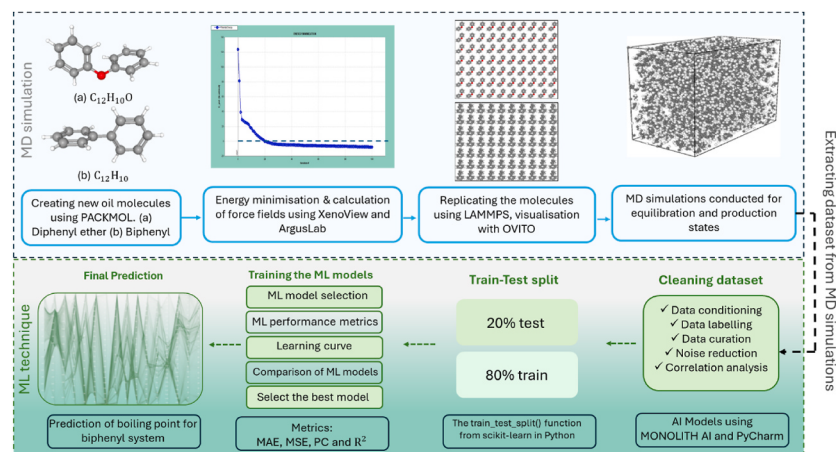


A hybrid molecular dynamics–machine learning framework for boiling point estimation in aromatic fluids

Amirali Shateri, Zhiyin Yang, Nasser Sherkat, Jianfei Xie^{*}

School of Engineering, University of Derby, DE22 3AW, UK

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Machine learning
Molecular dynamics simulation
Boiling point
Thermal fluids
Phase change

ABSTRACT

Precise estimation of boiling points in organic fluids is critical for designing efficient and safe thermal systems. This study presents a hybrid molecular dynamic (MD)–machine learning (ML) framework for boiling point estimation in two representative aromatic fluids: biphenyl ($C_{12}H_{10}$) and diphenyl ether ($C_{12}H_{10}O$). Two force fields, OPLS-AA and COMPASS, were tested in equilibrium MD simulations. OPLS-AA produced density predictions with a relative error below 2 % compared to experimental values, while COMPASS showed reduced accuracy at elevated temperatures. Boiling point was estimated using a density threshold method (yielding 525.66 K) and

^{*} Corresponding author.

E-mail address: j.xie@derby.ac.uk (J. Xie).

a thermodynamically rigorous inflection-point method (508.18 K), revealing $\sim 3.3\%$ deviation between boiling onset and completion. MD data were used to train and evaluate three regression models—Nearest Neighbours Regression (NNR), Neural Network (NN), and Support Vector Regression (SVR). The NNR model achieved the best match with MD data, predicting a boiling point of 524.97 K and density of 0.064 g/cm^3 . The NN model accurately estimated boiling temperature (525.3 K) but overestimated density, while SVR underestimated both. This work contributes a novel, interpretable MD–ML framework to integrate the inflection-point detection with data-driven model selection, offering a reproducible and accurate method for boiling point estimation that can be extended to other organic thermal systems.

1. Introduction

Accurate predictions of thermophysical properties of fluids are essential to design efficient and sustainable thermal fluid systems for many industrial applications, e.g., chemical processing and thermal management. The chemical makeup of heat transfer fluids (HTFs) along with their operating conditions determines their thermal stability and heat transfer efficiency. Previous studies show that thermal fluids degrade because of thermal and chemical stresses during normal operations, underlining the importance of fluid chemical decomposition and environmental conditions for long-term performance [1,2]. Rigorous system design practices are essential to maximize the service life of HTFs [3]. The design of heat exchangers and heaters demands that engineers meticulously evaluate the heat flux control, temperature gradients, and thermal boundary conditions, with specific attention to energy recovery units [4]. In scenarios where flame impingement is present localized overheating often happens and raises the risk of premature thermal degradation [5]. Proper fuel-to-air mixing becomes essential because incorrect mixing leads to changes in flame dynamics which amplify thermal stresses [6,7]. The stability of HTF fluids is seriously compromised by chemical contamination because contaminants exhibit poor thermal stability, leading to fluid breakdown and surface fouling at high temperatures [8]. Rigorous filtration and conditioning processes are essential for maintaining fluid purity, thus ensuring the efficiency and integrity of the thermal-fluid system [9].

Direct economic and operational benefits arise from precise boiling point predictions by enabling efficient system operation and energy savings while also improving process design for chemical reactors and energy production systems. Hosseinifar and Shahverdi [10] developed a six-point technique to create petroleum fluid distillation curves from minimal experimental data. Their method builds mathematical connections between distillation temperatures and physical properties to enable the prediction of true boiling point and ASTM D86 distillation curves. The model was evaluated against both literature data and newly obtained experimental data, demonstrating strong predictive capability with a mean absolute deviation of 1.98 %. Mukwembi and Nyabadza [11] developed an innovative model to predict alkane boiling points using graph theory. Their results showed that new computational methods surpass traditional single-parameter models by effectively modelling complex thermophysical interactions in organic thermal-fluid systems with potential relevance to biphenyl and diphenyl ether-based mixtures. Stratiev et al. [12] investigated the relationship between boiling points and viscosity in secondary vacuum gas oils, discovering that accurate boiling point predictions enhance viscosity modelling results. These models achieve enhanced predictive accuracy and improved design efficiency in petrochemical systems through the strong link between viscosity and boiling points. Additionally, recent developments in molecular dynamics (MD) simulations have given important new perspectives on interfacial heat transport mechanisms and boiling behaviour of low-boiling-point organic fluids. Su et al. [13] examined how different liquid film thicknesses and solid-liquid interaction forces along with initial temperatures affect the boiling dynamics of 1,1,1,2-tetrafluoroethane (R134a). Their research showed how film thickness determines the boiling mode because thinner films result in thin film boiling while thicker films lead to explosive boiling, emphasizing vapor layer thermal resistance as the main factor that affects the entire heat transfer mechanism. Su et al. [14] investigated how liquid sodium (Na) and R134a behaved during boiling when exposed to high heat flux conditions. MD simulations showed that liquid sodium maintained a lengthier nucleate boiling stage which featured more active bubble movement and achieved substantially higher heat flux than R134a. Furthermore, Lin et al. [15] focused on understanding of how vibration affects phase change dynamics in organic fluids with low boiling points. Their research indicated that vibration amplitude, frequency, and the wettability of surfaces determine the phase transition modes such as diffusive evaporation, nucleate boiling, and film boiling/cavitation. Their work enhances molecular-level knowledge of phase change processes which may lead to industrial cooling and thermal management advancements.

Despite significant progress in classical and computational boiling-point predictions, there remains an identifiable gap: current methods struggle to accurately assess the properties of thermal fluids such as biphenyl and diphenyl ether [16]. Industrial heat transfer applications often use these substances because they maintain thermal stability well while featuring low vapor pressure at high temperatures and good thermal conductivity. As diphenyl ether ($\text{C}_{12}\text{H}_{10}\text{O}$) boils at 528 K and biphenyl ($\text{C}_{12}\text{H}_{10}$) has a boiling range of 524–538 K, their combination in Dowtherm A provides exceptional heat-transfer qualities, low volatility, and enhanced thermal stability, making it ideal for high-temperature industrial applications [17,19]. Nevertheless, these fluids face several challenges, including potential thermal degradation together with environmental persistence and associated toxicity concerns. Boiling-point prediction methodologies must be both accurate and reliable to maintain thermal efficiency and protect environmental and occupational safety within thermal fluid systems that use these substances [18]. Traditional thermodynamic models and empirical correlations provide useful information, but they often fail to accurately predict boiling points near critical conditions or during phase transitions due to their inability to capture complex molecular interactions. In contrast, machine learning (ML) approaches have shown significant promise in improving predictive performance by learning intricate patterns from large, multi-dimensional datasets. Xu et al.

[20] demonstrated that Convolutional Neural Networks (CNNs) can accurately estimate thermophysical properties of organic compounds using 3D molecular structure representations. Similarly, Liu and Nouroddin [21] applied Extreme Learning Machines (ELM) and Ensemble Decision Tree Boosting (EDT Boosted) algorithms to predict refrigerant boiling points, showing ML's capacity to model nonlinearities inherent in molecular thermodynamics. Building on this foundation, Liu et al. [31] constructed a database of potential insulating gases and applied a Gradient Boosting Regression (GBR) model with RDKit descriptors to predict boiling points, achieving a determination coefficient (R^2) of 0.97 on the test set. Their use of SHAP (SHapley Additive exPlanations) further revealed that spatial interaction descriptors (e.g., I_{pc}) played key roles in the prediction, highlighting ML's ability to uncover structure–property relationships.

Parallel research has applied ML to predict a wide range of thermophysical properties in nanofluids and phase change materials (PCMs). Ullah et al. [32] proposed integrating artificial neural networks (ANN) in ML with computational fluid dynamics (CFD) pipelines to overcome limitations of empirical correlations when modelling complex nanofluids. Bhanuteja et al. [33] experimentally measured the thermal conductivity and viscosity of hybrid nanofluids and compared traditional correlations with ML models, demonstrating that ML could reduce prediction error to less than 4 %. Gao et al. [34] used Automatic Relevance Determination (ARD), k -nearest neighbours (KNN), and least absolute shrinkage and selection operator (LASSO) to model the thermal conductivity and latent heat of carbon-based nano-enhanced PCMs across 100 thermal cycles, achieving R^2 values of about 0.999. Similarly, Bakouri et al. [35] implemented multiple regressors, including Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), KNN, and Huber regression, to predict the latent heat and thermal conductivity enhancements in metal-based PCMs with strong statistical performance. Comprehensive reviews have reinforced these findings. Upadhyay et al. [36] provided a wide-ranging overview of ML models applied to boiling heat transfer, recommending feature reduction, algorithm selection, and physics-inspired modelling to improve interpretability and generalisability. Jirasek and Hasse [37] reviewed deep learning approaches that integrate MD simulation data with ML models, underlining their potential in predicting thermal conductivity and specific heat across ionic and molecular fluids.

The degradation of thermal fluids not only presents safety hazards but also leads to increased operational costs and maintenance demands. Repeated heating cycles that approach or exceed the fluid's boiling point accelerate chemical breakdown, reduce heat transfer efficiency, and shorten service life. Early prediction of boiling behaviour during the initial phases of system operation, when the fluid remains chemically stable, is therefore essential for effective performance management. Biphenyl and diphenyl ether were selected for this investigation due to their excellent thermal stability and widespread industrial use in high-temperature applications. Accurate boiling point estimation during early operational conditions supports proactive intervention, thereby reducing degradation risks and improving system reliability. While ML methods have shown promise in enhancing prediction capabilities beyond traditional empirical correlations, their integration with molecular-level simulation data remains limited.

To fill this gap, the present study develops a hybrid molecular dynamics–machine learning (MD–ML) framework for boiling point estimation in aromatic thermal fluids. The approach combines measurable thermodynamic properties extracted from equilibrium MD simulations, including density, kinetic energy, potential energy, and pressure–volume data, along with multiple supervised ML algorithms. A systematic evaluation of kernel-based, instance-based, and neural network (NN) models is carried out to identify architectures best suited for capturing thermophysical trends and phase-transition behaviour. Additionally, multiple force fields (OPLS-AA and COMPASS CLASS2) are employed to assess the reliability of MD-predicted thermophysical properties and their alignment with a diverse set of experimental benchmarks. Boiling points are estimated using both a fixed density threshold and an inflection-point method, allowing robust cross-validation of phase transition detection. Model accuracy is quantified using mean absolute error (MAE), mean squared error (MSE), R^2 , and Pearson correlation (PC) metrics. This integrative framework offers a reproducible and physically grounded strategy for boiling point prediction, while it also provides guidance on force field selection and ML model suitability for a wider range of thermal fluid systems.

2. Methodology

The study analysed biphenyl ($C_{12}H_{10}$) and diphenyl ether ($C_{12}H_{10}O$) molecules because they represent typical examples of emerging components in thermal oil heat-transfer systems. Biphenyl ($C_{12}H_{10}$) and diphenyl ether ($C_{12}H_{10}O$) undergo thermal degradation at extended high temperatures, resulting in multiple byproducts. Biphenyl degradation produces both low-boiling chemicals such as benzene (C_6H_6) and phenol (C_6H_5OH) and high-boiling oligomers such as terphenyls ($C_{18}H_{14}$) and quaterphenyls ($C_{24}H_{18}$). Diphenyl ether degrades into volatile phenol (C_6H_5) and benzene (C_6) compounds while creating dibenzofurans ($C_{12}H_8O$), which elevate viscosity levels and obstruct fluid movement [22].

Initially, the molecular structures for biphenyl and diphenyl ether were constructed using Packmol [23]. In all systems, a total of 490 molecules were utilized. Previous studies indicated that increasing the number of molecules did not significantly affect the prediction for density [24]. The simulation process was independently replicated three times in each scenario to remove any bias that may occur from random molecular positioning at the beginning. The energy minimisation process enabled atoms in each configuration to reach their minimum potential energy states which ensured accurate molecular arrangements. The study consistently assigned partial atomic charges through Xenoview [25] and ArgusLab [26] to ensure accurate depiction of electrostatic interactions in MD simulations. The MD framework utilized OPLS-AA (Optimized Potentials for Liquid Simulations-All Atom) and COMPASS CLASS2 (Condensed-phase Optimized Molecular Potentials for Atomistic Simulation Studies) force fields to determine the effects of force-field selection on thermophysical prediction accuracy. Deploying both force fields facilitated a comprehensive assessment of molecular interaction parameters, resulting in strong validation through available experimental data. Molecular structures, post-energy minimisation and replication, were combined using the *replicate* and *append* commands within the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software [27]. The simulation systems underwent structural verification through OVITO visualization

software [28], which confirmed proper assembly of molecular configurations and structural integrity. The simulation systems entered controlled equilibrium and production phases in LAMMPS after their preparation. The resulting MD-generated datasets included detailed thermodynamic parameters such as density, kinetic energy, potential energy, pressure, and volume, which are all essential indicators of phase behaviour. The use of two different force fields, OPLS-AA and COMPASS CLASS2 along with comparison against experimental data allows the determination of which force field provides superior accuracy for biphenyl and diphenyl ether, aiming to improve the reliability of predictive analysis for these substances.

