IMPERIAL

Towards the Autonomous Molecular Design and Characterisation of Lubricant Base Oils and Additives

By Mark Egheosa Chesley Iyen Ogbomo

A Thesis Submitted for the Degree of Doctor of Philosophy

June 2025

Imperial College London
Department of Mechanical Engineering

Declaration of Originality

I, Egheosa Ogbomo, declare that, to the best of my knowledge, the work in this thesis is my own. The work of others has been appropriately referenced. A full list of references is given in the bibliography. This thesis has not been submitted for the award of any other degree.

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence (CC BY NC-SA).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that; you credit the author, do not use it for commercial purposes and share any derivative works under the same licence.

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. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

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.

Abstract

Improving the performance of lubricant formulations is critical for enhancing the energy efficiency and durability of machine components. As internal combustion engines are progressively replaced by electric motors, lubricant formulations must meet new operational demands, such as improved thermal stability, electrical compatibility, and reduced environmental impact. Despite these needs, the discovery and optimisation of lubricant formulations remains largely empirical and dependent on costly trial-and-error experimentation. This thesis attempts to address these challenges through the development and application of computational methods to accelerate the molecular-level characterisation and design of lubricant base oils, and provide a deeper understanding of the nanoscale behaviour of a prevalent lubricant additive.

The first part of this thesis focusses on using molecular dynamics (MD) simulations to explore the mechanochemical decomposition mechanisms of tricresyl phosphate (TCP), a widely used anti-wear additive, under varying surface chemistries, temperatures, sliding speeds, and shear stresses.

The second part of this thesis explores the use of machine learning (ML) models for the prediction of hydrocarbon transport properties, assessing the impact of molecular representation, model selection, and dataset generation using MD simulations. The performance of these ML models is benchmarked against experimental measurements and MD simulations, and their limitations in generalising to out-of-distribution data are examined.

Finally, a multi-objective optimisation framework is developed that integrates interpretable ML with evolutionary algorithms to propose novel base oil molecules with enhanced properties compared to current industrial lubricant base oils.

The methods and findings presented in this thesis contribute to a deeper molecular-level understanding of lubricant base oil and additive behaviour, and demonstrate the potential of computational approaches to guide the design of next-generation high-performance and environmentally-friendly lubricants.

Dedication

Like so many others, my PhD experience has been a challenging one, financially, emotionally and socially. However, the breadth of experiences I've had and the plethora of people that I've met during my studies has made my experience feel truly unique and rewarding. For this, I must thank my academic supervisors Prof. Daniele Dini and Dr. James P. Ewen for giving me this opportunity in the first place, and for affording me the freedom to explore so many things beyond the main work that forms this thesis. Thank you also, James and Daniele, for your invaluable guidance, patience, trust, encouragement and expansive knowledge of the field, without which I would not even have a PhD thesis to dedicate.

Thank you to the staff and students of Imperial's Tribology group for maintaining such a welcoming and healthy academic environment, complete with lots of tasty food from around the world and Imperial's best five-a-side team. Thanks to Dr. Amran Mohammed, Dr. Carlos Ayestaràn Latorre, Dr. Jagjeevan Bhamra, Dr. Manoj Murali and Dr. Stavros Ntioudis who have made my time within the group especially memorable.

I was lucky enough to help run the Black Doctoral Network, which made me and so many others feel comfortable enough to express our authentic selves while studying at Imperial. For making the BDN such a wonderful reality, I must thank Kemi Oloyede, Maya Thabal-Herron, Naima Adan and Peter Bryan, as well as our chief supporters, Prof. Sara Rankin, Dr. Sunday Popo-Ola and Provost Dr. Wayne Mitchell.

I managed to try my hand in the start-up world too during Imperial's 2023 Venture Catalyst Challenge, an experience which has left me with a near unshakeable belief in myself. Thank you to those members of the Imperial Enterprise lab who facilitated this transformative experience for me, especially Euan Bell and Emily Goddard.

I've spent two years as a residential subwarden, which taught me that university life is so much more than just tests and dissertations. Thanks again to Naima Adan for the opportunity, and to all the members of Princes Gardens wardening team for making the experience so fulfilling.

Thank you to my wonderful friends: Antonia, Gideon, Jamie, Jean-Paul, Richard and all the others I don't have the space to mention by name, for reminding me that I have a wonderful support network, and that work must be accompanied by play.

To my wonderful partner, Kathryn Plummer, thank you for your patience, never-ending support, humour, tenderness and trustworthiness. I'm acutely aware that it is a rarity to find a partner with whom one is so compatible and can earnestly imagine a long and happy life with. I will not take what we have for granted. Thank you also to your family for embracing me as one of their own and providing me with so many fond memories. Let this dedication serve as a permanent declaration: Bubbles, you have an immutable place in my heart and I love you profoundly.

On the topic of love, my biggest thanks must go to those who showed me what love is: my father Duke, my mother Dele, and my brother Osasuyi. Thank you family, for keeping me grounded and giving me

purpose. Thank you Osasuyi, for always keeping me honest, helping the family when we needed it most, and inspiring me to be more cool. Thank you Mum and Dad for sacrificing so much for me and Osasuyi, to give us both the opportunities we've been so fortunate to have. Understanding only a fraction of the adversity that you have overcome to get our family to where we are today, has motivated me to persevere through countless challenging days and nights during this PhD. To you, my family, let my words serve not just as a thank you but as a series of promises. A promise that I'll never forget the timeless values of hard work, resilience, healthy ambition and adaptability, among countless others, that you've instilled in me. A promise to always place importance on family. A promise to always believe that God's timing is the best.

Contents

1 Introduction **16**

1.1 Motivation . . . 16

1.2 Research Goals . . . 17

1.3 Layout of Thesis . . . 18

2 Literature Review **19**

2.1 The Importance of Molecular Modelling and ML in Lubricant Design and Tribology . . . 19

2.2 Background on Lubrication/Tribology . . . 20

2.2.1 Lubrication Regimes and Important Lubricant Properties . . . 20

2.2.2 Base Oils . . . 22

2.2.3 Additives . . . 24

2.2.4 Surfaces . . . 25

2.2.5 Mechanochemistry Theory . . . 26

2.3 Molecular Dynamics . . . 27

2.3.1 The Need for MD . . . 27

2.3.2 Different Types of MD Simulation . . . 28

2.3.3 Force fields . . . 30

2.4 Genetic Algorithms . . . 33

2.4.1 Non-dominated Sorting Genetic Algorithm II . . . 34

2.4.2 Weighted Sum Genetic Algorithms . . . 36

2.5 Artificial Intelligence for Molecular Generation and Molecular Property Prediction . . . 37

2.5.1 AI and ML for Molecular Candidate Generation/Screening . . . 37

2.5.2 ML for Property Prediction . . . 38

2.5.3 Data Processing For ML . . . 43

2.6 Summary . . . 45

3 Effects of Surface Chemistry on the Mechanochemical Decomposition of Tricresyl Phosphate **46**

3.1 Introduction . . . 46

3.2 Methodology . . . 49

3.2.1 System Setup . . . 49

3.2.2 Simulation Details . . . 51
3.3 Results and Discussion . . . 53
3.3.1 Base Oil Effects on TCP Decomposition . . . 53
3.3.2 Surface Chemistry Effects on Shear Stress . . . 55
3.3.3 Surface Chemistry Effects on TCP Dissociation . . . 56
3.3.4 TCP Decomposition Mechanisms . . . 58
3.3.5 Surface Chemistry Effects on Polyphosphate Growth . . . 59
3.3.6 Rate Analysis . . . 61
3.3.7 3D Shear Stress-Temperature-Rate Constant Fits . . . 63
3.4 Conclusions . . . 65

4 Effect of Sliding Velocity on the Decomposition of Tricresyl Phosphate 67
4.1 Introduction . . . 67
4.2 Methodology . . . 70
4.2.1 DLC System Setup . . . 70
4.2.2 Simulation Details . . . 71
4.2.3 Force Field Details and Validation . . . 72
4.2.4 Bond Length Analysis . . . 72
4.3 Results and Discussion . . . 73
4.3.1 Nature of TCP Dissociation at Different Sliding Velocities . . . 73
4.3.2 Sliding Velocity Effects on Shear Stress . . . 77
4.3.3 Sliding Speed Effects on TCP Dissociation . . . 78
4.3.4 Sliding Speed Effects on Polyphosphate Formation and TCP Bond Lengths . . . 83
4.4 Conclusions . . . 88

5 A Comparison of Molecular Dynamics and Machine Learning Techniques for The Prediction of Lubricant Transport Properties 89
5.1 Introduction . . . 89
5.2 Methodology . . . 92
5.2.1 Choice of benchmarking molecules . . . 92
5.2.2 Fluid Transport Property Calculation using Equilibrium Molecular Dynamics . . . 93
5.2.3 ML Model Datasets . . . 96
5.2.4 Molecular Representations . . . 98
5.2.5 ML Model Training . . . 101
5.3 Results and Discussion . . . 105
5.3.1 Embeddings and Fingerprint Optimisation . . . 105
5.3.2 Best Representations for ML and best ML Performance . . . 110
5.3.3 A Comparison of ML and MD Techniques for the Prediction of Hydrocarbon Thermophysical Properties . . . 114
5.3.4 MD versus Experimental Dataset Model Performance . . . 117
5.3.5 Out Of Distribution (OOD) Prediction . . . 120

5.4 Conclusions . . . 124

6 Evolutionary Algorithms and Explainable Machine Learning for the Multi-Objective Optimisation of Lubricant Base Oils 127

6.1 Introduction . . . 127
6.2 Methodology . . . 131
6.2.1 Genetic Algorithm Implementation . . . 131
6.2.2 Genetic Algorithm Fitness Functions . . . 136
6.2.3 Saliency Study Techniques . . . 140
6.3 Results and Discussion . . . 141
6.3.1 SHAP Analysis . . . 141
6.3.2 Other Design Considerations and Constraints for Molecular Generation . . . 150
6.3.3 Genetic Algorithm Hyperparameter Selection . . . 152
6.3.4 Best Performing Candidates from WSGA Runs . . . 155
6.4 Conclusions . . . 158

7 Conclusions, Limitations and Further Work 159

7.1 Chapter 3 . . . 159
7.2 Chapter 4 . . . 161
7.3 Chapter 5 . . . 162
7.4 Chapter 6 . . . 164

Bibliography 167

A Supporting information for Chapter 3 206

B Supporting information for Chapter 4 214

C Supporting information for Chapter 5 220

D Supporting information for Chapter 6 230

List of Figures

2.1 Stribeck Curve showing different lubrication regimes. Reproduced from *'Tribology: friction and wear of engineering materials'*.[33] . . . 21

2.2 The activation barrier for a reaction a) in the absence of an external applied force and b) under the influence of an applied external force. Reproduced from *'Stress-augmented thermal activation: Tribology feels the force'*.[88] . . . 26

3.1 Molecular structure of tri-meta-cresyl phosphate. . . . 48

3.2 Side-view snapshots showing the model systems with TCP molecules confined between a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001)-SCV surfaces. Systems shown after minimisation and equilibration, but before heating, compression, and shear. Black dashed line shows the simulation box. Rendered using OVITO.[306] . . . 50

3.3 Change in the number of intact TCP molecules with time at 600 K and 1-5 GPa for a) 100% TCP system b) 50% TCP, 50% DIDA system. The dashed lines show the simulation data and the solid lines are exponential decay fits used to calculate the first-order rate constants, k. c) Change in k for TCP decomposition with shear stress, τ. The data with and without DIDA can be fit using Equation 1 with a single activation volume (13.6 ± 2.6 Å^3). Inset shows the same data versus $\ln(k)$. . . . 54

3.4 Variation in shear stress with changes in normal stress for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) surfaces. . . . 55

3.5 Number of intact TCP molecules with time at 4 GPa between 400-900 K (top row) and at 600 K between 1-5 GPa (bottom row) for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) surfaces. The dashed lines show the simulation data and the solid lines are exponential decay fits used to calculate the first-order rate constants, k. . . . 56

3.6 Change in the number of Fe-C bonds with simulation time at 600 K between 1-5 GPa for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(100). . . . 57

3.7 Variation in $\ln(k)$ for TCP decomposition with the change in the number of Fe-C bonds from 0-1 ns for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(100). . . . 58

3.8 Change in the number of P-O and C-O bonds with sliding time at 500 K and 3 GPa for the a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) systems. Solid lines show the change in the total number of bonds compared to TCP, dashed lines show bond formation and dotted lines show bond cleavage. . . . 58

3.9 Comparison of oligomer, toluene, and cresol quantities at 500 K and 3 GPa for the a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) systems. 60
3.10 Change in the number of polyphosphate chains with sliding time at 500 K and 3 GPa for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) systems. 60
3.11 Shear stress dependence of the TCP decomposition rate at different temperatures for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(100) systems. Insets show the same data with a logarithmic *y*-axis. 61
3.12 Kinetic compensation effect between the activation energy and pre-exponential factor for Iron, Iron Carbide, and Iron Oxide 63
3.13 3D fits of TCP decomposition rate as a function of temperature and shear stress for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) surfaces. The surfaces were fitted with the nonlinear least squares method. 63

4.1 Side-view snapshots showing the model systems with TCP molecules confined between a) α-Fe(110) and b) α-Fe_2O_3(0001) surfaces. Systems shown after minimisation and equilibration, but before heating, compression, and shear. Black dashed line shows the simulation box. Rendered using OVITO.[306] 70
4.2 Plots of different fitting methods used to determine the rate constant for the DLC-Fe_2O_3 system at 1 m/s, 5 GPa, R^2 for each fit shown in the legend. 74
4.3 Average variation in shear stress over time for different sliding speeds for (a) DLC-Fe_2O_3 and (b) DLC-α-Fe(110). 77
4.4 Comparison of TCP molecule decomposition at 5 GPa at different temperatures for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3 at 10 m/s. 79
4.5 Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3. 79
4.6 Comparison of Arrhenius and Bell model activation energies across the studied sliding speeds for a) α-Fe and b) α-Fe_2O_3 systems. Fit of the logarithm of the activation energy to the sliding speeds used in this study, extrapolated to order of magnitude of sliding speeds used in experiments for c) α-Fe and d) α-Fe_2O_3 systems. 81
4.7 Time evolution of polyphosphate formation at 500 K and 5GPa, for various sliding speeds for DLC-Fe_2O_3 system. 84
4.8 Comparison of bond length distributions as a function of sliding speed between DLC-Fe_2O_3 and DLC-α-Fe(110) systems for selected bond types. Mean length shown by green diamonds. 86

5.1 Chemical structures of the nine hydrocarbons analysed in this study. 92
5.2 Time-dependent viscosity traces for Squalane at 313 K and 373 K with the Einstein and Green-Kubo methods, using the SCIPCFF forcefield. 95
5.3 Parity plot of experimental viscosity vs NIST Webbook values. Inset shows zoom in of plot between 0 and 5 mPa.s. 96

5.4 Different molecular representations used to train ML models in this study, implemented for example molecule Butyrophenone. 98
5.5 Effect of different embedding methods and dimensionality on XGBoost prediction of hydrocarbon thermophysical properties. Best method-dimension combination for each combination shown in legend. Error bars show 1 standard deviation from mean from 5-fold cross validation. 106
5.6 Validation RMSE surface plots for XGBoost across bit size and radius for (a) Viscosity and (b) Thermal Conductivity at 40°C. 108
5.7 Comparison of hashed vs. unhashed fingerprint performance for the best-performing hashed configurations, retrained with unhashed fingerprints at the same radius. The top bar chart displays XGBoost results, and the bottom bar chart shows MLP results. Each bar group indicates the test RMSE for the hashed (left bar) and unhashed (right bar) fingerprints. 109
5.8 Parity plot comparing actual vs. predicted values for viscosity across three models: Transformer, XGBoost with molecular descriptors, and XGBoost with fingerprints. Marker shapes denote temperature (circles for 100°C and squares for 40°C). 112
5.9 Mean absolute percent deviation of simulation predictions from experimental values for density ($\bar{\delta}\rho$), viscosity ($\bar{\delta}\eta$) via both Green–Kubo (GK) and Einstein formalisms, and thermal conductivity ($\bar{\delta}\kappa$) via GK and Einstein. 115
5.10 Effect on training dataset size on prediction of viscosity and density at 40°C and 100°C. Each colored line corresponds to a specific model–representation combination, and "X" markers indicate the training percentage with the best RMSE. 118
5.11 Out-of-distribution and test RMSE for log viscosity as a function of the percentage of retained features for both temperatures. The best-performing models for each RMSE type are highlighted. 121
5.12 Average rank of OOD and test RMSE across different feature retention levels for 40°C and 100°C. Lower rank indicates better performance. The lowest rank shows the optimal feature subset balancing both generalisation and internal accuracy. 122
5.13 Combined parity plot showing predicted vs. actual log viscosity values for OOD molecules across two temperatures (40°C and 100°C). Marker shape distinguishes temperature (squares for 40°C, circles for 100°C), while color indicates test property category. Each model is trained using the optimal stable features from recursive feature elimination (RFE). RMSE and R^2 values reflect prediction performance for each temperature-specific model on its respective OOD subset. 123
5.14 Relationship between Mahalanobis distance and absolute prediction error for the OOD dataset when predicting log viscosity at 40°C and 100°C. Molecules are color-coded by test property and flagged as OOD using a 97.5% Mahalanobis distance threshold. A Pearson correlation coefficient is included in the legend to assess association between Mahalanobis distance and absolute error. 124

6.1 Overview of Genetic Algorithm Process 131

6.2 Overview of Molecule Crossover Process. . . . 136
6.3 KVI vs DVI values for oils in the NIST experimental dataset. . . . 137
6.4 Test RMSE versus percentage of features retained for Thermal Conductivity, Density, and Log Heat Capacity at 40°C and 100°C. Error bars denote one standard deviation. . . . 139
6.5 Balanced Accuracy and Matthews Correlation Coefficient (MCC) on the test set as a function of the percentage of features retained after RFE. . . . 142
6.6 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule biodegradability. . . . 143
6.7 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule DVI. . . . 144
6.8 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 40°C. . . . 146
6.9 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 40°C. . . . 148
6.10 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log heat capacity at 40°C. . . . 149
6.11 Evolution of the average niched score over generations for WSGA configurations across three repeats. The plot highlights the top 3 best-performing hyperparameter combinations with coloured lines and shaded standard deviation regions, with the performance of other hyperparamater combinations shown by grey lines. . . . 152
6.12 Best hypervolume achieved by each hyperparameter set used to run the NSGA-II algorithm, plotted against the generation at which that hypervolume was achieved. The 5 hyperparemeter sets that achieved the best hypervolumes are highlighted. Error bars indicate the standard deviation of hypervolume at that generation. . . . 154
6.13 Grid of top 30 molecules from WSGA runs, showing the target property from which that molecules was taken, as well as its biodegradability prediction. . . . 156

A.1 Change in shear stress with normal stress for 100% TCP and 50% TCP, 50% DIDA at 600K and 1-5 GPa. Friction coefficient, $\mu = 0.18$ (100% TCP), $\mu = 0.17$ (50% TCP, 50% DIDA) and Derjaguin offset, $\tau_0 = 0.79$ GPa (100% TCP), $\tau_0 = 0.20$ GPa (50% TCP, 50% DIDA). . . . 206
A.2 Number density profiles for the Fe, C, O, and P atoms in the a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) systems at 500 K and 3 GPa . . . 206
A.3 Variation in TCP dissociation rates for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) under various conditions. . . . 208
A.4 Change in the number of P-O and C-O bonds compared to TCP, for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) under various conditions. Solid lines represent total change, dashed lines show bond formation and dotted lines show bond cleavage. . . . 209
A.5 Comparison of P-O_{surf} bonds formed in the Fe_3O_4(100) system at a) 400 K, b) 600 K, and c) 800 K . . . 210

A.6 Change in the number of polyphosphate chains with sliding time for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) systems. . . . 210
A.7 Comparison of the logarithm of the reaction rates calculated from the NEMD simulations and those predicted using Equation 1. R^2 = 0.91 for α-Fe(110), R^2 = 0.88 for Fe_3C(010), and R^2 = 0.87 for Fe_3O_4(001). . . . 210
A.8 Arrhenius plots of ln(k) versus 1000/T for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) obtained at different pressures. From Equation 1, the intercept of the lines gives A and gradient gives E_a. . . . 211
A.9 Temperature dependence of ΔV^* for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals. . . . 211
A.10 Pressure dependence of E_a for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals. . . . 211
A.11 Pressure dependence of ln(A) for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals. . . . 212
A.12 3D fits of TCP decomposition ln(k) as a function of temperature and shear stress for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) surfaces. . . . 213

B.1 Plots of different fitting methods used to determine rate constant for the DLC-α-Fe(110) systems at 1 m s^{-1}, R^2 for each fit shown in the legend. . . . 214
B.2 Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds and temperatures for the DLC-α-Fe(110) model. . . . 215
B.3 Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds and temperatures for the DLC-Fe_2O_3 model. . . . 216
B.4 ime evolution of polyphosphate formation at 500 K and 5GPa, for various sliding speeds for DLC-α-Fe(110) system. . . . 217
B.5 Comparison of bond length distributions as a function of normal pressure between DLC-Fe_2O_3 and DLC-α-Fe(110) systems for selected bond types. . . . 218
B.6 Comparison of TCP molecule decomposition at 5 GPa at 0 m/s for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3. . . . 219

C.1 Effect of different embedding methods and dimensionality on MLP prediction of hydrocarbon thermophysical properties. Best method-dimension combination for each combination shown in legend. Error bars show 1 standard deviation from mean from 5-fold cross validation. . . . 220
C.2 Parity plots comparing actual vs. predicted values for density, heat capacity and thermal conductivity across three models: Transformer, XGBoost with molecular descriptors, and XGBoost with fingerprints. . . . 222

C.3 Comparison of MD and ML predictions against experiments. (a) Density predictions and (b) Thermal-conductivity predictions (Green–Kubo formalism). Circle markers: 313 K; square markers: 373 K; error bars show one-sigma uncertainties. 228
C.4 Predicted vs. experimental viscosities from molecular-dynamics (Green–Kubo) and machine-learning models at 313 K (circle markers) and 373 K (square markers). Left: full range; right: zoom into 0–10 mPa.s. Error bars show one-sigma uncertainties. 229

D.1 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 100°C. 230
D.2 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 100°C. 231
D.3 Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log heat capacity at 100°C. 231
D.4 Grid of top 20 molecules by crowding distance from best hyperparemeter combination using NSGA-II optimisation. 232

List of Tables

3.1 Calculated values of the activation energy, E_a, natural logarithm of the pre-exponential factor, ln(A), and activation volume, ΔV^*, for TCP between α-Fe(110), Fe_3C(010) and Fe_3O_4(100) surfaces from the 3D fits to Equation 1. The parameter ranges represent the 95 % confidence intervals from the 3D fits. 64

4.1 R^2 values for the different fitting methods at 500 K, 5 GPa for DLC-Fe_2O_3 across the investigated sliding speeds. 73

4.2 Activation Volumes for α-Fe and α-Fe_2O_3 systems at different temperatures at 10 m/s, calculated from simulations performed at temperatures ranging from 300 K to 600 K in 100 K intervals, under varying normal pressures (2-10 GPa with 2 GPa intervals). 80

5.1 Comparison of molecular family counts between the collated dataset of experimental values and the Kajita dataset. 97

5.2 Features included in 3D Molecular Graphs 100

5.3 Molecular representation - ML model combinations investigated in this study. 102

5.4 Best performing validation RMSE and corresponding embedding dimension for each property using PCA and UMAP embeddings with XGBoost and MLP models. Values are shown as *RMSE (dimension)*, and the lowest RMSE per property is bolded. 107

5.5 Best performing ECFP configurations by model and property. 108

5.6 RMSE and R^2 (in italics) of multitask transformer model for each property and temperature. 114

5.7 Best performing RMSE for each property and temperature. Non-bracketed numbers are best RMSE values; square brackets denote R^2; parentheses denote percentage improvement from 1% use of the training set if any; angle brackets denote the training split at which this performance was achieved. 117

6.1 Mutation types used in the GA with associated conditions and probabilities. 135

A.1 Elastic constant (C_{33}) values for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) slabs with ReaxFF. 207

A.2 Friction Coefficient, μ, for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces at different temperatures. 207

A.3 Derjaguin offset for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces at different temperatures. 207
A.4 ΔV^* for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces. 212
A.5 E_a for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces. 212
A.6 $\ln(A)$ for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces. 212

B.1 R^2 values for the different fitting methods at 500 K for DLC-α-Fe(110). 215
B.2 Arrhenius activation energies and pre-exponential factors for DLC-α-Fe(110) and DLC-Fe_2O_3. 216
B.3 Average Bell activation energies for DLC-α-Fe(110) and DLC-Fe_2O_3. 216

C.1 Test RMSE values for different ML models across molecular properties. Top 3 performers per property are highlighted. Overall rank is shown in the final column. 221
C.2 OOD test data with different atoms present in structures compared to training data. . . . 223
C.3 OOD test data with higher viscosities compared to training data. 224
C.4 MD/ML prediction on test data for thermal conductivity values. 225
C.5 MD/ML prediction on test data for viscosity values. 226
C.6 MD/ML prediction on test data for density values. 227

D.1 Molecular properties of top 20 molecules by crowding distance from best hyperparameter combination using NSGA-II optimisation. 233
D.2 Ranked properties of top 30 WSGA molecules with gold, silver, and bronze highlighting best performing molecules for that property. 234

List of Abbreviations, Acronyms and Symbols

Abbreviation/Symbol	Definition
ΔV^*	Activation Volume
η	Viscosity
κ	Thermal Conductivity
μ	Friction Coefficient
ρ	Density
σ	Normal Stress
τ	Shear Stress
A	Pre-exponential Factor
E_a	Activation Energy
k_B	Boltzmann Constant
t_{cut}	Bootstrapping Cutoff Time for Time Decomposition Method
ACF	Autocorrelation Function
AFM	Atomic Force Microscopy
AMBER	Assisted Model Building with Energy Refinement Force Field
AES	Auger Electron Spectroscopy
ATS	Autocorrelation of Topological Structure
CHARMM	Chemistry at Harvard Macromolecular Mechanics Force Field
CNN	Convolutional Neural Network
DFT	Density Functional Theory
DIDA	Diisodecyl Adipate
DLC	Diamond-Like-Carbon
DVI	Dynamic Viscosity Index
ECFP	Extended-Connectivity Fingerprint
EHL	Elastohydrodynamic Lubrication
GA	Genetic Algorithm
GAN	Generative Adversarial Network
GAT	Graph Attention Network
GK	Green-Kubo
GTL	Gas-to-Liquid Oil
kNN	k-nearest Neighbours
KVI	Kinematic Viscosity Index
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
MAPS	Materials and Process Simulation Softwate (SARL Scienomics)
MCC	Matthews Correlation Coefficient
MCTS	Monte Carlo Tree Search
MD	Molecular Dynamics
ML	Machine Learning
MLP	Multi-Layer Perceptron

Abbreviation/Symbol	Definition
MMFF94	Merck Molecular Force Field 94
MOOP	Multi-Objective Optimisation Problem
MPNN	Message Passing Neural Network
MSD	Mean Squared Displacement
MSE	Mean Squared Error
NEB	Nudged Elastic Band
NEMD	Nonequilibrium Molecular Dynamics
NIST	National Institute of Standards and Technology
NN	Neural Network
NPT	Isothermal-isobaric ensemble
NSGA-II	Non-dominated Sorting Genetic Algorithm II
NVE	Microcanonical Ensemble
NVT	Canonical Ensemble
OOD	Out-of-Distribution
OPLS	Optimized Potentials for Liquid Simulations
OVITO	Open Visualization Tool
PAO	Poly-α-olefin
PBC	Periodic Boundary Condition
PCA	Principal Component Analysis
PES	Potential Energy Eurface
PFPE	Perfluoropolyether
PPPM	Particle-Particle, Particle-Mesh
ReaxFF	Reactive Force Field
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
SAScore	Synthetic Accessibility Score
SATA	Stress-Augmented Thermally Activated
SCScore	Syntheic Complexity Score
SELFIES	Self-Referencing Embedded Strings
SHAP	Shapley Additive Explanations
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machines
TCP	Tricresyl Phosphate
TEM	Transmission Electron Microscopy
UMAP	Uniform Manifold Approximation and Projection
VAE	Variational Autoencoder
WSGA	Weighted Sum Genetic Algorithm
XGBoost	Extreme Gradient Boosting
XPS	X-ray Photoelectron Spectroscopy
ZDDP	Zinc Dialkyl Dithiophosphate

Chapter 1

Introduction

This chapter provides an introduction to the research project undertaken. The need for new methods of designing and characterising both known and novel lubricant base oils, and the need for gaining a deeper understanding of lubricant additive behaviour, is discussed. Research goals are set forth which describe the overarching aim of the research project. Finally, a breakdown of the thesis structure is provided to aid the reader.

1.1 Motivation

The development of high-performance lubricant formulations is essential to improve energy efficiency, particularly in the transportation sector. As internal combustion engine technologies evolve and electric vehicles become increasingly prevalent, there is a pressing need for next-generation lubricants tailored to the unique operational requirements of modern powertrains. For electric vehicles, this includes demands for reduced wear, improved thermal stability, stability in environments containing electric fields, and compatibility with new materials and components.[1] In parallel, growing awareness of the environmental consequences associated with conventional lubricants has driven a shift toward more sustainable alternatives to traditional lubricant formulations. There is rising interest in formulations derived from renewable feedstocks, as well as lubricants that are biodegradable and exhibit low toxicity, all while maintaining the stringent performance standards expected in industrial and automotive applications.[2] Despite this urgent need and industrial relevance, the development of novel lubricants remains a largely empirical and/or iterative process. Traditional formulation approaches rely heavily on trial-and-error experimentation, which is both resource-intensive and time-consuming. This presents a major bottleneck in the discovery and optimisation of improved lubricant formulations.[3]

Modern lubricants are complex formulations consisting of a blend of at least one base oil, in which ten or more different additives are typically dissolved.[4] The primary role of a lubricant is to maintain adequate viscosity across a wide range of operating temperatures, ensuring consistent and effective lubrication under varying system conditions. In addition to viscosity control, additives are included to perform a variety of critical functions, such as preventing corrosion and enhancing thermal and ox-

idative stability.[5,6] However, both base oils and particularly additives are surface active and chemically reactive.[7] The interactions between each other[8] and the confining surfaces[9] can lead to unpredictable behaviours that often negatively impact lubricant performance. Furthermore, there is a lack of consensus regarding which atomic-scale mechanisms dictate the macroscale performance of lubricant additives, especially in the case of anti-wear additives such as zinc dialkyldithiophosphate (ZDDP).[10] Hence, there also a need to build an understanding of the mechanisms that govern the behaviours of these base oils and additives. Given the complexities of tribological systems and the difficulty in obtaining in-situ atomic-scale information from inaccessible contacts under extreme temperature, pressure, and shear conditions, experimental efforts attempting to characterise these mechanisms are limited.[11,12] Hence, computational modelling across scales is playing an increasingly important role in tribology research.[13,14] For nanoscale processes, molecular simulations have uncovered several fundamental processes central to the properties and performance of lubricants.[15,16,17] This PhD project seeks to develop computational molecular design and characterisation techniques to accelerate the development of novel lubricant base oils and additives.

1.2 Research Goals

The primary goal of this project is to develop computational molecular design and characterisation techniques that lay a foundation for the accelerated development of novel lubricant base oils and additives. As mentioned above, lubricant formulations are normally composed of a base oil and a plurality of additives, with the base oil and additive each presenting different challenges regarding the development of novel design and characterisation techniques. For example, there are very few publicly available databases that provide information on important lubricant base oil properties, and even fewer for lubricant additives. With the above in mind, the following specific objectives for this study were proposed in order to address the overarching goal of the thesis with the constraints of available data:

- Investigate the mechanochemical decomposition of a prevalent lubricant additive across different surface chemistries, solvent environments, temperatures, and shear conditions to understand how nanoscale mechanisms influence macroscale tribological behaviour of lubricant additives.
- Evaluate the effectiveness of various ML and MD techniques in predicting key transport properties of known and novel hydrocarbons suitable for use as lubricant base oils.
- Use interpretable ML approaches to uncover molecular features that influence lubricant behaviour and support the rational design of next-generation lubricant formulations.
- Develop and apply multi-objective optimisation algorithms to identify base oil molecules with a desirable balance of performance characteristics.

Understanding how additives decompose under varying stress and temperature environments, and how they interact with different surfaces, helps identify the conditions under which base oils should enhance or suppress these reactions, enabling the design of base oils that support or stabilise additive behaviour. Computational techniques such as MD or ML offer a scalable route to propose and screen

large numbers of candidate molecules, while interpretable ML reveals the molecular features that influence performance, helping to close the gap between predictive accuracy and chemical insight. Finally, multi-objective optimisation algorithms integrate these models and insights to guide the discovery of base oils with desirable combinations of properties. Together, these elements form a coherent workflow: mechanistic understanding grounds predictive modelling, modelling enables large-scale exploration, and optimisation transforms insight into actionable molecular design. This enables a more rational and efficient path to next-generation lubricant formulations.

1.3 Layout of Thesis

This thesis is organised into an initial chapter, where the relevant literature is reviewed, subsequent results chapters, where the modelling techniques are applied and the results analysed, and a closing chapter, where the main findings of the project are summarised and conclusions are drawn.

In Chapter 2, the computational modelling techniques and relevant theory used in this study are presented. This includes an overview of lubricants and the properties critical to their performance, as well as a discussion of evolutionary algorithms, ML, and MD, with a focus on their application to the design and characterisation of base oils and additives.

In Chapter 3, the effects of surface chemistry on the mechanochemical decomposition of TCP is explored using MD simulations. The influence of different ferrous surface chemistries on the decomposition mechanisms and tribofilm formation is analysed in detail.

In Chapter 4, the mechanochemical decomposition of TCP under varying shear stresses and temperatures is investigated, where the relationship between sliding speed, bond elongation, and TCP decomposition is examined

In Chapter 5, a range of ML methods and molecular representation combinations are evaluated to determine the most performant combination for predicting hydrocarbon transport properties. The performance of the best model/representation pair is subsequently compared to MD simulations using different force fields. The feasibility of using datasets generated from MD to train ML regression models is also assessed. Finally, the limitations of these ML models are explored by evaluating their performance on molecules that lie outside the distribution of the training data.

In Chapter 6, multi-objective optimisation techniques are employed to identify base oil molecules with desirable properties. ML and evolutionary algorithms are used to predict and interpret the influence of molecular structure on macroscopic lubricant base oil properties.

Concluding remarks and suggestions for future work based on the work of the previous chapters are presented in Chapter 7.

Chapter 2

Literature Review

This chapter provides a brief overview of some of the fundamental tribology principles relevant to this study, and introduces the computational techniques employed in this thesis. Initially, this chapter introduces the importance of computational molecular modelling in lubricant design and tribology. Next, a discussion of the formulation of modern lubricants, including the roles of base oils and additives in complex tribological systems is given. Following this, an introduction to MD techniques commonly used to simulate tribological systems is provided, including both reactive and non-reactive force-fields. This chapter then presents an overview of evolutionary algorithms and ML approaches for molecular candidate generation and property prediction. Finally, different molecular representations and data preprocessing techniques used in ML workflows are reviewed.

2.1 The Importance of Molecular Modelling and ML in Lubricant Design and Tribology

Improving the performance and sustainability of lubricated systems is critical for reducing global energy consumption and greenhouse gas emissions.[18] Tribological contacts are estimated to account for 23% of global energy consumption, with 20% lost to friction and a further 3% due to wear-related component replacement.[19] Emissions from the transport sector alone accounted for 27% of the total GHG emissions in the UK in 2019, representing the largest sector contribution, where passenger car and taxi emissions accounted for 55% of the UK's transport sector emissions.[20] In the transportation sector worldwide, internal combustion engine (ICE) vehicles consume approximately 200 billion litres of fuel annually to overcome friction.[18] Holmberg et al. estimated that by employing novel tribological practices, friction losses in passenger cars could be reduced by 18% in the next ten years, with reductions of up to 61% achievable in the next 15 to 25 years, which in turn could yield worldwide savings of 174 billion euros, 117 billion litres of fuel, and 290 million tonnes of CO_2 emissions.[21]

There is broad consensus within the automotive industry that electric vehicles (EVs) will dominate new passenger car sales within the next two decades.[22] This transition is being actively supported by govern-

ment legislation, such as the UK's commitment to end the sale of new petrol and diesel vehicles by 2030, with a subsequent ban on hybrid vehicles that include internal combustion engines by 2035.[23] However, the shift to electric vehicles introduces new challenges in lubricant formulation. Lubricants for EV drive units must perform under a wide range of operating conditions. They must protect gears and bearings under high-torque, low-speed scenarios while also minimising churning losses during high-speed, low-torque operation, requiring fluids that balance both high and low viscosity properties.[1,24] Additionally, in many modern EVs, a single fluid is used for both gearbox lubrication and e-motor cooling, imposing further requirements, such as high specific heat capacity and compatibility with copper components.[25] While conventional hydrocarbon-based lubricants may still offer strong surface protection, they often fall short of meeting the full spectrum of performance needs in electric vehicle applications.

Given the complexity of tribological systems, where chemically reactive surfaces, formulated lubricants, and operating conditions interact, in situ experimental investigations represent a considerable challenge.[11,12] Molecular modelling offers a powerful means of investigating tribochemical mechanisms at the atomic scale.[15,16,17] Tools such as MD simulations have already been employed to investigate the behaviour of electric fields, additive chemistry and base oil chemistry at this scale.[26,27,28] Furthermore, ML has become a staple for both the property prediction and molecular generation of known and novel tribofluids.[29,30,31,32] As the need for energy-efficient and sustainable tribofluids continues to grow, molecular modelling will continue to provide a critical foundation for the rational development of next-generation lubricants.

2.2 Background on Lubrication/Tribology

2.2.1 Lubrication Regimes and Important Lubricant Properties

Lubrication Regimes

Lubricants are used to reduce friction and wear between surfaces in relative motion by creating a protective film that separates them.[34] This film can reduce or prevent asperity contact between the sliding surfaces resulting in lowered resistance to motion, reducing heat generation and surface damage. The reduction in friction caused by effective lubrication enhances the reliability and efficiency of mechanical systems[35] and reduces the risk of component failure.

A widely used model to describe friction in lubricated contacts is the Stribeck curve, which relates the coefficient of friction to the Hersey number, defined as $H = \frac{\eta v}{P}$, where η is the dynamic viscosity, v is the sliding velocity, and P is the applied load per unit length.[36] The Stribeck curve (illustrated by Fig. 2.1) describes the relationship between the coefficient of friction and the Hersey number, identifying four primary lubrication regimes: boundary, mixed, elastohydrodynamic (EHL), and hydrodynamic lubrication. In the boundary regime, the lubricant film is too thin to completely separate the surfaces, so friction is governed primarily by surface interactions (e.g., asperity contact).[37] The mixed regime represents a transition phase where partial surface contact occurs alongside some fluid film separation.[38] In the EHL regime, the pressure is high enough to elastically deform the contacting surfaces and generate a thin but continuous fluid film, significantly reducing friction.[39] Finally, in the hydrodynamic regime, a full fluid

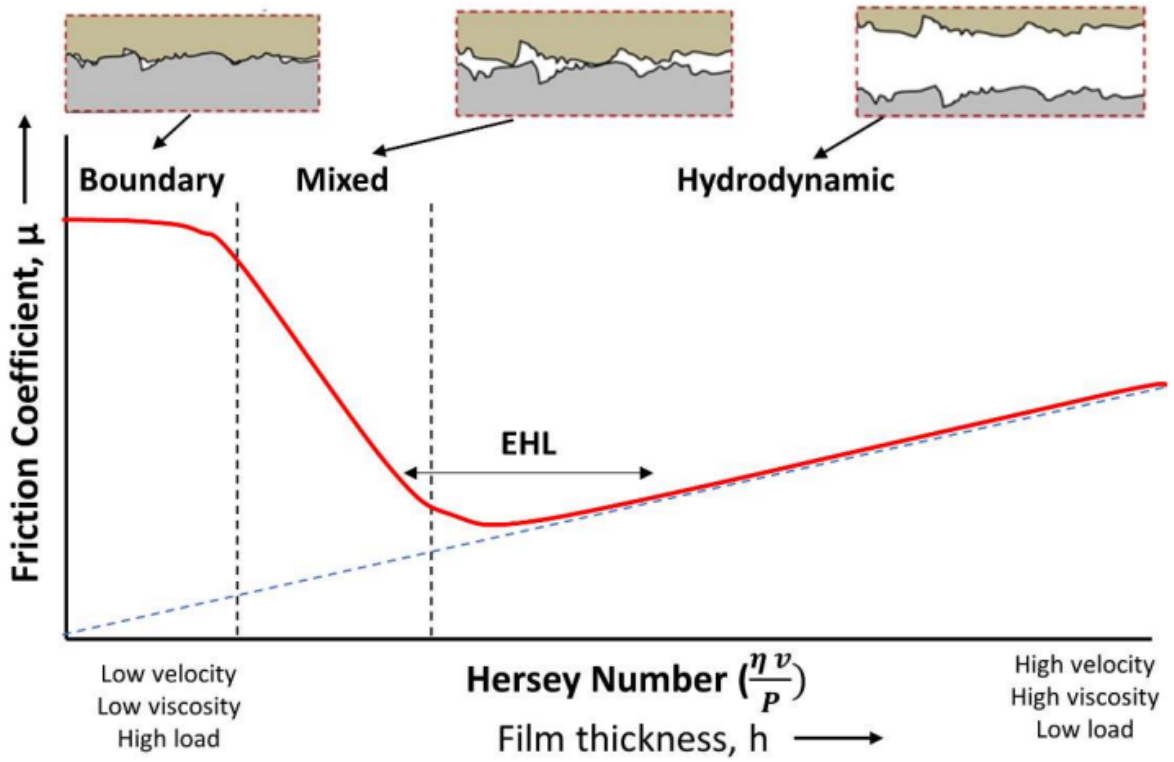


Figure 2.1: Stribeck Curve showing different lubrication regimes. Reproduced from *'Tribology: friction and wear of engineering materials'*.[33]

film separates the surfaces entirely, and friction arises from the internal shearing of the lubricant.[40]

Components with tribological contacts found in automotive vehicles such as gears and valve trains operate under mixed and boundary lubrication due to fluctuating loads and speeds.[33] Meanwhile, aviation systems featuring tribological contacts such as turbines and gearboxes are designed to maintain stable EHL or hydrodynamic lubrication to ensure reliable operation under extreme temperature and pressure conditions.[41] Electric vehicles (EVs) present entirely novel lubrication challenges due to the range of speeds that e-motors may operate at, as well as their frequent start-stop operation, meaning EV drive units often operate across boundary, mixed, and EHL lubrication regimes. Additionally, the absence of combustion reduces oil temperatures, affecting viscosity and the ability to maintain hydrodynamic films.[1,25]

Important Lubricant Properties

As discussed above, a well-formulated lubricant must optimise a combination of physical and chemical properties to ensure reliable performance across a range of operating conditions. Viscosity, the most fundamental property, quantifies a fluid's resistance to flow and is typically measured in centistokes (cSt) using capillary or rotational viscometers.[42] For automotive applications, typical kinematic viscosity values at 100°C range between 5 and 15 cSt,[43,44] while in aerospace applications, lower viscosities (e.g., 3–8 cSt) may be preferred for low-temperature flow and fuel economy.[45] A higher VI indicates that the

oil experiences less viscosity change over a given temperature range, which is crucial for maintaining effective lubrication during both cold starts and high-temperature operation. Synthetic oils typically exhibit VI values above 140, with high-performance oils often exceeding 160.[46,47]

Density affects energy losses through churning and impacts fuel economy with automotive lubricants typically falling within a density range of 840–1000 kg/m^3.[48,49,50] Thermal conductivity governs the ability of the lubricant to transfer heat away from contact zones. Higher values are beneficial for thermal regulation in high-speed gear systems and e-motor cooling, with typical values for automotive oils falling within the range of 0.13–0.16 W/m·K.[51] Heat capacity reflects the ability of the lubricant to absorb thermal energy without a large temperature rise where values between 1.8–2.3 kJ/kg·K are desirable to support effective cooling for automotive applications.[52]

Pressure stability, or high-pressure viscosity behaviour, ensures that the lubricant maintains a sufficient film thickness under extreme load. Seal compatibility characterises the chemical interaction between the lubricant and sealing materials ensuring that seals do not degrade, swell, or shrink. This is evaluated through standardised soak and tensile tests such as ASTM D471.[53] Lubricants must also balance oxidation stability, volatility, and additive solubility, especially in high-temperature/pressure applications seen in aviation, where long service life and resistance to breakdown are critical.[47] Optimising all of these properties is essential for ensuring performance, durability, and energy efficiency across all lubrication regimes. To this end, automotive and aerospace lubricants are complex chemical formulations primarily composed of a base oil, typically making up 75–95% of the total volume, in which a suite of performance-enhancing additives is dissolved, where the remaining 5–25% consists of a carefully balanced additive package to improve property deficiencies in the base oils.[54,55]

2.2.2 Base Oils

Base oils serve as the foundation of lubricant formulations and play a critical role in determining lubricant performance. Their primary function is to provide the correct viscosity for film formation while offering stability under mechanical, thermal, and oxidative stress. Base oils are broadly classified into mineral and synthetic groups based on their method of production.[56] Mineral base oils, derived from crude oil through refining processes such as solvent extraction, hydrocracking, and dewaxing, contain a mixture of paraffins (alkanes), naphthenes (cycloalkanes), and aromatics. While affordable and widely available, mineral oils have limited oxidative stability and volatility control compared to synthetics, and their performance can vary depending on the refining process and crude source.[47]

Synthetic base oils, produced via chemical synthesis or advanced processing of natural gas or crude oil fractions, can be tailored for enhanced performance characteristics. Global synthetic lubricant production is 3.5 million tons annually, representing around 11% of the total market, while semi-synthetic lubricants account for a further 2.1 million tons.[57] Among the most common synthetic base oils are alkanes, such as polyalphaolefins (PAOs) and gas-to-liquid (GTL) fluids. PAOs are made by oligomerising alpha-olefins, resulting in branched, saturated hydrocarbons with excellent low-temperature fluidity, high VI, and oxidation resistance. However, their low polarity can hinder additive solubility, meaining they often require blending with more polar fluids.[47] GTL base oils, derived from natural gas using the

Fischer–Tropsch process, offer similarly low volatility and thermal stability but share PAO's solvency limitations due to their highly linear, non-polar molecular structure.[47]

Alkylated aromatics, alkylated naphthalenes, alkylated cycloalkanes, and alkylated cycloalkenes are all synthetic hydrocarbons tailored to improve solvency, polarity, and oxidative stability. Alkylated aromatics, often used in compressor and refrigeration oils, exhibit good additive compatibility and thermal resistance.[47] Alkylated naphthalenes offer excellent hydrolytic and oxidative stability, making them suitable for high-temperature gear and turbine oils, although they tend to be more expensive and less biodegradable than PAOs.[47] Alkylated cycloalkanes and cycloalkenes provide improved lubricity and low-temperature performance but may exhibit moderate volatility depending on molecular weight and structure.[58] Alkenes, such as terminal olefins, are generally not used directly as base oils due to their chemical reactivity and poor stability, so they are mostly used as intermediates in synthetic base stock production (e.g., 1-decene trimer for PAO production).[59]

Synthetic oxygenates, including esters, aromatic esters, polyglycols, and ethers, introduce polar functional groups that significantly enhance lubricant properties. Esters, formed by reacting acids and alcohols, offer excellent lubricity, biodegradability, and thermal stability, and are widely used in aviation, biodegradable lubricants, and racing oils.[47] However, they are more prone to hydrolysis in the presence of moisture.[60] Aromatic esters provide even greater thermal and oxidative stability, suitable for turbine oils and jet engine lubricants, but can be more expensive. Polyglycols, which can be water-soluble or oil-soluble depending on their structure, provide high VI, low pour points, and fire resistance, making them ideal for hydraulic fluids and compressors. However, their poor compatibility with some seal materials and additives prevent their use in certain applications. Ethers offer outstanding low-temperature performance and oxidative stability, but are typically limited to niche applications due to their cost and synthetic complexity.[47]

Organohalogens such as perfluoropolyethers (PFPEs) represent a class of synthetic base oils that have garnered attention in the automotive sector for their exceptional chemical inertness, thermal stability, and resistance to degradation under extreme operating conditions. They are also non-flammable and exhibit strong oil- and water-repellent characteristics. For automotive lubrication purposes, PFPEs have been explored for use in engine oils, hydraulic fluids, and specialised greases where conventional hydrocarbons may degrade or fail, such as in high-temperature zones, electrified powertrains, and sealed bearing systems.[61] However, despite their functional advantages, PFPEs remain limited to niche automotive uses due to their high cost and incompatibility with certain elastomers and additive chemistries.[62]

Each of these base oil classes has distinct structural features that influence their viscosity behaviour, thermal stability, oxidation resistance, additive compatibility, and environmental impact. Their selection must be carefully matched to the operating conditions and performance requirements of the lubricant application, and as will be discussed later in this thesis, their molecular structure plays a central role in friction and wear behaviour at the nanoscale.

2.2.3 Additives

Additives can constitute up to 25% of a lubricant's volume and are included to enhance performance and meet specific regulatory or OEM requirements. Anti-foam additives reduce bubble formation and foam stability caused by surfactant-based additives, with high molecular weight silicone polymers commonly used for this role.[63] Antioxidants delay oxidative degradation of the base oil by neutralising reactive oxygen species, where common examples include aromatic amines, phenols, sulphur compounds, and organo-molybdenum compounds.[64] Corrosion inhibitors protect metal surfaces from oxidative and corrosive attack by adsorbing onto them, with typical examples including amine succinates, alkaline metal sulfonates, and benzotriazole.[65] Detergents prevent the formation of carbon-based deposits on engine surfaces at high temperatures. They are typically amphiphilic surfactants that form reverse micelles to neutralise acidic by-products.[66] Dispersants maintain the suspension of solid impurities within the lubricant, preventing sludge formation. They are also surfactant-like molecules, with succinimides being the most widely used.[66] Friction modifiers reduce friction under boundary lubrication by forming protective surface layers. Inorganic examples like molybdenum dithiocarbamate form MoS_2 films, while organic friction modifiers such as fatty acids, esters, and amines adsorb to surfaces to form monolayers.[7] Viscosity index improvers (VII) are high molecular weight polymers, such as polyalkyl-methacrylate (PMA) and olefin copolymers, that increase oil viscosity at high temperatures to ensure stable lubrication.[32] Pour point depressants lower the minimum temperature at which a base oil can flow by disrupting wax crystal formation, with low molecular weight PMA often used for this purpose.[66]

Anti-Wear Additives

Anti-wear additives, the primary additive type of interest to this thesis, protect mechanical components operating under high pressure and high-sliding speed conditions typical of internal combustion engines, electric drive units, and aerospace turbines.[67] The most widely used anti-wear additive in automotive lubricants is ZDDP, which has been in use since the 1940s due to its proven ability to form protective phosphate-based tribofilms on metal surfaces.[10] These films, typically 50–200 nm thick, are formed rapidly under rubbing conditions and consist largely of zinc and iron polyphosphates, which act as a barrier to wear while avoiding corrosive iron oxide formation.[10] ZDDP also offers antioxidant and corrosion-inhibiting functions, though its phosphorus and sulphur content contributes to deactivation and fouling of after-treatment systems such as diesel particulate filters.[68] To meet modern emissions regulations, low-SAPS (sulphated ash, phosphorus, sulphur) formulations are increasingly used, necessitating the development of alternative anti-wear chemistries. [67]

Ashless organophosphates, such as TCP, are employed in aerospace turbine oils due to their excellent thermal stability, and ability to form iron phosphate tribofilms.[68] Such phosphates can form films under sliding contacts that provide effective wear protection at high temperatures.[69] Organoboron compounds, metal-free thiophosphates, and functionalised phosphites, have all been investigated as potential replacements that offer comparable anti-wear performance while being more environmentally friendly.[70] In recent years, ionic liquids have also gained considerable attention as potential multifunctional lubricant additives.[71]

2.2.4 Surfaces

The nature of the surfaces in tribological contacts plays a critical role in determining the friction, wear, and lubrication behaviour within the system as surface chemistry, topography, and wettability all affect the way two surfaces interact under load and relative motion. These interfacial properties govern the formation and stability of boundary films, the degree of adhesion between sliding surfaces, and any chemical interactions that may occur. Understanding and controlling surface characteristics is therefore fundamental to optimising tribological performance in challenging environments.

Ferrous Surfaces

Iron and iron oxides are of significant interest in tribological research due to their prevalence in engineering systems, such as in steel components that operate within oxidative or corrosive environments. The Fe(110) surface, one of the most stable crystallographic planes of body-centred cubic (bcc) iron, is frequently used as a model substrate in computational studies.[72,73] Its atomic arrangement and surface reactivity make it a relevant system for investigating adhesion, chemical reactivity, and tribofilm formation. Under operating conditions involving exposure to air, water vapour, or lubricants, such iron surfaces readily oxidise, forming oxide layers that subsequently alter surface properties. These oxide films can either protect the underlying metal or contribute to wear, depending on their composition, structure, and interaction with the environment.[74]

Hematite (α-Fe_2O_3) and magnetite (Fe_3O_4) are among the most commonly encountered oxides that form on iron and steel. They have both been identified in the wear track of tribology experiments.[75,76] These surfaces exhibit distinct structural and electronic characteristics[77] that may influence their mechanical and chemical performance during sliding. Hematite, being harder, often contributes to increased friction,[78] whereas magnetite, with its mixed-valence structure and higher conductivity,[77] may facilitate charge transfer and accelerate tribochemical reactions. Both oxides exhibit anisotropic surface energies and terminations,[77] making their behaviour highly dependent on crystallographic orientation and environmental factors such as humidity, pH, and the presence of other chemical species in their respective environments like lubricants or wear debris.

Iron carbide (Fe_3C, cementite) is an intermetallic compound that plays a critical role in determining the mechanical and tribological properties of carbon steels and cast irons. Cementite is chemically stable and exhibits high hardness, making it an important part of wear-resistant materials. Its crystal structure is orthorhombic, with iron atoms occupying interstitial positions around carbon, resulting in a relatively brittle material that resists plastic deformation but may be prone to cracking under high contact stress.[79] Fe_3C contributes to abrasive wear resistance and can affect surface friction depending on its distribution and orientation. [80] It is less reactive than iron oxides in ambient environments, but under extreme conditions, such as high temperature or sliding-induced flash heating, it can undergo chemical interactions that influence tribofilm formation. The high mechanical strength and low reactivity of cementite has led to its use in cutting tools, bearings, and high-strength steels where both load-bearing capacity and wear resistance are required.[79]

DLC

One effective strategy for tailoring surface behaviour is the application of engineered coatings. Among these, diamond-like carbon (DLC) coatings have emerged as a widely adopted solution for reducing friction and improving wear resistance in demanding applications.[81,82] DLC is an amorphous form of carbon that combines some of the desirable properties of both diamond and graphite, including high hardness, chemical inertness, and low friction. DLC is used as a coating for various automotive components such as piston rings, fuel injectors and valve lifters.[83] Its tribological properties are closely linked to factors such as the ratio of sp^3 to sp^2 hybridised carbon atoms, the degree of surface hydrogenation, the presence of dopants, and chemical interactions with other species in its operating environment. These atomic-scale features influence surface energy, polarity, and reactivity, which in turn affect how DLC interacts with its environment during sliding.[35,84] In humid or aqueous conditions, for instance, the adsorption of water molecules can significantly modify the tribological response. Hydrophobic DLC films often exhibit lower adhesion and friction at low relative humidity,[85] whereas hydrophilic DLC surfaces display lower adhesion and friction under high humidity due to enhanced boundary film formation.[86] The wettability of DLC surfaces is therefore a critical parameter, and it can be tuned by altering the surface chemistry, including through doping with elements that increase surface polarity or energy.[87]

2.2.5 Mechanochemistry Theory

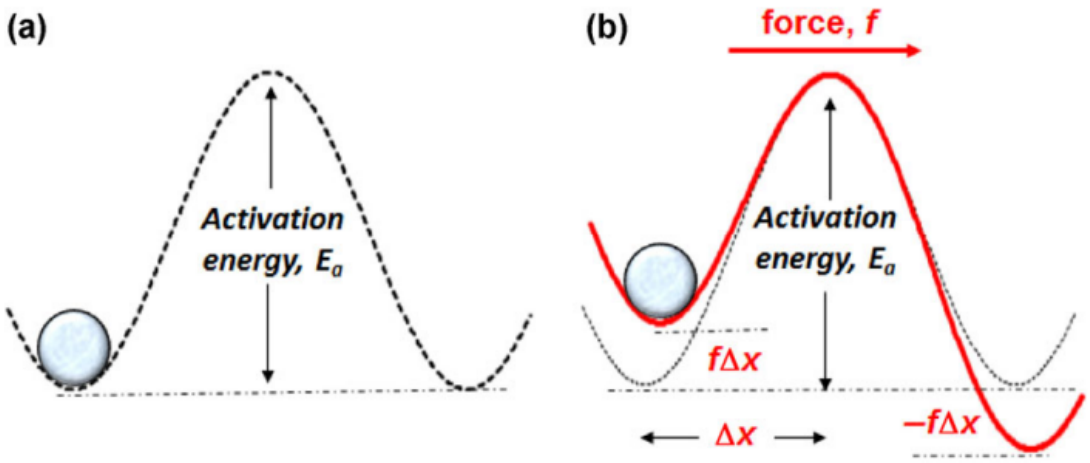


Figure 2.2: The activation barrier for a reaction a) in the absence of an external applied force and b) under the influence of an applied external force. Reproduced from *'Stress-augmented thermal activation: Tribology feels the force'*.[88]

Mechanochemistry refers to chemical reactions that are accelerated by mechanical forces such as stress, strain, or shear.[88,89] At tribological interfaces, such forces can significantly alter reaction pathways by lowering activation energy barriers (as illustrated by Fig. 2.2). Such reactions are known as stress-augmented thermal activated (SATA) processes.[88] This behaviour is described by the so-called Bell model,[90] which modifies the classical Arrhenius rate equation to account for mechanical stress.[88] When

a force f is applied along a reaction coordinate, the reaction rate becomes:

$$k = A\exp\left(-\frac{E_a - f\Delta x}{k_B T}\right) \tag{2.1}$$

where A is the pre-exponential factor, E_a is the activation energy, Δx is the displacement along the reaction coordinate, k_B is the Boltzmann constant, and T is the temperature.

In the case of applied shear stress τ, the equation takes the form:

$$k = A\exp\left(-\frac{E_a - \tau\Delta v}{k_B T}\right) \tag{2.2}$$

where Δv is the activation volume, quantifying how shear reduces the energy barrier.[88] Similar equations have been developed by Evans and Polanyi,[91] Eyring,[92] and others[93] to include the effects of applied stress on thermally activated processes. This theoretical framework is very useful for understanding mechanochemically-driven reactions, such as the formation of protective tribofilms in lubricants.[88] The same SATA framework is also useful for studying many other tribological processes, including lubricant rheology, dry friction, boundary lubrication, and nanoscale wear.[94] The SATA approach has already has already been employed in combination with both experimental and computational studies to provide insights into the formation of antiwear tribofilms.[27,95,96]

2.3 Molecular Dynamics

2.3.1 The Need for MD

Surface experimental surface analysis techniques such as atomic force microscopy (AFM), Auger electron spectroscopy (AES), X-ray photoelectron spectroscopy (XPS), time-of-flight secondary ion mass spectrometry (ToF-SIMS) and transmission electron microscopy (TEM) are widely employed in tribology to investigate surface composition, morphology, and chemical transformations.[97,98,99,100] Similarly, specialist equipment such as the ball-on-disc mini traction machines (MTMs) are widely used in tribology to measure friction and wear under controlled loads, speeds, and temperatures.[101]

While these techniques provide valuable insights, they are limited in their ability to access the confined interfaces where tribochemical reactions often occur. Many spectroscopic methods rely on photon or electron probes that cannot penetrate beyond the outermost atomic layers, making it challenging to study reactions taking place within confined, sliding interfaces.[11,12] Even advanced in-situ techniques face obstacles such as limited spatial or chemical resolution, restrictions to vacuum environments or non-representative contact geometries.[102] As a result, our molecular-level understanding of tribochemical phenomena remains incomplete, and the development of new lubricants, additives or coatings often proceeds through empirical or trial-and-error approaches. MD simulations offer a powerful complementary tool by enabling direct observation of atomic-scale interactions and chemical events under well-controlled sliding conditions, even at interfaces that remain experimentally inaccessible.[15]

2.3.2 Different Types of MD Simulation

There are several types of MD simulation, each tailored to specific physical conditions or modelling goals. In classical MD, atoms are treated as point particles interacting through predefined force fields, making it computationally efficient for relatively large systems of up to millions of atoms or up to microsecond timescales.[103] Reactive MD, using forcefields such as ReaxFF, allows for the making and breaking of chemical bonds, making it suitable for studying chemical processes like oxidation, decomposition, or tribofilm formation,[16] although it can be at least 20 times more computationally expensive than non-reactive MD.[104,105] With ab initio MD (AIMD), interatomic forces are derived from quantum mechanical calculations, enabling high-fidelity simulations of electronic structure.[17] Various types of AIMD have been used to study tribochemical processes.[106,107,108,109] However, such simulations are orders of magnitude more computationally expensive than ReaxFF simulations.[104] A relatively recent addition that bridges the cost and accuracy of AIMD and ReaxFF are machine learning potentials.[110] Although these have been applied to some systems in tribology,[111,112,113] they are still in their infancy compared to the aforementioned methods.

A further distinction is made between equilibrium molecular dynamics (EMD) and non-equilibrium molecular dynamics (NEMD). EMD focuses on time-averaged properties under equilibrium conditions, such as diffusion coefficients, viscosity, and radial distribution functions,[114] using techniques like the Green-Kubo formalism.[115,116] In contrast, NEMD introduces external gradients (e.g., shear, temperature, or pressure) to probe transport processes or response behaviour directly.[117] For lubricant research, EMD is well-suited for studying bulk fluid properties, while NEMD is essential for simulating shear-thinning behaviour, interfacial slip, and stress-dependent properties that are critical for understanding performance under relevant operating conditions.[15] One key limitation of NEMD simulations is that strong fields are required in order to obtain a sufficient signal-to-noise ratio over the available simulation time. In tribology simulations, this manifests as high shear rates or sliding velocities compared to experiments.[15] Methods such as the transient-time correlation function (TTCF) technique allow experimental shear rates and sliding velocities to be directly simulated, although thus far this has only been applied to relatively simple systems.[118,119] This thesis briefly explores the theory behind classical and reactive MD simulations and the reader is directed to *'Ab initio molecular dynamics: basic theory and advanced methods'*[120] for further information on AIMD methods.

Integrators

MD simulates the time evolution of atomic or molecular systems by numerically integrating Newton's equations of motion. One of the most commonly used integration algorithms is the velocity-Verlet method, which updates both positions and velocities at each timestep. It is favoured for its simplicity, good energy conservation properties, and its simple integration with several thermostat and barostat algorithms.[121,122] The velocity-Verlet algorithm suitable for systems with uniformly fast dynamics where a single, fixed timestep can adequately resolve all relevant motions. A less computationally-intensive alternative is the reversible reference system propagator algorithm (RESPA), which uses a multiple timestep integration approach that partitions forces into fast and slow components. By allowing

shorter timesteps for rapidly changing forces (e.g., bond stretching, angle bending) and longer timesteps for slowly evolving interactions (e.g., electrostatics, van der Waals interactions), RESPA enables more time-efficient simulations of complex systems without significantly compromising accuracy.[123]

Energy Minimisation, Ensembles, Thermostats and Barostats

MD simulations are conducted within thermodynamic ensembles each representing different physical conditions (e.g., constant temperature, pressure, volume, etc.). Each ensemble constrains a specific set of variables and is used depending on the physical situation being modelled. To conduct simulations within these ensembles, MD employs coupling mechanisms known as thermostats and barostats to regulate temperature and pressure, respectively.

Thermostats maintain the kinetic energy of a system thereby maintaining a target temperature, enabling simulations in ensembles such as the canonical ensemble (NVT) or the isothermal-isobaric ensemble (NPT). The Nosé–Hoover thermostat achieves temperature control by extending the system's Hamiltonian (the total energy function that governs the time evolution of the system's state) to include an additional degree of freedom representing a heat reservoir.[124,125] It has been widely used in the NEMD simulations of lubricants under periodic boundary conditions.[15,126] However, in confined systems, such as the thin fluid films trapped between solid slabs investigated in this study, direct application of this thermostat to the fluid can artificially suppress physical phenomena like shear-induced heating.[127]

The Langevin thermostat applies stochastic forces and frictional damping to mimic collisions with an imaginary heat bath, effectively removing excess energy in a more localised manner.[128,129] This makes it better suited for thermostatting simulations via the confining walls of a confined NEMD simulation, where it allows thermal gradients to form naturally in the fluid region without disturbing the underlying frictional dynamics.[130] Each thermostat thus presents trade-offs between statistical fidelity, numerical stability, and physical realism, and their use must be tailored to the geometry and objective of the simulation.

Barostats, such as the Nosé–Hoover and Berendsen methods, enable pressure control in MD simulations by dynamically adjusting the volume of the simulation box in response to internal pressure fluctuations, thereby maintaining constant pressure and allowing sampling of the NPT ensemble.[125,129,131] In bulk simulations of lubricants, the Nosé–Hoover barostat is commonly used to regulate the system volume in a manner consistent with extended Hamiltonian dynamics, enabling realistic density and phase behaviour. The Berendsen barostat regulates system pressure by continuously scaling the simulation box vectors and atomic coordinates to relax the pressure toward a reference value, following a first-order kinetic equation with a defined time constant. While the Berendsen barostat efficiently equilibrates pressure and stabilises the system, it artificially suppresses natural volume fluctuations, meaning it does not rigorously sample the NPT ensemble.[132]

In confined systems where fluids or additives are trapped between rigid slabs representing surfaces, pressure is more appropriately controlled by applying a constant normal force to the confining walls. This method more accurately represents mechanical loading conditions and avoids the unphysical artifacts

that can arise from fixing slab separations.[133] In this approach, the system is first energy-minimised with a larger initial gap between the slabs, and the normal force is applied only after minimisation. Because this sudden application of force can induce excessive oscillations in the slab separation, damping terms are often introduced to stabilise the system's response.[133] Hence, as with thermostatting, the choice of barostatting strategy depends on the geometry and intended physical representation of the simulated system.

In the microcanonical ensemble (NVE), energy is conserved and no thermostat or barostat is applied. This ensemble is good for studying fundamental dynamical properties such as thermal conductivity, where the natural fluctuation and propagation of energy should remain undisturbed by external controls.[114] The NVT ensemble maintains constant temperature using thermostats and is commonly used for equilibrium simulations where accurate control of thermal conditions is required, such as in the calculation of transport properties like shear and bulk viscosity through Green–Kubo relations.[114] The NPT ensemble also includes pressure control via barostats, allowing the simulation box to fluctuate in volume. This ensemble is especially well-suited for studying systems mimicking experimental conditions where both temperature and pressure are held constant, and is useful for calculating equilibrium density.[105,114]

To accurately compute transport properties such as self-diffusivity and viscosity in liquid systems, Maginn et al.[114] state that the NVE ensemble should be preferred NVT or NPT ensembles, as thermostats and barostats can artificially alter system dynamics and skew results. However, because properties are typically desired at specific temperatures and pressures, a sequence of ensemble simulations often used: first equilibrate density using NPT at desired temperature and pressure, then fix volume in NVT to match the average density from the NPT run, then finally perform production run in NVE using the configuration from the end of the NVT run.[114] While NVE remains best practice, NVT can yield statistically equivalent results for viscosity and diffusivity when using properly tuned velocity-scaling thermostats such as Berendsen, stochastic rescaling, or Nosé–Hoover with weak to moderate coupling.[134,135] In contrast, thermostats that randomise velocities, like the Andersen barostat[136] or strongly coupled Langevin methods, significantly distort transport properties and should be avoided.[114]

2.3.3 Force fields

In MD simulations, force fields are essential for approximating the potential energy surface of a molecular system without explicitly solving its electronic structure. Atoms, atom groups, or entire molecules are treated as classical particles, and their interactions are described using mathematical functions with tunable parameters. The customisability of these parameters provides the flexibility required to characterise a wide range of chemical systems. Force field parameters are derived from experimental data (Class I) or quantum mechanical calculations (Class II), optimised to reproduce the structural and dynamic behaviour of representative systems. Class I forcefields have been shown to outperform Class II forcefields for the prediction of alkane properties.[137] Machine learning (ML) force fields attempt to learn the statistical relation between chemical structure and potential energy without relying on a preconceived notion of fixed chemical bonds or knowledge about the relevant interactions.[138]

Depending on the level of resolution required, different levels of force field complexity can be employed.

All-atom (AA) force fields represent every atom explicitly, including hydrogen atoms, and offer detail that is useful for capturing intramolecular interactions. United-atom (UA) force fields simplify the model by grouping non-polar hydrogens with their parent carbon atoms, which reduces computational cost while retaining accuracy.[139] Coarse-grained (CG) force fields abstract collections of atoms into single interaction sites or 'beads', enabling the simulation of even larger systems or longer timescales.[140] The choice of force field resolution represents a balance between computational efficiency and chemical accuracy, and should be guided by the specific goals of the simulation. In the section below, we briefly review non-reactive (classical), reactive force fields, and ML force fields.

Classical Force fields

Classical force fields represent the total potential energy of a molecular system as the sum of bonded and non-bonded contributions. Bonded terms model interactions between atoms connected by covalent bonds, including bond stretching, angle bending, and torsional rotation. Non-bonded terms describe interactions between atoms that are not directly bonded, such as van der Waals and electrostatic forces.[141] The general form of a classical force field can be written as:[142]

$$E_{\text{total}} = \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] + \sum_{\text{non-bonded}} [V_{\text{LJ}}(r) + V_{\text{Coulomb}}(r)]$$

where k_b is the bond force constant, r is the bond length, and r_0 is the equilibrium bond length. k_a is the angle force constant, θ is the bond angle, and θ_0 is the equilibrium bond angle. V_n is the torsional barrier height, ϕ is the dihedral angle, n is the periodicity of the torsional potential (the number of energy minima in a full 360° rotation around the dihedral angle, and δ is the phase offset (the angular position of the minimum). The final term includes the non-bonded interactions. The van der Waals contribution is usually described by the Lennard-Jones potential and the electrostatic contribution described by the Coulomb potential. The Lennard-Jones potential can be written as:[143]

$$V_{\text{LJ}}(r_{ij}) = 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

where r_{ij} is the distance between atoms i and j, ε_{ij} is the depth of the potential well (representing the strength of the interaction), and σ_{ij} is the distance at which the inter-atom potential is zero. In practice, non-bonded interactions such as the Lennard-Jones potential are typically computed only within a fixed cut-off distance (commonly around 14Å).[105] Beyond this cut-off, the interaction energy is assumed to be negligible, significantly reducing computational cost without substantially affecting the accuracy of bulk properties in many systems.[105] The Coulomb potential can be written as:[144]

$$V_{\text{Coulomb}}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$

where q_i and q_j are the partial charges on atoms i and j, r_{ij} is the distance between them, and ε_r is the relative permittivity. As Coulombic interactions decay slowly with distance, their accurate computation requires long-range solvers. Methods such as Ewald summation[145] and its more efficient variant, the

particle–particle particle–mesh (PPPM) algorithm[146], are used to handle electrostatic interactions across periodic boundary conditions with manageable computational cost.

Popular classical force fields such as AMBER,[147] CHARMM,[148] and OPLS[149] use this general framework. Class II force fields such as COMPASS[150] and PCFF[151] include additional cross-terms to capture coupling effects between bonds, angles, and torsions. These cross-terms account for interactions such as bond–bond, bond–angle, angle–angle, bond–torsion, and angle–torsion coupling, as well as out-of-plane bending. By incorporating these terms, Class II force fields offer improved accuracy for condensed-phase simulations conformational flexibility, though this comes at the cost of increased computational expense.[152]

Reactive Forcefields

Reactive force fields enable the simulation of systems undergoing chemical reactions by dynamically updating bond connectivity during MD simulations. Unlike classical force fields, which do not account for bond formation or dissociation, reactive force fields use bond-order-based approaches to model chemically reactive processes, bridging the gap between classical molecular mechanics and quantum mechanical (QM) methods.[153,154]

Among the available reactive potentials, ReaxFF (Reactive Force Field) is one of the most widely used due to its general applicability across elements and phases.[154] It was developed to approximate quantum mechanical accuracy at a fraction of the computational cost, making it especially useful for simulating chemically reactive systems under nonequilibrium conditions at scales larger than those accessible using QM methods.[154] The total energy of a system simulated using ReaxFF is expressed as a sum of several components:[155]

$$E_{\text{system}} = E_{\text{bond}} + E_{\text{over}} + E_{\text{angle}} + E_{\text{tors}} + E_{\text{vdW}} + E_{\text{Coulomb}} + E_{\text{specific}}$$

Here, E_{bond} is a continuous bond energy function derived from interatomic distances and bond orders. E_{over} penalises over-coordination, ensuring chemically reasonable valency. E_{angle} and E_{tors} model angular and torsional strain respectively. E_{vdW} and E_{Coulomb} represent non-bonded van der Waals and electrostatic interactions, calculated between all atoms irrespective of connectivity. E_{specific} includes terms for lone pairs, hydrogen bonding, conjugation, and other system-specific corrections.

A central concept in ReaxFF is the use of a continuous bond order, which determines the strength and existence of a bond between two atoms based on their interatomic distance. The bond order between atoms i and j, denoted BO_{ij}, is typically given by:

$$BO_{ij} = BO_{ij}^{\sigma} + BO_{ij}^{\pi} + BO_{ij}^{\pi\pi}$$

Each term corresponds to contributions from sigma, pi, and double-pi bonding, respectively. These are

modelled as exponential functions of the interatomic distance r_{ij}, for example:

$$BO_{ij}^{\sigma} = \exp\left[p_{bo1}\left(\frac{r_0}{r_{ij}}\right)^{p_{bo2}}\right]$$

where r_0 is the bond distance at unit bond order, and p_{bo1}, p_{bo2} are fitted parameters. The bond order decreases smoothly as atoms move apart, allowing simulations to realistically account for bond formation and breaking. As charge distribution is dynamic in reactive systems, ReaxFF employs the charge equilibration (Qeq) method,[156] which updates atomic charges at each timestep by minimising the system's electrostatic energy subject to charge conservation. This enables ReaxFF to capture polarisation effects and long-range charge transfer without resorting to quantum calculations.[156]

ReaxFF can model atomic systems across the solid, liquid and gas phases with the same formalism, allowing simulations of phase transitions or mixed-phase interfaces. However, the accuracy of ReaxFF simulations depends heavily on the quality of the QM data used for the parametrisation, typically using density functional theory. These parameters are system-specific and their development is time-consuming and computationally intensive.[154] Despite these challenges, ReaxFF parameter sets have been developed for a range of lubricant and additive molecules, including alkanes and alcohols,[157] phosphoric acids,[158] phosphate esters,[159] and organo-sulphur compounds.[160] Using such force fields, systems at the nanoscale can be subjected to compression and shear in a similar manner to conventional confined tribological simulations, while tracking the resulting chemical reactions over timescales on the order of nanoseconds.[16]

ML force fields

ML force fields are generated by training models on high-quality reference data, such as energies and atomic forces computed from ab initio methods[138] While this approach is relatively new, some studies have attempted to generate such force-fields for lubricant-like molecules, such as the study by Hajibabaei et al., who created an ML force field for alkane and polyene hydrocarbons between C_8 and C_{64}.[161] ML force fields have also been used for tribochemical NEMD simulations of gallate esters[111,113] and hypercin, a precursor of carbon-based tribofilms.[112] Unlike conventional potentials, ML force fields do not require a predefined functional form or bonding topology, allowing them to flexibly capture complex, reactive chemical behaviour. Their main advantage is the ability to approach quantum chemical accuracy while maintaining computational efficiency comparable to classical force fields. However, their performance depends strongly on the diversity and quality of the training data, meaning they may fail to generalise reliably outside the domain covered by the reference set.[138]

2.4 Genetic Algorithms

Genetic algorithms (GAs) are population-based stochastic optimisation techniques inspired by the principles of natural selection and evolution.[162] They operate by evolving a population of candidate solutions over successive generations through genetic operators such as selection, crossover, and mutation. At each generation, individuals are evaluated according to a fitness function that quantifies their per-

formance with respect to a given objective. The most promising candidates are selected to generate offspring, allowing the population to converge toward optimal or near-optimal solutions. GAs are able to effectively explore large, complex, and non-convex search spaces where gradient information is unavailable or unreliable. They have been widely applied to problems in molecular design, drug discovery, materials screening, and several other fields, where traditional exhaustive search methods are computationally prohibitive.[163,164,165,166,167,168]

The design of lubricant base oils involves the simultaneous optimisation of multiple, often conflicting objectives such as viscosity, thermal conductivity, biodegradability, and synthetic feasibility. This inherently represents a multi-objective optimisation problem (MOOP), where the goal is to identify candidates that achieve the best possible trade-offs between target properties.

The general form of a MOOP can be expressed as:[169]

$$\begin{aligned} &\text{Maximise} \quad \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})] \\ &\text{where} \quad \mathbf{x} \in \mathscr{X} \end{aligned} \tag{2.3}$$

Here, x denotes the vector of variables that define a lubricant candidate such as molecular descriptors, topological features, or formulation parameters. The vector-valued objective function $f(x)$ consists of m scalar objectives, each corresponding to a target property or performance criterion. Additional preferences or limitations may be handled by explicit inequality and equality constraints, penalised objective terms or post-algorithm filtering of solutions.[170]

MOOPs can be addressed either by transforming them into a single-objective, such as through the weighted sum method, to yield a single optimal solution,[171] or by approximating the Pareto front: the set of non-dominated solutions where improving one objective necessitates the compromise of at least one other objective.[172]

2.4.1 Non-dominated Sorting Genetic Algorithm II

Given M objectives and a population of N individuals (candidate solutions), NSGA-II proceeds by identifying non-dominated solutions, promoting both convergence to the Pareto front and diversity among solutions. NSGA-II is designed to solve problems of the form defined in Eq. (2.3), using evolutionary operators and a non-dominated sorting procedure to approximate the Pareto front.[173]

A solution a is said to dominate another solution b (denoted $\mathbf{a} \prec \mathbf{b}$) if:

$$\forall i \in \{1,\ldots,M\}, \quad f_i(\mathbf{a}) \leq f_i(\mathbf{b}) \quad \text{and} \quad \exists j \in \{1,\ldots,M\} \text{ such that } f_j(\mathbf{a}) < f_j(\mathbf{b})$$

That is, a is no worse than b in all objectives and strictly better in at least one.[174] NSGA-II involves three main steps: non-dominated sorting, crowding distance calculation, and elitist selection. First, the population is sorted into Pareto fronts $\mathscr{F}_1, \mathscr{F}_2, \ldots$, where each front contains individuals that are non-dominated with respect to one another. The first Pareto front $\mathscr{F}_1$ consists of the best trade-off solutions, those for which no other individual in the population performs better across all objectives.

These represent the Pareto-optimal set, forming the approximation of the Pareto front for that set of objectives. Subsequent fronts $\mathscr{F}_2, \mathscr{F}_3, \ldots$ are formed by removing the current non-dominated solutions and repeating the process. Each front is ranked according to dominance such that:

$$\mathscr{F}_1 \prec \mathscr{F}_2 \prec \mathscr{F}_3 \prec \ldots$$

where $\mathscr{F}_1$ dominates all individuals in $\mathscr{F}_2$, and so on. This hierarchy guides the selection process in evolutionary algorithms, favouring solutions closer to the true Pareto front $\mathscr{F}_1$.

Next, a crowding distance d_i is assigned to each solution within a Pareto front to promote and preserve diversity among solutions. This metric estimates how isolated or crowded a solution is relative to its neighbours in the domain of potential solutions (the objective space), helping ensure that the search explores a wide range of trade-offs rather than clustering around a small region of solutions. For each objective j, the crowding distance of an individual i is computed as:[175]

$$d_i = \sum_{j=1}^{M} \frac{f_j^{i+1} - f_j^{i-1}}{f_j^{\max} - f_j^{\min}}$$

where M is the total number of objectives, f_j^{i+1} and f_j^{i-1} are the objective values of the neighbouring solutions adjacent to individual i when the population is sorted by objective j, and $f_j^{\max}$ and $f_j^{\min}$ are the maximum and minimum values of objective j within the current front. By summing the normalised differences across all objectives, d_i provides a scalar estimate of how 'spread out' a solution is compared to its neighbours. Individuals are then selected based on Pareto rank and crowding distance. The next generation is filled by iteratively adding individuals from the lowest-rank fronts, breaking ties using the crowding distance, with the total number of candidates passed to the next generation defined by the elitism parameter (a proportion of the total generation size).[176] Prior to evaluation, all objective values are often normalised to ensure comparability across different scales, supporting both the sorting and hypervolume calculation processes.[177]

The hypervolume (HV) metric is a widely used indicator for evaluating the quality of a set of non-dominated solutions in multi-objective optimisation.[178,179] It quantifies both convergence (closeness to the true Pareto front) and diversity (spread across the front) in a single scalar value. Specifically, hypervolume measures the portion of the objective space that is dominated by the Pareto front and bounded by a user-defined reference point r:

$$HV(P, \mathbf{r}) = \lambda \left(\bigcup_{\mathbf{f}_i \in P} [\mathbf{f}_i, \mathbf{r}] \right)$$

where P is a set of non-dominated solutions, with each f_i being a point in the objective space, [f_i, r] represents the hyper-rectangle formed between solution f_i and the reference point r, and λ denotes the Lebesgue measure, a generalisation of length (1D), area (2D), volume (3D), and so on, to higher dimensions.[180]

A larger hypervolume indicates that the solutions collectively dominate a greater portion of the objective space, implying better performance and coverage of the Pareto front.[181] To ensure meaningful comparison, objective values should be normalised to all fall in the range $[0,1]$, where the reference point r is chosen such that it lies worse than the worst expected value in each objective (e.g., r = [1.1, 1.1, ...] for normalised problems). In theory, hypervolume is also monotonic, meaning that adding a new non-dominated solution will never decrease the HV value, as it can only increase or remain unchanged.[182] In practice, this behaviour may not hold when using NSGA-II for multi-objective optimisation, especially in molecular design tasks. The algorithm may struggle to maintain population diversity or may converge prematurely, resulting in the selection of solutions that do not meaningfully expand the Pareto front.[183] As a result, hypervolume may stagnate or even decrease across generations, but this behaviour can be mitigated by the appropriate choice of hyperparamters.

2.4.2 Weighted Sum Genetic Algorithms

Weighted Sum Genetic Algorithms (WSGA) address multi-objective optimisation by converting multiple objectives into a single composite objective through a linear weighted combination.[184] Given M objectives and a population of N individuals (candidate solutions), WSGA optimises a scalarised objective function of the form:[185]

$$\text{Minimise (or Maximise): } F(\mathbf{x}) = \sum_{i=1}^{M} w_i f_i(\mathbf{x}),$$

where $x \in \mathscr{X}$ are the decision variables, $f_i(\mathbf{x})$ are the individual objective functions, and $w_i \geq 0$ are user-defined weights satisfying $\sum_{i=1}^{M} w_i = 1$. The scalarised function $F(\mathbf{x})$ defines a single-objective optimisation problem that can be solved using standard GA operators such as selection, crossover, and mutation. The choice of weights w_i directly controls the trade-offs between objectives where larger weights place more emphasis on optimising the corresponding property. By varying the weight vector w across runs, it is possible to sample different regions of the Pareto front as a single run of WSGA targets only a single combination of objectives. Considering the weighted sum formulation assumes linear trade-offs, it can struggle to find non-convex regions of the Pareto front.[186]

Selection pressure in WSGA is guided entirely by the scalarised function $F(\mathbf{x})$, and the evolutionary operators aim to minimise (or maximise) this value across generations. After initialisation, individuals are evaluated based on their composite fitness, and standard GA procedures (e.g., tournament selection,[187] roulette wheel selection[188]) are employed to choose parents for reproduction. Offspring are generated via crossover and mutation, inheriting and evolving characteristics of their parents, and the next generation is formed by selecting the best candidates based on the scalarised function value. Elitism is incorporated by preserving a subset of the best-performing individuals across generations. While relatively simple to implement, the relicance of WSGA on fixed weights requires careful selection to balance competing objectives appropriately, and no explicit mechanisms exist within the framework to promote diversity across solutions as in Pareto-based methods.[185] Methods such as niching allows for some incorporation of diversity in WSGA implementations.[189] Nonetheless, WSGA provides a computationally efficient and interpretable method for multi-objective optimisation tasks and is especially appropriate where the

approximate relative importance of objectives is known or can be reasonably assumed.

2.5 Artificial Intelligence for Molecular Generation and Molecular Property Prediction

In the drug-like compound chemical space, it has been estimated that there are between 10^{23} and 10^{60} potentially synthesizable molecules, while only around 10^{8} have been synthesized.[190] While tribological fluids may reside in a slightly different chemical domain, it is easily recognised that there are several undiscovered lubricant molecules, with the immense combinatorial space making exhaustive experimental exploration infeasible. ML models can learn complex structure-property relationships from experimental data or from data generated by computational methods such as MD. Once trained, these models can rapidly predict the performance of new compounds, prioritize high-potential candidates, propose novel molecular structures or formulations, and guide synthesis or simulation efforts, especially when compared to MD simulations, which are often orders of magnitude slower.[191] The following section reviews modern ML methods for molecular property prediction and candidate generation, training, and data processing techniques required to achieve optimal performance.

2.5.1 AI and ML for Molecular Candidate Generation/Screening

Various ML methods have been used to propose new molecules for further evaluation and potential synthesis. Architectures such as (graph) variational autoencoders (VAEs), generative adversarial networks (GANs), Monte Carlo tree search (MCTS), recurrent neural networks (RNNs), and transformers have been employed to generate lubricant-like molecules. For example, Kajita et al.[29] produced a MCTS algorithm used in conjunction with rapid MD simulations based on the Einstein-Helfand[192] method to generate lubricant molecules optimised for dynamic and kinematic viscosity. They generated 55,000 candidate molecules, from which one molecule was chosen for synthesis, which was found to perform better than currently available oils, although there was a variation of almost 20% between the predicted and actual values for dynamic viscosity index, density, and kinematic viscosity. Zhou et al.[3] employed a GAN trained with reinforcement learning to automatically generate new lubrication oil molecules with desired properties (viscosity, pour point and flash-point), coupled with an artificial neural network (NN) trained on a dataset of 67 esters to predict the properties of newly generated molecules. Liu et al.[193] used a database of 9,252 hydrocarbon molecules to train a VAE and MLP to simultaneously propose over 11 million novel hydrocarbon molecules and predict their fuel-related physicochemical properties based on a Simplified Molecular Input Line Entry System (SMILES) representation using the group contribution method.

For general-purpose molecule generation, Christofidellis et al.[194] proposed a multi-domain, multi-task transformer model trained on 26k molecule-description pairs from the ChEBI-20 dataset (initially presented by Edwards et al.[195]) allowing users to generate suggested molecular structures with desired properties in SMILES string format based on text prompts.[194] More recently, Pirnay et al.[196] presented a graph-based molecular design method that combines transformers with reinforcement learning, to enable

the sequential construction of chemically valid molecules while supporting structural constraints, when starting from existing structures. They found that this approach outperforms existing approaches on benchmark drug and solvent design tasks.[196] Meanwhile, Yue. et al.[197] evaluated six generative models for polymer design, showing that CharRNN, REINVENT,[198] and GraphINVENT[199] perform best when trained on real polymer data, while REINVENT excels with hypothetical polyimide datasets. When combined with reinforcement learning for optimising glass transition temperature (T_g), all three models showed strong performance, with CharRNN notably generating more novel polymers and REINVENT producing the most favorable distribution of high T_g candidates, albeit with fewer unique and valid structures.[197]

An alternative approach to directly generating new candidates is to train models to evaluate molecules based on desired properties, to then screen large pre-existing molecular databases for optimal candidates. This approach was taken by Zhou et al., who combined MD simulations with ML models to screen the PoLyInfo,[200] PI1M,[201] and Polyimides datasets for polymers, optimising for VI, thickening efficiency and synthetic accessibility score.[32] This study also leveraged recursive feature elimination (RFE) to identify 12 optimised features to train models to predict viscosity of mixtures containing the screened viscosity improvers. Similarly, Huang et al. screened the PoLyInfo anf PI1M datasets for polymers optimised for thermal conductivity, by training random forest models on optimised topological features and molecular descriptors, optimising the descriptors with random sequential feature selection.[202]

Rule-based approaches offer an alternative strategy for molecular candidate generation, relying on methods such as predefined chemical transformation rules, substructure libraries, or heuristic algorithms (e.g., GAs) that encode expert knowledge or physical principles to construct chemically valid and synthetically accessible molecules.[203] For example, combinatorial fragment assembly,[204] retrosynthetic templates,[205] and graph-rewriting rules[206] have been successfully used to explore targeted regions of chemical space. Unlike many ML-based methods, which require large and diverse datasets for effective training, rule-based approaches can function with limited data and offer greater interpretability and control over the design process.[207] This makes them valuable in highly specialised domains, such as lubricant design, where high-quality data may be scarce and synthetic feasibility is critical. Furthermore, rule-based methods can be tailored to include domain-specific constraints, such as enforcing or limiting the presence of desirable or undesirable functional groups. However, ML candidate generation methods still offer the advantages of being able to uncover non-obvious relationships in data that are difficult to encode as explicit rules. Furthermore, ML approaches reduce the risk of excessively biasing the candidate generation process with user-defined constraints, which can limit exploration and potentially exclude high-performing molecules that fall outside conventional chemical intuition.[191]

2.5.2 ML for Property Prediction

ML methods are widely applied in computational chemistry and materials science to predict the properties of molecules and formulations, which generally requires the supervised learning approach, where training data is labelled prior to being passed to a model.[208] These methods are used for both regression tasks, such as predicting viscosity or solubility, and classification tasks, such as determining whether a

compound is biodegradable or not. Several studies have already incorporated moderm ML techniques for the prediction of the wear rate, wear volume and coefficient .[208,209] For example, Echavarri Otero et al. used an artificial neural network (ANN) to predict the friction coefficient of PAO-6 under different thermal elastohydrodynamic lubrication regimes, achieving an average error of 2% (maximum error 10%) when compared to results measured from mini-traction-machine (MTM) experiments.[210] Panwar et al. developed Gaussian process regression-based models to predict the temperature-dependent density and dynamic viscosity of 305 hydrocarbons, with the best models predicting density and dynamic viscosity with coefficient of determination (R^2) values of 99.6% and 97.7% respectively.[31] Studies have also used ML to predict properties of mixtures. Bilodeau et al. developed a graph-based neural network architecture to predict the viscosity of binary liquid mixtures as a function of composition and temperature, with the model predicting viscosity with a mean absolute error of 0.043 and a root mean square error of 0.080 in log cP units.[30]

ML approaches can be grouped into classical and deep learning methods.[211] Classical models typically rely on manually engineered features and use algorithms such as linear regression, support vector machines (SVM),[212] decision trees,[213] random forests,[214] and gradient boosting.[215] These models are generally shallow, containing only a few 'layers' of computation, meaning that these model process input data through a limited number of sequential transformations, typically involving only one or two steps between input and output, without hierarchical feature extraction.[216] These techniques are often effective on small or moderately sized datasets and offer a level of interpretability, however, they may not capture complex non-linear relationships in high-dimensional data.[217] Whereas, deep learning models are characterised by architectures with multiple layers of non-linear transformations that allow the model to automatically learn feature representations from data.[216] Examples include MLPs,[218] convolutional neural networks (CNNs),[219] transformers,[220] and graph neural networks (GNNs)[221] and they are particularly effective for large and complex datasets.[222]

The use of ML for property prediction provides significant advantages as these models can reduce the need for costly and time-consuming experiments or simulations as they can make fast predictions for large numbers of candidates. They also have the ability to identify relationships between input features and target properties, which can support discovery efforts.[223] However, ML models also have limitations. Their performance depends heavily on the quality and diversity of the training data, where small or biased datasets can lead to overfitting and hence, unreliable predictions. In addition, deep learning models are often difficult to interpret, which can be a drawback in domains that require interpretablity.[222] Despite these challenges, ML has become an essential tool for data-driven property prediction and plays an increasing role in both screening and candidate design. The following section provides a background for 3 of the most relevant ML algorithms used in this study.

