Chinese Journal of
POLYMER SCIENCE

RESEARCH ARTICLE

https://doi.org/10.1007/s10118-025-3402-y
Chinese J. Polym. Sci. **2025**, *43*, 1749–1760

OpenPoly: A Polymer Database Empowering Benchmarking and Multi-property Predictions

Ji-Feng Wang, Yu-Bo Sun, Qiu-Tong Chen, Fei-Fan Ji, Yuan-Yuan Song, Meng-Yuan Ruan, and Ying Wang*

State Key Laboratory of Molecular Engineering of Polymers, Department of Macromolecular Science, AI for Polymer Science Research Center, Fudan University, Shanghai 200438, China

Abstract Advancing the integration of artificial intelligence and polymer science requires high-quality, open-source, and large-scale datasets. However, existing polymer databases often suffer from data sparsity, lack of polymer-property labels, and limited accessibility, hindering systematic modeling across property prediction tasks. Here, we present OpenPoly, a curated experimental polymer database derived from extensive literature mining and manual validation, comprising 3985 unique polymer-property data points spanning 26 key properties. We further develop a multi-task benchmarking framework that evaluates property prediction using four encoding methods and eight representative models. Our results highlight that the optimized degree-of-polymerization encoding coupled with Morgan fingerprints achieves an optimal trade-off between computational cost and accuracy. In data-scarce condition, XGBoost outperforms deep learning models on key properties such as dielectric constant, glass transition temperature, melting point, and mechanical strength, achieving R^2 scores of 0.65–0.87. To further showcase the practical utility of the database, we propose potential polymers for two energy-relevant applications: high temperature polymer dielectrics and fuel cell membranes. By offering a consistent and accessible benchmark and database, OpenPoly paves the way for more accurate polymer-property modeling and fosters data-driven advances in polymer genome engineering.

Keywords Polymer database; Polymer structure encoding; Property prediction; Functional reverse design; Benchmark models

Citation: Wang, J. F.; Sun, Y. B.; Chen, Q. T.; Ji, F. F.; Song, Y. Y.; Ruan, M. Y.; Wang, Y. OpenPoly: a polymer database empowering benchmarking and multi-property predictions. *Chinese J. Polym. Sci.* **2025**, *43*, 1749–1760.

INTRODUCTION

Artificial intelligence (AI) is emerging as a powerful engine driving innovation in polymer design and property prediction. The successful integration of AI in polymer science lies in the availability of high-quality, structurally standardized, and sufficiently large open-access polymer datasets.[1–3] However, current polymer data resources face three key challenges. First, experimental data remain sparse and unevenly distributed, which limits the performance of supervised learning models.[4,5] Second, property annotations often lack consistent polymer-property alignment and are seldom accompanied by standardized molecular representations, such as PSMILES.[6] Third, many existing databases suffer from limited accessibility and non-standard formats, hindering reproducibility and multi-model benchmarking.[7–11]

As summarized in Fig. 1, the state-of-the-art polymer databases exhibit significant limitations. For instance, PolyInfo[7] provides rich property data but has restricted accessibility. PI1M database[8] generates PSMILES from unstructured text but lacks property annotations. Khazana database[9] focuses on dielectric properties derived from DFT but stores data in CIF format, incompatible with mainstream ML pipelines. Open Macromolecular Genome (OMG)[10] offers millions of synthetically accessible polymer backbones, yet with no experimental properties. PolyID[11] provides graph-based learning data but omits reconstruction-friendly formats like PSMILES. Although large language models (LLMs) have recently been applied to automatically extract polymer data from literature,[12–14] they still struggle with chemical name disambiguation and require extensive manual correction to ensure quality.[15]

In addition to data availability, a unique challenges to develop models for polymers compared to small molecules, lies in the trivial intrinsic properties of polymers such as degree of polymerization (DP), copolymerization strategy, chain architecture, and processing conditions.[16] Experimental properties are highly sensitive to measurement protocols, leading to large inter-source variability.[17] While recent advances in transformer-based architectures and LLMs have shown promise in encoding polymer representations *via* chemical language inputs,[18–21] these models still face challenges in generalization, computational cost, and format sensitivity.[22] Even the state-of-the-art platform Polymer Genome[23] lack transparency in model training and preprocessing procedures, limiting their scientific reproducibility.

* Corresponding author, E-mail: wying@fudan.edu.cn

Special Topic: AI for Polymers
Received May 30, 2025; Accepted June 25, 2025; Published online September 10, 2025

© Chinese Chemical Society
Institute of Chemistry, Chinese Academy of Sciences

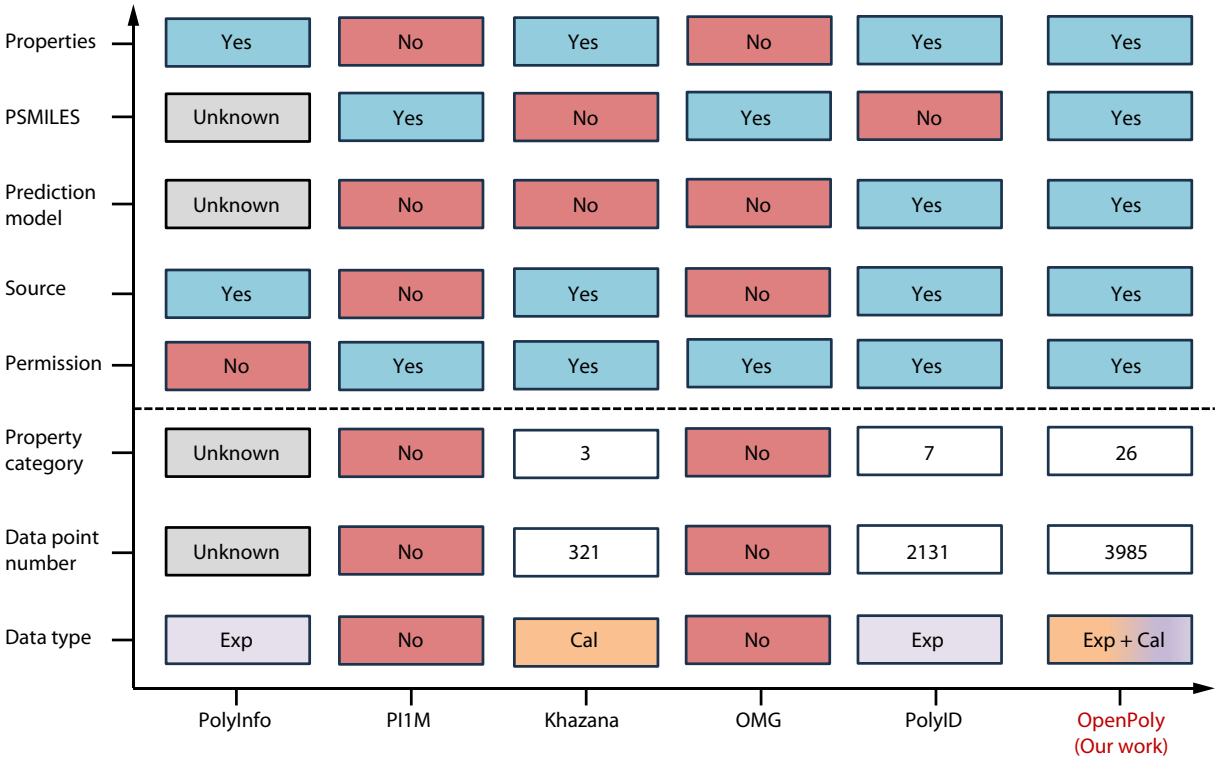

Fig. 1 Comparative overview of features for representative polymer databases[7–11]. "Properties" indicates the availability of experimental or computational property data in the databases; "PSMILES" denotes the inclusion of polymer structures encoded as PSMILES; "Prediction model" refers to the integration of machine learning models for property prediction; "Source" assesses the traceability of data entries *via* cited references; "Permission" reflects the openness of database access; "Property category" indicates the number of distinct property types covered for polymers; "Data type" distinguishes between experimentally measured ("Exp") and theoretically calculated ("Cal") data for properties.