MD simulation datasets needed thorough pre-processing before ML application which combined multiple steps including data conditioning, labelling, curation and elimination of outliers and noise together with correlation analysis. The cleaned data was divided into training and testing subsets in an 80–20 % ratio using the *train_test_split* method in Python [29,30]. Various ML models received training and evaluation treatment through the Monolith AI platform and PyCharm software. Throughout model training sessions learning curves were consistently tracked to identify and correct problems related to underfitting and overfitting. Model performance was comprehensively assessed through multiple metrics: mean absolute error (MAE), mean squared error (MSE), determination coefficient (R^2), and Pearson coefficient (PC) as the primary metrics. The chosen ML model demonstrated optimal results on every evaluation criterion and achieved high accuracy for predicting biphenyl boiling points, which supports early thermal fluid management interventions (see Fig. 1). The detailed configurations and parameter settings utilized in MD simulations, including force field parameters and equilibrium conditions, are further discussed in Sec. 3.1, providing clarity on MD simulation procedures. Sec. 3.2 elaborates upon the ML techniques employed, including their configuration, training, and the rationale for model selection, ensuring transparency and reproducibility of the methodology.

3. Configuration of MD and ML

3.1. MD simulations

3.1.1. Force field setup

All MD simulations were performed using LAMMPS, employing the real unit system to ensure consistency with experimental thermophysical data. Periodic boundary conditions were imposed in all three spatial dimensions to replicate bulk-phase behaviour, and a full-atom representation was used to accurately capture intermolecular interactions. The velocity-Verlet algorithm was employed to integrate the equations of motion, balancing computational efficiency and numerical stability [29]. Long-range Coulombic interactions were calculated using the particle–particle particle–mesh (PPPM) method with a precision of 1.0×10^{-4} . A cutoff distance of 1.2 nm was applied to nonbonded interactions to efficiently model both van der Waals and electrostatic forces.

To investigate the influence of potential models on boiling point estimation, two force fields were employed: OPLS-AA and COMPASS CLASS2. These force fields were manually implemented in LAMMPS through custom assignment of bond, angle, and dihedral parameters. For OPLS-AA, bonded interactions were defined using *dihedral_style opls*, and nonbonded terms were assigned using *pair_style lj/cut/coul/long*. For COMPASS CLASS2, the higher-order bonded terms were incorporated using *dihedral_style class2*, and nonbonded interactions were modelled using *pair_style lj/class2/coul/long*. Force field parameters were sourced from published studies involving organic and aromatic molecules [63,64,67], ensuring compatibility with the molecular structures of biphenyl and diphenyl ether used in this study, as detailed in Tables 1 and 2. This implementation ensured full compatibility with LAMMPS' force

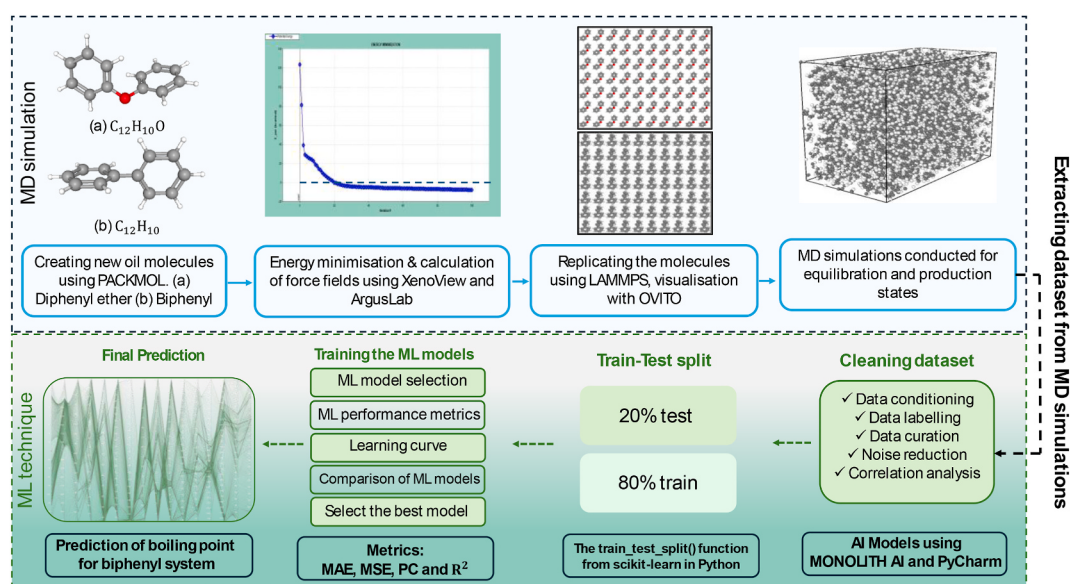


Fig. 1. Schematic of integrating MD simulation with ML algorithm to predict the boiling point of the thermal fluid.

field styles and accurately reproduced the molecular topologies of the target fluids.

The energy minimisation process was completed by setting force and energy convergence thresholds to 1.0×10^{-12} to achieve a well-prepared starting structure. All systems were initially equilibrated under the ensemble (NPT) at 300 K and 1atm pressure for 0.25ns using a Nose–Hoover thermostat and barostat to establish volumetric and thermal stability. The initial density was not manually set but allowed to emerge from pressure relaxation during the NPT equilibration stage. This strategy ensures that the system naturally evolves to its equilibrium volume and density at 300 K and 1atm without applying artificial constraints. Although the NPT equilibration was performed for 0.25ns, this duration was validated through preliminary trial simulations that showed rapid convergence of total energy, volume, and temperature within this period. To further confirm equilibrium, the atomic trajectories, energy drift, and volume evolution were analysed using LAMMPS output logs and OVITO visualization. The subsequent production stage in canonical ensemble (NVT) exhibited no systematic drift in thermodynamic observables, thereby confirming the system's relaxation to equilibrium before heating. This was followed by a 1ns heating phase under the NVT ensemble, where temperature was linearly increased to the target value in each case. These conditions ensured reproducible and stable MD trajectories across the full temperature range evaluated. Subsequently, for boiling point detection, a series of independent NVT simulations were performed at discrete target temperatures, with each system equilibrated separately to extract the corresponding equilibrium density.

The OPLS-AA force field models the total potential energy as a summation of bonded and non-bonded interactions [63–65]:

$$E_{\text{total}} = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedral}} + E_{\text{nonbonded}} \quad (1)$$

Bond stretching and angle bending interactions follow harmonic potentials:

$$E_{\text{bonds}} = \sum_{\text{bonds}} K_r (r - r_0)^2 \quad (2)$$

$$E_{\text{angles}} = \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \quad (3)$$

Torsional interactions are modelled using a Fourier series expansion:

$$E_{\text{dihedrals}} = \sum_{\text{dihedrals}} \left(\frac{1}{2} K_{\varphi_1} [1 + \cos(\varphi)] + \frac{1}{2} K_{\varphi_2} [1 + \cos(2\varphi)] + \frac{1}{2} K_{\varphi_3} [1 + \cos(3\varphi)] + \frac{1}{2} K_{\varphi_4} [1 + \cos(4\varphi)] \right) \quad (4)$$

Non-bonded interactions incorporate Coulombic interactions and Lennard-Jones (LJ) 12-6 potential:

Table 1
OPLS-AA force field parameters for biphenyl and diphenyl ether.

OPLS-AA Nonbonded parameters and atom types for biphenyl ₁ and diphenyl ether ₂					
Name	Element	Type	q_{OPLS}	$\sigma(\text{ \AA})$	$\varepsilon(\text{kcal/mol})$
C ₁	C	CH	−250	3.55	0.07
H ₁	H	H	240	2.42	0.03
C ₂	C	CH	−250	3.55	0.07
H ₂	H	H	+240	2.42	0.03
O ₂	O	O	−430	3.12	0.17
OPLS-AA Bond stretching parameters and bond types for biphenyl ₁ and diphenyl ether ₂					
Bond type	$K_r(\text{kcal.mol}^{-1}.\text{ \AA}^{-2})$			$r(\text{ \AA})$	
C ₁ -C ₁	469			1.4	
C ₁ -H ₁	367			1.08	
C ₂ -C ₂	469			1.4	
C ₂ -H ₂	367			1.08	
C ₂ -O ₂	320			1.41	
OPLS-AA Angle bending parameters and angle types for biphenyl ₁ and diphenyl ether ₂					
Angle type	$K_\theta(\text{kcal.mol}^{-1}.\text{rad}^{-2})$			$\theta(\text{deg})$	
C ₁ -C ₁ -C ₁	35			120	
C ₂ -C ₂ -C ₂	63			120	
C ₂ -C ₂ -O ₂	35			109.47	
C ₂ -C ₂ -H ₂	35			120	
C ₂ -O ₂ -H ₂	55			108.5	
H ₁ -C ₁ -H ₁	33			107.8	
OPLS-AA Torsion parameters and dihedral types for biphenyl ₁ and diphenyl ether ₂					
Dihedral type	K_{φ_1}	K_{φ_2}	K_{φ_3}	K_{φ_4}	
C ₁ -C ₁ -C ₁ -C ₁	1.3	−0.2	0.2	0.0	
C ₁ -C ₁ -C ₁ -H ₁	−1.552	0.0	0.0	0.0	
C ₂ -C ₂ -C ₂ -C ₂	1.3	−0.2	0.2	0.0	
C ₂ -C ₂ -C ₂ -H ₂	−1.552	0.0	0.0	0.0	
C ₂ -C ₂ -O ₂ -H ₂	−0.356	−0.174	0.492	0.0	

Footnote: subscripts '1' and '2' in atoms of carbon (C), hydrogen (H) and oxygen (O) refer to as the belonging of biphenyl or diphenyl ether, respectively.

Table 2
COMPASS Class 2 force field parameters for biphenyl and diphenyl ether.

COMPASS Pair coefficients for biphenyl 1 and diphenyl ether 2				
Element	ϵ (kcal/mol)			σ (Å)
C ₁	0.055			3.85
H ₁	0.02			2.42
C ₂	0.055			3.85
H ₂	0.02			2.42
O ₂	0.08			3.0
COMPASS Bond coefficients for biphenyl 1 and diphenyl ether 2				
Bond type	r (Å)	K_{r2} (kcal.mol ⁻¹ .Å ⁻²)	K_{r3} (kcal.mol ⁻¹ .Å ⁻²)	K_{r4} (kcal.mol ⁻¹ .Å ⁻²)
C ₁ -C ₁	1.53	299.67	-501.77	679.81
C ₁ -H ₁	1.1	345.0	-691.89	844.6
C ₂ -C ₂	1.53	299.67	-501.77	679.81
C ₂ -H ₂	1.1	345.0	-691.89	844.6
C ₂ -O ₂	1.41	320.0	-650.0	720.0
COMPASS Angle coefficients for biphenyl 1 and diphenyl ether 2				
Bond type	θ (deg)	$K_{\theta2}$ (kcal.mol ⁻¹ .Å ⁻²)	$K_{\theta3}$ (kcal.mol ⁻¹ .Å ⁻²)	$K_{\theta4}$ (kcal.mol ⁻¹ .Å ⁻²)
C ₁ -C ₁ -C ₁	120.0	39.516	-7.443	-9.558
C ₁ -C ₁ -H ₁	110.77	41.453	-10.604	5.129
C ₂ -C ₂ -C ₂	120.0	39.516	-7.443	-9.558
C ₂ -C ₂ -H ₂	110.77	41.453	-10.604	5.129
C ₂ -C ₂ -O ₂	109.5	40.0	-9.8	6.5

Footnote: subscripts '1' and '2' in atoms of carbon (C), hydrogen (H) and oxygen (O) refer to as the belonging of biphenyl or diphenyl ether, respectively.

$$E_{\text{nonbond}} = \frac{Cq_i q_j}{\epsilon_0 r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (5)$$

The force constants in these equations are K_r , K_θ , and K_{φ_n} ; the bond lengths, angles, and dihedral angles are represented by r , θ , and φ , respectively, and r_0 , θ_0 and φ_0 are the equilibrium values. There are two primary parts in E_{nonbond} : the first term, Coulombic interactions between partial charges, is represented by $\frac{Cq_i q_j}{\epsilon_0 r_{ij}}$, where q_i and q_j are the partial charges of atoms i and j , ϵ_0 is the vacuum permittivity, r_{ij}

is the distance between the atoms, and C is an energy-conversion constant. The second term, $4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$, characterizes the van der Waals interactions, where ϵ_{ij} and σ_{ij} are the energy and diameter parameters, respectively. Table 1 provides the selected values of parameters in this study [63,64,67].

COMPASS CLASS2 utilizes higher-order bonded potential terms to achieve better precision in condensed-phase simulations. The behaviour of bond stretching and angle bending is modelled with third- and fourth-order polynomial functions [66,67]:

$$E_{\text{bonds}} = K_{r2}(r - r_0)^2 + K_{r3}(r - r_0)^3 + K_{r4}(r - r_0)^4 \quad (6)$$

$$E_{\text{angles}} = K_{\theta2}(\theta - \theta_0)^2 + K_{\theta3}(\theta - \theta_0)^3 + K_{\theta4}(\theta - \theta_0)^4 \quad (7)$$

The modelling of dihedral and improper torsions utilizes higher-order expressions. The COMPASS CLASS2 version of the Lennard-Jones potential uses a 9-6 formulation:

$$E_{\text{Elec-LJ}} = \sum_{i < j} \frac{q_i q_j}{\epsilon_0 r_{ij}} + \sum_{i < j} \epsilon_{ij} \left[2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^9 - 3 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (8)$$

The force constants and bond lengths along with the Lennard-Jones parameters for carbon hydrogen and oxygen atoms match well with values found in previous research [66–68]. Table 2 displays the detailed parameter values for COMPASS CLASS2 force field.

The system was subsequently heated in a linear manner after equilibration, raising the temperature from 300 K to the target temperature over 1ns using the NVT ensemble. Two heating conditions were explored: The first set of OPLS-AA and COMPASS CLASS2 simulations heated the system from 300 to 600 K while the second set of OPLS-AA simulations reached a higher target of 700 K. The heating phase creates a controlled environment that enables researchers to examine boiling transition dynamics while directly comparing how different temperatures influence molecular interactions and phase behaviour.

3.1.2. Density

The system density ρ was calculated by dividing the total mass of thermal fluid by the volume of simulation box and is given by:

$$\rho = \frac{NM}{V_E N_A} \quad (9)$$

where N denotes the particle number, M denotes the molar mass, V_E represents the simulation cell's equilibrium volume at a NVE system, and N_A is the Avogadro's constant [69].

