



Article

Extrapolative Prediction of Propagation Rate Coefficient in Free-Radical Polymerization Based on $k_p(T, NI)$ -QSPR Model

Yajuan Shi¹, Xiaojie Feng², Fangyou Yan^{2,*}, Zheng-Hong Luo¹ and Yin-Ning Zhou^{1,*}¹ State Key Laboratory of Polyolefins and Catalysis, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China² School of Chemical Engineering and Material Science, Tianjin University of Science and Technology, Tianjin 300457, China* Correspondence: yanfangyou@tust.edu.cn (F.Y.); zhouyn@sjtu.edu.cn (Y.-N.Z.)**How To Cite:** Shi, Y.; Feng, X.; Yan, F.; et al. Extrapolative Prediction of Propagation Rate Coefficient in Free-Radical Polymerization Based on $k_p(T, NI)$ -QSPR Model. *Smart Chemical Engineering* **2026**, *2*(3), 8. <https://doi.org/10.53941/sce.2026.100008>

Received: 29 May 2026

Revised: 15 September 2026

Accepted: 16 September 2026

Published: 22 September 2026

Abstract: This work assesses the extrapolation performance of the $k_p(T, NI)$ -QSPR model for free-radical polymerization. Although QSPR models often achieve high interpolation accuracy, their reliability for predicting beyond the training domain, a situation common in monomer reactivity exploration, remains a critical challenge. To address this, prediction set is classified into interpolation and extrapolation categories using descriptor visualization. Extrapolation validation (EV) is employed to evaluate model reliability, defining an extrapolation degree (ED) to quantify how far test samples lie outside the training domain. Results demonstrate that the model achieves excellent interpolation accuracy but shows limited reliability when descriptors fall beyond the training set range. Descriptors suitable for forward or backward extrapolation are identified, with backward extrapolation yielding relatively better predictions for certain monomers. Uneven descriptor distributions are revealed as the primary cause of high ED and reduced extrapolation performance. This work provides a practical framework for assessing the applicability domain of the $k_p(T, NI)$ -QSPR model. More importantly, the established EV approach provides a transferable framework for applicability domain assessment and offers critical guidance for improving extrapolation robustness, thereby laying a foundation for extending such predictive methods to other temperature-dependent kinetic parameters in free-radical polymerization.

Keywords: QSPR model; propagation rate coefficient; extrapolation performance; Arrhenius parameter; free-radical polymerization

1. Introduction

In recent years, the rapid development of artificial intelligence (AI) has enriched the modeling of quantitative structure-property relationships (QSPR), which can figure out the complicated correlation between structure and molecular property well [1–5]. Attractively, supervised machine learning (ML) algorithms have been largely used for property prediction and process optimization in the chemical engineering community [6–15].

Molecular descriptors play a central role in these models by translating molecular structures into numerical representations. In addition to conventional descriptors that encode specific structural and physicochemical characteristics, fingerprint-based representations such as Extended-Connectivity Fingerprints (ECFPs) have also been widely used for structure-property modeling [16,17]. The choice of molecular representation determines the structural information available to the model and therefore may influence its applicability beyond the training



Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

domain. Although QSPR models often exhibit satisfactory predictive performance within their training domains, their reliability for extrapolative prediction of structurally or descriptively distinct molecules remains a critical challenge.

In the field of polymer reaction engineering (PRE), accurate kinetic parameters are extremely important for determining polymerization rate, the polymer molecular average properties, and the chemical composition of copolymer [18–21]. The propagation rate coefficient (k_p), as one of the significant rate coefficients in free-radical polymerization (FRP), is closely related to monomer structure [22,23] and can be obtained by experimental and computational methods [24–26]. As an IUPAC-recommended technique, the pulsed laser polymerization-size exclusion chromatography (PLP-SCE) approach has been widely employed to determine k_p values of a large range of monomers (e.g., methacrylates, acrylates, styrene, etc.) [27–32]. Common computational methods include quantum chemical calculation (QCC) [33,34] and kinetic model-based regression analysis [35–37]. QCC method for solution-phase systems is generally based on transition state theory, while reliable rate coefficients should be treated with care using certain corrections. The kinetic model-based regression analysis requires high-quality experimental data in advance to provide trustworthy results.

As an alternative approach, ML-driven strategies have been adopted to obtain key reaction/polymerization kinetic parameters [38–47]. For gaining knowledge of radical propagation, Nguyen and Bavarian employed a Graph Attention Network (GAN) method to train a multi-task learning model for calculating reactivity ratios of monomers (i.e., $r_1 = k_{11}/k_{12}$, a measure of the (co)monomer insertion tendency) in copolymerization; the interpretable characteristics of the Multi-Input-Multi-Output GAN model facilitate the accurate prediction of the reactivity ratios [48]. Also, for monomer reactivity ratios, Farajzadehahary et al. applied different ML methods, including artificial neural network (ANN), decision tree (DT), random forest (RF), and extreme gradient boosting (XGBoost), to fit the predictive targets. The results indicated that ANN shows better performance than other ML methods [49]. While for k_p estimation, our group developed a so-called k_p (T, *N*)-QSPR model based on the multiple linear regression (MLR) method, which enables the use of one unified formula to predict k_p values at variable temperatures. The calculated k_p values are in good agreement with the benchmarking k_p values measured by PLP-SEC, with $R^2 = 0.9974$ for the training set [50]. Later on, Van de Reydt et al. developed a ML model that enables correlation of k_p with basic molecular properties using the ridge regression (Ridge) method, which predicts k_p values with R^2 of 0.985 for various monomers at 25 °C. The authors further developed three additional models for predicting k_p values at 50/75/100 °C, respectively, which allows to calculate Arrhenius activation parameters [40].

Although the MLR and Ridge models exhibit satisfactory predictive performance within their respective training domains, their extrapolation capability for monomers with significantly different structures or descriptors lying outside the training set range has not been systematically evaluated. Extrapolation—predicting beyond the domain of the training data—remains a critical yet often overlooked challenge in scientific machine learning [1,2]. Although complex models typically excel at interpolation (e.g., random k-fold cross-validation), their reliability for extrapolation is frequently unsatisfactory [1]. Recent studies have proposed various strategies to assess and improve extrapolation performance. Muckley et al. found that simple linear models can outperform black-box algorithms in approximately 40% of extrapolation tasks, challenging the widely held assumption of an inherent trade-off between interpretability and extrapolation accuracy [51]. Wang and King introduced a pairwise ranking formulation specifically designed for extrapolation, which substantially outperformed standard regression in identifying samples with target values beyond the training range [52]. Ambrosio Garcia and Naets [53] developed a leave-boundary-out (LBO) data splitting method to explicitly select models with superior extrapolation performance for predictions at the boundaries of the training distribution. In the context of flame describing functions, McCartney et al. [54] demonstrated that the extrapolation performance of Gaussian processes degrades rapidly as the distance from the training domain increases, underscoring the importance of expanding the covered domain through strategic data selection. In the field of property prediction, Feng et al. developed an enhanced extrapolation degree analysis for each descriptor by incorporating the normal boiling point into their Antoine-based QSPR model, thereby enabling a more comprehensive assessment of the model's extrapolation capability [55]. Collectively, these studies highlight the necessity of systematically evaluating extrapolation performance and defining the applicability domain of predictive models.

