Is BigSMILES the Friend of Polymer Machine Learning?

Haoke Qiu[†,‡] and Zhao-Yan Sun[*,†,‡]

[†]*State Key Laboratory of Polymer Physics and Chemistry & Key Laboratory of Polymer Science and Technology, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China*

[‡]*School of Applied Chemistry and Engineering, University of Science and Technology of China, Hefei 230026, China*

E-mail: zysun@ciac.ac.cn

Phone: +86 (0431) 85262896

Abstract

Machine learning (ML) has become a powerful tool in polymer science, with its success strongly relying on effective structural representations of polymers. While the Simplified Molecular Input Line Entry System (SMILES) is widely used due to its simplicity, it was originally designed for small molecules and struggles to capture the stochastic nature of polymers. Recently, BigSMILES has been introduced as a more compact and versatile representation of polymer structures. However, the relative performance of SMILES and BigSMILES in polymer ML tasks remains unexplored. In this study, we systematically evaluate SMILES and BigSMILES across 12 polymer-related tasks, including property prediction and inverse design, utilizing convolutional neural networks (CNNs) and large language models (LLMs). Our results show that BigSMILES enables faster training times due to its reduced token complexity, and achieves comparable or superior performance to SMILES in certain predictive tasks. Moreover, BigSMILES more accurately encodes chemical information and monomer connectivity for copolymers within LLM frameworks. This work serves as a starting point for a comprehensive evaluation of SMILES and BigSMILES in polymer ML applications, highlighting the potential of BigSMILES to streamline and accelerate polymer informatics workflows, particularly for complex systems like copolymers and polymer composites. Looking ahead, advancing polymer representations to integrate polymer chain structure, phase morphology, and processing parameters will be crucial for capturing the multifaceted relationships between polymer structure and properties, driving more accurate and efficient modeling.

Introduction

The sustainable development of polymers has long been a key objective in polymer science. Numerous polymers have been developed across various fields, including aerospace,[1] environmental science,[2] energy device,[3] healthcare,[4,5] and others,[6,7] playing a central role in advancing modern technologies. However, achieving sustainability requires not only sustainable materials but also efficient and environmentally friendly development processes. Traditional trial-and-error approaches, often involving complex orthogonal experiments, are time-consuming and resource-intensive, highlighting the urgent need for streamlined polymer design and development workflows.

Machine learning (ML) methods have proven their efficiency and effectiveness in accelerating molecular and materials discovery.[8] A key determinant of ML is believed as how effectively the structural information of polymers is represented.[9–15] Early ML applications in polymer science primarily relies on numerical descriptors derived from cheminformatics tools such as RDKit[16] and Mordred.[17] While these numerical descriptors are effective in conventional ML models like Random Forest and Gaussian Process,[18,19] they are often atom/bond-wise, and struggle to capture both short-range and long-range interactions within polymers, limiting their performance in complex tasks. Graph neural networks (GNNs) have since gained increasing attention, offering the ability to capture molecular structural information at the atomic, bond, and functional group levels.[20–24] As the volume of molecular and polymer data continues to grow, GNNs and descriptor-based methods often face scalability challenges, since representing a single molecule often requires hundreds or even thousands of descriptors.[25–27]

String-based molecular representations, such as the Simplified Molecular Input Line Entry System (SMILES), provide an alternative approach to addressing these challenges. SMILES was introduced in 1988 by Weininger[28] and has become a cornerstone of molecular informatics due to its simplicity and compactness. Other string-based notations include the SYBYL Line Notation (SLN),[29] the Modular Chemical Descriptor Language (MCDL),[30] and the International Chemical Identifier (InChI).[31] Benefiting from recent advances in natural language processing (NLP) and Transformer-based models,[32–35] string-based (especially SMILES-based) ML workflows have achieved remarkable success in extracting high-dimensional chemical information. SMILES was originally designed for small molecules and faces inherent limitations when applied to polymers, particularly in representing stochasticity and polymerization sites.

To address these limitations, various extensions of SMILES have been developed for polymers. For instance, Polymer-SMILES (P-SMILES)[36] incorporates special symbols "*" to denote polymerization points, enhancing its utility in polymer-specific tasks, but still failing to represent more complex polymer structures and the inherent randomness of polymers. Recently, BigSMILES[37] was introduced as a more comprehensive representation capable of encoding a wide range of polymer structures, including random copolymers and block copolymers. BigSMILES has since gained significant attention within the polymer community[38–41] and is now becoming as the default representation in polymer databases such as the Community Resource for Innovation in Polymer Technology (CRIPT).[42] Despite its theoretical advantages, the practical performance of BigSMILES in ML tasks remains largely unexplored, leaving a critical gap in our understanding of its application relative to SMILES.

In this work, we present a systematic comparison of SMILES (or P-SMILES) and BigSMILES in polymer ML workflows, evaluating their performance across 12 diverse tasks, focusing on polymer property prediction and molecular generation tasks. Using convolutional neural networks (CNNs), deep neural networks (DNNs), and large language models (LLMs), we assess both representations in terms of prediction accuracy, training efficiency, and their ability to encode polymer inherent structures. Our results illustrate that while SMILES achieves competitive performance in certain tasks, BigSMILES enables shorter training times due to more concise encoding of chemical information, particularly for copolymer systems.

By bridging the gap in our understanding of SMILES and BigSMILES in polymer ML applications, this work provides a foundation for estimating polymer representations in data-driven workflows. As polymer datasets continue to expand, the efficiency gained by BigSMILES will have more potential to significantly advance sustainable polymer design and modeling practices. Future efforts should focus on developing next-generation polymer representations that integrate chain structures, aggregation behaviors, and processing conditions to further enhance the predictive power of ML models.