3.1.3. Viscosity

The viscosity measurements of biphenyl and diphenyl ether thermal fluids were conducted through equilibrium molecular dynamics (EMD) simulations based on the Green-Kubo method. The approach involves integrating the autocorrelation function of stress tensor components to connect viscosity with molecular momentum transport [38]. The canonical ensemble (NVT) was used in simulations to maintain constant temperature while keeping the system in equilibrium and enabling the collection of statistical data. The Green-Kubo relation determines shear viscosity η by integrating the autocorrelation function of the stress tensor's off-diagonal elements [39], and η is expressed as:

$$\eta = \frac{V}{k_B T} \int_0^\infty \langle P_{ij}(0) P_{ij}(t) \rangle dt \quad (10)$$

where V is the system volume, k_B represents the Boltzmann constant, and T denotes temperature alongside P_{ij} which signifies the off-diagonal pressure tensor component. The integral represents the time correlation function for pressure tensor components and demonstrates how intermolecular forces develop over time and affect overall viscosity. The total viscosity was obtained by averaging the contributions from different off-diagonal components of the stress tensor [40]:

$$\eta = \frac{1}{3} (\eta_{xy} + \eta_{xz} + \eta_{yz}) \quad (11)$$

where η_{xy} , η_{xz} , and η_{yz} correspond to viscosity components derived from the autocorrelation functions of the respective stress tensor elements. To ensure accurate viscosity calculations, a sufficiently long sampling time was used to capture the full relaxation behaviour of the stress autocorrelation function. Regular interval simulation data collection was followed by numerical integration using the trapezoidal rule for Green-Kubo integral approximation [38,40].

3.2. ML techniques

3.2.1. Data framework and structure

This research produced a comprehensive dataset generated via MD simulations to analyse the thermophysical properties of both biphenyl and diphenyl ether thermal fluids. The MD simulations ran for a full production period of 1ns while recording data every 1000fs. High temporal resolution characterizes this approach because it precisely captures dynamic molecular interactions and phase transition behaviours. The intricacy of atomic-scale systems necessitates the use of a structured methodology for data extraction to achieve accurate representation in ML applications. The dataset contains essential thermodynamic and molecular descriptors that enable analysis of system behaviour across different temperature ranges. The dataset contains temperature (Temp), density (Density), potential energy (PotEng), kinetic energy (KinEng), total energy (TotEng), pressure (Press), and system volume (Volume) that represents the product of dimensions L_x , L_y , L_z under periodic boundary conditions to track density changes accurately and reduce edge effects and finite-size artifacts throughout the simulation. The various parameters collectively reveal a multidimensional view of molecular dynamics by identifying essential factors of intermolecular forces and phase transitions. Although kinetic energy is directly related to the instantaneous temperature, the target variable in this study reflects a broader thermodynamic profile. The ML models were trained on raw MD-derived features, including kinetic energy, potential energy, density, total energy, and simulation time progression. As a result, the predicted temperature was interpreted not as a frame-level KinEng equivalent, but as a phase-sensitive,

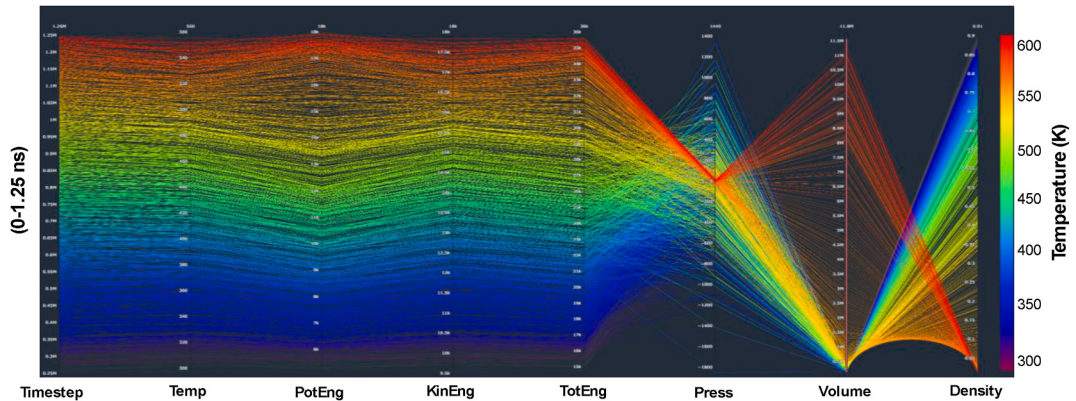


Fig. 2. Parallel coordinate plot of MD simulation data: key thermophysical properties were extracted from MD simulations over a 1.25ns simulation period. The data consists of 0.25ns for equilibration using NPT ensemble, followed by 1ns for the production state where thermodynamic properties were recorded every 1000fs. The colour gradient represents the temperature variations ranging from 300 to 600 K.

smoothed estimate that captures equilibrium trends across varying conditions. The learning process integrates multidimensional descriptors that collectively encode phase stability, energy accumulation, and boiling onset.

Analysing MD-generated datasets presents a significant challenge due to the inherent noise and nonlinear behaviour undergoing in molecular interactions [41,42]. Data quality improvements require rigorous preprocessing when energy distributions exhibit fluctuations alongside pressure and density variations. The research utilized noise reduction techniques to filter irrelevant data and retain only significant variations for further analysis by using Monolith AI platform and PyCharm. ML model training remained unbiased because normalizing variables to a common scale ensured dimensional consistency. The study used parallel coordinate plots to examine relationships among extracted features while identifying crucial influencing parameters. Fig. 2 demonstrates how temperature connects with energy distributions alongside pressure and density through parallel axes mapping. The visualization technique enables intuitive perception of dataset interdependencies, which helps analyse detection trends and anomalies. Each line in this visualization stands for a specific simulation moment, and temperature changes are shown through different colours. The analysis reveals a distinct pattern where rising temperatures show a connection with energy changes which then cause measurable fluctuations in pressure and density. The observed increase in line density during higher temperature conditions implies possible phase transition onset, highlighting the essential role of accurate boiling point predictions. Capabilities of the interactive data analysis enable the removal of outliers and inconsistencies and help produce reliable predictive modelling input. This visualization capability improves data comprehension while identifying essential factors that influence thermophysical properties in biphenyl and diphenyl ether systems. ML models derive their structure from the core dataset which undergoes an 80-20 % train-test division to achieve generalization. The training data enables the model to recognize hidden patterns, and the test set functions as an evaluation standard to validate performance [29,30]. Prior to ML model development, all MD-derived features were processed to remove statistical noise and outliers. Each feature trajectory (e.g., pressure, energy, density) was smoothed using a 5-point moving average to suppress the thermal fluctuations. Outlier detection was performed using the interquartile range (IQR) method: any data point falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ was excluded. $Q1$ represents the first quartile, the 25th percentile of the data, and IQR is the difference between the third quartile ($Q3$) and the first quartile ($Q1$). Features were then standardised using z-score normalisation. All preprocessing was conducted before the train-test split. This ensured that both the training and test datasets reflected uniformly conditioned thermodynamic signals. The effectiveness of this preprocessing was further validated through principal component analysis (PCA), which revealed the low feature redundancy and a structured temperature-driven variance profile. Predictive reliability improves when the dataset's extensive range and variety undergo thorough preprocessing. Using the *train_test_split* function from the *scikit-learn* library ensures that the dataset is split in a way that minimizes bias and maintains the consistency of data distribution across both sets. This study combines MD with ML methodologies to benefit from both approaches and achieve accurate computational boiling point predictions.

To examine the intrinsic dimensionality of the input feature space and identify dominant modes of variation, PCA was applied to the scaled ML input dataset. The first principal component (PC1) accounts for over 73 % of the total variance, while PC2 explains an additional 16 %, indicating that the system's behaviour is largely governed by one or two strongly correlated physical trends (see Fig. 3 (a)). The PCA scatter plot in Fig. 3(b) reveals a continuous temperature gradient along PC1, showing that thermally sensitive features, primarily kinetic energy, total energy, and density, dominate the input variance. This result confirms that temperature acts as a latent driver of feature correlations in the dataset. PC2 is orthogonal to PC1 and appears to capture pressure–volume variations. These observations justify the use of temperature-related thermophysical features in ML models for boiling point prediction and confirm that the input space is both low-dimensional and physically structured rather than noise-dominated. PCA loading vectors were also examined to assess redundancy among input features. Kinetic energy, total energy, and temperature-aligned features were all closely correlated along PC1, indicating that they contribute strongly to the thermal variance. Volume and pressure were grouped along PC2

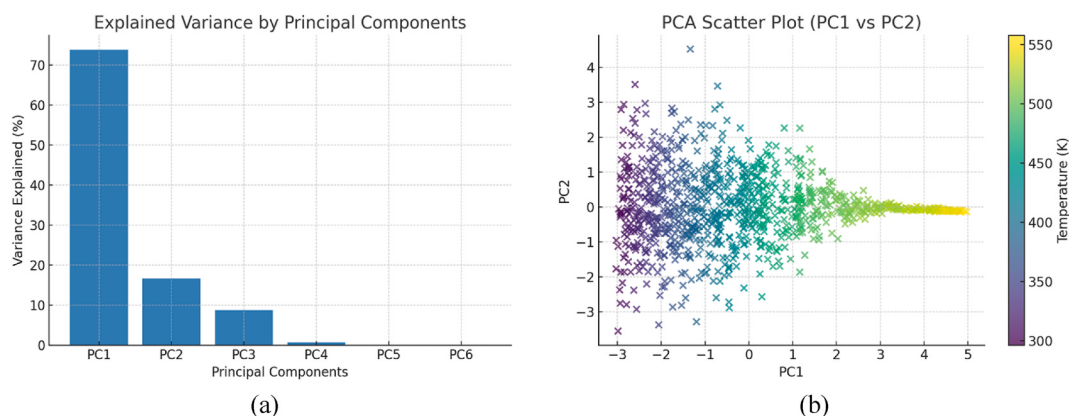


Fig. 3. Principal component analysis (PCA) of the ML input feature space: (a) Scree plot showing variance explained by each principal component, PC1 and PC2, together capturing ~90 % of the total variance and (b) PCA scatter plot (PC1 vs. PC2), colour-coded by simulation temperature. The clear gradient along PC1 indicates that temperature-sensitive thermophysical features such as kinetic energy, density and total energy are the primary drivers of variance in the input dataset.

and consistent with their low variability under the NVT ensemble. Although some correlation exists, particularly between kinetic energy and total energy, these variables represent distinct thermodynamic contributions and are thus retained to preserve model interpretability. Combining with the sensitivity analysis in Fig. 14, which confirms that pressure and volume have negligible influence while kinetic energy dominates the prediction, the findings in PCA validate our input selection strategy. No features were removed during training, but this analysis supports the potential for dimensionality reduction in future studies.

In addition to PCA, mutual information (MI) analysis was performed to quantify the direct informational contribution of each input feature to temperature prediction. Unlike PCA, which evaluates global linear variance across the input space, MI captures nonlinear dependencies between individual features and the target variable. As shown in Fig. 4, kinetic energy has the highest MI score as expected due to its direct thermodynamic link to temperature, but other features such as total energy, potential energy, and density also demonstrate substantial information content. This analysis reinforces that the model learns from a multidimensional thermodynamic signal, not just from KinEng alone. Volume and pressure exhibit lower MI values, keeping consistency with their limited variability under the NVT ensemble and PCA loadings on PC2. By considering the MI analysis and PCA, it can be confirmed that the dataset is both informationally rich and physically structured, supporting the chosen feature set for temperature-based property prediction. It is noted that the temp–volume correlation has been included not because volume is a dominant feature but to evidence that volume remains stable within each run and does not artificially track temperature under NVT. This rules out ensemble-related artifacts. Pressure is excluded because it is unconstrained in NVT and already known to be the least informative feature based on the MI analysis.

3.2.2. ML algorithms

The complexity of thermophysical interactions in thermal fluid systems requires the selection of robust regression models to effectively capture nonlinear relationships and improve generalisability. The present study utilized three distinct ML models, including Support Vector Regression (SVR), Neural Networks (NN), and Nearest Neighbour Regression (NNR), because each model demonstrated unique strengths in effective regression tasks. The Monolith AI platform supports the implementation process by enabling cross-validation procedures and hyperparameter optimization along with model comparison to achieve reliable predictive results. SVR functions as an effective regression method that focuses on reducing prediction errors while preserving simple model structure. SVR demonstrates excellent performance on datasets with complex nonlinear relationships, making it a suitable method for predicting boiling points from MD-generated thermodynamic parameters. Hyperparameter tuning through randomized search helped optimize SVR performance by comparing 10 distinct model configurations. The model applied a radial basis function (RBF) kernel to model nonlinear dependencies in data and tested different values for the regularization parameter (C), specifically {0.001, 0.01, 0.1, 1, 10, 100}. Additionally, the kernel coefficient (gamma) was tested using multiple settings: scale, auto, 0.0001, 0.01, and 0.1. A 5-fold cross-validation technique was employed in the present study to evaluate model stability with root mean squared error (RMSE) as the performance evaluation metric. The employed method prevented overfitting while allowing the model to perform accurately on new data.

NN were selected because they can learn sophisticated patterns from extensive datasets. NN were configured to evaluate multiple network architectures using hyperparameter tuning to optimize essential parameters for maximum accuracy, and a randomized search technique was implemented to test 10 different configurations by varying batch sizes {16, 32, 128, 512} and hidden layers in {1, 2, 3} quantities. The network could adjust its capacity to data complexity by selecting hidden layer sizes from a set of options: {20, 50, 75, 100, 150, 200, 300, 500}. Three activation functions were tested across three common nonlinear transformations, including rectified

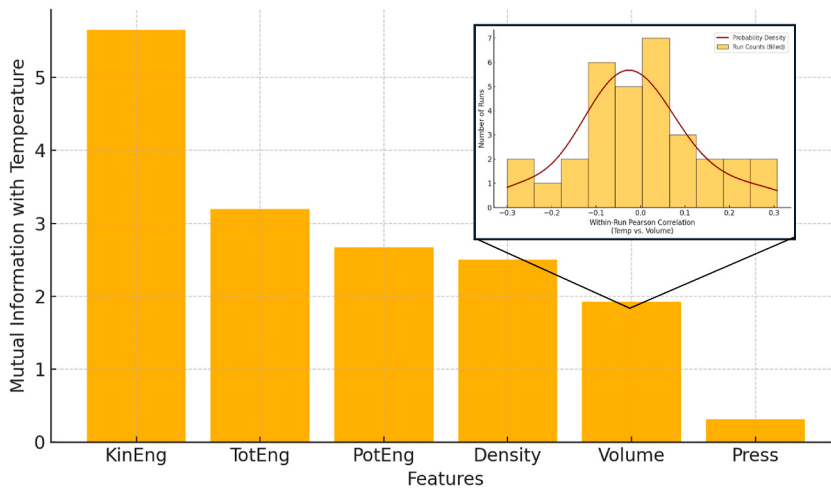


Fig. 4. Mutual information (MI) analysis of input features (KinEng, TolEng, PotEng, Density, Volume, and Press) in MD with temperature, demonstrating that temperature-related signals exist beyond kinetic energy. Inset shows within-run temp–volume correlations under NVT conditions, further confirming that the lower MI score for volume reflects physical ensemble constraints rather than noise.

linear unit (ReLU), exponential linear unit (ELU), and Swish. The model utilized dropout regularization at rates of 0.05 and 0.1 to counteract overfitting. The NN model also used 5-fold cross-validation with RMSE as the evaluation metric to maintain model reliability by avoiding overfitting and underfitting.

