

Research Article

Open Access



Phthalonitrile melting point prediction enabled by multi-fidelity learning

Beijian Xu^{1,#}, Xiao Hu^{1,#}, Haoxiang Lan^{1,#}, Tianyi Wang¹, Xin-Yao Xu¹, Chongyin Zhang^{2,*}, Jiaping Lin^{1,*}, Liquan Wang^{1,*}, Lei Du¹

¹Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory of Specially Functional Polymeric Materials and Related Technology (Ministry of Education), School of Materials Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.

²Shanghai Aerospace Equipment Manufacturing Co., Ltd., Shanghai 200245, China.

[#]Authors contributed equally.

***Correspondence to:** Dr. Chongyin Zhang, Shanghai Aerospace Equipment Manufacturing Co., Ltd., Huaning Road 100, Minhang District, Shanghai 200245, China. E-mail: chongyin1022@163.com; Prof. Jiaping Lin, Prof. Liquan Wang, Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory of Specially Functional Polymeric Materials and Related Technology (Ministry of Education), School of Materials Science and Engineering, East China University of Science and Technology, Meilong Road 130, Xuhui District, Shanghai 200237, China. E-mail: jlin@ecust.edu.cn; lq_wang@ecust.edu.cn

How to cite this article: Xu B, Hu X, Lan H, Wang T, Xu XY, Zhang C, Lin J, Wang L, Du L. Phthalonitrile melting point prediction enabled by multi-fidelity learning. *J Mater Inf* 2024;4:21. <https://dx.doi.org/10.20517/jmi.2024.27>

Received: 30 Jul 2024 **First Decision:** 24 Sep 2024 **Revised:** 28 Oct 2024 **Accepted:** 5 Nov 2024 **Published:** 14 Nov 2024

Academic Editors: Ming Hu, Lei Shen **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Phthalonitrile (PN) resins have been widely used in various fields for their excellent thermal stability and mechanical properties, but they suffer from poor processability due to their high melting point. Data-driven machine learning (ML) can assist in screening PNs with low melting points but is limited by the lack of experimental data. Using error correction and multi-fidelity co-training methods, we established two multi-fidelity models for predicting PN melting points. This work demonstrates that through multi-fidelity learning, limited experimental data can be effectively utilized with the assistance of all-atom molecular dynamics simulation data to establish ML-based property prediction models. A comparison between these two multi-fidelity prediction models was made, and the contribution of chemical units to the PN melting point was analyzed based on one of the models. Our work offers feasible ML tools for future designing PNs with good processability.

Keywords: Phthalonitrile, melting point, machine learning, molecular dynamics, multi-fidelity



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



INTRODUCTION

Phthalonitriles (PNs), as a representative heat-resistant thermosetting resin, bear exceptional thermal stability, chemical corrosion resistance, outstanding mechanical properties, flame retardancy, and low dielectric loss. These remarkable features contribute to their high utility in fundamental areas such as aerospace, military equipment, and shipbuilding^[1-4]. In recent years, PN-based materials have also demonstrated promising applications in new fields such as radiation-resistant and microelectronic devices^[5-9]. It is foreseeable that PNs will gain increased attention and exploration due to their unique performance advantages.

As thermosetting polymers, the highly cross-linked heterocyclic structure of PN resins guarantees their high glass transition and thermal decomposition temperatures. However, the melting points of PN molecules generally exceed 180 °C due to their rigid molecular structure and strong intermolecular forces^[10]. Evidence has shown that the melting point of bisphenol A-based PNs can reach 195 °C, while biphenyldiol-based PNs can achieve a higher melting point of 235 °C^[11]. High melting points can contribute to the heat resistance of PNs but become a limitation in processing. Some researchers regretfully believe that for industrial processing, the overmuch high melting point of most PNs, usually greater than 200 °C, can be their common drawback^[8,12]. In addition, due to the proximity of the high melting temperature to the polymerization onset temperature, the resin will polymerize immediately after melting, resulting in a very narrow processing window^[10,13]. Therefore, PNs with low melting points and excellent processing performance are anticipated to be discovered but remain challenging.

The melting points can be predicted through various methods. Molecular dynamics simulation has proven to be a reliable tool for predicting melting points. There are several examples regarding the study of the melting points of various matters using molecular dynamics simulations, demonstrating the versatility of this approach^[14-18]. However, the application of molecular dynamics simulation is usually limited by the accuracy of force fields and computational cost. As a technology of artificial intelligence, machine learning (ML) has gained attention for its advantages in predicting material properties^[19-24]. For example, Hu *et al.* utilized ML to construct a prediction model for the mechanical properties of epoxy resins and employed this model to predict and screen resins that balance tensile strength, tensile modulus, and elongation at break^[25]. Herein, we aim to utilize ML to establish a quantitative structure-property relationship (QSPR) for the melting point of PNs to assist in screening PNs with low melting points. Typically, a large and diverse dataset with high accuracy is crucial for accurate ML. However, “small data” challenges, such as limited experimental data, harrowing information extraction, and poor data quality, pose obstacles in ML. Collecting data from different sources, such as simulation work or experimental characterization of different methods, can help augment data. However, the different fidelity will lead to poor performance of the ML models. Multi-fidelity learning, which shares concepts similar to transfer learning in deep learning, can leverage data from different sources. Therefore, multi-fidelity learning is typically applied in scenarios where high-fidelity data is limited in quantity and time-consuming to obtain, but a large amount of low-fidelity data related to high-fidelity data is available. It is currently the most feasible strategy for addressing “small data” challenges. Xu *et al.* also faced a similar “small data” challenge when designing polycyanurates and thoroughly used limited experimental data with the assistance of simulation data to obtain a predictive model with excellent performance through multi-fidelity ML^[26].

In this study, we proposed a comprehensive strategy combining multi-fidelity data and ML to construct QSPR models, enabling accurate predictions of PN melting points. All-atom molecular dynamics (AAMD) simulations were extensively employed to augment the dataset. Following the concept of multi-fidelity learning, we employed two ML approaches - an error correction method (ECM) inspired by Xu *et al.* and a

multi-fidelity co-training method [MatErials graph network (MEGNet)] proposed by Chen *et al.* - to build QSPR models^[26,27]. The performances of two multi-fidelity QSPR models were compared. Additionally, we performed a fundamental gene analysis of the structure of PNs with low melting points and provided a brief recommendation for future molecule design.

METHODS

We first collected data on experimentally measured melting points of 58 PNs from existing literature. These collected experimental data are treated as high-fidelity data for ML. Since the experimental data is limited, we also generated an adequate volume of low-fidelity data to expand the total dataset. We constructed a sizeable candidate space of virtual PN structures through gene combination and selected 200 PNs from them randomly and evenly for AAMD simulations. The low-fidelity data on melting points were obtained through AAMD simulations.

