_V^* \left( \frac{\partial P^*}{\partial \rho^*} \right)_{T^*} \right] \quad (2.23c)$$

Finally, the Joule-Thomson coefficient can be obtained from the thermal expansion coefficient and the isobaric heat capacity. This coefficient can still diverge inside the binodal region. As shown in Figure 2.7 (f), the FE-ANN EoS shows a diverging behavior for the subcritical isotherms. This same diverging behavior is observed for the lowest temperature isotherm ( $T^* = 0.8$ ) predicted by the SAFT-VR-Mie EoS.

From the above, it is concluded that the FE-ANN EoS has been able to "learn" the residual Helmholtz free energy of  $(\lambda_r, \lambda_a = 6)$  Mie fluids using a thermodynamically-consistent data-driven approach. The correct behavior of the FE-ANN EoS has been shown here for  $\lambda_r = 12, 16$ , and  $20$ . However, the excellent agreement of this EoS with respect to molecular simulation data is also found for other Mie fluids with  $\lambda_r \in [7, 34]$ . The reader is referred to Section B.1 in Appendix B for calculations referring to other Mie fluids.

The unstable van der Waals pressure loops observed at subcritical conditions in Figure 2.6 suggested that the FE-ANN EoS can be employed to compute vapor-liquid equilibria (VLE). Phase equilibrium is obtained when two phases of different densities exhibit the same pressure and chemical potential at a given temperature. In practical terms, this corresponds to solving the system of Eqs. (2.7). The results for phase equilibria calculations are shown in Figure 2.8. In Figures 2.8 (a)-(c), it can be observed that the FE-ANN EoS (solid lines) predicts a correct phase envelope, vaporization pressure and vaporization enthalpy of the selected Mie Fluids ( $\lambda_r = 12, 16$ , and  $20$  and  $\lambda_a = 6$ ). The predicted phase equilibria results from the FE-ANN EoS are comparable to the SAFT-VR-Mie EoS (dashed lines), which is arguably one of the best molecular-inspired models for the Mie fluid. As observed from the parity plots, Figures 2.8 (d)-(e), the results from the FE-ANN EoS are in excellent agreement with the molecular simulation data. The Mean Average Absolute Deviation (MAPE) for the saturated liquid densities, vapor densities, equilibrium pressure, and vaporization enthalpy are 1.26, 4.23, 2.59, and 1.27, respectively. These deviations are similar to the ones obtained by the SAFT-VR-Mie EoS (1.50,

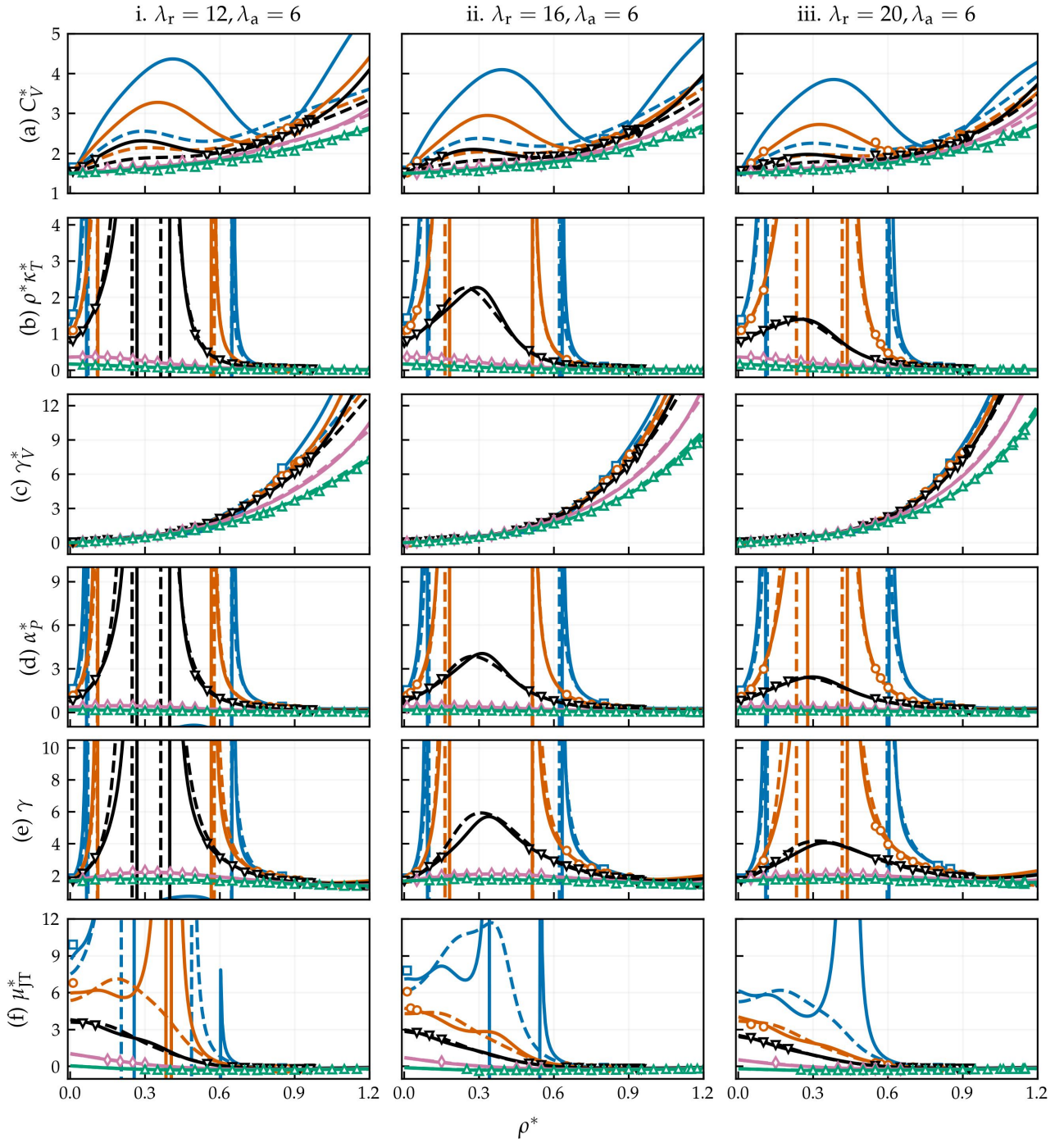


FIGURE 2.7: **Second-order derivative properties isotherms of selected  $(\lambda_r, \lambda_a = 6)$  Mie fluids.** i.  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid). ii.  $\lambda_r = 16, \lambda_a = 6$ . iii.  $\lambda_r = 20, \lambda_a = 6$ . (a) Isochoric heat capacity ( $C_V^*$ ). (b) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (c) Thermal pressure coefficient ( $\gamma_V^*$ ). (d) Thermal expansion coefficient ( $\alpha_P^*$ ). (e) Adiabatic index (heat capacity ratio,  $\gamma = C_P^*/C_V^*$ ). (f) Joule-Thomson coefficient ( $\mu_{JT}^*$ ). Molecular simulation data. Blue squares:  $T^* = 0.8$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results. The computed isotherms from the EoSs follow the same color convention as the simulation data.



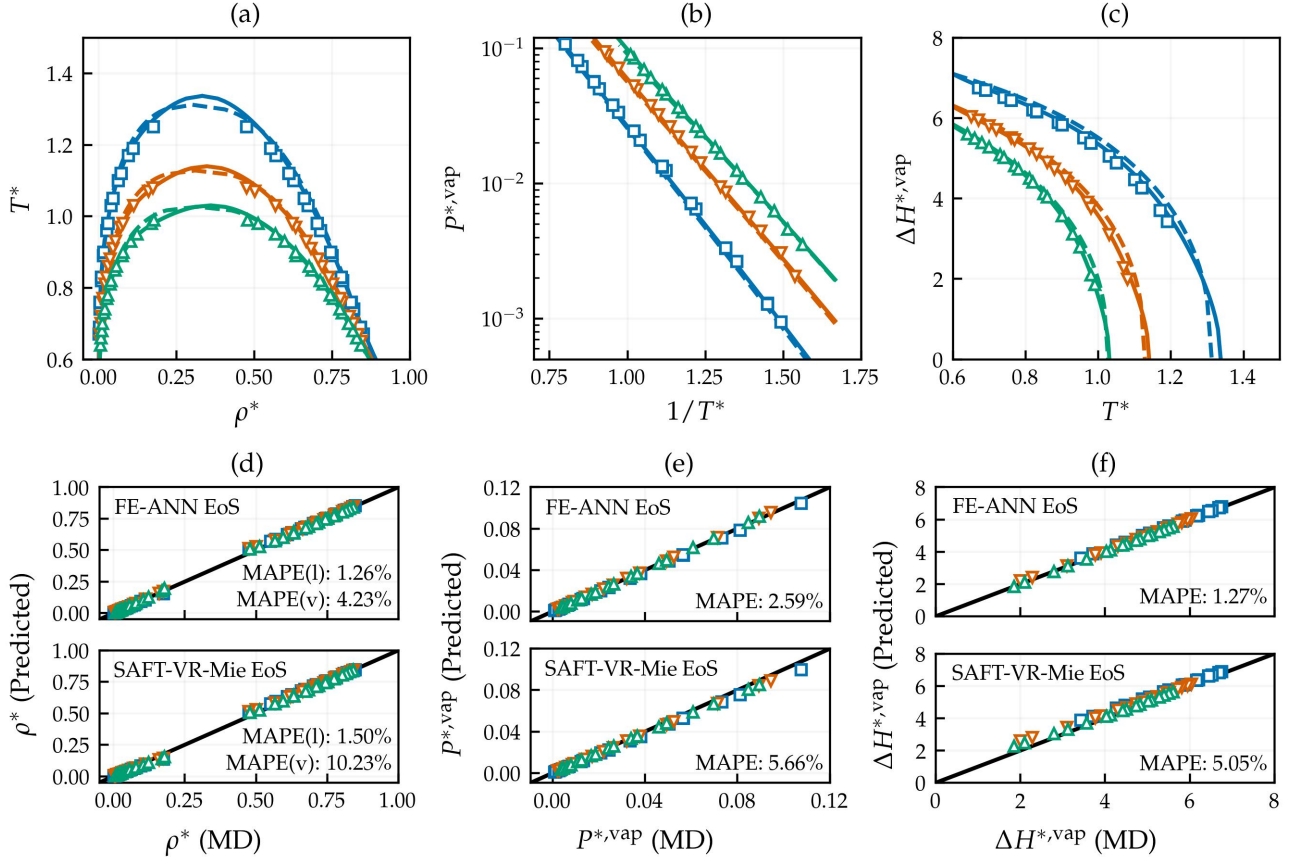


FIGURE 2.8: **Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** (a) Phase envelope. (b) Vaporization pressure. (c) Vaporization enthalpy. Molecular simulation data. Blue squares:  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid), orange downward triangles:  $\lambda_r = 16, \lambda_a = 6$ , and green upright triangles:  $\lambda_r = 20, \lambda_a = 6$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results. These VLE results are obtained by numerically solving the system of Eqs. (2.7). The predicted phase equilibria from the EoSs follow the same color convention as the simulation data. Parity plots. (d) Equilibrium densities. (e) Vaporization pressure. (f) Vaporization enthalpy. The top subplot refers to the FE-ANN EoS results, and the bottom subplot shows the SAFT-VR-Mie results. The symbols and colors follow the same convention as in (a), (b) and (c). The reported Mean Absolute Percentage Error (MAPE) is obtained as:  $\text{MAPE} = 100/N_P \sum_j^{N_P} |(y_j^{\text{pred}} - y_j^{\text{true}})/y_j^{\text{true}}|$ .

10.23, 5.66, and 5.05, following the same order). It is important to recall that the FE-ANN EoS has seen saturated data as single-phase PVT data but that no information about these points being in phase equilibria has been provided to the model during its training. Hence, these results are genuinely a prediction from the learned Helmholtz free energy surface.

It is encouraging that the FE-ANN EoS predicts the existence of a critical point. The critical point corresponds to an inflection point in the pressure-density diagram, and it is obtained as the point where the first- and second-order derivatives of the pressure with respect to the density vanish. This mathematical aspect implies that the isothermal compressibility diverges at the critical point. The critical point is obtained by numerically solving the system of Eqs. (2.6). The critical points predicted by the FE-ANN EoS and its comparison against published molecular simulation data [151, 164–166], and the SAFT-VR-Mie EoS predictions are shown in Figure 2.9. As observed in this figure, the FE-ANN EoS (solid line) correctly predicts that the critical temperature decreases, and the critical

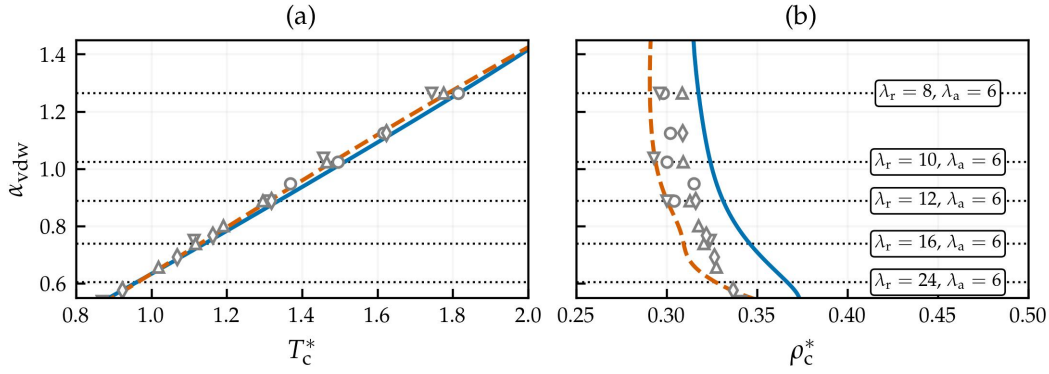


FIGURE 2.9: **Critical point of  $(\lambda_r, \lambda_a = 6)$  Mie fluids.** (a) Critical temperature ( $T_c^*$ ). (b) Critical density ( $\rho_c^*$ ). Grey empty symbols refer to published molecular simulation results: circles [151], upright triangles [166], downward triangles [164], and diamonds [165]. Blue solid lines refer to the predicted critical point FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed orange lines are SAFT-VR-Mie EoS [37] results. Critical points from the FE-ANN and SAFT-VR-Mie EoS are obtained by numerically solving Eqs. (2.6).

density increases when the repulsions of the system increase. This behavior is again observed in the simulation data (symbols) and the SAFT-VR-Mie EoS results (dashed line). The critical point is obtained from simulations by extrapolating the fit of a critical scaling equation [167] using equilibrium density data. The reported values should be used as a guide, not as an absolute truth. The FE-ANN EoS and SAFT-VR-Mie EoS correctly agree with the critical temperature. For the critical density, the SAFT-VR-Mie EoS is in better agreement with the available data, as the FE-ANN EoS slightly overestimates the critical density. Nevertheless, it must be noted that, as seen in the figures, the uncertainty of the critical density data is higher than that of the critical temperature data.

The physical consistency of an equation of state at different temperature and pressure conditions can be evaluated through Brown's characteristic curves [168]. These curves are related to specific temperature and pressure conditions for which a fluid exhibits the behavior of an ideal gas. This chapter considers the consistency of the following four curves: Amagat, Charles, Boyle, and Zeno. The Amagat curve (also known as the Joule inversion curve) is given by Eq. (2.24a). This curve takes place at extreme temperatures and pressures and encloses the other characteristic curves. In the low-pressure limit, the Amagat curve converges to the Amagat temperature, which is the temperature where the second virial coefficient ( $B_2^*$ ) exhibits a maximum value (i.e.,  $T_{\text{Amagat}}^* : dB_2^*/dT^* = 0$ ). The Charles curve is given by Eq. (2.24b). This curve is also known as the Joule-Thomson inversion curve, which refers to the conditions where the Joule-Thomson coefficient is zero. The Amagat curve encloses the Charles curve. The Charles curve starts at the saturation pressure curve and converges to the temperature where the second virial coefficient exhibits an asymptotic behavior (i.e.,  $T_{\text{Charles}}^* : dB_2^*/dT^* = B_2^*/T^*$ ) at the low-pressure regime. This temperature is known as the Charles temperature. The condition for the Boyle curve is given by Eq. (2.24c). The Boyle curve is expected to lie inside the Charles curve. This curve starts in the vicinity of the critical point and approaches the Boyle temperature (i.e.,  $T_{\text{Boyle}}^* : B_2^* = 0$ ) at the low-pressure limit. Finally, the Zeno curve is given by the conditions in which a fluid follows exactly the ideal gas equation for the pressure,  $P^* = \rho^* T^*$ , as shown in Eq. (2.24d). The Zeno curve intersects the Charles curve once and ends at the Boyle temperature in the low-pressure region. For further discussion about Brown's characteristic curves, the reader is referred to articles of

Deiters and De Reuck [89] and Stephan and Deiters [169].

$$\text{Amagat curve: } \left( \frac{\partial Z}{\partial T^*} \right)_{\rho^*} = 0. \quad (2.24a)$$

$$\text{Charles curve: } \left( \frac{\partial Z}{\partial T^*} \right)_{P^*} = 0. \quad (2.24b)$$

$$\text{Boyle curve: } \left( \frac{\partial Z}{\partial P^*} \right)_{T^*} = 0. \quad (2.24c)$$

$$\text{Zeno curve: } Z = 1 \quad (2.24d)$$

The faithful and correct reproduction of Brown's characteristic curves is a stringent consistency test for any equation of state [166, 170]. In Figure 2.10, Brown's characteristic curves and the second and third virial coefficients computed with the FE-ANN EoS for selected Mie fluid are shown. It can be observed that these curves exhibit the physically correct behavior described by Deiters and De Reuck [89]. This is, the Zeno and Boyle curves converge to the Boyle temperature in the low-pressure limit, the Charles curve reaches the Charles temperature at the low-pressure limit, the Zeno curve also intersects the Charles curve only once, and the Boyle, Charles, and Amagat curves do not cross each other. Additionally, the prediction of multiple stable Charles curves was studied using the FE-ANN EoS. The prediction of multiple Charles curves is a well-reported unphysical problem of some cubic and SAFT EoSs, and its presence implies the existence of multiple stable critical points [171]. The FE-ANN EoS does not predict this unphysical behavior over the studied range of density and temperature. For completeness, it is also observed that the virial coefficients ( $B_2^*$  and  $B_3^*$ ) are well-behaved, and as expected, they diverge at low temperatures. All of these conditions suggest that the trained FE-ANN EoS is physically consistent. This thermodynamically consistent behavior has also been observed for the SAFT-VR-Mie EoS [166, 169].

From Figure 2.10, it is also observed that the predicted Zeno, Boyle, and Charles curves predicted by FE-ANN EoS are in excellent agreement with published molecular simulation results [166]. The molecular simulation results reveal that the FE-ANN EoS accurately predicts the Boyle and Charles temperature. The Amagat curve predicted by the FE-ANN EoS is mostly in good agreement with the published data except at the high-pressure/low-temperature region, where the correlation of Khrapak et al. [151] suggests the Mie particle freezes. The molecular simulation results reveal that the FE-ANN EoS accurately predicts the Boyle and Charles temperature. The Amagat temperature is usually higher than the other two characteristic temperatures. As a reference, the Amagat temperature for the Lennard-Jones fluid is reported to be  $T^* = 25.15$  [169]. This value is far beyond the considered temperatures used in this work ( $T^* \leq 10$ ). Therefore, its calculation was not attempted. It must be emphasized that the FE-ANN EoS predicts the location and shape of the Brown's characteristic curves. The FE-ANN EoS has seen thermophysical properties related to these curves, but no explicit information about their shape and constraints has been enforced during the training process.

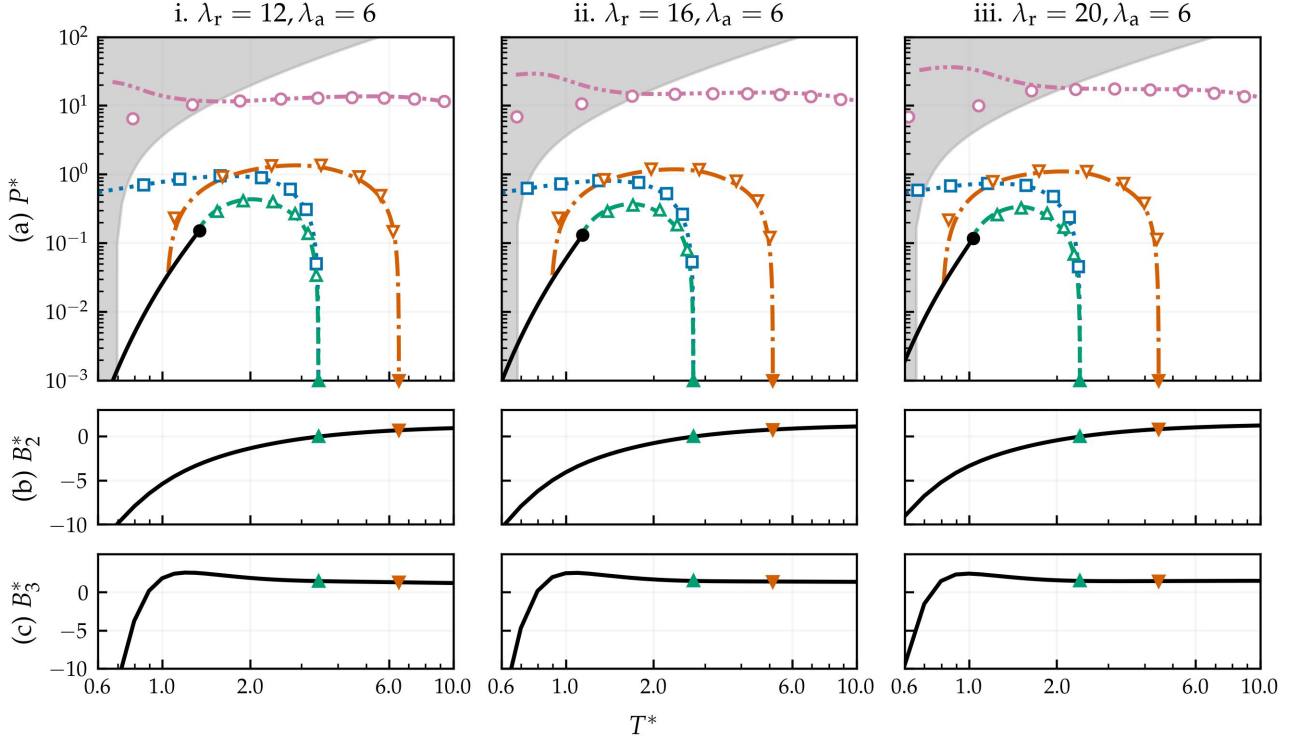


FIGURE 2.10: **Brown's characteristic curves and second virial coefficient of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** (a)  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid). (b)  $\lambda_r = 16, \lambda_a = 6$ . (c)  $\lambda_r = 20, \lambda_a = 6$ . Empty symbols are molecular simulation results obtained by Stephan and Urschel [166]. Pink circles: Amagat curve, orange downward triangles: Charles curve, green upright triangle: Boyle curve, and blue squares: Zeno curve. Lines are filled symbols are results obtained from the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Bottom plot: the black solid line is the third virial coefficient. Middle plot: the black solid line is the second virial coefficient. Top plot: the black solid line is the vaporization pressure, the black filled circle is the critical point, the pink dot-dot-dashed line is the Amagat curve, the orange dot-dashed line is the Charles curve, the green dashed line is the Boyle curve, and the blue dotted line is the Zeno curve. The filled green upright triangle is the Boyle temperature, and the orange downward triangle is the Charles temperature. The grey-shaded area represents the region where freezing is predicted by the correlation of Khrapak et al. [151], see Eq. (2.18).

## 2.6 Conclusions

This chapter presents a framework for developing thermodynamically consistent data-driven equations of state using an artificial neural network (FE-ANN EoS). A key aspect of this framework is the incorporation of physically inspired limits and constraints that force the functional form of the resulting Helmholtz free energy surface to be physically sound.

The capabilities of the "digital" approach are exemplified for the Mie fluid starting from data generated using high-throughput molecular dynamics simulations. It is observed that using only first-order derivative information (e.g., compressibility factor and internal energy) is insufficient to accurately reconstruct the Helmholtz free energy surface of the fluid. The model significantly benefits from being trained using second-order derivative properties, specifically cross-derivative properties, such as the thermal pressure coefficient and adiabatic index. However, it is necessary to quantify the trade-off in the weights assigned to each thermophysical property in the training loss function. It is observed that the training produces a well-behaved EoS when first-order properties have a higher relative weight than second-order properties. This behavior arises because second-order derivative functions are computed from molecular dynamics through the analysis of fluctuations and are inherently much noisier than the computed pressure and internal energy, which are directly during the simulation.

The FE-ANN EoS accurately describes the thermophysical properties on which it was trained. More importantly, due to the physical insights embedded in its formulation, the FE-ANN EoS can also predict thermophysical properties for which it was not trained. The FE-ANN EoS "discovered" an unstable region that allowed the prediction of vapor-liquid equilibria and the existence of a critical point. Additionally, the internal consistency of the FE-ANN EoS was successfully assessed through Brown's characteristic curves. It was recognized that the predicted characteristic curves exhibit the correct physical behavior described by Deiters and De Reuck [89]. The low-pressure regime of these curves also demonstrated the correct modeling of important temperature limits, such as the Boyle and Charles temperatures.

The FE-ANN EoS framework has proven to be a flexible, data-driven strategy capable of accurately learning the Helmholtz free energy surface of a target molecule (or pseudo-molecule). This framework is a promising alternative to classical empirical and theoretical-based equations of state. The potential of this framework is significant, not only in terms of the result accuracy but also in the speed at which simulation results can be integrated. Simulation data can be generated in a few months, and the FE-ANN EoS can be trained in just a few hours. The produced FE-ANN EoS performs on par with the monomer term of the latest version of the Statistical Associating Fluid Theory (SAFT) [37], which was the culmination of over three decades of research [172].

Finally, the FE-ANN EoS framework is built upon derivative data and well-established thermodynamic principles. The use of the FE-ANN EoS is here showcased for the Mie fluid. However, nothing prevents using the FE-ANN EoS for other contributions. The framework is suitable for studying regions where existing theory is likely to fail, such as under extreme temperature and density conditions. It must be noted that under these conditions, high-quality molecular simulation data is also more challenging to compute. Finally, the FE-ANN EoS can also be applied to other fluids, such as chain molecules or

fluids with more complex interactions, including dipolar or quadrupolar moments, point electrostatics, polarizability, induction, and molecular association.

## Chapter 3

# On the continuous modeling of fluid and solid states

### 3.1 Introduction

In 1874, J. C. Maxwell [173], inspired by the graphical thermodynamic methods of J. W. Gibbs [174], built a thermodynamic surface model that used volume ( $V$ ), entropy ( $S$ ), and internal energy ( $U$ ) as coordinates. A physical representation and a 2D projection of this thermodynamic surface are shown in Figure 3.1 (a) and (b) [175, Figure 26d]. The Maxwell thermodynamic surface schematically shows how the thermodynamic behavior of the different aggregation states (i.e., vapor, liquid, and solid states) of a simple substance could be represented by a single continuous surface. According to J.S. Rowlinson [176], this concept of continuity between states partially inspired J. D. van der Waals [28] during his doctoral studies and contributed to the development of his now well-known continuous equation of state (EoS) for fluid phases. The van der Waals EoS accounts for attractive and repulsive forces, which describe mechanically stable dilute (vapor) or dense (liquid) phases, depending on the density and temperature. More importantly, the interplay between these two forces results in a local maximum and a local minimum along a pressure isotherm at subcritical conditions. These stationary points, also known as spinodal points, define a region of mechanical instability where the compressibility is negative (i.e.,  $(\partial P / \partial \rho)_T < 0$ ).

This unstable behavior, also known as the van der Waals loop, has long been interpreted as a sign of the coexistence of two phases and is observed in the Maxwell thermodynamic surface. As observed in Figure 3.1, the thermodynamic surface of Maxwell not only suggests a continuous fluid transition but also that the solid-fluid transition could be modeled continuously. In this figure, several isobars are highlighted, namely  $P_2$ ,  $P_3$ ,  $P_4$ , and  $P_5$ . It can be observed that these isobars continuously describe different aggregation states and that, depending on the temperature, they might exhibit various types of phase equilibria. For example,  $P_2$  passes through the Solid-Vapor Equilibria (SVE) boundaries. Similarly, the  $P_4$  isobar can exhibit vapor-liquid equilibria (VLE) or solid-liquid equilibria (SLE) depending on the temperature.  $P_3$  corresponds to the triple pressure and indicates the coexistence of solid, liquid, and vapor phases (SLVE) at the triple temperature. Finally,  $P_5$  refers to the critical pressure, which passes through the endpoint of the VLE phase envelope. This critical isobar also crosses the SLE boundaries.



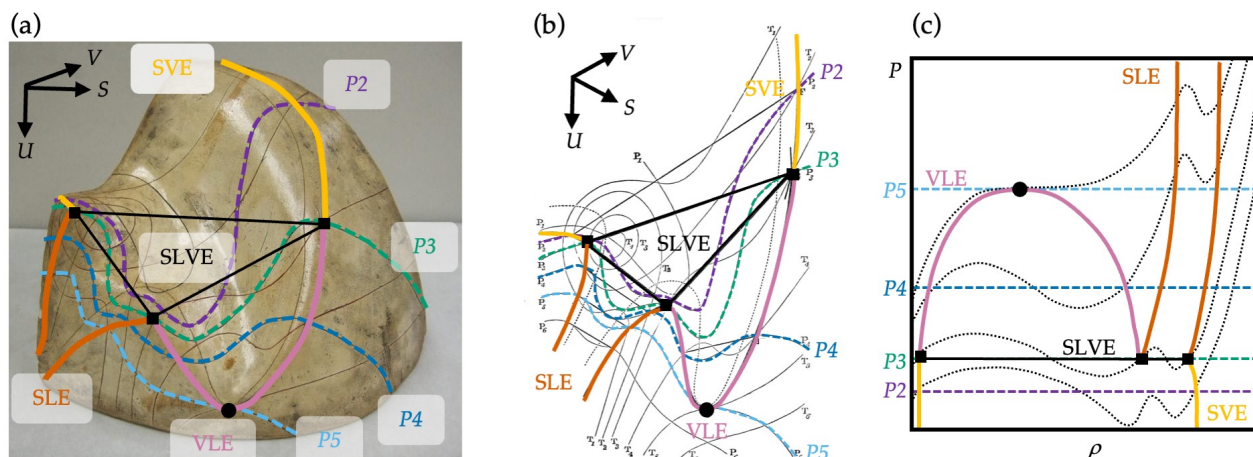


FIGURE 3.1: **Maxwell thermodynamic surface.** (a) Photograph of Maxwell thermodynamic surface (courtesy of Yale Peabody Museum, YPM HST 290012). (b) 2D projection of Maxwell thermodynamic surface [175, Figure 26d]. (c) Schematic representation of the Maxwell thermodynamic surface in the  $P - \rho$  diagram. The coordinates for the thermodynamic surface are internal energy ( $U$ ), volume ( $V$ ), and entropy ( $S$ ). The filled black circle is the critical point, the black solid line with filled black squares represents the Solid-Liquid-Vapor Equilibria (SLVE, triple point), the solid pink lines are the Vapor-Liquid Equilibria (VLE), the orange solid lines represent the Solid-Liquid Equilibria (SLE), the yellow solid lines refer to the Solid-Vapor Equilibria (SVE). The dashed lines represent isobars:  $P_2$  (purple),  $P_3$  (green),  $P_4$  (blue) and  $P_5$  (skyblue) with  $P_2 < P_3 < P_4 < P_5$ . The dotted lines in the right figure refer to corresponding pressure isotherms. Figure adapted from Ref. [1].

For clarity, the continuous pressure behavior observed in the Maxwell thermodynamic surface can be mapped onto a more familiar pressure-density ( $P - \rho$ ) projection, as shown in Figure 3.1 (c). In this projection, the isotherms associated with the phase coexistence are more relevant than the highlighted isobars, as the latter correspond to horizontal lines. These isotherms are represented by dotted black lines in the figure. This figure shows how, along the van der Waals loops for VLE, other solid-fluid equilibria types, such as SLE, SVE, and the triple point (SLVE), also require this mechanically unstable behavior.

Although these thermodynamic aspects are well understood, applying these ideas in such a way that makes them useful for engineering purposes (e.g., in process design) has not been straightforward. The crux of the matter is the solution of the underlying statistical mechanics, which relates the description of the forces between molecules to the macroscopically observed behavior. It is now accepted that the analytical solution of the statistical mechanics equations is out of reach except for the simplest of models, such as the ideal gas. Despite this, during the 20<sup>th</sup> century, the molecular modeling community has primarily focused on developing approximations and semi-empirical expressions to improve the modeling of fluids using EoSs [31, 42]. Among these are modifications of the van der Waals EoS [29, 30, 32, 34], the application of more rigorous theories to account for the attractive and repulsive interactions between molecules [37, 114, 119], as well as theories to model more complex interactions, such as chain formation and hydrogen bonding [35].

Compared to the modeling of fluid phases, the development of EoSs to describe the solid phase has been limited. The solid phase has been commonly modeled using an independent model linked to a suitable fluid model. In this fashion, the properties of the solid phase are obtained by starting from a fluid phase and performing a thermodynamic cycle through the phase changes [177]. Other



approaches consider a different theory to describe the solid phase. Among these are the lattice [178–180] and perturbation [164, 181] theories. In an alternative approach, Yokozeki [182], and later Mo et al. [183], introduced an empirical modification to a cubic EoS to handle the solid phase within a single model. This is achieved by enforcing a divergence of the EoS at the solid-fluid transition. This divergence allows computing SLE using Maxwell’s area reconstruction. Despite the practicality of this family of EoSs to model the entire phase diagram of pure molecules and mixtures, these EoSs do not obey an underlying molecular model. Moreover, Lee and Yoo [184] have pointed out that these EoSs exhibit an unphysical behavior, where a negative pressure dominates at high densities. This behavior contradicts molecular simulations and implies that tension dominates rather than compression at high densities.

Even though the idea of a continuous behavior of matter of Maxwell and Gibbs has been around for about 150 years, it has been scarcely discussed in the literature. In 1935, J. Frenkel [185] discussed structural similarities between liquid and the crystalline state near the melting point. Similarly, as observed in the Maxwell surface, J. Frenkel speculated that a pressure isotherm should exhibit a van der Waals loop in the fluid-solid transition region. Additionally, in molecular simulation strategies to compute SLE [164, 186, 187], it is recognized that metastable liquid and solid phases are accessible inside the SLE boundaries. The computer simulation of such metastable states shows the existence of the minimum and maximum pressure along an isotherm, again suggesting the existence of the van der Waals loop in this region. In a related context, Troncoso and Cerdeiriña [188] used a three-state Ising model to continuously model solid and fluid phases and showed a van der Waals loop at the solid-fluid transition. Despite the evidence that vapor, liquid, and solid states can be modeled using a single continuous equation of state, no such model has been developed yet.

In this chapter, the general Helmholtz Free Energy Equation of State based on Artificial Neural Networks (FE-ANN EoS) framework [3, 87] is used to develop an equation of state that continuously models the phase space across the entire density range, from the dilute gas to the crystalline phase. The results presented in this chapter are based on an already published article (Gustavo Chaparro and Erich A. Müller. “Development of a Helmholtz free energy equation of state for fluid and solid phases via artificial neural networks”. *Communications Physics* 7.1 (2024), p. 406. DOI: 10.1038/s42005-024-01892-3, Ref. [1]). The results presented here have been updated by training the EoS on a more extensive database and incorporating extra thermodynamically inspired constraints.

Contrary to the Maxwell thermodynamic surface ( $V$ - $S$ - $U$ ), the Helmholtz free energy as a function of density and temperature is chosen as a more appropriate thermodynamic potential. The FE-ANN EoS framework is based on ANNs, which are universal function approximators [53]. The FE-ANN EoS does not assume a mathematical functionality, but instead, it fits the Helmholtz surface that best describes thermophysical properties data while fulfilling well-accepted thermodynamic principles and constraints [116, 125]. The previous chapter explained and validated the FE-ANN EoS framework by generating an accurate and thermodynamically consistent EoS for the Mie fluid [90, 91]. The performance of this FE-ANN EoS is on par with the SAFT-VR-Mie EoS [37], which is arguably the best available theoretically-based EoS for this pseudo fluid. The domain of this framework is extended to handle the solid phase. The FE-ANN(s) EoS (an extra “(s)” is added to emphasize that this EoS has also been trained with solid-like data) is formulated to model the residual Helmholtz free energy and

exactly fulfills the ideal gas limit ( $\rho \rightarrow 0$ ) and Maxwell's relations.

$$A^{\text{res}} = \underbrace{\text{ANN}(\chi, \rho, \frac{1}{T}) - \text{ANN}(\chi, \rho = 0, \frac{1}{T})}_{\text{Model to be trained}} \quad (2.9b)$$

Where  $A^{\text{res}}$  is the residual Helmholtz free energy,  $\rho$  is the density,  $T$  is the temperature, and  $\chi$  are the molecular descriptors of the system (e.g., the molecular parameters of an interaction potential).

The general workflow for developing the FE-ANN(s) EoS is shown in Figure 3.2. This process is similar to the one explained in Figure 2.1 of Chapter 2. The first step of this process, Figure 3.2 (a), relies on Molecular Dynamics (MD) simulations to generate a database to train the FE-ANN(s) EoS. The MD simulations are performed at stable vapor, liquid, and solid state conditions of the phase space. These simulations provide the macroscopical thermophysical properties corresponding to a given state point of a collective of particles interacting on a determined Hamiltonian. Technical details about the database generation follow the ones explained in Section 2.4. In the second step, shown in Figure 3.2 (b), the FE-ANN(s) EoS explicitly models the residual Helmholtz free energy of the system of interest. The derivatives of the Helmholtz free energy are related to the available thermophysical properties as explained in Section 2.2. The FE-ANN(s) EoS is formulated and trained as described in Section 2.3. Upon training, the FE-ANN(s) EoS interprets the reference data and provides a mathematical model that accurately assesses the phase space. This formulation provides a simple but effective pathway for modeling the macroscopic properties of a given force field. As shown in Figure 3.2 (c), the FE-ANN(s) has learned the Helmholtz free energy of the entire phase space, allowing the EoS to model thermophysical properties of vapor, liquid, and solid phases, as well as to predict phase equilibria (SVE, SLE, VLE, and SLVE).

Building on the results of Chapter 2, the FE-ANN(s) EoS framework is also showcased for systems interacting through a truncated Mie potential [90, 91]. The truncated Mie potential is formulated as follows.

$$\mathcal{U}^{\text{Mie}} = \begin{cases} \mathcal{C}^{\text{Mie}} \epsilon \left[ (\sigma/r)^{\lambda_r} - (\sigma/r)^{\lambda_a} \right] & \text{if } r \leq 5\sigma \\ 0 & \text{if } r > 5\sigma \end{cases} \quad (1.1)$$

$$\mathcal{C}^{\text{Mie}} = \frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}$$

Where  $\mathcal{U}^{\text{Mie}}$  refers to the interaction energy between two particles,  $\epsilon$  is the energy scale,  $\sigma$  is the length scale, which is related to the diameter of the particle, and  $r$  is the center-to-center distance between two Mie particles. Finally,  $\lambda_r$  and  $\lambda_a$  control the steepness of the repulsive and attractive interactions. Similarly to the FE-ANN EoS generated in Section 2.5, the FE-ANN(s) EoS is formulated using reduced units (indicated by a "\*" superscript). The reader is referred to Table 2.1 for further details about reduced units. When using these units, the Mie particle is fully described by the interplay between the repulsive and attractive exponents. Moreover, due to the conformality of the exponents for fluid equilibrium properties [92], the Mie particle can be characterized by its van der Waals parameter

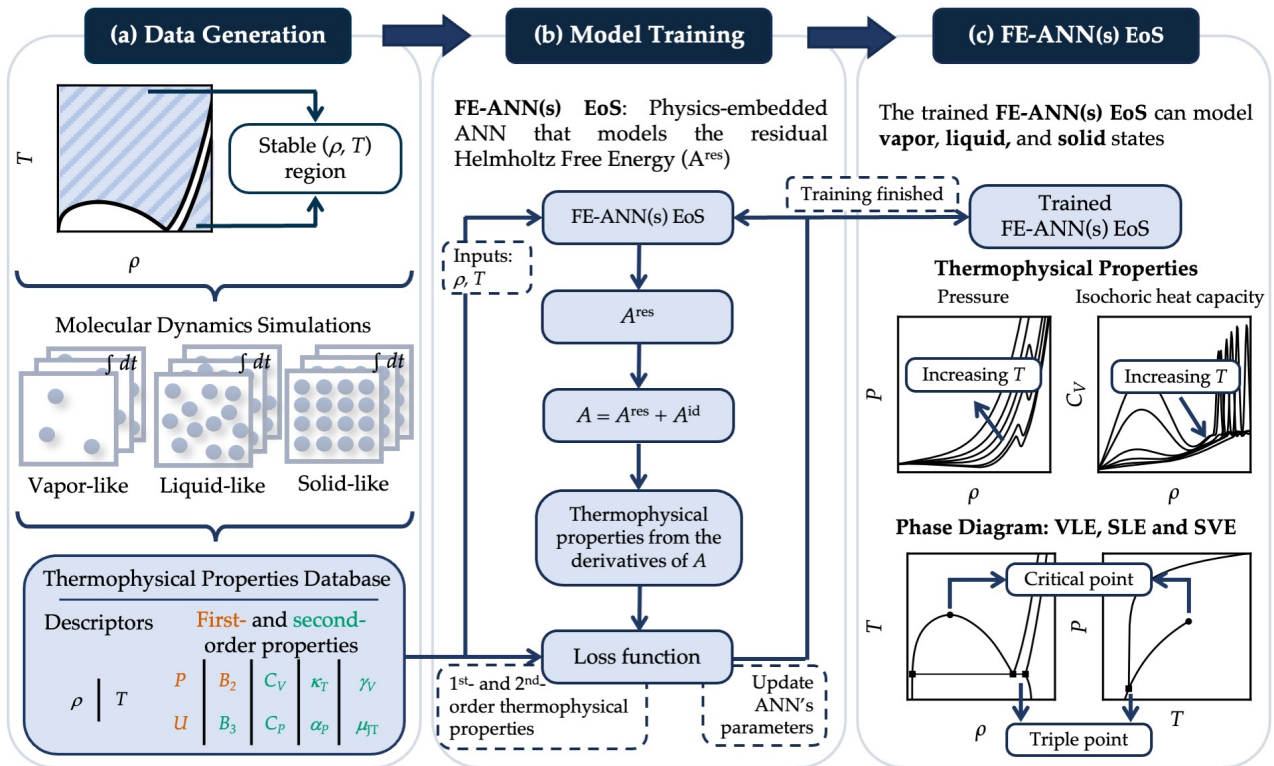


FIGURE 3.2: **Approach to develop a continuous equation of state for fluid and solid phases.** The approach consists of three main steps. (a) Firstly, molecular dynamics simulations are run to produce a thermophysical properties database. The produced database includes first-order properties such as the pressure ( $P$ ), second virial coefficient ( $B_2$ ), and internal energy ( $U$ ), as well as second-order thermophysical properties, such as the isochoric heat capacity ( $C_V$ ), thermal pressure coefficient ( $\gamma_V$ ), isothermal compressibility ( $\kappa_T$ ), thermal expansion coefficient ( $\alpha_P$ ), isobaric heat capacity ( $C_P$ ), Joule-Thomson coefficient ( $\mu_{JT}$ ) and third virial coefficient ( $B_3$ ). (b) Secondly, this database is used to train the FE-ANN(s) EoS, which is an artificial neural network-based equation of state that models the residual Helmholtz free energy ( $A^{\text{res}}$ ). The residual Helmholtz free energy is trained based only on its first- and second-order derivatives. (c) Finally, the trained FE-ANN(s) EoS can be used to model the thermophysical properties of vapor, liquid, and solid phases, as well as to predict phase equilibria (SVE, SLE, VLE, and SLVE). Figure adapted from Ref. [1].

( $\alpha_{\text{vdw}}$ ), Eq. (2.21). Hence, here the  $\alpha_{\text{vdw}}$  parameter is used as the descriptor  $\chi$  of Eq. (2.9b).

$$\alpha_{\text{vdw}} = \underbrace{\frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}}_{C^{\text{Mie}}} \left[ \frac{1}{\lambda_a - 3} - \frac{1}{\lambda_r - 3} \right] \quad (2.21)$$

The rest of this chapter is structured as follows. Section 3.2 details the extension of the Mie particle database to include solid phase data. Then, in Section 3.3, the FE-ANN(s) EoS for the Mie particle is developed. This section also checks the performance, correctness, and predictive capabilities of the FE-ANN(s) EoS. Finally, the main conclusions of this chapter are drawn in Section 3.4.

## 3.2 Mie particle thermophysical properties database

As explained in Figure 3.2 (a), the construction of the FE-ANN(s) relies on thermophysical properties of vapor, liquid, and solid states. In the case of the Mie particle, the thermophysical properties of an extensive region of the fluid phase space have already been curated in Chapter 2. For this reason, this section focuses on computing the thermophysical properties of crystalline face-centered cubic (FCC) Mie particles using molecular simulation methods. The Mie particle exhibits other crystal structures under more extreme conditions (e.g., higher densities or lower temperatures) [164, 189–191], but for simplicity, these are not considered here. The thermophysical properties described here are reported in reduced units, as indicated by a "\*" superscript. Refer to Table 2.1 for further details about reduced units.

All Mie particle molecular simulations in this section use a cutoff radius of  $5\sigma$  without tail corrections, a timestep of  $\Delta\tau^* = 0.002\tau^*$ , and periodic boundary conditions. Simulations performed in the NVT ensemble use the Nosé-Hoover thermostat [150] with a time constant of  $100\Delta\tau^*$ . Similarly, NPT simulations use the Nosé-Hoover thermostat-barostat [153] with time constants of  $100\Delta\tau^*$  and  $1000\Delta\tau^*$  for the thermostat and barostat, respectively. The thermophysical properties of each state point are obtained from a single simulation replica. Simulations are performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [147] (stable release June 2022).

### 3.2.1 Mie particle solid-liquid equilibria (SLE)

As shown in Figure 3.2 (a), the computation of single-phase solid-like thermophysical properties requires knowledge of the phase boundaries (VLE and SLE). Intuition about these boundaries guides the production of single-phase data by avoiding unstable states that will likely fail and produce misleading results. In Chapter 2, the VLE boundaries of the Mie particle were obtained using the temperature-quench method [148], and the liquid side of the SLE boundaries were approximated using the correlation of Khrapak et al. [151]. This correlation, Eq. (2.18), does not provide insights into the solid branch of the SLE. For this reason, the SLE of the Mie particle is fully computed here using the freeze method [164]. This method will be briefly described here. The reader is referred to the thesis of Ramrattan [164] for further details about the freeze method.

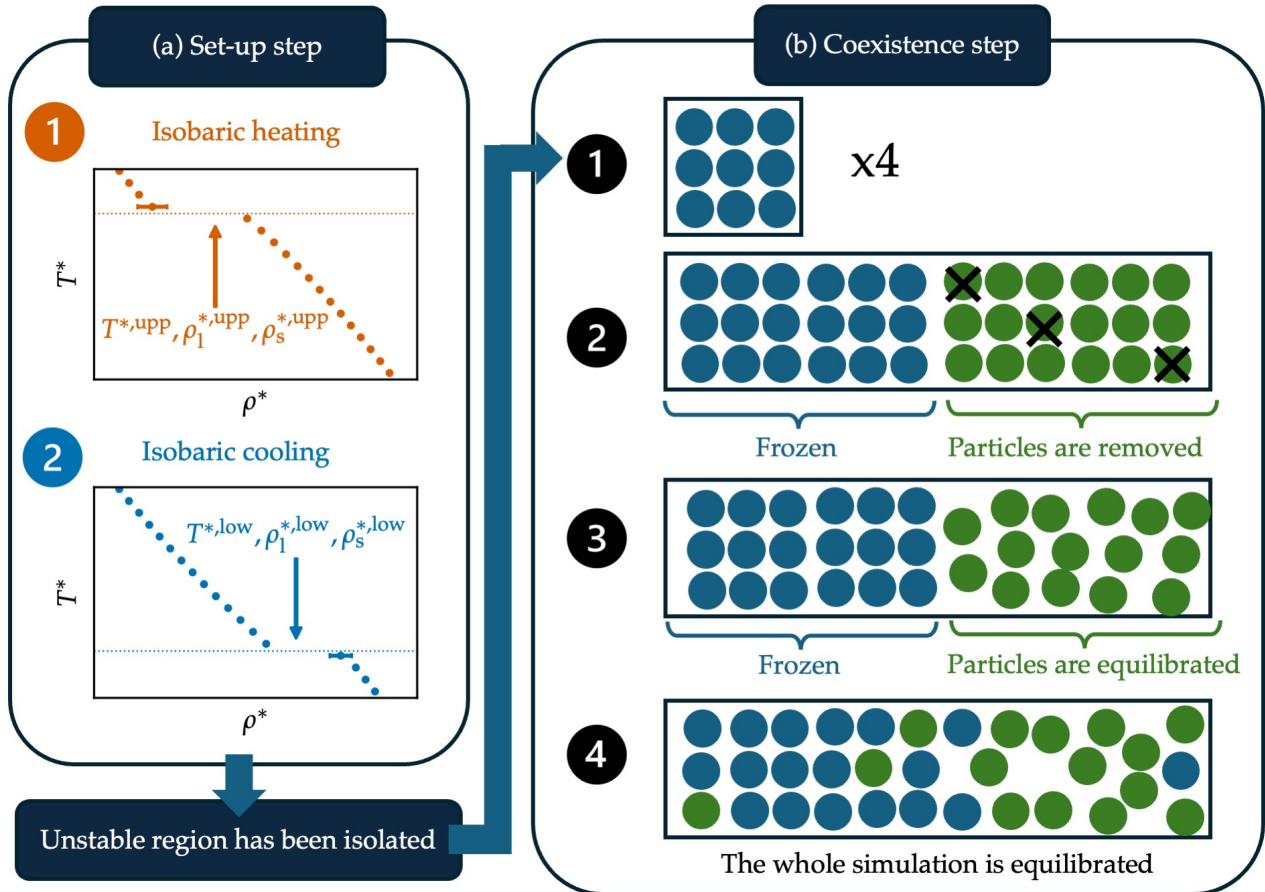


FIGURE 3.3: **Summary of the freeze method to compute SLE in molecular dynamic simulations.** **(a)** Set-up step. In this step, an initial solid is heated at constant pressure (i.e., isobarically) until it melts. From this process, an upper limit ( $T^{*,upp}, \rho_1^{*,upp}, \rho_s^{*,upp}$ ) for the SLE is obtained. Then, the melted fluid is isobarically cooled until freezing. From this process, a lower limit ( $T^{*,low}, \rho_1^{*,low}, \rho_s^{*,low}$ ) for the SLE is obtained. The SLE region is isolated from the set-up step, and good initial guesses are available to perform a simulation where both liquid and solid phases coexist. **(b)** Coexistence step. First, the initial solid initial solid configuration is replicated in the z-direction. Secondly, a solid and a liquid region are defined. Particles are deleted from the fluid region to achieve an unstable global density. Then, the simulation is equilibrated by maintaining the solid particles frozen. Finally, the entire simulation is equilibrated at the desired temperature. Statistics of the system are accumulated in this last step.

TABLE 3.1: Initial guesses for the coexistence step in the freeze method.

| Melting temperature ( $T^{*,\text{mel}}$ ) | Global density ( $\rho_{\text{global}}^*$ )                                                         | Note         |
|--------------------------------------------|-----------------------------------------------------------------------------------------------------|--------------|
| $T^{*,\text{upp}}$                         | $(\rho_s^{*,\text{upp}} + \rho_l^{*,\text{upp}})/2$                                                 | <sup>a</sup> |
| $T^{*,\text{upp}}$                         | $(\rho_s^{*,\text{upp}} + \rho_l^{*,\text{low}})/2$                                                 | <sup>b</sup> |
| $T^{*,\text{upp}}$                         | $(\rho_s^{*,\text{low}} + \rho_l^{*,\text{low}})/2$                                                 | <sup>a</sup> |
| $(T^{*,\text{upp}} + T^{*,\text{low}})/2$  | $(\rho_s^{*,\text{upp}} + \rho_l^{*,\text{upp}} + \rho_s^{*,\text{low}} + \rho_l^{*,\text{low}})/4$ | <sup>a</sup> |

<sup>a</sup> Initial guess proposed in this work. <sup>b</sup> Initial guess suggested by Ramrattan [164].

A schematic representation of the freeze method is shown in Figure 3.3. This method consists of two main steps: the set-up and the coexistence steps. The set-up step, Figure 3.3 (a), is needed to generate high-quality initial guesses of the SLE boundaries. This step is required to perform a simulation at a global unstable density. Contrarily to the simulation of VLE [148], the unstable region to achieve SLE is narrow due to the similar densities of the liquid and solid phases at equilibrium conditions. In this step, an initial simulation in the solid state is isobarically heated until melting and then cooled down until freezing. The phase changes are observed in this process by the "jumps" in the simulation density. It is recognized that this process exhibits hysteresis. Therefore, the melting and freezing conditions do not exactly match. Nevertheless, the melting temperature and an unstable density region have been isolated from this step, making it suitable to initialize simulations where liquid and solid phases coexist. Four possible initial guess combinations are considered here and detailed in Table 3.1. These multiple initial guesses are needed to obtain at least one successful SLE simulation in which both liquid and solid phases are well distributed.

In this work, the set-up step, Figure 3.3 (a), is performed by firstly initializing an FCC cubic simulation box with 2048 particles at a high density (i.e.,  $\rho^* = 1$ ). Initially, the simulation is equilibrated at a low temperature ( $T^* = 0.65$ ) in the NVT ensemble for  $10^6$  time steps. The pressure of the system is computed from this step and used to perform the isobaric heating and cooling process. This simulation is quickly heated isobarically to estimate the temperature ( $T^{*,\text{max}}$ ) where the initial solid melts. The initial solid simulation is heated from  $T^* = 0.65$  to  $T^* = T^{*,\text{max}} + 0.5$  in 20 equidistant steps. The same steps are repeated when isobarically cooling down the system. For each of these heating/cooling steps, the simulation is run for  $2.5 \cdot 10^5$  time steps in the NPT ensemble. The system statistics are collected every 1000 time steps. The equilibrated density, pressure, and temperature for each heating/cooling step are obtained by averaging the sampled data.

In the coexistence step of the freeze method, Figure 3.3 (b), the target is to perform a simulation in which both solid and fluid phases coexist. This procedure relies on having high-quality initial guesses of the melting conditions obtained from the set-up step. This coexistence simulation is performed in the NVT ensemble using a time step of  $\Delta\tau^* = 0.002\tau^*$  and considers the following steps.

1. A stable solid phase near the melting conditions is obtained from the set-up step. It is recommended to select the solid configuration used to compute the average global density. See Table 3.1.

2. The selected solid configuration is replicated four times in the z-axis. At this step, the simulation consists of  $4 \times 2048 = 8192$  particles.
3. The simulation is divided into two regions. One-half of the simulation, corresponding to the solid phase, is "frozen" (i.e., particles are not allowed to move). For the other half, corresponding to the liquid phase, particles are removed to achieve a global density ( $\rho_{\text{global}}^*$ ) that lies inside the unstable region. This global density is obtained from the set-up step. At this step, the simulation consists of approximately 8000 particles. The liquid phase and the pressure across the coexisting phase are equilibrated at this step.
4. The simulation is equilibrated at the desired melting temperature ( $T^{*,\text{mel}}$ ) for  $10^6$  time steps in the NVT ensemble. The solid phase is still frozen at this step.
5. The solid phase is unfrozen, and the entire simulation is equilibrated at the desired melting temperature ( $T^{*,\text{mel}}$ ) for  $10^6$  time steps in the NVT ensemble.
6. The entire simulation is integrated in the NVT ensemble at the melting temperature ( $T^{*,\text{mel}}$ ) for further  $8 \cdot 10^6$  time steps. In this production step, statistics of the system are accumulated every 1000 time steps.

The success of the coexistence step is assessed by analyzing the density profile of the simulation. If unsuccessful, the density profile will exhibit a poor phase distribution or contain only a single phase (liquid or solid-like). Those simulations are disregarded from further analysis. If successful, the simulation will exhibit coexisting liquid and crystal phases. The equilibrium densities ( $\rho_l^*$  and  $\rho_s^*$ ) are obtained by first identifying the liquid and solid regions in the density profile and then averaging their values accordingly. Similarly, the melting pressure ( $P^{*,\text{mel}}$ ) is taken as the normal pressure component of the pressure tensor ( $P_{zz}^*$ ).

The freeze method is used to determine the SLE boundaries of the target Mie particles ( $\lambda_r \in [7, 34]$  and  $\lambda_a = 6$ ) in this work. The SLE is computed within the regions of interest of this work (i.e.,  $\rho^* \lesssim 1.25$  and  $T^* \lesssim 10$ ). The freeze method uses the initial solid structure obtained from the set-up stage. Therefore, a single simulation is performed per initial guess.

### 3.2.2 Mie particle thermophysical properties computation and validation

Mie particles in the outer vicinity of the SLE boundaries are simulated to generate the first-order ( $P^*$  and  $U^*$ ) and second-order ( $C_V^*$ ,  $\gamma_V^*$ ,  $\kappa_T^*$ ,  $\alpha_p^*$ ,  $C_p^*$ , and  $\mu_{JT}^*$ ) thermophysical properties. The general procedure to compute these thermophysical properties follows the details covered in Section 2.4.1.

A cubic simulation box containing 4000 Mie particles is initialized as an FCC crystal. This box is significantly larger than needed to produce accurate results for a single-phase simulation due to the implementation of the periodic boundary conditions. Nevertheless, this system size is chosen to approach the thermodynamic limit as closely as possible. This simulation is firstly equilibrated in the NVT ensemble for  $2 \cdot 10^5$  time steps, then statistics of the systems are accumulated for further  $4 \cdot 10^6$  time steps in the production stage. The equilibrated NVT configuration is then integrated in the NPT ensemble. Similarly, as in the NVT step, the equilibration and production stages are performed for  $2 \cdot 10^5$  and  $4 \cdot 10^6$  time steps. The system statistics are collected every 1000 time steps

in each production stage. The thermophysical properties are obtained by analyzing the collected data. First-order properties, such as the pressure ( $P^*$ ) and internal energy ( $U^*$ ), are averaged from the collected data in the NVT ensemble. The LAMMPS code directly computes these properties. Second-order properties, such as the isochoric heat capacity ( $C_V^*$ ) and thermal pressure coefficient ( $\gamma_V^*$ ), are obtained by analyzing the fluctuations of the NVT simulation using Eqs. (2.19). Similarly, the isobaric heat capacity ( $C_P^*$ ), thermal expansion coefficient ( $\alpha_P^*$ ), and isothermal compressibility ( $\kappa_T^*$ ) are obtained by analyzing the fluctuations of the NPT simulation using Eqs. (2.20). Finally, the Joule-Thomson coefficient ( $\mu_{JT}^*$ ) is obtained using Eq. (2.3f) and the computed  $C_P^*$  and  $\alpha_P^*$ .