NNR functions as a non-parametric predictive model to determine outcomes by using the nearest data points in feature space. The relationship between thermodynamic properties such as density makes NNR a suitable benchmark for boiling point prediction. A randomized search strategy was employed to test 10 different configurations to optimize performance by altering the number of nearest neighbours (k), with values ranging from {1, 2, 3, 4, 5, 10, 20, 40}. This k parameter determines the prediction smoothing level, where lower values provide high sensitivity to local variations and higher values enhance generalization. The model's capability to capture local relationships while maintaining low variance was verified through a 5-fold cross-validation process that used RMSE as the evaluation metric.

The ranges for hyperparameter selection in all three models were chosen based on a combination of preliminary trial simulations, prior literature on regression problems in thermophysical modelling, and empirical tuning recommendations from widely used ML frameworks. For SVR, values of the regularization parameter (C) were ranged from 0.001 to 100 to explore a wide spectrum from strong regularization to high flexibility, while gamma settings (scale, auto, and small numeric values) were selected to capture both adaptive and fixed RBF kernel spreads. The NN architecture tested varying hidden layer sizes (20–500 neurons), batch sizes (16–512), and multiple activation functions to ensure coverage of both shallow and deeper networks suited for structured datasets. Dropout rates were selected conservatively to prevent overfitting in medium-sized models. For NNR, the k -values tested (1–40) were chosen to balance the sensitivity to local variation against generalisation across global patterns. These ranges were not arbitrary but informed by common practices in tuning ML models for physical regression tasks with structured numerical features.

3.2.3. Model evaluation

Multiple metrics have been utilized in the present study to evaluate the accuracy of ML models and determine which model most accurately predicts the boiling point of biphenyl, including mean absolute error (MAE), mean squared error (MSE), Pearson coefficient (PC), and determination coefficient (R^2). MAE measures the average size of prediction errors without considering whether they are positive or negative. The metric calculates the mean absolute deviation between predicted values and actual data points throughout the test sample and treats each deviation equally, and MAE is defined below [43]:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

where n is the number of observations, y_i is the actual value, $\hat{y}_i$ is the predicted value, and $|y_i - \hat{y}_i|$ is the absolute difference between the actual (MD results) and predicted values (ML predictions).

MSE measures the average of squares of the errors. It is more sensitive to outliers than MAE due to the squaring of each term, meaning that larger errors have a disproportionately large effect on MSE. MSE is given by Ref. [43]:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (13)$$

PC determines the strength and direction of the linear relationship between two variables by providing a value ranging from -1 to 1 . Value 1 represents a perfect positive linear correlation while -1 represents a perfect negative linear correlation and 0 indicates nonlinear relationship between variables. PC can be expressed by Ref. [44]:

$$\text{PC} = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (14)$$

In this case, $\bar{y}$ is the mean of the actual values, and $\bar{\hat{y}}$ is the mean of the predicted values. The deviation of actual values from the mean of actual values is denoted by $(y_i - \bar{y})$, and the deviation of predicted values from the mean of predicted values is denoted by $(\hat{y}_i - \bar{\hat{y}})$.

R^2 quantifies the percentage of the dependent variable's variance that can be predicted based on the independent variables. R^2 has a range of 0 – 1 , with higher values denoting greater model performance. R^2 can be defined as [45]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (15)$$

where the total variance in the actual values is measured by the sum of squares $(y_i - \bar{y})^2$, and the sum of the squared differences between the actual and predicted values $(y_i - \hat{y}_i)^2$. By measuring the prediction accuracy and error, these metrics together assist to assess the performance of the regression models and help choose the best model.

4. Results and discussion

4.1. Validation of MD simulation results

Integrating reliable data into ML models requires to evaluate the accuracy of MD simulation results by comparing them against experimental datasets. To avoid predictions which compromise ML model performance, suitable force field values guarantee accurate computations of thermophysical properties. Extensive comparisons between reference experimental data and generated density and viscosity values constituted part of the validation procedure. Experimental data from the NIST database [46] was compared with the estimated biphenyl density derived from OPLS-AA and COMPASS force fields. Fig. 5(a) illustrates a clear departure of the results between the COMPASS force field and experimental data as biphenyl density fluctuates with temperature. OPLS-AA follows the expected pattern; COMPASS does not replicate the exact temperature dependence of biphenyl density. At high temperatures such as 490 K COMPASS exhibit an unrealistically rapid decline in density levels. To assess the accuracy of OPLS-AA force field, Fig. 5(b) presents a thorough comparison between OPLS-AA forecasts and NIST values for a given temperature range. The strong agreement between the two datasets attests to OPLS-AA's fit for computing biphenyl density.

Table 3 shows percentage errors for both force fields over a wide range of temperatures, thereby supporting this selection by means of quantitative error analysis. Whereas OPLS-AA maintains error percentages below 2 % for all temperatures examined, the experimental results demonstrate that COMPASS consistently increases divergence from real values with increasing temperatures, displaying an error of 46.80 % at 498 K and 94.16 % at 523 K, respectively. The data show that OPLS-AA offers a realistic model for predicting the biphenyl density.

Furthermore, the thermophysical properties of diphenyl ether were validated by experimental viscosity data from Byers et al. [47] and density data from PubChem [48] and experimental study by Cabaleiro et al. [49]. Fig. 6(a) shows the simulated viscosity results from the OPLS-AA force field are in good agreement with the experimental data from Byers et al. Strong correlation is shown by the simulation data since it truly reflects the decreasing pattern of viscosity observed in experimental measurements with the increasing temperature, thus supporting theoretical predictions. Comparing OPLS-AA results with experimental data from PubChem reinforces the accuracy of the chosen force field by means of the density evaluation of diphenyl ether displayed in Fig. 6(b). The consistent values of diphenyl ether density predicted by OPLS-AA are reflected in both dataset's same linear behaviour. Although minor errors are inherent within a reasonable margin, the force field fairly reflects the thermophysical characteristics of diphenyl ether. In contrast, the COMPASS force field fails to provide accurate density predictions, reflecting greater deviations from experimental data compared with OPLS-AA. The close alignment of the viscosity computations for diphenyl ether with experimental data shows that the chosen values fairly capture the temperature-dependent behaviour of the fluid. The better accuracy of MD dataset qualifies it for ML studies.

4.2. Boiling point prediction by MD

In this study, the boiling point is estimated by analysing how the equilibrium density changes across a range of simulated temperatures, using a series of independent NVT simulations. At each target temperature, the system is equilibrated separately, and the corresponding densities are compiled into a discrete $\rho - T$ profile. The boiling onset is then defined thermodynamically as the temperature at which this curve exhibits its most rapid decline, i.e., the inflection point, corresponding to the minimum of the first-order derivative $d\rho/dT$. This inflection-based criterion provides a physically grounded signal of incipient vaporisation and offers a practical alternative to two-phase coexistence methods that require explicit interface tracking. Surface tension was not computed in this study,

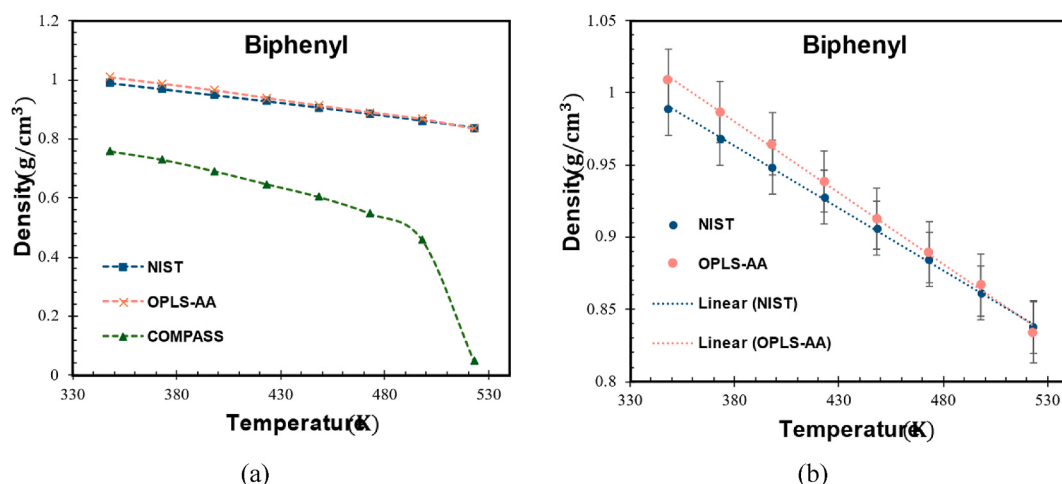
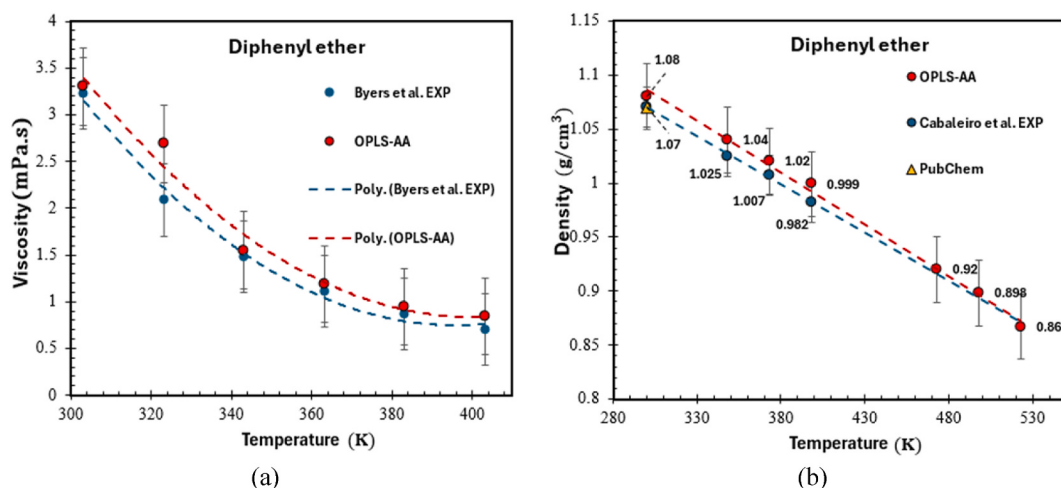


Fig. 5. Comparison of biphenyl density between MD simulation results and experimental measurements from NIST: (a) Density predictions of biphenyl using OPLS-AA and COMPASS force fields compared with NIST data and (b) Comparison between OPLS-AA and NIST data for biphenyl density.

Table 3

Error analysis of simulated density for biphenyl using OPLS-AA and COMPASS force fields.

Temp. (K)	ρ (g/cm ³)			Error (%)	
	NIST	OPLS-AA	COMPASS	OPLS-AA	COMPASS
348.13	0.9892	1.0087	0.7576	1.97	23.41
373.12	0.9685	0.9866	0.7293	1.87	24.70
398.12	0.9482	0.9646	0.6908	1.73	27.15
423.11	0.9277	0.9387	0.6471	1.19	30.25
448.11	0.9063	0.913	0.6042	0.74	33.33
473.11	0.8845	0.8897	0.5468	0.59	38.18
498.11	0.8618	0.8669	0.4585	0.59	46.80
523.11	0.8378	0.8339	0.0489	0.47	94.16

**Fig. 6.** Validation of MD simulated viscosity and density for diphenyl ether: (a) Comparison of viscosity results for diphenyl ether between OPLS-AA simulation results and experimental data and (b) Comparison of density for diphenyl ether between OPLS-AA simulation results and experimental data.

as the simulation protocol did not involve the vapor–liquid coexistence or interface formation. In molecular dynamics, surface tension emerges from the presence of a density gradient across an interfacial region, which requires two-phase configurations and spatially resolved slab geometries [58,65,70]. Since the current work relies solely on equilibrium bulk-phase trajectories under NVT, surface tension is not defined. Instead, boiling onset was inferred from the thermodynamic response of density, following protocols demonstrated in recent single-phase studies [59–61], which showed that inflection-based density analysis can reliably capture the phase transition behaviour without requiring interfacial metrics. Fig. 7 schematises this protocol. Each density point arises from an independently equilibrated NVT trajectory, and the boiling point is inferred from the curvature change in the $\rho - T$ trend rather than from a fixed density threshold or time-dependent trajectory. This approach circumvents the system size demands and stability limitations, which are typical of direct-coexistence simulations under NPT or Gibbs ensembles [58]. In fact, it aligns with modern single-phase strategies, where phase transition signatures are extracted from the equilibrium thermodynamic response functions. Gorfer et al. [59] demonstrated that inflection points in bulk properties such as density and surface tension can mark the onset of structural transformation, even without phase interfaces. Similarly, Wang and Hou [60] validated the use of NVT simulations for capturing thermal trends in density and vaporisation enthalpy. Moreover, Mohebbi et al. [61] confirmed that equilibrium NVT trajectories remain suitable for identifying the transition-like behaviour in complex fluids.

OPLS-AA force field was used to validate the capacity of MD simulations in estimating the boiling point of biphenyl because it has shown accuracy in reproducing their thermophysical parameters. Thus, this section focuses on assessing the boiling point predictions obtained from MD simulations using two different temperature ranges: OPLS-AA #1 (300–600 K) and OPLS-AA #2 (300–700 K). Both simulations were conducted for the same total duration of 1.25ns, ensuring consistent simulation conditions and enabling direct comparison of the effect of different temperature ranges on the predicted boiling point. Fig. 8(a) and (b) show the density profiles for OPLS-AA #1 and OPLS-AA #2, respectively. In all circumstances, the density falls gradually with the increasing temperature until a clear change indicates the beginning of phase change. Both simulations show a density value of about 0.06 g/cm³, implying that the phase transition from liquid to gas has been detected. However, the two scenarios experience different temperatures at which this density threshold is attained: The boiling point of OPLS-AA #1 is found at 525.66 K with a density of 0.0621 g/cm³, but the boiling point for OPLS-AA #2 moves higher to 534.19 K with a density of 0.0684 g/cm³.