In the AAMD simulation, the building of molecular structures, the assignment of structural information, and geometry optimization were first performed using Materials Studio software with a Condensed-phase Optimized Molecular Potentials for Atomistic Simulation Studies II (COMPASS II) force field. Then, the Moltemplate tool was employed to construct the initial simulation system with Dreiding force fields for AAMD simulation in the large-scale atomic/molecular massively parallel simulator (LAMMPS). A cooling process spanning from high to low temperatures was carried out for each simulation, and the density at each temperature was recorded. The melting point was determined by piecewise fitting of the density-temperature plot. In contrast with the experimentally measured melting points, the simulated melting points were regarded as low-fidelity data. In the end, a multi-fidelity dataset consisting of experimental and simulated melting points was established.

While Materials Studio is intuitive and straightforward, its computational efficiency is low. As the number of molecules increases and the simulation scale expands, the time and computational costs rise sharply. In an initial, unoptimized simulation, we even needed to spend a week to obtain only a small amount of usable data, which is inefficient. Although LAMMPS offers better algorithm optimization and parallel computing capabilities, it still requires sufficient computational power and theoretical knowledge, with significant learning costs that cannot be ignored. Overall, the AAMD simulation has limitations in computational cost. The overall simulation process is burdensome, making it difficult to scale up for larger applications, which is an issue that urgently needs to be addressed. On the other hand, the AAMD simulation can give low-fidelity data as compared with the experimental data. The details regarding preparing the multi-fidelity dataset can be referred to [Supplementary Sections 1-4](#). We adopted two approaches for the model construction based on the multi-fidelity dataset. Multi-fidelity learning, leveraging data with varied fidelities, offers a prediction model with a much lower computation cost than the AAMD simulation.

The first approach is an ECM. In this method, molecular descriptors for 200 selected PNs were calculated using the Mordred package and were dimensionally reduced based on the correlation analysis. We first utilized all simulated data to build a low-fidelity $Y_L(X)$ model using Gaussian process regression (GPR). GPR is chosen due to its status as a non-parametric model widely used in small-data tasks. More details of GPR can be found in [Supplementary Section 5](#). The GPR was performed with various kernel functions, where the dataset was randomly split into training, validation, and test sets in a ratio of 7:2:1. The optimal low-fidelity model $Y_L(X)$ was determined by careful consideration of correlation coefficient (R^2), mean absolute error (MAE), and mean square error (MSE) of all obtained models. Next, the difference between experimental and simulated melting points was calculated, and a difference model of $Y_D(X)$ was obtained using leave-one-out cross-validation (LOO). The final $Y_H(X)$ model was derived by adding $Y_D(X)$ to $Y_L(X)$.

that is, $Y_H(\mathbf{X}) = Y_L(\mathbf{X}) + Y_D(\mathbf{X})$.

The second approach is a multi-fidelity co-training method named MEGNet. The MEGNet was proposed by Chen *et al.*^[27]. We chose MEGNet because this graph neural network incorporates a global state attribute into the message-passing process during training, enabling multi-fidelity learning. In training, we incorporated experimental and simulated data into the training process simultaneously (co-training) by giving different global state attributes (0 or 1) to different-fidelity data. In MEGNet, simplified molecular input line entry systems (SMILES) were utilized for graph representation, and the model was built through MEGNet blocks and a series of message-passing and readout processes. The multi-fidelity dataset was divided into training and test sets in a ratio of 8:2, with the low-fidelity data removed from the test set. Five-fold cross-validation was performed to find the optimal hyperparameter combination, including learning rate, epochs, and batch size. A total of 50 splits were adopted in the ML, and the best-performing model was determined by comprehensively examining the R^2 , MAE, and root mean squared error (RMSE) of all generated models.

RESULTS AND DISCUSSION

The work is organized as shown in [Figure 1](#). We first prepared the dataset for multi-fidelity learning and then constructed two multi-fidelity models based on an ECM and a multi-fidelity co-training method based on MEGNet, respectively. The performance of these multi-fidelity models was then evaluated. Finally, we conducted gene analysis for PNs with low melting points and provided some essential discussions and recommendations about molecular design.

The primary focus of this work is to establish prediction models for PN melting points based on the multi-fidelity concept, and naturally, a high-quality and comprehensive multi-fidelity dataset is crucial. One prefers to conduct ML exclusively using high-fidelity experimental data, which is usually unrealistic for materials research. In this study, for example, we considered experimentally measured melting points, such as those obtained through differential scanning calorimetry (DSC), as high-fidelity data, but such data is rare and often includes information with low reliability, unknown origin, or related to mixtures. After extensively reviewing the relevant works, we collected 58 PNs with varied chemical structures and their melting points. These experimental data will be treated as high-fidelity data in the multi-fidelity dataset and fully utilized in subsequent ML processes. The details regarding the experimental melting point we collected can be found in [Supplementary Section 1](#).

Insufficient high-fidelity data necessitates supplementing low-fidelity data despite the additional data having a limited precision. Low-fidelity data does not require strict accuracy but must possess a sufficient volume, making AAMD simulations particularly suitable. By conducting AAMD simulations, we can obtain simulated melting points of numerous structures that have not been previously studied or discovered. The simulation data can serve as low-fidelity data to expand the total dataset. Before calculating the melting points, we designed a gene combination strategy to create numerous virtual PN molecules. During this, it is crucial to set a diverse range of structures and elements to ensure that models obtained by ML subsequently have excellent generality for complex molecules. Typically, a random gene combination may inevitably yield many chemically impractical molecules. To make the process more realistic and synthetically accessible, we designed three PN templates and identified the connection positions for chemical units (including spacers and substituents), as depicted in [Figure 2A](#). Ultimately, we obtained a candidate space of 148,964 PN molecules. The details regarding the creation of virtual PN molecules can be found in [Supplementary Section 2](#).

