Advancing Vapor Pressure Prediction: A Machine Learning Approach with Directed Message Passing Neural Networks

Yen-Hsiang Lin[a], Hsin-Hao Liang[a], Shiang-Tai Lin[a][b], and Yi-Pei Li*[a][b]*

[a] Department of Chemical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan.

[b] Taiwan International Graduate Program on Sustainable Chemical Science and Technology (TIGP-SCST), No. 128, Sec. 2, Academia Road, Taipei, 11529, Taiwan.

***Corresponding author:** Shiang-Tai Lin: stlin@ntu.edu.tw; Yi-Pei Li: yipeili@ntu.edu.tw

ABSTRACT

Background

Vapor pressure is a critical property in chemical and environmental engineering. Accurately predicting vapor pressure across a range of temperatures is vital for various applications, but traditional methods rely on critical property measurements or quantum mechanical calculations, which can be limiting, especially for new or under-characterized chemicals.

Methods

This study employs a machine learning model based on the directed message passing neural network (D-MPNN) architecture to predict the vapor pressure of organic molecules. Various strategies to incorporate temperature effects into the model are explored to improve prediction accuracy.

Significant findings

The D-MPNN model achieves significantly better accuracy than the traditional PR + COSMOSAC method, with a lower average absolute relative deviation (AARD) of 0.617 compared to 1.36 for the traditional method, using a dataset of 19,079 molecules. The machine learning approach offers a robust alternative that does not require additional critical property data or quantum mechanical calculations.

Keywords: Directed Message Passing Neural Networks (D-MPNN), Vapor Pressure Prediction, Phase Equilibrium

1. Introduction

Vapor pressure prediction is critically important in various scientific, industrial, and environmental fields [1]. It plays a vital role in designing chemical processes [2], maintaining product stability in industries [3], and understanding the behavior of volatile substances [4]. Additionally, vapor pressure significantly affects the release of environmental pollutants [5], the performance of materials at high temperatures [6], and phase equilibria calculations [7]. It also influences chemical reactions [8], safety assessments [9], climate studies [10], energy system design [11], and drug delivery mechanisms [12]. These extensive applications highlight its key role in diverse fields and informed decision-making. However, existing databases like Design Institute for Physical Properties (DIPPR) [13] and the Dortmund Data Bank (DDB) [14] offer limited vapor pressure data for chemicals, and with the rapid discovery of new species, accurate vapor pressure prediction models are essential.

Current methods for estimating vapor pressure include cubic equations of state (EOS), group contribution (GC) methods, and quantitative structure-property relations (QSPR) models. Cubic EOS, exemplified by the Soave-Redlich-Kwong EOS [15] and the Peng-Robinson (PR) EOS [16], offer accurate vapor pressure predictions through simple and computationally efficient equations. However, a limitation of these EOS lies in their reliance on parameters derived from experimental data (critical temperature, critical pressure, and acentric factor), which may not be available for newly discovered or high-temperature unstable compounds. GC methods approach this estimation by deconstructing a molecule into various predefined substructures and summing their contributions to calculate the vapor pressure of the substance [17-21]. QSPR models, extending beyond structural information, incorporate additional molecular descriptors, such as dipole moments, hydrogen bonding parameters, and molar volume, in their vapor pressure calculations [22-25]. A notable shortcoming of many traditional QSPR models is their focus on room temperature conditions, often neglecting the impact of temperature variations on vapor pressure. Recent advancements have addressed this issue by integrating neural networks into QSPR models, thereby enabling the estimation of vapor pressures across a range of temperatures [26, 27]. Furthermore, Tarjomannejad's innovative approach combines GC methods with neural network analysis, enhancing vapor pressure prediction accuracy using acentric factors, critical properties, and molecular structures inputted via a predefined group list [28]. Despite these advancements, current methods encounter predictive challenges, particularly with molecules that exhibit missing

Nomenclature	
AARD	average absolute relative deviation
afp	atomic fingerprints
ALD	average logarithmic deviation
COSMO	conductor-like screening model
D-MPNN	directed message-passing neural network
EE	equation embedded
EOS	equations of state
FFNN	feed-forward neural network
GC	group contribution
mfp-mean	mean-pooling molecular fingerprints
mfp-sum	sum-pooling molecular fingerprints
QSPR	quantitative structure-property relations
R^2	correlation coefficient
RMSE	root mean squared error
TC	temperature concatenated

group parameters. This necessitates the acquisition of additional data and the development of new groups or correction factors to expand the applicability and accuracy of these models [29].

In the realm of vapor pressure prediction for unexplored molecular structures, it is crucial to employ models that do not depend on experimental properties of the species in question, as such data are often unavailable. A quantum mechanically derived force field and workflow for predicting saturated vapor pressure of ALD precursors has demonstrated high accuracy without relying on experimental data, providing a robust method to support the design of novel precursors across diverse chemical systems [30]. Another prominent method fulfilling this criterion is the Conductor-like Screening Model (COSMO) [31]. Based on quantum mechanical implicit solvation calculations, COSMO excels in determining solvation free energy in the liquid phase. Building upon this, Hsieh and Lin have pioneered the PR + COSMOSAC EOS [32]. This novel model integrates COSMO-derived solvation free energies with the PR equation of state. Remarkably, its reliance solely on quantum chemical calculations for inputs allows it to proficiently predict vapor pressures over a range of temperatures and to estimate critical properties, making it exceptionally useful for compounds lacking experimental data. Recent developments indicate that incorporating

experimental data can further enhance the model's performance. Tsai and Lin have advanced the PR + COSMOSAC EOS by integrating experimental boiling points [33]. This enhancement has markedly improved the model's accuracy, achieving an average absolute relative deviation (AARD) of 0.58 for a dataset of 19,081 compounds with boiling points included, compared to an AARD of 1.41 in the absence of experimental data. These results highlight the potential for even greater precision in vapor pressure estimation methodologies.

Recent developments in machine learning have significantly impacted the field of molecular property prediction [34-39], contributing substantially to various domains such as drug design [40], chemical biology [41], retrosynthesis [42-44], and reaction engineering [45-50]. A key breakthrough in this area is the development of techniques that convert complex molecular structures into fixed-length representations [51, 52]. These featurization techniques are generally classified into three main categories, based on the dimensional aspect of the information they process [53]. The first category, one-dimensional representation, employs a linear string format for molecules, exemplified by the Simplified Molecular Input Line Entry System (SMILES) [54]. Techniques utilizing this 1D representation approach treat SMILES as a unique language and apply generic models from Natural Language Processing (NLP), such as transformers or BERT [55-59], for predictive tasks. In the realm of two-dimensional representations, molecules are visualized as graphs with atoms and bonds representing nodes and edges, respectively. This approach leverages various graph-based models, such as graph convolutional neural networks (GCNNs) and message-passing neural networks (MPNNs) [60-64], to extract molecular information. Three-dimensional representations offer a more detailed perspective, incorporating elements like bond lengths, angles, cis-trans isomerism, stereoisomerism, and the spatial arrangement of atoms. This additional detail has been demonstrated to enhance model performance in multiple applications [65-67]. However, employing 3D representations necessitates access to accurate structural data and accounts for the challenge of multiple possible conformations for a given molecule, where different geometries may yield distinct property values. These machine learning-based featurization techniques differ fundamentally from traditional GC or QSPR methods, which depend on manually defined chemical groups or descriptors. Instead, they autonomously extract critical features from molecular structures, enhancing property prediction capabilities. For instance, previous work has successfully applied graph convolutional models to predict species-dependent coefficients of the Antoine equation for vapor pressure prediction across different temperatures, demonstrating the

model's generalization capability to new molecules not included in the training phase [68]. Santana et al. applied similar graph convolutional models to the task of vapor pressure prediction, enhancing performance through a transfer learning technique from boiling point prediction [69].