To support the physical validity of MD-estimated boiling point and further justify the threshold-based results, an additional

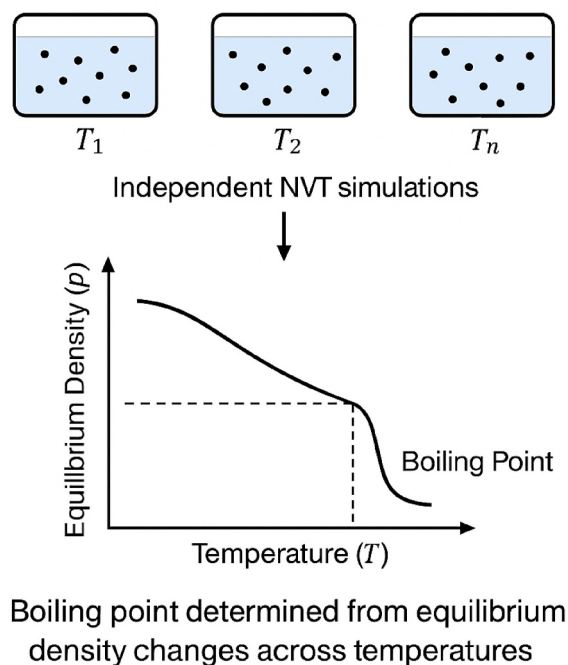


Fig. 7. Schematic workflow for estimating the boiling point from single-phase simulations. Independent canonical (NVT) simulations are performed at a series of fixed temperatures $T_1, T_2, \dots, T_n$. The equilibrium density obtained at each state point defines the isothermal $\rho - T$ profile. The boiling onset is identified with the inflection (largest negative slope) in this curve, corresponding to the temperature at which the liquid density begins its rapid collapse.

thermodynamic analysis is provided based on the curvature of the temperature-related density curve. Fig. 9 presents the spline-fitted $\rho - T$ trend obtained from MD data along with its first-order derivative $d\rho/dT$. The boiling onset is identified at the inflection point corresponding to the minimum of the derivative, which marks the temperature while the maximum density declines. This inflection-based approach does not rely on an arbitrary density threshold but instead reflects the underlying thermodynamic instability of the liquid phase. The inflection temperature is found to be 508.18 K, while the boiling point in MD, which was previously identified at a density threshold of 0.0621 g/cm^3 , is 525.66 K, resulting in a relative error of approximately 3.3 %. This level of deviation is within the typical uncertainty range for single-phase MD-based boiling point estimations, particularly in systems where interface effects are not explicitly resolved. While the inflection-based temperature represents the onset of boiling, the density threshold corresponds more closely to the completion or stabilisation of the boiling transition in NVT system. This distinction is consistent with MD thermodynamic expectations: the inflection reflects the earliest point of rapid structural destabilisation, whereas the threshold method tracks the effective disappearance of the liquid phase.

Fig. 10 presents a comparison between the predicted boiling point in MD with experimental data gathered from the NIST dataset [17,19,51–57]. The experimental values in Fig. 10 span 524–538 K, reaching good agreement with the predictions using both OPLS-AA #1 and OPLS-AA #2. This shows that the boiling point is still accurately predicted within the expected interval even with an expanded temperature range in OPLS-AA #2. Furthermore, the density threshold of $\sim 0.06 \text{ g/cm}^3$ found in both cases, corresponds with the expected behaviour of molecular systems experiencing a liquid-to-gas phase transition. As the temperature increases toward the boiling point, intermolecular forces become weaker, and molecules transition from the condensed liquid phase to the gaseous state, leading to a rapid decrease in density. This density drop signifies that the system has reached a threshold at which boiling occurs [50]. Although direct experimental validation of this exact density value for biphenyl at its boiling point is limited, consistency of this threshold in both simulations support its dependability as an indication of phase change. At high temperatures, the increased thermal energy enhances molecular motion, allowing molecules to overcome intermolecular forces and leading to an increase in intermolecular distance. This expansion of the fluid results in a phase transition, where molecular interactions are weak, illustrating that the system enters the vapor phase. The observed density reduction reflects this transition, indicating that the molecular arrangement shifts toward a less compact structure that characterizes the gas phase. These results demonstrate that the OPLS-AA force field not only accurately captures the thermodynamic characteristics of biphenyl but also efficiently predicts its boiling point within the expected experimental range, hence strengthening the robustness of the chosen simulation parameters. The corresponding values of boiling point from MD simulations and experimental sources shown in Fig. 10 are listed in Table S1 in Supporting Information (SI) for reference and accessibility.

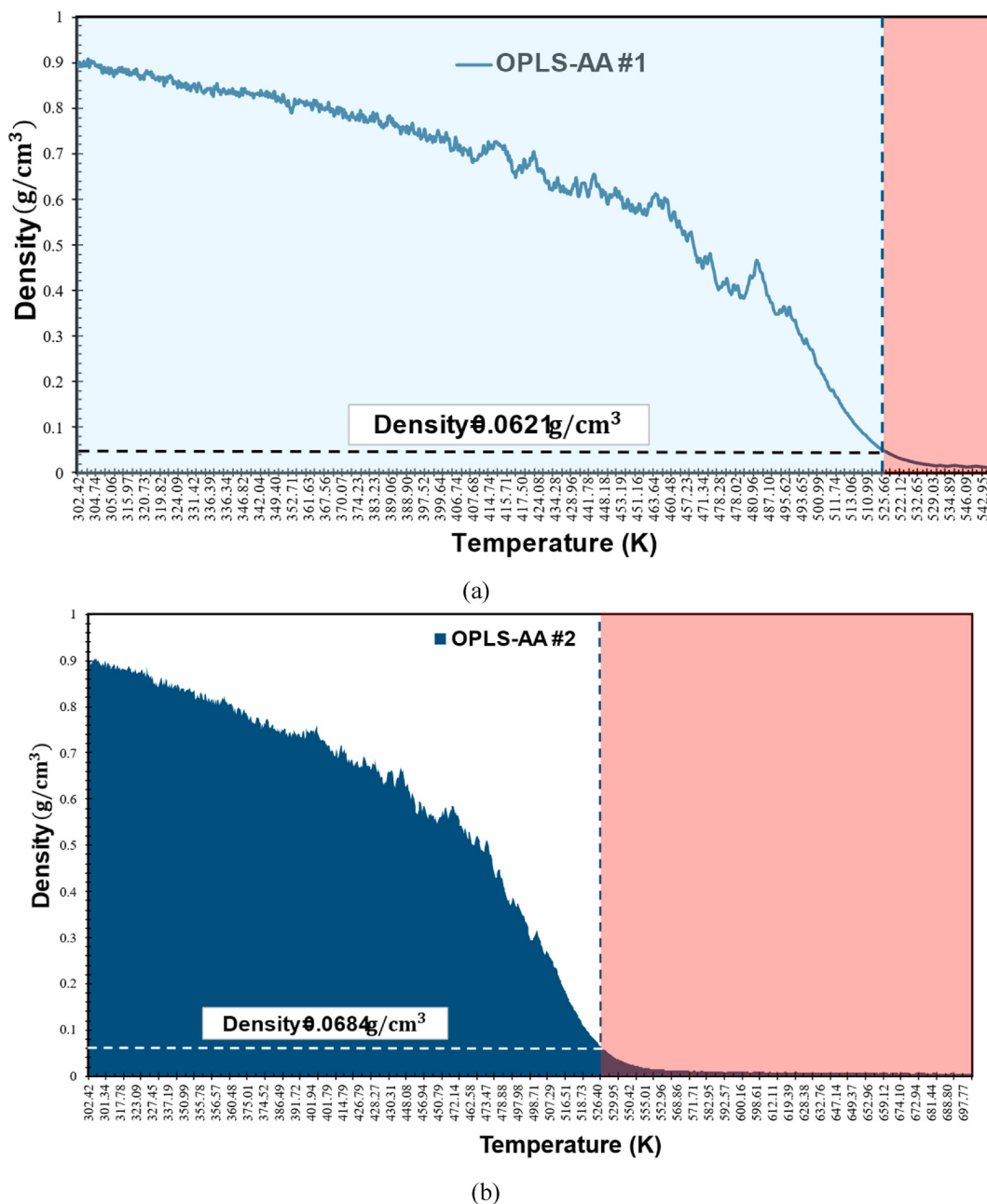


Fig. 8. Density variation of biphenyl using OPLS-AA force field at different temperature ranges: (a) OPLS-AA #1 with a temperature range of 300–600 K, detecting the boiling point at 525.66 K with a density of 0.0621 g/cm³ and (b) OPLS-AA #2 with a temperature range of 300–700 K, detecting the boiling point at 534.19 K with a density of 0.0684 g/cm³. The red-shaded region represents the post-boiling phase where density approaches to the values of gaseous phase.

4.3. ML predictions

4.3.1. Evaluation of ML models

Performances of three ML models, Nearest Neighbours Regression (NNR), Neural Network (NN), and Support Vector Regression (SVR), were assessed in temperature prediction using the MD simulation data. Table 4 shows the main performance metrics used to evaluate the predicted accuracy and dependability of every model in terms of primary metrics such as MAE, MSE, PC and R^2 . Along with the greatest PC (0.99944) and R^2 (0.99887), NNR showed the best accuracy among the evaluated models, attaining the lowest MAE (1.88681) and MSE (5.95195). This suggests that NNR is well-suited to capturing the combined thermodynamic behaviour of the system from multiple physical features. Its strong performance comes from its ability to learn local trends in the data without assuming

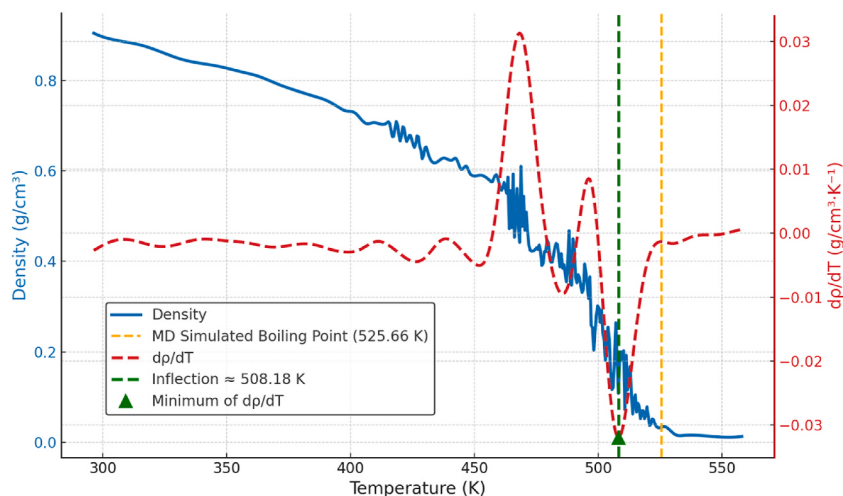


Fig. 9. Thermodynamic inflection-point detection of boiling onset from MD-derived density profile. Spline-fitted density curve (blue solid line), first-order derivative dp/dT (red dashed line), thermodynamic inflection point (green dashed line and filled triangle), and MD boiling point based on density threshold (yellow dashed line).

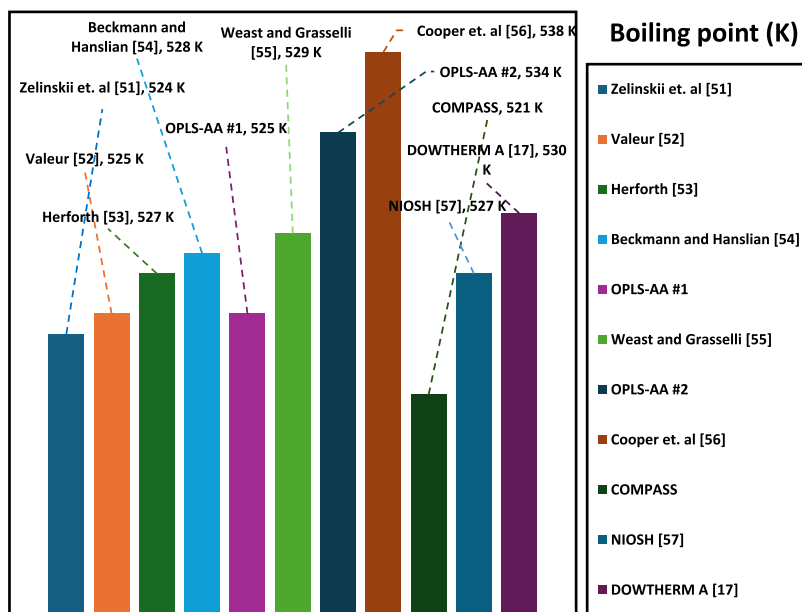


Fig. 10. Comparison of the boiling point of biphenyl between several force fields in MD simulations and experimental values. While COMPASS underestimates the boiling point, OPLS-AA #1 and OPLS-AA #2 predict boiling points within the confirmed experimental range of 524–538 K, closely corresponding with values from NIST and previous investigations [17,19,51–57].

Table 4

Performance metrics of ML models for prediction of system temperature.

Output	ML Model	Metrics			
		MAE	MSE	PC	R ²
Temperature	NNR	1.88681	5.95195	0.99944	0.99887
	NN	3.38837	16.39038	0.99958	0.99688
	SVR	12.3162	180.71225	0.99489	0.96557

a fixed mathematical form between inputs and the target. With higher MAE (3.38837) and MSE (16.39038) than NNR, the NN model also performed well. The NN model can efficiently understand intricate nonlinear interactions inside the dataset according to PC (0.99958) and R^2 (0.99688). Its lower accuracy in comparison with NNR, however, implies that NN model would need more training data or hyperparameter adjustment to produce higher performance. With greater MAE (12.3162) and MSE (180.71225) and comparatively reduced PC (0.99489) and R^2 (0.96557), the SVR model displayed the lowest performance. Its reduced accuracy implies that this model finds it difficult to generalize well for this dataset, probably because of the choice of kernel or constraints in capturing the fundamental data distribution. The bigger error figures suggest that SVR might not be the best option for temperature prediction in this work.