Multi-Layer Perceptron

Multi-layer perceptrons (MLP) have been widely employed as methods for predictive chemistry.[224,225,226] By emulating the learning processes of the human brain, they can predict the properties of existing and potential lubricant formulations. The rising popularity of neural networks, driven by growing recog-

nition of their predictive power, has led to a wide range of architectures combining various structural elements and training techniques. Most networks follow a common architecture comprising an input layer, one or more hidden layers, and an output layer. Each layer consists of nodes (neurons), with the input layer receiving and preprocessing training data, the hidden layers performing transformations, and the output layer presenting predictions. The hidden layers carry out the bulk of computation using a process known as forward propagation.[227] Each neuron in a hidden layer computes a weighted sum of its inputs, adds a bias term, and applies a non-linear activation function. This operation determines the strength of the signal passed to the next layer. Mathematically, in a single hidden-layer network is formulated as:

$$\mathbf{h} = \sigma(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \quad \hat{\mathbf{y}} = \phi(\mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}),$$

where $\mathbf{x}$ is the input vector, $\mathbf{W}^{(1)}, \mathbf{W}^{(2)}$ are weight matrices, $\mathbf{b}^{(1)}, \mathbf{b}^{(2)}$ are bias vectors, and $\sigma(\cdot), \phi(\cdot)$ are activation functions. This formulation describes a simple feedforward neural network with a single hidden layer. The input $\mathbf{x}$ is first linearly transformed using the weights and biases of the first layer, followed by the application of a non-linear activation function σ, which yields the hidden layer output $\mathbf{h}$. This intermediate representation is passed through another transformation and a second activation function ϕ, producing the final prediction $\hat{\mathbf{y}}$.

Activation functions introduce non-linearity into the network, enabling it to model complex, high-dimensional relationships that would be difficult to capture with only linear operations. This non-linearity is especially important in domains like molecular property prediction, where many input features interact to influence target properties. Common choices for activation functions include the sigmoid, hyperbolic tangent (tanh), and Rectified Linear Unit (ReLU).[228,229] The choice of activation function significantly impacts network performance, where an effective activation function maintains sensitivity to variations in input values and supports efficient gradient propagation during training. Each function determines whether and to what extent a neuron is activated based on the weighted input it receives, which directly shapes the flow of information through the network and contributes to the final prediction, by determining whether the neuron should be activated.[230]

Neural networks are trained by adjusting their parameters (weights and biases) to minimise a loss function that measures the difference between predicted and actual outputs. This process begins with pseudo-randomly generated parameters. The choice of initialisation has been shown to significantly affect the convergence and final performance of the model, as poor initial values may lead to vanishing or exploding gradients. To combat this, Xavier initialisation has become a widely adopted standard for balancing variance across layers and promoting stable training, by setting the initial weights based on the number of input and output units in each layer, helping to maintain the scale of gradients throughout the network.[231] The initialised parameters are then iteratively refined using some form of gradient descent optimisation, such as stochastic gradient descent (SGD) or Adam.[232,233] These algorithms adjust the model's weights by computing the gradient of the loss function with respect to each parameter and updating the parameters in the direction that reduces the error. While SGD updates weights based on small batches of data to improve efficiency, Adam incorporates adaptive learning rates and momentum terms to accelerate convergence and handle noisy gradients.[234] The gradients are calculated through backprop-

agation, where the chain rule is used to determine how much each parameter contributes to the overall error.[235] The learning rate determines how much the weights and biases are adjusted in each iteration. If the learning rate is too low, training is slow, whereas if it is too high, the model may fail to converge or become unstable. Proper tuning of the learning rate is important to ensure effective optimisation.

The loss function used during training depends on the type of task. For regression tasks, such as predicting continuous properties of lubricant formulations, common loss functions include mean error, (root) mean squared error, and mean absolute error. For classification tasks such as determining whether a formulation meets toxicity or biodegradability criteria, cross-entropy loss is commonly used, where this measure compares the predicted probability distribution for each cost with the true class labels.[236] The neural networks architecture allows easy adaptation to different data types and problem scales, and they serve as foundational benchmarks for evaluating more advanced deep learning architectures. As such, even relatively simple feedforward networks are frequently used to establish baseline performance in property prediction pipelines before exploring more specialised or task-specific models.[237]

XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and scalable ML algorithm widely used for regression and classification tasks in cheminformatics.[237,238,239,240] It is based on gradient-boosted decision trees, where an ensemble of weak learners is trained sequentially. Each new tree aims to correct the errors of the previous ensemble by minimising an overall objective function that combines a loss term and a regularisation penalty.[241]

The objective function at iteration t is given by:

$$\mathscr{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k),$$

where $l(y_i, \hat{y}_i^{(t)})$ is a differentiable convex loss function (e.g., MSE for regression, logistic loss[242] for binary classification) measuring the discrepancy between the true value y_i and the predicted value $\hat{y}_i^{(t)}$, and $\Omega(f_k)$ is a regularisation term penalising model complexity. The regularisation term is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2,$$

where T is the number of leaves in the tree f_k, and w is the vector of leaf weights. To efficiently optimise this objective, XGBoost employs a second-order Taylor expansion of the loss function around the current predictions. For a given training instance i, this expansion is:

$$l(y_i, \hat{y}_i + f(x_i)) \approx l(y_i, \hat{y}_i) + g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2,$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}$ is the first-order gradient and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$ is the second-order derivative (Hessian)

of the loss function. System-level optimisations such as parallelisation, sparsity-aware algorithms, and regularised tree pruning contribute to its high computational efficiency.[241] Parallelisation accelerates training by distributing the construction of decision trees across multiple CPU cores. Sparsity-aware algorithms efficiently handle missing or zero-valued features by learning default directions for sparse entries. Regularised tree pruning helps prevent overfitting by removing splits that do not improve the model's objective function beyond a specified threshold. These features make XGBoost distinctly effective for structured data tasks, including property prediction and molecular classification based on molecular descriptors.

Transformers

The transformer architecture introduced by Vaswani et al.[220] was originally developed for processing sequential data, in an attempt to overcome the vanishing gradient problem present when training recurrent neural networks.[243] It uses a self-attention mechanism to compute pairwise dependency values between all input tokens, enabling efficient modelling of both short-range and long-range interactions.

The scaled dot-product attention is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, and V are the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the keys. Multiple self-attention heads are run in parallel and concatenated to form multi-head attention:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O,$$

where each head is computed as $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, and W_i^Q, W_i^K, W_i^V, W^O are trainable parameter matrices of the transformer. Positional encodings are added to the input to preserve sequence order information, and the outputs are passed through residual connections (concatenation of input and output to preserve flow of gradients during backpropagation), layer normalisation, and positional feedforward layers, forming the building blocks of transformer encoder layers.

Transformer architectures have proven highly effective for modelling molecular sequences such as SMILES or SELFIES, where chemically related substructures may be far apart in token space.[244,245] Given an input sequence $X = \{x_1, x_2, ..., x_n\}$, the self-attention mechanism dynamically computes the contextual relevance between every pair of tokens. This enables the model to capture structural and functional relationships that are not necessarily reflected by the order of tokens in the input. Transformers are not limited to sequence representations; they can also be applied to molecular fingerprints or descriptor vectors. In these cases, the fixed-length input is reshaped or projected into a sequence-like form, allowing the model to learn complex and non-linear interactions between features.[246]

An interesting property of transformer models is their ability to support transfer learning, where models initially trained on large datasets can subsequently be fine-tuned using smaller, task-specific datasets.

In the chemical domain, such a process would start with pretraining on large chemical datasets using self-supervised learning objectives such as masked language modeling, followed by the pretrained model being fine-tuned on a smaller dataset such as predicting molecular properties in a specific chemical domain. Such an approach has already been taken to create foundational models in chemistry such as MolBert,[247] Chemformer,[248] and ChemBerta,[249] with each of these models subsequently being finetuned in later studies for downstream tasks such as predicting properties like toxicity, solubility and retrosynthesis pathways.[194,250,251,252] Several techniques have been developed to enable efficient transfer-learning approaches, including freezing earlier layers of the network, using adapter modules with fewer trainable parameters, and applying layer-wise learning rate decay.[253,254] Such methods aim to retain general chemical knowledge from pretraining while learning task-specific patterns, making transformers especially valuable for molecular property prediction in where data in specific chemical domains is limited.

2.5.3 Data Processing For ML

Data preprocessing is a critical step in developing accurate and reliable ML models for molecular property prediction. It ensures that input features are properly structured, relevant, and scaled for the learning algorithms. Effective data preprocessing can improve model performance, reduce overfitting, and enhance interpretability. The following section explores data processing techniques relevant to this study, including dimensionality reduction methods such as PCA and UMAP, SMARTS-based substructure matching and RFE.

Dimensionality Reduction Techniques

Dimensionality reduction techniques, such as UMAP,[255] t-SNE[256] and PCA[257] are often used in cheminformatics to reduce noise, and enable efficient visualisation and analysis of large dataset.[258] These methods have been used to identify underlying patterns in high-dimensional feature spaces, such as those formed by molecular descriptors or fingerprints, and can support classification, similarity analysis, and model training.[32,259,260,261]

UMAP (Uniform Manifold Approximation and Projection) is a non-linear dimensionality reduction method that is based on principles from manifold learning and topological data analysis. It begins by constructing a weighted graph (e.g., by using k-means clustering) that captures the local neighbourhood relationships between data points in the high-dimensional space. UMAP then learns a low-dimensional representation by optimising an objective function that preserves the structure of the original graph. This objective minimises the cross-entropy between the high-dimensional graph and its low-dimensional embedding, resulting in a projection that retains both local and global data structure.[255]

Principal component analysis (PCA) is a linear transformation technique that identifies the directions of maximum variance in the data and projects the data onto a lower-dimensional space along those directions. It is especially effective when the input features are continuous and highly correlated, which is often the case with physicochemical molecular descriptors.[262] PCA works by identifying the directions in the data that show the most variation and then reorienting the data along those directions. These

directions, or principal components, represent new axes that capture the most important patterns in the data. By selecting only the top few components that account for the majority of the variation, PCA reduces the number of features while retaining as much useful information as possible. The original data is then re-expressed in terms of these selected components, resulting in a simplified, lower-dimensional version of the dataset.[257]

Unlike supervised dimensionality reduction techniques such as Partial Least Squares (PLS), PCA and UMAP does not incorporate information from the target variable during dimensionality reduction.[257] This makes dimensionality reduction practical in situations where target labels are noisy, partially missing, or entirely unavailable. Its unsupervised nature allows it to identify the dominant patterns in the input feature space without relying on outcome-specific associations, which may vary across datasets or experimental conditions. As a result, such techniques may be less likely to overfit to correlations present in the training data that do not represent real-world trends, and may generalise better to out-of-distribution samples. In practice, this means these methods could potentially serve as a robust preprocessing method in molecular property prediction tasks where label quality is inconsistent or where the goal is to develop models that transfer across different domains or molecular classes. This hypothesis is tested in Section 5.3.5.

Recursive Feature Elimination

RFE is a feature selection technique that iteratively removes the least important features based on a model's weight or importance ranking.[263] RFE begins by training a base estimator, such as a linear model or support vector machine, on the full set of features and ranking them based on their contribution to the prediction, where the contribution of each feature is calculated during model training as the weight of that feature which minimises a predefined loss function. The absolute value is used to reflect the strength of a feature's influence on the prediction, regardless of whether it increases or decreases the output. After each iteration, the algorithm removes a specified number of the least important features and retrains the model on the remaining subset. This process continues recursively until the desired number of features k is reached. By evaluating model performance (e.g., cross-validation score) at each iteration, RFE aims to identify the minimal subset of features that yields optimal generalisation.[264] RFE has been applied to a plethora of use cases involving high-dimensional data, ranging from cheminformatics and medical applications, to cybersecurity and urban planning.[32,265,266,267]

SMARTS Substructure Matching

Identifying specific chemical substructures within molecules is essential for tasks such as molecular classification, prediction, and filtering based on molecular functional groups.[268] SMARTS (SMiles ARbitrary Target Specification) is a powerful language for specifying molecular substructures using a rich syntax that extends the SMILES representation.[269] Substructure matching with SMARTS involves parsing the SMARTS pattern into a query graph, where atoms and bonds are labelled with expressions that define element types, hybridizations, charges, aromaticity, and connectivity. Given a target molecule (represented as a molecular graph with nodes for atoms and edges for bonds), a graph isomorphism

search is performed to determine whether a subgraph of the molecule satisfies the constraints of the SMARTS query.[270] Atom and bond queries can include recursive patterns, logical operators (e.g., AND, OR, NOT), and ring membership constraints. For example, the SMARTS pattern [C;R3] matches a carbon atom in a 3-membered ring.

2.6 Summary

The state-of-the-art regarding the systems relevant to this thesis has been presented and open problems in their research identified. The development of energy-efficient, environmentally-friendly lubricants remains constrained by limited understanding of molecular-scale mechanisms and the slow, empirical nature of formulation. Molecular dynamics has emerged as a powerful tool for probing friction, wear, and tribochemical reactivity at relevant length and time scales, yet its accuracy is limited by the quality of the underlying force fields. Recent developments in reactive force fields, particularly ReaxFF parameterised using DFT data, have enabled MD simulations of anti-wear additives like TCP, providing new insights into tribofilm formation.

Surface effects remain a key area of interest. The chemistry and topology of ferrous and DLC surfaces govern interfacial interactions, yet their dynamic behaviour in the presence of additives and solvents is not well understood. Modelling these interactions accurately requires integration of mechanochemical theory with surface-specific models, particularly under shear. While DLC coatings have demonstrated excellent friction and wear performance, their interactions with polar additives and water remain problematic. Surface functionalisation through doping offers a pathway to tune these properties, and recent DFT studies have begun to elucidate the mechanisms responsible for their behaviour.

Progress has been made in using artificial intelligence to guide lubricant design. ML models trained on simulation or experimental data can predict critical fluid properties such as viscosity, thermal conductivity, and density, reducing reliance on costly experimental work. These models, when combined with generative methods and optimisation algorithms, enable efficient screening of vast chemical spaces. However, challenges remain in accurately representing molecular structures and generalising models to out-of-distribution candidates.

A combined approach using reactive MD, ML, and genetic algorithms is increasingly viable for rational lubricant design. A molecular-level understanding of lubricant–surface–additive interactions, grounded in accurate simulation and data-driven prediction, will enable the development of next-generation lubricants with improved performance and sustainability.

Chapter 3

Effects of Surface Chemistry on the Mechanochemical Decomposition of Tricresyl Phosphate

This chapter presents a study of the mechanochemical decomposition of tricresyl phosphate (TCP), a widely used lubricant additive, under nanoscale confinement using nonequilibrium molecular dynamics (NEMD) simulations with the ReaxFF force field. A systematic investigation is carried out across a range of surface chemistries, solvent environments, temperatures, and normal stress conditions to understand the molecular processes that influence additive performance. This chapter provides a foundation for linking nanoscale decomposition mechanisms to macroscale tribological behaviour, supporting the development of more effective lubricant formulations.

3.1 Introduction

The work described in this Chapter has previously been published in Ref.[271].

Mechanochemistry describes a wide range of phenomena where mechanical force influences chemical reactivity.[272] This includes reactions in ball mills, where energy is transferred from moving balls to the reactants during collisions of the balls with each other and the jar walls.[273] Mechanochemical reactions can be performed without solvents or heating, which can provide considerable environmental benefits over conventional routes.[274] The choice of ball and jar materials are known to be affect reactivity for a range of processes, such as cocrystallisation[275] and cycloaddition,[276] emphasizing the important role of the solid material. In some cases, the milling balls themselves are the catalyst of the reaction, which is known at direct mechanocatalysis.[277] Another example and particularly high-value application of mechanochemistry is lubricant additives that reduce wear in tribological contacts.[278] The most popular antiwear additive for automotive lubricants, zinc dialkyldithiophosphate (ZDDP), forms protective tribofilms on rubbing surfaces through mechanochemical processes.[95,96,279]

The tribofilm growth of ZDDP is strongly dependent on the nature of the rubbing surfaces.[9] Surfaces with a high elastic modulus, such as tungsten carbide,[279] require lower loads to form tribofilms than steel,[96] due to higher contact stresses. In experiments using aluminium-silicon alloys, thicker ZDDP tribofilms form on the silicon phase than the aluminium matrix, which is also due to the higher elastic modulus and thus stress.[280] As well as the mechanical properties of the rubbing surfaces, the surface energy can also affect ZDDP adsorption and tribofilm growth.[281] ZDDP forms thicker tribofilms on surfaces to which it readily adsorbs, such as steel, silicon nitride, and tungsten carbide surfaces, than on those where minimal adsorption occurs, such as diamond-like carbon (DLC) and silicon carbide.[282] Ion-implanting alloying elements into steel can affect ZDDP tribofilm growth, nickel promotes tribofilm formation, molybdenum and chromium slows growth, while vanadium and tungsten have no significant effect.[283] ZDDP forms thicker tribofilms on aluminium-magnesium alloys than on either of the pure metals due to the stronger adsorption of the ZDDP molecules on the alloy.[109] A recent study revealed how the interplay of mechanical and chemical factors determine the tribological behaviour of ZDDP on DLC, with hydrogen-free coatings of moderate elastic modulus being most wear-resistant, while still allowing ultralow friction.[108] Surface chemistry can also affect other mechanochemical processes inside rubbing contacts. For example, the rate of α-pinene tribopolymerisation is higher on palladium, copper oxide, nickel oxide, and chromium oxide than gold, silicon oxide, aluminium oxide, and DLC surfaces.[284] Moreover, the α-pinene tribopolymerisation rate is higher on bare silica surfaces than hydroxylated silica.[285] These observations can be explained by the fact that mechanochemical processes are accelerated when the confined molecules chemisorb to one[286] or both[107] sliding surfaces, since the shear stress applied through the sliding surfaces results in more strain within the strongly adsorbed molecules.

Phosphate esters lubricant additives have numerous industrial applications due to their stability, particularly in the presence of oxygen.[287] Their versatility in structure including the ability to incorporate a variety of substituent groups has led to their application as lubricant additives to improve the antiwear, boundary lubrication, antioxidant, anti-corrosion, anti-foaming, and metal deactivation performance.[68] Phosphate esters come in various forms, including aromatic, alkyl, thiophosphates, and metal-containing esters, and are readily accessible for commercial use.[68] One of the most popular of these additives is tricresyl phosphate (TCP), which has been widely employed in aviation lubricants for gas turbine engines for several decades.[288] The molecular structure of TCP is shown in Figure 3.1. Like ZDDP, TCP is an effective antiwear additive for steel contacts.[289,290] An additional benefit of TCP for aviation lubricants is that the tribofilms it forms on metal surfaces slow the degradation of synthetic neopentyl polyol ester base stocks.[291]

It was initially suggested that the ability of TCP to reduce wear on steel surfaces originated from polishing of the surfaces through the formation of iron phosphide.[288] However, later studies have shown that the major component of the tribofilms formed from TCP on steel was iron phosphate, which acted as a protective layer, rather than polishing the surfaces.[292,293,294] It has also been shown that either organic polyphosphates[295] or iron polyphosphates[296,297] are eventually formed from TCP inside tribological contacts, which are ultimately responsible for its antiwear performance on steel surfaces. At

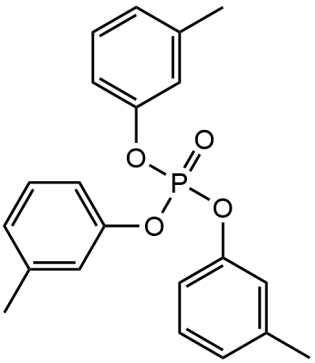


Figure 3.1: Molecular structure of tri-meta-cresyl phosphate.

high temperatures, TCP molecules form relatively thick (60–100 nm) thermal films on steel surfaces;[298] however, it forms thinner tribofilms ($\sim$ 20 nm) on rubbing surfaces.[290] This is much thinner that the ZDDP tribofilms that form on rubbing steel surfaces ($\sim$ 100 nm).[290]

The thermal decomposition of TCP has been studied on several different substrates. Wheeler and Faut[299] used X-ray photoelectron spectroscopy (XPS) to compare the adsorption and dissociation of TCP on gold and iron surfaces. On both substrates, adsorption saturated at about one monolayer, but adsorption was slower on gold than on iron.[299] TCP underwent non-dissociative adsorption on gold and desorbed upon heating, with no TCP left on the surface at 473 K.[299] On iron, no desorption occurred upon heating and additional interactions were observed between the substrate and the tolyl groups on the TCP molecule.[299] TCP decomposed on iron between 423 K and 473 K with one of the tolyl groups being removed to form iron phosphate.[299] The temperature range for decomposition is close to the temperature at which TCP significantly reduced the friction of rubbing steel surfaces in tribometer experiments.[294] It has since been shown that the thermal decomposition of TCP on iron surfaces occurs mainly via P-O bond scission to form methylphenoxy intermediates.[300] When TCP was heated to 800 K on iron, the methylphenoxy intermediates either desorbed as cresol via hydrogenation or decomposed further to generate tolyl intermediates.[300] Some of these tolyl intermediates desorbed as toluene via hydrogenation, but the majority decomposed resulting in hydrogen and carbon monoxide desorption and carbon deposition onto the iron surface.[300]

The steel surfaces inside tribological contacts are chemically heterogeneous and are covered by multiple different iron oxides.[75,76] From experiments conducted at high temperature (750 K), it was suggested that, due to acid-base reactions, TCP vapor preferentially reacts with iron in higher oxidation states, with reactivity increasing in the order: Fe < FeO < Fe_3O_4 < Fe_2O_3.[296] However, more recent experiments suggested that reactivity towards TCP increases in the order Fe_3O_4 < Fe < Fe_2O_3.[301] In both of these experimental studies, the iron surfaces will also form oxide layers that react with the TCP, rather than the native metal.[301] In addition to surface oxides, metal carbides are another important component of bearing steels.[302] A previous study of thermal decomposition found that TCP reactivity increases in the order: tungsten carbide < molybdenum carbide < chromium carbide < vanadium carbide.[303]

In recent years, experimental studies of TCP adsorption and decomposition have been complemented by

molecular simulations. For example, Khajeh et al.[159] used molecular dynamics (MD) simulations with a reactive force field (ReaxFF) to study the adsorption of TCP on hydroxylated amorphous Fe_3O_4. They found that, at lower temperatures ($\leq$ 500 K), chemisorption occurred mostly via Fe-C bond formation through the tolyl groups. At higher temperatures (600-700 K), a significant amount of Fe-O and O-P bonding also occurred, indicating chemisorption through the central phosphate group. Ewen et al.[304] compared the chemisorption and dissociation of TCP to trialkyl phosphates on Fe_3O_4(001) surfaces. In agreement with previous experiments,[300,305] they found that the triaryl phosphate TCP was much less reactive compared to the trialkyl phosphates. The MD simulations also showed that, while the main decomposition mechanism for trialkyl phosphates was C-O bond dissociation, the only bonds that broke in TCP were P-O bonds.[304] For the trialkyl phosphates, Ewen et al.[304] also compared the adsorption and dissociation on Fe_3O_4(001) to hydroxylated amorphous Fe_3O_4 and α-Fe surfaces. They found that the reactivity of the trialkyl phosphates increased in the order: hydroxylated amorphous Fe_3O_4 < Fe_3O_4(001) < α-Fe. Ayestarán Latorre et al.[27] studied the mechanochemical decomposition of trialkyl phosphates between sliding iron surfaces using non-equilibrium molecular dynamics (NEMD) simulations with ReaxFF. They found that secondary trialkyl phosphates were much more reactive than primary ones,[27] in agreement with previous experiments on ZDDP.[96] The influence of surface chemistry on the mechanochemical response of organophosphates is yet to be explored either experimentally or through NEMD simulations.

In this study, we use NEMD simulations with ReaxFF to compare the mechanochemical decomposition of TCP molecules confined between different ferrous surfaces over a range of temperatures and pressures. We observe significant differences in the mechanochemical reactivity of TCP on iron, iron oxide, and iron carbide surfaces. The variations in reactivity can be explained by differences in surface energy and stiffness, which can be accounted for in theoretical models for mechanochemistry.

3.2 Methodology

3.2.1 System Setup

We used NEMD simulations [307] to compare the mechanochemical reactivity of TCP on three ferrous surfaces, as shown in Figure 3.2. This is a molecule composed of three cresyl (methylphenyl) groups linked to a phosphate group. TCP is produced from cresol, which naturally contains a mixture of isomers with the methyl group in the ortho-, meta- and para-positions. The ortho-isomers of TCP are known to be toxic and modern production methods have reduced the concentration of these to very low levels.[308] Therefore, the TCPs used in current aviation lubricants generally contain mixtures of the meta-isomers and also some para-isomers.[308] The para-isomers were shown to have poorer lubricating properties compared to mixtures of isomers.[293] However, a recent study showed that the para-isomer is only slightly more reactive than the meta-isomer on ferrous surfaces.[301] In this study, we chose to study pure tri-meta-cresyl phosphate for simplicity; its molecular structure is shown in Figure 3.1.

All the systems were constructed using the Materials and Processes Simulations (MAPS) platform from Scienomics SARL. The three model surfaces were native iron, cementite, and magnetite. Symmetrical

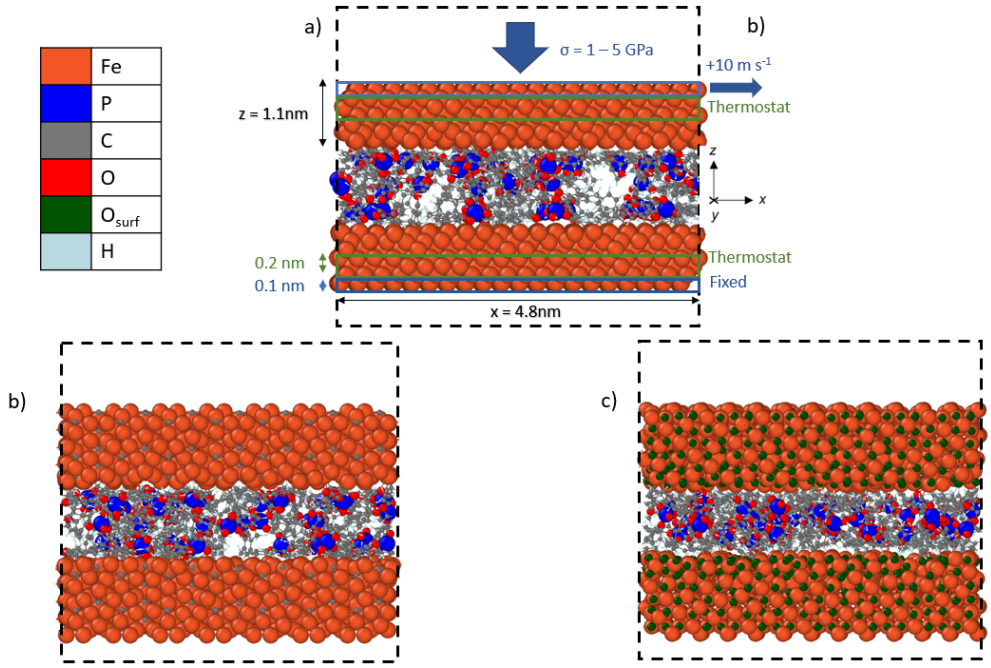


Figure 3.2: Side-view snapshots showing the model systems with TCP molecules confined between a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001)-SCV surfaces. Systems shown after minimisation and equilibration, but before heating, compression, and shear. Black dashed line shows the simulation box. Rendered using OVITO.[306]

contacts were used with the same material for the upper and lower surface. Native iron surfaces were considered due to the possibility that they are exposed to the additive following removal of surface oxides in rubbing steel contacts.[309] Cementite was chosen as a representative iron carbide surface because it was identified in previous thermal degradation experiments using TCP.[310] Magnetite was selected since previous experiments have shown that it is the major oxide formed on the surface of bearing steel following tribometer experiments.[75] We used an α-Fe(110) for the iron surface,[311] Fe_3C(010) as the cementite surface,[312] and Fe_3O_4(001) with a subsurface cation vacancy (SCV) reconstruction as the magnetite surface[313] due to their relatively low surface energies.

Periodic boundary conditions were applied in the x and y directions. Since the materials all had different lattice constants, it was not possible to create periodic surfaces with exactly the same dimensions. The α-Fe(110) surfaces had dimensions of x = 4.8, y = 4.8 and z = 1.1 nm. The Fe_3C(010) surfaces had dimensions of x = 4.9, y = 4.7 and z = 1.05 nm. The Fe_3O_4(001) surfaces had dimensions of x = 5.0, y = 5.0 and z = 1.1 nm. At the start of the simulations, two of each type of surface were separated by a 3.0 nm in the z-direction. A fixed boundary condition is applied in the z-direction.

To replicate tribometer experiments under full-film EHL conditions,[96,279,314] we ensured that no direct contact between the sliding ferrous surfaces occurred during the simulations. In most of the NEMD simulations, 48 TCP molecules were randomly inserted between the surfaces, leading to a fluid film thickness of approximately 1 nm. Experiments have shown that the molecular area of TCP is around 1.1 nm^2 at the air-water interface.[315] The molecular area of TCP on ferrous surfaces is likely to be similar

as on the air-water interface, as shown for other surfactants. Thus, the chosen number of molecules is representative of the surface coverage expected for one TCP monolayer adsorbed on each of the two confining surfaces (24 molecules/$\sim$25 nm^2) While the main application of TCP is as an additive dissolved in a base oil for aviation lubricants,[287] it can also be used as a lubricant as a pure substance in both the liquid phase[294] and the vapour phase for high-temperature applications.[316] Due to the significant additional computational expense of simulating non-reactive base oil molecules, these were excluded from the majority of our simulations. Although some reactive NEMD simulation studies of lubricant additive mechanochemistry have simulated the base oil molecules,[317,318] the vast majority have not.[27,106,107,319,320] Previous NEMD simulations of sulfur-based additives with ReaxFF have shown that there were only minor differences when unreactive base oils were included in the simulations.[317] To assess the effect of including base oil molecules in our systems, we performed a subset of simulations that contained 48 TCP molecules randomly mixed with 48 diisodecyl adipate molecules (DIDA) [298] between α-Fe(110) surfaces. This is the same molar ratio used in a previous NEMD simulations of additive-base oil systems.[317] Although the concentration of TCP used in commercial formulations (1-5%)[321] and tribometer experiments ($\leq$ 10%)[322] is typically much lower than 50%, this represents a reasonable compromise between computational efficiency and closeness to real-life applications.

3.2.2 Simulation Details

All the simulations were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) open-source software.[323] The velocity Verlet[121] integration scheme was used with a timestep of 0.25 fs.[159] The simulations were conducted in three phases: energy minimisation, equilibration, and then simultaneous heating, compression, and shear. The energy minimisation phase was conducted using the conjugate gradient method. For all the systems and conditions investigated, equilibration was performed at an ambient temperature of 300 K for 0.1 ns. During the equilibration phase, the two ferrous surfaces were brought closer together by applying a low pressure of 10 MPa. The system was pressurised by applying a constant normal force (z-direction) equally distributed across the topmost layer of atoms (0.1 nm) in the upper surface, as shown in Figure 3.2a. The lowermost layer of atoms (0.1 nm) in the bottom surface was kept fixed in the z-direction. The equilibration at 10 MPa ensured that a realistic liquid density of TCP (1.2$\pm$0.1 g cm^{-3}) was obtained prior to the heating, compression, and shear phase. Following the equilibration phase, the compression and shear phase of the simulations were run for 1 ns. Here, a constant sliding velocity of 10 m/s was applied to the outer layer of atoms in the upper surface in the x-direction, as shown in Figure 3.2a. This is representative of the sliding velocities found in lubricated components inside gas turbine engines and tribometer experiments with TCP, which have been performed at sliding velocities as high as 10 m/s.[321] The pressure was increased to the target value of 1-5 GPa using the same method as during the equilibration. The temperatures were increased to 400-900 K using a Langevin thermostat with a damping parameter of 25 fs.[324] The thermostat acted only in the y-direction and on the central layer of atoms (0.2 nm) in the upper and lower surfaces, as shown in Figure 3.2a. This approach enabled a temperature gradient to develop in the centre of the TCP film and heat to be dissipated at the surfaces, as would occur experimentally.[325] The thermostat was sufficiently efficient to ensure that the average system temperature was within 10 K of

the target value at the chosen sliding velocity. The selected range of temperatures and pressures spans those expected in applications of TCP such as antiwear additives in gearboxes for gas turbine engines[290] and as vapour-phase lubricants.[316]

The temperature, pressure, and sliding velocity were all increased simultaneously to capture their combined effects on the reaction rate. Some previous NEMD simulations of mechanochemical reactions with ReaxFF have isolated the effects of temperature, pressure, and sliding by using separate phases of the simulations.[286,319,326] The number of additive molecules in the NEMD simulations is finite, so separating the simulations into different phases (thermalisation, pressurisation, and sliding) means that fewer molecules are available to react in any subsequent phases. Conversely, in experimental mechanochemistry studies,[96,279,314] additive-containing lubricants are entrained between the sliding surfaces, continuously replenishing the number of additive molecules. Hence, in our NEMD simulations, we elected to increase the temperature, pressure, and sliding velocity to their target values concurrently.[27]

Force Field Details and Validation

We used the bond order-based ReaxFF potential for all of the NEMD simulations. Similar systems consisting of phosphorus-containing additives confined between iron surfaces have previously been studied with NEMD simulations using first principles methods.[106,327] However, bond order potentials are several orders of magnitude cheaper than first principles methods,[104] allowing much larger systems to be simulated under experimentally-relevant conditions. ReaxFF was originally developed by van Duin et al.[155] to study hydrocarbon reactivity and has been parameterised to model a wide range of chemical systems and processes.[154] Of particular relevance to this study, ReaxFF-based NEMD simulations are now routinely used to study tribochemical reactions.[16] The C/H/O/Fe/P ReaxFF parameters used in this study were developed by Khajeh et al.[159]. The parameters for Fe/O/H were originally developed by Aryanpour et al. [328], Fe/C by Zou et al. [329], P/O/C/H by Verlackt et al. [330], and C/H/O by Chenoweth et al. [157]. The point charges on the atoms vary dynamically during the simulations and were calculated using the charge equilibration (QEq) method.[156] Chemical bonding information was output every 1.0 ps, using a bond order cutoff of 0.3 to identify covalent bonds.[159] The combined C/H/O/Fe/P ReaxFF parameters [159] have been successfully applied to study the thermal and mechanochemical decomposition of phosphate esters with different alkyl and aryl substituents on several ferrous surfaces.[27,301,304] The ReaxFF parameters have also been validated against first principles calculations for the non-dissociative adsorption energy, C-O bond dissociation energy, and P-O bond dissociation energy for trialkyl phosphates on α-Fe(110).[304] The ReaxFF parameters for Fe/O/H have been shown to accurately reproduce the experimental lattice constants for α-Fe (within 1 %),[329] Fe_3C (within 0.5 %),[329] and Fe_3O_4 (within 3 %).[328] The ReaxFF parameters have also been shown to perform favourably compared to other many-body force fields in describing the mechanical properties of α-Fe.[331] At 300 K, the experimental elastic modulus is 166 GPa for α-Fe, [332] 174 GPa for Fe_3C, [333] and 186 GPa for Fe_3O_4(100). [334] With the ReaxFF parameters used here, [159,328,329] the bulk elastic modulus at 300 K is 136 GPa for α-Fe, [335] 148 GPa for Fe_3C, [335] and 129 GPa for Fe_3O_4. [336] Thus, at 300 K, the experimental bulk elastic modulus is underpredicted by 18 % for α-Fe, 15 % for Fe_3C, and 31 % for Fe_3O_4. Since previous density functional theory (DFT) calculations have highlighted the importance of contact stiffness to mechanochemical pro-

cesses, [337] we also calculated the elastic properties of thin single slabs of the three materials used in the NEMD simulations. For this calculation, the lowermost layer of atoms in each slab of material was fixed in the z-direction, while a constant normal force was applied the topmost layer of atoms to compress the system to 2 GPa. The temperature was varied between 300-900 K using the same thermostat as applied in the NEMD simulations. The corresponding elastic constant (C_{33}) was calculated from the strain fluctuations at constant normal stress.[338] The elastic constant results for the thin slabs are shown in Table A.1 . For the α-Fe(110) slabs, the elastic constant C_{33} decreases from 206 GPa at 300 K to 196 GPa at 900 K. The elastic constant C_{33} also decreases with increasing temperature for the Fe_3C(010) (from 200 GPa to 181 GPa) and Fe_3O_4(001) (from 220 GPa to 196 GPa) slabs. The higher stiffness of thin solid slabs compared to bulk solids was also observed in previous MD simulations and was attributed to geometrical confinement.[339] The decreasing stiffness with increasing temperature is also consistent with previous experiments. [332]

3.3 Results and Discussion

3.3.1 Base Oil Effects on TCP Decomposition

The primary application of TCP is as an antiwear additive for aviation lubricants where it is dissolved in a base oil.[287] Thus, to ensure that not simulating the base oil molecules did not significantly affect the mechanochemical decomposition mechanism, we performed a subset of simulations containing 50% TCP and 50% DIDA molecules. Despite some bonding of DIDA molecules to the surfaces, negligible decomposition of the ester base oil occurs during the simulations (<2 molecules). Previous experiments have suggested that ester base oil decomposition may become significant when it is heated for several hours with a stainless steel coupon at temperatures above about 600 K.[322] However, this process is much slower than TCP decomposition, which occurred experimentally at much lower temperatures.[322] The mechanochemical decomposition mechanism of TCP is also unaffected by the presence of base oil molecules. This observation is consistent with previous NEMD simulations of, which showed that the addition of dodecane base oil molecules did not affect the mechanochemical reaction pathway of sulfur-based lubricant additives on ferrous surfaces.[317]

For a given normal stress, the shear stress is much lower for the 50% TCP, 50% DIDA system than the 100% TCP system, as shown in by Fig. A.1. This reduction in shear stress is expected to due the lubricating effect of the low-friction ester base oil.[340] From previous mechanochemistry experiments of ZDDP tribofilm growth,[96,279,341] it is anticipated that the lower shear stress will result in lower reactivity for the 50% TCP, 50% DIDA than the 100% TCP system.

Figure 3.3a shows the change in the number of intact TCP molecules for 100% TCP at 600 K and 1-5 GPa, while Figure 3.3b shows the same for the 50% TCP, 50% DIDA system. The TCP molecules are counted as dissociated, i.e., no longer intact, if any bond other than a C-H is broken. The number of TCP molecules decreases faster for the pure TCP system than the 50% TCP, 50% DIDA system. For both systems, there is an exponential decay of the number of intact TCP molecules with time, which is indicative of a first-order reaction.[342] For first-order reactions, the reaction rate, $r = k[C]$, where k is the rate constant

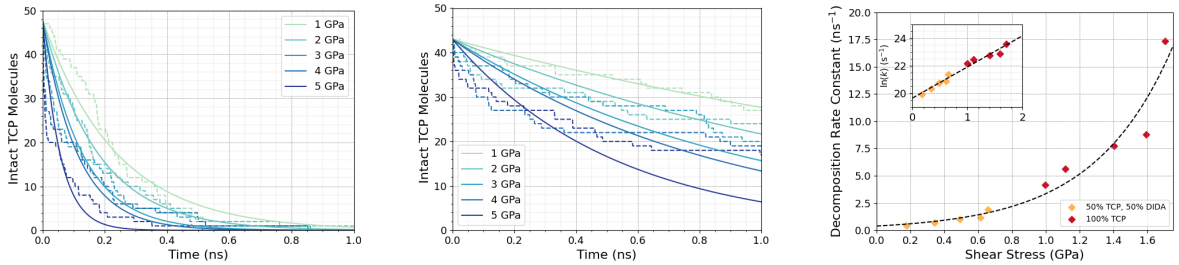


Figure 3.3: Change in the number of intact TCP molecules with time at 600 K and 1-5 GPa for a) 100% TCP system b) 50% TCP, 50% DIDA system. The dashed lines show the simulation data and the solid lines are exponential decay fits used to calculate the first-order rate constants, k. c) Change in k for TCP decomposition with shear stress, τ. The data with and without DIDA can be fit using Equation 1 with a single activation volume (13.6 ± 2.6 Å^3). Inset shows the same data versus ln(k).

and $[C]$ is the concentration of reactant. Previous NEMD simulations showed that the mechanochemical decomposition of trialkyl phosphate decomposition between α-Fe(110) surfaces also followed first-order kinetics.[27] In mechanochemistry experiments with ZDDP, the rate of tribofilm growth followed either zero-order or fractional-order kinetics. [95,96,314] The observation of zero-order kinetics indicates that the rate-determining step for tribofilm growth involves surface-adsorbed molecules, [96] which masks the underlying first-order kinetics. [343] In experiments, the surface-adsorbed additive molecules are continually replenished from the base oil solution and the bulk concentration does not significantly decrease.[96] On the other hand, there is a finite number of additive molecules in the NEMD simulations, which means that the decomposition rate inevitably decreases with time as fewer molecules become available to react.[27]

Figure 3.3c shows the change in the decomposition rate constants, k, with shear stress for the 100% TCP and 50% TCP, 50% DIDA systems. The k values were determined from the exponential decay fits shown in Figure 3.3a and Figure 3.3b and the shear stress values were taken from Fig. A.1. For all the systems and conditions shown in Figure 3.3, k increases exponentially with shear stress, suggesting that TCP decomposition is a stress-augmented thermally activated (SATA) processes. [88] The Bell model [90] can be used to describe the kinetics of SATA processes:

$$k = A \exp\left(-\frac{E_a - \tau \Delta V^*}{k_B T}\right) \tag{3.1}$$

where A is the pre-exponential factor, E_a is the activation energy, ΔV^* is the activation volume, and k_B is the Boltzmann constant. ΔV^* is sometimes interpreted as volume change that occurs when the transition state is formed from the reactants.[344] In our simulations, as well as previous experiments of ZDDP tribofilm growth, [95,96,314] ΔV^* is positive, indicating that the volume of the transition state is larger than the reactants, confirming that the rate-determining step involves bond breaking.[341] We use shear stress, τ, in Equation 3.1 rather than normal stress, σ. Previous experiments have shown that tribofilm growth is primarily controlled by shear stress and not the normal stress.[96,279,341] Moreover, from Figure 3.3c, it is clear that the higher shear stress in the 100% TCP system leads to larger k values than for the 50%

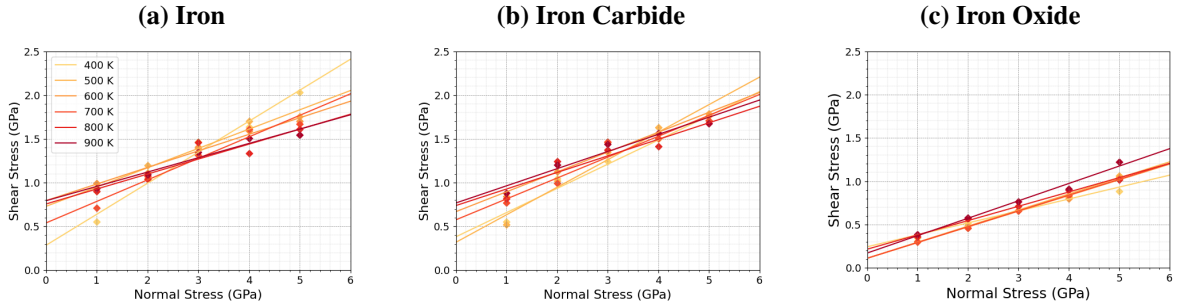


Figure 3.4: Variation in shear stress with changes in normal stress for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) surfaces.

TCP, 50% DIDA systems, even at identical normal stress Fig. A.1. The combined dataset rate constant *versus* shear stress data set (100% TCP and 50% TCP, 50% DIDA) can be described with a single fit to Equation 1 (dashed line in Figure 3.3c), yielding an activation volume of 13.6 ± 2.6 Å^3. This value is within the uncertainty range of the activation volume calculated for a 100% TCP system under the same conditions (10.7±6.2 Å^3). Thus, the reduction in k for the 50% TCP, 50% DIDA system compared to 100% TCP can be fully explained by the decrease in the shear stress, which is due to the lubricating effect of the low-friction DIDA molecules.[96,279,341] There do not seem to be any other interactions between the DIDA base oil molecules and the iron surface or TCP molecules that further directly reduce the reaction rate. Therefore, due to the much lower computational expense, we simulate pure TCP systems for the remainder of this study.

3.3.2 Surface Chemistry Effects on Shear Stress

We compared the change in the shear stress with normal stress for the iron, iron carbide, and iron oxide systems, as shown in Figure 3.4. The steady-state shear stress was block averaged over the final 0.5 ns of the NEMD simulations. At all the studied temperature and normal stress conditions, the shear stress is lower for the systems with TCP molecules sheared between iron oxide surfaces than iron and iron carbide surfaces. For all three surfaces, the shear stress increases linearly with normal stress with a finite intercept, as predicted by the extended Amontons-Coulomb law: $\tau = \mu\sigma + \tau_0$, where τ is the shear stress, μ is the friction coefficient, τ_0 is the Derjaguin offset, and σ is the normal stress.[345] The friction coefficient and Derjaguin offset for the different systems and temperatures are shown in Table A.2 and Table A.3 respectively). The friction coefficient, given by the gradient of the linear fits in Figure 3.4, is similar for all three surfaces. The Derjaguin offset, given by the intercept of the linear fit, is similar for iron and iron carbide surfaces, but is around 50 % lower for iron oxide. This is likely due to the higher surface energy and adhesion of TCP molecules on iron and iron carbide compared to iron oxide.[281] Previous DFT calculations suggested that the surface energy of α-Fe(110) (2.3 J m^{-2})[311] and Fe_3C(010) (2.2 J m^{-2})[312] is much larger than Fe_3O_4(001) (1.0 J m^{-2}).[346] For the iron and iron carbide systems, the friction coefficient decreases with increasing temperature (Table A.2); in contrast, the Derjaguin offset increases with increasing temperature (Table A.3). As the temperature increases, the amount of TCP decomposition will also increase,[304] which affects the temperature-dependence of the friction coefficient

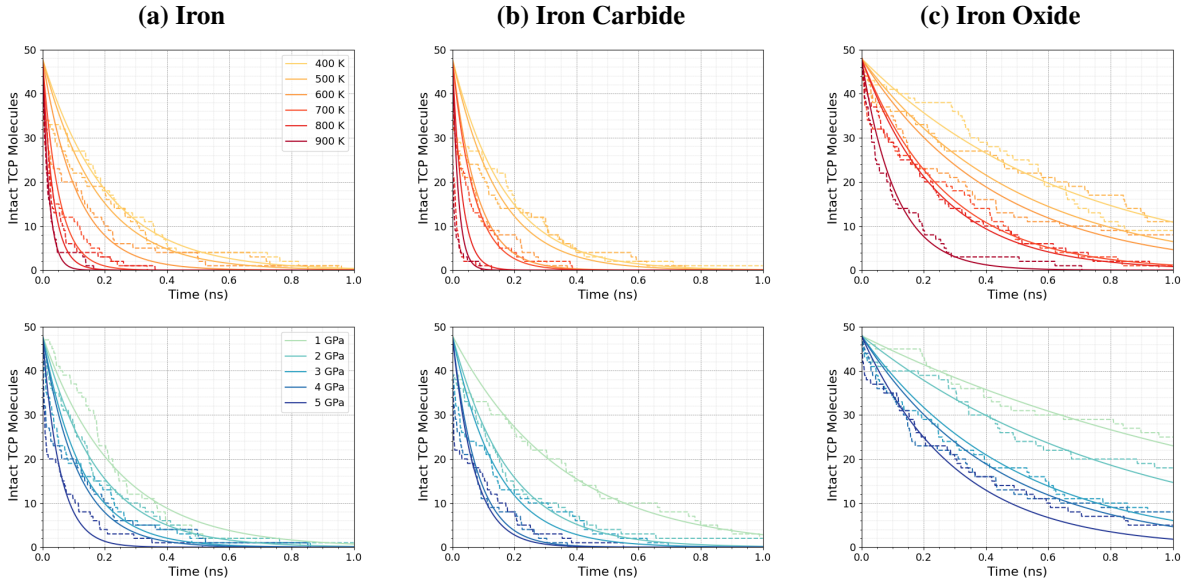


Figure 3.5: Number of intact TCP molecules with time at 4 GPa between 400-900 K (top row) and at 600 K between 1-5 GPa (bottom row) for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) surfaces. The dashed lines show the simulation data and the solid lines are exponential decay fits used to calculate the first-order rate constants, k.

and Derjaguin offset. For the iron oxide surfaces, the temperature has a negligible effect on the friction coefficient and Derjaguin offset, which is probably due to the much lower reactivity of TCP on this surface. [304] The addition of base oil molecules as well as TCP molecules between the iron surfaces leads to a large reduction in the Derjaguin offset (τ_0 = 0.79 GPa to τ_0 = 0.20 GPa), but a negligible change in the friction coefficient (μ = 0.18 toμ = 0.17), as shown in Fig. A.1. This observation is in agreement with previous NEMD simulations of iron surfaces lubricated by monolayers formed from polar lubricant additive molecules.[347] Consequently, if lubricant molecules were added to the iron carbide and iron carbide systems, the shear stress conditions would probably be similar for all the surfaces.

3.3.3 Surface Chemistry Effects on TCP Dissociation

Figure 3.5 shows the change in the number of intact TCP molecules with time for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces at constant normal stress (4 GPa) and varied temperature (400-900 K), and constant temperature (600 K) and varied normal stress (1-5 GPa). Results for other conditions are shown in Fig. A.3. The exponential decay curves are used to calculate the TCP decomposition rate constant, k, for each temperature and stress condition. Ayestarán Latorre et al.[27] investigated the mechanochemical decomposition of trialkyl phosphates between 300-500 K and 1-4 GPa for α-Fe(110) surfaces. Comparing the current results to this previous study, it is clear that triaryl phosphates such as TCP have much lower decomposition rate constants than trialkyl phosphates under the same conditions. This observation agrees with previous experimental[300,305] and simulation[304] studies of the thermal decomposition of phosphate esters on ferrous surfaces.

The TCP decomposition rate constants, k, vary significantly depending on the surface chemistry, with the fastest reactions being observed for the α-Fe(110) systems, followed by Fe_3C(010), and finally Fe_3O_4(100). This may be due to the differences in surface energies of the three ferrous surfaces, as discussed in the previous section.[311,312,346] Higher surface energies promote stronger molecular adsorption [348] and thus faster TCP decomposition. Indeed, DFT calculations have shown that adsorption is a crucial step the decomposition of TCP on iron surfaces.[349]

Previous high-temperature (750 K) thermal decomposition experiments suggested that TCP preferentially reacted with iron oxides that were in a higher oxidation state, i.e. Fe < FeO < Fe_3O_4 < Fe_2O_3.[296] This contrasts with first principles NEMD simulations of other lubricant additives, such as molybdenum dithiocarbamates, where native iron surfaces were found to be much more reactive than oxidised surfaces.[320] We observe much higher mechanochemical reactivity on the iron surface compared to the iron oxide surface, which is in agreement with the simulations,[320] but not the experiments.[296] This is probably because the iron used surfaces in the experiments[296] were also covered by an oxide layer rather than the native iron surfaces used in the simulations.[301]

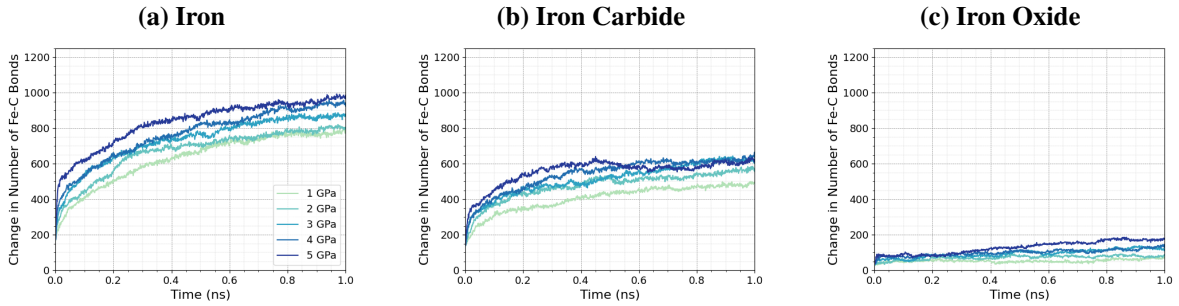


Figure 3.6: Change in the number of Fe-C bonds with simulation time at 600 K between 1-5 GPa for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(100).

No direct Fe-P bonds formed between the surfaces and TCP molecules in any of the NEMD simulations, but a large number of new Fe-O bonds form. This observation indicates that iron phosphate, rather than iron phosphide, is formed on the ferrous surfaces, which is consistent with observations in previous experimental studies.[292,293,294] Previous MD simulations investigating the thermal decomposition of TCP on hydroxylated amorphous Fe_3O_4 have shown that the initial step involved Fe-C bonding.[159] Indeed, DFT investigations have shown that benzene chemisorbs on iron oxide surfaces due to strong interactions between iron cations and the aromatic π-electrons.[350] To investigate this further, we quantified the change in the number of Fe-C bonds formed at different pressures and temperatures. The Fe-C bonds present in the Fe_3C(010) surfaces before exposure to TCP are not included in the analysis. Figure 3.6 shows the variation in the number of Fe-C bonds with time for the simulated TCP systems at 600 K between 1-5 GPa for α-Fe(110), Fe_3C(010), and Fe_3O_4(100). The number of Fe-C bonds at a given time increases as Fe_3O_4(100) < Fe_3C(010) < α-Fe(110), which is the same order observed for the rate constants of TCP decomposition on the three surfaces. We also see a clear trend of an increase in the number of Fe-C bonds, and therefore chemisorption of TCP, with an increase in pressure for all three surfaces. The increased bonding of the TCP carbon atoms on α-Fe(110) compared to Fe_3C(010) and

$Fe_3O_4(100)$ can be visualised in the number density profiles, as shown by Fig. A.2. The carbon peaks from the TCP molecules show increased overlap with the iron peaks for α-Fe(110) compared to the other surfaces.

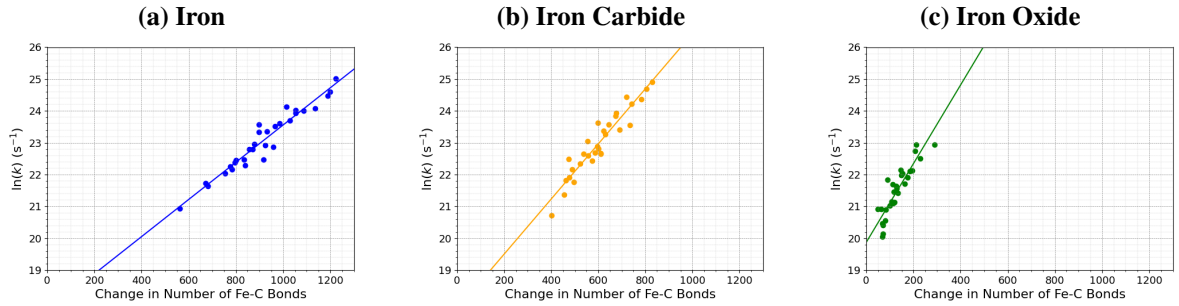


Figure 3.7: Variation in ln(k) for TCP decomposition with the change in the number of Fe-C bonds from 0-1 ns for a) α-Fe(110), b) $Fe_3C(010)$, and c) $Fe_3O_4(100)$.

Figure 3.7 shows the relationship between the number of Fe-C bonds formed and the TCP decomposition rate constant. There is a logarithmic increase in the TCP decomposition rate constant with the number of Fe-C bonds formed for all three surfaces. This implies that increased chemisorption of the molecules leads to faster dissociation.[281] Indeed, previous experiments have shown that ZDDP forms thicker tribofilms on surfaces to which it more strongly adsorbs.[282]

3.3.4 TCP Decomposition Mechanisms

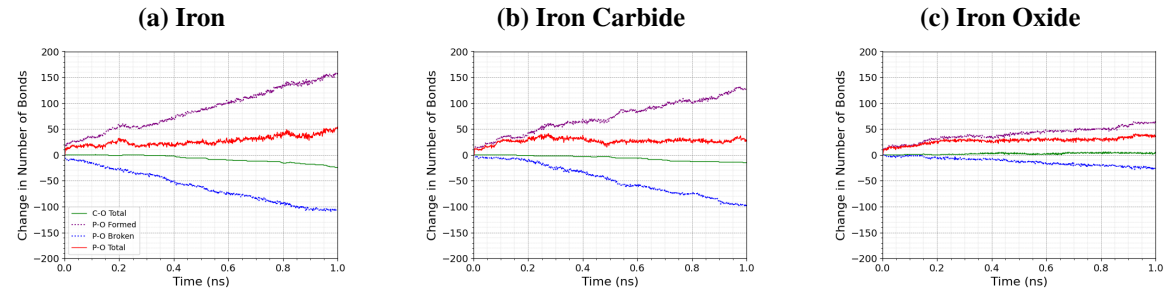


Figure 3.8: Change in the number of P-O and C-O bonds with sliding time at 500 K and 3 GPa for the a) α-Fe(110), b) $Fe_3C(010)$, and c) $Fe_3O_4(001)$ systems. Solid lines show the change in the total number of bonds compared to TCP, dashed lines show bond formation and dotted lines show bond cleavage.

Figure 3.8 shows the change in the total number of C-O and P-O bonds with sliding time for an intermediate temperature (500 K) and pressure (3 GPa). For the $Fe_3O_4(001)$ surfaces, bonds formed with surface oxygen atoms are included in the analysis. The red lines represent the total change in the number of bonds P-O bonds compared to the initial TCP molecules, the blue lines show the contribution from P-O bonds in TCP that are broken, and purple lines show new bonds that are formed. A negligible number of new C-O bonds form in the simulations and the total change is shown by the solid green lines. For all the surfaces, P-O cleavage precedes C-O cleavage and, by the end of the simulations, approximately five times as many P-O bonds were broken than C-O bonds, suggesting that P-O cleavage is the dominant

decomposition mechanism. These observation are consistent with previous MD simulations[301,304] and experiments[300] of the thermal decomposition of TCP. Conversely, in previous NEMD simulations of trialkyl phosphates confined between α-Fe(110) surfaces, the C-O bonds broke first and approximately twice as many C-O bonds were broken than P-O bonds.[27] The decomposition mechanism is therefore affected by the much stronger C-O bonds in triaryl phosphates compared to those in trialkyl phosphates. Moreover, unlike triaryl phosphates, some alkyl intermediates formed by C-O cleavage in trialkyl phosphates, such as butyl chains, can readily undergo β-hydride elimination reactions on ferrous surfaces to produce gaseous olefins.[300] In Figure 3.8, the number of C-O and P-O bonds broken for the α-Fe(110) and Fe_3C(010) surfaces is similar, but fewer bonds are broken for the Fe_3O_4(001) surfaces.

More P-O bonds are formed than are broken in the TCP molecules, leading to an overall increase in the number of P-O bonds, as shown by the solid red lines in Figure 3.8. This is due to the formation of P-O-P linkages between the phosphorus atoms, resulting in polyphosphate chain growth.[296,297] This process occurs mainly through nucleophilic substitution reactions, whereby P-O-R bonds in TCP (where R is a cresyl group) are replaced by P-O-P bonds.[27]

The total increase in the number of P-O bonds at the end of the 1 ns simulation is similar (40-50 bonds) for the α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces. This is because both the P-O cleavage (blue dotted lines) and P-O formation (purple dotted lines) reactions are accelerated by the more reactive surfaces (α-Fe(110)), as shown by the steeper lines in Figure 3.8. The total number of P-O bonds at the end of the simulation is also not sensitive to the temperature and pressure, as shown by Fig. A.4.[27]

For the Fe_3O_4(001) surface, nucleophilic attack from the surface oxygen atoms could increase the rate of P-O dissociation in the TCP molecules.[304] The variation in the number of P-O_{surf} bonds formed between phosphorus atoms in the TCP molecules and oxygen atoms in the Fe_3O_4(001) surface is shown Fig. A.5. At low temperature and pressure, the number of P-O_{surf} bonds formed is much lower than the total number of P-O bonds formed (Figure 3.8), suggesting that this process plays a negligible role in the nucleophilic substitution reactions. However, at high temperature and pressure, the number of P-O_{surf} bonds increases, such that it they represent a significant proportion (>10 %) of the total P-O bonds formed. Overall, surface oxygen atoms do not seem to promote the decomposition of TCP on ferrous surfaces, as was suggested from previous experiments.[296]

3.3.5 Surface Chemistry Effects on Polyphosphate Growth

Figure 3.9 shows the number of phosphate-containing oligomers, toluene, and cresol molecules formed from TCP at 500 K and 3 GPa on the three surfaces. On iron and iron carbide, the primary product of TCP decomposition is cresol, formed directly from P-O bond cleavage. Some toluene is also formed on these surfaces via C-O cleavage, but this occurs only after the formation of cresol. The number of oligomers formed is similar for all three surfaces. On iron oxide, phosphate-containing oligomers are the major product rather than cresol, due to the reduced rate of P-O cleavage reactions (Figure 3.8).

The length of the polyphosphate chains in tribofilms is important to their antiwear performance. [351] Figure 3.10 shows the change in the number of polyphosphate chains (dimers, trimers, tetramers, and $\geq$

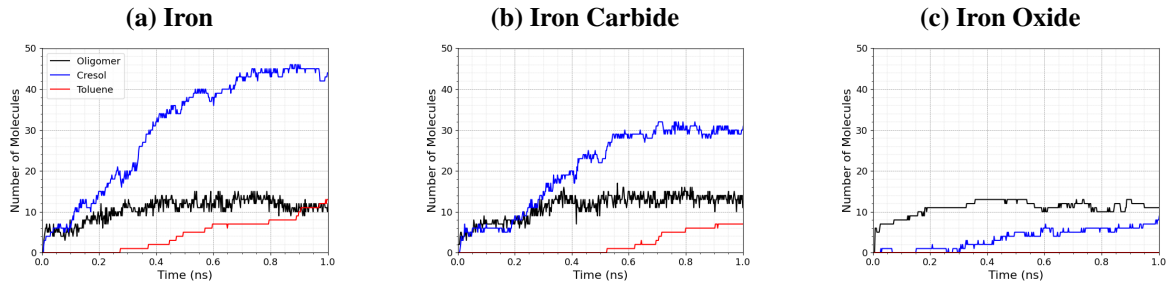


Figure 3.9: Comparison of oligomer, toluene, and cresol quantities at 500 K and 3 GPa for the a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) systems.

pentamers) with sliding time for the three studied systems. In all cases, the total number of polyphosphate fragments increases at short times before asymptoting towards a steady-state value of $\sim$10. For Fe_3C(010) and Fe_3O_4(001), the majority of TCP molecules form dimers, but longer chains, up to octomers, form in some cases. In general, longer polyphosphate chains are formed on the α-Fe(110) surfaces than Fe_3C(010) and particularly Fe_3O_4(001) surfaces. For α-Fe(110), a similar number of dimers and trimers form, with some of the chains growing up to octamers. The relatively short polyphosphate chains formed in our NEMD simulations are consistent with the average chain length of approximately 2.5 monomers found in tribometer experiments with ZDDP. [351] Much longer chains (up to 40 linked monomers) were formed in previous NEMD simulations with ReaxFF of phosphoric acid confined between sliding silica surfaces at very high temperature (1400 K).[158] The longer chains formed in these previous simulations could be due to the higher reactivity of phosphoric acid compared to TCP and the lack of iron atoms to catalyse the depolymerisation (P-O dissociation) reactions.

The difference in the polyphosphate chain length between the three surfaces is much more subtle than the change in TCP decomposition rates. This is probably because iron atoms or ions, which are most readily available in α-Fe(110), accelerate the depolymerisation of the polyphosphate chains by P-O bond cleavage. [351] Similarly, the polyphosphate chain length does not vary significantly over the wide range of temperatures and pressures studied in the NEMD simulations, as shown by Fig. A.6. This is because both the polymerisation (P-O formation) and depolymerisation (P-O dissociation) reactions are accelerated by a similar amount due to increased temperature and shear stress, as shown by Fig. A.4.

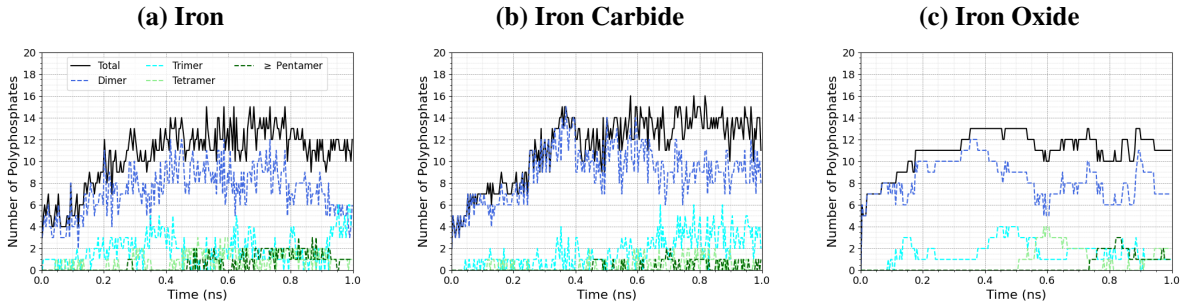


Figure 3.10: Change in the number of polyphosphate chains with sliding time at 500 K and 3 GPa for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) systems.

3.3.6 Rate Analysis

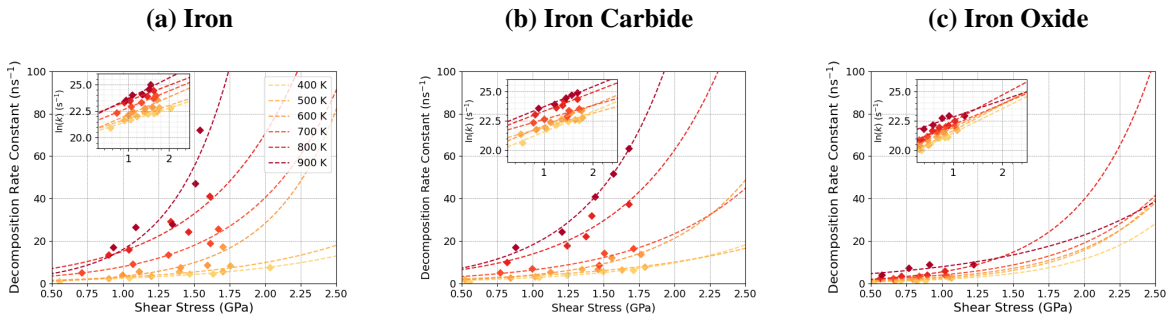


Figure 3.11: Shear stress dependence of the TCP decomposition rate at different temperatures for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(100) systems. Insets show the same data with a logarithmic y-axis.

Figure 3.11 shows the dependence of the TCP decomposition rate constant on shear stress for iron, iron carbide, and iron oxide surfaces. The Bell model [90] (Equation 1) accurately describes the rate of TCP decomposition with temperature and shear stress on the three surfaces, as shown by Fig. A.7. The activation volume, ΔV^*, is calculated from the 2D fits of the rate versus shear stress data to Equation 1 shown in the insets of Figure 3.11. With this 2D fitting method, ΔV^* is found to increase monotonically with temperature from 4.9 ± 1.7 Å^3 at 400 K to 17.5 ± 8.4 Å^3 at 900 K for TCP between the α-Fe(110) surfaces. In all cases, the activation volume is positive, which suggests that the rate determining step involves bond cleavage.[344] This observation is consistent with previous experiments of tribofilm growth from antiwear additives.[341] For the Fe_3C(010) surfaces, ΔV^* also increases with temperature from 6.6 ± 3.7 Å^3 400 K to 15.3 ± 2.4 Å^3 at 900 K, while for the Fe_3O_4(100) surfaces, ΔV^* increases from 8.3 ± 4.5 Å^3 at 400 K to 12.8 ± 6.6 Å^3 at 900 K. The calculated ΔV^* values are all consistent with the dimensions of single bonds. The molecular volume of TCP is approximately 500 Å^3, meaning that the activation volumes for TCP on the three surfaces range between 1 and 4 % of the molecular volume. This percentage range is is consistent with previous NEMD simulations using trialkyl phosphates (4 %), [27] while experiments using ZDDPs with different alkyl substituent showed somewhat larger percentages (7-17 %). [314]

Based on the 2D fits to Equation 1 shown in Figure 3.11, the activation volume is highly temperature-dependent, but there are no clear differences between the three surfaces due to the relatively large uncertainties. Previous experimental studies of mechanochemical tribofilm growth have treated the activation volume as a temperature-independent reaction constant; however, these studies considered a much smaller temperature range than the present study.[95,96,279,314,341] In our recent NEMD study of trialkyl phophate mechanochemical decomposition, we also observed that the the activation volume increased with increasing temperature.[27] This was partially explained through a reduction in the elastic modulus of the α-Fe(110) slabs.[27] A recent DFT study suggested that the contribution of the elastic deformation of the surrounding bulk to the activation volume of mechanochemical reactions can be significant and, in some cases, dominant.[337] The elastic modulus of all three surfaces decreases with increasing temperature (Table A.1). Since the activation volume is inversely proportional to the contact stiffness,[337] it is

expected that higher temperatures will lead to lower activation energies, as observed in Figure 3.11 and Table A.4. However, the increase in the activation volume (up to 350 %) is much larger than the decrease in the stiffness of the surfaces (up to 10 %) between 400-900 K. Thus, other factors, such as a decrease in stiffness of the confined TCP molecules and its decomposition products, must also contribute to the increased activation volume at higher temperature.

The activation energy, E_a, is calculated from the slope of the 2D fits of Equation 1 to the rate versus inverse temperature data illustrated by Fig. A.8. For TCP decomposition between the α-Fe(110) surfaces, E_a increases with pressure from 20.6 ± 5.5 kJ/mol at 1 GPa to 24.0 ± 6.5 kJ/mol at 5 GPa. For the Fe_3C(010) surfaces, E_a also increases from 19.7 ± 6.1 kJ/mol at 1 GPa to 24.3 ± 7.5 kJ/mol at 5 GPa. For the Fe_3O_4(100) surfaces, E_a again increases from 10.2 ± 7.2 kJ/mol at 1 GPa to 19.5 ± 3.63 kJ/mol at 5 GPa. A previous experimental study suggested that the thermal activation energy for TCP decomposition on stainless steel is approximately 60 kJ/mol.[305] Thus, the activation energy values are considerably underpredicted in our simulations compared to experiments,[305] which is probably due to the chromium-rich oxide layer formed on the surface of stainless steel in experiments.[352] Chromium oxide is known to be much less reactive towards TCP than iron oxide,[297,298,353] as was used in our NEMD simulations. Thus, the chromium oxide surface formed on stainless steel is expected to reduce the activation energy by a smaller amount than the iron oxides formed on conventional bearing steel.[75,76] From in-situ tribometer experiments, the activation energy for TCP decomposition on an iron substrate was estimated as 6.3 kJ/mol,[354] which is somewhat lower than our simulation results. This can be explained through the fact that the effect of shear stress on reducing the activation energy, as shown in Equation 1, was not considered in their calculations. As a result, the activation energy they calculated will be under-predicted.

The pre-exponential factor, A, is calculated from the intercept of the fits of Equation 1 to the rate versus inverse temperature data shown by Fig. A.8. This is often expressed as $A = Z\rho$, where Z is the frequency or collision factor, and ρ is the steric factor. The steric factor represents the fraction of successful collisions and is usually less than unity.[91] For TCP molecules between the α-Fe(110) surfaces, ln(A) increases with pressure from 25.4 ± 3.2 at 1 GPa and to 28.1 ± 3.86 at 5 GPa. For the Fe_3C(010) surfaces, ln(A) increases from 25.4 ± 3.6 at 1 GPa to 27.8 ± 4.7 at 5 GPa, while it increases from 22.6 ± 4.5 to 25.6 ± 2.1 for the Fe_3O_4(100) surfaces. The low-pressure values are in very good agreement with previous estimates of the pre-exponential factor for the thermal decomposition of TCP on stainless steel at atmospheric pressure, ln(A) = 22.7.[305]

Figure 3.12 shows the variation of ln(A) with E_a from the 2D fits to the rate versus temperature data. The concurrent increase in E_a with ln(A), which is known as the kinetic compensation effect, is often observed in heterogeneously catalysed processes.[343] Previous studies using ReaxFF to study the mechanochemical behaviours of trialkyl phosphates have also observed such an effect.[27] The presence of a compensation effect suggests that the changes in these parameters may have a mathematical rather than physical origin and does not necessarily indicate any changes to the reaction pathway.[355]

Experimental pre-exponential factors for surface reactions typically vary from 10^{10}-10^{16} s^{-1}. [356] Some

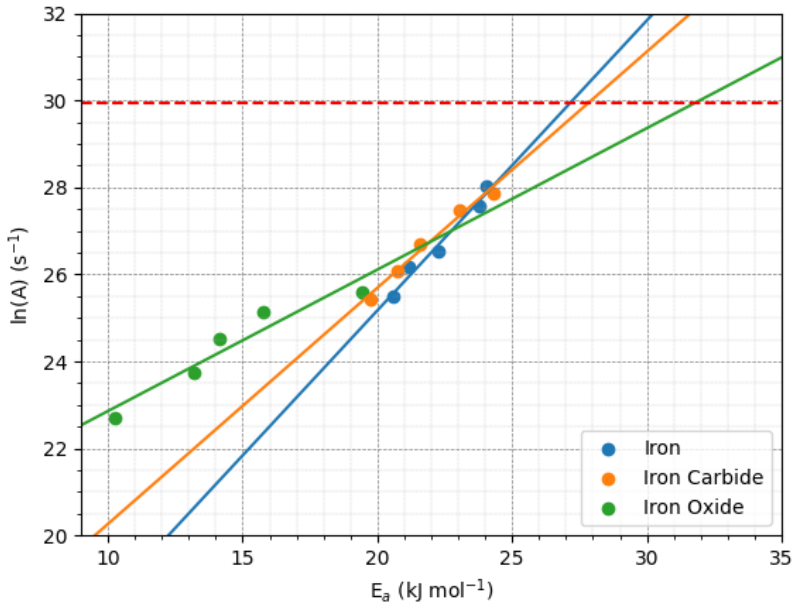


Figure 3.12: Kinetic compensation effect between the activation energy and pre-exponential factor for Iron, Iron Carbide, and Iron Oxide

previous mechanochemistry studies [95] have used a constant pre-exponential factor of 10^{13} s^{-1} to determine the activation energy. If a constant pre-exponential factor of 10^{13} s^{-1} ($\ln(A) = 30$) is used for all three surfaces (horizontal red dashed line in Figure 3.12, the activation energies of TCP decomposition for the three surfaces from the 2D fits increase in the order α-Fe(110) $\sim$ Fe_3C(010) < Fe_3O_4(001). Thus, the more reactive surfaces towards TCP give lower activation energies, which is consistent with previous experiments of α-pinene tribopolymerisation.[284]

3.3.7 3D Shear Stress-Temperature-Rate Constant Fits

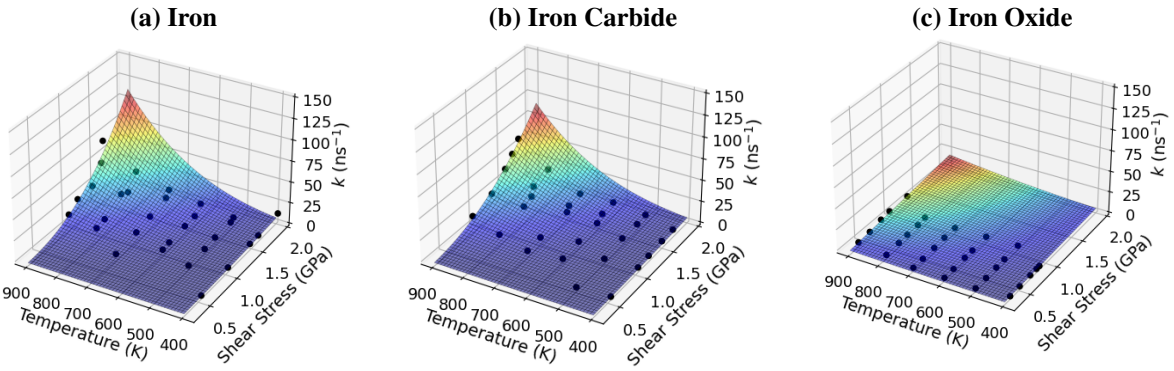


Figure 3.13: 3D fits of TCP decomposition rate as a function of temperature and shear stress for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) surfaces. The surfaces were fitted with the nonlinear least squares method.

Surface	E_a (kJ/mol)	ln(A) (s^{-1})	ΔV^* (Å^3)
α-Fe(110)	21.3±2.8	27.0±2.0	12.3±3.3
Fe_3C(010)	22.1±2.6	26.2±1.7	10.2±2.6
Fe_3O_4(100)	11.2±1.7	24.3±1.8	8.1±4.4

Table 3.1: Calculated values of the activation energy, E_a, natural logarithm of the pre-exponential factor, ln(A), and activation volume, ΔV^*, for TCP between α-Fe(110), Fe_3C(010) and Fe_3O_4(100) surfaces from the 3D fits to Equation 1. The parameter ranges represent the 95 % confidence intervals from the 3D fits.

To investigate the combined influence of temperature and stress on reactivity, the decomposition rate data was also plotted as a 3D surface and fitted to Equation 1. This approach provided significantly reduced uncertainties in the activation volume, activation energy, and pre-exponential factor compared to 2D fits in our previous NEMD study of trialkyl phosphate mechanochemistry.[27] The 3D fits for α-Fe(110), Fe_3C(010) and Fe_3O_4(100) are presented in Figure 3.13a), b), and c), respectively. To avoid biasing the 3D fits towards the higher decomposition rate constants, fits ln(k), rather than k were employed, as shown by Fig. A.12.[27] Using the 3D surfaces, we calculated ΔV^*, E_a, and A values across the entire range of temperatures and pressures studied, as shown in Table 3.1. Unlike the 2D plots, the 3D fitting method does not capture the changes in the Bell model parameters with temperature and pressure. However, compared to the 2D fits, the 3D fits led to reduced uncertainties in all three parameters in Equation 1,[27] as shown by Fig. A.9, Fig. A.10, and Fig. A.11). The 3D fits suggest that the activation energy increases in the order: Fe_3O_4(100) < Fe_3C(010) $\sim$ α-Fe(110), which is opposite to that expected from the relative reactivity of the surfaces. The smaller energy barrier for the less reactive oxide surfaces may be due to the availability of oxygen atoms to promote P-O dissociation through nucleophilic substitution reactions at the phosphorus atoms. On the other hand, the pre-exponential factor is also much smaller for the iron oxide surface, suggesting that the fraction of successful collisions is also much lower. This may be due to the decreased availability of iron atoms at the surface, which are essential to promoting the P-O dissociation reactions. The ΔV^* values increase in the order: Fe_3O_4(100) < Fe_3C(010) < α-Fe(110), suggesting that the more reactive surfaces lead to an increased shear stress dependence of the reactivity. The larger activation volume for α-Fe(110) than Fe_3C(010) and particularly Fe_3O_4(100) could be because the more reactive surfaces anchor the TCP molecules, leading to more efficient transmission of shear stress applied through the surfaces to strain within the molecules. Similar suggestions have been made from previous NEMD simulations in other systems, which have shown accelerated mechanochemical reaction when molecules are anchored to one [286] or both[107] sliding surfaces. Our observation of higher ΔV^* values on more reactive surfaces contrasts with previous experiments α-pinene tribopolymerisation, which showed that surfaces that promoted α-pinene chemisorption (e.g. copper oxide) gave smaller activation volumes than the ones only capable of physisorption (e.g. silicon oxide).[284] This discrepancy may be because the experiments studied bond-forming mechanochemical reactions, rather than bond-breaking as in the current simulations. The reaction yield was higher on the more reactive surfaces, implying that the accelerated reactivity for the more reactive surfaces seen in the experiments was mostly due to a reduction in activation energy (although this was not quantified).[284] Another factor contributing to the difference in activation volumes between the three surfaces could be the contact stiffness.[27,337] In our

NEMD simulations, at a given temperature, the stiffness of the three surfaces is quite similar (within 10 %), as shown by.Table A.1 At most temperatures, the contact stiffness increases in the order $Fe_3C(010)$ < α-Fe(110) < $Fe_3O_4(100)$. Since the activation volume is expected to be inversely proportional to the contact stiffness,[337] the increased stiffness of the $Fe_3O_4(100)$ surface may also contribute to its smaller activation volume compared to the other two surfaces. However, we expect that the differences in contact stiffness play a secondary role compared to the molecular adsorption strength in determining the variation in activation volume between the three surfaces.

3.4 Conclusions

We have studied the mechanochemical decomposition of TCP molecules confined between atomically-smooth iron, iron carbide, and iron oxide surfaces using NEMD simulations with ReaxFF. The iron surfaces are more effective in promoting TCP decomposition compared to iron carbide and particularly iron oxide surfaces, which correlates with the increase in Fe-C bonding between the molecules and the surface, as well as the surface energy. Once the molecules are adsorbed, decomposition of TCP proceeds through P-O bond dissociation for all of the surfaces to form cresol molecules. In some cases, a subsequent C-O bond dissociation within the cresol molecule leads to the formation of toluene. The rate of TCP decomposition increases exponentially with temperature and shear stress, indicating that this is a SATA process.

We show that the presence of DIDA base oil molecules does not affect the decomposition mechanism, but reduces the shear stress and thus the TCP decomposition rate. The change in the TCP decomposition rate constants with shear stress for the systems with and without DIDA can be described with a single fit to the Bell model with the same activation volume. This observation justifies the use of base oil-free NEMD simulations for the subsequent wide parameter study. As well as TCP decomposition, we also investigate the early stages of polyphosphate growth. Despite the fact that P-O bond formation is accelerated on the more reactive surfaces (iron), the rate of P-O bond cleavage is also increased by a similar amount, meaning that the total number of P-O bonds is comparable between the three surfaces.

The polyphosphate chain length is also similar between the surfaces, with a skew in the distribution to higher chain lengths (up to octomers) for the more reactive iron surfaces. 3D fits to the Bell model lead to activation energies and pre-exponential factors both increasing in the order: $Fe_3O_4(100)$ < $Fe_3C(010)$ $\sim$ α-Fe(110). This implies that, although the energy barrier is smaller for the less reactive oxide surfaces, the fraction of successful collisions is also much lower. This trend may be due to the decreased availability of iron atoms on the surface, which are essential to promoting the decomposition reactions.

The activation energies and pre-exponential factors calculated from the NEMD simulations are within the experimental range reported for iron and stainless steel (chromium oxide) surfaces. The activation volume from the 3D fits increases in the order: $Fe_3O_4(100)$ < $Fe_3C(010)$ < α-Fe(110), which suggests that higher surface energy causes the TCP molecules to be anchored to the surface, leading to more efficient transmission of shear stress mechanical energy. An additional contribution to the larger activation energy for α-Fe(110) surfaces could be the lower contact stiffness. We have shown that surface chem-

istry plays a central role in the mechanochemical decomposition of an industrially-important additive for aviation lubricants. The differences between the three ferrous surfaces can be rationalised through parameters from the Bell model.

The findings presented in this chapter advance the core research aim by elucidating atomic-scale mechanisms of additive decomposition under realistic tribological conditions. These insights lay the groundwork for predictive modelling of additive behaviour and offer guidance for the rational design of next-generation lubricant formulations.

Chapter 4

Effect of Sliding Velocity on the Decomposition of Tricresyl Phosphate

In this chapter, we use nonequilibrium molecular dynamics (NEMD) simulations with ReaxFF to investigate the mechanochemical decomposition of tricresyl phosphate (TCP) under nanoscale confinement between ferrous surfaces and diamond-like-carbon (DLC) surfaces, expanding on the studies of Chapter 3 by further isolating the effects of sliding speed and shear stress on TCP decomposition. Sliding speed and temperature are systematically varied to examine how molecular mechanisms of additive breakdown and tribofilm formation are influenced by sliding speeds and shear conditions. This study addresses the research objective of connecting nanoscale mechanochemical behaviour to the macroscopic performance of lubricant additives in real-world tribological applications.

4.1 Introduction

The commercial aviation industry was responsible for approximately 1.5% of global CO_2 emissions, or 425 Mt CO_2e in 2022.[357] The aviation industry is also responsible responsible for considerable amounts of non-CO_2 warming.[358] For example, jets that release condensates/vapour trails at high altitudes contribute to global warming, while the release of NO_x in the upper atmosphere results in greater amounts of ozone, resulting in extra warming effects.[359] As a result, it has been estimated that the aviation sector is actually responsible for 4.9% of anthropogenic warming.[360] As long as aviation is powered by combustion, these effects will continue to contribute, although some studies have pointed towards the use of synthetic fuels as a means to at least reduce condensate trails.[361] Hence, the development of strategies to mitigate emissions from the industry remains of significant commercial interest.[362] Decarbonising aviation poses a uniquely difficult challenge as aircrafts require high energy density for long-range travel, a requirement which current battery technology fails to satisfy adequately.[363] Weight reduction and tribological performance enhancement, leading to improved fuel efficiency energy loss reduction, are alternative approaches that can be used to reduce emissions in this sector.

Extensive research and development in the aerospace field has focused on the weight reduction of

aerospace components such as airframes, landing gear, turbine blades, bearings and gears.[364] For example, 60NiTi alloy, with its high hardness, compressive strength, low density and lower elastic modulus compared to standard materials, has seen its deployment in hydraulic rods, bearing, and connecting spline systems.[365,366,367,368,369] However, the versatility of the 60NiTi alloy possesses poor wear resistance properties in dry and boundary lubrication conditions, limiting its suitability for many of the extreme conditions present in aviation.[370,371] Applying coatings to the alloy has been shown to improve alloy performance in this regards and this approach has been widely used in the aerospace industry for improving other properties such as corrosion protection, oxidation resistance and water-proofing.[372,373] To this end, polymer-based coatings, such as deposited polydopamine (PDA) and polytetrafluoroethylene (PTFE), have been investigated as solutions. Miller et al.[374] found that in dry/boundary lubrication conditions, a 1.5 μm thick PDA/PTFE coating yielded an 85% reduction the coefficient of friction between 60NiTi alloy and Si_3N_4. While the COF of the 60NiTi alloy can be effectively reduced by the polymer-based coatings, the adhesion strength of such polymer coatings are poor, meaning that the coating can easily be worn away with shear stress.[375] Choudhury et al.[376] found that adding 0.25 wt% of graphite particles to PDA/PTFE coatings improved the adhesion strength between the coating and 60NiTi alloy, which increased the coating durability.

An effective alternative for such alloys is the Diamond-Like Carbon (DLC) coating, which has also has high hardness, excellent chemical inertness, and excellent tribological performances.[377,378,379,380] Specifically, hydrogenated-DLC has been investigated by multiple studies as it possesses excellent COF properties, coupled with favourable wear resistance in both vacuum and atmospheric environments.[84,381] Sharma et al. reported that DLC coatings exhibit low COF values in both dry and boundary lubrication regimes.[382] Erdemir et al. conducted friction and wear tests using steel ball–steel disc and sapphire ball–sapphire disc pairs in inert gas conditions (dry nitrogen and argon), finding that higher hydrogen content of the inert gas environment resulted in lower COF, regardless of the substrate.[383] Lugo et al. also found that the COF of H-DLC coatings in vacuum decreases with an increase in hydrogen content, suggesting that the hydrogen atoms in H-DLC coatings can passivate unbound σ-bonds on the contact surface of the coating, thereby preventing the formation of covalent bonds between opposing friction interfaces.[384,385,386] Dalibón, Wang, Amanov demonstrated that depositing DLC coatings on AISI 316L stainless steel, NiTi alloy, and Ti-6Al-4V alloy, respectively, significantly improves their friction and wear performance.[387,388,389] A greater understanding of the mechanisms that provide DLC coatings with favourable properties will help to design aviation tribological contacts with improved performance. Molecular dynamics simulations have already been employed with this in mind. For example, Wu et al.[390,391] investigated the tribological performance of nanocomposite DLC coatings (with WS_2 and WO_3 nanoparticles; WCS) using 2D MD simulations with the Lennard-Jones potential. The simulations suggested that the formation of surface tribolayers, dominated by graphitic carbon, tungsten trioxide and tungsten disulphide contribute significantly to the low friction properties of these coatings.[390,391] Lan et al. investigated amorphous DLC- SiDLC contacts with a Tersoff potential, finding that a transfer film forms between these surfaces during rubbing where increased silicon content can yield modest reduction in friction force.[392]

Lubricating oils used in aviation are subjected to relatively extreme operating conditions such as high temperatures (up to 550°C in wheel bearings),[393,394] high pressures (exceeding 3GPa),[395] and substantial radiation exposure.[396] Performance of traditional lubricant base oils and additives can be diminished in such conditions, increasing wear and friction and ultimately leading to premature system failure.[396] As discussed in Section 2.2.3 and Section 3.1, anti-wear additives can be used to prevent this. Considering the known properties of anti-wear additives and DLC coatings as discussed above, a natural question then arises: would the combination of the two yield even better tribological performance? TCP and its derivatives have been ubiquitous in aviation lubricants[287] since its introduction in the 1940s.[288] Over the last decade, several studies have highlighted the central role of mechanochemistry in the growth of tribofilms formed from the anti-wear additive zinc dialkyldithiophosphate (ZDDP).[95,96,279] TCP plays a similar role to ZDDP in automotive lubricants,[10] forming a protective tribofilm that prevents metal-metal contact to reduce wear, ensuring safety and reliability.[287] The reader is referred to Chapter 3 for a deeper discussion of the mechanisms that govern the utility of TCP as an anti-wear additive.