The quality of the computed thermophysical properties is assessed using the sanity tests described in Section 2.4.2. These tests verify that first-order properties computed in the NVT and NPT ensembles agree within a given tolerance, that only second-order properties computed with a low relative standard deviation ( $100|\sigma_x/\langle x \rangle| < 5$ ) are reported, and that there are no evident outliers in the database. Only data points that pass these sanity tests are included in the Mie particle database.

### 3.2.3 Mie particle database description

The Mie particle database considers systems with repulsive exponent  $\lambda_r \in [7, 34]$  and  $\lambda_a = 6$ . The final database consists of VLE, SLE, and thermophysical properties of the Mie particle. This thermophysical property database consists of the fluid database described in Section 2.4 plus approximately 4000 data points of thermophysical properties of FCC Mie crystals. The ranges and number of data points for each thermophysical property are detailed in Table 3.2.

The data points distribution of the Mie particle database is shown in Figure 3.4. For convenience, the **fluid data is shown in blue** and **solid data is shown in orange**. Compared to Figure 2.3, solid-phase data points populate the high-density region of the phase space. The phase distribution of the database using the temperature inverse is Figure 3.4 (b). It is observed that conveniently high-temperature data points are closely distributed. However, this also results in a sparser distribution of low-temperature data points. Finally, as shown in Figure 3.4 (c), each repulsive exponent includes approximately 800 to 1000 data points.

## 3.3 Continuous Artificial Neural Network based Equation of State for fluid and solid states (FE-ANN(s) EoS) for the Mie particle

In this section, the FE-ANN(s) EoS is trained to "learn" the residual Helmholtz free energy of Mie particles as obtained discretely from molecular simulations. The database includes MD data of the thermophysical properties of single phase fluid (gas, liquid, and supercritical) states, along with the stable Face Centered Cubic (FCC) crystal. No explicit information about the phase coexistence points is provided, as these are expected to be naturally determined from the resulting Helmholtz surface. In the FE-ANN(s) EoS, the Mie particle is described using the van der Waals parameter ( $\alpha_{vdw}$ ), Eq. (2.21), which varies from 0.53 ( $\lambda_r = 34$  and  $\lambda_a = 6$ ) to 1.47 ( $\lambda_r = 7$  and  $\lambda_a = 6$ ). In this version of the



TABLE 3.2: Mie particle database description.<sup>a</sup>

| Descriptor                             | Symbol         | # of data points | Min value | Max value |
|----------------------------------------|----------------|------------------|-----------|-----------|
| Repulsive exponent <sup>b</sup>        | $\lambda_r$    | –                | 7         | 34        |
| Attractive exponent <sup>b</sup>       | $\lambda_a$    | –                | 6         | 6         |
| van der Waals parameter <sup>c</sup>   | $\alpha_{vdw}$ | –                | 0.53      | 1.47      |
| Density                                | $\rho^*$       | –                | 1.00e-06  | 1.32      |
| Temperature <sup>d</sup>               | $T^*$          | –                | 0.58      | 10.06     |
| Thermophysical Property                |                |                  |           |           |
| Pressure                               | $P^*$          | 25750            | 7.80e-07  | 1.60e+02  |
| Compressibility factor <sup>e</sup>    | $Z$            | 25750            | 1.15e-04  | 32.74     |
| Internal energy                        | $U^*$          | 25750            | -9.22     | 22.70     |
| Second virial coefficient <sup>f</sup> | $B_2^*$        | 3485             | -9.91     | 1.57      |
| Third virial coefficient <sup>f</sup>  | $B_3^*$        | 3462             | -3.61     | 2.60      |
| Isochoric heat capacity                | $C_v^*$        | 21889            | 1.50      | 3.16      |
| Isobaric heat capacity                 | $C_p^*$        | 19876            | 2.50      | 23.98     |
| Adiabatic index <sup>g</sup>           | $\gamma$       | 19648            | 1.05      | 9.50      |
| Thermal pressure coefficient           | $\gamma_V^*$   | 21891            | 1.00e-06  | 19.02     |
| Thermal expansion coefficient          | $\alpha_p^*$   | 21260            | 1.52e-02  | 8.52      |
| Isothermal compressibility             | $\kappa_T^*$   | 21227            | 1.39e-03  | 1.31e+06  |
| Joule-Thomson coefficient              | $\mu_{JT}^*$   | 15313            | -0.50     | 11.85     |

<sup>a</sup> Descriptors and thermophysical properties are reported in reduced units. See Table 2.1. <sup>b</sup> Repulsive and attractive exponent of the Mie potential, Eq. (1.1). <sup>c</sup> van der Waals parameter is a descriptor of a Mie particle. See Eq. (2.21). <sup>d</sup> Temperature range used in molecular dynamics simulations. <sup>e</sup> Compressibility factor,  $Z = P^* / \rho^* T^*$ . <sup>f</sup> Virial coefficients obtained in the temperature range of 0.65 to 39.62. <sup>g</sup> Adiabatic index (heat capacity ratio),  $\gamma = C_p^* / C_v^*$ .

FE-ANN(s) EoS, the residual Helmholtz free energy is formulated as follows.

$$A^{*,res} = \text{ANN}(\alpha_{vdw}, \rho^*, \frac{1}{T^*}) - \text{ANN}(\alpha_{vdw}, \rho^* = 0, \frac{1}{T^*}) \quad (2.22b)$$

The implementation and training of the FE-ANN(s) EoS follows the guidelines given in Sections 2.3 and 2.5. The FE-ANN(s) EoS used in this section consists of 3 hidden layers of 30 neurons activated with the tanh activation function ( $f(x) = \tanh(x)$ ). The output layer is a single neuron using the linear activation function ( $f(x) = x$ ). This architecture is sufficient to accurately describe the target thermophysical properties while keeping the number of trainable parameters low. Further details about the hyperparameters of the FE-ANN(s) EoS can be found in Section A.2 of Appendix A. The FE-ANN(s) EoS is implemented using JAX [138] and Flax [161] and is trained using Optax, which is part of the DeepMind JAX Ecosystem [162].

### 3.3.1 FE-ANN(s) EoS: training and performance

The FE-ANN(s) EoS is trained using the reference thermophysical properties database and physically inspired constraints. The database is split using a 90/10 ratio for training and testing purposes. This split offers sufficient coverage of the phase space in both training and testing datasets. The

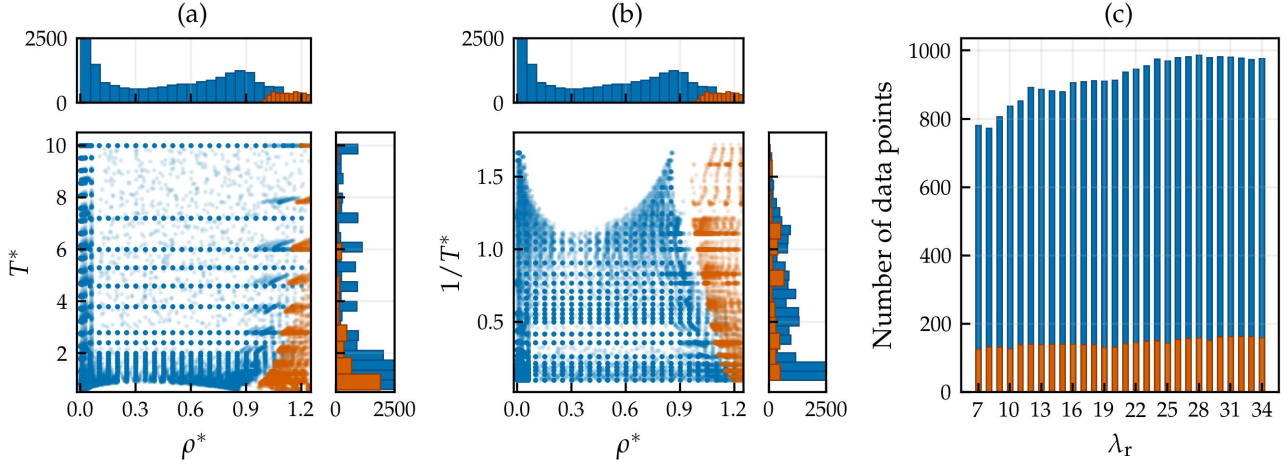


FIGURE 3.4: **Mie particle Thermophysical properties database distribution.** (a) Density-Temperature ( $\rho^* - T^*$ ) distribution. (b) Density-Temperature inverse ( $\rho^* - 1/T^*$ ) distribution. (c) Repulsive exponents ( $\lambda_r$ ) distribution. Blue data points and bars refer to fluid phase data and orange data points and bars refer to solid phase data.

training is performed using a batch size of 32. This small batch size value promotes the exploration of local minima of the loss function surface while training [192]. The training data includes first-order ( $Z = P^*/\rho^*T^*$ ,  $U^*$  and  $B_2^*$ ) and second-order ( $C_V^*$ ,  $\gamma_V^*$ ,  $\rho^*\kappa_T^*$ ,  $\alpha_P^*$ ,  $\gamma$ ,  $\mu_{JT}^*$ , and  $B_3^*$ ) derivative properties. The relationship of these properties with the Helmholtz free energy are given in Section 2.2.

The parameters of the FE-ANN(s) EoS (i.e., weights and biases) are trained to minimize the loss function ( $\mathcal{L}$ ) detailed in Eq. (2.11). This loss function considers two main contributions. The first contribution ( $\mathcal{L}^{\text{MSE}}$ ), Eq. (2.12) measured the square error of the thermophysical properties predicted by FE-ANN(S) EoS with respect to the reference data. In this loss function, each thermophysical property has a relative weight ( $w_i$ ), which determines the importance of the given property in the final EoS. The second contribution to the loss function ( $\mathcal{L}^{\text{pen}}$ ), Eq. (2.13), enforces a correct physical behavior to the FE-ANN(s) EoS. This contribution ensures that the isochoric heat capacity ( $C_V^*$ ), isothermal compressibility (as  $\rho^*\kappa_T^*$ ), and thermal pressure coefficient ( $\gamma_V^*$ ) must be positive for stable phases. The first two of these constraints are required by the *criteria of thermal* ( $C_V^* > 0$ ) and *mechanical* ( $\rho^*\kappa_T^* > 0$ ) stability [116, 125]. The objective of the last constraint is to preclude undesired crossing pressure isotherms (i.e.,  $\gamma_V^* > 0$ ). These penalty functions also include a relative weight to control their impact on the less function.

The training workflow of the FE-ANN(s) EoS follows the two-stage process described in Section 2.5. First, the FE-ANN(s) EoS is trained using only first-order properties for 10000 epochs using a learning rate  $\alpha = 10^{-4}$ . In this first stage, the relative weight ( $w_i$ ) of first-order properties is set to one, while the relative weight of the second-order properties and the penalties are set to zero. This first stage quickly generates a Helmholtz free energy surface that overfits the first-order properties while ignoring the deviations of second-order properties. Despite overfitting the first-order properties, the training parameters (weights and biases) are closer to a local minimum where the correct behavior of second-order properties can be amended. In the second training stage, the FE-ANN(s) EoS is further trained to a total of 50000 epochs using a learning rate  $\alpha = 10^{-6}$ . This small learning rate is required to ensure numerically stable training, as small changes in the weights and biases can lead to significant changes

in second-order properties. In this stage, the relative weight of first-order properties and the penalty functions are set to 1 (i.e.,  $w_i = 1$ ). Following the results of Section 2.5, the relative weight of second-order properties is set to 0.01. This small value is chosen so that second-order properties influence the loss function but do not bias the optimizer to fit the inherent noise of these thermophysical properties.

The deviations of the trained FE-ANN(s) EoS are summarized in the parity plots shown in Figure 3.5. Across the board, most data points lie near the diagonal of each subfigure. This behavior implies that the FE-ANN(s) EoS accurately models the target properties in accordance with the reference data. From Figures 3.5 (a)-(c), it is observed that first-order properties are modeled with lower dispersion than second-order properties. This behavior is expected due to the weighting used in training ( $w_i = 0.01$ ) and the inherent noise of the higher-order derivative properties data. The second-order derivative properties shown in Figures 3.5 (d), (e), and (f) are related to the imposed physical constraints on the EoS. From these figures, it is observed that for both training and testing datasets, the FE-ANN(s) EoS fulfills the *criteria of thermal* ( $C_V^* > 0$ ) and *mechanical* ( $\rho^* \kappa_T^* > 0$ ) stability. Additionally, unphysical crossing pressure isotherms are not observed (i.e.,  $\gamma_V^* > 0$ ). Finally, the FE-ANN(s) EoS performs similarly for both **seen (training)** and **unseen (testing)** data, which implies that the EoS generalizes well within the training phase space. As discussed in Chapter 2, the lack of overfitting results from training with derivative information and imposing constraints to ensure a physically correct Helmholtz free energy surface.

The parity plots shown in Figure 3.5 allow for a quick assessment of the performance and physicality of the FE-ANN(s) EoS. Additionally, it is mandatory to check the distribution of these errors in a density-temperature projection. The squared error distribution of first ( $Z$  and  $U^*$ ), second ( $C_V^*$  and  $\rho^* \kappa_T^*$ ), and cross ( $\gamma_V^*$ ) derivatives of the Helmholtz free energy surface are shown in Figure 3.6. As indicated by the dark purple data points, a large portion of the phase space is accurately modeled by the FE-ANN(s) EoS. There are two distinct regions of higher deviations. First, higher deviations are observed near the phase transitions (i.e., fluid-fluid or solid-fluid). In general, thermophysical properties can change considerably near phase boundaries. The FE-ANN(s) EoS is expected to "learn" these boundaries from single-phase data. Hence, some data points that lie in this region exhibit higher deviations. Secondly, in Figures 3.6 (a), (b), (c), and (e), higher deviations are observed at high densities in the region that is likely to be a solid phase. The magnitude of these properties ( $Z$ ,  $U^*$ ,  $C_V^*$ , and  $\gamma_V^*$ ) increases with density. For that reason, the squared error tends to be higher in that region. For completeness, key statistics of these errors are reported in Table 3.3. This table confirms that the deviations of the fluid and solid phase data points are similar for most target properties. Additionally, this table shows that the median error is lower than the mean for all the thermophysical properties. This statistical behavior suggests that the distribution of errors is skewed toward lower deviations, with higher errors being the exception.

In this section, it has been shown that the FE-ANN(s) EoS fulfills fundamental physical constraints and accurately models the reference thermophysical properties database of the Mie particle. In the next section, the FE-ANN(s) EoS will be used to unveil the continuous nature of the Helmholtz free energy (and its derived thermophysical properties) across the entire phase space.

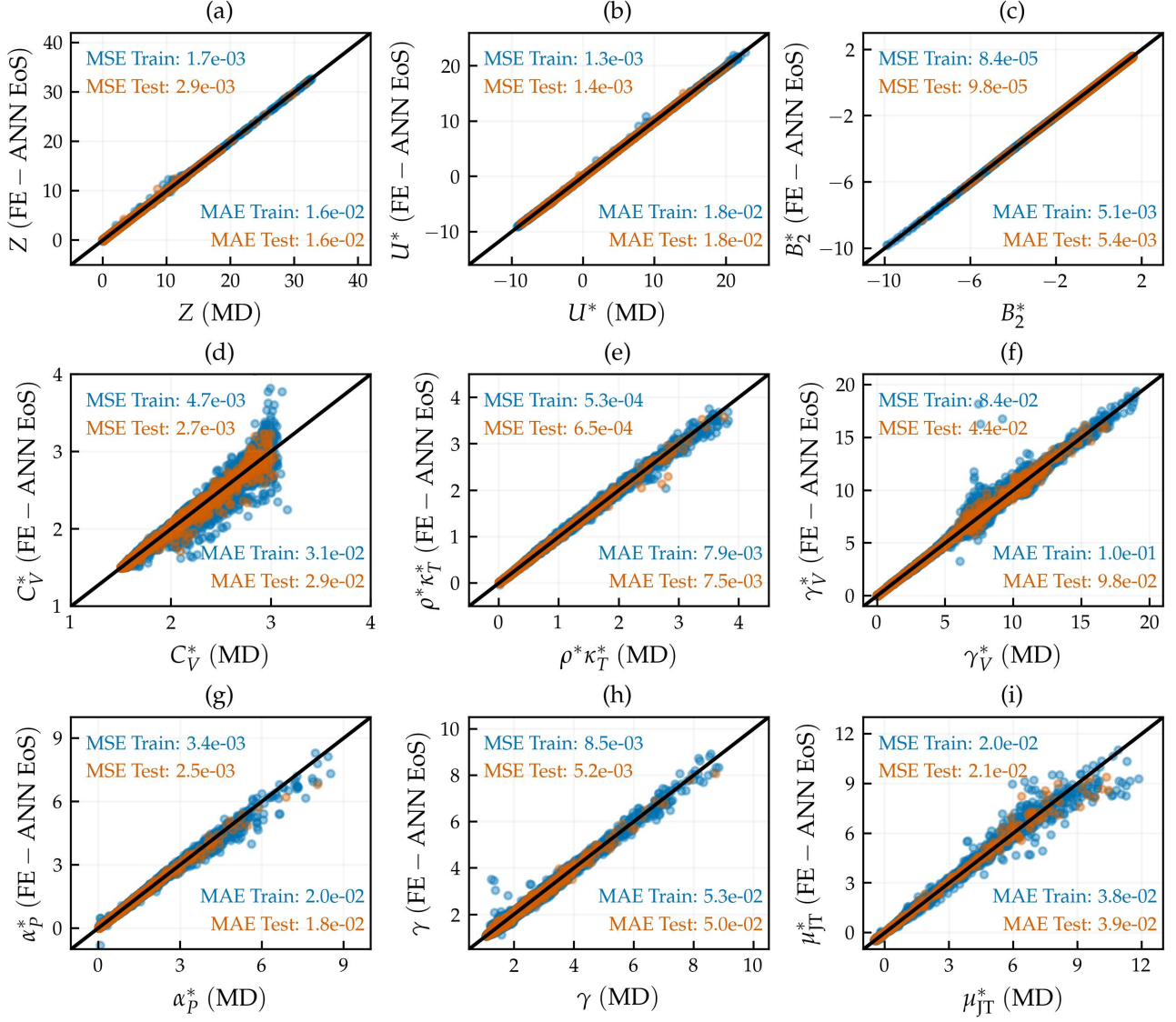


FIGURE 3.5: Parity plot of thermophysical properties computed with FE-ANN(s) EoS with respect to the reference database. (a) Compressibility factor ( $Z$ ). (b) Internal energy ( $U^*$ ). (c) Second virial coefficient ( $B_2^*$ ). (d) Isochoric heat capacity ( $C_V^*$ ). (e) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (f) Thermal pressure coefficient ( $\gamma_V^*$ ). (g) Thermal expansion coefficient ( $\alpha_P^*$ ). (h) Adiabatic index (heat capacity ratio,  $\gamma = C_P^*/C_V^*$ ). (i) Joule-Thomson coefficient ( $\mu_{JT}^*$ ). The reference data points were obtained as explained in Sections 2.4 and 3.2. Thermophysical properties obtained from molecular dynamics simulation include a "(MD)" in the x-axis label. The second virial coefficient was obtained using Eq. (2.17). The thermophysical properties are compared with the results of the FE-ANN(s) EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.01$  for second-order properties. Blue circles refer to the train data points and orange circles refer to the test data set. The reported Mean Squared Error (MSE) is obtained as follows:  $MSE = 1/N_P \sum_j (y_j^{\text{pred}} - y_j^{\text{true}})^2$ . The reported Mean Absolute Error (MAE) is obtained as  $MAE = 1/N_P \sum_j |y_j^{\text{pred}} - y_j^{\text{true}}|$ .

TABLE 3.3: Squared error distribution of selected thermophysical properties as obtained from the FE-ANN(s) EoS.<sup>a, b</sup>

| Property            | All data |         |         |         | Fluid data |         |         |         | Solid data |         |         |         |
|---------------------|----------|---------|---------|---------|------------|---------|---------|---------|------------|---------|---------|---------|
|                     | Min      | Max     | Median  | Mean    | Min        | Max     | Median  | Mean    | Min        | Max     | Median  | Mean    |
| Z                   | 7.0e-14  | 4.3e+00 | 5.3e-05 | 1.7e-03 | 7.0e-14    | 4.3e+00 | 3.3e-05 | 1.8e-03 | 1.3e-09    | 7.5e-02 | 4.9e-04 | 1.6e-03 |
|                     | 4.5e-13  | 3.4e+00 | 5.1e-05 | 2.9e-03 | 4.5e-13    | 3.4e+00 | 3.2e-05 | 3.2e-03 | 6.9e-09    | 4.3e-02 | 5.5e-04 | 1.4e-03 |
| $U^*$               | 3.9e-12  | 4.1e+00 | 1.1e-04 | 1.3e-03 | 3.9e-12    | 4.1e+00 | 7.2e-05 | 1.2e-03 | 1.1e-09    | 5.6e-02 | 5.3e-04 | 2.1e-03 |
|                     | 1.0e-13  | 1.0e+00 | 1.1e-04 | 1.4e-03 | 1.0e-13    | 1.0e+00 | 7.3e-05 | 1.3e-03 | 5.3e-09    | 2.7e-02 | 5.5e-04 | 2.1e-03 |
| $C_V^*$             | 3.2e-20  | 1.4e+01 | 2.7e-04 | 4.7e-03 | 3.2e-20    | 1.4e+01 | 1.8e-04 | 4.3e-03 | 2.4e-10    | 6.8e-01 | 1.3e-03 | 6.7e-03 |
|                     | 4.2e-18  | 1.8e-01 | 2.6e-04 | 2.7e-03 | 4.2e-18    | 1.8e-01 | 1.6e-04 | 2.2e-03 | 6.5e-08    | 1.1e-01 | 1.3e-03 | 5.7e-03 |
| $\rho^* \kappa_T^*$ | 1.8e-16  | 5.4e-01 | 1.4e-06 | 5.3e-04 | 6.7e-15    | 5.4e-01 | 3.2e-06 | 6.4e-04 | 1.8e-16    | 8.2e-05 | 7.2e-08 | 7.7e-07 |
|                     | 3.9e-13  | 3.4e-01 | 1.4e-06 | 6.5e-04 | 1.1e-12    | 3.4e-01 | 3.1e-06 | 7.8e-04 | 3.9e-13    | 2.1e-05 | 8.6e-08 | 5.0e-07 |
| $\gamma_V^*$        | 5.4e-28  | 3.4e+02 | 3.4e-04 | 8.4e-02 | 5.4e-28    | 3.4e+02 | 5.8e-05 | 6.0e-02 | 3.0e-08    | 1.7e+01 | 6.5e-02 | 2.2e-01 |
|                     | 6.0e-26  | 2.8e+00 | 3.0e-04 | 4.4e-02 | 6.0e-26    | 2.8e+00 | 4.5e-05 | 1.8e-02 | 2.0e-06    | 1.7e+00 | 5.5e-02 | 1.8e-01 |
| $\alpha_P^*$        | 1.4e-13  | 2.6e+00 | 3.4e-05 | 3.4e-03 | 3.1e-13    | 2.6e+00 | 4.2e-05 | 4.1e-03 | 1.4e-13    | 1.3e-02 | 1.7e-05 | 1.5e-04 |
|                     | 2.7e-11  | 1.5e+00 | 3.2e-05 | 2.5e-03 | 2.7e-11    | 1.5e+00 | 4.0e-05 | 3.0e-03 | 5.0e-11    | 5.4e-03 | 1.6e-05 | 1.0e-04 |
| $\gamma^*$          | 2.5e-12  | 3.3e+01 | 1.3e-03 | 8.5e-03 | 2.5e-12    | 3.3e+01 | 1.3e-03 | 9.2e-03 | 1.4e-10    | 1.0e+00 | 1.1e-03 | 5.3e-03 |
|                     | 6.8e-10  | 4.3e-01 | 1.2e-03 | 5.2e-03 | 6.8e-10    | 4.3e-01 | 1.3e-03 | 5.6e-03 | 1.7e-08    | 6.8e-02 | 8.4e-04 | 3.8e-03 |
| $\mu_J^*$           | 5.9e-13  | 9.1e+00 | 8.0e-05 | 2.0e-02 | 1.1e-11    | 9.1e+00 | 9.4e-05 | 2.5e-02 | 5.9e-13    | 9.0e-03 | 4.8e-05 | 2.2e-04 |
|                     | 4.2e-10  | 3.5e+00 | 8.5e-05 | 2.1e-02 | 4.2e-10    | 3.5e+00 | 1.0e-04 | 2.7e-02 | 8.6e-10    | 2.9e-03 | 5.7e-05 | 1.9e-04 |
| $B_2^{*c}$          | 5.3e-12  | 9.2e-03 | 7.6e-06 | 8.4e-05 | –          | –       | –       | –       | –          | –       | –       | –       |
|                     | 3.2e-15  | 4.2e-03 | 8.1e-06 | 9.8e-05 | –          | –       | –       | –       | –          | –       | –       | –       |
| $B_3^{*c}$          | 3.6e-13  | 1.2e-01 | 9.0e-05 | 5.4e-04 | –          | –       | –       | –       | –          | –       | –       | –       |
|                     | 1.6e-09  | 4.6e-02 | 8.8e-05 | 5.4e-04 | –          | –       | –       | –       | –          | –       | –       | –       |

<sup>a</sup> The squared error is computed as  $(y^{\text{pred}} - y^{\text{true}})^2$ . <sup>b</sup> A median lower than the mean indicates that error distribution is skewed towards low errors.

<sup>c</sup> The virial coefficient data is computed at zero density (i.e., fluid condition). The error distribution is reported just for the "All data" section to avoid repetition.



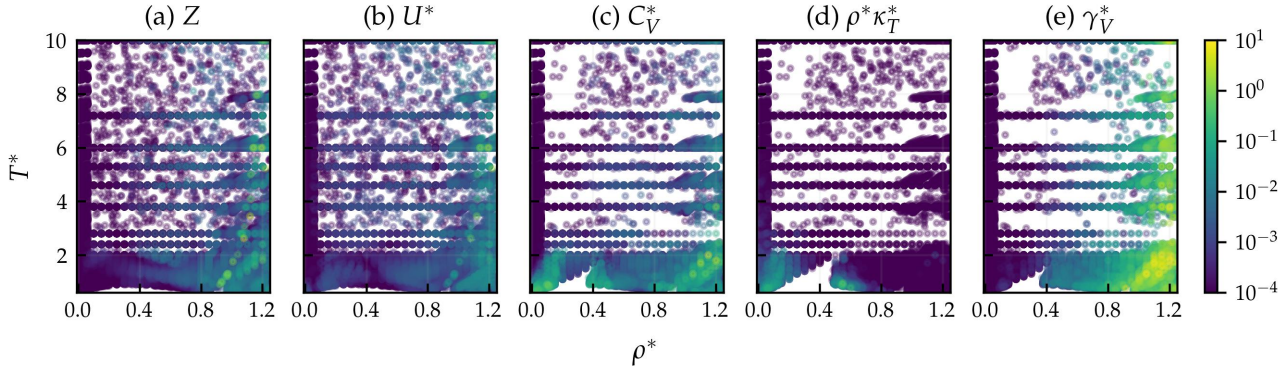


FIGURE 3.6: **Squared error distribution of the FE-ANN(s) EoS for selected first- and second-order derivative properties.** Thermophysical properties. (a) Compressibility factor ( $Z$ ). (b) Internal energy ( $U^*$ ). (c) Isochoric heat capacity ( $C_V^*$ ). (d) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (e) Thermal pressure coefficient ( $\gamma_V^*$ ). The squared error is indicated by the color bar and computed as  $(y^{\text{pred}} - y^{\text{true}})^2$ . Dark purple data points refer to low deviation, while yellow data points refer to high deviation with respect to the database.

### 3.3.2 FE-ANN(s) EoS: thermophysical properties modeling and prediction

In this section, the use of the FE-ANN(s) EoS will be showcased by predicting the thermophysical properties of three selected Mie particles ( $\lambda_r = 12, 16$ , and  $20$  and  $\lambda_a = 6$ ). The reader is referred to Section B.2 in Appendix B for results of other Mie particles with  $\lambda_r \in [7, 34]$  and  $\lambda_a = 6$ .

Several pressure isotherms of selected Mie particles are shown in Figure 3.7. This figure considers isotherms with temperatures ranging from  $0.65$  to  $7.2$ . For reference, for the Lennard-Jones particle, the triple ( $T_{\text{tr}}^*$ ) and critical ( $T_c^*$ ) temperatures are estimated to be  $0.69$  and  $1.321$ , respectively. [163] Therefore, this figure considers supercritical ( $T^* > T_c^*$ ), near-critical ( $T^* \approx T_c^*$ ), subcritical ( $T^* < T_c^*$ ), and near the triple temperature ( $T^* \approx T_{\text{tr}}^*$ ). The FE-ANN(s) EoS predictions (solid lines) are well-behaved and are in good agreement with the available data from MD simulations (empty circles). As a reference, the available data at other temperatures are shown with transparent circles. It can be observed that at high pressures and relatively low temperatures ( $T^* \leq 1.3$ ), the FE-ANN(s) EoS predicts the appearance of a van der Waals loop at high densities ( $0.8 < \rho^* < 1.1$ ). This van der Waals loop was previously observed in the Maxwell thermodynamic surface, Figure 3.1, and it is interpreted to be the region where the solid-liquid transition occurs. This van der Waals loop implies the existence of a metastable solid-fluid region that can be related to observable macroscopic phenomena such as the supercooling of fluids and superheating of solids [193].

The low-pressure regime is highlighted in the bottom zoomed-in plot of Figure 3.7. At these conditions, the commonly expected van der Waals loop is observed in the regions where the vapor-liquid transition happens. More importantly, when decreasing the temperature (dark blue isotherms), it is observed that the FE-ANN(s) EoS forecasts a region where instability is detected for both the vapor-liquid and solid-liquid transition. This type of isotherm suggests that the model predicts a temperature at which the three aggregation states coexist; i.e., the trained FE-ANN(s) EoS can model the triple point. Furthermore, at the lowest temperature ( $T^* = 0.65$ ), the isotherms also exhibit this "double van der Waals loop". However, it can be observed that an equilibrium tie-line for the solid-liquid coexistence would presumably take place at a negative pressure. This predicted behavior suggests that the SLE

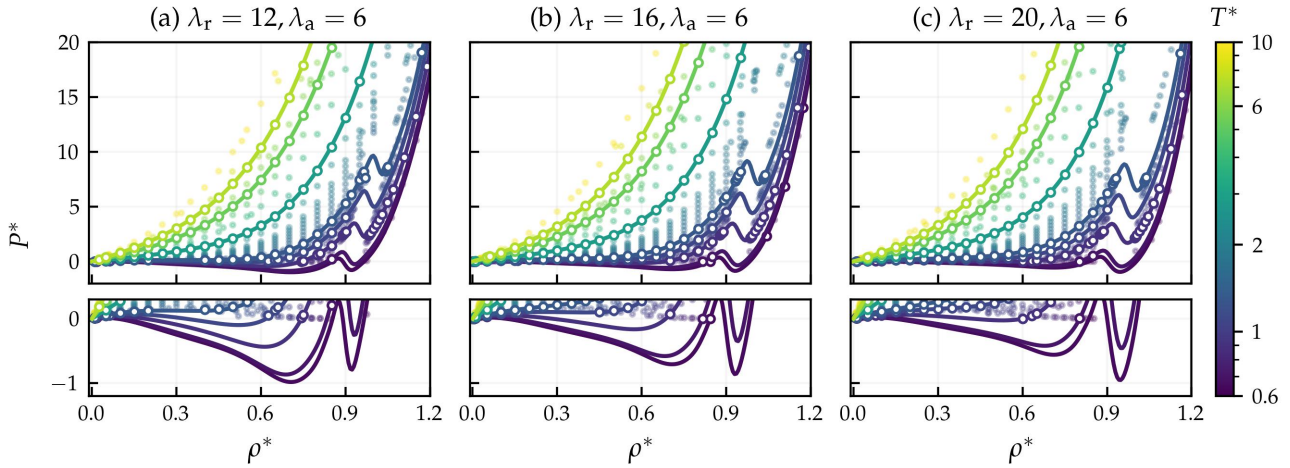


FIGURE 3.7: **Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** (a)  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones particle). (b)  $\lambda_r = 16, \lambda_a = 6$ . (c)  $\lambda_r = 20, \lambda_a = 6$ . Solid lines represent the following isotherms:  $T^* = 0.65, 0.7, 0.9, 1.1, 1.3, 2.8, 5.3$ , and  $7.2$  obtained with the FE-ANN(s) EoS. Solid empty circles represent the molecular dynamics (MD) result of the aforementioned isotherms. Transparent circles represent the data points available from the database. The coloring of lines and symbols refers to the temperature as indicated by the color bar.

is mechanically unstable at this temperature. Instead, as shown in Figure 3.8, the FE-ANN(s) EoS predicts a stable vapor-solid transition at this low temperature.

In summary, the FE-ANN(s) EoS predicts van der Waals loops, which define stable, metastable, and unstable regions. Although metastable phases have been observed in nature, stability limits (spinodal points,  $(\partial P^* / \partial \rho^*)_{T^*} = 0$ ) have been questioned in the literature [194]. The spinodal limits are a theoretical construct in EoSs based on the mean-field theory. Similarly, the spinodal limits observed by molecular simulations are system size dependent [195, 196]. It must be emphasized that the FE-ANN(s) EoS has "learned" from stable single-phase data points that have successfully passed rigorous sanity tests. The observed van der Waals loops are predicted by the EoS as a result of the imposed differentiable nature of the free energy surface. These van der Waals loops define unstable regions where different aggregation states coexist, enabling the FE-ANN(s) EoS to predict different types of phase equilibria (VLE, SLE, SVE, and SLVE).

The van der Waals loops define an accessible metastable region inside the equilibrium phase boundaries. This metastable region has been well-known for the vapor-liquid coexistence since the doctoral thesis of van der Waals [28]. In contrast, metastable supercooled liquids inside the SLE boundaries have been observed experimentally [197], but a discontinuous modeling approach prevented its study using equations of state. The FE-ANN(s) EoS [1] is the first continuous EoS capable of modeling this metastable behavior. The van der Waals loop for the solid-fluid transition implies limiting spinodal temperatures exist. There is no compelling evidence of the existence of such stability limits [198–200]. On the one hand, the spinodal limits appear for supercooled liquids where the repulsive core of the interaction potential exhibits a curvature change, which is not the case of the Mie particle [201]. On the other hand, the freeze method used to simulate SLE [164] (see Figure 3.3 (a) and Section 3.2), relies on finding these limiting temperatures to set-up a simulation that lies inside the unstable region. The FE-ANN(s) EoS suggests these limits exist merely based on single-phase (stable and metastable) data. Finally, molecular simulations have shown that the freezing of supercooling liquids is driven

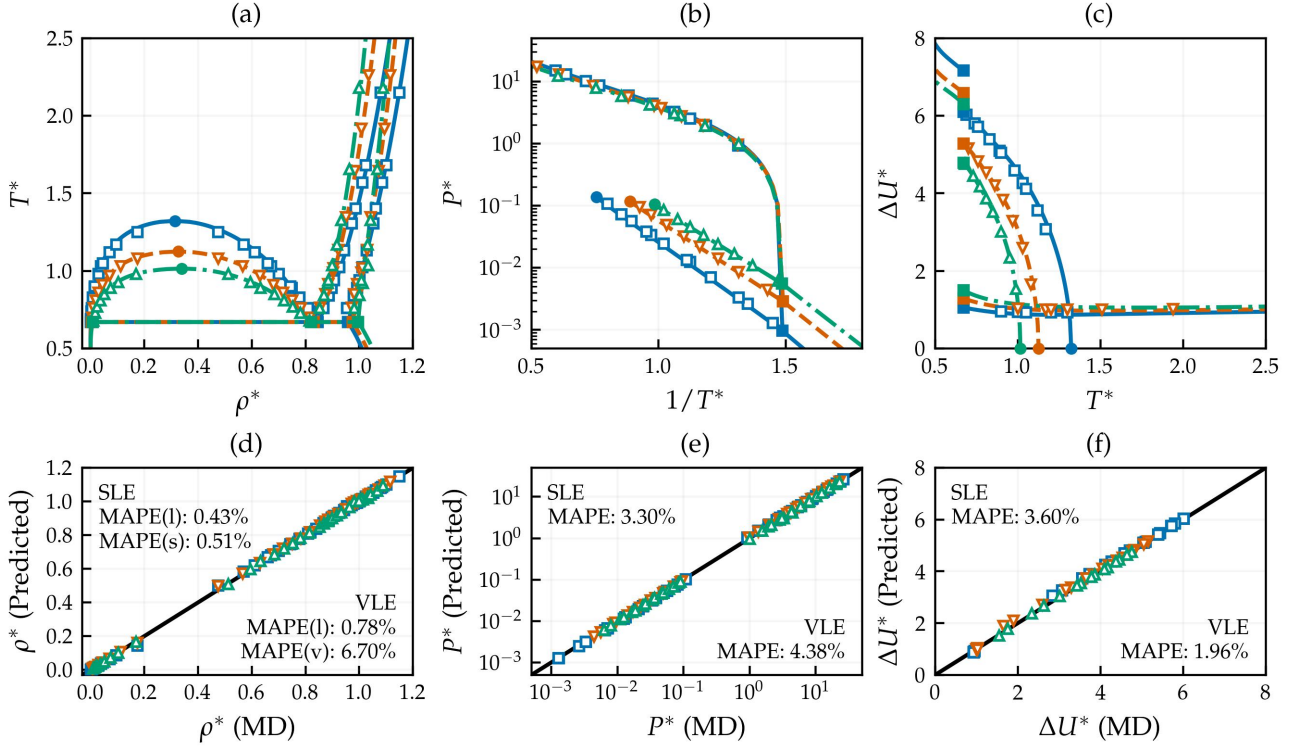


FIGURE 3.8: **Phase equilibria results of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** (a) Phase envelope. (b) Equilibrium pressure. (c) Internal energy change ( $\Delta U^{a \rightarrow b} = U^b - U^a$ ). Parity plots. (d) Equilibrium densities. (e) Equilibrium pressure. (f) Internal energy change. Blue empty squares and solid lines refer to  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones particle), orange empty downward triangles and dotted lines refer to  $\lambda_r = 16, \lambda_a = 6$ , green empty upright triangles and dashed lines refer to  $\lambda_r = 20, \lambda_a = 6$ . Lines and solid symbols refer to results obtained with the FE-ANN(s) EoS. The filled squares represent the triple points, and the filled circles represent the critical points. Empty symbols refer to Vapor-Liquid Equilibria (VLE) and Solid-Liquid Equilibria (SLE) from molecular dynamics (MD) simulations. The reported Mean Absolute Percentage Error (MAPE) is obtained as:

$$\text{MAPE} = 100 / N_P \sum_j^{N_P} |(y_j^{\text{pred}} - y_j^{\text{true}}) / y_j^{\text{true}}|.$$

by nucleation rather than by a spinodal decomposition [199]. This finding suggests that in a similar way as in condensation and vaporization [194], freezing is dominated by nucleation [199] and occurs before approaching the predicted spinodal limits.

Phase equilibria results predicted by the FE-ANN(s) are shown in Figure 3.8. Two-phase equilibrium (VLE, SLE, and SVE) is obtained when two phases of different densities exhibit the same pressure and chemical potential at a given temperature. Phase equilibria can be computed using the Maxwell equal area rule or equivalent by numerically solving the system of Eqs. (2.7). From Figure 3.8 (a), it is observed that the FE-ANN(s) EoS predicts a phase diagram that exhibits VLE, SLE, and SVE. The SLE tends to shift toward lower-density regions as the repulsive part of the potential becomes steeper (increasing repulsive exponent). Contrarily, the FE-ANN(s) EoS predicts that the melting pressure and internal energy change remain relatively insensitive to the repulsive exponent of the potential. In general, the predicted equilibrium density, pressure, and internal energy phase change are in good agreement with the available MD VLE and SLE results not used for training. The excellent predictive capabilities are confirmed in the parity plots, Figures 3.8 (c)-(f), where most data points lie in the diagonal and low deviations are observed.



It is reassuring to observe that the FE-ANN(s) EoS predicts important phase behavior like the critical and triple points without this information being built into the EoS formulation. The critical point is an inflection point in the  $\rho^* - P^*$  projection (i.e.,  $(\partial P^* / \partial \rho^*)_T^* = (\partial^2 P^* / \partial \rho^{*2})_T^* = 0$ ). This condition implies that derivative properties that depend on  $1 / (\partial P^* / \partial \rho^*)_T^*$  diverge at this condition. These properties include the isothermal compressibility, isobaric heat capacity, and thermal pressure coefficient. The critical point is obtained by numerically solving the system of Eqs. (2.6). As expected, no fluid-solid critical point is predicted within the FE-ANN(s) EoS density and temperature training range (i.e.,  $\rho^* < 1.3$  and  $T^* < 10$ ). Extrapolation beyond these limits is not recommended. Lastly, the FE-ANN(s) EoS also predicts a unique temperature where a stable a vapor (v), liquid (l), and solid (s) phase coexist. The triple point requires that these three phases have the same chemical potential and pressure, as shown below.

$$T_{\text{tr},v} = T_{\text{tr},l} = T_{\text{tr},s} \quad (3.1a)$$

$$P_{\text{tr},v} = P_{\text{tr},l} = P_{\text{tr},s} \quad (3.1b)$$

$$\mu_{\text{tr},v} = \mu_{\text{tr},l} = \mu_{\text{tr},s} \quad (3.1c)$$

This system of Eqs. (3.1) is numerically solved for the triple temperature ( $T_{\text{tr}}$ ), vapor density ( $\rho_{\text{tr},v}$ ), liquid density ( $\rho_{\text{tr},l}$ ) and solid density ( $\rho_{\text{tr},s}$ ). The pressure and chemical potential are obtained by evaluating Eqs. (2.1b) and (2.2c), respectively.

The critical and triple points of ( $\lambda_r, \lambda_a = 6$ ) Mie particles are shown in Figure 3.9. As observed in Figure 3.9 (a), the critical temperature decreases the higher the repulsive exponent. This trend is well-known and observed in other EoSs for the Mie fluid [3, 37, 113, 115]. This decrease is also observed for the triple temperature but at a much slower rate. The FE-ANN(s) EoS predicts how the VLE phase envelope shrinks for increasingly repulsive force fields and, under sensible extrapolation, predicts the vanishing of the VLE phase envelope (i.e.,  $T_c^* / T_{\text{tr}}^* = 1$ ) for highly repulsive potentials. This extrapolated point is indicated by a star symbol in Figures 3.9 (a) and (b). The FE-ANN(s) EoS predicts that the VLE vanishes for the Mie particle with  $\alpha_{\text{vdw}} = 0.36$  (equivalently to  $\lambda_r = 431.6$  and  $\lambda_a = 6$ ). This result is in excellent agreement with the value of  $\alpha_{\text{vdw}} = 0.38$  obtained by extrapolating the results of Ramrattan [164]. Additionally, this result is consistent with the behavior of purely repulsive systems such as the hard sphere [202] and highly repulsive colloidal systems [203]. Finally, as shown in Figure 3.9, the predicted critical and triple points by the FE-ANN(s) EoS are in excellent agreement with published molecular simulation results [151, 164–166, 204].

The FE-ANN(s) EoS also excels at modeling second-order properties. The isochoric heat capacity ( $C_V^*$ ) of selected Mie particles are shown in Figure 3.10. In general, the predicted heat capacity isotherms (solid lines) are in good agreement with respect to the MD data (empty circles). Similar to Figure 3.7, the transparent empty circle shows the available data points in the database, and the temperature of the isotherms is indicated by the color bar. The FE-ANN(s) EoS predicts two distinguishable maximums of the isochoric heat capacity. The first maximum is observed at the region that lies inside the VLE phase boundaries. This maximum is also predicted by cubic [28], molecular [37], and technical [115] EoSs and is observed in experimental [12] and simulation [3, 163] data. A second, sharper maximum of the isochoric heat capacity is predicted at the solid-fluid transition. It must be recalled that the FE-ANN(s)

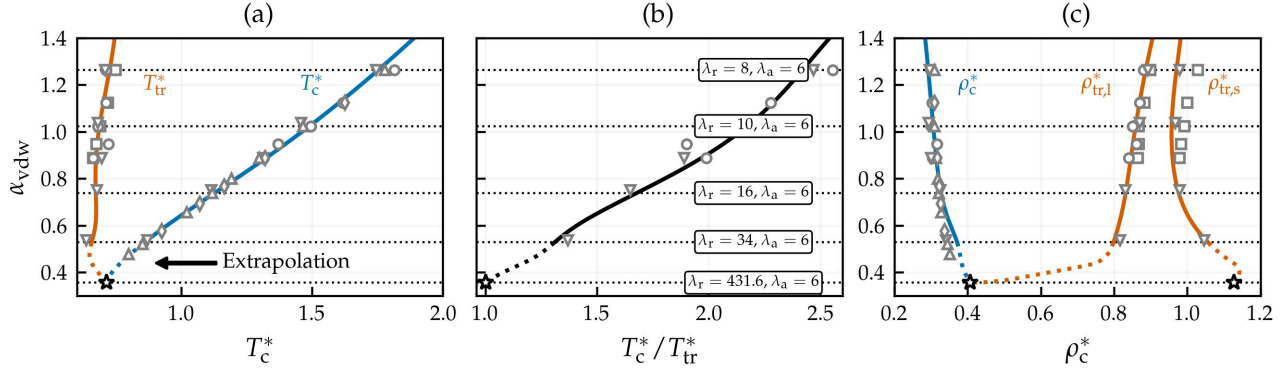


FIGURE 3.9: **Critical and triple points of  $(\lambda_r, \lambda_a = 6)$  Mie particles.** (a) Critical ( $T_c^*$ ) and triple ( $T_{tr}^*$ ) temperatures. (b) Critical and triple temperature ratio ( $T_c^*/T_{tr}^*$ ). (c) Critical ( $\rho_c^*$ ) and liquid and solid triple ( $\rho_{tr,l}^*$  and  $\rho_{tr,s}^*$ ) densities. Blue lines refer to a critical point property. Orange lines refer to a triple point property. Solid lines refer to calculations obtained from the FE-ANN(s) EoS in the training range ( $\lambda_r \in [7, 34]$ ). Dotted lines refer to extrapolations beyond the training range ( $\lambda_r > 34$ ). Star symbol refers to the point where the VLE vanishes (i.e.,  $T_c^* = T_{tr}^*$ ). Grey empty symbols refer to published molecular simulation results: squares [204], circles [151], upside down triangles [164], diamonds [165], and triangles [166].

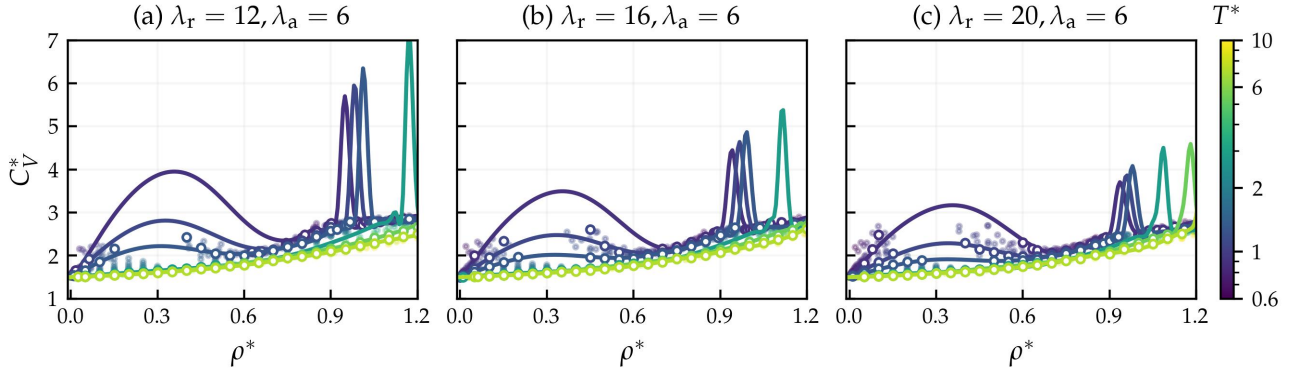


FIGURE 3.10: **Isochoric heat capacity of selected  $(\lambda_r, \lambda_a = 6)$  Mie particles.** (a)  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones particle). (b)  $\lambda_r = 16, \lambda_a = 6$ . (c)  $\lambda_r = 20, \lambda_a = 6$ . Solid lines represent the following isotherms:  $T^* = 0.9, 1.1, 1.3, 2.8, 5.3$ , and  $7.2$  obtained with FE-ANN(s) EoS. Solid empty circles represent the molecular dynamics (MD) result of the aforementioned isotherms. Transparent circles represent the data points available from the database. The coloring of lines and symbols refers to the temperature as indicated by the color bar.

EoS is formulated as a physics-embedded ANN trained only using derivative information and that this solid-fluid maximum is predicted from interpolating the MD data. This maximum is not observed from the MD simulation data, as no data is available between the transition region. Nevertheless, this predicted heat capacity maximum is experimentally well-known and is commonly employed to determine the melting temperature of real substances using differential scanning calorimetry [205]. Finally, the solid isochoric heat capacities are close to 3. This value is predicted by the Einstein solid at high temperatures [206] and has been empirically corroborated by the Dulong-Petit law [207].

The FE-ANN(s) EoS accurately models other second-order derivative properties. The prediction of the isothermal compressibility (as  $\rho^* \kappa_T^*$ ), thermal pressure coefficient ( $\gamma_V^*$ ), thermal expansion coefficient ( $\alpha_P^*$ ), adiabatic index ( $\gamma = C_P^*/C_V^*$ ) and Joule-Thomson coefficient ( $\mu_{JT}^*$ ) and their comparison to the available MD data are shown in Figure 3.11. Overall, the FE-ANN(s) EoS predicts well-behaved second-order properties isotherms that are in good agreement with the reference data. The FE-ANN(s)

EoS predicts spinodal points for the vapor-liquid and fluid-solid transitions that imply a divergent behavior for some properties, see the vertical lines in Figures 3.11 (a), (c) and (d). As explained in Eqs. (2.23) in Section 2.5, the isothermal compressibility, thermal expansion coefficient and adiabatic index exhibit this divergence, due to a division by  $(\partial P^* / \partial \rho^*)_{T^*} = 0$  at the spinodal points.

The consistency of the FE-ANN(s) EoS is assessed using Brown's characteristic curves [168]. These curves are related to the conditions where a fluid exhibits the behavior of an ideal gas. This work focuses on four curves, namely: the Zeno curve (Eq. (2.24d),  $Z = P^* / \rho^* T^* = 1$ ), the Boyle curve (Eq. (2.24c),  $(\partial Z / \partial P^*)_{T^*} = 0$ ), Charles curve (also known as Joule-Thomson inversion curve, Eq. (2.24b),  $(\partial Z / \partial T^*)_{P^*} = 0$ ,) and the Amagat curve (also known as Joule inversion curve, Eq. (2.24a),  $(\partial Z / \partial T^*)_{\rho^*} = 0$ ). The reader is referred to the work of Deiters and De Reuck [89] for a thorough discussion about the shape and behavior of these curves for an equation of state to be correct.

The phase diagram, Brown's characteristic curves, and virial coefficients of selected Mie particles ( $\lambda_r = 12, 16$ , and  $20$  and  $\lambda_a = 6$ ) are shown in Figure 3.12. It can be observed in Figure 3.12 (b) and (c) that the predicted second and third virial coefficients exhibit smooth behavior and correctly decay to negative values at low temperatures.

The predicted Brown's characteristic curves, shown in Figure 3.12 (a), follow the expected behavior in the fluid region and are in good agreement with published data [166]. The Zeno curve encloses the VLE curve, the Boyle curve finishes close to the critical point, the Charles curve encloses the Boyle curve and crosses the Zeno curve only once, and the Amagat curve encloses all the other curves. The Zeno and Boyle curves asymptotically approach the Boyle temperature ( $T_{\text{Boyle}}^*$ ) at the low-pressure regime. At this temperature, the second-virial coefficient is zero ( $B^* = 0$ ), as shown by the green upright triangle in Figure 3.12 (b). Similarly, at the low-pressure limit, the Charles curve approaches the Charles temperature ( $T_{\text{Charles}}^*$ ), which is characterized by an asymptotic behavior of the second-virial coefficient ( $(dB^* / dT^*) = B^* / T^*$ ). The Charles temperature is shown by the orange downward triangle in Figure 3.12. The Amagat curves should approach the Amagat temperature ( $T_{\text{Amagat}}^*$ ) at the low-pressure regime. This temperature is characterized by a maximum of the second virial coefficient (i.e.,  $(dB^* / dT^*) = 0$ ) and occurs at a high temperature, e.g.,  $T_{\text{Amagat}}^* \approx 25$  for the Lennard-Jones particle [163]. This temperature is far outside the training region ( $T^* \leq 10$ ) and is not shown here. Interestingly, for fluid-only EoSs, the Zeno and Amagat extend beyond the freezing conditions. At these conditions, the FE-ANN(s) EoS encounters the solid-fluid van der Waals loop (see Figure 3.7) and stationary/divergent conditions in second-order properties (see Figures 3.10 and 3.11) that prevent the calculation of the Brown's curve beyond the solid-fluid phase transition. Overall, the FE-ANN(s) EoS correctly passed this consistency test, further confirming the correctness of the model.

### 3.4 Conclusions

This chapter shows how the FE-ANN(s) EoS framework, which leverages modern machine learning models trained on comprehensive molecular dynamics data and well-established physical constraints, fulfills the longstanding aspiration of J.C. Maxwell and J.W. Gibbs to represent all thermodynamic states of matter with a single free energy surface. Although this methodology has been applied to a simple spherical model, the resulting FE-ANN(s) EoS provides valuable insights into the complex

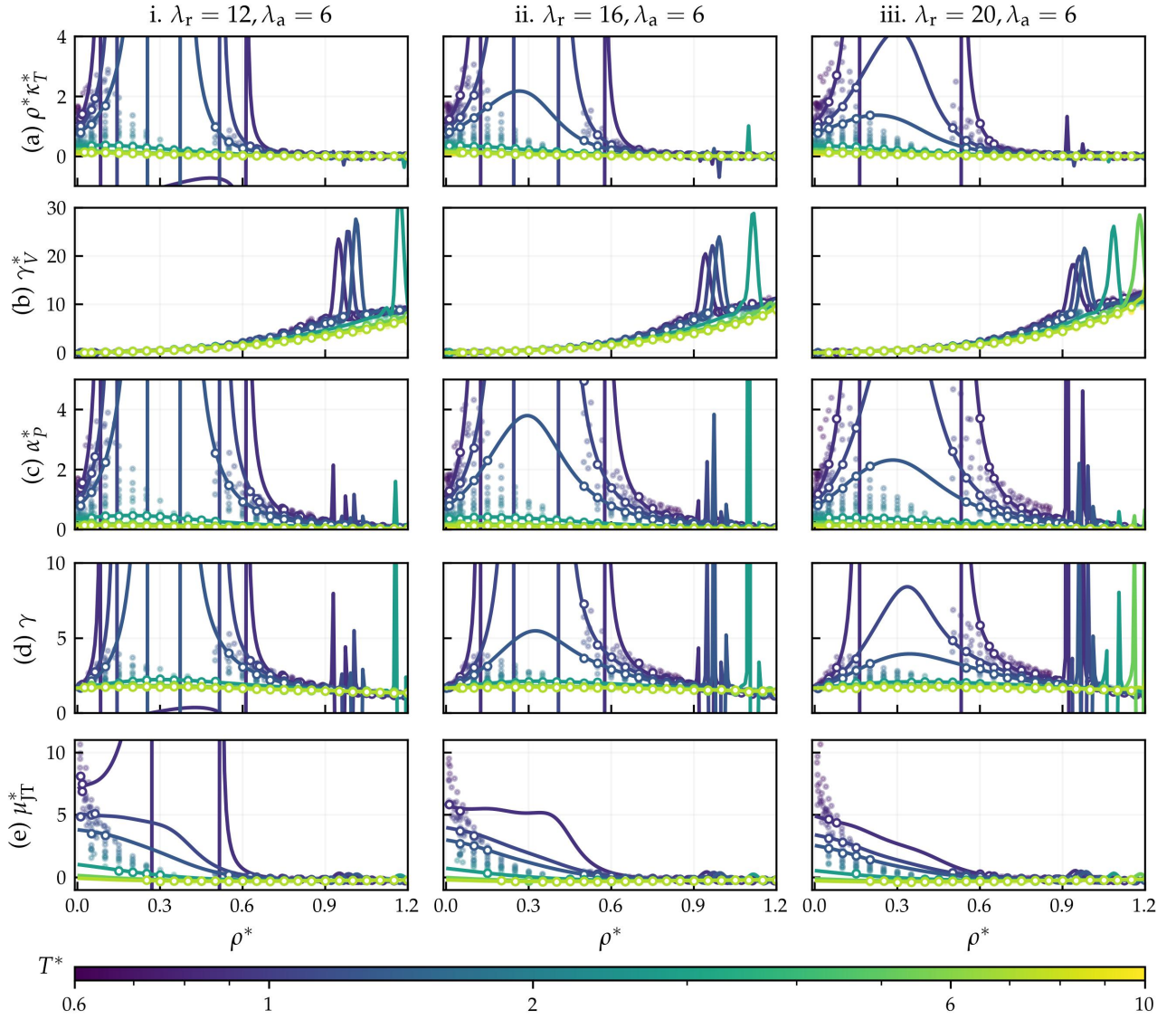


FIGURE 3.11: **Second-order properties of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** i.  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones particle). ii.  $\lambda_r = 16, \lambda_a = 6$ . iii.  $\lambda_r = 20, \lambda_a = 6$ . (a) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (b) Thermal pressure coefficient ( $\gamma_V^*$ ). (c) Thermal expansion coefficient ( $\alpha_p^*$ ). (d) Adiabatic index (heat capacity ratio,  $\gamma = C_p^*/C_V^*$ ). (e) Joule-Thomson coefficient ( $\mu_{JT}^*$ ). Solid lines represent the following isotherms:  $T^* = 0.9, 1.1, 1.3, 2.8, 5.3$ , and  $7.2$  obtained with FE-ANN(s) EoS. Solid empty circles represent the molecular dynamics (MD) result of the aforementioned isotherms. Transparent circles represent the data points available from the database. The coloring of lines and symbols refers to the temperature as indicated by the color bar.