To address these limitations, we present an open database OpenPoly with standardized structures and diverse property labels based on curated experimental and calculation results. Derived from extensive literature and public data, OpenPoly contains repeat-unit structures (in PSMILES format) aligned with experimentally measured properties, verified through human curation. On top of this dataset, we establish a benchmarking framework that systematically evaluates eight representative models across multiple property prediction tasks. Our results demonstrate that, under real-world conditions of data sparsity and heterogeneity, machine learning methods such as XGBoost outperform deep neural networks in both robustness and adaptability, especially in multi-task and low-data regimes.

The key contribution of this work lies in the construction of a unified, property-rich, and structure-standardized open polymer database; meanwhile, we establish a multi-property model benchmarking framework for standardized algorithmic comparison and offer a transparent and reproducible platform to support polymer genome research and high-throughput polymer screening tasks.

DATA COLLECTION AND CURATION

To construct a high-quality polymer property database, we adopt a multi-source data collection strategy integrating both automated and manual approaches. We retrieve the source data from existing polymer datasets,[9,11,19] handbooks,[24,25] technical sheets provided by commercial suppliers and the large language model (LLM)-based extraction dataset developed by Gupta *et al.*[15] To ensure structural accuracy and semantic consistency, each polymer entry underwent manual verification. PSMILES representations are corrected, standardized, and validated against their corresponding structural diagrams, as outlined in Fig. 2(a). To maintain data uniformity and facilitate machine-readable encoding, only homopolymers with well-defined repeat units are included in this study. Copolymers and crosslinked systems, due to their structural ambiguity and lack of standardized representations are excluded. During data integration, we encounter three challenges: (1) the same polymer property can be reported across multiple sources; (2) lightly modified polymers (*e.g.*, doped, end-functionalized, or partially copolymerized variants) are sometimes grouped under the same polymer name, potentially confounding property attribution; (3) differences in experimental protocols, measurement precision, and quality control introduce data variance with reduced reliability. To address these issues, we develop a standardized curation pipeline (Fig. 2b) as below.

(1) Data normalization and structural unification. All entries are formatted uniformly, including units and nomenclature. Modified polymers with minor alterations are classified based on their backbone structure to maintain dataset coherence.

(2) Cross-source integration and outlier filtering. For property values reported by multiple sources, we applied an in-

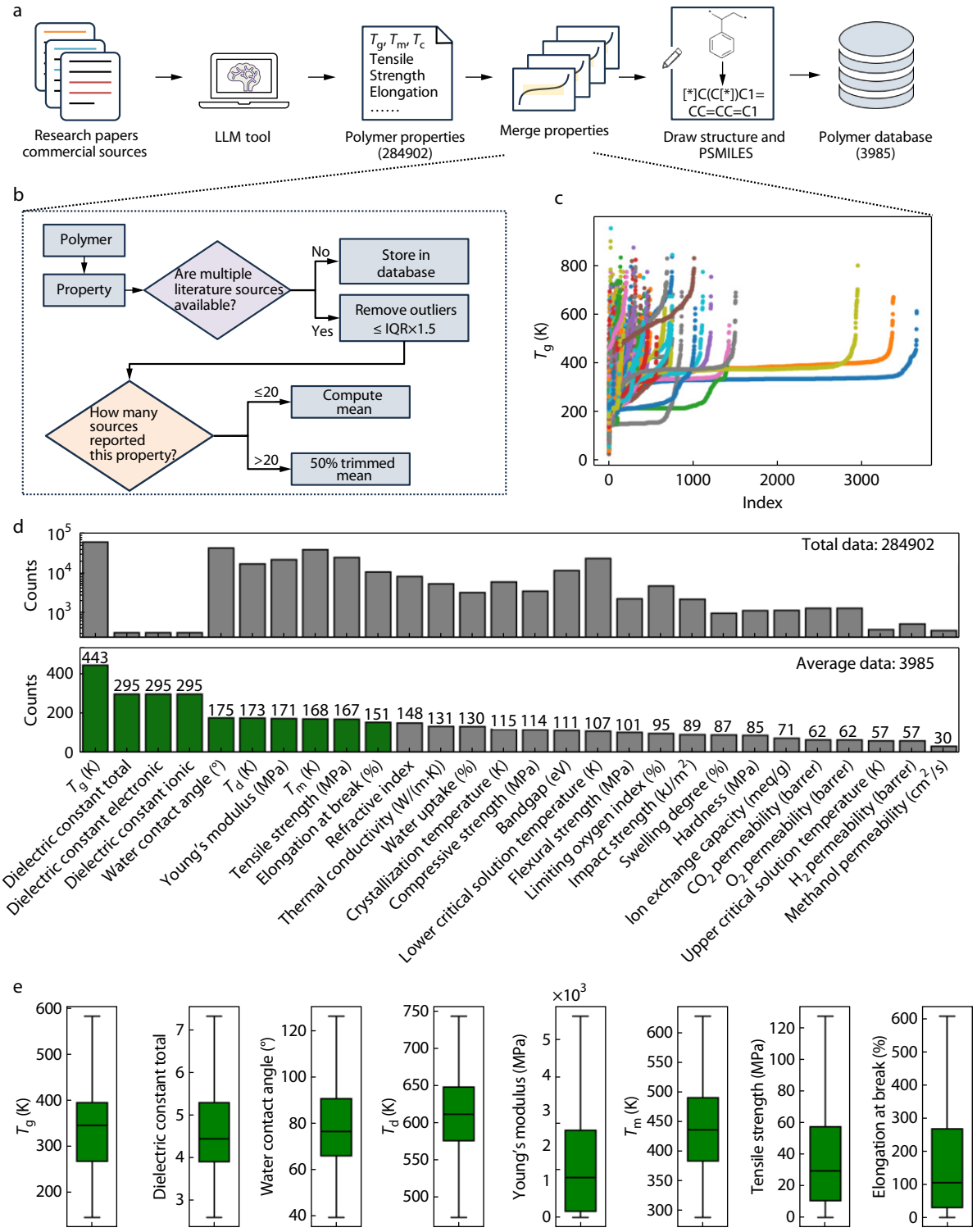

Fig. 2 Overview of the OpenPoly database construction and curation process. (a) Workflow for extracting and standardizing polymer property data from literature, databases, and commercial sources, including PSMILES verification and structural normalization. (b) Schematic of data integration procedures for cases with multiple sources reporting the same polymer–property pair, including outlier removal using the interquartile range (IQR) method and estimation of representative values. (c) Distribution of reported T_g for different polymers across different sources. (d) Distribution of polymer property records before and after data processing. The raw dataset (top) contains approximately 284,902 entries covering 745 polymers across 26 properties. The curated dataset (bottom) retains 3,985 unique polymer–property entries with standardized structures and validated values. (e) Distribution of property values for properties with more than 150 polymer–property data points retained after curation.

https://doi.org/10.1007/s10118-025-3402-y

terquartile range (IQR)-based filter to exclude extreme outliers (≥1.5×IQR).[26] The IQR method is a widely accepted nonparametric approach in robust statistics, particularly suitable for experimental datasets with unknown or non-Gaussian distributions. This approach is especially important in polymer science, where discrepancies due to differences in sample preparation, thermal history, or instrument calibration are common. To preserve data diversity, single-source values are retained directly to support small-sample learning tasks.

(3) Representative value estimation. For each polymer–property pair, we employ a data-dependent strategy. If polymer–property records ≥ 20, a 50%-trimmed mean is calculated. This estimator removes the top and bottom 25% of values before averaging, offering resistance to outliers while preserving the distribution's core. This estimator offers superior robustness for polymer datasets of moderate variability. Although this method is not extensively used in polymer informatics, analogous applications in other domains have demonstrated significant gains in statistical stability.[27,28] If polymer-property records ≤ 20, a simple arithmetic mean is used to maximize data utilization under limited information. Although more sensitive to extreme values, the mean allows for unbiased estimation when trimming is statistically inefficient due to small sample sizes. As exemplified by T_g in Fig. 2(c), polymers exhibit an S-shaped distribution across different sources, highlighting the need for aggregation to ensure value accuracy.