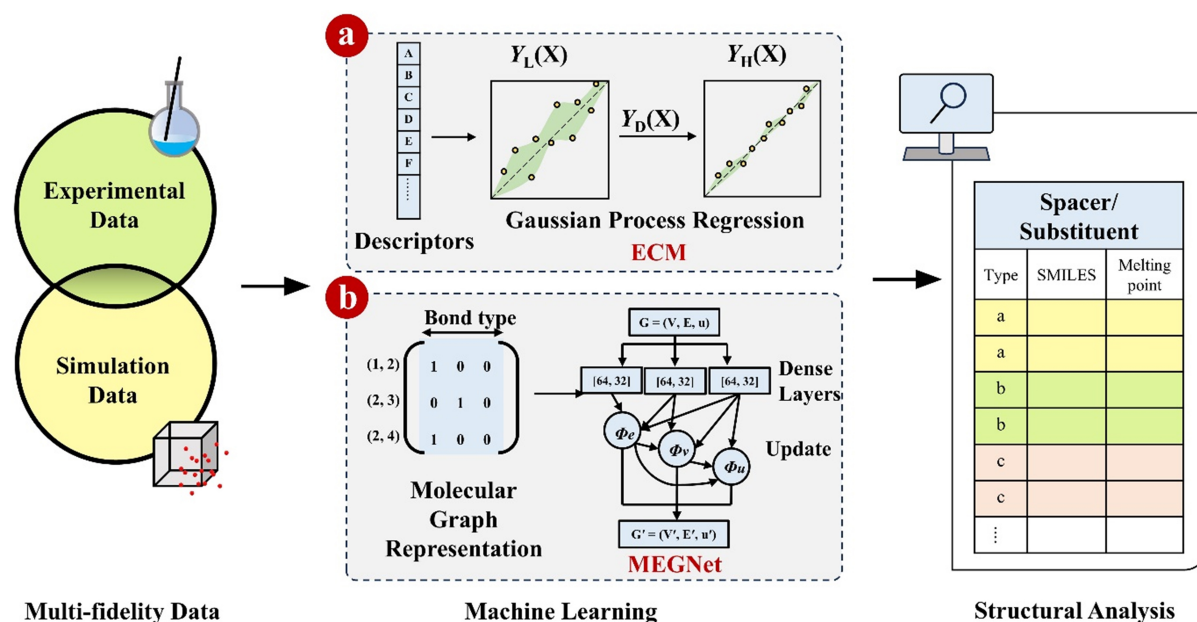


Figure 1. General implementation route of this work. Limited experimental data are expanded with simulated data, which are utilized to build multi-fidelity models by either (A) ECMs or (B) multi-fidelity co-training methods. Based on these models, we can analyze the contribution of spacers/substituents to PN melting points. ECMs: Error correction method; PN: phthalonitrile.

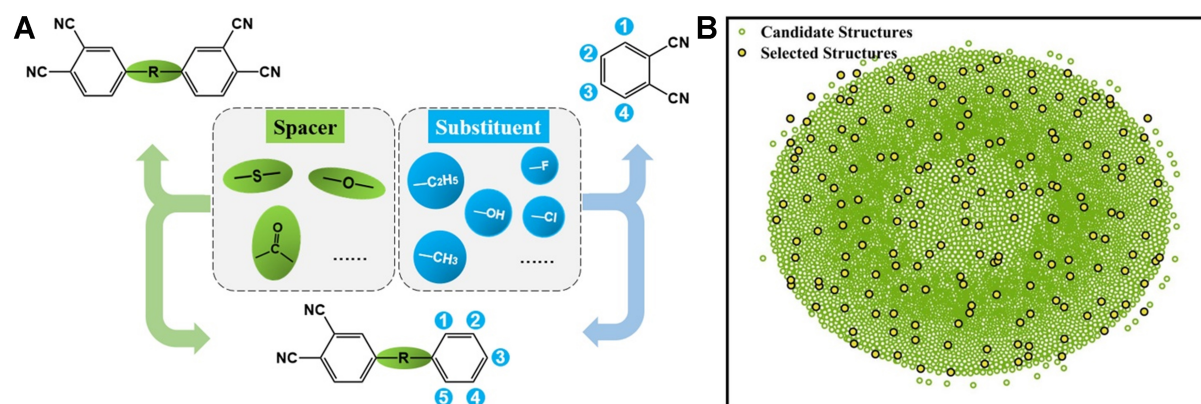


Figure 2. (A) Gene combination strategy for creating virtual PN molecules. Specific spacers and substituents are combined into the template to generate a vast candidate space; (B) Two-dimensional distribution of the sampling space and candidate space generated by *t*-SNE analysis. Sampling structures (in yellow) evenly cover the entire candidate space (in green), indicating that the sampling is representative and can avoid selecting too many similar structures. PN: Phthalonitrile; *t*-SNE: *t*-distributed stochastic neighbor embedding.

The key to ML lies in data acquisition. Ideally, we would like to obtain more experimental data, but we are constrained by the challenge of “small data”. While ensuring the accuracy and effectiveness of simulation methods, acquiring more simulation data to help ML capture deep underlying patterns is also a good option, though it requires balancing the computational cost of simulations. Using more data in ML can, in principle, lead to more accurate predictions. However, conducting high-throughput AAMD simulations for melting points demands significant computational power and time. Therefore, selecting the appropriate amount of simulation data for multi-fidelity ML can balance cost-effectiveness and accuracy better.

Here, we selected 200 PNs as representatives for the AAMD simulation. The selected molecules should be representative to ensure that the model we constructed possesses strong generalization capability. To demonstrate that our selection is extensive and evenly distributed, we employed *t*-distributed stochastic neighbor embedding (*t*-SNE) analysis to visualize the entire candidate space in two dimensions. The coordinates of substances in this two-dimensional space reveal their structural characteristics. As shown in [Figure 2B](#), the distribution of points in the candidate space spans a broad range, which proves the chemical diversity of the molecules we have created. Notably, the molecules we selected effectively cover the entire candidate space, indicating that the sampling is uniform. The PNs with similar chemical structural features can exhibit clustered distribution characteristics. Since the virtual samples generated through gene combinations in the *t*-SNE analysis bear diverse structures, they exhibit local similarity and global continuity, displaying a “uniform” distribution. For detailed information regarding *t*-SNE and random sampling, see [Supplementary Section 3](#).

Subsequently, we adopted an efficient AAMD simulation scheme capable of calculating the melting point for various PN molecules. We used this scheme to calculate melting points directly for 200 randomly selected molecules and 58 PNs we collected from the literature. Then, we examined the data distribution of experimental and simulated melting points. The results are shown in [Figure 3](#). The experimental data does not show a normal distribution. In contrast, the increased amount of simulation data overcomes such a problem, giving the combined total data greater plausibility and breadth.

To prove the reliability of our simulation scheme, we analyzed the correlation coefficient R^2 between experimental and simulated melting points. The result is shown in [Figure 4A](#). The R^2 of 0.885 indicates a strong correlation despite a disparity between the simulated and experimental results. The strong correlation confirms that the simulation scheme is suitable for generating low-fidelity data required for multi-fidelity learning. So far, we have established a multi-fidelity melting point dataset for PNs, which contains the experimental and simulated melting points for 58 PNs collected from the literature and the simulated melting points for 200 other randomly selected PNs.

ECM

We first used an ECM to construct the multi-fidelity model. In the ECM method, the fundamental knowledge was built based on low-fidelity data and further improved by the difference generated between low-fidelity and high-fidelity data. We can represent the ECM mathematically as^[21]

$$Y_H(\mathbf{X}) = Y_L(\mathbf{X}) + Y_D(\mathbf{X}) \quad (1)$$

Here, $Y_H(\mathbf{X})$ and $Y_L(\mathbf{X})$ represent high- and low-fidelity models, respectively. $Y_D(\mathbf{X})$ signifies the difference (\mathbf{D}) between $Y_H(\mathbf{X})$ and $Y_L(\mathbf{X})$.

Following the multi-fidelity learning framework, the QSPR model for melting points was established in two steps. The low-fidelity dataset was used in the first step to learn the general chemical rules between structure and melting point. Note that the low-fidelity dataset only includes simulated melting points. We used GPR with different kernel functions to build ML models. The GPR is a non-parametric regression that models the function space through a Gaussian process. The details of GPR can be seen in [Supplementary Section 5](#). R^2 , MAE, and MSE were utilized to evaluate the performance of the obtained low-fidelity model $Y_L(\mathbf{X})$. The results are shown in [Figure 4B](#) and [Supplementary Table 6](#). The R^2 between the simulated and the predicted melting points obtained by the optimal $Y_L(\mathbf{X})$ model, respectively, reaches 0.783, 0.605, and 0.671 on the training set, validation set, and test set, indicating that the $Y_L(\mathbf{X})$ model trained solely based on low-fidelity data can possess the ability to predict the melting point of PN, but the accuracy is not satisfactory.