In most jet aircraft, filtered recycled cabin air is mixed with heated and compressed bleed air extracted from the jet engine to supply the cabin air.[397] Thus, contamination can occur if turbine oil, hydraulic fluid, or other substances leak into the bleed air system.[398] Furthermore, aircraft hydraulic reservoir vents are sometimes connected to cabin air ventilation systems, potentially allowing for hydraulic oil aerosols to enter cabin air.[399] Turbine and hydraulic oils normally comprise a mineral or synthetic base oil containing organophosphate additives such as TCP or butyl phosphates, which are included to enhance anti-wear, anti-corrosion, flame retardant, and high-pressure properties.[399] Passengers and crew are exposed to measurable levels of triaryl phosphates such as TCP approximately 23% of monitored flights.[400] Furthermore, elevated exposure levels may occur in the event of engine seal wear or failure, facilitating greater quantities of lubricant-based contaminants entering cabin air. Symptoms of aerotoxic syndrome linked to such exposures include severe cognitive impairment, which poses flight safety risks when aircrew are significantly affected.[401] Taking the above into account, it is imperative that less toxic alternatives to TCP are found.[399]

To address this issue, several recent studies have investigated the use of alternative additives such as ionic liquids with base oils in conjunction with DLC coatings.[402,403,404] Another study by Zhou et al.[405] investigated the wear properties in DLC-Steel contacts, looking at how these properties varied in the presence of ZDDP vs ionic liquids with and without phosphate anions. Reciprocating steel balls were used against DLC-coated flat surfaces in lubricated environments and the study observed significantly higher wear on the steel counterface when phosphate additives were present, despite DLC's usual wear-reducing properties. This increased wear was not observed in systems with non-carbon coatings or phosphate-free additives, suggesting a carbon-catalyzed tribochemical mechanism with the phosphate group. AES, XPS, and TEM analysis revealed greater porosity in the tribofilms on the formed on steel surface in the DLC-phosphate systems. Zhou et al.[405] proposed that the excessive wear observed is due to the catalytic role of carbon in the DLC coating, which accelerates tribochemical reactions between phosphate additives and the steel ball surface causing material removal. Interestingly, this study found that higher concentrations of phosphate anions present in the additive led to greater amounts of ball

wear, and that phosphate containing additives without these anion groups, but featuring a carboxylate anion and a phosphonium cation did not produce increased wear as seen in the systems with the other two additives.[405,406]

To our knowledge, the behaviour of TCP between DLC coatings and ferrous surfaces has not been studied via MD simulation. Experimental studies have examined the performance of TCP in steel contacts under various conditions,[407,408,409] but no work has investigated its behaviour in the presence of DLC. Furthermore, the effect of shear stress on the tribochemical behaviour of TCP remains poorly understood, particularly in the context of DLC-coated contacts. Variations in shear stress can influence the rate and pathways of TCP decomposition and film formation,[271,410] but this relationship has not been systematically studied through experiments or simulations for DLC-ferrous surface contacts. MD simulations provide a way to examine the chemical processes that occur under different sliding conditions, including decomposition and film formation. This study uses NEMD simulations with the ReaxFF forcefield to analyse the mechanochemical response of TCP at DLC-ferrous surface interfaces under different temperatures and sliding speed-induced shear stresses. These simulations will help identify the conditions that influence TCP reactivity and its tribofilm formation pathways, which is critical for the design of future, less toxic anti-wear additives for the aviation industry.

4.2 Methodology

4.2.1 DLC System Setup

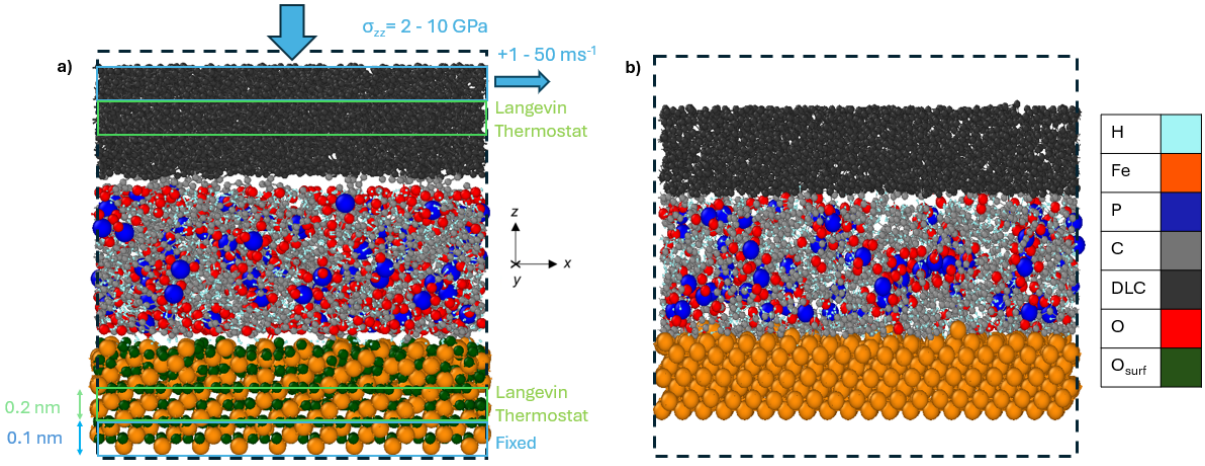


Figure 4.1: Side-view snapshots showing the model systems with TCP molecules confined between a) α-Fe(110) and b) α-Fe_2O_3(0001) surfaces. Systems shown after minimisation and equilibration, but before heating, compression, and shear. Black dashed line shows the simulation box. Rendered using OVITO.[306]

Fig. 4.1 shows side-view snapshots showing the model systems with TCP molecules confined between a) α-Fe(110) and b) α-Fe_2O_3(0001) surfaces. A total of 90 TCP molecules were confined between the DLC and ferrous surfaces resulting in average surface separation distances of approximately 18 Å. To generate an amorphous DLC block, the following procedure was used with ReaxFF. Starting from

a regular carbon block structure made using LAMMPS,[323] the system was first minimized using the conjugate gradient algorithm. Following minimization, the system was equilibrated at 300 K and 10 MPa to simulate thermal equilibrium conditions. The block was then heated to 8000 K under the NVT ensemble in 150 ps using a timestep of 0.25 fs. This ensured that the structure transitioned into a fully molten state promoting the breakdown of crystalline order. The system was then cooled from 8000 K to 300 K in the NVT ensemble in 150 ps. Afterwards, the system was annealed at 300 K for 400 ps under NVT conditions to allow the amorphous structure to stabilise without external pressure fluctuations. This process has been used by several studies to create amorphous DLC blocks for MD simulations.[411, 412,413,414] The dimensions of the final DLC were approximately 50.28 Å in the x-direction, 43.25 Å in the y-direction, and 29.99 Å in the z-direction. The Fe_2O_3 and α-Fe(110) slabs were matched to have the same x and y dimensions as the DLC block, with a thickness of 10.6 Å and 11 Å respectively. After the amorphisation process, a passivation step was applied to the DLC surface. A layer of H_2O molecules was introduced to the non-passivated DLC surface to reactively terminate under-coordinated carbon atoms. The passivation process was carried out using ReaxFF in LAMMPS under PBCs in the x and y directions, with a fixed boundary in z. The system was equilibrated at 300 K with Langevin thermostats applied near the top of the surface. Unreacted H or O atoms or H_2O molecules that did not bond with the DLC surface during equilibration were removed prior to the production stage. Similar processes have been used to passivate DLC surfaces in other molecular dynamics studies.[415]

4.2.2 Simulation Details

All the simulations were performed using the LAMMPS.[323] Simulations where performed with PBCs in the x and y directions, with a fixed boundary in z. The velocity Verlet[121] integration scheme was used with a timestep of 0.25 fs.[159] The simulations were conducted in three phases: energy minimisation, equilibration, and then simultaneous heating, compression, and shear. The energy minimisation phase was conducted using the conjugate gradient method. During the equilibration phase, the two surfaces were brought closer together by applying a low pressure of 10 MPa. The system was pressurised by applying a constant normal force in the z-dimension equally distributed across the topmost layer of atoms (0.1 nm) in the upper surface, as shown in Fig. 4.1a. The lowermost layer of atoms (0.1 nm) in the bottom surface was kept fixed in the z-direction. Equilibration at 10 MPa ensured that a realistic liquid density of TCP (1.2 ± 0.1 g cm^{-3}) was obtained prior to the heating, compression, and shear phase.[416]

Following the equilibration phase, the compression and shear phase of the simulations were run for 1 ns. For the DLC systems, constant sliding velocities of 1, 2.5, 5, 10, 20, 30 and 50 m/s were applied to the outer layer of atoms in the upper surface in the x-direction, as shown in Fig. 4.1a. These velocities were selected to span a broad range of sliding speed conditions relevant to tribological contacts found in aviation applications, such as those in turbine engines, gear systems, and bearings, where local sliding speeds can vary widely depending on operating conditions and component geometry. For example, sliding speeds in gas turbine engine components typically range from a few metres per second up to 100 m/s in turbine engines, as they have to operate at rotational speeds of between 15,000-60,000 rpm.[417] From a computational perspective, this range allows for investigation of the tribochemical response of TCP under different shear rates while remaining within the limits of simulation accuracy achievable

using ReaxFF in NEMD simulations.

The pressure was increased to 5 GPa using the same method as during the equilibration. The temperatures studied varied from 300-600 K in 100 K intervals using a Langevin thermostat with a damping parameter of 25 fs.[324] The thermostat acted only in the *y*-direction and on the central layer of atoms (0.2 nm) in the upper and lower surfaces, as shown in Figure 4.1a. This approach enabled a temperature gradient to develop in the centre of the TCP film and heat to be dissipated at the surfaces, as would occur experimentally.[325] The thermostat ensured that the average system temperature was within a maximum 35 K of the target value at the highest sliding velocity. The selected range of temperatures and pressures spans those expected in applications of TCP such as anti-wear additives in gearboxes for gas turbine engines[290] and as vapour-phase lubricants.[316]

The temperature, pressure, and sliding velocity were all increased simultaneously to capture their combined effects on the reaction rate. Some previous NEMD simulations of mechanochemical reactions with ReaxFF have isolated the effects of temperature, pressure, and sliding by using separate phases of the simulations.[286,319,326] The number of additive molecules in the NEMD simulations is finite, so separating the simulations into different phases (thermalisation, pressurisation, and sliding) means that fewer molecules are available to react in any subsequent phases. Conversely, in experimental mechanochemistry studies,[96,279,314] additive-containing lubricants are entrained between the sliding surfaces, continuously replenishing the number of additive molecules. Hence, for these NEMD simulations, we elected to increase the temperature, pressure, and sliding velocity to their target values concurrently.[27]

4.2.3 Force Field Details and Validation

We used the bond order-based ReaxFF potential for all of the NEMD simulations. The C/H/O/Fe/P ReaxFF parameters used in this study were developed by Khajeh et al.[159] The parameters for Fe/O/H were originally developed by Aryanpour et al.[328], Fe/C by Zou et al.[329], P/O/C/H by Verlackt et al.[330], and C/H/O by Chenoweth et al.[157] The point charges on the atoms vary dynamically during the simulations and were calculated using the charge equilibration (QEq) method.[156] Chemical bonding information was output every 1.0 ps, using a bond order cutoff of 0.3 to identify covalent bonds.[159] The combined C/H/O/Fe/P ReaxFF parameters [159] have been successfully applied to study the thermal and mechanochemical decomposition of phosphate esters with different alkyl and aryl substituents on several ferrous surfaces.[26,27,271,301,304] The ReaxFF parameters have also been validated against first principles calculations for the non-dissociative adsorption energy, C-O bond dissociation energy, and P-O bond dissociation energy for trialkyl phosphates on α-Fe(110).[304] The ReaxFF parameters for Fe/O/H have been shown to accurately reproduce the experimental lattice constants for α-Fe (within 1 %).[329]

4.2.4 Bond Length Analysis

To identify molecular events such as bonding to surfaces and inter-molecular reactions during reactive molecular dynamics simulations, a Python-based method was developed to parse and analyse LAMMPS ReaxFF input and output files. Similar analysis has been performed in previous NEMD simulations with ReaxFF.[418] Atom identities, atom types, and Cartesian coordinates were extracted from the LAMMPS

input data files, while atomic connectivity was obtained from ReaxFF bond order output files. This data was used to construct a molecular graph in which atoms are represented as nodes and bonds as edges. Each subgraph corresponds to a molecule, molecular fragment, or surface structure. Each node stores the atom's ID, element, Cartesian coordinates, and a unique label generated from its local bonding environment. Unique labels were assigned by hashing the atom's nearest-neighbour signature, allowing chemically distinct atoms of the same element to be differentiated, such as oxygen atoms in hydroxyl and phosphate groups. Bond lengths were assigned to each edge by calculating the Euclidean distance between bonded atoms, accounting for PBCs in the x and y-directions using the box dimensions parsed from the simulation input. Filtering operations were applied to exclude atoms or structures based on atom type or size, enabling targeted analysis such as the removal of hydrogen or bulk substrate atoms from the results.

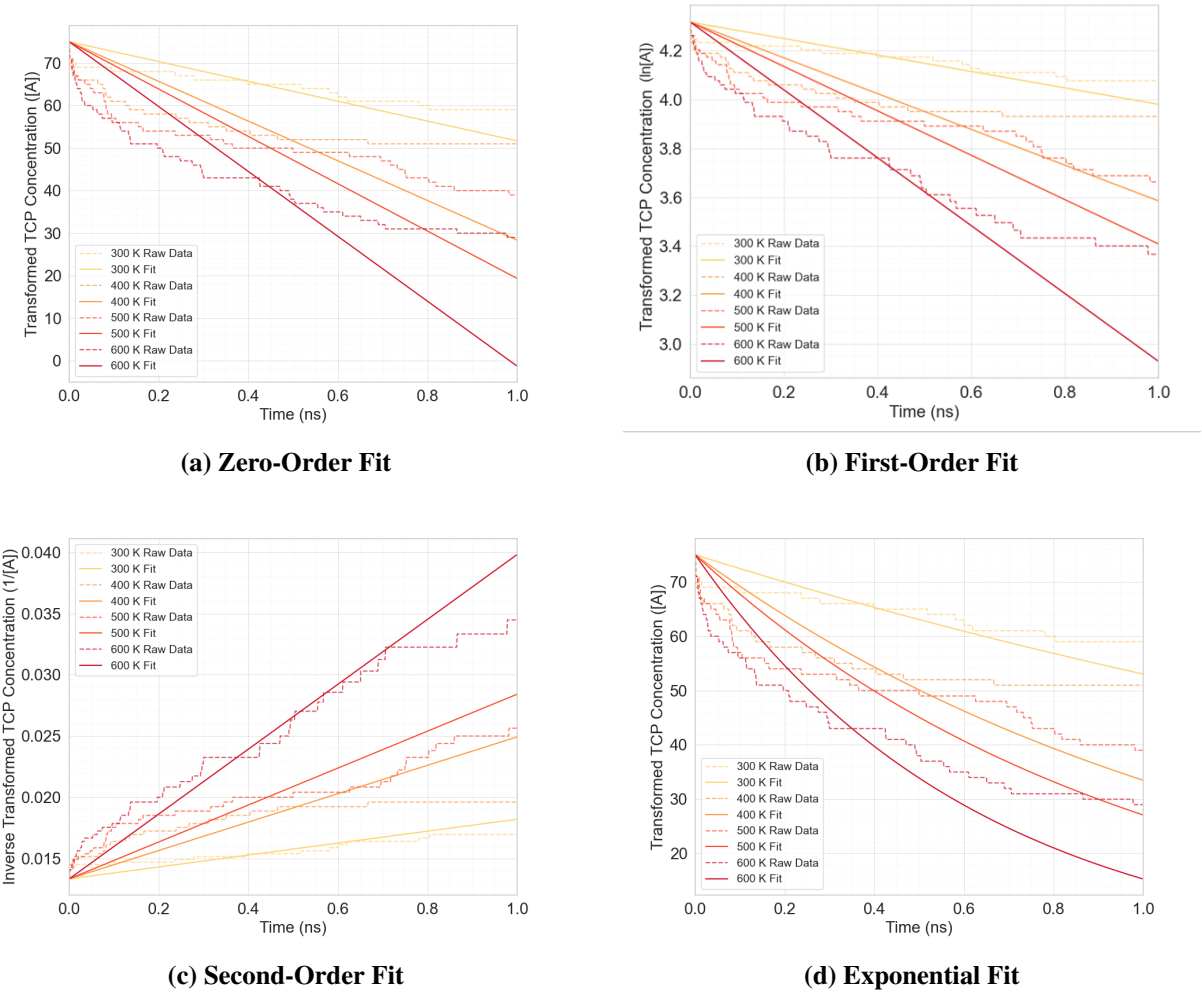
For multi-timestep simulations, bonded atom pairs were extracted at each timestep from ReaxFF bond files, and atomic coordinates were obtained from the corresponding LAMMPS trajectory files. This data was stored as dictionaries mapping atom IDs to Cartesian positions, knowing that atom identities remained fixed throughout the simulation. As atom types in the ReaxFF bond files and LAMMPS dump files may differ, atom types were cross-referenced against the original LAMMPS data file to maintain consistency across all sources. At each timestep, molecular subgraphs were reconstructed using the updated atom types, coordinates, and bond connectivity. Bond lengths were recalculated from atomic positions using Euclidean distances, accounting for PBCs. Unique atom labels were extended as new bonding environments appeared, including those involving surface interactions. Subgraphs were compared to those from the previous timestep to identify bond-breaking and bond-forming events. Subgraph membership for each atom was retained across timesteps, enabling the tracking of molecular identity through reactions or fragmentation. A conditional depth-first search (DFS) was implemented to capture extended interactions such as bonding between molecules and surfaces. This method expanded each molecular subgraph by tracing connected atoms, with optional constraints to halt traversal at specified node attributes, allowing targeted analysis of reactive structures. Bond length evolution was tracked over time for each bond type identified through neighbour signatures. These values were stored and used to analyse structural dynamics and reaction pathways throughout the simulation.

4.3 Results and Discussion

4.3.1 Nature of TCP Dissociation at Different Sliding Velocities

Model	0 m/s	1 m/s	2.5 m/s	5 m/s	10 m/s	20 m/s	30 m/s	50 m/s
Exponential	-0.1425	-0.2779	0.4715	0.9182	0.9848	0.9745	0.9916	0.9918
First-Order	-0.0884	-0.0427	0.5109	0.9486	0.9926	0.9562	0.9839	0.6015
Second-Order	0.4182	0.5500	0.9219	0.9006	0.8034	0.5778	0.3637	0.7747
Zero-Order	-0.8737	-1.1129	-0.6147	0.2949	0.4951	-0.5282	-0.6688	-2.1633

Table 4.1: R^2 values for the different fitting methods at 500 K, 5 GPa for DLC-Fe_2O_3 across the investigated sliding speeds.



(a) Zero-Order Fit

(b) First-Order Fit

(c) Second-Order Fit

(d) Exponential Fit

Figure 4.2: Plots of different fitting methods used to determine the rate constant for the DLC-Fe_2O_3 system at 1 m/s, 5 GPa, R^2 for each fit shown in the legend.

To quantify the rate constant of TCP dissociation under different sliding velocities, temperatures, and pressures, four kinetic models were fitted to the amount of TCP molecules remaining intact over time during the MD simulations: zero-order, first-order, second-order, and exponential decay. As the concentration at the start of the production run can be counted directly, fits were constrained to start from the initial known concentration. These rate constants were used to estimate decomposition rate constants for subsequent analysis using the Bell and Arrhenius formulations.[90]

The zero-order model assumes a constant (linear) rate of decomposition, independent of the number of remaining TCP molecules. The form of the zero-order model is given by:[419]

$$[A](t) = [A]_0 - kt, \tag{4.1}$$

where $[A](t)$ is the number of intact TCP molecules at time t, $[A]_0$ is the initial number, and k is the rate constant. Previous experiments investigating the performance of ZDDP have suggested that the

rate of tribofilm growth followed either zero-order or fractional-order kinetics.[95,96,314,420] As discussed in the previous chapter, the observation of zero-order kinetics indicates that the rate-determining step for tribofilm growth involves surface-adsorbed molecules,[96] which masks the underlying first-order kinetics.[343] As the supply of TCP is finite in our simulations, the reaction cannot follow zero-order kinetics indefinitely, unlike experimental conditions where a constant supply of the additive may be entrained into the contact. This means that a change in kinetics should be expected as the number of available reactant molecules and surface reactive sites decreases.

A first-order reaction proceeds at a rate that is directly proportional to the concentration of the reactant. The form of the first-order rate law is given by:[421]

$$\ln[A](t) = \ln[A]_0 - kt, \tag{4.2}$$

where $[A](t)$ is the number of intact TCP molecules at time t, $[A]_0$ is the initial number, and k is the rate constant. First-order reactions are characterised by a constant half-life, independent of the initial concentration, which reflects an exponential loss of reactant over time, which is likely given the finite amount of TCP molecules available to react in our systems.[27] The exponential model is mathematically equivalent to the first-order model but avoids linearisation, directly fitting to the intact TCP population using:[422]

$$[A](t) = [A]_0 \exp(-kt). \tag{4.3}$$

This form describes exponential decay, which arises when the rate of change is proportional to the current quantity. In confined systems such as TCP between sliding surfaces, this behaviour indicates that each molecule decomposes with a constant probability per unit time, independent of other molecules. As such, good agreement with exponential or first-order models would suggest that TCP dissociation under shear occurs predominantly through independent, thermally or mechanically activated events (e.g., SATA processes).

Second-order kinetics describe reactions where the decomposition rate depends on the square of the reactant concentration or the product of two reactant concentrations. In its simplest form, for a unimolecular second-order process, the fitting equation is:[423]

$$\frac{1}{[A](t)} = \frac{1}{[A]_0} + kt, \tag{4.4}$$

where $[A](t)$ is the number of intact TCP molecules at time t, $[A]_0$ is the initial quantity, and k is the second-order rate constant, with units of $M^{-1}s^{-1}$ (M = mol/litre). In this model, the rate increases with the square of the reactant concentration, so that a higher initial concentration leads to a faster rate of decomposition. This form of kinetics typically arises in reactions involving bimolecular interactions or cooperative processes such as surface-catalysed reactions involving two adsorbed species.[424]

As shown by Table 4.1, the best performing fitting method depends on the sliding velocity, and hence shear stress applied to the system (Fig. 4.3 for variation in shear stress with sliding speed). At lower speeds (0-2.5 m/s), the second-order model yields the highest R^2 values, suggesting that TCP decom-

position under less shear is dominated by cooperative mechanisms between reactant species. In other words, the good agreement with the second-order kinetics model would suggest that decomposition occurs through collisions between TCP molecules or requires activation involving pairs of TCP molecules or molecule–surface interactions.[424] However, as the sliding velocity increases and shear-driven dissociation becomes dominant, the system transitions away from second-order kinetics, as observed in the decreasing R^2 values for this model at higher velocities. At intermediate speeds (5-10 m/s), both first-order and exponential fits perform well, indicating decomposition progresses more uniformly and is increasingly dominated by single-molecule reactions. At the highest speeds (20-50 m/s), the exponential model consistently outperforms the others, implying that under strong shear, the reaction follows a first-order decay with a constant probability of bond breaking. Such trends have also been seen in other similar experimental and MD simulations.[27,88,271,410] This trend was observed across all temperatures studied and was consistent for both DLC–Fe_2O_3 and DLC–α-Fe systems, suggesting that shear stress is the dominant factor governing the decomposition kinetics across different surface chemistries.

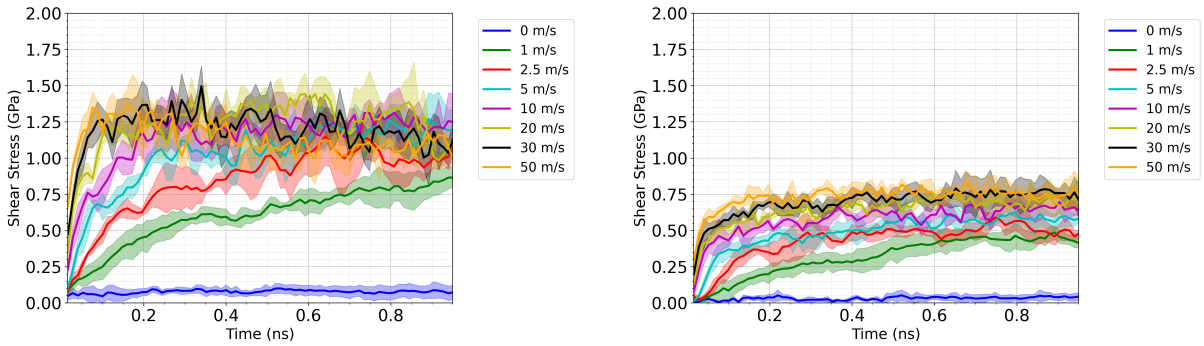
The Bell model[90] and its extensions are commonly used to describe force-dependent reaction kinetics in systems where an applied mechanical force lowers the activation barrier.[88] This framework is typically applied to unimolecular events such as bond cleavage, where the reaction rate can be expressed as an exponential function of the applied force. While the classical Bell model assumes a rigid energy barrier and a constant activation volume, recent extensions have introduced second-order effects to account for the force-induced deformation of the potential energy surface.[425] This extended-Bell model includes a quadratic dependence of the activation energy on force or stress, as shown in Eq. (4.5), where $E_{act}(F)$ is the force-dependent activation energy, E_0 is the zero-force activation energy, F is the applied force, $\Delta x^{\ddagger}$ is the activation length, and χ_I and χ_T are the curvatures of the initial and transition state potentials, respectively.

$$E_{act}(F) = E_0 - F\Delta x^{\ddagger} + \frac{F^2}{2}(\chi_I - \chi_T) \tag{4.5}$$

This approach is analogous to treatments in the Evans–Polanyi model[91] and Marcus theory,[426] where the energy landscapes are expanded as parabolas. In such models, the activation barrier is not only lowered by force but also reshaped, leading to the Hammond effect,[427] in which the transition state shifts toward the reactant state under increasing stress. This can be visualised as a compression of the x-axis in Fig. 2.2. Calculation of χ_I and χ_T would require techniques such as ab initio or DFT simulations, which is beyond the scope of this study. The first-order Bell model (Eq. (2.2)) can still be applied to describe systems such as ours which seemingly displaying second-order kinetics under low shear stresses. This approximation is valid under several conditions: (i) the reaction is rate-limited by a unimolecular step, such as the bond cleavage or surface adsorption of a single molecule; (ii) the applied force acts primarily on a single species and aligns with the reaction coordinate; (iii) the deformation of the potential energy surface under force is small such that higher-order corrections to the barrier shape are negligible, and (iv) the reaction appears pseudo-first-order, for example when the surface acts as a constant, non-depleting reactive partner.[93] However, at higher applied forces, or in cases where changes in the curvature of the energy landscape become significant, the extended-Bell model provides a more accurate description by incorporating second-order effects.[93]

As discussed in Chapter 3, the primary TCP dissociation mechanism observed in those systems was based on the adsorption of TCP to a surface followed by P-O bond cleavage where decomposition also increased with temperature and system shear stress, satisfying requirement (i). Furthermore, we also discussed that experimental studies of similar systems observed zero-order kinetics indicating that the rate-determining step for tribofilm growth involves surface-adsorbed molecules,[96] which masks the underlying first-order kinetics,[343] suggesting that requirement (iv) is most likely the case in our systems. No works seem to have investigated changes in the potential energy surface of TCP under different stresses, however, a study by Liu et al. showed that applied shear stress leads to conformational changes in PFPE lubricant molecules confined and sheared between ferrous surfaces, increasing their potential energy due to bond deformation.[428] They applied a constant velocity of 100 m/s and applied normal loads up to 2 GPa, observing changes of up to 75 eV in the potential energy increment between applied normal loads of 300 MPa and 1000 MPa. The shear stresses in these systems were not reported so we cannot know how large a difference in shear stress there was between these conditions, however, these findings do seem to suggest that application of even small shear stresses to a system like ours would cause a non-negligible change in the PES, violating requirement (iii). Further work could calculate of the potential energy surface without and without applied stress (e.g., via DFT) to remove ambiguity. Finally, for requirement (ii) to be fully satisfied in our case, the applied shear stress needs to predominantly act on a single species (i.e., a single bond type) aligned to the direction of applied stress.[429] In other words, we should see bond stretching of P-O bonds primarily in the x-axis of our systems; this is investigated below.

4.3.2 Sliding Velocity Effects on Shear Stress



(a) α-Fe(110) **(b) DLC-Fe_2O_3**

Figure 4.3: Average variation in shear stress over time for different sliding speeds for (a) DLC-Fe_2O_3 and (b) DLC-α-Fe(110).

Fig. 4.3 shows the average evolution of shear stress at different sliding speeds for the DLC-Fe_2O_3 and DLC-α-Fe(110) systems, at 5 GPa, across the different temperatures (300 K - 600 K), where the shaded region shows 1-σ variation across temperatures. Shear stress in the system is calculated by measuring the total lateral force acting on the fixed bottom surface in the direction of sliding, where force is divided by the surface area to obtain the shear stress. At 0 m/s, although the system is subject to a purely compressive load in the z-direction, non-zero shear stress values are still observed. This stress anisotropy is

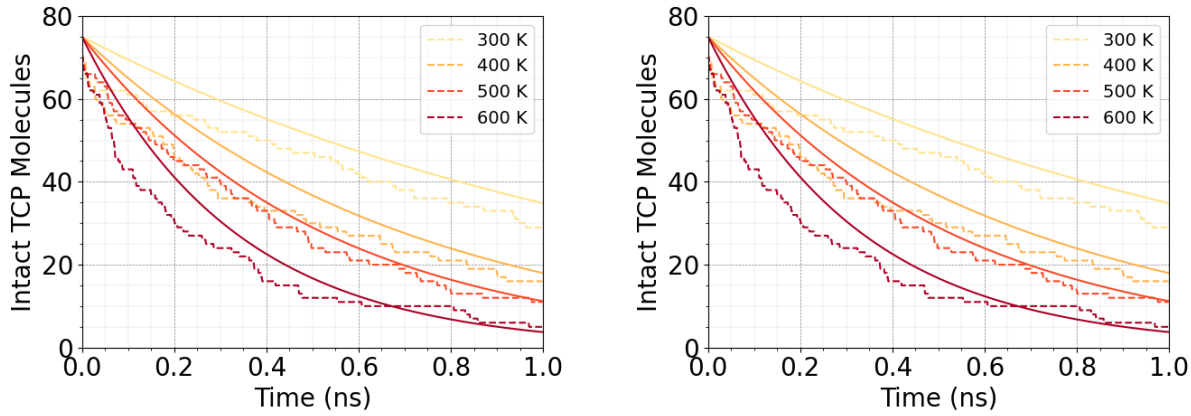
expected for uniaxial compression under applied normal stress.[430] The periodic boundaries allow lateral translation of the system, so transient shear stresses can develop without an imposed shear deformation. Additionally, we compute the absolute magnitude of the shear stress, meaning that instantaneous fluctuations in both positive and negative directions are treated equally. As a result, the average shear stress over time may be closer to zero than appears in the plot.

As illustrated by Fig. 4.3, the shear stress increases with sliding time, before plateauing towards a constant value. Faster sliding velocities take less time to reach a non-equilibrium steady state. Increased sliding speeds result in increased steady state shear stresses, as expected for boundary-lubricated systems.[431] Above approximately 20 m/s, there is a saturation in the shear stress. The observed saturation likely indicates the onset of a dynamic equilibrium at the sliding interface, where further increases in sliding velocity do not lead to proportional rises in resistance. This is probably because the molecules at the surface have reorganised into a more energetically favourable configuration with less resistance to sliding. This behaviour aligns with findings the study by Heyes et al.[432] who used molecular dynamics simulations to show that simple liquids under large shear rates exhibit structural layering aligned with the shear direction. Such rearrangement facilitates flow by reducing shear viscosity, thereby allowing the system to accommodate increasing deformation without significant rises in internal stress.[432]

For the iron systems, a modest reduction in shear stress is observed at higher velocities (30 and 50 m/s) as the simulations progress. One plausible explanation is the enhanced formation of tribofilms at these speeds on the iron surface, a process that could be accelerated more by the reactive α-Fe(110) surface (Fig. 3.10) than the for iron oxide. This is discussed in more detail in for these systems in Section 4.3.4). Shear thinning behaviour under higher strain rates caused by the higher sliding speeds may also contribute to reduced shear stress at these conditions.[433] The magnitude of shear stress also differs between iron and iron oxide surfaces at comparable speeds, with consistently higher values observed on iron. This behaviour is most likely attributed to the stronger adhesion of TCP molecules on iron than iron oxide due to its higher surface energy and chemical reactivity. This is reflected by the higher Derjaguin offsets shown by the bare iron surfaces compared to iron oxide surfaces (Table A.3).[281]

4.3.3 Sliding Speed Effects on TCP Dissociation

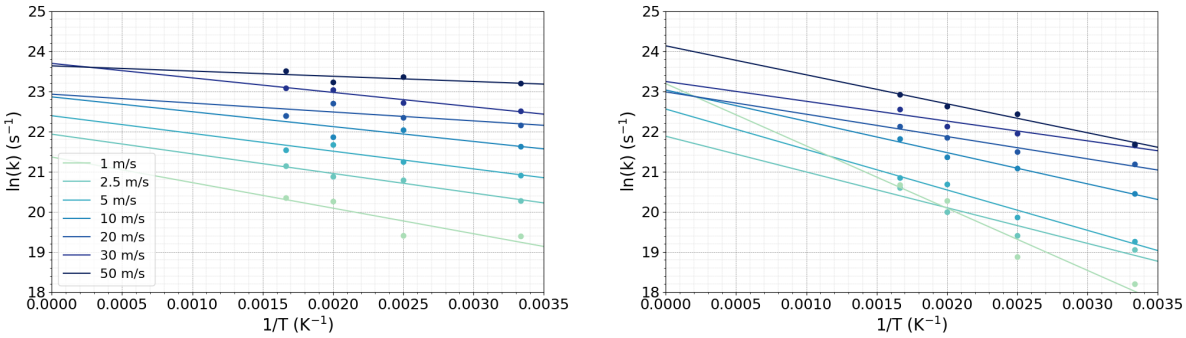
Fig. 4.4 shows a comparison of TCP dissociation at 5 GPa at different temperatures for the a) DLC-α-Fe(110) and b) DLC-Fe_2O_3 systems at 10 m/s, using an exponential fitting function (Eq. (4.3)). Fig. B.2 and Fig. B.3 show the TCP dissociation curves using exponential fits at 5 GPa for the different temperatures and sliding speeds investigated for the DLC-α-Fe(110) and DLC-Fe_2O_3 systems respectively. As shown by Fig. 4.4, there is slightly greater TCP dissociation in the DLC-α-Fe(110) systems than the DLC-Fe_2O_3 systems across all sliding speeds, although this difference is not as pronounced as the difference in dissociation between the α-Fe(110) and the Fe_3O_4 surfaces observed in Section 3.3.2. As discussed in Section 3.3.2, the difference in TCP dissociation between the different systems is most likely due to the higher surface energy and adhesion strength of TCP molecules on iron compared to iron oxide. This is demonstrated by Fig. 4.3, where the shear stresses are higher at all speeds in the DLC-α-Fe(110) systems. DFT studies have found that the surface energy of α-Fe_2O_3(0001) to be be-



(a) DLC-α-Fe(110) **(b) DLC-Fe_2O_3**

Figure 4.4: Comparison of TCP molecule decomposition at 5 GPa at different temperatures for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3 at 10 m/s.

tween 1.78 and 2.4 J m^{-2},[434,435] much closer to reported surface energy values of the α-Fe(110) (also 2.4 J m^{-2}), [311] potentially explaining why dissociation rates are more similar at comparable conditions for these two surfaces.



(a) DLC-α-Fe(110) **(b) DLC-Fe_2O_3**

Figure 4.5: Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3.

As shown by Fig. B.6, there is noticeable TCP dissociation when there is no sliding speed applied to the top surface (negligible shear stress). This behaviour could be attributed to a couple of different factors. The presence of normal load may increase the local pressure at the interfaces, which may induce mechanical deformation and expose reactive surface sites to catalyse the dissociation of TCP. This effect is closely linked to the contact stiffness of the materials involved, which depends on their elastic moduli (as discussed in Section 3.3.6). Stiffer interfaces such as those in the studied systems can exhibit contact stiffness high enough that there is localised stress transmission to the molecules.[436] This concentrated stress can stretch molecular bonds as shown by Fig. 4.8 and Fig. B.5, and lower the activation energy required for dissociation by promoting bond cleavage. Nevertheless, the rates of dissociation under normal stress only are lower than those where shear stress is applied.

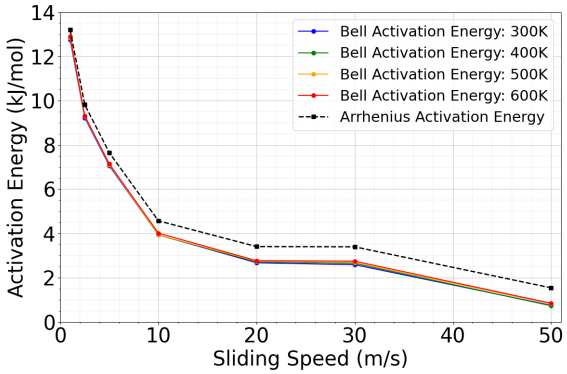
Temperature	α-Fe (Å^3)	α-Fe_2O_3 (Å^3)
300 K	6.492 $\pm$ 3.88	4.540 $\pm$ 2.50
400 K	7.197 $\pm$ 3.30	4.697 $\pm$ 2.61
500 K	8.607 $\pm$ 2.75	6.432 $\pm$ 2.53
600 K	9.889 $\pm$ 3.56	7.331 $\pm$ 2.71

Table 4.2: Activation Volumes for α-Fe and α-Fe_2O_3 systems at different temperatures at 10 m/s, calculated from simulations performed at temperatures ranging from 300 K to 600 K in 100 K intervals, under varying normal pressures (2-10 GPa with 2 GPa intervals).

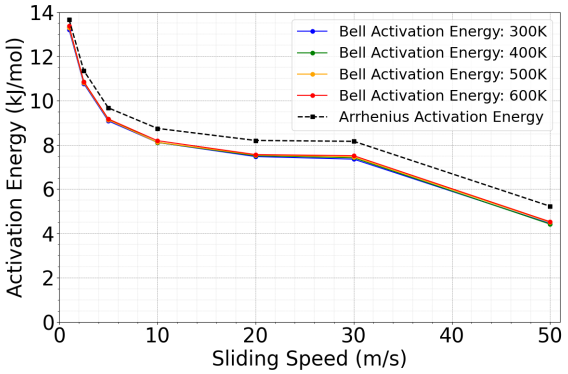
To calculate activation energies and pre-exponential factors using Eq. (2.2), an appropriate activation volume is required. As discussed in Section 3.3.6, to calculate the relevant activation volume for each sliding speed, we would need to perform simulations for a range of pressures at each of the studied temperatures for each sliding speed. The computational expense of these simulations makes such an approach impractical. Hence, we selected a single sliding speed (10 m/s) as a reference condition and performed simulations at multiple temperatures ranging from 300 K to 600 K, under varying normal pressures (2-10 GPa with 2 GPa intervals). For this speed, the average shear stress across the temperature range was used to represent the mechanical load, which, as shown by Fig. 4.3, remains in the same order of magnitude across all temperatures for all the sliding speeds in both systems. Activation volumes were then calculated at each temperature based on the temperature-dependent dissociation rate constants and the average shear stress. The resulting activation volumes were subsequently used to calculate activation energies and pre-exponential prefactors across the range of studied sliding speeds at the relevant temperatures. Given that the molecular-scale deformation and bond-breaking mechanisms of TCP are affected by local stresses rather than global shear stress (see Section 4.3.4 for more discussion), we postulate that the extracted activation volume remains valid for systems with comparable stress magnitudes. The alternative of using an average temperature to calculate a single activation volume from the different sliding speeds would overlook the non-linear dependence of the reaction rates on temperature.

Table 4.2 show the calculated activation volumes. The activation volumes are slightly higher for the DLC-α-Fe systems compared to the α-Fe_2O_3 systems, and are roughly comparable to those of the α-Fe values from Chapter 3 (Table A.4). The activation volume has been shown to span several orders of magnitude across different studies, ranging from about 0.2 Å^3 to 53 Å^3 for tribofilm growth of ZDDP.[420,437] This variation does not seem to result from differences in measurement scales or molecular species.[89,438] The ΔV^* can depend on the direction of sliding in relation to the crystallographic orientation of the surface,[439,440] hence, for the same molecule, ΔV^* can vary based on the reactivity of the substrate. There are several interpretations of ΔV^* including relating it to the size of atoms, ions, or bonds.[89,441,442] Other studies which have found larger values of ΔV^* relate it to molecular size,[89,438] such as the product of the area of a molecule and the distance over which force is applied.[279] Another view, often used in conjunction with molecular dynamics simulations, connects ΔV^* to the change in volume or bond length of molecules as they deform from their thermal equilibrium shape.[285,286,443] Taking the latter interpretation, the positive sign of the measured activation volumes suggests that the dissociation mechanisms in these systems are shear-driven rather than compression-driven.[344] Furthermore, the increase in magni-

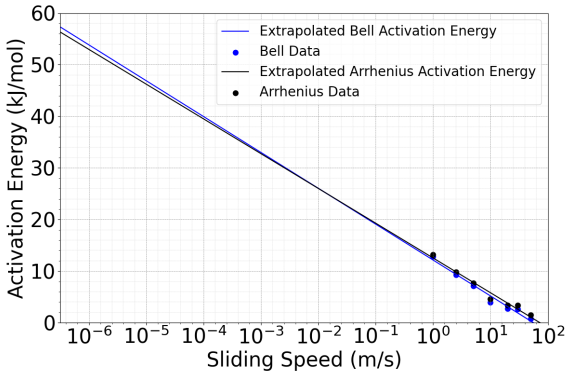
tude of these values with temperature suggest that higher temperature aids the TCP dissociation process. Whereas the higher activation volumes for the α-Fe compared to the α-Fe_2O_3 implies that stronger adhesion on iron promotes greater molecular deformation during TCP dissociation, leading to increased bond stretching or molecular volume expansion in the transition state.



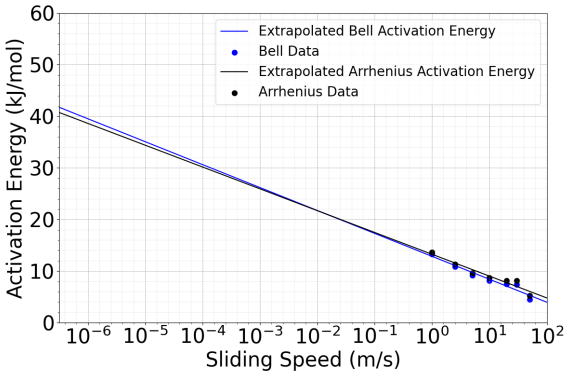
(a) DLC-α-Fe(110): Activation energy vs sliding speed across temperatures.



(b) DLC-Fe_2O_3: Activation energy vs sliding speed across temperatures.



(c) DLC-α-Fe(110): Extrapolated Arrhenius vs Bell activation energies.



(d) DLC-Fe_2O_3: Extrapolated Arrhenius vs Bell activation energies.

Figure 4.6: Comparison of Arrhenius and Bell model activation energies across the studied sliding speeds for a) α-Fe and b) α-Fe_2O_3 systems. Fit of the logarithm of the activation energy to the sliding speeds used in this study, extrapolated to order of magnitude of sliding speeds used in experiments for c) α-Fe and d) α-Fe_2O_3 systems.

Fig. 4.6a and Fig. 4.6b shows a comparison of Arrhenius and Bell model activation energies for the α-Fe(110) and Fe_2O_3 systems, while Appendix B shows the calculated Arrhenius activation energies for each speed, and pre-exponential factor values. Appendix B shows the average Bell activation energies for each speed for both systems. The Arrhenius activation energy is calculated from the slope of the 2D fits to the natural logarithm of the rate constants versus the inverse of temperature while pre-exponential factors are taken from the y-intercepts of these slopes, as illustrated by Fig. 4.5.[444] To calculate the Bell activation energies, we use the measured shear stresses for each system in Eq. (2.1) rather than the normal stress. For TCP decomposition in the DLC-α-Fe(110) system, Arrhenius E_a decreases with increasing sliding speed (and hence shear stress) from 13.2 to 1.54 kJ/mol, while Arrhenius E_a decreases from 13.66 to 5.23 kJ/mol in the DLC-α-Fe_2O_3 systems. Average Bell E_a decreases with increasing

sliding speed from 12.74 to 0.76 kJ/mol in the DLC-α-Fe(110) system, while average Bell E_a decreases from 13.26 to 4.43 kJ/mol in the DLC-α-Fe_2O_3 systems, providing further evidence that, as sliding speed increases, shear stress becomes the dominant driver of reaction. The lower activation energies observed for the DLC-α-Fe(110) systems compared to those calculated for the α-Fe(110) at a similar condition (24.3 kJ/mol at 10 m/s) systems in Chapter 3 may be attributed to a higher concentration of TCP molecules at the interface. Increased molecular concentration enhances surface coverage and the likelihood of reactive configurations, facilitating decomposition and therefore lowering the apparent activation energy. However, the use of 90 TCP molecules in the DLC-α-Fe(110) systems, compared to only 48 in the α-Fe(110) systems, may also result in improved statistical sampling. Thus, the reduced activation energies may reflect better sampling rather than intrinsic differences in system reactivity.

As sliding speed increases, we postulate that the rate at which mechanical forces are applied to molecular bonds rises, resulting in a higher frequency of activation attempts. In other words, at higher sliding speeds, the contact covers a greater sliding distance over the same time period leading to more opportunities for the relevant bonds to be cleaved due to the applied shear stress component. This interpretation is supported by a recent study conducted by Rana et al.[445] which extended the Evans-Polanyi perturbation theory[91] to quantify how normal and shear stresses influence tribochemical reaction rates at sliding interfaces. The authors investigated the shear-induced decomposition of ethyl thiolate species adsorbed on a Cu(100) single crystal surface using AFM under ultra-high vacuum conditions, finding that the y-intercept of a linear fit of the dissociation rate constant of the thiolate species across different shear stresses is higher than the dissociation rate constant from thermal decomposition only.[445] In other words, applied shear stress still lowers the activation energy even in the absence of an applied normal force. Their results demonstrate that the effective activation energy under shear is not solely governed by mechanical loading but also by changes in the intrinsic kinetics, resulting in an asymptotic rate constant that increases with sliding speed. This also aligns with the observed trends in the pre-exponential factors, where both systems have a general trend of increasing with sliding speed, and are generally comparable in magnitude between the two ferrous surfaces. Furthermore, increased thermal agitation at the sliding interface contributes to an enhanced reaction rate. This thermal agitation arises from frictional heating generated by the relative motion between the non-bonded TCP molecules and the ferrous surface, resulting in localised temperature increases that help overcome activation energy barriers. Indeed, at the lowest sliding speed, we see maximum increases in temperature of 2% from the target temperature, compared to 6% at the highest sliding speed. The greater reduction in activation energy for TCP decomposition on α-Fe compared to Fe_2O_3 with increasing sliding speed suggests higher stress sensitivity for this surface. This again could be attributed to stronger adhesion and surface reactivity of α-Fe, which leads to stronger molecular bonding to the α-Fe surface, facilitating more effective stress transfer to reactive bonds, lowering the energy barrier more significantly than on α-Fe_2O_3.[271]

A previous experimental study suggested that the thermal activation energy for TCP decomposition on stainless steel is approximately 60 kJ/mol.[305] This is close to the value obtained for our systems when extrapolated to experimentally-accessible sliding velocities (40-60 kJ/mol). In-situ ball-on-disc experiments conducted at 25 °C and 60 °C with a load of 14 N at a sliding speed of $5x10^{-5}$ m/s, estimated

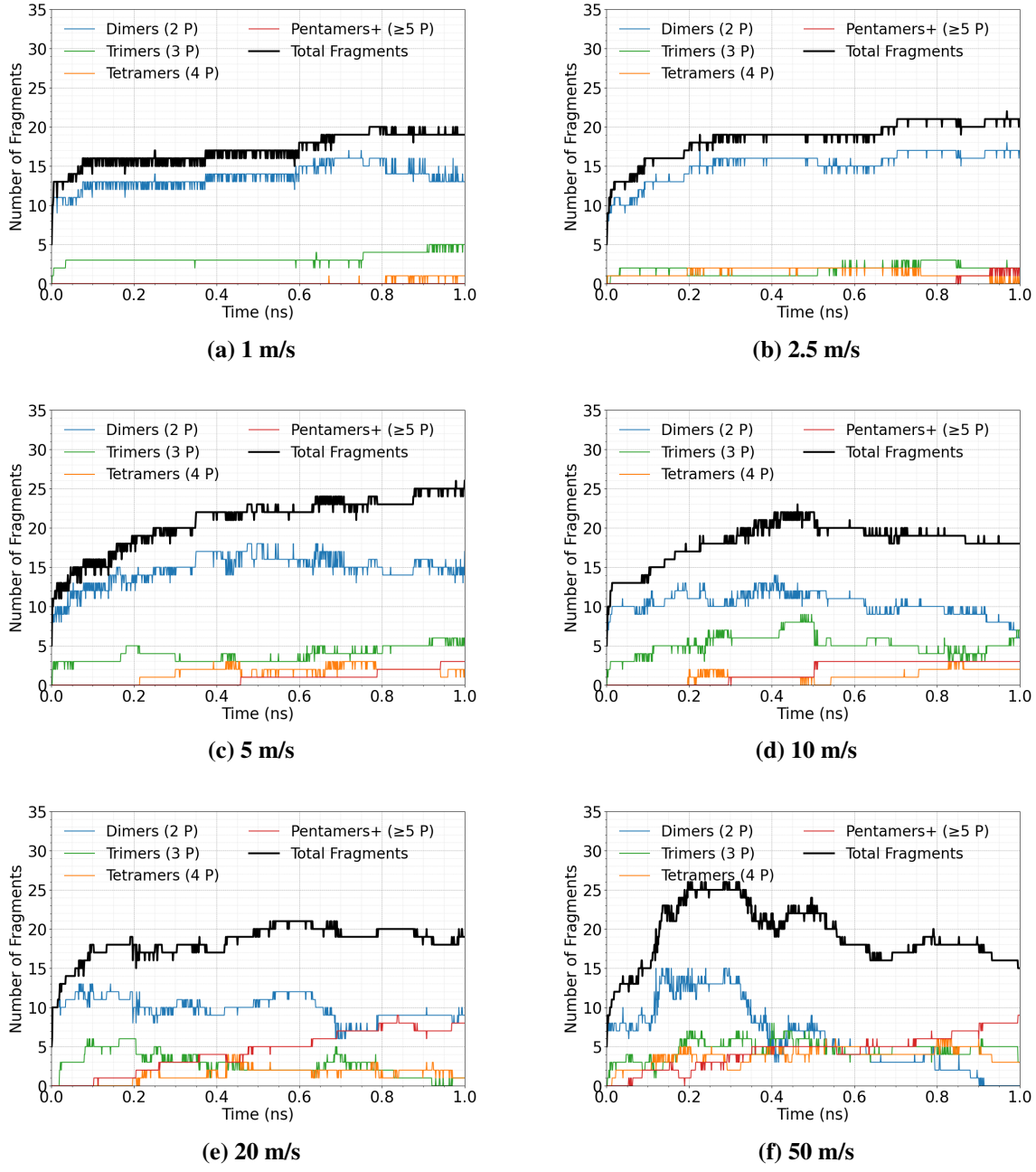
the activation energy for TCP decomposition on an iron substrate as 6.3 kJ/mol.[354] The ball was made from aluminium and the disk made from germanium and coated with a 20 mm thick iron layer.[354] Using Hertzian contact theory,[446] with estimated values for the elastic modulus and Poisson ratios for aluminium and iron,[447,448] we estimate that the normal load present in this system is approximately 0.7 GPa, while the estimated maximum shear stress is 0.22 GPa.

Comparison of NEMD results to experimental data remains challenging due to the significant differences in spatial and temporal scales between simulations and experiments.[16,286] To facilitate direct comparison between experimental activation energy calculated by Sasaki et al.[354] and our simulations, we perform a fit of the logarithm of the activation energy to the sliding speed, as illustrated by Fig. 4.6c and Fig. 4.6d for the α-Fe(110) and Fe_2O_3 systems respectively. At the relevant speed ($5x10^{-5}$ m/s), this extrapolation predicts an activation energy of approximately 50 kJ/mol for the DLC-α-Fe(110) and 37.5 kJ/mol for the DLC-Fe_2O_3 systems. While the extrapolated value for the DLC-α-Fe(110) shows a near 10-fold over-prediction compared to the reported experimental values, the experimental prediction is only based on two temperatures, meaning that there could be a considerable amount of uncertainty in the reported value. Furthermore, the chemical inertness of the DLC surface may reduce the amount of TCP dissociation compared to the aluminium-iron interface in the experiments, causing the higher activation energies. Second-order effects can may explain why linear extrapolation of measured reaction rates/activation energies to the nearer zero-stress limit fails to reproduce the experimentally observed activation energy of the reaction.[16] However, such extrapolations require large assumptions and introduce significant uncertainty in estimating the near zero-stress rate, such as in our case where we extrapolate across five orders of magnitude in sliding velocity. Therefore, it is often difficult to definitively determine whether second-order effects are present.[16]

4.3.4 Sliding Speed Effects on Polyphosphate Formation and TCP Bond Lengths

Polyphosphate Growth

Fig. 4.7 shows the time-dependent evolution of polyphosphate chain length at selected sliding speeds at 5 GPa for the DLC-Fe_2O_3 systems, while Fig. B.4 shows the same for the DLC-α-Fe(110) systems. As shown in the figures, both systems exhibit an increase in the number and length of formed polyphosphate chains with increasing sliding speed. For example, at 1 m/s in the DLC-α-Fe(110) system, no pentamers (polyphosphate chains with at least 5 P atoms present) are formed, whereas as many as 9 pentamers (or polyphosphate chains of even greater length) are formed in the 50 m/s simulations. At lower speeds, polyphosphates comprising 2-3 monomers are most prevalent, similar to a previous study finding that tribofilms formed by ZDPP possess an average chain length of 2.5 phosphate groups.[351] It should noted that TCP forms thinner tribofilms than ZDDP and takes longer to grow.[290] However, at higher speeds (>20 m/s), for both surfaces, polyphosphates with 5 or more P atoms start to become more prevalent, with polyphosphates containing up to 18 P atoms forming at the highest speed in the DLC-α-Fe(110) system. The formation of longer polyphosphate chains at higher sliding speeds suggests that the increased shear stress at higher sliding speeds, and the associated greater dissociation of the TCP molecules, results in a higher concentration of reactive fragments that can participate in tribofilm formation. Additionally,



(a) 1 m/s **(b) 2.5 m/s**

(c) 5 m/s **(d) 10 m/s**

(e) 20 m/s **(f) 50 m/s**

Figure 4.7: Time evolution of polyphosphate formation at 500 K and 5GPa, for various sliding speeds for DLC-Fe_2O_3 system.

the higher sliding speeds may facilitate greater mobility and transport of these reactive species within the system, enabling the formation of longer polyphosphate chains. This is supported by the observed increase in polyphosphate chain length in both the DLC–Fe_2O_3 and DLC–α-Fe(110) systems, despite the relatively small increases in measured shear stresses at speeds above 10 m/s (Fig. 4.3).

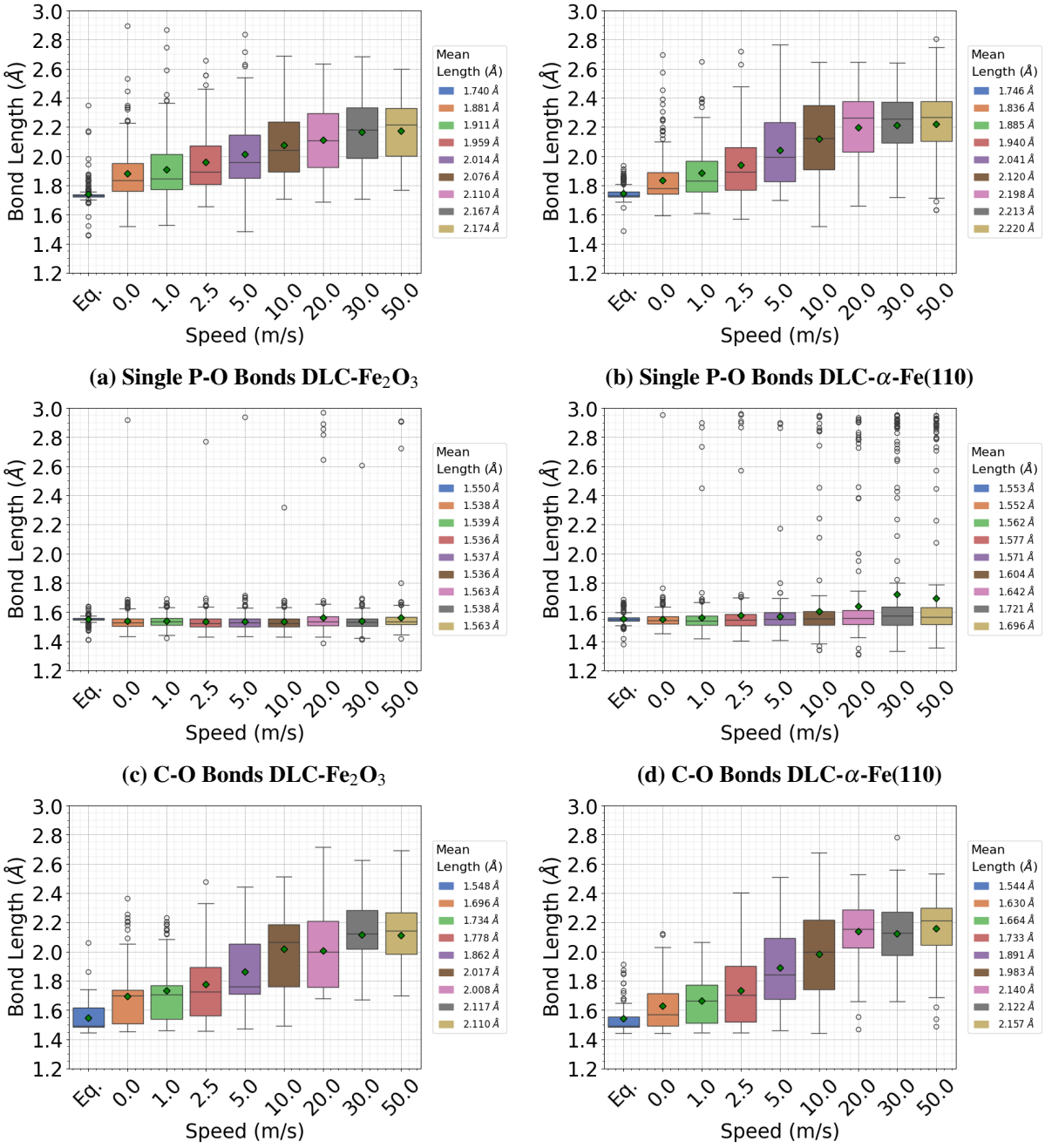
Similar to the observed polyphosphate film formation trends observed in Chapter 3, most sliding speed conditions show a consistent pattern in which the total number of polyphosphate chains stabilises be-

tween 18 and 25. However, at the highest sliding speeds, the number of polyphosphate chains present in the later stages of the simulation counterintuitively decreases, such as at the 50 m/s condition for the DLC-α-Fe(110) system, where the total number of polyphosphates drops below 15. This behaviour suggests that the formation of polyphosphate films in these simulations could be constrained by the finite supply of TCP molecules available for reaction rather than by any inherent chemical limitation on chain growth. Notably, the DLC–α-Fe(110) system consistently forms longer polyphosphate chains more quickly than the DLC–Fe_2O_3 systems at comparable sliding speeds. This quicker formation of longer polyphosphate chains in the DLC–α-Fe(110) systems can be attributed to the greater chemical reactivity of the α-Fe(110) surface, where iron atoms are more readily available to catalyse tribofilm formation.[287,449] These results are consistent with previous studies which have shown that ferrous surfaces promote the formation of iron-containing phosphate or polyphosphate films, while such films do not form as freely on inert substrates such as SiC or Si_3N_4, highlighting the essential catalytic role of the metal surface.[450,451]

TCP Bond Length Distributions

Two primary mechanisms have been proposed to explain the dissociation of TCP and the subsequent formation of anti-wear tribofilms.[287,349] In the first mechanism, the anti-wear performance is attributed to the generation of phosphoric acid following the cleavage of the C-O bonds connecting the phosphate core to the aryl (tolyl) groups, resulting in the formation of iron phosphate films.[293,295] Several variations of this mechanism have been proposed. In one case, the tolyl groups adsorb onto the iron surface before being released as toluene, leaving reactive phosphate species behind. In another, the tolyl groups are removed directly without surface adsorption. A third variation suggests that hydrogen atoms released during toluene formation remain on the surface. Finally, some pathways propose that the aryl groups are retained entirely, resulting in a phosphate precursor still fully substituted with tolyl groups.[349] The second mechanism involves thermal decomposition through the cleavage of P-O bonds within the phosphate core of TCP. This pathway results in the formation of methylphenoxy intermediates and the subsequent release of cresol. One variant proposes that cresol is released sequentially after each P-O bond is broken, while another suggests that all P-O bonds cleave first, followed by a collective release of the cresol groups.[296,300] In both cases, the phosphorus and oxygen residues facilitate the growth of a protective phosphate layer on the surface. For both mechanisms, the cleavage of either the C-O or P-O bonds is preceded by bond elongation as this stretching weakens the bond and lowers the activation energy required for dissociation,[452] making these bonds more reactive, thus serving as a key initial step in the decomposition of TCP.[453] To investigate the effect of applied stress on bond elongation, we measure the variation in bond length of the P-O single bonds, C-O single bonds, and P-O double bonds across the studied sliding speeds and at different normal pressures. The bond length distributions from the equilibration phase (at 300 K and 10 MPa normal load) are also included as a benchmark.

Fig. 4.8 shows a box plot comparison of bond length distributions across all the TCP molecules as a function of sliding speed for the DLC-Fe_2O_3 and DLC-α-Fe(110) systems for the P-O single bonds, C-O single bonds, and P-O double bonds at 500 K and 5 GPa normal load. Fig. B.5 shows a box plot comparison of the same bond distributions as a function of different normal pressures at 500 K and a



(a) Single P-O Bonds DLC-Fe_2O_3

(b) Single P-O Bonds DLC-α-Fe(110)

(c) C-O Bonds DLC-Fe_2O_3

(d) C-O Bonds DLC-α-Fe(110)

(e) Double P-O Bonds DLC-Fe_2O_3

(f) Double P-O Bonds DLC-α-Fe(110)

Figure 4.8: Comparison of bond length distributions as a function of sliding speed between DLC-Fe_2O_3 and DLC-α-Fe(110) systems for selected bond types. Mean length shown by green diamonds.

constant sliding speed of 10 m/s. These bond length values are calculated by taking the average length of the relevant bond(s) from each TCP molecule 50 ps before that bond is permanently cleaved (i.e., remains broken for greater than 50 ps after the initial break is detected), or if the relevant bond(s) do not break during the production run, the average length of the relevant bond(s) 50 ps before the end of the production run.[418] The average bond lengths of the P-O single bonds, C-O single bonds, and P-O double in the equilibrium phase for both systems are approximately 1.74 Å, 1.55 Å and 1.55 Å

respectively. Previous experiments have estimated C-O bond length to generally be 1.43 Å,[454] but can be expected to be in the region of 1.50 Å for C-O bonds in organophosphates like TCP.[454] Other studies have observed a bond length of 1.52 Å for P-O double bonds and 1.66 Å for P-O single bonds.[455,456] Thus, while the average bond lengths for the C-O single bonds and P-O double bonds are consistent with literature, the average lengths of the P-O single bonds seem to be overestimated. As these bond lengths are measured at 500 K, and it is known that elevated temperatures lead to greater bond lengths as the amplitude of atomic vibration at the equilibrium position increases,[457] this overestimation could be due to the elevated temperatures in the simulations compared to the temperatures that these experimental bond lengths are reported for.

As shown by Fig. 4.8, we observe significant increases in average bond lengths for both the P-O single and double bonds increase with increasing sliding speed, in both the DLC-Fe_2O_3 and DLC-α-Fe(110) systems, while the average bond-lengths of the C-O single bonds stay relatively unchanged. Notably, at 0 m/s (5 GPa normal load) there is an increase in the average bond length of the P-O single and double bonds compared to the equilibrium condition, counterintuitively suggesting that the application of compressive force alone can cause some bond stretching. However, as shown by Fig. B.5, varying the normal load does not lead to a significant increase in average bond lengths even at normal pressures of 10 GPa. There is some stretching of individual P-O single bonds in both systems in the absence of applied sliding, likely due to localised effects such as molecular reorientation due to uneven stress distributions in the bulk. The greater elongation of P–O bonds compared to C–O bonds provides further evidence that TCP dissociation is primarily driven by P–O bond cleavage, consistent with our prior study.[271] The observed larger elongation in P–O bonds compared to C–O bonds can be attributed to differences in bond energy, orbital overlap, and delocalisation. Previous studies have found that the P–O single bond has an average bond energy of 335 kJ/mol, the P-O double bond an energy of 544 kJ/mol, and the C–O single bond an energy of 358 kJ/mol.[458,459,460] Although the P-O double bond is the strongest among them, several factors make P-O bonds in the phosphate group susceptible to elongation. One major factor is the orbital-size mismatch as phosphorus possesses diffuse $3p$ orbitals, while oxygen possesses compact $2p$ orbitals, resulting in poor overlap and a bond that is more polarisable and flexible.[461,462] In contrast, the C-O bond benefits from $2p$-$2p$ overlap and is therefore less polarisable and more resistant to elongation, despite only a modest difference in bond energy compared to the P-O single bond. Additionally, in phosphate esters like TCP, the electronic structure is best described as a resonance hybrid involving delocalisation of π-electrons over multiple oxygen atoms, implying that no P-O bond is purely double or purely single.[463] As a result, partial double-bond character is distributed across all P-O bonds, and this delocalisation allows even the formally strongest P-O double bond to stretch significantly under thermal or mechanical stress. In ReaxFF simulations, where one P-O bond is explicitly designated as a double bond, it is important to recognise that this assignment is an approximation of a more complex electronic structure. Elongation of this labelled double bond is therefore not inconsistent with its higher bond energy.

4.4 Conclusions

We have investigated the mechanochemical decomposition of TCP molecules confined between DLC-α-Fe(110) and DLC-Fe_2O_3 interfaces using NEMD simulations with ReaxFF. The decomposition kinetics were quantified across a wide range of sliding speeds and temperatures under a constant normal pressure of 5 GPa. Four kinetic models were applied to the MD data, revealing that TCP decomposition transitions from second-order to first-order/exponential behaviour with increasing sliding speed, indicating a shift from cooperative to unimolecular shear-driven mechanisms. Exponential fits best described high-speed dissociation, suggesting a stochastic process governed by constant bond-breaking probability.

Shear stress was found to increase with sliding speed until saturation above 20 m/s, consistent with interfacial reorganisation. Activation volumes extracted from pressure-dependent simulations at 10 m/s revealed temperature-dependent increases and were consistently higher for α-Fe(110), implying more effective stress transfer on this more reactive surface. Rate constants derived from the simulations enabled estimation of Bell and Arrhenius parameters. The Bell activation energies decreased with increasing sliding speed, confirming shear-assisted thermal activation (SATA) as the dominant mechanism. The extrapolated activation energies were comparable to those reported experimentally for iron-based surfaces under low shear conditions.

Polyphosphate chain growth was more extensive and rapid on α-Fe(110) surfaces compared to α-Fe_2O_3, with chain lengths reaching up to 18 P atoms at 50 m/s, likely driven by increased molecular fragmentation and enhanced mobility of reactive intermediates at high speeds. Bond-length analysis revealed that P–O bonds experienced greater elongation under shear than C–O bonds, consistent with preferential P–O bond cleavage in TCP dissociation. These trends were more pronounced on iron due to stronger adsorption and surface interaction.

Overall, the simulations show that TCP decomposition is strongly dependent on shear stress, with surface chemistry influencing both kinetics and tribofilm formation. The results further underscore the relevance of surface reactivity for tribofilm growth and highlight the importance of shear-induced effects in mechanochemical additive performance.

Similar to the findings of Chapter 3, these results support the overall research goal by further developing our understanding of the stress-based mechanisms of additive decomposition across realistic tribological conditions, again providing a foundation for future predictive modelling of additive behaviour in different industrially-relevant environments.

Chapter 5

A Comparison of Molecular Dynamics and Machine Learning Techniques for The Prediction of Lubricant Transport Properties

In this chapter, the effectiveness of machine learning (ML) and molecular dynamics (MD) approaches for predicting key transport properties of hydrocarbons relevant to lubricant base oil design is assessed. Different ML models and molecular representations are benchmarked, and their predictions compared to MD simulations using a range of force fields, with the aim of identifying reliable and efficient methods for estimating properties such as viscosity, density, and thermal conductivity. ML model generalisation to novel chemical structures outside the training distribution is evaluated, and the limitations of using MD-generated data to train ML models is explored. This work contributes to the broader goal of enabling data-driven design and screening of new base oil candidates for advanced lubricant formulations. The insights are agnostic of chemistry, meaning the methods and insights of this study could also be applied to the characterisation of lubricant additives, such as those studied in Chapters 3 & 4.

5.1 Introduction

The accurate prediction of the thermophysical properties of hydrocarbons such as the density, viscosity, and thermal conductivity are important for a wide range of industries, including oil and gas, chemicals, and pharmaceuticals.[464,465] Here, the lack of reliable data is one of the largest stumbling blocks in rational molecular, formulation, and process design.[29,464,465,466] For example, lubricants are required to adhere to strict viscosity limits across a wide temperature range,[467] however, for the hydrocarbon molecules of relevance to automotive and aerospace applications (C_{10}-C_{50}),[468] viscosity-temperature data is still lacking. Viscosity data are available for many normal alkanes up to $\sim C_{24}$, but data availability for branched alkanes, cyclic hydrocarbons, aromatics, alkylated saturated cyclic hydrocarbons, and alky-

lated aromatics is much more sparse.[469] It is often possible to interpolate between missing entries with calculated predictions using a reliable equation of state, such as the perturbed chain-statistical associating fluid theory (PC-SAFT).[470] Several other predictive models are available for pure hydrocarbons,[471] with the most promising being group contribution and particularly entropy scaling methods.[472,473] Quantitative structure-property relationship (QSPR) models[474,475,476,477,478] have also been identified that are capable of predicting the viscosity of organic liquids. The main problem with all of these methods is their low transferability, with predictive performance being poor outside the molecules and conditions for which experimental data are available.[479] Thus, they are not always suitable for extrapolation beyond the experimentally gathered dataset to new molecules or conditions.

Over the last few decades, MD simulations have been established as a useful tool for predicting thermophysical properties and generating datasets.[480,481,482] Unlike the modelling methods above, MD simulations can be used to accurately predict the properties of any organic liquid, provided that the molecular structure is known and a suitable force field is available.[483,484] MD simulations are now routinely used to calculate the shear viscosity of single-component,[105,139,485,486] binary mixtures,[487,488,489,490,491] and multi-component[492,493,494] hydrocarbon fluids with high accuracy over a wide range of conditions. In most cases, however, these simulations have been restricted to small molecules, typically linear or branched alkanes, due to several limiting factors. Accurate MD simulations are computationally intensive, requiring long simulation times and significant computational resources to produce statistically reliable results, as simulating large, flexible, or highly cyclic molecules introduces further complexity due to slow conformational sampling and increased system size.[495] The accuracy of MD predictions also depends heavily on the quality of the force field used, which must be carefully parameterised for the specific class of molecules under study, limiting generalisability.[496]

In recent years, machine learning has also emerged as a useful to tool with which to predict the thermophysical properties of pure organic liquids and their mixtures.[497] Accurate viscosity prediction has been achieved using ML models, such as neural-networks for single-component,[498,499,500,501,502,503,504] binary mixtures,[30] and multicomponent[505] organic liquids. ML methods have also successfully been used to predict the viscosity of ionic liquids[506] and deep eutectic solvents.[507] As with QSPR models, ML model accuracy remains at the mercy of accurate training data availability, with small, sparse or noisy datasets all having adverse effects on model accuracy.[222] Multitask learning offers a potential solution to data scarcity by leveraging shared information across related property prediction tasks.[508] Some studies have combined MD and ML by using dynamic descriptors from the MD simulations to improve ML predictions.[31,240] MD has also been used to generate datasets to train ML models.[509,510] Both MD[511] and ML[512] been considered for this purpose for a long time, but only recent advances in hardware and software have made this type of study feasible.

As with QSPR and MD methods, ML methods also struggle to generalise beyond the domain of the training set. Recent studies have began benchmarking models on their OOD performance in different chemical domains[513] and proposing methods to improve OOD performance, such as synthetic data generation.[514] For example, Omee et al.[515] evaluated multiple GNN models on structure-based OOD materials property prediction tasks taken from the MatBench suite of benchmarking datasets,[516] re-

vealing that no single model achieved consistent state-of-the-art performance across all datasets or OOD scenarios.[515] Other studies have made attempts to classify inputs to deep learning models as OOD or not, in order to determine how much confidence should be placed in ML model outputs.[517,518] Hendrycks and Gimpel[519] proposed a method of detecting classification errors and OOD inputs using softmax prediction probabilities, showing that while softmax outputs are imperfect indicators of model confidence, they can still effectively distinguish between correct and incorrect or in- and out-of-distribution examples.[519] Even with more attention being paid to this problem, model OOD generalisation and characterisation remains an open problem, and there is currently no consensus on reliable strategies to assess or improve model performance when extrapolating to new molecules.

The choice of ML algorithm also plays a significant role in determining model accuracy and generalisability. Several benchmarking studies have assessed the performance of ML models for molecular property prediction.[516] While some benchmark studies have found that message passing neural networks (MPNNs) achieve state-of-the-art performance on property regression tasks for datasets such as QM9 and lipophilicity[240,520], other studies report that descriptor-based models such as support vector machines (SVMs) and XGBoost[237] often outperform graph neural networks for certain molecular property prediction tasks. Notwithstanding the lack of consensus regarding the preferred property prediction architecture, these benchmark datasets are more geared towards drug-like compounds[521,522] and do not contain many lubricant-like hydrocarbons, which typically feature larger, more flexible aliphatic chains and are likely governed by different structure-property relationships. To our knowledge, no studies have benchmarked a range of ML algorithms for the prediction of lubricant-like hydrocarbons of industrial interest in the contexts of OOD performance, sparse experimental data and simulation-derived datasets.

The choice of molecular representation has a direct impact on model performance and reliability. An effective molecular representation should correlate strongly with at least one target property and be capable of distinguishing between isomers. It should reflect local structural features of the molecule and be constructed in a way that is simple, interpretable, and independent of experimental measurements. Furthermore, good representations should also be non-redundant, meaning they do not encode trivial relationships with other descriptors, and should be easy to compute from the molecular structure. In addition, they should vary smoothly when small changes are made to the molecular structure, allowing models to capture continuous structure–property relationships.[523] Additionally, the representation should have uniqueness (each molecular structure relates to a single representation) and invertibility (each representation is associated with one single molecule).[222] Many studies have employed a wide range of molecular representations, including molecular fingerprints,[524] embeddings of molecular structures,[525] 3D molecular graphs,[240] string-based encodings such as SMILES and SELFIES[502] and physicochemical molecular descriptors[31] to train ML models for property prediction. A smaller number of works directly compare some of the different representation methods on the same dataset to evaluate how representation choice impacts model performance, highlighting that the ability of ML models to learn meaningful structure–property relationships is highly dependent on how molecular information is encoded.[237]

In this study, we focus on alkyl cycloalkanes and alkylated aromatic hydrocarbons, as well as model

alkanes and an ester, that are relevant to the development of fuels and lubricants. Such molecules are important components of automotive fuels,[526] aviation fuels,[527] and lubricants.[468,528] We evaluate a range of molecular representations in combination with a variety of ML models to predict key thermophysical properties at industrially relevant conditions, such as viscosity, density, and thermal conductivity, with the goal of determining the best ML-algorithm. We compare the performance of the best ML model to a range of commonly used MD force fields to characterise the predictive accuracy of these respective methods on unseen molecules within the applicability domain of the models. We also examine the utility of synthetic data from MD simulations and assess the generalisation performance of the best performing ML models on OOD prediction tasks. These studies aim to establish practical guidelines for the use of data-driven methods in the prediction of lubricant-like hydrocarbon transport properties.

5.2 Methodology

5.2.1 Choice of benchmarking molecules

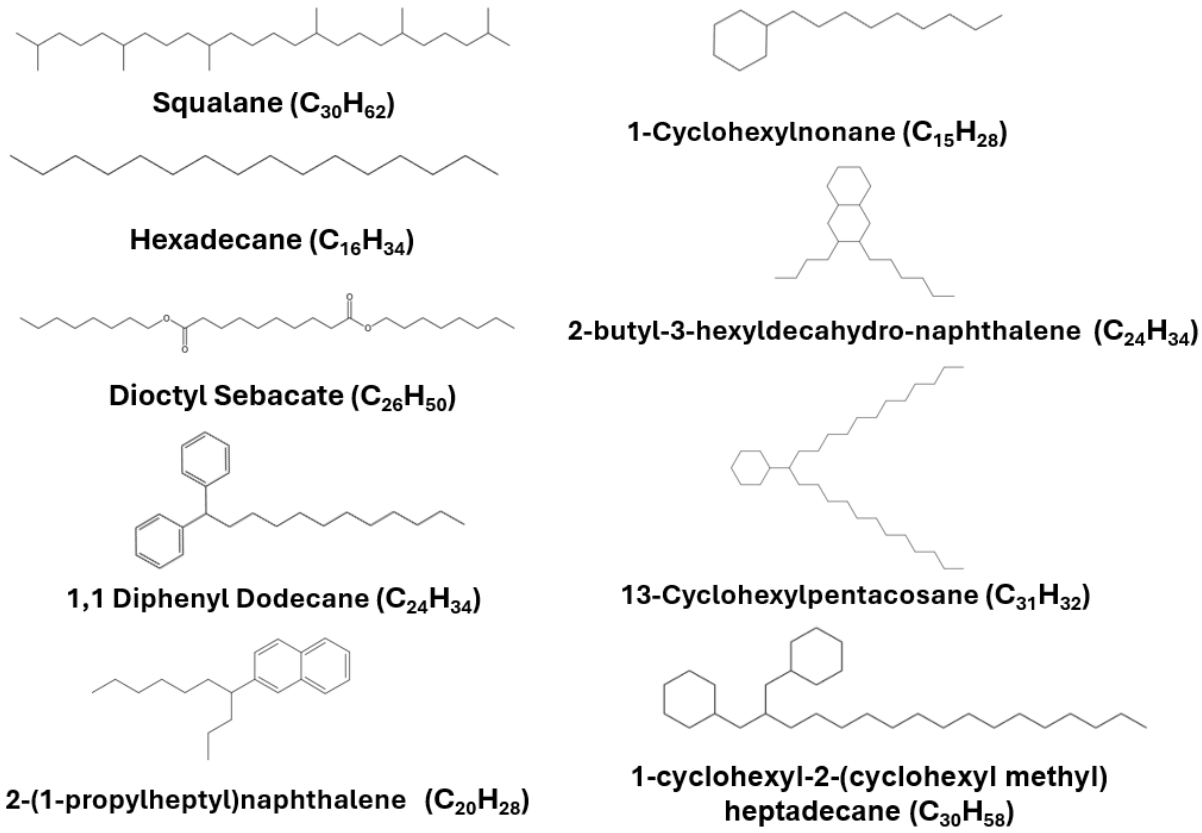


Figure 5.1: Chemical structures of the nine hydrocarbons analysed in this study.

The chemical structures of the nine molecules analysed in this benchmarking study are illustrated in Figure Fig. 5.1. Squalane is frequently used as a model lubricant in experiments [529,530] and also an important substance in cosmetics.[531] Similarly, *n*-hexadecane is frequently used as a model fluid in tribology studies, both experimentally [532] and for MD simulations .[533] Furthermore, hexadecane is often used in stability testing to understand how long-chain hydrocarbons behave under thermal and ox-

idative stress.[534] Dioctyl sebacate has been used as an lubricant base oil in aerospace and automotive applications[535] and due to its strong additive solubility, has also been used as an adjuvant for mineral oils or poly-alpha-olefins (PAOs).[536] Alkylated benzenes such as diphenyl-dodecane have been used as high performance compressor oils, and due to their excellent pour points and superior solvency for anti-wear additives and antioxidants, they offer advantages over PAOs and (gas-to-liquid) GTL oils in certain high-performance applications.[537] Furthermore, their structure can be tailored to achieve desired viscosity characteristics.[538] Cycloalkanes such as 1-cyclohexylnonane and 13-Cyclohexylpentacosane are of interest as they are used for jet engine lubricants and automotive transmission fluids, due to their chemical stability, high oxidation resistance and superior shear resistance.[539] Alkylated naphthalenes such as 2-(1-propylheptyl)napthalene offer excellent thermal and oxidative stability, high load-carrying capacity, and superior additive solvency, making them ideal for use in jet engine oils, high-performance automotive lubricants, industrial gear and compressor oils, and specialty dielectric fluids.[540] Collectively, these molecules include most of the features found in lubricant base oil molecules.

5.2.2 Fluid Transport Property Calculation using Equilibrium Molecular Dynamics

Viscosity

EMD simulations were used to calculate the density, viscosity, and thermal conductivity of the nine benchmarking molecules. EMD is most effective for fluids with relatively low viscosity, typically less than 20 mPa.s, although it has been successfully implemented for systems with viscosities approaching 50 mPa.s.[114] The best practice study by Maginn et al.[114] for computing viscosity with EMD recommends using the Green-Kubo formulation, whereas Hess et al.[541] state that the Einstein relation is more convenient due to its ability to ignore inaccuracies in long-time correlations by considering integrals over shorter times. The Green-Kubo formalism for calculating viscosity is given by :[115,116]

$$\eta = \frac{V}{k_B T} \int_0^\infty \langle \sigma_{\alpha\beta}(t_0)\sigma_{\alpha\beta}(t+t_0)\rangle_{t_0}\, dt \tag{5.1}$$

where η represents the viscosity, V is the system volume, k_B is the Boltzmann constant, T denotes the temperature, $\sigma_{\alpha\beta}(t)$ stands for the off-diagonal component of the stress tensor at time t (with α and β representing the x, y, or z direction in Cartesian coordinates), and $\langle \sigma_{\alpha\beta}(t_0)\sigma_{\alpha\beta}(t+t_0)\rangle$ denoting the ACF for the off-diagonal components of the stress tensor.

The Einstein (or Einstein-Helfand[192]) formalism provides an alternative approach for computing shear viscosity from EMD simulations. Rather than integrating the stress ACF, this approach tracks the mean square displacement (MSD) of a time-integrated stress tensor, offering robustness against noise of the long-term ACF.[541] The Einstein relation for shear viscosity is given by:

$$\eta = \lim_{t\to\infty} \frac{V}{2k_B T} \frac{d}{dt} \left\langle \left[\int_0^t \sigma_{\alpha\beta}(t')\, dt' \right]^2 \right\rangle \tag{5.2}$$

where η is the viscosity, V is the system volume, k_B is the Boltzmann constant, T is the temperature, and $\sigma_{\alpha\beta}(t')$ is the off-diagonal component of the stress tensor at time t'.

Thermal Conductivity

The Green-Kubo formalism is also applicable to the calculation of thermal conductivity κ, relying on the autocorrelation of the off-diagonal components of the bulk heat flux vector $\mathbf{J}_q$. This is expressed as:[115,116]

$$\kappa = \frac{1}{Vk_BT^2}\int_0^\infty \langle \mathbf{J}_q(t_0)\cdot\mathbf{J}_q(t+t_0)\rangle_{t_0}\,dt \tag{5.3}$$

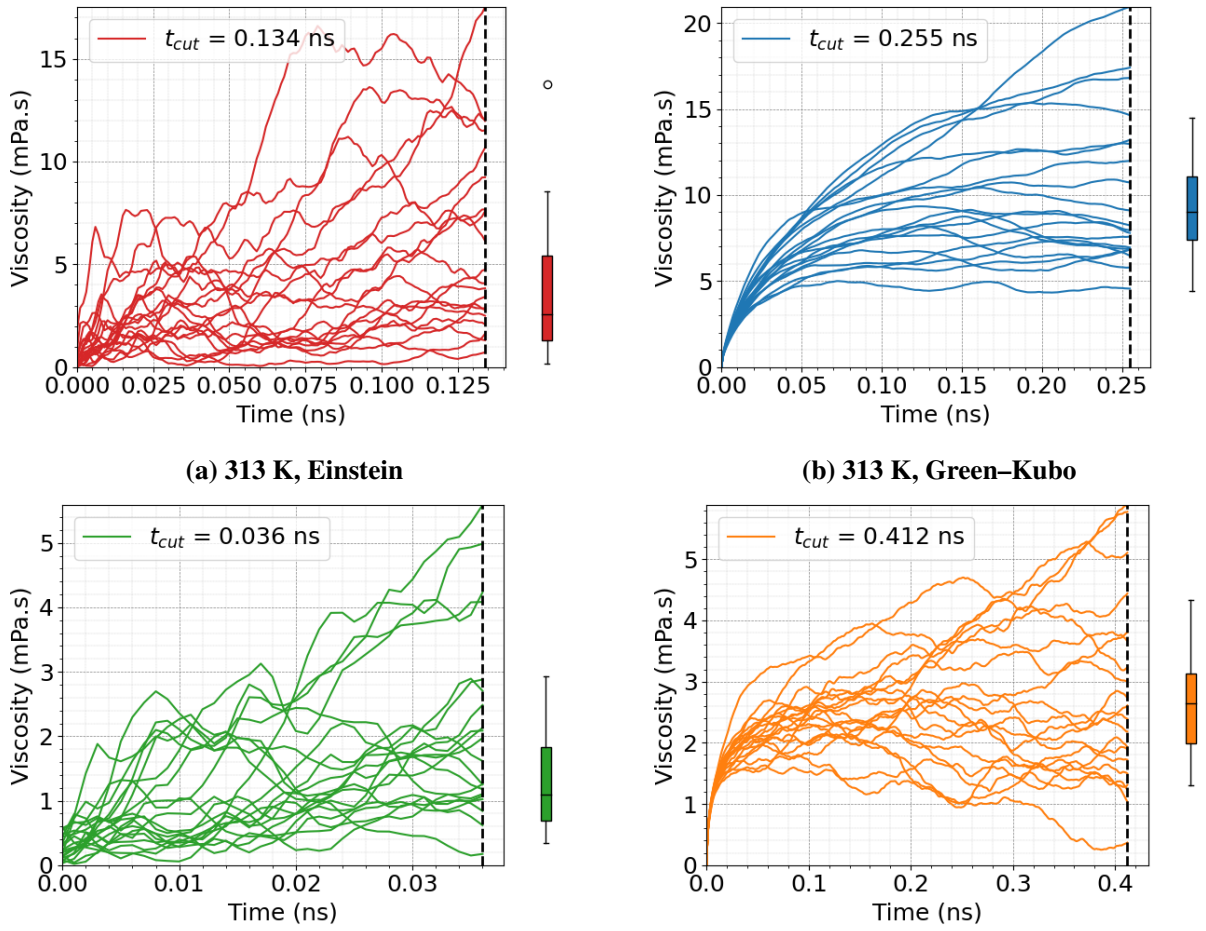
where κ is the thermal conductivity, $\mathbf{J}_q$ is the heat flux vector, and the angle brackets indicate a time-origin average of the ACF. Alternatively, thermal conductivity can be computed using the Einstein-Helfand approach.[192] The corresponding expression is:[542]

$$\kappa = \lim_{t\to\infty}\frac{1}{2Vk_BT^2}\frac{d}{dt}\left\langle\left|\int_0^t \mathbf{J}_q(t')\,dt'\right|^2\right\rangle \tag{5.4}$$

This method provides a diffusion-like representation of energy transport, where the time derivative of the MSD of the integrated heat flux yields the thermal conductivity.

A study by Zhang et al.[543] determined that incorrect viscosity values will always be computed from trajectories that are too short, regardless of the number of replicates used. Specifically, they found that viscosity values calculated from trajectories of 500 ps diverged from those calculated from trajectories of 1, 2, and 4 ns.[543] When studying the shear viscosity of alkanes, Payal et al. were able to compute accurate values using 20 independent trajectories.[544] Hence, to ensure adequate sampling time for each molecule, each molecule is simulated for 20 independent trajectories of 2 ns. Finite size effects may occur in small, dense systems as the restricted space limits configurational rearrangements.[545] However, multiple studies suggest that finite size effects are minimal when calculating viscosity from systems containing between 125 and 500 molecules,[543,546] while Daivis and Evans found that shear viscosity exhibits a weak dependence on system size.[547] To ensure a large enough system size, for each molecule, a cubic model with 3D PBCs containing around 15,000 atoms is created with Packmol.[548] To eliminate the likelihood of unphysical overlapping of molecules, Packmol insertion was conducted at low density, as has been done in previous benchmarking studies.[105] To enable system creation with Packmol and Moltemplate,[549] 3D structures of each candidate molecule were generated from their 2D SMILES representations, using the Experimental-Torsion Knowledge Distance Geometry (ETKDG) algorithm (version 3).[550] Datafiles are created using Moltemplate.[549]

All MD simulations were performed using LAMMPS.[323] The following protocol was used in the EMD simulations as it has been shown to enable successful calculation of lubricant viscosity.[31] The velocity-Verlet algorithm was used with an integration time-step of 1.0 fs.[121] Interactions between unlike atoms were evaluated using geometric mixing rules, and long range tail corrections to energy and pressure were included. A cutoff distance of 12Å was used for Lennard-Jones interactions. Long range electrostatic interactions were evaluated using the particle-particle particle-mesh (PPPM) method.[146] Energy minimization of the system was performed using the conjugate gradient method[551] to prevent large forces in the system at the start of the simulation. Following minimization, the system density was equilibrated at 1.0 atm and the target temperature for 1.0 ns in the NPT ensemble, utilising the Nosé–Hoover thermostat



(a) 313 K, Einstein

(b) 313 K, Green–Kubo

(c) 373 K, Einstein

(d) 373 K, Green–Kubo

Figure 5.2: Time-dependent viscosity traces for Squalane at 313 K and 373 K with the Einstein and Green-Kubo methods, using the SCIPCFF forcefield.

and barostat[125,131] with damping coefficients of 100 fs and 1000 fs respectively. The average density from the last 20 timesteps of this ensemble was taken to be the predicted density of the molecule. Subsequently, the simulation box was deformed for 0.5 ns in the NVT ensemble while maintaining a constant temperature until the fluid density matched the average density from the previous NPT simulations. Finally, production runs of 2 ns were carried out in the NVT ensemble for each molecule.

Time Decomposition Method

The time decomposition method (TDM)[543] was used to improve the reliability of calculated high viscosity values from the EMD simulations. The time evolution of the viscosity, $\eta(t)$, was recorded for the full duration of simulation production runs in the NVT ensemble. For the calculation of the autocorrelation function (ACF), a fast Fourier transform (FFT) technique was used.[552] Contributions from all six independent components of the pressure tensor $\sigma_{\alpha\beta}$ were averaged to determine the viscosity, as suggested by Maginn et al.[114] The cutoff time t_{cut} for the ACFs, introduced to mitigate long-term noise,

was determined from the ratio of standard deviation $\sigma_\eta(t)$ to average final viscosity from the trajectories with the following criterion:[543]

$$\frac{\sigma_\eta}{\eta_{final}} = 0.4$$

matching the criterion used by Schmitt et al.[105], who showed that ratios of σ_η/η_{final} between 0.2 and 0.8 do not materially impact the extracted viscosity values. The bootstrapping method,[543] with 200 iterations, was used to compute the final viscosity value and its uncertainty via the PyLatt software.[553] In this method, 200 combinations of the 20 simulations were generated randomly, allowing multiple occurrences of individual simulations. For each combination, an average viscosity $\bar{\eta}(t)$ was calculated, and the following function (Eq. 5.5) was fitted to $\bar{\eta}(t)$, where η_∞ represents the long-term viscosity limit and β_1 and β_2 are fitting parameters:

$$\eta(t) = \frac{\eta_\infty \left[\alpha\beta_1\left(1-e^{-t/\beta_1}\right) + (1-\alpha)\beta_2\left(1-e^{-t/\beta_2}\right)\right]}{\alpha\beta_1 + (1-\alpha)\beta_2} \tag{5.5}$$

This method is illustrated for the calculation of viscosity at 313 K and 373 K with the SciPCFF force field,[150,151] for both the Einstein and Green-Kubo correlations, by Fig. 5.2. A similar approach was used to calculate thermal conductivity values.