Fig. 2(d) illustrates the changes in data distribution before and after processing. The initial dataset contained about 2.8×10^5 raw entries covering approximately 700 polymers across 26 properties. After standardization and deduplication, we retain 3985 unique entries with clearly defined structures and high-confidence property annotations. To further assess the representativeness of the curated dataset, we display the statistical distributions of all properties with at least 150 polymer–property data points in Fig. 2(e). The box plots reveal broad parameter coverage across multiple property dimensions, including thermal, mechanical and dielectric characteristics. The diverse value ranges and interquartile spreads observed suggest that the OpenPoly database contains a wide property space of the accumulated polymers. This diversity is essential for enabling robust model training and benchmarking tasks with varying complexity and scale.

BENCHMARKING POLYMER PROPERTY PREDICTION MODELS

Despite the availability of a comprehensive polymer property dataset, accurate property prediction remains a critical challenge. On one hand, polymer datasets are inherently sparse, and variations in experimental protocols result in non-uniform data quality and limited sample sizes. On the other hand, the multiscale structural complexity of polymers—driven by factors such as DP, rendering difficulty to model the highly nonlinear structure–property relationships. Thus, it is essential to develop structurally informative and modeling-friendly representations of polymers. An ideal polymer encoding scheme should support efficient learning with low model complexity while remaining interpretable and scalable. However, we lack consensus and systematic comparisons of the diverse polymer representation, including molecular fingerprints, 2D and 3D graph-based formats for diverse property prediction tasks.

Here we design a unified benchmarking framework to evaluate the performance of multiple encoding schemes and model architectures in polymer property prediction. The framework incorporates eight representative models, spanning both traditional and advanced machine learning paradigms: XGBoost, multilayer perceptron (MLP), graph neural networks (GCN,[29] GAT[30]), pretrained molecular representation models (PolyBERT,[31] UniMol2[32]), invariant 3D geometric networks (Spherenet[33]), and the small-data-optimized TabPFN.[34] These models are evaluated using four distinct types of polymer representations: (1) Morgan fingerprints,[35] computed using RDKit with a radius of 2 and a dimensionality of 2048, were used as input features for XGBoost, MLP, and TabPFN models; (2) 2D molecular graphs, including node features such as atomic number, chirality, hybridization, aromaticity, and ring membership, are used as input for GCN and GAT; (3) 3D molecular graphs, which add spatial coordinates to the 2D molecular graphs are used for Spherenet and UniMol2; (4) Polymer-specific fingerprints are generated by PolyBERT (dimension = 600) and used as the regression head for downstream prediction.

As illustrated in Fig. 3(a), homopolymers are encoded from repeat-unit PSMILES with defined attachment points (*e.g.*, polyethylene as [*]CC[*]). To construct complete molecular structures, hydrogen atoms are added to cap the PSMILES, yielding canonical SMILES such as [H]CC[H]. These structures serve as the basis for generating the above representations. All models are trained using a consistent protocol and evaluated with standardized performance metrics across multiple property prediction tasks. This framework allows us to systematically compare encoding–model combinations, assess performance under data-scarce conditions, and identify robust, generalizable models for polymer property prediction.

To investigate how the number of repeat units affects prediction accuracy, we benchmarked eight representative models across 26 polymer properties using three representations of polymer chains with increasing repeat units—namely, N = 1 (N1), N = 2 (N2), and N = 4 (N4). Each representation was derived by repeating the monomer unit accordingly—for instance, polyethylene is represented as "CC" for N1, "CCCC" for N2, and "CCCCCCCC" for N4. These variations in repeat unit length affect the generated molecular structure and, consequently, may influence model learning behavior and performance stability. To quantify the impact of polymerization degree on model accuracy, we apply Bland–Altman analysis, which plots the difference in model performance between two encodings against their mean performance. This method helps reveal systematic biases, variability, and agreement intervals between encoding schemes.[36] To allow cross-property comparison, we first normalize the error metrics for each property using the following formulations:

$$\mathrm{NRMSE}_i = \frac{\sqrt{\mathrm{MSE}}}{y_{i,\max} - y_{i,\min}} \tag{1}$$

$$\mathrm{NMAE}_i = \frac{\mathrm{MAE}}{y_{i,\max} - y_{i,\min}} \tag{2}$$

For each property, we compute the average model performance (R^2, NRMSE, and NMAE) across all eight models for

each encoding (N1, N2, N4), then compared three encoding pairs: N1–N2, N1–N4, and N2–N4. As shown in Fig. 3(b), the N1–N2 and N1–N4 comparisons reveal substantial variability in performance, with R^2 differences exceeding 0.4 in some properties, indicating instability and inconsistency when using short-chain encodings (N1). In contrast, differences between N2 and N4 are minimal across all metrics, suggesting that performance stabilizes significantly when N ≥ 2. Interestingly, while N2 showed slightly lower R^2 and marginally higher NRMSE and NMAE compared to N4, it greatly reduces the computational costs due to shorter input sequences and lower memory usage during training. Overall, encoding polymers with N=2 offers an optimal balance between structural informativeness, prediction accuracy, and computational efficiency, and is recommended for general-purpose polymer property prediction tasks.

To minimize performance instability caused by small sample sizes, we limit our analysis to polymer properties with polymer–property data points > 150. This threshold was established based on two criteria: (1) data sufficiency, ensuring that each property has an adequate number of samples to support robust statistical analysis and reliable model training; and (2) practical relevance, prioritizing properties that are frequently considered in polymer design, evaluation, and industrial applications. As shown in Fig. 2(d), the number of data points per property follows a long-tailed distribution. Setting the cutoff at 150 ensures the inclusion of the top one-third most data-rich and application-relevant properties. Moreover, this threshold aligns with common practices in materials informatics, where approximately 10^2 samples are often regarded as the minimum size for effectively training widely used machine learning models.[37,38] This filtering yields a subset of 8 properties: glass transition temperature (T_g), melting point (T_m), decomposition temperature (T_d), tensile strength, Young's modulus, dielectric constant, elongation at break, and water contact angle. Based on these subsets, we systematically evaluate the predictive performance of eight representative models using three standard metrics: R^2, NRMSE, and NMAE. For each property prediction task, we rank the models based on each metric and compute an average rank by averaging the ranks across the three metrics for each model–property pair. The average rankings are visualized as a heatmap (Fig. 3c) and summarized in Fig. 3(d) to illustrate cross-task performance trends. Across majority of the prediction tasks, XGBoost and the small-dataset optimized TabPFN consistently outperform other models, frequently achieving the top two ranks. Notably, XGBoost shows the highest overall stability and generalization. The outperformance can be attributed to its ability to handle high-dimensional, sparse inputs through tree-based non-linear partitioning, which is particularly suited for small-sample, high-noise scenarios.[39,40] Furthermore, XGBoost exhibits inherent robustness to missing values, label noise, and data heterogeneity[41,42] for common challenges in experimental polymer datasets. TabPFN, while competitive, perform slightly below XGBoost. This may stem from its performance sensitivity to input dimensionality, as its optimal feature size is under 500, whereas our fingerprint vectors contain 2048 dimensions. Pretrained molecular representation models such as UniMol2 and PolyBERT demonstrate moderate performance, indicating partial transferability of semantic embeddings learned from molecules to polymer property prediction. However, recent studies suggest that pretraining on small-molecule data may limit model performance when applied directly to polymers.[43] By contrast, graph neural networks (GCN, GAT) and the 3D geometry-aware Spherenet exhibit poor performance, likely due to their inability to effectively capture structure–property dependencies under these data limited scenarios. MLP performs the worst overall, indicating limited capacity to extract meaningful features from high-dimensional polymer encodings.