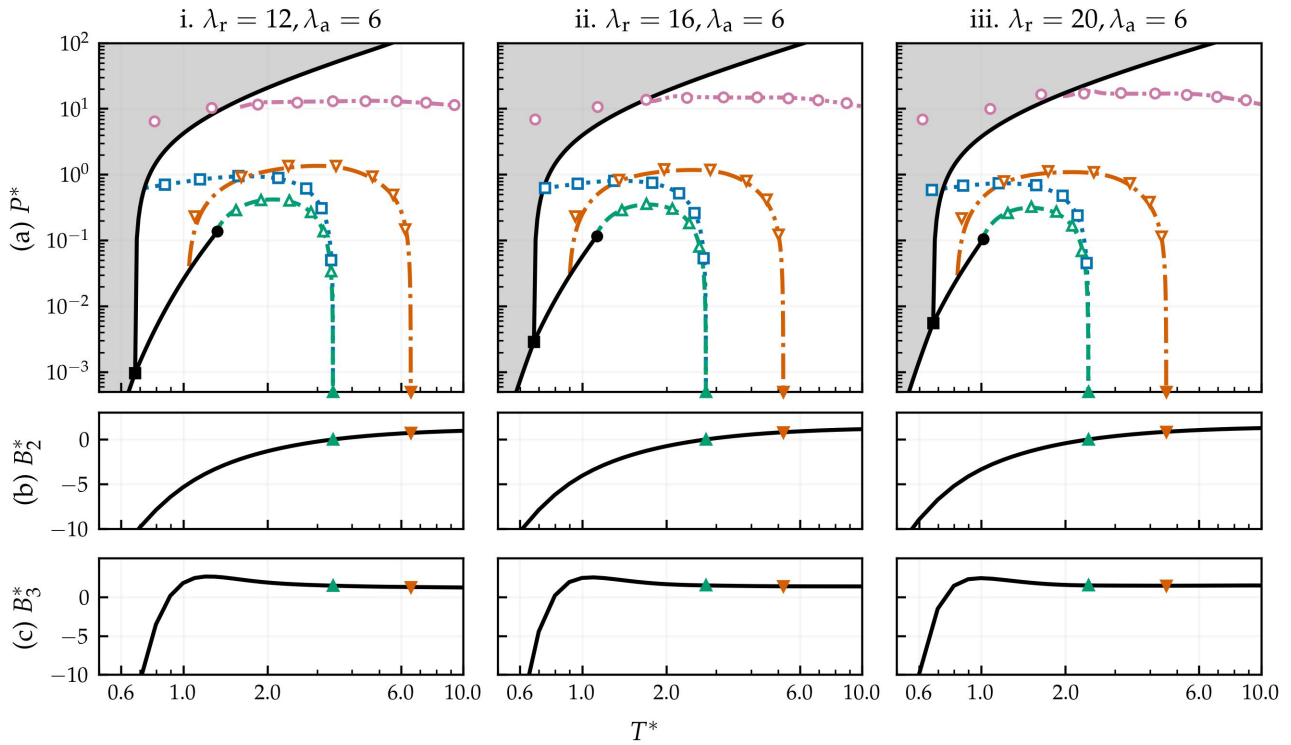


FIGURE 3.12: **Brown's characteristic curves and virial coefficients of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** i.  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones particle). ii.  $\lambda_r = 16, \lambda_a = 6$ . iii.  $\lambda_r = 20, \lambda_a = 6$ . Lines and filled symbols are results obtained with the FE-ANN(s) EoS. **(a)** Brown characteristic curves. The black solid lines represent equilibrium pressure (SVE, SLE, and VLE), blue dotted line is the Zeno curve, green dashed line is the Boyle curve, orange dot-dashed line is the Charles curve, and pink dot-dot-dashed line is the Amagat curve. The filled black square represents the triple point, the filled black circle represents the critical point, the filled upright green triangle is the Boyle temperature, and the filled downward orange triangle is the Charles temperature. The grey-shaded area represents the solid region. Empty symbols are molecular simulation results obtained by Stephan and Urschel [166]. **(b)** Second virial coefficient. **(c)** Third virial coefficient.

transition regions between aggregation states. To the best of my knowledge, the FE-ANN(s) EoS is the first EoS to continuously model the phase space across the entire density range from the gas phase to the crystalline phase. The framework represents a significant advance in the development of equations of state.

The robustness and flexibility of the FE-ANN(s) EoS framework are demonstrated for systems interacting through a truncated Mie potential. The FE-ANN(s) EoS identifies unstable and metastable regions which are characterized by van der Waals loops. These van der Waals loops indicate regions where the Mie particle exhibits stable vapor-liquid and solid-liquid equilibria. Moreover, at low temperatures, these van der Waals loops merge at negative pressures, suggesting the existence of a stable solid-vapor equilibrium. Furthermore, the FE-ANN(s) EoS predicts a unique pressure isotherm where vapor, liquid, and solid phases coexist, signifying a continuous description of the triple point. This EoS also accurately captures the critical point for the vapor-liquid phase transition and, as expected, it does not predict a critical point for the solid-liquid transition within the validity range of the model (i.e.,  $\rho^* \in [0, 1.3]$  and  $T^* \in [0.58, 10]$ ). Additionally, the FE-ANN(s) EoS predicts a shift of the solid-liquid equilibrium towards lower densities and a shrinking of the vapor-liquid phase envelope as the repulsive interactions in the system increase. In such cases, the FE-ANN(s) EoS suggests the eventual destabilization of the vapor-liquid equilibrium, potentially falling below the triple line, as previously hypothesized for highly repulsive systems [203].

Second-order derivative properties, such as isobaric heat capacity, thermal expansion coefficient, and isothermal compressibility, exhibit divergence within unstable regions. Conversely, the isochoric heat capacity exhibits maxima in the fluid-fluid and fluid-solid metastable regions. Such maxima are corroborated by experimental data, which further support the correctness of the FE-ANN(s) EoS predictions. Furthermore, the internal thermodynamic consistency of the model is successfully verified using Brown's characteristic curves.

The FE-ANN(s) EoS framework bypasses the need for a rigorous analytical resolution of the statistical mechanics problem. Instead, it embeds well-accepted principles into its formulation (e.g., Maxwell's relations, the ideal gas law, and criteria for thermal and mechanical stability), and it establishes a numerical relationship between the intermolecular potential (force field) and macroscopic thermodynamic properties obtained using high-throughput MD simulations. Notably, this physics-embedded machine learning framework can explore and generate thermodynamically consistent EoSs for systems lacking accepted theoretical descriptions. This capability is exemplified here for the solid-fluid transition, but it can also be used in systems with significant polarization effects, charge interactions, or multipolar moments.

Finally, the approach presented here has the potential to significantly impact the development and parameterization of future EoSs by incorporating solid-phase data as a target in the objective functions. Accurately describing the solid phase is crucial for various fields, including chemical process design (crystallization, supercritical extraction) [208] and modeling solid solutions [209].

## Chapter 4

# Simulation and data-driven modeling of the transport properties of the Mie fluid

### 4.1 Introduction

For design and engineering purposes, it is essential to accurately model both thermodynamic properties (e.g., phase equilibria and heat capacities) and transport properties (e.g., diffusivity, shear viscosity, and thermal conductivity) of a fluid (or pseudo-fluid) over a wide range of conditions. [14, 41] In this sense, the Mie potential, Eq. (1.1), is a simple semi-empirical interaction potential that has gained attention because of its flexibility in modeling thermophysical properties of a diverse range of fluids and fluid mixtures [91]. The thermodynamic properties of the Mie fluid have been successfully modeled through theoretical [37, 65, 113, 114], empirical [115], and machine-learning-based (detailed in Chapter 2 and Ref. [3]) Equations of State (EoS). By using these EoSs, the Mie potential has served as the basis for representing the phase equilibria of simple [97] and asymmetric [98] mixtures, biofuel blends [99, 100], active pharmaceutical ingredients (API) [101, 102], and cryogenics [95, 103]. In contrast to the success in modeling thermodynamic properties, the transport properties of the Mie fluid have been only scarcely studied in the open literature. For example, the Mie potential has been used as a model to describe transport properties, such as bulk [110] and shear viscosities [111, 112] of complex fluid systems using molecular simulations, but its modeling has been limited. From a modeling perspective, one of the main challenges is the absence of a widely accepted framework for describing transport properties.

Transport properties have been historically modeled using a combination of empirical, semi-empirical, and theoretical-based approaches. Correlations based on state conditions (density and temperature) have been extensively used by the U.S. National Institute of Standards and Technology (NIST) [12] to model transport properties of molecular fluids, such as the shifted Lennard-Jones fluid [210] as well as real fluids, see Refs. [211–213] to name a few. These empirical correlations can accurately model transport properties. However, in most cases, they lack generality, and extrapolation is not recommended.

Transport properties have also been modeled using semi-empirical approaches. For example, the free-volume theory [24] provides a link between an "effective" free volume and the shear viscosity. This theory has been successfully applied to pure fluids and fluid mixtures [214–216]. This theory relies

on three component-specific adjustable parameters and the liquid density, which can be predicted using an EoS. Similarly, the friction theory [25] relates the shear viscosity to the balance between attractive ( $P < 0$ ) and repulsive ( $P > 0$ ) pressures in a fluid system. These repulsive and attractive pressures are correlated with the shear viscosity using an empirical function. These pressures are obtained from an EoS. As an alternative approach, entropy scaling [217] has been proposed as a framework to link thermodynamic and transport properties. In this framework, a scaled transport property is considered a univariate function of the residual entropy. The residual entropy is obtained directly from a Helmholtz-based EoS using Eq. (2.1a). Although entropy scaling has its roots in an empirical observation, it has been derived from isomorph theory [26, 218]. One of the limitations of entropy scaling is that it is usually valid only for dense (liquid-like) phases. This framework exhibits divergent behavior in the low-density regime. Furthermore, the functionality of the scaled transport property with respect to the reduced entropy is not known and must be obtained empirically [219–222]. Following a similar approach, in thermodynamic (density) scaling, a transport property of a dense state is considered invariant with respect to the ratio of the scaled density to the temperature. Thermodynamic scaling also has its roots in isomorph theory [218].

Lastly, transport properties have also been modeled using theoretical approaches. The options here are scarce and mainly limited to the low-density region. The Chapman-Enskog theory [117] relies on solving the collision integral of an interaction potential and can predict the transport properties of a dilute gas ( $\rho \rightarrow 0$ ). The revised-Enskog theory (RET) [117, 223] extends the validity of the original Chapman-Enskog theory to denser states and has been successfully applied to predict the transport properties of the Mie fluid [48]. Jervell and Wilhelmson [48] showed that the experimental transport properties of selected gases (e.g., argon and methane) can be accurately modeled using molecular parameters (i.e.,  $\epsilon$ ,  $\sigma$ , and  $\lambda_r$  in Eq. (1.1)) fitted exclusively to thermodynamic (equilibrium) properties. Despite the promising results of the revised Enskog theory for Mie fluids [48], these predictions have yet to be verified by molecular simulation results.

Regardless of whether transport properties are modeled using an empirical, semi-empirical, or theoretical approach, it is necessary to rely on reference data for the transport properties of the Mie fluid to develop the models and validate the theories. In the literature, data on transport properties over a wide range of densities and temperatures have been collected and generated by Lautenschlaeger and Hasse [210] for the Lennard-Jones fluid. Šlepavičius et al. [224] generated self-diffusivity and shear viscosity data for a few Mie fluids. This database consists of approximately 1000 data points for the self-diffusivity and 270 data points for the shear viscosity for low to moderate densities and low to moderate supercritical temperatures. A comprehensive and self-consistent database for transport properties, including the thermal conductivity over a wide range of densities and temperatures, is still needed.

For these reasons, this chapter has two main objectives. The first objective is to develop a comprehensive database for the transport properties (self-diffusivity, shear viscosity, and thermal conductivity) of the Mie fluid for a wide range of density and temperature conditions using molecular dynamics simulations. The computation and description of this database will be covered in Section 4.2. The second objective is to adequately model the transport properties of the Mie fluid. Despite the advances provided by the revised Enskog theory [48], this theory is still limited to low to moderate



density conditions. Therefore, two modeling approaches based on machine learning will be tested in Section 4.3. The first empirical approach directly models a transport property based on its state conditions (density and temperature). This data-driven approach has already been used to correlate the transport properties of pseudo and real fluids [224–227]. The second semi-empirical approach uses entropy scaling to model the transport properties. As the functional relationship between the input variables ( $(\rho, T)$  or  $S$ ) and the transport property is not known beforehand, this is an ideal case for using machine learning algorithms to uncover the structure-property relationships between the transport property and the input variables. Finally, the findings of this chapter are summarized and concluded in Section 4.4.

The results presented in this chapter are based on an already published article (Gustavo Chaparro and Erich A. Müller. “Simulation and Data-Driven Modeling of the Transport Properties of the Mie Fluid”. *The Journal of Physical Chemistry B* 128.2 (2024), pp. 551–566. DOI: 10.1021/acs.jpcb.3c06813, Ref. [2]). The results presented here have been updated and include an extensive database and machine learning models trained on this updated dataset.

## 4.2 Mie fluid transport properties database

This section focuses on computing the transport properties (self-diffusivity, shear viscosity, and thermal conductivity) of fluid systems interacting through a truncated Mie potential, Eq. (1.1) using molecular dynamics (MD) simulations.

$$\mathcal{U}^{\text{Mie}} = \begin{cases} \mathcal{C}^{\text{Mie}} \epsilon \left[ (\sigma/r)^{\lambda_r} - (\sigma/r)^{\lambda_a} \right] & \text{if } r \leq 5\sigma \\ 0 & \text{if } r > 5\sigma \end{cases} \quad (1.1)$$

$$\mathcal{C}^{\text{Mie}} = \frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}$$

The Mie potential relies on several molecular parameters. The transport properties database is reported in reduced units, henceforth denoted by a "\*" superscript. Reduced units are obtained when setting the Boltzmann constant ( $k_B$ ), Avogadro constant ( $N_{\text{Av}}$ ), the monomer mass ( $M$ ), the potential well-depth ( $\epsilon$ ), and the characteristic size parameter ( $\sigma$ ) to one. In this case, the Mie potential is fully described by the repulsive ( $\lambda_r$ ) and attractive ( $\lambda_a$ ) exponents, which control the steepness of these interactions. The definition of relevant physical quantities and transport properties in reduced units are given in Table 4.1.

Similar to the Mie fluid thermophysical properties database obtained in Section 2.4, the transport properties database considers Mie fluids with repulsive exponents ranging from 7 to 34. The attractive exponent is taken as 6 in agreement with the London theory for dispersion forces [93] resulting in what is sometimes referred to as the  $(\lambda_r, \lambda_a = 6)$  Mie fluid [91]. The transport properties database includes fluid data points outside the Vapor-Liquid Equilibria (VLE) and Solid-Liquid Equilibria (SLE) boundaries. The VLE phase envelope is available from Section 2.4, and the SLE boundaries were estimated using the correlation of Khrapak et al. [151], Eq. (2.18).

TABLE 4.1: Definition of physical quantities in reduced units.

| Quantity             | Symbol   | Units                           | Reduced quantity <sup>a, b, c</sup>                                           |
|----------------------|----------|---------------------------------|-------------------------------------------------------------------------------|
| Time                 | $\tau$   | s                               | $\tau^* = \tau / \sigma \sqrt{M / N_{\text{Av}} \epsilon}$                    |
| Mass                 | $m$      | $\text{kg mol}^{-1}$            | $m^* = m / M$                                                                 |
| Density              | $\rho$   | $\text{mol m}^{-3}$             | $\rho^* = \rho N_{\text{Av}} \sigma^3$                                        |
| Temperature          | $T$      | K                               | $T^* = T k_{\text{B}} / \epsilon$                                             |
| Pressure             | $P$      | Pa                              | $P^* = P \sigma^3 / \epsilon$                                                 |
| Self-diffusivity     | $D$      | $\text{m}^2 \text{s}^{-1}$      | $D^* = D / \sigma \sqrt{N_{\text{Av}} \epsilon / M}$                          |
| Shear Viscosity      | $\eta$   | Pa s                            | $\eta^* = \eta \sigma^2 / \sqrt{\epsilon M / N_{\text{Av}}}$                  |
| Thermal Conductivity | $\kappa$ | $\text{W m}^{-1} \text{K}^{-1}$ | $\kappa^* = \kappa \sigma^2 / k_{\text{B}} \sqrt{N_{\text{Av}} \epsilon / M}$ |

<sup>a</sup> Superscript "\*" refers to a property in reduced units. <sup>b</sup>  $N_{\text{Av}} = 6.02214076 \cdot 10^{23} \text{ mol}^{-1}$  is the Avogadro constant [141],  $k_{\text{B}} = 1.380649 \cdot 10^{-23} \text{ J K}^{-1}$  is the Boltzmann constant [141] and  $M$  is the molar mass of a single monomer. <sup>c</sup>  $\sigma$  is the shape parameter, and  $\epsilon$  is the potential well depth of the Mie potential (See Eq. (1.1)).

The computation of transport properties using molecular dynamics simulations is detailed in Section 4.2.1. The curated transport properties database is then described and assessed against published data in Section 4.2.2.

### 4.2.1 Molecular dynamics simulations

#### Computation of transport properties

In molecular dynamics (MD) simulations, transport properties (self-diffusivity, shear viscosity, and thermal conductivity) can be obtained either from non-equilibrium (NEMD) or equilibrium (EMD) simulations. Both approaches have advantages and disadvantages. In NEMD, the simulation is perturbed in a way that the transport property is related to its fundamental definition, e.g., thermal conductivity from Fourier's law [228] and shear viscosity from the Newtonian fluid definition [229]. In the NEMD approach, the transport property is obtained in a transient state, which usually implies that the results can be obtained with shorter simulation times than with EMD. However, as an equilibrated simulation box is uniquely perturbed, either by a temperature or a momentum gradient, in general, only one transport property can be computed at a time [230]. Conversely, in EMD, the simulation evolves in time in its canonical state. Hence, all the transport properties can be obtained simultaneously. This characteristic makes using EMD to produce a comprehensive transport properties database convenient. The EMD method comes at the expense of a more extended simulation, as statistical uncertainty is considerable [231].

In EMD, two fundamentally equivalent methodologies exist to compute a transport property. In the Green-Kubo approach, Eq. (4.1a), a transport property ( $\Lambda^*$ ) is related to the integral of a time-correlation function of a dynamical variable  $\dot{\mathcal{A}}^*$ . Similarly, the Einstein method, Eq. (4.1b), relates the

same transport property to the mean-squared displacement (MSD) of the variable  $\mathcal{A}^*$ .

$$\Lambda^* = \int_0^\infty d\tau^* \langle \dot{\mathcal{A}}^*(\tau^*) \dot{\mathcal{A}}^*(0) \rangle \quad (4.1a)$$

$$\begin{aligned} \Lambda^* &= \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \left\langle \left[ \int_0^{\tau^*} \dot{\mathcal{A}}^*(\tau^*) d\tau^* \right]^2 \right\rangle \\ &= \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \underbrace{\left\langle [\mathcal{A}^*(\tau) - \mathcal{A}^*(0)]^2 \right\rangle}_{MSD} \end{aligned} \quad (4.1b)$$

In Eqs. (4.1), the  $\langle \cdots \rangle$  brackets refer to an ensemble average. Both the Green-Kubo and Einstein formulations require sampling the property  $\mathcal{A}^*$  (or  $\dot{\mathcal{A}}^*$ ) for a sufficiently long time to obtain reliable statistics. The integrand in the Green-Kubo approach method slowly decays to zero regardless of the simulation time [232, 233] and non-trivial time-decomposition methods are necessary to improve the reliability of the results [234]. The Einstein method has some practical advantages that make it suitable for producing results on a large scale. The validity of the Einstein method can be directly verified by plotting the MSD as a function of time in a log-log plot. The Einstein formulation is only valid if the slope of the MSD curve equals one [235]. This latter approach is particularly convenient for simple isotropic fluids and is used herein.

In the Einstein formulation, the self-diffusivity coefficient ( $D^*$ ) is obtained as the averaged mean squared displacement of all the particles.

$$D^* = \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \frac{1}{3N} \left\langle \sum_{j=1}^N \left[ \int_0^{\tau^*} \mathbf{v}_j^*(\tau^*) d\tau^* \right]^2 \right\rangle \quad (4.2a)$$

$$= \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \frac{1}{3N} \underbrace{\left\langle \sum_{j=1}^N [\mathbf{r}_j^*(\tau^*) - \mathbf{r}_j^*(0)]^2 \right\rangle}_{MSD} \quad (4.2b)$$

In Eq. (4.2),  $\tau^*$  is the correlation time,  $N$  is the number of particles, and  $\mathbf{r}_j^*$  and  $\mathbf{v}_j^*$  refer to the position and velocity vectors of the  $j$ -th particle, respectively. The velocity integral in Eq. (4.2a) is readily available when integrating molecular dynamics simulations. As shown in Eq. (4.2b), this integral corresponds to the position of the particles in the simulation.

The shear viscosity is directly related to the time correlation of the pressure tensor. For the case of anisotropic fluids, the shear viscosity ( $\eta_{ij}^*$ ) on the direction  $ij$  is obtained as follows.

$$\eta_{ij}^* = \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \frac{V^*}{T^*} \underbrace{\left\langle \left( \int_0^{\tau^*} P_{ij}^*(\tau^*) d\tau^* \right)^2 \right\rangle}_{MSD} \quad (4.3)$$

Here,  $V^* = N/\rho^*$  refers to the total volume of the simulation box in reduced units,  $T^*$  is the simulation temperature, and  $P_{ij}^*$  is the off-diagonal component of the pressure tensor in the  $ij$  direction. In the case of isotropic fluids, the statistics can be improved by computing the shear viscosity ( $\eta^*$ ) from the average of all of the components of the symmetric traceless pressure tensor ( $P_{ij}^{*,os}$ ) [236, 237], as shown

below.

$$\eta^* = \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{10} \frac{1}{2} \frac{V^*}{T^*} \underbrace{\left\langle \sum_{ij} \left[ \int_0^{\tau^*} P_{ij}^{*,os}(\tau^*) d\tau^* \right]^2 \right\rangle}_{MSD} \quad (4.4)$$

$$\text{where, } P_{ij}^{*,os} = \frac{P_{ij}^* + P_{ji}^*}{2} - \delta_{ij} \left[ \frac{1}{3} \sum_k P_{kk}^* \right]$$

In Eq. (4.4),  $\delta_{ij}$  is the Kronecker delta function. The extra factor of  $1/10$  in this equation results from the  $3/3$  contribution from each one of the six off-diagonal terms of the traceless pressure tensor, plus the  $4/3$  implicit contribution of each diagonal term (i.e.,  $6 \cdot 3/3 + 3 \cdot 4/3 = 10$ ).

Finally, the isotropic thermal conductivity ( $\kappa^*$ ) is related to the average heat flux in the simulation.

$$\kappa^* = \lim_{\tau^* \rightarrow \infty} \frac{d}{d\tau^*} \frac{1}{2} \frac{V^*}{T^{*2}} \underbrace{\left\langle \left[ \int_0^{\tau^*} J^*(\tau^*) d\tau^* \right]^2 \right\rangle}_{MSD} \quad (4.5)$$

In Eq. (4.5),  $J^*$  is the average heat flux. The heat flux vector ( $\mathbf{J}^*$ ) can be obtained for systems interacting via pairwise potentials as follows.

$$\mathbf{J}^* = \frac{1}{V^*} \frac{1}{2} \sum_{k=1}^N \mathbf{v}_k^* \left[ \mathbf{v}_k^{*2} + \sum_{j=1, j \neq k}^N \left[ \mathcal{U}_{jk}^* + \mathbf{r}_{jk}^* \cdot \mathbf{f}_{jk}^* \right] \right] \quad (4.6)$$

Here,  $N$  corresponds to the total number of particles,  $\mathbf{v}_k^*$  is the velocity vector of the particle  $k$  and  $\mathcal{U}_{jk}^*$ ,  $\mathbf{f}_{jk}^*$  and  $\mathbf{r}_{jk}^*$  are the interaction energy, force and distance vector between the particles  $j$  and  $k$ . Note that this equation is incorrectly reported in Eq. (S13) of Ref. [235] and Eq. (7) of Ref. [2]. In both references, this equation is missing the  $1/V^*$  factor.

### Finite-size effects

Molecular simulations require periodic boundary conditions to mimic infinite-sized systems [66, 70]. These artificially larger systems allow the computation of volumetric properties, such as those obtained in Chapter 2, at the thermodynamic limit. Contrarily, it is well known that the calculation of some transport properties, like the self-diffusivity coefficient, suffers from finite-size effects [235, 238–241]. The computed diffusion coefficient using Eq. (4.2) scales linearly with the inverse of the simulation box size [238, 239]. In the case of a cubic simulation box (i.e.,  $L^* = L_x^* = L_y^* = L_z^*$ ), Yeh and Hummer [239] derived the following expression to obtain the self-diffusivity of a virtually infinite-sized system ( $D_\infty^*$ ).

$$D_\infty^* = D^* + \underbrace{\frac{T^* \xi}{6\pi\eta^* L^*}}_{\Delta D_{\text{corr}}^*} \quad (4.7)$$

Here,  $D^*$  is the self-diffusivity obtained from Eq. (4.2),  $\Delta D_{\text{corr}}^*$  is the finite-size correction term,  $T^*$  is the temperature of the system,  $\xi$  is a dimensionless universal constant equal to 2.837298,  $L^* = \sqrt[3]{N/\rho^*}$  is the simulation box length, and  $\eta^*$  is the shear viscosity computed from Eq. (4.4). This equation requires the shear viscosity to be computed at the same conditions as the self-diffusivity. However,

as will be shown later in Section 4.2.2, the computed shear viscosity results tend to be noisier, which can potentially introduce noise into the extrapolated  $D_{\infty}^*$  value. Finally, unlike the self-diffusivity, the shear viscosity and thermal conductivity do not exhibit appreciable finite-size effects [235, 239].

### Molecular simulation setup

The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software (stable release June 2022) [147], alongside the On-the-fly Calculation of Transport Properties (OCTP) plugin [235], are used to compute the transport properties of the Mie fluid. The simulations are conducted using reduced units as defined in Table 4.1.

MD simulations are performed in a cubic simulation box with 4096 Mie particles. This relatively large number of particles is chosen to minimize any possible finite-size effects in two ways. Firstly, a large simulation box reduces the impact of the correction term in Eq. (4.7) in the extrapolated  $D_{\infty}^*$ . Secondly, it also minimizes finite-size effects of the shear viscosity and thermal conductivity, for which no analytical correction is available. The MD simulations are performed using a cutoff radius of  $r_c = 5\sigma$ , with no long-range corrections applied. The simulations are run with a timestep of  $\Delta\tau^* = 0.002\tau^*$ .

The simulation is first equilibrated for  $10^6$  timesteps in the canonical (NVT) ensemble using the Nosé-Hoover thermostat [150] with a time constant of  $100\Delta\tau^*$ . As recommended by Maginn et al. [231], the transport properties calculation should be performed in the microcanonical (NVE) ensemble to avoid external perturbations to the dynamics of the system. Therefore, once the system is equilibrated at a desired temperature, the simulation is integrated in the NVE ensemble for  $2.9 \cdot 10^6$  timesteps using the velocity Verlet algorithm [242]. During this production stage, statistics of the systems are accumulated every 1000 timesteps. The MSDs to compute self-diffusivity (Eq. (4.2b)), shear viscosity (Eq. (4.4)), and thermal conductivity (Eq. (4.5)) are obtained using the OCTP plugin [235]. This plugin efficiently samples the three required MSDs within a single simulation using the n-order algorithm proposed by Frenkel and Smit [70, Chapter 4]. As recommended by Jamali et al. [235], the MSD required for the self-diffusivity coefficient is sampled every 1000 timesteps, while the MSDs for the shear viscosity and thermal conductivity are sampled every 5 steps. This frequent sampling is required to obtain more accurate results in the integral of Eqs. (4.4) and (4.5). This integral is evaluated numerically using Simpson's rule. Finally, the MSDs are obtained for each simulation at ten different starting points, separated by  $10^5$  time steps.

### Database curation

The transport properties are obtained by analyzing the sampled MSDs using standard NumPy [154] and SciPy [146] functions. The general procedure consists of two steps. First, the region where the Einstein method is valid is identified by filtering the region where the slope of the MSD with respect to time is close to one in a log-log plot. This procedure is automated by fitting the natural logarithm of the MSD against the natural logarithm of time using a cubic spline. The derivative of the fitted spline is directly computed, and the region where the slope is  $1 \pm 0.075$  is filtered. From this region, the corresponding transport property is computed using either Eq. (4.2), (4.4), or (4.5), for the self-diffusivity, shear viscosity or thermal conductivity, respectively. The procedure is repeated for each of the ten computed

MSDs. Then, mean values and standard deviations of each transport property are obtained for each state point.

The correctness of the transport properties database is assessed using the same four sanity tests detailed in Section 2.4.2. The first test checks the consistency of a data point in different ensembles. In the second sanity test, high variance data points where the relative standard deviation (i.e.,  $100|\sigma_{\Lambda^*}/\langle\Lambda^*\rangle|$ ) is greater than 20% are disregarded. This relatively high tolerance is chosen to eliminate unreliable data points while maintaining sufficient phase space coverage, which is necessary for data-driven modeling. The modeling step will use regularization techniques to avoid overfitting due to the inherent noise of the transport property results. The third sanity test analyzes the RDF of data points close to the SLE boundaries. Data points that entered a glassy metastable state are disregarded. In the fourth and final sanity test, the `LocalOutlierFactor` [155] method implemented in `Scikit-Learn` [156] is used to disregard possible outliers. This test provides a score that measures the "outlierness" of a data point. These scores are fit to a probability distribution using the `fitter` [158] Python package, and data points outside the 95% confidence interval are disregarded [157].

#### 4.2.2 Transport properties database description

The Mie fluid transport properties database is summarized in Table 4.2. The database consists of transport properties for 28 Mie fluids that span for reduced densities ( $\rho^*$ ) ranging from  $5 \cdot 10^{-3}$  up to 1.26 and reduced temperatures ( $T^*$ ) from 0.6 up to 10.0. The attractive exponent ( $\lambda_a$ ) was set to 6, whereas the repulsive exponent ( $\lambda_r$ ) varied from 7 to 34. This combination of exponents corresponds to  $\alpha_{vdw}$  parameters, Eq. (2.21), in the range from 0.53 to 1.47, which cover industrially relevant low molecular weight fluids [94, 97]. After disregarding high-variance results and possible outliers, as mentioned in Section 4.2.1, the final database consists of 18140, 15529, and 14317 data points for the self-diffusivity, shear viscosity, and thermal conductivity, respectively. The phase space distribution for each transport property is shown in Figure 4.1. The database was purposely populated close to the phase envelope for temperatures up to 2.0. Higher temperature isotherms up to 10.0 are sparsely covered. Similar to the thermophysical properties database obtained in Chapter 2, the density distribution of the database is evenly distributed except at high densities and low temperatures. This unpopulated area at high densities corresponds to the region where Mie fluid freezes [92, 151]. Finally, from Figure 4.1, it can also be observed that this dataset covers approximately 600, 500, and 450 data points per Mie fluid for the self-diffusivity, shear viscosity, and thermal conductivity, respectively.

The database is assessed against a compiled transport properties database for the Lennard-Jones fluid ( $\lambda_r = 12$  and  $\lambda_a = 6$ ) [210]. In Figure 4.2, the uncorrected self-diffusivity, shear viscosity, and thermal conductivity computed in this work are compared against published transport properties at four different temperatures, namely,  $T^* = 1.2, 2.0, 6.0$  and  $10.0$ . In this figure, the source of the data point is identified by its marker, while the color denotes the temperature of the data point. Across the board, it can be observed that the computed transport properties are in good agreement with the published Lennard-Jones data. The shear viscosity and thermal conductivity results exhibit higher deviations (i.e., more noticeable error bars) than the self-diffusivity. The deviations tend to increase with temperature and density. This problem has been well-reported in the literature and has several underlying causes. Firstly, the self-diffusivity is averaged for all the particles in the simulation, which

TABLE 4.2: Mie fluid transport properties database description.<sup>a</sup>

| Descriptor                           | Symbol         | # of data points | Min value | Max value |
|--------------------------------------|----------------|------------------|-----------|-----------|
| Repulsive exponent <sup>b</sup>      | $\lambda_r$    | –                | 7         | 34        |
| Attractive exponent <sup>b</sup>     | $\lambda_a$    | –                | 6         | 6         |
| van der Waals parameter <sup>c</sup> | $\alpha_{vdw}$ | –                | 0.53      | 1.47      |
| Density                              | $\rho^*$       | –                | 5.00e-03  | 1.26      |
| Temperature                          | $T^*$          | –                | 0.60      | 10.00     |
| Transport Property                   |                |                  |           |           |
| Self-diffusivity                     | $D^*$          | 18140            | 3.40e-04  | 203.46    |
| Shear viscosity                      | $\eta^*$       | 15529            | 7.29e-02  | 20.72     |
| Thermal conductivity                 | $\kappa^*$     | 14317            | 0.36      | 40.45     |

<sup>a</sup> Descriptors and transport properties are reported in reduced units. Refer to Table 4.1 for further details. <sup>b</sup> Repulsive and attractive exponent of the Mie potential, Eq. (1.1). <sup>c</sup> van der Waals parameter is a descriptor of the Mie fluid. See Eq. (2.21).

leads to better-sampled statistics, see Eq. (4.2). Contrarily, the shear viscosity and thermal conductivity are "only" averaged for each direction when studying isotropic fluids. Finally, a further source of deviations for the shear viscosity and thermal conductivity comes from the numerical integration of Eqs. (4.4) and (4.5) using Simpson's rule [235]. In contrast, this integral is analytically available from the step integration algorithm for the self-diffusivity, refer to Eq. (4.2b). Secondly, it is well reported that different trajectories can lead to significantly different transport properties [231, 234]. It is mandatory to perform replicas when computing transport properties from molecular simulations. This work addresses this problem by using ten MSDs per state point. Despite the higher deviations in the shear viscosity and thermal conductivity, the database still shows clear trends that will allow the development of data-driven transport property models.

Finally, it must be recalled that the self-diffusivity suffers from finite-size effects, and its value can be extrapolated to an infinite-sized system using Eq. (4.7). The correction term requires the shear viscosity to be computed at the same conditions as the self-diffusivity, which implies that approximately only 2/3 of the available data points can be corrected. The relative importance of the correction term  $\Delta D_{\text{corr}}^*$  is shown in Figure 4.3. This figure shows the ratio of the correction term to the self-diffusivity ( $\% \Delta D_{\text{corr}}^* = 100 \Delta D_{\text{corr}}^* / D^*$ ) for different densities and temperatures. It can be observed from Figure 4.3 (a) that the relative correction term is more significant at high densities, where the self-diffusivity is small. Contrarily, as indicated by the colormap, the temperature does not play a significant role in the relative magnitude of the correction term. Finally, from Figure 4.3 (b), it can be observed that the finite-size correction term increases on average a 6% of the self-diffusivity computed from MD for the simulated system size. This correction term is not negligible and should be considered when modeling the self-diffusivity in practical applications.

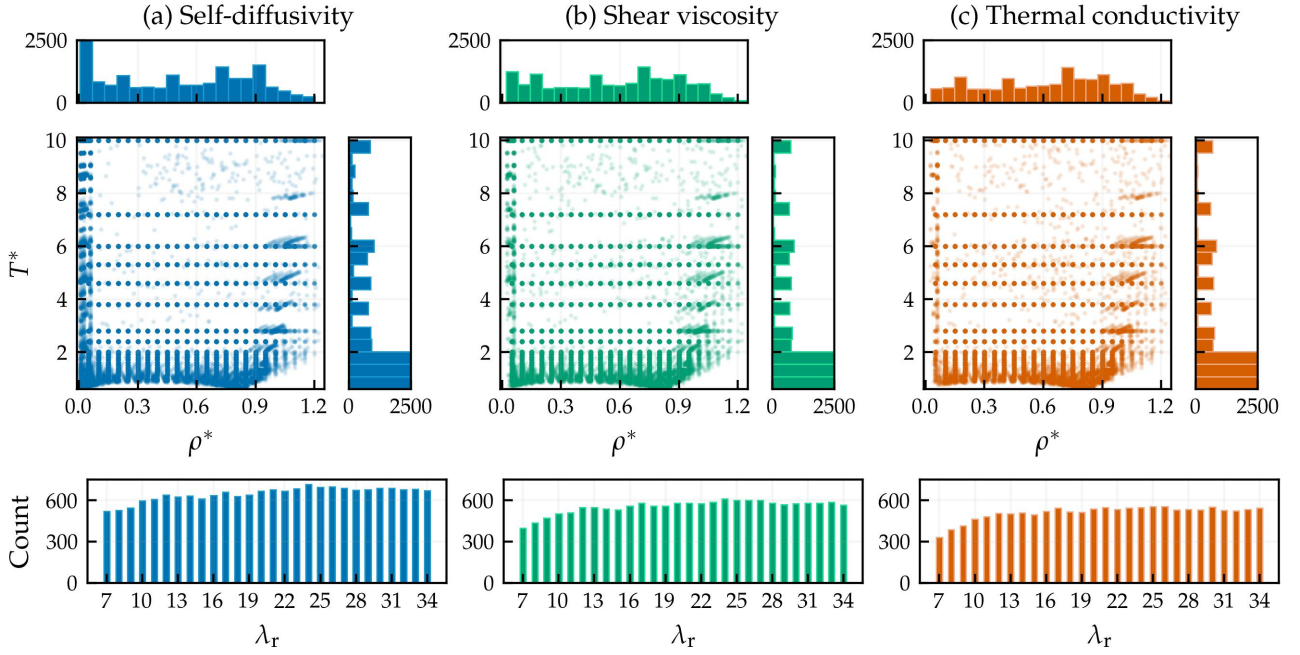


FIGURE 4.1: **Mie fluid transport properties database distribution.** (a) Self-diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Top plot: density-temperature ( $\rho^*$ - $T^*$ ) data points distribution. Bottom plot: repulsive exponent ( $\lambda_r$ ) distribution.

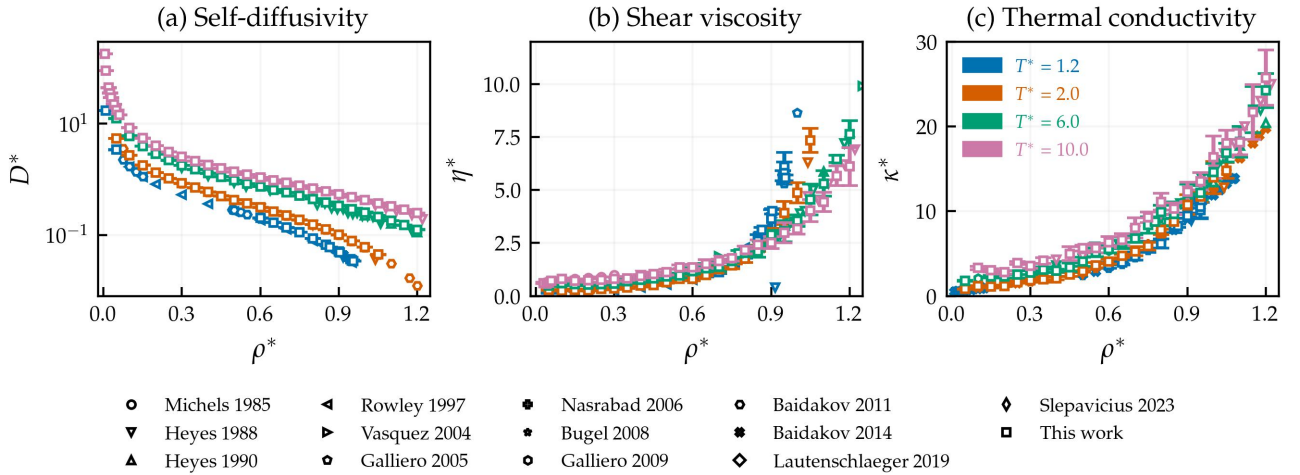


FIGURE 4.2: **Transport data simulation data of the Lennard-Jones fluid** ( $\lambda_r = 12$  and  $\lambda_a = 6$ ). (a) Self-diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Blue symbols:  $T^* = 1.2$ , orange symbols:  $T^* = 2.0$ , green symbols:  $T^* = 6.0$ , pink symbols:  $T^* = 10.0$ . Published molecular simulation data: circles [243], downward triangle [244], upright triangle [245], left-pointing triangle [246], right-pointing triangle [247], pentagon [248], "plus" sign [249], star [250], hexagon [251], side hexagon [252], "X" sign [253], diamond [210], thin diamond [224]. This work: squares. Errors bars refer to the 95% confidence interval.



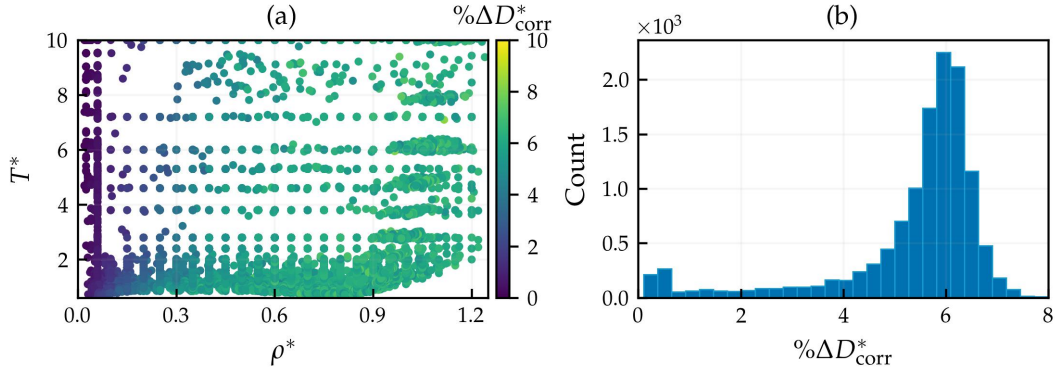


FIGURE 4.3: **Finite-size effect correction to the self-diffusivity.** (a) Ratio of the correction term to the self-diffusivity ( $\% \Delta D_{\text{corr}}^* = 100 \Delta D_{\text{corr}}^* / D^*$ ) as a function of density and temperature. (b) Histogram of the ratio of the correction term to the self-diffusivity ( $\% \Delta D_{\text{corr}}^*$ ).

### 4.3 Physically-inspired data-driven modeling of the Mie fluid transport properties

#### 4.3.1 Transport property data-driven models formulation

In general, a transport property ( $\Lambda^*$ ) of the Mie fluid is characterized by the molecular descriptors (i.e., molecular parameters) and by its state conditions (i.e., density and temperature). The Mie potential is fully characterized by its repulsive ( $\lambda_r$ ) and attractive ( $\lambda_a$ ) exponents when using reduced units. Moreover, without any loss of generality, it is convenient to describe the Mie fluid using the van der Waals parameter ( $\alpha_{\text{vdw}}$ ) [92].

$$\alpha_{\text{vdw}} = \underbrace{\frac{\lambda_r}{\lambda_r - \lambda_a} \left( \frac{\lambda_r}{\lambda_a} \right)^{\frac{\lambda_a}{\lambda_r - \lambda_a}}}_{\mathcal{C}^{\text{Mie}}} \left[ \frac{1}{\lambda_a - 3} - \frac{1}{\lambda_r - 3} \right] \quad (2.21)$$

In principle, a Mie fluid transport property can be represented as a function of the van der Waals constant ( $\alpha_{\text{vdw}}$ ), density ( $\rho^*$ ), and temperature ( $T^*$ ), i.e.,  $\Lambda^* = \Lambda^*(\alpha_{\text{vdw}}, \rho^*, T^*)$ . However, as pointed out previously in Chapter 2 [3], it is more convenient to consider an inverse of the temperature functionality ( $\Lambda^* = \Lambda^*(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$ ). The temperature scale shrinks by using the inverse of the temperature. This numerical transformation eases the interpolation efforts of a data-driven model. Additionally, the inverse of the temperature dependence is found in thermodynamic scaling [254], which has its roots in isomorph theory [218]. In thermodynamic scaling, a transport property is considered a function of the  $(\rho^*)^\delta / T^*$  ratio, where  $\delta$  is a scaling exponent.

Alternatively, Rosenfeld [217] suggested that a scaled transport property ( $\tilde{\Lambda}^*$ ) is a univariate function of the residual entropy ( $S^{*,\text{res}}$ ), i.e.,  $\tilde{\Lambda}^* = \tilde{\Lambda}^*(\alpha_{\text{vdw}}, -S^{*,\text{res}})$ . In this framework, a.k.a. entropy scaling,

the scaled transport properties are defined as in Eqs. (4.8).

$$\tilde{D}^* = \frac{(\rho^*)^{1/3} D^*}{\sqrt{T^*}} \quad (4.8a)$$

$$\tilde{\eta}^* = \frac{\eta^*}{(\rho^*)^{2/3} \sqrt{T^*}} \quad (4.8b)$$

$$\tilde{\kappa}^* = \frac{\kappa^*}{(\rho^*)^{2/3} \sqrt{T^*}} \quad (4.8c)$$

In the entropy scaling framework, the residual entropy can be obtained from an equation of state that explicitly models the Helmholtz free energy. For the Mie fluid, several options are available [37, 113, 115]. In this chapter, the residual entropy is obtained from the FE-ANN EoS developed in Chapter 2 [3]. The FE-ANN EoS is defined as follows.

$$A^{*,\text{res}} = \text{ANN}(\alpha_{\text{vdw}}, \rho^*, \frac{1}{T^*}) - \text{ANN}(\alpha_{\text{vdw}}, \rho^* = 0, \frac{1}{T^*}) \quad (2.22b)$$

This EoS evaluates the same Artificial Neural Network (ANN) twice, at the given density  $\rho^*$ , and  $\rho^* = 0$ . This formulation guarantees fulfilling the ideal gas law limit and Maxwell's relationships. As discussed in Section 2.5, the FE-ANN EoS accurately models the first- and second-order thermophysical properties of the Mie fluid over a vast range of densities and temperatures. The residual entropy is obtained using Eq. (2.1a), which consists on taking the isochoric temperature derivative of Eq. (2.22b) as follows:  $S^{*,\text{res}} = -(\partial A^{*,\text{res}} / \partial T^*)_{\rho^*}$ .

Independent of which modeling approach is chosen (i.e.,  $\Lambda^*(\alpha_{\text{vdw}}, \rho^*, 1/T^*)$  or  $\tilde{\Lambda}^*(\alpha_{\text{vdw}}, -S^{*,\text{res}})$ ), the exact functionality to link the descriptors and state variables to the transport property is usually not known. For this reason, empirical correlations based on exponential terms and polynomial expansions have been widely used in the literature [210, 219, 222]. In this chapter, the functionality of the transport property with respect to its variables is "learned" using artificial neural networks (ANNs), as follows.

$$\Lambda^* = \text{ANN}(\alpha_{\text{vdw}}, \rho^*, 1/T^*) \quad (4.9a)$$

$$\tilde{\Lambda}^* = \text{ANN}(\alpha_{\text{vdw}}, -S^{*,\text{res}}) \quad (4.9b)$$

Artificial neural networks [52] are a deep learning algorithm that can act as a universal function approximation [53]. A full description of artificial neural networks is given in Appendix A. For completeness, a brief description is repeated here. An ANN consists of a series of linear transformations followed by a non-linear transformation, also known as an activation function. An n-layered ANN is mathematically represented as follows,

$$\mathbf{y} = \underbrace{f_n(\mathbf{W}_n \cdots f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \cdots \mathbf{b}_n)}_{\text{ANN}(\mathbf{x})} \quad (4.10)$$

In Eq. (4.10),  $\mathbf{x}$  and  $\mathbf{y}$  are the input and output features of the ANN, respectively. The  $i$ -th layer in the ANN consists of a linear transformation done by the weight matrix  $\mathbf{W}_i$  and the bias vector  $\mathbf{b}_i$ . The dimensions of the weight matrix and bias vector are  $(m_i \times n_{i-1})$  and  $m_i$ , respectively, where  $m_i$  is the number of neurons on the  $i$ -th layer and  $n_{i-1}$  is the dimension of the output vector from the previous

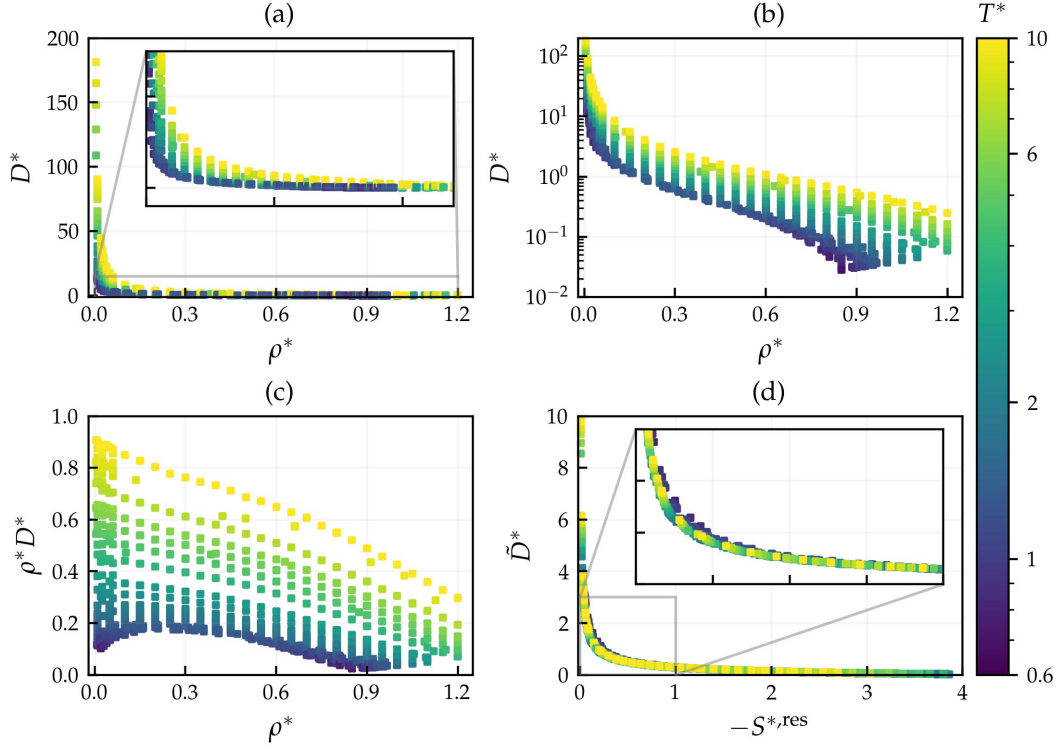


FIGURE 4.4: **Self diffusivity of the Lennard-Jones fluid** ( $\lambda_r = 12$  and  $\lambda_a = 6$ ). (a)  $D^*$  - linear scale. (b)  $D^*$  - semi-log scale. (c)  $\rho^* D^*$  - linear scale. (d) Reduced self-diffusivity,  $\tilde{D}^*$  - linear scale, see Eq. (4.8a). The colormap refers to the temperature.

layer. The linear transformation is followed by the activation function  $f_i$ , which adds non-linearity to the model [135]. In this work, the hidden layers use the tanh activation function ( $f(x) = \tanh(x)$ ) while the output layer can use either the linear ( $f(x) = x$ ) or softplus ( $f(x) = \ln(1 + \exp(x))$ ) activation functions. This choice of activation functions ensures that the output of the ANN is always continuous and differentiable with respect to the inputs.

It is important to consider the numerically convenient and physics-informed outputs for a data-driven model [3, 58]. In Figure 4.4, different representations of the uncorrected self-diffusivity of the Lennard-Jones fluid ( $\lambda_r = 12$  and  $\lambda_a = 6$ ) are shown. The temperature of each data point is indicated by the attached color bar. As can be observed from Figure 4.4 (a), modeling the self-diffusivity ( $D^*$ ) with a linear scale results in a challenging task. The value of the transport property diverges to infinity at low densities. As shown in Figure 4.4 (b), this numerical inconvenience can be bypassed using a semi-log scale. Similarly, modeling the  $\rho^* D^*$  product instead of the self-diffusivity itself is numerically convenient as the values are easily separable in the  $\rho^* - T^*$  space. Additionally, the values are bounded to a narrow range (0 and 1 for the available database), which acts as a regularization technique for a machine-learning model. Finally, from Figure 4.4 (d), it can be observed that the scaled self-diffusivity, Eq. (4.8a), is conveniently described as a function of the residual entropy. However, there is a range (zoomed region) where the residual entropy corresponds to multiple self-diffusivity values. This behavior has also been observed in other entropy scaling studies of the self-diffusivity [220, 221] and will likely produce conflicting regions when training a data-driven model. At this point, several possible outputs for the data-driven model can be considered to represent the self-diffusivity.

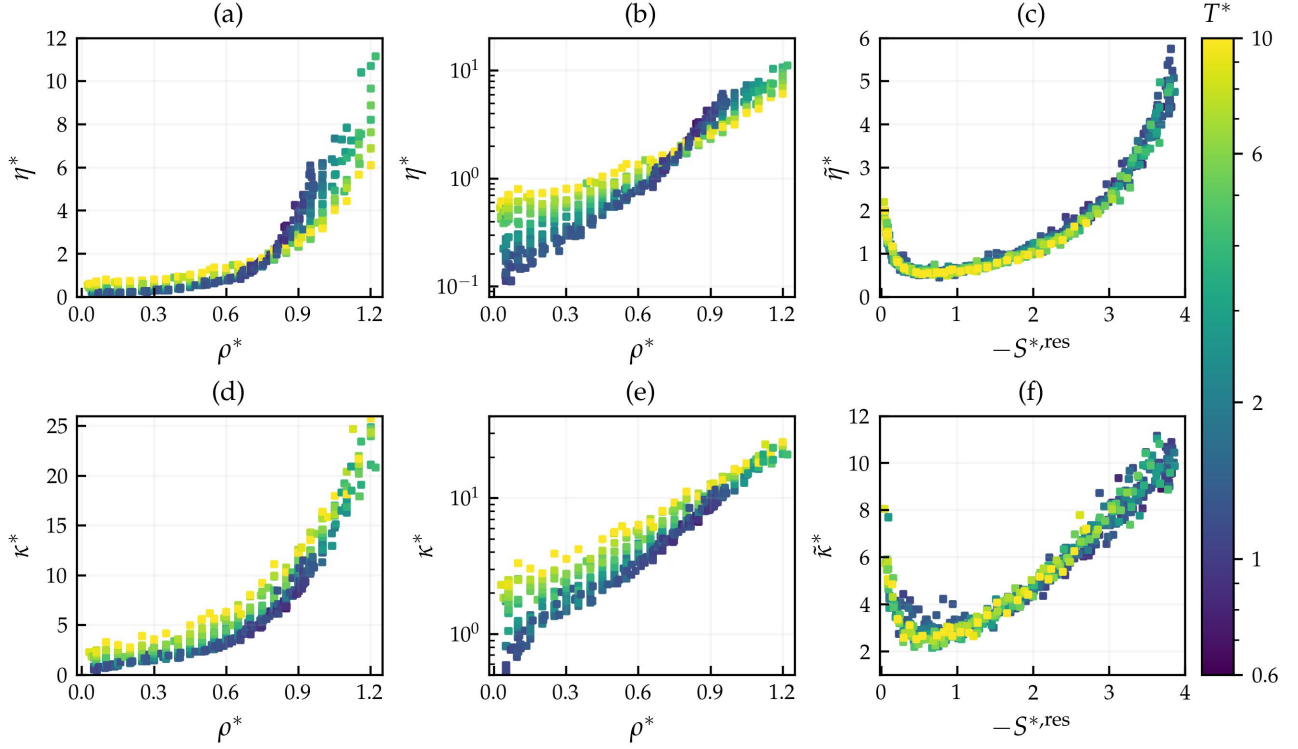


FIGURE 4.5: **Shear viscosity and thermal conductivity of the Lennard-Jones fluid** ( $\lambda_r = 12$  and  $\lambda_a = 6$ ). Shear viscosity. (a)  $\eta^*$  - linear scale. (b)  $\eta^*$  - semi-log scale. (c) Reduced shear viscosity,  $\tilde{\eta}^*$  - linear scale, see Eq. (4.8b). Thermal conductivity. (d)  $\kappa^*$  - linear scale. (e)  $\kappa^*$  - semi-log scale. (f) Reduced thermal conductivity,  $\tilde{\kappa}^*$  - linear scale, see Eq. (4.8c). The colormap refers to the temperature.

A similar analysis can be done for the shear viscosity and thermal conductivity. Different numerical representations of these transport properties of the Lennard-Jones fluid ( $\lambda_r = 12$  and  $\lambda_a = 6$ ) are shown in Figure 4.5. In contrast to the self-diffusivity, these properties tend to increase with density, and no divergences are observed when using a linear scale, see Figures 4.5 (a) and (d). Nevertheless, the use of a semi-log scale helps in two ways. As shown in Figures 4.5 (b) and (e), the semi-log scale shrinks the output scale, which acts as a regularizing step for an ANN. Additionally, a "simpler", more linear dependence is observed in the isotherms. The combination of both aspects will benefit the training of an ANN. Finally, from Figures 4.5 (b) and (e), it can be observed that the scaled shear viscosity and thermal conductivity can potentially be described as a univariate function of the residual entropy. These scaled properties show an expected level of dispersion due to the inherent noise of the transport property data.

In summary, two main transport property modeling approaches have been proposed here. The first one, Eq. (4.9a), directly uses the density and temperature as state variables, while the second one, Eq. (4.9b), uses entropy scaling. The representation of the target transport property (e.g., linear scale, semi-log scale) plays a key role in the success of the modeling. The different modeling approaches will be tested later in Section 4.3.4. Until now, the proposed data-driven models do not consider any special treatment to handle the dilute gas state ( $\rho^* \rightarrow 0$ ). However, this low-density region can be well-described from kinetic theory [27, 255]. The following section will focus on obtaining the dilute gas transport properties of the Mie fluid and will discuss how to include that information in the data-driven model.