Motivated by these findings, this study explores the application of graph-based property prediction models combined with embedded empirical equations for vapor pressure prediction. We utilized a directed message-passing neural network (D-MPNN) and conducted a comprehensive investigation of various pooling architectures and techniques to incorporate temperature effects into the model. This approach enabled us to develop a model that can learn molecular features essential for accurate vapor pressure predictions, without requiring additional inputs such as critical properties, acentric factors, or manually crafted descriptors that are typically needed in traditional models. To evaluate the performance of the D-MPNN model, we compared it with the PR + COSMOSAC method [33] and SIMPOL.1 [21]. While the PR + COSMOSAC method does not depend on experimental critical properties, it requires quantum mechanical calculations for each molecule. SIMPOL.1 is a group contribution method that predicts vapor pressure and enthalpy of vaporization of organic compounds as functions of temperature, without the need for experimental critical properties. In our experiments, the D-MPNN model outperformed the PR + COSMOSAC method without the inclusion of experimental boiling temperatures, achieving an impressive AARD of 0.617, and it also surpassed the result of SIMPOL.1. This result is comparable to the enhanced PR + COSMOSAC method that incorporates experimental boiling points [33]. The results demonstrate that the D-MPNN architecture can achieve accurate vapor pressure predictions with only molecular structures (atom connectivity) as input. This study highlights the potential of machine learning as a promising approach for accurate and efficient vapor pressure prediction without resorting to computationally expensive quantum chemical calculations or experimental data.

2. Methods

2.1. Model architecture

In this work, the employed model comprises two distinct components: the encoder block and the readout block, as illustrated in Fig. 1. The encoder block utilizes the D-MPNN architecture to transform the molecular structure into a vector representation. The D-MPNN algorithm begins with transforming molecular SMILES strings into molecular graphs, where atoms are represented

as vertices and bonds as edges. Initial atom and bond features are encoded based on properties such as atomic number, bond type, and stereochemistry. Directed edges are created to facilitate information flow, with feature vectors derived by concatenating atom and bond features. The D-MPNN updates directed edge features iteratively over a specified number of message-passing steps, allowing information transfer between neighboring edges. The hidden states of edges are aggregated to form atomic embeddings, which are subsequently used to generate molecular embeddings by summing over all atomic embeddings. Details of the D-MPNN architecture are elaborated in the work of Yang et al. [63] and Heid et al. [64].

As depicted in Fig. 2, these atomic hidden vectors can be combined through either sum-pooling or mean-pooling to form two types of molecular fingerprints: the sum-pooling molecular fingerprint (mfp-sum) and the mean-pooling molecular fingerprint (mfp-mean). These fingerprints represent the holistic vector representation of the molecule and are subsequently fed into a feed-forward neural network (FFNN) to predict molecular properties. Alternatively, the atomic hidden vectors can be treated as atomic fingerprints (afp) and directly inputted into the FFNN to estimate the contribution of individual atoms to the molecular property [35]. The aggregate of these atomic contributions yields the overall property of the molecule. Previous studies have highlighted that the choice of fingerprint representation (mfp-sum, mfp-mean, and afp) significantly influences the performance of the model [35], necessitating careful selection based on the targeted molecular property. In this study, we carefully evaluate the efficacy of these fingerprint representations in predicting vapor pressure, an aspect that, to our knowledge, has not been extensively explored previously.

Another challenging aspect of vapor pressure prediction is its dependence on temperature, which also distinguishes it from many other molecular property predictions. This necessitates modifying the architecture of molecular property prediction models to incorporate both molecular structure and temperature as inputs. In this study, we investigate two model architectures designed to integrate temperature effects into vapor pressure prediction: the Equation Embedded (EE) model (Fig. 1(b)) and the Temperature Concatenated (TC) model (Fig. 1(c)).

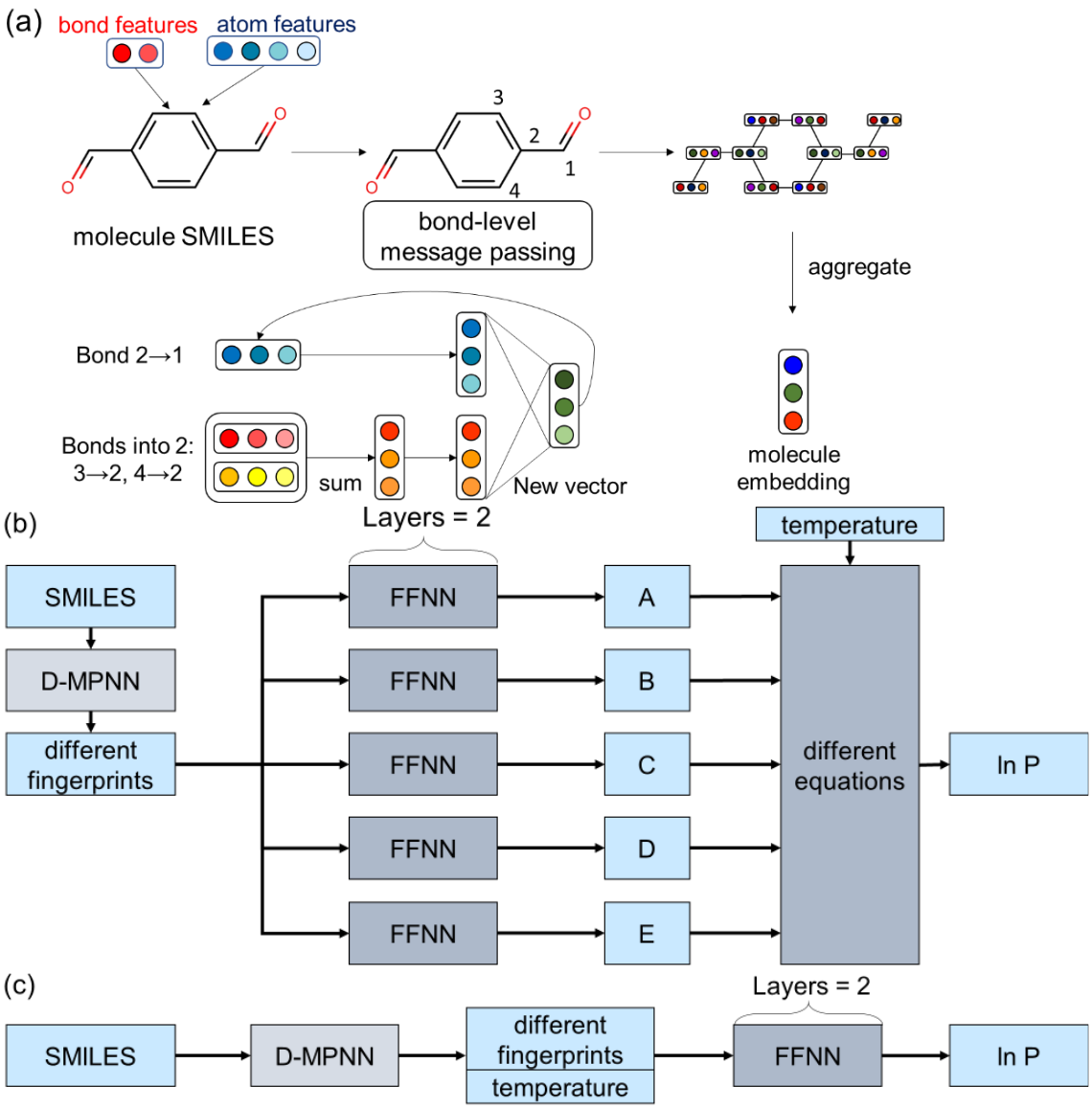

Fig. 1. The algorithm of D-MPNN and two approaches for integrating temperatures into vapor pressure prediction. (a) Workflow of the D-MPNN where molecular SMILES are converted into graphs, features are updated through message-passing steps, and embeddings are aggregated to generate molecular representations. (b) The equation embedded (EE) approach involves using the machine learning model to predict the constants (A, B, C, D, and E) of different empirical equations describing the link between vapor pressure and temperature. (c) The temperature concatenated (TC) model involves directly appending the temperature to the fingerprint representation generated by D-MPNN. This augmented input is then fed into a FFNN to predict vapor pressure at the specified temperature.