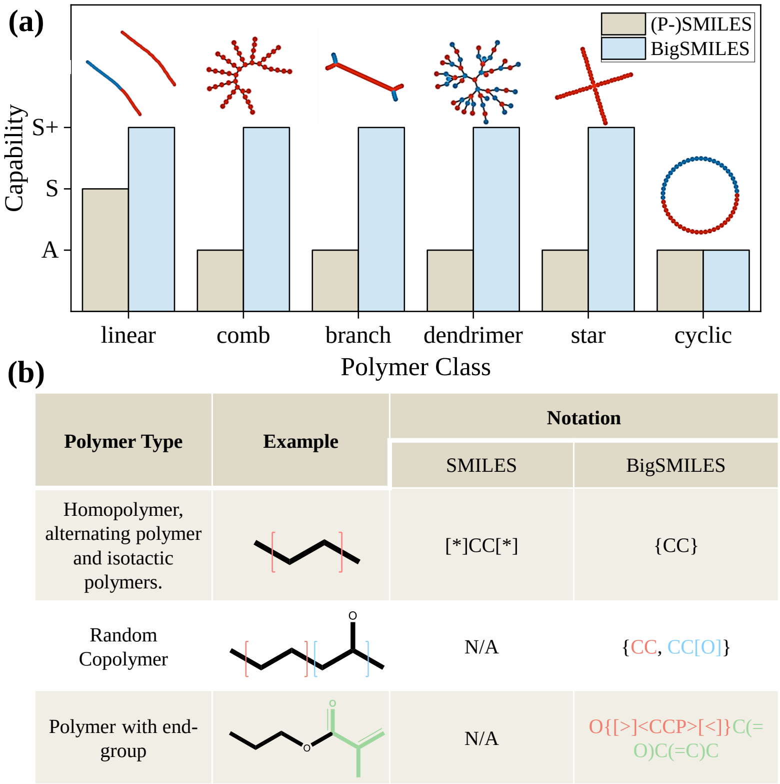
Results and discussion

Challenges of SMILES in representing complex polymers

Compared to descriptor-based and graph-based polymer representations, (P-)SMILES provides a notably concise way of encoding polymer chemistry, largely relying on the identification of a minimal repeating unit. However, this advantage diminishes when the repeating unit cannot be clearly determined, a limitation often encountered in complex polymer systems. Polymers are typically classified as homopolymers or copolymers based on their monomer combinations and can be further categorized into various structural forms, such as linear, comb, branched, dendrimeric, star-shaped, and cyclic architectures.[43] These diverse structural configurations introduce inherent randomness in many polymers, posing significant challenges for (P-)SMILES. Unlike small molecules with well-defined chemical structures, polymers often exhibit stochastic variations in their connectivity or composition, making it difficult for SMILES to concisely represent these structural complexities.

For effective polymer representation, an ideal method should encode the polymer chain using only the essential information about the repeating units and their connection patterns. However, for polymers without a well-defined minimal repeating unit, such as random copolymers, (P-)SMILES representations become excessively long and complex, as they must fully enumerate all repeating units and their connections to preserve structural accuracy. In constrast, BigSMILES addresses this challenge by introducing specific operators (e.g., "[O]") to indicate randomness at the end of repeating unit notations. This feature greatly simplifies the representation of complex systems like block copolymers, enabling concise yet informative descriptions.

To evaluate the ability of (P-)SMILES and BigSMILES to represent diverse polymer structures concisely, we compared the corresponding representation of six common polymer classes: linear, comb, branched, dendrimeric, star-shaped, and cyclic structures. Representation conciseness was categorized into three grades: methods capable of achieving concise representations for all polymer types were assigned a grade of **S**+; those that succeeded partially, such as adequately representing homopolymers but struggling with complex copolymers, were rated as **S**; and methods unable to provide concise representations of both homopolymers and copolymers were rated as **A**. The results of this assessment are presented in Figure 1(a).



Polymer Type	Example	Notation	
		SMILES	BigSMILES
Homopolymer, alternating polymer and isotactic polymers.		[*]CC[*]	{CC}
Random Copolymer		N/A	{CC, CC[O]}
Polymer with end-group		N/A	O{[>]<CCP>[<]}C(=O)C(=C)C

Figure 1: (a) Representation grades of (P-)SMILES/BigSMILES on six classes[43] of polymers. For simplicity, only binary copolymerization is illustrated (red and blue beads, respectively). (b) Examples of (P-)SMILES/BigSMILES on three typical polymers.

Several examples illustrating the differences in representation between (P-)SMILES and BigSMILES are shown in Figure 1(b). For instance, while (P-)SMILES can provide clear repersentations for linear homopolymers, the resulting string lengths are still longer than those generated by BigSMILES, as seen in the first row of Figure 1(b). For polymers containing multiple repeating units or featuring complex topologies, such as block copolymers or branched polymers, SMILES struggles to concisely encode structure details. In contrast, BigSMILES extends the functionality of string-based representations by introducing stochastic operators, allowing for a simplified yet coarse-grained encoding of such structures. This makes BigSMILES a more versatile and efficient tool for representing complex polymers, as demonstrated by Olsen et al.[37]

Property prediction of homopolymers using ML

To evaluate the applicability of (P-)SMILES and BigSMILES in practical polymer informatics workflows, we benchmarked their performance in property prediction tasks specific to homopolymers. ML models, including both conventional algorithms and deep learning architectures, are typically optimized for numerical data. For string-based data such as SMILES and BigSMILES, two primary approaches are used for encoding polymer chemical information directly from these strings (as illustrated in Figure 2).

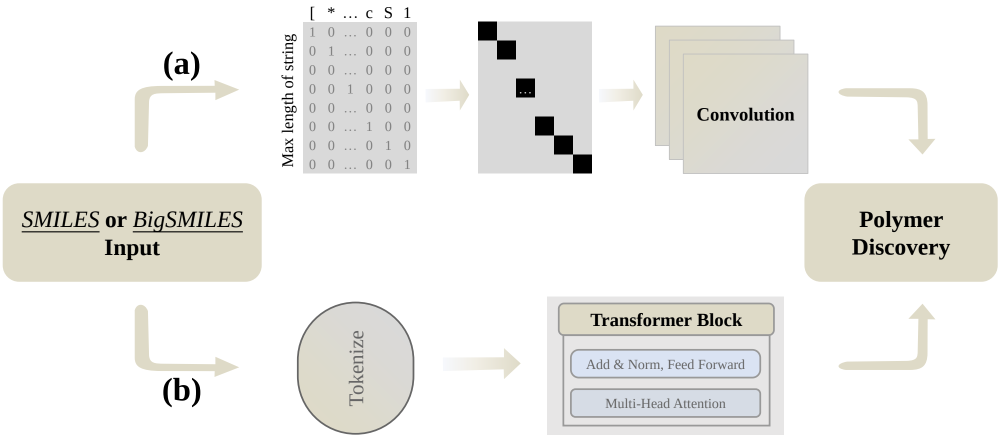


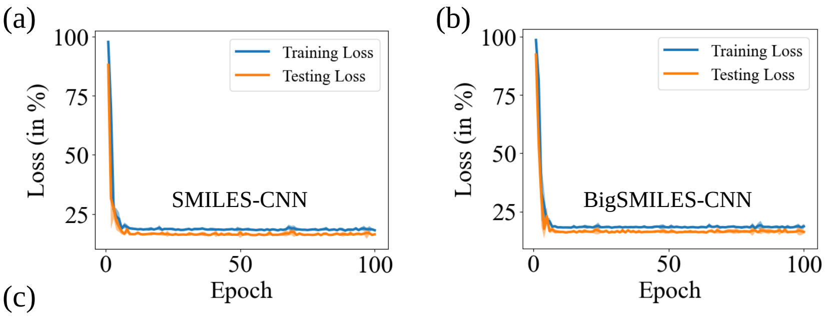
Figure 2: Two approaches that string-based polymer representation can be used as ML inputs. (a) Textual polymer representation is first transformed to images and then learned by CNNs. (b) Textual polymer representation is directly served as input of ML models, such as the LLMs.