Fig. 11 displays a comparison of the predicted values of system temperature by the three models against the actual temperature readings (e.g., MD results). NNR's great predictive accuracy is confirmed by the data points, which most nearly match the ideal prediction line. While the SVR model shows more departure from the ideal trend, therefore illustrating its poorer performance, which is closely followed by NN. These results show the need of choosing a suitable ML model for the temperature prediction. NNR has been selected for this work since its better performance shows high efficiency in managing the structured relationships in dataset. Although the NN model also shows promise, its accuracy may need further improvement by extra tuning. Moreover, SVR's performance implies that different strategies could be required to raise its predictive capacity in this setting.

For NNR, SVR, and NN, the learning curves in Fig. 12 show the link between training data size and model error. These results help clarify the generalizing capacity and learning efficiency of each model, further supporting the conclusions drawn in Table 4 and Fig. 11. As the amount of training data rises, the NNR model in Fig. 12(a) clearly and steadily shows a lower trend in both training and test errors. The smooth convergence in two error curves points to NNR's effective generalizing without notable overfitting. This is in line with founding in Tables 4 and in which NNR obtained the highest R^2 (0.99887) and the lowest MAE (1.89681) and MSE (5.95195). Moreover, Fig. 11 validates this tendency since the results by NNR are quite near the ideal prediction line, thereby supporting its accuracy for temperature prediction. Even as the magnitude of the training data grows, the SVR model in Fig. 12(b) shows notable fluctuations in test error. The test error does not steadily drop, unlike NNR, suggesting inadequate generalization and possible susceptibility to hyperparameter choice. While training error stays quite constant, the unpredictable test error points to SVR overfitting to patterns in the training data rather than learning a broader relationship. This observation is also in line with Tables 4 and in which SVR has the lowest R^2 (0.96557) together with the highest MAE (12.3162) and MSE (180.71225). Deviating further from the ideal prediction line, the predictions by SVR in Fig. 11 show greater variation than those of NNR and NN, thereby demonstrating its poorer predictive performance. The NN model in Fig. 12(c) shows that the test and training errors vary at smaller data percentages but progressively stabilize as more training data is used, therefore demonstrating progressive learning and lower overfitting. This trend is more consistent than SVR but partially less accurate than NNR since NN attained an MAE of 3.38837 and an MSE of 16.39038 in Table 4. Fig. 11 further indicates that, although with slightly higher variance than NNR, predictions by NN closely track the ideal prediction line, hence strengthening their modest dependability in temperature prediction.

4.3.2. Boiling point prediction by ML

Fig. 13 compares the predicted temperatures from three ML models (NNR, NN, and SVR) with the corresponding MD simulation densities across the investigated temperature range. This comparison evaluates whether the predicted temperatures reflect physically meaningful density–temperature behaviour near the boiling point. These models were trained to predict the simulation temperature from thermodynamic input features, and this figure indirectly illustrates how the predicted temperatures align with known physical trends, such as the decline in density with increasing temperature. Especially in the boiling-point region, where a rapid drop in density

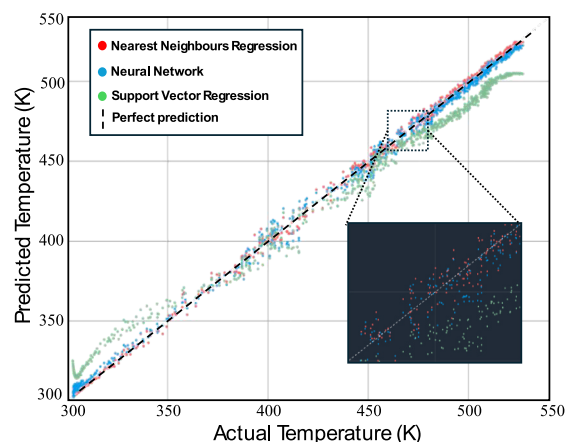


Fig. 11. Comparison of ML model predictions (NNR, NN, SVR) vs. actual value (MD simulations) for system temperature. The perfect prediction line represents an ideal match between predicted and actual values. The insets highlight areas where model performance varies significantly, particularly for SVR.

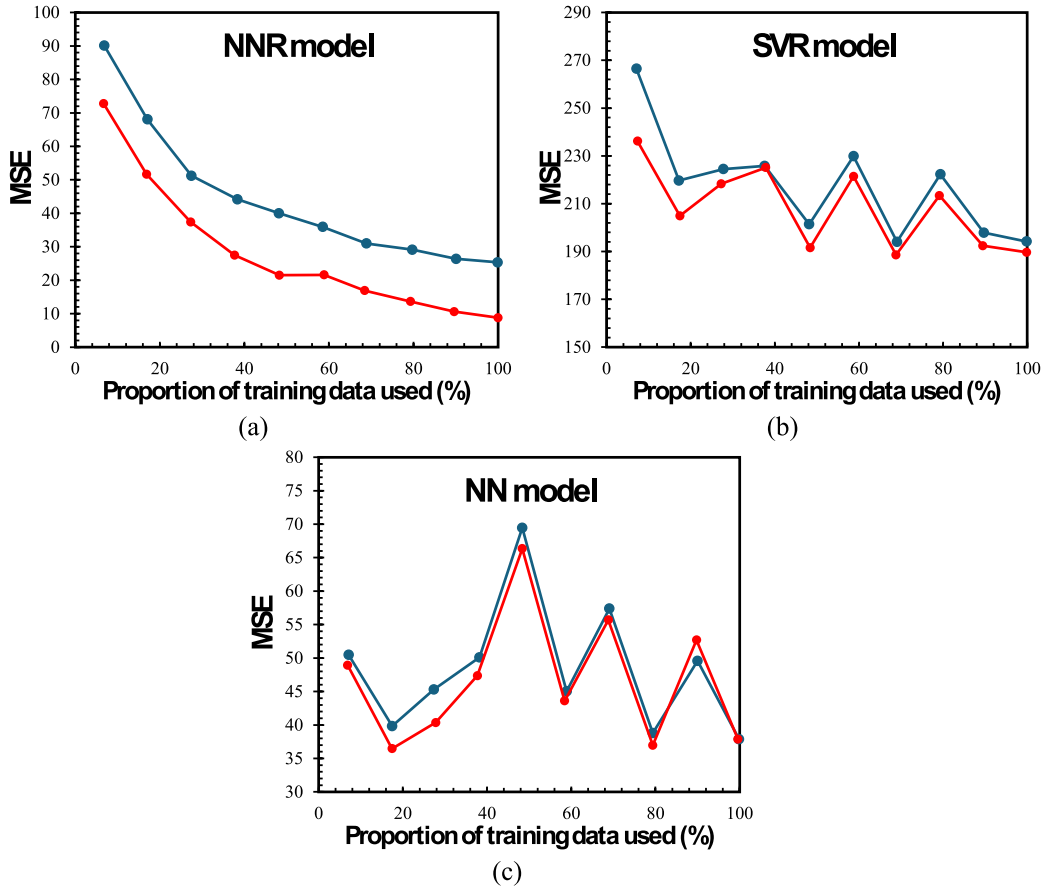


Fig. 12. Learning curves in ML models: (a) NNR; (b) SVR; and (c) NN. The test set error (blue) and training set error (red) reveal information on the generalizing capacity as well as learning efficiency of each model.

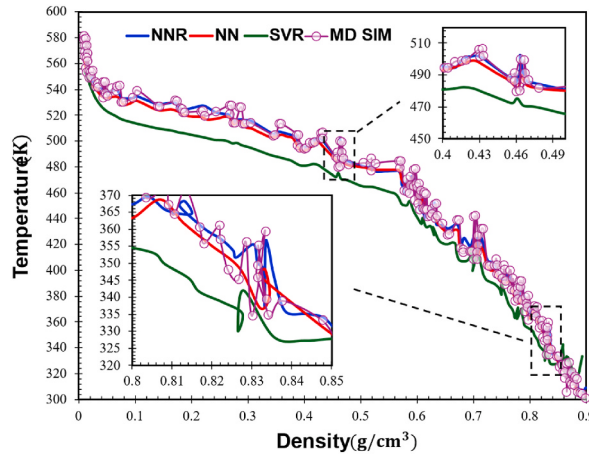


Fig. 13. Comparison of ML-predicted temperatures (NNR—blue line, NN—red line, SVR—green line) mapped onto known MD simulation density data (purple circles) across the investigated temperature range. This visualization tests whether the predicted temperatures preserve the physical density–temperature trend observed in MD. The top inset highlights the transition region near boiling, where NNR and NN most accurately capture phase-sensitive density changes. The bottom inset shows the general deviation between ML models and MD simulations.

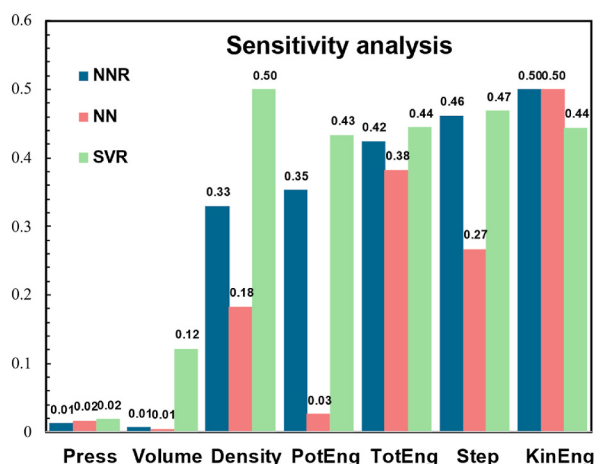


Fig. 14. Sensitivity analysis of ML models (NNR–blue bars, NN–pink bars, and SVR–green bars): effects of parameters on the temperature prediction.

occurs, this comparison allows for an indirect assessment of the models' ability to reflect phase-sensitive behaviour. In the liquid-phase regime, the temperatures predicted by NNR and NN, when mapped back onto the known MD ρ – T curve, correspond closely to the expected densities, indicating that these models are aligned with physical phase behaviour. The SVR model, however, predicts temperatures that, when matched against MD density values, correspond to higher-than-expected densities near the boiling region, suggesting weaker sensitivity to phase transition cues. The upper inset (470–510 K) highlights how NNR more accurately tracks the steep density decline near boiling, consistent with its better generalization seen in Table 4. Minor fluctuations in NNR and NN predictions may stem from noise inherent in complex thermodynamic signals near the phase boundary. Nonetheless, the alignment with MD densities suggests that these models effectively preserve key thermal patterns without requiring explicit phase labelling. These observations complement the regression metrics in Figs. 11 and 12, where NNR consistently shows the highest predictive accuracy and lowest error. While kinetic energy is known to correlate with temperature by first principles, the parallel coordinate graphs, PCA and MI analysis (see Figs. 2–4) demonstrate that the model performance is supported by multiple thermodynamic inputs, not only by kinetic energy.

Fig. 14 shows the results of sensitivity analysis, which presents a comparison of the effects of important physical parameters, including pressure (Press), volume (Volume), density (Density), potential energy (PotEng), total energy (TotEng), simulation step (Step), and kinetic energy (KinEng), on the temperature predictions by the NNR, NN, and SVR models. For all the identified important factors such as density (0.33), total energy (0.42), step count (0.46), and kinetic energy (0.50), the NNR model (blue bars in Fig. 14) shows a plausible sensitivity distribution because the kinetic energy is directly linked to temperature-related molecular motion. Furthermore, the large influence of step count is reasonable, as system relaxation and phase stability in MD simulations are affected by time development. The low values of pressure (0.01) and volume (0.01) are consistent with the fact that simulations were carried out in a NVT ensemble, where volume stays constant, and pressure changes are secondary. These findings confirm the robustness of NNR and further demonstrate that it is the most reliable model for the temperature prediction by appropriately prioritizing the kinetic energy over other parameters. Although density (0.18) is weighted lower than expected, the NN model (red bars in Fig. 14) shows a reasonable parameter priority ranking with kinetic energy as the most influential parameter. In the meantime, step count (0.27) and total energy (0.38) acquire higher effect, and NN's decreased sensitivity to density and greater reliance on time evolution (step count) create possible bias. This implies that, despite its balanced weighting across thermodynamic factors, NN would be less successful than NNR in identifying rapid changes in phase behaviour. The SVR model (green bars in Fig. 14) exhibits a different parameter priority ranking: density (0.50) is still the most crucial characteristic; PotEng (0.43), TotEng (0.44), and Step (0.47) have weights either similar or greater effects than KinEng (0.44) does. This contradicts basic thermodynamic principles, as kinetic energy is expected to have the greatest effect on the temperature prediction. Moreover, Volume (0.12) gets higher weight than those in other models; this poses a problem considering that the simulation runs in a NVT ensemble where volume is constant. This also explains SVR's lower performance in past evaluation criteria, indicating that it fails to properly differentiate between the fundamental and secondary thermodynamic impacts (e.g., Table 4 and Fig. 11). Although all MD simulations were conducted under NVT, the volume was not fixed globally across the dataset. Each system was independently equilibrated at 1atm using the NPT ensemble prior to NVT production, resulting in slightly different fixed volumes for each simulation at different temperatures. This allowed volume to appear as a varying input in the dataset, even though it remained constant within each simulation. The elevated sensitivity to volume seen in SVR model (green bars in Fig. 14) reflects a limitation of SVR itself rather than meaningful physical correlation. The elevated importance of volume in SVR model may stem from its known susceptibility to overfitting in small, structured datasets and its sensitivity to feature scaling and correlation effects [61,62]. In contrast, the NNR and NN models more accurately reflected expected thermodynamic behaviour, by assigning higher importance to kinetic and total energies, which are properties directly linked to temperature and phase stability, while appropriately down-weighting volume remains constant in NVT simulations. This discrepancy reinforces our rationale

for comparing different ML models. By including NNR (instance-based), NN (connectionist), and SVR (kernel-based), the study aimed to evaluate how different learning architectures handle phase-dependent thermophysical data. The NNR model consistently offered both high accuracy and physical consistency, confirming its suitability for boiling point estimation in aromatic fluids.