5.2.3 ML Model Datasets

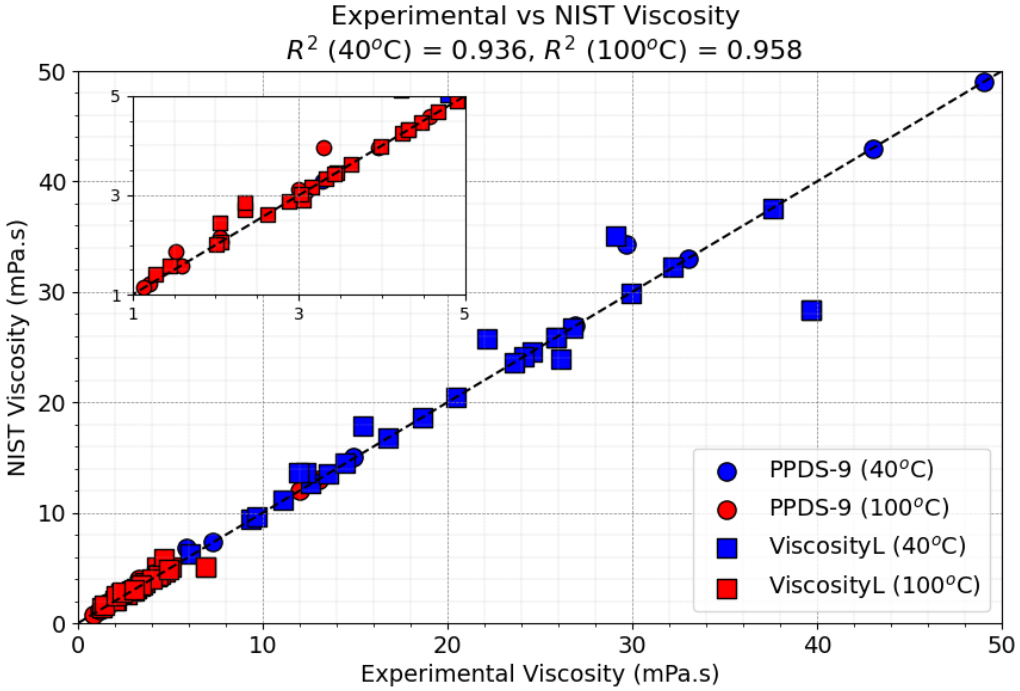


Figure 5.3: Parity plot of experimental viscosity vs NIST Webbook values. Inset shows zoom in of plot between 0 and 5 mPa.s.

To construct a robust and diverse dataset for model development and evaluation, we collected data for density, thermal conductivity, heat capacity, and dynamic viscosity at atmospheric pressure at both 40

°C and 100 °C for approximately 1,200 unique molecules. As EMD has been successfully implemented for systems with viscosities between 1 and 50 mPa.s,[114] which roughly corresponds to the range of viscosities desired for most aerospace and automotive applications, the dataset exclusively of molecules with viscosities falling within this range and both temperatures.[47] This dataset was compiled by integrating entries from peer-reviewed literature, the Landolt-Börnstein series,[554] and the NIST Chemistry WebBook,[555] resulting in a total of over 6,500 experimental data points. To benchmark the performance of machine learning models trained on viscosity values generated via molecular dynamics simulations against models trained directly on experimental values, we also used a processed dataset curated from the work of Kajita et al.,[29] who used rapid MD simulations in conjunction with the Einstein method to calculate viscosity values at 40 °C and 100 °C. The dataset from the Kajita study comprises over 45,000 lubricant-like molecules.

Chemical Family	**Experimental Dataset**	**Kajita Dataset**
Alkanes	160	0
Aromatics	281	30,939
Alkenes	50	461
Naphthalenes	106	0
Cycloalkanes and Cycloalkenes	140	1
Aromatic Esters	92	7,244
Esters and Ethers	359	13,031

Table 5.1: Comparison of molecular family counts between the collated dataset of experimental values and the Kajita dataset.

Table 5.1 shows a comparison of chemical family counts between the experimental dataset and the MD simulation-based Kajita dataset.[29] The counts for the experimental dataset were done manually. The counts for the Kajita dataset were estimated using substructure pattern matching and structural property checks applied to the SMILES representations of molecules. Molecules may belong to more than one chemical family due to the presence of multiple functional groups. Additionally, some molecules may not have been classified due to their unique or atypical structures, a common occurrence in molecular discovery studies where chemical diversity is high.

The NIST WebBook utilises a combination of property estimation methods to interpolate liquid viscosity values from experimental data, including PPDS9 (Physical Property Data Service) and ViscosityL correlations, which are based on semi-empirical and theoretical equations of state fitted from experimental values.[556] PPDS9 (or extended Andrade model[557,558]) offers polynomial and exponential fits derived from experimental measurements and ViscosityL (another extension of the Andrade model[557]) is a legacy correlation system used in NIST data systems. For liquid density values, the NIST WebBook uses the Tait equation.[559] For liquid thermal conductivity, the TransportPolynomial model was applied, utilising temperature-dependent polynomial fits to experimental thermal conductivity data. For heat capacity, a combination of the PolynomialHC,[560] Wilhoit[561] and REFPROP[562,563] models are used.

From the Landolt-Börnstein series,[564] in cases where values were not provided exactly at 40 °C or 100 °C, cubic spline interpolation using a minimum of three surrounding data points was employed to

estimate the desired values. Where only kinematic viscosity was reported, it was converted to dynamic viscosity using the relation:

$$\mu = \nu \cdot \rho$$

where μ is dynamic viscosity, ν is kinematic viscosity, and ρ is density. To assess consistency between data sources, a parity plot was generated from 5% of the viscosity dataset at both temperatures, comparing experimental viscosity values from Landolt-Börnstein to values from the NIST WebBook. Fig. 5.3 shows the parity plot of experimental viscosity and NIST Webbook Values. As shown by Fig. 5.3, the resulting coefficient of determination (R^2) was 0.936 for 40 °C data and 0.958 for 100 °C data, demonstrating that the NIST property relations yield values that are in excellent agreement with high-quality experimental sources. Other studies have benchmarked the accuracy of the NIST WebBook for density and thermal conductivity for a simple liquid, finding that the data has good agreement with the literature.[565] Viscosity and heat capacity values were normalised using the natural logarithm to improve regression performance during ML training.

5.2.4 Molecular Representations

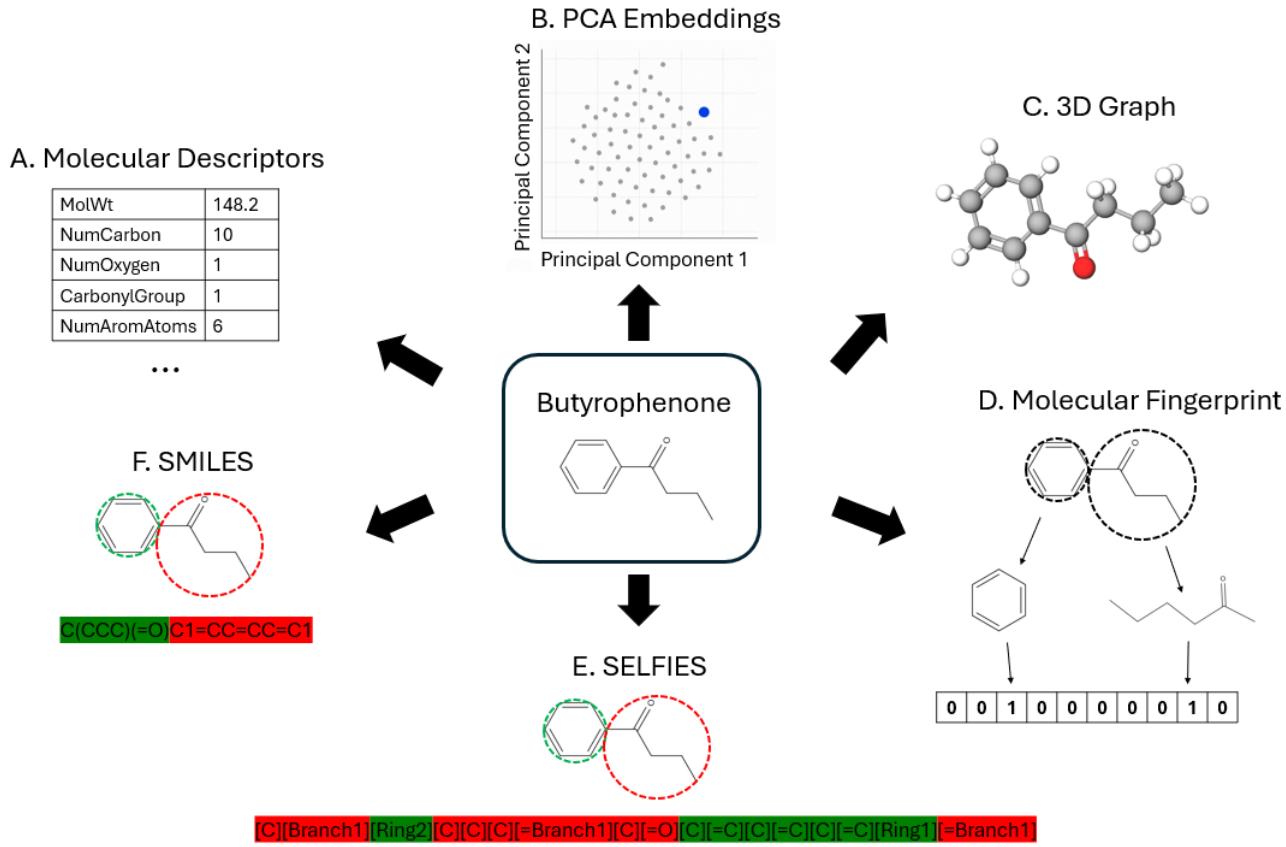


Figure 5.4: Different molecular representations used to train ML models in this study, implemented for example molecule Butyrophenone.

In this study, we compare the most commonly used molecular representations on our experimental dataset to see which representation allows ML models to perform best on transport property predic-

tion tasks. Figure 5.4 shows the different molecular representations used to train ML models in this study.

Molecular Fingerprints

Molecular fingerprints are normally presented in the form of a 2D binary string, where each value in the string represents a Boolean indicator of the presence or absence of a structural feature or substructure within the molecule. Each bit position corresponds to a predefined chemical fragment, such as a functional group, ring system, or atom environment, depending on the fingerprinting scheme. Common schemes for generating such fingerprints include MACCS keys,[566] which consist of 166 predefined structural keys, and the PubChem fingerprint, which encodes 881 substructures based on the PubChem subgraph definitions.[567] Another widely used and generally more performant method is the extended-connectivity fingerprint (ECFP), a famous example being the Morgan fingerprint.[568] ECFPs are generated by iteratively hashing atom-centered circular neighbourhoods up to a specified radius, encoding local substructures based on atomic properties and their connectivity to produce a fixed-length binary or count vector that captures molecular topology.

ECFPs have been shown to outperform other fingerprint types in molecular property prediction tasks and also enable ML interpretability through feature attribution methods such as SHAP.[569] A previous study has investigated the optimal parameters for generating ECFPs in the context of molecular similarity assessment,[570] but to our knowledge, no work has focused on optimizing ECFP generation specifically for property prediction of hydrocarbons using ML models. Hence, this study assesses a range of ECFP generation parameters to determine which allows ML models to perform best on regression tasks. In this study, ECFP fingerprints are generated using the Morgan Fingerprint generator implemented in RDKit.[571]

Dimensionality Reduced Embeddings

Dimensionality-reduced embeddings have emerged as a powerful approach for representing molecular structures, offering a compact and information-rich alternative to high-dimensional descriptors or fingerprints.[572] These methods reduce noise and redundancy in the data, often improving model performance and training efficiency by focusing on the most informative components.[572] Although PCA is commonly used in regression workflows, UMAP's non-linear dimensionality reduction may offer advantages when feature–target relationships are inherently nonlinear, as is often the case in molecular property prediction. Furthermore, the number of embedding dimensions retained can significantly impact model performance, as overly aggressive dimensionality reduction may discard relevant information while insufficient reduction may retain noise.

Therefore, we investigate the effect of the number of embedding dimensions on model accuracy to identify an optimal balance between dimensionality and predictive performance. Both PCA and UMAP embeddings are used to train ML models to determine which embedding method is better for ML property prediction. The embeddings were generated using the scikit-learn and UMAP-learn libraries in Python, based on molecular descriptors computed from SMILES strings of the molecules in the dataset.

Prior to applying the dimensionality reduction techniques, the descriptors were scaled using z-score normalisation.

Generating Molecular Graphs

Molecular graphs encode information about molecules by representing atoms as nodes and the bonds between them as edges. Nodes can be characterised by atom-level features such as atomic number, while the edges can incorporate bond-related characteristics such as bond type. Table 5.2 shows the features used to create molecular graphs in this study.

Feature Name	Description
atomic_number	Atomic number of the atom
degree	Number of directly bonded neighbours
valence	Valence of the atom
formal_charge	Formal charge assigned to the atom
aromatic	Boolean flag indicating aromaticity
mmff94_partial_charge	Partial atomic charge from MMFF94
hybridization	Hybridization state (e.g., $sp^3 = 3$)
total_num_h	Total number of hydrogens bonded to the atom
in_ring	Boolean flag for whether atom is in a ring
chiral_tag	Chiral tag indicating stereochemistry
electronegativity	Pauling electronegativity value
num_neighbour_C	Number of directly bonded carbon atoms
num_neighbour_O	Number of directly bonded oxygen atoms
bond_length	Euclidean distance between bonded atoms
is_conjugated	Boolean flag for bond conjugation
bond_type	Type of bond (e.g., single = 1)
bond_order	Order of the bond (e.g., single = 1)
bond_stereo	Bond stereochemistry code
rotatable	Boolean flag for rotatable bond
is_in_ring	Boolean flag for ring membership

Table 5.2: Features included in 3D Molecular Graphs

Character Embeddings for SMILES and SELFIES

SMILES strings are a molecular representation scheme in which information about a molecule is encoded into a sequence of ASCII characters. SMILES strings are not inherently unique due to the ambiguity in choosing the starting atom for string generation, meaning that a molecule with N non-hydrogen atoms can have at least N equivalent SMILES strings. To address this, Canonical SMILES were introduced, which apply a deterministic set of rules to produce a unique representation per molecule.[573] While this solves the problem of uniqueness, studies such as that by Arus-Pous et al.[198] have shown that ML models trained on canonical SMILES may overfit to the syntactic regularities of the format rather than learning the underlying chemical structure. Despite this limitation, SMILES remains a widely used representation due to its simplicity, ease of preprocessing, and compatibility with sequence-based models like CNNs, and Transformers. To overcome some of the structural limitations of SMILES,

the SELFIES (Self-Referencing Embedded Strings) representation was proposed.[574] SELFIES ensures 100% syntactic validity by design, meaning that any arbitrary sequence of SELFIES tokens corresponds to a chemically valid molecule, a property not guaranteed in SMILES.

In this study, both Canonical SMILES and SELFIES are explored as input representations for molecular property prediction, allowing us to evaluate how the choice of string encoding impacts model performance and generalisation. SMILES and SELFIES strings are converted into token embeddings before being used to train the ML models. SMILES strings in this study were generated from IUPAC names using an online IUPAC-to-SMILES converter.[575] SELFIES strings were generated from SMILES strings using the SELFIES API from the associated publication.[574]

Generating Molecular Descriptors

Molecular descriptors are numerical representations that capture specific chemical, physical, topological, or structural properties of molecules, and are widely used as input features in ML models for property prediction. These descriptors can encode both 2D and 3D information, ranging from simple counts of atom types and functional groups to more complex geometric, electronic, and quantum-chemical properties.[571,576]

In this study, we generated molecular descriptors for our experimental dataset using a combination of the Mordred and RDKit Python libraries.[576] Together, these tools yielded a total of 891 descriptors per molecule after filtering out invalid or missing entries and removing categorical data. Molecular descriptors were selectively scaled depending on the requirements of the ML model. Specifically, we applied feature scaling in the form of z-score normalisation to the models sensitive to the magnitude of input values (SVMs and kNNs) to ensure that all descriptors contributed comparably to distance-based or kernel-based computations. We did not scale the descriptors for the random forest (RF) and XGBoost models, as normalisation methods may negatively impact training by distorting the natural spacing and distribution of feature values, which can obscure chemically meaningful thresholds. This transformation can suppress rare but significant features, reducing the model's ability to retain important structure-property relationships for predicting OOD molecule properties.

5.2.5 ML Model Training

In this study, we evaluate a range of models using the collated dataset to identify the most suitable algorithm for property prediction. The model-representation combinations evaluated in this study are shown in Table 5.3. The classic ML models including random forests, k-nearest neighbour, and SVMs are implemented using the scikit-learn package in Python. The graph-based models, message passing neural networks (MPNNs) and graph attention networks (GATs), are implemented with the PyTorch Geometric library. The deep learning models, CNNs, MLPs, transformers, and the finetuned ChemBERTa model where trained using the Tensorflow and Pytorch libraries. XGBoost models were trained using the XGBoost python library.[241] To ensure fair comparison and optimal performance, a grid search was performed across a range of model-specific hyperparameters for each model. Grid search was chosen over Bayesian optimisation as it computationally feasible due to the relatively small size of the exper-

Model	SMILES	Descriptors	Fingerprints	Graphs	SELFIES	Embeddings
k-NN		✓	✓			✓
MLP	✓		✓		✓	✓
XGBoost		✓	✓			✓
SVM		✓	✓			✓
Transformer	✓				✓	✓
MPNN				✓		
Random Forest		✓				✓
GAT				✓		
CNN	✓				✓	
ChemBERTa	✓					✓

Table 5.3: Molecular representation - ML model combinations investigated in this study.

imental dataset. Models were trained using 5-fold cross validation, with a dataset train-validation-test split of 80:10:10.

Support Vector Machines

SVMs are widely used ML models suitable for both classification and regression tasks.[577] They are effective in high-dimensional spaces and can handle both linear and non-linear relationships through the use of kernel functions. The core principle of a SVM is to identify a decision boundary (hyperplane) that maximises the margin between data points of different classes in a classification problem. With support vector regression, the goal is to find a function that fits the data with as small a deviation as possible within a specified margin.[578] Grid search was conducted across the following range of hyperparameters: regularisation parameter $C \in \{0.01, 0.1, 1, 3, 5, 10\}$, epsilon-insensitive loss parameter $\varepsilon \in \{0.001, 0.01, 0.1, 0.2\}$, and kernel types {linear, polynomial, radial basis function (RBF), sigmoid}.

XGBoost

For an overview of the theory behind the XGBoost architecture, please refer to Section 2.5.2. Grid search was conducted across the following range of hyperparameters: number of estimators `n_estimators` $\in \{100, 200, 300, 500, 1000, 2000, 5000\}$, learning rate `lr` $\in \{0.01, 0.05, 0.1, 0.2, 0.5\}$, maximum tree depth `max_depth` $\in \{3, 4, 5, 6, 10, 20, 50, 100\}$, minimum child weight `min_child_weight` $\in \{1, 2, 3, 4, 5, 6\}$, and sampling ratios for both rows `subsample` $\in [0.7, 1.0]$ and columns `colsample_bytree` $\in 0.7, 0.8, 0.9, 1.0$.

Random Forest

Random Forest is a widely used ensemble learning method based on aggregating the predictions of multiple decision trees to improve model accuracy and robustness.[579] Unlike a single decision tree, which is prone to overfitting and sensitive to noise, Random Forests reduce variance by training each tree on a bootstrap sample of the data and introducing randomness in the feature selection process at each split. This approach leads to improved generalisation and stability across a wide range of prediction tasks, making it a strong baseline for both regression and classification problems.[214] Grid

search was conducted across the following range of hyperparameters: number of trees `n_estimators` $\in \{100, 200, 300, 500, 1000\}$, maximum tree depth `max_depth` $\in \{\texttt{None}, 10, 30, 50, 100, 300, 500\}$, minimum number of samples required to split a node `min_samples_split` $\in \{2, 5, 10, 15\}$, minimum number of samples required to be a leaf node `min_samples_leaf` $\in \{1, 2, 4, 6, 8\}$, and maximum number of features considered for splitting `max_features` $\in \{\texttt{'sqrt'}, \texttt{'log2'}\}$.

MLP

For an overview of the theory behind the MLP architecture, please refer to Section 2.5.2. Grid search was conducted across the following range of hyperparameters: optimiser type `optimiser` $\in \{\texttt{Adam}, \texttt{RMSProp}\}$, hidden layer dimensionality `hidden_dim` $\in \{16, 32, 64, 128\}$, number of hidden layers `num_layers` $\in \{1, 2, 3, 4, 5\}$, and batch size `batch_size` $\in \{8, 16, 32, 64\}$.

Convolutional Neural Networks

CNNs are a class of deep learning models originally designed for image and spatial data, but have also been applied to sequential and structured data such as SMILES strings or molecular fingerprints.[580] CNNs leverage convolutional layers to detect local patterns through sliding filters, enabling the model to learn hierarchical feature representations that which are useful for capturing spatial or sequential dependencies.[219] In molecular property prediction, one-dimensional (1-D) CNNs are often applied to representations such as SMILES or SELFIES, where local substructure patterns are informative for property prediction.[581] Grid search was conducted across the following range of hyperparameters: optimiser type `optimiser` $\in \{\texttt{Adam}, \texttt{RMSProp}\}$, hidden layer size `hidden_dim` $\in \{16, 32, 64, 128\}$, number of convolutional layers `num_layers` $\in \{1, 2, 3, 4, 5\}$, number of filters per layer `filters` $\in \{32, 64, 128\}$, and kernel size `kernel_size` $\in \{3, 5, 7\}$. We also varied the batch size `batch_size` $\in \{16, 32, 64, 128\}$ to explore its effect on training stability and generalisation.

Graph Attention Networks

GATs are a type of graph neural network that incorporates attention to weight the importance of neighbouring nodes during message passing. Unlike traditional graph convolutional models, which aggregate information from neighbours with equal or fixed weights, GATs learn attention coefficients that allow the model to focus on the most relevant local substructures for a given prediction task. This makes them very effective for learning on molecular graphs, where the significance of atomic neighbours may vary based on chemical context.[582] Grid search was conducted across the following range of hyperparameters: learning rate `lr` $\in [0.0001, 0.01]$, batch size `batch_size` $\in \{8, 16, 32\}$, activation function `activation` $\in \{\texttt{ReLU}, \texttt{LeakyReLU}\}$, optimiser `optimiser` $\in \{\texttt{Adam}, \texttt{RMSProp}\}$, number of hidden units `hidden_dim` $\in \{4, 8, 16, 32\}$, number of attention heads `head_size` $\in \{16, 32, 64, 128\}$, and number of layers `num_layers` $\in \{1, 2, 3, 4, 5\}$.

Message Passing Neural Networks

MPNNs are a class of graph neural networks designed for learning on molecular graphs. An MPNN operates on undirected molecular graphs $G = (V, E)$ with node features X_v and edge features E_{km}, and consists of two main phases: a message passing phase and a readout phase. During the message passing phase, information is propagated across the graph using message functions M_t and node update functions U_t to learn intermediate node representations, while the readout phase aggregates these representations using a readout function R to compute a fixed-length feature vector that models the molecular properties of interest.[583] MPNNs are well-suited for molecular property prediction tasks, as they can incorporate both node and edge features to learn chemically meaningful representations, and versions of this architecture, such as the directed-MPNN, have been shown in literature to provide superior performance to other deep learning architectures.[30,240,520] Grid search was conducted across the following range of hyperparameters: learning rate `lr` $\in [0.0001, 0.01]$, batch size `batch_size` $\in \{8, 16, 32\}$, activation function `activation` $\in \{$`ReLU`, `LeakyReLU`$\}$, optimiser `optimiser` $\in \{$`Adam`, `RMSProp`$\}$, number of message passing layers `num_layers` $\in \{1, 2, 3, 4, 5\}$, and hidden layer size `hidden_dim` $\in \{16, 32, 64, 128\}$.

Transformers

For an overview of the theory behind the transformer architecture, please refer to Section 2.5.2. Grid search was conducted across the following range of hyperparameters: `batch_size` $\in \{8, 16, 32\}$, `activation` $\in \{$`ReLU`, `LeakyReLU`$\}$, `optimiser` $\in \{$`Adam`, `RMSProp`$\}$, `hidden_dim` $\in \{5, 10, 20, 50\}$, number of transformer blocks `num_blocks` $\in \{1, 2, 3, 4, 5\}$, and self-attention head size `head_size` $\in \{16, 32, 64\}$.

ChemBERTa

ChemBERTa is a transformer-based model built on the RoBERTa architecture and pre-trained on large-scale chemical language datasets such as SMILES. The ChemBERTa family used in this study was trained on 77 million canonicalised and shuffled SMILES strings using multi-task regression (MTR), where the model simultaneously predicts 200 RDKit-computed molecular properties.[251] In this study, we fine-tuned the 46M parameter version of ChemBERTa on our hydrocarbon dataset using the default hyperparameter configuration provided in the original implementation.[251] Fine-tuning enables the model to adapt its internal representations to better reflect the characteristics of the new data, potentially improving both accuracy and generalisation, especially when domain-specific patterns differ from those seen during pretraining.[584] This transfer learning approach is powerful in chemical domains where labelled data is limited, but underlying structural relationships can still be captured through learned embeddings.[249] We fine-tuned ChemBERTa using the HuggingFace Trainer API, training separate models for each molecular property with early stopping and an 80:10:10 train/validation/test split, while unfreezing all model weights to allow full backpropagation through the pretrained transformer layers.

k-nearest neighbours

The k-nearest neighbours (kNN) algorithm is a simple yet effective non-parametric method used for both classification and regression tasks. It makes predictions based on the labels or values of the k most sim-

ilar data points in the training set, as measured by a distance metric in feature space.[585] It has been used for decades for molecular classification and regression tasks.[586,587] Grid search was conducted across the following range of hyperparameters: the number of neighbours `n_neighbours` $\in \{2,5,7,9,15,21\}$, weighting scheme `weights` $\in \{$`uniform`,`distance`$\}$, distance metric power parameter `p` $\in \{1,2,3\}$, and algorithm for neighbour search `algorithm` $\in \{$`auto`,`ball_tree`,`kd_tree`,`brute`$\}$. We also varied the leaf size `leaf_size` $\in \{10,20,30,40,50,100,200,500\}$, which can influence the efficiency of the underlying tree-based search algorithms.

5.3 Results and Discussion

5.3.1 Embeddings and Fingerprint Optimisation

Embeddings

Figure 5.5 shows the effect of different embedding methods and dimensionality on XGBoost prediction of hydrocarbon thermophysical properties. Embedding dimensionalities from 10 to 750 were evaluated for both PCA and UMAP to assess performance across a broad range of feature compression levels, using 5-fold cross-validation and an 80:10:10 train:test:validation split. This range aimed to explore the tradeoff of maintaining the diversity and distribution of chemical and physical characteristics represented by the molecular descriptors within the dataset, and minimising the risk of overfitting due to redundant or noisy features. A prior study found that as few as 12 RFE-optimised (and PCA validated) molecular descriptors were sufficient to accurately predict the viscosity of binary mixtures with an R^2 of 0.98 and a mean-squared error of 1.5.[32] As shown by Figure 5.5, there is some variance in the optimal embedding dimensionality across different properties and temperatures, but overall, PCA embeddings consistently outperformed UMAP across most properties and dimensionalities. This variance suggests that the minimum number of descriptors required to optimise performance is strongly dependent on the property. However, this variation in the best performing dimensionality for each property never exceeded 100 dimensions, suggesting that there may be redundancy in the predictive power in around 85% of molecular descriptors for these properties. We also observe that RMSE values were notably higher for for heat capacity and viscosity, even though these values were normalised. This may reflect greater non-linearity in the relationship between molecular structure and these properties.

PCA's linear assumption with dimensionality reduction by identifying axes of maximal global variance can create a strong synergy with models like XGBoost, which, while not linear themselves, rely on axis aligned splits and benefit from input features that expose monotonic relationships with the target variable.[588] The linear transformations produced by PCA often enhance the signal in dominant variance directions, most likely making the resulting embeddings more compatible for tree-based models. UMAP is designed to preserve local neighbourhood structure and optimise manifold (lower-dimensional space) topology for visualisation or clustering tasks, rather than to maintain global variance.[255] This can result in embeddings that obscure relationships between features and target variables, reducing compatibility with tree-based model models.

To investigate whether the superior performance of PCA is due to synergistic effects with XGBoost,

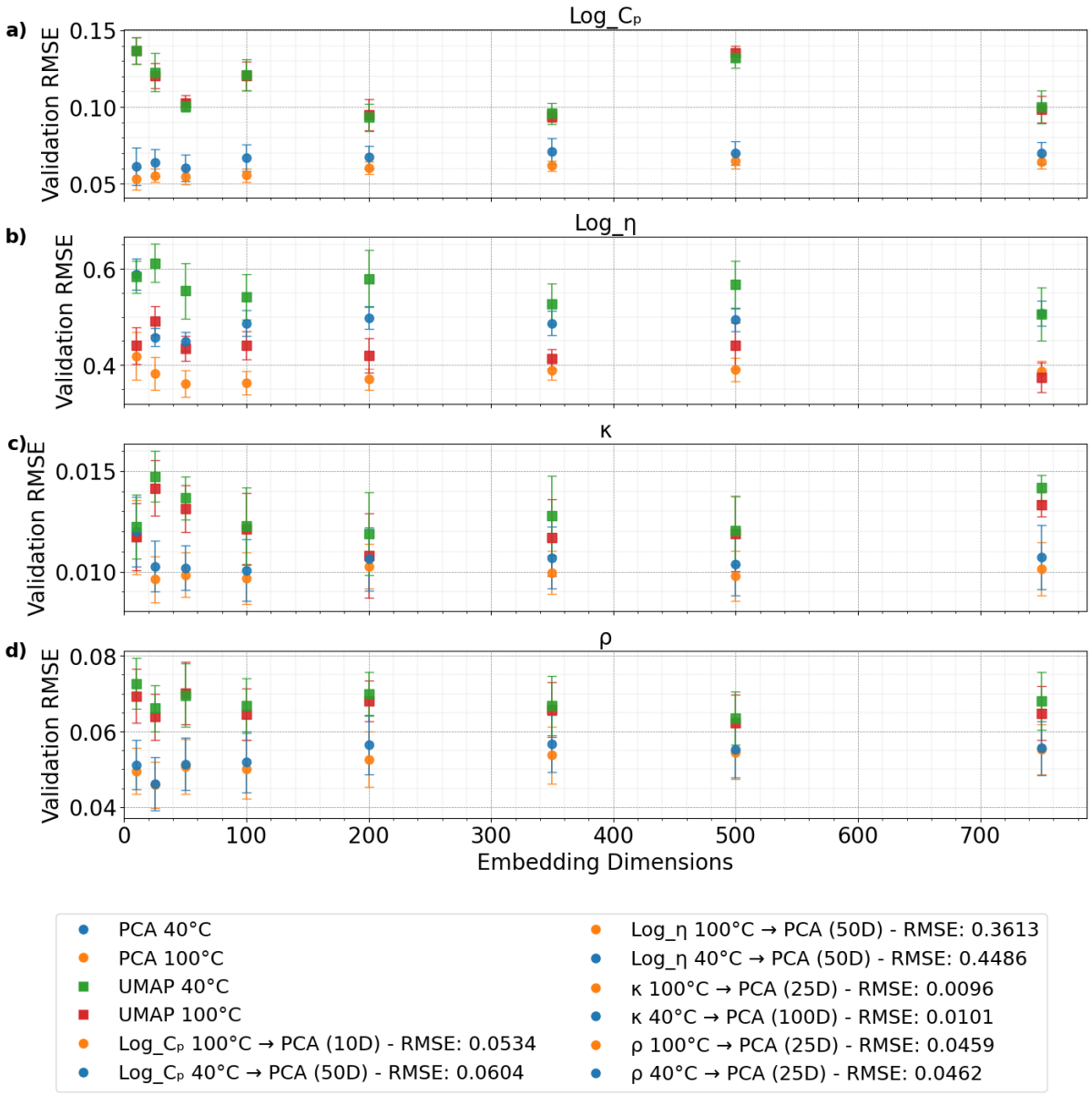


Figure 5.5: Effect of different embedding methods and dimensionality on XGBoost prediction of hydrocarbon thermophysical properties. Best method-dimension combination for each combination shown in legend. Error bars show 1 standard deviation from mean from 5-fold cross validation.

an MLP was trained with the same range of embeddings. As with the XGBoost models, most of the properties were best predicted with embeddings sizes of 100 dimensions or less, with the exception of viscosity at the two temperatures which needed much greater dimensionality to maximise the predictive performance of the MLP. Interestingly, while PCA often outperforms UMAP when used with tree-based models like XGBoost, we find that UMAP embeddings yield superior performance when used with MLPs, for 6 of the 8 properties (Figure C.1). This may be because MLPs can recover the non-linear patterns embedded in the locally structured manifolds UMAP creates, compensating for the global distortions introduced by UMAP and making it a more compatible embedding method for MLPs and other deep models, especially as it is likely that the true structure–property relationships are non-

linear and high-dimensional. Despite this, we observed that overall cross-validation RMSE values were consistently lower for XGBoost compared to MLP across all embeddings and properties, suggesting that XGBoost remains more robust for embedding-based molecular representations.

Property	**PCA XGBoost**	**PCA MLP**	**UMAP XGBoost**	**UMAP MLP**
Density 100°C	**0.046 (25)**	0.090 (750)	0.062 (500)	0.067 (100)
Density 40°C	**0.046 (25)**	0.086 (750)	0.064 (500)	0.067 (25)
Log Heat Capacity 100°C	**0.053 (10)**	0.365 (10)	0.094 (350)	0.145 (10)
Log Heat Capacity 40°C	**0.060 (50)**	0.471 (200)	0.093 (200)	0.165 (50)
Log Viscosity 100°C	**0.361 (50)**	0.398 (350)	0.374 (750)	0.497 (10)
Log Viscosity 40°C	**0.449 (50)**	0.468 (350)	0.505 (750)	0.631 (10)
Thermal Conductivity 100°C	**0.010 (25)**	0.028 (750)	0.011 (200)	0.016 (50)
Thermal Conductivity 40°C	**0.010 (100)**	0.035 (500)	0.012 (200)	0.016 (100)

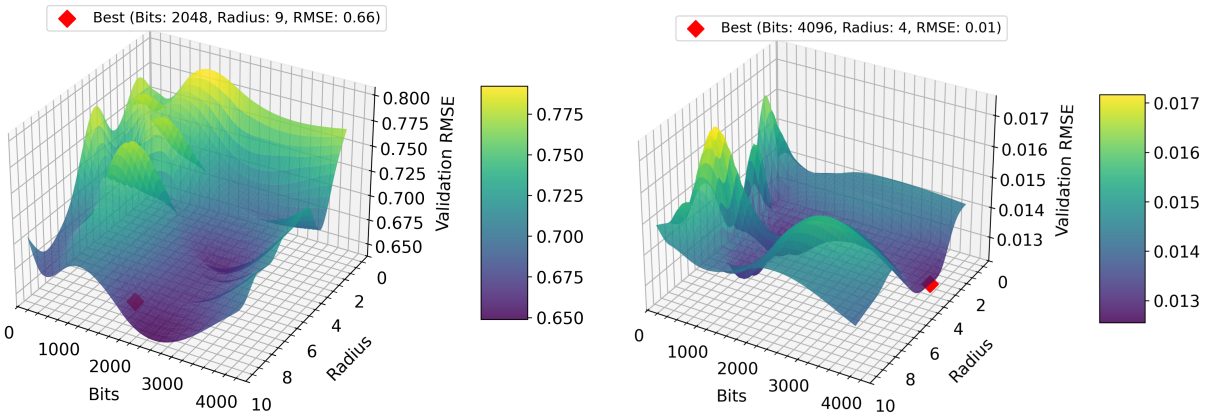
Table 5.4: Best performing validation RMSE and corresponding embedding dimension for each property using PCA and UMAP embeddings with XGBoost and MLP models. Values are shown as *RMSE (dimension)*, and the lowest RMSE per property is bolded.

Table 5.4 summarises the best-performing embedding dimension and validation RMSE for each property for the best performing model trained during cross-validation, for the XGBoost and MLP models and the two dimensionality reduction methods. There is a drop in the validation RMSE scores compared to the average cross-validation scores for the XGboost and MLP models suggesting some degree of overfitting. That aside, the XGBoost models trained with PCA embeddings consistently outperformed the other three combinations. We also note that many more dimensions are required to maximise performance of models trained on the UMAP embeddings compared to the PCA embeddings for the XGBoost models, and vice versa for the MLP models, further highlighting the incompatibility of certain dimensionality reduction techniques with certain model types. However, the use of UMAP emeddings did not consistently lead to significantly improved prediction performance by the MLP models compared to the models trained on PCA embeddings. Based on these findings, in the full benchmarking studies, we use PCA embeddings of 100 dimensions for all of the models.

Fingerprints

We evaluated how ECFP generation influences molecular property prediction using ML. We generated binary Morgan fingerprints for each molecule across a range of fingerprint radii (1–10) and bit sizes (256 to 4096). Varying bit sizes from 256 to 4096 allowed us to investigate how bit density and collision frequency frequency scale with fingerprint size, which becomes increasingly relevant in large molecules where many substructures must be uniquely encoded. XGBoost and MLP models, were trained to predict the eight properties using an 80:10:10 train/validation/test split, again with calculation of 5-fold cross-validation RMSE and validation RMSE.

Table 5.5 shows the best performing ECFP configurations for each property for the trained MLP and XGboost models. We observed that as fingerprint radius and bit size increased, the bit collision rate also tended to rise. This is expected as larger radii capture more extended atom environments, which introduces greater substructure diversity. Despite this rise in collisions, both XGBoost and MLP models



(a) Viscosity at 40°C **(b) Thermal Conductivity at 40°C**

Figure 5.6: Validation RMSE surface plots for XGBoost across bit size and radius for (a) Viscosity and (b) Thermal Conductivity at 40°C.

Model	Property	Bits	Radius	Cross Val. RMSE	Test RMSE	Bit Collisions
MLP	Density 100°C	1024	1	0.0662	0.0459	112
XGBoost	Density 100°C	4096	2	0.0575	0.0411	697
MLP	Density 40°C	2048	1	0.0645	0.0509	108
XGBoost	Density 40°C	4096	2	0.0588	0.0426	697
MLP	Heat Cap. 100°C	2048	3	0.3144	0.2875	108
XGBoost	Heat Cap. 100°C	1024	7	0.1707	0.1547	1024
MLP	Heat Cap. 40°C	4096	3	0.3048	0.2886	102
XGBoost	Heat Cap. 40°C	1024	9	0.1541	0.1585	512
MLP	T. Cond. 100°C	4096	4	0.0165	0.0141	2513
XGBoost	T. Cond. 100°C	4096	4	0.0149	0.0120	2513
MLP	T. Cond. 40°C	2048	4	0.0180	0.0146	122
XGBoost	T. Cond. 40°C	4096	4	0.0157	0.0128	2513
MLP	Viscosity 100°C	1024	9	0.4752	0.4480	1024
XGBoost	Viscosity 100°C	1024	9	0.4767	0.4075	512
MLP	Viscosity 40°C	1024	9	0.6301	0.6563	1024
XGBoost	Viscosity 40°C	2048	9	0.6528	0.6639	2046

Table 5.5: Best performing ECFP configurations by model and property.

exhibited improved predictive performance with higher bit sizes, regardless of the radius. This counter-intuitive effect could be explained by the fact that higher radii encode larger, and hence more diverse and informative, chemical substructures that are better suited for distinguishing molecular properties. At the same time, increasing the bit size reduces the probability of collision per bit by expanding the hashing space, which mitigates the downside of larger radii.

As shown by Table 5.5, the XGBoost models again outperform all of the MLP models. We find that the optimal fingerprint configuration varies by property, most likely because different molecular properties are influenced by different structural features and interaction patterns, as illustrated by Fig. 5.6. Thermal conductivity and density are often governed by local atomic environments such as atomic packing and

short-range intermolecular forces, which can be more effectively captured by fingerprints with smaller radii.[47] However, heat capacity and viscosity depend more heavily on global molecular features, such as overall molecular size, shape, branching, and flexibility, which influence the distribution of vibrational modes or resistance to flow.[47] These more complex are better represented by larger radii fingerprints that encode extended substructures and longer-range connectivity.

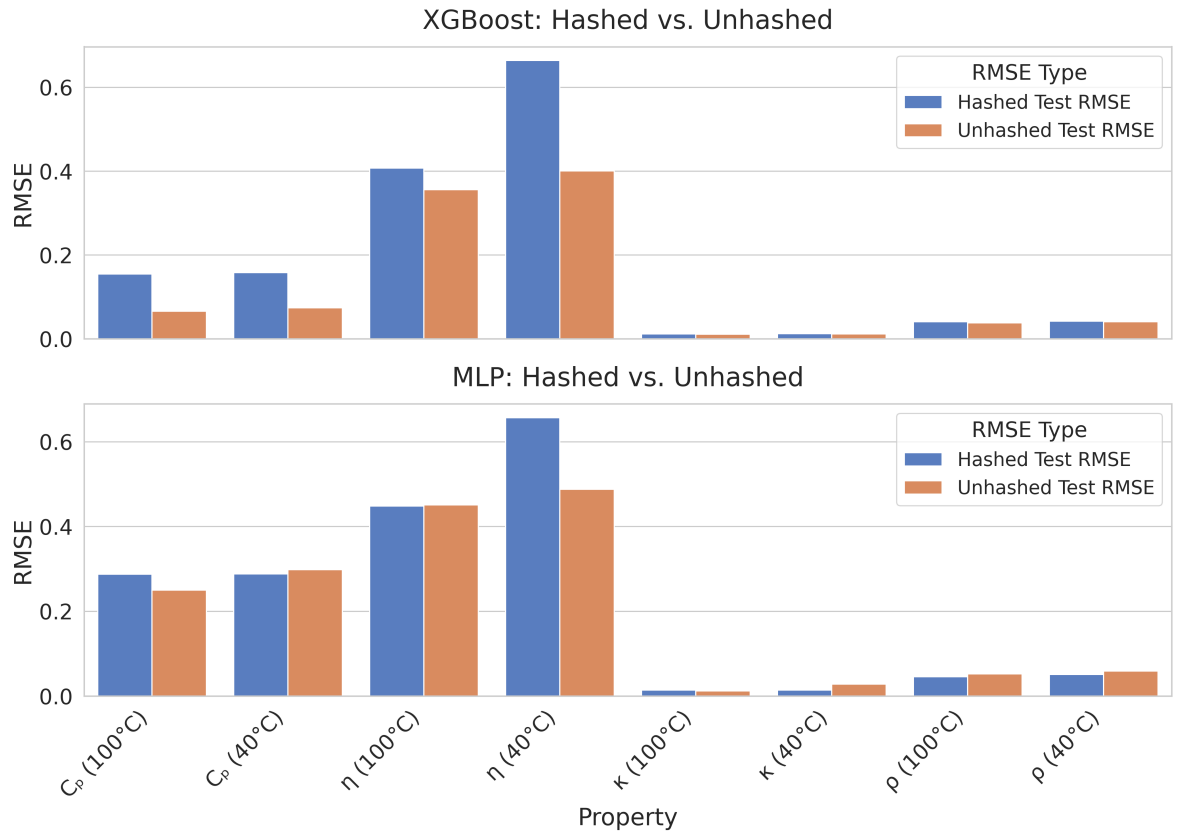


Figure 5.7: Comparison of hashed vs. unhashed fingerprint performance for the best-performing hashed configurations, retrained with unhashed fingerprints at the same radius. The top bar chart displays XGBoost results, and the bottom bar chart shows MLP results. Each bar group indicates the test RMSE for the hashed (left bar) and unhashed (right bar) fingerprints.

Hashed ECFP fingerprints limit interpretability due to bit collisions, which prevent direct tracing of model predictions to specific chemical features. To address this, unhashed Morgan fingerprints were used with a fixed radius, avoiding compression of substructure identifiers into a fixed-length bit vector. This approach eliminates bit collisions and ensures that each substructure retains a unique representation in the feature space.

Unhashed fingerprints present the tradeoff of potentially leading to very high-dimensional and sparse feature spaces, especially for large or diverse molecular datasets, increasing memory consumption and degrading performance. However, as shown in Fig. 5.7, models trained on unhashed fingerprints achieved comparable or better performance on the test set than those trained on hashed fingerprints. This improvement is likely arises to the complete removal of collisions, allowing the model to treat each

chemical environment as a distinct input and more effectively learn structure–property relationships, compensating for any sparsity in the representation. Based on these results, unhashed fingerprints with a radius of 8 were used in all benchmarking studies to improve the prediction of complex properties such as heat capacity and viscosity.

5.3.2 Best Representations for ML and best ML Performance

Table C.1 shows the test RMSE values for all the different ML models across molecular properties, when trained with different molecular representations as well as an overall rank of model performance compared to others. To calculate rank, RMSE values are first transformed via z-score normalisation and then inverted so that better performance (i.e., lower RMSE) corresponds to higher normalised performance. All predicted property performances are thus scaled to a [0, 1] range, allowing direct comparison. We then compute rank by averaging each model's normalised scores, with higher averages indicating better performance.

XGBoost and RF models, when used with molecular descriptors, consistently performed well across all transport properties. This aligns with previous studies,[237,589] which also found tree-based models to be reliable across datasets. These methods handle non-linear relationships and high-dimensional features effectively, making them suitable for the diverse information captured by molecular descriptors. Tree-based models are also less sensitive to data scales, irregular patterns in the target function and uninformative features.[590] XGBoost performed particularly well for predicting viscosity and heat capacity, likely due to the strong dependence of these properties on the types of features encoded by molecular descriptors. When comparing overall predictive performance across all properties for different representations, the trend for XGBoost performance is: descriptors > embeddings > fingerprints. This may be explained by the fact that while descriptors and embeddings can incorporate both structural and physicochemical information, fingerprints are purely structural, lacking explicit encoding of global molecular properties, which may limit their expressiveness for predicting more complex properties.

In contrast to the strong performance of tree-based models, the SVM and KNN models showed comparatively poor results when using molecular descriptors, even though inputs were scaled pre-training. This may be due to the high dimensionality and redundancy often present in descriptor sets, which can degrade the performance of distance based models like KNN and SVM models (the curse of dimensionality).[591] Although SVMs have shown strong results in other regression and classification benchmark studies,[237] their success often depends on careful feature selection, which may not generalise well to prediction tasks such as transport property prediction. The underlying structure–property relationships in the collated dataset most likely form a complex, non-linear landscape with overlapping and interdependent effects, which may challenge SVMs' ability to construct effective decision boundaries, even with tuned parameters. Interestingly, SVM models perform significantly better when paired with learned molecular embeddings rather than traditional descriptors, likely stemming from the fact that the constructed embeddings used had significantly fewer dimensions, suggesting that the more compressed nature of embeddings provides a more suitable input representation for SVMs.

The GAT and MPNNs models, which relied on 2D molecular graph representations consistently per-

formed worse compared to other model-representation combinations. This result is surprising, given the theoretical advantages of 3D structural information in modeling molecular properties, and that previous studies have found that graph ML methods trained on similar molecular graphs have shown strong performance.[30,520] One potential reason for this discrepancy is that graph-based models typically require larger datasets to learn effectively,[592] meaning that the datasets used in this study may not be large enough for graph based models to learn effectively. Another potential reason could be the method of generation of the input graphs used to train these models. While the combination of ETKDG and MMFF94 minimisation provides a fast and reasonable estimation of molecular geometry, these methods may not always capture the most energetically favourable conformeations, especially for flexible molecules or properties that are strongly dependent on molecular conformation. As a result, the graph-based models trained based on bond lengths and partial charges calculated in this way may be limited in their ability to fully represent the true spatial and electronic environments that govern complex transport properties. Creating base conformers through more chemically robust methods, such as density functional theory (DFT) could yield improvements.

The GAT models outperformed the MPNN when considering overall performance, suggesting architecture differences may have contributed to the performance gap. GAT uses attention mechanisms to dynamically weight the contributions of neighbouring nodes during message passing, which may allow it to identify relevant substructures and improve robustness to noise.[593] MPNNs, which tend to aggregate messages more uniformly across neighbours unless guided by carefully engineered features.[520] The GAT's attention mechanism most likely enabled more effective information weighting from neighbouring nodes, allowing the model to focus on the most relevant features when predicting properties, contributing to its superior performance compared to MPNNs. It is also possible that GAT's architecture was more robust across the tested hyperparameter space, while the MPNN may have required more careful tuning to reach its full potential.

Among the deep learning models evaluated, the transformer models using the PCA molecular embeddings achieved the best overall performance, and performed near comparably to the top-performing tree-based descriptor models. Transformers with precomputed embedding-based inputs likely outperformed MLPs and CNNs due to their ability to model long-range dependencies with its attention mechanism. It is surprising that a relatively small transformer model with fixed-length molecular embeddings outperformed ChemBERTa, a large pretrained transformer designed specifically for chemistry-related tasks, when paired with both SMILES strings and embeddings, especially given the success of transformer-based large language models in natural language processing when pretrained on massive datasets.[594] However, the performance gap may stem the fact that ChemBERTa was pretrained on SMILES strings, which while convenient, does not fully capture the spatial, electronic, or topological features of molecules that give them their properties. This suggests the SMILES format lacks sufficient structural information such that even a powerful pretrained model may struggle to generalise effectively. Another contributing factor could be a distributional mismatch between ChemBERTa's pretraining corpus and the downstream task dataset. The molecules seen during pretraining were mainly drug-like molecules,[251,567] differing substantially in both structure and property distribution from the

molecules in the hydrocarbon dataset. Moreover, our finetuning approach may have inadvertently reduced ChemBERTa's effectiveness, as we allowed all model weights to be updated during training, potentially overwriting useful pretrained features, especially if the finetuning dataset was small. In such cases, techniques like LoRA (Low-Rank Adaptation) could be more effective. LoRA freezes the main weights of the model and introduces small, trainable matrices that adapt the model to the downstream task without overwriting its core knowledge, often leading to better generalisation in finetuning tasks.[254] This approach has already been shown to be effective with protein large-language models.[595]

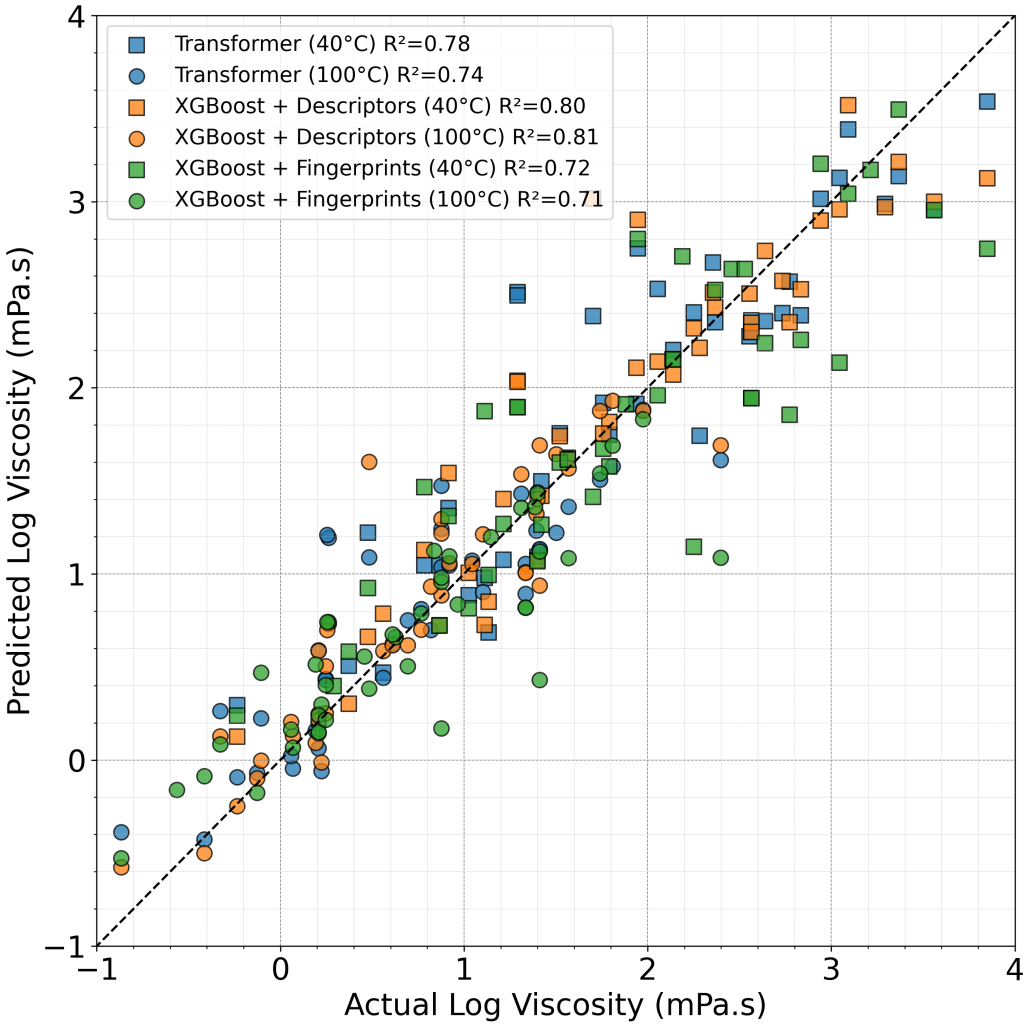


Figure 5.8: Parity plot comparing actual vs. predicted values for viscosity across three models: Transformer, XGBoost with molecular descriptors, and XGBoost with fingerprints. Marker shapes denote temperature (circles for 100°C and squares for 40°C).

Models trained on character-level embeddings of string-based molecular representations such as SMILES and SELFIES more often than not underperformed relative to models using molecular descriptors or PCA embeddings. Interestingly, the same model architectures trained on SMILES inputs often slightly outperformed their counterparts trained on SELFIES, despite the theoretical advantages of SELFIES in

guaranteeing valid molecular strings. This performance gap may stem from fundamental differences in how SMILES and SELFIES encode molecular structure. While SMILES strings are not guaranteed to produce valid molecules when randomly generated or mutated,[573] SMILES tends to preserve certain local structural patterns, such as rings, branches, or functional groups, in a way that maintains some degree of chemical locality. That is, substructures that are adjacent in chemical space are often close together in the SMILES string. This implicit locality may help deep models models learn chemically meaningful sub-patterns even from character-level data.

SELFIES strings, however, were designed to always decode to valid molecules regardless of the string structure.[574] While this robustness is highly beneficial in generative tasks, the encoding process tends to disrupt local chemical continuity. SELFIES tokens represent abstract chemical building blocks and may rearrange or distance related substructures within the sequence. As a result, two neighbouring substructures might appear far apart in the SELFIES representation, making it harder for models, especially those relying on convolutional or recurrent architectures, to learn useful patterns based on token proximity. This is best highlighted by the large difference in rank between the best performing CNN model trained on SMILES strings, and the best performing CNN model trained on SELFIES strings. Moreover, the additional abstraction in SELFIES results in longer sequences with more complex token vocabularies, which may increase the difficulty of learning effective embeddings, especially with limited data or smaller models.

To further analyse the performance of the top three models, parity plots were generated to visualise the performance of these models on the test set. Fig. 5.8 shows a parity plot for viscosity predictions at 40°C and 100°C for the top three performing models, while Fig. C.2 shows parity plots for predictions of the other properties at the two temperatures. Across the evaluated properties, the R^2 values generally indicate stronger predictive performance at 100°C compared to 40°C, though this trend varies by property. For heat capacity, all models perform exceptionally well at both temperatures, with XGBoost using descriptors achieving nearly perfect R^2 values (> 0.99), probably reflecting the lower variability of this property. Density shows notably lower R^2 values, especially for the transformer model which may be due to effects structure has on molecule density that embeddings fail to capture. Interestingly, fingerprint-based XGBoost models performed slightly better for density, suggesting that structures play a greater role in determining this property. For thermal conductivity and viscosity, moderate to high R^2 values are observed across models, with the descriptor-based model again outperforming others. In both cases, predictions at 100°C tend to have slightly higher R^2 values, potentially due to reduced property variance at higher temperatures. These trends highlight how both temperature and feature representation impact regression accuracy across different thermophysical properties.

Multitask learning enables models to share representations across related tasks, potentially improving generalisation, reducing overfitting, and enabling knowledge transfer between tasks with limited data.[596] Multiple studies have shown that multitask deep learning models can outperform single-task deep learning models on certain problems.[597,598] While models such as XGBoost cannot be readily adapted for multitask regression as they are inherently designed to optimise a single output target per tree, and lack a shared representation mechanism across tasks, deep learning models such as transformers are read-

Train Method	Density	Heat Capacity	Thermal Conductivity	Viscosity
Multitask 40°C	0.106 (*−0.895*)	0.136 (*0.829*)	0.121 (*−16.360*)	0.326 (*0.881*)
Multitask 100°C	0.093 (*−0.442*)	0.099 (*0.906*)	0.059 (*−3.763*)	0.302 (*0.823*)
Single-task 40°C	0.045 (*0.26*)	0.055 (*0.94*)	0.017 (*0.63*)	0.319 (*0.78*)
Single-task 100°C	0.045 (*−0.02*)	0.055 (*0.93*)	0.017 (*0.64*)	0.275 (*0.74*)

Table 5.6: RMSE and R^2 (in italics) of multitask transformer model for each property and temperature.

ily adaptable to this approach. To investigate whether a multi-task approach could improve predictive accuracy, we trained a transformer model to predict the 8 properties so far analysed in this study.

As shown by Table 5.6, we observed that the multitask model did not outperform any of the single-task models using the same transformer architecture. A potential reason for this could be that the use of identical molecular structures as inputs across tasks may limit the diversity of the learning signal and reduce the benefits of shared representations. Since each task involves predicting a different property of the same structure, the model may struggle to specialise internal representations for each property, leading to task interference.[596,599] This may be further exacerbated by task conflict, where gradients from different property predictions push the model in competing directions, or by the differences in task complexity, which we can see is quite large considering the different R^2 values across the different properties.

5.3.3 A Comparison of ML and MD Techniques for the Prediction of Hydrocarbon Thermophysical Properties

MD simulations using popular force fields such as LOPLS and COMPASS have been shown to accurately predict the thermophysical properties of linear alkanes[139] branched alkanes[105], and esters.[600]

However, fewer studies have investigated the accuracy of such forcefields for saturated and unsaturated cyclic molecules of sizes relevant to lubricant base oils. For example, a study by Kondratyuk et al. studied the efficacy of EMD methods at predicting the density and viscosity of mixtures of 1-methylnaphthalene and methylcyclohexane at 323K for a range of pressures, finding that this approach had good agreement with experimental values up to 100 MPa.[489] However, molecules of this size ($< C_{12}$) are unlikely to be able to serve as useful lubricant base oils. As demonstrated by the previous sections, ML approaches can provide accurate predictions of the thermophysical properties of a wide range of lubricant-like molecules, when trained on an appropriate dataset. Determining whether MD simulations or ML models offer greater predictive accuracy is important in the context of high-throughput molecular screening, where large numbers of candidate molecules must be evaluated in a short period of time. ML models can generate property estimates much more quickly than MD, but their utility depends on whether they can match or exceed the reliability of such physics-based simulations.

To determine which approach possesses better predictive accuracy, we conducted EMD simulations on nine lubricant molecules to assess the predictive performance of four all-atom force fields (LOPLS,[149,601,602] COMPASS,[150] AMBER,[603] and SciPCFF[150,151]), comparing these outputs to the predictions made by the best performing XGBoost models trained in the previous section for thermal conductvity,

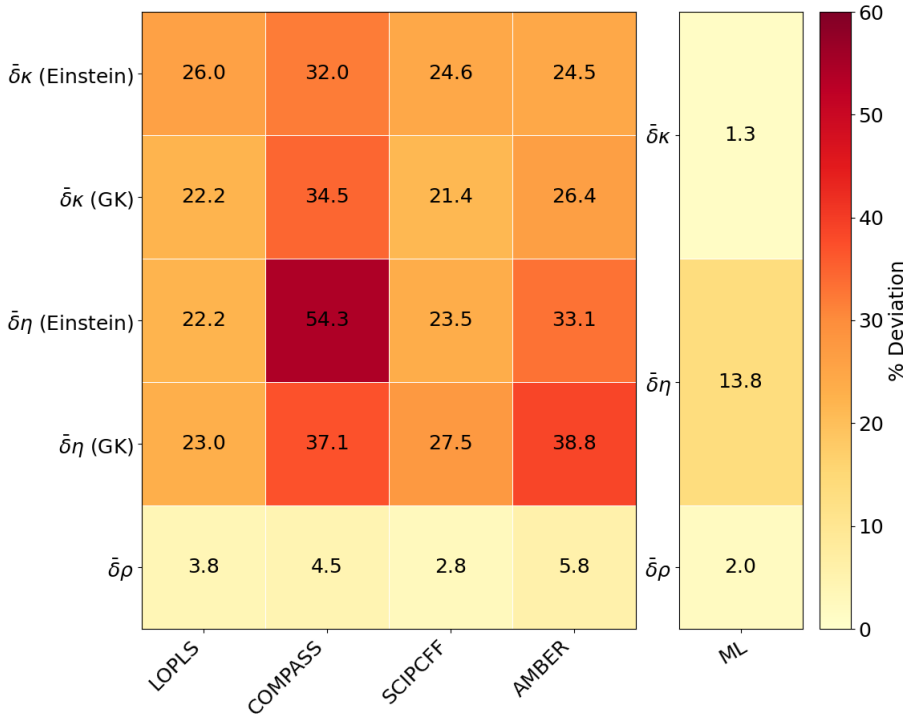


Figure 5.9: Mean absolute percent deviation of simulation predictions from experimental values for density ($\bar{\delta}\rho$), viscosity ($\bar{\delta}\eta$) via both Green–Kubo (GK) and Einstein formalisms, and thermal conductivity ($\bar{\delta}\kappa$) via GK and Einstein.

viscosity and density. These nine molecules were chosen due to their industrial relevance (see Section 5.2.1), and to ensure no data leakage, these molecules were removed from the datasets used to train the XGBoost models. Experimental values for the test molecules were taken from a combination of the NIST WebBook, the Landolt-Bornstein series, and values available in literature.[139,554,555,562,604,605,606] Parameters for the LOPLS simulations were taken from Jorgensen et al.[149], Price et al.[607], Sui et al.[601], and Pluhackova et al.[602] The LOPLS systems were parameterised using the MAPS platform from Scienomics SARL. The SciPCFF[150,151] and AMBER95[603] systems were also parameterised using the MAPS platform from Scienomics SARL. The COMPASS[150] systems were parameterised using the LUNAR suite of tools by Kemppainen et al.[608] For each simulation, we calculated density, thermal conductivity, and viscosity using the Green–Kubo and Einstein formalisms (detailed in Section 5.2.2) and compared the resulting predictions with both experimental measurements and predictions from the XGBoost models. Uncertainties for the XGBoost models were taken as the RMSE values from prediction on the test sets during training (see Table C.1).

Figure 5.9 shows the mean absolute percentage deviation from the experimental values for each prediction method (for both temperatures), calculated as:

$$\bar{\delta}X = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{X_i^{\text{pred}} - X_i^{\text{exp}}}{X_i^{\text{exp}}}\right| \times 100\%,$$

where X denotes the property of interest (density ρ, viscosity η, or thermal conductivity κ), and N is the

total number of predictions made. Fig. C.3a, Fig. C.3b and Fig. C.4 show experimental values against MD predictions (from both Green–Kubo and Einstein formalisms) and the XGBoost models for density, thermal conductivity and viscosity respectively, while Table C.4, while Table C.5 and Table C.6 show the property predictions for each of the molecules by each prediction method. As shown by Fig. 5.9, the XGBoost models significantly outperform the MD predictions for all of the properties, with absolute differences of 0.8%, 8.4% and 20.1% lower mean deviations than the next best MD approaches for density, viscosity and thermal conductivity respectively. The SciPCFF yields the predictions with the lowest mean deviation from the experimental values out of all the MD forcefields for density and thermal conductivity when using the the GK formalism. The LOPLS force field combined with the Einstein formalism performs best out of all the MD approaches for the prediction of viscosity, although there is not much difference between LOPLS and SCIPCFF when using either the Green-Kubo or Einstein formalisms. A recent extension, known as OPLS/2020,[609] with updated parameters for unsaturated hydrocarbons, alcohols, and ethers could potentially improve performance further. However, unlike LOPLS, the changes made in this parameterisation were not specifically for long chain lubricant-like molecules.

There is no significant trend of the Green-Kubo and Einstein methods outperforming each other, in agreement with previous studies showing that these methods provide similar estimates for viscosity.[114,610] Furthermore, there seems to be no pronounced trend of over- or underestimation by the MD methods for thermal conductivity or viscosity; however, MD often under-predicted the density values (Fig. C.3a). This could be an artefact of the NVT ensemble used to estimate the viscosity values. That notwithstanding, all MD forcefields seemed capable of providing density estimated with mean deviations no larger than 6% on average. A similar study investigating the performance of LOPLS and COMPASS for predicting the density of linear and branched alkanes reported average deviations of 0.8% when predicting density using COMPASS, and average deviations of 1.4% for LOPLS.[105] This study also reported also reported average deviations of 14.1% when predicting viscosity using COMPASS, and average deviations of 38.4% when using LOPLS.[105]

The relatively poor performance of all the forcefields compared to the XGboost models may be because the prediction of these properties using EMD simulations is more difficult for cyclic hydrocarbons than for linear and branched alkanes, due to differences in molecular structure and conformational flexibility. The ring structures in these systems introduces a wider range of intramolecular interactions and alters packing behaviour in the liquid phase.[611,612] Many common used force fields are parameterised for linear and branched alkanes and may not capture the structural and electronic features of cyclic compounds. As a result, simulations involving cyclic hydrocarbons may require force field refinement or reparameterisation to achieve more accurate predictions.[613] Despite these limitations, MD simulations are still very useful for scenarios where the system of interest lies outside the chemical or thermodynamic domain covered by the training dataset of the ML, such as predictions at a wider range of pressures and temperatures.

5.3.4 MD versus Experimental Dataset Model Performance

It is well known that the current availability of high-quality, large-scale experimental datasets in the field of tribology is insufficient to fully harness the potential of ML for property prediction.[614] Experimental data acquisition remains expensive, time-consuming, and often biased toward pre-existing, easily synthesisable compounds. In attempts to overcome this issue, recent research efforts have explored the generation of synthetic datasets via MD simulations, with some studies demonstrating the feasibility of using MD-derived data to augment experimental datasets, or serve as stand-alone training data for property prediction tasks.[31,32,240,615] Nevertheless, questions remain regarding the fidelity of MD-generated data and its utility in training ML models that generalise well to real-world test cases.

In our study, we investigated whether varying the amount of MD-generated data has a tangible effect on the performance of ML models, as measured by their root mean square error (RMSE) on a test set. To this end, we used the publicly available dataset from the study by Kajita et al.[29], which comprises MD simulations for approximately 48,000 small organic molecules in a similar region of chemical space to the experimental dataset used to train the ML models in the previous sections. The scale and homogeneity of this dataset allows it to serve as an ideal benchmark to evaluate whether increased volume of synthetic data can compensate for the potential inaccuracies inherent in MD simulations. We trained the top three model-representation combinations with increasing fractions of the Kajita MD dataset, using an 80:10:10 train:test:validation split. For example, models trained on 10% of the total dataset would be trained on approximately 3840 molecules, with a test set of 480 molecules, but validated on the same 4800 molecules as all other models. These MD-trained models were then evaluated on the same test set as the models in Section 5.3.2. As shown by Table 5.7, the MD-trained ML models across all training set sizes present substantially worse RMSE and R^2 values on the test set than the best performing models trained on the experimental dataset (Table C.1). Nevertheless, the performance of the model on the validation set is sensitive to the size of the dataset used to the train said model.

Property	Temp	Model	Validation RMSE	Test RMSE
Density	40°C	Descriptor	0.009 [0.86] (0%) ⟨1%⟩	0.065 [0.29] (+16.1%) ⟨10%⟩
Density	40°C	Fingerprints	0.017 [0.82] (+20.1%) ⟨50%⟩	0.067 [0.24] (+10.0%) ⟨100%⟩
Density	40°C	Transformer	0.014 [0.88] (+77.7%) ⟨50%⟩	0.086 [–0.24] (+21.3%) ⟨30%⟩
Density	100°C	Descriptor	0.011 [0.91] (0%) ⟨1%⟩	0.066 [0.28] (+25.7%) ⟨10%⟩
Density	100°C	Fingerprints	0.018 [0.82] (+26.8%) ⟨50%⟩	0.073 [0.11] (+20.7%) ⟨20%⟩
Density	100°C	Transformer	0.014 [0.90] (+87.4%) ⟨50%⟩	0.099 [–0.64] (+39.0%) ⟨20%⟩
Viscosity	40°C	Descriptor	0.067 [0.98] (+52.8%) ⟨100%⟩	0.650 [0.53] (+1.1%) ⟨100%⟩
Viscosity	40°C	Fingerprints	0.189 [0.87] (+47.7%) ⟨100%⟩	0.865 [0.16] (+10.8%) ⟨20%⟩
Viscosity	40°C	Transformer	0.061 [0.99] (+64.8%) ⟨100%⟩	1.255 [–0.77] (+5.1%) ⟨20%⟩
Viscosity	100°C	Descriptor	0.044 [0.99] (+54.1%) ⟨100%⟩	0.478 [0.56] (+2.6%) ⟨10%⟩
Viscosity	100°C	Fingerprints	0.148 [0.86] (+52.5%) ⟨100%⟩	0.667 [0.14] (+10.3%) ⟨10%⟩
Viscosity	100°C	Transformer	0.037 [0.99] (+54.4%) ⟨100%⟩	0.938 [–0.71] (+1.9%) ⟨10%⟩

Table 5.7: Best performing RMSE for each property and temperature. Non-bracketed numbers are best RMSE values; square brackets denote R^2; parentheses denote percentage improvement from 1% use of the training set if any; angle brackets denote the training split at which this performance was achieved.

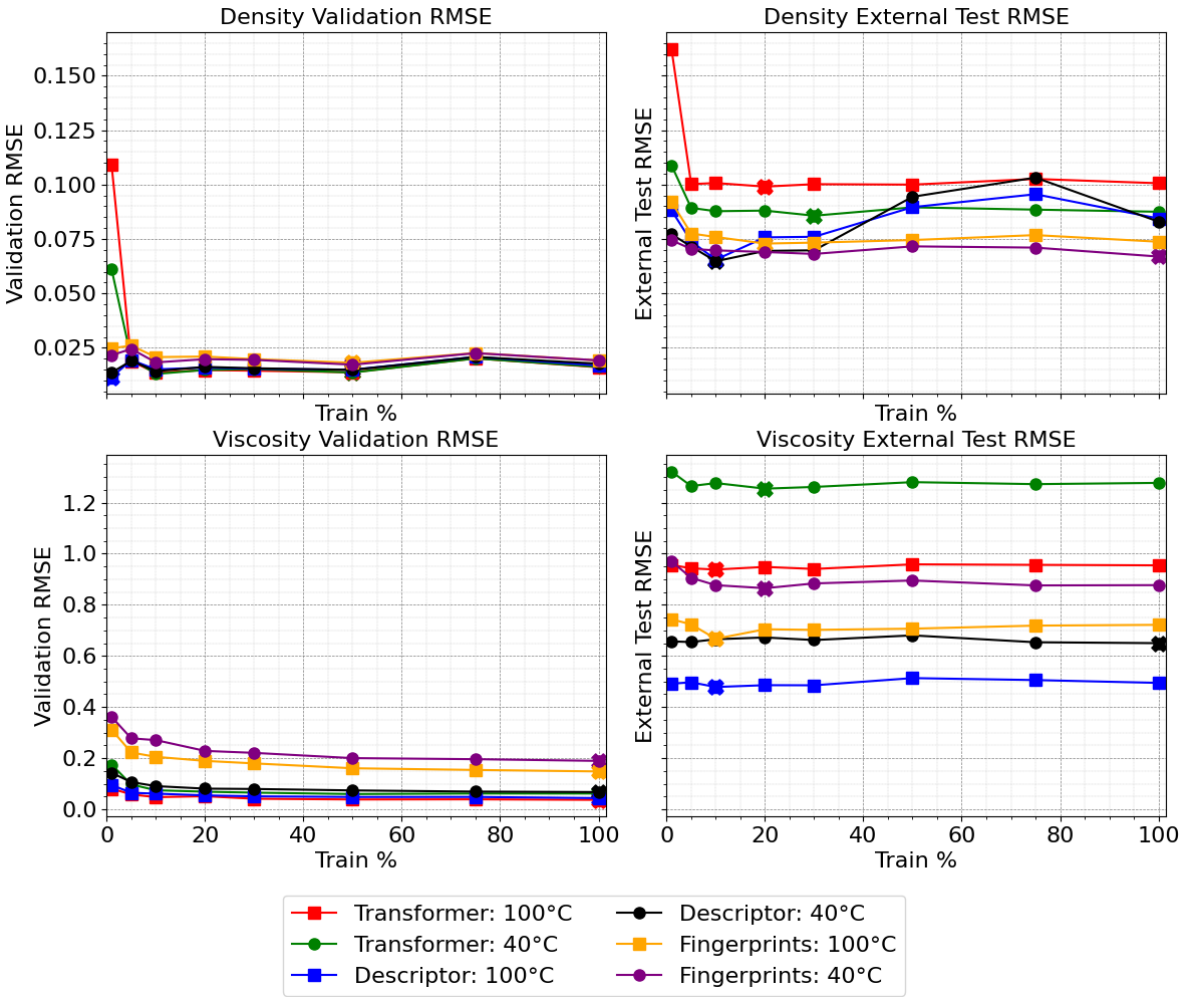


Figure 5.10: Effect on training dataset size on prediction of viscosity and density at 40°C and 100°C. Each colored line corresponds to a specific model–representation combination, and "X" markers indicate the training percentage with the best RMSE.

As Fig. 5.10 shows, increasing the size of training set almost always reduces RMSE for every model–representation combination when evaluated on both the MD-validation dataset and the experimental test set. Most models performed best when trained on 50% of the Kajita dataset when predicting density and 100% of the Kajita dataset when predicting viscosity, suggesting that beyond a certain point there are diminishing returns when training with more data for density prediction. A potential reason for this is that, as the training set grows, small measurement errors in density compound until they exceed a noise threshold, at which point this accumulated error acts like random noise and begins to degrade the model's ability to learn accurate relationships. The exception is for XGboost models trained on descriptors on the validation set to predict density, achieving its best RMSE (0.009 and 0.011 for 40°C and 100°C respectively) at just 1% of the training set, with additional training examples degrading its performance on the MD validation set. This trend does not translate to the performance of the model on the experimental test set, where this model performs best when trained on 10% of the Kajita dataset. This suggests the training on the very small set led to overfitting that falsely inflated the model accuracy.

When evaluated on the test set, there was more variation in the optimal amount of the Kajita dataset to maximise predictive performance, although most models performed best using 10-20% of the dataset. Percentage increases from the 1% dataset baseline were less pronounced for the experimental set than on the MD validation set, suggesting that the increase in training dataset size did not strongly correlate to improved performance on the test set. The transformer architectures trained on PCA embeddings consistently deliver far larger percentage improvements as more training examples are added, gaining up to 87% RMSE reduction, whereas the XGBoost-based descriptor and fingerprint models improved more modestly. This reflects the reliance of the transformer on an abundance of examples to learn distributed feature interactions, and shows that with limited data, it is more likely to overfit, as is possibly the case with other deep learning models evaluated in earlier sections of this study. XGBoost models saturate earlier, extracting most of their predictive power from relatively few samples and yielding smaller incremental benefits from additional training points.

More generally, the lack of significant effect on test set performance across different training dataset sizes suggests that scale alone does not guarantee improved predictive performance when the data originates from synthetic simulations. One contributing factor to this relatively poor performance could be that, despite its size, the Kajita dataset does not feature some of the functional groups present in some of the molecules in the test set. As shown by Table 5.1, the Kajita dataset does not feature any alkanes, cycloalkanes, alkenes or napthalenes, most likely a result of the use and/or parameterisation of the MCTS algorithm used to generate the population. The MCTS algorithm in the Kajita study was parameterised to optimise the upper confidence bound to maximise DVI, resulting in the algorithm failing to survey a sufficiently broad range of compounds, underscoring the need for a selection strategy that actively explores a more diverse span of the chemical space.[29]

Another reason for the performance discrepancy could be down to the specific force field employed, 'TEAM_MS', used in the Kajita study. This force field is not open-source, but has been validated on common organic molecules, polymers, drug-like molecules, zeolites, ionic liquids and metal oxides.[616] The poor performance of the MD-trained ML models in this case suggests that this force field parameterisation may not sufficiently capture the relevant intermolecular interactions or molecular conformations for accurate prediction of transport properties outside of the training set distribution. Classical force fields are parameterised using empirical or ab initio data for a limited set of molecular systems and may not generalise well for prediction of larger or chemically diverse molecules.[617,618]

To estimate liquid viscosity, the Kajita study[29] employed the shoving model, which describes flow as the result of molecules escaping their local cages by pushing against neighbours.[619] By combining this model with transition-state theory and a modification accounting for boiling point (van Velzen's model),[620] an Arrhenius equation is derived for viscosity predictions using drastically short MD simulations and SMILES-based boiling point estimates.[621] While computationally efficient, this approach may compromise the accuracy of the MD simulations, especially for properties like viscosity which are sensitive to long-timescale molecular behaviour. For instance, viscosity and diffusion coefficients are highly dependent on the dynamic behaviour of molecules over extended simulation timeframes, where, if these are truncated prematurely, the property estimates may be unreliable.[114] Of course, a more accurate

dataset might be obtained through full-length MD simulations with more sophisticated force fields (e.g., polarisable or machine-learned potentials);[31,105] however, the computational cost associated with such an undertaking would be prohibitive for datasets on the scale of tens of thousands of molecules. Therefore, while MD data generation remains a promising avenue, it cannot yet serve as a complete substitute for experimental data in training ML models without considerable methodological improvements.

5.3.5 Out Of Distribution (OOD) Prediction

It is well-known that ML models yield unreliable property predictions for molecules that lie outside the chemical domain of their training sets.[515,622] This limitation is problematic for molecular discovery pipelines, where generated compounds frequently occupy regions of chemical space not sampled during model training, leading to large and unquantified prediction errors. Experimental measurements for such molecules are scarce, which impedes evaluation of model generalisation. Furthermore, certain molecular representations are inherently unsuited to OOD property prediction. For example, fingerprint-based models will not be able to extrapolate reliably because they have only encoded substructures seen during training, meaning any novel fragments will not activate bits, leading to meaningless predictions. Tokenisations and embeddings based solely on string-based representations such as SMILES will not be able to generalise as they rely on short and long term sequential relationships that will be distorted by out of distribution atoms and functional groups.

Molecular descriptors, on the other hand, provide features which, by design, vary smoothly with subtle structural modifications, characterise properties that are relatively independent of chemistry (e.g., molecular volume), and span a broad range of chemistries, making this representation much more suitable for OOD prediction.[623] This has been demonstrated in the drug-discovery domain by Tossou et al, who demonstrated that OOD prediction is a covariate shift problem (where there is a difference in test and dataset distributions).[624] Based on this characterisation, they showed performance degradation when predicting on OOD molecules is much more dependent on representation than choice of model. They also showed that RDKit-based descriptors perform better than other representations on in-distribution test sets, and minimise performance degradation on OOD property prediction compared to other representations (fingerprints, molecular graphs, embeddings).[624]

However, descriptor sets can be high-dimensional and correlated, which risks overfitting and singular covariance estimates for smaller datasets unless carefully regularised. Moreover, many descriptors rely on empirical or quantum-chemical calculations calibrated on familiar chemistries, so completely novel structures may fall outside their calibrated ranges, leading to extrapolation errors. Beyond this, defining and quantifying out-of-distribution (OOD) status is itself challenging. Previous studies have leveraged Tanimoto similarity on fingerprints[625] or Euclidean distance in descriptor or cluster-based spaces.[624] However, when using fingerprints, novel compounds trigger zero or near-zero similarity scores, providing no useful gradient for extrapolation, while Euclidean distance in descriptor space may become dominated by a few high-variance axes.

With these limitations in mind, to quantify OOD error in our XGBoost-descriptor models, we assembled a small test set of tribofluids comprising atom types and viscosity ranges absent from the experimen-

tal training dataset (Table C.2 and Table C.3). We measure OOD status via Mahalanobis distance in descriptor space, which mitigates varying descriptor scales by normalising each feature and ensuring the covariance matrix captures inter-feature correlations rather than just magnitudes (as with Euclidean distance).[626] Mahalanobis distance is calculated by:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})},$$

where x is the descriptor vector of the query molecule, μ is the training-set mean vector, and Σ is the covariance matrix of the training descriptors. We flag any molecule for which

$$D_M^2(\mathbf{x}) > \chi^2_{p,0.975},$$

where p is the number of retained descriptors. In other words, we mark a molecule as an outlier if it is unusually 'far' in terms of descriptor distribution from the rest of the data. Specifically, if this descriptor distance is further away from the centre of the distribution of the training set than 97.5% of molecules in the set, it is flagged as OOD. As Mahalanobis distance assumes a multivariate-Gaussian descriptor distribution and requires an accurate covariance matrix, we apply recursive feature elimination (RFE), with 5-fold cross validation, to remove redundant or noisy descriptors in an attempt to improve performance.

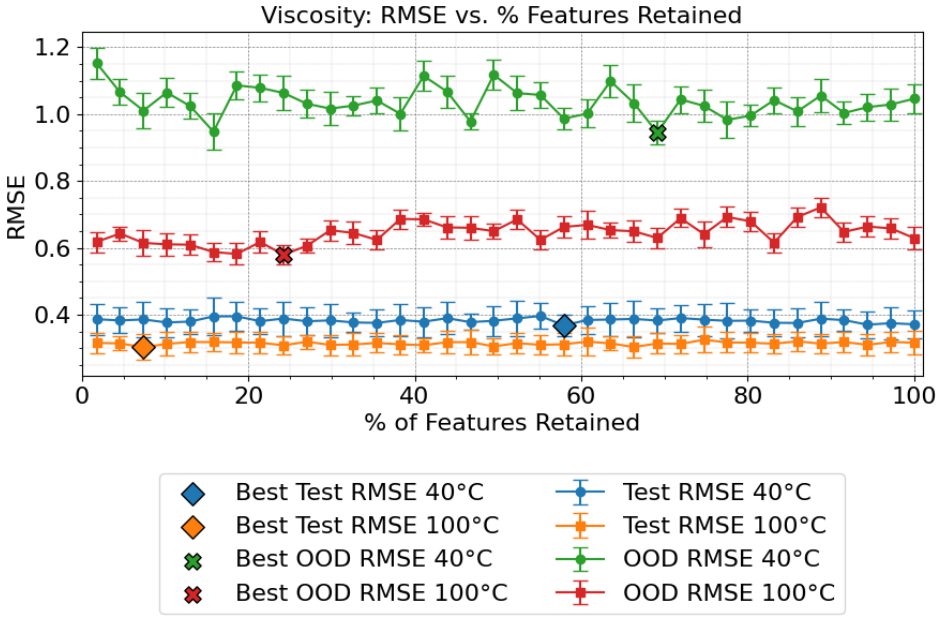


Figure 5.11: Out-of-distribution and test RMSE for log viscosity as a function of the percentage of retained features for both temperatures. The best-performing models for each RMSE type are highlighted.

As shown by Fig. 5.11, the optimal number of features to retain varies greatly depending on the temperature and whether test or OOD RMSE is the primary performance metric. However, across most of the feature retention range (above 10%, or around 90 features), the RMSE values remain relatively stable. This suggests that the majority of descriptors used to predict viscosity on molecules similar to that in the

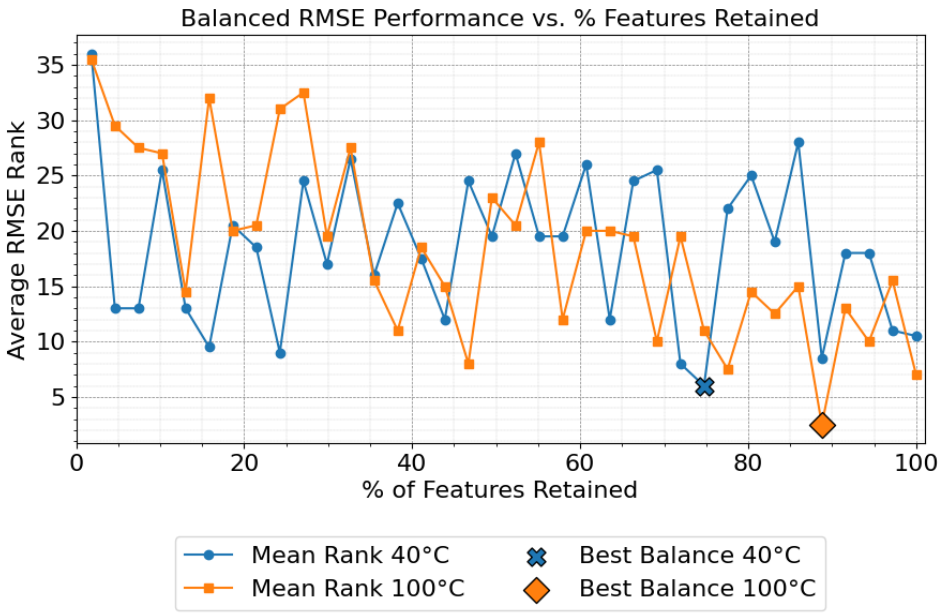


Figure 5.12: Average rank of OOD and test RMSE across different feature retention levels for 40°C and 100°C. Lower rank indicates better performance. The lowest rank shows the optimal feature subset balancing both generalisation and internal accuracy.

experimental dataset provide redundant information, such that removing a moderate number of features does not degrade model performance. A notable drop in both test and OOD performance is observed when fewer than 10% of features are retained, suggesting that at this extreme, critical descriptors likely begin to be excluded, leading to information loss and underfitting.

The optimal tradeoff between feature number and generalisation is achieved when retaining between 75% and 85% (between 660 and 760) of the features, as shown by Fig. 5.12. Fig. 5.11 makes it clear to see how OOD prediction performance is much worse than in-distrubution performance with a 100-200% increase in RMSE for both temperatures.

To investigate the causes of this significant drop-off in performance, we decomposed predictive performance via a parity plot showing the R^2 and RMSE performance of the models across molecules with higher viscosities and different atom types. As shown by Fig. 5.13, there is a large discrepancy between both the R^2 and RMSE values for the high viscosity and different atom groups for at both temperatures. The significantly worse performance of these models when predicting the viscosity at 40°C of high-viscosity molecules (RMSE=1.15, R^2=–5.98), compared to molecules with different atom types (RMSE=0.88, R^2=0.40), suggests that the trained model is more sensitive to the range of target values it sees during training than to OOD features. Upon visual inspection of the parity plot, it is clear that the trained model exclusively underestimates viscosity values for high-viscosity molecules. This indicates that extrapolation beyond the target variable domain may be more challenging for the model than generalising across chemical structure. At 100°C, the model achieves a lower RMSE for high-viscosity molecules (0.54) than for the different atom group (0.71), but the corresponding R^2 value is much worse (–1.54 compared to 0.32). This suggests the model may be consistently biased or fails to capture vari-

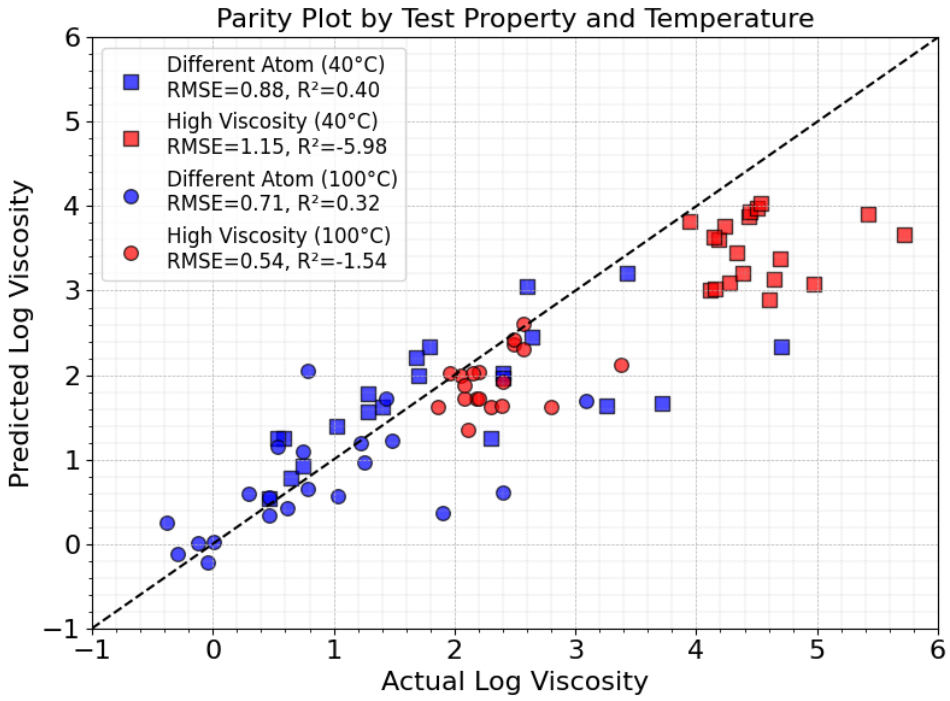


Figure 5.13: Combined parity plot showing predicted vs. actual log viscosity values for OOD molecules across two temperatures (40°C and 100°C). Marker shape distinguishes temperature (squares for 40°C, circles for 100°C), while color indicates test property category. Each model is trained using the optimal stable features from recursive feature elimination (RFE). RMSE and R^2 values reflect prediction performance for each temperature-specific model on its respective OOD subset.

ance within the high-viscosity subgroup, leading to large penalties in the R^2 score despite seemingly small prediction errors. In other words, this model performs well on this dataset in terms of RMSE as there is lower variance in the test set at this temperature.

Fig. 5.14 shows how absolute error varies with Mahalanobis distance. As shown by Fig. 5.14, 38 of the 40 OOD molecules are correctly identified as OOD using this distance metric. When stratifying the OOD molecules by category, a clear distinction emerges between high-viscosity molecules and those containing atom types absent from the training set. The high-viscosity molecules exhibit substantially lower Mahalanobis distances despite being harder to predict, while molecules with different atoms are effectively flagged with greater confidence as OOD by the Mahalanobis metric. This implies that Mahalanobis distance is more sensitive to shifts in feature distributions tied to atomic composition than to shifts in the distribution of the target variable. In other words, descriptors related to atom type dominate the covariance structure, leading to low Mahalanobis distances for high-viscosity molecules that share similar molecular stuctures with the training set, despite their extreme property values.