### 4.3.2 Transport properties of the dilute Mie gas

In the context of kinetic theory, a.k.a. Chapman-Enskog theory, the transport property of a dilute gas is related to an expansion based on collision integrals ( $\Omega^{(l,s)*}$ ) [27, 255]. For the explored temperature conditions in this work ( $T^* \leq 10$ ), the first-order approximation (denoted by  $\Lambda_{[1]}^*$ ) is sufficient to describe a transport property within a 1% error [27]. In this case, the transport properties of the dilute gas are obtained as follows.

$$\rho^* D_{[1]}^* = \frac{3}{8} \frac{1}{\Omega^{(1,1)*}} \sqrt{\frac{T^*}{\pi}} \quad (4.11a)$$

$$\eta_{[1]}^* = \frac{5}{12} \frac{1}{\Omega^{(2,2)*}} \sqrt{\frac{T^*}{\pi}} \quad (4.11b)$$

$$\kappa_{[1]}^* = \frac{25}{32} \frac{C_V^*}{\Omega^{(2,2)*}} \sqrt{\frac{T^*}{\pi}} \quad (4.11c)$$

In Eqs. (4.11),  $D_{[1]}^*$ ,  $\eta_{[1]}^*$  and  $\kappa_{[1]}^*$  are the first-order approximation for the self-diffusivity, shear viscosity, and thermal conductivity, respectively,  $C_V^* = C_V/R$  is the dimensionless isochoric heat capacity, equal to 1.5 for monoatomic molecules at zero density, and  $\Omega^{(l,s)*}$  refers to a collision integral [27, 255]. The calculation of the collision integral is not trivial, however, Fokin et al. [47] computed and correlated the values of  $\Omega^{(1,1)*}$  and  $\Omega^{(2,2)*}$  for  $\lambda_r - 6$  Mie fluids with  $\lambda_r \in [8, \infty]$  and  $T^* \in [0.4, 200]$ . In this correlation, the collision integral  $\Omega^{(k,k)*}$  is obtained as follows.

$$\ln \Omega^{(k,k)*} = -\frac{2}{\lambda_r} \ln T^* + \delta^{(k,k)} \ln \left( 1 - \frac{2}{3\lambda_r} \right) + \sum_{i=1}^6 a_i^{(k,k)}(\lambda_r) \left( \frac{1}{T^*} \right)^{(i-1)/2} \quad (4.12)$$

Here,  $\delta^{(1,1)} = 0$  and  $\delta^{(2,2)} = 1$ . The coefficients  $a_i^{(k,k)}(\lambda_r)$  are obtained from the expansion defined in Eq. (4.13). The parameters  $a_{ij}^{(k,k)}$  needed to evaluate this expansion are given in Table 4.3.

$$a_i^{(k,k)}(\lambda_r) = a_{i1}^{(k,k)} + \frac{a_{i2}^{(k,k)}}{\lambda_r} + \frac{a_{i3}^{(k,k)}}{\lambda_r^2} + \frac{a_{i4}^{(k,k)}}{\lambda_r^2} \ln \lambda_r \quad (4.13)$$

The dilute gas transport properties, defined in Eq. (4.11), and computed with the collision integral correlated in Eq. (4.12), can be embedded into a data-driven model in two ways. The first option is to generate data artificially by evaluating these equations and then adding that data when training the model. The second modeling option is formulated to exactly fulfill the first-order transport property approximation from kinetic theory. In this case, the data-driven models are formulated as follows.

$$\Lambda^* = \Lambda_{[1]}^* + \underbrace{\text{ANN}(\alpha_{\text{vdw}}(\lambda_r, 6), \rho^*, 1/T^*) - \text{ANN}(\alpha_{\text{vdw}}(\lambda_r, 6), \rho^* = 0, 1/T^*)}_{\Lambda^{*,\text{res}}} \quad (4.14a)$$

or,

$$\ln \Lambda^* = \ln \Lambda_{[1]}^* + \underbrace{\text{ANN}(\alpha_{\text{vdw}}(\lambda_r, 6), \rho^*, 1/T^*) - \text{ANN}(\alpha_{\text{vdw}}(\lambda_r, 6), \rho^* = 0, 1/T^*)}_{\ln \Lambda^{*,\text{res}}} \quad (4.14b)$$

These data-driven models are formulated to learn only the "residual" contribution of a transport property ( $\Lambda^{*,\text{res}}$ ). This formulation exactly fulfills the first-order kinetic theory approximation at zero density. The dilute gas limit is obtained using Eqs. (4.11) and the correlations of  $\Omega^{(1,1)*}$  and  $\Omega^{(2,2)*}$

TABLE 4.3: Coefficients  $a_{ij}^{(k,k)}$  needed to evaluate the collision integral of  $(\lambda_r, \lambda_a = 6)$  Mie fluids.<sup>a, b</sup>

| $a^{(k,k)}$ | $k = (1, 1)$                | $k = (2, 2)$                | $a^{(k,k)}$ | $k = (1, 1)$                | $k = (2, 2)$              |
|-------------|-----------------------------|-----------------------------|-------------|-----------------------------|---------------------------|
| $a_{11}$    | 0.0                         | 0.0                         | $a_{41}$    | 0.485 352                   | 0.697 682                 |
| $a_{12}$    | $-0.145\,269 \times 10^1$   | $0.113\,086 \times 10^1$    | $a_{42}$    | $0.245\,523 \times 10^2$    | $0.590\,192 \times 10^2$  |
| $a_{13}$    | $0.294\,682 \times 10^2$    | $0.234\,799 \times 10^2$    | $a_{43}$    | $-0.336\,782 \times 10^3$   | $-0.143\,670 \times 10^3$ |
| $a_{14}$    | $0.242\,508 \times 10^1$    | $0.310\,127 \times 10^1$    | $a_{44}$    | $0.814\,187 \times 10^2$    | $-0.123\,518 \times 10^3$ |
| $a_{21}$    | $0.107\,782 \times 10^{-1}$ | 0.0                         | $a_{51}$    | $-0.385\,355$               | $-0.564\,238$             |
| $a_{22}$    | 0.587 725                   | $0.551\,559 \times 10^1$    | $a_{52}$    | $-0.206\,868 \times 10^2$   | $-0.430\,549 \times 10^2$ |
| $a_{23}$    | $-0.180\,714 \times 10^3$   | $-0.137\,023 \times 10^3$   | $a_{53}$    | $0.132\,246 \times 10^3$    | 0.0                       |
| $a_{24}$    | $0.595\,694 \times 10^2$    | $0.185\,848 \times 10^2$    | $a_{54}$    | 0.0                         | $0.137\,282 \times 10^3$  |
| $a_{31}$    | $0.546\,646 \times 10^{-1}$ | $0.325\,909 \times 10^{-1}$ | $a_{61}$    | $0.847\,232 \times 10^{-1}$ | 0.126 508                 |
| $a_{32}$    | $-0.651\,465 \times 10^1$   | $-0.292\,925 \times 10^2$   | $a_{62}$    | $0.521\,812 \times 10^1$    | $0.104\,273 \times 10^2$  |
| $a_{33}$    | $0.374\,457 \times 10^3$    | $0.243\,741 \times 10^3$    | $a_{63}$    | $-0.181\,140 \times 10^2$   | $0.150\,601 \times 10^2$  |
| $a_{34}$    | $-0.137\,807 \times 10^3$   | 0.0                         | $a_{64}$    | $-0.747\,215 \times 10^1$   | $-0.408\,911 \times 10^2$ |

<sup>a</sup> Coefficients  $a_{ij}^{(k,k)}$  required in Eq. (4.13). <sup>b</sup> Coefficients correlated by Fokin et al. [47].

developed by Fokin et al. [47]. These residual models can be formulated using a linear, Eq. (4.14a) or a semi-log scale, Eq. (4.14b). In the latter, the transport property is obtained as  $\Lambda^* = \Lambda_{[1]}^* * \Lambda^{*,\text{res}}$ . Due to the restrictions of the collision integral correlations, these models are only valid for  $\lambda_r = 6$  Mie fluids with  $\lambda_r \geq 8$ .

It must be noted that the dilute-gas limit ( $\rho^* \rightarrow 0$ ) cannot always be incorporated into a transport property model. In the case of the self-diffusivity,  $D^*$  diverges at zero density, making this representation incompatible with the dilute gas data. However, the product  $\rho^* D^*$  is well defined at zero density, and this representation allows including this information. Similarly, the dilute gas limit data is incompatible with the entropy scaling approach due to a division by zero in the definition of the scaled transport properties, see Eqs. (4.8). That being said, the dilute gas transport property data will be included in the training of data-driven models whenever that limit is well-defined and no numerical divergences are expected.

### 4.3.3 Transport property data-driven models training and architecture

Several data-driven transport property models have been proposed. Transport properties can be directly modeled using Eqs. (4.9a), using entropy scaling as in Eq. (4.9b), or by modeling the "residual" transport property as defined in Eqs. (4.14). Regardless of the chosen modeling approach, these models rely on artificial neural networks and their trainable parameters, i.e., weights ( $\mathbf{W}_i$ ) and biases ( $\mathbf{b}_i$ ). These parameters must be optimized to reproduce the reference data and to fulfill physically-inspired constraints.

The training involves updating the weights and biases to minimize the following loss function ( $\mathcal{L}$ ).

$$\underset{\mathbf{W}_i, \mathbf{b}_i}{\text{minimize}} \mathcal{L} = \mathcal{L}^{\text{MSE}} + \mathcal{L}^{\text{dilute}} + \mathcal{L}^{\text{pen}} \quad (4.15)$$

The loss function considers three main contributions. The first contribution,  $\mathcal{L}^{\text{MSE}}$ , measures the squared error with respect to the reference molecular simulation data obtained in Section 4.2. This

contribution is computed as follows.

$$\mathcal{L}^{\text{MSE}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left( \Lambda_i^{*,\text{pred}} - \Lambda_i^{*,\text{true}} \right)^2 \quad (4.16)$$

Here,  $N_p$  is the number of data points, and  $\Lambda^*$  is the target transport property. The superscripts "pred" and "true" refer to the value predicted by the data-driven model and the value available from the database, respectively.

The second contribution to the loss function,  $\mathcal{L}^{\text{dilute}}$ , measures the squared error with respect to the dilute gas transport property as obtained in Section 4.3.2. The dilute gas data was obtained for  $(\lambda_r, \lambda_a = 6)$  Mie fluids with  $\lambda_r \in [8, 34]$  for reduced temperatures ( $T^*$ ) ranging from 0.6 to 10 using Eqs. (4.11). This loss function is computed as follows.

$$\mathcal{L}^{\text{dilute}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left( \Lambda_i^{*,\text{pred}}(\rho^* = 0) - \Lambda_{[1],i}^{*,\text{true}} \right)^2 \quad (4.17)$$

In Eq. (4.17),  $N_p$  is the number of data points,  $\Lambda^{*,\text{pred}}(\rho^* = 0)$  is the predicted transport property at zero density and  $\Lambda_{[1]}^{*,\text{true}}$  is the reference first-order approximation transport property obtained from kinetic theory.

Finally, the training can be improved by imposing physical-inspired constraints to fulfill expected physical trends. For example, the self-diffusivity is expected to be a decreasing function with respect to the density. The thermal conductivity, on the other hand, is a monotonically increasing function with respect to the density, except in the vicinity of the critical point where a maximum is expected. This critical enhancement has been experimentally observed in real molecular systems [256–258], and it is linked to the heat capacity maximum on the same region [259]. Similarly, the shear viscosity exhibits a critical enhancement in a narrow region close to the critical point. This effect can be safely neglected for practical applications [221]. That being said, the shear viscosity is expected to be an increasing function of the density near the critical point and a monotonically increasing function of the density for supercritical temperatures [258]. Although a data-driven model might be expected to discover these behaviors, it is recommended to enforce it using a physics-informed approach [86]. The following penalty loss,  $\mathcal{L}^{\text{pen}}$ , is considered during training.

$$\mathcal{L}^{\text{pen}} = \begin{cases} 1/N_p \sum_{i=1}^{N_p} \max [0, (\partial \Lambda^* / \partial \rho^*)_{T^*}]^2 & \text{if } \Lambda^* = D^* \\ 1/N_p \sum_{i=1}^{N_p} \min [0, (\partial \Lambda^* / \partial \rho^*)_{T^*}]^2 & \text{if } \Lambda^* \in [\eta^*, \kappa^*] \end{cases} \quad (4.18)$$

This penalty function, Eq. (4.18), enforces that the self-diffusivity is a decreasing function with respect to density (i.e.,  $(\partial \Lambda^* / \partial \rho^*)_{T^*} < 0$ ) and that the shear viscosity and thermal conductivity are increasing functions with respect to the density (i.e.,  $(\partial \Lambda^* / \partial \rho^*)_{T^*} > 0$ ). These penalties contribute to the loss function when an unphysical behavior is detected. These penalties are applied to supercritical data points where  $T^* > 1.1T_c^*$  to avoid conflicts with the critical enhancement observed near the critical region.

The training is performed using the Adam method [139]. This optimization method is a modified

gradient descent step where information from previous iterations is considered during the minimization. This method relies on the derivatives of the loss function with respect to the weights and biases. These derivatives are obtained using automatic differentiation [88]. The Adam method relies on the learning rate ( $\alpha$ ), which controls the step during the training. The dataset is split into training and validation sets using a 90/10 split ratio. The ANNs are trained using the training data set, while the validation data set is used to assess overfitting and if the data-driven model can generalize to unseen data. The training can be intensive in memory use. For this reason, the minimization is performed in mini-batches of data. An epoch is completed when the minimization has gone through the entire training data set.

The architecture of ANN in the transport property models, i.e., the number of layers, the number of neurons per layer and the activation function, the minimization algorithm, the learning parameter, and the number of epochs and size of the mini-batches, are hyperparameters of the ANN. In this chapter, the hyperparameters of the ANNs are optimized. However, the following standard practices are imposed to reduce the complexity of the hyperparameter optimization: the hidden layers use the tanh activation function and are restricted to have the same number of neurons, the learning rate is set to  $10^{-3}$ , the batch size is set 32, and the number of epochs is set to 1500. The selected learning rate and batch size are based on standard practices when working with artificial neural networks [78, 135]. Finally, the number of training epochs was selected to reach a plateau where no further improvement is noticed while optimizing the loss function. Finally, the output layer can use either the linear ( $f(x) = x$ ) or the softplus ( $f(x) = \ln(1 + \exp(x))$ ) activation function. The softplus activation function forces the output to be positive, which is required when directly modeling transport properties. Contrarily, the linear activation function can predict any real value, which is necessary when using a semi-log scale.

The following optimized architectures (number of layers and neurons per layer) are used to model each transport property. The architectures were chosen to accurately describe the target transport property while keeping the number of parameters low. Further details about the hyperparameter optimization are found in Section A.2 of Appendix A.

- Self-diffusivity: 2 hidden layers with 30 neurons each using the tanh activation function.
- Shear viscosity: 2 hidden layers with 30 neurons each using the tanh activation function.
- Thermal conductivity: 3 hidden layers with 30 neurons each using the tanh activation function.

The ANNs are implemented using Flax [161] and trained using Optax, which is part of the DeepMind JAX Ecosystem [162]. The hyperparameter tuning is performed using the Tree-structured Parzen Estimator [159] included in Optuna [160].

#### 4.3.4 Transport properties data-driven models performance and predictions

The results for the different trained transport property models based on ANNs are summarized in Table 4.4. The results from modeling  $\ln D^*$  and  $\rho^* D^*$  exhibit lower deviations than directly modeling  $D^*$ . This result is a direct consequence of how using regularization techniques facilitates the interpolating efforts of the ANN. This result was already anticipated from the different representations of the self-diffusivity shown in Figures 4.4 (b) and (c). For this transport property, the entropy scaling



approach produces higher deviations. As observed in Figure 4.4 (d), there are conflicting regions in which a single value of residual entropy can lead to multiple self-diffusivities. The dilute gas limit is adequately addressed using the  $\rho^* D^*$  formulation. Contrarily, the other representations ( $D^*$ ,  $\ln D^*$ , and  $\tilde{D}^*$ ) diverge at the zero density limit. Finally, the residual model ( $\rho^* D^{*,\text{res}}$ ), based on Eq. (4.14a), is the best-performing model with respect to the reference molecular simulation data. Additionally, this model exactly fulfills the first-order approximation from kinetic theory.

In the case of shear viscosity and thermal conductivity, the reported deviations of the data-driven models are higher than those of the self-diffusivity. These higher deviations are attributed to the inherent noisiness of the data. In general, entropy scaling leads to higher deviations due to the underlying limitations of the framework. Modeling scaled transport properties based on the residual entropy can be justified from isomorph theory [26]. In this theory, it is concluded that only the transport properties of a fluid interacting through an inverse power law (i.e.,  $\mathcal{U} = \epsilon (\sigma/r)^n$ ) are a monovariate function of the residual entropy [221]. For other fluids, such as the Mie fluid, the conditions are only met for dense (liquid) states [217, 220], and a divergent behavior is expected at low densities. There exist workarounds in literature to deal with low-density regions either by using a modified scaling [219], by using a modified input variable based on the residual entropy [222] or by modeling the "residual" transport property [260]. However, it has also been argued that using these workarounds negatively affect the description of the liquid phase transport property [220]. Thus, using entropy scaling over the entire density space is impractical for this application.

A direct correlation using density and temperature as inputs yields a better fit of the shear viscosity and thermal conductivity. It is observed that using a semi-log scale ( $\ln \Lambda^*$  or  $\ln \Lambda^{*,\text{res}}$ ) leads to lower deviations of both the molecular simulation and the dilute gas data. Using a semi-log scale causes the training outputs to be in the same order of magnitude, which prevents a loss function based on the mean squared error from being biased to fit only values with a high magnitude. The residual shear viscosity model ( $\ln \eta^{*,\text{res}}$ ), based on Eq. (4.14b), provides the best fit for the reference data, while exactly fulfilling the dilute gas limit. The performance of this model is followed by the  $\ln \eta^*$  model. A similar trend is observed for the thermal conductivity, but in this case, the  $\ln \kappa^*$  model exhibits the lower deviations. The performance of this model is followed by the residual thermal conductivity model ( $\ln \kappa^{*,\text{res}}$ ), which additionally fulfills the dilute gas limit.

It is important to emphasize that the results in Table 4.4 show that the transport property models perform similarly for seen (train) and unseen (test) data. This results implies the models will generalize well. The lack of overfitting is a consequence of regularizing the data, e.g., modifying the model to a convenient scale as well as imposing constraints that enforce a correct physical behavior, see Eq. (4.18).

A final clarification has to be made about the self-diffusivity models. As mentioned in Section 4.2.1, the self-diffusivity should be corrected for finite-size effects. The correction term proposed by Yeh and Hummer [239], Eq. (4.7), involves using the shear viscosity. The first problem here is that the sizes of the databases for the self-diffusivity and shear viscosity do not precisely match, which means that not all the self-diffusivity values could be corrected. Secondly, as pointed out before, the shear viscosity simulation results exhibit higher deviations. This noise can potentially contaminate the corrected self-diffusivity results. Therefore, the models reported in Table 4.4 were trained using the uncorrected

TABLE 4.4: Transport properties Mean Absolute Percentage Error (MAPE) from transport property models based on artificial neural networks (ANNs).

| Transport Property                | Target                    | Descriptors                                 | MAPE <sup>a</sup>    |             |                              |             |
|-----------------------------------|---------------------------|---------------------------------------------|----------------------|-------------|------------------------------|-------------|
|                                   |                           |                                             | MD data <sup>b</sup> |             | Dilute gas data <sup>c</sup> |             |
|                                   |                           |                                             | Train                | Test        | Train                        | Test        |
| Self-diffusivity <sup>d</sup>     | $D^{*f,h}$                | $\alpha_{vdw}, \rho^*, 1/T^*$               | 9.44                 | 8.25        | –                            | –           |
|                                   | $\ln D^{*g,h}$            | $\alpha_{vdw}, \rho^*, 1/T^*$               | 1.84                 | 1.15        | –                            | –           |
|                                   | $\rho^* D^{*f}$           | $\alpha_{vdw}, \rho^*, 1/T^*$               | 1.82                 | 0.94        | 0.14                         | 0.15        |
|                                   | $\rho^* D^{*,resg,i}$     | $\alpha_{vdw}(\lambda_r, 6), \rho^*, 1/T^*$ | <b>1.67</b>          | <b>0.79</b> | <b>0.00</b>                  | <b>0.00</b> |
|                                   | $\tilde{D}^{*f,h,j}$      | $\alpha_{vdw}, S^{*,res}$                   | 5.27                 | 4.40        | –                            | –           |
| Shear viscosity <sup>d</sup>      | $\eta^{*f}$               | $\alpha_{vdw}, \rho^*, 1/T^*$               | 7.53                 | 7.26        | 5.48                         | 5.73        |
|                                   | $\ln \eta^{*g}$           | $\alpha_{vdw}, \rho^*, 1/T^*$               | 6.07                 | 5.94        | 0.41                         | 0.41        |
|                                   | $\eta^{*,resg,i}$         | $\alpha_{vdw}(\lambda_r, 6), \rho^*, 1/T^*$ | 6.37                 | 6.32        | <b>0.00</b>                  | <b>0.00</b> |
|                                   | $\ln \eta^{*,resg,i}$     | $\alpha_{vdw}(\lambda_r, 6), \rho^*, 1/T^*$ | <b>5.86</b>          | <b>5.72</b> | <b>0.00</b>                  | <b>0.00</b> |
|                                   | $\tilde{\eta}^{*f,h,j}$   | $\alpha_{vdw}, S^{*,res}$                   | 6.14                 | 6.07        | –                            | –           |
| Thermal conductivity <sup>e</sup> | $\kappa^{*f}$             | $\alpha_{vdw}, \rho^*, 1/T^*$               | 6.34                 | 6.68        | 2.50                         | 2.42        |
|                                   | $\ln \kappa^{*g}$         | $\alpha_{vdw}, \rho^*, 1/T^*$               | <b>6.00</b>          | <b>6.20</b> | 0.69                         | 0.69        |
|                                   | $\kappa^{*,resg,i}$       | $\alpha_{vdw}(\lambda_r, 6), \rho^*, 1/T^*$ | 6.24                 | 6.48        | <b>0.00</b>                  | <b>0.00</b> |
|                                   | $\ln \kappa^{*,resg,i}$   | $\alpha_{vdw}(\lambda_r, 6), \rho^*, 1/T^*$ | 6.15                 | 6.34        | <b>0.00</b>                  | <b>0.00</b> |
|                                   | $\tilde{\kappa}^{*f,h,j}$ | $\alpha_{vdw}, S^{*,res}$                   | 6.86                 | 7.16        | –                            | –           |

<sup>a</sup> The reported Mean Absolute Percentage Error (MAPE) is obtained as  $MAPE = \frac{100}{N_p} \sum_{i=1}^{N_p} \left| \frac{y_i^{model}}{y_i^{true}} - 1 \right|$ .

The smallest MAPE deviations for the test and train datasets are highlighted in **bold**. <sup>b</sup> Molecular simulation data obtained in Section 4.2. <sup>c</sup> Dilute gas transport properties obtained using the Eqs. (4.11) and collision integral correlations of Fokin et al. [47], Eqs. (4.12). <sup>d</sup> The ANN consists of 2 hidden layers with 30 neurons, each using the tanh activation function. <sup>e</sup> The ANN consists of 3 hidden layers with 30 neurons, each using the tanh activation function. <sup>f</sup> The output layer uses the linear activation function. <sup>g</sup> The output layer uses the softplus activation function. <sup>h</sup> The transport property model diverges at zero density. <sup>i</sup> Residual transport property model as defined in Eqs. (4.14). <sup>j</sup> Entropy scaling model relies on the residual entropy computed from the FE-ANN EoS developed in Chapter 2.

self-diffusivity. The corrected self-diffusivity can be obtained afterward using the trained data-driven models.

$$D_{\infty}^* = D_{ANN}^* + \frac{T^* \xi}{6\pi \eta_{ANN}^* L^*} \quad (4.19)$$

In Eq. (4.19),  $\xi = 2.837298$ ,  $L^* = \sqrt[3]{N/\rho^*}$ , where  $N$  equals to the number of particles (4096 for this case) and  $D_{ANN}^*$  and  $\eta_{ANN}^*$  are the outputs of the artificial network models for the self-diffusivity and shear viscosity, respectively.

With this in mind, the rest of this section will focus on two models per transport property. Namely,  $\rho^* D^*$  and  $\rho^* D^{*,res}$  for the self-diffusivity,  $\ln \eta^*$  and  $\ln \eta^{*,res}$  for the shear viscosity, and  $\ln \kappa^*$  and  $\ln \kappa^{*,res}$  for the thermal conductivity. This selection of models allows to effectively compare the results for each transport property with and without imposing the dilute gas limit. The  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  do not enforce the dilute gas limit, but they learn it by incorporating  $\mathcal{L}^{dilute}$ , Eq. (4.17), in the loss function. These transport property models and the kinetic theory data are compared in Figure 4.6. This figure confirms that despite the dilute gas limit not being hard imposed in these transport property models, it is still correctly modeled for different repulsive exponents over the temperature range of interest

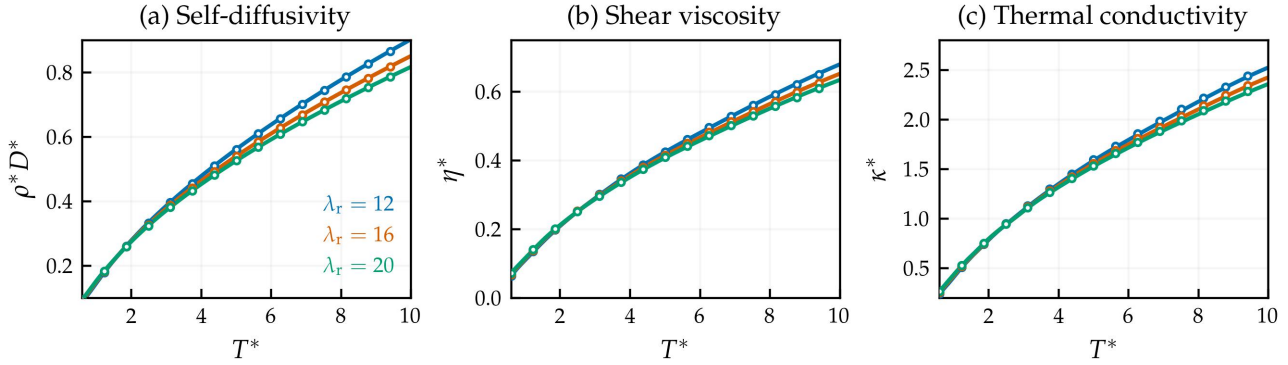


FIGURE 4.6: **Transport properties at zero density of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** (a) Self diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Empty circles are kinetic theory results (Eqs. (4.11)) using the collision integral correlations of Fokin et al. [47], Eqs. (4.12). Solid lines are the transport property models ( $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$ ), Eq. (4.9a)). Blue symbols and lines:  $\lambda_r = 12$ , orange symbols and lines:  $\lambda_r = 16$ , and green symbols and lines:  $\lambda_r = 20$ .

( $T^* \in [0.6, 10]$ ). Moreover, the deviations for the dilute gas data reported in Table 4.4 are all lower than 1% for the selected models. For reference, this corresponds to the estimated error in the first-order transport property approximation from kinetic theory. It can be concluded that both types of models correctly model the dilute gas limit.

The overall performance of these two model types can also be assessed against the available molecular simulation results. The parity plots and error distribution of the  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models are shown in Figure 4.7. A similar plot is available for the residual models,  $\rho^* D^{*,res}$ ,  $\ln \eta^{*,res}$  and  $\ln \kappa^{*,res}$ , in Figure 4.8. Across the board, both types of models exhibit similar behavior for the three transport properties. In subfigure (a), the results for the self-diffusivity show that almost all data points lie on the diagonal for a wide range of values. For the case of the shear viscosity and thermal conductivity, subfigures (b) and (c), it can also be observed that the points lie within the diagonal. As discussed previously, these data points are inherently noisier, and a more noticeable dispersion was expected. Finally, it can be seen that in all the ANN models, the unseen test data is modeled similarly to the training data, disregarding the possible overfitting of the models.

The trends in the relative deviations for each transport property can be identified in the density-temperature space. The distribution of relative errors is shown in subfigures (d), (e), and (f) of Figures 4.7 and 4.8. For clarity, data points with higher deviations are highlighted. From subfigures (d), it can be seen that most of the self-diffusivity data points are correctly described except for a few cases in the surroundings of the phase transitions (VLE or SLE). For the case of the shear viscosity and thermal conductivity, subfigures (b) and (c), it can be noticed that higher deviation points are found in the low-density region and close to the region where the vapor-liquid phase envelope should lie. Firstly, the shear viscosity and thermal conductivity increase with density. Therefore, higher relative deviations in low-density regions are also an artifact of calculating the deviation of a relatively small number. Secondly, higher relative deviations were expected for these transport properties due to the noisiness of the molecular simulation data.

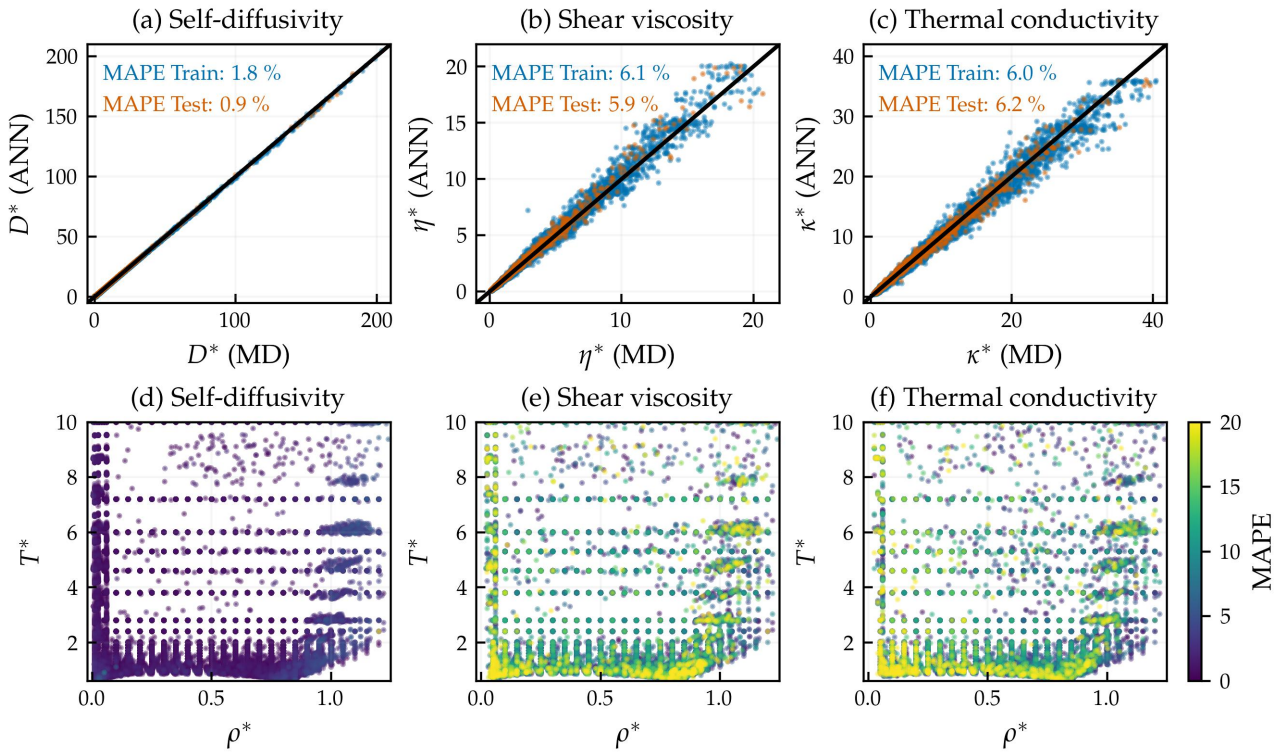


FIGURE 4.7: Parity plots and error distribution of transport property models based on ANNs ( $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$ ). (a) and (d): Self-diffusivity. (b) and (e): Shear viscosity. (c) and (f): Thermal conductivity. Parity plots: (a), (b), and (c). Blue circles refer to train data and orange circles refer to test data. Error distribution: (d), (e), and (f). The color of the data points refers to Mean Absolute Percentage Error (MAPE), which is obtained as

$$\text{MAPE} = \frac{100}{N_P} \sum_{i=1}^{N_P} \left| \frac{y_i^{\text{model}}}{y_i^{\text{true}}} - 1 \right|.$$

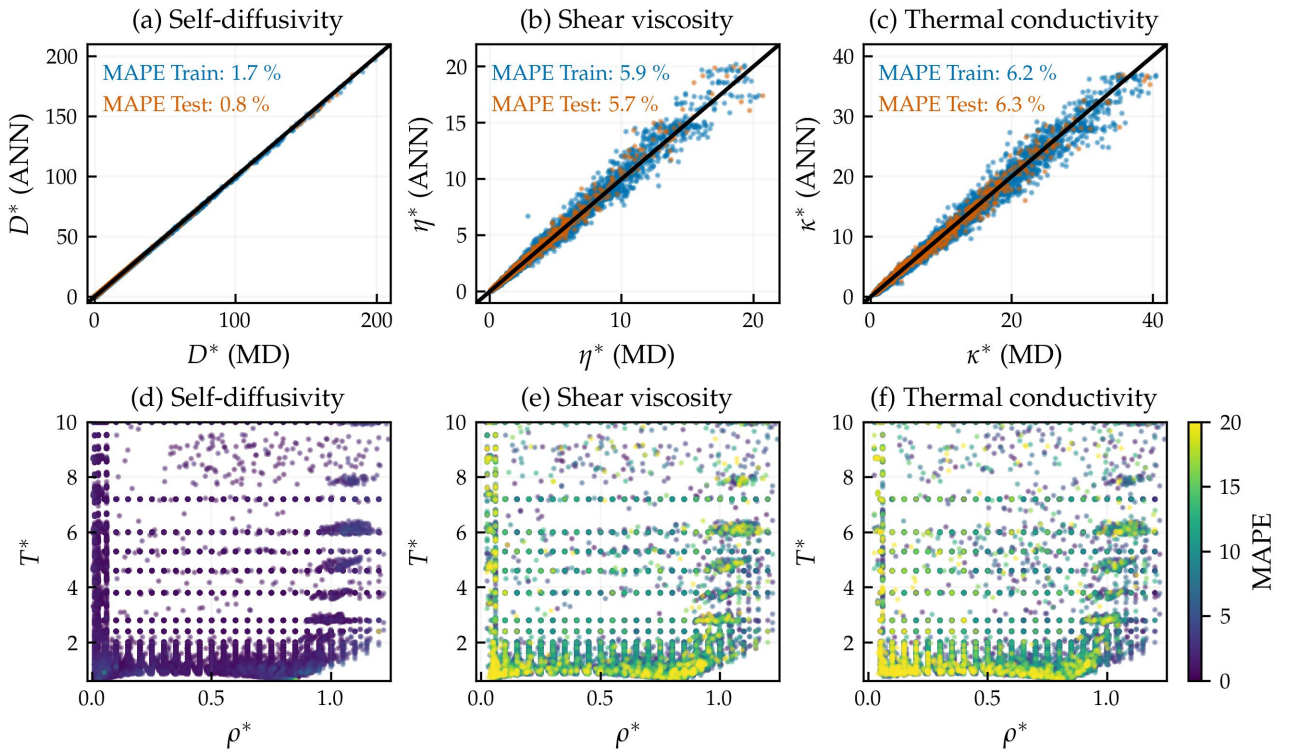


FIGURE 4.8: Parity plots and error distribution of transport property models based on ANNs ( $\rho^* D^{*,\text{res}}$ ,  $\ln \eta^{*,\text{res}}$  and  $\ln \kappa^{*,\text{res}}$ ). (a) and (d): Self-diffusivity. (b) and (e): Shear viscosity. (c) and (f): Thermal conductivity. Parity plots: (a), (b), and (c). Blue circles refer to train data and orange circles refer to test data. Error distribution: (d), (e), and (f). The color of the data points refers to Mean Absolute Percentage Error (MAPE), which is obtained as

$$\text{MAPE} = \frac{100}{N_P} \sum_{i=1}^{N_P} \left| \frac{y_i^{\text{model}}}{y_i^{\text{true}}} - 1 \right|.$$

Even though it is well known that machine learning models can achieve high accuracy in describing thermophysical properties [3, 58, 62, 224], it is also important to consider the consistency and physicality of the predictions [261]. In Figure 4.9, five predicted isotherms are shown for each transport property using the trained ANN models. This figure considers three selected Mie fluid ( $\lambda_r = 12, 16$ , and  $20$  and  $\lambda_a = 6$ ). For reference, in the case of the Lennard-Jones fluid ( $\lambda_r = 12$  and  $\lambda_a = 6$ ), the selected temperatures include two subcritical isotherms ( $T^* = 0.9$  and  $1.0$ ), one near-critical isotherm ( $T^* = 1.3$ ), and two supercritical isotherms ( $T^* = 2.8$  and  $6.0$ ). The modeled transport properties for other Mie fluids are shown in Section B.3 of Appendix B. Figure 4.9 includes the molecular simulation data as well as the results from the  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models (solid lines) and the residual  $\rho^* D^{*,res}$ ,  $\ln \eta^{*,res}$  and  $\ln \kappa^{*,res}$  models (dashed lines). As expected from the magnitude of the deviations, the models for self-diffusivity, shown in Figure 4.9 (a), follow a smooth trend across the entire density and temperature space. This behavior is even observed in regions with no data available, such as the region that lies inside the expected vapor-liquid phase envelope. From this figure, it can be noticed that both modeling approaches perform similarly in regions where data is available, with isotherms that practically overlap. The main discrepancies are observed at high densities and low temperatures, where a solid phase is expected. Finally, as expected, both models correctly predict that the self-diffusivity decreases with density.

The shear viscosity predicted isotherms are shown in Figure 4.9 (b). It observed that both modeling approaches behave similarly and are in good agreement with the molecular simulation data. Despite the more noticeable error bars of the reference data, the models predict an increasing shear viscosity with density. The isotherms are smooth and well-behaved. As expected, no artifacts are noticed near the critical point ( $\rho_c^* \approx 0.33$ ) [221]. Finally, in Figure 4.9 (c), the computed five isotherms for the thermal conductivity of the selected Mie fluids are shown. The predicted isotherms are smooth and non-oscillatory. Moreover, they are in good agreement with the error boundaries of the reference molecular simulation data. Both models,  $\ln \kappa^*$  and  $\ln \kappa^{*,res}$ , are in good agreement with each other except in regions with no available data. The first of them is at high densities and high temperatures, which is where a solid phase is expected. The second region is at near-critical conditions. Both models predict a critical enhancement for the subcritical and near-critical isotherms. However, the  $\ln \kappa^*$  model predicts a more noticeable maximum. No explicit guidance about the critical enhancement has been given to the models to predict this maximum. This behavior is a prediction of the models, which is directly learned from the available simulation data. Finally, both models correctly predict a monotonically increasing thermal conductivity for supercritical isotherms.

Up to this point, it has been proven that both modeling approaches,  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models based on Eq. (4.9a) and  $\rho^* D^{*,res}$ ,  $\ln \eta^{*,res}$  and  $\ln \kappa^{*,res}$  models based on Eq. (4.14), can correctly predict the self-diffusivity, shear viscosity and thermal conductivity of Mie fluids. These models exhibit similar behavior for the entire fluid density space. One key difference between these two modeling approaches is that the residual models rely on the collision integrals correlations developed by Fokin et al. [47]. These correlations are restricted to ( $\lambda_r, \lambda_a = 6$ ) Mie fluids. Conversely, under the assumption of the conformality of the Mie potential based on the  $\alpha_{vdw}$  parameter, the  $\rho^* D^*$ ,  $\ln \eta^*$ , and  $\ln \kappa^*$  models could, in principle, be applied to other ( $\lambda_r, \lambda_a$ ) Mie fluids. For this reason, the predictions of these models are compared against published data. This data includes the simulation data collected by



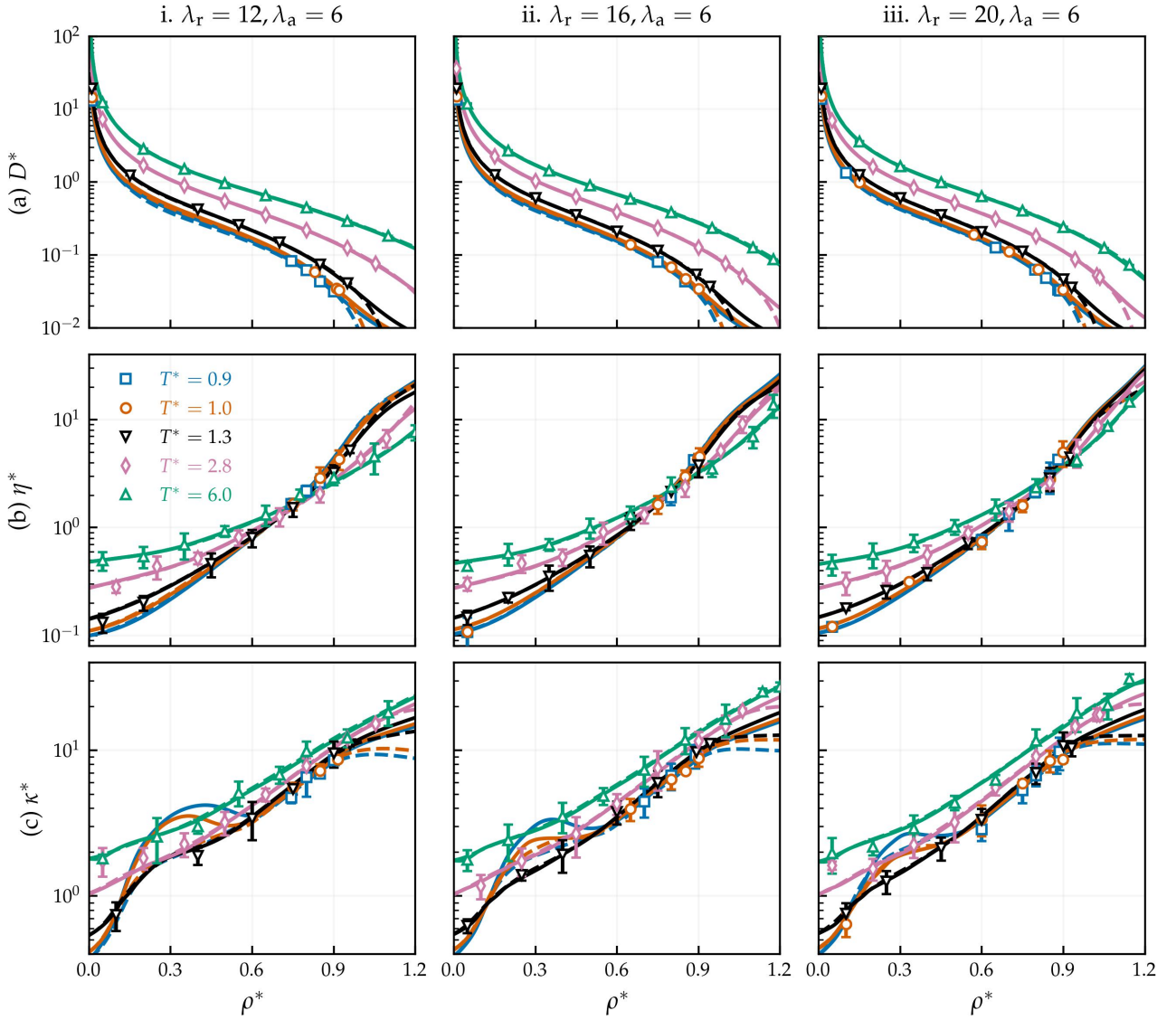


FIGURE 4.9: **Transport properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** i.  $\lambda_r = 12, \lambda_a = 6$  (Lennard-Jones fluid). ii.  $\lambda_r = 16, \lambda_a = 6$ . iii.  $\lambda_r = 20, \lambda_a = 6$ . (a) Self-diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Molecular simulation data. Blue squares:  $T^* = 0.9$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$ . Solid lines refer to the  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models, Eq. (4.9a). Dashed lines refer to the residual  $\rho^* D^{*,\text{res}}$ ,  $\ln \eta^{*,\text{res}}$  and  $\ln \kappa^{*,\text{res}}$  models, Eq. (4.14). The computed isotherms from the ANN-based models follow the same color convention as the simulation data.

Lautenschlaeger and Hasse [210] for the Lennard-Jones fluid and the data for self-diffusivity and shear viscosity for selected Mie fluids computed by Šlepavičius et al. [224]. It has to be considered that even though the same underlying interaction potential is being used, the molecular dynamics simulations differ in the cut-off radius and the number of particles. Hence, the results from the transport models are only partially comparable to the published data. Nevertheless, they still allow assessing the behavior of the trained models in untrained conditions. The parity plots and phase space distribution for the published data are shown in Figure 4.10. The symbols are colored based on the  $\alpha_{\text{vdw}}$  parameter, Eq. (2.21). For reference, this parameter is 0.88 for the Lennard-Jones fluid (green symbols in the figure). From the top subfigure of Figure 4.10 (a), it can be observed that most self-diffusivity data points are correctly described. The global relative deviation for this property equals 13.50%. The main source of discrepancy is attributed to data points at very high densities ( $\rho^* > 1.2$ ) [245], at metastable conditions [252], and at considerable small  $\alpha_{\text{vdw}}$  values [224], which are far from the training range ( $0.53 < \alpha_{\text{vdw}} < 1.47$ ).

Similarly, the predicted thermal conductivities are compared against the published Lennard-Jones data in Figure 4.10 (c). As can be observed from the figure, most data points are correctly described. The global relative deviation for this property is 4.1%. Under the inspection of the relative residuals, the more significant deviations are found for data points inside the unstable region for the Lennard-Jones fluid. These higher deviations directly result from comparing simulation results with different cut-off radius (e.g., Lautenschlaeger and Hasse [210] use shifted potential with  $r_c = 2.5\sigma$ ). The cut-off radius modifies the unstable region of the Lennard-Jones fluid [262, 263], allowing those simulations to access unstable regions (inside the VLE phase boundaries) under the cut-off radius used in this work ( $r_c = 5\sigma$ ). For reference, the VLE and SLE boundaries of the Lennard-Jones particle computed with the FE-ANN(s) EoS developed in Chapter 3 are shown in the bottom subfigure of Figure 4.10 (c). The errors in the unstable region can be considerable, as the  $\ln \kappa^*$  model predicts a maximum for subcritical temperatures near the critical density. This maximum was previously shown in Figure 4.9 (c).

Finally, the parity plot for the shear viscosity is shown in Figure 4.10 (b). There are evident discrepancies with the published Lennard-Jones data for this transport property. It is found that the main discrepancies are for data points that lie inside the vapor-liquid phase envelope. This behavior was expected from comparing simulations with different cut-off radius. More importantly, it becomes evident that the  $\ln \eta^*$  model fails to extrapolate to very low  $\alpha_{\text{vdw}}$  values [224], which are indicated by dark-blue colors. These values of the  $\alpha_{\text{vdw}}$  correspond to highly attractive exponents ( $\lambda_a$ ) in the order of 10 to 14. This result suggests that the transport property models do not extrapolate well. Moreover, this result suggests that the conformality of the Mie potential does not hold for transport properties. This conclusion is in agreement with the results of Stephan and Urschel [166], whose find that some extreme combinations of the repulsive-attractive exponents of the Mie potential may result in the breaking down of the corresponding states principle. The  $\alpha_{\text{vdw}}$  is still a convenient molecular descriptor of the Mie potential. However, the transport property models developed here should be suitable only for  $(\lambda_r, \lambda_a = 6)$  Mie fluids.

In this section, different data-driven modeling approaches for the self-diffusivity, shear viscosity, and thermal conductivity of the Mie fluid were tested. The developed models were formulated to address



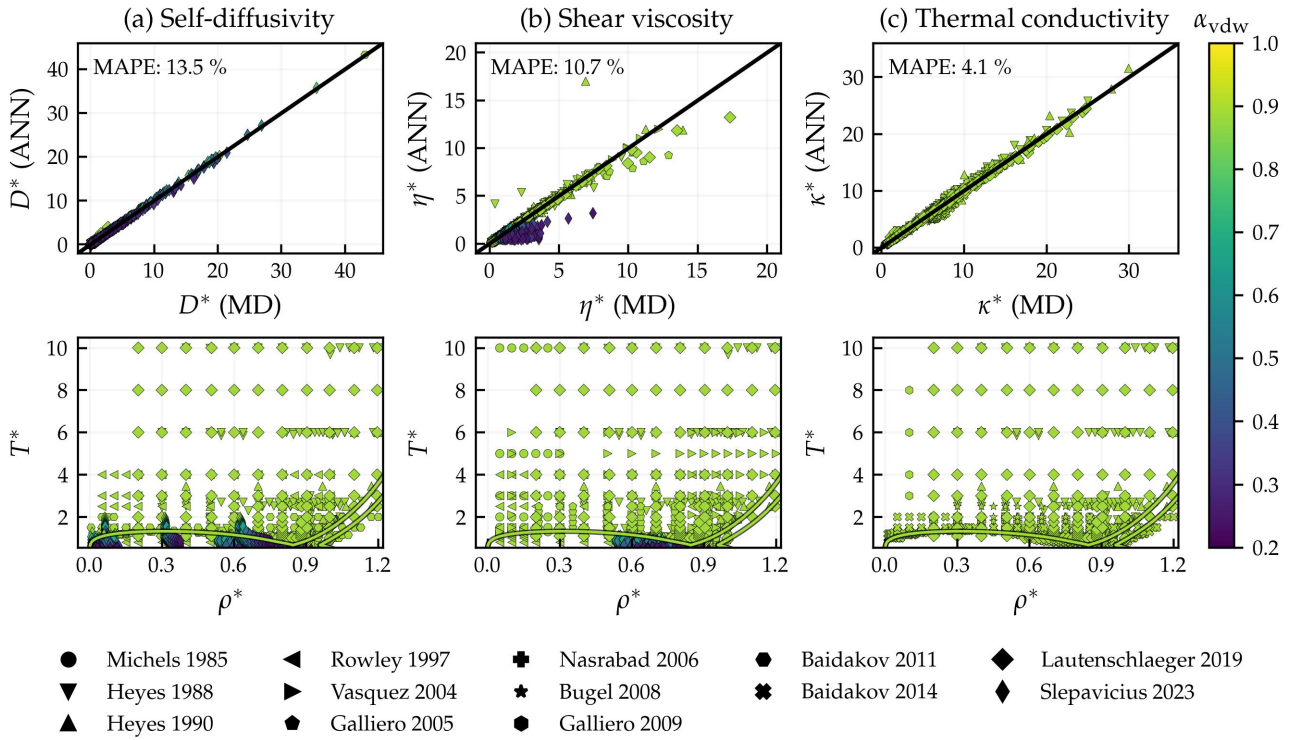


FIGURE 4.10: **Predicted transport property models against published molecular dynamics data.** (a) Self-diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Transport property models  $\rho^*D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models are based on Eq. (4.9a). Top figures: parity plot. Bottom figures: density-temperature distribution. Published molecular simulation data: circles [243], downward triangle [244], upright triangle [245], left-pointing triangle [246], right-pointing triangle [247], pentagon [248], "plus" sign [249], star [250], hexagon [251], side hexagon [252], "X" sign [253], diamond [210], thin diamond [224]. The colormap refers to the  $\alpha_{vdw}$  parameter of the Mie fluid (Eq. (2.21)). The green symbols refer to the Lennard-Jones fluid ( $\alpha_{vdw} = 0.88$ ). For reference, the vapor-liquid equilibria (VLE) and solid-liquid equilibria (SLE) of the Lennard-Jones particle are shown in the bottom figures. The VLE and SLE boundaries are obtained from the FE-ANN(s) EoS developed in Chapter 3.

the dilute gas limit, and they successfully model the molecular simulation data over a wide range of densities and temperatures. Additionally, the developed transport models follow expected physical trends. The transport property models were built on the premise of the conformality of the Mie potential. However, it is noticed that this conformality, based on the  $\alpha_{\text{vdw}}$  parameter, does not hold for transport properties. Therefore, using these transport property models exclusively for  $(\lambda_r, \lambda_a = 6)$  Mie fluids is recommended.

## 4.4 Conclusions

In this chapter, the self-diffusivity ( $D^*$ ), shear viscosity ( $\eta^*$ ), and thermal conductivity ( $\kappa^*$ ) of  $(\lambda_r, \lambda_a = 6)$  Mie fluids have been computed and modeled. The transport properties have been computed using equilibrium molecular dynamics and the Einstein relationship over a vast region of the phase space. The correctness of the computed transport properties database has been verified by comparison with published simulation results. The produced database will be useful for testing and verifying transport properties theories, such as modified scaling approaches [264] and those based on kinetic theory [48].

The transport properties of the Mie fluid have been modeled using physics-informed artificial neural networks. Two main modeling approaches were tested: a semi-empirical formulation based on entropy scaling or an empirical formulation based on density and temperature as input variables. For the former, it was found that a single formulation based on entropy scaling does not lead to satisfactory results over the entire density range, with issues mainly associated with divergence and incorrect scaling of the transport properties at low densities. Linearizing the data, e.g., modeling  $\rho^* D^*$  instead of  $D^*$ ,  $\ln \eta^*$  instead of  $\eta^*$ , and  $\ln \kappa^*$  instead of  $\kappa^*$ , has a regularizing effect and facilitates the interpolation efforts of the ANNs. These modeling approaches can also account for the dilute gas limit predicted by kinetic theory. This limit can be addressed in two ways: by including the zero-density data in the ANN training or by directly modeling the "residual" transport property. Regardless of which modeling approach is used, the trained ANNs can model both seen and unseen data over a wide range of densities and temperatures. Moreover, they predict the expected physical behavior. For example, the self-diffusivity model suggests that the diffusivity is a smooth decreasing function of density and an increasing function of temperature. The shear viscosity model is a well-behaved, increasing function of density, and the model for thermal conductivity exhibits the well-known critical enhancement.

The  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  and the residual  $\rho^* D^{*,\text{res}}$ ,  $\ln \eta^{*,\text{res}}$  and  $\ln \kappa^{*,\text{res}}$  models perform similarly in the fluid training region, i.e.,  $\alpha_{\text{vdw}} \in [0.53, 1.47]$ ,  $\rho^* \in [5 \cdot 10^{-3}, 1.26]$  and  $T^* \in [0.6, 10.0]$ . It must be noted that the residual models are valid only for  $(\lambda_r, \lambda_a = 6)$  Mie fluids. Contrarily, in principle, the  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models were constructed based on the conformality of the Mie potential through the  $\alpha_{\text{vdw}}$  parameter. This conformality could potentially allow the use of attractive exponents different from 6. However, the models may fail for unseen conditions, like at highly attractive Mie fluids that produce low  $\alpha_{\text{vdw}}$  values. It is noted that the conformality of the potential does not hold for transport properties. For this reason, it is recommended that the developed models be used exclusively for  $(\lambda_r, \lambda_a = 6)$  Mie fluids.

Finally, the developed data-driven transport property models allow assessment of the performance of previously published Mie fluid parameters for transport properties. Moreover, these transport property models can be used alongside an equation of state for the Mie fluid (such as the FE-ANN EoS developed in Chapter 2 [3] or the SAFT-VR-Mie EoS [37]) to parametrize the Mie potential using both equilibrium and transport properties.

## Chapter 5

# Parametrization of quasi-spherical molecules through the Mie potential

### 5.1 Introduction

Accurately representing the interactions at the molecular level is key to adequately describing and predicting thermophysical properties through molecular simulations. Molecular interactions are modeled using mathematical functions known as force fields. Historically, the development of force fields has relied on postulating semiempirical functions, where non-bonded (e.g., the Mie potential), bonded, angular, torsional, and electrostatic interactions are explicitly modeled [255, 265–267]. These arguably "simple" force fields provide a computationally efficient but incomplete representation of molecular interactions. While successful in many applications, their limitations must be acknowledged [15], especially when studying systems where polarization and electrostatic interactions are significant. Additionally, these classical force fields do not account for the local environment of the particles. Force fields based on machine-learning (data-driven) techniques [67] have been developed to address these challenges. These include force fields based on artificial neural networks [268, 269], Gaussian processes [270], and many-body expansions [271, 272]. Force fields can be used to represent molecules at different resolutions: All-Atom (AA), United-Atom (UA), and Coarse Grained (CG). In AA, all atoms in a molecule are explicitly accounted for. In UA, molecular groups are condensed into a single particle. For instance, a  $-\text{CH}_2-$  group can be treated as a single particle. Lastly, in CG, the molecular representation is simplified using larger fictitious particles that can group multiple functional groups. The CG approach simplifies the molecular representation, allowing the exploration of spatio-temporal scales closer to the macroscopic level. CG force fields have been parametrized using experimental thermophysical properties [123] and structural [273, 274] data. The coarse grained approach has been successfully used to model a wide variety of thermophysical properties, including phase equilibria [97, 98], interfacial properties [98, 109, 275], transport properties [111], and mechanical properties [104–106] of diverse molecular systems.

Regardless of the type of force field (i.e., classical or machine-learning-based) and its molecular resolution (AA, UA, or CG), the force field depends on molecular parameters which, along with the state conditions (i.e., density, temperature, and composition), define a Hamiltonian which can be translated into macroscopic observables through molecular simulations and statistical thermodynamics. The

appropriate parameterization of the force field (i.e., the fitting of the molecular parameters) is crucial for its success in predicting thermophysical properties. The molecular parameters can follow either a bottom-up [276, 277] or a top-down [123] approach. In bottom-up approaches, the force field is fitted to reproduce forces, bond distances, and angles from rigorous quantum mechanics calculations. This approach is the base for machine learning potentials, which aim to bridge *ab initio* accuracy to larger scales [68, 69]. Conversely, the top-down approach optimizes the molecular parameters directly to reproduce relevant experimental data. This method directly links a molecular model to its implied macroscopic thermophysical properties and is the most efficient from an engineering and design perspective. Conveniently, this approach can take advantage of the physics-informed artificial neural network models developed in Chapters 2, 3 and 4, which will be used henceforth.

The top-down approach has been used to parametrize (or fine-tune) force fields of different molecular resolutions [123, 265, 266]. Lennard-Jones himself pioneered the parametrization of the Mie potential, Eq. (1.1), using a top-down approach [278, 279]. At the time, it was recognized that the molecular parameters of the Mie potential could be fitted to either the second virial coefficient or the viscosity of the dilute gas. This work led to the well-known Lennard-Jones potential (i.e.,  $U^{\text{LJ}} = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ ). The results from Lennard-Jones already showed that even for simple molecules like Argon, there are trade-offs with respect to which property to fit [27]. In a modern approach, the top-down strategy can target multiple thermophysical properties. For example, the SAFT- $\gamma$ -Mie force field [123] is a coarse grained force field based on the Mie potential. The parameterization of this force field relies on the success of the SAFT-VR-Mie EoS [37] and its group contribution version (SAFT- $\gamma$ -Mie EoS) [65] to effectively map the volumetric thermophysical properties of homo- and heteronuclear Mie chains with respect to molecular simulation results. The SAFT- $\gamma$ -Mie force field has been successfully used to model a wide range of thermophysical properties of alkanes [280], carbon dioxide [281], water [275], and polymers [104–106].

The top-down approach has also been used to parameterize force fields based on the corresponding states principle. Mejía et al. [126] developed a corresponding states correlation that relies on the critical temperature ( $T_c$ ), saturated liquid density ( $\rho_l$  at  $T_r = 0.7$ ) and the Pitzer acentric factor ( $\omega = -1 - \log_{10}(P^{\text{sat}}/P_c)$  at  $T_r = 0.7$ ) to obtain the molecular parameters of homonuclear Mie chains. This approach, also known as M&M correlation, has been successfully applied to model the phase equilibria and interfacial properties of a diverse range of industrially relevant molecules, including alkanes, aromatics, gases, and solvents [94]. Moreover, this corresponding states principle has been applied to 6000+ molecular fluids, whose parameters have been freely published in the Bottled SAFT webpage [97]. The M&M correlation has also been extended to handle homonuclear rings [282]. The extension to ring structures is also implemented within the SAFT-VR-Mie EoS framework by introducing a simple modification to the chain term of the EoS. Alternatively to the M&M correlation, Hoang et al. [111] proposed a similar corresponding states correlation for homonuclear Mie chains. Interestingly, they proposed using the reduced liquid viscosity ( $\eta_l$  at  $T_r = 0.7$ ) as an additional parameter into the correlation. This augmented correlation suggests that transport properties could be employed in addition to the volumetric equilibrium properties to obtain more robust and transferable models.