The EE model integrates the machine learning model with the empirical equations to capture the relationship between vapor pressure and temperature. This approach allows the machine learning model to predict coefficients for the empirical equations based on molecular structures alone, thereby eliminating the need to incorporate temperature as a direct input. The equations we considered in this work include the Antoine equation [70]

$$\ln P = A - \frac{B}{T + C}, \tag{1}$$

the second-order group contribution equation derived from the Clausius-Clapeyron equation by Tu in 1994 [17]

$$\ln P = A + \frac{B}{T_s} - C \ln T_s - DT_s - \ln M, \tag{2}$$

the Riedel equation [71]

$$\ln P_r = A - \frac{B}{T_r} + C \ln T_r + DT_r{}^6, \tag{3}$$

and the Wagner25 equation [72]

$$\ln P_r = \frac{A(1 - T_r)^{1.0} + B(1 - T_r)^{1.5} + C(1 - T_r)^{2.5} + D(1 - T_r)^{5.0}}{T_r}, \tag{4}$$

where P is the vapor pressure, T is the temperature, T_s is the scaled temperature, M is the molecular weight, P_r is the reduced pressure, T_r is the reduced temperature, and A, B, C, and D are the empirical parameters. To refine the predictive accuracy of vapor pressure calculations without relying on critical property data, we adapt the Riedel and Wagner equations (Eqs. 3 and 4) by substituting $\ln P_r$ with $\ln P - \ln P_c$, where P_c represents the critical pressure. This modification allows the $-\ln P_c$ term to be incorporated into the constant A in the Riedel equation (Eq. 3) and introduced as an additional empirical constant E in the Wagner25 equation (Eq. 4). Furthermore, we replace the temperature T in the Antoine equation (Eq. 1) and the reduced temperature T_r in both the Riedel and Wagner25 equations with a scaled temperature T_s. This scaled temperature is obtained by dividing the actual temperature by the highest temperature value (T_h) in the dataset ($T_s = T/T_h$) to ensure that all scaled temperatures are within a normalized range of zero to one, which enhances the model's training stability [73, 74]. The reformed Antoine equation

$$\ln P = A - \frac{B}{T_s + C} \tag{5}$$

adapts the original equation by utilizing T_s, allowing for the normalization of temperature effects. The reformed Riedel equation

$$\ln P = A - \frac{B}{T_s} + C \ln T_s + D{T_s}^6 \tag{6}$$

and the reformed Wagner equation

$$\ln P = \frac{A(1 - T_s)^{1.0} + B(1 - T_s)^{1.5} + C(1 - T_s)^{2.5} + D(1 - T_s)^{5.0}}{T_s} + E, \tag{7}$$

include T_s to remove the dependency on critical temperature measurements. The coefficients (A, B, C, D, and E) for these equations (Eqs. 2, 5, 6, and 7) are derived using machine learning techniques, with models trained on molecular structures to predict these parameters. These derived constants, in conjunction with the scaled temperature T_s, are then employed to estimate vapor pressure across various temperature conditions.

In addition to the EE approach, we also explore the TC model, which introduces temperature directly into the machine learning framework to predict vapor pressure at different temperatures. As illustrated in Fig. 1(b), the TC model strategy involves appending the temperature of each data point to its molecular fingerprint, thereby expanding the fingerprint vector with an additional temperature dimension. These extended fingerprints are subsequently processed by the FFNN to predict vapor pressure at the given temperature. To ensure uniformity and enhance model stability, temperatures are scaled by the highest temperature value in the dataset, normalizing all values to a range between zero and one. Additionally, in the afp model depicted in Fig. 2(c), the machine learning framework is tasked with independently predicting each atom's contribution to the empirical constants (as per the EE approach) and directly to vapor pressure (as per the TC approach). These atomic-level contributions are then aggregated to derive the empirical constants and overall vapor pressure for the molecule. By comparing the performance of the EE and TC models, our research aims to identify optimal strategies for leveraging molecular structure and temperature data, ensuring accurate vapor pressure estimations across a broad spectrum of conditions.

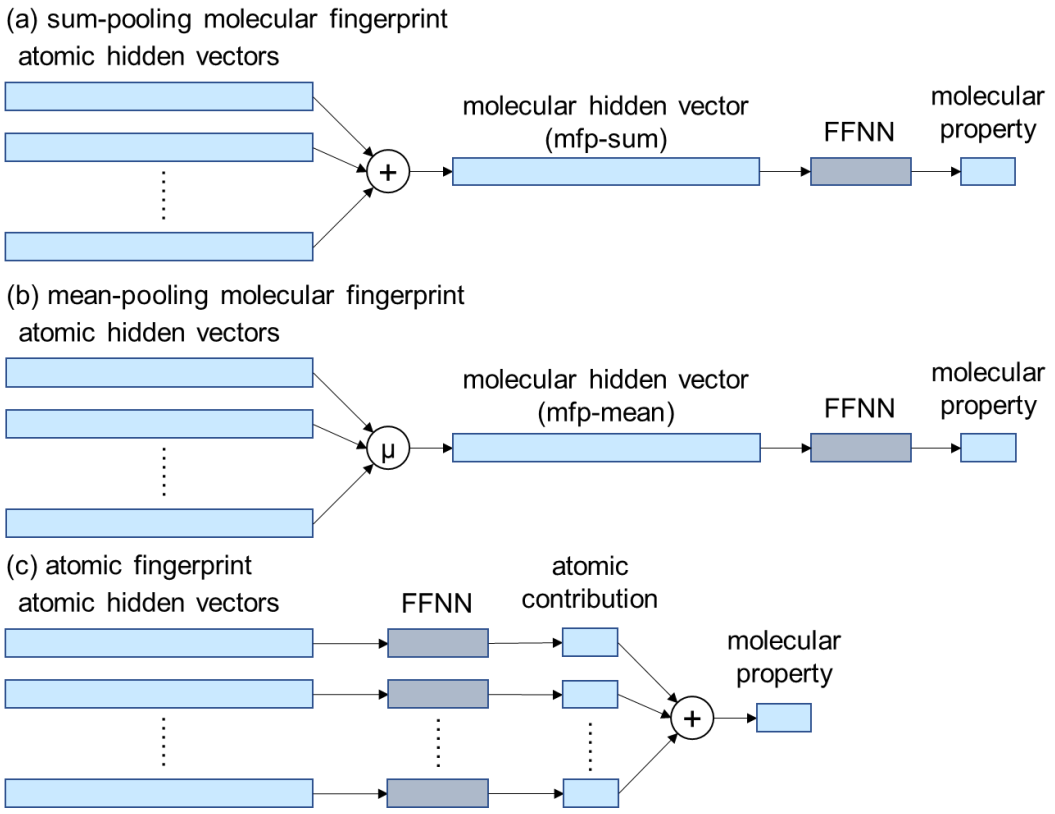

Fig. 2. Schematic representation of the methodology for deriving molecular fingerprints from atomic hidden vectors. Part (a) shows the generation of sum-pooling molecular fingerprints (mfp-sum), and part (b) depicts the creation of mean-pooling molecular fingerprints (mfp-mean), both of which capture the comprehensive vector representation of molecules for subsequent property predictions through a FFNN. Part (c) illustrates the approach of utilizing atomic hidden vectors as atomic fingerprints (afp), which are directly fed into the FFNN to evaluate the individual contributions of atoms to the molecular property of interest.

2.2 Data preparation

Following the methodology of Tsai and Lin [33], we gathered vapor pressure data for 19,079 unique organic molecules from the NIST-TRC databank available in Aspen Plus V11 [75]. This extensive dataset, documented using Wagner25 coefficients [72], critical temperature, and critical pressure, serves as the foundation for our machine learning models. The SMILES representations for all molecules are available on the GitHub repository.

To evaluate the performance of the machine learning model under different scenarios, two data splitting strategies were employed: molecule split and temperature split. The molecule split

strategy assesses the model's ability to generalize to unseen molecules. A ten-fold cross-validation approach was used, where 10% of the molecules in the dataset (19,079 molecules in total) were assigned to the test set in each fold, while the remaining 90% were used for training and validation. For molecules in the training or validation set, vapor pressures were calculated at ten evenly spaced temperatures within the valid range using the Wagner25 equation, resulting in 190,790 data points over all iterations. If the upper temperature limit of the valid range of a molecule surpassed 1,500 K, this limit was adjusted to 1,500 K (T_h=1,500K). It is noteworthy that the choice of a constant value T_h that is larger than most of the compounds in the database is advantageous since it allows one to obtain pseudo vapor pressure beyond the crtical point. The pseudo vapor pressure is important for estimation of Henry's constant of gaseous species in various solvents ($H_{\frac{i}{s}}(T) = \gamma_{\frac{i}{s}}^{\infty}(T)P_i^{vap}(T)$) [1, 76]. We also note that the choice of T_h does not affect the performance of our model, as the testing results with a 1,000 K limit (Table S1 and S2 in the Supporting Information) are very similar to those with a 1,500 K limit (Table 1 and 2). For molecules in the test set, vapor pressures were computed at three specific reduced temperatures: high T (Tr = 0.9), medium T (Tr = 0.65), and low T (Tr = 0.4), providing approximately 51,713 test data points over all iterations after excluding data outside the valid range, detailed in the distribution plots within Fig. 3(a).