The first approach converts the string-based data into binary image representations, enabling CNNs to extract chemical features.[44,45] The second approach takes inspiration from NLP and employs sequence-learning models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs),[46,47] and advanced architectures like Transformers and LLMs.[32,33,35] Given that SMILES is predominantly utilized for homopolymer informatics and BigSMILES demonstrates substantial advantages in describing complex polymer architectures,[37,39,41] our evaluation focuses on comparing their effectiveness specifically in homopolymer property prediction tasks to ensure a fair and consistent analysis.

Performance in CNN-Based Models

We first used binary image representations derived from SMILES and BigSMILES as inputs to a CNN model to predict the glass transition temperature (Tg) of various homopolymers. The dataset and model parameters were adopted from prior study.[46] To minimize variability, each configuration was trained for 100 epochs and repeated five times. Results were assessed using the relative absolute error (RAE), defined as the absolute percentage error of the predicted value relative to the true value.

Under identical training configurations, the SMILES-CNN achieved a test-set RAE of 16.46±0.12%, while the BigSMILES-CNN showed a comparable performance with an RAE of 16.57±0.28%. Both models achieved prediction errors within the experimentally accepted range for Tg, regarding experimental measurement uncertainties.[22] We further illustrate some random examples in Figure 3(c), highlighting the predictive accuracy of both methods, with predictions closer to true values emphasized in bold. These results suggest that both SMILES and BigSMILES have their own merits for homopolymer property prediction when using CNN-based models.



(c)

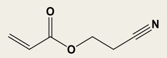
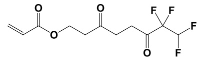
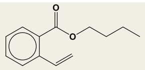
Polymer	Referred Tg/K	Predicted Tg/K	
		SMILES-CNN	BigSMILES-CNN
	277	345.38	**278.03**
	233	309.34	**274.46**
	339	**315.38**	278.03

Figure 3: Performance comparison of SMILES (a) and BigSMILES (b) in the CNNs. (c) shows the prediction results for three example polymers, illustrating that BigSMILES-CNN and SMILES-CNN exhibit comparable inference performance on the test set. Detailed prediction lists are provided in the Section 1 of the Supporting Information (S1).

Performance in LLM-Based Models

We then explore the application of these two text-based polymer representations in LLMs, which can directly process text-based inputs without additional preprocessing. Using the end-to-end polymer LLM, PolyNC, we fine-tuned the model on nine polymer property prediction tasks, including atomization energy (AE), bandgap (BG) of polymer chains and crystals, charge injection barrier (CIB), crystallization tendency (CT), electron affinity (EA), ionization energy (IE), CO2 permeability in membranes (log-scale), and Tg of polyimides. The input data were encoded as either SMILES or BigSMILES strings, with the corresponding property values as outputs.

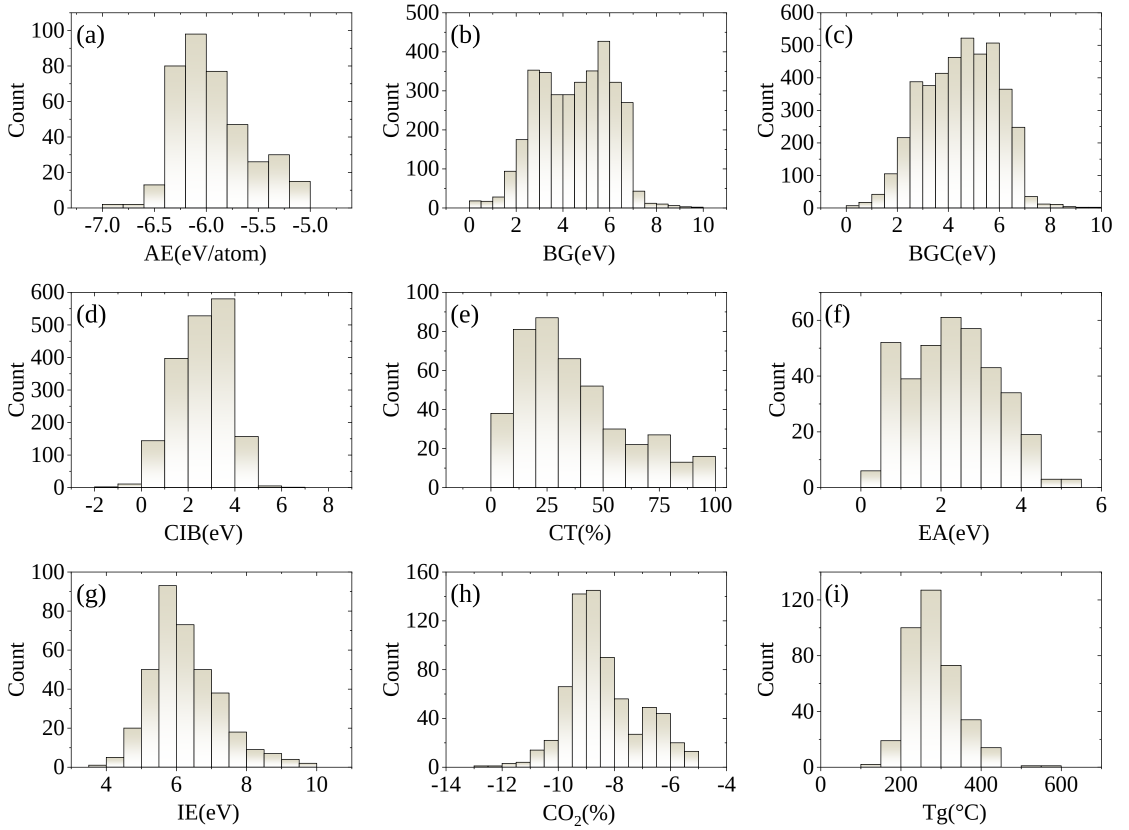


Figure 4: Data distribution. These datasets exhibit a fairly pronounced normal distribution.

Figure 4 shows the distribution of the datasets, which exhibit normal-like distributions, indicating balanced data suitable for ML modeling. These datasets cover very wide property ranges, such as crystallization tendencies spanning 0–100 and Tg values ranging from 150–450°C, typical for commonly studied polymers. The performance of fine-tuned models, evaluated using the mean absolute error (MAE), is summarized in Figure 5(a). For specific tasks, such as Tg prediction and CO2 permeability, BigSMILES-based models slightly underperformed compared to SMILES-based models. This discrepancy may stem from the pretraining stage of PolyNC, where SMILES served as the default polymer representation. However, in other tasks, BigSMILES-based models demonstrated comparable performance to SMILES-based ones, suggesting its broad utility in homopolymer informatics.