In addition to using equations of state or the corresponding states principle, force fields can also be

parametrized directly using molecular simulation results. For example, the OPLS-AA [265] and the TraPPE [266] force fields are primarily based on quantum mechanical calculations to define angular and torsion parameters, and they rely on experimental data to improve the accuracy of non-bonded interactions. Similarly, successful water models like the TIP4P/2005 [283] and the TIP4P/Ice [284] are primarily fitted to experimental data. Directly fitting molecular parameters of a force field is a tedious and error-prone process that can quickly get stuck in suboptimal molecular parameter solutions. This issue has been addressed using Bayesian optimization to guide the force field parametrization [285]. This data-driven approach considerably reduces the number of simulations required to parametrize the force fields and enables the targeting of multiple thermophysical properties.

That being said, this chapter focuses on the single-bead CG modeling approach for quasi-spherical molecules using the Mie potential, Eq. (1.1). Three main parametrization schemes are explored and compared. The first scheme follows the classical strategy for parametrizing force fields where only vapor-liquid equilibria and fluid data are considered. In this case, the fluid properties are described using the FE-ANN EoS developed in Chapter 2. The second scheme follows the corresponding states principle of Hoang et al. [111], where the liquid shear viscosity data is considered in addition to the VLE data. The shear viscosity information is obtained from the ANN-based model developed in Chapter 4. Finally, in most cases, the influence of solid-fluid equilibria is not considered when determining the molecular parameters of a force field, as solid-fluid equilibria is challenging to simulate. The third parametrization scheme includes solid-fluid data along the usual fluid phase equilibria using the FE-ANN(s) EoS developed in Chapter 3. The ultimate aim of this chapter is to assess the impact of thermophysical properties in parametrizing force fields.

The rest of the chapter is structured as follows. Section 5.2 details the different parametrization approaches and how the thermophysical properties are modeled. Then, Section 5.3 discusses the results for parametrizing quasi-spherical molecules using different target properties. The main conclusions of this chapter are summarized in Section 5.4.

## 5.2 Multi-property parametrization approaches

In this chapter, "simple" quasi-spherical molecules are described using the Mie potential, Eq. (1.1). The Mie potential is defined by a shape parameter ( $\sigma$ ), the energy well depth ( $\epsilon$ ), and the repulsive and attractive exponents ( $\lambda_r$  and  $\lambda_a$ ) to describe the thermophysical properties of molecules. The attractive exponent is set to 6 to account for London dispersion forces [93]. Therefore, a quasi-spherical molecule is described by three molecular parameters ( $\sigma$ ,  $\epsilon$ , and  $\lambda_r$ ) when using the Mie potential. These molecular parameters must be optimized to fit experimental data.

Three main parametrization approaches are considered to obtain the molecular parameters ( $\sigma, \epsilon, \lambda_r$ ) of the Mie potential. The general form of the objective function is defined as follows.

$$\text{OF}(\sigma, \epsilon, \lambda_r) = \sum_{k \in \mathcal{K}} \frac{w_k}{N_{p,k}} \sum_{i=1}^{N_{p,k}} \left| \frac{y_{k,i}^{\text{pred}}}{y_{k,i}^{\text{exp}}} - 1 \right| \quad (5.1)$$

TABLE 5.1: Parametrization approaches to obtain the Mie potential parameters.<sup>a</sup>

|                 | $P_{\text{VLE}}^b$ | $\rho_{\text{l,VLE}}^b$ | $\Delta H_{\text{VLE}}^b$ | $P_{\text{SLE}}^c$ | $P_{\text{SVE}}^d$ | $\Delta H_{\text{SVE}}^d$ | $\eta^e$ |
|-----------------|--------------------|-------------------------|---------------------------|--------------------|--------------------|---------------------------|----------|
| OF <sub>1</sub> | ✓                  | ✓                       | ✓                         | ×                  | ×                  | ×                         | ×        |
| OF <sub>2</sub> | ✓                  | ✓                       | ✓                         | ×                  | ×                  | ×                         | ✓        |
| OF <sub>3</sub> | ✓                  | ✓                       | ✓                         | ✓                  | ✓                  | ✓                         | ×        |

<sup>a</sup> Thermophysical properties considered in the objective function (OF) are indicated by a "✓" in a grey-colored cell. <sup>b</sup> Vapor-liquid Equilibria (VLE) properties are also referred to as "vaporization" properties. These include the vaporization pressure ( $P_{\text{VLE}}$ ), saturated liquid density ( $\rho_{\text{l,VLE}}$ ), and vaporization enthalpy ( $\Delta H_{\text{VLE}}$ ). VLE properties can be obtained either with the FE-ANN EoS (Chapter 2) or FE-ANN(s) EoS (Chapter 3), based on Eq. (5.2). <sup>c</sup> Solid-liquid equilibria (SLE) properties are also referred to as "melting" properties. Only the melting pressure ( $P_{\text{SLE}}$ ) is considered here. <sup>d</sup> Solid-vapor equilibria (SVE) properties are also referred to as "sublimation" properties. These include the sublimation pressure ( $P_{\text{SVE}}$ ) and sublimation enthalpy ( $\Delta H_{\text{SVE}}$ ). SLE and SVE properties are obtained using the FE-ANN(s) EoS. <sup>e</sup> The shear viscosity ( $\eta$ ) is modeled using the ANN-based models developed in Chapter 4, based on Eq. (5.4).

This objective function, Eq. (5.1), represents a weighted average of the relative deviations of various thermophysical properties " $k$ ". The thermophysical properties in the optimization ( $\mathcal{K}$ ) may include vapor-liquid equilibria data, shear viscosity data, and solid-fluid (solid-liquid and solid-vapor) equilibria data. In Eq. (5.1),  $w_k$  is the relative weight,  $N_{p,k}$  is the number of data points, and  $y_k^{\text{pred}}$  and  $y_k^{\text{exp}}$  are the predicted and experimental values of the property " $k$ ". A baseline case for parametrizing the Mie potential (OF<sub>1</sub>) is to consider only Vapor-Liquid Equilibria (VLE)-related properties, such as saturation pressure ( $P_{\text{VLE}}$ ), saturated liquid density ( $\rho_{\text{l,VLE}}$ ), and vaporization enthalpy ( $\Delta H_{\text{VLE}}$ ). This first objective function resembles the parametrization of SAFT EoSs [36, 37, 65]. The second objective function (OF<sub>2</sub>) considers the liquid shear viscosity ( $\eta$ ) in addition to the VLE data. Finally, the third objective function (OF<sub>3</sub>) considers equilibrium properties, including vapor-liquid and solid-fluid equilibria data. This latter objective function allows an assessment of the impact of the solid phase on the molecular parameters. The solid phase data includes the melting pressure ( $P_{\text{SLE}}$ ), sublimation pressure ( $P_{\text{SVE}}$ ) and sublimation enthalpy ( $\Delta H_{\text{SVE}}$ ). The details of the different objective functions are summarized in Table 5.1. The relative weight ( $w_k$ ) for all the thermophysical properties in the objective function is set to 1, except for the melting pressure, whose relative weight is set to  $w_k = 0.01$ . This relatively small weight is chosen to avoid biasing the objective function, as melting pressures are several orders of magnitude larger than all the other properties considered.

### 5.2.1 Experimental data acquisition

The objective functions based on Eq. (5.1) require experimental data. Pseudo-experimental data are obtained from the NIST TRC database [12]. The use of the data sets provided by NIST offers advantages over using discrete experimental data sets. The data reported by NIST have already been curated and fitted to accurate empirical functions. The pseudo-experimental data are smooth and free of outliers and "continuous" in nature as they allow the evaluation of properties over a wide range of state points. Furthermore, the data are internally and thermodynamically consistent.

Thermophysical properties under VLE conditions are generated within the range  $T_r = T/T_c \in [0.55, 0.95]$ . The VLE data includes the saturation pressure, saturated liquid density, vaporization

enthalpy, and saturated liquid viscosity. In addition, the thermal conductivity and speed of sound of the saturated liquid phase are generated to assess the transferability of the molecular parameters to unseen thermophysical properties. Similarly, the melting pressure data are generated in the  $T_r = T/T_c \in [0.55, 1.1]$  range, and the sublimation pressure and enthalpies are generated in the  $T_r = T/T_c \in [0.45, 0.55]$  range.

### 5.2.2 Thermophysical properties modeling

To assess the objective functions, it is necessary to evaluate the thermophysical properties corresponding to the Mie potential at a given state point as a function of the molecular parameters ( $\sigma$ ,  $\epsilon$ , and  $\lambda_r$ ). The thermophysical properties of the Mie particle have been modeled throughout this thesis using physics-informed Artificial Neural Networks (ANNs). The FE-ANN EoS, Eq. (2.22b), developed in Chapter 2 explicitly models the Helmholtz free energy of the Mie fluid and can predict VLE. Similarly, the FE-ANN(s) EoS developed in Chapter 3 extends the framework to additionally handle solid-fluid equilibria, such as SLE and SVE. The Helmholtz free energy is a convenient state function for computing thermophysical properties and phase equilibria. The reader is referred to Section 2.2 for further details about Helmholtz free energy thermodynamics. The FE-ANN/FE-ANN(s) EoSs are formulated using reduced units, see Table 2.1. The framework can be used for real molecules by converting the results into real units.

$$A^{\text{res}} = N_{\text{Av}} \epsilon A^{*,\text{res}} = N_{\text{Av}} \epsilon \left[ \text{ANN}(\alpha_{\text{vdw}}, \rho N_{\text{Av}} \sigma^3, \frac{\epsilon}{T k_{\text{B}}}) - \text{ANN}(\alpha_{\text{vdw}}, \rho = 0, \frac{\epsilon}{T k_{\text{B}}}) \right] \quad (5.2)$$

In Eq. (5.2),  $A^{\text{res}}$  is the residual Helmholtz free energy (in  $\text{J mol}^{-1}$ ),  $\rho$  is the molar density (in  $\text{mol m}^{-3}$ ),  $T$  is the temperature (in K). The shape parameter ( $\sigma$ , in m) and energy scale ( $\epsilon$ , in J) are the molecular parameters. Finally,  $N_{\text{Av}}$  and  $k_{\text{B}}$  are the Avogadro and Boltzmann constants [141]. The Mie potential is also described through the  $\alpha_{\text{vdw}}$  parameter. This parameter is defined below for  $\lambda_r, \lambda_a = 6$  Mie particles.

$$\alpha_{\text{vdw}}(\lambda_r, \lambda_a = 6) = C^{\text{Mie}} \left[ \frac{1}{3} - \frac{1}{\lambda_r - 3} \right] \quad (5.3)$$

The second objective function (OF<sub>2</sub>) requires predicting the shear viscosity of the Mie fluid. ANN-based models for the transport properties of the Mie fluid were developed in Chapter 4. The recommended model for the shear viscosity is based on Eq. (4.9a) and uses a semi-log scale. This model is formulated using reduced units and can be converted into real units as shown below in Eq. (5.4). See Table 4.1 for further details about transport properties reduced units.

$$\eta = \frac{\sqrt{\epsilon M / N_{\text{Av}}}}{\sigma^2} \exp[\ln \eta^*] = \frac{\sqrt{\epsilon M / N_{\text{Av}}}}{\sigma^2} \exp \left[ \text{ANN}(\alpha_{\text{vdw}}, \rho N_{\text{Av}} \sigma^3, \frac{\epsilon}{T k_{\text{B}}}) \right] \quad (5.4)$$

Here,  $\eta$  is the shear viscosity (in Pa s) and  $M$  is the molar mass (in  $\text{kg mol}^{-1}$ ).

The behavior of the speed of sound and thermal conductivity will also be studied to assess the transferability of the optimized molecular parameters ( $\sigma, \epsilon, \lambda_r$ ) to other thermophysical properties.



The speed of sound can be obtained by any EoS as follows.

$$w = \sqrt{\left[ \rho \kappa_T \left( \frac{C_V}{C_P} \right) M \right]^{-1}} \quad (5.5)$$

Here,  $w$  is the speed of sound (in  $\text{m s}^{-1}$ ),  $\kappa_T$  is the isothermal compressibility (in  $\text{Pa}^{-1}$ ) and  $C_V$  and  $C_P$  are obtained isochoric and the isobaric heat capacities (in  $\text{J K}^{-1} \text{mol}^{-1}$ ). The FE-ANN/FE-ANN(s) EoSs explicitly model the residual Helmholtz free energy and the ideal contribution can be accounted for analytically using Eq. (2.8). In this case, the isothermal compressibility can be obtained directly using Eq. (2.3c). It is recommended that the ideal and residual contributions be considered separately for the heat capacities.

$$C_V = C_V^{\text{id}} + C_V^{\text{res}} \quad (5.6a)$$

$$C_P = C_P^{\text{id}} + C_P^{\text{res}} \quad (5.6b)$$

Here, the residual contributions, indicated by a "res" superscript, are obtained from the FE-ANN/FE-ANN(s) EoSs using Eqs. (2.3a) and (2.3e). The ideal contributions ( $C_V^{\text{id}}$  and  $C_P^{\text{id}}$ ) are computed from the correlations from the DIPPR project 801 [13]. These correlations consider rotational and vibrational contributions to the ideal gas heat capacities that are not considered in the ideal gas law.

Finally, the thermal conductivity is obtained from the recommended ANN-based model trained in Chapter 4. This model uses reduced units and a semi-log scale. Therefore, the thermal conductivity of the Mie fluid is obtained as follows.

$$\kappa^{\text{ANN}} = \frac{k_B \sqrt{N_{\text{Av}} \epsilon / M}}{\sigma^2} \exp[\ln \kappa^*] = \frac{k_B \sqrt{N_{\text{Av}} \epsilon / M}}{\sigma^2} \exp \left[ \text{ANN}(\alpha_{\text{vdw}}, \rho N_{\text{Av}} \sigma^3, \frac{\epsilon}{T k_B}) \right] \quad (5.7)$$

In Eq. (5.7),  $\kappa^{\text{ANN}}$  is the thermal conductivity (in  $\text{W m}^{-1} \text{K}^{-1}$ ). This ANN-based model considers the thermal conductivity from non-bonded interactions. Real molecules also exhibit a thermal conductivity contribution due to bond vibrations. The ANN-based model does not account for this contribution. The following correction term has been proposed to address this issue [286].

$$\kappa^{\text{vib}} = \rho D \left( C_V^{\text{id}} - \# \text{ non-vibrational degrees of freedom} \cdot \frac{R}{2} \right) \quad (5.8)$$

The correction term considers the self-diffusivity (as  $\rho D$ ), which can be obtained from the ANN-based model in Chapter 4. The ideal gas isochoric heat capacity is obtained from the correlations of the DIPPR project 801 [13]. The number of non-vibrational degrees of freedom in a molecule depends on its molecular structure. Monoatomic molecules, such as ideal gases, have three non-vibrational degrees of freedom. Linear molecules, such as Carbon monoxide, have five non-vibrational degrees of freedom. Non-linear molecules, like Methane, have six non-vibrational degrees of freedom. The correction term  $\kappa^{\text{vib}}$  is small for liquid phases. For the molecules of interest of this chapter  $\kappa^{\text{vib}} / \kappa^{\text{ANN}} \lesssim 10^{-2}$ . Finally, the thermal conductivity is obtained by adding Eq. (5.7) and (5.8).

$$\kappa = \kappa^{\text{ANN}} + \kappa^{\text{vib}} \quad (5.9)$$

TABLE 5.2: Optimization details to fit the Mie potential molecular parameters.

|                | Units | Initial guess <sup>a</sup> |           | Optimization bounds <sup>b</sup> |           |
|----------------|-------|----------------------------|-----------|----------------------------------|-----------|
|                |       | Min value                  | Max value | Min value                        | Max value |
| $\sigma$       | Å     | 2.5                        | 3.5       | –                                | –         |
| $\epsilon/k_B$ | K     | 100                        | 300       | –                                | –         |
| $\lambda_r$    | –     | 9                          | 26        | 8                                | 34        |

<sup>a</sup> 32 initial guesses are generated in these ranges using Sobol' sequences [288]. <sup>b</sup> The repulsive exponent is bounded to be in the validity region of the ANNs-based models.

### 5.2.3 Multi-start optimization approach

The objective functions, Eq. (5.1), are minimized to find suitable parameters  $(\sigma, \epsilon, \lambda_r)$  to describe the phase behavior of a given molecule. The objective function is minimized using the Nelder-Mead method [287] implemented in the `minimize` function from the `scipy.optimize` Python library [146]. The Nelder-mead method, also known as the simplex method, is a derivative-free local minimization algorithm capable of handling bounded optimization. Depending on the initial guess, this optimization method can easily get stuck in a local minimum of the optimization surface. In order to find an optimal "global" set of molecular parameters, a multi-start optimization approach using 32 initial guesses generated through Sobol' sequences [288] is used here. This number of initial guesses was chosen to enhance the likelihood of reaching multiple local optimal solutions, from which the one with the lowest objective value is selected. The ranges of the initial guesses are detailed in Table 5.2. Lastly, to operate in the validity region of the ANN-based models, the repulsive exponent is restricted to be in the range  $\lambda_r \in [8, 34]$ .

## 5.3 Parametrization results for quasi-spherical molecules

The molecular parameters of the Mie potential have been determined for representative quasi-spherical molecules. Different parameter sets were obtained following the parametrization approaches outlined in Section 5.2. The optimized molecular parameters and their corresponding deviations are summarized in Table 5.3. This table also includes the parameters predicted by the M&M correlation [126]. Across the board, it is evident from this table that the VLE can be modeled with low deviations ( $\text{MAPE} \lesssim 5\%$ ) when it is the sole target property (i.e.,  $\text{OF}_1$ ). It must be noted that the parameters obtained from this objective function are in good agreement with the M&M correlation. This similarity results from the M&M correlation using the critical temperature, acentric factor, and liquid density to predict molecular parameters that correctly predict the VLE conditions at  $T_r = 0.7$ . However, this correlation is parametrized to predict a correct density behavior, and higher deviations ( $> 10\%$ ) are observed for the saturation pressure ( $P_{\text{VLE}}$ ).

Using  $\text{OF}_1$  (or the M&M correlation) provides a good description of the VLE. However, overfitting the VLE phase envelope commonly comes at the cost of increasing the repulsive exponent and interaction energy of the Mie potential. As will be shown later in Figures 5.3, 5.4, and 5.5, these apparent good fit results from stretching the validity range of a fluid EoS into regions where a solid phase would appear. This overfitting is not problematic if only fluid properties are considered, but it has implications if one wishes to transfer the molecular model to other obtain properties. For instance, an erroneous

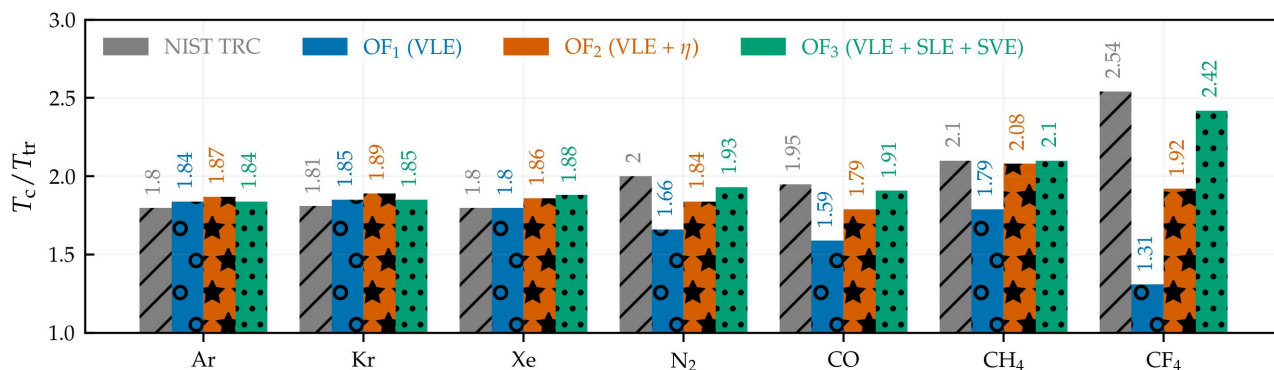


FIGURE 5.1: **Critical and triple temperature ratio of selected quasi-spherical molecules.** Grey bars with diagonal hatch refer to the value obtained from the NIST TRC database [12]. Blue bars with a circle pattern refer to the values obtained from OF<sub>1</sub> (VLE only). Orange bars with a star pattern refer to the values obtained from OF<sub>2</sub> (VLE + shear viscosity). Green bars with a dotted pattern refer to the values obtained from OF<sub>3</sub> (VLE + SLE + SVE).

prediction of the triple point greatly influences shear viscosity predictions, which are known to increase exponentially as they approach solid-phase conditions [289]. Therefore, it is convenient to use an EoS that also predicts the triple point to assess the validity of the molecular parameters. The ratio between the critical ( $T_c$ ) and the triple ( $T_{tr}$ ) points for the selected molecules is shown in Figure 5.1. From this figure, it is apparent that the  $T_c/T_{tr}$  ratio is well represented for ideal monoatomic gases when using only VLE data. Including the shear viscosity (OF<sub>1</sub>) or solid-fluid equilibria (OF<sub>3</sub>) data is necessary to predict this ratio for polyatomic molecules more effectively. These results set the stage for the main discussion below, which explains how achieving acceptable fits for certain thermophysical properties comes at the cost of producing an unreliable representation of others. The selected quasi-spherical molecules can be categorized into three major groups: monoatomic (Argon, Krypton, and Xenon), linear (Nitrogen and Carbon Monoxide), and tetrahedral (Methane and Tetrafluoromethane) molecules. The discussion will be done below based on these groups. Further results for molecules not covered in this section are available in the Appendix C.

### 5.3.1 Monoatomic molecules

Three monoatomic noble gases are considered here: Argon (Ar), Krypton (Kr), and Xenon (Xe). The results summarized in Table 5.3 show that the molecular parameters of each molecule are similar irrespective of which objective function is used. Consequently, the deviations of each thermophysical property fall within the same ranges. Nevertheless, interesting insights emerge when considering transport properties or solid-fluid equilibria data. It can be observed that including the shear viscosity in the objective function (OF<sub>2</sub>) tends to decrease the deviations in the properties related to SLE and SVE. It is understood that the shear viscosity and the description of the solid phase are sensitive to short-distance effects between molecules. The repulsive part of an interaction potential is relevant at these short distances. Therefore, improving the description of one property benefits the other. Notably, the optimized molecular parameters correctly predict the thermal conductivity and speed of sound of these molecules. The thermal conductivity is predicted accurately for Argon and Xenon (MAPE  $\kappa \lesssim 4\%$ ). For Krypton, the thermal conductivity is predicted with higher deviations, on the

TABLE 5.3: Optimized Mie molecular parameters for selected quasi-spherical molecules.<sup>a, b</sup>

| Component           |                  | MAPE <sup>c, d</sup> |                  |               |             |                |                  |              |             |                    |             |                 |             |
|---------------------|------------------|----------------------|------------------|---------------|-------------|----------------|------------------|--------------|-------------|--------------------|-------------|-----------------|-------------|
|                     |                  | $\sigma^e$           | $\epsilon/k_B^e$ | $\lambda_r^e$ | $P_{VLE}$   | $\rho_{l,VLE}$ | $\Delta H_{VLE}$ | $P_{SLE}$    | $P_{SVE}^f$ | $\Delta H_{SVE}^f$ | $\eta$      | $\kappa^{g, h}$ | $w^h$       |
| Argon               | M&M <sup>i</sup> | 3.41                 | 126.70           | 14.34         | 11.2        | <b>1.65</b>    | <b>1.58</b>      | 9.98         | 6.10        | 1.67               | 2.55        | 2.86            | <b>1.43</b> |
|                     | OF <sub>1</sub>  | 3.43                 | 124.98           | 13.52         | 1.13        | 2.71           | 3.29             | 5.75         | 4.43        | 1.57               | 0.85        | <b>1.16</b>     | 4.59        |
|                     | OF <sub>2</sub>  | 3.44                 | 123.19           | 13.16         | <b>0.90</b> | 3.31           | 2.70             | 4.21         | <b>0.34</b> | <b>0.57</b>        | 1.80        | 1.24            | 4.56        |
|                     | OF <sub>3</sub>  | 3.42                 | 124.62           | 13.53         | 1.90        | 2.04           | 2.23             | <b>3.68</b>  | 0.35        | 1.20               | <b>0.57</b> | 1.29            | 3.95        |
| Krypton             | M&M <sup>i</sup> | 3.63                 | 172.12           | 13.72         | 13.4        | <b>1.58</b>    | <b>1.71</b>      | <b>5.13</b>  | 16.7        | 0.46               | <b>1.63</b> | 15.5            | <b>1.87</b> |
|                     | OF <sub>1</sub>  | 3.67                 | 172.85           | 13.37         | 0.92        | 2.82           | 3.42             | 8.47         | 2.92        | 1.69               | 2.52        | 14.4            | 5.12        |
|                     | OF <sub>2</sub>  | 3.68                 | 169.42           | 12.89         | <b>0.59</b> | 3.86           | 2.56             | 8.69         | 3.79        | <b>0.31</b>        | 1.85        | 16.3            | 4.98        |
|                     | OF <sub>3</sub>  | 3.66                 | 172.84           | 13.44         | 1.77        | 2.13           | 2.63             | 7.65         | <b>0.36</b> | 1.53               | 2.39        | <b>14.3</b>     | 4.58        |
| Xenon               | M&M <sup>i</sup> | 3.96                 | 246.86           | 14.73         | 13.8        | <b>1.62</b>    | <b>1.85</b>      | 19.7         | 10.0        | 5.89               | 3.31        | 5.67            | <b>3.70</b> |
|                     | OF <sub>1</sub>  | 4.00                 | 245.07           | 14.00         | 0.93        | 3.21           | 3.88             | 17.4         | 19.0        | 6.34               | 2.62        | 4.48            | 7.32        |
|                     | OF <sub>2</sub>  | 4.01                 | 238.49           | 13.28         | <b>0.42</b> | 4.02           | 2.98             | <b>4.46</b>  | <b>7.13</b> | 4.10               | <b>2.18</b> | 2.23            | 5.96        |
|                     | OF <sub>3</sub>  | 4.00                 | 235.37           | 12.96         | 1.00        | 3.39           | 2.48             | 7.05         | 7.34        | <b>3.39</b>        | 3.18        | <b>1.76</b>     | 5.83        |
| Nitrogen            | M&M <sup>i</sup> | 3.65                 | 120.65           | 19.11         | 11.9        | <b>1.59</b>    | 2.82             | 69.2         | 44.8        | 10.9               | 17.3        | <b>7.81</b>     | <b>4.47</b> |
|                     | OF <sub>1</sub>  | 3.66                 | 115.19           | 16.30         | <b>1.15</b> | 1.89           | <b>2.67</b>      | 63.7         | 39.0        | 7.79               | 15.5        | 9.57            | 6.03        |
|                     | OF <sub>2</sub>  | 3.67                 | 105.83           | 13.57         | 3.03        | 4.76           | 2.69             | 53.8         | 11.1        | 1.58               | <b>2.70</b> | 17.0            | 4.56        |
|                     | OF <sub>3</sub>  | 3.62                 | 101.64           | 12.39         | 4.07        | 1.71           | 3.95             | <b>46.5</b>  | <b>1.44</b> | <b>0.59</b>        | 5.31        | 16.9            | 6.70        |
| Carbon Monoxide     | M&M <sup>i</sup> | 3.68                 | 130.95           | 20.67         | 11.8        | <b>1.16</b>    | <b>2.31</b>      | 116.6        | 45.5        | 8.07               | 15.7        | 3.51            | 7.22        |
|                     | OF <sub>1</sub>  | 3.69                 | 126.29           | 17.89         | <b>0.98</b> | 1.46           | 2.65             | 119.2        | 42.3        | 5.77               | 13.2        | <b>3.07</b>     | 8.70        |
|                     | OF <sub>2</sub>  | 3.63                 | 115.70           | 14.20         | 3.60        | 2.34           | 3.72             | 129.1        | 16.7        | <b>0.31</b>        | <b>12.3</b> | 7.09            | <b>5.61</b> |
|                     | OF <sub>3</sub>  | 3.65                 | 108.96           | 12.60         | 5.38        | 1.91           | 4.45             | <b>83.7</b>  | <b>3.21</b> | 3.88               | 17.4        | 13.2            | 5.88        |
| Methane             | M&M <sup>i</sup> | 3.74                 | 162.71           | 14.79         | 14.3        | <b>2.16</b>    | 3.65             | 65.8         | 27.0        | 9.12               | 19.6        | 5.48            | <b>1.40</b> |
|                     | OF <sub>1</sub>  | 3.78                 | 162.32           | 14.20         | <b>0.50</b> | 3.35           | <b>3.50</b>      | 66.1         | 38.8        | 9.93               | 24.8        | <b>3.16</b>     | 5.93        |
|                     | OF <sub>2</sub>  | 3.82                 | 139.81           | 10.83         | 5.25        | 7.48           | 4.58             | 38.3         | 9.82        | <b>0.64</b>        | <b>2.47</b> | 12.3            | 9.11        |
|                     | OF <sub>3</sub>  | 3.76                 | 139.96           | 10.68         | 4.62        | 3.29           | 5.55             | <b>35.9</b>  | <b>3.96</b> | 0.89               | 6.92        | 8.31            | 10.0        |
| Tetrafluoro-methane | M&M <sup>i</sup> | 4.38                 | 269.16           | 38.29         | 16.0        | 5.98           | 11.2             | 2858.6       | –           | –                  | 31.3        | 14.1            | <b>7.42</b> |
|                     | OF <sub>1</sub>  | 4.37                 | 266.77           | 34.00         | <b>4.10</b> | <b>1.97</b>    | <b>1.55</b>      | 4415.4       | –           | –                  | 44.8        | <b>9.52</b>     | 10.1        |
|                     | OF <sub>2</sub>  | 4.40                 | 191.17           | 12.59         | 17.0        | 11.0           | 12.4             | 797.1        | 114.9       | <b>6.93</b>        | <b>7.00</b> | 32.7            | 14.5        |
|                     | OF <sub>3</sub>  | 5.39                 | 153.76           | 8.00          | 68.2        | 49.4           | 15.3             | <b>341.7</b> | <b>14.8</b> | 15.6               | 36.9        | 55.1            | 21.4        |

<sup>a</sup> Thermophysical properties included in the objective functions are colored with a lightgrey background. Refer to Eq. (5.1) and Table 5.1 for further details about OF<sub>1</sub>, OF<sub>2</sub>, and OF<sub>3</sub>. <sup>b</sup> The objective functions are optimized using pseudo-experimental data obtained from the NIST TRC [12]. <sup>c</sup> The reported Mean Absolute Percentage Error (MAPE) is obtained as

$$\text{MAPE} = 100/N_P \sum_{i=1}^{N_P} |y_i^{\text{model}}/y_i^{\text{true}} - 1|.$$
 The reported MAPE include the vaporization pressure ( $P_{VLE}$ ), saturated liquid density ( $\rho_{l,VLE}$ ), vaporization enthalpy ( $\Delta H_{VLE}$ ), melting pressure ( $P_{SLE}$ ), sublimation pressure ( $P_{SVE}$ ), sublimation enthalpy ( $\Delta H_{SVE}$ ), shear viscosity ( $\eta$ ), thermal conductivity ( $\kappa$ ), and speed of sound ( $w$ ). <sup>d</sup> The smallest MAPE deviations per component are highlighted in **bold**.

<sup>e</sup>  $\sigma$ ,  $\epsilon/k_B$ , and  $\lambda_r$  are the molecular parameters of the Mie potential, Eq. (1.1).  $\lambda_a$  is set to 6.  $\sigma$  is reported in Angstroms ( $1 \text{ \AA} = 10^{-10}$  meters) and  $\epsilon/k_B$  is reported in Kelvin. <sup>f</sup> No MAPE is reported for the sublimation pressure and enthalpy when the pseudo-experimental data is below the validity of the FE-ANN(s) EoS, i.e.,  $T^* < 0.6$ . <sup>g</sup> The predicted thermal conductivity is obtained as in Eq. (5.9) which is based on Eq. (5.7) and (5.8). <sup>h</sup> The modeled thermal conductivity and speed of sound use the ideal gas heat capacities correlations from the DIPPR project 801 [13]. <sup>i</sup> Molecular parameters obtained using the M&M correlation [126]. This correlation relies on the critical temperature, acentric factor, and liquid density at  $T_r = 0.7$ .

order of 15%. Similarly, the Mie potential can predict the speed of sound with relative errors of approximately 5% for these ideal gases. The speed of sound is a stringent test for a set of molecular parameters, as it depends on multiple derivatives of the Helmholtz free energy. Moreover, the speed of sound can be accurately measured [290, 291], which enables the assessment of the limitations of the molecular model and its parameters.

The results from Table 5.3 are best interpreted graphically. The phase equilibria and selected thermophysical property results of Argon are shown in Figure 5.2. Similar figures for Krypton and Xenon can be found in Appendix C. Figure 5.2 considers the three sets of parameters along with two possible modeling approaches. Figures 5.2 (a) to (c) show the phase equilibria results modeled using the fluid FE-ANN EoS obtained in Chapter 2. This EoS can only predict VLE. Overall, a good agreement with the reference VLE data can be noted when using the OF<sub>1</sub> and OF<sub>2</sub> parametrization approaches. Figures 5.2 (d) to (f) consider the same set of molecular parameters but using the FE-ANN(s) EoS obtained in Chapter 3. This EoS can predict the entire phase diagram, including VLE, SLE, SVE, and the triple point. It can be observed that even when solid phase data is not included in the parametrization, all modeling approaches predict a similar behavior for the phase envelope, equilibrium pressure, and enthalpy of phase change. As a reference, these sets of parameters also accurately predict the triple temperature, which is marked by a solid black line in the figures.

The results for the speed of sound, shear viscosity, and thermal conductivity of the saturated liquid Argon are presented in Figures 5.2 (g) to (i). Similar to the phase equilibria results, the three parametrization approaches accurately predict these thermophysical properties. These findings further confirm that the Mie potential can accurately model the thermophysical properties of simple monoatomic gases. Notably, the optimal repulsive exponent for the noble gases is in the 13-13.5 range, as opposed to the value of 12 employed in the classical Lennard-Jones potential.

### 5.3.2 Linear molecules

While the different parametrization approaches did not significantly affect the molecular parameters of noble gases, a different scenario unfolds as the molecules become more complex. This section will discuss the effect of the parametrization approach on small linear molecules, such as Nitrogen (N<sub>2</sub>) and Carbon Monoxide (CO). From the results shown in Table 5.3 for these molecules, it can be observed that the VLE properties are better predicted with OF<sub>1</sub> at the expense of increasing the repulsiveness ( $\uparrow \lambda_r$ ) and energy ( $\uparrow \epsilon$ ) of the interactions. The shape parameter ( $\sigma$ ), which is closely related to the liquid density, remains within the same order of magnitude across the three parametrization approaches. Including the shear viscosity (OF<sub>2</sub>) or solid (OF<sub>3</sub>) data in the optimization results in decreasing the repulsive exponent and interaction energy of the potential. These two properties target molecular parameters in the same direction. It can be observed that there is a trade-off regarding which properties fit better with a simple pairwise potential. A better description of a transport property compromises the description of the equilibrium properties. Moreover, the melting pressure cannot be modeled correctly with the Mie potential. Presumably, the coarse grained approximation overlooks structural effects that influence the conformation of the crystal phase as well as the friction between molecules that affect the transport properties. These results indicate that the coarse grained

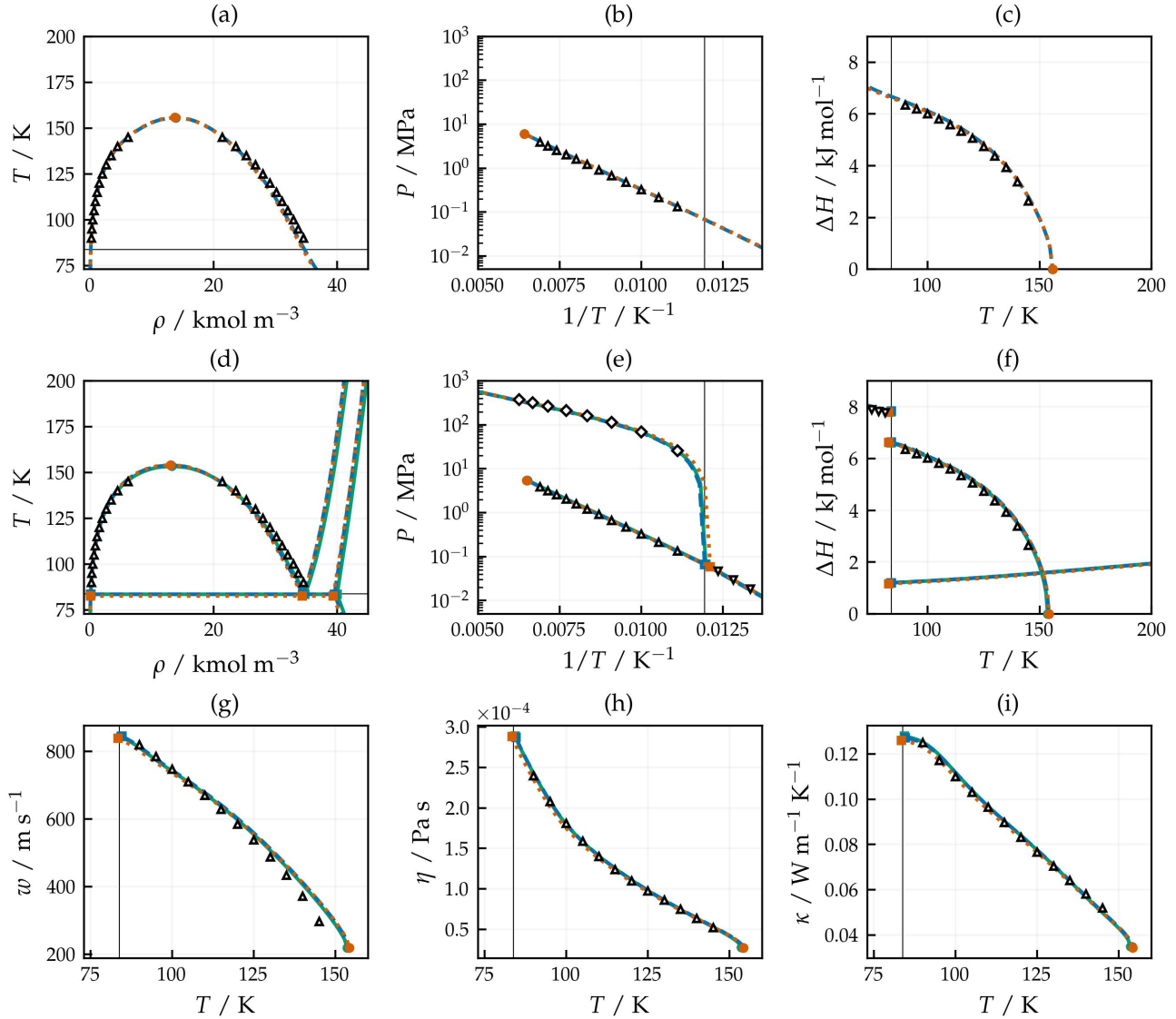


FIGURE 5.2: **Selected thermophysical properties of Argon.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SLE data. For reference, the solid black line refers to the triple temperature.

approximation is, in most cases, overfitted to a single property, and its extrapolation to other properties must be approached with caution.

The modeling results for Nitrogen are presented in Figure 5.3. Figures 5.3 (a) to (c) illustrate the VLE results using the fluid FE-ANN EoS. A good agreement can be observed for the VLE phase envelope, vaporization pressure, and vaporization enthalpy when using OF<sub>1</sub> or OF<sub>2</sub>. However, it is important to note that the fluid FE-ANN EoS is stretched to predict VLE in regions where the Mie particle would exhibit a solid phase. In Figures 5.3 (d) to (f), the entire phase diagram is modeled using the FE-ANN(s) EoS using the different parameter sets. These figures clearly demonstrate that the triple point predicted by the parameters obtained from OF<sub>1</sub> is above some experimental VLE data points. Under those conditions, the VLE becomes unstable, and the Mie particle exhibits SVE. Incorporating the shear viscosity or solid phase data decreases the repulsive exponent of the Mie potential and, consequently, lowers the triple point. The parameters from OF<sub>2</sub> and OF<sub>3</sub> model the VLE similarly to those from OF<sub>1</sub>, but they provide a better description of the triple point, SLE, and SVE. It is noteworthy that the order of magnitude of the melting pressure ( $\sim 10^3$  MPa) is considerably higher than the vaporization ( $\sim 10^{-1}$  MPa) and sublimation ( $\sim 10^{-3}$  MPa) pressures. Although the molecular parameters obtained from OF<sub>3</sub> predict the correct melting pressure trend, the reported relative errors are considerable ( $\sim 50\%$ ). Finally, the differences between the various molecular parameters result in noticeably different SLE densities. This information could be useful for distinguishing the optimal molecular parameters; however, this data is unavailable from the NIST TRC correlations [12].

The liquid speed of sound results of Nitrogen are shown in Figure 5.3 (g). It is observed here that the three parametrization approaches led to a similar description of this property. However, it should be noted that the described ranges for these properties depend on the predicted critical and triple points. For the case of the shear viscosity, shown in Figure 5.3 (h), it can be observed that including properties related to the repulsive part of the force field in the parametrization enhance the accurate description of this property. In this case, the parameters from OF<sub>2</sub> and OF<sub>3</sub> exhibit a similar behavior. This finding contrasts with the observations for the thermal conductivity, as shown in Figure 5.3 (i). These results graphically show how, in the coarse grained modeling approach, there are trade-offs regarding which properties to fit (shear viscosity or thermal conductivity in this case). Similar results and conclusions can be drawn for Carbon Monoxide from analyzing Figure C.3 in Appendix C.

### 5.3.3 Tetrahedral molecules

It has been shown that the Mie potential can accurately model the entire phase diagram of ideal gases. However, this flexibility of the potential vanishes when dealing with polyatomic molecules, such as linear molecules. This final section will further analyze the limitations of the Mie potential to model tetrahedral molecules, such as Methane (CH<sub>4</sub>) and Tetrafluoromethane (CF<sub>4</sub>). The results for these two molecules are shown in Figures 5.4 and 5.5 respectively. When only VLE information using the fluid FE-ANN EoS, figures (a) - (c), there is an apparent good fit of the phase envelope, saturation pressure, and vaporization enthalpy. Results such as these have led to claims that the Mie force field holds an advantage over the standard Lennard-Jones model, in that the use of the additional parameter ( $\lambda_r$ ) allows for a quantitative description of the fluid phase equilibria [292]. This is certainly the case, but these apparent "good" results emerge from predicting high repulsions and interaction energy

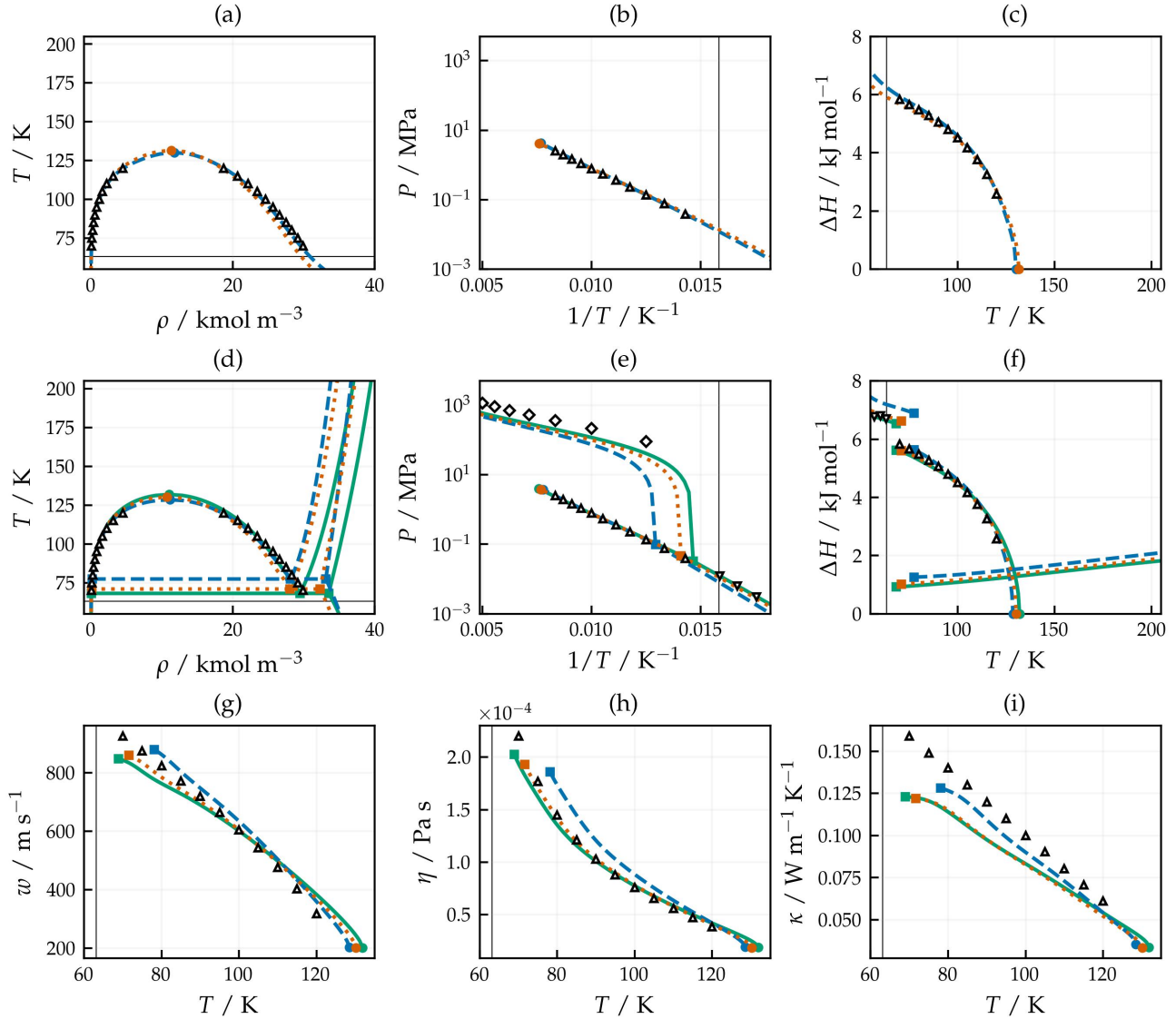


FIGURE 5.3: **Selected thermophysical properties of Nitrogen.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.



of the force field when using  $OF_1$ . In the case of Tetrafluoromethane, the optimization reaches the upper limit of the FE-ANN EoS with  $\lambda_r = 34$ . The overfitting of the VLE properties is evident when analyzing the high deviations of the shear viscosity and melting pressure in Table 5.3. As with the case of linear molecules, the use of the overfitted parameters unphysically predicts VLE in regions below the triple point. These regions are outside the validity range of the fluid FE-ANN EoS.

The effect of the various choices of molecular parameters when used with the FE-ANN(s) EoS and transport property models are shown in subfigures (d)-(i). For the case of Methane, Figure 5.4, it is observed that when including solid phase ( $OF_3$ ) or shear viscosity ( $OF_2$ ) in the parametrization, the VLE is correctly modeled. In these cases, the prediction of the triple point, SLE, and SVE are also improved. The parameters from  $OF_1$  result in overestimating the triple point, which results in instability in part of the VLE region. Both  $OF_2$  and  $OF_3$  objective functions lead to lower repulsions and interaction energies of the system and similar molecular parameters. In this context, it is seen that incorporating a thermophysical property related to the repulsive interactions benefits the description of the phase diagram and transport properties. This result is also observed for linear molecules and aligns with the corresponding states methodology proposed by Hoang et al. [111].

In the case of Tetrafluoromethane, Figure 5.5, the apparent limitations of the Mie potential become more apparent. In Figure 5.5 (d), (e) and (f), it is shown how the highly repulsive force field (i.e.,  $\lambda_r = 34$ ) obtained from  $OF_1$  lead to a phase diagram where a considerable part of the experimental VLE data points are below the triple line. Below this temperature, the Mie particle will exhibit a solid phase. Moreover, when incorporating solid ( $OF_3$ ) or shear viscosity ( $OF_2$ ) data, there is no clear optimal set of parameters that can adequately describe the phase equilibrium properties. Notably, all these parameter sets fail to describe the triple temperature of Tetrafluoromethane. It is not a surprise that other thermophysical properties, such as the speed of sound, Figure 5.5 (g), and thermal conductivity, Figure 5.5 (i), are not described correctly. Tetrafluoromethane ( $CF_4$ ), similar to Methane ( $CH_4$ ), is a nonpolar molecule due to the tetrahedral geometry of the molecule. However, the fluorine atoms in  $CF_4$  are highly electronegative and of bigger size than the hydrogen atoms in  $CH_4$ . The fluorine atoms in  $CF_4$  can lead to strong local polar moments, an overall octopolar moment, and steric effects that a simple single-bead coarse grained model does not take into account. This result unveils the limitations of the classical coarse grained modeling approach in describing more complex molecules. Coarse grained models can easily overfit properties like the VLE, but this parametrization does not necessarily transfer well to other thermophysical properties.

## 5.4 Conclusions

This chapter has showcased the limitations of the Mie potential for modeling quasi-spherical molecules. The target molecules include monoatomic (e.g., Argon, Krypton, and Xenon), linear (e.g., Nitrogen and Carbon monoxide), and tetrahedral (e.g., Methane and Tetrafluoromethane) molecules. Three different approaches were employed to parametrize these molecules. The first parametrization approach considered only vapor-liquid equilibria data. The second parametrization approach took into account the liquid shear viscosity and the vapor-liquid equilibria data. Finally, the third parametrization

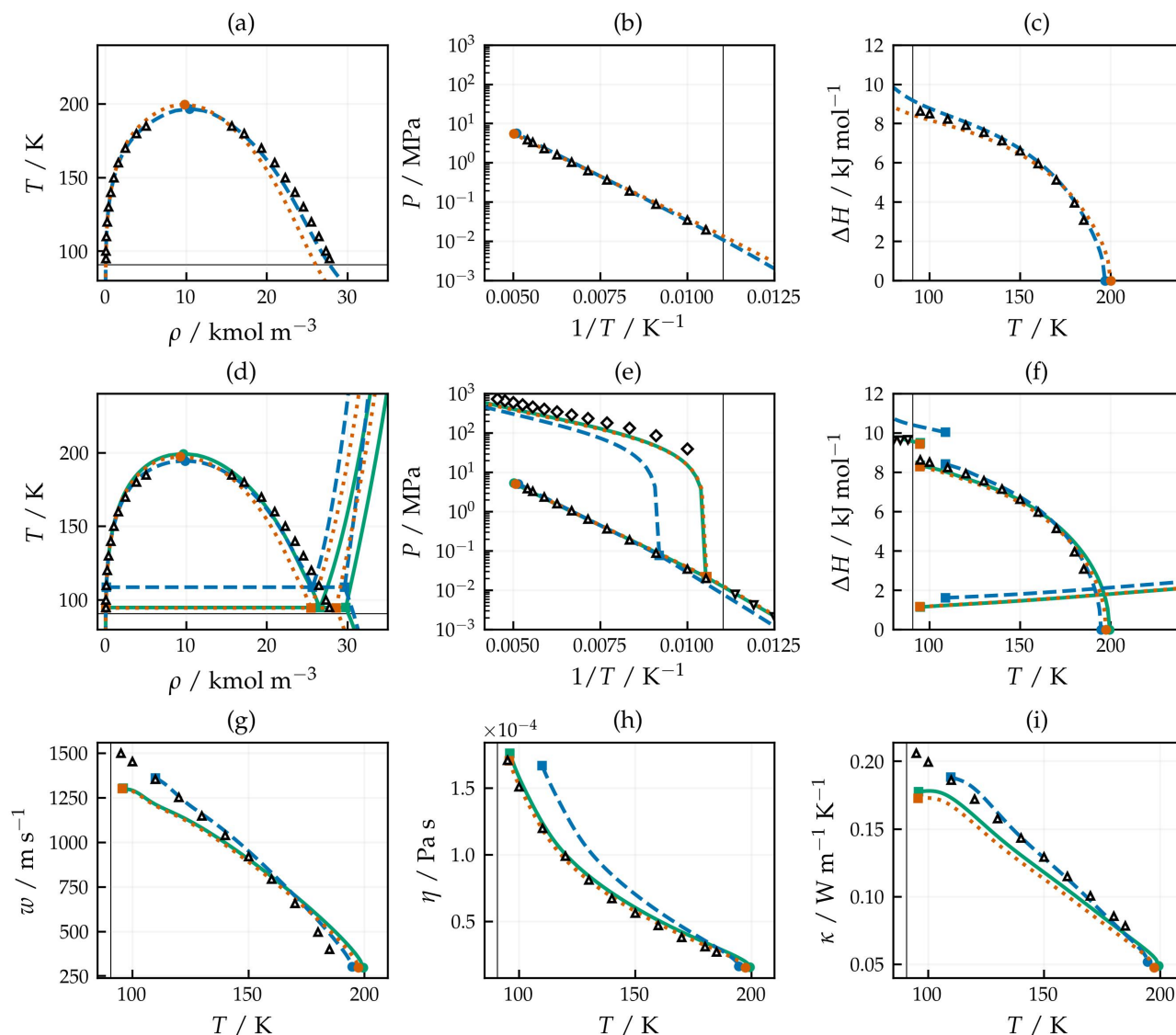


FIGURE 5.4: **Selected thermophysical properties of Methane.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.

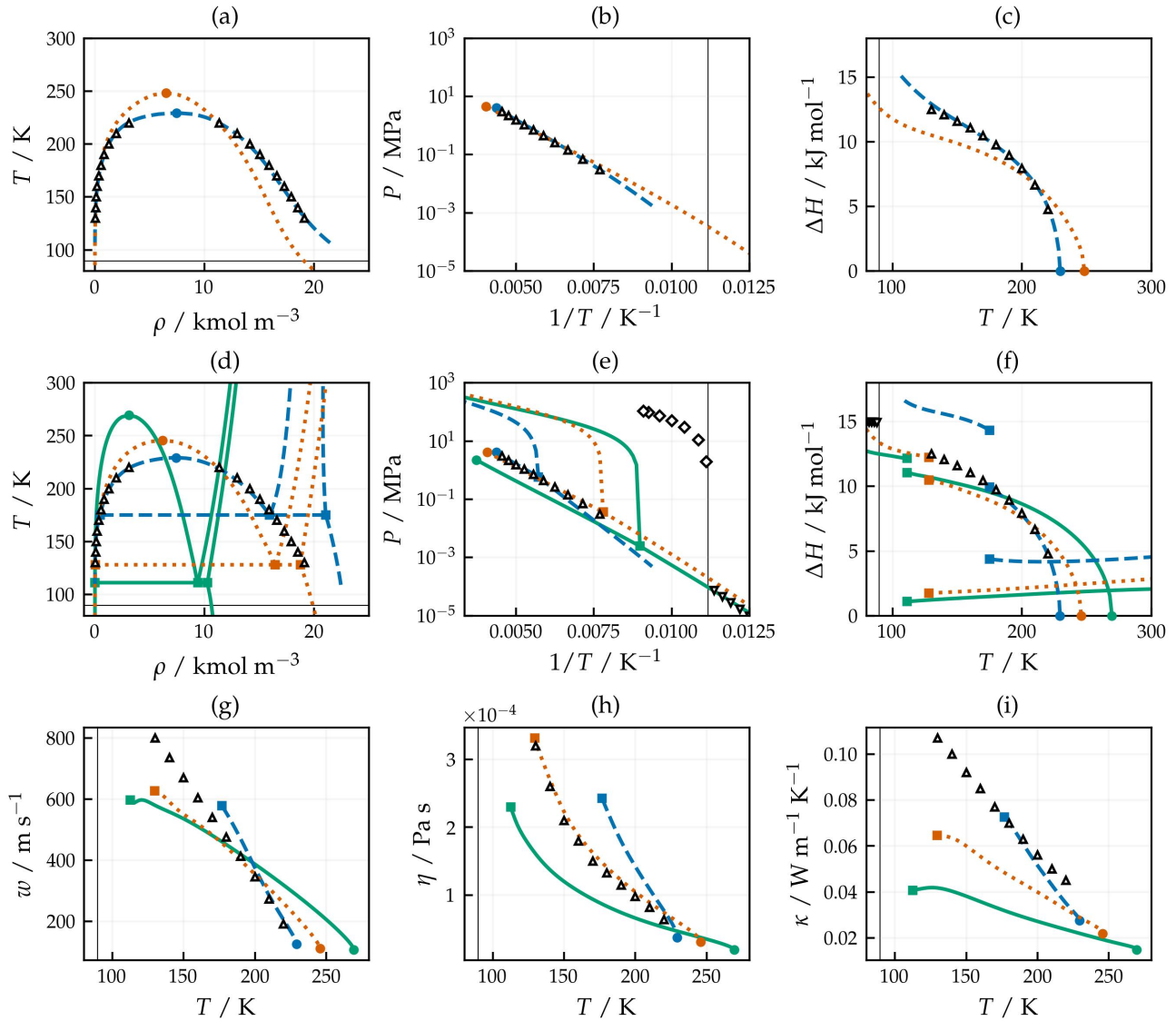


FIGURE 5.5: **Selected thermophysical properties of Tetrafluoromethane.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.

approach included solid-fluid (solid-liquid and solid-vapor) equilibria data in addition to the vapor-liquid equilibria data.

The Mie potential was found to accurately describe the entire range of properties of simple monoatomic molecules, such as Argon, Krypton, and Xenon. This result was validated irrespective of the thermophysical properties used to parametrize the Mie potential. However, the limitations of the Mie potential become evident beyond ideal gases. Vapor-liquid equilibria data can always be overfitted for linear and tetrahedral molecules when it is the only target property. This approach is a customary parametrization strategy in SAFT-like EoSs [36, 37, 172]. This commonly used parametrization approach causes a shift in the triple point, leading to instability in a region of the VLE phase envelope. Furthermore, this parametrization leads to suboptimal parameters that provide a poor description of other thermophysical properties, particularly transport properties. Unlike noble gases, which interact solely through van der Waals forces, polyatomic molecules can exhibit additional sources of non-ideality. Examples include dipolar and higher polar moments, polarization effects, and steric effects caused by molecular geometry. These interactions are not accounted for when modeling a polyatomic molecule using a coarse grained approach with a single bead per molecule.