Counterintuitively, the results reveal no strong positive correlation between Mahalanobis distance and absolute prediction error. This suggests that molecules deemed further from the training distribution in descriptor space are not necessarily harder for the model to predict accurately. A reason for this could be that the prediction error is driven by the relationship between molecular structure and property in ways

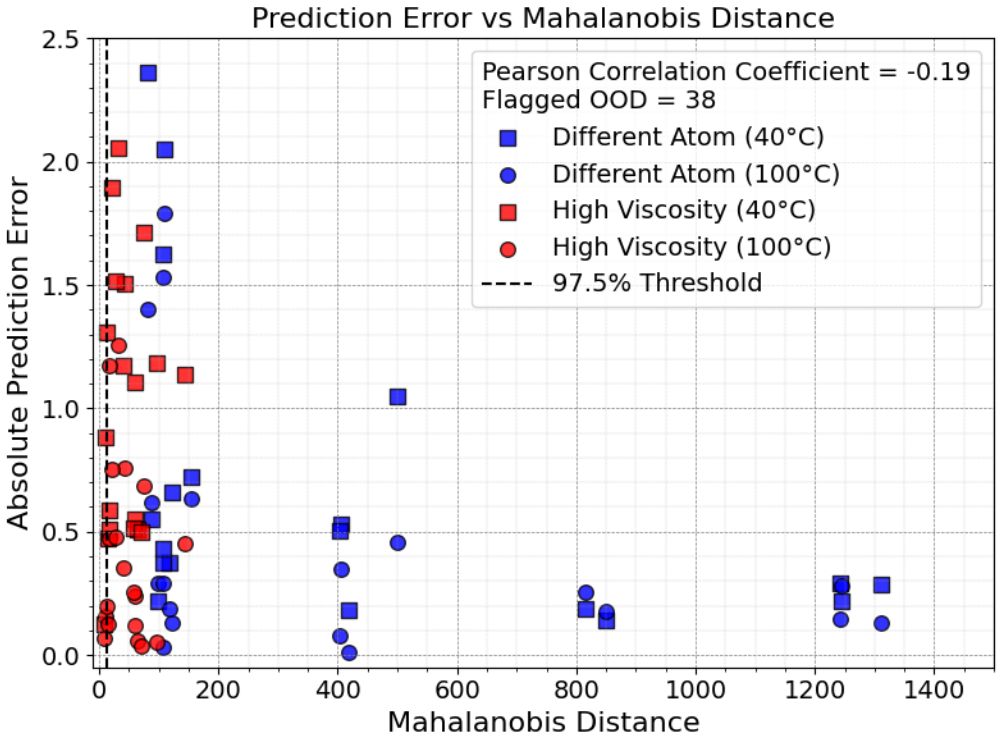


Figure 5.14: Relationship between Mahalanobis distance and absolute prediction error for the OOD dataset when predicting log viscosity at 40°C and 100°C. Molecules are color-coded by test property and flagged as OOD using a 97.5% Mahalanobis distance threshold. A Pearson correlation coefficient is included in the legend to assess association between Mahalanobis distance and absolute error.

that do not align linearly with descriptor variance. It must be noted that these findings may be influenced by the limited size of both the training and OOD datasets, which can lead to unreliable estimates of the feature covariance matrix, reduced descriptor coverage of chemical space, and potentially unstable Mahalanobis distance measurements that do not generalise well across under-represented molecular types. In practice, the molecular descriptors used in our study most likely do not conform to a multivariate Gaussian distribution, limiting the reliability of Mahalanobis distance alone. To improve robustness of OOD detection and quantification, future studies could pair Mahalanobis distance with a non-parametric kNN approach, where the average Euclidean distance from each query molecule to its k-nearest neighbours in a standardized (normalised) descriptor space is constructed from the training set. Molecules exhibiting mean distances exceeding a chosen threshold could then be defined as OOD.[624]

5.4 Conclusions

In this study, a range of ML methods and molecular representation combinations are evaluated to determine the most performant combination for predicting hydrocarbon transport properties. The performance of the best model-representation pair were compared to MD simulations using different force fields. The feasibility of using datasets generated from MD to train ML regression models was also assessed. Finally, the limitations of these ML models were explored by evaluating their performance on

molecules that lie outside the distribution of the training data.

Our evaluation of embedding and fingerprinting strategies highlights the importance of representation–model compatibility in molecular property prediction. PCA-based embeddings outperformed UMAP when used with tree-based models like XGBoost, likely due to their alignment with axis-aligned split strategies and their emphasis on global variance. UMAP embeddings proved more effective for MLPs in most cases, indicating their utility in capturing nonlinear, locally structured manifolds that deep learning models can exploit. Nevertheless, XGBoost models consistently outperformed MLPs across both embedding types, confirming their robustness across representation spaces. Similarly, fingerprinting experiments revealed that hashed ECFP performance improves with larger radii and bit sizes despite increased bit collisions, with XGBoost models again outperforming MLPs. Transitioning to unhashed fingerprints eliminated collision-induced ambiguity, further improving model performance and interpretability. These findings support the use of PCA-reduced descriptors and unhashed extended fingerprints for robust, interpretable, and generalisable ML models in the context of hydrocarbon transport property prediction.

Our benchmarking study confirms that tree-based ensemble models consistently deliver optimal predictive accuracy across a range of thermophysical properties when paired with molecular descriptors. The observed trends in the study emphasize the importance of representation–model synergy: descriptors excel for viscosity and heat capacity, whereas fingerprints perform better for properties more reliant on structure, such as density. Multitask learning did not surpass single-task performance, likely due to task interference and the overlapping molecular input space. These findings highlight that optimal predictive performance of chemical properties arises from a combination of robust feature engineering, carefully selected model architectures, and domain-informed design choices tailored to specific property types.

We compared the accuracy of ML and MD approaches for predicting the thermophysical properties of hydrocarbons. While MD simulations using commonly used force fields such as LOPLS, COMPASS, AMBER, and SCIPCFF were able to produce reasonable estimates for density, viscosity, and thermal conductivity, the XGBoost models consistently outperformed them in terms of predictive accuracy across all three properties. This performance gap was particularly pronounced for viscosity and thermal conductivity, where the ML models achieved significantly lower mean absolute percentage deviations from experimental values. The poorer performance of MD methods on these benchmark molecules is likely due to the presence of cyclic and substituted ring structures, which are more difficult to simulate accurately using standard force fields that have been predominantly parameterised for linear and branched hydrocarbons.

Our comparisons between MD-based and experimental datasets highlight the limitations of relying on large-scale MD-generated datasets for ML model training in molecular property prediction tasks. The inferior performance of models trained on MD data, despite its scale, suggests that careful attention must be paid to force field selection, molecular diversity, and simulation protocols. These results highlight the importance of high-quality experimental datasets and motivate the development of hybrid approaches that combine experimental data with carefully curated synthetic data for optimal model performance.

Our studies on the OOD performance of the best performing models demonstrates that molecular descriptor-based models face substantial limitations when extrapolating to novel chemical spaces. Mahalanobis distance fails to correlate strongly with absolute prediction error, suggesting that structural dissimilarity does not always equate to predictive difficulty. These results show that prediction errors are more sensitive to shifts in the target property distribution (e.g., high-viscosity compounds) than to atom-type dissimilarity.

This study lays the foundation for leveraging ML methods in combination with MD for the multi-objective optimisation of lubricant base-oils.

Chapter 6

Evolutionary Algorithms and Explainable Machine Learning for the Multi-Objective Optimisation of Lubricant Base Oils

In this chapter, a computational framework that combines interpretable ML and evolutionary multi-objective optimisation to support the design of next generation lubricant base oils is presented, building on the insights gained from the work of Chapter 5 regarding the best ways of leveraging ML for thermophysical property prediction. By leveraging experimental data and domain informed constraints, we explore how molecular structure influences key performance properties and use this understanding to guide the generation of high performing candidate molecules. The framework addresses both interpretability and optimisation, aligning with the broader goals of uncovering structure–property relationships and identifying base oils with optimal tradeoffs across competing design objectives.

6.1 Introduction

Historically, the design of new molecules and materials has been reliant on human intuition and trial-and-error approaches. For many important applications, such as synthetic lubricant base oils, only a relatively small number of molecule types and production processes have so far been utilised. Virtual screening has emerged as a powerful tool with which to accelerate the laborious process of molecular design for these types of applications.

The development of high-performance lubricant formulations remains an important part of research in the pursuit of improved energy efficiency in the automotive sector.[627] These lubricants play a critical role in minimising wear and friction between moving parts, directly impacting the energy consumption and longevity of machinery.[19] The fluids used in these applications are complicated mixtures consisting of blends of base fluids and additives designed to perform specific functions. In addition, increasing concerns about the environmental impact of lubricants, which encompass both their production and their

subsequent emissions into the environment, have prompted increased demand for the design and production of formulations from sustainable feedstocks that offer biodegradability and reduced environmental toxicity without compromising performance.[2]

The emergence of electric vehicles requires the development of new lubricant formulations tailored to their unique requirements.[1] Effective lubrication of electric vehicle drive units presents unique challenges in terms of lubricant formulation. An electric vehicle drive unit lubricating fluid is required to provide adequate protection to gear and bearing surfaces under high-torque, low-speed operating conditions, while minimizing churning losses at high speeds and low torques.[628] The former requires a high viscosity fluid, and the latter a low viscosity one. Furthermore, in most modern electric vehicles the same fluid is used to lubricate the gearbox and cool the electric motor, which imposes additional requirements on the fluid of having a high specific heat capacity, suitable thermal conductivity and favourable copper compatibility.[629] Although existing hydrocarbon-based oils may excel in some of these aspects, such as offering superior surface protection, they may no longer offer the optimum combination of properties to adequately satisfy this whole set of requirements.

Until recently, lubricant development has been mostly evolutionary, with empirical testing coupled with incremental adjustments to existing compounds being the method of choice for the improvement of overall performance. Such top-down methods are slow and costly, which means that novel techniques are required to keep up with the pace of technological advancement. To alleviate these drawbacks, more modern approaches to lubricant design have sought to incorporate various computational methods to design lubricants and/or predict their properties on the molecular scale.[3,29,503,630,631] The concept of computed-aided molecular design (CAMD) has been around for decades,[632] however, significant increases in the capability and accessibility of high performance computing, coupled with the development of novel cheminformatics and artificial intelligence algorithms has seen a surge in attempts to use CAMD methods to reduce the cost and time requirements of developing improved lubricant formulations.

Force field development and innovation has also played a key role in the widespread adoption of CAMD techniques. Since the force fields developed by Allinger et al. (MM2[633] MM3,[634] MM4[635]) in the latter stages of the 20th century with a focus on characterising small hydrocarbons, a plethora of all-atom (e.g., Drieding,[636] AMBER,[147] UFF,[637] CHARMM,[148] PCFF,[151] OPLS,[149] COMPASS,[150] REAXFF,[155]) united-atom (e.g., TraPPE,[638] Potoff,[639]), and coarse-grained (e.g., MARTINI[640]) force fields have been developed, allowing for the study of increasingly complex systems with MD simulations. It has been shown that the use of an accurate force field is critical to predicting the properties of long chain molecules such as those used in lubricants. A study by Ewen et al. comparing various open-source all-atom and united-atom force fields found that L-OPLS-AA[601] produced the most accurate estimates for the viscosity of hexadecane.[139] A more recent study by Schmitt et al. investigating the accuracy of various force fields in predicting the thermophysical properties (density, viscosity, diffusivity) of long linear and branched alkanes (n-decane, n-icosane, n-triacontane) found that COMPASS outperformed L-OPLS-AA in terms of prediction accuracy for thermal conductivity and viscosity. The study also noted that COMPASS was only marginally more expensive computationally than L-OPLS-AA and returned predictions that varied less from experimental data over several runs.[105]

Developments in equilibrium molecular dynamics (EMD) and non-equilibrium molecular dynamics (NEMD) techniques have further enabled the ability to predict the physical and chemical properties of molecules and their interactions at the molecular level. EMD has been used in several studies to predict important lubricant properties such as viscosity,[541,641,642,643] thermal conductivity,[644,645] diffusivity,[646,647] and pour point.[648] The Green-Kubo and Einstein-Helfand methods are two techniques that have been used extensively for extracting transport properties, such as viscosity, thermal conductivity and diffusivity from both equilibrium and non-equilibrium MD simulations at different scales[610] The Green-Kubo method relates transport coefficients to the time integral of the auto-correlation function (ACF) of a flux associated with the transport process being predicted.[649] In other words, by analysing the spontaneous fluctuations from equilibrium of parameters related to the target transport property. Whereas the Einstein-Helfand method offers an alternative approach by focusing on the mean squared displacement of particles or energy over time.[650] It has been observed that the Einstein method requires shorter trajectories to achieve the same statistical accuracy as the Green–Kubo formula on the meso scale.[651] EMD and NEMD has already been employed to aid the autonomous molecular design of lubricant oil molecules. For example, Chew et al. integrated physics-informed descriptors from molecular dynamic simulations to enhance the accuracy and interpretability of Graph based ML models to predict temperature-dependent viscosities of small organic molecules ($<C_{12}$).[240]

Despite these advances, it must be noted that MD simulations require significant computational resources to calculate properties such as viscosity, particularly at higher viscosities where the autocorrelation function converges slowly.[543] In NEMD simulations, the shear rate must be sufficiently low to achieve the Newtonian regime, often requiring extended simulation durations due to poor signal-to-noise ratios at low shear rates.[15] Furthermore, approaches such as the Green-Kubo method are subject to a large amount of variability because the plateau region of the Green–Kubo integral is difficult to identify unambiguously. Hence, as viscosity significantly increases under high pressures, predicting transport properties at these conditions becomes a complex and computationally demanding challenge.[652] Techniques such as the time decomposition method have been employed to address this issue.[543]

Advances in artificial intelligence (AI), specifically ML, along with their successful applications in adjacent fields such as drug discovery,[653] have inspired the tribology research community to adopt these techniques in their own research and development efforts. While ML offers significant potential for the computer-aided design of lubricant molecules, several drawbacks limit its practical application. A primary challenge is the lack of public availability of large, high quality datasets with full accurate lubricant properties across a range of lubricant families.[29,464,465,466] Datasets that are large enough to extract patterns from using ML often have missing data due to missing measuring equipment for the property in question or due to human error. Recently, Jeong et al. attempted to rectify this by using a graph-based imputation method, with a graph convolution network, to impute the missing data points in a lubrication formulation dataset.[654] The dearth of comprehensive experimental data on lubricant molecules hampers an ML model's ability to sufficiently generalise across the vast lubricant chemical space, often leading

to model overfitting, where the model performs well on training data but poorly on new, unseen data.[655] Furthermore, many ML models employed in cheminformatics also suffer from an inherent lack of integration with domain-specific knowledge, such as chemical valence rules, which are crucial for accurate and reliable molecule design. Combined with the fact that many studies have not interpreted the outputs of these ML architectures in practice,[223] their value as a replacement for traditional methods of lubricant design and characterisation is diminished. Hence, there is a clear need for methods of lubricant design and characterisation that address the aforementioned issues.

Genetic algorithms (GA) are a class of optimisation techniques inspired by the principles of natural selection and genetics. Genetic algorithms have shown great promise in various domains for their ability to search large, complex spaces by iteratively selecting, mutating, and recombining candidates to improve performance according to defined fitness functions over a series of generations.[656,657] Unlike traditional optimisation methods, GAs do not require a comprehensive dataset to initiate the search, making them attractive for applications like lubricant formulation design where empirical data is scarce and/or expensive to obtain. Moreover, GAs offer interpretability through the tracking of a molecule's 'genealogical' information, allowing researchers to understand the evolution of candidate molecules over generations and allowing researchers to retrospectively identify molecule features related to stronger candidate performance. Furthermore, GAs allow for researchers to explicitly define constraints on the generated candidates based on knowledge of the problem domain. Most importantly GAs allow for multi-objective optimization through the combination of the outputs of different fitness functions, allowing researches to directly compare tradeoffs between different properties of interest.

The idea of using GAs for molecular discovery has been around since the 90s. For example, Venkatasubramanian et al. used a weighted-sum genetic algorithm (WSGA), where molecules were represented as bit strings, to explore if the approach could identify an optimal target polymer from a constrained search space according to density, thermal expansion, specific heat capacity and other properties of interest.[658] They found that even with a theoretical search space of $1.1x10^{13}$ candidates, such an approach was still able to find a candidate that matched the performance of an optimised target. More recent studies have combined machine learning, molecular dynamics and genetic algorithms to optimise polymer density,[659] to minimise the energy of a polymer-polymer interface,[660] to optimise glass transition temperature,[661] and to maximise thermal conductivity.[662]

Certain studies have also employed genetic algorithms for optimisation tasks within the field of tribology. For example, Li et al. used a genetic algorithm to optimise the hyperparameters of a neural network trained to predict the friction coefficient and wear rate of nano-TiO_2 reinforced PA6 composites lubricated with solid lubricants, achieving a correlation index between the predicted and the experimental values for friction coefficient and wear rate of 0.992 and 0.998 respectively.[663] Bhaumik et al. used a genetic algorithm to optimise biodegradable lubricant formulations comprising vegetable oils base stocks with different nano-friction modifiers, using a neural network trained to predict a formulation's coefficient of friction as the fitness function.[664] Zhang et al. used a genetic algorithm to optimize the shape of concave surface textures under reciprocating sliding motion for improved friction coefficient, using a modified average flow Reynolds equation as a fitness function.[665]

Although GAs have been used in several optimisation tasks within the field of tribology, there have been no studies utilising them in conjunction with ML to directly optimise lubricant base oils. To address gap, we propose a novel approach that combines GAs with explainable ML for the multi-objective optimisation of lubricant base oil molecules. This approach incorporates key properties such as viscosity, thermal conductivity, viscosity index, synthetic complexity and molecule toxicity into the fitness function, with post-optimisation predictions of biodegradability for the best performing molecules. Through this approach, we seek to demonstrate the feasibility of genetic algorithms in optimising lubricant molecules.

6.2 Methodology

6.2.1 Genetic Algorithm Implementation

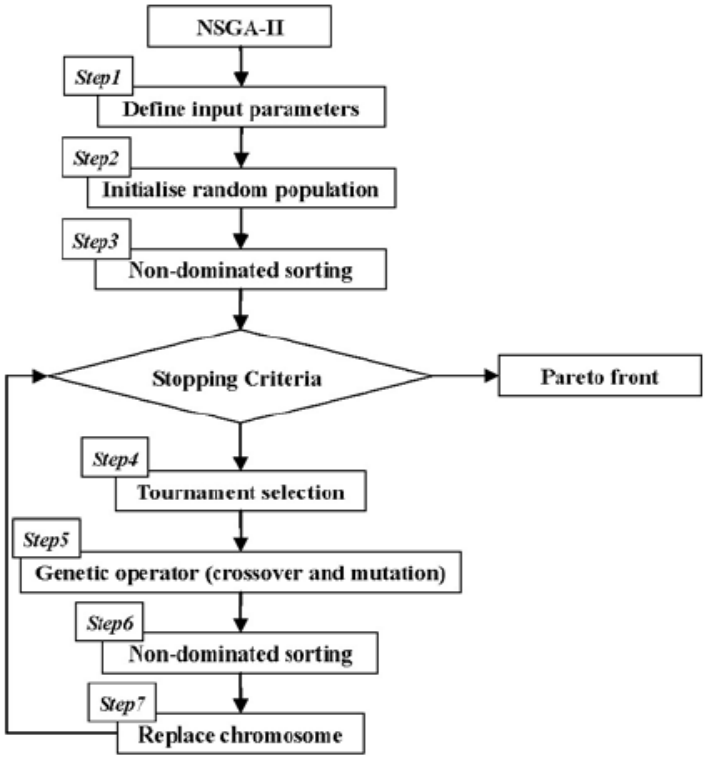


Figure 6.1: Overview of Genetic Algorithm Process

Figure 6.1 illustrates the steps followed by genetic algoritms such as NSGA-II. The first step of any GA involves the initialisation of a population of candidates. Each individual in a population is evaluated using one or several fitness functions which assess how well the individual meets the desired criteria. Based on their fitness scores, the best performing candidates in the population are selected to be passed on to a subsequent generation unchanged and for combination with other candidates to produce the remaining candidates for the subsequent generation. The selected 'parents' undergo crossover operations, exchanging segments of their 'genetic information' to produce offspring with new feature combinations. To maintain genetic diversity and prevent premature convergence, mutation operations may be performed on offspring, introducing random changes to offspring. The new generation of can-

didates, including those from crossover and mutation, replaces the old population, the fitness function is reapplied to the new generation, and the process repeats. This cycle of fitness evaluation, selection, crossover, mutation, and replacement continues until a predefined termination condition is met, such as a set number of generations or other convergence threshold.[666]

Each candidate in a population of a GA needs to be represented computationally in a way that allows for efficient crossover and mutation. The most common representations for GAs are molecular-based strings such as SMILES[667] and SELFIES[668] or molecular graphs.[669] In this study, canonical SMILES[573] is used to represent candidates as it provides a concise and unambiguous representation of molecular structures, allowing for straightforward encoding, mutation, and crossover operations. Additionally, the widespread adoption of SMILES in cheminformatics allows for the integration with existing chemical databases and software tools.

Genetic Algorithm Hyperparameters - Population Initialisation, Mutation Rate, Population Size, Number of Generations, Elitism and Parent Selection

When applying GAs for molecular generation and optimisation, several key hyperparameters can significantly influence the quality and diversity of the resulting chemical space. The population size (or generation size) defines how many molecules are evaluated per generation. The mutation rate controls how frequently random changes are introduced into candidate molecules. Elitism determines how many of the top-performing solutions are carried forward unchanged to the next generation. Careful tuning of these hyperparameters are essential to achieve an effective tradeoff between exploration and exploitation of the objective space.

A previous study has investigated the effects of the above hyperparamters for a single-objective optimisation problem.[670] However, the findings from this study may not directly translate to multi-objective optimisation frameworks such as NSGA-II or WSGA. The NSGA-II implementation in this study relies on Pareto-based ranking and crowding distance calculation, which may be sensitive to different population dynamics and selection pressures (e.g., a lack of Pareto-front selection pressure due to population size being smaller than the number of non-dominated solutions).[671] Similarly, the effectiveness of mutation and elitism rates in WSGA optimisation may depend on the chosen weighting scheme for each objective and the shape of the underlying Pareto front. Therefore, to identify the most effective configuration, we performed a grid search over relevant hyperparameter ranges for each GA scheme, where the maximum number of generations is set to 200. For NSGA-II, the hypervolume (HV) indicator is used to evaluate and compare the quality of generated solutions of the elite in each generation.[178,179] For the weighted-sum method, we tracked the average niched score for the elite portion of each generation's population.[672]

Parent selection, where candidates from the current population are chosen to produce offspring for the next generation, is another important component of GAs. A plethora of selection strategies exist for parent selection including k-way tournament,[187] roulette wheel,[188], stochastic universal sampling,[673] direct ranking (according to fitness score), and random. K-way tournament selection works by randomly selecting a subset of k individuals from a population and evaluating their fitness, where the fittest in-

dividual from each subset is chosen as a parent. This is repeated until the desired number of parents, normally two, is selected. A study reviewing the effect of GA parameters for optimising heterogeneous catalysts determined that between roulette wheel, ranking, threshold, and tournament selection methods, three-way tournament selection performed the best.[674] These findings were corroborated by a study by Greenstein et al.[670], which found that three-way tournament selection demonstrated the best balance between coverage of the chemical space and speedup of the GA when compared with other selection strategies, hence three-way tournament selection is used to select parents for offspring generation in this study.

There are two widely used strategies for selecting the initial population of candidates: seeding the initial population with good solutions,[666] or creating the initial population from a random selection of candidates within the search space.[675] Seeding the initial population with known good solutions may promote faster convergence towards an optimised solution similar to the initial population, however it may discourage the GA from exploring counter-intuitive candidates with desired properties. To avoid the GA exploring a limited section of the search space, while still allowing the GA to converge to an optimal solution quickly, the initial population used for the GA in this study is generated by performing random mutations to molecules that were part of the dataset used to train machine learning models.

Promoting Population Diversity through Molecular Similarity Calculation

NSGA-II uses crowding distance to promote population diversity by estimating the density of solutions around a given individual in objective space, favouring those located in sparsely populated regions to encourage a well-distributed Pareto front. To ensure that each objective contributes equally to this calculation, we normalise all objective values to fall within the range 0 to 1. Unlike NSGA-II, WSGA multiple objectives into a single scalar value, which eliminates the ability to explicitly assess solution spread in objective space, hence, alternative methods are needed to promote diversity and avoid premature convergence to a narrow region of the Pareto front. To prevent WSGA stagnation and premature convergence towards a single local minimum as mentioned above, niching is employed, where the formation of sub-populations is encouraged.

There are many methods for implementing niching, including clearing,[676] restricted tournament selection,[677] and speciation.[678] However, these methods suffer from excessive sensitivity to parameter selection and/or a requirement of subjective input, such as in the case of speciation, where sub-populations (niches) must be defined explicitly. Hence, the fitness sharing method, where the fitness score of candidates that are too similar to each other in terms of their molecular structure is adjusted, as proposed by Goldberg et al.[189] is used in this study. The fitness sharing function adjusted fitness score $f_i^{'}$ of a candidate i with original fitness score f_i is calculated as:

$$f_i^{'} = \frac{f_i}{m_i} \tag{6.1}$$

where m_i is calculated as:

$$m_i = \sum_{i=1}^{N} S(d_{ij}) \tag{6.2}$$

where N represents the population size, d_{ij} is a representative value for the contextual 'distance' between candidate i and candidate j (e.g., molecular structure), and $S(d)$ represents the sharing function, used to calculate the similarity between two candidates in a generation.

Several methods have been proposed for determining molecular similarity as this is an indispensable element of cheminformatics, with the majority of these methods based on molecular fingerprints as inputs.[679] Topological fingerprints,[680] structural keys such as MACCS keys,[566] and circular fingerprints such as ECFPs,[568] are popular fingerprinting techniques used widely in the literature for this purpose.[681] Circular fingerprints such as ECFP have been shown to perform consistently and outperform topological fingerprints and structural keys in representing the neighbourhood principle, stating that molecules within the same locality in structural space, have a higher chance of exhibiting similar properties.[682] Hence, ECFPs are chosen to encode molecular structures for similarity comparison. A study by Lee et al.[570] found that unique fingerprints can be generated for 100% of molecules in a dataset with a radius of 8, hence, a radius of 8 is used in this study to generate molecule ECFPs for molecular similarity calculations. Measures such as the Tanimoto coefficient, Dice coefficient and Cosine coefficient have been shown to be optimal for fingerprint-based similarity calculations.[683] The Tanimoto coefficient method is chosen as the sharing function due to its simple implementation, compatibility with molecular fingerprints and directly interpretable output format. The Tanimoto algorithm describes sets A and B as collections of fingerprint bits from the fingerprints of molecules A and B, respectively. The set AB consists of the bits that are common to both fingerprints. The Tanimoto coefficient ranges from 0, indicating that the fingerprints share no common bits, to 1, indicating that the fingerprints are exactly the same and is calculated as:

$$T(A,B) = \frac{c}{a+b-c} \tag{6.3}$$

where a is the number of bits set to 1 in the fingerprint of molecule A, b is the number of bits set to 1 in the fingerprint of molecule B and c is the number of bits set to 1 that are common to both fingerprints A and B.

Table 6.1 outlines the mutation operations used in the GA, along with the conditions that govern their application. Certain mutations are restricted to candidates above a certain number of heavy atoms, in order to prevent uncontrolled growth of molecule size. Conversely, other mutations are only applied to smaller molecules to encourage exploration of the molecular space in the intermediate molecular size range.

Mutations

Figure 6.2 shows the crossover protocol used in the offspring generation phase of the GAs. As shown by Figure 6.2, after two parents are chosen, each of the two parent molecules are fragmented at random points into two sub-fragments, with the largest sub-fragment from each parent used for combination. Sub-fragments are combined by creating bonds between two randomly selected atoms on each sub-fragment. Combined sub-fragments may then be subjected to one of six mutations: Atom Addition, Atom Removal, Atom Replacement, Bond Replacement, Fragment Addition or Fragment Removal.

Mutation	Conditions
Add Atom	Heavy atoms ≤ 40
Remove Atom	N/A
Replace Bond	N/A
Replace Atom	N/A
Add Fragment	Heavy atoms ≤ 40
Remove Fragment	Heavy atoms ≥ 40
Replace Candidate	Heavy atoms ≥ 40
Crossover	N/A

Table 6.1: Mutation types used in the GA with associated conditions and probabilities.

The atom removal and atom replacement mutations involve choosing an atom to be removed/replaced at random, where bond order between neighbouring atoms is maintained. With the atom replacement mutation, a bond in the target molecule is chosen at random and removed, an atom type is chosen at random from the list of potential atoms, then bonds of a random order are created between the atom(s) connected to the removed bond. The bond replacement mutation involves a bond in the target molecule being chosen at random and changed to a different bond type. The fragment addition process involves choosing a fragment for insertion from a predefined database of molecular fragments, where the chosen fragment has no more than one-third the number of heavy atoms as the target molecule, to prevent candidate molecules becoming too large too quickly in the GA process. A bond is then created between randomly chosen atoms on both the initial molecule and the chosen fragment. The database of molecular fragments used for the fragment addition (and removal) mutation was determined by choosing fragments present in the molecules of the collated dataset used in Chapter 5. The fragment removal process involves either random fragmentation of a target molecule into 3 sub-fragments, where the two largest sub-fragments are chosen for recombination by creating a bond between two randomly selected atoms on each sub-fragment, or removal of a targeted fragment. In addition to the six mutation operations, we include a 'Replace Candidate' operator that randomly replaces a non-elite candidate within a population with a molecule from the collated dataset used in Chapter 5, thereby reintroducing evolutionary paths originating from known molecules.

The probability that any combination of two sub-fragments undergoes a further mutation is determined by the mutation rate. Rule-based mutations on SMILES strings offer a simple, interpretable, and chemically valid way to explore chemical space without the need for complex molecular representations or large datasets. As opposed to operating on canonical SMILES strings, operating directly on graph-based or 3D representations may require additional computational overhead and chemical validation to ensure resulting structures are valid molecules. Moreover, while ML-based generative models (e.g., transformers, graph neural networks, diffusion models) can produce novel candidates, they often require large amounts of curated training data and may lack the interpretability and control needed for targeted chemical modifications. Rule-based SMILES mutations also provide more transparency in the evolutionary process and allow domain knowledge to be embedded directly into the mutation logic, which can be advantageous when searching for specific structural features or optimizing molecules for known properties. The calculation of ECFPs, Tanimoto coefficients and the implementation of molecule crossover

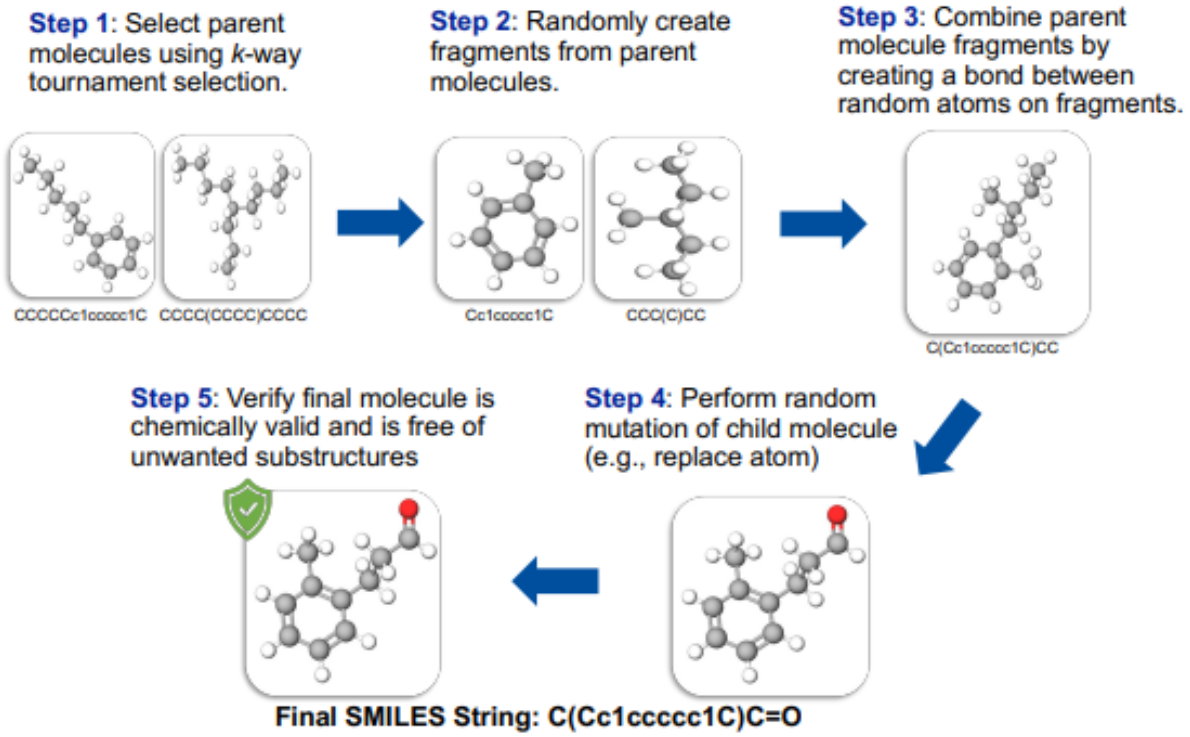


Figure 6.2: Overview of Molecule Crossover Process.

and mutation are performed using custom Python functions built with RDKit.[571]

6.2.2 Genetic Algorithm Fitness Functions

Viscosity Index

The viscosity index (VI) is a dimensionless number that indicates the change of a lubricant's viscosity with temperature, used to quantify the degree to which a lubricant's viscosity changes as the temperature changes. A higher viscosity index indicates that the lubricant's viscosity changes less with temperature fluctuations, which is desirable in most applications.[684] The most recognised method of calculating VI is according to the American Society for Testing and Materials (ASTM) D2270 standard.[685] According to this measure, VI is determined using the following equation:

$$VI_{ASTM} = \begin{cases} 100 \times \frac{L-\eta_k^{40^oC}}{L-H}, & \eta_k^{40^oC} > H \\ \\ \frac{\text{antilog}(N)-1}{0.00715} + 100, & \eta_k^{40^oC} < H \end{cases} \tag{6.4}$$

where:

$$N = \frac{\log(H) - \log(\eta_k^{40^oC})}{\log(\eta_k^{100^oC})} \tag{6.5}$$

where, $\eta_k^{T^oC}$ is the kinematic viscosity of the target oil at at temperature T, L is the kinematic viscosity at 40°C of a reference oil with a viscosity index of 0 with the same kinematic viscosity at 100°C as

the target oil and H is the kinematic viscosity at 40°C of a reference oil with a viscosity index of 100 with the same kinematic viscosity at 100°C as the target oil (all in mm^2/s). Reference oil values are obtained from the ASTM 2270 specification[685] and linear interpolation is used to obtain L and H values when $\eta_k^{40^oC}$ and/or $\eta_k^{100^oC}$ is between tabulated reference values. $ASTM_{VI}$ is undefined for molecules with a kinematic viscosity $\leq$2.0 mm^2/s, and for low viscosity molecules, this measure is known to underestimate VI.[686,687] As a result, the KVI will often be undefined for many important base oils. This is illustrated by Figure Fig. 6.3, which plots the relationship between KVI and DVI for the oils in the dataset used to train the XGBoost models. As shown by Fig. 6.3, many paraffins have an undefinded KVI (molecules with undefined KVI plotted as having a KVI of 0), illustrating the significant limitation of this measure. Hence, the Dynamic Viscosity Index (DVI), as proposed by Kajita et al. [29] is calculated for each molecule, using the following equation:

$$DVI = 220 - 7 \times 10^S \tag{6.6}$$

where:

$$S = -\log_{10}\left(\frac{\log_{10}(\eta^{40^oC}+1.2}{\log_{10}(\eta^{100^oC}+1.2}\right)/\log_{10}\left(\frac{27}{35}\right) \tag{6.7}$$

where η^T is the dynamic viscosity at temperature T, related to the kinematic viscosity through $\eta_k = \eta/\rho$, where ρ is the density of the liquid at temperature T. As shown by Fig. 6.3, DVI still allows for comparison of molecules that are undefined by the KVI measure.

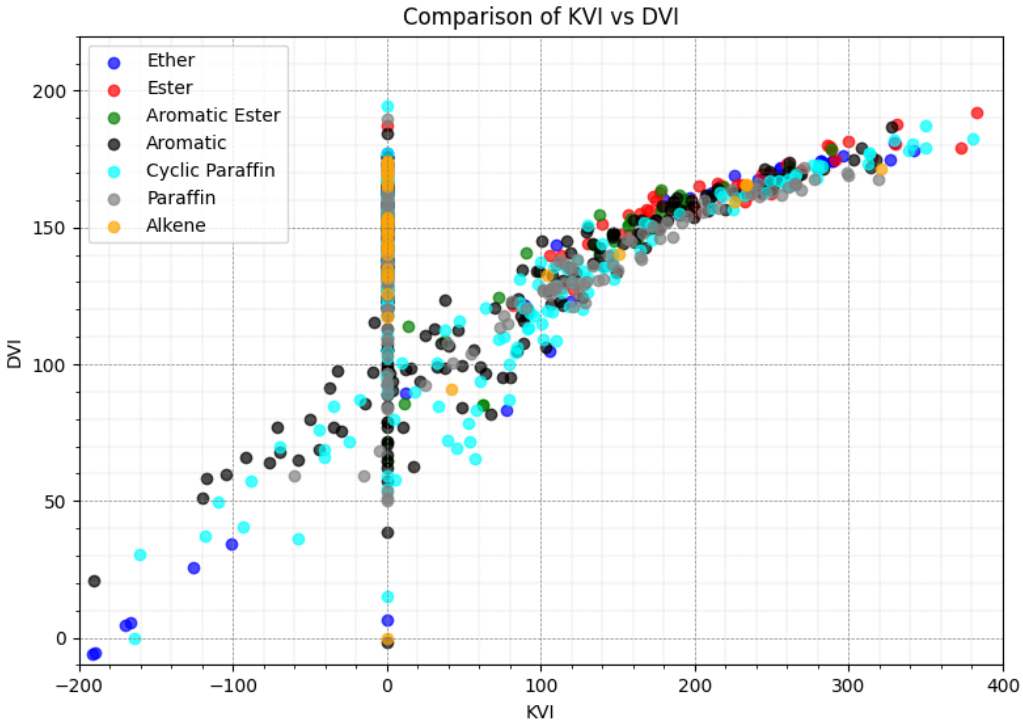


Figure 6.3: KVI vs DVI values for oils in the NIST experimental dataset.

Synthetic Complexity

Synthetic complexity is often overlooked as a property of interest in de-novo molecular design studies, often leading to the design of theoretically promising molecules that are impractical or impossible to synthesize. Synthetic complexity is considered to be broad, intuitive and subjective concept, meaning there is no widely accepted metric to quantify it.[688] Several studies have attempted to quantify synthetic complexity using various approaches, including deriving scores a small number of expert-selected structural descriptors,[689] deriving scores calculated from automated retro-synthetic analysis,[690] and calculating scores using fragment or substructure-level contributions, such as the Sythetic Accessibility (SA) score proposed by Ertl et al.[691] However, these methods all suffer from a reliance on explicit information about the entire sequence of reactions required to synthesize the compound, subjective rankings/scores provided by expert chemists, and/or comparison to a fixed set of purchasable compounds.[692] Furthermore, the SA score is also limited by its reliance on fragment frequency and structural heuristics, which can underestimate the feasibility of novel yet synthetically accessible molecules and over-penalise uncommon substructures not well represented in known compound databases.[691] Hence, in this study, synthetic complexity of generated candidates was evaluated by calculating the SCScore metric proposed by Coley et al.[692] using a neural network trained on 12 million reaction pathways from the Reaxys database.[693] The SCScore metric was chosen as it provides an easily interpreted score, compatibility with canonical SMILES, applicability to molecules from several domains, and ease/speed of calculation. The SCScore metric defines synthetic complexity as follows: if a synthetic route from compound A to compound C involves an intermediate B, then the reaction steps A $\rightarrow$ B and B $\rightarrow$ C should both be regarded as productive steps, which is reflected by an increase in synthetic complexity. SCScore takes the form of a numerical value linearly scaled to (1,5), with a score of 1 representing a molecule with low synthetic complexity, and a score of 5 representing the opposite.

Machine Learning Models for the prediction of Density, Viscosity, Heat Capacity and Thermal Conductivity

To evaluate viscosity, density, thermal conductivity and heat capacity of generated candidates, the best performing models from Chapter 5 were used. Namely, XGBoost models with molecular descriptors were trained using the optimised hyperparameters trained during grid search in the previous section. To improve computational efficiency, RFE was performed to remove noisy and redundant features. Fig. 6.4 shows the effect of RFE on the predictive performance of the trained models. The optimal number of features as determined by the RFE study were used to predict each property.

Toxicity

Toxic lubricants can have significant environmental and health impacts, leading to potential harm to humans, wildlife, and ecosystems.[694] Additionally, regulatory requirements and sustainability goals necessitate the development of lubricants that are safe and environmentally friendly.[695] Therefore, evaluating and minimizing toxicity is an essential aspect of designing safe lubricants. In this study, candidate molecule toxicity is predicted using an attention-based multiscale convolutional neural network pro-

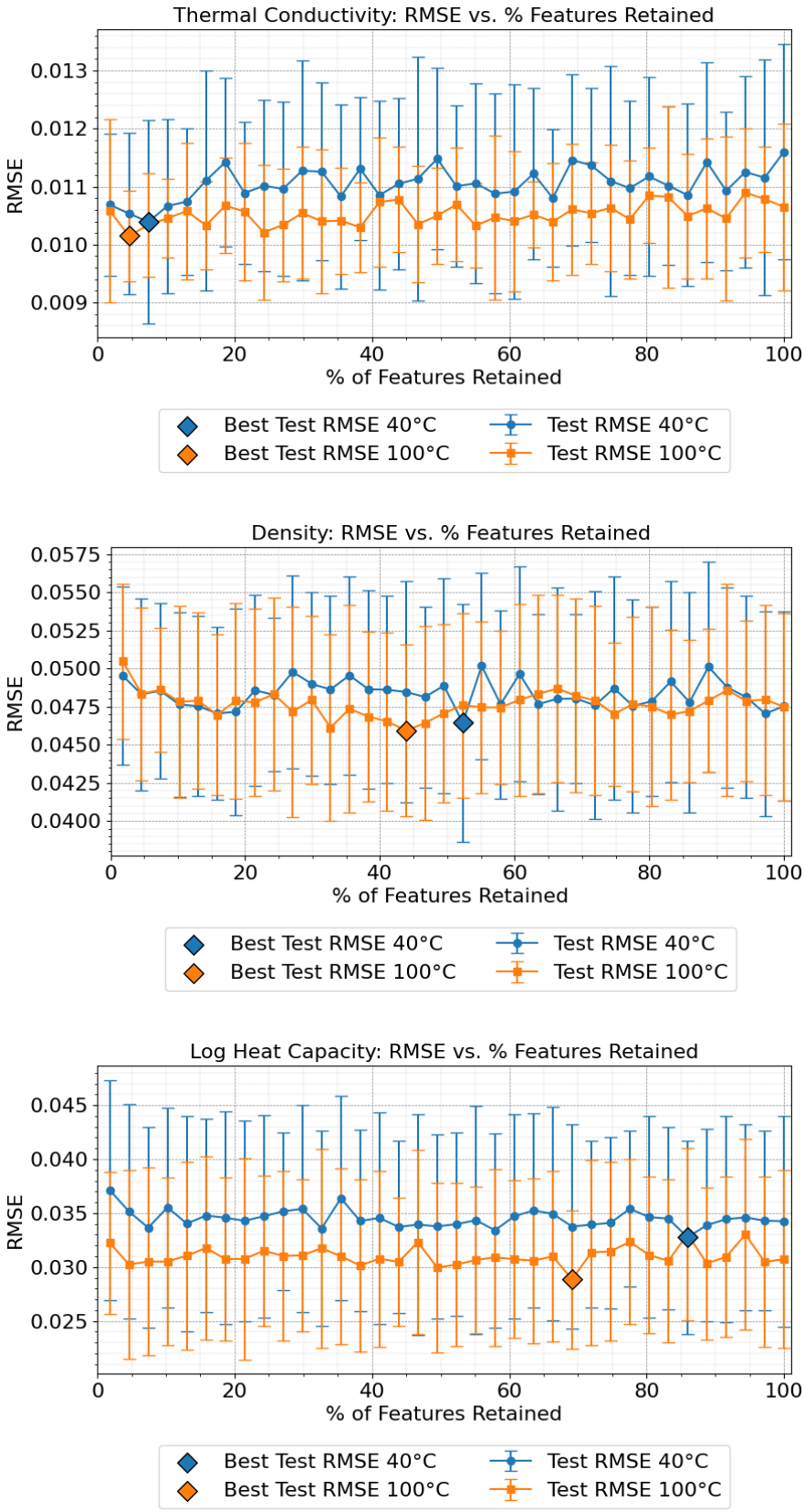


Figure 6.4: Test RMSE versus percentage of features retained for Thermal Conductivity, Density, and Log Heat Capacity at 40°C and 100°C. Error bars denote one standard deviation.

posed by Born et al.[696] This model was trained on over 12000 entries from the Tox21 dataset[697] related to a molecule's harmful affects on 12 targets related to hormone and stress response pathways, returning a score between 0 and 1 for each of the 12 targets, where 1 means toxic and 0 non-toxic. An average of each of the 12 target toxicity scores is taken to provide a single representative figure of the overall toxicity of the molecule.

Biodegradability

Biodegradability is a critical property for lubricants intended for environmentally sensitive applications, as it ensures that, when released into natural ecosystems, the substances can be broken down by microbial activity, thereby minimizing long-term ecological harm and reducing the risk of bioaccumulation or persistent pollution.[698] One of the most common methods is the OECD 301B test (Modified Sturm Test), which evaluates the percentage of carbon dioxide evolved as a result of microbial activity over a 28-day period, where a lubricant is generally classified as readily biodegradable if it achieves at least 60% mineralisation within this timeframe. Other commonly used biodegradability tests for lubricants include ASTM D5864, ISO 14593, and OECD 310.[699] As experimental tests for assessing biodegradability are often time-consuming and costly, several studies have explored the use of machine learning models to predict or classify molecules as biodegradable or non-biodegradable.[699,700,701] To assess candidates for biodegradability, we train a binary classification XGBoost model on the dataset used in the study by Yin et al. [702] The dataset contains 1306 not-readily biodegradable and 622 readily biodegradable chemicals.

6.2.3 Saliency Study Techniques

Shapley Additive Explanations (SHAP) provide a framework for interpreting the output of machine learning models by attributing predictions to individual input features based on cooperative game theory.[703] Its applicable to a wide-range of machine learning models has made it an indispensable tool for understanding the predictions made by such models in for a wide range of prediction tasks in tribology.[240,704,705,706] Given a model f and an input $x \in \mathbb{R}^d$ (d is the number of features), the prediction $f(x)$ can be decomposed into a sum of contributions from each feature:

$$f(x) = \phi_0 + \sum_{i=1}^{d} \phi_i,$$

where ϕ_0 is the expected model output and ϕ_i represents the marginal contribution of feature i. The Shapley value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right],$$

where S denotes a subset of features excluding i, and f_S is the model restricted to the feature subset S. Beyond attribution to individual features, SHAP analysis can reveal interaction effects, where the contribution of one feature depends on the value of another. By leveraging TreeSHAP, analysis can also be extended to compute SHAP interaction values, which decompose the prediction into main effects and

pairwise interactions:[703]

$$f(x) = \phi_0 + \sum_{i=1}^{d} \phi_{i,i} + \sum_{i=1}^{d} \sum_{j>i} \phi_{i,j},$$

where $\phi_{i,i}$ captures the main effect of feature i, and $\phi_{i,j}$ quantifies the interaction between features i and j. The SHAP interaction value $\phi_{i,j}$ is defined as

$$\phi_{i,j} = \frac{1}{2} \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i,j\}} \frac{|S|!(d-|S|-2)!}{(d-1)!} \left[f_{S\cup\{i,j\}}(x_{S\cup\{i,j\}}) - f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_{S\cup\{j\}}(x_{S\cup\{j\}}) + f_S(x_S) \right].$$

This enables the identification of which features most significantly influence model outputs and also the exploration of feature interdependencies. A core limitation of traditional SHAP analysis is that it assumes feature independence when estimating marginal contributions, meaning it may inaccurately attribute importance when features are highly correlated, leading to misleading or unstable explanations in datasets with strong feature dependencies.[707] TreeSHAP partially mitigates this limitation by incorporating feature dependencies through the structure of tree-based models, leading to more accurate and stable explanations.[708] In this work, we apply SHAP and TreeSHAP analysis to XGBoost models trained on RFE optimized sets of molecular descriptors, and molecular fingerprints. This approach allows us to systematically investigate which features contribute most significantly to the predicted properties of interest, as well as to explore how these contributions are modulated by interactions with other structural or physicochemical features.

6.3 Results and Discussion

6.3.1 SHAP Analysis

To better understand the trained models' behaviour, we first performed standard SHAP analysis for each target property, using beeswarm plots generated from our experimental test set, as suggested by a study suggesting best practices for SHAP studies.[707] Then, we conducted feature dependency analyses focusing on the three most influential SHAP descriptors for each property, along with associated structural features. This helped us identify three key properties that most strongly drive the variation in the top SHAP features for the various features. Definitions of all the molecular descriptors shown in the SHAP plots or mentioned in the text can be found in the Mordred documentation.[576]

Biodegrability

XGBoost classifiers were optimized via grid search across the same hyperparameter space used in the ML model optimisation studies. The dataset was split using a training:test:validation split of 80:10:10. Hyperparameter tuning was performed with three-fold cross-validation on the training set, evaluating models based on using balanced accuracy and Matthews correlation coefficient (MCC)[709] on the test set as in the original study the dataset was obtained from. The best-performing model from this training process achieved a balanced accuracy of 88.7%, with a Matthews Correlation Coefficient of 0.77. The best performing model from the study from which this dataset was taken, a Transformer-CNN algo-

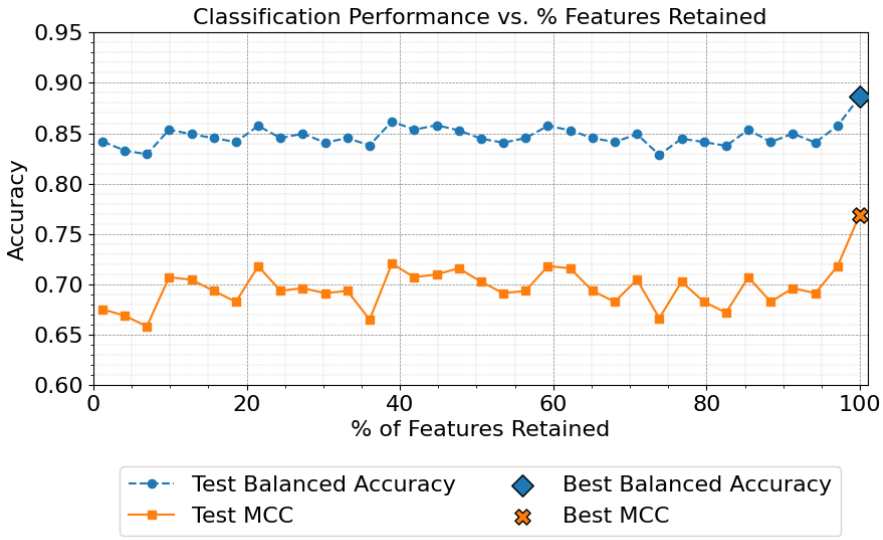


Figure 6.5: Balanced Accuracy and Matthews Correlation Coefficient (MCC) on the test set as a function of the percentage of features retained after RFE.

rithm, achieved a balanced accuracy of 86.29% and a Matthews correlation coefficient of 0.71, using a test set size of 25%.[702] It should be noted that our use of a 10% test set may partially contribute to the observed improvement, although the optimized representation selection and model hyperparameter tuning are likely the primary factors driving the enhanced performance. In the study by Yin et al, they trained a random forest model that achieved a balanced accuracy of 85.29% and a Matthews correlation coefficient of 0.65, suggesting that using a more powerful tree-based method could yield the results that we are seeing. As shown by Fig. 6.5, performing any amount of RFE consistently worsens performance, suggesting that the model relies on the full spectrum of molecular descriptors to classify molecules. Fig. 6.6 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule biodegradability.

As shown by the plot, the XPc-4dv descriptor, which encodes the four-dimensional spatial distribution of electronegativity-weighted atomic properties reflecting both molecular geometry and electronic structure, is the most influential feature for predicting biodegradability. Lower XPc-4dv values are associated with molecules that have more uniform and less polarisable electronic distributions, such as straight-chain alkanes and simple esters, which the model links to increased biodegradability. Conversely, higher XPc-4dv values indicate more complex or spatially clustered electron-rich regions, often seen in branched, cyclic, or aromatic compounds, which the model associates with reduced biodegradability, likely due to increased steric hindrance.

The Mor10 descriptor captures how atomic properties (atomic mass, electronegativity, polarisability, and partial charge) are spatially distributed across the molecule, focusing on interactions between atoms separated by approximately 10Å.[576] Lower Mor10 values are associated with linear structures like straight-chain alkanes, which the model predicts as improving biodegradability, while higher values correspond

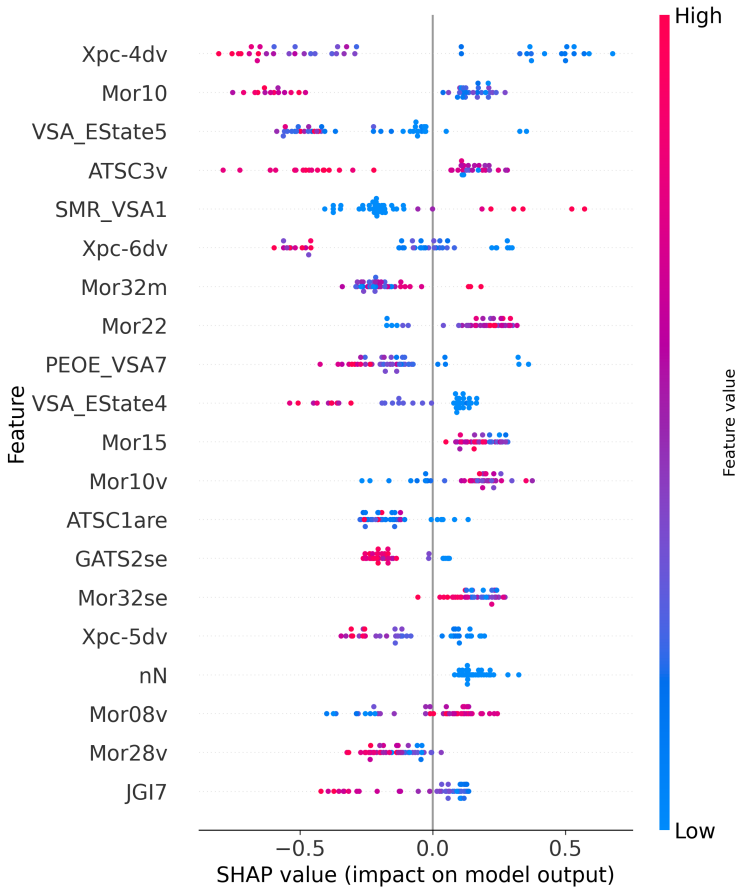


Figure 6.6: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule biodegradability.

to bulkier, more branched or cyclic molecules that the model associates with reduced biodegradability.

The VSA_EState descriptor combines van der Waals surface area (VSA) contributions with electrotopological state (EState) indices. Specifically, it quantifies the surface areas of atoms with intermediate electronic states.[710] A high SHAP importance value for VSA_EState suggests that the spatial distribution and electronic character of surface-exposed atoms strongly influences biodegradability predictions. Lower VSA_EState values generally reflect less electronically differentiated surface areas, as seen in molecules such as n-alkanes, and are often associated with model predictions of higher biodegradability. However, the model also shows that in some cases, low VSA_EState values correspond to non-biodegradable predictions, suggesting that low surface differentiation alone does not guarantee biodegradability. Similarly, while higher VSA_EState values, indicating more electronically complex surfaces, as found in esters or heteroatom-containing oils, are frequently linked to reduced biodegradability, the presence of reactive polar groups in esters may sometimes offset this effect, leading to greater biodegradability. Interestingly, a recent study by Dan et al., found that VSA_EState and Xpc-4dv (as well as SMR_VSA1) were among the top 20 most important features from toxicity prediction, suggesting some crossover between important features for predicting these properties.[711]

For Xpc-4dv, the features SMR_VSA1, AATSC0v, and Mor10 interact complementarily. The complementary interactions suggest that when considered together, these features reinforce spatial and electronic structural patterns that strengthen the model's prediction related to the role of electronegative regions in biodegradability. For Mor10, the features Xpc-4dv, Mor30m, and Mor03m interact complementarily. The complementary interactions imply that mass-weighted and geometric descriptors at multiple spatial scales consistently support the model's ability to use Mor10-related information for predicting biodegradability. Based on these findings, our molecular generation should increase the presence of structures with more uniform electronic distributions and lower spatial complexity such as straight-chain alkanes, to promote lower XPc-4dv, Mor10, and VSA_EState5 values and maximise biodegradability. Conversely, structures leading to higher values of these descriptors, including highly branched, cyclic, or electronically differentiated molecules, should be minimised unless reactive polar groups (such as esters) are deliberately introduced to improve biodegradability.

Dynamic Viscosity Index

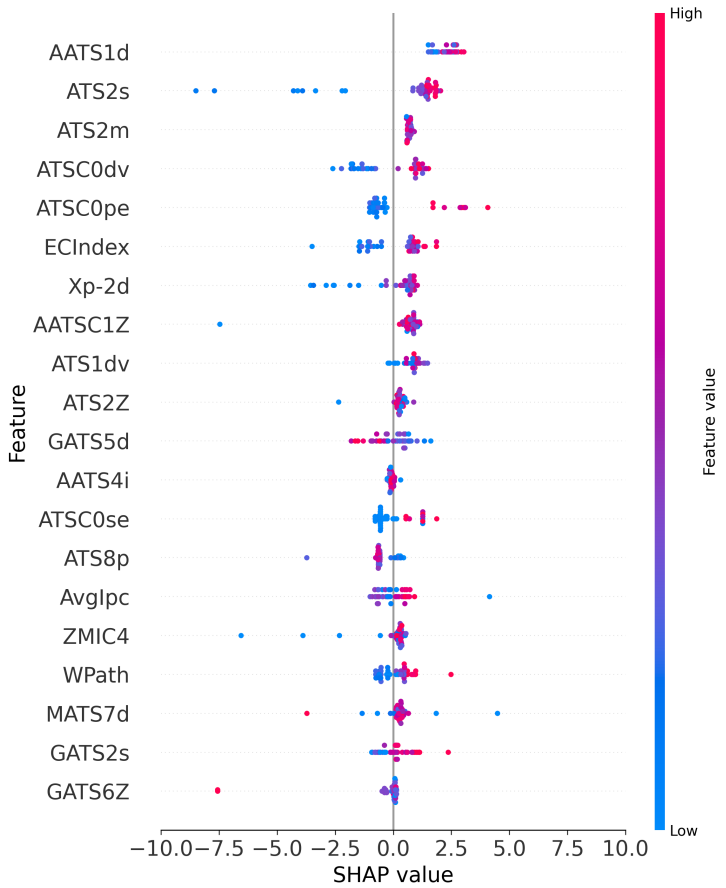


Figure 6.7: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule DVI.

To better understand the molecular features influencing molecule thermal stability, we trained an XGBoost model to predict DVI specifically for the purpose of performing SHAP analysis. As DVI is an empirical measure derived from viscosity changes with temperature as opposed to a directly measurable

molecular property, the degree to which a trained model and the resulting SHAP interpretations reflects true physical behaviour is limited. Nevertheless, this approach provides potentially valuable insights into structural features that correlate with improved DVI, offering useful guidance for molecular design and optimization in lubricant applications.

Fig. 6.7 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule DVI. The Autocorrelation of Topological Structure (ATS) family of descriptors, such AATS1d, ATS2s, ATS2m, ATSC0dv, and ATSC0pe, capture atom-based autocorrelations over molecular topologies, reflecting how specific atomic properties, such as mass, electronegativity, van der Waals volume, and polarisability, are distributed across different bond distances within a molecule.[712] These descriptors measure the degree of local and medium-range uniformity in structural and electronic environments, either directly (ATS) or through centered and normalised forms (ATSC). As shown by Fig. 6.7, higher ATS values, indicating greater consistency and regularity in atomic properties, are generally associated by the model with increased DVI, while lower ATS values are linked to reduced DVI. A notable exception to this seems to be the with ATS2m descriptor, which specifically quantifies the topological autocorrelation weighted by atomic mass, at a lag of 2, shows a positive SHAP contribution to DVI prediction at both low and high values. Low ATS2m values correspond to branched or cyclic hydrocarbon structures, while higher ATS2m values may correspond to long-chain or heavier molecules. This dual influence suggests that the model recognises structurally distinct features at both extremes as beneficial. Cyclic hydrocarbons and straight-chain alkanes typically exhibit higher ATS values due to their symmetric and repeating structures, while highly branched alkanes and ester-containing molecules tend to have lower ATS values because of localised variations in mass, polarity, and bonding environments. Molecules with lower ATS and corresponding DVI values may reflect the presence of atomic heterogeneity, which could disrupt phonon transfer.

The ECIndex descriptor captures the molecular eccentric connectivity index, which combines information about both the topological distance of atoms within its molecular graph (eccentricity) and the degree of atomic connectivity (number of bonded neighbors), reflecting how extended or compact a molecule's structure is based on how far atoms are from each other and how heavily connected they are.[713] Higher ECIndex values, indicating more linear, flexible, and uniformly connected structures, are associated with molecules that the model predicts as having higher DVI, while lower ECIndex values, related to more compact molecular structures are linked to reduced DVI. Molecules such as straight-chain alkanes or lightly branched hydrocarbons tend to exhibit higher ECIndex values due to their long, open-chain structures, while cyclic hydrocarbons, highly branched molecules and molecules with ring systems display lower ECIndex values because of their shorter internal distances and localised connectivity.

These findings are consistent with studies by researchers such as Verdier et al., who suggested that molecules with long alkyl chains, methyl branching in the centre of the molecule, low aromatic content, no ethyl branching and no tertiary carbons help to provide lubricants with a high kinematic viscosity index and low pour points.[684] Similarly, Lynch proposed a ranking of hydrocarbon families based on their effect on the kinematic viscosity index, in the following order: n-paraffins > lightly branched isoparaffins > multiply branched isoparaffins $\approx$ mononaphthenes with long side chains $\approx$ monoaromatics with long

side chains > diaromatics > polyaromatics ≈ polycyclic naphthenes with multiple short side chains.[714] The above results strongly suggest that molecular generation strategies should bias towards long alkyl chains, moderate central branching, and low aromatic or heteroatomic environments when proposing new molecules aiming to optimise for high VI performance.

Thermal Conductivity

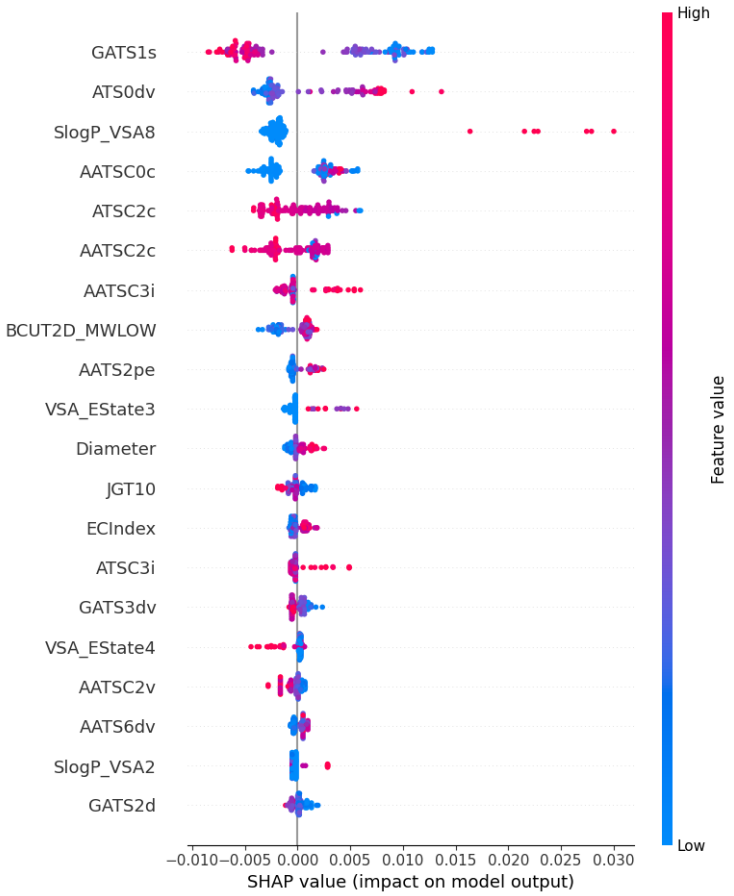


Figure 6.8: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 40°C.

Fig. 6.8 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 40°C, while Fig. D.1 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 100°C. The GATS1s descriptor captures the Geary autocorrelation of Sanderson electronegativity at a topological lag of 1, quantifying how much variability exists in the electronic environment between directly bonded atoms.[715] Higher values of GATS1s indicate greater local differences in electronegativity between neighboring atoms, reflecting a more electronically heterogeneous molecular structure. The model associates higher GATS1s values with lower thermal conductivity, likely because localised electronic asymmetry can disrupt phonon transport or reduce the efficiency of energy transfer through the molecular structure. Lower GATS1s values, typical of more electronically uniform molecules such as straight-chain alkanes, are associated with higher thermal conductivity, as consistent atomic environments promote better

vibrational energy propagation across molecular structures.[716]

ATS descriptors weighted by Gasteiger partial charge, such as ATS0dv and ATSC2c, quantify how charge distributions vary across molecular topologies at defined bond distances.[717] Higher AATSC0c values correspond to greater autocorrelation of partial charges, reflecting structured electronic environments with spatially non-uniform charge distribution. Molecules with higher values for this descriptor include alkylated aromatic hydrocarbons and linear esters, where polar functional groups are consistently positioned along the molecular structure. In contrast, linear alkanes exhibit lower AATSC0c values due to their electronically uniform nature and minimal differences in partial charge across C-C and C-H bonds. Despite these low values, many linear alkanes show relatively high thermal conductivity, indicating that AATSC0c does not correlate linearly with this property. TreeSHAP analysis reveals strong antagonistic interactions between AATSC0c and both VSA_EState and ECIndex. These interactions indicate that when partial charge heterogeneity is combined with geometric irregularity or compact molecular structure, as seen in highly branched or cyclic compounds, thermal conductivity transfer is reduced. These findings show that the influence of Gasteiger-weighted ATS descriptors on thermal conductivity depends not only on the magnitude of charge variation but also on how that variation aligns with the molecule's geometric and mass distribution.

The SlogP_VSA descriptor captures the total van der Waals surface area of molecular regions classified by defined ranges of atomic contributions to the octanol-water partition coefficient (logP), linking surface area to molecular hydrophobicity.[718] Higher SlogP_VSA values correspond to larger surface areas associated with hydrophobic atoms, reflecting longer, low-polarity molecular surfaces. In thermal conductivity prediction, higher values for this descriptor are associated with increased thermal conductivity. This relationship shows the effect of larger hydrophobic molecules on molecular packing, where more efficient packing promotes better vibrational energy transfer.[719] Lower SlogP_VSA values are observed in molecules with more polar or spatially irregular surfaces, including short-chain alcohols, esters with exposed carbonyl groups, or highly branched polar hydrocarbons. These structures are associated with reduced thermal conductivity. These results suggest that molecular generation strategies targeting improved thermal conductivity should prioritise structures containing long, linear or lightly branched alkyl chains, evenly distributed polar functional groups, and minimal ring systems.

Viscosity

Fig. 6.9 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 40°C, while Fig. D.2 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 100°C. The ATS2m and ATS3m descriptors are autocorrelation descriptors that encode molecular topology weighted by atomic mass, at a lag of 2 and 3 bonds respectively. These descriptors quantify how mass is distributed across the molecular graph, with higher values reflecting heavy atom mass at specific topological distances. The model associates higher ATS2m and ATS3m values with increased viscosity, suggesting that molecules with greater atomic mass spread over longer topological distances, such as branched and cyclic molecules have higher viscosities. Lower ATS2m and ATS3m values, normally seen in less branched and lighter

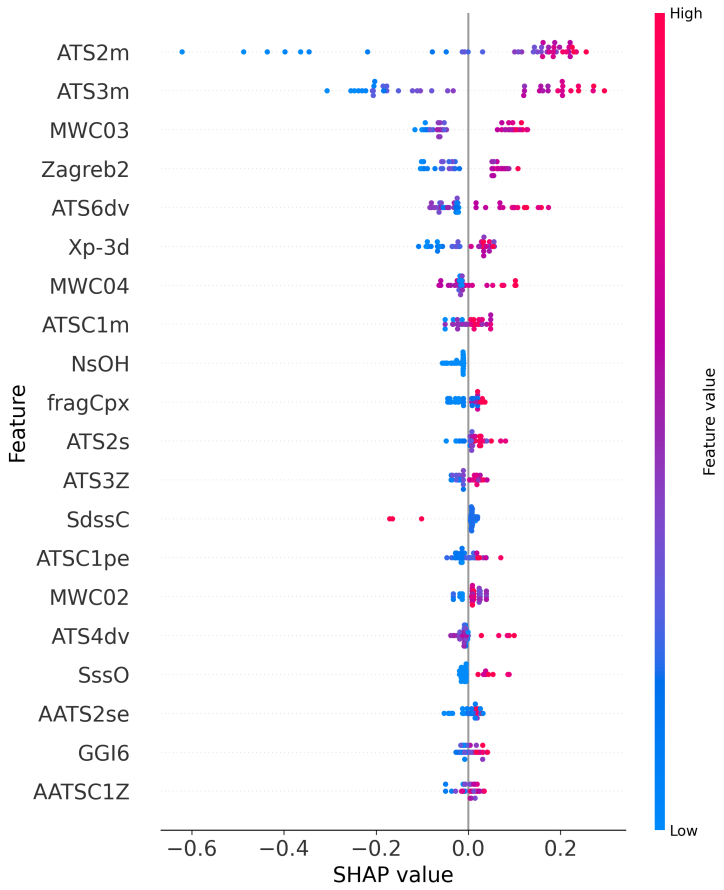


Figure 6.9: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 40°C.

molecules, are associated with lower viscosity.

The MWC03 descriptor represents a molecular walk count, capturing the number of distinct paths of length three between atom pairs within the molecular graph, reflecting molecular connectivity and complexity and emphasizing branching and cyclic structures at short topological ranges.[720] The model links higher MWC03 values with increased viscosity, indicating that molecules with more branching and denser connectivity patterns may hinder molecular mobility and increase viscosity. Lower MWC03 values, often found in linear or lightly branched molecules, are associated with lower viscosity.

The Zagreb2 index is a topological descriptor that measures molecular branching by summing the squares of the degrees (connectivity) of adjacent atom pairs across a molecule, emphasizing regions of high atomic connectivity, especially in branched or compact frameworks. The model associates higher Zagreb2 values with higher viscosity, suggesting that densely connected, branched molecules introduce greater resistance to flow. Conversely, lower Zagreb2 values are indicative of molecules with more linear, less branched structures, which tend to exhibit lower viscosity. These results suggest that molecular generation aimed at optimizing viscosity should prioritize structures with lower structural complexity, such as linear or lightly branched molecules.

Heat Capacity

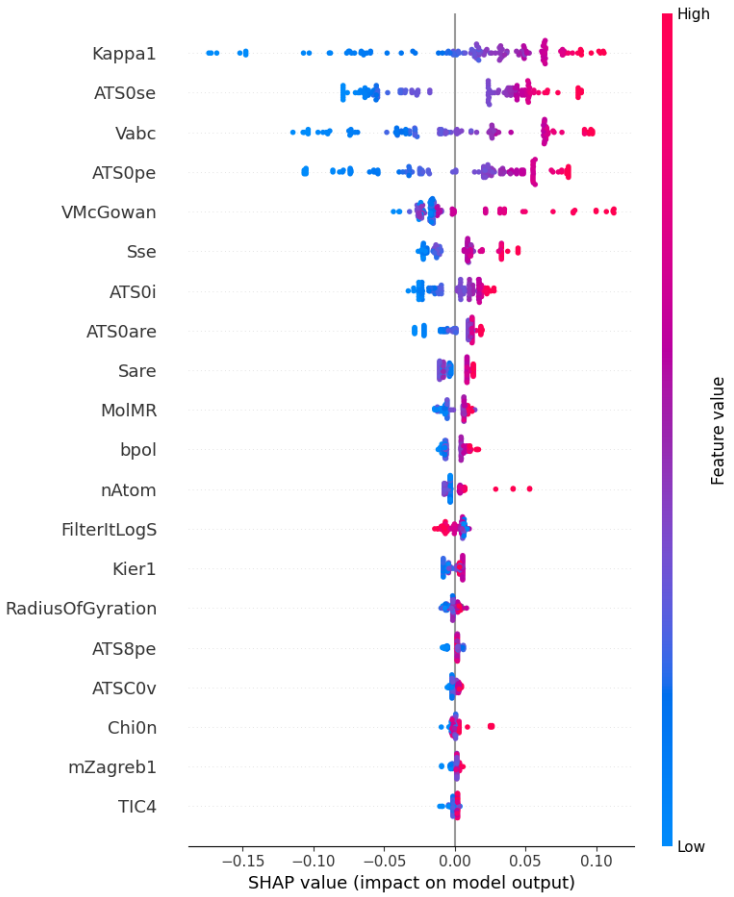


Figure 6.10: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log heat capacity at 40°C.

Fig. 6.10 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 40°C, while Fig. D.3 shows a beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 100°C. The Kappa1 descriptor (first order kappa shape index) quantifies molecular branching and shape based on the connectivity of atoms within a molecule.[721] In the model, higher Kappa1 values are associated with increased heat capacity, indicating that the presence of branching or cyclic functional groups increase heat capacity. This may be due to the fact that branched molecules often introduce additional vibrational modes or spatial configurations that influence heat retention or distribution characteristics. Conversely, lower Kappa1 values, typical of linear molecules, have lower heat capaciies, suggesting reduced capacity for such structures to store or modulate thermal energy. The ATS0pe and ATS0se descriptors are 0-lag 2D-autocorrelation measures weighted by atomic electronegativity (pe) and Sanderson electronegativity (se), respectively.[720] Higher ATS0pe and ATS0se values are associated with higher heat capacity by the model, likely because electronegative atoms, such as oxygen, introduce additional vibrational modes through polar bonds and electron-withdrawing groups.[722] These modes store thermal energy more efficiently, increasing the molecule's ability to absorb heat. Lower values of these descriptors,

typical of nonpolar hydrocarbons or structures lacking polar functional groups, correspond to lower heat capacity.[722]

The Vabc descriptor quantifies the van der Waals volume of a molecule by summing the volume contributions of individual atoms based on their van der Waals radii.[723] The model links higher van der Waals volumes to higher heat capacity, implying that molecules with larger atomic mass, often due to heavier atoms or branching, possess higher heat capacities, while molecules with lower van der Waals volumes, typically smaller or more compact molecules, exhibit lower heat capacities.

The VMcGowan descriptor estimates molecular volume based on group contributions, reflecting molecular bulk and atomic packing regardless of molecular geometry.[724] Higher VMcGowan values are linked to higher heat capacity. Smaller VMcGowan values, typical of compact molecules with low molecular weight, correlate with lower heat capacity, likely because such molecules have fewer vibrational modes. These results suggest that molecular generation aimed at optimizing heat capacity should prioritize structures containing electronegative atoms and polar functional groups, which contribute additional vibrational modes via polar bonds. Furthermore, designing molecules for enhanced thermal energy storage should focus on maximizing molecular volume while maintaining structural flexibility.

6.3.2 Other Design Considerations and Constraints for Molecular Generation

The SHAP analysis presented in the previous section highlights the conflicting nature of the molecular properties being optimized, revealing that structural features beneficial for one property may adversely affect another. For example, analysis highlighted that smaller, linear alkane molecules work best for minimising viscosity, however, larger linear molecules provide advantages for viscosity index. Furthermore, there are several other critical considerations when optimizing lubricant base oils to ensure both performance and longevity across a wide range of operating conditions. Oxidation stability is essential for resisting thermal and oxidative degradation over time. Aromatic compounds, including alkylated aromatics and naphthalenes, generally exhibit greater oxidative stability than paraffins, except when secondary carbons are directly bonded to the aromatic ring (e.g., methyl groups), which can reduce stability.[725] In contrast, long-chain and highly branched alkanes show poorer oxidation stability due to their increased susceptibility to radical initiation.[726,727] Compared to PAOs, dibasic esters, polyol esters, and polyalkylene glycols (PAG), alkylated naphthalenes have superior thermal and oxidative stability without additional antioxidants.[728,729,730] Dielectric constant affects a lubricant's insulating behaviour, which is especially important in electrically active environments such as those present in electric vehicles. Esters and ethers exhibit the highest dielectric constants followed by aromatic esters, while non-polar hydrocarbons such as alkanes and cycloalkanes show significantly lower dielectric values.[731] Additive compatibility ensures that functional additives, such as anti-wear agents, dispersants, or antioxidants, are effective within the base oil. Esters and ethers offer the highest compatibility due to their polar functional groups, which promote solubility and interaction with a broad range of additives. Aromatic esters and alkylated napthalenes also perform well, balancing polarity and hydrophobicity, while non-polar hydrocarbons like alkanes and cycloalkanes show relatively poor additive compatibility, often requiring additives.[47]

Polycyclic aromatic hydrocarbons (PAHs) are known to be toxic[732] and PAHs with four or more rings have been shown to be carcinogenic.[733] PAHs are emitted into the environment via the degradation of lubricating oils and, as a study by Worton et al. determined, as primary organic aerosol emissions.[539] Suitable lubricants tend to be derived from C16–C70 hydrocarbons with branched chains as branching leads to lower pour points, meaning oils with low pour points are suitable for low temperature applications.[734] Branched paraffins demonstrate similar viscosities to their non-branched counterparts, indicating that branching alone does not significantly alter viscosity. Whereas cycloparaffins with two or more fused rings, exhibit higher viscosities compared to n-paraffins (saturated hydrocarbon chains) of the same carbon number. The introduction of single aromatic rings into the hydrocarbon structure does not markedly change viscosity, however, the presence of multiple aromatic rings (3 or more) leads to a substantial increase in viscosity.[714] Additionally, the de-aromatization of multi-aromatic compounds results in a further rise in viscosity.[47] Non-aromatic hydrocarbons generally have better thermal conductivity compared to aromatic hydrocarbons as the ring structures of aromatic hydrocarbons limit the ease with which heat can be conducted. However, naphtenic hydrocarbons (e.g., cycloalkanes), are common in low temperature applications (below -30°C) for hydraulics and low temperature greases due to their superior pour points.[47]

Incorporating all of this information into the design process could easily overly constrain the search, reproducing already known molecules. Hence, there is a need to strike a careful balance between driving optimization and maintaining diversity, ensuring that the discovery process does not become confined to a narrow, non-novel region of the explorable chemical space. Therefore, to take all of these design considerations into account these without unduly biasing the generation process, some constraints were applied to the types of molecules that could be generated. Proposed candidates were limited to a maximum of 50 heavy atoms, limiting the likelihood of excessively high viscosities, which the trained models will struggle to accurately predict. Additionally, branched molecules were prioritized to reduce the likelihood of waxing at low temperatures by biasing the atom addition, fragment addition and fragment removal mutations to create bonds at non-terminal atoms. Similarly, the inclusion of rings and ester groups were encouraged to help promote better oxidation stability and dielectric constants. Proposed candidates were limited to a maximum of two aromatic rings to strike a balance between providing sufficient thermal and oxidative stability while maintaining low viscosity and thermal conductivity. Candidates with triple bonds were excluded due to their high energy state and elevated reactivity, which contribute to catalytic degradation processes in lubricants.[734] Rings with 3/4 members experience larger angle and torsional strain, rendering them more reactive and less stable while larger rings (7+ members) often exhibit increased conformational flexibility, leading to challenges in physical property control and synthetic accessibility.[46,735] Hence, only five and six-membered rings were allowed during molecular generation due to their optimal balance of stability and flexibility. SMARTS strings were used for substructure filtering as they allow for the precise description of molecular patterns, including atomic properties, connectivity, and bonding configurations.

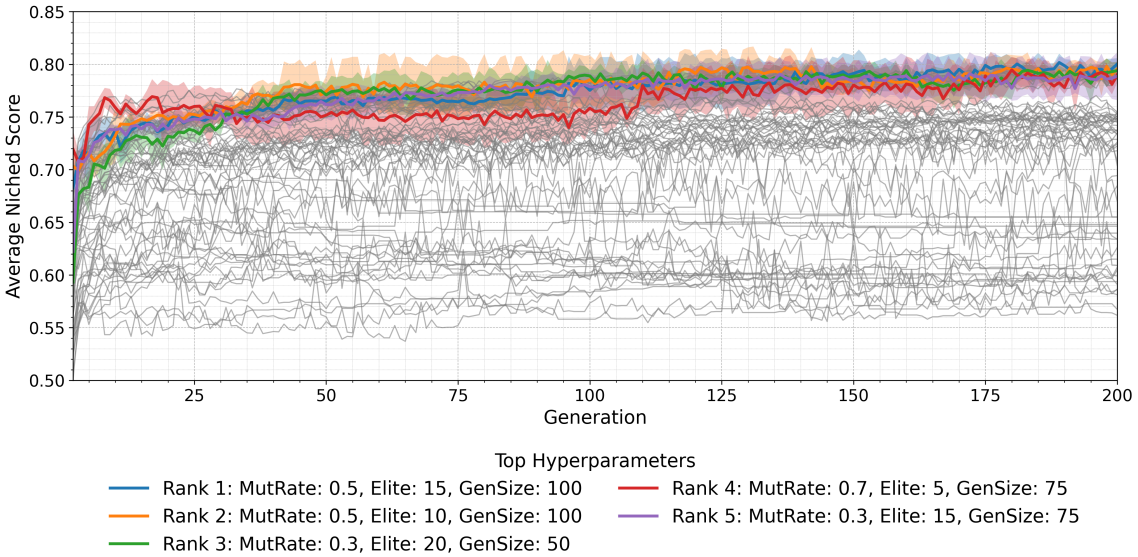


Figure 6.11: Evolution of the average niched score over generations for WSGA configurations across three repeats. The plot highlights the top 3 best-performing hyperparameter combinations with coloured lines and shaded standard deviation regions, with the performance of other hyperparamater combinations shown by grey lines.

6.3.3 Genetic Algorithm Hyperparameter Selection

Greenstein et al. studied the effects of varying GA hyperparameters on model performance on single-objective optimisation tasks, and how these hyperparameters affected a GA's ability to identify a known optimal solution in a chemical space.[670] They explored population sizes ranging from 16 to 96 candidates, observing that most configurations were able to identify the optimal solution. However, they showed larger populations consistently resulted in broader exploration of the target chemical space. Greenstein et al. also examined elitism rates between 5% and 80%, concluding that a rate of approximately 50% yielded the best performance in terms of GA speedup, chemical space coverage, and the ability to identify a known optimal solution from multiple candidates.[670] The same study investigated mutation rates from 1% to 90% in 10% increments, finding that a mutation rate around 40% provided an effective balance, encouraging sufficient diversity while still enabling local exploration of promising regions in the chemical space. To evaluate whether these parameters translate to MOOPs, we conducted a grid search over the following hyperparameters for both WSGA and NSGA-II, broadly matching the ranges explored in the Greenstein study: the number of elite individuals was varied across the set 5, 10, 15, 20, 25, the mutation rate was tested at 0.0, 0.3, 0.5, 0.7, 0.85, 0.95, and the generation size was explored at 20, 30, 50, 75, 100. Three repeats were conducted for each parameter combination to account for the stochastic nature of these algorithms. For both algorithms, the top 5 molecules from each generation were used for hypervolume and niched score calculations to ensure that generation size did not artificially boost or negate these metrics.

WSGA

Fig. 6.11 shows the evolution of the average niched score over generations for the tested WSGA configurations across the three repeats, with the top 5 performing hyperparameter combinations highlighted. Larger generation sizes often yield better-performing elite populations, as they enable more extensive sampling of the search space, increasing the likelihood of discovering high-quality solutions and maintaining diversity across generations. Moderate mutation rates, typically between 30% and 50%, help maintain structural diversity in the population while still allowing for the local refinement of high-performing candidates, thereby balancing exploration of new regions in the search space with exploitation of known promising areas without overwhelming the population with unproductive variations. Moderate levels of elitism, strike a balance between preserving strong candidates and allowing sufficient room for exploration through new individuals. This helps avoid premature convergence to suboptimal regions of the objective space.[736] The best-performing WSGA combinations in our experiments, mutation rates of 50% with 15 elites and a generation size of 100; 50% with 10 elites and a generation size of 100; and 30% with 20 elites and a generation size of 50, reflecting this balance.

While these trends are generally consistent with the findings of Greenstein et al., who also observed benefits from larger population sizes and moderate mutation rates, our optimal configurations involved lower elitism rates (10–20%) compared to their optimal 50% rate. This difference likely stems from the nature of the weighted sum framework. As multiple objectives are combined into a single scalar fitness value, the algorithm may converge prematurely to local optima that reflect specific weightings, especially if population diversity is not maintained. Lower elitism rates help mitigate this risk by allowing a larger fraction of the population to be replaced in each generation, increasing opportunities for novel individuals to enter the population and explore under-represented regions of the weighted objective space. The majority of hyperparameter combinations exhibited a plateau in performance after approximately 50 generations, with only the best-performing configurations continuing to show incremental improvements beyond this point. The averaged niched scores occasionally showed temporary declines, likely due to the introduction of diverse or exploratory candidates, followed by subsequent increases as the population adapted to incorporate beneficial traits of these new candidates.

NSGA-II

Fig. 6.12 shows the best hypervolumes achieved by each hyperparameter set used to run the NSGA-II algorithm. In our plot, we exclude the first ten generations for each run to remove noise caused by initialisation, focussing on the meaningful progression of hypervolume over time. As shown by Fig. 6.12, the highest mutation rates and generation sizes coupled with the smallest elitism rates yield better hypervolume performance. In many cases, the highest hypervolumes are achieved relatively early in the optimisation process. This suggests that the search quickly identifies diverse solutions, but struggles to refine or improve upon them in later generations.

One reason for this behaviour is that NSGA-II differs fundamentally from scalarisation methods such as WSGA. NSGA-II promotes non-dominated diversity rather than targeting a single composite objective, which can lead to stagnation in high-dimensional settings if the front quickly becomes saturated with

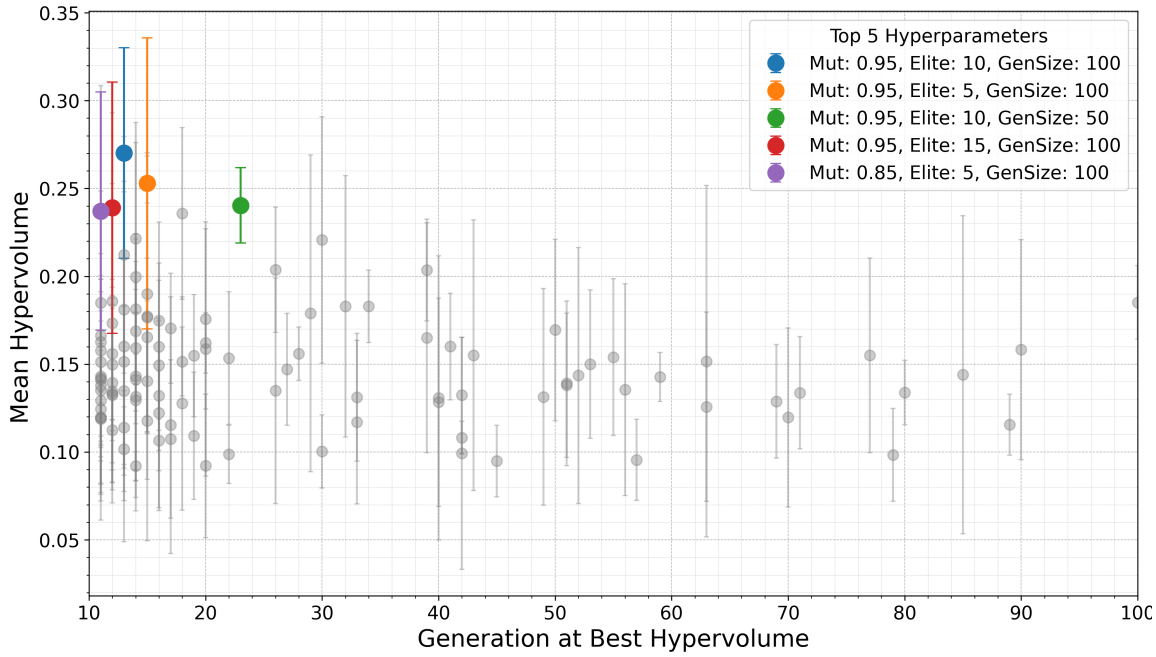


Figure 6.12: Best hypervolume achieved by each hyperparameter set used to run the NSGA-II algorithm, plotted against the generation at which that hypervolume was achieved. The 5 hyperparemeter sets that achieved the best hypervolumes are highlighted. Error bars indicate the standard deviation of hypervolume at that generation.

non-dominated solutions. This highlights the limitations of crowding distance method in its ability to distinguish between favourable and suboptimal molecules, as it relies solely on objective space spread and does not account for the absolute quality or relevance of individual solutions. In other words, a larger hypervolume does not always imply better quality solutions. In higher-dimensional problems such as ours, it becomes increasingly unlikely for one solution to completely dominate another. As a result, the Pareto front tends to grow in size, while the proportion of truly elite individuals diminishes. This reduces the effectiveness of metrics like crowding distance for evaluating solution quality. When the number of objectives increases, the expected number of non-dominated solutions can be estimated theoretically under the assumption of uniform sampling in objective space. One such estimate is given by:

$$\text{Pareto front size} \approx \frac{(\ln N)^{k-1}}{(k-1)!},$$

where N is the population size and k is the number of objectives.[737] For example, with $N = 100$ and $k = 6$, the estimated size of the Pareto front is,

$$\frac{(\ln 100)^5}{5!} = \frac{(4.605)^5}{120} \approx 18,$$

indicating that approximately 18% of the population is expected to lie on the first front. This suggests that while the front does grow with dimensionality, it does not necessarily dominate the entire population. In our case, an observed Pareto front size of 18% still represents a substantial portion of the

population and may reduce selection pressure, leading to stagnation in optimisation performance.[738] Fig. D.4 shows the top 20 molecules by crowding distance from best hyperparemeter combination using NSGA-II optimisation, while Table D.1 shows their molecular properties. As shown by Fig. D.4 and Table D.1, while these molecules represent the most diverse subset of molecules that are not dominated by any other solution, several of these solutions do not possess properties that would make them suitable as lubricant base oils. For example, while molecule 2 has a predicted viscosity at 40°C of 99.7 mPa.s, much higher than would be found in automotive lubricants, while molecule 3 has a thermal conductivity 0.08 W/m·K, much lower than the typical value range of 0.13–0.16 W/m·K.[51] These observations suggest that WSGA may be more suitable for this optimisation task, as it enables finer control of the tradeoffs of these conflicting objectives.

6.3.4 Best Performing Candidates from WSGA Runs

Using the optimised hyperparameters discovered from grid search in Section 6.3.3, we performed six runs of the WSGA algorithm. Five of these runs favoured individual properties: viscosity, thermal conductivity, heat capacity, toxicity, and DVI, by giving each property a weighting three times larger than all of the other properties in the weighted sum equation (Section 2.4.2). The sixth run applied an even weighting across all objectives. Each individual run spanned 200 generations, meaning approximately 102,600 unique molecules were screened. From each run, we selected the five best performing molecules based on their final niched score, resulting in a final group of thirty candidates from different parts of the relevant chemical space. The properties of these candidates were compared against five reference molecules widely used as either benchmarking fluids or as automotive base oils: PAO 4, hexadecane, diisodecyl adipate, squalane, and dioctyl sebacate. As biodegradability is a categorical property, it was not included in the WSGA optimisation process. Future studies can directly integrate biodegradability as a selection criteria in the WSGA pipeline, e.g., only admitting biodegradable molecules to a population. Fig. 6.13 illustrates the 30 molecules selected from the six independent runs as well as which run they were taken from and their predicted biodegradability. As predicted biodegradability was not applied directly as a selection criteria during optimisation, many of the top performing molecules are not biodegradable. Table D.2 shows how the top performers perform against each other and the reference molecules, by averaging their ranking across viscosity, thermal conductivity, heat capacity, SCScore, toxicity, and DVI. As shown by Table D.2 and Fig. 6.13, the best performing molecules broadly came from optimising for lower viscosity and lower toxicity, with 8 of 10 of the best overall performing molecules coming from these runs.

