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

The inherent randomness of polymers has long posed challenges for their representation learning in polymer machine learning (ML). The Simplified Molecular-Input Line-Entry System (SMILES) notation, which has excelled in small molecule research, unfortunately, struggles to flexibly capture the complexity of polymer structures, such as random block copolymers. Recently, BigSMILES and its extensions have paved the way for more accurate descriptions of polymer structures. However, whether BigSMILES outperforms SMILES in polymer ML workflows has yet to be systematically explored and demonstrated. To fill this scientific gap, we conducted extensive experiments investigating this question, encompassing a variety of polymer property prediction and inverse design tasks based on both image and text inputs. Our findings reveal that in 11 tasks involving homopolymer systems, BigSMILES-based ML workflows exhibit

performance comparable to or even exceeding that of SMILES, underscoring the utility of BigSMILES in representing polymer structures. Furthermore, BigSMILES offers a more compact textual representation compared to SMILES, significantly reducing the computational cost of model training, particularly for large language models. Through these comprehensive experiments, we demonstrate that BigSMILES can achieve performance on par with SMILES, while also facilitating faster model training and reducing energy consumption, which could have a substantial impact on a wide range of polymer tasks in the future, including property prediction (and classification) and polymer generation across various polymer types.

Introduction

Machine learning (ML) methods have proven their efficiency and effectiveness in advancing grand sustainable goals, such as accelerating molecular and materials discovery.[1] In recent years, a series of studies have focused on representation learning of materials to achieve higher accuracy in ML predictions.[2–8] Initially, scientists predominantly utilized descriptors calculated from cheminformatics tools, such as RDKit[9] and Mordred.[10] These numerical descriptors can be directly used as inputs for conventional ML models like Random Forest and Gaussian Process.[11,12] However, these descriptors are often atom- or bond-specific, making it difficult to capture both short-range and long-range interactions within molecules. Consequently, graph neural networks (GNNs) have garnered increasing attention in the molecular field.[13–17] Through graph convolution operations, GNNs extend the model's learning from the atomic and bond levels to the functional group level, extracting functional group-level descriptors, thereby achieving high-dimensional molecular representations.

Yet, with the exponential growth of molecular data,[18–20] descriptor-based and graph-based models have revealed potential storage issues, as representing a single molecule often requires hundreds or even thousands of descriptors, which could result in a breakdown in access to and storage of molecular data, especially in mobile personal computers. As a remedy,

textual representations of molecules can perfectly distinguish and record different molecular structures with minimal memory requirements. The textual representation of molecules is not a recent innovation. As early as 1988, Weininger ingeniously invented the Simplified Molecular-Input Line-Entry System (SMILES)[21] to represent molecular structures, which has since been widely adopted by the scientific community. There emerged other textual notations including the SYBYL Line Notation (SLN),[22] the Modular Chemical Descriptor Language (MCDL),[23] and the International Chemical Identifier (InChI).[24] Benefiting from the recent surge in text processing capabilities powered by Transformer-based models, text-based, especially SMILES-based, molecular property prediction models have received unprecedented attention and development.[25–28] The success of these models demonstrates the potential to extract high-dimensional, complex chemical information from textual representations.

To extend the functionality of molecular textual representations, such as SMILES, to the polymer domain, Ma et al. introduced the use of "*" symbols in SMILES to denote polymerization points within polymers, leading to the development of Polymer-SMILES (p-SMILES).[29] This concept has been widely adopted by polymer scientists. To more effectively capture the randomness and complexity of polymers, Lin et al. proposed BigSMILES,[30] a comprehensive and systematic representation method capable of describing any polymer structure, overcoming the limitations of other text-based polymer representations. Due to its powerful features, BigSMILES has been highly anticipated by the polymer community[31–34] and has become the default representation method in polymer databases such as the recently Community Resource for Innovation in Polymer Technology (CRIPT).[35] However, the performance of BigSMILES in polymer ML remains underexplored, leaving a gap between its theoretically powerful capabilities and its practical applications.

In this work, we aim to benchmark the performance of BigSMILES in polymer ML pipelines, comparing its relative advantages or potential disadvantages to SMILES (or p-SMILES). Since current cheminformatics tools have yet to integrate BigSMILES, it is not

possible to compute descriptors using BigSMILES, and therefore, we are unable to include descriptor-based tasks in this comparison. Our study explores typical input types including binary images induced by BigSMILES and using BigSMILES strings. We evaluated these in convolutional neural networks (CNNs), deep neural networks, and large language models (LLMs), focusing on polymer property prediction and molecular generation tasks.