The predictions of boiling points derived from NNR, NN, and SVR models are compared against the MD simulation results in Fig. 15. Evaluating the accuracy of these models depends on agreement with the MD-predicted boiling point (525.66 K) and consistency with the MD-derived density at the phase transition (0.062 g/cm^3). The NNR model predicts a boiling point of 524.97 K together with a density of 0.064 g/cm^3 , which agree well with the MD results, confirming the NNR's great capacity of phase transition prediction. A boiling temperature of 525.3 K predicted by NN model also fits well with the MD simulated result. However, its estimated density, 0.22 g/cm^3 , is higher than the MD-derived density, $\sim 0.062 \text{ g/cm}^3$. Based on the sensitivity analysis in Fig. 14, overestimation of the liquid-phase density near the boiling point by NN is potentially due to its higher sensitivity to step count and total energy. Although the boiling temperature stays accurate, the inaccurate density prediction undermines the reliability of NN in modelling the phase transition accurately. It can be also seen clearly from Fig. 15 that a boiling point of 502.78 K is predicted by the SVR model, which is well below the MD result of 525.66 K. Moreover, its predicted density of 0.16 g/cm^3 is also higher than the MD-derived density of 0.062 g/cm^3 . This further supports the earlier findings in Figs. 11 and 14, where SVR showed inaccurate predictions and an unreasonable ranking in the sensitivity analysis. Underestimating the boiling point implies that SVR struggles to appropriately capture the temperature-density relationship, which is crucial for the phase transition modelling. The most reliable and accurate model is NNR since its predicted density (0.064 g/cm^3) and boiling point (525.97 K) well match the MD simulation results. Although the NN model is still a good substitute, its inflated density prediction at phase transition causes possible inaccuracy in modelling the phase transition process. The SVR model is the least reliable because it underpredicts the boiling point and overpredicts density, clearly demonstrating its limited applicability for phase transition predictions.

The plots of surface prediction displayed in Fig. 16(a–d) reveal how the predicted temperature from the NNR model varies with changes in thermodynamic properties and how these changes elucidate the phase transition behaviour. Fig. 16(a) shows how rising potential energy results in weaker intermolecular forces, thereby promoting phase transition. Density serves as a crucial variable that sets the specific point, at which phase transition occurs. Fig. 16(b) reflects how the density changes during each simulation step. At the beginning of each simulation period, density stays stable, indicating that the liquid phase remains unchanged. As the system approaches to its boiling temperature, it undergoes a swift reduction in density, which signifies the transition from liquid to vapor phase. The model's accuracy is validated by the predicted decrease in density, which aligns with the typical pattern of the actual boiling process. Fig. 16(c) illustrates how temperature rises with fast molecular motion, indicating the direct link between kinetic energy and temperature. This is consistent with the sensitivity analysis results shown in Fig. 14 that kinetic energy has the greatest effect on the temperature prediction. The also demonstrates the effectiveness of NNR model to accurately capture the boiling behaviour by presenting the sharp increase in temperature, marking the transition from liquid to gas phase. Fig. 16(d) illustrates how the predicted temperature responses to changes in pressure and volume. It can be seen clearly that the predicted temperature remains almost unchanged, thereby confirming the sensitivity analysis presented in Fig. 14 that pressure and volume have little effects on the temperature prediction. It also indicates the accumulation of kinetic energy through simulation steps until it hits the boiling threshold. The accumulation of kinetic energy continues rising at a uniform rate until it reaches the necessary level to overcome intermolecular forces that enables boiling. Fig. 16(a–d) further evidence how the NNR model successfully emulates boiling physics by matching both experimental data and MD simulations results.

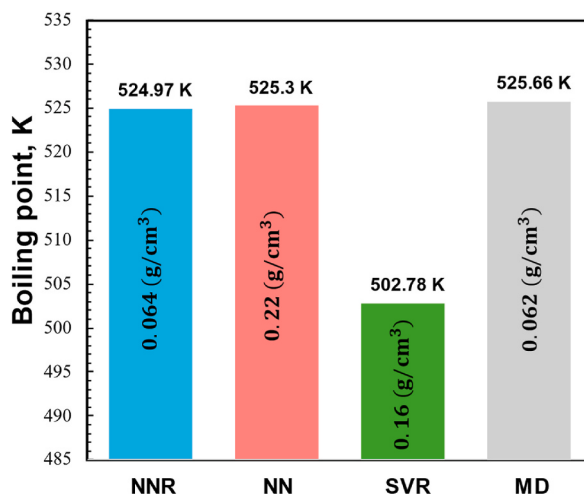


Fig. 15. Predictions of boiling point along with density by ML models compared with MD simulations.

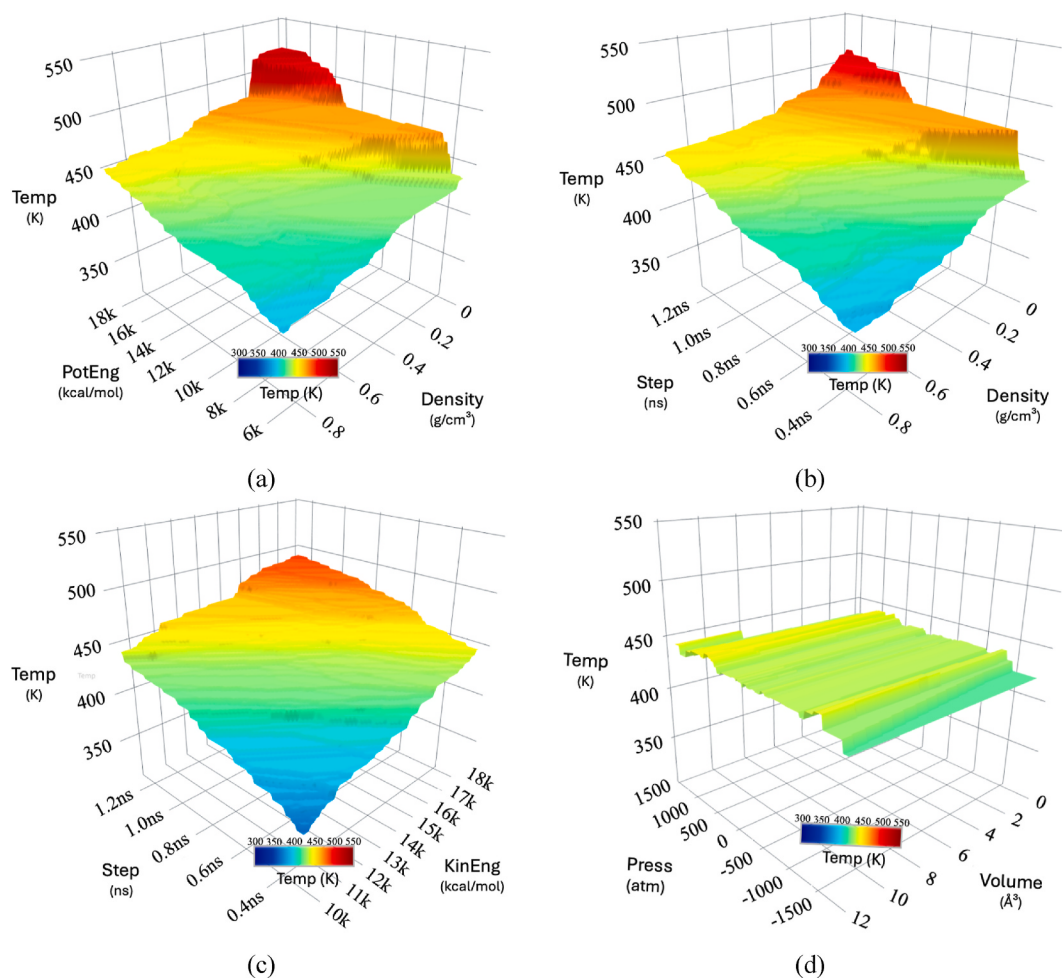


Fig. 16. Surface prediction models display how temperature changes with essential thermodynamic properties to estimate the boiling point: (a) Temperature as a function of density and potential energy; (b) Temperature evolution with simulation step and density; (c) Temperature variation with changes in kinetic energy and simulation step; and (d) Temperature response to changes in pressure and volume.

5. Conclusions

This study presented a hybrid molecular dynamics (MD)–machine learning (ML) framework for boiling point estimation in aromatic fluids, with a particular focus on biphenyl and diphenyl ether. Density predictions using the OPLS-AA force field aligned closely with experimental benchmarks, showing deviations below 2 %, while the COMPASS force field significantly overpredicted densities at elevated temperatures, with errors reaching 94.16 % at 523 K.

Three regression-based ML models were trained using features derived from equilibrium MD data. The Nearest Neighbours Regression (NNR) model demonstrated the best overall performance, predicting a boiling temperature of 524.97 K and a density of 0.064 g/cm³, which are close to the MD-derived values of 525.66 K and 0.062 g/cm³, respectively. Neural Network (NN) predictions also yielded accurate boiling points (525.3 K), but significantly overestimated density. The Support Vector Regression (SVR) model showed the weakest consistency, particularly under phase transition conditions, likely due to its sensitivity to volume, which was a fixed parameter in NVT simulations. These results were reinforced by learning curve of temperature–density relationship and feature sensitivity analysis, which confirmed NNR’s generalisability and robustness near phase boundaries.

In addition to agreement between ML prediction and MD simulation, an inflection-based thermodynamic analysis of the ρ – T curve identified the boiling onset temperature at 508.18 K. Although it is slightly lower than the threshold-based MD estimate (525.66 K), the ~3.3 % discrepancy is consistent with the distinction between onset and completion of boiling in equilibrium MD. The inflection-based point also aligned closely with the ML prediction, validating the physical soundness of the data-driven approach. This MD–ML framework offers an efficient and transferable methodology for predicting phase behaviour in high-performance working fluids. Potential applications include the screening of organic fluids for Rankine cycles, design of heat exchangers, thermal storage, and predictive control of phase-change-driven systems.

The models were trained on thermophysical data from biphenyl and diphenyl ether, which are two structurally similar aromatic

compounds characterised by delocalised π -systems, planar rigidity, and comparable van der Waals-driven phase transition mechanisms. It should be noted that the results are expected to generalize across similar aromatic fluids. However, extension to non-aromatic compounds with different dominant interactions (e.g., hydrogen bonding or strong polarity) will require retraining on a broader thermodynamic space. This will be explored in future work. Future research should broaden the dataset to include a more diverse range of molecular classes beyond aromatic compounds and investigate performance under varying pressure conditions. Integrating chemical reactivity modelling (e.g., via reactive force field in MD simulations) could further extend the framework's scope, enabling the simulation of thermal degradation pathways and intermediate species formation. Incorporating more advanced dimensionality-reduction or feature-selection methods may also help to streamline the ML pipeline, particularly as larger molecular datasets are introduced.

CRedit authorship contribution statement

Amirali Shateri: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Zhiyin Yang:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Formal analysis. **Nasser Sherkat:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis. **Jianfei Xie:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their gratitude to Monolith AI for granting permission to use their Monolith platform. Amirali Shateri would like to acknowledge the University of Derby for the PhD studentship (contract no. S&E_Engineering_0722) and the support provided.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csite.2025.106684>.

Data availability

Data will be made available on request.

References

- [1] S.A. Mostafavi, A. Rezaei, Modelling of the fired preheater of crude oil considering the effect of geometrical parameters on fuel consumption, *Heat Transf.* 51 (2022) 1425–1448, <https://doi.org/10.1002/htj.22358>.
- [2] R.A. Rezaei, M. Khezerloo, Enhancement of thermal efficiency in gas-fired heaters through a novel double-walled chimney design: experimental and CFD analysis, *J. Sustain. Energy* 2 (2023) 109–118, <https://doi.org/10.56578/jse020301>.
- [3] D.D. Mohite, A. Goyal, A.S. Singh, M.I. Ansari, K.A. Patil, P.D. Yadav, P.V. Londhe, Improvement of thermal performance through nanofluids in industrial applications: a review on technical aspects, *Mater. Today Proc.* (2024), <https://doi.org/10.1016/j.matpr.2024.04.083>.
- [4] R. Yang, Z. Deng, B. Li, Y. Chen, Thermal-flow-force-electrical coupling characteristics of microchannel cooling systems in high heat flux chips, *J. Appl. Phys.* 135 (2024), <https://doi.org/10.1063/5.0210958>.
- [5] Z. Li, P. Cheng, Q. Li, P. Cao, Study on the formation and maintenance mechanism of the stationary flame at the head of a LOX/GCH₄ pintle injector element, *Aero. Sci. Technol.* 145 (2024) 108899, <https://doi.org/10.1016/j.ast.2024.108899>.
- [6] A. Shateri, B. Jalili, S. Saffar, P. Jalili, D.D. Ganji, Numerical study of the effect of ultrasound waves on the turbulent flow with chemical reaction, *Energy* 289 (2024) 129707, <https://doi.org/10.1016/j.energy.2023.129707>.
- [7] A. Shateri, Z. Yang, Y. Liu, J. Xie, Data-driven optimization of turbulent kinetic energy and tumble-y in combustion engines: a comparative study of machine learning models, *Fuel* 389 (2025) 134590, <https://doi.org/10.1016/j.fuel.2025.134590>.
- [8] M.M. Awad, *Fouling of Heat Transfer Surfaces*, INTECH Open Access Publisher, London, 2011, pp. 505–542. <http://www.intechopen.com/books/heat-transfer-theoretical-analysis-experimental-investigations-and-industrial-systems/fouling-of-heat-transfer-surfaces>.
- [9] P. Li, C. Shu, J. Li, Y. Wang, Y. Chen, X. Ren, X. Liu, Thermodynamic investigation and economic evaluation of a high-temperature triple organic rankine cycle system, *Energies* 16 (2023) 7818, <https://doi.org/10.3390/en16237818>.
- [10] P. Hosseinfar, H. Shahverdi, Prediction of the ASTM and TBP distillation curves and specific gravity distribution curve for fuels and petroleum fluids, *Can. J. Chem. Eng.* 100 (2022) 3288–3310, <https://doi.org/10.1002/cjce.24335>.
- [11] S. Mukwembli, F. Nyabadza, A new model for predicting boiling points of alkanes, *Sci. Rep.* 11 (2021) 24261, <https://doi.org/10.1038/s41598-021-03541-z>.
- [12] D.S. Stratiev, S. Nenov, I.K. Shishkova, R.K. Dinkov, K. Zlatanov, D. Yordanov, L. Todorova-Yankova, Comparison of empirical models to predict the viscosity of secondary vacuum gas oils, *Resources* 10 (2021) 82, <https://doi.org/10.3390/resources10080082>.
- [13] D.D. Su, X.B. Li, H.N. Zhang, F.C. Li, Molecular dynamics simulation of interfacial heat transfer behaviour during the boiling of low-boiling-point organic fluid, *Int. J. Heat Mass Tran.* 220 (2024) 124962, <https://doi.org/10.1016/j.ijheatmasstransfer.2023.124962>.
- [14] D.D. Su, X.B. Li, H.N. Zhang, F.C. Li, Comparative study of boiling behaviour between liquid sodium and organic fluid under high heat flux: a molecular dynamics simulation, *Int. J. Heat Mass Tran.* 236 (2025) 126324, <https://doi.org/10.1016/j.ijheatmasstransfer.2024.126324>.