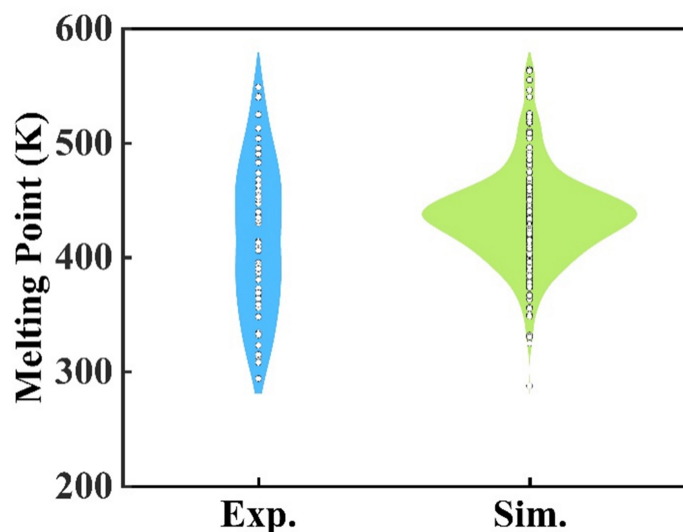


Figure 3. Violin plots of experimental and simulated melting points. The circles show the data points of experimental and simulated melting points, and the colored regimes denote the probability distribution. The blue and green regimes correspond to the experimental and simulated melting points, respectively.

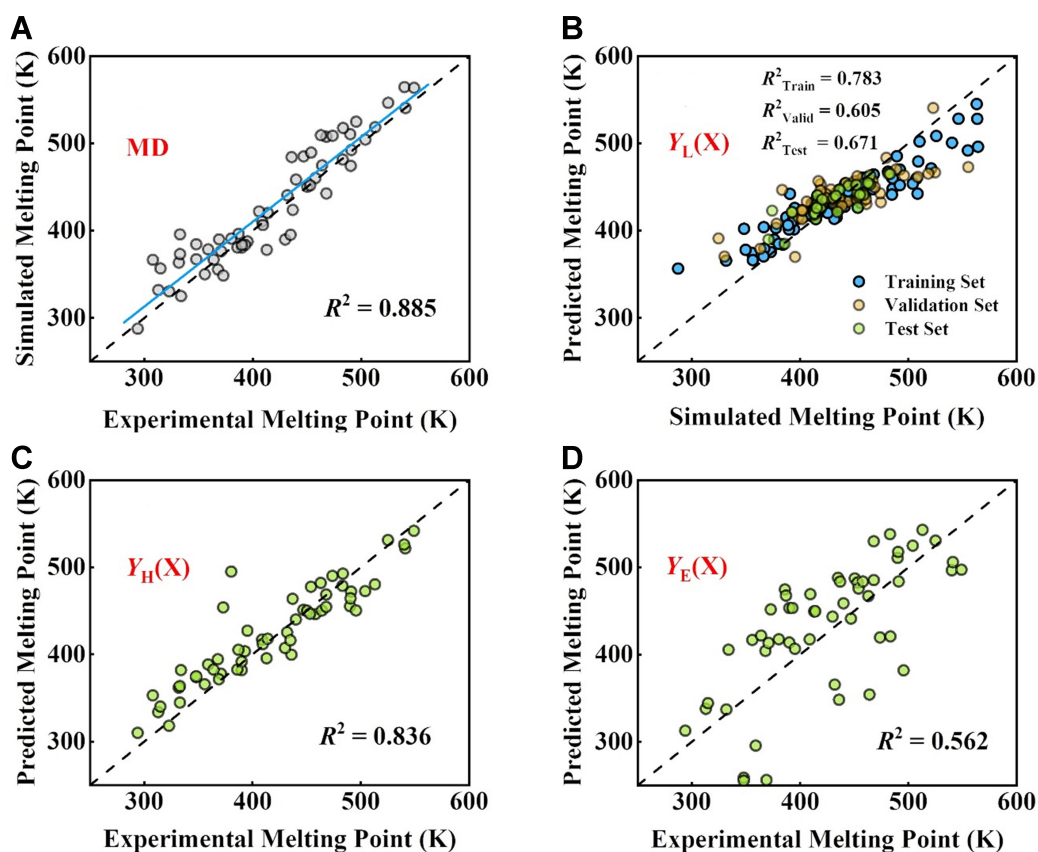


Figure 4. (A) The comparison between experimental and simulated melting points, where the blue solid line indicates a linear fit. The performances of (B) $Y_L(X)$ model, (C) $Y_H(X)$ model, and (D) $Y_E(X)$ model.

In the second step, we integrated high- and low-fidelity data to learn the correspondence between experimental and simulated results. Differences between experimental and simulated melting points were calculated, and similarly, using GPR with various kernel functions, we studied the correlation between structures and differences, i.e., $Y_D(X)$. Due to limited data, we employed LOO, which performs well on small datasets, to maximize data utilization. The $Y_L(X)$ and $Y_D(X)$ models with the best performance were selected, leading to $Y_H(X)$ via Equation (1). The R^2 and MAE were utilized to evaluate the performance of $Y_H(X)$, revealing an R^2 of 0.836 as depicted in [Figure 4C](#) and [Supplementary Table 7](#), demonstrating that the $Y_H(X)$ model exhibits good prediction accuracy for melting points.

We further employed neat data on experimental melting points to construct the $Y_E(X)$ model. The same procedure for building the $Y_D(X)$ model was used to construct the $Y_E(X)$ model. The advantage of multi-fidelity learning in a “small data” scenario was demonstrated by comparing it with the $Y_E(X)$ model. By comparing [Figure 4C](#) with [4D](#), one can conclude that under the same conditions, the $Y_H(X)$ model exhibits superior performance over the $Y_E(X)$ model. Through multi-fidelity learning, more effective utilization of limited experimental data was achieved by supplementing simulation data, resulting in a QSPR model with enhanced performance. For details regarding the construction and evaluation of the model through GPR, refer to [Supplementary Section 6](#).

Multi-fidelity co-training method

In addition to the ECM, we also attempted to establish a prediction model for PN melting points using a multi-fidelity co-training method named MEGNet. The MEGNet is a graph neural network model designed for materials science, embracing the message-passing neural network concept to transform polymers into graph-structured data^[27].

The construction of the QSPR model for melting point based on MEGNet can be divided into four steps. In the first step, we established the multi-fidelity dataset for input. Through software such as Kingdraw, we converted the PN structures into SMILES strings to facilitate computer reading and identification. Since there are two kinds of melting points for 58 PNs, one from experiments and the other obtained by simulations, we defined the fidelity global state as “0” and “1” to differentiate the simulated and experimental melting points. Such a treatment allows us to leverage the dataset for MEGNet straightforwardly. We removed the low-fidelity data as we evaluated the performance of the obtained model.