Molecule 27 exhibited the lowest viscosities of the candidates generated by the WSGA runs at both 40°C and 100°C. Molecule 25 and molecule 19 have predicted viscosities that would make them less effective as automotive lubricants, maybe due to the presence of extensive branching and cyclic structures that increase internal resistance to flow. However, thermal conductivity is highest in molecule 25 which has a plurality of polar groups and two aromatic rings, both of which support vibrational energy transfer via improved molecular packing. Molecule 6, which shows the lowest thermal conductivity values of the generated molecules, and contains relatively more branching which would likely interfere with efficient energy propagation across the molecular structure. Molecules 26 and 3 achieve the highest values for

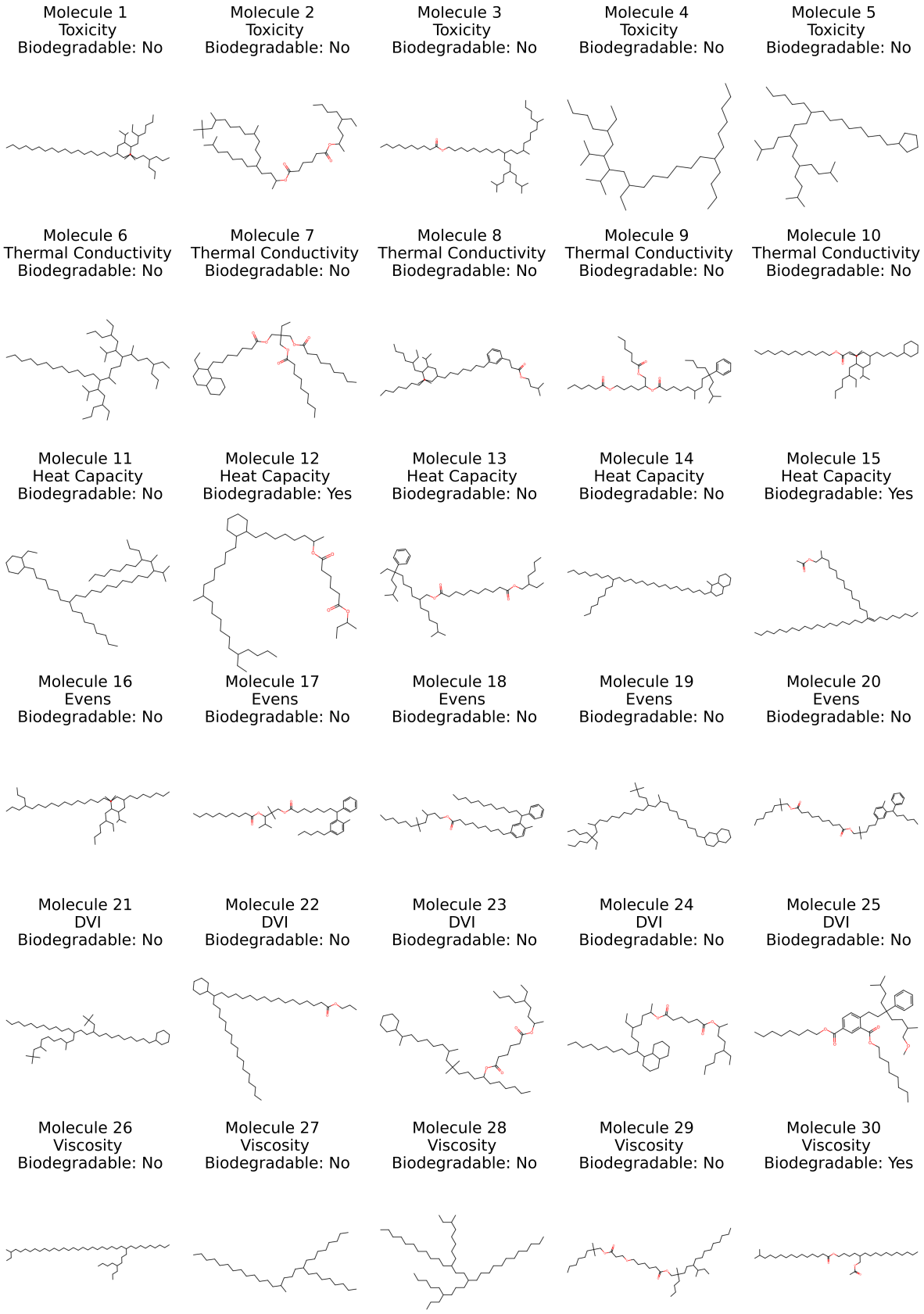


Figure 6.13: Grid of top 30 molecules from WSGA runs, showing the target property from which that molecules was taken, as well as its biodegradability prediction.

heat capacity. Their molecules feature long hydrocarbon chains/polar functional groups that contribute to greater thermal energy storage. Molecule 19 exhibits the highest dynamic viscosity index among all candidates. Its structure contains a long hydrocarbon chain with moderate branching and a fused ring functional group, which together contribute to a uniform distribution of mass and allowing it to maintain viscosity stability across a broad temperature range. The lowest toxicity values are observed in molecule 1, molecule 16, and molecule 6. These molecules contain saturated hydrocarbon chains with no aromatic rings or reactive heteroatoms. SCScore values are high for several top-performing molecules with molecule 15 and molecule 22 showing the lowest scores, which are still higher than all of the reference molecules, indicating greater synthetic complexity compared to the reference molecules. However, it should be borne in mind that like the SAScore, the SCScore will be biased to molecule reaction pathways present in its training dataset, meaning it could overestimate how difficult it is to synthesise molecules outside of its training distribution.

Overall, the highest-ranking molecules across all properties contain linear or lightly branched alkyl chains, minimal aromatic content, and a limited number of polar groups. These features support optimal viscosity, thermal stability, and toxicity while maintaining feasible synthetic complexity. These results emphasise the intrinsic conflict in simultaneously optimising all target properties. Molecule 19 demonstrates this clearly, achieving the highest dynamic viscosity index while ranking among the lowest across the other evaluated metrics. This highlights the challenge of designing molecules that perform well across multiple dimensions without compromising specific objectives.

All of the reference molecules are outperformed by the best WSGA candidates when considering performance across all of the metrics. Hexadecane and squalane show low viscosity and moderate thermal conductivity, reflecting their simple hydrocarbon structures. However, both exhibit relatively low heat capacity and low dynamic viscosity index, indicating limited thermal stability and narrower operating temperature ranges. Dioctyl sebacate and diisodecyl adipate achieve relatively low viscosities and moderately high thermal conductivities, but their heat capacity and DVI values remain below those of the highest-ranking generated molecules. PAO 4, while exhibiting good viscosity, demonstrates lower thermal conductivity and a less favourable toxicity score compared to several generated candidates.

The best-performing WSGA molecules achieve higher scores across a broader range of properties, including heat capacity and DVI, while maintaining comparable or better performance in viscosity and toxicity. Furthermore, the implemented constraints, while not directly captured in the comparative analysis, are expected to provide solutions with improved properties in other performance requirements such as enhanced oxidation resistance and additive compatibility, relative to the reference base oils. All of the generated candidates, fall within the applicability domain of the XGBoost models determined by Mahalanobis distance. This confirms that model predictions are reliable for the chemical space explored in this study. Overall, these results show that multi-objective molecular generation is capable of producing candidates that surpass conventional base oils in multiple performance metrics, while retaining structural features that remain within the bounds of synthetic feasibility.

6.4 Conclusions

This work presents an integrated framework for the multi-objective optimisation of lubricant base oils, combining evolutionary algorithms, machine learning models, and domain-specific design constraints. We show that WSGAs, when paired with predictive ML models trained on experimental property data, are capable of generating candidate molecules that outperform reference base oils across multiple performance criteria.

The use of SHAP analysis across multiple target properties facilitated insights into the molecular functional groups associated with desirable performance. Across biodegradability, thermal conductivity, viscosity, DVI and heat capacity, we identified relationships between structural descriptors and property predictions. These insights informed the design of constraints in the molecular generation process, biasing candidate exploration towards more favourable regions of the search space.

Our evaluation of hyperparameter configurations highlights that larger generation sizes, moderate mutation rates, and lower elitism rates improve multi-objective optimisation performance for both scalarised and Pareto-based approaches. These hyperparameters help maintain population diversity and enable exploration of high-performing regions of chemical space. Our comparison of NSGA-II and WSGA shows that NSGA-II, while effective at promoting non-dominated diversity, often struggles to maintain selection pressure in high-dimensional objective spaces. This is due to the increasing size of the Pareto front relative to the population, which reduces the effectiveness of crowding distance and leads to the retention of suboptimal solutions. WSGA provides finer control over tradeoffs between conflicting objectives, enabling the generation of molecules with more desirable property combinations.

The generated candidates demonstrate favourable DVI, thermal properties, and toxicity relative to reference fluids, while adhering to synthetic feasibility constraints. This work establishes a scalable and interpretable approach for multi-objective molecular design of lubricant base oils. By integrating sensible molecular constraints, model-driven insights, and optimisation strategies tailored to the demands of lubricant formulation, the framework presented here offers a practical tool for advancing the design of next-generation base oils.

Chapter 7

Conclusions, Limitations and Further Work

In this chapter, the main conclusions drawn from each study are summarised. A discussion of how each chapter addressed the relevant research objectives is provided. The limitations of the methodologies used and findings presented in each study are elucidated. Finally, work that could be done to extend each project further is discussed.

7.1 Chapter 3

Conclusions

In this chapter, we studied the mechanochemical decomposition of TCP molecules confined between atomically-smooth iron, iron carbide, and iron oxide surfaces using NEMD simulations with ReaxFF.

TCP decomposes more readily on iron surfaces compared to iron carbide and especially iron oxide, consistent with increased Fe-C bonding and higher surface energy. Adsorption is followed by P-O bond dissociation to yield cresol on all surfaces, with occasional subsequent C-O cleavage producing toluene.

The decomposition rate increases exponentially with temperature and shear stress, indicating a SATA mechanism. While DIDA base oil molecules do not alter the decomposition pathway, they reduce shear stress and thus lower the decomposition rate. Both systems, with and without DIDA, can be described by a single Bell model fit, supporting the use of base oil-free NEMD simulations for broader studies.

Polyphosphate growth initiates concurrently, with comparable P-O bond counts across surfaces due to offsetting rates of formation and cleavage. Chain lengths are similar overall, with iron surfaces exhibiting a skew toward longer chains.

3D Bell model fits show activation energy and pre-exponential factors increasing in the order: $Fe_3O_4(100)$ < $Fe_3C(010) \approx \alpha$-Fe(110). This trend reflects reduced reactivity and fewer accessible iron atoms on oxide surfaces, leading to lower reaction probabilities despite smaller barriers. Activation volumes follow

the same order, implying that higher surface energy enhances stress transmission by anchoring TCP more strongly. Lower contact stiffness may further contribute to the higher activation energy on α-Fe(110). Overall, surface chemistry governs the mechanochemical decomposition of TCP, with observed trends rationalised via Bell model parameters.

These findings directly address the overarching research objective by revealing how surface chemistry, temperature, and shear stress govern the mechanochemical decomposition of TCP at the nanoscale, thereby shaping its macroscopic tribological behaviour. The sensitivity of TCP reactivity to the type of ferrous surface underscores the importance of tailoring surface composition in tribological contacts to optimise additive performance. Additionally, the observed influence of base oil (DIDA) on stress transmission, despite its negligible effect on the underlying decomposition mechanism, highlights the need to consider how lubricant formulations modulate mechanical energy delivery to additives under load. This suggests that base oils act not only as solvents, but also as mediators of mechanical energy transfer. As such, their selection should consider both chemical compatibility and their influence on stress distribution at the molecular level.

Limitations and Further Work

While the findings of this study provide valuable insights into the effects of surface on the decomposition mechanisms of TCP, there are some limitations to this study that should be considered. Firstly, the simulations were performed on idealised, atomically smooth surfaces that do not reflect the topographical and chemical heterogeneity present in real tribological contacts, meaning that the effects of surface roughness on TCP have been neglected by these studies. In addition, the concentration of TCP molecules in the simulated base oil systems ($\approx 36\%$), were much higher than those found in commercial formulations (1-10%),[54,55] which may lead to overestimated dissociation rates and correspondingly lower activation energies. Furthermore, the simulations were limited to one nanosecond, which may be insufficient to capture slower tribochemical processes such as phosphate cross-linking or long-term film restructuring. Finally, the small number of TCP molecules present in these simulations may restrict the diversity of observed decomposition mechanisms and may limit the observation of longer polyphosphate films.

Future work should aim to bridge the remaining gaps between simulation and experiment. Experiments using chemically well-defined surfaces, such as iron oxides grown on atomically-smooth silicon wafers, would allow for more direct comparison with simulation results and provide validation of mechanistic predictions. Incorporating surface roughness and microstructural features into the NEMD simulations would enhance their realism and may uncover additional decomposition pathways or tribofilm formation mechanisms. For example, NEMD simulations of stearic acid on iron surfaces have shown that surfaces with greater levels of nanoscale RMS roughness slightly alter the efficacy of organic friction modifiers.[739] A similar approach for anti-wear additives could determine the efficacy of such additives in conditions closer to that experienced in real-world operating conditions. The integration of multiscale modelling approaches could extend simulation timescales and reveal longer-term tribochemical phenomena. For example, a recent study by Ntioudis et al. coupled NEMD with kinetic Monte Carlo methods, to

model the thermal decomposition of TCP and subsequent tribofilm growth.[740] Simulating systems with multiple additives and base oils at realistic concentrations would improve the relevance to practical lubricant formulations, though this will necessitate significantly larger simulation cells and computational resources. Finally, DFT and NEB calculations could be employed to refine activation energy estimates and provide a more detailed understanding of individual reaction steps involved in additive decomposition under applied shear stress. Similar approaches have already been taken to characterise the energies involved in different thermal decomposition pathways of meta- and para- TCP molecules exposed to ferrous surfaces.[301,349]

7.2 Chapter 4

Conclusions

In this chapter, we investigated the mechanochemical decomposition of TCP molecules confined between DLC-α-Fe(110) and DLC-Fe_2O_3 interfaces using NEMD simulations with ReaxFF, expanding on the studies of Chapter 3 by further isolating the effects of shear stress on TCP decomposition.

A transition from second-order to first-order kinetics with increasing speed indicated a shift from cooperative to unimolecular, shear-driven decomposition. At high speeds, exponential fits described the data well, suggesting stochastic bond dissociation.

Shear stress increased with speed before saturating above 20 m/s. Activation volumes were higher on α-Fe(110) and increased with temperature, reflecting more effective stress transfer. Bell activation energies decreased with speed, confirming SATA as the dominant mechanism. Extrapolated values agreed with experimental data for iron surfaces.

Polyphosphate growth was more extensive on α-Fe(110), with chains reaching 18 phosphorus atoms at 50m/s. P-O bonds elongated more under shear than C-O bonds, consistent with preferential P-O cleavage. These effects were enhanced by stronger adhesion on iron.

Overall, TCP decomposition is governed by shear stress and surface chemistry, which together control reaction kinetics and tribofilm formation under tribological conditions.

These results directly support the objectives of the thesis by revealing how sliding speed, temperature, and shear conditions influence the mechanochemical decomposition of a lubricant additive. The observed dependence of decomposition kinetics and tribofilm formation on sliding speed highlights the critical role of shear stress in controlling additive activation. As shear stress increases with sliding speed, the likelihood of bond scission events and molecular fragmentation rises, particularly on more reactive surfaces that effectively aid in transmitting mechanical energy to the additive. While this study focused on additive-only systems, the findings suggest that lubricant design should account for operating conditions that dictate shear regimes, such as speed and load. Base oil formulations and surface coatings should be co-optimised to ensure that the generated shear stress under realistic conditions enables effective additive activation.

Limitations and Further Work

While this study provides useful insights into the shear-driven decomposition of TCP at DLC-ferrous surface interfaces, a few limitations constrain the applicability of the findings. Similar to the limitations of Chapter 3, use of atomically smooth surfaces neglects the influence of surface roughness and topographical heterogeneity, which are intrinsic to real tribological contacts and likely influence local stress distribution and reactive site availability. Furthermore, the role of base oil was not explored under varying shear stress conditions, despite its potential to alter interfacial shear stresses and influence additive reactivity. Additionally, the activation volume used to calculate other mechanochemical parameters was calculated at a single sliding speed, meaning any variation in sliding speed that may affect the calculated Bell activation energies and pre-exponential factors was not fully assessed. The use of a first order Bell model was justified by empirical fits but not fully validated; ab initio calculations of the curvature of the potential energy surface would be required to determine whether second order corrections are necessary. Finally, the relatively small number of TCP molecules may have limited the diversity of the decomposition mechanisms observed, and constrained the formation of longer polyphosphate films.

Future work should focus on addressing the above limitations. Experimental studies performed across a broader range of sliding speeds and temperatures than currently available in literature, with comparable interface chemistry to that used in our simulations, would enable more direct validation of extrapolated activation parameters and support model calibration. Simulations involving other phosphate esters, such as alkyl phosphates, and a variety of base oil chemistries would help elucidate how substituent structure and lubricant composition influence decomposition mechanisms and tribofilm formation under shear. To improve the accuracy of activation volumes used for Bell model calculations, additional simulations should be conducted at each sliding speed under a range of compressive stresses, allowing for speed-specific activation volume estimates rather than relying on a single reference condition. Further validation of the used Bell model formulation should involve calculation of the potential energy surface curvature using density functional theory, to determine whether the use of the first order model remains appropriate or if second order corrections are required. Finally, incorporating surface roughness, surface heterogeneity, and increasing the number of TCP molecules in future simulations would improve simulation fidelity, providing deeper insight into the formation dynamics of TCP tribofilms.

7.3 Chapter 5

Conclusions

In this chapter, the effectiveness of machine learning (ML) and molecular dynamics (MD) approaches for predicting key transport properties of hydrocarbons relevant to lubricant base oil design was assessed. Different ML models and molecular representations were benchmarked, and their predictions compared to MD simulations using a range of force fields, with the aim of identifying reliable and efficient methods for estimating properties such as viscosity, density, and thermal conductivity. ML model generalisation to novel chemical structures outside the training distribution was evaluated, and the limitations of using MD-generated data to train ML models was explored.

PCA embeddings outperformed UMAP for tree-based models like XGBoost, while UMAP was more effective for MLPs. Across all cases, XGBoost models consistently achieved superior accuracy. Fingerprinting experiments showed improved performance with larger-radius, unhashed ECFP descriptors, supporting their use in interpretable and generalisable property prediction.

Tree-based models paired with descriptors provided optimal accuracy across thermophysical properties, with fingerprint representations being better-suited for structure-dependent targets like density, while descriptor representations worked better for viscosity, thermal conductivity and heat capacity. Multitask learning offered no benefit over single-task models, likely due to task interference. These findings highlight the importance of aligning molecular representations with model architecture for reliable property prediction.

ML models outperformed MD simulations in predicting density, viscosity, and thermal conductivity of hydrocarbons, particularly for complex molecules with cyclic structures. Standard force fields showed reduced accuracy, especially for viscosity and thermal conductivity, reflecting limitations in their parameterisation.

Models trained on MD-generated data underperformed compared to those trained on experimental data, underscoring the importance of dataset quality. This supports the use of hybrid training strategies that integrate experimental and curated synthetic data.

Out-of-distribution analysis revealed that descriptor-based models struggle to generalise to novel chemical spaces. Prediction errors correlated more strongly with shifts in target property values than with structural dissimilarity, indicating a need for caution when applying models beyond their training domains.

These findings directly support the objectives of the thesis by identifying the most effective combinations of ML/MD models and molecular representations for predicting hydrocarbon transport properties critical to lubricant base oil performance. The strong predictive accuracy of tree based models such as XGBoost, especially when used with full molecular descriptors, demonstrates their potential for screening both established and novel hydrocarbon structures. Our results show that ML models trained on MD-generated datasets underperform compared to those trained on experimental data, underscoring the importance of force field quality and molecular diversity in simulation-based workflows. Nevertheless, MD-generated datasets still provide a way to train ML models for property prediction in domains where data is scarce, such as lubricant additive properties and lubricant mixture properties. Furthermore, the limited extrapolation performance of descriptor-based models on OOD molecules highlights the need for caution when applying ML methods to unfamiliar chemical spaces.

Limitations and Further Work

This study benchmarks ML and MD techniques for predicting hydrocarbon transport properties, but several directions remain open for improvement and expansion.

Although extensive curation was undertaken, the dataset of $\approx$ 1,200 molecules remains small by mod-

ern ML standards. This relatively small dataset most likely limited the performance of the trained deep learning models, and constrained the evaluation of generalisation, particularly for out-of-distribution (OOD) molecules. Expanding datasets to include a broader chemical space and integrating high-quality experimental and synthetic data will be essential to improve representation learning, support the development of large-scale language models for chemistry, and enhance the generalisability and accuracy of property predictions. The ChemBERTa model underperformed relative to simpler models. Alternative fine-tuning strategies, such as low-rank adaptation (LoRA), could improve its utility. Better conformer generation methods such as DFT or force fields optimised for energetics could enhance graph-based learning. Future work should also explore more chemically informative graph representations and richer node and edge features.

This work focused exclusively on predicting properties of pure base oils at a narrow range of temperatures at atmospheric pressure. Real-world lubricants are often mixtures of base oils and additives exposed to a wide range of temperatures and pressures, so future studies should assess both ML and MD methods for such systems. Using improved parameterisation of MD force fields, particularly for cyclic and functionalised molecules, would provide a more up-to-date representation of the capabilities of EMD for transport property prediction. For OOD prediction, improved uncertainty quantification is needed. Mahalanobis distance is limited by its Gaussian assumptions; combining it with non-parametric methods such as k-nearest neighbours in normalised descriptor space may yield more robust OOD detection.

7.4 Chapter 6

Conclusions

This work presents an integrated framework for the multi-objective optimisation of lubricant base oils, combining evolutionary algorithms, machine learning models, and domain-specific design constraints. We show that WSGAs, when paired with predictive ML models trained on experimental property data, are capable of generating candidate molecules that outperform reference base oils across multiple performance criteria.

The use of SHAP analysis across multiple target properties facilitated insights into the molecular functional groups associated with desirable performance. Across biodegradability, thermal conductivity, viscosity, DVI and heat capacity, we identified relationships between structural descriptors and property predictions. These insights informed the design of constraints in the molecular generation process, biasing candidate exploration towards more favourable regions of the search space.

Our evaluation of hyperparameter configurations highlights that larger generation sizes, moderate mutation rates, and lower elitism rates improve multi-objective optimisation performance for both scalarised and Pareto-based approaches. These hyperparameters help maintain population diversity and enable exploration of high-performing regions of chemical space. Our comparison of NSGA-II and WSGA shows that NSGA-II, while effective at promoting non-dominated diversity, often struggles to maintain selection pressure in high-dimensional objective spaces. This is due to the increasing size of the

Pareto front relative to the population, which reduces the effectiveness of crowding distance and leads to the retention of suboptimal solutions. WSGA provides finer control over tradeoffs between conflicting objectives, enabling the generation of molecules with more desirable property combinations.

The best generated candidates demonstrate favourable DVI, thermal properties, and toxicity relative to reference fluids, while remaining synthetically feasible. This work establishes a scalable and interpretable approach for multi-objective molecular design of lubricant base oils.

These findings directly support the stated objectives by integrating interpretable ML with optimisation techniques to guide the rational design of lubricant base oils. SHAP analysis revealed how specific molecular features contribute to key properties such as viscosity, thermal conductivity, biodegradability, and heat capacity, providing actionable insights for molecular design. These insights were used to inform chemically sensible constraints in the candidate generation process, helping to focus exploration on promising areas of chemical space. The application of weighted sum genetic algorithms (WSGA) in conjunction with predictive ML models led to the discovery of molecules that outperform reference fluids across multiple objectives, while adhering to synthetic feasibility. Additionally, comparisons between WSGA and NSGA-II showed that WSGA enables more effective control over tradeoffs in high dimensional optimisation problems, which is critical for lubricant formulation where multiple properties must be balanced. This integrated approach demonstrates how interpretable ML and tailored optimisation strategies can be used together to support the development of advanced lubricant formulations with improved performance and sustainability.

Limitations and Further Work

This study demonstrates that combining genetic algorithms with explainable machine learning can effectively support multi-objective molecular optimisation. However, several limitations remain. The predictive accuracy of the machine learning models depends on the chemical diversity and coverage of the training dataset. Molecules that fall outside the model's applicability domain would most likely produce unreliable property predictions, particularly when they contain wholly novel structural features or underrepresented functional groups. The dataset used in this study does not include lubricant chemistries such as ionic liquids, polymeric fluids, or molecules containing elements like fluorine, boron, or nitrogen. These functional groups play important roles in thermal stability, polarity, and additive compatibility. Including a wider range of chemistries and explicitly modelling additive interactions would help focus the search space and improve model generalisability.

This study incorporates several important base oil properties that are often overlooked by other de-novo design pipelines, but other important properties were not explicitly modelled. These include oxidation stability, electrical properties such as dielectric constant and breakdown voltage, and low temperature performance indicators such as pour point. Standard molecular descriptors cannot reliably capture these properties, so predicting them may require physics-based methods such as MD simulations using reactive force fields like ReaxFF. While these approaches can offer better predictions of degradation, additive reactivity, and transport properties, they are also computationally expensive.

Genetic algorithms, although interpretable and adaptable, have intrinsic limitations. They may converge prematurely or explore the search space inefficiently, particularly in high dimensional objective spaces. Genetic algorithms based on weighted sum scalarisation, such as WSGA, rely on fixed objective weightings. This approach may oversimplify the problem when the tradeoffs between objectives are nonlinear. Future work could explore alternative optimisation strategies such as particle swarm optimisation,[741] Bayesian optimisation,[742] or reinforcement learning.[743] These alternatives may improve the search efficiency and offer better convergence behaviour.

It is also important to recognise the limitations of SHAP analysis. SHAP values reflect how a model uses features to make predictions as opposed to representing causal or chemically valid relationships. As such, interpretations may be biased by feature collinearity, non-linear interactions, or artefacts in model training. Moreover, reliance on test set predictions for SHAP analysis means that interpretation quality is contingent on generalisation performance and data representativeness.[744] We have shown in previous sections that the trained models perform well on predicting properties of unseen molecules within the training distribution (e.g., Fig. 5.8), however we have also shown that model performance significantly decreases when predicting on molecules outside of the training distribution (Fig. 5.14).

This study used Mordred and RDKit molecular descriptors, which may not fully capture complex structure–property relationships. Future work could employ richer and more expressive feature representations. Graph-based embeddings and quantum chemical descriptors may provide improved model accuracy and enable deeper insights into molecular behaviour.

The optimisation framework presented here focuses on single molecule candidates. In practice, commercial lubricants are complex formulations composed of multiple components, including base oils, viscosity modifiers, antioxidants, and other additives. Extending the optimisation approach to support multicomponent systems would better reflect the realities of industrial lubricant design. Such a framework could enable the simultaneous optimisation of base oils and additive systems.

Although synthetic complexity scoring was used to avoid impractical candidates, this approach remains a proxy for true feasibility of synthesis. More accurate assessments would require reaction prediction models or retrosynthesis tools that consider real-world reaction pathways. Finally, while this study identifies several promising molecules through computational screening, these molecules must undergo experimental validation to confirm their true usefulness. Laboratory synthesis, tribological performance testing, and long-term ageing studies will be necessary to verify that the predicted molecular properties translate into practical performance under realistic use conditions.

Bibliography

[1] Yan Chen et al. “Performance characteristics of lubricants in electric and hybrid vehicles: a review of current and future needs”. In: *Frontiers in Mechanical Engineering* 6 (2020), p. 571464.

[2] Paulina Nowak, Karolina Kucharska, and Marian Kamiński. “Ecological and health effects of lubricant oils emitted into the environment”. In: *International Journal of Environmental Research and Public Health* 16.16 (2019), p. 3002.

[3] Rui Zhou et al. ““Lubrication Brain”—A machine learning framework of lubrication oil molecule design”. In: *Tribology International* 183 (2023), p. 108381.

[4] Robert Ian Taylor. “Tribology and energy efficiency: from molecules to lubricated contacts to complete machines”. In: *Faraday discussions* 156.1 (2012), pp. 361–382.

[5] Deping Xia et al. “Research progress of antioxidant additives for lubricating oils”. In: *Lubricants* 12.4 (2024), p. 115.

[6] Haogang Li et al. “Extreme pressure and antiwear additives for lubricant: academic insights and perspectives”. In: *The International Journal of Advanced Manufacturing Technology* 120.1 (2022), pp. 1–27.

[7] Hugh Spikes. “Friction modifier additives”. In: *Tribology Letters* 60 (2015), pp. 1–26.

[8] Johan Guegan, Mark Southby, and Hugh Spikes. “Friction Modifier Additives, Synergies and Antagonisms”. In: *Tribology Letters* 67 (2019), p. 83. DOI: 10.1007/s11249-019-1198-z.

[9] A. Neville et al. “Compatibility between tribological surfaces and lubricant additives-How friction and wear reduction can be controlled by surface/lube synergies”. In: *Tribology International* 40 (10-12 2007), pp. 1680–1695. ISSN: 0301-679X. DOI: 10.1016/j.triboint.2007.01.019.

[10] H. Spikes. “The history and mechanisms of ZDDP”. In: *Tribology Letters* 17.3 (2004), pp. 469–489.

[11] P.M. Cann. “In-Contact Molecular Spectroscopy of Liquid Lubricant Films”. In: *MRS Bulletin* 33 (12 Dec. 2008), pp. 1151–1158. ISSN: 0883-7694. DOI: 10.1557/mrs2008.245.

[12] Tevis D B Jacobs et al. “Insights into tribology from in situ nanoscale experiments”. In: *MRS Bulletin* 44 (2019), pp. 478–486. DOI: 10.1557/mrs.2019.122.

[13] A.I. Vakis et al. “Modeling and simulation in tribology across scales: An overview”. In: *Tribology International* 125 (2018), pp. 169–199. ISSN: 0301679X. DOI: 10.1016/j.triboint.2018.02.005.

[14] Suhaib Ardah, Francisco J Profito, and Daniele Dini. “A comprehensive review and trends in lubrication modelling”. In: *Advances in Colloid and Interface Science* (2025), p. 103492. DOI: 10.1016/j.cis.2025.103492.

[15] JP Ewen, DM Heyes, and Daniele Dini. “Advances in nonequilibrium molecular dynamics simulations of lubricants and additives”. In: *Friction* 6 (2018), pp. 349–386.

[16] Ashlie Martini, Stefan J. Eder, and Nicole Dörr. “Tribochemistry: A Review of Reactive Molecular Dynamics Simulations”. In: *Lubricants* 8 (2020), p. 44. DOI: 10.3390/lubricants8040044.

[17] Huong T. T. Ta et al. “Computational Tribochemistry: A Review from Classical and Quantum Mechanics Studies”. In: *The Journal of Physical Chemistry C* 125 (2021), pp. 16875–16891. ISSN: 1932-7447. DOI: 10.1021/acs.jpcc.1c03725.

[18] Kenneth Holmberg and Ali Erdemir. “The impact of tribology on energy use and CO2 emission globally and in combustion engine and electric cars”. In: *Tribology International* 135 (2019), pp. 389–396.

[19] Kenneth Holmberg and Ali Erdemir. “Influence of tribology on global energy consumption, costs and emissions”. In: *Friction* 5 (2017), pp. 263–284.

[20] Department for Transport. *Transport and Environment Statistics - Autumn 2021*. https://www.gov.uk/government/statistics/transport-and-environment-statistics-autumn-2021. 2021.

[21] Kenneth Holmberg, Peter Andersson, and Ali Erdemir. “Global energy consumption due to friction in passenger cars”. In: *Tribology international* 47 (2012), pp. 221–234.

[22] Justin Rowlatt. *Why electric cars will take over sooner than you think*. Accessed: 2022-06-01. BBC. June 1, 2021. URL: https://www.bbc.co.uk/news/business-57253947.

[23] Department for Transport. *Government takes historic step towards net-zero with end of sale of new petrol and diesel cars by 2030*. https://www.gov.uk/government/news/government-takes-historic-step-towards-net-zero-with-end-of-sale-of-new-petrol-and-diesel-cars-by-2030. Accessed: 2025-05-01. 2020.

[24] Raj Shah, Blerim Gashi, and Andreas Rosenkranz. “Latest developments in designing advanced lubricants and greases for electric vehicles—An overview”. In: *Lubrication Science* 34.8 (2022), pp. 515–526.

[25] Simon C Tung, Mathias Woydt, and Raj Shah. “Global insights on future trends of hybrid/EV driveline lubrication and thermal management”. In: *Frontiers in Mechanical Engineering* 6 (2020), p. 571786.

[26] Zhaoran Zhu et al. “Effect of Electric Fields on the Decomposition of Phosphate Esters”. In: *The Journal of Physical Chemistry C* 128.38 (2024), pp. 15959–15973.

[27] Carlos Ayestarán Latorre et al. “Mechanochemistry of phosphate esters confined between sliding iron surfaces”. In: *Communications Chemistry* 4.1 (2021), p. 178.

[28] Roghayyeh Lotfi, Adri CT Van Duin, and Mousumi Mani Biswas. “Molecular dynamics simulations of perfluoropolyether lubricant degradation in the presence of oxygen, water, and oxide nanoparticles using a ReaxFF reactive force field”. In: *The Journal of Physical Chemistry C* 122.5 (2018), pp. 2684–2695.

[29] Seiji Kajita, Tomoyuki Kinjo, and Tomoki Nishi. “Autonomous molecular design by Monte-Carlo tree search and rapid evaluations using molecular dynamics simulations”. In: *Communications Physics* 3.1 (2020), p. 77.

[30] Camille Bilodeau et al. “Machine learning for predicting the viscosity of binary liquid mixtures”. In: *Chemical Engineering Journal* 464 (2023), p. 142454. ISSN: 13858947. DOI: 10.1016/j.cej.2023.142454.

[31] Pawan Panwar, Quanpeng Yang, and Ashlie Martini. “Temperature-dependent density and viscosity prediction for hydrocarbons: machine learning and molecular dynamics simulations”. In: *Journal of Chemical Information and Modeling* 64 (2023), pp. 2760–2774.

[32] Rui Zhou et al. “Exploring high-performance viscosity index improver polymers via high-throughput molecular dynamics and explainable AI”. In: *npj Computational Materials* 11.1 (2025), p. 52.

[33] Ian Hutchings and Philip Shipway. *Tribology: friction and wear of engineering materials*. Butterworth-heinemann, 2017.

[34] Peter J Blau et al. *Friction, lubrication, and wear technology*. ASM Materials Park, USA, 1992.

[35] Ali Erdemir and Jean Michel Martin. “Superior wear resistance of diamond and DLC coatings”. In: *Current Opinion in Solid State and Materials Science* 22.6 (2018), pp. 243–254.

[36] Xiaobin Lu, MM Khonsari, and ERM Gelinck. “The Stribeck curve: experimental results and theoretical prediction”. In: *Journal of tribology* 128.4 (2006), pp. 789–794.

[37] HA Spikes. “Boundary lubrication and boundary films”. In: *Tribology series* 25 (1993), pp. 331–346.
[38] HA Spikes. “Mixed lubrication—an overview”. In: *Lubrication Science* 9.3 (1997), pp. 221–253.
[39] Hugh Spikes and Zhang Jie. “History, origins and prediction of elastohydrodynamic friction”. In: *Tribology Letters* 56 (2014), pp. 1–25.
[40] Yukio Hori. *Foundations of hydrodynamic lubrication*. Springer, 2006.
[41] Robert L Fusaro. *Lubrication of Space Systems (c)*. Tech. rep. NASA, 1995.
[42] Atul Bhattad. “Review on viscosity measurement: devices, methods and models”. In: *Journal of Thermal Analysis and Calorimetry* 148.14 (2023), pp. 6527–6543.
[43] Artur Wolak et al. “Experimental investigation of the viscosity parameters ranges—Case study of engine oils in the selected viscosity grade”. In: *Energies* 13.12 (2020), p. 3152.
[44] Fuels and Lubricants TC 1 Engine Lubrication. *Engine Oil Viscosity Classification*. SAE International, 2021. DOI: https://doi.org/10.4271/J300_202104. URL: https://doi.org/10.4271/J300_202104.
[45] NATO Standardization Office (NSO). *MIL-PRF-23699F – Lubricating Oil, Aircraft Turbine Engine, Synthetic Base*. Department of Defense Performance Specification. 2005.
[46] George E Totten. *Handbook of lubrication and tribology: Volume I application and maintenance*. CRC Press, 2006.
[47] Leslie R Rudnick. *Synthetics, mineral oils, and bio-based lubricants: chemistry and technology*. CRC press, 2020.
[48] Castrol Limited. *Safety Data Sheet: Castrol EDGE 5W-30*. https://msdspds.castrol.com/ussds/amersdsf.nsf/0/528D332C949412A28025839B0057F37E/\protect\T1\textdollarfile/11448585.pdf. 2018.
[49] Shell Global. *Safety Data Sheet: Shell Helix Ultra 5W-40*. https://www.epc.shell.com/DocumentManagement/BlobDocumentDownload?DocId=161736716. 2020.
[50] ExxonMobil Corporation. *Safety Data Sheet: Mobil 1 5W-30*. https://www.in.gov/fssa/thehub/files/Mobil-1-5W-30.pdf. Accessed: 2024-05-04. 2016.
[51] Thermtest Inc. *Why the Thermal Properties of Engine Oils are Important for Maintaining a Healthy Car Engine*. Accessed: 2024-05-04. 2021. URL: https://thermtest.com/why-the-thermal-properties-of-engine-oils-are-important-for-maintaining-a-healthy-car-engine.
[52] NJ George et al. “Lubricating and cooling capacities of different SAE 20W–50 engine oil samples using specific heat capacity and cooling rate”. In: *Archives of Physics Research* 1.2 (2010), pp. 103–111.
[53] ASTM International. *ASTM D471-16a(2021): Standard Test Method for Rubber Property—Effect of Liquids*. https://store.astm.org/d0471-16ar21.html. Accessed: 2024-05-04. 2021.
[54] Debashis Puhan. “Lubricant and Lubricant Additives”. In: *Tribology in Materials and Manufacturing*. Ed. by Amar Patnaik, Tej Singh, and Vikas Kukshal. Rijeka: IntechOpen, 2020. Chap. 9. DOI: 10.5772/intechopen.93830.
[55] Oscar Lacroix-Andrivet et al. “Characterization of Base Oil and Additive Oxidation Products from Formulated Lubricant by Ultra-High Resolution Mass Spectrometry”. In: *Lubricants* 11.8 (2023). ISSN: 2075-4442. DOI: 10.3390/lubricants11080345. URL: https://www.mdpi.com/2075-4442/11/8/345.
[56] A Jackson. “Synthetic versus mineral fluids in lubrication”. In: *International Tribology Conference 1987, Melbourne, 2-4 December 1987: Preprints of Papers*. Institution of Engineers, Australia Barton, ACT. 1987, pp. 428–437.
[57] David Tsui. “Why synthetics are leading the lubricants market”. In: *Tribology and Lubrication Technology* 73 (2017), p. 24.

[58] Ying Wang et al. “Is replacement of aromatic compounds with alkyl ones really the effective way toward developing greener lubricants?” In: *Journal of Molecular Liquids* 397 (2024), p. 124001.

[59] LD Moore et al. “PAO-based synthetic lubricants in industrial applications”. In: *Tribology & Lubrication Technology* 59.1 (2003), p. 23.

[60] S Boyde. “Hydrolytic stability of synthetic ester lubricants”. In: *Journal of synthetic lubrication* 16.4 (2000), pp. 297–312.

[61] Bruno Améduri. “The promising future of fluoropolymers”. In: *Macromolecular Chemistry and Physics* 221.8 (2020), p. 1900573.

[62] Darrius Dias et al. “Forever chemicals, per-and polyfluoroalkyl substances (PFAS), in lubrication”. In: *Lubricants* 12.4 (2024), p. 114.

[63] Chenfei Ren et al. “Antifoaming agent for lubricating oil: Preparation, mechanism and application”. In: *Molecules* 28.7 (2023), p. 3152.

[64] Robert G Rowland, Jun Dong, and Cyril A Migdal. “Antioxidants”. In: *Lubricant Additives*. CRC Press, 2017, pp. 3–36.

[65] Chiara Gattinoni and Angelos Michaelides. “Understanding corrosion inhibition with van der Waals DFT methods: the case of benzotriazole”. In: *Faraday discussions* 180 (2015), pp. 439–458.

[66] Leslie R Rudnick. “Additives for industrial lubricant applications”. In: *Lubricant Additives* (2017), pp. 473–484.

[67] Hugh Spikes. “Low-and zero-sulphated ash, phosphorus and sulphur anti-wear additives for engine oils”. In: *Lubrication science* 20.2 (2008), pp. 103–136.

[68] David W Johnson and John E Hils. “Phosphate esters, thiophosphate esters and metal thiophosphates as lubricant additives”. In: *Lubricants* 1.4 (2013), pp. 132–148.

[69] W David Phillips and Neal Milne. “Ashless phosphorus–containing lubricating oil additives”. In: *Lubricant Additives*. CRC Press, 2017, pp. 157–196.

[70] Leslie R Rudnick. *Lubricant additives: chemistry and applications*. CRC press, 2009.

[71] Yan Zhou and Jun Qu. “Ionic liquids as lubricant additives: A review”. In: *ACS Applied Materials and Interfaces* 9 (4 2017), pp. 3209–3222. ISSN: 19448252. DOI: 10.1021/acsami.6b12489.

[72] Ram Subbaraman, Sanket A Deshmukh, and Subramanian KRS Sankaranarayanan. “Atomistic insights into early stage oxidation and nanoscale oxide growth on Fe (100), Fe (111) and Fe (110) surfaces”. In: *The Journal of Physical Chemistry C* 117.10 (2013), pp. 5195–5207.

[73] Thi D Ta et al. “Adsorption of normal-alkanes on Fe (110), FeO (110), and Fe2O3 (0001): Influence of iron oxide surfaces”. In: *The Journal of Physical Chemistry C* 119.23 (2015), pp. 12999–13010.

[74] Yi Zhu et al. “A study on iron oxides and surface roughness in dry and wet wheel- rail contacts”. In: *Wear* 328 (2015), pp. 241–248.

[75] Ryan D. Evans et al. “Transmission electron microscopy of boundary-lubricated bearing surfaces. Part I: Mineral oil lubricant”. In: *Tribology Transactions* 47 (3 2004), pp. 430–439. ISSN: 10402004. DOI: 10.1080/05698190490463286.

[76] Ryan D. Evans et al. “Transmission electron microscopy of boundary-lubricated bearing surfaces. Part II: Mineral oil lubricant with sulfur- and phosphorus-containing gear oil additives”. In: *Tribology Transactions* 48 (3 2005), pp. 299–307. ISSN: 10402004. DOI: 10.1080/05698190590965602.

[77] Gareth S Parkinson. “Iron oxide surfaces”. In: *Surface Science Reports* 71.1 (2016), pp. 272–365.

[78] M. M. De Oliveira et al. “Effect of iron oxide debris on the reciprocating sliding wear of tool steels”. In: *Wear* 426-427 (Apr. 2019), pp. 1065–1075. ISSN: 00431648. DOI: 10.1016/j.wear.2018.12.047.

[79] HKDH Bhadeshia. “Cementite”. In: *International Materials Reviews* 65.1 (2020), pp. 1–27.

[80] M. A. Moore. "The relationship between the abrasive wear resistance, hardness and microstructure of ferritic materials". In: *Wear* 28 (1 Apr. 1974), pp. 59–68. ISSN: 00431648. DOI: 10.1016/0043-1648(74)90101-X.

[81] Yonggang Meng et al. "A review of recent advances in tribology". In: *Friction* 8 (2020), pp. 221–300.

[82] Ali Erdemir and Christophe Donnet. "Tribology of diamond-like carbon films: recent progress and future prospects". In: *Journal of Physics D: Applied Physics* 39.18 (2006), R311.

[83] Makoto Kano. "Overview of DLC-coated engine components". In: *Coating technology for vehicle applications* (2015), pp. 37–62.

[84] A Erdemir. "The role of hydrogen in tribological properties of diamond-like carbon films". In: *Surface and Coatings Technology* 146 (2001), pp. 292–297.

[85] Zhe Chen et al. "Effect of humidity on friction and wear—A critical review". In: *Lubricants* 6.3 (2018), p. 74.

[86] Jia Wang et al. "Improving frictional properties of DLC films by surface energy manipulation". In: *RSC advances* 8.21 (2018), pp. 11388–11394.

[87] Haibo Sun et al. "Effects of element doping on the structure and properties of diamond-like carbon films: a review". In: *Lubricants* 11.4 (2023), p. 186.

[88] Hugh Spikes. "Stress-augmented thermal activation: Tribology feels the force". In: *Friction* 6 (1 2018), pp. 1–31. ISSN: 22237704. DOI: 10.1007/s40544-018-0201-2.

[89] Nicholas Hopper et al. "Activation Volumes in Tribochemistry; What Do They Mean and How to Calculate Them?" In: *Tribology Letters* 73.2 (2025), p. 40.

[90] George I Bell. "Models for the specific adhesion of cells to cells: a theoretical framework for adhesion mediated by reversible bonds between cell surface molecules." In: *Science* 200.4342 (1978), pp. 618–627.

[91] M. G. Evans and M. Polanyi. "Some applications of the transition state method to the calculation of reaction velocities, especially in solution". In: *Transactions of the Faraday Society* 31 (1935), pp. 875–894. ISSN: 00147672. DOI: 10.1039/tf9353100875.

[92] H Eyring. "Viscosity, plasticity, and diffusion as examples of absolute reaction rates". In: *The Journal of Chemical Physics* 4 (1936), pp. 283–291. DOI: 10.1063/1.1749836.

[93] Wilfred Tysoe. "On stress-induced tribochemical reaction rates". In: *Tribology letters* 65 (2017), pp. 1–16.

[94] H Spikes and W Tysoe. "On the Commonality Between Theoretical Models for Fluid and Solid Friction, Wear and Tribochemistry". In: *Tribology Letters* 59 (2015), p. 21. ISSN: 1023-8883. DOI: 10.1007/s11249-015-0544-z.

[95] N. N. Gosvami et al. "Mechanisms of antiwear tribofilm growth revealed in situ by single-asperity sliding contacts". In: *Science* 348.6230 (2015), pp. 102–106. DOI: 10.1126/science.1258788.

[96] Jie Zhang et al. "Mechanochemistry of zinc dialkyldithiophosphate on steel surfaces under elastohydrodynamic lubrication conditions". In: *ACS Applied Materials & Interfaces* 12.5 (2020), pp. 6662–6676.

[97] MP Seah. "The quantitative analysis of surfaces by XPS: A review". In: *Surface and Interface Analysis* 2.6 (1980), pp. 222–239.

[98] Chuan C Chang. "Auger electron spectroscopy". In: *Surface Science* 25.1 (1971), pp. 53–79.

[99] Franz J Giessibl. "Advances in atomic force microscopy". In: *Reviews of modern physics* 75.3 (2003), p. 949.

[100] David B Williams et al. *The transmission electron microscope*. Springer, 2009.

[101] Robert Ian Taylor. "Rough surface contact modelling—a review". In: *Lubricants* 10.5 (2022), p. 98.

[102] W Gregory Sawyer and Kathryn J Wahl. "Accessing inaccessible interfaces: in situ approaches to materials tribology". In: *MRS Bulletin* 33.12 (2008), pp. 1145–1150.

[103] Andreas W Gotz et al. “Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born”. In: *Journal of chemical theory and computation* 8.5 (2012), pp. 1542–1555.

[104] Steven J. Plimpton and Aidan P. Thompson. “Computational aspects of many-body potentials”. In: *MRS Bulletin* 37 (5 2012), pp. 513–521. ISSN: 0883-7694. DOI: 10.1557/mrs.2012.96.

[105] Sebastian Schmitt et al. “Comparison of force fields for the prediction of thermophysical properties of long linear and branched alkanes”. In: *The Journal of Physical Chemistry B* 127.8 (2023), pp. 1789–1802.

[106] Sophie Loehlé and Maria Righi. “Ab Initio Molecular Dynamics Simulation of Tribochemical Reactions Involving Phosphorus Additives at Sliding Iron Interfaces”. In: *Lubricants* 6 (2 2018), p. 31. ISSN: 2075-4442. DOI: 10.3390/lubricants6020031.

[107] Takuya Kuwahara et al. “Mechano-chemical decomposition of organic friction modifiers with multiple reactive centres induces superlubricity of ta-C”. In: *Nature Communications* 10 (2019), p. 151. DOI: 10.1038/s41467-018-08042-8.

[108] Valentin R Salinas Ruiz et al. “Interplay of mechanics and chemistry governs wear of diamond-like carbon coatings interacting with ZDDP-additivated lubricants”. In: *Nature Communications* 12 (1 2021), p. 4550. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24766-6.

[109] Stefan Peeters et al. “Adsorption and decomposition of ZDDP on lightweight metallic substrates: Ab initio and experimental insights”. In: *Applied Surface Science* 600 (2022), p. 153947. ISSN: 0169-4332. DOI: https://doi.org/10.1016/j.apsusc.2022.153947.

[110] Jörg Behler. “Perspective: Machine learning potentials for atomistic simulations”. In: *Journal of Chemical Physics* 145 (17 Nov. 2016). ISSN: 00219606. DOI: 10.1063/1.4966192.

[111] Huong T.T. Ta et al. “Ab initio Informed Machine Learning Potential for Tribochemistry and Mechanochemistry: Application for Eco–friendly Gallate Lubricant Additive”. In: *Computational Materials Today* (May 2024), p. 100005. ISSN: 29504635. DOI: 10.1016/j.commt.2024.100005.

[112] Alberto Pacini et al. “Advancing tribological simulations of carbon-based lubricants with active learning and machine learning molecular dynamics”. In: *The European Physical Journal Plus* 139 (6 June 2024), p. 549. ISSN: 2190-5444. DOI: 10.1140/epjp/s13360-024-05348-z.

[113] Huong Thi Thuy Ta et al. “Probing additives for green lubricants with the aid of machine learning molecular dynamics: The case of gallate molecules for aqueous solutions”. In: *Applied Surface Science* 695 (June 2025). ISSN: 01694332. DOI: 10.1016/j.apsusc.2025.162836.

[114] Edward J Maginn et al. “Best practices for computing transport properties 1. Self-diffusivity and viscosity from equilibrium molecular dynamics [article v1. 0]”. In: *Living Journal of Computational Molecular Science* 1.1 (2019), pp. 6324–6324.

[115] Melville S Green. “Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids”. In: *The Journal of Chemical Physics* 22.3 (1954), pp. 398–413.

[116] Ryogo Kubo. “Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems”. In: *Journal of the Physical Society of Japan* 12.6 (1957), pp. 570–586.

[117] B D Todd and P J Daivis. “Homogeneous non-equilibrium molecular dynamics simulations of viscous flow: techniques and applications”. In: *Molecular Simulation* 33 (2007), pp. 189–229. DOI: 10.1080/08927020601026629.

[118] Luca Maffioli et al. “Slip and stress from low shear rate nonequilibrium molecular dynamics: The transient-time correlation function technique”. In: *The Journal of Chemical Physics* 156 (18 May 2022), p. 184111. ISSN: 0021-9606. DOI: 10.1063/5.0088127.

[119] Luca Maffioli et al. “TTCF4LAMMPS: A toolkit for simulation of the non-equilibrium behaviour of molecular fluids at experimentally accessible shear rates”. In: *Computer Physics Communications* 300 (Nov. 2024), p. 109205.

[120] Dominik Marx and Jurg Hutter. *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press, 2009.

[121] Loup Verlet. "Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules". In: *Physical Review* 159.1 (1967), p. 98.

[122] Helmut Grubmüller et al. "Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions". In: *Molecular Simulation* 6.1-3 (1991), pp. 121–142.

[123] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. "Reversible multiple time scale molecular dynamics". In: *The Journal of chemical physics* 97.3 (1992), pp. 1990–2001.

[124] S NOSE. "A molecular dynamics method for simulations in the canonical ensemble". In: *Molecular Physics* 100.1 (2002), pp. 191–198.

[125] William G Hoover. "Canonical dynamics: Equilibrium phase-space distributions". In: *Physical Review A* 31.3 (1985), p. 1695.

[126] Jiaqi He, Huajie Tang, and Chenglong Wang. "Advances of molecular dynamics simulation in tribochemistry and lubrication investigations: a review". In: *Journal of Industrial and Engineering Chemistry* 126 (2023), pp. 1–19.

[127] Stefano Bernardi, BD Todd, and Debra J Searles. "Thermostating highly confined fluids". In: *The Journal of chemical physics* 132.24 (2010).

[128] Toni Schneider and E Stoll. "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions". In: *Physical Review B* 17.3 (1978), p. 1302.

[129] Herman JC Berendsen et al. "Molecular dynamics with coupling to an external bath". In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[130] Xin Yong and Lucy T Zhang. "Thermostats and thermostat strategies for molecular dynamics simulations of nanofluidics". In: *The Journal of chemical physics* 138.8 (2013).

[131] S Nosé. "A molecular dynamics method for simulations in the canonical ensemble". In: *Molecular Physics* 52.2 (1984), pp. 255–268.

[132] Qia Ke et al. "Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations". In: *Journal of Molecular Liquids* 365 (2022), p. 120116. ISSN: 0167-7322. DOI: https://doi.org/10.1016/j.molliq.2022.120116.

[133] Chiara Gattinoni et al. "Boundary-controlled barostats for slab geometries in molecular dynamics simulations". In: *Physical Review E* 90.4 (2014), p. 043302.

[134] George S Fanourgakis, JS Medina, and Rita Prosmiti. "Determining the bulk viscosity of rigid water models". In: *The Journal of Physical Chemistry A* 116.10 (2012), pp. 2564–2570.

[135] Joseph E Basconi and Michael R Shirts. "Effects of temperature control algorithms on transport properties and kinetics in molecular dynamics simulations". In: *Journal of chemical theory and computation* 9.7 (2013), pp. 2887–2899.

[136] Hans C Andersen. "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393.

[137] Marcus G Martin. "Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor–liquid coexistence curves and liquid densities". In: *Fluid phase equilibria* 248.1 (2006), pp. 50–55.

[138] Oliver T Unke et al. "Machine learning force fields". In: *Chemical Reviews* 121.16 (2021), pp. 10142–10186.

[139] James P Ewen et al. "A comparison of classical force-fields for molecular dynamics simulations of lubricants". In: *Materials* 9.8 (2016), p. 651.

[140] Katie A Maerzke and J Ilja Siepmann. "Transferable Potentials for Phase Equilibria- Coarse-Grain Description for Linear Alkanes". In: *The Journal of Physical Chemistry B* 115.13 (2011), pp. 3452–3465.

[141] Michael Levitt. "The birth of computational structural biology". In: *Nature Structural Biology* 8 (5 May 2001), pp. 392–393. ISSN: 10728368. DOI: 10.1038/87545.

[142] Miguel A González. "Force fields and molecular dynamics simulations". In: *École thématique de la Société Française de la Neutronique* 12 (2011), pp. 169–200.

[143] Johannes Lenhard, Simon Stephan, and Hans Hasse. "On the History of the Lennard-Jones Potential". In: *Annalen der Physik* 536.6 (2024), p. 2400115.

[144] Jane S Murray and Peter Politzer. "The electrostatic potential: an overview". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 153–163.

[145] Paul P Ewald. "Die Berechnung optischer und elektrostatischer Gitterpotentiale". In: *Annalen der physik* 369.3 (1921), pp. 253–287.

[146] Roger W Hockney and James W Eastwood. *Computer simulation using particles*. CRC Press, 2021.

[147] Scott J Weiner et al. "A new force field for molecular mechanical simulation of nucleic acids and proteins". In: *Journal of the American Chemical Society* 106.3 (1984), pp. 765–784.

[148] Kenno Vanommeslaeghe et al. "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *Journal of Computational Chemistry* 31.4 (2010), pp. 671–690.

[149] 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". In: *Journal of the American Chemical Society* 118.45 (1996), pp. 11225–11236.

[150] Huai Sun. "COMPASS: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds". In: *The Journal of Physical Chemistry B* 102.38 (1998), pp. 7338–7364.

[151] Huai Sun et al. "An ab initio CFF93 all-atom force field for polycarbonates". In: *Journal of the American Chemical society* 116.7 (1994), pp. 2978–2987.

[152] Jon R Maple et al. "Derivation of class II force fields. I. Methodology and quantum force field for the alkyl functional group and alkane molecules". In: *Journal of Computational Chemistry* 15.2 (1994), pp. 162–182.

[153] Michael F Russo Jr and Adri CT Van Duin. "Atomistic-scale simulations of chemical reactions: Bridging from quantum chemistry to engineering". In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 269.14 (2011), pp. 1549–1554.

[154] Thomas P Senftle et al. "The ReaxFF reactive force-field: development, applications and future directions". In: *npj Computational Materials* 2.1 (2016), pp. 1–14.

[155] Adri CT Van Duin et al. "ReaxFF: a reactive force field for hydrocarbons". In: *The Journal of Physical Chemistry A* 105.41 (2001), pp. 9396–9409.

[156] Anthony K Rappe and William A Goddard III. "Charge equilibration for molecular dynamics simulations". In: *The Journal of Physical Chemistry* 95.8 (1991), pp. 3358–3363.

[157] Kimberly Chenoweth, Adri C.T. Van Duin, and William A. Goddard. "ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation". In: *Journal of Physical Chemistry A* 112 (5 2008), pp. 1040–1053. ISSN: 10895639. DOI: 10.1021/jp709896w.

[158] Da Chuan Yue et al. "Tribochemistry of phosphoric acid sheared between quartz surfaces: A reactive molecular dynamics study". In: *Journal of Physical Chemistry C* 117 (48 2013), pp. 25604–25614. ISSN: 19327447. DOI: 10.1021/jp406360u.

[159] Arash Khajeh et al. "Statistical analysis of tri-cresyl phosphate conversion on an iron oxide surface using reactive molecular dynamics simulations". In: *The Journal of Physical Chemistry C* 123.20 (2019), pp. 12886–12893.

[160] Jejoon Yeon et al. “Development of a ReaxFF force field for Cu/S/C/H and reactive MD simulations of methyl thiolate decomposition on Cu (100)”. In: *The Journal of Physical Chemistry B* 122.2 (2018), pp. 888–896.

[161] Amir Hajibabaei et al. “Machine learning of first-principles force-fields for alkane and polyene hydrocarbons”. In: *The Journal of Physical Chemistry A* 125.42 (2021), pp. 9414–9420.

[162] Carlos M Fonseca and Peter J Fleming. “An overview of evolutionary algorithms in multiobjective optimization”. In: *Evolutionary computation* 3.1 (1995), pp. 1–16.

[163] Lothar Terfloth and Johann Gasteiger. “Neural networks and genetic algorithms in drug design”. In: *Drug Discovery Today* 6 (2001), pp. 102–108.

[164] Dominique Douguet, Etienne Thoreau, and Gérard Grassy. “A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm”. In: *Journal of computer-aided molecular design* 14 (2000), pp. 449–466.

[165] A Jahan et al. “Material screening and choosing methods–a review”. In: *Materials & Design* 31.2 (2010), pp. 696–705.

[166] Paul C Jennings et al. “Genetic algorithms for computational materials discovery accelerated by machine learning”. In: *NPJ Computational Materials* 5.1 (2019), p. 46.

[167] Lingaraj Haldurai, T Madhubala, and R Rajalakshmi. “A study on genetic algorithm and its applications”. In: *Int. J. Comput. Sci. Eng* 4.10 (2016), pp. 139–143.

[168] Annu Lambora, Kunal Gupta, and Kriti Chopra. “Genetic algorithm-A literature review”. In: *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE. 2019, pp. 380–384.

[169] Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. “Multi-objective optimization”. In: *Decision sciences*. CRC Press, 2016, pp. 161–200.

[170] Abdullah Konak, David W Coit, and Alice E Smith. “Multi-objective optimization using genetic algorithms: A tutorial”. In: *Reliability engineering & system safety* 91.9 (2006), pp. 992–1007.

[171] R Timothy Marler and Jasbir S Arora. “The weighted sum method for multi-objective optimization: new insights”. In: *Structural and multidisciplinary optimization* 41 (2010), pp. 853–862.

[172] Kalyanmoy Deb et al. “A fast and elitist multiobjective genetic algorithm: NSGA-II”. In: *IEEE transactions on evolutionary computation* 6.2 (2002), pp. 182–197.

[173] Kalyanmoy Deb et al. “A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II”. In: *International conference on parallel problem solving from nature*. Springer. 2000, pp. 849–858.

[174] Richard A Watson and Jordan B Pollack. “Symbiotic combination as an alternative to sexual recombination in genetic algorithms”. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2000, pp. 425–434.

[175] Carlo R Raquel and Prospero C Naval Jr. “An effective use of crowding distance in multiobjective particle swarm optimization”. In: *Proceedings of the 7th Annual conference on Genetic and Evolutionary Computation*. 2005, pp. 257–264.

[176] Haiming Du et al. “Elitism and distance strategy for selection of evolutionary algorithms”. In: *IEEe Access* 6 (2018), pp. 44531–44541.

[177] Marco Laumanns et al. “Combining convergence and diversity in evolutionary multiobjective optimization”. In: *Evolutionary computation* 10.3 (2002), pp. 263–282.

[178] Lyndon While et al. “Heuristics for optimizing the calculation of hypervolume for multi-objective optimization problems”. In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 3. IEEE. 2005, pp. 2225–2232.

[179] Johannes Bader and Eckart Zitzler. “HypE: An algorithm for fast hypervolume-based many-objective optimization”. In: *Evolutionary computation* 19.1 (2011), pp. 45–76.

[180] Joshua D Knowles, David W Corne, and Mark Fleischer. “Bounded archiving using the Lebesgue measure”. In: *The 2003 Congress on Evolutionary Computation, 2003. CEC’03.* Vol. 4. IEEE. 2003, pp. 2490–2497.

[181] Lucas Bradstreet. *The hypervolume indicator for multi-objective optimisation: calculation and use*. University of Western Australia Perth, 2011.

[182] Andreia P Guerreiro, Carlos M Fonseca, and Luís Paquete. “The hypervolume indicator: Computational problems and algorithms”. In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–42.

[183] Karl Bringmann and Tobias Friedrich. “Approximation quality of the hypervolume indicator”. In: *Artificial Intelligence* 195 (2013), pp. 265–290.

[184] Kalyanmoy Deb. “Multi-objective optimisation using evolutionary algorithms: an introduction”. In: *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, 2011, pp. 3–34.

[185] Michael TM Emmerich and André H Deutz. “A tutorial on multiobjective optimization: fundamentals and evolutionary methods”. In: *Natural computing* 17 (2018), pp. 585–609.

[186] Eckart Zitzler and Lothar Thiele. “Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach”. In: *IEEE transactions on Evolutionary Computation* 3.4 (1999), pp. 257–271.

[187] Christopher K Oei, David E Goldberg, and Shau-Jin Chang. “Tournament selection, niching, and the preservation of diversity”. In: *IlliGAL report* 91011 (1991).

[188] Adam Lipowski and Dorota Lipowska. “Roulette-wheel selection via stochastic acceptance”. In: *Physica A: Statistical Mechanics and its Applications* 391.6 (2012), pp. 2193–2196.

[189] David E Goldberg, Jon Richardson, et al. “Genetic algorithms with sharing for multimodal function optimization”. In: *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*. Vol. 4149. Cambridge, MA. 1987, pp. 414–425.

[190] Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. “Estimation of the size of drug-like chemical space based on GDB-17 data”. In: *Journal of computer-aided molecular design* 27 (2013), pp. 675–679.

[191] Keith T Butler et al. “Machine learning for molecular and materials science”. In: *Nature* 559.7715 (2018), pp. 547–555.

[192] Eugene Helfand. “Transport coefficients from dissipation in a canonical ensemble”. In: *Physical Review* 119.1 (1960), p. 1.

[193] Ruichen Liu et al. “Design of fuel molecules based on variational autoencoder”. In: *Fuel* 316 (2022), p. 123426.

[194] Dimitrios Christofidellis et al. “Unifying molecular and textual representations via multi-task language modelling”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 6140–6157.

[195] Carl Edwards, ChengXiang Zhai, and Heng Ji. “Text2mol: Cross-modal molecule retrieval with natural language queries”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 595–607.

[196] Jonathan Pirnay et al. “GraphXForm: Graph transformer for computer-aided molecular design”. In: *Digital Discovery* (2025).

[197] Tianle Yue et al. “Benchmarking study of deep generative models for inverse polymer design”. In: *Digital Discovery* (2025).

[198] Josep Arús-Pous et al. “Randomized SMILES strings improve the quality of molecular generative models”. In: *Journal of cheminformatics* 11 (2019), pp. 1–13.

[199] Rocío Mercado et al. “Graph networks for molecular design”. In: *Machine Learning: Science and Technology* 2.2 (2021), p. 025023.

[200] Shingo Otsuka et al. “PoLyInfo: Polymer database for polymeric materials design”. In: *2011 International Conference on Emerging Intelligent Data and Web Technologies*. IEEE. 2011, pp. 22–29.

[201] Ruimin Ma and Tengfei Luo. “PI1M: a benchmark database for polymer informatics”. In: *Journal of Chemical Information and Modeling* 60.10 (2020), pp. 4684–4690.

[202] Xiang Huang et al. “Exploring high thermal conductivity polymers via interpretable machine learning with physical descriptors”. In: *npj Computational Materials* 9.1 (2023), p. 191.

[203] Robert C. Glen and AWR Payne. “A genetic algorithm for the automated generation of molecules within constraints”. In: *Journal of Computer-Aided Molecular Design* 9 (1995), pp. 181–202.

[204] Robert P Sheridan and Simon K Kearsley. “Using a genetic algorithm to suggest combinatorial libraries”. In: *Journal of Chemical Information and Computer Sciences* 35.2 (1995), pp. 310–320.

[205] Zipeng Zhong et al. “Recent advances in deep learning for retrosynthesis”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 14.1 (2024), e1694.

[206] Pavel Polishchuk. “CReM: chemically reasonable mutations framework for structure generation”. In: *Journal of Cheminformatics* 12.1 (2020), p. 28.

[207] Xiangxiang Zeng et al. “Deep generative molecular design reshapes drug discovery”. In: *Cell Reports Medicine* 3.12 (2022).

[208] Andreas Rosenkranz et al. “The use of artificial intelligence in tribology—A perspective”. In: *Lubricants* 9.1 (2020), p. 2.

[209] Uma Maheshwera Reddy Paturi, Sai Teja Palakurthy, and NS Reddy. “The role of machine learning in tribology: a systematic review”. In: *Archives of Computational Methods in Engineering* 30.2 (2023), pp. 1345–1397.

[210] J Echávarri Otero et al. “Artificial neural network approach to predict the lubricated friction coefficient”. In: *Lubrication Science* 26.3 (2014), pp. 141–162.

[211] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. “Comparative study between traditional machine learning and deep learning approaches for text classification”. In: *Proceedings of the ACM Symposium on Document Engineering 2018*. 2018, pp. 1–11.

[212] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.

[213] J. Ross Quinlan. “Induction of decision trees”. In: *Machine learning* 1 (1986), pp. 81–106.

[214] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[215] Tianqi Chen et al. “Xgboost: extreme gradient boosting”. In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[216] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[217] Pedro Domingos. “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10 (2012), pp. 78–87.

[218] Rudolf Kruse et al. “Multi-layer perceptrons”. In: *Computational intelligence: a methodological introduction*. Springer, 2022, pp. 53–124.

[219] Keiron O’shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).

[220] Ashish Vaswani et al. “Attention is all you need”. In: *31st Conference on Neural Information Processing Systems*. Vol. 30. 2017.

[221] Zonghan Wu et al. “A comprehensive survey on graph neural networks”. In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[222] Daniel C Elton et al. “Deep learning for molecular design—a review of the state of the art”. In: *Molecular Systems Design & Engineering* 4.4 (2019), pp. 828–849.

[223] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. “Machine learning interpretability: A survey on methods and metrics”. In: *Electronics* 8.8 (2019), p. 832.

[224] Jennifer N Wei, David Duvenaud, and Alán Aspuru-Guzik. “Neural networks for the prediction of organic chemistry reactions”. In: *ACS central science* 2.10 (2016), pp. 725–732.

[225] Hüseyin Çamur and Ahmed Muayad Rashid Al-Ani. “Prediction of oxidation stability of biodiesel derived from waste and refined vegetable oils by statistical approaches”. In: *Energies* 15.2 (2022), p. 407.

[226] Marcus Olivecrona et al. “Molecular de-novo design through deep reinforcement learning”. In: *Journal of cheminformatics* 9 (2017), pp. 1–14.

[227] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. “Introduction to multi-layer feed-forward neural networks”. In: *Chemometrics and intelligent laboratory systems* 39.1 (1997), pp. 43–62.

[228] Yoshifusa Ito. “Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory”. In: *Neural Networks* 4.3 (1991), pp. 385–394.

[229] Abien Fred Agarap. “Deep learning using rectified linear units (relu)”. In: *arXiv preprint arXiv:1803.08375* (2018).

[230] Andrea Apicella et al. “A survey on modern trainable activation functions”. In: *Neural Networks* 138 (2021), pp. 14–32.

[231] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[232] Shun-ichi Amari. “Backpropagation and stochastic gradient descent method”. In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[233] Zijun Zhang. “Improved adam optimizer for deep neural networks”. In: *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee. 2018, pp. 1–2.

[234] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[235] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[236] Qi Wang et al. “A comprehensive survey of loss functions in machine learning”. In: *Annals of Data Science* 9.2 (2022), pp. 187–212.

[237] Dejun Jiang et al. “Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models”. In: *Journal of cheminformatics* 13 (2021), pp. 1–23.

[238] Ismail Babajide Mustapha and Faisal Saeed. “Bioactive molecule prediction using extreme gradient boosting”. In: *Molecules* 21.8 (2016), p. 983.

[239] Jing Dong et al. “XGBoost-based intelligence yield prediction and reaction factors analysis of amination reaction”. In: *Journal of Computational Chemistry* 43.4 (2022), pp. 289–302.

[240] Alex K. Chew et al. “Advancing material property prediction: using physics-informed machine learning models for viscosity”. In: *Journal of Cheminformatics* 16 (1 2024). ISSN: 17582946. DOI: 10.1186/s13321-024-00820-5.

[241] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[242] Jan Salomon Cramer. *The origins of logistic regression*. Tech. rep. Tinbergen Institute discussion paper, 2002.

[243] Sepp Hochreiter. “The vanishing gradient problem during learning recurrent neural nets and problem solutions”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[244] Lukasz Maziarka et al. “Molecule attention transformer”. In: *arXiv preprint arXiv:2002.08264* (2020).

[245] Dong Chen et al. “Algebraic graph-assisted bidirectional transformers for molecular property prediction”. In: *Nature communications* 12.1 (2021), p. 3521.

[246] Yu Rong et al. “Self-supervised graph transformer on large-scale molecular data”. In: *Advances in neural information processing systems* 33 (2020), pp. 12559–12571.

[247] Juncai Li and Xiaofei Jiang. “Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction”. In: *Wireless Communications and Mobile Computing* 2021.1 (2021), p. 7181815.

[248] Ross Irwin et al. “Chemformer: a pre-trained transformer for computational chemistry”. In: *Machine Learning: Science and Technology* 3.1 (2022), p. 015022.

[249] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. “ChemBERTa: large-scale self-supervised pretraining for molecular property prediction”. In: *arXiv preprint arXiv:2010.09885* (2020).

[250] Yunwu Liu et al. “MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction”. In: *Journal of Molecular Graphics and Modelling* 118 (2023), p. 108344.

[251] Walid Ahmad et al. “Chemberta-2: Towards chemical foundation models”. In: *arXiv preprint arXiv:2209.01712* (2022).

[252] Annie M Westerlund et al. “Do chemformers dream of organic matter? evaluating a transformer model for multistep retrosynthesis”. In: *Journal of Chemical Information and Modeling* 64.8 (2024), pp. 3021–3033.

[253] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *Journal of machine learning research* 21.140 (2020), pp. 1–67.

[254] Edward J Hu et al. “Lora: Low-rank adaptation of large language models.” In: *ICLR* 1.2 (2022), p. 3.

[255] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[256] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[257] WJ Dunn Iii, DR Scott, and WG Glen. “Principal components analysis and partial least squares regression”. In: *Tetrahedron computer methodology* 2.6 (1989), pp. 349–376.

[258] Papia Ray, S Surender Reddy, and Tuhina Banerjee. “Various dimension reduction techniques for high dimensional data analysis: a review”. In: *Artificial Intelligence Review* 54.5 (2021), pp. 3473–3515.

[259] Etienne Becht et al. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature biotechnology* 37.1 (2019), pp. 38–44.

[260] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. “Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study”. In: *International conference on image and signal processing*. Springer. 2020, pp. 317–325.

[261] David M Chan et al. “t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data”. In: *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. IEEE. 2018, pp. 330–338.

[262] Rasmus Bro and Age K Smilde. “Principal component analysis”. In: *Analytical methods* 6.9 (2014), pp. 2812–2831.

[263] Isabelle Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine Learning* 46.1 (2002), pp. 389–422.

[264] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[265] Jesse Jeremiah Tanimu et al. "A machine learning method for classification of cervical cancer". In: *Electronics* 11.3 (2022), p. 463.

[266] Yao Li et al. "Urban flood susceptibility mapping based on social media data in Chengdu city, China". In: *Sustainable Cities and Society* 88 (2023), p. 104307.

[267] Basim Ahmad Alabsi, Mohammed Anbar, and Shaza Dawood Ahmed Rihan. "CNN-CNN: dual convolutional neural network approach for feature selection and attack detection on internet of things networks". In: *Sensors* 23.14 (2023), p. 6507.

[268] Christiane Ehrt et al. "SMARTS. plus–a toolbox for chemical pattern design". In: *Molecular informatics* 39.12 (2020), p. 2000216.

[269] Daylight Chemical Information Systems. *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. 2019.

[270] Robert Schmidt et al. "Comparing molecular patterns using the example of SMARTS: theory and algorithms". In: *Journal of Chemical Information and Modeling* 59.6 (2019), pp. 2560–2571.

[271] Egheosa Ogbomo et al. "Effects of surface chemistry on the mechanochemical decomposition of tricresyl phosphate". In: *Physical Chemistry Chemical Physics* 26 (2024), pp. 278–292. ISSN: 1463-9076. DOI: 10.1039/D3CP05320B.

[272] Robert T O'Neill and Roman Boulatov. "The many flavours of mechanochemistry and its plausible conceptual underpinnings". In: *Nature Reviews Chemistry* 5.3 (2021), pp. 148–167.

[273] Evelina Colacino et al. "Processing and Investigation Methods in Mechanochemical Kinetics". In: *ACS Omega* 3 (8 2018), pp. 9196–9209. ISSN: 24701343. DOI: 10.1021/acsomega.8b01431.

[274] Stuart L James et al. "Mechanochemistry: opportunities for new and cleaner synthesis". In: *Chemical Society Reviews* 41 (2012), pp. 413–447. DOI: 10.1039/c1cs15171a.

[275] Adam A.L. Michalchuk, Ivan A. Tumanov, and Elena V. Boldyreva. "Ball size or ball mass-what matters in organic mechanochemical synthesis?" In: *CrystEngComm* 21 (13 2019), pp. 2174–2179. ISSN: 14668033. DOI: 10.1039/c8ce02109k.

[276] Rebecca A. Haley et al. "Nickel Catalysis in a High Speed Ball Mill: A Recyclable Mechanochemical Method for Producing Substituted Cyclooctatetraene Compounds". In: *ACS Sustainable Chemistry and Engineering* 4 (5 2016), pp. 2464–2469. ISSN: 21680485. DOI: 10.1021/acssuschemeng.6b00363.

[277] Suhmi Hwang, Sven Grätz, and Lars Borchardt. "A guide to direct mechanocatalysis". In: *Chemical Communications* 58 (2022), pp. 1661–1671. ISSN: 1359-7345. DOI: 10.1039/D1CC05697B.

[278] Alejandro Boscoboinik et al. "Measuring and modelling mechanochemical reaction kinetics". In: *Chemical Communications* 56 (56 2020), pp. 7730–7733. DOI: 10.1039/D0CC02992K.

[279] Jie Zhang and Hugh Spikes. "On the mechanism of ZDDP antiwear film formation". In: *Tribology Letters* 63.2 (2016), p. 24.

[280] Prashant Mittal et al. "In situ microscale study of tribology and growth of ZDDP antiwear tribofilms on an Al-Si alloy". In: *Tribology International* 151 (2020), p. 106419. ISSN: 0301679X. DOI: 10.1016/j.triboint.2020.106419.

[281] Christopher McFadden, Cristian Soto, and Nicholas D Spencer. "Adsorption and surface chemistry in tribology". In: *Tribology International* 30.12 (1997), pp. 881–888. ISSN: 0301-679X. DOI: https://doi.org/10.1016/S0301-679X(97)00075-3.

[282] Mao Ueda, Amir Kadiric, and Hugh Spikes. "ZDDP tribofilm formation on non-ferrous surfaces". In: *Tribology Online* 15 (5 2020), pp. 318–331. ISSN: 18812198. DOI: 10.2474/TROL.15.318.

[283] Mao Ueda, Amir Kadiric, and Hugh Spikes. “Influence of Steel Surface Composition on ZDDP Tribofilm Growth Using Ion Implantation”. In: *Tribology Letters* 69 (2021), p. 62. ISSN: 1573-2711. DOI: 10.1007/s11249-021-01436-8.

[284] Xin He and Seong H. Kim. “Surface Chemistry Dependence of Mechanochemical Reaction of Adsorbed Molecules - An Experimental Study on Tribopolymerization of α-Pinene on Metal, Metal Oxide, and Carbon Surfaces”. In: *Langmuir* 34 (7 2018), pp. 2432–2440. ISSN: 15205827. DOI: 10.1021/acs.langmuir.7b03763.

[285] Arash Khajeh et al. “Mechanochemical association reaction of interfacial molecules driven by shear”. In: *Langmuir* 34 (21 2018), pp. 5971–5977. ISSN: 15205827. DOI: 10.1021/acs.langmuir.8b00315.

[286] Jejoon Yeon et al. “Mechanochemistry at solid surfaces: polymerization of adsorbed molecules by mechanical shear at tribological interfaces”. In: *ACS Applied Materials & Interfaces* 9.3 (2017), pp. 3142–3148.

[287] Bihan Guan, Bernadeta A. Pochopien, and Dominic S. Wright. “The chemistry, mechanism and function of tricresyl phosphate (TCP) as an anti-wear lubricant additive”. In: *Lubrication Science* 28 (5 2016), pp. 257–265. ISSN: 15576833. DOI: 10.1002/ls.1327.

[288] Otto Beeck, JW Givens, and EC Williams. “On the mechanism of boundary lubrication. II. Wear prevention by addition agents”. In: *Proceedings of the Royal Society of London, Series A* 177.968 (1940), pp. 103–118.

[289] Du Hee Han and Masabumi Masuko. “Comparison of antiwear additive response among several base oils of different polarities”. In: *Tribology Transactions* 42 (4 1999), pp. 902–906. ISSN: 1547397X. DOI: 10.1080/10402009908982299.

[290] Jake Airey et al. “The effect of aviation anti-wear additives on tribofilm formation and micropitting propensity”. In: *Proceedings of the Institution of Mechanical Engineers, Part J* 237 (2023), pp. 1548–1567. ISSN: 2041305X. DOI: 10.1177/13506501231167621.

[291] Robert L. Jones, Harold Ravner, and Robert L. Cottington. “Inhibition of iron-catalyzed neopentyl polyol ester thermal degradation through passivation of the active metal surface by tricresyl phosphate”. In: *ASLE Transactions* 13 (1 1970), pp. 1–10. ISSN: 05698197. DOI: 10.1080/05698197008972276.

[292] FT Barcroft and SG Daniel. “The action of neutral organic phosphates as EP additives”. In: *Journal of Basic Engineering* 87 (1965), pp. 761–767.

[293] Douglas Godfrey. “The lubrication mechanism of tricresyl phosphate on steel”. In: *ASLE Transactions* 8.1 (1965), pp. 1–11.

[294] Owen D Faut and Donald R Wheeler. “On the mechanism of lubrication by tricresylphosphate (TCP)—the coefficient of friction as a function of temperature for TCP on M-50 steel”. In: *ASLE Transactions* 26.3 (1983), pp. 344–350.

[295] A. H. J. M. Gauthier, H. Montes, and J. M. Georges. “Boundary lubrication with tricresylphosphate (TCP). Importance of corrosive wear”. In: *ASLE Transactions* 25.4 (1982), pp. 445–455.

[296] CS Saba and NH Forster. “Reactions of aromatic phosphate esters with metals and their oxides”. In: *Tribology Letters* 12 (2002), pp. 135–146.

[297] David W Johnson et al. “Vapor-phase lubrication: Reaction of phosphate ester vapors with iron and steel”. In: *Chemistry of Materials* 14.9 (2002), pp. 3767–3775.

[298] B Acharya et al. “In situ, real time studies of thermal reaction film formation temperatures for iron and 304ss surfaces immersed in 5% tricresyl phosphate in base oil”. In: *Tribology International* 126 (2018), pp. 106–115.

[299] Donald R Wheeler and Owen D Faut. “The adsorption and thermal decomposition of tricresylphosphate (TCP) on iron and gold”. In: *Applications of Surface Science* 18.1-2 (1984), pp. 106–122.

[300] D Sung and AJ Gellman. “The surface chemistry of alkyl and arylphosphate vapor phase lubricants on Fe foil”. In: *Tribology International* 35.9 (2002), pp. 579–590.

[301] Arash Khajeh et al. “Thermal decomposition of tricresyl phosphate on ferrous surfaces”. In: *The Journal of Physical Chemistry C* 125.9 (2021), pp. 5076–5087.

[302] H. K. D. H. Bhadeshia. “Steels for bearings”. In: *Progress in Materials Science* 57 (2 2012), pp. 268–435. ISSN: 00796425. DOI: 10.1016/j.pmatsci.2011.06.002.

[303] David W. Johnson, John E. Hils, and Nelson Forster. “Interaction of polyol esters and phosphate esters with metal carbides”. In: *Tribology Letters* 42 (2 2011), pp. 223–232. ISSN: 10238883. DOI: 10.1007/s11249-011-9766-x.

[304] James P Ewen et al. “Substituent effects on the thermal decomposition of phosphate esters on ferrous surfaces”. In: *The Journal of Physical Chemistry C* 124.18 (2020), pp. 9852–9865.

[305] J. F. Makki and E. E. Graham. “Vapor phase deposition on high temperature surfaces”. In: *Tribology Transactions* 33 (4 1990), pp. 595–603. ISSN: 1547397X. DOI: 10.1080/10402009008981994.

[306] Alexander Stukowski. “Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool”. In: *Modelling and Simulation in Materials Science and Engineering* 18.1 (2009), p. 015012.

[307] J. P. Ewen, D. M. Heyes, and D. Dini. “Advances in nonequilibrium molecular dynamics simulations of lubricants and additives”. In: *Friction* 6 (4 2018), pp. 349–386. ISSN: 22237704. DOI: 10.1007/s40544-018-0207-9.

[308] Chris Winder and Jean Christophe Balouet. “The toxicity of commercial jet oils”. In: *Environmental Research* 89 (2 2002), pp. 146–164. ISSN: 00139351. DOI: 10.1006/enrs.2002.4346.

[309] Renguo Lu et al. “Deactivation Effect of Tricresyl Phosphate (TCP) on Tribochemical Decomposition of Hydrocarbon Oil on a Nascent Steel Surface”. In: *Tribology Letters* 33 (2009), pp. 1–8. DOI: 10.1007/s11249-008-9379-1.

[310] E. E. Klaus et al. “Structure of films formed during the deposition of lubrication molecules on iron and silicon carbide”. In: *Tribology Transactions* 33 (1 1990), pp. 25–32. ISSN: 1547397X. DOI: 10.1080/10402009008981926.

[311] P. Blonski and A. Kiejna. “Structural, electronic, and magnetic properties of BCC iron surfaces”. In: *Surface Science* 601 (1 2007), pp. 123–133. ISSN: 00396028. DOI: 10.1016/j.susc.2006.09.013.

[312] Wun C. Chiou and Emily A. Carter. “Structure and stability of Fe3C-cementite surfaces from first principles”. In: *Surface Science* 530 (1-2 2003), pp. 88–100. ISSN: 00396028. DOI: 10.1016/S0039-6028(03)00352-2.

[313] R. Bliem et al. “Subsurface cation vacancy stabilization of the magnetite (001) surface”. In: *Science.* 346 (6214 2014), pp. 1215–1218. ISSN: 10959203. DOI: 10.1126/science.1260556.

[314] Jie Zhang, James P Ewen, and Hugh A Spikes. “Substituent effects on the mechanochemical response of zinc dialkyldithiophosphate”. In: *Molecular Systems Design & Engineering* 7.9 (2022), pp. 1045–1055.

[315] George L Clark and James V Robinson. “A Study of Monolayers of Some Esters and Chlorinated Derivatives Possibly Useful as Lubricating Addition Agents”. In: *Journal of the American Chemical Society* 62 (1940), pp. 1948–1951.

[316] E. E. Graham and E. E. Klaus. “Lubrication from the vapor phase at high temperatures”. In: *ASLE Transactions* 29 (2 1986), pp. 229–234. DOI: 10.1080/05698198608981682.

[317] Karen Mohammadtabar et al. “Shear-driven reactions of organosulfur compounds on ferrous surfaces: A molecular dynamics study”. In: *Tribology International* 176 (2022), p. 107922. ISSN: 0301-679X. DOI: https://doi.org/10.1016/j.triboint.2022.107922.

[318] Sota Hayashi et al. “Pressure-assisted decomposition of tricresyl phosphate on amorphous FeO using hybrid quantum-classical simulations”. In: *Journal of Computational Chemistry* 44 (2023), pp. 766–776. ISSN: 0192-8651. DOI: 10.1002/jcc.27039.

[319] Karen Mohammadtabar et al. “Heat-, Load-, and Shear-Driven Reactions of Di-tert-butyl Disulfide on Fe(100)”. In: *The Journal of Physical Chemistry C* 123 (2019), pp. 19688–19692. ISSN: 1932-7447. DOI: 10.1021/acs.jpcc.9b05068.

[320] S. Peeters et al. “Importance of the catalytic effect of the substrate in the functionality of lubricant additives: the case of molybdenum dithiocarbamates”. In: *Materials Today Chemistry* 21 (2021), p. 100487. ISSN: 24685194. DOI: 10.1016/j.mtchem.2021.100487.

[321] L. Wang and R. J.K. Wood. “The influence of contact conditions on surface reaction layers formed between steel surfaces lubricated by an aviation oil”. In: *Tribology International* 40 (2007), pp. 1655–1666. ISSN: 0038092X. DOI: 10.1016/j.solener.2007.02.005.

[322] David W Johnson, Matthew Bachus, and John E Hils. “Interaction between lubricants containing phosphate ester additives and stainless steels”. In: *Lubricants* 1.2 (2013), pp. 48–60.

[323] Aidan P Thompson et al. “LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales”. In: *Computer Physics Communications* 271 (2022), p. 108171.

[324] T Schneider and E Stoll. “Molecular-dynamics study of a three-dimensional one-component model for distortive phase-transitions”. In: *Physical Review B* 17 (3 1978), pp. 1302–1322. ISSN: 0163-1829. DOI: 10.1103/PhysRevB.17.1302.

[325] Xin Yong and Lucy T. Zhang. “Thermostats and thermostat strategies for molecular dynamics simulations of nanofluidics”. In: *The Journal of Chemical Physics* 138 (2013), p. 084503. ISSN: 00219606. DOI: 10.1063/1.4792202.

[326] Fakhrul H. Bhuiyan, Seong H. Kim, and Ashlie Martini. “Reactive Molecular Dynamics Simulations of Thermal and Shear-Driven Oligomerization”. In: *Applied Surface Science* 591 (2022), p. 153209. ISSN: 01694332. DOI: 10.1016/j.apsusc.2022.153209.

[327] M Koyama et al. “Tribochemical reaction dynamics of phosphoric ester lubricant additive by using a hybrid tight-binding quantum chemical molecular dynamics method”. In: *Journal of Physical Chemistry B* 110 (35 2006), pp. 17507–17511. ISSN: 1520-6106. DOI: 10.1021/jp061210m.

[328] Masoud Aryanpour, Adri CT van Duin, and James D Kubicki. “Development of a reactive force field for iron- oxyhydroxide systems”. In: *The Journal of Physical Chemistry A* 114.21 (2010), pp. 6298–6307.

[329] Chenyu Zou, Adri CT Van Duin, and Dan C Sorescu. “Theoretical investigation of hydrogen adsorption and dissociation on iron and iron carbide surfaces using the ReaxFF reactive force field method”. In: *Topics in Catalysis* 55 (2012), pp. 391–401.

[330] CCW Verlackt et al. “Atomic-scale insight into the interactions between hydroxyl radicals and DNA in solution using the ReaxFF reactive force field”. In: *New Journal of Physics* 17.10 (2015), p. 103005.

[331] Liam S Morrissey et al. “Atomistic uniaxial tension tests: investigating various many-body potentials for their ability to produce accurate stress strain curves using molecular dynamics simulations”. In: *Molecular Simulation* 45.6 (2019), pp. 501–508.

[332] John J Adams et al. “Elastic constants of monocrystal iron from 3 to 500 K”. In: *Journal of Applied Physics* 100.11 (2006), p. 113530.

[333] J. Li et al. “Compression of Fe3C to 30 GPa at room temperature”. In: *Physics and Chemistry of Minerals* 29 (3 2002), pp. 166–169. ISSN: 03421791. DOI: 10.1007/s00269-001-0224-4.

[334] H J Reichmann and S D Jacobsen. “High-pressure elasticity of a natural magnetite crystal”. In: *American Mineralogist* 89 (2004), pp. 1061–1066.

[335] Md Mahbubul Islam et al. “Interactions of hydrogen with the iron and iron carbide interfaces: A ReaxFF molecular dynamics study”. In: *Physical Chemistry Chemical Physics* 18 (2 2015), pp. 761–771. ISSN: 14639076. DOI: 10.1039/c5cp06108c.

[336] Thi Dinh Ta et al. “Reactive Molecular Dynamics Study of Hierarchical Tribochemical Lubricant Films at Elevated Temperatures”. In: *ACS Applied Nano Materials* 3 (2020), pp. 2687–2704. DOI: 10.1021/acsanm.0c00042.

[337] Zhuohan Li and Izabela Szlufarska. “Physical origin of the mechanochemical coupling at interfaces”. In: *Physical Review Letters* 126.7 (2021), p. 076001.

[338] G. Clavier et al. “Computation of elastic constants of solids using molecular simulation: comparison of constant volume and constant pressure ensemble methods”. In: *Molecular Simulation* 43 (17 2017), pp. 1413–1422. ISSN: 10290435. DOI: 10.1080/08927022.2017.1313418.

[339] Katherine Sebeck, Chen Shao, and John Kieffer. “Alkane-Metal Interfacial Structure and Elastic Properties by Molecular Dynamics Simulation”. In: *ACS Applied Materials and Interfaces* 8 (26 2016), pp. 16885–16896. ISSN: 19448252. DOI: 10.1021/acsami.6b01665.