Our previous work [50] demonstrated the satisfactory interpolation accuracy of the k_p (T, *N*)-QSPR model, it did not address the model's reliability for monomers whose descriptors fall outside the training set domain. This gap is critical because extrapolation prediction, rather than interpolation, is the primary scenario in practical monomer screening. Therefore, clearly defining the applicability domain and quantifying extrapolation reliability are essential for the robust generalization of QSPR methods to new monomers. The present work systematically evaluates the extrapolation performance of our previously developed k_p (T, *N*)-QSPR model [50], following the workflow shown in Figure 1. Prediction set monomers are first classified into interpolation and extrapolation categories via descriptor visualization. Extrapolation validation (EV) is then applied to quantify the extrapolation

degree (ED) for each descriptor, distinguishing forward from backward extrapolation. The uneven distribution of descriptors is identified as the main cause of reduced extrapolation reliability. Furthermore, we determine which descriptors are suitable for forward or backward extrapolation and analyze the underlying reasons for the inaccurate k_p predictions of some new monomers. Beyond evaluating a specific model, this workflow provides generalizable guidelines for applicability domain assessment in QSPR modeling for predicting monomer reactivity, thereby offering a practical tool for researchers to judge the trustworthiness of predictions when screening structurally novel monomers.

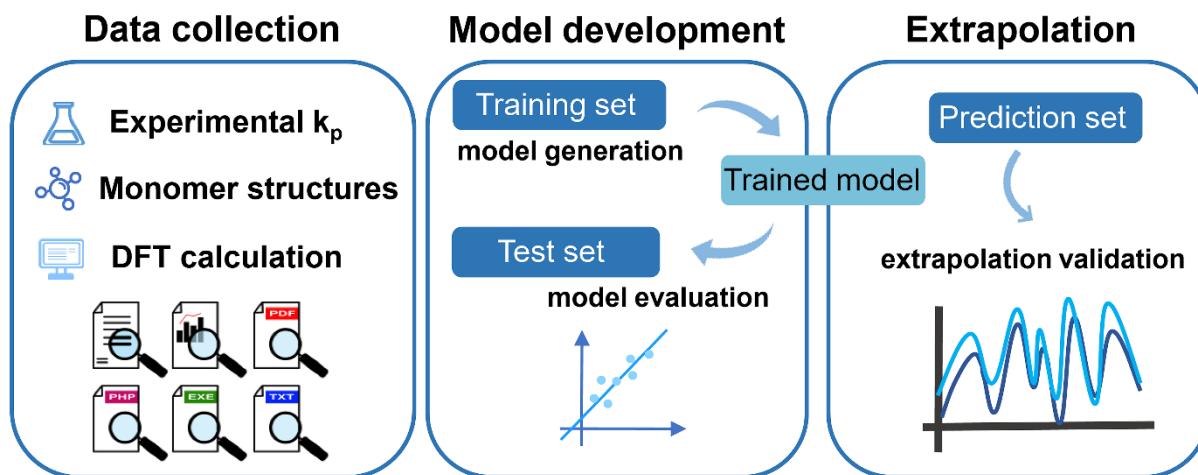


Figure 1. Workflow of assessing extrapolation performance of the k_p (T, NI)-QSPR model.

2. Methodology

2.1. Dataset and Features

Figure 2 presents the monomer structures used in the model development. The dataset covers linear and branched methacrylates, linear and branched acrylates, nitrogen-containing methacrylates, hydroxyl-containing (meth)acrylates, styrene, vinyl acetate, N-vinyl formamide, 1,3-butadiene, and acrylonitrile. The training and test sets of k_p of 33 monomers follow our previous work [50], and the prediction set containing 12 monomers is collected from Van de Reydt et al.'s work [40]. As such, at different temperatures, 26 monomers including 454 data points are used as the training set, 7 monomers covering 111 data points are used as the test set, and 12 monomers with 48 data points are used as the prediction set. To reduce the range and symmetrize the response, the k_p values are taken the natural logarithm. To exclude the influence of aqueous or organic systems on k_p values, the k_p values used in this work are from bulk polymerization. The training set is used to construct the models, the test set is employed for external validation, and the prediction set is used to assess the external predictive ability. The norm descriptors (NI_1 – NI_{12}) used in this work were derived from atomic distribution matrices constructed from monomer structures. Briefly, the calculation involves extracting atomic connection relationships and Euclidean spatial distances from the optimized molecular geometry, combined with atomic properties including natural population charges and electrostatic potentials obtained from quantum chemical calculations. Detailed procedures for descriptor calculation, including matrix construction and norm extraction, have been described in our previous work [50]. All data are listed in Supplementary Materials Table S1.

2.2. Descriptor Visualization

To evaluate the applicability of the model, we propose a method for descriptor visualization to judge whether the monomers belong to interpolation or extrapolation prediction. First, the input variables (i.e., norm descriptors) of the training set are normalized to a range from 0 to 1, establishing the training domain boundary (Figure 3). Second, the normalized norm descriptors of the prediction set are plotted on the same scale. Third, for each monomer in the prediction set, the values of descriptors are compared against the [0, 1] range: if all normalized norm descriptors for a given monomer lie within this range, the prediction for this monomer is classified as interpolation; otherwise, it is classified as extrapolation. As visualized in Figure 3, monomers iBoA (No. pre3), INA-A (No. pre4), and PnMA (No. pre9) are interpolation predictions with all normalized norm descriptors in the range from 0 to 1, while the other nine monomers (C17A, C21A, HCEA, HCPA, PhCEA, PhCPA, BuDE, AN, and NVF) belong to extrapolation predictions.