Efficiency and Tokenization Advantages of BigSMILES

We find an interesting observation during model fine-tuning, i.e., BigSMILES consistently required less computational time compared to SMILES for the same training configurations, shown in Figure 5(b). This efficiency advantage stems from the ability of BigSMILES to encode polymer structures using fewer tokens (highlighted in red in the figure) compared to most (P-)SMILES,[36,48] resulting in shorter input sequences. For instance, as illustrated in Figure 6, a polymer encoded as SMILES required 27 tokens, whereas the equivalent BigSMILES representation used only 24 tokens. Shorter token sequences reduce the size of self-attention and cross-attention matrices in Transformer-based architectures, thereby lowering computational costs and enabling faster model iterations.

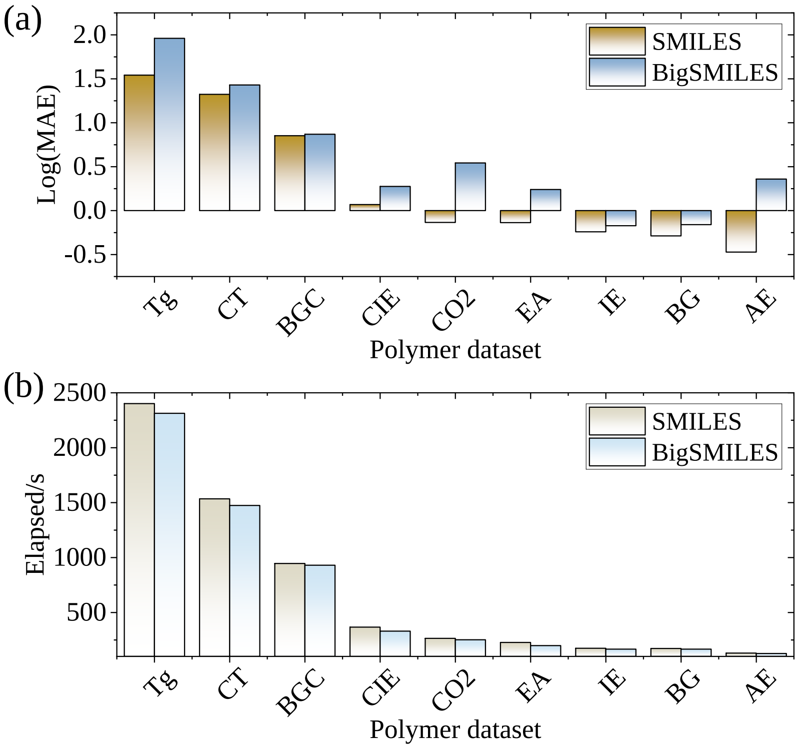
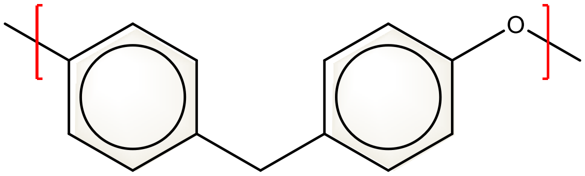


Figure 5: Performance of the two representation methods in fine-tuning PolyNC: (a) shows the model's MAE, and (b) displays the time taken for model fine-tuning (in seconds).



Item	SMILES	BigSMILES
Representation	[*]Oc1ccc(Cc2ccc([*])cc2)cc1	{<Oc1ccc(cc1)Cc2ccc(cc2)>}
Tokens	'▁[', '*', ']', 'O', 'c', '1', 'c', 'c', 'c', '(', 'C', 'c', '2', 'c', 'c', 'c', '(', '[', '*', ']', ')', 'c', 'c', '2)', 'c', 'c', '1'	'▁', '{<', 'O', 'c', '1', 'c', 'c', 'c', '(', 'c', 'c', '1)', 'C', 'c', '2', 'c', 'c', 'c', '(', 'c', 'c', '2)', '>', '}'
Length of Tokens	**27**	**24**

Figure 6: Encoding details of SMILES and BigSMILES using PolyNC's encoder. BigSMILES represents polymer structures with fewer tokens.

This tokenization efficiency is particularly important for large-scale datasets, where reduced sequence lengths translate to faster training times, lower energy consumption, and improved scalability. BigSMILES demonstrated a capability for more memory-efficient storage of polymer data compared to SMILES, which is increasingly important as polymer datasets continue to grow.[42] Note that SMILES and BigSMILES may achieve comparable levels of compactness when using the most concise representations of polymers. For instance, polyethylene can be represented as **CC** and *{CC}* for SMILES and BigSMILES respectively. However, BigSMILES excels in representing more complex polymer structures with inherent randomness.

Property Prediction of Copolymers Using ML

Now we focus on predicting the properties of copolymers, using polyhydroxyalkanoates glass transition data from the literature[49] as a case study. Following appropriate preprocessing (outlined in the Methods section), we generated SMILES-based and BigSMILES-based representations of the polymers for training LLM models. The prediction results on the same test set are shown in Figure 7(a). Both representations produced comparable accuracy, though a few structures exhibited higher prediction errors, resulting in increased statistical uncertainty (illustrated by the light blue shaded area in Figure 7(a)). These outliers suggest that the performance of both models may be improved with the inclusion of more training data.

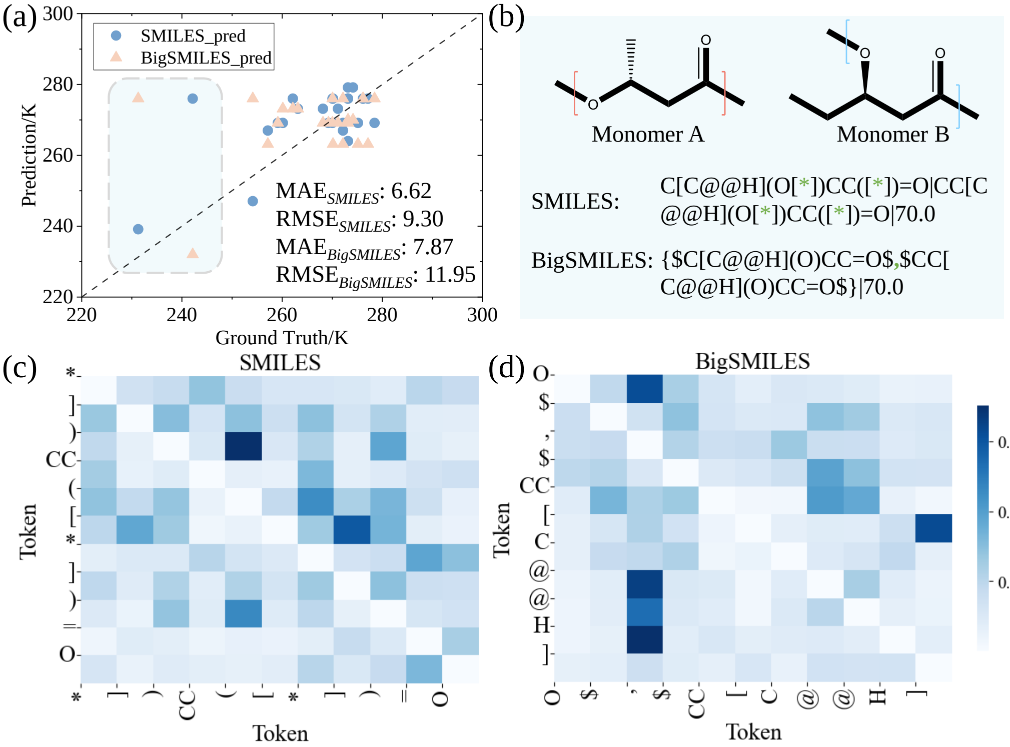


Figure 7: (a) Prediction performance on this copolymer task utilized SMILES and BigSMILES, respectively. (b) An example copolymer and their SMILES-based and BigSMILES-based input to each LLM. (c) Attention score of attention head 5 among tokens in SMILES-based LLM. (d) Attention score of attention head 5 among tokens in BigSMILES-based LLM.