In the second step, we defined the parameters for each component of MEGNet. Firstly, the number of stacked MEGNet blocks was set. The previous research revealed that without stacking (i.e., the number of stacked blocks is 1), the generalization ability is poor, and the performance is significantly weaker than the model with stacking. However, stacking too many blocks can lead to issues such as overfitting or computational inefficiency. Therefore, we defined three MEGNet blocks. In each MEGNet block, the number of layers in the fully connected layer was set to 2, with neurons of 64 and 32, respectively. The update function, a three-layer independent perceptron, has 64, 64, and 32 neurons. Subsequently, we defined the parameters for the seq2seq neural network (message readout module). We set the number of iterations for processing blocks to 2, with default values for other parameters. The readout atom and bond information matrix were merged with the global state attribute and passed through two fully connected layers with 32 and 16 neurons, respectively, for single-value output.

In the third step, a five-fold cross-validation was conducted to select the optimal hyperparameter combination. The dataset was randomly divided into a training set and a test set at a ratio of 8:2. Note that

the test set includes experimental and simulated data, and the testing can be performed for either experimental or simulated data. A five-fold cross-validation was carried out by evenly dividing the training set into five subsets. By comparing the average performance of the five models for each hyperparameter combination, we can determine the effectiveness of different hyperparameter settings. The results show that the best-performing hyperparameter combination has a learning rate of 0.0008, an epoch of 4, and a batch size of 2,000, with an average R^2 reaching 0.904.

In the fourth step, we constructed the QSPR model using the training set with the optimal hyperparameter combination identified in the third step. The entire process was iterated 50 times, and the performance was assessed with R^2 , MAE, and RMSE [Supplementary Table 12]. As shown in Figure 5A, The R^2 between the predicted melting point obtained from the optimal model and the true melting point within the dataset reaches 0.934 on the training set and 0.925 on the test set of experimental data, demonstrating excellent predictive accuracy of the model for the PN melting point. The R^2 between the predicted and experimental melting points for all the experimentally collected PNs is 0.954 [Figure 5B]. For detailed information regarding MEGNet, refer to Supplementary Section 7.

Comparison between ECM and MEGNet models

We made a comparison between the ECM model and the MEGNet model. The commonality between them lies in the application of the multi-fidelity concept. When using ML to construct QSPR models, facing the “small data” challenge, a dataset with too few pure experimental data may struggle to interpret the deep structure-performance relationship, potentially leading to an extreme model performance with inadequate generality. On the other hand, a simulation dataset inherently contains errors compared to the experimental values, which might even be amplified after ML, resulting in weak model generalization. In contrast, utilizing a mixed dataset composed of high-fidelity but rare experimental data and low-fidelity but inexpensive simulated data for ML is more reliable. In the current maturity of MDs, researchers do not need to conduct extensive experimental work to collect material properties, and instead, they can construct predictive models using multi-fidelity datasets, yielding models with satisfactory predictive capabilities.

Still, there are differences between the two approaches in various aspects. In describing molecular features, the former is represented through descriptors, while the latter is represented through molecular graphs. In terms of data utilization, in the former, experimental data is not directly used as high-fidelity data but is incorporated into the construction process of the high-fidelity model in the form of errors. In contrast, the latter can directly handle and learn from the mixed dataset. Regarding the operational process, the former has an evident stepwise processing mechanism, where the optimal low-fidelity model must be established first before high-fidelity learning can commence. In contrast, the latter goes more directly. Once the optimal hyperparameter combination is found, the process can be done without breaks. Comparing Figure 4C with 5B, we may find that the model built by MEGNet exhibits better performance than the ECM model. This is likely because, compared to the descriptors used in the former, the molecular graph used by MEGNet provides a clearer understanding of the deep rules in molecular structures. It aggregates information from graph nodes through message passing, resulting in a QSPR model with strong generalization capabilities, while on the contrary, the precision of the ECM model is constrained by various factors, such as the accuracy of descriptors, the rationality of feature engineering, and the precision of low-fidelity model.

We performed an error analysis to explain why the ECM model underperformed. Figure 6A and B shows the residuals for the $Y_L(X)$ and $Y_H(X)$ models, respectively. The $Y_L(X)$ model shows a significant residual when predicting a larger or smaller melting point, leading to poor performance as the melting points approach the experimental limits [Figure 6A]. This derivation also leads to the inaccuracy of the $Y_H(X)$

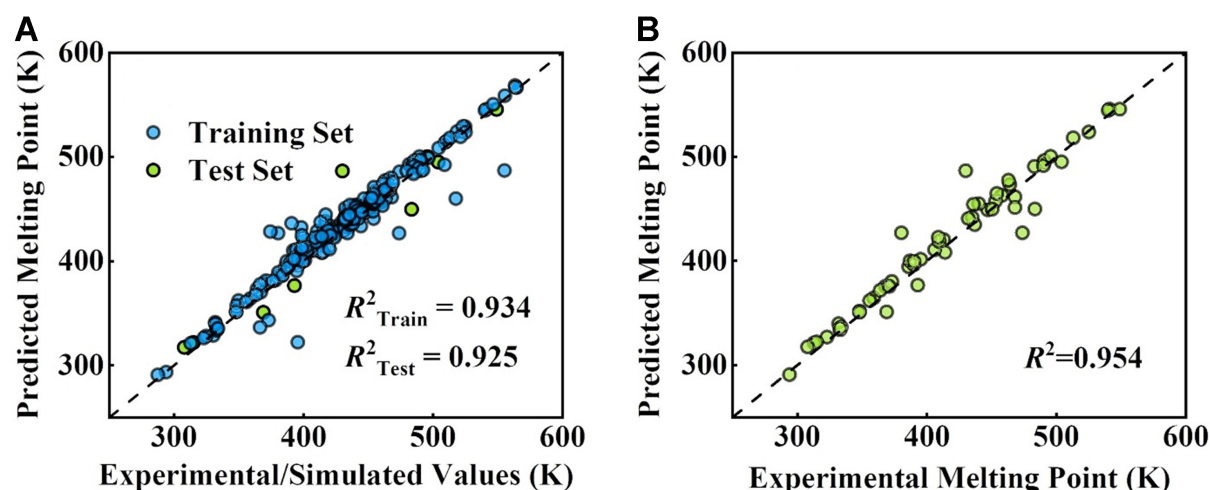


Figure 5. (A) The performance of the best MEGNet model; (B) The performance of the MEGNet model in predicting the melting point of PNs collected from the literature. MEGNet: MatErials graph network; PNs: phthalonitriles.