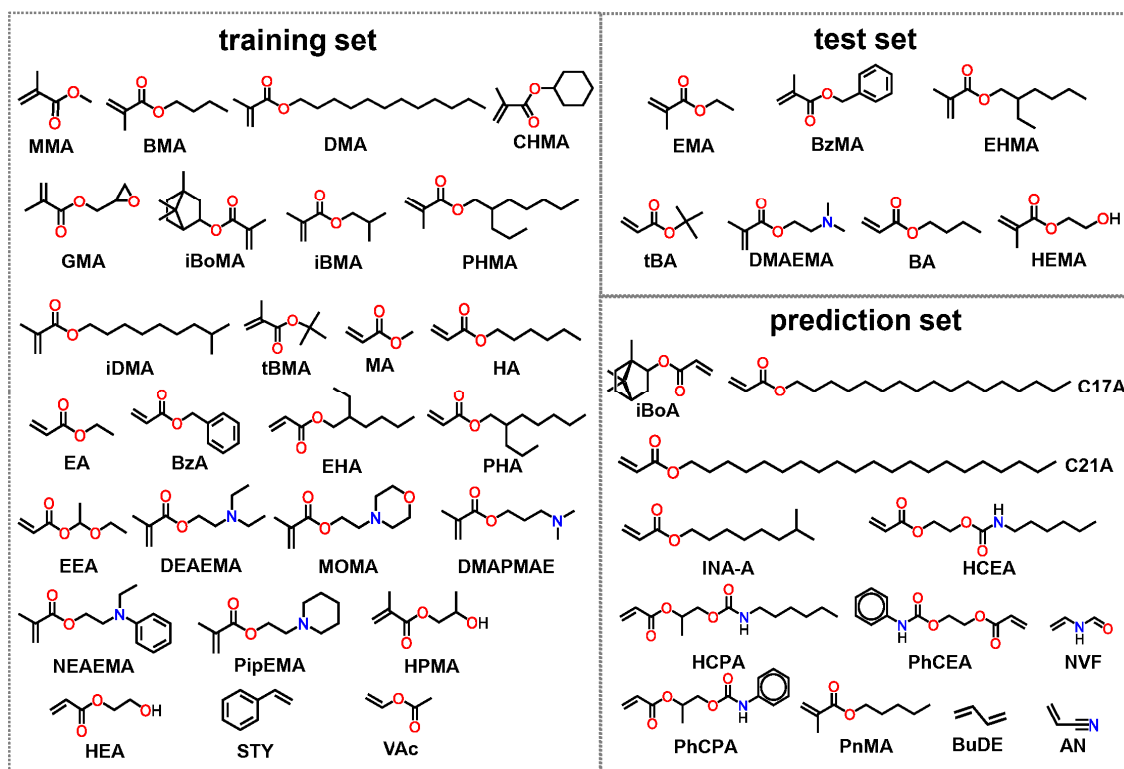


Figure 2. Monomer structures for training set, test set, and prediction set.

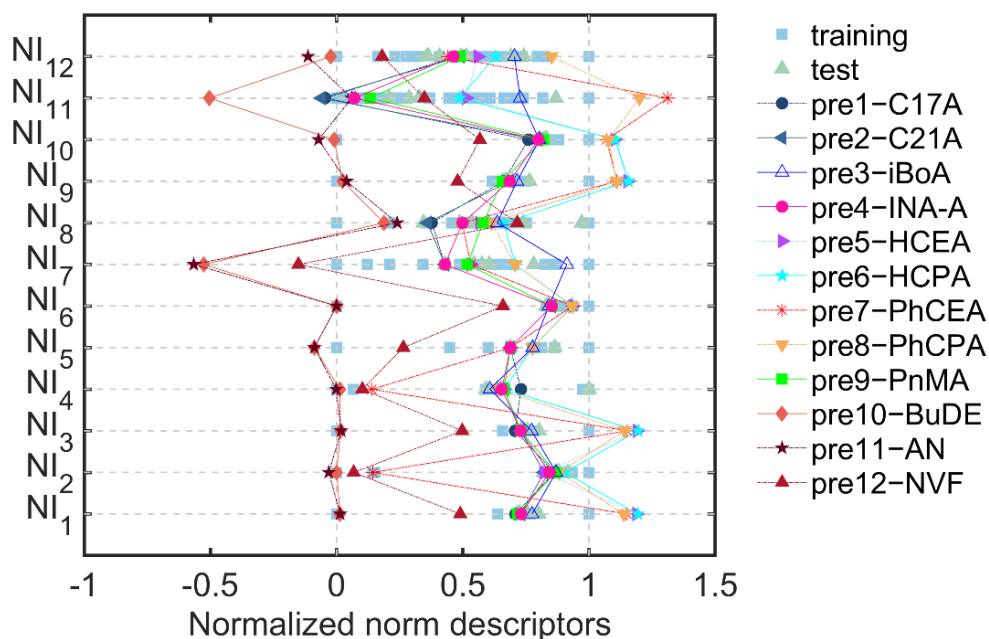


Figure 3. Distribution of normalized norm descriptors.

It is worth clarifying that the descriptor visualization method addresses a different question than conventional applicability domain assessment tools. The visualization method diagnoses whether a given monomer lies inside or outside the training descriptor domain (i.e., interpolation vs. extrapolation), based solely on individual descriptors. In contrast, conventional applicability domain methods, such as the Williams plot, leverage analysis, or distance-based metrics, generally characterize the applicability domain of the model based on the overall training dataset and model characteristics. These approaches are not designed to assess the extrapolation behavior of individual descriptors or to determine whether a specific descriptor of a new, out-of-domain compound undergoes forward or backward extrapolation. This method introduces a descriptor-level tool that directly identify extrapolation risk of model application.

2.3. Extrapolation Validation Analysis

Herein, extrapolation validation (EV) [56] is applied to evaluate the extrapolation ability of the linear-based models. This method sorts each independent variable of the training set, and then re-divides the training set and test set according to the sorted independent variable in a ratio of 4:1 (the first 80% is used as the training set_EV, and the last 20% is used as the test set_EV). Considering the influence of the independent variables, the model is evaluated using forward and backward extrapolation, where the samples are sequentially ordered according to the descriptor values in ascending and descending order, respectively. The extrapolation ability of the model is evaluated by refitting the model using the training set_EV and then analyzing the results of the test set_EV. In addition, the extrapolation degree (ED) is defined to quantitatively represent the degree of the independent variable for the test set_EV outside the domain of the training set_EV. When the performance of the test set_EV is good and the ED is high, it shows that even for the samples far away from the training set domain, the predicted value of the model is reliable. It should be noted that ED itself does not have a universal threshold for defining reliable extrapolation. The acceptable extrapolation range is model- and descriptor-dependent, and can be established by relating ED to the corresponding prediction error ($RMSE_{test}(EV)$). In this work, this relationship is used to identify descriptors suitable for forward or backward extrapolation, providing descriptor-specific guidance for model application.

2.4. Multiple Linear Regression (MLR)

The MLR model was chosen for this extrapolation analysis because its linear and additive structure enables transparent decoupling of individual descriptor contributions and does not require hyperparameter tuning. The method can provide an interpretable description of how the dependent variables affect the predicted variables [57,58]. The MLR model with 12- explainable variables and the related coefficients are shown in Table S2, the same as in our previous work [50]. The calculation of input variables is from the definition of norm descriptors in k_p (T, NI)-QSPR model, with a brief computational procedure shown in Figure 4, T and their detailed calculation procedure are available in the Supporting Information.

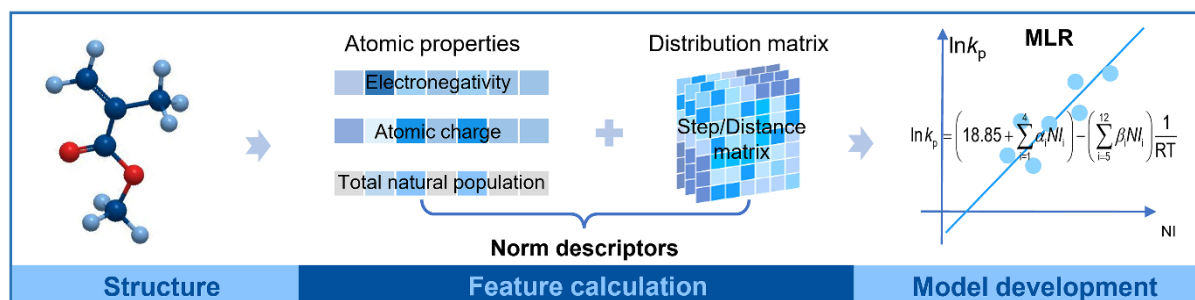


Figure 4. Procedure for MLR-based QSPR model.