[340] Jie Zhang, Andrew Tan, and Hugh Spikes. “Effect of Base Oil Structure on Elastohydrodynamic Friction”. In: *Tribology Letters* 65 (1 2017), p. 13. ISSN: 1023-8883. DOI: 10.1007/s11249-016-0791-7.

[341] Lu Fang et al. “What stress components drive mechanochemistry? A study of ZDDP tribofilm formation”. In: *Faraday Discussions* 241 (2023), pp. 394–412.

[342] Xingyu Chen et al. “ReaxFF reactive molecular dynamics simulations of mechano-chemical decomposition of perfluoropolyether lubricants in heat-assisted magnetic recording”. In: *The Journal of Physical Chemistry C* 124.41 (2020), pp. 22496–22505.

[343] Emil Roduner. “Understanding catalysis”. In: *Chemical Society Reviews* 43 (24 2014), pp. 8226–8239. ISSN: 14604744. DOI: 10.1039/c4cs00210e.

[344] Bo Chen, Roald Hoffmann, and Roberto Cammi. “The Effect of Pressure on Organic Reactions in Fluids - a New Theoretical Perspective”. In: *Angewandte Chemie International Edition* 56 (2017), pp. 11126–11142. ISSN: 0044-8249. DOI: 10.1002/angie.201705427.

[345] J P Gao et al. “Frictional forces and Amontons’ law: From the molecular to the macroscopic scale”. In: *The Journal of Physical Chemistry B* 108 (11 2004), pp. 3410–3425. ISSN: 1520-6106. DOI: 10.1021/jp0363621.

[346] Xiaohu Yu et al. “Fe3O4 surface electronic structures and stability from GGA+U”. In: *Surface Science* 606 (9-10 2012), pp. 872–879. ISSN: 00396028. DOI: 10.1016/j.susc.2012.02.003.

[347] S J Eder, A Vernes, and G Betz. “On the Derjaguin Offset in Boundary-Lubricated Nanotribological Systems”. In: *Langmuir* 29 (45 2013), pp. 13760–13772. ISSN: 0743-7463. DOI: 10.1021/la4026443.

[348] Saeedeh S. Tafreshi et al. “Adsorption of hydrazine on the perfect and defective copper (111) surface: A dispersion-corrected DFT study”. In: *Surface Science* 622 (2014), pp. 1–8. ISSN: 00396028. DOI: 10.1016/j.susc.2013.11.013.

[349] Eric Osei-Agyemang, Stephen Berkebile, and Ashlie Martini. “Decomposition mechanisms of anti-wear lubricant additive tricresyl phosphate on iron surfaces using DFT and atomistic thermodynamic studies”. In: *Tribology Letters* 66 (2018), pp. 1–22.

[350] N Y Dzade, A Roldan, and N H de Leeuw. “A Density Functional Theory Study of the Adsorption of Benzene on Hematite (α-Fe2O3) Surfaces”. In: *Minerals* 4 (1 2014), pp. 89–115. ISSN: 2075-163X. DOI: 10.3390/min4010089.

[351] Roman Heuberger, Antonella Rossi, and Nicholas D Spencer. “XPS study of the influence of temperature on ZnDTP tribofilm composition”. In: *Tribology Letters* 25 (3 2007), pp. 185–196. DOI: 10.1007/s11249-006-9166-9.

[352] C A Olsson and D Landolt. “Passive films on stainless steels - chemistry, structure and growth”. In: *Electrochimica Acta* 48 (2003), pp. 1093–1104. DOI: 10.1016/S0013-4686(02)00841-1.

[353] M. Abdelmaksoud, J. W. Bender, and J. Krim. “Nanotribology of a vapor-phase lubricant: A quartz crystal microbalance study of tricresylphosphate (TCP) uptake on iron and chromium”. In: *Tribology Letters* 13 (3 2002), pp. 179–186. DOI: 10.1023/A:1020155925183.

[354] Keiji Sasaki, Naruhiko Inayoshi, and Kohji Tashiro. “Friction-induced dynamic chemical changes of tricresyl phosphate as lubricant additive observed under boundary lubrication with 2D fast imaging FTIR-ATR spectrometer”. In: *Wear* 268 (7-8 2010), pp. 911–916. ISSN: 0043-1648. DOI: 10.1016/j.wear.2009.12.017.

[355] Patrick J Barrie. “The mathematical origins of the kinetic compensation effect: 1. The effect of random experimental errors”. In: *Physical Chemistry Chemical Physics* 14.1 (2012), pp. 318–326.

[356] H. Lynggaard et al. “Analysis of simple kinetic models in heterogeneous catalysis”. In: *Progress in Surface Science* 77 (3-4 2004), pp. 71–137. ISSN: 00796816. DOI: 10.1016/j.progsurf.2004.09.001.

[357] Monica Crippa et al. “GHG emissions of all world countries”. In: *Publications Office of the European Union, Luxembourg* 10 (2023), p. 953322.

[358] Piers M de F Forster, Keith P Shine, and Nicola Stuber. “It is premature to include non-CO2 effects of aviation in emission trading schemes”. In: *Atmospheric environment* 40.6 (2006), pp. 1117–1121.

[359] Russell W Stratton, Philip J Wolfe, and James I Hileman. “Impact of aviation non-CO2 combustion effects on the environmental feasibility of alternative jet fuels”. In: *Environmental science & technology* 45.24 (2011), pp. 10736–10743.

[360] A Murphy et al. “Roadmap to decarbonising European aviation”. In: *Transport & Environment: Brussels, Belgium* (2018).

[361] Patrick Schmidt et al. *Power-to-liquids: Potentials and perspectives for the future supply of renewable aviation fuel*. German Environment Agency, 2016.

[362] Nathan Gray et al. “Decarbonising ships, planes and trucks: An analysis of suitable low-carbon fuels for the maritime, aviation and haulage sectors”. In: *Advances in Applied Energy* 1 (2021), p. 100008.

[363] Ausilio Bauen et al. “Sustainable Aviation Fuels: Status, challenges and prospects of drop-in liquid fuels, hydrogen and electrification in aviation”. In: *Johnson Matthey Technology Review* 64.3 (2020), pp. 263–278.

[364] L Zhu, N Li, and PRN Childs. “Light-weighting in aerospace component and system design”. In: *Propulsion and Power Research* 7.2 (2018), pp. 103–119.

[365] Malcolm K Stanford, Fransua Thomas, and Christopher DellaCorte. *Processing issues for preliminary melts of the intermetallic compound 60-NiTiNOL*. Tech. rep. 2012.

[366] Christopher DellaCorte and Walter A Wozniak. “Design and manufacturing considerations for shockproof and corrosion-immune superelastic nickel-titanium bearings for a space station application”. In: *41st Aerospace Mechanisms Symposium*. GRC-E-DAA-TN5006. 2012.

[367] Christopher DellaCorte and Lewis E Moore. “The effect of indenter ball radius on the static load capacity of the superelastic 60NiTi for rolling element bearings”. In: *Society of Tribologists and Lubrication Engineers (STLE) Annual Meeting and Exhibition*. GRC-E-DAA-TN13462. 2014.

[368] Christopher DellaCorte and Michael Jefferson. “60NiTi intermetallic material evaluation for lightweight and corrosion resistant spherical sliding bearings for aerospace applications, report on NASA-Kamatics SAA3-1288”. In: *Tribology Frontiers Conference*. E-19059. 2015.

[369] Christopher DellaCorte, Fransua Thomas, and Olivia Ann Leak. “Tribological evaluation of candidate gear materials operating under light loads in highly humid conditions”. In: *Society of Tribologists and Lubrication Engineers Annual Meeting and Exhibition*. GRC-E-DAA-TN26629. 2015.

[370] Rabin Neupane and Zoheir Farhat. “Wear resistance and indentation behavior of equiatomic superelastic TiNi and 60NiTi”. In: *Materials Sciences and Applications* 6.7 (2015), pp. 694–706.

[371] Khashayar Khanlari et al. “Reciprocating sliding wear behavior of 60NiTi as compared to 440C steel under lubricated and unlubricated conditions”. In: *Tribology Transactions* 61.6 (2018), pp. 991–1002.

[372] XIA Long et al. “Effect of boron doping on waterproof and dielectric properties of polyborosiloxane coating on SiO2f/SiO2 composites”. In: *Chinese Journal of Aeronautics* 32.8 (2019), pp. 2017–2027.

[373] XIAO Jie et al. “Microstructure and high-temperature oxidation behavior of plasma-sprayed Si/Yb2SiO5 environmental barrier coatings”. In: *Chinese Journal of Aeronautics* 32.8 (2019), pp. 1994–1999.

[374] Charles Miller, Dipankar Choudhury, and Min Zou. “The effects of surface roughness on the durability of polydopamine/PTFE solid lubricant coatings on NiTiNOL 60”. In: *Tribology Transactions* 62.5 (2019), pp. 919–929.

[375] Samuel Beckford and Min Zou. “Wear resistant PTFE thin film enabled by a polydopamine adhesive layer”. In: *Applied Surface Science* 292 (2014), pp. 350–356.

[376] Dipankar Choudhury, Charles Miller, and Min Zou. “Tribological performance of PDA/PTFE+ graphite particle coatings on 60NiTi”. In: *Applied Surface Science* 527 (2020), p. 146731.

[377] Tai-Hua Zhang et al. “Micro/nanomechanical and tribological properties of thin diamond-like carbon coatings”. In: *Chinese Journal of Aeronautics* 16.1 (2003), pp. 47–51.

[378] Yeong Ju Jo et al. “Synthesis and electrochemical properties of Ti-doped DLC films by a hybrid PVD/PECVD process”. In: *Applied Surface Science* 433 (2018), pp. 1184–1191.

[379] Yefei Zhou et al. “Mechanical and tribological behaviors of Ti-DLC films deposited on 304 stainless steel: Exploration with Ti doping from micro to macro”. In: *Diamond and Related Materials* 107 (2020), p. 107870.

[380] Yefei Zhou et al. “Exploring long-run reciprocating Wear of diamond-like carbon coatings: Microstructural, morphological and tribological evolution”. In: *Surface and Coatings Technology* 405 (2021), p. 126581.

[381] Christophe Donnet and Ali Erdemir. *Tribology of diamond-like carbon films: fundamentals and applications*. Springer Science & Business Media, 2007.

[382] Neha Sharma et al. “Scratch resistance and tribological properties of DLC coatings under dry and lubrication conditions”. In: *Tribology International* 56 (2012), pp. 129–140.

[383] Ali Erdemir, OL Eryilmaz, and G Fenske. “Synthesis of diamondlike carbon films with superlow friction and wear properties”. In: *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films* 18.4 (2000), pp. 1987–1992.

[384] Frank Philip Bowden and AE Hanwell. “The friction of clean crystal surfaces”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 295.1442 (1966), pp. 233–243.

[385] C Su and J-C Lin. “Thermal desorption of hydrogen from the diamond C (100) surface”. In: *Surface science* 406.1-3 (1998), pp. 149–166.

[386] Ali Erdemir. “Genesis of superlow friction and wear in diamondlike carbon films”. In: *Tribology International* 37.11-12 (2004), pp. 1005–1012.

[387] Jigang Wang and Min Liu. “Study on the tribological properties of hard films deposited on biomedical NiTi alloy”. In: *Materials Chemistry and Physics* 129.1-2 (2011), pp. 40–45.

[388] Eugenia L Dalibón et al. “Soft and thick DLC deposited on AISI 316L stainless steel with nitriding as pretreatment tested in severe wear conditions”. In: *Diamond and Related Materials* 106 (2020), p. 107881.

[389] Auezhan Amanov et al. “Fretting wear and fracture behaviors of Cr-doped and non-doped DLC films deposited on Ti–6al–4V alloy by unbalanced magnetron sputtering”. In: *Tribology International* 62 (2013), pp. 49–57.

[390] J-H Wu et al. “Tribological behavior of WC/DLC/WS2 nanocomposite coatings”. In: *Surface and Coatings Technology* 188 (2004), pp. 605–611.

[391] J-H Wu et al. “Tribological characteristics of diamond-like carbon (DLC) based nanocomposite coatings”. In: *Wear* 259.1-6 (2005), pp. 744–751.

[392] Huiqing Lan, Takahisa Kato, and Can Liu. “Molecular dynamics simulations of atomic-scale tribology between amorphous DLC and Si-DLC films”. In: *Tribology international* 44.11 (2011), pp. 1329–1332.

[393] RL Johnson and Edmond E Bisson. *Bearings and lubricants for aircraft turbine engines*. Tech. rep. SAE Technical Paper, 1955.

[394] AR Lansdown and S Lee. “Aviation lubricants”. In: *Chemistry and technology of lubricants* (2010), pp. 345–373.

[395] Duncan Dowson et al. *Tribological research and design for engineering systems: proceedings of the 29th leeds-lyon symposium*. Elsevier, 2003.

[396] Peter Onu, Suday L Lawal, et al. “A mini review on Aerospace Engineering: Lubrication Challenges and Innovation Tribology”. In: *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*. IEEE. 2024, pp. 1–9.

[397] Paul E Baker et al. “Identifying safer anti-wear triaryl phosphate additives for jet engine lubricants”. In: *Chemico-biological interactions* 203.1 (2013), pp. 257–264.

[398] Chris Winder and Jean-Christophe Balouet. “The toxicity of commercial jet oils”. In: *Environmental research* 89.2 (2002), pp. 146–164.

[399] Kasper Solbu et al. “Organophosphates in aircraft cabin and cockpit air - Method development and measurements of contaminants”. In: *Journal of Environmental Monitoring* 13 (5 2011), pp. 1393–1403. ISSN: 14640325. DOI: 10.1039/c0em00763c.

[400] Derrick Crump, Paul Harrison, and Christopher Walton. “Aircraft cabin air sampling study; part 2 of the final report”. In: *Cranfield University Staff Publications, Cranfield Health* (2011).

[401] C Winder and JC Balouet. “Aerotoxic sydrome”. In: *Toxicology*. Vol. 164. 1-3. ELSEVIER SCI IRELAND LTD CUSTOMER RELATIONS MANAGER, BAY 15, SHANNON ... 2001, pp. 47–47.

[402] Mohammadamin Sadeghi et al. “Tribological behavior of doped DLC coatings in the presence of ionic liquid additive under different lubrication regimes”. In: *Coatings* 13.5 (2023), p. 891.

[403] Shahsharif Shaikh et al. “Impact of temperature variation on friction behaviour of rare earth-doped diamond-like carbon coatings with ionic liquid lubricants”. In: *Lubricants* 11.7 (2023), p. 302.

[404] T Omiya et al. “Tribological performance of Gd-DLC and Eu-DLC coatings in the presence of synthetic oils containing ionic liquid additives”. In: *Tribology Letters* 71.2 (2023), p. 65.

[405] Yan Zhou et al. “Does the use of diamond-like carbon coating and organophosphate lubricant additive together cause excessive tribochemical material removal?” In: *Advanced Materials Interfaces* 2.15 (2015).

[406] Yan Zhou et al. “Ionic liquids composed of phosphonium cations and organophosphate, carboxylate, and sulfonate anions as lubricant antiwear additives”. In: *Langmuir* 30.44 (2014), pp. 13301–13311.

[407] Y Yamamoto and F Hirano. “The effect of the addition of phosphate esters to paraffinic base oils on their lubricating performance under sliding conditions”. In: *Wear* 78.3 (1982), pp. 285–296.

[408] IL Goldblatt and JK Appeldoorn. “The antiwear behavior of TCP in different atmospheres and different base stocks”. In: *ASLE transactions* 13.3 (1970), pp. 203–214.

[409] Du Hee Han and Masabumi Masuko. “Elucidation of the antiwear performance of several organic phosphates used with different polyol ester base oils from the aspect of interaction between additive and base oil”. In: *Tribology transactions* 41.4 (1998), pp. 600–604.

[410] Qing Yang and Fangli Duan. “Tribological properties of phosphate ester confined between iron-based surfaces”. In: *Langmuir* 40.7 (2024), pp. 3738–3747.

[411] Yang Wang et al. “Tight-binding quantum chemical molecular dynamics study on the friction and wear processes of diamond-like carbon coatings: Effect of tensile stress”. In: *ACS applied materials & interfaces* 9.39 (2017), pp. 34396–34404.

[412] Carla de Tomas, Irene Suarez-Martinez, and Nigel A Marks. "Graphitization of amorphous carbons: A comparative study of interatomic potentials". In: *Carbon* 109 (2016), pp. 681–693.

[413] Tao Huang et al. "Molecular dynamics study on nano sliding behavior at DLC/AISI 304 interface". In: *Carbon* (2025), p. 120371.

[414] Jingxiang Xu et al. "Systematically study the tensile and compressive behaviors of diamond-like carbon". In: *Nanomaterials* 13.11 (2023), p. 1772.

[415] Shengxi Wang and Kyriakos Komvopoulos. "A molecular dynamics analysis of the effect of surface passivation on the adhesion, deformation behavior and structure stability of amorphous carbon ultrathin films". In: *Materials Letters* 289 (2021), p. 129435.

[416] Merck Life Science UK Limited. *Safety Data Sheet: Tritolyl phosphate*. https://www.sigmaaldrich.com/GB/en/sds/aldrich/268917. Product Number: 268917, Version 7.2, Revision Date: 2023-06-15, Print Date: 2024-08-31. 2023.

[417] Chris DellaCorte and Oscar Pinkus. *Tribological Limitations in Gas Turbine Engines: A Workshop to Identify the Challenges and Set Future Directions*. Tech. rep. NASA, 2000.

[418] Fakhrul H. Bhuiyan et al. "Shear-activation of mechanochemical reactions through molecular deformation". In: *Scientific Reports* 14 (1 2024), p. 2992. ISSN: 2045-2322. DOI: 10.1038/s41598-024-53254-2. URL: https://www.nature.com/articles/s41598-024-53254-2.

[419] Rachael Curtis, Jessica Martin, and David Cao. *2.10: Zero-Order Reactions*. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.10%3A_Zero-Order_Reactions. Accessed: 2025-05-18. 2023. URL: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.10%3A_Zero-Order_Reactions.

[420] Abdel Dorgham et al. "Single-asperity study of the reaction kinetics of P-based triboreactive films". In: *Tribology International* 133 (2019), pp. 288–296.

[421] Rachael Curtis, Cathy Nguyen, and Stephen Lower. *2.3: First-Order Reactions*. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.03%3A_First-Order_Reactions. Accessed: 2025-05-18. 2023. URL: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.03%3A_First-Order_Reactions.

[422] John Marzluff. *7.2: First-Order Reactions Show an Exponential Decay of Reactant Concentration with Time*. https://chem.libretexts.org/Workbench/Username%3A_marzluff@grinnell.edu/Unit_3%3A_Kinetics/7%3A_Chemical_Kinetics_Rate_Laws/7.2%3A_First-Order_Reactions_Show_an_Exponential_Decay_of_Reactant_Concentration_with_Time. Accessed: 2025-05-18. 2023.

[423] Lannah Lua, Ciara Murphy, and Victoria Blanchard. *2.8: Second-Order Reactions*. https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.08%3A_Second-Order_Reactions. Accessed: 2025-05-18. 2023. URL: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/02%3A_Reaction_Rates/2.08%3A_Second-Order_Reactions.

[424] Yuh-Shan Ho. "Review of second-order models for adsorption systems". In: *Journal of hazardous materials* 136.3 (2006), pp. 681–689.

[425] Sai Sriharsha M Konda et al. "Chemical reactions modulated by mechanical stress: extended Bell theory". In: *The Journal of chemical physics* 135.16 (2011).

[426] Rudolph A Marcus. "On the theory of oxidation-reduction reactions involving electron transfer. I". In: *The Journal of chemical physics* 24.5 (1956), pp. 966–978.

[427] George S Hammond. "A correlation of reaction rates". In: *Journal of the American Chemical Society* 77.2 (1955), pp. 334–338.

[428] Dongjie Liu et al. "Molecular dynamics simulation of the lubricant conformation changes and energy transfer of the confined thin lubricant film". In: *Chemical Engineering Science* 270 (2023), p. 118541.

[429] Evan Evans and Ken Ritchie. "Dynamic strength of molecular adhesion bonds". In: *Biophysical journal* 72.4 (1997), pp. 1541–1555.

[430] Sergey V Sukhomlinov and Martin H Müser. "Stress Anisotropy Severely Affects Zinc Phosphate Network Formation". In: *Tribology Letters* 69 (2021), p. 89. ISSN: 1573-2711. DOI: 10.1007/s11249-021-01462-6.

[431] S Campen et al. "On the Increase in Boundary Friction with Sliding Speed". In: *Tribology Letters* 48 (2 2012), pp. 237–248. ISSN: 1023-8883. DOI: 10.1007/s11249-012-0019-4.

[432] DM Heyes et al. "Time dependent nonlinear shear stress effects in simple liquids: A molecular dynamics study". In: *The Journal of Chemical Physics* 73.8 (1980), pp. 3987–3996.

[433] George Carson, Hsuan-Wei Hu, and Steve Granick. "Molecular tribology of fluid lubrication: shear thinning". In: *Tribology Transactions* 35.3 (1992), pp. 405–410.

[434] Thomas Stirner, David Scholz, and Jizhong Sun. "Ab initio simulation of structure and surface energy of low-index surfaces of stoichiometric α-Fe2O3". In: *Surface Science* 671 (2018), pp. 11–16.

[435] Nora H de Leeuw and Timothy G Cooper. "Surface simulation studies of the hydration of white rust Fe (OH) 2, goethite α-FeO (OH) and hematite α-Fe2O3". In: *Geochimica et Cosmochimica Acta* 71.7 (2007), pp. 1655–1673.

[436] RW Carpick, DY Sasaki, and AR Burns. "First observation of mechanochromism at the nanometer scale". In: *Langmuir* 16.3 (2000), pp. 1270–1278.

[437] Aydar Akchurin and Rob Bosman. "A deterministic stress-activated model for tribo-film growth and wear simulation". In: *Tribology letters* 65 (2017), pp. 1–9.

[438] Ashlie Martini and Seong H Kim. "Activation Volume in Shear-Driven Chemical Reactions". In: *Tribology Letters* 69 (2021), p. 150. ISSN: 1573-2711. DOI: 10.1007/s11249-021-01522-x. URL: https://doi.org/10.1007/s11249-021-01522-x.

[439] Chen Xiao et al. "Surface structure dependence of mechanochemical etching: Scanning probe-based nanolithography study on Si (100), Si (110), and Si (111)". In: *ACS applied materials & interfaces* 11.23 (2019), pp. 20583–20588.

[440] Chen Xiao et al. "Interplay between solution chemistry and mechanical activation in friction-induced material removal of silicon surface in aqueous solution". In: *Tribology International* 148 (2020), p. 106319.

[441] Jonathan R Felts et al. "Direct mechanochemical cleavage of functional groups from graphene". In: *Nature communications* 6.1 (2015), p. 6467.

[442] Nam-Seok Park et al. "Atomic layer wear of single-crystal calcite in aqueous solution using scanning force microscopy". In: *Journal of Applied Physics* 80.5 (1996), pp. 2680–2686.

[443] S Loehlé and MC Righi. "First principles study of organophosphorus additives in boundary lubrication conditions: Effects of hydrocarbon chain length". In: *Lubrication Science* 29.7 (2017), pp. 485–491.

[444] David Johns and Andra Hutton. *6.2.3.4: The Arrhenius Law - Arrhenius Plots*. 2024. URL: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_

Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/06%3A_Modeling_Reaction_Kinetics/6.02%3A_Temperature_Dependence_of_Reaction_Rates/6.2.03%3A_The_Arrhenius_Law/6.2.3.04%3A_The_Arrhenius_Law_-_Arrhenius_Plots.

[445] Resham Rana et al. "An Analysis of Shear-Dependent Mechanochemical Reaction Kinetics". In: *Tribology Letters* 72 (3 2024), p. 76. ISSN: 15732711. DOI: 10.1007/s11249-024-01879-9.

[446] Richard Gordon Budynas, J Keith Nisbett, et al. *Shigley's mechanical engineering design*. Vol. 9. McGraw-Hill New York, 2011.

[447] AZO Materials. *An Introduction to Iron*. https://www.azom.com/properties.aspx?ArticleID=619. 2025.

[448] AZO Materials. *What Are the Properties of Aluminum?* https://www.azom.com/properties.aspx?ArticleID=1446. 2025.

[449] Juli Felicio Luiz and Hugh Spikes. "Tribofilm formation, friction and wear-reducing properties of some phosphorus-containing antiwear additives". In: *Tribology Letters* 68 (2020), pp. 1–24.

[450] IN Lacey et al. "Thick antiwear films in elastohydrodynamic contacts. Part II: chemical nature of the deposited films". In: *ASLE transactions* 29.3 (1986), pp. 306–311.

[451] PA Bertrand. "Reactions of tricresyl phosphate with bearingmaterials". In: *Tribology Letters* 3.4 (1997), pp. 367–377.

[452] Allison G. Roessler and Paul M. Zimmerman. "Examining the Ways to Bend and Break Reaction Pathways Using Mechanochemistry". In: *Journal of Physical Chemistry C* 122 (12 2018), pp. 6996–7004. ISSN: 19327455. DOI: 10.1021/acs.jpcc.8b00467.

[453] Paul G Loncke and Paul J Berti. "Implications of Protonation and Substituent Effects for C- O and O- P Bond Cleavage in Phosphate Monoesters". In: *Journal of the American Chemical Society* 128.18 (2006), pp. 6132–6140.

[454] Frank H Allen and Anthony J Kirby. "Bond length and reactivity. Variable length of the carbon-oxygen single bond". In: *Journal of the American Chemical Society* 106.21 (1984), pp. 6197–6200.

[455] Werner EG Müller, Heinz C Schröder, and Xiaohong Wang. "Inorganic polyphosphates as storage for and generator of metabolic energy in the extracellular matrix". In: *Chemical reviews* 119.24 (2019), pp. 12337–12374.

[456] George A Lager and GV Gibbs. "Effect of variations in OPO and POP angles on PO bond overlap populations for some selected ortho-and pyrophosphates". In: *American Mineralogist: Journal of Earth and Planetary Materials* 58.7-8 (1973), pp. 756–764.

[457] David A Cardwell, David C Larbalestier, and Aleksander Braginski. *Handbook of Superconductivity: Fundamentals and Materials, Volume One*. Vol. 1. CRC Press, 2022.

[458] Tom Leadbetter Cottrell. "The strengths of chemical bonds". In: *(No Title)* (1954).

[459] B de B Darwent. *National standard reference data series*. 1970.

[460] Sidney W Benson. "III-Bond energies". In: *Journal of Chemical Education* 42.9 (1965), p. 502.

[461] Robert J. Lancashire. *Polarizability*. 2024. URL: https://chem.libretexts.org/Courses/University_of_Missouri/MU%3A__1330H_(Keller)/09._Molecular_Geometry_and_Bonding_Theories/9.4%3A_Covalent_Bonding_and_Orbital_Overlap.

[462] Tim Soderberg and LibreTexts Contributors. *9.4: Covalent Bonding and Orbital Overlap*. 2024. URL: https://chem.libretexts.org/Courses/University_of_Missouri/MU%3A__1330H_(Keller)/09._Molecular_Geometry_and_Bonding_Theories/9.4%3A_Covalent_Bonding_and_Orbital_Overlap.

[463] Tim Soderberg and LibreTexts Contributors. *9.1: Overview of Phosphate Groups*. Accessed: 2025-06-11. Licensed under CC BY-NC-SA 4.0. 2024. URL: https://chem.libretexts.org/Courses/Oregon_

Institute_of_Technology/OIT%3A_CHE_332_--_Organic_Chemistry_II_(Lund)/9%3A_Phosphate_Transfer_Reactions/9.1%3A_Overview_of_Phosphate_Groups.

[464] Eric Hendriks et al. "Industrial requirements for thermodynamics and transport properties". In: *Industrial and Engineering Chemistry Research* 49 (22 2010), pp. 11131–11141. ISSN: 08885885. DOI: 10.1021/ie101231b.

[465] Georgios M. Kontogeorgis et al. "Industrial requirements for thermodynamic and transport properties: 2020". In: *Industrial and Engineering Chemistry Research* 60 (13 2021), pp. 4987–5013. ISSN: 15205045. DOI: 10.1021/acs.iecr.0c05356.

[466] Elisa Conte, Rafiqul Gani, and Ka Ming Ng. "Design of formulated products: A systematic methodology". In: *AIChE Journal* 57 (9 2011), pp. 2431–2449. ISSN: 00011541. DOI: 10.1002/aic.12458.

[467] Masami Ishikawa et al. "Introduction of Fuel Economy Engine Oil Performance Target with New SAE Viscosity Grade". In: *SAE International Journal of Fuels and Lubricants* 9 (2 2016), pp. 374–382. ISSN: 19463960. DOI: 10.4271/2016-01-0896.

[468] Adam S. Gross. "Simple Physics-Guided Compositional Model for Kinematic Viscosity at 100 °C for Groups II and III Base Oils and Base Stocks". In: *Industrial and Engineering Chemistry Research* 61 (25 2022), pp. 9087–9098. ISSN: 15205045. DOI: 10.1021/acs.iecr.2c00576.

[469] Rajendar R. Mallepally et al. "Fluid properties at high pressures and temperatures: Experimental and modelling challenges". In: *Journal of Supercritical Fluids* 134 (2018), pp. 33–40. ISSN: 08968446. DOI: 10.1016/j.supflu.2017.12.003.

[470] Joachim Gross and Gabriele Sadowski. "Perturbed-chain SAFT: An equation of state based on a perturbation theory for chain molecules". In: *Industrial and Engineering Chemistry Research* 40 (4 2001), pp. 1244–1260. ISSN: 08885885. DOI: 10.1021/ie0003887.

[471] Haseen O. Baled et al. "Viscosity models for pure hydrocarbons at extreme conditions: A review and comparative study". In: *Fuel* 218 (2018), pp. 89–111. ISSN: 00162361. DOI: 10.1016/j.fuel.2018.01.002.

[472] Oliver Lötgering-Lin and Joachim Gross. "Group Contribution Method for Viscosities Based on Entropy Scaling Using the Perturbed-Chain Polar Statistical Associating Fluid Theory". In: *Industrial and Engineering Chemistry Research* 54 (32 2015), pp. 7942–7952. ISSN: 15205045. DOI: 10.1021/acs.iecr.5b01698.

[473] Ian H Bell. "Entropy Scaling of Viscosity-II: Predictive Scheme for Normal Alkanes". In: *J. Chem. Eng. Data* 65 (2020), pp. 5606–5616. DOI: 10.1021/acs.jced.0c00749.

[474] Ovidiu Ivanciuc et al. "Estimation of the liquid viscosity of organic compounds with a quantitative structure-property model". In: *Journal of Chemical Information and Computer Sciences* 39 (3 1999), pp. 515–524. ISSN: 00952338. DOI: 10.1021/ci980117v.

[475] Alan R. Katritzky et al. "Prediction of liquid viscosity for organic compounds by a quantitative structure-property relationship". In: *Journal of Physical Organic Chemistry* 13 (1 2000), pp. 80–86. ISSN: 08943230. DOI: 10.1002/(SICI)1099-1395(200001)13:1<80::AID-POC179>3.0.CO;2-8.

[476] Gregory W. Kauffman and Peter C. Jurs. "Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure-property relationships". In: *Journal of Chemical Information and Computer Sciences* 41 (2001), pp. 408–418. ISSN: 00952338. DOI: 10.1021/ci000139t.

[477] Remya Rajappan et al. "Quantitative structure-property relationship (QSPR) prediction of liquid viscosities of pure organic compounds employing random forest regression". In: *Industrial and Engineering Chemistry Research* 48 (21 2009), pp. 9708–9712. ISSN: 08885885. DOI: 10.1021/ie8018406.

[478] Guangqing Cai et al. “Quantitative Structure-Property Relationship Model for Hydrocarbon Liquid Viscosity Prediction”. In: *Energy and Fuels* 32 (3 2018), pp. 3290–3298. ISSN: 15205029. DOI: 10.1021/acs.energyfuels.7b04075.

[479] Chia Ming Chang et al. “The q-RASPR approach for predicting the property and fate of persistent organic pollutants”. In: *Scientific Reports* 15.1 (2025), p. 1344.

[480] Edward J Maginn. “From Discovery to Data: What Must Happen for Molecular Simulation to Become a Mainstream Chemical Engineering Tool”. In: *AIChE Journal* 55 (6 2009), pp. 1304–1310. DOI: 10.1002/aic.11932.

[481] E J Maginn and J R Elliott. “Historical Perspective and Current Outlook for Molecular Dynamics As a Chemical Engineering Tool”. In: *Industrial & Engineering Chemistry Research* 49 (7 2010), pp. 3059–3078. ISSN: 0888-5885. DOI: 10.1021/ie901898k.

[482] Marcus J. Tillotson et al. “Molecular modelling of the thermophysical properties of fluids: expectations, limitations, gaps and opportunities”. In: *Physical Chemistry Chemical Physics* 25 (2023), pp. 12607–12628. ISSN: 1463-9076. DOI: 10.1039/D2CP05423J.

[483] G Kaminski and W L Jorgensen. “Performance of the AMBER94, MMFF94, and OPLS-AA force fields for modeling organic liquids”. In: *Journal of Physical Chemistry* 100 (46 1996), pp. 18010–18013. ISSN: 0022-3654. DOI: 10.1021/jp9624257.

[484] C Caleman et al. “Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant”. In: *Journal of Chemical Theory and Computation* 8 (1 2012), pp. 61–74. ISSN: 1549-9618. DOI: 10.1021/ct200731v.

[485] J D Moore et al. “Rheology of lubricant basestocks: A molecular dynamics study of C30 isomers”. In: *J. Chem. Phys.* 113 (19 2000), pp. 8833–8840. ISSN: 0021-9606. DOI: 10.1063/1.1318768.

[486] Konstantinos D. Papavasileiou et al. “Molecular Dynamics Simulation of Pure n-Alkanes and their Mixtures at Elevated Temperatures Using Atomistic and Coarse-Grained Force Fields”. In: *J. Phys. Chem. B* 123 (2019), pp. 6229–6243. DOI: 10.1021/acs.jpcb.9b02840.

[487] Abhinav Verma et al. “Determination of Diesel Physical Properties at Injection Pressures and Temperatures via All-Atom Molecular Simulations”. In: *SAE International Journal of Fuels and Lubricants* 9 (3 2016), pp. 567–574. DOI: 10.4271/2016-01-2253.

[488] Dianne J Luning Prak et al. “Thermophysical Properties of Binary Mixtures of n-Dodecane with n-Alkylcyclohexanes: Experimental Measurements and Molecular Dynamics Simulations”. In: *J. Chem. Eng. Data* 64 (4 2019), pp. 1550–1568. DOI: 10.1021/acs.jced.8b01135.

[489] Nikolay D. Kondratyuk, Vasily V. Pisarev, and James P. Ewen. “Probing the High-Pressure Viscosity of Hydrocarbon Mixtures using Molecular Dynamics Simulations”. In: *Journal of Chemical Physics* 153 (2020), p. 154502. DOI: 10.1063/5.0028393.

[490] Frances D. Lenahan et al. “Viscosity, Interfacial Tension, and Density of Binary-Liquid Mixtures of n-Hexadecane with n-Octacosane, 2,2,4,4,6,8,8-Heptamethylnonane, or 1-Hexadecanol at Temperatures between 298.15 and 573.15 K by Surface Light Scattering and Equilibrium Molecular Dy”. In: *Journal of Chemical and Engineering Data* 66 (5 2021), pp. 2264–2280. ISSN: 15205134. DOI: 10.1021/acs.jced.1c00108.

[491] Xueming Yang et al. “A molecular dynamics simulation study of the densities and viscosities of 1,2,4-trimethylbenzene and its binary mixture with n-decane”. In: *Fluid Phase Equilibria* 562 (2022), p. 113566. ISSN: 03783812. DOI: 10.1016/j.fluid.2022.113566.

[492] M Aquing et al. “Composition Analysis and Viscosity Prediction of Complex Fuel Mixtures Using a Molecular-Based Approach”. In: *Energy and Fuels* 26 (4 2012), pp. 2220–2230. ISSN: 0887-0624. DOI: 10.1021/ef300106z.

[493] Motoaki Iwase et al. “Development of Digital Oil for Heavy Crude Oil: Molecular Model and Molecular Dynamics Simulations”. In: *Energy and Fuels* 32 (2018), pp. 2781–2792. DOI: 10.1021/acs.energyfuels.7b02881.

[494] Mingjun Shen et al. “Molecular dynamics study on the viscosity of hydraulic oil in the deep-sea environment”. In: *Journal of Molecular Liquids* 411 (2024), p. 125716. ISSN: 01677322. DOI: 10.1016/j.molliq.2024.125716.

[495] Stewart A Adcock and J Andrew McCammon. “Molecular dynamics: survey of methods for simulating the activity of proteins”. In: *Chemical reviews* 106.5 (2006), pp. 1589–1615.

[496] Venkatesh Botu et al. “Machine learning force fields: construction, validation, and outlook”. In: *The Journal of Physical Chemistry C* 121.1 (2017), pp. 511–522.

[497] Fabian Jirasek and Hans Hasse. “Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures”. In: *Annual Review of Chemical and Biomolecular Engineering* 14 (2023), pp. 31–51. DOI: 10.1146/annurev-chembioeng.

[498] John Homer, Sotos C Generalis, and John H Robson. “Artificial neural networks for the prediction of liquid viscosity, density, heat of vaporization, boiling point and Pitzer’s acentric factor. Part I. Hydrocarbons”. In: *Physical Chemistry Chemical Physics* 1 (1999), pp. 4075–4081.

[499] Takahiro Suzuki, Ralf Uwe Ebert, and Gerrit Schüürmann. “Application of Neural Networks to Modeling and Estimating Temperature-Dependent Liquid Viscosity of Organic Compounds”. In: *Journal of Chemical Information and Computer Sciences* 41 (3 2001), pp. 776–790. ISSN: 00952338. DOI: 10.1021/ci000154y.

[500] Diego Alonso Saldana et al. “Prediction of density and viscosity of biofuel compounds using machine learning methods”. In: *Energy and Fuels* 26 (4 2012), pp. 2416–2426. ISSN: 08870624. DOI: 10.1021/ef3001339.

[501] Pavao Santak and Gareth Conduit. “Predicting physical properties of alkanes with neural networks”. In: *Fluid Phase Equilibria* 501 (2019), p. 112259. ISSN: 03783812. DOI: 10.1016/j.fluid.2019.112259.

[502] Valentin Goussard et al. “A New Machine-Learning Tool for Fast Estimation of Liquid Viscosity. Application to Cosmetic Oils”. In: *Journal of Chemical Information and Modeling* 60 (4 2020), pp. 2012–2023. ISSN: 15205142. DOI: 10.1021/acs.jcim.0c00083.

[503] GC Loh et al. “Viscosity prediction of lubricants by a general feed-forward neural network”. In: *Journal of Chemical Information and Modeling* 60.3 (2020), pp. 1224–1234.

[504] Jae Young Kim, Salman A. Khan, and Dionisios G. Vlachos. “Similarity-Based Machine Learning for Small Data Sets: Predicting Biolubricant Base Oil Viscosities”. In: *Journal of Physical Chemistry B* (2024), pp. 11963–11970. ISSN: 15205207. DOI: 10.1021/acs.jpcb.4c06687.

[505] Maarten R. Dobbelaere et al. “Machine Learning for Physicochemical Property Prediction of Complex Hydrocarbon Mixtures”. In: *Industrial and Engineering Chemistry Research* 61 (24 2022), pp. 8581–8594. ISSN: 15205045. DOI: 10.1021/acs.iecr.2c00442.

[506] Mood Mohan et al. “High-Throughput Screening and Accurate Prediction of Ionic Liquid Viscosities Using Interpretable Machine Learning”. In: *ACS Sustainable Chemistry and Engineering* 12 (18 2024), pp. 7040–7054. ISSN: 21680485. DOI: 10.1021/acssuschemeng.4c00631.

[507] Mood Mohan et al. “Accurate Machine Learning for Predicting the Viscosities of Deep Eutectic Solvents”. In: *Journal of Chemical Theory and Computation* 20 (9 2024), pp. 3911–3926. ISSN: 15499626. DOI: 10.1021/acs.jctc.3c01163.

[508] Yu Zhang and Qiang Yang. “A survey on multi-task learning”. In: *IEEE transactions on knowledge and data engineering* 34.12 (2021), pp. 5586–5609.

[509] Zheng Gong et al. “Predicting Thermodynamic Properties of Alkanes by High-Throughput Force Field Simulation and Machine Learning”. In: *Journal of Chemical Information and Modeling* 58 (12 2018), pp. 2502–2516. ISSN: 1549960X. DOI: 10.1021/acs.jcim.8b00407.

[510] Nikhil V S Avula et al. “Building robust machine learning models for small chemical science data: the case of shear viscosity of fluids”. In: *Machine Learning: Science and Technology* 3 (4 2022), p. 045032. ISSN: 2632-2153. DOI: 10.1088/2632-2153/acac01.

[511] Juan J. De Pablo and Fernando A. Escobedo. “Molecular simulations in chemical engineering: Present and future”. In: *AIChE Journal* 48 (12 2002), pp. 2716–2721. ISSN: 00011541. DOI: 10.1002/aic.690481202.

[512] David M. Himmelblau. “Accounts of experiences in the application of artificial neural networks in chemical engineering”. In: *Industrial and Engineering Chemistry Research* 47 (16 2008), pp. 5782–5796. ISSN: 08885885. DOI: 10.1021/ie800076s.

[513] Jiashuo Liu et al. “Towards out-of-distribution generalization: A survey”. In: *arXiv preprint arXiv:2108.13624* (2021).

[514] Shiori Sagawa et al. “An investigation of why overparameterization exacerbates spurious correlations”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8346–8356.

[515] Sadman Sadeed Omee et al. “Structure-based out-of-distribution (OOD) materials property prediction: a benchmark study”. In: *npj Computational Materials* 10.1 (2024), p. 144.

[516] Alexander Dunn et al. “Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm”. In: *npj Computational Materials* 6.1 (2020), p. 138.

[517] Jingkang Yang et al. “Openood: Benchmarking generalized out-of-distribution detection”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 32598–32611.

[518] Peng Cui and Jinjia Wang. “Out-of-distribution (OOD) detection based on deep learning: A review”. In: *Electronics* 11.21 (2022), p. 3500.

[519] Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks”. In: *arXiv preprint arXiv:1610.02136* (2016).

[520] Kevin Yang et al. “Analyzing learned molecular representations for property prediction”. In: *Journal of chemical information and modeling* 59.8 (2019), pp. 3370–3388.

[521] Raghunathan Ramakrishnan et al. “Quantum chemistry structures and properties of 134 kilo molecules”. In: *Scientific data* 1.1 (2014), pp. 1–7.

[522] Ann M Richard et al. “The Tox21 10K compound library: collaborative chemistry advancing toxicology”. In: *Chemical Research in Toxicology* 34.2 (2020), pp. 189–216.

[523] Kunal Roy, Supratik Kar, and Rudra Narayan Das. *A primer on QSAR/QSPR modeling: fundamental concepts*. Springer, 2015.

[524] Yi Ding et al. “Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties”. In: *Journal of Molecular Liquids* 326 (2021), p. 115212.

[525] AS Barnard and G Opletal. “Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning”. In: *Nanoscale* 11.48 (2019), pp. 23165–23172.

[526] Dianne J. Luning Prak et al. “Biobased Diesel Fuel Analysis and Formulation and Testing of Surrogate Fuel Mixtures”. In: *Industrial and Engineering Chemistry Research* 57 (2 2018), pp. 600–610. ISSN: 15205045. DOI: 10.1021/acs.iecr.7b04419.

[527] Louis Edwards Caceres-Martinez et al. “Influence of blending cycloalkanes on the energy content, density, and viscosity of Jet-A”. In: *Fuel* 358 (2024), p. 129986. ISSN: 00162361. DOI: 10.1016/j.fuel.2023.129986.

[528] Mark Gough and Steven Rowland. “Characterization of Unresolved Complex Mixtures of Hydrocarbons from Lubricating Oil Feedstocks”. In: *Energy and Fuels* 5 (1991), pp. 869–874.

[529] Pinzhi Liu et al. “Lubricant shear thinning behavior correlated with variation of radius of gyration via molecular dynamics simulations”. In: *The Journal of Chemical Physics* 147.8 (2017).

[530] Vincent Burk et al. “Complementary experimental data and extended density and viscosity reference models for squalane”. In: *Journal of Chemical & Engineering Data* 66.5 (2021), pp. 1992–2005.

[531] Se-Kwon Kim and Fatih Karadeniz. “Biological importance and applications of squalene and squalane”. In: *Advances in food and nutrition research* 65 (2012), pp. 223–233.

[532] JY Leong et al. “Hydrodynamic friction reduction in a MAC–hexadecane lubricated MEMS contact”. In: *Tribology Letters* 49.1 (2013), pp. 217–225.

[533] Kaikai Ren et al. “Lubrication behavior of n-hexadecane on ZnO layer at the nanoscale: a molecular dynamic exploration”. In: *Tribology Letters* 70.3 (2022), p. 90.

[534] Sophie Campen et al. “Thermal degradation of n-hexadecane base oil and its impact on boundary friction and surface adsorption”. In: *Tribology International* 170 (2022), p. 107492.

[535] David S Ogunniyi. “Castor oil: a vital industrial raw material”. In: *Bioresource technology* 97.9 (2006), pp. 1086–1091.

[536] Yong Xu et al. “Formation of an adsorption film of MoS2 nanoparticles and dioctyl sebacate on a steel surface for alleviating friction and wear”. In: *Tribology International* 92 (2015), pp. 172–183.

[537] Margaret M Wu. “Alkylated aromatics”. In: *CHEMICAL INDUSTRIES-NEW YORK-MARCEL DEKKER*- (1999), pp. 195–214.

[538] Patrick Briot et al. “Synthetic oil with high viscosity index and having a low pour point”. EP1152050B1. 20031.

[539] David R Worton et al. “Lubricating oil dominates primary organic aerosol emissions from motor vehicles”. In: *Environmental Science & Technology* 48.7 (2014), pp. 3698–3706.

[540] Michel J Hourani et al. “Alkylated naphthalenes as high-performance synthetic lubricating fluids”. In: *Tribology transactions* 50.1 (2007), pp. 82–87.

[541] Berk Hess. “Determining the shear viscosity of model liquids from molecular dynamics simulations”. In: *The Journal of Chemical Physics* 116.1 (2002), pp. 209–217.

[542] Alper Kinaci, Justin B Haskins, and Tahir Çağın. “On calculation of thermal conductivity from Einstein relation in equilibrium molecular dynamics”. In: *The Journal of chemical physics* 137.1 (2012).

[543] Yong Zhang, Akihito Otani, and Edward J Maginn. “Reliable viscosity calculation from equilibrium molecular dynamics simulations: A time decomposition method”. In: *Journal of Chemical Theory and Computation* 11.8 (2015), pp. 3537–3546.

[544] Rajdeep Singh Payal et al. “Shear viscosity of linear alkanes through molecular simulations: Quantitative tests for n-decane and n-hexadecane”. In: *Molecular Simulation* 38.14-15 (2012), pp. 1234–1241.

[545] Kang-Sahn Kim et al. “Nature of intrinsic uncertainties in equilibrium molecular dynamics estimation of shear viscosity for simple and complex fluids”. In: *The Journal of Chemical Physics* 149.4 (2018), p. 044510.

[546] Othonas A Moultos et al. “System-size corrections for self-diffusion coefficients calculated from molecular dynamics simulations: The case of CO2, n-alkanes, and poly (ethylene glycol) dimethyl ethers”. In: *The Journal of Chemical Physics* 145.7 (2016).

[547] Peter J Daivis and Denis J Evans. “Transport coefficients of liquid butane near the boiling point by equilibrium molecular dynamics”. In: *The Journal of Chemical Physics* 103.10 (1995), pp. 4261–4265.

[548] Leandro Martínez et al. “PACKMOL: A package for building initial configurations for molecular dynamics simulations”. In: *Journal of Computational Chemistry* 30.13 (2009), pp. 2157–2164.

[549] Andrew I Jewett et al. “Moltemplate: A tool for coarse-grained modeling of complex biological matter and soft condensed matter physics”. In: *Journal of Molecular Biology* 433.11 (2021), p. 166841.

[550] Sereina Riniker and Gregory A Landrum. “Better informed distance geometry: using what we know to improve conformation generation”. In: *Journal of Chemical Information and Modeling* 55.12 (2015), pp. 2562–2574.

[551] John L Nazareth. “Conjugate gradient method”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 1.3 (2009), pp. 348–353.

[552] Vania Calandrini et al. “nMoldyn-Interfacing spectroscopic experiments, molecular dynamics simulations and models for time correlation functions”. In: *École thématique de la Société Française de la Neutronique* 12 (2011), pp. 201–232.

[553] Michael T. Humbert, Yong Zhang, and Edward J. Maginn. “PyLAT: Python LAMMPS Analysis Tools”. In: *Journal of Chemical Information and Modeling* 59 (4 2019), pp. 1301–1305. ISSN: 15205142. DOI: 10.1021/acs.jcim.9b00066.

[554] Otfried Madelung and Rainer Poerschke. *Der Landolt-Börnstein: Erfolgsgeschichte einer wissenschaftlichen Datensammlung im Springer-Verlag*. Springer-Verlag, 2010.

[555] Peter J Linstrom and William G Mallard. “The NIST Chemistry WebBook: A chemical data resource on the internet”. In: *Journal of Chemical & Engineering Data* 46.5 (2001), pp. 1059–1063.

[556] Sumnesh Gupta et al. “Industrial Expectations of a Pure Component Database for Thermodynamic and Transport Properties”. In: *Industrial and Engineering Chemistry Research* 61 (42 Oct. 2022), pp. 15514–15553. ISSN: 15205045. DOI: 10.1021/acs.iecr.2c01642.

[557] E. N. da C. Andrade. “The viscosity of liquids”. In: *Nature* 125 (3148 1930), pp. 309–310.

[558] Thomas E Daubert. “Evaluated Equation Forms for Correlating Thermodynamic and Transport Properties with Temperature”. In: *Industrial and Engineering Chemistry Research* 37 (1998), pp. 3260–3267. URL: https://pubs.acs.org/sharingguidelines.

[559] J. H. Dymond and R. Malhotra. “The Tait equation: 100 years on”. In: *International Journal of Thermophysics* 9 (6 1988), pp. 941–951. ISSN: 0195928X. DOI: 10.1007/BF01133262.

[560] Physical Property Data Service. *LOADER2 for Windows: User Manual and Reference Guide*. Physical Property Data Service, NEL. NEL. East Kilbride, Glasgow, UK, 1996.

[561] RC Wilhoit and KN Marsh. “Future directions for data compilations”. In: *International journal of thermophysics* 20 (1999), pp. 247–255.

[562] Eric W Lemmon, Marcia L Huber, Mark O McLinden, et al. “NIST standard reference database 23”. In: *Reference fluid thermodynamic and transport properties (REFPROP), version* 9 (2010).

[563] Marcia L. Huber et al. “The NIST REFPROP Database for Highly Accurate Properties of Industrially Important Fluids”. In: *Industrial and Engineering Chemistry Research* 61 (42 Oct. 2022), pp. 15449–15472. ISSN: 15205045. DOI: 10.1021/acs.iecr.2c01427.

[564] “Viscosity of Pure Organic Liquids and Binary Liquid Mixtures”. In: *Landolt-Börnstein*. Ed. by M.D. Lechner. Vol. 25. Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-56050-0. DOI: 10.1007/978-3-540-75486-2.

[565] Osama Marzouk. “Thermo physical chemical properties of fluids using the free NIST chemistry web book database”. In: *Fluid Mechanics Research International Journal* 1.1 (2017), pp. 14–18.

[566] Joseph L Durant et al. “Reoptimization of MDL keys for use in drug discovery”. In: *Journal of Chemical Information and Computer Sciences* 42.6 (2002), pp. 1273–1280.

[567] Sunghwan Kim et al. “PubChem substance and compound databases”. In: *Nucleic acids research* 44.D1 (2016), pp. D1202–D1213.

[568] David Rogers and Mathew Hahn. “Extended-connectivity fingerprints”. In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754.

[569] Raquel Rodríguez-Pérez and Jürgen Bajorath. "Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values". In: *Journal of medicinal chemistry* 63.16 (2019), pp. 8761–8777.

[570] Franklin Langlang Lee et al. "Comparison of machine learning methods towards developing interpretable polyamide property prediction". In: *Polymers* 13.21 (2021), p. 3653.

[571] Greg Landrum. *RDKit: Open-source cheminformatics*. 2024. DOI: 10.5281/zenodo.11396708. URL: https://www.rdkit.org.

[572] Mario Lovrić et al. "Should we embed in chemistry? A comparison of unsupervised transfer learning with PCA, UMAP, and VAE on molecular fingerprints". In: *Pharmaceuticals* 14.8 (2021), p. 758.

[573] Noel M O'Boyle. "Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI". In: *Journal of Cheminformatics* 4 (2012), pp. 1–14.

[574] Mario Krenn et al. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". In: *Machine Learning: Science and Technology* 1.4 (2020), p. 045024.

[575] A. Vasetenkov. *IUPAC to SMILES*. https://antvaset.com/iupac-to-smiles. Retrieved April 3, 2025.

[576] Hirotomo Moriwaki et al. "Mordred: a molecular descriptor calculator". In: *Journal of cheminformatics* 10.1 (2018), p. 4.

[577] Dakhaz Mustafa Abdullah and Adnan Mohsin Abdulazeez. "Machine learning applications based on SVM classification a review". In: *Qubahan Academic Journal* 1.2 (2021), pp. 81–90.

[578] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

[579] Mark R Segal. "Machine learning benchmarks and random forest regression". In: (2004).

[580] Zhen Li et al. "Deep learning methods for molecular representation and property prediction". In: *Drug Discovery Today* 27.12 (2022), p. 103373.

[581] Talia B Kimber, Maxime Gagnebin, and Andrea Volkamer. "Maxsmi: maximizing molecular property prediction performance with confidence estimation using smiles augmentation and deep learning". In: *Artificial Intelligence in the Life Sciences* 1 (2021), p. 100014.

[582] Tuan Le, Frank Noé, and Djork-Arné Clevert. "Equivariant graph attention networks for molecular property prediction". In: *arXiv preprint arXiv:2202.09891* (2022).

[583] Justin Gilmer et al. "Neural message passing for quantum chemistry". In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272.

[584] Neil Houlsby et al. "Parameter-efficient transfer learning for NLP". In: *International conference on machine learning*. PMLR. 2019, pp. 2790–2799.

[585] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[586] Lakshmi Mandal and Nanda Dulal Jana. "A comparative study of naive bayes and k-NN algorithm for multi-class drug molecule classification". In: *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE. 2019, pp. 1–4.

[587] Mohammad A Khanfar and Mutasem O Taha. "Elaborate ligand-based modeling coupled with multiple linear regression and k nearest neighbor QSAR analyses unveiled new nanomolar mTOR inhibitors". In: *Journal of chemical information and modeling* 53.10 (2013), pp. 2587–2612.

[588] Andrzej Maćkiewicz and Waldemar Ratajczak. "Principal components analysis (PCA)". In: *Computers & Geosciences* 19.3 (1993), pp. 303–342.

[589] Jun Xia et al. "Understanding the limitations of deep models for molecular property prediction: Insights and solutions". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 64774–64792.

[590] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35 (2022), pp. 507–520.

[591] Mario Köppen. "The curse of dimensionality". In: *5th online world conference on soft computing in industrial applications (WSC5)*. Vol. 1. 2000, pp. 4–8.

[592] Zhichun Guo et al. "Few-shot graph learning for molecular property prediction". In: *Proceedings of the web conference 2021*. 2021, pp. 2559–2567.

[593] Petar Veličković et al. "Graph attention networks". In: *arXiv preprint arXiv:1710.10903* (2017).

[594] Josh Achiam et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[595] Samuel Sledzieski et al. "Democratizing protein language models with parameter-efficient fine-tuning". In: *Proceedings of the National Academy of Sciences* 121.26 (2024), e2405840121.

[596] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[597] Jie Shen and Christos A Nicolaou. "Molecular property prediction: recent trends in the era of artificial intelligence". In: *Drug Discovery Today: Technologies* 32 (2019), pp. 29–36.

[598] Shengchao Liu et al. "Structured multi-task learning for molecular property prediction". In: *International conference on artificial intelligence and statistics*. PMLR. 2022, pp. 8906–8920.

[599] Tianhe Yu et al. "Gradient surgery for multi-task learning". In: *Advances in neural information processing systems* 33 (2020), pp. 5824–5836.

[600] Hanbo Zheng et al. "Investigation on molecular dynamics simulation for predicting kinematic viscosity of natural Ester insulating oil". In: *IEEE Transactions on Dielectrics and Electrical Insulation* 29.5 (2022), pp. 1882–1888.

[601] Shirley WI Siu, Kristyna Pluhackova, and Rainer A Bockmann. "Optimization of the OPLS-AA force field for long hydrocarbons". In: *Journal of Chemical Theory and Computation* 8.4 (2012), pp. 1459–1470.

[602] Kristyna Pluhackova et al. "Extension of the LOPLS-AA force field for alcohols, esters, and monoolein bilayers and its validation by neutron scattering experiments". In: *The Journal of Physical Chemistry B* 119.49 (2015), pp. 15287–15299.

[603] W D Cornell et al. "A second generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules". In: *Journal of the American Chemical Society* 117 (1995), pp. 5179–5197. ISSN: 0002-7863. DOI: 10.1021/ja00124a002.

[604] East Harbour Group. *Technical Specification: Dioctyl Sebacate*. https://www.eastharbourgroup.com/storage/download-files/dioctyl-sebacate-dos/Dioctyl-Sebacate-TDS-280-V1.pdf. Document Number: 280, Version: 1. 2022. URL: https://www.eastharbourgroup.com/storage/download-files/dioctyl-sebacate-dos/Dioctyl-Sebacate-TDS-280-V1.pdf.

[605] Sofia K Mylona et al. "Reference Correlations for the Density and Viscosity of Squalane from 273 to 473 K at Pressures to 200 MPa". In: *Journal of Physical and Chemical Reference Data* 43.1 (2014).

[606] Robert C Hardy. "Viscosity of n-Hexadecane". In: *Journal of Research of the National Bureau of Standards* 61.5 (1958), p. 433.

[607] M L P Price, D Ostrovsky, and W L Jorgensen. "Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field". In: *Journal of Computational Chemistry* 22 (13 2001), pp. 1340–1352. ISSN: 0192-8651. DOI: 10.1002/jcc.1092.

[608] Josh Kemppainen et al. "LUNAR: automated input generation and analysis for reactive LAMMPS simulations". In: *Journal of chemical information and modeling* 64.13 (2024), pp. 5108–5126.

[609] William L Jorgensen et al. "OPLS/2020 force field for unsaturated hydrocarbons, alcohols, and ethers". In: *The Journal of Physical Chemistry B* 128.1 (2023), pp. 250–262.

[610] EM Kirova and GE Norman. "Viscosity calculations at molecular dynamics simulations". In: *Journal of Physics: Conference Series*. Vol. 653. 1. IOP Publishing. 2015, p. 012106.

[611] Maxwell W Terban and Simon JL Billinge. "Structural analysis of molecular materials using the pair distribution function". In: *Chemical Reviews* 122.1 (2021), pp. 1208–1272.

[612] Panagiotis V Alatas, Dimitrios G Tsalikis, and Vlasis G Mavrantzas. "Detailed Molecular Dynamics Simulation of the Structure and Self-Diffusion of Linear and Cyclic n-Alkanes in Melt and Blends". In: *Macromolecular Theory and Simulations* 26.1 (2017), p. 1600049.

[613] Daniel J Cole et al. *The future of force fields in computer-aided drug design*. 2019.

[614] Abhishek T Sose et al. "A review of recent advances and applications of machine learning in tribology". In: *Physical Chemistry Chemical Physics* 25.6 (2023), pp. 4408–4443.

[615] Sereina Riniker. "Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences". In: *Journal of chemical information and modeling* 57.4 (2017), pp. 726–741.

[616] Aeon Technology. *Direct Force Field (DFF) Empowers Molecular Simulation*. https://www.aeontechnology.com/. Accessed: 2025-04-16. San Diego, CA, USA, 2025.

[617] Nikolay D Kondratyuk, Genri E Norman, and Vladimir V Stegailov. "Self-consistent molecular dynamics calculation of diffusion in higher n-alkanes". In: *The Journal of chemical physics* 145.20 (2016).

[618] ND Kondratyuk. "Comparing different force fields by viscosity prediction for branched alkane at 0.1 and 400 MPa". In: *Journal of Physics: Conference Series*. Vol. 1385. 1. IOP Publishing. 2019, p. 012048.

[619] Jeppe C Dyre. "Colloquium: The glass transition and elastic models of glass-forming liquids". In: *Reviews of modern physics* 78.3 (2006), pp. 953–972.

[620] D Van Velzen, R Lopes Cardozo, and H Langenkamp. "A liquid viscosity-temperature-chemical constitution relation for organic compounds". In: *Industrial & Engineering Chemistry Fundamentals* 11.1 (1972), pp. 20–25.

[621] Jeppe C Dyre and Wei Hua Wang. "The instantaneous shear modulus in the shoving model". In: *The Journal of chemical physics* 136.22 (2012).

[622] Zihao Wang et al. "Towards out-of-distribution generalizable predictions of chemical kinetics properties". In: *arXiv preprint arXiv:2310.03152* (2023).

[623] Daniel S Wigh, Jonathan M Goodman, and Alexei A Lapkin. "A review of molecular representation in the age of machine learning". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 12.5 (2022), e1603.

[624] Prudencio Tossou et al. "Real-world molecular out-of-distribution: Specification and investigation". In: *Journal of Chemical Information and Modeling* 64.3 (2024), pp. 697–711.

[625] Davide Boldini et al. "Effectiveness of molecular fingerprints for exploring the chemical space of natural products". In: *Journal of Cheminformatics* 16.1 (2024), p. 35.

[626] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. "The mahalanobis distance". In: *Chemometrics and intelligent laboratory systems* 50.1 (2000), pp. 1–18.

[627] Victor W Wong and Simon C Tung. "Overview of automotive engine friction and reduction trends–Effects of surface, material, and lubricant-additive technologies". In: *Friction* 4 (2016), pp. 1–28.

[628] Jiageng Ruan. *Design and verification of novel powertrain management for multi-geared battery electric vehicles*. University of Technology Sydney (Australia), 2016.

[629] Waleed Ahmed Abdalglil Mustafa et al. "A review on potentials and challenges of nanolubricants as promising lubricants for electric vehicles". In: *Lubrication Science* 34.1 (2022), pp. 1–29.

[630] Darinel Valencia-Marquez et al. "Computer aided molecular design coupled with molecular dynamics as a novel approach to design new lubricants". In: *Computers & Chemical Engineering* 156 (2022), p. 107523.

[631] Hanwen Wang et al. "Influence of ether group on viscosity and film lubrication of diester lubricants: Integrated quantitative structure–property relationship and molecular dynamics simulation methods". In:

Journal of Molecular Liquids (2024), p. 124254. ISSN: 01677322. DOI: 10.1016/j.molliq.2024.124254.

[632] Costas D Maranas. "Optimal computer-aided molecular design: A polymer design case study". In: *Industrial & Engineering Chemistry Research* 35.10 (1996), pp. 3403–3414.

[633] Norman L Allinger. "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms". In: *Journal of the American Chemical Society* 99.25 (1977), pp. 8127–8134.

[634] Norman L Allinger, Young H Yuh, and Jenn Huei Lii. "Molecular mechanics. The MM3 force field for hydrocarbons. 1". In: *Journal of the American Chemical Society* 111.23 (1989), pp. 8551–8566.

[635] Norman L Allinger, Kuohsiang Chen, and Jenn-Huei Lii. "An improved force field (MM4) for saturated hydrocarbons". In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 642–668.

[636] Stephen L Mayo, Barry D Olafson, and William A Goddard. "DREIDING: a generic force field for molecular simulations". In: *Journal of Physical Chemistry* 94.26 (1990), pp. 8897–8909.

[637] Anthony K Rappé et al. "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations". In: *Journal of the American Chemical Society* 114.25 (1992), pp. 10024–10035.

[638] Marcus G Martin and J Ilja Siepmann. "Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes". In: *The Journal of Physical Chemistry B* 102.14 (1998), pp. 2569–2577.

[639] Jeffrey J Potoff and Damien A Bernard-Brunel. "Mie potentials for phase equilibria calculations: Application to alkanes and perfluoroalkanes". In: *The Journal of Physical Chemistry B* 113.44 (2009), pp. 14725–14731.

[640] Siewert J Marrink et al. "The MARTINI force field: coarse grained model for biomolecular simulations". In: *The Journal of Physical Chemistry B* 111.27 (2007), pp. 7812–7824.

[641] Maurizio Mondello and Gary S Grest. "Viscosity calculations of n-alkanes by equilibrium molecular dynamics". In: *The Journal of Chemical Physics* 106.22 (1997), pp. 9327–9336.

[642] D Nevins and FJ Spera. "Accurate computation of shear viscosity from equilibrium molecular dynamics simulations". In: *Molecular Simulation* 33.15 (2007), pp. 1261–1266.

[643] Loukas I Kioupis and Edward J Maginn. "Molecular simulation of poly-α-olefin synthetic lubricants: impact of molecular architecture on performance properties". In: *The Journal of Physical Chemistry B* 103.49 (1999), pp. 10781–10790.

[644] E Lampin et al. "Thermal conductivity from approach-to-equilibrium molecular dynamics". In: *Journal of Applied Physics* 114.3 (2013).

[645] Rizky Ruliandini et al. "Thermal conductivity of trimethylolpropane trioleate (TMPTO) based bio-lubricant using molecular dynamic simulation". In: *AIP Conference Proceedings*. Vol. 2255. 1. AIP Publishing. 2020.

[646] N Kondratyuk, D Lenev, and V Pisarev. "Transport coefficients of model lubricants up to 400 MPa from molecular dynamics". In: *The Journal of Chemical Physics* 152.19 (2020).

[647] Kerstin Falk, Daniele Savio, and Michael Moseler. "Nonempirical free volume viscosity model for alkane lubricants under severe pressures". In: *Physical Review Letters* 124.10 (2020), p. 105501.

[648] Jannat Ahmed et al. "A novel method for fluid pour-point prediction by molecular dynamics simulations". In: *Tribology Transactions* 64.4 (2021), pp. 721–729.

[649] S Plumari et al. "Shear viscosity of a strongly interacting system: Green-Kubo correlator versus Chapman-Enskog and relaxation-time approximations". In: *Physical Review C* 86.5 (2012), p. 054902.

[650] RL Rowley and MM Painter. "Diffusion and viscosity equations of state for a Lennard-Jones fluid obtained from molecular dynamics simulations". In: *International Journal of Thermophysics* 18 (1997), pp. 1109–1121.

[651] Maria Panoukidou, Charlie R Wand, and Paola Carbone. "Comparison of equilibrium techniques for the viscosity calculation from DPD simulations". In: *Soft Matter* 17.36 (2021), pp. 8343–8353.

[652] S Bair. “The high pressure rheology of some simple model hydrocarbons”. In: *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology* 216.3 (2002), pp. 139–149.

[653] José Jiménez-Luna et al. “Artificial intelligence in drug discovery: recent advances and future perspectives”. In: *Expert Opinion on Drug Discovery* 16.9 (2021), pp. 949–959.

[654] Soohwan Jeong et al. “A novel graph-based missing values imputation method for industrial lubricant data”. In: *Computers in Industry* 150 (2023), p. 103937.

[655] Douglas M Hawkins. “The problem of overfitting”. In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004), pp. 1–12.

[656] John McCall. “Genetic algorithms for modelling and optimisation”. In: *Journal of Computational and Applied Mathematics* 184.1 (2005), pp. 205–222.

[657] Mahdi Ramezanizadeh et al. “A review on the utilized machine learning approaches for modeling the dynamic viscosity of nanofluids”. In: *Renewable and Sustainable Energy Reviews* 114 (2019), p. 109345.

[658] Venkat Venkatasubramanian, King Chan, and James M Caruthers. “Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm”. In: *Journal of Chemical Information and Computer Sciences* 35 (1995), pp. 188–195.

[659] Tarak K. Patra et al. “Neural-Network-Biased Genetic Algorithms for Materials Design: Evolutionary Algorithms That Learn”. In: *ACS Combinatorial Science* 19 (2 2017), pp. 96–107. ISSN: 21568944. DOI: 10.1021/acscombsci.6b00136.

[660] Venkatesh Meenakshisundaram et al. “Designing Sequence-Specific Copolymer Compatibilizers Using a Molecular-Dynamics-Simulation-Based Genetic Algorithm”. In: *Macromolecules* 50 (3 2017), pp. 1155–1166. ISSN: 15205835. DOI: 10.1021/acs.macromol.6b01747.

[661] Venkatesh Meenakshisundaram, Jui Hsiang Hung, and David S. Simmons. “Design rules for glass formation from model molecules designed by a neural-network-biased genetic algorithm”. In: *Soft Matter* 15 (39 2019), pp. 7795–7808. ISSN: 17446848. DOI: 10.1039/c9sm01486a.

[662] Tianhang Zhou et al. “Sequence-Engineering Polyethylene-Polypropylene Copolymers with High Thermal Conductivity Using a Molecular-Dynamics-Based Genetic Algorithm”. In: *Journal of Chemical Theory and Computation* 17 (6 2021), pp. 3772–3782. ISSN: 15499626. DOI: 10.1021/acs.jctc.1c00134.

[663] Duxin Li et al. “Hybrid neural network-based prediction model for tribological properties of polyamide6-based friction materials”. In: *Polymer Composites* 38.8 (2017), pp. 1705–1711.

[664] Shubrajit Bhaumik, Behanan Roy Mathew, and Shubhabrata Datta. “Computational intelligence-based design of lubricant with vegetable oil blend and various nano friction modifiers”. In: *Fuel* 241 (2019), pp. 733–743.

[665] Hui Zhang et al. “Improvement of tribological behaviors by optimizing concave texture shape under reciprocating sliding motion”. In: *Journal of Tribology* 139.1 (2017), p. 011701.

[666] Riccardo Leardi. “Genetic algorithms in chemistry”. In: *Journal of Chromatography A* 1158.1-2 (2007), pp. 226–233.

[667] Danielle C Hiener and Geoffrey R Hutchison. “Pareto optimization of oligomer polarizability and dipole moment using a genetic algorithm”. In: *The Journal of Physical Chemistry A* 126.17 (2022), pp. 2750–2760.

[668] AkshatKumar Nigam, Robert Pollice, and Alán Aspuru-Guzik. “Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design”. In: *Digital Discovery* 1.4 (2022), pp. 390–404.

[669] Jan H Jensen. “A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space”. In: *Chemical Science* 10.12 (2019), pp. 3567–3572.

[670] Brianna L Greenstein, Danielle C Elsey, and Geoffrey R Hutchison. “Determining best practices for using genetic algorithms in molecular discovery”. In: *The Journal of Chemical Physics* 159.9 (2023), p. 091501.

[671] Ayesha Khan and Abdul Rauf Baig. “Multi-objective feature subset selection using non-dominated sorting genetic algorithm”. In: *Journal of applied research and technology* 13.1 (2015), pp. 145–159.

[672] Jeffrey Horn, Nicholas Nafpliotis, and David E Goldberg. “A niched Pareto genetic algorithm for multiobjective optimization”. In: *Proceedings of the first IEEE conference on evolutionary computation. IEEE world congress on computational intelligence*. Ieee. 1994, pp. 82–87.

[673] Khalid Jebari, Mohammed Madiafi, et al. “Selection methods for genetic algorithms”. In: *International Journal of Emerging Sciences* 3.4 (2013), pp. 333–344.

[674] Silvia RM Pereira et al. “Effect of the genetic algorithm parameters on the optimisation of heterogeneous catalysts”. In: *QSAR & Combinatorial Science* 24.1 (2005), pp. 45–57.

[675] D Brynn Hibbert. “Genetic algorithms in chemistry”. In: *Chemometrics and Intelligent Laboratory Systems* 19.3 (1993), pp. 277–293.

[676] Alain Pétrowski. “A clearing procedure as a niching method for genetic algorithms”. In: *Proceedings of IEEE international conference on evolutionary computation*. IEEE. 1996, pp. 798–803.

[677] Georges R Harik et al. “Finding multimodal solutions using restricted tournament selection.” In: *ICGA*. 1995, pp. 24–31.

[678] Pablo A Estévez and Rodrigo E Caballero. “A niching genetic algorithm for selecting features for neural network classifiers”. In: *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998 8*. Springer. 1998, pp. 311–316.

[679] Andreas Bender and Robert C Glen. “Molecular similarity: a key technique in molecular informatics”. In: *Organic & Biomolecular Chemistry* 2.22 (2004), pp. 3204–3218.

[680] Raymond E Carhart, Dennis H Smith, and Rengachari Venkataraghavan. “Atom pairs as molecular features in structure-activity studies: definition and applications”. In: *Journal of Chemical Information and Computer Sciences* 25.2 (1985), pp. 64–73.

[681] Ingo Muegge and Prasenjit Mukherjee. “An overview of molecular fingerprint similarity search in virtual screening”. In: *Expert Opinion on Drug Discovery* 11.2 (2016), pp. 137–148.

[682] David E Patterson et al. “Neighborhood behavior: a useful concept for validation of “molecular diversity” descriptors”. In: *Journal of Medicinal Chemistry* 39.16 (1996), pp. 3049–3059.

[683] Dávid Bajusz, Anita Rácz, and Károly Héberger. “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?” In: *Journal of Cheminformatics* 7 (2015), pp. 1–13.

[684] Sylvain Verdier et al. “A critical approach to viscosity index”. In: *Fuel* 88.11 (2009), pp. 2199–2206.

[685] ASTM International. *ASTM D2270-93 Standard Practice for Calculating Viscosity Index From Kinematic Viscosity at 40 and 100°C*. 2016.

[686] MJ Covitch. “An improved method for calculating viscosity index (VI) of low viscosity base oils”. In: *Journal of Testing and Evaluation* 46.2 (2018), pp. 820–825.

[687] Jack Zakarian. “The Limitations of the Viscosity Index and Proposals for Other Methods to Rate Viscosity-Temperature Behavior of Lubricating Oils”. In: *SAE International Journal of Fuels and Lubricants* 5 (3 2012), pp. 1123–1131. ISSN: 19463952. DOI: 10.4271/2012-01-1671.

[688] Oscar Méndez-Lucio and José L Medina-Franco. “The many roles of molecular complexity in drug discovery”. In: *Drug Discovery Today* 22.1 (2017), pp. 120–126.

[689] Yuji Takaoka et al. “Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists’ intuition”. In: *Journal of Chemical Information and Computer Sciences* 43.4 (2003), pp. 1269–1275.

[690] Yevgeniy Podolyan, Michael A Walters, and George Karypis. “Assessing synthetic accessibility of chemical compounds using machine learning methods”. In: *Journal of Chemical Information and Modeling* 50.6 (2010), pp. 979–991.

[691] Peter Ertl and Ansgar Schuffenhauer. “Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions”. In: *Journal of Cheminformatics* 1 (2009), pp. 1–11.

[692] Connor W Coley et al. “SCScore: synthetic complexity learned from a reaction corpus”. In: *Journal of Chemical Information and Modeling* 58.2 (2018), pp. 252–261.

[693] Alexander J Lawson et al. “The making of reaxys—towards unobstructed access to relevant chemistry information”. In: *The Future of the History of Chemical Information*. ACS Publications, 2014, pp. 127–148.

[694] Paulo Renato Matos Lopes et al. “Toxicity and biodegradation in sandy soil contaminated by lubricant oils”. In: *Bulletin of Environmental Contamination and Toxicology* 84 (2010), pp. 454–458.

[695] EUROPEAN UNION. “Regulation (EC) No 1223/2009 of the european parliament and of the council”. In: *Official Journal of the European Union L* 342 (2009), p. 59.

[696] Jannis Born et al. “Chemical representation learning for toxicity prediction”. In: *Digital Discovery* 2.3 (2023), pp. 674–691.

[697] Raymond R Tice et al. “Improving the human hazard characterization of chemicals: a Tox21 update”. In: *Environmental Health Perspectives* 121.7 (2013), pp. 756–765.

[698] Ponnekanti Nagendramma and Savita Kaul. “Development of ecofriendly/biodegradable lubricants: An overview”. In: *Renewable and sustainable energy reviews* 16.1 (2012), pp. 764–774.

[699] NS Battersby. “Biodegradable lubricants: What does ‘biodegradable’really mean?” In: *Journal of Synthetic Lubrication* 22.1 (2005), pp. 3–18.

[700] Weihao Tang et al. “Development of models predicting biodegradation rate rating with multiple linear regression and support vector machine algorithms”. In: *Chemosphere* 253 (2020), p. 126666. ISSN: 0045-6535. DOI: https://doi.org/10.1016/j.chemosphere.2020.126666. URL: https://www.sciencedirect.com/science/article/pii/S0045653520308596.

[701] Zhixiong Zhan et al. “Prediction of chemical biodegradability using computational methods”. In: *Molecular Simulation* 43.13-16 (2017), pp. 1277–1290.

[702] Hongyan Yin et al. “Prediction and structure–activity relationship analysis on ready biodegradability of chemical using machine learning method”. In: *Chemical Research in Toxicology* 36.4 (2023), pp. 617–629.

[703] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[704] Jianpeng Wu et al. “Data-Driven Prediction of Coefficient of Friction in Wet Friction Components: A Model Development and Interpretability Analysis”. In: *Journal of Tribology* 147.7 (2025).

[705] Artur Movsessian, David Garcia Cava, and Dmitri Tcherniak. “Interpretable machine learning in damage detection using shapley additive explanations”. In: *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering* 8.2 (2022), p. 021101.

[706] Mood Mohan et al. “High-throughput screening and accurate prediction of ionic liquid viscosities using interpretable machine learning”. In: *ACS Sustainable Chemistry & Engineering* 12.18 (2024), pp. 7040–7054.

[707] Ana Victoria Ponce-Bobadilla et al. “Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development”. In: *Clinical and Translational Science* 17.11 (2024), e70056.

[708] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. “Consistent individualized feature attribution for tree ensembles”. In: *arXiv preprint arXiv:1802.03888* (2018).

[709] Davide Chicco and Giuseppe Jurman. “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC genomics* 21 (2020), pp. 1–13.

[710] Lemont B Kier and Lowell H Hall. “An electrotopological-state index for atoms in molecules”. In: *Pharmaceutical research* 7 (1990), pp. 801–807.

[711] Yuanyuan Dan et al. “Predicting the Toxicity of Drug Molecules with Selecting Effective Descriptors Using a Binary Ant Colony Optimization (BACO) Feature Selection Approach”. In: *Molecules* 30.7 (2025), p. 1548.

[712] Gilles Moreau and Pierre Broto. “The autocorrelation of a topological structure: a new molecular descriptor”. In: (1980).

[713] Vikas Sharma, Reena Goswami, and AK Madan. “Eccentric connectivity index: A novel highly discriminating topological descriptor for structure- property and structure- activity studies”. In: *Journal of chemical information and computer sciences* 37.2 (1997), pp. 273–282.

[714] Thomas R Lynch. *Process chemistry of lubricant base stocks*. CRC Press, 2007.

[715] Pierre Legendre. “Spatial autocorrelation: trouble or new paradigm?” In: *Ecology* 74.6 (1993), pp. 1659–1673.

[716] Ashutosh Giri, Connor J Dionne, and Patrick E Hopkins. “Atomic coordination dictates vibrational characteristics and thermal conductivity in amorphous carbon”. In: *npj Computational Materials* 8.1 (2022), p. 55.

[717] Johann Gasteiger and Mario Marsili. “Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges”. In: *Tetrahedron* 36.22 (1980), pp. 3219–3228.

[718] Scott A Wildman and Gordon M Crippen. “Prediction of physicochemical parameters by atomic contributions”. In: *Journal of chemical information and computer sciences* 39.5 (1999), pp. 868–873.

[719] Weijing Dai, Dorian Hanaor, and Yixiang Gan. “The effects of packing structure on the effective thermal conductivity of granular media: A grain scale investigation”. In: *International Journal of Thermal Sciences* 142 (2019), pp. 266–279.

[720] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

[721] Lowell H Hall and Lemont B Kier. “The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling”. In: *Reviews in computational chemistry* (1991), pp. 367–422.

[722] Bruce E Poling, John M Prausnitz, John P O’connell, et al. *The properties of gases and liquids*. Vol. 5. Mcgraw-hill New York, 2001.

[723] Yuan H Zhao, Michael H Abraham, and Andreas M Zissimos. “Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds”. In: *The Journal of organic chemistry* 68.19 (2003), pp. 7368–7373.

[724] Michael H Abraham and JC McGowan. “The use of characteristic volumes to measure cavity terms in reversed phase liquid chromatography”. In: *Chromatographia* 23.4 (1987), pp. 243–246.

[725] Arij Ben Amara, Skander Kaoubi, and Laurie Starck. “Toward an optimal formulation of alternative jet fuels: Enhanced oxidation and thermal stability by the addition of cyclic molecules”. In: *Fuel* 173 (2016), pp. 98–105.

[726] Adrian Venegas-Reynoso et al. “Oxidation Stability of Hydrocarbons: A Machine-Learning-Based Study”. In: *Energy & Fuels* 39.9 (2025), pp. 4361–4373.

[727] Karl Chatelain et al. “Wide range experimental and kinetic modeling study of chain length impact on n-alkanes autoxidation”. In: *Energy & Fuels* 30.2 (2016), pp. 1294–1303.

[728] Quang N. Le. “Lubricant composition of polyalphaolefn and alkylated aromatic fuid”. In: *U.S. Patent 5,602,086* (1997).

[729] Shuji Luo, Suzzy C Ho, and Margaret M Wu. “Alkylated Aromatics”. In: *Synthetics, Mineral Oils, and Bio-Based Lubricants* (2020), pp. 161–180.

[730] Peter Brown. “Synthetic basestocks (Groups IV and V) in lubricant applications”. In: *Tribology & Lubrication Technology* 59.9 (2003), p. 20.

[731] University of Washington. *Dielectric Constant of Common Solvents*. Accessed: 2025-05-01. URL: https://depts.washington.edu/eooptic/linkfiles/dielectric_chart%5B1%5D.pdf.

[732] P.K Wong and J Wang. “The accumulation of polycyclic aromatic hydrocarbons in lubricating oil over time — a comparison of supercritical fluid and liquid–liquid extraction methods”. In: *Environmental Pollution* 112.3 (2001), pp. 407–415. ISSN: 0269-7491. DOI: https://doi.org/10.1016/S0269-7491(00)00142-1.

[733] Ronald G Harvey. “Polycyclic hydrocarbons and carcinogenesis”. In: (1985).

[734] Walter Holweger. “Fundamentals of lubricants and lubrication”. In: *Tribology-fundamentals and advancements*. IntechOpen, 2013.

[735] Theo Mang and Wilfried Dresel. *Lubricants and lubrication*. John Wiley & Sons, 2007.

[736] Kenneth De Jong. “Evolutionary computation: a unified approach”. In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2017, pp. 373–388.

[737] Jon Louis Bentley et al. “On the average number of maxima in a set of vectors and applications”. In: *Journal of the ACM (JACM)* 25.4 (1978), pp. 536–543.

[738] Weijie Zheng and Benjamin Doerr. “Mathematical runtime analysis for the non-dominated sorting genetic algorithm II (NSGA-II)”. In: *Artificial Intelligence* 325 (2023), p. 104016.

[739] James P Ewen et al. “Nonequilibrium molecular dynamics simulations of stearic acid adsorbed on iron surfaces with nanoscale roughness”. In: *Tribology International* 107 (2017), pp. 264–273.

[740] Stavros Ntioudis et al. “A hybrid off-lattice kinetic Monte Carlo/molecular dynamics method for amorphous thin film growth”. In: *Computational Materials Science* 229 (2023), p. 112421.

[741] Markus Hartenfeller et al. “Concept of combinatorial de novo design of drug-like molecules by particle swarm optimization”. In: *Chemical biology & drug design* 72.1 (2008), pp. 16–26.

[742] Ksenia Korovina et al. “Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3393–3403.

[743] Zhenpeng Zhou et al. “Optimization of molecules via deep reinforcement learning”. In: *Scientific reports* 9.1 (2019), p. 10752.

[744] Xuanxiang Huang and Joao Marques-Silva. “On the failings of Shapley values for explainability”. In: *International Journal of Approximate Reasoning* 171 (2024), p. 109112.

Appendix A

Supporting information for Chapter 3

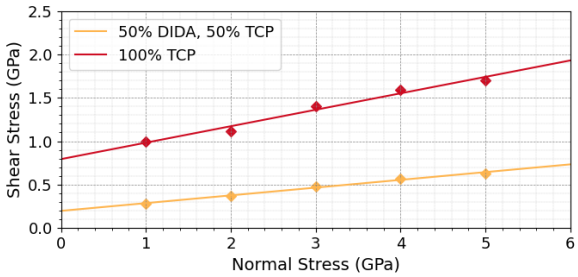


Figure A.1: Change in shear stress with normal stress for 100% TCP and 50% TCP, 50% DIDA at 600K and 1-5 GPa. Friction coefficient, μ = 0.18 (100% TCP), μ = 0.17 (50% TCP, 50% DIDA) and Derjaguin offset, τ_0 = 0.79 GPa (100% TCP), τ_0 = 0.20 GPa (50% TCP, 50% DIDA).

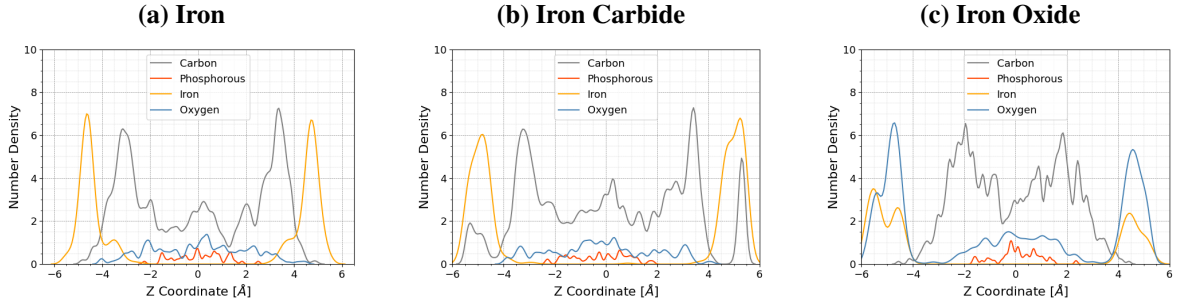


Figure A.2: Number density profiles for the Fe, C, O, and P atoms in the a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) systems at 500 K and 3 GPa

	Stiffness (GPa)						
Temperature	300 K	400 K	500 K	600 K	700 K	800 K	900 K
Iron	206	206	202	200	200	199	196
Iron Carbide	200	200	201	198	200	196	181
Iron Oxide	220	216	210	209	206	199	196

Table A.1: Elastic constant (C_{33}) values for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) slabs with ReaxFF.

	Friction Coefficient (μ)					
Temperature	400 K	500 K	600 K	700 K	800 K	900 K
Iron	0.35	0.22	0.18	0.25	0.17	0.16
Iron Carbide	0.28	0.31	0.23	0.24	0.19	0.20
Iron Oxide	0.24	0.24	0.23	0.23	0.26	0.27

Table A.2: Friction Coefficient, μ, for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces at different temperatures.

	Derjaguin Offset					
Temperature	400 K	500 K	600 K	700 K	800 K	900 K
Iron	0.28	0.72	0.79	0.53	0.75	0.79
Iron Carbide	0.38	0.31	0.66	0.57	0.73	0.77
Iron Oxide	0.26	0.27	0.31	0.32	0.34	0.34

Table A.3: Derjaguin offset for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces at different temperatures.

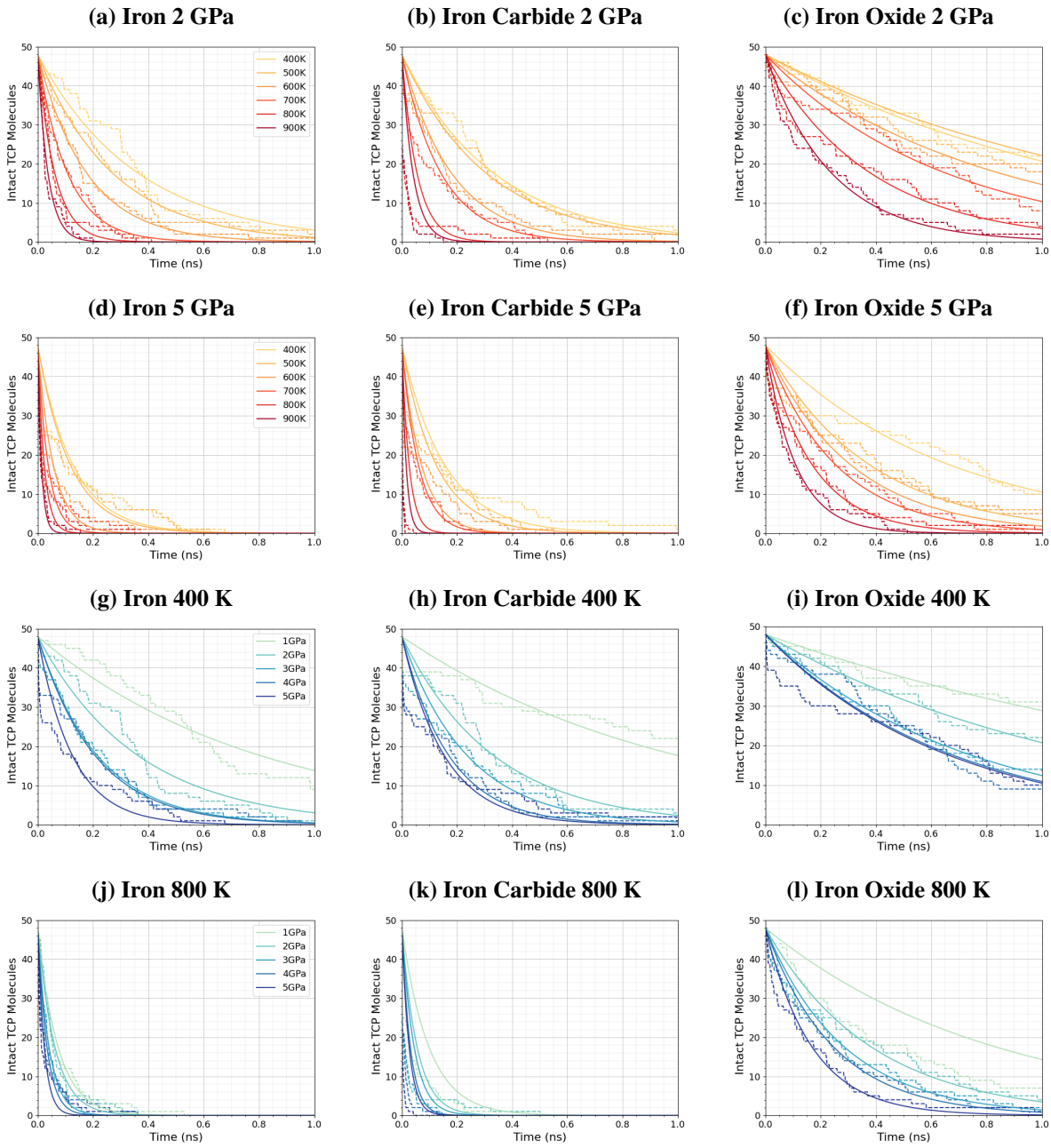


Figure A.3: Variation in TCP dissociation rates for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) under various conditions.