Our extensive results demonstrate that BigSMILES encapsulates sufficient polymer chemistry information, enabling ML models based on BigSMILES to achieve prediction accuracy that matches or even surpasses that of SMILES-based models. Surprisingly, in tasks involving LLMs, BigSMILES led to faster training speeds, which is a welcome benefit given the ever-growing volume of polymer data. By utilizing BigSMILES, we can complete polymer modeling tasks in shorter times and with lower energy consumption. This work bridges the knowledge gap in the polymer community regarding the performance of BigSMILES in ML tasks, providing objective evidence of BigSMILES' role in polymer ML. As polymer data continues to proliferate, the reduced memory usage and faster training speeds offered by BigSMILES are likely to bring significant value to a wide range of polymer tasks.

Results and discussion

Conventional ML models or deep learning models are typically used to handle tabular and numerical data. For character-based data like SMILES and BigSMILES, there are currently two primary approaches to directly learn polymer chemical information from these strings (as illustrated in Figure 1). The first approach involves converting the characters into binary images, from which CNNs can extract chemical knowledge.[36,37] The second approach employs deep learning models that can process sequence and text directly, such as recurrent neural networks (RNNs) and long short-term memory (LSTM),[38,39] and the recently leading and rapidly evolving Transformer and LLMs.[25,26,28]

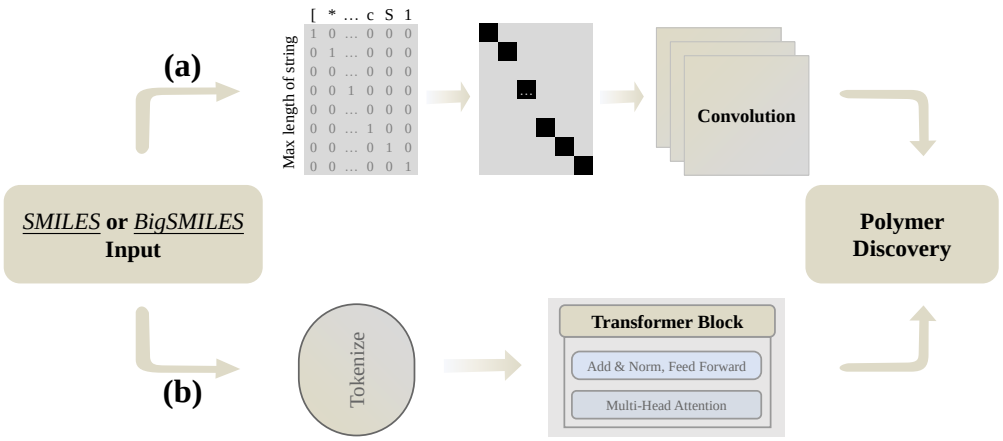
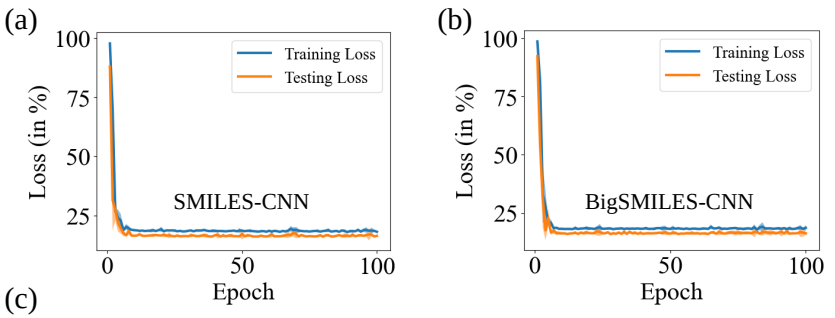


Figure 1: Two approaches that textual polymer representation can be used as ML inputs. (a) Textual polymer representation is first transformed to images and then learned by CNNs. (b) Textual polymer representation is directly served as input of LLMs after tokenization.

Regarding the first approach, we used binary images induced by SMILES and BigSMILES as inputs to a CNN model to predict the glass transition temperature of various polymers. The data and model parameters were adopted from this reference.[38] After 100 training epochs, ML models based on both types of string-based polymer representations converged. To eliminate randomness, we repeated each configuration five times, with the results presented in Figure 2. The loss function is the relative absolute error(RAE), which is the absolute value of the percentage error between the predicted value and the true value relative to the true value. After the same training, the SMILES-CNN achieved a RAE of 16.46±0.12% on the test set, while the BigSMILES-CNN performed a comparable RAE of 16.57±0.28%. This finding is more intuitively illustrated in Figure 2(c), where we present several random structures and the prediction results from models based on the two representations (with the predictions closer to the true values highlighted in bold).