2.5. Validation Metrics

Validation metrics are important criteria to evaluate the quality of the as-developed model. Herein, statistical parameters, including coefficient of determination (R^2 , Equation (S11)), root-mean-square error ($RMSE$, Equation (S12)), and mean absolute error (MAE , Equation (S13)) are employed to assess how well the as-developed models predict the k_p values. For an ideal model, R^2 should be close to 1, MAE and $RMSE$ should be 0, which indicates perfect forecasts [59,60].

3. Results and Discussion

3.1. $k_p(T)$ Prediction for Monomers in Prediction Set

The as-developed MLR-based QSPR model is used to assess the $\ln k_p$ values of 12 monomers at 25/50/75/100 °C. Figure 5 presents the calculated $\ln k_p$ values versus experimental $\ln k_p$ values for extrapolation and interpolation predictions. The statistical parameters are listed in Table 1, it can be shown from R^2 , $RMSE$, and MAE that the results for interpolation prediction are better than extrapolation prediction. It can be concluded that the MLR-based model achieves good interpolation performance ($R^2 = 0.9710$), while its extrapolation performance ($R^2 = 0.7950$) is moderate and requires careful interpretation, particularly for monomers whose descriptors lie far outside the training domain.

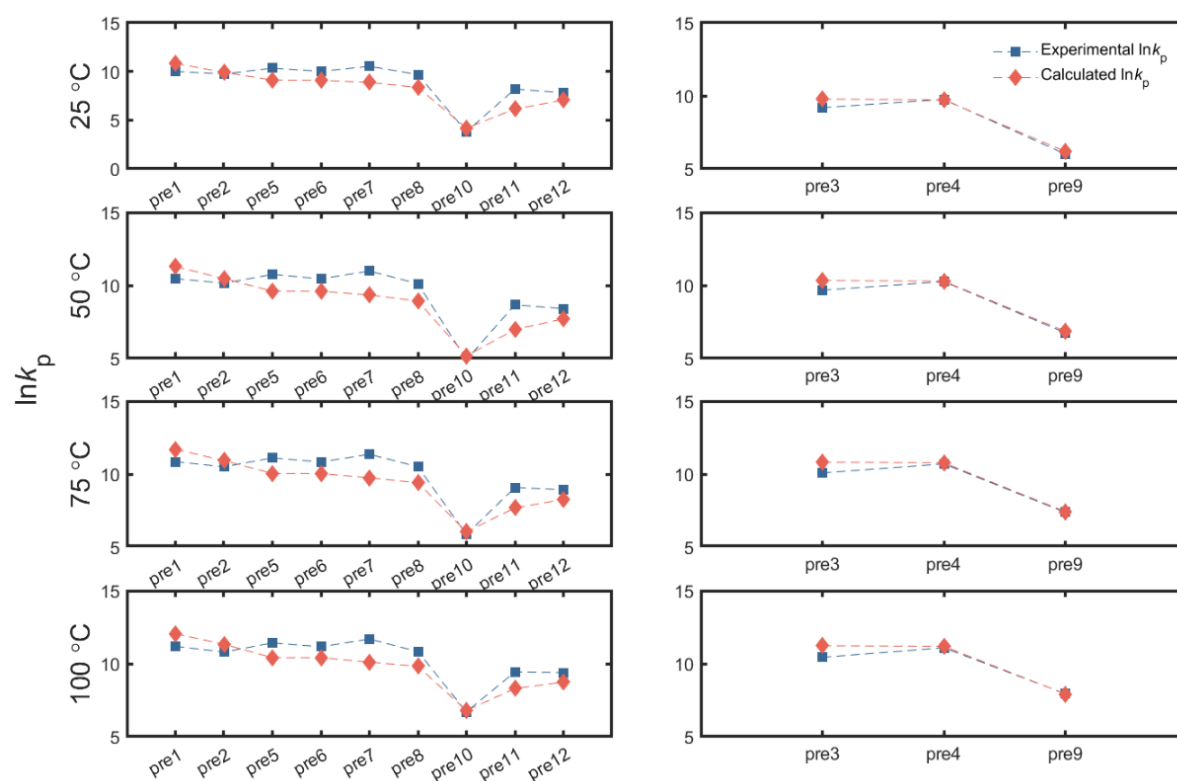


Figure 5. Calculated $\ln k_p$ values versus experimental $\ln k_p$ values for the prediction set.

Table 1. Statistical parameters for interpolation and extrapolation predictions.

Method	Extrapolation Prediction			Interpolation Prediction		
	R^2	RMSE	MAE	R^2	RMSE	MAE
MLR	0.7950	1.0474	0.9334	0.9710	0.4076	0.2780

3.2. Calculation of Arrhenius Parameters

The as-developed MLR-based QSPR model can not only predict the k_p values at variable temperatures but also provide the related Arrhenius parameters (i.e., pre-exponential factor, A , and activation energy, E_a) of $\ln k_p$ for various monomers. Herein, based on the MLR-based QSPR model, the Arrhenius parameters of 12 monomers are obtained through regressing the calculated $\ln k_p$ values. Figure 6 presents the experimental and calculated $\ln k_p$ values versus $1/T$, and their Arrhenius equations calculated by the MLR-based model. Table 2 lists the MAE and $RMSE$ for 12 monomers in the prediction set. It is suggested that the MLR-based QSPR model has near-perfect goodness-of-fit both in calculating $\ln k_p$ values and fitting Arrhenius parameters for the interpolation prediction set (i.e., iBoA, INA-A, PnMA, the five-pointed stars marked as blue squares in Figure 6). For the extrapolation prediction set, the calculated $\ln k_p$ values and Arrhenius parameters seem to have larger errors than the interpolation prediction monomers, which indicates that the application ability of the model is limited. The model, based on limited data, cannot accurately calculate the k_p value of all monomers, but the calculation of monomers with similar structures is accurate.

Table 2. Statistical parameters for the prediction set.