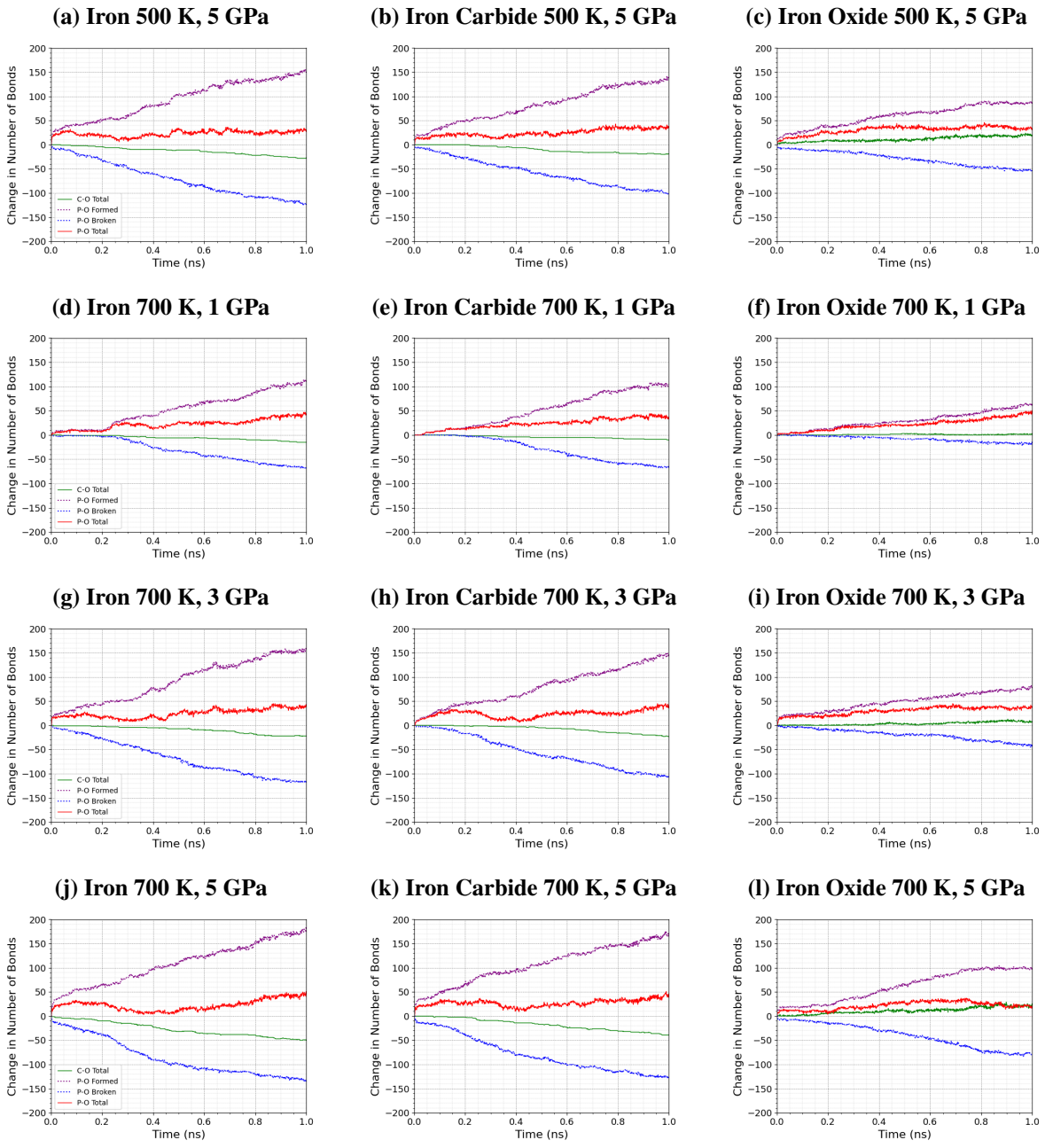


Figure A.4: Change in the number of P-O and C-O bonds compared to TCP, for α-Fe(110), Fe_3C(010), and Fe_3O_4(001) under various conditions. Solid lines represent total change, dashed lines show bond formation and dotted lines show bond cleavage.

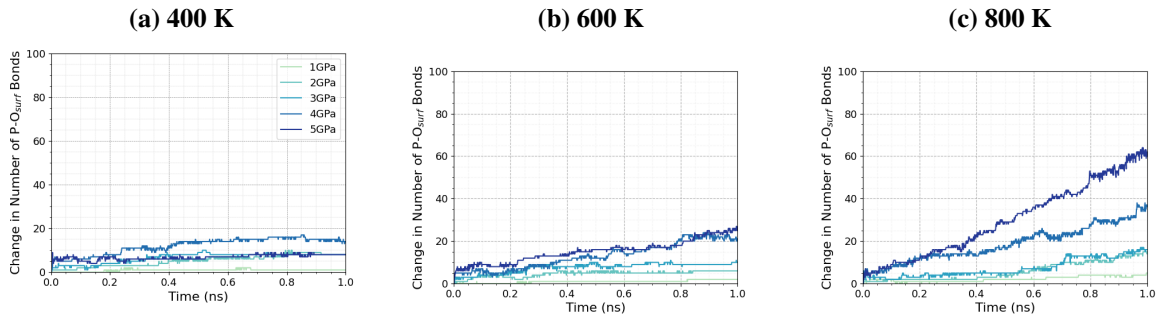


Figure A.5: Comparison of P-O_{surf} bonds formed in the Fe_3O_4(100) system at a) 400 K, b) 600 K, and c) 800 K

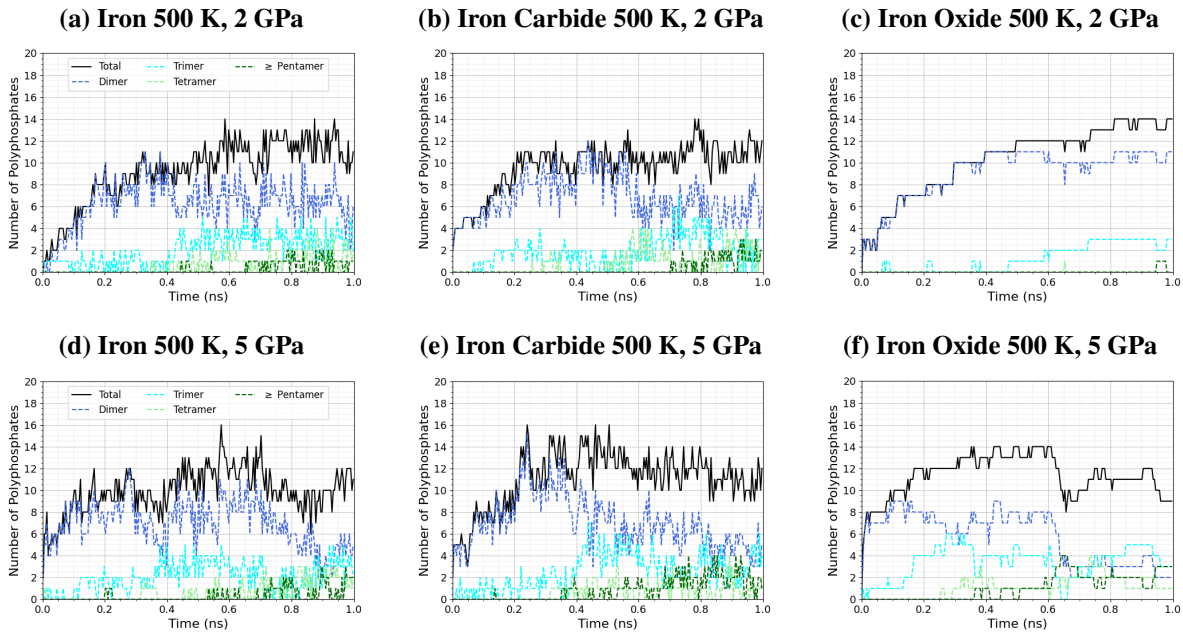


Figure A.6: Change in the number of polyphosphate chains with sliding time for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) systems.

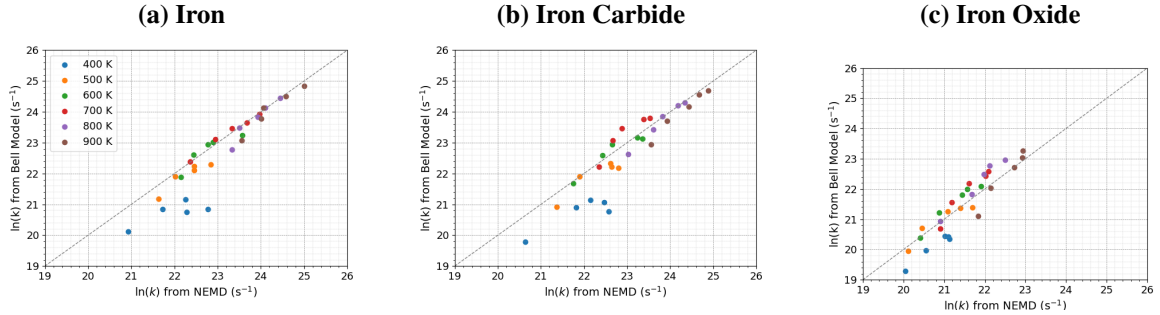


Figure A.7: Comparison of the logarithm of the reaction rates calculated from the NEMD simulations and those predicted using Equation 1. R^2 = 0.91 for α-Fe(110), R^2 = 0.88 for Fe_3C(010), and R^2 = 0.87 for Fe_3O_4(001).

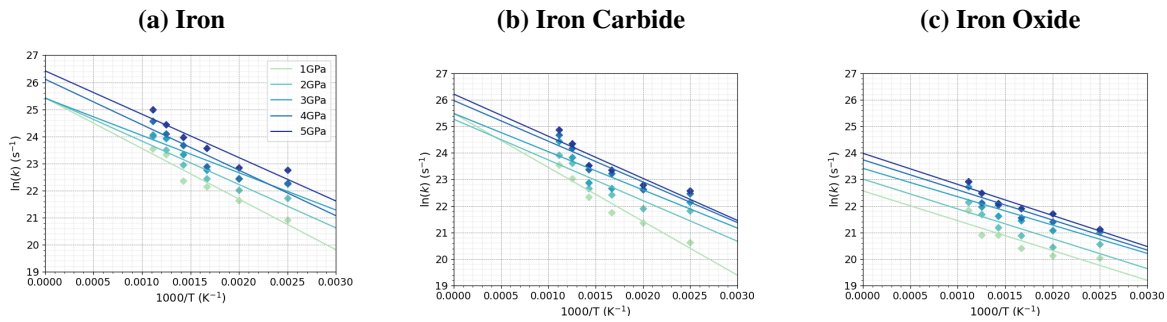


Figure A.8: Arrhenius plots of ln(k) versus 1000/T for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001) obtained at different pressures. From Equation 1, the intercept of the lines gives A and gradient gives E_a.

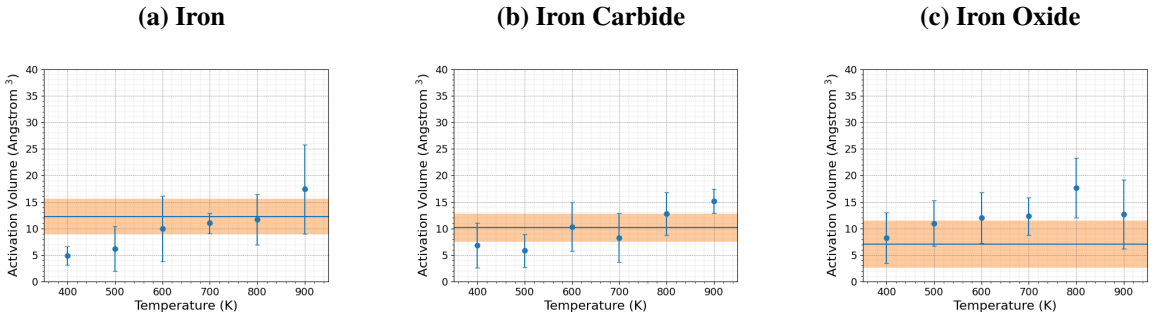


Figure A.9: Temperature dependence of ΔV^* for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals.

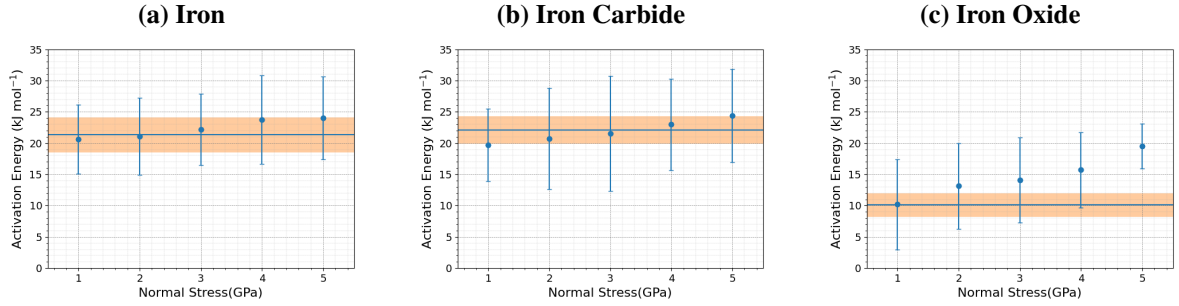


Figure A.10: Pressure dependence of E_a for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals.

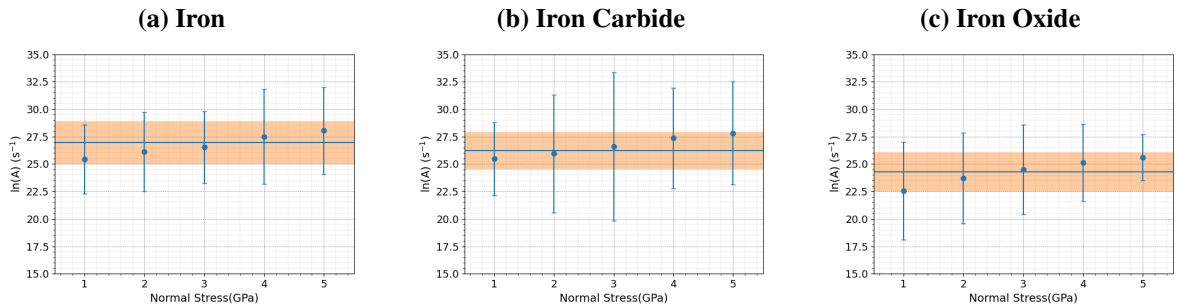


Figure A.11: Pressure dependence of ln(A) for a) α-Fe(110), b) Fe_3C(010), and c) Fe_3O_4(001). Blue points are from the 2D fits, orange are from the 3D fits. Vertical bars indicate 95 % confidence intervals.

	ΔV* (Å^3)					
	400 K	500 K	600 K	700 K	800 K	900 K
Iron	4.9 ± 1.7	6.2 ± 4.2	10.0 ± 6.2	11.0 ± 1.9	11.7 ± 4.7	17.4 ± 8.4
Iron Carbide	6.8 ± 4.2	5.5 ± 3.1	10.3 ± 4.6	8.3 ± 4.6	12.8 ± 4.0	15.2 ± 2.3
Iron Oxide	8.3 ± 4.5	10.9 ± 4.3	12.0 ± 4.8	12.6 ± 3.8	17.7 ± 5.6	12.8 ± 6.6

Table A.4: ΔV^* for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces.

	E_a (kJ mol^{-1})				
	1 GPa	2 GPa	3 GPa	4 GPa	5 GPa
Iron	20.6 ± 5.5	21.1 ± 6.2	22.2 ± 5.7	23.7 ± 7.1	24.0 ± 6.6
Iron Carbide	19.7 ± 5.8	20.7 ± 8.1	21.5 ± 9.2	23.0 ± 7.3	24.3 ± 7.4
Iron Oxide	10.4 ± 7.2	13.2 ± 6.9	14.1 ± 6.8	15.8 ± 6.1	19.4 ± 3.6

Table A.5: E_a for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces.

	ln(A)(s^{-1})				
	1 GPa	2 GPa	3 GPa	4 GPa	5 GPa
Iron	25.4 ± 3.1	26.1 ± 3.6	26.5 ± 3.3	27.5 ± 4.3	28.0 ± 3.9
Iron Carbide	25.4 ± 3.3	26.0 ± 5.4	26.6 ± 6.7	27.4 ± 4.6	27.8 ± 4.7
Iron Oxide	22.6 ± 4.5	23.7 ± 4.1	24.5 ± 4.1	25.1 ± 3.5	25.6 ± 2.1

Table A.6: ln(A) for TCP between α-Fe(110), Fe_3C(010), and Fe_3O_4(001) surfaces.

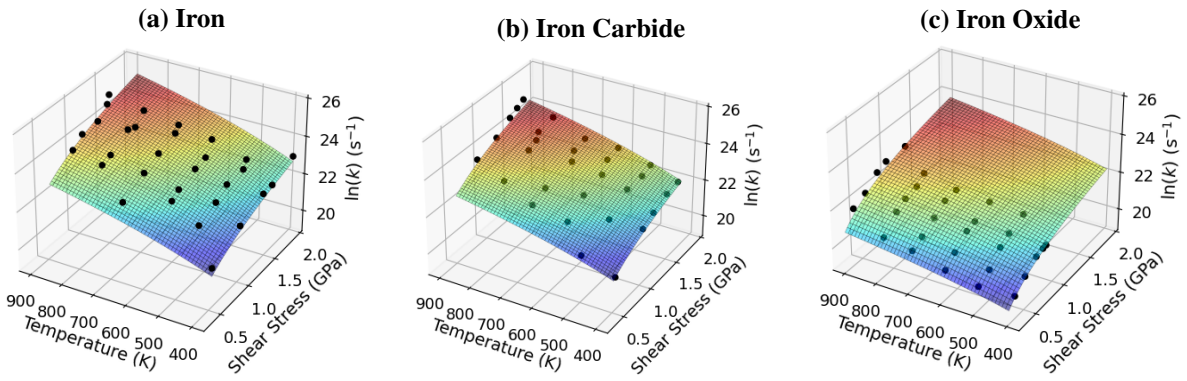
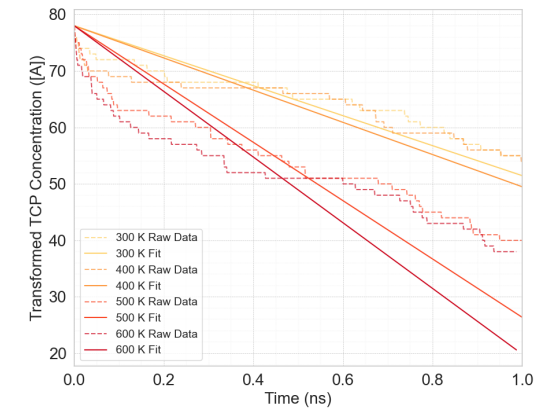


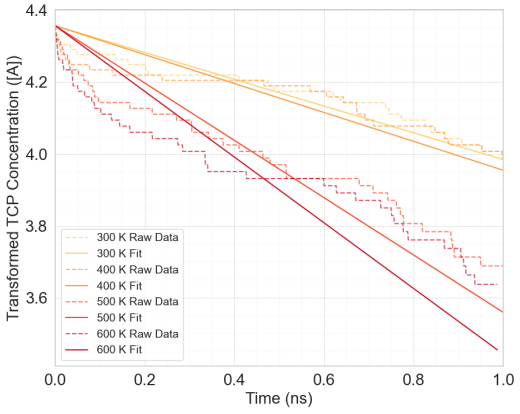
Figure A.12: 3D fits of TCP decomposition ln(k) as a function of temperature and shear stress for a) α-Fe(110), b) Fe_3C(010) and c) Fe_3O_4(100) surfaces.

Appendix B

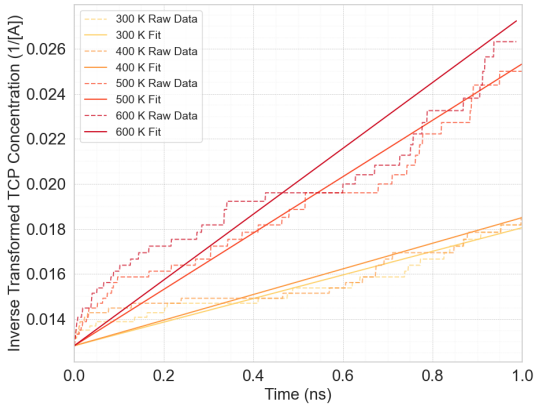
Supporting information for Chapter 4



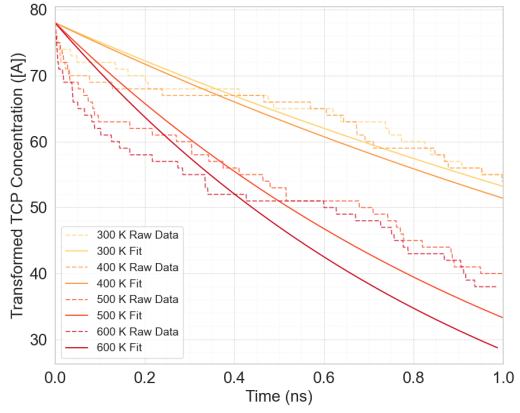
(a) Zero-Order Fit



(b) First-Order Fit



(c) Second-Order Fit

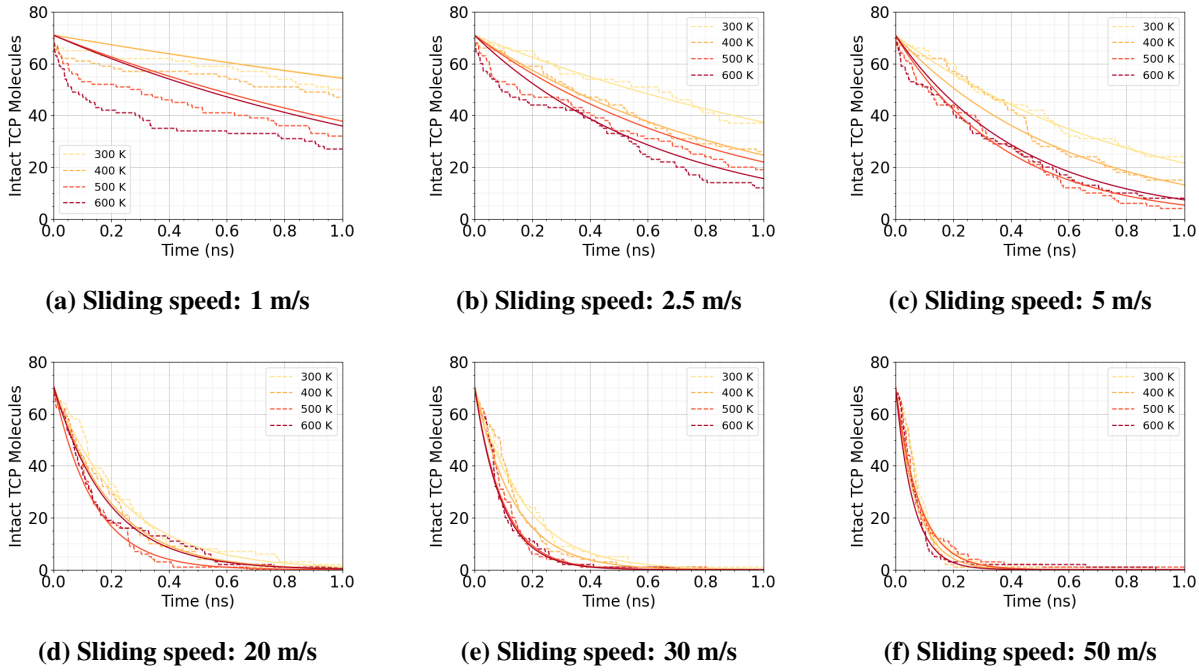


(d) Exponential Fit

Figure B.1: Plots of different fitting methods used to determine rate constant for the DLC-α-Fe(110) systems at 1 m s^{-1}, R^2 for each fit shown in the legend.

Model	0 m s^{-1}	1 m s^{-1}	2.5 m s^{-1}	5 m s^{-1}	10 m s^{-1}	20 m s^{-1}	30 m s^{-1}	50 m s^{-1}
Exponential	0.3753	0.6718	0.9735	0.9911	0.9889	0.9742	0.9753	0.9762
First-Order	0.3668	0.7525	0.9831	0.9865	0.9748	0.9531	0.3609	0.1541
Second-Order	0.5853	0.9208	0.9244	0.6961	0.5475	0.5464	0.8161	0.7948
Zero-Order	0.0716	0.3434	0.7891	0.7100	0.3493	-0.3744	-2.8131	-5.6675

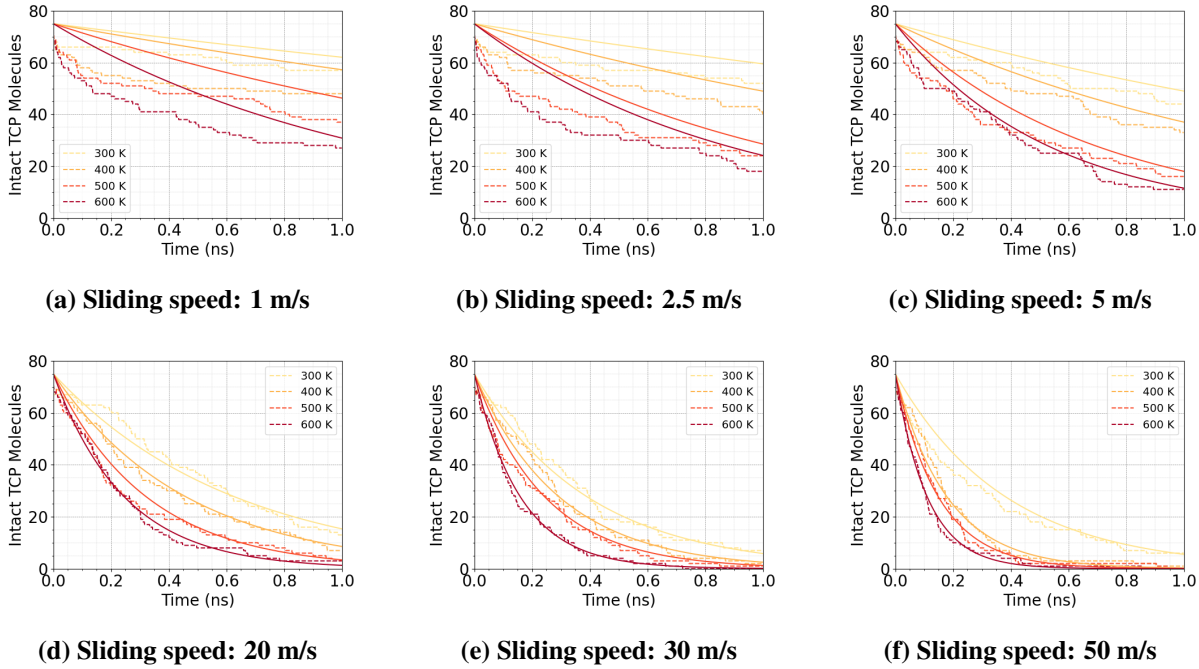
Table B.1: R^2 values for the different fitting methods at 500 K for DLC-α-Fe(110).



(a) Sliding speed: 1 m/s **(b) Sliding speed: 2.5 m/s** **(c) Sliding speed: 5 m/s**

(d) Sliding speed: 20 m/s **(e) Sliding speed: 30 m/s** **(f) Sliding speed: 50 m/s**

Figure B.2: Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds and temperatures for the DLC-α-Fe(110) model.



(a) Sliding speed: 1 m/s **(b) Sliding speed: 2.5 m/s** **(c) Sliding speed: 5 m/s**

(d) Sliding speed: 20 m/s **(e) Sliding speed: 30 m/s** **(f) Sliding speed: 50 m/s**

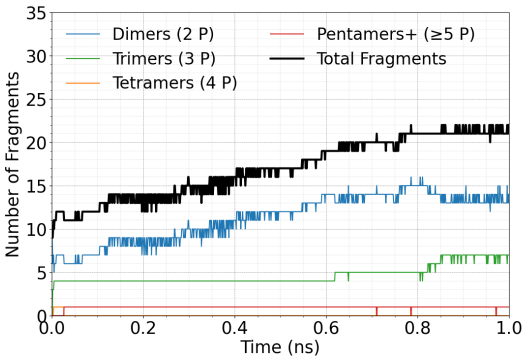
Figure B.3: Comparison of TCP molecule decomposition at 5 GPa at different sliding speeds and temperatures for the DLC-Fe_2O_3 model.

Speed (m/s)	**DLC-α-Fe(110)** E_a (kJ/mol)	**DLC-α-Fe(110)** ln(A) (s^{-1})	**DLC-Fe_2O_3** E_a (kJ/mol)	**DLC-Fe_2O_3** ln(A) (s^{-1})
1	13.20	21.35	13.66	23.20
2.5	9.81	21.91	11.35	21.88
5	7.61	22.39	9.67	22.56
10	4.56	22.84	8.75	23.04
20	3.40	22.93	8.20	22.00
30	3.39	23.65	8.16	23.26
50	1.54	23.62	5.23	24.15

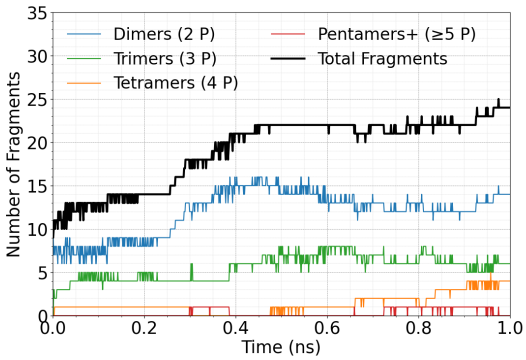
Table B.2: Arrhenius activation energies and pre-exponential factors for DLC-α-Fe(110) and DLC-Fe_2O_3.

Speed (m/s)	**DLC-α-Fe(110)** E_a (kJ/mol)	**DLC-Fe_2O_3** E_a (kJ/mol)
1	12.74	13.26
2.5	9.22	10.71
5	7.09	9.13
10	4.04	8.10
20	3.75	7.52
30	2.60	7.47
50	0.76	4.43

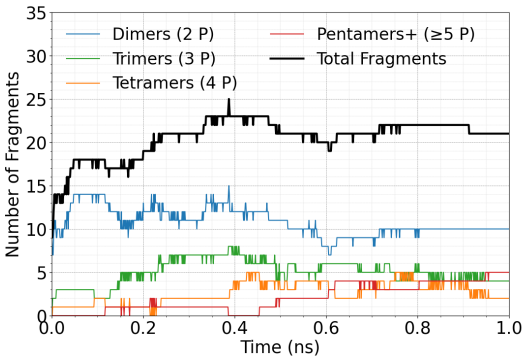
Table B.3: Average Bell activation energies for DLC-α-Fe(110) and DLC-Fe_2O_3.



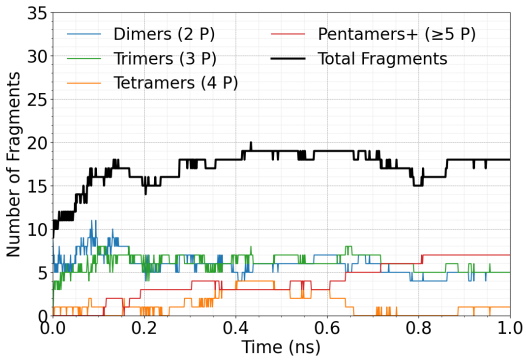
(a) 1 m/s



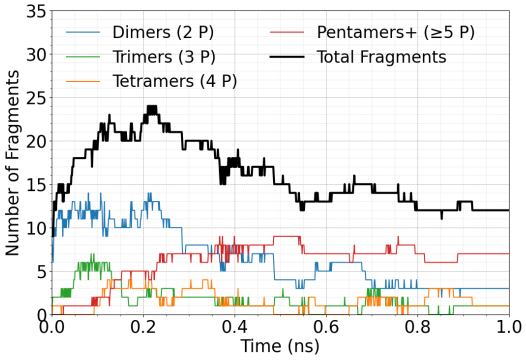
(b) 2.5 m/s



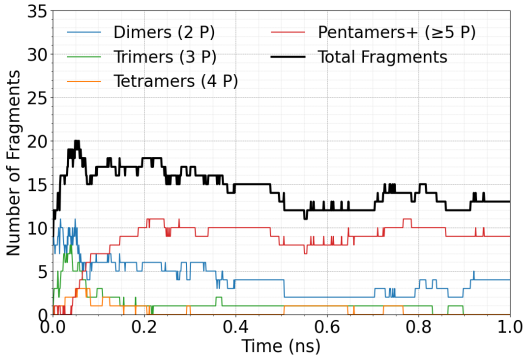
(c) 5 m/s



(d) 10 m/s

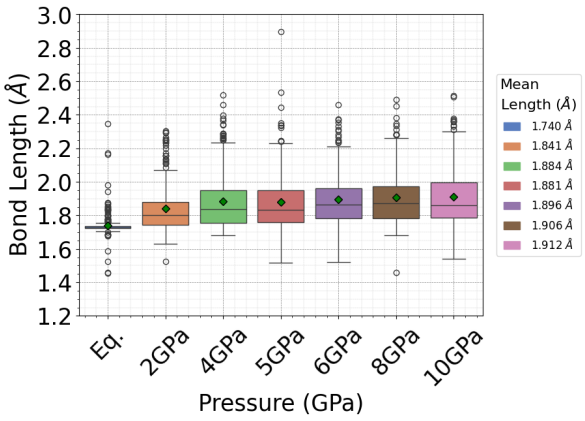


(e) 20 m/s

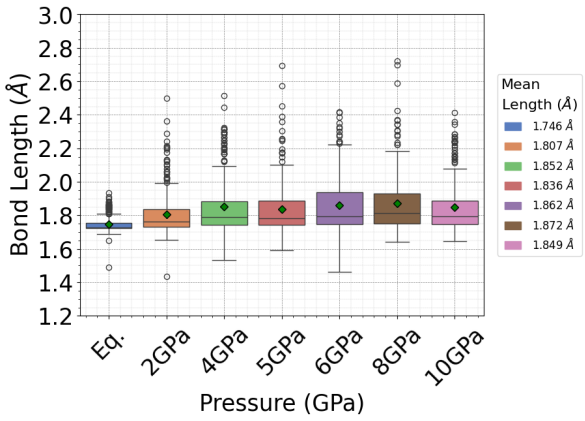


(f) 50 m/s

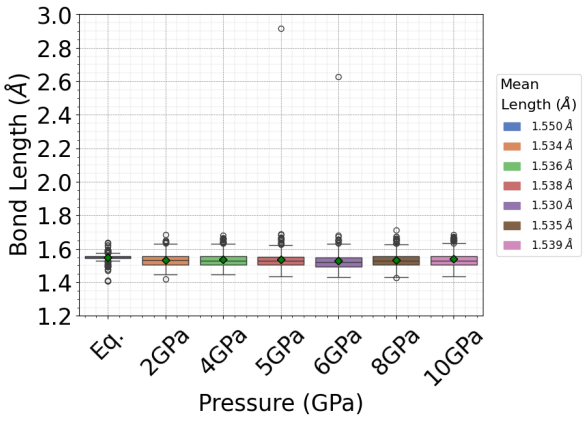
Figure B.4: ime evolution of polyphosphate formation at 500 K and 5GPa, for various sliding speeds for DLC-α-Fe(110) system.



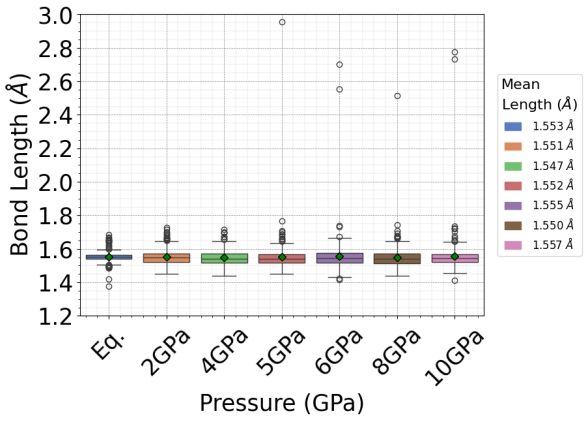
(a) Single P-O Bonds DLC-Fe_2O_3



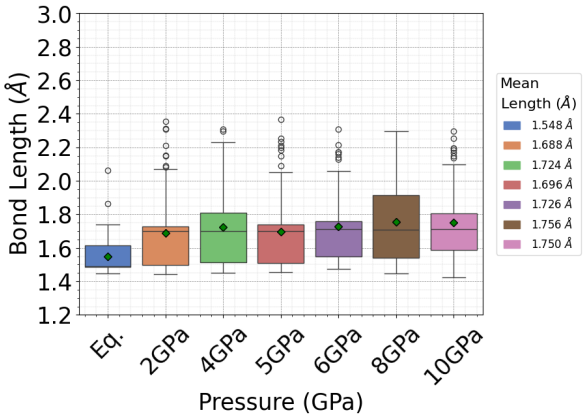
(b) Single P-O Bonds DLC-α-Fe(110)



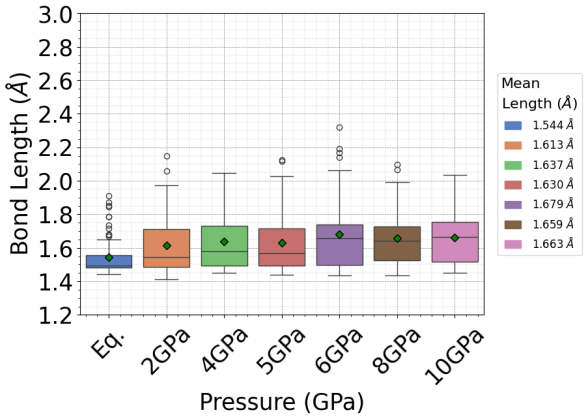
(c) C-O Bonds DLC-Fe_2O_3



(d) C-O Bonds DLC-α-Fe(110)

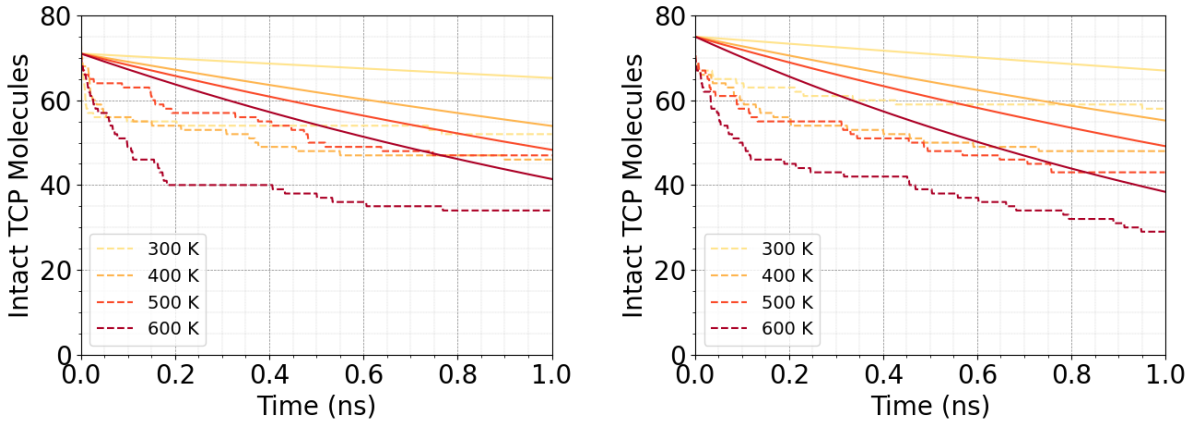


(e) Double P-O Bonds DLC-Fe_2O_3



(f) Double P-O Bonds DLC-α-Fe(110)

Figure B.5: Comparison of bond length distributions as a function of normal pressure between DLC-Fe_2O_3 and DLC-α-Fe(110) systems for selected bond types.



(a) DLC-α-Fe(110) **(b) DLC-Fe_2O_3**

Figure B.6: Comparison of TCP molecule decomposition at 5 GPa at 0 m/s for a) DLC-α-Fe(110) and b) DLC-Fe_2O_3.

Appendix C

Supporting information for Chapter 5

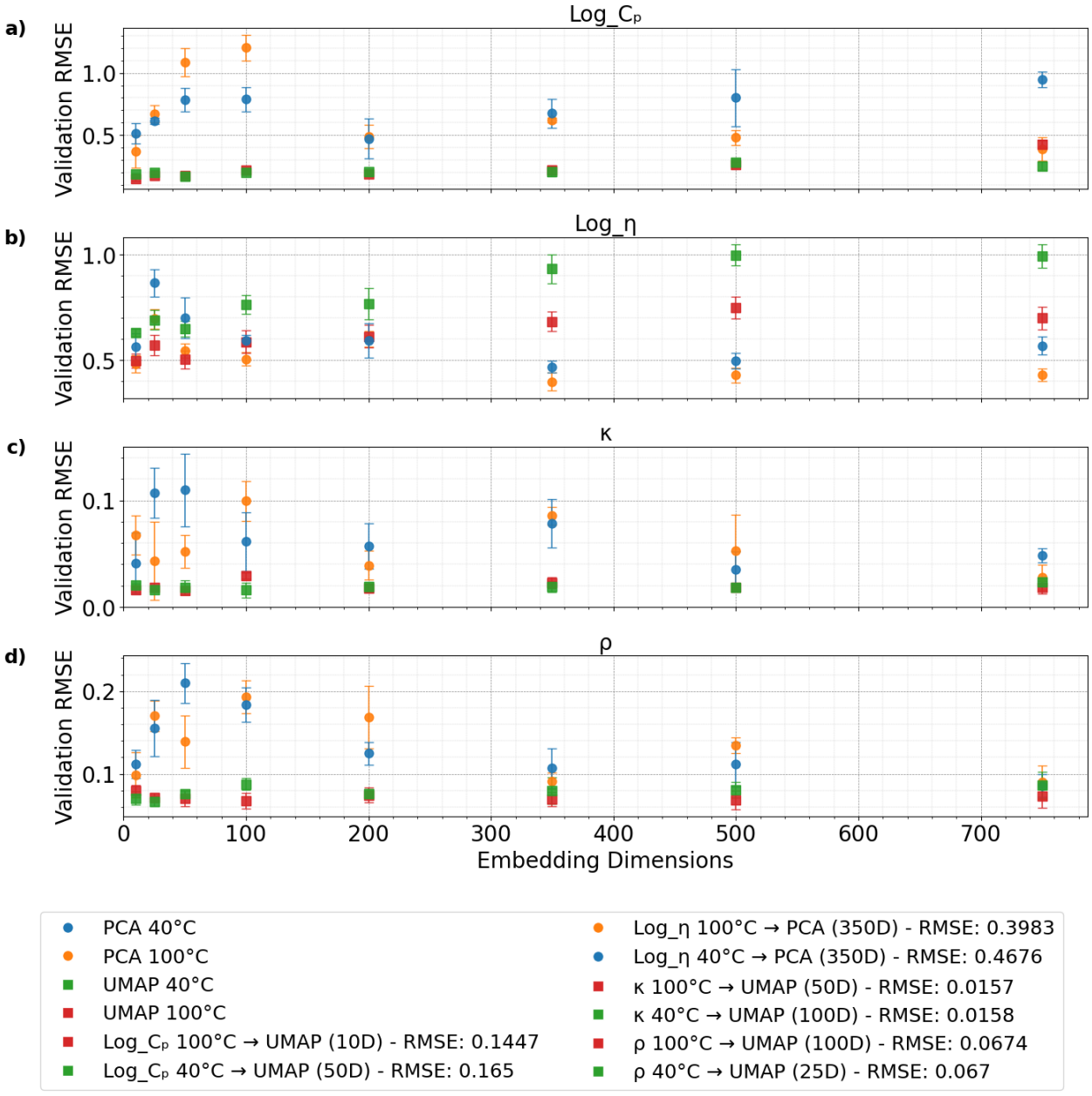


Figure C.1: Effect of different embedding methods and dimensionality on MLP prediction of hydrocarbon thermophysical properties. Best method-dimension combination for each combination shown in legend. Error bars show 1 standard deviation from mean from 5-fold cross validation.

Model	Viscosity 40°C	Viscosity 100°C	Density 40°C	Dens. 100°C	T. Cond. 40°C	T. Cond. 100°C	Heat Cap. 40°C	Heat Cap. 100°C	Rank
XGBoost + Descriptors	0.314	0.272	0.045	0.045	0.009	0.009	0.028	0.028	1
Random Forest + Descriptors	0.339	0.289	0.045	0.045	0.01	0.01	0.03	0.028	2
Transformer + Embeddings	0.319	0.275	0.045	0.045	0.017	0.017	0.055	0.055	3
XGBoost + Embedding	0.375	0.322	0.048	0.047	0.011	0.011	0.052	0.049	4
XGBoost + Fingerprints	0.351	0.388	0.034	0.035	0.009	0.009	0.112	0.113	5
CNN + SMILES	0.382	0.297	0.055	0.055	0.02	0.017	0.063	0.063	6
MLP + Embeddings	0.374	0.318	0.089	0.077	0.014	0.014	0.114	0.084	7
SVM + Embedding	0.346	0.288	0.07	0.068	0.036	0.032	0.1	0.1	8
Transformer + SMILES	0.355	0.326	0.055	0.055	0.02	0.017	0.055	0.071	9
SVM + Fingerprints	0.538	0.417	0.077	0.075	0.033	0.03	0.142	0.141	10
KNN + Descriptors	0.479	0.354	0.063	0.063	0.02	0.017	0.063	0.054	11
Transformer + SELFIES	0.458	0.383	0.055	0.055	0.02	0.024	0.084	0.063	12
ChemBerta + Embedding	0.43	0.327	0.077	0.069	0.029	0.028	0.097	0.121	13
MLP + SMILES	0.483	0.346	0.054	0.054	0.022	0.026	0.077	0.077	14
Random Forest + Embedding	0.544	0.413	0.056	0.055	0.015	0.015	0.15	0.15	15
KNN + Embedding	0.519	0.409	0.06	0.058	0.015	0.014	0.117	0.116	16
MLP + Fingerprints	0.435	0.413	0.122	0.105	0.016	0.0141	0.253	0.22	17
CNN + SELFIES	0.547	0.435	0.055	0.055	0.014	0.014	0.158	0.148	18
MLP + SELFIES	0.5	0.422	0.054	0.054	0.022	0.02	0.383	0.37	19
Random Forest + Fingerprints	0.665	0.505	0.052	0.051	0.015	0.014	0.17	0.171	20
ChemBerta + SMILES	0.356	0.319	0.184	0.1	0.11	0.077	0.479	0.536	21
SVM + Molecular Descriptors	0.854	0.616	0.1	0.095	0.032	0.03	0.315	0.308	22
MPNN + 3D Graph	0.637	0.495	0.055	0.089	0.028	0.026	0.29	0.29	23
GAT + 3D Graph	0.794	0.48	0.077	0.071	0.02	0.02	0.195	0.182	24
KNN + Fingerprints	1.07	0.769	0.089	0.089	0.02	0.017	0.263	0.26	25

Table C.1: Test RMSE values for different ML models across molecular properties. Top 3 performers per property are highlighted. Overall rank is shown in the final column.

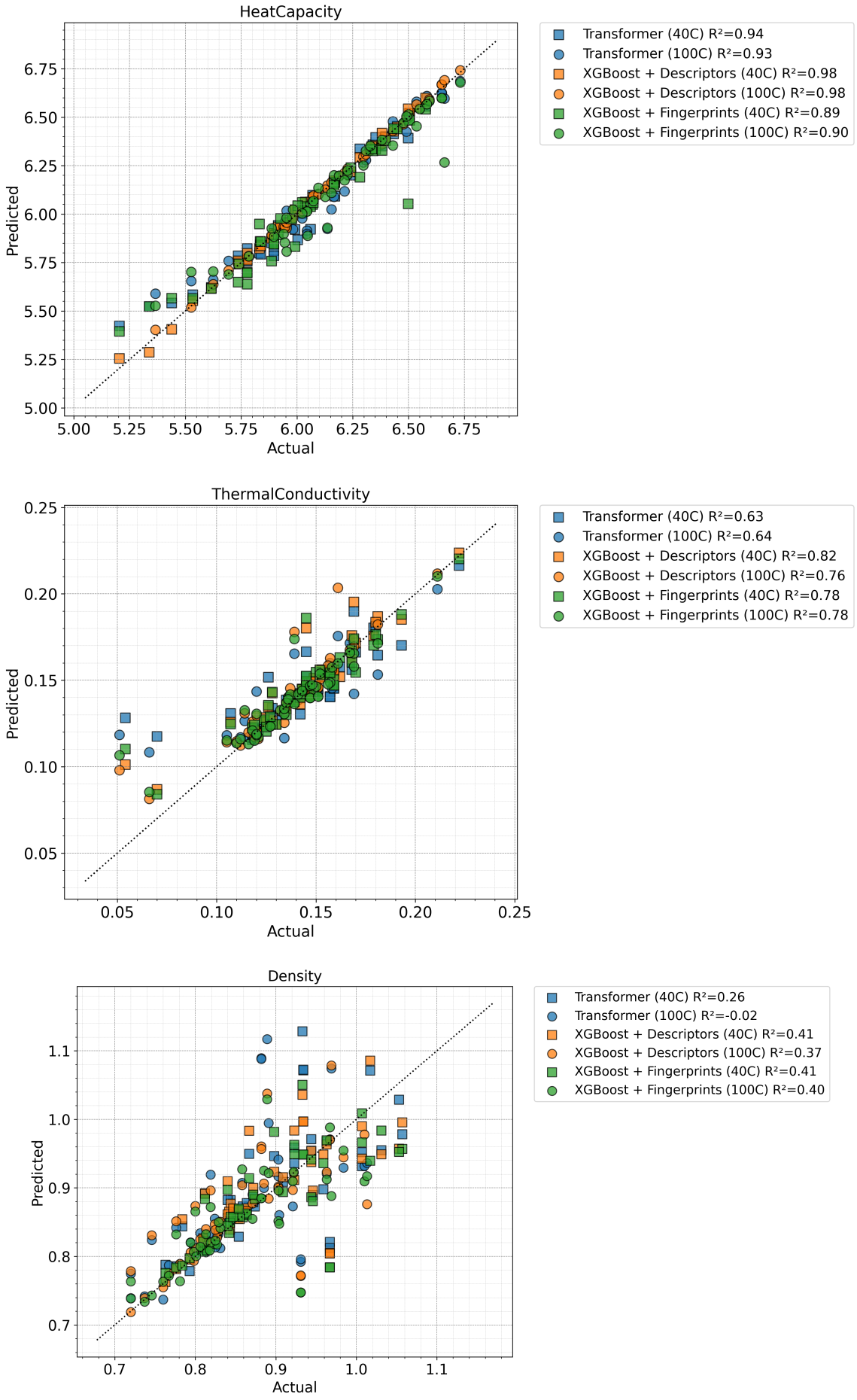


Figure C.2: Parity plots comparing actual vs. predicted values for density, heat capacity and thermal conductivity across three models: Transformer, XGBoost with molecular descriptors, and XGBoost with fingerprints.

Name	Test Property	Viscosity 40°C	Viscosity 100°C	Viscosity Prediction
Tri-Cresyl Phosphate	Different Atom	30.8	4.2	ViscosityL
diheptylamine	Different Atom	1.8	0.9	Landolt-Bornstein
triisopentylamine	Different Atom	1.7	0.7	Landolt-Bornstein
trihexylamine	Different Atom	2.8	1.8	Landolt-Bornstein
triheptylamine	Different Atom	4.0	1.3	Landolt-Bornstein
trioctylamine	Different Atom	6.0	1.7	Landolt-Bornstein
N-butyldodecanamide	Different Atom	26.0	6.7	ViscosityL
N-cyclohexyllinoleamide	Different Atom	110.0	22.0	PPDS 9
N,N-dimethyloleamide	Different Atom	11.0	3.5	ViscosityL
N-isobutyl-3,3-dipropylheptanamide	Different Atom	41.0	11.0	PPDS 9
N,N-dihexyloctanamide	Different Atom	11.0	3.4	ViscosityL
diisobutyl disulfide	Different Atom	1.6	0.8	PPDS 9
octyl disulfide	Different Atom	3.6	1.6	ViscosityL
didecyl disulfide	Different Atom	5.5	2.2	ViscosityL
diphenyl sulfide	Different Atom	10.0	2.8	ViscosityL
n-dodecyl sulfide	Different Atom	5.4	2.1	ViscosityL
1,6-bis(2-ethylhexylthio)hexane	Different Atom	14.0	4.4	ViscosityL
4,8-dithiadodecane	Different Atom	1.9	1.0	PPDS 9
4-thiahenicosane	Different Atom	3.6	1.6	ViscosityL
4-thiahexadecane	Different Atom	2.1	1.0	PPDS 9

Table C.2: OOD test data with different atoms present in structures compared to training data.

Name	Test Property	Viscosity 40°C	Viscosity 100°C	Viscosity Prediction
pentaerythritol tetrakis-(2-ethylbutanoate)	High Viscosity	61.4	6.5	Landolt-Bornstein
12-benzoyloxy-octadecanoic acid phenyl ester	High Viscosity	304.6	29.4	Landolt-Bornstein
pentaerythritol dihexanoate dicyclohexanecarboxylate	High Viscosity	99.9	10.0	Landolt-Bornstein
9,10,12,13-tetrakis-(propionyloxy)-octadecanoic acid methyl ester	High Viscosity	72.0	7.7	Landolt-Bornstein
6-(1-butyl-octadecyl)-1,2,3,4-tetrahydro-naphthalene	High Viscosity	51.9	7.1	Landolt-Bornstein
dipentaerythritol hexabutanoate	High Viscosity	64.0	8.8	Landolt-Bornstein
4-(1-butyl-octadecyl)-bicyclohexyl	High Viscosity	76.1	9.0	Landolt-Bornstein
5-biphenyl-4-yl-hexacos-4-ene	High Viscosity	66.0	9.0	Landolt-Bornstein
1-naphthalen-2-yl-1,2-dibutyl-eicosane	High Viscosity	69.0	8.6	Landolt-Bornstein
ethylene glycol bis-(12-acetoxy-stearate)	High Viscosity	84.0	12.0	Landolt-Bornstein
pentanedioic acid bis-(2,2,3,3,4,4,5,5-octafluoropentyl) ester	Different Atom	13.5	2.2	Landolt-Bornstein
2,2,4,15,17,17-hexamethyl-7,12-bis-(3,5,5-trimethylhexyl)-octadecane	High Viscosity	104.0	8.3	Landolt-Bornstein
propylene glycol bis-(12-acetoxy-stearate)	High Viscosity	85.0	12.0	Landolt-Bornstein
trimethylene glycol bis-(12-acetoxy-stearate)	High Viscosity	90.0	13.0	Landolt-Bornstein
diethylene glycol bis-(12-acetoxy-stearate)	High Viscosity	93.0	13.0	Landolt-Bornstein
7,16-dimethyl-9,11,12,14-tetrahexyl-docosane	High Viscosity	80.0	8.0	Landolt-Bornstein
1,10-bis-(naphth-1-yl)-decane	High Viscosity	226.0	11.0	Landolt-Bornstein
1,10-bis-(decahydro-naphth-1-yl)-decane	High Viscosity	145.0	10.9	Landolt-Bornstein
1-cyclohexyl-4-decahydro-naphth-1-yl-tetradecane	High Viscosity	109.0	8.0	Landolt-Bornstein
5-(1-phenethyl-tetradecyl)-1,2,3,4-tetrahydro-naphthalene	High Viscosity	63.1	16.4	Landolt-Bornstein

Table C.3: OOD test data with higher viscosities compared to training data.

Molecule	Temp (K)	Exp.	LOPLS	COMPASS	SCIPCFF	AMBER	ML
13-Cyclohexylpentacosane	313	0.157	0.245 ± 0.043	0.139 ± 0.024	0.169 ± 0.030	0.168 ± 0.029	0.156 ± 0.009
	373	0.148	0.189 ± 0.033	0.116 ± 0.020	0.126 ± 0.022	0.139 ± 0.024	0.145 ± 0.009
1,1 Diphenyl Dodecane	313	0.180	0.156 ± 0.027	0.101 ± 0.018	0.103 ± 0.018	0.133 ± 0.023	0.169 ± 0.009
	373	0.169	0.166 ± 0.029	0.102 ± 0.018	0.157 ± 0.028	0.114 ± 0.020	0.158 ± 0.009
1-cyclohexyl-2-(cyclohexyl methyl) heptadecane	313	0.135	0.188 ± 0.033	0.201 ± 0.035	0.167 ± 0.029	0.218 ± 0.038	0.137 ± 0.009
	373	0.128	0.134 ± 0.023	0.205 ± 0.036	0.137 ± 0.024	0.149 ± 0.026	0.131 ± 0.009
1-Cyclohexylnonane	313	0.136	0.148 ± 0.026	0.249 ± 0.044	0.207 ± 0.036	0.154 ± 0.027	0.134 ± 0.009
	373	0.128	0.154 ± 0.027	0.068 ± 0.012	0.188 ± 0.033	0.188 ± 0.033	0.125 ± 0.009
2-(1-propylheptyl)naphthalene	313	0.210	0.126 ± 0.022	0.088 ± 0.015	0.136 ± 0.024	0.164 ± 0.029	0.207 ± 0.009
	373	0.199	0.184 ± 0.032	0.107 ± 0.019	0.097 ± 0.017	0.121 ± 0.021	0.194 ± 0.009
2-butyl-3-hexyldecahydro-naphthalene	313	0.125	0.217 ± 0.038	0.148 ± 0.026	0.132 ± 0.023	0.138 ± 0.024	0.125 ± 0.009
	373	0.118	0.151 ± 0.026	0.192 ± 0.034	0.169 ± 0.030	0.181 ± 0.032	0.118 ± 0.009
Dioctyl Sebacate	313	0.183	0.148 ± 0.026	0.171 ± 0.030	0.178 ± 0.031	0.148 ± 0.026	0.179 ± 0.009
	373	0.173	0.125 ± 0.022	0.148 ± 0.026	0.170 ± 0.030	0.145 ± 0.025	0.168 ± 0.009
Hexadecane	313	0.143	0.166 ± 0.029	0.164 ± 0.029	0.139 ± 0.024	0.153 ± 0.027	0.145 ± 0.009
	373	0.132	0.134 ± 0.023	0.161 ± 0.028	0.111 ± 0.019	0.145 ± 0.025	0.134 ± 0.009
Squalane	313	0.123	0.128 ± 0.022	0.143 ± 0.025	0.139 ± 0.024	0.175 ± 0.031	0.123 ± 0.009
	373	0.115	0.127 ± 0.022	0.109 ± 0.019	0.129 ± 0.023	0.169 ± 0.030	0.117 ± 0.009

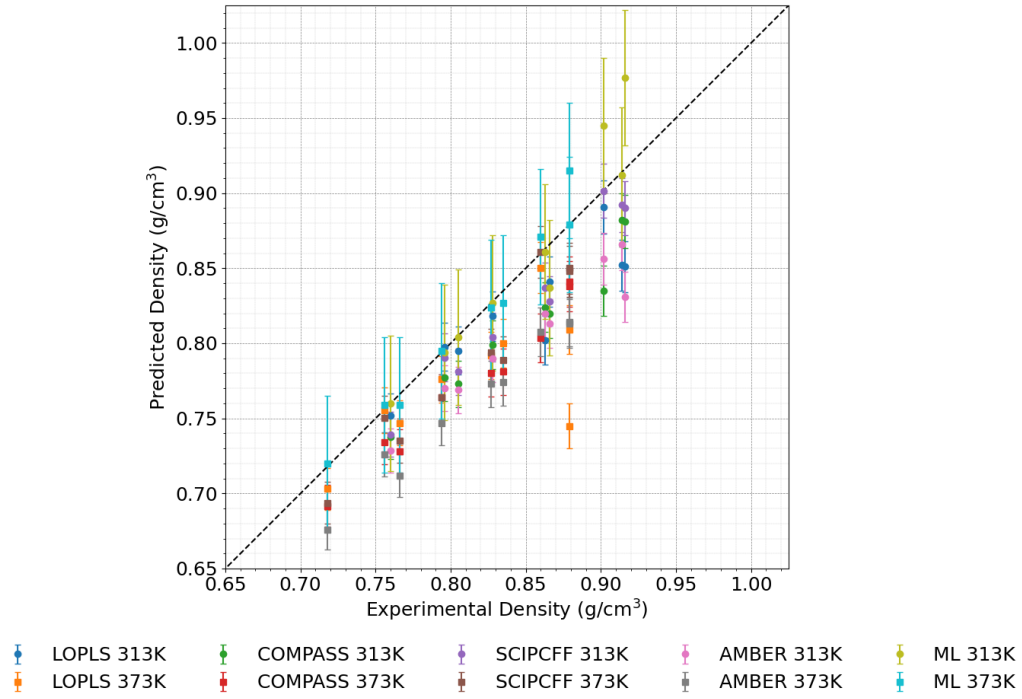
Table C.4: MD/ML prediction on test data for thermal conductivity values.

Molecule	Temp (K)	Exp.	LOPLS	COMPASS	SCIPCFF	AMBER	ML
13-Cyclohexylpentacosane	313	18.650	10.140 ± 1.268	16.521 ± 2.065	13.318 ± 1.665	28.258 ± 3.532	19.410 ± 1.369
	373	3.630	2.932 ± 0.366	2.485 ± 0.311	3.095 ± 0.387	3.874 ± 0.484	3.680 ± 1.313
1,1 Diphenyl Dodecane	313	14.500	25.693 ± 3.212	6.889 ± 0.861	12.090 ± 1.511	7.618 ± 0.952	18.630 ± 1.369
	373	4.000	3.584 ± 0.448	1.975 ± 0.247	1.701 ± 0.213	1.869 ± 0.234	3.120 ± 1.313
1-cyclohexyl-2-(cyclohexyl methyl) heptadecane	313	29.000	24.016 ± 3.002	21.288 ± 2.661	28.678 ± 3.585	47.364 ± 5.921	28.900 ± 1.369
	373	8.400	4.279 ± 0.535	3.874 ± 0.484	3.527 ± 0.441	7.526 ± 0.941	5.690 ± 1.313
1-Cyclohexylnonane	313	2.680	3.373 ± 0.422	2.111 ± 0.264	1.905 ± 0.238	1.629 ± 0.204	2.690 ± 1.369
	373	1.000	1.054 ± 0.132	0.712 ± 0.089	0.515 ± 0.064	0.915 ± 0.114	1.020 ± 1.313
2-(1-propylheptyl)naphthalene	313	13.344	11.645 ± 1.456	6.586 ± 0.823	7.381 ± 0.923	5.877 ± 0.735	14.150 ± 1.369
	373	2.224	2.827 ± 0.353	1.390 ± 0.174	0.679 ± 0.085	1.884 ± 0.236	2.450 ± 1.313
2-butyl-3-hexyldecahydro-naphthalene	313	10.295	9.251 ± 1.156	5.596 ± 0.700	9.870 ± 1.234	20.226 ± 2.528	14.168 ± 1.369
	373	1.907	2.520 ± 0.315	1.751 ± 0.219	1.408 ± 0.176	2.946 ± 0.368	2.440 ± 1.313
Dioctyl Sebacate	313	14.110	18.354 ± 3.544	13.353 ± 1.669	18.827 ± 3.728	13.193 ± 1.649	12.760 ± 1.369
	373	4.362	6.245 ± 0.781	4.054 ± 0.507	6.079 ± 0.760	3.329 ± 0.416	3.590 ± 1.313
Hexadecane	313	2.200	2.135 ± 0.267	1.377 ± 0.172	1.957 ± 0.245	2.148 ± 0.269	2.310 ± 1.369
	373	0.900	0.777 ± 0.097	0.506 ± 0.063	0.550 ± 0.069	0.790 ± 0.099	0.900 ± 1.313
Squalane	313	15.060	9.443 ± 1.180	3.245 ± 0.406	14.486 ± 1.811	23.795 ± 2.974	14.960 ± 1.369
	373	3.120	3.224 ± 0.403	4.060 ± 0.507	2.913 ± 0.364	4.789 ± 0.599	3.130 ± 1.313

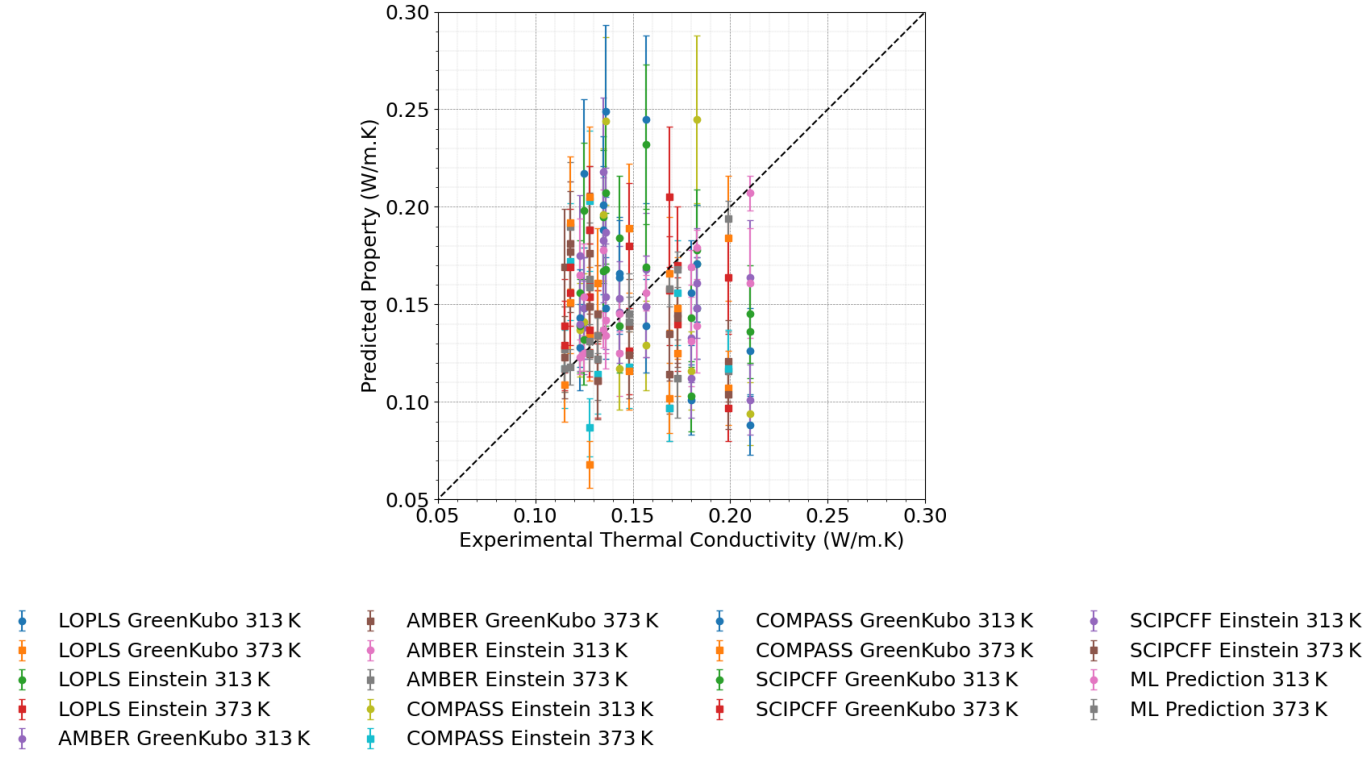
Table C.5: MD/ML prediction on test data for viscosity values.

Molecule	Temp (K)	Experimental	LOPLS	COMPASS	SCIPCFF	AMBER	ML
13-Cyclohexylpentacosane	313	0.828	0.818 ± 0.016	0.799 ± 0.016	0.804 ± 0.016	0.790 ± 0.016	0.827 ± 0.045
	373	0.794	0.776 ± 0.016	0.764 ± 0.015	0.764 ± 0.015	0.747 ± 0.015	0.795 ± 0.045
1,1 Diphenyl Dodecane	313	0.916	0.851 ± 0.017	0.881 ± 0.018	0.890 ± 0.018	0.831 ± 0.017	0.977 ± 0.045
	373	0.879	0.745 ± 0.015	0.838 ± 0.017	0.848 ± 0.017	0.813 ± 0.016	0.915 ± 0.045
1-cyclohexyl-2-(cyclohexyl methyl) heptadecane	313	0.866	0.841 ± 0.017	0.820 ± 0.016	0.828 ± 0.017	0.813 ± 0.016	0.837 ± 0.045
	373	0.835	0.800 ± 0.016	0.781 ± 0.016	0.789 ± 0.016	0.774 ± 0.015	0.827 ± 0.045
1-Cyclohexylnonane	313	0.805	0.795 ± 0.016	0.773 ± 0.015	0.781 ± 0.016	0.769 ± 0.015	0.804 ± 0.045
	373	0.766	0.747 ± 0.015	0.728 ± 0.015	0.735 ± 0.015	0.712 ± 0.014	0.759 ± 0.045
2-(1-propylheptyl)naphthalene	313	0.914	0.852 ± 0.017	0.882 ± 0.018	0.892 ± 0.018	0.866 ± 0.017	0.912 ± 0.045
	373	0.879	0.809 ± 0.016	0.841 ± 0.017	0.850 ± 0.017	0.814 ± 0.016	0.879 ± 0.045
2-butyl-3-hexyldecahydro-naphthalene	313	0.863	0.802 ± 0.016	0.824 ± 0.016	0.837 ± 0.017	0.820 ± 0.016	0.861 ± 0.045
	373	0.827	0.792 ± 0.016	0.780 ± 0.016	0.794 ± 0.016	0.773 ± 0.015	0.824 ± 0.045
Dioctyl Sebacate	313	0.902	0.891 ± 0.018	0.835 ± 0.017	0.902 ± 0.018	0.856 ± 0.017	0.945 ± 0.045
	373	0.860	0.850 ± 0.017	0.803 ± 0.016	0.861 ± 0.017	0.808 ± 0.016	0.871 ± 0.045
Hexadecane	313	0.760	0.752 ± 0.015	0.738 ± 0.015	0.739 ± 0.015	0.728 ± 0.015	0.760 ± 0.045
	373	0.718	0.703 ± 0.014	0.692 ± 0.014	0.694 ± 0.014	0.676 ± 0.014	0.720 ± 0.045
Squalane	313	0.796	0.798 ± 0.016	0.777 ± 0.016	0.791 ± 0.016	0.770 ± 0.015	0.794 ± 0.045
	373	0.756	0.755 ± 0.015	0.734 ± 0.015	0.750 ± 0.015	0.726 ± 0.015	0.759 ± 0.045

Table C.6: MD/ML prediction on test data for density values.



(a) Predicted vs. experimental density values at 313 K and 373 K.



(b) Predicted vs. experimental thermal conductivity values at 313 K and 373 K.

Figure C.3: Comparison of MD and ML predictions against experiments. (a) Density predictions and (b) Thermal-conductivity predictions (Green–Kubo formalism). Circle markers: 313 K; square markers: 373 K; error bars show one-sigma uncertainties.

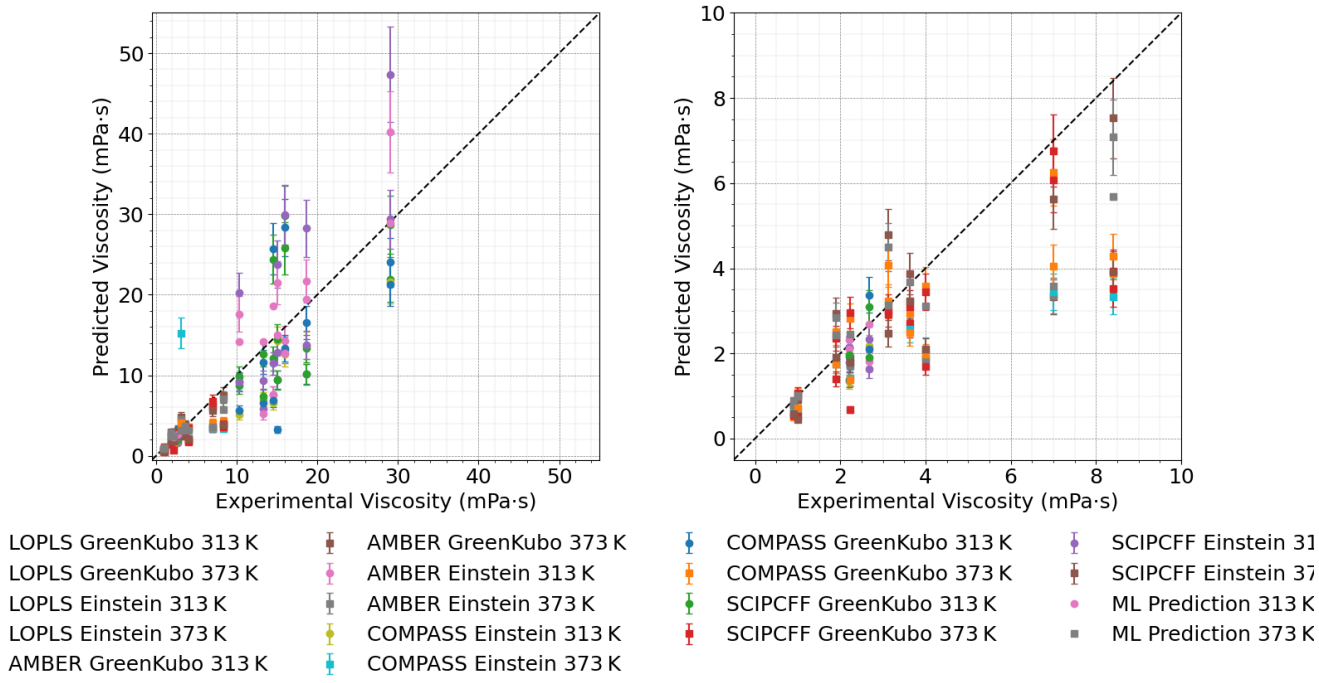


Figure C.4: Predicted vs. experimental viscosities from molecular-dynamics (Green–Kubo) and machine-learning models at 313 K (circle markers) and 373 K (square markers). Left: full range; right: zoom into 0–10 mPa.s. Error bars show one-sigma uncertainties.

Appendix D

Supporting information for Chapter 6

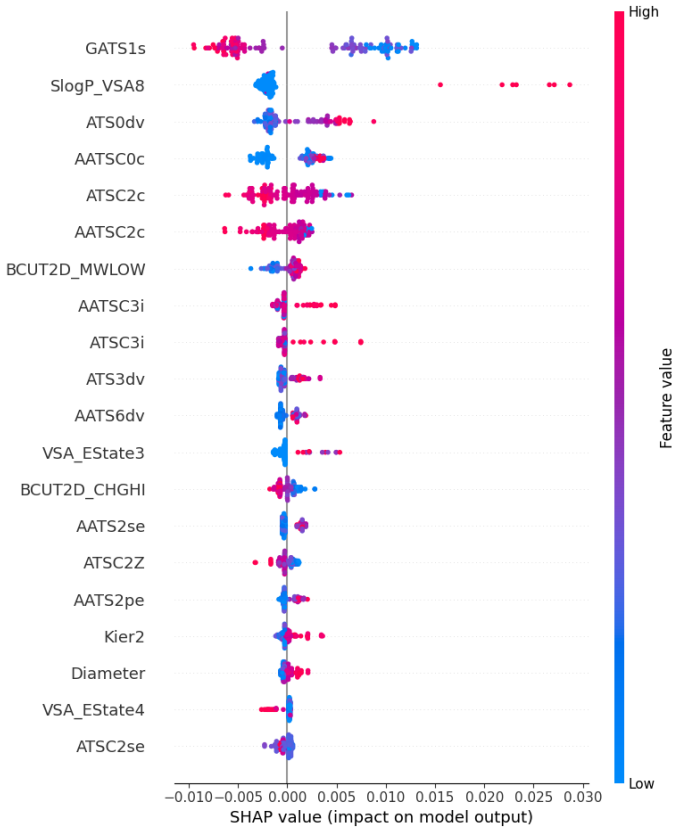


Figure D.1: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule thermal conductivity at 100°C.

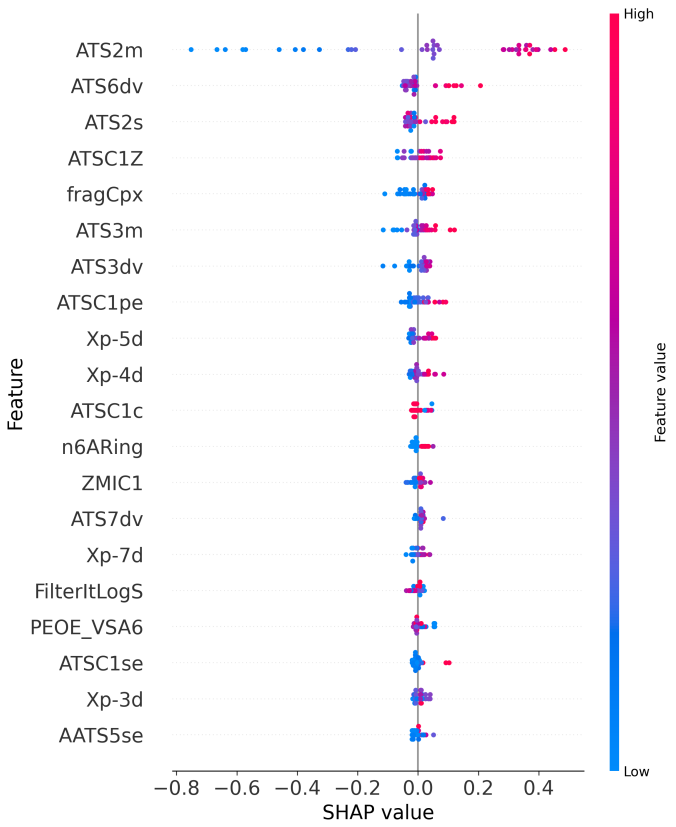


Figure D.2: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log viscosity at 100°C.

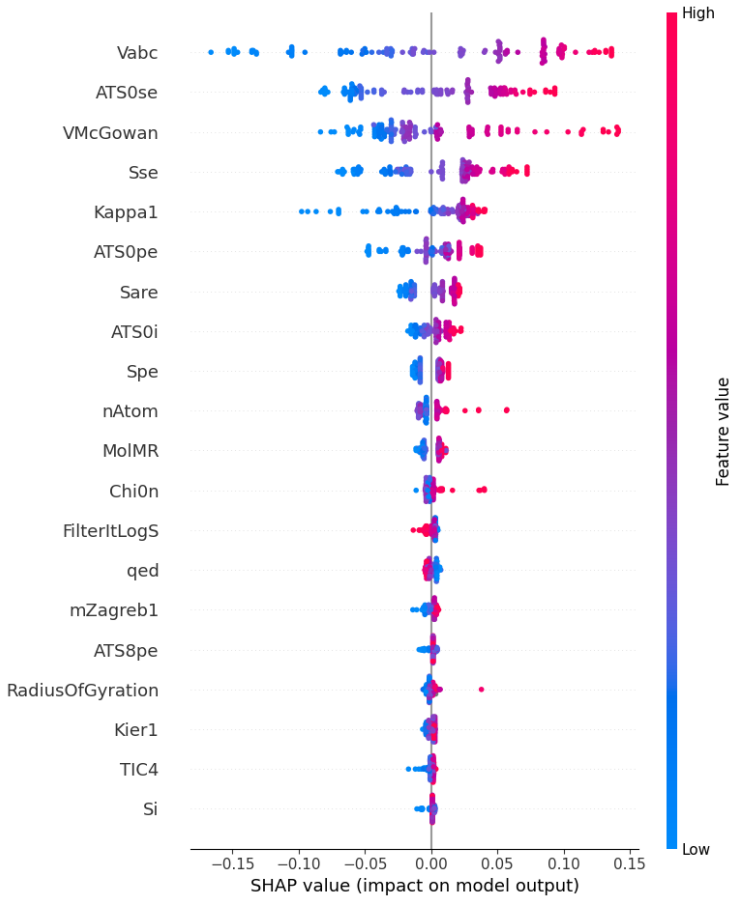


Figure D.3: Beeswarm plot showing the top 20 most influential molecular descriptors in predicting molecule log heat capacity at 100°C.

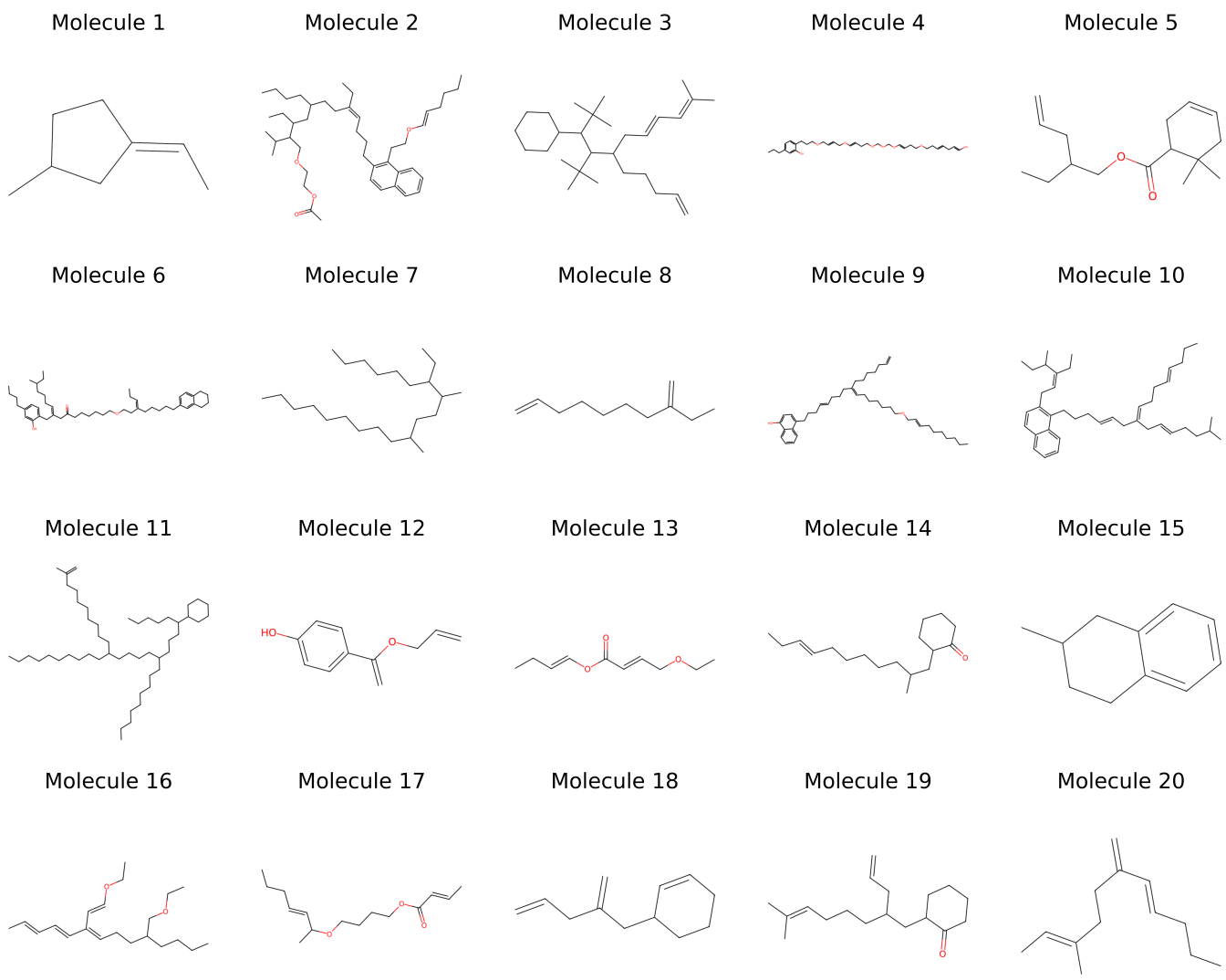


Figure D.4: Grid of top 20 molecules by crowding distance from best hyperparemeter combination using NSGA-II optimisation.

Molecule ID	ρ (100°C)	ρ (40°C)	η (40°C)	η (100°C)	κ (40°C)	κ (100°C)	Cp (40°C)	Cp (100°C)	DVI	Toxicity	SCScore
Molecule 1	0.7683	0.8278	0.566	0.408	0.1266	0.1171	184.145	219.298	0.0000	0.4689	0.4967
Molecule 2	0.9745	0.9891	99.725	12.185	0.1817	0.1707	864.808	1008.167	196.7946	0.3760	0.8173
Molecule 3	0.8129	0.8726	21.583	4.597	0.0834	0.0719	636.674	754.571	186.9606	0.4515	0.5883
Molecule 4	1.0111	0.9753	33.081	6.828	0.2212	0.2094	763.642	890.489	195.2466	0.9481	0.8182
Molecule 5	0.9868	0.9770	4.297	1.404	0.1454	0.1430	365.509	423.348	0.0000	0.5100	0.6032
Molecule 6	0.9728	0.9946	36.810	7.920	0.1772	0.1801	874.490	1007.114	197.8249	0.6335	0.9998
Molecule 7	0.7912	0.8188	7.323	1.939	0.1324	0.1247	581.736	675.171	79.4155	0.0909	0.7576
Molecule 8	0.7092	0.7416	0.521	0.280	0.1251	0.1197	248.002	289.769	0.0000	0.6208	0.4649
Molecule 9	0.9823	1.0199	44.428	11.875	0.2450	0.2330	862.761	1004.328	203.4480	0.8496	0.8015
Molecule 10	0.9379	0.9556	28.403	10.308	0.2127	0.2096	807.198	965.669	205.2706	0.6146	0.9489
Molecule 11	0.8039	0.8577	42.310	5.839	0.1283	0.1273	966.805	1090.595	185.6408	0.3326	0.6965
Molecule 12	0.9876	1.0075	11.720	2.107	0.1576	0.1525	221.164	253.346	67.2955	0.9165	0.3430
Molecule 13	0.9028	0.9174	3.767	1.114	0.1755	0.1582	244.969	287.368	0.0000	0.7129	0.3882
Molecule 14	0.8785	0.8962	5.353	1.702	0.1545	0.1428	398.436	470.428	16.6552	0.4433	0.6443
Molecule 15	0.8847	0.9267	1.812	0.683	0.1255	0.1161	188.440	220.728	0.0000	0.5846	0.5407
Molecule 16	0.9799	0.9832	6.537	1.824	0.1767	0.1737	534.738	581.828	49.3669	0.6601	0.5881
Molecule 17	0.9673	1.0267	6.242	1.775	0.1731	0.1587	364.277	425.460	30.2100	0.6213	0.4935
Molecule 18	0.8386	0.8924	1.185	0.638	0.1346	0.1231	236.727	269.316	0.0000	0.7042	0.4753
Molecule 19	0.9112	0.9484	9.510	2.128	0.1526	0.1391	379.258	447.803	98.9571	0.4967	0.5098
Molecule 20	0.7565	0.8205	0.577	0.389	0.1277	0.1179	292.652	327.237	0.0000	0.7021	0.5246

Table D.1: Molecular properties of top 20 molecules by crowding distance from best hyperparameter combination using NSGA-II optimisation.

Molecule ID	Rank	ρ (100°C)	ρ (40°C)	η (40°C)	η (100°C)	κ (40°C)	κ (100°C)	Cp (40°C)	Cp (100°C)	DVI	Toxicity	SCScore
Molecule 15	1	0.88	0.91	23.24	5.58	0.193	0.185	957.72	1083.24	190.64	0.17	0.69
Molecule 26	2	0.77	0.80	19.79	4.78	0.177	0.170	958.33	1102.57	186.60	0.11	0.85
Molecule 3	3	0.87	0.95	26.30	6.04	0.170	0.164	964.32	1089.77	191.63	0.10	0.83
Molecule 30	4	0.88	0.96	20.67	4.97	0.188	0.179	857.59	989.10	187.69	0.13	0.77
Molecule 28	5	0.77	0.77	22.25	4.68	0.164	0.160	958.68	1102.13	182.93	0.11	0.86
Molecule 22	6	0.94	0.97	32.69	7.77	0.179	0.174	958.58	1086.23	196.37	0.24	0.67
Molecule 27	6	0.79	0.81	13.98	3.97	0.181	0.175	847.49	987.04	183.77	0.11	0.85
Molecule 2	8	0.93	0.96	30.78	6.22	0.160	0.155	962.43	1090.76	190.27	0.12	0.84
Molecule 29	9	0.95	0.93	32.72	6.54	0.183	0.177	878.67	1023.09	191.12	0.16	0.86
Molecule 1	10	0.76	0.79	30.11	5.39	0.151	0.148	959.78	1099.37	184.39	0.09	0.88
Dioctyl Sebacate	11	0.86	0.90	14.01	3.15	0.183	0.173	655.00	755.00	174.14	0.24	0.44
Molecule 23	12	1.01	1.00	41.60	8.33	0.174	0.174	960.39	1088.23	195.24	0.18	0.90
Molecule 12	13	1.00	1.01	40.16	6.73	0.179	0.177	962.93	1090.19	189.11	0.18	0.98
Molecule 9	13	0.96	0.95	46.83	7.83	0.185	0.176	875.96	1010.32	192.05	0.17	0.83
Molecule 16	15	0.79	0.79	33.59	5.46	0.151	0.148	959.71	1102.22	182.90	0.09	0.88
Molecule 24	15	1.00	0.99	31.43	7.64	0.170	0.169	961.02	1088.01	196.42	0.29	0.90
Molecule 6	17	0.77	0.79	37.43	7.09	0.134	0.135	963.73	1098.33	191.93	0.08	0.86
Diisodecyl Adipate	17	1.01	1.05	13.46	3.49	0.177	0.168	655.00	755.00	172.06	0.19	0.56
Molecule 13	19	0.95	0.94	38.68	7.88	0.168	0.169	877.70	1024.70	194.65	0.14	0.83
Molecule 11	20	0.78	0.82	35.92	5.82	0.156	0.155	963.54	1091.80	184.82	0.15	0.99
Molecule 10	21	0.90	0.93	35.51	6.06	0.167	0.163	899.80	1049.85	186.89	0.20	0.86
Molecule 18	22	0.95	0.97	30.16	8.16	0.170	0.170	878.36	1023.10	198.52	0.28	0.96
Molecule 25	23	0.97	0.97	55.12	9.06	0.199	0.184	861.46	1010.77	194.14	0.20	0.90
Squalane	24	0.76	0.80	15.06	3.12	0.123	0.115	727.00	847.00	166.58	0.07	0.66
Molecule 5	25	0.81	0.85	23.72	4.58	0.158	0.157	858.29	1000.53	180.09	0.14	0.90
Molecule 21	25	0.79	0.82	36.60	6.80	0.144	0.139	963.06	1091.93	190.85	0.16	0.95
Molecule 7	27	0.99	0.95	53.40	8.70	0.184	0.177	869.30	1020.40	193.44	0.32	0.90
Molecule 8	28	0.94	0.94	33.42	7.08	0.156	0.155	878.56	1028.89	193.43	0.18	0.89
Molecule 4	29	0.78	0.78	29.78	4.62	0.143	0.143	845.53	978.85	175.38	0.09	0.84
Molecule 14	30	0.83	0.87	31.75	6.03	0.165	0.161	956.85	1087.23	188.53	0.37	0.97
Molecule 17	31	0.95	0.95	25.73	6.07	0.161	0.154	853.10	1013.35	192.14	0.24	0.93
Hexadecane	32	0.72	0.76	2.20	0.79	0.143	0.132	384.20	444.60	0.00	0.19	0.42
PAO 4	33	0.76	0.80	13.32	3.35	0.145	0.136	772.78	845.56	167.87	0.35	0.58
Molecule 20	34	0.95	0.94	33.36	6.41	0.167	0.168	864.06	1008.71	190.12	0.24	0.94
Molecule 19	35	0.84	0.86	35.86	9.29	0.134	0.137	961.96	1092.01	199.40	0.27	1.00

Table D.2: Ranked properties of top 30 WSGA molecules with gold, silver, and bronze highlighting best performing molecules for that property.