(c)

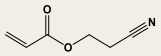
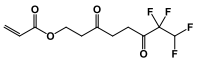
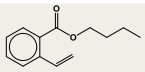
Polymer	Referred Tg/K	Predicted Tg/K	
		SMILES-CNN	BigSMILES-CNN
	277	345.38	**278.03**
	233	309.34	**274.46**
	339	**315.38**	278.03

Figure 2: 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).

Next, we explored the application of these two text-based polymer representations in LLMs, as LLMs can directly take text-based data as input. We fine-tuned PolyNC, an end-to-end polymer LLM, on nine polymer tasks. During the fine-tuning phase, the model input was either SMILES or BigSMILES, and the output was the corresponding property value. The nine polymer tasks included atomization energy (AE), bandgap of polymer chains (BG), bandgap of polymer crystals (BGC), charge injection barrier (CIB), crystallization tendency (CT), electron affinity (EA), ionization energy (IE), CO2 permeability in polymer membranes (CO2), and glass transition temperature of polyimides (Tg). The distribution of each dataset is shown in Figure 3. As can be seen, all datasets exhibit a fairly pronounced normal distribution, indicating that these datasets have well-distributed data, which is beneficial for subsequent ML modeling.

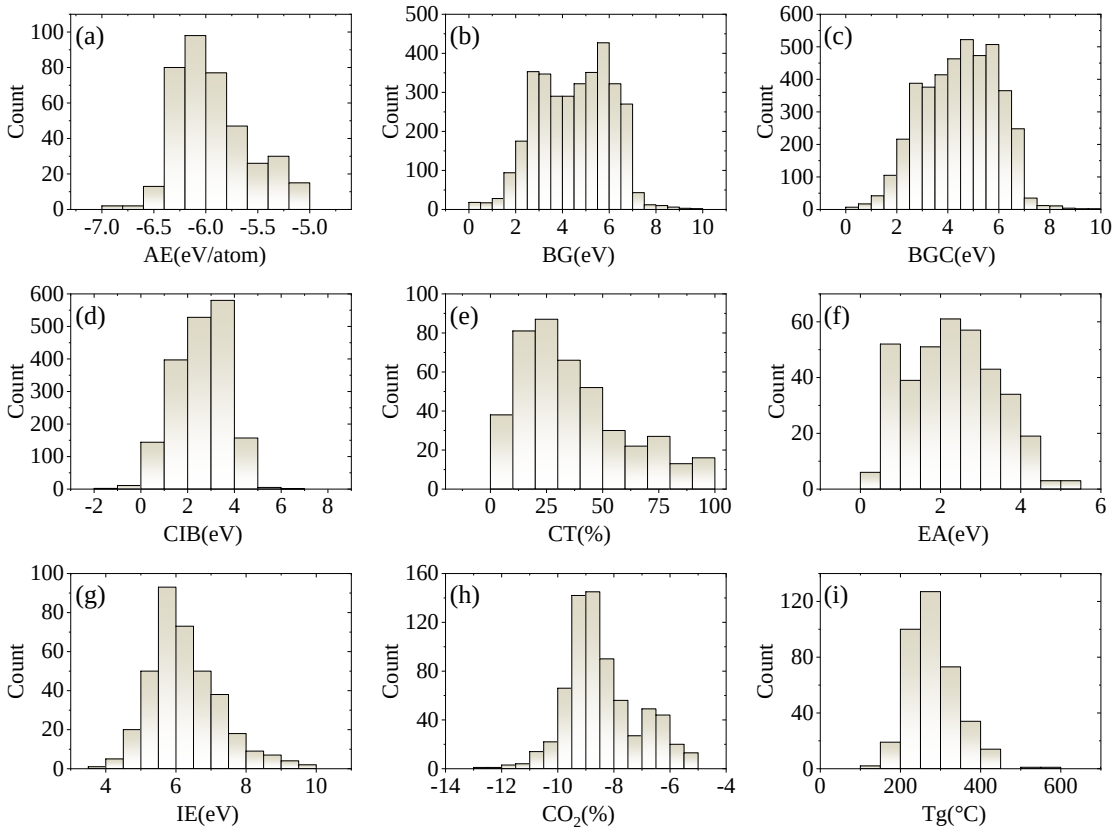


Figure 3: Data distribution. These datasets exhibit a fairly pronounced normal distribution.