Monomer	MSE	RMSE	MAE	Monomer	MSE	RMSE	MAE
Pre1-C17A	0.6971	0.8349	0.8347	Pre7-PhCEA	2.6777	1.6364	1.6361
Pre2-C21A	0.1071	0.3273	0.2854	Pre8-PhCPA	1.4961	1.2232	1.2156
Pre3-iBoA	0.4259	0.6526	0.6461	Pre9-PnMA	0.0208	0.1442	0.1187
Pre4-INA-A	0.0035	0.0592	0.0513	Pre10-BuDE	0.0938	0.3063	0.2813
Pre5-HCEA	1.3370	1.1563	1.1523	Pre11-AN	3.2125	1.7923	1.7474
Pre6-HCPA	0.7705	0.8778	0.8739	Pre12-NVF	0.5133	0.7165	0.7148

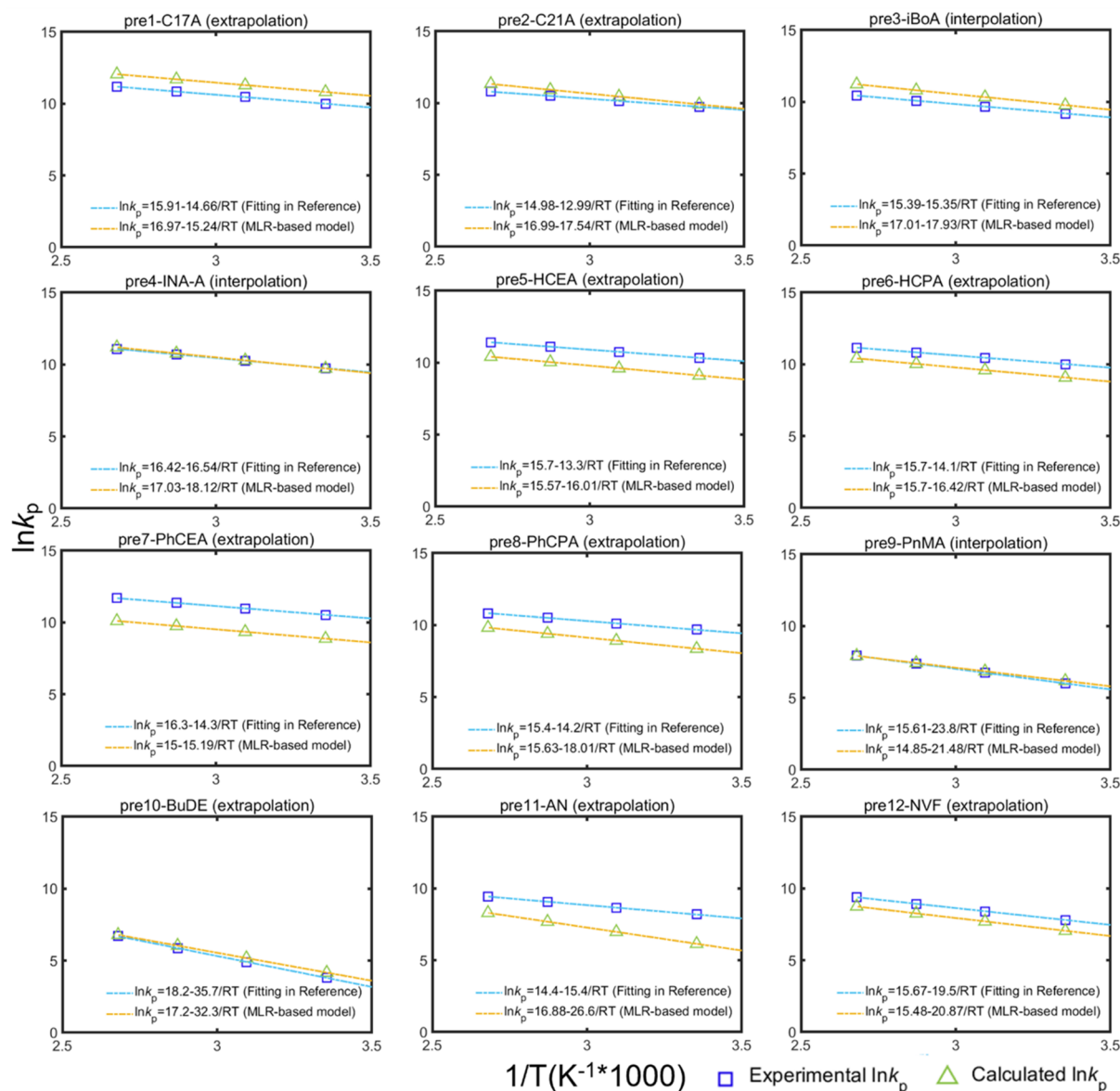


Figure 6. $\ln k_p$ values versus $1/T$ for 12 monomers in the prediction set.

3.3. EV Results

Figure 7 presents the EV result of the MLR model, with each element carrying specific diagnostic information. First, the red dotted line represents the standard deviation of the samples within the 95% confidence interval. The black dashed line represents the $RMSE$ of the training set. Second, the bar represents the $RMSE$ of the test set (Model) in EV of this descriptor. The circle represents the $RMSE$ of the test set (EV). Third, the left and right color bars indicate the extrapolation degree (ED), with darker colors representing higher ED values. Fourth, comparing these elements reveals that descriptors NI_5 , NI_6 , and NI_7 have much larger $RMSE_{test}(EV)$ than both $RMSE_{test}(model)$ and σ_{95} , indicating that they are unsuitable for backward extrapolation. Combined with the data distribution in Figure 3, it can be concluded that uneven descriptor distributions lead to high ED and poor extrapolation ability.

The descriptors NI_1 – NI_{11} are appropriate for forward extrapolation. However, descriptors NI_5 , NI_6 , and NI_7 have a much larger $RMSE$ in the test set (EV) than in the model test set (Model), so they are not suitable for backward extrapolation. As previously shown, Figure 3 suggests that 9 monomers (C17A, C21A, HCEA, HCPA, PhCEA, PhCPA, BuDE, AN and NVF) belong to extrapolation prediction. The descriptors for the 9 monomers during extrapolation are summarized in Table 3. The normalized descriptor NI_{11} of C17A and C21A is less than 0, thus descriptor NI_{11} for these two monomers belongs to backward extrapolation. The normalized descriptors NI_1 , NI_3 , NI_9 , and NI_{10} of HCEA and HCPA are larger than 1; thus, these descriptors for monomers HCEA and HCPA belong to forward extrapolation. Other monomer descriptors are also analyzed according to this method.