On the other hand, the temperature split strategy evaluates the model's performance when extrapolating to temperatures outside the training range. Vapor pressures for all 19,079 molecules were calculated at ten evenly spaced temperatures within the valid range, resulting in 190,790 data points. Data points corresponding to temperatures within the 400 K to 600 K range were used for training and validation (70,031 data points), while data outside this range were assigned to the test set (120,759 data points), with higher temperature data points (≥ 600 K, 56,309 data points) and lower temperature data points (≤ 400 K, 64,450 data points) analyzed separately to evaluate the performance of models when extrapolated to temperatures outside those in the training set. The distribution of these sets is visualized in Fig. 3(b). This split simulates the scenario of predicting vapor pressures at extreme temperatures beyond the training data.

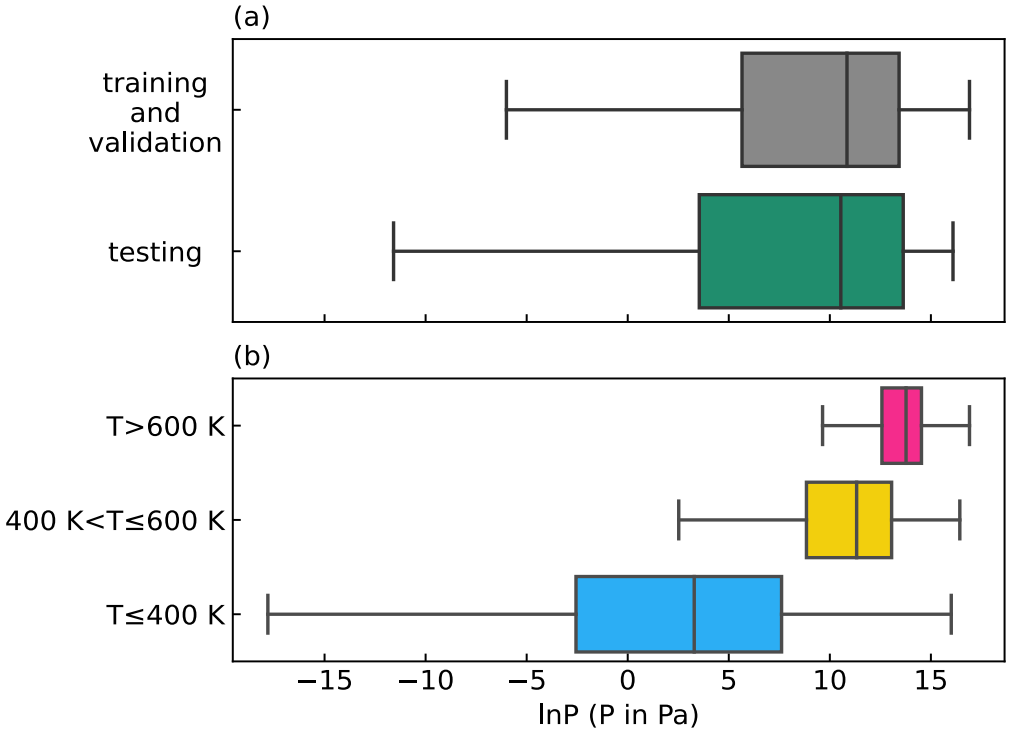

Fig. 3. Distribution of vapor pressure data. Panel (a) displays box plots (indicating minimum, lower quartile Q1, median, upper quartile Q3, and maximum values) for various molecule datasets, segmented by molecule split. Panel (b) shows box plots for distinct temperature ranges, categorized by temperature split.

2.3 Computational details

In the optimization of our machine learning model, we fine-tuned some hyperparameters to ensure optimal performance, leveraging the hyperopt package to navigate the hyperparameter space efficiently [77]. For the D-MPNN, the hyperparameter search space included a depth of message-passing steps from 1 to 5 and a hidden fingerprint size between 100 and 500. For the FFNN, the search covered 1 to 5 layers, with each layer's hidden size ranging from 100 to 500. These ranges were systematically explored to identify the optimal model configuration for our task. After conducting 50 iterative trials, the optimal hyperparameters were chosen based on superior model performance, with the specific configurations detailed in Table S3 of the Supporting Information, while other hyperparameters not explicitly mentioned retained their default values [64]. To bolster prediction accuracy further, we employed a model ensemble strategy. This approach amalgamates the outputs of three separate models, each sharing the same architecture

but differentiated by unique random seed initializations. This diversity in initialization aids in mitigating model bias and variance, leading to more robust predictions.

The training of the models utilized the average logarithmic deviation (ALD) as the loss function, defined as:

$$\mathrm{ALD} = \frac{1}{N}\sum_{i=1}^{N}\left|\ln P_{i,pre} - \ln P_{i,ref}\right| \tag{8}$$

where N represents the total count of vapor pressure data points, with $P_{i,pre}$ and $P_{i,ref}$ denoting the predicted and reference vapor pressure values, respectively. This metric effectively quantifies the deviation between model predictions and actual measurements, guiding the model towards higher precision. To prevent overfitting, an early stopping mechanism halts training if the ALD of the validation set does not improve over 30 consecutive epochs, ensuring the generalizability of the model. The selection of the best model is based on achieving the lowest ALD score on the validation set.

Ten-fold cross-validation was implemented for the molecule split dataset. This method ensures comprehensive model evaluation by guaranteeing that each data point in the testing set is predicted without the model having been exposed to the same molecule in the training or validation phases. Such validation practices are crucial for affirming the ability of the model to accurately predict vapor pressure for unseen molecules, thereby confirming the effectiveness and reliability of the predictive framework. To maintain consistency and enable direct comparisons with existing methods, specifically the PR + COSMOSAC models [33], we adopt the average absolute relative deviation (AARD) as our primary performance metric

$$\mathrm{AARD} = (e^{ALD} - 1). \tag{9}$$

This conversion from ALD to AARD allows for a more intuitive understanding of the prediction error in relative terms, facilitating a straightforward comparison with the established PR + COSMOSAC benchmarks.

To comprehensively evaluate our model's performance, we included additional metrics beyond AARD: the correlation coefficient (R^2), root mean squared error (RMSE), and maximum error. The R^2 is calculated using the following formula:

$$R^2 = \frac{\sum_{i=1}^{N}\left(\ln P_{i,pre} - \ln \bar{P}_{pre}\right)\left(\ln P_{i,ref} - \ln \bar{P}_{ref}\right)}{\sqrt{\sum_{i=1}^{N}\left(\ln P_{i,pre} - \ln \bar{P}_{pre}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\ln P_{i,ref} - \ln \bar{P}_{ref}\right)^2}} \tag{10}$$

where $\ln \bar{P}_{pre}$ and $\ln \bar{P}_{ref}$ represent the average value of the predicted and reference vapor pressure values, respectively. This metric measures the strength and direction of the linear relationship between predicted and actual values, with a value of 1 indicating a perfect positive correlation. It helps assess how well the model captures the trend in data. The RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\ln P_{i,pre} - \ln P_{i,ref}\right)^2} \tag{11}$$

RMSE calculates the square root of the average squared differences between predicted and actual values. It provides insight into the magnitude of prediction errors, with lower values indicating better model accuracy. The maximum error is defined as:

$$Maximum\ error = \max_{i=1\sim N}\left|\ln P_{i,pre} - \ln P_{i,ref}\right| \tag{12}$$

This metric represents the largest absolute difference between predicted and actual values in the dataset, highlighting the worst-case error. It is useful for understanding the model's performance in extreme cases or identifying outliers.

3. Results and Discussion

3.1 Molecule split performance

The testing performances of different model architectures on molecule-split data reveal nuanced insights into their accuracy for predicting vapor pressure across varying molecular structures, as detailed in Table 1. The Antoine EE model showed significantly higher AARD values compared to other models, which aligns with expectations given the known limitations of the Antoine equation across a wide temperature range [70]. This result highlights the restricted utility of the equation in accurately capturing vapor pressure variations. In terms of other temperature input methods, whether incorporating temperature directly or adapting empirical equations, the overall prediction errors were similar across different approaches. This observation suggests that the specific method of temperature integration has a minimal impact on the ability of a model to generalize to new chemical species. However, a consistent trend observed across all models is an increase in AARD as the temperature decreases. This pattern reflects the complexities of accurately measuring and predicting vapor pressure at lower temperatures, where vapor pressure values significantly drop, introducing a greater degree of uncertainty in the reference measurements.