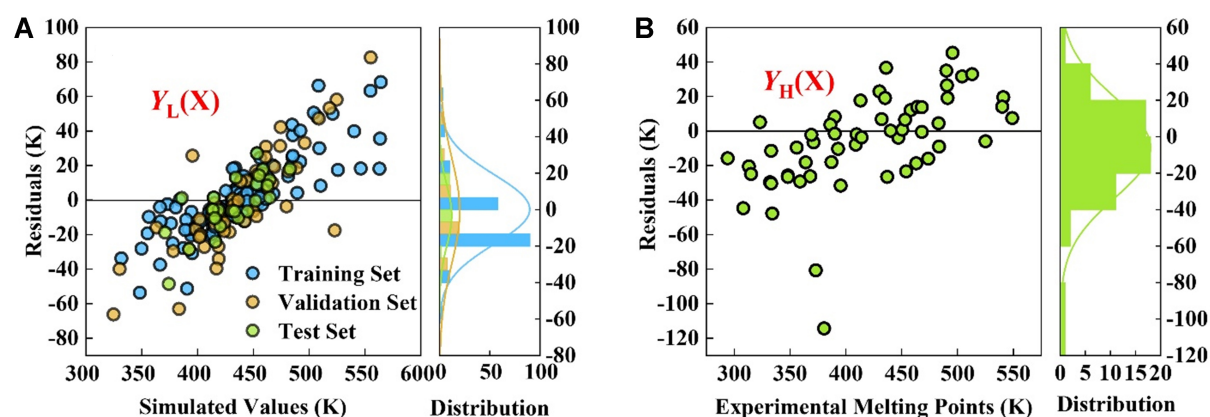


Figure 6. The residual plot and its distribution plot for the (A) $Y_L(X)$ model and (B) $Y_H(X)$ model.

model in predicting a larger or smaller melting point [Figure 6B]. This is the main reason that the ECM model performs poorly in predicting melting points beyond the experimental range.

One can see from Figure 7 that the predictive results from two multi-fidelity models are not in good agreement. The melting points predicted by ECM are located between 370 and 470 K, while those predicted by MEGNet range from 300 to 560 K. The ECM model almost lost the ability to predict the melting points beyond the experimental range (training data). It is foreseeable that MEGNet may become a hotspot for material performance prediction due to its powerful capabilities in the future.

Structural analysis based on MEGNet models

At this point, we have established the relationship between structures and melting points for PNs. The primary goal of this study is to utilize the constructed QSPR model for PNs to screen structures with excellent processability, i.e., those with low melting points. In the above, we obtained a candidate space comprising 148,964 PN structures, and with the help of the optimal MEGNet model using the multi-fidelity co-training strategy, we can predict the melting points for all candidate structures, forming a considerable dataset that instills confidence for statistical analysis. It is widely recognized that the performance of PN

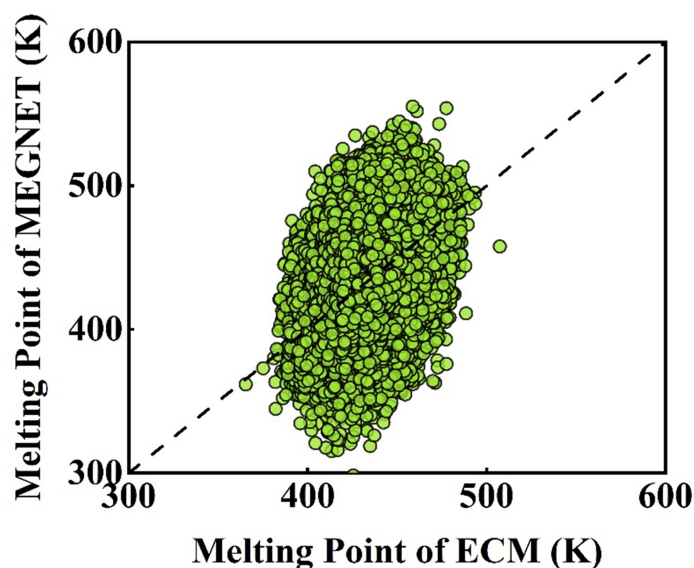


Figure 7. Capability comparison of the MEGNet and ECM model in predicting virtual PNs. The melting points predicted by the ECM model are located between 370 and 470 K, while those predicted by the MEGNet model range from 300 to 560 K. MEGNet: MatErials graph network; ECM: error correction method; PNs: phthalonitriles.

resins is influenced significantly by the spacer (R group) structures^[28]. Therefore, we analyzed the ten R groups involved in our designed candidate space. Initially, we arranged all 148,964 melting points in ascending order and selected the top 20% as preferred structures with lower melting points. Subsequently, we designed an algorithm to separate spacers from these preferred structures and statistically analyzed their occurrence frequency. The result is shown in Figure 8A. It can be simplistically inferred that R groups with higher occurrence frequencies may positively contribute to reducing the melting point of PNs. However, it is essential to acknowledge that the contributions of different structures to performance are relative. Therefore, we wish to propose an alternative approach to capture this relativity. We designed algorithms to summarize the PN structures with the same R group and their melting points and then plotted their cumulative distribution function (CDF) curve [Figure 8B]. The structures to the left of the average CDF curve are considered relatively favorable for lowering the melting point, while those to the right impede lowering the melting point. We also computed the integral area enclosed by each CDF curve and the average CDF curve, as depicted in Figure 8C in a histogram, providing a quantitative estimate of their relative contributions.

From Figure 8A-C, we learned that PN molecules with spacers such as ester bonds and ether bonds^[29] possess rotational freedom, thus having a lower melting temperature due to good flexibility. The introduction of silicon can usually lower the melting point of the monomer^[30], such as siloxane structure, and for other siloxane-like structures, the relatively weak silicon-oxygen bonds may lead to diminished attractive forces between siloxane molecules, reducing intermolecular interactions and causing a decrease in the melting point^[31-33]. In cases where phosphine oxide structures exist in the spacer, some reports suggest introducing phosphine oxide-containing structures may lower the melting point^[31-33]. Conversely, the presence of specific spacer structures with strong intermolecular forces (sulfone, peptide bonds, fluoromethyl group), conjugation effects (carbonyl group), and rigid segments (benzene) may lead to an increase in the melting point.