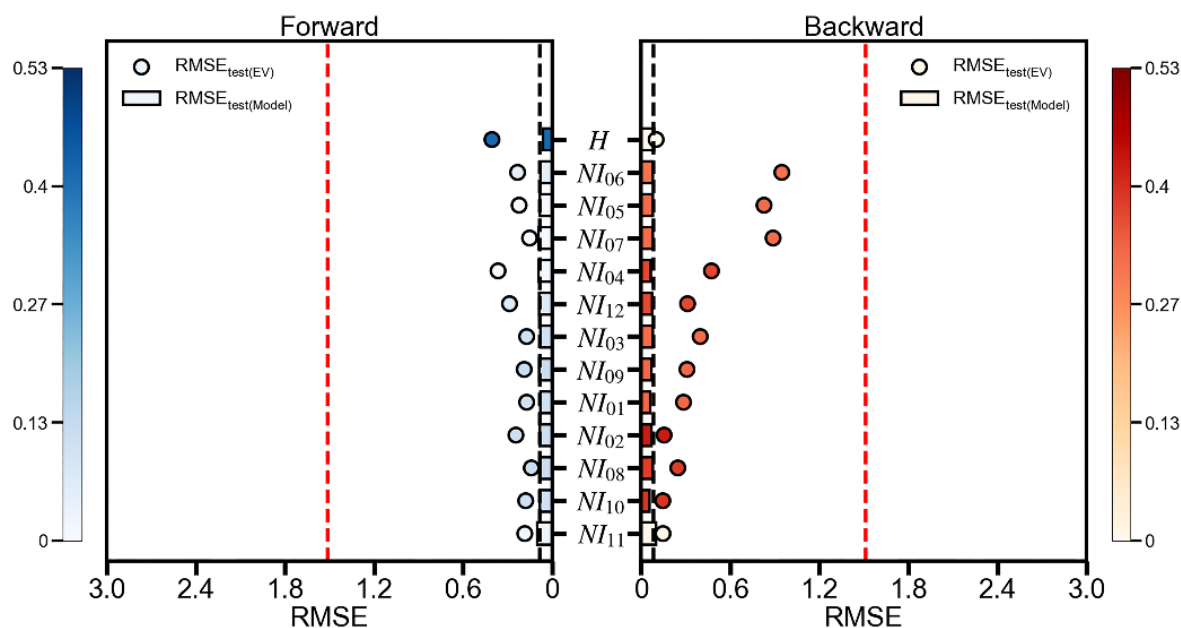


Figure 7. EV results for the MLR-based QSPR model with explanations detailed in the text (*H* denotes as the leverage value for the descriptors).

Table 3. Summary of descriptors for the 9 monomers during extrapolation.

Monomers	Forward Extrapolation	Backward Extrapolation	Interpolation
C17A	-	NI_{11}	$NI_1, NI_2, NI_3, NI_4, NI_5, NI_6, NI_7, NI_8, NI_9, NI_{10}, NI_{12}$
C21A	-	NI_{11}	$NI_1, NI_2, NI_3, NI_4, NI_5, NI_6, NI_7, NI_8, NI_9, NI_{10}, NI_{12}$
HCEA	$NI_1, NI_3, NI_9, NI_{10}$	-	$NI_2, NI_4, NI_5, NI_6, NI_7, NI_8, NI_{11}, NI_{12}$
HCPA	$NI_1, NI_3, NI_9, NI_{10}$	-	$NI_2, NI_4, NI_5, NI_6, NI_7, NI_8, NI_{11}, NI_{12}$
PhCEA	$NI_1, NI_3, NI_9, NI_{10}, NI_{11}$	-	$NI_2, NI_4, NI_5, NI_6, NI_7, NI_8, NI_{12}$
PhCPA	$NI_1, NI_3, NI_9, NI_{10}, NI_{11}$	-	$NI_2, NI_4, NI_5, NI_6, NI_7, NI_8, NI_{12}$
BuDE	-	$NI_5, NI_7, NI_{10}, NI_{11}, NI_{12}$	$NI_1, NI_2, NI_3, NI_4, NI_6, NI_8, NI_9$
AN	-	$NI_2, NI_5, NI_7, NI_{10}, NI_{12}$	$NI_1, NI_3, NI_4, NI_6, NI_8, NI_9, NI_{11}$
NVF	-	NI_7	$NI_1, NI_2, NI_3, NI_4, NI_5, NI_6, NI_8, NI_9, NI_{10}, NI_{11}, NI_{12}$

To quantitatively assess the unevenness of descriptor distributions, the variance confidence intervals of each descriptor in the training set were calculated (Table S3). The interval widths vary from 0.0299 (NI_9) to 0.0938 (NI_{11}). For NI_5 and NI_6 , although their interval widths are moderate (0.0372 and 0.0387, respectively), their normalized values are concentrated near the upper boundary of the training domain (Figure S1, close to 1). This implies that the training set contains very few monomers with low NI_5 or NI_6 values, leaving the lower range of these descriptors largely unrepresented. Consequently, as shown in Figure 3, when a test monomer (e.g., BuDE) possesses NI_5 and NI_6 values below the training minimum (equal to 0), the model lacks relevant chemical information in that region, leading to unreliable predictions. Furthermore, for NI_7 , the variance interval is relatively wide (i.e., 0.0677), yet the NI_7 value (i.e., -0.5659) for monomer AN (shown in Figure 3) is much lower than the minimum value (i.e., 0) in the training set. This represents an extreme extrapolation distance, placing AN far outside the descriptor space covered by the training data. Even though NI_7 is relatively well distributed across the training set, the extreme distance from the training domain exceeds the model's reasonable extrapolation capacity, resulting in inaccurate k_p predictions.

The monomers for which the k_p values are inaccurately predicted in extrapolation can be summarized as following. The descriptors NI_5 , NI_6 , and NI_7 are not suitable for backward extrapolation. Consequently, monomers BuDE, AN, and NVF, which involve backward extrapolation of these descriptors, yield inaccurate predictions of their k_p values. For monomers HCEA, HCPA, PhCEA, and PhCPA, the presence of multiple forward extrapolation descriptors leads to large errors for the prediction of k_p values. In contrast, NI_{11} is suitable for backward extrapolation, which is associated with accurate k_p prediction.

4. Conclusions

This study establishes a systematic framework for evaluating the extrapolation capability of QSPR models, using our previously developed k_p (T, NI)-QSPR model for free-radical polymerization as a benchmark. Descriptor visualization enables the classification of 12 new monomers into interpolation and extrapolation categories. The analysis demonstrates significantly better statistical performance for interpolation predictions, confirming its high accuracy for monomers structurally similar to the training set. Critically, the extrapolation validation analysis reveals that uneven descriptor distributions are associated with high extrapolation degrees, which may contribute to reduced prediction reliability. We further identify that descriptors NI_1 – NI_{11} are suitable for forward extrapolation, while NI_5 , NI_6 and NI_7 are unsuitable for backward extrapolation. More importantly, beyond the specific model evaluation, the primary contribution of this work is the transferable EV-based workflow for assessing and quantifying extrapolation reliability. This framework offers a practical protocol for defining the applicability domain of QSPR models in polymerization kinetics and provides critical guidance for improving extrapolation robustness. It also lays a foundation for extending such predictive methods to other temperature-dependent kinetic parameters (e.g., initiation and termination rate coefficients) in free-radical polymerization, where extrapolation beyond known monomer spaces is frequently required.

It should be acknowledged that the prediction set in this work is limited to 12 monomers, primarily due to the current scarcity of experimental k_p data for FRP systems. Therefore, extrapolation evaluation is recommended before applying the developed k_p (T, NI)-QSPR model to a broader range of new monomers. The extrapolation robustness of the model may be further enhanced through targeted sample augmentation, appropriate machine-learning algorithm selection, and optimization or expansion of molecular descriptors. Meanwhile, the EV framework provides a useful tool for assessing extrapolation reliability and guiding the application of the model beyond its original chemical domain.

Supplementary Materials

The additional data and information can be downloaded at: <https://media.sciltp.com/articles/others/2609211652221647/SCE-26050235-SI.pdf>. The dataset used for modelling, the definition of statistical parameters, and the calculation for prediction set by the MLR-based QSPR model are provided in the Supporting Information.