Comparing the mean-pooling and sum-pooling approaches, as well as the atomic fingerprint (afp) model, the mean-pooling method generally yields higher AARD values. This less favorable outcome can be attributed to the mean-pooling function's tendency to average hidden atom features, which results in a loss of crucial information about molecular size [35]. Given that vapor pressure is directly related to molecular size [19], retaining this information is essential for accurate predictions. Conversely, the afp model, which sums the contributions of individual atoms, maintains the proportionality between predicted properties and molecular size but faces limitations. Specifically, the assumption of consistent atom contributions across similar chemical structures does not always hold, especially for vapor pressure predictions where the influence of specific functional groups can vary with molecular size [19]. The sum-pooling approach, on the other hand, inherently accounts for molecular size, enhancing prediction accuracy. Unlike mean-pooling, it does not dilute molecule size representation through averaging and does not assume a simplistic additivity of fragment contributions without considering the overall molecular context. As a result, the sum-pooling method often achieves the lowest prediction error among the models evaluated.

In this subsection, the ability of the machine learning model to generalize to unseen molecules was evaluated using a data-splitting method specifically designed to mimic real-world scenarios where the target molecules are entirely absent from the training and validation sets. This approach ensures that the model's predictive performance is assessed on truly novel compounds. While the results demonstrate that the model performs effectively within the chemical space covered by the dataset, it is important to note that its generalization is inherently limited to the diversity of the training data. Molecules with atom types, structural motifs, or functional groups entirely absent from the dataset may present challenges, highlighting the importance of dataset diversity in model development.

3.2 Temperature split performance

In our study, the temperature split analysis was carried out to delve into the performance of models in extrapolating vapor pressure predictions at temperatures not covered in the training set. This evaluation was crucial for understanding how each model architecture manages the challenge of predicting vapor pressure across a broad temperature spectrum, thereby highlighting their respective strengths and limitations in terms of temperature variability.

The results of this analysis, detailed in Table 2, reveal that the Antoine EE models

demonstrated the largest errors among the models tested. This finding aligns with observations from the molecule split analysis, reinforcing that the Antoine equation might lack the necessary robustness to model the temperature dependence of vapor pressure accurately across a wide range of temperatures. The shortcomings of the Antoine equation in this context suggest its limited applicability for tasks requiring extensive temperature range predictions.

Conversely, the mfp-sum Wagner EE model emerged as the most accurate, with an AARD of 0.672, underscoring the reformed Wagner equation (Eq. 7) as the most suitable for integration into a deep learning framework for vapor pressure prediction. The superior performance of the Wagner EE model can be anticipated considering the training set data points were generated using the Wagner25 equation (Eq. 4) based on experimental critical properties. This congruence between the training methodology and the Wagner equation's inherent capacity to encapsulate temperature effects on vapor pressure evidently contributes to the enhanced predictive accuracy of the model.

We note that empirical equations, such as the Wagner25 equation, have long been used to predict vapor pressures with high accuracy by fitting to experimental data. While these equations are highly effective for known compounds, their applicability is limited to cases where the required constants, derived from experimental measurements, are available. This dependency restricts their use for novel compounds, where no experimental data exist. In contrast, the machine learning model presented in this study addresses this limitation by directly linking molecular structure to vapor pressure predictions. By learning patterns from a diverse dataset, the model eliminates the need for predefined constants and enables rapid predictions for new compounds solely based on molecular structures. To assess the model's ability to generalize to unseen compounds, we validated its performance on an independent experimental dataset curated by Santana et al. [69], containing vapor pressure measurements for 1,852 molecules. After filtering out overlaps with our training data, the final test set included 1,273 unique compounds. We evaluated the mfp-sum model's performance across various EE and TC models (Fig. S1 of the Supporting Information). Among these, the Wagner EE model demonstrated slightly superior accuracy, likely due to the greater number of coefficients in the Wagner formula (Eq. 7) compared to other equations. This validation highlights the model's capability to provide meaningful predictions and demonstrates its potential for application to compounds beyond those included in the training dataset.

Table 1. Comparison of AARD from various model architectures for molecule split data. The bold symbols indicate the models with the best performance under each temperature point.

model	Antoine EE			Tu EE			Riedel EE			Wagner EE			TC		
fingerprint	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum
overall	5.207	2.716	4.140	0.653	0.699	0.658	0.670	0.753	0.638	0.661	0.701	**0.617**	0.687	0.722	0.630
Tr = 0.4	8.890	6.493	17.55	1.687	1.765	1.602	1.717	1.896	**1.598**	1.711	1.812	**1.598**	1.842	1.904	1.626
Tr = 0.65	5.825	2.271	2.120	0.484	0.481	0.477	0.497	0.569	0.470	0.492	0.532	0.453	0.502	0.521	**0.447**
Tr = 0.9	2.873	1.413	2.066	0.250	0.323	0.300	0.265	0.313	0.264	0.252	0.265	**0.234**	0.251	0.286	0.257

Table 2. Comparison of AARD from various model architectures for temperature split data. The bold symbols indicate the models with the best performance within each temperature region.

model	Antoine EE			Tu EE			Riedel EE			Wagner EE			TC		
fingerprint	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum	afp	mfp-mean	mfp-sum
overall	2056	61.06	4674	2.183	1.952	5.066	6.252	4.403	9.426	0.695	0.760	**0.672**	3.398	6.205	2.741
T ≤ 400 K	20784	413.5	6293	2.770	2.425	9.224	2.222	2.754	2.851	1.122	**1.011**	1.132	6.211	14.75	3.381
T > 600 K	145.6	6.098	3328	1.624	1.492	2.343	17.32	7.187	31.51	0.312	0.511	**0.266**	1.501	1.949	2.123

3.3 Performance comparison with previously proposed methods

In this section, we focus on evaluating the performance of our optimal model, the mfp-sum Wagner EE model, against three previously proposed methods: PR-pred, PR-pred-expTb, and PR-exp, as delineated by Tsai and Lin [33]. Each of these methods represents a distinct approach to predicting vapor pressure, ranging from reliance on purely computational inputs to the integration of extensive experimental data.

The PR-pred method stands out for its use of quantum mechanical solvation calculations to estimate energy and molecular volume parameters for the PR EOS, eliminating the need for experimental T_c, P_c, and ω [33]. This approach offers a theoretical model that bypasses the requirement for experimental critical properties. Enhancing the prediction accuracy of the PR + COSMOSAC EOS, the PR-pred-expTb method incorporates the experimental normal boiling point into the model [33]. This addition aims to improve the model predictions by using one experimental vapor pressure data at the normal boiling temperature. Conversely, the PR-exp method relies on the PR EOS but utilizes experimental values for T_c, P_c, and ω to calculate vapor pressure [16, 33]. This method represents the use of two experimental vapor pressure data points (one at the critical point and the other at a reduced temperature of Tr=0.7), resulting in high accuracy at the cost of more experimental input.

The selection of these four methods for comparison spans a spectrum from models independent of experimental measurements to those heavily reliant on such data. Our analysis targets the same set of testing molecules as utilized by Tsai and Lin [33], allowing for a direct comparison of model performances. However, it is important to note that the original paper used a weighted average AARD, which leads to slight discrepancies between our calculated AARD and the original reported data. This comparative study aims to elucidate the trade-offs between computational predictions and the necessity of experimental measurements in achieving accurate vapor pressure estimations.

Fig. 4 shows that the AARD for the PR model diminishes as the incorporation of experimental data increases, aligning with Tsai and Lin's findings [33]. A key observation from our study is the ability of the mfp-sum Wagner EE model to significantly outperform the PR-pred method (AARD 0.617 vs 1.36), which, like our machine learning approach, does not rely on experimental data inputs. This advantage underscores the potential of machine learning in accurately predicting vapor pressures

for novel chemical species without available experimental measurements. Additional metrics, including R^2, RMSE and maximum error, are provided in Table S4 of the Supporting Information. While the maximum error indicates that the mfp-sum Wagner EE model has larger outliers compared to the PR-pred method, as shown in the first row of the parity plot in Fig. 4, both R^2 and RMSE demonstrate that the mfp-sum Wagner EE model achieves lower overall prediction errors. We also compared our model with the group contribution method SIMPOL.1 [21], implemented by Zhao [78], which is another vapor pressure model that does not need experimental critical properties. The performance of SIMPOL.1 is shown in Fig. S2 of the Supporting Information (AARD 102), which is much worse than the other methods.