The various proposed parametrization approaches indicate that incorporating properties such as the shear viscosity or solid-fluid equilibria data leads to similar molecular parameters. This behavior results from the fact that both of these properties are related to short-range effects. The shear viscosity is related to intermolecular friction, which is significant at short distances. Similarly, the solid phase and crystal structure are directly related to relatively short-distance interactions. In contrast, vapor-liquid equilibria depends on the balance between repulsive and attractive interactions. This interplay can be captured by multiple sets of molecular parameters that correctly describe the vapor-liquid equilibria region. These parameters do not necessarily extrapolate well to other thermophysical properties. Therefore, it is advisable to include properties related to the repulsive interactions of the system when parametrizing force fields. This observation aligns with the corresponding states methodology proposed by Hoang et al. [111], which considers the shear viscosity in the force field parametrization.

Finally, the parametrization approaches presented in this chapter rely on physics-informed artificial neural networks to model thermophysical properties. These thermophysical property models, developed in Chapters 2, 3, and 4, are equations of state that are derived directly from discrete molecular simulation data and physics-inspired constraints employing machine learning techniques. These models serve as reliable surrogates for molecular simulation results and represent a promising path for parametrizing force fields using a top-down approach. It should be noted that the observations made in this chapter are limited to simple spherical geometry and London dispersion interactions. Future work should aim to extend the framework to incorporate additional interactions and molecular geometries.

## Chapter 6

# Conclusions and outlook

Thermophysical properties are of great relevance for scientific and engineering purposes. The modeling of these properties has traditionally been done using empirical, semi-empirical, and theoretically based methods. This thesis proposes a paradigm shift by modeling thermophysical properties using thermodynamically-guided machine learning methods. These methods are demonstrated for a system interacting through the well-characterized Mie potential.

In Chapter 2, a physics-embedded artificial neural network framework that explicitly models the Helmholtz free energy (FE-ANN EoS) was proposed. The FE-ANN EoS is formulated to fulfill the ideal gas law and Maxwell's relationships. This EoS is trained using derivative information obtained through high-throughput molecular dynamics simulations. The FE-ANN EoS additionally considers penalties to fulfill well-established principles, such as the principle of mechanical and thermal stability. The FE-ANN EoS can learn the Helmholtz free energy from discrete single-phase data. Moreover, due to its physics-inspired formulation, it can predict previously unseen behavior, such as vapor-liquid equilibria. Finally, the thermodynamic consistency of the FE-ANN EoS was verified using Brown's characteristic curves. The FE-ANN EoS framework has been validated for the Mie fluid. However, the flexibility of the data-driven approach allows this framework to study other complex nonideal behaviors, such as permanent dipoles, charged systems, and extreme phase space conditions.

In Chapter 3, the FE-ANN EoS framework is extended to handle the crystal phase (FE-ANN(s) EoS). The FE-ANN(s) EoS retains the physics-inspired formulation of the FE-ANN EoS but is trained using single-phase data from solid and fluid states. The FE-ANN(s) EoS is capable of accurately modeling thermophysical properties. Moreover, it can predict the entire phase diagram, including vapor-liquid, solid-liquid, and solid-vapor phase envelopes. The FE-ANN(s) EoS also predicts the existence of unique critical and triple points. This EoS exhibits correct physical trends for the behavior of the isochoric heat capacity in unstable regions, the vanishing of the vapor-liquid equilibria for highly repulsive systems, and accurately predicts Brown's characteristic curves. The FE-ANN(s) EoS exhibits a thermodynamically correct behavior, and it is the first equation of state capable of continuously handling fluid and solid states. This type of EoS supports the premise of Maxwell and Gibbs that a single continuous thermodynamic surface spans the entire phase space. Finally, the FE-ANN(s) EoS showcases the versatility of using a hybrid physics-informed machine learning modeling approach and how it is possible to go beyond the state of current theoretical approaches.

Thermophysical properties include volumetric (equilibrium) and transport properties. Unlike volumetric properties, no widely accepted theory exists for modeling transport properties. For this reason, Chapter 4 has two main objectives. First, in this chapter, a reference database for the self-diffusivity, shear viscosity, and thermal conductivity of the Mie fluid is computed from high-throughput molecular dynamics simulations. This database aims to serve as a validation tool for testing modeling approaches for transport properties. Additionally, these transport properties were successfully modeled using physics-informed artificial neural networks. These neural networks account for the dilute gas limit and correctly capture the density dependence of the transport properties. The developed artificial neural network models aid in assessing the performance of single bead coarse grained models based on the Mie potential.

Finally, in Chapter 5, the equations of state (FE-ANN and FE-ANN(s) EoS) and transport property models developed in Chapters 2, 3, and 4 are used to test different parametrization approaches for quasi-spherical molecules based on the Mie potential. The parametrization approaches are considered, considering only vapor-liquid equilibria data, vapor-liquid equilibria and liquid shear viscosity data, and vapor-liquid, solid-liquid, and solid-vapor equilibria data. The limitations of the Mie potential become evident when shear viscosity or solid-fluid equilibria data are included in the parametrization. Beyond the simplest cases, trade-offs are observed with respect to which property to fit. It is always possible to overfit the vapor-liquid equilibria data at the cost of increasing the repulsion and interaction energy of the system. However, these molecular parameters are not transferable to other thermophysical properties beyond ideal gases. Interestingly, including shear viscosity data or solid-fluid equilibrium data tends to yield similar molecular parameters. This behavior is expected as both properties are related to short-range interactions, which are relevant to the repulsive part of an interaction potential. Including either of these two properties leads to molecular parameters that generalize better.

## 6.1 Suggested future work

The results presented in this thesis represent the foundations for using physics-informed data-driven methods for modeling thermophysical properties. However, the methods presented in this work are limited to handling simple spherical particles, which limits the scope of applicability of these methods to simple quasi-spherical molecules. A few ideas are suggested in this section to extend the current framework to handle more complex molecules and mixtures.

### 6.1.1 Machine learning-based SAFT EoS

In the context of the Statistical Association Fluid Theory (SAFT), the free energy of a fluid is obtained as a sum of different contributions [172]. As a baseline, the free energy will consider a kinetic (ideal) contribution of non-interacting particles. In the SAFT framework, the residual (nonideal) Helmholtz free energy considers contributions due to non-bonded interaction ( $A^{\text{monomer}}$ ), chain formation ( $A^{\text{chain}}$ ), and for molecular association ( $A^{\text{association}}$ ). These contributions to the free energy can be added or neglected depending on the nature of the molecule. The chain term can be omitted if the molecular representation comprises a single sphere. Likewise, the association term can be neglected if the

molecules cannot associate (i.e., cannot form hydrogen bonds). Other corrections to the free energy can be considered if necessary. Other contributions include dipolar [38], quadrupolar [39] and ion [40] contributions.

$$A^{\text{res}} = A^{\text{monomer}} + A^{\text{chain}} + A^{\text{association}} \quad (6.1)$$

The FE-ANN EoS developed in Chapter 2 provides an accurate function for obtaining the monomer contribution ( $A^{\text{monomer}}$ ) of Mie monomers to the free energy. However, there is no restriction on using the framework to model other contributions to the free energy.

The chain contribution is typically obtained in the limit of infinite association from Wertheim's theory [35]. This term depends on the radial distribution function at contact, which is obtained as follows:

$$\frac{A^{\text{chain}}}{RT} = -(m_s - 1) \ln g(\sigma) \quad (6.2)$$

In Eq. (6.2),  $R$  is the ideal gas constant,  $T$  is the absolute temperature,  $m_s$  is the number of chain segments, and  $g(\sigma)$  is the radial distribution function evaluated at the molecular diameter. The function  $g(\sigma)$  can be interpreted as a measure of the likelihood that two monomers remain at a distance of  $\sigma$ . This information can be obtained from standard molecular simulation techniques [66, 70]. The function  $g(\sigma)$  can be modeled using a physics-informed method inspired by the FE-ANN EoS.

$$\ln g(\sigma) = \text{ANN}(\chi, \rho, \frac{1}{T}) - \text{ANN}(\chi, \rho = 0, \frac{1}{T}) \quad (6.3)$$

The proposed formulation, Eq. (6.3), relies on the molecular descriptors ( $\chi$ , e.g., the molecular parameters of the force field), and the density ( $\rho$ ) and temperature ( $T$ ). This formulation satisfies the ideal gas limit, where  $g(\sigma) = 1$ . The  $A^{\text{chain}}$  term derived from this ANN-based model for  $g(\sigma)$  must satisfy thermodynamic consistency tests, such as the *criteria of thermal and mechanical stability* [116, 125]. These additional constraints must be incorporated when developing the machine-learning-based SAFT EoS.

The association terms are also implicitly related to the structural information of the association sites. The association term can be obtained from Wertheim's theory as follows.

$$\frac{A^{\text{association}}}{RT} = \sum_{a=1}^s \left( \ln X_a - \frac{X_a}{2} \right) + \frac{s}{2} \quad (6.4)$$

Here,  $s$  represents the number of association sites per molecule, and  $X_a$  is the fraction of sites  $a$  that are not bonded at the given density and temperature. The solution for the non-bonded sites is given from the following mass-action relation.

$$X_a = \frac{1}{1 + \rho \sum_{b=1}^s X_b \Delta_{ab}} \quad (6.5a)$$

where,

$$\Delta_{ab} = \int g(r_{12}) [\exp(-\mathcal{U}(r_{12})/RT) - 1] dr_{12} \quad (6.5b)$$

The fraction of non-bonded association sites depends on the association strength ( $\Delta_{ab}$ ). The association strength depends on the selected interaction potential ( $-\mathcal{U}(r_{12})$ ) between the association sites and its radial distribution function  $g(r_{12})$ . The square-well potential has been commonly used to model the

association sites [35, 37, 132]. However, the Lennard-Jones and Mie potential can also be considered [293]. The calculation of the association strength ultimately depends on the radial distribution function  $g(r_{12})$ . This structural information can be obtained from molecular simulations and fitted to a physics-informed artificial neural network. Similar to the chain term, this ANN must fulfill well-established thermodynamic constraints [116, 125].

Incorporating the chain and association contributions into the current FE-ANN EoS will transform it into a machine-learning-based SAFT EoS. This EoS should be capable of handling chain molecules, like alkanes and polymers, as well as molecules that exhibit hydrogen bonding, such as alcohols and water.

### 6.1.2 Molecular descriptors embedded into the FE-ANN EoS

The FE-ANN EoS is a flexible framework for developing EoS from discrete thermophysical properties data. As described by Eq. (2.9b), the FE-ANN EoS explicitly models the Helmholtz free energy based on a molecular descriptor ( $\chi$ ) and the state conditions ( $\rho$  and  $T$ ).

$$A^{\text{res}} = \text{ANN}(\chi, \rho, \frac{1}{T}) - \text{ANN}(\chi, \rho = 0, \frac{1}{T}) \quad (2.9b)$$

Throughout this thesis, the FE-ANN EoS has been showcased for the Mie potential, which can be described by the van der Waals parameters ( $\alpha_{\text{vdw}}$ ). The FE-ANN EoS framework can be extended to more complex molecules by providing suitable molecular descriptors. Three potential routes to produce molecular descriptors will be described below.

In the machine learning community, several approaches to computationally describe molecules exist. For example, the RDKit [294] project allows computing numerical descriptors of molecules. These descriptors capture several molecular attributes, such as molecular volume, surface area, and electronic distribution. By default, RDKit computes a comprehensive list of over 200 descriptors, some of which may exhibit correlations. For example, RDKit descriptors have been successfully used to study the impact of terminal groups in lubricant monolayers [295]. Similarly, these descriptors can be used as the  $\chi$  vector in the context of the FE-ANN EoS.

Molecules can also be represented as graph structures, making them suitable for modeling with Graph Neural Networks (GNNs). Graphs structures consider "nodes" and "edges". The nodes represent the vertices of the graph, and the edges represent the connections between different nodes. In the context of a molecular representation, each atom is explicitly considered a node in the graph, and bonds are interpreted as the edges of the graph. GNNs can be used to produce a numerical representation of a molecule. GNNs have already been used to model thermophysical properties [63] and to formulate thermodynamically consistent activity coefficient models [84]. In the context of the FE-ANN EoS, a GNN can be trained to output a machine-learned molecular descriptor vector  $\chi$  as required by Eq. (2.9b).

Finally, molecular structures can be represented as strings. The self-referencing embedded string (SELFIES) [61] and simplified molecular input line entry system (SMILES) [60] are examples of well-defined molecular representations using strings. The attention mechanism of transformers can process



these strings [77] to encode molecular representations. This procedure has already been successfully used to predict molecular parameters of traditional thermodynamic models [80–82]. Therefore, this can also represent a suitable pathway to produce the molecular descriptor vector ( $\chi$ ) required by the FE-ANN EoS.

Regardless of the chosen approach for obtaining a numerical molecular description, extending the FE-ANN EoS to real molecules requires thermophysical property data. This data can be obtained from curated databases [12, 296] or high-throughput molecular simulations.

### 6.1.3 Application of the FE-ANN EoS to other molecular fluids

Finally, the FE-ANN EoS framework has been showcased for the Mie potential. This choice allowed for the comparison of the FE-ANN EoS with a theoretically based EoS, such as the SAFT-VR-Mie [37]. However, the FE-ANN EoS framework can be applied to model other non-bonded interactions, especially for systems without a theoretical representation of the free energy. These systems include, for example, the Buckingham potential, the Stockmayer fluid, and systems with charged species.

Similar to the Mie potential, the Buckingham potential considers repulsive and attractive interaction between particles. It has been argued that the exponential dependence on the repulsive part of the Buckingham potential is more physically sound than the  $r^{-\lambda_r}$  dependence in the Mie potential [27]. Therefore, the phase behavior can be better captured by using this interaction potential. The Buckingham potential is defined by the following equation.

$$\mathcal{U}^{\text{Buckingham}} = \epsilon \left[ \frac{\eta}{\alpha - \eta} \exp \left[ \alpha \left( 1 - \frac{r}{\sigma} \right) \right] - \frac{\alpha}{\alpha - \eta} \left( \frac{\sigma}{r} \right)^\eta \right] \quad (6.6)$$

In Eq. (6.6),  $\epsilon$  corresponds to the potential well depth,  $\sigma$  is the shape parameter,  $r$  is the center-to-center distance between two Buckingham particles. The repulsion and attractive parts of the potential are governed by the  $\alpha$  and  $\eta$  parameters, respectively. The attractive parameter ( $\eta$ ) is commonly set to six due to the London dispersion forces [93]. Hence, the Buckingham potential has three adjustable parameters ( $\epsilon$ ,  $\sigma$ , and  $\alpha$ ) to model intermolecular interactions.

The Buckingham potential is a substitute for non-bonded interactions modeled by the Mie potential. However, for some relevant molecules, additional contributions to the free energy must be considered. For example, a permanent dipolar moment is essential for modeling solvents like ethers. The dipolar contribution is explicitly accounted for in the Stockmayer fluid [297], whose interactions are detailed below.

$$\mathcal{U}^{\text{Stockmayer}}(r_{ij}) = 4\epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right] - \frac{\mu^2}{4\pi\epsilon_0 r_{ij}^3} [2 \cos \theta_i \cos \theta_j - \sin \theta_i \sin \theta_i \cos \phi_{ij}] \quad (6.7)$$

In this interaction potential, the first term represents the Lennard-Jones potential, while the second term accounts for dipole-dipole interactions. This term is governed by the dipole moment ( $\mu$ ), the polar angle of the dipoles ( $\theta$ ), and the azimuthal angle between the dipoles ( $\phi_{ij} = \phi_j - \phi_i$ ). In this equation,  $\epsilon_0$  is the vacuum permittivity.

Additionally, correctly modeling systems with electrolytes is relevant for several industrial and biological processes [14]. In the case of charged systems, the molecular simulation should include an explicit Coulomb potential as defined below.

$$\mathcal{U}^{\text{Coulomb}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (6.8)$$

In Eq. (6.8), the interaction energy ( $\mathcal{U}^{\text{Coulomb}}$ ) depends of the charges  $q_i$  and  $q_j$  of the species and its separation  $r_{ij}$ .

The process of modeling these contributions follows the same pipeline described in Figure 2.1. This pipeline produces data on discrete thermophysical properties from molecular dynamics simulations. These interaction potentials are readily implemented in standard molecular simulation packages, such as LAMMPS [147]. Subsequently, the molecular simulation data can be used to train a physics-informed neural network for modeling the Helmholtz free energy. In the context of contributions that rely on a base interaction potential, like the Stockmayer fluid or for charged species, it is recommended to formulate these contributions to the free energy using a perturbation approach, as this will allow considering these contributions as needed.

#### 6.1.4 Extension of the FE-ANN EoS to mixtures

The FE-ANN EoS framework developed in this thesis models the free energy of pure species. The framework can be extended to mixtures using two main independent approaches. The first approach relies on combining rules similar to the ones used in molecular-based equations of state [172]. In this case, the total Helmholtz free energy is defined as follows,

$$A(\mathbf{x}, \rho, T) = A^{\text{id}}(\mathbf{x}, \rho, T) + \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} x_i x_j A_{ij}^{\text{res}}(\chi_{ij}, \rho, 1/T) \quad (6.9)$$

In this formulation, the total Helmholtz free energy of a mixture ( $A$ ) is a function of the composition ( $\mathbf{x}$ ), density ( $\rho$ ), and temperature ( $T$ ). The total free energy can be divided into an ideal contribution ( $A^{\text{id}}$ ) and a residual contribution. The residual term is commonly taken as a quadratic sum of the pairwise interactions ( $A_{ij}^{\text{res}}$ ) [37, 65, 298]. These residual contributions are obtained using the FE-ANN EoS formulated in Eq. (2.9b). In this case, each molecule is described by like ( $\chi_{ii}$ ) and unlike ( $\chi_{ij}$ ) molecular parameters. For the Mie particle, the like molecular parameters ( $\sigma_{ii}, \epsilon_{ii}, \lambda_{r,ii}$ ) can be obtained as described in Chapter 5. For the unlike parameters, it is customary to rely on combining rules. It is expected that Eq. (6.9) will perform correctly for species of similar size. However, it might fail to correctly describe asymmetric systems. Moreover, the correctness of a quadratic term to predict the expected behavior of the virial coefficients has been criticized in open literature [299].

An alternative data-driven approach based on the thermodynamics of mixtures is proposed below.

$$A(\mathbf{x}, \rho, T) = \sum_{i=1}^{N_c} x_i A_i + \underbrace{RT \sum_{i=1}^{N_c} x_i \ln x_i}_{\Delta A^{\text{id}}} + A^{\text{ex}} \quad (6.10)$$

In this case, the total Helmholtz free energy of a mixture is obtained from three main contributions. The first contribution ( $\sum_i x_i A_i$ ) explicitly considers the Helmholtz free energy of the individual species as obtained by the FE-ANN EoS. The second contribution ( $\Delta A^{\text{id}}$ ) accounts for the free energy obtained from an ideal mixing process. The final contribution refers to the excess Helmholtz free energy ( $A^{\text{ex}}$ ), which can be modeled using a data-driven approach through ANNs and must be learned directly from mixture data. This formulation as a whole must consider thermodynamic limits when formulated and trained. For example, Eq. (6.10) must follow the *criteria of thermal* and *mechanical* stability [116], and the Gibbs-Duhem equation [131]. Finally, the formulation must simplify to a pure component if the mole fraction is  $x_i = 1$ . This condition requires the excess Helmholtz free energy to be zero for pure species.

# Bibliography

- [1] Gustavo Chaparro and Erich A. Müller. “Development of a Helmholtz free energy equation of state for fluid and solid phases via artificial neural networks”. *Communications Physics* 7.1 (2024), p. 406. DOI: 10.1038/s42005-024-01892-3.
- [2] Gustavo Chaparro and Erich A. Müller. “Simulation and Data-Driven Modeling of the Transport Properties of the Mie Fluid”. *The Journal of Physical Chemistry B* 128.2 (2024), pp. 551–566. DOI: 10.1021/acs.jpcc.3c06813.
- [3] Gustavo Chaparro and Erich A. Müller. “Development of thermodynamically consistent machine-learning equations of state: Application to the Mie fluid”. *The Journal of Chemical Physics* 158.18 (2023), p. 184505. DOI: 10.1063/5.0146634.
- [4] Robert H. Perry and Don W. Green, eds. *Perry’s chemical engineers’ handbook*. 7th Edition. McGraw-Hill, 1999. ISBN: 978-0-07-115448-2.
- [5] John Francis Richardson, John Hadlett Harker, and John Rayner Backhurst. *Chemical engineering: particle technology and separation processes*. 5th Edition. Butterworth-Heinemann, 2002. ISBN: 978-0-7506-4445-7.
- [6] Phillip C. Wankat. *Separation process engineering: includes mass transfer analysis*. 3rd Edition. Prentice Hall, 2012. ISBN: 978-0-13-138227-5.
- [7] Christos N. Markides, André Bardow, Michel De Paepe, Carlo De Servi, Joachim Groß, Andrew J. Haslam, Steven Lecompte, Athanasios I. Papadopoulos, Oyenyi A. Oyewunmi, Panos Seferlis, Johannes Schilling, Patrick Linke, Hua Tian, and Gequn Shu. “Working fluid and system optimisation of organic Rankine cycles via computer-aided molecular design: A review”. *Progress in Energy and Combustion Science* 107 (2025), p. 101201. DOI: 10.1016/j.pecs.2024.101201.
- [8] Hussam Jouhara, Alina Żabnieńska-Góra, Bertrand Delpech, Valentina Olabi, Tala El Samad, and Abdulnaser Sayma. “High-temperature heat pumps: Fundamentals, modelling approaches and applications”. *Energy* 303 (2024), p. 131882. DOI: 10.1016/j.energy.2024.131882.
- [9] Signe K. Ratkje, Dick Bedeaux, Eivind Johannessen, and Joachim Gross. *Non-equilibrium thermodynamics for engineers*. 2nd Edition. World Scientific, 2017. ISBN: 978-981-320-030-2.
- [10] Robert L. Mott and Joseph A. Untener. *Applied fluid mechanics*. 7th Edition. Pearson, 2016. ISBN: 978-0-13-255892-1.
- [11] Frank P. Incropera, David P. DeWitt, Theodore L. Bergman, and Adrienne S. Lavine, eds. *Principles of heat and mass transfer*. 7th Edition. Wiley, 2013. ISBN: 978-0-470-50197-9.
- [12] Peter Linstrom. *NIST Chemistry WebBook, NIST Standard Reference Database* 69. 1997. DOI: 10.18434/T4D303. URL: <http://webbook.nist.gov/chemistry/> (visited on Dec. 24, 2024).

- [13] W. Vincent Wilding, Richard L Rowley, and John L Oscarson. "DIPPR® Project 801 evaluated process design data". *Fluid Phase Equilibria* 150-151 (1998), pp. 413–420. DOI: 10.1016/S0378-3812(98)00341-0.
- [14] Georgios M. Kontogeorgis, Ralf Dohrn, Ioannis G. Economou, Jean-Charles De Hemptinne, Antoon Ten Kate, Susanna Kuitunen, Miranda Mooijer, Ljudmila Fele Žilnik, and Velisa Vesovic. "Industrial Requirements for Thermodynamic and Transport Properties: 2020". *Industrial & Engineering Chemistry Research* 60.13 (2021), pp. 4987–5013. DOI: 10.1021/acs.iecr.0c05356.
- [15] Marcus J. Tillotson, Nikolaos I. Diamantonis, Corneliu Buda, Leslie W. Bolton, and Erich A. Müller. "Molecular modelling of the thermophysical properties of fluids: expectations, limitations, gaps and opportunities". *Physical Chemistry Chemical Physics* 25.18 (2023), pp. 12607–12628. DOI: 10.1039/D2CP05423J.
- [16] Eric W. Lemmon and Roland Span. "Chapter 12: Multi-parameter Equations of State for Pure Fluids and Mixtures". *Applied Thermodynamics of Fluids*. Ed. by Anthony R. Goodwin, Jan Sengers, and Cor J Peters. The Royal Society of Chemistry, 2010, pp. 394–432. ISBN: 978-1-84755-806-0. DOI: 10.1039/9781849730983-00394.
- [17] Hans-Joachim Kretzschmar and Wolfgang Wagner. *International Steam Tables: Properties of Water and Steam based on the Industrial Formulation IAPWS-IF97*. Springer Berlin Heidelberg, 2019. ISBN: 978-3-662-53217-1. DOI: 10.1007/978-3-662-53219-5.
- [18] O. Kunz and W. Wagner. "The GERG-2008 Wide-Range Equation of State for Natural Gases and Other Mixtures: An Expansion of GERG-2004". *Journal of Chemical & Engineering Data* 57.11 (2012), pp. 3032–3091. DOI: 10.1021/jc300655b.
- [19] Eric W. Lemmon, Mark O. McLinden, and Wolfgang Wagner. "Thermodynamic Properties of Propane. III. A Reference Equation of State for Temperatures from the Melting Line to 650 K and Pressures up to 1000 MPa". *Journal of Chemical & Engineering Data* 54.12 (2009), pp. 3141–3180. DOI: 10.1021/jc900217v.
- [20] Henri Renon and J. M. Prausnitz. "Local compositions in thermodynamic excess functions for liquid mixtures". *AIChE Journal* 14.1 (1968), pp. 135–144. DOI: 10.1002/aic.690140124.
- [21] Grant M. Wilson. "Vapor-Liquid Equilibrium. XI. A New Expression for the Excess Free Energy of Mixing". *Journal of the American Chemical Society* 86.2 (1964), pp. 127–130. DOI: 10.1021/ja01056a002.
- [22] Denis S. Abrams and John M. Prausnitz. "Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems". *AIChE Journal* 21.1 (1975), pp. 116–128. DOI: 10.1002/aic.690210115.
- [23] Ulrich Weidlich and Juergen Gmehling. "A modified UNIFAC model. 1. Prediction of VLE, hE, and  $\gamma_{\infty}$ ". *Industrial & Engineering Chemistry Research* 26.7 (1987), pp. 1372–1381. DOI: 10.1021/ie00067a018.
- [24] A. Allal, C. Boned, and A. Baylaucq. "Free-volume viscosity model for fluids in the dense and gaseous states". *Physical Review E* 64.1 (2001), p. 011203. DOI: 10.1103/PhysRevE.64.011203.
- [25] Sergio E. Quiñones-Cisneros, Claus K. Zéberg-Mikkelsen, and Erling H. Stenby. "The friction theory (f-theory) for viscosity modeling". *Fluid Phase Equilibria* 169.2 (2000), pp. 249–276. DOI: 10.1016/S0378-3812(00)00310-1.

- [26] Jeppe C. Dyre. "Perspective: Excess-entropy scaling". *The Journal of Chemical Physics* 149.21 (2018), p. 210901. DOI: 10.1063/1.5055064.
- [27] Thomas M. Reed and Keith E. Gubbins. *Applied statistical mechanics: thermodynamic and transport properties of fluids*. Chemical Engineering Series. McGraw-Hill, 1973. ISBN: 978-0-07-051495-9.
- [28] Johannes Diderik van der Waals. "Over de continuïteit van den gas- en vloeistofoestand". PhD thesis. University of Leiden, 1873.
- [29] Otto. Redlich and J. N. S. Kwong. "On the Thermodynamics of Solutions. V. An Equation of State. Fugacities of Gaseous Solutions." *Chemical Reviews* 44.1 (1949), pp. 233–244. DOI: 10.1021/cr60137a013.
- [30] Ding-Yu Peng and Donald B. Robinson. "A New Two-Constant Equation of State". *Industrial & Engineering Chemistry Fundamentals* 15.1 (1976), pp. 59–64. DOI: 10.1021/i160057a011.
- [31] Georgios M. Kontogeorgis, Xiaodong Liang, Alay Arya, and Ioannis Tsivintzelis. "Equations of state in three centuries. Are we closer to arriving to a single model for all applications?" *Chemical Engineering Science: X* 7 (2020), p. 100060. DOI: 10.1016/j.cesx.2020.100060.
- [32] André Péneloux, Evelyne Rauzy, and Richard Fréze. "A consistent correction for Redlich-Kwong-Soave volumes". *Fluid Phase Equilibria* 8.1 (1982), pp. 7–23. DOI: 10.1016/0378-3812(82)80002-2.
- [33] Ilya Polishuk, Ilya Bourgin, and Naor Cohen. "IMPLEMENTATION OF MATHEMATICA FOR DEVELOPMENT AND APPLICATION OF EOS MODELS. III: CALCULATION OF HEAT CAPACITIES, ENTHALPY AND ENTROPY OF EVAPORATION, AND SPEED OF SOUND". *Chemical Engineering Communications* 196.11 (2009), pp. 1291–1299. DOI: 10.1080/00986440902900048.
- [34] Yohann Le Guennec, Romain Privat, and Jean-Noël Jaubert. "Development of the translated-consistent tc-PR and tc-RK cubic equations of state for a safe and accurate prediction of volumetric, energetic and saturation properties of pure compounds in the sub- and super-critical domains". *Fluid Phase Equilibria* 429 (2016), pp. 301–312. DOI: 10.1016/j.fluid.2016.09.003.
- [35] Walter G. Chapman, Keith E. Gubbins, George Jackson, and Maciej Radosz. "New reference equation of state for associating liquids". *Industrial & Engineering Chemistry Research* 29.8 (1990), pp. 1709–1721. DOI: 10.1021/ie00104a021.
- [36] Joachim Gross and Gabriele Sadowski. "Perturbed-Chain SAFT: An Equation of State Based on a Perturbation Theory for Chain Molecules". *Industrial & Engineering Chemistry Research* 40.4 (2001), pp. 1244–1260. DOI: 10.1021/ie0003887.
- [37] Thomas Lafitte, Anastasia Apostolakou, Carlos Avendaño, Amparo Galindo, Claire S. Adjiman, Erich A. Müller, and George Jackson. "Accurate statistical associating fluid theory for chain molecules formed from Mie segments". *The Journal of Chemical Physics* 139.15 (2013), p. 154504. DOI: 10.1063/1.4819786.
- [38] Joachim Gross and Jadran Vrabec. "An equation-of-state contribution for polar components: Dipolar molecules". *AIChE Journal* 52.3 (2006), pp. 1194–1204. DOI: 10.1002/aic.10683.
- [39] Joachim Gross. "An equation-of-state contribution for polar components: Quadrupolar molecules". *AIChE Journal* 51.9 (2005), pp. 2556–2568. DOI: 10.1002/aic.10502.

- [40] Spiros Kournopoulos, Mirella Simões Santos, Srikanth Ravipati, Andrew J. Haslam, George Jackson, Ioannis G. Economou, and Amparo Galindo. "The Contribution of the Ion–Ion and Ion–Solvent Interactions in a Molecular Thermodynamic Treatment of Electrolyte Solutions". *The Journal of Physical Chemistry B* 126.47 (2022), pp. 9821–9839. DOI: 10.1021/acs.jpcc.2c03915.
- [41] Eric Hendriks, Georgios M. Kontogeorgis, Ralf Dohrn, Jean-Charles De Hemptinne, Ioannis G. Economou, Ljudmila Fele Žilnik, and Velisa Vesovic. "Industrial Requirements for Thermodynamics and Transport Properties". *Industrial & Engineering Chemistry Research* 49.22 (2010), pp. 11131–11141. DOI: 10.1021/ie101231b.
- [42] Øivind Wilhelmsen, Ailo Aasen, Geir Skaugen, Peder Aursand, Anders Austegard, Eskil Aursand, Magnus Aa. Gjennestad, Halvor Lund, Gaute Linga, and Morten Hammer. "Thermodynamic Modeling with Equations of State: Present Challenges with Established Methods". *Industrial & Engineering Chemistry Research* 56.13 (2017), pp. 3503–3515. DOI: 10.1021/acs.iecr.7b00317.
- [43] Andrés Mejía, Erich A. Müller, and Gustavo Chaparro Maldonado. "SGTPy: A Python Code for Calculating the Interfacial Properties of Fluids Based on the Square Gradient Theory Using the SAFT-VR Mie Equation of State". *Journal of Chemical Information and Modeling* 61.3 (2021), pp. 1244–1250. DOI: 10.1021/acs.jcim.0c01324.
- [44] Pierre J. Walker, Hon-Wa Yew, and Andrés Riedemann. "Clapeyron.jl: An Extensible, Open-Source Fluid Thermodynamics Toolkit". *Industrial & Engineering Chemistry Research* 61.20 (2022), pp. 7130–7153. DOI: 10.1021/acs.iecr.2c00326.
- [45] Ian H. Bell, Ulrich K. Deiters, and Allan M. M. Leal. "Implementing an Equation of State without Derivatives: teqp". *Industrial & Engineering Chemistry Research* 61.17 (2022), pp. 6010–6027. DOI: 10.1021/acs.iecr.2c00237.
- [46] Philipp Rehner, Gernot Bauer, and Joachim Gross. "FeO<sub>s</sub>: An Open-Source Framework for Equations of State and Classical Density Functional Theory". *Industrial & Engineering Chemistry Research* 62.12 (2023), pp. 5347–5357. DOI: 10.1021/acs.iecr.2c04561.
- [47] L. Fokin, V. Popov, and A. Kalashnikov. "Analytical presentation of the collision integrals for the (m-6) Lennard-Jones potential in the EPIDIF data base". *High Temperature* 37.1 (1999), pp. 49–55.
- [48] Vegard G. Jervell and Øivind Wilhelmsen. "Revised Enskog theory for Mie fluids: Prediction of diffusion coefficients, thermal diffusion coefficients, viscosities, and thermal conductivities". *The Journal of Chemical Physics* 158.22 (2023), p. 224101. DOI: 10.1063/5.0149865.
- [49] Fabian Jirasek and Hans Hasse. "Perspective: Machine Learning of Thermophysical Properties". *Fluid Phase Equilibria* 549 (2021), p. 113206. DOI: 10.1016/j.fluid.2021.113206.
- [50] Fabian Jirasek and Hans Hasse. "Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures". *Annual Review of Chemical and Biomolecular Engineering* 14.1 (2023), pp. 31–51. DOI: 10.1146/annurev-chembioeng-092220-025342.
- [51] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.
- [52] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". *The Bulletin of Mathematical Biophysics* 5.4 (1943), pp. 115–133. DOI: 10.1007/BF02478259.

- [53] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". *Neural Networks* 2.5 (1989), pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.
- [54] Jiaqi Ding, Nan Xu, Manh Tien Nguyen, Qi Qiao, Yao Shi, Yi He, and Qing Shao. "Machine learning for molecular thermodynamics". *Chinese Journal of Chemical Engineering* 31 (2021), pp. 227–239. DOI: 10.1016/j.cjche.2020.10.044.
- [55] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. "Machine Learning for Molecular Simulation". *Annual Review of Physical Chemistry* 71.1 (2020), pp. 361–390. DOI: 10.1146/annurev-physchem-042018-052331.
- [56] Yuanqi Du, Arian R. Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L. Blundell. "Machine learning-aided generative molecular design". *Nature Machine Intelligence* 6.6 (2024), pp. 589–604. DOI: 10.1038/s42256-024-00843-5.
- [57] Lisa Joss and Erich A. Müller. "Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB". *Journal of Chemical Education* 96.4 (2019), pp. 697–703. DOI: 10.1021/acs.jchemed.8b00692.
- [58] Kezheng Zhu and Erich A. Müller. "Generating a Machine-Learned Equation of State for Fluid Properties". *The Journal of Physical Chemistry B* 124.39 (2020), pp. 8628–8639. DOI: 10.1021/acs.jpcb.0c05806.
- [59] Zheng Gong, Yanze Wu, Liang Wu, and Huai Sun. "Predicting Thermodynamic Properties of Alkanes by High-Throughput Force Field Simulation and Machine Learning". *Journal of Chemical Information and Modeling* 58.12 (2018), pp. 2502–2516. DOI: 10.1021/acs.jcim.8b00407.
- [60] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36. DOI: 10.1021/ci00057a005.
- [61] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". *Machine Learning: Science and Technology* 1.4 (2020), p. 045024. DOI: 10.1088/2632-2153/aba947.
- [62] Abdulelah S. Alshehri, Anjan K. Tula, Fengqi You, and Rafiqul Gani. "Next generation pure component property estimation models: With and without machine learning techniques". *AIChE Journal* 68.6 (2021), e17469. DOI: 10.1002/aic.17469.
- [63] Jan G. Rittig, Qinghe Gao, Manuel Dahmen, Alexander Mitsos, and Artur M. Schweidtmann. "Graph Neural Networks for the Prediction of Molecular Structure–Property Relationships". *Machine Learning and Hybrid Modelling for Reaction Engineering*. Ed. by Dongda Zhang and Ehecatl Antonio Del Río Chanona. Royal Society of Chemistry, 2023, pp. 159–181. ISBN: 978-1-83916-563-4. DOI: 10.1039/BK9781837670178-00159.
- [64] K.G. Joback and R.C. Reid. "ESTIMATION OF PURE-COMPONENT PROPERTIES FROM GROUP-CONTRIBUTIONS". *Chemical Engineering Communications* 57.1 (1987), pp. 233–243. DOI: 10.1080/00986448708960487.



- [65] Vasileios Papaioannou, Thomas Lafitte, Carlos Avendaño, Claire S. Adjiman, George Jackson, Erich A. Müller, and Amparo Galindo. "Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments". *The Journal of Chemical Physics* 140.5 (2014), p. 054107. DOI: 10.1063/1.4851455.
- [66] Michael P. Allen and Dominic J. Tildesley. *Computer simulation of liquids*. 2nd Edition. Oxford university press, 2017. ISBN: 978-0-19-880319-5.
- [67] Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. "Machine Learning Force Fields". *Chemical Reviews* 121.16 (2021), pp. 10142–10186. DOI: 10.1021/acs.chemrev.0c01111.
- [68] Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, and Angelos Michaelides. "Machine Learning Potential for Hexagonal Boron Nitride Applied to Thermally and Mechanically Induced Rippling". *The Journal of Physical Chemistry C* 124.40 (2020), pp. 22278–22290. DOI: 10.1021/acs.jpcc.0c05831.
- [69] Christoph Schran, Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, Ondrej Marsalek, and Angelos Michaelides. "Machine learning potentials for complex aqueous systems made simple". *Proceedings of the National Academy of Sciences* 118.38 (2021), e2110077118. DOI: 10.1073/pnas.2110077118.
- [70] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. 2nd Edition. Computational science series 1. Academic Press, 2002. ISBN: 978-0-12-267351-1.
- [71] Georgios M. Kontogeorgis and Georgios K. Folas. *Thermodynamic Models for Industrial Applications: From Classical and Advanced Mixing Rules to Association Theories*. 1st ed. Wiley, 2010. ISBN: 978-0-470-69726-9. DOI: 10.1002/9780470747537.
- [72] Fabian Jirasek, Rodrigo A. S. Alves, Julie Damay, Robert A. Vandermeulen, Robert Bamler, Michael Bortz, Stephan Mandt, Marius Kloft, and Hans Hasse. "Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion". *The Journal of Physical Chemistry Letters* 11.3 (2020), pp. 981–985. DOI: 10.1021/acs.jpclett.9b03657.
- [73] Julie Damay, Fabian Jirasek, Marius Kloft, Michael Bortz, and Hans Hasse. "Predicting Activity Coefficients at Infinite Dilution for Varying Temperatures by Matrix Completion". *Industrial & Engineering Chemistry Research* 60.40 (2021), pp. 14564–14578. DOI: 10.1021/acs.iecr.1c02039.
- [74] Nicolas Hayer, Fabian Jirasek, and Hans Hasse. "Prediction of Henry's law constants by matrix completion". *AIChE Journal* 68.9 (2022), e17753. DOI: 10.1002/aic.17753.
- [75] Fabian Jirasek, Robert Bamler, Sophie Fellenz, Michael Bortz, Marius Kloft, Stephan Mandt, and Hans Hasse. "Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions". *Chemical Science* 13.17 (2022), pp. 4854–4862. DOI: 10.1039/D1SC07210B.
- [76] Hiroaki Matsukawa, Masayuki Kitahara, and Katsuto Otake. "Estimation of pure component parameters of PC-SAFT EoS by an artificial neural network based on a group contribution method". *Fluid Phase Equilibria* 548 (2021), p. 113179. DOI: 10.1016/j.fluid.2021.113179.
- [77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2023. DOI: 10.48550/arXiv.1706.03762.

- [78] Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Cham: Springer International Publishing, 2024. ISBN: 978-3-031-45467-7. DOI: 10.1007/978-3-031-45468-4.
- [79] OpenAI. *ChatGPT*. 2025. URL: <https://chat.openai.com/chat> (visited on Apr. 24, 2025).
- [80] Benedikt Winter, Clemens Winter, Johannes Schilling, and André Bardow. “A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing”. *Digital Discovery* 1.6 (2022), pp. 859–869. DOI: 10.1039/D2DD00058J.
- [81] Benedikt Winter, Clemens Winter, Timm Esper, Johannes Schilling, and André Bardow. “SPT-NRTL: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients”. *Fluid Phase Equilibria* 568 (2023), p. 113731. DOI: 10.1016/j.fluid.2023.113731.
- [82] Benedikt Winter, Philipp Rehner, Timm Esper, Johannes Schilling, and André Bardow. “Understanding the language of molecules: predicting pure component parameters for the PC-SAFT equation of state from SMILES”. *Digital Discovery* (2025). DOI: 10.1039/D4DD00077C.
- [83] Galen T. Craven, Nicholas Lubbers, Kipton Barros, and Sergei Tretiak. “Machine learning approaches for structural and thermodynamic properties of a Lennard-Jones fluid”. *The Journal of Chemical Physics* 153.10 (2020), p. 104502. DOI: 10.1063/5.0017894.
- [84] Jan G. Rittig, Kobi C. Felton, Alexei A. Lapkin, and Alexander Mitsos. “Gibbs–Duhem-informed neural networks for binary activity coefficient prediction”. *Digital Discovery* 2.6 (2023), pp. 1752–1767. DOI: 10.1039/D3DD00103B.
- [85] Thomas Specht, Mayank Nagda, Sophie Fellenz, Stephan Mandt, Hans Hasse, and Fabian Jirasek. “HANNA: Hard-constraint Neural Network for Consistent Activity Coefficient Prediction”. *Chemical Science* 15.47 (2024), pp. 19777–19786. DOI: 10.1039/D4SC05115G.
- [86] M. Raissi, P. Perdikaris, and G.E. Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. *Journal of Computational Physics* 378 (2019), pp. 686–707. DOI: 10.1016/j.jcp.2018.10.045.
- [87] David Rosenberger, Kipton Barros, Timothy C. Germann, and Nicholas Lubbers. “Machine learning of consistent thermodynamic models using automatic differentiation”. *Physical Review E* 105.4 (2022), p. 045301. DOI: 10.1103/PhysRevE.105.045301.
- [88] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0.
- [89] U.K. Deiters and K.M. De Reuck. “Guidelines for publication of equations of state—I. Pure fluids”. *Chemical Engineering Journal* 69.1 (1998), pp. 69–81. DOI: 10.1016/S1385-8947(97)00070-3.
- [90] Gustav Mie. “Zur kinetischen Theorie der einatomigen Körper”. *Annalen der Physik* 316.8 (1903), pp. 657–697. DOI: 10.1002/andp.19033160802.
- [91] Johann Fischer and Martin Wendland. “On the history of key empirical intermolecular potentials”. *Fluid Phase Equilibria* 573 (2023), p. 113876. DOI: 10.1016/j.fluid.2023.113876.
- [92] N.S. Ramrattan, C. Avendaño, E.A. Müller, and A. Galindo. “A corresponding-states framework for the description of the Mie family of intermolecular potentials”. *Molecular Physics* 113.9 (2015), pp. 932–947. DOI: 10.1080/00268976.2015.1025112.

- [93] F. London. "The general theory of molecular forces". *Transactions of the Faraday Society* 33 (1937), 8b. DOI: 10.1039/tf937330008b.
- [94] Carmelo Herdes, Tim S. Totton, and Erich A. Müller. "Coarse grained force field for the molecular simulation of natural gases and condensates". *Fluid Phase Equilibria* 406 (2015), pp. 91–100. DOI: 10.1016/j.fluid.2015.07.014.
- [95] Ailo Aasen, Morten Hammer, Åsmund Ervik, Erich A. Müller, and Øivind Wilhelmsen. "Equation of state and force fields for Feynman–Hibbs-corrected Mie fluids. I. Application to pure helium, neon, hydrogen, and deuterium". *The Journal of Chemical Physics* 151.6 (2019), p. 064508. DOI: 10.1063/1.5111364.
- [96] Andrew J. Haslam, Alfonso González-Pérez, Silvia Di Lecce, Siti H. Khalit, Felipe A. Perdomo, Spiros Kournopoulos, Maximilian Kohns, Tom Lindeboom, Malak Wehbe, Sara Febra, George Jackson, Claire S. Adjiman, and Amparo Galindo. "Expanding the Applications of the SAFT- $\gamma$  Mie Group-Contribution Equation of State: Prediction of Thermodynamic Properties and Phase Behavior of Mixtures". *Journal of Chemical & Engineering Data* 65.12 (2020), pp. 5862–5890. DOI: 10.1021/acs.jced.0c00746.
- [97] Åsmund Ervik, Andrés Mejía, and Erich A. Müller. "Bottled SAFT: A Web App Providing SAFT- $\gamma$  Mie Force Field Parameters for Thousands of Molecular Fluids". *Journal of Chemical Information and Modeling* 56.9 (2016), pp. 1609–1614. DOI: 10.1021/acs.jcim.6b00149.
- [98] Andrés Mejía, Marcela Cartes, Gustavo Chaparro, Esther Feria, Felipe J. Blas, José Manuel Míguez, Jesús Algaba, and Erich A. Müller. "Phase equilibria and interfacial properties of selected methane + n-alkane binary mixtures". *Journal of Molecular Liquids* 341 (2021), p. 116918. DOI: 10.1016/j.molliq.2021.116918.
- [99] Marcela Cartes, Gustavo Chaparro, and Andrés Mejía. "Measurements and modeling of the phase equilibria for the n-hexane + ethanol + dibutyl ether ternary mixture and its dibutyl ether sub-binary mixtures at 94 kPa". *Fuel* 297 (2021), p. 120361. DOI: 10.1016/j.fuel.2021.120361.
- [100] Andrés Mejía, Marcela Cartes, Alejandra Velásquez, Gustavo Chaparro, and Vilma Sanhueza. "Vapor-liquid phase equilibria, liquid densities, liquid viscosities and surface tensions for the ternary n-hexane + cyclopentyl methyl ether + 1-butanol mixture". *Fluid Phase Equilibria* 558 (2022), p. 113444. DOI: 10.1016/j.fluid.2022.113444.
- [101] Sara A. Febra, Thomas Bernet, Corin Mack, John McGinty, Iyke I. Onyemelukwe, Stephanie J. Urwin, Jan Sefcik, Joop H. Ter Horst, Claire S. Adjiman, George Jackson, and Amparo Galindo. "Extending the SAFT-  $\gamma$  Mie approach to model benzoic acid, diphenylamine, and mefenamic acid: Solubility prediction and experimental measurement". *Fluid Phase Equilibria* 540 (2021), p. 113002. DOI: 10.1016/j.fluid.2021.113002.
- [102] Malak Wehbe, Andrew J. Haslam, George Jackson, and Amparo Galindo. "Phase behaviour and pH-solubility profile prediction of aqueous buffered solutions of ibuprofen and ketoprofen". *Fluid Phase Equilibria* 560 (2022), p. 113504. DOI: 10.1016/j.fluid.2022.113504.
- [103] Ailo Aasen, Morten Hammer, Erich A. Müller, and Øivind Wilhelmsen. "Equation of state and force fields for Feynman–Hibbs-corrected Mie fluids. II. Application to mixtures of helium, neon, hydrogen, and deuterium". *The Journal of Chemical Physics* 152.7 (2020), p. 074507. DOI: 10.1063/1.5136079.

- [104] Guadalupe Jiménez-Serratos, Carmelo Herdes, Andrew J. Haslam, George Jackson, and Erich A. Müller. "Group Contribution Coarse-Grained Molecular Simulations of Polystyrene Melts and Polystyrene Solutions in Alkanes Using the SAFT- $\gamma$  Force Field". *Macromolecules* 50.12 (2017), pp. 4840–4853. DOI: 10.1021/acs.macromol.6b02072.
- [105] Maziar Fayaz-Torshizi and Erich A. Müller. "Coarse-Grained Molecular Simulation of Polymers Supported by the Use of the SAFT- $\gamma$  Mie Equation of State". *Macromolecular Theory and Simulations* 31.1 (2022), p. 2100031. DOI: 10.1002/mats.202100031.
- [106] Maziar Fayaz-Torshizi and Erich A. Müller. "Coarse-grained molecular dynamics study of the self-assembly of polyphilic bolaamphiphiles using the SAFT- $\gamma$  Mie force field". *Molecular Systems Design & Engineering* 6.8 (2021), pp. 594–608. DOI: 10.1039/D1ME00021G.
- [107] Panagiotis E. Theodorakis, Edward R. Smith, and Erich A. Müller. "Spreading of aqueous droplets with common and superspreading surfactants. A molecular dynamics study". *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 581 (2019), p. 123810. DOI: 10.1016/j.colsurfa.2019.123810.
- [108] Guadalupe Jiménez-Serratos, Tim S. Totton, George Jackson, and Erich A. Müller. "Aggregation Behavior of Model Asphaltenes Revealed from Large-Scale Coarse-Grained Molecular Simulations". *The Journal of Physical Chemistry B* 123.10 (2019), pp. 2380–2396. DOI: 10.1021/acs.jpcc.8b12295.
- [109] José Matías Garrido, Andrés Mejía, Manuel M. Piñeiro, Felipe J. Blas, and Erich A. Müller. "Interfacial tensions of industrial fluids from a molecular-based square gradient theory". *AIChE Journal* 62.5 (2016), pp. 1781–1794. DOI: 10.1002/aic.15190.
- [110] Frederike Jaeger, Omar K. Matar, and Erich A. Müller. "Bulk viscosity of molecular fluids". *The Journal of Chemical Physics* 148.17 (2018), p. 174504. DOI: 10.1063/1.5022752.
- [111] Hai Hoang, Stéphanie Delage-Santacreu, and Guillaume Galliero. "Simultaneous Description of Equilibrium, Interfacial, and Transport Properties of Fluids Using a Mie Chain Coarse-Grained Force Field". *Industrial & Engineering Chemistry Research* 56.32 (2017), pp. 9213–9226. DOI: 10.1021/acs.iecr.7b01397.
- [112] Richard A. Messerly, Michelle C. Anderson, S. Mostafa Razavi, and J. Richard Elliott. "Improvements and limitations of Mie  $\lambda$ -6 potential for prediction of saturated and compressed liquid viscosity". *Fluid Phase Equilibria* 483 (2019), pp. 101–115. DOI: 10.1016/j.fluid.2018.11.002.
- [113] Anja Reimer, Thijs Van Westen, and Joachim Gross. "Physically based equation of state for Mie  $v$ -6 fluids". *The Journal of Chemical Physics* 158.16 (2023), p. 164506. DOI: 10.1063/5.0141856.
- [114] Thijs Van Westen and Joachim Gross. "Accurate thermodynamics of simple fluids and chain fluids based on first-order perturbation theory and second virial coefficients:  $uv$ -theory". *The Journal of Chemical Physics* 155.24 (2021), p. 244501. DOI: 10.1063/5.0073572.
- [115] Sven Pohl, Robin Fingerhut, Monika Thol, Jadran Vrabec, and Roland Span. "Equation of state for the Mie ( $\lambda_r$ ,6) fluid with a repulsive exponent from 11 to 13". *The Journal of Chemical Physics* 158.8 (2023), p. 084506. DOI: 10.1063/5.0133412.
- [116] Michael Modell and Robert C. Reid. *Thermodynamics and its applications*. 2nd Edition. Prentice-Hall international series in the physical and chemical engineering science. Prentice-Hall, 1983. ISBN: 978-0-13-915017-3.

- [117] Sydney Chapman and Thomas G. Cowling. *The Mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction and diffusion in gases*. 3rd Edition. Cambridge Univ. Pr, 1991. ISBN: 978-0-521-40844-8.
- [118] Ingo Müller. *A History of Thermodynamics: The Doctrine of Energy and Entropy*. 1st Edition. Springer Berlin Heidelberg, 2007. ISBN: 978-3-540-46226-2. DOI: 10.1007/978-3-540-46227-9.
- [119] Norman F. Carnahan and Kenneth E. Starling. "Equation of State for Nonattracting Rigid Spheres". *The Journal of Chemical Physics* 51.2 (1969), pp. 635–636. DOI: 10.1063/1.1672048.
- [120] Ya Song Wei and Richard J. Sadus. "Equations of state for the calculation of fluid-phase equilibria". *AIChE Journal* 46.1 (2000), pp. 169–196. DOI: 10.1002/aic.690460119.
- [121] Jean-Pierre Hansen and Ian R. McDonald. *Theory of Simple Liquids*. Elsevier, 2013. ISBN: 978-0-12-387032-2. DOI: 10.1016/C2010-0-66723-X.
- [122] Keith E. Gubbins. "Perturbation theories of the thermodynamics of polar and associating liquids: A historical perspective". *Fluid Phase Equilibria* 416 (2016), pp. 3–17. DOI: 10.1016/j.fluid.2015.12.043.
- [123] Erich A. Müller and George Jackson. "Force-Field Parameters from the SAFT- $\gamma$  Equation of State for Use in Coarse-Grained Molecular Simulations". *Annual Review of Chemical and Biomolecular Engineering* 5.1 (2014), pp. 405–427. DOI: 10.1146/annurev-chembioeng-061312-103314.
- [124] J. D. Van Der Waals. "Thermodynamische Theorie der Kapillarität unter Voraussetzung stetiger Dichteänderung". *Zeitschrift für Physikalische Chemie* 13U.1 (1894), pp. 657–725. DOI: 10.1515/zpch-1894-1338.
- [125] Herbert B. Callen. *Thermodynamics and an introduction to thermostatistics*. 2nd Edition. Wiley, 1985. ISBN: 978-0-471-86256-7.
- [126] Andrés Mejía, Carmelo Herdes, and Erich A. Müller. "Force Fields for Coarse-Grained Molecular Simulations from a Corresponding States Correlation". *Industrial & Engineering Chemistry Research* 53.10 (2014), pp. 4131–4141. DOI: 10.1021/ie404247e.
- [127] Carmelo Herdes, Åsmund Ervik, Andrés Mejía, and Erich A. Müller. "Prediction of the water/oil interfacial tension from molecular simulations using the coarse-grained SAFT- $\gamma$  Mie force field". *Fluid Phase Equilibria* 476 (2018), pp. 9–15. DOI: 10.1016/j.fluid.2017.06.016.
- [128] Gerard Alonso, Gustavo Chaparro, Marcela Cartes, Erich A. Müller, and Andrés Mejía. "Probing the Interfacial Behavior of Type IIIa Binary Mixtures Along the Three-Phase Line Employing Molecular Thermodynamics". *Molecules* 25.7 (2020), p. 1499. DOI: 10.3390/molecules25071499.
- [129] Amulya K. Pervaje, Christopher C. Walker, and Erik E. Santiso. "Molecular simulation of polymers with a SAFT- $\gamma$  Mie approach". *Molecular Simulation* 45.14 (2019), pp. 1223–1241. DOI: 10.1080/08927022.2019.1645331.
- [130] Stanley H. Huang and Maciej Radosz. "Equation of state for small, large, polydisperse, and associating molecules". *Industrial & Engineering Chemistry Research* 29.11 (1990), pp. 2284–2294. DOI: 10.1021/ie00107a014.
- [131] Michael L. Michelsen and Jørgen M. Møllerup. *Thermodynamic models: fundamentals & computational aspects*. 2nd Edition. Tie-Line Publications, 2007. ISBN: 978-87-989961-3-2.

- [132] Erich A. Müller and Keith E. Gubbins. "An Equation of State for Water from a Simplified Intermolecular Potential". *Industrial & Engineering Chemistry Research* 34.10 (1995), pp. 3662–3673. DOI: 10.1021/ie00037a055.
- [133] A. J. Masters. "Virial expansions". *Journal of Physics: Condensed Matter* 20.28 (2008), p. 283102. DOI: 10.1088/0953-8984/20/28/283102.
- [134] Hugo Segura and Jaime Wisniak. "Calculation of pure saturation properties using cubic equations of state". *Computers & Chemical Engineering* 21.12 (1997), pp. 1339–1347. DOI: 10.1016/S0098-1354(97)00016-1.
- [135] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2nd Edition. O'Reilly, 2019. ISBN: 978-1-4920-3264-9.
- [136] Monika Thol, Gabor Rutkai, Andreas Köster, Rolf Lustig, Roland Span, and Jadran Vrabec. "Equation of State for the Lennard-Jones Fluid". *Journal of Physical and Chemical Reference Data* 45.2 (2016), p. 023101. DOI: 10.1063/1.4945000.
- [137] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. Version Number: 2. 2016. DOI: 10.48550/ARXIV.1603.04467.
- [138] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Nécule, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. 2018. URL: <http://github.com/jax-ml/jax>.
- [139] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. URL: <http://arxiv.org/abs/1412.6980>.
- [140] Li Yang and Abdallah Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice". *Neurocomputing* 415 (2020), pp. 295–316. DOI: 10.1016/j.neucom.2020.07.061.
- [141] D. B. Newell, F. Cabiati, J. Fischer, K. Fujii, S. G. Karshenboim, H. S. Margolis, E. De Mirandés, P. J. Mohr, F. Nez, K. Pachucki, T. J. Quinn, B. N. Taylor, M. Wang, B. M. Wood, and Z. Zhang. "The CODATA 2017 values of  $h$ ,  $e$ ,  $k$ , and  $N_A$  for the revision of the SI". *Metrologia* 55.1 (2018), pp. L13–L16. DOI: 10.1088/1681-7575/aa950a.
- [142] Artit Hutem and Sutee Boonchui. "Numerical evaluation of second and third virial coefficients of some inert gases via classical cluster expansion". *Journal of Mathematical Chemistry* 50.5 (2012), pp. 1262–1276. DOI: 10.1007/s10910-011-9966-5.
- [143] Andrew J. Schultz. *Etomica: B3 via fast Fourier transforms*. 2005. URL: <https://www.etomica.org/apps/virial/b3> (visited on Oct. 31, 2024).