Author Contributions

Y.S.: Writing—original draft, Writing—reviewing and editing, Methodology, Data curation, Visualization, Investigation; X.F.: Investigation, Methodology; F.Y.: Writing—review & editing, Software, Supervision, Methodology, Validation, Funding acquisition, Conceptualization; Z.-H.L.: Supervision, Resources; Y.-N.Z.: Writing—review & editing, Software, Supervision, Methodology, Validation, Data curation, Funding acquisition, Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding

This work was financially supported by Advanced Materials-National Science and Technology Major Project (2025ZD0619604) and the National Natural Science Foundation of China (22578332, 22222807).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data will be made available on request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

- Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260.
- Gil, Y.; Greaves, M.; Hendler, J.; et al. Amplify scientific discovery with artificial intelligence. *Science* **2014**, *346*, 171–172.
- Shi, Y.; Yan, F.; Jin, J.; et al. Advances in calculation of kinetic parameters in free-radical polymerization by data-driven methods. *Curr. Opin. Chem. Eng.* **2025**, *48*, 101141.
- Li, J.; Zhao, T.; Yang, Q.; et al. A review of quantitative structure-activity relationship: The development and current status of data sets, molecular descriptors and mathematical models. *Chemom. Intell. Lab. Syst.* **2025**, *256*, 105278.
- Le, T.; Epa, V.C.; Burden, F.R.; et al. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.* **2012**, *112*, 2889–2919.
- Shi, Y.; Wang, J.; Wang, Q.; et al. Supervised machine learning algorithms for predicting rate constants of ozone reaction with micropollutants. *Ind. Eng. Chem. Res.* **2022**, *61*, 8359–8367.
- Pan, I.; Mason, L.R.; Matar, O.K. Data-centric Engineering: Integrating simulation, machine learning and statistics. Challenges and opportunities. *Chem. Eng. Sci.* **2022**, *249*, 117271.
- Chen, L.; Liu, G.; Zhang, Z.; et al. Machine learning and molecular design algorithm assisted discovery of gas separation membranes exceeding the CO₂/CH₄ and CO₂/N₂ upper bounds. *Chem. Eng. Sci.* **2024**, *291*, 119952.
- Zhang, J.; Wang, Q.; Shen, W. Hyper-parameter optimization of multiple machine learning algorithms for molecular property prediction using hyperopt library. *Chin. J. Chem. Eng.* **2022**, *52*, 115–125.
- Chen, Y.; Ma, S.; Lei, Y.; et al. Ionic liquid binary mixtures: Machine learning-assisted modeling, solvent tailoring, process design, and optimization. *AIChE J.* **2024**, *70*, e18392.
- Chen, G.; Song, Z.; Qi, Z.; et al. A scalable and integrated machine learning framework for molecular properties prediction. *AIChE J.* **2023**, *69*, e18185.
- Ma, Y.; Gao, Z.; Shi, P.; et al. Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. *Front. Chem. Sci. Eng.* **2022**, *16*, 523–535.
- Fiosina, J.; Sievers, P.; Drache, M.; et al. Polymer reaction engineering meets explainable machine learning. *Comput. Chem. Eng.* **2023**, *177*, 108356.
- Raven, D.B.; Chikkula, Y.; Patel, K.M.; et al. Machine learning & conventional approaches to process control & optimization: Industrial applications & perspectives. *Comput. Chem. Eng.* **2024**, *189*, 108789.
- Kunz, M.R.; Yonge, A.; Fang, Z.; et al. Data driven reaction mechanism estimation via transient kinetics and machine learning. *Chem. Eng. J.* **2021**, *420*, 129610.
- Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- Kwon, H.; Ali, Z.A.; Wong, B.M. Harnessing Semi-Supervised Machine Learning to Automatically Predict Bioactivities of Per- and Polyfluoroalkyl Substances (PFASs). *Environ. Sci. Technol. Lett.* **2023**, *10*, 1017–1022.
- Buback, M. Free-radical propagation rate coefficients. *Can. J. Chem. Eng.* **2023**, *101*, 5087–5098.
- Wang, T.T.; Zhou, Y.N.; Luo, Z.H.; et al. Beauty of Explicit Dispersity (Đ) Equations in Controlled Polymerizations. *ACS Macro Lett.* **2023**, *12*, 1423–1436.
- Wang, T.T.; Wu, Y.Y.; Luo, Z.H.; et al. “Living” Polymer Dispersity Quantification for Nitroxide-Mediated Polymerization Systems by Mimicking a Monodispersed Polymer Blending Strategy. *Macromolecules* **2020**, *53*, 10813–10822.
- Kockler, K.B.; Haehnel, A.P.; Junkers, T.; et al. Determining free-radical propagation rate coefficients with high-frequency lasers: Current status and future perspectives. *Macromol. Rapid Commun.* **2016**, *37*, 123–134.
- Beuermann, S.; Harrisson, S.; Hutchinson, R.A.; et al. Update and Critical Reanalysis of IUPAC Benchmark Propagation Rate Coefficient Data. *Polym. Chem.* **2022**, *13*, 1891–1900.
- Van Herck, J.; Harrisson, S.; Hutchinson, R.A.; et al. A Machine-Readable Online Database for Rate Coefficients in Radical Polymerization. *Polym. Chem.* **2021**, *12*, 3688–3692.
- Marien, Y.W.; Edeleva, M.; Van Steenberge, P.H.M.; et al. Chapter Three—Exploiting the pulsed laser polymerization-size exclusion chromatography technique to retrieve kinetic parameters in radical polymerization: State-of-the-art and future challenges. In *Advances in Chemical Engineering*; Academic Press: Cambridge, MA, USA, 2020; pp. 59–95.
- D’Hooge, D.R.; Van Steenberge, P.H.M.; Reyniers, M.F.; et al. The strength of multi-scale modeling to unveil the complexity of radical polymerization. *Prog. Polym. Sci.* **2016**, *58*, 59–89.
- Zhou, Y.N.; Li, J.J.; Wang, T.T.; et al. Precision polymer synthesis by controlled radical polymerization: Fusing the progress from polymer chemistry and reaction engineering. *Prog. Polym. Sci.* **2022**, *130*, 101555.
- Li, Y.; Wang, Q.; Lim, K.H.; et al. Characterization technique for kinetic coefficients of free radical polymerization. *CIESC J.* **2023**, *74*, 559–570.