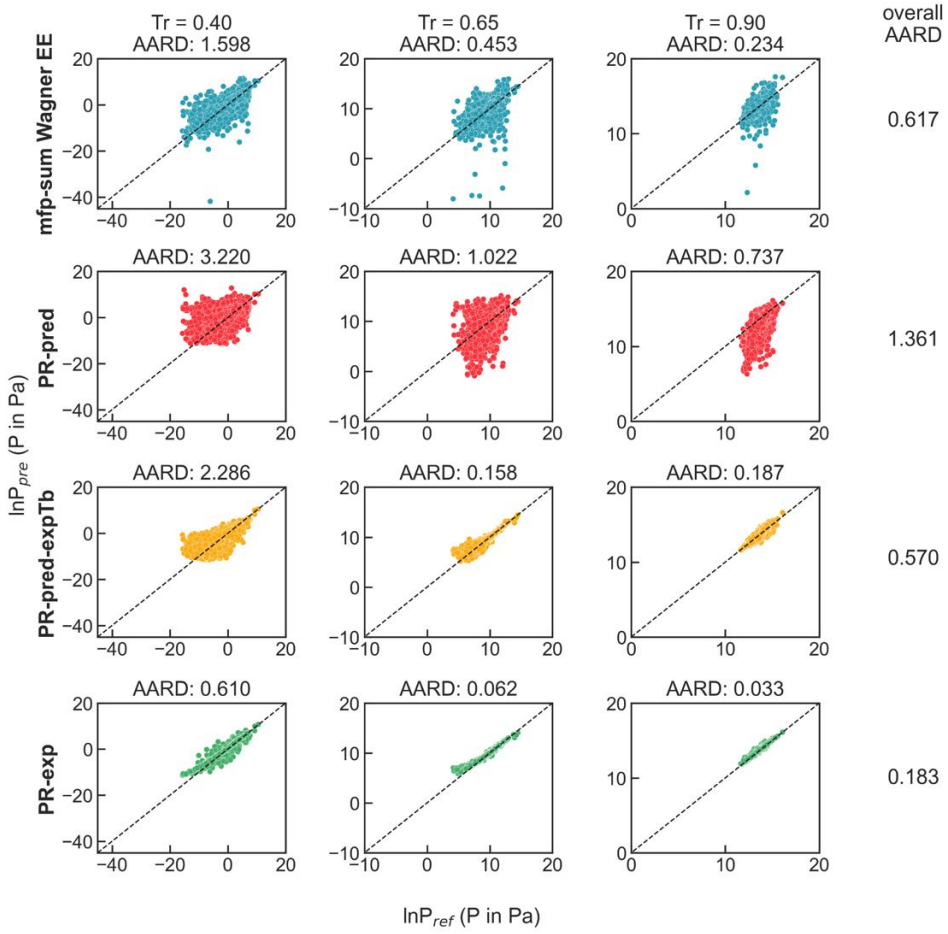

Fig. 4. Parity plots of the reference vapor pressures and the values predicted by four methods, the mfp-sum Wagner EE model, PR-pred, PR-pred-expTb, and PR-exp at three temperatures, Tr = 0.40, Tr = 0.65 and Tr = 0.90. The title of each subplot displays the AARD for that method at a specific temperature, and the right column presents the overall AARD for each method, encapsulating their predictive performance across the

temperature spectrum.

Moreover, the machine learning model demonstrates a remarkable capability by achieving an overall AARD (0.617) nearly on par with that of the PR-pred-expTb method (0.570), despite the latter's reliance on experimental boiling point data. Compared to PR-based methods, this highlights the utility of the machine learning approach in scenarios where direct experimental data are lacking, affirming its value in predicting vapor pressures of new chemical entities. Furthermore, the absence of a need for computationally demanding quantum mechanical calculations for each molecule positions the machine learning model as a viable tool for extensive virtual screening in molecular design projects.

To further understand the model's effectiveness across different molecule types, we divided the test set molecules into three categories: hydrogen bonding (14,725 data points), polar non-hydrogen bonding (19,994 data points), and nonpolar non-hydrogen bonding (16,994 data points). A detailed analysis, as shown in Fig. 5, revealed that the hydrogen bonding subset consistently exhibited the highest prediction error across all models and methods. This was particularly pronounced for the PR-pred model, indicating its limitations in capturing electrostatic interactions. In contrast, the mfp-sum Wagner EE model showed better resilience to the presence of hydrogen bonds and polar molecules, suggesting its partial capability in handling electrostatic interactions. This observation underscores the potential of machine learning techniques in predicting vapor pressures for a diverse molecule set. However, our analysis also uncovered outliers in machine learning model predictions, especially among nitrogen-containing molecules with multiple rings. These outliers increased the AARD for this subgroup, as shown in Fig. S3 of the Supporting Information, highlighting a challenge in accurately predicting vapor pressures for complex structures. This pattern, aligning with prior research [33], suggests avenues for model refinement to improve accuracy. Specifically, the difficulties in predicting properties of nitrogen-containing molecules indicate opportunities for enhancing the model, possibly by incorporating more data or refining model features to better understand such intricate chemical species.

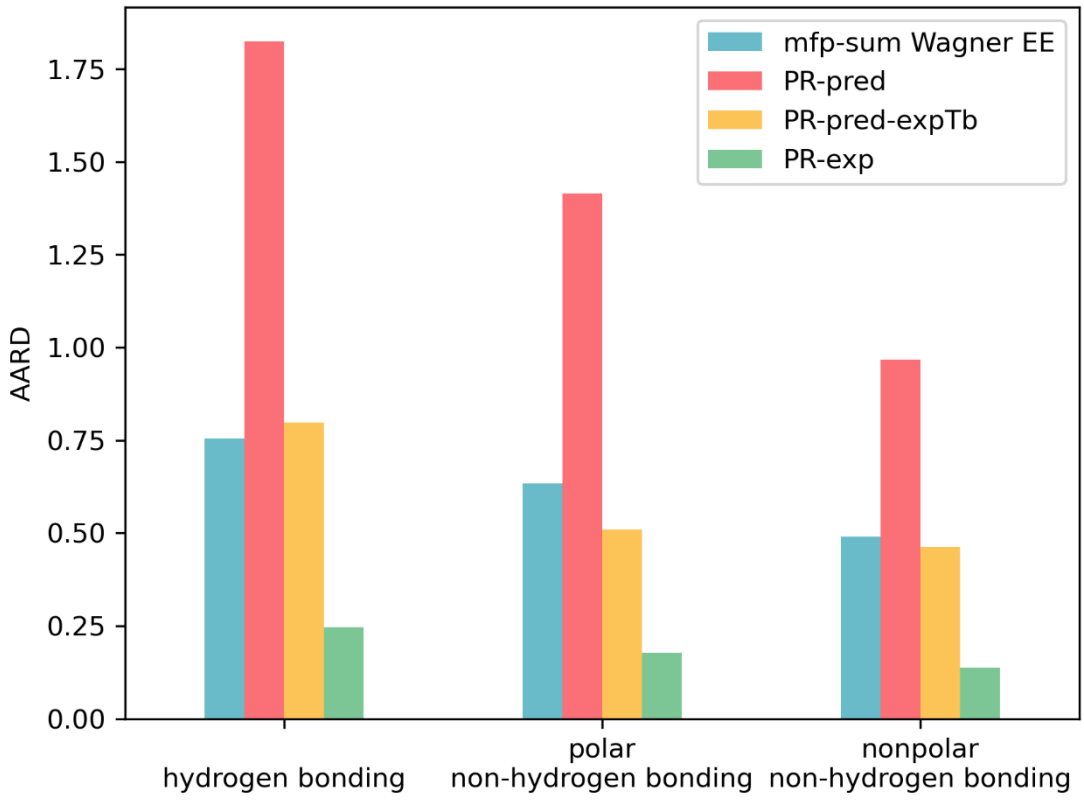

Fig. 5. Performance comparison of different models across molecule types. The bar color signifies different methods: blue represents mfp-sum Wagner EE, red corresponds to PR-pred, yellow depicts PR-pred-expTb, and green denotes PR-exp. Molecule types are distinguished based on the presence of hydrogen bonding and polarity.