After fine-tuning with the same configurations (fine-tuning details are provided in the Methods section), the performance of the models based on the two textual representations is shown in Figure 4. We used the mean absolute error (MAE) of the predictions on the test set as the evaluation metric (Figure 4(a)). It can be observed that models using BigSMILES as the representation method exhibited performance comparable to those using SMILES, with slightly higher MAEs across various tasks. This may be influenced by the pretraining of PolyNC, as SMILES was used as the structural representation for polymers during the pretraining phase.

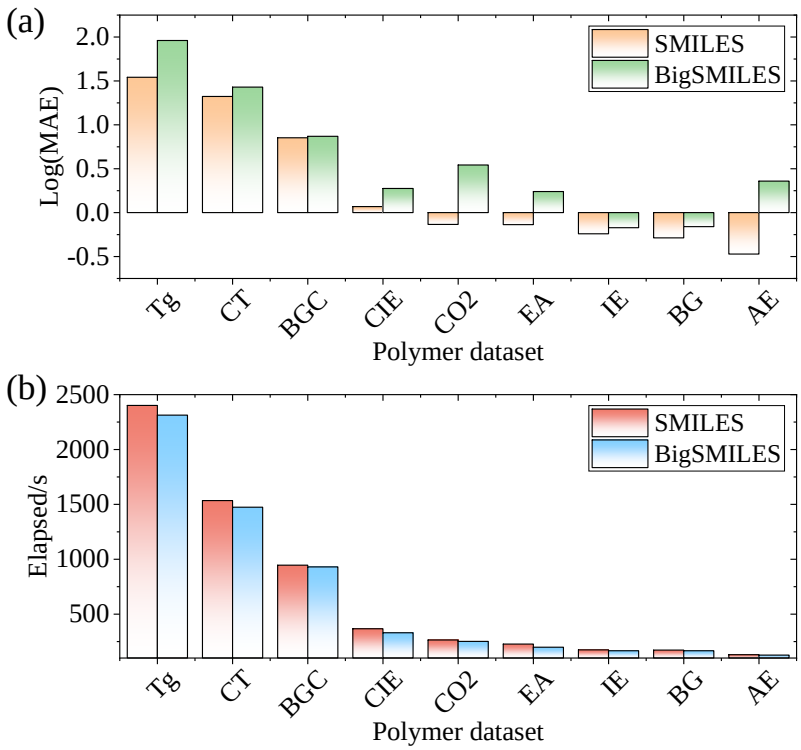
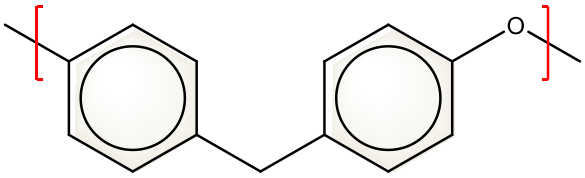


Figure 4: Performance of the two representation methods in fine-tuning PolyNC: (a) shows the model's MAE, and (b) displays the time taken for model fine-tuning (in seconds).

Remarkably, it is noteworthy that when using BigSMILES as polymer representation, the time required for fine-tuning the model was consistently shorter than that of SMILES. By analyzing the datasets corresponding to SMILES and BigSMILES, we found that BigSMILES can represent polymer structures with fewer tokens. For example, as shown in Figure 5, after encoding by PolyNC's encoder, the SMILES-based representation required 27 tokens, whereas the BigSMILES-based representation needed only 24 tokens. The reason lies in BigSMILES' ability to convey polymer connectivity information with fewer tokens (highlighted in red in the figure). This ability and superiority are especially significant for large datasets, where shorter training times translate to lower energy consumption and faster model iteration.



Item	SMILES	BigSMILES
Representation	[*]Oc1ccc(Cc2ccc([*])cc2)cc1	{<Oc1ccc(cc1)Cc2ccc(cc2)>}
Tokens	'▁[', '*', ']', 'O', 'c', '1', 'c', '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 5: Encoding details of SMILES and BigSMILES using PolyNC's encoder. BigSMILES represents polymer structures with fewer tokens.

We also explored the performance of BigSMILES in polymer generation tasks. The training details of the model were consistent with our previously trained SMILES-based polymer generation model, PolyTAO,[20] with the only difference being the substitution of SMILES with BigSMILES. Using data from PI1M,[29] we trained a polymer generation model based on BigSMILES. However, this model exhibited a lower capability in generating valid BigSMILES structures. This may suggest that while BigSMILES represents polymerization sites with fewer tokens, the chemical information conveyed by this representation is weaker than that of p-SMILES. Further research is required to address this, including but not limited to refining the syntax of BigSMILES and developing a custom tokenizer specifically for BigSMILES.