To further assess model robustness, we plot the distribution of R^2 scores for all models across the eight properties (Fig. 3e). Even for the best-performing XGBoost, the best R^2 is approximately 0.8 with average value approaching > 0.5, suggesting that reliable predictions are achieved only for a few tasks. For most other models, R^2 values are under expected, indicating a lack of meaningful structure-property mapping. This poor generalization can be largely attributed to data scarcity. The small sample sizes and presence of measurement errors or outliers exacerbate overfitting and numerical instability, particularly in deep models. In contrast, XGBoost and TabPFN exhibit narrower distributions in NRMSE and NMAE, confirming their superior robustness and error control. Collectively, these results reveal a clear trend: there is a strong interplay between model complexity and data availability. Under severely limited data conditions, simpler and more robust models such as XGBoost demonstrate clear practical advantages over complex neural architectures.[44,45]

Following the comprehensive ranking of all model-property combinations, we identify a subset of high-confidence prediction pairs that demonstrated both superior accuracy and robustness. Specifically, combinations with an Average Rank of 1 and a test-set R^2 ≥ 0.6 are considered reliable. According to this criterion, four polymer properties were selected: dielectric constant, T_g, tensile strength, and Young's modulus. For each of these tasks, XGBoost emerge as the top-performing model. As illustrated in Figs. 4(a) and 4(d), these properties exhibit clear structure–property trends. For example, in Fig. 4(a), XGBoost achieved an R^2 of 0.87, with MAE = 0.21 and MSE = 0.07 on the test set in predicting the dielectric constant. The predicted values closely align with the experimental measurements along the diagonal, indicating strong fitting and generalization capability. Similarly, T_g (Fig. 4b) predictions yield high linear correlations, with R^2 values of 0.82, demonstrating that even with limited data, the Morgan fingerprint-based XGBoost model effectively captures the structural determinants of thermal properties. In predicting the mechanical property, XGBoost also shows promising performance with R^2 = 0.78 for tensile strength (Fig. 4c) and R^2 = 0.65 for Young's modulus (Fig. 4d). Despite the high variability and wide dynamic range of these properties, the model successfully suppresses the influence of outliers and maintain close alignment between predictions and ground truth. Notably, residual errors for these tasks are significantly lower than those observed with deep learning models, highlighting XGBoost's superior robustness.

Nonetheless, the current modeling framework also

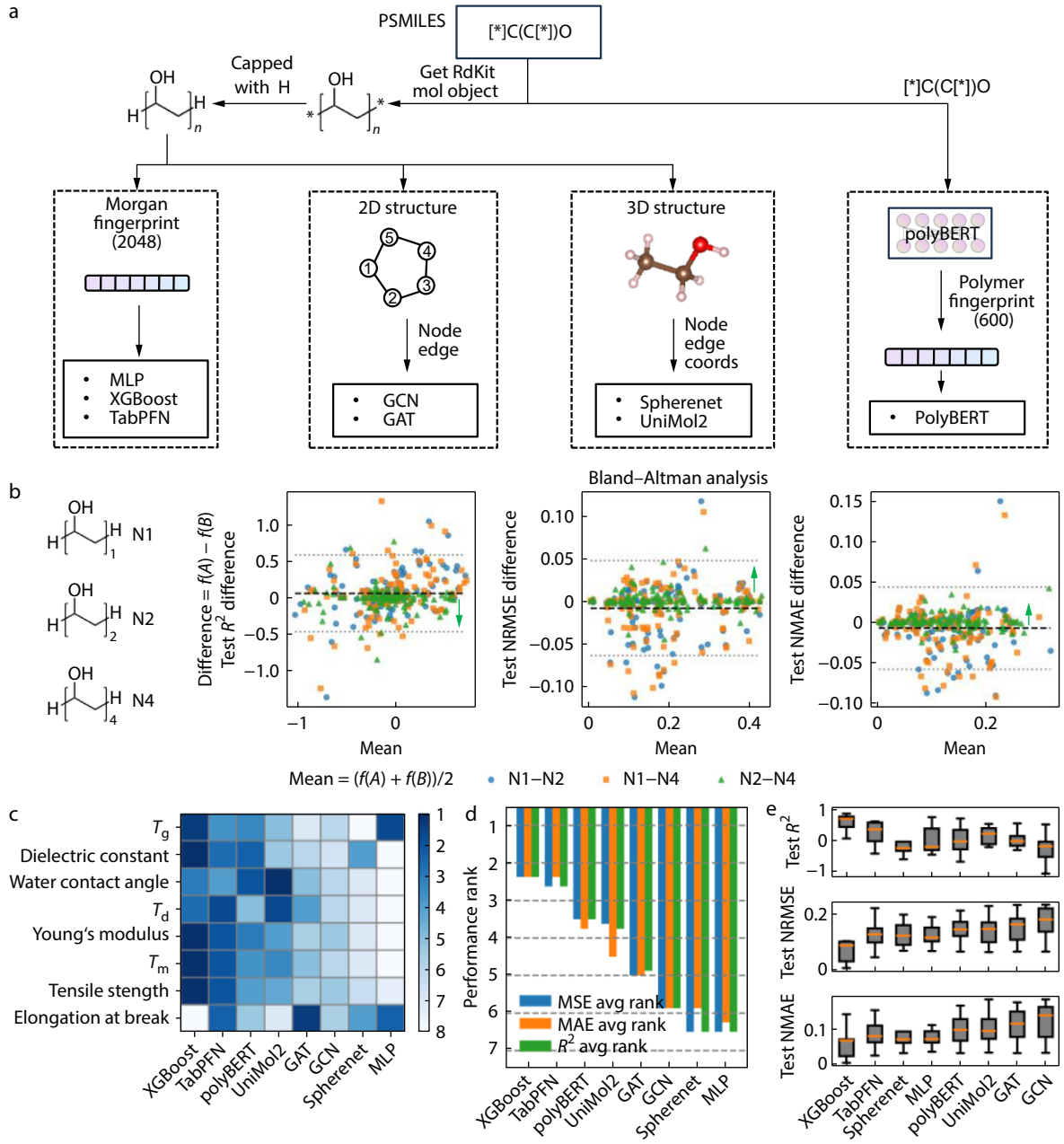

Fig. 3 Benchmarking framework and performance evaluation for property prediction of polymers. (a) Overview of the benchmarking pipeline, including four polymer encoding strategies and eight representative models used for property prediction. (b) Bland–Altman analysis of the effect of polymerization degree (N = 1, 2, 4) on model performance. Each point represents the average performance difference across models and metrics for a specific property. (c) Heatmap of average rankings for each model–property pair across test R^2, NRMSE, and NMAE. Darker colors indicate better rankings. (d) Bar plot showing the average ranking of each model across the three metrics. (e) Box plots of model performance distributions across eight properties, evaluated using test R^2, NRMSE, and NMAE. All model performances were evaluated on the test set.

presents limitations due to its simplified polymer representation and data preprocessing strategy. In particular, it omits critical contextual factors—most notably the DP, which strongly influences bulk properties such as tensile strength and melting point. For example, low- and high-molecular-weight polyethylene exhibit distinct physical behaviors, yet this distinction cannot be captured without DP information. As a result, the model performs reliably only when molecular weight distributions are narrow or consistently reported. In addition, the model does not consider variations in experimental conditions (*e.g.*, temperature, strain rate), which can significantly affect property values—especially mechanical properties that are highly protocol-dependent. By contrast, properties like T_g and dielectric constant are typically measured under more standardized conditions, yielding more consistent data and better predictive performance, as reflected in Fig. 4. These limitations largely stem from the scarcity and heterogeneity of DP and test-condition metadata in the