To further assess whether the models captured relevant chemical trends from the two representations, we analyzed their ability to interpret polymer connectivity. In SMILES, the "*" character signifies connectivity between monomer or repeat units, while in BigSMILES, this connectivity is represented by a comma (","). As a comparative case, we selected a copolymer system for which both models achieved highly accurate predictions (with an error of only 0.1 K). The chemical composition of this copolymer system and its respective representations are shown in Figure 7(b).

To interpret how the models processed these representations, we conducted attention analysis, in which the relative importance of each token during inference can be revealed, with higher attention scores indicating greater significance. We analyzed the fifth attention head for both models, as this head mainly focuses on inter-token interactions rather than self-attention (detailed in S2). For clarity, we visualized local attention patterns for the 15 tokens surrounding the connectivity token ("*" or ","). We find notable differences in how each representation influenced attention. In SMILES, the "*" character localized attention around itself, as shown in Figure 7(c). However, the comma (",") in BigSMILES facilitated attention across a broader range of tokens (Figure 7(d)). This is arising from the chemical specificity of the comma in BigSMILES. Unlike "*", which is often interpreted as a special atom in SMILES, the comma "," in BigSMILES is often recognized as a token with unique chemical significance, enabling the model to better generalize its meaning (see S2 for detailed attention maps).

Performance on Polymer Generation Tasks

We also explored the performance of BigSMILES in polymer generation tasks, training a BigSMILES-based model under the same framework as our previously developed SMILES-based generation model PolyTAO[27] (This new model is referred as BigSMILES-based PolyTAO). The training parameters were also kept identical, with the only difference being the substitution of SMILES with BigSMILES as input. As illustrated in Figure 8, the BigSMILES-based PolyTAO exhibited faster convergence during training and achieved lower loss values on a test set of approximately 20K samples, indicating its potential superiority in polymer generation tasks.

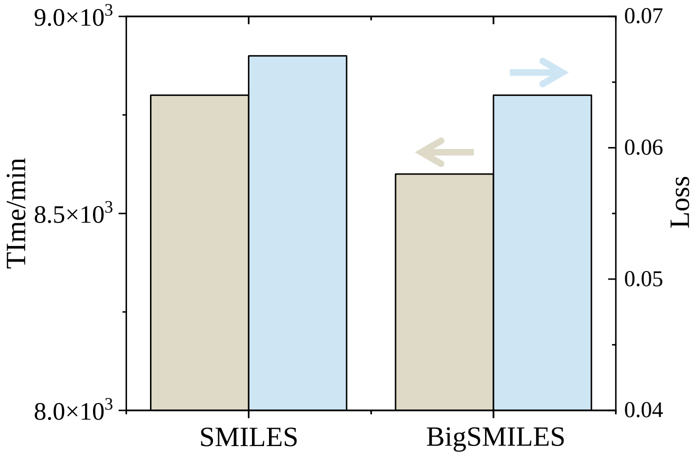


Figure 8: Model performance of BigSMILES-based PolyTAO comparing to SMILES-based PolyTAO.

Similar to the SMILES-based PolyTAO, the input features for BigSMILES-based PolyTAO consisted of 15 predefined physicochemical properties extracted from the RDKit package. These include 'MolWt', 'HeavyAtomCount', 'NHOHCount', 'NOCount', 'NumAliphaticCarbocycles', 'NumAliphaticHeterocycles', 'NumAliphaticRings', 'NumAromaticCarbocycles', 'NumAromaticHeterocycles', 'NumAromaticRings', 'NumHAcceptors', 'NumHDonors', 'NumHeteroatoms', 'NumRotatableBonds', and 'RingCount' (see S3 for details). By using these predefined features as input, the model generates polymer structures that align with the specified properties. Figure 9(a) presents a repeating unit from the test set that satisfies the input properties. Figures 9(b) to 9(f) display the repeating units of five polymers generated by the model in top-5 mode. Each subplot shows the values of the 15 predefined properties for the corresponding molecule, with 'MolWt' in text form for clarity. The results reveal a strong alignment between the repeating unit of the generated polymer and the input property specifications (Figure 9(a)). In addition, the model also explores broader chemical space, generating structures (the structures in Figures 9(b) - (f)) that differ significantly from the one in the test set. These results also demonstrate the impressive capabilities of BigSMILES-based PolyTAO in polymer generation tasks.