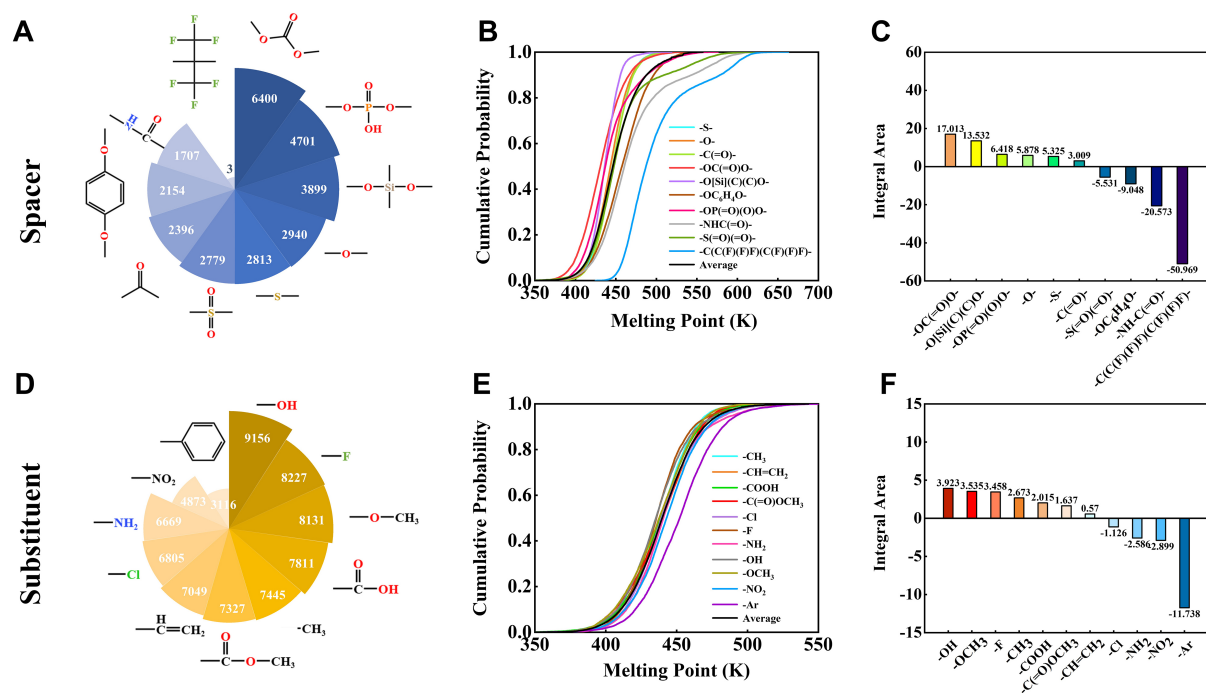


Figure 8. Analysis of gene contribution to melting points for (A-C) spacers and (D-F) substituents. The frequency of occurrence of each spacer (A) and substituent (D) in preferred structures is represented numerically in the sectors of the Nightingale rose diagram. The CDF curves are given for spacers (B) and substituents (E). The integrated area enclosed by the CDF curve of each spacer (C) and substituent (F) and the average CDF curve was represented as a histogram to describe the impact of the chemical unit on the melting point. CDF: Cumulative distribution function.

In addition to spacers, the substituent is another crucial factor influencing melting points. Following the same approach for obtaining Figure 8A-C, we derived frequency statistics [Figure 8D], CDF curves [Figure 8E], and integral areas [Figure 8F] for each substituent in the preferred structures. With no contradiction to experience, the introduction of large rigid groups (benzene), strongly electronegative groups (such as nitro and halogen), and groups capable of forming hydrogen bonds (amino) could lead to an increase in the melting point. Note that the melting point changes upon fluorine and chlorine substitution exhibit opposite trends. One possible explanation is that smaller but more electronegative fluorine atoms alter the inter- and intra-molecular interaction environment.

Finally, we can offer some suggestions on molecule design. If one desires to reduce the melting point of PNs, introducing flexible groups to increase the flexibility of the main chain in the molecular structure, such as ether bonds, can be effective. Introducing silicon-containing structures such as siloxanes will likely lower the melting point and serve as a suitable alternative. Introducing some unique structures, such as phosphorus oxide, may also reduce the melting point, but more experimental research on its impact must be strengthened. Conversely, introducing rigid and large structural units will significantly increase the melting point of PNs. Substituents with high electronegativity often increase the melting point in most cases and should be avoided during synthesis. Additionally, attention should be paid to the influence of introduced chemical units on the melting point when combining other properties into the materials. A good example is the introduction of fluorine structures on the main chain to enhance the dielectric performance of PN, which may substantially increase the melting point to over 500 K^[34].

To design PN with a specific melting point in the future, one can refer to the method in the section of “Structural analysis based on MEGNet models” and the gene combination method shown in [Figure 2](#). The methodology for industrial partners to design PN-based materials is as follows: (1) change the R and substituent groups; (2) carry out high-throughput prediction; and (3) screen the appropriate PN according to the threshold, synthesis feasibility, and cost. As the PN candidate is selected, the melting point is measured for the synthesized PN, and the experimental validation is performed by comparing with the predicted and experimental values. The immediate step is to update the model to improve the performance further by incorporating new data from experimental validation.

The principle involved in constructing our model can, in principle, be extended to studying other material properties, serving as a model for performance research that can be further promoted or integrated, ultimately achieving a powerful, universal predictive capability. However, the challenge lies in the fact that, despite considering applicability and generality as much as possible during the construction of the predictive model, its generalization ability remains significantly insufficient when faced with the vast and diverse material space. We only selected random combinations of ten linking groups and 11 substituents, including nine elements on three templates as the learning objects. This limits the types of materials that can be predicted and is far from enough to cover other substances. Predictive accuracy cannot be guaranteed if we forcefully extend our model. The future direction of this research could focus on further broadening the applicability of the predictive model by breaking the template constraint, allowing the prediction of any property of materials with arbitrary elements, configurations, and chemical units, aiming for universality.

CONCLUSIONS

In this work, we established two multi-fidelity models for predicting the melting point of PNs. Confronted with limited experimental data, we expanded the dataset using all-atom molecular dynamic simulation. Incorporated in the concept of multi-fidelity learning, we effectively utilized the high-fidelity experimental data and low-fidelity simulation data and stepwise or collectively applied them in ML approaches. The combined dataset effectively addressed the “small data” dilemma, and the well-trained multi-fidelity model demonstrated excellent generalization capabilities, which can assist in screening a range of PNs with low melting points. Structural analysis yielded insights into the contributions of chemical units to lowering the melting points, offering guidance for the future design of such materials. This work can contribute to accelerating the discovery of easy-to-processing PN resins.

DECLARATIONS

Authors' contributions

Writing - original draft, methodology, investigation, formal analysis, data curation: Xu B, Hu X

Writing - original draft, investigation: Lan H

Software, methodology, investigation: Wang T, Xu XY

Writing - review and editing, conceptualization: Zhang C

Writing - review and editing, project administration, conceptualization: Lin J

Writing - review and editing, supervision, project administration, formal analysis, methodology, conceptualization: Wang L

Supervision, conceptualization: Du L

Availability of data and materials

The original data is provided in the [Supplementary Materials](#). Full details include collection of experimentally measured melting points of PNs, generation of candidate space for PNs through gene combination, *t*-SNE, and random sampling, molecular dynamic simulation of PN melting points, Gaussian

process regression, constructing models by Gaussian process regression, and graph machine learning, MPNN and MEGNet.

Financial support and sponsorship

This work is supported by the National Key R&D Program of China (2022YFB3707302), the National Natural Science Foundation of China (22173030, 52394271), and Shanghai Scientific and Technological Innovation Projects (22ZR1417500).

Conflicts of interest

Zhang C is affiliated with Shanghai Aerospace Equipment Manufacturer Co., Ltd., while the other authors have declared that they have no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2024.