https://doi.org/10.1007/s10118-025-3402-y

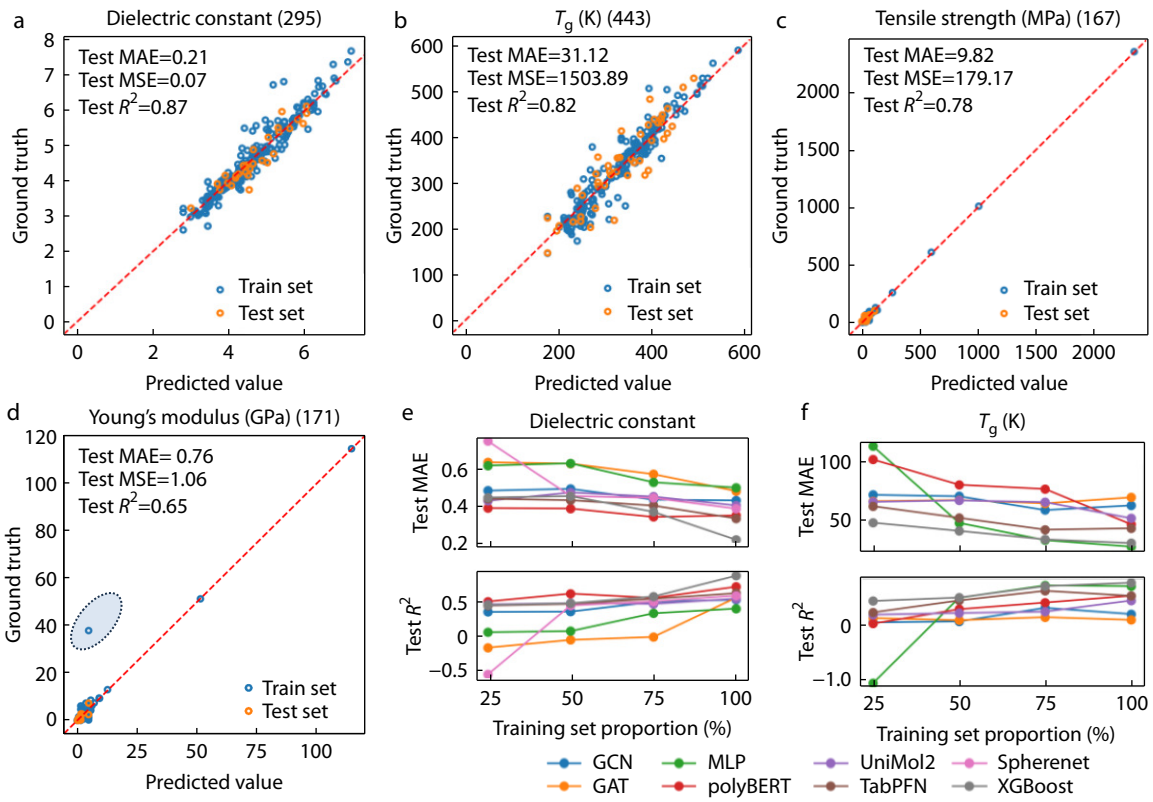

Fig. 4 Predictive performance and data efficiency of machine learning models for key properties. (a–d) Scatter plots showing model predictions versus experimental values for six representative properties: dielectric constant (a), T_g (b), tensile strength (c), and Young's modulus (d). Blue boxes indicate representative outliers in the dataset. (e, f), Impact of training set size on model performance for dielectric constant (e) and T_g (f). Eight models are evaluated using 25%, 50%, 75%, and 100% of the training data, with consistent validation and test sets across all settings. Spherenet fails to converge on the T_g prediction due to numerical instability.

literature, as well as the absence of standardized encoding schemes. Addressing this challenge may require (1) automated extraction of such metadata using LLMs, and (2) the development of text-aware models capable of integrating unstructured experimental context into the prediction pipeline.

To further investigate the data efficiency and scalability of different models, we select T_g and dielectric constant as two representative property prediction tasks. These properties are chosen due to their relatively large sample sizes, which allow the construction of multiple training subsets without introducing severe statistical bias. Their performance under varying data sizes also assesses the robust of the models for predicting polymer properties with limited data availability. For each task, we construct training subsets comprising 25%, 50%, 75%, and 100% of the full training data, while keeping the validation and test sets fixed to ensure comparability. As shown in Figs. 4(e)–4(f), all models exhibit consistent trends with steadily improved R^2 as the training size increases. This reflects a strong data-dependence in polymer property prediction, consistent with trends observed in broader materials informatics.[37]

We observe that T_g and dielectric constant predictions based on XGBoost achieve a relatively high R^2 (about 0.5) even with just 25% of the training data, and continued to improve as the dataset expanded, ultimately outperforming all other models when trained on the full dataset. In contrast, although other models also benefit from larger data, their gains are modest, with flatter R^2 increasements. This discrepancy likely stems from the limited capacity of these models under sparse sampling and high-dimensional feature representations—*i.e.*, they rely more heavily on full coverage of the feature space and are more sensitive to outliers under low-data conditions.[46]

Together, these findings reveal that model performance is governed not only by the model architecture, but also by a complex interplay among training set size, property-specific nonlinearity, and the informativeness of the structural encoding. In data-scarce scenarios, XGBoost demonstrates superior adaptability due to its high data efficiency and robustness to noise. In contrast, deep learning models require substantially more data and enriched structural-property information to realize their full predictive potential. It is important to note, however, that the current dataset exhibits incomplete and non-uniform property coverage across polymers. Specifically, each polymer is labeled for only a subset of target properties, and the subset varies between entries. This results in a structurally imbalanced and sparsely populated property-structure matrix, which may bias models toward overrepresented property combinations and structural motifs. To further address these limitations, we propose two complementary strategies. First, as the OpenPoly database continues to grow through ongoing curation and data integration, which will help mitigate bias and sparsity in a natural, data-driven manner. Second, we plan to implement a pretraining–finetuning

framework in future work. This approach aims to first learn generalizable structural representations through self-supervised learning, and subsequently adapt them to specific property prediction tasks *via* fine-tuning. By decoupling representation learning from the requirement of fully labeled data, this strategy is expected to improve model robustness and generalization, particularly in settings with sparse annotations or limited data availability.

Building on the predictive capabilities established in the benchmarking phase, we further investigate the practical utility of the OpenPoly database in guiding polymer screening and structural design for actual applications. Specifically, we target two classes of energy-related materials: high temperature polymer dielectrics (HTPDs) and proton exchange membrane fuel cells (PEMs). To enable comprehensive evaluation, we employ the top-performing XGBoost model to impute missing experimental values within the database, thereby facilitating high-throughput screening across the entire polymer space. The property thresholds employed for each application are illustrated in Fig. 5.

HTPDs have emerged as critical materials for applications in harsh environments, particularly in high-temperature capacitors, flexible electronics, and integrated sensors, owing to their excellent thermal stability, high energy density, and scalable processability.[47,48] Guided by these insights, we select dielectric constant, T_g, and tensile strength as key screening metrics, capturing polarity, thermal rigidity, and mechanical resilience, respectively. As shown in Fig. 5(a), three promising candidate polymers are identified, all exhibiting high T_g values (> 500 K), moderate-to-high dielectric constants (about 6.2–7.2), and tensile strength exceeding 600 MPa. These favorable properties stem from well-defined structural motifs: rigid aromatic units such as phenylene and thiophene rings that enhance backbone stiffness and suppress segmental mo-