To further evaluate the model's performance across a wide temperature range, we analyzed and plotted the average logarithm deviation (ALD) alongside the data frequency for each temperature range, as illustrated in Fig. S4 of the Supporting Information. The results demonstrate that the model achieves reliable prediction accuracy within the temperature range of 200 to 1000 K, where the data density is relatively high. However, outside this range, specifically at temperatures below 200 K and above 1000 K, the error increases. This limitation is inherent to data-driven models, which rely heavily on the availability and distribution of training data. Additionally, in the low-temperature region, vapor pressure measurements are particularly challenging, often leading to potential biases in the experimental data. This contributes to the reduced accuracy of the model in these regions (below 200 K). Conversely, in higher temperature regions (above 1000 K), the generally higher quality of experimental data enables the model to perform comparatively well. These findings emphasize the critical

role of data availability and quality in ensuring accurate vapor pressure predictions and highlight the model's practical applicability within temperature ranges where sufficient and reliable data are available.

4. Conclusions

This study presents a significant advancement in the field of vapor pressure prediction, leveraging the power of machine learning to address a critical challenge in chemical and environmental sciences. We introduced a machine learning-based model, employing the D-MPNN architecture, to predict the vapor pressure of organic molecules across a wide range of temperatures. This machine learning approach diverges from conventional methodologies that typically depend on either direct experimental data or the intensive computational demands of quantum mechanical calculations. Notably, our model outperforms the existing PR + COSMOSAC method, achieving an impressive AARD of 0.617 for a large testing set of 19,079 molecules. This performance significantly surpasses that of the standard PR + COSMOSAC method, which records an AARD of 1.36.

The findings of this study underscore the efficacy of machine learning in extracting and utilizing complex molecular features for property prediction, bypassing the need for explicit critical properties or extensive experimental data. This capability is particularly valuable for novel or unexplored chemical species, for which such data may not be readily available. Furthermore, our research highlights the potential of integrating temperature effects into molecular property predictions, a critical factor for accurate vapor pressure estimation across various applications. Future work could focus on enhancing the interpretability of the machine learning model, providing deeper insights into the relationships between molecular structure and predicted properties. This study opens new avenues for research in molecular property prediction, offering insights that could lead to significant advancements in chemical engineering, environmental science, and beyond.

CRediT authorship contribution statement

Yen-Hsiang Lin: Writing – original draft, Visualization, Software, Methodology, Data curation. **Hsin-Hao Liang:** Validation, Data curation. **Shiang-Tai Lin:** Writing – review & editing. **Yi-Pei Li:** Writing – review & editing, Supervision, Resources,

Funding acquisition.

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.

Data availability

The dataset on vapor pressure is restricted by third-party constraints. A detailed list of the molecules studied, along with the code utilized in this research, is accessible at https://github.com/fatcat0322/chemprop_VP.

Acknowledgments

Y.P.L. is supported by Taiwan NSTC (113-2628-E-002-017-MY3 and 113-2622-8-002-015-SB) and the Higher Education Sprout Project by the Ministry of Education in Taiwan (113L891305). We are grateful to the National Center for High-performance Computing (NCHC) and the Computer and Information Networking Center at NTU for the support of computing facilities. During the preparation of this work, the authors used ChatGPT in order to correct grammatical mistakes and enhance the fluency of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Sandler SI. Chemical, biochemical, and engineering thermodynamics. John Wiley & Sons; 2017.
[2] Luyben WL. Distillation column pressure selection. Separation and Purification Technology 2016;168:62-7.
[3] Fujii A, Hermann ER. Correlation between flash points and vapor pressures of organic compounds. Journal of Safety Research 1982;13(4):163-75.
[4] Bilde M, Barsanti K, Booth M, Cappa CD, Donahue NM, Emanuelsson EU, et al. Saturation vapor pressures and transition enthalpies of low-volatility organic molecules of atmospheric relevance: from dicarboxylic acids to complex mixtures. Chemical reviews 2015;115(10):4115-56.
[5] Paasivirta J, Sinkkonen S, Mikkelson P, Rantio T, Wania F. Estimation of vapor pressures, solubilities and Henry's law constants of selected persistent organic pollutants as functions of temperature. Chemosphere 1999;39(5):811-32.
[6] Abrefah J, Olander D, Balooch M, Siekhaus W. Vapor pressure of Buckminsterfullerene. Applied Physics Letters 1992;60(11):1313-4.

[7] Mixon F, Gumowski B, Carpenter B. Computation of vapor-liquid equilibrium data from solution vapor pressure measurements. Industrial & Engineering Chemistry Fundamentals 1965;4(4):455-9.
[8] Xue Z, Thridandam H, Kaesz HD, Hicks RF. Organometallic chemical vapor deposition of platinum. Reaction kinetics and vapor pressures of precursors. Chemistry of materials 1992;4(1):162-6.
[9] Chen Q-S, Wegrzyn J, Prasad V. Analysis of temperature and pressure changes in liquefied natural gas (LNG) cryogenic tanks. Cryogenics 2004;44(10):701-9.
[10] Schneider T, O'Gorman PA, Levine XJ. Water vapor and the dynamics of climate changes. Reviews of Geophysics 2010;48(3).
[11] Schmidt M, Gutierrez A, Linder M. Thermochemical energy storage with CaO/Ca (OH) 2–Experimental investigation of the thermal capability at low vapor pressures in a lab scale reactor. Applied Energy 2017;188:672-81.
[12] Vervaet C, Byron PR. Drug–surfactant–propellant interactions in HFA-formulations. International journal of pharmaceutics 1999;186(1):13-30.
[13] Thomson G. The DIPPR® databases. International journal of thermophysics 1996;17:223-32.
[14] Onken U, Rarey-Nies J, Gmehling J. The Dortmund Data Bank: A computerized system for retrieval, correlation, and prediction of thermodynamic properties of mixtures. International Journal of Thermophysics 1989;10:739-47.
[15] Soave G. Equilibrium constants from a modified Redlich-Kwong equation of state. Chemical engineering science 1972;27(6):1197-203.
[16] Peng D-Y, Robinson DB. A new two-constant equation of state. Industrial & Engineering Chemistry Fundamentals 1976;15(1):59-64.
[17] Tu CH. Group-Contribution Method for the Estimation of Vapor-Pressures. Fluid Phase Equilibria 1994;99:105-20. https://doi.org/Doi 10.1016/0378-3812(94)80025-1.
[18] Coutsikos P, Voutsas E, Magoulas K, Tassios DP. Prediction of vapor pressures of solid organic compounds with a group-contribution method. Fluid phase equilibria 2003;207(1-2):263-81.
[19] Moller B, Rarey J, Ramjugernath D. Estimation of the vapour pressure of non-electrolyte organic compounds via group contributions and group interactions. Journal of Molecular Liquids 2008;143(1):52-63.
[20] Nannoolal Y, Rarey J, Ramjugernath D. Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions. Fluid Phase Equilibria 2008;269(1-2):117-33.
[21] Pankow JF, Asher WE. SIMPOL. 1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. Atmospheric Chemistry and Physics 2008;8(10):2773-96.
[22] Katritzky AR, Wang Y, Sild S, Tamm T, Karelson M. QSPR studies on vapor pressure, aqueous solubility, and the prediction of water− air partition coefficients. Journal of Chemical Information and Computer Sciences 1998;38(4):720-5.
[23] Godavarthy SS, Robinson Jr RL, Gasem KA. SVRC–QSPR model for predicting saturated vapor pressures of pure fluids. Fluid phase equilibria 2006;246(1-2):39-51.