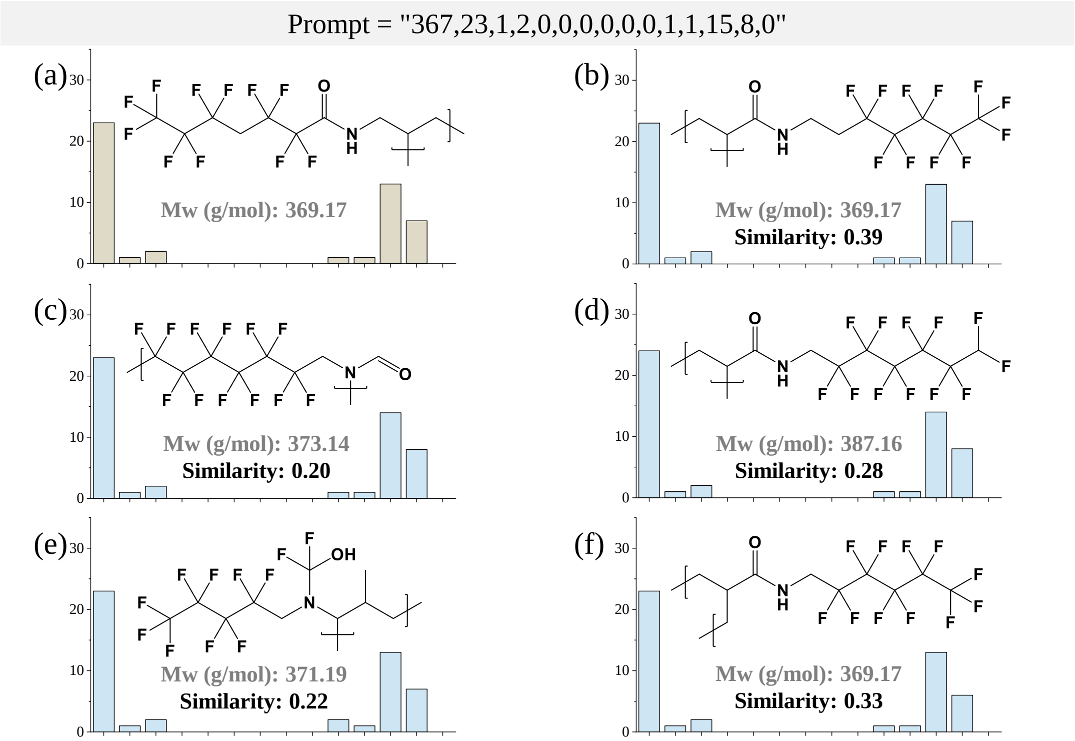


Figure 9: Generation performance of BigSMILES-based PolyTAO. In each subplot, the bars on the horizontal axis represent 'HeavyAtomCount', 'NHOHCount', 'NOCount', 'NumAliphaticCarbocycles', 'NumAliphaticHeterocycles', 'NumAliphaticRings', 'NumAromaticCarbocycles', 'NumAromaticHeterocycles', 'NumAromaticRings', 'NumHAcceptors', 'NumHDonors', 'NumHeteroatoms', 'NumRotatableBonds', and 'RingCount' from the RDKit package, while the vertical axis represents their respective values. The polymer repeating unit in subplot (a) is sourced from the test set. The polymer repeating units in (b) to (f) are generated novel molecules based on the properties of the molecule in (a).

Discussion and Conclusion

In this study, we systematically evaluated the performance of SMILES and BigSMILES, two widely used polymer representations, across various ML tasks for both homopolymers and copolymers. The results demonstrate that BigSMILES achieves comparable performance to SMILES in these tasks, highlighting its potential as an alternative representation for polymer machine learning workflows. Note that our present work only utilized the stable version of BigSMILES. Recently developed updated versions like G-BigSMILES may offer improved accuracy in polymer property prediction tasks. Future studies could explore the utility of these advanced versions to further enhance polymer ML workflows.

Another key finding of this study is that ML workflows based on BigSMILES generally require shorter training times compared to those based on SMILES, particularly in LLM scenarios. This advantage may stem from the streamlined syntax of BigSMILES, which reduces computational consumption without sacrificing chemical information. For instance, in the copolymer property prediction task, BigSMILES retained essential chemical details while facilitating faster model convergence. As the scope of polymer discovery and virtual design continues to expand, the size of training datasets for polymer ML will largely grow. Recent estimates by Li et al. suggest that the candidate space for polyimides alone could reach nearly 2×10^{12} compounds.[26] Given this immense chemical space, the use of BigSMILES could significantly accelerate the development of polymer ML pipelines, particularly in forward screening (e.g., property prediction) and inverse design (e.g., on-demand polymer generation) paradigms.

A particularly compelling advantage of BigSMILES is its ability to succinctly describe complex polymer structures, such as copolymers, which are often poorly represented by SMILES. This limitation restricts current polymer ML workflows, which primarily focus on homopolymers, as SMILES struggles to efficiently encode the structural diversity and stochasticity inherent in copolymers and other complex architectures. As polymer ML expands into these more complex domains, the limitations of SMILES cannot be ignored, requiring for more versatile representations. BigSMILES addresses this gap, and its adoption is likely to grow among polymer scientists, especially as its functionality is integrated into cheminformatics tools like RDKit for descriptor computation.

To further promote the adoption of BigSMILES and minimize the repetitive labor and energy consumption involved in converting SMILES to BigSMILES, we developed a refined database of nearly one million polymer BigSMILES representations based on the PI1M dataset. This new resource, named PI1M-BigSMILES, is freely available at `https://github.com/hkqiu/SMILES-vs-BigSMILES/blob/main/PI1M-BigSMILES.zip`. We also adapted the SMILES-based polymer generation model PolyTAO[27] and created the first BigSMILES-based polymer generation model, which is now available on Hugging Face (`https://huggingface.co/hkqiu/PolyTAO-BigSMILES_Version`).

In conclusion, our findings highlight the great potential of BigSMILES in polymer ML workflows. By enabling more concise and versatile representations, BigSMILES not only accelerates training but also expands the horizons of polymer informatics to encompass more complex structures. With the ongoing development of advanced BigSMILES variants and the creation of complementary resources like PI1M-BigSMILES, the field is well-positioned to leverage these tools for both scientific discovery and practical applications.

Methods

Batch Conversion to BigSMILES

At present, local chemical structure drawing tools do not support the direct extraction of BigSMILES. However, recent advancements from research groups led by Prof. Olsen and Prof. Seok have enabled the interconversion between molecular structure/SMILES and BigSMILES.[50,51] These tools were utilized to obtain the millions of BigSMILES entries involved in this work. Specifically, `BigSMILES\_homopolymer`[51] was employed for the conversion of homopolymer SMILES to BigSMILES, and the structure-to-BigSMILES tool developed by Olsen et al. was utilized for other polymer architectures.[50]

Text-Induced Image Convolutional Neural Network for Property Prediction

For a CNN architecture, we employed optimal network parameters based on the configuration described in prior work.[44] The network consisted of two convolutional layers with convolutional kernel sizes of (3, 3). The first layer included 256 kernels, and the second layer included 128 kernels (see ref.[44] for details). A fully connected layer with 100 neurons was used for feature extraction and regression. Pooling layers were used for down-sampling with kernel sizes of (3, 3). The input to the CNN was an image representation of polymer SMILES or BigSMILES, with the image dimensions defined as $w \times h$, where w corresponds to the length of string list and h corresponds to the maximum string length. Training was conducted on the *dataset1* of glass transition temperatures for polystyrenes and polyacrylates, as described in previous study.[44] An 80:20 train-test split was used, and the implementation was carrid out using PyTorch (version 1.12.1+cu113).

Large Language Model-Based Property Prediction

Homopolymers

Several excellent pre-trained polymer language models, such as TransPolymer,[32] polyBERT,[33] and PolyNC,[35] are available for polymer informatics tasks. In this study, PolyNC was selected as the base model for fine-tuning due to its end-to-end architecture and compatibility with polymer text descriptions. For each fine-tuning task, polymer text representations (SMILES or BigSMILES) were used as input, and the target property values served as the output. The hyperparameters used for fine-tuning are summarized in Table 1. All fine-tuning experiments were performed on four NVIDIA RTX 3090 GPUs.

Table 1: Hyperparameters during model fine-tuning.