28. Beuermann, S.; Buback, M. Rate coefficients of free-radical polymerization deduced from pulsed laser experiments. *Prog. Polym. Sci.* **2002**, *27*, 191–254.
29. Buback, M.; Gilbert, R.G.; Hutchinson, R.A.; et al. Critically Evaluated Rate Coefficients for Free-Radical Polymerization, 1. Propagation Rate Coefficient for Styrene. *Macromol. Chem. Phys.* **1995**, *196*, 3267–3280.
30. Beuermann, S.; Buback, M.; Davis, T.P.; et al. Critically Evaluated Rate Coefficients for Free-Radical Polymerization, 2. Propagation Rate Coefficients for Methyl Methacrylate. *Macromol. Chem. Phys.* **1997**, *198*, 1545–1560.
31. Olaj, O.F.; Bitai, I.; Hinkelmann, F. The Laser-Flash-Initiated Polymerization as a Tool of Evaluating (Individual) Kinetic Constants of Free-Radical Polymerization, 2. The Direct Determination of the Rate of Constant of Chain Propagation. *Makromol. Chem.* **1987**, *188*, 1689–1702.
32. Vir, A.B.; Marien, Y.W.; Van Steenberge, P.H.M.; et al. From n-butyl acrylate Arrhenius parameters for backbiting and tertiary propagation to β -scission via stepwise pulsed laser polymerization. *Polym. Chem.* **2019**, *10*, 4116–4125.
33. Moscatelli, D.; Dossi, M.; Cavallotti, C.; et al. Ab Initio Calculation of the Propagation Kinetics in Free Radical Polymerization: Chain Length and Penultimate Effects. *Macromol. Symp.* **2007**, *259*, 337–347.
34. Edeleva, M.; Van Steenberge, P.H.M.; Sabbe, M.K.; et al. Connecting Gas-Phase Computational Chemistry to Condensed Phase Kinetic Modeling: The State-of-the-Art. *Polymers* **2021**, *13*, 3027.
35. Trigilio, A.D.; Marien, Y.W.; Van Steenberge, P.H.M.; et al. Gillespie-Driven kinetic Monte Carlo Algorithms to Model Events for Bulk or Solution (Bio)Chemical Systems Containing Elemental and Distributed Species. *Ind. Eng. Chem. Res.* **2020**, *59*, 18357–18386.
36. Arraez, F.J.; Xu, X.; Van Steenberge, P.H.M.; et al. Macropropagation Rate Coefficients and Branching Levels in Cationic Ring-Opening Polymerization of 2-Ethyl-2-oxazoline through Prediction of Size Exclusion Chromatography Data. *Macromolecules* **2019**, *52*, 4067–4078.
37. Wu, Y.Y.; Figueira, F.L.; Van Steenberge, P.H.M.; et al. Bridging principal component analysis and method of moments based parameter estimation for grafting of polybutadiene with styrene. *Chem. Eng. J.* **2021**, *425*, 130463.
38. Yu, X.; Wang, X. A support vector machine model for the prediction of monomer reactivity ratios. *Polym. Sci. Ser. B* **2011**, *53*, 528.
39. Xu, J.; Wang, L.; Zhang, H.; et al. Quantitative structure-property relationships studies on free-radical polymerization chain-transfer constants for styrene. *J. Appl. Polym. Sci.* **2012**, *123*, 356–364.
40. Van De Reydt, E.; Marom, N.; Saunderson, J.; et al. A Predictive machine-learning model for propagation rate coefficients in radical polymerization. *Polym. Chem.* **2023**, *14*, 1622–1629.
41. Shi, Y.; Han, J.P.; Yan, F.; et al. In Silico Prediction of the Temperature-Dependent Decomposition Rate Coefficients of Symmetrical Azo Compounds. *Ind. Eng. Chem. Res.* **2023**, *62*, 17461–17472.
42. Ishola, N.B.; McKenna, T.F.L. Machine learning techniques for the prediction of polymerization kinetics and polymer properties. *Can. J. Chem. Eng.* **2024**, *102*, 2228–2243.
43. Datta, S.; Dev, V.A.; Eden, M.R. Hybrid genetic algorithm-decision tree approach for rate constant prediction using structures of reactants and solvent for Diels-Alder reaction. *Comput. Chem. Eng.* **2017**, *106*, 690–698.
44. Datta, S.; Dev, V.A.; Eden, M.R. Developing non-linear rate constant QSPR using decision trees and multi-gene genetic programming. *Comput. Chem. Eng.* **2019**, *127*, 150–157.
45. Zhou, Y.; Ren, Y.; Cui, M.; et al. Sonoelectrochemical system mechanisms, design, and machine learning for predicting degradation kinetic constants of pharmaceutical pollutants. *Chem. Eng. J.* **2023**, *478*, 147266.
46. Chakkingal, A.; Janssens, P.; Poissonnier, J.; et al. Multi-output machine learning models for kinetic data evaluation: A Fischer-Tropsch synthesis case study. *Chem. Eng. J.* **2022**, *446*, 137186.
47. Ouyang, B.; Zhang, D.; Chen, Z.; et al. Enhancing quantitative structure-property relationship models by integrating complete molecular structure with deep learning. *Eng. Chem. Eng.* **2026**, *20*, 16.
48. Nguyen, T.; Bavarian, M. Machine learning approach to polymer reaction engineering: Determining monomers reactivity ratios. *Polymer* **2023**, *275*, 125866.
49. Farajzadehahary, K.; Telleria-Allika, X.; Asua, J.M.; et al. An artificial neural network to predict reactivity ratios in radical copolymerization. *Polym. Chem.* **2023**, *14*, 2779–2787.
50. Shi, Y.; Yu, M.; Liu, J.; et al. Quantitative structure–property relationship model for predicting the propagation rate coefficient in free-radical polymerization. *Macromolecules* **2022**, *55*, 9397–9410.
51. Muckley, E.S.; Saal, J.E.; Meredig, B.; et al. Interpretable models for extrapolation in scientific machine learning. *Digit. Discov.* **2023**, *2*, 1425–1435.
52. Wang, Y.; King, R.D. Extrapolation is not the same as interpolation. *Mach. Learn.* **2024**, *113*, 8205–8232.
53. Ambrosio Garcia, F.; Naets, F. Beyond Limits: Enhancing the Extrapolation Performance of Regression Models by Leaving the Boundary Out. *Mach. Learn.* **2025**, *114*, 285.
54. McCartney, M.; Haeringer, M.; Polifke, W. Comparison of Machine Learning Algorithms in the Interpolation and Extrapolation of Flame Describing Functions. *J. Eng. Gas Turbines Power* **2020**, *142*, 061009.

55. Feng, X.; Xiong, J.; Wang, Q.; et al. Boiling point enhanced vapor pressure prediction by Antoine-based quantitative structure–property relationship. *Chem. Eng. Sci.* **2025**, *306*, 121269.
56. Yu, M.; Zhou, Y.N.; Wang, Q.; et al. Extrapolation validation (EV): A universal validation method for mitigating machine learning extrapolation risk. *Digit. Discov.* **2024**, *3*, 1058–1067.
57. Hastie, T.; Tibshirani, R.; Friedman, J. Linear Methods for Regression. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 43–99.
58. Alexopoulos, E.C. Introduction to multivariate regression analysis. *Hippokratia* **2010**, *14*, 23–28.
59. Roy, K.; Kar, S.; Das, R.N. Chapter 7—Validation of QSAR Models. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; Academic Press: Boston, MA, USA, 2015; pp. 231–289.
60. Gramatica, P.; Sangion, A. A historical excursus on the statistical validation parameters for QSAR models: A clarification concerning metrics and terminology. *J. Chem. Inf. Model.* **2016**, *56*, 1127–1131.