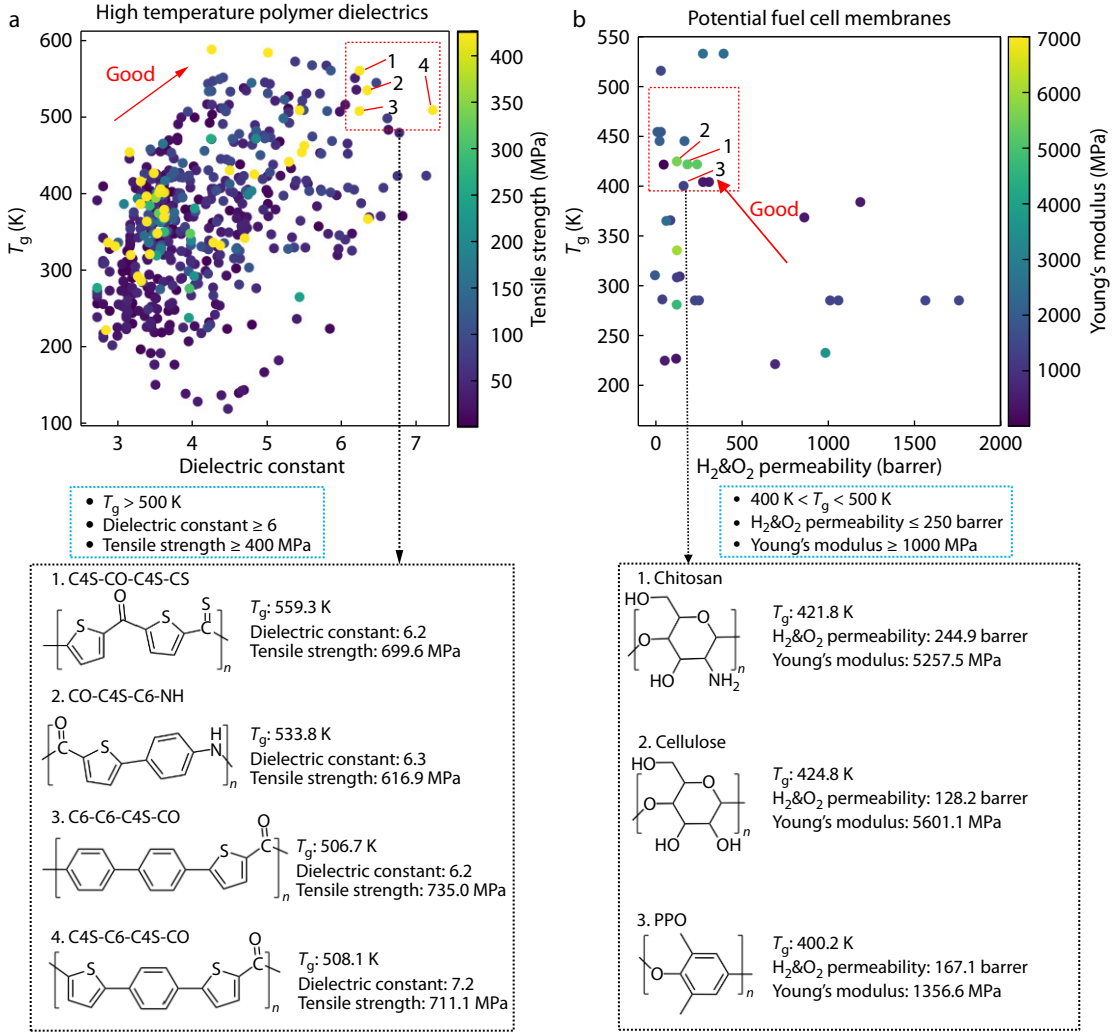

Fig. 5 Screening of high-performance polymers for high temperature polymer dielectrics and fuel cell membranes. (a) Candidate polymers screened for application in high temperature polymer dielectrics, visualized in a property space defined by dielectric constant and T_g. (b) Candidate polymers for fuel cell membranes, plotted in a property space defined by the total permeability of hydrogen and oxygen gases (H_2&O_2 permeability) and T_g. The corresponding repeat unit structures and key property values of these polymers are shown below. Only polymers with well-balanced properties suitable for the target applications were retained, while those with evidently incompatible characteristics were excluded. Blue boxes indicate the predefined screening thresholds.

https://doi.org/10.1007/s10118-025-3402-y

tion, and electron-withdrawing groups like carbonyl (C=O) and thiocarbonyl (C=S) functionalities that elevate dipolar polarizability and thus increase dielectric constant. This design philosophy aligns with the established molecular principles for high-temperature dielectrics, which emphasize the integration of thermally robust, conformationally restricted backbones with high-permittivity functional groups.[49,50] Although the selected materials have not yet been experimentally applied in HTPD systems, their constituent features resemble those found in high-performance aromatic polyketones and polyetherketones, which are known to exhibit superior dielectric properties above 473 K.[51] Accordingly, the newly identified polymers represent promising candidates for HTPDs and warrant further synthetic tuning and experimental validation.

The design of polymer membranes for PEM fuel cells involves a more intricate set of requirements due to the high-temperature, hydrated environment inherent to fuel cell operation. Ideal materials must combine low gas permeability (to minimize H_2 or O_2 crossover), moderate T_g (balancing thermal stability and processability), and adequate stiffness (to prevent membrane embrittlement).[52] We therefore construct a three-dimensional screening space using H_2&O_2 permeability (the sum of H_2 and O_2 permeability), T_g, and Young's modulus as selection criteria. As illustrated in Fig. 5(b), three well-balanced candidates emerge, including chitosan, cellulose, and poly(2,6-dimethyl-1,4-phenylene oxide) (PPO). Chitosan and cellulose are renewable, biodegradable polymers that offer excellent thermomechanical stability and have been extensively investigated in the context of green membrane technologies. PPO, on the other hand, is widely recognized for its superior thermal stability, mechanical strength, and tunable chemical structure. Notably, PPO can be readily functionalized with ion-conducting groups such as quaternary ammonium or imidazolium moieties to enhance proton or hydroxide conductivity.[53] Recent experimental advances further validate the suitability of these materials under practical fuel cell conditions. For instance, Han *et al.*[54] demonstrated that crosslinked PPO-based membranes maintain mechanical robustness (tensile strength > 10 MPa after acid doping) and exhibits a thermal degradation onset above 473 K. Similarly, Guccini *et al.*[55] developed carboxylated cellulose nanofibre membranes with a hydrogen permeability of approximately 60 Barrer—substantially lower than that of Nafion—while preserving a high dry-state Young's modulus of 15 GPa. Moreover, Swaghatha *et al.*[56] reported that a MoS_2–NiO–Co_3O_4 nanocomposite-filled chitosan membrane achieves an even lower hydrogen permeability of 20.43 barrer, along with excellent tensile strength (25.63 MPa) and operational stability at 470 K.

Together, these two case studies demonstrate the broader potential of OpenPoly as a data-driven platform for targeted polymer discovery. By coupling standardized experimental data with predictive machine learning models, OpenPoly enables the transition from performance prediction to application-specific molecular design, offering a scalable and extensible framework for accelerating materials innovation across energy, environmental, and engineering domains.

CONCLUSIONS

We present OpenPoly, a curated, standardized and open-access polymer property database comprising 3985 structure–property pairs across 26 experimental properties, establishing a reliable foundation for AI-driven polymer prediction. Leveraging this dataset, we develop a multi-task benchmarking framework spanning four encoding strategies and eight representative models. XGBoost consistently outperforms deep learning models under data-scarce and noisy conditions, demonstrating superior robustness and efficiency. Notably, Bland-Altman analysis further supports that encoding polymers with N = 2 using Morgan fingerprints provides sufficient structural resolution while minimizing computational cost, making it a practical default for most property prediction tasks. Beyond benchmarking, we demonstrate that OpenPoly effectively supports application-specific polymer selection by enabling performance-driven candidate screening. Despite its advantages, OpenPoly currently focuses on homopolymers and inherits limitations from heterogeneous experimental data. Future efforts will be focusing on expanding polymer structural diversity, improving data uniformity, and developing unified encoding schemes for generalizable, multi-task learning. We demonstrate that OpenPoly provides both a reproducible data resource and a standardized benchmark for advancing polymer informatics.

METHODS

Molecular Encoding Strategies

Morgan fingerprint encoding: Polymer PSMILES representing a single repeat unit was first expanded to N=2. A Morgan fingerprint with a radius of 2 was then computed using RDKit and flattened into a 2048-dimensional vector. This representation served as the input feature for XGBoost, TabPFN, and MLP models.

2D graph-based encoding: Using the end-capped SMILES of each polymer, a molecular graph was constructed *via* RDKit. Node features were represented by five-dimensional vectors encoding atomic number, chirality, hybridization state, aromaticity, and ring membership, yielding a tensor of shape [N, 5], where N denotes the number of atoms. Edge connectivity was defined as fully connected, with edge indices of shape [2, E], where E denotes the total number of edges. This encoding was used as input for GCN and GAT models.