Hyperparameter	**Configuration**
batch_size	80
epochs	100
learning_rate	1e-5
warmup_ratio	0.2
epsilon	1e-8

Copolymers

The copolymer dataset was sourced from the literature,[49] containing information on polyhydroxyalkanoate homopolymers and copolymers. This dataset included the SMILES for monomer A, the SMILES for monomer B, and the corresponding composition ratio. BigSMILES representations were derived according to established syntax rules.[37] To represent copolymers for the LLM, SMILES-based input was formatted as: "$\text{SMILES}_{MonomerA}|\text{SMILES}_{MonomerB}|\text{ratio}$", and BigSMILES-based input as "BigSMILES|ratio". The dataset was split 80:20 into training and test sets.

For model training, we used pre-trained weights from `https://huggingface.co/GT4SD/multitask-text-and-chemistry-t5-base-standard`, instead of PolyNC, to avoid interference from SMILES-based pretraining in PolyNC. Training was conducted for 200 epochs with a peak learning rate of 5×10^{-6}, using a cosine decay schedule with a 20% warm-up period. Each batch contained five samples.

Data and code availability

The training data of the CNN task and copolymer task can be accessed in prior work.[44,49] Nine properties of polymers during fine-tuning of LLMs were collected from these references.[22,24,52–55] The PI1M dataset[36] uesd for training the polymer generation model is publicly available at `https://github.com/RUIMINMA1996/PI1M`.

Our pre-trained generation model is publicly available at `https://huggingface.co/hkqiu/PolymerGenerationPretrainedModel` (SMILES version) and `https://huggingface.co/hkqiu/PolyTAO-BigSMILES_Version` (BigSMILES version). Any other data and code related to reproducing the results will be provided promptly upon request.

Acknowledgement

We thank the support from the National Key R&D Program of China (No. **2022YFB3707303**), and the National Natural Science Foundation of China (No. **52293471**). The work is also supported by the hardware in the Network and Computing Center in Changchun Institute of Applied Chemistry, Chinese Academy of Sciences.

Supporting Information Available

References

(1) Ma, P.; Dai, C.; Wang, H.; Li, Z.; Liu, H.; Li, W.; Yang, C. A Review on High Temperature Resistant Polyimide Films: Heterocyclic Structures and Nanocomposites. *Compos. Commun.* **2019**, *16*, 84–93.

(2) Toland, A.; Tran, H.; Chen, L.; Li, Y.; Zhang, C.; Gutekunst, W.; Ramprasad, R. Accelerated Scheme to Predict Ring-Opening Polymerization Enthalpy: Simulation-Experimental Data Fusion and Multitask Machine Learning. *J. Phys. Chem. A* **2023**, *127*, 10709–10716.

(3) Gurnani, R.; Shukla, S.; Kamal, D.; Wu, C.; Hao, J.; Kuenneth, C.; Aklujkar, P.; Khomane, A.; Daniels, R.; Deshmukh, A. A.; Cao, Y.; Sotzing, G.; Ramprasad, R. AI-assisted Discovery of High-Temperature Dielectrics for Energy Storage. *Nat. Commun.* **2024**, *15*, 6107.

(4) McDonald, S. M.; Augustine, E. K.; Lanners, Q.; Rudin, C.; Catherine Brinson, L.; Becker, M. L. Applied Machine Learning as a Driver for Polymeric Biomaterials Design. *Nat. Commun.* **2023**, *14*, 4838.

(5) Gao, B.; Li, D.; Li, X.; Duan, R.; Pang, X.; Cui, Y.; Duan, Q.; Chen, X. Preparation of

biocompatible, biodegradable and sustainable polylactides catalyzed by aluminum complexes bearing unsymmetrical dinaphthalene-imine derivatives via ring-opening polymerization of lactides. *Catal. Sci. Technol.* **2015**, *5*, 4644–4652.

(6) Fan, B.; Kan, Y.; Chen, B.; Han, S.; Gao, Z. A soybean adhesive with excellent hygrothermal resistance and enhanced mildew resistance via optimal synthesis of polyamidoamine–epichlorohydrin resin. *Int. J. of Adhes. Adhes.* **2022**, *118*, 103197.

(7) Lyu, Q.; Li, M.; Zhang, L.; Zhu, J. Structurally-Colored Adhesives for Sensitive, High-Resolution, and Non-Invasive Adhesion Self-Monitoring. *Nat. Commun.* **2024**, *15*, 8419.

(8) Szymanski, N. J. et al. An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials. *Nature* **2023**, *624*, 86–91.

(9) Ma, R.; Liu, Z.; Zhang, Q.; Liu, Z.; Luo, T. Evaluating Polymer Representations via Quantifying Structure–Property Relationships. *J. Chem. Inf. Model.* **2019**, *59*, 3110–3119.

(10) Tao, L.; Varshney, V.; Li, Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. *J. Chem. Inf. Model.* **2021**, *61*, 5395–5413.

(11) Aldeghi, M.; Coley, C. W. A Graph Representation of Molecular Ensembles for Polymer Property Prediction. *Chem. Sci.* **2022**, *13*, 10486–10498.

(12) Gurnani, R.; Kuenneth, C.; Toland, A.; Ramprasad, R. Polymer Informatics at Scale with Multitask Graph Neural Networks. *Chem. Mater.* **2023**, *35*, 1560–1567.

(13) Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.* **2022**, *4*, 127–134.

(14) Ji, Z.; Shi, R.; Lu, J.; Li, F.; Yang, Y. ReLMole: Molecular Representation Learning Based on Two-Level Graph Similarities. *J. Chem. Inf. Model.* **2022**,

(15) Qiu, H.; Zhao, W.; Pei, H.; Li, J.; Sun, Z.-Y. Highly Accurate Prediction of Viscosity of Epoxy Resin and Diluent at Various Temperatures Utilizing Machine Learning. *Polymer* **2022**, *256*, 125216.

(16) Landrum, G., et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*, 31.

(17) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminformatics* **2018**, *10*, 4.

(18) Yang, J.; Tao, L.; He, J.; McCutcheon, J. R.; Li, Y. Machine Learning Enables Interpretable Discovery of Innovative Polymers for Gas Separation Membranes. *Sci. Adv.* **2022**, *8*, eabn9545.

(19) Xu, X.; Zhao, W.; Hu, Y.; Wang, L.; Lin, J.; Qi, H.; Du, L. Discovery of Thermosetting Polymers with Low Hygroscopicity, Low Thermal Expansivity, and High Modulus by Machine Learning. *J. Mater. Chem. A* **2023**, 10.1039.D2TA09272G.

(20) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *13*, 2224–2232.

(21) Lee, C.-K.; Lu, C.; Yu, Y.; Sun, Q.; Hsieh, C.-Y.; Zhang, S.; Liu, Q.; Shi, L. Transfer Learning with Graph Neural Networks for Optoelectronic Properties of Conjugated Oligomers. *J. Chem. Phys.* **2021**, *154*, 024906.

(22) Qiu, H.; Qiu, X.; Dai, X.; Sun, Z.-Y. Design of Polyimides with Targeted Glass Transition Temperature Using a Graph Neural Network. *J. Mater. Chem. C* **2023**, *11*, 2930–2940.