- [15] X.W. Lin, L.F. Zhang, Y. Jiang, Y. Liang, Z.F. Zhou, Molecular insights into vibration-induced phase change dynamics of low-boiling-point refrigerant, *Int. J. Heat Mass Tran.* 233 (2024) 126053, <https://doi.org/10.1016/j.ijheatmasstransfer.2024.126053>.
- [16] P. Yu, H. Liu, J. Xue, P. Li, Y. Zhang, D. Che, Assessment of the estimation methods for the normal boiling point of liquid organic hydrogen carriers, *Int. J. Hydrogen Energy* 62 (2024) 904–912, <https://doi.org/10.1016/j.ijhydene.2024.03.131>.
- [17] A. Dowtherm, *Heat Transfer Fluid*, Productive Technology data, 1997.
- [18] D. Szczesna, K. Wiczczonek, J. Jurewicz, An exposure to endocrine active persistent pollutants and endometriosis—a review of current epidemiological studies, *Environ. Sci. Pollut. Control Ser.* 30 (2023) 13974–13993, <https://doi.org/10.1007/s11356-022-24785-w>.
- [19] National Institute of Standards and Technology (NIST), NIST chemistry WebBook: NIST standard reference database 69, 2025-03-10, <https://webbook.nist.gov/cgi/cbook.cgi?ID=C92524&Units=SI&Type=TBOIL>, 2025.
- [20] Y. Xu, X. Huang, C. Li, Z. Wei, M. Wang, Predicting structure-dependent properties directly from the three dimensional molecular images via convolutional neural networks, *AIChE J.* 68 (2022) e17721, <https://doi.org/10.1002/aic.17721>.
- [21] B. Liu, M. Karimi Nouroddin, Application of artificial intelligence approach to predict the normal boiling point of refrigerants, *Int. J. Chem. Eng.* 2023 (2023) 6809569, <https://doi.org/10.1155/2023/6809569>.
- [22] C. Jung, Thermal stability of used eutectic mixture of biphenyl and diphenyl ether, in: AIP Conference Proceedings, 2815, AIP Publishing, 2023, October, <https://doi.org/10.1063/5.0149109>.
- [23] L. Martínez, R. Andrade, E.G. Birgin, J.M. Martínez, PACKMOL: a package for building initial configurations for molecular dynamics simulations, *J. Comput. Chem.* 30 (2009) 2157–2164, <https://doi.org/10.1002/jcc.21224>.
- [24] R.S. Freitas, Á.P. Lima, C. Chen, F.A. Rochinha, D. Mira, X. Jiang, Towards predicting liquid fuel physicochemical properties using molecular dynamics guided machine learning models, *Fuel* 329 (2022) 125415, <https://doi.org/10.1016/j.fuel.2022.125415>.
- [25] S. Shenogin, R. Ozisik, XenoView v. 3.4; Vemmer (2007).
- [26] M.A. Thompson, A QM/MM molecular dynamics study of the potential of mean force for the association of K⁺ with dimethyl ether in aqueous solution, *J. Am. Chem. Soc.* 117 (1995) 11341–11344, <https://doi.org/10.1021/ja00150a036>.
- [27] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, *J. Comput. Phys.* 117 (1995) 1–19, <https://doi.org/10.1006/jcph.1995.1039>.
- [28] A. Stukowski, Visualization and analysis of atomistic simulation data with OVITO—the open visualization tool, *Model. Simulat. Mater. Sci. Eng.* 18 (2009) 015012, <https://doi.org/10.1088/0965-0393/18/1/015012>.
- [29] A. Shateri, Z. Yang, J. Xie, Machine learning-based molecular dynamics studies on predicting thermophysical properties of ethanol–octane blends, *Energy Fuels* 39 (2025) 1070–1090, <https://doi.org/10.1021/acs.energyfuels.4c05653>.
- [30] M. Ihme, W.T. Chung, A.A. Mishra, Combustion machine learning: principles, progress and prospects, *Prog. Energy Combust. Sci.* 91 (2022) 101010, <https://doi.org/10.1016/j.pecs.2022.101010>.
- [31] W. Liu, J. Zha, M. Ling, D. Li, K. Shen, L. Cheng, A boiling point prediction method based on machine learning for potential insulating gases, *Chem. Phys.* 588 (2025) 112447, <https://doi.org/10.1016/j.chemphys.2024.112447>.
- [32] A. Ullah, M. Kilic, G. Habib, M. Sahin, R.Z. Khalid, K. Sanaullah, Reliable prediction of thermophysical properties of nanofluids for enhanced heat transfer in process industry: a perspective on bridging the gap between experiments, CFD and machine learning, *J. Therm. Anal. Calorim.* 148 (12) (2023) 5859–5881, <https://doi.org/10.1007/s10973-023-12083-7>.
- [33] S. Bhanuteja, V. Srinivas, C.V. Moorthy, S. Jai Kumar, B. Lakshmiipathi Lakshmiipathi Raju, Prediction of thermophysical properties of hybrid nanofluids using machine learning algorithms, *Int. J. Interact. Des. Manuf.* 18 (9) (2024) 6559–6572, <https://doi.org/10.1007/s12008-023-01293-w>.
- [34] Y. Gao, I.M. Shigidi, M.A. Ali, R.Z. Homod, M.R. Safaei, Thermophysical properties prediction of carbon-based nano-enhanced phase change material's using various machine learning methods, *J. Taiwan Inst. Chem. Eng.* 148 (2023) 104662, <https://doi.org/10.1016/j.jtice.2022.104662>.
- [35] M. Bakouri, H.S. Sultan, S. Samad, H. Togun, M. Goodarzi, Predicting thermophysical properties enhancement of metal-based phase change materials using various machine learning algorithms, *J. Taiwan Inst. Chem. Eng.* 148 (2023) 104934, <https://doi.org/10.1016/j.jtice.2023.104934>.
- [36] A. Upadhyay, S.K. Hazra, A. Assam, R. Raj, Review of the current status and the potential of machine learning tools in boiling heat transfer, *Numer. Heat Tran., Part B: Fundamentals* 85 (11) (2024) 1489–1532, <https://doi.org/10.1080/10407790.2023.2266770>.
- [37] F. Jirasek, H. Hasse, Perspective: machine learning of thermophysical properties, *Fluid Phase Equilib.* 549 (2021) 113206, <https://doi.org/10.1016/j.fluid.2021.113206>.
- [38] B. Hess, Determining the shear viscosity of model liquids from molecular dynamics simulations, *J. Chem. Phys.* 116 (2002) 209–217, <https://doi.org/10.1063/1.1421362>.
- [39] R. Hafner, G. Guevara-Carrion, J. Vrabec, P. Klein, Sampling the bulk viscosity of water with molecular dynamics simulation in the canonical ensemble, *J. Phys. Chem. B* 126 (2022) 10172–10184, <https://doi.org/10.1021/acs.jpcc.2c06035>.
- [40] M. Panoukidou, C.R. Wand, P. Carbone, Comparison of equilibrium techniques for the viscosity calculation from DPD simulations, *Soft Matter* 17 (2021) 8343–8353, <https://doi.org/10.1039/D1SM00891A>.
- [41] A. Shateri, Z. Yang, J. Xie, Utilizing artificial intelligence to identify an optimal machine learning model for predicting fuel consumption in diesel engines, *Energy AI* 16 (2024) 100360, <https://doi.org/10.1016/j.egyai.2024.100360>.
- [42] S. Gupta, A. Gupta, Dealing with noise problem in machine learning data-sets: a systematic review, *Procedia Comput. Sci.* 161 (2019) 466–474, <https://doi.org/10.1016/j.procs.2019.11.146>.
- [43] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.* (GMD) 7 (2014) 1247–1250, <https://doi.org/10.5194/gmd-7-1247-2014>.
- [44] Y. Mu, X. Liu, L. Wang, A Pearson's correlation coefficient based decision tree and its parallel implementation, *Inf. Sci.* 435 (2018) 40–58, <https://doi.org/10.1016/j.ins.2017.12.059>.
- [45] N.R. Draper, H. Smith, *Applied Regression Analysis*, 326, John Wiley & Sons, 1998.
- [46] Robert D. Chirico, Biphenyl: Experimental and Derived Thermodynamic Properties, National Institute of Standards and Technology, 2025, <https://doi.org/10.18434/nds2-3211>.
- [47] C.H. Byers, D.F. Williams, Viscosities of pure polyaromatic hydrocarbons, *J. Chem. Eng. Data* 32 (1987) 344–348, <https://doi.org/10.1021/jc00049a018>.
- [48] National Center for Biotechnology Information, PubChem compound summary for CID 7583, diphenyl ether. <https://pubchem.ncbi.nlm.nih.gov/compound/Diphenyl-ether>, 2025.
- [49] D. Cabaleiro, M.J. Pastoriza-Gallego, M.M. Piñeiro, J.L. Legido, L. Lugo, Thermophysical properties of (diphenyl ether+ biphenyl) mixtures for their use as heat transfer fluids, *J. Chem. Therm.* 50 (2012) 80–88, <https://doi.org/10.1016/j.jct.2012.02.001>.
- [50] J.F. Xie, S.S. Sazhin, B.Y. Cao, Molecular dynamics study of condensation/evaporation and velocity distribution of n-dodecane at liquid–vapour phase equilibria, *J. Therm. Sci. Technol.* 7 (2012) 288–300, <https://doi.org/10.1299/jtst.7.288>.
- [51] N.D. Zelinskii, I.N. Tits, M. Gaverdovskaya, Data from the National Institute of Standards and Technology (NIST) standard reference database 69, NIST Chemistry WebBook, 1926.
- [52] A. Valeur, Data from the National Institute of Standards and Technology (NIST) standard reference database 69, NIST Chemistry WebBook, 1903.
- [53] L. Herforth, Excitation of fluorescence of organic substances by α -Particles and γ -Rays, *Ann. Phys. (Leipzig)* 69 (1950), 7, 312, Data from The National Institute of Standards and Technology (NIST) Standard Reference Database NIST Chemistry WebBook.
- [54] E. Beckmann, R. Hanslian, Z. Anorg Chem, Data from the National Institute of Standards and Technology (NIST) standard reference database 69, NIST Chemistry WebBook, 1913.
- [55] CRC Handbook of Data on Organic Compounds, in: 2nd Edition (Ed.), Data from the National Institute of Standards and Technology (NIST) Standard Reference Database 69, vol. 1, NIST Chemistry WebBook, Boca Raton, FL, 1989.
- [56] A.R. Cooper, C.W.P. Crowne, P.G. Farrell, Gas-liquid chromatographic studies of electron-donor-acceptor systems, *trans. Faraday soc* 69, NIST Chemistry WebBook, 1967.

- [57] Occupational Safety and Health Administration (OSHA). Chemical data for biphenyl. U.S. Department of Labor. <https://www.osha.gov/chemicaldata/740..>
- [58] A. Trokhymchuk, J. Alejandre, Computer simulations of liquid/vapor interface in lennard-jones fluids: some questions and answers, *J. Chem. Phys.* 111 (18) (1999) 8510–8523, <https://doi.org/10.1063/1.480192>.
- [59] A. Gorfer, C. Dellago, M. Sega, High-density liquid (HDL) adsorption at the supercooled water/vapor interface and its possible relation to the second surface tension inflection point, *J. Chem. Phys.* 158 (5) (2023), <https://doi.org/10.1063/5.0132985>.
- [60] J. Wang, T. Hou, Application of molecular dynamics simulations in molecular property prediction. 1. density and heat of vaporization, *J. Chem. Theor. Comput.* 7 (7) (2011) 2151–2165, <https://doi.org/10.1021/ct200142z>.
- [61] A. Mohebbi, Z. Jayhani, H. Dorrani, H. Bagheri, A new method based on binary mixture concept for prediction of ionic liquids critical properties using molecular dynamics simulation, *Sci. Rep.* 15 (1) (2025) 7545, <https://doi.org/10.1038/s41598-025-91633-5>.
- [62] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.* 14 (2004) 199–222, <https://doi.org/10.1023/B:STCO.0000035301.49549.88>.
- [63] B. Chen, M.G. Martin, J.I. Siepmann, Thermodynamic properties of the williams, OPLS-AA, and MMFF94 all-atom force fields for normal alkanes, *J. Phys. Chem. B* 102 (1998) 2578–2586, <https://doi.org/10.1021/jp9801065>.
- [64] M.L. Price, D. Ostrovsky, W.L. Jorgensen, Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field, *J. Comput. Chem.* 22 (2001) 1340–1352, <https://doi.org/10.1002/jcc.1092>.
- [65] J.F. Xie, S.S. Sazhin, B.Y. Cao, Molecular dynamics study of the processes in the vicinity of the n-dodecane vapour/liquid interface, *Phys. Fluids* 23 (2011), <https://doi.org/10.1063/1.3662004>.
- [66] H. Sun, COMPASS: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds, *J. Phys. Chem. B* 102 (1998) 7338–7364, <https://doi.org/10.1021/jp980939v>.
- [67] H. Sun, COMPASS: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds, *J. Phys. Chem. B* 102 (1998) 7338–7364, [https://doi.org/10.1016/S1089-3156\(98\)00042-7](https://doi.org/10.1016/S1089-3156(98)00042-7).
- [68] S. Wang, K. Hou, H. Heinz, Accurate and compatible force fields for molecular oxygen, nitrogen, and hydrogen to simulate gases, electrolytes, and heterogeneous interfaces, *J. Chem. Theor. Comput.* 17 (2021) 5198–5213, <https://doi.org/10.1021/acs.jctc.0c01132>.
- [69] C. Ji, X. Yang, Y. Ma, H. Chi, J. Xie, Molecular dynamics study on thermophysical properties of K₂CO₃ molten salt-based SiO₂ nanofluids using buckingham potential framework, *J. Mol. Liq.* 407 (2024) 125231, <https://doi.org/10.1016/j.molliq.2024.125231>.
- [70] B.Y. Cao, J.F. Xie, S.S. Sazhin, Molecular dynamics study on evaporation and condensation of n-dodecane at liquid–vapor phase equilibria, *J. Chem. Phys.* 134 (16) (2011), <https://doi.org/10.1063/1.3579457>.