3D graph-based encoding: Building upon the 2D encoding, spatial edges were added based on atomic distances within a 5 Å cutoff, forming a 3D molecular graph. The 3D cartesian coordinates were generated from RDKit's initial geometry and subsequently optimized using the UFF force field.[57] This encoding was employed for Spherenet models, which require spatial information for geometric message passing.

Polymer-specific fingerprint encoding: Polymer PSMILES strings were fed directly into the pretrained PolyBERT model to obtain 600-dimensional embedding vectors. These embeddings were used for downstream property regression tasks.

Model Training Protocols

To ensure consistent evaluation across models, the dataset was randomly split into training, validation, and test sets at an 8:1:1 ratio. All models were trained, tuned, and tested on the

https://doi.org/10.1007/s10118-025-3402-y

same data partitions. In addition, to address the inherent imbalance in the distribution of property values—especially for properties with long-tailed or multimodal distributions—we implemented a stratified sampling strategy based on value ranges. Specifically, for each property, its numerical value range was divided into 10 equal-width bins, and samples were proportionally drawn from each bin to construct the training, validation, and test subsets. This approach ensures that both low-frequency extreme values and high-density central regions are adequately represented during model training and evaluation, thereby enhancing model robustness and reducing the risk of bias toward dominant value regimes.

Training of deep learning models

All deep learning models, including GCN, GAT, MLP, SphereNet, PolyBERT, and UniMol2 were trained using PyTorch version 2.5.1 on CUDA-enabled GPUs. Training followed a supervised learning paradigm with independent data loaders for training, validation, and testing phases. To prevent overfitting, early stopping was employed based on validation performance. Models were optimized using the Adam optimizer (learning rate = 5×10^{-4}, weight decay = 5×10^{-4}), and MSE as the loss function. After each epoch, models were evaluated on the validation set with gradient computation disabled. Validation MSE was used to monitor learning and trigger three-stage early stopping criteria: (1) If no improvement was observed for 20 consecutive epochs, the learning rate was reduced by a factor of 10. (2) Training was terminated if the learning rate fell below 1/1000 of its original value with no further improvement. (3) The model checkpoint was updated whenever the validation loss reached a new minimum. Upon convergence, the best-performing model (based on validation MSE) was reloaded and evaluated on an independent test set.

Training of the XGBoost model

XGBoost was implemented using the XGBRegressor class and served as a baseline model. Hyperparameter optimization was conducted using Optuna, a Bayesian optimization framework, to minimize validation MSE. The search space included the key hyperparameters, including maximum tree depth, learning rate, number of boosting rounds, minimum sum of instance weight in a child, minimum loss reduction required for further partitioning, sampling ratios for rows and features, L1 and L2 regularization terms. Each trial involved model fitting on the training set and evaluation on the validation set. A total of 100 optimization trials or a 600-second time limit was used to identify the optimal hyperparameter set. The final model was retrained on the entire training set using the best parameters and evaluated on the test set using MSE as the performance metric.

Training of the TabPFN model

The TabPFN model was implemented *via* the official AutoTabPFN Regressor for property regression. Fine-tuning was performed on the training set starting from pretrained weights. Input features were identical to those used in the XGBoost model.

Deep Learning Architectures

MLP: The MLP model takes Morgan fingerprint vectors (dimension = 2048) as input and outputs scalar property predictions. The architecture comprises three fully connected layers with hidden dimensions of 256 and 64. Each hidden layer is followed by batch normalization, a ReLU activation, and a dropout layer (dropout rate = 0.3). The final output layer is a linear transformation mapping the last hidden state to the target property.

Graph Neural Networks (GCN and GAT): Both GCN and GAT models were implemented using standard modules from the PyTorch Geometric library. Each network consists of three graph convolutional layers that operate on molecular graph inputs with node features and edge indices. The GCN uses 64-dimensional hidden layers, while the GAT employs 4 attention heads and 256-dimensional hidden representations. All intermediate layers are followed by eLU activations. A final linear layer outputs the property prediction.

Spherenet: The Spherenet architecture was adopted from the DIG official repository (https://github.com/divelab/DIG/tree/dig-stable). It operates on 3D molecular graphs with Cartesian coordinates and predicts scalar properties. Default hyperparameters and model settings were used without further tuning.

PolyBERT: PolyBERT was implemented using the pretrained model available at https://github.com/Ramprasad-Group/polyBERT. Given the polymer PSMILES string as input, PolyBERT generates a 600-dimensional embedding vector that is subsequently fed into a linear regression head for property prediction. All pretrained parameters were kept frozen during training.

UniMol2: UniMol2 predictions were obtained using the official UniMol2-1.1B checkpoint (https://github.com/deepmodeling/Uni-Mol/tree/main/unimol2). The model encodes terminally capped SMILES into a 1536-dimensional molecular representation. A linear regression head was added on top for scalar property prediction. All pretrained weights were frozen during evaluation.

Conflict of Interests

The authors declare no interest conflict.

Data Availability Statement

All data supporting this study are openly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Code Availability: The model code, training scripts, and evaluation workflows are publicly accessible at: https://github.com/WangGroupFDU/Openpoly_benchmark

OpenPoly Database: OpenPoly database can be accessed *via* the link: https://cleanenergymaterials.cn/polymer/polymer_database/experiment_polymer_database

Online prediction service: To support practical applications, pretrained models are available at: https://cleanenergymaterials.cn/polymer/polymer_predict_page

Users can access the web interface to predict polymer properties by uploading tabulated input data. Access to the platform requires user registration and login authentication.

ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (Nos. 92372126 and 52373203)

and the Excellent Young Scientists Fund Program. The computations in this research were performed using the CFFF platform of Fudan University.