(23) Queen, O.; McCarver, G. A.; Thatigotla, S.; Abolins, B. P.; Brown, C. L.; Maroulas, V.; Vogiatzis, K. D. Polymer Graph Neural Networks for Multitask Property Learning. *npj Comput. Mater.* **2023**, *9*, 90.

(24) Qiu, H.; Wang, J.; Qiu, X.; Dai, X.; Sun, Z.-Y. Heat-Resistant Polymer Discovery by Utilizing Interpretable Graph Neural Network with Small Data. *Macromolecules* **2024**, *57*, 3515–3528.

(25) Ohno, M.; Hayashi, Y.; Zhang, Q.; Kaneko, Y.; Yoshida, R. SMiPoly: Generation of a Synthesizable Polymer Virtual Library Using Rule-Based Polymerization Reactions. *J. Chem. Inf. Model.* **2023**, *63*, 5539–5548.

(26) Yue T, L. Y., He J PolyUniverse: Generation of a Large-scale Polymer Library Using Rule-Based Polymerization Reactions for Polymer Informatics. *ChemRxiv* **2024**,

(27) Qiu, H.; Sun, Z.-Y. On-Demand Reverse Design of Polymers with PolyTAO. *npj Comput. Mater.* **2024**, *10*, 273.

(28) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(29) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 2294–2307.

(30) Gakh, A. A.; Burnett, M. N. Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules. *J. Chem. Inf. Comput. Sci* **2001**, *41*, 1494–1499.

(31) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*.

(32) Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: A Transformer-based Language Model for Polymer Property Predictions. *npj Comput. Mater.* **2023**, *9*, 64.

(33) Kuenneth, C.; Ramprasad, R. polyBERT: A Chemical Language Model to Enable Fully Machine-Driven Ultrafast Polymer Informatics. *Nat. Commun.* **2023**, *14*, 4099.

(34) White, A. D. The Future of Chemistry Is Language. *Nat. Rev. Chem.* **2023**, *7*, 457–458.

(35) 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.

(36) Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684–4690.

(37) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523–1531.

(38) Lin, T.-S.; Rebello, N. J.; Lee, G.-H.; Morris, M. A.; Olsen, B. D. Canonicalizing BigSMILES for Polymers with Defined Backbones. *ACS Polymers Au* **2022**, *2*, 486–500.

(39) Zou, W.; Martell Monterroza, A.; Yao, Y.; Millik, S. C.; Cencer, M. M.; Rebello, N. J.; Beech, H. K.; Morris, M. A.; Lin, T.-S.; Castano, C. S.; Kalow, J. A.; Craig, S. L.; Nelson, A.; Moore, J. S.; Olsen, B. D. Extending BigSMILES to non-covalent bonds in supramolecular polymer assemblies. *Chem. Sci.* **2022**, *13*, 12045–12055.

(40) Yan, C.; Feng, X.; Wick, C.; Peters, A.; Li, G. Machine learning assisted discovery of new thermoset shape memory polymers based on a small training dataset. *Polymer* **2021**, *214*, 123351.

(41) Schneider, L.; Walsh, D.; Olsen, B.; De Pablo, J. J. Generative BigSMILES: An Extension for Polymer Informatics, Computer Simulations & ML/AI. *Digit. Discov.* **2023**, 10.1039.D3DD00147D.

(42) Walsh, D. J.; Zou, W.; Schneider, L.; Mello, R.; Deagen, M. E.; Mysona, J.; Lin, T.-S.; de Pablo, J. J.; Jensen, K. F.; Audus, D. J.; Olsen, B. D. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. *ACS Cent. Sci.* **2023**, *9*, 330–338.

(43) Jiang, S.; Dieng, A. B.; Webb, M. A. Property-Guided Generation of Complex Polymer Topologies Using Variational Autoencoders. *npj Comput. Mater.* **2024**, *10*, 139.

(44) Miccio, L. A.; Schwartz, G. A. From Chemical Structure to Quantitative Polymer Properties Prediction through Convolutional Neural Networks. *Polymer* **2020**, *193*, 122341.

(45) Nguyen, T.; Bavarian, M. A Machine Learning Framework for Predicting the Glass Transition Temperature of Homopolymers. *Ind. Eng. Chem. Res.* **2022**, *61*, 12690–12698.

(46) Chen, G.; Tao, L.; Li, Y. Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. *Polymers* **2021**, *13*.

(47) Goswami, S.; Ghosh, R.; Neog, A.; Das, B. Deep learning based approach for prediction of glass transition temperature in polymers. *Materials Today: Proceedings* **2021**, *46*, 5838–5843, International Conference on Advances in Materials Science, Communication and Microelectronics.

(48) Phan, B. K.; Shen, K.-H.; Gurnani, R.; Tran, H.; Lively, R.; Ramprasad, R. Gas Permeability, Diffusivity, and Solubility in Polymers: Simulation-experiment Data Fusion and Multi-Task Machine Learning. *npj Comput. Mater.* **2024**, *10*, 186.

(49) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59*, 5013–5025.

(50) Deagen, M. E.; Dalle-Cort, B.; Rebello, N. J.; Lin, T.-S.; Walsh, D. J.; Olsen, B. D. Machine Translation between BigSMILES Line Notation and Chemical Structure Diagrams. *Macromolecules* **2024**, *57*, 42–53.

(51) Choi, S.; Lee, J.; Seo, J.; Han, S. W.; Lee, S. H.; Seo, J.-H.; Seok, J. Automated BigSMILES Conversion Workflow and Dataset for Homopolymeric Macromolecules. *Sci. Data* **2024**, *11*, 371.

(52) Afzal, M. A. F.; Browning, A. R.; Goldberg, A.; Halls, M. D.; Gavartin, J. L.; Morisato, T.; Hughes, T.; Giesen, D. J.; Goose, J. E. High-Throughput Molecular Dynamics Simulations and Validation of Thermophysical Properties of Polymers for Various Applications. *ACS Appl. Polym. Mater.* **2020**, *3*, 620–630.

(53) Kuenneth, C.; Rajan, A. C.; Tran, H.; Chen, L.; Kim, C.; Ramprasad, R. Polymer Informatics with Multi-Task Learning. *Patterns* **2021**, *2*, 100238.

(54) Kamal, D.; Tran, H.; Kim, C.; Wang, Y.; Chen, L.; Cao, Y.; Joseph, V. R.; Ramprasad, R. Novel high voltage polymer insulators using computational and data-driven techniques. *J. Chem. Phys.* **2021**, *154*, 174906.

(55) Phan, B. K.; Shen, K.-H.; Gurnani, R.; Tran, H.; Lively, R.; Ramprasad, R. Gas permeability, diffusivity, and solubility in polymers: Simulation-experiment data fusion and multi-task machine learning. 2024; `https://arxiv.org/abs/2406.14809`.