- [144] N. A. R. Hussein and S. M. Ahmed. "Virial coefficients for the square-well potential". *Journal of Physics A: Mathematical and General* 24.1 (1991), pp. 289–305. DOI: 10.1088/0305-4470/24/1/035.
- [145] Richard J. Sadus. "Second virial coefficient properties of the  $n - m$  Lennard-Jones/Mie potential". *The Journal of Chemical Physics* 149.7 (2018), p. 074504. DOI: 10.1063/1.5041320.
- [146] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nature Methods* 17.3 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- [147] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. In 'T Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. "LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales". *Computer Physics Communications* 271 (2022), p. 108171. DOI: 10.1016/j.cpc.2021.108171.
- [148] F. Martínez-Veracoechea and E.A. Müller. "Temperature-quench Molecular Dynamics Simulations for Fluid Phase Equilibria". *Molecular Simulation* 31.1 (2005), pp. 33–43. DOI: 10.1080/08927020412331298991.
- [149] Erich A. Müller, Åsmund Ervik, and Andrés Mejía. "A Guide to Computing Interfacial Properties of Fluids from Molecular Simulations [Article v1.0]". *Living Journal of Computational Molecular Science* 2.1 (2020), p. 21385. DOI: 10.33011/livecoms.2.1.21385.
- [150] William G. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". *Physical Review A* 31.3 (1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.
- [151] Sergey A. Khrapak, Manis Chaudhuri, and Gregor E. Morfill. "Freezing of Lennard-Jones-type fluids". *The Journal of Chemical Physics* 134.5 (2011), p. 054120. DOI: 10.1063/1.3552948.
- [152] Enrique De Miguel and George Jackson. "The nature of the calculation of the pressure in molecular simulations of continuous models from volume perturbations". *The Journal of Chemical Physics* 125.16 (2006), p. 164109. DOI: 10.1063/1.2363381.
- [153] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. "Constant pressure molecular dynamics algorithms". *The Journal of Chemical Physics* 101.5 (1994), pp. 4177–4189. DOI: 10.1063/1.467468.
- [154] Charles R. Harris, K. Jarrod Millman, Stéfan J. Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. Van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández Del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. "Array programming with NumPy". *Nature* 585.7825 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.
- [155] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. "LOF: identifying density-based local outliers". *ACM SIGMOD Record* 29.2 (2000), pp. 93–104. DOI: 10.1145/335191.335388.
- [156] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. "Scikit-learn: Machine Learning in Python".

- Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>.
- [157] Myeongsu Kang. “Machine Learning: Anomaly Detection”. *Prognostics and Health Management of Electronics*. Ed. by Michael G. Pecht and Myeongsu Kang. 1st Edition. Wiley, 2018, pp. 131–162. ISBN: 978-1-119-51532-6. DOI: 10.1002/9781119515326.ch6.
- [158] Thomas Cokelaer. *fitter*. 2020. URL: <https://github.com/cokelaer/fitter/> (visited on Nov. 1, 2024).
- [159] Shuhei Watanabe. *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*. 2023. URL: <http://arxiv.org/abs/2304.11127>.
- [160] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. “Optuna: A Next-generation Hyperparameter Optimization Framework”. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701.
- [161] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. *Flax: A neural network library and ecosystem for JAX*. Version 0.6.11. 2023. URL: <http://github.com/google/flax> (visited on Nov. 12, 2024).
- [162] DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. *The DeepMind JAX Ecosystem*. 2020. URL: <http://github.com/google-deepmind> (visited on Nov. 12, 2024).
- [163] Simon Stephan, Monika Thol, Jadran Vrabec, and Hans Hasse. “Thermophysical Properties of the Lennard-Jones Fluid: Database and Data Assessment”. *Journal of Chemical Information and Modeling* 59.10 (2019), pp. 4248–4265. DOI: 10.1021/acs.jcim.9b00620.
- [164] Nina Soraya Ramrattan. “Simulation and theoretical perspectives of the phase behaviour of solids, liquids and gases using the Mie family of intermolecular potentials”. PhD thesis. Imperial College London, 2013. DOI: 10.25560/18808.
- [165] Stephan Werth, Katrin Stöbener, Martin Horsch, and Hans Hasse. “Simultaneous description of bulk and interfacial properties of fluids by the Mie potential”. *Molecular Physics* 115.9 (2017), pp. 1017–1030. DOI: 10.1080/00268976.2016.1206218.
- [166] Simon Stephan and Maximilian Urschel. “Characteristic curves of the Mie fluid”. *Journal of Molecular Liquids* 383 (2023), p. 122088. DOI: 10.1016/j.molliq.2023.122088.
- [167] H. Eugene Stanley. *Introduction to phase transitions and critical phenomena*. Clarendon Press, 1971. ISBN: 978-0-19-851257-8.
- [168] E. H. Brown. “On the thermodynamic properties of fluids”. *Bulletin de l’Institut International du Froid Annexe* 1 (1960), pp. 169–178.
- [169] Simon Stephan and Ulrich K. Deiters. “Characteristic Curves of the Lennard-Jones Fluid”. *International Journal of Thermophysics* 41.10 (2020), p. 147. DOI: 10.1007/s10765-020-02721-9.



- [170] Maximilian Urschel and Simon Stephan. "Determining Brown's Characteristic Curves Using Molecular Simulation". *Journal of Chemical Theory and Computation* 19.5 (2023), pp. 1537–1552. DOI: 10.1021/acs.jctc.2c01102.
- [171] Hugo Segura, Thomas Kraska, Andrés Mejía, Jaime Wisniak, and Ilya Polishuk. "Unnoticed Pitfalls of Soave-Type Alpha Functions in Cubic Equations of State". *Industrial & Engineering Chemistry Research* 42.22 (2003), pp. 5662–5673. DOI: 10.1021/ie020828p.
- [172] Clare McCabe and Amparo Galindo. "Chapter 8: SAFT Associating Fluids and Fluid Mixtures". *Applied Thermodynamics of Fluids*. Ed. by Anthony R. Goodwin, Jan Sengers, and Cor J Peters. The Royal Society of Chemistry, 2010, pp. 215–279. ISBN: 978-1-84755-806-0. DOI: 10.1039/9781849730983-00215.
- [173] James Clerk Maxwell. *The Scientific Letters and Papers of James Clerk Maxwell: Volume 3, 1874-1879*. Ed. by Peter Michael Harman. Cambridge University Press, 2002. ISBN: 978-0-521-25627-8.
- [174] Josiah Willard Gibbs. "A method of geometrical representation of the thermodynamic properties of substances by means of surfaces". *Transactions of the Connecticut Academy II* (1873), pp. 382–404.
- [175] James Clerk Maxwell. *Theory of Heat*. 4th Edition. Longmans, Green, and co., 1875.
- [176] Johannes Diderik van der Waals. *On the Continuity of the Gaseous and Liquid States*. Ed. by John Shipley Rowlinson. Dover Publications, Inc., 2004. ISBN: 978-0-486-49593-4.
- [177] John M. Prausnitz, Rüdiger N. Lichtenthaler, and Edmundo Gomes de Azevedo. *Molecular thermodynamics of Fluid-Phase Equilibria*. 3rd Edition. Prentice Hall PTR, 1999. ISBN: 0-13-977745-8.
- [178] Mary Bradburn. "The thermodynamics of crystal lattices: III. The equation of state for a face-centred cubic lattice". *Mathematical Proceedings of the Cambridge Philosophical Society* 39.2 (1943), pp. 113–127. DOI: 10.1017/S030500410001776X.
- [179] J. P. Martin Trusler. "Equation of State for Solid Phase I of Carbon Dioxide Valid for Temperatures up to 800 K and Pressures up to 12 GPa". *Journal of Physical and Chemical Reference Data* 40.4 (2011), p. 043105. DOI: 10.1063/1.3664915.
- [180] Tage W. Maltby, Morten Hammer, and Øivind Wilhelmsen. "Equation of State for Solid Argon Valid for Temperatures up to 300 K and Pressures up to 16 GPa". *Journal of Physical and Chemical Reference Data* 53.4 (2024), p. 043102. DOI: 10.1063/5.0237497.
- [181] C. Vega, F. J. Blas, and A. Galindo. "Extending Wertheim's perturbation theory to the solid phase of Lennard-Jones chains: Determination of the global phase diagram". *The Journal of Chemical Physics* 116.17 (2002), pp. 7645–7655. DOI: 10.1063/1.1465397.
- [182] A. Yokozeki. "Analytical Equation of State for Solid-Liquid-Vapor Phases". *International Journal of Thermophysics* 24.3 (2003), pp. 589–620. DOI: 10.1023/A:1024015729095.
- [183] Chaoping Mo, Guangdong Zhang, Zhiwei Zhang, Daibo Yan, and Sen Yang. "A Modified Solid-Liquid-Gas Phase Equation of State". *ACS Omega* 7.11 (2022), pp. 9322–9332. DOI: 10.1021/acsomega.1c06142.
- [184] Ju Ho Lee and Ki-Pung Yoo. "Comments on "Analytic Equation of State for Solid-Liquid-Vapor Phases" (Int. J. Thermophys. 24, 589 (2003))". *International Journal of Thermophysics* 32.3 (2011), pp. 553–558. DOI: 10.1007/s10765-011-0943-9.

- [185] J. Frenkel. "Continuity of the Solid and the Liquid States". *Nature* 136.3431 (1935), pp. 167–168. DOI: 10.1038/136167a0.
- [186] M. Parrinello and A. Rahman. "Crystal Structure and Pair Potentials: A Molecular-Dynamics Study". *Physical Review Letters* 45.14 (1980), pp. 1196–1199. DOI: 10.1103/PhysRevLett.45.1196.
- [187] Umang Agarwal and Fernando A. Escobedo. "Mesophase behaviour of polyhedral particles". *Nature Materials* 10.3 (Mar. 2011), pp. 230–235. DOI: 10.1038/nmat2959.
- [188] Jacobo Troncoso and Claudio A. Cerdeiriña. "Ising model for the freezing transition". *Physical Review E* 109.1 (2024), p. 014123. DOI: 10.1103/PhysRevE.109.014123.
- [189] Alex Travesset. "Phase diagram of power law and Lennard-Jones systems: Crystal phases". *The Journal of Chemical Physics* 141.16 (2014), p. 164501. DOI: 10.1063/1.4898371.
- [190] Andrew J. Schultz and David A. Kofke. "Comprehensive high-precision high-accuracy equation of state and coexistence properties for classical Lennard-Jones crystals and low-temperature fluid phases". *The Journal of Chemical Physics* 149.20 (2018), p. 204508. DOI: 10.1063/1.5053714.
- [191] Peter Schwerdtfeger and Antony Burrows. "Cuboidal bcc to fcc Transformation of Lennard-Jones Phases under High Pressure Derived from Exact Lattice Summations". *The Journal of Physical Chemistry C* 126.20 (2022), pp. 8874–8882. DOI: 10.1021/acs.jpcc.2c01255.
- [192] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*. 2017. URL: <http://arxiv.org/abs/1609.04836>.
- [193] Xian-Ming Bai and Mo Li. "Differences between solid superheating and liquid supercooling". *The Journal of Chemical Physics* 123.15 (2005), p. 151102. DOI: 10.1063/1.2085147.
- [194] K. Binder. "Double-well thermodynamic potentials and spinodal curves: how real are they?". *Philosophical Magazine Letters* 87.11 (2007), pp. 799–811. DOI: 10.1080/09500830701496560.
- [195] Luis G. MacDowell, Vincent K. Shen, and Jeffrey R. Errington. "Nucleation and cavitation of spherical, cylindrical, and slablike droplets and bubbles in small systems". *The Journal of Chemical Physics* 125.3 (2006), p. 034705. DOI: 10.1063/1.2218845.
- [196] Kurt Binder, Benjamin J. Block, Peter Virnau, and Andreas Tröster. "Beyond the Van Der Waals loop: What can be learned from simulating Lennard-Jones fluids inside the region of phase coexistence". *American Journal of Physics* 80.12 (2012), pp. 1099–1109. DOI: 10.1119/1.4754020.
- [197] Pablo G. Debenedetti and Frank H. Stillinger. "Supercooled liquids and the glass transition". *Nature* 410.6825 (2001), pp. 259–267. DOI: 10.1038/35065704.
- [198] Federica Trudu, Davide Donadio, and Michele Parrinello. "Freezing of a Lennard-Jones Fluid: From Nucleation to Spinodal Regime". *Physical Review Letters* 97.10 (2006), p. 105701. DOI: 10.1103/PhysRevLett.97.105701.
- [199] Lawrence S. Bartell and David T. Wu. "Do supercooled liquids freeze by spinodal decomposition?" *The Journal of Chemical Physics* 127.17 (2007), p. 174507. DOI: 10.1063/1.2779036.
- [200] Måns Elenius and Mikhail Dzugutov. "Evidence for a liquid-solid critical point in a simple monatomic system". *The Journal of Chemical Physics* 131.10 (2009), p. 104502. DOI: 10.1063/1.3213616.

- [201] Pablo G. Debenedetti, V. S. Raghavan, and Steven S. Borick. "Spinodal curve of some supercooled liquids". *The Journal of Physical Chemistry* 95.11 (1991), pp. 4540–4551. DOI: 10.1021/j100164a066.
- [202] William G. Hoover and Francis H. Ree. "Melting Transition and Communal Entropy for Hard Spheres". *The Journal of Chemical Physics* 49.8 (1968), pp. 3609–3617. DOI: 10.1063/1.1670641.
- [203] Daan Frenkel. "Colloidal Encounters: A Matter of Attraction". *Science* 314.5800 (2006), pp. 768–769. DOI: 10.1126/science.1135544.
- [204] Alauddin Ahmed and Richard J. Sadus. "Solid-liquid equilibria and triple points of n-6 Lennard-Jones fluids". *The Journal of Chemical Physics* 131.17 (2009), p. 174504. DOI: 10.1063/1.3253686.
- [205] International Organization for Standardization. *ISO 11357-3:2018. Plastics – Differential scanning calorimetry (DSC) – Part 3: Determination of temperature and enthalpy of melting and crystallization*. Version 3. 2018. URL: <https://www.iso.org/standard/72460.html> (visited on Jan. 6, 2025).
- [206] A. Einstein. "Die Plancksche Theorie der Strahlung und die Theorie der spezifischen Wärme". *Annalen der Physik* 327.1 (1907), pp. 180–190. DOI: 10.1002/andp.19063270110.
- [207] Alexis Thérèse Petit and Pierre Louis Dulong. "Researches on some important Points of the Theory of Heat". *Annals of Philosophy* 14 (1819), pp. 189–198. URL: <https://www.biodiversitylibrary.org/item/53907> (visited on Dec. 27, 2024).
- [208] J.D. Seader, Ernest J. Henley, and D. Keith Roper. *Separation Process Principles: Chemical and Biochemical Operations*. 3rd Edition. John Wiley & Sons Inc., 2011. ISBN: 978-0-470-48183-7.
- [209] Harry Seltz. "Thermodynamics of Solid Solutions. II. Deviations from Raoult's Law". *Journal of the American Chemical Society* 57.3 (1935), pp. 391–395. DOI: 10.1021/ja01306a002.
- [210] Martin P. Lautenschlaeger and Hans Hasse. "Transport properties of the Lennard-Jones truncated and shifted fluid from non-equilibrium molecular dynamics simulations". *Fluid Phase Equilibria* 482 (2019), pp. 38–47. DOI: 10.1016/j.fluid.2018.10.019.
- [211] M. L. Huber, R. A. Perkins, A. Laesecke, D. G. Friend, J. V. Sengers, M. J. Assael, I. N. Metaxa, E. Vogel, R. Mareš, and K. Miyagawa. "New International Formulation for the Viscosity of H<sub>2</sub>O". *Journal of Physical and Chemical Reference Data* 38.2 (2009), pp. 101–125. DOI: 10.1063/1.3088050.
- [212] Arno Laesecke and Chris D. Muzny. "Reference Correlation for the Viscosity of Carbon Dioxide". *Journal of Physical and Chemical Reference Data* 46.1 (2017), p. 013107. DOI: 10.1063/1.4977429.
- [213] M. J. Assael, A. E. Kalyva, S. A. Monogenidou, M. L. Huber, R. A. Perkins, D. G. Friend, and E. F. May. "Reference Values and Reference Correlations for the Thermal Conductivity and Viscosity of Fluids". *Journal of Physical and Chemical Reference Data* 47.2 (2018), p. 021501. DOI: 10.1063/1.5036625.
- [214] F. Llovell, R. M. Marcos, and L. F. Vega. "Free-Volume Theory Coupled with Soft-SAFT for Viscosity Calculations: Comparison with Molecular Simulation and Experimental Data". *The Journal of Physical Chemistry B* 117.27 (2013), pp. 8159–8171. DOI: 10.1021/jp401307t.
- [215] Yunhao Sun, Gulou Shen, Christoph Held, Xin Feng, Xiaohua Lu, and Xiaoyan Ji. "Modeling Viscosity of Ionic Liquids with Electrolyte Perturbed-Chain Statistical Associating Fluid Theory and Free Volume Theory". *Industrial & Engineering Chemistry Research* 57.26 (2018), pp. 8784–8801. DOI: 10.1021/acs.iecr.8b00328.

- [216] Marcela Cartes, Gustavo Chaparro, Gerard Alonso, and Andrés Mejía. "Density and viscosity of liquid mixtures formed by n-hexane, ethanol, and cyclopentyl methyl ether". *Journal of Molecular Liquids* 359 (2022), p. 119353. DOI: 10.1016/j.molliq.2022.119353.
- [217] Yaakov Rosenfeld. "Relation between the transport coefficients and the internal entropy of simple systems". *Physical Review A* 15.6 (1977), pp. 2545–2549. DOI: 10.1103/PhysRevA.15.2545.
- [218] Jeppe C. Dyre. "Hidden Scale Invariance in Condensed Matter". *The Journal of Physical Chemistry B* 118.34 (2014), pp. 10007–10024. DOI: 10.1021/jp501852b.
- [219] Ian H. Bell, Richard Messerly, Monika Thol, Lorenzo Costigliola, and Jeppe C. Dyre. "Modified Entropy Scaling of the Transport Properties of the Lennard-Jones Fluid". *The Journal of Physical Chemistry B* 123.29 (2019), pp. 6345–6363. DOI: 10.1021/acs.jpcb.9b05808.
- [220] Ian H. Bell. "Probing the link between residual entropy and viscosity of molecular fluids and model potentials". *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4070–4079. DOI: 10.1073/pnas.1815943116.
- [221] Ian H. Bell. "Entropy Scaling of Viscosity—I: A Case Study of Propane". *Journal of Chemical & Engineering Data* 65.6 (2020), pp. 3203–3215. DOI: 10.1021/acs.jced.0c00209.
- [222] Aghilas Dehlouz, Romain Privat, Marc Bonnissel, and Jean-Noël Jaubert. "An experiment-design methodology for the selection of optimal experimental conditions for the correlation of transport properties". *Fluid Phase Equilibria* 572 (2023), p. 113829. DOI: 10.1016/j.fluid.2023.113829.
- [223] Joseph O. Hirschfelder, Charles F. Curtiss, and Robert Byron Bird. *Molecular theory of gases and liquids*. Wiley, 1964. ISBN: 978-0-471-40065-3.
- [224] Justinas Šlepavičius, Alessandro Patti, James L. McDonagh, and Carlos Avendaño. "Application of machine-learning algorithms to predict the transport properties of Mie fluids". *The Journal of Chemical Physics* 159.2 (2023), p. 024127. DOI: 10.1063/5.0151123.
- [225] Joshua P. Allers, Jacob A. Harvey, Fernando H. Garzon, and Todd M. Alam. "Machine learning prediction of self-diffusion in Lennard-Jones fluids". *The Journal of Chemical Physics* 153.3 (2020), p. 034102. DOI: 10.1063/5.0011512.
- [226] Joshua P. Allers, Chad W. Priest, Jeffery A. Greathouse, and Todd M. Alam. "Using Computationally-Determined Properties for Machine Learning Prediction of Self-Diffusion Coefficients in Pure Liquids". *The Journal of Physical Chemistry B* 125.47 (2021), pp. 12990–13002. DOI: 10.1021/acs.jpcb.1c07092.
- [227] Todd M. Alam, Joshua P. Allers, Calen J. Leverant, and Jacob A. Harvey. "Symbolic regression development of empirical equations for diffusion in Lennard-Jones fluids". *The Journal of Chemical Physics* 157.1 (2022), p. 014503. DOI: 10.1063/5.0093658.
- [228] Florian Müller-Plathe. "A simple nonequilibrium molecular dynamics method for calculating the thermal conductivity". *The Journal of Chemical Physics* 106.14 (1997), pp. 6082–6085. DOI: 10.1063/1.473271.
- [229] Florian Müller-Plathe. "Reversing the perturbation in nonequilibrium molecular dynamics: An easy way to calculate the shear viscosity of fluids". *Physical Review E* 59.5 (1999), pp. 4894–4898. DOI: 10.1103/PhysRevE.59.4894.

- [230] Martin P. Lautenschlaeger, Martin Horsch, and Hans Hasse. "Simultaneous determination of thermal conductivity and shear viscosity using two-gradient non-equilibrium molecular dynamics simulations". *Molecular Physics* 117.2 (2019), pp. 189–199. DOI: 10.1080/00268976.2018.1504134.
- [231] Edward J. Maginn, Richard A. Messerly, Daniel J. Carlson, Daniel R. Roe, and J. Richard Elliot. "Best Practices for Computing Transport Properties 1. Self-Diffusivity and Viscosity from Equilibrium Molecular Dynamics [Article v1.0]". *Living Journal of Computational Molecular Science* 2.1 (2020). DOI: 10.33011/livecoms.1.1.6324.
- [232] M. A. Van Der Hoef and D. Frenkel. "Long-time tails of the velocity autocorrelation function in two- and three-dimensional lattice-gas cellular automata: A test of mode-coupling theory". *Physical Review A* 41.8 (1990), pp. 4277–4284. DOI: 10.1103/PhysRevA.41.4277.
- [233] M. H. J. Hagen, C. P. Lowe, and D. Frenkel. "Long time tails in stress correlation functions". *25 Years of Non-Equilibrium Statistical Mechanics*. Ed. by J. J. Brey, J. Marro, J. M. Rubí, and M. San Miguel. Vol. 445. Springer Berlin Heidelberg, 1995, pp. 240–249. ISBN: 978-3-540-59158-0. DOI: 10.1007/3-540-59158-3\_47.
- [234] Yong Zhang, Akihito Otani, and Edward J. Maginn. "Reliable Viscosity Calculation from Equilibrium Molecular Dynamics Simulations: A Time Decomposition Method". *Journal of Chemical Theory and Computation* 11.8 (2015), pp. 3537–3546. DOI: 10.1021/acs.jctc.5b00351.
- [235] Seyed Hossein Jamali, Ludger Wolff, Tim M. Becker, Mariëtte De Groen, Mahinder Ramdin, Remco Hartkamp, André Bardow, Thijs J. H. Vlugt, and Othonas A. Moulτος. "OCTP: A Tool for On-the-Fly Calculation of Transport Properties of Fluids with the Order-  $n$  Algorithm in LAMMPS". *Journal of Chemical Information and Modeling* 59.4 (2019), pp. 1290–1294. DOI: 10.1021/acs.jcim.8b00939.
- [236] Peter J. Daivis and Denis J. Evans. "Comparison of constant pressure and constant volume nonequilibrium simulations of sheared model decane". *The Journal of Chemical Physics* 100.1 (1994), pp. 541–547. DOI: 10.1063/1.466970.
- [237] Craig M. Tenney and Edward J. Maginn. "Limitations and recommendations for the calculation of shear viscosity using reverse nonequilibrium molecular dynamics". *The Journal of Chemical Physics* 132.1 (2010), p. 014103. DOI: 10.1063/1.3276454.
- [238] Burkhard Dünweg and Kurt Kremer. "Molecular dynamics simulation of a polymer chain in solution". *The Journal of Chemical Physics* 99.9 (1993), pp. 6983–6997. DOI: 10.1063/1.465445.
- [239] In-Chul Yeh and Gerhard Hummer. "System-Size Dependence of Diffusion Coefficients and Viscosities from Molecular Dynamics Simulations with Periodic Boundary Conditions". *The Journal of Physical Chemistry B* 108.40 (2004), pp. 15873–15879. DOI: 10.1021/jp0477147.
- [240] Seyed Hossein Jamali, Remco Hartkamp, Christos Bardas, Jakob Söhl, Thijs J. H. Vlugt, and Othonas A. Moulτος. "Shear Viscosity Computed from the Finite-Size Effects of Self-Diffusivity in Equilibrium Molecular Dynamics". *Journal of Chemical Theory and Computation* 14.11 (2018), pp. 5959–5968. DOI: 10.1021/acs.jctc.8b00625.
- [241] Alper T. Celebi, Seyed Hossein Jamali, André Bardow, Thijs J. H. Vlugt, and Othonas A. Moulτος. "Finite-size effects of diffusion coefficients computed from molecular dynamics: a review of what we have learned so far". *Molecular Simulation* 47.10 (2021), pp. 831–845. DOI: 10.1080/08927022.2020.1810685.

- [242] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649. DOI: 10.1063/1.442716.
- [243] J.P.J. Michels and N.J. Trappeniers. "Molecular dynamical calculations of the viscosity of Lennard-Jones systems". *Physica A: Statistical Mechanics and its Applications* 133.1 (1985), pp. 281–290. DOI: 10.1016/0378-4371(85)90067-6.
- [244] David M. Heyes. "Transport coefficients of Lennard-Jones fluids: A molecular-dynamics and effective-hard-sphere treatment". *Physical Review B* 37.10 (1988), pp. 5677–5696. DOI: 10.1103/PhysRevB.37.5677.
- [245] D.M. Heyes and J.G. Powles. "Information theory applied to the transport coefficients of Lennard-Jones fluids". *Molecular Physics* 71.4 (1990), pp. 781–800. DOI: 10.1080/00268979000102111.
- [246] R. L. Rowley and M. M. Painter. "Diffusion and viscosity equations of state for a Lennard-Jones fluid obtained from molecular dynamics simulations". *International Journal of Thermophysics* 18.5 (1997), pp. 1109–1121. DOI: 10.1007/BF02575252.
- [247] V. R. Vasquez, E. A. Macedo, and M. S. Zabaloy. "Lennard-Jones Viscosities in Wide Ranges of Temperature and Density: Fast Calculations Using a Steady-State Periodic Perturbation Method". *International Journal of Thermophysics* 25.6 (2004), pp. 1799–1818. DOI: 10.1007/s10765-004-7736-3.
- [248] Guillaume Galliéro, Christian Boned, and Antoine Baylaucq. "Molecular Dynamics Study of the Lennard-Jones Fluid Viscosity: Application to Real Fluids". *Industrial & Engineering Chemistry Research* 44.17 (2005), pp. 6963–6972. DOI: 10.1021/ie050154t.
- [249] Afshin Eskandari Nasrabad, Rozita Laghaei, and Byung Chan Eu. "Molecular theory of thermal conductivity of the Lennard-Jones fluid". *The Journal of Chemical Physics* 124.8 (2006), p. 084506. DOI: 10.1063/1.2166394.
- [250] Mathilde Bugel and Guillaume Galliero. "Thermal conductivity of the Lennard-Jones fluid: An empirical correlation". *Chemical Physics* 352.1 (2008), pp. 249–257. DOI: 10.1016/j.chemphys.2008.06.013.
- [251] Guillaume Galliero and Christian Boned. "Thermal conductivity of the Lennard-Jones chain fluid model". *Physical Review E* 80.6 (2009), p. 061202. DOI: 10.1103/PhysRevE.80.061202.
- [252] V.G. Baidakov, S.P. Protsenko, and Z.R. Kozlova. "The self-diffusion coefficient in stable and metastable states of the Lennard-Jones fluid". *Fluid Phase Equilibria* 305.2 (2011), pp. 106–113. DOI: 10.1016/j.fluid.2011.03.002.
- [253] Vladimir G. Baidakov and Sergey P. Protsenko. "Metastable Lennard-Jones fluids. II. Thermal conductivity". *The Journal of Chemical Physics* 140.21 (2014), p. 214506. DOI: 10.1063/1.4880958.
- [254] Stephanie Delage-Santacreu, Guillaume Galliero, Hai Hoang, Jean-Patrick Bazile, Christian Boned, and Josefa Fernandez. "Thermodynamic scaling of the shear viscosity of Mie  $n$ -6 fluids and their binary mixtures". *The Journal of Chemical Physics* 142.17 (2015), p. 174501. DOI: 10.1063/1.4919296.
- [255] Geoffrey C. Maitland. *Intermolecular forces: Their origin and determination*. Clarendon Pr, 1981. ISBN: 978-0-19-855611-4.

- [256] A. Michels, J.V. Sengers, and P.S. Van Der Gulik. "The thermal conductivity of carbon dioxide in the critical region". *Physica* 28.12 (1962), pp. 1216–1237. DOI: 10.1016/0031-8914(62)90135-0.
- [257] B. J. Bailey and K. Kellner. "The thermal conductivity of argon near the critical point". *British Journal of Applied Physics* 18.11 (1967), pp. 1645–1647. DOI: 10.1088/0508-3443/18/11/420.
- [258] J. V. Sengers. "Transport properties of fluids near critical points". *International Journal of Thermophysics* 6.3 (1985), pp. 203–232. DOI: 10.1007/BF00522145.
- [259] J. V. Sengers. "Behavior of Viscosity and Thermal Conductivity of Fluids Near the Critical Point". *Critical Phenomena: Proceedings*. Critical Phenomena. U.S. Department of Commerce, National Bureau of Standards, 1966, pp. 165–178.
- [260] Ian H. Bell. "Entropy Scaling of Viscosity—II: Predictive Scheme for Normal Alkanes". *Journal of Chemical & Engineering Data* 65.11 (2020), pp. 5606–5616. DOI: 10.1021/acs.jced.0c00749.
- [261] Andres Carranza-Abaid, Hallvard F. Svendsen, and Jana P. Jakobsen. "Thermodynamically consistent vapor-liquid equilibrium modelling with artificial neural networks". *Fluid Phase Equilibria* 564 (2023), p. 113597. DOI: 10.1016/j.fluid.2022.113597.
- [262] B. Smit and D. Frenkel. "Vapor-liquid equilibria of the two-dimensional Lennard-Jones fluid(s)". *The Journal of Chemical Physics* 94.8 (1991), pp. 5663–5668. DOI: 10.1063/1.460477.
- [263] B. Smit. "Phase diagrams of Lennard-Jones fluids". *The Journal of Chemical Physics* 96.11 (1992), pp. 8639–8640. DOI: 10.1063/1.462271.
- [264] Denis Saric, Ian H. Bell, Gabriela Guevara-Carrion, and Jadran Vrabec. "Influence of repulsion on entropy scaling and density scaling of monatomic fluids". *The Journal of Chemical Physics* 160.10 (2024), p. 104503. DOI: 10.1063/5.0196592.
- [265] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids". *Journal of the American Chemical Society* 118.45 (1996), pp. 11225–11236. DOI: 10.1021/ja9621760.
- [266] Marcus G. Martin and J. Ilja Siepmann. "Transferable Potentials for Phase Equilibria. 1. United-Atom Description of *n*-Alkanes". *The Journal of Physical Chemistry B* 102.14 (1998), pp. 2569–2577. DOI: 10.1021/jp972543+.
- [267] Andrea Hemmen and Joachim Gross. "Transferable Anisotropic United-Atom Force Field Based on the Mie Potential for Phase Equilibrium Calculations: *n*-Alkanes and *n*-Olefins". *The Journal of Physical Chemistry B* 119.35 (2015), pp. 11695–11707. DOI: 10.1021/acs.jpccb.5b01354.
- [268] Jörg Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". *Physical Review Letters* 98.14 (2007), p. 146401. DOI: 10.1103/PhysRevLett.98.146401.
- [269] Christoph Schran, Krystof Brezina, and Ondrej Marsalek. "Committee neural network potentials control generalization errors and enable active learning". *The Journal of Chemical Physics* 153.10 (2020), p. 104105. DOI: 10.1063/5.0016004.
- [270] Volker L. Deringer, Davide M. Proserpio, Gábor Csányi, and Chris J. Pickard. "Data-driven learning and prediction of inorganic crystal structures". *Faraday Discussions* 211 (2018), pp. 45–59. DOI: 10.1039/C8FD00034D.
- [271] Dávid Péter Kovács, Cas Van Der Oord, Jiri Kucera, Alice E. A. Allen, Daniel J. Cole, Christoph Ortner, and Gábor Csányi. "Linear Atomic Cluster Expansion Force Fields for Organic Molecules:

- Beyond RMSE". *Journal of Chemical Theory and Computation* 17.12 (2021), pp. 7696–7711. DOI: 10.1021/acs.jctc.1c00647.
- [272] Dávid Péter Kovács, Ilyes Batatia, Eszter Sára Arany, and Gábor Csányi. "Evaluation of the MACE force field architecture: From medicinal chemistry to materials science". *The Journal of Chemical Physics* 159.4 (2023), p. 044118. DOI: 10.1063/5.0155322.
- [273] Kevin R. Hadley and Clare McCabe. "On the Investigation of Coarse-Grained Models for Water: Balancing Computational Efficiency and the Retention of Structural Properties". *The Journal of Physical Chemistry B* 114.13 (2010), pp. 4590–4599. DOI: 10.1021/jp911894a.
- [274] Timothy C. Moore, Christopher R. Iacovella, and Clare McCabe. "Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion". *The Journal of Chemical Physics* 140.22 (2014), p. 224104. DOI: 10.1063/1.4880555.
- [275] Olga Lobanova, Carlos Avendaño, Thomas Lafitte, Erich A. Müller, and George Jackson. "SAFT- $\gamma$  force field for the simulation of molecular fluids: 4. A single-site coarse-grained model of water applicable over a wide temperature range". *Molecular Physics* 113.9 (2015), pp. 1228–1249. DOI: 10.1080/00268976.2015.1004804.
- [276] Tristan G. A. Youngs, Mario G. Del Pópolo, and Jorge Kohanoff. "Development of Complex Classical Force Fields through Force Matching to ab Initio Data: Application to a Room-Temperature Ionic Liquid". *The Journal of Physical Chemistry B* 110.11 (2006), pp. 5697–5707. DOI: 10.1021/jp056931k.
- [277] Jaehyeok Jin, Alexander J. Pak, Aleksander E. P. Durumeric, Timothy D. Loose, and Gregory A. Voth. "Bottom-up Coarse-Graining: Principles and Perspectives". *Journal of Chemical Theory and Computation* 18.10 (2022), pp. 5759–5791. DOI: 10.1021/acs.jctc.2c00643.
- [278] J. E. Jones. "On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature". *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 106.738 (1924), pp. 441–462. DOI: 10.1098/rspa.1924.0081.
- [279] J. E. Jones. "On the determination of molecular fields. —II. From the equation of state of a gas". *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 106.738 (1924), pp. 463–477. DOI: 10.1098/rspa.1924.0082.
- [280] Sadia Rahman, Olga Lobanova, Guadalupe Jiménez-Serratos, Carlos Braga, Vasilios Raptis, Erich A. Müller, George Jackson, Carlos Avendaño, and Amparo Galindo. "SAFT- $\gamma$  Force Field for the Simulation of Molecular Fluids. 5. Hetero-Group Coarse-Grained Models of Linear Alkanes and the Importance of Intramolecular Interactions". *The Journal of Physical Chemistry B* 122.39 (2018), pp. 9161–9177. DOI: 10.1021/acs.jpcc.8b04095.
- [281] Carlos Avendaño, Thomas Lafitte, Amparo Galindo, Claire S. Adjiman, George Jackson, and Erich A. Müller. "SAFT- $\gamma$  Force Field for the Simulation of Molecular Fluids. 1. A Single-Site Coarse Grained Model of Carbon Dioxide". *The Journal of Physical Chemistry B* 115.38 (2011), pp. 11154–11169. DOI: 10.1021/jp204908d.
- [282] Erich A. Müller and Andrés Mejía. "Extension of the SAFT-VR Mie EoS To Model Homonuclear Rings and Its Parametrization Based on the Principle of Corresponding States". *Langmuir* 33.42 (2017), pp. 11518–11529. DOI: 10.1021/acs.langmuir.7b00976.
- [283] J. L. F. Abascal and C. Vega. "A general purpose model for the condensed phases of water: TIP4P/2005". *The Journal of Chemical Physics* 123.23 (2005), p. 234505. DOI: 10.1063/1.2121687.



- [284] J. L. F. Abascal, E. Sanz, R. García Fernández, and C. Vega. "A potential model for the study of ices and amorphous water: TIP4P/Ice". *The Journal of Chemical Physics* 122.23 (2005), p. 234511. DOI: 10.1063/1.1931662.
- [285] Bridgette J. Befort, Ryan S. DeFever, Garrett M. Tow, Alexander W. Dowling, and Edward J. Maginn. "Machine Learning Directed Optimization of Classical Molecular Modeling Force Fields". *Journal of Chemical Information and Modeling* 61.9 (2021), pp. 4400–4414. DOI: 10.1021/acs.jcim.1c00448.
- [286] Zhi Liang and Hai-Lung Tsai. "The vibrational contribution to the thermal conductivity of a polyatomic fluid". *Molecular Physics* 108.13 (2010), pp. 1707–1714. DOI: 10.1080/00268976.2010.489520.
- [287] Fuchang Gao and Lixing Han. "Implementing the Nelder-Mead simplex algorithm with adaptive parameters". *Computational Optimization and Applications* 51.1 (2012), pp. 259–277. DOI: 10.1007/s10589-010-9329-3.
- [288] I.M Sobol'. "On the distribution of points in a cube and the approximate evaluation of integrals". *USSR Computational Mathematics and Mathematical Physics* 7.4 (1967), pp. 86–112. DOI: 10.1016/0041-5553(67)90144-9.
- [289] Lingru Zheng, J.P. Martin Trusler, Fernando Bresme, and Erich A. Müller. "Predicting the pressure dependence of the viscosity of 2,2,4-trimethylhexane using the SAFT coarse-grained force field". *Fluid Phase Equilibria* 496 (2019), pp. 1–6. DOI: 10.1016/j.fluid.2019.05.017.
- [290] Karsten Meier and Stephan Kabelac. "Speed of sound instrument for fluids with pressures up to 100 MPa". *Review of Scientific Instruments* 77.12 (2006), p. 123903. DOI: 10.1063/1.2400019.
- [291] Carsten Wedler and J. P. Martin Trusler. "Speed of Sound Measurements in Helium at Pressures from 15 to 100 MPa and Temperatures from 273 to 373 K". *Journal of Chemical & Engineering Data* 68.6 (2023), pp. 1305–1312. DOI: 10.1021/acs.jced.3c00083.
- [292] Carlos Avendaño, Thomas Lafitte, Claire S. Adjiman, Amparo Galindo, Erich A. Müller, and George Jackson. "SAFT- $\gamma$  Force Field for the Simulation of Molecular Fluids: 2. Coarse-Grained Models of Greenhouse Gases, Refrigerants, and Long Alkanes". *The Journal of Physical Chemistry B* 117.9 (2013), pp. 2717–2733. DOI: 10.1021/jp306442b.
- [293] Simon Dufal, Thomas Lafitte, Andrew J. Haslam, Amparo Galindo, Gary N.I. Clark, Carlos Vega, and George Jackson. "The A in SAFT: developing the contribution of association to the Helmholtz free energy within a Wertheim TPT1 treatment of generic Mie fluids". *Molecular Physics* 113.9 (2015), pp. 948–984. DOI: 10.1080/00268976.2015.1029027.
- [294] *RDKit: open-source cheminformatics*. Version 2024.09.05. URL: <https://www.rdkit.org> (visited on Feb. 26, 2025).
- [295] Co D. Quach, Justin B. Gilmer, Daniel Pert, Akanke Mason-Hogans, Christopher R. Iacovella, Peter T. Cummings, and Clare McCabe. "High-throughput screening of tribological properties of monolayer films using molecular dynamics and machine learning". *The Journal of Chemical Physics* 156.15 (2022), p. 154902. DOI: 10.1063/5.0080838.
- [296] *Dortmund Data Bank*. 2025. URL: [www.ddbst.com](http://www.ddbst.com) (visited on Mar. 11, 2025).
- [297] Jens Staubach and Simon Stephan. "Characteristic curves of the stockmayer fluid: Molecular simulation and equation of state modeling". *Fluid Phase Equilibria* 592 (2025), p. 114314. DOI: 10.1016/j.fluid.2024.114314.

- [298] Clare McCabe and George Jackson. "SAFT-VR modelling of the phase equilibrium of long-chain n-alkanes". *Physical Chemistry Chemical Physics* 1.9 (1999), pp. 2057–2064. DOI: 10.1039/a808085b.
- [299] Jan Hrubý. "A General Model for Thermodynamic Properties of Fluid Mixtures Based on Helmholtz Energy Formulations for the Components. Virial Expansion and Reduction to van der Waals Mixing Rules". *International Journal of Thermophysics* 44.8 (2023), p. 130. DOI: 10.1007/s10765-023-03237-8.
- [300] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". *Communications of the ACM* 60.6 (2017), pp. 84–90. DOI: 10.1145/3065386.
- [301] Waseem Rawat and Zenghui Wang. "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review". *Neural Computation* 29.9 (2017), pp. 2352–2449. DOI: 10.1162/neco\_a\_00990.
- [302] L. Marquez, T. Hill, R. Worthley, and W. Remus. "Neural network models as an alternative to regression". *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*. Vol. iv. Kauai, HI, USA: IEEE Comput. Soc. Press, 1991, pp. 129–135. DOI: 10.1109/HICSS.1991.184052.
- [303] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. *On the Number of Linear Regions of Deep Neural Networks*. 2014. URL: <http://arxiv.org/abs/1402.1869>.
- [304] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Vol. 9. Proceedings of Machine Learning Research, 2010, pp. 249–256. URL: <https://proceedings.mlr.press/v9/glorot10a.html>.
- [305] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. URL: <http://arxiv.org/abs/1502.01852>.
- [306] Nitish Shirish Keskar and Richard Socher. *Improving Generalization Performance by Switching from Adam to SGD*. 2017. URL: <http://arxiv.org/abs/1712.07628>.
- [307] Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. *Imposing Hard Constraints on Deep Networks: Promises and Limitations*. 2017. DOI: 10.48550/ARXIV.1706.02025.
- [308] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. "An Efficient Approach for Assessing Hyperparameter Importance". *Proceedings of the 31st International Conference on Machine Learning*. International Conference on Machine Learning. Vol. 32. China, 2014, pp. 754–762. URL: <https://proceedings.mlr.press/v32/hutter14.html>.

## Appendix A

# Data modeling with artificial neural networks

### A.1 Artificial neural networks description

This appendix provides the basic background for Artificial neural networks (ANNs). The reader is referred to the books of Géron [135] and Bishop and Bishop [78] for a thorough discussion about ANNs.

ANNs are a versatile machine learning algorithm inspired by how biological neurons communicate and propagate information between them. ANNs are suitable for classification [300, 301] and regression [302] problems. ANNs are universal function approximators [53] and were first introduced by McCulloch and Pitts [52] in 1943. ANNs rely on several adjustable parameters, which were difficult to tune at the time. It was not until 1986 that the backpropagation method [88] was proposed as a feasible training algorithm. This method allowed the efficient computation of derivatives at computer precision and facilitated the training of ANNs for practical applications. Nowadays, increasing computer capabilities of Central Processing Units (CPUs), Graphics Processing Units (GPUs), and Tensor Processing Units (TPUs) have made ANNs a standard engineering tool for data correlation [57].

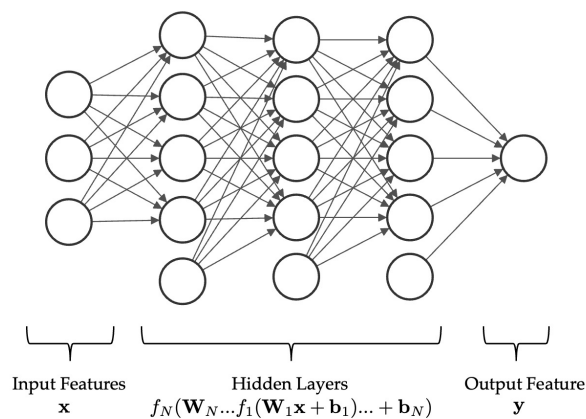


FIGURE A.1: Schematic representation of an Artificial Neural Network (ANN).

Mathematically speaking, an  $N$ -layered ANN consists of a series of  $N$  linear transformations followed by a nonlinear transformation, also known as the activation function. A schematic representation of an ANN is shown in Figure A.1. An ANN consists of three main parts, namely: an input layer where a feature vector  $\mathbf{x}$  is supplied, an output layer that produces the vector  $\mathbf{y}$ , and the hidden layers where the ANN uses trainable parameters and nonlinear transformation to map from  $\mathbf{x}$  to  $\mathbf{y}$ . The mathematical representation of an ANN is given below in Eq. (A.1).

$$\mathbf{y} = \text{ANN}(\mathbf{x}) = f_N(\mathbf{W}_N \dots f_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \dots + \mathbf{b}_N) \quad (\text{A.1})$$

In Eq. (A.1)  $\mathbf{x}$  and  $\mathbf{y}$  are the input and output features of the ANN. In this equation, each layer applies a linear transformation  $\mathbf{W}_i \mathbf{x}_{i-1} + \mathbf{b}_i$ . The weight matrix ( $\mathbf{W}_i$ ) and the bias vector ( $\mathbf{b}_i$ ) are the trainable parameters of the  $i$ -th layer of the ANN. The weight matrix is an  $(m \times n)$  dimensional matrix, where  $m$  is the number of neurons in the layer and  $n$  is the dimension of the input vector. The bias vector is of dimension  $m$ .

An activation function,  $f_i$ , follows the linear transformation of each layer. The role of the activation function is considerable. This function adds non-linearity to the ANN and provides the flexibility to model complex data behavior [135]. A summary of commonly used activation functions is given in Table A.1. For regression problems, it is recommended to use a linear activation function (i.e.,  $f(x) = x$ ) in the output layer. The linear activation function is equivalent to the identity function and can produce any real value. If the target feature is always positive, using the softplus activation function (i.e.,  $f(x) = \ln[1 + \exp(x)]$ ) ensures the predicted values are always in that domain. For binary classification problems, it is recommended to use the sigmoid activation function (i.e.,  $f(x) = 1/[1 + \exp(-x)]$ ) as the output activation function. The output of this function is bounded between 0 and 1, which facilitates the classification into two categories given a threshold. Based on a similar idea, the softmax activation function ( $f(x)_i = \exp(x_i) / \sum \exp(x)$ ) can be used for multiclass classification. For the case of the hidden layers, the ReLU activation function (i.e.,  $f(x) = \max(0, x)$ ) is a standard choice due to its low computational cost, which translates into faster training. However, the ReLU function is not continuous, and the final ANN result can be interpreted as a piece-wise hyperplane function [303]. This behavior might not be desired for continuous data. In this case, it is preferable to use the tanh activation function (i.e.,  $f(x) = \tanh(x)$ ).

### Artificial Neural Network training

The training of an ANN consists of updating the weights ( $\mathbf{W}_i$ ) and biases ( $\mathbf{b}_i$ ) of each layer to improve the performance of the model. For example, to improve classification accuracy or reduce the squared error in a regression problem. The weight matrices are usually randomly initialized [304, 305], whereas the biases are initialized as a zero vector. The training of the ANN is performed by minimizing a loss function,  $\mathcal{L}$ . In regression problems, which are of interest to this thesis, the loss function is based on the Mean Squared Error (MSE), as shown below.

$$\underset{\mathbf{W}_i, \mathbf{b}_i}{\text{minimize}} \mathcal{L}(\mathbf{W}_i, \mathbf{b}_i) = \frac{1}{N_p} \sum_{i=1}^{N_p} (y_i^{\text{pred}} - y_i^{\text{true}})^2 \quad (\text{A.2})$$

TABLE A.1: Common activation functions used in ANNs.

| Activation function | Definition                                             | Output range        | Comments                                                                                                                                                                                   |
|---------------------|--------------------------------------------------------|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| linear              | $f(x) = x$                                             | $(-\infty, \infty)$ | Recommended for the output layer in regression problems                                                                                                                                    |
| tanh                | $f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ | $[-1, 1]$           | Centered at zero, considered superior to sigmoid function. It can also saturate if the input is too big or too small.                                                                      |
| ReLU                | $f(x) = \max(0, x)$                                    | $[0, \infty)$       | Default activation function for fast training. The function is not differentiable, and it approximates the data by superpositioning of hyperplanes.                                        |
| softplus            | $f(x) = \ln[1 + \exp(x)]$                              | $[0, \infty)$       | Differentiable version of ReLU activation function. It is useful for regression in which the output value is always positive.                                                              |
| sigmoid             | $f(x) = \frac{1}{1 + \exp(-x)}$                        | $[0, 1]$            | Useful for binary classification. It saturates quickly if the input is too big or too small. In the case of saturation, the gradients tend to zero while training, implying slow training. |
| softmax             | $f(x)_i = \frac{\exp(x_i)}{\sum \exp(x)}$              | $[0, 1]$            | Useful for multiclass classification.                                                                                                                                                      |

In Eq. (A.2),  $\mathbf{y}_i^{\text{true}}$  and  $\mathbf{y}_i^{\text{pred}}$  refer to the  $i$ -th sample of the true and predicted output feature. In principle, this minimization problem can be solved by any optimization method. However, due to the high dimensionality of the problem and the computational cost associated with higher-order derivatives, gradient descent-based methods are preferred. A gradient descent step is illustrated below.

$$\theta_{j+1} = \theta_j - \alpha \frac{\partial \mathcal{L}}{\partial \theta_j} \quad (\text{A.3})$$

In Eq. (A.3),  $\theta_i$  refers to a parameter of the ANN (i.e., weight or bias) at step  $j$ , and  $\alpha$  is the learning rate. The derivatives of the loss function with respect to these parameters ( $\partial \mathcal{L} / \partial \theta_i$ ) are obtained using the backpropagation method [88]. This method computes derivatives at computer precision, and it is available in modern machine learning libraries, such as TensorFlow [137] and JAX [138]. The minimization step is often modified to include information about previous iterations, known as momentum. The Adam [139] optimizer is an example of an optimization algorithm with momentum. It should be noted that even with this "improvement", there is no guarantee of achieving better performance during training [306]. For this reason, the optimizer selection is a hyperparameter to choose before training an artificial neural network.

Regardless of which optimization algorithm is selected, the learning rate ( $\alpha$ ) is one of the most relevant hyperparameters during training. The learning rate must be tuned for a specific task. On the one hand, small values of  $\alpha$  imply the parameters of the ANN are slightly updated on each optimization step, see Eq. (A.3). This conservative approach implies a slow training. On the other hand, big values of  $\alpha$  can cause an ill-behaved optimization, where the loss function exhibits an oscillating behavior. In extreme cases, a large learning rate can cause the training to diverge. Typically, the learning is chosen

as a power of ten, with common values ranging from  $10^{-4}$  to  $10^{-2}$  [135].

Additionally, due to memory constraints, the ANN training is usually performed using mini-batches of data. It has been shown that using "small" mini-batches helps the optimizer move between local minima. Conversely, "large" mini-batches lead to a sharper definition of the loss function surface, which can cause the optimizer to become stuck in a local minimum [192]. The batch size is commonly chosen as a power of two, with typical values ranging from  $2^5$  to  $2^8$ . The number of epochs refers to how many times the optimizer passes through the entire dataset. The number of epochs to achieve satisfactory training will vary depending on the task. A customary strategy to select the number of epochs is to analyze the loss function with respect to the number of epochs. Ideally, the number of epochs must be chosen so the loss function reaches a plateau and any extra improvement is marginal. The batch size and the number of epochs are training hyperparameters of the model.

Finally, the ability of ANNs to learn complex patterns from data comes at the risk of overfitting. In this case, an ANN performs exceptionally well for the data used for training, but it does not generalize well for "unseen" data points. The overfitting of an ANN is easily verified by splitting the database into a train and a test set. The ANN is trained using only the training set, but the loss function is also evaluated for the test set. If the loss of the training set is considerably lower than the loss of the test set, then the model is overfitting the data. There are regularization techniques to avoid overfitting. For example, the Ridge or Lasso regularizations [135] add a penalty to the loss function based on the magnitude of the weights ( $\mathbf{W}_i$ ) of the ANN. These techniques preclude large weights and decrease the effective degrees of freedom of the ANN, resulting in better-behaved models. Alternatively, as described in the next section, the regularization can be inspired by prior (physics-inspired) knowledge.

### Physics-informed artificial neural networks

Artificial neural networks have been extensively used for scientific and engineering purposes. In these cases, it is tangible to add extra constraints to ensure the data follows a correct physical behavior [50]. These constraints can be embedded into the ANN formulation. For example, the FE-ANN EoS proposed in Chapter 2 fulfills the ideal gas law and Maxwell's relationships by design. A similar approach has been proposed for an ANN-based activity coefficient model [84] to fulfill the Gibbs-Duhem equation. Moreover, artificial neural networks have also been formulated to fulfill known limits of activity coefficient models [85]. Embedding physical constraints in an ANN model acts as a regularization technique that prevents overfitting to nonphysical solutions.

Alternatively, soft constraints can be added to the loss function during the training of the ANN [86]. These penalty functions, like those described in Eq. (2.13) and Eq. (4.18), are added alongside the primary loss function to enforce correct physical behavior while training the model parameters. These soft constraints are not guaranteed to be fulfilled at the end of the training and the correctness of the model must be checked. Alternatively, more complex approaches can impose physical behavior as hard constraints, but they are limited to simplified scenarios where the soft-constrained approach performs similarly [307].

## Hyperparameter optimization of artificial neural networks

The performance of an ANN for a given task depends on several hyperparameters. The hyperparameters related to the architecture include the number of hidden layers, the number of neurons per layer, and the activation function used in each layer. Similarly, the training of the ANN depends on the selection of the training algorithm, learning rate, use of regularization techniques, batch size, and total number of epochs. These hyperparameters must be tuned for a given task. This procedure is non-trivial as it requires handling different variable types. For example, some hyperparameters can be treated as continuous variables (e.g., the learning rate), others are integer values (e.g., the number of hidden layers), and others are choices (e.g., the optimization algorithm). Several methods have been proposed to solve this problem efficiently. These include using a grid search, random search, genetic algorithms, and Bayesian optimization. The reader is referred to the review of Yang and Shami [140] for a thorough discussion of these methods.

### A.2 Hyperparameter optimization of the physics-informed artificial neural network models

This section details the hyperparameter optimization (HPO) of the artificial neural networks used in this thesis. The hyperparameter optimization was performed using the Tree-structured Parzen Estimator [159] method implemented in black-box optimization library *optuna* [160]. This method efficiently samples the hyperparameters using Bayesian optimization. In order to keep the complexity of the optimization low, the HPO aims to find only a suitable architecture (number of hidden layers and neurons per layer) to model a target property. The results are reproducible by controlling the seed to initiate the ANN parameters. The importance of these hyperparameters was assessed using the fANOVA method [308] included in *Optuna*. Other hyperparameters related to the training are chosen from standard practices related to ANNs [78, 135]. For example, the training of the ANN uses the Adam optimization method [139], a batch size of 32 is used for training, the learning rate is set  $\alpha = 10^{-3}$  and the number of epochs to 1000. The selected learning rate and number of epochs allow to train ANNs quickly. This quick training benefits the exploration of the Bayesian optimization algorithm to find a suitable architecture to describe a given thermophysical property.

Finally, the ANN models in the HPO are trained using 90% of the available data and compared against each other using the remaining 10% of it. Due to the size of the datasets, this split is sufficient to provide a similar phase space distribution between the training and testing datasets.

### FE-ANN EoS from Chapter 2

The FE-ANN EoS explicitly models the residual Helmholtz free energy. As explained in Chapter 2, the FE-ANN EoS is trained based on the derivatives of the Helmholtz free energy. The architecture of the ANN (number of neurons and neurons per layer) used in the FE-ANN EoS is optimized exclusively to first-order thermophysical properties of the Mie fluid. These properties include the pressure ( $P^*$ ), internal energy ( $U^*$ ), and the second-virial coefficient ( $B_2^*$ ). No second-order properties or penalty functions were considered during this process. The training of these test models uses a learning rate

$\alpha = 10^{-3}$  and 1000 epochs. These values differ from the ones used in Chapter 2 as the FE-ANN EoS trained in this chapter considers second-order properties and penalty functions. This extra information requires a smaller learning rate to ensure a stable optimization. Due to the slow training, in Chapter 2, it is required to increase the number of epochs.

The hyperparameter optimization results are summarized in Figure A.2. From Figure A.2 (a), it can be observed that the number of neurons is the most relevant variable to consider when choosing the ANN architecture. From Figures A.2 (b) and (c), it can be observed that using a single hidden layer is not enough to capture the target first-order properties. The Helmholtz free energy can be similarly modeled using 2, 3, 4, or 5 layers when considering more than 20 neurons per layer. Based on this information, the FE-ANN EoS uses three hidden layers and 30 neurons per layer. This architecture offers sufficient flexibility to model the Helmholtz free energy without relying on an excessive number of parameters.

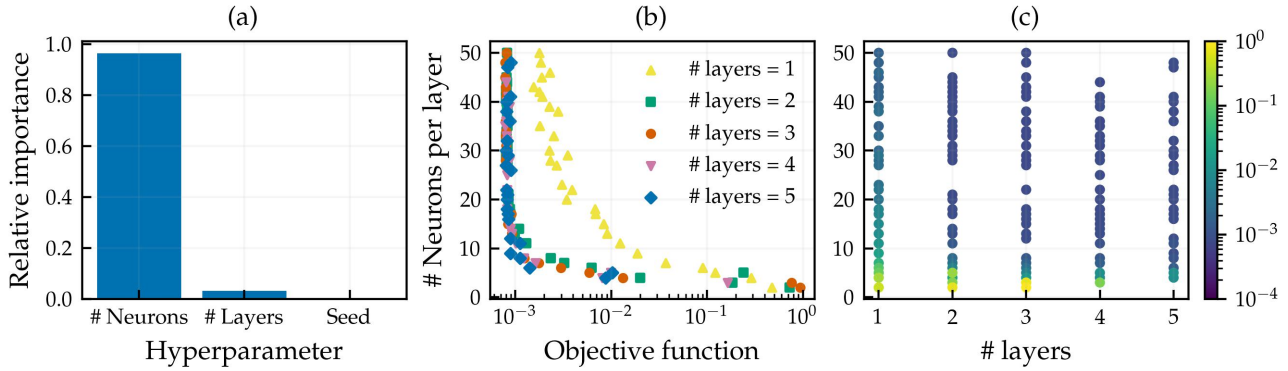


FIGURE A.2: **Hyperparameter optimization results of the FE-ANN EoS developed in Chapter 2.** (a) Hyperparameter relative importance. (b) Pareto frontier of the objective function against the number of neurons per layer. (c) Heatmap of the objective function for different ANN architectures.

### FE-ANN(s) EoS from Chapter 3

The hyperparameter optimization of the FE-ANN(s) EoS follows the same principles as the FE-ANN EoS described above. The architecture of the FE-ANN(s) EoS is optimized by analyzing the performance of different test models. These models are trained using a learning rate  $\alpha = 10^{-3}$  for 1000 epochs and considering only first-order thermophysical properties of the Mie particle. No second-order properties or penalty functions are considered in this optimization. As a note, the FE-ANN(s) EoS developed in Chapter 3 considers second-order properties and penalty functions when training. This extra information forces to use a lower learning rate and train for more epochs.

The hyperparameter optimization results for the FE-ANN(s) EoS are summarized in Figure A.3. In this case, it is also observed that the number of neurons per layer is the most important hyperparameter to consider. From Figures A.3 (b) and (c), it can be observed that the FE-ANN(s) EoS requires at least three hidden layers with 20 or more neurons to model the data accurately. The FE-ANN(s) EoS can use the same architecture as the FE-ANN EoS to model the Helmholtz free energy across the entire density space. The FE-ANN(s) EoS uses three hidden layers with 30 neurons each.