REFERENCES

1 Wang, Y. Application-oriented design of machine learning paradigms for battery science. *npj Comput. Mater.* **2025**, *11*, 89.
2 Liu, Y.; Madanchi, A.; Anker, A. S.; Simine, L.; Deringer, V. L. The amorphous state as a frontier in computational materials design. *Nat. Rev. Mater.* **2025**, *10*, 228–241.
3 Ge, W.; De Silva, R.; Fan, Y.; Sisson, S. A.; Stenzel, M. H. Machine Learning in Polymer Research. *Adv. Mater.* **2025**, *37*, 2413695.
4 Audus, D. J.; de Pablo, J. J. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* **2017**, *6*, 1078–1082.
5 Li, Y. Q.; Jiang, Y.; Wang, L. Q.; Li, J. F. Data and machine learning in polymer science. *Chinese J. Polym. Sci.* **2023**, *41*, 1371–1376.
6 Sha, W.; Li, Y.; Tang, S.; Tian, J.; Zhao, Y.; Guo, Y.; Zhang, W.; Zhang, X.; Lu, S.; Cao, Y. C. Machine learning in polymer informatics. *InfoMat* **2021**, *3*, 353–361.
7 Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, **2011**; IEEE: pp 22-29.
8 Ma, R.; Luo, T. PI1M: a benchmark database for polymer informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684–4690.
9 Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A polymer dataset for accelerated property prediction and design. *Sci. Data* **2016**, *3*, 1–10.
10 Kim, S.; Schroeder, C. M.; Jackson, N. E. Open macromolecular genome: Generative design of synthetically accessible polymers. *ACS Polymers Au* **2023**, *3*, 318–330.
11 Wilson, A. N.; St John, P. C.; Marin, D. H.; Hoyt, C. B.; Rognerud, E. G.; Nimlos, M. R.; Cywar, R. M.; Rorrer, N. A.; Shebek, K. M.; Broadbelt, L. J. PolyID: Artificial intelligence for discovering performance-advantaged and sustainable polymers. *Macromolecules* **2023**, *56*, 8547–8557.
12 Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured information extraction from scientific text with large language models. *Nat. Commun.* **2024**, *15*, 1418.
13 Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **2024**, *15*, 1569.
14 Kong, J.; Panapitiya, G.; Saldanha, E. Extracting material property measurements from scientific literature with limited annotations. *J. Chem. Inf. Model.* **2025**, *65*, 4906–4917.
15 Gupta, S.; Mahmood, A.; Shetty, P.; Adeboye, A.; Ramprasad, R. Data extraction from polymer literature using large language models. *Commun. Mater.* **2024**, *5*, 269.
16 Jiang, S.; Dieng, A. B.; Webb, M. A. Property-guided generation of complex polymer topologies using variational autoencoders. *npj Comput. Mater.* **2024**, *10*, 139.
17 Martin, T. B.; Audus, D. J. Emerging trends in machine learning: a polymer perspective. *ACS Polym. Au* **2023**, *3*, 239–258.
18 Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. *npj Comput. Mater.* **2023**, *9*, 64.
19 Qiu, H.; Liu, L.; Qiu, X.; Dai, X.; Ji, X.; Sun, Z. Y. PolyNC: a natural and chemical language model for the prediction of unified polymer properties. *Chem. Sci.* **2024**, *15*, 534–544.
20 Agarwal, S.; Mahmood, A.; Ramprasad, R. Polymer solubility prediction using large language models. *ACS Mater. Lett.* **2025**, *7*, 2017–2023.
21 Liu, N.; Jafarzadeh, S.; Lattimer, B. Y.; Ni, S.; Lua, J.; Yu, Y. Harnessing large language models for data-scarce learning of polymer properties. *Nat. Comput. Sci.* **2025**, *5*, 245–254.
22 Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
23 Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **2018**, *122*, 17575–17585.
24 Brandrup, J.; Immergut, E. H.; Grulke, E. A.; Abe, A.; Bloch, D. R. *Polymer Handbook*, 4th ed, Wiley, New York, **1999**, p. 705–763
25 Mark, J. E. *Polymer Data Handbook*, 2nd ed, Oxford University Press, New York, **2009**, p. 114-973
26 Hoaglin, D. C.; Iglewicz, B.; Tukey, J. W. Performance of some resistant rules for outlier labeling. *J. Am. Stat. Assoc.* **1986**, *81*, 991–999.
27 Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nat. Commun.* **2023**, *14*, 1403.
28 Finegan, D. P.; Vamvakeros, A.; Tan, C.; Heenan, T. M. M.; Daemi, S. R.; Seitzman, N.; Di Michiel, M.; Jacques, S.; Beale, A. M.; Brett, D. J. L.; et al. Spatial quantification of dynamic inter and intra particle crystallographic heterogeneities within lithium ion electrodes. *Nat. Commun.* **2020**, *11*, 631.
29 Jiang, B.; Zhang, Z.; Lin, D.; Tang, J.; Luo, B. Semi-supervised learning with graph learning-convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2019**; pp 11313–11320.
30 Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, **2017**.
31 Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun*. **2023**, *14*, 4099.
32 Wang, Z.; Gao, Z.; Zheng, H.; Zhang, L.; Ke, G. Exploring molecular pretraining model at scale. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 46956–46978.
33 Liu, Y.; Wang, L.; Liu, M.; Lin, Y.; Zhang, X.; Oztekin, B.; Ji, S. Spherical message passing for 3d molecular graphs. In *International Conference on Learning Representations (ICLR)*, **2022**.
34 Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S. B.; Schirrmeister, R. T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326.
35 Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
36 Doğan, N. Ö. Bland-Altman analysis: a paradigm to understand correlation and agreement. *Turk. J. Emerg. Med.* **2018**, *18*, 139–141.
37 Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. *npj Comput. Mater.* **2023**, *9*, 42.
38 Zantvoort, K.; Nacke, B.; Görlich, D.; Hornstein, S.; Jacobi, C.; Funk, B. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digital Med.* **2024**, *7*, 361.
39 Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520.
40 McElfresh, D.; Khandagale, S.; Valverde, J.; Prasad C, V.; Ramakrishnan, G.; Goldblum, M.; White, C. When do neural nets outperform boosted trees on tabular data. *Adv. Neural Inf.*

Process. Syst. **2023**, *36*, 76336–76369.
41 Aydin, Z. E.; Ozturk, Z. K. Performance analysis of XGBoost classifier with missing data. *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)* **2021**, *2*, 2021.
42 Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; Tabona, O. A survey on missing data in machine learning. *J. Big Data* **2021**, *8*, 140.
43 Zhang, P.; Kearney, L.; Bhowmik, D.; Fox, Z.; Naskar, A. K.; Gounley, J. Transferring a molecular foundation model for polymer property predictions. *J. Chem. Inf. Model.* **2023**, *63*, 7689–7698.
44 Ye, X.; Zhang, Z.; Lan, H.; Zhang, C.; Wang, L.; Lin, J.; Xu, X.; Xu, Y.; Du, L.; Tian, X. Design of thermosetting polymers with high thermal stability and enhanced processability *via* ML-assisted material genome approach. *Macromolecules* **2025**, *58*, 5090–5100.
45 Kazemi-Khasragh, E.; González, C.; Haranczyk, M. Toward diverse polymer property prediction using transfer learning. *Comput. Mater. Sci* **2024**, *244*, 113206.
46 Joshi, C. K.; Bodnar, C.; Mathis, S. V.; Cohen, T.; Lio, P. On the expressive power of geometric graph neural networks. In *International conference on machine learning*, **2023**; PMLR: pp 15330–15355.
47 Li, H.; Zhou, Y.; Liu, Y.; Li, L.; Liu, Y.; Wang, Q. Dielectric polymers for high-temperature capacitive energy storage. *Chem. Soc. Rev.* **2021**, *50*, 6369–6400.
48 Li, X.; Hu, P.; Jiang, J.; Pan, J.; Nan, C. W.; Shen, Y. High-temperature polymer composite dielectrics: energy storage performance, large-scale preparation, and device design. *Adv. Mater.* **2025**, *37*, 2411507.
49 Yang, M.; Ren, W.; Jin, Z.; Xu, E.; Shen, Y. Enhanced high-temperature energy storage performances in polymer dielectrics by synergistically optimizing band-gap and polarization of dipolar glass. *Nat. Commun.* **2024**, *15*, 8647.
50 Zhang, Q.; Xie, Q.; Wang, T.; Huang, S.; Zhang, Q. Scalable all polymer dielectrics with self-assembled nanoscale multiboundary exhibiting superior high temperature capacitive performance. *Nat. Commun.* **2024**, *15*, 9351.
51 Kamishima, T.; Inagaki, K.; Nukazuka, A.; Iwata, T.; Enomoto, Y. Synthesis and characterization of aromatic polyketones and polyetherketones derived from divanillic acid *via* Friedel–Crafts acylation. *Eur. Polym. J.* **2025**, *228*, 113823.
52 Tran, H.; Gurnani, R.; Kim, C.; Pilania, G.; Kwon, H. K.; Lively, R. P.; Ramprasad, R. Design of functional and sustainable polymers assisted by artificial intelligence. *Nat. Rev. Mater.* **2024**, *9*, 866–886.
53 Liu, L.; Chu, X.; Liao, J.; Huang, Y.; Li, Y.; Ge, Z.; Hickner, M. A.; Li, N. Tuning the properties of poly(2,6-dimethyl-1,4-phenylene oxide) anion exchange membranes and their performance in H_2/O_2 fuel cells. *Energy Environ. Sci.* **2018**, *11*, 435–446.
54 Han, Y.; Xu, F.; Ji, J.; Li, Y.; Chu, F.; Lin, B. Phosphoric acid-doped cross-linked poly(phenylene oxide)-based membranes for high temperature proton exchange membrane fuel cells. *Int. J. Hydrogen Energy* **2024**, *50*, 1417–1426.
55 Guccini, V.; Carlson, A.; Yu, S.; Lindbergh, G.; Lindström, R. W.; Salazar-Alvarez, G. Highly proton conductive membranes based on carboxylated cellulose nanofibres and their performance in proton exchange membrane fuel cells. *J. Mater. Chem. A* **2019**, *7*, 25032–25039.
56 Swaghatha, A. A. K.; Cindrella, L. Improved proton conductivity in MoS_2–NiO–Co_3O_4 filled chitosan based proton exchange membranes for fuel cell applications. *Mater. Chem. Phys.* **2022**, *290*, 126654.
57 Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.