REFERENCES

- Jia Y, Bu X, Dong J, et al. Catalytic polymerization of phthalonitrile resins by carborane with enhanced thermal oxidation resistance: experimental and molecular simulation. *Polymers* 2022;14:219. [DOI](#) [PubMed](#) [PMC](#)
- Zhang H, Yan Z, Yang Z, Mu Q, Peng D, Zhao H. Synthesis, curing and thermal properties of the low melting point phthalonitrile resins containing glycidyl groups. *Polym Bull* 2023;80:725-38. [DOI](#)
- Mouritz A, Gellert E, Burchill P, Challis K. Review of advanced composite structures for naval ships and submarines. *Compos Struct* 2001;53:21-42. [DOI](#)
- Bai S, Sun X, Chen X, Yu X, Zhang Q. Synthesis and properties of a thioether bonded phthalonitrile resin. *Mater Today Commun* 2020;24:101352. [DOI](#)
- Liu C, Zhang B, Sun M, et al. Novel low-melting bisphthalonitrile monomers: Synthesis and their excellent adhesive performance. *Eur Polym J* 2021;153:110511. [DOI](#)
- Wang H, Wang J, Guo H, et al. A novel high temperature vinylpyridine-based phthalonitrile polymer with a low melting point and good mechanical properties. *Polym Chem* 2018;9:976-83. [DOI](#)
- Peng W, Yao F, Hu J, et al. Renewable protein-based monomer for thermosets: a case study on phthalonitrile resin. *Green Chem* 2018;20:5158-68. [DOI](#)
- Gu H, Gao C, Du A, et al. An overview of high-performance phthalonitrile resins: fabrication and electronic applications. *J Mater Chem C* 2022;10:2925-37. [DOI](#)
- Wu Z, Wang S, Zong L, Li N, Wang J, Jian X. Novel phthalonitrile-based composites with excellent processing, thermal, and mechanical properties. *High Perform Polym* 2018;30:720-30. [DOI](#)
- Nechausov S, Aleksanova A, Morozov O, Bulgakov B, Babkin A, Kepman A. Low-melting phthalonitrile monomers containing maleimide group: synthesis, dual-curing behavior, thermal and mechanical properties. *React Funct Polym* 2021;164:104932. [DOI](#)
- Hu Y, Weng Z, Qi Y, et al. Self-curing triphenol A-based phthalonitrile resin precursor acts as a flexibilizer and curing agent for phthalonitrile resin. *RSC Adv* 2018;8:32899-908. [DOI](#) [PubMed](#) [PMC](#)
- Dominguez DD, Keller TM. Properties of phthalonitrile monomer blends and thermosetting phthalonitrile copolymers. *Polymer* 2007;48:91-7. [DOI](#)
- Babkin AV, Zdobinov EB, Bulgakov BA, Kepman AV, Avdeev VV. Low-melting siloxane-bridged phthalonitriles for heat-resistant matrices. *Eur Polym J* 2015;66:452-7. [DOI](#)
- Zhang J, Feng Y, Yuan H, Feng D, Zhang X, Wang G. Thermal properties of C17H36/MCM-41 composite phase change materials. *Comput Mater Sci* 2015;109:300-7. [DOI](#)
- Qiao Z, Feng H, Zhou J. Molecular dynamics simulations on the melting of gold nanoparticles. *Phase Transit* 2014;87:59-70. [DOI](#)
- Li JF, Zhao XP, Liu J. Molecular dynamics simulations on melting of aluminum. *Appl Mech Mater* 2013;423-6:935-8. [DOI](#)
- Ganz E, Ganz AB, Yang LM, Dornfeld M. The initial stages of melting of graphene between 4000 K and 6000 K. *Phys Chem Chem Phys* 2017;19:3756-62. [DOI](#) [PubMed](#) [PMC](#)

18. Liu Y, Lai W, Yu T, Ma Y, Guo W, Ge Z. Melting point prediction of energetic materials via continuous heating simulation on solid-to-liquid phase transition. *ACS Omega* 2019;4:4320-4. [DOI](#)
19. Du S, Zhang S, Wang L, Lin J, Du L. Polymer genome approach: a new method for research and development of polymers. *Acta Polym Sin* 2022;53:592-607. [DOI](#)
20. Yang Z, Nie W, Liu L, Xu X, Xia W, Xu W. Applications of machine learning methods in the studies of polymer glass formation. *Acta Polym Sin* 2023;54:432-50.
21. Gong X, Jiang Y. Advances and challenges of machine learning in polymer material genomes. *Acta Polym Sin* 2022;53:1287-300. [DOI](#)
22. Li Y, Jiang Y, Wang L, Li J. Data and machine learning in polymer science. *Chin J Polym Sci* 2023;41:1371-6. [DOI](#)
23. Habibi-Yangjeh A, Pourbasheer E, Danandeh-Jenagharad M. Prediction of melting point for drug-like compounds using principal component-genetic algorithm-artificial neural network. *Bull Korean Chem Soc* 2008;29:833-41. [DOI](#)
24. Qu N, Liu Y, Liao M, et al. Ultra-high temperature ceramics melting temperature prediction via machine learning. *Ceram Int* 2019;45:18551-5. [DOI](#)
25. Hu Y, Zhao W, Wang L, Lin J, Du L. Machine-learning-assisted design of highly tough thermosetting polymers. *ACS Appl Mater Interfaces* 2022;14:55004-16. [DOI](#) [PubMed](#)
26. Xu X, Zhao W, Hu Y, et al. Discovery of thermosetting polymers with low hygroscopicity, low thermal expansivity, and high modulus by machine learning. *J Mater Chem A* 2023;11:12918-27. [DOI](#)
27. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater* 2019;31:3564-72. [DOI](#)
28. Zeng K, Yang G. Phthalonitrile matrix resins and composites. In: *Wiley Encyclopedia of Composites*. Wiley; 2011. pp. 1-14. [DOI](#)
29. Laskoski M, Dominguez DD, Keller TM. Synthesis and properties of a bisphenol A based phthalonitrile resin. *J Polym Sci A Polym Chem* 2005;43:4136-43. [DOI](#)
30. Shi X, Bai S, Ji P, Naito K, Yu X, Zhang Q. A phthalonitrile resin with a low melting point and high storage modulus containing high-density aromatic ether bonds. *ChemistrySelect* 2020;5:12213-7. [DOI](#)
31. Liu K, Li Y, Tao L, Liu C, Xiao R. Synthesis and characterization of inherently flame retardant polyamide 6 based on a phosphine oxide derivative. *Polym Degrad Stabil* 2019;163:151-60. [DOI](#)
32. Łaskiewicz B. Preparation of organophosphorus polyurethane block copolymers. *J Appl Polym Sci* 1967;11:2295-301. [DOI](#)
33. Jain P, Choudhary V, Varma IK. Effect of phosphorus content on thermal behaviour of diglycidyl ether of bisphenol-A/phosphorus containing amines. *J Therm Anal Calorim* 2022;67:761-72. [DOI](#)
34. Sastri SB, Keller TM. Phthalonitrile polymers: cure behavior and properties. *J Polym Sci A Polym Chem* 1999;37:2105-11. [DOI](#)