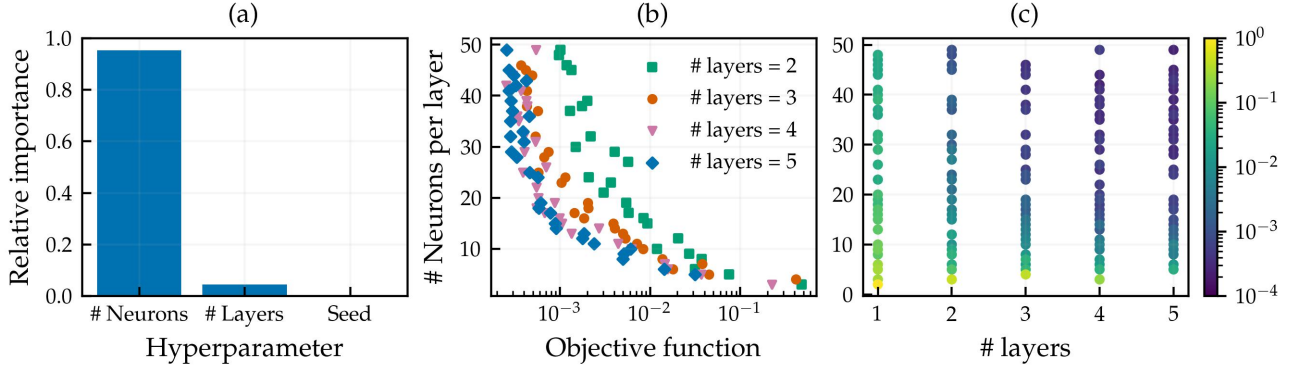


FIGURE A.3: **Hyperparameter optimization results of the FE-ANN(s) EoS developed in Chapter 3.** (a) Hyperparameter relative importance. (b) Pareto frontier of the objective function against the number of neurons per layer. (c) Heatmap of the objective function for different ANN architectures.

### Transport property models from Chapter 4

The architecture of the ANN-based models for transport properties is optimized by training different test models. The test models are trained using a learning rate  $\alpha = 10^{-3}$  for 1000 epochs and are trained directly on the transport property data. The target output features are  $\rho^* D^*$ ,  $\ln \eta^*$ , and  $\kappa^*$  for the self-diffusivity, shear viscosity, and thermal conductivity, respectively.

The results of the hyperparameter optimization are summarized in Figures A.4, A.5, and A.6. From these figures, it can be observed that several combinations of hidden layers and neurons per layer led to similar results. However, as shown in Figure (a), the number of neurons per layer is the most relevant hyperparameter. The architecture of each model is chosen so that transport properties can be adequately modeled while keeping the number of total adjustable parameters low. For this reason, the self-diffusivity and shear viscosity use two hidden layers with 30 neurons each, and the thermal conductivity uses three hidden layers and 30 neurons each.

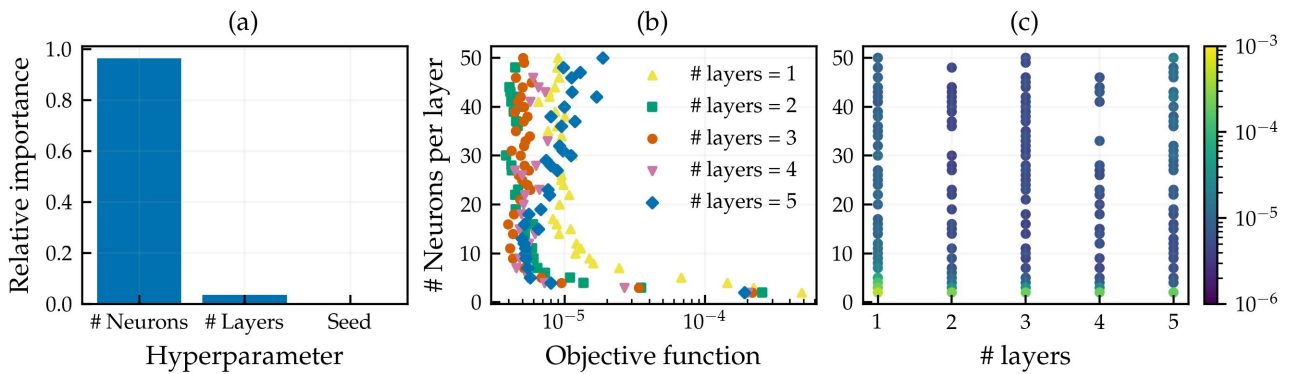


FIGURE A.4: **Hyperparameter optimization results of the self-diffusivity ANN-based model developed in Chapter 4.** (a) Hyperparameter relative importance. (b) Pareto frontier of the objective function against the number of neurons per layer. (c) Heatmap of the objective function for different ANN architectures.

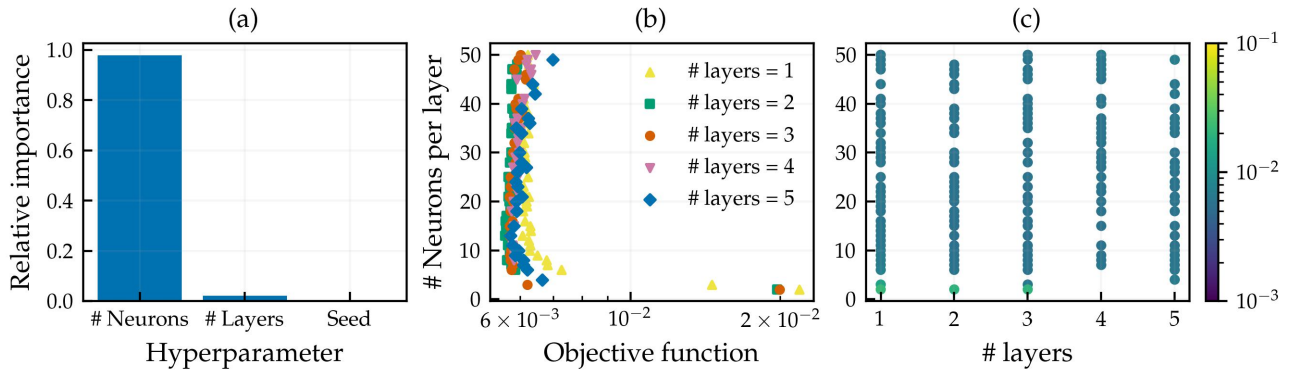


FIGURE A.5: **Hyperparameter optimization results of the shear viscosity ANN-based model developed in Chapter 4.** (a) Hyperparameter relative importance. (b) Pareto frontier of the objective function against the number of neurons per layer. (c) Heatmap of the objective function for different ANN architectures.

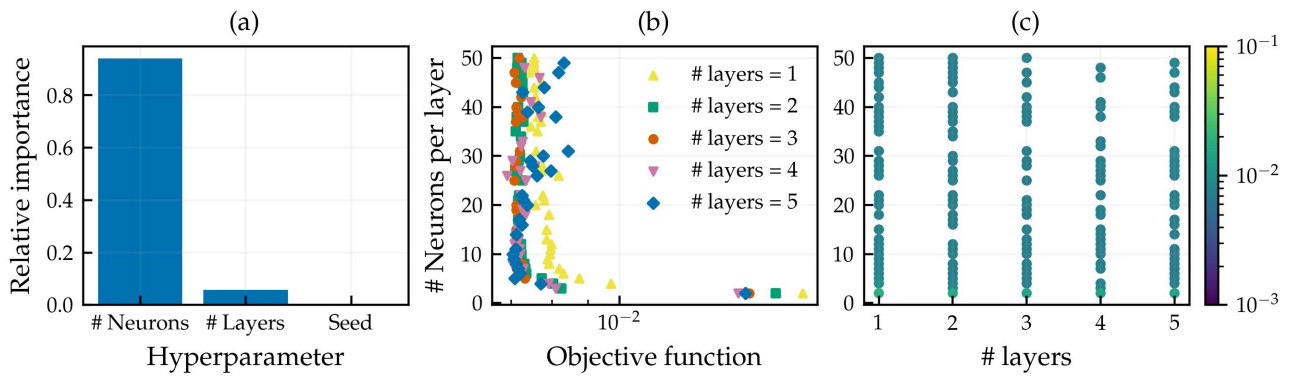


FIGURE A.6: **Hyperparameter optimization results of the thermal conductivity ANN-based model developed in Chapter 4.** (a) Hyperparameter relative importance. (b) Pareto frontier of the objective function against the number of neurons per layer. (c) Heatmap of the objective function for different ANN architectures.

## Appendix B

# Thermophysical properties modeling of selected Mie particles

This appendix aims to complement the thermophysical property modeling results of Chapters 2, 3, and 4 for additional Mie particles. These chapters focus on Mie particles with  $\lambda_r=12, 16, \text{ and } 20$ , and  $\lambda_a=6$ . This appendix considers Mie particles with  $\lambda_r=10, 18, \text{ and } 26$ , and  $\lambda_a=6$ .

- Section B.1 illustrates pressure isotherms, second-order properties isotherms, phase equilibria (VLE) results, and Brown's characteristic curves predicted by the FE-ANN EoS developed in Chapter 2.
- Section B.2 illustrates pressure isotherms, second-order properties isotherms, phase equilibria (VLE, SLE, and SVE) results, and Brown's characteristic curves predicted by the FE-ANN(s) EoS developed in Chapter 3.
- Section B.3 illustrates the self-diffusivity, shear viscosity, and thermal conductivity predicted by the ANN-based models developed in Chapter 4.

### B.1 FE-ANN EoS results for the $\lambda_r = 10, 18$ , and $26$ , and $\lambda_a = 6$ Mie fluids.

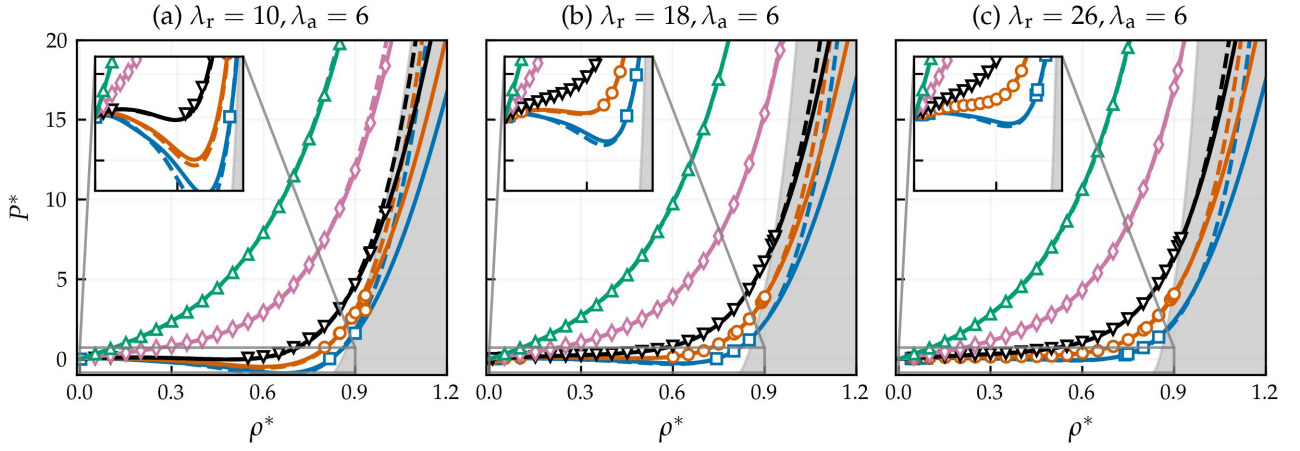


FIGURE B.1: **Pressure isotherms of selected  $(\lambda_r, \lambda_a = 6)$  Mie fluids.** (a)  $\lambda_r = 10, \lambda_a = 6$ . (b)  $\lambda_r = 18, \lambda_a = 6$ . (c)  $\lambda_r = 26, \lambda_a = 6$ . Molecular simulation data. Blue squares:  $T^* = 0.8$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results. The computed isotherms from the EoSs follow the same color convention as the simulation data. The grey-shaded area represents the region where freezing is predicted by the correlation of Khrapak et al. [151], see Eq. (2.18).

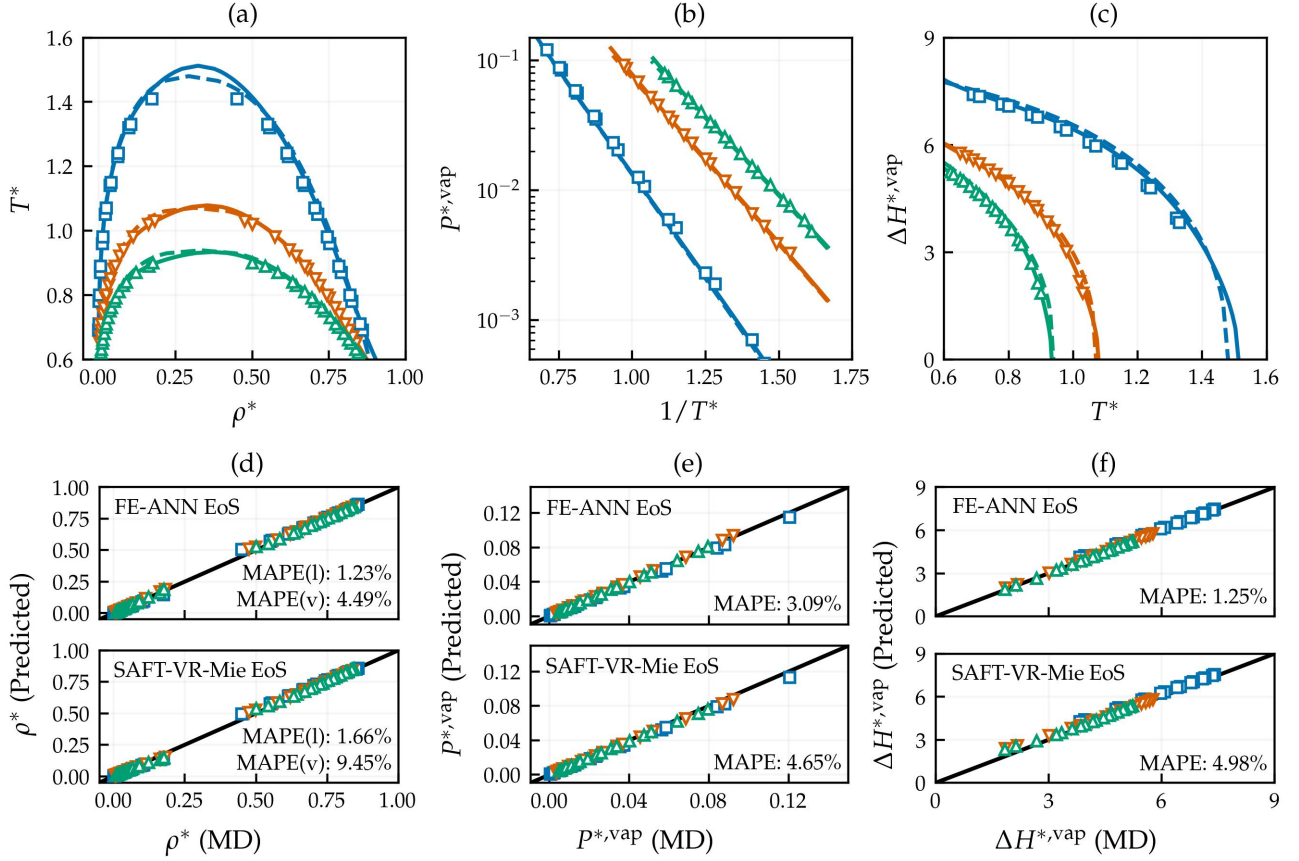


FIGURE B.2: **Phase equilibria results of selected  $(\lambda_r, \lambda_a = 6)$  Mie fluids.** (a) Phase envelope. (b) Vaporization pressure. (c) Vaporization enthalpy. Molecular simulation data. Blue squares:  $\lambda_r = 10, \lambda_a = 6$ , orange downward triangles:  $\lambda_r = 18, \lambda_a = 6$ , and green upright triangles:  $\lambda_r = 26, \lambda_a = 6$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{\text{vdw}}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results. These VLE results are obtained by numerically solving the system of Eqs. (2.7). The predicted phase equilibria from the EoSs follow the same color convention as the simulation data. Parity plots. (d) Equilibrium densities. (e) Vaporization pressure. (f) Vaporization enthalpy. The top subplot refers to the FE-ANN EoS results, and the bottom subplot shows the SAFT-VR-Mie results. The symbols and colors follow the same convention as in (a), (b) and (c). The reported Mean Absolute Percentage Error (MAPE) is obtained as:  $\text{MAPE} = 100/N_P \sum_j^{N_P} |(y_j^{\text{pred}} - y_j^{\text{true}})/y_j^{\text{true}}|$ .

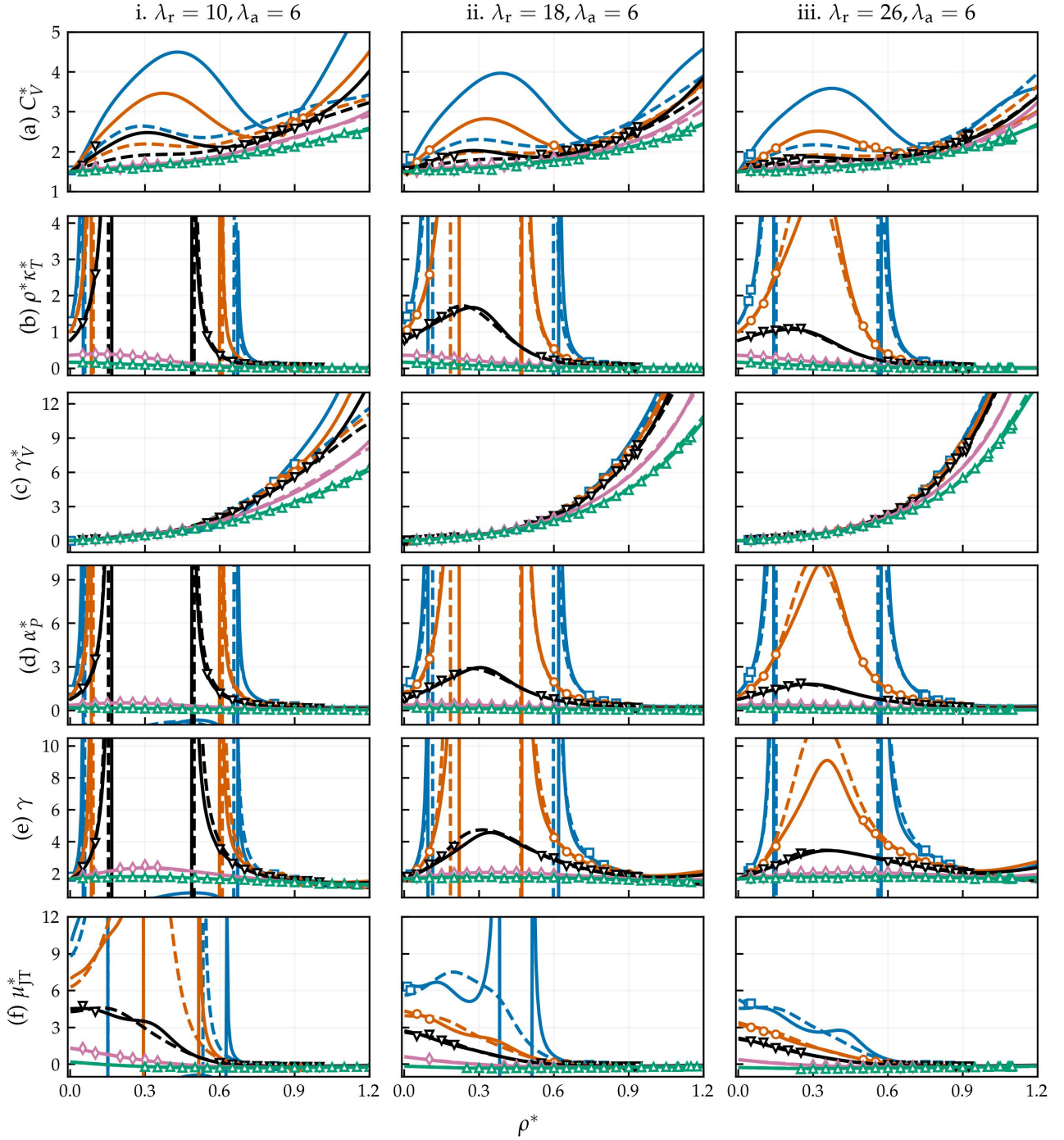


FIGURE B.3: **Second-order derivative properties isotherms of selected  $(\lambda_r, \lambda_a = 6)$  Mie fluids.** i.  $\lambda_r = 10, \lambda_a = 6$ . ii.  $\lambda_r = 18, \lambda_a = 6$ . iii.  $\lambda_r = 26, \lambda_a = 6$ . (a) Isochoric heat capacity ( $C_V^*$ ). (b) Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). (c) Thermal pressure coefficient ( $\gamma_V^*$ ). (d) Thermal expansion coefficient ( $\alpha_P^*$ ). (e) Adiabatic index (heat capacity ratio,  $\gamma = C_P^*/C_V^*$ ). (f) Joule-Thomson coefficient ( $\mu_{JT}^*$ ). Molecular simulation data. Blue squares:  $T^* = 0.8$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$ . Solid lines are results of the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Dashed lines are SAFT-VR-Mie EoS [37] results.

The computed isotherms from the EoSs follow the same color convention as the simulation data.



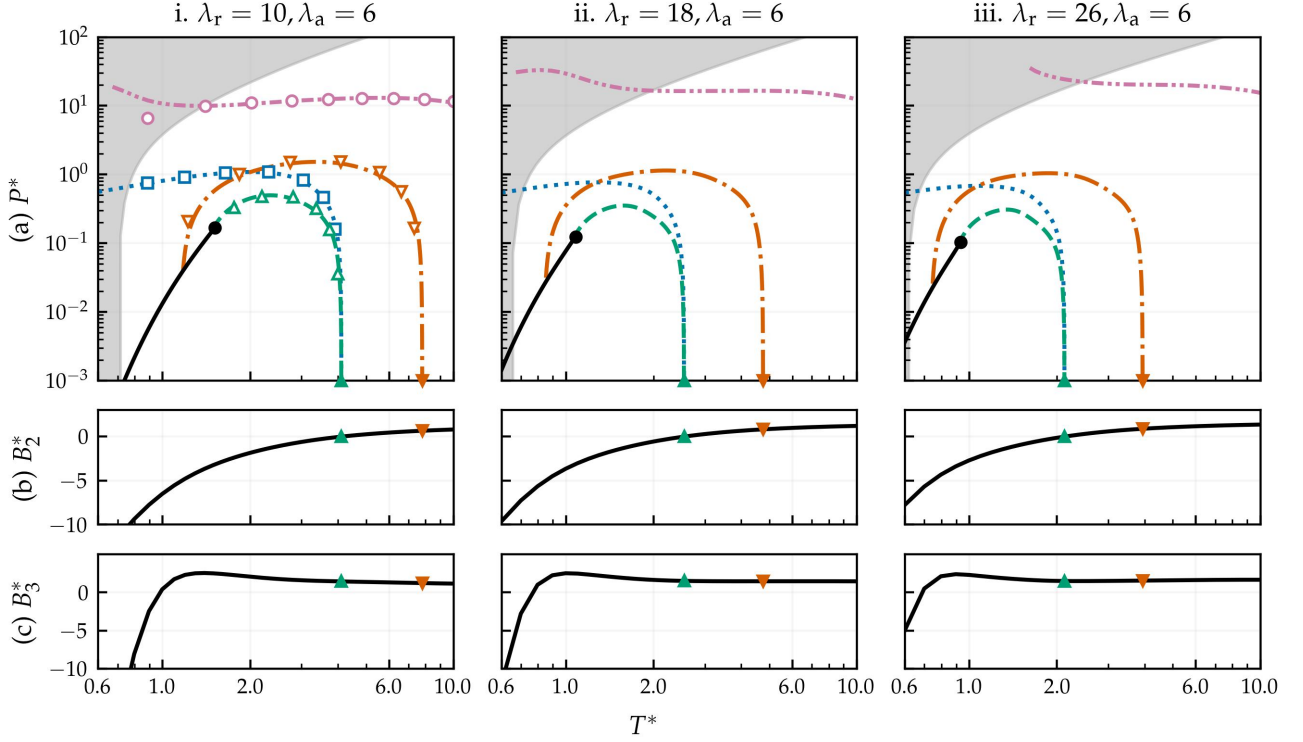


FIGURE B.4: **Brown's characteristic curves and second virial coefficient of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** (a)  $\lambda_r = 10, \lambda_a = 6$ . (b)  $\lambda_r = 18, \lambda_a = 6$ . (c)  $\lambda_r = 26, \lambda_a = 6$ . Empty symbols are molecular simulation results obtained by Stephan and Urschel [166]. Pink circles: Amagat curve, orange downward triangles: Charles curve, green upright triangle: Boyle curve, and blue squares: Zeno curve. Lines with filled symbols are results obtained from the FE-ANN EoS ( $\alpha_{vdw}, \rho^*, 1/T^*$ ) (Eq. (2.22b)) trained using a relative weight  $w_i = 0.05$  for second-order properties. Bottom plot: the black solid line is the third virial coefficient. Middle plot: the black solid line is the second virial coefficient. Top plot: the black solid line is the vaporization pressure, the black filled circle is the critical point, the pink dot-dot-dashed line is the Amagat curve, the orange dot-dashed line is the Charles curve, the green dashed line is the Boyle curve, and the blue dotted line is the Zeno curve. The filled green upright triangle is the Boyle temperature, and the orange downward triangle is the Charles temperature. The grey-shaded area represents the region where freezing is predicted by the correlation of Khrapak et al. [151], see Eq. (2.18).

## B.2 FE-ANN(s) EoS results for the $\lambda_r=10, 18$ , and $26$ , and $\lambda_a=6$ Mie particles.

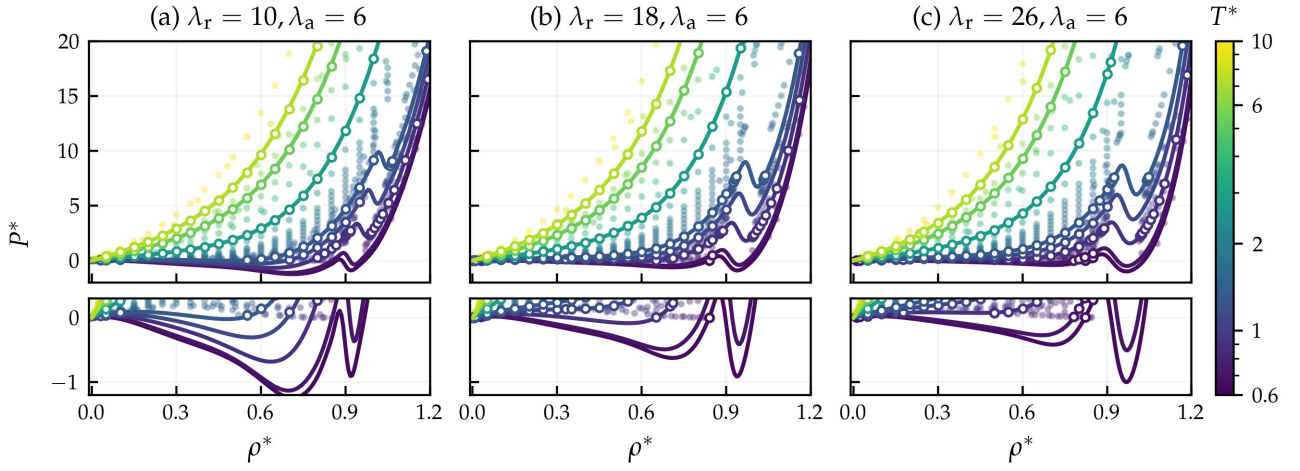
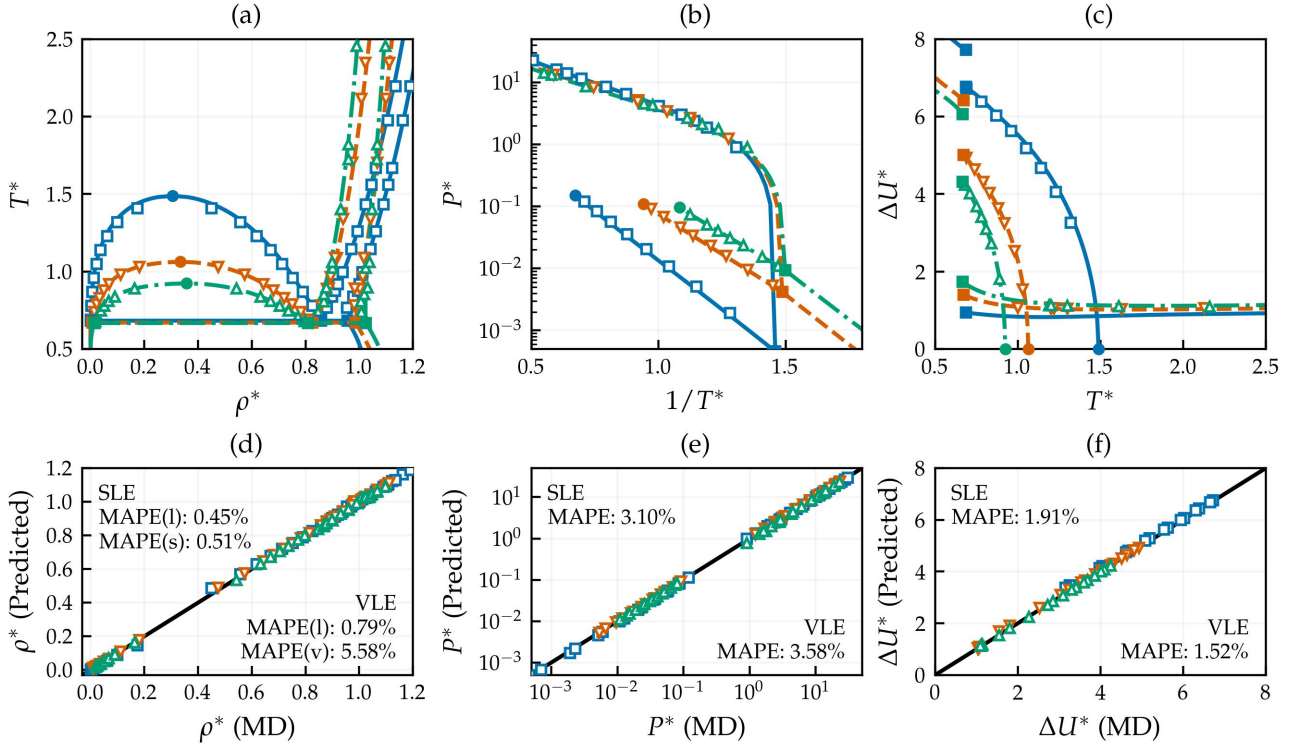


FIGURE B.5: **Pressure isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** (a)  $\lambda_r = 10, \lambda_a = 6$ . (b)  $\lambda_r = 18, \lambda_a = 6$ . (c)  $\lambda_r = 24, \lambda_a = 6$ . Solid lines represent the following isotherms:  $T^* = 0.65, 0.7, 0.9, 1.1, 1.3, 2.8, 5.3$ , and  $7.2$  obtained with the FE-ANN(s) EoS. Solid empty circles represent the molecular dynamics (MD) result of the aforementioned isotherms. Transparent circles represent the data points available from the database. The coloring of lines and symbols refers to the temperature as indicated by the color bar.





$$\text{MAPE} = 100/N_P \sum_j^{N_P} |(y_j^{\text{pred}} - y_j^{\text{true}})/y_j^{\text{true}}|.$$

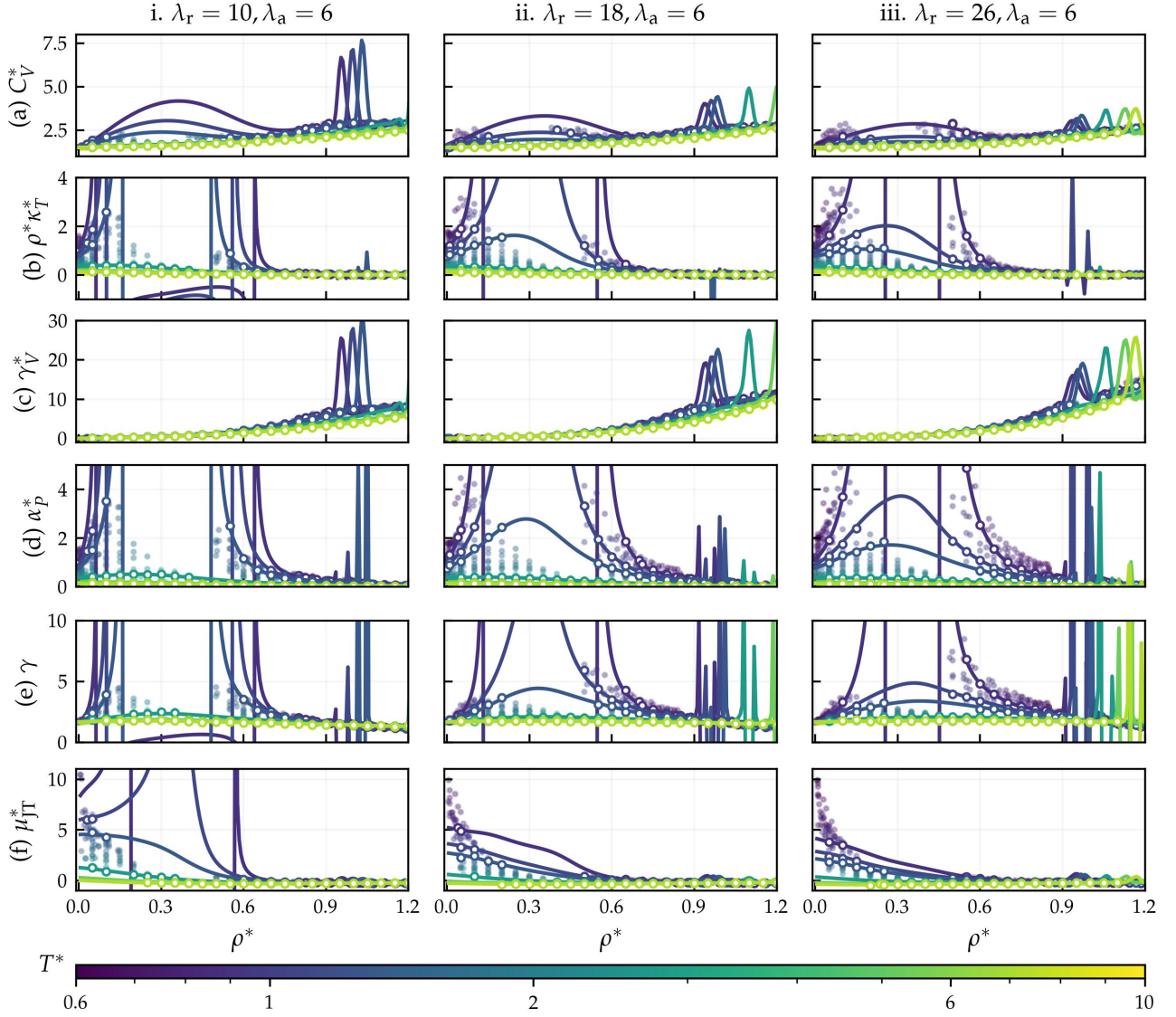


FIGURE B.7: **Second-order properties of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** **i.**  $\lambda_r = 10, \lambda_a = 6$ . **ii.**  $\lambda_r = 18, \lambda_a = 6$ . **iii.**  $\lambda_r = 26, \lambda_a = 6$ . **(a)** Isochoric heat capacity ( $C_V^*$ ). **(b)** Isothermal compressibility (as  $\rho^* \kappa_T^*$ ). **(c)** Thermal pressure coefficient ( $\gamma_V^*$ ). **(d)** Thermal expansion coefficient ( $\alpha_P^*$ ). **(e)** Adiabatic index (heat capacity ratio,  $\gamma = C_P^*/C_V^*$ ). **(f)** Joule-Thomson coefficient ( $\mu_{JT}^*$ ). Solid lines represent the following isotherms:  $T^* = 0.9, 1.1, 1.3, 2.8, 5.3$ , and  $7.2$  obtained with FE-ANN(s) EoS. Solid empty circles represent the molecular dynamics (MD) result of the aforementioned isotherms. Transparent circles represent the data points available from the database. The coloring of lines and symbols refers to the temperature as indicated by the color bar.

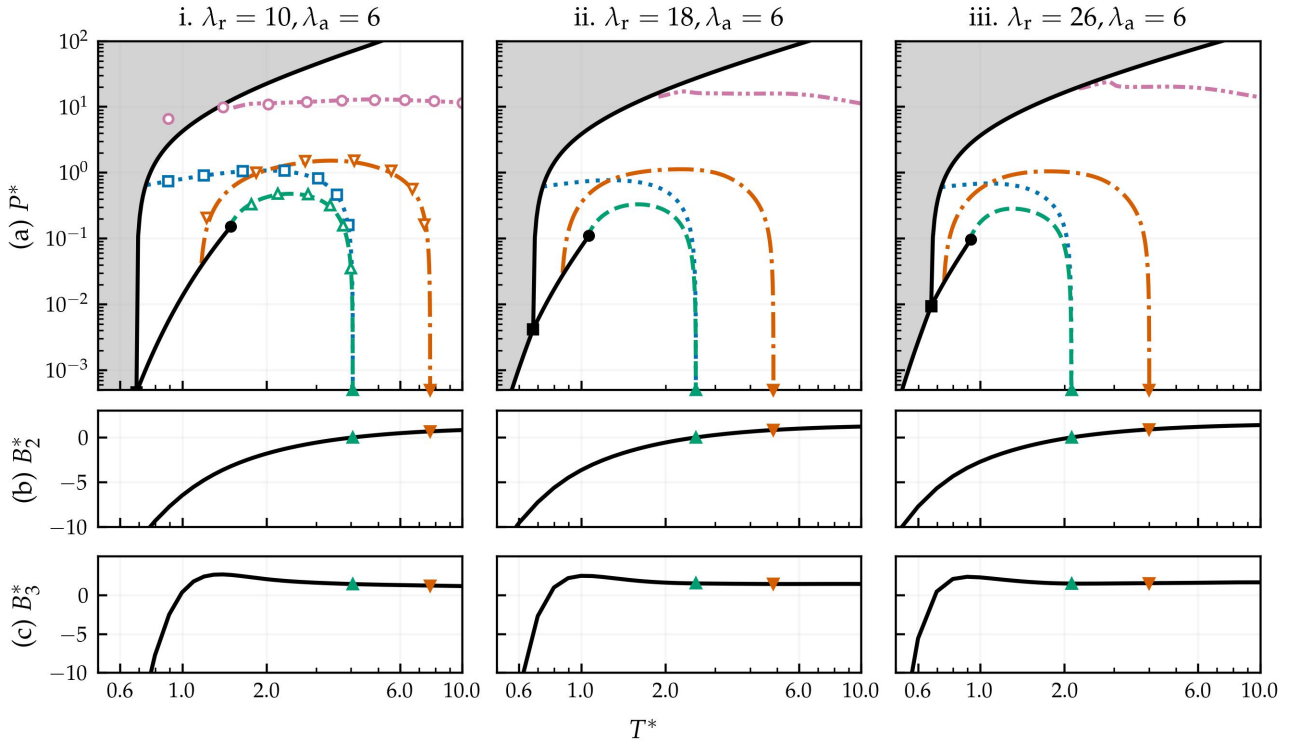


FIGURE B.8: **Brown's characteristic curves and second virial coefficient of selected ( $\lambda_r, \lambda_a = 6$ ) Mie particles.** i.  $\lambda_r = 10, \lambda_a = 6$ . ii.  $\lambda_r = 18, \lambda_a = 6$ . iii.  $\lambda_r = 26, \lambda_a = 6$ . Lines and filled symbols are results obtained with the FE-ANN(s) EoS. **(a)** Brown characteristic curves. The black solid lines represent equilibrium pressure (SVE, SLE, and VLE), blue dotted line is the Zeno curve, green dashed line is the Boyle curve, orange dot-dashed line is the Charles curve, and pink dot-dot-dashed line is the Amagat curve. The filled black square represents the triple point, the filled black circle represents the critical point, the filled upright green triangle is the Boyle temperature, and the filled downward orange triangle is the Charles temperature. The grey-shaded area represents the solid region. Empty symbols are molecular simulation results obtained by Stephan and Urschel [166]. **(b)** Second virial coefficient. **(c)** Third virial coefficient.

### B.3 Transport properties results for the $\lambda_r=10, 18$ , and $26$ , and $\lambda_a=6$ Mie fluids.

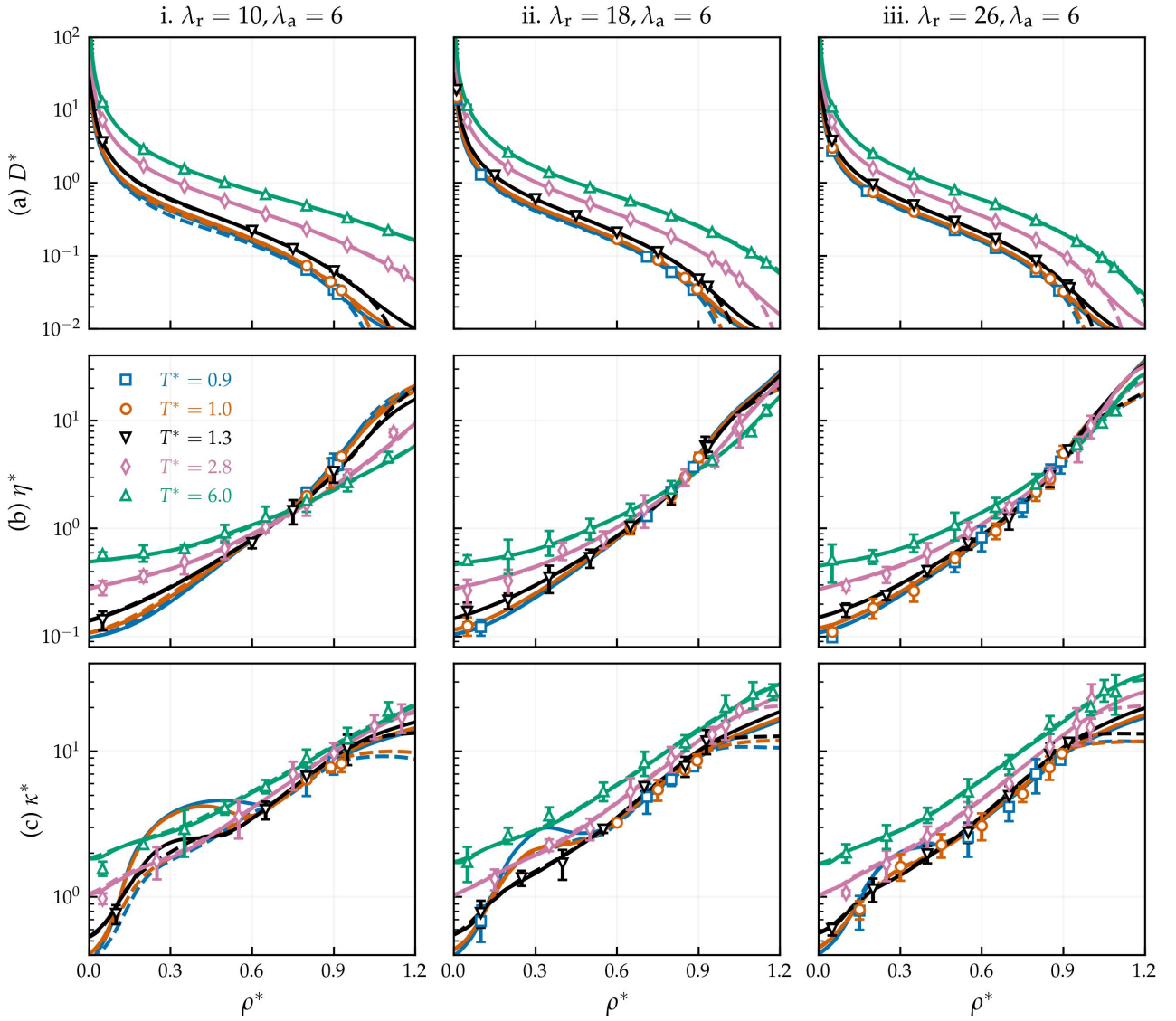


FIGURE B.9: **Transport properties isotherms of selected ( $\lambda_r, \lambda_a = 6$ ) Mie fluids.** i.  $\lambda_r = 10, \lambda_a = 6$ . ii.  $\lambda_r = 18, \lambda_a = 6$ . iii.  $\lambda_r = 26, \lambda_a = 6$ . (a) Self-diffusivity. (b) Shear viscosity. (c) Thermal conductivity. Molecular simulation data. Blue squares:  $T^* = 0.9$ , orange circles:  $T^* = 1.0$ , black downward triangles:  $T^* = 1.3$ , pink diamonds:  $T^* = 2.8$  and green upright triangles:  $T^* = 6.0$  Solid lines refer to the  $\rho^* D^*$ ,  $\ln \eta^*$  and  $\ln \kappa^*$  models, Eq. (4.9a). Dashed lines refer to the residual  $\rho^* D^{*,res}$ ,  $\ln \eta^{*,res}$  and  $\ln \kappa^{*,res}$  models, Eq. (4.14). The computed isotherms from the ANN-based models follow the same color convention as the simulation data.

## Appendix C

# Further parametrization results of quasi-spherical molecules

This appendix presents the modeling results for Krypton (Figure C.1), Xenon (Figure C.2), and Carbon Monoxide (Figure C.3). These figures present two possible modeling approaches that use different molecular parameter sets. Figures (a), (b), and (c) show VLE results predicted by the FE-ANN EoS developed in Chapter 2. Figures (d) to (g) show phase equilibria results (VLE, SLE, and SVE) obtained using the FE-ANN(s) EoS developed in Chapter 3. This EoS is also used to predict the speed of sound in Figure (g). Finally, the shear viscosity and thermal conductivity results obtained using the ANN-based models from Chapter 4 are shown in Figures (h) and (i).

The following parametrization approaches were considered.

- $OF_1$ : VLE only
- $OF_2$ : VLE + shear viscosity
- $OF_3$ : VLE + SLE + SVE

Refer to Chapter 5 for a discussion of the impact of different thermophysical properties on the resulting molecular parameters.

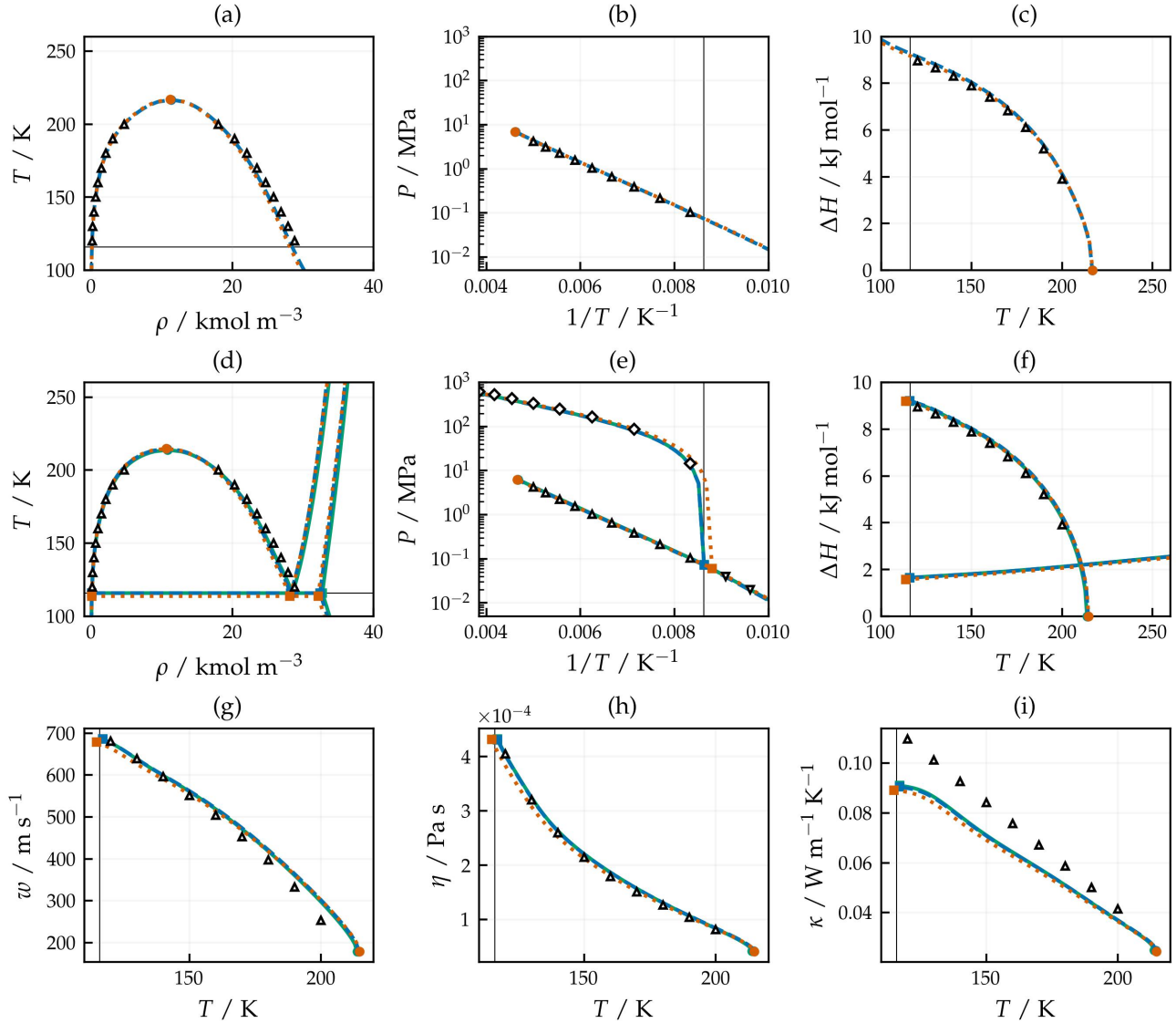


FIGURE C.1: **Selected thermophysical properties of Krypton.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.



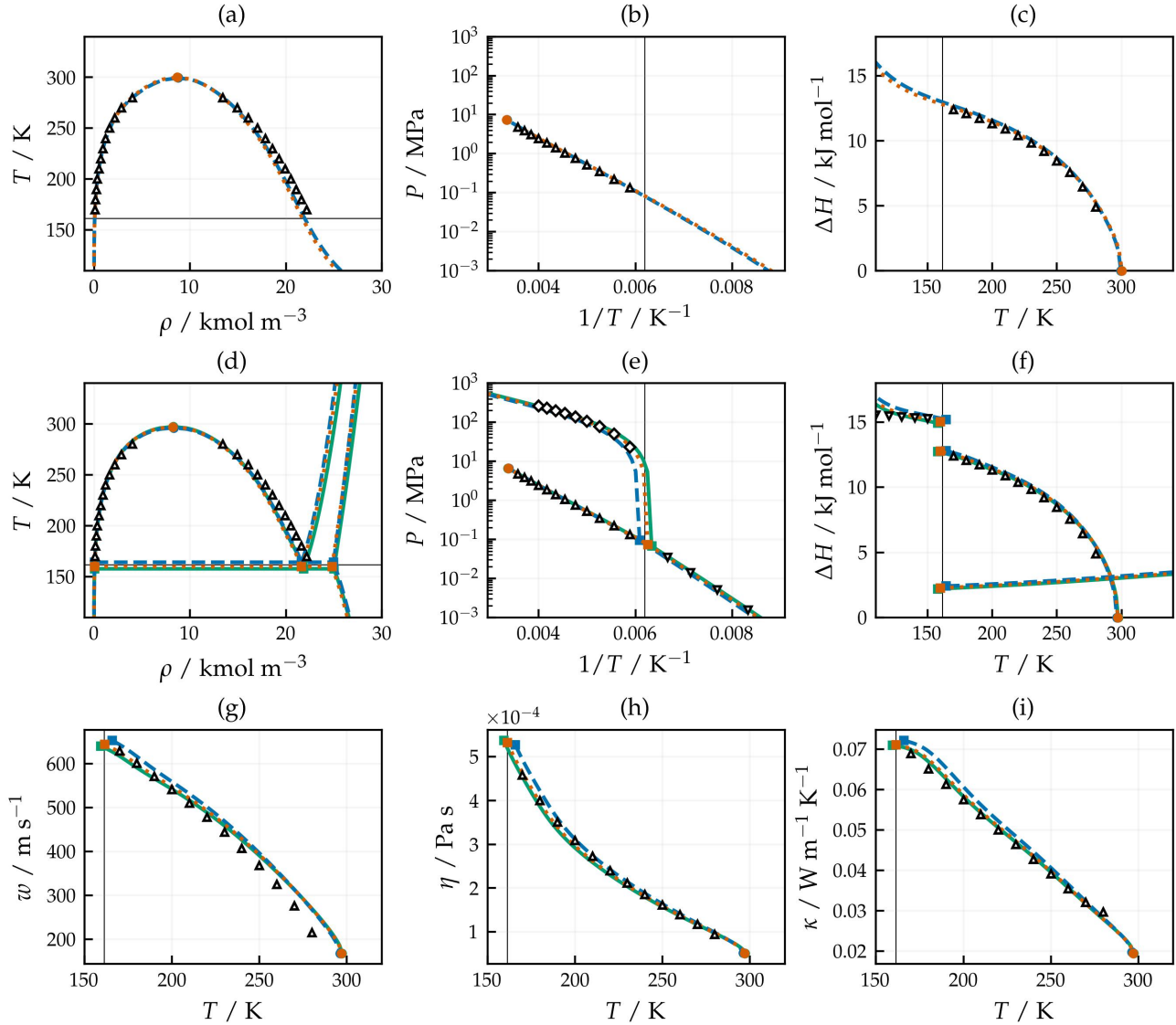


FIGURE C.2: **Selected thermophysical properties of Xenon.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with OF<sub>1</sub> (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with OF<sub>2</sub> (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with OF<sub>3</sub> (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.

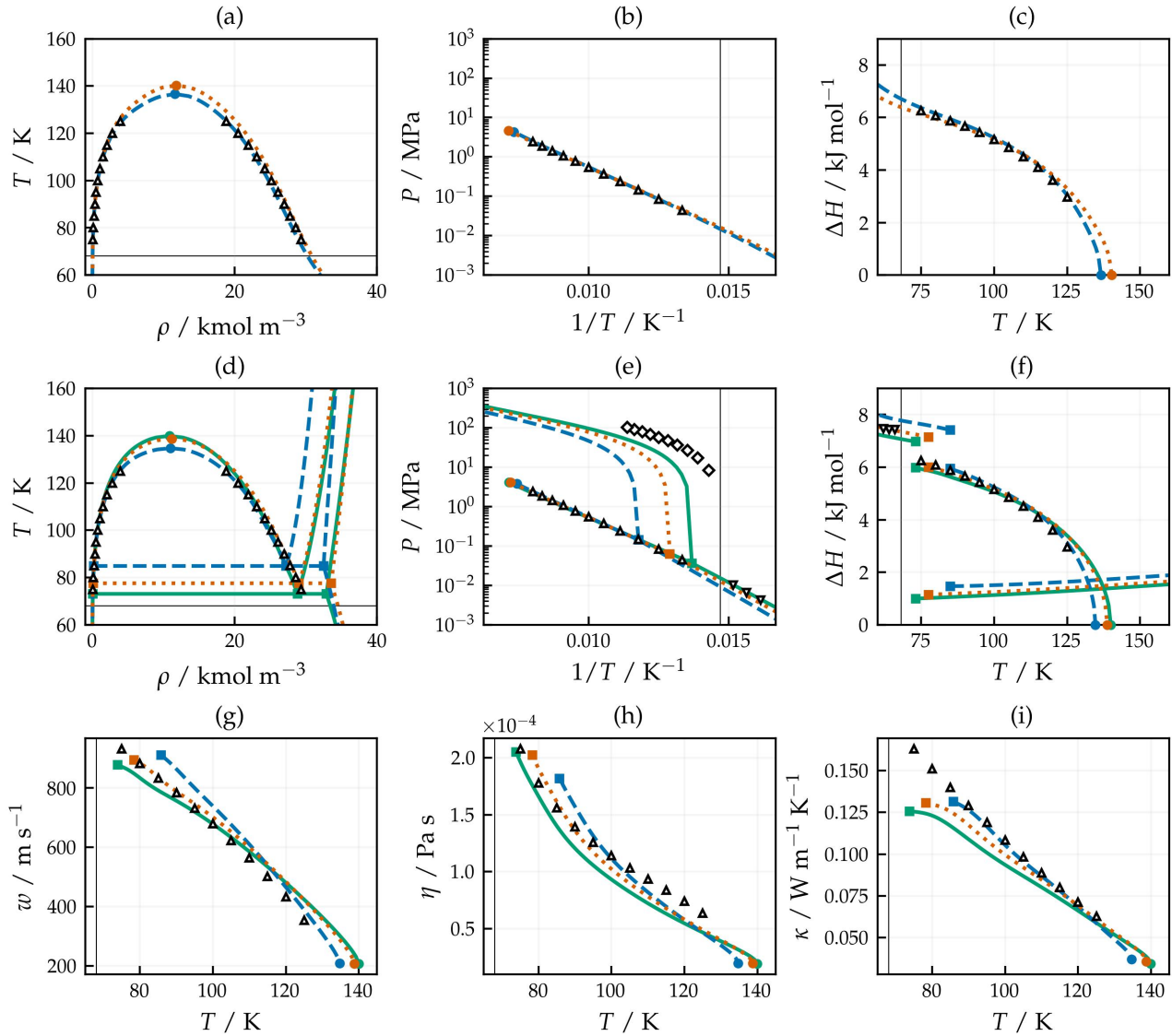


FIGURE C.3: **Selected thermophysical properties of Carbon monoxide.** (a) and (d) Phase envelope. (b) and (e) Clapeyron plot. (c) and (f) Enthalpy of phase change. (g) Speed of sound. (h) Shear viscosity. (i) Thermal conductivity. (a) - (c) Phase equilibria modeled using the FE-ANN EoS developed in Chapter 2. (d) - (g) Phase equilibria and speed of sound modeled using the FE-ANN(s) EoS developed in Chapter 3. FE-ANN/FE-ANN(s) EoSs are based on Eq. (5.2). (h) and (i) The shear viscosity and thermal conductivity are modeled using the ANN-based models developed in Chapter 4, based on Eqs. (5.4) and (5.9). Dashed blue lines and symbols refer to results using parameters optimized with  $OF_1$  (VLE only). Dotted orange lines and symbols refer to results using parameters optimized with  $OF_2$  (VLE + shear viscosity). Solid green lines and symbols refer to results using parameters optimized with  $OF_3$  (VLE + SLE + SVE). Refer to Eq. (5.1) and Table 5.1 for further details about the objective functions. Filled circle is the critical point. Filled square is the triple point. Pseudo-experimental data obtained from NIST TRC [12]. Upright triangles are VLE data, diamonds are SLE data, and downward triangles are SVE data. For reference, the solid black line refers to the triple temperature.