[24] Katritzky AR, Slavov SH, Dobchev DA, Karelson M. Rapid QSPR model development technique for prediction of vapor pressure of organic compounds. Computers & chemical engineering 2007;31(9):1123-30.
[25] Gharagheizi F, Eslamimanesh A, Ilani-Kashkouli P, Mohammadi AH, Richon D. QSPR molecular approach for representation/prediction of very large vapor pressure dataset. Chemical engineering science 2012;76:99-107.
[26] Kühne R, Ebert R-U, Schüürmann G. Estimation of vapour pressures for hydrocarbons and halogenated hydrocarbons from chemical structure by a neural network. Chemosphere 1997;34(4):671-86.
[27] Yaffe D, Cohen Y. Neural network based temperature-dependent quantitative structure property relations (QSPRs) for predicting vapor pressure of hydrocarbons. Journal of chemical information and computer sciences 2001;41(2):463-77.
[28] Tarjomannejad A. Prediction of the liquid vapor pressure using the artificial neural network-group contribution method. Iranian Journal of Chemistry and Chemical Engineering (IJCCE) 2015;34(4):97-111.
[29] Dearden JC. Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point. Environmental Toxicology and Chemistry: An International Journal 2003;22(8):1696-709.
[30] Odinokov A, Son W-J, Yakubovich A, Park JY, Jung Y. Ab Initio Prediction of Vapor Pressure for Diverse Atomic Layer Deposition Precursors. Journal of Chemical Theory and Computation 2024;20(14):6144-51.
[31] Klamt A, Schüürmann G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. Journal of the Chemical Society, Perkin Transactions 2 1993(5):799-805.
[32] Lin S-T, Hsieh C-M, Lee M-T. Solvation and chemical engineering thermodynamics. Journal of the Chinese Institute of Chemical Engineers 2007;38(5-6):467-76.
[33] Tsai CC, Lin ST. Improved vapor pressure prediction from PR+ COSMOSAC EOS using normal boiling temperature. AIChE Journal 2023;69(3):e17997.
[34] Li Y-P, Han K, Grambow CA, Green WH. Self-evolving machine: A continuously improving model for molecular thermochemistry. The Journal of Physical Chemistry A 2019;123(10):2142-52.
[35] Chen L-Y, Hsu T-W, Hsiung T-C, Li Y-P. Deep Learning-Based Increment Theory for Formation Enthalpy Predictions. The Journal of Physical Chemistry A 2022;126(41):7548-56.
[36] Yang C-I, Li Y-P. Explainable uncertainty quantifications for deep learning-based molecular property prediction. Journal of Cheminformatics 2023;15(1):13.
[37] Muthiah B, Li S-C, Li Y-P. Developing machine learning models for accurate prediction of radiative efficiency of greenhouse gases. Journal of the Taiwan Institute of Chemical Engineers 2023;151:105123.
[38] Cheng Y-H, Sung I-T, Hsieh C-M, Lin L-C. Module-based machine learning models using sigma profiles of organic linkers to predict gaseous adsorption in metal-organic frameworks. Journal of the Taiwan Institute of Chemical Engineers 2024;165:105728.
[39] Li S-C, Wu H, Menon A, Spiekermann KA, Li Y-P, Green WH. When Do

Quantum Mechanical Descriptors Help Graph Neural Networks to Predict Chemical Properties? Journal of the American Chemical Society 2024;146(33):23103-20.
[40] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug discovery today 2018;23(6):1241-50.
[41] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Molecular systems biology 2016;12(7):878.
[42] Segler MH, Waller MP. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry–A European Journal 2017;23(25):5966-71.
[43] Schreck JS, Coley CW, Bishop KJ. Learning retrosynthetic planning through simulated experience. ACS central science 2019;5(6):970-81.
[44] Jeong J, Lee N, Shin Y, Shin D. Intelligent generation of optimal synthetic pathways based on knowledge graph inference and retrosynthetic predictions using reaction big data. Journal of the Taiwan Institute of Chemical Engineers 2022;130:103982.
[45] Chiu P-H, Yang Y-L, Tsao H-K, Sheng Y-J. Deep learning for predictions of hydrolysis rates and conditional molecular design of esters. Journal of the Taiwan Institute of Chemical Engineers 2021;126:1-13.
[46] Meuwly M. Machine learning for chemical reactions. Chemical Reviews 2021;121(16):10218-39.
[47] Chen L-Y, Li Y-P. Machine Learning Applications in Chemical Kinetics and Thermochemistry. In: Machine Learning in Molecular Sciences, Springer; 2023, p. 203-26.
[48] Chen L-Y, Li Y-P. Enhancing chemical synthesis: a two-stage deep neural network for predicting feasible reaction conditions. Journal of Cheminformatics 2024;16(1):11.
[49] Chen L-Y, Li Y-P. Machine learning-guided strategies for reaction conditions design and optimization. Beilstein Journal of Organic Chemistry 2024;20(1):2476-92.
[50] Chen L-Y, Li Y-P. AutoTemplate: enhancing chemical reaction datasets for machine learning applications in organic chemistry. Journal of Cheminformatics 2024;16(1):74.
[51] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling 2010;50(5):742-54.
[52] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation 1965;5(2):107-13.
[53] Pattanaik L, Coley CW. Molecular Representation: Going Long on Fingerprints. Chem 2020;6(6):1204-7.
[54] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences 1988;28(1):31-6.
[55] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems 2017;30.
[56] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint

arXiv:181004805 2018.
[57] Honda S, Shi S, Ueda HR. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. arXiv preprint arXiv:191104738 2019.
[58] Wang S, Guo Y, Wang Y, Sun H, Huang J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. 2019, p. 429-36.
[59] Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:201009885 2020.
[60] Duvenaud DK, Maclaurin D, Iparraguirre J, Bombarell R, Hirzel T, Aspuru-Guzik A, et al. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems 2015;28.
[61] Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. Journal of chemical information and modeling 2017;57(8):1757-72.
[62] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. Neural Message Passing for Quantum Chemistry 2017:1263-72.
[63] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing Learned Molecular Representations for Property Prediction. Journal of Chemical Information and Modeling 2019;59(8):3370-88. https://doi.org/10.1021/acs.jcim.9b00237.
[64] Heid E, Greenman KP, Chung Y, Li S-C, Graff DE, Vermeire FH, et al. Chemprop: A machine learning package for chemical property prediction. Journal of Chemical Information and Modeling 2023.
[65] Kajita S, Ohba N, Jinnouchi R, Asahi R. A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks. Scientific reports 2017;7(1):1-9.
[66] Kuzminykh D, Polykovskiy D, Kadurin A, Zhebrak A, Baskov I, Nikolenko S, et al. 3D molecular representations based on the wave transform for convolutional neural networks. Molecular pharmaceutics 2018;15(10):4378-85.
[67] Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, et al. Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence 2022;4(2):127-34.
[68] Lansford JL, Jensen KF, Barnes BC. Physics-informed Transfer Learning for Out-of-sample Vapor Pressure Predictions. Propellants, Explosives, Pyrotechnics 2023;48(4):e202200265.
[69] Santana VV, Rebello CM, Queiroz LP, Ribeiro AM, Shardt N, Nogueira IB. PUFFIN: A path-unifying feed-forward interfaced network for vapor pressure prediction. Chemical Engineering Science 2024;286:119623.
[70] Thomson GW. The Antoine equation for vapor-pressure data. Chemical reviews 1946;38(1):1-39.
[71] Riedel L. Eine neue universelle Dampfdruckformel Untersuchungen über eine Erweiterung des Theorems der übereinstimmenden Zustände. Teil I. Chemie Ingenieur Technik 1954;26(2):83-9.
[72] Ambrose D, Ewing M, Ghiassee N, Ochoa JS. The ebulliometric method of

vapour-pressure measurement: vapour pressures of benzene, hexafluorobenzene, and naphthalene. The Journal of Chemical Thermodynamics 1990;22(6):589-605.
[73] Huang L. Normalization Techniques in Deep Learning. Springer; 2022.
[74] Cabello-Solorzano K, Ortigosa de Araujo I, Peña M, Correia L, J. Tallón-Ballesteros A. The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis. International Conference on Soft Computing Models in Industrial and Environmental Applications. Springer; 2023, p. 344-53.
[75] Aspen Plus. Aspen Technology, Inc; 2019.
[76] Smith FL, Harvey AH. Avoid common pitfalls when using Henry's law. Chemical engineering progress 2007;103(9):33-9.
[77] Bergstra J, Yamins D, Cox DD. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. Proceedings of the 12th Python in science conference. Citeseer; 2013, p. 20.
[78] Zhao Q. TCIT_thermo; Available from: https://github.com/zhaoqy1996/TCIT_thermo/. [Accessed May 2024].