Discussion and Conclusion

In this concise and timely study, we systematically explored the performance of SMILES and BigSMILES, two polymer representations, across multiple polymer ML tasks. We found that

BigSMILES exhibits performance comparable to SMILES. Although BigSMILES slightly underperforms SMILES in some tasks, its ability to succinctly describe complex polymer structures is a distinct advantage that SMILES lacks. Future efforts should focus on refining the syntax rules of BigSMILES to enhance its richness in polymer chemical information. As polymer ML ventures into uncharted territories, the limitations of SMILES in representing polymer structures increasingly constrain its utility and scope. Consequently, polymer ML pipelines based on BigSMILES are likely to attract greater attention and adoption among polymer scientists, with BigSMILES' functionality progressively improving—especially through its integration into cheminformatics tools for descriptor computation.

Another interesting finding of this study is that polymer ML workflows based on BigSMILES consistently required shorter training times compared to those based on SMILES, particularly in large language model scenarios. This advantage stems from the streamlined syntax of BigSMILES. As more polymers are discovered and virtually designed, the datasets for polymer ML training will continue to grow. Recent estimates by Li et al. suggest that the candidate space for polyimides alone could reach nearly $2*10^{12}$ compounds. Therefore, using BigSMILES as a representation could significantly accelerate the construction of polymer ML pipelines, whether in forward screening paradigms or inverse design paradigms. We also adapted the SMILES-based polymer generation model PolyTAO [20] and trained the first polymer generation model based on BigSMILES, but further improvements are needed for the model to better learn the BigSMILES syntax and generate more valid representations. The current model is available on Hugging Face (`https://huggingface.co/hkqiu/PolyTAO-BigSMILES_Version`).

Methods

Batch Conversion to BigSMILES

At present, local chemical structure drawing tools do not yet support direct extraction of BigSMILES. Fortunately, research groups led by Prof. Olsen and Prof. Seok have recently developed tools to interconvert molecular structure/SMILES and BigSMILES,[40,41] enabling the acquisition of the millions of BigSMILES entries involved in this work. `BigSMILES\_homopolymer`[41] can convert SMILES of homopolymers to BigSMILES, while for other polymer structures, the structure-to-BigSMILES tool developed by Olsen et al. was utilized.[40]

Text-Induced Image Convolutional Neural Network

Here, we employed the optimal network parameters reported in this literature,[36] comprising 2 convolutional layers, a fully-connected layer with 100 neurons, and convolutional and pooling kernels of size (3, 3). The first convolutional layer had 256 kernels, while the second had 128 (see ref.[36] for details). The shape of image is $w * h$, where w denotes the length of string list and h denotes the max length of these strings. The *dataset1* from this previous work, a dataset of glass transition temperatures for polystyrenes and polyacrylates, was used for training. The training-to-test split ratio was 0.8:0.2. The implementation was carried out using PyTorch (version 1.12.1+cu113).

Large Language Model Fine-tuning

There are many excellent pre-trained polymer language models available, such as TransPolymer,[25] polyBERT,[26] and PolyNC.[28] Here, we chose to use PolyNC as the base model, as it is a native end-to-end architecture that is convenient for building polymer text-description-based property prediction models. For each fine-tuning task, we used the polymer text representations (SMILES or BigSMILES) as the model input, and the target property values as the output.

The hyperparameters used for the fine-tuning process are provided in Table 1. This fine-tuning was implemented using 4 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

Data and code availability

The training data of the CNN task can be accessed in this ref.[36] The nine properties of polymers during the fine-tuning of LLMs were collected from these references.[15,17,42–45] The PI1M dataset uesd for training the polymer generation models is publicly available at `https://github.com/RUIMINMA1996/PI1M`.

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

(2) 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.

(3) 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.

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

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

(6) 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.

(7) 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**,

(8) 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.

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

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

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

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

(20) Qiu, H.; Sun, Z.-Y. On-Demand Reverse Design of Polymers with PolyTAO. *ChemRxiv* **2024**,

(21) 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.

(22) 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.

(23) 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.

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

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

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

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

(28) 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.

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

(30) 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.

(31) 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.

(32) 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.

(33) 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.

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

(35) 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 Central Science* **2023**, *9*, 330–338.

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

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

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

(39) 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.

(40) 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.

(41) 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. *Scientific Data* **2024**, *11*, 371.

(42) 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.

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

(44) 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.

(45) 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`.